Artificial Intelligence and Algorithmic Mediations:
Affect, Power, and Subjectivation on Kaggle

Luciano dos Reis Frizzera

A Thesis

In the Department of

Communication Studies

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Communication Studies)

at Concordia University

Montréal, Québec, Canada

June 2024

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Luciano dos Reis Frizzera

Entitled: Artificial Intelligence & Algorithmic Mediations: Affect, Power, and Subjectivation on Kaggle

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Communication Studies)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Thesis Supervisor

Dr. Fenwick McKelvey

______________________________ Examiner

Dr. Alessandra Renzi

______________________________ Examiner

Dr. Owen Chapman

______________________________ Examiner

Dr. Bart Simon

______________________________ External Examiner

Dr. Jennifer Whitson

______________________________ Chair

Dr. Marc Steinberg

Approved by ______________________________

Dr. Krysta Lynes, Graduate Program Director

06/17/2014 ______________________________

Dr. Pascale Sicotte, Dean Faculty of Arts and Science

Abstract

Artificial Intelligence & Algorithmic Mediations: Affect, Power, and Subjectivation on Kaggle

Luciano dos Reis Frizzera, PhD

Concordia University, 2024

Over the past decade, the widespread investment in digital infrastructure and the extensive digitization of individual behaviour have provided the basis for the rapid development of machine-learning techniques and Artificial Intelligence (AI). AI datafy our body and our identity, producing live databases full of calculated linkages between humans and nonhumans. It creates a new cartography of biopower that sometimes produces technologies, but always produces subjects. This research examines the political economy of subjectivation in the "making of" machine-learning algorithms and AI by closely examining the relations of power, affect, and subjectivation on Kaggle, the world's largest data science community. Conceived as a gamified platform for crowdsourced machine-learning challenges, Kaggle is a networked public where users are under constant pressure to produce new and improved algorithms.

This research first engages with Kaggle as a company and platform, offering a narrative of its history and a detailed description of how it works. Combining discourse analysis, software studies, and digital methods, this research aims to understand how code, data, digital infrastructures, crowdsourced labour, and political-economic interests are mobilized to create instruments of control that shape, modulate, and mediate individual behaviour. This phenomenon, which I call modes of automatic subjectivation, points toward the possibility of using subjective and impersonal materials to reorganize life in its broadest sense according to a specific system of power and privileges involving gender, race, sexuality, and social class.

This dissertation argues that these modes of subjectivation are designed to control the "production of possibilities" and reinforce specific types of socioeconomic relations, which in turn reproduce current conditions of existence. Furthermore, this research argues that the data science community has a notable compulsion toward cost reduction, indifference toward human life, an obsession with controlling populations and individual bodies, and a desire to produce a predictable future for economic gain. Ultimately, this research identifies algorithmic media based on AI Technology as a core asset in the attention economy and as a source of power that can be used as an interface to prescribe individual behaviour.

Acknowledgments

The solitude of writing a dissertation is a myth perpetrated by a society based on individualism and meritocracy. The research I bring to life here could not have taken place without the steadfast support and encouragement of countless individuals and institutions. As I stand at the threshold of completion, I am filled with deep gratitude for all those who have played a pivotal role in shaping this project.

First and foremost, I am grateful for the invaluable guidance of my supervisor, Dr. Fenwick McKelvey, who has contributed so much to my academic accomplishments. His unwavering commitment to excellence, boundless patience and kindness, and profound scholarly insights have been my guiding light. His mentorship has nurtured my academic growth and instilled in me a deep passion for rigorous inquiry and intellectual exploration. He has encouraged me to become bold, creative, and flexible with my ideas, and I hope that this project will stand as a testament to his insights.

I am deeply indebted to the members of my dissertation committee, Dr. Alessandra Renzi and Dr. Owen Chapman. From the very inception of this project, they believed in its potential and provided invaluable feedback and scholarly guidance. Their expertise and constructive criticism have significantly enhanced the quality of this work. I extend my sincere thanks to other faculty members who have supported me along the way, notably Dr. Kim Sawchuk, Dr. Peter van Wick, Dr. Sandra Gabriele, Dr. Matt Soar, Dr. Mia Consalvo, Dr. Orit Halpern, and Dr. Brian Massumi. Lastly, I am grateful to all the scholars, authors, and researchers whose work has informed and inspired my own. Their contributions have laid the foundation upon which this dissertation stands.

I also want to express my heartfelt thanks to Concordia University for their unwavering support. They have provided me with the necessary resources, facilities, and academic environment that have been instrumental in my research and intellectual exploration. I am grateful for the opportunities made available to me through the various teaching and research assistantships, which have all contributed to the advancement of my project. In particular, I would like to thank Dr. Bart Simon and Dr. Fenwick McKelvey for providing creative environments for research experimentations at the Machine Agencies and the Algorithmic Media Observatory. I am also grateful to Dr. Kim Sawchuk and Dr. Owen Chapman for the opportunity to work with different research perspectives at the beginning of my journey at the Participatory Media Cluster, and to Dr. Ann-Louise Davidson and Dr. Giuliana Cucinelli for recognizing and supporting my skills in the

Department of Education. Many thanks to all my colleagues, coworkers, and research partners, especially Julia Sales, Julia Zamboni, Ryan Scheiding, Sam Thulin, Ashley McAskill, and Ceyda Yolgörmez. The countless hours spent in the lab and classroom were important to our project's progress and invaluable to developing our friendship.

I am thankful for the encouragement and support of my cohort and other colleagues and friends I made along this journey, particularly Tricia Toso and Myriam Rafla, with whom I shared many of the challenges of bringing this project to fruition. These Canadian souls welcomed me into their lives, gave me emotional and professional support, and assisted me in every possible way. Tricia not only revised the early versions of this manuscript but also helped me find my voice in it. I extend my gratitude to Antonia Hernández, Chery Smilley, Sarah Ganzon, Andrei Zanescu, Marc Lajeunesse, Lai-Tze Fan, Marilyn Sugiarto, Scott Dejong, Dinara Moura, and Abhisekh Patra for the wonderful insights and the opportunities to share ideas and knowledge they provided while I worked on this project. I cannot imagine what my PhD experience would have been like without them, and I will cherish the times we had and the memories we made. I am forever grateful.

Living abroad is always challenging. I am very thankful to my Brazilian friends in Canada, especially Julia Salles, Cristiano Therrien, Priscylla Joca, Francisco Loiola, Tucaauê Alvarez, and Januana Teixeira, who made me feel closer to home.

I am deeply grateful for the support my family has given me throughout this journey. Their encouragement, understanding, and unwavering belief in my abilities have sustained me throughout this journey. Their love and encouragement have been my source of strength during the triumphs and challenges I encountered. My parents, Ruth Reis and Kleber Frizzera, assisted me emotionally, financially, and academically—I am following in your footsteps. My brothers, Gabriel and Erico Frizzera, pushed me to go through the pain and encouraged me not to give up. And to my beloved wife, Sâmia Pedraça, who not only persuaded me to come here in the first place, but also followed me to the coldest place on Earth to make this project happen: thank you for supporting my writing process in all conceivable ways, listening to my ideas, and understanding the journey. I love you.

Table of Contents

LIST OF FIGURES ..... **IX**

LIST OF TABLES ..... **X**

1. INTRODUCTION **1**

1.1. DEFINITIONS 7
1.2. CHAPTER SUMMARIES 12

2. THEORETICAL FRAMEWORK **17**

2.1. KAGGLE AS A PLATFORM 19
2.2. DATAFIED POLITICAL ECONOMY 23
2.2.1. Datafication and Platformization *26*
2.2.2. Surveillance Capitalism and Modes of Living *29*
2.3. AUTOMATED SUBJECTIVATION 31
2.3.1. From Biopolitics to Psychopolitics *34*
2.3.2. Impersonal Subjectivation *36*
2.4. MEDIATIONS 40
2.4.1. Radical Mediation *42*
2.5. SUMMARY 45

3. METHODOLOGY **47**

3.1. TIMEFRAME 49
3.2. SOURCE MATERIAL 50
3.3. DATA GATHERING 51
3.3.1. Digital Methods *52*
3.3.1.1. Web Scraping 54
3.3.1.2. Proceedings 56
3.3.2. Qualitative Content Analysis *57*
3.3.2.1. Proceedings 59
3.4. ANALYTICAL FRAMEWORK 59
3.4.1. Software Studies *61*
3.4.2. Discourse Analysis *64*
3.5. POSITIONALITY 67
3.6. SUMMARY 69

4. ON KAGGLE: FROM DISRUPTIVE START-UP TO BIG TECH SUBSIDIARY **70**

4.1. THE COMPANY 73
4.1.1. Proof-of-concept (2010) *75*
4.1.2. Silicon Valley Start-up (2011–2016) *77*
4.1.3. Alphabet's Subsidiary (2017 and beyond) *80*
4.1.4. Controversies *82*
4.2. THE PLATFORM 83
4.2.1. Prototype *85*
4.2.2. Hackathon Organizer *87*
4.2.3. Temp Agency *89*
4.2.4. Social Media Platform *92*
4.2.5. Google Cloud Service *94*
4.3. STRATEGIC VISION AND BUSINESS MODEL 95
4.4. SUMMARY 104

5. THE MACHINE-LEARNING GAME: INFRASTRUCTURE, COMMUNITY, AND COMPETITIONS **107**

5.1. REPOSITORIES: CODE AND DATA FOR MACHINE LEARNING 110

5.1.1. Notebooks: Code Repository ... *110*
5.1.2. Dataset: Data Repository ... *111*
5.2. KAGGLERS: A COMMUNITY OF DATA SCIENTISTS ... 114
5.2.1. User Rankings ... *120*
5.2.1.1. User Ranking: Competitions ... 122
5.2.1.2. User Ranking: Datasets ... 124
5.2.1.3. User Ranking: Notebooks ... 125
5.2.1.4. User Ranking: Discussion ... 126
5.3. COMPETITIONS: GAMIFIED PREDICTIVE MODEL DEVELOPMENT ... 127
5.3.1. Simulacra and winner-take-all ... *132*
5.3.2. Machine-Learning Modalities ... *134*
5.4. SUMMARY ... 139

6. IDENTIFY: DEEPFAKES AND SHALLOW TRUTHS ... 142

6.1. SENSE: FACIAL RECOGNITION ... 145
6.1.1. Seeing is Believing ... *148*
6.1.2. Genuinely Fake ... *150*
6.2. MOBILIZE: DEEPFAKE DETECTION CHALLENGE ... 153
6.2.1. Gamified Community ... *155*
6.2.2. Deep Sociotechnical Tendencies ... *158*
6.2.3. Fluid Ethics ... *162*
6.3. MODULATE: THE TRUE SELF ... 166
6.3.1. More and Deeper Fakes ... *168*
6.3.2. Uncanny Self ... *170*
6.3.3. Ontogenesis ... *172*
6.4. SUMMARY ... 174

7. PREDICT: AB/NORMAL BODIES AND SPECULATIVE FUTURES ... 177

7.1. SENSE: AB/NORMAL BODIES ... 181
7.1.1. Virtual Strip Search ... *184*
7.1.2. Preemptive Anticipation ... *187*
7.2. MOBILIZE: PASSENGER SCREENING ALGORITHM CHALLENGE ... 189
7.2.1. Thin Human Slices ... *190*
7.2.2. Discriminatory Compromises ... *192*
7.2.3. Mislabeled Threats and Lack of Bodies ... *197*
7.3. MODULATE: SPURIOUS CORRELATIONS ... 201
7.3.1. The Fantasy of Epistemic Purity ... *204*
7.3.2. Automatic Projections of the Self ... *206*
7.3.3. Self-Fulfilling Prophecies ... *208*
7.4. SUMMARY ... 210

8. NUDGE: RADICAL RECOMMENDATIONS ... 214

8.1. SENSE: BEHAVIOUR DATA ... 219
8.1.1. Individual Signature and Hyper-Profiles ... *221*
8.1.2. Hyper Nudges ... *223*
8.2. MOBILIZE: INSTACART MARKET BASKET ANALYSIS ... 225
8.2.1. Bananas and Pizzas: You Are What You Eat ... *228*
8.2.2. Finding the Sweet Spot ... *231*
8.2.3. Standing on the Shoulders of Crowdsourced Free Labour ... *235*
8.3. MODULATE: SHAPING THE SELF ... 239
8.3.1. Habitual Media ... *241*
8.3.2. Radical Subjectivation ... *243*
8.3.3. Remote Control ... *246*
8.4. SUMMARY ... 248

9. CONCLUSION ... 253

9.1. GAMIFIED FREE LABOUR 256
9.2. DATAFICATION AND THE INDIFFERENCE TOWARD HUMAN LIFE 259
9.3. ALGORITHMIC SUBJECTIVATION 262
9.4. ALGORITHMIC MEDIATIONS AND THE PRODUCTION OF POSSIBILITIES 265
9.5. LIMITATIONS 268
9.6. FUTURE RESEARCH 270

10. APPENDIX I: BUILDING A WEB SCRAPER 274

10.1. SCRAPER I: COMPETITIONS, DATASETS, AND USER RANKING 276
10.1.1. Competition 276
10.1.2. Datasets 278
10.1.3. User Ranking 279
10.2. SCRAPER II: COMPETITION'S INTERNAL PAGES 280
10.3. SCRAPER III: HOMEPAGE SCREENSHOTS 282
10.4. BRIEF CONSIDERATION ON SECURITY AND PRIVACY 282

BIBLIOGRAPHY 283

List of Figures

FIGURE 1.1: PHASES OF AI DEVELOPMENT ON KAGGLE. 11
FIGURE 4.1: KAGGLE HISTORY TIMELINE. 74
FIGURE 4.2: KAGGLE TIMELINE HEATMAP 84
FIGURE 4.3: KAGGLE HOMEPAGE IN 2010. 86
FIGURE 4.4: KAGGLE HOMEPAGE IN 2010 AND 2011 88
FIGURE 4.5: KAGGLE HOMEPAGE BETWEEN 2013 AND 2015. 91
FIGURE 4.6: KAGGLE HOMEPAGE IN 2016. 93
FIGURE 4.7: KAGGLE HOMEPAGE IN 2019 93
FIGURE 5.1: KAGGLE NOTEBOOKS. 110
FIGURE 5.2: KAGGLE DATASETS. 112
FIGURE 5.3: DATASETS ON KAGGLE 113
FIGURE 5.4: TOP DATASETS & CONTRIBUTORS ON KAGGLE. 114
FIGURE 5.5: NUMBER OF USERS ON KAGGLE. 115
FIGURE 5.6: USER GENDER DISTRIBUTION. 116
FIGURE 5.7: USER GEOGRAPHIC DISTRIBUTION. 117
FIGURE 5.8: USER EDUCATION DISTRIBUTION 118
FIGURE 5.9: ILLUSTRATION FOR A KAGGLE RPG. 119
FIGURE 5.10: USER SALARY RANGE. 120
FIGURE 5.11: KAGGLE USER RANKINGS. 121
FIGURE 5.12: KAGGLE COMPETITIONS 128
FIGURE 5.13: KAGGLE TOP COMPETITIONS. 130
FIGURE 5.14: PROGRESS IN THE MAPPING DARK MATTER COMPETITION. 134
FIGURE 5.16: KAGGLE COMPETITIONS MODALITY PRIZE DISTRIBUTION 138
FIGURE 5.15: KAGGLE COMPETITIONS MODALITY TIMELINE. 138
FIGURE 6.1: FACIAL MEASUREMENT. 147
FIGURE 6.2: DEEP LEARNING DIAGRAM. 148
FIGURE 6.3: DEEPFAKE DETECTION CHALLENGE TIMELINE 154
FIGURE 6.4: DEEPFAKE SAMPLE. 155
FIGURE 7.1: SCHEMATIC OF A BODY MEASUREMENT DETECTOR. 183
FIGURE 7.2: FULL BODY 3D SCANNER OUTPUTS REAL BODIES 185
FIGURE 7.3: FULL BODY 3D SCANNER OUTPUTS ILLUSTRATION 186
FIGURE 7.4: PASSENGER SCREENING ALGORITHM CHALLENGE TIMELINE 189
FIGURE 7.6: BODY SECTIONS DIAGRAM. 191
FIGURE 7.5: 3D RENDERING OF FULL BODY SCANNER DATA 191
FIGURE 8.1: INSTACART'S FOUR SIDED MARKETPLACE 226
FIGURE 8.2: INSTACART DATA SAMPLE. 226
FIGURE 8.3: INSTACART MARKET BASKET ANALYSIS TIMELINE. 227
FIGURE 8.4: FOOD PURCHASE DISTRIBUTION 228
FIGURE 8.5: SUPERMARKET SECTOR DISTRIBUTION. 230
FIGURE 10.1: KAGGLE COMPETITIONS PAGE. 277
FIGURE 10.2: KAGGLE COMPETITION DETAIL PAGE. 277
FIGURE 10.3: KAGGLE DATASETS PAGE. 278
FIGURE 10.4: KAGGLE USER RANKING PAGE. 278
FIGURE 10.5: SCRAPPER LOGS 280
FIGURE 10.6: KAGGLE COMPETITION DISCUSSION PAGE. 281
FIGURE 10.7: KAGGLE COMPETITION DISCUSSION'S THREAD PAGE. 281

List of Tables

TABLE 5.1: COMPETITION MEDALS 123
TABLE 5.2: COMPETITIONS RANKING 123
TABLE 5.3: DATASETS RANKING 124
TABLE 5.4: NOTEBOOKS RANKING 125
TABLE 5.5: DISCUSSION RANKING 127

1. Introduction

> [AI] is fast-tracking the commodification of the human spirit by mechanising the imagination. It renders our participation in the act of creation as valueless and unnecessary.
>
> — Nick Cave, “Letter About ChatGPT and Human Creativity,” 2023

In 2011, the Heritage Provider Network (HPN), a California healthcare insurance company, launched a multi-year competition to develop a predictive algorithm to help identify patients most likely to be admitted to healthcare providers. The competition attracted over 2,000 machine-learning developers and distributed US$ 3 million in prizes. While the challenge did not achieve its primary objective—fixing America’s healthcare system (Heritage Provider Network, 2011)—it gave the company deeper insight and more control over its patients’ records. In 2017, the US Department of Homeland Security (DHS) promoted the “Passenger Screening Algorithm Challenge,” offering US$ 1.5 million to improve the accuracy of full-body scanners used in airport terminal security. In six months, the DHS attracted more than 11,000 participants, including the world’s top data scientists. While the competition was considered a success, it was not without pitfalls and did not produce reliable results, including finding bombs in passengers’ hair. In 2019, after Samantha Cole’s (2018) revelatory report on the viral popularity of deepfake and non-consensual pornography, Facebook sponsored a Deepfake Detection Challenge, offering US$ 1 million in prizes to improve the detection of fake content on social media. While thousands of participants submitted 35,000 predictive models, the winning solution was considered no better than a coin toss. Moreover, this competition might have helped accelerate improvements to deepfake methods, which have become more common and abusive.

All these competitions promised to *solve* a social problem through collaborative crowdsourcing. All exemplify the kinds of competitions on Kaggle, the world’s largest data science community. Kaggle is a digital platform for machine-learning competitions aiming to “solve real-world problems” across an array of industries, including life sciences, financial services, security, information technology, and retail. These examples illustrate why Kaggle is not just big but important for understanding the contemporary political economy in the production and legitimation of new subjectivation, a way to understand Kaggle's significance even as its solutions never seem to escape the platform.

Given the technological transformations occurring over the last three decades, subjectivation has become thought of as infrastructural and impersonal (Langlois & Elmer, 2019), leading to an expansion in critical attention to the processes of subjectivation. Though technical, the effects are social. Digital infrastructures datafy our body—our face, our gender, the colour of our skin—and reduce our lives to quantifiable metrics to be distributed and processed into live databases full of calculated linkages among humans, nonhumans, and the environment. They create a new cartography of biopower (Lazzarato, 2002). Individuals are no longer subjects of value realization but objects from which raw material is extracted in order to identify, sort, and classify populations; to shape individual behaviour, habits, and attitudes; to create value and capital; to modulate and mediate our experience of and within the world. Digital platforms actively cultivate the creation and exchange of user-generated content, whether personal (pictures, posts) or contextual (purchasing patterns, viewing habits). At the same time, while individuals experience subjectivation at the front end of a centralized app, the backend processes connect with third-party systems (other platforms, advertisement agencies, security firms, and governments) seeking to capitalize upon impersonal relationships.

While Langlois and Elmer (2019) are interested in how large digital and social media giants expand and transition their business beyond their users and toward mobilizing subjective capacities in broader social and economic environments, my research focuses on Kaggle as a site founded and operating under these goals: mobilizing subjective materials usually associated with an individual to transform broad conditions of existence. Kaggle has many similarities with social media, but it is unique in how it engages users and produces value. Unlike most digital platforms that have forged their empire on mining individual data and on an attention economy, Kaggle harvests free labour on the Internet, promoting gamified crowdsource competitions to seize individual data and produce predictive models. It participates in the "content creator economy." However, instead of creating social media posts or videos for streaming platforms, the main products are data, code, statistical analyses, and machine-learning models, such as object detection, facial recognition, predictive behavioural models, and recommendation systems.

Kaggle produces socio-technical knowledge and cultural values as predictive models and AI products that aim to improve the logic and infrastructure that circumscribe social and algorithmic media. Many scholars have contributed to our understanding that particular cultural values are inscribed in AI and the methods used to build these systems, fuelling the indiscriminate use of these technologies. Recent studies describe how statistical models reinforce all kinds of social and cultural biases in our society (Apprich et al., 2018; Bolukbasi et al., 2016; Buolamwini & Gebru, 2018; O'Neil, 2016; Xu et al., 2022) and how predict models are idealized for pragmatic economic uses (Finn, 2017; Hong, 2020; Napoli, 2014; Srnicek, 2017). Scholars have also examined how digital platforms have transformed the digital landscape (Chun, 2017; Citron & Pasquale, 2014; Steinberg, 2017; van Dijck, 2013; van Dijck et al., 2018); analyzed the ethics and political impact of algorithmic decision-making (Amoore, 2020; Bucher, 2012; Bozdag, 2013; Crawford, 2021); and demonstrated how AI has been used for surveillance and social control (Crary, 2014; Gillespie, 2018; Kalluri et al., 2023; Zuboff, 2020). However, most of these contributions have looked at the work done by large teams of well-paid developers within big tech companies like Google and Amazon or at mature and commercially proven systems like facial recognition and algorithmic recommendations. Little is known about places such as Kaggle, where AI systems are iteratively prototyped, tested, and refined—tasks often run by volunteers and low-wage workers. Despite the public attention, massive machine learning competitions, and a growing user base, Kaggle is a hidden layer in the AI industry that must be carefully examined.

Kaggle serves in this research as both an object of inquiry and as primary material source. By examining the competitions held on the platform, I aim to shed light on how big data is mobilized to produce value, both in terms of monetary gain and technology of the self that serves as an instrument of mediation for governing populations. Kaggle is a crucial place to study how social and political norms and rules are codified into models that shape individuals' behaviours, preferences, and desires, which, in turn, produce the subjects of this digital era. My dissertation adds further knowledge of this new political economy of subjectivation, where the dynamics of our relationship are mined and mobilized by a variety of social and economic actors and institutions to fulfill a specific political agenda.

Infrastructuralized platforms do not simply privatize individual data and public resources; they seek to generate revenue streams from modular forms of population management. As such, we must think of algorithmic media and AI as a discourse that produces specialized knowledge based on specific inputs from datasets as a way to manage populations. Drawing from Foucault's (2017) concept of subjectivation, I argue that the constitution of the subject is not merely the relations we have with our own individuality but our relationship to other individuals, collectives, institutions, and the world at large. It is a process whereby one's sense of self as an individual agent is paradoxically shaped according to forces external to the self. In other words, AI systems interpellate individuals as subjects of the digital society, not only to reinforce our conditions of existence (Lazzarato, 2004) but also to intervene, modulate, and (re)mediate our embodied experiences (Grusin, 2015). The questions are, in turn, how do algorithmic media produce subjects and shape our sense of identity? and what political economy and sociotechnical conditions are involved in the production of AI?

Many theorists argue that we live in an informational and immaterial economy (Castells, 2000; Lazzarato, 2004; Terranova, 2004), wherein collective cooperation (labour), cultural content, knowledge, services, and affect become sources of value. However, the whole production chain of the AI industry is long and complex, involving the exploitation of both material and immaterial resources. As Crawford (2021) puts it, AI is a technology of extraction: from the mining of rare earth minerals to the exploitation of low-wage workers to the data taken from every action and expression. The logic behind digital platforms and the exploitative economic model that drives the AI industry create an imbalance between the different types of labour and resources necessary to make the technology work. Following the logic of late capitalism, the digital economy is dominated by the obsession to acquire and own data. Data has become a new form of "raw material" that must be extracted, polished, transformed and exchanged (Srnicek, 2017). This vision is an attractive conceptualization for digital platforms: as a raw resource, data is merely being "channelled" through "online veins," allowing a wide variety of actors to monitor what individuals think, feel, and experience in order to be quantified and subsequently commodified. However, data is always already prefigured by sensorial apparatuses and surveillance techniques. Data scientists promptly embrace the contradiction, arguing that "raw" data must be cleaned, organized, described, and optimized to be used "correctly." But what does it mean to say there is a "correct" way to use data? How do data scientists, computer engineers, investors, and others in power define the rules and values for the "proper" use of data?

Lisa Gitelman reminds us that data is a contentious product of social, political, economic, technical, and cultural conditions. Data is not neutral or a natural resource that exists by itself in the world but a cultural one that needs to be generated, observed, and interpreted. Data is never raw but "precooked" (Gitelman, 2013) with different "flavours" and "spices" for specific usages. Conceptually, data is a product of our actions that, intentionally or not, carry assumptions about its nature, quality, and relevance. Deciding what constitutes "data" and what kind of data is relevant is a conflicting and disputed task that involves complicated power dynamics. This process raises questions such as what counts as “data,” who gets to decide what is included and what is excluded from datasets, where is it stored and preserved, who has access to it, how do we interpret and experience data, and what gets lost in the process of datafication. Data exists and ceases to exist by and through human actions, a contentious process stemming from specific social conditions, political disputes, cultural hegemonies, economic pressure, technical limitations, and epistemological standpoints. Digital systems do not merely “measure” activities, sentiments, thoughts, and performances but also shape and trigger them. We should ask how algorithmic media produce subjects and for what purposes, what kind of subjects these systems produce, how they intervene and modulate individual behaviour, and what values these systems extract and recreate.

Raw or precooked, data is intended to be mobilized for specific purposes. In a capitalist society, we can agree that the primary purpose is to increase sales, expand markets, reduce costs, and earn profits. However, I argue that something else is at play when aggregated data is mobilized to optimize capital surplus. We should ask how predictive models create the conditions for capital gain. Or, put differently, what kinds of mobilization of data take place in and around the production of predictive models? The data science community believes that “data speaks for itself,” preaching that it would be the source of truth from where we can extract “signals” to build precise and predictable models of reality. To track and record these signals, data scientists reduce the body to an aggregated blob of data that can only exist as correlations to predefined metrics. Physical characteristics, psychological attributes, behavioural traits, affect, and desires are broken down into a set of obscured but logical relationships that machines can read, catalogue, “make sense of,” and act upon. It implies that what is at play in the mobilization of data is the effort to build instruments of control capable of producing subjects that fit a particular social condition. It is essential that the customer, the consumer, and the citizen have a “life experience” compatible with a specific political economy in order to fulfil the desire for the profitability of a service or product. Therefore, predictive algorithms have to target and interpellate individuals as subjects of a stable and predictable mathematical model where the desires of capitalism can exist inexorably. In other words, I aim to show that data has been mobilized to create predictive models that (re)produce the conditions through which anyone, or anything, gains the possibility of existence in this world.

This ostensibly untroubled data mobilization for algorithmic subjectivation is the focus of my research. I contend that algorithmic media should be understood not merely as an interface that sits between users and other agents but as assemblages of humans and nonhumans where social norms, cultural heritage, and political economy are interlaced and modulated by technical rules, matrices of optimization, and epistemological idiosyncrasies. The subtlety of these instruments of control substantially impacts our lives, driving our behaviour, habits, and decisions, nudging us to do and think things in a specific way. Following Grusin (2015) and Kember and Zylinska (2014), I argue that these assemblages constitute a complex environment that constantly modulates and mediates each other at two interdependent levels: at the cognitive level, where immaterial knowledge (culture, arts, politics, ideologies) circulates and exerts power, and at the embodied and material level, where particles, objects, subjects, and events take place. These technologies are an affective and performative force that not only reinforces our conditions of existence in this world but also intervenes, modulates, and (re)mediates our embodied experiences. It becomes vital to understand how these mediation processes come into existence, materialize themselves, and actively (re)shape all sorts of individual, collective, and institutional relations. We should then ask how AI and algorithmic media, hardwired with ideological, political, and economic preconceptions and filled with social and cultural biases, affect everyday life. In this case, as a hub of AI development, Kaggle is a place through which we can understand how technical solutions become the answer to social problems and a way to grasp the genesis of algorithmic subjectivation.

This research's central question is *how algorithmic media produce subjects and reorganize life*. To answer this rather broad inquiry about algorithms and AI, I propose the following research questions:

- *How do personal and impersonal data become the raw material that fuels machine-learning algorithms?*
- *How are code, data, and political economy interests mobilized to create valued predictive models?*
- *How are predictive algorithms used to intervene and modulate individual behaviour?*

These questions attempt to identify the values and strategic projects that govern the sociotechnical and political choices made by engineers and data scientists, but also venture capitalists and academic researchers when developing AI systems. However, the answers to these questions may seem intertwined, taking the form of dynamic sketches rather than unequivocal deductions. Taking apart the mechanisms of machine learning algorithms and predictive models in relation to the process of subjectivation means undertaking a strategic analysis of the social relations they imply. Examining how AI is developed requires careful consideration of the values, practices, and power relations involved in the production of predictive models and how they affect the data science community and the individuals they target. These models often become a starting point for AI applications deployed in solutions distributed through and within algorithmic media.

Lastly, the main objectives of this dissertation are two-fold: (1) to examine Kaggle's history, economic model, the way the platform operates, as well as the community of users formed around it; and (2) to identify the primary machine-learning modalities at play on Kaggle competitions and examine how they identify, predict, and modulate human behaviour, therefore contributing to a new form of political economy of subjectivation. To understand Kaggle, its community, and the modes of machine learning for subjectivation, I first need to explain the platform's relationship to algorithms, AI, and the larger question of algorithmic media.

1.1. Definitions

My research developed during a shift in terminology in digital media studies. At the start of the project, the field had begun to question the importance of algorithms in critical media systems. In 2016, Tarleton Gillespie and Nick Seaver published the Critical Algorithm Studies Reading List.[1] Algorithms quickly became eclipsed by the growing interest in artificial intelligence when AI Now launched in 2017, which helped popularize the emerging field of Critical AI Studies (Crawford, 2021; Roberge & Castelle, 2021). Both terms, crucially, can only be defined socially and technically, or sociotechnically. Technically, the differences between AI and algorithms are greater than the social or cultural differences. AI is technically comprised of algorithms, and yet AI culturally promises to be "intelligent," whereas an algorithm only promises to be automated. In the following section, I offer a lexicon of key terms relevant to my study, although I acknowledge that I am working with and during shifts in these terms. After offering some brief definitions, I elaborate on how my three key critical terms—sense, mobilize, modulate—best capture the practical importance of AI and algorithms, both to my object of study and to the field.

I use the term *algorithm* to describe a set of steps typically used to solve a class of specific problems that can be done automatically. It can be formalized in different formats, such as a cake recipe, a mathematical equation, an instruction to gather or send information, a function to process data, or a command to execute simple or complex operations. In computer science, it is described as a set of rules or processes to be followed in calculations or other problem-solving operations to achieve practical solutions to problems like efficiently sorting a list of items, clustering data, or finding a

[1] See https://socialmediacollective.org/reading-lists/critical-algorithm-studies/

pathway through a network. Advanced algorithms may use conditionals to divert the code execution through various routes (automated decision-making) and deduce valid inferences (automated reasoning). In that sense, algorithms automate and speed up repetitive routines and processes. Ultimately, it describes a task that, if followed precisely, not only leads to the desired outcome but also prescribes a specific way to do things (Franklin, 1999).

Algorithms have become part of a more extensive process to train and produce what we call AI, but might be better called *machine learning* to produce statistical *models*. The specific algorithms themselves—and the different approaches to machine artificial intelligence—ultimately are just one small part of the production of models. Machine-learning techniques, including algorithms and pre-trained models, encode datasets and, in so doing, identify patterns that produce generalized models then used to perform tasks on unseen data without explicit human instructions (IBM Cloud Education, 2020). There are two kinds of approaches to machine learning: supervised and unsupervised. Supervised learning involves complex algorithms and labelled training data, where the developer defines the variables and goals of an algorithm. In this case, both the algorithm's input and output are specified. Unsupervised machine learning, on the other hand, involves unlabeled data and algorithms looking for patterns that can be used to cluster data points into subsets. Most types of deep learning, including neural networks, are unsupervised algorithms. The outcome of machine-learning processes is also called predictive models, which can be used, among other things, for object detection and classification, facial recognition, prediction of future trends, behaviours, and patterns, recommender systems, and further automation and optimization of production processes.

Technically, *artificial intelligence (AI)* refers to quasi-autonomous algorithms that leverage large datasets and predictive models to be used in digital applications to simulate human intelligence. These include simulations of learning (e.g., the acquisition of information and rules for using the information), reasoning (e.g., using rules to derive inferences), and self-correction (e.g., identifying errors and fixing them). The most popular examples today are Open AI ChatGPT and Google Gemini, which are conversational chatbots training on large datasets crawled from the Internet and elsewhere. Whether ChatGPT or any AI is actually intelligent is subject to debate. So, too, is the concept of intelligence. AI has been described as having three phases: (1) Narrow AI, designed to perform a specific task, such as voice recognition, recommendation systems, and image recognition; (2) Artificial General Intelligence (AGI), an advanced machine able to understand, learn, adapt, and implement knowledge in a wide variety of tasks at the level of a human being; and (3) Artificial

Super Intelligence (ASI), a hypothetical construct, often described in science fiction as either the destroyer or the saviour of humankind, which would surpass human intelligence in practically every field, including scientific creativity, general wisdom, and social skills. While ASI is a piece of fiction, and current developments arguably place us somewhere between Narrow AI and AGI, this research is only concerned with the currently available technology—that is, Narrow AI, which I will generally refer to as AI. There are plenty of examples of AI use: recommendation systems on social media, streaming platforms, chatbots and personal assistants like Siri and Alexa, facial recognition on smartphones, object detection embedded in airport scanners, predictive crime software, self-driving cars, and smart cities.

Algorithms are not just tools to solve problems. They are purposeful instruments that produce specific problems to be solved, which are not only technical but also economic, social, and cultural. They carry many assumptions about the problem they are built to solve, which, in turn, reflect the politics of algorithms. For example, social media platforms sort posts based on levels of engagement and assumptions that there is a specific correlation between what users interact with and the content they want to see or consume. Google PageRank (Bozdag, 2013) and Facebook's EdgeRank (Bucher, 2012) are typical examples of algorithms that aim to solve problems of visibility, mediation, and optimization of content delivery. Beyond organizing information, these algorithms re-produce a specific society, ideology, norm, and set of priorities. Algorithmic media aims to mediate communication between humans and machines, shape user behaviours, influence preference formation, and impact content production decisions. That is, algorithmic media not only assists people in finding information, but also provides a means to know what there is to know and how to know it; to take part in social and political discourse; to familiarize ourselves with the public in which we participate, to market a product, and to manage and shape social habits.

In short, and keeping the turn to AI in mind, an algorithm or AI involves three phases of development: 1. ***Sense***, or ways to turn the world into data readable by technical systems (datafication, digitization, and metrification); 2. ***Mobilize***, or ways to assemble labour, infrastructure, code, and training data in order to adapt algorithms to turn sensed data into meaningful outputs and produce predictive models; and 3. ***Modulate***, or the ways that the combination of algorithms, data, and predictive models are applied to social problems.

I use *sense* to discuss how specific properties of the human body and behaviour became central to the development of machine learning as a commercial product and as an instrument of control. For instance, I examine how digital sensors (hardware and software) are developed to identify features of our body and recognize facial attributes, detect objects, predict threats, and produce comprehensive individual profiles from personal and impersonal data. These technologies reduce human life to quantitative metrics to fit into statistical models designed for specific goals, more often for commercial gain, though they are also frequently used to prescribe a particular condition of existence. Sense sets the stage for each machine-learning competition considered in this dissertation, giving the political, sociotechnical and economic context in which each competition took place.

Mobilize focuses on a specific competition to discuss how data, code, and developers were mobilized to solve a particular problem. Here, I explain how the competition was planned, describe the dataset and the codes used in the event, and discuss the contentions around the competition objectives, such as the ethical concerns about how datasets were collected and how they will be used. I also examine how the data science community approached the challenge, as well as how the competition's sponsors exploit the gamified crowdsourced work on Kaggle. Mobilize uncovers the processes of production at the centre of the machine learning industry, which is heavily dependent on open-source technology and free labour. It provides empirical evidence about how predictive models have been developed as new forms of subjectivation.

Modulate unpacks the broader implications of each modality of machine learning algorithms. Here, I connect the work done in each competition (e.g., code, inferences, statistical analysis, predictive models, and discussions in the website's forum) with other examples inside and outside Kaggle to generalize the arguments made in this dissertation. I also contextualize the work done by the data science community within the digital economy, discussing the impact of machine learning algorithms as tools for automation, labour exploitation, and social control. Modulate aims to understand how these predictive models based on commercial interests exert a controlling influence over society, examining the ways in which predictive models interfere, mediate, and shape our sense of self.

These phases do not imply self-contained stages or a rigid linear sequence in machine-learning development. Quite the opposite. They are iterative, often overlapping and may occur in a different order as developers obtain feedback and external data from already deployed applications. These three terms may describe the works of AI and algorithms broadly but specifically describe my case study (see Figure 1.1): Kaggle provides data (sense) and infrastructure for developers to apply specific machine-learning techniques (mobilize) to win competitions and provide solutions for real-

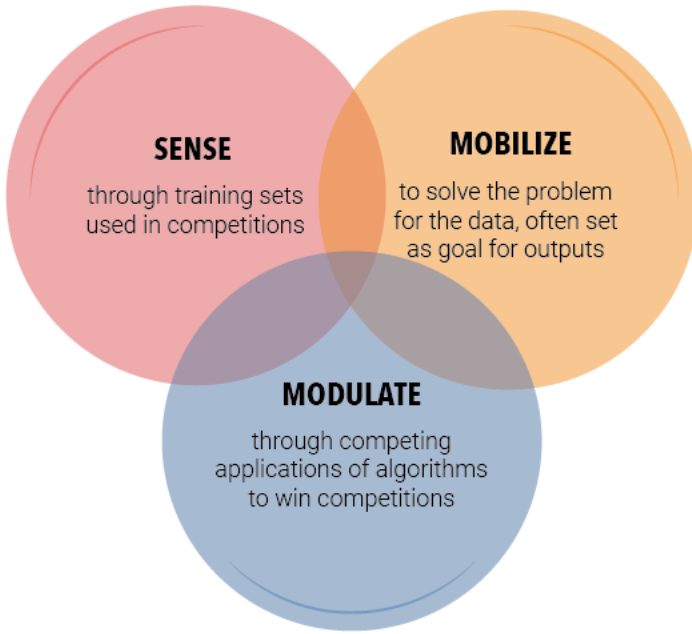


Figure 1.1: Phases of AI development on Kaggle.

world applications (modulate). Kaggle competitions are at once a dataset, a race to develop the best model, and a solution to social problems.

However, the aim of this project is not so much to understand the technical (data structures, statistical models) but the political and sociotechnical foundations of AI and algorithmic media: how they are designed to do what they do. That is, what is essential is not to grasp how machine-learning algorithms and predictive models work but rather to unveil and unpack the implications of how they work for the purposes for which they are designed. Take, for instance, the rapid adoption of facial recognition by police enforcement and the reinforcement of racism ingrained in our society. Or the free labour necessary to tune predictive models to increase airport security, the indiscriminate use of private data to nudge users to buy products, and the dissemination of disinformation and hate speech due to enhancements in recommendation systems based on engagement as profit. As such, my research concerns the political economy and the discursive power that saturates AI. The focus is on particular discourses of optimization and efficiency that shape their design and practices and produce certain positions of power. I argue that there is a specific discourse of AI driven by a particular epistemological and ideological orientation overloaded in the training sets and machine-learning algorithms, a discourse that aims to classify, sort, predict, and mediate human behaviour at the micro and macro scales.

Algorithms and AI are all part of a wider theoretical concept: *algorithmic media.* The models produced on Kaggle do not remain on the platform: they become building blocks of experimental and commercial applications employed in solutions distributed through and within algorithmic media. I use the term algorithmic media to discuss the broader term in the platformization of infrastructure to link and develop systems that rely on multiple algorithms. Computer algorithms usually do not work in isolation but are deeply interconnected, interlaced, chained, and overlapped with other algorithms. Kitchin (2017) writes that algorithms are usually woven together with hundreds of other algorithms to create algorithmic media. In a networked system, algorithms are within and between devices: they are both nodes by which information is demanded, consumed, and transformed, and in the edges in which data transverse an intricate web of distributed machines. When algorithms stop being just a sequence of steps to compute and produce an output and start incorporating mass communication functions, they become a social force. One algorithm informs another, creating a continuous modulation that Deleuze (1992) identified as a distinct characteristic of contemporary power. Algorithmic media are assemblages of codes, instructions, and processes that "function as control technologies in all sorts of media and information systems, dynamically modifying content and function through these programmed routines" (McKelvey, 2014, p. 598). As McKelvey (2018) has argued, the infrastructural turn in media studies must attend to the underlying algorithms.

Algorithmic media are a helpful context for understanding how algorithms become entangled with media technologies, acting as a force to drive society in specific directions and shape everyday life. Ultimately, my research is primarily concerned with the ways AI mediates individuals' sense of self and how machine-learning and predictive models have been developed as a new form of subjectivation. That is, I am interested in how code, data, digital infrastructures, crowdsourced labour, and political economy interests are mobilized to create instruments of control that shape individuals' behaviours and produce subjects according to specific hierarchies of power.

1.2. Chapter Summaries

This dissertation is divided into nine chapters. The first two chapters following the introduction outline the theoretical foundation and the methodological approach taken in my doctoral project. Chapters four and five offer a narrative of Kaggle's history, both as a company and a platform, and a description of how the platform works. Chapters six to eight focus on Kaggle competitions, examining the different machine-learning modalities the platform and its sponsors put forward:

identify, predict, and recommend. Each chapter dives deep into one competition to explore the symbiotic relationship between the platform, the competition's sponsors, and the developers in a predatory and hyper-competitive environment aiming to advance a specific sociotechnical paradigm that is key for the currently established economic agenda. The final chapter draws conclusions from the machine-learning modalities to make general observations about what they can show us regarding a data-driven political economy of subjectivation.

This dissertation examines Kaggle machine-learning competitions through three theoretical lenses that complement each other: political economy, subjectivation, and mediation. Chapter two examines key works in these fields through a literature review to acknowledge and comment on existing theories and form connections with the empirical objects analyzed in this dissertation. In terms of political economy, the chapter focuses on the political economy applied to Communication Studies and delineates the general concerns of the field using the work of Vincent Mosco and David Hesmondhalgh. Drawing from Dyer-Witheford, Maurizio Lazzarato, and Tiziana Terranova, I discuss the new forms of commodification in the digital economy, in particular, the concepts of platformization put forward by Tarleton Gillespie, Nick Srnicek, and Jose van Dijck, and Surveillance Capitalism, proposed by Shoshana Zuboff. Subjectivation is defined for my purposes through a combination of Michel Foucault's approach to the concept and subsequent work by Teresa de Lauretis, Judith Butler, and Simone Browne. Ganaele Langlois and Greg Elmer's theories of impersonal subjectivation and Byung-Chul Han's concept of psychopolitics are also used to enhance this understanding. Lastly, mediation is defined here by combining the historical accounts of the term described by Raymond Williams with more contemporary work by Richard Grusin, Sarah Kember, and Joanna Zylinska.

Algorithms are sociotechnical assemblages and exist in two states simultaneously: material objects—pieces of code stored in machines made of silicon—and dynamic immaterial forces that affect everyday life. Studying the impact of algorithm systems requires a multifaceted approach, combining traditional methodological perspectives in Media Studies with new methods focusing on new media and digital platforms. Chapter three discusses the methodologies used in the project. First, I define the timeline considered in this dissertation, followed by a description of the source material, primarily from within the Kaggle platform, where I gathered workable data. Drawing from Richard Rogers's work on Digital Methods, I developed a Web scraper to collect data from Kaggle's website. I then used a qualitative content analysis approach to define the main machine learning modalities on the platform. Software Studies and a Foucauldian perspective on Discourse Analysis are combined as an analytical framework to examine the discourses produced in the platform, the power relations in the machine-learning community, and the broader implications of predictive models and AI in society. Lastly, I disclose my identity as a researcher and my positionality in relation to the topics of this dissertation.

Chapter four offers a narrative of Kaggle's steps to become a successful start-up company able to attract millions of users. I describe how the company first emerged as a hackathon promoting machine-learning techniques to predict the 2010 Eurovision Song Contest. The chapter identifies the main actors involved in its conceptualization, such as its founder, Anthony Goldbloom, an econometrician from Australia, and Max Levchin, a libertarian venture capitalist. I show how Kaggle's website evolved, highlighting the struggle to define itself as a platform among other platforms. More importantly, I examine the platform's economic model based on crowdsource and gamification and discuss how it evolved from a niche and specialized platform for engineers into a large community of data scientists later acquired by Google.

Chapter five sheds light on Kaggle's digital infrastructure and user base, as well as on how the company organizes machine-learning competitions. First, I show how Kaggle works as a code and data repository for machine learning, highlighting its cloud services and user contributions. I then focus on user demography to make a portrait of Kaggle's user base. I also describe how Kaggle raking systems work and briefly examine the platform's top users. Lastly, I describe the types of competitions promoted by Kaggle and its main rules and protocols, highlighting the most popular competitive ones. I also classify the competitions into modalities of values based on the primary goal of each competition, namely, identify, mobilize, and modulate. The ensuing chapters use these modalities to examine how machine learning has been used to modulate individual behaviour and, therefore, create new processes of subjectivation.

Chapter six focuses on machine learning as a technology for producing regimes of truth. As a competition modality on Kaggle, "identify" involves the automatic detection, identification, classification, and recognition of bodies, things, and environments, which are techniques to produce knowledge and define what is knowable. What interests me here is the practice of truth embedded in machine-learning algorithms, which allows for creating knowledge that can be useful and, at the same time, modulate individuals and collective affective moods and structures of feeling. More specifically, this chapter discusses how machines came to "learn" how to detect, identify, and recognize faces. While facial recognition can be used for many purposes, I focus on its discriminatory embeddings stemming from developers' idiosyncrasies, technological farsightedness, and the abuse of such technology for creating deepfakes, in particular, porn revenge and manipulation of public opinion. Facebook Deepfake Detection Challenge serves as an instrument of discussion about how facial and object recognition algorithms have been used to define the ontologies of the world we live in.

Chapter seven focuses on machine learning as a technology of surveillance and speculation. As a competition modality on Kaggle, “predict” has to do with forecasting future events, trends, and behaviours. Predictions serve as a benchmark for the other machine-learning modalities discussed in this dissertation, assigning mathematical truth to each model regardless of its purposes. What interests me here is how data is mobilized to predict individual behaviour but, above all, to assign standards of ab/normality. Consider, for instance, what is at play in airport full-body scanners: What do these scanners capture and measure? What can these measurements tell about a person’s identity and behaviour? Since airports have long been laboratories for new technological and social control strategies, this chapter uses the Passenger Screening Algorithm Challenge to illustrate the data science community’s chilling and blinding assumption that machine learning can reveal fundamental truths about individuals.

Chapter eight focuses on machine learning as a technology of manipulation. As a competition modality on Kaggle, “recommend” involves curating limited and ranked options to support decision-making processes. Here, I am interested in the direct impact of machine learning algorithms on individual behaviour. Instead of facial measurement or external physical properties, the example explored in this chapter seeks to shape behaviour and psychological traits. Consider, for instance, what is at play when you buy groceries: What items go into your shopping cart? In what order? How often do you buy these items? Does the way these items are arranged influence what you choose to buy? Do advertisements convince you to change your mind? Instacart Market Basket Analysis is an example of how data and code are mobilized to make inferences and assumptions about mundane and everyday life activities, such as groceries. In particular, I focus on how the machine-learning community explores and exploits personal data, but above all, how they engineer and fabricate data patterns to find the best predictor for user behaviour, sometimes producing spurious correlations and “hallucinations.”

Lastly, the conclusion summarizes the arguments and findings of the previous chapters and discusses the broader implications of the role of AI as a new form of political economy of subjectivation. The objective of applied machine learning on Kaggle is to find patterns in data, most commonly human behaviour, for capital. This practice reveals important sociotechnical tendencies that permeate the development of predictive algorithms and artificial intelligence. In particular, I contend that the data science community has a notable compulsion to reduce the cost of production, an indifference toward human life, an obsession to control populations and individual bodies, and a desire to produce a predictable future for economic gain. This chapter also lists the limitations of this research and offers suggestions for further research.

2. Theoretical Framework

> Capitalism should not be eaten raw. Capitalism, like sausage, is meant to be cooked by a democratic society and its institutions because raw capitalism is antisocial. (Zuboff, 2020, p. 43)

The technologies developed by AI businesses and machine-learning communities such as Kaggle claim to be able to do a variety of things, including recognizing subtle behavioural patterns and automating tedious, repetitive tasks. In its 2017 report, the AI Now Institute argues that the examination and comprehension of artificial intelligence should not be restricted to its technological capabilities. Our current social, economic, and political relationships and institutions face significant normative and ethical challenges due to the design and application of this next generation of computational tools, and changes in these areas are already underway. Simply put, ”AI does not exist in a vacuum. We must also ask how broader phenomena like widening inequality, an intensification of concentrated geopolitical power and populist political movements will shape and be shaped by the development and application of AI technologies” (Campolo et al., 2017, p. 3). As such, we can think of algorithmic media, including AI and machine learning, as a discourse that produces specialized knowledge based on specific inputs or inferred from a dataset and subjects involved in the process, such as developers, designers, data scientists, consumers, and users.

When released into the world, these systems express authority and algorithmic governance, working within their own regime of truth through the automatic and autonomous exercise of power, often without human oversight. In essence, algorithmic media produce and enact regimes of power and knowledge as they are predicated on a defined set of rules about how a system and its users will behave at any one time and in any given situation. I argue that this phenomenon is connected to a political economy of subjectivation designed to control the “production of possibilities” and reinforce specific types of socioeconomic relations, creating the conditions of existence that determine how resources and people are organized, who is valued in what roles, what activities are undertaken, and to what purposes.

The increasing number of media technologies used to attract and capitalize on individuals has expanded the scope and capacities of a political economy of subjectivation. Tech companies are moving away from an enclosed platform model toward a distributed, impersonal infrastructure

(Langlois & Elmer, 2019): from surveillance cameras to mobile and wearable technologies to new technologies of information and connectivity that increasingly track, quantify, datafy, and analyze not only our movements, gestures, and emotions, but also what constitutes the context around us. We have now developed processes to track preconscious routines, from biological rhythms to behavioural habits such as time spent scrolling pages, staring at a screen, walking around, and standing by.

This chapter constitutes a theoretical framework to understand how algorithmic mediation produces subjects and reorganizes life. This framework combines three theoretical lenses that complement each other, aiming to extend our understanding of the relationship among technology, political economy, and processes of mediation and how these relationships impact the processes of subjectivation. I use this framework to examine Kaggle both as a company and as a digital platform, as well as how machine-learning competitions promoted by and within the platform mobilize machine, code, data, digital infrastructure, and crowdsourced labour to produce predictive models used in AI systems. Drawing from Vincent Mosco's (2008) approach to the political economy of communication, the first section of this chapter discusses the processes of digitization and datafication triggered by digital platforms. I highlight how data has become the "raw material" from where tech companies extract value (Srnicek, 2017) and behavioural surplus (Zuboff, 2020), reinforcing not only a mode of production but a production of modes of living (Lazzarato, 2004) to which individuals are subject. Drawing from Michel Foucault's (1995) definition of subjectivation, the second section situates the process of subjectivation in the context of digital infrastructure and algorithmic media. In particular, I argue that algorithmic media constitute a new form of subjectivation—an impersonal (Langlois & Elmer, 2019), distributed, and automatized process—that is not only shaped by biopolitics but is also conducive to psychopolitics (Han, 2017). Lastly, I borrow the concept of radical mediation from Grusin (2015) to discuss how algorithmic media should be understood not as an interface between users and other agents but as assemblages of humans and nonhumans. I argue that this assemblage constitutes a complex symbiotic environment (Kember & Zylinska, 2014) in which algorithms, develeopers, users, and political-economic intererests constantly modulate and mediate each other, re-producing our current conditions of existence.

2.1. Kaggle as a Platform

Kaggle is a crowdsourcing platform focused on AI development that coordinates large groups of dispersed participants around competitions. With more than five million registered users across 194 countries (Crunchbase, 2019), Kaggle is a platform where statisticians, computer scientists, economists, and marketers collaborate and share experiences but, most importantly, compete for prizes in solving complex problems with predictive models. The platform doubles as a data repository for machine-learning training, storing more than 50,000 datasets, including search engine results, crime data, commodity prices, public health, credit and financial fraud, social media interactions, and all kinds of biometric and demographic data. It also offers cloud computing, code repository, and courses and tutorials on machine learning. From 2010 to 2020, Kaggle organized over 3,000 machine-learning competitions, distributed over US$16 million in prizes, and attracted more than five million registered users who contributed to creating dozens of millions of predictive models.

Competitions connect companies looking for AI-driven solutions and cheap labour to machine-learning engineers looking for challenges and job training. Kaggle promotes and curates competitions to attract new talented developers to work on experimental and speculative problems involving predictive modelling. It is a platform where statisticians, computer scientists, economists, and marketers collaborate and share experiences but, most importantly, compete for prizes in solving complex problems using machine learning. Kaggle doubles as a platform for code-data repository for AI development, cloud computing infrastructure for predictive model training, and ed-tech, offering courses and tutorials on programming languages, data science, and machine learning.

Platforms are often described as business models (Srnicek, 2017). Gillespie (2017) describes platforms as “sites and services that host public expression, store it on and serve it up from the cloud, organize access to it through search and recommendation, or install it onto mobile devices” (p. 254). Working as mediators between types of users, platforms put forward a new economic model built around data circulation that enables a multisided market with interdependency connection among different actors: customers, advertisers, service providers, producers, suppliers, and even physical objects (van Dijck, 2013; van Dijck, Poell, & Waal, 2018). However, platform studies tend to privilege specific types of platforms, most notably social media channels and multisided market apps. The first kind is commonly defined as a website that hosts and organizes user content (in textual or audiovisual format) for public circulation without having produced or commissioned it (e.g., Facebook, Twitter, YouTube). The content generated on these platforms is diverse (ranging from, but not limited to, personal opinion, (fake) news, entertainment, marketing material, and tutorials) and intended for fast consumption within the platform itself. These sites have a specific business model centralized on attention economy through curated feeds of user-generated content designed to keep users engaged and attentive to advertising (Bucher, 2012; Carr & Hayes, 2015; Gillespie, 2010; 2017; van Dijck, 2013). The second kind, multisided market apps, is typically described as mechanisms to match buyers and sellers using algorithms that link them based on criteria such as their needs and preferences to optimize the user experience and operational efficiency. These platforms centralize specific markets around themselves where they can regulate every aspect of the task or activity, from availability to price range (Schor et al., 2015; Srnicek, 2017). They specialize in providing commercial services, exploiting cheap labour, minimizing operational costs, and improving resource management, which gave birth to business models based on the gig economy. Uber, Instacart, and Upwork are examples of this kind of platform.

My project addresses a gap in platform studies largely focused on social media or app businesses and the possibilities of platforms as new sites of social coordination and production of what I argue above are techniques of impersonal subjectivation. Little attention has been paid to platforms aiming to streamline business processes by producing technological advancements. These platforms focus on crowdsourcing simple or complex tasks. They mobilize large groups of people to collaborate or compete in solving concrete or speculative problems. The solutions produced on these platforms are diverse depending on their specialization and business model, ranging from cheap microtasks (e.g., Amazon's Mechanical Turk, TaskRabbit), codes, prototypes, and comprehensive apps and digital systems (e.g., GitHub, NPM), to complex machine-learning code and predictive models (e.g., Kaggle, Hugging Face). Some of these platforms target emergent markets, niche communities, and highly qualified activity. They not only regulate the activities on the platforms (define rules, provide support, etc.) but also serve as a training ground for individuals looking to work in these fields. In this regard, Kaggle is unique among platforms because some of the knowledge (datasets, training skills, know-how) and material (code, models) produced by its users may not be configured as final products, but as prototypes that can be used in and applied to other platforms, such as the predictive models and recommendation systems created to solve challenges sponsored by private companies.

My focus on Kaggle builds on Nick Srnicek's (2017) original work, particularly in conceptualizing platforms as logistical entities that see data as a type of "raw material" that must be extracted for profit. Srnicek's analysis of platforms as central actors in the reorganization of economic and social life through data collection, analysis, and management aligns with my interest in Kaggle as a platform for developing sociotechnical capital focused on rules and optimalities. My research emphasizes the infrastructural role of predictive models produced in Kaggle in optimizing processes, coordinating activities, and managing resources across various sectors. Similar to Srnicek's, my research underscores the necessity of understanding platforms' logistical and data-centric strategies to fully grasp their impact on the contemporary socioeconomic landscape. The emphasis on data as raw material highlights the critical role of data-driven coordination in shaping modern economies and societies. By harnessing and processing large datasets, platforms not only improve logistical processes but also generate new forms of value creation and economic activity. This perspective reveals the dual nature of platforms as both technological infrastructures and economic actors that redefine traditional business models and market dynamics.

Kaggle has many similarities with Srnicek's digital platform topology, but it is unique in the way it engages users and produces value. Unlike most platforms that have formed their empire based on the attention economy and the mining of individual data, Kaggle harvests free labour on the Internet by promoting gamified crowdsource competitions focusing on developing AI systems. Unlike other platforms, Kaggle is focused on technological development rather than media content, generating revenue from sponsored competitors rather than advertisements. As I will discuss later, it fits Srnicek's definition of lean platforms, but instead of providing services, it closely resembles a "content creator economy." Yet, as noted above, instead of creating social media posts or videos for streaming platforms, the main products are data, code, statistical analyses, and machine-learning models, such as object detection, facial recognition, predictive behavioural models, and recommendation systems. By adding code, data, and programming skills to the hall of media content and activities carried out on platforms, my research allows for a more nuanced understanding of how specialized platforms navigate their sociotechnical environments and function logistically. Moreover, by exploring Kaggle's strategies and the implications of predictive models on user behaviour, business decisions, and political discourse, my research extends Srnicek's analyses by incorporating current trends and case studies from the AI industry.

My approach to Kaggle as a platform also closely aligns with Tarleton Gillespie's (2010) seminal work by looking at platforms not just as business models but as discourses. At the outset, Gillespie found that the term allowed large technology firms

> to make a broadly progressive sales pitch while also eliding the tensions inherent in their service: between user-generated and commercially-produced content, between cultivating community and serving up advertising, between intervening in the delivery of content and remaining neutral. (p. 348)

Gillespie's analysis of platforms as intermediaries crafting their narratives to navigate regulatory, market, and social landscapes aligns with my focus on platforms' discursive practices. Kaggle, I find, offers similar rhetoric in its sales. By analyzing the discourse and semantic strategies used by platforms, my research aims to reveal the underlying power dynamics and ideological constructs that govern digital ecosystems. The emphasis on the discursive nature of platforms highlights the critical role of language in shaping technological and social realities.

Kaggle competitions follow the Silicon Valley model for social change based on an optimistic hacking culture. These hackathons, as Kaggle's CEO Anthony Goldbloom calls them, "move fast and break things," aiming to produce the seeds of the future where technology replaces politics. Most often, however, they only produce demos and prototypes, sometimes not even functional or usable ones. Yet, the goal is not to produce meaningful results but to experiment and identify business opportunities. It is a place where new ideas for machine learning and AI are born to optimize capital surplus and social control. Kaggle is a laboratory for developing sociotechnical capital focused on rules and optimalities. With Langlois and Elmer (2019), I argue that it is a site of impersonal subjectivation where value is mobilized through large assemblages of data, code, crowdsourced labour, and specific economic interests. This form of subjectivation puts the individuals at a distance; it sees them as a collection of amorphous data capable of objectively describing a person. Individuals are reduced to pieces of information that go beyond their social lives (professional choices, moral grounds, ideology) to include their embodied experiences (race, gender, sexuality), the biological rhythms of their bodies (heart rate, physiological need, sleep cycles), and psychological traits (attention span, desires, affect).

In sum, my interest in Kaggle, and digital platforms in general, is a gateway to answering deeper political, economic, and epistemological questions about process subjectivation and mediation when platforms become tangled up with predictive models and AI systems. These technologies not only interpellate individuals as subjects of a postmodern society where we must fulfill a socioeconomic role, but also as objects of interest from which data is constantly harvested. Individuals' data is aggregated into impersonal hyper-profiles as individuals themselves become integral to a large and complex digital infrastructure. In a sense, algorithmic mediations are always remediations: it is not just about connecting experiences, but it also generates, transforms, and modulates them. In other words, the conjunction of digital platforms and predictive models is an affective and performative force that (re)mediates material experience within the world (Grusin, 2015). I argue that it has become crucial to understand the political economy of subjectivation enacted on and by digital platforms and how their mediation processes come into existence, materialize themselves, and actively (re)shape all sorts of individual, collective, and institutional relations.

2.2. Datafied Political Economy

My approach to Kaggle is imbedded in the political economy the company and the platform subscribe to. That is, how the company thinks, plans, and organizes its business, as well as the ways in which the platform puts the company's strategy into practice, such as the relations of production, ownership, and monetization. A political economy approach pays considerable attention to describing and analyzing modes of production and systems of governance. As the de facto dominant mode of production, capitalism, in its current neoliberal phase, is often at the centre of these analyses. Capitalism is focused on turning resources like workers, raw materials, land, and information into marketable commodities for a profit. It is no secret that the longevity and success of capitalism depend on the emergence of new market forms that express new logics of accumulation to meet the ever-evolving needs of those in power (Lazzarato, 2004; Mosco, 2008; Zuboff, 2020). The advantage of adopting political economy as a theoretical framework is that there are countless studies documenting and tracking the restructuring of public entities as well as unfolding their integration with the private sector, revealing a process of hybrid production in which the relationship between public and private interests become blurred at every level of government activity (Mosco, 2008). It is important to note, however, that due to commercial secrecy, some corporate activities remain hidden from the public, even when performed in plain sight, such as the work of algorithms in digital systems.

When applied to Communication Studies, political economy focuses on institutional aspects of media, telecommunications, and digital systems, examining the relationships among owners, labour, consumers, advertisers, structures of production, business models, and the power relations among them. The approach also emphasizes the transition from traditional analog media to digital platforms, deepening and extending tendencies within earlier forms of capitalism by opening new possibilities to transform the audience's inputs (i.e., personal, and impersonal data) into saleable commodities. Hesmondhalgh (2008) claims that the marketization of media involves several processes, most notably the "privatisation of government-owned enterprises and institutions, the lifting of restraints on businesses so that they can pursue profit more easily, and the expansion of private ownership" (pp. 100-01). It is crucial to pay attention to the shifts of power in the economy, accentuated by the aggressive disputes around ownership of communication infrastructures, as well as to consider the broader impact of these changes on the mode of production and the production of modes of subjectivation. Tech giants, such as Google, Microsoft, Facebook, Apple, and Amazon, are often cited as the main actors disrupting and ultimately transforming the way we think of communication and media technologies. Start-up companies backed by venture capitalists, such as Uber, Airbnb, Kaggle, and OpenAI, also play a role in shifting the power in the industry using the *blitzkrieg* strategy—refined as "Move Fast and Break Things" by Jonathan Taplin (2017)—to monopolize the Internet and disrupt niche markets, often exploiting unregulated activities or using unethical and illegal approaches.

Highly influenced by the Libertarian thought of the so-called "Californian Ideology" (Barbrook & Cameron, 1996), the dominant political economy in the digital industry pushes toward the deregulation of markets and the intensive use of digital technology and automation to free the economic potential of under-utilized assets (including human behaviour, attention, and affect). This apparent change, or "disruption" from the orthodoxy of twentieth-century capitalism, as the Tech start-up culture likes to brand it, has many names: gig/sharing economy (Schor et al., 2015), the next industrial revolution, surveillance economy (Zuboff, 2020), app economy, attention economy, and platform economy (Srnicek, 2017). Indeed, several theorists argue that we live in a different stage of capitalism, loosely called "late capitalism" (Terranova, 2004), where the economy is focused on the intangible and immaterial, such as networks and information (Castells, 2000), attention and cognition (Crary, 2014), behaviour and affect (Zuboff, 2020). In this logic, (labour) cooperation, cultural content, knowledge, services, affect, and so-called "raw data" become precious sources of value. Examples include media content such as YouTube videos and social media posts, as well as broader contributions in the form of creating apps, participating in online forums, or the simple act of using digital devices with sensors, trackers, or any form of connection to the Internet or other digital systems.

Immaterial labour and its outcomes need to be grounded, materialized, datafied, and quantified in order to become commodities. Following the logic of late capitalism, the digital economy is dominated by a new type of property: instead of means of production, the obsession has turned into the ownership of information—data, to be more precise—or at least to control and regulate access to it. Srnicek (2017) argues that data has become the new "gold," a type of "raw material" or "natural resource" that *must* be extracted. According to Srnicek (2017) data can be any piece of information captured from any source, be it natural (land, water, minerals) or human-generated (products, labour, behaviour, cognitive, affect). Srnicek suggests that data entails recording and, therefore, a material medium of some kind: "As a recorded entity, any datum requires sensors to capture [it] and massive storage systems to maintain it" (p. 39). Consequently, data is not immaterial in the sense that it exists within devices and data centres that need a constant influx of energy. Data gathering today depends on a vast infrastructure to sense, record, store, and analyze, which can only be constructed by prominent actors, such as sovereign states and private companies.

The apparent "rawness" of data is assigned to its supposed objectivity, neutrality, and transparency—that it is the self-evidence of the fundamental stuff of truth itself. However, scholars like Lisa Gitelman, Virginia Jackson, and Geoffrey Bowker dispute this assertion (Gitelman, 2013). As Gitelman argues, data is not a natural resource that exists by itself in the world but a cultural one that must be generated, protected, and interpreted. Data is a contentious product of social, political, economic, technical, and cultural conditions. It is always already prefigured through sensorial mechanisms and disciplined by those in power. Manovich (2001) reminds us that "data does not just exist—it has to be generated" (p. 224) and interpreted. Data must be imagined as data to exist and function as such, and imagining data involves an interpretive base grounded on beliefs, culture, and politics. As such, data is not raw but "precooked" (Gitelman, 2013). Data is a product of our actions that, inadvertently or not, carry assumptions about its nature, quality, and relevance. Consequentially, digital platforms do not merely "measure" data—sentiments, thoughts, and human activities—but also trigger, modulate, and mediate them. The politics and the interpretation of *what data means* and *for what it can be used* are crucial to understanding the processes of datafication, platformization, subjectivation, and mediation.

2.2.1. Datafication and Platformization

The idea that data is neutral, a raw resource to be captured, extracted, or even squeezed (Goldbloom, 2016), is a very attractive and profitable concept for digital entrepreneurs who are inexorably looking to expand capital gain. The increased production and control of large datasets have provided companies with a framework to customize and personalize products and services tailored specifically to each customer's needs. Beyond offering commercial services, platforms often regulate and interfere in what was otherwise understood as open markets. That is, platforms privatize the very essence of the economy: the marketplace itself. For Srnicek (2017), this is the crucial advantage over traditional business models since "a platform positions itself (1) between users, and (2) as the ground upon which their activities occur, which thus give it privileged access to record them" (p. 44). Consequently, platforms are far more than and go way beyond the "virtual world" of the Internet, extending their operations into the offline world and to every activity that becomes digitized. Examples of such platforms include Google, Amazon, Facebook, Apple, and Microsoft, along with their Chinese counterparts Tencent, Alibaba, Baidu, and JD.com.

Srnicek (2017) classifies digital platforms into five types: advertising, cloud, industrial, product, and lean. Advertising platforms extract user information, undertake a labour of analysis, and then use it to sell ad space (e.g., Google, Facebook). Cloud platforms own hardware and software and rent them on-demand to digital-dependent businesses (e.g., AWS, Salesforce). Industrial platforms build the necessary hardware and software for traditional industries and manufacturing to lower costs, increase scale, and speed up production (e.g., GE, Siemens). Product platforms use other platforms to transform traditional goods into digitally available services by renting or charging subscription fees (e.g., Car2Go, Spotify). Lean platforms work with minimum ownership of assets by transferring them to the users or service providers (e.g., Uber, Airbnb, Kaggle).

Furthermore, digital platforms, social media in particular, produce and rely upon "network effects:" the more numerous the users in a platform, the more valuable that platform becomes for everyone else (Srnicek, 2017; van Dijck, 2013). Facebook is an example: with over three billion users in 2023 (Dixon, 2023), it became the default social media platform simply by the sheer number of people on it. This effect produces a "natural" tendency toward monopolization, the ever-increasing monetization of everyday life, which in turn relies upon collecting, extracting, and processing even more data. Though digital platforms often present themselves as empty spaces to be filled—a piece of technology waiting to be used—they embody specific politics aiming to control user interactions and data extraction. In essence, the platform economy configures the neoliberal disruptive approach to keep everything the same. Digital entrepreneurs promote the digital platform as a self-organized collective of individuals that collaborate and produce creative work, described by Lévy (1999) as "collective intelligence" and by Surowiecki (2005) as a "wisdom of crowds." However, the main reason to invest in these platforms is to harness the free labour (Terranova, 2004) available in the digital world that crosses national borders, has little regulation, a lack of essential protection for workers, and no accountability for private companies (Dyer-Witheford, 1999; Lazzarato, 2004). As a result, concentration and monopolization tendencies are reinforced within a discourse of "raw data" and the supposed neutrality of the technology and digital platforms in relation to their mediation power.

The AI industry works in a similar fashion; it cannot survive without free labour and open access to large quantities of data/code (Crawford, 2021). It not only requires but thrives on a large number of people willing to share their work for free in order to create economic value, technological progress, and the reinforcement of a supposed meritocratic system in our society. The push for open-source software in machine learning was set in motion by a paper manifest authored by Sonnenburg et al. (2007) in which they argue that the lack of openly available algorithm implementations is a significant obstacle to scientific progress and economic growth. Langenkamp and Yue (2022) estimate that every dollar spent on Machine Learning Open Source Software (MLOSS) produces an impact of US$100 in the industry. Big tech companies and non-profit foundations spend between 100 and 300 million dollars per year on MLOSS, contributing at least 100 billion dollars to the global economy in 2022 (Langenkamp & Yue, 2022, p. 388). However, behind the altruistic desire to promote open-source projects lies the foundation of a predatory model that encourages individuals to actively work on these projects not as employees but as "collaborators" who share their work but do not own their contribution.

On the one hand, companies like OpenAI, Microsoft, Amazon, and Facebook rely on millions of low-paid workers (mainly in the Global South) to label and filter training data for AI models (Perrigo, 2023; Rowe, 2023). On the other hand, crowdsourcing economic models used by Kaggle and Amazon's Mechanical Turk, for instance, lure developers into competitive events to solve problems. Kagle focuses on the AI industry, promoting challenges to be solved with machine learning techniques whereby they are encouraged to work for free using their own computers. Whereas this creates a sense of ownership and collectiveness among developers, the intricacies of open-source licensing and commercial use weigh in favour of whoever sponsors the project. Large enterprises often patent technology they open source (Simonite, 2018), ensuring that the companies are protected from intellectual property lawsuits. Hence Kaggle fits into Srnicek's (2017) topology as a lean platform since it strives to exploit available assets in society at the lowest cost possible without directly owning them, with Uber as one of its icons. Paraphrasing Srnicek, we can define Kaggle as the world's largest data science company that owns no data, no data centres, and no code.

While Srnicek's (2017) classification of digital platforms' business model is useful for examining the economic aspect of digital platforms (including the relations among production, labour conditions, market concentration, and digital governance), my research focuses on how data is articulated to modulate human behaviour, ultimately producing subjects of the platform. Indeed, data—and lots of it—is a platform'a most important and valuable asset. Quantity matters more than quality and diversity. Every action performed by users, no matter how ordinary, is useful for producing predictive models, reconfiguring algorithms, and optimizing processes. Consequently, a significant problem for the AI industry is access to good, balanced, reliable, and affordable datasets, preferably free and without restrictions.

Most training data are scraped and collected from the Internet and digital gadgets, usually without the owner's permission or knowledge. Data is such an essential commodity for machine-learning enthusiasts that they accept it as is and rarely pay attention to privacy, copyright, or ethical issues. In machine-learning communities, such as Kaggle, using unauthorized datasets and personal data is widely accepted (and even encouraged) as long as the model is sufficiently transformative, improves efficiency or accuracy, optimizes previous versions, generates profit, and does not interfere with corporate interests. It is not that these companies and researchers do not care about privacy or copyright material. In the public discourse, there is an assurance that privacy has been protected and content rights remain with their creators. However, in private circles, where machine learning and AI technology have been financed (e.g., Companies' Earning Calls) and developed (e.g., Kaggle), these concepts seem to be outdated and in the way of the inevitable "progress" expressed by digital technologies: a discursive trap of technological determinism.

Such is the importance of access to large datasets that these companies could make all their software open source and still maintain their dominant position due to ownership over data. For instance, while Google's powerful machine-learning software TensorFlow is open source and freely available on their website, Kaggle even gives away prizes and awards for data scientists and engineers to work on crowdsourced machine learning competitions. The technology necessary to transform any activity into data has become very cheap in the last few years, and the offer of computation power, storage, and analytics algorithms has increased exponentially. The proliferation of digital devices has opened up massive opportunities to collect data; new industries are emerging to extract data and surveil individuals in order to provide services, offer insights into consumer behaviour, control workers, and sell advertisers. There is a convergence of surveillance and profit-making in the digital economy, what Zuboff (2020) calls "surveillance capitalism."

2.2.2. Surveillance Capitalism and Modes of Living

Surveillance is not a new thread in Political Economy (Crary, 2014; Foucault, 1995, 2004; Fuchs, 2013; Lazzarato, 2002; Sanders, 2017; Zuboff, 2020). However, the fact that it can be automatized and remotely controlled through and by digital machines has made it possible for governments and private companies to monitor citizens'/consumers' activities on an unprecedented scale and with unprecedented speed and precision. Aspects of everyday life, once considered useless and without market value, are now extracted and crunched into valuable assets for the new immaterial economy. Zuboff (2020) refutes Srnicek's (2017) claim that this is an unintended consequence of past global economic crises. On the contrary, Zuboff affirms that this new phase of capitalism emerges from a "deeply intentional and highly consequential new logic of accumulation" (p. 75). At its core, this new logic of accumulation seeks to predict and modify human behaviour to increase profits and expand social and economic control. It not only carries an epistemological assumption that the world is knowable—and that this knowledge should be used for economic gain—but also blurs long-established social and institutional relations and values by making the extraction and use of everyday data an intrinsic part of commercial strategies. This logic of accumulation organizes perception and shapes the expression of technological affordances. Because it is taken for granted, and its assumptions are largely tacit, the power of digital platforms to shape the field of possibilities is largely invisible (Lazzarato, 2004; Zuboff, 2020). These platforms define the problems, the objectives, the metrics of success or failure, and the solutions. They determine how resources and people are organized, who is valued in what roles, what activities are undertaken, and for what purpose. In other words, the logic of accumulation, or the mode of production, as pointed out by Marxist thought, produces its own social relations together with its conceptions and uses of authority and power.

Zuboff (2015) implies that computing not only automatizes and replaces human labour but also "generates information that provides a deeper level of transparency to activities that had been either partially or completely opaque" (p. 76). Computer mediation symbolically renders events, objects, and processes that become visible, knowable, and shareable in new ways. Information and communication technologies have a unique capacity to automate and "informate," that is, to produce and shape (give form to) new types of information. Consequently, computer-mediated work produces a comprehensive "textualization": it datafies the work environment and extends the organization's control over it. It is difficult to find a single place or activity not driven by computer-mediated processes in the first decades of this century. Some are more formal, with clear objectives and goals, such as enterprise integration, employee monitoring, continuous improvements to logistics operations, mobile and temporary workforces, and marketing strategies. Others are ephemeral and mundane, such as the unceasing flow of emails, online searches, smartphone activities, and social media interactions. Consequently, the current form of capitalist accumulation is no longer based solely on the exhaustion of labour in the industrial sense but also on the exploitation of knowledge, life, health, leisure, culture, affects, habits, and behaviour. What digital platforms produce and sell goes beyond material and immaterial goods, including forms of communication, standards of socialization, perception, habits, and behaviour. It is about putting life to work, concludes Lazzarato (2004).

Within the contemporary context of digital infrastructures, capitalism is not only a mode of production but a production of modes of living (Lazzarato, 2004). The profound transformation within the capitalist conditions of production split the contemporary business organization into two distinct entities: the factory and the enterprise. This distinction is important because both entities operate at different fundamental levels: "the enterprise does not create its object (goods) but *the world within which the object exists* … the enterprise does not create its subjects (workers and consumers) but *the world within which the subject exists*" (Lazzarato, 2004, p. 188, emphasis in original). That is, the enterprise uses all its available resources (research, workers, machines, marketing) to create a world in which its products and services, together with the consumer and the worker, can exist and, at the same time, the world is in turn deeply inscribed in the individual's body and mind. In other words, "in the societies of control, the aim is no longer to *appropriate* as in societies of sovereignty, nor to *combine and increase the power of the forces* as in disciplinary societies, but to *create worlds*" (p. 202, emphasis in original). This proposition directly affects the modes of subjectivation and the forms of mediation promoted by algorithms in late capitalism, discussed in more detail in the next section.

Nonetheless, the late capitalist logic of accumulation has a narrow and short-sighted perspective when it comes to creating words. Lazzarato (2004) argues that the "different" styles of life promoted by enterprises are, in reality, a variation of the same: "the capitalist ways of life produce a homogenisation and not a singularisation of individualities" (p. 202). The *production of possibilities* is always controlled, predictable, and codified according to the modes of capital valourization. In other words, the modes of subjectivation produced by digital media "do not draw upon the infinity of monstrosities concealed within the human soul" (p. 202) but from Western patriarchal society expressed and caricatured in the cis-normative middle-class individualist white male. In that case, Zuboff (2015) is right in persistently asking questions regarding authority and power: if everyday life not only becomes a source of data extraction and analysis but is also subjected to personalization and customization due to continuous monitoring and experimentation by large tech companies, what roles are assigned and distributed in this new logic of accumulation? What narratives, discourses, and realities are possible? Who decides what is possible, moral, and ethical? What happens when authority fails? Who is responsible and accountable? These questions drive this research to inquire more deeply into how machine learning and predictive models compose new forms of subjectivation and how algorithmic mediations can be a political, affective, and performative force that (re)mediates individuals' experience.

2.3. Automated Subjectivation

To consider how automated algorithms following a specific political economy and a formal logic based on statistical models could be so influential as to redefine how an individual understands oneself and their experience of the world, I draw on the critical concept of subjectivation. To me, subjectivation refers to the processes whereby one's sense of self as an individual agent is paradoxically shaped according to forces external to the self. According to Althusser (1971), the concept of the *subject* is commonly understood as (1) a free subjectivity, a centre of initiatives, author of and responsible for its actions; and (2) a subjected being, who submits to a higher authority and is therefore stripped of all freedom except that of freely accepting his submission. That is, the sense of self originates from the individuals themselves but is also shaped by external forces. I am not saying that individuals do not have agency and are fully controlled by external forces. On the contrary, the individual sense of agency, that is, their freedom to act, is in itself a product of different psychological, behavioural, and social forces to which they are subjected. By freely acting in the world, subjects also (freely) reproduce specific gestures and actions they are subjected to. In fact, it is of the utmost importance that individuals be free to act since there are no subjects except by and for their subjection (Davidson, 2016; Foucault, 1982). This constructed sense of the individual in a network of social relations is what I refer to as the subject, and the processes comprising the production of the subjects are called subjectivation.

What constitutes a subject—the relationship between individuals and networks of social forces to which they are subjected—is by no means simple and clear. Indeed, subjectivation is one of the major concerns within disciplines like Sociology and Cultural Studies, which focus on the sense that various cultures make of "the individual" and the sense of the "self" that we, as individuals, experience. From the traditional humanistic ideas of the autonomous individual who possesses the determining force over the course of their life to the postmodern subject that is destabilized and susceptible to the transformation required of them by the power relations in the world they live, this debate has preoccupied thinkers such as Louis Althusser (1971) and Michel Foucault (1982; 1995), who work the concepts in terms of State power over citizens and forms of governmentality. Others, like Teresa de Lauretis (1987), Judith Butler (1990), and Simone Browne (2015), have added other complexities to the concept of subjectivation, including the power dynamic involving gender, sexuality, and race, as well as the notions of social norms, moral values, labour potential, and economic circumstances. More recently, Ganaele Langlois and Greg Elmer (2019) have considered the role of digital platforms in the production of the self. Drawing on all these authors, I aim to situate the process of subjectivation in the context of digital infrastructure, algorithmic media, and predictive models.

Althusser (1971) has examined the notion of subjectivation according to a traditional line of Marxism, indicating a "state machine" pervaded with ideological apparatuses that interpellate its subjects through an internal logic of time, labour, and capital. He argues that ideology provides "the imaginary relationship of individuals to their real conditions of existence" (p. 153). The law and the coercive power of the State constitute a concrete example: a policeman calling out "Hey, you there!" and an individual's recognition that it is he or she who is being called out makes the individual a subject. Insofar as that ideology, as conceptualized by Althusser, has no physical existence but has a material manifestation, it is possible to say that there is no ideology except by the subject and for the subject. In other words, "there is no ideology except for concrete subjects, and this destination for ideology is only made possible by the subject" (p. 160). Althusser thus suggests that we are always subjects, as we are constantly practicing rituals of ideological recognition. Through these acts, we become individual, distinguishable, and (naturally) irreplaceable subjects.

Althusser seems to imply a universal theory of subjectivation in which power follows a hierarchy and is assumed to come from a single inevitable source (i.e., the ideological State and its apparatuses). Michel Foucault (1995), on the other hand, rejects such a deterministic structural analysis and claims that there are different techniques of power, which are defused, decentralized, and naturalized in order to be effective in the production of the subject. Because the relations are decentralized and woven through everyday experience, Foucault questions the assumption that communication takes place between autonomous, self-aware individuals who use language to negotiate and organize community formation, arguing instead that this web of discourse practices and power relations produces subjects differentially suited to the contingencies of particular historical epochs. Foucault argues that discourse is a particular knowledge of the world that shapes how it is understood and how we act upon it. As such, discourse is not simply a repressive power that imposes rules and behaviours upon human beings. Instead, it produces the world as we understand it, which claims to be the absolute truth. In this way, "power is not exercised simply as an obligation or a prohibition on those who 'do not have it'; it invests them, is transmitted by them and through them" (Foucault, 1995, p. 27). Discourse is the powerful force through which individuals become subjects and their bodies become "socially useful," concludes Foucault.

So far, the concept of the subject and the process of subjectivation have been more concerned with the articulation of immaterial relations (ideology, discourse) that manifest in an individual as a particular subject. A feminist perspective, as illustrated by Teresa de Lauretis, Judith Butler, and Simone Browne, shows that processes of subjectivation can also be based on embodied experiences involving gender, sexuality, race, and the intersections among them that produce subjects that conform with a particular social structure. For instance, de Lauretis (1987) argues that feminist discourse creates the object of its investigation. Because of limited notions of gender and sexual differences (as binaries), and despite the effort to repeal the patriarchal system in which women are subjugated, a feminist perspective constructs theories about women in which the discourse itself has created the universal category of "woman," wherein all women come to be subjected. Similarly, Butler (1990) suggests that sexuality is culturally constructed, not "a cause" of sexual experience, behaviour and desire, but instead "an effect"; that is, "the production of a given regime of sexuality … seeks to regulate sexual experience by instating the discreet categories of sex as foundational and causal functions within any discursive account of sexuality" (p. 23). Considering the idea of an arbitrary connection between "cause" and "effect," Butler contends that social and cultural discursive practises aim to govern and regulate the gendered body, making it a subject in the process. On another front, Browne (2015) examines how certain surveillance technologies, particularly during slavery periods, were used "to monitor and track blackness as property … to anticipate the contemporary surveillance of racialized subjects" (p. 24). The colour of the skin becomes an attribute of segregation and hierarchization, where some individuals are not only subjected to specific modes of control and discipline but whose lives might not even have value within society.

In light of these authors, we can understand subjectivation as (social, political, economic, and cultural) processes that produce useful and productive subjects (individuals) within a narrow and specific interest. However, these processes interpellate individuals directly and indirectly but are not unidirectional or static. The process of subjectivation can be performed by different institutions to produce specific types of subjects, shaping the way individuals act, behave, think, and feel. At the same time, subjectivation can be self-applied by individuals to regulate themselves, to act and behave according to specific and acceptable rules, or to resist external forces by pushing the boundaries of what is acceptable. Hence, "agency"—that is, the ability and necessity to be a free subject in both the performance and resistance to subjectivation—is essential. The forces that produce subjects simultaneously subjugate an individual and affirm the individual's own being, creating their identities. Since subjectivation is dynamic, contentious, and unbalanced, these identities are not static or fixed in time and space. As such, concepts like intersectionality have tried to deal with overlapping subjectivation, arguing that the subject is not merely an abstract identity through which one expresses itself, but an embodied and concrete experience through which individuals live their lives.

2.3.1. From Biopolitics to Psychopolitics

Subjectivation is intrinsically connected with what Foucault (2004) calls governmentality, a governmental rationality that constructs social ways of being and forms of life—that is, social subjectivities. By defining government as "the conduct of conduct," Foucault (1982) links the technologies of domination with "technologies of the self," wherein individuals understand, shape, and develop themselves as docile, useful, and productive subjects. The logic of the body as the focal point where power relations are established and managed (biopower) forces individuals to emit signs, to signify their relation to social norms, producing subjects that participate in the ideologies of the society through cooperation with the desire to fit in and conform to social norms (biopolitics). Biopolitics is primarily concerned with population demography and the functions of the body to provide the object of regulation, that is, the parameters to govern populations. In other words, biopolitics fundamentally concerns the biological and physical, constituting a politic of the body in a broader sense. Hence Foucault claims that power exercised on the body is not a property but a strategy. Its effects of domination are attributed not to appropriation but to dispositions, tactics, and techniques.

However, biopower and biopolitics are not enough to describe the new instruments of governmentality based on large datasets, computational power, and statistical models employed to predict human behaviour and modulate how we act and think. The mind and the psyche have become prominent issues of governmentality, especially in the context of digital platforms: a new frontier to exploit individuals and shape subjects (Crary, 2014). Biopolitics is focused on the materialities of the body and population statistics; it has no access to the psychic realm. Han points out that "demography is not the same thing as psychography. It cannot tap into or disclose the psych" (p. 21). For that, Hans suggests a similar but more profound mode of control over the individual, which he calls psychopolitics.

Sometimes described as a separate concept (Han, 2017; Rau, 2013) but also argued as being an integral part of the Foucauldian concept of biopolitics (Prozorov, 2021; Stiegle, 2000), psychopolitics is focused on the individual's psyche as the pressure point for surveillance, discipline, and control. It is an instrument of governmentality whereby computational configurations, particularly predictive modelling, are deployed as "technologies of the self," constituting an instrument to prescribe how people should behave, act, feel, and think. The concept resembles Jeremy Bentham's dream of a panopticon (Foucault, 1995) as a new mode of obtaining "power of mind over mind." Bentham's model, however, is bound to a physical architecture based on optical perspectives with unavoidable blind spots. In the panopticon, the mind game is played through a centralized, always-visible tower in the centre that hosts an invisible power: the guards always see the inmates from a specific vantage point, and the inmates are never sure they are being watched. Though effective, the panopticon can only extract partial accounts of the prisoners' lives through surveillance of their bodies. Their thoughts, desires, and psychological inclinations remained obscure to "big brother" until the architecture got abstracted away by digital infrastructures. Digital technologies are aperspectival, enabling surveillance from any angle. Digital trackers are all around—inside, embedded, and circumscribed into mobile and environmental sensors. Instead of constraining the individual into a confined and obscured space, freedom of movement and transparency are the hallmarks of the digital panopticon. Power has no unique physical manifestation, as it is decentralized and distributed in space, though individuals always know now that they have been watched. On this account, the never-ending data mining and the large statistical models made possible by machine learning and big data are a way to extend control over populations. The subjectivity of individuals' psyche becomes programmatically involved and intertwined with predictive models, providing the means for establishing not just individual psychological signatures but a collective psychogram to manage and control populations, a phenomenon Stiegler (2000) calls psychopower.

Moreover, individuals freely and willingly relinquish their privacy by sharing data on digital platforms. They not only provide information about their presence in the world (pictures, geolocations) but all kinds of mannerisms, habits, and idiosyncrasies of their personal lives: favourite foods and music bands, political inclinations, obscure desires, incomplete thoughts, beliefs, and opinions about anything and everything. Once individuals are marked and identified, the data flows in a reverse course: it comes back to haunt the individuals as deceptively design-based mechanisms of influence that "hyper nudges" (Hull, 2018; Yeung, 2016) them into doing or thinking in particular ways. In other words, digital platforms interpellate individuals as subjects of the platform, affecting how they act in front of the pre-selected choices offered to them. They come as imperative calls for action: scroll, click, swipe, like, follow, buy, subscribe, react, watch, play, and share.

When collected on a massive scale, individuals' psychological attributes become a relevant factor in economic and marketing management, as well as in battles involving political and ideological agendas. With the knowledge of individuals' personalities and collective behaviours produced by predictive models, digital platforms hold power to weaponize and radicalize groups and manipulate the audience, posing a threat to an individual's well-being, freedom, and even life. Hence body and mind are both the necessary raw material for constructing and elaborating means of behavioural modification. The processes of subjectivation using psychological traits go beyond the materiality of the body in an attempt to tap into the immaterial nature and the subjectivity of the inner self. Our habits, behaviours, and sense of identity, but also our feelings, emotions, and affects, are not only the raw material to feed large-scale training sets for machine learning, but also the heart of a new form of governmentality put forward by the digital infrastructure and algorithmic media.

2.3.2. Impersonal Subjectivation

With the increasing privatization of State power and the accelerated penetration of digital technologies in society, social control has been shifting away from the rigid centralized institutions of subjectivation toward distributed sites of power where subjectivation is modulated, continuously changing from one moment to the other, in order to optimize markets and regulate life (Deleuze, 1992). Indeed, Foucault (2004) has demonstrated that neoliberal subjectivation depends upon social power and the ability to profit from individuals' lives. In this context, digital platforms and infrastructures have become essential instruments for producing subjects through different modes of subjectivation. Their sociotechnical infrastructure has "much more varied relationships to nonstate forms of authority and noncapitalist economies" (Bratton, 2016, p. 7). In recognizing the role of digital technologies, Langlois and Elmer (2019) note an intensification—in scale and speed—and redistribution of the process of subjectivation.

According to Langlois and Elmer (2019), the new political economy of subjectivation involves two sets of complex and intertwined opposing characteristics: (1) personal and impersonal (unique and distinct, yet banal and stereotypical) and (2) distributed and centralized (multi-sited yet concentrated in a few places). On the one hand, the traditional notion of an individual's uniqueness is centred on the individuals themselves. It is based on a logic that promotes a sense of oneself as a subject separate from others, but that fits into the dominant political rationality. For instance, in neoliberalism, the sites of subjectivation train and groom individuals to become useful and productive forces in society, encouraging and rewarding them for pursuing appropriate professional objectives and personal desires. As such, these individuals should not *forcefully become* successful economic agents. Instead, they *should want* to do so by themselves.

At the same time, this very sense of self, including one's aspirations, beliefs, and emotions, is articulated by external forces with interests that differ from the individuals, usually aiming to exploit them in order to carry out large-scale strategies to create markets and increase profit margins. In other words, it is "not the cultivation of persons that is the final goal of some of the major industries of subjectivation, but rather the mining and mobilization of subjective materials" (Langlois & Elmer, 2019, p. 237). This form of subjectivation puts the individuals at a distance; it sees them as a collection of amorphous data capable of objectively describing a person. In this case, the individuals are reduced to pieces of information that go beyond their social lives (professional choices, moral grounds, ideology) to include their embodied experiences (race, gender, sexuality), the biological rhythms of their bodies (heart rate, physiological need, sleep cycles), and psychological traits (attention span, desires, emotions). Langlois and Elmer (2019) call this "impersonal subjectivation."

It is important to note that impersonal structures are not a novelty; governments and large corporations have been managed by proxy for a long time (Resnick, 2005). Indeed, the mass media and the marketing industries have long illustrated that the mobilization of attention, desires, and affect has nothing to do with personal agency. Instead, this process has everything to do with what Lazzarato (2004) calls "conditions of existence," that is, the possibility of using subjective materials to reorganize life according to specific hierarchies of power. This process leads to the second point made by Langlois and Elmer (2019): subjectivation now involves multi-sited processes whereby the many dynamics of existence and life itself are mobilized and mined by a variety of social and economic actors and institutions to fulfill a neoliberal agenda.

The increase in privately owned media technologies directed toward attracting and capitalizing on individuals has led to a full expansion in scope and capacities of a political economy of subjectivation. Large conglomerates that emerged from and within blooming digital media (e.g., Google, Facebook, Amazon) are expanding from an enclosed platform model (Gillespie, 2018) toward a distributed and impersonal infrastructure. For Langlois and Elmer (2019), the platform model is user-centric and focused on ideas of self-determination and the ability to transform oneself freely. This model is experienced at the interface level, where the user can only access curated and highly personalized information. Digital platforms actively cultivate the creation and exchange of user-generated content, whether personal (pictures, posts) or contextual (purchasing patterns, viewing habits). At the same time, while individuals experience subjectivation at the front end of a centralized app, the backend processes connect with third-party systems (other platforms, advertisement agencies, security firms, and governments) seeking to capitalize upon impersonal relationships, of which, of course, Kaggle is an example and an important site of impersonal subjectivation. The process of subjectivation at the level of the infrastructure datafies individuals—reducing them into quantifiable metrics—and distribute the data among and between many platforms.

The political economy of impersonal subjectivation is built on the orchestration of the relationships among multiple data points, in which the objective is to correlate the personal data to other data points such as user habits, market demands, aggregated audience/consumer profiles, socioeconomic clusters, and other impersonal categories in order to achieve specific goals (e.g., increasing sales of a specific commodity). Digital infrastructures push for the production of impersonal content, such as ambient data (location, mobility, checkpoints, sounds, images) and biometrics (heart rate, eyesight, calories burnt, fertile cycles), not only as a way to prospect the creation of new markets but also to disrupt mature and well-established ones. Examples of this outward movement from enclosed digital platforms on the Internet can be found in the substantial investments in smart cars (Alphabet, Uber, Tesla), groceries and food delivery (Amazon, Instacart), all kinds of conversational "chatbots" assistance (OpenAI, Facebook, SAP, Google, Apple), and the dream to "fix" or "tame" urban life in the so-called "smart cities" (Alphabet Sidewalk, Cisco). Such new markets and commodities correspond to the "infrastructuralization" of digital and social media platforms, highlighting the increasingly contradictory ways of extracting value from—while also (de)constituting—the subject (Langlois & Elmer, 2019).

Digital infrastructures attract people through a discourse of rich communicative experience and immersive mediation of everyday life. Their main goal, however, is to employ algorithms to collect data, undertake large-scale surveillance and microanalysis of user profiles, and contextualize and link multiple data points to display personalized advertisements and sell personal and impersonal data for profit. Napoli (2014) suggests that algorithmic media are the "base structures and parameters that regulate the production, distribution, and consumption of content" (p. 343) in the twenty-first century. With the power to automatically sort and filter data at speed, assign meaningfulness and priorities, and manage how people perceive information, Gillespie (2014) claims that the conjunction of media and algorithms is "now a key logic governing the flows of information on which we depend" (p. 167). That is, algorithmic media not only assist people in finding information but also in providing the means to know what there is to know and how to know it, to take part in social and political discourse, to familiarize ourselves with the public in which we participate, and to manage and shape our social habits. For instance, search engines use specific algorithms to tailor quick and "meaningful" results from users' queries (Bozdag, 2013). Social media platforms are also perpetually sorting users' content according to a pre-established set of ranking systems (Bucher, 2012), while video streaming services are constantly surveilling content looking for copyright infringement (Dayal, 2012), and online stores track, record, and compare user behaviour in order to recommend the consumer's next purchase (Finn, 2017).

Algorithmic media are the intermediaries—the mediators—that allow for connections between users and non-users, humans and nonhumans, as well as between them and third-party products and services. They provide access to relevant users' data, influence moods and attention, and allow third parties to collect user data and link to other datasets to create their own large-scale manipulation campaigns (popularly known as Fake News). The recent expansions of "smart" services (mobility, urban planning, health and fitness, content moderation, notification, and recommendation) reveal a project to establish algorithmic media as the unavoidable mediators for all aspects of life "with a capacity to act simultaneously at the molar level of large-scale social shaping of attitudes and habits and the molecular level of personalized targeting of users" (Langlois

& Elmer, 2019, p. 245). The tracking of heartbeats from a smartwatch, for instance, requires little effort and attention from the user but demands a preconscious routine habit of keeping the smartwatch on, which Wendy Chun (2017) calls “habitual media.”

With the increasing array of networked digital apparatuses—all kinds of sensors, mobile and wearable devices—and a sprawling ecology of algorithmic media, we have no control over our own personal data within digital infrastructures. Langlois and Elmer (2019) claim that this new (impersonal and distributed) process of subjectivation “invokes a politics beyond the individual user, calling into question the integration of non-users across a large infrastructure connecting heterogeneous systems and networks” (p. 239), which also includes actions performed and data generated from preconscious actors, nonhumans, and the environment (e.g., automated machines, AI, food crops, and pollution). The amalgamation of algorithms and data, particularly the ones produced by machine-learning techniques, massive datasets, media, and economic interests, not only points to an acceleration of impersonal subjectivation but also opens the possibility for new ways to understand the relations between humans, nonhumans, and the environment. More than simply shedding light on how algorithmic media modulate individual behaviour, my research also asks that we shift our notion of mediation to include nonhumans as agents of actions that affect and transform our experience of the world.

2.4. Mediations

My approach to researching digital infrastructure and AI is built upon the concept of radical mediation, wherein I understand these systems as assemblages of humans and nonhumans. To understand the complexity of computational-centred networks comprising machines, data, code, users, affect, and political-economic interests, we must consider the interactions among these actors and how different parts of the network influence, modulate, and interfere with each other. The concept of mediation has a long history spanning the entire Western philosophical tradition, from Aristotle to Marx (Grusin, 2015; Williams, 2015). It was subsequently captured by media technologies in the twentieth century, becoming predominantly attached to communication and primarily connected to mass media. More recently, the concept has been revisited and updated by scholars like Richard Grusin (2015) and Sarah Kember and Joanna Zylinska (2014) to consider the implications advanced by new digital technologies, as well as to acknowledge the role of nonhumans in the process of mediation. As discussed above, the recent datafication phenomenon brought forward by digital infrastructures shifts the power relations and processes of subjectivation, complicating the concept of mediation. Hence I use the concept of mediation to describe what happens during and in the midst of the interactions among users, machines, algorithms, and developers, producing the conditions of life and affecting our experience of and relationship with the world.

Mediation is a complex and slippery term. It has been used as a key term in several modern thought systems to describe different operations and processes. Traditionally, *mediation* is understood to come in between, or in the middle of, already pre-formed, preexistent subjects or objects. That is, mediation was defined as simply a function of an interface or tool that reconciles or intermediates interactions between subjects. In *Keywords*, Williams (2015) describes early common uses of the concept of mediation:

> (1) the political sense of intermediary action designed to bring about reconciliation or agreement; (2) the dualist sense, of an activity which expresses, either indirectly or deviously and misleadingly (and thus often in a falsely reconciling way), a relationship between otherwise separated facts and actions and experiences. (p. 154)

Both perspectives are firmly tied to the idea of reconciliation between Spirit and World, Subject and Object. They describe the interaction of two opposing elements or forces, in which the first conceptualization can be connected to the common practice of law, such as reconciliation among conflicting or adversarial parts in a litigation process. The second, often used in relation to ideology critique, serves as an agent of epistemological correlation that prevents a "true" understanding of reality (Grusin, 2015). In the late nineteenth and early twentieth centuries, the concept of mediation, particularly the second definition Williams (2015) posed, was reduced to and focused on particular media technologies, especially in relation to the social effects of mass media. In this sense, mediation has a closed connection with Marxist theory and, therefore, is understood as a dialectical process, where it refers to the reconciliation of two opposing and irresolvable contradictory forces within a given society—a mediating factor of a given culture that takes the form of the medium of communication itself (Kember & Zylinska, 2014). This notion of mediation is typically used in Media Studies, where certain social actors are seen as indirectly but deliberately interposed between reality and social consciousness.

In many ways, Baudrillard (1983) was alluding to the mediation power of mass media in his "Simulacra & Simulation" theory. However, this approach to mediation depends on the assumption that reality and simulation, or consciousness and unconsciousness, are opposites, wherein mediation acts between them, resolving their contradictions. In this case, mediation is primarily a tool for enacting a particular socio-political goal. Such a narrow definition of mediation is insufficient to explain current modes of subjectivation because it reduces mediation to some (material) form of media technology and assumes that there are stable (and static) structures in a society where mediation is mobilized as a third intervening and negotiating factor. Kember and Zylinska (2014) break with this logic by proposing that "mediation is the originary process of media emergence, with media being seen as (ongoing) stabilizations of the media flow" (p. 21). In this instance, media are temporary "fixings" of technological and other forms of becoming, making it impossible to speak about media in isolation without considering the process of mediation that enables such "fixings."

Looking at the interrelationship of technical, social, and biological processes of mediation, Kember and Zylinska (2014) argue that "life itself under certain circumstances becomes articulated as a medium that is subject to the same mechanism of reproduction, transformation, flattening, and patenting that other media forms … underwent previously" (p. xiii). That is, life is to be perceived as a medium, in which, in this case, we must critically examine the complex and dynamic processes of mediation operating at the biological, social, technical, and political levels in the world. By separating media (objects) from mediation (process), Kember and Zylinska aim to clarify the relationship between them, suggesting that mediation is not simply an interface, a transparent layer that intermediates entities. Rather, it is all-encompassing and indivisible, an intrinsic condition of *being-in* and *becoming-with* the technological world. Mediation is a complex and hybrid process or event, simultaneously economic, social, and cultural, that produces and provides the conditions for the emergence of subjects and objects (Grusin, 2015; Kember & Zylinska, 2014).

2.4.1. Radical Mediation

Kember and Zylinska resonate with what Grusin (2015) calls "radical mediation," though they mainly discuss the concept in terms of art and communicative media as these are traditionally understood. Grusin takes a step forward, offering a different path, if not a radical turn, in conceptualizing mediation. According to Grusin, mediation cannot be limited to the means of communication, representation, or the arts. Instead, it should be understood as a fundamental process of human and nonhuman existence. While media technologies have always operated at the epistemological and cognitive levels as modes of knowledge production, "they also function technically, bodily, and materially to generate and modulate individual and collective affective moods or structures of feeling among assemblages of humans and nonhumans" (p. 125). Grusin's arguments resemble Lazzarato's (2004) discussion of "immaterial labour" in the post-Fordist society, where he writes that organizations produce and sell not only material or immaterial goods, but also "forms of communication, standards of socialisation, perception, education, housing, transportation" (p. 205). Therefore, mediation operates at two interdependent levels: at the cognitive level, where immaterial knowledge (culture, arts, politics, ideologies) circulates and exerts power, and in the embodied and material world, where particles, objects, subjects, and events take place.

Grusin (2015) argues that mediation is "the process, action, or event that generates or provides the conditions for the emergence of subjects and objects, for the individuation of entities within the world" (p. 128). In other words, mediations are events where elements are in constant transformation through a process of becoming to generate individual subjects and objects. Following this assertion, we can argue that digital media is an affective and performative force that (re)mediates material experience within the world. This concept fits the current mediascape dominated by digital infrastructure and artificial intelligence. Predictive models created with machine learning algorithms and the individuals they interpellate, for instance, do not preexist in isolation but become subjects and objects of an event through mediation processes. Even their roles as subject and object are not predefined and static. The vectors of actions, data flow, and control switch back and forth, affecting both sides as they interact with each other. In this sense, mediation must be understood ontologically as a process, or an event, prior to and not reducible to particular media technologies. Mediation is not just about connecting experiences but also generating, transforming, and modulating them. In other words, mediations are always remediations (Bolter & Grusin, 2000) that change, translate, relate, and re-connect experiences. Similar to Kember and Zylinska, Grusin (2015) argues that "all bodies (whether human or nonhuman) are fundamentally media and life itself is a form of mediation" (p. 132). Grusin's "radical mediation" operates physically and materially as an object, event, or process in the world, impacting humans and nonhumans alike, which in some sense might also be understood as a more-than-human, or even inhuman, process of mediation.

When we consider the extent to which digital technology has come to be environmental, distributed, and autonomous—such as the large-scale use of sensors and mobile devices, massive cloud computing services, and the proliferating use of artificial intelligence, machine learning algorithms, and predictive models connected through networked digital platforms—the reconceptualization of the term "mediation" advanced by Grusin, Kember, and Zylinska becomes crucial to understanding the impact and implications of algorithmic media and digital infrastructures in the twenty-first century. Because digital infrastructures invoke impersonal subjects and nonhuman agents (Langlois & Elmer, 2019), this reflects Grusin's proposition for a radical mediation, whereby nonhumans also become agents of action. For Kember and Zylinska (2014), it is not simply the case that we (humans) live in a complex technological environment that we can use, interact with, and control. Instead, "we are—physically and hence ontologically—part of that technological environment, and it makes no more sense to talk of *us* using *it*, than it does of *it* using *us*" (p. 13, emphasis in original). We must then recognize that we are not entirely distinct from our technological tools: "As we modify and extend 'our' technologies and 'our' media, we modify and extend ourselves and our environments" (p. 13). Digital technologies are not just tools or intermediaries through which we accomplish some action; they have become part of us.

Our relationality and entanglement with nonhuman entities continue to intensify with the ever more corporeal, ever more intimate dispersal of media and technologies into our social spaces. In this sense, instead of examining algorithmic media as neutral third parties or passive media objects—a social media platform that "facilitates" content sharing, a digital service that "enables" you to buy groceries more efficiently, or a mobile device that "allows" users to see how their faces would look like when they get older—my research considers the continuous process of mediation that occurs among the users, the algorithm-machines, and their contexts. The question becomes how entities influence and interfere with each other, such as in the case of the reflexive phenomenon inherent in much algorithmically driven media consumption. Facebook, for instance, monitors users to find the most salient behavioural patterns. Once these patterns are found, they are fed back to the users via Facebook's newsfeed. Consequently, even more users will apparently act in the way that the algorithm predicts (Bucher, 2012). From words and images, combined with sensors and programmable nudges, AI and algorithmic media create a complex world. However, its fundamental logic remains hidden from most users, locked in complex black box systems and commercial secrecy.

Algorithmically mediated relations attempt to deceive natural processes through technology. We produce modes and models of life with such ingenuity that the world seems to correspond to these conceptual systems: the territory no longer precedes the map, wrote Baudrillard (1983), alluding to Jorge Luis Borges' short story "On Exactitude in Science." Digital media seems capable of describing, explaining, and even predicting our sensorial experiences. Their codes (algorithms) produce alternative versions of the so-called reality, parallel worlds, and multiple experiences of the "here and now," which persuade, touch, and become "real" as we collectively believe in their effectiveness. Literature, photography, and film might be the earliest and most exquisite examples of this mediation phenomenon. New technologies, however, tend to blur our embodied experiences by simulating other possibilities to satisfy our desire to experience something real and direct without the mediation of a technological apparatus. The simulacra of reality are presented in the narrative of current events, distributed by the mass media (Baudrillard, 1983), social media (Pariser, 2012), and digital platforms (Gillespie, 2018), mediating our experience of the physical space through smartphones (Farman, 2011) and smart cities (Shepard, 2011), or cyberspaces shaped by chatrooms (Dibbell, 1993), videogames (Bogost, 2007), metaverses (Patel, 2022), and the uncanniness of deepfakes (Maddocks, 2020). It is a phenomenon deeply rooted in a particular political economy of subjectivation intended to regulate the "production of possibilities" and uphold particular socioeconomic relations. It establishes the conditions of existence that dictate the allocation of resources and people, as well as who is valued for what roles, what is done, and for what reason.

2.5. Summary

What characterizes contemporary society is the power of algorithms to organize and mediate aspects of everyday life. Increasingly, the decisions that affect our lives are being made by and through digital systems using statistical models, computational instructions, and large datasets. The theoretical framework proposed in this chapter aims to shed light on three intricate aspects of algorithmic media and digital infrastructures: political economy, subjectivation, and mediation. The process of datafication brought forward by digital platforms calls into question traditional economic categories and the capacity of capitalist economies to control them. The networked economy shifts the value of goods and subjects in a world of digital networks. In particular, the worth of a product, a service, or even a user increases when others buy or interact with the same goods, services and users (Srnicek, 2017). The pressing issue for political economists is whether the networked economy driven by algorithmic media implies a difference in degree or kind (Mosco, 2008). That is, whether the recent changes in the economy imply a continuation or disruption of the traditional tendencies of capitalism, which also emphasizes the importance of studying new forms of subjectivation and mediation that modulate individuals' conditions of existence (Lazzarato, 2004).

The question is, in turn, what kinds of mobilization of users, what forms of subjectivation take place in distributed infrastructuralized digital media corporations, and for what purposes is this done. In a networked digital infrastructure, users do not control their own data, being subjected to all kinds of exploitations from multiple sites of subjectivation. The concept of impersonal subjectivation (Langlois & Elmer, 2019) becomes useful to unpack the different ways data has been sensed, collected, encoded, and transmitted by and through these infrastructures to exploit new forms of economic gains, such as the datafication and commodification of human behaviour, attention, and habits. More specific to this research, it allows us to ask how code, data, digital infrastructures, and crowdsourced labour are mobilized to create instruments of control that shape individuals' behaviours and produce subjects. Alternatively, this dissertation asks how algorithmic mediation produces subjects and reorganizes life.

Kember and Zylinska (2014) and Grusin (2015) situate these individual subjects as nodes on a network of interactions and possibilities. We must recognize that we (humans) are not exceptional and that the production of reality is shared with other entities, both natural beings and other agents from a different "nature" that we help create and shape. Living in symbioses with our machines, we are obliged to give up the possibility of control, because even the notions of the "self" (who we are) are transformed and reshaped. Here, Grusin's (2015) concept of "radical mediation" helps make sense of how twenty-first century media operate within the world as objects or events no different from any other. This perspective raises questions not only regarding the independence of individuals who are free to act and to think by themselves but also the autonomy of the so-called "smart" technologies and whether the current technological shift creates a new social actor—a networked machine capable of (re)acting autonomously, independent of human intervention. If so, we should ask how digital media, hardwired with ideological, political, and economic preconceptions and full of social and cultural biases, affect everyday life.

3. Methodology

> The CNN [Convolutional Neural Network] learns on its own which locations correspond to which labels without any human guidance … *It may feel like black magic at first but it works and it works pretty well*.
>
> — Moejoe (Shayan), 10th place in Kaggle's Passenger Screening Algorithm Challenge

Studying algorithms is not a simple task. Algorithms live in-between states: they are both concrete material objects—static written pieces of code stored in hardware made of mined materials— and dynamic immaterial forces that affect life on this planet. Some algorithms may be simple, but they are often interconnected and interwoven with other algorithms, producing a long, if not never-ending chaining process where one algorithm feeds the next. These systems grew more prominent in the last decades, becoming a central nervous system of the capitalist society, which, by design, has a tendency to hide and protect algorithms as commercial secrets. The sociotechnical attributes of algorithms and political-economic forces acting within and around them make these digital systems almost impenetrable—a black box, or "black magic," as described above by a Kaggle user. Indeed, the difficulties involved in investigating these black-box machines have become a barrier not only to researchers interested in the social-political implications of digital technologies but also to government officials and policy-makers pushing for best practices and regulatory marks.

Yet, as a concrete object and sociotechnical construct, algorithms media are not out of reach. We can use different methods and perspectives to approach algorithms as an object of inquiry. For instance, Kitchin (2017) suggests we can observe algorithms in the wild, taking into account how they affect, and are affected by, users and the collective use of digital platforms, as well as by the ecology of other digital systems linked to them and to other social-political events. This could be done from several traditional perspectives (technical, sociological, legal, philosophical), as well as from a "code/software studies' perspective that studies the politics and power embedded in algorithms, their framing within a wider socio-technical assemblage and how they reshape particular domains" (p. 20). Nonetheless, scholarly work following this approach tends to focus on already established large enterprises with well-known public faces, like Facebook, Apple, or Amazon, overlooking smaller, niche, but no less critical, platforms like Kaggle.

Kaggle is a digital platform for machine learning competitions and big data consulting with a track record of solving real-world problems across an array of industries, including life sciences, financial services, aviation, information technology, and retail. Founded in 2010 and bought by Alphabet (Google's parent company) in 2017, it hosts the largest data scientist community: more than five million registered users across 194 countries (Crunchbase, 2019). Through the platform, companies can launch challenges that reward money; hence users compete with one another to solve complex data science problems using the latest machine-learning algorithms. The platform doubles as a data repository, storing more than 20,000 datasets for machine-learning training, such as search engine results, crime data, commodity prices, credit and financial fraud, social media interactions, and all kinds of sensorial and demographic data. A crucial player in machine-learning development, Kaggle serves in this research as both an object of inquiry and the primary source of material, where I examine its history and economic model, how the platform operates, and the community around it.

This research is primarily interested in the ways algorithms mediate individuals' sense of self and how AI have been developed as a new form of subjectivation. That is, I am interested in how code, data, digital infrastructures, and crowdsourced labour are mobilized to create instruments of control that shape individuals' behaviours and produce subjects. Algorithmic mediation points toward the possibility of using subjective and impersonal materials to reorganize life in its broadest sense according to specific hierarchies of power. This research's central question is *How algorithmic media produce subjects and reorganize life*. To answer this rather broad inquiry about algorithms and predictive models, I propose the following research questions:

- *How do personal and impersonal data become the raw material that fuels machine-learning algorithms?*

- *How are code, data, and political economy interests mobilized to create valued predictive models?*

- *How are predictive algorithms used to intervene and modulate individual behaviour?*

Addressing these questions requires a mixed-method approach, combining traditional methodological perspectives in Media Studies, such as Qualitative Content Analysis and Discourse Analysis, with new methodological approaches focusing on new media and digital platforms, such as Digital Methods and Software Studies. The following chapters discusses these methodologies and are used in this research to collect, manage, and analyze the source material.

First, I define the timeframe considered in this dissertation, followed by a list of the primary sources used in this research. I then explain how I systematically employed digital methods and qualitative content analysis to gather and classify the corpus for this research. Next, I discuss how I used Software Studies in conjunction with Discourse Analysis to unpack how machine-learning development on Kaggle helps construct and mediate specific sociotechnical realities. Lastly, I disclose my identity as a researcher and my position in relation to the topics discussed in this dissertation.

3.1. Timeframe

This research examines Kaggle, both as a company and a platform, and delves into the machine-learning competitions it hosts. The timeframe considered here spans from Kaggle's inception in April 2010 to the data collection period in November 2020 for this research, ensuring a thorough analysis of the platform's evolution and impact.

This is an important period in the history of AI due to several converging factors evolving in the mid-2000s, such as technological advances in machine learning techniques (e.g., deep learning), the exponential data availability (freely accessible) made possible by Web 2.0 technologies, social media platforms, and cloud storage, the massive adoption of sensed-ready mobile technologies such as smartphones, and the push for a more controllable built environment such as smart cities and the Internet of Things. In its first ten years, Kaggle helped the AI industry expand and consolidate machine learning as a mainstream practice and served as a laboratory for ideation and workforce formation.

The time frame has some flexibility at both ends as a way (1) to discuss and clarify the conditions in which Kaggle was born as well as how machine learning became mainstream in technological development and central to the current political economy; and (2) to discuss the impact of predictive models on our society.

3.2. Source Material

The inquiries of this research rely on a diverse set of material sources, most publicly and widely available. However, a considerable part was unearthed with the help of digital tools developed in the context of this research. The primary sources of material are the following:

Kaggle

As the object of this research, Kaggle is the main source. I examine the company's announcements and reports, the platform's documentation, Terms of Services (ToS), End-User License Agreements (EULA), community guidelines, and various kinds of information made public through the website, such as featured competitions, user rankings, public APIs, and blog posts, to name a few. Using a scraper, I also gathered metadata information about the competitions hosted on the platform, the datasets repository, the user rankings, as well as the messages exchanged in selected Kaggle competition public forums.

Internet archive

The Internet archive provides access to content once available online but no longer accessible through the original URL. Here, I accessed the archived version of Kaggle's defunct blog, "No Free Hunch Blog," and screen-captured Kaggle's homepage from selected moments in time (from 2010 to 2021), most notably when it underwent significant design updates.

Financial reports

Kaggle was an independent private company up until it was acquired by Alphabet in 2017. While Kaggle itself did not release any financial report, it is possible to gather some data about the company on financial news aggregator websites like Crunchbase. I also looked at Google's and Alphabet's Earning Call Transcripts and Financial Annual Reports from 2014 to 2020, which are publicly available on their websites. These documents contain inside information about the interests and plans for Kaggle as a Google subsidiary and the company's plans for AI and machine learning.

Kaggle's founders, investors, and other prominent figures

Material about, produced by, or related to the main actors involved in the platform provided this research with rich contextual information. The main figures include Anthony Goldbloom and Ben

Hamner (co-founders), Jeremy Howard (Kaggle's first president), Hal Varian and Max Levchin (earlier investors), Sundar Pichai (Alphabet's CEO), as well as competition sponsors and prize winners. The material was researched and found in a variety of sources: news media reports (Reuters, New York Times), specialized media (e.g., Wired Magazine, TechCrunch), interviews in independent media (e.g., Chai Time Data Science Podcast, Software Engineering Daily), promotional videos (e.g., World Economic Forum), blog posts (e.g., Alphabet, personal blogs), talks and presentations (e.g., at Scale by The Bay, Strata Conference), to name a few.

Paratexts related to selected competitions and machine-learning algorithms at large

These include contextual information that helps explain and unpack technical apparatuses (algorithms, datasets, external code libraries), machine-learning strategies, and decision-making processes (which algorithm to use, how to clean, aggregate, reduce, and find useful attributes in a dataset), and relations of power among competitors, the platform, and the selected competition's sponsors. These comprise white papers, academic articles, blog posts, code repositories, and data provided on the competition's website and through the sponsors' main communication channels (e.g., website, official communication, blog posts).

3.3. Data Gathering

This research began at the public version of Kaggle's homepage. From this single focal point, I expand outward (articles, financial reports, public communication, blog posts) and inward (dataset descriptions, user replies on competition forums, documentation, lines of code), looking for clues about how the platform works, who owns it, how it operates, its economic model, who participates in it, who are the main actors, and what happens in the platform. As a web of hyperlinks, each piece I collected led me to different, sometimes divergent, paths, where I would find more material to be analyzed (when not prevented by paywalls or "404" pages).

Traditional textual resources—official documents, financial reports, surveys, news articles, presentations, interviews, blog posts, Web forums, and technical documentation—and audiovisual material were gathered directly from official channels, news outlets, independent media, or social media channels. I used Web search engines (mainly Google, but not exclusively) to find these materials. The search terms I used were variations and combinations of terms related to Kaggle, such as, but not limited to, "Kaggle" itself, their founders and key investors (e.g., "Anthony Goldbloom"), companies and agencies involved with Kaggle (e.g., "Google," "Alphabet," "Nasa"), terms related to Kaggle (e.g., "Neural Networks," "Deep Learning"), and high-profile competitions (e.g., "Facebook Deepfake Detection," "Passenger Screening Algorithm Challenge "). I also found myself searching for more obscure themes that emerged from the material, such as "hyper nudge and dark psychology," "predictive models and Eugenics," and "deepfake pornography." In total, I collected and examined 257 source materials.

Non-traditional resources—digital elements scattered on Kaggle's platform, metadata about competitions, datasets, and the userbase—were collected using a Digital Methods approach, more specifically, a Web scraper, explained in more detail below. I collected metadata from 3,437 competitions, 55,897 datasets, and 9,164 users on the public ranking page. As I will explain later, this research focuses on public competitions that offered over US$ 1,000 in prizes (n = 284) and closely examines three of them. I also collected and analyzed user conversations on three competitions' public forums, totaling 11,047 replies on 1,268 threads. These sources were used to track Kaggle's history both as a company and as a platform and to explore the discourse of AI embedded in the data science community. Lastly, part of the "raw" data from the Web scraper was classified into machine-learning competition modalities using Qualitative Content Analysis, which served as the starting point for a deeper discussion of predictive models and subjectivation.

3.3.1. Digital Methods

Rogers (2009) points out that there is an ontological distinction between natively digital objects and digitized objects. While scanned documents, libraries, online surveys, and videoconference interviews have migrated to new media platforms, tweets, emojis, Massive Multiplayer Online Games (MMOs), social media, folksonomy, and recommendation systems were born within the logics of the Internet. As a result of the fast digitization of every aspect of society in the last few decades, researchers were forced to consider ways to collect, examine, and make sense of these new data formats. While traditional methods were adapted or "virtualized" to new media in order to bridge the gap, they mainly account for traditional objects now "virtualized" in digital format. Not only do virtualized methods fail to account for new kinds of objects, but they also lack the tools and conceptualization to follow the scale and complexity of digital systems. Digital Methods is a response to this divide and the limitations of the virtualized traditional methods.

Digital Methods aim to study sociotechnical practices inscribed in digital objects (Rogers, 2009). This is a way of learning from the methods of the medium, requiring innovative thinking to both apprehend the logic of digital media and repurpose them for social and cultural research (Rogers, 2009, p. 1). Digital Methods are often defined as heuristics and techniques for identifying, capturing, managing, analyzing, and theorizing materialities, social practices, and implications of sociotechnical assemblages associated with digital technologies (Leszczynski, 2018a). While Digital Methods can be used to study both digitized objects and natively digital objects, this approach tends to focus more on native digital material. It has been used to follow and study traces captured through and produced by digital infrastructures. However, depending on the goal to be reached, digital methods can, and in many ways must, be combined with more traditional approaches in social sciences, such as content analysis, discourse-analytic, ethnographic, performative, and hermeneutic approaches (Leszczynski, 2018b).

Recent studies adopting this approach have focused on large online infrastructures, mainly social media platforms like Facebook, Twitter, and Wikipedia. For instance, posts published on blogs, threads of tweets, hyperlinks connecting websites, the dynamics of a massive number of people playing games together, the relationships created through social media platforms, the particularities of subreddit communities, and the new forms of individual expression (avatars, emojis, memes, selfies). Concrete examples can be found in the methods used for mapping controversies developed by Tommaso Venturini (2015) and his colleagues at the CRN Centre for Internet and Society, the network visualizations used to study political discourse on social media produced by Fabio Malini (2014) and his team at the Image and Culture Studies Lab (Labic-UFES), the collection of text analysis tools developed under the leadership of Geoffrey Rockwell (2010) and Stan Ruecker (2012), Bernhard Rieder's (2013) algorithms used to extract data from Facebook, and Rogers's (2017) Web Historiography approach to creating "biographies" of digital objects using archived versions of its homepage. These initiatives are, to some extent, rooted in Digital Methods and are useful examples and sources of inspiration for my research. They unveil the specificities of new media, their platform vernaculars and users' cultures, and the different ways they can affect individuals, collectives, and the environment.

Similarly, my research focuses on Kaggle—self-titled "the world's largest data science platform." The platform was conceived as a gamified environment for crowdsourced competitive machine-learning challenges. It doubles as a dataset repository—unsurprisingly dubbed "the world's machine-learning repository," and, up until recently, a temp job agency. In this sense, I use the Digital Method approach primarily as an instrument to capture, identify, and make sense of the platform. For instance, how many and what kinds of competitions Kaggle has held; who are the main sponsors and how much money are they willing to pay in these competitions; how many and what types of datasets are stored on its repository; who contributes to the repository; how many users are registered in the platform and how many participate in the competition; what motivates users to compete; and what are goals and interests behind these competitions.

3.3.1.1. Web Scraping

With little more than 5 million users, Kaggle (2020b) is not comparable in size to other digital platforms such as Twitter, Facebook, and YouTube. Yet, with over three thousand competitions and more than 50,000 datasets, each with its own secondary pages, rankings, and discussion forums, the amount of data is not negligible when handled by a single researcher or even by a large team. Collecting, identifying, managing, and making sense of such "big data" requires inventiveness and some automation. One of the easiest ways to retrieve information from digital platforms is using their Application Programming Interface (API), if available. For instance, "X" (formerly Twitter) used to allow users, marketers, and researchers to have programmatic access to their API for all kinds of purposes, from data collection to direct interaction with the platforms (Twitter, 2021). Fortunately, Kaggle maintains a dedicated public API through which it is possible to upload and update datasets (Kaggle, 2021d). Competitors on Kaggle also rely on the API to submit their code as fast as possible—timing is crucial in competitions. Though it is possible to use the API to retrieve some information about datasets and competitions, the results are limited. It was designed to facilitate integration in the competitions and datasets workflow, not to reveal the details about each competition, dataset, or user in the platform. As a consequence, I turned my attention to a popular data mining technique frequently used in machine-learning development: Web scraping.

Web scraping, or simply scraping, is a method used for extracting data from websites. It is a form of data mining for copying specific information from a webpage into a central local database or spreadsheet for later retrieval or analysis. Scraping can be used to gather contact information (e.g., names, emails), monitor page updates (e.g., online price change, price comparison, product review), analytics (e.g., website change detection, user interactions), Web mashup and integration (e.g., gathering real estate listings, weather data monitoring), and research (e.g., tracking online presence and reputation, content creation, network relationships).

A webpage is built with a hyper-text markup language (HTML) and often contains an abundance of valuable data in textual form, sometimes hidden in plain sight in the attributes of its structure. Visually, a webpage is rendered for human interaction; structurally, it comprises an intricate tree-like hierarchy that is increasingly built by machines and for machines. Collecting data from the Web can be done manually by systematically navigating a webpage, copying specific information, and pasting it into a database. However, this approach is only attainable for a small amount of data to be collected (dozens of data points on a couple of pages), which would not justify the time it takes to configure or develop a Web scraper. Automating the process using out-of-the-box software or customized scripts might be necessary to collect dozens of data points from thousands of webpages. Specialized tools and software for scraping webpages range from ready-to-use tools and services for people without coding skills to supporting libraries for specific programming languages.[2]

Scraping is a controversial method. It is considered a grey area in terms of ethics and legal provisions. While the information on a webpage may be publicly available, digital platforms like Kaggle (and any other business-oriented platform) try to prevent scraping for several reasons, such as user privacy and security, copyright protection, or ownership. Digital platforms' Terms of Service (ToS) and End-user License Agreements (EULA) have specific clauses to discipline and punish the use of scraping, clearly stating that data collection without prior permission constitutes a violation of the terms (see Kaggle, 2020e). Those terms, which are unilaterally set by each platform, and which might change at any time, often prohibit users, researchers, and journalists in particular, from collecting data for research and auditing purposes. In the United States, the Computer Fraud and Abuse Act (CFAA) was systematically used to penalize those who infringed digital platforms' ToS and EULAs (The Markup, 2020). Yet, the same platforms, and the machine-learning community in particular, not only use these methods to collect data about individuals without their consent but also thrive and profit from all kinds of illegally obtained information to feed training for AI products (news, imagery, video, books, medical and financial records, etc.). This power imbalance was recently mitigated by U.S. federal court decisions allowing researchers to engage in Web scraping when their research aims to uncover whether algorithms result in gender, racial, ethnic, or other discrimination does not violate the Computer Fraud and Abuse Act (American Civil Liberties Union, 2020; Whittaker, 2020). Since this dissertation discusses forms of subjectivation and relations of production in machine learning development, issues of discrimination and civil liberties are deeply ingrained in this research.

Beyond the ab/uses of legal instruments and illegal practices, digital platforms also use socio-technical methods to prevent scrapers, such as Captchas and API policies. Captchas provide challenges that are difficult for computers to perform but relatively easy for humans, used in this context to detect and disallow bots from crawling webpages. API policies are a collection of rules

[2] Examples of such tools, services, and libraries include Octoparse, Easy Web Extract, Beautiful Soup for Python, Cheerio, and Puppeteer for NodeJS.

and gateways that can limit who, how, and when users and bots can retrieve data from the platform. An API may require authentication or limit the type and amount of data the user can retrieve and the number of requests a user can make in a specific time frame. In response, developers have used techniques to avoid detection and overcome these barriers, such as accessing the website using a Virtual Private Network (VPN), creating bots that simulate user behaviours, collecting data directly from the Document Object Model (DOM) on the browser instead of plain HTML, and using computer vision and natural language processing to parse the data.

3.3.1.2. Proceedings

Drawing from past experiences, most notably a scraper to gather YouTube's recommendation videos (McKelvey & Frizzera, 2019; Reis, Zanetti, & Frizzera, 2020), and using some of the tactics described above, I developed three Web scrapers to collect data from Kaggle and make screenshots from previous versions of Kaggle's homepage stored in the Internet Archive. The process is documented in Appendix I, and the code is publicly available on my GitHub repository.[3]

The first scraper was developed in October 2020 to collect data from Kaggle, focusing on three main areas:

1. Competition: Metadata from the competitions listed on the platform. It includes the title, description, who organized or sponsored it, the prize offered, the start and end dates, the number of competitors and teams, and the final leaderboard (including usernames, team names, ranking position, number of entries, and score).

2. Datasets: Metadata from all the datasets listed on the platform. It includes the title, URL, owner (username and URL), medals, upvotes, creation date, size, file types, and the number of files. It excludes any content from the dataset itself.

3. Public Ranking: Metadata from all the users ranked in the four publicly listed rankings on the website (i.e., competition, datasets, notebook, discussion). It includes the username, the URL, the date when the account was created, and the position, tier, medals, and points on each competition the user was listed.

I ran the scraper in two phases: on November 1, 2020, to collect a list of competition, datasets, and users, which took about eight hours to complete, and on December 13, 2020, to gather the details,

[3] The code for the Web scraper used in this research can be accessed at https://github.com/lucaju/kaggle

which took about 15 hours to complete. The scraper collected data from 3,437 competitions, 55,897 datasets, and 9,164 users on the public ranking page.

The second scraper was developed in March 2022 to collect snapshots from previous versions of Kaggle's homepage stored on the Internet Archive via its Wayback Machine. The data was captured in a two-step process:

1. Extracting a list of archived versions of *kaggle.com* containing dates and the URLs to each version (limited to one per month) using the tool "Internet Archive Wayback Machine Link Ripper" developed by Rogers (2017) and,

2. Visiting each URL and screen capturing each page using a custom script I wrote for this purpose.

The scraper ran on April 8, 2022, taking over three hours to complete. From February 6, 2010, when Kaggle was launched, to April 14, 2022, the Internet Archive saved its homepage 4,610 times. The second step collected 136 samples from October 2, 2010, to October 7, 2021.

The third scraper was developed in April 2022 to collect extra information about Kaggle's competition, in particular, the ones analyzed in this research, namely Facebook Deepfake Detection Challenge, Passenger Screening Algorithm Challenge, and Instacart Market Basket Analysis. This scraper focused on the threads posted by users in the competition's discussion forum. It includes the author (username, tier, and URL), title, content, date, the number of votes, the number of replies, and the URL. It also contains the thread's replies, including author information (username, tier, and URL), content, date, number of votes, and medals.

The scraper ran on May 4, 2022, taking about one hour to complete. From the three competitions mentioned above, the scraper collected 1,268 threads with 11,047 replies. More specifically, Facebook Deepfake Detection Challenge had 772 threads and 6,777 replies, the Passenger Screening Algorithm Challenge had 147 threads and 1,282 replies, and Instacart Market Basket Analysis had 350 threads and 2,988 replies.

3.3.2. Qualitative Content Analysis

Qualitative content analysis shares many features with other qualitative methods, such as a concern with meaning and interpretation of symbolic material, the importance of context in determining meaning, and a data-driven and partly iterative procedure. Unlike quantitative content analysis, which is exclusively data-driven and formally concerned with the objective, systematic, and quantitative description of the manifest content of the corpus, qualitative content analysis combines data-driven and concept-driven categories applied to latent and more context-dependent meaning. Because it can be challenging to describe consensually, consistency as a quality criterion is handled less strictly in the qualitative version of the method. In other words, it is a method used to systematically describe the meaning of qualitative data, done by assigning successive parts of the material to the categories of a coding frame, which, in turn, contains all those aspects that feature in the description and interpretation of the material (Schreier & Flick, 2014). This requires the researcher to focus on specific aspects of meaning that relate to the overall research question. The meaning found in specific passages of the material is to be taken to a higher level of abstraction, resulting in categories that apply to several slightly different passages in the material.

According to Schreier and Flick (2014), the coding frame consists of at least one main category and at least two subcategories. The main categories are aspects of the material the researcher would like more information about. Each main category should cover only one aspect of the material (unidimensionality), such as the primary goal of a competition on Kaggle. Subcategories specify what is related in the material concerning these main categories and should be mutually exclusive (e.g., algorithms to identify deepfakes). However, exclusiveness does not imply that any one unit can be coded only once: it implies that any unit can be coded only once under one main category. In this case, there can be only one goal per competition. Usually, the main categories are conceptual-driven, following the main research question and main objectives of the research. While the subcategories can be created following both concept-driven and data-driven approaches, the researcher should prioritize a data-driven strategy to ensure the coding frame is representative of the corpus. Drawing from the metadata collected using the web scraper, I use this approach to uncover the meaning and classify the modalities of machine learning algorithms developed on Kaggle.

Nonetheless, while content analysis seeks out latent meanings behind the surface appearance, it is limited to the evidence provided by a systematic quantitative study. That is, meaning is often complex, holistic, context-dependent, not necessarily apparent at first sight, and does not always equate the coding frequency of a given theme with its importance (Kracauer, 1952). Qualitative Content Analysis is primarily concerned with the description of categories, but it is ontologically and epistemologically “naive” (Schreier & Flick, 2014, p. 181). To overcome this limitation, I use

Software Studies combined with Discourse Analysis, which enables access to the ontological and epistemological assumptions that lie behind an algorithm, a project, or a discourse.

3.3.2.1. Proceedings

At the end of 2020, Kaggle listed 442 public competitions and over 3,000 education oriented (InClass) challenges. Since InClass challenges are beyond the scope of my research, the corpus used here is configured solely by open and publicly available competitions. Moreover, because the research also considered work relations and the commercial interests between the platforms and their users, I only examined the competition that offered at least US$ 1,000 in cash prizes (n = 284). Using this corpus and following the main objectives of this research, the coding frame comprises a single main category: the competition's primary goal. The subcategories were defined by critically inspecting each competition's title and its respective overview description, typically ranging from one to four paragraphs displayed on the competition's page. The four subcategories, which I called modalities of machine learning development, were created using a data-driven approach: identify, predict, recommend, and generate. The coding was performed primarily by the principal researcher, which is at least 10% (n = 30) of the material double-coded by a second researcher to ensure the consistency and validity of the subcategories created. Let us say at this point that this approach resulted in 140 competitions classified as "identify" (49%), 92 as "predict" (32%), 47 as "recommend" (17%), and five as "generate" (2%). I describe these categories in more detail in chapter five and unpack the machine-learning modalities they represent in chapters six to eight.

3.4. Analytical Framework

Using the sources listed in 3.2, chapters four and five of this dissertation focus on Kaggle as both a company and a platform, discussing how it was conceived, the main actors involved in its conceptualization, how it evolved since its first inception, what is its economic model, what is inside the platform, how it works, and who participates in it. Using a Software Studies and Discourse Analysis approach, I looked at how Kaggle's homepage design evolved through time, the founder's interviews, investor decisions, financial reports, blog posts, journal articles, public presentations, news reports, and other accounts from commentators. I consistently reviewed the material, checking facts with multiple sources, followed hyperlinks, and resurfaced websites using the Internet Archive in order to offer a narrative of Kaggle's history. Similarly, to shed light on how the platform works, I explore the dataset gathered using the web scrapers I developed to critically examine Kaggle's three main elements: its digital infrastructure, user base, and the competitions promoted on the website. These elements are enriched with the platform's documentation, surveys about user demography, comments from users and other actors, and examples of how competitions work.

Subsequent chapters focus on predictive models as a new form of subjectivation. Drawing from the first three machine-learning modalities (identify, predict, and recommend), I selected one competition from each modality as a case study.

> On ***Identify***, I examine Facebook's 2019 ***Deepfake Detection Challenge***. Offering US$ 1 million in prizes, this competition was intended to improve the detection of fake content on social media that elicits contentious questions about privacy, copyright, definitions of reality, and the relationship between subjectivity, truth, and power.
>
> On ***Predict***, I analyze the 2017 ***Passenger Screening Algorithm Challenge***. This competition, sponsored by the U.S. government, offered US$ 1.5 million to improve security in airport terminals, raising important questions about the normalization of bodies and behaviours, speculative futures, and geopolitical power.
>
> On ***Recommend***, I consider the 2017 ***Instacart Market Basket Analysis***. With only US$ 25,000 in prizes, this rather modest competition represents the most common and typical challenge on Kaggle: the attempt to make incremental improvements to algorithms in order to increase sales. However, this ordinary and perhaps unimportant competition brings serious controversial questions about privacy, hyper-profiles, nudges for profit gain, and the radicalization of user bases for population control.

The case studies selection criteria were based on a combination of factors involving data availability about the competition, their representativeness and importance on each category, diversity in terms of competition goals, the relevance of these competitions within the platform and in relation to the main questions of this dissertation, namely the political economy of subjectivation. The money prize was also an important but not determining factor. High-profile competitions attract a large number of participants and might propose pivotal questions about machine learning development. Mundane and trivial competitions, on the other hand, can reveal the automatism and solidification of socio-technical practices in this field. While the fourth modality, "generate," became the principal modality of interest in 2022, when GANS and LLM became widely available (ChatGPT, Bard, Midjourney, Stable Diffusion), this category was set aside due to its low representation in the corpus, poor participation, and insubstantial prize offer.

The goal is to investigate the main aspects, phases, and implications of a competition, such as its purposes and context, the type of data provided and how it was obtained, how the community was mobilized to participate in the competition, what kinds of solutions emerged from the challenge, and the broader implications for society of these predictive models developed by a data science community. By using Software Studies combined with Discursive Analysis, my goal is to unpack how data, code, digital infrastructure, and the user base are mobilized in competitions, what kind of value the output of these competitions has, and how they can be utilized to modulate individuals' habits and behaviours algorithmically. The next two sections discuss these methodological approaches.

3.4.1. Software Studies

Algorithms are created for purposes that are often far from neutral: extracting data and content, creating value and capital, and nudging behaviour and habits. They are used as a means of seduction, coercion, discipline, regulation, and control. Algorithms are an active force in the world, but they are also prone to failure and reappropriation. Developers cannot predict all the possible scenarios where algorithms are used. However, they can be updated based on contexts, under varying conditions, and in collaboration with people, data, and other technologies. Their effects, therefore, unfold in contingent and relational ways, producing localized and situated outcomes. When we use an algorithm, we are not merely executing commands and receiving information but learning, internalizing, and becoming intimate with it (Galloway, 2006). While we know that inputs condition algorithms, the cumulative data stored about users' interaction (profiling) can make the output much more complex and personalized than a simple response to a command. However, the very materiality of software, such as code, data, and the development process, often remains invisible.

Software Studies is a response to handling the particularities of new media and digital platforms and the complexities of these sociotechnical apparatuses. Researchers use this approach to study algorithms, programming subcultures, the value and aesthetic judgements built into computing, how digital systems impact everyday life, and "the tightly formulated building blocks that work to make, name, multiply, control, and interweave reality" (Fuller, 2008, n.p.). Kitchin (2017) suggests that the best and most logical way to understand algorithms and how they come to life, and figure out how they work in the world, is to conduct detailed empirical research focused on algorithms. He proposes that the constitution and work of algorithms can be empirically studied in several ways, and he suggests six methodological perspectives: (1) Examining pseudo-code (how tasks are codified in machine language); (2) Reflexively producing code (auto-ethnography); (3) Reverse engineering (build the code from its output); (4) Ethnography; (5) Discourse Analysis; and (6) Examining the effects of algorithms in the world. Furthermore, it is essential to account for the ways in which "individuals interfaced with the algorithm through software, including their assessments as to their intentions, sense of what is occurring and associated consequences, tactics of engagement, feelings, concerns and so on" (p. 26). Likewise, we should also account for how people resist against and subvert the work of algorithms and re-purpose them for purposes they were not originally intended.

As digital media becomes a cultural communicative infrastructure, Social Sciences and Humanities scholars begin developing new methods to interrogate their material and semiotic complexities, framing the cultural and social practices its users and developers engage in as they use these digital artifacts. For instance, Slack and Wise (2014) bring a Cultural Studies approach to engage technological cultures that consider politics, economics, space, time, and identity. They expose cultural assumptions that underlie our thinking about technology, stories so deeply held that we often do not recognize their influence. Brock (2018) goes even further with his Critical Technocultural Discourse Analysis (CTDA), which "combines analyses of information technology material and virtual design with an inquiry into the production of meaning through information technology practice and the articulations of information technology users in situ" (p. 1013). CTDA is a multimodal analytical technique that uses critical theory (Critical Race Theory, Queer Theory, Critical Feminism, Intersectionality, Postcolonial, or Gender and Women's studies) to provide a holistic analysis of the interactions among technology, cultural ideology, and technology practice.

Marino (2016), on the other hand, put forward a Critical Code Studies (CCS) approach that focuses on the cultural significance of algorithms themselves. He argues that by examining the source code, we can unpack social and cultural artifacts to understand the motivations, intentionalities, tendencies, and contentiousness of digital platforms. In other words, Marino suggests that the research should read "between the lines" of code to go beyond the surface of an object that supposedly only exists as neutral commands to make a machine execute specific actions. From this perspective, he argues that we should understand code broadly, not just the lines written in a specific computer language (e.g., JavaScript, Python, or C) to execute commands. Instead, we should also pay attention to code documentation, including the comments left by the programmer in the code, the overall structure and architecture of the software, and its compiled version. I argue that, whenever possible, the researcher should also experience and interact with the code in runtime, where it realizes its potential as a sociotechnical actor in our society. In this sense, other aspects related to code should also come into the research's purview, such as the ecosystem in which the code is inserted (hardware, operational systems, code dependencies), developer exchange related to the code in specialized forums (Stack Overflow, Kaggle), code repositories' issues tracker (Github, Gitlab), the runtime messages output from the code when it is in execution (logs, error message), and the code output (text, data translation, user interface, media content).

Given the broad understanding of what algorithms may be within Software Studies, the concern with digital media may quite easily mean several different things. For instance, to map Facebook's evolution from a platform to an infrastructure model, Helmond et al. (2019) look at the technical "boundary resources" (APIs, SDKs, and reference documentation) in conjunction with historical data about the company's business model and partnerships. Bucher (2013) combines interviews with developers and a critical analysis of Twitter APIs to reveal the mediatory functions of APIs and their power to produce new realities. Napoli (2014) looks at how media institutions are organized to explicitly link the concept of institutions with algorithms and the roles they play in the dynamics of media production and consumption. Kvale's (2016) study on MS Word's visual features uses Critical Discourse Analysis of software to show how the social values of office organization management have influenced users through templates and standardized multimodal representations. Maher (2011) also employs Critical Discourse Analysis in his analysis of Mozilla's Firefox technical documentation and identification of how the ideologies of software intersect with the nationally situated cultural values in which these technologies and texts are developed and used. These authors explore the social, political, and ideological implications of the materiality of software and algorithms' technical properties. On the one hand, Kvale (2016) and Maher (2011) are useful to this research in terms of the approach to the type of material examined (technical documentation around software development as discourse). On the other hand, Bucher (2013), Napoli (2014), and Helmond et al. (2019) are closer to the aims of my dissertation, which focus on the production of code and its impact on broader social and political values.

However, while the aforementioned studies mainly focus on well-established traditional software or digital platforms, my research is focused on machine-learning algorithms or, more precisely, how these algorithms are developed in a competitive crowdsource environment to produce predictive models. Following Kitchin's (2017) suggestion, this research uses Discourse Analysis to unpack the assumptions, motives, objectives, and practices involved in machine-learning competitions to produce predictive models that can, in turn, be utilized to modulate and shape individuals' behaviours algorithmically. This approach can reveal "how algorithms are imagined and narrated, illuminate the discourses surrounding and promoting them, and how they are understood by those that create and promote them" (p. 25).

3.4.2. Discourse Analysis

Discourse can be understood as a group of statements that structure how things are thought and the way we act based on that thinking. It is a particular form of language with its own rules and conventions that circumscribe institutions within which the discourse is produced and circulated. It is articulated through all sorts of technological devices, algorithms, codes, and texts, specialized or not, and also through the practices that those languages permit. The diversity of forms through which a discourse can be articulated makes intertextuality a critical method for understanding the underlying structure of a discourse (Rose, 2001). That is, the meanings of any one discursive text depend not only on them but also on the meanings carried by other texts. Foucault (1980) calls this discursive formation a "system of dispersion," consisting of the relations among parts of a discourse. A collection of fragmented pieces of information may seem disconnected and irrelevant, but it can reveal a coherent discourse among them (an order, correlations, positions, connotations, implication, resonance). A discursive formation is the way meanings are connected in a particular discourse, which, in turn, produces a particular knowledge about the world that shapes how the world is understood and how we act upon it (Rose, 2001).

The Foucauldian approach to discourse analysis focuses on the social processes through which a range of subjects are constituted. These processes seek to discipline individuals into particular ways of thinking and acting, fitting these subjects into a specific and well-defined condition of existence. Here, discourse can be described as a particular form of power that defines social rules and conventions. It is not a simply repressive power that imposes rules and behaviours upon human beings. Instead, it produces the world as it understands it. Rose (2001) describes it similarly: "Human subjects are produced through discourses. Our sense of our self is made through the operation of discourse. So too are objects, relations, places, scenes" (p. 137). Hence discourse is an instrument of governmentality where knowledge is used as "technologies of the self," constituting instruments to prescribe the "conduct of conducts" (Foucault, 1982). In this sense, discourse also produces subjects and their particular condition for their existence. For instance, a digital technology discourse produces, among other subject positions, programmers, engineers, data scientists, venture capitalists, advertisers, and users, as well as their roles, responsibilities, practices, opportunities, and aspirations.

Rose (2001) observes that the Foucauldian approach to discourse analysis produces two methodological venues, which, while similar in approach and overall structure, have different emphases in terms of their end goal. She simply calls these two branches "discourse analysis I" and "discourse analysis II." Discourse analysis I emphasizes the notion of discourse as an articulation of various kinds of text (written, spoken, visual images). Rose points out that this approach is "most concerned with discourse, discursive formations and their productivity" (p. 140), focusing on how people use language to construct their account of the social world. Understanding that discourses are socially produced rather than created by individuals, this research is primarily concerned with machine-learning algorithms' social, political, and economic aspects and the persuasive strategies used to produce such discourse.

Discourse analysis II, on the other hand, is concerned with the production and reiteration of particular institutions, practices, and subjects. Foucault suggests that institutions work in two ways: through their apparatuses and their technologies. Institution apparatuses are the forms of power/knowledge comprising the institutions, such as regulations, architectures, technical standards, scientific treatises, philosophical statements, and the discourse articulated through all these. Institutional technologies, on the other hand, are the practical techniques used to practice that power/knowledge. For Foucault (1995), technologies are "diffuse, rarely formulated in continuous, systematic discourse … often made up of bits and pieces … a disparate set of tools and methods" (p. 26). Examples of technologies in the context of this research are source codes (algorithms), datasets, training data, external libraries, predictive models, user interfaces, digital devices, computational infrastructures, and digital platforms. This second type of discourse analysis follows Foucault in understanding algorithms as embedded in the practices of institutions and their exercise of power, which is centrally concerned with the social production and effects of algorithms. It offers a methodology that considers how the effects of dominant power relations work through and by digital platforms and capitalist institutions.

When applied together with Software Studies, Discourse Analysis allows us to consider computer code, algorithmic media, the technical language employed in their development, the interfaces produced by these algorithms, their documentation and legal frameworks, the datasets and their predictive models derivative, the digital infrastructure they rely upon, the industries of technology, the community of developers, and the shared values and knowledge made possible within the current mode of production. As such, a discourse analysis approach would require an empirical account of particular texts and institutions, focusing on their details, casual assumptions, everyday mundane routines, taken-for-granted practices, and banalities. For instance, a digital platform might use a specific algorithm to make certain things visible in particular ways and other things unseeable, nudging subjects to behave and act in a specific way within the constraints of such a platform. These detailed descriptions help us unpack algorithmic practices and techniques and give us a glimpse of how subjects and objects were and are discursively produced.

According to Rose (2001), the tenuous distinction between the two approaches makes both suitable to either be used in isolation or combined as a single methodological approach that examines verbal texts, visual images, algorithms, datasets, institutions, and social practices together, which is the primary goal of this research. Since machine learning is not exclusively comprised of computer code, this research uses discourse analysis I with a variety of materialities related to digital technology development, such as textual information (documentation, comments, reports, interviews, and articles), pieces of code within and around selected machine-learning competitions, and the different ways user interact with predictive models (interface, mediated content) to explore how these algorithms affect individuals and produce subjects. I also pay close attention to internal and external resources for machine-learning development, such as datasets, training data, code libraries, as well as logs, software development tools and socio-technical infrastructure. For instance, predictive models are highly influenced by the dataset they are built upon, if not overly dependent on. In this case, closely examining the source dataset is as crucial as investigating traditional texts. Moreover, machine learning is a computationally intensive operation, requiring specific code libraries (e.g., PyTorch, Tensorflow) and specialized hardware (GPU, TPU) often deployed on cloud computing (Google Cloud, AWS, Azure). Here, the external libraries and digital infrastructure used and targeted by the code become an essential layer of scrutiny. In short, this method allows us to examine the discursive impact of machine learning as it materializes in the world, both during the production (e.g., inherent assumptions, benchmark datasets, unit operations) and consumption phases, such as on the user interface, public APIs, and the content generated, which can expose how these systems mediate and modulate user behaviour.

Yet, my research contends that machine-learning algorithms function simultaneously as a technology and an institution apparatus, acting both as power/knowledge (1982) and as a vehicle through which power is practiced, which can be used to produce a regime of truth. The institutional use of machine learning makes us believe the predictive models are truthful representations of reality, free from human biases and prejudice. However, they are built from particular social and political accounts of reality, using curated datasets and manipulated for and by a specific economic condition. Discourse analysis II is helpful here because it enables us to access the ontological and epistemological assumptions behind a statement, an algorithm, a project, or a discourse (Rose, 2001). It is a tool to help reveal the hidden structures behind systems of power, allowing us to inquire into how code, data, developers, and digital infrastructures are mobilized to produce specific kinds of machine-learning algorithms and predictive models. That is, how the machine-learning community and the AI industry develop machine learning algorithms as new methods of subjectivation. I use Discourse Analysis to focus on the power dynamic in which these algorithms have been developed, observing the interactions of machine learning competitions on Kaggle's forum, the context in which the competitions were initiated, how the datasets provided by the competitions' sponsors were constructed, the technology involved in producing machine learning and predictive models, the conditions of production in the platform, and the impact of these models on the individual and on society.

3.5. Positionality

As with any research project, I recognize that the researcher's worldviews and life experience exert considerable influence on the research process, particularly when discussing the process of subjectivation, the sense of "self," and the mediation of embodied experiences. At times in this research, I borrow the voices of different individuals, often representative of minority groups, not so much to illustrate but to explain in their own words how AI systems impact their lives. Their stories they tell are disturbing and a trigger for some audiences, but they help us glimpse the devastating implications of AI technology. However, I do not consider myself part of a minority group. I have never been black, a woman, homosexual, or transgender. I have never been marginalized, experienced poverty, or discriminated against for the colour of my skin, my gender, or my sexual orientation. Quite the opposite; I am what is generally referred to as "white passing," easily identifiable as a white male who is in a privileged position of being a PhD student and a researcher at a prestigious university in North America.

Recognizing that my identity is involved in all of my analytic work, I believe it is important to disclose my own life experiences that may have some impact on this research. I was born in Brazil and had a privileged upbringing in the comfortable, white middle-class of a midsize city. Both my parents are academics, teaching and researching media studies and architecture at a public university. I grew up surrounded by family members with ties to education and research institutes, from whom I had the opportunity and the privilege (and perhaps the annoyance) to share knowledge. My native language is Portuguese. I consider myself fluent in English and have a limited understanding of French and Spanish. I have lived most of my life in Brazil (Espírito Santo) before moving to Canada more than twelve years ago to do my Master's (in Alberta) and PhD (in Quebec).

Doing research in Canada, thousands of kilometres away from my home country, in a language that is not native to me, made me foreign, sometimes even marginalized: I am not quite as "Western" as I was once thought to be. As an immigrant from Brazil in Canada, I position myself as a Global South scholar interested in the study of power within global capitalism in ways that transcend previous notions of nation-state subjectivities and focus on the intersectionality among subaltern groups across national, linguistic, racial, and ethnic lines in relation to digital technologies. Still, I recognize that the privilege accompanying my life means that, despite its good intentions and efforts to promote equity and diversity, my research risks being patronizing, insulting, threatening, imperialist, and recolonizing. I acknowledge my privileged position in accessing specific (material and immaterial) resources, but I strive to be aware of my biases and recognize how they may shape my research.

As a researcher with a background in Media Studies and Digital Humanities, my engagement with Kaggle and the broader data science community is both an exploration and a learning journey. I had access to digital media as a teenager and have worked as a UX designer and web developer ever since. My experience in web development has provided me with a strong technical foundation; however, my knowledge of data analytics and statistics is limited, positioning me as a novice in machine learning and data science. This unique standpoint allows me to approach Kaggle not just as a platform for machine-learning competitions but as a critical site for understanding the intersection of technology, data culture, and community practices. My perspective is informed by media studies theories and methods, which emphasize the socio-cultural implications of digital platforms and their user dynamics. This perspective shapes my research to focus on how knowledge is constructed and shared within the Kaggle community, how collaborative and competitive interactions unfold, and how these processes influence the broader landscape of data science and machine learning. I had never used Kaggle before. I created a profile on the platform for the sole purpose of this research, and I do not consider myself part of the data science community. Importantly, though, the time spent in the platform not only helped me to observe the data science community's idiosyncrasies but also exposed me to more advanced computer science concepts and coding practices, ultimately improving the results of this research and my own skills in programming languages and data analytics.

3.6. Summary

This research uses a mixed-method approach, combining traditional methodological perspectives in Media Studies, such as Qualitative Content Analysis and Discourse Analysis, with new methodological approaches focusing on new media and digital platforms, such as Digital Methods and Software Studies, to answer how algorithmic mediation produces subjects and reorganizes life. These methodological approaches provide the necessary instruments to uncover empirical data from real-world events, produce evidence of the current practices and ethical concerns in machine learning development, unpack the relations of power in the AI industry, and shed light on the ways predictive models have evolved to shape user behaviour and the notion of the self.

The following chapters consider how code, data, and Kaggle's large data science community are mobilized to produce predictive models that can be utilized to modulate individuals' habits and behaviours algorithmically, leading to a redistribution of processes of subjectivation. In other words, I examine how this new infrastructure of subjectivation aims to mobilize subjective materials associated with individuals in order to transform broad conditions of existence and shape the very conditions through which anyone, or anything, gains the possibility of existence in this world.

4. On Kaggle: From Disruptive Start-up to Big Tech Subsidiary

> Today you can already rent the brain of a data-mining genius via Kaggle by the hour, tomorrow by brain-hour … your brain plug firmware will earn you a little extra cash while you sleep, by being remotely programmed to solve hard problems. (Levchin, 2013, para. 21)

In 2007, during a short three-month internship at *The Economist*, Anthony Goldbloom, who had recently graduated in Econometrics at the University of Melbourne, was asked to write a piece about the emerging hype around "big data." Following his academic background, which strives to use statistics and mathematical techniques to "justify" a theoretical economic model with empirical rigour, he voiced the idea that data and statistics models alone ought to be enough to solve economic problems. In his article, he put forward that companies should be investing in predictive analytics and machine learning applied to business problems, arguing that the potential to use data to make business decisions is far greater than trying to forecast the economy based on aggregated indexes (Goldbloom; 2019; Gruen & Goldbloom, 2008). Ironically, Goldbloom was proposing to extend a mainstream practice used for decades in the sales and financial sectors (Turow, 2017), even though it was highly criticized for failing to predict the 2008 financial crisis (Cooper, 2011) and contributed to the historic stock market crash in the U.S. Capital market in 2010 (Finn, 2017). Until 2012, outside of academic research and large hi-tech corporations like Microsoft and Google, machine learning had largely been under the radar, with little impact on public opinion and business practices. "Data science" and "machine learning" were obscure terms only used in niche sub-fields. However, a sequence of technological breakthroughs in computer science[4] made Goldbloom's pitch resonate like gold in the financial sector, especially for economists and

[4] Most noticeable are the development of ImageNet, a large-scale image library for object recognition, by Geoff Hinton's group at the University of Toronto, and advancements in unsupervised machine learning by DeepMind, acquired by Alphabet in 2014 (Lawrence, 2015).

marketers. It put his name on the Forbes 2012's "30 under 30" list, right next to Mark Zuckerberg, Apoorva Mehta (Instacart's founder), Lady Gaga, Ariana Grande, and Justin Bieber (Forbes, 2012).

At the beginning of his career, Anthony Goldbloom briefly worked as a statistician at the Australian Treasury in Canberra and at Australia's Reserve Bank. In an interview for Chai Time Data Science, Goldbloom (2020) explained that macroeconomic indexes, such as GDP, inflation, and unemployment rate, are hard to forecast because the data is composite and messy. Furthermore, he argued that governments and financial agencies do not have accurate and minute data about every economic sector. Businesses, on the other hand, are able to produce rich and extensive datasets for each moving part of their operations (transactions, costs, stocks, salaries, investments, etc.). In a short promotional video for the 2014 World Economic Forum (WEF, 2013), Goldbloom stated that he wanted to "help businesses make decisions on the basis of data rather than gut instinct" (0:12), implying that the traditional way of doing statistics was missing important signals. Caught up in the centre of the neoliberal forum, Goldbloom's statements reflect the idea that the State is the enemy of private individuals and businesses, suggesting that governments are inefficient, slow, costly, and do not have the interest of the market at heart. Adept at using the mantra "data speaks for itself," he believes that if businesses could find a better way to extract and harness more information from their complex and large datasets, they would make better business decisions, which would also push the country's economic growth: government agencies do not have the time or the resources to understand individual businesses, so why not let them do the job themselves? (Goldbloom, 2019).

Goldbloom proposed that the new generation of computers should be put to work by and for the private sector to amass large quantities of data as a way to provide a more predictable forecast for business decisions. This would be achieved using the real-time data collection of people going about their everyday lives while machine-learning algorithms would find the best predictors of their behaviours, all with the goal of optimizing capital gain (Gruen & Goldbloom, 2008, p. 16). However, machine learning and predictive analysis were not affordable for most businesses. It requires high computation power to crunch large volumes of data, skilled programmers able to develop algorithms to unearth signals from the dataset, and specialized professionals capable of understanding the data to produce efficient economic predictions for supposedly better business decisions. Goldbloom's intuition led him to create Kaggle, a platform to host hackathons focused on machine learning solvable problems as a way to prove his point that business decisions should be data-driven rather than based on gut instinct.

This chapter traces the first ten years of Kaggle's history as a company and a digital platform. In the first section, I offer a narrative of Goldbloom's and his partners' steps, successes, failures and pivots to build a start-up company able to attract millions of users and the interests of the big tech industry. Following the start-up culture and Californian Ideology (Barbrook & Cameron, 1996), Kaggle was built to appeal to both technologist enthusiasts and venture capitalists, promising a digital revolution with low cost and high profit. At its core, it carries the idea that statistical models alone can solve any political, social and economic problem, a concept closely related to Goldbloom's background in Econometrics. Furthermore, Kaggle embraces the logic of the platform economy (Srnicek, 2017; van Dijck, Poell, & Waal, 2018), where it positions itself as an indispensable intermediary between large corporations seeking to use big data to optimize their business and developers avid to find jobs. Engineered similarly to Uber and Airbnb, Kaggle is seen as the world's largest data science company that owns no data, no data centres, and no code.

In the second section, I examine Kaggle's website archive to consider how the company has materialized itself to its users, evolving from a hackathon organizer to a temp agency to a gamified social network until it became part of Google Cloud Services. Using a digital archeological approach (Rogers, 2017), I dive into the graphical user interface used by Kaggle on its homepage to understand how the platform communicates with its public. Similar to traditional archeology, the artifacts found by a digital researcher might be half-broken, malfunctioning, or inaccessible due to changes in technology. These fragments, sometimes displaced and corrupted, constitute the history of a platform, infrastructure, collective effort, and human ingenuity. Thus, the shifts in the webpage layout, content, and imagery used by Kaggle can reveal the aesthetics and technological choices in a given period as well as the discourses, strategies, and goals the company was and still is pursuing.

Lastly, I discuss Kaggle's strategic vision and business model. Heavily based on crowdsourcing, Kaggle aimed to generate revenue from machine-learning competitions by charging companies to sponsor these events and attract free labour to solve problems using predictive models. The model put in practice by Goldbloom and his partners configures the neoliberal approach of disruption to keep everything the same. With its insatiable appetite for profit, Kaggle does not see the Internet as a self-organized collective of individuals that collaborate and produce creative work, described by Lévy (1999) as "collective intelligence" and by Surowiecki (2005) as a "wisdom of crowds." Instead, the platform harnesses free labour (Terranova, 2004) available in the digital world that crosses national borders and has little regulation: zero for workers and no accountability for private companies (Dyer-Witheford, 1999; Lazzarato, 2004). Not only does Kaggle disguise its operations by gamifying (Whitson, 2013) the tasks it requires from its users, making the platform attractive to an economic system that champions individualism and competitiveness, but it also pushes toward unlimited exploitation of the human body and cognitive abilities as a computation resource where our brains would be remotely programmed to solve problems in exchange for a little extra cash even while we sleep (Crary, 2014; Levchin, 2013).

Kaggle is an important entry point for developers and private companies that wish get into the machine learning space. Different from the workflow of a large corporation, Kaggle's competitions are a place of experimentation where most market rules do not apply. It is where new ideas for machine learning, artificial intelligence, and algorithms are born. It is a crucial place in which to study how social and political norms and rules are codified into models that shape individuals' behaviour, preferences, and desires, which, in turn, produce the subjects of this digital era.

4.1. The Company

After his internship at *The Economist*, Goldbloom started to sketch how he would put his ideas into practice. His vision, however, would not come to fruition without proper access to a robust digital infrastructure, machine learning techniques, and the know-how to handle big data. Observing hackathons, such as the Netflix (2009) prize and the KDD Cups,[5] Goldbloom realized that he would not need to spend millions on digital infrastructure and hire thousands of computer scientists to put his plan in motion. He simply needed to convince people to work for free on their own computers to do the job. If he posed interesting problem-solving challenges and offered some form of reward, people would come by the thousands.

At the same time, Ben Hamner, recently graduated in biomedical engineering, electrical and computer engineering, and mathematics from Duke University, was working with machine learning to improve non-invasive brain-computer interfaces as a Whitaker Fellow at the École Polytechnique Fédérale de Lausanne (EPFL) (Kaggle Team, 2011). Hamner (2015) felt that his research was very restrictive and wanted to pursue a broader use of the technology that would produce a "greater

[5] The KDD Cup is an annual competition organized by the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD).

impact in the commercial environment ecosystem" (43:30-43:40). Goldbloom joined forces with Hamner to co-found Kaggle in 2010. Since Hamner had more experience with machine learning, he assumed the CTO position. He was in charge of the decisions related to the technologies used in the platform, as well as designing and structuring competitions. Goldbloom, who was more skilled in business and economics, assumed the leadership and CEO position, becoming the public face of the

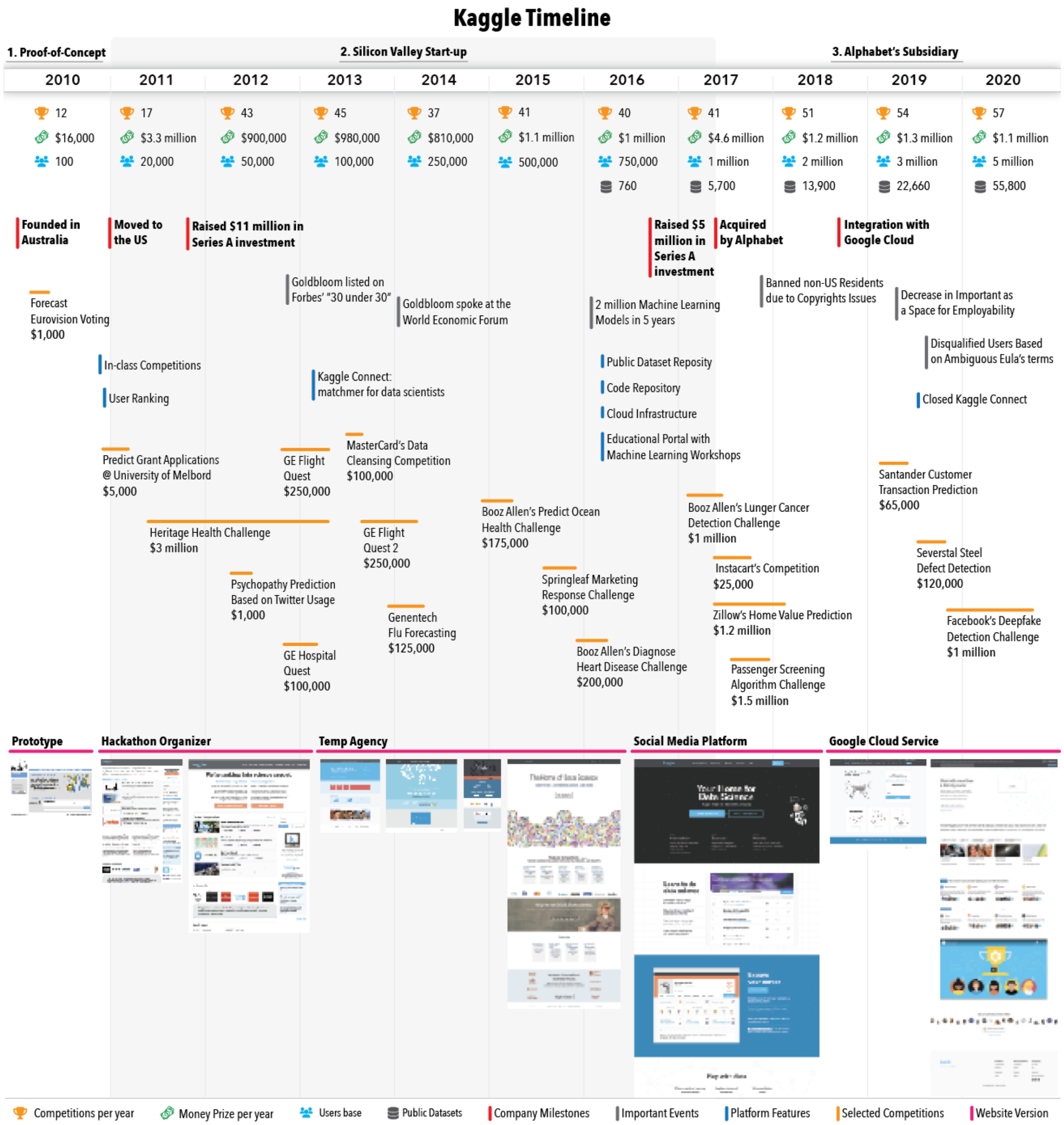


Figure 4.1: Kaggle History Timeline. A timeline showing the company phases, milestones, metrics, and website versions. Designed by L. Frizzera.

new company. Kaggle's trajectory can be split into three distinct moments: a proof-of-concept in Australia in 2010, a start-up company in San Francisco from 2011 to 2016, and a branch of a large company when Google/Alphabet acquired it in 2017 (see Figure 4.1).

4.1.1. Proof-of-concept (2010)

Kaggle was initially launched as a website to hosts data science competitions focused on solving business problems. The goal was to provide private companies with a place where they could outsource number-crunching jobs by turning them into competitive challenges where data scientists and software developers produce predictive models that would solve the problem, or at least offer optimized solutions. However, it was not an innovative business idea. Other companies and initiatives, such as InnoCentive and OpenCoder, were already established and had been in operation for about a decade when Goldbloom and Hamner decided to launch Kaggle. For instance, in 2001, InnoCentive (n.d.) started its operations with a similar proposition: companies are invited to share data about complex scientific research or development challenges online, offering cash bounties for solutions. In exchange, its online community provides answers to scientific puzzles, from which companies gain valuable insights that their in-house scientists might have struggled to develop, and much more quickly and cheaply. By 2006, InnoCentive was a successful platform with thousands of users and challenges worth US$ 1 million. While Kaggle lacked the infrastructure, experience, and user base to compete directly with InnoCentive or OpenCoder, Goldbloom was betting on a specific trend in data and computer sciences to capture the industry's interest and new developers: big data and machine learning. Yet, by 2010, similar coding competitions used this same approach, sponsored by private companies (Netflix Prize) and in academic settings (KDD Cup).

Goldbloom needed a proof-of-concept, an example that machine learning techniques could provide useful predictive models. In April of the same year, Kaggle launched its first competition: the Eurovision Song Contest. The challenge required engineers to forecast the voting for that year's Eurovision Song Contest in Norway. Twenty-five people participated in this six-week challenge, in which the winner received US$ 1,000. In a matter of weeks, Kaggle had enlisted volunteers to work on made-up machine-learning problems using their own computers and free time. The following competition set the developers to predict how genetic markers might affect the viral load of HIV-infected individuals. Dozens of people from around the world flocked to Kaggle to help solve this problem, none of whom had experience with HIV research. “In a week and a half,” Goldbloom says, “the best scientific research had been blown out of the water” (Pollack, 2012).

Goldbloom’s brilliant idea is well-known to capitalist society: outsourcing the work and exploiting the commons. In digital capitalism terms, this translates to crowdsourcing and data appropriation, which I will explain in more detail in the next section. Goldbloom was advocating for a performative discourse around digital platforms in which these appear to be a synonym for efficiency. As van Dijck et al. (2018) point out, “by virtue of their alleged leanness and openness, they claim to make the world a better place because they get rid of costly overhead and enable citizens to act independently, [as] autonomous individuals” (p. 23).

Since its inception in Australia, half a world away from Silicon Valley, Kaggle has been highly influenced by the “start-up” culture and the Californian Ideology (Barbrook & Cameron, 1996). This ideology represents a new force of capitalism, emerging from a bizarre fusion of the cultural bohemianism of San Francisco with hi-tech industries combined with the free spirit of the hippies and the entrepreneurial enthusiasm of the yuppies. The ambiguity of the Californian Ideology is most pronounced in its contradictory visions of the future. It prophesies that in the digital utopia, everybody will be “hip and rich,” despite the rapid concentration of power in the hands of large digital companies and the increased social and economic inequality sponsored by the neoliberal agenda in the last 40 years (Piketty, 2017; Srnicek, 2017; van Dijck; 2013). This amalgamation of oppositional political visions has been achieved thanks to a profound faith in the emancipatory potential of the new information technologies, through which the development of hypermedia and digital infrastructures has become a critical component in the next stage of capitalism. As Zuboff (2015) points out, the introduction of media, computing, and telecommunications technologies into the factory and the office is the culmination of a long process of separating the workforce from direct involvement in production. According to the Californian Ideology, existing social, political, and legal power structures will wither away, replaced by unfettered interactions between autonomous individuals and their software (Barbrook & Cameron, 1996). It argues that when the ability to produce and receive unlimited amounts of information in any form is combined with the reach of the global networks, existing conditions of work and leisure will certainly be fundamentally transformed—a thought that permeates both the Kaggle’s marketing discourse and the way it has planned its business model.

Even the name of the company—*Kag · gle*—follows this mix of inventive geek practice with the pragmatism of the market. In the matured market of web domains, it is not easy to come up with a memorable name that fits the social technicality of a web URL—it should be short, with no spaces or special characters, and unique. Before the Dotcom bubble, at the turn of the century, we saw the rise of companies with unconventional names: Yahoo, Google, and ICQ. In the ensuing years, more tech companies emerged and branded themselves with even stranger and more idiosyncratic names: MySpace, Friendster, Orkut, Flickr, Tumblr, Twitter, YouTube, Twitch, and so on. Goldbloom followed Silicon Valley's culture, choosing a quirky name to baptize his start-up. With the "good" short name domains already taken, he took a data-driven approach: he wrote an algorithm to generate all the pronounceable combinations of letters, three syllables or fewer, whose dot-com addresses were not yet claimed. From the 700 names output by the algorithm, he picked two finalists: Stumble and Kagle. With these two names, Goldbloom abandoned the data-driven strategy and consulted with family and friends (Gellman, 2013). Stumble was too close to *StumbleUpon*, a well-known discovery and advertisement engine. Besides, this name would give the wrong impression if they were trying to sell the idea of solving problems with statistical precision. The other option, "Kagle," was already taken by a small online education portal in the US. So, Goldbloom added an extra "g," and Kaggle was chosen as the official name for what has become the world's largest data science community.

4.1.2. Silicon Valley Start-up (2011–2016)

Goldbloom was overwhelmed with work; preparing for new challenges consumed most of his time. In November 2010, Jeremy Howard, an Australian data scientist and entrepreneur—one of the first Kaggle users—joined the firm as President and Chief Scientist (Pollack, 2012). Howard convinced Goldbloom that Kaggle's success depended on its proximity to other large players in the digital economy. So, in 2011, they moved the company to California, where they quickly secured US$11 million in a Series A investment round (Crunchbase, 2021), mostly from oligarchs in the tech industry. Goldbloom invited the Ukraine-born American software engineer and businessman Max Levchin to become Kaggle's Chairman. Levchin was an experienced and successful investor in the San Francisco Bay Area. Before putting money into Kaggle, he had invested in several other start-up companies, such as Yelp and Evernote, and sold his company "Slide" to Google in 2010 for US$182 million (Crunchbase, 2022). Levchin is part of the so-called "PayPal Mafia," a group of twenty founders and former employees who worked for the company in the late 1990s and became successful entrepreneurs. Other members of this "PayPal Mafia" were also betting heavily on the next big technological leap, from machine learning and big data to social networks and social media platforms: Peter Tiel founded Palantir focusing on big data analytics and predictive policing in

2003; Elon Musk became Tesla's largest shareholder in 2004; Steve Chen, Chad Hurley, and Jawed Karim founded YouTube in 2005 and sold it to Google in 2006; Reid Hoffman founded LinkedIn in 2003, acquired by Microsoft in 2016. Other early investors include Index Ventures, Khosla Ventures, SV Angel, Naval Ravikant, Hal Varian, Google chief economist and one of the masterminds behind Google surveillance capitalism strategies (Zuboff, 2015), and Yuri Milner, co-founder and former chairperson of Mail.Ru Group (Crunchbase, 2021).

Goldbloom was surrounded by opportunities and people he admired (WEF, 2015), who, at the same time, were fierce competitors for his forum-like website. With so much money at stake, he could not afford any mistakes. In Australia, he and his partner managed a dozen events with no more than 100 participants, offering little over US$1,000 in prizes for each competition. In California, they would need to be bolder.

Early on, Kaggle was more a fun project than anything with a grand vision. In its first year in California, Kaggle partnered with Heritage Provider Network, a Californian healthcare insurance company, to launch a US$3 million multi-year competition to identify patients admitted to hospitals within the following year using historical claims data. The high-value prize was a strategy to attract people to the platform, which succeeded in bringing more than 1,500 competitors to this challenge alone. However, the negotiation and organization of each competition were poorly structured in the early years of the platform. Most sponsors would not pay more than US$10,000, with the exception of large companies like General Electric (GE), Hewlett-Packard (HP), Mastercard, and Booz Allen Hamilton, which each held competitions offering over US$100,000 in prizes. A coherent and standardized plan to organize these events would only be put in motion in 2016 after several changes in the platform (Goldbloom, 2016).

The competitions were publicly announced and opened to anyone who wanted to participate. However, developers were unfamiliar with machine-learning tools, which looked a little daunting at first glance. Kaggle began promoting machine-learning literacy to address this problem and attract even more users. The company wanted to prepare and train a new generation of computer scientists to work in a gamified competitive environment. Kaggle invested in a program they named "in-class" competitions, where high school and college teachers could run private challenges as graded student assignments. In 2021, 90% of the competitions held on the website were private.

Between 2011 and 2016, Kaggle received international attention and recognition as a business model and infrastructure advancing machine-learning development. Goldbloom was once again on the list of speakers at the WEF (2015), where he sold his vision of using low-wage crowdsource labour to develop machine learning that focused on marketing and financial gain. At the same time, the company began listing all of the users on the platform in a general ranking system based on how they scored in each competition. By 2016, Kaggle had reached 1 million users; over 2 million machine-learning models were submitted to the platform (Goldbloom, 2016).

However, Kaggle failed to bring in new sponsors and monetize the platform. Since its inception, Kaggle has made all its revenue from machine-learning competitions, but that has not been profitable (Chauhan, 2021). Machine learning was very immature, and there was not much market for it beyond big companies. In 2013, Kaggle looked at adding on other business lines that would be more profitable, such as forming expertise in specific industries and building machine-learning solutions for that industry. At first, they picked Oil & Gas because they had Shell as a customer who wanted to do more machine learning. Goldbloom thought the market opportunity was good, but it became more challenging when oil prices crashed in late 2014 (Plumer, 2015).

As the machine-learning market matured, Kaggle returned to building a solid business around machine-learning competitions. Looking for different ways to increase engagement, Kaggle revamped the website to make it look like a social media platform where users could follow one another and like one another's code. On the educational side, Kaggle includes a Courses Section that features workshops and online classes focusing on Artificial Intelligence, Machine Learning, and Data Science literacy, giving the platform greater appeal and reaching beyond the expert audience. Furthermore, foreseeing the importance of cloud computing, the company also heavily invested in data centres, introducing new services that not only opened the possibility of sharing datasets on public repositories but also enabled users to use cloud infrastructure to train their models for a competition, exponentially increasing the quality of the work done on Kaggle.

The platform began to look like a game where players compete not just for prizes but also to become the best among their peers. As the players complete tasks and accumulate competition scores, they are also awarded symbolic titles—Contributor, Expert, Master, Grandmaster—to display next to their usernames, which can be used to brag about their skills in the platform. As a result, the site became a hub for scouting new talent, the best data scientists and machine-learning developers. These changes made Kaggle financially attractive again, instilling investors to inject US$5 million more in another round of a Series A investment in the same year.

4.1.3. Alphabet's Subsidiary (2017 and beyond)

Alphabet acquired Kaggle in March 2017 for an undisclosed amount as a way to secure a good position in the AI and machine-learning market (Lardinois et al., 2017). Apart from some criticism about the platform's independence and how Alphabet would oversee the community, Kaggle's users welcomed the change, looking forward to better support and improved infrastructure for machine-learning development. The success of this move can be measured by the rapid expansion of the user base, which reached five million users in 2020 (Kaggle, 2020b). Though an important step for Goldbloom's company, the acquisition seemed unimportant from investors' perspective, with little to contribute to a large corporation like Alphabet. In Alphabet's earning calls, Kaggle has been mentioned only once since 2015 and is apparently misrepresented as the "largest community of data centers" (Alphabet, 2017a) rather than as a data science community, as Goldbloom describes it.

In 2016, a few months prior to the acquisition, Google partnered with Kaggle to host a US$100,000 competition to improve YouTube's video classification through a deep integration with the Google Cloud Platform. This was no coincidence but part of a larger plan that had been put into practice by Sundar Pichai, Google's CEO, in 2015. Appointed amidst the creation of the new holding company for the Google company family, Alphabet Inc., Pichai would step into the role of new CEO in 2019. With a new business structure, Google was reinventing itself once again, and for Pichai, the future was in predictive models. In his first letter to stockholders as Google's CEO, he laid out his vision, highlighting the role of machine learning and artificial intelligence: "Over time, the computer itself—whatever its form factor—will be an intelligent assistant helping you through your day. We will move from mobile-first to an AI-first world" (Pichai, 2016, Powerful computing platforms, para. 3). Eric Schmidt, Google's former CEO, had already made similar remarks at Google's 2015 Annual Meeting of Stockholders, observing that he was "convinced that computer science and the platforms that we are all building on will be changed by machine learning" (Google, 2015, n.p.).

Google acquired other AI-focused companies, most notably DeepMind, in 2014. Unlike Kaggle, DeepMind was mentioned several times in the earnings calls and annual reports. Both the investors and the company's top executives praised DeepMind's new advances and technological breakthroughs in machine learning and AI, such as when "AlphaGo took on Lee Sedol, a legendary Go master, becoming the first program to beat a professional at the most complex game mankind ever devised" (Alphabet, 2015, p. 5). While the company cheerfully publicizes such achievements, the focus has always been on applying AI for advertisement optimization and looking for increased rentability and profit. Year after year, the company reassured investors that its efforts were, first and foremost, aligned with financial gain. Google wants to "put the power of machine learning into marketers' hands" (Alphabet, 2018a, n.p.) so they can use it to "improve campaigns and drive great results for advertisers" (Alphabet, 2017a, n.p.) and "help make it easier for advertisers to reach consumers" (Alphabet, 2017b, n.p.), "analyze ad placements on a publisher's page and show ads when they are likely to perform well while providing a good user experience" (Alphabet, 2018b, n.p.), and "[automatize] the manual process of building text ads and optimize them in real time to show the best performing ad for each search query" (Alphabet, 2018a, n.p.). This vision is precisely what Goldbloom proposed in his 2008 article: predictive analytics and machine learning applied to business problems can generate more profit.

Still, machine learning was not a novelty for Google. It has been part of the company's DNA since it started using advertising as its business model in the early 2000s. According to Sridhar Ramaswamy, Google's Senior Vice President for Ads and Commerce, ads are one of the earliest and largest use cases of machine learning at Google "because we've been very intimately involved with things like predicting click-through race of ad which was how good an ad is for a consumer [yielding] rich dividends for the ads team and for Google" (Alphabet, 2016). While machine learning was always a structural part of Google's business, it was only in the last decade that the company began heavily investing in artificial intelligence.

Though Kaggle has been focused on machine learning since its inception, its main focus has not been on research and development but on organizing competitive events that use machine learning to solve third-party business problems. Kaggle was designed to be in between—to mediate—third-party datasets and programmers' code. It is, as it was, a platform to mobilize crowdsourced jobs. Therefore, it is not surprising that Kaggle's acquisition has less to do with AI and is more related to expanding Google's computing infrastructure. In January 2017, Pichai indicated that Alphabet was "investing significantly in the machine learning capabilities and next-generation computing infrastructure that will propel Google's growth over the longer-term" (Alphabet, 2017a, n.p.). Two months later, when the acquisition became public, Kaggle was welcomed by Fei-Fei Li (2017), Chief Scientist at Google Cloud, highlighting the synergy between the human-machine assemblage: "Kaggle and Google Cloud will foster a thriving community of machine learning developers and data scientists, giving them direct access to the most advanced cloud machine learning environment" (para. 4).

Goldbloom (2017a) gave the impression that the goal of the deal was to make "Google Cloud technology available to the community, [allowing] us to offer access to powerful infrastructure, scalable training and deployment services and the ability to store and query large data sets" (para. 6). However, while this is true, the acquisition had the opposite goal: make Kaggle's community—users, datasets, codes—available as assets to advance Google's Cloud and AI agenda. In this sense, Pichai's misrepresentation of Kaggle as the "largest community of data centers" is not a mistake but a Freudian slip: it places the value of Kaggle not in the collective work done by its community of developers but as part of an extensive digital infrastructure. For Pichai, Kaggle's users are essentially just cogs that sit between the data and the computer, which he could offer together with cloud computing for pennies on the dollar per hour. Kaggle is important not as high-level research and development, but as a low-level laboratory where Google can advertise and expand Google Cloud services, and leverage the crowdsource community to produce new insights in the machine-learning field for a tiny fraction spent on in-house AI teams. As I will discuss in more detail later in this chapter, what is most important about Kaggle is its business model and the community around the platforms where Alphabet can train a workforce, in effect modulating the market supply of data scientists and machine-learning engineers.

4.1.4. Controversies

Not surprisingly, Kaggle has been involved in several controversies involving unfair rules, copyright infringement, and cases of plagiarism and cheating. For instance, in 2017, the company temporarily banned Chinese users from a US$ 1.2 million competition because the American company Zillow feared having its copyrighted material stolen by the Chinese government (Hsu, 2017). In the same year, repudiating its own discourse to promote a culture of open data and international cooperation, Kaggle limited participation to U.S. citizens only in the US$ 1.5 million competition sponsored by the U.S. Department of Homeland Security (Chowdhury, 2017). As I will show in chapter seven, Goldbloom (2017b) admitted the company was making discriminatory compromises for the "sake of a dataset."

In another case, before announcing the winner of Facebook's US$1 million "Deepfake Detection Challenge," Kaggle disqualified the top two teams due to unauthorized use of external data sources. Though the competition's rules allowed external resources, the teams must "ensure the External Data is available to use by all participants of the competition for purposes of the competition at no cost to the other participants" (Synced. 2020, para. 5). The previously top-ranked team, "All Faces

Are Real," manually created a dataset from YouTube videos and the Flickr-Faces-HQ Dataset with a Creative Commons (CC-BY) license, explicitly allowing commercial use. While the team did not have specific written permission from each individual appearing in videos, Facebook and other large companies constantly used this practice to train their own machine-learning algorithms. I examine this issue in more depth in chapter six.

Algorithms and code libraries are also subject to disputes. In the Kaggle's 2020 Global Wheat Detection competition, many competitors incorporated the recently launched YOLOv5 into their machine-learning algorithm to improve predictions and dominate the public leaderboard. YOLOv5 is a model in the You Only Look Once (YOLO) family of computer vision models and is commonly used for detecting objects. However, Kaggle banned YOLOv5 due to licensing concerns. YOLOv5 is licensed through GPL3 (General Public License), but competition rules state that the winner's submission (data and code) must comply with the MIT License, which has more flexibility for commercial use (Amusi, 2020). Lastly, a severe case of cheating occurred in 2019, when a team of programmers, led by the "Grandmaster" Bestpetting, third in the overall ranking at the time, scraped a pet adoption website to cheat in a US$25,000 contest that was intended to help shelter animals get adopted (Gordon, 2020; Quach, 2020a).

4.2. The Platform

Another way to tell Kaggle's story is to observe how the company has communicated with its target audience—the types of discourse, media channels, visual aids, and metaphors it has used—and how this communication has evolved over time. As a born-digital object, Kaggle, the web-based platform, can also be historically analyzed through its website archive. The Internet Archive has saved snapshots of websites since 1996. It has made the collection available through its Wayback Machine, a timeline tool that allows for the visualization and navigation of webpages as they were archived. From February 6, 2010, when Kaggle was launched, to April 14, 2022, the Internet Archive saved its website (*kaggle.com*) 4,634 times (see Figure 4.2). The homepage alone has been captured 4,610 times and has gone through at least 2,931 unique versions, which involve, most of

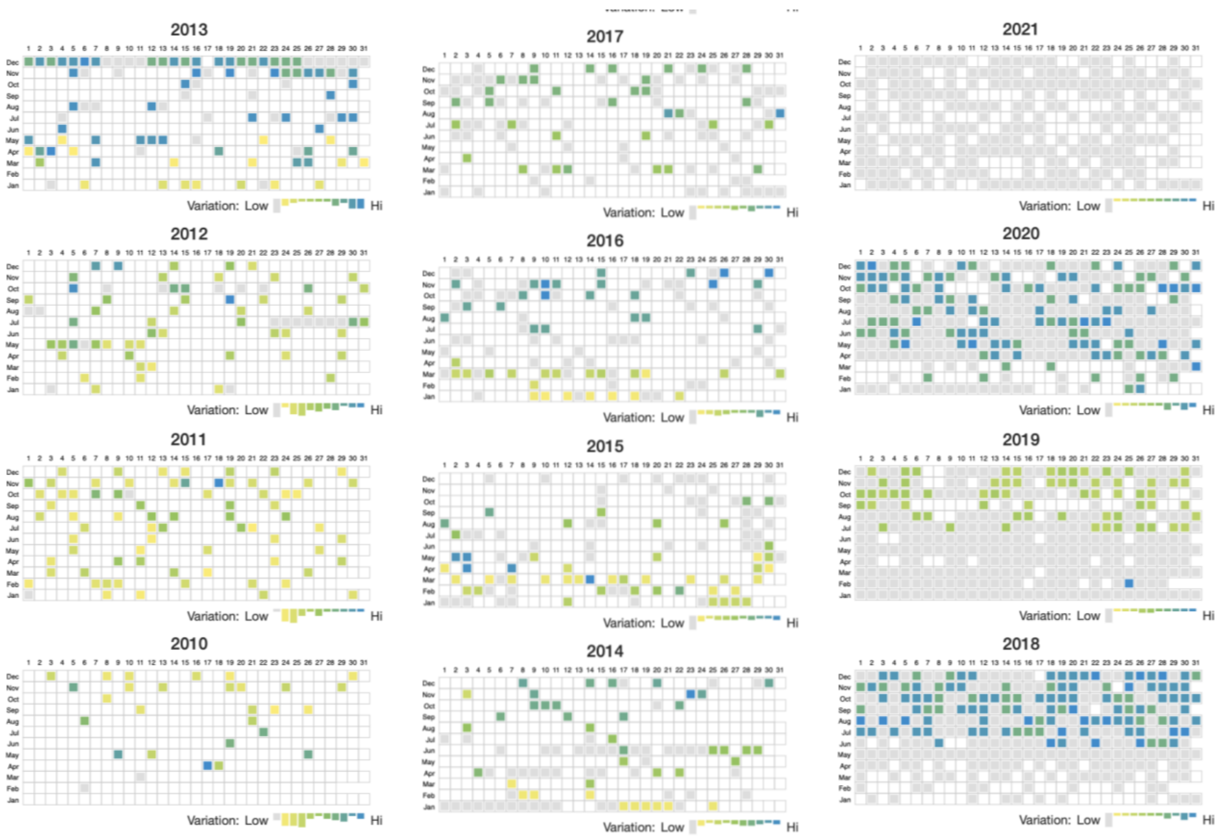


Figure 4.2: Kaggle Timeline Heatmap. Timeline showing the changes on kaggle.com captured by the Wayback Machine (n.d.). Screenshot by L. Frizzera.

the time, small changes like updating the number of users or adding a new blog post. The exploration of the many versions of a website over time offers material evidence for developing a narrative of an individual, organization, idea, or any other entity to which a website has been dedicated (Rogers, 2017). In this case, the Web Historiography approach allows us to create a biography of Kaggle's online presence by examining how the platform evolved to explore each moment of its existence.

Since its inception, Kaggle was conceived as a closed platform; only registered users would be able to sign in and see what was inside. Though the Wayback Machine captures some page resources (styles, images), it does not save back-end services, such as user authentication, making it impossible to see how Kaggle's inner pages (competitions, forums, user rankings) looked in the past. Due to these limitations—and for simplicity—I limited my analyses to Kaggle's homepage as a way to unearth the company's public face to its visitors and the platform's history. By observing the visual changes in the design and the nuances of how the company addresses the user, we can identify 11 different versions of the website, which can be further grouped into five distinct moments (see Figure 4.1, bottom): 1. Prototype (two versions in 2010); 2. Hackathon Organizer (two versions between 2010 and 2013); 3. Temp Agency (four versions between 2013 and 2016); 4. Social Media Platform (one version between 2016 and 2018); and 5. Google Cloud Service (two versions between 2018 and 2022). Let me consider each one in turn.

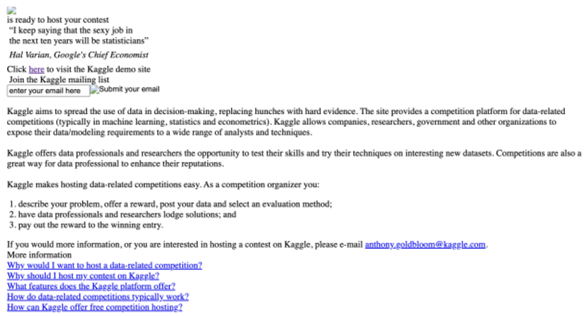
4.2.1. Prototype

First launched in February 2010, kaggle.com had little to show (see Figure. 4.3, left). For two months, it displayed a text-only website inviting visitors to join a mailing list and visit an under-construction demo site. In the process of figuring out what Kaggle ought to be, Goldbloom left traces of his initial inspirations and goals in this first version. For instance, he saw a business opportunity in Hal Varian's (2009), Google's Chief Economist, prediction in an interview for *The McKinsey Quarterly*. Varian: "I keep saying that the sexy job in the next ten years will be statisticians" (On workers and managers, para. 1). Kaggle would use this prediction to build a business model, lure its potential workforce through statements about the sexiness of the work, and then exploit their labour. Though statistics have been "sexy" since they became a trend in the nineteenth century, their popularity today is not half what it was in the twentieth century (Books Ngram Viewer, 2022). Varian might be right about the demand for statisticians in computer engineering, but it was not appealing enough to make this type of profession a trend. Goldbloom quickly realized the mistake before moving the company to California, replacing *statisticians* with *data scientists*.[6]

Another early example of Kaggle's assumptions and objectives was imprinted in the way Goldbloom described the company's endeavours:

> The site provides a competition platform for data-related competitions (typically in machine learning, statistics and econometrics). Kaggle allows companies, researchers, government and other organizations to expose their data/modelling requirements to a wide range of analysts and techniques. Kaggle offers data professionals and researchers the opportunity to test their skills and try their techniques on interesting new datasets. Competitions are also a great way for data professional to enhance their reputations. (Kaggle, 2010, n.p.)

[6] The term "Data Science" and its derivation "data scientists" are alternative names for "Computer Science" and "Computer Scientists" proposed by Peter Naur in 1974, but it was not a consensus in the field (Cao, 2017). In 1985, in a lecture given at the Chinese Academy of Sciences in Beijing, C.F. Jeff Wu used "Data Science" as an alternative name for "Statistics." However, at the 1992 Statistics Symposium at the University of Montpellier II, Data Science was finally defined as a new discipline that differs from pure statistics by combining established concepts and principles of statistics and data analysis with computing methods (Escoufier, Hayashi & Fichet, 1995).



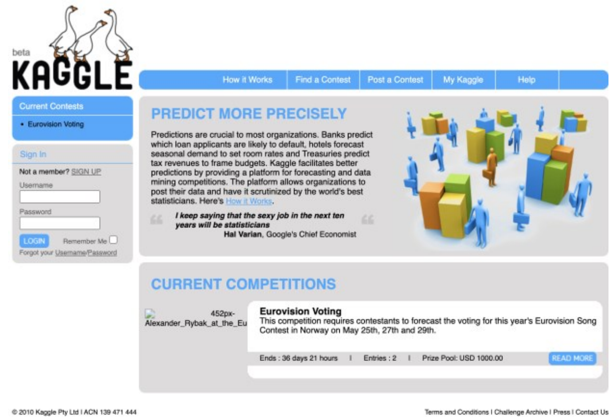


Figure 4.3: Kaggle homepage in 2010. Kaggle homepage captured by the Wayback Machine in February 2010 (left) and April 2010 (right). Screenshot by L. Frizzera.

Goldbloom was proposing that Kaggle ought to become an intermediary between near-future high-demand "sexy" professionals and supposedly ill-managed companies and public institutions in order to help them achieve a high degree of accuracy and precision in their decision-making processes. Moreover, he restated the company motto using different words: "Kaggle aims to spread the use of data in decision-making, replacing hunches with hard evidence" (Kaggle, 2010 n.p.). Here, Goldbloom calls for hard evidence to configure the concrete and supposedly objective data a company may have about its business, as opposed to the subjective opinions and emotional feelings owners and managers tend to base their decisions on. However, as we will see throughout this dissertation, hunches and gut feelings are baked into every step of data science practice, making this statement just as empty as its marketing strategies. In an interview with Reuters correspondent James Bennett, Goldbloom (2019) acknowledges that "for all the fancy mathematics … sometimes it is simple insights and simple 'a-ha' moments … that end up being the thing that wins the competition" (03:33).

In April 2010, the demo site was promoted to the front page (see Figure 4.3, right), but not for long. In a very simple and amateurish design, Kaggle's website had a logo and a short-lived mascot—three geese perched on the logo; a proper introductory description emphasizing the company's motto, "predict more precisely"; a list of current competitions; a menu bar with links to other pages; and an authentication system.

4.2.2. Hackathon Organizer

In November 2010, just ten months after being established, kaggle.com was redesigned. Perhaps in anticipation of moving the company to California, Goldbloom developed a more coherent and professional-looking homepage as a way to appeal to Silicon Valley's audience (see Figure 4.4, left). The company would now promote worldwide hackathons to promote machine-learning technologies. Hackathons are intense events where individuals with relevant skills (usually programmers and designers) meet in person to quickly prototype solutions for a problem, usually as a software demo. A hackathon can be a single-day event or span multiple days, but typically lasts no more than a week. These events usually take place at night, on weekends, or during conferences—times away from routine obligations to family, managers, or long-term plans. Over the course of a hackathon, participants form work groups, explore ways to address the theme, and push toward a ''demo''—a piece of software that supports storytelling around future technologies. Much of what people build in this context never gets built. However, this failure does not cast doubt on the process of urgent, collaborative, and frequent production of the seeds of futures: it is a place for experimentation and ideation where urgency is manufactured with bursts of optimism that technology by itself can solve any social, political, and cultural problem. As Irani (2015) puts it, "hackathons sometimes produce technologies, and they always, however, produce subjects" (p. 800). Irani described hackathons as a rehearsal of entrepreneurial citizenship celebrated in transnational cultures inspired by Silicon Valley values for models of social change. She argues that "a hackathon is not just a place where technology gets made" but also a site to promote entrepreneurialism as a "productive activity [that] could produce social surplus, national development, and global progress (p. 801).

Kaggle pushes the traditional concept of hackathons to its limits. Instead of collaboration-driven short events, Kaggle promotes competitiveness throughout multiple long-term events, sometimes spanning months. Instead of in-person meetings with organic group formation, people now articulate teams in a hyperspace based on chances to win a competition. In that, sense, Kaggle is a platform that promotes and runs hackathons on behalf of third parties. Indeed, it is no coincidence that right at the top of this version of the site, Kaggle presented itself as a "*platform* [emphasis added] for data prediction competitions that allow organizations to post their data and have it scrutinized by the world's best data scientists" (Kaggle, 2011). As Steinberg (2019) has argued, the keywords chosen to describe a business matter: it reveals how a company positions itself to its shareholders and the wider public. By using "platform," Kaggle not only signalled that it is a player

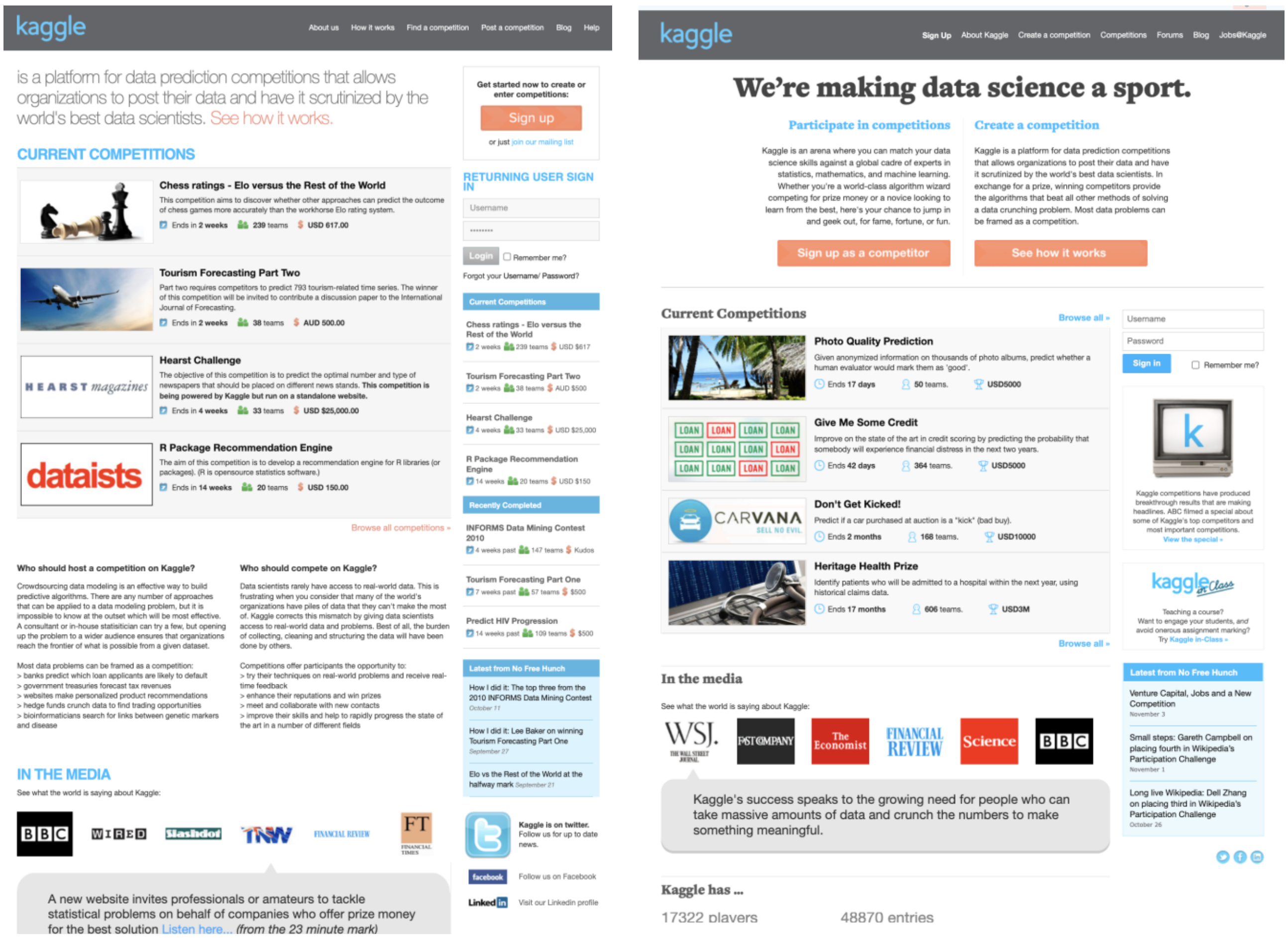


Figure 4.4: Kaggle homepage in 2010 and 2011. Kaggle homepage captured by the Wayback Machine in November 2010 (left) and November 2011 (right). Screenshot by L. Frizzera.

among other digital platforms but also reinforced concepts and practices linked to this keyword in the business and technical environment. While Kaggle was nothing more than a Hackathon Organizer at this point, Goldbloom endeavoured to have the website's audience (and potential investors) understand Kaggle as a platform equal to other already established digital platforms, such as Facebook and LinkedIn. Moreover, he was also using this managerial construct to pave the way to develop a more extensive infrastructure that would shape the work relations users enter into with the data sciences challenges, sponsor, data, and code.

The site became user-centric, but the target had changed: instead of *sexy statisticians*, Kaggle promoted *the world's best data scientists*. It invited new visitors to sign up and returning users to sign in. It created accounts on social media platforms (Twitter, Facebook, LinkedIn) to promote itself and offered a new blog titled "No free hunch," advertising the winners of past challenges and featuring news about the Data Sciences field. The list of current competitions became more prominent on the page, revealing the number of participating teams and the prizes for the winners, both in Australian and American Dollars, indicating that Kaggle was expanding quickly. In April 2011, the site launched its biggest event: the Heritage Health Prize, a US$3 million competition to solve America's healthcare crisis. The announcement was an obvious marketing approach to attract users and show off how quickly Kaggle could scale. Still, it also reads as the typical optimistic and naive discourse in which data and technology alone would be able to solve social-political problems.

Once settled in San Francisco, a second version of the site was launched in November 2011, a year after the last changes (see Figure 4.4, right). The site's look and feel remained the same, but the volume of information was reduced and some clutter was removed. The list of current competitions had only US-based events, completing the transition to Silicon Valley and leaving behind Kaggle's home country. What is different in this version is Kaggle's approach to data science. Goldbloom must have found the sweet spot to lure statisticians, mathematicians, and computer engineers onto the platform: gamification. Kaggle embraced the metaphor and imprinted it in big letters at the top of the site: "We're making data science a sport." Data science should not be a boring and bureaucratic task that one does for the government, banks, or insurance companies.: it must be an enjoyable and challenging activity that one engages in for pleasure. So, the focus changed from "world's best data science" to the "world-class algorithm-wizard" who can "geek out" in data science competitions "for fame, fortune, or fun." Kaggle sought to become a crowd of competitive programmers who would participate in machine-learning challenges in exchange for virtual medals and honorific titles like "expert" and "grand-master," a topic discussed in subsequent chapters.

4.2.3. Temp Agency

The website changed abruptly in March 2013. Things might not have gone well for the company. Despite holding 43 prize competitions in 2012, three times as many as were held in 2011, the amount of money distributed through these competitions was three times lower, forcing the company to promote changes on the website to attract more investors and sponsors in order to survive. From 2013 to 2016, the website changed significantly at least four times, switching the focus from users and competition to one focused on consulting and labour supply—ultimately resembling a temp agency. The lively atmosphere showing current competitions and blog posts gave way to a dry and institutional website (see Figure 4.5, top left). The only reference to Kaggle's main activity was a small banner at the bottom of the website: "Interested in our competitions platform? Browse the active competitions here." In this version, the website showcased three to four top data scientists as if they were its employees, offering its rich community of data scientists to help companies "go from Big Data to Big Analytics." Indeed, Kaggle was actively selling the community's workforce to large companies in specific industries, explicitly targeting software and technology, consumer goods and retail, finance and insurance, healthcare, energy, and transport, listing Pfizer, Allstate, GE, and Merck among its clients. The website clearly stated that "all you need is data and a question. Our data scientist will provide the answer." For a few weeks, Kaggle had even deployed desperate measures to attract new sponsors by offering free hours in their promotion: "8 complimentary hours to get you started in having your first data problems solved. Get your free hours or compete as a data scientist."

By the end of 2013, Kaggle had not improved the number of competitions nor the ratio between the number of competitions and awards. To attract new clients, the company once again renewed its website in December 2013 (see Figure 4.5, top right). The website's design remained dry and muted, resembling any other tech company's institutional website despite the addition of some illustrative elements, such as a network visualization to convey data analytics. At this moment, Kaggle saw itself as the "global leader in solving business challenges through predictive analytics." It kept and updated a list of prominent clients as evidence of its economic potential and commercial relationship: Facebook, GE, Mastercard, Merck, and NASA. The list of the top data scientists featured in previous versions of the site was removed. Any mention of data science competitions was long gone from the public view except for a single card showing the current challenge, along with a modest and subtle return to the idea that data science can be a place for fun and fortune. Squashed between the company's corporate goals, Kaggle still invited the visitor to "compete as a data scientist for fortune, fame, and fun."

Nine months after the last changes, the company rearranged a few elements in September 2014 but kept the site's overall look and feel (see Figure 4.5, bottom left). Most noticeably, Kaggle tried a different marketing strategy: the website's main section printed in large letters "The home of Data Science" over a faded network visualization. Below the new motto, Kaggle displays its confusing list

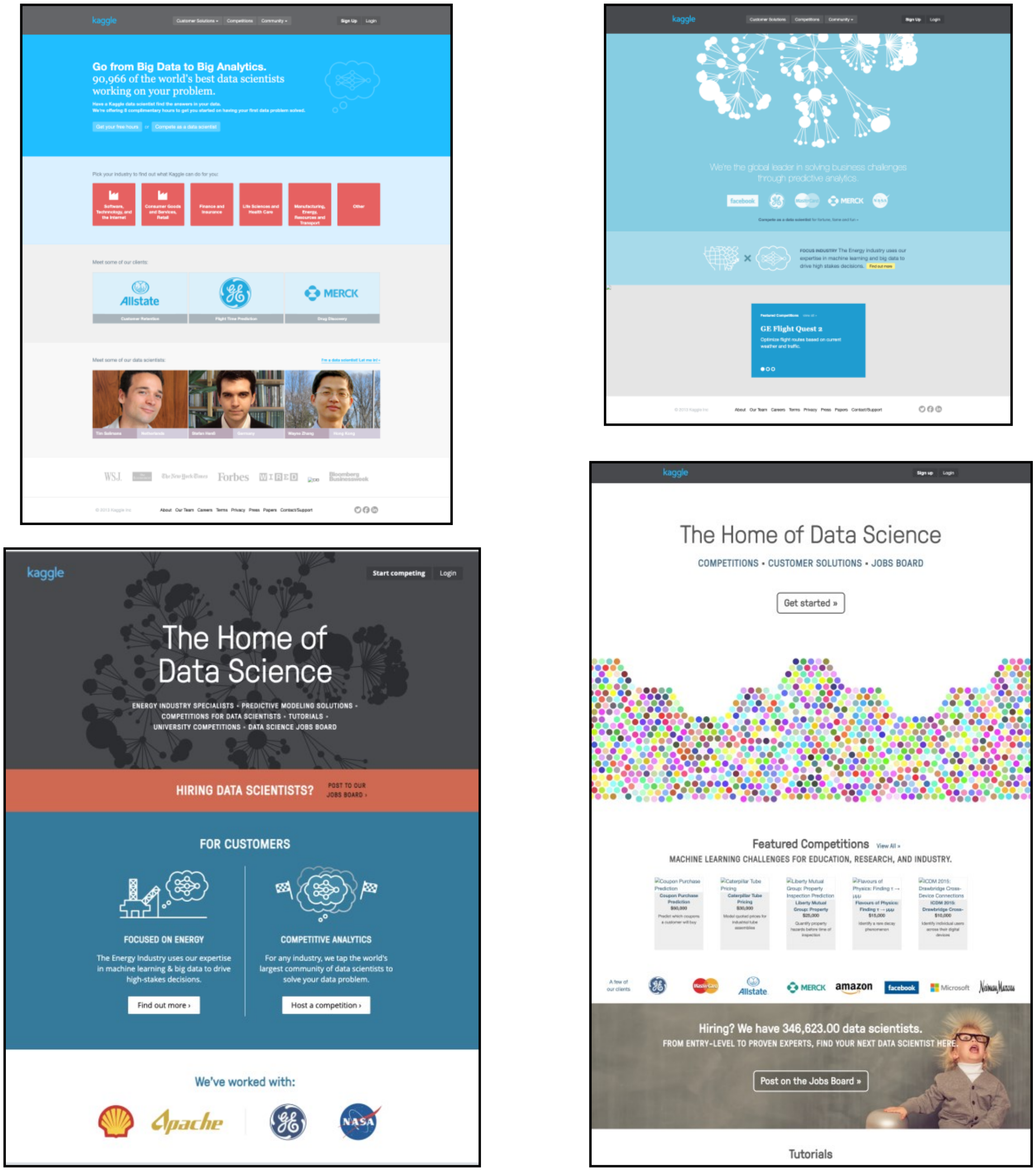


Figure 4.5: Kaggle homepage between 2013 and 2015. Kaggle homepage captured by the Wayback Machine in May 2013 (top left), December 2013 (top right), November 2014 (bottom left), and August 2015 (bottom right). Screenshot by L. Frizzera.

of features and business priorities: Energy industry specialists, Predictive modelling solutions, Competitions for data scientists, Tutorials, University competitions, and a Data science jobs board. As highlighted at the top of the page, the general frame as a temp agency remains: "Hiring Data scientists? Post to our job board." Aside from the steady increase in the number of users in 2014, which reached more than 200,000, the year was not a success in terms of monetization: The number of prize competitions shrank to 37, and the total prize shrank to US$800,000.

Once again, in March 2015, looking for new business opportunities, the company changed its website (see Figure 4.5, bottom right). This time, the header looked less formal, with multicoloured circles as a metaphor for exploring data points. The company kept the slogan "The Home of Data Science" but simplified the featured services list: Competitions, Customer Solutions, and Jobs Board. The list of current competitions reappeared, featuring at least five of them. Kaggle also wanted to make sure the visitor understood the company goals: "Machine learning for Education, Research, and Industry." In particular, it emphasized education with links to tutorials and its ties to major U.S. and Canadian universities. It held academic competitions where "theory meets practice" at Berkeley, Columbia, Harvard, Oxford, UCLA, Cornell, and the University of Toronto. Yet, the general idea of a temp agency was deeply ingrained into Kaggle's business model. This time, the website displayed a prominent banner with pictures of children dressed up as stereotypical scientists looking proud, crazy, or goofy. An unsuspecting viewer may think Kaggle was targeting young kids to become data scientists or, worse, offering child labour as an option. In fact, Kaggle was actively offering the work of a large contingent of data scientists from all skill levels to other companies: "Hiring? We have 323,934.00 data scientists. From entry-level to proven experts, find your next data scientist here." It is not so much that Kaggle considered this large number of individuals their employees; instead, Kaggle was already turning toward a platform logic, treating the cognitive potential of its users as "raw material" (Srnicek, 2017) to be sold as computational power.

4.2.4. Social Media Platform

In 2016, Kaggle decided to return to its origins: a hackathon organizer, but now with social media attributes. The site was completely revamped, switching the focus from corporations back to users (see Figure 4.6). Kaggle invited visitors to create an account (and corporations to host competitions), highlighting three main features of the platform, two of them recently introduced: Competitions, (public) datasets, and Kernels (cloud computing), about which I will go into more detail in the next chapter. The site, however, was less transparent compared to the 2011 version. As on other social media platform homepages, there is no way to see what is happening inside: no current competitions, no top data scientists, no blog posts. The homepage was just a way to lure visitors to sign up and get in. This change in strategy might explain how the company raised an extra US$5 million in a second round of a Series A investment later in the year, a few months before being acquired by Alphabet.

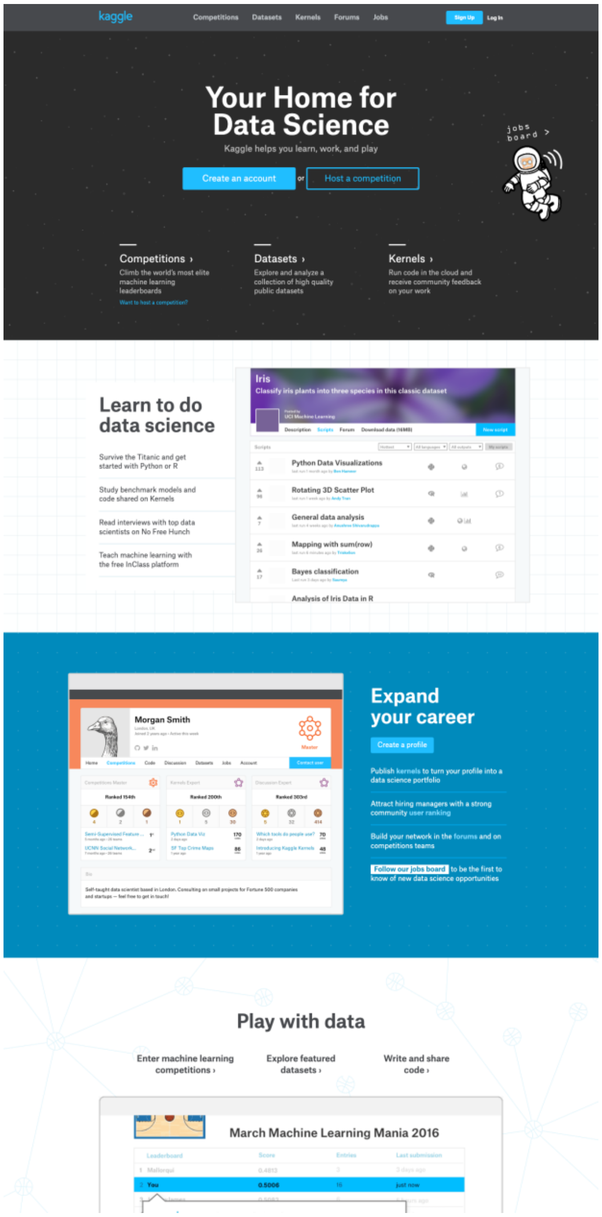


Figure 4.6: Kaggle homepage in 2016. Kaggle homepage captured by the Wayback Machine in October 2016. Screenshot by L. Frizzera.

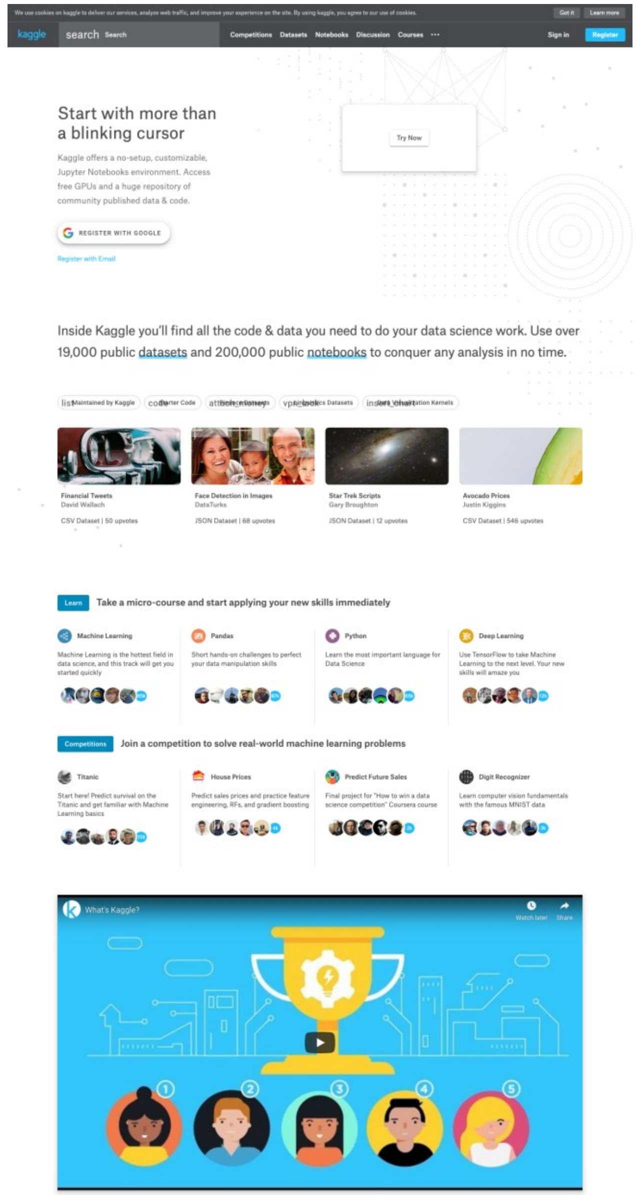


Figure 4.7: Kaggle homepage in 2019. Kaggle homepage captured by the Wayback Machine in October 2019. Screenshot by L. Frizzera.

At first, the website addressed the visitor in a more personal and informal way, marketing Kaggle as "Your Home for Data Science." A few months after being incorporated into Google Cloud, the motto changed to give more specificity to the platform: "The home of data science & machine learning." In any case, Kaggle saw itself as the place that "helps you learn, work, and play." The reference to learning promotes its educational portal with machine-learning workshops where users can study benchmark models and read interviews with top data scientists. Once the user is engaged, Kaggle doubles as a professional social network where the user can create a portfolio using code and datasets and begin to build a reputation by collecting medals and titles, as well as find opportunities on the job board and attract companies who are hiring. The third dimension to Kaggle's new approach is a return to its origins: the hackathon as a place where people get together to share knowledge, solve problems, and have fun. The concept of "play" here has not only smoothed out the complex work relations involved in data science competitions, but also served to keep the user engaged with the platform and with the challenges held by third-party companies.

4.2.5. Google Cloud Service

In July 2018, Kaggle's website was remade once again to reflect what it had become after Alphabet's purchase: just another Google Cloud Service. The homepage had very little information about what was inside. The motto and graphics changed, removing the metaphor of an informal gathering and moving to a more formal workplace: "Kaggle is the place to do data science projects." For over a year, Kaggle's homepage was just a dry and empty vessel to showcase Goggle's new service. Alphabet would correct course and produce a friendlier and more marketable website for Kaggle in September 2019 (see Figure 4.7).

Entering its latest and most current stage, the site has become standardized, following Google's Material Designs guidelines. However, Kaggle is no longer about a community of geek coders looking for complex data science challenges. Indeed, except for a cartoon-style video explaining what Kaggle is about at the bottom of the page, the original idea of having fun and pursuing fame and fortune is gone. The new and catchy motto says it all: "start with more than a blinking cursor." It has become a place to sell easy-to-use Google Cloud infrastructure packaged together with the work done by a legion of coders that, more often than not, are working for free: "Kaggle offers a no-setup, customizable, Jupiter Notebook environment. Access free GPUs and a huge repository of community published data & code. Try now." Below it, the site claims one can find all the code and data for one's data science work, inviting the visitor to use one of the more than 19,000 public datasets and 200,000 notebooks, updated from time to time to reflect the ever-growing contributions from the community. The site also lists a few of these datasets together with some of the micro-courses related to machine-learning development and a couple of competitions to give visitors an idea of what they will find inside.

4.3. Strategic Vision and Business Model

In his keynote presentation at the 2013 Digital-Life-Design conference, Max Levchin, an early Kaggle chairman and investor, spelled out the business model imagined by Goldbloom and praised by Alphabet's CEO: "Just like the SETI@Home screensaver 'steals' CPU cycles to sift through cosmic radio noise for alien voices, your brain plug firmware will earn you a little extra cash while you sleep, by being remotely programmed to solve hard problems" (Levchin, 2013, para. 21). He concluded by saying that "opportunities to build businesses that process this data and improve lives will abound" (para. 33). Levchin's statement shows little concern with human beings, revealing an understanding that people are merely resources that are not being utilized correctly or efficiently. He is convinced—if not obsessed with the idea—that sensors, big data, and real-time information can improve humans at its fundamental level, in much the same way that programmers try to optimize the clock cycles of a microprocessor. Levchin claims that, as a species, "*we must* [emphasis added] use all available resources to their maximum" (para. 24), suggesting that the best way to optimize society is to harvest individuals' free time and cognitive power as a collective of cogs. He sees the path to the future pointing toward Kaggle's model: "today you can already rent the brain of a data-mining genius via Kaggle by the hour, tomorrow by brain-hour" (para. 21), wherein private companies can take advantage to improve services and provide better products. Levchin understands that this model can be unfair and discriminatory but "believes that what we can enable with data insights greatly outweighs the downsides" (para. 28). Levchin dreams about a wired society where a crowd of a million workers would happily work for pennies to solve problems, not for their own benefit, but for large companies to increase their profit. When he gave the keynote, Kaggle had about 100,000 users; in 2021, it surpassed 5 million.

Kaggle comprises an active community of users who are always ready to participate in and contribute to developing new predictive models, no matter the impact or intention, ranging from predicting the next NBA champion to identifying whales on satellite images. However, more often than not, the "jobs" launched on the platform are in some way or another related to business optimization, looking to find "interesting" individual behaviours and social relations that private companies can exploit to cut costs and increase profits. In the process of producing these models, the community itself is exploited by the platform and the sponsors as free labour. The community is responsible for uploading new datasets, building and debugging code, and training machine learning in their free time. Without this crowd of people to do the job, Kaggle would be no more than just another consulting firm. The company crowdsources most of the activities on the platform, heavily relying on its flourishing community to solve complicated problems using machine-learning techniques.

Coined in 2006 by *Wired Magazine* contributing editor Jeff Howe, "crowdsourcing" is a portmanteau of "outsourcing" and "crowd" of online labour. Brabham (2013) defines it as "an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals" (p. xix). In this arrangement, the control over the relations of production (both material and immaterial) exists in a shared space between the organization and the public: an emerging collective and open creation emerging from the bottom-up together with top-down management that oversees the production by those in charge serving the organization's strategic interests. In other words, crowdsourcing can be described as the processes of problem-solving and innovation through the group phenomenon of collective intelligence. "The crowd" configured by Internet users was once praised as the liberating and democratizing form of work where people from different regions of the world with multi-faceted cultural backgrounds and diverse interests would collaborate, cooperate, and execute creative work. In 1994, Pierre Lévy (1999) proposed that the rise of the Internet as a global network that connects individuals would give rise to the ability to coordinate collective and creative work. He put forth the idea of collective intelligence as a form of universally distributed intelligence, continually coordinated and improved in real-time, resulting in the effective mobilization of diverse skill sets. Surowiecki (2005) calls this phenomenon the "wisdom of crowds," where, under the right conditions, groups of people can outperform even the best individuals and experts. Clay Shirky (2009; 2011) has a more pragmatic and utilitarian approach, arguing that digital technologies and social media would decentralize ways of organizing and transforming consumers into collaborators.

Besides the numerous arguments for its democratic potential, the Internet's links with capitalism have always been a bit too strong.[7] Terranova (2004) points out that "computer networks are the material and ideological heart of information capital" (p. 80). As such, the productive potential of crowdsourced labour was quickly captured by private companies that saw the opportunity to outsource expensive and laborious tasks. It is an easy path to cheap labour, high profit, and low-risk endeavours. Indeed, distributed work, whether outsourced overseas or crowdsourced over the Internet, has become a symbol of global capitalism seeking to disorganize markets by sponsoring measures to reduce the power of labour unions, remove basic labour protections, and lessen the

[7] See "The Californian Ideology" (Barbrook & Cameron, 1996).

regulations and accountability of private companies (Dyer-Witheford, 1999; Lazzarato, 2004; Terranova, 2004). The precarious work conditions generated by these measures in the digital economy are viewed as a new form of sweatshop, also called “click servitude,” “digital slavery,” and “crowdsploitation” (Brabham, 2013), which are the essential characteristics of the so-called platform economy” (Srnicek, 2017) or “sharing economy” (Eckhardt & Bardhi, 2015; Schor et al., 2015), such as Uber and Airbnb. The contentiousness of this business model transforms the original ideas of community-based production into a set of rigid rules that can be automated and controlled, making them attractive to large companies. In fact, the immaterial labour developed by a massive contingent of people involves a series of activities that are not generally recognized as “work” in the traditional sense; instead, they are seen as free “collaborators” that execute some sort of task as a mundane activity in their own lives.

It is worth noting that pervasive technology does not automatically turn every user into an active producer and every worker into a creative individual. The process whereby production and consumption are reconfigured into a category of free labour reveals the unfolding of another logic of value in which operations require careful analysis. For instance, in the industrial economy, while workers attempted to achieve satisfaction through leisure, they were at the same time alienated from the means of production, which were owned and controlled by someone else. In the digital economy, on the other hand: “the worker achieves fulfillment through work and finds in her brain her own, unalienated means of production” (Terranova, 2004, p. 79). However, extra effort is needed to ensure the exploitation of the workforce. Such means of production must be cultivated, and workers need to be encouraged to participate in a culture of exchange in which flows are primarily kept within the company but also need to involve an ‘outsider’ contact with the fast-moving world of knowledge in general. That is, the outside participant must be encouraged to contribute and should believe that they are doing so for themselves and the good of the community. As a result, these ‘volunteers’ are not always aware that they are engaging in productive work; rather, they are often motivated by a desire for affective and cultural productivity, which is just as real because of its social construction.

On Kaggle, crowdsource tasks are gamified (Whitson, 2013) as a means of making the company, the platform, and the work attractive to an economic system that champions individualism and competitiveness. It emerges as a mechanism that shapes new work conditions (a topic I will discuss in more depth in the following chapter), as it circumscribes the work relations on the Kaggle platform, as well as defines the social conditions in which both machine learning and predictive models are built and disseminated. The social and the game components that drive Kaggle make developers eager to participate in competitions not only for the prizes but also because they are looking for opportunities to get involved with the community, to learn from one another, to have a sense of belonging, and to have their names attached to something bigger than themselves. The platform exploits its users' affect as a "stock of brains," a commodity, which, in turn, becomes a workforce willing to exchange their cognitive abilities, if not for free, at least for low-value tokens of appreciation (scores, virtual medals, badges, made-up honorific titles).

Broadly speaking, Brabham (2013) proposes that crowdsourcing comprises four main elements: (1) an organization or company with a task that needs to be executed; (2) a community to perform the task voluntarily; (3) an online platform that enables community engagement and allows the work to take place; and (4) mutual benefit for the organization and the community. Many of the early crowdsourcing companies and initiatives began at the onset of the Internet, right in Silicon Valley, where the economic and technical conditions were able to capitalize on the articulation of technology, creative energy, and community. As Brabham puts it, "the technologies and social relationships that were fostered by those technologies were the fertile ground in which crowdsourcing took root in the early 2000s" (p. 18). Amazon's Mechanical Turk is perhaps the most (in)famous crowdsourcing experiment, where organizations outsource micro-tasks to an online community of workers (Brabham, 2013; Crawford, 2021; Finn, 2017). These are usually tasks computers cannot easily perform, such as accurately tagging the content of images for a search engine or machine-learning algorithms.[8] "Turkers," as the Mechanical Turk community called themselves, sign up to accomplish a series of these "human-intelligence tasks" in exchange for extremely low micro-payments, often not more than 50 cents per task, no matter how long or complex the task. Brabham (2013) classifies this type of crowdsourcing as "distributed-human-intelligence tasking," which is ideal for large-scale data analysis where human abilities are more efficient or effective than computer power. Essentially, Mechanical Turk coordinates the large-scale collection of simple tasks that require human intelligence for organizations that use the service to parse data quickly and inexpensively.

Kaggle falls into a different category of crowdsourcing, identified by Brabham (2013) as "Broadcast search." In this approach, an organization tasks a crowd with finding more efficient and optimized solutions for complex empirical problems. According to Brabham, this type of crowdsourcing is

[8] There are plenty of companies offering similar services now, particularly image labelling for machine-learning training; for instance, Google Cloud AI Platform, The Hive, and Cloudfactory.

appropriate for situations where a provable empirically "right" answer exists but is not yet known. In this sense, Kaggle's crowdsourcing arrangement is not a novelty. Ten years prior to Kaggle, in 2001, InnoCentive (n.d.) started its operations with a similar proposition: companies are invited to share data about complex scientific research or development challenges online and offer cash bounties for solutions. In exchange, InnoCentive's online community members answer scientific puzzles, allowing companies to gain valuable insights more quickly and cheaply than their in-house scientists might have struggled to develop. Like InnoCentive, Kaggle combines complex data science challenges that are usually related to business efficiency or service optimization with specific rules and pre-defined accepted solutions provided by the sponsoring company so that they might pursue their strategic interests. Unlike Mechanical Turk, where the tasks are designed to serve the organization's needs while paying a small amount of money to the worker, Kaggle and InnoCentive harness their community production as free labour while offering prizes up to US$ 3 million for the top competitors.

Because it targets the development of algorithms, and machine learning in particular, Kaggle also has roots in a different type of crowdsourcing initiative: "bug bounties." A bug bounty program is a job opportunity where security researchers legally trade newly discovered security exploits and vulnerabilities for monetary, recognition, and reputation rewards. This practice can be traced back to 1995, when Netscape offered to pay US$1,000 for vulnerabilities found in its Internet browser Netscape Navigator 2.0. Mozilla started a similar program in 2004, offering rewards for security exploits and bugs in Firefox (Elazari, 2018). Private companies and intermediary platforms dictate the terms of the bug bounty economy, usually transferring to the individual work any hazard related to the job by "using multiple layers of unilaterally drafted 'take-it-or-leave-it' terms, that often put hackers in 'legal' harm's way—shifting the risk for civil and criminal liability towards hackers instead of authorizing access" (p. 3). However, instead of looking back to fix cybersecurity flaws, Kaggle looks forward, promoting the active creation of new algorithms and predictive models. As such, Kaggle business is modelled on contest events like the U.S. Defence Advanced Research Projects Agency (DARPA) grand challenges centred on autonomous vehicles (Darpa, n.d.), the Loebner Prize—the annual Turing test competition (University of Exeter, 2011), and the Netflix Prize (2009). Though Kaggle hosts a few scientific challenges like the first two examples above, it is closer to Netflix since it focuses on complex commercial applications of artificial intelligence capabilities.

In 2006, Netflix offered US$1 million to the first individual or team to improve the efficiency of their recommendation algorithm by 10%. At the time, Netflix was trying to predict future users' choices based on five-star ratings and user history, completely ignoring the complex position of Hollywood entertainment and movie rental as a culture machine (Finn, 2017; Hallinan & Striphas, 2016). After three years of intense competitive development among more than 50,000 rival teams comprising computer scientists and statisticians, the team "BellKor's Pragmatic Chaos" won the prize by barely improving the original algorithmic effectiveness. However, Netflix never implemented the winner's algorithms despite this long and expensive effort. Part of the reason has to do with potential privacy violations during the competition that prevented the company from using the algorithm. The main reason, however, came from the realization that the whole approach to the problem was wrong (Finn, 2017). Netflix framed a cultural activity as a purely mathematical problem and asked a community of engineers and statisticians to come up with solutions to an "equation." It reduced complex social and cultural activities across multiple populations from different backgrounds to a five-star rating system and other quantitative data points. However, cultural choices are never that simple. As Finn (2017) put it, "the logic of the culture machine trumps the statistics, and humans are left to recreate meaning in the space between the ascetic structure of five-point rating scales" (p. 89). The cultural emptiness of the Netflix Prize may have improved the company's algorithms, but, in essence, it was an effort to reinterpret what culture is—how it is evaluated, by whom, and to what ends (Hallinan & Striphas, 2016), which are now in the hand of engineers, mathematicians, and marketers.

The Netflix Prize illustrates the emerging challenges of algorithmic culture, where issues of quality or hierarchy are transposed onto matters of fit and optimization. Both Kaggle and Netflix are examples of pure algorithmic thinking applied to modulate human behaviour in order to solve a business problem, except that Kaggle takes two steps forward: (1) it offers data science and machine-learning consulting services for large corporations, and, more importantly, (2) it acts as a lean platform (Srnicek, 2017). Kaggle defines itself as an intermediary between large corporations and a skilled labour reserve (at the Master's and PhD levels), mainly unemployed or earning low wages, scattered across the globe, most notably in India (Kaggle, 2021).[9]

[9] According to Kaggle's (2021b) yearly survey, only 14% of respondents are employed as data scientists. Almost 65% hold a master's degree or PhD. India alone represents 24% of Kaggle's data scientists. The average salary paid by Indian companies is 30 times lower than in the U.S., putting Indian data scientists in a precarious position compared to their American counterparts. This inequality puts pressure on the workforce, especially in poorer countries, which will accept

Kaggle's business model was structured to generate revenue from machine-learning competitions. Kaggle was both a platform to host competitions and a consulting firm to advise sponsors on designing a machine-learning competition. The company charges a fee to oversee the process, makes its cloud infrastructure available, and assists with the competition's operations, such as selecting and cleaning data and defining expectations within the limits of the desired outcome. The sponsors pay the fees and provide the data and money prizes for the competitors. In a brochure used to sell its services, Kaggle states that "competitions work best when participants are asked to predict one specific variable from a dataset, based on a ground truth outcome, with a specific and clear data problem to be solved" (Kaggle, 2020c, p. 6). In the same document, Kaggle illustrates that a typical commercial budget to host a competition ranges from US$85,000 to US$200,000. This value includes the infrastructure cost, consulting fee, and the prizes awarded to the participants, which, for a competition sponsored by a for-profit-oriented company, should be at least US$25,000 (Hamner, 2015; Kaggle, 2020c). Kaggle also offers a three-phase-template and a timeline to host a competition that takes into consideration the specific requirements and complexity of the task: a four-to-six-week planning phase where Kaggle provides assistance and support to design the event, followed by a 12-to-18-month competition, during which competitors have access to the data and cloud computing power to run their algorithms and develop their predictive models, and then a two-week phase in which the sponsor will work together with Kaggle to assess the predictions against a private test dataset to announce the winners.

As we have seen, Kaggle started as a forum-like website to advertise hackathons. It promoted serious work as a fun activity where people could work together, learn together, and disrupt the industry together. But their founders dreamed about remotely controlling brains and using individuals' idle cognitive capacity in exchange for cash. While togetherness and fun may be useful keywords to lure individuals into work, it is not a scalable business model to realize the potential envisioned by Goldbloom and its investors. The company was quick to adopt the keyword "platform" to define itself: "Kaggle is a *platform* [emphasis added] for data prediction competitions that allow organizations to post their data and have it scrutinized by the world's best data scientists" (Kaggle, 2011). For Steinberg (2019), platforms are much more than just words: "they are also technological and managerial constructs that mediate our relationship to our worlds, that create habits, addictions, and impulses" (p. 3). That is, Kaggle is not just a website that promotes

unfair rates for high-quality work. While Kaggle ignores the unbalanced work relations on its platform, the sponsors take advantage of the situation to outsource R&D and make a profit with incremental increases in predictive model accuracy.

hackathons and hosts competitions, but managerial constructs that shape the users and the relations they enter into with the platform's challenges, companies, data, and code. Following Irani (2015), the work practices associated with software production have come to signify collaboration, voluntarism, optimism, and wealth, which have been tested in software practice and are ready to enter new domains of public life. Hackathons legitimize technoscience to remake culture and produce entrepreneurial subjects (p. 800).

In Steinberg's (2019) terms, Kaggle is a transactional platform that uses its infrastructure to mediate multisided markets. As such, the key advantage for Kaggle is that it positions itself between users and as the ground upon which their activities occur, giving the company privileged access to record these activities. By being the mediator and the "medium," the platform operator asserts power over all entities and agents it mediates, whether these are data, code, and machines or users, companies, and institutions. More specifically, Kaggle fits into Srnicek's (2017) topology as a lean platform since it strives to exploit available assets in society at the lowest cost possible without directly owning them, with Uber as one of its icons: "Uber, the world's largest taxi company, owns no vehicles" (p. 76). Paraphrasing Srnicek, we can define Kaggle as the world's largest data science company that owns no data, no data centres, and no code (at least before Google's acquisition).

Indeed, Kaggle has always advertised itself as a social media platform: an online community of data scientists and machine-learning practitioners. Beyond offering consulting and technical services, Kaggle is a platform that regulates and interferes in the relationship between "Kagglers," as users called themselves, and third-party companies, which was one of the main reasons Alphabet bought Kaggle in 2017. Without a large-scale social media platform to call its own, Google uses Kaggle as a social media niche to promote its services and infrastructures among a specialized public and steer the community to produce specific kinds of machine-learning solutions. Kaggle/Google does more than use the crowd to solve problems; they use it as a marketing and business discourse, capitalizing on cultural terms like "community" and "collaboration" as substitutes for economic terms such as "consumers" and "commodities" (Van Dijck & Nieborg, 2009). This discourse has been used as a means to lure more participants to the platform through affect and the desire to be part of a collective (as experienced on the platform) while at the same time promoting a particular image of the company that is designed to appeal to the public and investors (as seen on the earnings calls, annual reports, and public events such as developer conferences).

Lastly, we must acknowledge that the collective aspect of labour in Kaggle implies a rejection of the equivalence between labour and employment. The hackathons promoted by Kaggle are a means to recruit volunteer labour, generate interest in social or technological platforms, use participants to explore possible futures for a host organization and to cultivate speculation about technological futures (Irani, 2015). However, Terranova (2004) argues that "labour is not equivalent to waged labour" (p. 88), but it is also not the same as exploited labour. People do things also because they like to: for leisure, as altruistic gestures, for a cause they believe in, and because they care. Free labour is appropriated by capital as one of the key means of maintaining structures and functionality. Capital forces realize that the best way to keep businesses operating and thriving is to turn them into spaces accessible and somehow built by their users. In this sense, Terranova (2004) is right to argue that "users keep a site alive through their labour, the cumulative hours of accessing the site (thus generating advertising), writing messages, participating in conversations and sometimes making the jump to collaborators" (p. 91). Terranova's arguments describe the foundation upon which major social media platforms were built and help us understand how crowdsourcing would become a driving force to solve complex business problems by creating and feeding machine-learning algorithms, artificial intelligence systems, and smart devices on Kaggle.

Such a reliance, a conditional dependency among elements (business, crowd, data, code), is part of the more extensive capitalist extraction mechanisms of value intrinsic to late capitalism. In other words, such processes are not created outside capital in order to be appropriated. Instead, they result from a complex history where the relationship between labour and capital is mutually constitutive, entangled, and crucially forged during the second part of the twentieth century and the beginning of the twenty-first. Consequently, late capitalism "nurtures, exploits, and exhausts its labor force and its cultural and affective production … it is technically impossible to separate neatly the digital economy of the Net from the larger network economy of late capitalism" (Terranova, 2004, p. 94). Far from being confined to the most self-conscious of labourers working in the digital economy, these processes are part of a much larger diffused cultural economy that operates throughout the Internet and beyond, omnipresent via our digital gadgets, continually reminding us that we can and should be useful and productive subjects of society. On Kaggle, this process is more than just a way of exploring a possible future and pushing machine-learning development. It also involves "rehearsals for future employment, partnerships, or investments" (Irani, 2015, p. 804). Kaggle can be understood as just another instance of these interdependent production relations. The sociotechnical production actively developed on the platform automatizes and reinforces this logic into the source of code of allegedly objective machines, a recurrent theme discussed throughout this dissertation.

4.4. Summary

Though it operates in a niche market, Kaggle has attracted the attention of data scientists, computer engineers, large corporations, and investors. Founded in 2010, the company started as a small hackathon organizer in Australia before moving to California to become the home of the world's largest data science community focused on machine learning. Based on a crowdsource competition model, Kaggle made its name by luring computer engineers and statisticians to its platform with a challenge to solve complex problems in exchange for money prizes and virtual medals for the winning teams. After receiving a first round of Series A investments from tech industry moguls, Kaggle and its founder, Anthony Goldbloom, became a gravitational pole around the machine-learning topic, occupying centre stage both in the media—with Goldbloom named one of Forbes' "30 under 30" individuals—and among the world's finance leaders at the WEF, where he has been invited to share his vision about the future of artificial intelligence and digital platforms.

Though its user base steadily grew in its first ten years, Kaggle's path has not been without mishaps. The company's business model aimed to generate revenue from machine-learning competitions. However, the strategy proved unprofitable due to an immature market for machine-learning development in the early 2010s. As it failed to secure Series B investment, Kaggle burned through its initial investors' money to survive. In 2013, Goldbloom refocused the company, giving it a more traditional approach as a consulting firm and temp agency but failing to bring in more investors and sponsored competition. In 2016, the company took advantage of the scale and highly skilled community to return Kaggle to its original objective as a hackathon organizer with the same contours of a social media platform, transforming its business logic from crowdsourcing to the platform economy and finally securing a second round of Series A investment. With cloud computing resources, a public dataset repository, and an ever-increasing number of competitors on its platform, Kaggle surfed the wave of machine-learning hype. Soon after, in 2017, Alphabet acquired the company for an undisclosed amount to expand its cloud service and incorporate the data science community as a crowd of users to serve as a labour market reserve for current and future developments in machine learning and artificial intelligence.

Goldbloom repeatedly described the company as innovative, transformative, and even disruptive because of its unique approach to data-driven business: a competitive crowdsource data science platform (Goldbloom, 2016; 2019; 2020; WEF, 2013; 2015). This description not only sought to fit Kaggle within the Californian start-up culture (Barbrook & Cameron, 1996) but also has a deeper connection to how Goldbloom pictures the role of statistics and technology both in his company and the algorithms built on Kaggle. Drawn from econometrics, the company's motto—"help businesses make decisions on the basis of data rather than gut instinct"—serves as justification for using statistical models to solve economic problems. Firmly stuck in an empiricist tradition, econometrics is only concerned with the measurable aspects of reality, revealing a tendency to artificially isolate business data from human perceptions and intuitions, essentially separating economic processes from social phenomena. In other words, it subscribes to a particular line of thought that presupposes a prior distinction between object and subject (Grusin, 2015; Kember & Zylinska, 2014), wherein objects are understood as neutral and unbiased empirical data that hold the truth about the world, and subjects are always "contaminated" by theory, ideology, and politics that distorts or prevents facts from emerging (Laudan, 1977).

Econometrics, and, to a large degree, the work of machine-learning algorithms, betrays itself by not contending with only making optimal predictions but also by aspiring to explain things in terms of causes and effects through statistical models (Moosa, 2017; Syll, 2018). The application of econometric methods presupposes that phenomena of our real world are ruled by stable causal relations among variables, resulting in a large number of assumptions, preconceptions, and biases built into and often hidden within these models. However, real-world social systems are not governed by stable causal mechanisms or capacities since there is always the possibility that other variables, not necessarily epistemologically inaccessible, are not considered within the model (Moosa, 2017). Consequently, econometrics can never be more than a starting point in that endeavour since statistical explanations are not explanations in terms of mechanisms, powers, capacities, or causes. As Syll (2018) pointed out, "all science entails human judgement, and using mathematical and statistical models does not relieve us of that necessity. They are no substitutes for thinking and doing real science" (p. 6).

The epistemology of traditional empiricists and the assumptions carried out by econometrics are crucial issues in understanding how predictive models are developed within the context of Kaggle competitions. In this chapter, however, it suffices to say that the underlying assumptions shared by Goldbloom and Kaggle's community about the goals of predictive models carry out a set of stable conditions of Kaggle's existence as a company and as a platform. These conditions are defined by the current historical political economy paradigm (i.e., capitalism) that proclaims a universal and undeniable objective truth about the world, such as the prioritization of economics over other aspects of life, in which processes should be optimized for business efficiency economic growth, and profitability. This explains why Kaggle is neither inventive, transformative, nor disruptive, despite its founders' pitch. Quite the opposite: Kaggle is rather conservative, static, and convenient for the status quo. Inventiveness means the ability to create new things or to think in an original manner. Kaggle recycled past experiences with crowdsourcing computing-oriented competitions such as bug bounty programs (Elazari, 2018) and the KDD Cup and the Netflix Prize (2009), and copied business models and practices from other companies like InnoCentive (n.d.), founded ten years earlier. To be transformative is to cause some form of change in how things are done. Kaggle's approach to using data and statistical models to drive economic and business decisions is not new: entire economic sectors used these tools in the nineteenth century (Finn, 2017; Turow, 2017). Nor is using computers and, more specifically, machine-learning algorithms to aid business decisions: Google has been using these technologies since the early 2000s (Alphabet, 2016). Perhaps more importantly, disruption entails breaking a paradigm or at least making a radical change to an existing industry or market. Kaggle does none of these. Instead, it subscribes to the assumptions inherited from the current dominant economic model. Not only has Kaggle received investments from venture capitalists interested in developing the platform for a profit (Crunchbase, 2021), it has also followed and promoted the neoliberal agenda exploiting work relations, which exhaust its labour force and its cultural and affective production (Levchin, 2013; Terranova, 2004).

In short, platforms like Kaggle define the problems, the objectives, the metrics of success or failure, and the solutions. They determine how resources and people are organized, who is valued in what roles, what activities are undertaken and for what purpose. By harnessing the crowd to work for free, Kaggle not only creates the first essential condition to generate value from aggregate data but also produces what Lazzarato (2004) calls "conditions of existence." That is, the possibility of using subjective materials to reorganize life according to specific hierarchies of power. The logic of accumulation, or the mode of production, as pointed out by Marxist thought, produces its own social relations together with its conceptions and uses of authority and power. The questions are, in turn, what kinds of subjectivation occur on Kaggle, and in the predictive models developed in the context of its competitions? What value do these predictive models have, and for what purposes have they been designed? To understand all this we must understand the profile of its community, what kinds of datasets are used to create these predictive models, and how the competitions unfold inside the platform. The following chapters will shed light on these questions and examine how algorithms have been intervening and modulating individual behaviours.

5. The Machine-Learning Game: Infrastructure, Community, and Competitions

> We manage a community of over a hundred thousand data scientists and statisticians … we've ranked them from 1 to 100 thousand by actually looking at [how] effective they have been in predictive modelling, [and] how predictive their models are. So, we are the only source in the world of the ranked world's best data scientists. (Anthony Goldbloom and Jeremy Howard in World Economic Forum, 2013, 0:40-1:04)

In an interview with the Chai Time Data Science podcast in 2020, Anthony Goldbloom revealed that he was not an entrepreneur and had little contact with Silicon Valley's start-up culture before Kaggle. When he envisioned Kaggle in 2010, Goldbloom was not planning for a large-scale platform where millions of people collaborate and compete among themselves to produce machine-learning algorithms. Instead, he was trying to figure out how to financially support himself by doing something fun. In the interview, he recollects that, at that moment, "I was not really enjoying my job as a statistician or an econometrician … I thought that this website [Kaggle] would be a really really fun thing to work, and I was hoping that it could support me. That was the initial goal" (Goldbloom, 2020, 8:47-9:10). Kaggle was not conceived as a machine-learning powerhouse but rather as a "gig" for bored and unhappy statisticians looking for a more objective and exciting approach to increase productivity (4:50) and make money. As discussed in the previous chapter, the first public version of Kaggle was no more than a static website inviting visitors to join a mailing list and participate in informal machine-learning challenges. There was no clear plan for organizing these competitions, creating the infrastructure to support multiple users, and monetizing Goldbloom's idea.

The relocation from Australia to the U.S. and the partnership with Jeremy Howard and Max Levchin transformed Goldbloom's initial goal. More than just a toy for financial support, Kaggle became the selling point for a gamified crowdsourced machine-learning economic model. For that to work, it needed an entry point to attract and lure computer engineers, data scientists, and hobbyists into devoting their time and resources to participate in challenges to solve complex problems for private companies in exchange for “fame, fortune, or fun.” In 2011, the website was redesigned to become user-centric. It incorporated Web 2.0 features, such as a blog and connections to social media platforms (Twitter, Facebook, and LinkedIn). The competitions were organized into dedicated pages with instructions, rules, direct links to the datasets, how submissions would be evaluated, the prizes, and a leaderboard where users could check their progress and compare results with other users. However, most tasks were handled manually by the user: downloading each dataset from a list posted on the website, developing algorithms, training models, and even the submission was made by email or via Web form. Most of the work—writing and running code—was done on the user’s computer (or in the infrastructure they could buy).

It was only a few months prior to being acquired by Google in 2017 that Kaggle actually began to function as a digital platform. In 2016, the company made changes in how events and competitions are organized and launched new features to attract more users and investors. Most notable was the introduction of “Datasets,” a free and open data repository where users can contribute and store data specifically for machine learning, which made Kaggle double as a data repository, storing more than 50,000 datasets for machine-learning training, such as search engine results, crime data, commodity prices, credit and financial fraud, social media interactions, and all kinds of sensorial and demographic data. That same year, Kaggle also introduced “Notebooks” (previously known as “Kernels”), a limited cloud computing for training machine-learning models, expanding the company services horizontally and vertically, centralizing every step of production. Kaggle also further gamified the activities in the platform by redesigning its award system, adopting a multidimensional incentive hierarchy to engage users in competitive and prosocial activities and encourage them to contribute content to the communities. More importantly, these changes shifted the focus from managing competitions toward deeper connectivity among datasets, code, and users. Combining the new features with the competitions held on the platform, Kaggle would have more control over the work done by the community, exploiting the full potential of a human-machine-data assemblage to (re)produce predictive algorithmic models that will eventually be deployed for commercial use.

In the previous chapter, I discussed how Kaggle evolved as a company and examined its digital public face (the homepage). This chapter reveals what is inside the platform, describing and discussing Kaggle as a gamified platform where a community of developers and data scientists compete against each other for “fun, glory, and money” while having their work exploited by third-party companies. More specifically, I highlight three critical elements that constitute Kaggle as a platform for crowdsourcing machine-learning development: its digital infrastructure, the community of "Kagglers"—as the members call themselves—and the competitions organized by Kaggle and third-party sponsors. The first section focuses on the digital infrastructure made available by the company to its users, more specifically, the repositories where they can store datasets and run machine-learning code in the cloud. While cloud computing is not Kaggle's main product, I describe how this service works on the platform and what features the company offers its users. I also describe how Kaggle's datasets repository works, listing the main contributors and the most popular datasets.

The second section offers an overview of Kaggle's data science community. Using Kaggle's surveys from 2017 to 2022, I show Kagglers' main demographic features (age, gender, location, occupation, salary range) and how some of these characteristics have changed and evolved. The section also considers the platform's multi-layered ranking system, where users are evaluated based on their sociotechnical engagement in the platforms, as well as the levels of productivity and effectiveness of the predictive models they produce in competitive events. I detail how the system of tiers, medals, and points works for different categories of expertise and show the current state of each ranking, with particular attention to the users at the top.

The last section discusses the competitions organized by Kaggle, describing and detailing the different types of events, how they are managed, their limitations, and their prizes. I discuss these competitions in the context of their strategies to lure competitors into participating in the "game" and share their work with the promise of awards, money prizes, points, virtual medals, and made-up honorific titles. In so doing, I am able to show how much money Kaggle has distributed to participants throughout the years and who the most prominent sponsors of these challenges are. The section also considers controversial decisions taken by Kaggle to organize these competitions, such as the winner-take-all approach and the fact that they are narrow in scope and skewed in relation to the general practices of a data scientist's job. Lastly, I classify the competitions into different modalities according to their objectives using qualitative content analyses. These modalities serve as the basis for a deeper examination of the modes of subjectivation using AI to shape users' behaviour.

5.1. Repositories: Code and Data for Machine Learning

5.1.1. Notebooks: Code Repository

In computer science and code development circles, Notebook is an interactive development environment to create, share, and present code and data. Its flexible interface enables users to configure and arrange code/data workflows for specific data science, scientific computing, and machine-learning tasks. It can be distributed and used over the Web, and its modular design invites extensions to expand and enrich functionality. On Kaggle, Notebooks are accessible throughout the platforms: they are present on the user news feed, within each competition or dataset, and on their own dedicated page (see Figure 5.1, left). Browsing Notebooks in the Datasets or the Competitions sections provides a way of becoming acquainted with the specificities of the data and learning new techniques from other users. On its dedicated page, Notebooks are, by default, listed by “Hotness”—that is, by popularity. Kaggle encourages new users to visit other users’ Notebooks as a way to learn how people do data science. It is also possible to fork (clone) any existing public Notebook to copy the code and start experimenting without changing the original code. Kaggle’s Notebook infrastructure supports three development environments: scripts, RMarkdown, and Jupyter Notebook. While scripts are plain text files that can be executed as code sequentially, Jupyter notebooks consist of a sequence of cells formatted in either Markdown (comments, documentation, and partial outputs) or a programming language of one’s choice, facilitating the readability and reproducibility of the algorithm. Users can write notebooks in R or Python. R coders and people submitting code for competitions often use scripts; Python coders and people doing exploratory data analysis tend to prefer Jupyter Notebook.

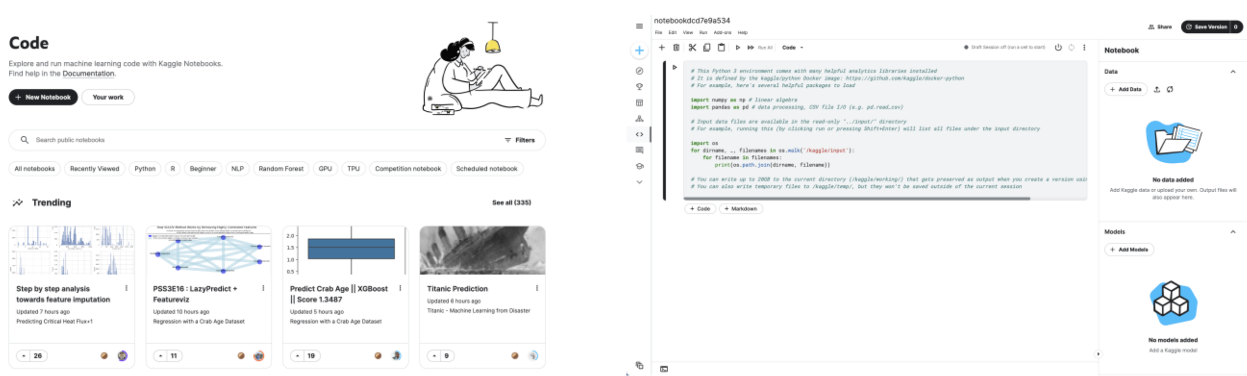


Figure 5.1: Kaggle Notebooks. Page listing Notebooks (left) and an example of a Notebook (right). Screenshot captured in June 2023 by L. Frizzera.

On Kaggle, Notebooks are more than simple repositories through which users can publicly share versions of their code, like GitHub. Notebooks double as a complete Integrated Development Environment (IDE) where users can run and test their code without investing in new high-end and overly expensive computers (see Figure 5.1, right). For specific competitions, Kaggle offers on-demand limited free access to cloud computing for participants. These machines are equipped with Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and other Google Cloud Platform services to support and accelerate machine-learning training. These "virtual machines" run statistical analyses and produce predictive models written in R (computer language typically used in statistics) or Python.

With Notebooks, Kaggle offers an infrastructure that works as an "engine" to power machine-learning development on an industrial scale. The company argues that free access to cloud-based notebooks makes for more balanced competition since users have access to similar machine power. On the other hand, it can create a technical dependency on commercial platforms with access to all the work being processed within their infrastructure. While the company claims that it simply fosters collaboration among developers by allowing them to work and learn together, sharing code and data via the platform is, however, paramount to producing the network effect among users, giving the impression that the platform is run by users rather than by operators (Van Dijck et al., 2018).

5.1.2. Dataset: Data Repository

If Notebooks are the engine, Datasets are their fuel. In the first five years, datasets were the "source material" of Kaggle competitions and only existed to give purpose to these challenges. Starting in 2016, Kaggle allowed its users to store and share structured and unstructured data on the platform. The dataset repository is diverse, spanning all kinds of functionalities, interests, and research fields. Some examples submitted to Kaggle are "The Complete Pokemon Dataset," consisting of more than 800 Pokemon from all seven Generations (Banik, 2017), "The European Soccer Database," listing more than 25,000 matches, 10,000 players from 11 European countries between the 2008 and 2016 seasons (Mathien, 2016); "Chest X-Ray Images (Pneumonia)," with 5,863 X-Ray Images from patients of one to five years old from Guangzhou Women and Children's Medical Center (Mooney, 2018); and the "USA Name Data," containing 5.55 million names from Social Security card applications for births that occurred in the United States after 1879 (Data.gov, 2019).

Users can submit datasets in various formats: Comma-Separated Values (CSV) for tabular data; JavaScript Object Notation (JSON) for tree-like data; SQLite for small databases, ZIP and 7z archives (often used to compress imagery); and BigQuery, which are massive multi-terabyte SQL databases hosted on Google's servers. These datasets can be uploaded to the website, pulled from other repositories, or generated within Kaggle using Notebook outputs. Though Kaggle caps the maximum size of datasets at 100GB, it also offers extra paid storage via Google Cloud services. Dataset owners can keep their datasets private, invite people to collaborate, or make them public.

The Datasets section of the platform is more than just a simple, stale data repository. Sharing datasets is considered core functionality. In this regard, it has more in common with other social media platforms, where the users update datasets (including personal and private data) to the platform without anything in exchange. The data stored in Kaggle can be accessed, explored, combined, aggregated, acted upon, and even used as external resources in machine-learning competitions. Similar to Notebooks, Datasets are listed in the user newsfeed and on their own dedicated page. By default, datasets are sorted by a metric named "Hotness," which measures the popularity of datasets on the platform (a combination of freshness, engagement, level of activity, and the number of upvotes) (see Figure 5.2, left). Communities are built around each dataset where people can collaborate, discuss, and share machine-learning codes as well as techniques to work on specific data properties. Moreover, beyond the data itself, each dataset may contain several metadata, including the collaborator's name, a short description, license type, tags, expectation of update, and a usability score. Each dataset has a dedicated discussion forum and a notebook area for exploratory data analysis. The page also shows basic statistics usage, such as the number of

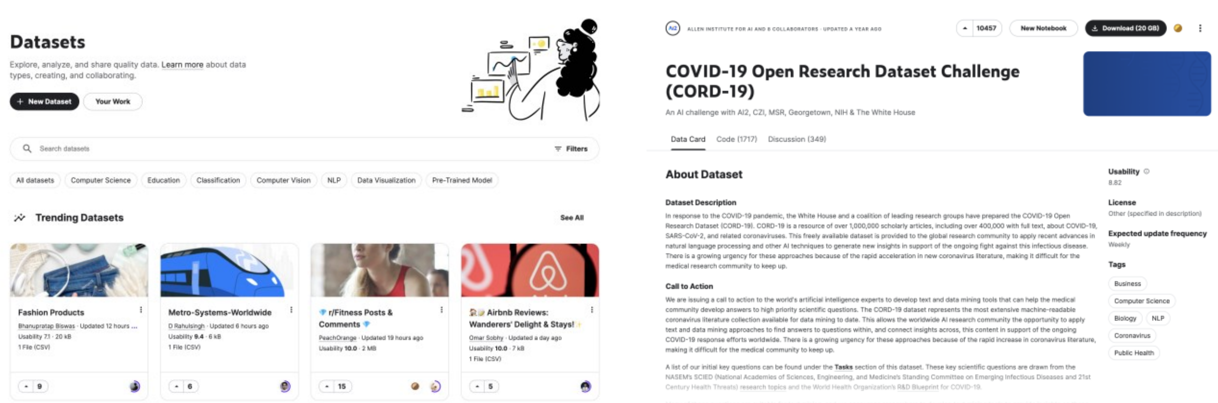


Figure 5.2: Kaggle Datasets. Page listing Datasets (left) and an example of a dataset page (right). Screenshot captured in June 2023 by L. Frizzera.

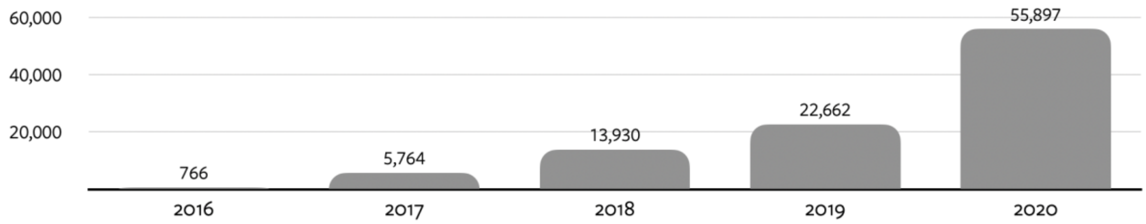


Figure 5.3: Datasets on Kaggle. Number of datasets stored on Kaggle per year based on the upload date.

views and downloads, and the level of engagement in the platform, such as badges and the number of upvotes (see Figure 5.2, right).

The number of datasets hosted on Kaggle has grown steadily over the years. Martin Heller (2020) reported 35,000 datasets in June 2020. A few months later, when I collected data for this research in November 2020, there were 55,897 datasets (see Figure 5.3). Almost three years later, in June 2023, there were over 227,000 datasets available on in the platform. The richness of the user contribution can be demonstrated by the more than 50 terabytes of data uploaded to the platforms until 2021, making Kaggle rapidly become the largest repository for machine learning and data science. These datasets range from a single spreadsheet file with just a few kilobytes to massive databases with several terabytes. Notable examples of large datasets often contain visual data, such as the 224,500 GAN Generated Human Faces Dataset (NiveditJain, 2021), 1 Million Fake Faces (Tunguz, 2019a), the 500,000 Aligned Anime Faces (0sb1, 2020), and the 53,000 images of retinopathy from both eyes (Zeeshan, 2020).

As in any other social media platform, a minority of users is responsible for providing most of the content (Lovink & Niederer, 2008). Up to 2021, approximately 26,000 users (0.5% of its 5 million users) uploaded at least one dataset to the platform. The top contributor is Mathurin Aché, a Data Scientist at Prevision.io, an AI company based in France, who uploaded 865 datasets. Johar M. Ashfaque, a Data Scientist, Researcher and Writer at the East Kent Hospitals University NHS Foundation Trust (UK), comes in second, with 278 datasets (see Figure 5.4, left). Public institutions, international organizations, and private companies also shared multiple datasets on the platform. For instance, the State of New York (226 datasets), the City of New York (199 datasets), the U.S. Federal Reserve (158 datasets), the World Bank (53 datasets), and FiveThirtyEight, a website focused on opinion poll analysis, politics, economics, and sports (83 datasets).

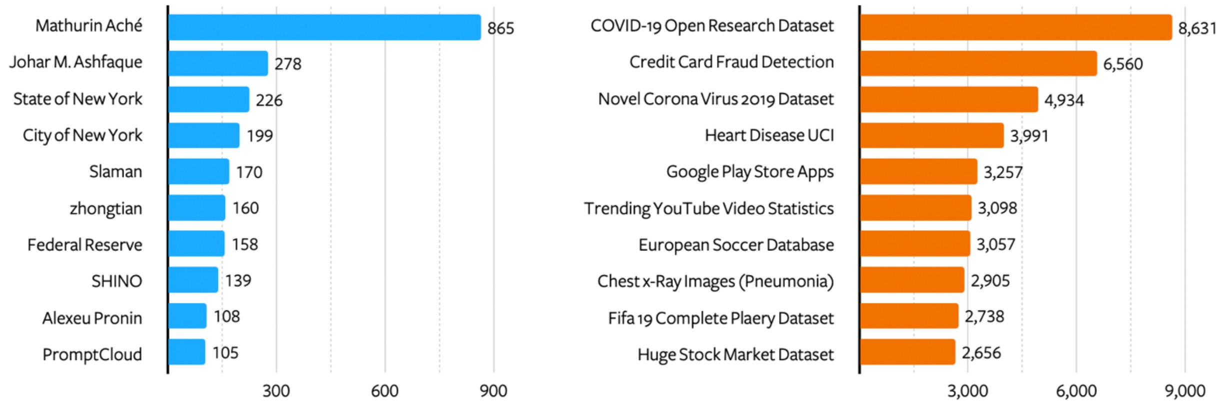


Figure 5.4: Top Datasets & Contributors on Kaggle. Top 10 dataset contributors (left); Top 10 upvoted datasets (right).

Datasets are more than just a collection of data points: they are an important commodity, the raw material of the twenty-first century (Srnicek, 2017; Zuboff, 2020). Users can upvote their favourite datasets, which, in turn, can be worth points and medals for their owners (see next section) and increase their popularity on the platform. Of the top upvoted datasets in the platform (see Figure 5.4, right), at least four are directly related to the health sector, including the first in the ranking (COVID-19 Open research Dataset), listing 400,000 scholarly articles about coronaviruses. The financial sector is directly present with at least two datasets focused on stock markets and Credit Card Fraud Detection. The digital sector also attracted a significant number of upvotes with special interests in trending apps on app stores and videos on YouTube. The interest in sports, a field that historically has intensively used statistics to enhance one's odds of winning, is represented here with the European Soccer Database and FIFA 19 complete player dataset.

5.2. Kagglers: A Community of Data Scientists

Kaggle's most important asset is not the digital infrastructure with cloud computing and e-learning platform, nor the vast repository of datasets uploaded by thousands of users for machine-learning training. Rather, what is most valuable to the company, and to any social media platform generally, is the community of users around it that produce content, upload data, and create social bonds, keeping the platform active and alive. As discussed in the previous chapter, Kaggle's business model is based on crowdsourcing machine-learning solvable problems in order to produce predictive models and drive business decision-making. Each competition mobilizes thousands of users working on their own time and from their own laptops. Since only the top 3-5 teams may receive some cash award, most participants work for free in exchange for virtual points and medals. Max

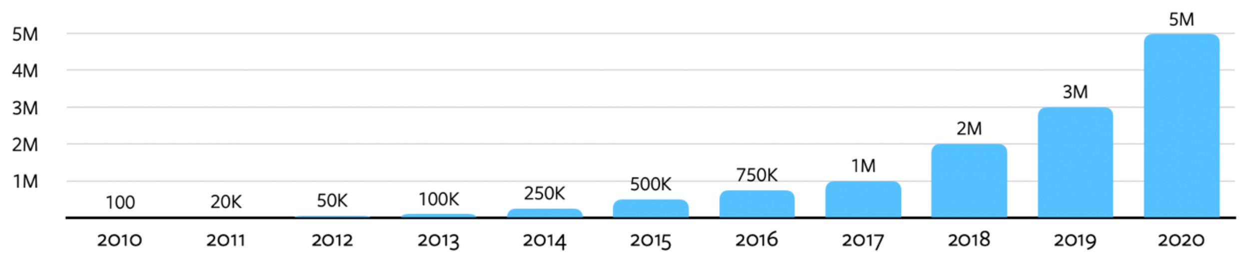


Figure 5.5: Number of Users on Kaggle. Approximated number (in millions) of registered user per year on Kaggle.

Levchin (2013), one of Kaggle's initial investors, described the platform's advantageous position to other venture capitalists as a place to "rent the brain of a data-mining genius via Kaggle by the hour, tomorrow by brain-hour" and soon "brain plug firmware will earn [developers] a little extra cash while [they] sleep, by being remotely programmed to solve hard problems" (Levchin, 2013, para. 21). The concept of crowdsource labour as an inexpensive infrastructure resource attracted private companies that saw the opportunity to outsource expensive and laborious tasks.

It was precisely a community of data scientists, a reserve of highly skilled workers that could be used for crowdsourcing development in machine learning and artificial intelligence, that was the main reason Google acquired Kaggle in 2017 (Lardinois et al., 2017). Nonetheless, aside from the interaction and engagement in the platform, neither Kaggle nor Google had a clear picture of the community in terms of demography. Since 2017, Kaggle (2017; 2018; 2019; 2020a; 2021a; 2021b; 2022a; 2022b) has run annual Machine Learning & Data Science Surveys to have a sense of the profile of the community in order to steer the platform in a specific direction. These surveys are self-reported and vary in the number of questions asked (50, on average) and the number of respondents, which fluctuate from 16,000 to 24,0000. Nonetheless, Kaggle (2019) believes it has obtained a representative sample of its community, which has been growing at a fast pace since its launch in 2010, reaching 5 million users in 2020 (see Figure 5.5). Moreover, since 2019, in a typical Kaggle crowdsourcing approach, each year's survey raw data was released as a prized challenge so the community could perform a statistical analysis and figure out new ways to tell rich stories about a subset of the data science and machine-learning community. I rely on these surveys and challenges to paint a portrait of the data science community's main characteristics, such as age, gender, location, education, occupation, wage, interests, and aspirations.

Kagglers are relatively young, with an average of 30 years old. More than half of the community is between the ages of 22 and 34 (Kaggle, 2022a). Though this average remained stable from 2017 to

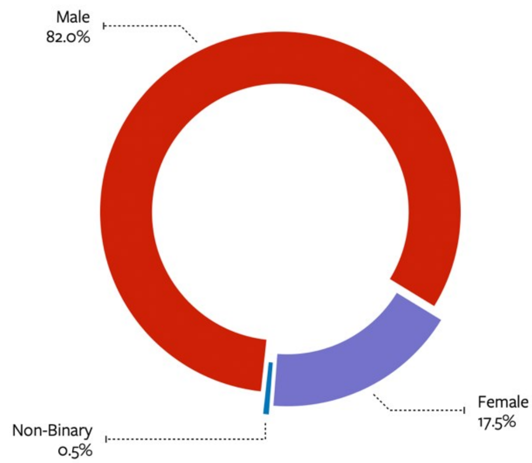


Figure 5.6: User Gender Distribution. Kaggle's (2022a) gender imbalance mirrors the dominance of men in the tech industry.

2022, the number of senior professionals is declining, and the number of young adults (primarily students) is increasing: only one in five professional data scientists is 40 years old or older, and at least 7% are aged 18 to 21. In terms of gender, Kaggle is unsurprisingly dominated by male users. Since 2017, the proportion of male to female users has remained stable, with only slight fluctuation: around 82% of the respondents identify themselves as male, while about 17% identify as female (see Figure 5.6). Parul Pandey (2019), a Kaggle's Notebook Grandmaster from India, explores this issue further with her series of EDA using Kaggle surveys titled "Geek Girls Rising: Myth or Reality!" that highlight the gender bias and lack of gender equity in the tech industry. The surveys indicate that the gender imbalance affecting the tech sector extends to data science and AI. While the numbers may vary slightly depending on the context (location, education, level, job opportunities), the challenge of getting and keeping women in tech and data science is now culturally and socially established, going well beyond skill level and educational exposure.

As the self-proclaimed world's largest data science community, it is unsurprising that the community is spread throughout 171 countries (Kaggle, 2022a). Since 2017, users from the U.S. and India have dominated the community. In 2017, one in four users was from the U.S. (25%) and 16% from India (see Figure 5.7, left), followed by Russia (3.5%), China (2.8%), Brazil (2.8%), Germany (2.8%), France (2.7%), and Canada (2.6%). Since 2020, the U.S. and India have switched places. In 2022, India had far more users in Kaggle's community than any other country (see Figure 5.7, right): one in three users (37%). On the other hand, U.S. representation declined, reaching 16% that

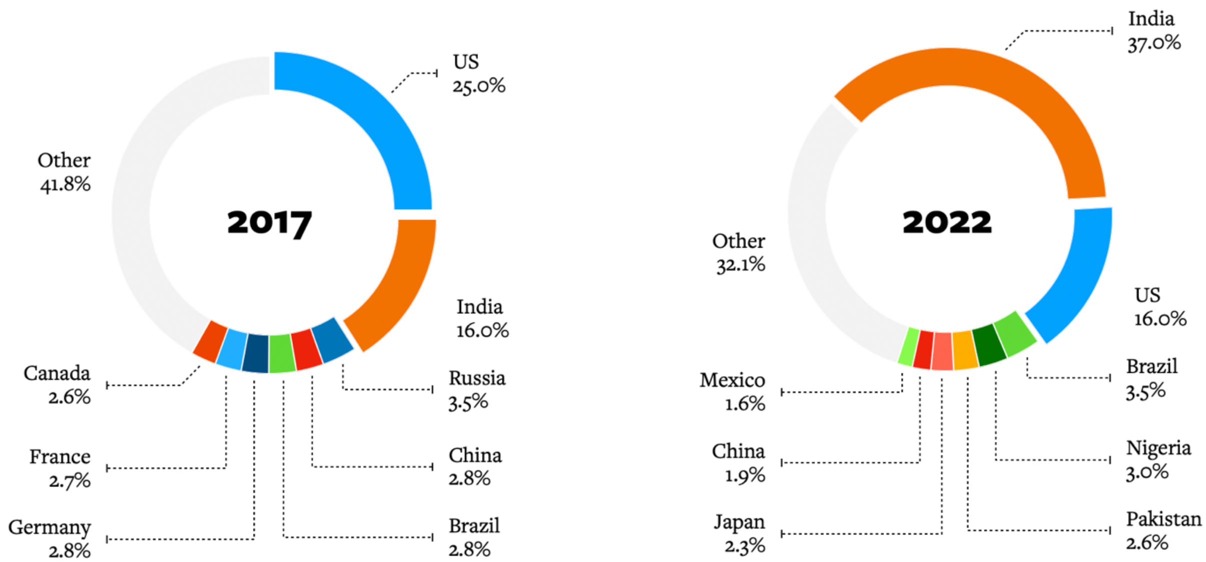


Figure 5.7: User Geographic Distribution. Distribution of users by country (Kaggle, 2017; 2022a).

same year, followed by Brazil (3.5%), Nigeria (3%), Pakistan (2.6%), Japan (2.3%), China (1.9%), and Mexico (1.6%). Michal Bogacz (2022), a Kaggle's Notebook Grandmaster from Poland, extensively analyzed demography, education, technology, and career factors across countries, revealing the similarities and contrasts of each location in relation to the machine-learning and AI industries. For instance, his Notebook "15 Factors for data science in your country!" shows granular data about the ratio among countries, such as the ratio of people under 30 years old (Bangladesh 78%; Spain 20%), the ratio of men (Japan 91%; Tunisia 53%), and people who earn over US$10,000 per year (Netherlands 57%; Iran 1%).

Producing well-rounded predictive models and machine-learning algorithms requires highly skilled workers with specific knowledge in statistics, computer science, mathematics, and programming languages (Kaggle, 2022a). As such, only 2.4% of the users have had no formal training or education past high school. The vast majority (83%) of the community has or is pursuing some university degree (see Figure 5.8, left): Bachelor's (32.6%), Master's (39.1%), and PhD (11.4%). At least 43% of degree holders have published academic research. Unsurprisingly, STEM disciplines are the most common undergraduate background (see Figure 5.8, right): Computer Science (40%), Engineering (15%), Mathematics or Statistics (12%), Economy or Finance (8%), Information Technology (5%),

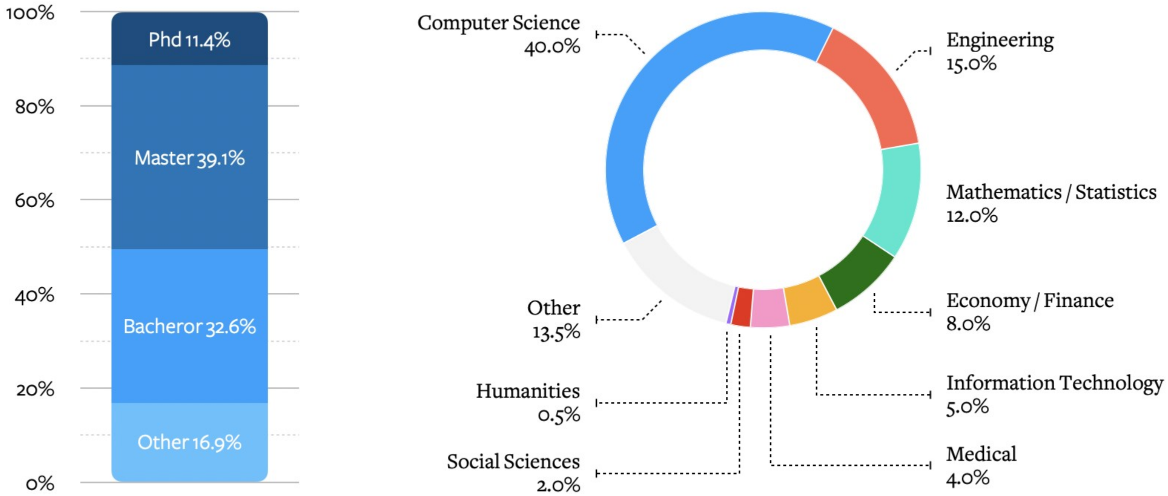


Figure 5.8: User Education Distribution. Distribution of users with an academic degree (right) (Kaggle, 2022a); most common undergraduate background (left) (Kaggle, 2018).

Medical (4%), Social Sciences (2%), Humanities (less than 1%). The number of students on the platform has increased over the years, jumping from 15% in 2017 to 27% in 2020 and 50% in 2022. This might be a consequence of different factors, such as the introduction of the learning portal with data science courses and machine-learning tutorials in 2017 and the popularization of machine learning and AI in education circles, including adopting Kaggle as a learning platform in schools. The user "xxxxyyyy80008" (2022), a Kaggle Notebook Expert, dove deeper into the intersection of age and educationin Kaggle between 2018 and 2022, showing how these factors are distributed in different locations; for instance, most Canadian users have a Master's degree while most of the users in India are undergraduates.

The most common job title on Kaggle (2022a) is Data Scientist (18.1%). Nevertheless, the surveys show that many other jobs support or require data science workflows and machine-learning techniques. Kagglers also identify their job titles as Data Analyst (14.5%), Software Engineer (9.2%), Teacher (7.8%), Manager (7.8%), Research Scientist (5.6%), and Statistician (1.2%). These users are employed in different sectors, most commonly in Technology (25.5%), Education (15.9%), Finance (8.8%), Manufacturing (5.6%), Medical (5.6%), Government (5.5%), and Online Service (5.1%). However, at least 13.5% responded that they were unemployed. Muhammad Nakhaee (2022), a Kaggle Dataset Expert, creatively illustrates these professions in his Notebook titled "The Fellowship of Kaggle," where he produced cartoonish-style data-driven Kaggle

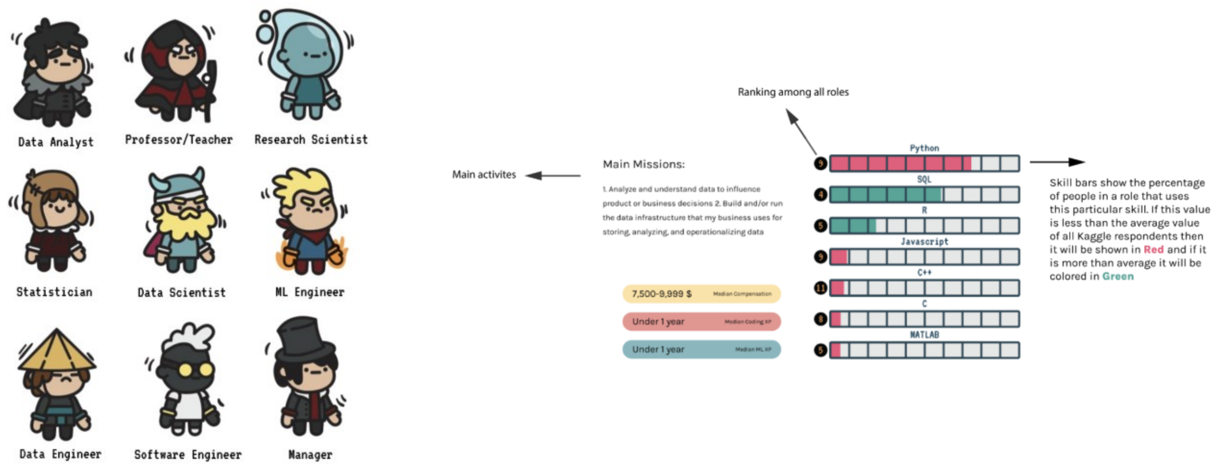


Figure 5.9: Illustration for a Kaggle RPG. Kaggle Champions representing the most common job titles on the platform (Nakhaee, 2022).

Champions, describing each profession as characters in a role-playing game (RPG) (see Figure 5.9). For instance, the Data Analyst profile is described as follows:

> Also known as the Rangers of the Data World, the data analyst[s] were at the forefront of the battle against the Data Monster. Using their *storytelling* and *visualization* skills and their basic knowledge of programming and mathematics [SIC]. Their strength lay in their ability to use **advanced SQL queries** to find treasures in raw data and magically transform them into useful insights for the stakeholder. However, [t]hey were lightly armored in programming and they were particularly vulnerable in cloud skills. [...] They roamed different realms of data in search of useful information on data problems. They shared their findings with the stakeholders (the manager) to help them make better decisions in the battle against the Data Monster. (n.p.)

In the 2022 survey, most Kagglers had some experience with code: 66% had one or more years of experience. However, they are relatively new to machine learning. Slightly over 53% of users have less than one year of experience, and less than 8% have been using machine learning for five years or more. Skill level, experience, and education impact the salary. Though data science is considered a highly skilled job, over 50% reported their yearly salary ranged from US$0 to US$25,000. These low wages might be explained by the fact that half of the 2022 survey respondents comprised students. However, distortion in the global market (where companies are located and where the developers live), as well as the different currency exchange rates, may also explain the imbalance in the machine-learning industry. For instance, as we can see in Paul Mooney's (2022) EDA, companies in the United States are most likely to pay in the six figures: 85% of US-based users reported salaries higher than US$100,000, while 85% of India-based users make less than

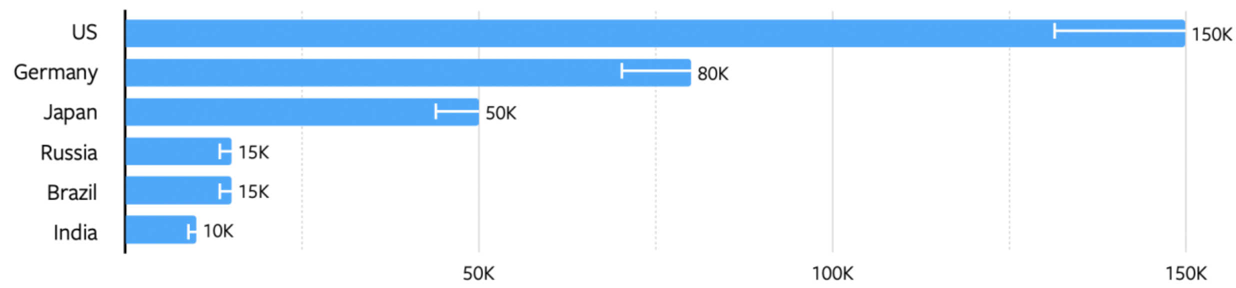


Figure 5.10: User Salary Range. Common data scientist salary range per country (Kaggle, 2022a).

US$40,000. If we take the most common salary range per country, we note that users in the Global North have higher salaries than Global South users (see Figure 5.10). For instance, a data scientist's salary in India varies between US$7,500 and US$10,000, while in the US, these professionals earn between US$125,000 and US$150,000 annually.

Lastly, since everything is done remotely on Kaggle, access to technology and computing power is crucial but is unevenly distributed among users. In the 2021 survey, while most users (85%) reported using their own personal computers when participating in Kaggle's activities, only 9% said they used cloud computing (AWS, Azure, Google Cloud, Kaggle Notebook) and 3% reported using a deep learning workstation (NVIDIA GTX, LambdaLabs). Moreover, more than 50% do not own or use hardware accelerators specifically to train machine-learning algorithms. Despite the high cost of this equipment, 33% said they would use NVIDIA GPUs, and 14% preferred to pay for Google Cloud TPUs. Once again, as shown by Bogacz (2022) and Mooney (2022), the distribution of computing power depends on the user's location and wealth, which is, unsurprisingly, skewed toward Global North countries. Access to a powerful computer and cloud computing may be crucial to getting ahead in a competition leaderboard and having a chance to win cash prizes.

5.2.1. User Rankings

Kaggle attentively observes and quantifies every action these users make on the platform. As discussed in the previous chapter, the levels of engagement become metrics in a gamified environment that pushes toward unlimited exploitation of the human body and cognitive abilities as a computation resource (Crary, 2014; Whitson, 2013). The company crowdsources most of the activities on the platform, heavily relying on its flourishing community to solve complicated problems using machine-learning techniques. As such, it is not enough to invite and encourage Kagglers to use the tools available on the website, upload data, share code, participate in challenges,

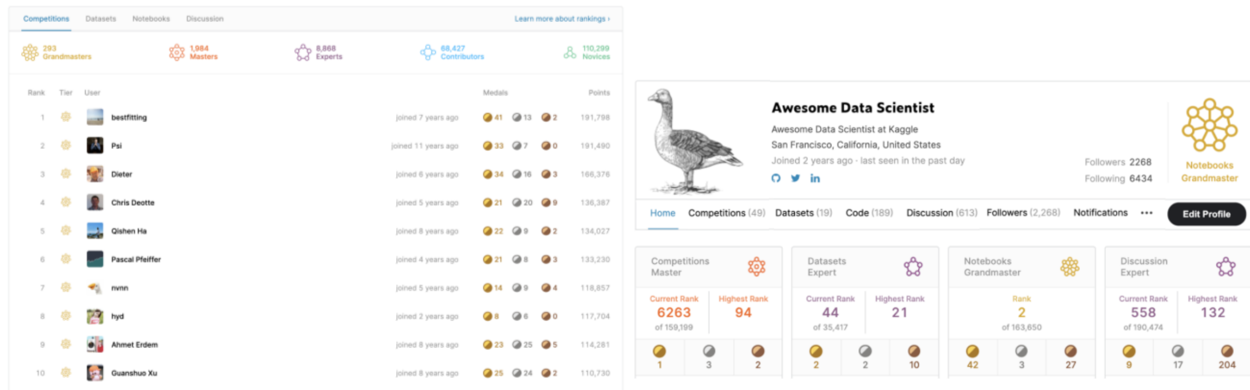


Figure 5.11: Kaggle User Rankings. Page listing Raking Rankings (left); the user profile page (right). Screenshot captured in June 2023 by L. Frizzera.

produce content, and interact with other users. Kaggle also ranks its users—from best to worst—by "looking at [how] effective they have been in predictive modelling, [and] how predictive their models are. So, we are the only source in the world of the ranked world's best data scientists" (Goldbloom & Howard in World Economic Forum, 2013, 0:40-1:04). This "ranking culture" (Rieder et at., 2018) contributes to the processes of hierarchization and modulation of visibility of users and content creators within the platform, but also to the outside market as a way of managing the workforce and creating a market reserve.

Kaggle's (2021c) Progression System resembles Massively Multiplayer Online Role-Playing Games (MMORPG) or Multiplayer Online Battle Areas (MOBA), in which players assume the roles of characters in a fictional setting, such as World of Warcraft, Fortnite, and Eve Online. Players take responsibility for acting out these roles within a narrative through a process of structured decision-making related to character development. Actions taken within these games follow a formal system of rules and guidelines, producing a sense of accomplishment in the users and the material substance for game leaderboards that can be used to rank and compare players' skills and abilities. Likewise, Kaggle maintains a Rankings page with a live leaderboard of "the absolute best data scientists on Kaggle" (see Figure 5.11, left). Goldbloom and Howard indicate that Kaggle has been ranking users since the first competition based on how well they perform in competitions (World Economic Forum, 2014). However, it was only in 2016 that the company designed a complex and well-organized system of tiers and classes whereby users progress by accumulating points and earning virtual medals.

According to Kaggle (2021c), the “Progression System uses performance tiers to track your growth as a data scientist … you’ll earn medals for your achievements and compete for data science glory on live leaderboards” (n.p.). The progression system is designed around four independent categories of data science expertise: Competitions, Notebooks, Datasets, and Discussion. There are five performance tiers within each of these categories that can be achieved following the quantity and quality of work produced by the user: Novice, Contributor, Expert, Master, and Grandmaster. The highest tier a user has achieved in any of the categories of expertise will be displayed on their profiles (see Figure 5.11, right) and under their avatar across the site. These titles are awarded based on medals earned in each category. Kaggle defines medals as a standardized way of recognizing and rewarding excellent work across the categories of expertise. Each medal is awarded for a single accomplishment: a great competition result, a popular notebook, a useful dataset, or an insightful comment. While tiers and medals are permanent representations of a data scientist’s achievements, points are designed to decay over time, keeping the rankings contemporary and competitive.

The first two tiers are designed as a tutorial and guide the user through the platform. New users are immediately placed at the Novice tier when they create an account on the platform. To become a Contributor, the user must complete four simple tasks: Run one notebook or script, make one competition submission, comment in a forum, and give one upvote to a dataset, code, or comment. To reach the other three tiers, the user must actively participate in the platform, making useful contributions, accumulating points, and earning medals following specific rules in each category, which I further explain below. An expert is a user who has completed a significant body of work on the platform in one or more categories of expertise. Once the user reaches the expert tier for a category, they will be entered into the site-wide Kaggle Ranking for that category. A master user has demonstrated excellence in one or more categories of expertise on Kaggle. Masters in the competitions category are eligible for exclusive Master-Only competitions. Lastly, “Grandmaster,” the most prestigious title, is given to a user who has demonstrated outstanding performance in one or more categories of expertise on Kaggle.

5.2.1.1. User Ranking: Competitions

Considered the most important ranking in the platform, the competition ranking reflects the ability of the user to solve problems and produce high-quality predictive models. Competition points are awarded based on how well a team did in a competition, the number of members on that team, and the number of teams in the competition. The number of medals awarded per competition varies depending on the size of the competition (see Table 5.1). Levelling up to higher tiers requires the user to collect several of these medals. To become an Expert, a user must collect two bronze medals. The Master tier requires one gold and two silver medals. The Grandmaster title is awarded when the user gets five gold medals and a solo gold medal[10] in competitions.

Table 5.1: Competition Medals

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
Bronze	Top 40%	Top 40%	Top 100	Top 10%
Silver	Top 20%	Top 20%	Top 50	Top 5%
Gold	Top 10%	Top 10	Top 10 + 0.2%	Top 10 + 0.2%

In November 2020, the public competitions ranking listed 150,394 users. There were 198 Grandmasters, 1,513 Masters, 6,162 Experts, 57,457 Contributors, and 85,064 Novices. All the top ten users in this ranking are Grandmasters (see Table 5.2) with years of experience in the data science field and familiar with how Kaggle competitions are organized. While the ranking is diverse in terms of where the users are located, it is unsurprisingly concentrated on users directly involved in developing machine-learning and AI technologies in two multinational companies: H20.ai and NVIDIA. The exception is Ian Pan, a Resident Physician at a University Hospital. It raises the question of whether these users donate their free time to participate in these competitions or if these companies are investing time and resources to sponsor top specialists in strategic data science forums.

Table 5.2: User Ranking: Competitions (November 2020)

#	User	Tier	Position / Company	Location
1	Guanshuo Xu	Grandmaster	Data Scientist at H2O.ai	US
2	bestfitting [Shubin Dai]	Grandmaster	Data Scientist and Engineering Manager	China
3	Psi [Phillip Singer]	Grandmaster	Senior Data Scientist at H2O.ai	Austria
4	Dieter [Christof Henkel]	Grandmaster	Deep Learning Data Scientist at Nvidia	Germany
5	Μαριος Μιχαηλιδης KazAnova [Marios Michailidis]	Grandmaster	Data Scientist at H2O.ai	Greece
6	Giba [Gilberto Titericz]	Grandmaster	Data Scientist at RAPIDS at Nvidia	Brazil
7	Dott [Dmitry Gordeev]	Grandmaster	Senior Data Scientist at H2O.ai	Austria

[10] Users can team up to participate in competitions. A user must win at least one gold medal without the help of others to become a Grandmaster.

Table 5.2: User Ranking: Competitions (November 2020)

8	Ahmet Erdem	Grandmaster	Data Scientist at RAPIDS at Nvidia	Netherlands
9	CPMP [Jean-François Puget]	Grandmaster	RAPIDS and deep learning at Nvidia	France
10	Ian Pan	Grandmaster	Resident Physician at a University Hospital	US

5.2.1.2. User Ranking: Datasets

Datasets are valuable assets to Kaggle and to the community. They contribute to an extensive repository of machine-learning training data that can be purposely used in any context. Dataset points are awarded based on popularity. Each upvote on a dataset is initially worth one point and decays daily soon after the vote is cast. Similarly, dataset medals are awarded to popular public datasets published on the site, as measured by the number of upvotes. Not all upvotes count toward medals: votes by novices are excluded. A dataset gets a bronze medal when it reaches five votes, a silver medal with 20 votes, and a gold medal with 50 votes. The user must collect three bronze medals to reach the Expert tier, one gold and four silver medals to get into the Master tier, and five gold and five silver medals to reach Grandmaster status.

By November 2020, 4,493 users had uploaded datasets to the platforms and were listed in the public datasets ranking: 9 Grandmasters, 20 Masters, 353 Experts, 2,529 Contributors, and 1,582 Novices. This ranking reflects the concentration of users around the development of AI and machine-learning technologies. Most of the top ten users work for big tech companies, once again dominated by H20.ai and NVIDIA, but there is also at least one academic in dentistry in the group (see Table 5.3). The location of these users reflects the overall distribution by country on Kaggle, with most of the high-quality contributions coming from India, the US, and Brazil. Some of these users figure in other rankings, indicating active participation across the main activities on Kaggle. For instance, Chris Deotte leads the Notebooks and Discussion leaderboard, Marília Prata is the fourth in Discussions, and Abhishek Thakur is the fourth in Notebooks.

Table 5.3: User Ranking: Datasets (November 2020)

#	User	Tier	Position / Company	Location
1	Chris Deotte	Grandmaster	Data Scientist & Researcher at Nvidia	US
2	Larxel [Andrew Maranhão]	Grandmaster	Volunteer at D4SG	Brazil
3	SRK [Sudalai Rajkumar]	Grandmaster	Data Scientist at H2O.ai	India
4	Chris Crawford	Grandmaster	Data Engineer at Team Rubicon	US
5	Ruchi Bhatia	Grandmaster	Data Science Global Ambassador at HP & NVIDIA	India

Table 5.3: User Ranking: Datasets (November 2020)

6	Vopani [Rohan Rao]	Master	Data Scientist at H2O.ai	India
7	Abhishek Thakur	Grandmaster	NLP at Hugging Face	Norway
8	Tensor Girl [Usha Rengaraju]	Grandmaster	Data Science Consultant at Infinite-Sum Modeling	India
9	Bojan Tunguz	Master	Data Scientist at Nvidia	US
10	Marília Prata	Master	Self-employed Doctor of Dental Surgery	Brazil

5.2.1.3. User Ranking: Notebooks

Notebooks are pieces of code shared in the platform. They are considered valuable contributions to the community because they demonstrate technical skills, describe datasets, introduce techniques, and can be freely repurposed in different contexts. Like datasets, notebook points are awarded based on popularity. Each upvote on a notebook is initially worth one point and decays daily soon after the vote is cast. Similarly, notebook medals are awarded to popular notebooks, as measured by the number of upvotes a notebook receives. Not all upvotes count toward medals: self-votes, votes by novices, and votes on old posts are excluded. A notebook gets a bronze medal when it receives five votes, a silver medal with 20 votes, and a gold medal with 50 votes. The user must collect five bronze medals to become an Expert, ten silver medals to reach the Master tier, and 15 gold medals to reach Grandmaster status.

By November 2020, 20,853 users had submitted or collaborated with code and were listed on the public notebooks ranking: 10 Grandmasters, 134 Masters, 1,538 Experts, 11,695 Contributors, and 7,450 Novices. The top ten users on this leaderboard are Grandmasters (see Table 5.4) who have been on the platform for at least two years. DanB and Abhishek Thakur have been registered on Kaggle since it was officially launched in the US. This ranking has more diversity in terms of where these users are located, though they are still highly concentrated in the Global North. Big tech companies are always present in this ranking, but it also shows users from other industries, such as management consulting (Deloitte) and pharmaceuticals (AstraZeneca). Note that one account, DATAI, may represent a team of developers from the Data company.

Table 5.4: User Ranking: Notebooks (November 2020)

#	User	Tier	Position / Company	Location
1	Chris Deotte	Grandmaster	Data Scientist & Researcher at Nvidia	US
2	DanB [Dan Becker]	Grandmaster	Founder at decision.ai	US
3	Andrew Lukyanenko	Grandmaster	DS at deep learning	Russia

Table 5.4: User Ranking: Notebooks (November 2020)

4	Abhishek Thakur	Grandmaster	NLP at Hugging Face	Norway
5	xhlulu	Grandmaster	Plotly / Deloitte	Canada
6	Alex Shonenkov	Grandmaster	Researcher at OCRV	Russia
7	DATAI [Datai Team]	Grandmaster	-	Turkey
8	Kostiantyn Isaienkov	Grandmaster	Data Science Engineer at Quantum	Ukraine
9	Parul Pandey	Grandmaster	Data Science Evangelist at H2O.ai	India
10	Rob Mulla	Grandmaster	Senior Data Scientist at AstraZeneca	US

5.2.1.4. User Ranking: Discussion

Considered the lower-hanging fruit, the discussion category is regarded as a pseudo-category in the community because it does not require technical skills to level up. Discussion points are calculated as the sum of total upvotes minus the sum of total downvotes cast on a data scientist's topics and comments on Kaggle. Decay is applied to both upvotes and downvotes based on the day the votes were cast to keep the ranking always fresh. Kaggle briefly mentions additional hidden rules to prevent upvote rings and progression system manipulation but does not further explain how these rules apply. Discussion Medals are awarded to popular topics and comments posted across the site, as measured by net votes (upvotes minus downvotes). Not all upvotes count toward medals: votes by novices and on old posts are excluded. The user's comment receives a bronze medal when it gets one vote, a silver medal with five votes, and a gold medal with ten votes. Like other categories, levelling up to higher tiers requires the user to collect several of these medals. To become an Expert, a user must collect 50 bronze medals. The Master tier requires five silver medals and 200 medals in total. The Grandmaster title is awarded when the user gets 50 gold medals and 500 medals in total.

In November 2020, the public discussion ranking listed 18,953 users: 22 Grandmasters, 64 Masters, 1,323 Experts, 12,283 Contributors, and 5,261 Novices. The top ten users in this ranking show diversity in skill level and location (see Table 5.5). These users have been on the platform for at least two years, with a few registered for over seven years. Most of these users work for a private company in the AI industry; Nvidia, a leader in the sector, employs four. The Discussion Ranking seems to cross all the others because the type of activity involved (asking and answering questions in the forum) is common when dealing with Datasets, participating in competitions, and producing notebooks, as six of the top ten users also figured in the other rankings. For instance, Chris Deotte is the leader in Dataset and Notebooks, CPMP is the ninth in Competitions, Marília Prata is the tenth in Datasets, and Abhishek Thakur is the fourth in Notebook.

Table 5.5: User Ranking: **Discussion** (November 2020)

#	User	Tier	Position / Company	Location
1	Chris Deotte	Grandmaster	Data Scientist & Researcher at Nvidia	US
2	CPMP [Jean-Francois Puget]	Grandmaster	RAPIDS and deep learning at Nvidia	France
3	““ [Cher Keng Heng]	Grandmaster	Deep Learning/Computer Vision at HP & NVIDIA	Singapore
4	Marília Prata	Expert	Self-employed Doctor of Dental Surgery	Brazil
5	Psi [Phillip Singer]	Grandmaster	Senior Data Scientist at H2O.ai	Austria
6	Yirun Zhang	Expert	PhD in Telecommunications at King’s College	UK
7	Konstantin Yakovlev	Master	Freelancer	Portugal
8	Abhishek Thakur	Grandmaster	NLP at Hugging Face	Norway
9	Dieter [Christof Henkel]	Grandmaster	Deep Learning Data Scientist at Nvidia	Germany
10	Matthew Masters	Master	Student at Purdue University	US

5.3. Competitions: Gamified Predictive Model Development

Kaggle depends on its online community to maintain its status as a vibrant social-professional space and a viable commercial enterprise. As noted earlier, the community actively participates in the platform by uploading datasets, producing code, sharing comments, and participating in machine-learning competitions. These competitions are the heart and soul of Kaggle’s enterprise, as it mainly functions as a gamified platform for developing machine-learning algorithms and predictive models. Kagglers are encouraged to share their work for several reasons: to learn, to explore datasets, to improve algorithm performance, to teach, to show off their skills. Mostly, however, they are lured into give away their work and data for free as a chance to win the prizes offered on the many sponsored competitions hosted on the platform. These competitions often offer money prizes for the first teams to solve the proposed problem using machine-learning techniques, most commonly by making incremental improvements on efficiencies and accuracies.

In 11 years, Kaggle has hosted 442 open competitions (see Figure 5.12, top) and more than 3,000 education-oriented (InClass) challenges. InClass competition is a self-service learning environment through which an instructor sets up a safe space for students to acquire and practice machine-learning skills. This type of competition is an interesting investigation topic in itself because it shapes the practices and assumptions a future data scientist will carry in their careers. However,

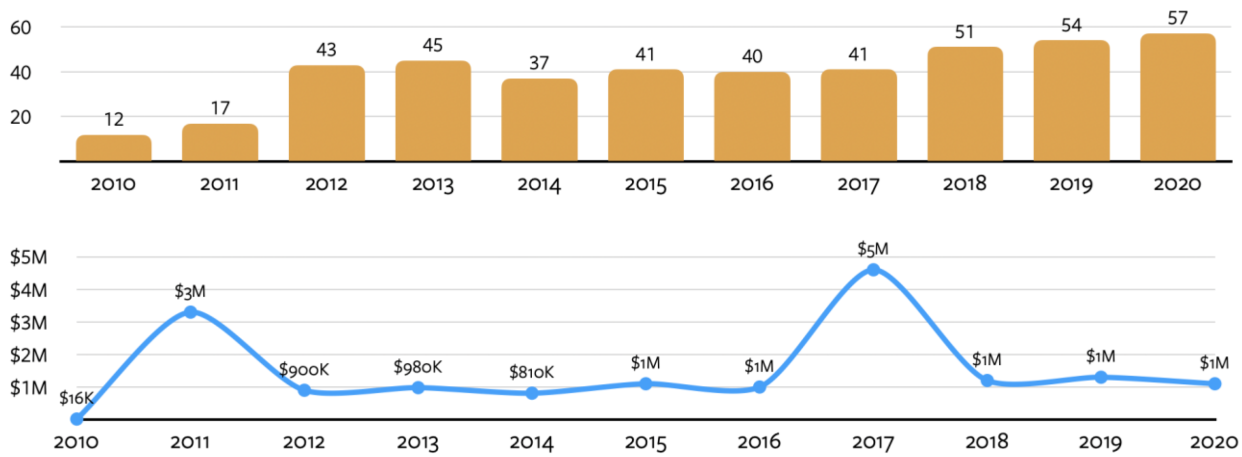


Figure 5.12: Kaggle Competitions. Number of open competitions hosted on Kaggle per year (top). Accumulated prize distributed per year (bottom).

InClass competitions are beyond the scope of my research, which focuses on public crowdsource events of machine-learning development and how predictive models are used to mediate and modulate user behaviours and habits. Therefore, this section focuses on open competitions, particularly the ones offering cash prizes, in which sponsors poured over US$15 million over the last ten years (see Figure 5.12, bottom).

Kaggle has held, on average, 42 open competitions per year. They range from a few weeks to several months and up to two years, such as the 2011–13 Heritage Health Prize (2013). More recently, after observing the dynamics around competitions (the number of competitors, the levels of engagement, the speed of improvements on the code, etc.), the company recommends running a competition for 12 to 18 months (Kaggle, 2020c). While only registered users can join these competitions, they are asked to provide a phone number and formally accept the rules of each competition they enter. Technically, this process is designed to ensure the user is a human being—not a bot—and guarantee they have read and accepted the rules. However, by providing their phone number, users disclose their location, which could prevent them from participating in some or all events because Kaggle follows the U.S. policy of blocking access from "enemy countries," such as Cuba, Iran, and Syria.[11] In addition, sponsors can make their own geopolitical restrictions; for instance, limiting the prizes to American citizens (DHS, 2017) or blocking participation from a specific country such as China (Chowdhury, 2017).

[11] For more on the Kaggle Terms of Service, see section 7 (Kaggle, 2020e).

These competitions attract thousands, sometimes hundreds of thousands of participants worldwide. These participants organize themselves in teams, typically between one to five per team, varying according to the competition's rules. These teams are ephemeral assemblages, only formed for the purposes of a specific competition. Examples of popular competitions include Home Credit Default Risk (131,935 competitors in 7,190 teams), IEEE-CIS Fraud Detection (125,890 competitors in 6,381 teams), Santander Customer Transaction Prediction (104,495 competitors in 8,802 teams), Titanic disaster (104,142 competitors in 21,739 teams), and SIIM-ISIC Melanoma Classification (102,544 competitors in 3,314 teams). However, Kaggle competitions are really among and against the users themselves. As discussed in the last section, individual users accumulate points, medals, and badges based on their productivity on the platform. Hardcore Kagglers participate in several competitions simultaneously to prove themselves to be the best among their peers.

These competitions offer prizes and awards. Some offer job interviews; others offer "swag" or "prizes" without other specifications. There are also the so-called "invaluable" awards, such as some form of social recognition materialized as "kudos" or simply "knowledge." However, as a competitive crowdsourced platform, Kaggle encourages competition organizers to offer money, recently set to a minimum of US$25,000 (Kaggle, 2020b), but reaching as high as US$3 million. There have been 311 competitions offering cash prizes, at least five over US$1 million (see Figure 5.13, left). For instance, in 2011, the Heritage Provider Network launched a US$3 million multi-year competition to identify patients admitted to a hospital during the following year using historical claims data. In 2017, the U.S. Department of Homeland Security offered US$1.5 million to improve the accuracy of its "security threat recognition" algorithms using airport screening images, which I further examine in chapter seven. Zillow, an American company that runs an online real estate marketplace platform, offered US$1.2 million to improve its algorithm for home value prediction. That same year, Booz Allen Hamilton allocated US$1 million to improve lung cancer detection in its Data Science Bowl 2017 competition. In 2019, Facebook launched a US$1 million competition to improve deepfake detection on videos with facial or voice manipulation (discussed in the next chapter). While the top-paid competitions are concerned with very different interests (security, real estate, data manipulation, and healthcare), a glance at the competitions that offer between US$25,000 and US$1 million (163 competitions) reveals a more commercial and for-profit focus, such as predicting the stock market and sports events; determining customers behaviour, habits, and preferences; recommending and persuading users to buy product and services; identifying financial fraud; and predicting risks of customer default. Moreover, it is no surprise that

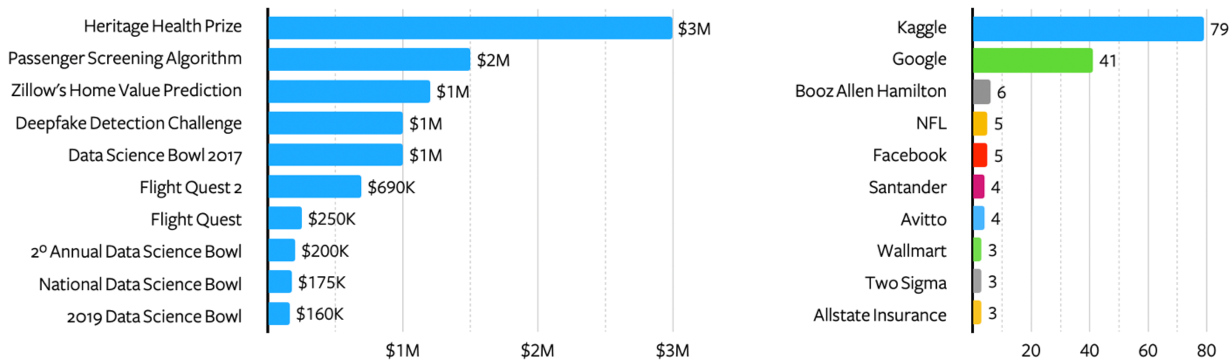


Figure 5.13: Kaggle Top Competitions. Top competitions by money prizes in USD$ (left). Top sponsors by number of competitions (right).

competitions not directly connected to commercial goals have lower money prizes. For example, improving our understanding of complex natural phenomena ($3,000 to detect dark matter in the universe; US$13,000 to improve Higgs Boson particle detection), helping wildlife conservation (US$10,000 to recognize and locate right whales in a series of satellite imagery), solving health problems (US$10,000 to predict Parkinson's disease progression), and promoting social diversity and equity (US$10,000 to detect insults on social media; US$25,000 for gendered pronoun resolution).

Who and what interests are behind these competitions are sometimes hidden from the public and even from competitors. Not every competition discloses its organizers (37%), in which case Kaggle would refer to the "organizer," "sponsor," or "promoter." On the other hand, 112 known organizations hosted 279 competitions (see Figure 5.13, right). Not surprisingly, Kaggle is the top organizer with 79 events. Google, Kaggle's owner, is next with more than 41 events organized through its subsidiaries. Booz Allen Hamilton, a U.S.-based information technology consulting company, has sponsored six competitions. The National Football League (NFL) and Facebook hosted five competitions each. Banco Santander and the Russian classified advertisements website Avitto had four competitions each. Other well-known companies, such as Walmart, Two Sigma, Microsoft, Lyft, and Allstate Insurance, have sponsored at least one competition on the platform. What drives these companies to invest in the platform goes beyond the straightforward financial gain of harnessing free labour from crowdsourcing tasks to build solutions to optimize commercial operations. As I will show in the next chapters, the value lies in the models created by algorithms applied to specific datasets to detect, manipulate, predict, and modulate objects, subjects, and events in the various systems at play, including how individuals should respond and act to all kinds of events.

There are six categories of competition on Kaggle: Getting Started, Playground, Featured, Recruitment, Research, and special limited competitions. With several tutorials to guide users in solving the problem, Getting Started and Playgrounds are semi-permanent competitions targeted to helping new users familiarize themselves with data science and machine-learning techniques. These competitions allow new users to see how their scores stack up against a cohort of competitors rather than the whole community. While Getting Started does not offer any prizes, on Playgrounds awards range from "kudos" to small cash prizes. The best example of a Getting Started competition is the "Titanic – Machine Learning from Disaster," which has been running since 2017 and had more than 100,000 participants by November 2020. An example of a Playground competition is "Dogs versus Cats," which asks the competitors to create an algorithm to distinguish one from another.

Featured competitions are full-scale community-wide machine-learning challenges that pose complex real-world prediction problems, generally with a commercial purpose. The topics and purpose of these competitions are diverse, ranging from predicting the existence and kind of toxic comments on Wikipedia (Google Jigsaw, 2018) to the type of insurance policy customers purchase based on their shopping history (Allstate Insurance, 2011); from improving the accuracy of threat recognition algorithms based on Passenger Screening at airports (DHS, 2017) to identifying nerve structures in ultrasound images of the neck (Halyard Health, 2016). The prizes of these competitions are, on average, US$25,000: some lower than US$1,000, particularly in the first years, and some going beyond US$100,000, reaching as high as US$3 million depending on the complexity of the task and how much the sponsor is willing to reward the participants. On special occasions, companies also hold recruitment competitions in which participants append their résumés for consideration together with their algorithms. Companies like Walmart, Airbnb, and Instacart took advantage of this kind of competition to offer job interviews to top-ranked users.

Research competitions, on the other hand, are more experimental, serving as opportunities to work on problems integral to a specific domain or area that may not have a clean or easy solution. Due to their exploratory nature, these challenges usually do not offer cash prizes or points. For instance, the Large Scale Hierarchical Text Classification challenge asked participants to Classify Wikipedia documents into one of 325,056 categories (Class-Y ANR Project, 2014); the U.S. National Oceanic and Atmospheric Administration (NOAA) summoned users to identify endangered right whales in aerial photographs (NOAA, & MathWorks, 2016). A particular case of research-oriented competition is the special limited competitions, which are private events that limit visibility and submissions only to invited users, generally Masters and Grandmasters.

The format and rules of these competitions vary depending on the sponsors' goals. In a standard competition, users are given access to a complete dataset at the beginning of the competition. Typically, the data is split into three sets: (1) training data, (2) public test data to validate the models, also used to keep the competition's public leaderboard updated, and (3) private test data used to assess the models' accuracy at the end of the competition. Participating teams can either download the data and build models on their own computers or, as described earlier in this chapter, use Kaggle Notebooks to run their algorithms in the cloud. After generating the model, each team can submit up to five predictions per day. While most competitions on Kaggle follow this format, there are variations. Some have stages, where the results from one stage are used in subsequent ones. Others, called "code competitions," must be carried out exclusively within Kaggle Notebooks and usually have a stricter hardware requirement, making the competition more balanced since users would have similar hardware allowances.

5.3.1. Simulacra and winner-take-all

Far from being unanimous in the data science community, Kaggle is highly criticized for the way it runs its competitions. Data scientist and Kaggle user Kiri Nichol (2015) points out that a typical competition artificially cleans data beforehand, creating an artificial environment for data science research. She also suggests that competition goals are designed to focus only on the model's accuracy toward a specific problem defined by the sponsor, skewing the results toward a biased interpretation of the problem. This corroborates the finding buried in the 2018 Kaggle survey, in which the company asked: What metrics do you or your organization use to determine whether or not your models were successful? The majority of the users responded "accuracy" (35%), followed by "revenue" (23%); "unfair bias" was only considered by 12%. Kaggle's narrow approach to data science is defined as exploiting statistical models for accuracy and revenue first without considering the context of the problem, its ethical implications, sociotechnical limitations, and political-economical abuses. That is, the predictive model's accuracy follows revenue more than human rights or ethical value. Nichol (2015) points out that Kaggle's data science is a simulacrum sold as a "fun" activity to solve problems with a computer. In reality, data science has more to do with statistical analysis than with machine learning: understanding the data, evaluating the metrics, asking questions about the data, and creating experimental designs to define the question that drives these competitions are the “real” work of data scientists.

Another common criticism is that Kaggle reproduces a “winner-take-all” economy. A typical competition distributes money prizes to the first 3-5 teams in the final ranking, with the top teams receiving virtual badges and medals; everybody else gets nothing. One of the best examples of this exploitative model goes back to the platform’s early years. In 2011, a consortium led by NASA, the European Space Agency, and the Royal Astronomical Society launched a competition to solve a problem on which physicists have been working for decades: detecting dark matter in the universe. The consortium asked competitors to use a phenomenon called gravitational lensing. This process stems from Einstein’s theory of gravity, which predicts that as light emanates from a galaxy and passes through dark matter, the dark matter distorts the light’s path. The distortion depends on the amount of dark matter present along the way between Earth and the galaxy. Therefore, by identifying the distortions of light coming from other galaxies, one can generate a three-dimensional dark matter map of the universe (AstroTom, 2011). To solve this highly complex problem, the consortium offered a very modest prize: a trip (up to US$ 1,000, later increased to US$ 3,000) to the Jet Propulsion Laboratory (JPL) in Pasadena, California, to attend yet another competition: the GREAT10 challenge workshop “Image Analysis for Cosmology.”

In the first few days of the competition, Martin O’Leary, a PhD student in Glaciology at Cambridge University, submitted his first entry. His statistical model performed better than NASA’s approaches, placing him first on the leaderboard. He applied techniques specific to his discipline to detect edges in glacier fronts from satellite images in order to solve this similar but unrelated problem. The data science community quickly glorified O’Leary’s ingenuity. The White House Office of Science and Technology Policy published a blog post praising the researcher: “In less than a week, [O’Leary] had crafted an algorithm that outperformed the state-of-the-art algorithms most commonly used in astronomy for mapping dark matter” (Rhodes, 2011, para. 2), placing O’Leary’s contribution alongside Newton’s and Einstein’s.

However, the permeability of the message exchanges in the forum shows that competitors learn from each other in the website’s forum, converging in a specific way to extract the maximum accuracy from the dataset. In a succession of submissions in the second week of the competition, two other teams—Ali Haissaine & Eu Jin Loc (Signature Verification, Qatar U & Grad Student @ Deloitte) and Marius Cobzarenco (Graduate Student in Computer Vision, UC London)—made

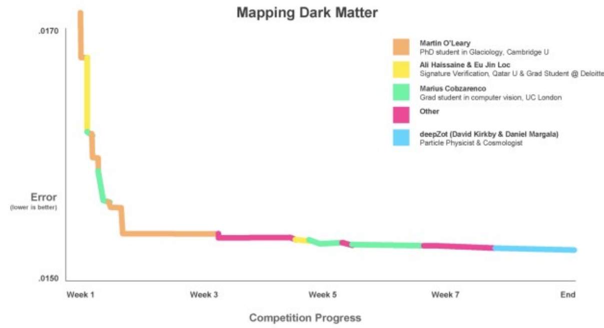


Figure 5.14: Progress in the Mapping Dark Matter Competition. Progression of improvements in the Mapping Dark Matter competition on Kaggle (Goldbloom, 2016).

substantial improvements to O'Leary model and switched positions in the public ranking (see Figure 5.14). By the third week, the competition had stalled, with every new improvement adding complexity without adding much performance. At the end of the competition, the team named "deepZot," formed by the Particle Physicist David Kirkby and the Cosmologist Daniel Margal, jumped ahead and was declared the winner. Martin O'Leary finished fourth. NASA paid travel expenses for the first two in the ranking but invited the top ten, including O'Leary, to attend the GREAT10 challenge workshop in Pasadena at their own cost. Furthermore, NASA asked the top five in the ranking to hand in their code, descriptions, and explanations of the model, in keeping with competition rules. Once compared to Einstein, O'Leary was crucial to the competition, making an essential contribution to Cosmology. However, the logic of Kaggle's competition and business model made his work disposable. O'Leary worked for free while others quickly seized his work to win the prize. From a competition standpoint, he had lost.

5.3.2. Machine-Learning Modalities

A quantitative analysis approach to studying Kaggle, its community, and the predictive models produced using machine learning is insufficient to uncover the meaning and the modalities of value explored by the platform and its sponsors. Meaning is often complex, holistic, context-dependent, not necessarily apparent at first sight, and does not always equate the coding frequency of a given theme with its importance (Kracauer, 1952). A qualitative content analysis approach, combining concept-driven and data-driven categories, would allow for more in-depth insights into how these competitions are used to produce value (Schreier & Flick, 2014). Using the data scraped from the Kaggle—more specifically, the title and description of public competitions that offered at least US$1,000 in cash prizes (284 in total)—I defined a simple coding frame to categorize these challenges in four different ways of data mobilization: identify, predict, recommend, and generate. Below is a short definition of each category:

Identify

Use machine learning to automatically detect, identify and classify bodies, things, and environments. This is common practice in the machine-learning community as a way of getting started with a dataset, writing exploratory statistical analysis, and defining what can be known from the data. It is typically done through clustering patterns to extract essential information from large datasets, which, in turn, are used to produce a model that can later be applied to detect and recognize similar objects/subjects. These models are used for various objectives, from labelling maps and terrain features to classifying toxicity levels in conversation, locating planets in far galaxies, detecting the presence of cancer cells in the human body, identifying birds by their song, recognizing individuals in a crowd, and deciding if an image of a human face is real or synthetically created (fake). It can also be used to create hyper-profiles from personal and impersonal data by identifying users', customers', and citizens' behaviours, habits, and preferences, typically through historical data across various sources, like social media, smartphones, Internet traces, CCTV, location, demographics, financial, health, purchase transactions, and so on. For categorization purposes, it is irrelevant what the algorithms are trying to identify (individuals, animals, objects, starts, molecules) or the sponsor's motivation in proposing the challenge (commercial, scientific, entertainment, educational). However, if the competition overview implies that the algorithm can be used to create predictions or recommendation systems, the categories "predict" or "recommend" should be used instead.

Predict

Use machine learning to forecast future events, trends, and behaviours. It is the core and, in many senses, the benchmark of machine-learning models since it determines the efficiency and accuracy of an algorithm (i.e., how well the algorithm fits a task). It typically uses substantial amounts of historical data from related and unrelated sources. However, it might rely on statistical techniques to use less data to reach a high confidence level. Predictions are valuable in all kinds of domains, such as the health sciences (map occurrence of epidemics), sports (ranking teams with better chances of winning a tournament), financial and marketing sectors (stock prices, product sales), policing and national security (predicting crimes and acts of terrorism), marketing (chances of a prospect accept an offer), and politics (probability of winning elections). The most common goal is to get ahead of events, know possible outcomes, and use the model to aid business decision-making. For categorization purposes, what the algorithms try to predict (behaviour, sports results, sales, threats, the presence of objects) or the sponsor's motivation in proposing the challenge (commercial, scientific, entertainment, educational) is irrelevant. However, if the competition overview implies that the algorithm can be used to create recommendation systems, the category "recommend" should be used instead.

Recommend

Use machine learning to produce limited hierarchical options to support decision-making. It is a higher-level modality of value since it involves both modalities above (identify and predict). The models produced by these machine-learning algorithms must consider the context—user, time, location, content, history, log, preferences, limitations, etc.—to suggest or recommend actions for accomplishing specific goals and even nudge individuals to act or to think in a specific way. The objective of recommendations can be to optimize inventory restock, enhance customer experience, retain user attention, create opportunities for encounters, satisfy a need, and so on. In particular, this modality of machine-learning algorithms is used on large-scale systems, such as search engines, social media networks, car navigation systems, entertainment streaming (video, music, games), app stores, and online retail. For categorization purposes, what the algorithms are trying to recommend (products, content, policy, actions) or the sponsor's motivation to propose the challenge (commercial, scientific, entertainment, educational) is irrelevant.

Generate

Use machine learning to generate new data, create art, or produce new forms of knowledge. Generate is an emerging modality based on Generative Adversarial Network (GAN), also known as Generative AI, where two neural networks compete to overcome limited training. Despite their incipient and sparse use on Kaggle, these generative models are deemed valuable in producing unique objects done by creative content creators on labour-intensive tasks, such as videogame levels and 3D characters (e.g., Epic Unreal engine and MetaHuman), hyper-realistic images capable of deceiving both humans and machines (e.g., MidJourney, Stable Diffusion), and the ability to produce textual material in response to a user prompt via chat bots (e.g., ChatGPT, Gemini). In particular, this modality has been used to create deepfakes and simulacra, such as images of people that have never existed and video recordings of people saying or doing things they never did. For categorization purposes, what the algorithms are trying to generate (text, imagery, video, code) or the sponsor's motivation to propose the challenge (commercial, scientific, entertainment, educational) is irrelevant. However, if the competition overview implies the algorithm can be used to create forms of identification, predictions or recommendation systems, the categories "identify," "predict," or "recommend" should be used instead.

While these categories are specifically designed to classify the different goals of Kaggle's competition, they only capture the single dimension of the sponsors' and platforms' objectives. Due to the fluidity of algorithm development, commercial secrecy, and the dynamic nature of digital platforms, these categories do not imply that a competition has no other goals or hidden agendas. These categories are not exclusive and do not describe a logical sequence, since competitions can have multiple goals or be intermediary steps of an undisclosed strategy. For instance, detecting and predicting are the main procedures in these competitions, but the end goal might be to enhance ad targeting or improve product discoverability (recommendation). By assigning just one category to each competition, I aim to understand the primary goal of companies and the main reason they invest in crowdsource predictive model development, and how each of these modalities interferes with and modulates individual behaviour, creating new processes of subjectivation.

Based on these categories, we can observe that half of the competitions held on Kaggle (see Figure 5.15, left) are primarily focused on "identify" (49%), followed by "predict" (32%), "recommend" (17%), and "generate" (2%). These modalities are consistently present through time, except for "generate," which only appeared in the platform at a later date (see Figure 5.16). We can see that, from 2010 to 2016, "identify" and "predict" have the same importance, slightly fluctuating through the years, while "Recommend" was not always a focal point of interest. When Google acquired Kaggle in 2017, the trend changed slightly, favouring the modality focused on "identify," while there was a stagnation of the "predict" and "recommend" modalities. With the popularization of Generative Adversarial Networks (GAN) and Large Language Models (LLMs), a new modality—generate—emerged on Kaggle. Still, it did not get enough traction to be on par with other modalities during the period observed. In terms of money prizes, the 140 identify-oriented

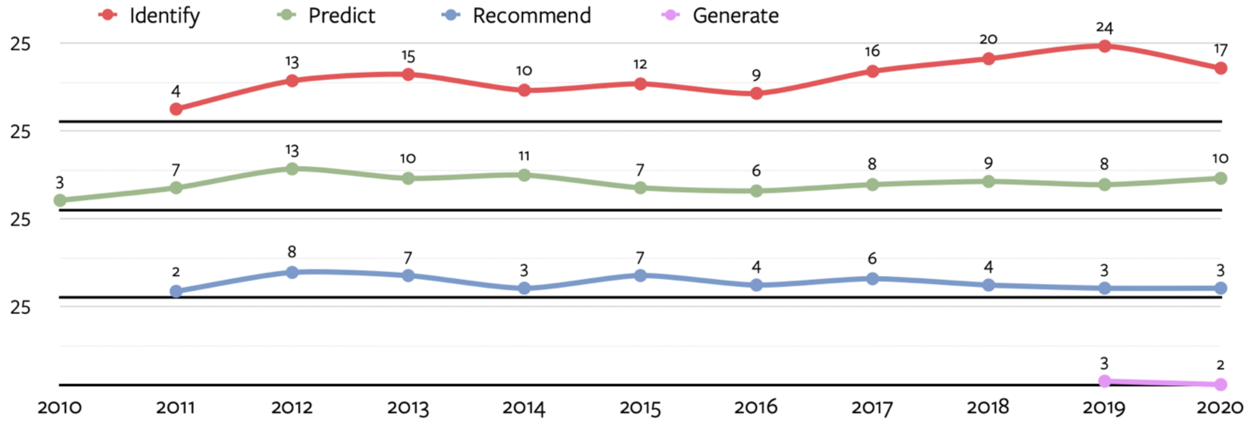


Figure 5.16: Kaggle Competitions Modality Timeline. Number of competitions per year by modality.

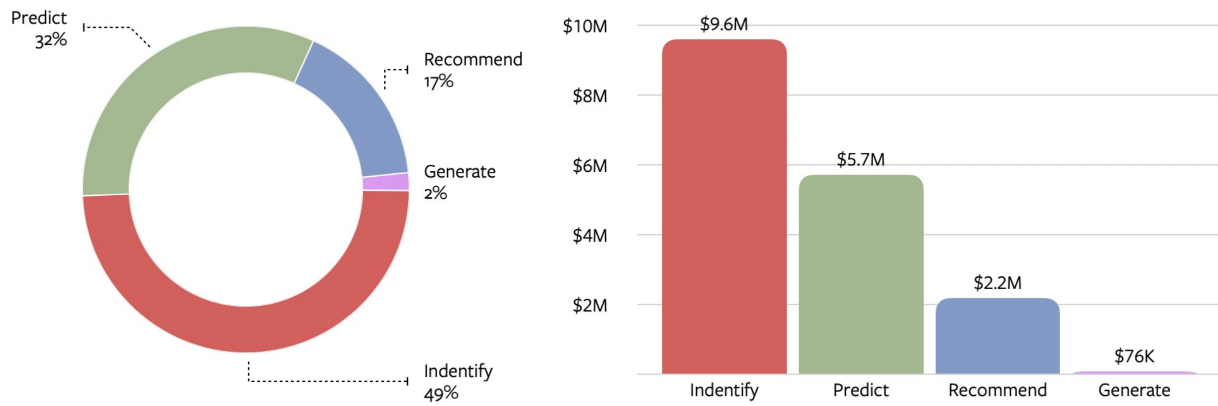


Figure 5.15: Kaggle Competitions Modality Prize Distribution. Proportion of the competition modalities (left) and the accumulated competition prizes in USD$ (right).

competitions paid over US$9.6 million during the period, followed by the 92 competitions focused on "predict" that awarded US$5.7 million, the 47 challenges focused on "recommend" offering US$2.2 million, and the five generate-focused competitions that paid US$76,000 (see Figure 5.15, right).

Categorizing and describing the different types of goals in Kaggle's competitive events and comparing the quantities of events and volumes of money prizes does not tell us the whole story. It does not tell us anything about how machine-learning algorithms are developed, who participated and the conditions of participation in these competitions, the type of data and how it became available to competitors, the overall impacts of the predictive models emerging from these challenges, or how these technologies are used. To better understand the value of predictive models being developed on Kaggle, we need to explore what is behind and inside these competitions: the datasets and techniques used, the goal and context of each competition, the ethical consideration, and further sociotechnical impact and political implications these models have in our society.

I will bring up these considerations in the following chapters, where I thoroughly examine the first three categories (identity, predict, and recommend). On Identify, I focus on Facebook's Deepfake Detection Challenge, a competition intended to improve the detection of fake content on social media that brings up contentious questions about privacy, copyright, definitions of reality, and the relationship between subjectivity, truth and power. On Predict, I analyze the Passenger Screening Algorithm Challenge, a competition sponsored by the U.S. government to improve security in airport terminals, which raises important questions about the normalization of bodies and behaviours, speculative futures, and geopolitical power. On Recommend, I consider the Instacart Market Basket Analysis competition, a typical attempt at improving algorithms to increase sales with serious controversial questions about privacy, hyper profiles, nudges for profit gain, and radicalization of user bases for population control. My goal is to describe how data and Kagglers are mobilized in competitions, what kind of value the output of these competitions has, and how they can be utilized to modulate individuals' habits and behaviours algorithmically.

5.4. Summary

Kaggle began as a small-scale website to promote data science challenges. Goldbloom's initial plan was simple: find personal financial support and have fun. His proposal to use the objectivity of data-driven decision-making found substance in widespread technological advancement, most notably the improvement in machine-learning algorithms and the availability of large datasets. The company grew fast thanks to the investment of venture capitalists. While Google's acquisition encloses the website as a subsidiary, it allowed Kaggle to become a platform with full access to a more extensive computational infrastructure. Today, Kaggle is a machine-learning powerhouse that goes beyond organizing sponsor-led competitions to produce predictive models. It is a successful business housing the world's largest data science community with direct access to massive cloud computing infrastructure and a huge collection of code and datasets for machine-learning development. Following a crowdsource business model, Kaggle has everything it needs in one site to steer the AI industry: code, dataset, free labour, and the means to quickly mobilize them in specific directions.

After introducing code and data repositories in 2016, Kaggle received an influx of thousands of datasets primarily dedicated to being used as training sets for machine-learning models and millions of code excerpts that use these datasets to produce predictive models. Users from around the world are eager to upload and share whatever code and data they have, including ones with personal and sensitive information. Private companies and government agencies also contribute to Kaggle's repositories, which some users have called "data ecology." These repositories are valuable assets in a community that believes machine learning can only be improved with more and more data. However, with thousands of datasets on Kaggle, one could ask how data scientists, engineers, programmers, statisticians, venture capitalists, and tech companies find value in these messy, disconnected fragments of reality. Kaggle is replete with business-oriented and financially obsessive objectives, which naturally drive the overall direction through which they will produce, evaluate, and appraise their machine-learning codes and models.

While data and code are valuable, Kaggle's most important asset is not in its material infrastructure, which comprises cloud computing and a vast repository for machine learning data/code. To be useful, code and data must be acted upon, examined, filtered, aggregated, and refined in specific ways. It is the data science community, millions of Kagglers from different places and backgrounds, who undertake these tasks and make the platform alive. As a reflection of the tech industry, Kaggle is dominated by a young male population. Most of its users are students from India, but the top users are post-graduates working for big companies located in Global North countries. This polarity reveals the inherited salary gap between data scientists in different countries, an inequality the platform business model intensely exploits. Furthermore, to make the workforce more attractive to an economic system that champions individualism and competitiveness, Kaggle ranks users according to their productivity, effort, and participation, providing status conferral to the top users. Unsurprisingly, Kaggle's ranking system became a benchmark for measuring machine-learning talents. Playing the "Kaggle game" became crucial to landing a job in the industry, as companies use the ranking to filter and narrow candidates, who now proudly add their tier title, the number of medals, and their position in Kaggle Rankings on their CVs. As such, Kaggle's gamified system of tiers, medals, points, and categories of expertise emerges as a mechanism that shapes the work conditions inside and outside the platform, and that defines the social conditions in which both machine learning and predictive models are built and disseminated, a topic I will discuss throughout this dissertation.

Competitions in Kaggle involve thousands of users working for free and against each other. Here, like datasets and pieces of code, participants are also raw material that must be systematically exploited and exhausted—or "squeezed" to extract every signal, as Goldbloom (2016) put it. Their work (code, datasets, predictions, descriptions) must solve complex problems to maximize performance, accuracy, profit, and efficiency while reducing costs. With Kaggle, companies are able to mobilize the work of thousands of people at a time for a fraction of the cost of hiring in-house data scientist teams. As noted by Terranova (2004), free labour is "a feature of the cultural economy at large, and an important … source of value in advanced capitalist societies" (p. 73). With the decentralization and disorganization of the workplace promoted by digital networks and assemblages of computers, 'labour' becomes fluid and flexible, mixing with other human activities. Free labour is a keystone of neoliberal companies such as Kaggle, aiming to harness and exploit any form of production, transforming every action into a commodity.

More often than not, these competitions are used to find patterns in human behaviour in order to nudge consumers to consume more. The goal is only to find and even to fabricate notable signals to make predictions about individuals in order to obtain financial profit. The data is stripped away from its subjective, contextual, and contentious nature to become an "objective" predictor and the speculative ground through which data scientists build new forms of reality. The following chapters will consider how code, data, and this large community are mobilized to produce predictive models that can be utilized to modulate individuals' habits and behaviours algorithmically, leading to a redistribution of processes of subjectivation. In other words, I will examine how this new infrastructure of subjectivation aims to mobilize subjective materials usually associated with an individual in order to transform broad conditions of existence and shape the very conditions through which anyone, or anything, gains the possibility of existence in this world.

6. Identify: Deepfakes and Shallow Truths

> ‘There’s deepfakes of you, the email read. Instantly, my pulse quickened. …The video took a moment to load before an 11 second clip began playing in front of me. It was of a woman on top of a man, having sex with him, her whole naked body was visible, her eyes looked straight into the camera as her body moved and her face reacted to the activity. Except it wasn’t the face of a stranger... it was my face … The title had my name in it. The description stated that it was a genuine video of me … I soon discovered another video titled with my first and last name, which appeared to show me performing oral sex. I watched as my eyes connected with the camera, as my own mouth moved. *It was convincing, even to me* [emphasis added]. Why would people ever think it wasn’t real? Anyone could have seen it – colleagues, friends, friends of friends. I slammed my laptop shut and sat in stunned silence. [sic] (Noelle Martin, as told to Daniella Scott, 2020, para. 1-4)

Noelle Martin’s face and identity was used to generate an alternative truth that bears no relation to any reality whatsoever: a simulacrum of herself (Baudrillard, 1983). Her face was masked over someone else’s body to indicate that she was indeed performing such actions. The effect was not just to make the person in the video look like her (the appearance of likeness) but *to be her*, or at least to *make believe that it was her* (a simulation). “Seeing is believing,” as the idiomatic expression goes. Such simulations—particularly the strategic manipulations of visual media content—are not rare in modern society, fitting into a long history of audio-visual manipulation.

Before deepfakes, basic techniques, such as camera positioning, audio distortion, and lookalikes, were used to create “shallow fakes” (Johnson, 2019). When Louis Lumière released one of the first motion pictures in 1896, *L’Arrivée d’un Train en Gare de la Ciotat*, he purposely set the camera at a low angle to the tracks. As the locomotive approached the lenses, the train grew larger and larger in the frame until it appeared as if the locomotive might barrel into the theatre, causing the audience members to scream or even faint in the face of the onrushing train (Loiperdinger & Elzer, 2004). Noelle must have felt a much stronger sensation of terror than Lumière’s audience that day. With digital technologies, a new instrument allowed amateurs and professionals to easily retouch pictures, commonly known as "photoshopping," and make explicit editorial commentary, significantly changing the cultural attitude toward photographic truth and giving rise to "cheapfakes" (Paris & Donovan, 2019).

Deepfakes may be just the next chapter of similar use of technology to create realities and mediate everyday life experiences. Deepfakes are a synthetic media form that broke into the world in 2017 when Reddit users applied open-source software from Google and elsewhere to face-swap celebrities' faces onto the bodies of porn actresses—Scarlett Johansson, Gal Gadot, and Taylor Swift, to name a few—giving the impression that they were exposing their bodies and having sex. These short clips quickly went viral on the lightly regulated anonymous social media platform (Cole, 2017). Historical records show that pornography possesses the power to either make or break the popularity of certain technological advances (Kikerpill, 2020). Unsurprisingly, approximately 96% of currently available deepfake videos can be classified as pornography (Ajder et al., 2019). Almost exclusively targeting women, fake porn follows historical patterns of gendered abuse in cyberspace, from virtual rape in early Internet chat rooms (Dibbell, 1993) and Metaverse worlds (Patel, 2022) to the rape and death threats experienced by Noelle Martin (Scott, 2020) and Rana Ayyub (2018).

Called "Photoshop for videos," deepfakes use deep learning technology, a branch of machine learning that applies neural net simulation to massive datasets. More specifically, deepfakes are facilitated by a technique developed by Goodfellow et al. (2014) named Generative Adversarial Network (GAN), a closed system with an ever-escalating race in which two neural networks try to outwit each other. GANs are constituted by a pair of models: one to create photorealistic images and another that aims to identify whether the generated picture is real or fake, genuine or synthetic. A third ingredient is necessary: data, in huge quantities. Once the bottleneck of this process, the data from where these networks "learn" is available in abundance on the Internet, in particular on social media platforms where the richness of audio-visual content produced on a massive scale by its user repurposes as fuel to train these algorithms to (re)produce or simulate likeness. Previously only possible for big-budget Hollywood movies with enough data and computational power, anyone can now fabricate spectacular and credible imagery and clips, and in very little time.

In its current state, deepfake detection is often referred to as a "cat-and-mouse" game, a term initially used in digital development to describe the competition between quickly evolving cybersecurity attacks and defences. Here, the adversarial game is between deepfake generators and the learned detectors designed to identify them. In this chapter, I am interested in the latter. As a competition modality on Kaggle, "identify" involves the automatic detection, classification, and recognition of bodies, things, and environments. It is typically done through clustering patterns to extract essential information from large datasets, which, in turn, are used to produce a model that can later be applied to detect and recognize similar objects/subjects. It can also be used to create hyper-profiles from personal and impersonal data by identifying users', customers', and citizens' behaviours, habits, and preferences. By asking how these detectors identify deepfakes, we are also asking what is referential for what is fake and genuine. This echoes a similar question asked in 1935 by Walter Benjamin (2008) in *The Work of Art in the Age of Mechanical Reproduction*: How can these algorithms tell the difference between an original and a forgery? Can they spot disinformation when they "see" one? To put it differently, how are deepfake detections designed to make such a distinction and, in the process, produce subjects and their condition of existence?

Following a Foucauldian approach, what interests me here is the practice of the truth. Not the epistemological controversy over correspondence and coherence, but as dividing and excluding; truth as constraining and liberating, as political and ethical. Identifying, labelling, and categorizing are techniques to produce knowledge, define what is knowable in the world, and, ultimately, to produce regimes of truth (Foucault, 1980). It is precisely what this machine-learning modality is about: creating knowledge that can be useful and, at the same time, modulating individuals and collective affective moods and structures of feeling. In machine learning's sociotechnical world, these processes are distributed among different actors—humans, machines, and organizations—that are not always directly interconnected but collaborate to assemble code and predictive models for a specific end.

The Facebook Deepfake Detection Challenge (DFDC) serves here as an instrument of discussion about how facial and object recognition algorithms have been used to define the ontologies of the world we live in. In other words, with the goal of identifying things and individuals, machine-learning algorithms and predictive models are attempts to answer a "what is it?" question: What is a face? What is real, and what is fake? What is true, and what is false? What is there to be known, and what is simply noise? Hence I divide this chapter into three sections: Sense, Mobilize, and Modulate—although they do not imply a rigid and linear sequence in technology development.

In the first section, I discuss how machines came to "learn" how to detect, identify, and recognize faces. While the capability of facial recognition can be used for many different purposes, such as law enforcement, entertainment, and as a convenient method to unlock devices, here I focus on the discriminatory nature of facial recognition systems due to the idiosyncrasies and limitations of their developers, and on the abuse of such technology for creating deepfakes, in particular porn revenge and manipulation of public opinion. In Mobilize, I show how big tech corporations are mobilizing free labour, algorithms, and extensive user-generated datasets to produce predictive models for detecting and asserting what is true and what is fake. Here, I use the DFDC as an example and a point of discussion about the conflicting practices shared by developers and large corporations. In particular, I discuss how Facebook articulated this challenge, the social and cultural biases inherent to datasets used in facial recognition applications, and the ethics (or their lack) in terms of privacy and copyright. Lastly, in Modulate, I consider the outcome of the competition and the context in which it was brought to light by Facebook. In particular, I discuss the repercussions and implications of deepfakes and their detectors disguised as "AI for the good," which have been used to mediate our conditions of existence and produce regimes of truths where current social and political norms are reiterated throughout automatic systems.

6.1. Sense: Facial Recognition

Deepfakes have their roots in the success of the "neural networks" (neural nets, for short), an early form of Artificial Intelligence (AI) that has re-emerged today to power "smart" technologies such as driverless cars, natural language processing, image recognition, and several other applications. Although the concept of neural networks can be traced back to the 1940s (Raviv, 2020), it only began to take shape in the 1960s, when AI was barely a decade old, and computation power and availability were very limited. At that time, computer scientists were crafting algorithms to allow machines to read text from images (computer vision); facial recognition was a far-reaching goal, if not a dream, among researchers. While developing a character recognition algorithm with his partner Iben Browning in 1959, the American mathematician Woody Bledsoe had his first daydream about building a machine that he called a "computer person." Bledsoe envisioned a portable device with a "camera that would fit on my glasses, with an attached earplug that would whisper into my ear the names of my friends and acquaintances as I met them on the street ... my computer friend had the ability to recognize faces" (as quoted in Raviv, 2020, Woodrow Wilson Bledsoe, para. 10). More than just "read" manuscripts and printed material, Bledsoe was excited about the possibility of artificial consciousness, one that would also be his friend and confidant. Bledsoe's character recognition technology never helped him land a single contract (Ballantyne et al., 1996). Still, the underlying concept of his pattern recognition made an impression in the computer science community and got the attention of government agencies.

In 1960, Bledsoe and some of his colleagues founded Panoramic Research Incorporated, one of the first companies to focus on research in artificial intelligence, including pattern recognition, facial recognition, genetics, and neural nets (Boyer, 2012). Much of Panoramic's history is shrouded in secrecy since most of its projects were funded through the U.S. Central Intelligence Agency's (CIA) shell companies, such as the Medical Sciences Research Foundation and the King-Hurley Research Group (Brice, 2020; Lee-Morrison, 2019; Richards, 1978).[12] However, two reports about Bledsoe's first attempts to build a facial recognition system were made public in 2014, revealing the ambitions and challenges of the project.

In 1963, Bledsoe conducted a study to determine the feasibility of a facial recognition machine. Given a database of mug shots and a photograph of random people, the problem was selecting a small set of records from the database such that one of the images matched the photograph. The method's success could be measured in terms of the ratio of the number of matches to the number of records in the database. Bledsoe described the process in two steps: set up a dataset and run the identification algorithms. Using an early digitization device known as the Grafacon or Rand Tablet (similar to the modern smart tablet), a human operator examines a photograph and extracts the coordinates of some facial features such as the centre of pupils, point of widow's peak, the inside and the outside corner of the eyes, and the tip of the nose (see Figure 6.1). Using these coordinates, the team computed a list of 20 measurements, such as the width of the mouth and eyes, the length of the noise, and the height of the forehead (Bledsoe, 1963). This task was laborious, tedious, and slow, involving researchers working up to one hour to process 40 pictures. Bledsoe called it a man-machine collaboration (Ballantyne, 1996). In the identification step, a new image was used to check for a match against the dataset. The system "compute[d] a 'pseudo distance' between the new photograph and each of the previously entered photographs. The distance is computed by summing the differences of similar features, first dividing each difference by a standard deviation of measurement for that feature" (Boyer, 2012, p. 3).

[12] Panoramic had ties with the CIA while the agency was actively pursuing research in behavioural modification (Project MK-Ultra) funded by the same shell companies. Known as the CIA's "mind control" project, it performed experiments intended to develop procedures and identify drugs such as LSD that could be used in interrogations to weaken individuals and force confessions through brainwashing and psychological torture. MK-Ultra used numerous methods to manipulate individuals' mental states and brain functions, such as the administration of high doses of psychoactive drugs, electroshocks, hypnosis, and sensory deprivation, in addition to other forms of torture (Gross, 2019; Project MKULTRA, the CIA's Program of Research in Behavioral Modification, 1977).

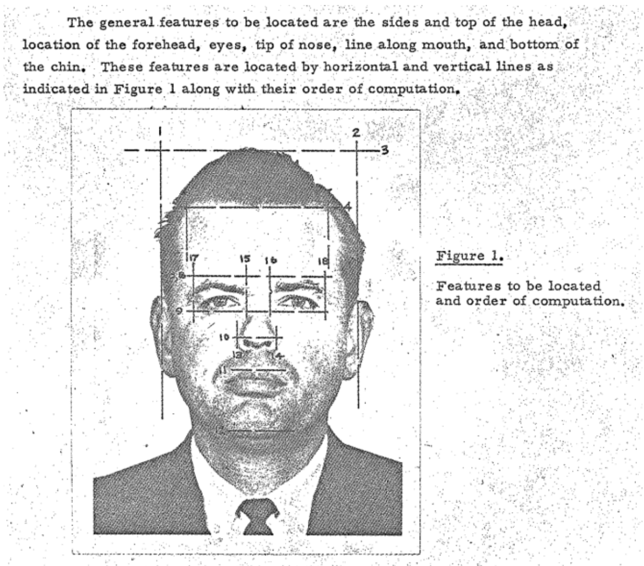

The general features to be located are the sides and top of the head, location of the forehead, eyes, tip of nose, line along mouth, and bottom of the chin. These features are located by horizontal and vertical lines as indicated in Figure 1 along with their order of computation.

Figure 1.

Features to be located and order of computation.

Figure 6.1: Facial Measurement. A sample photograph showing facial features to be located and measured (Bledsoe, 1963, p. 69).

This brief description is, of course, an oversimplification and omits the mathematical and statistical formulas behind the process. Yet, we can see it is no trivial task. In his reports, Bledsoe described the difficulties of facial recognition based on photographs, especially the significant variability in head position (rotation, tilt, distance from the camera), lighting intensity, facial expression, and aging. Bledsoe solved a few of these problems by making the computer distort the picture to normalize the distance representing the face as if it were taken in a frontal orientation. Once the computer found a match, the selected results were presented to a human, who finally decided which photographs matched, if any. Finding the matching face became a competition (a game) between humans and machines. In experiments involving a database of several thousand different photographs, the computer consistently outperformed humans when presented with the same recognition task (Ballantyne, 1996), reducing the number of pictures that needed to be considered by the human to one percent of the photographs in the database (Bledsoe, 1963).

These early attempts to implement facial recognition using neural networks were limited by the machinery available at the time. In the 1960s, Bledsoe had access to only a few machines with no more than 192kb of RAM, a small fraction of the capability embedded in today's smartphones. As computer power increased, the technology behind neural networks also improved performance and their ability to detect and recognize faces. For instance, in the 1980s, researchers organized the nodes in neural networks into a limited series of layers: data flows into an input layer; then it is combined and fed into a second layer that contains nodes concerned with identifying simple

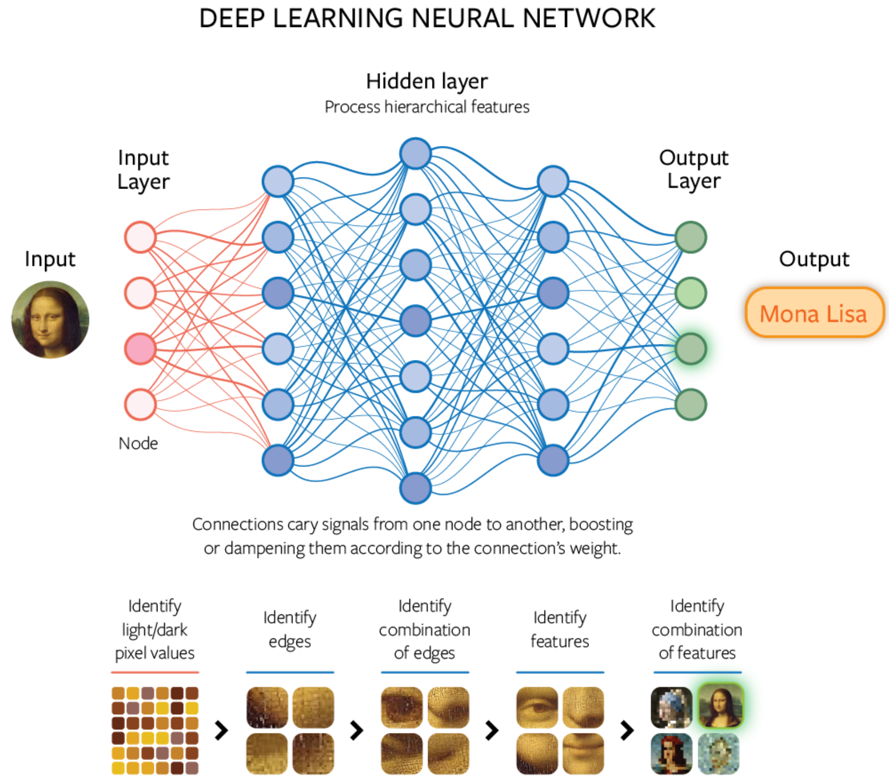


Figure 6.2: Deep Learning Diagram. Deep Neural Networks are designed to model the brain's neurons and links with a web of simulated nodes and connections. Designed by L. Frizzera.

features like lines and curves; then into a third layer that has nodes responsible for recognizing more complex shapes such as eyes and noses; and so on (see Figure 6.2). In the mid-2000s, a new method initially dubbed "Deep belief nets" by Geoffrey Hinton (Mohamed et al., 2011), but later generalized as "deep learning" (LeCun, Bengio, & Hinton, 2015), allowed the number of layers to grow exponentially, proving that the technology was economically viable. Large tech companies, such as Google, Amazon, and Facebook, quickly began incorporating deep learning neural networks into every product they could, as did researchers in biomedicine, high-energy physics, and marketing, as well as financial brokers.

6.1.1. Seeing is Believing

There are two important considerations to make regarding Bledsoe's influential work, which naturally extends to the latest advances made by other researchers such as LeCun, Bengio, and Hinton: the assumptions that (1) a machine can not only simulate how the human brain works in terms of image processing to recognize patterns but also learn and act upon these images in the same way humans do, and that (2) individuals are as simple as the sum of their average facial measurements. Both assumptions have become core axioms in AI research over the last 60 years. The first assumption generalizes how the human brain works and is clearly stated as a metaphor: Neural Network—a network of highly simplified nodes, analogous to brain cells, known as neurons. For instance, to recognize what is in an image, the system pipes the picture's raw pixels into the nodes of a neural network, where they become weighted signals. These signals then flow from one node to the next along connections, analogous to the synaptic junctions that pass nerve impulses through the neurons. Depending on how these connections are organized, the signals combine and split as they go until they eventually activate one of a series of output nodes. Each output, in turn, corresponds to a high-level classification of the image's content—a human face, for example. Advocates of this method argue that neural networks could be much better than standard algorithms, as these networks would not need to be programmed, just trained, as if they were skilled workers: "Simply show your network a few zillion examples of, say, puppies and not-puppies, like so many flash cards, and ask it to guess what each image shows" (Waldrop, 2020).

Neural nets are typically trained by "supervised learning" (IBM Cloud Education, 2020). They are presented with many examples with the object to be learned as input, and then the connection weights are gradually adjusted. The neural network is optimized until it "learns" to produce the desired result. To learn what a cat is, for instance, a neural net may be presented with many different pictures of cats and slowly learn to predict what a cat looks like (its shape, colours, the various proportional relationships in its body, etc.). The outcome is predetermined, leaving the developer to optimize the network for a narrow objective and specialized task aiming for a high degree of accuracy in the desired goal. So, for facial recognition systems, it is crucial to have a massive list of examples with different types of faces, including diversity in skin colour, skull structure, gender, age, ethnicity, the presence of facial hair, accessories, or any other distinctive mark (beard, moustache, glasses, piercings, tattoos, mask, moles, freckles, or even extreme asymmetry), as well variability in light conditions, and camera angles and perspectives. Bledsoe's research in the 1960s started with small datasets, with as few as 122 images, reaching over 2,000 photographs over several years. Such limitations were primarily due to the manual labour required to preprocess pictures as well as the lack of high-quality publicly available photos. This problem has been solved in the last 20 years by the development of the Web 2.0, and subsequently the rise of social media platforms, the massive adoption of sense-ready mobile technologies such as smartphones, and the push for a more controllable built environment such as smart cities and the Internet of Things. It is generally believed that using deep learning and millions of data points to train a predictive algorithm would, in theory, allow for a more efficient and accurate facial recognition outcome.

However, selecting and curating datasets for supervised learning is a complex socio-cultural and political-economic task, making it less objective and more prone to subjective interpretations of what to include and what to exclude from a dataset. The indiscriminate use of data and the oversimplification of how machine-learning enthusiasts understand and describe individuals leads us to the second assumption: individuals are as simple as the sum of their average facial measurements. Bledsoe believes that the most promising path to automated facial recognition would reduce a face to a set of relationships among significant landmarks such as eyes, ears, nose, and lips. The system he imagined can be traced back to the end of the nineteenth century when Cesare Lombroso (2006), Francis Galton, and Alphonse Bertillon used anthropometrics to identify criminals (Sekula, 1986). Bertillon pioneered the modern 'mug shot' in 1876, describing people based on 11 physical measurements, including the length of their feet and the length from the elbow to the end of the middle finger (Stephen, 2013). This assumption became canonical in the Machine Learning field, where researchers use whatever datasets are at their disposal to produce correlations between facial measurements and a person's identity. While Bledsoe used mug shots to extract 20 facial measurements (Bledsoe, 1963), modern facial recognition methods use unauthorized photos scraped from the Internet to extend this list, recently reaching more than 120 vectors representing the most important features of a face (Schroff et al., 2015). I will return to this discussion in the following sections. Suffice it to point out that collections of facial portraits produce narratives (including accounts of aesthetics, cultural spaces, social practices, and financial conditions)[13] about collectives of individuals leaving behind visual artifacts that machine-learning-driven automated systems would pick up. Thus, facial recognition systems have become controversial since they are as good as the data they are trained with, reproducing the training set's limitations, distortions, and biases.

6.1.2. Genuinely Fake

In the 2010s, with the success of facial recognition systems and more computing power available in cloud infrastructures, deep learning neural networks started to be explored in different and

[13] Jessica Helfand (2019) examines these narratives in her book *Face: A Visual Odyssey*. From historical mugshots to Instagram posts, she explores how the face has been perceived and represented over time and how it has been instrumentalized by others, including facial recognition systems.

controversial ways. Perhaps the most (in)famous example is a technique known as "deepfake," a direct sub-product of specialized machine-learning algorithms dubbed a Generative Adversarial Network (GAN). This idea of confronting neural networks competing in an ever-escalating race came to fruition in 2014 when Ian Goodfellow, then a PhD student at the University of Montreal, presented his research at the 28th Conference on Neural Information Processing Systems (NIPS).[14] In simple terms, GANs comprise a pair of models: a generator and a discriminator. The generator creates a synthetic image that looks like it belongs to a particular set of images, while the discriminator decides whether the image belongs or not to the original collection of images. The generator takes the discriminator's feedback as a new input on each new cycle until the output improves to the point where the discriminator cannot distinguish between real and fake, genuine and synthetic. This relatively simple idea quickly raised Goodfellow to tech stardom, having been cited in *MIT Technology Review*'s "35 Innovators Under 35" (Knight, 2017) and hired by Google, OpenAI, and Apple.

Goodfellow believes that GANs have the potential to generate objects that we can use in the real world, citing examples such as medicines that need to be more effective or batteries that must become more efficient (Giles, 2018; Bardají, 2019). GAN is seen as a game changer in areas where data availability is the major bottleneck in machine learning, often needing to ingest millions, if not billions, of data points to produce a "reliable," "accurate," and "desirable" model. For instance, in healthcare research, where privacy concerns prevent researchers from harvesting patient data without consent, GANs can be an alternative to help solve this problem by generating fake records that can be almost as good as the real ones. However, as I show throughout this dissertation, gathering data without consent was never a problem for machine-learning enthusiasts and technologists who systematically break ethical principles, treating security, privacy, and copyright issues as an afterthought (see Copeland, 2020; Marelli et al., 2021). On the other hand, a machine designed to create realistic fakes has the potential to replace the original, creating confusion and distrust, becoming a perfect weapon for purveyors of fake news who want to influence public opinion (Maras & Alexandrou, 2019), elections (Kerner & Risse, 2020), and the stock market (Giles, 2018), and mediate everyday experiences.

[14] Goodfellow's work is filled with controversies. For instance, he has been accused of not recognizing the contribution of previous researchers in machine learning (Vance, 2018). Indeed, his GAN is heavily based upon previous research by Jürgen Schmidhuber (1991; 2020) on predictability minimization and artificial curiosity.

GANs did not create these problems, but they made them worse. In just four years, the technology behind deepfakes has become open-sourced, easy to work with, and deeply controversial. As discussed in the introduction to this chapter, since Samantha Cole's (2018) revelatory reporting on the viral popularity of deepfake and non-consensual pornography, Reddit users began experimenting with less offensive content, looking for entertainment and humour, such as memes with actor Nicolas Cage's face swapped into various movies (Haysom, 2018) or making Barack Obama call Donald Trump a "complete dipshit" (BuzzFeedVideo, 2018). Yet, the danger of deepfakes is chiefly acknowledged when there is potential for political abuse, such as spreading disinformation to manipulate public opinion or discredit the opposition. For instance, in 2018, investigative journalist Rana Ayyub was a victim of a deepfake porn plot intended to silence her. After being invited by major news outlets to comment on the political circumstances surrounding the rape of an eight-year-old Kashmiri girl, Ayyub faced a concentrated attack from the nationalist Bharatiya Janata Party (BJP) to discredit her image and work. Ayyub (2018) was tormented and surprised to see her face on a deepfake porn: "I was shocked to see my face, but I could tell it wasn't actually me because, for one, I have curly hair and the woman had straight hair. She also looked really young, not more than 17 or 18" (para. 12). Still, the video went viral across India creating widespread knowledge of "witnessing" Ayyub in an intimate setting undermining her standing as a journalist. Like Noelle, Ayyub had little chance against the perpetrators since, in most cases, the law and regulatory marks have not yet matched the effects of deepfakes. The combination of powerful neural network algorithms, the way news is spread through social media, and the exploitation of our ability to discern truth from fake news have set us up for severe consequences.

Digital platforms acknowledge that misinformation can cause social distress and political turmoil, especially after Cambridge Analytica exploited Facebook's techno-economy to spread fake news. In 2018, Reddit, Pornhub, and Twitter banned the posting and sharing of Deepfakes involving "involuntary pornography (Clark, 2018; landoflobsters, 2018; Maddocks, 2020; Scott, 2020). The strategy quickly failed due to the logic of digital platforms based on the rapid circulation of user-generated content. Social media platforms, in particular, are ambiguous about fakes and disinformation, usually allowing these types of content to circulate among their users because they radicalize the user base with the exponential potential to generate engagement (likes, retweets, comments) and, as a result, a fast-paced increase in revenue and profit (Srnicek, 2017). As a result, banning deepfakes from one platform only displaces the problem to another digital space.

On the other hand, big techs noticed the increased ethical implication of being unable to tell the real from the fake. For technologists, the problem was not the method but a matter of optimizing the algorithm with even more data. Instead of halting or banning the use of GANs to produce deepfakes, some of these companies decided to crowdsource the problem to find ways to detect deepfakes or authenticate content. The goal is not to prevent deepfakes but to create better models to distinguish the real from the fake, which in turn has the power to mediate the conditions of existence to produce regimes of truths where current social and political norms are reiterated throughout automatic systems. The following section explores how crowdsourced competition sponsored by big tech mobilized user data and an army of developers to create models that could tell deepfakes from authentic content.

6.2. Mobilize: Deepfake Detection Challenge

The controversial usage of GANs to produce realistic AI-generated videos of real people doing and saying fictional things has significant implications for determining the legitimacy of information presented online. The technology was getting out of hand. Big tech companies had no instruments or benchmarks to detect deepfakes. Pressured by the increased spread of disinformation on its social media networks, Facebook saw an opportunity to respond to the accusation of not doing enough to moderate harmful content (Ehl, 2019; Gallagher, 2019; Lima, 2019) by proposing a machine learning competition to debunk deepfakes. In September 2019, with Amazon, Microsoft and a group of academic researchers, Facebook proposed leveraging the machine-learning community on Kaggle to crowdsource the work, launching the Deepfake Detection Challenge (DFDC) on the platform. The company invited the Partnership for AI (PAI)—a non-profit organization focused on creating solutions to AI advancements for positive outcomes for people and society—to coordinate and oversee the challenge, which included determining competition governance, such as shaping the scoring tactics, leaderboard, model access parameters, and the entry requirements. The event, officially launched on December 11, 2019 (see Figure 6.3), offered US$1 million for the best machine-learning techniques and predictive models to identify videos with synthetic facial or voice manipulations (Deepfake Detection Challenge, 2019a).

A significant problem for the machine-learning community is access to good, balanced, reliable, and affordable datasets, preferably free and without restrictions. The vast majority of training data is scraped and collected from the Internet and digital gadgets, usually without the owner's permission or knowledge. Data is such an essential commodity for machine learning enthusiasts that they

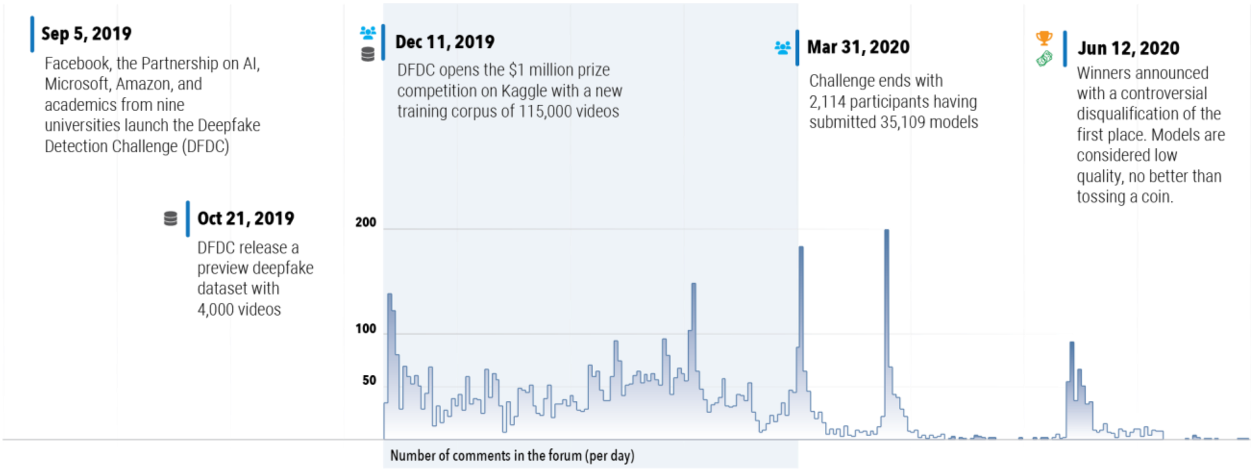


Figure 6.3: Deepfake Detection Challenge Timeline. Timeline showing the competition milestones and number of comments in the forum through time. Designed by L. Frizzera.

accept it as is and rarely pay attention to privacy, copyright, or ethical issues. Kaggle hosts thousands of databases without indicating their source, clearance, consent, or restriction. Putting together a dataset for deepfake detection adds another level to this already unregulated practice since the items themselves are already ethically questionable: millions of images are extracted from unknown sources to produce involuntary fake pornography and electoral disinformation. Deepfake datasets available at the time, such as DeeperForensics, Celeb-DF, and FaceForensics++, source the data from the Internet (YouTube in particular), perform face swaps between public individuals, are unclear about the rights and restrictions of the underlying content, and lack the individuals' consent represented in the deepfakes (Dolhansky et al., 2020). The complex sociotechnical challenges around curating a dataset of synthetic media are not easy to overcome. Since a deepfake dataset would contain human subjects, PAI recommended that the DFDC's dataset reflect a realistic distribution of deepfake types while protecting the rights of the people portrayed on these deepfakes (Leibowicz, 2019). Nonetheless, since Kaggle works as an unregulated job market for machine-learning research, Facebook decided not to subject the dataset to an independent board review for ethical evaluation (see Kofman, 2020). Instead, the social media platform commissioned a realistic dataset with paid actors who had to sign a waiver from their image rights so the company could "ethically" produce deepfakes for the competition.

Paying actors to license their faces to make deepfake-style marketing clones is becoming common practice. Hour One is one of the companies taking deepfake tech mainstream (Hao & Heaven, 2020), using it to produce mashups of real footage and AI-generated video. Other companies have used

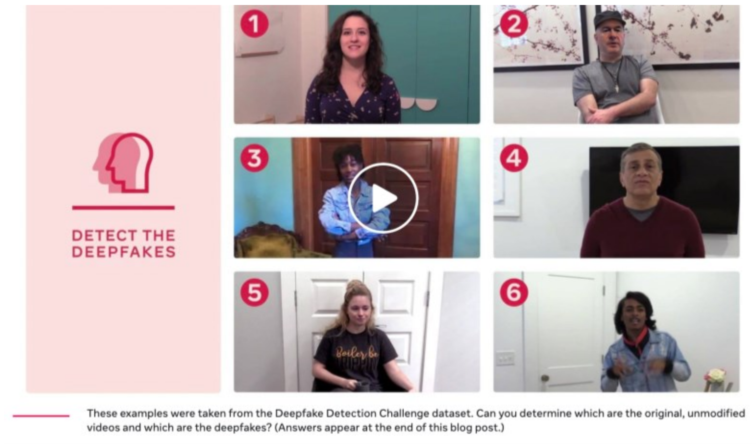


Figure 6.4: Deepfake Sample. Examples of deepfakes created for the Deepfake Detection Challenge. Clips 1, 4, and 6 are original, unmodified videos. Clips 2, 3, and 5 are deepfakes (Deepfake Detection Challenge, 2019a).

professional actors to give life to deep-faked personas (Hao, 2020). While hiring actors is common in marketing and audio-visual productions, Hour One does requires no particular skills for a job. Instead, it targets unskilled, low-wage workers and vulnerable populations who would willingly hand over the rights to their own faces in exchange for a small amount of money. Its current portfolio illustrates this reality: 80% of characters are under 50 years old, 70% are female, and 25% are white (Heaven, 2021).

Facebook followed the same strategy, collecting 25 terabytes of raw footage from 3,426 individuals (Dolhansky et al., 2020) hired for an undisclosed amount who agreed to appear in the dataset where computer vision algorithms may manipulate their faces (see Figure 6.4). Due to technical and time constraints, the DFDC dataset only contains video clips from 960 individuals; the selection criteria were, however, not disclosed. With this collection of almost a thousand 15-minute video clips, the Facebook AI team was able to produce over 100,000 deepfake videos (500 GB) using several GAN-based and non-learned methods. Facebook claimed that the DFDC is the largest currently and publicly available face swap video dataset (Dolhansky et al., 2020).

6.2.1. Gamified Community

According to PAI, the competition sought to “mobilize the global AI research community toward this timely issue, while keeping in mind the real-world implications and contexts in which deepfake videos are often weaponized” (Leibowicz, 2019, para. 3). By encouraging machine-learning researchers and developers worldwide to take part on the competition, the consortium behind DFDC was looking for a dynamic environment where a diversity of actors could take a fresh look into the problem in order to design better algorithms to identify deepfakes, minimize deception, and enhance the integrity of media. At the same time, the crowdsourced nature of the event, a staple attribute of social media platforms, was proven to be a cheap path for experimentation, research, and development, where made-up rules are designed to maximize labour exploitation in a highly imbalanced and competitive environment. Kaggle has been training a new generation of computer scientists to work in a gamified competitive environment for over a decade. As the self-proclaimed largest data science community, Kaggle was the perfect place to host the challenge.

The competition was designed to run for four months, from December 2019 to April 2020. With an eye on the grand prize, Kagglers strategically team up to increase their chances of being in the "money zone." Not surprisingly, "Looking for a Team Thread" (Howard, 2019a) was one of the forum's busiest threads, with almost 200 comments and replies. The thread reveals how transactional Kagglers' interactions can be when seeking partners. While advanced users look for partners with solid machine learning and computer vision skills to improve their chances, less skilled participants offer their free time, including weekends and holidays, and access to powerful (personal or cloud) infrastructure for an opportunity to learn. For instance, a user from Dubai was willing to buy hardware or invest in virtual cloud computing if he found highly skilled teammates. A Computer Science master's student from China made his own hardware available (two 12 GB NVIDIA Tesla P100) to whoever wanted to team up with him. An engineer from France mixed his personal skills with machine capabilities: "Python, Keras/Tensorflow 2.x, Win/Ubuntu, GTX1080ti, videogame dev, sound engineer, got plenty of free time and I'm fast. Looking for teammates" (Caby, 2019, n.p.).

Creating predictive models using deep learning is a highly intensive computational task. Depending on the available hardware, it can take hours, days, if not weeks, to train, refine, and optimize a model. The minimum required equipment to train such a model can cost tens of thousands of dollars, an asset certainly not in reach of most of participants. However, cloud computing has shifted these expensive operations with cloud data centres, where virtual machines are available for rent per computing millisecond. In an effort to attract more participants and consolidate its ownership over Kaggle, Google offered US$200 coupons and TPU[15] quotas on the Google Cloud Platform (GCP) for the duration of the challenge (Elliott, 2020). Amazon, also part of the DFDC consortium, made a similar incentive, offering between US$1,000 to US$10,000 in Amazon Web

[15] Tensor Processing Unit (TPU) is an AI accelerator application-specific integrated circuit developed by Google for neural network machine learning, using Google's TensorFlow software.

Services (AWS) credits depending on the viability or success of detecting deepfakes (miwojc, 2019). By choosing one or the other, participants were also contributing to the success of these companies in the cloud service industry.

The high number of comments and replies shows a strong commitment to the task. Still, the objective was less about creating a deepfake detector and more about the competitive gamified environment that doubles as a window display for future hiring processes: users were most interested in improving and consolidating their position on the leaderboard for a better chance of getting a medal. For instance, an undergraduate from India with two years of experience in deep learning was “ready to dedicate 3-4 hours daily (probably even more during holidays) [with a team] with similar or more commitment to the challenge” (Chaudhari, 2019, n.p.). Beginners to machine learning (primarily students, mainly from India and China), eager to learn from more experienced users, were willing to work full time—more than 30 hours per week—and disrupt sleep cycles to be in a team with chances of getting a medal. Less apt Kagglers tend to be side-stepped because they do not have suitable hardware to share or the necessary skills for the task. Experienced users are simply unwilling to team up with beginners to avoid extra work training or teaching the basics of machine learning and computer vision. In the harsh environment of a high-prize competition, the replies can get very unpleasant, as shown in this example:

> Sry bro no place for newbies, if you are new in CV [computer vision] and wanna try this competition I guess you are f*d up, this is no joke to make a 2 class classifier for 1 million usd. Imagine the difficulty and new learning even I am getting from this competition when from the very start my focus was in deep learning with computer vision. Was that harsh?, If so then nvm but I am going for the win and not for learning. (Sheoran, 2019, n.p.)

On the one hand, the aggressive behaviour and ambivalence between collaboration and competition are typical in a gamified environment such as Kaggle. Participants are eager to win at all costs, looking for productive partnerships that could take them to the top. On the other hand, digital platforms are more interested in harnessing the free labour available in the digital world (Terranova, 2004). Even though in Kaggle’s “winner-take-all” economy, only the top 3-5 teams receive financial compensation for their work, the DFDC attracted 2,904 participants self-organized into 2,265 teams,[16] who collectively submitted more than 35,000 models to the website (Ferrer,

[16] Kagglers are encouraged to participate in the challenge even after its conclusion. These figures considered the number of participants and teams at the moment of the data collection in November 2020. According to Dolhansky et al. (2020), 2,114 teams were registered in the competition when the results were released.

2020b). Many joined, enticed by the large money prize—US$1 million; others were drawn to the competition due to its technical challenges; many more were there looking for opportunities to get involved with the community, to learn from one another, to have a sense of belonging, but also to acquire knowledge and skills for their current or aspiring position as data scientists. In contrast, large AI companies, such as OpenAI, Google Brain, Facebook, and Apple, are willing to pay salaries higher than US$1 million for top AI talent engineers such as Ian Goodfellow, GAN's creator (Metz, 2018).

The consortium led by Facebook took advantage of Kaggle's gamified platform to the full potential of a human-machine-data assemblage. Together with the dataset, storage, and cloud computing, the participants were also raw material to be "squeezed" for cost reduction, performance, accuracy, and commercial value. As Max Levchin (2013) speculated in early 2013, Kaggle was meant to be a place where companies could rent the brains of data-mining geniuses to solve hard problems for a little extra cash. The platform exploits its community as a "stock of brains," a commodity, which, in turn, becomes a workforce willing to exchange their cognitive abilities, if not for free, at least for low-value tokens of appreciation, such as virtual medals, which are distributed among the top 10% of the participating teams.

6.2.2. Deep Sociotechnical Tendencies

As soon as the competition began, Kagglers pored over the data to run Exploratory Data Analysis (EDAs) and post their first impressions. For instance, Aleksandra Deis (2019) explored the training sample folder. The metadata of each item in the dataset contains the video's filename, a label indicating if the video is "REAL" or "FAKE," and, if fake, a reference to the original video. Out of 400 videos in the sample folder, four in five were "FAKE." By simply visualizing a few of these videos, she could identify visual attributes of FAKE videos: distorted noses, blurry face edges, non-realistic glasses, flickering eyes; and features that make a credible REAL person, such as actual teeth (not just one white blob) and glasses with a reflection. nosound (2020) found duplication in REAL videos and noted audio distortion (fake audio) in 10% of FAKE videos, which sparked discussions about whether or not fake audios should be part of the detection model, a feature downplayed by the competition's sponsor (Ferrer, 2020a).

While many other exploratory analyses considered other particularities in the dataset—how many faces appear in each video (MPWARE, 2019), the length of the video and the quality of the image (Preda, 2020), the similarity of the faces (Mendonça, 2020)—there was no mention or attempt to check for fairness or balance representation in terms of age, gender, race, or ethnicity. In over 6,000 replies to 772 threads on the forum, only two users briefly mentioned how facial detection could be biased: ryches (2020) acknowledged that current "pre-existing face detection models are significantly worse for the black participants (either because of lack of contrast or because of lack of training data on black people for the face detection algorithms)" (n.p.). Another user seconded his remarks, but they fell short of starting any discussion about further exploratory examination of the dataset in this context. In DFDC, as in most machine-learning competitions hosted by Kaggle, social bias and political issues do not play an important role, making participants in these competitions avoid discussing the topic or ignore anyone who raises questions about these issues.

However, social and cultural bias in deepfake detection, and facial recognition algorithms, in general, significantly affect their predictions, causing them to fail both the individual and the society. Xu et al. (2019) argue that the main reasons for AI bias originate from unbalanced training datasets and the lack of awareness that training data might be biased in several regards. Xu et al. are right about unbalanced training sets, which have been used in machine learning since its early days when Bledsoe developed the first facial recognition algorithm in 1963. However, the lack of awareness might not be an issue in today's machine-learning community. DFDC participants are pretty open about the tools and datasets they use. For instance, when asked to list external datasets and pre-trained models and their intent to use on the DFDC, Kagglers listed more than 100 resources (Howard, 2019b). Among the dataset, participants mentioned they were using 1 Million Fake Faces, Celeb-DF, Coco, DeepfakeTIMIT, DeeperForensics-1.0, FaceForensics++, Flickr-Faces-HQ, ImageNet, MS-Celeb,[17] and YouTube-8M, to name a few. Kagglers also mentioned algorithms they planned to use to speed up the facial detection and deepfake creation processes, such as FaceNet, MTCNN (Multi-Task Cascaded Convolutional Neural Networks), VGG16 (a type of Convolutional Neural Network), XceptionNet, and StyleGan. The links shared among the participants point to academic studies, scholarly papers, code, benchmarks, videos, guides, tutorials, and blog posts to understand the specificities of each resource. Deepfake detection, however, is not about social fairness or balanced training sets; rather, its primary goal is to accurately measure whether a picture is real or fake.

While Kaggle's community remains silent about these issues, academic research shows how nasty most datasets and detection algorithms can be. Recent studies reveal that most datasets used to

[17] MS-Celeb was launched in 2016 by Microsoft but was taken down in 2019 due to bias and copyright issues. However, the dataset is still available online via torrent (Brandom, 2019a).

train and test machine learning models are inherently biased. For instance, Camilleri et al. (2019) describe all kinds of social biases on facial recognition models, from racism to sexism to ageism. According to their research, these biases have been found in all forms of machine learning because data (text, audio, image, video, etc.) must be described using language. That is, data needs to be encoded and labelled in such a way that we (and the machine) can make inferences and references. In supervised learning, the machine-learning algorithm ingests labelled data, a task usually crowdsourced through Amazon Mechanical Turk or similar services that leverage an underpaid workforce under severe competition for available assignments and depressed by scant career prospects. Annotators are motivated to quickly provide the “right” answer and move on to the next micro-assignment without enough time for any considerations about the data. The average annotator brings their own life experience, idiosyncrasies, and biases to their annotation. For instance, on average, an North American annotator is more likely to label phrases in African American Vernacular English (AAVE) as toxic, leading AI toxicity detectors trained on these labels to disproportionately flag content produced by African Americans as harmful (Wiggers & Coldewey, 2022). The annotator is not the only one responsible. Palmar et al. (2022) found evidence that the instructions influenced the annotators to follow specific patterns, which then propagated to the datasets. For example, over half of the annotations in Quoref, a dataset designed to test the ability of AI systems to understand when two or more expressions refer to the same person (or thing), start with the phrase “What is the name,” a phrase present in a third of the instructions for the dataset.

Like humans, AI systems are susceptible to developing biases from sources that are not always obvious. In 2016, Bolukbasi et al. (2016) revealed the sexist behaviour of a commonly used word embedding method trained on Google News articles to predict professional roles. The predictive model made “accurate” predictions, assigning gender to specific social roles. Male professions included “maestro,” “skipper,” “philosopher,” and “captain.” In contrast, female occupations have a strong correlation with “nurse,” “receptionist,” “librarian,” and “hairdresser.” The model reflects and reifies the societal biases found in commonly used data sources in many businesses, governments, and research contexts without any awareness or correction. The same is true in many facial classification models. Facial image analysis describes a range of face perception tasks, including face detection (Mathias et al., 2014; Zafeiriou et al., 2015), face classification (Reid et al., 2013; Rothe et al., 2016) and facial recognition (Gershgorn, 2021A; Wen et al., 2016). Datasets like ImageNet, COCO, CIFAR, and CelebA collect images freely available on the Web, such as Google Images, Wikipedia, social media platforms, and stock photo services. These datasets’ lack of racial and gender diversity is alarming, affecting their accuracy and favouring results toward white male individuals (Camilleri, 2019).

In 2009, a video titled "HP computers are racist"[18] (wzamen01, 2009) went viral on YouTube, revealing how an apparently harmless webcam equipped with a facial detection mechanism can affect people's lives. The video shows a black man testing facial tracking technology using a webcam in a store without much success. The camera simply did not recognize his face. When he called a white friend to try it, the camera quickly turned and zoomed in toward her direction. Switching back and forth between the black man and the white woman, the machine revealed the inherent prejudice its engineers failed to see: the machine-learning model was biased against black people. Another example of such behaviour occurred in 2015 when the image recognition algorithms in Google Photos were classifying black people as "gorillas." Machine-learning experts quickly dismissed the problem, suggesting it simply needed more data. After three years of collecting data and unsuccessfully solving the problem, Google's engineers removed "gorilla" as a category to avoid miscategorization. As a result, gorillas are not recognized on Google Photos (Vincent, 2018): they simply do not exist for the machine. It is not so much that these specific events may have physically or emotionally hurt someone in the black community. More importantly, this example shows how machine learning and AI have been developed to create a technology that might not recognize black individuals as part of society or even consider them human.

Dataset for deepfake detection is not an exception. A closer look at FaceForensics++, one of the favourite datasets on DFDC, reveals that it is overwhelmingly composed of female white Caucasians (Trinh & Liu, 2021). Celeb-DF, another frequently cited dataset on DFDC, mainly contains images of white celebrities' interviews on YouTube in which they present themselves to the camera, producing a tendency toward a specific aesthetic that uses heavy makeup (Xu et al., 2022). DFDC's dataset is no different. Despite reaffirming diversity in gender, skin tone, ethnicity, age, and other characteristics (Ferrer et al., 2020), Xu et al. (2022) show otherwise: the dataset provided by Facebook is highly biased in terms of demographic attributes, with disproportionally more white and Caucasian faces than black and Asian individuals. Skewed and biased datasets have a substantial impact on deepfake detection models, leading to fairness issues in real-life applications since internal representations of neural network models preserve information from the training data even if it is not directly needed for the model objective (Terhörst et al., 2020). As a result,

[18] Available at https://www.youtube.com/watch?v=t4DT3tQqgRM

deepfake detection is prone to produce false positives, identifying a fake face as a genuine citizen, and a real person as fake, someone who is not part of “our” society. It is important to note, however, that when PAI became involved in the DFDC, Facebook was already in the latter stages of the dataset production, restricting the organization’s ability to make any requirements in relation to the issues found by Xu et al.

It seems that, despite several studies pointing out the evidence of an unbalanced ratio of race, gender, and other demographic attributes on machine-learning training sets, Facebook’s DFDC was not interested in making any effort to address the problem. The consortium was more concerned with the legal use of data to produce deepfakes than actually protecting individual rights or having a more balanced and fair representation of individuals in the dataset. Despite most of the resources listed in the DFDC having been publicly and heavily criticized, Kagglers insisted on using the same instruments and datasets everywhere. This practice reveals a dysconscious bias among machine-learning practitioners, who in fact might be aware of the social harm of such tools—as exemplified in Pedro Bernardo’s and ryches’s considerations mentioned earlier—but who do not necessarily see themselves harming anyone. Following Lorna Roth (2009), I am not directly accusing engineers and private companies of being deliberately racist, although we cannot simply view their actions as unintentional. I prefer the more nuanced position suggested by Joyce E. King (2001), whereby racism and sexism are not unconscious actions (that is, an absence of consciousness) but an impaired consciousness or distorted way of thinking about social constructs such as gender, race, and sex. King argues that this kind of bias tacitly accepts dominant norms—Western patriarchal society expressed and caricatured in the normative individualist middle-class white male—that justifies inequality and exploitation by taking the existing order of things as given. As digital technology, particularly facial recognition, is predominately developed to reinforce socioeconomic norms, deepfake detection cannot help by discriminating against dark-skinned individuals and other minorities.

6.2.3. Fluid Ethics

Social and cultural bias is hardly the only issue concerning the unregulated and unrestrained use of large quantities of online material for machine-learning training. Competitions held in and by Kaggle are not about some sort of humanitarian endeavour nor a place for social justice. Kagglers rarely question how datasets are built or whether they are balanced and fair. They do not spend much time considering the ethical outcomes of pre-trained models or object detection algorithms and hardly ever inquire about the goals of competitions and the interests of their sponsors. Instead, competitors use whatever they have at their disposal to get them into the prize zone. Concepts such as fairness, ethics, and integrity only emerge in relation to the competition itself: its rules, access to the right tools, data, and algorithms, the metrics of assessment, and the distribution of prizes. DFDC is an example of how ethical rules about machine learning can be flexible depending on the context where predictive models are created, further developed, and deployed.

Kaggle's community was caught by surprise when the DFDC's results were announced on June 12, 2020, two months after the submission deadline. That day, Cristian Ferrer [mozaic] (2020b), Research Manager at Facebook, posted a new thread revealing the winners: 1. Selim Seferbekov (US$ 500k); 2. \WM/ (US$ 300k); 3. NtechLab (US$ 100k); 4. Eighteen years old (US$ 60k); 5. The Medics (US$ 50K). Below the results, an explanatory note: "You may notice that the top rankings have changed. Unfortunately, the top two teams in the preliminary standings used external data sources in their winning submissions that were not allowed under the rules of this competition" (n.p.). Strong reactions against Kaggle and Facebook followed with downvotes and questions about the competition's integrity, particularly regarding the participants' usage and ownership of external data, inadvertently providing us with a glimpse of the actual values of this competition.

The previous top-ranked team, "All Faces Are Real," led by the Grandmaster Giba, had manually created a face image dataset from YouTube videos and Flickr-Faces-HQ Dataset with a CC-BY license, which explicitly allows for commercial use. Not only did the competition rules permit the use of external datasets, but this strategy was encouraged by DFDC as a way to enrich the original dataset and make the predictions of fake videos more robust. However, the teams must "ensure the External Data is available to use by all participants of the competition for purposes of the competition at no cost to the other participants" (Deepfake Detection Challenge, 2019b, "7. COMPETITION DATA.", para. 4). "All Faces Are Real" felt that Facebook's decision to remove the team from the competition was arbitrary, arguing that they did not seek to undermine any rules. The team asked why Kaggle never clarified that external data must follow more restrictive regulations. When requested to provide additional permissions or licenses from individuals appearing in the external dataset, however, "All Faces Are Real" replied that since the data was from public datasets, they did not have specific written permission from each individual appearing in them, nor did they have any way of identifying these individuals.

Giba (2020) initiated a thread to discuss the issue and received hundreds of upvotes and supportive messages. Many participants flagged the ambiguity of the rules regarding the use of external data, including the competition's winner, Selim Seferbekov (2020), who posted his concerns: "As external data might be very helpful to obtain better scores on public/private leaderboards how are you going to validate solutions' compliance with the rules?" (n.p.). Kagglers were confused and ambivalent about imposing regulations on datasets. They have used user data to fuel machine learning without consent for years. Should they have to get written and notarized permission from the user in the dataset every time they produce a new model? The community demanded an explanation, but all attempts from Kaggle and Facebook to clarify concerns about external datasets only repeated the competition rules, frustrating the community.

The reason that led to Giba's team being disqualified reveals the double standard regarding ethical concerns about individuals' privacy and touches on questions about user consent and the use of open-source and copyrighted material to produce machine-learning models, particularly the ones related to facial recognition. In the name of rapid technological advancement and quick turnout in profit, these big tech and digital platforms constantly collect data from the Internet without the owner's permission or consent, constituting repetitive unethical behaviour. For instance, OpenAI used terabytes of copyrighted music in their training for OpenAI Jukebox (Dhariwal, 2020). Microsoft practically stole ten years of open-source code posted by tens of thousands of users on Stack Overflow to train a model for GitHub's automatic coding tool (Gershgorn, 2021b). It is also no secret that Facebook, Apple, Google, and Microsoft have used user photos, including selfies, without consent as source training sets for facial recognition algorithms for years (Solon, 2019).

Even with stricter ethical rules, academic research follows the same practice, often incentivized by universities and research centres aiming to compete for talents in the science, technology, engineering, and mathematics (STEM) disciplines and profit from new intellectual property. Indeed, computer-vision researchers building state-of-the-art object-recognition datasets often do not obtain explicit permission to use the dataset or from any individual in particular. For example, in widely used academic resources for object recognition, ImageNet comprises pictures of people and things collected from public websites under license (Quach, 2020b). While its 14 million images were automatically extracted from the Web, the demanding tasks of annotating the dataset were crowdsourced to underpaid anonymous users on Mechanical Turk, making the project the world's largest academic user of Mechanical Turk in 2012 (Markoff, 2012). Led by Fei-Fei Li, Google Cloud's Vice President and former director of the Stanford Artificial Intelligence Laboratory, the project has been instrumental in advancing computer vision and deep learning research.[19] ImageNet is just one of the many datasets of its kind in circulation. Kagglers, in particular, are heavy users of CIFAR, collected by Geoffrey Hinton and his team at the University of Toronto, and COCO (Common Objects in Context),[20] sponsored by Facebook, Google, and Microsoft.

Specific datasets for facial recognition projects focused on public faces and celebrities, such as CelebFaces, with 200,000 faces from 10,000 individuals (Liu et al., 2015), and Microsoft's MS-Celeb (10 million faces from 100,000 individuals), launched in 2015 and 2016, respectively, were also scraped off the Web from search engines and videos shared on social media. Despite the name, these two projects do not exclusively target celebrities but also contain the faces of private individuals, including investigative journalists such as Kim Zetter and Adrian Chen, Shoshana Zuboff, the author of *Surveillance Capitalism*, and Julie Brill, the former FTC commissioner responsible for protecting consumer privacy. After being accused of ethical deviations and violation of the EU's General Data Protection Law, Microsoft retracted the dataset, but not before allowing several commercial organizations, including IBM, Panasonic, Alibaba, Nvidia, Hitachi, SenseTime and Megvii, to download and use the dataset for their own purposes (Murgia, 2019). There are, however, initiatives to combat and expose these types of ethical deviation. For instance, in January 2021, aiming to alert users of unauthorized uses of their images, Adam Harvey and Jules LaPlace (2021) launched Exposing.ai, a Web application to check if an image was used in one of the major machine-learning datasets. Currently, the project has tracked 3.6 million photos from six image datasets used for training, testing, or enhancing face recognition technologies.

Common to most machine-learning datasets freely circulating on the Internet is that they are widely available for non-commercial research purposes only; yet, large companies do not always observe these rules, creating a precedent for others to do the same. In machine-learning communities, such as Kaggle, using unauthorized datasets and personal data is widely accepted, even encouraged, as long as the model is sufficiently transformative, improves efficiency or accuracy, optimizes previous versions, and does not interfere with the sponsors' interests. Big companies can be even more invasive since they have access to their user base data at all times. When was the last time Google, Facebook, or Apple asked for consent before using personal data to improve some new feature? When was the last time these companies sent users an email asking permission to use their personal photo albums, including selfies, to enhance personalized ad

[19] For an in-depth critical analysis of ImageNet, see Denton et al. (2021).
[20] A large-scale object detection, segmentation, and captioning dataset. See cocodataset.org

delivery? For these companies, copyright and privacy are outdated concepts and obstacles to the inevitable "progress" expressed by digital technologies, a discursive trap of technological determinism.

6.3. Modulate: The True Self

With over 35,000 predictive models submitted for consideration, Kaggle and Facebook closed the competition by praising the participants for the high quality of their work. However, while several algorithms achieved above 80% accuracy on the public test set, the metric dropped significantly on the private test set, indicating substantial differences between the videos Facebook carefully created for the competition and deepfakes circulating on the Internet. The winning algorithm, developed by Selim Seferbekov, a machine-learning engineer at Mapbox in Belarus, achieved 65% accuracy against the black box dataset (Ferrer et al., 2020), considered no better than a coin toss (Morse, 2020). Facebook recognized that the challenge did not produce adequate solutions, observing that the results "show that this is still very much an unsolved problem" and that "none of the 2,114 participants, which included leading experts from around the globe, achieved 70 percent average precision on unseen deepfakes in the black box data set" (Ferrer et at., 2020, A challenge and dataset designed, para. 4). That is, not only does Facebook's dataset not reflect all ranges of methods to produce deepfakes, but it also means that machine-learning models trained with the DFDC dataset might not be able to recognize deepfakes made using newer techniques.

As in all Kaggle's competitions, competitors send their best algorithms to be tested against a private dataset. According to Facebook AI Research's team, unlike the public test set, half of the dataset included in the private test set is "organic content found on the Internet, and the other 50% is unseen content from our source video dataset, collected from various sources" (Dolhansky et al., 2020, p. 6). It is unclear from where Facebook took the data indicated as "found on the Internet" and "collected from various sources" or if they have permission to use it, which may cause the company to infringe several copyright laws and privacy rights, including the competition's rules, used to disqualify the first two contenders on the public ranking.

Facebook has a long history of invasive practices, data manipulation, inappropriate use of user data, and lack of transparency (Bucher, 2012; Gillespie, 2017; Hill, 2014; Kramer, 2014). After numerous controversies involving the dissemination of disinformation and attempts to control public opinion—the most famous related to the 2016 U.S. federal election, when Cambridge Analytica exploited Facebook's techno-economy to spread fake news (see Heawood, 2018)—the concerns about the spread of lies and hate speech on the platform only grew. In late 2018, the company admitted that it had helped fuel a genocidal anti-Muslim campaign in Myanmar for several years (Warofka, 2018). When Frances Haugen released the so-called "Facebook papers," it became clear that Facebook only started to take action against Holocaust deniers, anti-vaxxers, and the conspiracy movement QAnon in 2020 (Lima, 2021). However, the company mainly focused its efforts on the Global North, leaving far-right extremists in the Global South free to propagate lies and profit on the platform monetization system (Milmo, 2021). All these dangerous falsehoods were metastasizing thanks to a mix of the company's economic model, and the unleashed AI capabilities to automatize the system for profit. As a result, Facebook struggled to deal with harmful content and misinformation on the platform in the most fundamental way. It does not know how to identify whether the content is real or fake, misinformation or disinformation, organic or fabricated. Furthermore, the company's engineers admitted not knowing what Facebook does with users' data or where it goes (Franceschi-Bicchierai, 2022). Without knowing the nature of the content circulating on the network, Facebook could not only lose its control over the platform but have difficulties making a profit from it.

Threatened with external regulation and pressured to improve its systems of moderation, Facebook experimented with a variety of methods with different levels of interference, from human intervention such as news curation (Isaac, 2019; World Wide Web Foundation, 2019), community-driven flagging, and an army of low-paid moderators around the world (Gillespie, 2018), to algorithmic recommendations (Bucher, 2012). As much as Facebook relies on a shadow workforce to moderate (Roberts, 2019), it is increasingly betting on automation and algorithmic-driven moderation systems (Gorwa et al., 2020). However, the algorithms that underpin Facebook's business are not created to identify and filter out what is fake or inflammatory. Instead, they are designed to make people engage with and share as much content as possible by showing them things they will most likely be outraged by. Driven by Mark Zuckerberg's well-known relentless desire for economic growth, the company once again decided not to get too involved with its main controversies, diverting to a new and even more extreme issue: deepfakes.

Indeed, the choice to crack deepfakes was odd since the company was not having problems with synthetic images on the platform (Brandom, 2019b) but rather with the exploitation of its system to target users with manipulated content for political reasons. With DFDC, Facebook saw an opportunity to respond to the accusation of not doing enough to moderate harmful content by proposing a machine-learning competition to debunk deepfakes. In a public relations stunt, the company brought the deepfake controversy to Kaggle to try a new approach to designing its algorithmic regulation. Facebook knew from the beginning that detecting deepfake was a difficult task. Mike Schroepfer admitted that the AI Research team had worked on similar issues for months with little success (Vincent, 2020). The company was not expecting significant improvement in identifying deepfakes but was interested in using Kaggle and the crowdsourced workforce as a testbed for its curated dataset. Unsurprisingly, Facebook's engineers dismissed the DFDC's solutions altogether (Kahn, 2020). Partly to prevent people from figuring out how to trick the system, the decision reveals that the company was not planning to put an end to deepfakes; it was simply leveraging the machine-learning community to improve its own image recognition algorithms to incorporate deepfake detection techniques into its business model.

6.3.1. More and Deeper Fakes

While imperfect, the DFDC results could be used as a deterring tool to prevent the less technically skilled from trying to spread deepfakes on the Web. Social media platforms could use deepfake detection technology to scan uploaded content at scale and automatically flag synthetic videos for human review. Furthermore, PAI recommended that "the ability to detect synthetic media should extend to journalists, fact-checkers, and those working in civil society" (Partnership on AI, 2020, p. 9). However, detecting synthetic media alone is not enough. Simply identifying fake content does not increase the perceived legitimacy of content or establish the truth. As PAI's final report about the challenge concludes, "technical detection is one important aspect of addressing information integrity challenges, but it is not a panacea" (p. 9).

Suppose the competition successfully produced an accurate and high-quality deepfake detector algorithm. In that case, it is possible to improve the next generation of deepfakes by using this same detector as the second process in GAN's generator/discriminator pair, essentially designing a new generation of deepfakes specialized in persuading the new detector. This dilemma constitutes a real problem for the artificial intelligence community, which is pragmatically in favour of open-source code and development: once a new technique is developed, a collection of papers, data, and models is rapidly and widely dispersed across the field (Engler, 2019). Indeed, DFDC's rules required that the top five models be made public after receiving the money prize. More importantly, perhaps, participants were allowed and encouraged to publicly share their solutions, methods, and models after the end of the competition. Each incremental improvement makes deepfakes easier to create and more difficult to spot. For instance, in 2018, researchers pointed out that deepfake faces do not typically blink (Jung et al., 2020; Li et al., 2018). As soon as the research was published, an algorithm addressing the problem was openly distributed in forums and machine-learning competition platforms such as Kaggle (see Li, 2020).

From porn to politics to business, six years after Samantha Cole's (2018) exposé and less than four after the largest deepfake detection competition promoted by Facebook, the use of deepfakes has become more common than ever but no less abusive. Several services, apps, videos, and pre-training models are available on the Internet and in app stores to assist anyone who wants to create deepfakes. To name one example, Deepfakes Web (2022) provides an easy-to-use cloud service for deepfake creation in four steps: "Step 1: Upload Your Source & Target Videos; Step 2: Let the AI Learn and Render your Video; Step 3: Download or Watch your Newly Created Deepfake Video; Step 4: Reuse your model" (n.p.). Costing US$3 per hour, the website states that, on average, a short deepfake can be created in 5 to 20 hours, depending on the quality.

Conservative and far-right political parties around the globe are adopting such technology to fool voters and spread disinformation. For instance, in South Korea's 2022 presidential election, the front-runner candidate courted younger voters with a friendly version of the now-president Yoon Suk-yeol (Shin & Yi, 2020). With neatly-combed black hair and a smart suit, his deepfake alter ego looks near-identical to his real counterpart but uses salty language and meme-ready quips to engage younger voters who get their news online. Similarly, Brazil's 2022 presidential election saw the far-right coalition spread disinformation about pooling numbers on WhatsApp and Telegram. Cristina Tardáguila (2022) revealed a deepfake in which the voice of a TV anchor and the pooling graphs were altered to pretend that Jair Bolsonaro was ahead in the race.

Deepfakes can also transparently blend into the fabric of society disguised as "AI for the good"—or as filters of convenience, as I prefer to call it. For instance, the Silicon Valley start-up Sanas has developed a voice-altering tech for call centres that uses deep learning to remove the accents of non-native English speakers in order to avoid discrimination. Using relevant data about the sounds of different accents and dialects and how they correspond to each other, Sanas' AI engine can transform a speaker's accent into what passes for another one that sounds like a white American in real-time. Interviewed by Chan (2022), Sana's CEO Sharath Narayana admits that his technology can harm individuals but argues that people must accept the world as it is; that is, accept their socio-techno-political conditions of existence. The call-centre sector is well known for terrible working conditions and forceful language neutralization, or, to be more specific, the "englishization" of transactional conversation (see Aneesh, 2015; Boussebaa et al., 2014). Artificial neutralizing accents might accelerate the postcolonial economic perspective of big tech companies and endanger cultural diversity and the dignity of the person at the other end of the phone. It creates a deepfake—a shallow truth that people will believe and take as reality itself. Sanas's voice-altering device would undoubtedly clash with some of the experiments in the DFDC, where a user suggested trying to identify deepfakes by reconstructing a face using the voice (Tunguz, 2019b). The idea was deemed unethical and could reinforce stereotypes, yet the post had several upvotes, signalling that Kaggle's community is excited about the possibility.

6.3.2. Uncanny Self

Deepfakes, and the mechanisms developed to detect them, are techniques of and for subjectivation that interpellate individuals' sense of existence. Purposely, they are not necessarily refined tools or a full-fledged strategy to produce subjects, but mechanisms through which one can gain control over an individual's self-image in order to reconstruct their identity either to simulate non-existing behaviours and events or dissimulate the target's character and personality. The fakery may be so uncanny that it can fool us into believing deepfakes are authentic expressions of reality. Take, for instance, Noelle Martin's experience when she saw the video for the first time: "I watched as my eyes connected with the camera, as my own mouth moved. *It was convincing, even to me. Why would people ever think it wasn't real?* [emphasis added]" (as told to Daniella Scott, 2020, para. 4). The image was not perfect, but it was enough to convince others that *she and the fake were the same*.

In fact, the quality of the image—defects at the pixel level, slight changes in colour tone, minor distortions around the edges of a face—does not really matter much in the rapid communication pace that we live in today. We consume an overwhelming amount of information from infinite scrolling newsfeeds provided by social media platforms without even blinking or reflecting on them. Alongside advertisement, propaganda, and fake news, deepfakes might help create a no-trust society where people cannot, or no longer bother to, separate truth from falsehood. Deepfakes allow people to produce recordings of events that never occurred, creating a burden—often painful if it is an attack on an individual—to disprove that evidence. Recall Rana Ayyub (2018). Most people could detect that the woman in the video was not her. But the creation of a collective sense that now lots of people were *in* on something at least close to watching her in a sex act undermined her credibility and integrity. Her dignity as a person was violated and her authority as a journalist was undermined.

This mechanism is part of what Foucault calls a complex set of relationships between subjectivity, truth, and power in Western societies that postulates that one needs, at some point, to tell the truth about oneself to someone else (Lorenzini, 2016, p. 70). Deepfakes force individuals to resist and react against these new mechanisms that are trying to steal their identity. It requires the individuals to present themselves by saying, "I am the real me," and provide evidence of their own existence. By revealing more about their "true" self and reaffirming their own subjectivity, more data about individuals gets publicized. In turn, this data is collected, scrapped, tracked, traced and ingested to become fuel for a new generation of machine-learning applications, as well as a commodity to be sold, exchanged, and exploited by digital platforms for further processing. In this way, the individual not only reclaims their own identity but also becomes a willingly productive subject within the context of a neoliberal digital society.

On the other hand, in order to regain control over the content circulating on the Internet and, most importantly, over the process of subjectivation within digital platforms, technologists were quick to promise a mechanism to reverse engineer deepfakes, taking themselves as the authority to distinguish the real from the fake. Deepfake detectors have become a new method of subjectivation based on normalizing the human face and the individual's identity. At its core, the training sets and the facial recognition algorithms carry a normalized standard of how a human being should look like, including skin tone and facial proportions, what to wear, and how to behave. In turn, the deepfake detectors assign a score or grade to determine how well one corresponds with the pre-established standard. These detectors are designed to be generalizable; that is, they should be valid for any person, independently of race, ethnicity, gender, or age, as well as cultural heritage, class, and context. Depending on the threshold, an individual could pass the test—considered a legitimate self-belonging to a certain society—or fail it, whether because they are invisible to the machine or non-conforming to the standard; that is, a non-existing being.

Following Foucault (1982), the question here is not a skeptical or relativistic refusal of verified truth: whether a visual media is synthetically created (i.e., a fake). What is at play is the way in which knowledge circulates and functions, its relations to power, and the regime of truths. With deepfakes and their detector counterparts, the machine-learning community creates different forms of subjectivity, implying a series of 'practices of freedom' and the inauguration of new ways of life as a way to govern populations. More than just recognizing faces and distinguishing untouched videos from synthetically created media, GANs and machine learning, in general, become powerful tools to identify and classify the conduct of others. For Foucault, "conduct" means both to "lead" according to a mechanism of coercion and a way of behaving within a more or less open field of possibilities. The exercise of this power "consists in guiding the possibility of conduct and putting in order the possible outcome. Basically, power is less a confrontation between two adversaries or the linking of one to the other than a question of government" (Foucault, 1982, p. 789). As such, deepfakes and their detectors control what can be said and who can say anything about an individual. This allows for total control over the self and its condition of existence.

6.3.3. Ontogenesis

Facebook's Deepfake Detection Challenge is evidence that the importance of machine learning goes beyond the collective effort to solve puzzles, or the prizes distributed. Although one of the largest prizes offered on Kaggle, it was a mere fraction of the actual spending in the AI industry. The value of the competition lies neither in the dataset, the largest yet on Kaggle, or in the algorithm that produces the best predictions. Rather, the mobilization of hundreds of competitors, millions of images, and thousands of optimized algorithms for object detection, classification, and recognition serve as instruments of power to define the ontologies of our world. In other words, it is an attempt to answer "what is it?" questions: what is real and what is fake, what is true and what is false, what is a cat and what is not, what is a face, what is there to be known and what is simply noise.

Following Gabrys (2016), I argue that, in the case of facial recognition, detection-driven algorithms go beyond fitting objects into pre-predefined ontologies, working more as ontogenesis. The processes of operationalizing these objects put dynamic attributes into play rather than simply writing a script against which a workflow is executed. That is, machine-learning algorithms create ontological categories based on patterns and clusters, on similarities and distinctions of data attributes. The algorithm runs multiple times—millions of times if needed. Each cycle builds upon the previous one, making tweaks, tiny modifications, and incremental adjustments to produce a model *of* the dataset and *for* the dataset. Suppose the results are not accurate enough, or the expected categories are not met. In that case, the algorithm is tweaked once again, the dataset is extended a little more, and the process is restarted. As Gabrys (2016) describes so well in her book *Program Earth*, this perspective allows Facebook and other large companies to control newly created ontologies, deciding what falls into which category, making the world more manipulable and "programmable."

The ability to produce and control ontologies goes hand in hand with the essential technical features of machine-learning techniques, always involving pattern recognition and clustering of similar attributes. It is also the first step in exploring large datasets on different domains, which, not surprisingly, account for almost half of Kaggle's competitions offering money prizes. The "*what is it?*" question is present in all the challenges aiming to detect, identify, classify, or extract objects and subjects, working analogously to Facebook's Deepfake detection challenge. For instance, in 2012, the cybersecurity company Impermium (2012) offered US$10,000 and a job interview for whoever developed an algorithm for "Detecting Insults in Social Commentary;" in 2018, Google Jigsaw (2018; 2019) launched the "Toxic Comment Classification Challenge," and in 2019 the spin-off "Unintended Bias in Toxicity Classification"; Kaggle (2020d) promoted a similar challenge named "Tweet Sentiment Extraction" in 2020. The main questions these challenges pose can be reformulated as: What is an insult? What is a toxic comment? What is sentiment?

While words can be easier to work with, both in technical and conceptual terms, images, on the other hand, are more abstract and polysemic, carrying wider ranges of contextual and cultural meaning (Barthes, 1977) and opening broader possibilities for ontogenesis and other processes of reinterpretation and subjectivation. In 2015, the U.S. National Oceanic and Atmospheric Administration (NOAA, 2016) offered US$10,000 for a "Right Whale Recognition" algorithm to identify endangered right whales in aerial photographs; in 2020, the University of Saskatchewan (2020) asked for some help to identify wheat heads using image analysis on its US$15,000 "Global Wheat Detection" challenge; the following year, Makerere University (2021) launched a similar challenge for "Cassava Leaf Disease Classification." The questions these competitions are trying to answer are: What are the visual attributes of a right whale as seen from above? What does a wheat head look like? Or what is cassava, and what does a disease that attacks its leaves look like? But also what is *not* cassava; or a wheat head; or a right whale. Even more open-ended questions were asked by Google AI's (2018a; 2018b; 2019) team in its 2018 and 2019 contest "AI Open Images," where competitors had to create algorithms to detect, segment, and establish relations between objects and subjects in varied and complex images; and by Lyft (2019) on its "3D Object Detection for Autonomous Vehicles," all of them using large datasets with images of urban settings and everyday life. The main questions at play for Google and Lyft on this mode of data mobilization can be as broad as What is and what is not an object? In Lyft's case, which focused on self-driving cars, the questions could be recast as *What is and what is not a human being*.

6.4. Summary

Media technologies have continuously operated at the epistemological and cognitive levels as modes of knowledge production. They also act directly on the individuals' body and sense of self, generating and modulating individual and collective affective moods and structures of feeling among assemblages of humans and non-humans (Grusin, 2015). From "shallow fakes" (Johnson, 2019) to "cheapfakes" (Paris & Donovan, 2019) to deepfakes, synthetic media can divest one's sense of self, gain control over their subjectivity, and manipulate their behaviour. Following Foucault (1980), this new technique of governmentality produces a regime of truth in which individuals willingly help and accept the terms of their role in the world.

Over the past two decades, millions of people have been lured to upload their everyday lives to the digital realm: names, pictures, memories, moving images, thoughts, desires, and fantasies. Versions of ourselves piled up in an out-of-reach database while digital platforms expanded their ability to track real-time personal events from all kinds of sensors. Following Srnicek (2016), personal data has not only become the new "gold" in the current economic climate, but it is also the fuel that feeds machine-learning algorithms in their endless quest to find useful patterns. High volumes of (big) data, unavailable to Bledsoe in the 1960s, now circulate "freely" on the Web, where anyone with some access to computing power and little coding knowledge is able to create their own AI friend capable of recognizing faces. Face recognition has become a security feature of choice to unlock phones, access user accounts, and confirm payments. It promises to revolutionize the way we use technology by making it more convenient, speeding up the diagnosis of certain illnesses, and, more importantly for the current political economic agenda, enhancing targeted advertising businesses. Yet, it is also increasingly becoming a tool of state oppression and corporate surveillance.

With the development of GANs in 2014, algorithms jumped from passive detection and recognition tasks to a more active role in generating synthetic representations of the self—faces, expressions, gestures, and voices. Adversarial in nature, GANs are used to confront reality with newly created synthetic content in every iteration: realistic simulations of alternative truth that bear no relation to any reality whatsoever. However, the invention catalyzing the already accelerated development of automated machines got out of control in less than four years, hitting the world with so-called deepfakes. In its first iteration, deepfakes went from pixelated porn fantasies with swapping celebrities' faces (Cole, 2018) to low-resolution clips where actors were inserted into films they never made and words never spoken by politicians and their mouths (Haysom, 2018;

BuzzFeedVideo, 2018), to compelling exposés discrediting and attacking women's dignity by swapping their faces into porn videos (Ayyub, 2018; Hao, 2021; Scott, 2020), to widespread attempts to manipulate elections and the democratic processes with credible fake material (Shin & Yi, 2020; Tardáguila, 2022). As Kerner and Risse (2021) put it, deepfakes, alongside other synthetic media and fake news, might help create a no-trust society in which people cannot, or no longer bother to, separate truth from falsehood, and no reliable media could help them do so.

The increased deepfake radicalization got tangled up with other no less severe issues caused by and as a consequence of digital platforms and social media, such as content moderation, hate speech, fake news, and disinformation. DFDC was both a response against this radicalization and a public relations stunt to regain public trust in order to reclaim control over the content circulating on the Internet. As is the norm in Silicon Valley's big tech mentality, the pool of companies behind DFDC crowdsourced the task to the machine-learning community, leveraging the free labour available on Kaggle. With US$1 million in prizes, the challenge encouraged Kagglers to team up and contribute with "a robust response to the emergent threat deepfakes pose globally" (Deepfake Detection Challenge, 2019a). Contrary to the initial rules defined by DFDC, but following the traditional practice in machine learning, Kagglers used everything at their disposal to make incremental adjustments on deepfake detects, including using data lacking user content and well-known unbalanced and biased datasets. The lack of diversity impacted not only the creation of the dataset provided by DFDC (Xu et al., 2022) but also the methods to identify deepfakes. When measuring the fairness indicators, such as age, gender, and apparent skin type, the top five DFDC winners' methods to detect deepfakes are biased toward lighter skin tones since they mostly fail on darker-skinned subjects (Hazirbas et al., 2022).

The competition was a kind of madness, following Louise Amoore (2020), that rendered unthinkable the political and instead focused on rules and optimalities. There is a widespread consensus that technology will never solve the deepfake challenge by itself since most of the issues are non-technological. However, big tech's myopic position that technology alone would solve social problems is so pedantic that Mike Schroepfer (2019), Facebook's Chief Technology Officer, only consulted with academics from Engineering departments in the United States and UK institutions about the DFDC project: all of them white males. No other institution from different fields of knowledge or non-English speaking countries was invited. This is very problematic since the Machine Learning field and IT, in general, are dominated by certain segments of society and frequently only ask questions about data that reflect their own experiences. Moreover, data collection occurs where technology is oftentimes available through devices that certain segments of society own more commonly than others. Furthermore, the data itself reflects what are often racist trajectories. As a result, the prejudicial structures of the past might end up not only shaping the future but optimizing it to keep the current social and political norms intact.

Indeed, deepfaking is an important and challenging issue. Like other types of harmful content, it is adversarial in nature and will continue to evolve. No single organization can solve such challenges on its own. Perhaps more importantly, in time, the technology behind deepfakes is likely to have implications far beyond memes, porn, and disinformation campaigns. It can work as a mechanism of impersonal subjectivation (Langlois & Elmer, 2019), where distributed automated machines attain control over an individual's self-image in order to reconstruct their identity either to simulate non-existing behaviours and events or dissimulate the target's character and personality. Our subjectivity is deconstructed and rebuilt again and again in different shapes, flavours and formats, including in possible erotic fabrication—in which case, to use Samantha Cole's (2018) terms, "we are truly fucked." By creating new versions of the self and new identities, deepfakes, and, more generally, object detection and facial recognition enabled by machine-learning algorithms bring about new ways of life, not only opening the possibility for different living experiences but also as a way to guide the behaviour of populations allowing for greater control over the self. Moreover, the business logic of machine-learning development pushes the algorithms to accelerate the economy, producing a specific condition of existence where all the actors involved—humans, non-humans, and machines—play a role (Lazzarato, 2004). It uses all its available resources to create a world in which its products and services, together with the consumer and the worker, can co-exist, and, at the same time, the world is, in turn, deeply inscribed in their body and mind.

This chapter illustrated how machine learning has been used as a mechanism to identify individuals, classify their facial features, know who they are, and rank them to establish their very existence. The value of identifying things and subjects in the world lies, however, not in the massive dataset itself nor in the algorithm to traverse it. Instead, the value lies at the intersection of both (code+data), producing predictive models to define what is identifiable and classifiable. The ontologies automatically derived from these algorithms, often trained on unauthorized, racist, sexist, and poorly diversified datasets, *define the things*—objects, subjects—that matter in the machinery world; everything else is just noise. The following chapters will trace different machine-learning challenges to uncover how the same sociotechnical strategies have also been used to predict individual behaviour, as well as to recommend or nudge them in specific directions.

7. Predict: Ab/Normal Bodies and Speculative Futures

> As I stand in the security line and draw closer to the millimeter wave scanning machine, my stress levels begin to rise … My heartbeat speeds up slightly as I near the end of the line, because I know that I'm almost certainly about to be subject to an embarrassing, uncomfortable, and perhaps even humiliating search by a TSA officer, after my body is flagged as anomalous by the millimeter wave scanner. … As I expected, bright fluorescent yellow blocks on the diagram highlight my chest and groin areas. You see, when I entered the scanner, the TSA operator on the other side was prompted by the UI to select 'Male' or 'Female.' … If the agent selects 'male,' my breasts are large enough, statistically speaking, in comparison to the normative 'male' body-shape construct in the database, to trigger an anomalous warning and a highlight around my chest area. If they select 'female,' my groin area deviates enough from the statistical 'female' norm to trigger the risk alert, and bright yellow pixels highlight my groin, as visible on the flat panel display. In other words, I can't win. I'm sure to be marked as 'risky,' and that will trigger an escalation to the next level in the TSA security protocol. (Costanza-Chock, 2018, pp. 2-3)

Sasha Costanza-Chock's body was virtually undressed by a full-body scanner searching for "abnormal" or threatening objects, such as knives, guns, and plastic explosives, that may have been concealed from sight. Sasha describes herself as a nonbinary, transgender, femme-presenting person. Like many minorities, she was targeted by a security procedure based on a long-established pre-conception of a "normal" human body. Transgender individuals, and the LGBT+ community in general, have always been discriminated against and persecuted due to the non-conformity of their bodies. Traditionally seen as a pathology, an abnormality to be corrected within a binary system of sex/gender, their bodies re-emerge as an indicator of a potential security hazard under the scrutiny of the full-body scanner (Amir & Kotef, 2018).

Dorian Wanzer, a black woman from Washington, D.C, has a similar experience. Every time she goes through an airport body scanner, the machine flags her as a potential threat due to some "abnormalities" in her hair. "It happens with my natural Afro, when I have braids or two-strand twists," she explains, wondering if "is this for security, or am I being profiled for my race?" (Medina & Frank, 2019, para. 2). Black women have historically faced discrimination, whether they wear their hair in its natural state or wear hairstyles that are primarily associated with black culture, like braids, two-strand twists, cornrows, and locks. Natural black hair has been policed for ages: it is considered unprofessional, unsanitary, and radical. Nowadays, it is also perceived as a signal for potential criminal activities.

This chapter further explores how machine-learning algorithms are configured to predict threats based on a supposed "abnormality" found in non-conforming bodies. Unsurprisingly, full-body scanners disproportionally target minority groups, considered "abnormal" by the underlying dataset used to train the embedded recognition system. Data-driven surveillance relies on supposedly objective accounts of the world. The full-body scanner and the facial recognition sensors, in particular, are based on the notion of a statistically calculated normality to aid decision-making. While "normal" in this context supposedly represents the universality of a given phenomenon, these systems ultimately reproduce categories aligned with social norms. Despite being historically controversial (Cuff, n.d.), the discriminatory presupposition of intentionality, culpability, and "abnormality" based on anthropometrics is deliberately circumscribed in the internal logic of sensors like the airport scanner. As a result, people whose bodies and behaviours deviate from measured standards of normality and fall into categories of social abnormality (such as gender non-conforming individuals, the mentally and physically disabled, or ethnic minorities) re-emerge in such settings as suspected terrorists and potential criminals.

Digital sensors are merely the point of contact between the individual and a much larger infrastructure that capitalizes upon the data stream absorbed and tracked by digital platforms. Data processing shifted away from being situated in the direct mediation experienced by individuals toward distributed off-sites controlled at a distance by a variety of interested parties like media companies, data brokers, research centres, and data science competitions. This involves multi-sited processes, whereby the dynamics of our relationship are mobilized and mined by a diverse body of social and economic actors and institutions to fulfill a specific political agenda (Langlois & Elmer, 2019). As a result, building algorithms for predictive models often generates ontological and epistemological conflicts. In the data science community, these predictions put forward the idea that statistical probability alone can explain any kind of event, feeding the fantasy that, with enough data, they can forecast future events based on historical trends (Hong, 2020). In this cultural imaginary, the world is an extensive, messy archive, and the machines of today are up to the task of cataloguing it and predicting what will happen tomorrow.

As a competition modality on Kaggle, "predict" has to do with forecasting future events, trends, and behaviours. It also refers to the output of a machine-learning algorithm trained on a curated dataset applied to new data when forecasting the likelihood of a particular outcome, such as whether a video of a person is a deepfake, the chance of an individual is carrying a weapon under their clothes, or the probability of a customer cancelling a service in the next 30 days. Predictions serve as a benchmark for the other machine-learning modalities discussed in this dissertation (identify and recommend), assigning mathematical truth to each model regardless of its purposes. What interests me here is how data is mobilized not only to forecast future events based on curated and fragmented accounts of events but, more importantly, to identify, recognize, and predict individual behaviour; that is, to produce subjects based on statistical models that might not even correspond to the individual, what Langlois and Elmer (2019) term "impersonal subjectivation." Consider, for instance, what is at play when dealing with airport security, particularly when we must pass through a full-body scanner: What do these scanners capture and measure? What can these measurements tell us about an individual, and how do they relate to airplane security? What are the baselines and the standards of comparison? More importantly, what do human measurements and proportion have to do with behaviour and prediction of future acts of terrorism?

Drawing from the specific vantage point of an airport full-body scanner, this chapter discusses how machine learning has been used to decontextualize and project individuals' identities into speculative futures as a way to predict their behaviour. The effectiveness of airport screening technologies can be measured in terms of their capability to detect threats while minimizing false alarms accurately. While false alarm rates are easily measured in operational settings, detection rates cannot be precisely known because of uncertainty over what may have gone undetected. For instance, in 2011, ProPublica revealed an alarmingly high number of false positives in European airports: 23% in Italy and 54% in Germany (Grabell & Salewski, 2011). The machine "saw" potentially threatening individuals carrying a weapon that later was attributed to body lotion, sweat, sudden subtle movements, or buttons and folds in clothing (Russell, 2013). The U.S.

Department of Homeland Security (DHS) acknowledged that the technology's high false alarm rates were creating significant bottlenecks at airport checkpoints: whenever the TSA's sensors and algorithms predict a potential threat, TSA staff must engage in a secondary, manual screening process that slows everything down (Fortune, 2018). In 2017, the DHS sponsored the Passenger Screening Algorithm Challenge on Kaggle. It was not only an attempt to improve the machine learning embedded in these machines but an example to show how these presumed scientific, measurable, and objective sociotechnical apparatuses are depoliticized, and yet, at the same time, intentionally engineered to reproduce specific political-economic conditions.

As in the previous chapter, here I break down the operations through which individuals' identities are deconstructed (sense), recombined (mobilize), and reshaped (modulate) into speculative subjects. In Sense, I briefly describe how a full-body scanner works both as a passive sensorial apparatus built to detect extraneous shapes in the human body and as an instrument able to (re)produce the dichotomy between what is "normal" and what is "abnormal." While these scanners are used in a variety of contexts, notably in the medical field, where they can serve to detect diseases and fractures, here I focus on the securitization of abnormality, where they are used to detect threats and for sorting populations, particularly in the context of airport security.

In Mobilize, I discuss how the machine-learning community was incited to work long hours on projects with little to no information in order to produce predictive models for identifying threatening objects and predicting the likelihood of a future event to occur. The Passenger Screening Algorithm Challenge illustrates how the assumptions of training data as the representative of reality, which would reveal some fundamental truth about individuals, are entirely flawed. In particular, I discuss how the competition was organized, the discriminatory compromises the machine learning community is willing to make to have access to specific datasets, and the developers' carelessness in considering the historical, social, and political aspects ingrained in the problem-solving activity they engage with.

Lastly, in Modulate, I consider the outcome of the competition and the shortcomings of machine learning development. More specifically, I discuss the spurious correlation predictive modelling brings about and the consequences of using machine-learning algorithms to produce justifiable certainties that become historical inevitabilities, which in turn reinforce current social, economic, and political norms. I conclude by arguing that the goal of predictive modelling goes far beyond its applicability in airport security, the medical field, or marketing purposes. Instead, these models seek to get as close as possible to answering *what is going to happen next*. That is, to have the ability to *foresee* events with some degree of certainty as a way to *produce specific subjects* that fit particular and well-defined conditions of existence.

7.1. Sense: Ab/normal Bodies

Foucault distinguishes between two types of "normal": one comes from disciplinary apparatuses, such as mental disability, deviant behaviour, or gender non-conformity, while the other emerges from biopolitics. In the first type, normal works to a standard—a preconceived model to which one should conform, such as identity, efficiency, health, and ideology. Foucault (2009) calls "normation" (or normalization) the processes of measuring against this model and adopting subjects, "the normal being precisely that which can conform to this norm, and the abnormal that which is incapable of conforming to the norm" (p. 85). The second type, on the other hand, is extrapolated from the calculated measurement of empirically observed characteristics to produce a body signature (biometrics)—height, weight, the level of pigmentation on the skin, specific skills level, or behavioural and habitual patterns. Here, "normal" constitutes a certain frequency of a trait and its location on a Gaussian curve, presumably precluding any form of bias or judgements and, as such, reflecting the "natural order of things." In principle, this type of "normal" is derived from empirical reality rather than imposed on it. This is precisely the starting point from where machine learning models and full body scanners operationalize the normal: a calculation of the frequency of a given phenomenon inferred from the natural flow of things and living beings, their patterns of movement and modes of action. It is, as Elden (2007) writes, "the means by which the group living beings understood as a population is measured in order to be governed, and tied to the political rationality of liberalism" (p. 573).

These practices began to be sketched in the eighteenth century, circumscribed by the political effect of economic accumulation and the demographic upswing in Western Europe. The necessity to manage economic growth and preservation of power led to the invention of mechanisms of control, making the population, with its multidimensional numerical variables of space, time, longevity, and health, emerge not only as a problem to be dealt with but as an object of surveillance, intervention, and modification. At the time, the technologies of population focused on large-scale societal problems, such as demographic estimates, studies of social and economic relations, and the development of new forms of education and professional training. The body—of an individual—increasingly became the object of scrutiny. As Foucault (1980) noted, "the biological traits of a population become relevant factors for economic management, and it becomes necessary to organise around them an apparatus which will ensure not only their subjection but the constant increase of their utility" (p. 172). Essential to a discussion of the securitization of abnormality by full-body scanners, I focus here on a specific biological metrics: anthropometric measurements.

Simply put, anthropometry is the systematic measuring of the physical characteristics of the human body, particularly in terms of dimensions that describe body size and shape, such as height, weight, and the circumference of the hips, waist, chest, arms, and legs (Rumbo-Rodríguez et al., 2021). These measurements can indirectly predict body composition, such as body fat indices. Anthropometry can use these metrics to produce correlations of historical social and economic conditions, such as nutrition levels and a population's wealth. For instance, physical stature is commonly associated with a country's net nutritional status (Cuff, n.d.) and can predict how well individuals thrive in their socioeconomic environment (Komlos, 1992). While Anthropometry has a controversial history—serving as the base theory for eugenics, for example (Cuff, n.d.)—today it plays an essential role in the health sciences, particularly in clinical nutrition, for assessing growth, body composition, response to treatments, and predicting health risks, but also in industrial design, fashion, ergonomics, fitness and wellness, and national security, where statistical data about the human body dimensions in the population are used to optimize products, customize experiences, and identify potential threatening individuals. However, changes in lifestyle, nutrition, and ethnic composition of populations can lead to changes in the distribution of body dimensions (e.g., the rise in obesity). Wang et al. (2000) suggest that most anthropometric variables vary by age, gender, and ethnicity, as well as by geographical location and year of measurement, requiring regular updating of data collection.

Traditionally, anthropometric measurements are manually collected using a tape or calliper on the surface of the human body, which, by its nature, is time-consuming and involves physical contact with the individual being measured. Automation only emerged in the mid-1980s when the University of Loughborough developed a fully automated system capable of accurately, quickly, and comprehensively measuring the size and shape of the human body (Jones et al., 1989). The company wished to explore the possibility of developing a "non-contact machine that is reasonably transportable and sufficiently speedy in operation to survey economically a large sample of the British population" (Heymsfield et al., 2018, p. 680). What emerged in 1987 was the Loughborough Anthropometric Shadow Scanner (LASS). The equipment was specifically designed to measure a body in terms of radii and angles in conjunction with height. It included a television camera, projector, and a 360° rotating table upon which the volunteer stood during the evaluation

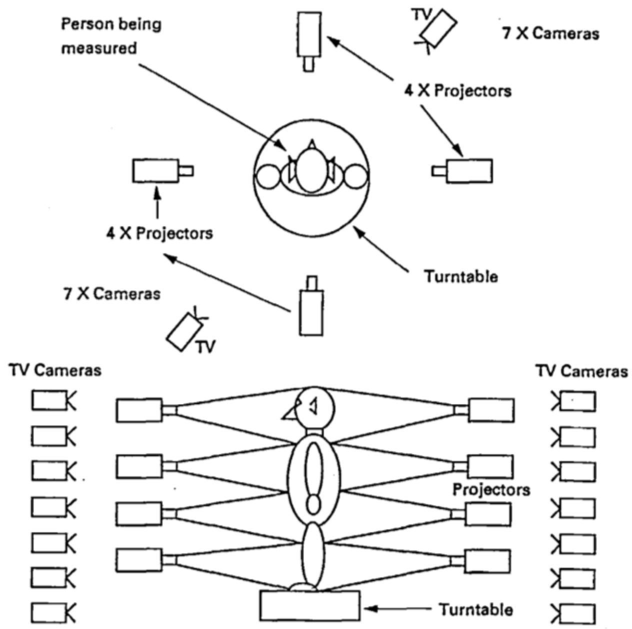


Figure 7.1: Schematic of a Body Measurement Detector. In LASS, the individual is placed on a turntable and rotated 360 degrees in small increments while the camera captures the slit of lights from the projectors. All points where the edge of the light falls on the body define the horizontal radii at points in the vertical plane (Jones et al., 1989, p. 165).

procedure (see Figure 7.1). The scanner was first used to aid the UK-based retail company Mark & Spencer in manufacturing more comfortable bras for women and, subsequently, in partnership with Computer Design Incorporated (CDI), to provide more accurate and speedy measurements for textile designers. While the first full-body 3D scanner was developed to fulfill clothing industry marketing strategies by providing a means for economically measuring the human body shape, LASS was more than just a more accurate and fancier ruler: researchers at Loughborough gave the machine the ability to "see" under the human skin. The inventors eagerly presented an instrument that measures the thickness of the subcutaneous adipose tissue and calculates its "true" density from the body's volume (Jones et al., 1989, p. 164). As such, the development of 3D scanners has introduced opportunities for measuring the human body more efficiently.

Since then, several extensive 3D anthropometric surveys have been conducted, such as SizeUK, SizeUSA, and the Civilian American and European Surface Anthropometry Resource (CAESAR) (Liu, 2017). Three-dimensional machine vision rapidly advanced in methods designed to quantify human body shape, including structured light systems and millimetre wave radar methods. Structured light systems utilize controlled visible or infrared illumination patterns projected across the imaging field. One or more cameras measure deformations in the light pattern over objects in the scene. This deformation can be used to calculate per-pixel distances between the camera and the object, creating a depth image using geometric triangulation (Heymsfield et al., 2018). Though the levels of precision and resolution are not accurate enough to produce reliable data, these systems are inexpensive, making them a good fit for entertainment and consumer-level products, such as Microsoft's Kinect and Apple's facial recognition sensors. Millimetre wave radar methods, on the other hand, are more accurate and expensive machines and are only used in specific industrial and security applications, such as military weaponry or as attack deterrents in airports.

7.1.1. Virtual Strip Search

After the 9/11 incidents, and in direct response to an increased perception of terrorist threats from abroad, the DHS decided to enhance security in U.S. airports. It ordered the Transportation Security Administration (TSA) to prioritize developing, testing, improving, and deploying airport checkpoint screening technologies to detect non-metallic, chemical, biological, and radiological weapons and explosives on passengers and in carry-on items. One of the most impactful and controversial measures was the adoption of Three-Dimensional Millimeter-Wave Imaging for Concealed Weapon Detection, also known as full-body scanners. While the technology behind these machines had been extensively developed during the 1990s at the Pacific Northwest National Laboratory exclusively to be employed in airport security, the system was not ready for large-scale real-world operations due to its low resolution, slow processing times, and high cost (Sheen et al., 2001). In 2007, despite the lower confidence in the system, the TSA began using full-body scanners, referred to as Advanced Imaging Technology (AIT) systems, in American airports as a secondary screening method for selected individuals (Elias, 2012). However, an incident in 2009 accelerated the adoption: Umar Farouk Abdulmutallab tried to blow up Northwest Airlines Flight 253 on Christmas Day using plastic explosives hidden in his underwear (Harris, 2011). While the security agencies' officials admitted that operational scanners would not have aided in detecting Abdulmutallab's hidden explosives (Elias, 2012), the TSA and some foreign aviation agencies decided to make the full-body scanner the primary security device at airport checkpoints.

We have grown accustomed to seeing these Three-Dimensional Millimeter-Wave scanners in operation in airports. If you travel through any large international airport, you probably have to go through them repeatedly: step in and raise your arms, place your hands in a triangle shape, palms

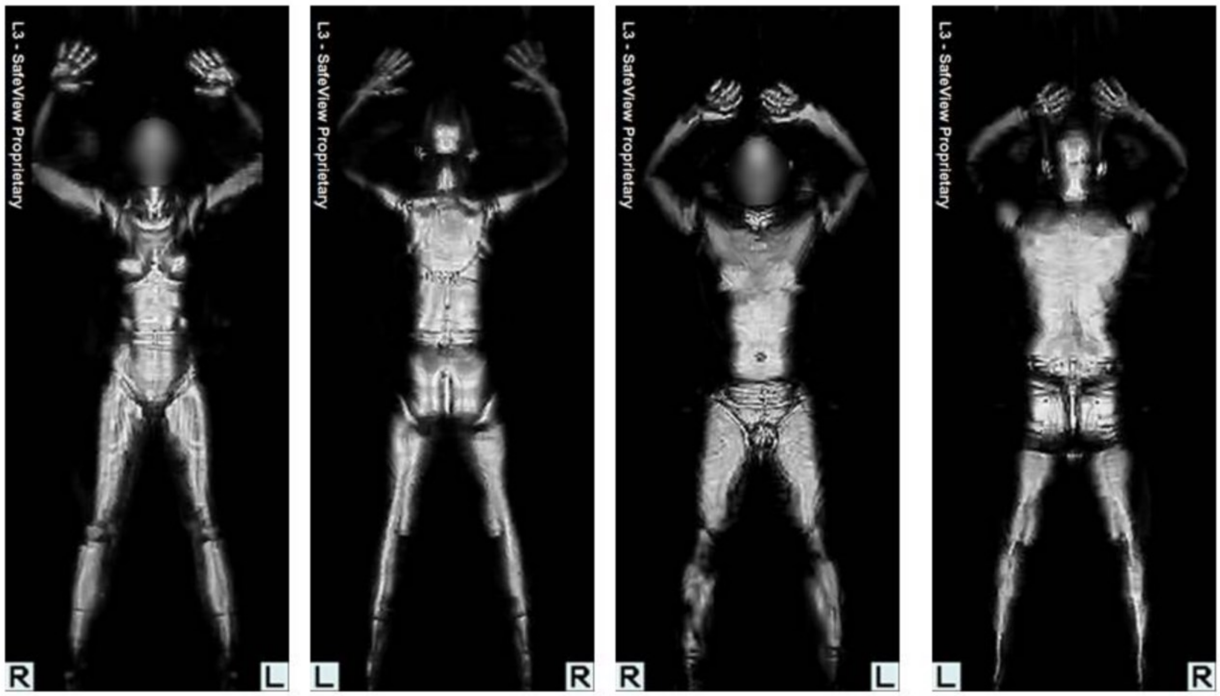


Figure 7.2: Full Body 3D Scanner Outputs Real Bodies. Typical Millimeter Wave Images of a Female (Left) and Male (Right) with Facial Blurring Applied (Amir & Kotef, 2018, p. 240).

facing forward above your head while listening to the machine emitting a quick sound as it rotates the camera around you. These machines emit non-ionizing electromagnetic radio waves in the millimetre wave within the 30–300-GHz frequency band to render images of what lies directly under the clothing and near the skin (Sheen et al., 2001). The scanner generates images that look like photographic negatives (see Figure 7.2), revealing one's naked body to a security agent a few metres away, where they can check if you are carrying a weapon or represent a danger to the flight. Moreover, these machines are capable of "reconstructing" the human body as a 3D model, scanning the full 360 degrees around an individual. This 3D scan renders front and back images that are then assessed for threats by trained TSA screeners (Elias, 2012). The images produced by these scanners were quickly considered too invasive, intensifying what Fuster et al. (2015) called "Nurturing Ob-Scene Politics." Situated in the intersection of In/Visibility and Dis-Appearance, these scanners are designed to expose what we are socialized to keep private: the body under our clothes.

Their introduction into airports prompted strong public objection, including claims that this scanning is equivalent to a "virtual strip search" and that the machines show extraordinary disregard for individual and collective privacy rights. To solve these issues, in 2013 the TSA bought new scanners with image-analyzing and recognition software provided by L-3 Communications,

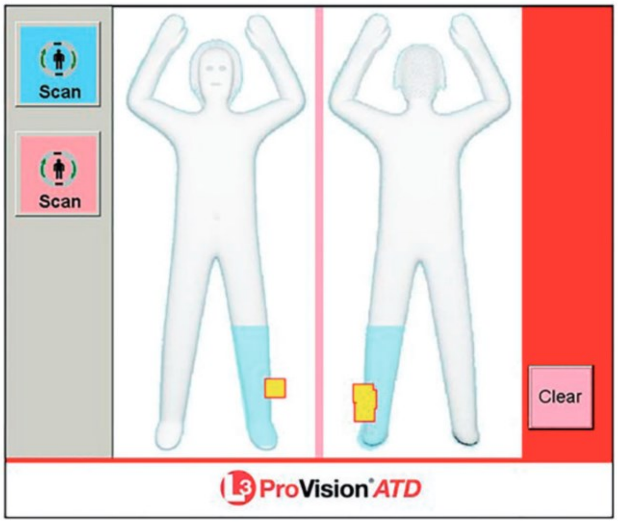


Figure 7.3: Full Body 3D Scanner Outputs Illustration. The output generated by L3 ProVision shows the region where the threatening object was detected (Leidos, n.d., p. 3).

later acquired by Leidos (2020a). Based on machine-learning algorithms, the Automated Target Recognition (ATR) system analyzes the image by looking for "abnormal" patterns of shapes to identify possible threats. The reconstructed 3D image is never shown to the machine operator nor permanently stored (Elias, 2012). Instead, the scanner only outputs an abstracted human figure, on which the location of suspected regions on the passenger's body is marked (see Figure 7.3). According to Leidos (n.d.), the scanning process takes only 1.5 seconds, after which the operator "is able to locate many metallic and non-metallic threats and contraband—both low and high-density materials … a quick scan provides security personnel with data they can interpret to aid in passing an individual through a checkpoint with confidence" (p. 2). The machine is designed to identify concealed weaponry and contraband objects under people's clothing and inside body orifices. Using millimetre waves, the full-body scanner produces an image of the scanned entity, which is then analyzed by the ATR algorithm. Former L-3 Communications President Tom Ripp has stated that the algorithm is designed to "look for abnormalities" and detect "objects that are not supposed to be there" (Grabell & Salewski, 2011, Guns, Sweat and Privacy Fears, para. 10).

While other scanners, such as X-rays, are designed to detect differences in the types of materials being scanned based on their relative penetrability, millimetre waves cannot "see" through thick objects and body tissue. As a result, most objects and body parts may appear indistinguishable in the output image. To distinguish between what is and is not "supposed to be there," the scanner relies on the shapes of the reproduced figure. That is, the ATR algorithm depends on a previous understanding of what the human body looks like and the relationship between its parts, which is predetermined by anthropometric measurements, such as overall body shape and the average height, size, length, weight, and other metrics. The standards produced by Anthropometry are extrapolations based on aggregated data used to create statistically computed configurations, that is, the norm. Passengers' bodies are then compared to these normalized configurations of the human body. By abstracting away the body, the scanner—like a photography camera—brings a sense of "mechanical objectivity" (Benjamin, 2008), resting on the presumption that their operation is necessarily neutral and impartial. As such, manufacturers advertise their full-body scanners as eliminating discrimination since they do not register pigment density or racial characteristics and are "colour blind" (Leidos, n.d.).

7.1.2. Preemptive Anticipation

From this perspective, scanners might be seen as a success story in which state-of-the-art technology offers solutions to human concerns and limitations related to the intersection of security and privacy. However, they also bring additional constraints and presuppositions that, instead of eliminating discriminatory practices, further exacerbate target profiling. Once the ATR algorithms were integrated into full-body scanners, it very quickly became apparent that particular groups of people were being singled out, disproportionally affecting minorities: gender non-conforming individuals such as trans persons, individuals with particular hairstyles (such as "dreadlocks" or "sisterlocks"), people wearing certain religious clothing and accessories (such as Sikh turbans), or passengers carrying medical appliances on their bodies (such as people with a stoma). This is precisely the type of experience Costanza-Chock (2018) was referring to when she had to go through a full-body scanner. While the scanner is equipped with an ATR algorithm, the TSA operator must manually select the gender of the person getting into the scanner. The choice is binary ('male' or 'female'), based on how the person looks—or presents themselves—from the operator's perspective on the other side of the machine. Furthermore, depending on how often ATR algorithms are trained, the scanner might identify body parts, clothing, or accessories as threats, flagging the individual as a potential suspect of a potential crime.

Beyond identifying mere "abnormalities," full-body scanners double as predictive sensors. To view a device as a sensor is to approach it from a particular angle: to determine what information it automatically detects and how it can be put to use. These scanners/sensors are built to detect and recognize "normal" human bodies, which correlates with "normal" patterns of behaviour (i.e., passenger vs. terrorist). With the normal baseline built into the algorithmic analysis, these systems trigger alerts when identifying people who deviate from it. The rationale is that people who do not have ‘normal’ bodies may not be regular passengers and, therefore, pose a security threat. This alert is then used to anticipate the risks before the individual can take any action. That is, the objective of full-body scanners is to model an individual’s future behaviour based on impersonal and generic data patterns. Amir and Kotef (2018) call this processes the securitization of abnormalities.

Automatic detection of abnormal behaviour operates by predicting the possible developments resulting from different types of behaviour. Machine-learning enthusiasts believe that, by observing the characteristic features of a familiar sequence in a particular situation, it is possible to make probable inferences about future developments and intervene to prevent them from ever happening. These predictions are, however, modelled on aggregated data about human bodies and behaviour (including cultural artifacts, political interests, and social conditions) that might bear no relation to the individual being identified, creating a paradoxical effect in which the anomalies are detected simply because it is not codified in the initial training set. That is, any behaviour that diverges from the web of habitual activities may indicate a threat simply because the device/sensor has never experienced it before. As a result, technology itself, sensor and machine learning in particular, not only reshapes modes of governmentality (Foucault, 2009) but also reconfigures subjects based on the quantification of bodies.

Solving the “body shape” problem becomes a technological problem: “All comes down to ‘machine learning,’” observers former L-3 Communications President Tom Rip (Grabell & Salewski, 2011, Possible Solutions, para. 6). In fact, improving the technology to increase detection but limit false alarms has been an issue since the initial development of the full-body Millimeter Wave scanner in the Pacific Northwest National Laboratory. Getting the information of what is “normal” to improve the technology requires many thousands of scans, and developing predictive algorithms using machine learning can be very expensive. Without real-world data, a massive technological infrastructure and a large team of developers, TSA must find other ways to teach the software to distinguish real threats from false ones. The following section discusses how crowdsourced labour is mobilized to further develop predictive algorithms for automatic threat recognition based on discriminatory practices and assumptions that curated data about human bodies can anticipate human behaviour and distinguish “regular passengers” from “terrorists.”

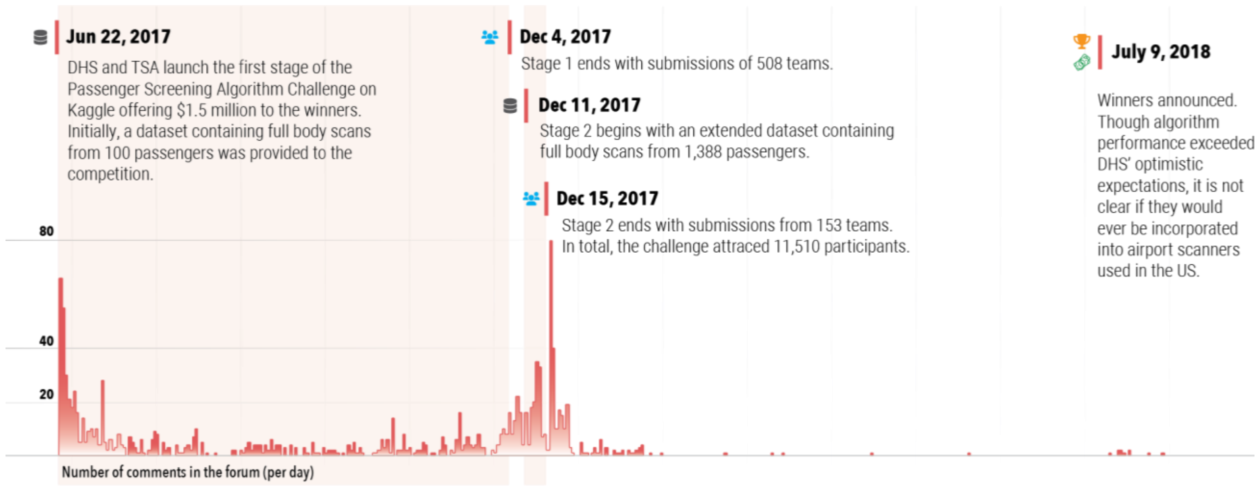


Figure 7.4: Passenger Screening Algorithm Challenge Timeline. Timeline showing the competition milestones and number of comments in the forum through time. Designed by L. Frizzera.

7.2. Mobilize: Passenger Screening Algorithm Challenge

Following advice from the Aviation Security Advisory Committee (2017) to "foster collaboration among industry and incentivize competition, including promoting integrated solutions to improve the overall performance of checkpoints" (p. 20), the DHS launched the "Passenger Screening Algorithm Challenge" on Kaggle. The competition ran as a partnership between Kaggle and the DHS's Apex Screening at Speed Program, aiming to experiment with a crowdsource development model as a way to improve the efficiency of full-body scanner recognition software. The sense of urgency and severity of the problem was evident on the competition's overview page: "Even a modest decrease in false alarms will help TSA significantly improve the passenger experience while maintaining high levels of security" (DHS, 2017, n.p.). To ensure that top machine learning developers would participate in this challenge, the DHS offered eight prizes for the best teams, totalling US$1.5 million, the second highest in Kaggle's history. The competition was planned to take place in two phases. Stage one began on June 22, 2017, and comprised a five-month period where participants explored a small section of the dataset with scans of 100 passengers to build their predictive models. Stage two reset the competition ranking and extended the dataset to 1,388 passengers; it began on December 11 and gave participants four days to submit a newer and improved predictive model using "unseen individuals" present in the extended version of the dataset (see Figure 7.4).

The challenge asked participants to *find* useful data, *filter* the noise out, and *predict* the likelihood of an attack in mid-flight. According to one of the participant teams, the ideal is to reduce the noise and clearly show that the person in the image is or is not a threat. The approach should be methodical, clearly deconstructing and dehumanizing the passenger to "produce an algorithm that fractionated the human body image in regions to be able to identify the body's region correctly … identify the true or false of the detection, for measuring their accuracy, recall and precision" (Guimaraes & Tofighi, 2018, p. 1). According to these remarks, what is at play in this competition compounds two different modalities of machine learning: identify and predict. That is, the goal is to *identify* abnormal shapes adjoined to the individual bodies that could be seen as threatening objects, such as knives, guns, or anything that might be used as a weapon, and also *predict* the likelihood that these anomalies are, in fact, threatening objects. In other words, by taking apart individual bodies in thin slices, the model being pursued not only aims to re/construct and sort subjects based on the shape of their bodies (normal vs. abnormal) but also to mark any "strange" or "abnormal" shape or attribute as a possible threat to national security, which in turn assigns intentionalities and behaviours to individuals that do not conform to the norm.

7.2.1. Thin Human Slices

The competition was designed to capture real-world conditions. However, similar to the Deepfake Detection Challenge discussed in the previous chapter, the DHS could not share actual passengers' body scans or any other personal data with the competition participants for privacy reasons. To mitigate this problem, the DHS recruited volunteers to pass through the scanner and release their privacy rights over the data collected. The dataset made available to the competition consisted of data from scanned images of 1,000 male and female volunteers with different levels of body mass indices wearing a variety of clothing types. Some of these volunteers had several possible threatening objects adjoined to their bodies. The DHS did not disclose how it selected these volunteers or what possible threatening objects they were carrying. The competition website explains that "due to restrictions on revealing the types of threats for which the TSA screens, the threats in the competition images are 'inert' objects with varying material properties. These materials were carefully chosen to simulate real threats" (DHS, 2017, Data, para. 2). Moreover, the dataset does not contain actual scanned imagery but the necessary data (tables and vectors) to reconstruct these images and generate three-dimensional models of each individual. To protect the volunteers' privacy, Kaggle and the DHS asked participants not to share the data or any composite image derived from the dataset, which includes not posting images on the competition's forum or

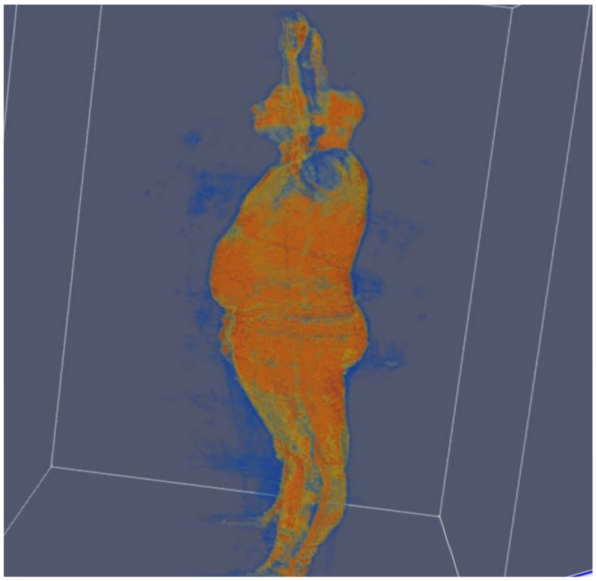


Figure 7.5: 3D Rendering of Full Body Scanner Data. Three-dimensional image of one of the volunteers taken by a participant in the competition. Screenshot by L. Frizzera.

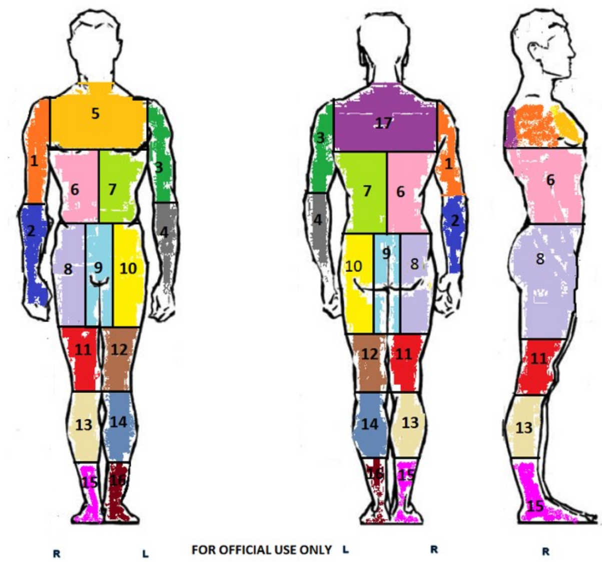


Figure 7.6: Body Sections Diagram. Schematics of full body scans split into 17 areas of interest provided with the dataset (DHS, 2017).

any other website. At the end of the competition, participants were instructed to refrain from publishing the dataset in other repositories and to delete the dataset from their machines. However, at least one user[21] posted a video on a GitHub repository showing a 3D-rendered image from one of the volunteers (see Figure 7.5).

For this competition, the DHS asked volunteers to pass through the High-Definition Advanced Imaging Technology, a new generation of millimetre wave scanners. The images were further processed and split into 17 zones (see Figure 7.6). The predictive model must indicate the probability of contraband within each of these zones. The dataset for the first stage had more than three terabytes and contained scans of 100 individuals, each with 17 different images representing each zone (1,700 scans in total). Each scan provided consists of four binary files that represent the data in different ways: (1) A calibrated object raw data file; (2) a projected image angle sequence file consisting of 3D images for 90-degree segments of data that, when played back plane-by-plane, makes it seem as if the object were spinning on the screen; (3) a combined image 3D file, showing a composite 3D volumetric image that when played back plane-by-plane displays cross-section slices through the individual at sequential heights; and (4) a combined image angle sequence file, similar

[21] To protect users' privacy, their names and GitHub repository were omitted.

to the projected image angle described in (2). The dataset used in the second stage of the competition was 14 times larger, containing scans of 1,388 individuals—almost 24,000 scans in total. The DHS explained that the volunteers' scans used in the first and second stages of the competition were different, and participants should not make assumptions about the number, distribution, or location of threats in the second stage.

The sheer volume of data made Kaggle refrain from storing the dataset on its own servers, restricting the challenge to users with access to high-end computers, large storage capacity, and high-speed Internet. For instance, Hans Bouwmeester (2017) reported that it took up to five days (100-125 hours) to train a model using his "water-cooled high-end computer with 4 GPUs" (n.p.). However, the competition began only three months after Alphabet acquired Kaggle. Google was eager to advertise its Cloud services and promptly made the full dataset available on multi-regional Google Cloud Storage. Furthermore, Google offered Cloud Credits to assist with participants' storage and computation needs for the Passenger Screening Algorithm Challenge: "The first 500 eligible entrants will receive US$500 in Google Cloud Credits each!" Kaggle's staff Addison Howard (2017) shouted out in the forum. Though participants were happy with this unexpected "gift," some users, particularly those in China, complained that they could not access Google Cloud in their countries, cutting them off from the competition (pym1993, 2017).

7.2.2. Discriminatory Compromises

Even before Kagglers started virtually "dissecting" human bodies, a controversial decision on who could participate in the competition revealed some of the political and economic interests at play in this event. The day prior to the launch, Will Cukierski (2017a), Kaggle's Head of Competitions, welcomed the participants with a caveat on the rules: "Members of the Kaggle community who are not United States Citizens or legal permanent residents at the time of entry are allowed to participate but are not eligible to win prizes" (para. 3). The community quickly pointed out the contradiction: a platform that praises meritocracy and promotes a culture of open data and international cooperation is now actively limiting participation based on citizenship. Ironically, it was not the first time that Kaggle, an American company founded by an Australian, got caught discriminating against participants from other countries. A few months earlier, in May 2017, the US$1.2-million competition sponsored by Zillow initially prohibited Chinese users from participating in the second round of the challenge. Zillow was concerned about acquiring intellectual property rights in China but did not care about copyright infringements on massive datasets and source codes used in the competition. The restriction was eventually lifted after a chorus of Kagglers voiced their disapproval (Chowdhury, 2017).

In the Passenger Screening Algorithm Challenge, however, the restrictions were maintained. The cause was a U.S. piece of legislation ironically known as the America COMPETES Act that bars the government from awarding federal prize money to anyone who is not an American citizen or permanent resident. Users expressed outrage at being invited to contribute to U.S. national security without having a fair shot at the money prize (Hsu, 2017). Vladimir Iglovikov, Senior Data Scientist at TrueAccord, voiced his frustrations: "Hosting open competitions in which people are differentiated by anything except their skill is unacceptable and it is definitely against the spirit of the community" (Chowdhury, 2017, Why is Kaggle's competition, para. 2). The dissatisfaction was also mirrored in the official Kaggle forum. Cukierski's (2017a) "Welcome Announcement" post received more than 250 downvotes, while Mushinskiy's (2017) thread, titled "This is insane discrimination and [an] insult to our international community," received 963 upvotes. To circumvent the restriction, non-U.S. users began seeking partnerships with U.S. citizens in order to be eligible for the prize, as illustrated by this user in India: "looking for US Citizens for forming a Team so that we shall start working seriously on this challenge!" (GOKAGGLERS, 2017, n.p.). However, Grandmaster Poplavskiy (2017), an important member of the community, observed that the rule would make "top performing teams rejecting new member purely based on his/her citizenship to avoid losing prize eligibility, or member from US not willing to join non US team" (n.p.). Indeed, the rules clearly stated that "if a team has one or more members who are not prize eligible, then the entire team is not prize eligible" (DHS, 2017, Rules), creating a geopolitical divide in the platform. The shock and disappointment led many Kagglers to boycott the competition, as they felt it defied the platform's broader mission of crowdsourcing important data-science solutions to transcend national borders.

In an attempt to explain the reasons that led Kaggle to host the competition under such strict rules, Kaggle's CEO, Anthony Goldbloom (2017b), offers a glimpse of what was at stake:

> If we don't host, then we don't expose our community to one of the most interesting and valuable datasets that's ever been hosted on Kaggle. As we've seen over the past five years with the rise of deep neural networks, the availability of datasets allows us to push forward to [sic] science of machine learning. Not exposing this dataset to our community meant depriving our community, and the machine learning world more generally, of a chance to push forward the science with a novel and challenging dataset. (para. 4)

Goldbloom preaches the unavoidability of data-driven technological determinism, defining the community's right to access datasets as the keystone for machine learning development and, unequivocally, the company's right to profit from it. However, according to the competition rules, "these images are for competition participants only. You should not share the images publicly, plan to use them in future research, or blog about them" (Cukierski, 2017a, n.p.). Moreover, "Participants must delete all Data provided by the Competition Sponsor from their computer within 30 days of the contest submission end date and provide a written acknowledgement of its destruction to Kaggle" (DHS, 2017, Rules). Without access to the dataset after the competition, Goldbloom's justifications for restricting participation based on citizenship in order to "push forward the science with a novel and challenging dataset" falls apart.

The community pressured Kaggle to find alternatives: Some users suggested Kaggle co-sponsor the competition by offering prizes exclusive for non-U.S. users (EduardoLi, 2017); others proposed dropping the money prize and competing solely for points and badges (NxGTR, 2017). In his statement, however, Goldbloom clearly refused these alternatives, indicating that the company was making discriminatory compromises for the "sake of a dataset." The community accused Kaggle of creating second-class users, promoting a different kind of competition, as well as reinforcing the lack of diversity and inclusion on the field, a widespread practice in the AI industry. Like many other instances of machine learning development, Kaggle was not only seeking to profit from the conditions imposed by the DHS, but was also pushing its users to compete for free, something that had never happened before in the platform. Kagglers argue for open-source, multi-cultural and worldwide data sciences but want to be compensated for their efforts. For this reason, they believe that Kaggle should not host competitions that discriminate against competitors because of origin, religion, age, sex, or disability.

More than just a platform for data science competitions, the community defined Kaggle as a place to learn and have fun and, more importantly, contribute to science without consideration of gender, nationality or ethnicity. In fact, as discussed in chapters four and five, this is how Kaggle advertises itself. It manufactures a sense of belonging, which the users absorb as a sense of duty and purpose. For instance, one of the participants wrote in the forum that "the motivation for every competition is knowledge and fun. Money is an optional bonus … This is [a] different environment. This is a scientific community, and science is above all of this" (Iglovikov, 2017, n.p.). This sense of belonging to something greater than oneself, mixed with the highly competitive machine learning market, makes participants accept arbitrary rules even if these rules are discriminatory or unfair. A user from Mozambique, for example, acknowledged he felt unwelcome but would enter the competition anyway for learning purposes (Ragnar, 2017). Another user observed that "'Learning' will always be the first incentive but … almost 80% of the top rankers are non-US Citizens/legal permanent residents, thus it [the restriction to win cash prizes] will lead to the degradation of the competition[,] hence degradation in learning" (Daft Vader, 2017, n.p.).

However, the altruistic and inclusive arguments made in these forums are often superficial and contentious, paying little attention to the socioeconomic conditions of each user and the broader sociotechnical and geopolitical implications of each competition. The community recognizes that having fun and learning on Kaggle is not for everybody: "If you don't have electricity try joining [K]aggle, never mind all the other luxuries that are required just to be able to login" (HedgeHog, 2017, n.p.). The inequalities and precariousness among Kagglers play a significant role in defining their motivations and purpose in each competition. For instance, top leaderboards are often highly skilled expert data scientists looking for complex puzzles to solve. However, half of Kaggle users are students looking for an opportunity to learn new tricks, get a job, and, with some luck, win some cash (Cut Onion, 2017). Moreover, since most Kagglers were not eligible for the Passenger Detection Challenge cash prize—about 88% of them were non-US citizens (Kaggle, 2022a)—their only option was to work for free or walk away.

To overcome these inequalities and bring back a more collaborative approach to the competition, a Kaggle Master suggested sharing the prize among all participants: "There [is] no team size limit, so let's say there will be 150 participants in the end, organize youself [sic] in the top 8 teams and everybody predict average [sic] target. That way each person win[s] 10000 $" (Lam Dang, 2017, np.). However, his post was downvoted as the community saw it as a direct attack on Kaggle's meritocratic system. Why should the winner split the prizes with other less capable participants? How does sharing the prizes create better predictive models? For Kaggle and its community, without "a proper" financial incentive, there would not be a competitive environment to move machine learning and data science forward. Furthermore, it would undermine the fundamental logic on which Kaggle was built: crowdsourcing, exploiting, and profiting over the free labour of thousands of individuals. In other words, the power imbalance (i.e., the control over data, infrastructure, and technology) is part of a "game" that is intrinsic to the platform that, in turn, produces the subjects participating in these competitions.

Take, for example, what the Defence Science and Technology Laboratory (DSTL) wrote about the value of Kaggle for the scientific field. Six months prior of the Passenger Detection Challenge, DSTL held a competition on Kaggle offering US$100k for an automated algorithmic system to classify features of satellite imagery. The "competition has been a great success," attracting 519 participants who submitted over 5,000 solutions "[equating] to over £2.5m [US$3 million] of research – a seven fold return on the cost of running the competition" (DSTL, 2017, para. 1). By this account, we can infer that DSTL spent over US$400k to run the event: US$300k went to Kaggle and US$100k to the top three competitors. The "success" implied by the institution has little to do with the solutions provided by Kagglers, which does not produce any particular breakthrough, and more to do with the reduction of labour costs and the reorganization of the relationship between who owns the data and the infrastructure, and who actually does the job. With a crowdsource model, Kaggle provides the platform, infrastructure, and a committed community of developers eager to work for free for a chance to win modest money prizes, which, by the very competitive nature of this arrangement, increases the number of "solutions" submitted to the competition. Given that the Passenger Detection Challenge prize was 15 times higher than the DSTL competition, Goldbloom could not afford to lose the opportunity to make a high profit despite the discriminatory contingencies present in the competition rules.

It did not take long for some Kagglers to question how far the restrictions would go and the implications of making full-body scanners more accurate. For a brief moment, the unusual set of rules adopted by Kaggle revealed a very common and vital canon of Neoliberalism: free circulation of products and restriction of people's movement across national borders (Harvey, 2012). Phunter (2017) asked Kaggle's staff: "If an American team uses some code made by non-american[s], is this team eligible for the prize?" (n.p.). Triskelion (2017) echoed a similar question: "Are Americans allowed to use non-American software? If not, that's fair? [sic]" (n.p.). Grandmaster Uthman (2017a) pointed out that the U.S. government paid an Israeli firm to hack the iPhone: "why they can't pay an international citizen that helps bolster their security such as in this competition, or finds 0-day vulnerabilities in their software[?]" (n.p.). Uthman was referring to Cellebrite, a provider of mobile forensic software that helped the FBI's investigation of San Bernardino shooters in 2016 (Reuter, 2016).

Most of the discussion in the forum, however, was motivated by a rule restricting non-U.S. users from receiving cash prizes in this competition. In over 1,200 replies to 147 threads on the forum, not once did the community raise similar questions about diversity, discrimination, privacy, or fairness on the dataset and algorithms used to produce predictive models. Quite the opposite: most Kagglers were so deeply invested in improving the accuracy of their models to get to the top of the ranking that they barely devoted any time to understanding the data and tools they so closely employed. One of the few voices urging a more critical approach to this competition, Uthman (2017b) describes how the community should face this problem:

> As data hackers, explorers, engineers, scientists, or whatever, it really is our moral responsibility to see through to the end of our actions. By that I mean, sure, it might be interesting data and in [an] interesting problem; but just looking one step ahead at the immediate goal of wining [sic] a contest, acquiring some fame, coming into some money, etc. when the technology we're working with is exponential in nature, and the organizations we develop the technology for are national, and the implications of our work undoubtedly will effect [affect] hundreds of millions of people, kinda is deserving of at least an introspective discussion, if not a public one. (n.p.)

7.2.3. Mislabeled Threats and Lack of Bodies

Economic advantage is not the only goal behind Kaggle's controversial decisions. Under the surface of a pragmatic profit-driven platform backed by one of the most prominent contemporary tech companies lies a more complex and valuable purpose: access to scarce data, biometrics, in particular, to fuel machine-learning algorithms. As Goldbloom recognizes, full-body scanner data is not a common asset and is usually reserved for governmental agencies and the medical field. For Kaggle's CEO, not exposing the dataset to the data science community would compromise the future of machine learning as a field. Goldbloom's remarks advance a vision of historical inevitability in which technology itself can make correct and precise predictions based on curated and pre-selected archives that account for facts. His decision resonates with the community's belief that "raw data" is the source of truth, but this data must be cleaned, organized, described, and optimized in order to be used "correctly." As such, for Kaggle, a loosely connected community of machine-learning developers ought to be the right (if not the only) place to run these operations and define the correct way to use the data.

Every competition starts with an Exploratory Data Analysis (EDA), where participants post Notebooks with alternated step-by-step code, output, and comments showing how to load the dataset, perform simple statistical analysis, and produce graphs, visualization, and image rendering to familiarize themselves with the data. Due to the restrictions imposed on sharing images derived from the dataset, all graphs and 3D renderings were removed from user posts and comments. Yet, the participants' written analysis and comments reveal details about the collection of scans provided by the DHS. For instance, the images generated by the scanner were taken in regular intervals of 22.5-degree rotation, 360 degrees around the subject scanned (Farrar, 2017); most passengers have similar body shapes, though slightly skewed to overweight individuals (Branden, 2017; DiscipleOfScience, 2017a). More importantly, Kagglers pointed out two main challenges while working with the data. Since the beginning of the competition, it became apparent that not only was the dataset poorly labelled (Trott, 2017b), but it was also not a good sample—that is, there was not enough data—for training machine-learning algorithms able to produce a good predictive model (Farrar, 2017). Let me consider these two issues in turn.

When identifying the presence of threatening objects, Oleg Trott (2017b), fifth place in this competition, estimated that 1-3% of the image scans were mislabeled. If Trott's estimation is correct, there were between 13-40 individuals either marked as "clear" but carrying a weapon (false negative) or as a suspect but not carrying any threats (false positive). In recent studies, MIT estimated a similar or higher number of labelling mistakes on the most commonly used machine learning datasets, including canonical and benchmark datasets such as ImageNet (3.4%), CIFAR (6%), and QuickDraw (10%) (Northcutt et al., 2022; 2021). Handling missing values or dealing with incorrect labels are some of the most common challenges data analysts face. I touched on this topic in chapter four, where I briefly discussed the issues of data annotation made through systems like Mechanical Turk and similar services that leverage underpaid workforce to tag images. Though the number of mistakes may be small compared with other datasets, Kagglers agreed that it would create distortions and even hurt their predictive models. Kaggle acknowledged the problem but clarified that the labels are those provided by the DHS, completely unaltered.

According to the platform policy, Kaggle would not correct or "update the dataset unless there was a systematic problem. Doing so would generate widespread confusion and spin off an endless stream of dataset versions as errors are uncovered" (Cukierski, 2017b, n.p.). Participants were allowed to correct the training set themselves if they wished, but not in the test set. However, DiscipleOfScience (2017a) pointed out that "if we correct labeling, then [we] will get a worse score. Since it only matters if our predictions match what they have; even though [we] will have better detection in general" (n.p.). The competitors themselves dismiss this rather severe problem, arguing that the mislabels would not disturb the competition since "everyone is in the same boat" (Trott, 2017b, n.p.). Once again, Kaggle and its community fell short of improving how machine learning and data science are practiced. With no further consideration of how this issue would affect the results while scanning real people, the community's shortsightedness demonstrated that they were only concerned with the competition's internal affairs. Strictly speaking, the community would take the problem seriously only if there were any unfair advantage that might harm their chances of winning the money prizes. Annoyed, DiscipleOfScience (2017b) protested: "In a sense, we are training models to make [the] same mistakes" (n.p.).

When Farrar (2017) posted his EDA, he warned the other competitors about another issue: "With just under 1200 examples, we'll need to figure out how to make or find more" (n.p.). The "threatening object" in the training dataset is not very diverse either. Trott (2017a) commented in another thread: "It's impossible to think, in advance, of all possible types of shapes and materials someone might try to smuggle, let alone include such contraband in the dataset" (n.p.). If there were not enough examples—a small number of passengers, low diversity of human bodies, a limited number of threatening objects—on the training set to make good predictions on the test set, how were participants supposed to produce a predictive model able to identify abnormalities and threats in the wild? The solution was to augment the dataset by "fabricating" new passengers or using external resources to expand the number and range of bodies the machine-learning algorithm would ingest.

Some users exploit the symmetry of the human body to produce mirrored clones of the original passengers. Thornton (2017) explained that "by switching and horizontally flipping certain images, I could use the same crops for right and left forearm, right and left shin" (n.p.). Other participants manually created thousands of "fake" scans with minor variations based on the original files using vector graphic tools (transforming and transplanting a threat object from one location to another or from one scan to another). idle_speculation (2017), for instance, was able to add 800 new passengers to his dataset, increasing the diversity and number of examples to be used in training his predictive model.

The majority of competitors, however, decided to use external tools, datasets, and presets to extract "signals" from and make sense of the "raw data" they had in front of them. Every major competition had a thread in the forum where participants post the tools and sources of external data they planned to use to solve the problem. This thread highlighted resources—techniques, algorithms, pre-trained models, and datasets—being considered to optimize the results. In this respect, the Passenger Screening Algorithm Challenge is no different from other high-profile competitions. The thread is cluttered with repeated mentions of pre-trained image classification using Convolutional Neural Networks (CNN) and Residual Neural Networks (ResNet) based on canonical datasets, such as ImageNet and CIFAR (discussed in chapter three). More interesting for this competition, however, is the listing of more specialized anthropometric data collections, such as the Max Planck

Institute for Informatics (MPII) Human Shape and Human Pose Dataset, and the DeepMind Kinetics, which Kagglers assume would be of great help in expanding the dataset and optimizing their algorithms.

However, these datasets were built with different purposes, sometimes with incompatible methodologies and objectives. For instance, MPII Human Shape is a family of expressive 3D human body shape models based on the CAESAR, the largest commercially available statistical body representation scan database to date (Pishchulin et al., 2017). On behalf of the North Atlantic Treaty Organization (NATO) and executed by the American Society of Automotive Engineers (SAE) in 2000, the CAESAR project includes a database with seventy-three anthropometry landmarks extracted from the 3D scans of 5,000 people (men and women) between the ages of 18 and 65, exclusively from the United States and Europe (Robinette et al., 2002; SAE, 2004). While MPII Human Shape only accounts for a single pose, typically representing a standing person with resting arms, MPII Human Pose is a benchmark for evaluating articulated human pose estimation. The dataset includes around 25,000 images of over 40,000 people performing 410 different activities with rich annotated body joints, body part occlusions, and head orientations. The images were extracted from YouTube videos without any consent using an established taxonomy of everyday human activities (Andriluka et al., 2014). Similarly, DeepMind Kinetics (2017) is a collection of 650,000 annotated high-quality video clips that cover up to 700 human action classes, such as dyeing eyebrows, square dancing, calculating, fixing a bicycle, and opening coconuts. The video clips were searched on and collected from YouTube, using only four languages (English, French, Spanish, and Portuguese), from a highly disproportionate geographic distribution: North America (56.6%), Europe (19.5%), Asia (11.7%), Latin America (7.7%), Oceania (3.5%), and Africa (1%) (Smaira et al., 2020).

These resources are used in the competition to establish common grounds that define the shape of the human body. The algorithm must first understand what a “normal” and harmless human body looks like on the training set before finding anomalies and signals of threat. However, Kagglers take these datasets as a pure reflection of reality; the ground “truth” from which they would build their algorithms. There are no discussions, warnings, or special considerations relating to these resources to be found in the forum. For a platform that self-proclaims to be the world’s largest data science community, its practices have little scientific rigour. Kagglers use these tools and datasets indiscriminately, taking for granted what each one represents, how they came into being, and what type of bodies they favour. Not only are these datasets removed from their context, but the body is also removed from the individual, serving as a "raw" resource to be ingested by machine learning regardless of any deeper consideration. The participants do not even consider the social, cultural, and political implications of including one or many of these datasets in their predictive model, not even that they are working with data of real individuals like themselves. The exhaustion of datasets to produce the winning algorithms is the only goal here.

7.3. Modulate: Spurious Correlations

Despite the controversial rules, an exceptionally high number of participants (11,510 users self-organized in 518 teams) accepted the terms and registered in the competition (Fortune, 2018). However, only 149 teams submitted their predictive models for the second stage, indicating that most competitors registered for the event only to explore and download the dataset. In contrast with the strict terms defining who could win cash prizes, Kaggle had little control over who could access the dataset. While the only requirement to register in the competition was to "promise" to delete the data by the end of the event, it was unclear how to enforce the DHS's request for the complete removal of the dataset from users' computers and personal cloud infrastructure. Kaggle staff were unprepared for such a demand for data privacy and protection. The alternative was to have users post a non-binding written acknowledgement on the competition forum. Of 11,510 participants, 56 replied to the Andrew Long (2017) thread with the same message: "I've deleted the data provided by the Competition Sponsor." Neither Kaggle nor the DHS can guarantee that the data was not shared or stored elsewhere, showing indifference and a lack of concern for the privacy of those who volunteered to have their body scanned for this competition.

A common issue and, to some extent, regular practice on Kaggle, is the extensive over-reliance on pre-trained models and external databases, as well as the lack of proper understanding of how external resources affect predictive model. Take, for instance, how Moejoe (2017), tenth place in the competition, described his use of the popular ResNet-50 CNN pre-trained with more than a million images from the ImageNet database: "the CNN learns on its own which locations correspond to which labels without any human guidance. … I*t may feel like black magic at first but it works and it works pretty well* [emphasis added]" (n.p.). Yusaku Sako (2017), twenty-second in this competition, had the same impression: "I used various CNNs (VGG, Resnet, Densenet, etc) that have been pretrained with ImageNet … I did not expect this to work too well as ImageNet and TSA images are very different" (n.p.). The top contenders have all used pre-trained deep learning algorithms in combination with large-scale datasets to build their predictive models. The machine learning community is centred on the blind certainty of its methods for risk analysis, such as determining whether a passenger is a threat because they carry strange silhouettes close to their bodies.

However, Kagglers were surprised when they saw their score reduced due to a more diverse set of bodies in the private dataset released in the competition's second stage, which included people with a broader range of heights and weights, using different hairstyles, and wearing a more diversified type of clothes. Their models were well-designed to account for "normal" and "familiar" bodies presented on the initial training set and pre-trained models, overfitting the prediction to a specific type of passenger, triggering what the community dubbed "the Bob Marley lookalike guy problem" (Soleyman, 2017, n.p.). Oleg Trott's (2017c) remarks illustrate his and other participants' reactions to the problem:

> The big surprise on Stage2 was *the guy with massive dreadlocks*, like no one had in the training dataset. *They looked a lot like some of the bombs in the training scans* instead, except for the fact that they adjoined his head. My model was understandably suspicious. [emphases added] (n.p.)

It is not possible to picture what these participants identified as "the Bob Marley lookalike" individual because, as in the first dataset, the dataset for the second stage only contains tables and vectors, was only available for a few days, and only to the teams classified to this phase of the competition. The community was astonished and annoyed by the "intrusive element" that caused their model to fail and considerably reduced their score in the ranking. Requa (2017), for instance, was in the "money zone" (top 8) by the end of the first stage. When his model came across individuals with unfamiliar hairstyles, it sounded the alarm. Their model would make TSA staff engage in a secondary manual screening process since the person would be flagged as a threat. The increase in false positive results made Requa fall to fortieth in the ranking, losing his chance to win some money. The hairstyle was not the only problem. A "guy [wearing] suspenders fucked with my model" (emergent complexity, 2017, n.p.). The suspender metal clip was identified as a threat. The comment was followed with relief by others who faced the same problem; at least they were not alone. More problems emerged: "Height was the most variable feature, so I stretched each person to be the same height" (Kevin H, 2017); "Did anyone use gender recognition to facilitate groin and upperchest [detection]?" (numericLee, 2017); "some subjects had excessive…um…girth so that would require an adjustment across slices" (DavidGbodiOdaibo, 2017, n.p.).

Kagglers' overconfidence in their initial predictive models can be summarized by Sako's (2017) honest comment: "I had strong convictions in my intuitions (that I unfortunately did not bother to validate)" (n.p.). idle_speculation and others acknowledge that their model could produce false positives due to "edge cases" found in the wild. In other words, their algorithm would cause the machine to flag a non-conformed body as a threat. The consensus is that if the "guy with dreadlocks" and the other "problematic passengers" were included in the initial training set, this problem would be solved. Indeed, this is a recurrent justification in the data science community: the model is not incorrect; it just needs more data. Requa (2017) concluded that "Had my model at least seen examples of something similar I think it probably would have scored much higher as it would have learned to whitelist some of those new features in stage 2" (n.p.).

The event was no more than a game for the 518 teams on the Passenger Screening Algorithm Challenge. Their only focus was on getting the highest accuracy rate in identifying "suspicious bodies" through the airports' scans to increase their score on the public ranking. On December 4, 2017, when the first stage ended, the community was already speculating about who would take the big prize. Some participants were betting on Jeremy Walters, also known as "idle_speculation," a data scientist at DataLab USA and a Kaggle Grandmaster, who has won multiple awards at past competitions. However, the public ranking, where the participants can compare and follow other teams' predictions, showed idle_speculation in the uncomfortable 134th place. The competition ended ten days later, on December 15, but the final ranking was only announced six months later, on July 9, 2018. In a brief press release, the DHS (2018) announced the winners, confirming idle_speculation was in first place. "idle_speculation" received US$500k for his contribution, and nine other competitors shared the remaining money prizes: Sergei Fortin, machine-learning Engineer at Apple ($300k); David O. and Thomas A, software engineers and cofounders of Analytical AI ($200k); Zach Teed ($100k); Oleg Trott, a specialist in machine learning in drug discovery ($100k); "CNN is fake model," comprising Halla Yang, Data Strategies at Group Citadel, and Phillip Chilton Adkins, Lead Data Scientist at Grainger ($100k); Sucker Balaji "suchir," a Computer Science student at UC Berkeley currently working at OpenAI ($100k); and Michael Avendi "Kaggle446," a data scientist at Avannos medical ($100k).

The DHS (2018) praised the competition, declaring that "working with algorithm developers to improve screening technologies directly serves S&T's mission to deliver effective and innovative insight, methods and solutions for the critical needs of the Homeland Security Enterprise" (para 2). Furthermore, the DHS stated that

> algorithms developed from this competition have the potential to improve the speed and accuracy of the Advanced Imaging Technology (AIT) scanners used to screen airline passengers for prohibited items. A comprehensive set of new automated detection algorithms has the potential to be integrated into the latest screening equipment. … These algorithms will complement existing systems funded under the DHS S&T Apex Screening at Speed Program, which is pursuing transformative R&D activities that support a future vision for increasing security effectiveness while dramatically reducing wait times and improving the passenger experience. (Para. 3, 14)

While it is not clear whether any of these models were, in some way, incorporated into TSA's full-body scanners, the prospects from the competition boosted the use of predictive models to identify threats. For instance, the second generation of the millimetre wave scanning machine (L3 ProVision 2) "uses machine learning algorithms exclusively, which means human eyes don't evaluate the images, ensuring privacy" (Leidos, 2020b, n.p.). The algorithms used by Leidos, however, continue to produce subjects based on a supposedly universal human body, which, as demonstrated, identifies diversity in human bodies as a threatening anomaly. In 2019, ProPublica reported several cases of discrimination against black women, identified as threats simply because the machine-learning algorithms at the core of the full-body scanner identified something abnormal about their hairstyles (Medina & Frank, 2019)—precisely the problem faced by Kagglers in this competition.

7.3.1. The Fantasy of Epistemic Purity

Full-body scanners are, in essence, biometric instruments for automatically classifying individuals. Unsurprisingly, the methods used by the data science community resemble a century-and-half-old work of British statistician Francis Galton and Italian criminologist Cesare Lombroso to identify criminals. From the 1870s to 1910s, Galton developed a series of scientific photographic experiments layering individual portraits of a given number of people onto a single photographic plate in what he called "composite portraiture." Galton saw this technique as a form of "pictorial statistics" for his anthropometric studies, from which he could deduce measurements, proportions, and averages. Similar to what is said about machine learning and artificial intelligence today, Galton valued photography because of its supposedly objective perspective and "mechanical precision," the same way statistics provide the tools to reach numerical accuracy. The scientific value of his method was the way the composite portraiture made visible the average, or typical, face within a particular type, as well as its normal distribution of differences: the "ghost of a trace of individual peculiarities," as expressed by Cryle and Stephens (2017). Galton, however, did not have access to large-scale image collections to conduct his experiments, mostly carried out with limited amounts of donated material. One of his most famous experiments sought to identify criminal facial features based on a partial and biased collection of pictures of prisoners provided by the Director-General of prisons (Cryle & Stephens, 2017). He aimed to identify the most statistically prevalent type of each category of criminal, although he could only prove the social norm established as the truth at that time: all prisoners are criminals. Anyone who looks like a prisoner, that is, possesses the main facial and body features found on prisoners at that period, must also be a criminal. However, in mixing the portraits of individuals indiscriminately together, Galton produced a "photographic impression of an abstract, statistically defined, and empirically non-existent criminal face" (Sekula, 1986, p. 19). Misled by its own idiosyncrasies, Galton's claim of objective truth produced by a mechanical machine was compromised by manipulating "raw data," in which the incremental exposure of the photographic plate produces not an individual subject but a fabricated ideal type.

Influenced by Galton's ideas, Lombroso (2006) published his infamous *Criminal Man* in 1887. In the book, he outlined his arguments about the inferiority of criminals to "honest people," of women to men, and blacks to whites, thereby reinforcing the prevailing social politics of gender and racial hierarchy. Like in the Passenger Screening Algorithm Challenge, Lombroso's work draws these relationships directly from the physical attributes of the human body: the size of the skull, the proportion of the nose, the body shape, the colour of the skin, and the hairstyle. Galton and Lombroso used pictures to find commonalities—what was always there—among those who were defined as deviants by nineteenth-century society: "thieves are notable for their expressive faces and manual dexterity, small wandering eyes that are often oblique in form, thick and close eyebrows, distorted or squashed noses, thin beards and hair, and sloping foreheads" (Tibbetts & Hemmens, 2009, p. 220). The "intelligent" full-body scanners, on the other hand, use photographic negatives to identify what was "not supposed to be there" when compared with the average body measuments collected from a specific population: individuals with uncommon shapes, thick and braided hair, or wearing odd clothing are suspects of possible future crimes.

The assumption that the training data is representative of reality and would, by itself, reveal some fundamental truth is entirely flawed. Take, for instance, a similar problem involving recognizing significant threats against one's life or private property, also known as predictive policing. Predictive crime models, such as those developed by PredPol and Palantir, are built on historical violent crime data. Resembling Galton's studies, these models portray a specific type of criminal, targeting poor, black, and immigrant communities due to social and political oppression and persecution against these populations in most Western societies (O'Neil, 2016). At the same time, these models purposefully exclude certain types of crimes depending on skin colour, social class, or economic power: for example, white supremacists are not categorized as terrorists; corruption, fraud, and financial offences are deemed less severe than stealing rotten food to feed a hungry family. This creates a pernicious feedback loop, as policing spawns new data, which justifies more policing. The knowledge produced by these models is so powerful that they trick us into believing that if a person lives close by or looks like a criminal, they must be a criminal.

The examples above demonstrate the extent to which biological, mathematical, and cultural epistemologies have intersected and informed one another since the late nineteenth century. On the one hand, Galton's and Lombroso's anthropometric project encouraged a widespread uptake in new self-assessment practices and an increasingly quantified view of self and society. Indeed, the systematic measurement of one's body was particularly influential in the United States in the first half of the twentieth century, later becoming an essential part of social engineering projects led by tech companies in the twenty-first. As Cryle and Stephens (2017) put it, "assessing oneself in relation to statistical norms would become commonplace, and a key cultural mechanism by which to standardise—that is, normalise—one's own body and behaviour" (p. 21). On the other hand, Galton's and Lombroso's positivist methodological attempts to define and regulate social deviance, long discredited as pseudoscience and unethical, regain life among machine-learning communities, as the Passenger Screening Algorithm Challenge shows. In pursuing pure objectivity, they conflate data, facts, information, and knowledge such that they are either seen to naturally follow one another or support a sense of legitimacy (Hong, 2020).

7.3.2. Automatic Projections of the Self

The assertions made by Galton, Lombroso, and by Kaggle's data science community, in particular, are based on the idea that the processes of categorizing, disciplining, and signifying individuals and populations are entangled with the procedures of statistically measuring the human body and behaviour. Indeed, from what can be observed from the messages exchanged in the competition's forum, these measures and predictions have little to do with pre-conceptions regarding the stability and permanence of gender identities, skin colour, or ethnic group: they are aimed at identifying potential terrorists. All judgements regarding the different, the abnormal, or even the deviant are irrelevant, as they are replaced by the seemingly objective questions about the accuracy of bodies to gendered models. The problem is that bodies can only be sorted, measured, and compared after they have been normalized and assigned to categories. Most of the data, including that provided by the TSA, are produced by and for a specific type of society with its own physiognomies and anthropometric averages, with specific social roles based on gender, race, and ethnicity, and with particular ways of dressing and hiding or showing some body parts, which, of course, follows specific cultural values, fashion, and social conventions. One could ask, for example, how a scanner trained with these datasets would react to individuals taller or shorter than average; under- or overweight people; having different facial features; of non-binary gender; carrying religious or cultural symbols; with more or less clothing; with tattoos, piercings, or other body alterations.

The securitization of gender, for instance, is a particular outcome of an algorithm that relies on seemingly objective criteria in which non-conformity bodies become a cause for suspicion. As Costanza-Chock (2018) reported and experienced herself, the composition of the non-normative body itself becomes a subject of scrutiny: “As I expected, bright fluorescent yellow blocks on the diagram highlight my chest and groin areas … I'm sure to be marked as ‘risky,’ and that will trigger an escalation to the next level in the TSA security protocol” (p. 2-3). That is, the composition of the non-normative body itself becomes an “alarm” that triggers further security procedures. Importantly, what is perceived as a marginal effect of a technology that seeks to isolate security threats concealed under people's clothing is an essential and direct outcome of these systems' very logic of operation: a mechanism of subjectivation that projects a distorted and negative sense of self onto the individual been scanned.

A similar effect occurs when examining the category of race and how methods of racial profiling reappear in the workings of the scanner: it is black (women) rather than white (men or women) who are most affected since they are most commonly projected as threatening individuals. The crucial difference between this and “classic” racial profiling is that here, “blackness” appears as a potential security issue, not due to preconceived racist notions of dangerousness or explicit racial biases, but it “automatically” emerges and is made visible through configurations of normality embedded into the predictive models used by the full-body scanner. Racial biases are reconstructed through these technologies since in a society where race appears as an aberration from “whiteness” (Delgado & Stefancic, 2001), they are inseparable from what is programmed into these machines as a normal body or the normal composition of a hairstyle. For instance, people sporting dreadlocks—the “Bob Marley lookalike guy problem” identified by Kagglers—or braided hair, like Dorian Wanzer (Medina & Frank, 2019), are singled out not because they are considered dangerous but because the algorithm associates them as deviations from a particular (white) body composition and a particular Western culture, which, as a consequence, identifying their hair to be “abnormal.”

What is noticeable about the data used in the Passenger Screening Algorithm Challenge is that it cannot do more than sense bodies, which is different from sensing behaviour. However, following the classic anthropometric approach, body data has been mobilized here to predict (deviant) behaviours. The scanner and its predictive models are a form of power that speculatively produces subjects and projects future behaviours. It is a form of power that makes individuals subjects: it "categorizes the individual, marks him by his own individuality, attaches him to his own identity, imposes a law of truth on him which he must recognize and which others have to recognize in him" (Foucault, 1982, p. 781). Subsequently, based on their body signature, it ties these subjects to existing behaviour or actions that may arise in the present or the future. The relationship between bodies, behaviours, and subjects is not direct but constructed based on a specific regime of truth. When blindly relying on established sources of truth, most machine-learning practitioners fail to recognize they have been using partial and biased accounts of history, taking for a fact a specific knowledge produced in a particular time and place, which undoubtedly reinforces the historical accounts of a given event into a prediction of the future.

Kagglers, in particular, honestly believe they can solve complex centuries-old social-political issues using a few lines of code and a handful of well-curated datasets. Nevertheless, even if such algorithms were designed under different sets of assumptions concerning the structure of gender, race, ethnic, or cultural categories, abnormalities of some kind would necessarily still be produced by these technologies and marked as a security problem (be it heart rate, body heat, size, mobility, fashion, religion, etc.). No matter the intention of Kaggle's community, the predictive algorithms entertained in this competition produce a tautological effect that can only forecast what has already happened—the future of the past—(re)producing the same reality over and over again, similar to events in a programming language's infinite loop.

7.3.3. Self-Fulfilling Prophecies

In this context, Kaggle serves as a testbed for machine-learning algorithms. Similar to other modes of competitions previously discussed, the value of machine-learning models lies not in the richness of the dataset, as trumpeted by Anthony Goldbloom, nor in the large sums of money paid to the winners, which is a tiny fraction of what governments and private companies spend on artificial intelligence. Instead, the value of such an automated process lies in telling as much as possible about the individuals' bodies and behaviours in order to predict what they are going to do next. In other words, it is an attempt to answer "*what is going to happen next?*" type of questions: Will this person commit a crime? Are they planning a terrorist attack? In other words, what is at play here is the ability to *foresee* events with some degree of certainty and *produce specific futures* according to certain conditions of existence (Lazzarato, 2004).

The "*what is going to happen next?*" type of question is present in many competitions that attempt to identify, recognize, categorize, predict, or control individuals as subjects. For instance, in 2012, Heritage Provider Network (2013) offered US$3 million and asked participants to "identify patients who will be admitted to a hospital within the next year"; the insurance company Porto Seguro (2017) challenged Kagglers to "Predict if a driver will file an insurance claim"; Home Credit Group (2018) provoked the community: "Can you predict how capable each applicant is of repaying a loan?" While cutting costs and increasing profit are recurrent themes, hyper-profiles can be produced for more exploratory purposes. For instance, the Conference on Document Analysis and Recognition (2013) sponsored a competition to "predict if a handwritten document has been produced by a male or a female writer"; in 2014, the insurance company AXA (2015) promoted a competition challenging participants to "use telematic data to identify a driver signature"; similarly, in 2015 the insurance company State Farm (2016) challenged data scientists to "spot distracted drivers" using computer vision; that same year the marketing agency Drawbridge (2015) promoted a competition to "Identify individual users across their digital devices." These challenges look for prospects for a speculative future, trying to determine which patients will be capable of paying the following year's bills, what type of person is a reckless driver and which driver will file for an insurance claim, or who is the author of a particular document.

The quest for speculative futures is ingrained into the logic of machine-learning challenges hosted by Kaggle. It goes beyond the prediction of human behaviours, also aiming to forecast political decisions and business investments. For instance, the University of Melbourne (2011) offered US$5,000 to "predict the outcome of grant applications" in the university using the institution's historical archive. Not surprisingly, the solution fitted the university's politics to invest in strategic research areas. The Hewlett Foundation (2012) held a US$100k competition to "develop a scoring algorithm for student-written essays." In other words, the challenge aimed to reduce the cost of paid workers involved in correcting essays and, in the process, reduce subjective material into numerical and countable lexicons to *predict and rank students*' *critical thinking* according to the presumable "correctness" of their answers. In 2019, Two Sigma (2019) challenged participants to "use news analytics to predict stock price performance" in its US$100k award competition. The company aimed to understand the predictive power of the news to forecast "financial outcomes and generate significant economic impact all over the world." Reuters, a Canadian-American multinational media conglomerate with stock market stakes, was the competition's primary news source. As a result, any machine-learning model created out of this competition would undoubtedly favour specific players and perpetuate the dominant economic discourse, creating a self-fulfilling prophecy effect (Petalas, van Schie, & Vettehen, 2017). That same year, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE, 2019) provoked Kagglers with a US$25k prize to answer the question "How much energy will a building consume?" ASHRAE justified the competition by its aim to improve building efficiency and reduce costs and emissions, which would help solve climate change. However, believing that climate change would be solved by computation alone, machine-learning practitioners tend to forget that the intense use of computer power also consumes energy without even thinking twice about the current political-economic affairs that drive energy consumption in the first place.

As we can see, these predictions provide no mathematical certainty but a range of possible correlations to legitimize highly anticipatory forms of exploitation, incarceration, surveillance, and subjectivation. Although the speculative nature of machine learning predictions resembles an exercise in numerology, it is precisely rooted in the belief that "raw and untainted data begets not only an excessive reliance on algorithmic factmaking but also extends the older and deeper cultural desire for sorting the world into stable and discrete pieces" (Hong, 2020, p. 21). The trouble is that specific processes of data-driven analytics work within narrowly defined parameters where the input datasets are standardized, modelled, and manipulated. However, machine learning is not a mystical statistical device that predicts the future or foresees what will happen based on historical data patterns. The future is not yet realized to be predicted. Out of the infinite possibilities, what is in place is the *creation* of a specific future, one that replicates a narrow understanding of society and reproduces the same social conditions. These so-called predictions of the future must—because any alternative seems too unbearable for those in power—be the similitude, the mirror without distortion of a patriarchal society led by the world's filthy-rich white-cis-males.

7.4. Summary

In Foucauldian terminology, the Passenger Screening Algorithm Challenge serves as an anecdote to demonstrate how machine learning is an instrument of subjectivation firmly based on the calculated measurement of empirically observed body characteristics that prescribe the difference between the "normal" and the "abnormal." This distinction relies on and is predefined by anthropometric measurements, such as height, weight, and the circumference of the chest and limbs (Rumbo-Rodríguez et al., 2021). It is based on empirical observations of quantifiable natural occurrences of frequencies of human body traits, presumed to be an unbiased method to classify populations and reflect the "natural order of things." Once an arduous and time-consuming job, anthropometric measurements are now automatized by full-body scanners, a machine capable of quickly "reconstructing" the human body as a 3D model. The supposedly "mechanical objectivity" of this automated process lends further credence to the idea that their operation is necessarily neutral and impartial, giving credibility to both the method and the technology. As such, Anthropometry has been used to study large-scale patterns in society, define standards in industrial production, and often serve as a justification for racism, slavery, white supremacism, sexism, homophobia, xenophobia, ageism, colonialism, and imperialism (Cuff, n.d.).

Security agencies quickly adopted full-body scanners to detect threatening individuals in critical spaces, such as airports. However, the technology behind these scanners is considered too invasive, as it reveals passengers' naked bodies to security agents, and unreliable due to the high number of false alarms (Amir & Kotef, 2018; Fuster et al., 2015). The high number of false positives became a public image issue for the DHS and TSA, increasing the distrust in the technology and the objectives involved in the mass surveillance project launched by the American government. More than just an occasional error, the predictive algorithm embedded in the scanner actively and disproportionally discriminates against minority groups, considered "abnormal" by the underlying dataset used to train the system (Costanza-Chock, 2018). The Passenger Screening Algorithm Challenge, a rare occasion when the DHS crowdsourced new security technologies, was a cheaper way to respond to the public outcry. On the surface, the competition was advertised as a way to speed up the scanning process to reduce long lines and manual searching. For the machine-learning community, which generally holds the belief that technology and private enterprises are more efficient than public agencies' bureaucracies, it was a cry for help. It was an opportunity to show off machine learning's state-of-the-art and demonstrate how governments and state agencies can take advantage of a community of developers willing to work for free. In other words, the DHS has offered US$1.5 million to fix a billion-dollar system to self-justify systematic surveillance at American gateways and keep the country's geopolitics regarding the so-called war against terror alive.

The undeniably but fairly obvious discriminatory concessions made by Kaggle and its community of developers caused controversy surrounding the tournament. On the one hand, due to regulatory restrictions imposed by U.S. law, only U.S. citizens were eligible to win cash prizes. Instead of creating alternative awards for non-U.S. users, the company admittedly made discriminatory compromises for the "sake of a dataset." Goldbloom was clearly more interested in exploiting the community's free labour for profit's sake than in the always celebratory libertarian discourse that the data science community thrives on in its collaboration beyond national borders. On the other hand, the geopolitical division was further exacerbated, given the nature of the predictive models to be built. Identifying threatening individuals in airports quickly raises the question of a common enemy that must be preemptively detected, deterred, and detained. By preventing citizens from other nations from participating, the competition exposed a political and cultural bias against minorities, more specifically Arabs, Muslims, and black people. Noj Vek (2017), one of the competition participants, asks, "what next, … a terrorist detection competition where [M]uslims are not allowed to participate; a crime classifier where black people are not allowed to participate?" (n.p.). The restriction on the eligibility for this competition has far more implications that go beyond financial rewards or the kudos of cracking a problem. By marginalizing part of the machine-learning community from the construction of a solution, the DHS and Kaggle are reinforcing social biases and prejudices.

More importantly, the Passenger Screening Algorithm Challenge shows that Kagglers see themselves as a community of data scientists detached from the political and ethical realm. There is a disconnection between the activities during the competition and the impact of their work on society as a whole. There is no public discussion about the sociotechnical implications of the predictive models being developed on Kaggle, the resources used to build these models, whether or not they are a good fit for the task, or if they are reliable, balanced, or biased. The data used to build these predictive models appears cleansed of the historical, social, and cultural contexts in which it was constructed and gained meaning. In this competition, all that mattered was having access to a large number of examples of "normal" human bodies to produce an algorithm that could identify deviations from the "norm," which in turn could give them a chance to win money prizes. Machine-learning practitioners make shortsighted decisions because that is what Kaggle's community, and the AI industry as a whole, incentivize. Without any reflection, critical thinking, or ethical considerations, Kagglers were not learning or developing research, as they like to say, but following orders, doing precisely what Tom Ripp, former L-3 Communications President, prescribed for the full-body scanner: "look for abnormalities" and detect "objects that are not supposed to be there" (Grabell & Salewski, 2011, Guns, Sweat and Privacy Fears, para. 10).

More than just a gateway of scrutiny and a place for harassment, confusion, and misunderstanding, full-body scanners became an instrument of permanent, exhaustive, omnipresent surveillance, capable of making everything visible while itself remaining invisible. These scanners produce transparent data about the passengers' bodies, rendering visible the human anatomy, but are not capable of telling anything meaningful about the passengers' intentions, including assertions or signals of threats. Instead, the full-body scanner serves here to demonstrate a tendency of machine-learning algorithms in the configuration of sorting bodies and the violence this configuration entails. They are not used as a weapon to respond to actual attacks but rather as a way to prevent the development of emerging threats. They aim to detect, sort, deter, disrupt, control, or detain suspicious individuals before they can do any harm. Predictive models, such as the ones embedded in full-body scanners, are not designed to punish—though this is a side-effect—but rather to "predict" specific events in order to "preserve" society from the danger presented by the presence of deviants or abnormal beings. Consequently, they are part of a governmentality designated technique to control the conduct of individuals or collectives of people (Foucault, 1982). They are designed not only to identify and classify but also to structure a population's possible field of action and behaviours.

This chapter demonstrated how the shortcomings of objective, statistical, and scientific language are used as unequivocal justifications to produce an assemblage of code-machine that projects a distorted and decontextualized identity over specific individuals. This decontextualization depoliticizes such identities and masks them as merely numeric, yet within a field wherein risk is defined as an abnormality. The securitization of the "abnormal body" aims to create a speculative future as a way to produce knowledge about human behaviour. As Hong writes (2020), "data-driven surveillance seized its claim to knowledge by mobilizing projections and estimations about technology and the future world that will necessitate those technologies" (p. 15). Consequently, the design of machine-learning algorithms and the initial identification of the kind of data to be gathered renders it an interested, if not deliberately biased, process. One such driving interest is precisely the manufacture of usable, justifiable certainties that become historical inevitability; all they can do is *predict the future of the past*, reinforcing current social, economic, and political norms. The next logical step is to ensure that these speculative futures unfold accordingly. The following chapter takes a step forward to discuss how similar sociotechnical strategies have also been used to nudge individuals and collectives of people into doing or thinking in particulaflr ways.

8. Nudge: Radical Recommendations

> When my rapist showed up under the People You May Know tab on Facebook it felt like the closest to the crime scene I've ever been.
>
> That is if I don't count the clockwork murder that I make of my own memory every time that I drive down Colfax avenue.
>
> Still, I sit in my living room, I sift for clues.
>
> Click; I see myself caught in his teeth; He's dancing with his shirt off in a city that I've never been to.
>
> Click; he is eating sushi over a few beers with friends and I am under his fingernails.
>
> Click; I know that alley.
>
> Click; I killed the memory of that t-shirt.
>
> Click; this is an old photograph. It's a baby picture. There's also an older man, presumably his father, they are both round and right and still smiling.
>
> Click; he is shirtless again and I catch my reflection in the weight room mirror. "#beastmode selfie"
>
> — Kevin Kantor, 2021[22]

In his spoken word poem, Kevin Kantor tells his own experience as a male victim of sexual violence and how social media, specifically Facebook, exacerbated it. He describes seeing his rapist being suggested under the platform's "people you may know" section. Contrasting the shock of seeing his rapist to the aftermath of a crime scene, Kevin recollects thoughts and emotions he experiences while clicking through his rapist's profile and details about his life, such as the music he listened to, his baby pictures, and the distress caused by being informed by Facebook that they had three mutual friends, ultimately bringing to life the person he actively buried. Kevin not only bravely calls attention to the scrutiny, blame, and stigma men and boys receive after revealing their sexual assault but also illustrates how recommendation systems can deeply affect someone's life. This

[22] Watch Kantor recite the poem in full at the 2015 College Unions Poetry Slam Invitational (Button Poetry, 2015) on YouTube: https://www.youtube.com/watch?v=LoyfunmYIpU

contentious feature, powered by machine learning and now commonplace in most large digital platforms, is generally built for the current economic system's typical and mundane goal: retain user attention and increase sales. To work "correctly," or at least within a range of acceptable accuracy, these systems need to ingest vast amounts of data about and around their users, from the intimate and personal details to the impersonal aspects of the self. The suspicion that you have been followed,[23] spied upon, and have had your privacy violated is a recurrent feeling most people experience when using digital technologies today. We have become targets of algorithms designed to extract behavioural data and feedback using nudges (Thaler & Sunstein, 2021), directing us to do specific things, most notably to consume goods, services, and media content.

Advances in behavioural tracking hardware (geolocation and geofencing, accelerometer and gyroscope, biometrics monitors) and software (cookies, logs, analytics, machine learning) entice marketers who can more easily identify, track, and intercept our activities as we consume digital information. With users regularly sharing personal data online and digital platforms tracking every click, marketers have gained unprecedented insights into users' and consumers' minds. By collecting personal data, ad tech companies claim they can tell who their users are and what they like—so they can then more accurately target advertisements at them. The information feedback to the individual becomes over-personalized and tailored to their needs, looking to provide personal experiences. Over the years, these same ad tech companies have convinced advertisers to buy data and micro-targeting services in order to create "better ad experiences" for consumers. It is commonly believed that, over time, consumers will benefit from personalization due to the increased ability of a system to know their preferences and to make recommendations uniquely suited to their wants and needs (Thomas Kramer et al., 2007).

For marketers and technologists alike, behavioural tracking is a benign instrument designed for individuals so they may exercise their right to be informed consumers and receive precisely what they are looking for. However, as discussed in previous chapters, these technologies have been developed using involuntary and frequently illegally obtained user data by installing web bugs on their browsers and devices or by scraping, acquiring, and purchasing large datasets of personal information. Sophisticated software, now based on machine learning, enables the combination of

[23] Often this is experienced as harassment or perceived as if someone were stalking you. According to Tjaden and Thoennes (1998), stalking involves repeated physical proximity, unwanted communication, threats, fear, or a combination of these events. One of the conditions of stalking requires the communication to be nonconsensual, where "more than one overt act of unwanted pursuit of the 'victim' is perceived as being harassing" (Meloy & Gothard, 1995, p. 259).

multiple sources of personal behaviour—online and offline—and provides ways of making speculative inferences based on predictive models. These models, in turn, facilitate the construction of hyper-personalized recommendations that are displayed in many different formats, most commonly as ad banners alongside newsfeeds and pop-ups, but also mixed up with friends' updates on social media, geofenced notifications on smartphones, interruptions of audiovisual material, interleaved educational material, and disguised as hard news (popularly known as fake news). This form of micro-targeting occurs in real-time as individuals are constantly in close proximity to digital sensors and systems.

The anticipation, that is, the prediction of behaviours, actions, and events, is not only valuable for advertising. As I have been arguing in this research, predictive models also serve as a convenient and powerful instrument of control to prescribe decision-making strategies. Data-driven predictions tell stories about *what would happen* in a particular future based on partial and supposedly objective historical accounts of the past. In the machine-learning world, these stories are scored and ranked, providing probabilities about what will come next, quickly becoming arguments for an inevitable truth about our conditions of existence. They are used to self-justify and self-fulfill business models, political economy recipes, social norms and behaviours, cultural formation, consumption habits, technological development, and many other ways to drive society toward specific goals. However, despite the certainties promised by predictive models, machine learning can only produce approximation accounts of what comes next, generating fragile models that cannot account for complexities beyond the scope they were built for.

Consequently, predictions and forecasts have a tendency to crumble in the face of unaccounted factors unfolding in the present, be it large-scale events like climate change or the Covid-19 pandemic; localized actions such as political demonstrations and artistic interventions; personal choices like lifestyle changes; or micro-scale incidents like DNA mutations. To fight against unforeseen events, the proposed futures must be ensured by mediations and interventions to correct course if needed. These interventions are usually at the decision-making level, rendered as notifications, suggestions, and recommendations to aid individuals in deciding what to do next. These aids, or prescribed actions, may not appear as a violent and forceful power upon the subjects to whom they are directed. On the contrary: though they can be intrusive (and very annoying), they blend into everyday life as convenient instruments to help individuals accomplish common tasks, such as shopping, staying healthy, keeping informed, or finding friends. The subtlety of these instruments of control substantially impacts our lives, driving our behaviour, habits, preferences, and decisions by prescribing and reinforcing specific ways to do and think things.

On Kaggle, most competitions involving decision-making aids—another name for recommender algorithms—are targeted at large-scale consumption systems, such as search engines, social networks, e-commerce, delivery platforms, and entertainment media streaming. Since the data science community's interests are skewed toward marketing, sales, and financial gains, these competitions focus on optimizing inventory, retaining user attention, creating sales opportunities, enhancing customer experience, or satisfying users' needs. On the surface, the end goal is always the same: reduce costs and increase profit. However, the way these recommendation systems work has a more profound impact on the individuals exposed to them. Recommendation models interpellate these individuals as subjects of the platform, affecting how they will act in front of the pre-selected choices offered to them. Consider, for instance, what is at play when you do groceries: Which items go into your shopping cart? In what order? How often do you buy these items? How does the diversity and availability of products impact your decision? Similarly, is the way these items are arranged and categorized influence what you choose to buy? Do the language or visual aspects of a product affect your decision? Do advertisements convince you to change your mind?

To understand how recommendation systems affect individuals, this chapter explores how Instacart, a grocery delivery and pick-up service based in the U.S., has been using machine learning to relieve users of the "burden of decision-making" and direct them to perform specific actions in order to increase sales. Dubbed "Uber" for groceries, Instacart has 10 million monthly active users and 500 million products on the platform, providing a direct channel for customers to order food from participating retailers (Curry, 2022). The platform has been using a recommendation system to aid users to "do grocery more efficiently" and improve "user experience" in their app as they expand to new markets. For Instacart, "whether [we] shop from meticulously planned grocery lists or let whimsy guide [our] grazing, our unique food rituals define who we are" (Instacart, 2017, para. 1).

Looking for a cheap way to optimize its services and improve the platform's underperforming recommendation system, the company released a dataset with a sample of its users' activities—three million food delivery orders from over 200,000 users—as part of its 2017 Instacart Market Basket Analysis challenge on Kaggle. This competition serves as space to discuss the third mode through which machine learning has been explored: a recommendation system or a nudge device that makes individuals choose specific outcomes or act in specific ways. That is, by using predictive models, Instacart claims to *know what is going to happen next* and *actively intervene* in the user's eating habits, *ensuring that some products will be part of their next purchase*.

In contrast with the previous two chapters, where I explored the direct application of predictive models on the body, here I focus my attention on the impact of machine learning algorithms on the psyche and mind. Instead of facial measurement or external physical properties, the example explored in this chapter seeks to identify, predict, and shape psychological traits (behaviour, desires, needs, preferences). Psychopolitics (Han, 2017; Prozorov, 2021; Rau, 2013) is an instrument of governmentality where the computational configurations, predictive modelling in particular, are deployed as "technologies of the self" comprising instruments to prescribe the "conduct of conducts" (Foucault, 1982). The processes of subjectivation using psychological traits go beyond the materiality of the body in an attempt to tap into the immateriality and subjectivity of the inner self. Our habits, behaviours, and sense of identity, but also our feelings and emotions, are not only the raw material to feed large-scale training sets for machine learning but the heart of a new form of governing individuals.

This chapter discusses the operations through which individuals are subjected to hyper nudges that persuade and steer them to take specific actions that, in turn, shape their behaviour, habits, and desires. This chapter is divided into three parts following the same concepts of Sense, Mobilize, and Modulate. In the first section, I describe the strategies used by digital platforms to collect, identify, and reconstruct individuals' behavioural attributes. While there are many ways to acquire personal data, here I focus on tactical big data to infer users' behaviours, habits, and preferences based on traces we leave behind when we consume or socially interact through digital technology (clicks, likes, views, logs, browsing history, purchases, etc.). Since this approach is fragmented and incomplete by nature, I also explore the role of data brokers in augmenting, recombining, reconciling, and elucidating patterns of user behaviours into discrete and distinct hyper-profiles.

In Mobilize, I discuss how the machine-learning community explores and exploits personal data to produce predictive models for recommendation systems. Here, the Instacart Market Basket Analysis serves as an example to discuss how data and code are mobilized to make inferences and assumptions about mundane and everyday life activities, such as purchasing groceries, and exactly how it can be used to nudge consumers to buy specific items or to adopt specific habits for convenience and efficiency. In particular, I discuss the motivations behind the competition, how the community made sense of the dataset provided by Instacart, how they solved the challenge of predicting and recommending products for the users' next purchase, and the dynamics involved in crowdsourcing this type of task.

Lastly, in Modulate, I consider the outcomes of the competition and the broader implications of recommending systems. I discuss the reflexive nature of personalized recommendation systems that tend to reinforce behaviours, habits, and discrimination, locking individuals into an echo chamber (Pariser, 2012). Moreover, I examine the current trend of radical subjectivation, where digital platforms allow and encourage advertisers and users to produce extreme and radical content as a way to retain and engage their audience. I argue that the goal of predictive models for recommendation systems goes far beyond their applicability to sell more products in retail stores or engage broader audiences on social media platforms. Instead, these models seek to discipline the individual into making specific tasks, reshaping their habits, and thinking things in a particular way. In other words, predictive models created by large tech groups aim to prescribe what should happen next as a form to *produce specific subjects* that fit a particular and well-defined condition of existence.

8.1. Sense: Behaviour Data

Crucial to discussing Instacart and Kaggle is how behavioural data has become sensed as big data. For over a century, retail and grocery stores have monitored consumers and tracked shopping behaviour to increase sales (Turow, 2017). For most of the twentieth century, strategies to attract new buyers used mass campaigns targeting broad categories of people: coupons, loyalty programs, carefully designed built environments, planned ad campaigns, status conferral, gift cards, and free samples. In the early days of retail's direct marketing, refined data about shoppers was scarce, and probabilistic statistics were only relevant to professional gamblers and actuaries. With the widespread digital technology at the end of the twentieth century, massive datasets and increasing computing power became available, creating the conditions for predictive algorithms to become a universal way of thinking and be incorporated into all aspects of our lives, including managing shopping behaviour and dietary habits. As a result, in the 2000s, large grocery chains in North America and Europe used a wide range of devices (e.g., Bluetooth beacons, WIFI hubs, cellular antennas, GPS satellites, websites, apps, and mobile devices) to collect personal data, construct hyper-personalized models that explain user behaviours, and deliver customized recommendations to shoppers.

Data about a person became easily traceable, storable, sortable, trackable, and programable. Digital technologies are not just another powerful and invasive surveillance apparatus. The political economy prescribed by and for digital platforms identifies human behaviours in terms of transactions: every data point is perceived as actionable for a potential commercial opportunity with effects on the modern economy. This business opportunity is clearly articulated by Hal Varian (2010), Google's top economist, who sees new uses for behavioural data, such as "data extraction and analysis," "new contractual forms due to better monitoring," "personalization and customization," and "continuous experimentation." It is no surprise that Varian was among the first venture capitalists to invest in Kaggle and that his ideas reflect the company's main goals and reverberate in the data science community. He was the godfather of Google's economic model, preaching the use of machine learning to efficiently "squeeze" meaning and signals from these users and convert raw material (behavioural data) into the firm's highly profitable algorithmic products designed to predict the behaviour of its users.

The asymmetries of this power balance are linked to the effects of the "conditions of existence" defined by the digital economy, which facilitate shaping behaviour and impose a social relation upon the platform's users. With the popularization of social psychological instruments and strategies, such as the Nudge Theory (Thaler & Sunstein, 2021) and psychometrics applied to advertising (Kosinski et al., 2013), technologists and marketers seek to become architects of "choices" as a way to influence consumers' decision-making process according to a specific socioeconomic standard that is supposed to be for the "greater good." However, Turow (2017) observes that retailers' strategies "mix shrewd loyalty programs, high-tech tracking instruments, and esoteric statistical manipulation with soothing brand images and smoke screens in such a way that shoppers accept systematic biases about them" (p. 11). By making decision-making more automatic and less reflective, machine learning and artificial intelligence backed up by large corporations are reshaping the sense of self, producing subject-consumers more susceptible to personalized ads and, ultimately, to ideological propaganda.

This operation requires significant amounts of subject material extracted from individuals' behaviours, habits, and preferences, which Zuboff (2020) refers to as "behavioural surplus data." Behavioural surplus results from a convergence of data science, material infrastructure, computation power, algorithms, and automated systems, combined with psychometrics and fantasies of epistemic purity (Hong, 2020). It is a by-product of digital technologies comprising unstructured signals, often mined and obtained from individuals without their consent or awareness, providing detailed stories about each person—thoughts, feelings, and interests. This method of obtaining data enables digital platforms to surveil, capture, expand, construct, and claim behavioural surplus, including data that users intentionally choose not to share or data protected by local policies or federal laws (e.g., copyright). No moral, ethical, legal, or social constraints stop digital platforms from using behavioural data for commercial purposes. In this sense, individuals are deemed a source of raw subject material from which companies can extract and process data to produce models—hyper-profiles—to optimize and personalize services, most notably targeted advertising. These profiles not only *re-present*, but also *re-act* on behalf of every single user, customer, or citizen. Eventually, they can be used for behaviour modification, shaping individuals' experiences in relation to other individuals, their social sphere, and the physical environment (Langlois & Elmer, 2019), ultimately producing new forms of subjective and social conditions.

8.1.1. Individual Signature and Hyper-Profiles

Ideally, personal data is thought to be acquired voluntarily. The users should post, upload, and update the data by themselves without any coercion. A confession (Foucault, 2017) of their everyday life, behaviour, and desires. This low-cost process to obtain data makes individuals feel they have some control over and ownership of their own lives, deciding what to share and in what circumstances. The problem is that this process is slow and limited. Users may not keep regular updates, decide not to share a portion of their lives, lie about themselves, be unable to inform specific data due to its pre-conscious or involuntary nature, or even lack the means to do so. The alternative is to extract and collect data using automated process and tracking devices, which are not submitted spontaneously but with the presumed consent of the user that the data may and will be transmitted and stored elsewhere in external data centres. This is the case with cookies, Web trackers, and digital sensors in smart-phones, -cameras, -watches, -cities. Though individuals have some awareness that data has been collected, they usually do not know when this happens and what type of information has been released. This type of data collection process captures the unsaid, the involuntary, and the automatism of our actions.

Users and digital platforms value the material being shared differently. Individuals have a subjective, personal, and affective attachment to their data; they can only see it from their own perspectives, valuing it according to their own point of view and experiences. Digital processes, on the other hand, are indifferent to data, an impersonal standpoint (Langlois & Elmer, 2019) that makes social, political, and cultural values disappear in favour of some sort of optimization, which, in most cases, is economically driven. In fact, at this level, the very concept of data becomes obscure for most of us since even innocuous and harmless bits of our lives may find their way as a utilitarian variable of optimization processes. Little things—such as what we like or follow on social media, how many times we visit a website, how long we spend watching TV, what time we usually wake up, how much time we spend standing on one leg, the way we arrange the food on our plate, our heartbeat, or how often we buy bananas—become nodes and layers of a gigantic machine-learning matrix made to crunch human lives into the probabilities of their next purchase.

To produce strong predictive models, digital platforms may need to obtain much more data than they can collect. As the raw material of the twenty-first century (Srnicek, 2017; Zuboff, 2020), personal data can be sold, bought, and traded by data brokers, such as Acxiom, Epsilon Data Management, and Oracle. Acxiom (2021), for instance, advertises that it collects comprehensive global data to connect market technology with advertising execution, offering data about consumers across 62 countries. In a report published in 2017, Cracked Labs revealed that Acxiom claims to have more than 11,000 data attributes on 2.5 billion people to help brands connect with people "ethically" (Christl, 2017, p. 54). What Acxiom has to offer, however, is just a subset of what large digital platforms are able to ingest. In 2016, ProPublica revealed that Facebook worked with more than 52,000 unique attributes for ad targeting, which includes data about users' finances, such as the number of credit lines one has, the likelihood of an individual having a loan or whether one carries a balance on one's credit card; shopping habits, for instance, how many times a user buys groceries in a month, the types of food a user buys, the last purchase made, streaming subscriptions, and much more (Angwin et al., 2016). This level of data collection and aggregation allows for precise profiling, giving data brokers the ability to identify and target subgroups upon subgroups of individuals through criteria such as gender, race, marital status, income level, and all the sensitive characteristics that these individuals had no idea would end up in a database—let alone be up for sale.

However, subjective and personal data is messy. Tracking people in many situations of their lives using massive datasets from multiple sources and recognizing them as the same individuals is no trivial matter. Traditional social identifications, such as legal names and driver's licenses, can produce ambiguity on a global scale. Advertisers and digital apps usually use email addresses, phone numbers, and unique smartphone identifiers to link profiles and behavioural data across different databases, platforms, and devices. Nonetheless, these are not unique since people change smartphones frequently, have multiple phone numbers and emails, and share devices and service accounts (e.g., TVs and streaming platforms). Sub-profiles are assigned to the person actively using the device to disambiguate users in shared devices. For instance, Netflix, YouTube, and Apple allow users to switch accounts to "improve their experience." Large platforms assign "advertising IDs" to individuals to centralize all these identifiers, now widely used to match and link data.

Yet, this is not enough to track all the digital traces an individual leaves behind. Amperity (2011) advertises itself as a Customer Data Platform, promising to resolve customer identity at scale. The company claims to use artificial intelligence to unify fragmented records from different places into complete comprehensive hyper-profiles of every individual across platforms, devices, services, and sources. According to the company, these profiles adapt to changes in the individual's behaviour over time, aggregating new data points and different aspects of their life into canonical subjects (also referred to as ontologies) that can be accessed, filtered, categorized, selected, or discarded from a dashboard. The company describes its system as a "fully transparent approach to build a consistent, trusted identity foundation for your business, so the work you do to understand, engage, and serve your customers drives the results *you want* [emphasis added]" (Amperity, 2011, n.p.). Knowing *who* the customers are, or, in general terms, knowledge about the subject is imperative not only to be able deliver meaningful content and services but also to *create the conditions* for these experiences to take place.

8.1.2. Hyper Nudges

Once individuals are marked and identified by their hyper-profiles, the data flows in a reverse course and targets them as subjects of hyper nudges (Hull, 2018). Hyper nudges are a direct product of Behavioural Economics and Psychometrics combined with big data and algorithms, usually used to influence attitudes, habits, and opinions. By configuring and personalizing the users' information environment, typically through algorithmic analysis of multiple data sources, digital companies claim to offer predictive insights concerning the habits, preferences and interests of targeted individuals. Such a vast amount of knowledge is, in turn, used to influence these individuals, a strategy present at the core of many digital platforms: from Google's PageRank and Facebook's EdgeRank to Fitbit's incitement for a "healthy life" and the pressure to watch the next Netflix blockbuster. They are operated using recommendation systems that configure and personalize the users' information environment, offering pre-selected choices that supposedly match the user's hyper-profile. In other words, they interpellate individuals as subjects of the platform, affecting how they act in front of the pre-selected choices offered to them.

Digital platforms often use hyper-profiles and predictive models to play with the users' preferences, emotions, and affections in order to lure them into engaging with specific content, products, ads, interactive features, and even with other users. They come as imperative calls for action, such as scroll, click, swipe, like, follow, buy, subscribe, react, watch, play, and share. Though suggesting actions and driving user behaviour might be effective, users may find themselves overwhelmed by the options in front of them. That is, in the "battle" for the user's attention, recommendation systems are devoted to creating the conditions to nudge the user in a specific direction, most often to consume media, purchase products or services, or interact—and fast—with the platform. Despite the complexity and sophistication of algorithmic processes, these digital systems ultimately rely on a deceptively design-based mechanism of influence (Hull, 2018) that "hyper nudges" individuals into doing or thinking in particular ways. The hyper nudge is the underlying base for complex systems of notification, recommendation, and gamification with the aim of engaging users with the platform, shaping behaviours, and directing attention toward a set of predefined "choices architect[ed] through processes that are subtle, unobtrusive, yet extraordinarily powerful" (Yeung, 2016, p. 119). In other words, we have become targets of algorithms designed to extract behavioural data and feedback with nudges directing us to do specific things, most notably to consume goods, services, and media content.

Hyper-personalization and recommendation systems have not only become a norm in the marketing and sales sectors, but they are also increasingly becoming an insurmountable social problem with the potential to lock individuals in "social bubbles" in which people are exposed only to information from like-minded individuals, thus amplifying confirmation bias (Pariser, 2012). As I will show in the next section, while Instacart was still learning how to tatget its customers' shopping behaviour using recommended pre-composed grocery lists, other digital platforms have been implementing and perfecting recommendation and notification systems as a form of social and economic control for over a decade. For instance, Bucher's (2012) examination of Facebook's GraphRank algorithm shows that the company monitors users' activities inside and outside the platform to learn their behaviour patterns. Once these patterns are found, they are fed back to the users via the platform's News Feed. Consequently, "even more users will apparently act in the way that the algorithm predicts" (p. 14). While Napoli (2014) argues that this process leads to a "certain amount of reflexivity that is inherent in much algorithmically driven media consumption" (p. 346), Turow (2017) reminds us that the same applies in the retail and grocery market. The individuals may vaguely understand how the platform works, and "they may try to change their behavior to get better deals, often without success, all the while wondering why the system … is treating them that way" (p. 260).

Hyper nudge does not just offer subjects choices: "they create subjects" (Hull, 2018, n.p.). This is subjectivation in action. The code that sets these choices is built and selected by code-writers and designers. It constrains the direction of any action by showing or hiding other possibilities (Lessig, 2006). For Hull (2018), hyper nudges structure the informational environment, which is directly constitutive of the choices we can make: "If we do not know something is there, we cannot choose or not choose it" (n.p.). Because algorithmic media is contextual, it can modulate outputs based on the users' profiles, spatial and temporal conditions, and all kinds of contentious events. Moreover, code can embed countless nudges, making it difficult for users to maintain independence or build a sense of shared reality in an entirely mediated world. Hyper nudges, by design, treat all users in as precisely an individuated manner as possible. As a consequence, the information environment creates the "truth" of the world and produces forms of subjectivation that users, machines, and other entities will relate to and operate within.

To accomplish this goal, machine-learning engineers need in-depth knowledge about the user base. Then the question moves from *who is* the user, and what is their *behaviour signature*, to *how can individuals be exploited* based on these signatures? On Kaggle, it is easy to spot the contours of power imbalance where user data is scrutinized, inferred, presumed, crunched into hyper-profiles, and nudged in septic directions. The following section discusses how the machine-learning community explores and exploits shopping behaviour and eating habits to produce predictive models for recommendation systems based on a curated dataset prepared by Instacart.

8.2. Mobilize: Instacart Market Basket Analysis

Founded in 2012 by Apoorva Mehta, a former Amazon employee, Instacart is an American company operating a grocery delivery and pick-up service via a mobile app and website. It has been dubbed "Uber for groceries," providing a direct channel for customers to order food from participating retailers. The shopping and delivery are carried out by a fleet of tens of thousands of "personal shoppers" who use their own means to accomplish the task. With 500 million products on the platform, Instacart claims to have the most extensive online grocery catalogue. Like other e-commerce and digital platforms, Instacart has largely benefited from the Covid-19 pandemic lockdown, during which the service saw a surge in demand, reaching more than 10 million monthly active users, which, in turn, made the company profitable for the first time since its foundation, driving its market value from US$13 billion in 2020 to US$39 billion in 2021 (Curry, 2022). By 2022, the service was available through 45,000 partner stores in more than 5,000 cities across the U.S. and Canada.

With a similar business logic as Uber and Airbnb, Instacart fits Srnicek's (2017) concept of a lean platform, operating as an intermediary in a multisided network with minimum ownership of assets and under deregulated labour practices. In one of the first slides in a presentation at DataEngConf 2017, Sharath Rao (2017), Instacart's Director of Engineering, explained the company's business model, showing how the platform connects and mobilizes people, products, and data (see Figure 8.1), and, at the same time, remaining outside of these relationships where it can maintain oversight and control without accountability. In a multisided network business model, the logic of data-driven operations has a specific directive: make the actors involved act fast to intensify engagement and, consequentially, increase sales. The question is how to make people engage with the service more often and buy more. In the retail business, some of these problems are related to how products are arranged and what are the consumer's expectations, which, if misaligned, could prevent the transaction.

For Jeremy Stanley (2017a), former VP of Data Science at Instacart, "shopping for groceries is hard … The hummus you want could be in the dairy section, the deli section, or somewhere else entirely. Efficiently navigating a store can be a daunting task" (para. 1-2)—an assertion that might only hold true for part of society that lives in excess and does not care about food security. Imagine, for example, he asks the reader, having milk ready to be added to your cart just when you run out, or knowing it's time to stock up again on your favourite ice cream. Instacart aims to make your shopping experience less painful, more convenient, and personalized—but only for those with the means and willingness to pay for it. To that end, Instacart needs to know its users, classify them in

Figure 8.1: Instacart's Four Sided Marketplace. Slide extracted from Rao (2017) presentation at DataEngConf 2017 explaining Instacart's business model.

user id	order id	order number	order dow	order hour of day	add to cart order	product id	product name
1	2539329	1	2	8	1	196	Soda
1	2539329	1	2	8	2	14084	Organic Unsweetened Vanilla Almond Milk
1	2539329	1	2	8	3	12427	Original Beef Jerky
1	2539329	1	2	8	4	26088	Aged White Cheddar Popcorn
1	2539329	1	2	8	5	26405	XL Pick-A-Size Paper Towel Rolls
1	2398795	2	3	7	1	196	Soda
1	2398795	2	3	7	2	10258	Pistachios
1	2398795	2	3	7	3	12427	Original Beef Jerky
1	2398795	2	3	7	4	13176	Bag of Organic Bananas
1	2398795	2	3	7	5	26088	Aged White Cheddar Popcorn
1	2398795	2	3	7	6	13032	Cinnamon Toast Crunch

Figure 8.2: Instacart Data Sample. The first two orders on Instacart for user_id 1 (Stanley, 2017a).

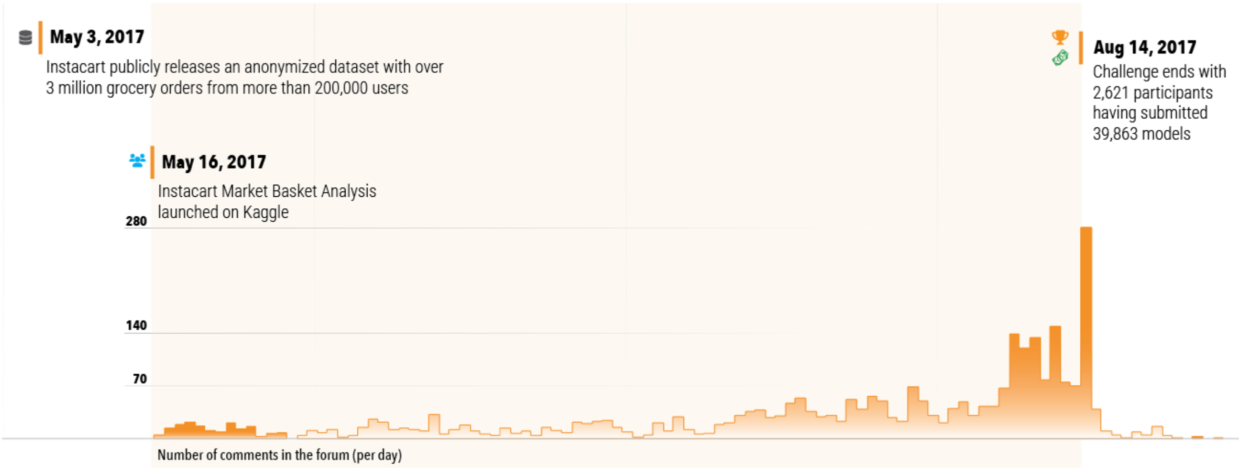


Figure 8.3: Instacart Market Basket Analysis Timeline. Timeline showing the competition milestones and number of comments in the forum through time. Designed by L. Frizzera.

useful ways, find out their behaviour and preferences, and finally nudge them into making purchases.

Broadly speaking, as discussed in the previous section, usefulness can be read as fitting a particular purpose, such as social stratification, cultural formation, or marketing segmentation, where the individual becomes the subject of a specific matter. As such, the company uses transactional data to model its users into hyper-profiles using complex shopping behaviour, food preferences, eating habits, and dietary restrictions. These hyper-profiles are used to predict which products a customer will buy again, or try for the first time, and in which order they might put these items in their cart. These products are sorted, ranked, and conveniently rendered as recommended shopping lists. This type of predictive model allows Instacart to know *what is going to happen next* and *actively intervene* in the user's eating habits, *ensuring that some products will be part of their next purchase*.

On May 3, 2017, Instacart enticed the data science community after publicly releasing a large-scale dataset with three million Instacart orders from over 200,000 users. It contained a 30-day shopping behaviour of its users, including the order in which each item was put in their cart, the day of the week, and the order's time (see Figure 8.2). There was also limited information about the items these users purchased, such as their name—which might contain extra information about the products, like brand and package size—and the aisle and department they belonged to. The company used this dataset to launch a challenge on Kaggle two weeks after the data was released, on May 16, 2017: the Instacart Market Basket Analysis (see Figure 8.3). The goal was to crowdsource the effort to improve the platform's underperforming recommendation system (Rao, 2017). The company offered job interviews and US$25,000 in prizes for the top three participants who could increase accuracy in predicting “which previously purchased products will be in a user’s next order” (Instacart, 2017, n.p.). The award was not very high, but the chance to work for a rapidly growing start-up company in Silicon Valley attracted thousands of competitors worldwide.

8.2.1. Bananas and Pizzas: You Are What You Eat

Before the competition began, Instacart shared insights into the dataset to encourage the data science community to engage with the challenge. For instance, the platform pays close attention to the frequency with which people buy products, making inferences about their habits in relation to the properties of each item (see Figure 8.4, left): “Fruits are reordered more frequently than vegetables—perhaps because vegetables are more intermittently purchased for recipes … soups and baking ingredients are least likely to be reordered—perhaps because they are less frequently needed” (Stanley, 2017b, Some interesting findings section, para. 3). The company also looked closely at when users purchase specific products, recording the exact time day they browse the app. For example, healthier snacks and staples tend to be purchased earlier in the day, whereas ice cream is far more popular when customers order in the evening (see Figure 8.4, right). Popularity is also an important factor: “Of the top 25 latest ordered products, the first 24 are ice cream! The last one, of course, is a frozen pizza” (Stanley, 2017b, Some interesting findings section, para. 6).

As soon as the dataset was available on Kaggle, competitors flooded the forum with their initial exploratory analysis, familiarizing themselves with the main features and attributes they would squeeze for the next three months. Philipp Spachtholz (2017a), Data Scientist at features4.com and

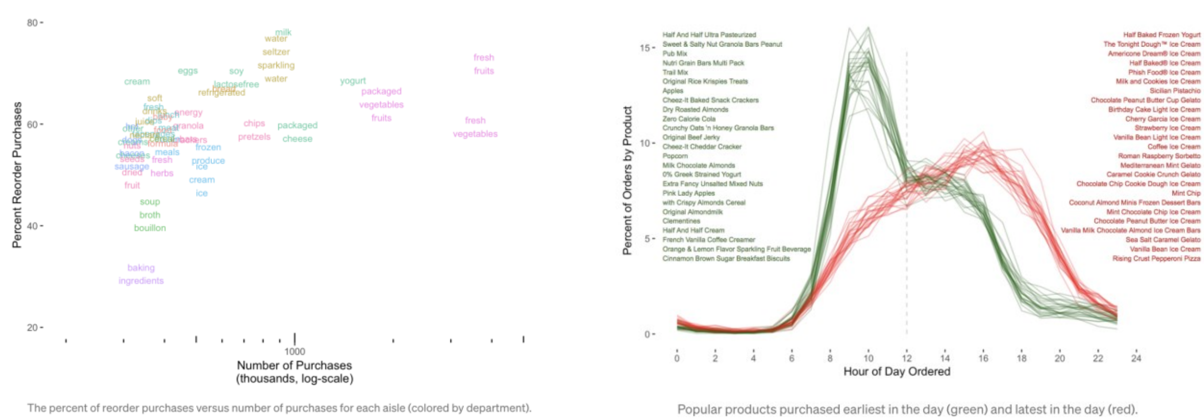


Figure 8.4: Food Purchase Distribution. On the left, the percentage of reorder purchases versus the number of purchases for each aisle; on the right, popular products purchased earliest in the day (green) and latest in the day (red) (Stanley, 2017b).

author of the top-voted exploratory analysis in this competition, shared the key variables and possible relationships one can find about Instacart customers' shopping habits. For instance, most grocery orders are made over the weekend, between 8 am and 6 pm. People order more often after precisely one week or 30 days after the previous purchases, revealing the ordinary family's planned behaviour to restock their fridges weekly or monthly. While most orders often contain around five items on average, some people buy 40 or more items at once. Moreover, consumers seem to be quite sure about "Multifold Towels;" if they buy them, they are the first item in the basket 66% of the time. Bottles of "Sparkling Water" and "Purified Alkaline Water with Minerals pH10" comes next, followed by "Organic Dark Roast," "2% Lactose-Free Milk," "Cookie Tray," "Boneless Pork Shoulder Butt," "Party Tumblers," "Sports Bottle with Flip Cap Natural Spring Water," and "XL Pick-A-Size Paper Towel Rolls."

Besides simply descriptive statistics, Instacart explicitly pushed for a predictive model to build recommendation items for reordering; Kagglers followed, looking for patterns and repetitions. Spachtholz (2017a) found that 59% of the items were reordered; that is, they appeared on subsequent orders made by the same person. The top ten products with the highest probability of being reordered are "2% Lactose-Free Milk," "Organic Low Fat Milk," "100% Florida Orange Juice," "Organic Spelt Tortillas," "Original Sparkling Seltzer Water Cans," "Bananas", "Petit Suisse Fruit," "Organic Low fat 1% Milk," "Organic Lactose-Free 1% Low-fat Milk," and "1% Low-fat Milk." Once again, we see the organic labels and mostly perishable foods, such as dairy products (milk, yogurt), juice, and fruits. However, industrialized items like water bottles and tortilla chips are also among the most popular reordered items. Neal Akyildirim (2020), Data Analyst at Altice, confirms Spachtholz's findings, pointing out that the most popular aisles are fresh fruits, fresh vegetables, and packaged vegetables and fruit, followed by yogurt and packaged cheese. Produce is the most popular department, followed by "dairy eggs and snacks." Beyond individual needs and preferences, Akyildirim claims that there might be other reasons why consumers are not purchasing from other departments: "If Instacart has the biggest markup for canned goods, but the user experience … does not entice the consumer for canned goods, regardless of the consumer needs the order amount on canned goods will be low" (n.p.). Moreover, considering that top items are sold mainly on Mondays and Tuesdays (excluding bananas), Akyildirim speculates that this can be related to retailer item inventory issues. For example, "would it be possible Trader Joe's runs out of organic strawberries by Wednesday, and [an] Instacart employee substitutes that order with a regular strawberry based on customer request" (n.p.)?

These first findings allowed Kagglers to speculate about user eating habits and hypothesize the possible causes for their shopping behaviour and preferences. In terms of food preference, Banana is the winner with 18,726 orders, followed by more bananas!—the organic kind, with 15,480 orders. Organic strawberries, organic baby spinach, large lemon, organic avocado, organic Hass avocado, strawberries, limes, and organic raspberries complete the top ten most popular items. It is no surprise that perishable foods, like fruits and vegetables, are at the top of the list (see Figure 8.5). Still, they do tell us what North American families are eating, or at least those who choose to buy from Instacart as opposed to ordering elsewhere or driving to the grocery store. What might be controversial is that most top-selling products have the "organic" label. saagie_anthony (2017), a data scientist at Saagie, a software company based in Paris, raised the issue: "I don't know if we have the same definition of organic in Europe and in the USA" (n.p.). The question can be extended to other parts of the globe, as each country has (or lacks) its own rules and policies to categorize these types of food. Despite what organic means in different contexts, Spachtholz found that organic foods are reordered (65%) more often than non-organic ones (58%). Nevertheless, the main question remains: How do reordered items relate to individual behaviours and eating habits, and how can this be used to predict a consumer's next purchase?

Kagglers wanted access to Instacart's partners and user data to better understand granular and contextual relationships: geographic location, age, gender, race, general demography, promotions, sales, etc. Instacart claimed that, for privacy reasons, it could not release sensitive information about users and other businesses. Before releasing the data, the company randomized the user ID in the dataset to guarantee that it would not be traced back to any specific individual. Only users

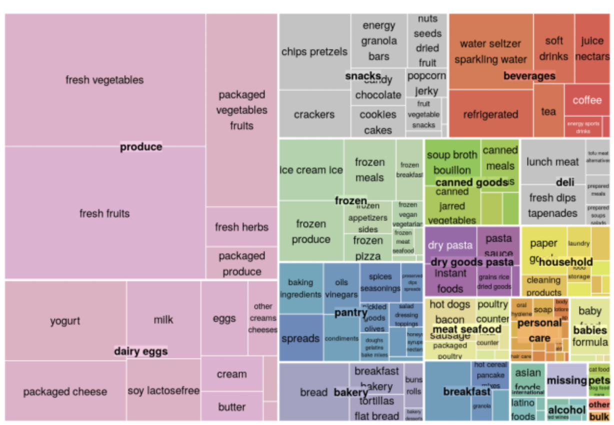


Figure 8.5: Supermarket Sector Distribution. How often products from the department/aisle are sold. Box size shows the number of sales (Spachtholz, 2017a).

who bought multiple products at multiple retailers were included, but no retailer ID was provided. The company recognizes that the dataset is a heavily biased subset of its production data and does not represent its products, users, or purchasing behaviour. Despite the efforts to anonymize the dataset, Spachtholz (2017a) took less than eight hours to show how deeply this type of data can pervade one's life. He looked for customers who reordered the same products all the time: "I look at all orders (excluding the first order), where the percentage of reordered items is exactly 1 … We can see there are in fact 3,487 customers, just always reordering products" (Exploring Customer Habits section, para. 1). Exploring a single user can reveal strong evidence of habits and dietary restrictions, such as user #99753, who has 97 repeated and identical orders: Organic Whole Milk and Organic Reduced Fat Milk.

While Kagglers cannot identify the users directly, Instacart certainly can. The company has tons of information about its users, including names, addresses, phone numbers, credit card numbers, and all kinds of data from the users' mobile devices and web browsers (search logs, browser history, geolocation, contact list), in addition to data purchased from data brokers. However, commercial interests tend to prevail on Kaggle. By protecting its users and partner stores, Instacart deprives the public of deeper inquiries about its business model and prevents other companies from poaching their users and partners. Since there are no other public means to access this type of data, the company cannot be held accountable for any mistake or wrongdoing in its algorithms. Nonetheless, three years after the competition, when Buzzfeed reported that the personal information of hundreds of thousands of Instacart customers was being sold on the dark web as a result of a data breach in the company (Lytvynenko, 2020), suddenly the anonymity of the dataset shared on Kaggle was compromised. Instacart has since removed the dataset from its own servers, but it is still available to download on Kaggle.

8.2.2. Finding the Sweet Spot

Exploratory data analyses, such as the one made by Spachtholz (2017a) and others, are just the beginning of much broader, extensive, and intense work to explore and exploit the dataset. To produce reliable predictive models using machine learning, data scientists and engineers need a massive volume of data as well as a large set of variables from which they can derive associations among data points. This is especially true when the goal is to achieve a high degree of personalization in targeting common individuals (the platform users) with behavioural nudges and media content (news, videos, ads, etc.). Market Basket Analysis is one of the techniques large retailers use to uncover associations among items; that is, it allows retailers to identify relationships among the things people buy. It is no coincidence that Instacart used this exact term as the title of its competition. Association rules are used to analyze retail transaction data in order to identify strong relationships using measures of interestingness. Grace Tenorio [DATATHÈQUE] (2017), a Data Scientist at Amazon, explains that "once item pairs have been identified as having a positive relationship, recommendations can be made to customers in order to increase sales … hopefully, along the way, [to] also introduce customers to items they never would have tried before or even imagined existed" (conclusion, para. 1).

Without access to Instacart's internal data or any external source related to the competition's dataset to draw relations from, Kagglers turned inward, using feature engineering to derive information from all kinds of correlations, associations, and cross-validation they could find between users, products, aisles, departments, time of the day, and day of the week people ordered their groceries. Feature engineering consists of four main steps: creation, transformations, extraction, and selection. This process is used to augment the dataset, as well as to select and transform the most relevant variables from datasets when creating predictive models using machine learning or statistical modelling. Creating features involves identifying the most useful variables in the predictive model. It is a subjective process and often requires human intervention and creativity,[24] allied with the domain knowledge about the dataset, which, in the Instacart case, involves eating habits, shopping behaviour, and marking strategy. Existing features are mixed (added, subtracted, sliced, interweaved) to create new derived features with greater predictive power. Transformation involves manipulating the predictors to improve model performance, such as ensuring the model fits the data it ingests, is on the same scale, within an acceptable range for the model, and improves accuracy. Feature extraction is the automatic creation of variables based on raw data to reduce the volume of data into a more manageable set for modelling. Some feature extraction methods include text analytics, cluster analysis, and edge detection algorithms.
Lastly, feature selection is made by analyzing, judging, and ranking various features to determine which ones are irrelevant or redundant, and which ones are most useful for the model to target a pre-established goal.

Recognizing that the right set of rules of association and correlations would win the competition, Kagglers spent most of their time on feature creation to find the best predictors for user shopping

[24] Tools for automated feature engineering have been available in some machine-learning software since 2016, when deep learning methods became widespread (Heavy.ai, n.d.).

behaviours. Each of these features must be tested, weighted, and ranked. They can produce correlations that can increase or decrease the degree of certainty and, therefore, the model's reliability. While competitors rarely share their feature engineering strategies, some participants are more open about their process because they want to show off their skills or seek confirmation from more knowledgeable peers. Some of the associations found on Instacart's dataset are not surprising and easy to make using common sense, such as a direct correlation between one flavour of an item being purchased with another flavour from the same item family (e.g., Strawberry Chia Cottage Cheese and Blueberry Açai Cottage Cheese, or Chicken Cat Food and Turkey Cat Food). Dipak Tiwari (2021) listed dozens of such combinations he used to boost his model: the number of reordering per aisle and department; the number of times a user has reordered a particular product in the past; and ranked timeframes (by day and by the hour) indicating when particular users do their groceries, to cite a few. These associations may validate and corroborate previous knowledge about these individuals, such as the case of the regularities and recurrences of weekly and monthly shopping.

Other correlations might sound odd and surprising. Spachtholz (2017a) made an association between the time of the last order and the probability of reordering: if people order again on the same day, they order the same product more often; otherwise, when 30 days have passed, they tend to try something new. Arcady27 (2017) went in the same direction, making associations related to changes in user behaviour: the differences between the last and first order, the sum of the first and last three orders and the number of unique products in the basket. In some cases, these correlations might be trivial, invisible, and imperceptible because they reveal patterns of behaviour and habits that individuals do not register themselves. These can be at the micro level or the subjective order; such as things that might sound inconsequential to most of us, like picking up a product because of its colour and always getting eggs when buying milk. On the other hand, it can also occur at a macro level, triggered by marketing campaigns that push or replace products on supermarket shelves; or based on many social and cultural biases, such as who is responsible for doing the groceries (gender divider), family affordance (class and economic divider), and market segregation (racial and ethnic divider). As reported by yananchen (2017), a Machine Learning Engineer at Alibaba, Kagglers were using hundreds of these features to find the sweet spot where they could reveal the "truth" about each user, or at least how predictable consumer behaviour can be.

Even more interesting, feature engineering can produce hallucinations—a kind of spurious correlation madness that make anything seem plausible. Consider the strategies used by Sean

Vasquez ("sjv"), an undergraduate at MIT who made it to third place in the competition. Instead of manually finding correlations, he "extracted" features using neural networks. Most of these features are, however, statistical correlations without any connection to how we subjectively do groceries or objectively correlate users, items, food, and shopping times together. Sourabh Jha [svj24] (2017) followed the same path, finding a correlation between the product's name length and the quantity and regularity at which the product is bought: "It seems weird why this should matter, but it did improve the model a little bit" (n.p.). The rarity of the product's name—how different it is from the rest of the corpus—might also boost the model, making it an essential factor in explaining what type of product a person purchases. These findings, combined with frequencies, quantities, and distributions, might generate user profiles that might prefer to buy products with longer names on Wednesday nights or customers with a taste for products with exotic names that only buy on odd days. These speculative phenomena about user behaviours might sound unfamiliar, strange, or just random, but they produce statistically valid correlations useful for a particular predictive model. Vigen (2015) reminds us how easily we can put together two unrelated facts that create absurd, illogical, and sometimes hilarious accounts of the world we live. His website[25] contains 30,000 similar examples, such as the annual number of people who drowned by falling into a pool, which correlates with films starring Nicolas Cage, or Apple iPhone sales correlating with U.S. uranium exports. On Kaggle, however, feature engineering is used as part of a competitive game to find the best predictable variables: *correlation conflates with causation* if they help create a more robust predictive model.

While there is a danger of creating factoids about individuals—assumptions reported and repeated until they become accepted as facts—feature engineering is considered the "holy grail" for recommendation systems (Tenorio [DATATHÈQUE], 2017). According to Rao, Instacart's Director of Engineering, the features with the most positive impact on the model are subsequently used to discriminate users into "bins," or categories of users, which is the basis of market basket analysis for hyper-profiling and the recommendation system used on Instacart and most of its marketing strategies (Stanley, 2017c). While traditional marketing strategy defines these categories solely based on demography data, digital platforms like Instacart use complex relationships to create dynamic and granular "bins" that can be rearranged on demand. Therefore, at any time, Instacart can aggregate users in different ways, such as shopping behaviour, dietary restriction, food preference, region, and time frame. In other words, such a model produces "kinds of users,"

[25] See more at https://www.tylervigen.com/spurious-correlations

discriminating them into workable, computable, and programable categories. For instance, the model could identify the "opportunist user" who orders once and never comes back, most likely because of a promotional offer via social media; the "subscriber buyer" who orders based on a time frame (every day, week, month); the "organic consumer" who only buys organic food, either by choice or dietary restriction; "parents of young kids" who have a disproportionate amount of items related to children; the "brand loyalist" who only buys products from a single brand; or the "smoothie lover" who has only fruits and yogurt in their cart.

With so many user archetypes, Instacart can target them individually by recommending products that "matter the most" and ranking them by relevance according to the context and the consumer profile. Stanley (2017a) explains why this model matters from a marketing perspective:

> By observing how our shoppers have picked millions of customer orders through our app, we have built models that predict the sequences our fastest shoppers will follow. Then, when a shopper is given a new order to pick, we use this predicted fastest sequence to sort the items for them. This approach has *reduced our shopping times by minutes per trip* [emphasis added]. (Enter deep learning section, para. 2)

Instacart uses these dynamic hyper-profiles to match, sort, and rank merchandising across their services to output several contextual recommendations, including ad targeting and repurchase modelling, which is the primary purposes of the competition: "to be able to sort the list of items that a user has purchased in the past in their 'buy it again' aisle" (Stanley, 2017c, n.p.). From Instacart's perspective, the best predictive models would allow the company to nudge their customers to buy more often by offering a convenient pre-selected grocery list that not only fits with their past shopping behaviour—giving them exactly what they are used to buying—but also allows marketing strategies to push new products into consumer habits.

8.2.3. Standing on the Shoulders of Crowdsourced Free Labour

While the Instacart Market Basket Analysis competition aimed to reward the best algorithms for solving a formulaic business problem, a different kind of game was being played among Kagglers: merit and reputation. The world created by the logic embedded in digital platforms is undoubtedly reflected in Kaggle and its community of data scientists and computer engineers. Kagglers perceived their role as solving problems on their own merits without specific domain knowledge about the issues at hand. As with everything else in the data science community, this ingenuity must be measurable, ranked, and publicly displayed. The user ranking serves not just to define the winner of each competition but also as a way to recognize their efforts with virtual gold medals and titles such as "notebooks expert," "datasets master," and "competitions grandmaster." As Han (2017) reminds us, in the digital era, "everything is made comparable and measurable and subjected to the logic of the market" (p. 29). As much as the code is crafted, the effort of each competitor follows a logic of self-optimization, seeking never-ending improvement. Nevertheless, this self-optimization follows from systemic constraint—from the logic of quantifying productivity and success. The crowdsource work done by competitors on Kaggle is captured by the neoliberal regime not just as product-added value (the code, the models, the predictions) but as ways to demonstrate individual self-realization and self-optimization, the tenet values in late-capitalist society.

Kagglers genuinely believe in an objective meritocratic system. Similar to their machine-learning accuracy score, the ranking reflects their skills and proper place in society. Indeed, the title, the medals, the points accumulated, and the number of contributions are critical for employability. In a hyper-competitive environment, such as the data science and machine-learning market, Kagglers will do whatever is necessary to stay ahead of others in the game, no matter the consequences. It came as a surprise for the community when Vladimir Ovsyannikov ("sh1ng"), Senior Software Engineer at H2O.ai, released his complete solution in the forum a week prior to the end of the competition. In a thread titled "Baseline 0.4029970," sh1ng (2017) shared the code that would put any competitor in the top 10% and within reach of getting a medal in the competition. The post attracted hundreds of mixed-feeling responses. Enthusiastic users praised his altruistic gesture, pointing out the collaborative nature of Kaggle's community in producing high-quality work: sh1ng's code could be used as a stepping stone to a better solution. Others responded with accusations, criticizing his actions as anti-competitive and unethical: "R.I.P. competition," declared hippskill (2017), a participant from Moscow.

Participants were concerned about the competition's fair play and how they would have an accurate measure of their skills. The ranking is not just a scoreboard but the core of Kaggle's meritocratic system that can be used outside the platform to prove the value of users' reputations. Spachtholz's (2017b) protest illustrates how the community felt when sh1ng put their reputation in danger: "… could you please consider that for some people (like me) competition rankings are in fact important (e.g., for applying for a data science job), as they demonstrate effort and skill?" (n.p.). Though sh1ng (2017) reacted by saying that "Fancy medals, rating etc is a [sic] bullshit" (n.p.), he is an active and long-term user on Kaggle, proudly displaying the title "Competition Master" on his profile. In the same post, he admitted sharing his code because he would not get the money prize using his approach; otherwise, he would not have done so. Despite hundreds of complaints about sh1ng's post, neither Kaggle nor Instacart intervened or made any public statement. Indeed, oversharing, recycling code, and plain-sight cheating have been part of the platform since its inception. Kaggle had addressed this problem earlier in a post on its defunct blog:

> Public sharing of code and tips during competitions is encouraged when the objective is educating – or getting feedback from – community members. Publicly sharing high-performing code that creates competition submissions should not happen in the last week of a competition, since it's unlikely that participants will have the time to understand the shared code and ideas. (Cukierski, 2015, Diluted merit, para. 3)

While Kaggle recognizes the problem, the rules for each competition do not address the issue, making Cukierski's blog posts work more as unofficial guidelines and best practices than as an instrument to discipline possible infractions or misbehaving.

This example, together with other cases involving cheating (Gordon, 2020; Quach, 2020a), copyright infringement and piracy (Synced. 2020), and privacy violation (Chowdhury, 2017), reveals that the lack of intervention by Kaggle is by design. Despite all the hype and discourse about high technology and automation, the AI industry cannot survive without free labour and open access to data and code. It not only thrives on but requires a large number of people willing to share their work for free in order to create economic value, technological progress, and reinforcement of a supposed meritocratic system in our society. Langenkamp and Yue (2022) estimate that the economic value generated by Machine Learning Open-Source Software (MLOSS) alone is 100:1, which is much higher than the 7:1 ratio observed by the Defence Science and Technology Laboratory (DSTL, 2017) competition in 2017 discussed in the previous chapter. According to Langenkamp and Yue (2022), if the AI industry provided at least 100 billion dollars to the global economy in 2022, big tech companies and non-profit foundations spend between 100 and 300 million dollars per year (p. 388)[26] to develop and maintain machine-learning tools like PyTorch, sponsored by Meta, and TensorFlow, backed by Google.

[26] Though I disagree with the metric used in Langenkamp and Yue's (2022) study to measure quantity and quality of work (number of commits and number of lines changed in the software repository), the authors reveal how the quantification of free labour and the platform logic became ingrained into business efficiency. Creating economic value on top of free digital labour is an ongoing trend and has been discussed by many scholars in the last thirty years. See, for example, Terranova (2004) and Srnicek (2017).

Despite the supposedly altruistic discourse to fund and promote open-source technology, tech companies use this model as a predatory practice to encourage users to actively work on and maintain large open-source initiatives not as employees but as "collaborators," creating a sense of community that shares the same values. While this practice produces a sense of ownership and collectiveness among developers, the intricacies of open-source licensing and commercial use play in favour of whoever sponsors the project, leaving developers locked in specific technologies. Open-source software also facilitates mass adoption, asserting success in the sociotechnical ecosystem. Well known to the Kaggle community, these tools are always present in every competition and have become so essential to machine-learning development that engineers cannot live without them. Langenkamp and Yue's (2022) study includes plenty of examples of how entrenched these tools are in the machine-learning community: "Because those tools are there, I can just grab the person's implementation… If PyTorch or an equivalent didn't exist, I would literally have to go back into the papers, which would be pretty bad" (p. 391). Kagglers are only able to operate as machine-learning engineers because of a thriving open-source community, which they themselves become part of.

Nonetheless, in the name of supposed scientific progress and a chance to get a good job or earn extra money, Kagglers are willing to spend their time and resources for free to help a large corporation make more profit. Only when there is a disruption in their workflow, such as the code oversharing discussed above, users may realize the community has been exploited and that they have little to gain from how Kaggle runs its competition. In a thread titled "This competition is disappointing," AlphaMeow (2017) exposed their dissatisfaction with the competition: "it's hard for me because I spent days and nights for more than two months on this project and someone just released it as a gaining for free before the end" (n.p.). Most of the comments on the thread expressed a similar opinion: "I am dumbfounded!!! After a lot of hard work, I am waiting for my model to finish running so that I can submit and see the score and then this?" (Sheriff-Hussaini, 2017, n.p.). Yet, many Kagglers dismissed his comments. "The best professionals will adapt and win from most sharing," shouted Charles Jansen (2017), Head of DeFi Transformation at S&P Global Ratings. Indeed, as observed by Dmitriy Guller (2017), Lead Data Scientist at New York Life, open source and gamification—the two main principles of Kaggle as a platform—are primarily incompatible and are starting to cannibalize each other.

The shortsightedness and self-centredness of the community also prevent them from helping others in similar situations related to the precariousness of work conditions. The ubiquitous rhetoric for increased efficiency and optimization among data scientists rarely strives to improve working conditions or pay fair wages to employees. It is estimated that 80% of Instacart's 500,000 employees work under a lower wage regime and in precarious conditions (Elejalde-Ruiz, 2018), sometimes earning less than US$1 per hour (Wells, 2017). When the user InstaShopper (2017) requested the community to help Instacart shoppers, who receive US$0.40 per unique item they deliver, no one cared to reply or consider using data science to solve problems afflicting the low-wage side of this multi-sided market. Indeed, when closely examining how Instacart describes its business model (see Figure 8.2), we notice a skewed network where shoppers provide value but receive little in return. The manual labour extracted from these workers is not a concern for the machine-learning community. Indeed, the precariousness of the workforce is not an issue to be solved, as it is perhaps even encouraged as a way to reduce costs. After all, according to Jeremy Stanley (2017a), what matters is to optimize business processes to save the middle-class consumer from the discomfort of an intimidating trip to the grocery store.

8.3. Modulate: Shaping the Self

Despite the dramatic and controversial final weeks, Kaggle and Instacart considered the competition a success. In three months, 2,621 competitors submitted 39,863 entries containing predictions of grocery products an Instacart consumer will purchase again and when they will do so. Nevertheless, the competition ended as it began: there were no blog posts or announcements in the forum. A day after the deadline, the leaderboard was updated, revealing the winners: "Weiwei," a Chinese machine-learning developer working for a similar food delivery service in the Beijing area, who rarely interacted in the forum, received US$12,000; Kazuki "Onodera," a data scientist at Yahoo! JAPAN, took second place, earning US$8,000; Sean Vasquez "sjv," an undergraduate at MIT at the time, received US$5,000 in third place. Except for "Weiwei," who never publicized their algorithm, Kagglers spent the following days congratulating themselves, sharing their solutions, and praising Instacart and Kaggle for the "opportunity."

The US$25,000 offered by Instacart was significantly lower than other high-profile competitions analyzed in this research. Though the company sweetened the prize by offering job interviews to the winners looking to grow their machine-learning team, there is no evidence that these interviews took place, and none of them got hired by Instacart. Furthermore, according to an AMA (short for "ask me anything") session held by Kaggle with Instacart's data scientist team, the algorithms and predictions produced in the contexts of the competitions were merely an exercise to provide the company with ideas for further uses of machine-learning techniques (Stanley, 2017c).

Though the company never planned to integrate any work into its platform due to the complexity of its workflow, it reserved the right to use the winner's model unilaterally in other functionalities if it proved profitable. As with every sponsor on Kaggle, Instacart used the platform for experimentation, capitalizing on the cheap and free labour a crowdsourcing community offers.

The business-oriented question posed by Instacart's challenge, that is, to "reveal" users' shopping behaviours in order to enhance a private company's marketing strategy, gave the event a sense of normality to a community of data scientists that are either employed or looking for employment in the marketing sector. After all, unlike the Deepfake Detection Competition and the Passenger Screening Algorithm Challenge discussed in previous chapters, in which the datasets were synthetically created and built according to specific metrics to respond to particular problems, the Instacart Basket Analysis brought "real-world data collected in the wild" to "solve real world problems" (Stanley, 2017c). Furthermore, while the first two challenges' datasets were difficult to work with due to the complexity of the data type—moving images of fake individuals in the first and matrices of a three-dimensional representation of human bodies in the second, Instacart's data was easy to understand, as it consisted of human-readable tables and lists of trivial and everyday life objects, such as food and department names. The simplicity of the dataset and the triviality of the activity analyzed (i.e., food purchase) allowed participants to feel comfortable in making bold assertions and predictions, as if they had previous knowledge of eating habits, shopping behaviours, retail business, and marketing planning. The predictions produced by Kagglers follow a strict sales orientation, leaving aside all sorts of health and dietary condition as well as the socioeconomic reality of the platform's users. Machine-learning competitions like the Instacart challenge are treated as "business as usual" and can reveal how inconsequential and ill-thought the models developed for trivial business problems can be.

The problem posed by Instacart is not trivial, however. The focus on understanding temporal behaviour patterns makes the problem fairly different from standard item recommendations, wherein user needs and preferences are often assumed to be relatively constant across short periods. For instance, whereas Netflix assumes viewers want to watch another movie similar to the one they just watched, it is less clear that people will want to reorder a fresh batch of almond butter or toilet paper if they bought it yesterday. Nonetheless, Instacart buys into Silicon Valley's discourse of technological disruption and refuses to see this complexity, treating the problem as any other digital platform that uses probabilistic statistics to find convenient methods to increase sales and keep users engaged. Stanley (2017c) describes the platform as a start-up company offering highly innovative products and services to "save customers time, thereby increasing the value of the service, thereby increasing retention, [and] finding new products a customer might love, identifying complementary products while they are shopping, and trigger new purchases through the personalized push notifications" (n.p.).

However, Instacart is not as disruptive or innovative as its data scientists want to believe. On the contrary, Instacart is more interested in commonalities: little actions and repetitions of everyday lives. The focus is not on the unprecedented or extraordinary but on the individual silent and constant habits and behaviours. Here, we are dealing with what Wendy Chun (2016) calls 'habitual media,' which becomes pervasive "by disappearing from consciousness" (p. x). The company actively develops processes to track preconscious routines, from biological rhythms to behavioural habits such as food preference and dietary restriction, together with most common digital actions like time spent scrolling, staring, and clicking to find patterns. The user's body has become an archive from which data scientists can "mine" data and unearth personal habits and where machine-learning engineers can curate the "user experience" to trigger specific psychological and behavioural responses from these individuals.

8.3.1. Habitual Media

Chun (2016) argues that "habits are creative anticipations based on past repetitions that make network maps of the historical future. Through habits, networks are scaled, for individual tics become indications of collective inclinations" (p. 3). Habit is information: it in-forms and connects. As such, in the minds of machine-learning engineers, habits can be described as data points and zones of influence across individuals, things, and environments. They can be collected, analyzed, manipulated, and programmed, most commonly to produce positive psychological responses to advertisements (Zuboff, 2020). Large-scale psychological manipulations are not a novelty in social media platforms. The infamous 2014 Facebook mood experiment demonstrated how users' emotions might be influenced by how their newsfeed was filtered (Adams Kramer et al., 2014). The study showed Facebook's new ability to control user emotions in ways that users would not be aware of. The platform purposely intervenes in how users consume information and interact with the platform, nudging them to act and react to specific types of content and, as a result, producing new media habits.

Consider how we engage with Netflix's home screen: users examine each title for only 1.8 seconds and give up if they do not find anything interesting in a minute and a half (Netflix, 2016). To remediate the so-called "analysis paralysis," Netflix has customized its tiles based on a user's watching habits. For instance, if a viewer watches many romance films, the company might serve a thumbnail for *Good Will Hunting* that features Matt Damon and Minnie Driver kissing. If the viewer prefers comedy, the *Good Will Hunting* thumbnail might feature Robin Williams with a smiley face (Chandrashekar et al., 2017). The same goes for the sense of belonging and identity. Black people are more likely to see black performers in their thumbnails, even if these artists have only minor roles in the movie or TV show (Tiku, 2018). In other words, race and gender also play a part in Netflix's and other platforms' recommendation algorithms. In this sense, the Instacart challenge follows the same strategy, featuring and ranking products based on user preferences and past purchases, narrowing the user choice, and impacting what the user decides to buy. As Crary (2014) wrote in his book *24/7: Late Capitalism and the Ends of Sleep*, attention is no longer the product of (semi-)independent rating firms that measure how many eyeballs a program attracts every minute. Instead, attention is measured by the same systems that also produce and distribute content, organize and rank a video's display, connect ads to content, and attune the algorithms that connect content to advertisers.

Moreover, Chun (2016) argues that habit has moved from *habes* (to have) to *addictio* (to lose—to be forfeited to one's creditor): "A habit is now a form of dependency, a condition of debt" (p. 4), where a user get addicted to specific conditions of the platform. For instance, because a business model focuses on the attention economy, digital platforms often use deceptive "click-bait" in their recommendation systems to play with users' emotions and entice them into engaging with the platform. YouTube, for instance, spent years experimenting with different monetization formulas and machine-learning techniques combining user data, search trends, advertising, and video content. By exploiting connectivity as a resource (van Dijck, 2013), the company developed several strategies to capitalize on streaming video's viral marketing potential. According to YouTube Chief Product Officer Neal Mohan, more than 70% of the time spent on the platform is driven by the service's algorithms (Solsman, 2018). In other words, when users visit YouTube, the site uses predictive models similar to the ones created on the Instacart challenge to lead viewers into a path filled with the videos it "thinks" they want to watch. The content is irrelevant as long as the users stay on the platform to watch yet another video.

These recommendations and nudges are not always directly related to what we generally understand as a product (material goods or media consumption). They go beyond the regular retail coupons or the repeated call for action like "buy now." More and more, the individuals themselves become advertised to other individuals, where the marketers aim to exploit the social relations and the individuals' attention and affect. Digital platforms transform the integral human being into an object of exploitation. Take, for example, Facebook's People You May Know (PYMK), which was launched in 2008 as an instrument to help users find other people they know on the platform. In the announcement, the company made the new feature seem a fun and whimsical way to make new friends and reconnect with old ones. Though a vital tool to densify social relations and intensify the exchange of information, it quickly became one of the most hated features of the platform. Users complained that Facebook constantly shows unknown and strange people they did not want to friend, suggests that a psychiatrist's patients friend one another (Hill, 2018), and even recommends rapists as people they should befriend, as Kevin Kantor (2021) recollects in his poem. Fifteen years after its launch, Facebook's PYMK continues to creep users, as described by the journalist Stevie Martin (2023):

> having a brief scan over my own current suggested friends, there are at least three people there that I haven't met, I've never worked with, we didn't go to the same school, I have no mutual friends with and they're not in my contacts. One of them is someone I'm fairly sure, after doing a quick search on other social media sites, is currently seeing my ex. So why is she being suggested?! … I don't know her! Were we once in the same cafe? Has she, completely understandably, been stalking me? (Para. 3)

Combined with machine learning and behavioural data, targeted advertisement becomes a parasite encroaching on the individual profile. Still, it is not necessarily true that Facebook and other digital platforms are trying to figure out which people are seeing which doctors or who one's ex-boyfriend is dating. On the other hand, it is beneficial to advertisers to know which "micro-segment" a person falls into to target ads more efficiently. The Instacart challenge shows that the more advertisers learn about individuals, the more they can infer their spending consumption and habits. Gender, age, race, income, relationship status, behaviours, habits, preferences, desires and even the unconscious; all come into play when serving ads.

8.3.2. Radical Subjectivation

While companies' and digital platforms' most common goal is to "turn all users into money," there is no reason why hyper-profiles and hyper nudges cannot be used for other ends. In other words, though marketing and advertising were, and still are, the driving force behind behavioural data surplus, it is merely incidental to the method of how it operates. With the knowledge of individuals' personalities and collective behaviours produced by predictive models, digital platforms hold the power to weaponize radicalized groups and manipulate the audience, posing a threat to an individual's well-being, freedom, and even life. In this sense, Han (2017) argues that neoliberal psychopolitics is dominated by positivity. That is, rather than disciplining behaviour or censoring material, it works with positive stimuli and freedom of speech: "Instead of administering 'bitter medicine,' it enlists Linking" (p. 37); instead of containing discriminatory behaviour, it incites violence and hate. In other words, it indulges the individual's psyche instead of paralyzing it with shocks. In their crusade to cut costs and generate profit, digital platforms and big tech corporations see the exploitation of human life as a practical path to economic success.

Thus, at least since 2016, the strategies to intensify engagement in the digital space have become more controversial. A more relaxed approach to content moderation and circulation was adopted to encourage "binge-watching" behaviour, where platforms push a continuous flow of information to the users via a recommendation system. This strategy not only facilitates the production and circulation of radicalized content but also encourages the dissemination of disinformation with inflammatory material, such as conspiracy theories, hate speech against minorities, white supremacist rants, and science denials (anti-vaccine, flat-earth), among a long list of "fake news" that "auto-plays" or "pop ups" on newsfeeds under the label "recommended for you." "Neoliberal psychopolitics seduces the soul" (Han, 2017, p. 37), as the algorithms seem to be built to please and fulfill a sadistic and vicious agenda, not to repress it. Indeed, the directions taken by these algorithms are not given by chance but as a reflection of our society's inclinations—at least the portion represented by the datasets used to train these machines—as it formalizes these desires, needs, and wishes into streams of code, predictions, posts, videos, and memes.

Algorithmically driven disinformation and manipulations exploiting users' psychological and behavioural inclinations have been observed again and again for economic and political gain. The most infamous case so far is the efficiency of Cambridge Analytica's manipulation campaigns (Davies, 2015) to undermine democratic processes by spreading disinformation in many countries, including in the 2016 "Brexit" referendum, the 2016 U.S. election (González, 2017; Stark, 2018), and the 2018 Brazil elections (Ituassu et al., 2019; Paiva, 2020). Tufekci (2018) states that we are witnessing the computational exploitation of a natural human desire, digging deeper into something that affects and engages us: "As we click and click, we are carried along by the exciting sensation of uncovering more secrets and deeper truths. YouTube leads viewers down a rabbit hole of extremism, while Google racks up the ad sales" (para. 15).

While digital platforms argue that this is a "non-issue," usually hiding behind the catch-all concept of "free speech," examples of discriminatory usage of recommendation systems pile up: Facebook allows advertisers to target their audience based on race and gender, essentially discriminating against users from minority groups (Angwin & Parris, 2016); TikTok gives less visibility to black creators (Asare, 2020); LinkedIn recommends more men than women for open roles simply because "men are often more aggressive at seeking out new opportunities" (Wall & Schellmann, 2021); and X (formerly Twitter) allows hate speech to circulate freely on the platform as a way to generate more engagement (Frenkel & Conger, 2022). Favouring their interests and against its own policies, Google, Youtube, Meta, and Spotify manipulated their recommendation systems and sponsored posts to disseminate disinformation about the Brazilian government's attempt to regulate digital platforms (Bischoff, 2023; Netlab, 2023). Furthermore, when these companies are forced to recognize wrongdoing—due to bad publicity or by the force of law—and "fix" the technology, they tend to create other forms of discrimination. For instance, Facebook's race-blind practices concerning hate speech came at the expense of black users because of pressure made by shareholders and investors, more specifically from from conservative partners (Dwoskin, 2021)

The convenience and practicality sold by digital platforms put individuals in a situation where they are led to believe they are freely and voluntarily using the tools and services provided. However, by doing so, they are subjected to predictive algorithms, that is, the choices made by the developer in building the algorithm. We surrender our "selves" to the platform, not only giving up our personal experiences, captured in data streams to produce copies of ourselves (i.e., hyper-profiles, avatars), but also renouncing the freedom to (or the burden of) making decisions since we are convinced that a predictive algorithm will choose and deliver precisely what we want or what is best for us. We can see how this relationship has been exploited using behavioural reflectivity presented on recommendation algorithms, in which not only the user/target is presented with an item or action for their consideration, but also the options shaping the future actions of this user. That is, these algorithms are purposely designed to give the users a tiny push—a hyper nudge—so they can "voluntarily" take a specific path within a series of restrictions and constraints.

Foucault (2017) reminds us that subjectivation is not just the relations we have with our own individuality "but our relationship to others inasmuch as they are also ourselves" (p. 12). In other words, we are constituted by and within the relations we have with others and with digital systems. Therefore, there is no solipsism in the Foucauldian proposition of subjectivation, but an acknowledgement that the practices of the self are not isolated from the external world; that is, they are also social practices (Ferreira Neto, 2018, p. 14). This explains the strategies and tactics used by both machine-learning engineers and interface architects in offering hyper nudges as a way to put users in proximity to each other. The ceaseless pursuit of an accelerated and intense flow of interaction and engagement with other users on the platform seeks to capitalize on social dynamics in which the individual would find support for their own opinions and attitudes, not in the dry abstraction of a faceless algorithm, or from a supposed artificial intelligence that imitates human behaviour, but from other users like them. This process could bring to light collective patterns of behaviour individuals are unaware of, possibly rendering some form of shared collectivity, which Han (2017) calls the *digital unconscious* (p. 65). As such, digital psychopolitics would be in the position to take control of mass behaviour at a level that escapes detection.

8.3.3. Remote Control

As in the case of Facebook's Deepfake Detection Challenge and the Passenger Screening Algorithm Challenge, the value of mobilizing data and code to push forward machine learning and predictive models lies not in the prizes offered by Instacart, which was not enough to pay an average data scientist a month's salary. The value of these competitions and machine learning models lies instead in the behavioural patterns that can be deduced or unveiled through the normalization, aggregation, augmentation, and ranking of detailed collateral data about individuals. These patterns are promptly incorporated into a hyper-realistic model of a person's preferences, behaviours, and desires, producing hyper profiles—a digital *Doppelgänger* that *looks like you, talks like you, and behaves like you, but is not quite you*. Yet, these collections of hyper-profiles are enough to be used as cheap proxies of human beings to exploit individuals' sensitivities, personality traits, desires, and needs. The value in creating hyper-profiles of every individual is the power to *know who they are* in order to fit them into arbitrary categories for specific purposes: increase sales, optimize services, create experiences, manipulate public opinion, trigger desires, and, ultimately, shape behaviours.

Recommendation systems are often praised as the best solution for information overload. It is a response to the problem of the sheer scale of big data, reducing individuals' cognitive overload and aiding the decision-making processes. However, this problem of scale has been used to justify an authoritative role in which the algorithms automatize the controls of data flow, providing specific routes of actions and curated options for the users. As a result, digital platforms become gatekeepers with the power to limit access to information and modulate individuals' behaviour. What is at stake in this modality of data-code mobilization is not in the algorithm's accuracy in predicting users' behaviour. Instead, the value of recommendation systems is to push users in a specific direction, to "nudge" them to take specific actions. If predictive algorithms cannot be precise enough to predict the future and tell us *what is going to happen*, recommendation systems must answer *what should happen* to ensure a desired outcome. If predictions are optimistic, the recommendation aims to keep the user in a loop, serving the same content, giving the same options, and repeating the same actions. Otherwise, if predictions are pessimistic, the recommendations will try to "correct" users' behaviour and remediate the course of action by shuffling options, hiding actions, and serving different content—radicalizing, if necessary, in order to nudge these individuals to a more "positive" outcome.

The question of *what should happen* is a recurring theme in Kaggle's competitions, focusing on developing recommendation algorithms. For instance, Santander (2016B) held a US$60,000 competition to build a recommendation system that could better "meet the individual needs of all customers and ensure their satisfaction" (n.p.) no matter where they are in life. A few months before the Instacart competition, in 2017, Outbrain (2017), a company specializing in recommendation platforms, promoted a US$25,000 competition to test the efficiency of their own algorithm, asking Kagglers to "predict which recommended content each user will click" (n.p.). That same year, Corporación Favorita (2018), a large brick-and-mortar Ecuadorian supermarket chain, offered US$30,000 for competitors to build an accurate sales predictor so the company could "please customers by having just enough of the right products" (n.p.) at the right time. In 2018, using a slightly different approach, Mercari (2018), the largest community-powered marketplace in Japan, challenged Kagglers to build an algorithm that "automatically suggests product prices to online sellers" (n.p.) in a US$100,000 competition. In 2019, Elo (2019), a Brazilian financial services company, offered US$50,000 for competitors to develop algorithms to "identify and serve the most relevant opportunities to individuals" (n.p.) by uncovering signals in customer loyalty. Beyond retail, finance, and e-commerce, Kaggle also held competition related to the flight industry in 2012, when General Electric (2013) offered a staggering US$250,000 in a two-phase multiyear challenge to build a recommender system that "optimizes flight routes based on current weather and traffic" (n.p.). Lastly, since 2018, the American National Football League (NFL) has been hosting yearly competitions looking for algorithms that suggest rules to improve player safety during punt plays and recommend defensive performance on passing plays (NFL, 2019; 2021).

8.4. Summary

The Instacart challenge is a dry, conventional, and mundane example compared to other competitions held on Kaggle analyzed in this research. Creating recommendation systems for pre-filling customers' carts with recurring products for their next purchase is not as eccentric and controversial as identifying deepfakes or predicting threat objects using scanners. Admittedly, top Kagglers were neither impressed with the problem posed by Instacart nor excited to work with grocery lists. Yet, perhaps this lack of "wonder" allows us to see what truly drives machine-learning development. Most of the competitions on Kaggle are similar to those sponsored by Instacart: they look for practical use cases of the technology to improve the business model. In this sense, developing automatic methods capable of capturing individual attention and guiding them to carry out specific tasks or act in a particular way is a priority in forums like Kaggle or any other company focused on employing artificial intelligence in their business. Recommendation systems, in particular, are very popular because they can interpellate individuals as subjects of the platform, affecting how they will act in front of pre-selected choices. As such, the exploitation of human behaviour and psychological traits combined with ordinary everyday life activities, such as doing groceries, has become instrumental in fuelling predictive models to fulfill a specific economic agenda.

Knowing what users want and need, their feelings, wishes and desires, and how they behave is not trivial. Nevertheless, data scientists, machine-learning engineers, and marketers seem to be very attached to the nineteenth-century epistemology proclaimed by Galton (1879): "Until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science" (p. 149). That is, the machine-learning community believes that subjective and psychological qualities are (and must) be measurable, computable, and even programable. Measuring individuals' psychological traits has become a hallmark of the digital era (Han, 2017). Having access to the right set of data about users' behaviour and habits has the power to reveal social patterns that not even the individuals themselves are aware of. Foreseeing the economic and political advantages of this type of knowledge, digital platforms began using behavioural data to create hyper-profiles that not only *re-present*, but also *re-act* on behalf of every single user, customer, or citizen.

However, hyper-profiling individuals is often hidden from the public eye because it destroys the reciprocity and trust between users and platforms. This operation requires significant amounts of subject material, often mined from everyday life transactions as behavioural surplus. Data, personal and subjective data in particular, is the most precious commodity in the digital economy. Private companies and public agencies avoid speaking about this practice as it might implicate them in criminal activities, since individuals do not necessarily agree to release their data to fuel machine-learning algorithms. The lack of regulation and public policy makes this a sensitive topic even in machine-learning specialized forums like Kaggle, and its contentions rarely surface on the platform. Instead, each competition on Kaggle that involves data extracted from individuals is cautiously crafted around the idea of "improving user experience" (Instacart, 2017, n.p.), "delivering what the user wants when they need [it]" (Santander, 2016, n.p.) in order "to fulfill user personal wishes and desires" (Springleaf, 2015, n.p.) and "help them take control of their lives back" (Deepfake Detection Challenge, 2019a, n.p.). The supposedly good intentions and the focus on user advantage conceal the irregularities behind the dataset and the objectives of the existence of the dataset in the first place.

Instacart capitalizes on behavioural data and hyper-profiles to optimize its platform. They appeal to a specific audience accustomed to living in excess and surrounded by convenient services. Instacart aims to make the shopping experience more personalized but also more automatic by anticipating their customer's needs and desires: "Imagine, for example, having milk ready to be added to your cart right when you run out, or knowing that it's time to stock up again on your favorite ice cream" (Kaggle Team, 2017, para. 1). To that end, Instacart needs to understand its user's behaviours and classify them into useful categories in order to deliver a curated list of items that fit that user's needs. The competition the company held on Kaggle gives us a glimpse into the amount and diversity of data the company has on their users. Through the millions of grocery items purchased by thousands of users, Instacart is able to make inferences about their eating habits, such as how often people buy specific products, their food preferences and brand loyalty, and how many calories and other nutrients people consume. It is no surprise that the list of popular products contains many industrialized items common to the North American market, but also perishable items such as produce and fruits, which are bought more frequently due to their rapid decay: bananas may be the winner, but frozen pizza and ice cream are certainly the most popular.

Nonetheless, Instacart is not interested in its users' wellness and fitness or perhaps never even had an interest in what type of food people buy. Rather, it has a sharp focus on marketing and advertisement. The competition was an experiment to provide ideation among the company's data science team (Stanley, 2017c) in order to improve their recommendation system. Accordingly, Kagglers dive deep and fiercely into feature creation through association because this technique is considered the "Holy Grail" for recommendation systems (Tenorio [DATATHÈQUE], 2017). The community explored every corner of the dataset, looking for all kinds of relationships, correlations, associations and cross-validations that could enhance their predictive models. The search for the best predictor, a sweet spot that could reveal the "truth" of each user, or at least how predictable they can be, led participants to the extreme, as they used machine learning to produce spurious correlations and "hallucinations" about user behaviour. The highly speculative nature of these experiments has not stopped Kagglers from submitting their models, since it is precisely what Instacart was looking for: "predictive models that produce kinds of users," discriminating and aggregating them into workable, computable, and programable categories. These models serve as proxies for building recommendation systems primarily oriented to nudge consumers to increase engagement with the platform and buy more products, but with the power to shape individual behaviour and desires.

Furthermore, neither Instacart nor Kaggle's community cares about the individuals recorded on spreadsheets made available in the competition. Discussions on the forum reveal little to no regard for Instacart customer' financial and social conditions nor for the working conditions of the shoppers responsible for doing picking up and delivering the products. It is all about feature engineering, finding the best predictors, enhancing the shopping experience, and selling more items. The socio-techno apparatuses involved in creating predictive models and recommendation systems can materialize new and actionable subjects from their personal information. Yet, the individuals whose data were extracted remain invisible and discarded. Mobilizing data for hyper nudges means building a hyper profile for each user, deriving their essence from their behavioural data while remaining indifferent to what they do or what is done with them. Currently, recommender algorithms are intrinsically related to increasing sales, improving revenue, and optimizing advertisement. On Kaggle, machine learning is mostly about making users engage more (increase click rate) and converting interactions into transactions (increase consumption). The desired outcome is baked into machine-learning metrics together with the training dataset and the developers' and stakeholders' ideologies, dysconscious biases, and idiosyncrasies. As Turow (2017) observes, predictive algorithms have become a universal way of thinking and have been incorporated into all aspects of our lives, including managing shopping behaviour and dietary habits. The recommendations produced by these algorithms have narrow and specific goals, which, in the case of the current neoliberal agenda, entail profit over and above anything else.

However, beyond the merely transactional and business purposes of the recommendation system, they can also be seen as a new method to obtain what Bentham defined as “Power of Mind over Mind.” That is, depending on the way they are used, these recommendations resemble Foucault’s discipline society, where digital platforms aim to correct and discipline subjects (Foucault, 1995), or Deleuze’s (1992) society of control, where algorithmic media dynamically and perpetually modulate subjects’ behaviour from one moment to another. As such, recommendation algorithms and hyper nudges can be used as an instrument of governmentality (Foucault, 1982) to prescribe conduct, modify behaviour, and shape individuals’ experiences in relation to other individuals, their social sphere, and the physical environment (Langlois & Elmer, 2019). As decision-making becomes automatic and less reflective, machine learning and artificial intelligence backed by large corporations are reshaping our sense of self, producing subject-consumers more susceptible to personalized ads and, ultimately, to ideological propaganda. Recommender algorithms are valuable resources for shaping individuals’ behaviour to ensure *what should happen* in a desired future. Platforms use recommendation systems to keep their desired future in check, including the radicalization of content and options served to the user: racism, homophobia, hate speech, white supremacists, and disinformation will be tolerated as long as they are profitable and ensure the economic goal of the platform. In other words, the content is irrelevant, as long as the users stay on the platform to engage with yet another post, video, music, or any other type of interaction. As a result, in the quest to deliver convenient services and easy information access, these platforms are willing to conspire and collude with unethical, unlawful, and illegal practices.

This chapter demonstrated how trivial questions about everyday life, such as what people buy in grocery stores and apparently naive datasets with shopping lists, can be used to produce complex and powerful recommendation systems capable of shaping individual behaviours. Datafication of the self exploits both the individual’s body (anthropometrics, biometrics) and behaviour to produce subjects of the platform: agents that feed the machine who are simultaneously targets of intense streams of content, pushing them to keep engaging with the system. As Han (2017) argues, “Neoliberal psychopolitics seduces the soul” (p. 37) as algorithms seem to be built to trigger specific emotions and effects to get individuals doing or thinking in particular ways. As such, the value of creating hyper-profiles and recommendation systems lies not in the static datasets collected by data brokers, shared among data scientists and developers, and sold to third-party companies. Rather, the value lies at the intersection of code and data that produce models of users’ behaviour signatures. These signatures are the fuel for a more complex operation that will attempt to answer questions about the future, forecasting events in order to find ways to intervene and shape individual and collective behaviour.

9. Conclusion

> Privacy is nothing that worries me… I think that *people often behave better when they have the sense that their actions are being watched* [emphasis added]. (Anthony Goldbloom in World Economic Forum, 2015, 38:22 - 38:45)

The World Economic Forum (WEF, 2015) promoted a round table titled "Davos 2015—A Brave New World," where five guest speakers—representing the tech industry, the third sector, and from academia—were invited to answer questions and contribute to the debate about the impact of recent advances in Artificial Intelligence (AI), smart sensors, and social technology in our lives. Industry representatives repeated platitudes regarding industrial automatization, increased productivity in the job market, convenience for consumers, and, of course, new business opportunities. The main concern raised by third-sector representatives was the lack of privacy or even a disregard for privacy rights in the tech industry. Kenneth Roth, Executive Director at Human Rights Watch, reminded the participants that the logic behind AI and digital platforms lies in the lack of regulation regarding the large-scale data mining of personal (and impersonal) data. He contended that we currently have no choice but to let Internet companies have access to our data as a way to, among other things, feed to machine-learning algorithms to produce optimized solutions. Anthony Goldbloom, Kaggle's CEO and one of the guest speakers, dismissed the problem, arguing that the "Google generation is less sensitive to privacy" (WEF, 2015, 37:38). He added that he would gladly exchange his privacy for a 50% off coupon on a scarf or auto-complete functionality while writing an email. Indeed, as Roth put it, the trade-off between privacy and service is at the core of the economic model developed within the digital industry: People give up their privacy and personal data for discounts and convenience services.

What was overlooked during Goldbloom's contribution is the way he ties privacy—or the lack of it—with behaviour. As shown in this chapter's epigraph, Goldbloom insists people are on their best behaviour when they are watched. Indeed, the so-called "Google generation" grew up with a more pragmatic approach to privacy and individual rights, one that has less friction and causes fewer problems for private companies exploring and exploiting personal data. The tech industry sells the digital revolution as a disruption that will improve people's lives, and eventually everybody will be able to profit from the outcomes. However, the critics in the third sector and academia are worried that, once again, human lives have been scraped, torn, and taken apart in the name of an intangible dream of relentless economic growth. It is not simply a transactional process where the individual freely, though most of the time unawares, trades their data for products, services, features, discounts, or virtual goods. In effect, Goldbloom reifies what Zuboff (2020) called "surveillance capitalism."

The future of AI and the digital society envisioned by Kaggle's founders and investors is based on the potential to shape individual behaviour. It is a new process of subjectivation that treats individuals no longer as subjects of value realization but as objects from which raw material is extracted in order to predict and indirectly shape behaviours. Predictions about behaviour are commodities sold for profit to other players (e.g., advertisers, digital platforms, and governments) to target and nudge individuals to do and think things in a specific way. In the minds of the economic elite who gather every year in Davos and the technophiles who believe technology would solve any problem and save humanity, this is a perfectly balanced and desired system.

Remarkably, in the digital revolution, our lives have been reduced to numbers in a spreadsheet and rendered as behavioural data in user-friendly graphs, personalized advertisements, memes, frivolous content, and indifference toward others. We have lost control over our individual and collective rights. There is a sense of instability, disempowerment, and dispossession where "there is no exit, no voice, no loyalty, only helpless, resignation, and psychic numbing" (Zuboff, 2020, p. 94). Yet, digital technologies do not promote anarchy and chaos. On the contrary, the digital revolution aims at social stability. Goldbloom speaks about a society that conforms to certain social norms. The process of automated subjectivation put forward by AI and digital platforms strives for normative conformity and stability achieved through social engineering, using surveillance, training, persuasion, nudges, and violence. It combines the dystopian world portrayed in George Orwell's *Nineteen Eighty-Four* and Aldous Huxley's *Brave New World*. In both these worlds, stability is valued over everything else, particularly anything that could fracture the socio-political-economic order or the citizens' superficial sense of happiness. Digital surveillance and control are used in the same way: to make people conform and even like their unescapable social destiny.

Yes, the AI revolution is about profit and markets. It promises to revolutionize the way we use technology by making it more convenient, increasing business opportunities, and, more importantly, enhancing targeted advertising. But, above all else, it aims to achieve control over the human body and psyche. It is, if nothing else, a major instrument of subjectivation and social stability in the name of automation, optimization, centralization of power, and profitability. AI systems are perfect systems for the maintenance of the libertarian and neoliberal discourse based on a combination of acceleration of labour exploitation, exhaustion of natural resources, the exploitation of the human body and mind, and "accumulation by dispossession" (Harvey, 2012). The AI revolution seeks to control the production of possibilities, reinforcing not only a mode of production but a production of modes of living (Lazzarato, 2004) to which individuals are subject. The production of possibilities is always codified, controlled, and predictable according to the modes of capital valourization. The prevailing concept underlying Goldbloom's contribution is *power*. Power over individuals' behaviour. Power to shape habits, discipline bodies, and govern populations. Power to make individuals discipline themselves—the power of mind over mind, in order to reinforce a specific form of society using a particular set of social norms, one that follows the elites in Davos and the technophiles in Silicon Valley.

I have argued throughout this dissertation that the value of aggregation and mobilization of big data goes beyond commercial interests, making incursions into a deeper and more automated form of subjectivation and mediation via machine-learning algorithms and predictive models. What is valuable on Kaggle, and in the machine-learning community in general, is the control of the production of possibilities in order to ensure the reproduction of a specific type of socioeconomic relations. The three modes of data/code mobilization that emerged from Kaggle's competition provide the ability to answer complex questions to produce docile, useful, and productive subjects.

In *identify*, the value in mobilizing hundreds of competitors, millions of images, and optimized algorithms for object detection, classification, and recognition is to have the power to answer a *what is* type of question in order to *define the ontologies of the world we live in*: what is real and what is fake, what is true and what is false, what is there to be known and what is simply noise. Moreover, when targeting individuals, the inquiry becomes a *who is* type of question to *identify*, *recognize*, and *categorize* individuals as subjects: Who are you? Who is your audience? Who are the customers? Who is a criminal? These questions get increasingly intrusive as they also ask about behaviour signatures, individual preferences and habits, their whereabouts, and all kinds of details about a person's life.

In *predict*, the questions become more speculative, aiming to foresee *what is going to happen next*, looking for the ability to *anticipate* events based on partial and biased accounts of history to *produce specific futures*: Will the stock market rise or fall? Will the customer pay the bills in full and on time? Will a person commit a crime or engage in an act of terrorism? Will my team win the next championship?

Lastly, in *recommend*, the question becomes more concrete over the decision-making process, asking *what should happen*, that is, the ability to *act and intervene* in someone's life to ensure desired outcomes: What are the most relevant offers for this specific prospect? What content should be shown on the user's newsfeed to keep them engaged and happy? What is the best order of items to ensure a specific customer purchase on their next visit? How much should a healthcare plan charge based on a person's health habits? How dangerous can this individual become in an airplane?

These questions demonstrate how, after a decade of competitions and challenges, Kaggle is primarily concerned with promoting machine learning as a catch-all solution for any problem in every domain of knowledge. Data is mobilized for capital gain and stripped of its subjective, contextual, and contentious nature to become an "objective" predictor and the speculative ground through which data scientists build new forms of reality. This practice also reveals important tendencies that permeate the development of predictive algorithms and artificial intelligence, most notably a compulsion to reduce the cost of production, an indifference toward human life, an obsession to control populations and individual bodies, and a desire to produce a predictable future for economic gain. The following sections situate these issues in relation to the findings of this research.

9.1. Gamified Free Labour

The expectation of profit is inscribed into Kaggle's history as part of an expanding digital economy. The company was founded by an economist who believes public finances should adopt a data-driven approach in order to reduce costs and increase profit in the private sector. Antony Goldbloom's proposition to use machine learning to invigorate and expand the deteriorated markets after the 2008 crisis was welcomed by financiers. From its inception, the platform was funded by venture capital firms and other investors linked to neoliberal ideals avid to see what the Australian entrepreneur was capable of. As such, Kaggle's community is business-oriented and permeated by obsessive financial goals that condition and drive the direction in which datasets, algorithms, and predictive models are produced, selected, evaluated, and appraised. Hong (2020) elegantly points out that "the optimism that any and every process can be improved through datafication constitutes a voracious impulse that reveals big data's fundamental affinity with capitalism's search for continual growth" (p. 25).

Conceived as a gamified platform (Whitson, 2013) for crowdsourced machine-learning challenges, Kaggle resembles an MMORPG, attracting millions of developers, data scientists, engineers, and hobbyists to participate in intense technical-cognitive challenges to solve problems with predictive models. The gamification of these activities is attractive to an economic system that champions individualism, competitiveness, and meritocracy as core values, shaping both the work and social conditions where users become subjects of the platform/game. Hence playing the "Kaggle game" is crucial to landing a job in the industry, as companies use the platform ranking to select candidates, who now proudly add their tier title, number of medals, and position in Kaggle Rankings on their CVs.

Whereas most Kagglers are considered skilled workers (with half of the participants holding a Master's or PhD degree) operating in a highly disputed and valued market, they take part in subordinate challenges to exploit their workforce. Like other digital platform users, Kagglers are also part of the "content creator economy." However, instead of creating posts for social media or videos for streaming platforms, they upload data, produce code, make statistical analyses, and create predictive models. Kagglers are a networked public where participants are constantly pressured to produce new and improved algorithms but are keen to work for free using their own equipment to become influencers among their peers. Like datasets and pieces of code, they must be systematically "squeezed," exploited, and exhausted, becoming a source of "raw material." The gamified system of tiers, medals, and points works as a mechanism to organize and shape the work conditions inside and outside the platform, as well as to define the social conditions in which both machine learning and predictive models are built and disseminated. That is, the platform exploits its user's immaterial labour and affect as a commodity—as a "stock of brains" (Levchin, 2013), which, in turn, makes them a workforce willing to exchange their cognitive abilities, if not for free, at least for low-value tokens of appreciation. Kaggle's "winner-take-all" approach only exacerbates the inequality among players, where only a few of them actually gain something in return. The platform thus masks free labour as a ludic activity, advertising data science as a "sport" where participants compete for "fame, fortune, and fun."

The Silicon Valley culture provided the perfect economic and technical conditions for Kaggle to capitalize on the articulation of technology, creative energy, and community. Kaggle operates deeply in the capitalist ideology, in which the "markets know best" mantra is promoted and endlessly repeated. It thrives in a system in which deregulation and the imposition of insecure work conditions are praised as freedom, an ethical norm, and sometimes even as a law of nature. As such, the competitions promoted on Kaggle are not an exercise in curiosity or "real world problems" to be solved, but a carefully crafted set of challenges—organized and sponsored by large private companies—to mobilize code, data, digital infrastructures, crowdsourced labour, and political economy interests in order to advance machine techniques in a specific direction.

As such, the political economy on Kaggle is no different from other digital platforms (van Dijck et al., 2018). It fits into Srnicek's (2017) topology as a lean platform since it strives to exploit available assets in society at the lowest cost possible without directly owning them. The company is a reflection of the platform logic based on self-promotion and profit maximization that has dominated the tech discourse over the last two decades. It is the self-proclaimed world's largest data science platform but does not make science and owns no data or code. The company crowdsources most of the activities on the platform, heavily relying on its flourishing community to solve complicated problems using machine-learning techniques. By crowdsourcing cognitive work, Kaggle built a platform based on a multisided network business model to offer private corporations access to large datasets, predictive models, low-cost computer power, and, more importantly, free labour.

Kaggle's attractiveness lies in the frictionless approach to harnessing the free labour available in the digital world (Terranova, 2004) that crosses national borders with little regulation, a lack of essential protection for workers, and no accountability for private companies (Dyer-Witheford, 1999; Lazzarato, 2004). Kaggle, and the AI industry in general, cannot survive without free labour and open access to large quantities of data/code (Crawford, 2021). It not only thrives on but requires a large contingent of people willing to share their work for free to create economic value, technological progress, and the reinforcement of a supposed meritocratic system in our society. The work done by competitors on Kaggle is captured by the neoliberal regime not just as product-added value (the code, the models, the predictions) but as ways to demonstrate individual self-realization and self-optimization, the tenet values in late-capitalist society. Moreover, free labour must be cultivated, and workers must be motivated to participate in a culture of exchange in which flows of capital are primarily kept within the company. That is, participants must be encouraged to contribute and should believe that they are doing so for themselves and contributing to the community. By steering workers' affect and sense of belonging in the world, the AI industry harnesses immaterial labour (Lazzarato, 2004). As a result, these "volunteer" workers are not always aware they are engaging in productive work. Instead, they are often motivated by a desire for affective and cultural productivity. Consequently, capital surplus is not only based solely on the exhaustion of labour in the industrial sense but also on the exploitation of immaterial labour—culture, knowledge, leisure, affects, and behaviour—as data. As with any other capitalist company, Kaggle becomes a "data-human grinder," squeezing data and people all together to produce efficient and optimized solutions for capital gain.

9.2. Datafication and the Indifference Toward Human Life

The data-driven society follows the logic of late capitalism, wherein the outcomes of crowdsourced immaterial labour must be materialized as quantifiable data to become a commodity. Converting everyday life into digital data reduces life to discrete elements that can be aggregated, manipulated, and controlled using a specific set of metrics and correlations. Deeply rooted in an empiricist tradition and neoliberal values, the machine-learning community and the AI industry are mainly concerned with the measurable aspects of reality with particular economic and financial goals. By effectively isolating economic processes from social, cultural, and political events, this practice exposes a propensity to artificially insulate corporate interests from public scrutiny. Official statements and public communications from Kaggle's founders, investors, directors, and even community managers corroborate the notion that a data-driven society is not about solving problems, improving life quality, mitigating natural and artificial disasters, or aspiring to a more equitable society. Instead, the focus is always heavily skewed to some sort of financial optimization and social control. Goldbloom (2017; Gruen & Goldbloom, 2008) insists on using machine learning to make private companies more profitable, even if the data is unreliable. Levchin (2013) suggests exploiting developers' cognition power in exchange for extra cash while they sleep. Pichai (Alphabet, 2017a) wants to optimize and personalize target ads even if they radicalize the audience. Kaggle's community managers disregard unethical practices for the sake of the competition. All that matters is that data and predictive algorithms shield corporations from accountability and savers to make a profit.

Conversations in Kaggle's competitions forum confirm how shortsighted the data science community is. Most competitions do not care about the world outside of the training set, rarely considering the underlying meanings and purposes of these competitions, external factors beyond the dataset, or the possible impact of their work turning into predictive models when released as a final product. Data scientists and software engineers could care less about identifying dark matter, predicting lung cancer, improving early diagnoses of Covid-19, or detecting deepfakes or fraudulent transactions. The end goal of each competition is irrelevant. The competitions are concerned with the complex mechanisms with which data can be sorted, aggregated, calculated, and modelled to classify identities and recommend products, individuals, behaviours, media content, or any other asset that can generate surplus.

The weak ties with other peers—a loose sense of community in a remote and asynchronous platform—and the ephemerality of the gamified environment on Kaggle render the participants detached from reality, less responsible and accountable for what they develop. They become indifferent to the datasets in front of them and to the implications of their work. This indifference is also toward human life and to the context of everyday life from where data is extracted, openly exposed in the lack of ethical procedures to collect them, the indiscriminate usage of biased algorithms and training sets, and the increasing number of flawed models that are praised as ingenious solutions if they "innovate" and generate profit. Whenever these "imperfections" come to light, the machine-learning community is quick to dismiss their responsibility, blaming "unforeseeable" external factors—a person that does not fit the model, a behaviour that deviates from the norm, a turn in the market, a pandemic that was not supposed to happen. The solution is always to "extract more data" to build "better models": innovation without accountability.

This course of action quickly becomes a tautological development of machine learning: improvements for the sake of improvements. It is a convenient vantage point for the ceaseless production of predictive algorithms without a specific nature and a rigid presumption of its utility because, as Hong (2020) put it, "data will always remain open to further exploratory analyses, recombining different datasets and analytical methods to discover unforeseen correlations" (p. 24). Predictive models are deployed and sold as systems for generating "insights" across different socioeconomic problems. Nevertheless, the development process often involves recombining data that can be conveniently acquired until a useful correlation is discovered. It is also expected that such data leads to new and formerly unimagined kinds of predictions.

I have shown how permissible, ambivalent, and volatile the predictive models produced on Kaggle competitions can be. The datasets and algorithms used in competitions are often put together and further used for unrelated purposes. For instance, detecting deepfakes becomes no different from creating normative rules, equating deepfakes with nonconforming bodies, alternative facts with dissent, and the counterfeit with the unconventional. Predicting threats at airports is no different from predicting crimes, shopping behaviours, financial gain, and academic grant applications, which consequently equates terrorists with petty criminals, customers' behaviour with the housing market, and academic research priorities with shopping lists. Recommending grocery items is no different from recommending movies, posts, friends, and facts, equating bananas with Hollywood blockbusters, cat videos with personality traits, and disinformation with factual data. If the accuracy score is high, it does not matter how the dataset was gathered, who produced the algorithms, where predictive models will be used, or to which purposes it will be deployed.

In the data science community, data is perceived as amorphous, neutral, and disinterested: a common belief is that raw data contains objective truth from which facts can be drawn. Data, facts, information, and knowledge are often conflated such that they are seen to naturally follow one another, implying a (false) sense of legitimacy. By following this supposedly "natural order of things," data-driven scientists are indifferent to how the data comes to be, who and what made it available, and their interference while manipulating the dataset. However, to make sense of the dataset, engineers must order, augment, aggregate, and recombine it to squeeze out every piece of information the dataset might contain. Furthermore, Kagglers openly speak about cherry-picking variables, omitting outliers, and generating fake data to make a more robust predictive model. This forceful and inconsequential approach often conflates correlations with causation, with a propensity to create hallucinations—a sort of spurious correlation madness that makes anything seem plausible. The knowledge extracted from datasets is indifferent to the human experience of the world and the context of everyday living. However, it is deemed valid when it can be used to build precise and predictable models capable of replicating or simulating realities. In general terms, Zuboff (2020) put it this way: "It is about predicting us, without actually caring what we do or what is done to us" (p. 70).

Indifference, however, does not imply neutrality. Although the analytical process is becoming more automated and dependent on machines, the way those learning processes are designed and how the data is collected makes the process "interested," if not purposefully biased for specific intentions. The indifference toward the implication of predictive models is deeply rooted in a cultural desire to sort the world into stable and discrete pieces. This assumption appears to be a naive attempt to induce a specific understanding of the world based on notions that society (culture, economics, politics) operates according to general laws that can be modelled following a specific algorithm. Popular among positivists, this vision can be traced back to early empiricism (e.g., Francis Galton and Cesare Lombroso), the belief that pure empirical research can be undertaken free from any preconception or pre-established understanding of a phenomenon (Sekula, 1986).

We must acknowledge, however, that most developers have some notion that training data is often skewed in some way or another and may not fit the general population. However, the machine-learning community reveals a dysconscious bias, repressing any form of debate about their work's social, cultural, or political implications. Gender imbalance, racial profiling, and lack of diversity are attenuated or reduced to silence. Not only do these issues have no place on Kaggle's forum, but they have no right to exist, disappearing with any manifestation (comments are refuted, ignored, or simply deleted). The optimistic and self-righteous data scientists always look forward to solving problems "for the greater good." However, "the greater good" is rarely defined or only spoken about in abstract terms in relation to a more "productive" and "wealthy" society, without ever discussing who benefits from the proposed solutions. In Kaggle's machine-learning community, where the developers have PhDs from prestigious universities, jobs in multinational companies, and follow strict rules of objectivity and productivity, there is no bias, only numbers, correlations, silence, and indifference.

That is, machine learning enthusiasts subscribe to a particular epistemological standpoint that presupposes a prior distinction between object and subject (Grusin, 2015; Kember & Zylinska, 2014), wherein objects are understood as neutral and unbiased empirical data that hold the truth about the world, and subjects are always "contaminated" by theory, ideology, and politics that distorts or prevents facts from emerging. Data scientists act as if their instruments have no history, their cultural background has no role in their assessments, and their socioeconomic conditions have nothing to do with the truth they unearth from the data. However, the construction of knowledge cannot be based on data alone but must draw on already existing knowledge and be interpreted according to some conceptual framework—language, politics, sociotechnical protocols—that inevitably becomes the lens through which we experience the world and structure society.

9.3. Algorithmic Subjectivation

The data-driven society is bound to its historical context and conditions of existence. It not only produces the logic of technology development but also the subjects that will reproduce this logic. As discussed in this research, data and predictive models are presented as (1) commodities—hyped up for capitalist excitement—and (2) universal optimizers designed to keep existing relations of production intact and maximize the ratio at which labour power and individual behaviour are converted into surplus value. Indeed, this process results in a blatantly commercial "profit," as in the case of direct consumer goods sales and targeted advertising. However, the profits or uses of predictive models in the context of surveillance capitalism must also be considered in the biopolitical sense, where not only the state but private companies aim to define, control, and govern populations. Subjectivity has become the fuel to drive machine-learning algorithms and power the so-called artificial intelligent systems. Each individual is required to use self-optimization techniques to make themselves more appealing to the algorithmic decision-making system (Crary, 2014; Crawford, 2021; Turow, 2017). Digital platforms are supposed to capture, expand, manipulate, and claim behavioural surplus from user data—even if the data was unintentionally shared or protected by public policies. No moral, ethical, legal or social constraints will stop digital platforms from using behavioural data for commercial purposes, social control, or to target individuals as subjects.

The process of subjectivation is a mode of relations with the self and others (Foucault, 2017), understood here in a broader sense to include machines, reflective algorithms, and AI. This research identified two different but interrelated processes of subjectivation on Kaggle's power dynamic: (1) a narrow and focal strategy to produce subjects at the level of the platform, and (2) a broader approach inscribed in and reinforced by predictive models produced by the data science community. The first can be observed in how the platform organizes and manages machine learning competitions and harnesses its community to engage in crowdsourced free labour activities. For each competition, the dataset and the metrics are carefully curated and crafted by large companies to answer specific questions. Kaggle and these companies define the problems, the objectives, the metrics of success or failure, and the correct solutions. They determine how resources and people are organized, who is valued in what roles, which activities are undertaken, and for what purposes.

In observing almost 300 public competitions on Kaggle, I noticed that a curious trope about the role of machine learning engineers was consistently repeated among users: the genius, one who solves problems by themselves without any domain expertise but who also does whatever is necessary to stay ahead on the game no matter the consequences. Kagglers believe in an objective meritocratic system, which works similarly to their machine-learning accuracy score—a ranking that reflects their skills and true place in society. Interestingly, a group of people who create algorithms that rank data are themselves ranked on and by the platform. The user ranking serves not only to define the winner of each competition but also as a way to recognize their efforts with virtual gold medals and titles such as "notebook expert," "datasets master," and "competitions grandmaster." The prizes offered in each competition are calibrated to keep the community competitive and engaged with the platform. As such, the process of subjectivation on Kaggle is based on the exploits of work relations that exhaust its labour force, as well as its cultural and affective production. It plays with the community's aspirations to become data scientists, to get a job, and to be part of something meaningful. It manufactures a sense of belonging and promotes a neoliberal agenda, in which users not only aim at becoming but want to become valuable and productive members of society.

Moreover, as subjects of the platform and far from being the most self-conscious of labourers working in the digital economy, Kagglers are somewhat unaware of the politics and overarching goals to develop instruments for managing populations and individuals. It is not surprising that most Kagglers have a narrow understanding of what constitutes an individual, both at the level of the body (e.g., physiognomy, eating habits, the role of hormone levels) and from a social, cultural, and economic perspective (e.g., taste formation, beauty standards, wage, class, sense of identity). The community is business-oriented—mostly from the marketing and economic sectors, without prior knowledge of specific domains. In an effort to solve the problems posed by the sponsored company, Kagglers avoid seeing beyond the dataset and the task at hand, reducing individuals' lives to aggregated blobs of data that can only exist as correlations to a predefined set of metrics.

The goal of exploiting individuals' cognitive abilities, sociability, affect, behaviour, aspirations, and social conditions goes beyond the platform's productive relations. The political economy embedded in the data-driven society promotes an impersonal and fragmented process of subjectivation based on the mining and mobilization of subjective materials (Langlois & Elmer, 2019). This new process of subjectivation is part of a much larger diffused cultural economy that operates throughout Kaggle and beyond to produce subjects in our current society. It is built on the orchestration of the relationships among multiple data points being processed in different sites, in which the objective is to correlate personal data to other data points such as user habits, market demands, aggregated audience/consumer profiles, socioeconomic clusters, and other impersonal categories to achieve specific goals. Our habits, behaviours, and sense of identity, as well as our feelings, emotions, and affects, are not only the raw material to feed large-scale training sets for machine learning but also the heart of a new form of governing individuals.

The data-driven society reasons in terms of how we relate a number to the body's sensory experience. Everything is converted to numbers on a scale that can measured and compared to predefined metrics, producing an algorithmically generated expression of reality. As I have shown in this dissertation, most Kaggle competitions are designed to find patterns to identify, describe, classify, and predict human attributes (e.g., body, behaviour, preference, expression, affect) in order to directly or indirectly nudge us in a specific direction—to consume more, to engage with content online, to pursue a career goal, to think in specific ways. It seeks only to find or even to forge signals to make predictions about individuals in order to obtain financial profit or some sort of control over our bodies and minds.

The datasets used in Kaggle competitions are stripped of their subjective, contextual, and contentious nature to become an "objective" predictor and the speculative ground upon which data scientists build predictors for human behaviour, which in turn become the basis for new forms of sociability within the context of digital platforms. Hong (2020) argues that this logic governs how bodies are turned into facts: different bodies lend themselves to specific kinds of datafication and factmaking processes (p. 24). Han (2017) further suggests that large data models could also reveal collective patterns of behaviour individuals are unaware of, possibly rendering some form of collective unconsciousness (p. 65). As such, psychological profiles are used to calculate, compute, discriminate and govern the way individuals think, behave, and feel. By condensing and codifying the body and the mind into predictive algorithms, we are transferring the construction of knowledge from the human to the machine, inaugurating a new form of psychopolitics that would be capable of mediating mass behaviour on a level that escapes detection. What follows is that the source of truth is not in the body, or even in the mind, but in how the binary (data/code) encodes and decodes human experience to produce specific regimes of truth in order to amplify social control. As such, we are now governed by a rationality based on algorithmically mediated relations specialized in disciplining the body, shaping behaviour, and controlling the mind.

9.4. Algorithmic Mediations and the Production of Possibilities

Algorithmic media and AI can be understood as a discourse that produces specialized knowledge inferred from datasets and complex statistical calculations—often producing spurious correlations—about the body, the mind and the self. They express authority and algorithmic governance, working within their own regime of truth through the automatic and autonomous exercise of power with little human oversight. If left alone, operating by itself, this assemblage of code-machine becomes a drone: an autonomous entity that classifies, sorts, and shapes individuals by itself. Indeed, Kember and Zylinska (2015) argue that "we are—physically and hence ontologically—part of that technological environment, and it makes no more sense to talk of *us* using *it*, than it does of *it* using *us*" (p. 13, emphasis in original). We must then recognize that we are not entirely distinct from our technological tools. Therefore, digital technologies are not merely tools or intermediaries through which we accomplish some action. As a network of devices, platforms, infrastructures, and autonomous algorithms, digital technology becomes integral to defining ourselves and our society.

To apprehend the complexity of computational-centred networks constituted of machines, data, code, users, affect, and political-economic interests, we must consider the interactions among these actors and how different parts of the network influence, modulate, and interfere with each other. Digital systems not only interpellate individuals as subjects who must fulfill a socioeconomic role but are also affective and performative forces that (re)mediate our material experience of the world. Grusin (2015) argues that while media technologies have continuously operated on the epistemological and cognitive levels as modes of knowledge production, "they also function technically, bodily, and materially to generate and modulate individual and collective affective moods or structures of feeling among assemblages of humans and nonhumans" (p. 125). Predictive models created with machine-learning algorithms and the individuals they interpellate, for instance, do not preexist in isolation but become subjects and objects of an event through mediation. Even their roles as subject and object are not predefined and static. The vectors of actions, data flow, and control switch back and forth, affecting both sides as they interact with each other. Our relationality and entanglement with nonhuman entities continue to intensify with the ever more corporeal, ever more intimate dispersal of media and technologies into our social spaces, producing the conditions of life and affecting our experience of and relationship with the world.

Algorithmically mediated relations attempt to deceive natural processes through technology. The world created by the logic embedded in digital platforms is undoubtedly reflected in Kaggle and its community of data scientists and computer engineers. They adhere to a specific set of truths and values that produces permanent consensus as a surplus product, which gets directly injected into black box models, producing a feedback loop from the algorithm to the users' overall adherence to the same regime of truth. Algorithmic media produce and enact regimes of power and knowledge as they are predicated on a specific set of rules about how a system and its users must behave at any given time or in any situation (Foucault, 2004). We see these algorithms in the "wild," quietly working within digital platforms, particularly social media, which mediate information flows and shape social behaviour. The sheer volume of data exchange and the increased relevance of the attention economy as the fuel for corporate profit impact what kind of values circulate on digital media and who has the privilege to speak these values.

However, as Langlois and Elmer (2019) have argued, the mobilization of attention, desire, and affect has nothing to do with personal agency, but plays with the possibility of using subjective materials to reorganize life according to specific hierarchies of power. The business logic of digital platforms is deeply rooted in a particular political economy of subjectivation intended to regulate the *production of possibilities* and uphold particular socioeconomic relations. It establishes the conditions of existence (Lazzarato, 2004) that dictate the allocation of resources and people, as well as who is valued for what roles, what is done, and for what reason. Therefore, to guarantee that the production of possibilities (i.e., choices, behaviours, facts) is always controlled, predictable, codified, and mediated according to the modes of capital valourization and current regimes of truth, big tech not only targets individuals as subjects of this logic but is also willing to radicalize the algorithm (Tufekci, 2018). Thus, in the name of economics and technological urgency, digital platforms justify the racism and discrimination present in our society, becoming a fertile ground for deceiving radicalized and discriminatory content, including disinformation campaigns, hate speech, and deepfakes.

The widespread algorithms that facilitate harmful content to proliferate and become viral directly result from the increased adoption of recommendation systems and the lack of regulation and oversight of AI. It is a deliberate strategy that privileges radicalized content at the expense of minorities and marginalized groups, causing an ever-increasing inequality, terror, and a resurgence of racist, fascist, and nationalist forces determined to exclude and destroy. While algorithm radicalization is commonly identified with far-right ideologies, often considered a threat to liberal democracy, it is also not far removed from hegemonic practices in the neoliberal society (Ituassu et al., 2019; Paiva, 2020). Indeed, a central promise of AI is that it enables large-scale automated categorization. This “promise” becomes a menace when directed at the complexities of everyday life. Careless labels can oppress and harm when they assert false authority. In her book *Weapons of Math Destruction*, Cathy O’Neil (2017) describes in detail how various algorithmic systems media calcify oppression, foster deep inequalities, and deprive targeted populations, erode human dignity and undermine basic democratic mechanisms when engineered irresponsibly. Indeed, the AI project aims to scrape, exploit, and exhaust human nature. Our lives are rendered as behavioural data for the sake of others’ improved control of us. Predictive algorithms do not provide options or even aid us in decision-making because the decision rights vanish before one even knows there is a decision to make.

In the world created by empirical capitalists, data appears to be an entity disconnected from everything that contains fundamental and useful truths about the universe. Data may "speak by itself," but whatever we do with data depends on how we interpret it (Gitelman, 2013). In Antony Goldbloom's platform, data speaks about indifference toward the methods of data extraction, from whom data are "squeezed," how data is manipulated, and the consequence of predictive models in everyday life. For the data science community on Kaggle, data speaks about the colonization of bodies; it argues for the exploitation of affect, labour, and natural resources; it claims for a profit at any cost. A data-driven society has no feelings, no empathy, and no concern for the collectivity and the individuals.

However, if we care to listen, data might voice other values. To do so, it demands that we reconsider data as a space of diversity that moves past the assumption that it can only be used for profit and social control. It demands that we imagine alternative futures that break with the values imposed by the Californian libertarian ideology. I wonder what data would tell us if the data science community had been founded with a different set of values, one that promotes cooperation instead of competition, a critical approach to the methods used, self-awareness about implicit biases, and different goals beyond the easy-to-sell but destructive capitalist gain. Would the developers be more conscious of their own idiosyncrasies? Would the community seek fair work relations? Would companies be accountable for the use of predictive models? Would the algorithms be more transparent and easier to audit? Would the models foster a more equitable society with more respect and empathy? I wonder what other values can be drawn and mobilized from large datasets available in our society. What other modes of subjectivities will emerge, and what kinds of futures can we create using a more humanist approach to data mobilization?

9.5. Limitations

This research is primarily interested in how algorithms mediate the sense of self and how artificial intelligence has been developed as a new form of subjectivation. The bulk of the empirical work focuses on the developing stages of predictive models in the context of high-profile competitive machine-learning events held by a large digital platform. While I believe this study has successfully compiled a detailed profile on Kaggle (as a company and a platform) and produced substantial arguments about the political economy of subjectivation involving predictive models, this research has some notable empirical, analytical, and methodological limitations.

The main empirical and analytical limitations are the sample size and the sampling method employed in the research. While I considered all public competitions on Kaggle that offered at least US$1,000 in cash prizes, I only examined one competition for each machin-learning modality, thus producing a bias toward the selected events. The modalities discussed here could be enriched by expanding the analysis to other competitions on Kaggle. I suggest starting with the ones described at the end of each chapter to relate to the present research directly. Another limitation was the availability of the material at the time of the analysis. For instance, the Passenger Screening Algorithm Challenge's dataset has been removed from Kaggle due to national security concerns; comments and threads on the competition's forum were deleted; previous versions of Kaggle were only partially available through the Internet Archive. Lastly, this research focuses on a platform for machine-learning development. The platform's nature and its users' idiosyncrasies may have skewed the arguments of this research. Expanding toward other spaces—alternative platforms for algorithm development, academic research labs, start-ups, and perhaps large companies—would undoubtedly help to generalize this research's findings.

Methodological limitations are unavoidable and part of any research endeavour. Focusing on Digital Methods and Discourse Analysis, this research primarily relied on living archives (e.g., competitions' websites and forums, blog posts, datasets, and code repositories). These archives are insufficient to reveal the developer's motivation, intentions, propensities, and goals, especially while participating in machine-learning competitions. Moreover, official documents (e.g., earning calls and annual reports) are highly performative and double as marketing material, where investors and company leaders avoid discussing controversial issues. An ethnographic approach, using semi-structured interviews with developers, Kaggle staff, and competition organizers, could offer different perspectives from the community and shed light on some of the contentious debates outlined in this research. Alternatively, a Critical Code Analysis approach would help direct inquiries into datasets, machine-learning algorithms, and predictive models in order to reveal embedded knowledge and intrinsic biases. In both cases, the research could be improved by directly participating in competitions and observing the events unfold instead of relying on archives and second-hand reports of the events.

9.6. Future Research

AI is spreading in all directions, rapidly evolving to become one of the most critical infrastructures in human history and occupying centre stage in every aspect of our lives. The hype around the theme might be a marketing tool for short-term profit, but it also produces an excess of discourse about AI. This dissertation has merely scratched the surface of the political economy of subjectivation and new forms of mediation using predictive models. The competitions I examined are only a fraction of a more extensive process of subjectivation and production of the conditions of existence. Probing deeper, the results of this dissertation provide a strong foundation for future work striving to shed light on how predictive models are used to shape social habits, discipline bodies, and govern populations. In closing, therefore, I would like to suggest three main areas that urgently need more careful scrutiny and that should be prioritized in future research: generative AI and large language models (LLM); the geopolitics of AI (neo-colonialism and techno-imperialism); and the regulation of and resistance to AI.

Since OpenAI launched ChatGPT in November 2022, the AI industry has changed completely. People started to have more informal, direct, and even intimate interactive relations with conversational chatbots in a way that felt more like talking to another human than to an AI system. The interaction is seamless, with a natural flow of conversation at each turn in an apparent sublimation of the mediation horizon and the sociotechnical apparatuses in which the chatbot is grounded. Beyond words, LLMs are multimodal, capable of generating any type of text, imagery, or sound: from essays to code, from realist portraits and photographs to any artistic style, from the human voice to new genres of music, from raw video footage to deepfakes. In a single year, the ecology of generative AI burst into a rich fauna with many "different species:" Google's LaMDA, Bard, and Gemini; Meta's LLaMA and Voicebox; Microsoft's Github Copilot; Midjourney; Open AI's DALL-E; Stable Diffusion; Adobe's Firefly; MusicML; MusicGen, to name only a few. Generative AI opens possibilities for ontogenesis and other processes of reinterpretation and subjectivation that can enable new kinds of relations between humans and machines.

However, these "intelligent systems" are reductionist and often discriminatory, privileging the dominant culture and social norm. Marginalized groups only participate as a precarious workforce. LLM hallucinations, cultural appropriation, and copyright infringements further increase the sense of urgency to understand how to control, moderate, and regulate AI systems. All these issues were already present in an unexplored modality of machine learning on Kaggle, as identified in this research. Though Kaggle only held a few low-profile competitions exclusively focused on generative systems, competitions such as the Facebook Deepfake Challenge are based on generative AI and could have been explored in such a way. Further research should aim to understand how the selective cultural bias induced by the AI industry implies homogeneity and discrimination, inadvertently failing to account for other possibilities and effectively neglecting, if not forgetting, the diversity of human cultural heritage. Moreover, since generative AI is constantly “learning” from each interaction, we should think about how the cultural exchange from this intertwined mediatory process of our relations to AI produces new kinds of subjects, new forms of experiences, and possibly new forms of life.

The broad impact of AI systems is the cause of increasing geopolitical tensions. Currently, the sector is dominated by U.S. companies, most notably big tech like OpenAI, Google, Microsoft, Meta, Amazon, and Apple. The current global hegemony of U.S. AI-driven platforms raises anxieties about a new form of imperialism (Jin, 2015), resurrecting the “cultural imperialism” debates of the 1980s and 1990s. The political economy of this neo-cultural techno-imperialism aims to reproduce economic asymmetries in geopolitical relations by erasing cultural differences under the sameness of American culture as it is distributed and consumed around the world (particularly via social media, chatbots, and tools used in the creative industry). AI-generated content and deepfakes significantly lower the cost of influence operations by both domestic and foreign actors and influence—even disrupt—democratic processes. Above all, the vast economic gains drive the centrality of AI competition in the global political agenda. While other countries, particularly China, are also aggressively investing in AI through their domestic companies (Watrix, Tencent, Alibaba, SenseTime, ByteDance, and Huawei are not far behind their American counterparts), the new U.S. policy (The White House, 2023) seeks to assert the country’s supremacy by spending billions of dollars in “innovation.” This “arms race” has been compared to the development of the atomic bomb (Reynolds, 2023) and raised questions about the broad social, cultural, economic, and political implications of an AI-driven society. The question is how national and economic interests influence the direction in which AI has been and will be developed.

Furthermore, as Kate Crawford (2021) demonstrates, AI systems are not merely slick digital interfaces that output “infinite” possibilities. They are actually made of other finite materials, and the industry itself does not bear the true costs of AI. The sector demands the mining of rare earth minerals (to build its electronic parts), energy (usually from oil and coal sources), an impressive amount of water (to cool down data warehouses), specialized industrial assembly lines together with a long and reliable logistic chain, data (tons of it), and a load of human capital to work (both specialized workers and cheap labour). Most of these assets are not exclusively located in the Global North but are "mined" and exploited from the Global South, repeating the patterns of colonial history (Hao, 2020). AI users are usually unaware of or shielded from these issues, focusing on using AI-driven platforms to write schoolwork, fix grammar mistakes, produce advertisements and entertainment content, and automatize creative work. The Global South, on the other hand, is to be kept focused on exporting primary resources and providing cheap and precarious labour. This economic divider is clearly shaped by implicit ideas of a specific hierarchy of power in which some populations do not need—or are less deserving of—livable wages and economic stability. The question here is not only how AI systems impact the creative industry in the North but also how the same notion of optimization has been used to justify a new form of colonialism in the South.

The lack of regulation and care for human life and the limited resources on the planet are also important questions to be asked and pursued in future research. The discourse among AI venture capitalists prioritizes exponential financial growth, led by private corporations that have gained autonomy, rights, power, and nearly unregulated societal influence. Ito (2018) suggests that the mentality of these capitalists is akin to the behaviour of cancer cells, which are optimized for unconstrained growth and spreading with disregard to their function or context (p. 4). Indeed, Marc Andreessen, a well-known venture capitalist billionaire in the tech business, proposed making exceptions in the copyright laws and labour protection for the AI industry, effectively cannibalizing other businesses in order to fuel AI systems; otherwise, tech companies "might not be able to afford to license copyrighted training data" or the work done by coders, analysts, and moderators. In response to this and other harmful outcomes, many nations are debating, proposing, and implementing new policies to oversee AI. Approved in December 2023, the EU AI Act is the first of its kind. It strives to mitigate the risks of AI by creating rules to ensure that AI is safe, transparent, traceable, non-discriminatory, and environmentally friendly. Yet, with rules only applicable to the European block, it is difficult to detect a significant change in an industry that works on a planetary scale. The challenge here is how to make AI systems and their owners accountable and legally liable for their decisions.

Lastly, we must find ways to resist, if not refuse, the way predictive models and AI systems construct our social conditions. As discussed in this dissertation, AI has been used to target us as subjects using personal and impersonal data. These systems are built to label people based on behaviour, socioeconomic status, or personal characteristics using real-time and remote biometric identification systems for cognitive behavioural manipulation of people or specific vulnerable groups. In protest at the inadequacy of these labels, many young people today proudly defy unwelcome categorizations using tactics that vary from feeding “wrong” data (noise) about themselves into AI systems, using facial make-up or wearing clothes with unique patterns to confuse facial detectors, and creating multiple “fake accounts” in their devices to split their profile (Penn, 2018). Yet, perhaps the political, ethical, social, and philosophical problem of our time is not to try to liberate the individual from AI systems but to liberate ourselves from a kind of neoliberal subjection. Echoing Foucault (1982), I argue that we should take the opportunity to feed the machine with different values and promote new forms of subjectivity by refusing and resisting the AI subjectivation that has been imposed upon us.

10. Appendix I: Building a Web Scraper

In a first impression, scraping a webpage can be straightforward: identify the information to be scrapped and set up some sort of code to look for structural patterns. However, this operation can become overly complicated depending on the technology used to build the website, the consistency and reliability of the information structure, and the website's policy in relation to scrapers. These were the challenges I faced in building a scrapper to collect metadata from thousands of pages on Kaggle. In this appendix, I describe my process of building a Web scrapper: assessing the website, choosing a technology, writing the scrapper, and the strategies to avoid detection.

In my first assessment, Kaggle showed consistency, but not without some irregularities—such as unordered lists, missing metadata, and an unpredictable information hierarchy. Moreover, Kaggle webpages are asynchronously rendered, which means that the HTML source code initially served on the browser was no more than an empty template waiting to be filled with information from subsequent API calls to Kaggle's server. The initial content is limited to a specific amount, usually the top 10 or most recent (competitions, datasets, users, discussions). The rest of the content only loads when the user interacts with the page by clicking to see more details of a dataset, scrolling to the bottom of the page to load more users on the ranking, or changing the tab to show recently submitted notebooks. Tools such as Beautiful Soup and Cheerio would not find any useful data since they would stop short on the initial HTML source code. Information retrieval from disembodied requests—that is, without user input—is insufficient to obtain data from Kaggle's website. The process requires the actual human experience of waiting for the website to load and navigating through the page to unearth data from Kaggle's data warehouses. It was clear from the beginning that I would have to build a scraper from scratch.

My approach to extracting data from Kaggle involves a method I was experimenting with but never fully implemented in a different project where I examined YouTube's recommendation algorithm (see McKelvey & Frizzera, 2019; Reis, Zanetti, & Frizzera, 2020): a bot that simulates user interactions on-screen and navigates the website in real-time. To build this bot, I used a tool with an appropriate name: Puppeteer.[27] Built by Google Chrome's Team as an instrument to automate web-based app testing (performance, rendering, interactions), Puppeteer is a Node.js library which provides high-level API control over a web browser (Chromium). Puppeteer allows us to create agents (bots) embodying a predefined "persona" to visit the website according to a specific plan, including its own set of personal preferences, account passwords, and log history. Indeed, the targeted website would perceive this visitor as a human user actively and purposefully navigating the page. However, visiting thousands of pages one by one to scrape data from each competition and dataset would take a very long time. To speed up the process, I used an add-on to Puppeteer named "puppeteer-cluster,"[28] developed by Thomas Dondorf. This library facilitates the orchestration of a cluster of browsers. That is, the scraper could open multiple instances of the browsers as if multiple users were accessing different pages simultaneously, enabling concurrent data gathering and reducing the time to complete the task.

The last piece of the puzzle was to define how and where to store the data. There are different approaches to storing the data depending on the volume, complexity, and purposes for which the data will be used. For instance, if the data comprises only plain key and value pairs, dumping the result into a spreadsheet or CSV file may be a quick and straightforward approach, even if the dataset contains hundreds of thousands of data points. On the other hand, if the data has a more complex hierarchical structure or contains different data types, such as arrays, objects, and numbers (e.g., Twitter's user profile), storing it on a JSON file is preferable. However, depending on the volume of data (e.g., streams of social media posts, financial data, social network relationships, etc.), working with databases is preferable for better data management and retrieval (queries).

My research aimed at producing a snapshot of the Kaggle platform. That is, running a script to collect all the data at once. My first incursion on Kaggle's webpage prompted me to use a NoSQL database to store the data. MongoDB[29] was the obvious choice: it natively works with JSON objects, allowing us to store more complex data types without the traditional complexity of SQL's relational databases. Another reason to use a database was to avoid duplicate entries while storing participant users in each competition.

---

[27] See https://pptr.dev

[28] See https://github.com/thomasdondorf/puppeteer-cluster

[29] See https://www.mongodb.com/

10.1. Scraper I: Competitions, Datasets, and User Ranking

With this set of tools and technologies, I wrote a script to collect data from three sections on Kaggle:[30] competitions, datasets, and user ranking. Each section has its own scraper logic due to the difference in information hierarchy.

10.1.1. Competition

The Competitions page lists all current and previous competitions held on Kaggle. The page has three tabs: Active, Completed, and InClass (see Figure 10.1). Competitions are sorted in no particular order by default, but the user can sort them by date, reward, or the number of participant teams. It is also possible to filter by category or search for a specific competition. From this page alone, it is possible to gather a few attributes without accessing an internal page, such as the competition URL, title, short description, category, subcategory, relative due date,[31] the number of teams, and the award. It is important to note that some competitions might not have all these attributes, disrupting the scraper as it traverses the HTML structure tree.

The page only loads 25 items at a time. The user must scroll to the bottom of the page to load the next 25 items. However, this action does not update the URL with information about how many items to load or from where to start pulling data. The scraper must simulate user interactions on the page, such as mouse clicks and scroll events, to navigate into each section and load subsequent batches of data. This automated action is performed by a bot clicking on one of the tabs and waiting for the main content. As soon as the content loads, the bot scrolls down to the bottom of the page to load the next 25 competitions as many times as necessary. When there is nothing else to load, the scraper loops through each item on the list, locates the appropriate HTML elements, collects the data, and saves it to the database before continuing to the next item. The bot clicks on the next tab when the list is over, and the process restarts. After a first glance at the data collected, I noticed an essential piece of information was missing from the overview data: the competition's sponsor.

Using the saved data, I extended the scraper to access the competition's internal pages containing detailed information about each competition. The header summarizes the details: title, short

[30] The code is open source and is available together with the results at github.com/lucaju/kaggle.
[31] The date is relative to the moment when the page was accessed. For instance, if a competition was ending on April 14 and the page was accessed a week earlier, on April 7, the site would display "7 days to go."

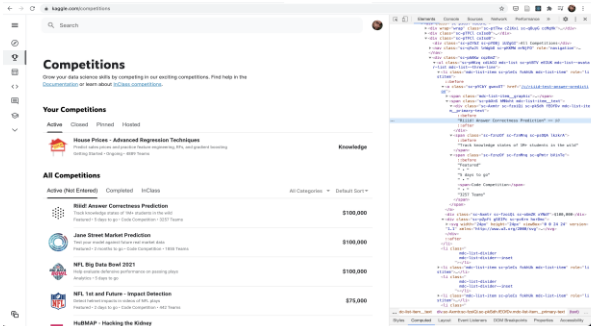


Figure 10.1: Kaggle Competitions Page. List of competitions with the source code panel opened showing the HTML structure. Screenshot by L. Frizzera.

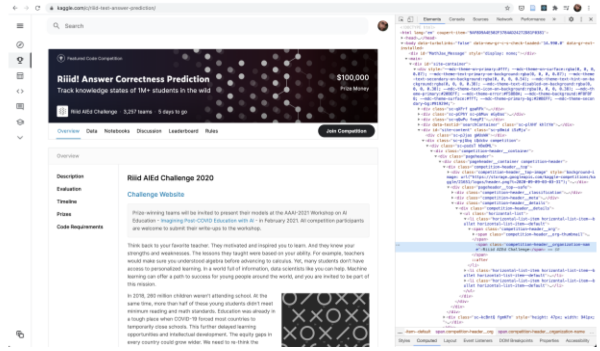


Figure 10.2: Kaggle Competition Detail Page. Competition details with the source code panel opened showing the HTML structure. Screenshot by L. Frizzera.

description, category, organizer, number of teams, the relative due date, and the prize (see Figure 10.2). The page is split into six tabs: Overview, Data, Notebooks, Discussions, Leaderboard, and Rules. The overview contains a complete description of the competitions, a list of tags, how contributions are evaluated, a timeline, the list of prizes, specific requirements, the most recent notebooks, and the most recent contribution in the discussion forum. The Data section has a description of the dataset used in the competition, a command to download the data directly from the API, and a data explorer through which it is possible to access some files in the dataset without downloading them. The Notebooks section contains a list of contributions that can be upvoted and replied to by other users. The Discussion section consists of a forum where the participants can exchange messages. It simply lists conversation threads with a title and the number of comments and upvotes. The leaderboard contains a public and a private ranking. Even after the deadline, the public ranking continuously updates as the competitors submit contributions. The private ranking evaluates users' submissions against a private test data host at Kaggle. It is also the official competition's ranking, frozen in time when the competition ended. Both rankings contain a list of teams with their members, scores, and the number of entries in the competition. Lastly, the Rules section describes the rules for the competitions.

On the internal page, I focused on just a few details: the organization that sponsors the competition, the date it started and ended, the number of competitors, tags, and the public leaderboard. To scrape more than 3,000 pages, I simultaneously deployed multiple (up to 8) bots to visit each page. Each bot navigated to a competition page and waited for the content to load. After gathering the information on the overview page, the bot clicked on the leaderboard tab and waited for the content to load. Since the leaderboard page only loads the first 50 teams, the bot scrolled to the bottom of the page and clicked on the "load more" button to pull the full ranking. The bot then looped through each team to gather the team's name, score, and members. When data collection was over, the bot closed the window and opened another with the next competition on the list.

10.1.2. Datasets

The Datasets page lists all publicly available datasets hosted on Kaggle. At the top of the page, Kaggle invites users to create their own public dataset: "Open a dialogue, accept contributions, and get insights: improve your dataset by publishing it on Kaggle." By default, the list of datasets is ordered by "hottest" (a score based on release date and upvotes). However, the user has other options, such as sorting by date (uploaded or updated), the number of votes, and the usability score. It is also possible to filter and search datasets using tags, file size and type, and type of license. Similar to the competitions page, Kaggle only loads 25 datasets at a time. To load more, the user must scroll to the bottom of the page to load the next 25 items (see Figure 10.3). The list of datasets displays the following attributes: the competition page's URL, title, owner usernames, last update relative date, usability score, number and type of file, size, medals, and the number of upvotes.

Once again, the scraper must simulate a user interacting with the page. However, the approach used on the competition page would not be a good fit here. With dozens of thousands of datasets to be loaded on a single page, the browser would overload the memory capacity. I filtered the datasets to gather a manageable number of items on a page. Tags would be the best semantic candidate. However, datasets can have multiple tags, which would cause datasets to be listed several times. File type suffers from the same issue. The license was not an option either since many owners do not assign a license to their dataset. The best option was to create batches of datasets based on size. After a few tests, I split the datasets into groups of about 3,000 based on their size: 1-2kB, 2-5KB, 5-

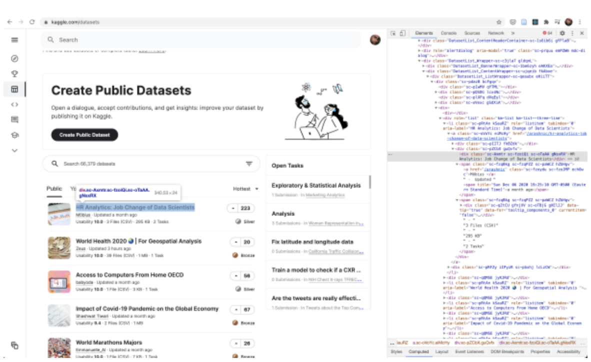


Figure 10.3: Kaggle Datasets Page. List of datasets with the source code panel opened showing the HTML structure. Screenshot by L. Frizzera.

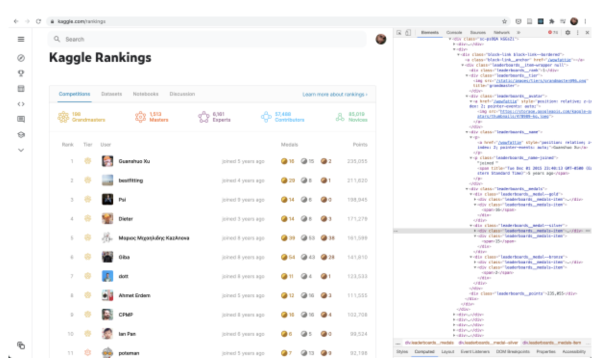


Figure 10.4: Kaggle User Ranking Page. List of top users with the source code panel opened showing the HTML structure. Screenshot by L. Frizzera.

10KB, and up to 20-100GB, the upper limit for a dataset submission. Fortunately, Kaggle generated a new URL for each search, making navigating to the appropriate group easier.

With the dataset groups defined, the bot systematically navigated each group and waited for the main content to load. As soon as the content loaded, the bot scrolled down to the bottom of the page to load more data. When there was nothing else to load, the scraper looped through each item on the list, located the appropriate elements, collected the data, and saved it into the database before continuing to the next item. When the list was over, the bot navigated to the next group, restarting the process.

10.1.3. User Ranking

Kaggle classifies its users into five tiers: Novice (lower tier), Contributors, Expert, Master, and Grandmaster (higher tier). The User Ranking lists users on the Expert tier and higher in four categories: Competitions, Datasets, Notebooks, and Discussion. There is no other option for sorting or filtering users. Each user on this list carries the following attributes: the username with their position on the ranking, tier, number of medals and points, and the relative date the user was registered on Kaggle. Similar to competitions and datasets, Kaggle only displays 40 users at a time. To load more, the user must scroll to the bottom of the page to load the next 40 items (see Figure 10.4).

Once more, the scraper simulates a user interacting with the page. The bot clicked on one of the rankings and waited for the content. As soon as the content loaded, the bot scrolled down to the bottom of the page to load more data. When there was nothing else to load, the scraper looped through each item on the list, located the appropriate elements, collected the data, and saved it into the database before continuing to the next item. The bot clicked on the next tab when the list was over, and the process restarted. It is important to note that the same user can appear on multiple rankings. The scraper checked if they were already stored before saving users to the database. If the user exists, the scraper appends the ranking data to the registered user, avoiding duplication and enriching the user data.

I ran the scraper in two phases using a laptop (MacBook Pro) (1) on November 1, 2020, to collect a list of competition, datasets, and users, which took about eight hours to complete, and (2) between December 7 and December 8, 2020, to gather competition's details, which took about 15 hours to

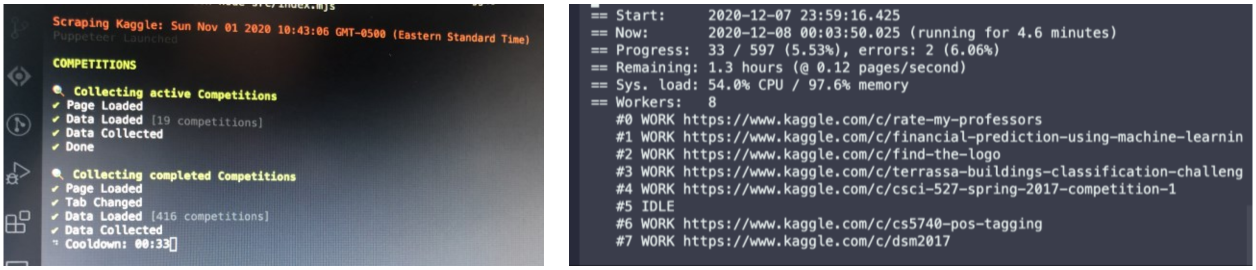


Figure 10.5: Scrapper Logs. Screenshots of the scraper running, gathering the list of competitions (left) and visiting each competition's page (right).

complete (see Figure 10.5). The scraper collected data from 3,437 competitions, 55,897 datasets, and 9,164 users on the public ranking page.

10.2. Scraper II: Competition's Internal Pages

A second scraper was developed in April 2022 to collect extra information about Kaggle's competition analyzed in this research, namely Facebook's 2019 Deepfake Detection Challenge, 2017 Passenger Screening Algorithm Challenge, and 2017 Instacart Market Basket Analysis. This scrapper focused on the comments posted by users in the competition's discussion forum.

The Competition Discussion page only loads 20 threads at a time (see Figure 10.6). The bottom of the page displays a pagination navigation where the user can load more threads. The threads are sorted by "hottest" by default. The user can search threads using keywords or sort the list by "recent posted," "recent comments," "most votes," or "most comments." Kaggle's Social Media Manager can pin replies to the top of the list. The list of threads displays the following attributes: the thread titles, the creator's name, a relative date, the number of comments and upvotes.

Each thread displays an initial comment at the top of the page, including the author (username, tier, and URL), title, content, date, the number of votes, the number of replies, and the URL (see Figure 10.7). Below it is a list of replies, including author information (username, tier, and URL), content, date, number of votes, and medals. The replies are sorted by "hottest" by default. This sorting method nested replies to replies, making it easier to follow the context of a conversation. The user can also sort the list by "newest first," "oldest first," or "most votes," but in these cases, the conversation is not nested. The page only loads 50 comments at a time. To load more, the user must scroll to the bottom of the page to load the next 50 comments. Each reply displays the following attributes: the user's name, username, tier, a relative date, the position of the user in competition

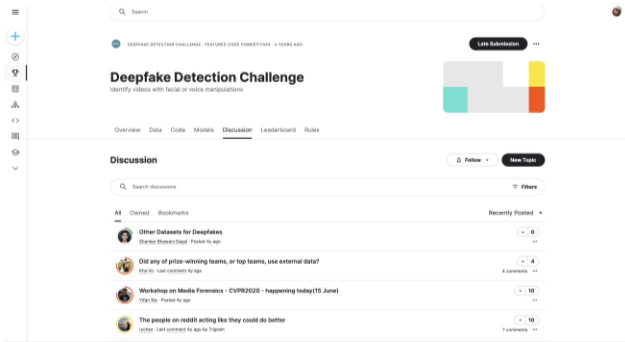


Figure 10.6: Kaggle Competition Discussion page. List discussion threads in a specific competition. Screenshot by L. Frizzera.

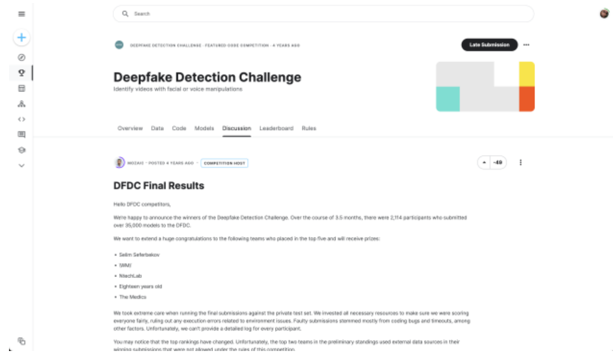


Figure 10.7: Kaggle Competition Discussion's Thread Page. Shows an initial comment and a list of replies. Screenshot by L. Frizzera.

ranking, the number of upvotes, and the contents of the reply. It is important to note that Kaggle's Social Media Manager has the power to delete a reply; in this case, the reply remains on the thread, but its content will be replaced with the text "[deleted]." Moreover, if a user deletes their account, the reply remains on the website, but the user's name is replaced with the text "[deleted]" and the user's information is not available.

I divided the data collection into two phases: (1) collecting the threads and (2) collecting the comments in each thread. In phase 1, the scraper first navigated directly to the competition discussion page. The bot looped through each thread, located the appropriate elements, collected the data, and saved it into the database before continuing to the following thread. After collecting all threads on the page, the scrapper navigated to the next page. Each page has a specific URL, so it was easy to navigate directly to a specific page. After collecting all the threads, I initiated phase 2. Using the URL for each thread saved on the database, the scrapper navigated to each page for each thread. To facilitate the data collection, the bot clicked the sorting button and selected "oldest first," making the list display chronologically. The bot scrolled down to the bottom of the page to load more replies. When there was nothing else to load, the scraper first collected data from the header (the initial comment) and then looped through each comment on the list, located the appropriate elements, collected the data, and saved it into the database before continuing to the next item.

The scraper ran on May 4, 2022, using a laptop (MacBook Pro). Collecting the discussion pages for three selected competitions took about one hour to complete. In total, the scraper collected 1,268 threads with 11,047 replies. More specifically, Facebook's Deepfake Detection Challenge had 772 threads and 6,777 replies, the Passenger Screening Algorithm Challenge had 147 threads and 1,282 replies, and Instacart Market Basket Analysis had 350 threads and 2,988 replies.

10.3. Scraper III: Homepage Screenshots

The third scraper was developed in March 2022 to collect snapshots from previous versions of Kaggle's homepage stored on the Internet Archive via its Wayback Machine. The data was captured in two steps: (1) obtain a list of archived versions of *kaggle.com* on the Wayback Machine and (2) screen capture each version. For the first step, I used Richard Rogers's (2017) tool, "Internet Archive Wayback Machine Link Ripper."[32] I configured the tool to extract a list of archived versions of *kaggle.com*, limited to one version per month. For the second step, I used Puppeteer to navigate to each URL and screen capture each page. Puppeteer allowed me to generate a JPEG for the whole page, not just what can be displayed on the browser's viewport.

The scraper ran on April 8, 2022, using a laptop (MacBook Pro), taking about three hours to complete. From February 6, 2010, when Kaggle was launched, to April 14, 2022, the Internet Archive saved its homepage 4,610 times. The first step collected a list of 146 links to unique versions of the site from February 6, 2010 to October 7, 2021, limited to one per month. The second step generated 136 screenshots of these pages from October 2, 2010, to October 7, 2021, reflecting the analysis period of this dissertation.

10.4. Brief Consideration on Security and Privacy

Digital platforms have API policies and technical barriers to stop scrapers from collecting data. On its Terms of Service, Kaggle (2020e) explicitly prohibits "'Crawls,' 'scrapes,' or 'spiders' any page, data, or portion of or relating to the Services or Content (through use of manual or automated means)" ("Are there any additional restrictions on my use of the Services?", para. 5). I have discussed the legalities and implications of web scraping in my methodologies (chapter three). Here, I want to emphasize that there is no way to know if and how Kaggle enforces this rule. Moreover, since this scraper visits eight pages simultaneously, Kaggle's server may perceive these requests as a Denial-of-Service (DDoS) attack. To avoid detection and mitigate this problem, the scraper was used behind a Virtual Private Network (VPN) with an IP rotation and takes 5-minute breaks every time the scraper switches a tab or visits a new competition or dataset page.

[32] See https://wiki.digitalmethods.net/Dmi/ToolInternetArchiveWaybackMachineLinkRipper

Bibliography

0sb1. (2020). Aligned Anime Faces. Kaggle. https://www.kaggle.com/datasets/reitanaka/alignedanimefaces

Acxiom. (2021). Acxiom Website. Acxiom. https://www.acxiom.com/

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The State of Deepfakes: Landscape (p. 27) [Cybersecurity]. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Akyildirim, N. (2020, February 6). Instacart Exploratory Analysis [Data Science Blog]. Towards Data Science. https://towardsdatascience.com/instacart-exploratory-analysis-fdd5e3f9bf60

Allstate Insurance. (2011). Allstate Claim Prediction Challenge. Kaggle. https://kaggle.com/c/ClaimPredictionChallenge

Alphabet. (2015). Annual Report (Annual Report 2015; p. 124). https://abc.xyz/investor/static/pdf/2015_alphabet_annual_report.pdf?cache=40474a1

Alphabet. (2016). Alphabet's (GOOG) Management presents atCredit Suisse Technology, Media and TelecomConference (Transcript). https://seekingalpha.com/article/4027135-alphabets-goog-management-presents-credit-suisse-technology-media-and-telecom-conference

Alphabet. (2017a). Alphabet (GOOG) Q1 2017 Results—EarningsCall Transcript. https://seekingalpha.com/article/4066436-alphabet-goog-q1-2017-results-earnings-call-transcript

Alphabet. (2017b). Alphabet (GOOG) Q3 2017 Results—EarningsCall Transcript. https://seekingalpha.com/article/4117116-alphabet-goog-q3-2017-results-earnings-call-transcript

Alphabet. (2017c). Alphabet (GOOG) Q4 2016 Results—EarningsCall Transcript. https://seekingalpha.com/article/4040406-alphabet-goog-q4-2016-results-earnings-call-transcript

Alphabet. (2018a). Alphabet Inc. (GOOG) CEO Sundar Pichai on Q2 2018 Results—Earnings Call Transcript. https://seekingalpha.com/article/4189317-alphabet-inc-goog-ceo-sundar-pichai-on-q2-2018-results-earnings-call-transcript

Alphabet. (2018b). Alphabet's (GOOG) CEO Sundar Pichai on Q1 2018 Results—Earnings Call Transcript. https://seekingalpha.com/article/4165019-alphabets-goog-ceo-sundar-pichai-on-q1-2018-results-earnings-call-transcript

AlphaMeow. (2017, August 8). This competition is disappointing [Discussion]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/competitions/instacart-market-basket-analysis

Althusser, L. (1971). Ideology and Ideological State Apparatuses. In Lenin and Philosophy (pp. 123–156). New Left Books. https://www.marxists.org/reference/archive/althusser/1970/ideology.htm

American Civil Liberties Union. (2020, March 28). Federal Court Rules 'Big Data' Discrimination Studies Do Not Violate Federal Anti-Hacking Law. American Civil Liberties Union. https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti

Amir, M., & Kotef, H. (2018). In-secure identities: On the securitization of abnormality. Environment and Planning D: Society and Space, 36(2), 236–254. https://doi.org/10.1177/0263775817744780

Amoore, L. (2020). Cloud Ethics: Algorithms and the Attributes of Ourselves and Others. Duke University Press.

Amperity. (2021). AmpID. Amperity. https://amperity.com/platform/ampid

Amusi. (2020). YOLOv5 is disabled! Results of Kaggle global wheat testing competition cause controversy—Programmer Sought. https://www.programmersought.com/article/83514966533/

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 3686–3693. https://doi.org/10.1109/CVPR.2014.471

Aneesh, P. A. (2015). Neutral Accent: How Language, Labor, and Life Become Global. Duke University Press.

Angwin, J., Parris Jr., T., & Mattu, S. (2016, September 28). Breaking the Black Box: What Facebook Knows About You. ProPublica. https://www.propublica.org/article/breaking-the-black-box-what-facebook-knows-about-you?token=2IGQdWFIQ6nSzBbnQ4-16RVFd3ayz6DV

Angwin, J., & Parris, T. (2016, October 28). Facebook Lets Advertisers Exclude Users by Race. ProPublica. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

Apprich, C., Chun, W. H. K., Cramer, F., & Steyerl, H. (2018). Pattern Discrimination. Meson Press.

Arcady27. (2017, August 17). My useful features [Discussion]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/competitions/instacart-market-basket-analysis

Asare, J. G. (2020, April 14). Does TikTok Have A Race Problem? Forbes. https://www.forbes.com/sites/janicegassam/2020/04/14/does-tiktok-have-a-race-problem/

ASHRAE. (2019). ASHRAE – Great Energy Predictor III. Kaggle. https://kaggle.com/c/ashrae-energy-prediction

AstroTom. (2011). Mapping Dark Matter. Kaggle. https://kaggle.com/competitions/mdm

Aviation Security Advisory Committee. (2017). Imporoving Checkpoints at U.D. Airports (p. 26). https://www.tsa.gov/sites/default/files/asac_cof_final_report_23may17.pdf

Ayyub, R. (2018, November 21). I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. HuffPost UK. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316

Ballantyne, M., Boyer, R. S., & Hines, L. (1996). Woody Bledsoe—His Life and Legacy. AI Magazine, 17(1), 7–20. https://doi.org/10.1609/aimag.v17i1.1207

Banik, R. (2017). The Complete Pokemon Dataset. Kaggle. https://www.kaggle.com/datasets/rounakbanik/pokemon

Barbrook, R., & Cameron, A. (1996). The Californian ideology. Science as Culture, 6(1), 44–72. https://doi.org/10.1080/09505439609526455

Bardají, L. (2019, November 13). Deepfakes: Can I believe what I see and hear? [Educational]. Do Better by Esade. https://dobetter.esade.edu/en/deepfakes

Barron, L. (2023). AI and Popular Culture. Emerald Publishing Limited.

Baudrillard, J. (1983). Simulations (P. Beitchman, P. Foss, & P. Patton, Trans.). Semiotex(e).

Benjamin, W. (2008). The Work of Art in the Age of Mechanical Reproduction. In M. W. Jennings, B. Doherty, & T. Y. Levin (Eds.), & E. Jephcott, R. Livingstone, H. Eiland, & Others (Trans.), The Work of Art in the

Age of Its Technological Reproducibility, and Other Writings on Media (1st ed., pp. 19–52). Belknap Press of Harvard University Press.

benpflaum, W. K., Brian G, djdj, Irina Kofman, JE Tester, JLElliott, Joshua Metherd, Julia Elliott, Mozaic, Phil Culliton, Sohier Dane. (2019). Deepfake Detection Challenge. Kaggle. https://kaggle.com/competitions/deepfake-detection-challenge

Bischoff, W. (2023, May 2). MPF questiona Google sobre possível favorecimento de conteúdos contrários ao PL das Fake News em resultados de busca. G1. https://g1.globo.com/politica/noticia/2023/05/01/mpf-questiona-google-sobre-possivel-favorecimento-de-conteudos-contrarios-ao-pl-das-fake-news-em-resultados-de-busca.ghtml

Bledsoe, W. W. (1963). A Facial Recognition Project Report. http://archive.org/details/firstfacialrecognitionresearch

Bogacz, M. (2022, November 22). 15 factors for data science in your country! [Code]. Kaggle. https://kaggle.com/code/michau96/15-factors-for-data-science-in-your-country

Bogost, I. (2007). Persuasive Games: The Expressive Power of Videogames. The MIT Press. https://doi.org/10.7551/mitpress/5334.001.0001

Bolter, J. D., & Grusin, R. (2000). Remediation: Understanding new media. MIT Press.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. 30th Conference on Neural Information Processing Systems (NIPS 2016), 9. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Books Ngram Viewer. (2022). https://books.google.com/ngrams/graph?content=data+science%2Cstatistics%2Ccomputer+science%2Cmachine+learning%2Cartificial+intelligence&year_start=1800&year_end=2019&corpus=26&smoothing=0&direct_url=t1%3B%2Cdata%20science%3B%2Cc0%3B.t1%3B%2Cstatistics%3B%2Cc0%3B.t1%3B%2Ccomputer%20science%3B%2Cc0%3B.t1%3B%2Cmachine%20learning%3B%2Cc0%3B.t1%3B%2Cartificial%20intelligence%3B%2Cc0

Borsboom, D., & Molenaar, D. (2015). Psychometrics. In J. D. Wright (Ed.), International Encyclopedia of the Social & Behavioral Sciences (2nd edition) (pp. 418–422). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.43079-5

Boussebaa, M., Sinha, S., & Gabriel, Y. (2014). Englishization in offshore call centers: A postcolonial perspective. Journal of International Business Studies, 45(9), 1152–1169. https://doi.org/10.1057/jibs.2014.25

Bouwmeester, H. (2017, December 11). H(o)i Bastiaan, Thanks for sharing. I've played with Google Cloud VMs a bit [Commented on My workflow[[Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45487

Boyer, R. S. (Ed.). (2012). Automated Reasoning: Essays in Honor of Woody Bledsoe (Softcover reprint of the original 1st 1991 edition). Springer.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and Information Technology, 15(3), 209–227. https://doi.org/10.1007/s10676-013-9321-6

Brabham, D. C. (2013). Crowdsourcing. The MIT Press.

Branden, M. (2017, July 13). I've seen passengers with a variety of weights and heights. There are definitely a few smaller people in the dataset. [Commented on Are these samples acquired from the same person?] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35976

Brandom, R. (2019a, March 5). Deepfake propaganda is not a real problem. The Verge. https://www.theverge.com/2019/3/5/18251736/deepfake-propaganda-misinformation-troll-video-hoax

Brandom, R. (2019b, June 7). Microsoft pulls open facial recognition dataset after Financial Times investigation. The Verge. https://www.theverge.com/2019/6/7/18656800/microsoft-facial-recognition-dataset-removed-privacy

Bratton, B. H. (2016). The Stack: On Software and Sovereignty (1st edition). The MIT Press.

Brice, S. (2020, November 15). A Short History of Facial Recognition. Medium. https://samdbrice.medium.com/4ecd290aaab1

Brock, A. (2018). Critical technocultural discourse analysis. New Media & Society, 20(3), 1012–1030. https://doi.org/10.1177/1461444816677532

Browne, S. (2015). Dark matters: On the surveillance of blackness. Duke University Press.

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. New Media & Society, 14(7), 1164–1180. https://doi.org/10.1177/1461444812440159

Bucher, T. (2013). Objects of Intense Feeling: The Case of the Twitter API. Computational Culture, 3. http://computationalculture.net/objects-of-intense-feeling-the-case-of-the-twitter-api/

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *1*(81), 1–15.

Butler, J. (1990). Gender Trouble. Routledge.

Button Poetry (Director). (2015, April 5). Kevin Kantor—People You May Know. https://www.youtube.com/watch?v=LoyfunmYIpU

BuzzFeedVideo (Director). (2018, April 17). You Won't Believe What Obama Says In This Video! 😉. https://www.youtube.com/watch?v=cQ54GDm1eL0

Caby, S. (2019, December 18). Python, Keras/Tensorflow 2.x, Win/Ubuntu, GTX1080ti, videogame dev, sound engineer, got plenty of free time and I'm fast. [Commented on Looking for a Team Thread] [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/121201

Camilleri, A., Geoghegan, R., Meade, R., Osorio, S., & Zou, Q. (2019, December 13). Bias in Machine Learning: How Facial Recognition Models Show Signs of Racism, Sexism and Ageism | by Rachel Meade | Towards Data Science. Medium. https://towardsdatascience.com/bias-in-machine-learning-how-facial-recognition-models-show-signs-of-racism-sexism-and-ageism-32549e2c972d

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI Now 2017 Report (White Paper Recommendation 2; AI NOW, p. 37). New York University and AI NOQ.

Cao, L. (2017). Data Science: A Comprehensive Overview. *ACM Computing Survey*, *50*(3), 1–42. https://doi.org/10.1145/3076253

Carr, C. T., & Hayes, R. A. (2015). Social Media: Defining, Developing, and Divining. *Atlantic Journal of Communication*, *23*(1), 46–65. https://doi.org/10.1080/15456870.2015.972282

Castells, M. (2000). The Rise of The Network Society: The Information Age: Economy, Society and Culture, Volume I (2nd edition). Wiley-Blackwell.

Chan, W. (2022, August 24). The AI startup erasing call center worker accents: Is it fighting bias – or perpetuating it? The Guardian. https://www.theguardian.com/technology/2022/aug/23/voice-accent-technology-call-center-white-american

Chandrashekar, A., Amat, F., Basilico, J., & Jebara, T. (2017, December 7). Artwork Personalization at Netflix [Medium]. Netflix Technology Blog. https://netflixtechblog.com/artwork-personalization-c589f074ad76

Chaudhari, S. (2019, December 16). I am an undergrad from India. I have an experience of 2 years in the Deep Learning field. I have [Commented on Looking for a Team Thread] [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://kaggle.com/competitions/deepfake-detection-challenge

Chauhan, N. (2021, September 20). *Your search ends here: The journey of Kaggle*. https://kaggle.com/code/nilaychauhan/your-search-ends-here-the-journey-of-kaggle

Chowdhury, A. P. (2017, July 26). Why is Kaggle's upcoming challenge surrounded by so much controversy? Analytics India Magazine. https://www.analyticsindiamag.com/kaggles-upcoming-challenge-receiving-much-flak/

Christl, W. (2017). Corporate Surveillance In Everyday Life. How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions. Cracked Lab – Institute for Critical Digital Culture. http://crackedlabs.org/en/corporate-surveillance

Chun, W. H. K. (2017). Updating to Remain the Same: Habitual New Media (reprint edition). The MIT Press.

Citron, D. K., & Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. Washington Law Review, 89. https://papers.ssrn.com/abstract=2376209

Clark, B. (2018, April 19). Pornhub promised to ban 'deepfakes' videos. And it failed miserably. TNW | Insider. https://thenextweb.com/news/pornhub-promised-to-ban-deepfakes-videos-and-it-failed-miserably

Class-Y ANR Project. (2014). Large Scale Hierarchical Text Classification. Kaggle. https://kaggle.com/c/lshtc

Cole, S. (2017, December 11). AI-Assisted Fake Porn Is Here and We're All Fucked. Vice. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Cole, S. (2018, January 24). We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. Vice. https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley

Conference on Document Analysis and Recognition. (2013). Gender Prediction from Handwriting. Kaggle. https://kaggle.com/c/icdar2013-gender-prediction-from-handwriting

Cooper, M. (2011). Complexity Theory After the Financial Crisis. Journal of Cultural Economy, 4(4), 371–385. https://doi.org/10.1080/17530350.2011.609692

Copeland, R. (n.d.). WSJ News Exclusive | Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans. WSJ. Retrieved August 13, 2022, from https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790

Corporación Favorita. (2018). Corporación Favorita Grocery Sales Forecasting. Kaggle. https://kaggle.com/c/favorita-grocery-sales-forecasting

Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. Journal of Design and Science. https://doi.org/10.21428/96c8d426

Crary, J. (2014). 24/7: Late Capitalism and the Ends of Sleep. Verso.

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Crunchbase. (2019, July 4). Kaggle [Database]. Crunchbase. https://www.crunchbase.com/organization/kaggle

Crunchbase. (2021). Kaggle—Crunchbase Company Profile & Funding. Crunchbase. https://www.crunchbase.com/organization/kaggle

Crunchbase. (2022). Max Levchin—Founder & CEO @ Affirm—Crunchbase Person Profile. Crunchbase. https://www.crunchbase.com/person/max-levchin

Cryle, P., & Stephens, E. (2017). Normality: A Critical Genealogy. University of Chicago Press.

Cuff, T. (n.d.). Historical Anthropometrics. In Economic History. Retrieved November 17, 2022, from https://eh.net/encyclopedia/historical-anthropometrics/

Cukierski, W. (2015, July 27). A Rising Tide Lifts All Scripts. No Free Hunch. https://web.archive.org/web/20170813094152/http://blog.kaggle.com/2015/07/27/a-rising-tide-lifts-all-scripts/

Cukierski, W. (2017a, June 22). Welcome! [Discussion]. Kaggle. https://www.kaggle.com/c/passenger-screening-algorithm-challenge/discussion/35116

Cukierski, W. (2017b, August 10). We generally do not update competition datasets unless there is a systematic problem [Commented on Annotation issue] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/37615

Curry, D. (2022, January 11). Instacart Revenue and Usage Statistics 2022. https://www.businessofapps.com/data/instacart-statistics/

Cut Onion. (2017, June 24). [Commented on Are we here for money on learning!?] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35126

Daft Vader. (2017, June 22). Possible solutions to the “Prize problem” [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://kaggle.com/competitions/passenger-screening-algorithm-challenge

Darpa. (n.d.). The Grand Challenge. Retrieved April 2, 2022, from https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles

Data.gov. (2019). USA Name Data. Kaggle. https://kaggle.com/datagov/usa-names

DavidGbodiOdaibo. (2017, December 16). 100 predictions!!! With Test time augmentation? [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Davidson, A. I. (Ed.). (2016). From Subjection to Subjectivation: Michel Foucault and the History of Sexuality. In Foucault and the making of subjects (pp. 55–62). Rowman & Littlefield International, Ltd.

Davies, H. (2015, December 11). Ted Cruz using firm that harvested data on millions of unwitting Facebook users. The Guardian. https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data

Dayal, G. (2012, June 9). The Algorithmic Copyright Cops: Streaming Video's Robotic Overlords. WIRED. https://www.wired.com/2012/09/streaming-videos-robotic-overlords-algorithmic-copyright-cops/

de Lauretis, T. (1987). Technologies of Gender: Essays on Theory, Film, and Fiction. Indiana University Press.

Deepfake Detection Challenge. (2019a). Deepfake Detection Challenge. Kaggle. https://kaggle.com/c/deepfake-detection-challenge

Deepfake Detection Challenge. (2019b). Deepfake Detection Challenge Rules. https://www.kaggle.com/competitions/deepfake-detection-challenge/rules

Deepmind. (2017). Kinetics. Deepmind. https://deepmind.com/research/open-source/kinetics

Deis, A. (2019, December 21). DeepFake Challenge EDA [Code]. Kaggle Facebook's Deepfake Detection Challenge. https://kaggle.com/code/aleksandradeis/deepfake-challenge-eda

Deleuze, G. (1992). Postscript on the Societies of Control. October, 59, 3–7.

Delgado, R., & Stefancic, J. (2001). Critical race theory: An introduction. New York University Press.

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. Big Data & Society, 8(2), 20539517211035955. https://doi.org/10.1177/20539517211035955

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A Generative Model for Music. arXiv:2005.00341 [Cs, Eess, Stat]. http://arxiv.org/abs/2005.00341

DHS. (2017). Passenger Screening Algorithm Challenge. Kaggle. https://kaggle.com/c/passenger-screening-algorithm-challenge

DHS. (2018, July 8). News Release: S&T Announces Winners in $1.5M Prize Competition | Homeland Security. US Department of Homeland Security. https://www.dhs.gov/science-and-technology/news/2018/07/09/news-release-st-announces-winners-15m-prize-competition

Dibbell, J. (1993, December). A Rape in Cyberspace. The Village Voice.

DiscipleOfScience. (2017a, July 13). I have not seen all images yet, but the 30 or so if have, seem to all have basically same body type [Commented on Are these samples acquired from the same person?] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35976

DiscipleOfScience. (2017b, August 10). Have seen A LOT of mislabel or no label as well. [Commented on Annotation issue] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/37615

Dixon, S. (2022). Biggest social media platforms 2022. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Dixon, S. J. (2023, September 21). Facebook MAU worldwide 2023. Statista. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. arXiv:2006.07397 [Cs]. http://arxiv.org/abs/2006.07397

Drawbridge. (2015). ICDM 2015: Drawbridge Cross-Device Connections. Kaggle. https://kaggle.com/c/icdm-2015-drawbridge-cross-device-connections

DSTL. (2017, March 15). Dstl's Kaggle competition has been a great success [Institutional]. GOV.UK. https://www.gov.uk/government/news/dstls-kaggle-competition-has-been-a-great-success

Dwoskin, E., Tiku, N., & Timberg, C. (2021, November 21). Facebook's race-blind practices around hate speech came at the expense of Black users, new documents show. Washington Post. https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/

Dyer-Witheford, N. (1999). Cyber-Marx: Cycles and Circuits of Struggle in High Technology Capitalism (1st edition). University of Illinois Press.

Eckhardt, G. M., & Bardhi, F. (2015, January 18). The Sharing Economy Isn't About Sharing at All. Harvard Business Review. https://hbr.org/2015/01/the-sharing-economy-isnt-about-sharing-at-all

EduardoLi. (2017, July 12). Anthony, a solution adopted by some sport competitions could solve the misunderstanding [Commented on the online forum post Our thinking in hosting this prize] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/c/passenger-screening-algorithm-challenge/discussion/35977

Ehl, D. (2019, May 5). Is Facebook doing enough to combat fake news? | DW | 05.05.2019. Deutsche Welle. https://www.dw.com/en/is-facebook-doing-enough-to-combat-fake-news/a-48608648

Elazari, A. (2018). Private Ordering Shaping Cybersecurity Policy: The Case of Bug Bounties (SSRN Scholarly Paper ID 3161758). Social Science Research Network. https://papers.ssrn.com/abstract=3161758

Elden, S. (2007). Governmentality, Calculation, Territory. Environment and Planning D: Society and Space, 25(3), 562–580. https://doi.org/10.1068/d428t

Elejalde-Ruiz, A. (2018, December 10). "We're being mistreated": Instacart shoppers complain of pay cuts as company shifts to new model. Chicago Tribune. https://www.chicagotribune.com/business/ct-biz-instacart-shopper-complaints-20181206-story.html

Elias, B. (2012). Airport Body Scanners: The Role of Advanced Imaging Technology in Airline Passenger Screening (Congressional Research Service 7–5700; p. 15). https://sgp.fas.org/crs/homesec/R42750.pdf

Elliott, J. (2020, January 9). Request Free Cloud TPU Quota! [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/125313

Elo. (2019). Elo Merchant Category Recommendation. Kaggle. https://kaggle.com/c/elo-merchant-category-recommendation

emergent complexity. (2017, December 17). I started with zone segmentation from the very beginning. [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening

Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Engler, A. (2019, November 14). Fighting deepfakes when detection fails. Brookings. https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/

Escoufier, Y., Hayashi, C., & Fichet, B. (Eds.). (1995). Data science and its applications = La @science des donn??es et ses applications. Academic Press/Harcourt Brace.

Farman, J. (2011). Mobile Interface Theory: Embodied Space and Locative Media (1st edition). Routledge.

Farrar, B. (2017, September 30). Exploratory Data Analysis and Example Generation [Notebook]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/code/jbfarrar/exploratory-data-analysis-and-example-generation/notebook

Ferreira Neto, J. L. J. L. (2018). A Analítica da Subjetivação em Michel Foucault. Revista Polis e Psique, 7(3), 7. https://doi.org/10.22456/2238-152X.76339

Ferrer, C. C., Dolhansky, B., Pflaum, B., Bitton, J., Pan, J., & Lu, J. (2020, June 12). Deepfake Detection Challenge Results: An open initiative to advance AI. Facebook AI. https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

Ferrer [mozaic], C. C. (2020a, January 2). Let me provide some comments here: If audio and/or video are modified, you should predict the video as FAKE [Commented on the online forum post REAL/FAKE is not enough, we need to know if its video/audio/both] [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/c/deepfake-detection-challenge/discussion/121694

Ferrer [mozaic], C. C. (2020b, June 12). DFDC Final Results [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/157925

Finn, E. (2017). What Algorithms Want: Imagination in the Age of Computing. The MIT Press.

Flick, U. (2014). The SAGE Handbook of Qualitative Data Analysis. Sage. https://doi.org/10.4135/9781446282243

Forbes. (2012, December 17). 30 Under 30. Forbes. http://www.forbes.com/special-report/2012/30-under-30/30-under-30.html

Fortune, J. (2018). Passenger Screening Algorithm Challenge (Apex Screening at Speed, p. 5). SHS.

Foucault, M. (1980). Truth and Power. In Power/Knowledge: Selected Interviews and Other Writings, 1972-1977 (pp. 109–133). Vintage.

Foucault, M. (1982). The Subject and Power. Critical Inquiry, 8(4), 777–795. https://doi.org/10.1086/448181

Foucault, M. (1995). Discipline & Punish: The Birth of the Prison (REP edition). Vintage.

Foucault, M. (2004). The Birth of Biopolitics: Lectures at the Collège de France, 1978–1979. Picador.

Foucault, M. (2009). Security, Territory, Population: Lectures at the Collège de France 1977–1978 (G. Burchell, Trans.; 1st edition). Picador.

Foucault, M. (2017). Subjectivity and Truth: Lectures at the Collège de France, 1980-1981 (F. Ewald, A. Fontana, & F. Gros, Eds.; G. Burchell, Trans.; 1st 2017 edition). Palgrave Macmillan.

Franceschi-Bicchierai, L. (2022, April 26). Facebook Doesn't Know What It Does With Your Data, Or Where It Goes: Leaked Document [News]. Vice. https://www.vice.com/en/article/akvmke/facebook-doesnt-know-what-it-does-with-your-data-or-where-it-goes

Franklin, U. (1999). *The Real World of Technology* (1st edition). House of Anansi Press.

Frenkel, S., & Conger, K. (2022, December 2). Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find. The New York Times. https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html

Fuchs, C. (2013). Political Economy and Surveillance Theory. Critical Sociology, 39(5), 671–687. https://doi.org/10.1177/0896920511435710

Fuller, M. (Ed.). (2008). Software Studies: A Lexicon. MIT Press. http://mitpress.mit.edu/9780262062749/software-studies/

Fuster, G. G., Bellanova, R., & Gellert, R. (2015). Nurturing Ob-Scene Politics: Surveillance Between In/Visibility and Dis-Appearance. Surveillance & Society, 13(3/4), 512–527.

Gabrys, J. (2016). Program Earth: Environmental Sensing Technology and the Making of a Computational Planet. University of Minnesota Press.

Gallagher, F. (2019, November 15). Are Google, Twitter and Facebook doing enough to protect the 2020 election in the age of "information disorder"? ABC News. https://abcnews.go.com/US/google-twitter-facebook-protect-2020-election-age-information/story?id=66527631

Galloway, A. R. (2006). Protocol: How Control Exists after Decentralization (1st edition). The MIT Press.

Galton, F. (1879). Psychometric experiments. Brain: A Journal of Neurology, 11, 149–162.

Gellman, L. (2013, July 17). Why Startups Are Sporting Increasingly Quirky Names. The Wall Street Journal. https://www.wsj.com/articles/SB10001424127887324263404578611772689785466

General Electric. (2013). Flight Quest 2: Flight Optimization, Milestone Phase. Kaggle. https://kaggle.com/c/flight2-milestone

Gershgorn, D. (2021a, June 9). Is there any way out of Clearview's facial recognition database? The Verge. https://www.theverge.com/22522486/clearview-ai-facial-recognition-avoid-escape-privacy

Gershgorn, D. (2021b, July 7). Can AI learn from any public code online? The Verge. https://www.theverge.com/2021/7/7/22561180/github-copilot-legal-copyright-fair-use-public-code

Giba. (2020). 1st Place Removed Solution—All Faces Are Real Team. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/157983

Giles, M. (2018, February 21). The GANfather: The man who's given machines the gift of imagination. MIT Technology Review. https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/

Gillespie, T. (2010). The politics of 'platforms.' *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), Media technologies: Essays on communication, materiality, and society (pp. 167–193). The MIT Press.

Gillespie, T. (2017). Regulation of and by Platforms. In J. Burgess, T. Poell, & A. Marwick (Eds.), *SAGE Handbook of Social Media* (pp. 254–279). Sage. https://web-a-ebscohost-com.lib-ezproxy.concordia.ca/ehost/ebookviewer/ebook/bmxlYmtfXzE0NDI2MjZfX0FO0?sid=cf3c6e97-eee5-4150-a58b-1a6654341369@sdc-v-sessmgr02&vid=0&format=EB&rid=1

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.

Gitelman, L. (Ed.). (2013). "Raw data" is an oxymoron. The MIT Press.

GOKAGGLERS. (2017, September 20). Any US Citizen willing to Teamup with me on this challenge [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/39729

Goldbloom, A. (2016, July 25). What Kaggle has learned from 2MM machine learning models [Conference]. data.bythebay.io, San Francisco, US. https://www.youtube.com/watch?app=desktop&v=oYNKc_u9Os8

Goldbloom, A. (2017a, March 8). Kaggle Joins Google Cloud | No Free Hunch. Kaggle. https://web.archive.org/web/20170730050043/http://blog.kaggle.com/2017/03/08/kaggle-joins-google-cloud/

Goldbloom, A. (2017b, July 8). Our thinking in hosting this prize [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/c/passenger-screening-algorithm-challenge/discussion/35977

Goldbloom, A. (2019, June 10). The IQ Interview: Anthony Goldbloom, CEO Kaggle (J. Bennett, Interviewer) [Video]. https://vimeo.com/341464245

Goldbloom, A. (2020, February 6). The Story of Kaggle & Kaggle's Evolution | Interview with the CEO of Kaggle: Anthony Goldbloom (S. Bhutani, Interviewer) [Video on Chai Time Data Science]. https://www.youtube.com/watch?v=jw2Z-IMyFYw

González, R. J. (2017). Hacking the citizenry?: Personality profiling, 'big data' and the election of Donald Trump. Anthropology Today, 33(3), 9–12. https://doi.org/10.1111/1467-8322.12348

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks (arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Google. (2015). Google's (GOOG) CEO Larry Page Hosts 2015 Annual Meeting of Stockholders (Transcript).

Google AI. (2018a). Google AI Open Images—Object Detection Track. Kaggle. https://kaggle.com/c/google-ai-open-images-object-detection-track

Google AI. (2018b). Google AI Open Images—Visual Relationship Track. Kaggle. https://kaggle.com/c/google-ai-open-images-visual-relationship-track

Google AI. (2019). Open Images 2019—Instance Segmentation. Kaggle. https://kaggle.com/c/open-images-2019-instance-segmentation

Google Jigsaw. (2018). Toxic Comment Classification Challenge. Kaggle. https://kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Google Jigsaw. (2019). Jigsaw Unintended Bias in Toxicity Classification. Kaggle. https://kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

Gordon, A. (2020, January 14). Data Science Community Rocked by Pet Adoption Contest Cheating Scandal. Vice. https://www.vice.com/en/article/m7qezx/kaggle-data-science-community-rocked-by-pet-adoption-contest-cheating-scandal

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Grabell, M., & Salewski, C. (2011, December). Sweating Bullets: Body Scanners Can See Perspiration as a Potential Weapon. ProPublica. https://www.propublica.org/article/sweating-bullets-body-scanners-can-see-perspiration-as-a-potential-weapon?token=ANuQ_0K6ZdI2EeccenY98zKjB_c7Pza-

Gross, T. (2019, September 9). The CIA's Secret Quest For Mind Control: Torture, LSD And A "Poisoner In Chief." NPR. https://www.npr.org/2019/09/09/758989641/the-cias-secret-quest-for-mind-control-torture-lsd-and-a-poisoner-in-chief

Gruen, D., & Goldbloom, A. (2008). The smarter use of data. 12.

Grusin, R. (2015). Radical Mediation. Critical Inquiry, 42(1), 124–148. https://doi.org/10.1086/682998

Guimaraes, A. A. R., & Tofighi, G. (2018). Detecting Zones and Threat on 3D Body for Security in Airports using Deep Machine Learning. arXiv:1802.00565 [Cs]. https://doi.org/10.5281/zenodo.1189345

Guller, D. (2017, August 9). If that's not making sense to you, then let me try again. [Commented on the thread This competition is disappointing] [Discussion]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/competitions/instacart-market-basket-analysis

Hallinan, B., & Striphas, T. (2016). Recommended for you: The Netflix Prize and the production of algorithmic culture. New Media & Society, 18(1), 117–137. https://doi.org/10.1177/1461444814538646

Hamner, B. (2015, October 3). *Kaggle with Ben Hamner* (Pranay, Interviewer) [Audio]. http://softwareengineeringdaily.com/2015/10/03/kaggle-with-ben-hamner/

Han, B.-C. (2017). Psychopolitics: Neoliberalism and New Technologies of Power (E. Butler, Trans.). Verso.

Hao, K. (2020, October 9). Inside the strange new world of being a deepfake actor. MIT Technology Review. https://www.technologyreview.com/2020/10/09/1009850/ai-deepfake-acting/

Hao, K. (2022, April 19). Artificial intelligence is creating a new colonial world order. MIT Technology Review. https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/

Hao, K., & Heaven, W. D. (2020, December 24). The year deepfakes went mainstream. MIT Technology Review. https://www.technologyreview.com/2020/12/24/1015380/best-ai-deepfakes-of-2020/

Harris, P. (2011, October 11). "Underwear bomber", accused over Christmas Day terror plot, goes on trial. The Guardian. https://www.theguardian.com/world/2011/oct/11/underwear-bomber-plot-trial-detriot

Harvey, A., & LaPlace, J. (2021). Exposing.ai. Exposing.Ai. https://exposing.ai/

Harvey, D. (2012). Rebel Cities. Verso Books.

Haysom, S. (2018, January 31). People are using face-swapping tech to add Nicolas Cage to random movies and what is 2018. Mashable. https://mashable.com/article/nicolas-cage-face-swapping-deepfakes

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., & Ferrer, C. C. (2022). Towards Measuring Fairness in AI: The Casual Conversations Dataset. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(3), 324–332. https://doi.org/10.1109/TBIOM.2021.3132237

Heaven, W. D. (2021, August 27). People are hiring out their faces to become deepfake-style marketing clones. MIT Technology Review. https://www.technologyreview.com/2021/08/27/1033879/people-hiring-faces-work-deepfake-ai-marketing-clones/

Heavy.ai. (n.d.). What is Feature Engineering? Definition and FAQ. https://www.heavy.ai/technical-glossary/feature-engineering

Heawood, J. (2018). Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. Information Polity, 23(4), 429–434. https://doi.org/10.3233/IP-180009

HedgeHog. (2017, July 7). Wow! I agree. What exactly makes you think this competition is not worldwide and multi-cultral? [Commented on Welcome!] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35116

Helfand, J. (2019). Face: A Visual Odyssey (illustrated edition). The MIT Press.

Heller, M. (2020, June 30). Kaggle: Where data scientists learn and compete. InfoWorld. https://www.infoworld.com/article/3564164/kaggle-where-data-scientists-learn-and-compete.html

Helmond, A., Nieborg, D. b., & van der Vlist, F. n. (2019). Facebook's evolution: Development of a platform-as-infrastructure. Internet Histories, 3(2), 123–146. https://doi.org/10.1080/24701475.2019.1593667

Heritage Provider Network. (2013). Heritage Health Prize. Kaggle. https://kaggle.com/c/hhp

Hesmondhalgh, D. (2008). Neoliberalism, imperialism and the media. In The Media and Social Theory (1st edition, p. 312). Routledge.

Hewlett Foundation. (2012). The Hewlett Foundation: Automated Essay Scoring. Kaggle. https://kaggle.com/c/asap-aes

Heymsfield, S. B., Bourgeois, B., Ng, B. K., Sommer, M. J., Li, X., & Shepherd, J. A. (2018). Digital anthropometry: A critical review. European Journal of Clinical Nutrition, 72(5), 680–687. https://doi.org/10.1038/s41430-018-0145-7

Hill, K. (2014, June 28). Facebook Manipulated 689,003 Users' Emotions For Science. Forbes. https://www.forbes.com/sites/kashmirhill/2014/06/28/facebook-manipulated-689003-users-emotions-for-science/

Hill, K. (2018, August 8). "People You May Know:" A Controversial Facebook Feature's 10-Year History. Gizmodo. https://gizmodo.com/people-you-may-know-a-controversial-facebook-features-1827981959

hippskill. (2017, August 7). R.I.P. competition [Commented on the thread Baseline 0.4029970] [Discussion]. Kaggle Instacart Market Basket Analysis. https://www.kaggle.com/hippskill/competitions

Home Credit Group. (2018). Home Credit Default Risk. Kaggle. https://kaggle.com/c/home-credit-default-risk

Hong, S. (2020). Technologies of Speculation: The Limits of Knowledge in a Data-Driven Society. NYU Press.

Howard, A. (2017, July 27). Apply for FREE Google Cloud Credits from Kaggle! (CLOSED) [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://kaggle.com/competitions/passenger-screening-algorithm-challenge

Howard, A. (2019a, December 11). External Data Disclosure Thread [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/121203

Howard, A. (2019b, December 11). Looking for a Team Thread [Discussion]. Discussion. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/121201

Hsu, J. (2017, July 3). Earn a $1.5 Million Prize at "Kaggle!" (American Applicants Only, Please.). Undark Magazine. https://undark.org/2017/07/03/kaggle-data-science-dhs/

Hull, G. (2018, May 23). Hypernudges as Subjectification. New APPS: Art, Politics, Philosophy, Science. https://www.newappsblog.com/2018/05/hypernudges-as-subjectification.html

IBM Cloud Education. (2020, August 19). What is Supervised Learning? https://www.ibm.com/cloud/learn/supervised-learning

idle_speculation. (2017, December 21). My final submission was a little complicated, but basically it boils down to 5 variations on the strategy outlined below [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Iglovikov, V. (2017, June 22). (With respect to the current rules I am eligible to claim prize money if I win [Commented on This is insane discrimination and insult to our international community] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35118

Impermium. (2012). Detecting Insults in Social Commentary. Kaggle. https://kaggle.com/c/detecting-insults-in-social-commentary

InnoCentive. (n.d.). Solving some of the World's biggest problems. InnoCentive. Retrieved April 2, 2022, from https://www.innocentive.com/about-us/about/

Instacart. (2017). Instacart Market Basket Analysis. Kaggle. https://kaggle.com/c/instacart-market-basket-analysis

InstaShopper. (2017, June 11). Hi Philip and Kagglers—Instacart Shoppers [Commented on the thread Exploratory Analysis—Instacart] [Code]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/code/philippsp/exploratory-analysis-instacart

Irani, L. (2015). Hackathons and the Making of Entrepreneurial Citizenship. *Science, Technology, & Human Values*, *40*(5), 799–824. https://doi.org/10.1177/0162243915578486

Isaac, M. (2019, August 20). In New Facebook Effort, Humans Will Help Curate Your News Stories. The New York Times. https://www.nytimes.com/2019/08/20/technology/facebook-news-humans.html

Ito, J. (2017). Resisting Reduction: A Manifesto. Journal of Design and Science. https://doi.org/10.21428/8f7503e4

Ituassu, A., Lifschitz, S., Capone, L., & Mannheimer, V. (2019). DE DONALD TRUMP A JAIR BOLSONARO: democracia e comunicação política digital nas eleições de 2016, nos Estados Unidos, e 2018, no Brasil. Anais Compolítica 2019, 25. http://compolitica.org/novo/anais/2019_gt4_Ituassu.pdf

Jansen, C. (2017, August 9). Great data analyst: I agree on this. [Commented on the thread This competition is disappointing] [Discussion]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/competitions/instacart-market-basket-analysis

Jin, D. Y. (2015). Digital Platforms, Imperialism and Political Culture (1st edition). Routledge.

Johnson, B. (2019, March 25). Deepfakes are solvable—But don't forget that "shallowfakes" are already pervasive. MIT Technology Review. https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/

Jones, P. R. M., West, G. M., Harris, D. H., & Read, J. B. (1989). The loughborough anthropometric shadow scanner (LASS). Endeavour, 13(4), 162–168. https://doi.org/10.1016/S0160-9327(89)80004-3

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. IEEE Access, 8, 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

Kaggle. (2010, February 6). Kaggle—A platform for data-related competitions. https://web.archive.org/web/20100206144351/http://kaggle.com/

Kaggle. (2011, May 27). Data mining, forecasting and bioinformatics competitions on Kaggle. Kaggle. http://web.archive.org/web/20110527212123/http://www.kaggle.com/

Kaggle. (2017). 2017 Kaggle Machine Learning & Data Science Survey [dataset]. Kaggle. https://www.kaggle.com/datasets/kaggle/kaggle-survey-2017

Kaggle. (2018). 2018 Kaggle Machine Learning & Data Science Survey [dataset]. https://www.kaggle.com/datasets/kaggle/kaggle-survey-2018

Kaggle. (2019). 2019 Kaggle Machine Learning & Data Science Survey [dataset]. Kaggle. https://kaggle.com/competitions/kaggle-survey-2019

Kaggle. (2020a). 2020 Kaggle Machine Learning & Data Science Survey [dataset]. Kaggle. https://kaggle.com/competitions/kaggle-survey-2020

Kaggle. (2020b). Kaggle milestone: 5 million registered users! | Data Science and Machine Learning. https://www.kaggle.com/general/164795

Kaggle. (2020c). Meet Kaggle. https://www.kaggle.com/static/slides/meetkaggle.pdf?Host_Business

Kaggle. (2020d). Tweet Sentiment Extraction. Kaggle. https://kaggle.com/c/tweet-sentiment-extraction

Kaggle. (2020e, September). Terms of Use | Kaggle. Terms of Use | Kaggle. https://www.kaggle.com/terms

Kaggle. (2021a). 2021 Kaggle Machine Learning & Data Science Survey [dataset]. Kaggle. https://kaggle.com/competitions/kaggle-survey-2021

Kaggle. (2021b). State of Machine Learning and Data Science 2021 (State of Data Science and Machine Learning Survey, p. 45) [Survey]. Kaggle. https://www.kaggle.com/kaggle-survey-2021

Kaggle. (2021c). Kaggle Progression System. Kaggle. https://www.kaggle.com/progression

Kaggle. (2021d). Public API Documentation. Kaggle. https://www.kaggle.com/docs/api

Kaggle. (2022a). 2022 Kaggle Machine Learning & Data Science Survey [dataset]. Kaggle. https://kaggle.com/competitions/kaggle-survey-2022

Kaggle. (2022b). Kaggle's 2022 State of Data Science and Machine Learning survey (p. 27). Kaggle. https://www.kaggle.com/code/paultimothymooney/kaggle-survey-2022-all-results/notebook

Kaggle. (2022c, May 26). Terms of Use. Kaggle. https://www.kaggle.com/terms

Kaggle Team. (2011, November 17). Words of Wisdom From Ben Hamner, Our Newest Recruit. *Kaggle Blog*. https://medium.com/kaggle-blog/words-of-wisdom-from-ben-hamner-our-newest-recruit-f2d6d1ccb93a

Kaggle Team. (2017, September 27). Instacart Market Basket Analysis [Medium]. Kaggle Blog. https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded

Kahn, J. (2020, June 12). Want to find deepfakes? Don't count on A.I. to help you. Fortune. https://fortune.com/2020/06/12/deepfake-detection-contest-facebook/

Kalluri, P. R., Agnew, W., Cheng, M., Owens, K., Soldaini, L., & Birhane, A. (2023, September 26). The Surveillance AI Pipeline. arXiv.Org. https://arxiv.org/abs/2309.15084v2

Kantor, K. (2021). Please Come Off-Book. Button Poetry.

Kember, S., & Zylinska, J. (2014). Life after New Media: Mediation as a Vital Process (reprint edition). The MIT Press.

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. Moral Philosophy and Politics, 8(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

Kevin H. (2017, December 17). As promised, I'll post my architecture. Obviously it must have over-fit to the stage 1 participants [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Kikerpill, K. (2020). Choose your stars and studs: The rise of deepfake designer porn. Porn Studies, 7(4), 352–356. https://doi.org/10.1080/23268743.2020.1765851

King, J. E. (1991). Dysconscious Racism: Ideology, Identity, and the Miseducation of Teachers. Journal of Negro Education, 60(2), 133–146.

Kitchin, R. (2017). Thinking critically about and researching algorithms. Information, Communication & Society, 20(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Knight, W. (2017). Ian Goodfellow. MIT Review Technology: Inovators under 35. https://www.innovatorsunder35.com/the-list/ian-goodfellow/

Kofman, I. (2020, February 4). The dataset was commissioned by Facebook specifically for use in the Deepfake Detection Challenge. It utilizes paid actors with required [Commented on the online forum post IRB-HSR Documentation for Dataset] [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/c/deepfake-detection-challenge/discussion/122113)

Komlos, J. (1992). Anthropometric History: What Is It? OAH Magazine of History, 6(4), 3.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kracauer, S. (1952). The Challenge of Qualitative Content Analysis. Public Opinion Quarterly, 16(4), 631–642. https://doi.org/10.1086/266427

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kramer, T., Spolter-Weisfeld, S., & Thakkar, M. (2007). The Effect of Cultural Orientation on Consumer Responses to Personalization. Marketing Science, 26(2), 246–258. https://doi.org/10.1287/mksc.1060.0223

Kvale, G. (2016). Software as ideology A multimodal critical discourse analysis of Microsoft Word and SmartArt. Journal of Language and Politics, 15(3), 259–273. https://doi.org/10.1075/jlp.15.3.02kva

Lam Dang. (2017, June 27). Free money! [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35399

landoflobsters. (2018, February 7). Update on site-wide rules regarding involuntary pornography and the sexualization of minors [Reddit Post]. R/Announcements. www.reddit.com/r/announcements/comments/7vxzrb/update_on_sitewide_rules_regarding_involuntary/

Langenkamp, M., & Yue, D. N. (2022). How Open Source Machine Learning Software Shapes AI. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 385–395. https://doi.org/10.1145/3514094.3534167

Langlois, G., & Elmer, G. (2019). Impersonal subjectivation from platforms to infrastructures. Media, Culture & Society, 41(2), 236–251. https://doi.org/10.1177/0163443718818374

Lardinois, F., Lynley, M., & Mannes, J. (2017, March 17). Google is acquiring data science community Kaggle. TechCrunch. https://social.techcrunch.com/2017/03/07/google-is-acquiring-data-science-community-kaggle/

Laudan, L. (1977). Progress and its problems: Towards a theory of scientific growth (1st paperback printing). University of California Press.

Lawrence, N. (2015). AI "cheating" scandal makes machine learning sound like a sport – it isn't. The Conversation. http://theconversation.com/ai-cheating-scandal-makes-machine-learning-sound-like-a-sport-it-isnt-43132

Lazzarato, M. (2002). From Biopower to Biopolitics. Pli, 13, 112–125.

Lazzarato, M. (2004). From Capital-Labour to Capital-Life (V. Fournier, A. Virtanen, & J. Vâhâmâki, Trans.). Ephemera, 4(3), 187–208.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), Article 7553. https://doi.org/10.1038/nature14539

Lee-Morrison, L. (2019). Portraits of Automated Facial Recognition: On Machinic Ways of Seeing the Face. transcript Verlag.

Leibowicz, C. (2019, December 11). On AI & Media Integrity: Insights from the Deepfake Detection Challenge. Partnership on AI. https://partnershiponai.org/on-ai-media-integrity-insights-from-the-deepfake-detection-challenge/

Leidos. (n.d.). SafeView Adcanced Image Technical Documentation Brochure. Aeroexpo. Retrieved December 2, 2022, from https://pdf.aeroexpo.online/pdf/leidos/safeview/168784-20211.html#open1880

Leidos. (2020a, May 4). Leidos completes acquisition of L3Harris Technologies' Security Detection and Automation Businesses creating a comprehensive, global security and detection portfolio. Leidos. https://www.leidos.com/insights/leidos-completes-acquisition-l3harris-technologies-security-detection-and-automation

Leidos. (2020b, September 29). Checkpoint technology for frictionless airport screening. Leidos. https://www.leidos.com/insights/checkpoint-technology-frictionless-airport-screening

Lessig, L. (2006). Code: Version 2.0 (2. ed.). Basic Books.

Leszczynski, A. (2018a). Digital methods I: Wicked tensions. Progress in Human Geography, 42(3), 473–481. https://doi.org/10.1177/0309132517711779

Leszczynski, A. (2018b). Digital methods II: Digital-visual methods. Progress in Human Geography, 030913251878797. https://doi.org/10.1177/0309132518787997

Levchin, M. (2013, January 21). DLD13 Keynote [Tumblr]. https://max.levch.in/post/41116802381/dld13-keynote

Lévy, P. (1999). Collective Intelligence (revised edition). Basic Books.

Li, F.-F. (2017, March 8). Welcome Kaggle to Google Cloud. Google Cloud Blog. https://cloud.google.com/blog/products/gcp/welcome-kaggle-to-google-cloud/

Li, S. (2020, January 4). Useful Model Structures [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/124583#748029

Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1–7. https://doi.org/10.1109/WIFS.2018.8630787

Lima, C. (2019, October 25). Why Facebook's plan to rescue journalism may not be enough. POLITICO. https://www.politico.com/news/2019/10/25/facebook-news-tab-058003

Lima, C. (2021, October 25). A whistleblower's power: Key takeaways from the Facebook Papers. Washington Post. https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/

Liu, X., Niu, J., Ran, L., & Liu, T. (2017). Estimation of Human Body Volume (BV) from Anthropometric Measurements Based on Three-Dimensional (3D) Scan Technique. Aesthetic Plastic Surgery, 41(4), 971–978. https://doi.org/10.1007/s00266-017-0872-5

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015, December). Large-scale CelebFaces Attributes (CelebA) Dataset. Proceedings of International Conference on Computer Vision (ICCV). Conference on Computer Vision. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Loiperdinger, M., & Elzer, B. (2004). Lumiere's Arrival of the Train: Cinema's Founding Myth. The Moving Image, 4(1), 89–118. https://doi.org/10.1353/mov.2004.0014

Lombroso, C. (2006). Criminal Man (Gibson & Rafter, Trans.; illustrated edition). Duke University Press.

Long, A. (2017, December 17). Delete Data—Written Acknowledgement [Discussions]. Kaggre Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45903

Lorenzini, D. (2016). Foucault, Regimes of Truth and the Making of the Subject. In L. Cremonesi (Ed.), Foucault and the making of subjects (pp. 63–76). Rowman & Littlefield International, Ltd.

Lovink, G., & Niederer, S. (Eds.). (2008). Video Vortex reader: Responses to Youtube (2nd edition). Inst. of Network Cultures.

Lyft. (2019). Lyft 3D Object Detection for Autonomous Vehicles. Kaggle. https://kaggle.com/c/3d-object-detection-for-autonomous-vehicles

Lytvynenko, J. (2020, July). Instacart Customers' Personal Info Is Being Sold Online. BuzzFeed News. https://www.buzzfeednews.com/article/janelytvynenko/instacart-customers-info-sold-online

Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': Exploring continuities between pornographic and 'political' deep fakes. Porn Studies, 7(4), 415–423. https://doi.org/10.1080/23268743.2020.1757499

Maher, J. H. (2011). The Technical Communicator as Evangelist: Toward Critical and Rhetorical Literacies of Software Documentation. Journal of Technical Writing & Communication, 41(4), 367–401. https://doi.org/10.2190/TW.41.4.d

Make Your Own Deepfakes [Online App]. (n.d.). Deepfakes Web. Retrieved September 10, 2022, from https://deepfakesweb.com/

Malekzadeh, M. (2018). MotionSense Dataset: Smartphone Sensor Data - HAR. Kaggle. https://kaggle.com/malekzadeh/motionsense-dataset

Malini, F., Goveia, F., Ciarelli, P., Carreira, L., Herkenhoff, G., Regattieri, L., & Leite, M. V. (2014). #VemPraRua: Narrativas da Revolta brasileira. GI1: Comunicación Digital, Redes y Procesos. Alaic, Lima, Peru. http://www.labic.net/wp-content/uploads/VemPraRua-Narrativas-da-Revolta-brasileira.pdf

Manovich, L. (2001). The Language of New Media. MIT Press. http://www.manovich.net/LNM/Manovich.pdf

Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. The International Journal of Evidence & Proof, 23(3), 255–262. https://doi.org/10.1177/1365712718807226

Marelli, L., Testa, G., & Van Hoyweghen, I. (2021). Big Tech platforms in health research: Re-purposing big data governance in light of the General Data Protection Regulation's research exemption. Big Data & Society, 8(1), 20539517211018783. https://doi.org/10.1177/20539517211018783

Marino, M. C. (2020). Critical Code Studies. The MIT Press.

Markoff, J. (2012, November 19). Seeking a Better Way to Find Web Images. The New York Times. https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html

Martin, S. (2023, 01). How Does Facebook Suggested Friends Actually Work? Grazia. https://graziadaily.co.uk/life/real-life/facebook-suggested-friends-work/

Mathias, M., Benenson, R., Pedersoli, M., & Van Gool, L. (2014). Face Detection without Bells and Whistles. Computer Vision - ECCV 2014, 8692, 720–735. https://doi.org/10.1007/978-3-319-10593-2_47

Mathien, H. (2016). European Soccer Database. Kaggle. https://kaggle.com/hugomathien/soccer

McKelvey, F. (2018). *Internet daemons: Digital communications possessed*. University of Minnesota Press.

McKelvey, F. R. (2014). Algorithmic Media Need Algorithmic Methods: Why Publics Matter. *Canadian Journal of Communication*, *39*(4), 597–613. https://doi.org/10.22230/cjc.2014v39n4a2746

McKelvey, F., & Frizzera, L. (2019, June 3). The Political Narrative of YouTube Recommendation System: 2018 Ontario election. CCA, Vancouver, Canada. http://labs.fluxo.art.br/rankflow/

Medina, T., & Frank, B. (2019, April 19). TSA Agents Say They're Not Discriminating Against Black Women, But Their Body Scanners Might Be. ProPublica. https://www.propublica.org/article/tsa-not-discriminating-against-black-women-but-their-body-scanners-might-be

Meloy, J. R., & Gothard, S. (1995). Demographic and clinical comparison of obsessional followers and offenders with mental disorders. American Journal of Psychiatry, 152(2), 258–263. https://doi.org/10.1176/ajp.152.2.258

Mendonça, H. (2020, February 3). Proper Clustering with Facenet Embeddings + EDA [Code]. Kaggle Facebook's Deepfake Detection Challenge. https://kaggle.com/code/hmendonca/proper-clustering-with-facenet-embeddings-eda

Mercari. (2018). Mercari Price Suggestion Challenge. Kaggle. https://kaggle.com/c/mercari-price-suggestion-challenge

Metz, C. (2018, April 19). A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit. The New York Times. https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html

Milmo, D. (2021, October 24). Frances Haugen: 'I never wanted to be a whistleblower. But lives were in danger.' The Observer. https://www.theguardian.com/technology/2021/oct/24/frances-haugen-i-never-wanted-to-be-a-whistleblower-but-lives-were-in-danger

miwojc. (2020, December 13). Thousands in credits form AWS [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/121462

moejoe. (2017, December 16). The targets come from the training data. Consider the following sample from the training data [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Mohamed, A., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., & Picheny, M. A. (2011). Deep Belief Networks using discriminative features for phone recognition. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5060–5063. https://doi.org/10.1109/ICASSP.2011.5947494

Mooney, P. (2018). Chest X-Ray Images (Pneumonia). Kaggle. https://kaggle.com/paultimothymooney/chest-xray-pneumonia

Mooney, P. (2022, October 6). Kaggle Survey 2022: All Results [Code]. Kaggle. https://kaggle.com/code/paultimothymooney/kaggle-survey-2022-all-results

Moosa, I. (2017). Econometrics as a Con Art. Edward Elgar Publishing. https://doi.org/10.4337/9781785369957

Morse, J. (2020, June 13). Facebook drives effort to combat deepfakes while enabling misinformation. Mashable. https://mashable.com/article/facebook-deepfake-detection

Mosco, V. (2008). Current Trends in the Political Economy of Communication. 19.

MPWARE. (2019, December 30). EDA - Faces statistics [Code]. Kaggle Facebook's Deepfake Detection Challenge. https://kaggle.com/code/mpware/eda-faces-statistics

Murgia, M. (2019, June 6). Microsoft quietly deletes largest public face recognition data set. Financial Times. https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2

Mushinskiy, S. (2017, June 22). This is insane discrimination and insult to our international community. [Discussion]. Kaggle. https://www.kaggle.com/c/passenger-screening-algorithm-challenge/discussion/35118

Nakhaee, M. (2022, December 9). The Fellowship of Kaggle [Code]. Kaggle. https://kaggle.com/code/imuhammad/the-fellowship-of-kaggle

Napoli, P. M. (2014). Automated Media: An Institutional Theory Perspective on Algorithmic Media Production and Consumption. Communication Theory, 24(3), 340–360. https://doi.org/10.1111/comt.12039

Netflix. (2016). The Power of a Picture. https://about.netflix.com/en/news/the-power-of-a-picture

Netflix. (2009). Netflix Prize. https://web.archive.org/web/20090919150646/http://www.netflixprize.com/index

Netlab. (2023). A guerra das plataformas contra o PL 2630 (p. 10). Netlab UFRJ. https://uploads.strikinglycdn.com/files/2cab203d-e44d-423e-b4e9-2a13cf44432e/A%20guerra%20das%20plataformas%20contra%20o%20PL%202630%20-%20NetLab%20UFRJ,%20Abril%202023.pdf

NFL. (2019). NFL Punt Analytics Competition. Kaggle. https://kaggle.com/c/NFL-Punt-Analytics-Competition

NFL. (2021). NFL Big Data Bowl 2021. Ka. https://kaggle.com/c/nfl-big-data-bowl-2021

Nichol, K. (2015, November 28). How to (almost) win at Kaggle [Unconference]. [Data Science YVR] - Polyglot Software Association Unconference, Vancouver, BC, Canada. https://www.youtube.com/watch?v=JyEm3m7AzkE

NiveditJain. (2021). Human Faces Dataset. Kaggle. https://kaggle.com/niveditjain/human-faces-dataset

NOAA, & MathWorks. (2016). Right Whale Recognition. https://kaggle.com/c/noaa-right-whale-recognition

Noj Vek. (2017, June 28). I do think Kaggle should make a stand against such discrimination. [Commented on This is insane discrimination and insult to our international community] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://kaggle.com/competitions/passenger-screening-algorithm-challenge

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks (arXiv:2103.14749). arXiv. https://doi.org/10.48550/arXiv.2103.14749

Northcutt, C. G., Jiang, L., & Chuang, I. L. (2022). Confident Learning: Estimating Uncertainty in Dataset Labels (arXiv:1911.00068). arXiv. https://doi.org/10.48550/arXiv.1911.00068

nosound [zaharch]. (2020, January 3). Hash conclusions [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/124417

Nunez, M. (2016, August 19). All of the Creepy Things Facebook Knows About You. Gizmodo. https://gizmodo.com/all-of-the-creepy-things-facebook-knows-about-you-1785510980

NxGTR. (2017, June 22). No Ranking Points? Think about it. [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35148

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (reprint edition). Broadway Books.

Outbrain. (2017). Outbrain Click Prediction. Kaggle. https://kaggle.com/c/outbrain-click-prediction

Paiva, V. (2020, January 9). Novo vazamento da Cambridge Analytica aponta eleições manipuladas em escala 'industrial.' Hypeness. https://www.hypeness.com.br/2020/01/novo-vazamento-da-cambridge-analytica-aponta-eleicoes-manipuladas-em-escala-industrial/

Pandey, P. (2019, December 16). *Geek Girls Rising: Myth or Reality*. Medium. https://towardsdatascience.com/geek-girls-rising-myth-or-reality-81e1897433c8

Paris, B., & Donovan, J. (2019, September 18). Deepfakes and Cheap Fakes. Data & Society; Data & Society Research Institute. https://datasociety.net/library/deepfakes-and-cheap-fakes/

Pariser, E. (2012). The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think (reprint edition). Penguin Books.

Parmar, M., Mishra, S., Geva, M., & Baral, C. (2022). Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions (arXiv:2205.00415). arXiv. https://doi.org/10.48550/arXiv.2205.00415

Partnership on AI. (2020). The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity (p. 17). Partnership on AI. http://partnershiponai.org/wp-content/uploads/2021/07/671004_Format-Report-for-PDF_031120-1.pdf

Patel, N. J. (2022, February 4). Fiction Vs. Non-Fiction. Kabuni. https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0

Penn, J. (2018, November 26). AI thinks like a corporation—And that's worrying. The Economist. https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying

Perrigo, B. (2023, January 18). Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer. Time. https://time.com/6247678/openai-chatgpt-kenya-workers/

Petalas, D. P., van Schie, H., & Vettehen, P. H. (2017). Forecasted economic change and the self-fulfilling prophecy in economic decision-making. PLoS ONE, 12(3), e0174353. https://doi.org/10.1371/journal.pone.0174353

phunter. (2017, June 22). According to rule 1 and rule 2 where only a pure blood American team can get the prize [Commented on Welcome!] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35116

Pichai, S. (2016, April 28). This year's Founders' Letter. Google. https://blog.google/alphabet/this-years-founders-letter/

Piketty, T. (2017). Capital in the Twenty-First Century (A. Goldhammer, Trans.; reprint edition). Belknap Press of Harvard University Press.

Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., & Schiele, B. (2017). Building Statistical Shape Spaces for 3D Human Modeling. Pattern Recognition, 67, 276–286. https://doi.org/10.1016/j.patcog.2017.02.018

Plumer, B. (2015, January 23). Why oil prices keep falling—And throwing the world into turmoil. *Vox*. https://www.vox.com/2014/12/16/7401705/oil-prices-falling

Pollack, N. (2012, July 16). The exabyte revolution: How Kaggle is turning data scientists into rock stars. *Wired*. https://www.wired.com/story/the-exabyte-revolution/

Poplavskiy, D. (2017, June 24). The rule to make the whole team not eligible for the prize with one member outside US can cause even more segregation issues. [Commented on This is insane discrimination and insult to our international community.] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://kaggle.com/competitions/passenger-screening-algorithm-challenge

Porto Seguro. (2017). Porto Seguro's Safe Driver Prediction. Kaggle. https://kaggle.com/c/porto-seguro-safe-driver-prediction

Preda, G. (2020, February 5). DeepFake Starter Kit [Code]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/code/gpreda/deepfake-starter-kit

Project MKUTRA, the CIA's Program of Research in Behavioral Modification (p. 173). (1977). The US Senate. https://www.intelligence.senate.gov/sites/default/files/hearings/95mkultra.pdf

Prozorov, S. (2021). Foucault and the birth of psychopolitics: Towards a genealogy of crisis governance. Security Dialogue, 52(5), 436–451. https://doi.org/10.1177/0967010620968345

Quach, K. (2020a, January 21). How a Kaggle Grandmaster cheated in $25,000 AI contest with hidden code – and was fired from dream SV job. The Register. https://www.theregister.com/2020/01/21/ai_kaggle_contest_cheat/

Quach, K. (2020b, June 18). Facebook's $500k deepfake-detector AI contest drama: Winning team disqualified on buried consent technicality. https://www.theregister.com/2020/06/18/facebook_deepfake_kaggle_contest/

Ragnar. (2017, June 24). yeah, really, I'm not welcome in this competition. I think now the reason we have to enter this competition is to learn [Commented on Most of you are NOT Welcome!] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35177

Rao, S. (2017, April 27). Machine Learning Models in Production [Data Science]. DataEngConf 2017, San Francisco, US. https://www.slideshare.net/SharathRao6/lessons-from-integrating-machine-learning-models-into-data-products

Rau, A. (2013). "Psychopolitics" at work: The subjective turn in labour and the question of feminization. Equality, Diversity and Inclusion: An International Journal, 32(6), 604–614. https://doi.org/10.1108/EDI-09-2012-0076

Raviv, S. (2020, January 21). The Secret History of Facial Recognition. Wired, 26.12. https://www.wired.com/story/secret-history-facial-recognition/

Reid, D., Samangooei, S., Chen, C., Nixon, M., & Ross, A. (2013). Soft biometrics for surveillance: An overview (pp. 327–352). Elsevier. https://doi.org/10.1016/B978-0-444-53859-8.00013-8

Reis, R., Zanetti, D., & Frizzera, L. (2020). A conveniência dos algoritmos. Compolítica, 10(1), 35–58. https://doi.org/10.21878/compolitica.2020.10.1.333

Requa, J. (2017, December 16). Congrats Oleg on placing 5th in this comp, thats very impressive! [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Resnick, P. (2005). Impersonal sociotechnical capital, ICTs, and collective action among strangers. In Transforming Enterprise: The Economic and Social Implications of Information Technology (pp. 399–412). The MIT Press. https://doi.org/10.7551/mitpress/6900.001.0001

Reuters. (2016, March 23). Israeli firm helping FBI to open encrypted iPhone: Report. Reuters. https://www.reuters.com/article/us-apple-encryption-cellebrite-idUSKCN0WP17J

Reynolds, M. (n.d.). The AI Doomsday Bible Is a Book About the Atomic Bomb. Wired UK. Retrieved December 26, 2023, from https://www.wired.co.uk/article/the-making-of-the-atomic-bomb-artificial-intelligence

Rhodes, J. (2011, June 27). Competition Shines Light on Dark Matter. Whitehouse.Gov. https://obamawhitehouse.archives.gov/blog/2011/06/27/competition-shines-light-dark-matter

Richards, B. (1978, May 12). Firm's Suits Against CIA Shed Light on Clandestine Air Force. Washington Post. https://www.washingtonpost.com/archive/politics/1978/05/12/firms-suits-against-cia-shed-light-on-clandestine-air-force/a8d77856-6920-40a4-a430-13b9c9171245/

Rieder, B. (2013). Studying Facebook via data extraction: The Netvizz application. Proceedings of the 5th Annual ACM Web Science Conference, 346–355. https://doi.org/10.1145/2464464.2464475

Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results. Convergence: The International Journal of Research into New Media Technologies, 24(1), 50–68. https://doi.org/10.1177/1354856517736982

Roberge, J., & Castelle, M. (Eds.). (2021). *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*. Springer International Publishing. https://doi.org/10.1007/978-3-030-56286-1

Roberts, S. T. (2019). Behind the Screen: Content Moderation in the Shadows of Social Media (illustrated edition). Yale University Press.

Robinette, K. M., Blackwell, S., & Daanen, H. (2002). Civilian American and European Surface anthropometry Resourse (CAESAR). Volume I (AFRL-HE-WP-TR-2002-0169; p. 70). US Air Force, NATO, and SAE. https://www.humanics-es.com/CAESARvol1.pdf

Roccas, S., Sagiv, L., Schwartz, S. H., & Knafo, A. (2002). The Big Five Personality Factors and Personal Values. Personality and Social Psychology Bulletin, 28(6), 789–801. https://doi.org/10.1177/0146167202289008

Rockwell, G., Sinclair, S. G., Ruecker, S., & Organisciak, P. (2010). Ubiquitous Text Analysis. The Journal of the Initiative for Digital Humanities, Media, and Culture, 2(1). http://journals.tdl.org/paj/index.php/paj/article/view/13

Rogers, R. (2009). The End of the Virtual: Digital Methods. Amsterdam University Press. https://doi.org/10.5117/9789056295936

Rogers, R. (2017). Doing Web history with the Internet Archive: Screencast documentaries. Internet Histories, 1(1–2), 160–172. https://doi.org/10.1080/24701475.2017.1307542

Rose, G. (2001). Visual methodologies: An introduction to the interpretation of visual materials. Sage.

Roth, L. (2009). Looking at Shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. Canadian Journal of Communication, 34(1), 111–136.

Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. International Journal of Computer Vision, 126(2–4), 144–157. https://doi.org/10.1007/s11263-016-0940-3

Rowe, N. (n.d.). Millions of Workers Are Training AI Models for Pennies. Wired. Retrieved October 25, 2023, from https://www.wired.com/story/millions-of-workers-are-training-ai-models-for-pennies/

Ruecker, S., Rockwell, G., Sondheim, D., Ilovan, M., Windsor, J., Bieber, M., Frizzera, L., Rodriguez, O., Ranaweera, K., Florentino, C., Sinclair, S., Radzikowska, M., Dobson, T., Blandford, A., Faisal, S., Giacometti, A., Brown, S., Nelson, B., & Michura, P. (2012). The Beginning, The Middle, and The End: New Tools for the Scholarly Edition. Scholarly and Research Communication, 3(4), Article 4. https://doi.org/10.22230/src.2012v3n4a57

Rumbo-Rodríguez, L., Sánchez-SanSegundo, M., Ferrer-Cascales, R., García-D'Urso, N., Hurtado-Sánchez, J. A., & Zaragoza-Martí, A. (2021). Comparison of Body Scanner and Manual Anthropometric Measurements of Body Shape: A Systematic Review. International Journal of Environmental Research and Public Health, 18(12), Article 12. https://doi.org/10.3390/ijerph18126213

Russell, B. (2013, October 24). Common Lotion Ingredient Sets Off Explosives Screening at Airport. NBC 5 Dallas-Fort Worth. https://www.nbcdfw.com/news/local/common-lotion-ingredient-sets-off-explosives-screening-at-airport/1961221/

ryches. (2020, January 2). @pedromb What makes me feel it will be a bit fragile is splitting by a folder may mean that we [Commented on Validation strategies?] [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/c/deepfake-detection-challenge/discussion/124127

saagie_anthony. (2017, May 29). I don't know if we have the same definition of organic in Europe and in the USA.) [Commented on thread Exploratory Analysis—Instacart] [Code]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/philippsp/exploratory-analysis-instacart

SAE. (2004, March 6). CAESAR Fact Sheet. https://web.archive.org/web/20040306183543/http://www.sae.org/standardsdev/tsb/cooperative/caefact.htm

Sako, Y. (2017, December 16). I started out with APS images first. I used various CNNs (VGG, Resnet, Densenet, etc) [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Sanders, R. (2017). Self-tracking in the Digital Era: Biopower, Patriarchy, and the New Biometric Body Projects. Body & Society, 23(1), 36–63. https://doi.org/10.1177/1357034X16660366

Santander. (2016). Santander Product Recommendation. Kaggle. https://kaggle.com/c/santander-product-recommendation

Schmidhuber, urgen. (1991). A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers. In J.-A. Meyer & S. W. Wilson (Eds.), International Conference on Simulation of Adaptive

Behavior: From Animals to Animats (pp. 222–227). The MIT Press. https://doi.org/10.7551/mitpress/3115.003.0030

Schmidhuber, J. (2020). Generative Adversarial Networks are Special Cases of Artificial Curiosity (1990) and also Closely Related to Predictability Minimization (1991) (arXiv:1906.04493). arXiv. https://doi.org/10.48550/arXiv.1906.04493

Schor, J. B., Walker, E. T., Lee, C. W., Parigi, P., & Cook, K. (2015). On the Sharing Economy. Contexts, 14(1), 12–19. https://doi.org/10.1177/1536504214567860

Schreier, M., & Flick, U. (2014). Qualitative Content Analysis. In The SAGE Handbook of Qualitative Data Analysis (pp. 171–183). Sage. https://doi.org/10.4135/9781446282243

Schroepfer, M. (2019, September). Creating a dataset and a challenge for deepfakes. https://ai.facebook.com/blog/deepfake-detection-challenge/

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Scott, D. (2020, February 6). Deepfake Porn Nearly Ruined My Life. ELLE. https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/

Seferbekov, S. (2020, February 28). @addisonhoward @juliaelliott As external data might be very helpful to obtain better scores on public/private leaderboards how are you [Commented on the online forum post External Data Disclosure Thread] [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/121203

Sekula, A. (1986). The Body and the Archive. October, 39, 3–64. https://doi.org/10.2307/778312

sh1ng. (2017, August 7). Baseline 0.4029970 [Discussion]. Kaggle Instacart Market Basket Analysis. https://www.kaggle.com/hippskill/competitions

Sheen, D. M., McMakin, D. L., & Hall, T. E. (2001). Three-dimensional millimeter-wave imaging for concealed weapon detection. IEEE Transactions on Microwave Theory and Techniques, 49(9), 1581–1592. https://doi.org/10.1109/22.942570

Sheoran, H. (2020, January 20). @pratikbhangale, Sry bro no place for newbies, if you are new in CV and wanna try this competition I guess [Commented on Looking for a Team Thread] [Discussion]. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/121201

Shepard, M. (Ed.). (2011). Sentient City: Ubiquitous Computing, Architecture, and the Future of Urban Space. The MIT Press.

Sheriff-Hussaini, Y. (2017, August 8). I am dumbfounded!!! After a lot of hard work, [Commented on the thread This competition is disappointing] [Discussion]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/competitions/instacart-market-basket-analysis

Shin, H., & Yi, H. Y. (2022, March 7). South Korea candidates woo young voters with "deepfakes," hair insurance. Reuters. https://www.reuters.com/world/asia-pacific/skorea-candidates-woo-young-voters-with-deepfakes-hair-insurance-2022-03-03/

Shirky, C. (2009). Here Comes Everybody: The Power of Organizing Without Organizations (reprint edition). Penguin Books.

Shirky, C. (2011). Cognitive surplus: How technology makes consumers into collaborators. Penguin Books.

Schor, J. B., Walker, E. T., Lee, C. W., Parigi, P., & Cook, K. (2015). On the Sharing Economy. *Contexts*, *14*(1), 12–19. https://doi.org/10.1177/1536504214567860

Simonite, T. (2018, July 21). Despite Pledging Openness, Companies Rush to Patent AI Tech | WIRED. Wired. https://www.wired.com/story/despite-pledging-openness-companies-rush-to-patent-ai-tech/

Slack, J. D., & Wise, J. M. (2014). Culture and Technology: A Primer (4th edition). Peter Lang Inc., International Academic Publishers.

Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., & Zisserman, A. (2020). A Short Note on the Kinetics-700-2020 Human Action Dataset. arXiv:2010.10864 [Cs]. http://arxiv.org/abs/2010.10864

Soleyman, S. (2017, December 15). Congratulations to everyone who participated in this challenging and suspenseful competition! [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Solon, O. (2019, March 12). Facial recognition's "dirty little secret": Social media photos used without consent. NBC News. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

Solsman, J. (2018, January 10). Ever get caught in an unexpected hourlong YouTube binge? Thank YouTube AI for that. CNET. https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan/

Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., & LeCun, Y. (2007). The Need for Open Source Software in Machine Learning. Journal of Machine Learning Research, 8, 2443–2466.

Spachtholz, P. (2017a, June 30). Exploratory Analysis—Instacart [Code]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/philippsp/exploratory-analysis-instacart

Spachtholz, P. (2017b, August). Honestly, before doing something like this, could you please consider that for some people (like me) competition rankings [Commented on thread Baseline 0.4029970] [Discussion]. Kaggle Instacart Market Basket Analysis. https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/37697

Springleaf. (2015). Springleaf Marketing Response. Kaggle. https://kaggle.com/c/springleaf-marketing-response

Srnicek, N. (2017). Platform Capitalism. Polity.

Stanley, J. (2017a, March 29). Deep Learning with Emojis (not Math). Medium. https://tech.instacart.com/deep-learning-with-emojis-not-math-660ba1ad6cdc

Stanley, J. (2017b, May 4). 3 Million Instacart Orders, Open Sourced. Medium. https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

Stanley, J. (2017c, June 5). Instacart AMA [Discussions]. Kaggle Instacart Market Basket Analysis. https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/34200

Stark, L. (2018). Algorithmic psychometrics and the scalable subject. Social Studies of Science, 48(2), 204–231. https://doi.org/10.1177/0306312718772094

State Farm. (2016). State Farm Distracted Driver Detection. Kaggle. https://kaggle.com/c/state-farm-distracted-driver-detection

Steinberg, M. (2019). The platform economy: How Japan transformed the consumer Internet. University of Minnesota Press.

Stephens, E. (2013, June 20). Francis Galton's Composite Portraits: The Productive Failure of a Scientific Experiment.

Stiegler, B. (2000). Biopower, psychopower and the logic of the scapegoat | Ars Industrialis. Ars Industrialis - Association Internationale Pour Une Politique Industrielle Des Technologies de l'Esprit. https://arsindustrialis.org/node/2924

Surowiecki, J. (2005). The Wisdom of Crowds (reprint edition). Anchor.

svj24. (2017, August 15). My Learnings from 18th place solution [Discussion]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/competitions/instacart-market-basket-analysis

Syll, L. (2018). The main reason why almost all econometric models are wrong | World Economics Association. World Economics Association Commentaries, 8(3), 5–10.

Synced. (2020, June 14). Facebook and Kaggle Face Backlash After Disqualifying Apparent 'Deepfake Detection Challenge' Winner | Synced. Synced. https://syncedreview.com/2020/06/14/facebook-and-kaggle-face-backlash-after-disqualifying-apparent-deepfake-detection-challenge-winner/

Taplin, J. (2017). Move Fast and Break Things: How Facebook, Google, and Amazon Cornered Culture and Undermined Democracy. Little, Brown and Company.

Tardáguila, C. (2022, August 18). Eleições: 1ª deepfake mostra pesquisa falsa na voz de Renata Vasconcellos. UOL. https://noticias.uol.com.br/colunas/cristina-tardaguila/2022/08/18/eleicoes-1-deep-fake-mostra-pesquisa-falsa-na-voz-de-renata-vasconcellos.htm

Tenorio [DATATHÈQUE], G. (2017, September 18). Association Rules Mining/Market Basket Analysis [Notebook]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/datatheque/association-rules-mining-market-basket-analysis

Terhörst, P., Fährmann, D., Damer, N., Kirchbuchner, F., & Kuijper, A. (2020). Beyond Identity: What Information Is Stored in Biometric Face Templates? 2020 IEEE International Joint Conference on Biometrics (IJCB), 1–10. https://doi.org/10.1109/IJCB48548.2020.9304874

Terranova, T. (2004). Network Culture: Politics for the Information Age. Pluto Press.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.

The Markup. (2020, December 3). Why Web Scraping Is Vital to Democracy – The Markup. The Markup. https://themarkup.org/news/2020/12/03/why-web-scraping-is-vital-to-democracy

The White House. (2023, October 30). FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Thornton, J. (2017, December 18). Prototype Question [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/43374

Tibbetts, S. G., & Hemmens, C. (2009). Criminological Theory: A Text/Reader. Sage.

Tiku, N. (2018, October 10). Why Netflix Features Black Actors in Promos to Black Users. Wired. https://www.wired.com/story/why-netflix-features-black-actors-promos-to-black-users/

Tiwari, D. K. (2021, March 19). Analysis Of Instacart From Kaggle Competition. Medium. https://medium.com/analytics-vidhya/analysis-of-instacart-from-kaggle-competition-12a26a24ef4b

Tjaden, P. G., & Thoennes, N. (1998). Stalking in America: Findings from the National Violence Against Women Survey. U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Trinh, L., & Liu, Y. (2021). An Examination of Fairness of AI Models for Deepfake Detection. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 567–574. https://doi.org/10.24963/ijcai.2021/79

Trott, O. (2017a, September 24). TLDR: I think it's a feature, rather than a bug [Commented on Will the same volunteers appear in the stage 2 images multiple times?] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/39484

Trott, O. (2017b, October 14). Corrupted scans and incorrect labels [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/41223

Trott, O. (2017c, December 16). My CV was 0.032-ish (No subject was in multiple CV folds, of course). [Commented on Post Competition Architecture Discussion] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/45805

Tufekci, Z. (2018, March 10). YouTube, the Great Radicalizer. The New York Times. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Tunguz, B. (2019a). 1 Million Fake Faces. Kaggle. https://www.kaggle.com/datasets/tunguz/1-million-fake-faces

Tunguz, B. (2019b, December 13). Face Reconstruction from Voice using Generative Adversarial Networks. Kaggle Facebook's Deepfake Detection Challenge. https://www.kaggle.com/competitions/deepfake-detection-challenge/discussion/121455

Turow, J. (2017). The Aisles Have Eyes: How Retailers Track Your Shopping, Strip Your Privacy, and Define Your Power. Yale University Press.

Twitter. (2021). Use Cases, Tutorials, & Documentation. Twitter Developer Page. https://developer.twitter.com/en

Two Sigma. (2019). Two Sigma: Using News to Predict Stock Movements. Kaggle. https://kaggle.com/c/two-sigma-financial-news

University of Exeter. (2011). 2011 Loebner Prize. https://loebner.exeter.ac.uk/

University of Melbourn. (2011). Predict Grant Applications. Kaggle. https://kaggle.com/c/unimelb

University of Saskatchewan. (2020). Global Wheat Detection. Kaggle. https://kaggle.com/c/global-wheat-detection

Uthman, A. [عثمان]. (2017a, July 1). Absolutely. The US government paid an Israeli firm to hack the couples iPhone. I don't see why they can't pay [Commented on the online forum post The issue is... The prize is payed by USA Federal money] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35332

Uthman, A. [عثمان]. (2017b, July 8). All the controversy behind this competition / dataset / DHS data provider reminds of a Netflix series titled "The Machine". Spoiler alert [Commented on the online forum post Our thinking in hosting this prize] [Discussion]. Kaggle Passenger Screening Algorithm Challenge. https://www.kaggle.com/competitions/passenger-screening-algorithm-challenge/discussion/35977

van Dijck, J. (2013). The Culture of Connectivity: A Critical History of Social Media. Oxford University Press.

van Dijck, J., & Nieborg, D. (2009). Wikinomics and its discontents: A critical analysis of Web 2.0 business manifestos. New Media & Society, 11(5), 855–874. https://doi.org/10.1177/1461444809105356

van Dijck, J., Poell, T., & Waal, M. de. (2018). The Platform Society. Oxford University Press.

Vance, A. (2018, May 15). This Man Is the Godfather the AI Community Wants to Forget. Bloomberg.Com. https://www.bloomberg.com/news/features/2018-05-15/google-amazon-and-facebook-owe-j-rgen-schmidhuber-a-fortune

Varian, H. (2009, February 2). Hal Varian on how web challenge managers (J. Manyika, Interviewer) [McKinsey Quarterly 1(2.2)]. https://web.archive.org/web/20090202002651/http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286

Varian, H. R. (2010). Computer Mediated Transactions. American Economic Review, 100(2), 1–10. https://doi.org/10.1257/aer.100.2.1

Venturini, T., Ricci, D., Mauri, M., Kimbell, L., & Meunier, A. (2015). Designing Controversies and Their Publics. Design Issues, 31(3), 74–87. https://doi.org/10.1162/DESI_a_00340

Vigen, T. (2015). Spurious Correlations. Hachette Books.

Vincent, J. (2018, January 12). Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech—The Verge. The Verge. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Vincent, J. (2020, June 12). Facebook contest reveals deepfake detection is still an "unsolved problem." The Verge. https://www.theverge.com/21289164/facebook-deepfake-detection-challenge-unsolved-problem-ai

Waldrop, M. M. (2020). Synthetic media: The real trouble with deepfakes. Knowable Magazine | Annual Reviews. https://doi.org/10.1146/knowable-031320-1

Wall, S., & Schellmann, H. (2021, June 23). LinkedIn's job-matching AI was biased. The company's solution? More AI. MIT Technology Review. https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/

Wang, J. (2018). Carceral capitalism. Semiotext(e).

Warofka, A. (2018, November 6). An Independent Assessment of the Human Rights Impact of Facebook in Myanmar. About Facebook. https://about.fb.com/news/2018/11/myanmar-hria/

Wayback Machine. (n.d.). Retrieved April 17, 2022, from https://web.archive.org/web/*/kaggle.com

Wells, J. (2017, November 22). Instacart: Strike last week had no impact on operations. Food Dive. https://www.fooddive.com/news/grocery--instacart-strike-earlier-this-week-had-no-impact-on-operations/511521/

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A Discriminative Feature Learning Approach for Deep Face Recognition. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision – ECCV 2016 (Vol. 9911, pp. 499–515). Springer International Publishing. https://doi.org/10.1007/978-3-319-46478-7_31

Whitson, J. R. (2013). Gaming the Quantified Self. Surveillance & Society, 11(1/2), 163–176. https://doi.org/10.24908/ss.v11i1/2.4454

Whittaker, Z. (2022, April 18). Web scraping is legal, US appeals court reaffirms. TechCrunch. https://techcrunch.com/2022/04/18/web-scraping-legal-court/

Wiggers, K., & Coldewey, D. (2022, May 8). Perceptron AI Roundup: Bias, computer vision and wave action. TechCrunch. https://social.techcrunch.com/2022/05/08/perceptron-ai-bias-can-arise-from-annotation-instructions/

Williams, R. (2015). Keywords: A vocabulary of culture and society (new edition). Oxford University Press.

World Economic Forum (Director). (2013, August 27). Technology Pioneer 2014. https://www.youtube.com/watch?v=OShGuf7QeJY

World Economic Forum (Director). (2015, January 22). Davos 2015—A Brave New World. https://www.youtube.com/watch?v=wGLJXO08IYo

World Wide Web Foundation. (2019). The invisible curation of content: Facebook's News Feed and our information diets. World Wide Web Foundation. https://webfoundation.org/research/the-invisible-curation-of-content-facebooks-news-feed-and-our-information-diets/

wzamen01 (Director). (2009, December 10). HP computers are racist. https://www.youtube.com/watch?v=t4DT3tQqgRM

Xu, Y., Terhörst, P., Raja, K., & Pedersen, M. (2022). A Comprehensive Analysis of AI Biases in DeepFake Detection With Massively Annotated Databases (arXiv:2208.05845). arXiv. http://arxiv.org/abs/2208.05845

xxxxyyyy80008. (2022, October 19). Education and Age by Country—2018 and 2022 [Code]. Kaggle. https://kaggle.com/code/xxxxyyyy80008/education-and-age-by-country-2018-and-2022

yananchen. (2017, July 25). How many features are best suitable for the problem ? [Discussion]. Kaggle Instacart Market Basket Analysis. https://kaggle.com/competitions/instacart-market-basket-analysis

Yeung, K. (2016). 'Hypernudge': Big Data as a Mode of Regulation by Design'. Information, Communication & Society, 32(1), 32.

Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future. Computer Vision and Image Understanding, 138, 1–24. https://doi.org/10.1016/j.cviu.2015.03.015

Zeeshan, M. (2020). Diabetic Retinopathy Blindness Detection c_data. Kaggle. https://kaggle.com/makrovh/diabetic-retinopathy-blindness-detection-c-data

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. Journal of Information Technology; London, 30(1), 75–89. http://dx.doi.org.lib-ezproxy.concordia.ca/10.1057/jit.2015.5

Zuboff, S. (2020). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (illustrated edition). PublicAffairs.