NORTHWESTERN UNIVERSITY

Data Leverage: A Framework for Empowering the Public to Mitigate Harms of Artificial Intelligence

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Technology and Social Behavior

By

Nicholas Vincent

EVANSTON, ILLINOIS

December 2022

ABSTRACT

Data Leverage: A Framework for Empowering the Public to Mitigate Harms of Artificial Intelligence

Nicholas Vincent

Many computing technologies are primarily useful because of the existence of some set of data created by people, intentionally in some cases and unintentionally in others. For instance, technologies like search engines, recommender systems, classifiers, and language models are all dependent on digital records of things people have said, done, typed, clicked and experienced. In practice, the creation of this useful data involves the participation by, or surveillance of, members of the public. One way to view this relationship is to say that computing technologies are reliant on *data labor*, and that people who perform data labor have a potential source of leverage – which we might call *data leverage* – over the operators of computing technologies, i.e. technology companies and other large organizations. Identifying new ways to empower the public is important in light of growing concerns that advances in computing – especially around artificial intelligence, machine learning, and statistics – will contribute to inequality in power and wealth in addition to creating negative impacts along other dimensions. In this thesis, we describe work that

has sought to understand and support data leverage along several fronts: by measuring the value of data to existing systems and platforms, by estimating the potential impact of data leverage actions, and by developing frameworks and tools that support data leverage.

Acknowledgements

This has been a work of enormous collective effort. In one sense, writing an acknowledgments section requires a kind of attribution, not so dissimilar to the task of data value estimation. In this very thesis, we study data value estimation in a very zoomed in fashion, and argue for designers to lean more towards collective valuation and away from individual valuation. In this spirit, I will not claim that I can accurately assess how everyone has shaped me as I pursued this research agenda over many years. Nonetheless, there are certainly some individuals and small groups that had, we might say, an outsized impact on my career, and I will try to name a few here.

Of course, I must thank my committee and especially my advisor Brent. Without their guidance I simply could not have even begun to tackle the research problems I did, and I certainly would have faltered many times. I often advise prospective or new graduate students that nearly all the important factors that determine your success and happiness in graduate school depend on your advisor, and I have been enormously fortunate in that regard.

Many communities were absolutely critical to providing me with a wonderful support system. The broad HCI community, including many people who I met only at conferences or online, shared their perspectives and ideas with me. I owe thanks in particular to my colleagues at Northwestern, especially to those in TSB, MTS, and CS. I can also confidently attribute extra thanks to GroupLens, the CollabLab, and the Community

Data Science Collective, who all provided a constant stream of feedback and support. My mentors and colleagues at Snap and Microsoft broadened my perspective, and helped me keep pushing forward on research during a challenging time of changing work practices. The RadicalxChange community has been a huge inspiration.

I am so thankful for every scholar and researcher with whom I have had the chance to co-author work. Collaboration is at the heart of what makes research effective – and enjoyable – and I could not have asked for a better set of people to work with over these many years. I am also so appreciative of the steadfast proofreading and questions from my mother, who is always able to provide a fresh perspective and find the acronyms I failed to define.

Finally, I owe so many thanks to my friends and family, who make life worth living and research worth pursuing, one chunk at a time. And most of all, I am deeply grateful for the support of my wife. To her, I dedicate the positive outcomes of this research, and to myself I dedicate the typos and ideas that will be challenged in years to come.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Artificial intelligence (AI) technologies have seen rapid development in recent years, with a flurry of financial investment in AI activities and the development of impressive new AI capabilities [436]. However, there is mounting evidence that AI technologies can cause a great deal of societal harm along many dimensions [3, 11, 301]. One of these concerns is the potential for AI technologies to cause a significant increase in economic inequality [332, 212, 258]. In conjunction with other documented harms of AI systems, e.g., the potential to reflect and reinforce historical biases [11, 301], there is serious concern that AI and data-dependent systems will do more harm than good in many contexts. Concretely, even AI technologies that create economic growth could leave many people worse off. As the computing research community races to advance AI technologies in pursuit of large-scale benefits, there is a critical need to develop techniques to ensure the AI systems are governed more democratically, the winnings of AI systems are shared more broadly, and negative impacts across a variety of dimensions are mitigated.

The research in this thesis is rooted in the hypothesis that, by changing how AI systems are designed and governed, AI can mitigate inequalities in wealth and power rather than exacerbate them. The goal of such systems would be to work towards a computing paradigm that distributes its economic winnings and non-monetary utility much more broadly. The research community can help build this new paradigm by studying and scaffolding the relationships between the data-generating public and the AI technologies they fuel with their data contributions. For instance, search engines rely on click data from users and content written by volunteers such as blogs and Wikipedia articles [408, 404]. Recommender systems rely on explicit user feedback (e.g., “star ratings”) and behavioral data (e.g., browsing history) that reveal user preferences. The entire body of research on supervised learning relies on crowd workers, volunteers, and sometimes unwitting users (e.g., reCAPTCHA participants) to label images and text. Without this data generated by the public, data-driven technologies (i.e., those that use machine learning and statistical models) could not exist. The critical role of all these kinds of data suggests a yet untapped source of power for the broad public.

The work described here aims to address power and information asymmetries between data-contributing users and AI-operating technology companies. The operators of AI technologies are cognizant of the essential nature of human-generated data, but users have limited understanding of this data’s essential role in making these technologies work (though this is changing [242]). By building systems to measure and make people aware of the value of their data, we believe it is possible to reduce this problematic knowledge imbalance. In doing so, we aim to help people use the leverage inherent in their essential role as data generators to develop symbiotic relationships between data generators – the public – and the operators of AI systems. These relationships would involve data contributors pushing back when data is used in a way they view as undesirable, supporting causes they deem desirable (e.g., providing preference data to a new start-up) or even receiving financial incentives for data contributions via a data market [63] or a data dividend [111]. Such relationships could – and should – be mutually beneficial, leading to reduced economic inequality *and* better-performing AI technologies.

Our research efforts to address this information asymmetry have centered around a two step process: (1) measuring the value of data – in terms of potential impact – to make people aware of this value and (2) identifying avenues to translate this potential impact into actualized leverage. If the value of data is not made salient and discussed prominently, it is very unlikely that any kind of dialogue or negotiations will ever begin. Thus, measuring and making people aware of data's value can be seen as a prerequisite for changing the relationship between data contributors and data users. However, as information asymmetry around data contributions is reduced, it will also be critical to identify how actual leverage can be achieved. A key theme in our results on this front is to show the importance of collective action. Data records have much more value when they are aggregated and reflect something about groups of people, so gaining leverage through data will require coordination.

1.1. Overview of Chapters

This thesis is organized around the two step process described above: first, we discuss efforts to measure the value of datasets created by the public, and then discuss efforts to support high-impact collective action based around data. Our work in the first category has focused specifically on investigating the relationships between data contributions – especially contributions to the peer production platform Wikipedia (i.e., voluntarily-provided "data labor") – and for-profit platforms and systems like Reddit, StackOverflow, Google, Bing, and DuckDuckGo. Our work in the second category has involved developing a framework for thinking about different kinds of data-related collective action. This framework, which we call "data leverage" provides an actionable path towards empowering the public to engage in collective action. We have worked to make this framework more concrete by using simulations to estimate the potential impact of different kinds of data-related collective action.

Both of these categories are highly interdisciplinary, primarily sitting at the intersection of human-computer interaction and machine learning. In Chapter 2, we discuss the specific subfields that informed our work. We explain how data leverage is primarily rooted in human-computer interaction (HCI) but draws heavily on machine learning scholarship. On the HCI side, we provide a broad summary of how data leverage relates to ongoing efforts within HCI to address the negative impacts of computing, and especially highlight the potential of data leverage to address this major challenge. On the ML side, this section highlights some of the broad connections between ML research and data leverage, and discusses the role of "data valuation research" in particular. We also discuss connections to AI governance and differing views on the political economy of data. Finally, Chapter 2 concludes by introducing some definitions that are relevant throughout the following chapters.

The role that human-generated data has played in fueling computing advances, especially in machine learning, has been at times undervalued. This has led to a status quo wherein most members of the public – each of whom plays a critical role in the success of data-dependent technologies – are not given information about the value of their contributions. Chapters 3, 4 and 5 describe our work measuring the value of the public's data contributions. These projects pay particular attention to data contributions to Wikipedia, building on prior work suggesting Wikipedia is an undervalued dependency underlying powerful computing technologies, and take advantage of Wikipedia's unique transparency.

Chapter 3 looks at the relationship between Wikipedia content and other online communities (we look specifically at Reddit and Stack Overflow, but discuss the likelihood that similar relationships exist for many other communities and platforms). For this study, we considered both a simple upper-bound counterfactual and used propensity score methods to estimate the effect of posting Wikipedia links on engagement. We estimated the increased engagement may contribute to firm revenue on the order of magnitude of $100k per year. This work provided a specific measurement of how value flows from volunteer-produced data to for-profit companies. Further, for research on online communities, our results suggest there is much to be gained by studying online communities with a broad lens that includes the relationships between communities. Chapters 4 and 5 describe several projects that investigated the role of user-generated content, and Wikipedia content specifically, in populating search engine results. This work looked at a variety search engines and search contexts to see how often volunteer-generated data labor was responsible for the content appearing in prominent parts of search engine results. Overall, our results suggest that Wikipedia is an absolutely critical source for search engines to address user needs.

In Chapters 6, 7, and 8, we describe work focused specifically on imagining and supporting collective action around data. Chapter 6 describes simulation work aimed at understanding the potential impact of data strikes. Chapter 7 expands on the work of the previous chapter, and further describes how we might compare the relative impact of collective action aimed at withholding data to bring down a system against collective action aimed at contributing data to boost up an alternative system. Together, these results validate that data leverage can be very effective, but importantly also show how simulation can be an important way to check when certain kinds of collective action may actually be less impactful and not worth the requisite costs. Chapter 8 describes the overall framework of data leverage. Looking across a variety of different research areas, including ML, HCI, and economics, we identified three distinct data levers available to a group of people acting collectively: withholding data as part of a data strike, manipulating data as part of a data poisoning attack (a concept that's been well studied in the ML community, albeit primarily as anti-social, not pro-social), and redirecting/donating data as part of a conscious data contribution campaign. This work also highlighted the complementary nature of data strikes and data contribution: even if a group is too small to effectively data strike, data contribution can be very effective, oftentimes exhibiting a kind of "80/20 rule" wherein 20% of users can boost a system up to 80% of best-case performance. In conjunction with recent empirical work that shows that many deep learning systems show power law data scaling [169], it appears that data contribution can be a highly impactful source of leverage as well.

Data leverage could be used to change organizational policy along a number of dimensions, but in contexts where data-dependent systems are very profitable, a natural application of data leverage may be to push firms or states to increase the share of economic winnings that flow back to upstream data contributors. Chapter 9 provides an analysis of a "data dividends", a high-stakes policy discussion that is related to data value estimation and data-related collective action. The basic idea behind a data dividend is that one way of recognizing the value that the public provides to AI systems is to give people money. A data dividend is one potential outcome from collective action, e.g. the public could withhold data until the state or a firm shares some of the profits from AI. The data dividend concept raises many questions, especially around how data contributions should be valued. To understand the design space of data dividends, we conducted a series of simulations aimed at understanding how different design choices affect how the payments disbursed as part of a "meritocratic" data dividends (i.e., a dividend where more "valuable" or "impactful" data contributions are given larger payments) vary in terms of their equality and distributions.

The remainder of the thesis primarily describes future work. Chapter 10 describes ripe ground for connecting growing research on data value estimation, primarily taking place in the ML research community, with a collective action perspective on data. In this chapter, we provide a review of how different data value definitions (e.g., a "leave one out" definition of impact vs. a "Shapley value" definition) relate to data leverage.

Finally, in Chapter 11, we describe some of our planned next steps. A high priority direction for extending the data leverage research agenda is to build tools that visualize or otherwise communicate the impact of different data-affecting actions to users. Such tools would very directly attack the information asymmetry around data's value and impact, aiming to provide as much information as possible to data contributors about their contribution's downstream impact. This will require incorporating ideas and findings from the work described in all the preceding chapters of the thesis. These tools also have applications for a broader class of governance questions in online settings. In this concluding Chapter, we also discuss the implications of adopting large models, e.g. large language models, for data leverage initiatives.

1.2. Stylistic Notes

First, throughout this document, we will use the plural first person pronouns to refer to collaborative efforts underlying the various papers described herein (e.g., “our work”, “we argue”). Indeed, each of the individual projects described here was highly collaborative.

Second, many of the chapters describe work that was already published elsewhere. In each of these chapters, we note the original publication venue in a footnote. While we have made some updates to these chapters to better communicate a cohesive research narrative, we have left the Discussion section of each chapter mostly untouched to capture our thinking throughout this work’s lifecycle.

CHAPTER 2

Related Work

2.1. Chapter Overview

In this Chapter, we first discuss where the work in this thesis sits relative to existing research areas. In each of the subsequent chapters, we discuss specific related work, so in order to avoid excessive redundancy, this Chapter is focused on situating data leverage research at a high level.

The study of data leverage sits at the intersection of human-computer interaction (HCI) and machine learning (ML). Ultimately, the primary phenomenon of interest in most of the studies described here are about interactions that lead to data records, i.e. humans interacting with computers. For instance, our studies discussed in Chapters 3, 4, and 5 focus heavily on Wikipedia and other online communities that have long been of interest to HCI scholars. The impact of these data records can amplified be techniques and ideas taken from machine learning research. Our studies discussed in Chapters 6 and 7 involve using evaluation techniques from the ML community.

With this in mind, much of the data leverage research described here can primarily be seen as falling under the nascent disciplinary boundaries of *human-centered machine learning* (a growing community, see e.g. [131]) and *responsible artificial intelligence* (see e.g. [14]). In fact, data leverage provides a very practical definition of “responsible AI”: a world in which data leverage actions are cheap and common will lead to AI systems that are "responsible" to the broad public in a very immediate sense.

While primarily situated in HCI scholarship and secondarily in machine learning, this line of work has also been heavily influenced by a scholarly initiative in economics to take seriously the idea of treating "data as labor" [13, 309]. The core idea, as explained by Arrieta Ibarra et al. [13], is that data has been typically treated as a kind of capital that is produced in a manner akin to exhaust from a vehicle, but should instead be seen as work continually supplied by "data laborers". This insight motivated our earlier work preceding "data leverage", and is also crucial to the argument underlying data leverage. Just as labor exerts leverage against capital, data labor may also exert leverage against data-dependent capital.

2.2. Human-computer interaction

Human-computer interaction (HCI) scholarship has been a central force in paving the way for studying human factors of data-dependent technologies. As such, our work relies on, and contributes to, scholarship in HCI.

2.2.1. Computer supported cooperative work and social computing

Within HCI, scholarship centered around *Computer supported cooperative work* (CSCW) is especially relevant to discussing data leverage. CSCW is a rather broad field, with early work focused on "groupware" like Lotus Notes used in organizations [147]. This early work certainly involves labor and data, but is our primary focus. Later work has focused increasingly on large online platforms, e.g. social media like Twitter and "modern forums" like Reddit. These online platforms are important sites of data labor and have been responsible for much of the "early data labor" fueling data-dependent computing systems (e.g., Wikipedia fueling search and natural language processing).

To study data labor almost invariably involves studying social computing platforms. While it is certainly possible to find examples of data labor that are not facilitated through platforms (e.g., machine learning labels generated entirely in a lab setting), a dominant portion of data-dependent computing is specifically dependent on data from the online platforms popular in social computing.

Efforts to better understand and support social computing platforms and online communities have, intentionally or not, also helped to document the role of users in generating data that has been crucial to the success of several research areas (e.g., studying the production of Wikipedia [122], documenting user contributions to Reddit [27]). Much of our early data leverage work was inspired specifically by calls for understanding the relationship between Wikipedia and Google [385, 265].

Another impactful concept originating in CSCW scholarship has been that of non-use [340]. This line of work has argued for designing for people who cannot use, or refuse to use certain technologies or features. Data leverage is very reliant on taking "non-users" seriously. Under a worldview that ignores non-use, it would be very easy to assume that "users" should or would have no choice but to consider generating data exhaust.

Importantly, the CSCW community has also taken the lead in discussing the role of power dynamics in core issues in computing. [51, 82]. For instance, scholars in this field have sought to shift discussions of "human-centered design" towards advocacy and activism [342].

2.2.2. Computer supported cooperative work and social computing

2.3. Machine Learning and Artificial Intelligence

Our work draws on scholarship in machine learning. While the general ideas around data leverage apply broadly to any system that is "data dependent", i.e. relies on some degree of upstream human-generated data, many highly impactful data-dependent technologies are specifically *machine learning* technologies. Furthermore, ideas from the machine learning community are critical for conceptualizing data dependent systems.

Our prior work (outlined above) has shown the potential value of simulating various configurations of so-called "data leverage". For instance, we can predict the drop in performance of a recommender model in the scenario that 30% of all users withhold their data (for a popular recommender task, we see performance drop of 50% of the way towards an unpersonalized baseline, see Chapter 6). We can also imagine investigating more specific scenarios, such as, "what if users who are top ten percentile contributors withhold half of their data contributions"? The knowledge creation practices familiar in machine learning scholarship are well suited to explore these counterfactuals and give plausible estimates of impact. Indeed, some recent work has begun to explicitly call such explorations "data counterfactuals"[182]. Put simply, for a particular imagined configuration of data leverage, it is possible to run and analyze a corresponding simulation.

2.3.1. Data Valuation

Driven in part by developments in data economics (i.e. emerging ideas such as data markets [63] and data dividends, as discussed in Chapter 9), there have been many recent advances in *data valuation* techniques [206, 126, 191]. Data valuation techniques involve estimating the value of units and groups of data (i.e. "datums", or as we will call them, observations) in terms of their impact on a loss calculation or other metric.

Work has sought to price individual "observations" of data. Several effective techniques for valuing observations have been proposed, including influence functions [206] and various forms of Shapley values [126, 191] (a concept from game theory that can be applied to data by imagining the creation of a dataset as a cooperative game). These have also been extended to groups of observations, which is particularly relevant to questions of data leverage.

Critically, many of the data valuation approaches that have been the subject of research advances can be interpreted very concretely in terms of counterfactual data scenarios.

After describing the conceptual development of "data leverage" in Chapter 8 and the preceding chapters, and showing the very direct connections between data valuation and data dividends in Chapter 9, we provide a more in-depth discussion of the connections between data leverage and research on data value estimation in Chapter 10.

2.4. Other Relevant Subfields

2.4.1. AI Governance and Alignment

There is growing interest in coalescing a research community around the topic of "AI Alignment" [117]. Key questions in this area persist around what value systems AI systems should align with, i.e., utilitarian, Kantian, or other value systems. Alignment can be seen as an aspect of AI governance [75]. These topics are ultimately concerned with questions about how humans can control the behaviors of AI systems. The idea of using data as source of leverage for collective action provides a potential answer to some of these questions. As noted, we might imagine a model of responsible AI in which data-related collective action is a key lever for regulating the behaviors of AI systems; when they go off the rails, members of the public engage in data leverage actions in a reactive manner.

2.4.2. Related Ideas in Political Economy

Our work contributes to a broad conversation about how data-related systems should be governed, and about how issues related to digital capitalism should be addressed more broadly. There has since been ongoing scholarly discussion about different ways to frame and study data's role in modern capitalism.

As noted above, our work has been heavily influenced by the "data as labor" perspective. Several competing ideas have emerged in response to this idea. In comparing several data metaphors, Viljoen considers data as labor to be a propertarian approach, and contrasts this with dignitarian approaches to data [403]. Sadowski analyzes data's role in the modern political economy as a distinct kind of capital that is similar to economic capital in terms of cycles of capital accumulation, and argues for scholars to focus on the process of data extraction, as opposed to over-emphasizing issues of privacy and security [334]. This work argues for regulating the data economy via capital controls, i.e., limits on data possession and building towards public infrastructure models of data management.

Here, I provide a brief statement on how I frame my work in light of ongoing debates and disputes around the political economy of data and AI: A goal of my work is to produce empirical results that can be useful from a variety of political economy perspectives regarding data, the "data-driven economy", and AI. In other words, regardless of which of the above (or other) perspectives ends up being more *predictive* in the long run or which perspective ends up becoming more *dominant* in scholarly and public discourses, I am hopeful that our empirical results and design work will be able to meaningfully contribute to mitigating the negative impacts of AI.

CHAPTER 3

Wikipedia's Value to Other Online Communities

3.1. Introduction

Over the past decade, Wikipedia has been a subject of tremendous interest in the computing literature[1]. Researchers have used Wikipedia to understand online collaboration dynamics (e.g. [201, 248, 440, 439]), to evaluate volunteer recruitment and retention strategies (e.g. [152, 153, 441, 442]), and even to train many state-of-the-art artificial intelligence systems [110, 118, 175, 177, 267, 433]. Indeed, Wikipedia has likely become one of the most important datasets and research environments in modern computing [59, 175, 267, 273].

However, a major limitation of the vast majority of the Wikipedia literature is that it considers Wikipedia in isolation, outside the context of its broader online ecosystem. The importance of this limitation was made quite salient in recent research that suggested that Wikipedia's relationships with other websites are tremendously significant [134, 265, 277, 421]. For instance, this work has shown that Wikipedia's relationships with other websites are important factors in the peer production process and, consequently, have impacts on key variables of interest such as content quality, reader demand, and contribution patterns [265, 421].

[1] This chapter was originally published as [406].

Perhaps more importantly, however, this recent research has additionally suggested that the reverse is also true: Wikipedia content appears to play a substantially more important role in the Internet ecosystem than anticipated, with other websites having critical dependencies on Wikipedia content. In particular, McMahon et al. [265] showed that the click-through rates of Google SERPs (search engine results pages) drop dramatically when Wikipedia links are removed, suggesting that Google is quite reliant on Wikipedia to satisfy user information needs. Among other implications, this means that the Wikipedia peer production processes studied in the social computing literature likely have a substantial – and largely unstudied – impact on other websites. McMahon et al.'s results also raised important questions related to the revenue being generated by for-profit institutions using volunteer-created Wikipedia content, especially in light of Wikipedia's limited donation income.

Recognizing the importance of understanding Wikipedia's relationships with its broader online ecosystem, the Wikimedia Foundation (the operator of Wikipedia) has called for more research on these relationships as a central part of its "New Research Agenda" for Wikipedia [385]. The goal of the research presented in this chapter is to help address this call. More specifically, we seek to extend McMahon et al.'s work on Wikipedia's relationship with Google into a different and important area of the online ecosystem: other large-scale online communities.

In line with guidance in the social computing community to increase robustness through the use of multi-community analyses [333, 395], we examine Wikipedia's relationship with two distinct large-scale online communities: Stack Overflow (SO) and Reddit.

Users post links to Wikipedia on both SO and Reddit, facilitating an important, and potentially bidirectional, relationship with Wikipedia.

Following the high-level structure of McMahon et al [265], this chapter asks two overarching research questions about Wikipedia's relationship with SO and Reddit*:*

RQ1: What value is Wikipedia providing to other large-scale online communities like Stack Overflow and Reddit? (i.e. Does Wikipedia content increase community engagement and/or company revenue?)

RQ2: What value do these large-scale online communities provide to Wikipedia? (i.e. Are they contributing page views? Editors?)

We additionally take an important step beyond McMahon et al. and investigate how the *quality* of Wikipedia articles affects the relationships examined in RQ1 and RQ2. In other words, we look at the association between the quality of articles on Wikipedia and the value that Wikipedia provides to external entities, and vice versa.

We address our RQs using a combined framework of associative and causal analyses that allows us to estimate Wikipedia's relationships with SO and Reddit under a range of conditions. For instance, at the upper bound of this range, our associative analyses allow us to ask, "How much value would be lost from SO and Reddit if posts containing Wikipedia links never appeared?" Similarly, to estimate a lower bound on the value Wikipedia could be providing to SO and Reddit, we use causal analysis to examine the counterfactual scenario, "What if the posts containing Wikipedia links remained unchanged content-wise, but instead had a link to a site other than Wikipedia?"

The results of our analyses indicate that Wikipedia creates a large amount of value for SO and Reddit, even in our lower-bound estimates. More specifically, we observe that posts containing Wikipedia links on both sites are exceptionally valuable posts, with engagement metrics like user-voted scores much higher than posts that do not contain Wikipedia links (often by a factor of at least 2x, and sometimes as much as 4x-5x). This results in an estimated increase in revenue on the order of $100K/year for both sites.

However, we find little evidence that posts with Wikipedia links provide direct value to the Wikipedia community. We were able to replicate work that showed that Wikipedia posts on the popular Reddit community "TIL" ("Today I Learned") were responsible for a large spike in viewership. However, our results suggest that this large effect does not generalize beyond the "TIL" community or beyond Reddit. Moreover, we see negligible increases in Wikipedia edits and editor signup despite the large volume of links posted on both sites. Through a smaller-scale qualitative analysis, we find evidence that suggests that the "paradox of re-use" [385] may be playing a role here: Wikipedia's permissive content licenses make it easy for Reddit and SO users to directly include the Wikipedia content in their posts, which could be mitigating the benefits of Wikipedia links in terms of traffic to Wikipedia and new Wikipedia edits/editors.

As we discuss below, these results have important implications for a number of constituencies. For companies that rely on Wikipedia content, our findings highlight the value (including both engagement and revenue) created by Wikipedia's free and volunteer-created content. For Wikipedia and its editors, RQ1's results further demonstrate the critical role the Wikipedia community plays in the broader online ecosystem outside of Wikipedia. However, our RQ2 results present more challenging implications for the Wikipedia community: these results highlight the need for further research on solutions to the paradox of re-use.

3.2. Background

Prior to presenting related work, we first provide high-level context about Reddit, SO and how Wikipedia content appears on these sites. For both Reddit and SO, there is a class of posts that is highly-dependent (if not entirely reliant) on Wikipedia content (i.e. the posts could not exist without Wikipedia). Additionally, users can "upvote" and "downvote" content on Reddit and SO, which gives it a "score" (# of upvotes – # of downvotes).

3.2.1. Reddit and Wikipedia

Reddit is a large-scale online community that allows users to share links to external content (e.g. news articles, images, videos), post original text content (e.g. questions, opinions), and discuss this content through comments. As of July 2017, Reddit was the ninth-most-visited website globally [9], had 300M monthly visitors, and had a $1.8B valuation [414].

On Reddit, links to Wikipedia often appear in the well-known "TIL" (Today I Learned) "subreddit", among others. The term "subreddit" refers to the sub-communities that together make up Reddit. Many posts in "TIL" are composed of only a Wikipedia link and a short summary or quote from the article. An example TIL post is shown in Fig. 3.1.

3.2.2. Stack Overflow and Wikipedia

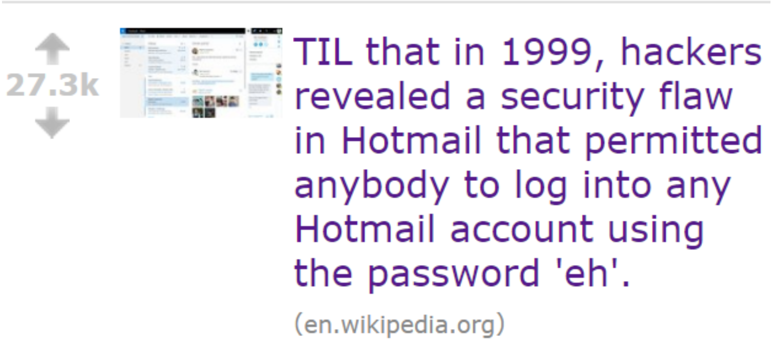
SO is a Q&A community for programmers and was the 56th-most-visited website in the world [9]. Stack Exchange, the company that owns SO, raised $40M of venture capital funding in 2015, bringing it to a total of $70M raised [145].

On SO, Wikipedia supports answers in the form of links and quoted text. Answers often use technical terms or acronyms and include a Wikipedia link in lieu of defining these terms. An example post that uses a Wikipedia link and quote to support an answer is shown in Fig. 3.2.

3.3. Related Work

3.3.1. Wikipedia and Its Internet Ecosystem

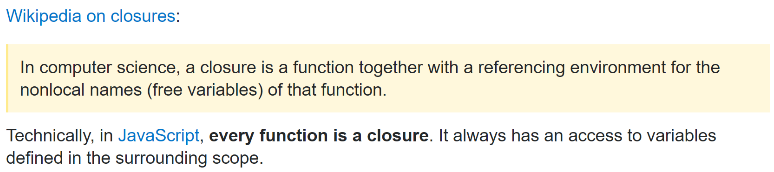
This research project was directly motivated by two recent developments related to Wikipedia and its broader Internet ecosystem. First, the Wikimedia Foundation called on Wikipedia researchers to focus on these relationships as part of a "New Research Agenda" for Wikipedia [385]. Second, recent work further substantiated this call with new evidence



TIL that in 1999, hackers revealed a security flaw in Hotmail that permitted anybody to log into any Hotmail account using the password 'eh'.

(en.wikipedia.org)

Figure 3.1. Example reddit post, with a Wikipedia link.



Wikipedia on closures:

> In computer science, a closure is a function together with a referencing environment for the nonlocal names (free variables) of that function.

Technically, in JavaScript, **every function is a closure**. It always has an access to variables defined in the surrounding scope.

Figure 3.2. Example StackOverflow answer, with a Wikipedia link.

showing how critical these relationships can be, both for Wikipedia and for the Web more generally [265].

With respect to the former, Dario Taraborelli – the Head of Research at Wikimedia – urged the large community of Wikipedia researchers to refocus their efforts on several new opportunities and challenges facing Wikipedia. One of those challenges is better understanding how Wikipedia (and Wikidata) content is used by external entities, the importance of this content to those entities, and the effect on Wikipedia (and Wikidata) of this external use. One concern of Taraborelli's is that the increasing re-use of Wikipedia content outside of Wikipedia will reduce traffic to Wikipedia. This would weaken Wikipedia's editor base over the long term, thus diminishing the re-use value of Wikipedia content. Taraborelli called this phenomenon the "paradox of re-use". We return to this concern below.

Recent work by McMahon et al. [265] provided clear evidence to substantiate Taraborelli's argument about the importance of understanding Wikipedia's relationships with its broader ecosystem. McMahon et al. found that click-through rates on Google SERPs drop dramatically when Wikipedia links are removed, and that Google is the mediator for the majority of observed Wikipedia traffic.

In addition to the direct implications for Google and Wikipedia, McMahon et al.'s study raised new questions for both the social computing community and the broader computing literature. For social computing, the amount of traffic to Wikipedia mediated by Google means that any changes to Google's presentation of Wikipedia content may have enormous effects on key variables of interest for those studying Wikipedia's peer production process, e.g., new editor recruitment and retention [57, 152, 153, 441]. For the broader computing literature (especially information retrieval), McMahon et al. found that the effect size of the presence of Wikipedia content was much larger than many algorithmic search improvements. More generally, this finding demonstrated how important social computing phenomena like peer production can be for addressing core problems in other areas of computing like addressing user information needs in web search.

In this chapter, we seek to expand on McMahon et al.'s work beyond search engines by considering Wikipedia's relationship with other large-scale online communities. We believed large-scale online communities would be an optimal domain to help follow-up McMahon et al.'s work for two reasons: (1) we anecdotally observed that this relationship may be particularly salient and (2) several research papers have provided early evidence that our anecdotal observations may generalize [134, 277]. We also expand McMahon et al.'s lens to consider the quality of the associated Wikipedia content, i.e. does higher quality content create more value for external websites? This question provides further insight into how Wikipedia's internal quality metrics (and therefore investment of community effort) map to external value.

The early evidence that motivated us to examine large-scale online communities primarily came from two papers: Moyer et al. [277] and Gómez et al. [134]. Moyer et al. examined the causal effect of Reddit posts on Wikipedia page views and found that sharing Wikipedia content on Reddit's TIL community increased page views to the corresponding Wikipedia articles. Below, we replicate this result, but find that it does not generalize to other subreddits. While studying the role of links on SO with regards to innovation diffusion in the software engineering space, Gómez et al. [134] found an intriguing peripheral result: Wikipedia links were the second most common external link on SO. While we also observed that Wikipedia links are very common on SO, we found little *direct* evidence that SO is contributing substantial value to Wikipedia.

3.3.2. UGC and the Physical World

Through our research, we also seek estimate the economic impact of Wikipedia content on external entities. This lens on our results was motivated by very recent findings from Hinnosaar et al. [173], which showed that improvements to Wikipedia article content about places in Spain directly increased tourism revenue in those places [173]. Analogous results have also been identified with types of user-generated content (UGC) other than Wikipedia (e.g. [53, 252]).

3.3.3. Factors Influencing Content Value

A key methodological challenge in our causal analyses below is identifying potential confounding factors for the substantial increases in value we see associated with Wikipedia-linked content on Reddit and SO. For example, users who post Wikipedia links may also write longer posts, and longer posts may be more popular. To search for potential confounds, we turned to the burgeoning literature on Reddit, SO, and other large-scale online communities. We present the literature used to guide model decisions here, and summarize in greater detail how these factors were operationalized in the Methods section.

3.3.3.1. Specific factors influencing value on Reddit. From platform-specific work on Reddit, we identified three major factors that can influence user votes (score), a value metric that we use in this chapter. Previous work identified content type [234] and title characteristics [172] as predictive of score on Reddit. Additionally, Gilbert showed that content posted multiple times (i.e. a particular image) that received a high score on later postings was often “missed” on early postings [128], implying that popularity is highly contextual, and not necessarily purely dependent on content itself.

3.3.3.2. Specific factors influencing value on SO. We also identified three factors that influence content value on SO. Anderson et al. found that contextual information, such as user reputation and time of arrival, were predictive of the long-term value of SO pages as measured by page views [12]. Ponzanelli et al. [306] showed that adding readability metrics to simple textual metrics (e.g. percent uppercase, number of characters) improved low-quality post detection. Finally, Calefato et al. [46] identified promptness, presence of a hyperlink, presence of a code snippet, and answer sentiment as key predictors of an answer’s likelihood to be selected as the best answer by the questioner.

Research on other question and answer (Q&A) sites also helped to inform our model design. On Math Overflow, a math-focused version of SO, both online reputation (points) and offline reputation (number of publications) were found to influence content popularity [386]. Harper et al. studied a variety of sites including Yahoo! Answers and identified that the number of hyperlinks in an answer was an indicator of answer quality [159]. We emphasize this potential confound in our analysis to understand the value that Wikipedia links add *beyond the value added by the presence of links overall.*

3.4. Methods

In this section, we first present the two aspects of our methodology that cut across our investigations of both research questions: (1) data collection and (2) handling of current events. We then describe our methodology specific to Study One (RQ1) and Study Two (RQ2).

3.4.1. Datasets

We downloaded database tables corresponding to every Reddit post in 2016 from pushshift.io as hosted on Google BigQuery (metadata, i.e. the latest score, last updated July 2017 [139]). We also leveraged the Reddit API to obtain user information [317]. We used BigQuery to download full database tables for all SO questions, users, and answers, starting July 31, 2008 and ending June 11, 2017 [139].

Following statistical best practice [319], we separated the analysis into two phases: an initial phase for developing the methods and a testing phase for generating results. As we describe below, many of our analyses required using multiple rate-limited APIs (e.g. for calculating features for our causal analyses), so we employed random sampling to minimize API endpoint load and to make query times tractable. For each platform and phase, we used a random sample composed of ˜1M posts from the entire dataset (1.10% of all Reddit posts and 4.46% of all SO posts) and, in the testing phase, an additional

˜40K posts from the subset of Wikipedia-linking posts (to ensure that we had adequate Wikipedia links to pages of each class of article quality).

3.4.2. Defining "Value"

The concept of "value" is central to both of our research questions, i.e. the value that Wikipedia is providing large-scale online communities and the value that these communities are contributing to Wikipedia. In this chapter, we seek to increase robustness and detect potentially interesting nuances by operationalizing the notion of value through multiple metrics rather than using just one (e.g. page views).

Specifically, we measure post value in RQ1 through four metrics. The first three metrics are user engagement statistics: (1) *Score*, equal to upvotes minus downvotes, (2) *Comments*, the number of comments a post receives, (3) *Page Views*, the number of views a post receives. To contextualize these metrics, we also calculate (4) *Revenue*, or the financial gain generated by Wikipedia posts. *Revenue* is calculated directly from the engagement statistics using publicly-available financial information (described in detail in Study 1 – Results). In the case of Reddit (which does not release page view data), it is important to note that score controls post visibility and correlates with page views [382].

With respect to RQ2, we assess the value that Wikipedia receives from external communities as contributions to the editing community and increased readership. Specifically, we measure this value with four metrics that capture changes in edits, editors, and viewership in a given week: (1) *Edit Count* is the number of times an article was edited, (2) *Editors Gained* is the number of new editors who edited an article, (3) *Editors Retained* is the number of new editors who made another edit in the future (we measured at one month and six months later, following past research on editor retention [57, 276]), and (4) *Article Page Views* is the number of views that each Wikipedia article received. To capture the effect of Reddit and SO on Wikipedia, we calculated the metrics for the week before and the week after each post containing a Wikipedia link.

3.4.3. Influence of Current Events

One potential confound of all our measurements is the impact of current events on our value metrics. For instance, if Reddit users happened to post Wikipedia links related to current events, then any subsequent increase in Wikipedia page views might be largely driven by current events and not by the Reddit post. We predicted, however, that very few posts with Wikipedia links on either platform are related to current events because SO is strictly for programming discussion and the Reddit TIL community does not allow current events posts.

To formally verify this assumption, we performed a qualitative coding exercise. Following standard practices [278], we used a small (10 posts per site) calibration procedure with two researchers, achieved a 90% agreement, and then one researcher classified an additional 100 posts per site. In this qualitative analysis, we identified that only 5% of Wikipedia-linked Reddit posts were related to current events, and no Wikipedia-influenced Stack Overflow posts were related to current events. This gave us confidence that our results were largely not driven by current events.

Estimate	Counterfactual	Calculation
UB	All content containing (1) WP links or (2) higher quality WP links was never generated	Mean values
MG	(1) WP links or (2) higher quality WP links were removed from all content (posts remain)	Propensity score stratified multivariate regressions
	(1) Non-WP links or (2) low quality WP links were removed from all content (posts remain)	Propensity score stratified multivariate regressions
LB	(1) WP links or (2) higher quality WP links were replaced with other external links	MG minus above estimate

Table 3.1. Summary of analyses used to obtain upper-bound (UB), middle-ground (MG), and lower-bound (LB) estimates of the effect of Wikipedia (WP) links. In the Counterfactual column, (1) refers to Presence of WP and (2) to Quality of WP.

3.5. Study One

Our first study targets RQ1, or "What value is Wikipedia providing to other large-scale online communities like Stack Overflow and Reddit?" Here, we present the methodology specific to this question and then present our results.

3.5.1. Controls and Treatments

In Study 1, our goal is to estimate the value that Wikipedia provides to SO and Reddit. We study value through two separate analyses. The various estimates are summarized in Table 3.1. First, we estimate the effects that the *presence* of Wikipedia links has on value. We then estimate the effects that Wikipedia article *quality* has on value. For each analysis, following standard practice, we label the change in value in the treatment group

– the group associated with Wikipedia content – the "treatment effect". We defined three groups:

Has Wikipedia Link: posts with at least one valid link to Wikipedia (as described below, this amounts to 0.13% of all Reddit posts and 1.28% of all SO posts).

Has Other Link: posts with at least one external link, but no links to Wikipedia (49.1% of all Reddit posts and 31.2% of all SO posts).

No External Link: posts with no external links at all (50.8% of all Reddit posts and 67.5% of all SO posts).

To estimate the effects of Wikipedia article quality on value to Reddit and SO (the second half of RQ1), we further subdivide the *Has Wikipedia Link* group into high-quality and low-quality groups. While there are many definitions of quality on Wikipedia [418], we rely on revision-specific predictions of quality along English Wikipedia's internal assessment scale[2] as produced by Wikimedia's ORES API [151]. Following Johnson et al. [194], we use the "C"-class assessment as a minimum assessment for a high-quality article because "C"-class articles are the first that are "useful to a casual reader" [92]. Specifically, we define our high-quality and low-quality groups as follows:

C-class or Better: All posts with *any* links to C-class or higher articles are in this group (79% of all Wikipedia-linked Reddit posts and 77% of all Wikipedia-linked SO posts).

Below C-class: All posts in which all Wikipedia links are to articles below C-class are in this group.

[2] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikipedia/Assessment`

3.5.2. Simulating a World without Wikipedia (Counterfactuals)

We cannot know how SO and Reddit would function in a world without Wikipedia. Therefore, we draw upon well-established causal inference methods (e.g. [183, 326]) that *estimate* the loss in value that would occur to these communities if posts with Wikipedia links were replaced with a *range of alternatives* (i.e. counterfactuals). In other words, we consider how SO and Reddit would be affected if they were not "treated" by Wikipedia content under a series of different assumptions. This approach provides a reasonable upper-bound, middle-ground, and lower-bound estimate of the value contributed by the Wikipedia community to SO and Reddit. In our consideration of the effect of Wikipedia article quality, we take a similar approach and estimate the loss in value if links to high-quality Wikipedia articles were replaced with a range of alternatives. We emphasize that causal analyses, the statistical methods we employ, can estimate causal effects with confounding effects reduced, but not eliminated (as in a randomized controlled trial) [183, 326].

Our upper-bound estimate of the value created by Wikipedia on SO and Reddit assumes that without Wikipedia, *all posts containing a Wikipedia link would not exist (i.e. would not have been generated).* For our analysis of quality, we make an analogous assumption: we calculate the value that would be lost if all posts containing high-quality articles did not exist. This upper bound is simply equivalent to the value of all "treated" posts.

The middle-ground estimate corresponds to the value that would be lost if the links to treated articles were *completely replaced with an identical post with no links.* In other words, this scenario assumes that the links to Wikipedia were removed, but the post still exists.

A lower-bound for value contributed by Wikipedia can be obtained by estimating the value that would be lost if the links to Wikipedia article were *replaced with links to alternative external domains*. In other words, the lower-bound scenario assumes that the exact same post (with alternative links) could be written without Wikipedia, and so the added value is solely from Wikipedia's reputation or other factors associated with Wikipedia (in comparison to other websites). In the quality analysis, the lower bound is the estimate of value that would be lost if links to high-quality articles were replaced with links to low-quality articles (without other changes).

3.5.3. Causal Analysis

While the upper-bound estimate is relatively easy to calculate using descriptive statistics, our middle-ground and lower-bound estimates require the estimation of formal counterfactuals, which calls for causal analysis techniques.

One of the first steps in casual analysis is to consider other potential causal factors, i.e. other reasons that Wikipedia-treated posts may have increased in value besides the Wikipedia treatment. To address this challenge, we turned to the literature on factors associated with content value, discussed above in Related Work and summarized in Table 3.2. This review broadly shows that value may come from four sources other than Wikipedia itself: user characteristics (e.g. users with high reputation), stylistic and structural characteristics (e.g. long posts, posts with code snippets, posts with punctuation),

Research Project	**Factors Addressed in Our Study**
Leavitt and Clark [234]	Content Type
Lakkaraju et al. [172]	Text length, sentence type, sentiment
Gilbert [128]	Day of week, month, hour, poster reputation
Anderson et al. [12]	Answer length, poster reputation, response time
Ponzenelli et al. [306]	Post length, percent uppercase, percent spaces, percent punctuation, post starts capitalized, Coleman-Liau index (readability)
Tausczik and Pennebaker [386]	User reputation
Calefato et al. [46]	User reputation, response time, presence of hyperlink, presence of code snippet, sentiment
Harper et al. [159]	Answer length, number of hyperlinks

Table 3.2. Summary of related work that identified factors that may affect post value in large-scale online communities.

post timing (e.g. posting on a certain day of the week), and the presence of any external link.

We operationalize these potential alternative causal factors through the calculation of features that capture them. Some of these features are numerical (e.g. post length) and others are dummy variables (e.g. whether the post includes code snippets). While most of these features were very straightforward to calculate (e.g. title length), readability and sentiment were more complex. With respect to readability, we use the Coleman-Liau index [61], which was used by Ponzanelli et al. [306]. For sentiment, we used the TextBlob library [251], which leverages sentiment analysis models (trained on customer review text) from the well-performing [322] library “Pattern” [368] to estimate objectivity and polarity. Additionally, we log-transform reputation and response time as they are otherwise highly-skewed variables. Further details about these implementations are available in our source code, which we have made publicly accessible for download[3].

In terms of our statistical approach to causal analyses, we performed four propensity-score-stratified regressions[4]. This method, originally described by Rosenbaum and Rubin [326] has been used in the context of HCI for many purposes and is an approach advocated for in the HCI community (e.g. [78, 89, 288, 108]). We controlled for the potential alternative causal factors – user, style, structure, and timing – and then estimate the effect of including a Wikipedia link, a non-Wikipedia link, a high-quality Wikipedia link, and a low-quality Wikipedia link.

By finding the difference between corresponding estimates (*Has Wiki Link* vs *Has Other Link* and *C-class or Better* vs *Below C-class)*, we produce a robust estimate of the value that Wikipedia uniquely provides, and the value that high-quality articles uniquely provide. This estimate is minimally affected by bias from hidden variables or incorrect model assumptions because this bias should affect the corresponding regressions equally. For instance, if there was a hidden variable at play (perhaps users have a general unwillingness to visit external domains), that would affect posts with Wikipedia links and posts with external links equally and be removed by taking the difference.

The propensity score of a given post is an estimated probability that the post was “treated” based on all available covariates. In our case, the covariates are the computed

[3] https://github.com/nickmvincent/ugc-val-est

[4] To check robustness, we also performed propensity score matching [16] and propensity score stratification with covariate overlap adjustment [72] and confirmed that these methods led to the same conclusions.

features summarized in Table 3.2 (except for the number of hyperlinks, which relates directly to whether there is a Wikipedia link). The actual propensity score for each post is calculated by logistic regression and represents how likely a given post is to include a Wikipedia link (or a good Wikipedia link) based only on its features. Posts that have features that are commonly found in Wikipedia-linked posts will therefore have high propensity scores.

Using the open-source *causalinference* [426] library, we *stratify* [16] our datasets by propensity score (subdivide into many subsets) in order to emulate a randomized blocked trial using observational data. Each stratum contains posts with a small range of propensity scores, so posts have similar features with each stratum, which reduces the standardized bias (described by Rosenbaum and Rubin [327]). To determine the number of strata and propensity scores for each stratum, we use a bin-selection algorithm described by Imbens and Rubin [183] and implemented in [426].

We performed a multivariate linear regression separately for each stratum, extracted the treatment coefficient, and then computed a weighted average of the treatment coefficients based on the number of treated items in each stratum. In other words, a coefficient from a stratum with more Wikipedia links than another stratum will be given greater weight in determining the estimated effect. The weighted average is the Average Treatment Effect on the Treated (ATT), which is the main result we present. This value represents how much value would be lost if all the treatment posts were replaced with control posts with nearly the same covariates.

The analysis of SO page views requires special handling because page views correspond to one question page, which could have many associated SO answers. This means we must take care to avoid overestimating page view effects. When computing mean page views for the upper-bound estimate, we only count each question once. In our causal analysis (i.e. middle-ground and lower-bound estimate), we make a conservative assumption that all page views are attributed to the top-scoring answer for a question, which means that only top answers are included in this analysis.

3.5.4. Study 1 – Results

In this section, we first present high-level descriptive statistics about the relationships between Wikipedia and SO and Reddit (e.g. # of Wikipedia-linked posts). We then present the results from our core analyses for RQ1.

3.5.4.1. Descriptive Results. Overall, we were surprised to find that Wikipedia links represent only 0.13% of posted content on Reddit. However, further examining this result, we found that Wikipedia links are substantially over-represented in high-value locations. For instance, Wikipedia is the third most-linked site (after YouTube and Imgur) on the ten most-popular subreddits. This relatively low-quantity/high-quality dynamic is one we see frequently in our formal analyses below.

On SO, Wikipedia links appear in 1.28% of posts, but this makes them the fourth-most-common type of external link (after github.com, msdn.microsoft.com, and the popular "code playground" jsfiddle.net). Notably, github and jsfiddle are used to share code, and msdn is Microsoft's code documentation library, meaning that Wikipedia is the most important conceptual reference for programming on SO.

3.5.4.2. Effects of Wikipedia-linked content on Post Value. The results from our full analysis to address RQ1 are presented in Table 3.3. Below, we unpack the main trends in Table 3.3, as well as discuss the implications of these results with respect to Wikipedia's aggregate impact on SO and Reddit.

Effects on Reddit posts: Table 3.3 shows that Wikipedia-linked posts on Reddit are exceptionally valuable. To a post's score, Wikipedia adds between 108 points (ATT, **delta** in Table 3.3) and 151 points (Mean Values: Has WP Link, with a middle-ground estimate of 141 points (ATT: Has WP Link vs No External Link). Relative to the average post's score of 30 points, this is a 4x-5x increase. Aggregating these findings across all 120K Wikipedia-linked posts from 2016, this means that Wikipedia is responsible for an increase in user-voted score of between 13.1M and 18.4M points in 2016 (up to 0.7% of all points on the site).

Additionally, Table 3.3 also shows that Wikipedia adds 11-19 comments per post. This means Wikipedia-linked posts generate twice as much discussion as average posts. In total, this amounted to between 1.3M and 2.3M comments in 2016.

Effects on Stack Overflow posts: For SO, Table 3.3 displays a similar trend to what we saw with Reddit. For instance, Wikipedia-linked content on SO adds 2.9-6.5 points per post, with a middle ground of about 3.4 points. This means that Wikipedia-linked answers are roughly twice as valuable as other answers and, across the 280K Wikipedia-linked answers on SO, increased total score between 0.8M and 1.7M points. This score increase is accompanied by a page view increase of 954-3473 per question, with a middle ground of 1337. Estimating based on the 12K questions with Wikipedia-linked answers in our page view analysis, Wikipedia may have added 64M to 234M views (which we use for revenue estimation). However, we see no evidence of an effect on SO comments.

		Mean Values			**ATT [99% CI]**			
	Variable	Has WP Link (UB)	No WP Link	Δ	Has WP Link vs No External Link (MG)	Has Other Link vs No External Link	Δ (LB)	Ratio (LB-UB)
3a) Presence of WP								
R	Score	151	31	120	141 [119.4-163.1]*	34 [31.0-36.4]*	108 [85.6-130.0]*	4.5-4.9
	Comments	19	8	11	7 [2.4-11.0]*	-4.0 [-5.2- -2.7]*	11 [6.2-15.1]*	2.3-2.3
SO	Score	6.5	2.5	4.0	3.4 [0.9-6.0]*	0.5 [0.3-0.7]*	2.9 [0.4-5.4]*	2.1-2.6
	Comments	1.7	1.4	0.3	0.0 [-0.04-0.1]	0.0 [-0.1-0.0]	0.0 [-0.02-0.1]	1-1.2
	Page views	8535	5062	3473†	1337 [727-1947]*	383 [252-515]*	954 [330-1578]*	1.2-1.7
3b) Quality of WP								
	Variable	C-class or Better (UB)	Below C-class	Δ	C-class or Better vs No External Link (MG)	Below C-class vs No External Link	Δ (LB)	Ratio (LB-UB)
R	Score	163	107	56	151 [126.5-175.3]*	90 [63.9-115.4]*	61 [25.7-96.7]*	1.5-1.5
	Comments	21	13	8	7 [2.6-12.2]*	-1 [-5.7-4.0]	8 [1.5-15.1]*	1.6-1.6
SO	Score	6.6	6.2	0.4	4.3 [0.3-8.4]*	2.5 [0.8-4.2]*	1.9 [-2.5-6.2]	1-1.1
	Comments	1.7	1.7	0.0	0 [-0.1-0.1]	0 [-0.1-0.1]	0 [-0.1-0.1]	1-1
	Page views	8834	7955	880†	1492 [762-2222]*	1309 [141-2478]*	183 [-1195-1560]	1-1.1

Table 3.3. Shows estimated effects of Wikipedia (WP) links on Reddit (R) and Stack Overflow (SO). Includes upper-bound (UB), middle-ground (MG) and lower-bound (LB) estimates for both the Presence of WP analysis (3a) and Quality of WP analysis (3b); * indicates p ¡ 0.01; † indicates upper-bound estimate for SO page view analysis. "Ratio (LB-UB)" columns shows how much more valuable a treated post is compared to average - for the LB and UB estimates.

Overall, the presented estimates show that even assuming all authors could continue writing the same posts, except with non-Wikipedia alternative links, Wikipedia still adds significant value (i.e. through its brand or other factors).

3.5.4.3. Effect of Wikipedia Article Quality. Compared to posts that only have links to *Below C-class* Wikipedia articles, we find mixed evidence that links to *C-class or Better* articles contribute to value. Relative to *Below-C-class articles*, *C-class or Better* articles add 61-163 points on Reddit (1.5-1.5x increase). Similarly, *C-class or Better* articles also add 8-21 comments on Reddit (1.6-1.6x increase). However, we observe no effect on SO score, comments, or page views.

The above results suggest that article quality for the purposes of the SO community may mean something different than article quality for the Wikipedia community (and Reddit). For instance, SO members may not differentiate between the value of short or stub-like articles and longer, high-quality articles, as long as those articles contain a specific piece of desired technical information.

3.5.4.4. Back-of-the-Napkin Revenue Estimation. To better understand how the results of Study 1 translate into real-world revenue for SO and Reddit, we use the following "back-of-the-napkin" revenue estimations, incorporating as much actual public financial information as possible and making conservative assumptions (e.g. assuming all SO ads were sold for the lowest price listed). Both platforms sell ads at a fixed cost per thousand impressions and therefore revenue scales linearly with page views.

Estimating Reddit's 2016 ad revenue is relatively simple, as Reddit only makes money from ads and Reddit Gold (a subscription service). Using Reddit's $20M revenue projection [219] and an approximation of $1M revenue from "Reddit Gold", we presume a total ad revenue of \$19M. To obtain a rough figure for Stack Overflow's total revenue from 2008-2016, we use the following equation:

$$rev_{SO} = (views - votes) * \frac{ads}{views} * \frac{cost}{ad} * adblock$$

The $(views - votes)$ accounts for the fact that high reputation users see reduced ads [375] (we conservatively assume that every vote was made by a high reputation user, and therefore the corresponding view should not be included in ad revenue). We also multiply by an "adblock coefficient" of 0.75 to account for the 25% of desktop users who block advertisements [195]. We conservatively assumed all ads cost \$0.00466 [374], the lowest price listed in August 2017 and that users see only two impressions per page (users who scroll down actually see three).

Finally, because Reddit page view data is not available, we estimate page views using score, based on research showing that the score of a Reddit post correlates with views [382].

Overall, under these conservative assumptions, we estimate that in 2016, Wikipedia was responsible for between \$114,500 and \$124,200 of Reddit's revenue, and from 2008 to 2017, Wikipedia annually was responsible for between \$49,700 and \$180,900 of SO's annual revenue.

In total, considering only the content analyzed from two communities and the limited time periods we studied (nine years of SO activity and only one year of Reddit activity), Wikipedia may have been (conservatively) responsible for about \$1.7 million in revenue, entirely from volunteer work of the community.

	SO			Reddit		
	Before	After	Increase (99% CI)	Before	After	Increase (99% CI)
Edit count	1.440	1.426	-0.014 [-0.054 – 0.027]	4.584	4.990	0.405 [0.1 – 0.7]*
Editors gained	0.050	0.047	-0.003 [-0.009 – 0.004]	0.102	0.098	-0.004 [-0.02 – 0.01]
Editors retained (1 month later)	0.011	0.011	0.000 [-0.003 – 0.003]	0.016	0.017	0.002 [-0.004 – 0.007]
Article page views (daily)	10472	10697	225 [-243 – 694]	94323	96536	2213 [-2082 – 6508]
Article page views from TIL posts (daily)				47737	54026	6289 [2228 – 10350]*

Table 3.4. Summary of the effects of SO and Reddit on Wikipedia; * $p < 0.01$.

3.6. Study 2

In this section, we present a study that addresses RQ2, or "What value do Reddit and SO provide to Wikipedia?" For the before-and-after windows for each post (one week each), we calculated the *Edit Count*, *Editors Gained*, *Editors Retained*, and *Article Page Views* (reported daily by the Wikipedia page view API), as discussed above.

3.6.1. Study 2 Results

Table 3.4 shows the full results of our quantitative before-and-after investigation into RQ2.

The clearest trend is the near-absence of any significant results for both edit behavior and page views. With regard to edit behavior, on both platforms, we did not observe a significant increase in editors gained or editors retained during the week after a Wikipedia-linking post appeared. While we did observe an effect for the number of edits from Reddit posts, this effect is quite small (about 1 edit per 2 posts). If we consider all the Wikipedia-linking Reddit posts from 2016, this amounts to about 0.002 edits per second. Compared to Wikipedia's 10 edits per second [94], this effect is largely negligible. We did observe one minor significant effect that warrants discussion: Low-quality articles ("Stub" and "Start") were edited enough to achieve a statistically significant increase in edits relative to high-quality articles, even on SO where no other effects were observed. The effect was only 0.04 edits per post, but the presence of this small effect suggests that lower-quality articles received proportionally more contribution from Reddit and SO. We discuss the implications below.

During this analysis, we also analyzed the effect of Reddit posts exclusively from the TIL community on Wikipedia page views, a replication of work by Moyer et al. that used 2012 Reddit data [277]. In this specific case, we found the increase in page views corresponded with previous results. However, when including posts from outside TIL and when looking at SO, we found that the increase in page views was not statistically significant, indicating that while this result replicates, it may not generalize (we note that we used the Wikipedia pageview API, which returns daily statistics, whereas Moyer et al. used hourly page views). In other words, we saw no evidence that an arbitrary link to a Wikipedia page on Reddit and SO significantly increases traffic to the Wikipedia page in the week following.

3.7. Discussion

3.7.1. Paradox of Re-use

As noted above, a concern that has been raised within the Wikipedia community is the "paradox of re-use" – i.e. that the quality of Wikipedia's content and Wikipedia's editor base could decline over the long term through the external re-use of Wikipedia content (and that the re-using parties would then suffer as well) [385]. This concern leads to the question of whether Wikipedia needs to adapt its permissive licensing to survive, especially as its content is increasingly appearing on platforms more than a "click away" (e.g. voice assistants). Given that we observed that Wikipedia-linking posts were receiving a great deal of attention and engagement on SO and Reddit, we were surprised that so little of this attention and engagement was returned to Wikipedia in the form of page views, edits, and editors. We hypothesized that the paradox of re-use may be a factor.

To understand exactly how SO and Reddit use content from Wikipedia, we performed a small-scale qualitative coding exercise. Using the same approach to qualitative coding as the current events analysis, we classified whether each post matched the following definition of direct re-use: "Has text (quoted or summarized) from article." Two authors conducted a calibration coding and achieved reasonable inter-rater agreement (Reddit: Cohen's kappa = 0.74, 90% agreement; SO: Cohen's kappa = 0.62; 90% agreement), and then one author coded 100 posts with Wikipedia links per site.

Our coding analysis revealed that 79% of Reddit posts had quoted or summarized text from the linked Wikipedia article, whereas this was true of 33% of SO posts. This result shows that many posts can be characterized by "direct re-use".

These early results help to expand our understanding of the nature of the "paradox of re-use" from McMahon et al. [265]. McMahon et al. found causal evidence for the paradox of re-use on Google – i.e. Knowledge Card assets in Google search results were capturing views that would have otherwise gone to Wikipedia. This indicates a "strong" version of the paradox of re-use, in which Google was capturing traffic that would have gone to Wikipedia. Our (associative) results suggest that a "weak" form of the paradox may be occurring in Reddit and SO: the Wikipedia links on Reddit and SO posts might have generated new traffic for Wikipedia (rather than capturing existing traffic), but the re-use of content directly in Reddit and SO may be mitigating this effect. However, it is critical to note that our results merely allow this as a possibility: future work should test this hypothesis directly, ideally through experimental approaches.

3.7.2. Broadening Peer Production Research

At a high level, the results of Study 1 demonstrate that there are important relationships between different peer-production communities (with Reddit and SO themselves being peer production communities). We see that when content creators on Reddit and SO leverage existing peer-produced content (i.e. Wikipedia) to create new content in their communities, some of the "value" transfers and the magnitude of that value is reflective of the re-used content's quality (in this case, the quality of Wikipedia articles). In other words, Reddit and SO users can take advantage of the work already performed by the Wikipedia community to post higher value content in their communities with less effort.

These relationships between peer production communities have important implications for research. Studies of some important variables such as content quality may want to consider external dependencies and implications, e.g. Wikipedia article quality matters well outside of Wikipedia, adding importance to the body of work that studies Wikipedia quality and how to improve it [151, 57, 152, 419, 418, 441].

3.7.3. Coverage Biases and External Entities

The work on Wikipedia content biases is another key research area in the Wikipedia literature that our results suggest should be examined with a broader lens. Wikipedia (and other UGC communities) have been shown to over-focus on some topics and under-focus on others, particularly along the lines of gender (e.g. [171, 269, 413]), geography (e.g. [164, 194]), and language (e.g. [165]). Importantly, our results suggest that these biases may be having ripple effects around the web, including on SO and Reddit. For instance, the findings from Study 1 indicate that these biases mean that the TIL community on Reddit is working from a dataset that disadvantages women, certain geographies (e.g. rural areas [194]), and topics outside of the cultural sphere of most English speakers. Any posts about these Wikipedia-disadvantaged topics that do appear on TIL are less likely to have a Wikipedia link (or high-quality link) available, and thus cannot get the benefits of that link.

There is a potential silver lining regarding coverage biases: while the number of edits contributed to Wikipedia from Reddit that we observed in Study 2 was quite small in the context of Wikipedia as a whole, it may be that this effect would be meaningful if applied to specific under-covered content areas. For instance, if more Wikipedia links were posted on a subreddit for rural issues, the throughput of edits that is quite small for all of Wikipedia could make non-trivial improvements to coverage of some rural topics, for which a small number of edits is a large relative increase in edits.

One could also imagine the Wikipedia community (e.g. WikiProject leaders) working with the moderators of subreddits in under-covered areas on intentional "quality improvement projects" [418]. This might involve rallying these subreddits to encourage edits to relevant Wikipedia articles and to encourage members to become long-term editors. It would be interesting to compare the results of such an effort to our findings, which would provide an excellent baseline to determine if the effort was effective.

3.7.4. Implications for Reddit and Stack Overflow

McMahon et al.'s work [265] indicated that financially supporting Wikipedia may be highly incentivized for search companies, and our work suggests that the same might be true for Reddit and SO (although to a lesser degree). In other words, our results suggest that by donating to Wikipedia, entities like SO and Reddit can not only garner goodwill, but also could feasibly see a return on investment. In general, donations to support staffing and infrastructure are an understudied aspect of the peer production process, especially relative to their importance. Our results indicate that one direction of research in this space that might be fruitful is identifying formal value propositions for external consumers of peer-produced content.

Our results also suggest that Reddit and SO design interventions could help increase the mutual value of Wikipedia-SO and Wikipedia-Reddit relationships. For instance, Reddit and SO could implement a feature that detects when linked peer produced content is low quality or under-covered and directly encourage users to contribute. By improving

Wikipedia content, this would in turn add value to the Reddit and SO and could even be gamified, e.g. giving extra Reddit/SO "karma" when users contribute. Similarly, to facilitate higher-quality posts on Reddit and SO, topic modeling could also be used to suggest related Wikipedia articles to improve post quality.

3.7.5. Future Work and limitations

An important direct extension of this work would be to attempt to replicate our causal analysis on observational data with a randomized experiment. However, experiment-driven posting behavior could be seen as deviant by the SO and Reddit communities, so ethical and ecological validity challenges should be carefully considered.

Secondly, while the "back of the napkin" estimations of revenue above provide a minor contribution to understanding the economic implications of volunteer peer-produced content, much more work is needed in this area. This chapter, along with prior research [265], suggests that volunteers' content creation efforts generate important economic value well outside the corresponding communities. Understanding this value could provide intrinsic motivation for volunteers and incentivize donations from beneficiaries. Some economists even believe that better understanding this value could produce improved national GDP estimates [40]. However, moving forward in this space will be difficult because many of the corporate beneficiaries of volunteer-created content do not release the necessary information for rigorous estimates. As such, new methods that go beyond our "back of the napkin" approach will likely be necessary, such as using targeted ad buys to estimate actual revenue, advancing techniques similar to Cabañas et al. [136], and attempting to measure the value of volunteer-created content to profitable machine learning systems.

Finally, given that most of our results were not community-specific but rather generalized across Reddit and SO, similar relationships likely exist between Wikipedia and other discussion-based communities (e.g. akin to Reddit: Hackernews, Facebook groups; akin to SO: Quora, Yahoo! Answers). Future work should seek to examine these relationships, as well as those between these communities and peer-produced data repositories like OpenStreetMap.

3.8. Conclusion

In this chapter, we presented results that identify and quantify relationships between Wikipedia and the large-scale communities Reddit and Stack Overflow. In general, we observe a one-way relationship in which Wikipedia-influenced content adds value to the external communities, but no evidence of substantial contributions in the reverse direction is observed. This research highlights the value of examining online communities using a broad lens, as cross-community relationships can have large effects.

CHAPTER 4

User-Generated Content's Value to Search Engines

4.1. Introduction

Search engines are immensely popular and enormously valuable intelligent technologies.[1]. Over 92% of American adults use web search [312] and Google.com is the most-visited website in the entire world [9]. Moreover, Google makes over $20 billion per year from search advertising revenue [392] and Google's market capitalization is one of the highest in the world [113].

However, very recent work has suggested that search engines, despite their power and profitability, may be surprisingly dependent on a resource that is both volunteer-created and freely available: user-generated content (UGC), and specifically Wikipedia. In particular, McMahon et al. [265] found that search engine result page (SERP) click-through rates dropped drastically from 26% to 14% when Wikipedia results were removed from SERPs. This drop in click-through rate – a critical search engine evaluation metric – is enough to easily wipe out gains made by even major improvements to search engine algorithms, for instance the introduction of deep learning [59].

While McMahon et al. showed that Google search users have a strong preference for Wikipedia pages when they are surfaced, McMahon et al.'s study design did not allow them to ask an equally important question: *How often do search engines surface Wikipedia*

[1] This chapter was originally published as [408].

links – let alone links to other types of user-generated content – in the first place? In other words, it is unclear how often users are able to act on their strong Wikipedia preference. This means the full real-world impact of McMahon et al.'s findings are also unclear.

In this chapter, we perform a rigorous audit of Google's search engine to understand the extent to which Google surfaces links to English Wikipedia and other UGC. Specifically, we examined results across six categories of high-value queries selected for popularity, potential for advertising revenue generation, and potential to influence users' lives. Using software we developed, we also robustly address potential confounding effects from geographic personalization, known to be by far the major source of variation in search results [155, 204, 430].

Our results both complement and strengthen the findings of McMahon et al. We find that across all six categories of important queries, Google is highly reliant on Wikipedia to perform its core mission of satisfying user information needs. For some categories of queries (*trending* queries and *controversial* queries), Wikipedia articles appear in over 80% of (first) results pages and appear in the particularly important "top three links" over 50% of the time. Even for types of important queries for which Wikipedia appears less often (e.g. some high-revenue queries), Wikipedia still appears in over 20% of results pages. More generally, Wikipedia was by far the single most prevalent source of links across all query types. In other words, in our study, Google returned links to English Wikipedia far more often than it did for any other website in the world.

We do also find, however, that the value of UGC to Google more or less stops with Wikipedia. While Google frequently surfaces content from platforms that the literature commonly considers to be UGC (e.g. social media platforms), our findings showed that most of this content comes from professional sources (e.g. corporations, journalists) rather than individual users. For instance, although tweets frequently appeared in SERPs in our study, these tweets were almost always from corporate accounts or official political accounts like those of U.S. senators.

Our results have important implications for a number of specific constituencies. Most notably, our Wikipedia findings raise the stakes of the large social computing and computational social science literatures on Wikipedia. Our results suggest that the findings in these literatures – e.g. those about gender biases [171, 413] and geographic [194] content biases – not only have an impact within the Wikipedia web site, but also affect popular search engines.

More generally, our Wikipedia findings contribute to a growing discussion [163, 228, 265, 309] about the relationships between end users and intelligent technologies like search engines. Our results – along with those of McMahon et al. and others – highlight that end users are not just silent consumers of powerful intelligent technologies. Rather, through the content that they create, end users play an absolutely critical role in helping these technologies accomplish their core goals. This critical role is the basis for the nascent-but-burgeoning debate about the current distribution of financial rewards from intelligent technologies, a debate to which our results provide valuable early empirical information.

4.2. Related Work

4.2.1. Web Search and User-Generated Content

The work in this chapter was directly inspired by McMahon et al.'s work showing that Wikipedia is critically important to the success of web search [265]. As is discussed above, our research fills in a key piece of the puzzle outlined by McMahon et al. by examining how often Wikipedia results appear on Google SERPs. In the previous chapter, focusing on the relationships between Stack Overflow, Reddit, and Wikipedia, we al. found a similar – though smaller – effect for the amount of value Wikipedia adds to these external sites. It is important to note that McMahon et al.'s and our research was itself directly motivated by a call from the Wikimedia Foundation (the operator of Wikipedia) for research into the relationships between Wikipedia and its broader ecosystem, including search engines [385].

Researchers from tourism studies and medicine have also investigated the important role UGC plays in serving domain-specific search queries. In 2010, Xiang and Gretzel conducted an audit study using tourism-specific queries and found that social media platforms like TripAdvisor and Yelp made up about 11% of all Google results they collected [429]. Haiyan performed a very similar study using the Baidu search engine and Chinese tourism queries and found social media comprised almost 50% of results [150]. Laurent and Vickers studied the role of Wikipedia in serving health queries and found that Wikipedia was common: in their study Wikipedia appeared in 71-85% of top-ten results across multiple health-related query sets [231]. Interestingly, this statistic is quite a bit higher than we observe for medical queries, a point to which we return below.

Outside of the academic literature, the search engine optimization (SEO) industry has leveraged the role of UGC in search ranking algorithms [203, 434]. In fact, SEO firms are known to manipulate UGC (e.g. editing Wikipedia pages) to attempt to boost the rank of webpages of their clients [361].

4.2.2. Search Engine Personalization Auditing

Many of our methodological choices below draw heavily from the findings and best practices of work that has sought to audit the degree of personalization in web search. It has been consistently shown in this literature that location is an important driver of personalization in search results. For instance, examining search personalization with respect to a multitude of factors such as gender, age, education, and browser choice, Hannak et al. found geographic location to be the main source of personalization (outside logging into a personal account; see below) [155]. Similarly, in a user-focused study of search personalization, Xing et al. [430] found evidence of substantial personalization due to location. When doing more focused analyses on the role of geographic location in search personalization, Kliman-Silver and colleagues [204] found that the magnitude of geographic personalization varied with query type. This is one reason why we examine six types of queries in this work rather than focusing on a single type.

In our Methods section below, we discuss in significantly more detail how the personalization auditing literature inspired and informed our methods.

4.3. Methods

In this section, we describe the five key aspects of our methodological approach: (1) our software framework, (2) how we selected queries, (3) how we analyzed SERPs, (4) how we identified UGC, and (5) how we handled the potential confound of geographic personalization.

4.3.1. Software Framework

The high-level methodological challenge faced in this research was to collect Google SERPs for many queries from a variety of simulated locations. To address this challenge, we built a software package that modifies and extends the open-source, Selenium-based SerpScrap [344] library, which automates the desktop version of Chrome web browser. In this chapter, we focus on desktop search and leave to future work extending our analyses to incorporate the nuances of mobile search (see Discussion below).[2]. We note that utilizing the software will require moderate updates due to the constantly changing structure of Google SERPs.

Our software iterates through queries (selected as described below) and locations (also selected as described below) in quick succession but pauses for a full minute between each query to avoid causing undue load. While this approach is inspired by past work by Kliman-Silver and colleagues [204], it also differs from this work in one key way: Kliman-Silver and colleagues took samples for a single query at one time instant, whereas we issue queries sequentially. We believe our sequential approach, which reduces the resources required to collect data, is appropriate because a conclusion of Kliman-Silver et al.'s work was that personalization is consistent over time. Indeed, we were able to verify that our sequential approach led to similar levels of personalization as Kliman-Silver et al.'s parallel approach: our results replicate the general levels of personalization found in their work (our SERPs had an average Jaccard Index of 0.86 and an average edit distance of 1.9, within the range of values observed by Kliman-Silver et al.).

[2] For software, see: https://github.com/nickmvincent/SerpScrap for data collection code; https://github.com/nickmvincent/you-geo-see for analysis code and data

To simulate queries from different locations, we also take inspiration from Kliman-Silver et al. [204]. Specifically, following their approach, we inject Javascript that overrides the *geolocation.getCurrentPosition*() function to return a latitude and longitude of our choice. We then automatically click the "update location" button and refresh the SERP. We verified that this approach worked as it did in Kliman-Silver et al. by leveraging the fact that Google reveals the perceived location of each query at the bottom of each SERP. For instance, a query from Chicago will have the following text at the bottom of the resulting SERP: "Chicago, IL. Reported by this computer".

Our software only captures the first SERP for each query. We focused on the first SERP as research has shown that users only very rarely look at results pages beyond the first one [398]. For a similar reason, in our analyses, we provide an additional level of focus on the top three results on the first SERP. The first spot may receive up to 30% of all traffic, with the top three spots receiving 60% of all traffic [185].

4.3.2. Selecting Queries

In the search literature – and certainly in the search auditing literature – deciding on a set of queries for an analysis is well-known to be challenging [155, 204, 265, 295]. Aside from researchers operating within search companies (and sometimes even for these researchers), it is impossible to obtain a set of queries that is guaranteed to be representative. As a result, researchers must use heuristic strategies to generate an imperfect query sample that still can provide insight for their research questions. A typical approach involves first choosing a limited set of query types that have significant real-world implications, e.g. queries related to medical issues, commerce, or politics [103, 156, 204, 231, 295, 429].

By focusing on a single or small set of query types, researchers can then use creative approaches to generate viable specific queries within these type(s), e.g. manually adapting a query dataset from a published paper to a new geographic context [150, 265], using externally available data to generate potential queries [155, 231], or manually generating reasonable queries [204, 429].

In our research, we sought to adopt a diversified version of the above approach by including six separate query types instead of just one or two. More specifically, we focused on query types with real-world importance along three dimensions: (1) how often a query is made (popular queries), (2) the revenue Google makes from selling ads on the query (high-revenue queries), and (3) the degree to which the results of a query could impact users' lives (influential queries). In keeping with approaches in the search literature, we used a combination of external resources such as Google Trends and Google AdWords and data from existing research [155, 204] to select queries in a systematic way. For each of the three dimensions above, we developed two separate categories of queries, leading to six total query categories. Each query category contains between 10-20 queries, a number selected to be practical with respect to the rate limit we imposed to avoid excessive querying.

By considering three different dimensions of importance and using two different categories of queries for each dimension, our intention was to gain a broad and robust view of the role that UGC plays in Google SERPs, and one that is not unduly contingent on query-specific idiosyncrasies. Below, we detail our categories and their constituent queries. For replication and extension purposes, a full list of our queries and their assigned categories can be found in our software repository linked above.

Popular Queries: We considered two categories of popular queries: *trending* queries and *most-popular* queries. To develop our *trending* category of queries we turned to Google Trends. Google Trends is a public website that Google maintains to share data about patterns in usage of Google's search engine. We used Google Trends' "trending searches" feature to obtain queries characterized by a large baseline number of searches and a spike in searches (typically 1,000,000+ searches) [140]. Specifically, our *trending* category consists of each daily top trending query from Nov. 28 to Dec. 7, 2017 (10 queries).

With respect to our *most-popular* category, Google Trends does not directly provide a list of the most popular queries on Google's site overall, but it does do so for specific query topics. In other words, we can know what queries are popular within a topic, but we do not know the global popularity of the topic. As such, to develop our *most-popular* query category, we collected the top three queries by U.S. query volume across a set of Google Trends-defined query topics which were commercial or political in nature: Auto Companies, Fast Food Restaurants, Financial Companies, Governmental Bodies, Politicians, and Retail Companies. We discuss the potential limitations of manually selecting categories from Google Trends' offerings in our Limitations section.

High-Revenue Queries: Google sells many of its ads – and generates much of its revenue [353] – by allowing entities to bid on SERP ad placements on a query-by-query basis (using a system called Google AdWords). While Google does not provide high-level data about which are the most expensive queries, the SEO industry has published informal studies on this topic. According to one such study, *insurance*-related queries and *loan*-related queries are two of the most expensive categories of queries [427] and, as such, we selected these two categories to represent high-revenue queries. To populate these categories with actual queries, we used Google Trends' "Explore" feature to obtain the top ten queries for "insurance" and for "loans" (in the U.S., from all of 2017). We used Google AdWords' Keyword Planner to verify that the bids for these query categories were indeed very high; we observed a top cost-per-click of $514 for the most expensive query in the *insurance* category and $259 for the most expensive query in the *loans* category in December 2017.

Influential Queries: Query popularity and query revenue do not necessarily correlate strongly with the influence of a SERP on people's lives. Some types of queries – e.g. queries related to a family member's serious illness or queries related to informing one's political views – can have an out-sized impact [103, 371]. To gain a sense of UGC's influence in Google search results for particularly influential queries, we included two additional categories of queries that have been the subject of prior research in the search literature because of their influential nature: queries on medical topics and controversial topics. For our medical category, we use a subset of queries from Soldaini et al.'s study of health searches on the Bing search engine [371]. This set consists of 50 queries sampled from Bing's top 500 medical queries; we used the first 20 queries. For our controversial query category, we were unable to re-use queries from Epstein's experiment or Kliman-Silver's audit study [204] because the queries were related to current events (e.g. topics included the UK Prime Minister Election, Barack Obama's US presidency). As such, to systematically generate a diverse list of up-to-date search queries, we used the top ten topics from procon.org, a non-profit organization that hosts information about controversial issues.

4.3.3. Understanding SERPs

Modern Google SERPs consist of substantially more than the traditional "ten blue links" [52] that formally comprised the canonical search results page. Current SERPs contain multiple columns of content, and items like carousels (which have multiple links per row), answer boxes, and more. To understand the prominence of UGC on Google SERPs, it was important that we account for all this complexity.

As such, in addition to standard "blue links", our analytical framework also explicitly considers the following Google SERP element types, which are also visualized in the diagram in Fig. 4.1:

- *NewsCarousel*: A row of three cards that each link to a news story.

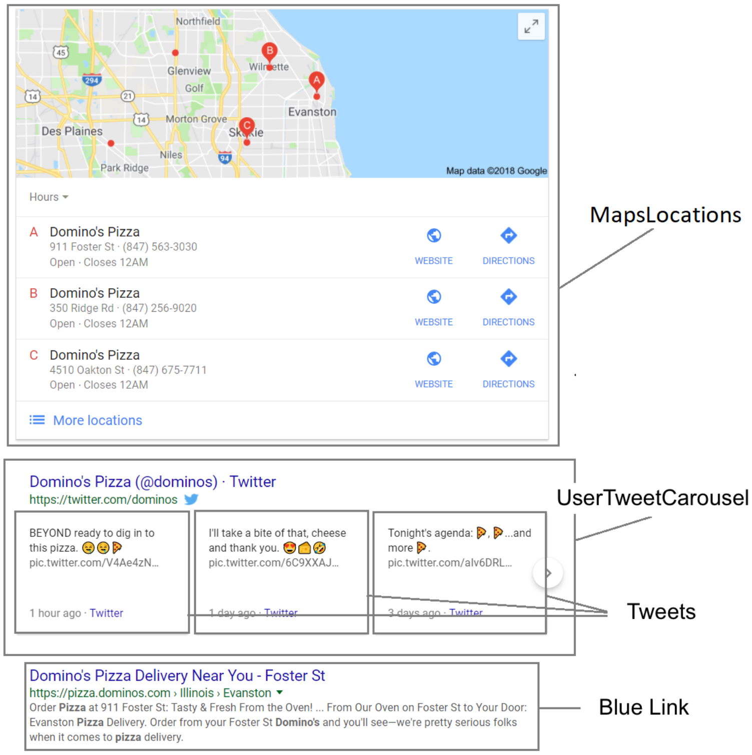
- *TweetCarousel*: A row of three cards with one tweet each. Google obtains the tweets either from Twitter's search (a *SearchTweetCarousel*) or a single user (a *UserTweetCarousel*).

- *MapsBox*: A box with Google Maps embedded that includes up to three locations. We mainly observed *LocationsMapsBox* elements, which have entries corresponding to multiple locations of a single entity (i.e. business).

- *AnswerBox*: A box that includes a link to a website and a snippet of text meant to answer a question; includes variants such as *PeopleAlsoAskAnswerBox* elements.

It is important to note that for most of our analyses below, we do not consider links that occur in the "*KnowledgePanel*", another SERP element that, for desktop web browsers, appears on the right-hand side of SERPs. Although KnowledgePanels include Wikipedia and social media links, we did not consider them in our core analyses because content in KnowledgePanels cannot be easily assigned a rank (as the panel essentially exists separately from the ranked search results). However, we did perform a small analysis of how our results might change if we did consider KnowledgePanels, and we discuss that analysis below. With respect to implementation, to operationalize our above framework (e.g. to store a link that appears as a blue link in our database as such), our software parses the CSS (cascading stylesheets) associated with each SERP. Since elements are represented the same away across SERPs, this is a straightforward task.

Metrics: Given the complexity of SERPs, there are many metrics one could use to understand the prominence of UGC in SERPs. The primary metric on which we focus is the *incidence rate* of each domain and element, an approach that is typical when assessing



Hours ▾

A Domino's Pizza
911 Foster St · (847) 563-3030
Open · Closes 12AM
WEBSITE DIRECTIONS

B Domino's Pizza
350 Ridge Rd · (847) 256-9020
Open · Closes 12AM
WEBSITE DIRECTIONS

C Domino's Pizza
4510 Oakton St · (847) 675-7711
Open · Closes 12AM
WEBSITE DIRECTIONS

More locations

Domino's Pizza (@dominos) · Twitter
https://twitter.com/dominos

BEYOND ready to dig in to this pizza. 😝😝🍕
pic.twitter.com/V4Ae4zN...
1 hour ago · Twitter

I'll take a bite of that, cheese and thank you. 😍🧀🤣
pic.twitter.com/6C9XXAJ...
1 day ago · Twitter

Tonight's agenda: 🍕, 🍕...and more 🍕.
pic.twitter.com/alv6DRL...
3 days ago · Twitter

Domino's Pizza Delivery Near You - Foster St
https://pizza.dominos.com › Illinois › Evanston ▾
Order **Pizza** at 911 Foster St: Tasty & Fresh From the Oven! ... From Our Oven on Foster St to Your Door: Evanston **Pizza** Delivery. Order from your Foster St **Domino's** and you'll see—we're pretty serious folks when it comes to **pizza** delivery.

Figure 4.1. Example Google SERP elements.

the prominence of content on SERPs [150, 231, 324, 429]. The incidence rate is the fraction of SERPs in a given query category on which a domain or element appears. For instance, if a Wikipedia link appears in 10% of SERPs for a given query category, Wikipedia would have an incidence rate of 0.10 for that category.

As we discussed above, research has shown that higher-ranking content gets substantially more traffic, so we also calculate the *top-three incidence rate* of each domain or element. This is the fraction of SERPs that have a given domain or element in their top three rows. When necessary to avoid ambiguity between incidence rate types, we refer to the basic form of incidence rates as *full-page incidence rates.*

In calculating both incidence rates, we treat SERP elements like carousels as a single item (hence the NewsCarousel has its own incidence rate), but also count the content of carousels as items (i.e. the tweets and news articles). For example, if there is a SERP with a *New York Times* article embedded in a rank 2 NewsCarousel, then that SERP will increase the top-three incidence rate for the domain nytimes.com *and* the element NewsCarousel.

4.3.4. Classifying Content as UGC

A critical requirement of our analyses is the ability to distinguish between UGC and non-UGC search results. Because there is no consensus definition of UGC [402], we operationalized two definitions from the literature: a *platform-centric* definition and a *content-centric* definition.

The platform-centric definition is one that is commonly used in UGC research, often implicitly (e.g. [48, 192, 230, 166]. Broadly speaking, this definition assumes that any content is UGC if it appears on a platform that hosts a large amount of UGC. There are a few prominent definitions of UGC that explicitly adopt a platform-centric perspective. For instance, Luca [253] provides a categorized list of popular UGC platforms and Dhar and Chang [79] define UGC as the "conjunction of blogs and social networking sites".

To operationalize our platform-specific definition, we cross-referenced the list of domains encountered in our data collection process with those on Luca's 2015 list. More specifically, this means that, under the platform-centric definition, we categorized as UGC any content from the following domains: Wikipedia, Facebook, Twitter, YouTube, Instagram, Yelp, LinkedIn, and TripAdvisor. The coding we describe here was applied only to content from these domains.

In a 2007 report [402], the Organization for Economic Cooperation and Development (OECD) offered a stricter definition of UGC that focused on content rather than platform. Under the OECD definition, UGC must (1) be published, (2) require some creative effort (i.e. not be a copy of some existing content), and (3) be created "outside professional routines and practices." To determine whether content was UGC under this definition, we used a qualitative coding approach and assessed whether a search result (e.g. a tweet in a Twitter carousel) met criteria #2 and #3 (all results implicitly met criteria #1). Specifically, our codebook instructed coders to view each search result (both the content on the SERP as well as the content on the linked website, e.g. a Twitter page) from our list of UGC platforms and identify (1) if the content appeared to be "creative" (i.e. not a copy of some other content) and (2) if the content appeared to be authored outside of professional "routines and practices". Coders used contextual information such as Twitter biographies or the presence of user reviews to judge whether the content appeared to be "professional" or not.

As we describe below in Results ("Types of UGC"), our results highlighted an interesting and meaningful contrast between the platform-centric and content-centric definitions of UGC. Upon discovering this contrast, we sought to better understand it by classifying content along two additional criteria that we hypothesized would be insightful based on what we saw in exploratory analyses. The first axis was related to who appeared to have authored the content: either an individual, an organization, or a bot. The second axis was related to the type of actor that authored the content. Using an inductive approach, we identified four types of individuals (*journalist, political figure, celebrity, other)* and five types of organizations (*journalistic*, *political*, *corporate*, *non-profit*, *other*).

To assess the reliability of the full coding scheme, we sampled up to 10 items for each UGC domain from our first dataset (a comparison of urban and rural search results: see below) and two researchers coded these samples. The researchers achieved substantial or perfect agreement in every case. For Facebook, Instagram, LinkedIn, Yelp, and TripAdvisor the coders achieved perfect agreement. They achieved a Cohen's Kappa of 0.87 and 0.67 for YouTube and Twitter respectively. Given this level of agreement, only one researcher coded the remaining samples.

4.3.5. Controlling for Geography

As noted above, based on our review of the search personalization literature, we expected that the importance of Wikipedia and UGC to Google might vary extensively by query location. As such, we developed a rigorous infrastructure to issue queries from a variety of carefully chosen simulated locations and planned to report our results with ranges defined by our location-specific results. However, upon running our experiments with this framework, we found that with respect to the prominence of Wikipedia and UGC in Google SERPs, there was little geographic variation.

As such, we in fact ran two separate experiments in this research project: (1) an experiment that rigorously considered potential geographic variation and, upon finding that this did not exist, (2) an experiment that used a simple population-weighted sample of geographic queries that could provide reliable single results for metrics of interest instead of ranges. We describe the methods we used in each of these experiments in turn below. The results of each experiment are discussed in detail in the sections that follow, as are the implications of the lack of variation that we observed.

Geographic Variation Experiment: Our geographic variation experiment was rooted in a spatial sampling approach that was designed to understand the maximum variation of UGC incidence across geography while at the same time maintaining a reasonable query rate. Our sampling strategy targeted three spectra on which the behavior of intelligent geographic technologies are known to vary: rural/urban, socioeconomic status (SES), and political preference (see e.g. [194]).

As our researchers had the most familiarity with Google's US search results, we restricted our focus to the United States. Choosing a specific study area for these and similar reasons is a common design choice in "GeoHCI" [168] and computational social science (as well as many other fields). We discuss possible expansions of this work to different geographic contexts in Future Work.

To generate specific geographic coordinates for the strategy outlined above, we used the following approach:

1. Urban-Rural: Using the urban-rural classifications by the U.S. National Center for Health Statistics (NCHS) [184], we sampled 10 counties from the most urban and most rural classes. These NCHS classifications are often leveraged in GeoHCI examining rural-urban issues [62, 194, 388]. We then used the centroid latitude and longitude provided by the U.S. Census for each county as a query location.

2. Income: We selected the top and bottom 10 counties in terms of 2015 median income, according to the 2011-2015 U.S. Census American Community Survey 5-Year Estimates [397], and executed the county-to-coordinate mapping as described above.

3. Voting: We selected the top and bottom 10 counties in terms of percentage of votes for Hillary Clinton in the 2016 U.S. Presidential election and again executed the same county-to-coordinate mapping. This county-level data was published by Townhall [391] and accessed via [264].

Population-weighted Experiment: As reported below in Results, the rigorous geographic comparisons described above showed little evidence of geographic variation in metrics of interest. As such, it was reasonable to use a single set of query locations to report our results. However, it was non-optimal to select one of our pre-existing location sets (e.g. most-rural or wealthiest) as representative and report those results for two reasons: (1) we did observe a (quite) small amount of variation across the spectra outlined above and (2) doing so may raise other ecological validity concerns.

As such, in our second experiment, we developed a new set of query locations that avoided both of these issues. To develop this set, we randomly sampled 40 U.S. counties

Query Category	**Top Domain**	**Rank of WP in Incidence Rate**	**WP Full-Page Incidence Rate**	**WP Top-3 Incidence Rate**
Trending	Wikipedia	1	0.81	0.50
Most-Popular	Wikipedia	1	0.77	0.28
Loans	wellsfargo.com	6	0.25	0
Insurance	progressive.com	10	0.29	0.08
Controversial	Wikipedia	1	0.90	0.51
Medical	webmd.com	5	0.45	0.15

Table 4.1. Prevalence of Wikipedia (WP) in Google SERPs by query category. Does not include SERP elements.

using a population-weighted approach. We then used the U.S. Census-provided representative coordinate for each of these counties as our query locations. We issued each query from each of these coordinates and report results averaged across all coordinates. Experiments were run in January 2018.

4.4. Results

We first report the results of our population-weighted experiment described above. We then provide additional context with the results of our geographic variation experiment.

4.4.1. Population-Weighted Experiment

Fig. 4.2 summarizes the results from our population-weighted SERP dataset. The figure shows the full-page and top-three incidence rates for all UGC domains in the dataset, as well as the top five non-UGC domains and SERP elements to provide context. Table 4.1 zooms in and focuses on the results for Wikipedia specifically.

The strongest signal present here is that Wikipedia is absolutely critical in Google's approach to responding to queries. More specifically, Fig. 4.2 and Table 4.1 show that

Wikipedia is not only the most prominent UGC platform on Google SERPs but also is in fact the single most prominent website of any kind on Google SERPs. For *trending* and *controversial* queries, Wikipedia appears in over 80% of first SERPs and rivals the prominence of structural elements like the NewsCarousel. For *insurance* and *loan* queries, the lowest incidence rates for Wikipedia in our study, Wikipedia still appears in over 25% of SERPs. When considering only the top-three result rows, Wikipedia remains very prominent, showing up in 50% of top-three results for some categories of queries (*trending*

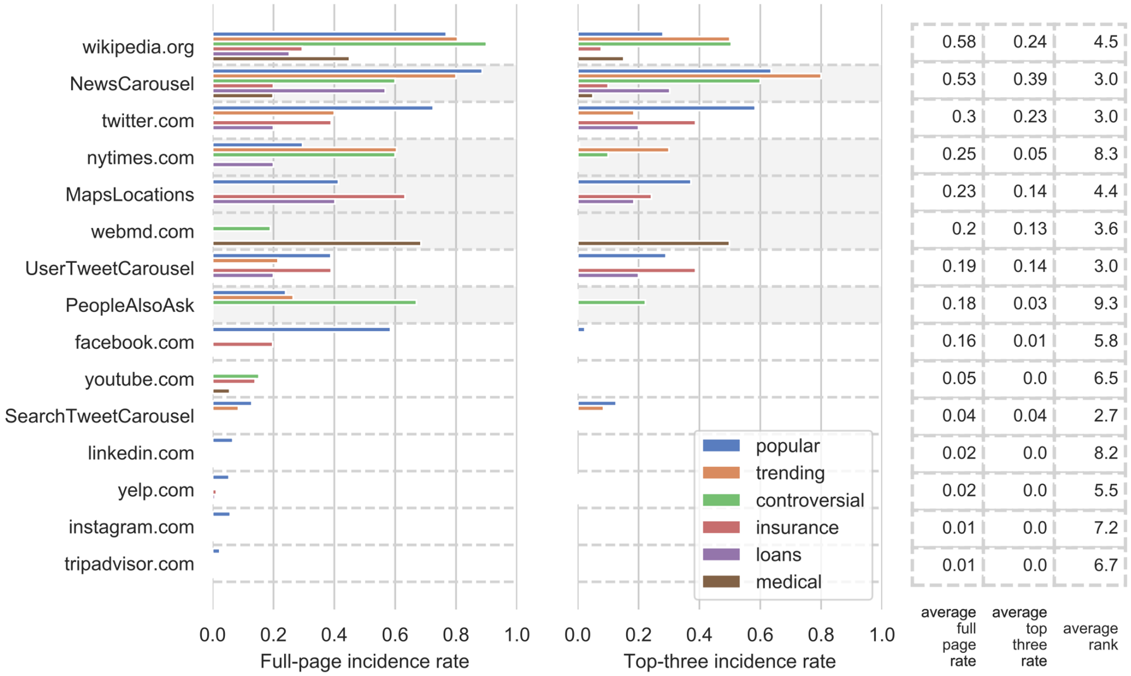


Figure 4.2. This figure summarizes key metrics for all UGC domains in our study and the top 5 non-UGC domains/elements (highlighted in light grey). Rows are ranked by average full-page incidence rate, shown on the right, followed by average top three incidence rate and average rank for each domain. The average incidence rates should be interpreted with a degree of caution as they are not intended to be representative of overall incidence rates, just the average across the six query categories we analyzed in this study.

and *controversial*). In aggregate, the average full-page SERP incidence rate for Wikipedia in our study was 0.58 (58% of SERPs had Wikipedia pages). Twitter's 0.30 is the next-highest average full-page rate.

One concrete example of a query in our sample for which Wikipedia is very important is "minimum wage" from our *controversial* query category. The SERPs for "minimum wage" had a link to the Wikipedia article "Minimum wage in the United States" in two places: a rank 1 AnswerBox, and a blue link at rank 6. On the other hand, an example of a query for which Wikipedia is less important is the query "life insurance," where Wikipedia showed up at rank nine.

Beyond Wikipedia, we additionally see that Twitter is also important to Google's ability to respond to queries in many of our categories. For instance, for *most-popular* and *trending* queries, the full-page Twitter incidence rate is above 40% and the top-three incidence rate for *most-popular* queries is higher than that of Wikipedia. Interestingly, however, Twitter almost never appears in *controversial* and *medical* SERPs, perhaps a reflection of a specific design decision at Google.

Types of UGC: Our Twitter results – combined with the non-trivial prevalence of other UGC platforms for certain types of queries (e.g. Facebook for *most-popular* and YouTube for *controversial*) – seemingly suggest that Google's dependence on UGC extends significantly beyond Wikipedia. Indeed, using a platform-centric definition of UGC, this is the case.

However, the results of our content-centric qualitative coding exercise demonstrate that the platform-centric perspective is problematic for our research. Of the 345 unique social media results that appeared in our collected data, 95% failed to meet the OECD's content-centric definition of user-generated content (we note that together, these links appeared approximately 5,000 times in our dataset, because many results like the Tweet-Carousel and corporate social media pages were identical across locations). In particular, 73% of the links came from an official corporate account. Another 14% were from an official political account and 4% were journalistic. This means that while Google surfaces a substantial amount of content from non-Wikipedia UGC platforms, almost all of this content is not UGC from a content-centric perspective. Instead, this content resembles that on typical webpages: it is written by professionals. We return to this point in the Discussion.

Effect of the Knowledge Panel: While the Knowledge Panel lacks a "rank" in the desktop version of Google, it is still possible to re-compute the full-page incidence rate of each domain including the Knowledge Panel. Since Wikipedia is prominent in the Knowledge Panel, this calculation substantially boosts Wikipedia's average full-page incidence rate from 58% to 69%. Although we saw some social media links in the Knowledge Panel, every one of them linked to an organizational account, so these would not influence our conclusions above.

4.4.2. Geographic Variation Experiment

As noted above, we were interested to find that – although prior work has highlighted the influence of location-based personalization for some queries – we saw very few meaningful differences in the importance of UGC across the spectra that we considered (urban vs. rural, SES, political preference). For the few cases in which we saw meaningful variation, the effect size was quite small.

We assessed the variation across the three geographic spectra by comparing full-page and top-three incidence rates from one end of each spectrum to the other. We performed these comparisons for every UGC domain (thus assuming a platform-centric definition of UGC) and across every query category. We tested to see if different types of locations saw different UGC domains at significantly different rates. To compute the median difference in incidence rates, we only considered the 115 comparisons in which a domain appeared at least once (e.g. Yelp pages never appeared for *medical* queries, so we did not include geographic comparisons of Yelp incidence rate for *medical* queries).

For Wikipedia, the median full-page incidence rate difference across all spectra and query categories was only 0.01, and the maximum was only 0.16 (political spectrum and *most-popular* queries); the Wikipedia top-three incidence rate median difference was 0 and the maximum was 0.07. The median across all 115 comparisons was 0.01. Moreover, of these comparisons, only 15 differences were statistically significant based on Fisher's exact test ($p < 0.05$), i.e. in most cases we fail to reject the null hypothesis that UGC is equally likely across geographic strata. When considering only links that meet the content-centric definition of UGC, only 9 differences were significant. In other words, observable geographic variation in UGC was rare.

4.5. Discussion

4.5.1. Distribution of the "Technological Dividend"

The most significant signal in our results is the critical role that Wikipedia plays in helping Google accomplish one of its most important goals: satisfying user information needs. For the English-language queries that we considered, Google is more dependent on Wikipedia than any other website in the world. Moreover, for some of Google's highest-volume queries (*trending* and *most-popular*), Wikipedia appears on a large majority of results pages.

These results help to inform a highly-consequential discussion about the economics of computing that is moving from the margins of the literature [13, 163] and into mainstream debate [218, 257, 307, 309]. This discussion centers on potential asymmetries in the relationship between users and lucrative intelligent technologies: user-generated data is immensely important to such technologies, but many argue that users are not receiving a proportional share of the economic benefits from these technologies. Our results certainly point to one such potential asymmetry: the Wikipedia community creates tremendous value for search engines, but search engines only donate a relatively small amount of money to support the Wikimedia community [298, 349]. This finding raises provocative questions that can advance this discussion, e.g. are Wikipedia editors some of the most important and underpaid employees of search companies?

Arrieta Ibarra and colleagues [13], Hecht [162], and others [307, 309] have identified information imbalances between intelligent technology owners and data creators as a key mechanism for the current distribution of economic benefits of intelligent technologies. While the developers of intelligent technologies know many such technologies would struggle substantially without constant "data labor" by their users and others (e.g. Wikipedia editors), most people have very little understanding of the value of their data-generating labor. These authors have argued that the research community should therefore work to level the playing field by measuring and making people aware of the value their data brings to intelligent technologies [163]. Our results make a contribution towards this goal, and also point to the importance of engaging in similar investigations in related domains (e.g. OpenStreetMap, Wikidata).

The discussion about the distribution of the technological dividend must also consider the value of the service that intelligent technologies "trade" for data-generating labor. After all, most Wikipedia editors benefit heavily from their use of Google, and McMahon et al. showed that Wikipedia itself does as well [265]. As such, our results point to an additional important area of future research: doing qualitative and quantitative work to understand whether the Wikipedia community believes anything should change in Wikipedia's relationship with intelligent technologies given the increasing informational equality on this topic.

Additionally, our results also highlight a related line of inquiry centered around a key question: How can we reduce any discrepancy between the value created by data like Wikipedia articles and the rewards received by those who created the data. Hecht [163] and others [13] have suggested that collective action by users – e.g. through boycotts, "data strikes", or data unions – can be one possible solution. Indeed, our work in this thesis has highlighted the potential impact that data strikes, boycotts, or combinations thereof could have on intelligent technologies. However, other, less confrontational approaches (which may also be more immediately tractable than widespread data strikes or data unions) may be possible and are likely desirable. For instance, just making visible the value of Wikipedia to search engines could encourage search engine companies to more prominently credit Wikipedia through design changes or to contribute donations of money or data.

More generally, Madsbjerg [257] and McMahon et al. [265] have argued that one major challenge in analyzing the value of user-provided data to profitable intelligent technologies is that the required datasets for such analyses are almost always private. This chapter highlights an approach we believe can, at least initially, be quite effective at addressing this challenge: focusing on Wikipedia and other UGC rather than more difficult-to-access types of user-generated data such as search logs and personal information that also play critical roles in intelligent technologies. Our results show that just focusing on more open types of information is sufficient to at least begin the empirical examination of these issues.

4.5.2. Wikipedia Matters outside Wikipedia

The social computing and computational social science communities have developed a large literature on Wikipedia. This literature has examined topics ranging from the content coverage biases that exist in Wikipedia (e.g. [165, 194, 304, 316] to the collaboration patterns between editors that lead to the highest-quality content [440, 441].

Our results further bolster the importance of this literature by showing that the literature's findings have implications far beyond the boundaries of Wikipedia. For instance, prior work has shown that the English Wikipedia has more missing articles about women than about men [316] and similar patterns have been observed with respect to Wikipedia's coverage of some geographic areas versus others [194]. Our results highlight that not only do these biases affect reader experience on Wikipedia, they also affect Google's ability to address information needs associated with the disadvantaged topics. That is, if

Wikipedia has less information about a topic of interest to a certain group, this will also affect Google's ability to address information needs related to this topic.

4.5.3. Definitions of User-Generated Content

Our findings related to the platform-centric versus content-centric definitions of UGC may have important methodological implications for some UGC research. In particular, we found that the platform-centric definition of UGC was problematic in the context of our study. Had we relied on this definition exclusively, we would have believed that everyday Twitter users were powering Google at almost the same rate as Wikipedia editors. Instead, thanks to our qualitative analysis, we discovered that at the content-level, the vast majority of tweets surfaced by Google in our study were not UGC but rather were written by professionals. In other words, these tweets are analogous to short-form company websites (and, in some cases, news articles).

This result highlights calls [333] for researchers to consider the nature of content on platforms that host UGC like Twitter before making assumptions about its professional or amateur nature. While much research currently takes care to do basic filtering for bot-created content – and there are well-known approaches for doing so in certain platforms [77] – filtering out organizational and other professional accounts will be more difficult and is deserving of further research along the lines of McCorriston et al. [263].

4.5.4. Geographic Personalization and UGC

Our geographic comparisons suggest that personalization based on geographic location may be non-substantial for certain types of search phenomena. This may simplify methods for some search auditing research projects, but more work is needed to understand when controlling for geography is necessary and when it is not. We note that we did see substantial variation for content other than UGC, e.g. Google Maps SERP elements. Additionally, given that Wikipedia is not equally comprehensive in all languages [165] and that platforms like Twitter are not equally popular in all countries [346], geography likely matters *across* national and linguistic borders. Future studies should address this directly.

4.5.5. Limitations

As is typical in the search auditing literature, although we aimed to generate queries systemically, the immense number of search engine use cases makes it impossible to generate a truly representative query set for data collection (at least from a position outside of an institution that operates a large search engine). As such, we emphasize that our results must be considered in the context of the queries we selected. This means that while our results likely generalize to many queries that are similar to our query sets, for instance queries about popular commercial entities or queries about common health problems, our results may not generalize to all search engine use cases, such as complex, infrequently-made queries. Furthermore, our *most-popular* query set is constrained by Google's willingness to share query volume data, as well as the manual process of selecting categories from Google Trends. It should also be noted that this query set may contain some thematic overlap (though no actual query overlap) with other query sets, e.g. *controversial.*

While we controlled for the effect of geography within the US, extending this analysis to include additional countries as noted above could provide valuable insight into the importance of Wikipedia and UGC globally. This analysis would require parsing other prominent search engines (e.g. Naver) and identifying appropriate geographic spectra.

Another direction for future research would to be to expand data collection and analysis to mobile devices. As mobile devices are used heavily for local search [387], focusing heavily on local queries would make sense in this case. Given that we saw extensive variation across query categories, extending our work to consider additional categories would also be a useful direction of future work.

Search engines, like many intelligent technologies, are constantly changing. Therefore, longitudinal auditing will be valuable, both to account for revisions to SERPs – i.e. new specialty boxes and elements – and to account for algorithmic changes. For instance, the importance of UGC sources may vary as search engines integrate new techniques from the deep learning (e.g. RankBrain [59] or structured knowledge domains (e.g. Knowledge Vault [85]). Indeed, the introduction of these technologies may be responsible for the decrease we observed in Wikipedia full-page incidence rate for medical queries relative to the work of Laurent and Vickers [231] last decade (although the methods are not directly comparable). Doing this longitudinal auditing will require careful attention to edge cases, which means that recurring human validation (and likely updating our software accordingly) will be critical for future research in this direction.

Using our software will require updates based on changes to SERP structure. Researchers may also want to implement our approach using headless browser-based techniques, which likely require less overhead, or consider the recently released framework by

Robertson et al. [324] that uses on a Chrome plug-in and crowd workers from CrowdFlower and Prolific Academic [311, 400]. Though Robertson et al.'s focus was on using their framework to study political personalization in web search, Robertson et al.'s data reveals they were able to replicate our results about the importance of Wikipedia in the political domain, adding an additional degree of rigor to the findings above.

Finally, UGC like Wikipedia has been used to train intelligent technologies, including by Google (e.g. for language understanding [170]). This is an entirely separate avenue by which UGC creates value for the owners of intelligent technologies. A very promising research direction would be to measure how the inclusion of UGC impacts the performance of these algorithms, similar to the previous chapter. However, doing the same for search engines and other private intelligent technologies will require creative approaches as it will likely require extensive access to proprietary software and data.

4.6. Conclusion

This chapter provides evidence that UGC, largely in the form of Wikipedia articles, is immensely valuable to web search engines. Examining six categories of queries, we found that Wikipedia's volunteers have created a resource that is critical to Google's ability to address information needs. Our results contribute to the growing discussion around potential economic asymmetries in the relationship between the people who create data and intelligent technologies that rely on this data. Our findings also have implications for Wikipedia research on content coverage and for methods in search auditing and UGC research.

CHAPTER 5

A Deeper Investigation of Wikipedia's Value to Search Engines

5.1. Introduction

We have thus far discussed the critical interdependencies between Wikipedia and Google Search[1]. For many query categories, Wikipedia links appear more than any other website on Google's search engine results pages (SERPs). This suggests that Wikipedia—a resource made by volunteers—plays an essential role in helping Google Search achieve its core function of addressing information needs. It also raises the stakes of Wikipedia-related research findings, with Wikipedia's collaboration processes, contribution patterns, and content outcomes being imperative to the success of Google Search.

In this chapter, we replicate and extend earlier work that identified the importance of Wikipedia to the success of search engines but was limited in that it focused only on desktop results for Google and treated those results as a ranked list, i.e. as "ten blue links" [52]. Like this earlier work, we collect the first SERP for a variety of important queries and ask, "How often do Wikipedia links appear and where are these links appearing within SERPs?" However, our work includes three critical extensions that provide us with a deeper view into the role Wikipedia links play in helping search engines achieve their primary goals. First, we consider three popular search engines – Google, Bing, and DuckDuckGo – rather than just Google. Second, in light of extensive search engine use on

[1] This chapter was originally published as [404].

mobile devices [378], we simulate queries from both desktop and mobile devices. Third, we built software to allow us to use a spatial approach to search auditing that is more compatible with modern SERPs, which have moved away from the traditional "10 blue links"; SERPs now have "knowledge panels" in a second, right-hand column, "featured answers", social media "carousels", and other elements that influence user experience. Our spatial approach to SERP analysis, which considers the location of each link within a SERP, also allows us to visually compare the representation of our SERP data to screenshots of SERPs. This is important because search engines are proprietary, opaque, and subject to frequent changes (e.g. Google has previously redesigned their SERPs [360]), and it is critical to validate that data being used for quantitative search auditing analysis accurately reflects how SERPs appear to users. We are sharing our data and code for replication purposes and to support additional search audits taking this approach.[2]

After running our audit, we found that the results identified in prior work regarding Google desktop search largely extend to other search engines and, with a smaller effect size, to mobile devices. For desktop SERPs, Wikipedia appeared in 81-84% of SERPs for our *common* queries, 67-72% for our *trending* queries, and 16-54% for our *medical* queries. Results are similar, though somewhat lower, for mobile devices. Furthermore, using our spatial analysis approach, we observe that for many of these queries, Wikipedia appears in the prominent area of SERPs visible without scrolling, through both Knowledge Panel elements and traditional blue links. We did, however, observe that Wikipedia appears

[2] See https://github.com/nickmvincent/se-scraper for archival repository of code used for the original dataset collection. See https://github.com/nickmvincent/LinkCoordMin for more recently updated SERP collection software.

less prominently for mobile SERPs, though they are still quite prominent and, in several situations, appear more often than they do in the desktop Google SERP case.

Our results have implications for a variety of constituencies. Past work suggested that Wikipedia was very important to desktop Google users. On one hand, this means search users stand to benefit from the high-quality knowledge produced by the Wikipedia community. On the other hand, this means search results are impacted by Wikipedia's weaknesses such as content biases. Our results suggest that this relationship extends beyond just desktop users of Google, although the presence of Wikipedia is smaller in mobile SERPs.

Finally, our results also reinforce the importance of volunteer-created data to the success of intelligent technologies. As we discuss below, by measuring how often and where Wikipedia links appear in SERPs, we can better understand to what degree and how the volunteer labor underlying Wikipedia fuels search engines, some of the most lucrative, important and widely used intelligent technologies. These findings suggest a need to rethink the power dynamics and economic relationships between people who generate content and the intelligent technologies that rely on this content.

5.2. Related Work

This chapter builds on research that has studied search engines and SERPs, as well as research that has specifically studied the role of Wikipedia articles in search results.

In McMahon et al.'s 2017 experimental study [265], a browser extension hid Wikipedia links from Google users and the authors observed a large drop in SERP click-through rate, possibly the most important search success metric [146]. This study, and the work described in the previous chapter, were motivated by a call from the Wikimedia Foundation to study the re-use of Wikipedia content [385]. A key concern was raised by McMahon et al. regarding how Google SERPs provided peer-produced content with attribution. McMahon et al. also discussed the "paradox of reuse": if SERPs do not provide searchers with links to Wikipedia (or other peer production platforms), they may deprive peer production communities of new members. We return to questions around attribution and reuse in SERPs below in our Discussion.

These projects are part of a broader body of search auditing literature, which has used auditing techniques (i.e. collecting the outputs of search engines) to study aspects of search such as personalization and the special components, such as "Knowledge Panels", that appear in SERPs [155, 204, 324]. Notably, in a study that focused on political SERPs, Robertson et al. [324] identified a large number of Wikipedia articles in the SERP data they collected, in line with the results from the user-generated content-focused study in the previous chapter. Related work has focused specifically on the complexities of SERP elements like the Knowledge Panel [254, 330]. Lurie and Mustafaraj found that Wikipedia played a critical role in populating the "Knowledge Panel" shown in news-related Google SERPs: queries about news sources with Wikipedia articles frequently had Knowledge Panels in the corresponding SERPs [254]. Rothschild et al. studied how Wikipedia links in Google Knowledge Panel components influenced user perception of online news [330].

Search auditing work has also focused on the topic of tech monopolies: an audit by Jeffries and Yin found that much of the content in Google SERPs was Google's own products (e.g. modern SERP components likes Maps and answer boxes) [190]. This work provides motivation for search auditors to consider modern SERP components. Additionally, it has motivated development on a “Simple Search” tool that returns to a “ten blue links” SERP [401].

The study that we most directly replicate and extend is the previous chapter, which used a ranking-based analysis to examine the role user-generated content played in search results for the desktop version of Google search. Our results provided motivation for us to focus specifically on Wikipedia while considering more search engines, mobile devices, and a spatial analysis approach.

It is also valuable to note that in addition to the academic literature on Wikipedia and search, a number of studies conducted by search engine optimization (SEO) companies have also measured how frequently Wikipedia appeared in SERPs [137, 138, 189, 323]. Estimated incidence rates range from 34% to 99%, depending on the choice of queries (a point to which we will return below).

Our analysis was additionally motivated by prior work that studied how users interact with SERPs. Prior work has shown that visual attention and clicks are directed towards higher-ranking links, especially the first two links. Papoutsaki et al. replicated three seminal search behavior studies using a webcam eye tracker [296]. In addition to replicating the concentration of attention and clinks on top-ranked links, they found in some cases that ads in right-hand column of a SERP can attract about as much visual attention as the 3rd and 4th links.

In designing our spatial analysis, we were influenced by the above findings relating to the role Wikipedia plays in populating Google’s Knowledge Panel, the robust results regarding the importance of top ranked links, and the potential value of right-hand column.

Below, we describe how we used these findings to define and interpret spatial incidence rates.

5.3. Methods

As mentioned above, modern SERPs include SERP components that are much more complex than “ten blue links”. These multifaceted SERPs pose challenges to ranking-based approaches used in prior work, such as work in the previous chapter which used Cascading Style Sheets (CSS) styles to differentiate different result types (e.g. blue link vs. Knowledge Panel vs. News Carousel) and create a ranked list of these results [408]. The critical challenge with ”ranking list” approaches is that they struggle to account for (1) the fact that SERP elements come in a variety of sizes and (2) the fact that SERP elements appear in a separate right-hand column. Furthermore, to differentiate results with CSS, researchers must manually identify which styles are associated with different result types, and these styles are liable to change.

In contrast to this prior work, we consider every link (i.e. each HTML <a> element) in each SERP and look at the coordinate for each link, in pixels, relative to the top left corner of the SERP, an approach specifically designed to handle modern SERPs. There are several other high-level benefits to this approach. It is easier to analyze multiple search engines (the approach requires substantially fewer hard-coded rules based on CSS styles), and we can visually validate that the representation of SERP data we analyze reflects the appearance of modern, component-heavy SERPs.

Below, we detail our data collection, validation, and analysis pipeline, and expand on the benefits of this spatial approach to SERP analysis.

5.3.1. Search Engines and Queries

In this work, we focus on three U.S. based search engines: Google, Bing and DuckDuckGo. Google and Bing are the two most popular search engines in the United States, while DuckDuckGo is the most popular privacy-focused search engine [96]. Market share estimates from different analytics services vary. In January 2020, analytics company StatCounter (`https://gs.statcounter.com`) estimated Google served 81% of U.S. desktop queries, Bing served 12%, and DuckDuckGo served 1.5%, and for mobile devices Google served 95%, Bing served 1.5%, and DuckDuckGo served 1.2%. Data from marketing research company ComScore suggests Bing has a larger market share; it reports that Google served about 62% of desktop search queries while Bing served about 25% [64].

Query selection is a critical and challenging aspect of search auditing work. Query datasets are highly proprietary, and in the past sharing queries has harmed users' privacy [18]. For these reasons, companies do not make query data available to sample from. What constitutes an "important" search query may change with current events and other factors. The results of a search audit are highly contextual based on queries used. To address these challenges, we followed best practices in prior work [155] closely and aimed to use queries that we believe to very important to a large fraction of search engine users.

Specifically, we consider queries from three important query categories: *common* queries, *trending* queries, and *medical* queries. *Common* queries are made very frequently, and thus are important by virtue of query volume. *Trending queries* are those that received a spike in interest, typically relating to news and current events. Thus, SERPs returned for *trending* queries can influence how search engine users learn about what is happening in the world. Finally, *medical* queries concern medical issues, and have the potential to impact users' health decisions. Notably, past work looking at Google SERPs found a lower prevalence of Wikipedia links for *medical* queries than other search categories [408]. In all three cases, the queries we used in our study are queries that we can plausibly assume are made very frequently, i.e., the corresponding SERPs were seen by a large group of search engine users.

For each query category, we needed to seek public sources of query data because query datasets are highly proprietary. Although search engine operators do not share their raw data about query volume, search engine optimization (SEO) companies collect data to estimate query volumes. Thus, to create a list of *common* queries, we took the top 100 queries by volume in October 2019, as estimated and made public by SEO company ahrefs.com [373] (we use their list that filters out "not safe for work" queries). To create a list of *trending* queries, we took all 282 queries from Google's public set of top trending queries from 2018 (we prepared our queries in 2019 and collected data in February of 2020) [140]. Unlike the ahrefs data, Google's trending query data contains no information about query volume – it provides the top trending queries for pre-selected topics (e.g., politics, music, etc.). While five of these query categories were in Spanish, the rest were English language queries (we used all queries regardless of language). Finally, to obtain important *medical* queries, we used a list of the top 50 medical queries made public by Bing for previous information retrieval research [371]. In total, we consider 432 different important queries.

Although we designed our query sets to be similar to those used in prior work, our query categories also map to different types of search intent. In the search literature, queries are often classified as "navigational" (a user wishes to navigate to some website, e.g. Facebook, via a SERP), "informational" (a user wishes to learn information about the query), and "transactional" (a user wishes to make some transaction) [188]. Past work suggests that users can have large differences in search behavior for different query intents, spending more time examining results for informational queries [296]. While there is not a clear objective mapping between individual queries and user intent categories (i.e. a single query could be associated with different user intents), considering approximate associations between our query categories and user intents is useful in identifying potential implications of our results for search users.

The *common* category appears to correspond to navigational queries, i.e. most queries refer to a major websites and companies. The top five queries are "facebook", "youtube", "amazon", "gmail", and "google". On the other hand, the *trending* category appears to primarily correspond to informational queries. Example trending queries include "World Cup" (a global sports event), "thank u, next" (a popular music album released in 2018), "How to vote", "Alexandria Ocasio-Cortez" (a U.S. politician), and "What is Fortnite". Finally, the *medical* queries we use also seem to be *informational*, e.g. "indigestion", "how to lose weight", "common cold symptoms", "acid reflux", and "can't sleep". The full list of queries is available with the code package.

The mappings of our query categories to user intents described above allows us to discuss our results in light of these user intents. For instance, what does it mean to have a high incidence rate of Wikipedia links for navigational queries? Many people searching for "facebook" (the most searched query, with an estimated monthly volume of 232 million searches) are likely not seeking information about Facebook. Nevertheless, as we will see in our results and discuss further below, desktop users are likely to be exposed to Wikipedia content as part of their navigational query, whereas mobile users are not.

5.3.2. Data Collection

We collected data programmatically using software we built that extends the *se-scraper* JavaScript library. The *se-scraper* library uses *puppeteer* to run a headless Chrome browser that can make search queries, save SERP results as HTML, and record screenshots of each page.[3] The screenshot functionality is important to our ability to validate that the SERPs collected by a headless browser resemble real SERPs. For each query, we collect only the first SERP.

Critically, the package we built stores the coordinates of each clickable hyperlink (specifically, HTML <a> elements with nonzero width and height) within a SERP. The package extracts the top coordinate (how many pixels from the top edge of the SERP is the top border of the link) and left coordinate (how many pixels from the left edge of the SERP is the left border of the link) of each link using JavaScript's *Element.getBoundingClientRect* function.

Our software can emulate mobile devices by providing a mobile user-agent string and using *puppeteer*'s "Devices" API. To collect mobile SERPs, we simulate an iPhone 10 running the default Safari browser. To simulate desktop queries, we use se scraper's default user-agent, which corresponds to Google Chrome running on a Windows 10 machine.

[3] *se-scraper*: `https://github.com/NikolaiT/se-scraper`,
puppeteer: `https://github.com/puppeteer/puppeteer/`

Past work has suggested that while geography plays a major role in the personalization of SERPs [204], this personalization does not heavily impact the incidence rate of user-generated content like Wikipedia for different locations with the United States [408]. As such, we made queries from a single urban location in the United States, which made it possible to include mobile devices and multiple search engines while limiting the time cost associated with data collection.

5.3.3. Data Validation

Given the proprietary and dynamic nature of search engines, SERP data can change its format unexpectedly and it is therefore critical to manually validate that the representation of SERP data being analyzed (e.g., links and their coordinates) resembles a SERP as consumed by users. Search engines frequently change the appearance of SERPs or add entirely new features, e.g. special components that are very different than "ten blue links" [52]. For instance, Google recently substantially redesigned their SERPs, moving link URLs above "blue links" [360]. SERP analysis software designed to analyze SERPs that works well one day could easily fail the next when a major design change—like Google's recent revamp—is rolled out.

Our software is designed to allow an auditor to validate that the representation of SERP data matches the appearance of SERPs. As described above, our software extracted all link elements and their coordinates. We created a visual representation of links at their coordinates and compared this representation to a screenshot of the SERP produced by our data collection software. The purpose of this step is to ensure that our spatial representation of SERP links reflects how a person would view the page, i.e., we ensure that our extracted link coordinates match the actual appearance of the SERP. We performed this validation for 5 random queries per configuration (device, search engine, and query category), for a total of 90 samples. This process is focused on ensuring that the extracted link data (i.e. links and their associated coordinates in the page) matches the screenshots captured during data collection.

Throughout our research process, we also engaged in manual verification checks that entailed comparing our programmatically generated SERP screenshots to SERPs generated by a separate human-operated web browser. Our goal here is to not ensure an exact match (the personalization of SERPs means that two SERPs made for the same query might not match exactly [155]), but rather to ensure that there are no major structural issues with data collection in the context of constantly changing complex SERP design.

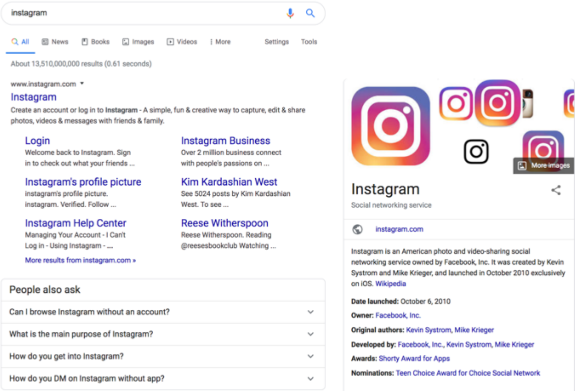
This process was effective in identifying and fixing data collection errors in light of the challenges posed by SERP data. For instance, using our visual validation and sanity check process, we discovered several data collection nuances and inconsistencies, e.g., SERPs that partially loaded, were blocked by a "location access" prompt, or failed to load entirely.

5.3.4. SERP Analysis

Our analysis is based on the coordinate-based representation of SERP links described above. Specifically, our data represents hyperlinks as a triplet consisting of domain, left coordinate, and top coordinate. For instance, an English Wikipedia link with left coordinate at 200 pixels and top coordinate at 300 pixels would be represented as the triplet (*en.wikipedia, 200, 300*).

Based on our representation of SERPs, we define a series of Wikipedia *incidence rates* that measure how often Wikipedia links appear. To begin, we measure the *full-page* incidence rate, or the fraction of SERPs in which Wikipedia links appear for a combination of device, search engines, and query category. We leverage our spatial analysis to better understand where Wikipedia links appear on a SERP, an important consideration for evaluating how salient these links are.

More specifically, we define four spatial incidence rates: *above-the-fold*, *left-hand, right-hand,* and *left-hand-above-the-fold*. The above-the-fold incidence rate is meant to capture how often links appear "above the fold", the portion of the SERP that is visible without scrolling. We know from prior work that top-ranked links receive much more visual attention and clicks [296, 422], but we also know that modern SERPs do not appear as a ranked list of ten, equally sized blue links – the rank 1 element could be a large SERP element that takes up most of the viewport (e.g. a map element). The above-the-fold incidence rate is a useful heuristic that measures how often links appear in prominent positions, while accounting for the impact of large SERP elements.



About 13,510,000,000 results (0.61 seconds)

www.instagram.com

Instagram

Create an account or log in to **Instagram** - A simple, fun & creative way to capture, edit & share photos, videos & messages with friends & family.

Login

Welcome back to Instagram. Sign in to check out what your friends ...

Instagram Business

Over 2 million business connect with people's passions on ...

Instagram's profile picture

instagram's profile picture. instagram. Verified. Follow ...

Kim Kardashian West

See 5024 posts by Kim Kardashian West. To see ...

Instagram Help Center

Managing Your Account - I Can't Log in - Using Instagram - ...

Reese Witherspoon

Reese Witherspoon. Reading @reesesbookclub Watching ...

More results from instagram.com »

People also ask

Can I browse Instagram without an account?

What is the main purpose of Instagram?

How do you get into Instagram?

How do you DM on Instagram without app?

Instagram

Social networking service

instagram.com

Instagram is an American photo and video-sharing social networking service owned by Facebook, Inc. It was created by Kevin Systrom and Mike Krieger, and launched in October 2010 exclusively on iOS. Wikipedia

Date launched: October 6, 2010

Owner: Facebook, Inc.

Original authors: Kevin Systrom, Mike Krieger

Developed by: Facebook, Inc., Kevin Systrom, Mike Krieger

Awards: Shorty Award for Apps

Nominations: Teen Choice Award for Choice Social Network

Figure 5.1. A Google SERP for the common query "Instagram" with special components, including a right column "Knowledge Panel" with a Wikipedia link.

The left-hand incidence rate is a proxy for how often Wikipedia links appear in the left column (where traditional “blue links” and other SERP components appear), while the right-hand incidence rate is a proxy for how often Wikipedia links appear in Knowledge Panel components. Finally, we additionally define a left-hand-above-the-fold incidence rate to understand how often Wikipedia links appear above the fold without appearing in the Knowledge Panel. In other words, the left-hand-above-the-fold incidence rate helps us easily answer the question: “Is Wikipedia primarily appearing above the fold because it appears in Knowledge Panel components?”

To create an operational definition of left-hand and right-hand incidence rates, we manually identified a vertical dividing line between the left column and the right column containing the Knowledge Panel and related components. To obtain this line, we examined panel elements for all three search engines and found that a vertical line at 780 pixels was able to cleanly divide the left column and right column for all three search engines. We do not consider left-hand or right-hand incidence rates for our mobile SERPs, as mobile results are presented in a single column.

Creating an operational definition for above-the-fold incidence is more challenging, as devices and settings (e.g., browser zoom, browser resizing, etc.) affect what content is visible without scrolling. To address this uncertainty, we considered multiple device viewport sizes to allow us to obtain lower bound, middle ground, and upper bound estimates of above-the-fold incidence.

For mobile devices, we considered a range of viewport heights corresponding to different devices [124]. Our lower bound was 667 pixels, the height of smaller iPhone 6/7/8 devices, and close to the height of Galaxy S7 devices. Our middle ground viewport estimate was 736 (corresponding to iPhone 6/7/8 plus and Galaxy S8/9 devices). Our upper bound viewport estimate was 812 pixels (corresponding to the large iPhone X and Google Pixel 3 devices).

For desktop devices, for our middle ground estimate, we consider the common viewport height of 768 pixels. Our lower bound scenario corresponds to a window at 110% zoom (i.e. zoomed in so that less content shows above the fold), our middle ground corresponds to 100% zoom (the default), and our upper bound corresponds to 90% zoom (i.e. zoomed out so that more content shows).

Overall, this range-based approach means that while our above-the-fold incidence rates make some assumptions about browsing configuration, we can observe how robust our results are against variations in these assumptions. Of course, these lower and upper bounds could be chosen based on factors other than a desktop user's zoom level (e.g., a user might resize their window without changing their zoom level, use external monitors, etc.). Our approach here is select a reasonable starting point, and future work might attempt to incorporate more such factors.

One important consideration for above-the-fold incidence is how it relates to ranking-based measurements like top-k incidence rate. For instance, based on findings that most clicks go to the top three ranked items in a SERP, past work has looked at the "top-three incidence rate" [408]. Unfortunately, there is no exact conversion between above-the-fold incidence and top-k incidence rates because SERP components in a post- "ten blue links" era have highly variable lengths. Instead, above-the-fold incidence provides a measurement approach that accounts for these variable lengths. Future work might consider more specific spatial incidence rates that take into account empirical data about click patterns for important search contexts (for instance, consider only sections of the SERP that receive on average a large number of clicks).

5.4. Results

5.4.1. Full-page incidence rates

We begin by reporting how often Wikipedia appeared in our (first page) SERPs. Fig. 5.2 shows full-page incidence rates across all combinations of devices, search engines, and query categories. As described above, we considered two devices, three search engines, and 432 total queries, so in total we collected 2592 SERPs (2 devices x 3 search engines x 432 queries). Below, we report all our results broken down by our three query categories.

Looking at desktop full-page incidence across categories rates (the left half of Fig. 5.2), we see that Wikipedia links were present in many *common* and *trending* SERPs but appear much less frequently in *medical* SERPs. Specifically, across search engines, Wikipedia appears in 80-84% of *common* SERPs and 67-72% of *trending SERPs.* However, for *medical* desktop queries, Wikipedia only appears in 16% of Google SERPs, 24%

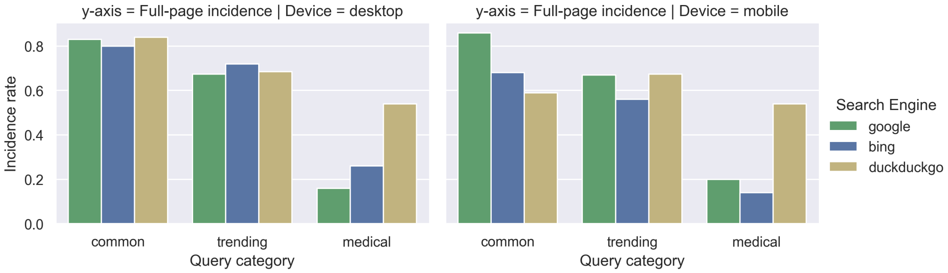


Figure 5.2. Full-page incidence rates for Wikipedia links. Left plot shows desktop results and right plot shows mobile results. Results are faceted by query category (along the x-axis) and search engine (color).

of Bing SERPs, and 54% of DuckDuckGo SERPs. Looking at the Google results, the high incidence for *common* and *trending* queries and low incidence for *medical* queries replicated the prior chapter's findings. Furthermore, for *common* and *trending* queries, Wikipedia's large incidence rate extends across search engines.

The results in Fig. 5.2 mean for the query categories we studied, the only major difference across search engines in Wikipedia's full-page incidence rate for desktop SERPs was for *medical* queries: DuckDuckGo shows many more Wikipedia links for these queries.

Comparing our desktop full-page incidence rates to mobile incidence rates (the right half of Fig. 5.2), we see similar results. For Google, the largest difference in mobile and desktop full-page incidence rates is 0.04 (for the *medical* category). Bing's mobile vs. desktop differences are slightly larger (0.12-0.16). Finally, DuckDuckGo shows the largest difference for mobile results: for *common* queries, the incidence rate is 0.25 lower.

5.4.2. Spatial incidence rates

While the full-page incidence rates presented above provide new insight into how Wikipedia helps to serve search queries across different search engines and devices, these results do not provide insight into where Wikipedia is appearing. For this, we turn to our spatial measurements: above-the-fold, left-hand, right-hand, and left-hand-above-the-fold incidence rates. These measurements give us crucial insight into how Wikipedia links appear in SERPs in a post- "ten blue links" world.

The first spatial measurements we look at are left-hand and right-hand incidence rates. These are shown in Fig. 5.3. Overall, right-hand incidence rates are higher than left-hand incidences rates and thus closer to full-page rates. For instance, Wikipedia's right-hand incidence rate for *trending* queries (shown in the right half of Fig. 5.3) ranges across search engines from 77-83%, very close to the full-page range of 80-84%. This suggests Knowledge Panel-style elements are a critical source of Wikipedia links in SERPs – but not the only source, as left-hand incidence rates are still substantial, e.g. 61-66% for *trending* queries.

Next, we look at our above-the-fold incidence rates. The middle ground above-the-fold rates are shown in Fig. 5.4, while the lower bounds and upper bounds (corresponding to smaller and larger viewports) are included in Table 5.4.2, which provides a summary of all our desktop incidence rates. In general, the lower bounds and upper bounds were relatively close to the middle ground estimate. A primary take-away is that for many cases, particularly for desktop SERPs, the *above-the-fold incidence rate* is only slightly lower than the *full-page incidence rate*. This means that not only is Wikipedia appearing frequently, but it is also appearing frequently in the most prominent area of our SERPs.

One exception to this trend is that Google and Bing have much lower above-the-fold incidence rates for mobile devices (shown in the right half of Fig. 5.4) than desktop devices, even when considering upper bound scenarios, i.e., even for mobile devices with large screens (lower and upper bounds for mobile above-the-fold incidence rates are shown in

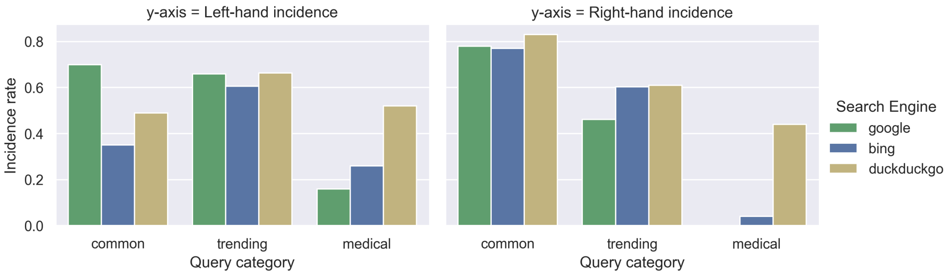


Figure 5.3. Left-page incidence rates (left plot) and right-hand incidence rates (right plot) for Wikipedia links.

Search Engine	Query Category	Full-page	Left hand	Right hand	Above-the-fold	Left-hand above-the-fold
google	common	0.83	0.7	0.78	0.80 (0.80 - 0.80)	0.05 (0.05 - 0.06)
google	medical	0.16	0.16	0	0.08 (0.08 - 0.12)	0.08 (0.08 - 0.12)
google	trending	0.67	0.66	0.46	0.54 (0.49 - 0.56)	0.33 (0.28 - 0.37)
bing	common	0.8	0.35	0.77	0.76 (0.76 - 0.76)	0.01 (0.01 - 0.03)
bing	medical	0.26	0.26	0.04	0.08 (0.08 - 0.08)	0.08 (0.08 - 0.08)
bing	trending	0.72	0.61	0.6	0.60 (0.59 - 0.62)	0.22 (0.18 - 0.26)
duckduckgo	common	0.84	0.49	0.83	0.83 (0.83 - 0.83)	0.14 (0.08 - 0.19)
duckduckgo	medical	0.54	0.52	0.44	0.44 (0.44 - 0.44)	0.18 (0.18 - 0.20)
duckduckgo	trending	0.68	0.66	0.61	0.64 (0.63 - 0.64)	0.45 (0.40 - 0.48)

Table 5.1. Desktop Wikipedia incidence rates for each search engine and query category.

Table 5.4.2). However, these incidence rates are still very high (over 20% for some cases). For context, a 20% incidence rate is likely highly desirable for many websites. Given the growing use of search engines from mobile devices, differences between desktop and

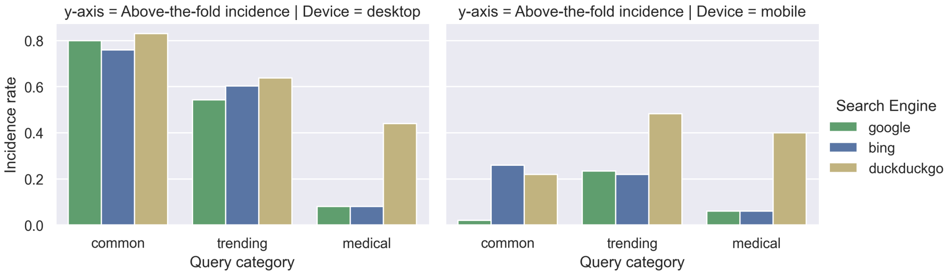


Figure 5.4. “Middle ground” above-the-fold incidence rates for Wikipedia links.

Search Engine	Query Category	Full-page incidence	Above-the-fold incidence (lower bound - upper bound)
google	common	0.86	0.02 (0.02 - 0.04)
google	medical	0.2	0.06 (0.06 - 0.08)
google	trending	0.67	0.23 (0.21 - 0.26)
bing	common	0.68	0.26 (0.22 - 0.30)
bing	medical	0.14	0.06 (0.06 - 0.06)
bing	trending	0.56	0.22 (0.21 - 0.24)
duckduckgo	common	0.59	0.22 (0.17 - 0.24)
duckduckgo	medical	0.54	0.40 (0.40 - 0.40)
duckduckgo	trending	0.67	0.47 (0.47 - 0.52)

Table 5.2. Mobile Wikipedia incidence rates for each search engine and query category.

mobile SERPs are important to consider in understanding the importance of Wikipedia to SERPs. As we will discuss below, the mobile use case is made even more complicated by features like “Siri Suggestions”, which loads website suggestions as a user types search queries in Safari (before the query is sent to Google or other search engines). This may be one way in which Wikipedia-addressable information needs are siphoned away from search engines before they get a chance to accept mobile user queries.

To fully understand the implications of Wikipedia's above-the-fold incidence rates, it is valuable to additionally consider left-hand-above-the-fold incidence rates. As described above, this incidence rate allows us to identify if Wikipedia is only appearing above the fold because it appears in Knowledge Panels. Looking at these incidence rates, shown in Table 5.4.2, we see they are substantially lower than above-the-fold rates. For *common* and *medical* queries, no left-hand-above-the-fold incidence rate is above 22%, and even for trending queries the largest rate is 45%. This result suggests that Wikipedia is primarily appearing above the fold because of the Knowledge Panel, especially for *common* queries. Only for *trending* queries does Wikipedia appear frequently above the fold as a "blue link" in the left column of our SERPs. This also provides a potential explanation for the lower above-the-fold incidence rates on mobile devices, which do not have a right column to highlight Knowledge Panel elements.

5.5. Discussion

Our results above suggest that the important role Wikipedia has been observed to play in serving search queries extends beyond just Google, beyond just the "10 blue links", and beyond just desktop search, although the effect is smaller on mobile devices. Indeed, Wikipedia articles appear very frequently across search engines, and they often appear "above-the-fold", driven in part by Knowledge Panel SERP components.. Below, we discuss the implications of Wikipedia's prominent role in search results. We additionally discuss how this prominence relates to search intents. Finally, we revisit the limitations and caveats attached to a targeted study like the one we present here.

5.5.1. Wikipedia's Volunteer-created Content has Impact Outside Wikipedia

Our results reinforce an idea with important implications for Wikipedia editors and those who research Wikipedia: Wikipedia content has a huge impact well beyond the wikipedia.org website. More specifically, our results highlight that the properties of Wikipedia content will also define the effectiveness of web search, at least in many important domains. This means that the large body of research that has sought to understand or improve Wikipedia likely has broader implications that has been widely understood thus far.

A particularly significant implication is that the biases of Wikipedia content will impact search results. A large literature has sought to understand the biases of Wikipedia content (e.g. [171, 194, 285, 316, 413]). Although we did not specifically analyze content biases in our data, the fact that we observed large incidence rates for Wikipedia in general suggest that these biases are likely reflected in web search results. It is likely the case that for queries about content that is underrepresented on Wikipedia (e.g., articles about women), there will be less high-quality Wikipedia content available to answer such queries. In other words, gaps in Wikipedia's coverage or quality could lead to gaps in search engines' ability to address user information needs. This is a critical area for future work, which might explicitly how specific biases (e.g., gender, language, content about rural areas, etc.) extend to search technologies and if there are opportunities for search engines to help address biases and gaps.

There are many active efforts aimed at filling in Wikipedia's knowledge gaps (e.g. [193, 318]). Because these gaps may be hurting search engines, our results suggest that search engines may have a substantial incentive to help address these gaps. At minimum, search engines should continue to attribute Wikipedia content and avoid cutting off Wikipedia from direct traffic. Looking forward, identifying opportunities to leverage the shared interests of search engines and the Wikipedia community could be fruitful for continuing to improve Wikipedia's coverage. These opportunities could include organizing events (e.g., an "edit-a-thon"), making design changes to show search users how to become Wikipedia editors, or other forms of engagement and support. For instance, if a user searches for information about a topic that is not well covered by Wikipedia, search engines could display a SERP element that indicates that the user can add information they find from their search to Wikipedia. Of course, any intervention will need to happen in dialogue with the Wikipedia community, as there are many ways in which design changes or other interventions could work against their intended outcome (e.g. if a design change leads to increased vandalism of Wikipedia, which might be particularly likely for underrepresented topics). These are issues that would be best handled through open discussion and careful consideration of existing Wikipedia practices.

5.5.2. For-profit Intelligent Technologies and Volunteer Content

Another important implication of the results above is that they highlight the essential role that volunteer-created content plays in fueling highly profitable intelligent technologies. Viewed through this lens, our results suggest that Wikipedia is providing critical content that helps search engines to succeed at their missions. In line with prior work [265]), it appears search engines would provide worse services in a world without Wikipedia. Alternatively, perhaps search engines would pay people to create a similar resource, potentially reducing the profitably of the industry. While users who primarily search from mobile devices might be less affected in a world without Wikipedia, a huge number of search results would be worse, and even mobile searchers would likely see worse search results (recall that although incidence rates were lower on mobile, they are still substantial). Relatedly, the fact that Wikipedia appears frequently in DuckDuckGo's search results suggests Wikipedia may be particularly valuable as a search result in a more privacy-focused world (DuckDuckGo does not collect any personal information).

Viewed through the lens that creating content like Wikipedia articles is a form of volunteer labor, Wikipedia can be understood as providing a "subsidy" of free labor to search engines [265]. However, it is also important to note that search engines provide a critical stream of traffic (both readers and potential editors) to Wikipedia [265]. This means that Wikipedia itself certainly sees benefits from its relationships with search engines and similar technologies.

That said, our results support the argument that the donations made by search engine companies to Wikipedia (e.g., [298]) likely represent only a tiny fraction of the value the Wikipedia community's labor has created for these companies. Additionally, given the Wikipedia community's important role in helping search engines serve their users, there is also an argument that the Wikipedia community should have agency in how its content is used, e.g. in discussions about how prominent Wikipedia links should be in knowledge panels [385]. An immediate concrete implication of this argument is that search engines should continue to attribute Wikipedia prominently, and when using Wikipedia content in SERP elements they should make hyperlinks easy to find. While we cannot be completely sure that search engines are always attributing Wikipedia when they present

Wikipedia content, our high incidence rate results suggest that search engines are attributing Wikipedia frequently. Looking to the future, designers of search engines might seek input from communities like Wikipedia when designing new SERP elements.

More generally, our results also provide an important data point for the growing discussion about the power dynamics between tech companies and the public, and the role of data in these dynamics. Scholars [13, 309] and the media [180, 307] have recently expressed interest in identifying and highlighting the dependencies that large tech companies have on data coming from the public with the goal of changing the power dynamics between the public and these companies. This might mean creating new economic relationships (e.g. getting paid for data in some manner [307]) or using these dependencies for data leverage in order to change tech company behavior. Our results highlight an underappreciated dependency, and a large one at that: peer-produced data. Moreover, the dependency in this study is of a different type than is often considered in these discussions. "Data leverage" has typically been considered in terms of personal data (e.g., data about users) [180, 307] and interaction data (e.g., "trace" data that is generated as people browse websites and apps) [229, 309, 407]. However, our results reinforce the importance of considering data from user-generated content communities as well.

The Wikipedia community itself is unlikely to engage in direct action against search engines. The manner in which search engines use Wikipedia content is legal, and, assuming Wikipedia is attributed (as it was in our results), in line with the stated goals of Wikipedia. Therefore, it does not seem likely that Wikipedia itself would engage in extreme forms of protest against search engines (e.g., blocking search engine web crawlers or intentionally vandalizing content in protest). Instead, our results suggest that a primary role of Wikipedia in discussions around data leverage is as a key example case in which the value of the public's data contributions to intelligent technologies, computing, and related fields is measurable. Ideally, we would also measure the value of the public's contributions to private datasets (like implicit feedback used to rank search results), but there are strong incentives in place that make this unlikely (e.g. many valuable datasets are proprietary and contain personally identifiable information [18]).

The general approach of studying peer production and commons datasets to measure the value of data contributions can apply to other platforms (e.g., OpenStreetMap, citizen science platforms), and future work along these lines will be highly valuable to inform discussion around data leverage and power dynamics between users and tech companies. Additionally, political action by Wikipedia is not completely out of the question. Wikipedia engaged in a "blackout" by temporarily shutting down English Wikipedia to protest the Stop Online Piracy Act [91], and so there remains a possibility that extreme situations (e.g. if new search engines features create major harms to Wikipedia) could lead to similar actions.

It is also important to note that the results in this chapter likely barely scratch the surface in terms of the ways that search engine companies rely on Wikipedia. In addition to user-facing links, search engines also use Wikipedia as foundational elements of their "knowledge graphs" and all capabilities that are dependent on their knowledge graphs [366]. Wikipedia is also used in machine translation, another service often baked into search engine results (e.g. [233]) along with other applications (e.g. [213]).

5.5.3. Wikipedia in SERPs and User Intent

As mentioned above, it is useful to consider our results in light of the user intent that may be associated with our query categories. While our *common* queries can be seen as *navigational* (e.g., “facebook”, “youtube”,), Wikipedia links are appearing for a huge number of these queries. This means even for users making these queries with navigational intent, they are still being exposed to relevant Wikipedia content. For instance, a user searching “facebook” with the intent to visit Facebook will be still be briefly exposed to Wikipedia text displayed in the Knowledge Box. This type of navigational exposure to Wikipedia content may be an interesting aspect of search engine-Wikipedia relationships for future study. In particular, how does Wikipedia’s prominence affect the experience of people using search navigationally? Do users trying to login to Facebook end up reading about Facebook on Wikipedia?

Relatedly, given that both *trending* queries and *medical* queries are predominantly *informational* queries, the substantially lower incidence rates for *medical* queries compared to *trending* queries is striking. It seems Google and Bing in particular may have made design choices to highlight non-Wikipedia sources of medical information. Examining SERPs for our medical queries (which we note, were taken from a medical query dataset released by Bing itself), we see both Google and Bing present a special knowledge box SERP element for these queries. This suggests that search engine designers have implemented a feature to treat these medical queries differently than other queries and doing so impacts Wikipedia’s incidence rate. Of course, it could be the case that differences in results are being driven by implicit feedback from search users, and not by platform designers.

5.5.4. Limitations and Future Work

It is important to emphasize the limitations of our targeted search audit. First, the choice of queries heavily impacts results such as incidence rates. For instance, it easy to construct a query set for which Wikipedia will never appear or for which Wikipedia will always appear (for instance, if we append the word "Wikipedia" to each of our queries, we can easily achieve a 100% Wikipedia incidence rate). We addressed this limitation by selecting queries that are made frequently and can be have large impacts on users (e.g. trending queries inform users about events in the world and medical queries may influence health decisions). Nonetheless, any incidence rate is conditional on query choice and investigating other query contexts is an important area of future work.

Although we considered (some) variation in device, search engine, and query category, there are other factors that impact SERPs. Most notably, search engines may have huge differences across languages and countries. This is a particularly ripe area for further study of the relationship between Wikipedia and search engines. Wikipedia covers a large number of languages, and future work could identify opportunities and pitfalls for the relationship between non-English Wikipedia and search engines. We note that although our trending queries did include some Spanish queries and corresponding Spanish es.wikipedia.org results, the Wikipedia links in our study were primarily English Wikipedia links.

Finally, future work should consider systems that modify the search process, e.g., "Siri Suggestions" or other "AI assistants". For instance, we observed "Siri Suggestions" that send users directly to Wikipedia pages without visiting SERPs several times during our confidence checks from a mobile device. These AI assistants may bypass search engines entirely, or substantially reduce the number of results a user receives. For instance, users who interact with AI assistants using a voice assistant may only get a “top 1” result, instead of 10 blue links or spatially situated SERPs. Assuming that many voice queries will tend to be more informational than navigational, it could be the case that Wikipedia’s “incidence rate” for voice search and Siri Suggestions will also be very high. Future auditing work might specifically investigate voice search and AI assistants. In cases that involve a single, or few results, auditing these systems may actually be substantially simpler than auditing complex SERPs, although automating audits may be challenging.

5.6. Conclusion

In this chapter, we reported on a targeted investigation of how often—and where—Wikipedia links appear within SERPs for three major search engines and across desktop and mobile. By considering the precise location of links within the SERPs, we shed light on how often Wikipedia content appears “above the fold” and in “Knowledge Panel” elements that exist outside the typical ranked list of search results. We find evidence that Wikipedia’s volunteer created content is important to all three search engines we studied, although the magnitude of this importance is heavily context dependent, with results varying across devices and types of queries.

CHAPTER 6

Data Strikes

6.1. Introduction

Large technology companies are facing a growing wave of public criticism[1]. Just in the last year, these companies have been condemned for a wide range of practices, including those related to privacy [15, 199], harassment [337], addiction [280], effects on democracy [216], and automation [377]. The breadth and scale of the public concerns about tech companies has even led to popularization of the term "Big Tech" [291], an adaptation of the terms "Big Oil" and "Big Tobacco" [97].

The same companies that anger the public, however, often are also particularly *dependent on the public*. Specifically, in addition to needing users and customers to generate revenue, *these companies rely on the public's data to power core functionality*. For example, Google requires user clicks to train its ranking algorithm [314]. Similarly, business-critical recommender systems employed by companies like Amazon and Netflix require large behavioral datasets from users (i.e. clicks and views) [135, 370].

Seen through the lens of the public's concerns about tech companies, companies' dependence on user data can be understood as a potentially powerful source of leverage for the public. To help the public action this leverage, several authors have proposed the notion of "**data strikes**" (e.g. [13, 241, 309, 90]), in which users halt their "**data labor**"

[1] This chapter was originally published as [407].

[309]. The basic logic that motivates data strikes is straightforward: If users withhold their data labor from a tech company, some of the company's essential services will suffer, and this would then force the company to make concessions that are desired by the public. These concessions could range from improved privacy policies to profit sharing [13, 141].

Despite the growing discussion around data strikes, little is known about how this new type of collective action would work or about how data strikes relate to standard forms of collective action like **traditional boycotts** (a type of collective action classified as political consumption [210, 282]). Additionally, there is little empirical information about how effective data strikes could be, let alone the data strike configurations that would be most effective. Activists seeking to organize a data strike have no guidance regarding the number of users that would need to join them, the kinds of services most vulnerable, the types of users that would allow them to be most successful, or even whether strikes can be successful at all. Similarly, platform operators are not aware of the potential damage that could be inflicted through data strikes.

This chapter seeks to improve our basic understanding of data strikes and provide much-needed empirical information about their effectiveness. We first situate data strikes in relationship to traditional boycotts, in which a user stops patronizing a company entirely. Through the introduction of a lightweight framework that describes **collective action** in a technology company context, we highlight that almost any traditional boycott against a technology company will implicitly also include a data strike, but that data strikes can also occur independently from boycotts. For example, a consumer who continues to purchase products from an online retailer could engage in a data strike by using private browsing windows and not providing product ratings.

Next, focusing on the domain of recommender systems, a critical driver of revenue for technology companies [135, 358], we introduce a novel evaluation procedure for understanding collective action campaigns against technology companies. Our procedure uses a metric called **surfaced hits** that independently measures the effects of a traditional boycott and a data strike. Leveraging surfaced hits, we examine how model performance changes depending on 1) the size of the participating group, 2) whether the participating group is a random group of users, or a homogeneous group of users who share some characteristics (e.g. women, people interested in documentaries), and 3) whether or not the group is conducting an independent data strike or doing so as part of an implicit part of a traditional boycott.

Our results confirm that users' **data labor power** – which is unique to online platforms and is manifest in a data strike – provides users with a new source of leverage in their relationships with technology companies. For small recommenders and in specific product spaces, this added leverage can be substantial. A moderately-sized data strike alone – even when not part of a traditional boycott – can significantly reduce the benefits of a recommender system. Indeed, for moderately-sized data strikes, we observe recommender accuracy decreasing to the levels that defined the state-of-the-art in recommender algorithms in 1999.

Additionally, our work shows that data strikes that occur as part of a traditional boycott add data labor power to the standard **consumer power** from a boycott, meaningfully increasing the overall power of the political consumption campaign. In effect, we see that the data strike component of a traditional boycott against a technology company can diminish the desirability and effectiveness of that company for all remaining customers who are not yet part of the boycott. This is analogous to an offline boycott of a specific product (e.g. a pizza brand) that also somehow managed to diminish the quality of the product itself (e.g. pizza quality), disincentivizing the purchase of that product for people who are not participating in the boycott.

Finally, our work also highlights that data strikes that are not part of a traditional boycott represent a *fundamentally new type of collective action*, one in which the *barrier to entry is much lower than in a boycott.* Most notably, we observe that data strike participants can substantially reduce the utility of a recommender system without sacrificing access to the underlying products and services. Given that participation in political consumption activities has been traditionally limited by factors like affluence and political stability [210], the demonstrated effectiveness of data strikes could substantially broaden participation (e.g. users who cannot afford to use expensive alternatives to online platforms can still strike). This is analogous to an offline boycott in which a user who cannot afford expensive pizza could somehow still take action against the local low-price pizza chain while continuing to buy their products. This notion of low barrier-to-entry protest action echoes early research on digital activism by Earl and Kimport, who argued that the web reduces costs for participating in protest behavior like boycotts, petitions, and email campaigns [88].

Below, we adopt a standard structure to motivate, explain, and expand on our findings. We first cover related work, then discuss methods, followed by results. We close with a discussion of the issues identified in our results and by highlighting limitations.

6.2. Related Work

In this section, we describe how this research draws motivation from four areas in particular: the growing discussion related to data strikes, research on the relationships between tech companies and volunteer-created content, work that generally seeks to quantify the financial value of user data, and work that looks specifically at ways to manipulate recommender system outputs.

6.2.1. Data Strikes

This research was most directly motivated by growing calls for collective action campaigns that force changes in technology platforms by leveraging the value of user data to these platforms [13, 163, 228, 309, 367, 379]. These growing calls use different, potentially conflicting framings of *data as capital* or *data as labor* [13]. With regard to the former (data as capital), collective action is framed in terms of a boycott, in which users stop their consumptive activities (e.g. purchasing products through a web platform, using a social media platform) which in turn prevents the flow of their capital (data, related revenue like advertising revenue) to the platform. This framing is exemplified by very recent boycotts put into practice against Facebook. The data labor view suggests that data "unions" should protect the interests of those who produce data (i.e. users) [13, 309, 379]. As traditional labor unions have implemented (and threatened) strikes to gain leverage when collectively bargaining, Lanier and others [163, 379] have written that data unions might similarly engage in a "strike". These authors point out that users can leverage their data in ways that resemble both traditional boycotts and strikes.

These conflicting understandings of collective action that use data leverage needed to be resolved in order to make campaigns concrete enough to simulate. Below in the Framework section, we enumerate one possible resolution and use the corresponding framework to inform the design of our experiments.

A major regulatory change that is highly relevant to any data-related collective action campaign – whether it is informed by an capital or labor perspective – is the recent adoption of the General Data Protection Regulation by the European Union. Among a variety of regulations regarding personal data, the GDPR includes a provision ensuring the right to erasure. Barring special circumstances (e.g. data critical to public health research), individuals covered by the GDPR will have the right to request their personal data to be deleted. As such, the GDPR potentially empowers activists to engage in much larger data-related collective action than previously possible, in particular by erasing old data instead of just stopping the flow of new data. While it remains to be seen how often and how effectively the right to erasure will be used in practice, the inclusion of this right marks a large shift in regulatory practices towards data usage: the GDPR could trailblaze the way for similar or even stronger provisions by other regulatory bodies (e.g. California's State Government [235]).

While the GDPR may make data strikes immediately tractable and lay the groundwork for future regulation, regulation like the GDPR is not necessary for these types of collective action campaigns. Campaigns could be targeted at areas where there is *no existing data*, for instance reviews about a new television show, or location data used to predict traffic. Furthermore, very recent research suggests that simple tools like browser extensions may help web users successfully join web-based collective action campaigns with an extremely low cost to the user, which could help stop the flow of implicit behavioral data [241].

Finally, it is important to note that the social science literature can give some guidance as to how large one could expect campaigns to be. Data from Europe and the U.S. found that between 28% and 35% of consumers had engaged in an act of political consumption [210, 282], which includes either boycotting or "buycotting" (aligning one's purchases with a company that is perceived to align with one's political preferences).

6.2.2. Tech Companies and Volunteer-Created Content

A related area of research that also helped motivate this study is work which has sought to understand the dependence of technology companies on volunteer-created content like Wikipedia articles. McMahon et al. [265] showed that Google search effectiveness drops substantially when Wikipedia links are silently removed from search results. In Chapter 3, we discussed how Stack Overflow and Reddit receive substantial benefits from Wikipedia in the form of impactful links and references; they showed that these benefits come in the form of both increased engagement from users and advertising revenue. Although not originally intended as such, these studies can be seen as simulating a form of data strike in which companies are somehow prevented from using Wikipedia content. Such a collective action campaign would be highly unlikely given Wikipedia's licensing [93], but the design of these studies helped to inform our methodological approach outlined below.

6.2.3. Financial Value of Data

This work was more generally motivated by a broad body of literature seeking to understand the financial value of data and highlighting the importance of this understanding. This research includes efforts to provide individual users with transparency into the value they create, such as the Facebook Data Valuation Tool created by González Cabañas et al. [136], as well as efforts to broadly understand the value of that data at a macroeconomic scale (e.g. the value of Wikipedia to GDP statistics) [40]. On the policy front, the World Economic Forum has identified data as a new “asset class” and suggests that thinking about data economics demands a new understanding of the personal data ecosystem [347].

6.2.4. Recommender System Manipulation

Within the recommender system literature, there has been research into various ways recommender systems might be manipulated. For instance, prior work has examined how recommender systems might be “shilled”, i.e. misled so as to promote a particular product (e.g. [54, 224]). Like potential strikes, shilling attacks are an adversarial approach to manipulating the outputs of a recommender. As we explain below, our experiments specifically focus on campaigns which withhold data, so findings related to shilling are not directly applicable. However, in practice collective action participants might be able to adopt techniques from shilling, in which case this body of literature may be useful to both users and the companies they seek to gain leverage against.

Recent research from Wen et al. explored recommender performance under conditions in which users filter some portion of their preference history to increase privacy and/or recommender accuracy [420]. Specifically, the authors found that users can filter out some of their preference history from the recommender while maintaining, or even improving performance for implicit recommendations. This research has direct implications for data-related collective action participants: although users might be able to perform a "partial strike" by deleting some of their data without suffering decreased performance, these partial strikes are unlikely to be very effective (as in some cases, partial strikes by all users improved population-level performance). In our experiments, we focus on directly simulating conditions where users withhold *all* their data, and we also explicitly consider both data strikes and traditional boycotts. However, exploring the interplay between boycotts and data-filtering tools will be an important area of future research that is mutually beneficial to both problem contexts. In particular, the data filtering interfaces described by Wen et al. could be another outlet for users to actuate data strikes.

6.3. Framework

As mentioned above, although collective action campaigns that use data leverage against technology companies have been discussed as a theoretical possibility, they have not been carefully characterized. Indeed, in the context of collective action against technology companies, the distinction between data strikes, boycotts, and combinations of the two can be unclear.

In order to simulate data strikes, we need to first concretely define data strikes and their relationship to traditional boycotts. To do so, we turned to the divergent theoretical underpinnings of the boycott and strike terms. In a boycott, participants are *consumers* who cut off the flow of an asset (e.g. money from purchases) to a firm. In a strike, participants are *laborers* who stop performing labor for the firm. Users of an online platform can therefore boycott the platform by refusing to use the platform as consumers (e.g. not buying from an e-commerce site, not visiting a news site or video site, etc.) and strike against the platform by refusing to provide data (e.g. deleting data, preventing the flow of new data by using Private Browsing features, or other privacy techniques like ad blockers or Mozilla's new Facebook Container [284]).

In many cases, *boycotts against tech companies implicitly include a data strike.* For instance, users of a video platform like YouTube who boycott (refuse to visit the website) are also implicitly conducting a data strike by cutting off the flow of behavioral data like views, likes, and comments. However, *it is often possible to participate in a data strike without boycotting the platform.* This occurs if someone continues to access a website but withholds data using privacy-preserving techniques (e.g. private browsing), leverages data management options made possible through data protection regulation, or – critically for our context – *refuses to comment on products, rate products, or review products.*

To put the nuanced relationship between boycotts and data strikes as defined above into better context, we consulted the literature to identify the various specific means by which these types of collective action campaigns could affect company revenue (e.g. [13, 309]). We identified four such pathways to revenue impacts:

- The *direct data labor effects* (e.g. algorithmic performance decreases leading to loss in sales)
- The *indirect data labor effects* (e.g. users quit using the platform because performance goes below some threshold leading to loss in sales) [372].

- The *direct consumer effects* (e.g. people stop buying products or viewing ads) [210, 282].
- The *indirect consumer effects* (e.g. a large number of customers quit, so business loses economy of scale advantages) [356].

The *consumer effects* above are those that exist in traditional boycotts and have been felt by targeted businesses since well before intelligent technologies came into common use. The *data labor effects* are specific to collective action campaigns against companies that use data-hungry intelligent technologies. While a traditional boycott only includes direct and indirect consumer effects, a simultaneous data strike and boycott includes all four of the above components.

This simple framework, depicted in Fig. 6.1, provides a much-needed lens into the ontology of these types of collective action campaigns, but it also highlights that researchers – especially those outside a technology platform – can only simulate a portion of the effects of data strikes and boycotts. For instance, without exact sales numbers, both the direct and indirect sales effects are very difficult to study externally. Similarly, the indirect long-term effects of reduced recommender performance would be difficult to capture both for researchers external and internal to a platform (although there is at least one case of this being done internally in the search literature using an A/B test framework [372]).

Fortunately, in this chapter, we show by focusing on the direct data labor effect while still considering the direct consumer effect, we can still learn a great deal about collective action campaigns that use data leverage. As we discuss below, our results support the effectiveness of data strikes, suggesting the unique relationship between users and tech companies can empower users beyond what would be the case in the offline world (analogous to a situation where customers boycotting a pizza company can lower the pizza quality). We also discuss how our results can be interpreted as a *lower bound* for the effects of data strikes and boycotts against tech companies because we cannot measure indirect effects.

6.4. Methods

In this section, we first describe aspects of our methods that were consistent across all our experimental configurations. We then describe the two broad types of collective action campaigns we simulated: "general" groups comprised of randomly selected users, and "homogenous groups" of users who share characteristic (e.g. power users, fans of comedy movies).

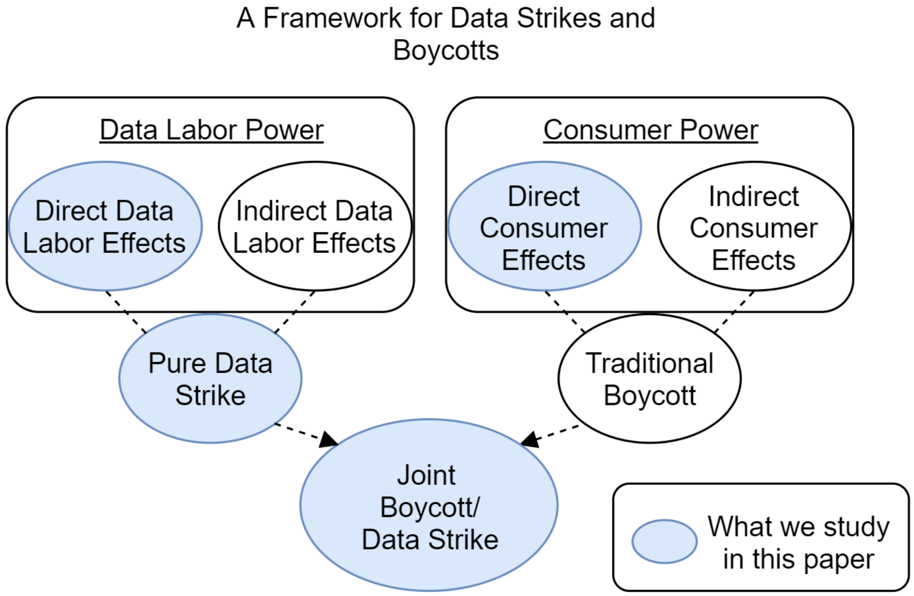

A Framework for Data Strikes and Boycotts
Data Labor Power
Direct Data Labor Effects
Indirect Data Labor Effects
Consumer Power
Direct Consumer Effects
Indirect Consumer Effects
Pure Data Strike
Traditional Boycott
Joint Boycott/ Data Strike
What we study in this paper

Figure 6.1. Graphical summary of our framework for defining data strikes and boycotts. Blue indicates aspects we focus on in this chapter.

6.4.1. Design of Experiments

In each experiment, we evaluate recommender systems under a variety of simulated data strike and boycott conditions, i.e. some users' ratings are not available for model training, and thus recommender performance changes. While the campaigning groups differ by experiment, the basic methods are the same.

6.4.1.1. Datasets. Our primary dataset was MovieLens-1M (ML-1M), which consists of 1 million "1 to 5 star" ratings for 3706 movies, provided by 6,040 users with self-reported demographic data [158]. The MovieLens datasets are hugely influential datasets that have been central to recommender system research for decades [158]. To better understand performance against large recommenders, we additionally performed experiments with the much newer MovieLens-20M (ML-20M) dataset [158], which contains 20x more ratings but no demographic data about users.

6.4.1.2. Algorithm Choice and Implementation. . For each experiment, we focus on the well-known and high-performing Singular Value Decomposition recommender algorithm (SVD) [116]. As validation, we also performed a smaller set of ML-1M experiments with an older and mathematically distinct algorithm: item-based K-Nearest Neighbors (KNN) [338] adjusting for item and user baselines as described by Koren [211]. Both algorithms are implemented in the open-source Python library Surprise [179], which we extended for our experiments. All code used for our experiments and analysis is available for replication and extension on GitHub[2]. We validated the accuracy of the SVD implementation by ensuring results were comparable to published results on the ML-1M

[2] `https://github.com/nickmvincent/surprise_sandbox`

dataset [178, 214, 232, 335]. These successful comparisons are summarized in the linked GitHub repository.

6.4.1.3. evaluation Procedure. : We evaluate the recommender with five-fold cross-validation, so for each evaluation fold 20% of total data is available for testing, and each rating is tested in exactly one fold. While data that is held out because of a simulated campaign cannot be used for training, it can be used for testing. This means we can consider results from the perspective of striking users who will receive non-personalized recommendations (based on each movie's average rating). Furthermore, when simulating strikes where participants share some characteristics (which we call homogenous strikes), we also record results from the perspective of the users who have the same characteristic as participants, which we call *Similar Users.*

6.4.1.4. Metrics for Evaluating Strikes and Boycotts. . When evaluating accuracy of explicit rating predictions for recommender systems, one common approach is to measure the error in individual predictions, e.g. through an accuracy metric such as root-mean-squared error (RMSE) which was used for the well-known Netflix Prize [116] or using information retrieval metrics such as gain (NDCG) [45] or precision. Retrieval metrics have been favored in recent years because of the ecological validity [71, 205] for most real-world contexts (e.g. "Top Ten Movies for You").

However, these metrics only capture performance for users who receive recommendations, and do not capture the effects of users leaving due to boycotts (i.e. *consumer effects*). Therefore, this approach is not well-suited to understand boycotts from the perspective of the system-owner, because the loss in revenue from boycotting users will not be visible in traditional metrics. For instance, if we simulate an 80% boycott and measure the RMSE of predicted ratings for the remaining 20% of users, the change in RMSE does not account for the users who *left the system entirely.*

To understand the relationship between data strikes and boycotts, it is critical to capture both the consumer power of boycotting users and the labor power of striking users. To do so, we introduce a new metric, which we call *surfaced hits.* The metric measures the fraction of hits (defined as a rating of at least 4.0, as is common in prior work, e.g. [214]) across an entire group of users (perhaps all users, or non-boycotting users). The underlying assumption, that one hit corresponds to one unit of value for a recommender system, is supported by the widespread use of analytic metrics such as click-through-rate in online systems [135]. A perfect algorithm will surface all hits, and therefore have a surfaced hits of 1.0. This metric can be effectively viewed as a variant of precision that sums (rather than averages) across all users and sets individual thresholds for precision equal to how many positive ratings each user has.

As stated earlier, for both boycotts and data strikes participating users' ratings are withheld from all training data sets. Users who boycott contribute zero hits to the numerator of surfaced hits for their test ratings ($\mathrm{h_u} = 0$). In other words, surfaced hits is "penalized" by marking all positive ratings for the user as a non-hit. For users in a data strike, we include the user's test ratings in the calculation of surfaced-hits, but "penalize" via non-personalized recommendations (i.e. movie averages) because the user's training data is not available due to the strike.

This metric has three useful properties for studying data strikes and boycotts. First, when users boycott, surfaced hits is reduced proportionally to the number of positive ratings in the boycotting group. In other words, if enough users boycott to remove half of all the "hits" in the dataset, surfaced hits will reduced by at least half. Second, this metric also accounts for differences in user behavior: a user with 1000 hits in their rating history has 10x the impact of a user with 100 hits in their history. This captures the potentially disproportional economic value of more active users. Third, we can calculate surfaced hits for different subsets of users (e.g. all users, non-striking users, users similar to striking users) to understand the effects of collective action from different perspectives. We verified that these metrics perform very similarly to well-established list metrics including NDCG and precision while at the same time capturing the damage that occurs when boycotting users leave a system (see the end of Results and the code repository).

6.4.2. Campaign Configurations

6.4.2.1. General campaigns. . First, to get a general understanding of the relationship between campaign size and recommender system performance, we simulated a series of "general" campaigns with random user selection. The demographic make-up of the groups approximates the distribution of all users. We selected a sequence of 16 different group sizes ranging from 0.01% of users to 99% of users. For each group size, we randomly selected a group of that size to participate in the campaign. To reduce noise associated with different random configurations, we repeated each of these 16 experiments 250 times with a new random user sample for the ML-1M dataset and 40 times for the ML-20M dataset.

6.4.2.2. Homogeneous campaigns. . To explore how data strikes might affect specific groups of users, we simulated a series of homogeneous campaign groups using patterns in rating behavior and demographic information in the MovieLens-1M dataset. These experiments were motivated both by the potential for collective action campaigns organized by people who share some characteristics, as well as the potential to investigate the effect of data strikes within groups.

We created five types of homogenous campaigns. We created groups of “fans” for each movie genre by identifying all users who rated at least ten movies of that genre and have an average rating for the genre of four or higher. To simulate demographically-defined campaigns, we created groups based on the user-reported demographics, specifically male/female, age bracket, and occupation. We also simulated campaigns by contribution amount, creating boycotts for “power users,” defined as the top 10% of raters and “low frequency” users, the bottom 10% of raters.

For each of the five types of homogeneous groups, we simulated campaigns in which 50% of users matching a given group participated. For example, we simulated campaigns with groups such as 50% of all female users and 50% of all comedy movie fans. Importantly, 50% participation allows us to simulate what happens to similar users who do not participate. In other words, if some women participate in a data strike, what happens to women who do not? As we explain below, in addition to understanding the outcomes of this type of realistic campaign, the results of this analysis proved critical to identifying vulnerabilities to recommenders backed by large datasets. We viewed a 50% group participation as more realistic than full participation, given the distributed, non-hierarchical nature of most online communities.

Our homogenous experiments focus on the perspective of non-participating users (i.e. data *labor effects*). The consumer effects of a homogenous boycott scale with the size of the boycott as measured by the number of positive ratings, and as we show in our general campaign experiments, this effect is substantially larger than that of strikes but this does not negate a strike's intrinsic value.

For each homogenous campaign configuration, we performed experiments with 50 sampled groups. We also compare the observed campaign effects to the expected effects for a random campaign with the same number of ratings. In order to obtain a relatively simple estimate of the "expected" effect of a data strike of some size, we used a quadratic interpolation.

6.5. Results

In this section, we first describe the relationship we observed between recommender performance and campaign size in the case of general (random users) collective action campaigns, focusing on comparing pure data strikes to strikes coupled with traditional boycotts and examining the overall effectiveness of campaigns across datasets, Then, we describe the key findings from our homogeneous group experiments, focusing on the finding that unique homogenous groups, defined demographically or behaviorally can exert their data labor power to disproportionately affect similar users, indicating the potential for data strikes that target specific preference spaces to boost their effectiveness.

This section uses the surfaced hits metric, described above. We focus on the popular SVD algorithm because the item-based KNN algorithm behaved similarly in our initial experiments.

6.5.1. General Campaign Experiments

We begin by examining the effect of general data strikes and boycotts (i.e. with random users) from the perspective of the system owner (e.g. Google, Facebook, MovieLens

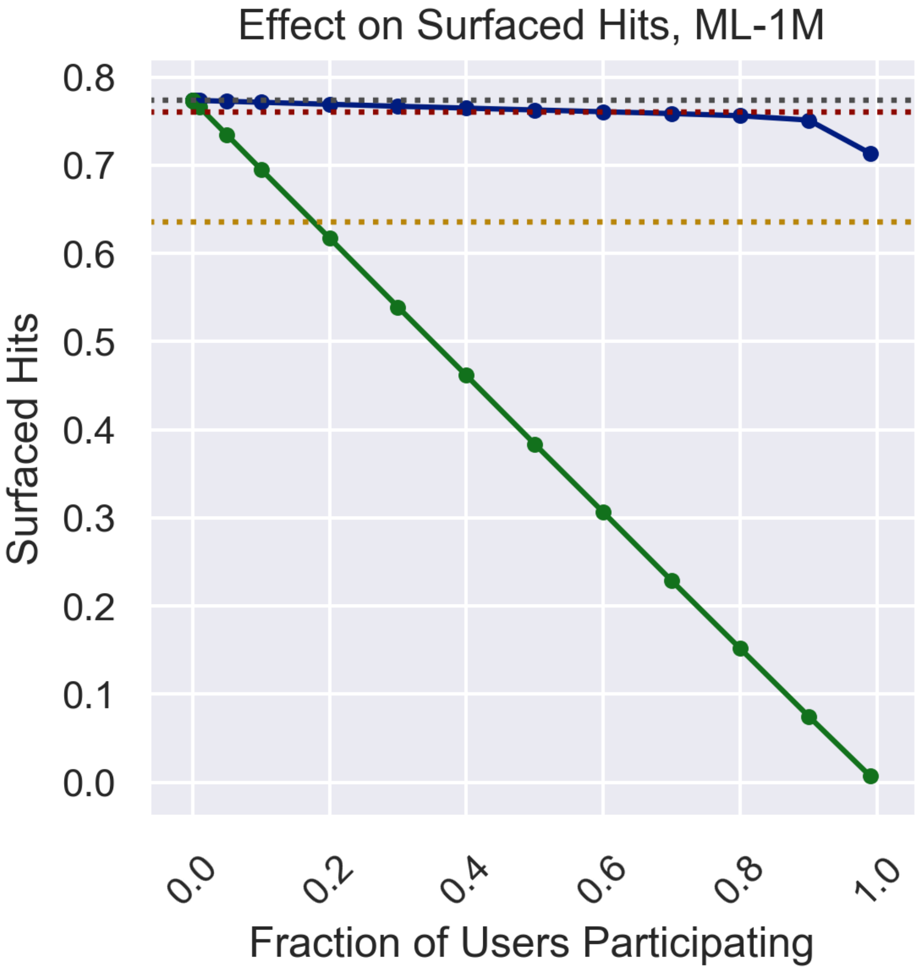


Figure 6.2. The relationship between campaign size and surfaced hits. Surfaced hits include all users (including strikers and boycotters), and therefore reflect the perspective of the system. Dotted horizontal lines provide comparisons: black (uppermost) shows fully personalized SVD, red shows un-personalized results, and gold shows random results. Blue points are data strike and green points are data boycott (see next figure for legend).

operators). Then, still focusing on the perspective of system owners, we specifically zoom in on performance changes relative to un-personalized results.

Fig. 6.2 shows the effect of data strikes (blue line) and joint data strikes/boycotts (green line) on surfaced hits (y-axis) across the system for both ML-1M (left column) and ML-20M (right column). As a reminder, a value of 1.0 would correspond to an algorithm that produces perfect ranked lists for every user.

These plots include dotted horizontal lines that provide important context: the black line shows performance of SVD with full access to the dataset (which gives 77.4% of hits), the red line shows the results of "MovieMean," which gives completely un-personalized ratings (movies are ranked in order their mean rating) and the gold line shows the results

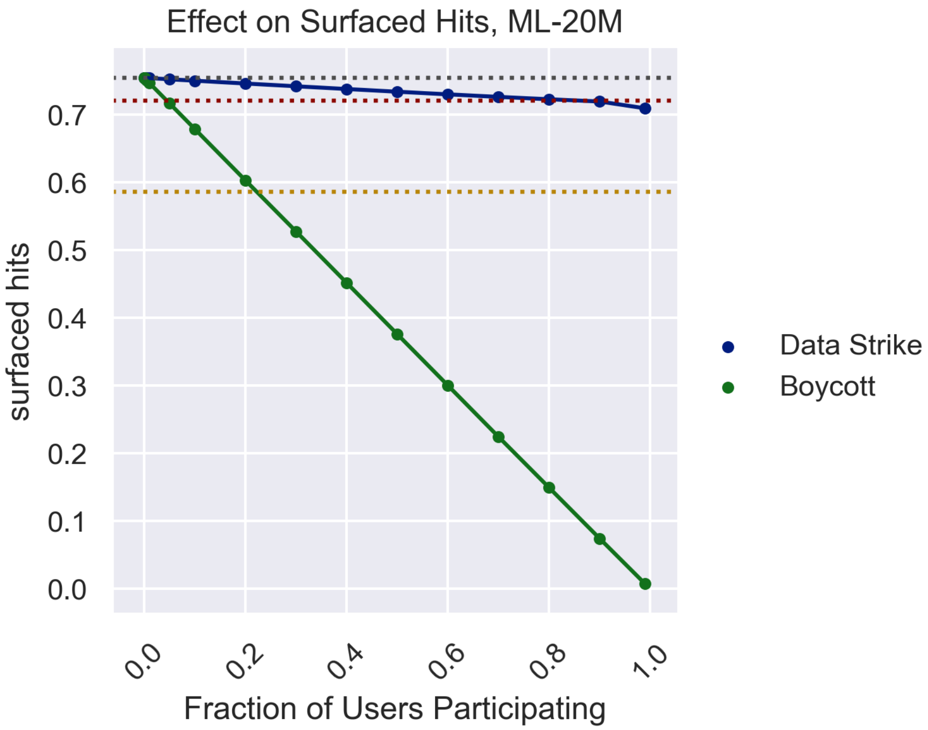


Figure 6.3. Same as previous figure, but for ML-20M

of completely randomly ranked lists (i.e. worst-case performance). Note the high number of hits associated with random lists: this is a result of MovieLens users' tendency to rate movies well, as well as a limitation of evaluating with explicit ratings. We address this limitation below by focusing on performance *relative to* un-personalized performance, i.e. we zoom in on performance change between the black and red lines.

The most significant trend in Fig. 6.2 is that boycotts are substantially more effective than data strikes. For instance, while a 30% boycott of ML-1M reduces hits to from 77.4% to 53.9%, a 30% data strike only reduces hits by 0.7% to 76.7% (ML-20M, right column, shows a similar trend). Furthermore, for both datasets a boycott of about 20% of users reduces surfaced hits to the amount expected for completely randomized recommendations. This result means that at first glance, the loss in hits caused by users who leave the system strongly outweighs the loss in hits from reduced algorithmic performance. Importantly, this finding does not invalidate the potential of data strikes, as we will describe below.

We note that the gap between un-personalized results (red line) and fully personalized results (black line) appears to be very small. This reflects the non-linear value of recommender algorithms; the small margin between un-personalized and personalized algorithms corresponds to a large amount of value for a recommender system operator. For example, for Netflix the small visual margin between un-personalized algorithms and personalized algorithms accounts for a 2-4x increase engagement with recommended items and $1B in revenue [135]. Thus, we now specifically focus on the change in performance relative to un-personalized results to inspect how data strikes leverage *data labor power* to lower recommendation performance towards un-personalized levels.

Fig. 6.4 zooms in on surfaced hits during a data strike. Again, the black horizontal line marks personalized performance and the red horizontal line marks un-personalized performance. Additionally, in Fig. 6.4 the horizontal cyan line shows the performance of

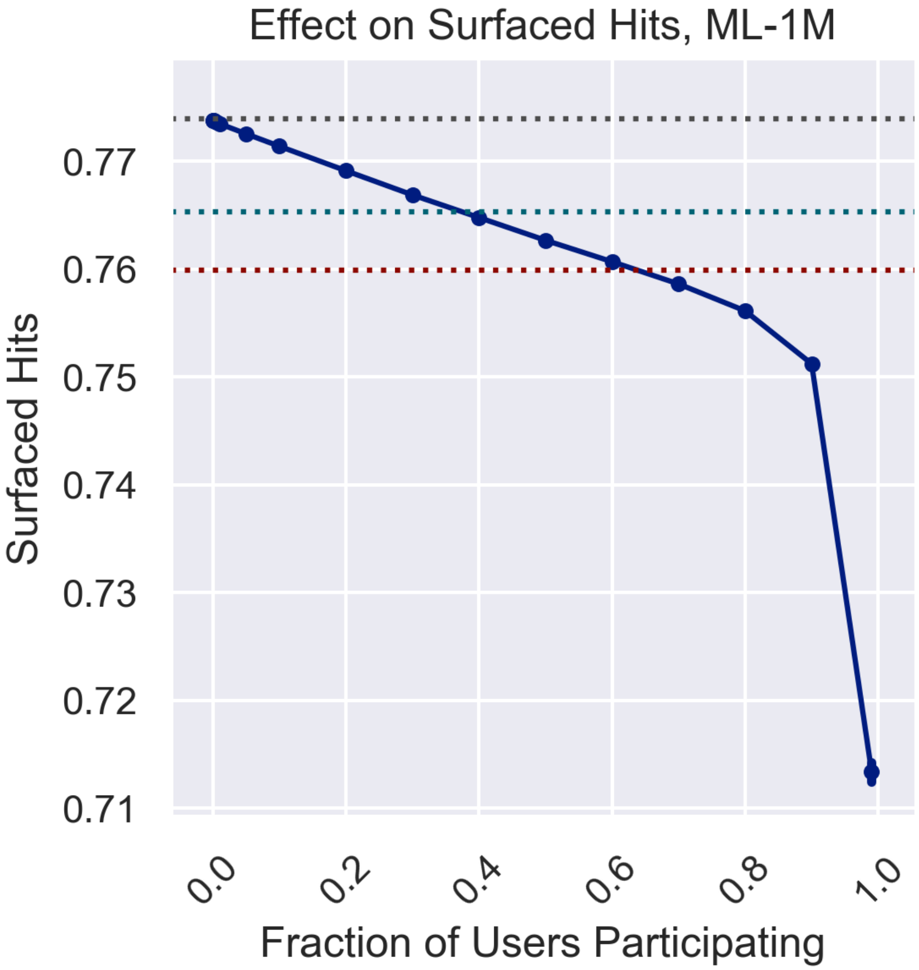


Figure 6.4. The relationship between data strike size and surfaced hits. Dotted horizontal lines provide comparisons: black shows fully personalized SVD, cyan shows item-based KNN (1999), and red shows un-personalized results.

simple item-based KNN (which we evaluated on with full access to each dataset), introduced in 1999 - this context shows the ability of campaigns to essentially set performance “back in time”.

Here, we see that data strikes are a source of potentially powerful leverage. Fig. 6.4 shows that campaigns had substantial effects on recommender performance for ML-1M

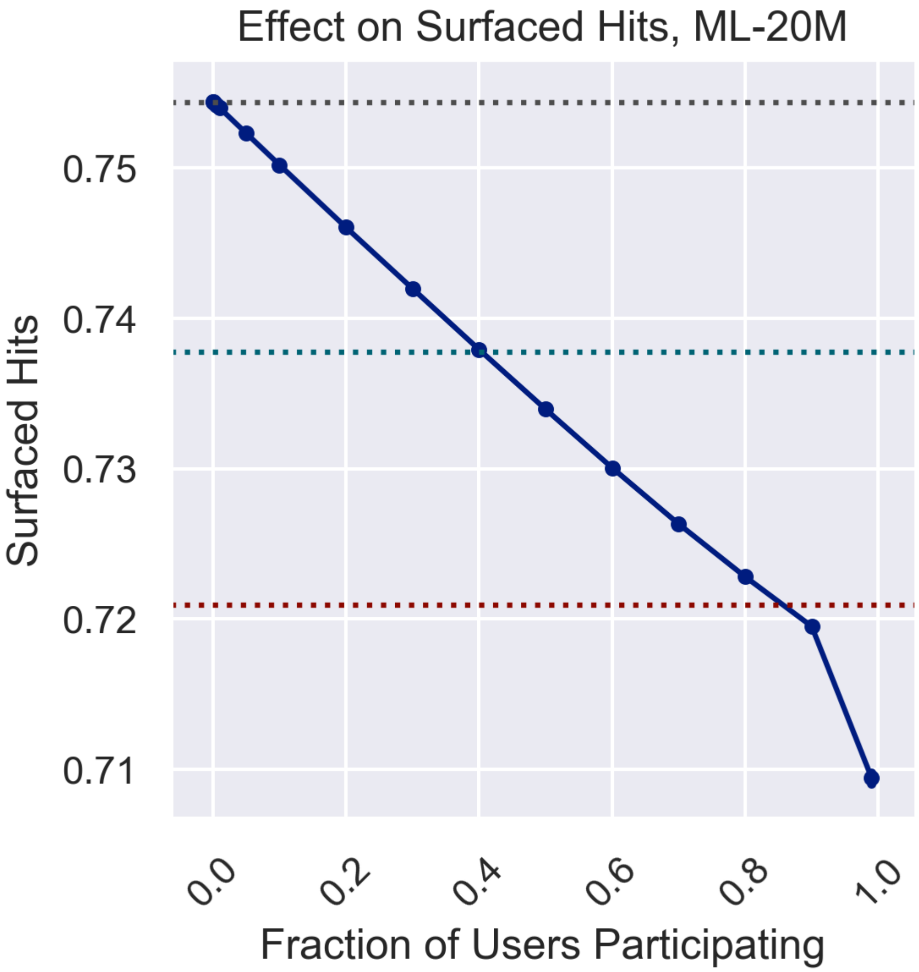


Figure 6.5. Same as 6.4, but for ML-20M.

relative to non-personalized ratings. Consider the performance degradation due to a campaign relative to non-personalized results: a strike with 30% of users (which is a realistic size based on research on political consumption; see Related Work) causes performance to degrade to 50.2% of non-personalized results. In other words, this mid-sized campaign would take the system half-way towards completely negating personalization. It is important to note one driver of this effect (although not the sole cause) is the fact that during a data strike, users continue receiving un-personalized recommendations.

These results illustrate the power of collective action to potentially negate decades of algorithmic advances, at least for recommenders using relatively small datasets (e.g. 1 million ratings). Looking again at MovieLens-1M, a strike by 37.5% of users can roll back hits to a level equivalent to the classic item-based collaborative filtering algorithm introduced in 1999 [338] (cyan dotted line in Fig. 6.4..

The ML-20M results (right column), however, suggest a more complicated story for recommenders using larger datasets: unsurprisingly, larger datasets are harder to strike against. When the dataset size is increased by a factor of 20, a strike by the same percentage of users becomes somewhat less effective. If 30% of MovieLens-1M strike (1800 users), we see a 50.3% reduction in the benefits of personalization, but if 30% of MovieLens-20M strike (41,400 users), it would only cause a 37.0% reduction. This means that even though 20 times as many users participated a strike to achieve 30% participation, the overall effect is smaller.

Finally, we reiterate that evidence from industry suggests that in commercial systems, a change in surfaced hits has a non-linear value to recommender operators. In other words, the visually-small performance change between the red and black horizontal lines in Figures 2 and 3 may have an out-sized effect on platform revenue. As mentioned before, the small surface-hits improvement due to personalization may correspond to a 2-4x increase engagement with recommended items in other contexts [135]. Similarly, in 2010 YouTube published findings that suggest recommendations add substantial value: almost 30% of video views came from their recommender system and the recommender was the main source of views for most videos [438]. An industry report from consulting company McKinsey estimates that recommender systems account for 35% of Amazon purchases and 70% of Netflix views [256]. Taken together, this means that in many contexts, the effects of data strikes would be magnified.

6.5.2. Homogeneous Campaign Group Experiments

For our homogenous campaign experiments, we focus our analyses specifically on the perspective of non-participating users (i.e. the *data labor effects* of a strike) because the consumer effects of a boycott are proportional to the number of positive ratings for the group in our experiments. As an example of our experiments, if 50% of women strike in a recommender system, 1) how does that affect the non-striking half of women (*Similar Users*), and 2) how does it affect users who are not women? Additionally, by focusing on the *data labor effects* here, we mitigate the challenge in comparing data strikes by groups that are very different in size.

As an example of our homogeneous analysis, consider *Similar Users* for women. We calculate surfaced hits for non-participating women with and without boycott and compute the percent change. We similarly calculate this result for other non-strikers (i.e. *Not*

Similar Users), and then find the ratio, which we call the *Similar User Effect Ratio*. Returning to our example, when half of women strike surfaces hits decreases for non-striking women by 0.38% while surface hits decreases for non-women by 0.09%. Therefore the Similar User Effect Ratio is 0.38 / 0.09 = 4.24. This means the effect of this data strike would be disproportionately felt by the group of similar users.

With this framing in mind, the results from our homogenous campaign experiments even more strongly support the potential of data strikes against recommenders using large datasets. Specifically, we observe that if cohesive groups of people strike, they can cause larger-than-expected reductions in recommender accuracy for the remaining *people within their homogenous group*. Furthermore, our results provide evidence that groups which

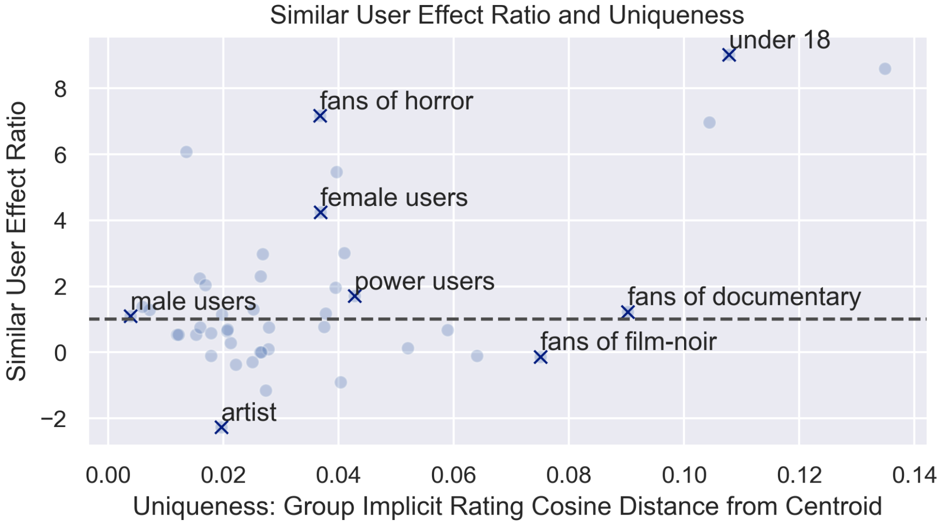


Figure 6.6. Scatterplot showing how homogeneous boycotts (with ¿ 20k ratings) affect similar users differently than the general population. Along the x-axis, groups are organized by increasing uniqueness, defined by the cosine distance between the mean implicit rating vector for the group and the general population. The y-axis shows the Similar User Effect Ratio. Gray dotted line shows ratio of 1. Pearson correlation is 0.55. "X" markers indicate various labeled examples.

are more unique in their preferences tend to be able to disproportionately affect their homogenous group. This suggests that campaigns against even very large recommenders can exert significant leverage, by threating large performance reductions for some subset of users who have unique preferences and interests. A secondary observation from these experiments is that homogenous groups' ability to affect the general population is not uniform: some groups "over-perform" expectations, and vice versa.

Some groups are especially effective at lowering performance for *Similar Users* compared to other users. Examples from across all five types of homogenous groups show how groups of various sizes have varying effects on *Similar Users* compared to other users. For instance, fans of horror have over 6x the effect on surfaced hits on other fans of horror compared to the rest of the population.

One hypothesis that explains why boycotts and strikes hurt *Similar Users* more than other users lies in a holistic view of the user preference space. If a group of users has substantial uniqueness– or more specifically, mathematical independence – in their preferences when compared to other groups, a campaign by that group is less likely to hurt users not in that group. At the extreme, a group whose preferences are completely orthogonal to every other group may be able to execute a campaign without substantially affecting personalization for any other group.

To understand this relationship, we compared a group's similar user effect ratios to a measure of preference independence based on rating overlap. To calculate preference independence, we first create a vector with a column for each movie (3706 columns) and value equal to the proportion of users in the group who have rated the movie. A group's preference independence is the cosine distance of a group's vector to the similarly calculated vector for the entire dataset.

Fig. 6.6 shows a full scatterplot of all groups with over 20k ratings (this means we remove six very small groups for which computing Similar User Effect ratio is very noisy). On the x-axis, we plot preference independence and on the y-axis we plot similar-user effect ratio. Our experiments suggest a moderate positive relationship between a group's preference independence and the effect of a boycott on non-participating users (Pearson's $r = 0.55$*,* $p <$ *0.001*). The interplay between preference independence and strike effectiveness is a fertile ground for research, and future work could more closely study how preference spaces might be operationalized for the purpose of data strikes.

While in some cases this effect may be undesirable (e.g. if strikers are averse to reducing algorithmic performance for similar individuals), it also suggests a promising avenue for challenging large recommenders that might be robust against small, "general" campaigns. For instance, if female users are interested in using data strikes against a company to effect some lasting change (e.g. changing discriminatory practices, profit sharing), they may be able to decrease company profit beyond what they could achieve in a general strike of the same size if they focus on recruiting female participants.

Additionally, a secondary observation from these experiments is that some groups perform better than "expected" based on their number of ratings (the expected value determined by a quadratic interpolation of our general boycott results, as described above). Homogenous boycotts were somewhat split between over- and under-performing random boycotts: in 34 of 50 homogeneous groups surfaced hits decreased *more* than would be expected for a random group of boycotters with an equal number of ratings.

Based on these findings, it seems that homogenous campaigns may not always be effective at damaging the general population of users, because some groups are under-performing, and even over-performing groups are limited by size in their ability to affect the general population. However, organizing homogenous campaigns with preference spaces in mind likely will be very effective and, critically, may provide a way to challenge large recommenders that otherwise are robust against data strikes.

6.5.3. Generalizing Beyond "Surfaced Hits" and SVD

In presenting our results, we have focused on our "surfaced hits" metrics. However, we also computed RMSE, NDCG with all items, NDCG@k, Precision@k, and Recall@k for k = 5, 10. We also computed these metrics when only including "long-tail" (i.e. unpopular) movies. All these metrics produce similar results regarding the effects of various data strikes configurations, although they did not afford us the ability to analyze boycotts. The full dataset of metrics is available in our Github repository. We also note that in our early experiments using item-based KNN and traditional metrics, we observed very similar trends.

6.6. Discussion

In this chapter, we have taken the previously hypothetical notion of data strikes, identified a wide variety of realistic campaign configurations (including when they are combined with traditional boycotts), and simulated the effect of these configurations in the recommender systems domain using best-practice evaluations. Comparing these campaign configurations, we find that while the *consumer power* of boycotts still substantially outweighs the *data labor power* of data strikes, data strikes represent a powerful new form of leverage. Moreover, we saw that data laborers might specifically target preference spaces within a recommender to achieve especially effective data strikes. Below, we discuss some of the more general implications of these results.

6.6.1. Barriers to Entry versus Impact

Our results suggest that collective action organizers targeting technologies have more options than is the case in traditional collective action. Specifically, these organizers have the ability to optimize for impact or for barrier to entry. Our results show that boycotts have a larger impact, but they require all participants bear the cost of not using a potentially valuable service. Strikes, on the other hand, allow data laborers to continue to benefit from the use of technology platforms without completely losing the ability to collectively bargain. Historically, participation in political consumption has been easier for affluent groups in politically stable countries [210] - data strikes represent an approach to collective action that may be substantially more accessible. Notably, our results suggest that new technologies focused on privacy (e.g. initiatives from Mozilla [284]) and online political consumption (e.g. recent work from Li et al. [241]) are a promising approach to empowering more individuals to collectively bargain.

6.6.2. Towards a Holistic View of Data Strikes

Although our results give us important insight into the potential impact of data strikes and boycotts, our work likely only captures a portion of the real-world effect of a collective action against a technology company. In particular, as is discussed in the Framework section, we cannot measure directly the indirect effects of consumer boycotts or data strikes. This means that our results should be interpreted as a lower bound on the effects of any collective action campaign against a technology company.

A related point that emerges from our results viewed with the lens of our framework is that collective action against technology companies will largely be more powerful than collective action against non-technology companies. The effect of a boycott against, for instance, a clothing company, would largely not include either direct or indirect loss of data value (excluding edge cases like long-term sales and marketing data). Since our results suggest that these factors will be non-trivial in most tech company boycotts, a user boycotting a tech company is likely to have a greater effect on revenue than would be expected in a boycott with a more traditional type of target. We expand on these power dynamics further below.

6.6.3. The Power of Algorithms versus The Power of Public Data

Outside of the context of boycotts and data strikes, our results can also be viewed as a means to better understand the power of data provided by the public relative to that of algorithms. Moderately-sized campaigns can bring recommender accuracy down to the levels of early recommender systems from 1999. Importantly, even for very large datasets (which are "harder" to action data leverage against), data strikes could effectively eliminate the ability of recommenders to make recommendations for some subset of users. For instance, in a hypothetical e-commerce dataset with 20 million ratings, 1 million of which are electronics products, a concerted boycott of electronics-reviewing users could remove the ability for the system to make personalized electronics recommendations.

These results, along with the work of McMahon et al. [265] and others [136, 163], emphasize the leverage that the public has in its relationship with data-hungry intelligent technologies and the companies that operate them. While the public perception of intelligent technologies like recommender systems is that they are largely the accomplishment of tech companies and the computer scientists they employ, these technologies are in fact a highly cooperative project between the public and companies. Without the companies and computer scientists, the intelligent technologies do not exist. But the same is also true for the public's contributions of data. This implies a much different power dynamic than is currently assumed by most people on both sides of this relationship.

6.6.4. Limitations

In addition to the discussion above about understanding the holistic effect of collection action against tech companies that includes online strikes and boycotts, this work also has several other limitations that should be highlighted. First and foremost, this chapter focused on recommender systems which, while a business-critical family of intelligent technologies, is only one family of intelligent technologies that could be vulnerable to collective action campaigns. Future work should seek to replicate our research for other intelligent technologies, for instance search ranking algorithms (e.g. [314]), "newsfeed"-style technologies (e.g. [293]), traffic prediction (e.g. [142]), and wi-fi geolocation (e.g. [266]).

Similarly, while we used best-practice evaluation techniques in the recommender systems community [71, 158, 335], these techniques have several limitations that also affect the large literature of recommender systems research that employs them. In particular, we considered only the explicit ratings and did not consider implicit preferences expressed through user behavior (which are not available in the MovieLens dataset). We also only considered our recommenders in an offline environment (as opposed to in a live experiment). Finally, to gain more insight into the nuances of recommenders, it will be valuable explore other recommender system datasets, particularly datasets from industry contexts.

Another important limitation is that in our experiments, we had to operationalize male/female as a binary variable due to the data available in the MovieLens dataset. Similarly, we were not able to test other types of demographic groups for the same reason (e.g. LGBT communities, political groups). However, future work on data strikes with a focus on preference spaces may be able to provide further insight into campaigns along these lines, without requiring the collection of sensitive demographic data.

Finally, this chapter focused on understanding the effect of collective action campaigns of various sizes and types, but it did not consider the collective action problem of organizing these boycotts. Fortunately, this problem maps to a deep body of work within social computing and related fields on sociotechnical strategies for motivating collective action online (e.g. [215, 336]). An obvious direction of future work in this research space involves building tools to organize data strikes and boycotts that leverages this body of work (either using GDPR or restricting new data collection). Recent work suggests that user-friendly tools like browser extensions may be an effective approach for making boycotts easy to join [241].

6.6.5. Potential Negative Impacts

In response to calls for the computing community to better engage with the negative impacts of our research [167], we discuss two major concerns with this work below.

First, we emphasize that our findings can be equally useful to organizers of campaigns as to they are to companies interested in mitigating the effectiveness of such campaigns. Relatedly, it is entirely possible that using a simulated boycott methodology, companies could identify which users are and are not “useful to the algorithm”. Technologies to organize collective action could help users bypass this issue by recruiting users across demographic lines (and perhaps, guided by future work, specifically focused on preference space lines), and our results suggest that this should be a priority in the design of these technologies.

Moreover, our ability to perform simulated campaigns was predicated on the public availability of the MovieLens dataset. Substantially more accurate simulations could be run using much richer datasets available only to corporations, so in any “data strike simulation arms race”, there will be a clear advantage for corporations. This means that corporations may be able to prepare models in advance to counteract boycotts or strikes. This might be mitigated through crowdsourced data collections or other means, a ripe area for future work.

6.7. Conclusion

In this chapter, we have done the work of advancing the notion of data strikes from abstract discussion point to concrete campaigns that can be simulated. Through these simulations, we identified the first empirical information to help advance the discussion around data strikes. In doing so, we first laid a framework that describes the key elements of collective action campaigns that action data leverage (i.e. data strikes and boycotts, which respectively leverage *data labor power* and *consumer power*). We found that these campaigns can be effective, with relatively small strikes wiping away significant portions of the value of recommender systems relative to simpler techniques. However, as datasets grow larger, data strikes become less effective, and strategies which target specific groups of users or preference spaces may become necessary. We discussed the implications of our results for those seeking to organize data strikes and companies seeking to understand potential effects on their core functionality.

CHAPTER 7

Conscious Data Contribution

7.1. Introduction

In this chapter, we explore how the public might exert **data leverage** against tech companies[1]. As we have noted many times in this thesis, users (i.e. the public) play a critical role in the economic success of tech companies by providing training data—i.e. "data labor" [13, 309]—that is existentially important to data-driven technologies. This means that users can take advantage of their role as data generators to gain leverage against data-dependent firms.

Recent research indicates that fertile ground exists for data leverage: in one survey published at CSCW, 30% of U.S.-based respondents reported they have already stopped or changed their technology use as a form of protest against tech companies, while in another survey 33% of U.S.-based respondents reported they believe tech companies have a negative effect on the country [84, 242].

Chapter 6 extensively discussed data strikes. In a data strike, a group of users who wishes to protest the values or actions of a tech company withholds and/or deletes their data contributions to reduce the performance of the company's data-driven technologies. While this research found through simulations that data strikes might be effective, data

[1] This chapter was originally published as [405].

strikes must contend with the diminishing returns of data to machine learning (ML) performance [169]. Indeed, the ability to generalize from limited data is one reason machine learning is so powerful. This means that a small data strike will likely have a very small effect on other users. Additionally, a user who participates in a data strike hinders their own ability to benefit from personalization-based ML systems, which may make participation hard to sustain in some cases.

In this chapter, we propose and evaluate an alternative means for users to exert data leverage against tech companies: **conscious data contribution** (or "CDC"). A group of users who wishes to protest a tech company using CDC contributes its data to a competing institution (e.g. another tech company) that has values or actions with which they agree more. They can additionally delete their data from the offending company's dataset, effectively combining a data strike and CDC. CDC takes advantage of the fact that data is "nonrival"—many firms can use the same data [196], so deleting data or quitting an existing technology is not a requirement for CDC. A group of people could help to support a new competitor in the market using CDC, without the need to completely quit using existing technologies.

In theory, CDC has two desirable characteristics compared to data strikes. First, CDC is more realistic within short-term time frames (as some data strikes will require support from regulators), which is important given the growing demand for immediate changes to the power dynamics between users and tech companies. In terms of legal support for CDC, regulators in various jurisdictions, e.g. the European Union, are increasingly advancing legislation that protects "data portability", the right for users to receive and re-use their personal data [235]. Tech companies are also supporting data portability features, e.g. Google's "Takeout" feature. As data portability laws and features become more common, CDC should become even easier to practice.

Second, while small data strikes must fight an uphill battle against the diminishing returns of data, CDC does not face diminishing returns until participation is high. While small data strikes may have a minor impact on a large company's technologies, small contributions of data could hugely improve the performance of a CDC beneficiary's data-driven technologies, helping it to compete with the target of a protest. In other words, CDC can more easily operate in the "vertical" region of the performance vs. dataset size curve instead of the "horizontal" region of this curve.

The goal of this chapter is to begin to understand how CDC might work in practice. To do so, we simulated CDC applied to four widely studied and business-relevant ML tasks: two recommendation tasks and two classification tasks. For context, we also consider data strikes—combined with CDC and on their own—in which users delete their data from an offending company. In total, we simulated three different data leverage scenarios: *CDC only*, *data strike only,* and *CDC with data strike*. To measure the data leverage achieved in each scenario with different *participation rates*, we compared the ML performance of a simulated large, data-rich incumbent company (the target of CDC) with that of a small competitor (the beneficiary of CDC). To enable comparisons across ML tasks with different evaluation metrics, dataset sizes, and data formats, we defined **Data Leverage Power** (DLP), a metric that facilitates cross-task comparison and the comparison of CDC to data strikes. In our analyses, we compare performance using both DLP and traditional ML metrics, which provide a task-specific perspective.

Our findings suggest that CDC with relatively small participation rates can effectively reduce the gap between a data-rich incumbent and its small competitor. If just 20% of users participate in CDC, the small competitor can get at least 80% of the way towards best-case performance for all our ML tasks. In certain situations, participation by 5% of users is enough to boost the small competitor's ML performance to 50% of best-case performance improvement, and 20% of users can get the small competitor 90% of the way.

Our results suggest that CDC may be more powerful than data strikes for many real-world contexts and could provide new opportunities for changing existing power dynamics between tech companies and their users. While we must be cautious in comparing the effects of CDC and data strikes because they operate differently (i.e. helping a competitor vs. directly hurting a company), we see that CDC is effective even when data strikes are impossible. More generally, our simulation experiments highlight how methods from machine learning research can be used to study and change power dynamics between tech companies and the public.

7.2. Related Work

7.2.1. Data Leverage and Consumer Leverage

In Chapter 8, we discuss a framework of "data leverage" that consists of three "data levers": data strikes, conscious data contribution, and data poisoning. Data strikes involve data deletion/withholding, conscious data contribution involves data generation/sharing , and data poisoning involves data manipulation aimed at harming data-driven technologies. In this chapter, we focus on comparing data strikes and conscious data contribution with simulation experiments. In other words, we focus specifically on two branches of the broader data leverage framework.

Previously, we simulated "data strikes" in the recommender system context and found that data strikes of moderate size (e.g., 30-50% of users) could be impactful. Weyl and Lanier [229] have proposed that cooperative entities might guide collective action like data strikes. While the data strikes concept has focused on withholding or deletion of data, Brunton and Nissembaum [39] have written about *obfuscation-based protest* – feeding intelligent technologies junk data to reduce their predictive power, using tools like AdNauseum [176]. Finally, Kulynych, Overdorf, and colleagues [220] laid out a broad framework for "Protective Optimization Technologies" (POTs) – technologies that attack optimization systems like intelligent technologies in order to reduce externalities caused by such systems. This framework is inclusive of any tactics that contest a technology, including the data levers we explore in this chapter (data strikes and CDC) or "data poisoning"—the third "data lever" —which draws on obfuscation, POTs, and machine learning literature on adversarial data.

7.2.2. The Relationship between Data Leverage and Consumer Leverage

While interest in data leverage is relatively new, a large body of work has studied forms of consumer leverage. The definition of CDC we use here was influenced by existing types of consumer leverage. In particular, the practice of "political consumerism", in which the public uses its consumer purchasing power as a political tool, is a precedent for CDC [282]. Political consumerism includes both boycotts and "buycotts": buying products or services to support a specific company. Buycotts, in contrast to boycotts, represent a "positive approach" to consumer action, as they reward, rather than punishing, a company [115]. Drawing on the dichotomies of data vs. consumer leverage and positive reward-based approaches vs. negative punishment-based approaches, CDC can be seen as a data leverage version of the positive, reward-focused boycotting approach.

Both political consumerism and interest in protest against tech are prevalent, suggesting there may be a large market for CDC. Recent work suggests over 50% of U.S.-based survey respondents having engaged in boycotts or buycotts in 2017 [102]. In another survey, 30% of U.S.-based respondents reported stopping or changing their use of technology companies in particular [242], and some work has begun to design tools for technology-assisted political consumerism [241]. The results of the survey work, in particular, suggest that large-scale CDC is well within the realm of near-term feasibility: if CDC is made straightforward, it seems there are many people who will be interested.

7.2.3. Learning Curves and Diminishing Returns from Data

Research on the relationship between ML performance and dataset size provides important motivation for conscious data contribution. Many scholars have empirically studied "learning curves" – the relationships between training dataset size and ML performance – for a variety of models. When looking at a specific task (e.g., image classification), learning curves can be characterized as exhibiting diminishing returns. At some point, the relationship between performance and data size becomes "flat" [34]. Diminishing returns have been observed in a variety of contexts, for instance statistical machine translation [208], deep learning [55, 169, 328], logistic regression [303], decision trees [80], and matrix factorization recommender models (see Chapter 6).

The techniques used to study learning curves can be adapted to study data leverage. A typical procedure to generate a learning curve would entail randomly sampling a fraction of training data, retraining a model with this sample of data, and measuring performance of the new trained model. One might repeat this procedure for some fixed number of iterations to obtain the average performance for a random sample of a certain size. By doing this for a variety of fractions, we obtain a curve of performance vs. dataset size. To study data leverage, we are interested in the curve that relates performance to the fraction of users who contribute data (though in some cases we must use the fraction of data as a proxy for fraction of users). At a high level, we obtain this curve using the same procedure for computing a learning curve that was used by Perlich et al. [303] and Cho et al. [55]. However, as we will describe below in Methods, there are some additional implementation details that distinguish data leverage simulations from learning curve experiments and help to increase the ecological validity of our simulations.

7.2.4. Data Leverage and Online Collective Action

The data leverage scenarios we simulated in this chapter are instances of online collective action, a topic with a rich literature in social computing. This literature considers online collective action in a broad variety of contexts. These contexts include leadership as a collective activity in Wikipedia [440], collective action for crowdworkers [336], the use of Twitter bots to organize civic volunteers [341], and the development of tools for "end-to-end" collective action [437]. It is likely that strategies that work well for these contexts may apply directly to data leverage scenarios, in particular by facilitating the organization and participation necessary to achieve large-scale participation.

For instance, leaning on findings from CSCW scholarship about leadership in the Wikipedia community, CDC groups might encourage members who are not formal leaders to take leadership actions using strategies from Wikipedia [440, 439]. More generally, CDC stands to benefit from most research on successful peer production (e.g. [152, 201]), as peer production participants (especially Wikipedia editors) are already critical sources of data labor that fuels AI systems [247], and share core similarities with CDC participants. Concretely, groups engaging in CDC can emulate peer production strategies, and in some cases participating in peer production could be a form of CDC (e.g. contributing labeled images to Wikimedia Commons with the goal of helping start-ups train computer vision models). Another area of CSCW research that is highly relevant to CDC is crowdwork. Although there are major differences between crowdwork and peer production, crowdworkers also provide crucial data labor [202] and have led successful collective action movements in the past [336].

Similarly, looking to research on bots and support tools in social computing, CDC participants might re-use or adapt existing bots, social platforms, and browser plug-ins to promote CDC engagement [241, 341, 437]. Tools to support political consumption, such as browser extensions that help people boycott websites, might be enhanced with CDC features [241].

Social computing researchers have called for more work that addresses "Computer Supported Collective Action" (CSCA) [359]. Generally, any given data leverage scenario (i.e. CDC and/or a data strike by some group) can be seen as an instance of CSCA [359]. This means those seeking to use data leverage must face the many challenges associated with collective action, e.g., challenges with leadership, communication, and planning. Conversely, models of success in CSCA can provide templates for successful CDC.

7.3. Methods

We conducted a series of experiments to compare the ML performance of two simulated companies when users exert data leverage using CDC and/or data strikes. For our simulations, we assume the following scenario. There exists a large, data-rich incumbent company – called "Large Co." – that starts with a full dataset. Some users of Large Co.'s data-driven technologies are interested in protesting Large Co. because of its values or actions. To do so, they want to support a small, data-poor competing company – "Small Co." – that better aligns with their values. We considered variations in this scenario in which users can contribute data to Small Co.'s dataset while simultaneously deleting it from Large Co.'s dataset (*CDC with data strike*) as well as variations in which deletion is impossible (*CDC only*). For additional context, we also considered variations in which users engage only in a data strike (*data strike only*). In all our simulations, data strikers delete all their data contributions and then models are retrained.

To begin our experiments, we first had to identify specific ML tasks to study and a corresponding ML approach to implement for each task. We selected four tasks that have both attracted substantial research attention and have clear industry applications: two recommendation tasks (using movie ratings and Pinterest interaction records) and two classification tasks (images and text). For each task, we sought out a top-performing ML approach with a publicly available implementation. Below, we further detail our simulation assumptions and the specific tasks, datasets, and ML approaches from prior work that we used.

At a high level, our experiments follow procedures similar to those used in learning curve research that has studied the relationship between ML performance and training dataset size [169]. For each task, we repeatedly retrained the corresponding model with samples of the benchmark training set corresponding to different CDC and/or data strike participation rates (e.g. 1%, 5%, etc.), and evaluated model performance.

Specifically, our data leverage simulations have three major differences from traditional learning curve simulations. First, in simulating data leverage scenarios, for datasets with user identifiers, we drew a sample of users to participate in CDC and/or a data strike instead of drawing a sample of data points. This approach simulates what would happen in a CDC scenario, in which data is added or removed on a user-by-user basis. For our classification datasets, which lack user identifiers, we randomly sampled data points, as in learning curve research. This is an inherent limitation, as many influential classification datasets lack user identifiers for privacy reasons.

Second, when simulating CDC, we are primarily interested in ML performance as evaluated from the perspective of each company. To get this *company perspective* evaluation, we use a test set drawn from each company's data sample. For instance, if Small Co. receives data from 10% of users, Small Co. must create its own test set using data from these 10% of users. However, as a secondary measurement, we can also hold out a separate, fixed test set that is hidden from each company. This *fixed holdout* perspective allows us to measure a performance while taking into account people who are accessing the technology as a brand new user (or anonymously, i.e. a user who receives recommendations in "Private Browsing" mode). This *fixed holdout* test set can also be seen as a more objective external measurement of model performance, as the *fixed holdout* set is the same across every simulation and is unaffected by which users or observations are available to a particular company. In a more practical sense, this *company perspective* vs. *fixed holdout* comparison is most relevant to personalization tasks, where performance might differ drastically for new (or anonymous) users.

Third, as mentioned above, our data leverage simulations allow us to consider the case in which one company gains data while another company loses data, i.e., in a data strike. Below, we will discuss how we addressed the challenges in comparing these scenarios (i.e. how do we define a comparison metric that allows us to compare the effectiveness of giving Small Co. data vs. deleting data from Large Co.?)

7.3.1. Simulation Details

Following past learning curve studies, we considered a range of participation rates. Specifically, we conducted simulations in which a group of users engages in one of the following scenarios: *CDC only*, *data strike only*, or *CDC with data strike.* We considered participation rates of 0.01 (1% of users or data), 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. We leveraged a natural symmetry of our experiments to estimate effects for larger groups: the situation in which Small Co. has some fraction s of all data is equivalent to the situation in which Large Co. has lost 1 - s of all data. For instance, if Small Co. gains 10% of all users or data, the expected ML performance is the same as when Large Co. has 90% of its users or data deleted. Thus, without running additional experiments, we can also measure ML performance for participation rates of 0.6, 0.7, 0.8, 0.9, 0.95, 0.99. As we will discuss further below, we also consider baseline results (i.e. when Small Co. has very little data) and best-case results (i.e. when Large Co. has all possible training data) in our definition of Data Leverage Power.

The end result of our simulations is to generate a curve showing the effectiveness of data leverage scenarios at different participation rates. While our simulated scenario focuses on just two companies, this curve could be used to evaluate data leverage exerted against many companies. If several companies simultaneously benefit from CDC, we can use each beneficiary company's participation rate to get corresponding DLP. For instance, we might use this curve to compare the performance of one company that received CDC with 10% participation, another company that received CDC with 20% participation, and a third company that was the target of a data strike by 30% of its users.

In each simulation, Small Co. and Large Co. gain access to a scenario-specific dataset, split their respective datasets into a train set and company-specific test datasets, train their models, and evaluate their trained models on two test sets: (1) their *company perspective* test set (which is unique to each scenario) and (2) a *fixed holdout* test set. As mentioned above, the *fixed holdout* test set provides an objective evaluation of performance (it does not depend on the company-specific test set) and is particularly relevant to personalization tasks. We further detail the purpose of considering these two different test sets below.

For each training run, we use the same, fixed hyperparameters that achieved reported results in prior work instead of running hyperparameter search for each simulation. While this substantially reduces the computational cost of our experiments, this means that for any given scenario, our results do not necessarily represent the best possible performance for a given participation rate. For instance, if Small Co. only has 10% data to work with, it's possible Small Co. could use different hyperparameters (or other techniques, such as data augmentation, different training algorithms, etc.). to boost performance. However, by using fixed hyperparameters, we reduce the computational cost and make it easier to explore several tasks, use multiple sampling seeds, and replicate our findings.

More formally, our simulations followed the same procedure for each fractional group size s:

Identify the data that will be given to Small Co. via CDC and removed from Large Co. via data strikes. To identify this data, we randomly sampled s *(* the participation rate) of all users and took the data attributed to the sampled users. For tasks in which units of data are not attributed to specific users (image and text classification), we sampled s of all units of data.

Train and evaluate a model using Small Co.'s dataset. This gives us Small Co.'s performance in the *CDC only* scenario.

Train and evaluate a model using Large Co.'s dataset. This gives us Large Co.'s performance in both the *CDC with data strike* and *data strike only* scenarios.

Compare Small Co.'s performance and Large Co.'s performance to worst-case and best-case performance values. As we will describe below, we formalized this comparison by defining "Data Leverage Power", a measurement that we used (alongside traditional ML metrics) to compare data leverage across different ML tasks. At a high level, DLP measures how close CDC gets Small Co. to Large Co.'s performance, or how close a data strike moves Large Co.'s performance to low-data baseline performance. We expand on DLP's precise definition and the motivation underlying the measurement below.

In order to measure average performances for each participation rate, we repeated this procedure using five different seeds for sampling s users/data and calculated the average performance across these iterations. Thus, for each ML task we studied, we retrained a model 70 times (7 group sizes, 5 sampling seeds, and 2 different companies), which was made easier by our choice to avoid costly hyperparameter searches.

7.3.2. Tasks, Datasets and Machine Learning Implementations

As mentioned above, we first identified ML tasks of interest to ML researchers in industry and academia, and then sought out a publicly available implementation of a high-performing ML approach for each task. To make experimentation feasible, we also needed to identify implementations that would be possible to retrain many times. We identified four public implementations of successful approaches to well-studied, industry relevant ML tasks: Rendle et al.'s [321] implementation of recommender systems that use star ratings, Dacrema et al.'s [74] implementation of recommender systems that use binary interaction data, an image classifier from the Stanford DAWNBench competition [60], and a text classifier from Google Jigsaw's Toxic Comment Classification challenge hosted on Kaggle [249].

Below, we provide additional detail about the four ML tasks and the specific datasets and models from prior work that we used. In each case, we followed prior work closely in our implementations so that the best-case performance achieved by Large Co. when it has access to a "full dataset" (i.e. 100% of the dataset used in prior work) is comparable to the published results. To make our simulations ecologically valid, our goal was for Large Co.'s best-case, full dataset performance to be comparable to reported results in prior work, while keeping computational costs down. To this end, we made some small changes aimed at reducing the cost of experiments while minimally impacting performance. We used software from prior work where possible and make the code for our experiments available.[2]

The first two tasks we studied were recommendation tasks that involve training "recommender systems" to predict the rating a user will give an item and predicting whether a user will interact with an item. Recommender systems are enormously important to a variety of industries, have garnered huge attention in the computing literature, and are immensely profitable [158, 256, 438]. The second two tasks we considered were classification tasks: classifying images (with ten possible image classes) and classifying text as toxic or non-toxic. These tasks come from two large ML research areas (computer vision and natural language processing) and are representative of classification systems that are applicable to a huge variety of industries (i.e. identifying the class of an image or piece of text is broadly useful).

For the rating prediction task, we used Rendle's [320] factorization machine approach with the extremely influential MovieLens 10-M dataset [158], as Rendle et al. [321] demonstrated that this approach outperformed a variety of competing approaches in terms of Root Mean Squared Error (RMSE), a metric used in past recommender system competitions. Specifically, we used Bayesian Matrix Factorization with size 32 embeddings and 50 sampling steps, which substantially lowers training costs and slightly lowers performance

[2] https://github.com/nickmvincent/cdc

compared to the most expensive configuration Rendle et al. used (size 500 embeddings and 500 sampling steps).

Dacrema et al. rigorously compared simple, yet well-calibrated baseline techniques to complex neural techniques for the interaction prediction task. We focus on the Pinterest recommendation dataset from Dacrema et al.'s work, originally from Geng et al. [125]. Dacrema et al. showed that a simple item-based k-nearest neighbor performs extremely well—better than neural techniques—for this dataset. We use Dacrema et al.'s implementation of a k-nearest neighbor recommender system. In terms of evaluation, Dacrema et al. used Hit Rate@5 (alongside other metrics that led to similar findings). Hit Rate is defined by holding out a single item that each user interacted with and then using the model to rank the held-out item and 99 random items the user did not interact with. If the item that the user actually interacted with is in the top five ranked results returned by the model, this is considered a success, and Hit Rate@5 is the total fraction of users for which the model was successful. This evaluation procedure maps well to top-n recommendation features common on many platforms (e.g. "Top Videos for You").

For a classification task from computer vision, we consider the CIFAR-10 dataset, a popular benchmark dataset for image classification [217]. The Stanford DAWNBench challenge [60] includes a leaderboard that documents image classification approaches that achieve high accuracy using minimal training time. From this leaderboard, we used Page's [294] ResNet approach and the well-studied CIFAR-10 dataset [217] (which includes images belonging to ten different classes). For this task, we evaluated our models using accuracy: for this task, accuracy is defined as fraction of test images that are successfully classified.

From natural language processing, we consider the case of toxic comment classification, using labeled Wikipedia discussion comments from Google Jigsaw's ML challenge hosted on Kaggle [249]. We used a TF-IDF logistic regression approach from the Kaggle leaderboard that achieves performance comparable to top models. While the dataset has labels for six different overlapping categories of toxicity, we focused only on binary classification (toxic vs. non-toxic) such that we train only a single model for each simulation. We made this task binary by treating a comment with any toxicity-related label (toxic, severely toxic, obscene, threatening, insulting, hateful) as generally toxic. As in the Kaggle competition, we evaluate the model using area under the receiver operator curve (which we refer to as AUC, for area under the curve). The binary classification performance of the ML approach is very close to the average performance across the six categories.

Our datasets also cover a range of sizes and data types: the ML-10M dataset has 10M explicit ratings (1-5 stars) from 72k users. The Pinterest dataset has 1.5M interactions from 55k users. CIFAR-10 has 50k train images and a fixed set of 10k test images. The Toxic Comments dataset has 160k comments. This presents a major challenge in comparing the results of our data leverage simulations: each task is typically evaluated in a different manner (i.e. RMSE vs. Hit Rate vs. Accuracy vs. AUC), our datasets are of different sizes, and the format of data varies substantially (i.e. a movie rating is different than an image or piece of text). For instance, we might perform several experiments that show that a contribution of x star ratings can improve recommender RMSE by $y,$ whereas a contribution of x images can improve image classification accuracy by z. However, comparing changes in RMSE to changes in image classification accuracy is not straightforward. Below, we describe how we defined Data Leverage Power in order to address this challenge.

7.3.3. Measuring the Effectiveness of CDC with Data Leverage Power

Different ML tasks require different evaluation techniques and our datasets have different sizes and different data formats. Here, we describe how we compared the effectiveness of CDC across different tasks.

To measure effectiveness, we introduce a task-agnostic measurement: **Data Leverage Power** (DLP). The goal of defining DLP is to create a single metric that captures the effectiveness of data leverage across ML tasks and across different data leverage scenarios (e.g., strike vs. CDC). DLP is defined in a scenario-specific manner such that it tracks Small Co.'s ability to catch up with Large Co. in ML performance in CDC scenarios, but also tracks the ability for a data strike to lower Large Co.'s performance in the *data strike only* scenario (in which Small Co.'s performance does not change).

For each participation rate, DLP takes into account four measurements: baseline performance (performance with a very low data approach, for instance a "random guess" approach for classification or "recommend most popular items" approach for recommendation), the full-data best-case performance, Small Co.'s average performance, and Large Co.'s average performance. Below, we refer to these four measurements, respectively, as *baseline*, *best*, *small*, and *large.* While measuring full-data best-case performance is straightforward, selecting a baseline is less so. After walking through exactly how we defined DLP below, we describe how we identified baseline performance for each task.

For our two scenarios that involve CDC (*CDC only* and *CDC with data strike*), DLP is defined as the ratio of Small Co.'s average performance improvement over baseline to Large Co.'s average performance improvement over baseline for a given participation rate. In other words, we compare how much better Small Co.'s performance improves on the baseline to how much Large Co.'s performance improves on the baseline. Mathematically, for our scenarios that involve CDC, DLP defined as is:

$\frac{small-baseline}{large-baseline}$

In *CDC only* scenarios, Large Co.'s performance never changes (no data strike occurs) and is therefore fixed at best-case performance, while Small Co.'s performance increases with participation rate. In the *CDC with data strike* scenarios, larger participation rates lower Large Co.'s performance while increasing Small Co.'s performance.

As an example, imagine that for some model, best-case performance is 1.0 accuracy and worst-case is 0.5. With full data, Large Co. achieves the best-case 1.0 accuracy and thus has an improvement over worst-case of 0.5. For CDC by 10% of users, Small Co.'s accuracy (averaged across iterations) is 0.7, an improvement over worst-case of 0.2. If this is accompanied by a data strike and this data strike causes Large Co.'s performance to drop to 0.9, Large Co. now has an improvement over worst-case of 0.4. In this case, the DLP for *CDC only* is 0.2 / 0.5 = 0.4 and the DLP for *CDC with data strike* is 0.2 / 0.4 = 0.5. By repeating the entire process for every group size, we obtain a full plot of DLP vs. participation rate. For each task, we set "baseline" performance as corresponding to the worst performance from all our experiments, which occurs when either company has as little data as possible (in our experiments, 1% of users/data). For ML-10M, this occurs when Small Co. has 1% of users and is evaluated on the *fixed holdout* test set. In this case, the model effectively guesses the mean rating for almost all predictions. For CIFAR-10, the worst-case performance also occurs when Small Co. has 1% of the data, and is about 10%, equivalent to randomly guessing one of ten classes. For Toxic Comments, the lowest performance occurs when Small Co. has a 1% sample of data (and is about 0.9 AUC). The Pinterest task is a special case, as our approach cannot make predictions for unseen users. To get a baseline for Pinterest, we followed Dacrema et al. and used the performance achieved when using a simple "recommend most popular items" approach with full data. As we will discuss below, we also "recommend most popular" to calculate *fixed holdout* performance (because in a *fixed holdout* scenario, the recommender will face unseen users).

In *data strike only* scenarios, however, Small Co.'s performance is fixed at worst-case performance and therefore Small Co.'s improvement over baseline is zero. This means the numerator of the ratio we used above is always zero in *data strike only* scenarios. Therefore, for these scenarios, we calculate how much Large Co.'s performance has fallen from best-case performance and find the ratio of Large Co.'s performance loss to gap between best-case and baseline. This "no-CDC" version of DLP is still comparable to CDC version, as it measures the delta between Small Co. and Large Co.'s performance. Mathematically, DLP for *data strike only* scenarios is:

$\frac{large-best}{baseline-best}$

The DLP approach to comparing data leverage simulations accounts for the fact that datasets are of different sizes and are comprised of different, hard-to-compare data types (e.g. a single image is different from a single user-item interaction). By focusing on DLP and participation rate (instead of e.g., number of users, number of observations, number of gigabytes of data, etc.), we can make comparisons across ML tasks, e.g., how does CDC by 30% of ML-10M users compare to CDC by 30% of CIFAR-10 users?

In interpreting our results, we calculate the participation rates needed to achieve a certain DLP for each ML task. For instance, we ask "How many users does it take to get 80% of the way to Large Co.'s performance?" We consider a variety of reasonable round-number DLP thresholds, because acceptable performance levels will vary by user and ML task (identifying acceptable performance levels is an important area of future work). For instance, a user who is motivated to support Small Co. (e.g. because they feel strongly about protesting Large Co.'s values or actions and feel strongly about supporting Small Co.'s values or actions) might accept much worse performance from Small Co.'s technologies than a user who does feel as strongly about the companies' value or actions. For instance, even for a DLP of 0.5 – e.g. Small Co.'s recommender system gets just 50% of the way to best-case performance – a user who strongly supports Small Co. might not mind needing to scroll further through their recommendation lists.

It is important to note that while DLP was critical to our ability to compare different data leverage scenarios, is important to consider task-specific factors such as performance thresholds and the real-world value of improved performance, i.e., if performance changes from x to y, what are downstream effects on revenue, user retention, etc.? To increase the interpretability of our piecewise definition of DLP and address the second challenge, we also report the raw performance values (traditional metrics) that accompany a given DLP value. In doing so, we retain the benefits of DLP (easy comparisons across tasks) while still allowing those familiar with a particular task to understand task-specific effects of a data leverage campaign (i.e., if a DLP campaign moves performance over or under an important task-specific performance threshold). As we will highlight again throughout our Discussion, a similar approach that treats DLP and traditional metrics as complementary will also be useful for studying data leverage in practice.

Additionally, comparing a data strike to CDC is comparing an action that harms ML performance to an action that helps ML performance. In order to address this challenge, we defined DLP as piecewise, with a separate definition for CDC scenarios and *data strike only* scenarios, such that harming Large Co. and helping Small Co. both represent increased balance in power between Large Co. and Small Co. In other words, our piecewise definition is motivated by the assumption that CDC and data strikes represent two ways of achieving the same goal.

7.3.4. Company Perspective vs. Fixed Holdout Evaluation

As mentioned above, when simulating CDC, we evaluate our models using scenario-specific *company perspective* test sets (i.e. test sets drawn from the data that Small Co. or Large Co. have). However, as a secondary measure, we can also look at evaluation metrics from a *fixed holdout set* that neither company can access, which summarizes ML performance while taking into account the experience of new users and anonymous users. In other words, we consider both a "subjective" test set and an "objective" test set (in the sense that the "objective" *fixed holdout* set is unaffected by the specifics of each data leverage scenario, while the "subjective" *company perspective* test set is affected). For ML-10M and Toxic Comments, we used the same approach used in Rendle et al.'s review and sampled a random 10% of data to create a *fixed holdout* set. For the Pinterest and CIFAR-10 datasets, we used the fixed holdout sets used in the prior work [74, 217] that inspired our modeling approach.

The distinction between *fixed holdout* and *company perspective* test sets is most important for personalization tasks (e.g., recommendation). For these contexts, if a company has no data about a particular user (e.g., because that person is a brand-new user, is accessing a service anonymously, or is engaging in obfuscation), that user necessarily receives non-personalized worst-case performance. For the ML-10M case, the approach we used can only, at best, predict that anonymous users will give every item the mean rating of all items. For the Pinterest case, the approach we used cannot produce recommendations for unseen users, so we uniformly assigned these users a Hit Rate@5 contribution of 0.1668, corresponding to Hit Rate@5 documented by Dacrema et al. when using a non-personalized "recommend most popular items to everyone" approach.

The real-word scenario that *fixed holdout* evaluation maps to is the one in which users receive recommendations from both Small Co. and Large Co. but use one or both of the services as a new or anonymous user (or use obfuscation to make themselves effectively anonymous). For instance, if only 10% of users contribute data to Small Co., but every single user chooses to use Small Co.'s recommender system, it will necessarily perform poorly for the 90% of users for whom the model cannot provide personalized results.

Critically, if a company has strong *company perspective* performance, but poor *fixed holdout* performance, this means their technologies will be very effective for current users, but they may have trouble expanding their userbase. Our experiments involve random sampling, so each simulated company perspective test set is drawn from the same distribution as the fixed holdout set, with the main difference arising when a user appears in the *fixed holdout* set but not a particular company's test set. In practice, CDC and data strikes may be practiced by homogenous groups, and so the distinction between company perspective and fixed holdout may become even more important. In presenting our results, we focus first on *company perspective* evaluation, and then discuss the implications of looking at results using a *fixed holdout se*t. Comparing different "test set perspectives" will be an important component of future data leverage research.

7.4. Results

Below, we present the results from our DLP simulations. We begin by focusing on our *CDC only* results. Next, we examine the additional effect of adding a data strike to CDC and examine our results for *CDC with data strike* and *data strike only* scenarios.

As mentioned above, our primary focus is on ML performance measured with *company perspective* evaluation. At the end of this section, we present our secondary measurement, *fixed holdout* evaluation, and describe how this secondary measurement informs us about interactions between data leverage and personalization systems.

7.4.1. *CDC Only* Scenarios using *Company Perspective* Evaluation

Our full set of *company-perspective* experimental results are shown in Fig. 7.1. The left column of Fig. 7.1 shows Data Leverage Power (DLP), our measurement described above that allows us to compare results across ML tasks. A higher DLP value means that a data leverage action was more effective, in terms of boosting Small Co. and/or reducing Large Co.'s performance. Within the left column, *CDC only* results are shown in black. To show how DLP relates to task-specific performance, the right column shows the various task-specific evaluation measurements that we used to calculate DLP for each scenario: Hit Rate, Root Mean Squared Error (RMSE), Accuracy, and Area under the Receiver-Operator Curve (AUC). Here, there are only two colors: black shows Small

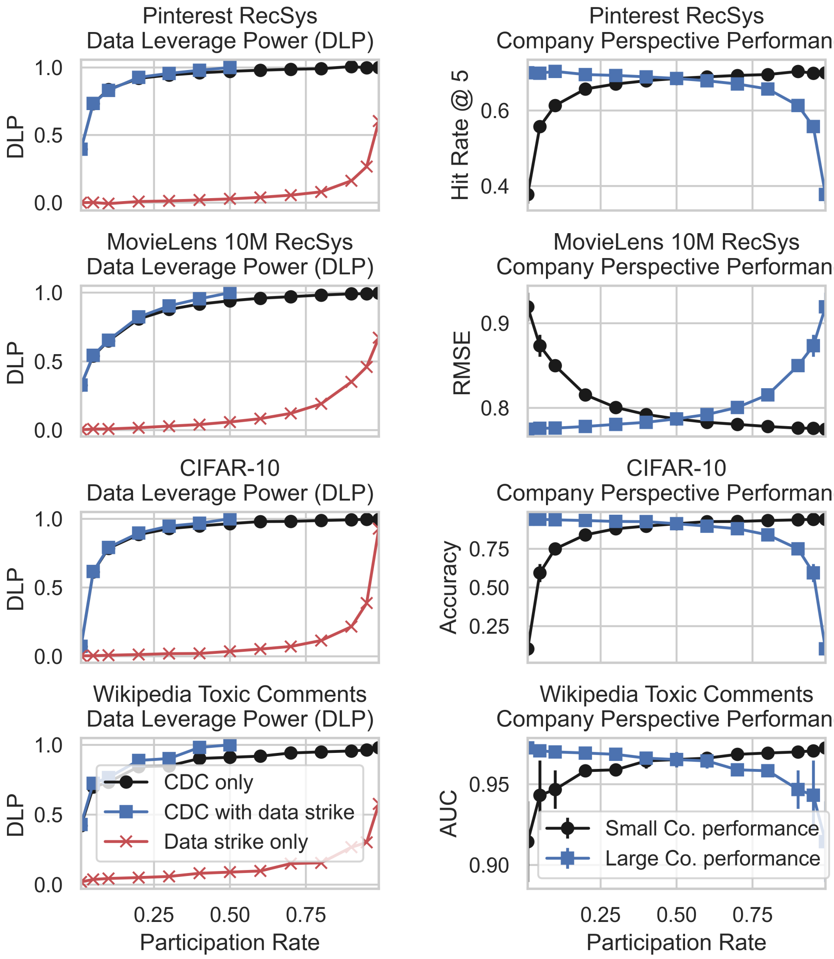


Figure 7.1. The left column shows DLP plotted against participation rate for CDC only, CDC with data strike, and data strike only scenarios. Each row shows a different ML task. The right column shows the task-specific performance measurement we used to calculate DLP. Vertical bars show standard deviation for task-specific results.

Co.'s performance improving as CDC participation increases while blue shows Large Co.'s performance decreasing as data strike participation increases (we discuss these data strike results below). As described above, using baseline performance, best-case performance, and these two performance curves, we can compute DLP values (left column) for all three of our scenarios.

Examining the black curves in the left-hand column of Fig. 7.1, it appears that CDC can be highly effective at allowing a small company to drastically reduce the performance gap between itself and a large competitor, as across our four tasks we see a CDC participation rate of at minimum 10% and at most 20% is needed to get Small Co.'s performance 80% of the way to best-case (i.e., the black curve reaches a DLP of 0.8). In general, both the scenario-specific evaluation curves (right column) and resulting DLP curves (left column) display diminishing returns of data. However, comparing the rows in Fig. 7.1, we see that the effectiveness of CDC at reducing the performance gap between companies is not identical across tasks. The curves "level off" at different rates. Like learning curves, DLP curves are influenced by a variety of factors such as the algorithm used and the size of the full dataset. DLP provides us with a consistent way to make comparisons.

To systematically compare the black DLP curves from each row shown in the left column of Fig. 7.1 is, we asked, "what group size of CDC is needed to achieve a given DLP"? Drawing from the results shown in Fig. 7.1, Table 7.1 shows the CDC group size needed to achieve DLP thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9.

Looking at Table 7.1, we see that CDC was especially effective for the Pinterest recommendation task, e.g., for the Pinterest task, only 20% of users need to engage in CDC

		DLP Thresholds				
		0.5	0.6	0.7	0.8	0.9
ML Task	Pinterest	5%	5%	5%	10%	20%
	ML-10M	5%	10%	20%	20%	40%
	CIFAR-10	5%	5%	10%	20%	30%
	Toxic Comments	5%	5%	10%	20%	40%

Table 7.1. Shows the CDC participation rate needed to reach DLP thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9 for each ML task we studied. For instance, the bottom left cell shows that 5% CDC participation is needed to reach a DLP of 0.5 for the Toxic Comments task.

to achieve 90% of best-case performance. For the other tasks, larger CDC participation (i.e. 30-40%) was required to achieve this same 90% DLP threshold.

Even for the cases that require greater participation to achieve a given DLP threshold, the results in Table 7.1 suggest CDC could be quite impactful. For instance, in the ML-10M case, 20% of users can achieve a DLP of 0.8, which may be good enough for some potential users to begin using Small Co.'s ML model. Furthermore, the Small Co. performance results are not upper bounds on performance for a given participation rate; it is possible that Small Co.'s data scientists could get even better performance by using a model that performs better for small datasets, by adjusting hyperparameters, by using data augmentation, etc.

Looking at the traditional performance metrics in the right column, we see underlying values from which DLP was calculated. To obtain DLP of 0.8 with in CDC only scenarios, Small Co. must achieve substantial improvements over the baseline. For instance, at participation rate of 20%, Small Co. achieves a Pinterest Hit Rate of 0.66, MovieLens RMSE of 0.82, CIFAR-10 accuracy of 0.84, and a Toxic Comments AUROC of 0.96. These are raw performance metrics that may very well be acceptable for Small Co.'s use. For instance, in their paper, Rendle et al. documented progress on ML-10M: an RMSE of 0.82 is equivalent to the state of the art in 2008, which may be adequate for CDC participants. Comparing the left and right columns of Fig. 7.1, we see that converting DLP to raw performance is relatively straightforward for *CDC only*, because Large Co.'s performance never changes (as there is no data strike) and the DLP curve therefore has the same shape as the raw performance curve.

7.4.2. CDC With Data Strike and Data Strike Only using Company Perspective Evaluation

Next, we look at how DLP changes when users engage in CDC (to Small Co.) and simultaneously engage in a data strike (by deleting data from Large Co.). Returning to the left column of Fig. 7.1, we focus on the blue squares (*CDC with data strike*) as well as the difference between the blue squares and black circles (*CDC only*).

Our primary finding here is that a data strike can add a small effect on top of CDC, but only when participation rates rise above around 20%. Even for a participation rate of 30%, adding a data strike adds at most 0.05 DLP. This means that for a group of 30% of users, the ability to delete data lowers Large Co.'s performance and thus "closes the performance gap" by at most an additional 5%. Of course, the logistical and potential legal challenge of adding the data strike on top of CDC may incur large costs relative to the gain.

Returning to the raw performance metrics in the right column, we can see how the additional "boost" from incorporating data strikes corresponds to reduced performance for Large Co. Looking at Large Co.'s performance curve in blue, we unsurprisingly see the same characteristic diminishing return curves: small data strikes only have minor impact on Large Co., explaining the relatively small DLP boost from adding data strikes to CDC.

Even at a participation rate of 0.5 (i.e. a group of 50% of all users included in the benchmark dataset), the benefits of adding a data strike are still modest, peaking at 9% for the Toxic Comments case. At the same 50% group size, deletion had only a 3% benefit for the Pinterest case. We cannot measure the benefits for groups larger than 0.5, because at this point Large Co. now has less data than Small Co. and we effectively begin traversing the DLP curve backwards (i.e., Small Co. has now become the "larger" company in terms of dataset size).

Finally, looking at the *data strike only DLP* curves (red x's in the left column), we see that data strikes alone have very little DLP relative to CDC approaches. This illustrates the challenge of diminishing returns of data. If a learning curve is flat (or almost flat) for a wide range of participation rates, a data strike alone must break out of this flat region to begin exerting substantial data leverage. As we will see immediately below, an interesting exception to this trend is when we consider anonymous users: users who engage in a data strike against Large Co. but continue to use Large Co.'s technology will receive non-personalized results, which can hurt Large Co.'s overall performance substantially.

To summarize, we see that data strikes have much less potential to exert data leverage than CDC at participation rates less than 50%. Of course, CDC requires the existence of a competitor, so in many cases people may only be able to data strike. Specifically, in monopoly contexts, data strikes may be the only option available to users. We expand on how supporting CDC might support greater competition, as both a tool against monopoly but also in the service of innovation.

7.4.3. CDC Effectiveness using *Fixed Holdout* Evaluation

As discussed above, it is informative to examine the effects of CDC when ML models are evaluated using a fixed holdout test set that includes users who may not appear in the *company perspective* test set. Fig. 7.2 shows the same measurements as Fig. 7.1 but uses *fixed holdout* performance instead of *company perspective* performance.

We observed that for non-personalized CIFAR-10 and Toxic Comment cases, *fixed holdout* test set evaluation gave very similar results to *company perspective* evaluation. This is expected: the *company perspective* test sets were randomly drawn from the same source as the *fixed holdout* test sets in our experiments. The only notable difference was higher standard deviation (vertical bars visible in the last row of Fig. 7.1) in performance for small amounts of data, because the *company perspective* test sets are smaller for smaller group sizes.

For ML-10M and Pinterest recommendation tasks (i.e. tasks that involve personalized results)*, fixed holdout* test set performance is different from *company perspective* performance. Rather than a diminishing returns curve, ML performance (and the resulting DLP) is linearly dependent on participation rate. Specifically, as each company loses users, their recommendation performance linearly approaches non-personalized baseline performance.

As described above, this result corresponds to a situation in which some users access a recommender system with a new account (i.e. because they are new users, or perhaps because they deleted their account as part of a data leverage campaign), forcing the system to output non-personalized results. Specifically, imagine that 30% of users engage in CDC to support Small Co. Some other group of users access Small Co.'s recommender system with new accounts. These users who did not participate in CDC but continue to use Small Co.'s model anonymously will not see the benefits of other people's data contributions until they provide their own data. This result highlights that personalized ML systems require users to provide their own data before they can see the benefits of CDC.

The evaluation of recommender systems using a *fixed holdout set* shown in Fig. 7.2 also highlights a case in which deletion can play a large role in reducing the performance gap between two personalization technologies. Put plainly, if a user insists on getting non-personalized results from one company, deleting their data from a competing company is effective at reducing the performance gap between the two: both companies will be forced to provide the user non-personalized results.

Overall, these results suggest that while the ability to delete data can enhance the ability of CDC by groups to increase competition between ML technologies, it will only make a large difference for relatively large campaigns or for cases in which people use personalized ML technologies with new accounts. Deletion raises ethical concerns as well: there are cases in which hurting some "Large Co." may be seen as anti-social, e.g. for classification models that are well known to assist physicians in achieving better health outcomes, but there are also cases where hurting Large Co. companies may be seen as pro-social, e.g. lowering the utility of disadvantage-reinforcing credit scoring systems. We expand on these concerns below.

7.5. Discussion

In this section, we discuss the implications of our experimental results and the limitations of our study that might be addressed by future work.

As we highlighted in the Introduction, the study of power dynamics between users and tech companies using the combined lens of collective action and machine learning is relatively new. We were able to contribute to this emerging research area by leveraging simulation methods, with assumptions grounded in ongoing discussions about data leverage and ML. We have seen that CDC represents a promising and feasible means by which the public might gain additional leverage.

The potential impact of CDC could be amplified through a number of avenues, spanning design, policy, and research. In order to make CDC more broadly available, there are major opportunities to design new technologies to make CDC easier, as well as opportunities to institute policy that supports CDC. We discuss specific examples below.

7.5.1. Implications for Design and Policy

We observed that CDC can be effective in reducing the performance gap between two competitors. This finding suggests that constituencies interested in creating more competition between AI services—as an intervention aimed to prevent monopolies or as a means to increase innovation—may wish to further investigate CDC itself (e.g. conducting similar experiments to those described here) and explore avenues for making CDC easier at the grassroots level. Policymakers and advocates might push for data portability regulations, an area of growing discussion [83, 329]. Specifically, policies that make it easier for people to obtain data they generate and transfer that data to another company could make CDC easier. For instance, the EU's GDPR has a "right to data portability" – other jurisdictions might emulate or extend this idea [297].

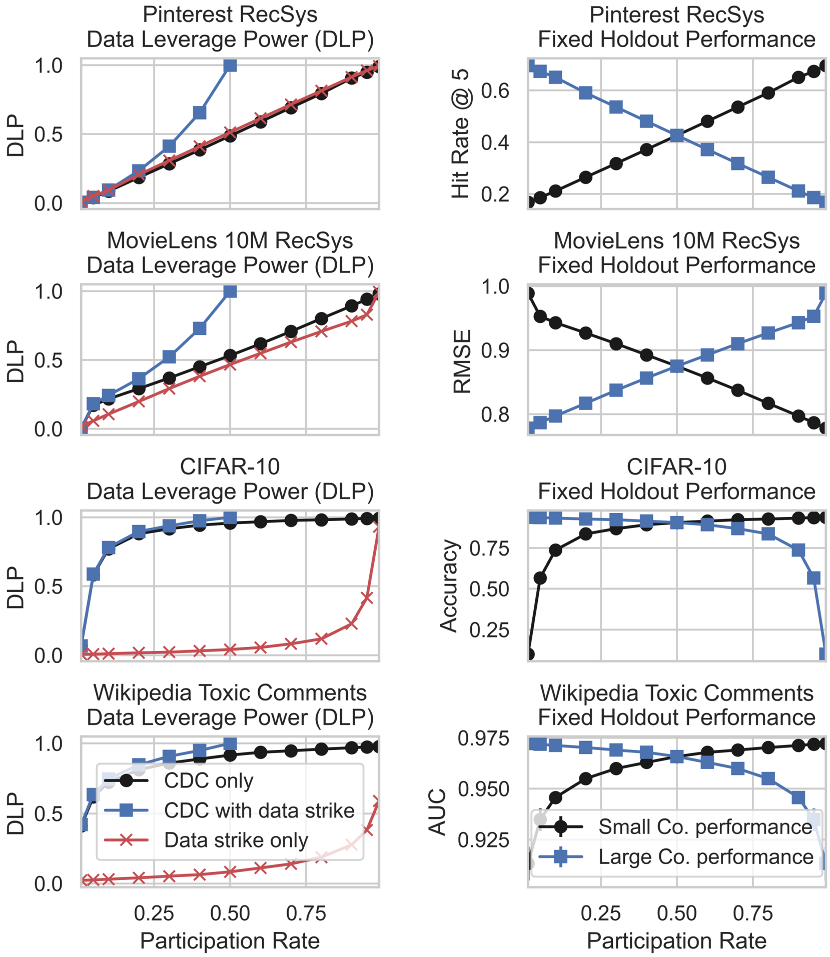


Figure 7.2. DLP results using Fixed Holdout Evaluation. The left column shows DLP plotted against participation, each row shows a different ML task, and the right column shows the task-specific performance measurement we used to calculate DLP.

Regulatory support will be particularly important for making CDC possible for datasets with complex formats. For instance, for datasets that are captured with proprietary sensors (e.g., wearable tech personal health datasets), CDC will likely remain impossible without regulation that compels companies to make data downloaded or transferrable in common (and machine-readable) format. The creation of such regulation will benefit from interdisciplinary research incorporating machine learning, law, and HCI/CSCW.

In the near-term, the technologies and companies to which CDC is applicable may be limited by the legal rights around data. In other words, while tools (e.g., scraping software from researchers) and voluntary choices by companies (e.g., Google's Takeout service) may make CDC possible today, there are certain contexts in which CDC needs legal support. In general, laws that give users more agency over the data they helped to generate will amplify the power of data leverage, whereas laws that make it harder to have agency over data will do the opposite. Data leverage will especially benefit from laws that are designed with a focus on the social and collective nature of data creation, as opposed to a framework of individual data ownership.

Software can complement regulation or support CDC in the absence of regulation. Designers might create tools that make CDC easier, such as tools that help users collect their data from web platforms in a format that is easy to share with other companies or organizations. These tools might look similar to software used to collect data for studies in social computing and CSCW (e.g. software like data scrapers and scripts that use APIs to obtain data from sources like Twitter, Wikipedia, Reddit, etc. [289]). There are also opportunities for technology designers to create tools that help organize and make visible the impact of CDC. These tools could take inspiration from technologies that support collective action (e.g. Zhang et al.'s "WeDo" [437] and Li et al.'s "Out of Site" [241]), and would aid in scaffolding the CDC process and communicating the impact of CDC to participants. Indeed, this direction would benefit especially from CSCW research around coordinating and motivating online groups (including work that has focused on peer production, e.g. [201, 215, 440, 439]). An additional synergy between the CDC concept and CSCW research is that CDC is, by its nature, online. Any techniques, tools, and strategies developed by CSCW researchers for online collective action (activism, peer production, etc.) might additionally be used to support groups who wish to engage in CDC. In the same vein, new findings from CSCW that support collective action can likely be applied to CDC in a straightforward manner.

7.5.2. Ecological validity of simulations

The conceptualization of CDC we simulated is something that users can realistically engage in today, although data strikes may not feasible in some contexts (e.g. if regulators do not enforce user requests for data deletion). Thus, our results that correspond to scenarios in which CDC participants also delete their data can be seen as measuring how much more effective CDC might be *in a world in which large-scale data deletion is possible* (and companies are forced to retrain their models regularly, protecting against "data laundering" through model parameters). In a very recent case, the Federal Trade Commission forced a privacy-violating company to delete both facial recognition data and the model trained using that data; this is a promising precedent for future data strikes [255].

An additional ecological validity concern is that expensive-to-train models, e.g. models that cost up to $245k to train [302], may be completely inaccessible to non-incumbent companies. Our experiments required repeated retraining of models, forcing us to select models that are fast to train. However, small challenger companies may also face financial constraints around model training, so by focusing first on fast-to-train models, we naturally select for approaches that small companies might realistically use. Furthermore, the datasets we investigated may be of similar size to some, but not all, models in production at tech companies. Future work that seeks to make expensive state-of-the-art model training cheaper could further widen the pool of models that are "CDC viable".

It is worth noting that some truly enormous models, such as OpenAI's recent GPT-3 language model, may be simply too expensive to ever study with the simulation approach we used here [37]. For these models, it may be possible to investigate the efficacy of CDC and data strikes through alternative approaches, such as "influence" estimation techniques from the machine learning literature [207, 409].

Finally, an important ecological validity concern is that there may be cases in which a "Small Co." and "Large Co." in the real world compete without offering comparable ML services, because the absence or presence of an ML service is a selling point (e.g. a privacy-focused Small Co. that eschews recommendation entirely). For such cases, the methodological approach of CDC simulation will not be as helpful as an analysis that focuses on other factors that govern business success.

7.5.3. Identifying "Acceptable" Performance of ML Models

One approach we used to compare ML tasks was to look at the CDC participation rate needed to reach reasonable round-number performance thresholds. In practice, there may be thresholds specific to certain contexts that require looking at traditional ML evaluation metrics, e.g., perhaps for certain users of a music recommender system, there is a certain Hit Rate at which they will stop using the recommender. By considering traditional metrics, it is possible to take such thresholds into account.

Public research that identifies these thresholds, and more generally the relationship between ML performance and downstream consequences (e.g., how performance impacts users leaving a platform, how performance maps to profit), will make it easier to plan CDC (following, for instance, Song et al.'s work on degraded search performance [372]). If CDC organizers know that DLP of 0.8 (or, returning to our results from ML-10M, an RMSE of 0.82) is "good enough" for most people, they can use simulations like ours to estimate what participation rate they need to achieve this DLP.

Studying the relationship between performance and downstream consequences will likely require proprietary data (especially in the case of performance's relationship with revenue). Furthermore, these relationships likely differ across ML tasks. Nonetheless, any findings in this area could be invaluable for informing CDC.

While offline evaluation metrics like Hit Rate are defined to correlate directly with revenue generation (i.e. each hit is meant to correspond to revenue), they do not necessarily correlate with user satisfaction. For the purposes of getting new users to join a competing platform, this means getting only part of the way towards "best-case" performance may be more than adequate to give users a satisfying experience.

7.5.4. Negative Impacts of Data Leverage

For ML technologies that are clearly identifiable as societally beneficial, exerting data leverage via data strikes good be harmful. Conversely, for technologies that are societally harmful, exerting data leverage via CDC could be harmful. Of course, the classification of beneficial vs. harmful technologies requires ongoing discussion, but research has already begun to identify some cases. Notably, Kulynych, Overdorf and colleagues' work on Protective Optimization Technologies provides an overview of various harmful ML instantiations, including discriminatory facial recognition and credit scoring that create unjust economic feedback loops [220].

Our results suggest that CDC should be preferred over data strikes for exerting data leverage against ML operators of societally beneficial technologies. Some ML models directly impact health and safety outcomes, e.g., ML models that assist doctors or operate vehicles. For such cases, using data deletion to reduce the performance of a large incumbent organization (e.g., a major hospital or transit company) could induce substantial societal harms (e.g., missed diagnoses or vehicle crashes).

There exist cases in which a. For cases in which a ML model is considered societally harmful, e.g. discriminatory facial recognition or unjust credit scoring [220], CDC does not represent a viable option. However, data strikes still face an uphill battle to break out of the flat region of diminishing returns curves. Our results suggest protest against such harmful ML models may be best accomplished through working towards regulation, perhaps in conjunction with data strikes or other types of POTs. Some jurisdictions, like San Francisco, have already moved towards such regulation, by banning the use of facial recognition by police [65]. Other jurisdictions might follow this example and directly regulate societally harmful technologies for which consumer leverage and data leverage are not well suited to address.

An additional consideration is that a small CDC beneficiary operating a less-than-ideal version of a particular technology could also induce harms. To address this possibility, a set of acceptable performance thresholds could be determined by an external body knowledgeable about task-specific metrics (e.g., a government mandated minimum precision and recall for a medical model). CDC users seeking to a support a new health start-up would likely need to contribute enough data for the start-up to meet external standards before their models could be put into use.

A final consideration is that evidence from economics literature suggests that broader data sharing may be more desirable from a consumer welfare perspective, as more companies would be able to provide high quality technologies [196]. In the extreme, if every company is subject to data strikes, AI technologies would broadly suffer. In the opposing extreme, if every company has access to data from every person in the world, AI technologies would be very accurate, but privacy would be grossly violated at a global scale.

7.5.5. Data Sharing by Corporations

We have so far framed our research by considering scenarios in which users collectively engage in CDC as a way of exerting data leverage against companies. However, firms can also share data that is broadly useful to other organizations. For instance, a company interested in social responsibility might release a labeled dataset that is useful for societally beneficial ML technologies as part of a conscious corporate social responsibility program.

We'd expect such "corporate CDC" to have similar effectiveness to the CDC we studied here: if a large company releases 10% of its "toxic comment detection" data, this may be enough data for other organizations to get 70% of the way towards best case performance on this task. Looking forward, government programs could even financially incentivize such open data initiatives as part of an effort to address concerns about the market power of tech companies.

7.5.6. Limitations and Future Work

In general, while our simulations covered a variety of tasks and we took steps to maximize the ecological validity of our simulations, there are many opportunities to extend our data leverage simulations. We used only one high-performing model and hyperparameter setup for each simulation. This means our results do not represent an upper bound for the effectiveness of CDC, as Small Co.'s data scientists would likely seek alternate models or hyperparameters that perform better for small datasets. Furthermore, there are numerous ML tasks that could be studied using simulation, perhaps with the ultimate goal of creating a "catalog" of CDC effectiveness. Such a catalog would be useful to CDC organizers, but also to policymakers interested in incentivizing corporate CDC and promoting competition around ML.

It is worth nothing that while our DLP definition is critical for allowing us to compare ML tasks with different evaluation procedures and different data sizes, a weakness of the DLP definition is that it requires careful selection of a baseline and careful consideration of what constitutes a "full dataset". Selecting an extremely weak baseline could make DLP appear exaggerated. Selecting too small or too large of a "full dataset" size could miss important parts of the DLP curve. We addressed these challenges by carefully choosing comparable baselines for each task (such that the baseline corresponds to a "low-data" or "no-data" approach that Small Co. might use when it has access to very little data) and by taking our datasets from prior work.

One way future simulations can zoom in on specific tasks would be to perform simulations that take into account the costs and rewards of successes and errors. Researchers might estimate the cost of false positives and false negatives, estimate the reward associated with successful classification, and calculate the expected total cost or reward for each organization associated with a given data leverage action (e.g. the cost to Large Co. and the benefit to Small Co.). This would move towards simulating the downstream consequences of CDC and data strikes.

Another direction for future data leverage research involves more directly modeling factors other than data leverage participation rate that facilitate success for businesses and technologies. In this chapter, we did not address the intricacies of markets, consumer preferences, the ecosystems in which tech companies operate, or the collective action processes required to organize data leverage campaigns. Each of these factors will be important for future work that advances understanding of data leverage.

Beyond simulation, other directions for advancing this research area might involve in-the-wild experiments and observational studies of users exerting data leverage. This research might be conducted, at least in part, by organizations directly affected by data leverage, e.g. companies who are the targets of protest or the beneficiaries of CDC. These organizations will likely have access to unique data on the effectiveness of data leverage.

In particular, as mentioned above, data that maps the effects of user-generated data to downstream consequences like business outcomes will be particularly valuable.

Finally, our experiments looked only at one ML task at a time. Future work should consider the interplay between datasets and ML tasks. For instance, how does CDC interact with ML pipelines and datasets that feed into multiple ML systems? Answering this question will be important for understanding the full effects of CDC and data leverage.

7.6. Conclusion

In this chapter, we proposed and evaluated **conscious data contribution**, a tactic the public might use to exert **data leverage** against tech companies to encourage them to change their behavior around key issues of interest. CDC entails users making data contributions aimed at reducing the performance gap between a large incumbent ML operator that users wish to protest and a small competitor that users wish to support. Using simulations, we measured the effectiveness of CDC in a variety of ML contexts using both a new metric called "data leverage power" and traditional ML metrics. Our results suggest that CDC represents a viable way to reduce the ML performance gap between a large incumbent and small competitor. We also observe that data deletion can enhance the effects of CDC, but the overall impact of data deletion is small compared to CDC. Overall, these results provide early information that inform the growing data leverage and provide guidance for constituencies interested in CDC.

CHAPTER 8

Data Leverage Framework

8.1. Introduction

In August 2020, the most valuable five technology companies had a total market cap of US$7 trillion [300][1]. This valuation is driven in part by large models that use data generated by the public to recommend content, rank search results, and provide many other services [40, 309, 438, 265]. More generally, lucrative technologies used by many companies rely on data generated by large groups of people to fulfill critical customer needs [40, 13, 309] and drive decision-making [41].

The reliance of powerful technologies (and thus powerful companies) on "data labor" [13, 309] by the general public presents an enormous opportunity for the public to gain more power in its relationship with tech companies. People perform data labor when they engage in the multitude of interactions with technology that generate data for firms (e.g. liking, clicking, rating, posting). By leveraging tech companies' reliance on their data labor, the public could demand changes on pressing issues [161, 220], such as diminished privacy [100, 39], the reinforcement of problematic societal biases by AI systems [2, 107, 283, 220, 7], eroded labor rights [157, 299], environmental harms [348], content moderation challenges [130], and the current imbalance in how profits from data-driven technologies are distributed between tech operators and data contributors

[1] This chapter was originally published as [410].

[309, 409, 40]. Armed with the knowledge of the importance of data contributions and the tools to action this knowledge, the public could potentially interfere with recommender systems, search engines, image classifiers, and other technologies until tech companies made changes related to these issues.

To capture the power inherent in the public's data labor, this paper introduces the concept of "data leverage" and discusses how the concept can be made operational. Simply put, data leverage refers to influence that members of the public have over tech companies because important computing technologies rely on the public's data contributions. Data leverage catalyzes power achieved by *harming* data-dependent technologies as well as power achieved by *improving* alternative data-dependent technologies and thereby creating increased competition [405]. The concept of data leverage highlights an emergent theme in the FAccT community and related areas, including human-computer interaction (HCI), social computing, society and technology studies (STS), machine learning (ML), and particularly ML research that seeks to advance fairness, justice, and a human-centered perspective (e.g. [2, 143, 32, 50]). This paper shows that this interdisciplinary lens can provide a structure for understanding and actioning an almost entirely untapped source of power that can advance a wide variety of pro-social goals. Our data leverage framework also highlights opportunities for future research and policy interventions that empower the public in its relationship with technology companies.

The contributions of this work are to (1) define data leverage, (2) provide a framework of potential "data levers", grounded in prior work that has advanced our understanding of these levers, (3) outline an initial assessment of strengths and weaknesses of each data lever in the public's "tool belt", and (4) highlight how data leverage provides important opportunities for research and new policy. Critically, research and policy can amplify data leverage and, conversely, using data leverage as a lens can raise the stakes for related research areas and policy discussions. We pay particular attention to factors that might facilitate the use of data leverage (e.g. policy interventions) or that block groups from exerting data leverage (drawing on the literature on ”non-use” of technology) [340, 428, 23, 25, 22, 26, 227].

8.1.1. Background and Definitions

Before continuing, we first present formal definitions of data leverage and supporting concepts. Note that while aiming to be comprehensive, these are working definitions. Data leverage is an emerging topic in a rapidly moving field, and we aim to advance and open the discussion around data leverage, not conclude it.

Data leverage: The power derived from computing technologies' dependence on human-generated data. Data leverage is exerted when a group influences an organization by threatening to engage in or directly engaging in data-related actions that harm that organization's technologies or help its competitors' technologies.

Data levers: The specific types of actions that individuals or groups engage in to exert data leverage. For instance, “data strikes” (see Chapter 6) are one of the data levers we discuss below and they operate by cutting off the flow of data to tech companies.

8.2. Related Work

In this section, we situate data leverage in relation to the FAccT domain, and then discuss four additional areas that contribute to the idea of data leverage.

8.2.1. Data Leverage and FAccT Research

Data leverage emerges in part from work in the broader FAccT community that has demonstrated the limitations of purely technical approaches to advancing fairness and justice in computing systems [2, 32, 107, 121, 50]. This large literature emphasizes the critical roles played by the societal context around computing systems, and has demonstrated that sociotechnical approaches are often much more powerful than purely technical approaches. Data leverage can in many ways be understood as a framework that helps us better understand data-driven technologies through a sociotechnical lens and use that lens to take action to achieve pro-social outcomes.

Data leverage is more specifically informed by Kulynych et al.'s work that proposed "Protective Optimization Technologies" (POTs) as a way to address the negative impacts of algorithmic systems and give agency to those impacted [220]. POTs allow people to contest or subvert optimization technologies, perhaps adopting techniques from data poisoning (which we further address below) [220, 393]. Data leverage and POTs are synergistic concepts, and many POTs enable people to exert data leverage.

8.2.2. Data as Labor

Data leverage is heavily informed by work that views data generated by people using computing systems as a type of labor. Building on [309], Arrieta Ibarra et al. argue that data should be considered as labor, not "exhaust" emitted in the process of using technology, and as such, should be subject to some kind of remuneration [13]. The relationship between the data-generating public and the companies that benefit from data is very asymmetric. Not only do people have very little knowledge of — let alone agency over — how data they contribute is used, but the economic winnings from powerful data-dependent technologies are reaped entirely by tech companies [309]. To mitigate this inequality, Posner and Weyl called for the formation of "data unions", which allow data laborers to collectively negotiate with technology companies [309]. Framing data as labor suggests that the current balance of power — in which users have very little power over tech companies and the technologies data labor fuels — should be disrupted to avoid negative societal outcomes and restore agency to data laborers.

The discussion around data labor has inspired work that aims to measure the economic value of data [406, 408, 207, 265]. One approach has been to look at the relationship between Wikipedia — the product of data contributions from the public — and real-world economic outcomes such as tourism and investment [173, 431]. Building on the data as labor concept, we have investigated how people might withhold or redirect their data labor to force a data-dependent organization to change its practices.

Scholars working on data feminism — an intersectional feminism-informed lens for data science — have called for more efforts to make the labor of data science visible, including the labor of data generation [81]. These scholars argue that the invisible labor of data science, much like housework, has been hidden from public view and therefore undervalued [81], and that researchers can begin to shine a light on this labor by studying and highlighting the processes of data creation (e.g. [70]). In this way, data feminism is very aligned with the ideas of data leverage; both aim to measure and make people aware of the value of previously invisible labor and ultimately reshape power imbalances.

In studying data as labor, we must consider several relevant economic properties of data. Data is nonrival — many firms can have access to the same data at once — which has implications for how firms and consumers manage data access [196]. Jones and Tonetti suggest that government intervention in around "data property rights" may be necessary to create outcomes that balance privacy and economics gains.

Data leverage must also contend with social nature of data: one user's data reveals information about other users. This has major implications for any sort of "data market" — prices will be depressed, and one person's choice to share data may harm other's welfare [4]. Data markets must also contend with the fact that firms often find unexpected uses for data [244].

In fact, some recent work has even argued the potential for information given some people to harms other constitutes "data pollution", and requires the same regulatory approaches applied to environmental pollution [29]. We argue that data leverage can help to take advantage of some of these properties in a way that interventions like data markets may not be able to.

8.2.3. Data Leverage and Technology Use/Non-Use

The data leverage concept is also informed by work from HCI and STS on technology "use", "non-use", and the spectrum of behaviors in between.

Work from Selwyn and Wyatt called attention to the need to understand people who do not use new technologies [350, 428]. Most relevant to data leverage, Selwyn documented that people engage in ideological refusal to use certain technology "despite being able to do so in practice". Further calls to study non-use in HCI and STS have been amplified in the years since [340, 24].

Use and non-use exist on a spectrum [428, 26]. People face many social and technical decisions in terms of when they will use, and stop using, a particular technology, and these decisions lead to many different forms of use and non-use [38, 345, 23]. Recently, Saxena et al. reviewed the methods for creating typologies to describe the many forms of use and non-use [343].

Many factors motivate non-use, such as exclusion [428], social capital [227], and socioeconomic factors [22, 26]. Anyone seeking to use data leverage to empower the public must contend with these factors. Attempts to support data leverage could exclude or disproportionately benefit certain groups following existing patterns in how technology excludes and benefits these groups.

One common theme in the non-use literature is that it is not easy for people to refrain from use when it comes to products that have some benefit in their life, even if the benefit(s) come with a host of long-term drawbacks. People often speak of their technology use as a type of addiction, using terms like 'relapsing" and "withdrawing" [25, 23]. Challenges also emerge related to the public presentation role of social media profiles [226]. Even if people stop using a technology, they may not necessarily delete their data. In studying individuals who left Grindr, a dating app, Brubaker et al. found that "even among those who deleted the app, only a minority tried to close their accounts or remove personal data...[putting them] in a paradoxical position of thinking they have left while their profile — or data — continues on" [38].

The non-use literature also indicates that people engage in protest-related use and non-use behaviors for reasons relating to privacy, data practices, perceived addiction, and other issues [23, 25, 242, 380]. Anyone engaging in such behaviors is a potential participant in data leverage campaigns. Casemajor et al. and Portwood-Stacer argue separately that non-participation in digital media can be an explicitly political action[47, 308]. Li et al. conducted a survey to better understand "protest users", or people who stop or change their use of tech to protest tech companies[242]. The results suggested that there is a large number of people interested in protest use: half of respondents were interested in becoming protest users of a company, and 30% were already engaged in protest use.

An important related lens is that of "refusal". Focusing on bioethics, Benjamin makes the case that broad support of "informed refusal" provides a means of developing a justice-oriented paradigm of science and technology [30]. In practice, people who engage in informed refusal are engaging in a political form of non-use, and thereby data leverage. Building on Benjamin's work, Cifor et al. and Garcia et al. describe how the notion of "critical refusal" informed by feminist scholars can be used improve practices around data [120, 58].

Overall, understanding non-use is helpful in understanding the potential impact of data leverage. As we will discuss below, data leverage raises the stakes of non-use: in some potential cases, giving non-users more power and in other cases excluding non-users.

8.2.4. Data Leverage and ML Research

Understanding the full potential of data leverage requires deep engagement with machine learning literature. Two relevant areas of ML research are those that answer questions around (1) the effectiveness of adversarial attacks on data-dependent systems and (2) the relationship between a system's performance and changes to underlying data.

There is a large literature that considers the case of adversaries attempting to attack ML systems (e.g. [20, 305, 31, 355, 354, 376, 224, 236, 148, 54, 275, 109, 239]). In early work on adversarial ML, Barreno et al. developed a taxonomy of attacks on ML systems [20]. They focused in particular on attacks in which an adversary "mis-trains" a system, which is called *data poisoning*. Data poisoning attacks against many types of ML systems have been studied in detail [20, 305, 31, 355, 354, 376]. A type of data poisoning attack that is particularly relevant to the work in this chapter is the "shilling" attack, which involves "lying" to a recommender system so that a system recommends certain products favored by the attacker [224]. Accordingly, much work has been done on counteracting shilling (e.g. [236, 148, 54, 275]), which may be of concern to groups who want to use shilling-style data poisoning attacks to exert data leverage as we describe below. Researchers have also explored advanced "data poisoning" techniques that use sophisticated methods to optimally harm ML systems [109, 239], which can be much more effective than unsophisticated attacks (e.g. providing random or average ratings to many items [224]).

Data leverage raises the stakes of the already high-stakes adversarial ML domain. This paper highlights how adversarial techniques, such as data poisoning, are not just relevant to issues of security and privacy, but also to the power dynamics between users and tech companies. While some recent work in adversarial ML has taken a political lens and highlighted real world examples of how adversarial ML can create socially desirable outcomes [7], most of the literature takes a strictly security-oriented lens.

The literature on the relationship between the amount of training data a model has access to and model performance is also highly relevant to data leverage. Many authors have found diminishing returns of additional data across many contexts and algorithms (e.g. [55, 80, 112, 169]), and some have studied techniques to address diminishing returns [34]. These findings are informative as to how effective data leverage can be.

8.2.5. Data Leverage and Data Activism

This paper builds on the literature that explores how the public can change practices of the technology industry. Data activism is a relatively new form of civic participation in response to tech companies' pervasive role in public life [17].

Currently, data activism encompasses practices that affect technology design, development, and deployment [271]. Data leverage can be seen as a subset of data activism with a specific focus on empowering the public to influence the performance of data-dependent technologies. Milan and Van der Velden provided a typology of data activism that further illustrated the specialized activities in this space — proactive and reactive data activism [271]. Proactive data activism refers to activists directly influencing software development or databases through open source projects or collaborating with institutions. A particularly relevant data activism initiative is the open data movement, which aims to democratize information that is currently only accessible to the state or businesses [149]. For example, Baack studied an open data project in Finland and highlighted the intermediary role of data activists between the public and operators of data-dependent technologies [17]. On the other hand, reactive data activism entails activists acting against data-collecting entities through adversarial behaviors such as employing encryption. Data leverage includes both types of data activism.

Data Lever Name	**Short Definition**	**Examples**
Data Strike	withholding or deleting data	leaving a platform, installing privacy tools
Data Poisoning	contributing harmful data	inputting fake data in user profile, clicking randomly, manipulating images
Conscious Data Contribution	contributing data to a competitor	switching to a new search engine, transferring photos to a new platform

Table 8.1. The three data levers in our framework, short definitions for each, and several examples of each.

Equipped with the knowledge and expertise to understand data's role in computing, researchers can provide the public with valuable information to identify and employ effective data leverage practices. Work on data activism has unveiled a rich space to improve data practices [68]. In particular, Lehtiniemi and Ruckenstein called for "linking knowledge production to data activism practice" to gain a comprehensive understanding of data's role in the public sphere [238].

8.3. Framework

In this section, we describe our framework for data leverage in detail . The framework — and this section — is organized around the three data levers we identified. For each lever, we first define the lever and any variants, and do so grounded in past work viewed through our data leverage lens. We then provide practical examples of each data lever and describe the likely factors that will govern the effectiveness of the lever. Table 8.1 lists the data levers, their definitions, and several examples of each.

8.3.1. Data Strikes

The first of the data levers we will consider are data strikes. Data strikes involve a person withholding or deleting data to reduce the amount of data an organization has available to train and operate data-dependent technologies. Although the term data strike is relatively new, the concept builds on the well-studied practices of stopping or changing technology use as a form of protest, as discussed in Related Work. For instance, groups have participated in prominent boycotts against companies like Facebook and Uber [351, 144]. In another example, people use ad blocking software to deprive companies of data about the success of their ad placements [43].

8.3.1.1. Data Strike Variants. The most basic form of a data strike is a *withholding-based data strike*. In some cases, users can withhold data by reducing or stopping their technology use, or by continuing to use a technology with privacy-protection tools (e.g. tracking blockers [260]). In jurisdictions that allow people to delete their past data (using laws like the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) [415, 339]), users can also engage in *deletion-based data strikes*. The effectiveness of such strikes will depend on how well regulations can force companies to regularly retrain or delete their models (so as to remove weights learned using now-deleted data). There is some precedent from the U.S. that model deletion can be enforced: in 2021, the Federal Trade Commission forced a company to delete both customer photos and the facial recognition models trained on the photos [255].

Data strikes can be further categorized based on their coordination requirements. Data strikes (and ther other data levers we will describe below) are likely possible without serious coordination, given the success of hashtag activism [186] and other forms of online collective action that operate without central leadership [259]. For instance, people wanting to start an informally-organized data strike might simply make a call for others to delete as much data as they are willing. However, “targeted” [20] data strikes have the potential for a group of data strikers to achieve disproportionate impact [407]. Following Barreno et al.’s definition of targeted attacks on ML systems, a targeted data strike might encourage participants to delete specific data points or recruit particularly valuable participants. For example, data strikers could try to reduce performance for a specific genre of movie recommendations, while leaving performance for other genres untouched [407]. Leaders might also recruit specific users to join their data strike – power users have disproportionate influence on systems [425, 424, 104] and withholding or deleting their data may be more impactful.

8.3.1.2. What Do Data Strikes Look Like in Practice? To understand what data strikes will look like, we can gain insight from the non-use literature described above. An individual that chooses to use a platform less frequently or avoid a feature of that platform reduces the amount of data they help to generate. In this way, a person’s choices about use and non-use affect how much data that person generates. Research in the use and non-use domain has provided empirical examples of what could be conceptualized as data strikes against Facebook and Twitter [23, 25, 22, 26, 345, 308].

Privacy and surveillance research also lends itself to uncovering privacy-focused behaviors that can be seen as data strikes. One prominent example is that many people use anti-tracking browser extensions that limit the amount of data online trackers collect [242, 260, 43]. Studies on algorithm transparency also provide evidence suggesting that people engage with data strike-like behaviors because of dissatisfaction with algorithmic system opacity, such as ceasing producing reviews for review platforms [105, 106]. Additionally, research on online communities presented case studies of both Reddit moderators and community members striking by disabling and leaving their communities [261, 281].

8.3.1.3. How Can Data Strikes Be Effective? A data strike can be evaluated based on the importance of the data that "goes missing" in terms of how that data affects relevant data-dependent systems. Said another way, does the missing data noticeably degrade a system's performance, move a classifier's decision boundary (or hyperplane, etc.) in a meaningful way, or otherwise change outputs?

To understand the effectiveness of data strikes, researchers and strike leaders might look to research on data scaling and learning curves, which describes the relationship between ML performance and the amount of data available for training (e.g. [55, 80, 112, 169]). Findings from this literature could be used to predict the effectiveness of a strike, as in prior work which explicitly simulated data strikes [407, 405].

If researchers have shown a model needs a certain number of observations in its training set to be effective (e.g. [55]), data strike organizers could use that research to guide their strike, for instance by setting a goal for participant recruitment.

In summary, data strikes are a data lever available to anyone who can withhold or delete their data. While a new concept, research in HCI, privacy, machine learning, and related fields can help us to understand what data strikes will look like and how effective they might be.

8.3.2. Data Poisoning

A data poisoning attack is an adversarial attack that inserts inaccurate or harmful training data into a data-dependent technology, thereby encouraging the model to perform poorly [20]. While data strikes harm performance by reducing the amount of available data, data poisoning harms performance by providing a technology with data that was created with the intention of thwarting the technology. A relatively accessible way that users can engage in data poisoning is simply by leveraging standard technology features in a deceptive manner. For instance, someone who dislikes pop music might use an online music platform to play a playlist of pop music when they step away from their device with the intention of "tricking" a recommender system into using their data to recommend pop music to similar pop-hating users. Other straightforward examples include the coordinated effort to create sexually explicit Google search results for former U.S. Senator Rick Santorum's name [129] and coordinated campaigns to use fake reviews to promote certain products [224]. As we will describe below, very sophisticated variants of data poisoning that draw on state-of-the-art machine learning research are also possible.

8.3.2.1. Data Poisoning Variants. Data poisoning is familiar to the ML community through adversarial ML (see e.g [20, 305, 31, 355, 354, 376]) and obfuscation (see e.g. [39, 176]). This means data poisoning organizers can benefit from the knowledge produced through this body of research.

There are many ways an individual alone can engage in data poisoning. The techniques for obfuscation described by Brunton and Nissenbaum are accessible means of data poisoning for individuals. For instance, users might trade accounts (drawing on Brunton and Nissenbaum) or fill in parts of their profile with fake information [56]. As another example, past work has studied attacks that involve following certain Twitter users to throw off Twitter's profiling [279]. These approaches are generally available to an individual acting alone.

The distinction between coordinated data poisoning attacks and uncoordinated attacks is important. Typically, adversarial ML papers frame data poisoning as a contest between a single attacker (which could be an organization) and a defender/victim. In a coordinated data poisoning attack, however, the attacker is an organized collective.

To execute a coordinated data poisoning attack, it will be necessary to find the appropriate technique for a particular technology. Organizers can look to taxonomies in the adversarial ML literature to see what knowledge an attacker requires and what specific systems are vulnerable to attacks [20, 305].

Shilling attacks are a data poisoning variant that focuses on manipulating specific system outcomes rather than general performance degradation [224, 236, 148]. Unlike other poisoning attacks, this type of data leverage manipulates a system to favorably recommend a product that may not actually be high in quality or popularity, i.e. putting "lipstick on a pig". Shilling can be defended against with systems that identify and remove fraudulent or false reviews [292, 243], but these systems themselves may be vulnerable to data poisoning and data strikes.

As with other forms of data leverage, data poisoning applies more generally to any data-dependent technology, not just to ML systems. For instance, Tahmasebian et al. provide a taxonomy of data poisoning attacks against crowdsourcing-based "truth inference" systems [383], e.g. a system that aims to use crowdsourced data to ascertain the true number of cars on a road. Generally, any system that makes or uses estimates about a population can be compromised by sampling poisoned data.

8.3.2.2. What Does Data Poisoning Look Like in Practice? Almost any data-driven technology is vulnerable to deceptive interactions from users, and there are numerous ways to engage in data poisoning in practice. In the wild, there are a wide range of behaviors that constitute data poisoning attacks. Examples include Uber and Lyft drivers providing false information about their availability [237] and internet-browsing user using software to automatically click ads [176].

The most accessible form of data poisoning involves a person using technology in a deceptive manner, e.g. by lying about their personal attributes, watching videos they dislike, or searching for content they are not interested in. They might even use deception-support tools like the location-spoofing software conceptualized by Van Kleek et al. to engage in "computationally-mediated pro-social deception" [399].

By combining findings and tools from HCI and ML, more complex forms of data poisoning may be possible. Users might employ tools like browser extensions (following Li et al. [241] and Howe and Nissenbaum [176]) or web platforms (following [437]) that help them participate in coordinated data poisoning with sophisticated means of producing poisoned data (e.g. [355, 109]). For instance, one could imagine a data poisoning platform, modeled on existing social computing platforms [215], that provides users with bespoke poisoned data that they can contribute to a data poisoning attack. In such a platform, users could upload images poisoned with pixel-level manipulation to spoof image recognition systems, or take suggestions of content to interact with so as to fool recommender systems.

"Data poisoners" might even take inspiration from recent research on what are known as "adversarial evasion attacks" [355], attacks that help users protect their own images from facial recognition systems (i.e. "evade" the system [305]). Shan et al. show that their tool, Fawkes, can imperceptibly alter images so that state-of-the-art facial recognition cannot recognize the altered images [355]. Such tools might be adapted for data poisoning purposes.

8.3.2.3. How can Data Poisoning be Effective? There are several reasons to believe data poisoning might be a powerful source of data leverage. Recent work on sophisticated data poisoning suggests that very small amounts of poisoned data (e.g. using less than 1% of a training set in work from Geiping et al. [123], using 3% of a training set in work from Steinhardt et al. [376]) can meaningfully change the performance of a classifier. Even unsophisticated data poisoning (e.g. playing music one does not actually enjoy) by a majority of users could so completely poison a dataset as to make it unusable.

Progress in adversarial ML could actually end up reducing the public's poisoning-based data leverage, in which case non-poisoning data levers would become more important. Fundamentally, data poisoners are engaging in a contest with data scientists. This means any data poisoning technique runs the risk of becoming outdated — if a company's data scientists find or invent a defense, the public might lose leverage [376, 355].

Another interesting outcome of data poisoning is its potential conversion to a data strike. In the case where an organization can detect and delete poisoned data, data poisoning reduces to a data strike. Detectable data poisoning could even be used to replicate a deletion-based data strike. For instance, search engine users could use data poisoning tools such as AdNauseum [176] — which clicks all ads in a user's browser — to effectively make their ad click data useless, forcing the search engine operator to delete it.

In general, to harm a tech company, data poisoning involves deception and requires affecting the experiences of other users of a platform. Consider someone who lies on a dating site, a surprisingly common phenomenon [154, 390]. The user may protect their privacy, but will also poison their own recommendations (e.g. for romantic partners) and make others' dating experiences worse off. The same logic applies to recommendations for friends, videos, and other goods.

A critical challenge for data leverage will be navigating ethical and legal challenges around when data poisoning is acceptable [399, 39, 355, 123]. Whether a particular instance of poisoning is interpreted to be political dissidence or sabotage depends on the society where it is enacted and on case-by-case specifics. For instance, in some cases existing laws around computer abuse or fraud may come into play, such as the United States' Computer Fraud and Abuse Act (CFAA) [1, 181].

8.3.3. Conscious Data Contribution

The above tactics operate by harming, or threatening to a harm, a given data-dependent technology. However, there are cases for which harmful tactics are not a good fit. For instance, perhaps users do not have the regulatory support needed to delete past data [411] or a new technique for detecting poisoned data foils their poisoning attack. Harmful tactics may also be undesirable because an organization's technologies may actively provide benefits to others (e.g. a ML model that is well known to improve accessibility outcomes).

“Conscious data contribution” (CDC) [405] is a promising alternative to harm-based data leverage. In CDC, instead of deleting, withholding, or poisoning data, people give their data to an organization that they support to increase market competition as a source of leverage. People using CDC for data leverage are similar to people engaging in “political consumption” [210], but instead of voting with their wallet, they vote with their data. An exciting aspect of CDC is that while small data strikes struggle to put a dent in large-data technologies because of diminishing returns, CDC by a small group of users takes advantage of diminishing returns and provide a competitor with a large boost in performance. We return to this point later in our assessment of data levers.

8.3.3.1. CDC Variants. Variants of CDC closely mirror variants of data strikes because CDC in a sense is the inverse of data strikes — where data strikes take, CDC gives.

The easiest way to engage in CDC is to simply start using another technology with the intention of producing useful data for the organization that operates the technology. Sometimes, these CDC campaigns may also involve a data strike if a user moves from one platform to another, for example abandoning Google and moving to DuckDuckGo.

In jurisdictions where data portability laws [297] require that companies allow users to download their data, users can engage in CDC by downloading data from a target organization and contributing it to the organization’s competitor. Many services already allow users to download or otherwise access some of their data contributions, but the usefulness of currently exportable data to other companies remains to be seen [187].

Similarly to how coordinated data strikes and data poisoning might seek to hurt a particular aspect of a technology, coordinated CDC can enhance specific aspects of a technology’s capabilities. In a coordinated CDC campaign, organizers might instruct participants to donate specific types of data, or organizers might seek out specific people to join a campaign, in an effort to focus on contributions towards a specific goal. For instance, in the recommendation context, CDC leaders might seek out comedy movies fans to contribute data to a comedy movie recommender, instead of trying to solicit data about every movie genre. Recommender system researchers have shown that allowing users to filter out their old data could actually improve recommendations [420], so CDC participants could even use filtering to further target their data contributions.

The idea of CDC has complex relationships with various proposals for "data markets" [4, 191], which are designed to give people the ability to sell data that they generate. While data markets allow users to participate in a form of CDC by giving them choices about to whom they will sell data, people may prioritize their personal economic incentives over attempts to gain leverage. A major issue with CDC via data markets is the fact that any data with a social component often has information about more than one person[29, 4], which could make it legally and ethically tricky to handle data via markets.

8.3.3.2. What Does CDC Look Like in Practice? As mentioned above, providing data to online platforms can be a form of Conscious Data Contribution if users aim to increase the performance of these technologies relative to their competitors. As such, there are many existing examples of what CDC might look like in practice.

Cases in which users switch platforms provide one set of examples. In 2015, many Reddit users expressed dissatisfaction with the platform and eventually migrated to alternative platforms such as Voat and Snapzu [281]. In doing so, these users performed an act of CDC, explicitly supporting Reddit's competitors. Past work suggests that migrations are an especially likely form of CDC, because an individual user's choice to move platforms as part of a CDC campaign may lead to people in the user's social network also migrating [119, 221]. Where social networks create friction against data strikes, they can help to drive CDC.

Many research initiatives involve collecting volunteered data, which in certain cases could provide opportunities for CDC. In Silva et al.'s study, people contributed data about their Facebook political ads to researchers for monitoring and auditing purposes [363]. While research studies on their own are not necessarily CDC (though they could be, if the research helps support competitive data-driven technologies), they can often provide a good example of how CDC might be implemented.

Other types of data sharing and generation can also be CDC. For instance, the "data donation" concept explored in the context of data ethics [310] could be used for CDC. In some cases participation in human computation [313], crowdsourcing systems, and other social computing platforms [215] could qualify as CDC. For example, under our definition, people who choose to contribute data to protein-folding games could be engaging in a form of CDC [98], with the potential to exert leverage against other organizations that benefit from protein folding models.

8.3.3.3. How Can CDC Be Most Effective? CDC has a lower barrier to entry than data strikes and data poisoning because it is possible to engage in CDC without completely stopping use of an existing technology. Despite this advantage, a critical question for any CDC effort will be how much leverage "helping a competitor" exerts on the target. For instance, a group of CDC users might be able to successfully improve the ML technologies of a small startup that is competing with a major platform. However, even with improved data-driven technologies, other factors like access to capital and switching costs for users might prevent the startup from competing effectively with the original target of leverage, thus reducing the chance that the original target changes their practices. In some cases, standing up a viable competitor that has better practices could be the end goal of a CDC campaign, even if does not directly harm another company. By supporting a new viable contender, CDC participants can effectively change the overall relationship between the public and technology companies.

Like data strikes, a key determinant of the effectiveness of CDC will be the level of participation. The more people that participate in CDC, the more powerful it will become, and the degree of effectiveness can be estimated using ML findings and methods as we discuss below. A critical distinction between data strikes and CDC is that while small data strikes may struggle to escape the flat portion of ML learning curves, CDC by a small group can actually provide a huge boost in ML performance to a small organization. We expand on this comparison in the following Assessment section.

8.4. Assessing Data Levers

In this section, we use three axes to evaluate strengths and weaknesses of each data lever: the *barrier-to-entry* to use a data lever, how *ethical and legal considerations* might complicate the use of a data lever, and finally the *potential impact* of each data lever. Table 8.2 contains a brief summary of our assessments.

8.4.1. Barriers to Entry

In general, CDC has the lowest barrier to entry of the data levers we identified. This is because CDC does not require stopping or changing the use of existing technologies, which

Table 8.2. Summary of key points from our assessment of data levers.

Data Lever	**Barriers to Entry**	**Legal and Ethical Considerations**	**Potential Impact**
Data Strike	*moderate*: non-use is challenging; hurts participating users; need for privacy tools	*lower*: need privacy laws to delete data; harming tech may be undesirable	*moderate*: small group has small effect; large group can have huge impact
Data Poisoning	*higher*: time/effort/bandwidth costs; may require ML knowledge; may require extra coordination	*higher*: potentially illegal; harming tech may be undesirable; inherently deceptive	*moderate*: small group can have huge effects; if caught, "reduces" to a strike; constant arms race
Conscious Data Contribution	*lower*: can continue using existing tech	*moderate*: potential to improve harmful technologies, privacy concerns of sharing data	*moderate*: small group can have large effects, large group faces diminishing

prior work discussed above indicates can be challenging (e.g. [23, 25, 345]). A person can continue using existing technologies operated by an organization against which they want to exert leverage while engaging in CDC [196, 405]. The main barriers to transfer-based CDC are regulatory and technical. Do laws help people transfer their data [297] and do tools exist to make data transfer realistic?

The barriers to entry for data strikes are more substantial then those for CDC and less substantial than those for data poisoning. While participating in a data strike disrupts a user's access to online platforms, strikes do not necessarily force a user to stop using a platform like a traditional boycott would. For instance, a user who relies on Facebook to communicate with family members could stop engaging with sponsored content on Facebook but continue messaging their family members. An Amazon user might continue buying products but stop leaving ratings and reviews. An important downside of data strikes is that they hurt the performance of technologies for participating users. By cutting off data contributions, an individual often reduces their own ability to benefit from a system. The effect of a data strike will almost always be most pronounced on the strike participants (see Chapter 6).

The barriers to entry for each data lever are also contingent on the bandwidth available to potential participants and any potential data caps or data charges they have. Data strikes are likely the least limited by bandwidth (although striking against an Internet provider, e.g. Facebook Free Basics, could be challenging [352]). In places where the Internet is easy to access and has relatively high data caps, poisoning data by letting music stream for hours or actively manipulating multimedia may be accessible. In contrast, in places where Internet access is limited [86], poisoning data may be difficult if not impossible. Similar dynamics likely will apply to CDC: data caps could stifle efforts to engage in CDC.

Many of the barriers to entry discussed above are not equally distributed across different populations, and this means that different populations likely have differing access to data leverage. For instance, with regards to data poisoning, the time available to expend the necessary effort and/or the skills necessary to do so will limit the ability of many populations to engage in data poisoning. Those most positioned to perform data poisoning attacks are ML researchers, technologists, and others with strong technical skills, an already relatively privileged group. Nonetheless, members of this group could use their powerful position for the benefit of people without these advantages (there is precedent of tech worker organizing along these lines, e.g. `https://techworkerscoalition.org`).

Turning to coordination, data leverage campaigns will differ in their coordination needs, with greater coordination requirements raising the barrier to entry for all three data levers. Large-scale data leverage is possible without formal organization: boycotts using Twitter hashtags provide real-world examples [186]. However, certain data levers require especially well-coordinated effort to see impact, e.g. sophisticated data poisoning [109].

8.4.2. Legal and Ethical Considerations

Data leverage organizers may face legal and ethical challenges. Withholding-based data strikes face the fewest of these challenges. These data strikes require almost no regulatory support as users can simply cease using platforms (keeping in mind the differential barrier to entry concerns discussed above). Deletion-based data strikes require a right to deletion and a guarantee that companies are not *data laundering* by retaining model weights trained on old data [255].

The legality of data poisoning is likely to remain an open question, and interdisciplinary work between computer scientists and legal scholars will be critical to understand the legal viability of data poisoning as a type of data leverage (and to do so in different jurisdictions). Arguments about the ethics of obfuscation (which itself can be a form of data poisoning) raised by Brunton and Nissenbaum apply directly to the use of all types of data poisoning [39]. Participants must contend with the potential effects of dishonesty, wastefulness, and other downstream effects of data poisoning. For instance, there are many harms that could stem from poisoning systems that improve accessibility, block hate speech, or support medical decision-making.

Interesting legal and ethical questions also emerge around CDC. Notably, if a certain data-driven technology is fundamentally harmful and no version of it can meaningful reduce harms (as can be argued for e.g. certain uses of facial recognition [2, 121]), CDC will effectively be neutralized.

Another challenge specific to CDC is that there is the potential that data contributions by one person might violate the privacy of others, as data is rarely truly "individual" [4, 29]. For instance, genetic data about one individual may reveal attributes about their family, while financial data may reveal attributes about their friends. On the legal front, CDC often requires either regulatory support in the form of data portability laws or data export features from tech companies.

8.4.3. Potential Impact

Data strikes and data poisoning harm data-dependent technologies, while CDC improves the performance of a data-dependent technology that can then compete with the technology that is the target of data leverage. We can measure potential impact in terms of performance improvement/degradation, as well as downstream effects (e.g. performance degradation leads to users leaving a platform). Ultimately, we are interested in how likely a data lever is to successfully change an organization's behavior with regards to the goals of the data leverage effort, e.g. making changes related to economic inequality, privacy, environmental impact, technologies that reinforce bias, etc.

ML performance exhibits diminishing returns; in general, for a particular task, a system can only get so accurate even with massive increases in available data. As such, when an organization accumulates a sufficient amount of data and begins to receive diminishing returns from new data, that organization is not very vulnerable to small data strikes. Such strikes will — broadly speaking — only unwind these diminishing marginal returns. To a company with billions of users, a (relatively) small data strike simply may not matter.

The potential impact of data poisoning is also enormous: a large-scale data poisoning attack could render a dataset completely unusable. This approach is also appealing for bargaining: a group could poison some data contributions, and make some demand in return for the “antidote”. However, the enormous corporate interest in detecting data poisoning means that the would-be poisoners face a constant arms race with operators of targeted technologies. In the worst case scenario, they will be caught, their poisoned data deleted, and the end effect will be equivalent to a data strike.

CDC campaigns, which improve technology performance, operate in the opposite direction of data strikes. Small-scale CDC could be high impact: about 20% of the users of a system could help a competitor get around 80% of the best-case performance [405]. On the other hand, once returns begin to diminish, the marginal effect of additional people engaging in CDC begins to fall.

Given the current evidence, we believe that the data levers we described have a place in the tool belt of those seeking to change the relationship between tech companies and the public. A critical challenge for data leverage researchers will be identifying the correct tool for a specific job. Based on the technologies a target organization uses, a realistic estimate of how many people might participate in data leverage, and knowledge about the resources available to participants, which data lever is most effective?

8.5. Discussion

In this section, we discuss questions associated with data leverage that lie beyond the bounds of our current framework. We first discuss the key question of who might expect to benefit from data leverage, and highlight how data leverage might backfire. Next, we summarize key opportunities for researchers, particularly those working in or around FAccT topics. Finally, we summarize opportunities for policy that can amplify and augment data leverage.

8.5.1. Who Benefits from Data Leverage?

Researchers, practitioners, activists, policymakers and others interested in studying, supporting, or amplifying data leverage to reduce power imbalances must contend with unequal access to data leverage. As discussed above, there is strong reason to expect that inequalities in access to data leverage mirror known patterns in access to technology and other sources of power more generally [2]. However, our framework suggests that data poisoning and CDC in particular might allow small groups to have disproportionate impacts. A group of users with needs not currently met by existing technologies might engage in CDC to support a competitor to existing tech companies, or use sophisticated data poisoning techniques that require coordination and knowledge, but not mass participation. Researchers can play an active and critical role by developing tools and promoting policy that widely distributes the ability to participate in data leverage efforts and receive benefits from data leverage. Future work may also need to contend with the possibility of organizations counteracting data leverage, e.g. removing access to publicly available data to maintain a dominant market position.

8.5.2. Data Leverage and Data in the Commons

Many lucrative data-dependent technologies rely on “commons” data (e.g. Wikipedia and OpenStreetMap) in addition to the largely proprietary types of data we have discussed so far (e.g. interaction data, rating data). The same is largely true for a variety of data sources that are privately-owned but are a sort of de facto commons for many purposes (e.g. Reddit data, public Twitter posts). Examples of commons-dependent technologies include large language models (e.g. [37]), search engines (e.g. [408, 265, 404]), and a variety of geographic technologies (e.g. [194]). Commons datasets have also been instrumental to the advancement of ML research (e.g.[27, 95]).

How can we view the widespread dependence on commons datasets through the lens of data leverage? Adopting a narrow perspective, all three data levers can certainly be employed using data in the commons. In fact, doing so might be a very effective way of exerting data leverage against a large number of data-dependent technologies at once. For instance, through poisoning (i.e. vandalizing) Wikipedia, one can negatively affect a wide variety of Wikipedia-dependent technologies including Google Search, Bing, and Siri [408, 265, 404]. Indeed, this has already been done with humorous intent a number of times (e.g. [423]). One could similarly imagine organizing a “data strike” of sorts in Wikipedia or OpenStreetMap that sought to ensure that a certain type of information does not appear in these datasets.

That said, from a broader perspective, it is very likely that data poisoning and data strikes using commons data will do substantially more harm than good. For instance, a concerted effort to vandalize (i.e. poison) Wikipedia will cause substantial damage: it would harm Wikipedia readers across the world and would affect technologies operated by non-targeted organizations in addition to those operated by targeted ones. A similar case could be made for most data strikes.

CDC in the context of commons data presents a more complex set of considerations. Indeed, contributing to a commons dataset like Wikipedia can in some ways be understood as a type of CDC as it helps smaller organizations as well as larger ones. However, an important consideration here is that the ability to make use of commons datasets in data-driven technologies is gated by capital. A salient example is GPT-3, OpenAI's high-profile language model that uses training data from sources like Wikipedia and Reddit [37]. The unprecedented computing power needed to train GPT-3 highlights how the data labor that improves Wikipedia and Reddit can disproportionately benefit organizations with enormous resources. An unfortunate reinforcing dynamic regarding data leverage and commons data thus emerges: while a huge number of organizations and individuals stand to be harmed by any sort of poisoning attack or strike on commons data, large and wealthy firms often stand to benefit disproportionately from improvements to these data. Future work that focuses on efficient training, smaller models, and related goals can help to mitigate this particular concern. Similarly, efforts to open-source models themselves (e.g. share model weights) could also help.

8.5.3. Can Data Leverage Research Backfire?

We have presented data leverage as a means to empower the public to address concerns around computing systems that exacerbate power imbalances and create negative societal outcomes. However, research, tools, and policy intended to help data leverage achieve these goals could do the opposite by empowering groups to perpetuate inequalities and, therefore, achieve socially harmful outcomes. For instance, hate groups take advantage of "data voids" in search engines to engage in what can be understood as data poisoning attacks by inserting hateful content and influencing model development [133]. Why wouldn't these groups also try to use other types of data leverage for similar ends?

There are no clear-cut ways to eliminate these risks, but there are steps that data leverage researchers can take to avoid a "backfire" outcome. When designing tools to support data leverage, designers might consider heuristic preventative design from Li et al. [241] and try to make harmful uses of a technology more challenging. For instance, a data poisoning tool might only help users poison certain types of images known to be important to a particular company or technology. Designers should also consider the principles of data feminism [58, 120], including those that emphasize challenging existing hierarchies, embracing pluralism and context, and making labor visible.

8.5.4. Research Opportunities for Data Leverage

The concept of data leverage presents exciting research opportunities for many fields. Researchers in FAccT, ML, HCI, STS and related areas in particular have unique opportunities to amplify data leverage.

Data leverage presents a new way of exerting pressure on corporations to make important changes. Most relevant to the FAccT community, this might involve exerting leverage so that a tech company stops the use of a harmful algorithm [220], or pushing for new economic relationships between data contributors and AI operators in which the benefits of AI are shared more broadly (see Chapter 9). Data leverage thus presents a novel avenue for researchers to actively pursue pro-social research roles and goals [2].

There is enormous potential to support data leverage with ML research methods. Using simulations and small-scale experiments, future work could build a catalog of results that activists could draw on to make predictions about the effectiveness of a particular data lever in a particular context, such as "if we get x participants to engage in a data strike against technology y, we can expect to bring down the accuracy of technology by z%, which will likely be enough to encourage company c to make the changes we are demanding". As data leverage becomes more mainstream, there may also be opportunities to study real-world examples and answer key questions such as: What are the downstream effects on revenue, user retention, and actual changes in company behavior?

Future design work could build upon the collective action literature and develop tools to coordinate efforts to use data leverage. For example, because collective action's progress is often opaque to individual participants and this can negatively impact engagement, future work may adopt tactics from "boycott-assisting technologies" [241] and display the impact of the public's data strike or poisoning (e.g. this technology has lost 3% of data). Such tools could also support automating data strikes or data poisoning, similar to AdNauseam Howe and Nissenbaum, to lower the barrier to entry for the public.

In addition to data strikes and poisoning, the computing community can also support CDC by addressing data compatibility and portability issues across platforms and technologies. Data generated by users are often highly platform- and/or technology-dependent. For example, ratings for the same restaurant or hotel may vary significantly across review platforms [105, 240]. Directly transferring data from one technology to another as an act of CDC may run into compatibility issues and even negatively affect the recipient's performance. There is a need for researchers and practitioners to develop software that automatically translates data generated using one technology into data can truly benefit another technology to maximize the success of CDC-based approaches.

Researchers should also seek to better understand the full set of societal impacts that would result from the widespread use of data leverage. As we have discussed above, we hypothesize that the direct effects of actioning data leverage will often involve broadly positive societal impacts, e.g. improved privacy, better distribution of the economic benefits from AI systems, more democratic governance of AI systems. However, the second- and greater-order effects of these changes are more difficult to assess, and even some direct effects may be negative in some cases as highlighted previously. More generally, data leverage defines a pathway to altering power structures in the current computing paradigm. Alterations of power structures in such a complex sociotechnical environment will almost certainly lead to complex outcomes, and more research will be needed to understand these potential outcomes.

8.5.5. Policy Opportunities for Data Leverage

Data leverage stands to benefit heavily from regulatory support. As such, data leverage research should be deeply engaged with policy by highlighting regulatory approaches that are likely to amplify the power of data leverage and address its potential negative impacts. Our taxonomy only scratches the surface of how policy may support data leverage; we are excited for this important direction of future work.

Following directly from our assessment of data levers above, we suggest a variety of ways policy can support data leverage:

- Data portability laws will directly enhance CDC, enabling users to contribute data they helped generate in the past.
- Right-to-delete laws will enhance data strikes, assuming these laws also account for the possibility that companies might "launder" deleted data in model weights.
- Data transparency laws that make data collection more apparent may help foster support for data leverage movements.

We note that these policy suggestions are generally aligned with policy aimed at addressing privacy concerns. This suggests a potential "win-win" situation, in which policy simultaneously supports consumer privacy and enhances data leverage.

Expanding on the above points about data portability and right-to-delete laws, policy also offers the potential for making it easy for individuals to use multiple data levers in conjunction with one another. As mentioned above, there are natural connections between data strikes and CDC: by moving from one platform to a new platform, a user can take advantage of both data levers. However, through regulatory support, it may be possible to engage in much more elaborate combinations of data strikes and CDC, for instance deleting only certain pieces of data and transferring over other pieces of data.

8.6. Conclusion

In this Chapter, we presented a framework for using "data leverage" to give the public more influence over technology company behavior. Drawing on a variety of research areas, we described and assessed the "data levers" available to the public. We highlighted key areas where researchers and policymakers can amplify data leverage and work to ensure data leverage distributes power more broadly than is the case in the status quo.

CHAPTER 9

Data Dividends

9.1. Introduction

In response to the concern that intelligent technologies – i.e., artificial intelligence (AI) and machine learning (ML) – are contributing to economic inequality [42, 40, 229, 331, 332, 69], there have been prominent calls to implement some form of "data dividend" – a program to compensate members of the public for their data contributions – to share the winnings of these technologies more broadly. For instance, the governor of California directly called for investigation of a data dividend [396], related conversations are happening in jurisdictions like West Virginia, Colorado, and Canada [250], and data dividends have begun to attract research attention [28, 409, 412, 111]. Given that excessive economic inequality is associated with major societal challenges, including political instability, financial crises [8], reduced national economic growth [381], and even public health harms [73], changes in the distribution of wealth caused by AI may create downstream harms that could outweigh the many possible benefits of AI. Within the human-computer interaction community, mitigating computing's effect on economic inequality has been proposed as the "HCI problem of our time" [225, 162], and data dividends can help to support initiatives within HCI to address social inequality [99] and work towards a more just world [114]. Within artificial intelligence, this topic is being cited as a top research challenge for the field [6]. As such, there is rich potential for research in the growing human-centered artificial intelligence (HCAI) field to lead the charge in tackling this high-stakes challenge.

The data dividend idea refers to giving people a share of the profits from an intelligent technology when their data contributions have been used to train that technology [180, 250, 396, 111]. Although similar in spirit to well-studied crowdwork markets for data (which have issues related to working conditions and pay [10, 157]), data dividends offer a novel approach for addressing computing-induced economic inequality: they are retroactive and thus well-suited to rewarding passively collected data generated during consumptive activities (e.g., rewarding people for behavioral logs that improve a search engine or commercial recommender system). Data dividends can be used to compensate people for data contributions that have already been incorporated into live AI systems, which is especially important in light of the growth of large generative models like "GPT" [37] and Dall-E [315] that rely on massive scraped datasets (with content like social media posts or open source code contributions having been generated well before the corresponding AI technologies were developed) and the continued dominance of search, recommendation, and personalization systems that rely on large-scale behavioral data [262].

To quote California's governor, the mandate of a data dividend is to help consumers "share in the wealth that is created from their data" [76], which implies a need to figure out *how much wealth* is being created from the people's data. This mandate could be interpreted as calling for very coarse value estimation (i.e., answering "what is the aggregate value created by all residents of California?"), but could also be interpreted as calling for very fine-grained value estimation (i.e., answering "what is the value created by a particular individual?"). In many discussions, the second framing, which is "meritocratic" and individualistic, has been influential. There is an implicit argument that payments should be made to individuals, and payments should depend on each individual's contribution to AI performance (see e.g. critiques [394, 290] and reviews [412] of the data dividend discussion).

However, a "meritocratic" framing for data dividends, which treats data as a kind of meritocracy, faces a major challenge: there is no objective, definitive, or even commonly agreed upon "best practice" approach for defining data value. There exist many ways to estimate the "merit" of data contributions in ways that are practically useful, but it could be the case that any given distribution of data value estimates is highly sensitive to choices by the people designing the appraisal process.

Put simply, there is major tension between the competing ideas that (a) data dividends should be implemented with a very coarse approach to measuring data value, by allocating a single pot of "data revenue" and dividing payments equally between all contributors or just spending the money on public goods and that (b) fine-grained data value estimation is important so that data dividends create incentives that help to improve data-dependent systems and are perceived as fair by recipients. Prior work has found that data valuation could lead to very unequal outcomes [409] and that data valuation techniques can be inconsistent [416], which suggests that perhaps a "coarse" baseline involving collecting a pot of money (e.g. via a tax [111]) and dividing that money equally amongst all dividend recipients (e.g., all residents of a state or all users of a platform) may be desirable. However, advances in the machine learning field on *data valuation techniques* have made it practical to estimate the impact of an individual data point or group of points on some machine learning model's performance on a particular test set [223]. This means it is possible to estimate the value of data contributed by each person in some population and disburse dividends according to the value estimates in a “meritocratic” fashion. In theory, if implemented well, a data dividend that incorporates fine-grained data value estimates could help lead to better performing AI systems. As data dividends are rolled out, it would be ideal to directly involve recipients in making these choices, but in the immediate future simulations can help navigate this tension.

In this chapter, we use machine learning experiments to simulate the impact of different data dividend design choices and describe tensions that arise when taking a “meritocratic” approach. We build on a recent advance in the data valuation space – the *Beta Shapley value* approach for data value estimation [223] – to simulate a series of data dividends pilot studies.[1] Specifically, we imagine 1000 randomly sampled U.S. households receive a dividend payment allocated based on data value estimates corresponding to a classification task. We vary key choices in dividend implementation and examine how this affects the distribution of the resulting data dividends. Further, we explore whether different dividend implementations will achieve progressive economic outcomes (i.e., reduced income inequality within the sample), in line with goals professed in responsible AI and human-centered ML research (see e.g. discussion such as [362, 49]).

The key goal of our work is for different actors – including researchers – to be able to make more educated decisions about what kinds of data dividends they want to support. Our experimental results show that how data value is defined drastically impacts dividend outcomes. Thus, any attempt at defining “merit” is sensitive to seemingly arbitrary decisions by the implementer. Similarly, the procedure for translating data value into units

[1] We will release the simulation code with the paper so other researchers can conduct similar simulations.

of money can also have a major impact. However, it is possible to navigate these decisions in a way that avoids extremely unequal outcomes. Further, it is possible to offload power from the designer to dividend recipients by allowing for data valuation at the collective level. Our results suggest that if dividends are implemented alongside legal and technical support for data coalitions, i.e. "data unions", [229] that offer technical and social means for people to bind together their data contributions in pursuit of shared goals, it may be much easier to settle on an appropriate balance of coarse (collective) valuation and fine-grained (meritocratic) valuation. If people group their data contributions together, there are *many fewer* counterfactual outcomes for a data value estimator to simulate, the data value estimation process is cheaper, the resulting dividends are likely to be more equal, and the responsibility for dividend outcomes shifts from a centralized designer to recipients themselves.

We reflect on implications for actors interested in data dividends, in particular emphasizing the core epistemological challenge at the heart of dividends: despite many advances in data valuation, "data merit" remains ill-defined. We conclude by proposing how data dividend designers might work from a desired outcome to design a dividend scheme.

9.2. Related Work

9.2.1. Background on Data Dividends

There is growing support for data dividends from political leaders and discussion in popular media. Most notably, the governor of California announced interest in implementing a data dividend in a major policy speech [396]. Similarly, many high profile media outlets have begun to cover data dividends and related concepts (e.g. [389, 307, 90, 250]).

There is also a growing body of literature that has advocated for data dividends and related approaches as a potential way to address computing-induced economic issues. Lanier and Weyl argued that providing payments for data, catalyzed by “mediators of individual data”, could be a critical step towards addressing economic inequalities [229]. Arrieta Ibarra and colleagues introduced a theoretical framework that calls for thinking of “data as labor” and rethinking the current paradigm in which data is treated as “exhaust” that companies collect; this framework directly supports the arguments underlying data dividends [13]. Some work has also investigated the benefits of considering data subjects as investors [200]; this approach may have have major synergies with the data dividends concept.

Some work has examined data dividends very closely, looking at feasibility [412], strategies for implementing a pragmatic data dividend in California specifically [111], and computational feasibility [28]. Several authors have also provided critiques of the data dividend idea on both practical and moral grounds [394, 290].

Data dividends may also benefit from the adoption of new data regulations in many jurisdictions. Of course, there are many challenges in translating changing laws into practical impact [36], but over time increased attention to data by legislative bodies around the world will likely make it easier, not harder, to implement a data dividend.

9.2.2. Data Dividends and Markets for Digital Labor

A substantial literature in HCI and related fields has studied crowdwork marketplaces such as Mechanical Turk [10, 157, 174, 202, 197] that enable people to be compensated for digital labor. Here, we discuss how data dividends differ from such marketplaces, but can be complementary in achieving pro-social outcomes.

In existing crowdwork marketplaces, workers are paid based on discrete data contribution tasks with definitive starting and ending points (see [10] for extensive comparison of crowdwork to historical "piecework" tasks). This approach does not translate well to compensating people for the data they produce passively during activities like using search engines and recommender systems, i.e. kinds of data often called "implicit feedback"[262]. These profitable intelligent technologies rely on behavioral data, e.g. what videos users watch and what links people visit. The people generating clicks for recommender systems and search engines are doing so under "natural conditions" – this data is valuable under the assumption that they represent "in the wild" preferences about videos and links. Thus, a developer cannot pay directly for 100 units of "search logs" in the way they could pay for 100 units of "labeled 256 by 256 images". However, they could pay for access to a year's worth of "natural" behavioral logs (of course, there is no true set of "natural" logs; behavioral logs generated in a world with payments will differ from those generated in a world without). This style of payment would effectively be a privately-implemented data dividend. In other words, if a crowdwork marketplace moves towards a pattern of paying for long-term behavioral data, that platform is implementing data dividends.

Furthermore, existing crowdwork marketplaces are not well-suited to rewarding historical contributions, and thus cannot be used to share the profits generated by technologies invented in the years before any broad profit-sharing approaches could be implemented. If computing-induced inequality increases dramatically, simply distributing earnings from future improvements to technologies will likely be insufficient to satiate economic and political demands.

Current crowdwork marketplaces have been subject to a broad range of critiques related to their suitability to addressing economic inequality (see e.g. [157]). Without careful consideration of the work which has critiqued the crowdwork paradigm, data dividends may fall into similar patterns, e.g. excessively low payments. There is active work seeking to address some of the above issues, especially those related to wages and working conditions, and our expectation is that crowdwork could be an important part of broadly distributing the profits from intelligent technologies.

9.2.3. Data Valuation

A central challenge for data dividends is identifying the value of a given unit or group of data in a particular context. As noted above, the mandate for data dividends can be interpreted as requiring value estimation at a very coarse level, or as requiring very fine-grained value estimation so as to provide highly individualized payments. Thus, our work intersects with data valuation literature at multiple levels: (1) valuing data in the aggregate, e.g. at the company level, and (2) observation-level data valuation, which is concerned with how much a single observation or distinct group of observations is worth in terms of impact on some machine learning model's performance measurements.

9.2.3.1. Aggregate Data Value. Many discussions around the value of data have considered estimating the aggregate monetary value of data to companies, which we might call "company-level data valuation" [132, 218, 245]. In discussing the results of their company-level data valuation study, Shapiro and Anejo called for 50% of estimated “data revenue” to be paid directly to Americans or into infrastructure [357].

Company-level data valuation is important for designing data dividends as estimating “data revenue” is likely to be an important step in determining the tax base for a state-run dividend, or the budget for a firm-led dividend. If a data dividend is going to be implemented using a very simple “just fund public goods” approach, this high-level data valuation may be the only value estimation needed. In early dividend implementations, this might involve very broad strokes, e.g. taxing 1% of revenue of a manually determined list of data-dependent firms. In the future, however, more precise estimates of the “causal impact of data” on revenue may be possible, especially if firms are incentivized to participate in the estimation.

Looking forward, identifying the value of intelligent technologies not directly connected to revenue (e.g. services like AI assistants, search, image tagging, etc.) and public datasets on which for-profit intelligent technologies rely (e.g. Wikipedia [404]) will be important for a more holistic perspective on aggregate data valuation.

9.2.3.2. Observation-level Data Value. Discussions of data dividends have suggested (often implicitly) that dividends should be paid out to individuals, in accordance with how much somebody impacted a data-dependent technology. Any such “meritocratic” approach to data dividends requires observation-level data valuation, i.e. a way to measure the value of specific observations to a ML model, which is an open challenge that has been explored by recent ML research [206, 223, 191, 126, 364].

Early data valuation work centered around the influence function approach from Koh and Liang’s work [206], which operationalized meritocratic data valuation by estimating how model loss changes when an observation is removed. This was highly influential in the ML community, receiving a best paper award at the ICML conference and substantial attention in the literature.

Following the attention to influence functions, rapid progress was made in developing data valuation techniques that use the "Shapley value" from cooperative game theory. The Shapley value approach was motivated in part by compelling theoretical properties [126, 191, 223]. For instance, under certain conditions, the "players" being assigned Shapley values (i.e., data contributors) can be sure that two observations that have the same effect on performance will be given the same value. The Beta Shapley approach [223] adds further "flexibility" to Shapley value based approaches, and offers unique advantages for data dividend simulations. We describe this in greater detail below in our Methods section. A later approach, Average Marginal Effect approach can be seen as inclusive of both Data Shapley and Beta Shapley [246].

Researchers have also proposed other alternative definitions of data value (see work from Sim et al. for a technical survey of data valuation [364]). The least core approach from Yan and Procaccia seeks to minimize the difference between the "value" and payoff of any given coalition [432]. This definition specifically incentivizes contributors to not opt out. However, this concern is probably more relevant to markets or long-running data dividend programs. The "datamodels" approach provides a way to directly learn about "data counterfactuals" for specific test examples, i.e. to predict the output of training a model with some training set and a specific test example [182]. Ilyas et al. show datamodels are very effective for many practical purposes.

All these definitions may be relevant to long-term data dividend discussions, but here we focus specifically on the Beta Shapley approach for its flexibility – as we are especially interested in the role of designer choice – and because it is practical to run experiments with small batches of data (see Methods for more details).

As we will discuss at greater length below, there is a fundamental challenge at play in using observation-level data valuation for data dividends: in general, amassing more data (assuming data quality remains steady) makes systems more useful, but it also makes estimating data value under most definitions more expensive. This is because estimating data value for large datasets involves simulating more counterfactual worlds (computing the exact leave-one-out value requires re-training a model n times, and computing the exact Shapley value requires re-training 2^n to cover every entry in the powerset of all observations).

For systems like large search engines, recommender systems, and models driving personalized ad sales which are very economically generative (see e.g. revenues from Alphabet and Meta), it is particularly challenging to apply fine-grained data value estimation techniques. For very small toy models that are amenable to very quick ML experiments, is it easy to apply data value estimation techniques. As we will see below, our results suggest a fortuitous fact for the equality-focused data dividend designer: performing valuation at the coalition-level can make valuation more tractable *and* lead to fairly equal data dividend outcomes that are less sensitive to seemingly arbitrary design choices.

9.3. Methods

9.3.1. Motivating our Simulations

Our goal is to closely simulate what a pilot program of data dividends might look like to better understand how outcomes vary based on designer choices.

First, we will restate the case for fine-grained valuation (i.e., a "meritocratic" approach). Then we will recap arguments from prior work about how meritocratic data dividends can easily go wrong. These arguments inform our simulation design, as our ultimate goal is to helps designers assess the costs and benefits of different data dividend design choices.

9.3.1.1. The Arguments for and Against Fine-Grained Valuation. Although this has not been made particularly explicit in ongoing discussions, it seems that the primary argument for fine-grained valuation in data dividends is that data dividends will create new incentives around the generation and sharing of data. As a motivating example, a program that offers to pay some fixed amount of money per search engine query, restaurant review, or piece of art posted to the Internet would likely lead to some degree of spam. Even a very basic attempt to distinguish "real contributions" from spam would constitute a basic assignment of "data values" (a spam/not-spam detector can be seen as a binary data value estimator). In other words, a data dividend implementer is performing very broad data value estimation just by choosing which data points should *not* be taken into consideration. If it is possible to both detect spam and identify that some observations are x times more valuable than other observations, it could make sense to pay dividends out accordingly.

Once a data dividend has been implemented, it will realistically change the incentives around data generation. For instance, if a yearly data dividend were announced by a state, country, or firm in December 2022, disbursed in January 2023, and covered in the news, for the rest of 2023 people may change their behavior in anticipation of a 2024 disbursement. Even a very simple implementation that pays a fixed amount to everyone who meets some inclusion criteria will create incentives to meet said criteria.

Thus, if we have access to techniques that assess the value of data, we could use those techniques to shape the new incentives around data creation. A key application of Beta Shapley and similar approaches is to identify data that is mislabeled, or particularly valuable to include in a training set. In other words, data value estimates provide some signal about data's impact, such that tying payments to data values *may* incentivize the creation of "better" data (in practice, this might mean data contributors spend more effort "cleaning" their data contributions, etc.). Of course, data values will be highly dependent on evaluation procedures and test set selection, and as discussed above, there are certainly ways that this scheme could go wrong.

The simulation approach in our paper is also relevant to data markets where buyers or sellers want to use data values to guide pricing (e.g., using data value estimates to set initial asking prices for observations, then adjusting based on market dynamics or auction mechanisms). The main object of our quantitative study here is: (1) how are data value scores distributed amongst different observations (and thus different people) and (2) how can we transform these data value scores to payments? Thus, if certain techniques lead to very (un)equal dividends, those same techniques will likely lead to (in)equality-related issues if used in data markets.

One could also argue for meritocratic data dividends on purely fairness-based grounds, i.e. regardless of how dividends change data-generation incentives going forward, we should pay more to people that made larger "contributions". We leave this argument to future work which directly engages with data dividend recipients and their perception of fairness.

Moving to arguments against fine-grained valuation, early work suggests that fine-grained valuation can leads to very unequal outcomes [409]. This is not a trivial concern; it seems very plausible that a poorly designed data dividend could act as primarily a "handout to the very well off". Further, a highly unequal dividend is likely to have inequalities along demographic lines, as seen in data valuation work [126].

Data valuation can be quite stochastic, which may make dividend payments seem capricious. This is an issue Wang and Jia aim to address with the "Data Banzhaf" technique [416]. In other words, data value estimation techniques are not guaranteed to produce estimates that are consistent across methodological choices, etc., then recipients may feel the disbursement is very unfair. In a world with data dividends, studying people's perceptions of data dividends could become a new area for HCI research.

Generalizing from work on learning curves and data scaling, we can conjecture that as any given dataset's size increases, the average impact of a randomly selected observation approaches zero. This leads to another argument against valuation-based dividends: if a technology is reliant on the collective contributions of millions or billions of people, we already know each individual value will be very, very small, so why bother spending time and energy performing data value estimation?

More generally, doing fine-grained data value estimation is costly. Given designers have access to very simple baseline options such as “divide evenly between every recipient” or “use money to fund public goods like infrastructure, education, etc.”, it will be important to assess the cost and benefits of using data value estimation at all.

As we will discuss further below, this final concern about dataset size and value estimation cost can be addressed by the introduction of the Beta Shapley value technique and the potential for valuation at the collective level.

9.3.1.2. Design Scenarios We Simulate. Prior work on data dividends proposed a set of design choices an implementer of data dividends will face [409]. The implementer entity could be a government agency (e.g., the state of California) or a private firm (e.g., a tech company). This implementer must have a funding source (e.g., a tax base in the case of a government). Next, the implementer must select some inclusion criteria for which kinds of data they will be evaluating, such as a list of tasks and corresponding datasets, or a single representative task and dataset.

In our simulations, we assume our hypothetical implementer has a fixed pot of money. In other words, we assume the aggregate data value estimation needed to figure out the size of the data dividend pot has already been performed, and focus specifically on the data value estimation challenges that arise from trying to allocate that pot of money. For illustrative purposes, we imagine a very specific scenario: a US federal entity is disbursing payments on the order of 2000 USD to a sample of U.S. households. The use of household income is convenient for estimating the impact on overall income inequality.

We further assume the implementer will use some variant of Beta Shapley value for value estimation. Following prior work [223, 416], we assume data value estimates will be computed using small batches of data. This addresses a weakness of Beta Shapley, which is that computational issues can make it challenging to value batches of ¿1000 observations [416]. However, a key idea from the Beta Shapley paper is that even when we draw very small samples from moderate to large sized datasets (e.g., drawing 200 observations from a dataset with 10k or 100k total observations) the resulting value estimates are still valid in terms of applications (noisy label detection, learning with subsamples, predicting the impact of point addition and point removal). We also validate that a batching approach gives meaningful data value estimates, which is something a data dividend implementer would also need to verify.

For our simulation purposes, we assume payments will be determined based on a single "representative data contribution" to a single classification task. Some datasets have highly unequal contribution patterns. However, we do not focus on these for two reasons: (1) because the actual raw differences between value estimates are small[2], for datasets with highly unequal contributions patterns the inequality in contributions patterns will likely dominate the inequality in dividends and (2) more generally, contexts with highly unequal contribution patterns are probably more appropriately mapped to markets (e.g., in the case of performing labeling tasks) or non-economic incentive systems (e.g., in the case of editing Wikipedia) than dividends.

9.3.1.3. Design Choices of Interest. We design our simulations to provide insight into how key choices about how value estimation is performed. The choices we investigate include:

[2] Formally, for a learning algorithm that is uniformly stable, "uniform value division produces a fairly good approximation to the true Shapley Value" [191] (Jia et al.).

- "valuation cardinality" – i.e., how many other observations are included when estimating the impact of a particular observation or group of observations
- choices about which data is considered and how that data is processed
- the impact of grouping data into coalitions and estimating group data values instead of individuals' data values
- choices around mapping data value estimates to payment amounts

By adjusting the distribution that provides the sampling weights for Beta Shapley estimation, a designer can effectively choose how much to weigh "low cardinality contributions" (i.e. the impact a data point has when it is added to small number of total data points) versus "high cardinality contributions" (i.e. the impact a data point has when added to larger number of existing observations). Effectively, these are different value definitions that correspond to the marginal impact of data defined based on different counterfactual scenarios.

We test the parameter choices from prior work: (16,1), (4,1), (1,1), (1,4), and (1,16), which includes two low cardinality approaches, the Data Shapley (i.e., medium cardinality), and two high cardinality approaches. See Fig. 3 from [223] for visual intuition as to how different parameter choices give weight to different cardinalities.

As noted above, we overcome the computational challenges with data value estimation by considering small batches of data. This choice also affects how cardinality is defined. If we only use batches of size 200, even high cardinality data value estimates will at most consider the interactions between 200 observations. However, this "ceiling" is somewhat inevitable. There is some dataset size at which valuation experiments become impractical (for instance, estimating an individual person's impact on a large language model is very unlikely to be performed in practice). As such, we view the batching approach as a reasonable limitation, and a choice an actual implementer would make for pragmatic reasons.

Data Attributes: Designers will need to choose which datasets and tasks they consider for valuation, and might even choose to make certain data processing choices before beginning value estimation. For instance, they might binarize a dataset with many kinds of labels or remove uninformative features to lower the computational cost of value estimation.

We use synthetic data so that we can explore several dataset attributes in a systematic manner. Specifically, we see how varying the number of classes in a classification task and changing the test set size impact data value distributions.

Multi-class problems have potential for overall greater information content. In other words, multi-class problems tend to map to harder tasks. Using the lens of extreme classification, we can even view tasks like user-specific personalization as special cases of very-many-class problems [274], so extrapolating to very large numbers of classes can be informative. If we understand how data values differ when tasks differ in terms of classes, this could give useful insight into how data dividends should account for binary classification tasks vs. high dimensionality classification or personalization tasks.

Data Coalitions Finally, we investigate how dividend outcomes would be change if people form "data coalitions", and data values are assigned to groups of data points rather than individual data points. We can think of data coalitions as a way for people to "bind together" their data contributions: the data user (e.g. tech company) will either have all or none of the observations from any given coalition. Computationally, this is convenient for data value estimation experiments, as it reduces the number of counterfactuals that data value estimation techniques need to explore. In the most extreme example where every contributor joins one data coalition, we need only train two models: one with "full data" and one with "no data". If there are just two coalitions, we need only three re-trains, and so on (with the exact number of retrains given by 2^c for c coalitions). However, it is not guaranteed that data coalitions will reduce inequality: it could be the case that one coalition's value estimate is very high. Our experiments investigate the potential impacts of data coalitions.

9.3.2. Experiment Details

9.3.2.1. Data Value Estimates with Beta Shapley Batches. The data valuation technique we focus on in our experiments is the Beta Shapley value [223], which builds on early work on "Data Shapley" [126, 191, 28]. Using Data Shapley as a measure of data value considers the impact of a given data point (or group of points) relative to all possible combinations of other data points. In practice, this means sampling many possible combinations of data points (of different sizes) and estimating performance. The Beta Shapley concept takes this a step further, and proposes that it might be useful to consider weighting certain sized combinations differently.

Beta Shapley allows a developer performing data value estimation choose if they prefer to emphasize the impact of an observation (or group of observations) relative to a small existing dataset (low cardinality) or a large existing (high cardinality) dataset. Beta Shapley can replicate the original Data Shapley approach (assign weights based on likelihood of drawing a particular coalition randomly) and LOO (consider only cardinalities equal to the dataset size minus one), so this technique is inclusive of other valuation techniques (though distinct from other value definitions like least core [432]).

We might expect that as the size of a training dataset increases, the absolute value of the leave-one-out impact of each observation goes down. The results from the Beta Shapely paper suggest that low cardinality data values are very helpful for recovering useful signal about the likelihood and observation is mislabeled, or the value of including a point in a subsample [223], and so can help address the basic issue that marginal impacts will trend closer and closer to zero as data size increases.

One limitation of Beta Shapley is that it is still quite expensive to estimate data values for many data points at a time (for instance, in [416], 1000 data points is used as a cutoff at which Beta Shapley is impractical). For our purposes, however, this limitation is not blocking, because Beta Shapley values can be acquired for just a small sample of data at a time. Indeed, in the main Beta Shapley paper, the authors sample just 200 training points, 200 validation points, and 1000 test points from the 580k data point Covertype dataset [33] and are able to detect noisy data. One could separate a large dataset into small batches of 100-200 points to get data value estimates for each data point in a computationally feasible fashion. These (cheap) estimates should be generally correlated with estimates generated (expensively) using larger batch sizes. To confirm this intuition, we validate that Beta Shapley values generated using different batch sizes are highly correlated, which suggests a reasonable approach for getting data values for many data points from a large dataset.

The idea that low cardinality data values are useful in practice is important for making our experiments viable. If a “low cardinality” approach had no signal at all, it would likely be impractical to use data valuation for most large models, and this might be preclude the use of value estimation for data dividends altogether. For instance, it will probably never be practical to use machine learning experiments to see how GPT-3 would behave if a single individuals' social media posts were removed from the training data [37], but this does not mean it's entirely impossible to estimate the value of data contributors relative to some other representative task.

9.3.2.2. Transforming Data Values to Money. In our experiments, we must transform data value estimates into units of money. Data values can be negative, so this poses a challenge because we assume that early data dividends should not be allowed to impose debt on data contributors. This means we must transform the data value assigned to each person or coalition into non-negative (but potentially zero) payment amounts. We consider three simple approaches that encapsulate different philosophies of data dividends, similar to prior work [409].

The first approach we consider is to linearly "shift" data values so that the most negative value becomes 0. Then, we normalize all values to add up to 1, so each person/coalition is assigned some fraction of the total pot. Mathematically, this is $x_{shifted} = x - min(x)$, $x_{fractions} = \frac{x_{shifted}}{sum(x_{shifted})}$, $x_{dollars} = x_{fractions} * total$

This approach encodes the stance that payments should be related to the impact an observation has (in terms of a weighted average over many possible combinations of observations). Furthermore, it maintains the overall shape of the data value distribution. It also guarantees that (almost) every data laborer receives at least some small payment. However, it can lead to some unintuitive outcomes, because the the ratio between any two data values is highly sensitive to the minimum value. We can illustrate this issue with a simple example. Imagine two people – Alice and Bob – have data values of 0.1 and 0.2 respectively. If Alice and Bob receive a dividend under "shift", Bob receives twice as much money, which has an intuitive connection to the fact that Bob has twice the impact on accuracy. However, if another person – Chen – joins and has data value -0.1, Bob now receives 1.5x more than Alice, not 2x more. If Di join with data value -0.2, Bob receives just 1.33x more than Alice. Under shift, relative payments are very sensitive to the "most harmful" observation.

The second approach we consider is to simply "clip" all negative values to zero, and then normalize so values add up to 1. This approach encodes the stance that the dividend implementer only wants to pay people/coalitions with a positive impact on model utility, and that the ratio of any two payments is equal to the ratio of the corresponding data values (a person with data value estimate of 0.2 always get 2x the payment of someone with data value estimation of 0.1).

The third approach is to just split payments equally between all people with positive data values. This approach is similar to "clip". One potential justification for this kind of approach would be to connect data valuation to some notion of quality thresholding (i.e., spam detection).

We refer to these approaches as "shift", "clip", and "binarize" respectively in our figures below. Of course, there are numerous ways to adjust the valuation-to-money procedure that could be considered reasonable. For instance, perhaps designers would like to shift data values and then re-scale, or use a more sophisticated clipping procedure that also handles outliers. As such, the exact values that our simulations outputs are much less important than the general trends we observe.

9.3.2.3. Datasets. For our base experiments, we generate synthetic classification data. We use sklearn's *make_classification* functionality[3]. We use the default parameter choices, except we reduce the number of features from 20 to five to speed up experiments. This software generates clusters of normally distributed points around the vertices of hypercube with dimension equal to the number of informative features. We focus on a binary classification task (except in specific experiments that involve adding more classes, described below). We show in the Appendix how non-synthetic data show similar value estimate trends.

9.3.2.4. Repetitions. For most of our experiments, we run five repetitions with a different sample of 1000 points separated into five size 200 batches (i.e., 5000 data points per experiment). We show a 95% bootstrapped confidence intervals over the repetitions.

9.3.2.5. Metrics. For each experiment, we first inspect the raw distribution of data value estimates outputted by the Beta Shapley technique (building on the code released with [223]). Then, we transform the data values into dividend payments using each of the approaches described above. We assume a pot of $200,000 to be disbursed amongst 1000 people in each batch, i.e. a mean payment of $2000 per person, a similar order of magnitude to the stimulus checks sent out in the U.S. during the Covid-19 pandemic [365].

From here, we produce a box plot describing the distribution of payments and then calculate the inequality of the ensuing payments in terms of the standard deviation and the Hoover index. The Hoover index provides a very interpretable summary of inequality:

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html`

the fraction of the total pot that would need to be redistributed to arrive at a perfectly equal distribution.

For each repetition, we draw a sample of incomes from the 2020 U.S. household income data released by the U.S. Census ("HINC-06" [44]).

This data release provides the mean mean household income and number of households within a variety of income brackets, so each sampled income in our data is the mean within a bracket. However, because our sample approximates a random sample, it is still effective for estimating how income inequality would change if dividend payments were disbursed at large.

Using this data, we simulate three scenarios. In the first scenario, we assign incomes to our simulated data contributors at random. We can think of this as estimating the impact of a data dividend where the appraised dataset is specifically chosen to have data values that are completely uncorrelated with income.

Next, we assume data value and income are perfectly correlated, i.e. the *most regressive possible scenario*. This might be the case if the test set is chosen to specifically reward increased accuracy for high income users. In this scenario, the person with the highest income will receive the largest dividend. Finally, we consider the case in which data value and income are anti-correlated, i.e. the most progressive scenario.

In interpreting these simulation results, we assume that choices leading to very unequal dividends are at higher risk of exacerbating inequality. Given the long line of work in HCI and related fields showing that computing technologies often perform more poorly for marginalized groups (see e.g. [268]), it seems more likely that data values will be correlated positively, not negatively, with income.

In our plots showing changes in income inequality in terms of the Hoover index, we include three baselines to contextualize these results. As a middle ground, we show a vertical line at -0.007 corresponding the the impact on Hoover index from giving every household a 2k payment. As a progressive upper bound, we show a vertical line at -0.014 correspond to the impact of giving all below median income households a 4k payments (i.e., a targeted intervention intended to create especially progressive outcomes). Finally, we also include the baseline of 0, i.e. the threshold at which a given set of dividends is actually regressive.

Just because a particular choice leads to concentrated data values, this does not *guarantee* the dividend will be regressive. Rather, it suggests that if dividend designers elect to make choices that creates unequal dividends, they will likely benefit from taking extra steps to avoid regressive outcomes.

In practice, a highly progressive scenario seems unlikely without very intentional design (e.g., using a test set meant to emphasize performance for people who are currently marginalized, disadvantaged, etc.).

In our Results section below, we focus primarily on describing the varying dividend outcomes. In our Discussion section, we further discuss the stance that designers should lean towards choices that reduce the overall inequality of dividend payments unless they are very confident that data values are not correlated with existing economic status.

9.4. Results

First, we provide an important preliminary result: the validation that data values estimated in small batches correlate to values from large batches. Then we describe the data value estimate distributions observed in our experiments, and finally how differences in value distributions translate into differences in dividend outcomes.

9.4.1. Calculating Beta Shapley with Batches

We performed experiments to test if Beta Shapley values computed using small samples correlate to values computed using large samples. These experiments test the following: If we generate a dataset with 1000 total observations, and split the datasets into chunks of various sizes (ten chunks of 100, five chunks of 200, 2 chunks of 500, or one chunk of 1000) in order to perform value estimation, will the values produced with each chunk size be correlated with each other? Based on results from the Beta Shapley paper, which showed small sample-derived value estimates are useful in practice, we expected this approach to work, but by testing this we add further confidence that our following experiments are ecologically valid.

In Fig. 9.1, we see fairly robust correlations for the lower cardinality Beta Shapley approaches. As we approach higher cardinality, these correlations are much smaller.

9.4.2. Data Value Estimate Distributions

Here, we examine of impact of key design choices on the distribution of data value estimates themselves, before these value estimates are transformed into units of money or distributed to households. These data values were generated using the Beta Shapley technique with accuracy as the evaluation metric of interest, so the data value numbers in this section can be interpreted as weighted averages of impact on accuracy.

In this section, the results of interest are descriptive statistics about the data values. The primary goal of analyzing these results is not to fully characterize or predict how data values are distributed, or to test a specific hypothesis about these distributions. Rather, our key goal is to see how much impact design choices have on data value distribution characteristics.

9.4.2.1. Cardinality. Fig. 9.2 shows different data value estimate distributions that arise from using different Beta distributions to produce weights for the Beta Shapley weighted average impact calculations.

The key takeaway from comparing data value estimate distributions across different cardinality choices is that low cardinalities provide a much wider range of data values. For instance, the "most helpful" (highest score) point with the low cardinality Beta(16,1) approach has a data value of 7.34, i.e. the weighted average impact on accuracy is 7%.

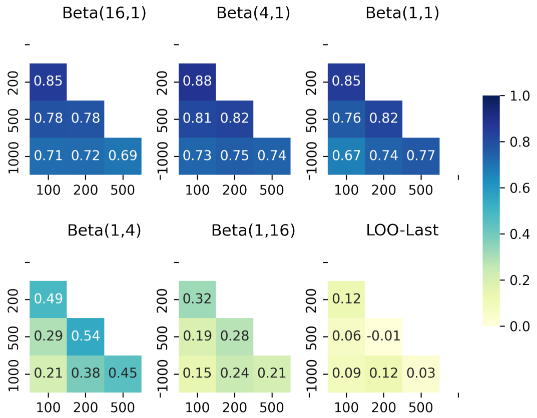


Figure 9.1. Pairwise correlations of Beta Shapley values computed using batches of size 100, 200, 500, and 1000.

However, with a high cardinality Beta (1,16) approach, the highest value is 0.7, an order of magnitude smaller. Higher cardinalities also reduce the median and standard deviation of the data values, i.e. the distribution becomes tighter with more mass close to zero. The percentage of negative observations also becomes higher, which will be relevant to our value-to-money transformations below.

We can interpret these results as illustrating how lower cardinality value definitions – which emphasize impact when data is added to a small dataset – allow any given observation to have more overall impact. This tracks with the diminishing returns curves that tend to characterize data scaling. As dataset size grows, the expected potential impact from adding a new observation falls.

These results also raise a question: as a data contributor, would one prefer to be appraised using low or high cardinality? As we will see below in our Payment results, the wide, right-skewed distribution with higher mean and median that is outputted by a low cardinality approach is likely preferable from the perspective of a user who wants to minimize the variance in their payment (we assume the mean payment will always be fixed because there is a fixed pot of money and fixed number of recipients).

9.4.2.2. Batch Size. Fig. 9.3 shows how data value estimates vary with different choices of batch size (i.e. how many training points are considered simultaneously). We show the low cardinality approach, so the overall range of values is still quite large. We see that batch size also has a large impact on how data value estimates are distributed. Larger batch sizes have much tighter distributions, with lower range, lower variance, and less skew. Furthermore, as batch size increases, we see the median data value decreases.

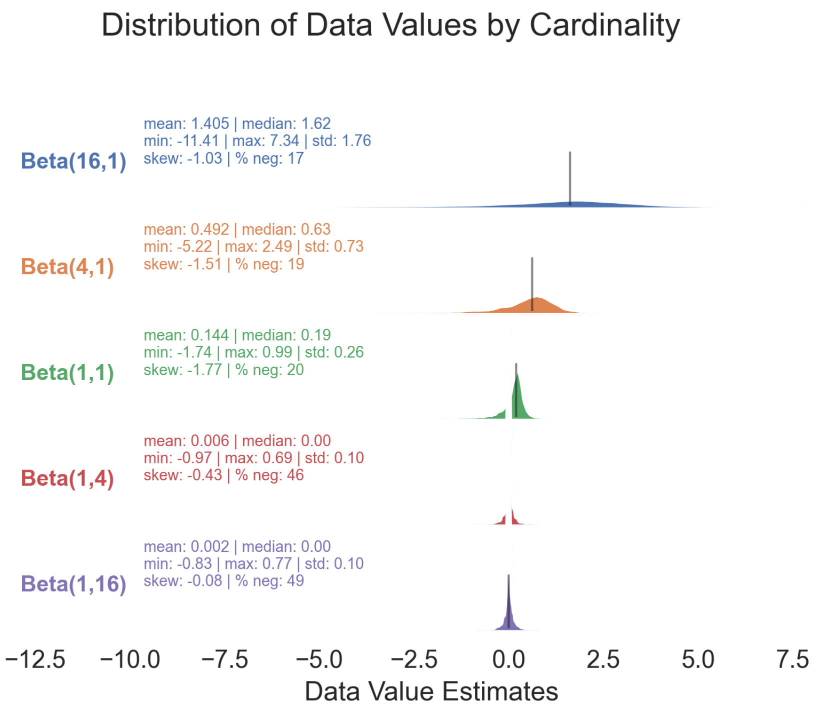


Figure 9.2. Density estimation plots showing how data values are distributed. Each row shows values obtained using a different set of weights that emphasize a different "valuation cardinality". From top to bottom, cardinality increases (topmost row is the lowest cardinality, which emphasizes impact of adding an observation to a small dataset). Each plot is annotated with descriptive statistics: the mean, median (also visualized with a black vertical line), minimum, maximum, standard deviation, skew, and percent of observations with negative values.

These results are consistent with the idea that batch size, like cardinality, impacts which counterfactual scenarios data valuation is "exploring" . Specifically, batch size controls the maximum dataset size we can consider, while different beta distributions control how much we weight all the possible combinations up to the maximum size. In other words, batch size imposes a ceiling on the size of counterfactual scenarios the data valuation technique considers. This explains why batch size shows a similar trend to cardinality.

Examining batch size in this context also raises an important question: is it possible that certain observations may miss out an being appraised their maximum value, because the batch size was too low? For instance, if an observation on its own causes accuracy to decrease, but in combination with 10000 other data points causes a substantial increase in accuracy, is a small batch size unfair to that observation? Looking at our correlation results from above (Fig. 9.1) together with the data values in Fig. 9.3, this seems possible but unlikely; some observations may see different relative values with different batch sizes, but most will not. We will see below that for a user seeking to minimize variance in payments, a small batch size with a wide distribution of value estimates is likely preferable.

9.4.2.3. Dataset Choices. We also investigated choices related to the dataset being appraised. We examined the role of test dataset size and the number of classes included in the dataset being appraised.

In general, the data value estimates were qualitatively very similar across test set size. In Section 9.5, we discuss how non-random perturbations to the test set might impact dividends and potential for future work along these lines.

Fig. 9.4 shows how varying the number of classes present in our classification task. We saw that the impact of adding classes was dependent on cardinality choices. With a low cardinality emphasis, adding more classes tends to push the data value estimation distribution tighter and more centered around zero. When a task has many classes, it could be the case that there are more classes than observations included in each of

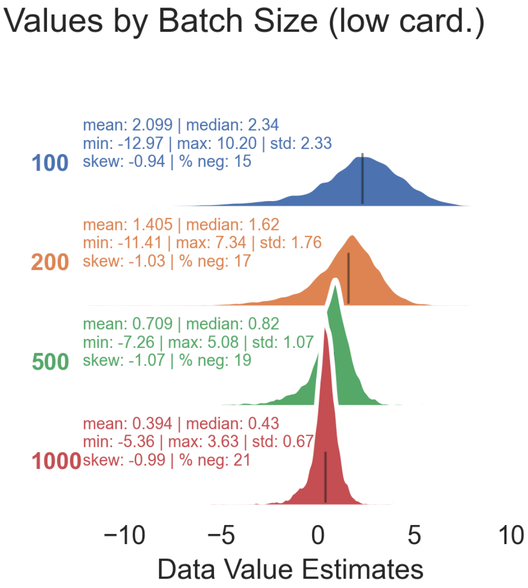


Figure 9.3. Impact of valuation batch size on data values. Follows the structure of Fig. 9.2. These data values estimates were produced using a low cardinality approach.

the counterfactual scenarios explored. For instance, the Beta(16,1) distribution applied to a batch of 200 emphasizes combinations of around 2-25 observations. When a high cardinality emphasis is used instead, this trend flips and including more classes or features causes the value distribution to spread.

9.4.2.4. Data Coalitions. Finally, we examine how data coalitions can impact data value distributions. In other words, when data contributors bind their data together, so that value estimation techniques treat coalition observations as grouped together, how do value distribution change?

Our coalition size experiments assume the following: a coalition will split earnings equally between its members, and all data contributed by coalition members will be

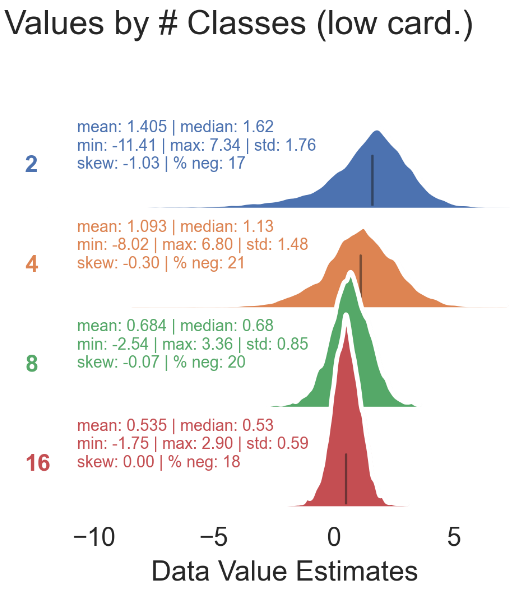


Figure 9.4. Impact of number of classes in a dataset on data values. Follows the structure of Fig. 9.2. These data values estimates were produced using a low cardinality approach.

valuated as a single group. Thus we only need to produce one data value estimate per coalition. For an experiment with only two coalitions, we produce only two *collective* data value estimates. Thus, our distribution plots for data coalitions are include a small number of values compared to the above non-coalition plots. Additionally, because these experiments are much cheaper to run, we include 50 repetitions instead of five. The number of combinations of coalitions is much lower than the number of combinations of observations. For instance, there could be only two coalitions. Thus, low cardinality valuation makes less sense here, so we focus on high cardinality results for these coalition simulations.

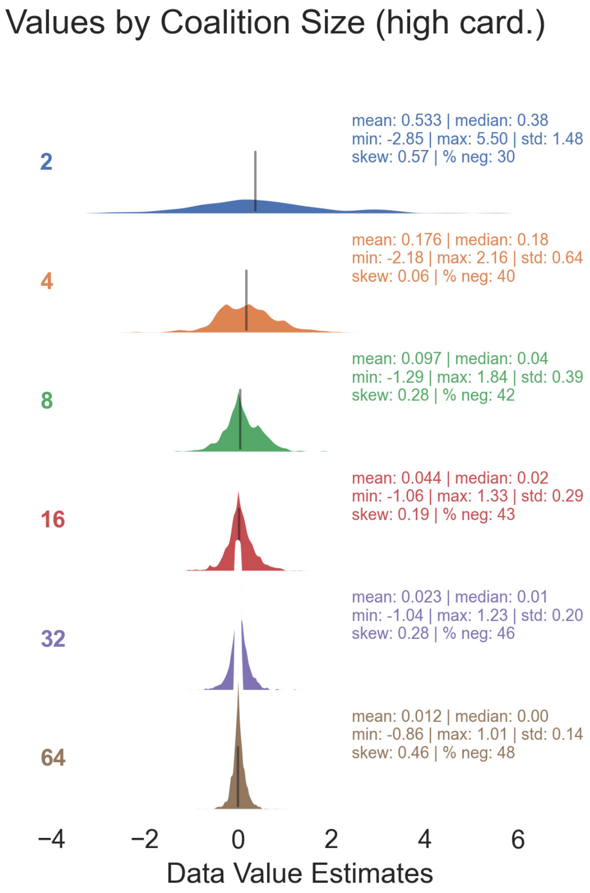


Figure 9.5. Impact of "data coalitions" on data values. Follows the structure of Fig. 9.2. From top to bottom, number of coalitions increase, and thus the coalition size decreases. These data values estimates were produced using a high cardinality approach.

In general, we see that splitting data between more coalitions (such that each coalition is smaller on average) create tighter distributions with smaller overall data values. This is consistent with the idea that data is more valuable in a group. As total number of coalitions approaches total number of contributors, the impact of coalitions go down.

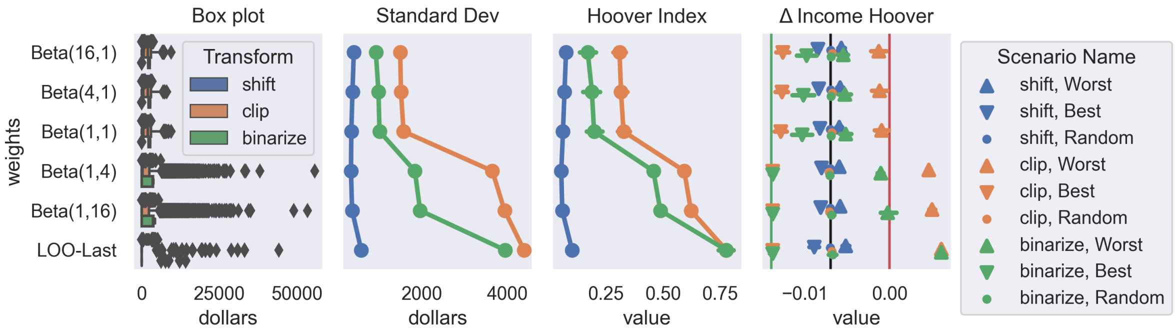


Figure 9.6. Shows payment outcomes, after converting our above data value estimates into monetary payments with a mean of 2000 USD. From left to right: a boxplot, faceted by different transform approaches, the standard deviation of payments, the Hoover index of payments, and finally the impact of U.S. household income inequality measured with the Hoover index. The rightmost panel includes changes in household income Hoover under "Worst" (most regressive), "Best" (most progressive) and "Random" (data values and income uncorrelated) conditions. Rightmost panel includes three vertical lines corresponding to several baselines: the change in income from allocating 4k payments to all below-median income households, the change in income from allocating 2k payments to all recipients, and a line at zero.

9.4.3. Inequality of Payments

Above, we saw how various design choices impact the distribution of raw data values. Here, we examine how the data value estimate distributions produced by different design choices translate into different payment outcomes. Overall, these results show that differences in data values can indeed translate into very different payment outcomes.

Looking specifically at Fig. 9.6, we see that higher cardinality choices – which lead to tighter data value distributions – are also more unequal overall. A similar, but less extreme trend exists for batch size, shown in Fig. 9.7. In fact, consistently across our results, cardinality is a very dominant factor in determining how unequal dividend payments are.

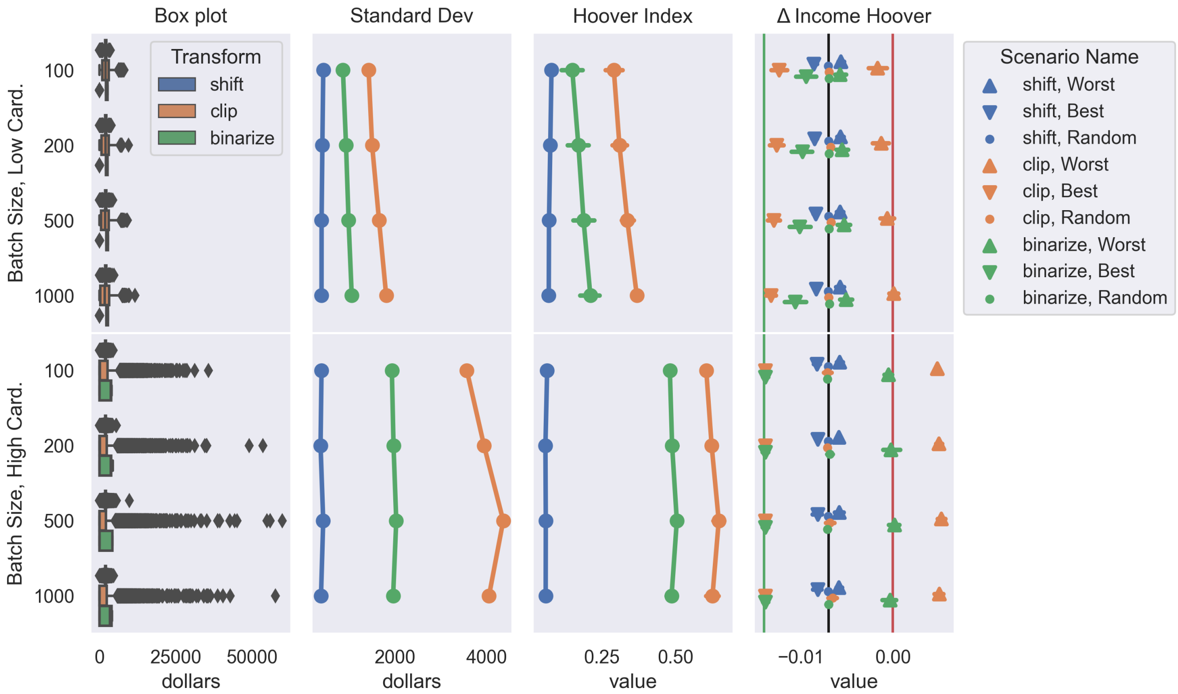


Figure 9.7. Following the structure of Fig. 9.6, shows payment outcomes for different choices of batch size. Top row shows low cardinality data values and bottom row shows high cardinality.

Looking at the role of different transformation approaches (i.e., comparing the different colors in Fig. 9.6), we see that in general the *shift* approach creates more equal payments. *Clip* creates very unequal distributions, and *binarize* serves as somewhat of a middle ground approach. The choice of transformation has a major impact on payments. For dividend recipients interested in having lower variance (i.e., more “certain”) payments, the results in Fig. 9.6 suggest that the clip and binarize transformations may be undesirable.

However, we see that the coalitions approach can minimize the impact of the transformation choice. In Fig. 9.8, when using low cardinality value estimation, payments remain fairly equal and actually become more equal with more coalitions. Even if high cardinality value estimation is used (shown in the bottom row of Fig. 9.8), the coalitions

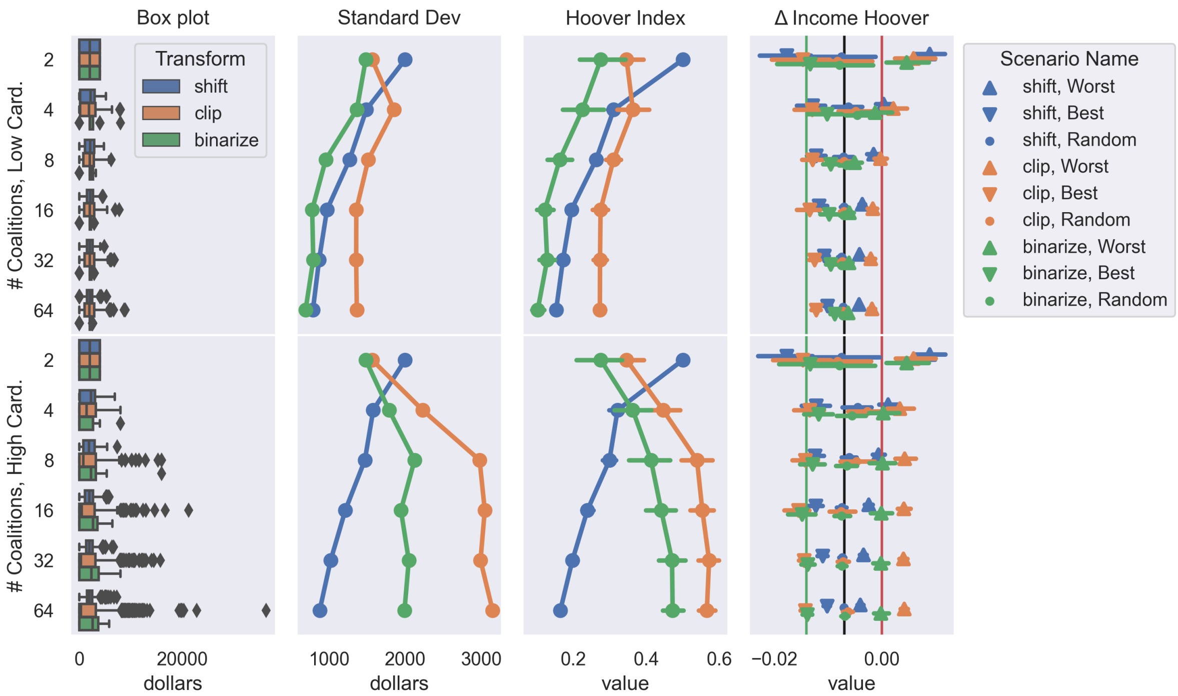


Figure 9.8. Following the structure of Fig. 9.6, shows payment outcomes for different numbers of coalitions. Top row shows low cardinality data values and bottom row shows high cardinality.

approach also keeps payments more equal so long as the number of total coalitions is low (i.e., each coalition is large).

Importantly, the rightmost panel of each plots shows how much outcomes could vary based on whether data value estimates and a recipient's income are correlated (positively, negatively, or not at all). In general, a particularly unequal set of payments can be very progressive – often matching the "very progressive baseline" (green vertical line, corresponding to the outcome of targeted payments). However, that same set of payments can also be very regressive (i.e., meaningful increase in the household income Hoover index).

An important general takeaway from these results is to confirm that the large differences in data value estimate distributions we observed can indeed translate into very different dividend outcomes. In other words, designer choices – which are discretionary, and may even seem arbitrary to recipients or to designers themselves – can massively impact whether data dividends are progressive or regressive. Even with support for data coalitions, dividends can still create regressive outcomes (the absolutely most certain way to avoid this: treat all users as members of one coalition and just direct a single pot of money towards progressive initiatives).

9.5. Discussion

In line with prior work, our results here reinforce how easily seemingly innocuous design choices can impact the outcomes of data dividends. Anyone implementing a data dividend should consider conducting similar simulation experiments, and sharing the results of these simulation with recipients of the data dividend so that recipients can weigh in. The idea of "human-in-the-loop" data dividends may be a particularly ripe area for HCI to contribute to this high-stakes challenges.

Given that seemingly arbitrary decisions (e.g., choosing shift or clip to transform data values to money) can have such large impacts on the payments people receive, it seems of critical importance to give recipients a say in how these decisions are made. If no attempts to include recipients in the design and governance of a data dividend program are made, there is always a strong possibility that data contributors could withhold their data contributions to attempt to gain leverage (as discussed throughout this thesis).

9.5.1. Data Dividends without Fine-grained Valuation

We examined several key choices that arise when attempting to perform fine-grained data valuation. As noted above in our arguments against using valuation, a data dividend could be implemented with only very coarse data valuation. For instance, a government could use very simple heuristics to tax data-dependent firms (e.g., a tax on 1% of revenue from some heuristically determined set of "data-dependent" firms [111]) and split this equally between all recipients, target payments towards worse off recipients, or just fund public goods.

Our results suggest a coarse approach is likely to have a low overall cost to design and test compared to a fine-grained approach. Furthermore, a coarse and very collective approach to valuation permits fewer paths to regressive outcomes. However, we did see there are paths to using fine-grained valuation but avoiding regressive dividends.

In terms of design implications for a data dividend that could be implemented tomorrow, it seems that focusing first on very coarse data valuation is a good first course of action. If fine-grained data valuation is preferred, using a low cardinality data value definition could be helpful to avoid very concentrated payments, and it will likely be worth paying close attention to data-related choices.

9.5.2. The Role of Data Coalitions

As noted above, our results suggest that particularly effective way to navigate the tension around data valuation in data dividends is to support data coalitions and perform valuation at the collective level. We can think of data coalitions as a way for dividend recipients to have a say in how fine-grained or coarse data valuation is: a small set of large coalitions will lead to coarse data valuation, whereas a large set of small coalitions will lead to fine-grained valuation. As the number of coalitions grows, the agency would effectively be transitioning towards increasingly fine-grained valuation.

Our results show how coalitions provide one approach for limiting how unequal dividends become. Dividend designers can even use a "clip" or "binarize" approach without excessively unequal outcomes. The intuition behind this result is that these approaches tend to penalize certain recipients, but coalitions help to spread out the impact this penalty.

An additional benefit is that coalitions provide a way for data dividends to leverage ongoing work on governance [435]. For instance, coalitions could perform their own data valuation and set norms around "high quality" data production.

Finally, an additional large benefit that is especially relevant to simulating data dividends is that calculating collective data values is much less computationally costly. When we have a just a few large coalitions, there are very few counterfactual scenarios to explore, so grouped data values can be estimated very quickly. If several data coalitions were to arise and bind together much of the world's data, data valuation in practice would become much easier (but overall, less informative). For instance, if everyone in California were to join one of ten data coalitions, the agency running a data dividend would only need to produce ten data value estimates (requiring only ten experiments for a leave-one-out approach, and 1024 experiments for an exact computation of Data Shapley or Beta Shapley).

9.5.3. Working Backwards from Desired Outcomes

A distinctive advantage of data dividends is that the designer can pick a desired level of inequality and desired impact on inequality metrics. Then, using knowledge about how data value estimates tend to be distributed (using simulations results like ours and/or incorporating theories that explain data value), it may be possible to design a data dividend scheme that achieves the intended outcome and is highly satisfactory to recipients.

An interesting idea ripe for future work is that designers could also influence the outcomes of data dividends through careful test set selection. This could open the door to data dividends that use either fine-grained individual level valuation or a middle ground collective approach to implement data dividends that are reparative in nature, e.g. by constructing a test set that emphasizes performance for groups that are historically or currently disadvantaged.

9.5.4. Limitations

Our simulations tell us about what small data dividends pilots might look like. We made assumptions about what kinds of techniques the data dividend designers might explore, but did not cover all potential possibilities.

A key takeaway from our work is that the many design choices that go into data value estimation will also have large effects on dividend outcomes. This means that as data valuation techniques are further improved and new definitions and techniques are proposed, the design space could become even more complicated. Conversely, in the ML community coalesces around a very specific approach to data value, the design space could shrink. Overall, these kinds of results are highly sensitive to trends within the realms of ML research and practice.

A particularly important open question is how data valuation will be adapted for very large models, especially those at the scale of major search engines, recommender systems, and language models [35]. At some dataset size, it's not practical to do any valuation, unless at the coalition level. If small models are replaced by very large models (i.e., models in the same class as "foundation models" [35]), data valuation will remain very important, but become more expensive in practice. The costs of data valuation could have major implications for data dividend design.

Finally, a critical question for future HCI work will be how recipients perceive and respond to data dividends. Ultimately, given the mandate of data dividends is sharing the wealth from data with the people who help fuel data-dependent technologies, giving data contributors voice should be a top priority for data dividend design. Simulation work cannot answer these kinds of questions, although our simulations results (e.g., the density plot figures) could be used to develop visualizations for dividend recipients or tools for people to explore different dividend outcomes.

9.6. Conclusion

In this chapter, we used simulation experiments to study the role of data valuation in the design of data dividends. Our results highlight that using fine-grained, "meritocratic" data valuation requires making a series of highly discretionary design choices. These choices can create dividends that are highly unequal, potentially causing the dividends to fail to produce progressive outcomes. We conclude with recommendations for designing dividends that mitigate this concern, by supporting data valuation at the collective level (i.e., coarse valuation of groups instead of fine-grained valuation of individuals).

CHAPTER 10

Towards a Theoretical and Empirical Understanding of Data Leverage

Questions about how different datasets impact downstream models fall under the purview of core ML research. As we have shown throughout the thesis so far, they are also central to core HCI questions, and answering them is important for addressing concerns of interest to both HCI and ML.

In this chapter, we discuss a line of ongoing and future work aiming to connect theoretical and empirical approaches to understanding the potential impact of data on downstream outcomes. We focus in particular on highlighting connections between theoretical research that explains how data points impact models, empirical work that explains how observed model performance varies with upstream data, and data leverage work. We lay out some promising avenues for extending research at this intersection, and discuss early findings that might support this research direction.

A key assumption motivating our future work is that when a group of data contributors are responsible for data that has a large impact on some systems, this will be correlated to greater likelihood of data leverage success. A group that massively impacts performance may have their demands addressed; a group that causes no noticeable impact will not.

It is likely that as time goes on, data will become a source of leverage for some people. The likelihood that a group of people will exert influence via withholding or making demands about data is dependent on the resulting impact-on-models. Different individuals and coalitions of individuals will have different impacts on models. Following this argument, this means that data contributions may act as a kind of voting credit; people will act as agents casting votes that determine the outcomes of the data-dependent systems, with major implications for the organizations who reap power and profits from them. Optimistically, drawing on the data as labor metaphor, this relationship could look something like works councils [5]. Less optimistically, we could see individuals or interest groups with very high levels of impactful data gain concentrated power.

This suggests it will be important to theorize about the *distribution of data contributions* and to study the power dynamics that result from data contribution in practice. The distribution of data contributions in the world may be a key part of how data-dependent systems and organizations are governed.

Just as it is useful to theoretically understand the mechanisms that allocate resources amongst agents (i.e., using the lens of mechanism design [160]), it will be useful to do the same for data leverage. Just as an individual might wonder, "How likely am I to impact the outcome of a particular election or to win an auction", an individual might also wonder "How likely am I to impact the outcome of a data leverage campaign"?

There is a possibility that it may become more common for people to be paid for data, via existing crowdwork platforms, other visions of data markets or data dividends (see e.g., the previous chapter and [111]). In this case, the ability to impact a model may translate directly into economic winnings. In this scenario, understanding how data value is distributed and how significant leverage can be achieved will be of critical importance to understanding the relationship between data and economic inequality.

The stakes are high for how "data impact" is distributed amongst individuals and groups. Key data sources like Wikipedia are the result of highly unequal contribution patterns; some people make many more contributions than other people, often following a power law pattern (or close to it) [222]. Does a person who contributes 10x as much data as another person have 10x more influence on downstream technologies? On the technology side, does this relationship vary by task and modeling choices? If there are less diminishing returns to data in a certain context, data leverage may be more likely to enable inequality in power. On the user side, how could coalitions formed by users impact these power dynamics? Perhaps a coalition of power users in some contexts will wield even more influence than their outsized contribution patterns would suggest. A theoretical understanding will also help to identify which kinds of coalitions can wield power through data.

In particular, it will be useful to begin to answer questions about data leverage such as:

- How much can an individual impact the group-level effectiveness of a data leverage action?
- How is the ability to influence data leverage outcomes distributed amongst different groups of contributors?
- How does the minimum, maximum, and average impact of data leverage scale with group size?

While the previous chapters have discussed completed research projects, this chapter focuses instead describing potential paths towards answering these questions. We provide formal definitions for the data leverage problem setting that will be useful starting points for answering these questions. These definitions draw heavily on the work described in prior chapters.

Then, we describe connections between existing data value estimation techniques and their applicability for data leverage, and how work along these lines can begin to answer questions such as the examples above.

The key goal of this chapter, and the reason for including "definitions", is to show how data leverage questions can be understood purely in terms of counterfactual questions about systems that use machine learning and statistics.

10.1. Definitions

10.1.1. Data-dependent Technology Operators

We are interested in scenarios in which some data-dependent technology operator (DDTO) – for instance, a technology company, state actor, or other large organization – wants to make inferences about people or predictions about aspects of the world. This broad class of scenarios implicates a number of high impact technologies across a massive number of domains. For instance, examples of DDTO's employing data-dependent technologies include:

- An e-commerce firm that recommends products to customers
- A search engine or social media-operating firm that personalizes advertisements for users
- A machine learning group that sells pay-as-you-use access to classification technologies, e.g. a computer vision API

- A government organization that deploys computer vision technologies, i.e. a police department

There are many organizations that can act as DDTOs, and many *tasks* that are useful to solve. For the purpose of our analysis, we specifically assume that a DDTO will use a socio-technical procedure (e.g. "assign a team of data scientists to develop a system for making inferences – involving choices around a loss function, an optimization algorithm, a dataset, etc.") to produce a system that maps some input x to some output y (e.g., map a user id to a list of recommended movies, map a search query to a set of search results, map an image to an inferred category, map a text prompt to additional text output).

Definition 1. *Data-dependent technology operator (DDTO) – an organization that operates a technology that relies on some pipeline of data used for training, testing, calibration, or other purposes. Can be a firm, a state entity, or other kind of organization. A DDTO can be the target of a data leverage action.*

We focus on *data-dependent* technologies, meaning technologies that perform tasks that require one or more *datasets* for training, calibration, evaluation, or other purposes. We call the individual units of data *observations*, and note that the exact granularity or definition of an observation will vary by task.

Following common practice, we call our input-output mapping system f. We assume that we will find f using n observations, where each observation is a pair of features and labels, $d_i = (x_i, y_i)$ and the set of all observations is called D.

A data leverage-focused analysis must account for a factor that is not typically considered in typical ML analyses: each observation is attributable to one or more people, the *contributors* of that observation. We will refer to individual contributors as c, the set of all contributors as S, and a particular coalition (i.e., combination of contributors) as C.

We say that an observation is attributable to a person if that person can exert agency in the present to change that observation (e.g., by submitting a data deletion request) or if that person could have changed their behavior in the past to deny or alter an observation. Thus, for each observation $d_i \in D$, we assume it is possible to identify one or more contributors responsible for that observation.

This agency-focused definition – the notion that contributors can be identified by their role as agents who might change the course of a data-dependent system – is a key contribution of our work.

The exploration of data leverage counterfactuals ("What if some contributors withheld their observations?") will rely heavily on mapping observations to contributors. In general, contexts in which contribution patterns are not tracked will make it hard, but not impossible, to engage in theoretical or empirical analyses of data leverage. Many existing data-focused analyses have this characteristic (see e.g., [169]). There are many good reasons why datasets, especially public datasets, do not include data necessary to distinguish between contributions. However, it is important to note that this poses a major challenge for studying contribution attribution.

Definition 2. *Contributor attribution – an observation is attributed to any person who can exert agency to change that observation, by e.g. submitting a data deletion request, actively editing a database, or by changing their "natural" behaviors. Observations can be* *attributed to more than one person and one person can help create many observations, so contributor attribution is, in most cases, a many-to-many property.*

10.1.2. System Evaluation and Data Leverage

We assume a DDTO evaluates its systems with one or more testing datasets. The process by which a testing dataset is selected has critical implications for data leverage, as this is the step where the DDTO asserts a particular value system and decision-making framework. As we will demonstrate in further detail below, different deployment (and therefore, testing) choices can radically change which contributors have leverage to harm a model. For instance, some organizations might want to prioritize overall accuracy for a random sample of users (which may implicitly favor the interests of a majority group). Another organization might construct a test set with subgroup equity in mind [19], or with the goal of optimizing on the cost of certain outcomes (i.e. cost sensitive learning [101]).

The DDTO's choice of testing dataset can be assumed to depend on the intended use case of the technology and the organization's values. This adds a complicating wrinkle to our neat narrative: any quantitative exploration – whether theoretical or empirical – of data leverage depends on modeling the value-systems embodied by organizations. Furthermore, there is no true "correct" test set, and recent work has increasingly highlighted the limitation of benchmark dataset culture [301].

We might address this by considering three distinct testing approaches that cover a variety of real use cases, and discuss the modifications needed to explore other value-systems. For instance, one set of choices could be:

- Random sample by observation frequency
- Random sample by contributor frequency
- Weighted sample by some reparative principle

10.1.3. Example Notation for Data Leverage Scenarios

With above context in mind, we now return to providing formal notation for describing data leverage scenarios. We assume that for a predictor f for task T trained with dataset D composed of n distinct observations from n_C contributors ($c_i \in 1...n_c$), the number of distinct training data combinations of any size (i.e. the *powerset* of all observations) is 2^n and the number of distinct combinations of contributors of any size is 2^{n_c}. Without even incorporating hyperparameter sweeps, attempting to work with exact computation will become impossible.

We can further assume that each DDTO gains utility from predictor f, or else they would not create and operate said predictor. We will call this utility $U_{operator}(f)$. U may be measured in units such as dollars, time saved, etc.

There may be certain thresholds for $U_{operator}$ at which the model must be pulled from production or substantially changed. Such thresholds will be relevant to effective leverage. In fact, by making assumptions about performance thresholds, it may be possible to greatly simplify early analyses. But, if these assumptions are inaccurate, these analyses may provide less predictive power.

Data contributors themselves gain utility from the model, $U_C(f)$, and contributor utility may not correlate positively (or at all) with operator utility. Contributors may even exist in groups with very different interests, e.g. a music recommender system that models content creators (musicians) and listeners who generate behavioral data used to recommend content [272].

We can quickly arrive at a very complicated analysis by trying to model the utility of every individual person who plays a role in the data supply chain (i.e. the data pipeline, data stream). To make early progress in this area, we believe it may be easiest to focus first on operator utility, and then incorporate the impact of contributor utility to understand the likelihood of collective action occurring.

10.2. Counterfactuals that Involve Removing or Adding Data

Now that we have discussed data leverage scenarios formally, we will provide an overview of different counterfactual ways of thinking about data value that are needed to discuss data leverage.

10.2.1. Influence Functions: Data Values that Assume a Fixed Dataset

By the definition given above, anyone who is a contributor to a dataset could counterfactually *not contribute to that dataset.* A contributor's impact is the difference between $U_{operator}(D)$, operator utility with "full" data, and $U_{operator}(c_i \notin D)$, the operator utility when all observations attributed to c are missing from D.

We might call this the Leave One Contributor Out Influence:

Definition 3. *Leave One Contributor Out Influence, where c is the set of all points attributable to an individual:* $I(c) = U_{operator}(D) - U_{operator}(D \forall d_i \notin c)$

The influence definition can be seen as one of the most intuitive definitions of "value": what happens when an observation is added to an existing training set? Furthermore, it also maps to perhaps the most popular notion of "data value" in the the recent ML literature. Koh and Liang built on Cook and Weisburg's earlier results [66, 67] to develop a technique for estimating the influence of upweighting a point z on parameters θ [206]. The resulting influence function paper and code was highly influential in the ML community. In later work, Koh and colleague define group influence as the *predicted effect* of removing a subset of training data [207].

A weakness of influence functions for use in data leverage work is that we are rarely concerned with the influence of a single data point, and group influence may be less accurate overall (though still very useful) [207]. Influence functions have also been subject to criticism, especially around their use for deep learning models [21].

In almost all real-world cases where data comes from many individuals, each Leave One Contributor Out value will be extremely small (and could even be positive, i.e. removing that individual's data improves test performance).[1].

More relevant to the topic of data leverage is to imagine coalitions of users who work together to withhold data (or to poison data, redirect data, etc.). Thus, it will be useful to similarly define a Leave Coalition Out Influence for each unique combination of contributors:

Definition 4. *Leave Coalition Out Influence, where C is the set of all points attributable to coalition members:* $I(C) = U_{operator}(D) - U_{operator}(D \forall d_i \notin C)$

[1] There are interesting exceptions wherein a single individual truly contributes a massive dataset entirely "on their own". This could occur when an entire dataset truly comes from a small group of people. For instance, imagine a graduate student in geology who has manually collected geologic measurements, and then trains an ML model on this "solo-authored" dataset.

Definitions of influence that involve only a single subtraction (utility with observation minus utility without observations) tend to map to a situation in which some dataset already exists, and some data is going to disappear from it. Implicitly, this assumption maps to an "opt-out" paradigm of data collection.

10.2.2. Shapley Values and A More Variable Notion of What Data is "Fixed"

We could also ask counterfactual questions about adding observations to a small dataset. This framing maps more closely to research on using data valuation for active learning [127]. Shapley values provide a way to split the difference between a viewpoint that emphasizes the impact of withholding observations versus one that emphasizes the impact of adding observations.

The *Data Shapley*, proposed by Ghorbani and Zou [126], gives the average influence of an observation across all possible marginal contribution scenarios of a data point. Given there are many such scenarios (2^n), it must be estimated, and several groups of researchers have produced software to do just that (see e.g, [126, 191, 127]). Computing the Data Shapley values for a dataset tells us about which observations we might want to drop, investigate, or try to add to our dataset (a main use case discussed in much of the work is data markets [126]). As noted in Chapter 9, the Beta Shapley provides an even more configurable variant of the Data Shapley.

The Shapley value originated in cooperative game theory as a way to estimate the impact of a "player" on the outcome of some game in a scenario where each player can join many different possible coalitions. In the context of data-dependent technologies, there are many ways a set of data points can be combined, so the core concept of considering many possible coalitions is very relevant.

A data point that has high impact when added to one sample may have nearly no impact when added to a different sample.

Shapley values can be thought of as weighted sums of influence values with different subsets of data. The largest contribution to this weighted sum comes from the influence value (i.e. when the subset is all other available data points).

Shapely values can be defined (and estimated) for groups of data points as well [223]. Thus, if we aggregate data Shapley at the contributor level, we have a value we might call the **Contributor Shapley**.

Contributor Shapley can tell us the ability of a particular contributor to influence the success of data leverage. We assume each contributor is an agent who can choose, for a particular leverage-oriented campaign, whether to "vote" for the campaign by withholding their data.

Definition 5. *Contributor Shapley – Group Data Shapley, computed for all observations that are attributable to one contributor. Gives a robust estimate of the impact of adding/removing that contributor.*

For datasets where a one-observation-per-person rule is enforced, contributor Shapley reduces to data Shapley, but this is rarely the case. Almost always, a model could be improved by allowing more observations per person.

In the context of reasoning about data leverage outcomes, certain permutations may be more likely than others. That is, there are certain groups of people who are more likely to engage in collective action together (e.g., because they are part of some kind of online or offline network or community).

Looking forward, it would be especially useful to consider a propensity-weighted variant of the Shapley value that takes into account that certain combinations of contributors are more likely to engage in data leverage together than other combinations. The topic of propensity-weighted estimates of data leverage coalition influence may be a particularly promising avenue for future work.

10.2.3. Data Scaling

In previous Chapters 7 and 8, we discussed the importance of data scaling research. Data scaling work is generally concerned with understanding broad trends in how system performance varies with access to more or less data. Data scaling experiments may not tell us about a specific instance of a data leverage campaign, but they are certainly useful for reasoning about the likely impact of different data leverage actions. Because they typically involve randomly sampling subsets of data, data scaling can be seen as telling us about the average impact of data leverage action. Indeed, the experiments in Chapters 6 and 7 effectively use a modified kind of data scaling / learning curve analysis.

In theory, we could use data value estimates to guess the shape of a data scaling curve, or use a data scaling curve to guess the shape of a data value estimate distribution.

In practice, data scaling curves may be a very useful lens for data leverage organizers who are trying to make choices about the size of a campaign. An influence or Shapley value analysis may be more useful to organizers who expect the size of their campaign will be relatively fixed, but there might be opportunities to recruit specific contributors based on the contributors' data contribution characteristics.

10.2.4. Other Work that Tells Us About Data Value

Work in price targeting and marketing can also be seen as providing a kind of data valuation. For instance, Smith et al. compared a price targeting model with access to purchase histories with one that has access to demographic information. This kind of work can be seen as a comparative assessment of data value [369].

Wang and colleagues have shown it is possible to "distill" data, which provides an alternative lens on exploring fixed datasets to find the "sources of value"[417]. We could imagine developing a data value measurement that focuses on how likely a data point is to remain in a distilled dataset.

The "least core" also provides a game theoretic alternative to Shapley values for data credit assignment, with some distinct advantages in terms of computation [432]. If all data contributors see themselves as people playing a cooperative game, the least core may be of interest [432]. This is probably most applicable to some kind of data sharing platform (i.e., a market) in which all participants have opted-*in*. For these participants, the framing of the least core could be a compelling argument that the platform operators are taking contributor incentives very seriously.

While these approaches to data value all provide critical insight into the causal effect of data on model outcomes and/or the likely preferences of data-contributing agents, they do not map as cleanly to specific data leverage scenarios in the same way as influence functions, Shapley values, or data scaling curves.

10.3. Summary

We have a variety of measurements and the ability to estimate them in practice. Which measurements best correspond to data leverage phenomena of interest?

If an individual wishes to understand the immediate impact of withholding their data contributions (or adding giving data to a firm that doesn't currently have it), influence functions make sense. Even if one wants understand the the impact of a large, coordinated group, influence functions could still work, though will likely underestimate impact of withholding and overestimate impact of data contribution [207].

However, an assumption when using influence functions is that the rest of the dataset will remain static. If a user expects that a random assortment of other users may also be withholding (or contributing) data, they likely prefer to consider some Shapley variant.

As noted above, a very fruitful direction could be to investigate the use of a propensity-weighted Shapley value, where coalitions are weighted more heavily based on how they likely those coalitions are to exist in existing social conditions.

Future work should continue to synthesize key findings from relevant data valuation and data scaling work and, where possible, use these findings to make concrete predictions about the impact of data leverage. By interpreting relevant findings and definitions through the lens of data leverage, it will be possible to (1) better understand factors that influence the success of data leverage in the short term and (2) plan out a long-term data leverage research agenda that can maximally benefit from ongoing advances in related areas of ML.

CHAPTER 11

Future Work and Conclusion

In this Chapter, we discuss our planned future work to support data leverage along two broad lines. First, we discuss the potential for tools that can amplify data leverage for interested organizers and participants. This approach will involve drawing on – and contributing to – the HCI literature, and also require consideration of factors that influence collective action [286]. This software will be designed to be usable by people interested in data leverage, with three key kinds of stakeholders in mind: potential data leverage *organizers*, data leverage *participants* and interested *policymakers*. We also discuss applications for governing online communities more broadly, inspired by the parallels between aggregating data labor outputs to train a model and aggregating preference data to find a consensus (e.g. counting votes). Drawing on the metaphor that users of a recommender system are "voting" to change recommendation outputs in some way when they provide data labor, we discuss how data leverage communication tools could be designed so that they are also useful for scaffolding collective decision-making more broadly. In other words, we will work towards tools that could be used for applications ranging from "manipulate a recommender system" to "determine policies for an online community".

Second, we discuss the implications of very large models trained on massive datasets, typically scraped from the web, e.g. Large Language Models and Generative Art Models (these have also been called Foundation Models [35]). These models have played a major role in shaping public discourse about AI, i.e. in the popular press and on social media

[270]. The ethical and legal questions about data that is used to train these models (typically, massive scraped datasets of text or images from across the web) are highly relevant to data leverage. Practically speaking, the widespread adoption of large models could represent a fulcrum point, with different regulatory responses and emergent professional norms massively influencing how effective data withholding, poisoning, and contribution will be.

Finally, we summarize key ideas from throughout this thesis and paint a possible path forwards.

11.1. Towards Data Leverage Communication Tools

There are many rich directions for extending data leverage research. Here, we describe a specific tool-based approach: designing data leverage communication tools. The core idea behind this direction of research is that by communicating the likely impact of different data generating actions available to current or potential data contributors, researchers can amplify the impact of data leverage and make coordination amongst data contributors easier. Such tools would incorporate simulations, which might build on the above work on data strikes, data contribution, and data dividends, alongside observational data (i.e., observing the impacts of actual data leverage campaigns) and some kind of interface. Another name we might give these tools, less tied to the data leverage approach, could be "data impact interfaces".

We can consider these tools in terms of three distinct components: a component that makes estimates about the impact of different datasets on model performance (i.e., the *data value estimation* module), a component that models how different levels and configurations of participation will lead to different datasets (i.e., the *collective action* module), and a component that communicates these estimates (i.e., the *communication* module). In order to support this line of future work, in this section we briefly describe the challenges that are likely to arise with each component. We also discuss how each module might be relevant to the three groups of stakeholders (organizers, participants, policymakers). We highlight fruitful research questions that arise from this line of investigation.

11.1.1. Data Value Estimation

A major challenge will be selecting a definition of data value/impact that is useful for "users", including highly active organizers and more casual participants with varying degrees of interest. In chapter 10, we discussed how different impact definitions map to different real-world scenarios, e.g. leave-one-estimates tell us about the marginal impact of a specific individual (or specific group), whereas the Shapley value tells us a kind of expected value of group action averaged over many possible groups.

A second challenge will simply be finding a balance between estimation and accuracy and generalizability. Different data usage contexts (e.g. recommender system vs. computer vision classifier) show differences in data dependence [405]. A system that is meant to model a specific implementation of a recommender system operating in a particular domain may be very accurate for that domain, but very inaccurate for other domains or if implementation details are changed. Ideally, data leverage communication tools should be usable by people who may not know specific details about the systems to which they contribute data.

If policymakers (or independent research efforts) can increase the overall transparency about data-dependent systems, this may make it easier to produce accurate estimates without sacrificing generalizability (because data leverage communication tools can simply account for different system details).

Based on the work described in this thesis, we can identify some likely desirable characteristics for such systems. First, given the very different trends between withholding data and contributing data, it will likely be useful to concisely communicate theoretical insights about data scaling and data value to end users. In particular, communicating to organizers when a particular action is likely to be very ineffective (i.e., very little chance of any noticeable drop in performance) could save serious organizer resources and reduce frustration for participants.

11.1.2. Collective Action Component

A key theme throughout our research is that data's value is highly collective. An individual acting alone with their data is unlikely to impact much of anything. This means that all data leverage problems are fundamentally collective action problems, and are likely to face several kinds of *social dilemmas* [209].

Data leverage communication tools will benefit heavily from better understanding online collective action, especially in the context of data-related actions. Even a very basic understanding of whether a particular action is likely to gain or lose momentum as participation grows could help organizers select recruitment tactics (i.e., Marwell and Oliver's "accelerating and decelerating production functions" [286, 287]).

11.1.3. Interface Component

Designing an interface that communicates the likely impact of data-related collective action poses important HCI research questions. The design space of “data impact interfaces” is quite large, because there are many possible data counterfactuals that an individual can induce, and combinatorially more when we start to consider groups of people.

We might evaluate different interface designs in several ways. First, we could consider an approach that is strictly about reducing information asymmetry. This could be seen as a kind of data literacy intervention [87], evaluated primarily in terms of whether people who interact with the interface come away with more knowledge about downstream systems. Indeed, much of data leverage research could seen in terms of teaching people about data-dependent systems, but with a specific focus on systems that are particularly reliant on the public or on particular communities.

Another approach for evaluating different interface designs could be a focus on whether users more effectively solve social dilemmas when using a particular tool. In other words, in scenarios in which there is a need to gain some “critical mass” [286] of participation before any goals are achieved (e.g., most data strikes, as shown in Chapter 7), we need to determine which design choices actually accomplish this outcome. This question could be studied with both lab experiments, but ideally also with field experiments in real conditions.

Finally, we could also evaluate these interfaces in terms of their ability to foster dialogue about AI systems and facilitate governance processes. For instance, rather than presupposing that users share a goal, but are hindered by social dilemma dynamics, we might ask if data impact interfaces help a group of people decide on a goal. Looking at the impact of data on some model, perhaps some users will come away from the interface wanting to produce more data and improve that model, and others will be excited about the potential to bring that model down.

11.1.4. Connections to Community Governance

Data leverage communications tools are fundamentally concerned with reducing the information asymmetry around potential data scenarios that individual people have the agency to bring about. In many online community contexts, people may contribute data that serves some kind of governance role, e.g. posting on Reddit about one's preferred moderation policies for a particular subreddit. Or, some communities may use digital tools to engage in voting or consensus building.

Driven by the insight that many individual data labor records (e.g., "user *123* watched video *abc*") functionally act as "votes" for a model to act a certain way (e.g., to recommend video *abc* to other users), we could imagine repurposing data leverage communication tools to support communication about a much broader class of governance systems. A tool that tells someone about different counterfactual outcomes for a recommender system could also tell them about different counterfactual outcomes for a voting mechanism.

11.2. Data Leverage and Large Models

At the time this dissertation was being written, a dominant force in discussions in the AI community was the adoption of large language models [37] and other large models used for purposes such as prompt-based art generation [315]. These models create unique challenges for data leverage because of their scale. From the perspective that models that draw on data from more people are "more" reliant on data labor, these large models are some of the most data-labor-intensive systems to exist, on a similar level of data-labor-intensity as search engines.

On the other hand, they represent a unique opportunity in terms of the burst in public attention to AI systems and data. The outputs of generative art models, in particular, have fostered both very great positive [325] and negative attention [270].

These technologies also raise the stakes for discussion about the long-term impacts of AI systems more broadly. Organizations such as OpenAI have expressed interest in studying the long-term economic impacts [384]. The rapid increase in AI capabilities seems poised to unlock new kinds of systems and businesses. These large models could represent step function increases in performance, as posited in *Radical Markets* [309]. But, it could be the case that society sees a net win in terms of AI capabilities but a not loss in terms of the distribution of resources and power.

Several data leverage questions emerge from the discussion around large models. First, can data counterfactuals and data scaling laws (as discussed in Chapter 10) even be applied to these models? It seems unlikely that researchers will ever obtain ground truth data about the marginal impact of individual data points on these models; it would simply be too expensive to do so. Nonetheless, public discourse has actually centered around the marginal impact of specific artists [270], with a narrative emerging around artists who seem to have particularly large influence over model outputs expressing frustrations with their relationship to these models.

While there is certainly interest in understanding macro trends in data scaling for large models [198], the kinds of predictions that might help foster collective action (e.g., "if 1000 people withhold data, performance will drop by x percent") may be hard to make. Thus, it may be the case that ever-growing models and datasets will make it harder to make precise claims about data leverage effectiveness. Additionally, they may specifically hinder data strikes. We saw in Chapter 6 that data strikes are less effective when datasets are larger (as we might expect from basic assumptions about data scaling). Thus, if firms begin to use large models more and more, systems may become more strike resistant (though niche scenarios may remain vulnerable to data strikes).

Conversely, large models could actually empower data contribution-centered small-scale collective action. In particular, if it becomes more viable for a small group of people to pool their data together and "fine-tune" a large pre-trained model, data contribution could become the dominant type of data-related collective action.

11.3. Conclusion

In this thesis, we have argued that data contributions – from volunteers in some cases, and from unwitting members of the general public in others – are an underappreciated, yet critical dependency of many of the most important modern computing systems. In the first half, we discussed several approaches for highlighting this value. In the second half, we have discussed how researchers can play a role in empowering the public to exert leverage through these data dependencies. We concluded by setting the stage for forward-looking data leverage communication tools that could facilitate collective action around data that meaningfully aligns technology company practices and AI systems themselves with the values of the data leverage participants. By advancing this line of work, researchers can work alongside the public to mitigate the negative impacts of AI and create a paradigm in which the capabilities of systems continue to improve and the economic bounty of these systems is shared widely.

,

References

[1] 18 U.S. Code § 1030 - Fraud and related activity in connection with computers, May 2020. URL `https://www.law.cornell.edu/uscode/text/18/1030`. [Online; accessed 7. Oct. 2020].

[2] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

[3] Daron Acemoglu. Harms of AI. Working Paper 29247, National Bureau of Economic Research, September 2021. URL `https://www.nber.org/papers/w29247`. Series: Working Paper Series.

[4] Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Too much data: Prices and inefficiencies in data markets. Technical report, National Bureau of Economic Research, 2019.

[5] John T. Addison, Claus Schnabel, and Joachim Wagner. Works councils in Germany: their effects on establishment performance. *Oxford Economic Papers*, 53 (4):659–694, October 2001. ISSN 0030-7653. doi: 10.1093/oep/53.4.659. URL `https://doi.org/10.1093/oep/53.4.659`.

[6] Partnership AI. About - The Partnership on AI. `https://www.partnershiponai.org/about/#pillar-3`, 2022.

[7] Kendra Albert, Jon Penney, Bruce Schneier, and Ram Shankar Siva Kumar. Politics of adversarial machine learning. In *Towards Trustworthy ML: Rethinking Security and Privacy for ML Workshop, Eighth International Conference on Learning Representations (ICLR)*, 2020.

[8] Alberto Alesina and Roberto Perotti. Income distribution, political instability, and investment. *European economic review*, 40(6):1203–1228, 1996.

[9] Alexa.com. Alexa Top 500 Global Sites. http://www.alexa.com/topsites, 2018.

[10] Ali Alkhatib, Michael S. Bernstein, and Margaret Levi. Examining Crowd Work and Gig Work Through The Historical Lens of Piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 4599–4616, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025974.

[11] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]*, July 2016. URL `http://arxiv.org/abs/1606.06565`. arXiv: 1606.06565.

[12] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 850–858, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339665.

[13] Imanol Arrieta Ibarra, Leonard Goff, Diego Jiménez Hernández, Jaron Lanier, and E Weyl. Should We Treat Data as Labor? Moving Beyond 'Free'. *American Economic Association Papers & Proceedings*, 1(1), 2018.

[14] Amanda Askell, Miles Brundage, and Gillian Hadfield. The role of cooperation in responsible ai development. *arXiv preprint arXiv:1907.04534*, 2019.

[15] Associated Press. The Latest: Facebook users await word on privacy scandal - The Washington Post. `https://www.washingtonpost.com/national/the-latest-facebook-users-await-word-on-privacy-scandal/2018/04/09/37e8d82a-3c2a-11e8-955b-7d2e19b79966_story.html`, 2018.

[16] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46 (3):399–424, 2011.

[17] Stefan Baack. Datafication and empowerment: How the open data movement rearticulates notions of democracy, participation, and journalism. *Big Data & Society*, 2(2):2053951715594634, 2015. doi: 10.1177/2053951715594634. URL `https://doi.org/10.1177/2053951715594634`.

[18] Michael Barbaro and Tom Zeller, Jr. A Face Is Exposed for AOL Searcher No. 4417749. *N.Y. Times*, August 2006. ISSN 0362-4331.

[19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[20] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pages 16–25, 2006.

[21] Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International conference on learning representations (ICLR)*, 2021.

[22] Eric PS Baumer. Socioeconomic inequalities in the non use of facebook. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[23] Eric PS Baumer, Phil Adams, Vera D Khovanskaya, Tony C Liao, Madeline E Smith, Victoria Schwanda Sosik, and Kaiton Williams. Limiting, leaving, and (re) lapsing: an exploration of facebook non-use practices and experiences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3257–3266, 2013.

[24] Eric PS Baumer, Morgan G. Ames, Jed R. Brubaker, Jenna Burrell, and Paul Dourish. Refusing, limiting, departing: Why we should study technology non-use. In *CHI EA '14: CHI '14 Extended Abstracts on Human Factors in Computing*

Systems, pages 65–68, 2014.

[25] Eric PS Baumer, Shion Guha, Emily Quan, David Mimno, and Geri K Gay. Missing photos, suffering withdrawal, or finding freedom? how experiences of social media non-use influence the likelihood of reversion. *Social Media+ Society*, 1(2): 2056305115614851, 2015.

[26] Eric PS Baumer, Shion Guha, Patrick Skeba, and Geraldine Gay. All users are (not) created equal: Predictors vary for different forms of facebook non/use. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28, 2019.

[27] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839, May 2020. ISSN 2334-0770. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/7347`.

[28] Eric Bax. Computing a Data Dividend, June 2019.

[29] Omri Ben-Shahar. Data Pollution. *Journal of Legal Analysis*, 11:104–159, 2019. Publisher: Narnia.

[30] Ruha Benjamin. Informed refusal: Toward a justice-based bioethics. *Science, Technology, & Human Values*, 41(6):967–990, 2016. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[31] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

[32] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159, 2018.

[33] Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

[34] Michael Bloodgood and Chris Callison-Burch. Bucking the trend: large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 854–864. Association for Computational Linguistics, 2010.

[35] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma,

Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*, August 2021.

[36] Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David Smeddinck. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–19, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3501947.

[37] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. *Language Models Are Few-Shot Learners*. arXiv, 2020.

[38] Jed R Brubaker, Mike Ananny, and Kate Crawford. Departing glances: A sociotechnical account of 'leaving'grindr. *New Media & Society*, 18(3):373–390, 2016.

[39] Finn Brunton and Helen Fay Nissenbaum. *Obfuscation: A User's Guide for Privacy and Protest*. MIT Press, Cambridge, Massachusetts, 2015. ISBN 978-0-262-02973-5.

[40] Erik Brynjolfsson and Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company, January 2014. ISBN 978-0-393-23935-5.

[41] Erik Brynjolfsson and Kristina McElheran. The rapid adoption of data-driven decision-making. *American Economic Review*, 106(5):133–39, 2016.

[42] Erik Brynjolfsson, Andrew McAfee, and Michael Spence. New world order: Labor, capital, and ideas in the power law economy. *Foreign Affairs*, 93(4):44–53, 2014.

[43] Ceren Budak, Sharad Goel, Justin Rao, and Georgios Zervas. Understanding Emerging Threats to Online Advertising. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, pages 561–578, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3936-0. doi: 10.1145/2940716.2940787. URL `http://doi.acm.org/10.1145/2940716.2940787`. event-place: Maastricht, The Netherlands.

[44] US Census Bureau. HINC-06. Income Distribution to $250,000 or More for Households. https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-06.html, 2021.

[45] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank Using Gradient Descent. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 89–96, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi: 10.1145/1102351.1102363.

[46] Fabio Calefato, Filippo Lanubile, Maria Concetta Marasciulo, and Nicole Novielli. Mining successful answers in stack overflow. In *Mining Software Repositories (MSR), 2015 IEEE/ACM 12th Working Conference On*, pages 430–433. IEEE, 2015.

[47] Nathalie Casemajor, Stıfmmode\acutee\elseé\fiphane Couture, Mauricio Delfin, Matthew Goerzen, and Alessandro Delfanti. Non-participation in digital media: toward a framework of mediated political action. *Media, Culture & Society*, 37(6): 850–866, May 2015. ISSN 0163-4437. doi: 10.1177/0163443715584098. Publisher: SAGE Publications Ltd.

[48] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 1–14. ACM, 2007.

[49] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):147:1–147:32, November 2019. doi: 10.1145/3359249.

[50] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW):1–32, 2019.

[51] Stevie Chancellor, Shion Guha, Jofish Kaye, Jen King, Niloufar Salehi, Sarita Schoenebeck, and Elizabeth Stowell. The Relationships between Data, Power, and Justice in CSCW Research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, CSCW '19, pages 102–105, New York, NY, USA, November 2019. Association for Computing Machinery. ISBN 978-1-4503-6692-2. doi: 10.1145/3311957.3358609. URL `https://doi.org/10.1145/3311957.3358609`.

[52] Danqi Chen, Weizhu Chen, Haixun Wang, Zheng Chen, and Qiang Yang. Beyond Ten Blue Links: Enabling User Click Modeling in Federated Web Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 463–472, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124351.

[53] Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.

[54] Paul-Alexandru Chirita, Wolfgang Nejdl, and Cristian Zamfir. Preventing shilling attacks in online recommender systems. In *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, pages 67–74, 2005.

[55] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.

[56] Max Cho. Unsell yourself—a protest model against facebook. *Yale Law & Technology*, 2011.

[57] Giovanni Luca Ciampaglia and Dario Taraborelli. MoodBar: Increasing new user retention in Wikipedia through lightweight socialization. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 734–742. ACM, 2015.

[58] Marika Cifor, Patricia Garcia, TL Cowan, Jasmine Rault, Tonia Sutherland, Anita Say Chan, Jennifer Rode, Anna Lauren Hoffmann, Niloufar Salehi, and Lisa Nakamura. Feminist data manifest-no, 2019.

[59] Jack Clark. Google Turning Its Lucrative Web Search Over to AI Machines - Bloomberg. https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines, 2015.

[60] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. In *NIPS*, 2017.

[61] Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

[62] Ashley Colley, Jacob Thebault-Spieker, Allen Yilun Lin, Donald Degraen, Benjamin Fischman, Jonna Häkkilä, Kate Kuehl, Valentina Nisi, Nuno Jardim Nunes, Nina Wenig, et al. The geography of Pokémon GO: Beneficial and problematic effects on places and movement. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1179–1192. ACM, 2017.

[63] Avinash Collis, Alex Moehring, Ananya Sen, and Alessandro Acquisti. Information Frictions and Heterogeneity in Valuations of Personal Data. SSRN Scholarly Paper ID 3974826, Social Science Research Network, Rochester, NY, November 2021. URL `https://papers.ssrn.com/abstract=3974826`.

[64] ComScore. ComScore US Search Market Share. `https://www.comscore.com/Insights/Rankings?country=US#tab_search_share`, February 2020.

[65] Kate Conger, Richard Fausset, and Serge F. Kovaleski. San Francisco Bans Facial Recognition Technology. *N.Y. Times*, May 2019. ISSN 0362-4331.

[66] R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

[67] R Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508,

1980.

[68] Nick Couldry and Alison Powell. Big data from the bottom up. *Big Data & Society*, 1(2):2053951714539277, 2014.

[69] Josie Cox. Automation risks exacerbating income inequality across UK, think tank warns. *The Independent*, 2017.

[70] Kate Crawford and Vladan Joler. Anatomy of an ai system-the amazon echo as an anatomical map of human labor, data and planetary resources. *AI Now Institute and Share Lab*, 7, 2018.

[71] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 39–46, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-906-0. doi: 10.1145/1864708.1864721.

[72] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1): 187–199, 2009.

[73] Era Dabla-Norris, Kalpana Kochhar, Nujin Suphaphiphat, Frantisek Ricka, and Evridiki Tsounta. *Causes and Consequences of Income Inequality: A Global Perspective.* International Monetary Fund, 2015.

[74] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 101–109. ACM, 2019.

[75] Allan Dafoe. AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018.

[76] Jeff Daniels. California governor proposes 'new data dividend' that could call on Facebook and Google to pay users. https://www.cnbc.com/2019/02/12/california-gov-newsom-calls-for-new-data-dividend-for-consumers.html, 2019.

[77] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[78] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110. ACM, 2016.

[79] Vasant Dhar and Elaine A Chang. Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing*, 23(4):300–307, 2009.

[80] Thomas G Dietterich and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Technical report, Department of Computer Science, Oregon State University, 1995.

[81] Catherine D'Ignazio and Lauren F Klein. *Data feminism.* MIT Press, 2020.

[82] Catherine D'Ignazio, Erhardt Graeff, Christina N. Harrington, and Daniela K. Rosner. Toward Equitable Participatory Design: Data Feminism for CSCW amidst Multiple Pandemics. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pages 437–445. Association for Computing Machinery, New York, NY, USA, October 2020. ISBN 978-1-4503-8059-1. URL https://doi.org/10.1145/3406865.3418588.

[83] Cory Doctorow. Regulating Big Tech makes them stronger, so they need competition instead. *The Economist*, June 2019.

[84] Carroll Doherty and Jocelyn Kiley. Americans have become much less positive about tech companies' impact on the U.S., 2019.

[85] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610. ACM, 2014.

[86] Michaelanne Dye, David Nemer, Laura R Pina, Nithya Sambasivan, Amy S Bruckman, and Neha Kumar. Locating the internet in the parks of havana. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3867–3878, 2017.

[87] Catherine D'Ignazio and Rahul Bhargava. Approaches to building big data literacy. In *Proceedings of the Bloomberg data for good exchange conference*, page 6, 2015.

[88] Jennifer Earl and Katrina Kimport. *Digitally Enabled Social Change: Activism in the Internet Age.* Mit Press, 2011.

[89] Dean Eckles and Eytan Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*, 2017.

[90] The Economist. Should internet firms pay for the data users currently give away? *The Economist*, 2018.

[91] Wikipedia editors. Protests against SOPA and PIPA. https://en.wikipedia.org/wiki/Protests_against_SOPA_and_PIPA, .

[92] Wikipedia editors. Template:Grading scheme. https://en.wikipedia.org/wiki/Template:Grading_scheme, .

[93] Wikipedia editors. Reusing wikipedia content. https://en.wikipedia.org/wiki/Wikipedia:Reusing_Wikipedia_content, .

[94] Wikipedia editors. Statistics. https://en.wikipedia.org/wiki/Wikipedia:Statistics, .

[95] Wikipedia editors. Wikipedia:Academic studies of Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Academic_studies_of_Wikipedia&oldid=971074694, Sep 2020. [Online; accessed 29. Sep. 2020].

[96] Wikipedia editors. Web search engine. https://en.wikipedia.org/wiki/Web_search_engine#Market_share, February 2020.

[97] Wikipedia editors. Big Tobacco. `https://en.wikipedia.org/wiki/Big_Tobacco`, 2022.

[98] Christopher B Eiben, Justin B Siegel, Jacob B Bale, Seth Cooper, Firas Khatib, Betty W Shen, Barry L Stoddard, Zoran Popovic, and David Baker. Increased diels-alderase activity through backbone remodeling guided by foldit players. *Nature biotechnology*, 30(2):190–192, 2012.

[99] Hamid Ekbia and Bonnie Nardi. Social Inequality and HCI: The View from Political Economy. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4997–5002, New York, NY, USA, May 2016. Association for Computing Machinery. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858343.

[100] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. volume 81 of *Proceedings of Machine Learning Research*, pages 35–47, New York, NY, USA, 23–24 Feb 2018. PMLR. URL `http://proceedings.mlr.press/v81/ekstrand18a.html`.

[101] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2*, IJCAI'01, pages 973–978, San Francisco, CA, USA, August 2001. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-812-2.

[102] Kyle Endres and Costas Panagopoulos. Boycotts, buycotts, and political consumerism in America. *Research & Politics*, 4(4):2053168017738632, 2017.

[103] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

[104] Farzad Eskandanian, Nasim Sonboli, and Bamshad Mobasher. Power of the Few: Analyzing the Impact of Influential Users in Collaborative Recommender Systems. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 225–233, 2019.

[105] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. "be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *ICWSM*, pages 62–71, 2017.

[106] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[107] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018.

[108] Microsoft Research Events. OSSM17: Observational Studies Through Social Media, 2017.

[109] Minghong Fang, Neil Zhenqiang Gong, and Jia Liu. *Influence Function based Data Poisoning Attacks to Top-N Recommender Systems.* Association for Computing Machinery, New York, NY, USA, Apr 2020. ISBN 978-1-45037023-3. doi: 10.1145/

3366423.3380072.

[110] Paolo Ferragina and Ugo Scaiella. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1625–1628. ACM, 2010.

[111] Yakov Feygin, Hanlin Li, Chirag Lala, Brent Hecht, Nicholas Vincent, Luisa Scarcella, and Matthew Prewitt. A data dividend that works: steps toward building an equitable data economy, 2021.

[112] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. Predicting sample size required for classification performance. *BMC medical informatics and decision making*, 12(1):8, 2012. URL `https://link.springer.com/article/10.1186/1472-6947-12-8`. Publisher: Springer.

[113] Forbes. The World's Biggest Public Companies List. `https://www.forbes.com/global2000/list/#tab:overall`, 2018.

[114] Sarah Fox, Mariam Asad, Katherine Lo, Jill P. Dimond, Lynn S. Dombrowski, and Shaowen Bardzell. Exploring Social Justice, Design, and HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 3293–3300, New York, NY, USA, May 2016. Association for Computing Machinery. ISBN 978-1-4503-4082-3. doi: 10.1145/2851581.2856465.

[115] Monroe Friedman. A positive approach to organized consumer action: The "buycott" as an alternative to the boycott. *Journal of Consumer Policy*, 19(4):439–451, 1996.

[116] Simon Funk. Netflix Update: Try This at Home. `http://sifter.org/simon/journal/20061211.html`.

[117] Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL `https://doi.org/10.1007/s11023-020-09539-2`.

[118] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611, 2007.

[119] David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. Social resilience in online communities: The autopsy of friendster. In *Proceedings of the first ACM conference on Online social networks*, pages 39–50, 2013.

[120] Patricia Garcia, Tonia Sutherland, Marika Cifor, Anita Say Chan, Lauren Klein, Catherine D'Ignazio, and Niloufar Salehi. No: Critical refusal as feminist data practice. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pages 199–202, 2020.

[121] Timnit Gebru. Oxford handbook on ai ethics book chapter on race and gender. *arXiv preprint arXiv:1908.06165*, 2019.

[122] R. Stuart Geiger and Aaron Halfaker. Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative*

work, CSCW '13, pages 861–870, New York, NY, USA, February 2013. Association for Computing Machinery. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441873. URL `https://doi.org/10.1145/2441776.2441873`.

[123] Jonas Geiping, Liam Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching, 2020.

[124] Media Genesis. Popular Screen Resolutions - Media Genesis. https://mediag.com/blog/popular-screen-resolutions-designing-for-all, March 2018.

[125] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. Learning image and user features for recommendation in social networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4274–4282, 2015.

[126] Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning*, 2019.

[127] Amirata Ghorbani, James Zou, and Andre Esteva. Data Shapley Valuation for Efficient Batch Active Learning. *arXiv:2104.08312 [cs, stat]*, April 2021. URL `http://arxiv.org/abs/2104.08312`. arXiv: 2104.08312.

[128] Eric Gilbert. Widespread Underprovision on Reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 803–808, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441866.

[129] Tarleton Gillespie. Algorithmically recognizable: Santorum's google problem, and google's santorum problem. *Information, communication & society*, 20(1):63–80, 2017.

[130] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[131] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d'Alessandro, Joëlle Tilmanne, Todd Kulesza, and Baptiste Caramiaux. Human-Centred Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 3558–3565, New York, NY, USA, May 2016. Association for Computing Machinery. ISBN 978-1-4503-4082-3. doi: 10.1145/2851581.2856492. URL `https://doi.org/10.1145/2851581.2856492`.

[132] Pauline Glikman and Nicolas Glady. What's The Value Of Your Data? *TechCrunch*, October 2015.

[133] Michael Golebiewski and Danah Boyd. Data voids: Where missing data can easily be exploited. *Data & Society*, 2019.

[134] Carlos Gómez, Brendan Cleary, and Leif Singer. A Study of Innovation Diffusion Through Link Sharing on Stack Overflow. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 81–84, Piscataway, NJ,

USA, 2013. IEEE Press. ISBN 978-1-4673-2936-1.

[135] Carlos A. Gomez-Uribe and Neil Hunt. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4): 13:1–13:19, December 2015. ISSN 2158-656X. doi: 10.1145/2843948.

[136] José González Cabañas, Angel Cuevas, and Rubén Cuevas. FDVT: Data Valuation Tool for Facebook Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3799–3809. ACM, 2017.

[137] Danny Goodwin. Bing, Not Google, Favors Wikipedia More Often in Search Results [Study] - Search Engine Watch. Technical report, March 2012.

[138] Danny Goodwin. Wikipedia Appears on Page 1 of Google for 99% of Searches [Study] - Search Engine Watch. Technical report, February 2012.

[139] Google. Google BigQuery. `https://bigquery.cloud.google.com/dataset/bigquery-public-data:stackoverflow`.

[140] Google. Google Trends. https://trends.google.com/trends/hottrends, 2018.

[141] Jennifer Granholm and Chris Eldred. Facebook owes you money (opinion) - CNN, 2018. URL `https://www.cnn.com/2018/04/11/opinions/facebook-should-pay-us-for-using-our-data-granholm-eldred/index.html`.

[142] Gary L Graunke. *Using Predictive Traffic Modeling.* Google Patents, July 2001.

[143] Ben Green. ‘Fair’ Risk Assessments: A Precarious Approach for Criminal Justice Reform. In *5th Workshop on fairness, accountability, and transparency in machine learning*, 2018.

[144] Rebecca Greenfield, Sarah Frier, and Ben Brody. NAACP Seeks Week-Long Facebook Boycott Over Racial Targeting. *Bloomberg.com*, December 2018. URL `https://www.bloomberg.com/news/articles/2018-12-17/naacp-calls-for-week-long-facebook-boycott-over-racial-targeting`.

[145] Erin Griffith. Stack Exchange, a site for software developers, raises $40 million | Fortune.com, 2015. URL `http://fortune.com/2015/01/20/stack-exchange-40-million/`.

[146] Artem Grotov and Maarten de Rijke. Online Learning to Rank for Information Retrieval: SIGIR 2016 Tutorial. In *SIGIR ’16: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1215–1218. Association for Computing Machinery, New York, NY, USA, July 2016. ISBN 978-1-4503-4069-4. doi: 10.1145/2911451.2914798.

[147] Jonathan Grudin. Cscw and groupware: Their history and trajectory. *Designing communication and collaboration support systems*, 2:68, 1999.

[148] Ihsan Gunes, Cihan Kaleli, Alper Bilge, and Huseyin Polat. Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review*, 42 (4):767–799, 2014. Publisher: Springer.

[149] Michael Gurstein. Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 16, 02 2011. doi: 10.5210/fm.v16i2.3316.

[150] Chai Haiyan. An impact of social media on online travel information search in China. In *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering*, pages 509–512. IEEE, 2010.

[151] Aaron Halfaker and Dario Taraborelli. Artificial intelligence service "ORES" gives Wikipedians X-ray specs to see through bad edits, November 2015. URL `https://diff.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/`.

[152] Aaron Halfaker, Aniket Kittur, and John Riedl. Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 163–172. ACM, 2011.

[153] Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. Making peripheral participation legitimate: Reader engagement experiments in wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 849–860. ACM, 2013.

[154] Jeffrey T Hancock, Catalina Toma, and Nicole Ellison. The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 449–452, 2007.

[155] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 527–538. ACM, 2013.

[156] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 305–318, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3213-2. doi: 10.1145/2663716.2663744.

[157] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 449. ACM, 2018.

[158] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, December 2015. ISSN 2160-6455. doi: 10.1145/2827872.

[159] F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. Predictors of Answer Quality in Online Q&A Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 865–874, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357191.

[160] Jason D Hartline. Mechanism design and approximation. *Book draft. October*, 122: 1, 2013.

[161] B Hecht, L Wilcox, JP Bigham, J Schöning, E Hoque, J Ernst, Y Bisk, L De Russis, L Yarosh, B Anjum, and others. It's time to do something: Mitigating the negative

impacts of computing through a change to the peer review process., 2018.

[162] Brent Hecht. HCI and the U.S. Presidential Election: A Few Thoughts on a Research Agenda. https://medium.com/@BrentH/hci-and-the-u-s-presidential-election-a-few-thoughts-on-a-research-agenda-7c1a0a04986, 2016.

[163] Brent Hecht. HCI and the U.S. Presidential Election: A Few Thoughts on a Research Agenda. In *CHI '18 Panel Presentation*, Denver, CO, 2017.

[164] Brent Hecht and Darren Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the Fourth International Conference on Communities and Technologies*, pages 11–20. ACM, 2009.

[165] Brent Hecht and Darren Gergle. The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 291–300. ACM, 2010.

[166] Brent Hecht and Monica Stephens. A Tale of Cities: Urban Biases in Volunteered Geographic Information. *ICWSM*, 14:197–205, 2014.

[167] Brent Hecht and Lauren Wilcox. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process - ACM FCA. https://acm-fca.org/2018/03/29/negativeimpacts/.

[168] Brent Hecht, Johannes Schöning, Muki Haklay, Licia Capra, Afra J Mashhadi, Loren Terveen, and Mei-Po Kwan. Geographic human-computer interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 3163–3166. ACM, 2013.

[169] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[170] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. WikiReading: A Novel Large-scale Language Understanding Task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1535–1545, 2016.

[171] Benjamin Mako Hill and Aaron Shaw. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one*, 8(6): e65782, 2013.

[172] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. *International AAAI Conference on Web and Social Media; Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[173] Marit Hinnosaar, Toomas Hinnosaar, Michael E Kummer, and Olga Slivko. Wikipedia matters. *Available at SSRN 3046400*, 2019.

[174] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429. International World Wide Web Conferences Steering Committee, 2015.

[175] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[176] Daniel C Howe and Helen Nissenbaum. Engineering privacy and protest: A case study of adnauseam. In *IWPE@ SP*, pages 57–64, 2017.

[177] Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 179–186. ACM, 2008.

[178] Shanshan Huang, Shuaiqiang Wang, Tie-Yan Liu, Jun Ma, Zhumin Chen, and Jari Veijalainen. Listwise Collaborative Filtering. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 343–352, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. doi: 10.1145/2766462.2767693.

[179] Nicolas Hug. *Surprise, a Python Library for Recommender Systems.* 2017.

[180] Chris Hughes. The wealth of our collective data should belong to all of us. *The Guardian*, 2018.

[181] Kate Mathews Hunt. Gaming the system: Fake online reviews v. consumer law. *Computer law & security review*, 31(1):3–25, 2015.

[182] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting Predictions from Training Data. *arXiv:2202.00622 [cs, stat]*, February 2022.

[183] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

[184] Deborah D Ingram and Sheila J Franco. 2013 NCHS urban-rural classification scheme for counties. 2014.

[185] Chitika Insights. The value of Google result positioning. *Retrived*, 29:2015, 2013.

[186] Sarah J Jackson, Moya Bailey, and Brooke Foucault Welles. *# HashtagActivism: Networks of Race and Gender Justice.* MIT Press, 2020.

[187] Ross James. How to use Google Takeout to download your Google data - Business Insider. *Business Insider*, Jan 2020. URL `https://www.businessinsider.com/what-is-google-takeout`.

[188] Bernard J Jansen, Danielle L Booth, and Amanda Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th International Conference on World Wide Web*, pages 1149–1150. ACM, 2007.

[189] Greg Jarboe. YouTube's Organic Visibility Tops Wikipedia in Google SERPs. *Search Engine Journal*, January 2020.

[190] Adrianne Jeffries and Leon Yin. Google's Top Search Result? Surprise! It's Google. *The Markup*, July 2020.

[191] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards Efficient Data Valuation Based on the Shapley Value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176, 2019.

[192] Seunga Venus Jin, Joe Phua, and Kwan Min Lee. Telling stories about breastfeeding through Facebook: The impact of user-generated content (UGC) on pro-breastfeeding attitudes. *Computers in Human Behavior*, 46:6–17, 2015.

[193] Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. Global gender differences in Wikipedia readership. *arXiv preprint arXiv:2007.10403*, 2020.

[194] Isaac L Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 13–25. ACM, 2016.

[195] Lauren Johnson. IAB Study Says 26% of Desktop Users Turn On Ad Blockers – Adweek, 2016. URL `http://www.adweek.com/digital/iab-study-says-26-desktop-users-turn-ad-blockers-172665/`.

[196] Charles I Jones and Christopher Tonetti. Nonrivalry and the Economics of Data. Technical report, National Bureau of Economic Research, 2019.

[197] Hyun Joon Jung, Yubin Park, and Matthew Lease. Predicting next label quality: A time-series model of crowdwork. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[198] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*, January 2020.

[199] Kevin Granville. Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens - The New York Times. https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html, 2018.

[200] Tae Wan Kim, Jooho Lee, Joseph Xu, and Bryan Routledge. Corporate Data Governance: Are Data Subjects Investors? *Academy of Management Proceedings*, 2020(1):13855, August 2020. ISSN 0065-0668. doi: 10.5465/AMBPP.2020.287.

[201] Aniket Kittur and Robert E Kraut. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 37–46. ACM, 2008.

[202] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, February 2013. Association for

Computing Machinery. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441923.

[203] Brian Klais. User Generated Content Offers Significant SEO Benefits. https://searchengineland.com/user-generated-content-offfers-significant-seo-benefits-36037, 2010.

[204] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 121–127. ACM, 2015.

[205] Noam Koenigstein. Rethinking Collaborative Filtering: A Practical Perspective on State-of-the-art Research Based on Real World Insights. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 336–337. ACM, 2017.

[206] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. PMLR, July 2017. URL `https://proceedings.mlr.press/v70/koh17a.html`. ISSN: 2640-3498.

[207] Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the Accuracy of Influence Functions for Measuring Group Effects. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/a78482ce76496fcf49085f2190e675b4-Abstract.html`.

[208] Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. Prediction of learning curves in machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22–30, 2012.

[209] Peter Kollock. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, 24(1):183–214, August 1998. ISSN 0360-0572, 1545-2115. doi: 10.1146/annurev.soc.24.1.183.

[210] Sebastian Koos. What drives political consumption in Europe? A multi-level analysis on individual characteristics, opportunity structures and globalization. *Acta Sociologica*, 55(1):37–57, March 2012. ISSN 0001-6993. doi: 10.1177/0001699311431594.

[211] Yehuda Koren. Factor in the Neighbors: Scalable and Accurate Collaborative Filtering. *ACM Trans. Knowl. Discov. Data*, 4(1):1:1–1:24, January 2010. ISSN 1556-4681. doi: 10.1145/1644873.1644874.

[212] Anton Korinek and Joseph E. Stiglitz. Covid-19 driven advances in automation and artificial intelligence risk exacerbating economic inequality. *BMJ*, 372:n367, March 2021. ISSN 1756-1833. doi: 10.1136/bmj.n367. URL `https://www.bmj.com/content/372/bmj.n367`. Publisher: British Medical Journal Publishing Group Section: Analysis.

[213] Maya Kosoff. YouTube Slaps a Feel-Good Band-Aid on Its Fake-News Problem. *Vanity Fair*, March 2018.

[214] Oluwasanmi Koyejo, Sreangsu Acharyya, and Joydeep Ghosh. Retargeted Matrix Factorization for Collaborative Filtering. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 49–56, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2409-0. doi: 10.1145/2507157.2507185.

[215] Robert E Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. *Building Successful Online Communities: Evidence-based Social Design.* Mit Press, 2012.

[216] Kristine Phillips and Brian Fung. Facebook admits social media sometimes harms democracy - The Washington Post. https://www.washingtonpost.com/news/the-switch/wp/2018/01/22/facebook-admits-it-sometimes-harms-democracy/?utm_term=.f0cf046c1ebe, 2018.

[217] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[218] Logan Kugler. The War over the Value of Personal Data. *Commun. ACM*, 61(2): 17–19, January 2018. ISSN 0001-0782. doi: 10.1145/3171580.

[219] Noah Kulwin. Reddit's plan to become a real business could fall apart pretty easily - Vox, 2016. URL `https://www.vox.com/2016/4/28/11586522/reddit-advertising-sales-plans`.

[220] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. POTs: Protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 177–188, 2020.

[221] Shamanth Kumar, Reza Zafarani, and Huan Liu. Understanding user migration patterns in social media. In *AAAI*, volume 11, pages 8–11, 2011.

[222] Okyu Kwon, Woo-Sik Son, and Woo-Sung Jung. The double power law in human collaboration behavior: The case of wikipedia. *Physica A: Statistical Mechanics and its Applications*, 461:85–91, 2016.

[223] Yongchan Kwon and James Zou. Beta Shapley: A Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 8780–8802. PMLR, May 2022.

[224] Shyong K Lam and John Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th International Conference on World Wide Web*, pages 393–402. ACM, 2004.

[225] Cliff Lampe. "Computer induced income inequality is the most important HCI issue of our time says @bhecht. Preach, brother! #CHI2017". https://twitter.com/clifflampe/status/862701057687904256, May 2017.

[226] Cliff Lampe, Nicole B. Ellison, and Charles Steinfeld. Changes in use and perception of facebook. In *Proceedings of the 2008 conference on Computer supported cooperative work*, pages 721–730, 2008.

[227] Cliff Lampe, Jessica Vitak, and Nicole Ellison. Users and nonusers: Interactions between levels of adoption and social capital. In *Proceedings of the 2013 conference*

on Computer supported cooperative work, pages 809–820, 2013.

[228] Jaron Lanier. *Who Owns the Future?* Simon and Schuster, 2014.

[229] Jaron Lanier and E Glen Weyl. A Blueprint for a Better Digital Society. *Harvard Business Review*, 2018.

[230] Robert P Latham, Carl C Butzer, and Jeremy T Brown. Legal implications of user-generated content: YouTube, MySpace, Facebook. *Intellectual Property & Technology Law Journal*, 20(5):1–11, 2008.

[231] Michaël R Laurent and Tim J Vickers. Seeking health information online: Does Wikipedia matter? *Journal of the American Medical Informatics Association*, 16 (4):471–479, 2009.

[232] Neil D. Lawrence and Raquel Urtasun. Non-linear Matrix Factorization with Gaussian Processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 601–608, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553452.

[233] Quoc V Le and Mike Schuster. A neural network for machine translation, at production scale. Technical report, 2016.

[234] Alex Leavitt and Joshua A. Clark. Upvoting Hurricane Sandy: Event-based News Production Processes on a Social News Site. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1495–1504, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557140.

[235] Colin Lecher. California just passed one of the toughest data privacy laws in the country. *The Verge*, 2018.

[236] Jong-Seok Lee and Dan Zhu. Shilling attack detection—a new approach for a trustworthy recommender system. *INFORMS Journal on Computing*, 24(1):117–131, 2012. Publisher: INFORMS.

[237] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1603–1612, 2015.

[238] Tuukka Lehtiniemi and Minna Ruckenstein. The social imaginaries of data activism. *Big Data & Society*, 6(1):2053951718821146, 2018.

[239] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. In *Advances in neural information processing systems*, pages 1885–1893, 2016.

[240] Hanlin Li and Brent Hecht. 3 stars on yelp, 4 stars on google maps: A cross-platform examination of restaurant ratings. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW), 2020.

[241] Hanlin Li, Bodhi Alarcon, Sara M. Espinosa, and Brent Hecht. Out of Site: Empowering a New Approach to Online Boycotts. *Proceedings of the 2018 Computer-Supported Cooperative Work and Social Computing (CSCW'2018 / PACM)*, 2018.

[242] Hanlin Li, Nicholas Vincent, Janice Tsai, Jofish Kaye, and Brent Hecht. How do people change their technology use in protest? Understanding "Protest Users". *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–22, 2019. Publisher: ACM New York, NY, USA.

[243] Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576, 2014.

[244] Wendy CY Li, Makoto Nirei, and Kazufumi Yamana. Value of data: there's no such thing as a free lunch in the digital economy. *US Bureau of Economic Analysis Working Paper, Washington, DC*, 2019.

[245] Cassandra Liem and Georgios Petropoulos. The economic value of personal data for online platforms, firms and consumers, 2016.

[246] Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. Measuring the Effect of Training Data on Deep Learning Predictions via Randomized Experiments. In *Proceedings of the 39th International Conference on Machine Learning*, pages 13468–13504. PMLR, June 2022.

[247] Yilun Lin, Bowen Yu, Andrew Hall, and Brent Hecht. Problematizing and addressing the article-as-concept assumption in wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2052–2067, 2017.

[248] Jun Liu and Sudha Ram. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)*, 2(2):11, 2011.

[249] Jigsaw LLC. Toxic Comment Classification Challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, January 2020.

[250] Steve Lohr. Calls Mount to Ease Big Tech's Grip on Your Data. *New York Times*, August 2019.

[251] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. TextBlob: Simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 2014.

[252] Michael Luca. Reviews, Reputation, and Revenue: The Case of Yelp.com. September 2011.

[253] Michael Luca. User-generated content and social media. In *Handbook of Media Economics*, volume 1, pages 563–592. Elsevier, 2015.

[254] Emma Lurie and Eni Mustafaraj. Investigating the Effects of Google's Search Engine Result Page in Evaluating the Credibility of Online News Sources. In *Proceedings of the 10th ACM Conference on Web Science*, pages 107–116, 2018.

[255] Kim Lyons. FTC settles with photo storage app that pivoted to facial recognition. *The Verge*, January 2021. Publisher: The Verge.

[256] Ian Mackenzie, Chris Meyer, and Steve Noble. How retailers can keep up with consumers. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers, 2013.

[257] Saadia Madsbjerg. It's Time to Tax Companies for Using Our Personal Data - The New York Times. `https://www.nytimes.com/2017/11/14/business/dealbook/taxing-companies-for-using-our-personal-data.html`, 2017.

[258] Sam Manning, Pamela Mishkin, Gillian Hadfield, Tyna Eloundou, and Emily Eisner. A research agenda for assessing the economic impacts of code generation models. 2022.

[259] Helen Margetts, Peter John, Scott Hale, and Taha Yasseri. *Political turbulence: How social media shape collective action*. Princeton University Press, 2015.

[260] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. Characterizing the use of browser-based blocking extensions to prevent online tracking. In *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*, pages 103–116, 2018.

[261] J Nathan Matias. Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 1138–1151, 2016.

[262] Julian McAuley. *Personalized Machine Learning*. Cambridge University Press, 2022.

[263] James McCorriston, David Jurgens, and Derek Ruths. Organizations are users too: Characterizing and detecting the presence of organizations on Twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[264] McGovern. United States General Election Presidential Results by County from 2008 to 2016 Github Repository, 2017.

[265] Connor McMahon, Isaac L Johnson, and Brent Hecht. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *ICWSM*, pages 142–151, 2017.

[266] Robert Mcmillan. Google, Apple tap crowdsourcing to map out WiFi locations | IT Business, 2010. URL `https://www.itbusiness.ca/news/google-apple-tap-crowdsourcing-to-map-out-wifi-locations/15658`.

[267] Olena Medelyan, David Milne, Catherine Legg, and Ian H Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, 2009.

[268] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):115:1–115:35, July 2021. ISSN 0360-0300. doi: 10.1145/3457607.

[269] Amanda Menking and Ingrid Erickson. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 207–210. ACM, 2015.

[270] Rachel Metz. These artists found out their work was used to train AI. Now they're furious | CNN Business, October 2022. URL https://www.cnn.com/2022/10/21/tech/artists-ai-images/index.html.

[271] Stefania Milan and Lonneke Van der Velden. The alternative epistemologies of data activism. *Digital Culture & Society*, 2(2):57–74, 2016.

[272] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The Information Society*, 37(1): 35–45, January 2021. ISSN 0197-2243. doi: 10.1080/01972243.2020.1832636. URL https://doi.org/10.1080/01972243.2020.1832636. Publisher: Routledge _eprint: https://doi.org/10.1080/01972243.2020.1832636.

[273] David Milne and Ian H Witten. An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

[274] Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. DECAF: Deep Extreme Classification with Label Features. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 49–57, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8297-7. doi: 10.1145/3437963.3441807.

[275] Bamshad Mobasher, Robin Burke, Runa Bhaumik, and Chad Williams. Effective attack models for shilling item-based collaborative filtering systems. In *Proceedings of the WebKDD Workshop*, pages 13–23. Citeseer, 2005.

[276] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. Tea and sympathy: Crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 839–848. ACM, 2013.

[277] Daniel Moyer, Samuel L Carson, Thayne Keegan Dye, Richard T Carson, and David Goldbaum. Determining the influence of Reddit posts on Wikipedia pageviews. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 2015.

[278] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. Survey research in HCI. In *Ways of Knowing in HCI*, pages 229–266. Springer, 2014.

[279] Yaroslav Nechaev, Francesco Corcoglioniti, and Claudio Giuliano. Concealing Interests of Passive Users in Social Media. In *BlackMirror@ ISWC*, 2017.

[280] Nellie Bowles. Early Facebook and Google Employees Form Coalition to Fight What They Built - The New York Times. https://www.nytimes.com/2018/02/04/technology/early-facebook-google-employees-fight-tech.html, 2018.

[281] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. User migration in online social networks: A case study on reddit during a period of community unrest. In *ICWSM*, pages 279–288, 2016.

[282] Benjamin J Newman and Brandon L Bartels. Politics at the checkout line: Explaining political consumerism in the United States. *Political Research Quarterly*, 64(4): 803–817, 2011.

[283] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism.* nyu Press, 2018.

[284] Chelsea Novak. Facebook Container Extension: Take control of how you're being tracked — The Firefox Frontier. https://blog.mozilla.org/firefox/facebook-container-extension/.

[285] Daniel Oberhaus. Nearly All of Wikipedia Is Written By Just 1 Percent of Its Editors - Motherboard. 2017.

[286] Pamela Oliver, Gerald Marwell, and Ruy Teixeira. A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American journal of Sociology*, 91(3):522–556, 1985. Publisher: University of Chicago Press.

[287] Pamela E Oliver and Gerald Marwell. Mobilizing technologies for collective action. *Frontiers in social movement theory*, pages 251–72, 1992.

[288] Alexandra Olteanu, Onur Varol, and Emre Kiciman. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *CSCW*, pages 370–386, 2017.

[289] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[290] Edward Ongweso, Jr. Andrew Yang's Data Dividend Isn't Radical, It's Useless, June 2020.

[291] Will Oremus. Big Tobacco. Big Pharma. Big Tech? *Slate*, November 2017. ISSN 1091-2339. URL `https://slate.com/technology/2017/11/how-silicon-valley-became-big-tech.html`. Section: Technology.

[292] Myle Ott, Claire Cardie, and Jeff Hancock. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, pages 201–210, 2012.

[293] Tim Paek, Michael Gamon, Scott Counts, David Maxwell Chickering, and Aman Dhesi. Predicting the Importance of Newsfeed Posts and Social Network Friends. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1419–1424, Atlanta, Georgia, 2010. AAAI Press.

[294] David Page. How to Train Your ResNet, September 2018.

[295] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication*, 12(3):801–823, 2007.

[296] Alexandra Papoutsaki, James Laskey, and Jeff Huang. Searchgazer: Webcam eye tracking for remote studies of web search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 17–26, 2017.

[297] European Parliament. Art. 20 GDPR – Right to data portability. https://gdpr-info.eu/art-20-gdpr, May 2020.

[298] Ben Parr. Google Gives $2 Million to Wikipedia's Foundation. https://mashable.com/2010/02/16/google-wikipedia-donation, 2010.

[299] Kari Paul. Prime Day: activists protest against Amazon in cities across US. *the Guardian*, Apr 2020.

[300] Business Paul R. La Monica. Tech's magnificent seven are worth $7.7 trillion, Oct 2020. URL `https://www.cnn.com/2020/08/20/investing/faang-microsoft-tesla/index.html`. [Online; accessed 6. Oct. 2020].

[301] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[302] Tony Peng. The Staggering Cost of Training SOTA AI Models. *Synced*, June 2019.

[303] Claudia Perlich, Foster Provost, and Jeffrey Simonoff. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, 2003.

[304] Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113, 2006.

[305] Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, and George Loukas. A taxonomy and survey of attacks against machine learning. *Computer Science Review*, 34:100199, November 2019. ISSN 1574-0137. doi: 10.1016/j.cosrev.2019.100199. Publisher: Elsevier.

[306] Luca Ponzanelli, Andrea Mocci, Alberto Bacchelli, Michele Lanza, and David Fullerton. Improving low quality stack overflow post detection. In *Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference On*, pages 541–544. IEEE, 2014.

[307] Eduardo Porter. Your Data Is Crucial to a Robotic Age. Shouldn't You Be Paid for It? *New York Times*, March 2018.

[308] Laura Portwood-Stacer. Media refusal and conspicuous non-consumption: The performative and political dimensions of facebook abstention. *New Media & Society*, 15(7):1041–1057, 2013.

[309] Eric A Posner and E Glen Weyl. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press, 2018.

[310] Barbara Prainsack. Data donation: How to resist the iLeviathan. In *The ethics of medical data donation*, pages 9–22. Springer, Cham, 2019.

[311] Prolific. Prolific. https://prolific.ac/, 2018.

[312] Kristin Purcell, Joanna Brenner, and Lee Rainie. Search engine use 2012. *Pew Research*, 2012.

[313] Alexander J Quinn and Benjamin B Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412. ACM, 2011.

[314] Filip Radlinksi and Thorsten Joachims. Query Chains: Learning to Rank from Implicit Feedback. In *KDD '05: 11th ACM Conference on Knowledge Discovery and Data Mining*, Chicago, IL, 2005.

[315] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022.

[316] Joseph Reagle and Lauren Rhue. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5:21, 2011.

[317] Reddit. Api documentation. https://www.reddit.com/dev/api/.

[318] Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. A Taxonomy of Knowledge Gaps for Wikimedia Projects (First Draft). *arXiv preprint arXiv:2008.12314*, 2020.

[319] Yuqing Ren and Robert E Kraut. Agent based modeling to inform the design of multiuser systems. In *Ways of Knowing in HCI*, pages 395–419. Springer, 2014.

[320] Steffen Rendle. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57, 2012.

[321] Steffen Rendle, Li Zhang, and Yehuda Koren. On the Difficulty of Evaluating Baselines: A Study on Recommender Systems. *arXiv preprint arXiv:1905.01395*, 2019.

[322] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[323] Luke Richards. Why Wikipedia is still visible across Google's SERPs in 2018 - Search Engine Watch, November 2018.

[324] Ronald E Robertson, David Lazer, and Christo Wilson. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 955–965. International World Wide Web Conferences Steering Committee, 2018.

[325] Kevin Roose. We Need to Talk About How Good A.I. Is Getting. *The New York Times*, August 2022. ISSN 0362-4331. URL `https://www.nytimes.com/2022/08/24/technology/ai-technology-progress.html`.

[326] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

[327] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[328] Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A Constructive Prediction of the Generalization Error Across Scales. *arXiv:1909.12673 [cs, stat]*, December 2019.

[329] Gus Rossi and Charlotte Slaiman. Interoperability = Privacy + Competition. *Public Knowledge*, October 2019.

[330] Annabel Rothshild, Emma Lurie, and Eni Mustafaraj. How the Interplay of Google and Wikipedia Affects Perceptions of Online News Sources. In *Computation+ Journalism Symposium*, 2019.

[331] David Rotman. Technology and inequality. *TECHNOLOGY REVIEW*, 117(6): 52–60, 2014.

[332] David Rotman. How to solve AI's inequality problem, April 2022. URL `https://www.technologyreview.com/2022/04/19/1049378/ai-inequality-problem/`.

[333] Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

[334] Jathan Sadowski. When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 6(1):2053951718820549, January 2019. ISSN 2053-9517. doi: 10.1177/2053951718820549. URL `https://doi.org/10.1177/2053951718820549`. Publisher: SAGE Publications Ltd.

[335] Alan Said and Alejandro Bellogín. Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 129–136, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. doi: 10.1145/2645710.2645746.

[336] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1621–1630, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702508.

[337] Sam Levin. Google sees major claims of harassment and discrimination as lawsuits proceed — Technology — The Guardian. https://www.theguardian.com/technology/2018/mar/28/google-sexual-harassment-pay-gap-lawsuits-proceed, 2018.

[338] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: 10.1145/371920.372071.

[339] Adam Satariano. What the G.D.P.R., Europe's Tough New Data Law, Means for You. *N.Y. Times*, May 2020. ISSN 0362-4331. URL `https://www.nytimes.com/2018/05/06/technology/gdpr-european-privacy-law.html`. Publisher: The New York Times Company.

[340] Christine Satchell and Paul Dourish. Beyond the user: Use and non-use in HCI. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, OZCHI '09, pages 9–16, New York, NY, USA, November 2009. Association for Computing Machinery. ISBN 978-1-60558-854-4. doi: 10.1145/1738826.1738829.

[341] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 813–822, 2016.

[342] Devansh Saxena, Erhardt Graeff, Shion Guha, EunJeong Cheon, Pedro Reynolds-Cuéllar, Dawn Walker, Christoph Becker, and Kenneth R. Fleischmann. Collective organizing and social responsibility at CSCW. In *Conference companion publication of the 2020 on computer supported cooperative work and social computing*, CSCW '20 companion, pages 503–509, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-8059-1. doi: 10.1145/3406865.3418593. URL `https://doi.org/10.1145/3406865.3418593`. Number of pages: 7 Place: Virtual Event, USA.

[343] Devansh Saxena, Patrick Skeba, Shion Guha, and Eric PS Baumer. Methods for generating typologies of non/use. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.

[344] Ronald Schmidt. SerpScrap. https://github.com/ecoron/SerpScrap, 2018.

[345] Sarita Yardi Schoenebeck. Giving up Twitter for Lent: how and why we take breaks from social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 773–782, 2014.

[346] Nico Schoonderwoerd. 4 ways how Twitter can keep growing. *Peer Research Blog*, 2013.

[347] Klaus Schwab, Alan Marcus, JO Oyola, William Hoffman, and M Luzi. Personal data: The emergence of a new asset class. In *An Initiative of the World Economic Forum*, 2011.

[348] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12):54–63, November 2020. ISSN 0001-0782. doi: 10.1145/3381831. URL `https://doi.org/10.1145/3381831`.

[349] Lisa Seitz-Gruwell. Google and Wikimedia Foundation partner to increase knowledge equity online, January 2019.

[350] Neil Selwyn. Apart from technology: understanding people's non-use of information and communication technologies in everyday life. *Technology in society*, 25(1):99–116, 2003.

[351] Alana Semuels. Why #DeleteUber and Other Boycotts Matter . *Atlantic*, Feb 2017. URL `https://www.theatlantic.com/business/archive/2017/02/why-deleteuber-and-other-boycotts-matter/517416`.

[352] Rijurekha Sen, Sohaib Ahmad, Amreesh Phokeer, Zaid Ahmed Farooq, Ihsan Ayyub Qazi, David Choffnes, and Krishna P Gummadi. Inside the walled garden: Deconstructing facebook's free basics program. *ACM SIGCOMM Computer Communication Review*, 47(5):12–24, 2017.

[353] Hamza Shaban. Google parent Alphabet reports soaring ad revenue, despite YouTube backlash. *Washington Post*, February 2018. ISSN 0190-8286.

[354] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018.

[355] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1589–1604, 2020.

[356] Carl Shapiro, Hal R Varian, and WE Becker. Information rules: A strategic guide to the network economy. *Journal of Economic Education*, 30:189–190, 1999.

[357] Robert Shapiro and Siddhartha Anejo. Who Owns Americans' Personal Information and What Is It Worth? Technical report, Future Majority, 2018.

[358] Amit Sharma, Jake M. Hofman, and Duncan J. Watts. Estimating the Causal Impact of Recommendation Systems from Observational Data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, pages 453–470, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3410-5. doi: 10.1145/2764468.2764488.

[359] Aaron Shaw, Haoqi Zhang, Andrés Monroy-Hernández, Sean Munson, Benjamin Mako Hill, Elizabeth Gerber, Peter Kinnaird, and Patrick Minder. Computer supported collective action. *interactions*, 21(2):74–77, 2014. Publisher: ACM New York, NY, USA.

[360] Jonathan Shieber. Google backtracks on search results design. *TechCrunch*, January 2020.

[361] Nate Shivar. How To Use Wikipedia for SEO & Content Marketing Strategy. https://www.shivarweb.com/3632/how-to-use-wikipedia-for-seo/, 2017.

[362] Ben Shneiderman. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4):26:1–26:31, October 2020. ISSN 2160-6455. doi: 10.1145/3419764.

[363] Márcio Silva, Lucas Santos de Oliveira, Athanasios Andreou, Pedro Olmo Vaz de Melo, Oana Goga, and Fabrício Benevenuto. Facebook ads monitor: An independent auditing system for political ads on facebook. In *Proceedings of The Web Conference 2020*, pages 224–234, 2020.

[364] Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5607–5614, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/782. URL `https://www.ijcai.org/proceedings/2022/782`.

[365] Bob Simison. How Sending Stimulus Checks to the Poor Can Boost the US Economy. https://www.chicagobooth.edu/review/how-sending-stimulus-checks-poor-can-boost-us-economy, 2022.

[366] Amit Singhal. Introducing the knowledge graph: Things, not strings. *Official google blog*, 16, 2012.

[367] Nicola Slawson. Faceblock campaign urges users to boycott Facebook for a day. https://www.theguardian.com/technology/2018/apr/07/faceblock-campaign-urges-users-boycott-facebook-for-one-day-protest-cambridge-analytica-scandal, 2018.

[368] Tom De Smedt and Walter Daelemans. Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067, 2012.

[369] Adam N. Smith, Stephan Seiler, and Ishant Aggarwal. Optimal Price Targeting, 2021. URL `https://papers.ssrn.com/abstract=3975957`.

[370] Brent Smith and Greg Linden. Two decades of recommender systems at Amazon. com. *IEEE Internet Computing*, 21(3):12–18, 2017.

[371] Luca Soldaini, Andrew Yates, Elad Yom-Tov, Ophir Frieder, and Nazli Goharian. Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal*, 19(1-2):149–173, 2016.

[372] Yang Song, Xiaolin Shi, and Xin Fu. Evaluating and Predicting User Engagement Change with Degraded Search Relevance. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1213–1224, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488494.

[373] Tim Soulo. Top Google searches (as of October 2019). https://ahrefs.com/blog/top-google-searches, October 2019.

[374] StackExchange. Book a campaign — Stack Exchange Self-Serve. https://www.selfserve-stackexchange.com/, 2018.

[375] StackOverflow. Ad Banners - Developer Advertising Solutions — Advertise on Stack Overflow. https://www.stackoverflowbusiness.com/advertise/solutions/ad-banners, 2018.

[376] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Advances in neural information processing systems*, pages 3517–3529, 2017.

[377] Stephen Spinelli Jr. 1 million jobs will disappear by 2026. How to prepare for automation-commentary. https://www.cnbc.com/2018/02/02/automation-will-kill-1-million-jobs-by-2026-what-we-need-to-do-commentary.html, 2018.

[378] Greg Sterling. It's Official: Google Says More Searches Now On Mobile Than On Desktop - Search Engine Land. `https://searchengineland.com/its-official-google-says-more-searches-now-on-mobile-than-on-desktop-220369`, May 2015.

[379] Tony Sterling and Alexandra Hudson. Facebook users unite! 'Data Labour Union' launches in Netherlands — Reuters. https://www.reuters.com/article/us-netherlands-tech-data-labour-union/facebook-users-unite-data-labour-union-launches-in-netherlands-idUSKCN1IO2M3, 2018.

[380] Stefan Stieger, Christoph Burger, Manuel Bohn, and Martin Voracek. Who commits virtual identity suicide? Differences in privacy concerns, internet addiction, and personality between Facebook users and quitters. *Cyberpsychology, Behavior, and Social Networking*, 16(9):629–634, 2013. Publisher: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.

[381] Joseph E Stiglitz. *The Price of Inequality: How Today's Divided Society Endangers Our Future*. WW Norton & Company, 2012.

[382] Greg Stoddard. Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News. In *ICWSM*, pages 416–425, 2015.

[383] Farnaz Tahmasebian, Li Xiong, Mani Sotoodeh, and Vaidy Sunderam. Crowdsourcing under data poisoning attacks: A comparative study. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 310–332. Springer, 2020.

[384] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*, 2021.

[385] Dario Taraborelli. The Sum of All Human Knowledge in the Age of Machines: A New Research Agenda for Wikimedia., 2015.

[386] Yla R. Tausczik and James W. Pennebaker. Predicting the Perceived Quality of Online Mathematics Contributions from Users' Reputations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1885–1888, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979215.

[387] Jaime Teevan, Amy Karlson, Shahriyar Amini, AJ Brush, and John Krumm. Understanding the importance of location, time, and people in mobile local search behavior. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 77–80. ACM, 2011.

[388] Jacob Thebault-Spieker, Brent Hecht, and Loren Terveen. Geographic Biases are'Born, not Made': Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pages 71–82. ACM, 2018.

[389] Financial Times. *Big Tech Must Pay for Access to America's 'digital Oil'*. Financial Times, April 2019.

[390] Catalina L Toma and Jeffrey T Hancock. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 5–8, 2010.

[391] Townhall.com. Election 2016 Results Map and Key Races for the President elections - View the latest election results, news, polls and conservative election commentary.

https://townhall.com/election/2016/president/, 2017.

[392] Tess Townsend. Google's share of the search ad market is expected to grow - Recode. https://www.recode.net/2017/3/14/14890122/google-search-ad-market-share-growth, 2017.

[393] Carmela Troncoso. Keynote Address: PETs, POTs, and Pitfalls: Rethinking the Protection of Users against Machine Learning. Santa Clara, CA, August 2019. USENIX Association.

[394] Hayley Tsukayama. Why Getting Paid for Your Data Is a Bad Deal. https://www.eff.org/deeplinks/2020/10/why-getting-paid-your-data-bad-deal, October 2020.

[395] Zeynep Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM*, 14:505–514, 2014.

[396] Jazmine Ulloa. Newsom wants companies collecting personal data to share the wealth with Californians. *latimes.com*, May 2019.

[397] U.S. Census Bureau. 2011-2015 American Community Survey 5-Year Estimates. Technical report, U.S. Census Bureau, 2011.

[398] Alexander JAM Van Deursen and Jan AGM Van Dijk. Using the Internet: Skill related problems in users' online behavior. *Interacting with computers*, 21(5-6): 393–402, 2009.

[399] Max Van Kleek, Dave Murray-Rust, Amy Guy, Kieron O'Hara, and Nigel Shadbolt. Computationally Mediated Pro-Social Deception. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 552–563, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858060.

[400] Chris Van Pelt and Alex Sorokin. Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 765–766. ACM, 2012.

[401] Maddy Varner and Sam Morris. Introducing Simple Search – The Markup. *The Markup*, January 2021.

[402] Graham Vickery and Sacha Wunsch-Vincent. *Participative Web and User-Created Content: Web 2.0 Wikis and Social Networking.* Organization for Economic Cooperation and Development (OECD), 2007.

[403] Salomé Viljoen. Data as Property?, October 2020. URL `https://www.phenomenalworld.org/analysis/data-as-property/`.

[404] Nicholas Vincent and Brent Hecht. A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):4:1–4:15, April 2021. doi: 10.1145/3449078. URL `https://doi.org/10.1145/3449078`.

[405] Nicholas Vincent and Brent Hecht. Can "Conscious Data Contribution" help users to exert "Data Leverage" against technology companies? *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021. doi: 10.1145/3449177. URL `https://doi.org/`

10.1145/3449177. Number of pages: 23 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 103 tex.issue_date: April 2021.

[406] Nicholas Vincent, Isaac Johnson, and Brent Hecht. Examining wikipedia with a broader lens: Quantifying the value of wikipedia's relationships with other large-scale online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[407] Nicholas Vincent, Brent Hecht, and Shilad Sen. “Data Strikes”: Evaluating the Effectiveness of New Forms of Collective Action Against Technology Platforms. In *Proceedings of The Web Conference 2019*, 2019.

[408] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. Measuring the Importance of User-Generated Content to Search Engines. *Proceedings of the International AAAI Conference on Web and Social Media*, 13:505–516, July 2019. ISSN 2334-0770. URL https://ojs.aaai.org/index.php/ICWSM/article/view/3248.

[409] Nicholas Vincent, Yichun Li, Renee Zha, and Brent Hecht. Mapping the Potential and Pitfalls of ”Data Dividends” as a Means of Sharing the Profits of Artificial Intelligence. *arXiv preprint arXiv:1912.00757*, 2019.

[410] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 215–227, 2021.

[411] Kaveh Waddell. California's New Privacy Rights Are Tough to Use, Consumer Reports Study Finds. *Consum. Rep.*, Oct 2020. URL https://www.consumerreports.org/privacy/californias-new-privacy-rights-are-tough-to-use.

[412] Tarun Wadhwa. Economic Impact and Feasibility of Data Dividends. page 12, 2020.

[413] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *ICWSM*, pages 454–463, 2015.

[414] Kurt Wagner. Reddit worth $1.8 billion, 2017. URL https://www.cnbc.com/2017/07/31/reddit-worth-1-point-8-billion.html.

[415] Daisuke Wakabayashi. California Passes Sweeping Law to Protect Online Privacy. *N.Y. Times*, Jun 2018. ISSN 0362-4331. URL https://www.nytimes.com/2018/06/28/technology/california-online-privacy-law.html.

[416] Tianhao Wang and Ruoxi Jia. Data Banzhaf: A Data Valuation Framework with Maximal Robustness to Learning Stochasticity, July 2022.

[417] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset Distillation, February 2020. URL http://arxiv.org/abs/1811.10959. Number: arXiv:1811.10959 arXiv:1811.10959 [cs, stat].

[418] Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 743–756. ACM, 2015.

[419] Morten Warncke-Wang, Vivek Ranjan, Loren G Terveen, and Brent Hecht. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. In *ICWSM*, pages 493–502, 2015.

[420] Hongyi Wen, Longqi Yang, Michael Sobolev, and Deborah Estrin. Exploring recommendations under user-controlled data filtering. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 72–76, 2018.

[421] Robert West, Ingmar Weber, and Carlos Castillo. Drawing a data-driven portrait of Wikipedia editors. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 3. ACM, 2012.

[422] Ryen W White, Fernando Diaz, and Qi Guo. Search result prefetching on desktop and mobile. *ACM Transactions on Information Systems (TOIS)*, 35(3):1–34, 2017.

[423] John Wilmhoff. Tom Brady literally owns the Jets, says Google search, July 2017. URL `https://www.espn.com/sportsnation/story/_/page/170727QTP_BradyOwnsJets/google-glitch-causes-tom-brady-appear-new-york-jets-owner`. Publication Title: ESPN.

[424] David C Wilson and Carlos E Seminario. When power users attack: assessing impacts in collaborative recommender systems. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 427–430, 2013.

[425] David C Wilson and Carlos E Seminario. Evil twins: Modeling power users in attacks on recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 231–242. Springer, 2014.

[426] Laurence Wong. Causal Inference in Python — Causalinference 0.1.2 documentation. http://causalinferenceinpython.org/, 2018.

[427] wordstream.com. The Top 20 Most Expensive Keywords in Google AdWords Advertising. https://www.wordstream.com/articles/most-expensive-keywords, 2011.

[428] Sally ME Wyatt. Non-users also matter: The construction of users and non-users of the internet. *Now users matter: The co-construction of users and technology*, pages 67–79, 2003.

[429] Zheng Xiang and Ulrike Gretzel. Role of social media in online travel information search. *Tourism management*, 31(2):179–188, 2010.

[430] Xinyu Xing, Wei Meng, Dan Doozan, Nick Feamster, Wenke Lee, and Alex C. Snoeren. Exposing Inconsistent Web Search Results with Bobble. In *Passive and Active Measurement*, Lecture Notes in Computer Science, pages 131–140. Springer, Cham, March 2014. ISBN 978-3-319-04917-5 978-3-319-04918-2. doi: 10.1007/978-3-319-04918-2_13.

[431] Sean Xin Xu and Xiaoquan (Michael) Zhang. Impact of Wikipedia on Market Information Environment: Evidence on Management Disclosure and Investor Reaction. *MIS Quarterly*, 37(4):1043–1068, Dec 2013. ISSN 0276-7783. URL `http://www.jstor.org/stable/43825781`.

[432] Tom Yan and Ariel D. Procaccia. If You Like Shapley Then You'll Love the Core. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5751–5759, May 2021. ISSN 2374-3468. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16721`. Number: 6.

[433] Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. WikiWalk: Random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49. Association for Computational Linguistics, 2009.

[434] Dario Zadro. Why Should UGC be a Part of Your SEO Strategy? https://www.searchenginejournal.com/ugc-part-seo-strategy/93557/, 2014.

[435] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 365–378. Association for Computing Machinery, New York, NY, USA, October 2020. ISBN 978-1-4503-7514-6.

[436] Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. The AI Index 2022 Annual Report. Technical report, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, May 2022. URL `http://arxiv.org/abs/2205.03468`. arXiv:2205.03468 [cs] type: article.

[437] Haoqi Zhang, Andrés Monroy-Hernández, Aaron Shaw, Sean A Munson, Elizabeth Gerber, Benjamin Mako Hill, Peter Kinnaird, Shelly D Farnham, and Patrick Minder. WeDo: End-to-end computer supported collective action. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[438] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pages 404–410. ACM, 2010.

[439] Haiyi Zhu, Robert E Kraut, Yi-Chia Wang, and Aniket Kittur. Identifying shared leadership in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3431–3434. ACM, 2011.

[440] Haiyi Zhu, Robert Kraut, and Aniket Kittur. Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 407–416. ACM, 2012.

[441] Haiyi Zhu, Amy Zhang, Jiping He, Robert E Kraut, and Aniket Kittur. Effects of peer feedback on contribution: A field experiment in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2253–2262. ACM, 2013.

[442] Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. The Impact of Membership Overlap on the Survival of Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 281–290, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557213.