Votes, Genes, Grades: Deconstructing Big Data's Rhetorical Reach

By

Kathleen Daly Weisse

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(English)

at the

UNIVERSITY OF WISCONSIN-MADISON

2019

Date of final oral examination: July 16, 2019

This dissertation is approved by the following members of the Final Oral Committee:
Christa Olson (chair), Associate Professor, English
Michael Bernard-Donals, Vice Provost, Full Professor, English
Kate Vieira, Associate Professor, Curriculum and Instruction
Joan Fujimura, Full Professor, Sociology

Table of Contents

Abstract

This dissertation interrogates the rhetorical implications of widespread applications of Big Data—a research method that deploys large-scale analytics to produce knowledge and predict or determine outcomes from massive, complex data sets. I employ rhetorical analysis to examples of Big Data as applied in three different sectors within the modern American landscape: politics, academia, and biotechnology. In tracing these examples, which range from the 2016 United States presidential campaign and election to learning analytic software in higher education to direct-to-consumer genetic testing services, I argue that the cultural cachet of Big Data research has empowered data analytic companies to sell their services as resources for revealing otherwise inaccessible truths about people and populations, despite obvious gaps in methodological regulation, demographic representation, and data quality. This dissertation shows that Big Data is rhetorically successful, because it removes sites of accountability while at the same time claiming to be a means of establishing accountability. Like many technological innovations that have come before it, Big Data often perpetuates existing social inequalities and reinforces partisanship; however, because the term "Big Data" increasingly connotes scientific credibility, arguments made using large-scale data analytics can, intentionally or unintentionally, inscribe in hegemonic structures a sense of objectivity. By exposing the power structures undergirding and supported by commonplace applications of Big Data technologies, this dissertation creates a rhetorical framework for approaching Big Data research ethically by emphasizing accountability and transparency.

Acknowledgements

This dissertation would not have been possible without the help of my mentors, colleagues, friends, and family who supported me along the way. Thank you to my dissertation committee for your keen insights and feedback on my drafts. I especially want to thank my chair, Christa Olson, for helping me figure out my own path through this program, for pushing me to work hard while also stressing the importance of taking care of myself, and for always making me think and laugh so hard. Thanks also for gently pulling me out of the clouds but for always knowing what I was getting at when I was up there.

Thank you to my mentor, Brad Hughes. Working with you as the Assistant Director of Writing Across the Curriculum was one of the most intense and rewarding jobs I've ever had. During our weekly meetings, you pushed me to articulate my ideas, to think critically about my decisions, and to take agency over the WAC program in whatever way I thought most valuable. Your dedication to teaching with writing is contagious, and with your guidance, I began to think of myself as an expert in something for the first time in my life.

I want to thank my colleagues at UW-Madison: Annika Konrad, Angela Zito, Tori Peters, Neil Simpkins, Meg Marquardt, Brandee Easter, Brenna Swift, Calley Marotta, Elisabeth Miller, Chris Earle, Leigh Elion, Kim Moreland, Gin Schwartz, and many others. And I would be remiss not to thank my colleagues from my time at Illinois State University: Julie Jung, Kellie Sharp-Hoskins, Erin Frost, Chris Mays, Karly Grice, and Marie Moeller. To my best friend in Madison, Rebecca Galvan, thanks for letting me witness you write your dissertation and paying me to edit it. Thanks for getting ordained and officiating my wedding. Thank you to my in-laws, Kathy and David Weisse, for cheering me on like I was your own.

Thank you to my mom, Barbara Daly, for saving all the things I wrote when I was a kid. She always asking to read my stuff, my syllabi, papers, after all this time. You even once completed an assignment I created! I want to thank you for caring so much and for being so interested, for always listening, for sometimes worrying, for letting me be independent while always being right there to hug me and comfort me when it was all just too hard to handle. I love you, Mom.

Thank you to my dad, David Daly, for taking me to the library and to all the bookstores when I was a kid, for filling our home with stacks and stacks of books, for teaching me that you should always carry a book around in case whatever you're doing gets kinda boring. This whole time I know you wanted me to be a pilot but at least I'm really into books. You have the most amazing hunger for the written word, and I'm so grateful that you passed it down to me. And I'm so glad I passed my love of cats up to you! I love you, Dad.

Thank you to my brother, Jack Daly. Right after I moved to Madison, you came to visit and we both explored Madison for the first time. We went to the Willy St. Festival, the Farmer's Market, the Contemporary Art Museum, and the Terrace—we experienced so many Madison firsts together. Those early days of exploring with you were so exciting and light and fun, and the memories I have of them are so strong, that even though you don't live here, when I think of Madison, I think of you. How cool!

Thank you to my forever friend, Katy Daly, for pulling all-nighters in undergrad when I wrote my first rhetoric paper (I think I got a C!). Thanks for introducing me to a lot of the music that's been the soundtrack to my academic life. The bands that have gotten me through grad school. I'm so glad to have you as a friend and now a sister-in-law.

Thank you, Rachelle Yates, for always having my back, for supporting me from near and far, for never judging me and for always believing in me. Thanks for always reminding me that life is better with sushi and pets.

Finally, I want to thank Travis Weisse. I met you the first day of the very last class I ever took and by the time the semester was over, we were living together. Six months later, you helped me through the most difficult loss of my life, and it was then that you really became a part of my family. Later that year, when I was working on my prelims essay and told you I was going to quit because everything was the worst and so terrible, you hugged me, let me cry it out, and said, "Okay…or you could finish it." And then you sat down with me and we finished it. And somewhere in there you let me open all of my Christmas presents up early, which helped a lot. You contributed so much emotional labor to this project, and I'm so glad that I found you in my last class. You're my number one reader, my safety net, my trivia teammate, my sous chef, my forever partner, and my best friend.

Chapter One

Introduction

In their *Wall Street Journal* article published in 2018, Stephanie Stamm, Tripp Mickle and Jessica Kuronen tell a story about two fictional friends, Sally and Kristen, who got together one night for pizza and a movie (Stamm et al.). The narrative, which takes place over the course of 4 hours, thoroughly excavates all of the many personalized digital traces the pair generates up to and through their evening together. The authors count no fewer than 53 digital traces that the women left behind through their simple encounter. When Sally used her Amazon Echo to ask Alexa to open her Domino's app and order a pizza, for instance, Amazon pulled payment information data from her previous transaction history, gathered data about her location, device type, and vocal characteristics, and logged the content of her request into its system. The Domino's app also gathered data from the interaction, including information about the content of her order (e.g. the type and quantity of pizza ordered), a transcript of her order (created and supplied by Alexa), the last four digits of her credit card number, and various device and settings data. When Kristen ordered the movie through her Apple TV, Apple gathered data about her Apple ID, payment information, Internet bandwidth information and purchase history. And when the two took a selfie together using Sally's phone and Sally then uploaded it to Facebook, Facebook's facial recognition software collected visual and location data from the photo, and the Facebook App gathered data from her phone, including the device type and operating system, battery level, Bluetooth signal status, app and file names and types, IP address, and more. The article was written to suggest that companies are harvesting more data from consumers than they know, and underscores the length and complexity of digital privacy policies, emphasizing that it

is unreasonable for the average consumer to become expert enough to decipher any, let alone all, of these agreements.

This dissertation picks up the narrative thread from stories like Sally and Kristen's, but instead of detailing all of the different ways in which data is (often unwittingly and/or non-consensually) generated and archived, I trace the consequences of the conclusions that are drawn, via Big Data and algorithmic analysis, about the people whose data is harvested.

To illustrate this point, let's continue the metaphor from the above article. The day following Sally and Kristen's movie night, Sally gets a push notification from her personalized news app with a link to a review of the movie they watched last night. When she sees the notification, Sally is surprised. It was Kristen, after all, who had ordered the movie and it was on Kristen's TV that they watched it together. She considers the source of the notification (a third-party news app), before shrugging it off as coincidence. She goes to clear the notification, but then, noticing the headline, decides it is worth skimming after all. The article is from a progressive online news source and offers a feminist critique of the movie. Sally nods her head in agreement as she reads; the argument of the piece mapped clearly onto her pre-existing sentiments about the movie's themes. This single keystroke then leads Sally's news app to provide even more content like the article she clicked on, which creates a feedback loop that continues until her newsfeed has been completely saturated with tailored content. Kristen also receives a push notification related to the movie from a third-party news app. She notices the lead actress's name in the headline and clicks on the article, which takes her to a listicle of the actress's 'best-dressed moments.' After scrolling through a few photos before she sees a sponsored advertisement about how to lower her student debt. Having debt herself, she clicks on the ad and is routed to a debt-management company's website.

In both Sally and Kristen's digital experiences, there are a few clear moments where we can see that digital content mediators are using data generated and collected from the previous evening's activities to inform and structure their future digital encounters. We have all probably experienced similar moments of clarity while browsing online: you purchase a new phone charging cord on Amazon and then when you log onto Facebook, you see an advertisement for the exact same phone charging cord you just purchased. These moments reveal something significant, something that was left out of the original WSJ article: companies are using the data they collect to draw assumptions and make decisions about who people are, what they like, how they have behaved in the past, and how they are likely to behave in the future. More specifically, companies use algorithms that sift through aggregated data and draw meaning from the information available—algorithms, as Chris Ingraham explains, "make rhetorical choices" (62). These choices, which may or may not map on to the actual preferences of the person, will, nevertheless, have implications for their future choices.

Additionally, and importantly, these moments reveal that data environments are porous—data that is generated and collected by one company, along with all of the things that that company learns about you, are not bound to that company, but rather bought and sold freely between companies. As data moves across the marketplace, it creates a disjointed cloud of digital traces that follows you around, and which any potential data buyer can assemble into a sophisticated (even psychographic) profile. Though nothing in the original story from the *Wall Street Journal* or in my narrative follow up directly suggested that either woman was in debt or held feminist values per se, Big Data algorithms triangulate from all available data and make assumptions based on regional or national patterns to fill in gaps. This profile, depending on who

or what is behind the algorithm, can fundamentally change how you encounter objects in the digital ecosystem, and, perhaps more importantly, how digital environments treat you.

Even providing third party apps or social media platforms with information as simple as location data, age, or recent purchases can trigger changes in how the algorithm interprets your data trails, and thus what content it makes available to you. Beyond the level of individual users, these clouds of digital traces left by every user in every location and built up over years of digital traffic, may have other, more systematic effects. By associating individuals (and their full data profiles) with other individuals, locations, groups, and activities, algorithmic judgments about individual users can bleed into judgments about every other user, location, group, and activity that that user encounters as well. Importantly, the personalized user profiles that algorithms construct are necessarily fragmented and imperfect. Though it may seem as though algorithms regularly predict user behavior with spooky accuracy, the times they are poorly calibrated and target users with ads that are irrelevant are so common as to be mundane. Though targeted ads for online shopping are *generally* harmless, the Big Data technologies through which they are deployed are not.

Clarifying the Term "Big Data"

Despite the prominence of Big Data in popular and professional discourses, a precise meaning of the term is surprisingly difficult to pin down. For the computer scientists who first used the term in the late 1990s, it referred to data sets that were simply too large (too many bits of information) for extant data processing methods and software. Other computer scientists have since used the term to stand in as a measure for the sum total of all data on Earth, tracking the exponential increase of data production enabled by the Internet and its attendant needs for data

storage. In contemporary parlance, however, Big Data tends to imply very large (but no longer impossible to process) datasets and the methodologies that are required to digest them. This version of Big Data is typically described as having four major characteristics (the "Four V's"): volume (massive amounts of data), velocity (data demands immediate analysis), variety (data comes in different forms and types), and variability (data is almost always incomplete, incongruent, and/or inconsistent). While this alliterative list is helpful in illustrating how complex datasets can be in modern computing, it focuses too narrowly on the 'data' that makes up 'Big Data,' and does little to clarify the actual processes and procedures involved in doing Big Data analytics. Further, it sidelines discussions about how the networks through which meaning is made from Big Data processes are contingent upon the ways in which data are analyzed, interpreted, and translated, all of which change according to different contexts, mechanisms, stakeholders, and research agendas.

In this dissertation, therefore, I will follow the lead of information scientists danah boyd and Kate Crawford, who, in their article "Critical Questions for Big Data," explain that "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets" (663). Conceptualized as a system of agencies, activities, and actants, however, Big Data—with a capital 'B' and 'D'—is much more than data; it is a process of meaning making, which wields rhetorical power through its capacity to make meaning from incomprehensibly large and complex sets of data, and its potential to reveal otherwise inaccessible knowledge about people and populations. Under this definition, Big Data is not a consistent or even singular phenomenon; rather, it changes according to where and from whom datasets are being collected, who is advocating for which type of data analysis, and what the ultimate purpose is for these practices.

Common Categories of Big Data Analytics

In *The Data Revolution,* Rob Kitchin categorizes modern Big Data analytics into four distinct types: analyses aimed at description, explanation, prediction and prescription. The first type, descriptive analytics, involve the use of data mining and pattern recognition to describe past events and determine *what happened*. In the retail sector, a company might use descriptive analytics to assess past sales or identify rates of returns over a given period of time; in healthcare contexts, these same analytic strategies can be used, for instance, to assess the average amount of time patients checking into a hospital spend in the waiting room before they are seen by a medical professional.

The second type of analytics, explanatory analytics, typically center around questions of why something happened and, more-so than with descriptive analytics, are aimed at meaning-making tasks. Explanatory analytics often involve the use of visual representations of data, or data visualizations, to identify patterns across large sets of data. Typically deployed to assess *why something happened*, exploratory analytics can help companies identify why, for example, there was a recent spike (or drop) in sales of a particular product.

The third type of analytics, predictive analytics, relies on statistical analyses that use historical data (akin to data used in descriptive analytics) to forecast future trends or events and to predict what *will happen*. Predictive analytics are often based on predictive modeling practices wherein existing datasets are compared with data drawn from surveys or focus groups; the aggregated data can then be used to make predictions about how people might behave in the future. Despite being based in probabilities, predictive analytics' potential for helping

organizations cut costs and increase efficiency has motivated its widespread uptake, with applications arising in business, government, healthcare, and education sectors, among others.

Finally, the fourth type of Big Data analytics, prescriptive analytics are used to determine what *should be done*. Typically deployed alongside simulation and optimization technologies, prescriptive analytics use findings from predictive analysis to determine future goals and actions. Prescriptive analytics use descriptive analytic data to make suggestions about the best possible options moving forward (similarly to how Google Maps analyzes traffic data to make suggestions as to the fastest possible route between locations).

There is significant overlap between all four types of analytics and organizations often use a combination of Big Data analytics to achieve their particular goals. Further, the results from descriptive analytics and explanatory analytics often feed into or inform how organizations deploy predictive analytics and prescriptive analytics. And the meaning that is generated in one iteration of analytics can shape how analytics are structured in the future (results from one application can change what kinds of questions should be asked or what kinds of "meaning" should be most valued, etc.). So while it is useful to recognize each type of analytics as having discrete applications and methods, it is important that these four types of analytics are also understood inter-relationally.

Importantly, each of these analytics relies on data mining and algorithms. In general, algorithms are just automated sets of instructions to be run by a computer. In the context of Big Data, however, algorithms typically sift through large sets of data to find patterns, trends, and connections that fit within the parameters of 'meaning' set forth by the algorithm's creator. In order to find 'meaningful information,' however, Big Data algorithms do not simply look for patterns that have been pre-programmed by the researchers who designed them. Rather, their

purpose is to identify patterns across enormous swaths of data to reveal potential associations between seemingly disparate data points, variables, or data sets.

To be effective, all Big Data analytics require that researchers critically reflect on the quality of data being used, the methodological assumptions being made, or the potential social and/or political contexts and consequences of their application. When data is deployed to help reach some desired end goal (whether in politics, genetics, or education, as this dissertation will show), the problem gets framed in terms that are more amenable to computational analysis (though they may not resemble the thing they are trying to measure in reality), creating a feedback loop where inputs are increasingly tailored to the analytical standards preferred by the algorithm. Ideally, Big Data algorithms isolate possible associations between variables that researchers can subject to greater scrutiny later (to determine a causal link, for instance). However, the conclusions Big Data algorithms draw are often left to stand alone, a fact which, under certain conditions, can be problematic.

All four types of Big Data analytics enumerated above are primarily rhetorical in that they are each aimed at identifying patterns in data with which to make arguments. Additionally, as this dissertation shows, because Big Data (through various types of analytics) generates effects, induces action and attitudes, and informs the shaping of reality, it can be studied rhetorically.

The Promise of Big Data

For Big Data purveyors, it seems, the technology has near limitless capacities to understand the world. Scholars Elish and boyd draw attention to the ways in which Big Data technologies are "invoked as the solution to otherwise intractable social, political, and economic

problems, and seem to promise efficiency, neutrality, and fairness—ideals that are often viewed as impossible to achieve through individual human or organizational decision-making processes" (17–8). Companies offering large-scale data analytic services have often boasted about the contextual flexibility and widespread applicability of Big Data technologies to function alongside such different research initiatives and contexts as polling, disease tracking, social media analytics, marketing and sales analysis and more. The "hype and promise" Big Data technologies (which has largely emanated from the business-sector), has granted Big Data significant rhetorical power and contributed to exaggerated public perceptions of Big Data's capacities, wherein the potential benefits these technologies afford "extend far past [their] methodological capabilities" (Elish and boyd 2).

Despite being ultimately unfulfillable, the assurance that Big Data will continually uncover hidden things about the world via digital traces has enabled Big Data technologies to occupy a central place in what STS scholar Sheila Jasanoff has called the sociotechnical imaginary ("Containing the Atom"). Jasanoff defines sociotechnical imaginaries as "collectively held, institutionally stabilised, and publicly performed visions of desirable futures," applied to the formation, implementation, and effects of technological progress (4). Importantly, sociotechnical imaginaries are not bound to the realm of imagination, but rather, are actively called upon to inform and prescribe real future trajectories for technological development. In other words, the ways in which technologies are forecasted to develop in the future (including the promise of as-yet unattainable capabilities) directly informs the ways in which technologies are made (and sold) in the present.

Whereas Jasanoff is concerned with the cross-national comparison of nuclear policies, I will primarily use the concept of the sociotechnical imaginary as a framing device to emphasize

the rhetorical forces at play in the development of Big Data technologies. Within the context of this project, then, sociotechnical imaginaries help remind us that the drivers of Big Data technologies are themselves deeply enmeshed in socio-political environments (and their respective power structures). Accordingly, how Big Data technologies get developed and deployed is directly informed by the ideologies and values of the dominant imaginaries. Within the Big Data imaginary, shared notions of what Big Data technologies should achieve in the future foster an exaggerated public understanding of Big Data's efficacy as well as exaggerated perceptions of Big Data's social and political value. Far from benign, the dominant sociotechnical imaginaries surrounding Big Data draw attention toward the promise of Big Data in theory and away from the material consequences of Big Data in practice.

The Challenges of Interrogating Big Data

Algorithmic Opacity and the Persuasiveness of Blackboxed Machines

A central feature of Big Data's persuasiveness is the illusion of objectivity and accuracy constructed around its methodological reliance on non-human mechanisms (i.e. automated data mining and algorithmic computation) for selecting and organizing data. Corporate and civic rationales for employing Big Data regularly point to its efficacy for making sense of the world through passive, objective knowledge production systems. Assumptions that Big Data findings are produced solely via non-human apparatuses for passive data collection feed into a cultural narrative of Big Data as inherently objective and grounded in the Truth (Bijker; Barad; boyd and Crawford). What these uncritical frameworks of absoluteness fail to account for, however, is that the apparatuses through which Big Data is collected, analyzed, interpreted, and translated are contingent upon a plurality of subjective actors, agencies, and institutions. Unfortunately, as

Elish and boyd remind us, "once deployed for public use, recommendations, predictions, and classifications produced by technical systems are often accepted as uncontroversial until a result challenges socially constructed assumptions" (15).

Because methodological transparency and accountability are not typically prioritized in Big Data research—where algorithms are often proprietary and thus removed from public view—its findings have been made to stand in for reality with relative ease. Raw data sets are often only accessible once mediated through technologies designed to process and analyze large sets of data, such as data visualization tools. But even then, these processes are difficult, if not impossible, to trace and identify post-analysis. Big Data's opacity has enabled the de-emphasis of methodological accountability among data analysts, because, as Sociologist Nathan Jurgenson argues, "By moving the truth-telling ability from the researcher to data that supposedly speaks for itself, Big Data implicitly encourages researchers to ignore conceptual frameworks like intersectionality or debates about how social categories can be queered rather than reinforced."

Within our current knowledge economy, where Big Data technologies are removed from public oversight (because the algorithms behind Big Data are blackboxed and proprietary) and Big Data applications and outputs are presumed true until proven otherwise, there are few pathways leading into Big Data through which we can interrogate these systems. Yet, we must interrogate them, because, left unquestioned, Big Data denies possibilities for locating difference while empowering and expanding interlocking political and economic systems.

Recognizing the challenges that Big Data's opacity and the pervasiveness of its claims of objectivity present, while remaining committed to the need for critical examination of these systems, I propose a turn to a different kind of analytics: rhetorical analysis. In his explanation of the primary activities of digital rhetoric, Douglas Eyman defines rhetorical analysis as a method

for "uncovering and interrogating ideologies and cultural formation." Likewise, rhetorical analysis offers us a framework for revealing the rhetorical functions of Big Data that might otherwise be obscured or unnoticed. When applied to the discourses circulating around Big Data, rhetorical analysis offers a way into the characteristics, affordances, and constraints of Big Data without necessitating access to raw data or fluency in computational and/or algorithmic procedures.

A Note about Big Data in Composition and Rhetoric Scholarship

The implications that Big Data has for scholars in composition and rhetoric loom large, but scholarship on Big Data in composition and rhetoric remains relatively scant. The majority of the extant scholarship comes from researchers advocating for the application of Big Data technologies within the discipline. Faced with institutional imperatives for meeting the massive interdisciplinary push to adopt and create new digital tools, scholars in composition and rhetoric have begun experimenting with Big Data in their research. Much of the scholarship about Big Data in our field has been driven by both the desire to foster "a broader and deeper set of shared interests and intellectual commitments" between digital humanities and rhetorical studies as well as institutional imperatives for meeting the massive interdisciplinary push to adopt and create new digital tools (Ridolfo and Hart-Davidson).[i] Similarly, in rhetorical studies, new materialist scholars such as Casey Boyle, Thomas Rickert, and Jeff Pruchnic argue that Big Data could serve to expand rhetoric's horizons. They offer a worldview in which Big Data provides ample rhetorical opportunities, but they rarely consider the ethical consequences of such an expansion. Aimed at enticing researchers to use methods from data science, these calls typically remain at the level of disciplinary implementation, often alluding to the promise Big Data holds for our

discipline's research and pedagogical agendas. Like many other information technologies and tools, Big Data offers new methods to produce knowledge; however, striking a balance between the promise of Big Data and the values central to our field presents an interesting challenge.

Many scholars in our field demonstrate a strong attunement to questions of access and dedication to diversity and inclusivity. However, barriers to access to Big Data technologies (and to the actual data) place limitations on who can use Big Data as a method. And, although Big Data enables scholars in composition and rhetoric to identify larger patterns of activity in their research than they would be able to see otherwise, individuals are depersonalized and, in many ways, dehumanized, in the process of becoming data. Scholars like Rice seek to trouble Big Data's new role in the world. In her recent *Philosophy and Rhetoric* essay "The Rhetorical Aesthetics of More: On Archival Magnitude," Rice interrogates the "faulty logic" of Big Data in which bigger and more numerous data sets yield better and more accurate information. This logic, she argues, "has led to some rather substantial big data failures" (28) Although these critical perspectives on Big Data gesture at the kind of rhetorical approach that I am arguing for in my dissertation, the extant studies are limited in scope and make up only a portion of the already small selection of scholarship on Big Data in the field.

For the remainder of this literature review, I will divide this interdisciplinary scholarship into two camps: scholarship that addresses Big Data in research and scholarship that addresses Big Data in teaching.

Big Data & Research

In considering data analytics' potential for composition and rhetoric research, Nathan Johnson argues that Big Data methods have "stronger currency" for funding and publication

initiatives because of their capacity to make discipline-specific research more legible and credible for institutional and financial stakeholders than qualitative research methods typical of composition and rhetoric research (101). One Big Data technique in particular, data visualization, has been presented as an efficient and effective way to present composition and rhetoric research to a range of audiences and stakeholders. Scholarship calling for increased engagement with digital humanities has (quite bluntly) claimed that data visualization, along with other Big Data techniques, can be leveraged to motivate public interest and investment in composition and rhetoric research in an era of widespread devaluing of the humanities (Johnson 101; Losh; Manovich and Gold).

Along these same lines, Big Data has been pitched as a tool for constructing new ways of making sense of work being done in the field. Mueller, Johnson, and Borgman have emphasized the potential for Big Data and data visualization to reframe how composition and rhetoric scholars understand the boundaries of our field. Mueller argues that, as a reflexive method, this type of Big Data research offers a new perspective for "seeing" how knowledge circulates within the discipline. In his essay "Modeling Rhetorical Disciplinarity: Mapping the Digital Network," Johnson presents factor mapping as a method for identifying previously unconsidered factors influencing the publication and circulation of academic texts from composition and rhetoric. Factor mapping is a technique that uses algorithmic procedures to synthesize pre-existing data and create visualizations, or "maps," of that synthesized data. Johnson explains that factor mapping enables scholars process pre-existing data from a composition and rhetoric text corpus (i.e. titles, headings, publication locations, and citations along with other available metadata) in order to uncover hidden factors at play in the field's publication practices and the influence that institutions, geography, and social relationships have for the production and circulation of

knowledge within the discipline (106). Johnson's argument is similar to others' pitching data visualization as a way to reassemble and re-see our research, bringing new perspectives to data that would otherwise remain inaccessible (Borgman; Ball et al.).

In addition to these advocates, there are many scholars who push back against applications of Big Data in composition and rhetoric research. Big Data's magnitude can appear antithetical to those values central to disciplines within the humanities who have worked hard to shed light on non-majority research subjects. Bethany Nowviskie attributes the resistance to Big Data as a part of a larger pattern within the humanities to value "small data," such as the data gathered through close-reading and careful analysis of an individual narrative, over Big Data. She explains, "Humanities scholars have made major theoretical advances and practical advances in the cause of social justice, by bringing forward carefully observed and exquisitely described little examples. Our small data add nuance and offer counter-narratives to views of history and the arts that would otherwise fall out along blunter lines" (Nowviskie). This argument for narrative over or alongside statistical data has been made long before Big Data was realized as a potential method for academic research, as statistical studies' coarser grain of analysis obscured some of the intimate details qualitative research values most. However, unlike traditional approaches to statistical data, Big Data methods draw conclusions from less rigorously defined samples and make fewer demands for researcher accountability. Further, although Big Data enables scholars in composition and rhetoric to identify larger patterns of activity in their research than they would be able to see otherwise, many would argue that, in the process of becoming aggregated data, individuals are depersonalized and, in many ways, dehumanized.

Big Data & Teaching

In addition to scholarship detailing the affordances and constraints of Big Data as applied to disciplinary research initiatives, much scholarship about Big Data focuses on the potential for data analytics to be used in the classroom (Lang and Baehr; Griffin and Minter; Comer and White). Notably, this literature addresses Big Data as providing composition instructors with new pedagogical possibilities for tracing students' writing and learning processes via data collected with online courses and course management systems. For instance, in their article "Adventuring into MOOC Writing Assessment: Challenges, Results, and Possibilities," Denise Comer and Edward White discuss the implementation of Big Data in Massive Open Online Courses (MOOCs) and the implications these courses hold for the assessment of student writing (Comer and White). While Comer and White emphasize the danger of focusing on the Big Data produced by large courses at the expense of focused attention on individual students, they, like Anson, argue that large-scale data analytics are inevitable in higher education. Rather than avoid engagement with these types of courses out of principle, Comer and White push researchers and instructors to devise assessment practices that have the potential to align with pedagogical values *and* can operate in online learning contexts.

Importantly, these pedagogical possibilities are typically pitched as *future* possibilities that demand critical, theoretical attention before they can be realized in the classroom. Some key issues delaying the implementation of Big Data in the classroom include inconsistent technology training and support for instructors, concerns about students' and teachers' security and privacy concerns, and inadequate professional development and pedagogical preparation that is tailored for online writing instruction, among others. Additionally, Big Data raises questions about access, diversity, and inclusion (I will elaborate on these concerns in the next section). For

research, both within and outside of the university, barriers to access to Big Data technologies (and to the actual data) place limitations on who can use Big Data as a method (boyd and Crawford).

Limitations of Current Scholarship

While the need to recognize and contend with the complexities of data-driven research has begun to gain traction in digital rhetoric and composition studies, principally via approaches that examine the specificities of Big Data in practice, too often such approaches fail to contend with asymmetries of power that exist between differentially embodied human beings. Most of the scholars advocating for Big Data, for pedagogical and research purposes, seek to disrupt "institutional hostilities" toward Big Data, arguing that large-scale data science methods should be recognized as an approach suitable for research in composition and rhetoric (Losh 286). Tracing how scholars talk about the limitations of Big Data, however, illuminates the big concerns that Big Data presents for the field. Although including limitation sections in research are valuable in that they can foster a sense of accountability for both researchers and readers, advocates seeking to mobilize Big Data in their local environment tend to rubber stamp their papers limitation sections that often mention the following:

1. Limits of sampling to be accurately representative (sampling bias)
2. Challenge of seeing smaller patterns
3. Risk of reinscribing power dynamics and deepening gaps in privilege and access

While many scholars list these limitations, they rarely go into detail about the implications they pose for research and teaching. This trend of providing only a shallow articulation of limitations risks misrepresenting Big Data's efficacy. The lack of critical attention to the limits of Big

Data—especially given the gravity of the potential consequences that accompany these limitations—resonates with arguments made by non-academic advocates of Big Data who see Big Data through an objectivist lens. And, although there are scholars, like Anson, Ball, Johnson, and Losh, who advise their readers to use a mixed-methods framework that draws from disciplinary methods and Big Data methods to mitigate the potential risks of Big Data, explanations of *how* to engage in this kind of mixed-methods framework often fail to move beyond pointing to the need for future research to adequately develop an appropriate methodological framework.

Examining Big Data's Promise through Rhetorical Analysis

Rhetorical analysis, as conceptualized for this project, directs attention to two elements central to the Big Data phenomenon: 1) the factors motivating the investment in and adoption of Big Data; and 2) the material consequences arising from Big Data technologies and methods once implemented and put into practice. In using rhetorical analysis as a lens through which to interrogate Big Data, this dissertation answers Jessica Reyman's call for more scholarly interest in "examining (un)ethical interface and platform design and data practices, exploring informed responses and actions challenging unethical practices, and theorizing about what frameworks and approaches for ethical human-machine collaborations might look like" ("The Rhetorical Agency of Algorithms").

In developing my methodological framework for this project, I follow the lead of rhetorical scholar Barbara Warnick who, in her book, *Critical Literacy in a Digital Era,* challenges the transparency of new media, as well as the unquestioned and uncritical movement that the digital inspires. By employing a critical rhetorical lens to the "persuasive discourse about

technology" circulating in popular media, Warnick argues, we can make visible the ways that these persuasive discourses and their accompanying, but often unspoken, ideological assumptions, "affec[t] how we think" about the technologies we use:

> In regard to discourse about new technologies, we need to consider what claims are credible, what evidence is accurate, and which spokespersons are truly acting in the public interest. We also should recognize explicitly how advocates and writers use narratives, myths, forms of language, and visual images to tell their stories. Through critical examination of these features, we can begin to see what ideologies are at work and whose interests are being served by the discourse. (*Critical Literacy*)

Accordingly, this dissertation examines the discourses about Big Data via a critical rhetorical framework, paying explicit attention to moments where the promises ascribed to Big Data (to tell us something new about the world, to save money, to increase efficiency, to personalize) function both to justify its adoption and to defer reckoning with its failure.

Dissertation Overview

In what follows, I use three case studies to juxtapose the *promises* against the *practices* of Big Data. Although not wholly representative of the Big Data phenomenon (accounting for all the ways Big Data informs contemporary experiences of the world is impossible), this combination of case studies creates possibilities for seeing and interrogating the larger networks of data and Big Data analytics that play an increasingly powerful role in shaping our on- and offline interactions with and in the world, while also providing insight into different ways of framing the systems and processes undergirding Big Data.

In chapter two, I interrogate the implementation of Big Data-driven campaigning practices over the past twenty years. My analysis examines the development of national voter files and the rise of social media electioneering, focusing specifically on the algorithmic limitations and biases of national voter databases and the political misuse of social media advertising services for voter suppression operations. By tracing the role that Big Data plays in contemporary US electioneering (namely as a funnel for exclusionary politics and as a tool for deploying manipulative and invasive campaign tactics), this chapter demonstrates the political, social, and cultural implications of Big Data for our collective understanding of democracy.

In chapter three, I examine Big Data as biotechnology, focusing specifically on the personal genomics company, 23andMe and the issues of demographic representation that undergird the company's collection, analysis, and distribution of genomic data. My analysis focuses on 23andMe's privacy policy and research consent documents, material aimed at consumer-level engagement including 23andMe.com and the company's advertising campaigns, as well as the company's research publications and documents detailing their corporate partnerships. This chapter demonstrates the power of Big Data to discipline bodies and dictate narratives of individual identity, as well as ideas of familial and community belonging.

Finally, in chapter four, I investigate the implementation of Big Data in education, with a focus on initiatives (from both a national and an institutional level) aimed at deploying learning analytics in online learning spaces in higher education and the implications that these data-driven assessment practices have for how we understand learning. I focus primarily on Instructure, the learning technology company behind the learning management system Canvas, and demonstrate that although learning analytics are designed to enhance student learning, in practice, these

mechanisms cover over the different ways that students actually learn and engage with course material.

The issues raised in each chapter should not be understood as particular to their respective contexts of healthcare, elections, and education; they are illustrative of Big Data practices and applications more broadly. Accordingly, each chapter follows a pattern which I argue is representative of many Big Data applications. First, stakeholders adopt Big Data for its promise to help them achieve some future goal (e.g. to secure votes, to decode genes, and to assess learning). Second, because Big Data requires data, data brokers make new promises about how the users (or "data donors") will benefit from the individualizing capabilities of the algorithm. Then, because Big Data technologies are owned by privatized companies, their data collection procedures, algorithms, and results are blackboxed. When the stakeholders achieve their goal (Big Data is popular for a reason, after all), it comes at the expense of the algorithm's democratizing promises. Importantly, the systems implemented, despite failing according to their own standards, are nevertheless continually reinfused with the same optimistic assurances of capacities to come, inevitably leading to the reinforcement and entrenchment of pre-existing social inequalities. In each case, I show that, by removing sites of accountability while claiming to be a means of establishing accountability, Big Data guarantees its own success as well as its own failure.

Chapter Two

When Data Trumps Democracy: Big Data in American Presidential Campaigns

When it comes to Big Data and government, "utopian claims abound" (Elmer et al. 14). Big Data claims to be a revolutionary tool that will make all levels of government and politics more productive, more efficient, more economical, and more democratic (Richards; Morabito). Motivated by the success of data-driven content marketing in the commercial sector, and enabled by increasingly efficient and affordable means for data storage and processing, political investment in Big Data has skyrocketed. While data-driven initiatives are visible in a range of government programs from healthcare to national security to environmental protection, over the past decade, the mark that Big Data has made on the American political landscape has become particularly striking within the context of contemporary electioneering. As evidenced by the most recent election cycles, Big Data campaigning has become the norm in US electoral politics, with political hopefuls from both sides of the aisle turning to Big Data to determine the best possible path to victory.

While presidential candidates in the United States have long sought to understand voters by way of data (e.g. through census data and polling results), the integration of Big Data-driven electoral politics has shifted campaign practices. For the purposes of this chapter, I will focus on four key effects of Big Data's intervention in politics: first, candidates and their campaign staff are no longer dependent on survey data and polls to determine campaign strategy; now, campaigns, and perhaps more importantly, political parties, have turned their focus towards the project of amassing large, dynamic voter databases. Second, campaigns have moved away from mass communication strategies and toward micro-targeting techniques that draw on personal

data bought from data-brokers as well as publicly-available voter data. The move from macro- to micro-communication has resulted in a third shift, which is the increase in the allocation of campaign funds for data analytics. And finally, candidates have turned to social media as a primary venue for voter outreach, displacing the popularity of traditional offline methods for reaching voters (e.g. mailers and newspaper advertisements). The surge of electoral investment in Big Data has been fueled by the promise that, through data, candidates and parties can better understand individual voters and, thus, become better equipped to accurately represent their constituents upon election into office. In practice, however, this promise falls short. The failure of data-driven politics to live up to its promise is not due to algorithmic limitations or mechanistic error. Rather, Big Data fails in electoral politics because the goal is not democratic representation—the goal is to win elections (Kreiss).

These shifts in campaign practices have met backlash on the grounds of constituent privacy, and the ubiquity of Big Data in elections has been critiqued as a form of surveillance, specifically, voter surveillance (Rubinstein; Bennett). While it is important to address issues of privacy that accompany these shifts in electoral politics, I argue that it is equally important to recognize the rhetorical implications that Big Data campaigning has for collective imaginings of democracy. With each of the shifts in campaign strategies enumerated above, presidential candidates further eschew values central to democratic participation, sacrificing transparency for efficiency and equal representation for a politics by exclusion. The amalgamation of Big Data has fundamentally shifted the scope of electoral politics toward increasingly invasive and exclusive campaign practices, imbuing candidates with the power to harness data to manipulate and skew select voters' perceptions of the electoral landscape, while actively suppressing others. Like the better-known application of Big Data to gerrymandering, Big Data in political

campaigning functions, not as a tool for fostering political participation or equal representation, but as a mechanism for entrenching divisive, hyper-partisan politics.

The Rise of Big Data Electioneering

When Barack Obama was elected president in 2008 and re-elected in 2012, his wins were widely attributed to the well-orchestrated deployment of targeted marketing strategies devised using Big Data algorithms (Newman; Stirland et al.). While Obama's 2008 run has been hailed as the first successful Big Data-driven campaign, the groundwork for his campaign was actually laid four years earlier by Democratic presidential candidate, Howard Dean. Although ultimately unsuccessful, Dean's 2004 presidential campaign was lauded for being the most technologically advanced and digitally innovative campaign the Democratic party had yet seen (*Rhetoric Online*). In 2003, Dean became the first Democratic nominee to refuse federal matching funds. Instead, Dean's campaign staffers used the internet as their primary space for political fundraising, soliciting many small donations from a large online audience. In the end, the campaign raised over 50 million dollars from over 350,000 individual donors (Hindman).

By uniting grassroots organizing and internet-enabled voter and donor outreach, Dean's 2004 campaign illuminated new possibilities for electioneering in 21st century America. Discussions about the implications of internet-driven campaigning sparked public speculation that the role of the internet would eventually extend far beyond the scope of online fundraising efforts, affecting campaign decisions more broadly (Stromer-Galley and Baker). Take for instance, the following quote from the New York Times' Glen Justice, in an article covering Dean's campaign:

Many who envision a day when the Internet plays a more prominent role believe it will

> actually influence the campaign's message when a candidate decides to run. Already, Internet consultants are beginning to join campaign managers, media specialists and other top advisers at the head table. The candidates who emerge may even change, as challengers play to the online audience's affection for insurgents with a biting message. (Justice)

Justice's prescient analysis signals an impending paradigm shift in American electoral politics toward what Daniel Kreiss, in his book *Prototype Politics*, calls "technology-intensive" campaigning. This type of campaigning, Kreiss argues, emphasizes "personalized and socially embedded forms of electioneering," and uses a combination of Big Data analytics, online social networks, and platform technologies (5).

After dropping out of the 2004 race, Dean founded the progressive political action committee Democracy for America (DFA), which he used to promote grassroots fundraising efforts among "other like-minded Democratic candidates and organizations" (*Democracy for America*). Built upon the online grassroots organizing strategies, resources, and data infrastructure from the Dean campaign, DFA conducted fundraising efforts and provided both voter and volunteer outreach services for its endorsed candidates (including Obama) and organizations.[ii] In 2005, when Dean was elected DNC chairman, he helped propel technological developments for electoral politics at the party-level. Under Dean's leadership, the DNC focused efforts toward building the party's digital infrastructure for online grassroots fundraising and establishing partnerships with technology and data-centric political organizations (many of which were headed by former staffers from the Dean campaign).

Motivated by past successes with Internet-enabled campaigning, Dean enlisted Voter Activation Network (VAN), a private company specializing in voter databases, to create

VoteBuilder, a centralized voter database and campaign interface for connecting Democratic campaigns with voters (Blake). While VoteBuilder helped unify and streamline the party's data collection efforts, it essentially became a centralized repository for years of inconsistent and fractured data collected from dozens of different campaigns. These issues in data quality meant that the DNC would need to conduct thorough data maintenance in order to improve the quality and accuracy of their party's voter data. Efforts to clean up the DNC's voter data were initially unsuccessful. However, legislative developments in the aftermath of the 2000 US Presidential Election—when former Vice President Al Gore's calls for a recount in Florida sparked national debates about the efficacy of state voting systems—provided VAN (and other national voter database specialists) with a seemingly perfect solution to their messy data problem.

The Help America Vote Act and the Development of National Voter Files

Investigations into the 2000 election raised questions about the efficiency and accuracy of using paper, punch, and lever (i.e. analog) ballots while also revealing significant inconsistencies between state voter registration practices (Toobin). Subsequent calls for national election reform led to the 2002 passage of the Help America Vote Act (HAVA), a measure designed to assist state governments in increasing democratic participation and accounting for all voters. In addition to supporting and partially funding the development and implementation of new voting technologies, HAVA required all states (with the exception of North Dakota, which has no voter registration requirement) to construct and maintain a "single, uniform, official, centralized, interactive computerized statewide voter registration list" (US Congress).

While the original goal of HAVA was to make political participation more accessible and accurate to the American public, the primary beneficiaries of this piece of electoral legislation

are not individual voters as much as political parties and campaigns. In the decades prior to HAVA's passage, political campaigns primarily used surveys and geographical measures to guide campaign decision making. In the 1970s, for instance, campaigns began using historical records of precinct-level election data to make predictions as to how that precinct would perform in an upcoming election. They then would allocate campaign resources toward targeting supporter- and swing-precincts with campaign communications (e.g. mailers and rallies). In the 1980s, campaigns began to conduct geodemographic targeting, a method that uses Census data combined with geographic marketing data to group people according to similarities in demographic data and behavioral patterns, essentially "leverag[ing] the assumption that people choose their neighborhoods according to their lifestyles" (Endres and Kelly 4). Although geodemographic targeting provided campaigns with different strategies for grouping voters than were possible with precinct-targeting strategies, both geodemographic- and precinct-targeting strategies are geographically bound. This meant that they often missed individuals whose political beliefs did not line up with the majority of their neighbors. The development of state voter registration files, however, meant that campaigns could begin identifying voters at the individual-level.

State voter registration lists act as the infrastructural skeleton upon which partisan and commercial organizations, like VAN, develop national voter databases, like VoteBuilder. While there are some differences in terms of what types of data are collected and published in their respective voter registration lists, the majority of the state voter files include the following information for each registered voter: name, residency information, gender, and date of birth, as well as a unique numerical identifier. Some southern states' voter files also include racial identity (Alabama, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and

Tennessee). In addition to the data included in each HAVA-mandated state voter list, these partisan voter databases, or national voter files, contain data drawn from a range of governmental and commercial sources, data purchased from commercial data brokers, and data mined from campaign activities. Through the use of personal identifiers, like name and address, contact information, IP address, and device IDs, campaigns have been able to compile these datasets and construct individual-level voter profiles (Rubinstein).

It is important to note that while the HAVA state voter registration files did ease some of the challenges of voter data collection, building the infrastructure for a national voter database remains no small feat. To construct a sustainable voter database independently, a candidate would need to devote a substantial amount of their campaign's time, funding, and resources toward building digital infrastructure and gathering data. This process would either need to start over each time or be passed down from campaign to campaign. It makes sense, then, that the mainstreaming of Big Data-driven campaigns has largely been driven by political parties. Because parties are not bound by election cycles or term limits, they can take a long-term approach to building data infrastructure. Further, due in part to their status as non-profit organizations as well as their organizational strength, political parties are uniquely positioned to develop sustainable Big Data systems. This flexibility has enabled party networks to combine data from a multiplicity of sources to develop massive voter databases that are dynamic, meaning they can be updated over time (Kelly and Hamm). These sophisticated data infrastructure systems and detailed voter databases have become some of the most valuable campaigning resources that a political party can provide (Hersh 209). Because the data infrastructure required to be a competitive political party is contingent upon having a strong and stable party network with extensive budgetary capacities, fringe parties are at an immediate disadvantage, which

further ensconces the power of the United States' two-party system.

Keeping in mind the context of the Dean campaign, HAVA's passage, and the rise of national partisan voter databases, I want to return now to my earlier discussion about Obama's Big Data-driven campaigns. In 2008, Obama was the first Democratic candidate to benefit from the DNC's newly streamlined, HAVA-enabled voter database. Like Dean, Obama's fundraising was a grassroots effort that relied on internet-enabled donor, volunteer, and voter outreach efforts. During both the 2008 and 2012 election cycles, Obama's campaign was able to use VoteBuilder to capitalize on the popularity of social media sites, the increasing ubiquity of mobile connectivity, and new developments in data analytics in the commercial sector (Stromer-Galley). Obama hired a significant number of staffers from the tech industry (including a number of Dean's former staffers), whose expertise in data, analytics and digital media he harnessed to develop intricate voter databases and a sophisticated microtargeting system through which to conduct voter outreach efforts. Over the course of the campaign's tenure, Obama's network of 2.2 million volunteers gathered an unprecedented amount of data on individual voters (McKenna et al.). In addition to collecting data from every interaction that campaign staffers and volunteers had with voters, including phone calls, in-person canvassing, and mailings, Obama's team collected extensive amounts of data from the campaign's social media sites and email systems, as well as, and perhaps most importantly, the campaign's website, BarackObama.com. Built to maximize the quantity and quality of data available to the campaign, this website was central to the campaign's data collection efforts. According to two of Obama's staffers, Matt Compton and Andrew Brown (who went on to serve as the Digital Director and Technology Directors of the DNC, respectively), the website's digital infrastructure was a groundbreaking development for connecting digital and in-person campaign efforts:

> In 2012, when you signed up for information at BarackObama.com, a piece of technology code named Airwolf compared your voter profile against a data model to determine your support score, and then automatically delivered a message from a local campaign organizer reminding you to cast a ballot, sign up to make calls, or volunteer to canvass your neighborhood. It made the connection between the online and offline worlds of the campaign nearly seamless and helped to make the best field program in the history of American politics a little bit more efficient, a little bit better. (Compton)

The campaign added to its own in-house data gathering operation data from voter registration, the census, and political polling, as well as data purchased from third-party data brokers, credit monitoring services, and cable service providers. In addition to using VoteBuilder, the campaign partnered with a number of up and coming data analyst firms that helped facilitate the analysis and application of the data gathered. The product of their extensive data mining and analysis efforts was an interactive voter database. These databases were primarily used to make predictions about voter behaviors, attitudes, and receptivity (Kitchin). In 2013, the Obama campaign gave all of its data and tech resources to the DNC, helping expand the depth and scope of VoteBuilder.

Party Databases, Party Rules

Because voter databases, like VoteBuilder, and their underlying algorithmic systems operate within party ecosystems, they are thus also mechanisms of partisanship, and the parameters of their use are reflective of party values. According to a recent PEW study, voter databases are predominantly used “to help political practitioners more effectively and efficiently

engage with potential voters," or, more specifically, "to identify potential supporters and to communicate with them, either to influence their candidate choice, mobilize them to turn out to vote, or both" (Igielnik et al. 41). Voter databases are, essentially, mechanisms of classification. As Safiya Umoja Nobel explains in her book *Algorithms of Oppression*, "Classification systems...are part of the scientific approach to understanding people and societies, and they hold the power biases of those who are able to propagate such systems" (Noble 116). The variables of classification built into these voter databases determine what and who is seen and heard.

Across the board, the most valuable variable of classification a voter database can offer is political party affiliation. However, partisanship can be difficult to trace. Because the HAVA voter registration databases were created at the state level, different state governments have had the power to decide what data to include in their voter databases—some states voted to include party affiliation data, while others did not. The gaps in party affiliation data between state databases have left analysts searching for any other data, or *combination* of data, that might signal a voter's partisan identity.

To assist in the identification of supporters, campaigns commonly use a data analysis technique called predictive modeling. Typically developed by campaign data analysts, predictive models crunch available data—including data gathered from state voter lists, donor rolls, and survey results, as well as, and perhaps most importantly, data that citizens provide to campaigns directly—to generate three types of individual-level predictive scores: a support score, responsiveness score, and behavior score (Nickerson and Rogers). Support scores predict how likely an individual is to support a particular candidate or issue. Responsiveness scores predict how an individual will respond to targeted campaign messages. Behavior scores predict what behaviors, or activities, an individual is likely to engage in, like turning out to vote, donating,

volunteering, or signing a petition. While support scores are useful for identifying people with strong partisan affiliations and determining which issue(s) they care about most, responsiveness and behavior scores can be used to decide which voters are worth reaching out to and which are not given the campaign timeline and budget realities.

Predictive modeling provides campaigns with a data-driven system for streamlining personalized voter outreach efforts. Instead of relying on precinct-level data to determine which voters to contact, campaigns (and parties) can segment the population into categories according to individual predictive scores. They can then select categories of voters to target with campaign messages (and which messages to target them with), and categories of voters to ignore. By limiting their contact with voters who are unlikely to support them or unlikely to turn up to vote, and by increasing their contact with voters whose scores show a high likelihood of supporting the candidate and/or issue *and* a high likelihood of turning out to vote, campaigns can theoretically improve their chances of winning the election. The race to mobilize voters with high support scores has meant that campaigns in recent years have spent little to no resources toward targeting swing voters and inactive, but registered voters (Nickerson and Rogers).

As is the case with all predictive modeling techniques, the quality of a predictive score is contingent upon the quality of the underlying database—"the knowledge logic of algorithms relies on the nature of the databases that serve as their foundations, the patterns of what is included and what is excluded" ("Rhetorical Agency of Algorithms" 114). Despite the level of access that government officials have to personal data, national voter databases are neither infallible nor are they fully comprehensive. When campaigns use aggregated predictive scores, for example, to facilitate large-scale campaign decision making, these decisions are necessarily based on "a simplified and distorted version of the electorate that is based on the data available

to them" (Hersh 12). When campaigns populate their voter databases with data gathered from within the campaign and party itself (using email lists, donor rolls, and/or social media profiles), they necessarily end up with more data from their supporters than from non-supporters. Regardless of how fine-tuned a campaign's predictive models may be, as long as voter databases are developed within a partisan framework, resultant predictive scores will necessarily skew toward voters who identify with that party. In the 2018 study on the efficacy of voter databases, PEW researchers found that, "[t]he kinds of people who are most likely to be missed in the voter files when they move do not constitute a random subset of the population, but instead are more likely to be younger, less educated, poorer and nonwhite" (Igielnik et al. 42). Because they are dependent upon voter registration lists, national voter databases often exclude the same people—namely black and Hispanic voters—whose voting rights have been suppressed by restrictive voting laws and other structural barriers to voter registration (Jackman and Spahn).

Despite the fact that these Big Data-enabled gaps in the voter registration system coincide with (and reinforce) existing structural inequalities, progressives and conservatives alike continue to use voter registration data to decide which voters matter, which voices matter, and which issues matter. In doing so, political parties and campaigns actively perpetuate and contribute to racial inequalities in voter registration and democratic participation. The logistical challenges of identifying and locating unregistered voters, new voters, and voters who have recently moved has further skewed voter databases to favor individuals who are older, white, less mobile, and consistently politically active. As PEW researchers warn, "If efficiency in the use of campaign resources is a principal goal of practitioners (rather than engaging new or irregular voters), voter files could produce greater inequality in participation by making it easier for campaigns to avoid 'wasting' effort on younger or poorer voters who may have a low propensity

to participate in the first place" (Igielnik et al. 49).

In an era when much of contemporary campaign decision making is contingent upon whatever algorithmic structures are employed to process voter data, the advent of dynamic voter databases and the drive to use predictive modeling to capitalize on partisanship has contributed to already widespread political polarization. Tracing the role of voter databases in campaign decision making, Hersh notes, "The data environment will affect a campaign's perceptions, the perceptions will affect their strategies, and the strategies will affect which voters are mobilized into the political process" (Hersh 128). As mechanisms of classification, partisan voter databases and their underlying algorithmic structures function rhetorically to shift the interpretive lens through which campaigns see and understand voters. This shifting of interpretive lens distorts the underlying electoral reality and becomes a self-fulfilling prophecy, following the common pattern of Big Data systems more broadly. When algorithms like support scores, for example, are built into the Big Data system as a heavily-weighted category, voters are treated differently. Further, when campaigns open up access to the voter data they have obtained, freely and otherwise, to volunteers across the nation—usually in efforts to standardize canvassing strategies and increase the effectiveness of local voter outreach efforts—this interpretive lens is spread through party supporters. As more and more campaigns employ predictive scoring to decide which citizens to mobilize, the cumulative effects pose significant risks to our democratic state that extend far beyond the campaigns cycle.

The 2016 Presidential Campaign Cycle

The 2016 election cycle saw a near-universal turn towards Big Data. Data-driven campaigning flooded the electoral landscape, as large-scale data analytics were taken up by

presidential candidates who, like Obama, sought to tilt the electoral college in their favor by using Big Data algorithms to target specific voter bases. Hillary Clinton's campaign was largely a data driven effort; the campaign employed a substantial data team headed by Elan Kriegel, the co-founder of the data analytics company BlueLabs and former senior staff member for the Obama campaign's analytics team. Although Donald Trump initially expressed little interest in Big Data analytics (Pace and Colvin), in the summer prior to the election, the Trump campaign (along with other Republicans, like Ted Cruz and Ben Carson) enlisted the services of the now-infamous data management firm Cambridge Analytica (Federal Election Commission).

While managing the Trump campaign, Cambridge Analytica used questionably-obtained Facebook data to target voters off the platform, which understandably led some lawmakers and ethicists to sound the alarm. While the federal investigation into Cambridge Analytica revealed a number of the company's ethical violations, it was mainly focused on issues of user privacy in the context of the 2016 election. To understand the place of Cambridge Analytica in the larger context of Big Data electioneering, it is important to consider the company's history and influence more broadly.

Cambridge Analytica

Founded and funded by tech-billionaire Robert Mercer (the highest single campaign donor in the 2016 election) and Stephen Bannon, Cambridge Analytica originally made a name for itself advertising state of the art psychographic profiling and targeting. Cambridge Analytica framed its campaign management system as the ultimate campaigning toolbox, complete with five campaign services: research, data integration, audience segmentation, targeted advertising, and evaluation. Cambridge Analytica sold their flagship product, "Audience Segmentation," not only as a method for categorizing individuals based on data correlations, but as politified,

predictive data science. In a presentation on Big Data electioneering, Cambridge Analytica's former CEO Alexander Nix explained, "We segment your electorate into distinct audiences using predictive analytics, a form of artificial intelligence that takes into account the behavioral conditioning of each individual to create informed forecasts of future behavior" (Concordia). Upon completion of "Audience Segmentation," clients could then enlist Cambridge Analytica's "Targeted Advertising" services, which Nix pitched as a foolproof tactic of persuasion: "We don't need to guess at which creative solution may or may not work…we can understand exactly which messages and going to appeal to which audiences before the creative process starts." Once a subset of voters was identified as being receptive to a particular message, advertisements could then be carefully tailored to match that intensely narrow audience.

The methods employed by the company during the 2016 election cycle were not developed in-house. Rather, Cambridge Analytica drew inspiration from research that began almost a decade prior at Cambridge University's Psychometrics Centre by Michal Kosinski and David Stillwell. Kosinski and Stillwell's work focused on the predictive power of data science, specifically the use of data analytics to predict information about individuals based on digital records: "People may choose not to reveal certain pieces of information about their lives, such as their sexual orientation or age, and yet this information might be predicted in a statistical sense from other aspects of their lives that they do reveal" (Kosinski et al.). Their research came out of a Facebook app that Stillwell, the current deputy director of Cambridge University's Psychometrics Centre, created in 2007, before beginning his graduate school career. The app, "MyPersonality," worked by aggregating data mined from Facebook profiles, friend lists, and "Likes" along with data collected from a standard OCEAN psychometric test to make predictions about individual users' identities, from gender, race, age, socioeconomic status and

sexual orientation, to political affiliation and level of education (Matz et al.).

In 2013, Stillwater and Kosinski, along with Thore Graepel, a Microsoft researcher who has since become the Research Lead for Google DeepMind, published research findings from a collaborative study on the MyPersonality app (Kosinski et al.). Using data from over 58,000 volunteers (e.g. users who had downloaded the app, accepted the user agreement, and connected their MyPersonality profile to their Facebook profile), the researchers were able to construct extensive demographic profiles. Using social data as well as results from psychometric tests, their proposed model would use Facebook Like data to predict individual users' psycho-demographic profiles; these profiles were then compared with users' psychometric test results. The researchers found that their model was able to predict individual user's psychodemographic profiles with a significant degree of accuracy: "The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait 'Openness,' prediction accuracy is close to the test–retest accuracy of a standard personality test." Follow ups with study participants further demonstrated the model's efficacy (Matz et al.).

Kosinski and Stillwell's research garnered attention from a Cambridge Analytica representative, who, in 2013 approached the researchers about a potential partnership, wherein Cambridge Analytica would gain access to their model as well as to the entire MyPersonality database. Kosinski and Stillwell declined the offer. In later publications, the researchers explain that the ultimate goal of their work with MyPersonality was not to actually apply the model for commercial purposes, but rather, to demonstrate the predictive power of psychometric profiling and to caution others about the risks of applying these data analysis techniques for the purposes

of psychological targeting and mass persuasion: “On the one hand, this form of psychological mass persuasion could be used to help people make better decisions and lead healthier and happier lives. On the other hand, it could be used to covertly exploit weaknesses in their character and persuade them to take action against their own best interest, highlighting the potential need for policy interventions” (Matz et al.).

After failing to recruit Kosinski and Stillwater, Cambridge Analytica reached out to another data scientist from Cambridge University’s Psychometric Center, Aleksander Kogan, who *did* agree to partner with the company. With funding from Cambridge Analytica, Kogan developed a new Facebook app called This Is Your Digital Life, which used the same data mining techniques as the MyPersonality app and successfully harvested data from over 50 million Facebook profiles. This data was then transferred to Cambridge Analytica, where it was repackaged as a comprehensive campaign management system.

Social Media Electioneering

While Cambridge Analytica has now become synonymous with exploitative campaigning, the mobilization of social data for political persuasion operations is not unique to Trump’s campaign. In fact, the politicization of social data is far from a new development in contemporary electioneering. For upwards of a decade, presidential campaigns have been using social media for official and unofficial political advertising. This type of campaigning, which Bossetta terms “social media electioneering,” has been largely restricted to platforms with active, extensive and diverse user bases and high functionality, like Facebook, Instagram, Twitter, and Snapchat (Bossetta).

Although social media platforms offer their social networking services to the public for free, the majority of these platforms *do* rely on user data to generate revenue. The fact that social

media sites capitalize on user data to make money is counterintuitive to commonplace understandings of the goals of a social media platform: "While we tend to think of social media corporations as free services to facilitate social connectivity among users, these corporations should be understood as platforms to deploy strategies to get users to produce data" (Elmer et al. 4). The technological affordances of these platforms and their underlying algorithmic structures, not to mention the vast amounts of readily-accessible data they house about individual users, make these platforms appealing venues through which to conduct voter outreach. Further, the digital architecture underlying these platforms' in-house advertising services affords campaigns the potential to "monitor and harvest users' digital traces and appropriate them for decisions relating to persuasion or mobilization initiatives" (Bossetta 477). And despite claims of neutrality, tech companies like Facebook and Twitter have played just as active a role in shaping the political landscape as their more decisively political Big Data counterparts like Cambridge Analytica.

By integrating mechanisms for collecting, analyzing, and sorting user data directly into their platform's algorithmic structure, social media developers have been able to amass enormous user databases, which they leverage to sell sophisticated and profitable suites of micro-targeted advertising services. These services rely on cookies, digital tracers that follow and report on users' activities and behaviors as they move within and across platforms. Before joining a social media platform, for example, users are asked to consent to the use of cookies; any time a user uses the platform thereafter (e.g. changes a user setting or preference, interacts with another user, clicks a link, or searches for a page), the cookie records data and sends it to the platform's website and server. Platforms can then use aggregated cookie data to train search engine algorithms, to identify patterns in behavior, to conduct personalized and micro targeted

advertising, and to streamline user navigation. And, as I explained in the introduction, user data can be shared across platforms, essentially becoming fodder for complex targeting algorithms that mediate everything a user is likely to encounter across many platforms (Hess). To better illustrate how cookie data works in the context of social media electioneering, we can look at the digital infrastructure undergirding one of the most popular tools for social media electioneering: Facebook Ads.

During the 2016 election cycle, Facebook was by far the most popular platform for social media electioneering—the platform's suite of targeted advertising services, especially the built-in matching, targeting, and analytics features, were especially appealing for campaigns looking to streamline their voter outreach efforts. When using Facebook Ads to deploy targeted political messages, all campaign teams followed roughly the same process. First, as I mentioned before, campaigns used aggregated predictive scores to construct audience groups for campaign targeting. Then, campaigns matched individual voters who had been selected for campaign targeting to their respective Facebook profiles. To do this, campaigns merged voter data (i.e. voter files, in-house campaign records, and donor rolls) with personally-identifying data (e.g. email addresses, phone numbers) purchased through third-party data brokers. They then uploaded this data to Facebook Ads using the "Custom Audiences" feature, which enabled campaigns to identify individual users. In addition to selecting a Custom Audience, campaigns used the platform's "Lookalike Audiences" feature. This feature uses algorithmic filtering to identify individual users whose data trails indicate that they are likely to behave similarly to users in a select Custom Audience group(s). Finally, campaigns used Facebook's "Core Audience" feature, which allowed them to select predetermined categories of users to view select ad campaigns. Facebook provides a number of different Core Audience categories for campaigns

to choose from, including demographic categories, such as age, location, gender, education, financial status, job, ethnicity, as well as interest and behavior categories, such as "Recently moved," "US politics (very liberal)," "Multicultural Affinity: African American (US)," and "Technology early adopters." Facebook Ads' 'Core Audience' groups are constructed using predictive analytics; just as data analysts for political campaigns use predictive scoring to segment the population into similar subsets of voters, Facebook Ads uses predictive algorithms to sort cookie data and create clusters of users grouped according to similarities in the data they produce (e.g. posts, comments, likes) as well as the data mined from their online activities (e.g. watching a video, clicking, or even lingering on a post or page) ("Advertising Policies").

In addition to their matching and targeting services, Facebook Ads also provides analytic services that enabled campaigns to track how audience groups respond to targeted advertisements. Essentially, these services use cookie data to track how users "behave" when presented with a particular ad. Whenever users like, share, click, flag, or dismiss a targeted ad, their cookies generate data, which are then aggregated across audience groups and analyzed to determine the effectiveness of that particular ad campaign. Explanatory analytics gauge the ad's relative effectiveness, prescriptive analytics suggest changes, and machine learning algorithms recalibrate the predictive models.

Partisan Filter Bubbles

As is now widely known, the spike in campaigns using social media platforms to disseminate campaign messages has resulted in increasingly impenetrable digital 'filter bubbles,' wherein individuals are less likely to encounter diverse or challenging opinions. When voters' only interactions with political candidates center exclusively on issues that the voter has been

pre-determined to support, there is little opportunity for voters to broaden or deepen their political understanding, let alone change their attitudes. As Aaron Hess explains in his article on digital rhetorical identification, “Political information is bound up in our own digital identification process, meaning that we are reinforced and rewarded for believing in the things that we (likely) already believe. In other words, if a user believes in one side of a political argument, it is likely that digital cookies and their algorithms will continue to find information that is relevant to and supportive of that position” (Hess 13). Making social data partisan, and enabling companies to sell based on partisan identity, creates not just a filter bubble whereby information is partisan, but an entire social ecosystem where every taste, image, icon, artifact is imbued with partisan meaning. Likewise, as data scientist Rob Kitchin explains, “[A]s we come to use and rely on databases and data infrastructures to make sense of and do work in the world, our discursive and material practices adapt and mutate in response to them" (24).

The cumulative effect of targeting voters via social media platforms has resulted in users’ becoming further enmeshed within algorithmically-generated partisan bubbles that Mary Stuckey argues “reflect specific understandings of the nation and who among its citizens are most welcomed and valued” (670). In her discussion of the implications that partisan understandings have on individual perceptions of political reality, Stuckey claims:

> [O]ur political worlds are made up of different and contesting views of the best locus of political authority and competing depictions of the polity, which are grounded in the political realities we witness and in which we participate. These depictions lead to and reinforce different political hierarchies, which are justified though differing understandings of our political myths. (Stuckey 667)

Expanding Stuckey’s argument to the context of Big Data campaigning reveals the

rhetorical power of social media electioneering. Earlier, I argued that, as partisan mechanisms of classification, voter databases (and their underlying algorithmic structures) shift the interpretive lens through which campaigns, parties, and candidates understand citizens to privilege (predictions of) partisanship. Likewise, when the messages that a voter receives from political campaigns are embedded with and resonate with all of the other content that is being threaded through their social media feed, the interpretive lens through which citizens see their social networks becomes increasingly influenced by partisan values.

Voter Suppression Operations

Recently, campaigns have experimented with using targeted advertising on social media to strategically obscure or reveal different elements of a candidate's platform to different categories of voters. While this type of strategic messaging is not new, when used in the context of social media filter bubbles, the effects can be especially powerful and divisive. In 2016, for instance, Donald Trump's campaign conducted what Brad Parscale, the digital media director for Trump's 2016 campaign (and his newly named 2020 campaign manager), termed "voter suppression operations" using Facebook's advertising feature (Green and Issenberg). One of these voter suppression operations involved the production of a political advertisement claiming, "Hillary Thinks African Americans are Super Predators." This ad was distributed to select African American male voters using Facebook's "dark post" feature, which allows advertisers to control the audience for their content, or as Parscale explained, "Only the people we want to see it, see it." Obviously, the "Super Predator" ad was not created with the intention of persuading African Americans men to vote for Trump. Rather, it was designed to dissuade African American men from voting for Hillary Clinton, essentially suppressing votes. Because Facebook's dark posts quite literally remain invisible to those voters who are not meant to see

them, however, these voter suppression operations largely remained hidden from public sphere.

The implications for democracy are substantial. Rather than educating voters about a candidate's platform, this tactic works to diminish voters' agency to make informed decisions at the polls. Further, when deployed for political gain, this strategy decreases candidates' accountability, making following through with platform goals irrelevant. These issues have been amplified in recent years as political actors from all sectors of state and federal government have increasingly turned to national voter databases and Big Data analytics to manage constituent engagement while in office. This trend is facilitated by third-party companies who work with members of Congress and congressional staffers to construct databases that can be used to tag, track, and organize records of constituent communications, including emails, phone calls, letters, and contact via social media. Leidos Digital Solutions, Inc., for instance, specializes in digital communication services for government offices within the US. Their feature product, Intranet Quorum (IQ), is used by over half of Congress, as well as 40% of US governors. In addition to offering communication management systems for elected officials and governing bodies, IQ provides users with a legislative/demographics analytics tool, aimed at linking messages from constituents to voter databases. The IQ service is pitched as a way to connect more effectively with constituents:

> Our legislative/demographics analytics tool, LegiStats, provides aggregated demographic data on your constituents broken down by state, Congressional district and county. Detailed data points like voting history, education levels, estimated income and many more are presented in graphical format allowing for easy interpretation of complex information. LegiStats gives you the ability to tie Individual legislative actions to related bills, assign to a staffer for monitoring and even track success rates. ("Legislative Tracking")

This company and many others argue that by segmenting constituents based on voting history, income level, race, and other categorical markers of identity, elected representatives can better address constituents in context. Importantly, these constituent data management systems draw on the same voter databases that are used for campaigning purposes, and thus include data from HAVA-required state voter registration lists, as well as data from donor rolls, campaign website data, consumer data and data purchased from data brokers, among others. Through these systems, elected officials have the option to decide when and how to engage with which constituents using markers of partisan identity. Rather than bring representatives and constituents closer, these systems risk further fragmenting the public along partisan lines.

Conclusion

One of the central ironies of Big Data's intrusion onto the political landscape is that the policies and laws around voter records are, by and large, written in a way that privileges individual rights and access to democratic participation. However, this central goal is fundamentally undermined when voter records are not used in a way that reflects the spirit of these policies. Big Data firms, and by extension the politicians who adopt their services, promise

to individualize politics, but individualizing politics is a double-edged sword. The same technology that affords parties the ability to reach out to underprivileged groups with unique and targeted messages that might motivate them to vote, can be equally deployed to identify, based on data, who is important and who is not—effectively silencing rather than amplifying individual voices.

In many ways, Big Data is the technology of neoliberalism. It commodifies social data, making sociality amenable to the marketplace. When these tools are then deployed in political contexts, the political sphere adapts market rather than civic ideals. The same demographic profiling that allows companies to reach their target markets, also creates efficient gaps in political exposure and information. Of course, even in a marketing context, consumers can hardly be said to be the substrate of Big Data, because through Big Data, people become disembodied, impersonal, fungible data. Through the interpretive lens of Big Data electioneering, at least in its current formulation, individuals are not voters, constituents, or citizens; they are instead conglomerates of manipulatable data points inscribed with partisan markers.

This is not an inevitable state of affairs. Politicians have the power to use these powerful technologies in accordance with civic virtues, but it requires a more critical eye toward the technologies themselves—understanding that technologies import their own hidden sets of values, which need to be actively interrogated and overcome—as well as the laws and policies that, in their current state, fail to promote ethical and democratic electioneering.

Chapter Three

Mapping Your (Genetic) Identity: 23andMe's Personal Genome Services and Sales

In October 2018, Senator Elizabeth Warren publicly shared the results of her DNA ancestry test with the hopes of silencing decades of criticism (most recently and vocally, Trump's derision of Warren by the epithet "Pocahontas") about her alleged Cherokee heritage ("ElizabethForMA"). Despite Warren's intentions to support tribal interests and boost national awareness of violence against Native Americans, the use of DNA testing to 'prove' her own Native American heritage had risky implications for tribal sovereignty. While the use of genetic testing to parse and quantify heritage is now a widely accepted practice in white America, the logics of DNA ancestry testing run counter to tribal understandings of identity as cultural, rather than simply biological. Cherokee Nation Secretary of State Chuck Hoskin Jr. spoke out against Warren's test and called attention to the dangers that the careless handling of these tests poses for sovereign nations:

> Sovereign tribal nations set their own legal requirements for citizenship, and while DNA tests can be used to determine lineage, such as paternity to an individual, it is not evidence for tribal affiliation. Using a DNA test to lay claim to any connection to the Cherokee Nation or any tribal nation, even vaguely, is inappropriate and wrong. It makes a mockery out of DNA tests and its legitimate uses while also dishonoring legitimate tribal governments and their citizens, whose ancestors are well documented and whose heritage is proven. ("Cherokee Nation")

While Warren did not use her DNA test results to seek any personal advantages beyond the political gain from putting a controversy to rest, her decision to broadcast the results publicly

gave political weight to genetic testing as a valid method for determining a person's ancestral identity, thus disempowering Native American tribes who have collectively refused to contribute their genetic data to public databases because of the risk that genomics poses for Indigenous rights. When Warren released her results, she assumed (wrongly) that they would be objective and speak for themselves, justifying her claim to authentic Cherokee heritage. Instead, the perceived low percentage of indigenous heritage in the senator's genetic composition score was the source of much mockery and derision among her political opponents. Senator Lindsey Graham told Fox News hosts that he was going to take his own test and try to "beat" her results (presumably by having a greater percentage of indigenous ancestry in his own report), despite the fact that he—unlike Warren—did not identify as having indigenous ancestry (Strange). While this media stunt was supposed to prove that Warren could not identify as Native American, it also defended the dominant (white American) narratives that genetic tests not only can, but should, act as an unassailable basis of cultural identification (TallBear). In the hands of both Elizabeth Warren and her opponents, genetic tests became weapons for ideological assertion, effectively granting white people the power to arbitrate authentic indigenous ancestry regardless of how tribes and tribal members themselves determine kinship and tribal membership.

Though Warren's ancestry test is just one (albeit very public) example of the cultural authority granted to genetic testing to adjudicate with objectivity the sticky matters of race and identity, it speaks to the much broader potential of these technologies to subtly or overtly shape societal notions about who "we" are. This authority to reshape the meaning of cultural markers of identity should not be understood as a given, but an active construction that emerges from conscious decisions about how to advertise and apply new genetic technologies, especially when

those technologies interact with the kinds of problematic analytic algorithms common to Big Data applications.

The Datification of Personal Genetics

In the wake of such high-profile and publicly funded genetic testing projects like the Human Genome Project and HapMap in the late 1990s and early 2000s, over the past decade, genetic testing services have become cheaper and more readily available thanks to the emergence of private direct to consumer genetic testing (DTCGT) companies. As genetic testing kits entered the public marketplace, so too did new possibilities for using data generated from genetic material to make meaning.

This chapter will examine the rhetorical and material effects of Big Data in personal genomics, focusing specifically on the DTCGT services offered by the privately-held biotechnology and genomics company 23andMe. In comparing the way that 23andMe frames its services to consumers with the way 23andMe's services actually work, I show how the company's marketing of individual empowerment and scientific progress functions to obscure their Big Data research biases and non-representative sampling practices. Blackboxing their methods ensures that 23andMe's success as a company is not contingent upon the accuracy of their results or the quality of their scientific methodologies. Rather, 23andMe's success is dependent on their capacity to exploit the cultural and social value of genetic material to persuade consumers to recognize themselves in the genetic information they are given. 23andMe's efforts toward building public trust around their DTCGT services have been wildly successful because their services are sold to consumers for their alleged cultural import. The results of their genetic tests are explicitly advertised as the cornerstones to one's own identity,

mediated through heritage as excavated by biotechnology while promising, in the redefinition of history and cultural memory, to uncover a kind of genetic destiny—to inform us of who we will become (Stevens). In other words, 23andMe has been successful, not just at selling genetic tests, but at selling access to a deeper understanding of the self, both individually and in the context of a larger, global network of genetically-interconnected bodies. But this illusion of access to the self covers over the company's manipulative and, in some cases, outright harmful data practices.

First, I explore 23andMe's advertising and branding to showcase how the company uses personal appeals to build customer trust. Then, I demonstrate how the company's commitment to enrolling its customers in genetic research yields misleading claims, through which, however, we can gain insight into the company's plans to mobilize their data in the future. Though enrolling consumers (and gaining their trust) is clearly essential to the company's business model, providing customers with an accurate or valuable product is not a major priority. To demonstrate this, I compare the process through which 23andMe harvests and encodes users' genetic data with the content of their privacy policy and research consent forms. I argue that these documents highlight a critical conflict between the company's stated aims of connecting users with valuable self-knowledge and the company's broader profit motives as a data broker. Finally, I use the FDA's 2012 ban on 23andMe's medical reports to highlight the serious risks these reports have for patients and the alleged "empowerment," the company says they provide. Despite having limited FDA approval today, I argue that the risks of these reports are as great as ever.

In the penultimate section, I offer a case study of the company's genetic ancestry analysis feature: specifically on disparities in the racial makeup of 23andMe's reference population database. By exploring the myriad limitations of 23andMe's services across the racialized genetic data-divide, I demonstrate how 23andMe's Big Data methodologies both rely on and

perpetuate Western logics that locate genetic data as the centerpiece of both individual and collective identity. Ultimately, I argue that, left unaccounted for, 23andMe and other personal genetic testing services perpetuate distorted understandings of genomic science that, in turn, work to sustain white (narratives of) superiority—including outright white supremacy.

23andMe: 'Welcome to You'

Founded in 2006, 23andMe markets personal genetic testing services at affordable prices, fanning the flames of an already intense cultural fascination with the gene (Nelkin and Susan Lindee). With a corporate mission of advancing scientific knowledge of human heritage for the greater public good, 23andMe promises to offer consumers unprecedented access to their own genomes for a modest fee without requiring the involvement of medical professionals, genetic counselors, or insurance providers. By acquiring and comparing genetic samples from millions of customers, 23andMe also promises to help speed up the discovery process in genetics, and thus play a part in revolutionizing understandings of biological and ancestral identity. Over the past thirteen years, 23andMe has gained more than ten million customers and its database of genetic and self-reported phenotypic data is the largest of its kind in the world ("About Us - 23andMe Media Center").

23andMe's genetic testing services are pitched to potential customers as a relatively simple process. In exchange for $99 and a test tube full of saliva, 23andMe will provide you with information about your ancestry. For an additional $100, customers can receive health and wellness reports, which include information about health conditions, physical and sensory traits, physical wellness, and possible carrier status for certain diseases and medical conditions. These

prices are not necessarily standard, however, as the company provides an abundance of discounts for both service packages.

When customers receive or buy the kit, it comes neatly packaged in a small white box that is decorated with brightly colored graphical representations of the 23 chromosomes from which the company has taken its name. Once the kit is opened, 23andMe immediately confronts its consumers with charged rhetorics of individuality and personalization: the topmost welcome card contains large gray text that greets the customer with the company's simple (trademarked) phrase, "Welcome to you" (*DNA Genetic Testing & Analysis - 23andMe*). Already in this opening exchange, 23andMe has inserted itself as a gatekeeper between "you" as a customer, and the true "you" as revealed by your genetics. In welcoming customers to themselves, the company quite literally invites its customers to identify with the test results the company provides. These messages are repeated frequently before, during, and after DNA testing. There are two major reasons for this: 1) the appeal to the personal builds trust and primes customers to believe in the results of the test; and 2) situating themselves as purveyors and arbiters of self-knowledge encourages customers to identify with their results and take a more active investment in their research program. It is worth mentioning that these tactics toward personalization strongly resemble those appeals made by other Big Data firms (e.g. in Facebook's advertising services).

Aside from the aforementioned tagline, the kit itself is rather innocuous. It includes a green greeting card (the front of the card reads, "Hi. Let's get started"; inside is the 23andMe hashtag and Twitter handle, customer care contact information, and large text that reads "We are excited for you to begin this journey"), step-by-step instructions for the saliva collection process, a small plastic saliva collection tube, and a pre-paid mailing label addressed to 23andMe's third-party laboratory. Nowhere inside this kit are the company's data-driven research agenda, the

privacy policy, the caveats to the medical information they will eventually provide, or the limitations of the accuracy of their sampling techniques outlined for their customers. For customers who purchased the kit for themselves, they may be familiar with the company's product. However, because 23andMe has marketed its product as an ideal gift to give family members (and has held special pricing events explicitly for gift-giving and/or family-centric holidays like Mother's Day, Father's Day, Christmas, and Thanksgiving), there is an increased likelihood that many people who are using 23andMe's services did not purchase the kit themselves (Schencker). In the case of the genetic testing gift recipient, they may have very little to no knowledge of the company's practices and the efficacy of their genetic testing services. And because the testing kit itself contains minimal information about the company's practices, it is possible that the first time these customers encounter any information about the company's practices is when they are agreeing to the privacy policy and research consent forms that are required to gain access to the test results. However, by this time, they may have already submitted their sample and been promised access to self-knowledge, and thus been thoroughly incentivized to agree to these documents.

Importantly, each kit must be registered before the laboratory will begin processing the DNA sample. To register their kit, customers are asked to disclose personally identifying information, including legal names, contact information, and self-reported ethnicity (which I will discuss later in the chapter). They are also presented with the company's privacy policy, agreement to which is required for all consumers, as well as multiple 'optional' research consent forms. The research consent forms are presented as opportunities for customers to take an active role in shaping the future of genetic research: "23andMe was founded to empower individuals and accelerate research. As a 23andMe customer, you are a partner in this mission. You have the

opportunity to participate in genetic research, which could contribute to revolutionary findings in human diseases, conditions, and traits" ("About Us"). From the customer's perspective, 23andMe's central goal is to give individuals access to their personal genetic information and to help them help themselves by giving them a way to participate in research as citizen scientists.

The main portal through which customers are able to participate in research is through the "Research" feature on the 23andMe website, which consists of three separate subsections: "Questions," "Insights," and "Publications" The "Questions" subsection presents users with numerous surveys and questions about health, lifestyle, wellness, diet, and others. Essentially, 23andMe uses these surveys and questions to collect self-reported data about health, environment, and lifestyle, which they then aggregate with individuals' genetic data. To make meaning from this aggregated data, 23andMe's researchers employ explanatory analytics: using algorithms that locate patterns between shared survey answers and their corresponding genomes to identify potentially genetically influenced traits. Responding to these survey questions is a prerequisite for customers to visit the "Insights" portion of the website, which provides users with "interesting data and findings" from 23andMe's research. Essentially, the company withholds research insights from the individual consumer until certain participation demands have been met.

The Research Insights results are among the most transparently misleading elements of the entire 23andMe experience. Here are some examples of "insights" that 23andMe customers might receive:

- "Based on your genetics and other factors, you are more likely to prefer chocolate ice cream over vanilla ice cream."
- "Based on your genetics and other factors, you are more likely to have stretch marks."

- "People with your genetics in their 20s wake up on average around 8:57 am on their days off."
- "Based on your genetics and other factors, you have about equal chances of being able or unable to match a musical pitch."
- "Based on your genetic makeup, you are 67% more likely to go skydiving than other 23andMe customers." (23andMe)

Some of these are clearly intended as entertaining bits of genetic trivia. By listing "skydiving" as an activity one can have encoded in one's genetic code despite the obvious lack of planes in the history of human development, we can only assume this is a tongue-in-cheek gesture to encode a more abstract trait (but one highly valuable to insurance companies): the degree to which one's genetic profile matches those of other risk-takers, and thus the likelihood that "you" will engage in risky behavior. This could more easily be dismissed as a joke were it not for the fact that these more stylized behavioral results are situated in between more direct genetic examples like hair and eye color, thus diluting the fact (for customers who lack genetic literacy) that these are not equivalent predictions in terms of their scientific footing. By juxtaposing the skydiving example with the more obviously true claim 'you have brown eyes,' 23andMe encourages customers to see truth in their reports. Accordingly, these reports place limits around how individuals can engage with their genetic test results, positioning them to think and act in particular ways depending on their "unique" genetic disposition.

As another example, consider 23andMe's new 'Fear of Public Speaking' report. "Are you dreading having to make a speech this wedding season? View your Fear of Public Speaking report to see if your DNA may be partly to blame." Though this report is framed around public speaking, it is allegedly a measure of fight or flight response. With the other personalized,

targeted rhetoric the company deploys, it is clear that 23andMe wants its customers to identify with at least some of these claims (which operate like horoscopes), despite there being shaky evidence (and plenty of ethical problems) with assigning genetic bases for behavioral traits (especially highly contextualized and culture-specific ones).

The third tab on the Research section of the customer portal, the "Publications" section, includes links to 23andMe's research publications. At the top of the publications section is a banner saying how many publications the user has contributed to—in other words, how many publications have used data drawn from that user's data. When the company has published new research in scientific journals, 23andMe customers receive updates which work to inflate the user's perception of 23andMe's credibility and the importance of their contribution to scientific knowledge. It is important to note that when customers receive updates about these new publications, they are directed to a 23andMe blog post that summarizes the research. These summaries often leave out key information about research methodology and study limitations. 23andMe explains the purpose of its customer-facing research portal as follows:

> By answering online survey questions and allowing researchers to combine your genetic data with millions of other data points, you can help drive scientific and medical discoveries. At regular intervals, you will be presented with research insights. These may show how your responses compare to other 23andMe users or may showcase 23andMe discoveries made possible thanks to the contribution of customers like you. (23andMe Customer Care)

> Milestones help mark your progress through the 23andMe Research experience, giving us an opportunity to tell you about the impact you are having and deliver interesting insights…Insights are a way for us to share interesting data and findings from 23andMe with you. These can be early discoveries 23andMe has made, or more background on the genetics of a particular topic or trait. You earn an Insight every time you hit a milestone. Keep in mind that they are preliminary and are meant for informational purposes only. (23andMe Customer Care)

Essentially, to incentivize users to provide more personal data, 23andMe tells consumers that, by completing surveys and answering questions, they are making valuable contributions to medical and scientific research for which they will be rewarded.

In their article "The Gift of Spit and the Obligation to Return it, Harris et al argue that by "present[ing] research participation as a form of gift exchange," 23andMe implies a "social bond" between the company and consumer (Harris, Wyatt, et al. 236). This bond persuades consumers to equate the value of themselves gaining access to new genetic information with the value of 23andMe gaining access to their data (Saukko). By framing their operation as an exchange through which the consumer is getting the better end of the deal, 23andMe persuades consumers away from questioning the profitability of the company's data-driven research (Harris, Wyatt, et al.). Building on this scholarship, I contend that, by presenting their services as a type of participatory research platform through which consumers can take an active role in genetic research, 23andMe has been able to masquerade as a civic-minded mediator of groundbreaking medical research, while subtly reaping enormous profits off of the harvested data of those individuals they claim to serve.

Although 23andMe claims that they see consumers as "research partners," the people who have consented to having their genetic data used for research purposes have little agency in determining the company's research agenda (*Research - 23andMe*). Additionally, because of the conditions of the research consent document, consumers are agreeing to waive their legal rights to push back against 23andMe's future research agendas, even if these agendas conflict with earlier claims made by the corporation. By presenting customers with a privacy policy as well as multiple research consent forms, the company makes research participation seem separate from customer's gaining access to their genetic information. Yet, the privacy policy specifically stipulates that by agreeing to the privacy policy, users are also agreeing to their 'Aggregated Information' (which includes genetic information that has been stripped of personal identifiers) being shared with third party service providers (*DNA Genetic Testing & Analysis - 23andMe*). While this stipulation is alluded to in many different sections of the privacy policy, it is perhaps most clearly articulated in the section titled "What happens if you do NOT consent to 23andMe Research?" which reads as follows:

> If you choose not to complete a Consent Document or any additional agreement with 23andMe, your Personal Information will not be used for 23andMe Research. However, your Genetic Information and Self-Reported Information may still be used by us and shared with our third party service providers to as outlined in this Privacy Statement. (*DNA Genetic Testing & Analysis - 23andMe*)

In other words, even if customers do not agree to the research consent document, their genetic data and self-reported data can still be used for 'research' purposes, albeit research conducted by 'third-party service providers,' and not necessarily for the purposes of peer-reviewed publication.[iii] By making customer access to genetic information contingent upon their

agreement to the privacy policy, and by situating genetic information as a part of the self, 23andMe prevents its customers from gaining self-knowledge until they agree to the privacy policy, further asserting the company as the gatekeeper to genetic identity. Regardless, the majority of customers agree to both the privacy policy and the multiple research consent forms: according to the company's website, around 85% of customers have consented to having their data used for research purposes, making 23andMe one of the most powerful players in genetic research ("23andMe for Scientists").

23andMe's promise (of accelerating research and empowering individuals) is threaded through every aspect of how 23andMe presents its genetic testing services to customers, to the point that it is almost impossible to see past the promise and understand how 23andMe's testing actually works. To help alleviate this gap, the next section will focus on unpacking how 23andMe's genetic testing services work, starting with an explanation of how 23andMe's scientists transform genetic material into genetic data, and how that genetic data is translated into genetic information for the consumer.

Genetic Test Results: From SNP to Self

Within the realm of genomic science, a number of different methods are employed to garner meaning from genetic material. Each human being has roughly three billion base pairs (letters in the genetic alphabet) in their genome. However, because sequencing all of these pairs can be expensive and because many large portions of the genome are non-coding, which is to say that they are not used by the body to produce proteins and so yield few interesting results, genetic scientists often look only at portions of the genetic record. To maximize the efficiency of the direct-to-consumer product, producing the greatest number of interesting results for the

lowest cost, DTCGT companies typically limit their analysis to particular sites within the genome where genetic variations are known to have occurred. To this end, 23andMe uses a method of genetic sequencing called single nucleotide polymorphism, or SNP, genotyping.

Unlike Whole Genome Sequencing, which investigates the entire genome, or even Whole Exome Sequencing, which analyzes those portions of genes that encode proteins (which contain many known disease variants—about 1% of the entire genome), SNP genotyping examines just 0.1% of the genome, targeting specific sites within the genome where genetic loci have been statistically and computationally linked to particular phenotypic traits and diseases (Stevens). Using a DNA chip, or microarray, genetic scientists compare discrete single nucleotide polymorphisms (SNPs) from one individual's DNA to SNPs from a reference database of previously collected and analyzed genomes called a r*eference population*. Sophisticated algorithms are then deployed to detect common genetic variants associated with certain diseases, conditions and/or traits. We can think of a SNP as a decontextualized bit of biological matter with specific meaning in the body, a kind of "bio-atom," denatured and data-fied to make it amenable for comparison with other such decontextualized bits drawn from the reference population data. Once analyzed, customers' saliva is stored in 23andMe's biobank; the DNA that was extracted from the saliva sample is then analyzed, and the resultant genetic data is stored in the company's genetic database.

The meaning that 23andMe draws from these SNP comparisons changes according to the purposes of the analysis. When constructing a customer's health report, 23andMe uses scientific assumptions about the meaning of particular genetic sites and their variations. For the ancestry report, they use assumptions about mutation speeds over large timescales between groups and assumptions about migration patterns that rely on Western narratives of geographic and temporal

belonging (discussed in more detail later). And for the traits report, they use internal databases of other 23andMe customers' self-reported survey data. Despite having different purposes, all of these different meaning making strategies rely on the same centralized repository of data. The contents of this centralized database extend far beyond customer genetic data. To supplement their internal biobank and genetic databases, 23andMe mines data from population biobanks, cohort studies, and genome databases; clinical and public health records; and they collect data through social media, fitness trackers, health apps, and biometric data sensors. Finally, and perhaps most unexpectedly, 23andMe uses cookies to track customers' browsing activity before and after visiting the customer portal website. Additionally, 23andMe recently introduced a centralized 'health hub' for customers to upload and link their personal medical records, lab results, and prescription information to their 23andMe profile (*Research Studies - 23andMe*).

"Lost" in Translation: The Limits of 23andMe's Genetic Reports

When 23andMe customers receive their genetic test reports, the degree to which a SNP and its comparability is an artifact of biotechnological procedures is largely erased (Stevens; Pálsson). In her research on how DTCGT services translate genetic test results for consumers, anthropologist Minna Ruckenstein explains, "23andMe inscribes each participant with a genetic identity based on numbers, calculations, graphs, and charts. Users have no access to how data about them are analyzed in order to evaluate its validity; they are merely empowered by a non-transparent logic of numbers to assist their quest for self-knowledge" (Ruckenstein 1027). To compensate for this apparent evidentiary opacity, 23andMe's genetic reports are ripe with visually appealing and "personalized" data visualizations that purport to translate complex genetic information for a non-expert audience.

As customers wade through health and lifestyle advice, medical recommendations, and trait predictions, along with copious trivialities ("Based on your genetic makeup, you are 67% more likely to go skydiving than other 23andMe customers."), they are repeatedly presented with calls to engage with their genetic reports on a physical, emotional, sensorial, and familial level. For instance, 23andMe's Genetic Health report and Carrier Status report adopt rhetorics of medical advocacy, stressing individual empowerment, while the Traits report takes a more playful approach to individualization, emphasizing the genetic quirks that make you, you. This focus on individuality is accompanied by rhetorics of belonging, connection, and sharing, including aggressive banner advertisements encouraging users to purchase genetic testing kits for family members, so they can "share and compare" their 23andMe results with each other. Through the lens of 23andMe's data mining practices, we can see that, beyond building trust by word of mouth referrals, by encouraging customers to gift kits to family members, 23andMe is asking customers to aid in their acquisition of more genetic material and thus help enrich the company's data holdings. Without evidence of the data infrastructures and algorithmic processes through which 23andMe's genetic testing services depend and operate, 23andMe's genetic test results and reports appear as artifacts of the genetic self that have been unproblematically traced and interpreted by neutral (and relatively infallible) scientific and computational mediators.

The reason that 23andMe is able to provide results without explaining their data or methodologies is largely a product of the company's privatized status. Unlike publicly funded genetic research, which requires scientists to submit data to a public database, 23andMe's privatized status means it is not required to submit the genetic data it collects to an open access repository. 23andMe has further obscured their methodologies by claiming that, because they are dealing with sensitive data (genetic, consumer, personally identifying data), it would be

impossible for the company to both submit their data for review while also keeping that personal data secure and maintaining customer privacy. Though the company insists that genetic profiles are de-identified when data is used for 'research' and sent to third party analysts, the fact that this data is genetic, and therefore, uniquely identifiable means this guarantee of anonymity is little more than a ruse. With no possibility of true anonymity and no real mechanism to opt out of 23andMe's Big Data research agenda once customer samples have been submitted, privacy—in any meaningful sense—is impossible. Under the company's guidelines then, privacy essentially becomes a powerful mechanism to obscure the realization of 23andMe's promises of empowerment.

By projecting an ethos of cooperation in the spirit of progress, the company persuades its customers that the research to which their data is contributing is socially beneficial. It is possible, even likely, however, that these research projects may eventually be used in ways that harm people, whether the customer themselves, their families, or some other class of people entirely. Take for instance the recent capture of the Golden State Killer using personal-genetic testing data. Law enforcement officials were able to upload DNA samples taken from an old crime scene onto an online genetic database called GEDmatch (a website where individuals can link their genetic test results from different testing companies, including 23andMe, to find relatives). Using this service, they were able to identify matches in the databases, which linked the DNA sample to similar DNA from other profiles (i.e. potential relatives). These matches allowed investigators to limit the number of suspects in the area where (and when) the crime was committed to the point where they eventually identified a suspect. While the capture of a high-profile, violent criminal seems like a boon to forensic genetic science, its applicability to other contexts (e.g. predictive policing) is troubling. Additionally, the example of the Golden State

Killer draws attention to the porous nature of genetic data and illuminates the inconsistencies between 23andMe's promise of protecting individual-level genetic data and their practice of using traditional structures of consent (e.g. individual-level consumer privacy policies). Because genes are not only representative of individuals, but are shared between entire families, nations, and/or ethnic groups, it is impractical, if not entirely impossible, for all affected parties to give full consent, and is thus irresponsible for 23andMe to claim that their services will help protect persons genetic privacy. In a *New York Times* article about the case, a law professor at NYU was quoted as having said, "Suppose you are worried about genetic privacy...If your sibling or parent or child engaged in this activity online, they are compromising your family for generations" (Kolata and Murphy).

We will see the damages of this dilemma for 23andMe's ancestry report later, but for now, I want to briefly pivot to one final weakness in their model for helping people understand their bodies in terms of medical risk.

The Dangers of False Results and Claims of Patient Empowerment

It should be clear now that 23andMe's two major promises, to 'accelerate research' and to 'empower individuals,' are mutually incompatible under the company's current business model. Perhaps the best example of 23andMe's failure to achieve both of these goals at once is in the context of the company's medical claims. In their examination of the promise of personal genomic medicine, Juengst et al explain the inability of "empowerment rhetoric" employed by DTCGT service providers like 23andMe to frame the ground-breaking possibilities that personalized genomic medicine affords, not only for the benefit of individual patients but also for increasing public good: "What is revolutionary about this kind of medicine, its advocates

maintain, is that it promises to resolve [the healthcare] crisis by simultaneously increasing the ability to be 'personalized,' 'predictive,' 'preventive,' and 'participatory'" (Juengst et al. 2). They argue that these claims jar with the realities of genetic science, namely because the benefits accompanying genetic information can only be accessed and utilized by people who are making medical decisions (e.g. doctors, care givers, medical providers). Further, the authors claim that, in terms of health care decision making, genomic science provides two types of valuable information: "(1) pharmacogenomic information about a patient's chances of responding well or poorly to a therapeutic regimen, and (2) genomic susceptibility information about a patient's chances of resisting or succumbing to other environmental or degenerative health threats" (Juengst et al. 5) These benefits, the authors argue, do not clearly map onto the patient's role, and that "neither kind of information obviously has a role in empowering patients that is analogous to its role in understanding, preventing, or treating disease" (Juengst et al. 4). However, when individuals are presented with the promise of empowerment, they might not see the limits of this empowerment and might not understand with what information they are being empowered. Whether DTCGT providers' promises (that personalized genomic medicine will empower individuals) are actually fulfilled, the authors argue, seems a moot point for providers (Juengst et al.).

Looking back on the history of the company offers a bit more perspective on the gaps between what 23andMe promises and what it provides, and the risks of false empowerment that arise when customers do not understand their genetic information. When 23andMe initially began offering personal genetic testing services in 2008, they provided consumers with data for hundreds of genetic traits and dispositions. As their customer base grew, the company was met with growing concerns from regulatory agencies about perceived gaps in the regulation and

quality control of their personal genome services (Anderson). On November 13, 2013, five years after 23andMe had began selling their services, the US Food and Drug Administration (FDA) issued a warning letter to Anne Wojcicki, 23andMe's CEO, explaining that because the company's Saliva Collection Kit and Personal Genome Service were "intended for use in the diagnosis of disease or other conditions or in the cure, mitigation, treatment, or prevention of disease, or is intended to affect the structure or function of the body," 23andMe was legally required to get FDA approval prior to marketing their product in the United States (FDA, "Warning Letter from the FDA to Anne Wojcicki,").

In addition to citing the fact that 23andMe had been marketing their product for five years already and had not taken the appropriate steps to get regulatory approval, the FDA pointed to the potential consequences that accompanied gaps in regulation and control. Specifically, they pointed to the potentially negative consequences arising from consumers using the DTCGT results to make decisions about their health. With no certified medical professional required to interpret and digest genetic test results (such as a genetic counselor), the FDA explained, consumers may be motivated to seek unnecessary treatment or discontinue medically-prescribed treatment. Throughout the letter, the FDA emphasized concerns about potential psychosocial harms of false-positives and false-negatives.

In addition to revealing regulatory concerns about 23andMe's product, the letter gives us a sense of the tenor of the FDA's ongoing relationship with 23andMe. One portion of the letter in particular raises red flags about the company's attitude towards public health, and I believe it is worth quoting here at length:

> As part of our interactions with you, including more than 14 face-to-face and teleconference meetings, hundreds of email exchanges, and dozens of written communications, we provided you with specific feedback on study protocols and clinical and analytical validation requirements, discussed potential classifications and regulatory pathways (including reasonable submission timelines), provided statistical advice, and discussed potential risk mitigation strategies. As discussed above, FDA is concerned about the public health consequences of inaccurate results from the PGS device; the main purpose of compliance with FDA's regulatory requirements is to ensure that the tests work. However, even after these many interactions with 23andMe, we still do not have any assurance that the firm has analytically or clinically validated the PGS for its intended uses, which have expanded from the uses that the firm identified in its submissions… You have not worked with us toward de novo classification, did not provide the additional information we requested...and FDA has not received any communication from 23andMe since May. Instead, we have become aware that you have initiated new marketing campaigns, including television commercials that, together with an increasing list of indications, show that you plan to expand the PGS's uses and consumer base without obtaining marketing authorization from FDA. (FDA, "Warning Letter from the FDA to Anne Wojcicki")

With this letter, the FDA banned 23andMe from marketing their genetic health services in the US. In the years following the FDA ban, however, 23andMe continued to provide its customers with non-health related genetic information, such as information about ancestry and physical/sensory features (even though false results about ancestry can also have dramatic psychosocial impacts). Further, to make up for the business they lost in the US, in 2014,

23andMe began to pursue research locations outside of the United States and has since sold testing kits to customers in Canada and the UK. It is important to note that even though 23andMe was not able to market or provide health related genetic reports while the FDA's ban was in place, they were still able to collect saliva samples from US consumers wanting to know more about their ancestry. And because it is impossible to simply parse out the portions of DNA related to ancestry, 23andMe continued to receive whole DNA samples from consumers. Thus, 23andMe was able to continue Big Data-driven medical research during the entirety of the FDA's ban.

In the years following the FDA ban, 23andMe began conducting "user comprehension testing" with the hopes of alleviating public and regulatory concerns that customers were at risk of misinterpreting and/or misunderstanding their genetic test results (Yuji et al.). In April 2017, the FDA partially lifted their ban, approving 23andMe's marketing of direct-to-consumer tests for ten different genetic diseases and conditions (FDA, *FDA Allows Marketing of First Direct-to-Consumer Tests That Provide Genetic Risk Information for Certain Conditions*).[iv] As justification for reversing their stance on 23andMe's marketing of genetic testing kits, the FDA explained, "A user study showed that the 23andMe GHR tests' instructions and reports were easy to follow and understand. The study indicated that people using the tests understood more than 90 percent of the information presented in the reports." However, proving that their customers *understand* their genetic test results does not mean that 23andMe's genetic tests pose any less risks for customers than they did when the FDA initially banned their product. Revealingly, the FDA approval also contained the following warning: "Risks associated with use of the 23andMe GHR tests include false positive findings, which can occur when a person receives a result indicating incorrectly that he or she has a certain genetic variant, and false

negative findings that can occur when a user receives a result indicating incorrectly that he or she does not have a certain genetic variant. Results obtained from the tests should not be used for diagnosis or to inform treatment decisions. Users should consult a health care professional with questions or concerns about results" (FDA). And while the 23andMe website now includes more stipulations about how customers should approach their 23andMe genetic test results, these warnings (which are often presented in superscript) are easily overshadowed by the company's appeals to the personal and the continued emphasis on self-knowledge (i.e. 'Welcome to You').

Despite their promises of using genetic science to accelerate research and empower individuals, it is vital to remember that, as a corporation, 23andMe is, fundamentally, a data-driven venture. The actions and attitudes of 23andMe as described in the FDA's warning letter and as demonstrated in the company's activities following the FDA's ban shed light on the motivations undergirding 23andMe's services. While 23andMe and other personal genetic testing services appeal to potential consumers by situating genetic materiality as the centerpiece of personal identity, their profit- and big data-driven research agendas reveal individuals as conglomerates of raw genetic data available to be mined.

The ethereal promise of Big Data coupled with the neoliberal drive toward ultimate personalization makes 23andMe's genetic reports powerful rhetorical tools for persuading people to begin to understand their bodies in a new, quantifiable light. However, in her book *Bodies in Flux*, Christa Teston reminds us that the meaning of genetic data is always "in flux." As the technologies and methods for doing genetic science undergo change, and as new genetic data is made available for reference and/or comparison, alternative possibilities for creating meaning from genetic data continually emerge. Genetic flux, however, runs counter to the logics of

patient empowerment and identity at the heart of 23andMe's DTCGT brand, which posit genetic information as material evidence of the self.

The cultural narrative that genetic information is a kind of human blueprint portrays DNA as a static "book of life," full of revealed truths about the self, not just the body (Kay). Yet this cultural idea runs counter to 23andMe's genetic test results (and indeed much of genetic science), which are continually fluctuating, updating, and changing as the database upon which each individual's results are calculated and calibrated grows. There is thus a mismatch between the fluid nature of the database and the objective-seeming reality the company creates, because while the genetics are in flux, the identification of the self with genetic test results (which 23andMe encourages) is largely not. It defies the logic of the company's own branded marketing that one's ancestry composition can substantially change, yet this occurs regularly to the chagrin of confused customers (Gates). The fluidity and impermanence of the company's conclusions is not sufficiently apparent to its customers when the company distributes its results; it is only by repeatedly logging in and witnessing the change in one's personal results over time that this intrinsic feature of 23andMe is evident at all. By pitching their services as an opportunity to "Live in the Know" without accounting for the limits of genomic knowledge, 23andMe actively fosters public misunderstandings of genomic science. Further, because 23andMe's success is contingent upon the myth of genetic determination, it matters little whether or not 23andMe's promises hold up once customers receive their genetic test results, because the company has primed its customers (who may already be so inclined) to read their results through a personalized lens.

Thus far, our discussion of 23andMe has been confined to the company's marketing tactics, methods, and regulatory relationships. The dangers I have outlined in their project pertain

mostly to the company's deceptive data-mining practices and the potential for their research reports to mislead. These dangers have become particularly potent because of the cultural relationship people have (and that 23andMe encourages them to have) with the gene as a marker for identity. However, as the opening narrative about Senator Elizabeth Warren demonstrated, not all 23andMe customers enter the database equally. Reading through 23andMe's publications reveals a not-so-surprising research bias, namely that the majority of their research only uses data collected from participants of European descent.

Racial Imbalances in 23andMe's Reference Population Datasets

While 23andMe markets their genetic tests as being for everyone, their reference population data and customer databases are woefully misrepresentative of global diversity ("Reference Populations"). While there are a number of larger, systemic social and political factors that contribute to skewed racial representation in 23andMe's database—including minority groups' historically-rooted mistrust of medical professionals and fears that genetic data in particular could be exploited for racialized discrimination and harm—there are also significant internal barriers to access that have been built into the genomic and algorithmic processes undergirding 23andMe's operation (Lee). Perhaps the largest systemic barrier is the way that 23andMe's reference populations are defined. Participation in the reference population is restricted to include only those individuals who can "prove" their heritage, a process which requires having documentation showing that all four biological grandparents were all born in the same country or geographic area. By restricting their reference populations to individuals with "proven" or "provable" ancestral pedigrees, 23andMe systematically disadvantages people of color, whose families are vastly more likely to have been affected by the historical, lineage-

disrupting forces of slavery and colonialism. By restricting participation in the reference population to certain individuals with homogeneous (but in some cases relatively recent) heritage, 23andMe constructs distorted genetic representations of particular regional populations under the implausible assumption that the modern boundaries of certain (especially European) geographic regions imply a historical purity of heritage.

Gaps in 23andMe's reference population data has ripple effects that surge through the rest of the utility of 23andMe's services since the accuracy of genomic information and the possible benefits of genomic research are restricted to those groups with robust reference populations: mainly European countries. The lack of data from and research on persons of non-European ethnicity means that the quality of service provided for white, European customers is not shared with its non-white/non-European customers. All but two of the company's seven genetic health reports and all but two of the 40 carrier status reports are labeled as being relevant *exclusively* for people of European descent. Further, because the wellness and traits reports are dependent on data from other 23andMe customers and because 23andMe's customer base is heavily skewed toward white individuals of European descent, those few results that do apply to non-white/non-European populations are less accurate than those provided to their white counterparts. As explained in the previous section, inaccurate genetic test results can have serious consequences; people have been known to take preventive action according to the medical information provided by 23andMe's services, and with a systematic accuracy bias favoring white populations, their genetic tests risk unnecessarily and disproportionately jeopardizing the health and wellbeing of populations of color. Further, because they draw from the same centralized databases, many of the same limitations that make the company's personalized genetic reports misleading pertain to the company's research endeavors as well

("Reference Populations").

Despite inequalities in demographic representation, the sheer quantity of data that 23andMe has harvested has exponentially increased the perceived value of the company's databases. Their title as steward of the world's largest genetic database has granted them undue weight in deciphering the content and meaning of global genetic data. The size of 23andMe's database seems to grant the company a sheen of objectivity. Regardless of the systematic flaws in their data collection and analysis mechanisms, the number of samples they have taken has made the company essential to conducting population-level genetic research that could not be accomplished with smaller and/or publicly funded databases.

In the brief explanations that the company offers about the lack of diversity in their databases, there is a striking lack of attention paid to the company's own positionality. Rather than recognize minority communities as having agency in deciding not to share their DNA, or account for the company's problematic marketing strategies and mapping practices both of which exclude non-western narratives and non-western family structures, the company explains that these gaps in demographic representation are simply a result of data deficits: "There are some genetic ancestries that are inherently difficult to tell apart, typically because the people in those regions mixed throughout history or have a shared history, or we might not have had enough data to tell them apart" ("Reference Populations"). And while 23andMe has publicly expressed a commitment to diversifying their reference population databases, the diversity initiatives the company has created present numerous other barriers to access that perpetuate those same structural inequalities the initiatives are supposedly designed to address. For instance, enrollment in 23andMe's African Ancestry stipulates that participants are US citizens over the age of 18 with four grandparents from the same region, are able to "read and write in English,"

and “have access to the internet” (*23andMe Global Genetics Project*). Here and in other initiatives, the burden of access is applied with undue pressure on and labor for minority populations.

The multifaceted discourses of identity, bodies, and data at play in 23andMe’s operation presents us with an opportunity to examine how the company’s appeals to the personal work to cover over the racial unrepresentativeness of their genetic database.

Ancestry for Some: The Datafication of Colonialism and Race

Just as 23andMe centers genetics as a primary mode of identifying personal risk for disease in a way that seeks primacy over other socio-spatial modes of identifying health risks, the company also centers genes as the primary indicator of what they term ‘ancestral identity.’ A core difficulty of employing Big Data to determine ancestry is the fuzziness of the concept of ancestry itself. Ancestry, for 23andMe, is a quantifiable and measurable quality that emerges as a coefficient from comparison of one’s SNPs to their (skewed and incomplete) database. Yet even with perfect representation in the reference population, 23andMe’s system (the rhetorical bundle of genomic data as personal truth) fails to describe what many people actually consider to be their heritage. The definition of ancestry itself is one that is mostly recognizable among white European immigrants, with a narrowly conceived notion of one’s ancestors having prior membership in a previous national community. Not only does this overlook other sociogeographical circumstances like the Atlantic slave trade and tribal membership that defy national boundaries, it also overlooks other culturally specific modes of defining heritage itself (Padawer). In some cultures, for instance, lateral relations among friends, community members, or adoptive kin (as opposed to 'vertical' familial ones) may hold more weight as determinants of

heritage and relatedness than those imposed by hardline genetic materiality. Long before the advent of genetics, families were defined largely by culturally specific notions of kinship, which may include close friends, community or tribe members, adopted children, etc. Even today, the correlation between genetics and heritage is obscured by the many widely available forms of non-genetic parenthood, such as sperm and egg donation or surrogacy, that are otherwise obscured or ruled irrelevant to one's ancestry within the 23andMe system. Not only does this limit the number of people to whom these genetic conclusions are valuable, but it also flattens broader cultural conversation about kinship more generally.

The consequences of 23andMe's whitewashed mechanisms for determining heritage are perhaps most evident in 23andMe's Ancestry Composition report. Labeled a "State-of-the-Art Geographic Ancestry Analysis," this report is presented as a Big Data-driven "story of who you are and how you're connected to populations around the world" (*DNA Ancestry Test*). Customers' ancestry analysis results are delivered in the form of a color-coded map of the world, a pie chart, and table that breaks down ancestry by percentage. These statistics-heavy data visualizations evoke what Jenny Rice explains as a "sense of something coherent, a sense that possibly transcends the individual pieces of datum that are contained within that aesthetic whole" (Rice 29). For Western audiences, the cohesive clarity of the ancestry composition report is made stronger by the use of mapping imagery. 23andMe's algorithm populates their map of global ancestry using SNPs from "proven" genetic reference populations. 23andMe then draws on the correlation between their genetic reference populations' data and reported geographic ancestry to predict and confirm the genetic ancestry of new 23andMe customers. Disintegrating national, cultural boundaries (traditional forms of identity construction) in favor of a disembodied, dislocated, and depersonalized calculus that simultaneously promises to link "us" to "our"

ancestors—people who have never been tested but who are nevertheless represented as ghosts in the algorithm, their genetic contributions filled in by equations. An invitation at the top of the report to "Trace your heritage through the centuries and uncover clues about where your ancestors lived and when," shifts agency away from 23andMe and onto the customer, who, equipped with the output of their DNA test, can now embark on their genealogical investigation. The report is accompanied by the following brief explanation:

> We determine your Ancestry Composition using only the information in your DNA and the DNA of other people with known genetic ancestries (our reference datasets). Your 23andMe reports are always based on your genetics. If your results describe you perfectly, that just goes to show you how much your DNA really tells about you! (*DNA Ancestry Test*)

Bundled as a playful exercise in genealogy, this explanation of 23andMe's Ancestry services covers over the violent realities of colonial migration, while simultaneously dismissing any need for methodological discussion (if these results describe you perfectly, then they must be true, and if your results jar with your perception of yourself, they're still true, because the DNA says so). Agency and responsibility for determining ancestry is attributed to "your genetics." The algorithms and mathematical models are simply neutral mediators, extracting truth from your genetics for you. As Stevens claims, "This is *Homo statisticus*—a human constructed from the statistical residue of his or her genome" (Stevens 221).

Returning to our opening discussion of Elizabeth Warren's DNA testing shows us that there is much at stake in these conceptions of how kinship relations are defined, especially for vulnerable or indigenous populations. Far from revealing ancestral and biological "Truth," these algorithms serve to empower and expand dominate socio-cultural and economic power

structures, imposing an order on heritage and ancestry that universalizes white Western logics of genetic identity and asserts dominant definitions of kinship as biological. The use of Western mathematical models to identify connections between individuals has allowed 23andMe to justify their placing the entire human population under the same definition of kinship. In addition to encouraging users to purchase genetic testing kits for family members so they can "share and compare" their 23andMe results, customers are offered access to "DNA Relatives," 23andMe's online tool for finding and connecting you with "new" genetic relatives. Further, 23andMe collapses historical and social specificity with respect to place of origin by locating 'ancestry' within a narrow window of time, again reflecting a white Western logic of how far kinship extends.

'Gene Talk': Navigating Genetic Test Results

The confidence that customers place in the company is evident in how customers talk about their ancestry reports, a type of discourse that Ruckenstein calls "gene talk" (Ruckenstein). Digital personal narratives of genetic testing, or what Harris et al refer to as "autobiologies," have become increasingly popular (Harris, Kelly, et al.). One of these videos, posted by "Try Guys," features a group of four men (the Try Guys) who have submitted their saliva to 23andMe for analysis (BuzzFeedVideo). In the video, the men are receiving and reacting to their own and to each other's genetic test results. The experience is facilitated by a 23andMe service representative who goes through each man's test results individually.

Before revealing their test results, they are asked to talk through what they expect to find. One of the Try Guys, Zach, says that he expects his results to show that he is "100% an Ashkenazi Jew" and explains that "being Jewish is something that I'm very proud of; I think it

has shaped who I am and who I aspire to be." However, when the 23andMe representative reads his genetic ancestry report to him, she says that he is 88.7% Ashkenazi Jewish. Zach responds, saying "I thought I was full blood," and the representative semi-jokingly says "Well, someone lied to you." The group laughs and then the representative continues breaking down the remaining 11% of Zach's report. After hearing the rest of his report, Zach seems to contradict his earlier explanation of what he had hoped to find, by saying, "Is it weird that I'm, like, disappointed that I'm not more of a mix?" to which another Try Guy, Eugene—the only person of color in the group—asks the 23andMe representative, "Do you always come across white people being disappointed they're not more of a mix?" This question is brushed aside, and the next two Try Guys receive their results (both reports claim that they are over 98% European) (BuzzFeedVideo).

Before the final Try Guy, Eugene, receives his results, he says that he expects his report to show that he is 100% Korean. However, when the representative reads his report, it says that he is 63% Korean, 20% Japanese, and 13.1% Chinese. The gravity of this revelation clearly hits Eugene hard. Looking shocked, he says, "These are not insignificant numbers. These are whole chunks of huge parts of relatives….I'm just shocked. I've always been told I'm Korean." The video does not linger long on Eugene's reaction before cutting to a playful discussion of whether or not the Try Guys are related (they're not). However, Eugene's genetic report is brought up again at the end of the video when, revealingly, we hear one of the Try Guys ask, "Wait does that mean that Eugene can make all of the Japanese and Chinese jokes that he wants and he's totally in the clear?"

Although 23andMe presents ancestry data in geographical terms, these results are very much racially coded. And 23andMe knows this. On the podcast *Recode*, 23andMe's founder

Anne Wojcicki discussed the effect that getting unexpected ancestry results has on how people understand their racial identities:

> One thing I didn't anticipate about 23andMe was how much learning your genetic information changes your sense of identity…it really changes how people view themselves and how they are connected to the world. There are a few small examples of people who say 'Oh, wow, look at me, I'm so white' … But more often than not, you get stories of people who were, sort of, white supremacists, but then they get their DNA, and it turns out they're not as white as they thought…most of us out there are not just French or Scottish or Jewish…It's diversity…What I love is seeing that map of who you're connected to…But now we're seeing stories of people who say 'I used to be a bigot and now I'm not. (*Recode*)

Wojcicki's comment about white supremacists refers to a trend among alt-right whites who enlist 23andMe services for the purpose of proving their whiteness (Panofsky and Donovan). These narratives have appeared on online forums affiliated with alt-right ideologies, like 4Chan. Wojcicki's dismissal of the racially violent motivations behind white supremacy group's decision to enlist the company's services is illustrative of 23andMe's systematic disregard for the material and cultural consequences that the company's services hold for historically underprivileged individuals and communities.

Conclusion

The disparity I have shown between what 23andMe promises and what they provide points to an urgent need for 23andMe to account for the material and cultural consequences their services hold for historically underprivileged individuals and communities. The cultural valuing

of Big Data research has enabled the company to call their research scientific and to publish questionable research findings as science, despite obvious gaps in methodological regulation, demographic representation, and data quality. One of the very real concerns is that they hope not just to detect valuable patterns in genetic information (which could have consequences for ethnic or racial groups or people at risk of developing chronic diseases), but actually patent human genes.

In the same way that voter databases have been used to erase the participatory processes for democratic representation by abstracting voters/citizens into commoditized data, 23andMe (and many other Big Data firms) has abstracted people into similarly commoditized data, effectively erasing and universalizing their contingent identities, and subsuming them into its difference-less neo-imperialist platform. 23andMe has been able to sell access to its databases to the highest bidder and make subjective selections as to which researchers and research agendas the company will support. And, because the company is exempt from data regulation requirements involved in public genetic data research, the sheer size of their database has given the company's researchers the power to publish findings that are essentially unverifiable. Just as social media companies like Facebook harvest users' social data to sell advertising, 23andMe collects customers' genetic and phenotypic data to sell to various stakeholders. The strategic partnerships that the company has curated over the past decade, including those with pharmaceutical giants like GlaxoSmithKline and Pfizer, point to a decidedly (and predictably) neoliberal agenda (Philippidis).

In her discussion of the rhetorical power afforded by the magnitude of an archive, Rice has pithily argued that, "Much like the truther who feels as if an abundance of shared and circulated images confirms evidence of a conspiracy, big data analysts can sometimes connect

too strongly the links between magnitude and scale" (37). Drawing on Hawhee's argument that the goal of epideictic rhetoric is to create rhetorical magnitude, Rice argues that the megethos of big data analytics evokes a sense of evidentiary weight for arguments extending from the networks of data in question. Because 23andMe's genetic database is so large, their results and research findings become valuable by default. In other words, 23andMe's success is contingent upon the commonplace assumption in Big Data research that equivocates quantity with quality, seeing bigger data as being inherently better data.

Chapter Four

When the Goal Becomes Assimilation: Learning Analytics in Higher Education

In 2006, the Commission on the Future of Higher Education released a report interrogating and criticizing universities for failing to adequately prepare students for the demands of their future careers, citing concerns shared by employers that new college graduates "lack[ed] the critical thinking, writing and problem-solving skills needed in today's workplaces" (Spellings 3). The report's authors situated this as a national concern, explaining, "As other nations rapidly improve their higher education systems, we are disturbed by evidence that the quality of student learning at U.S. colleges and universities is inadequate and, in some cases, declining." Additionally, they noted that this is a crisis, not just of learning, but of institutional transparency, explaining that the current "lack of useful data and accountability hinders policymakers and the public from making informed decisions and prevents higher education from demonstrating its contribution to the public good" (4). In the years following, higher education leaders were met with increased demands from public and financial stakeholders to provide empirical evidence demonstrating the effectiveness of their institutions' efforts toward educating future members of the nation's workforce. In response to these calls for greater accountability, university administrators across the US have since scrambled to provide quantified evidence of the education rigor of their programs and to demonstrate that students at their institutions are achieving desirable learning outcomes.

Compiling the data necessary to analyze an entire institution's educational outcomes is a large, complex endeavor requiring significant labor and resources. Many administrators turned to

Big Data, specifically learning analytics, hoping to make the institution-wide assessment of student learning a more manageable, affordable, and efficient endeavor. The Society for Learning Analytics Research defines learning analytics as "the measurement, data collection, data analysis, and reporting of data about learners and their contexts" (SoLAR). For the purposes of this chapter, I will use the term "learning analytics" to refer to any tools and/or methods for using automated algorithms to make meaning from large and complex sets of data generated from user activities in digital learning environments. Learning analytics is comprised of two different types of Big Data analytics, each of which is used to meet a specific goal. To enhance student learning, universities deploy predictive learning analytics, while to automate student learning outcomes assessment (both to feed back into the university to improve teaching and to demonstrate the university's progress to national accreditation agencies), universities use descriptive learning analytics.

To deploy either kind of learning analytics (predictive or descriptive) at the institutional-level, universities need educational data. While universities often use centralized student information systems to collect and store student data, this process is usually reserved for basic information like demographics, course enrollment, and grades. This kind of data offers little to no insight about what or how well students are *learning* in those classes, and it does not offer much data about what happens at the course-level. So while this data is useful to universities for understanding retention rates, graduation rates, and other similar metrics, what universities need for learning analytics is data generated in the classrooms themselves.

Fortunately, for some universities there was already a trove of educational data being actively generated and stored from students' online activities while using their institution's learning management system. Learning management systems (LMSs)—also known as course

management systems, distributed learning systems, content management systems, instructional management systems, or learning platforms—are online learning systems and platforms designed to create and host digital learning environments for both face-to-face and online courses (Salisbury). Because LMSs are typically designed for university-wide implementation and are thus used in courses across disciplines and colleges, more often than not, a university's LMS is also its largest and most comprehensive repository of educational data.

US colleges and universities began heavily investing in LMSs in the years following the turn into the 21st century, when open-source LMSs with internal networks, like Moodle, were introduced onto the higher education marketplace. In their 2005 article critiquing the impact of new LMSs on higher education, Coates, James and Baldwin identify a number of factors motivating initial adopters of LMSs, including administrative and educational desires to increase the "efficiency of teaching" at their respective institutions, both in terms of cost and quality, to meet growing demands for greater access to higher education spaces, and to meet "new student expectations for advanced technologies" (23). Bringing together pedagogical tools and course management services into a networked online space, these early LMSs opened up new possibilities for teaching and engaging with students both on- and off-campus. Coates, James, and Baldwin note that, for administrators, the introduction of LMS technology offered new (albeit expensive) opportunities for streamlining educational and institutional operations using data-driven decision making. Equipped with the data infrastructure afforded by an LMS, administrators could "reduce course management overheads, reduce physical space demands, enhance knowledge management, unify fragmented information technology initiatives within institutions, expedite information access, set auditable standards for course design and delivery and improve quality assurance procedures" (Coates et al. 24). Essentially, LMSs promised to

empower university administrators with "a hitherto undreamt-of capacity to control and regulate teaching" (25). This promise went far in persuading administrators to embrace LMSs, which was significant given that LMSs have high start-up costs and rely on relatively new (and thus risky) networked technologies.

As more universities adopted LMSs, the LMS market grew and became competitive, with software developers continuously adapting to meet the challenges of new developments in education. When universities began offering more online and blended courses, for instance, developers started implementing more features for multimodal teaching and learning into their LMS programs. Likewise, when accreditation agencies began asking for more quantifiable data on student success and having access to technologies and methods for conducting learning analytics became a de facto necessity for universities, LMS developers began growing (and LMS vendors began stressing) the learning analytic capacities of their LMSs. Now, the LMSs that are the most successful are those that are pitched by developers and educational associations as state-of-the-art, data-powered mechanisms for helping administrators, educators, and students increase educational accessibility and foster student success while simultaneously streamlining course management and data mining efforts.

Administrative decisions to adopt LMS technologies are still motivated, at least in part, by the logistical affordances these systems provide. However, the biggest driver of LMS adoption now is the belief (shared among administrators and stakeholders) in the promise (espoused by software developers, educational non-profits, and researchers from the newly formed fields of Learning Analytics and Educational Data Mining) that LMSs, through their learning analytic capacities, hold the solution to both higher education's data problem and its assessment problem. With these tools, universities can leverage educational data toward

increasing institutional effectiveness and efficiency. Unsurprisingly, given that the learning management market is dominated by for-profit companies and thus driven by market demands for growth and profit, LMS vendors dig into this promise, situating themselves as gatekeepers of learning analytics, and thus gatekeepers to the promise of Big Educational Data.

As recent scholarship critiquing the Big Data boom in higher education has shown, however, once such programs are integrated, LMSs almost always fall short of their initial promise (Crawford; Milner; Reyman; McKee). Even the National Institute for Learning Outcomes Assessment (NILOA)—an organization aimed at helping university and college administrators and educators, as well as accreditors and educational associations use assessment data to enhance learning and empower learners—recognizes the shortcomings of LMS learning analytics. In their 2018 report, which details findings from a nationwide survey of undergraduate universities, researchers found that, "[w]hile assessment-related technologies hold promise of assisting with alignment and integration of learning across the institution, meaningful implementation remains elusive" (Jankowski et al. 3). The report's authors explain that, when asked about LMS technologies, survey respondents expressed concerns as to how their institutions were integrating these technologies into the classroom. They note that, "Provosts indicated they were unsure how to implement software solutions in a manner that would fit with the institutional culture they were trying to support and build connections within and across the institution" (Jankowski et al. 4).

Although these concerns point to real gaps in how universities implement data-driven assessment technologies (including how administrators and instructors are trained to use these systems), they leave unquestioned the assumption that these assessment technologies actually have the capacity to accurately and adequately measure student learning. In fact, in much of the

research on data-driven assessment technologies, the limitations of learning analytics are framed as temporally-bound problems that will be solved once technology inevitably progresses (the scholarly equivalent of a raised hand vaguely gesturing toward the future) (Kinshuk; Aguilar; Siemens and Long).

This chapter pushes beyond these initial, limited critiques by arguing that the promise at the heart of learning analytics—namely that by tracking and measuring learning (as it is defined and produced using externally-developed Big Data manipulations) education can become more personalized and that institutions can be held more accountable to concrete standards—is necessarily unfulfillable. Despite the myriad conveniences LMS platforms afford instructors (especially in terms of streamlining the management and distribution of course materials), I argue that their underlying algorithmic structures for analyzing data are incapable of accounting for the complex and multi-faceted realities of student learning. Further, I argue that, while the original goals motivating the turn to data-driven learning outcomes assessment were informed by principles central to the already ongoing assessment movement in higher education, when Big Data (via learning analytics) was brought to bear on student learning, it put the concept of assessment (and with it, the concept of learning) into question (Kuh and Ewell).

The bulk of this chapter is oriented around the question of what counts as learning in these digitally-mediated and data-driven systems. I look at two examples of educational software through which learning analytics are being deployed to enhance student learning and facilitate data-driven learning outcomes assessment: Canvas LMS and Calibrated-Peer Review. In my analysis of these two programs, I trace the process of translating user data to (allegedly) successful learning outcomes and the deformation of learning that occurs as a result. In doing so, I show how the mechanisms underlying learning analytics work to decenter critical and

intentional pedagogy in favor of quick algorithmic judgments. These methods, I argue, enact new contexts for teaching within which difference is framed as a barrier to student success and, ultimately, learning.

The (Unfulfillable) Promise of Learning Analytics

Learning analytics as deployed in LMSs rely on data mined from user activities on the platform. This includes student-generated data including data drawn from assignment submissions, exam and quiz answers, as well as data mined from online activities, such as page views, clicks, and timestamps. Machine learning algorithms can sift through this data to identify traces of "student learning" that can then be aggregated across assignments in a course, across many students in a single course, and/or across many courses in an institution, and used as evidence of individual learning performance and level of engagement.

Accordingly, learning analytics is methodologically contingent upon the belief that an objective measure of student effort can be gleaned from digital traces, or data. It is no coincidence that this belief resonates with beliefs driving the Big Data rush in business; developers and educational advocates explicitly acknowledge that the turn to data-driven educational assessment and learning analytics is directly informed by the growing culture of data-driven decision making emanating from within the tech and commercial industries. For example, in 2012, the US Department of Education released the report "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics," documenting the application of Big Data practices in higher education, and calling for higher education leaders to embrace "the culture of data-driven decision making" (Bienkowski et al. 8). As justification and support for their call, the report's authors cited recent successes in the commercial sector with

companies using Big Data technologies (specifically, technologies for data mining and predictive modeling) to provide users with personalized content. They argued that those same Big Data technologies, repackaged as 'educational data mining,' and 'learning analytics,' could be leveraged for use in education, specifically in the context of personalized learning.

The report's authors claimed that learning analytics could "predict" students' chances of success by comparing data said to represent a student's digital engagement with predetermined standards for what a successful or unsuccessful student's individual digital engagement should look like: "Using these measures, teachers can distinguish between students who are not trying and those who are trying but still struggling and then differentiate instruction for each group" (Bienkowski et al. 20). This claim assumes educational data that has been mined and analyzed is wholly representative of an individual student's experience and engagement with digital material. Such an assumption presents obvious blind spots and problems of representativeness. For instance, the data points that get counted as learning do not clearly map onto disciplinary understandings of what learning looks like. Further, the programs and data infrastructures undergirding learning analytic systems cannot account for students or educators whose activities do not register as digital signals.

The report's authors imagine a future wherein learning analytics is fully integrated into online educational spaces. In this imagined, digital academic utopia, learning analytics opens up new possibilities for enacting personalized learning:

> Educational data mining and learning analytics [can be] used to research and build models in several areas that can influence online learning systems, including what a learner knows, what a learner's behavior and motivation are, what the user experience is like, and how satisfied users are with online learning. At the simplest

> level, analytics can detect when a student in an online course is going astray and nudge him or her on to a course correction. At the most complex, they hold promise of detecting boredom from patterns of key clicks and redirecting the student's attention. (Bienkowski et al. vii)

To put this into perspective, consider the examples that the authors' provide of the types of data points that are collected by learning softwares: "minutes spent on a unit, hints used, and common errors" (Bienkowski et al. 20). By claiming that learning analytics can extract information about students' affective experiences (in this case "boredom") from the data generated by their online activities (in this case 'patterns of key clicks'), the report's authors reveal a connection between the promise of learning analytics to personalize learning and the promise, explored in Chapter 2, of the marketing analytics to personalize advertising—the authors even go so far as to advocate for the institution of psychographic student profiles akin to those deployed by Facebook and Cambridge Analytica.

Stephen Aguilar, an education and psychology professor at the University of Southern California, frames learning analytics' promise to personalize student learning as a solution to the problem of "teaching to the average," when he writes:

> Learning analytics-driven educational technologies will move us away from thinking about the 'average' student, while simultaneously empowering those invested in education to focus on the individual needs of their students. These technologies have the potential to personalize learning, and are well positioned to contribute to socially just and equitable student outcomes because they do away with the notion of an 'average' student. Instead, they allow teachers, administrators, and other stakeholders to personalize the learning experiences of students. (Aguilar)

While I agree there is value in doing away with the notion of the "average" student, Aguilar's grand claims about the potential of learning analytics to help individualize learning fail to adequately account for the realities of data-driven student assessment, as well as the reality of students' experiences. Importantly, learning analytics largely depends on the practice of clustering—a Big Data method that detects patterns across many variables and creates groups, or clusters, based on similarities. In other words, learning analytics does not create a system for seeing students as individuals; rather it works by creating categories of students. In creating clusters of students, learning analytics enacts the very practice that it is posited to alleviate: identifying an "average" student profile within a cluster. And while Aguilar concedes that there are a number of ethical and methodological issues with Big Data practices, he assures readers that these problems are the product of temporary gaps in technological capacity.

However, by failing to account for the harm that Big Data commits in the present during these allegedly transitional phases, Aguilar reinforces the same unfounded promise responsible for the rise of Big Data in the first place. The notion that learning analytics will accomplish these promises given sufficient time to develop more sophisticated data gathering and processing techniques mirrors the critical rhetorical appeal that drives Big Data adoption writ large. If, as I argue, Big Data's success hinges on the fabled eventuality of as-yet undeveloped analytic power, it neither can nor will ever attain the equalizing and accommodating capacities assured by its guarantors. In fact, as I have shown in the previous chapters, these promises openly conflict with the mechanisms by which Big Data creates profit, and thus succeeds. Arguments like Aguilar's effectively kick the can down the road for future generations to grapple with these consequences, further enabling Big Data's destructive capacities.

As with the two previous chapters, the problem here is not that there is not enough data or that the analysis is not sophisticated enough, or that these systems are just preliminary attempts to use tools that will eventually, with greater refinement, accomplish the tasks set before them. Rather, the problem is that assessing learning with these technologies demands that learning itself be re-defined and reconfigured so as to be measurable by such a tool. Although the initial promise of learning analytics is grounded in rhetorics of personalized learning, this promise comes with the caveat that the mechanisms through which they are deployed place constraints around what can be counted as learning. While all methods for assessment require that learning be reshaped to fit the assessment model, the process of reshaping learning is more exaggerated when using learning analytics for data-driven learning outcomes assessment, and less clearly connected to common learning goals like critical thinking and deep reading. Just as instructors have to reshape their instructional materials and pedagogical approaches to fit within the predetermined structure of their institution's LMS or any other learning analytics-based application, students, too, have to make adjustments to how they approach the learning process. These patterns of refitting and reshaping learning to meet the demands of an LMS's predetermined structure creates a feedback loop. Over years of continued use and refinement of educator and student behavior to meet its constraints, the system creates data and, thus, motivates those behaviors that are the most amenable to the data generation and analysis functions it has been designed to fulfill. Importantly, the system never re-aligns itself with the student learning outcomes that the system was originally intended to measure and refine (Kuh and Ewell). Data thus becomes an end unto itself.

To more clearly illustrate what qualifies as learning in learning analytics contexts, and thus more clearly illustrate the risks that these systems pose for how we understand learning, I

will explore the data infrastructure and instructional practices behind the most popular and fastest-growing LMS for higher education in the US: Instructure's Canvas.

Learning Analytics in the Canvas LMS

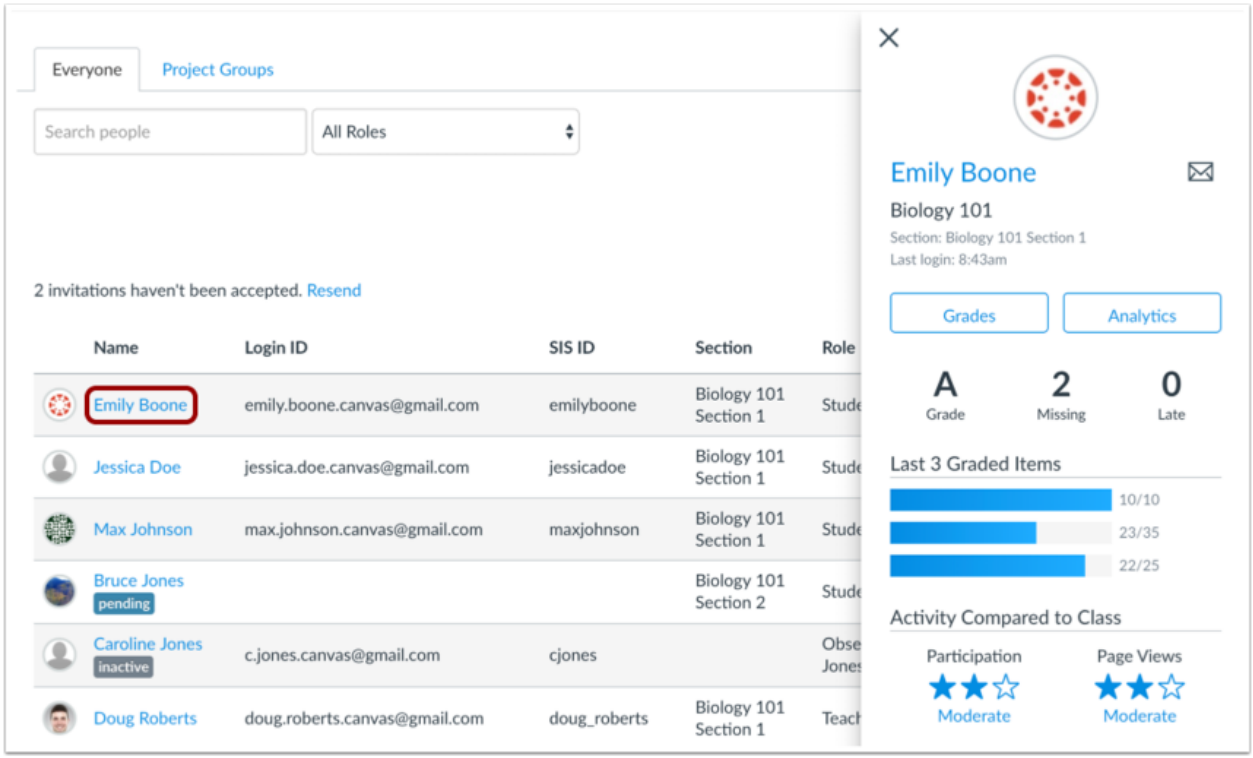
Developed and launched in 2011 by for-profit company Instructure, Canvas is a cloud-based LMS marketed for use in both K-12 and higher education contexts. What distinguished Canvas from other LMSs early on was that it operated as a Software as a Service (SaaS), a subscription-based and centrally hosted model of software licensing and development. The SaaS model means that users can access the Canvas software online, rather than through a downloaded, offline program. Similar to other SaaS like Google Drive and OneDrive, both of which operate via the cloud, Canvas's infrastructure makes its program easy for users to access and easy for developers to update. While Canvas is not currently the only SaaS operated LMS on the market, it was the first LMS on the market to offer cloud-based services capable of conducting large-scale data analytics and harnessing educational data to assess student learning. In the years following Canvas's release, while other LMS providers struggled to integrate similar data functionality into their services, Canvas was able to make its way to the forefront of LMS technologies and gain a significant advantage early on. This advantage has continued to give Canvas a leg up—as of 2019, Canvas is used by over 30% of higher education institutions in the US and is the fastest growing LMS on the market (*Canvas*).

One of Canvas's premiere features is the tools it provides for data analytics. Marketed as being "like Moneyball for student success instead of baseball"—referencing the wild success of the Oakland A's data-driven roster in 2002—Canvas's analytics are designed to serve a number of different functions aimed at bettering the quality of student education (*Instructure*). In

explanations of the potential benefits that universities can reap from using Canvas's analytics, the feature that Instructure emphasizes most is the LMS's capacity to help identify "at risk" students, which the program defines as "at risk of dropping out of a course, program, or institution" (Canvas, *Glossary*).

The main mechanisms through which instructors using the Canvas LMS are supposedly able to identify 'at risk' students are the 'Course Analytics' feature, which includes compilations of data from all of the students in a single course and/or all of the students in multiple sections of the same course, and the 'Student Analytics' feature, which includes data from individual students enrolled in each course. Both the course analytics and the student analytics rely on user-generated data. When an instructor or student logs onto Canvas, their activities and actions generate data that is collected, categorized, and analyzed via the LMS's underlying algorithm. This data is categorized and aggregated to produce various analytics, which are presented to certain users (i.e. instructors) as data visualizations representing student engagement and progress.

Canvas's data visualizations take the form of bar charts, line graphs, and various other graphical representations of data. While the course and student analytics are largely similar in their graphical representation, Canvas also offers a student "Context Card" which includes a more simplistic view of an individual student's analytics. In addition to providing the student's current grade, number of missing and complete assignments, and grades on the three most recent assignments, it includes a section titled "Activity Compared to the Class." As the image below shows, these "activities" are represented as two, three-star rating visualizations that show the individual student's page views and participation data in comparison to their classmates:



Graphic 1: Context Card https://community.canvaslms.com/docs/DOC-12705-415255479

Both of the minimalist, star-rating data visualizations are offered to instructors without any details as to what exactly these visualizations mean, or evidence of the mechanisms, data, or methods used in their production. The data used to construct the "Page Views" visualization is (if we are generous in our analysis) relatively straightforward in terms of what is being rated and compared (i.e. the number of discrete page views from each student's account, which are also made available to instructors in more detail via a timestamped log of each time that a student has accessed the Canvas course page). The data that the "Participation" visualization is meant to represent, however, is not as implicitly clear for instructors using the context card feature.

Despite the lack of explanation or context, however, the familiarity and clarity of the star-rating system grants the data visualization rhetorical power, encoding a particular kind of student success as a nudge to instructors. What counts as success and how success is represented in the

context card feature is very much contingent upon what is encoded into the Canvas algorithm. While Canvas does not provide an explanation of the context card mechanism on the Canvas portal, they do explain it on their Canvas Community website as a "simplified overview of a student's progress" that is based on grades from the course Gradebook and "standard page view and participation activity in course analytics" (Canvas Doc Team).

To create participation scores, Canvas' system compares the data that each individual student's account activity generates with the equivalent data from their classmates' accounts. The following table shows the user actions that generate the data that the participation analytics are based on:

User Actions that Generate Student Participation Data	
• Commenting on an announcement • Submitting an assignment • Submitting a quiz • Initiating a quiz	• Joining a web conference • Creating a wiki page • Posting a discussion comment • Loading a collaboration page

Once data has been generated, it is then aggregated and fed into an algorithm that scores student participation in relation to their peers. The resultant participation scores are then presented to instructors in the form of a three-star rating system labeled "Participation."

The explanation provided on the Canvas Community website also includes an important qualifier as to the quality of the data represented in the context card feature: Canvas's mobile app is not configured to collect data generated by student activities and actions (Canvas Doc Team).[v] In other words, because the algorithm used to create student analytics cannot account for the mobile app's limited data functionality, for those students who mainly use the Canvas mobile app (which may be for a number of reasons, from individual preference to having limited access

to wifi, laptop, or desktop), their student analytics will be skewed. In the space where student-generated data might appear given mobile functionality, there will instead be (potentially significant) gaps in logged activity. When their data is run through Canvas's predictive models, these students can potentially receive lower participation and page view ratings than their peers whose activity data has been successfully harvested via the Canvas website. While instructors may be able to account for these gaps in some other way, for instance, by asking students which type of device they use to access Canvas and then taking the device-type into consideration when assessing participation, this correction is unlikely given that these issues in data quality are not made readily apparent to instructors using the course and student analytics functions on their Canvas course page. It is also worth pointing out that, by posting this explanation of the Context Card on a page external to the Canvas Website, developers are working against their own narratives that LMSs are self-contained systems. Even if instructors were able to find and access the information that is posted on the external Canvas Community website, they would be hard pressed to find detailed explanations about how Canvas's learning analytics work.

Just as Canvas's analytics fail to account for data generated via the mobile app, they also inevitably fail to account for non-digital activities. For instance, if a student downloads the course assignments and syllabus, or prints out a PDF of course materials, they may return to that printed or downloaded document many more times throughout the semester. However, because their pageview and participation data will only show that they have visited the page once, Canvas's analytics will rate that student's activities as being less than, say, another student in the course who only accesses course material via a web browser.

These largely unaddressed issues with the quality and evenness of student data are quite problematic, especially considering that Canvas posits its course and student analytics features as

capable of predicting and preventing "at risk" students. Consider the following hypothetical example: An instructor using Canvas notices that a student has been automatically flagged by the system's course and student analytics features as potentially 'at-risk' (for instance, by highlighting their name in red in the gradebook). The instructor clicks on the student's context card and sees that, according to the system's analytics, the student has a low pageview ranking (say, one star out of three), a low participation ranking (two stars out of three), and has not submitted anything for an assignment that is now overdue. If the course is small enough in terms of student enrollment and if the semester is far enough along that the instructor knows the student personally, the instructor might realize that the student has not shown up to class for the past few sessions. Wanting to investigate further, the instructor checks that student's activity records and finds that the student has not generated any new data for two weeks. Assuming that this means that the student has not accessed their Canvas page for the same period of time that they have been absent from class sessions, the instructor could then triangulate that perhaps the student is experiencing some distress and send a follow-up email. If the class is large or it is early in the semester, and the instructor is not familiar with the student, however, the likelihood of them recognizing this student as being at risk drops significantly.

Now, imagine that there is a student who has been regularly attending class, but experiences perpetual anxiety about her performance, leading her to check the course's Canvas page frequently. Because she has generated a lot of data on the Canvas website, she is not flagged by the system. Her high participation and page view rankings mask her real degree of at-risk-ness, preventing the same kind of outreach the more "obviously" struggling student would receive. There are a number of factors that could contribute to a false positive or negative in Canvas's analytics: mental health, technical difficulties or limitations, group work, or offline

("analog") work. Identifying and correcting a false positive or negative is difficult, however, given how Canvas's Course and Student Analytics are structured. While analytics for individual students are made available to instructors, those same analytics are not available within the student view. In other words, students cannot see the data that they themselves have generated. On the one hand, opacity could be a benefit because students are less likely to game the system by artificially inflating their data. However, it poses an even larger ethical dilemma: without disclosing the types of assessment mechanics of the Canvas website to students in the course syllabus for example, students may not be aware how (or even *that*) their digital behavior is influencing not only their instructor's perception of them but potentially their course grade as well. Leaving students unaware of how their activity is being represented to instructors renders them unable to address inconsistencies, errors, or gaps that may arise across their own Canvas data.

Similar technologies are used to conduct instructor analytics using data generated from users assigned the role of "Instructor." The following table lists some of the types of data used to conduct instructor analytics:

User Actions that Generate Instructor Participation Data	
● Posting an announcement ● Commenting on an announcement ● Updating an assignment's settings or description ● Updating the course calendar	● Joining a web conference ● Creating a wiki page ● Posting a discussion comment ● Loading a collaboration page

Depending on the course settings, instructor analytics can be made available to lead instructors who can use it to compare other instructors in a course (i.e. teaching assistants) or across

multiple sections of the same course. These analytics are not clearly accessible to the individual instructor herself, but they are available for administrators, raising the same concerns as student analytics. And while a full accounting of instructor analytics is outside the scope of my current focus, it is important to recognize that learning analytics as deployed in Canvas's LMS reconfigures 'successful' teaching as well as learning.

Questions about the quality of data are rarely at the forefront of institutional discussions around LMS adoption. In the majority of universities and colleges where Canvas is being implemented, there has already been another LMS in place. Accordingly, institutional conversations surrounding Canvas often remain centered upon the logistical challenges of *transitioning* to Canvas from another system, like Blackboard, Moodle, etc. Unsurprisingly, these conversations are often dominated by comparative claims from students, teachers, and administrators hailing and lamenting various affordances and constraints of their institutions outgoing and incoming LMSs. And the majority of this comparative work starts and stops at the level of the user interface, with the increasingly consequential data collection and processing mechanisms often being left out of the discussion.

The lack of attention being paid by students and instructors toward the Big Data-end of Canvas is in part a product of the way that data privacy policies gets configured in LMS software. Much of the literature on the potential for the misuse of personalized data focuses on the right of the platform to harvest and manipulate individuals' identifiable, personal information. To mitigate the most serious risks, many argue, the system should implement a robust policy of informed consent such that users are aware of what data is being collected when and to what ends it is being applied (Smith et al.; Flores et al.; O'Neill). Many further insist that users then be given the right to opt out of such data aggregation. Large-scale platform companies

like Facebook have begun making certain types of automated data collection 'optional' for users; however, if data analytics, and thus data, are critical to a platform's functionality, users who opt out are, in doing so, agreeing to expect a lower quality, potentially unstable product and service from the company. In the case of LMSs this gets complicated

Digital rhetoric scholar Jessica Reyman argues that serious issues of consent arise when universities and instructors require students to use any particular web-based educational programs or platform, like an LMS. In these contexts, informed consent is not really possible because students and instructors have little agency in choosing their institution's LMS (Wintrup; Reyman). Instructors do have some choice in how big of a role LMS will play in their courses—for instance, whether they want to use the LMS strictly as a tool for distributing and storing course material or if they want to develop their whole course to fit within the LMS structure).

For students in all cases, informed consent of LMSs data practices and policies becomes tacit upon enrollment. In an interview with *Present Tense*, Reyman explains,

> Problems arise when students are compelled to use certain technological systems, like the online learning management systems Blackboard and Canvas, as a requirement for a course. Each system comes with terms of use, which typically collect data from users and apply it—in aggregate—to improve their services and technologies. The issue in these situations is not necessarily what data is collected or even how it is used, but that the data is collected without permission and without opt-out options. (Edwards and Gelms)

When universities subscribe to a particular LMS, they are not only giving consent for their own institutional data to be harvested, but they are also granting consent on behalf of their staff, faculty, and students. This practice of granting consent-by-proxy raises important ethical issues around data practices, especially in terms of what data is made visible and for whom. These

issues are compounded when we consider the issues with data quality illustrated earlier; namely, that many students and instructors are unaware of the data being collected.

It is important to recognize that, while instructors may not currently be using learning analytics to track and assess students' learning progress (and while these features may not yet be perceived as critical to the system's functionality), as they become more familiar with the Canvas platform and begin to facilitate more of their teaching via the Canvas LMS, there may be more widespread uptake among educators of features like the student context card. As high enrollment courses and online-only courses become more and more prevalent on college campuses (a parallel change that is also a result of increased demands for greater efficiency and access in higher education), instructors may find themselves ever more inclined to use Canvas's learning analytics to gauge their students' progress and effort, their assessments will be, whether they know it or not, tied directly to the capacity of the LMS to track and analyze student data. And as Coates, James, and Baldwin argued in their 2005 critique of LMS, "LMS are not pedagogically neutral technologies, but rather, through their very design, they influence and guide teaching" (27). Without more educator and learner oversight, Canvas itself is likely to shape teaching toward the kinds of things that Canvas values, which risks amplifying the negative effects in terms of non-intended ways of learning.

The idea of tracking student activity for assessment purposes resonates with a new movement in writing assessment toward labor-based contract grading. At their most basic level, labor-based grading contracts are a form of writing assessment that privileges student work, or labor, done for the course (i.e. reading, writing, reflecting, discussing, assessing) over subjective judgements from instructors and peers as to the quality of student writing. Essentially, the more labor that students do for the course, the better their grade will be. Scholarship from Asao Inoue

frames labor-based contract grading as a powerful tool for anti-racist writing assessment. Inoue argues that, because labor-based contracts count all labor as equally valuable when determining student grades, they help "build equity among diverse students with diverse linguistic competencies since it is a grading system that does not depend on a particular set of linguistic competencies to acquire grades" (132).

Mapping the ideas central to labor-based contract grading onto the LMS learning analytic model, we can see a number of parallels emerge. Both learning analytics and labor-based contract grading are framed rhetorically as pedagogical tools for making classroom environments more inclusive and for helping empower student learners to achieve course learning goals. Further, both assessment practices use records of student activity to gauge student progress toward course learning goals. This comparison is not meant to alleviate concerns about data-driven assessment, nor is it meant as a critique of labor-based contract grading. By drawing attention to the ways in which certain elements of the learning analytics model resonate with practices central to labor-based contract grading, I aim instead to make the rhetoricity of student data more visible.

While both models of assessment are built upon the same promise, namely, that they can help instructors teach more equitably and effectively, their underlying methodologies reveal stark differences. For instance, when constructing a labor-based grading contract, instructors and students have agency to decide what counts as labor. When using course and student analytics on an LMS like Canvas, however, students are granted no agency to decide what data counts as effort or labor, nor can they intervene in their own assessment. Despite both models holding the same promise, learning analytics' reliance on predetermined standards for assessing learning and opaque methods for surveilling student work means that, as a mode of assessment, learning

analytics ends up constraining rather than empowering student learners. Returning to the notion of using learning analytics as a lens into students' affective experience, we can see how these systems might create risky learning environments wherein individual students, coded as users, are compared. The next section will explore how the effects and risks of accompanying learning analytics shift when applied at the course level.

Calibrated Peer Review: Learning Analytics at the Course-Level

Calibrated Peer Review (CPR) is a learning analytics system designed for (but not restricted to) instructors in the sciences as both a resource for fostering student learning around disciplinary-genres and a tool for conducting data-driven assessment of student learning. Far narrower than an LMS, CPR operates, not just at the classroom-level, but at the assignment-level. The CPR program was developed by UCLA faculty Orville Chapman and Arlene Russell who saw the potential of learning analytics to motivate instructors in the sciences to integrate more writing assignments into their undergraduate classrooms. While Chapman and Russell recognized the value of using writing as a tool for fostering student learning, they also knew that teaching with writing effectively required significant instructional labor, and would likely be unmanageable for instructors of large-lecture courses. In short, they sought to use learning analytics to mediate the burden of writing-intensive instructional labor (Russell).

Chapman and Russell argued that the CPR system could provide a more customizable approach to using learning analytics to meet the goal of teaching students how to be effective reviewers of their peers' papers. When used in conjunction with carefully designed assignments and discussions about giving effective feedback, CPR is alleged to improve student writing skills and reviewing abilities (Russell). In addition to being sold as a data-driven system for helping

foster student learning, CPR is framed as a "versatile" software from which administrators can mine valuable educational data and better meet accreditation standards for learning outcome assessment (which increasingly emphasize the use of writing assignments as a vital tool for fostering student learning across the disciplines) (Carlson and Berry). This is a sharp distinction from other applications of learning analytics, which aim to measure student learning across all disciplines simultaneously.

After a student submits a draft to the CPR program, they are given access to a calibration module, where there are presented with three calibration papers written (by the instructor or adapted from another instructor) specifically for the CPR program. The calibration papers consist of three different versions of the same paper, each of varying quality. The high-, average-, and low-quality (or "the good, the bad, and the ugly") sample papers are responses to the same assignment that the student was initially given, and each of these papers has been evaluated by the instructor using calibration questions. These questions are written to align directly with the assignment's rubric and serve as a guide for helping students learn about the content, genre, structure, and style of the assigned paper as they move through the calibration process. Like the calibration papers, the calibration questions are written by and entered into the program by the course instructor (or adapted from another instructor's model).

For each of the three calibration papers, students answer the calibration questions and assign a rating. CPR then assigns a reviewer competency index based on a comparison of the student review to the instructor review of each paper. To "pass" students' reviews of the sample papers must align with the instructor's review. If they line up, students are deemed "calibrated." If not, they have to repeat the exercise until their reviews are adequately normalized and then they can submit their work through the program. Once calibrated, students are given access to

three of their classmates' papers (randomly assigned and anonymous), which they review and rate using the calibration questions. Students are then presented with their own papers, which they review and rate according to the same guidelines. At the end of this process, students are graded on how well their reviews and ratings both of theirs and their peers' papers align with the reviews and ratings of their peers.

As a tool for peer review, CPR provides a number of benefits. Calibration questions and the instructor evaluation of the calibration papers can help students better understand the goals of assignments as well as their instructor's expectations (and what counts as "good" writing within their particular disciplinary genre) (Russell). And like traditional peer review, CPR provides a space where students can engage with their peers' writing and deepen their understanding of course content. Introducing learning analytics into this process, however, sacrifices many of the key benefits of traditional peer review because of the limited scope and rigid demands set forth in the structure of the system itself. By alleviating instructor responsibility for assessment and replacing it with data processing, CPR still morphs what counts as learning to meet the constraints of its algorithm. Even without the imposition of disciplinary uniformity inherent in an LMS, the kind of datafied assessment that CPR employs still jeopardizes the learning outcomes it was designed to address. By structuring peer review within an isolated, anonymized space where students have been pre-calibrated to give only certain kinds of feedback, the program fails to accomplish the functions that are most desirable about peer review, such as helping build course community, exposing students to their peers' ideas (via both reading and discussing), and helping prepare students to more successfully integrate meaningful revisions into their final drafts (van den Berg et al.). In CPR, the learning analytic model functions by prescribing a standard for what successful digital behavior looks like in online courses and assessing students

based on how well their digital activity patterns fit or do not fit within the set standard. Anomalies in how students read and understand their peers' and their own writing are flagged and penalized, fostering a pedagogical context in which difference is framed as a barrier to student success and, ultimately, learning (Volz and Saterbak). Essentially, within this learning analytic context, where differing interpretations are presumed wrong, assimilation rather than diversity in thinking becomes the central goal.

As this chapter has shown, data-driven learning assessment, and learning analytics more broadly, necessitate higher education agencies, including instructors, students, and administrators, try to normalize not only what success looks like in the classroom and university, but also how students can move across educational spaces and how instructors can engage with students. The standards of success built into LMS and other learning analytic equipped platforms are grounded in subjective claims with real material consequences. As Trevor Pinch argues, "Standards are rarely simply technical matters; they are powerful ways of bringing a resolution to debates that might encompass different social meanings of a technology. Standards are set to be followed; they entail routinized social actions and are in effect a form of institutionalization" (Pinch, 473). Not only does this limit the visibility of non-digital actors, but it simultaneously promotes a fabricated perspective of student experience, because the algorithmic outputs of the system are always already contingent on subjective agencies that produced the parameters for data interpretation. To understand the risks that these issues present for learners, I want to conclude this chapter by turning attention to another Big Data development in higher education, this one in the context of university marketing, recruitment, and admissions.

The Rise of Big Educational Data

In late-1990's, Baylor University began experimenting with using large-scale data analytics to redefine their admissions and recruitment strategies. Using Big Data methods like predictive modeling, Baylor administrators hoped to motivate more applications from "desirable" students while keeping their admissions marketing budget low. They began by constructing databases of prospective students that tabulated such standard admissions fare as what high school they attended, their SAT scores, and extracurricular activities alongside more context-specific data like whether or not they had conducted a campus visit or attended any university events. Using algorithms, Baylor admissions officials compared their prospective students' data with comparable data they had collected about current students from when they were prospective students. They also integrated data from current students' college careers thus far, including grades, retention rates, demographic information, campus extracurriculars, and more. Then, the prospective student data was run through a predictive model that computationally categorized, ranked, and assigned each prospective student a predictive score. A higher score signified that a student had a higher likelihood of being admitted to the university than others and that that student had a higher likelihood of succeeding once admitted. These algorithmically generated, predictive scores were used to determine targeted marketing strategies akin to those used in social media electioneering as discussed in the previous chapter. For example, a prospective student's predictive score would be used to determine how many mailings that student would receive from Baylor University, as well as what quality of mailing they would receive: "the top 75 percent were sent the more expensive print and mail viewbook while the bottom 25 percent would receive only a reply card" (Campbell et al.).

Enrollment-focused Big Data analytics like those developed at Baylor have since been taken up by higher education institutions across the country where they have been used to guide decision making around admissions and recruitment strategies. Today, Big Data applied to recruitment and admissions allows universities to be more discreetly selective while also continuing to target students who have been algorithmically identified as fitting within the spectrum of desirable applicants. With enrollment-focused predictive modeling, universities can create user profiles of prospective student recruits (akin to those used in the 2016 Presidential campaign to identify supporters likely to vote) to assess how well a potential student's predicted profile maps onto the models—built by the university—of ideal students. 'Propensity for learning' is a heavily weighted factor in these data-driven decisions, and algorithms are trained to identify individuals who have a high potential for learning per the university's standards. When learning analytics moves beyond assessing student learning, to include the assessment of prospective students' potential for learning, it has effectively pre-figured every meaningful contributor not only to classroom compatibility but campus diversity as well, not only reconfiguring learning, but reconfiguring learners as well.

Chapter Five

Conclusion

Big Data research, technologies, and applications are in dire need of accountability and transparency. Only by actively interrogating the discourses and outputs of Big Data applications can we begin to poke holes in this Big Data imaginary, which, as it is allowed to continue unchecked and unquestioned, becomes increasingly damaging as the number of applications grows. Perhaps the largest Big Data-driven threat on the horizon is the so-called Internet of Things, a menage of billions of personal devices designed to gather invasive personal data on their users continuously. It is obvious that the same process I have outlined in this dissertation will govern this new world: stakeholders will continue to buy and sell data and the information drawn from that data will be used to inform ever more decisions about ever more features of our lives. Coupling this imminent development with another newly emergent technology, Big Data AI—which proponents claim will have enough anonymized data to fully train algorithms to make immediate actionable decisions without any human intervention—shows these technologies will only become more invasive if we so allow them. Now, more than ever, it is important to think critically about how to intervene in Big Data applications that threaten to cut across social complexity, cover over diversity, and deny change.

Enacting a 'Principle of Proximity'

Developing a model for resisting the creep of Big Data is difficult for many reasons. Perhaps one of the most insidious barriers to resistance comes from the academy's own complicity with Big Data-driven enterprises governing the tools of research. While I was conducting research for and writing this dissertation, I became acutely aware of the inborn

ironies of having to conduct research using the very technologies I sought to interrogate. Before I even finished the first chapter of my dissertation, the data-fied version of it was bought and sold. I was targeted with, literally, hundreds of ads for 23andMe on Facebook, Twitter, and Instagram. I received countless push notifications from Google News like 'Make Big Data Work For Your Business!" and "Is Big (Brother) Data Watching You Sleep?"

Although I experimented briefly with a number of note-taking applications and writing software, I wrote the majority of this dissertation in Google Docs in the Google Chrome browser while logged onto my personal Google account. I conducted the majority of my research using my own devices (computer and phone), and I connected to no less than 100 different wireless networks, both private and public, local and international. I toggled between browser tabs related to my dissertation and tabs related to everything else in my life. I navigated to and from 23andMe's website from my personal email, my work email, and my social media feeds. I researched and wrote the second chapter of this dissertation while planning my wedding. Peering back at my browsing history from last Spring, there is an almost manic pattern between searches, with queries like "Why no voter registration in North Dakota?" interspersed with queries like "Why are weddings so expensive?"

After casting my vote for the 2016 Presidential Election, I posted a selfie to Facebook and tagged my local polling place. I also submitted my DNA to 23andMe's laboratory and I used Canvas to teach three different courses (and for one of those, I used it a lot). I am in close proximity to my technologies of research. While my close proximity to Big Data is in part a product of existing in contemporary networked life, contributed to by decisions I made prior to developing this dissertation project (setting up and maintaining a Facebook account, for instance), my proximity to my research was also necessary for my project. To find out how

23andMe framed genetic test result information to customers, for instance, I first needed to gain access to 23andMe's customer portal, which is only possible for customers who have signed the privacy policy. Thus, in order to study 23andMe's Big Data practices, I had to expose my data and myself to potential risks. What these risks were and what degree of riskiness they presented was only made clear to me once I was in the system.

In "Googling the Archive," Janine Solberg writes that "Understanding the degree to which particular technologies advance our historical inquiries (not to mention ethical and theoretical commitments) demands that we reflect on the technologies themselves and the ways their structures 'push back' against our inquiries" (Solberg, 61). Pushing back against uncritical uses of digital tools and technologies, Solberg's 'principle of proximity' demands that composition and rhetoric historians—and, I would argue (and I think Solberg would agree), composition and rhetoric scholars broadly—interrogate the interaction *between* researcher and technology. While both researcher and technology are active agents in the research process, they are separated by differences in information complexity, understanding, and literacy. Put differently, there is an inherent divide between computational structure and process, and researcher experience and understanding. At the very least, this separation necessitates paying attention to the mediation, materiality, and positionality of both the researcher and the technologies of research. Solberg concludes her article with a claim and a question. She writes, "If ever there was a kairotic moment for change, experimentation, and reinventing our research tools, then we are in it. The question is, 'What, if anything, are we going to do about it?'" At what point do we attend to the failure of Big Data to drive real social change? And how?

Throughout this dissertation, I have explored three different case studies to make the argument that Big Data is fundamentally rhetorical, and by understanding this essential

rhetoricity, we can escape the false binary between accepting technologies uncritically and rejecting them wholesale. Yet rhetoricity, taken by itself, can be a slippery category that evades precise definition. After all, there are multiple levels on which a complex network of technologies like Big Data can be said to operate rhetorically. The definition I proposed in this dissertation, therefore, assigns rhetoricity—at the highest level of abstraction—to the entire sociotechnical system that upholds Big Data and its promises. Such a broad view of Big Data is necessary, I argue, precisely because resisting the worst impulses of the technocrats who propel Big Data while subscribing to its services for the basic functions of our daily lives and work routines requires being especially clear about the root of Big Data's power. This broad view of Big Data builds on yet challenges the predominating conceptions of digital materiality in the field.

Following the scholarship from new materialism, digital rhetoric scholars have emphasized the degree to which digital systems cannot be understood solely as a product of their human creators. Much of the attention in this domain rightly falls on algorithms. Because algorithms make meaning out of traces of human and non-human activities, they are said to be rhetorical (Brooke, Brown, Reyman). While the decisions that algorithms make are partly mediated by human actors, these scholars argue that algorithms themselves exhibit agency in their capacity to shape and inform the conditions of rhetorical situations. Much in the same way that a rhetor works to identify the best available means of persuasion based on their audience, context and purpose, algorithms search for patterns in available data and make decisions (i.e. about what meaning to draw out of datasets, about what content to generate or populate) with the goal of "persuad[ing] users toward particular engagements" (Brock and Shepherd).

In *Ethical Programs*, Jim Brown situates algorithmic activities in the space between database and narrative. In this liminal space, he argues, algorithms perform rhetorical procedures and act as "rhetorical devices" that make decisions and mediate the translation of data into narratives and vice versa. He explains, "If narrative presents a single and ordered path through information, database allows for multiple (even contradictory) paths to exist at once." Brown asserts that "the interpretation of data will always require narratives," and warns that as more data becomes available, the line between database and narrative becomes increasingly blurry. To address this blurriness, Brown posits rhetoricians as ideal candidates for participating in the development of methods for enacting "ethical programs." While I follow Brown's call for rhetorical mediation of the digital, I propose that this mediation requires a different scale of intervention. Brown and other digital rhetoric scholars are undoubtedly right about algorithmic agency and the manner in which digital technologies create new meanings, but confining our deployment of rhetorical tools to analyses of singular technologies like the algorithm fails to grapple with the (considerably larger) public spectre of the technology.

Big Data extends far beyond its technical definition and capacities. In the introduction to this dissertation, I explained how Big Data, once a technically useful term designating especially large and complex datasets, now has more to do with process: squeezing large datasets to discern every possible correlation. Yet, in each of my case studies, I have shown how Big Data has been exalted far beyond the merits of its technical abilities through the tantalizing and inescapable promise that through data we can know each other and ourselves. Likewise, Sheila Jasanoff has recently argued that:

> The bare term "data" tends to sanitize the world of observation, erasing from view the observational standpoints and associated political choices that accompany any compilation of authoritative information. The notion of 'big data,' even more sweeping in its scope and ambition, implies a panoptic viewpoint from which the entire diversity of

> human experience can be seen, catalogued, aggregated, and mined, so that the narratives derived from the data speak as if for themselves, compelling reasonable people to action (Jasanoff 2017).

This exaltation that "the data speaks for itself" is not only the driving force behind the widespread adoption of these technologies writ large, it is also the wellspring from which excuses for Big Data's transparent failures arise. Yet, neither of these phenomena emerges with any clarity from the algorithms themselves, or any other technical components for that matter. In fact, the narrative mediation of databases that digital scholars point to as a potential site for intervention exist at the same level as the narratives from my case studies that ultimately failed to uphold the larger promises of Big Data for the people who donate their data. By limiting our analyses to discrete technical processes through which particular Big Data programs operate (for instance, examining what motivates a particular algorithm to generate particular narratives), we are also cutting off possibilities for accounting for the mythical truth-telling aura bestowed upon technological networks and sustained by institutional imperatives for efficiency.

The interventions of rhetoric and writing studies scholars in Big Data need not be restricted to analyzing algorithms or technical components because Big Data's rhetoricity is more complicated than the built-in logics of the machine. Big Data's persuasive power is separate and distinct from its perceived capacity to reveal truths about the world. Its persuasive power is, however, directly tied to the discourses surrounding and driving public trust and investment in Big Data systems. These discourses, I argue, are the key to Big Data's capacity to reshape the interpretive lens through which we read and navigate the world. Even before data is analyzed, the discourses surrounding Big Data work to inform and shape how we experience identity and community, how we learn and teach, how we conceive of our own bodies, and how we interact with the past, present, and future. So, while it is important to recognize and

interrogate the rhetorical work that algorithms perform, Big Data's rhetoricity fundamentally cannot be explained at the level of the database or its attendant algorithms when the discursive drivers (the promises that companies make) of Big Data are divorced from the ways that these technologies are put into practice.

Importantly, Big Data often succeeds at meeting the promises that are made to its real stakeholders (i.e. party leaders, pharmaceutical companies), namely the promise that equates Big Data with greater empowerment. While the promise of Big Data to empower gets articulated differently to accommodate particular stakeholder goals (i.e. winning elections, dominating marketplaces), it is always about leveraging data as a means for gaining control of a system. The way that this promise is met is contingent, not on the technologies of analysis, but on the rhetorical power that accompanies Big Data in the sociotechnical imaginary. Big Data spreads to ever greater applications largely because of its success in accomplishing the tasks set before it by these stakeholders, and continues to encourage investment despite its failures to protect the public.

If we are being charitable, we might understand the unevenness in the fulfillment of Big Data's promises to be largely due to the fact that Big Data systems lack the necessary context through which to understand its own data (and, importantly, data cannot provide its own context). In my discussions of blackboxed and proprietary algorithms, I argue Big Data's promise along with its opacity to outsiders and subsequent claims to veracity through volume that discursively neutralizes its tendency to make errors, to fail to account for certain people and communities, and to discriminate. In Chapter 2, I argued that Big Data was initially pitched in political contexts as both a tool for helping campaigns win votes as well as a voter-driven method for grassroots campaigning and a legislative tool for increasing state accountability for

voter accessibility. In Chapter 3, Big Data was first introduced as a tool for empowering patients through individualized health care and helping make strides toward mapping the entire human genome; and in Chapter 4, Big Data was pitched as an instructional tool for helping instructors personalize student learning and helping administrators more effectively measure student learning progress. In each case, Big Data ultimately fails to achieve the goals of personalization, democratization, and individualization— essentially, the promises made to the people whose data is being harvested and analyzed.

That being said, I do not believe that Big Data is inherently bad—there are productive and useful applications of Big Data—as Jasanoff argues, some of our biggest contemporary problems are only visible because of Big Data. Big Data indeed has the capacity to help identify otherwise invisible problems, but when Big Data is relied upon as the sole method of research and/or inquiry or when it is positioned as the solution to a problem, it necessarily distorts reality to match the limitations of its programming. Beyond getting out of the binary between promise and provision, we must also transcend the binary of technology as an intrinsic social good or evil to move beyond the notion that Big Data must be applied to everything or to nothing.

Because Big Data (and thus algorithms) are so ubiquitous and because so much of our social, professional lives are mediated by Big Data technologies, its effects can be difficult to recognize, making alternative possibilities for digital landscapes difficult to imagine. When people are replaced by data and meaning is algorithmically generated, it places limits on what is possible for the human narrative. When Big Data applications are blackboxed, it becomes harder to see from outside the system the effects that these systems have (or even that they exist at all). As users, ‘fortunately,’ we aren't totally outside the system. We have a way in, through Big

Data's paratext, through the user interface, and through critical reflection of our own experiences in the Big Data ecosystem.

A Note on Analyzing Big Data Systems

To contend with the limits and risks of Big Data applications, I see it as vital that rhetoricians turn our gaze to the narratives driving Big Data applications. As this project has demonstrated, focusing our efforts on the critical examination of the discourses that build up Big Data's power in the sociotechnical imaginary is a valuable way that rhetoric and writing studies can productively intervene in the Big Data phenomenon. My research process helped train my eyes to recognize traces of Big Data outputs when they appeared on my screen. Here were the most useful things that I did to build my data literacy:

1. I read through privacy policies and user agreements, (pribot.org is a useful tool that highlights key sections of privacy policies that speak to what data will be collected and how that data will be used).
1. I skimmed government records, legislative documents, legal codes, and patent laws.
2. I compared customer-facing and investor-facing webpages and marketing materials for 23andMe and Canvas, along with numerous third-party data analytics firms, to see how these services were pitched differently to different stakeholders.
3. I attended lectures and events on Big Data and data analytics sponsored through the University's Holtz Center, the ISchool, the Wisconsin Institute for Discovery, and DoIT Academic Technology.
4. I did deep dives into community forums

5. I methodically combed through user portals, clicking every link and reading all of the fine text. I took screenshots and downloaded PDFs of various pages in case they were taken down or changed, and made heavy use of the Wayback Machine at archive.org.

Importantly, I also integrated my research into my teaching and learned how my students navigated various Big Data environments themselves. In a 'Rhetoric and Power Online' course that I taught in Spring 2018, I dedicated a unit to exploring and analyzing Big Data as a magnifier, transmitter, and limiter of power. For their final project, I asked students to analyze and interrogate the privacy and data collection policies and practices that inform their everyday digital lives, beginning with keeping a log of their digital movements and activities over the course of 24 hours. Students then investigated the data collection practices and policies of an online platform of their choice (e.g. Facebook, Twitter, Snapchat, Tinder, Google) and then penned critical reflection essays in which they interrogated their own experiences as users of that platform. In these reflections, students grappled with questions of privacy, agency, and power, and imagined alternative approaches to social networking and platform development in which users' privacy and methodological transparency are prioritized above corporate agendas. I have attached the assignment as an addendum. By pushing students to grapple with the real consequences and implications that Big Data has on how they understand and move through the world, instructors can turn their classrooms into spaces with high potential for social and political intervention into the problems addressed in this dissertation.

Works Cited

23andMe. *23andMe Global Genetics Project*. https://www.23andme.com/global-genetics/. Accessed 2 July 2019.

---. *DNA Ancestry Test, Find DNA Relatives - 23andMe*. https://www.23andme.com/dna-ancestry/. Accessed 2 July 2019.

---. *DNA Genetic Testing & Analysis - 23andMe*. https://www.23andme.com/about/privacy/. Accessed 29 June 2019.

---. *Research - 23andMe*. https://www.23andme.com/research/. Accessed 2 July 2019.

23andMe CEO Anne Wojcicki from Recode Decode. https://www.stitcher.com/podcast/vox/recode-decode/e/56809847. Accessed 2 July 2019.

"23andMe for Scientists | Accelerating Genomics Research." *23andMe for Scientists*, https://research.23andme.com/. Accessed 2 July 2019.

"23andMe Tests New Ancestry Breakdown in Central and South Asia - 23andMe Blog." *23andMe Blog*, 3 Apr. 2019, https://blog.23andme.com/ancestry/23andme-tests-new-ancestry-breakdown-in-central-and-south-asia/.

"About Us - 23andMe Media Center." *23andMe Media Center*, https://mediacenter.23andme.com/company/about-us/. Accessed 2 July 2019.

Advertising Policies. https://www.facebook.com/policies/ads/restricted_content/political. Accessed 29 Oct. 2018.

Aguilar, Stephen J. "Learning Analytics: At the Nexus of Big Data, Digital Innovation, and Social Justice in Education." *TechTrends*, vol. 62, no. 1, Jan. 2018, pp. 37–45, doi:10.1007/s11528-017-0226-9.

Anderson, Emily E. "Direct-to-Consumer Personal Genome Services: Need for More

Oversight." *The Virtual Mentor: VM*, vol. 11, no. 9, Sept. 2009, pp. 701–08, doi:10.1001/virtualmentor.2009.11.9.pfor1-0909.

Ball, Cheryl E., et al. "The Boutique Is Open: Data for Writing Studies." *Networked Humanities: Within and Without the University*, ceball.com, http://ceball.com/wp-content/uploads/2013/11/NHUK-chapter-rhetoric.io-PREPRINT.pdf.

Barad, Karen. "Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter." *Signs: Journal of Women in Culture and Society*, vol. 28, no. 3, The University of Chicago Press, Mar. 2003, pp. 801–31, doi:10.1086/345321.

Bennett, Colin. "Voter Surveillance, Micro-Targeting and Democratic Politics: Knowing How People Vote Before They Do." *SSRN Electronic Journal*, 2014, doi:10.2139/ssrn.2605183.

Bennett, Colin J. "Trends in Voter Surveillance in Western Societies: Privacy Intrusions and Democratic Implications." *Surveillance & Society*, vol. 13, no. 3/4, 2015, pp. 370–84, doi:10.24908/ss.v13i3/4.5373.

Bienkowski, Marie, et al. "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief." *US Department of Education, Office of Educational Technology*, vol. 1, 2012, pp. 1–57.

Bijker, Wiebe E. *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*. MIT Press, 1997.

Blake, Aaron. "DNC Holds National Training as It Rolls out New Voter File." *TheHill*, 15 Aug. 2007, https://thehill.com/homenews/campaign/680-dnc-holds-national-training-as-it-rolls-out-new-voter-file.

Borgman, Christine L. "The Digital Future Is Now: A Call to Action for the Humanities." *Digital Humanities Quarterly*, vol. 3, no. 4, escholarship.org, Jan. 2010,

https://escholarship.org/uc/item/0fp9n05s.

Bossetta, Michael. “The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election.” *Journalism & Mass Communication Quarterly*, vol. 95, no. 2, SAGE Publications Inc, June 2018, pp. 471–96, doi:10.1177/1077699018763307.

boyd, danah, and Kate Crawford. “Critical Questions for Big Data.” *Information, Communication and Society*, vol. 15, no. 5, Routledge, June 2012, pp. 662–79, doi:10.1080/1369118X.2012.678878.

BuzzFeedVideo. *YouTube*. Youtube, 2017.

Campbell, John P., et al. “Academic Analytics: A New Tool for a New Era.” *EDUCAUSE Review*, vol. 42, no. 4, Educause, 2007, p. 40, https://er.educause.edu/articles/2007/7/academic-analytics-a-new-tool-for-a-new-era.

Canvas Doc Team. *How Do I View a Context Card for a Student in A... | Canvas LMS Community*. https://community.canvaslms.com/docs/DOC-12709-4152698664. Accessed 26 June 2019.

Canvas the Learning Management Platform | Instructure. https://www.instructure.com/canvas. Accessed 3 July 2019.

Carlson, Patricia A., and Frederick C. Berry. “Calibrated Peer Review and Assessing Learning Outcomes.” *Frontiers in Education Conference*, vol. 2, STIPES, 2003, p. F3E – 1, http://www.icee.usm.edu/ICEE/conferences/FIEC2003/papers/1066.pdf.

Cherokee Nation Responds to Senator Warren’s DNA Test. https://cherokee.org/News/Stories/Archive_2018/20181015_Cherokee-Nation-responds-to-Senator-Warrens-DNA-test. Accessed 30 June 2019.

Coates, Hamish, et al. “A Critical Examination of the Effects of LMS.” *Tertiary Education and Management*, vol. 11, 2005, pp. 19–36.

Comer, Denise K., and Edward M. White. “Adventuring into MOOC Writing Assessment: Challenges, Results, and Possibilities.” *College Composition and Communication*, vol. 67, no. 3, National Council of Teachers of English, 2016, p. 318.

Commission, Federal Election, and Others. “Help America Vote Act of 2002.” *Public Law*, 2002, pp. 107–252.

Compton, Matt. “What’s Really Happening With the Technology and Data That Helped President Obama Win.” *Medium*, Soapbox, 25 Feb. 2014, https://medium.com/soapbox-dc/whats-really-happening-with-the-technology-and-data-that-helped-president-obama-win-706865a211b8.

Concordia. *Cambridge Analytica - The Power of Big Data and Psychographics*. Youtube, 2016, https://www.youtube.com/watch?v=n8Dd5aVXLCc.

Democracy for America : Home. http://democracyforamerica.com. Accessed 2 July 2019.

Edwards, Dustin W., and Bridget Gelms. *Vol. 6.3: Special Issue on the Rhetoric of Platforms*. http://www.presenttensejournal.org/wp-content/uploads/2018/03/Vol.-6.3-Special-Issue-on-the-Rhetoric-of-Platforms-%E2%80%93-Present-Tense.pdf.

Elish, M. C., and danah boyd. “Situating Methods in the Magic of Big Data and AI.” *Communication Monographs*, vol. 85, no. 1, Routledge, Jan. 2018, pp. 57–80, doi:10.1080/03637751.2017.1375130.

ElizabethForMA. *YouTube*. Youtube, 2018.

Elmer, Greg, et al. *Compromised Data: From Social Media to Big Data*. Bloomsbury Publishing USA, 2015.

Endres, Kyle, and Kristin J. Kelly. "Does Microtargeting Matter? Campaign Contact Strategies and Young Voters." *Journal of Elections, Public Opinion and Parties*, vol. 28, no. 1, 2017, pp. 1–18, doi:10.1080/17457289.2017.1378222.

Eyman, Douglas. *Digital Rhetoric: Theory, Method, Practice*. University of Michigan Press, 2015.

FDA. *FDA Allows Marketing of First Direct-to-Consumer Tests That Provide Genetic Risk Information for Certain Conditions*. 2017, https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-direct-consumer-tests-provide-genetic-risk-information-certain-conditions.

---. "Warning Letter from the FDA to Anne Wojcicki,." *FDA Medical Bulletin: Important Information for Health Professionals from the U.S. Food & Drug Administration*, 2013.

Flores, Lisa A., et al. "Dynamic Rhetorics of Race: California's Racial Privacy Initiative and the Shifting Grounds of Racial Politics." *Communication and Critical/Cultural Studies*, vol. 3, no. 3, Routledge, Sept. 2006, pp. 181–201, doi:10.1080/14791420600841351.

Gates, Henry Louis, Jr. "Tracing Your Roots: Why Did My DNA-Test Results Change?" *The Root*, The Root, 18 Aug. 2017, https://www.theroot.com/tracing-your-roots-why-did-my-dna-test-results-change-1797974267.

Glossary: Basic Analytics Terms: Introduction to Analytics with Canvas. https://learn.canvas.net/courses/1208/pages/glossary-basic-analytics-terms?module_item_id=161592. Accessed 26 June 2019.

Green, Joshua, and Sasha Issenberg. "Inside the Trump Bunker, With Days to Go." *Bloomberg News*, Bloomberg, 27 Oct. 2016, https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go.

Griffin, June, and Deborah Minter. “The Rise of the Online Writing Classroom: Reflecting on the Material Conditions of College Composition Teaching.” *College Composition and Communication*, vol. 65, no. 1, National Council of Teachers of English, 2013, pp. 140–61, http://www.jstor.org.ezproxy.library.wisc.edu/stable/43490811.

Harris, Anna, Susan E. Kelly, et al. “Autobiologies on YouTube: Narratives of Direct-to-Consumer Genetic Testing.” *New Genetics and Society*, vol. 33, no. 1, Mar. 2014, pp. 60–78, doi:10.1080/14636778.2014.884456.

Harris, Anna, Sally Wyatt, et al. “The Gift of Spit (And the Obligation to Return It) How Consumers of Online Genetic Testing Services Participate in Research.” *Information, Communication and Society*, vol. 16, no. 2, Taylor & Francis, 2013, pp. 236–57, https://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.701656.

Hersh, Eitan D. *Hacking the Electorate: How Campaigns Perceive Voters*. Cambridge University Press, 2015.

---. “Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy by Daniel Kreiss.” *Technology and Culture*, vol. 58, no. 2, 2017, pp. 608–09, doi:10.1353/tech.2017.0066.

Hess, Aaron. “You Are What You Compute (and What Is Computed For You): Considerations of Digital Rhetorical Identification.” *Journal of Contemporary Rhetoric*, vol. 4, 2014, http://contemporaryrhetoric.com/wp-content/uploads/2017/01/Hess8_1.pdf.

Hindman, Matthew. “The Real Lessons of Howard Dean: Reflections on the First Digital Campaign.” *Perspectives on Politics*, vol. 3, no. 1, Cambridge University Press, Mar. 2005, pp. 121–28, doi:10.1017/S1537592705050115.

Igielnik, Ruth, et al. “Commercial Voter Files and the Study of US Politics.” *Pew Research*

Center Report, Washington, DC, 2018.

Improving Learning | Student Success | Canvas Platform. https://www.instructure.com/canvas/higher-education/empowering-faculty/improving-learning. Accessed 26 June 2019.

Ingraham, Chris. "Toward an Algorithmic Rhetoric." *Digital Rhetoric and Global Literacies: Communication Modes and Digital Practices in the Networked World*, IGI Global, 2014, pp. 62–79, doi:10.4018/978-1-4666-4916-3.ch003.

Inoue, Asao B. *Labor-Based Grading Contracts: Building Equity and Inclusion in the Compassionate Writing Classroom*. CSU OPEN Press, 2019.

Jackman, Simon, and Bradley Spahn. "Unlisted in America." *Unpublished Paper. Accessed*, 2015, http://images.politico.com/global/2015/08/20/jackman_unlisted.pdf.

Jankowski, Natasha A., et al. "Assessment That Matters: Trending toward Practices That Document Authentic Student Learning." *National Institute for Learning Outcomes Assessment*, ERIC, 2018, https://eric.ed.gov/?id=ED590514.

Jasanoff, Sheila, and Sang-Hyun Kim. "Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea." *Minerva*, vol. 47, no. 2, June 2009, p. 119, doi:10.1007/s11024-009-9124-4.

---. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. University of Chicago Press, 2015.

Johnson, N. "Modeling Rhetorical Disciplinarity: Mapping the Digital Network." *Rhetoric and the Digital Humanities*, books.google.com.

Juengst, Eric T., et al. "Personalized Genomic Medicine and the Rhetoric of Empowerment." *The Hastings Center Report*, vol. 42, no. 5, Sept. 2012, pp. 34–40, doi:10.1002/hast.65.

Jurgenson, Nathan. “View From Nowhere.” *The New Inquiry*, 9 Oct. 2014, https://thenewinquiry.com/view-from-nowhere/.

Justice, Glen. “The Nation; Howard Dean’s Internet Push: Where Will It Lead?” *The New York Times*, 2 Nov. 2003, https://www.nytimes.com/2003/11/02/weekinreview/the-nation-howard-dean-s-internet-push-where-will-it-lead.html.

Kay, Lily E. *Who Wrote the Book of Life?: A History of the Genetic Code*. Stanford University Press, 2000.

Kelly, John E., and Steve Hamm. “Handling Big Data.” *Smart Machines*, 2013, pp. 43–68, doi:10.7312/columbia/9780231168564.003.0003.

Kinshuk. *Designing Adaptive and Personalized Learning Environments*. Routledge, 2016.

Kitchin, Rob. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE, 2014.

Kolata, Gina, and Heather Murphy. “The Golden State Killer Is Tracked Through a Thicket of DNA, and Experts Shudder.” *The New York Times*, 27 Apr. 2018, https://www.nytimes.com/2018/04/27/health/dna-privacy-golden-state-killer-genealogy.html.

Kosinski, Michal, et al. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 15, Apr. 2013, pp. 5802–05, doi:10.1073/pnas.1218772110.

Kreiss, Daniel. *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. Oxford University Press, 2016.

Kuh, George D., and Peter T. Ewell. “The State of Learning Outcomes Assessment in the United States.” *Higher Education Management and Policy*, vol. 22, no. 1, OECD, 2010, pp. 1–20,

https://www.oecd-ilibrary.org/education/the-state-of-learning-outcomes-assessment-in-the-united-states_hemp-22-5ks5dlhqbfr1.

Lang, Susan, and Craig Baehr. "Data Mining: A Hybrid Methodology for Complex and Dynamic Research." *College Composition and Communication*, vol. 64, no. 1, National Council of Teachers of English, 2012, p. 172.

Lee, Sandra Soo-Jin. "Race, Risk, and Recreation in Personal Genomics: The Limits of Play." *Medical Anthropology Quarterly*, vol. 27, no. 4, Dec. 2013, pp. 550–69, doi:10.1111/maq.12059.

Legislative Tracking | Intranet Quorum 2018. https://www.intranetquorum.com/capitol-hill/legislative-tracking. Accessed 29 Oct. 2018.

Lohr, Steve, and Natasha Singer. "How Data Failed Us in Calling an Election." *The New York Times*, vol. 10, 2016, p. 2016.

Losh, Elizabeth Mathews. *Virtualpolitik: An Electronic History of Government Media-Making in a Time of War, Scandal, Disaster, Miscommunication, and Mistakes*. MIT Press, 2009.

Manovich, L., and M. K. Gold. *Debates in the Digital Humanities*. Edited by null, vol. null, 2011.

Matz, S. C., et al. "Psychological Targeting as an Effective Approach to Digital Mass Persuasion." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 48, Nov. 2017, pp. 12714–19, doi:10.1073/pnas.1710966114.

McKenna, Elizabeth, et al. *Groundbreakers: How Obama's 2.2 Million Volunteers Transformed Campaigning in America*. Oxford University Press, 2015.

Morabito, Vincenzo. "Big Data and Analytics for Government Innovation." *Big Data and Analytics*, 2015, pp. 23–45, doi:10.1007/978-3-319-10665-6_2.

Nelkin, Dorothy, and M. Susan Lindee. *The DNA Mystique: The Gene as a Cultural Icon*. University of Michigan Press, 2010.

Newman, Bruce I. *The Marketing Revolution in Politics: What Recent U.S. Presidential Campaigns Can Teach Us About Effective Marketing*. University of Toronto Press, 2016.

Nickerson, David, and Todd Rogers. *Political Campaigns and Big Data*. 25 Feb. 2014, doi:10.2139/ssrn.2354474.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

Nowviskie, Bethany. "What Do Girls Dig?" *Debates in the Digital Humanities*, U of Minnesota Press, 2012, pp. 235–40.

O'Neill, Cathy. "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy." *Nueva York, NY: Crown Publishing Group*, 2016.

Padawer, Ruth. "Sigrid Johnson Was Black. A DNA Test Said She Wasn't." *The New York Times*, 19 Nov. 2018, https://www.nytimes.com/2018/11/19/magazine/dna-test-black-family.html.

Pálsson, Gísli. *Anthropology and the New Genetics*. Cambridge University Press, 2007.

Panofsky, Aaron, and Joan Donovan. *Genetic Ancestry Testing among White Nationalists*. SocArXiv, 2017, https://osf.io/7f9bc/files/osfstorage/5995a3449ad5a1025322287e/.

Philippidis, Alex. "Patient Investment: GSK, 23andMe Forge $300 Million Alliance to Develop Drugs via Genetics." *Clinical OMICs*, vol. 5, no. 5, Mary Ann Liebert, Inc., publishers, Sept. 2018, pp. 8–10, doi:10.1089/clinomi.05.05.08.

"Reference Populations." *23andMe Customer Care*, https://customercare.23andme.com/hc/en-us/articles/212169298-Reference-Populations. Accessed 2 July 2019.

Research Studies - 23andMe. https://you.23andme.com/research/studies/human_api/. Accessed 2 July 2019.

Reyman, Jessica. "Copyright, Distance Education, and the TEACH Act: Implications for Teaching Writing." *College Composition and Communication*, vol. 58, no. 1, National Council of Teachers of English, 2006, pp. 30–45, http://www.jstor.org/stable/20456921.

---. "The Rhetorical Agency of Algorithms." *Theorizing Digital Rhetoric*, 2017, pp. 112–25, doi:10.4324/9781315203645-11.

---. "User Data on the Social Web: Authorship, Agency, and Appropriation." *College English*, vol. 75, no. 5, National Council of Teachers of English, 2013, pp. 513–33, http://www.jstor.org/stable/24238250.

Rice, Jenny. "The Rhetorical Aesthetics of More: On Archival Magnitude." *Philosophy & Rhetoric*, vol. 50, no. 1, Penn State University Press, 2017, pp. 26–49, doi:10.5325/philrhet.50.1.0026.

Richards, Gregory. *Big Data and Analytics Applications in Government: Current Practices and Future Opportunities*. CRC Press, 2017.

Ridolfo, Jim, and William Hart-Davidson. *Rhetoric and the Digital Humanities*. 2015, doi:10.7208/chicago/9780226176727.001.0001.

Rubinstein, Ira. "Voter Privacy in the Age of Big Data." *SSRN Electronic Journal*, 2014, doi:10.2139/ssrn.2447956.

Ruckenstein, Minna. "Keeping Data Alive: Talking DTC Genetic Testing." *Information, Communication and Society*, vol. 20, no. 7, Routledge, July 2017, pp. 1024–39, doi:10.1080/1369118X.2016.1203975.

Russell, Arlene A. "Calibrated Peer Review-a Writing and Critical-Thinking Instructional Tool."

Teaching Tips: Innovations in Undergraduate Science Instruction, vol. 54, National Science Teachers Association Arlington, 2004.

Salisbury, Lauren E. "Just a Tool: Instructors' Attitudes and Use of Course Management Systems for Online Writing Instruction." *Computers and Composition*, vol. 48, June 2018, pp. 1–17, doi:10.1016/j.compcom.2018.03.004.

Saukko, Paula. "Shifting Metaphors in Direct-to-Consumer Genetic Testing: From Genes as Information to Genes as Big Data." *New Genetics and Society*, vol. 36, no. 3, Routledge, July 2017, pp. 296–313, doi:10.1080/14636778.2017.1354691.

Schencker, Lisa. "Planning to Give 23andMe or AncestryDNA Kits as Gifts? Read This First." *Chicago Tribune*, Chicago Tribune, 14 Dec. 2018, https://www.chicagotribune.com/business/ct-biz-genetic-tests-for-christmas-1216-story.html.

Siemens, George, and Phil Long. "Penetrating the Fog: Analytics in Learning and Education." *EDUCAUSE Review*, vol. 46, no. 5, EDUCAUSE. 2011, p. 30,

Smith, M., et al. "Big Data Privacy Issues in Public Social Media." *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 2012, pp. 1–6, doi:10.1109/DEST.2012.6227909.

Solberg, Janine. "Googling the Archive: Digital Tools and the Practice of History." *Advances in the History of Rhetoric*, vol. 15, no. 1, 2012, pp. 53–76, doi:10.1080/15362426.2012.657052.

Spellings, Margaret. *A Test of Leadership: Charting the Future of US Higher Education*. US Department of Education, 2006.

Stamm, Stephanie. "How Pizza Night Can Cost More in Data Than Dollars." *WSJ*, Wall Street

Journal, 10 Apr. 2018, https://www.wsj.com/graphics/how-pizza-night-can-cost-more-in-data-than-dollars/.

Stevens, Hallam. *Life Out of Sequence: A Data-Driven History of Bioinformatics*. University of Chicago Press, 2013.

Stirland, Sarah Lai, et al. "Obama's Secret Weapons: Internet, Databases and Psychology." *Wired*, WIRED, Oct. 2008, https://www.wired.com/2008/10/obamas-secret-w/.

Strange, Bill. "Lindsey Graham Weighs In On Elizabeth Warren's DNA Test | Fox Sports 975." *Fox Sports 975*, iheart.com, 17 Oct. 2018, https://foxsports975.iheart.com/content/2018-10-16-lindsey-graham-weighs-in-on-elizabeth-warrens-dna/.

Stromer-Galley, Jennifer. "2012: Data-Driven Networked Campaigning." *Presidential Campaigning in the Internet Age*, 2014, pp. 140–70, doi:10.1093/acprof:oso/9780199731930.003.0006.

Stromer-Galley, Jennifer, and Andrea B. Baker. "Joy and Sorrow of Interactivity on the Campaign Trail: Blogs in the Primary Campaign of Howard Dean." *The Internet Election: Perspectives on the Web in Campaign 2004*, Rowman & Littlefield Lanham, MD, 2006, pp. 111–31.

Stuckey, Mary. "American Elections and the Rhetoric of Political Change: Hyperbole, Anger, and Hope in U.S. Politics." *Rhetoric and Public Affairs*, vol. 20, no. 4, 2017, p. 667, doi:10.14321/rhetpublaffa.20.4.0667.

TallBear, Kim. *Native American DNA: Tribal Belonging and the False Promise of Genetic Science*. U of Minnesota Press, 2013.

Teston, Christa. *Bodies in Flux: Scientific Methods for Negotiating Medical Uncertainty*. University of Chicago Press, 2017.

Toobin, Jeffrey. *Too Close to Call: The Thirty-Six-Day Battle to Decide the 2000 Election*. Random House, 2002.

United States. Congress. *Help America Vote Act of 2002: Conference Report (to Accompany H.R. 3295)*. 2002.

van den Berg, Ineke, et al. "Designing Student Peer Assessment in Higher Education: Analysis of Written and Oral Peer Feedback." *Teaching in Higher Education*, vol. 11, no. 2, Routledge, Apr. 2006, pp. 135–47, doi:10.1080/13562510500527685.

Volz, Tracy, and Ann Saterbak. "Students' Strengths and Weaknesses in Evaluating Technical Arguments as Revealed through Implementing Calibrated Peer Review™ in a Bioengineering Laboratory." *Across the Disciplines*, vol. 6, no. 1, WAC Clearinghouse. Colorado, 2009, https://wac.colostate.edu/ATD/technologies/volz_saterbak.cfm.

Warnick, Barbara. *Critical Literacy in a Digital Era: Technology, Rhetoric, and the Public Interest*. Routledge, 2001, https://www.taylorfrancis.com/books/9781135638283.

---. *Rhetoric Online: Persuasion and Politics on the World Wide Web*. Peter Lang, 2007.

Wintrup, Julie. "Higher Education's Panopticon? Learning Analytics, Ethics and Student Engagement." *Higher Education Policy*, vol. 30, no. 1, Springer, 2017, pp. 87–103, https://link.springer.com/article/10.1057/s41307-016-0030-8.

Yuji, Koichiro, et al. "23andMe and the FDA." *The New England Journal of Medicine*, vol. 370, no. 23, June 2014, p. 2248, doi:10.1056/NEJMc1404692.

Unit 3 Portfolio: **Data Policies & Design Rhetorics**

30% of Final Grade Portfolio Due Fri., May 3rd

The final portfolio consists of multiple short assignments, each of which asks you to analyze and interrogate the privacy and data collection policies and practices that inform our everyday digital lives. You will explore data collection and use in two contexts: 1) your personal, everyday data footprint and 2) the data policies and practices of a particular online platform. You will also learn key principles of visual and design rhetoric, and practice translating complex ideas to a public audience through the construction of an infographic.

Part 1: Taking An Inventory Of Your Personal Data Footprint

Due: Monday, April 8 at the beginning of class

Terms-of-use policies that describe data collection and use are required by law, but these are lengthy and difficult to understand when read at all. Even more problematic is the fact that everyday users are often led to believe that the data they contribute is a vital and even beneficial component of the services they seek from a given platform. To complicate these shared assumptions about the data collection practices and policies that undergird our everyday digital lives, you will each keep a log of your digital movement/activity over the course of **24 hours**. We will then work in class to unpack the various layers and types of data generated and collected during this time period.

Your internet use log may be submitted as a bulleted list, a spreadsheet, a table, or some other readable format. Because the volume of internet/internet-connected device use will vary from person to person, there is no set minimum or maximum requirement for how many entries you include in your log. However, you should strive to be as detailed and accurate as possible as your log your internet footprint. I will not be collecting your logs (that would just be too ironic). Rather, I'll go around in class and check for completion.

Part 2: Analyzing a Platform's Data Collection Policies and Practices

Due: Monday, April 15

Select a platform (Facebook, Google, Netflix, Amazon, etc.) to analyze. Write an informal explanation (~500 words) of that platform's data collection policies and practices, addressing the following questions:

- What are the default settings of this system?
- What modifications to these settings are possible?
- How easy are these settings to manage?
- What are the terms of using this system?
- What are users agreeing to when they use this service?
- How are these terms (and any changes to them) communicated to users?

Part 3: Discussion and Reflection

Due: Wednesday, April 24th

In an age when participation in so many life activities—including commerce, education, civic discourse, personal communication—require users to relinquish rights to their personal data and content, norms regarding the responsible and ethical collection, management, circulation, and use of content and data need to change. Although internet users have some degree of agency in choosing among available settings and services, these choices are quite limited, and are ultimately controlled by the technology developers and providers. Therefore, the most pressing change that needs to happen is with our shared expectations for how such systems are designed, and what accountability we expect from software companies and service providers to offer ethical systems.

Based on your work from the first two parts of this assignment, write a short (~500 words) reflective essay that addresses the following questions: How might we, as a culture, hold software developers and companies responsible and accountable for designing systems that enact a different ethic, one that consider users' privacy and ownership rights? What broader issues would we need to address in order to enact ethical data practices? What systemic/structural changes might you propose to help users become more responsible and active participants rather than passive consumers of platforms and services?

Part 4: Create an Infographic

Rough Draft Due: Friday, April 26 - BRING TO CLASS

Final Infographic + Annotated Reflection Due: Friday, May 3

For the final part of this unit, you will create an infographic that pulls together your findings from Part 2 and your argument from Part 3. You will decide the audience for your final production as well as the central goal(s) that you want your infographic to meet. Be sure to keep your audience, purpose, and context in mind when deciding how to approach the design and content of your infographic. You will have the opportunity to get feedback from peers on a rough draft of your infographic. When making your infographic, you may want to use a program like Canva or easel.ly to get started.

Annotated Reflection

In addition to your final infographic, you will be asked to submit an annotated version of your infographic that details the rhetorical choices that went into your production. This will allow you to account for your design choices, to explain how your infographic meets the needs of your audience and achieves your intended purpose, and to unpack the narrative of your production process. We will discuss the logistics of this annotated reflection piece during class.

Evaluation Criteria

Part 1: Digital Log - 10 points *(completion grade)*

Part 2: Platform Analysis - 20 points

- Does your explanation demonstrate a thorough understanding of the platform's data policies and practices?
- Do you speak to the framing questions included in the assignment description?
- Does your explanation of the platform's data practices follow a clear, logical order?

Part 3: Discussion and Reflection - 20 points

- Does your discussion/reflection clearly identify relevant issues involved in the current standard for platforms' user agreements?
- Do you use examples to back up your claims about the issues underlying most user agreements?
- Does your essay propose systemic/structural changes that adequately respond to those issues?
- Do you provide a clear explanation of how these systemic/structural changes would work to foster more ethical software/tech design and practices?

Part 4: Infographic - 50 points

Argument 20 points	Argument/central thesis is clearly presented to and appropriate for the infographic's audience; argument is clearly threaded through the entire infographic.
Organization & Clarity 10 points	Content and visuals offer a clear and logical path for the audience to follow; amount of information provided is appropriate to the infographic genre.
Cohesion 10 points	Visuals and text clearly connect and reflect the central theme of the infographic.
Design/Aesthetics 5 points	Design is visually inviting and asthetically appealing; design choices (font, color, imagery, format) reflect the main argument/theme of the infographic; any whitespace is intentional
Citations 5 points	Sources are provided where appropriate (can use any citation style; may include footnotes with links; in-text citations are used when quoting).

Notes

[i] It is important to note that Ridolfo has addressed issues of access and ethics in digital spaces in *Digital Samaritans: Rhetorical Delivery and Engagement in the Digital Humanities.* U of Michigan P, 2015. While Ridolfo does not explicitly address Big Data, his ideas can inform questions of access and ethics being asked in scholarship on Big Data.

[ii] In addition to providing candidates with access to its large membership (over one million members as of this year), DFA provides workshops and resources for candidates, campaigns, and activists to learn about, build, and enact effective grassroots campaign strategies. http://www.democracyforamerica.com/about_trainings

[iii] According to the 23andMe Main Research Consent Form: "Specifically, 23andMe Research refers to research aimed at publication in peer-reviewed journals and research funded by the federal government (such as the National Institutes of Health - NIH)."

[iv] As of this writing, 23andMe offers customers a number of different genetic testing services, including information about Genetic Health Risks, Ancestry, Wellness, Carrier Status, and Traits. As of August 2017, the "Genetic Health Risks" report and the "Carrier Status" reports are the only reports that include FDA approved genetic information.

[v] "Mobile data is not included unless a user accesses Canvas directly through a mobile browser, or if a user accesses content within the mobile app that redirects to a mobile browser."