PERSPECTIVES ON DATA FOR GOOD: THE EMERGENCE OF EMBODIED DATA DISCOURSES

by

Kimberly Gardner

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Public Policy and Administration

Boise State University

August 2023

© 2023

Kimberly Gardner

ALL RIGHTS RESERVED

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Kimberly Gardner

Dissertation Title: Perspectives on Data for Good: The Emergence of Embodied Data Discourses

Date of Final Oral Examination: 07 April 2023

The following individuals read and discussed the dissertation submitted by student Kimberly Gardner, and they evaluated her presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Jen Schneider, Ph.D.	Chair, Supervisory Committee
Stephen Crowley, Ph.D.	Member, Supervisory Committee
Eric Lindquist, Ph.D.	Member, Supervisory Committee
Michael Ekstrand, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Jen Schneider, Ph.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

DEDICATION

I dedicate this dissertation to my children, Ruth and Robert Gardner.

ACKNOWLEDGMENTS

I would like to acknowledge the National Science Foundation's Big Data Hubs and Spokes program members for providing funding and access for this research.

I would also like to acknowledge my family for their endless patience and support during the whole dissertation journey. This undertaking would not have been possible without the support of my parents, Gena and Michael Waller, who reliably stepped in to watch over my children when I needed time and space to write.

I also want to acknowledge my Committee Chair, Dr. Jen Schneider, for her endless patience, deep insights, guidance, and careful editing.

Finally, I especially want to acknowledge my husband, Stewart Gardner, who walked with me during this project, carrying more than his share of stresses, worries, and burdens but also sharing in the joy and excitement of ideas, inspirations, and small victories.

ABSTRACT

In policy sciences, data have traditionally been a tool used by scientists and technocrats to guide state policy. Boundaries around what counts as data generally fall along traditional understandings that data are neutral, objective, and abstracted from individual bodies and experiences. Unfortunately, this understanding of data has a history of perpetuating harmful social hierarchies and, especially in the era of "big data", mirroring our racial and gendered prejudices (Kitchin, 2014). More recently, however, data have been claimed as a tool by a different kind of actor operating in a unique environment. These new actors, such as some police officers and citizen activists, are negotiating and redefining who is considered a data expert and what we understand data to be.

These conversations between traditional and novel understandings of data can be seen within the data for good movement, where actors from a broad range of backgrounds and training come together for the purpose of advancing some notion of social good. Given the history of data perpetuating social harms such as racial discrimination, how can these relatively new understandings of data promote the social good while avoiding data harms? Or, how can data be used to promote the social good? Using the theoretical framework of Data Feminism, the data from participant interviews suggests that shifting understandings of data rely on the emergence of the concept of embodiment. This research highlights the differences in how embodiment manifests in two dissimilar sites: Measure Austin, a non-profit advocacy organization for people of color, and the Big Data Hubs program within the National Science Foundation. The findings suggest that data for social good presents as a space where data advocates negotiate between embodied and disembodied meanings of data and where embodiment is more significant for street level bureaucracy and citizen activists.

The dissertation suggests that “embodied data” offers an alternative to the predominance of market-driven data approaches. This research ends with a discussion for how policy studies could benefit from incorporating the concept of embodiment in research related to data systems, including artificial intelligence and machine learning.

TABLE OF CONTENTS

DEDICATION iv

ACKNOWLEDGMENTS v

ABSTRACT vi

LIST OF TABLES xi

LIST OF ABBREVIATIONS xii

CHAPTER ONE: INTRODUCTION 1

Data generations: A brief history 5

First Generation: Knowledge is power 5

Second Generation: Data are power 8

Third Generation: Embodied data 10

Table 1.1 Data Generations 12

The concept of embodiment 14

Social good 16

A rising tide lifts all boats 17

A seat at the table 18

Chapter overview 21

CHAPTER TWO: LITERATURE REVIEW 25

Critical Data Studies (CDS) 26

Definitions and boundaries 27

Data harms and mitigations 30

Data activism 34

Data Feminism and neoliberalism 35

CHAPTER THREE: METHODS 40

Research design 40

Modified Grounded Theory 41

Site Description: Big Data Hubs and Measure Austin..................................41

Measure Austin 47

Data collection 50

Sampling 50

Semi-structured interviews 51

Transcription 52

Iterative data analysis 53

Confidentiality 56

Limitations..57

CHAPTER FOUR: DATA ADVOCATES AND COLLABORATIVE START-UPS 59

The Data Advocate 60

Misfits 67

Collaboration and the Start-Up Model 69

The start-up model of collaboration 69

CHAPTER FIVE: DISEMBODIED AND EMBODIED DISCOURSES 87

Disembodied discourses 89

Disembodied voice: Authoritative and common...90

Disembodied sight: a god’s eye view 92

Opening up to embodiment: the “Big Data” revolution and playing with categories 95

Embodied discourses 97

Embodied voice: multiple narratives 99

Embodied sight: a drone’s eye view 100

Data discourses and strategies for cooperation: Boundary objects and data thermostats.. 101

CHAPTER SIX: THEORIZING THE SOCIAL GOOD 107

Social good: A rising tide lifts all boats.. 109

Social good: A seat at the table 117

Code Switching 122

CONCLUSION... 130

REFERENCES 135

LIST OF TABLES

Table 1.1 Data Generations 12

Table 2.1 Seven principles of data feminism 36

Table 2.2 Data for social good versus co-liberation 37

Table 3.1 Interview format 52

Table 3.2 Primary-cycle codes 54

Table 4.1 Dream team: Roles within collaborative environments 77

Table 4.2 Problem children 81

Table 6.1 Data for social good versus co-liberation 109

Table 6.2 Social good: A rising tide lifts all boats 112

Table 6.4 From data ethics to data justice 123

LIST OF ABBREVIATIONS

NSF	National Science Foundation
HUBs	Big Data Hubs
CDS	Critical data studies
DF	Data feminism
PE	Policy entrepreneur

CHAPTER ONE: INTRODUCTION

Society is awash in controversial technologies that can radically disrupt, harm, empower, and thrill. In only two months' time, ChatGPT, a chatbot mimicking human communication in writing essays and even telling jokes, (ChatGPT, 2022) has thrown universities into a scramble to set policy and provide guidance for teachers (Marche, 2022; Wingard, 2023; Sanchez, 2023), and fueled further speculations about language bots replacing white collar jobs (Lowrey, 2023; Burton & Confino, 2023; Marr, 2023). It is evident that technologies change our society in complex and unpredictable ways. Much of the public debate about emergent technologies often takes the form of a utopian/dystopian binary, framing questions in terms of how technologies will either destroy or improve society. This research attempts to avoid this binary, asking rather how data can be employed toward a social good.

Through a qualitative comparative case study of the data for social good movement within the National Science Foundation's Big Data HUBs and Spokes (HUBs) and the grassroots organization, Measure Austin (Measure), this research finds that actors in both sites (a term used to describe the area or field of focus) are negotiating meanings of data and the social good in ways that can potentially create the conditions in which political deliberation and meaningful relationships can form. I argue that one of the key concepts in their renegotiation of data and the social good is embodiment. This research highlights the differences in how embodiment manifests in Measure Austin and the

HUBs, suggesting that data for social good presents as a space where data advocates negotiate between embodied and disembodied meanings of data and where embodiment is more significant for street level bureaucracy and citizen activists. The dissertation suggests that "embodied data" offers an alternative to the predominance of market-driven data approaches. This research ends with a discussion for how policy studies could benefit from incorporating the concept of embodiment in research related to data systems, including artificial intelligence and machine learning.

One impediment to honestly assessing how data can be used for social good is the problem of the frame we use to explain our situation: we are implicitly asked to choose either a utopian or a dystopian narrative (Townsend, 2013; O'Neil, 2016). According to the utopian narrative, data and data technologies like algorithms that can predict behaviors are finally "smart" enough to fix problems that have long been intractable to society. Data and algorithms are imbued with a kind of superhuman or magical quality. For example, if we have enough data then we can predict crime and respond before the crime takes place thereby averting damage, suffering, and loss. Or, with enough data, mental health apps could effectively map and predict shifts in mood and recommend interventions that help boost mental well-being. An abundance of data also means that companies can predict individuals' consumer behaviors, which means that individual consumers will only see advertisements and news articles that apply to them, creating a direct connection between supply and demand with incredible accuracy and efficiency. The tech-utopian narrative paints a picture of data technologies liberating society from faulty and irrational human decision making.

The tech dystopian narrative, on the other hand, describes a society of trapped individuals, at the whim of "technologies-run-amok" or controlled by nefarious and powerful groups. For those who adopt the dystopian view, efforts to reduce crime through prediction will inevitably result in an erosion of rights. The mental health app will be used by insurance companies to raise health insurance premiums and deny access to healthcare. And the individualized link between consumer and supplier will result in "echo chambers" and political polarization. According to the dystopian narrative, data technologies will inevitably corrode democratic institutions and processes and exploit individuals, as the controversies surrounding Facebook and the 2016 presidential election illustrate (Solon, 2016).

The binary utopian/dystopian frame-choice leaves little room for human agency to thoughtfully respond and intervene. These versions of technological determinism, where social arrangements are determined by technology, assume that data and technology are autonomous or "have taken a life of its own as if it were an out-of-control Frankenstein monster (Hess, 1997, p. 124). According to technological determinism, the momentum is too fast; the technology too inscrutable; the path already set. The risk of ascribing to this idea is that we abdicate our ability to change how data is used. Either data will ultimately be for the social good, in which case, *we should get out of Facebook's way*. Or data will ultimately destroy the social good, and because our efforts are unlikely to have a significant effect, *we should get out of Facebook's way*. From this perspective, we forfeit our agency.

Furthermore, this dichotomy tends to run on the assumption that technologies are static. Data, from this perspective, means one thing to everyone because data is detached from socially constructed meanings. Data act on us but not the other way around. However, as the concept of co-production has shown, the relationship between data and people is a reciprocal one (Jasanoff, 2006; Hallberg & Kullenberg, 2019; Miller & Wyborn, 2020). Consequently, researchers could potentially miss the alternative explanation that as more people with diverse levels of expertise and backgrounds adopt data as a tool for making sense of the world and taking action, the meanings and understandings of data will be renegotiated and changed. Data is not simply a technocratic tool; it is potentially a democratic one. My research is interested in this dynamic.

My research investigates common and emerging meanings of data and the social good within the Measure and HUBs sites. To organize the findings of this research, I first situate data within a data generations metaphor. Like generations between children, parents, and grandparents, each new generation is unique but also carries with it some of the past. Discourses about what counts as data operate in a similar way. The generations can broadly be mapped onto different periods of time. Data comes with history, or as Trevor Barnes argues, big data comes with big history (Barnes, 2013). It is within these broader understandings that we see negotiations taking place. Second, I introduce two metaphors for organizing discussions around understandings of social good: data as "rising tide" that raises all boats, and data as the means by which the disenfranchised gain "a seat at the table." Finally, I introduce the concept of embodiment, which I argue, challenges assumptions of disembodied discourses found in the first and second generations of data and suggests "a seat at the table" version of social good. I end the introduction with an overview of the following chapters in the dissertation.

Data generations: A brief history

This section provides an overview of the data generations metaphor used to situate further discussions of disembodied and embodied data discourses found throughout my interviews. The expanded understanding of data from neutral and objective to embodied can roughly be mapped onto different historical periods, although these various understandings can, and often do, overlap throughout different historical periods. In this section, I will outline a data generations framework that illustrates the rough historical periods of three different *understandings of data along* with the *most prominent actors* and the *important attributes of data*. Although the data generations framework is not the focal empirical finding from this research, it is nevertheless a relatively simple, and therefore useful, heuristic used to organize my findings, which describe the constellation of different understandings of data and the most important actors involved in using data at my two study sites. I will start each generation by discussing the underlying assumptions at work. I will then detail the most important attributes of data and end by identifying the important actors. The first historical period can be characterized around the idea that, "Knowledge is Power", the second-generation shifts to "Data are Power," and the third-generation, I argue, shifts our understanding of data to "Data are Embodied."

First Generation: Knowledge is power

First-generation data discourses can be characterized by the idea that "knowledge is power." Within this understanding, the goal of knowledge is to rationalize and ultimately control nature and politics, both of which are imagined as chaotic. It rationalizes nature and politics, with the hope of imposing order on both. In the words of Machiavelli's Prince, knowledge produces "mastery over nature" (Machievelli, *The Prince*). Data serves as a tool in acquiring and building new bodies of knowledge. This kind of understanding is referred to by many scholars as an ideology of modernity. This is what Daniel Rosenberg means when he says, "The beginning of data's relationship to society begins with the beginning of modernity" (Rosenberg, 2013). This knowledge- is-power logic falls in line with James Scott's analysis of the role of legibility in modern states (Scott, 1998). According to Scott, legibility "is a condition of manipulation" (p. 183) that describes a state's efforts to arrange populations in ways that simplify society to make traditional state functions like taxation more efficient. An obvious example comes from the founding of the United States where the census was one of the primary concerns written into its Constitution in Article 1, section 2. One of the primary concerns of the new American government was to understand the population they were governing and to distribute political power based on that knowledge. The purpose of knowledge production, by collecting, analyzing, and interpreting data, is to understand what levers to pull in the particular circumstances one wants to change or maintain.

Within the "knowledge is power" logic, data are simplifications of messy political and social processes and patterns. Data help form the basis of knowledge that everyone more or less agrees on. Data are therefore thought of as pre-factual and given. This provides those in power with objective facts to inform policy decisions (Rosenberg, 2013). If we return to the census example, we imagine that counting does not change depending on who is in power. A person is a person. But despite the appearance of neutrality and complete objectivity, projects such as these have been and continue to be highly politicized and controversial (boyd, 2019). It should also be noted that this clause in the Constitution also held the infamous 3/5 Clause, counting only 3/5 of Black Americans as important for apportionment and tax calculations.

In order to accomplish projects of legibility, states needed far more information in the form of standardized metrics, thus increasing incentives to create stable categories and standard processes of data collection and analysis. According to the ideology of modernity and the dominant method of scientific investigations, categories formed necessary boundaries within which to test theories and ultimately provided foundations for building up society's knowledge about the human and natural worlds. According to the scientific method, data serves knowledge production in that it either confirms or disproves theories from the scientific community that seek to explain why certain things happen the way they do.

Although the ideas at play during these generational discourses are not bound by time periods, the data attribute of scarcity as well as the key actors are. When first generation data discourse first emerged, before the second-generation "Data Revolution" we have recently come into, data were often difficult to generate and expensive to collect, analyze and store. It was often a lack of data that presented the greatest problems to knowledge production. Consequently, government entities largely remained in control of knowledge production and this knowledge was put to use in the service of rationalizing large populations for the purpose of power distribution. And yet, because of the dearth and expense of data, governments were often limited in their capacity to exploit data.

Second Generation: Data are power

If the first-generation motto is, “knowledge is power,” then the second-generation motto could easily be, “data are power.” The shift from knowledge to data, I would argue, is possibly the result of the sheer amount of data now available and the computational power we have to interpret and use that data. Before data may have been one tool among many to acquire knowledge but it seems to be a tool that is taking ever more space in our toolbox, possibly displacing others, leading some to claim that our society is becoming “datafied” (Schäfer & Es, 2017). While many parts of the first-generation ideology have been carried over into the 2000s, key shifts have changed our understanding of what data are and what they can and should be used for. Many of these shifts come from the new physical reality of how the data are produced, collected, stored, and analyzed. In short, the growth of computational power largely accounts for our current data reality and has opened up new possibilities for how we understand data and its purpose. Furthermore, this new computational power has been largely exploited by big tech companies such as Amazon and Google to create new markets through individualized and targeted ads and recommendation systems. A new massive economic system around data has grown as a result.

Recently, we have seen an explosion in the volume of data produced and collected. As part of relatively mainstream social dialogue, we imagine data differently than first-generation discourses do. Data are now relatively easy to collect and they saturate our decision making on markets and policy. Data are often described as ubiquitous and “big”. Data have reached the size of petabytes, the root of which means, unfathomable (Gitleman, 2013). Metaphors such as “data exhaust” and “data tsunamis” or “data floods" describe society as saturated with computational data that is constantly being produced by our phones, internet searches, and wearable devices. In other words, this generation is characterized as “the data revolution” (Kitchin, 2014) and the age of big data (Mayer-Schönberger & Cukier, 2014). The sheer volume and ubiquity of data is one critical feature that distinguishes first and second-generation discourses.

The explosion of data that we produce within this generation is computational, or machine readable. And the ability to process extremely large amounts of data allows for modeling and simulations of incredibly complex systems and processes. It also means that many processes, such as hiring or the distribution of Health and Welfare benefits, have become automated with tech companies benefiting from the development and sale of their machine learning algorithms and recommender systems. These systems, the code for which are proprietary and “black boxed” (Mühlbacher et al., 2014; Metaxa et al., 2021), are then used by government entities with little oversight. The consequences of this lack of oversight have been documented to perpetuate existing biases in society. Virginia Eubanks describes this new phenomenon as “automating inequality” (Eubanks, 2018) and Cathy O’Neil, writing on the economic collapse of 2008 as a result of automated trading systems on Wall Street, refers to these automated tools as, “Weapons of Math Destruction” (O’Neil, 2016). During this second-generation of data, the government employees who used data systems for analysis were replaced with automated tools: algorithms. These algorithms are “fed” by huge volumes of data that are gathered, not by the government, but by large tech companies. Data-driven decision making or data-augmented decision making looks more like computational and automated decision making by machine instead of a human case worker.

The previous generation data discourses may have focused on “why questions” to explain where changes could be made but the second-generation often focuses more on the predictive value that data generate. This predictive capacity of data has called into question the relationship between data and research. According to some analysts (Anderson, 2008), scientists’ reliance on theories to generate knowledge is no longer necessary. Data are not tethered to a theoretical context; their sheer volume makes this impossible. Rather, data are purely mathematical and only later are placed in a theoretical context. Consequently, explaining why something works the way it does is not the point of data any longer. The purpose of data is merely to show that it works.

Third Generation: Embodied data

While 1st and 2nd generation understandings of data shared the underlying view that data are neutral, abstract and objective, I suggest that a third-generation discourse– emerging now, in some areas of practice– slightly shifts our understanding of data as embodied, a concept I discuss in more depth below. Data are understood in more democratic terms. Consequently, our expectations and policy demands of data have changed as well.

Many of the characteristics and promises of this new generation of data are as yet unknown; this dissertation attempts to map some contours of this emerging discourse. What we can see is a citizen and activist demand that the data being collected and used by private companies and government organizations like the NSA or local police departments be made open, available, and easily scrutinized by corporate and government watchdog groups. We also see communities with larger groups of mainstream activists targeting their activism at the data itself. Not only are activists demanding data transparency but they are also demanding software or algorithmic transparency (Koene et al., 2019). At the same time, government caseworkers are pushed out of the government jobs by automation, we also see attempts to bring the government back in as a major player in the data game. Government entities are experimenting with governance models, mostly public/private partnerships, to create financially sustainable ways for governments to use more of the data tools and technologies developed by the private tech industry. Furthermore, data scientists are pushing back on the claim that data are no longer tethered to theory. If data are to be used for a social good then context is key and the question of why something is happening has forced its way back into the data discourse.

As I argued above, we must not forget that first-generation data discourses of control are still with us. These are discourses that may ebb and flow but never completely disappear. For example, first-generation discourses rooted in controlling populations, or making populations “legible,” are alive and well in new technologies of digital identification. For two decades during the war in Afghanistan, the U.S. military kept a detailed biometric database, the Afghan Personnel and Pay System (APPS), of the members connected to the Afghan military. The database included incredibly sensitive information including facial images, iris scans, and fingerprints. But political power in Afghanistan largely runs on vast and complicated webs or networks of community and tribal connections. So the database also included a “genealogy” of tribal connections, with information on individual’s lineage as well as their connections to tribal leaders. In effect, the database rationalized the complexity of Afghan society for the purpose of militarizing it. Unfortunately, since the departure of the U.S. military, the Taliban may now have access to this information. If so, the task of the Taliban targeting reprisals against those who provided aid to the U.S. military may be much easier than it otherwise would have been (Guo & Noori, 2021).

Table 1.1 Data Generations

Frame	Perceived Attributes of data	Key actors
knowledge is power	objective neutral stable categories explanation scarcity disembodied	Governments
Data are power	Objective neutral prediction abundance computational	Tech industry and finance
Embodied data	“strong objectivity” abundant embodied	Government tech industry and finance non-profit and activist

First and second generations of data map onto ideologies of control and power, or rationalization of the state and the market. I argue that these generations rely on disembodied discourses. Although third generation embodied discourses challenge the first two, it is possible that the third generation has inherited some of the ideologies from the previous two generations. Organizations under the data for social good umbrella are potentially at risk of perpetuating social disparities supported and exacerbated by first and second generations.

Drawing on work in the subfield known as “data feminism,” my research investigates this problem by asking, what are opportunities for data to produce a more robust social good? Data feminism is the application of intersectional feminist approaches to the study of science and technology. Researchers Catherine D'Ignazio and Lauren Klein, authors of *Data Feminism* (2020) offer principles from which to investigate core questions about technologies, who they benefit, and how. I will detail these principles and their work in my own research in the literature review chapter. The purpose of this approach, as with all critical approaches, is to not only to point out potential problems with the use of technologies but more importantly to suggest how to make these technologies more democratic and of benefit to those at the margins of society. As such, this approach rejects utilitarian understandings of social good and justice which tend to emphasize the good of the greatest number rather than of the most vulnerable.

Researchers in the field of Critical Data Studies have exposed numerous breaches and betrayals of our current data- driven systems (Friedman & Nissanbaum, 1996; O'Neil, 2016; Brayne, 2017; Noble, 2018; boyd & Golebiewski, 2019), sometimes resulting in real-world policy changes (Thamkittikasem et al., 2019). Studies like these ask us to break down the technology and the data sets from which the technologies are built. This leaves us feeling uneasy about the possibilities for data to produce something positive. This dissertation addresses this gap. While most scholars in Critical Data Studies literature tend to look more at how existing and developing technical systems perpetuate social ills, by considering whether and how those systems could be a potential for social good, my dissertation explores how data for social good is expressed and employed in different contexts.

The concept of embodiment

The purpose of this section is to introduce the concept of embodiment, how it broadens our understanding of data, and its impact on the ongoing conversations around data and the social good. To say that data is just about counting things is to miss a great deal of what produces data and, in turn, what data produce. In other words, this definition of data misses the complexity of our relationship with data (Jasanoff, 2006). Critical geographer and critical data studies scholar Rob Kitchin describes the common understanding of data as "raw material produced by abstracting the world into categories, measures and other representational forms…that constitute the building blocks from which information and knowledge are created" (Kitchin, 2014, p. 1). Using this definition, we might believe that data are neutral, objective, and disinterested, with no real political or relational attachments that might cloud decision making. In the fields of public policy and administration, this understanding of data is often at play in the debates surrounding the scientific administration of the state (Wilson, 1887; Simon, 1976; Kettl, 2000; Kettl, 2017; Abma & Noordegraff, 2003; Pirog, 2014). For example, the data collected through the census is not often thought of as having any political loyalties, affiliations, or opinions. According to the common understanding of data, the census is simply a count of people in the United States derived from a statistical model rather than a count derived from political dramas and horse trading. It is a technical system rather than a political one. The concept of embodiment draws on the insights of Langdon Winner (1980) and Geoffrey Bowker and Susan Leigh Star (1999) that technology always has a politics and therefore challenges the idea that data is politically neutral and takes place outside and beyond politics.

Rather than situating data as politically neutral, the concept of embodiment asks that we embrace an understanding of data as created, gathered, stored, analyzed and used by human beings who are at all times subject to political and social forces. In introducing the concept of embodiment, feminist theorist Linda Martin-Alcoff defines it as, "…the idea that there is a constitutive relationship of the lived body to thought, to knowledge, and to ethics, taking leave of the modern idea that bodies can be left behind as the mind does its work" (Martin-Alcoff, 2013, p.1). The concept of embodiment means that we imagine data coming up from and out of physical bodies instead of descending from on high. Bodies are located in a particular time at a particular place where senses like smell, touch, taste, and sound are the primary ways we experience and know the world. Furthermore, our bodies contain not only the physical senses but also cultural expectations (Young, 1980; Fotopoulou, 2019). For example, a black body carries with it different cultural expectations than a white body.

The concept of embodiment developed in the early 19th century as a response and critique to traditional scientific methods of inquiry.

An example from current events illustrates what is at stake in understanding data-as-embodied. In 2016, ProPublica published findings on racial biases found in a risk assessment scoring algorithm used in the court system (Angwin et al., 2016). It looked at how a risk assessment score, an algorithm used by judges to determine a defendant's risk of recidivism, ranked defendants. They found that the algorithm labeled black bodies as posing a higher risk of recidivism than white bodies. In one case, the algorithm determined that an 18-year-old black woman, Brisha Borden, who had no previous criminal record, was at higher risk than a 41-year-old white male, Vernon Prater, who had a previous criminal record for armed robbery. He had been caught shoplifting from a Home Depot. Borden, on the other hand, had taken a child's bike for a "joy ride". Although it is not clear if the judge acted on the information from this score, what is interesting is that two years after the algorithm's initial prediction, Borden had not been charged with any new crimes but Prater was serving an eight-year sentence for breaking and entering and theft. In this case, and in many others, the algorithm's predictive power, imbued with racialized biases, proved incorrect and harmful, with serious real-world consequences that were unhelpful at best and damaging at worst.

The ProPublica investigation revealed the risk assessment algorithm showed significant racial disparities, such as incorrectly rating black defendants as future criminals at nearly twice the rate as white defendants. Decision makers assumed the algorithm was neutral and that data did not carry with it cultural biases and expectations about black and white bodies. They were wrong. Because data come from human society, data carry cultural expectations and biases into the datasets and into the algorithms trained on those data sets. This has far-reaching implications for policy makers and governments that decide to adopt and implement data driven technologies as well as the regulations needed to ensure tech companies do not perpetuate or even "optimize" some of our most pernicious biases and prejudices.

Social good

My research question asks how data can be used for the social good. Practitioners vary in their use of the term social good. Throughout interviews, at least two different visions of social good emerged. For my research, I try to acknowledge this distinction while maintaining the integrity of data advocates' own views of what constitutes social good. A lack of clarity around this concept can lead to confusion and competing goals, even if all individuals and organizations involved are operating under the "data for social good" banner, as both organizations in this research understand themselves to be. The purpose of this section is to explore the different understandings of social good by introducing two broad and dominant views present in my research: "A rising tide lifts all boats" and "A seat at the table."

A rising tide lifts all boats

"A rising tide lifts all boats" is one dominant metaphor in our society that works to make sense of and justify social good. It implies that we all have our own individual boats. These boats can differ dramatically from one another; one can be an old, leaky kayak and another could be a giant yacht. From our own individual and radically different boats, we are all subject to rising and falling tides. For the purpose of this research, the most important insights from this metaphor are that for something to be a social good:

1. It does not need to increase a person's or community's access to democratic decision-making. The people in the leaky kayak need not be invited to spend time with the people on the yacht. And they need never have anything but the leaky kayak.
2. It does not suggest the necessity for human agency, at least not any agency from those outside locations of political power; the tide will rise whether those outside locations of power cast a vote or not. A data for social good project, under this metaphor, could consist of a collaboration between a university's computer

science program and the city's department of transportation to create a cleaner public transit system that results in less pollution in the air, which ideally benefits the greatest number of residents possible.

A seat at the table

In contrast to the previous metaphor, having "a seat at the table" implies individuals or communities sitting together at one table where decisions regarding the community are made, rather than separate tables of different size, beauty, and location. This other dominant metaphor suggests that for something to qualify as a social good, it must consist of a mechanism to increase participation from members of the community, particularly those who do not generally have access to political power. This understanding of social good, similar to theories of participatory democracy (Irvin & Stansbury, 2004; Beer, 2009; Ruijer et al., 2017), would shift the transportation and cleaner air example above to include participation from communities where pollution is already an issue.

Although both metaphors can describe different approaches to data for social good, the consequential difference is political power: who has it and who does not. According to the first understanding of social good, changing a group's position in society, from having less political power to more political power, is not significantly relevant to affecting a larger social good. It assumes that a more efficient transportation system with vehicles that emit fewer pollutants will benefit everyone, regardless of their education levels. What this understanding neglects to address is that those with less education will probably not benefit *as much* from policy changes as those with higher socioeconomic status, and they potentially never will. Without including communities who do not often have access to political power, a social good project is at risk of "data colonialism" (Couldry & Mejias, 2019; Ricaurte, 2019; Segura & Waisbord, 2019) and could run counter to democratic ends, whether or not this was the intention of the data for good organization (Brown, 2017).The concept of embodiment also informs a more robust version of social good, aligning more with the seat at the table metaphor than the rising tide metaphor of social good. This more participatory form of decision making includes community members more often left out of formal decision making processes. A version of social good influenced by the concept of embodiment favors the lived experiences of those affected by policy. Because of the differences between the HUBs and Measure sites, the concept of embodiment plays different roles in each. The next section provides a brief overview of findings before moving on to an overview of each chapter.

This dissertation uses a qualitative comparative design to understand and critique data for good projects within the National Science Foundation HUBs, a national government funding agency, and Measure Austin, a non-profit focused on criminal justice and education reform. As a relatively new phenomenon, data for good is ripe for critique which requires a "thick description" for understanding new phenomena. My research suggests multiple avenues for producing a social good. But as with all technologies, context matters. The lessons learned from comparing the two sites suggests that Measure Austin provides an example of how embodiment impacts efforts in data for social good. Its particular strength comes from a concrete understanding of social good as well as third-generation "embodied data" understanding of data that works alongside first and second-generation discourses. Furthermore, Measure Austin seems to suggest that data is more than just information; it potentially acts as a boundary object (Bowker &

Star, 1999) around which data advocates from organizations traditionally in conflict with each other are able to “lower the emotional temp” of the room and achieve some level of cooperation, and in some cases, develop friendships that help build and maintain trust as more contentious conversations come to the fore. In this case it may be that the relationships matter more than the data when producing a social good; it is not necessarily the data itself that produces a social good, but the relationships built upon a shared zeal for both the data and for collaboration. Measure Austin data advocates also face challenges in using data to produce a social good. Achieving a social good with the use of data, particularly in a collaborative environment, requires a baseline of trust between groups. It is not apparent that once that trust is established between traditionally contentious groups that data will be able to maintain it. If organizations are seen as having their own political agenda, disputes about the quality of the data collected and its interpretation can be undermined even when operating under a more participatory, seat at the table, version of social good.

The lessons learned from the Measure Austin site suggest ways in which we might critique the HUBs’ use of data for social good rhetoric. Data for social good in the HUBs context is at risk of reproducing social disparities or missing the mark of social good when the social good is ill-defined and when funding mechanisms incentivize the adoption of business models of financial sustainability. These challenges are particularly salient for the NSF where its organizational structure and history encourage the challenges listed above.

Chapter overview

My research contributes to the ongoing conversations about data and the social good within Critical Data Studies (CDS) by marrying it with the work on embodiment found in digital and data feminism. The literature review chapter pays particular attention to the concept of embodiment and how this concept contributes to the growing awareness of how data impacts our current social and political realities, particularly within the fields of public policy and administration, which do not engage with the concept of embodiment in data. This work offers an important opportunity to bring CDS into public policy and administration contexts, where the application of data systems appears to be proceeding without enough forethought or reflection.

I explore the potentials of data for social good by using a qualitative comparison design with a modified grounded theoretical framework. This chapter includes examples of interview questions and details the process by which key sensitizing concepts came out of interviews and site visits. In-depth case studies allow a researcher access to data that is rich, layered, and complex. The strength of case studies lies in this approach's ability to permit new themes and concepts to emerge from the perspective of the individual actors most involved in creating the phenomenon (Luker, 2009). In the case of data for good, vast amounts of money are being spent through organizations like the National Science Foundation's BD Hubs and these organizations have tremendous power in setting the agenda for other data for social good programs in universities and non-profits.

Before moving on to an analysis of embodied and disembodied data discourses, I introduce the actors and the environment that make embodied discourses possible. This chapter is meant to accomplish two things: it provides a description of the actors, which I call data advocates, and the collaborative environment within which the actors of this research are operating. For the purposes of this research, data advocates are elites or leaders from multiple sectors of society who are actively involved in applying data, data science, and data technologies to solve social issues. Not all data advocates are trained in computer science but all share an enthusiasm for using tools of data collection and analysis. As such data advocates are important in championing agendas of data driven decision making and data driven policy making in governments, non-profits, and activist organizations. I devote a chapter to describing the environment and the actors within it to provide the necessary context for the following substantive chapters.

To explore the possibilities of data producing a social good, we first have to understand what we mean by data and how we envision what data is supposed to do. My research shows two seemingly opposed understandings of data: disembodied data, where data are objective, and embodied data, where data are always tethered to physical bodies. The first discourse characterizes data as sitting above politics, values, and emotion. The alternative discourse sees data as situated in value-laden contexts and elevating emotion. This concept of embodiment is a key concept for analyzing data for social good projects. Data are always collected "from the ground," and therefore always political. One finding that emerged from interviews is that embodied discourse is relatively new and is used in tandem with discourses of disembodiment, which have their roots in first-generation data collection and analysis. This could indicate a possibility that disembodied discourses as well as embodied discourses could both be utilized to produce a social good, but at the very least they often persist together, even though they may be inherently in conflict.

This chapter discusses the different understandings of social good between the Measure Austin and NSF BD sites. My findings suggest that data advocates from the Measure Austin site more clearly define social goods in terms of policy gains for minoritized groups as well as a “do no harm” principle. Data advocates from the NSF Hubs, on the other hand, tend to articulate social goods broadly and largely allow end users to define their own version of social good or social impact. This chapter then analyzes why these different articulations of social good make sense in each site and how that could impact the potential for data producing a social good. The chapter ends with a discussion of how data can produce a social good by focusing on the positive lessons learned from the Measure Austin site and applying those lessons to the NSF BD Hubs and suggesting how the NSF could incorporate some of the practices from the Measure site to bolster their data for social good efforts. I also discuss potential challenges for the Measure site, paying particular attention to how using the languages from the first-generation of data technologies can impact social good efforts, how data collection efforts might be used by other powerful organizations, and the challenge of maintaining trust in highly contentious policy areas. Ultimately data as simply information does not produce a social good. However, the relationships created through the use of data can facilitate movement toward producing a social good. Maintaining that trust and those relationships will continue to present challenges in data for social good efforts.

The concluding discussion details the limitations of this research, focusing specifically on the limitations of small sample sizes required for an in-depth investigation. Future research would be needed to shed light on testing the effectiveness and prevalence of concepts such as lowering the emotional temperature of the room as well as embodied language with larger and possibly experimental research designs. The utility of such concepts could help organizations to structure collaborative social good projects in more effective and meaningful ways that bring more community members to the table in a participatory process that challenges existing power structures.

CHAPTER TWO: LITERATURE REVIEW

I situate my own research in the emerging field of Critical Data Studies (CDS) whose goal it is to understand data in terms of its relationship to power structures and hierarchies. Much of the CDS literature deals with how data have been used as a weapon against communities at the margins. More recently, this literature, recognizing that technology can be empowering as well as disempowering, is exploring how data can be used to mitigate harms and empower the marginalized. By using the relatively new theoretical framework of data feminism, my research adds to the growing literature on how data can be used to empower communities by comparing how differently situated groups understand the concept of data for social good. Literature in public policy and administration (PPA) would greatly benefit from incorporating perspectives and insights from CDS that analyze data and technology as a part of a larger political and social context, a perspective often lacking in PPA, especially regarding data and big data policy discussions (Jarmin & O'Hara, 2016; Kettl, 2017).

This chapter provides an overview of the relevant CDS literature on defining data, documenting and mitigating "data harms", and data activism. I then turn to a discussion of how data feminism can contribute to the ongoing discussion within CDS and end by showing how public policy and administration could deepen knowledge of the role of data in the policy process and within policy making institutions by incorporating work from CDS and data feminism that critically examines the impact of data within society.

Critical Data Studies (CDS)

CDS emerged as a critical response around the hype of the perceived revolutionary disruption of “big data”. This hype sought to cast big data as radically different from small data, effectively ushering in a new era with new ways of building knowledge. In a national bestselling book, Victor Mayer-Schonberger and Kenneth Cukier (2014) of *The Economist* declared big data to be a revolution that will transform how we live, work, and think. The authors claimed that, “not only is the world awash in more information than ever before, but that information is growing faster. The change of scale has led to a change of state. The quantitative change has led to a qualitative one” (p.6). The vast quantity of data changes markets, governing, and even seems to undermine the role of causality in science (Anderson, 2008).

Victor Mayer-Schonberger and Kenneth Cukier package these shifts as inevitable and beyond our control. Following the insight from Langdon Winner that technologies have a politics (1980), CDS scholars would dispute that framing, asserting instead that technologies are not autonomous from society or politics and that data are, “a form of power” (Iliadis & Russo, 2016). At stake is the recognition of data’s political power and its impact on the individual, or “the ways in which data are generated, curated, and how they permeate and exert power on all manner of forms of life (Ilidadis & Russo, 2016, p.2). CDS frameworks attend not only to how data shape society but also to the possibilities for human agency to interrupt and shape data and their interpretation (Dalton et al., 2016). By explicitly rejecting the inevitability of data technologies (often considered a form of technological determinism), CDS literature critically examines the collection, use, and framing of data with the explicitly normative goal of investigating the ways in which people can participate and intentionally shape the development and use of data.

Definitions and boundaries

One question that arises in this regard concerns the difference between data and big data. Starting from the premise that data are a form of power and therefore shape and are shaped by political and social forces, CDS responds to the exaggerated hype of big data by contextualizing the technology within existing social and political structures and histories to explore the ways in which practices around big data may or may not differ from past small data practices. By engaging with definitions and boundaries of data, CDS offers concepts and categories that inform my research on actors' different understandings of data.

CDS scholars maintain that in crucial ways the distinction between data and big data is not necessarily as significant –issues associated with small data are still with us. A great of deal of work in CDS explores the boundaries and definitions of (big) data, asking questions about big data epistemology and methodology (Milan & Trere, 2019). dana boyd & Kate Crawford (2012) remind researchers that big data is a fluid concept that will change the way we define knowledge. They state that we must therefore, "Ask difficult questions of Big Data's models of intelligibility before they crystallize into new orthodoxies" (p. 666). Kitchin (2014) disputes the supposedly revolutionary overthrow of the traditional scientific method (Anderson, 2008). Data and their meanings always come from a context. Kitchin acknowledges that recognizing patterns inductively is useful, but denies that this is most effectively done without the prior guidance of theories, hypotheses, and models. “The only way of tackling big data is to know what you are or may be looking for " (Kitchin, 2014, p.437).

While big data may not be an example of a Kuhnian paradigm shift, it does create new opportunities and vulnerabilities through insights previously unavailable, particularly in the arena of markets (Floridi, 2012). Expanding on this insight, Robin Wagner-Pacifici, John Mohr, and Ronald Breiger (2015) point out that the agents revealed in big data are conceptually different from those of traditional small-n statistics. Where mainstream social science “liquifies everyone in a homogenous soup despite how much they varied individually,” big data’s granularity reveals that we are all normally deviant (p. 6). Nevertheless, despite the large-N characteristic of big data and the granularity of analysis it allows, including its temporal variability, big data remains a spatially situated or contextualized science (Kitchin & McArdle, 2016; Dalton & Thatcher, 2015).

CDS scholars are also interrogating the myth of revolutionary overthrow by situating big data in history. Trevor Barnes and Matthew Wilson argue that denying the history of science and technology is a, “classic modern move. The past is ignored because nothing should hamper or limit what is to come” (Barnes & Wilson, 2014, p. 1). They claim that big data is merely a continuation of the social physics movement developed by William Warntz, which grounded on the belief in large scale data sets, artificial computing power, and mathematical modeling constitute knowledge. Rieder & Simon (2016) argue that big data is successfully expanding into the public sector under the guise of “evidence-based policy making and decision making” because society has progressively transferred their trust from flesh and blood individuals to faceless institutions and finally to disembodied data (Rieder & Simon, 2016). This insight of where we place our trust and whether data can assist in building a sense of trust between adversarial groups becomes crucial in the question of whether data can produce a social good, as I discuss in detail in chapter 6.

In many cases, the "bigness" of the data is not the most relevant point. This is mirrored in my own research, where the bigness of the data is often discussed and the label "big data" is used, but many of the issues discussed by my participants, such as who should own data or how to communicate the data with diverse stakeholder groups, cut across small and big data practices (Fotopoulou, 2021). For example, if data for a data set were collected with suspect methods or metrics, then the use of that data set, whether big or small, is not as relevant.

Rather than organizing research around big and small data, Kitchin and Tracey Lauriault (2015) provide a useful concept, the data assemblage, as a way to organize research and policy questions around data. The growth of big data plus the development of digital data infrastructures leads to numerous questions of how power is shaped and expressed in the nature of data, how those data are produced, organized, analyzed and used. Drawing on the works of Michel Foucualt (1995), Susan Star and James Griesemer (1989), and Geoffrey Bowker and Susan Star (1999), Kitchin and Lauriault (2014) suggest that the data assemblage "encompasses all of the technological, political, social and economic apparatuses and elements that constitutes and frames the generation, circulation and deployment of data" (p. 1). The data assemblage is defined as "a complex socio-technical system, composed of many apparatuses and elements that are thoroughly entwined whose central concern is the production of a data (p. 6).” These data and assemblages are mutually constituted and bound by context. The concept of the data assemblage offers a way to think about data that encompasses big and small data structures and practices. Insights into the novelty of “big” data and its difference from “small” data show us that as data become more embedded in society and expand in use, possibilities for abuse against communities and opportunities for empowerment also expand. Ultimately, the bigness of computational data systems both stem from and contribute to some of the same issues as “small” data..

Data harms and mitigations

By analyzing the “darker side of data assemblages” CDS scholars are documenting how data are employed in ways that produce pernicious effects on society: dataveillance and the erosion of privacy; profiling and social sorting; predictive profiling or what they label, anticipatory governance; and control creep (p. 8). This line of work is categorized as “data harms and mitigations”. One of the first issues to be addressed by CDS scholars under data harms is the computer code used to build the software of algorithms and recommender systems. Given how pervasive, banal, and paradoxically mythical algorithms and recommender systems have become, CDS scholars have been shaping a research agenda around the software side of big data assemblages. Some scholars are attempting to map the ethics debate around algorithms and artificial intelligence (Mittelstadt et al., 2016; Etzioni & Etzioni, 2017), and virtual reality (Egliston & Carter, 2021). Scholars in fields of education are also adopting CDS frameworks as they attempt to incorporate the analysis of power within projects of data literacy (Sander, 2020; Pangrazio & Selwyn, 2021). Other scholars attempt to “open the black box” of the algorithm, with the idea that increasing transparency is a fundamental democratic good (Christin, 2017; Berk et. al., 2021; Anqwin et. al., 2016). As a data analyst turned data activist, Cathy O'Neil (2016) labels black box algorithms as, "Weapons of Math Destruction". Not only are the algorithms opaque to anyone without a Phd in computer science, but they are also devoid of due process for the redress of possible mistakes. If a teacher is fired for scoring low on an algorithm-driven assessment, that teacher has no legal means of addressing her accuser. The "algorithm" did it, and we often perceive data to be neutral.

This is a problem, according to Cathy O'Neil (2016) because the algorithm lacks a feedback loop where mistakes of the algorithm's output are fed back into the algorithm as a corrective. Responding to this concern, some CDS scholars return to the black box metaphor and suggest developing research platforms that open and audit algorithmic systems (Metaxa et al., 2021).

The research agenda to hold algorithms accountable through audits and transparency is more complicated when technology is analyzed from the perspective of the assemblage, or from the perspective of how the technology was shaped and is shaping the socio-political context. Making use of his concept of data assemblage, Rob Kitchin (2016) situates the algorithm as the performative software that overtly and covertly shapes individuals' and groups' lives in society (p.6). Because algorithms are part of the assemblage, the kind of transparency required to "open the black box" of algorithms goes far beyond the technology itself. For example, if researchers are investigating the values written into source codes that create algorithms, access must be granted to the coding teams as those codes are written. This is made more difficult by the fact much of the source code is privately controlled. Furthermore, as Christian

Sandvig (2016) and Cathy O'Neil (2016) remind us, algorithms never simply stand alone but are entwined with many other algorithms, constituting an algorithmic system, where "authorship is collective, made, maintained, and revised by many people with different goals at different times" (Kitchin, 2016, p. 7). Complicating a critical examination of algorithms is their "ontogenetic nature", or that algorithms are constantly becoming and are never static or fixed. There exists no mechanical machine to grab hold of and open. Many algorithms are self-learning, or change their own code in response to more data that change the categorization of "opacity" (Burrell, 2016). Furthermore, algorithms, once introduced to the public, go through domestication, where users change the purposes of the algorithm once it is embedded into everyday use (Kitchin, 2016, p.6).

The complexity around transparency does not end with understanding the code of the algorithm. CDS scholars are not in complete agreement regarding transparency (Levy & Johns, 2016; Annany & Crawford, 2016). The idea of transparency as simply a democratic public good can easily be weaponized or used to reassure the public that the government is doing all it can to solve the problem (Levy & Merritt Johns, 2016; Pozen, 2018). Mike Ananny and Kate Crawford (2016) question whether the ideal of transparency is even possible within the "black box" understanding of algorithms. The problems associated with transparency include:

- Transparency can reveal corruption but if no action follows, public cynicism or even wider spread corruption may follow (p.6).
- Transparency can be harmful when it exposes vulnerable groups of people (p.7)

- Transparency invokes neoliberal models of agency or the market model of enlightenment where the burden of collecting and making sense of information falls on the individual (p. 8).

Instead of accountability through transparency, they suggest a process of constructive engagement. This is a key insight for policy-makers, non-profits, and scholars interested in using data for the social good.

One major implication of the increased use of algorithms is the tendency for these technologies to optimize the status quo, or existing power hierarchies that are supported by negative prejudices of people of color, in some cases rendering people of color invisible. Joy Buolamwini, a researcher at MIT, documented how facial recognition software did not recognize her face until she wore a white plastic mask (Buolamwini & Gebru, 2018). Other scholars have shown how algorithms automate inequalities, hurting the poor and most vulnerable (Eubanks, 2018), or how algorithms continue regimes of oppression (Noble, 2018) including “the new Jim Crow” (Benjamin, 2020).

CDS scholars have documented extensive evidence that shows racial bias in algorithmic systems. The societal implications of racial bias in data and algorithmic systems are significant, particularly in the fields of medicine and criminal justice. Algorithms used in telehealth have shown a propensity to consider people of color in need of less care than white people (Vyas et al., 2020) and to misdiagnose cancer in people of color at higher rates than whites, denying access to crucially needed health care interventions. In the field of criminal justice, Legal scholar Andrew Ferguson called into question the viability and accuracy of “predictive policing” (Ferguson, 2016). Scholars have also shown how problematic historical crime data feeds racially biased algorithms that saturate low-income neighborhoods with police presence, often resulting in stop and frisk policing (Goel et al, 2016). These biased data sets also feed criminal risk assessment scores that result in higher incarceration rates for people of color than their white counterparts (Angwin et al., 2016; Eckhouse et al., 2019).

Data harms extend beyond optimizing existing biases and prejudices regarding gender and race into the heart of the political process. Fake news, bots, voter suppression, and filter bubbles produce a great many social harms that threaten the integrity of the democratic process as the well-documented controversy of Cambridge Analytica illustrates (Shah, 2018). CDS scholars continue to document the ways in which our own long held prejudices and tendencies are written into code and mirrored back to us in our data. Data for good, as it manifests in different contexts, is subject to these same risks like perpetuating existing social hierarchies.

Data activism

Just as data for good is susceptible to reproducing existing social hierarchies such as racism, it is also the case that data can be used in ways that increase democratic participation. The field of data activism operates on similar premises as CDS regarding the contextuality of data, but it shifts the focus from criticism of harms to how data can be used positively to empower communities. Data activism combines concepts from Science and Technology Studies with ideas from Social Movement Studies (Milan & Velden, 2016) to explore the potential of using data for social good. Data for good projects can be described as a kind of activism"that utilizes the data infrastructure as an enabling method" (Gutierez, 2018, p. 21). The literature in data activism focuses on the possibilities for using data to promote social good through promoting increased participation and holding governments accountable. Data activism is both a social practice and a theoretical construct that is used as a “heuristic useful to explore how people engage politically with big data and massive data collection” (Milan & Gutierrez, 2015).

Of particular importance to my own research is the insight that data activism began as an elitist sub-field within grassroots activism but has become more diffused to include non-skilled actors. This is consistent with my understanding of the concept of the *data advocate,* which I describe in detail in chapter 3. Many data advocates, although not skilled in the methods of data science, ascribe to the authority of data to clarify issues and convince others.

Data Feminism and neoliberalism

Both CDS and data activism analyze data as a form of power that is involved with the processes and hierarchies of society. However, this dissertation proceeds from the premise that more work in CDS needs to be done unpacking specifically how data for good projects fit into this analysis. For example, a CDS perspective would interrogate the ways in which the social good label could be manipulated by those in power to maintain the status quo. Only a handful of researchers have analyzed data for good, and then primarily to point to the vagueness of the data for good label or the superficiality of its appeal to potential funding sources (Catlett & Ghani, 2015; Hooker, 2018; Williams, 2020).

The present research fills this gap in CDS by employing the theoretical framework of data feminism to compare and contrast the two dissimilar sites of Measure

Austin and the NSF HUBs with regard to their understanding of data and the social good. The remaining sections will describe the principles of data feminism in more detail and how those principles can be used to interrogate data technologies and understandings of social good, paying particular attention to the concept of embodiment and how the concept of embodiment challenges neoliberalism. This contributes to the work in CDS that is attempting to describe the changing and expanding role of data in society.

Data feminism uses seven guiding principles to analyze power structures. All seven principles make up an approach that informed my research, but the most important for my research is the concept of embodiment. The importance of embodiment emerged from my interviews with data advocates and the data feminism literature later helped me to theorize it. The table below shows the seven principles and their definitions.

Table 2.1 Seven principles of data feminism

Examine Power	Data feminism begins by analyzing how power operates in the world
Challenge Power	DF commits to challenging unequal power structures and working toward justice
Elevate emotion and embodiment	DF teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living-feeling bodies in the world
Rethink binaries and hierarchies	DF require us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression
Embrace Pluralism	DF insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, indigenous and experiential ways of knowing
Consider Context	DF asserts that data are not neutral or objective. They are products of unequal social relations, and this context is essential for conducting accurate, ethical analysis
Make labor visible	The work of data science, like all work in the world, is the work of many hands. DF makes this labor visible so that it can be recognized and valued.

(D’Ignazio & Klein, 2020, p. 140)

In their book introducing data feminism, D'ignazio and Klein aim part of their analysis at "data for social good" projects and show how the principles of data feminism challenge us to think more subtly about this vague term. Data feminism proposes a way to think about data for good as data for co-liberation (Table 2.2). This way of thinking about social good is both broader and more concise than other ways of viewing social good. Some would consider a data project to be a social good project if it sought to match data scientists with non-profits. Considered from the perspective of Data Feminism, the "matchmaker" version of social good is inadequate; it risks perpetuating existing power structures by following a charity model of social good and does not ultimately change power dynamics. Data projects with co-liberation as the goal must be specific about existing power dynamics and must seek to include traditionally marginalized voices (D'Ignazio & Klein, p. 141).

Table 2.2 Data for social good versus co-liberation

	Data for Social Good	Data for Co-Liberation
Leadership by members of minoritized groups working in community		X
Money and resources managed by members of minoritized groups		X
Data owned and governed by community		X
Quantitative data analysis is "ground truthed" through a participatory, community-centered data analysis process		X
Data scientists are not rock stars and wizards but rather facilitators and guides		X
Data education and knowledge transfer are part of the project design		X
Building social infrastructure is part of project design		X

(D'Ignazio & Klein, 2020, p. 140)

Different groups showing interest in the concept of social good do not share a common definition, an issue I will address in detail in chapter six of this work. What is relevant here is that this ambiguity opens the data for social good movement up to forms of market co-optation.

This is particularly salient with regard to neoliberalism where, according to scholars who study this phenomenon, one version of how society functions and ought to function has a tendency to obscure and /or co-opt any alternatives.

The danger, according to critic of neoliberalism Wendy Brown (2017), is not simply that democratic institutions are corrupted by "dark money" (Mayer, 2016) but rather that political reasoning and political character are changed into economic ones. In other words, neoliberal logic economizes non-economic spheres. According to Brown, neoliberalism is a kind of logic. "A normative order of reason developed over three decades into a widely and deeply disseminated governing rationality, neoliberalism transmogrifies every human domain and endeavor, along with humans themselves, according to a specific image of the economic" (Brown, 2017, p. 10). Furthermore, neoliberalism is a form of reason that actively configures all aspects of existence in economic terms (Brown, 2017, p. 17).

The role of data supporting neoliberal logic in the policy process is explored by Deborah Stone who, instead of using the neoliberal term, describes a similar form of reason as the market model. Data used in service to the market model, according to Stone, is part of the effort to rationalize the policy process and politics more broadly by, "rescuing public policy from the irrationalities and indignities of politics" (Stone, 2012, p.9). Under the market model, data, as numbers and measurements, abstract from indignities such as human emotion. Borrowing from Stone's analysis, I refer to this type of data throughout my research as disembodied, and I argue that embodied data offers an alternative to the market model and thus a challenge to neoliberalism.

Although work has been done analyzing what evidence-based policy making consists of and the role of data within government process (Heitmueller et al., 2014; (Höchtl et al., 2016; Jarmin & O'Hara, 2016; Kettl, 2017; Ingrams, 2019; El-Taliawi et al., 2021), literature in policy and administration does not theorize how embodiment could impact how data is used and viewed within policy making and administrative processes. While Stone's work does capture the notion of contextuality, or a kind of general embodiment, it does not give sufficient attention to the particular context of located human beings that embodiment does. My research begins to address this gap by applying the concept of embodiment to evidence-based and collaborative processes at a national funding agency level and at the local government level. Embodiment challenges neoliberal logic in three ways: it expands 1) narrow definitions of who is included as an expert, 2) ways of imagining and understanding data, and 3) the way we understand social good. In other words, the actor, the environment, data, and social good are all renegotiated and formed in part by the adoption of embodied discourses, which the three findings chapters will address in detail.

CHAPTER THREE: METHODS

Research design

This research uses a comparative qualitative case study approach to develop a description of the landscape around the relatively new phenomenon of “data for social good” to answer the question of how data can be used to produce a social good. In-depth case studies allow a researcher access to data that is rich, layered, and complex. The strength of case studies lies in this approach’s ability to permit new themes and concepts to emerge from the perspective of the individual actors most involved in creating the phenomenon (Luker, 2009). Ultimately, the goal of a qualitative case study is to arrive at a “thick description” of the phenomenon under study (Tracy, 2012). To achieve a more nuanced understanding, this research compares two dissimilar cases who understand themselves to be using data for the social good in some way. These cases are referred to as sites, a term used to describe the area or field of focus (Tracy, 2012). The first site, National Science Foundation Big Data Hubs (HUBs), is a federally funded institution with considerable institutional and expert resources. The second site, Measure Austin, began as a grass roots, citizen activist organization. After a discussion of how I have used modified grounded theory and a preliminary description of the data advocate, I describe each site in detail, followed by a discussion of my participant role in each site and the access this has allowed me during the study. I then describe the methods used and end with a discussion of the limitations of this study.

Modified Grounded Theory

Grounded theory is understood as an approach that is inductive and driven by the data as opposed to organizing and analyzing data through the lens of a pre-existing theory (Charmaz, 2006). The approach is used when investigating a new phenomenon with little existing data or theoretical frameworks. The strength of the approach is theoretical and conceptual development, particularly in areas with little existing data. The data for good phenomenon qualifies as a relatively new development, particularly from the grassroots perspective. My research therefore develops theory while introducing new concepts such as code switching and data thermostats. But there is also existing work on data and society coming out of CDS, as my literature review suggests. My approach, therefore, can be described as using a modified grounded theory approach, which brings collected data and its emergent themes into conversation with existing concepts borrowed from other theoretical frameworks. These "sensitizing concepts" are "interpretive devices that serve as a jumping off point" (Tracy, 2012, p.28) and attune or sensitize researchers to certain concepts at the beginning of the data collection and analysis process. In other words, a modified grounded theory approach does not dictate that data be analyzed within a pre-existing theoretical framework, nor does it leave the researcher without solid ground from which to begin.

Site Description: Big Data Hubs and Measure Austin

The data advocates identified in this research are affiliated with either the National Science Foundation (NSF)'s Big Data Hubs network or Measure Austin. Both sites share the same goal: to advance the use of data and data science. Both sites also share the view that this goal can best be achieved through cross-sector collaborations.

Data advocates, a term I use to describe actors in my sites who are leaders from multiple sectors of society who are actively involved in applying data, data science, and data technologies to solve social issues such as poverty - a concept I develop in detail in chapter 4, come together in collaborative infrastructures to share technical and/or substantive expertise.

It is not uncommon for data advocates in the academic or tech industry sectors to volunteer their expertise to less experienced data advocates in the government sector. The stated goal of the NSF Hubs is to play the matchmaker between data advocates and these sectors.

The stated goal of Measure Austin is similar in its commitment to data and collaboration, to bridge divisions through research and public education in active partnership with local communities to address complex social problems. Measure began with a focus on policing and communities of color in Austin, TX, but has expanded to address complex social issues such as the school to prison pipeline and healthcare disparities among women of color.

I chose these two sites because my previous work with the West Big Data Innovation Hub granted me access to the wider Hubs network and to Measure Austin. My early interactions with these groups suggested the two cases would indeed offer a study in contrasts while also providing insight into how data and society are being co-produced in new ways.

National Science Foundation Big Data Hubs group (HUBs)

In 2015, the National Science Foundation created the Big Data Hubs network with the purpose of advancing the use of data science through collaboration with multi-sector actors from diverse academic departments, tech industry, governments, and nonprofits. The interdisciplinary and multi-sector focus is relatively rare and new for the NSF. Beginning in 2013, formal work groups and informal discussions led to a focus at the NSF on the promises of "Big Data" tools and technologies, particularly the promise to governments of greater efficiency and lower costs. The problem, as identified by some Hubs data advocates, was framed as a disconnect between the data tech industry and local, state, and national governments. Data science and technology existed but the government was slow to adopt. It would take a matchmaker of sorts, or as one HUBs data advocate described, a data Yenta, to increase adoption across a wider group of organizations. Consequently, the focus of the BD Hubs was on collaborative strategies rather than primary research. (This focus on collaboration over primary research later created confusion for many applying for NSF funding in the BD Hubs program and by those at the NSF evaluating those proposals). According to conversations with HUBs data advocates, only by matching tech solutions with governments in need of those solutions would data science and data technology adoption increase.

The solicitation for the first round of BD Hubs (Program Solicitation NSF 18-562) in March 2015 came out of NSF's Directorate for Computer and Information Science and Engineering (CISE) and called for four regional BD Hubs – Midwest, South, Northeast, and West. The regional distribution was justified early on as a way to make the Hubs more accessible to groups who wanted to participate in Hubs events. The general thinking was that the social issues that big data could address with the greatest impact would probably take place at the state and local levels, so geographical proximity, based on the Census Regions of the United States, would encourage projects that used data science toward solving those local and regional problems. The following year, all four regional Hubs solicited planning proposals to pilot future Spokes activities. The success of the planning grants was determined by the quality of Spokes proposals submitted the following year.

According to several NSF Hubs data advocates, in addition to a regional focus, the BD Hubs concept was also an attempt to incorporate a “tech start-up” approach to big data and data science adoption. This meant that NSF program directors and other leaders avoided any top-down, “thou shalt” commandments and embraced a diversity of methodological approaches and perspectives from each region of the Hubs. According to NSF Hubs data advocates, the strength of the tech start-up approach would be in its flexibility to adapt to these diverse approaches and follow those with the greatest promise and traction. Furthermore, this flexibility would allow for sectors traditionally not involved with data science to engage in a meaningful way. For example, one Hubs data advocate pointed out that an issue like building a smart city infrastructure cannot be accomplished by any one sector alone. The challenge was to find a means by which multiple sectors might work together on solving smart city challenges. Somewhat reminiscent of the Silicon Valley ethos to “fail fast”, the Hubs as tech start-ups could play and experiment with how those collaborations would take place. Each Hub therefore adopted its own diverse portfolio of projects and activities.

By November 15 of 2015, the NSF released its Big Data Spokes (BD Spokes) solicitation and by September of 2016, 10 Spokes and 10 spokes planning grants were awarded. As initially envisioned by the Hubs leaders and program directors, the spokes were to work closely with the Hubs in identifying and defining specific collaborative projects where data science could play a key part in the solution. The partnerships built through these collaborative spokes projects could be thought of as nodes. By the end of 2016, the BD Hubs and spokes infrastructure was created.

From the perspective of many NSF Hubs data advocates, the tech start-up approach allowed each of the Hubs a great deal of autonomy in its governance structure. Each Hub decided on its leadership team, the role and background of executive directors, deputy directors, the size of its steering committee and its staff. This created some diversity between the Hubs.

Initially the structure of the network was supposed to resemble a conceptual wheel, with inter-reliant hubs and spokes. As the initial round of funding played out, though, the Hubs and spokes interacted far less than originally envisioned and the image of Hubs and spokes transformed into something more resembling an orbit with various degrees of gravitational pull. For example, some PIs of spokes and planning projects kept in close contact with Hubs leadership team throughout their projects, relying on Hubs infrastructures and existing personal networks. In my own involvement with the West Big Data Innovation Hub, my research team relied heavily on connections with local police departments that the West Hub leadership already had in place. Spokes projects would often dictate the extent to which they wanted involvement with the Hubs as PIs in the spokes were individually funded by the NSF and autonomous from the Hubs.

The Nodes: Data advocates working within the Hubs network fade in and out. For example, a non-profit group may have worked closely with the Hubs during a “hackathon” event, creating a lasting project, but then unplugged or faded out from the Hub after the completion of the project or the piece of the project that was of direct concern to that non-profit. The concept therefore of network is fitting in the sense that not all participants in the Hubs are “members”, which would imply an in or out, 0 or 1, demarcation. Rather, participants in this site choose when and where to participate for a variety of different reasons and therefore fade in and out of the network which made tracking extremely difficult,

The purpose of the Hubs is to “play the matchmaker” between data scientists in academia and tech industry on the one hand and governments and nonprofits in need of data science technologies and techniques on the other. This matchmaking role remains explicitly broad, at least in the first “start-up” phase, in order to attract the widest possible audience. One of the challenges of a data-focused network is to advertise events and activities in a way that attracts those with no experience in data science and also to attract data scientists who feel comfortable with a collaborative and interdisciplinary approach. The Hubs accomplish this by playing with language in their advertisements and heavy use of social media sites like Twitter. This is one of the reasons why the Hubs collaborative environment is composed of data advocates who may not necessarily be skilled in data science or even basic data analysis. Nevertheless, as data advocates, they believe that data can provide unique solutions to pre-existing and long-lasting problems and work with the NSF Hubs in the hopes of gaining data science expertise

Since the purpose of the Hubs is to advance data science through collaboration, a majority of efforts for the data advocates at leadership levels is the creation and maintenance of the relationships necessary for projects to move forward, which means that a lot of time is dedicated by Hubs leadership to phone calls, planning meetings, events coordination, maintaining websites, and sending updates.

Measure Austin

One of the more interesting developments in the wider application of data is the adoption of data methods and technologies by activist and non-profit groups. Much of technology innovation and adoption is spearheaded by national agencies such as the NSF or the NIH. I have selected an activist/non-profit site as a comparison group with the HUBs group because it is relatively rare for a grassroots activist group to spearhead data driven efforts. During a conversation with a potential HUBs partner in the field of policing, I was made aware of Measure Austin, an activist group in Austin Texas who was working closely with the Austin police department to build community trust with better data practices. Here I offer a detailed description of the Measure Austin site.

In 2015 Austin Police Chief Art Acevedo of the Austin Police Department attended the announcement of the Obama Administration's White House Police Data Initiative. The purpose of the initiative was to build trust between communities of color and police departments by using data science to inform policing practices and open data policies to hold police accountable to the public. The same day Chief Acevedo was in D.C, Jameila (Meme) Styles, member of the Austin Justice Coalition, stood on a stage with the Austin City Mayor in an attempt to make sense of yet another fatal shooting of a young unarmed black man by an Austin PD officer.

The audience, composed of members from social justice groups, such as Black Lives Matter, and Austin community members, demanded that the Austin Police Department recognize its own role in perpetuating racism through its training programs and policing practices. According to the citizen groups present, this tragedy, like many before it, was the fault of the city and its police. In response, the mayor argued that efforts in “community-oriented policing” were starting to address the problem in a productive way. Styles responded with a suggestion that surprised those on stage and the audience: “Show me the data.” The quality of the data held by the police department and the discrepancy between it and the lived experience of the cops and the members of the community meant that the data on community-oriented policing, and thus the efforts in community-oriented policing, could not be trusted. Meme Styles created Measure Austin in 2015 with the express purpose of collaborating with police officers and members of the community to create metrics of community-oriented policing that could be owned and trusted by the police and the minority communities they served. Two years later, Measure Austin introduced community-developed metrics and a training program to the Austin Police Department.

In 2018, Measure Austin held its first annual Big Data and Community Policing Conference which brought together police officers from around Texas, community groups, academics, and a new police organization dedicated to increasing the use of data and evidence in policing – The American Society for Evidence Based Policing – who also held their first annual conference that same year.

Located in Austin, Texas, Measure started as a citizen-activist grassroots organization relying on unpaid volunteers. It is now a registered non-profit with six employees. Its governance structure is much simpler than that of the HUBs, consisting of a president, vice president, chief research officer, chief financial officer, research development assistant, and a director of community engagement and partnerships. Measure Austin is led by Styles, who is founder and president.

Measure’s stated goals are more concrete than the HUBs. Their primary commitment is to equity. Thus the focus of Measure is more on social justice than simply social good. Although the organization began its work in the field of criminal justice, it now works in the areas of education justice, health justice, and economic justice.

Participant Observer

My access to both sites evolved out of my work as part of a research team on the Big Data Policing Planning grant. I had led three workshops over the course of a year and attended three All Hands Meetings, and consequently had worked with members of the West Hub throughout the year. One of the workshops was held in Austin, TX, with Measure Austin’s Big Data and Community Policing Conference. I have continued to check in with Measure’s work and NSF Hubs. As a participant in both groups, I often feel invested in the success of each group and advocates from both groups have expressed an interest in mine. One challenge with conducting interviews with Hubs and Measure Austin participants was situating my level of expertise and explaining exactly the goals of my research. My expertise is not that of a data scientist or even of a data advocate. Every participant in the HUBs and Measure groups has to negotiate this space to some extent, but I felt that in my case, the fact that I was present but not a data advocate made my involvement awkward at times. Furthermore, although I align with efforts to decolonize methods of data collection and its use, I recognize that my position as a White, educated, middle class, cis gendered woman influences my interactions with my participants in ways that may reproduce existing social hierarchies.

Data collection

Sampling

My research question necessitated that I use sampling practices representative of a larger phenomenon (data for social good), rather than sampling representative of a population (Luker, 2009. p.103). I conducted purposive sampling for my initial list of interviews with data advocates. This method of sampling is widely used in qualitative research (Tracy, 2012) and identifies appropriate individuals based on the parameters of the project's research questions, goals, and purposes" (Tracy, 2012, p.134). For this study, the parameters of the sample included leaders from the formal structure of each organization. For the NSF Hubs and Measure Austin, leadership was identified via my previous work with both organizations. Not all potential participants from the initial purposive sample agreed to be interviewed. This generated a relatively small sample of potential participants.

Snowball sampling was then used to expand the size of the sample and provided the additional benefit of revealing the data advocates' professional networks, both formal and informal. This is critical to my research, which asks about the development of "data for good" within collaborative environments. The collaborative environments in both sites function by using both formal and informal structures and relationships. In snowball sampling, participants are asked for recommendations of other data advocates with whom they have worked and who would be able to lend their own expertise about the topic. This, in effect, reveals an "organic" social network.

Recommendations came from almost all advocates, and in some cases, participants recommended advocates I had already interviewed. This revealed a tight network of partners in both the BD Hubs group and the Measure Austin group. In the end, I interviewed 22 individuals.

Semi-structured interviews

I conducted semi-structured interviews with 22 data advocates. Interviews lasted between 30 and 90 minutes with most lasting around 60 minutes. The questions I developed prompted participants to reflect on their opinions and experiences related to my research question: What do different groups mean by data for social good? I asked questions such as, “Why is data applied more broadly now than 10 years ago?”, and “Several data advocates have said collaboration is an absolute necessity to using more data effectively. What are your thoughts on that?” I also asked follow-up questions tailored to each conversation. The interview guide provided consistency across my interviews with both sites, which better allowed me to compare the sites. The guide was also flexible enough to allow for an organic conversation to develop when appropriate. Unfortunately, maintaining a balance of structure proved a challenge in many cases, particularly when the interview shifted into a more emotional space. For example, many data advocates wanted to discuss the role that gender played in the success of collaborative data projects, even when the question was not part of my initial interview protocol. However, these “tangential” discussions often led to important insights in the study, as I discuss in the findings chapters.

IRB consent forms were sent via email attachment to all data advocates before the interview began. Signed consent forms were either attached to my email and sent back or sent as a picture to my personal cell phone. Documents were then printed, signed by me, and then sent back via email to the data advocate. Of 22 interviews, 20 agreed to a recorded interview. I typed detailed notes during the interviews of the two cases where data advocates declined the recording. Unfortunately, two recordings also failed during the interview without my realizing although hand notes taken during the interviews were saved. The process for transferring data from the recorder to a secured Dropbox account immediately following the interview resulted in realizing my mistake immediately following the interview. I endeavored to recreate as much of the original conversation as possible by writing down what we had discussed. Due to the regional spread of data advocates in my sample and the prohibitive cost of travel, I conducted 19 interviews over the phone and two over Skype, and only one interview in person. Although these mediated interviews did not provide some of the details present in face-to-face interviews such non-verbal and embodied data, they did allow greater flexibility for scheduling, which presented one of the greatest access challenges. Most interviews had to be rescheduled at least once before the interview actually took place.

Table 3.1 Interview format

Interviews conducted over the phone	19
Interviews conducted over Skype	2
Interviews conducted in person	1

Transcription

I manually transcribed all recorded interviews myself verbatim. For my first six transcriptions, I listened to my audio recorder and typed into a Word document on my laptop. This resulted in detailed transcriptions that included, for example, all pauses, hesitations and laughter. Later, I purchased a transcription pedal and transcribed the interviews into the Express Scripts program, which sped up the transcription process. Because a more detailed transcription is not necessarily better than a less detailed one

(Tracy, 2012) and I had analyzed five of the six transcripts which resulted in emerging codes and sensitizing concepts, I stopped transcribing pauses, hesitations, and laughter for the last 10 data advocate interviews. Transcriptions were then moved from Express Scripts program and stored in Nvivo software on my own password-protected computer and on to a secure Dropbox account. I then assigned each participant an identification number and deleted names from my transcripts.

Data Analysis

Iterative data analysis

I used an iterative approach to data analysis, where I alternated between an inductive or grounded analysis and a deductive analysis of applying existing theories and concepts (Tracy, 2012). The strength of the iterative approach is that the researcher continually revisits her data and recasts her analysis as the picture of the phenomenon becomes clearer. This meant that throughout the data analysis process I continued to both code new interviews and revisit my older transcripts. This also meant that while collecting new data, I was developing analytic codes, or codes that move beyond description to analysis, and connecting those codes with existing theoretical constructs such as experimental work on how data might or might not lower emotional responses (Marcus et al., 2008; Coleman & Wu, 2010; Sumartojo et al., 2016; Amrute, 2019), theoretical concepts from literature on surveillance economies (Zuboff, 2019) and neoliberal critiques (Brown, 2017). Below, I provide a detailed description of my descriptive and analytic codes as well as examples from my codebook and an analytic memo (Appendix A).

In this first part of my analysis, I printed the interviews and word-by-word, line-by-line, generated primary cycle codes. Coding by this method helps the researcher avoid imposing her own “motives, fears, or unresolved personal issues” onto the participants and the data (Charmaz, 2006, p. 133). It is also an effective tool for recasting the familiar in a new light. Using data to solve social problems can be framed as nothing out of the ordinary. Organizations and governments have been using data for centuries to inform their decisions from the census (boyd, 2019) to “evidence-based policy-making (Shine & Bartley, 2011). The problem with glossing over “data for social good” as yet another iteration of using data like any other is that it obscures any potential differences or novelties in the process of developing, using, and thinking about data and data-intensive technologies and their relationship to existing social structures and processes.

Primary cycle coding revealed descriptions of what was happening and who was involved. Many of the primary codes for this research are in my own words, for example, “shifting identity,” where the professional identity of an individual is replaced with the data advocate identity. But as much as possible, I attempted to use the participants’ own language, or *in vivo* codes. For example, data advocates used the term, “bridging” to describe why data is considered an effective language between contentious groups. Below I have included an excerpt from my codebook of descriptive codes. I have deleted the examples column to protect the identities of my participants.

Table 3.2 Primary-cycle codes

Primary code	description	comments
Identity shifting	Events where individuals from different sectors think of themselves primarily	This possibly suggests the presence of a boundary organization rather than something like a policy network. Even though some examples exist of policy

	as "data lovers or advocates" instead of their sector.	changes and purpose is generally spoken about as policy intervention at some level, a key component seems to be this identify as a data lover rather than a cop or an activist. This identity shift can be a threat to a cop's job.
Lowering emo temp/distancing from the individual	Data, especially data storytelling, is seen as a tool that allows individuals from different sectors, even those traditionally opposed to each other, to work together more effectively by making the conversation less emotional – it's about the pattern rather than individual story	Data are effective at changing culture and policy because it can remove personal and emotional characteristics and feelings from the conversation. However, on multiple occasions of "data storytelling" this role of data is somewhat played down. Emotions are fanned and heightened in some story contexts such as generating community outcry against an existing policy or practice. The relationship between the personal/emotional/anecdotal and data is complex.
SG=relative	Social good is however your partner defines it.	This is similar to the political philosophy of liberalism – which suggests that the only way to deal with different definitions of something as volatile as "the good" is to attempt to remain as neutral as possible. It has been critiqued that neutrality in the good is not possible and only glosses over the exercise of market power and those who have access to the market. Does this translate into data for social good as relative?

The initial codes provided my research with sensitizing concepts to pursue in further data gathering and analysis. It was also informed by interview protocol and shifted my research away from its initial focus on networks and toward a focus on data for social good. I organized my initial codes under conceptual themes or analytic codes by paying attention to commonly occurring themes. Secondary or analytic codes serve as explanations of why something is happening by identifying patterns in the data and comparing those concepts with existing theoretical frameworks. In this stage of analysis, researchers move beyond description and attempt to explain what is happening. Patterns of data are then connected to existing relevant theories which then re-inform the data which has already been collected as well as further data collection efforts. For example, the data collected and analyzed at this stage of my research suggested the relevance of theories of neoliberalism (Brown, 2017; Zuboff, 2019) and boundary organizations (Bowker et al., 2015) resulting in analytic codes such as: “Data as boundary object”, “Metric construction as demos”, and “community building as demos”. Analytic codes were further developed by the use of analytic memos.

Analytic codes were explored and developed through the writing of analytic memos which are used to reflect and compare the emerging data between interviews. In this way, analytic memos provide guidance for further data collection and analysis, and ultimately for developing theory or explaining a particular phenomenon as opposed to remaining in a descriptive posture.

Confidentiality

Both the Measure site and the HUBs site are relatively small where members have worked together and know each other. This creates a risk where participants may be able to identify each other and absolute confidentiality cannot be guaranteed. This was explained to participants. To protect confidentiality as much as possible, I only identify participants as data advocates from either the Measure site or the HUBs site. Consequently, HUBs participants are not identified at any other level than HUBs data advocates, obscuring whether HUBs participants are staff or awardees. While this created a quandary between transparency and confidentiality, this research proceeded earring on the side of confidentiality.

Limitations

One limitation of this research is that a small sample size of 22 participants makes applying these findings to other situations difficult. However, as a qualitative researcher I am interested in generating theory or what Luker calls, a project of discovery, rather than testing existing theory (Luker, 2009. P. 125). It is possible that findings from this research are generalizable to a certain phenomenon, emerging embodiment in data discourses. However, all the participants interviewed perceive data as a legitimate and authoritative tool. Perceptions around how data can be used to promote the social good may then be limited to actors predisposed to using data in the first place.

study.

CHAPTER FOUR: DATA ADVOCATES AND COLLABORATIVE START-UPS

Measure Austin and HUBS are characterized by two new and interesting phenomena: the emergence of the *data advocate* as an important kind of actor, and a collaborative "start-up" environment. Both phenomena contribute to making space for a broadened understanding of who counts as a valued expert at the table and why. I refer to this as part of a third-generation environment. And yet, as the generations metaphor suggests, actors mostly characterized as operating within third generation discourse still carry over traits from the first- and second-generation discourses.

I attribute the broadened understanding of who counts as an expert to the greater attention paid to embodiment in the third generation, which, in general, attends more to context, going as far as to consider individual lived experience. Embodiment elevates the individual's lived experience as its own kind of expertise, valued in data for good projects. Both the actors and environments can be considered as a third-generation manifestation of the relationship between data and society. Because of the structure and difference in purpose between my two sites, embodiment shows up more in Measure than in the HUBs.

In their book *Data Feminism*, Catherine D'Ignazio and Lauren Klein (2021) argue that data expertise has historically been defined in terms of what I describe as first and second-generation discourses or ideologies. In earlier data generations, experts in the field of data science consist of those formally trained in data analysis and visualization, reproducing a culture that understands data as objective and politically neutral, devoid of judgment and even human emotion. Because data is often considered an abstract representation of something out in the world, it is by definition separated from the body and therefore separated from an individual's perception of their own experiences. Embodiment challenges the first generation's threshold of expertise because data advocates are considered expert, not simply when formally trained in data science, but because they have lived experience or knowledge that makes them uniquely qualified to participate in guiding the whole data collection and analysis process.

In the first part of this chapter, I will explain these two phenomena. The phenomena elaborated in chapters five and six–embodied discourses of data and a seat-at-the-table version of social good–will prove to depend upon the existence of the data advocate and the collaborative environment. I next discuss some typical obstacles faced in collaborative efforts toward the social good, one of which can be associated with disembodiment. In the latter part of the present chapter, before moving on, I will discuss some issues for collaborative networks that arise from traits inherited from the first- and second-generation discourses.

The Data Advocate

Whereas in the earlier generations the primary advocates for data were government officials and industry leaders – both exhibiting a top-down orientation—the present third generation is marked by the emergence of a broader group of actors within collaborative environments. The actors I interviewed come from a far wider range of locations than in the past, including activist organizations, non-profits, street-level and more senior administrators in government agencies, and groups of citizens–essentially, what we might think of as those making up civil society. Many are not trained in data science, but still qualify as experts in some respect. I call these actors *data advocates*.

I define data advocates as leaders from multiple sectors of society who are actively involved in applying data, data science, and data technologies to solve social issues such as poverty, police use of force, lack of equitable transportation, or access to safe drinking water. The data advocates from my sites are typically highly educated and often politically active, yet often fit the broader interpretation of expertise that typifies third generation discourse, with its greater attention to voices from different locations, and from different lived experiences.

The data advocate is similar to the concept of the policy entrepreneur (PE) in policy studies literature (Kingdon, 2011; Mintrom & Norman, 2009; Fowler, 2022) and is used within various policy studies frameworks, including: policy streams, institutionalism, punctuated equilibrium, and advocacy coalitions to explain how and why policy changes. PEs are, "advocates who are willing to invest their resources—time, energy, reputation, money—to promote a [policy] position in return for anticipated future gain in the form of material, purposive, or solitary benefits" (Arnold, 2021). The participants in this research are certainly "willing to invest their resources" to promote the increased use of data tools and techniques for the social good. The PE concept is often employed in the context of a specific policy domain, for example, in healthcare (Cohen & Horev, 2017), in carbon pricing (Narassimham et.al., 2022), or urban growth (Ramirez et. al., 2022). It is also the case that the PE is a concept used to explain policy change.

However, the actors I am describing are not promoting a specific policy gain or policy change. Data advocates are not promoting policy domain specific changes. Data can be used across all sectors and within all policy domains. The data advocate is involved in creating a data culture, or changing the status quo culture of government and activist organizations to be more supportive of data driven practices. Furthermore, I use the concept of data advocate in lieu of PE to underscore the role of embodiment in broadening expertise in data driven processes.

The concept of the *data advocate* is not widely used in critical data studies or policy studies. By developing the theoretical construct of the data advocate, this chapter contributes a concept that unifies disparate descriptions of actors who are building future data infrastructures. This chapter introduces this new concept into the literature and identifies data advocates' self-understanding as actors who are loyal to data--those who believe in the power of data to contribute to a social good–from diverse backgrounds disrupting the status quo but who also understand their identity as potentially isolating, turning data advocates into "misfits."

There are just a few places in the literature where something like the notion of a data advocate is discussed. In the field of Educational Technology, scholars have developed the concept of the educational data advocate strategy to increase their agency in educational data such as educational apps and platforms (Arantes & Buchanan, 2022). This is different from the concept of a data advocate in digital transformation leadership, who assists leadership in building a data driven culture (McCarthy et.al., 2021). This is somewhat similar to an o*pen data advocate*, who works on issues of data transparency, calling for open data sets, especially those held by government entities (Doerr, 2017).

Unlike policy domain-specific advocates in education, or open data advocates, who push for specific gains in radical data transparency, my concept of data advocates' goals are broader, working toward connecting data sources in governments and the private sector and applying data and data technologies to solve "wicked problems" (Head & Alford, 2015; Lonngren & van Poeck, 2021). In this case, wicked problems can be defined as problems that are so complex that no one sector, especially sectors centering technology and science, is equipped to solve them. For data advocates, data are an underutilized and powerful tool to add to the toolbox in solving these wicked problems, but they also believe that in order to find a solution, actors from different sectors and organizations need to collaborate. Since the actors come from multiple sectors with different backgrounds, a concept like the data advocate groups these diverse actors according to their similarities and allows us to see their work as a cohesive whole.

Data loyalists from diverse backgrounds disrupting the status quo

Data advocates from both sites hail from diverse backgrounds; they are interested in disrupting the status quo, and they exhibit data loyalty. Data loyalty refers to a commitment to data-driven approaches, such as evidence-based policing, implemented in non-profit/activist organizations, private industry, and government. They share a commitment to the power of data to make sense of difficult and complex problems and point the way to possible solutions that do the least harm possible to communities. Data advocates from both sites view data as a powerful tool that is able to challenge long held assumptions and biases and is thus able to transcend and challenge the prejudices or path dependencies that often inform and drive policy.

For many data advocates, this interest in centering data was often described as "always there". As one Measure data advocate explained, "I've always been interested in data and knew from even before I got my master's [degree] that we needed a better system in the police department–a system where we could track things, and store them and query [them]" (M10) For many data advocates, this commitment can turn into a passion and has meant devoting much of their spare time to data projects while holding down a full-time job elsewhere, for example, in a police department. Volunteering one's time to data projects was much more prominent in the Measure site than in the HUBs, where funding was directed to full-time paid staff. As one Measure data advocate said, "In 2015 we had no funding at all. Not a dime! We were all volunteers and just very grassroots" (M4). They continued, "We all have day jobs, so this is just what we do because we know it needs to be done" (M4).

In data for good projects, data loyalty is connected to the purpose or problem one is trying to address. Data loyalty is oriented toward real-world applications rather than fundamental research per se. Loyalty to data leads one to believe that without a well-defined purpose or problem that matters to communities, the research is empty. As was especially the case for Measure data advocates, publication did not necessarily motivate research. Because so many data advocates from the Measure group held outside full-time positions outside their data for good work, they differentiated themselves from academics by calling themselves "pracademics". "We're not a special interest group or just researchers," as one participant put it, "we don't care about getting published to get published. We need to see the rubber hit the road and see real changes" (M6). Although the term was not used in the HUBs group, the idea of the pracademic aligns with how data advocates in this group understood their use of data for social good,

> I think there's a difference between what one might do in the social good space versus in other types of areas of data science. So, in some areas of data science, say I'm looking at images of the sky, maps of the sky that I want to just pull out, [I would ask] what are some of the prominent features? Maybe then I would do a deep learning analysis and see what falls out, right? …But I think when you're doing data science in the social good space, it is driven by the questions, and the questions are defined by the people for whom these issues matter and who are going to be making the policy… (N3)

The "pracademic" is loyal to data but only when data is anchored to an important question, defined by those most affected by policy.

Data advocates also come from diverse backgrounds. This is especially true at Measure. Unlike other data-intensive fields such as research and data science, data advocates do not necessarily come from backgrounds that include formal technological or research education and training. Consistent with the third generation of data discourse perspective, a much broader range of embodied lived experience counts as expertise within the data for good culture. "Academics just don't see what we see. And so we are able to drive that research because we're in it, we live it" (M10). Because academics are too far removed from the experience, they are not as qualified to drive the research.

Although all HUB advocates interviewed are trained as data scientists or researchers, their understanding of who belongs in the group is also much broader. One HUB advocate explained,

> You have to have folks who are at varying stages of your local city government departments of transportation, for example, who are never going to become data scientists but need to have varying degrees of understanding and ability to work with teams and vice versa. The data scientists need to understand enough about the domain they're in to be able to work effectively with practitioners. (N6)

In the Measure group, all data advocates come from backgrounds in community advocacy, activist groups, and policing. Even the one interview participant (M11) representing the private tech industry held a graduate degree in criminal justice.

Data advocates from both sites also described themselves as disrupting the status quo in some way. This could include disrupting the internal culture of a bureaucracy, traditional data-sharing jurisdictions, or common practices and beliefs. For example, “It's replacing peoples’ intuition. Sometimes we think something is a certain way and the data just don’t support that at all” (N3). This data advocate explained their role in part, as replacing status quo thinking with data evidence. Measure advocates from a policing background often described internal policing culture as data-averse and bound to tradition. Successfully disrupting and changing the culture sometimes required sensitivity to the status quo in order “to help departments walk the line between the conservative (police officers and union) and the more progressive side” (M10). Measure data advocate M6 described traditional police culture as something like *The Matrix.* “Have you ever seen *The Matrix*? There are some people who are just married to the Matrix and will do whatever they can to protect it.” According to data advocates from the Measure site, policing policies tend to be led by popular trends unsupported by data and once adopted, change is difficult. As one Measure data advocate described their organization, “we’ve always been [kind of]a superficial organization. You know, what looks good or sounds good, whatever is the flavor of the day” (M3). Measure data advocates would like to disrupt this perceived common practice of adopting “flavor of the day” policies. If data were used to study the outcomes from these policies, according to Measure data advocates, then only policies that performed well would continue to be implemented and the “flavor of the day” policies would be thrown out.

Another way of disrupting the status quo is by opening up existing data sources and sharing data. One of the major challenges for data advocates was gaining access to already existing datasets, particularly those held by different levels of governments with different jurisdictions. Data advocates from both sites understood a core piece of their work to consist of “breaking down data silos” for the purpose of data sharing and in many cases, opening those datasets to the general public. Part of this effort was described in terms of open data policies, particularly in instances where trust had been broken between citizens of a certain community and government. “I encourage any police agency to be transparent and open with their data. People have a right to see anyway. It’s not really our data; it’s their data and we’re just managing it” (M1).

For many data advocates in both sites, the disruption leads to outcomes generally perceived as good, such as more rational policies that align with stated goals. However, one possible consequence of the commitment to disruption is being perceived as an outsider and feeling alienated within one’s organization or movement–feeling like a misfit. The problem was especially prominent in the Measure group.

Misfits

Although the concept of data advocate is surprisingly similar across my two dissimilar sites, one important difference between the two groups is that data advocates from Measure were more likely to describe their identity as misfits, in addition to disruptors. To be a data advocate can mean isolation within one’s larger organization or criticism and rejection from larger activist networks, especially those focused on increasing police accountability or decreasing funding to police forces. One Measure advocate described the reaction from Black Lives Matter activists in this way, "I was ostracized by my fellow activists when I went on that panel talking about data. They looked down on me and it took months for me to prove that this actually mattered, that black data matters" (M4). According to this data advocate, other activists reacted poorly to an agenda that elevated data as a means to racial equality. For these activists, focusing on the data was a distraction from what really mattered.

Measure advocates from policing backgrounds expressed similar concerns and experiences. Interviewees articulated that one could lose one's job or pass up advancement by using data to promote evidence-based policies. Several Measure data advocates from policing backgrounds talked about how being part of the group meant, "putting my badge on the line" (M10), or being perceived by their fellow officers as, "part of some cult" (M10). This role as a misfit and potential disruptor of the status quo means that some of these advocates are seen as potential threats by other members of the group. M6 described an experience where, "I was kind of a threat to [someone in the police department]. M10 described their experiences in a similar way where leadership, "pooh-poohed a lot of the things I did. So I had to do things kind of quietly, and I always tell people, do things quietly. Partner with one or two other people and just do it." Feelings of isolation within one's own field were exacerbated by a lack of support from leadership.

Collaboration and the Start-Up Model

In addition to featuring the data advocate, third generation negotiations of the meanings of expertise, data, and the social good also take place within a different kind of environment that is highly collaborative. As one HUBs data advocate put it, "It is a foundational value of the Hubs that collaboration is inherently better. More people collaborating is a public good" (N4). As was the case with data advocates, the collaborative environments across the two sites share significant similarities. Both view their collaborative efforts as experimental and use language similar to the start-up model typical of locations like Silicon Valley. Both groups share a commitment to collaboration as a means to more accurate information and eliminating both statistical bias and socially constructed biases. This section details the collaborative environment as described by data advocates' interview responses.

The start-up model of collaboration

Both sites shared an understanding of collaboration reminiscent of Silicon Valley start-ups where experimentation is prized. In the HUBs, data advocate N1 described this environment as follows, "NSF had tried to essentially encourage lots of different methodologies and approaches. [They looked]at a bunch of different perspectives and they weren't being very top down [or] 'thou shalt …'For better or worse there was this diversity of different approaches being set out from the get-go." One of the benefits of the start-up model described by this data advocate, echoing private sector discourses, is that it created space for creative solutions and ideas. "One of the big pluses was that we were able to see other very creative and resourceful executive directors and co-executive directors being brought on and sort of developing their own style and background, [which] forces a different and complementary culture of each Hub, a portfolio of different activities." Or as N2 described, "There are pros and cons to the NSF approach, but one of the big pros is that you get to see somewhat different trajectories of different ideas that come from different institutions and from people with different backgrounds."

Interviewees perceived the deliberate adoption of an experimental start-up model as necessary because the structure, at least of the Hubs, was new to the National Science Foundation and to most of the PIs applying for its grants. No one knew precisely what would work. N5 described the Hubs as an experiment in expanding the network in data science for applied research, fostering partnerships in industry, government, and academia. Unfortunately, according to this data advocate, most of the PIs were used to conducting "lone-wolf" research in fundamental science funded by the NSF. Or as another HUBs advocate described, "…we really need a long-standing organization of these partnerships for them to talk to each other, hold workshops, get people to share data knowledge and resources with each other. How do you actually create a partnership where the end goal is to get people working together who don't normally work together?" (N7).

One might reasonably expect that styling collaboration after start-ups would be less welcome at Measure than at HUBs, since that language aligns better with bureaucratic or corporate settings than with street-level activism. One HUBs data advocate suggested that the start-up model might be seen as problematic for non-profit, activist groups: "That's what happens in the start-up model as well. You fail as quickly as possible. That's exactly what I was talking about. You use your preferred parlance in the

Silicon Valley arena, so I use the ‘fail fast’ mantra, but then you get social groups who hear that and say, ‘Wait, that's not right!’” (N4).

However, perhaps surprisingly, Measure data advocates did not seem put-off by start-up rhetoric. Five of the Measure data advocates I interviewed used start-up language in much the same way as HUBs advocates. For example, in discussing one of their research partners, M10 used the “fail fast” mantra as a description of the research culture stating, “The model, which is rapid testing, rapid research…the model is that this can be easy peasy, one-page snapshots. Tell us what your trial was, tell us the outcomes, and let’s move on. If this doesn’t work, then it’s a fail forward, fail truthfully, fail fast, and just move on [situation].” Data advocates from both sites understand their model of collaboration in terms of start-ups and experiments.

The only data advocate who voiced concern that the startup model would not create a space that nurtured healthy creativity and experimentation–an advocate from HUBS!--called it “magical thinking”. This advocate said that the Silicon Valley startup model led to “coopertition” rather than cooperation, investing cooperation with underlying competition and tension, especially without more intentional and deliberate guidance from leadership (N5). Coopertition models are largely reminiscent of second generation data discourses. This interviewee’s comment suggests that the persistence of traits from earlier generations of data discourse into the present, can sometimes complicate third generation efforts in data for good. I will return to this issue in the final section of this chapter.

Collaboration and accuracy

A second prominent feature of the collaborative environment preferred at these sites is the commitment that more diverse perspectives at the table actually provide a more detailed and accurate body of knowledge than do collaborations that are more homogeneous. This recognition that scientific accuracy is connected to the diversity of embodied, lived experiences encourages collaborations among individuals with different kinds of expertise. Advocates from both sites discussed accuracy in more traditional terms as leading to quality research but also in terms of eliminating bias and including multiple points of view for providing context and a broader picture. For example, when asked about how members from different sectors benefited from plugging in to the HUBs network, one data advocate responded,

> I think it goes back to the way the scientific method is supposed to work…that you try to test the inverse, then you try to test the positive of your hypothesis, and you open it up for others to reproduce your results and to challenge them. So with collaboration [...] comes different perspectives, and if you're doing it right, a diversity of perspectives. And it also just protects against some myopic views that some can have if they get too far into what you're doing. (N5)

Another HUBs advocate responded in a similar way, "I would say it's about promoting collaboration in support of research. Beef up the collaboration aspect, [and] the more you'll improve research… broader, more diverse research" (N6). Measure advocates also understood that collaboration with diverse stakeholders created more accurate and higher quality research. "What we've found is that analysts are really good at being analysts, but they have no idea how law enforcement works, and so they don't understand the data or what it means. That's where it helps to have different people who see different things, who have different lenses" (M2).

Interviewees also articulated that engaging many more voices could be thought of as eliminating bias, especially social bias, thus making datasets more accurate. Unlike with statistical bias, which refers to there being a problem with data collection and sample sizes, the concern over social bias in both sites referred more to prejudices and assumptions made about groups of people. For example, they might be concerned with how communities of color are affected by or engage in data collection and use practices, or how falling back on preferred solutions could lead to tech solutionism. Social bias is related to statistical bias (e.g., a sample that only includes a college-age demographic is not unproblematically generalizable to the larger population), yet the two groups seemed to be speaking to larger issues of how cultural prejudices and assumptions inform the entirety of the data process.

This is a particularly clear example of the dominance of the third-generation data discourse's preoccupation with recognizing the inseparability of data from context. As data advocates negotiate meanings and uses of data within these collaborative environments, issues like social hierarchies and how those hierarchies can harm or improve social good are prevalent. As a Measure advocate described, "we elevate data to address what we call disparities, and the way we do that is we bring together all these multiple stakeholders around either health justice, education justice, criminal justice or economic justice" (M4). Or as a HUBs advocate explained,

> I think it's imperative [for] those leaders to surround themselves with diverse teams so they don't forget about the implicit bias they're baking into things just based on their own experience and what they think other people want. So, in my example of the food desert, maybe people don't need whole foods at their convenience store. (N6)

Dream teams and problem children

If including more perspectives is essential for higher quality data and research, then deciding who should be included in the research design becomes both an opportunity and a challenge. Both sites talked about who would be included in their “Dream Team” and also talked about potential “Problem Children”. For example, one of the most important members of the dream team for both sites is that of the relationship manager, or “Data Yenta” (N6). However, the HUBs site focused much more on how that role has been institutionalized and the gender implications for the role. By contrast, HUBs advocates identified a certain way of thinking, “Engineering Brain,” as a major obstacle to collaboration. “Engineering Brain” seems to stand for a data expert primarily informed by an earlier data discourse, who may still understand data as disembodied and may not explicitly value different points of view. In addition, both Measure and HUBs advocates identified tech vendors and volunteers as well as political leaders as potentially problematic, due to how network members might become too dependent on them and potentially beholden to their interests. Both “Dream Team” and “Problem Children”-type roles are explored in this section.

The Data “Dream Team”

Dream Team roles were dependent on how data advocates defined expertise. Advocates from both sites identified data experts as necessary members of any data driven team, but with slightly different understandings of how that expertise should be defined. Not surprisingly, advocates from the HUBs site identified data experts as highly trained data scientists or quantitative researchers. All HUBs participants came from data-intensive backgrounds with years of technical training, such as high-speed computing, data analytics, computer science, neuroscience, and data science. Advocates from the Measure site, however, tended to include a broader swath of actors within the data expert sphere. As described in the Data Advocates section, many of the advocates in the Measure site came from policing and activist backgrounds with no formal training in data science or high-speed computing. Many, however, did have training in statistics and research design. Measure utilized outside groups when more advanced quantitative analysis or app development was needed.

Another important role in the data network "Dream Team' is that of the Translator. Unfortunately, this role seems to be rarely staffed and largely informal, as is evidenced by this data advocate saying, "If only this position were funded, but it never is" (N1). This role is necessary to a team because the languages spoken by data experts and subject domain experts can be radically different. For example, one interviewee explained,

> What we've found is that analysts are really good at being analysts, but they have no idea how law enforcement works and so they don't understand the data or what it means. Our data acquisition team not only prepares the data for them and cleans it, but they actually make sense of it for those analysts[...]. But they also have a meeting with a data acquisition person prior to even starting. That person can explain all the nuances and all the unusual things, because in law enforcement in this country there's no standardization of data (M2).

In the Measure site, this Translator role is talked about as another kind of relationship manager, translating between police departments, data experts, and activists, sometimes even managing political dynamics within a police department. According to one Measure interviewee, "Relationship managers help departments walk the line between the conservative (police officers and union) and the more progressive side" (M2).

A third role that was seen as necessary on a team was that of Domain Expert who possessed knowledge of practices within certain communities or in certain policy areas such as transportation, policing, or education. The Domain Expert can mean two things: often, it refers to an academic expert who studies and publishes on policy domains. But it can also refer to the lived experience of a professional, such as a police officer, or an individual's experiences as part of a community, such as an individual who is a member of a minoritized community, where their experience is similar to the end-user experience. In the HUBs site, the Domain Expert often translates into "interdisciplinary research", where data scientists and data-driven researchers work in teams with social scientists or researchers from the humanities. Although some mention was made about including the people most affected by data for good projects as a subject expert, HUBs advocates did not include lived experiences as a kind of expertise as much as data advocates in the Measure site. Advocates from the Measure site more often described the lived experiences of front-line police officers and civilian community members as crucial pieces of evidence in building data for good projects and therefore conceptualized lived experience as necessary expertise.

However, not everyone agreed that lived experiences as expertise were needed or even helpful. When responding to a question about including activists in conversations around data projects, one Measure advocate expressed concern over their perceived lack of data expertise, responding, "My initial gut reaction is like, no! But…what I would actually say like yes, if they actually get it right" (M3). This speaks to the ambiguity of who is considered expert enough in data to qualify as a member of a data for good Dream Team. Including advocates without any formal training in data could potentially create a situation where data is interpreted according to a political agenda or for self-interest. Even though Measure advocates categorized lived experience as expertise more than HUBs advocates, they echoed HUBs advocates by identifying Domain Experts as academics. Partnerships with University Criminal Justice departments were therefore critical in data for good projects. Measure interviewees thus had an implicit typology of expertise that informed the kinds of teams they felt would be most effective.

Table 4.1 Dream team: Roles within collaborative environments

Roles	Description	Purpose
Translator	Someone who can "speak" the languages of domain expert and data expert	Facilitates effective communication between data experts and domain experts in data for good projects
Data Expert	Formally trained data scientists or quantitative researchers.	Performs the "high lift" technical aspects of data for good projects such as research design, app development, or analyzing large data sets
Domain Expert	These individuals can be professionally trained in policy areas such as policing, or these are individuals directly affected by policies – the lived experience or user expert.	Performs the "high lift" of context in data for good projects. Orients the data for good project toward benefiting those most affected
Data Yenta	The matchmaker and relationship maintainer between data sources and data advocates	This person or people facilitates initial connections between data sources and data advocates and between data advocates. This person is also often crucial in maintaining those relationships.

The three roles described above are accompanied by a fourth role, which deserves special attention–that of the "Data Yenta". It is especially interesting because, as a role, it is frequently gendered in a way the other roles might not be. As described above, data advocates are not all computer scientists. Not all data advocates have a background in data analysis, research, or statistics. For example, many data advocates in this research became engaged in data science and research because their advocacy strategies by other means proved ineffective. But it is also the case that highly skilled data scientists in their role as a data advocate had very little expertise in a substantive field such as policing or racial justice. This mismatch of skills has necessitated partnerships and collaborations between data advocates with varying degrees of data literacy and subject expertise. Consequently, the necessity to build and maintain relationships became crucial to the success of these two groups. The HUBs itself is an organization that was formally institutionalized and funded in order to manage such relationships. On the other hand, for the Measure group, this task of matchmaker remained informal.

The Data Yenta role carries gendered dynamics that map onto previous data discourse generations, but that persist in third generation discourse. Because two sites reflect first, second, and third generations to different extents, these gendered dynamics manifest differently at each site. Here I detail the matchmaking, or Data Yenta, role in each site, ending with a comparison of the gendered dynamics of this role in each site.

The term "Data Yenta" came from an interview with a HUBs data advocate trying to explain the mission of the NSF HUBs network: "We all kind of play a data yenta" (N6). Other HUBs advocates explained, "We spark relationships." The HUBs proposal calls explicitly for relationship building and collaboration. One Hubs interviewee said, "The whole point of collaboration was to link different sectors and this was really the goal of the HUBs and Spokes idea" (N4). For the HUBs site, Date Yentas not only connected people but also found the "hidden data everywhere" (N7), connecting datasets and databases with the right dream team. Even though HUBs data advocates described the whole HUBs network as a kind of "uber" Data Yenta, I use the term to describe a type of actor within the network in order to highlight their work building and maintaining relationships and the gender dynamics at play, particularly in the HUBs site.

Unlike the HUBs site, data advocates in the Measure site did not explicitly discuss an institutionalized Data Yenta role. Data Yentas in Measure are much more informal. Much like the HUBs site, Measure data advocates described key individuals in the network as those who connected or “collected people.” M10 described one Data Yenta this way: “You know she goes out and collects people, so that’s how I was brought into the fold.” The role of the networker in the Measure site was extremely important in connecting data advocates across the United States, particularly in police departments where data advocates often described themselves as isolated misfits.

The Data Yenta role is thus one of the most important roles in both sites, connecting data advocates to each other and connecting data resources to those who want to incorporate more data driven practices. I discuss this role at length here given that Data Feminism is interested in gender dynamics, and my analysis of the Data Yenta role brings some of these gender dynamics to the fore. What Rob Kitchin (2014) says about databases can also be said about data discourses: they “are expressions of power/knowledge and they enact and reproduce such relations” (p. 22). It is possible that the hierarchy at the level of the National Science Foundation effectively reproduces patriarchal roles and relations at the HUBs. This is especially the case with how emotional labor can be rendered invisible or not fully recognized by incentive structures. Emotional labor, often done by women in organizations, is assigned and taken up by women (Jackson, 2019). Although some scholars frame this division of labor as natural, feminist scholars have argued that this assumption that women are more naturally predisposed to cultivating and maintaining relationships stems from essentialism. This concept is used to describe the assumption that women are biologically aligned for caretaking and roles that require a great degree of empathy. That a majority of Data Yentas are women suggests that this view of essentialism could be at work in the HUBs, at least implicitly.

Emotional labor is not often recognized in NSF grant proposals. It is both hard to define and hard to quantify. However, the Hubs network was intentionally created to incentivize collaborations between experts who do not often work together. This requires dedicated paid staff who are in a position to connect different data users and data sources. In other words, creating and maintaining relationships was the purpose of the Hubs infrastructure. Most of the Hubs directors who often played the role of Data Yentas, at least in phase 1, were women, while a majority of the Principle Investigators were men. As one HUBs advocate explained, "This put many women in the awkward position where we were their peers and yet we did what they asked us to do…they treated some of the women like secretaries" (N5). Although the labor was not "invisible" in the same way as emotional labor often is, the academic incentives still in place, such as journal publications, clashed with the relationship building purpose of the Hubs. Building relationships is not easily translated to a Curriculum Vitae. This mismatch of incentives created, in some instances, an unhealthy dynamic between Hubs advocates and the principal investigators of the grants who were able to publish material.

"Problem children"

In addition to articulating the different types of roles that might make up a "dream team," study participants also identified types of roles that present the greatest challenges to effective collaboration, creating obstacles and dependencies. I've come to call these kinds of roles "problem children," and have provided a typology of them in Table 4.2. The three "'problem children" most frequently identified by both sites are: those exhibiting engineering brain, tech volunteers and vendors, and political leadership.

Table 4.2 Problem children

Problem Child	Description	Problem
Engineering Brain	A reductionist tendency to break problems apart and solve each problem or part on its own, often at the expense of the bigger picture or deeper context.	Obstacles in collaborating with others, especially outside of engineering fields
Tech Vendors and Volunteers	Data for good projects, especially those in "scrappy" organizations and networks depend on data expertise in the form of vendors and volunteers	Potential of hijacking the data for good agenda Potential for tech solutionism
Political leaders	Data for good agendas, resources, and support are dependent on political leaders.	Changing political administrations shifting agendas and resources away from data for good projects

HUBs data advocate N4 shared an insight that described the dynamics of cross-sector collaboration as one of disrupting "Engineering Brain". This concept was discussed by other HUBs advocates but did not come up in interviews with Measure advocates. Engineering brain refers to the tendency of engineers to work alone on discrete projects where those projects are broken down into pieces and each piece solved separately from the other pieces of the project. HUBs data advocates tended to describe the actors within the HUBs networks as either engineers or those working in a social field, such as education, policy, activism, non-profit, and government. The "social field" actors were described as naturally predisposed to collaboration, while engineers tended to work alone on discrete projects.

According to HUBs advocates, this tendency to work alone or only with other engineers is not a sustainable model for incorporating more data-driven practices. Nor does the tendency to break down problems into smaller projects that are each solved independently of each other offer a sustainable way to tackle "wicked" social problems such as chronic homelessness or inequitable transportation systems. HUBs advocates described the HUBs collaboration model as one that seeks to disrupt "engineering brain". To the extent that engineers do collaborate, the model tends to be an expand-and-contract model, where actors come together to initially collaborate, and then each retires to their own corners to solve little pieces of the project. The HUBs collaborative model intervenes at the key moment after contraction and requires engineers to come back to the table and collaborate multiple times throughout the project. It also requires collaborations with data advocates who are not engineers. Instead of expanding and contracting only once, the HUBs collaborative model goes through multiple iterations of expanding and contracting as a means of disrupting engineering brain.

Collaborating in data for good projects requires individuals trained in using data. Often this meant partnering with volunteers from the tech industry or with tech vendors who sponsored conferences. One problem, as described by N4, was that it could create a dynamic where both problem framing and solutions would be steered by tech volunteers and vendors. "Depending on how many people in each of these groups show up, then you get your agenda less from analysis and more on physical presence." (N4). Furthermore, N4 explained, tech volunteers come from a background where technical solutions are applied. Therefore, tech volunteers tend to apply technical solutions inappropriately where a different policy solution might be more effective and appropriate. For example, a technical solution to reducing pedestrian fatalities from vehicles is to install traffic cameras in neighborhoods. This reduces potential for overuse of force by police officers, but it also increases surveillance of the neighborhood and overburdens residents with traffic tickets. In contrast, a non-technical policy solution to install more sidewalks would have neither of these negative effects, all while creating a safer environment for pedestrians.

Data advocates from both sites pointed to the challenges of private tech vendors and the problems of building dependence on data science volunteers. As one Measure advocate explained,

> When you buy an RMS system for your agency, it's literally millions of dollars,so you're kind of at the mercy of who you purchase from. We're finding that agencies are trying to go back to RMS vendors and say, 'We need this kind of information in our report templates.' And the vendor will come back and say, no problem, that will be another $350,000 (M2).

In part, data advocates described these dependencies as impacting the network's autonomy and control over the research agenda.

The last problem child identified by data advocates from both sites was political leadership. For the HUBs, this meant a dependency on shifting agendas at the national level, whereas advocates in the Measure site focused on shifting agendas at the local level as a result of a change in political leadership. One HUBs advocate in particular stressed the importance of public/private partnerships as the means by which data for good projects could be sustained over time and be insulated from changes in national administrations: "It can't be the government pushing this all the time because the federal government changes, the administration changes…and big data is too important to leave it to different presidential administrations" (N7).

The link to private industry was seen as a way to promote a more sustainable organization that did not rely solely on government funding and would not remain at the whim of changing administrations. Data advocates in the Measure group talked about the importance of political leadership in advancing their own agendas. Several mentioned that without buy-in from the top, data driven efforts would ultimately fail:

> I always start at the top–always! That meant going directly to the mayor or to congresspeople and not their staffers. Social media has been my friend in terms of making those initial relationships, but definitely starting at the top. The [police] chief is accountable to the mayor, and the mayor is accountable to the [city] council, so getting that buy in makes that much easier (M4).

When asked whether data for social good efforts can be incorporated into police departments without political buy-in, one Measure advocate responded, "It doesn't work. End of discussion. It doesn't work" (M6).

Intrusion of first and second discourses on embodiment

The language of the collaborative environment of data for good projects through the lens of start-ups provided the license and freedom to experiment with different kinds of events and forms of collaboration, in person and virtual, such as using a Slack channel, Twitter, or policy roundtables at in-person events. However, at least at the HUBs site where competition for resources is an integral part of their activities, framing collaborations in terms borrowed from the tech industry also presents challenges.

The data advocate from HUBs who characterized the startup model as "magical thinking" and cooperation as "coopertition" was telegraphing these concerns. Because this site is embedded in a larger context (NSF) that is perhaps dominated by the first and second generation data discourse, as well as neoliberal market logics that are hegemonic in the wider world, it is susceptible to intrusion from that wider context. The example above of the perpetuation of patriarchy in the assignment of the Data Yenta role is one example.

Indeed, embodiment challenges the startup model of collaboration. As detailed above, data advocates from both sites framed their collaborative environments in terms of the Silicon Valley model of start-ups. Wendy Brown (2017) identifies the start-up model as part of a broader trend in neoliberal politics and economies. According to Brown, neoliberalism is "a peculiar form of reason that configures all aspects of existence in economic terms (Brown, 2017, p.17). It is "a contemporary phenomenon in which rule transmutes into governance and management in the order that neoliberal rationality is bringing about" (Brown, 2017, p.20). Neoliberal rationality, as a cultural shift, devours everything, including our notions of justice, equity, and democracy, transforming these values into economic returns on investment. The risk in data for good projects is that the good, understood through neoliberal rationality, is reduced to economic concepts. As Brown explains, "The conduct of government and the conduct of firms are now fundamentally identical; both are in the business of justice and sustainability, but never as ends in themselves. Rather, 'social responsibility', which must itself be entrepreneurialized, is part of what attracts consumers and investors" (Brown, 2017, p.27). The whole rationale of the start-up is to attract investors through intense competition, or in the case of the HUBs site, "coopertition".

Embodiment challenges the start-up model because it challenges the neoliberal rationality that reduces culture to economic terms. Embodiment is "…the idea that there is a constitutive relationship of the lived body to thought, to knowledge, and to ethics, taking leave of the modern idea that bodies can be left behind as the mind does its work" (Martin-Alcoff, 2013). Within this concept of embodiment is the elevation of the body and its locatedness and embeddedness in a cultural web not easily reduced to ROI. Justice, social good, and equity, from embodied reason, is not a marketing tool to attract investors but political and social values shared within communities, of importance regardless of their attractiveness to investors.

Embodiment challenges and exists alongside first- and second-generation data ideologies. In the next chapter, I explore how data advocates understand data itself. One implication of third generation embodied data is the elevation of personal, lived experience, filtered through or into more traditional data frameworks and the authority those frameworks provide which allows for interesting negotiations between first, second, and third generation data discourses.

CHAPTER FIVE: DISEMBODIED AND EMBODIED DISCOURSES

How data can be used to promote social good depends, in part, on how data are understood. This chapter analyzes both disembodied and embodied discourses as they show up in both Measure and HUBs data advocate interviews. The findings of this research suggest that data advocates from both groups rely on disembodied discourses that understand data as having a voice that speaks with authority in a language we can all understand and as having a god-like sight unclouded by personal agendas and emotions. But advocates also use embodied discourses, understanding data's "voice" as lending authority to diverse, lived experiences and sight that sees more like a drone–mobile, locatable, able to be manipulated, but also potentially weaponized–and less like a god. Furthermore, data advocates from the Measure site use third generation embodied discourses more than data advocates from the HUBs site and yet, data advocates from the Measure site are adept at moving between the disembodied and embodied discourses, blurring the dichotomy, suggesting that both discourses have a positive role to play in data for social good projects. Furthermore, these negotiations around the meaning of data provide even more evidence of the inseparability of data and politics, opening up possibilities for human agency and political action within existing and future data infrastructures.

Tensions between embodied and disembodied data discourses point to both the kinds of work data can do in the social good space and alert us to potential pitfalls. When organizations use the "data for social good" label to push, organize, and frame (or sell) their agenda, it is often assumed that data simply speak to the particular issue defined as a good, or as a problem impeding some social good. But we know that context matters.

Chronic homelessness, for example, is not a shared experience; those who are living in that circumstance experience the problems and the policy interventions in a different way than those who have never experienced homelessness. Nor do people of color experience policing or crime in the same way as white or caucasian people. Discourses that discount or ignore data's embodiment, and insist on data's disembodied objectivity and neutrality, run the real risk of ignoring how bodies (Black bodies, female bodies, transgender bodies, poor bodies) interact in socio-political systems and can, despite good intentions, reproduce social harms. "Switching" between disembodied and embodied discourses may help in reproducing data's authority while at the same time avoiding the naïve or imagined objectivity from disembodied discourses that can reproduce social harms.

For the purposes of explaining what follows, it is helpful to relate disembodied and embodied discourses to policy scholar Deborah Stone's descriptions of the market and the polis as alternative models of society (Stone, 2012). The market model seems to presume something like Adam Smith's "the invisible hand," which presumes that individuals pursuing individual self-interest will eventually accumulate into a larger public good, in a mechanistic way. Under this philosophy, society is created when a number of individual rational consumers come together to make exchanges. According to Stone, the project to understand the dynamics and phenomena of these exchanges is the 'rationality project'. The rationality project "worships objectivity and seeks modes of analysis that will lead to the objectively best results for society" (p.10). This harmonizes perfectly with the concept of disembodied data discourse, in which the categories under analysis tend to be viewed as fixed and stable, and data is naively employed to measure and test for efficiencies.

Stone's polis model, on the other hand, understands society to be a community of embodied persons whose interests converge and diverge in complex ways in different and shifting contexts, so that achieving a social good entails a process of negotiating toward compromises and solutions. According to Stone, Embodied data discourse is consistent with the polis model of society.

While the sites I studied mostly strove toward a polis model view of their own communities, they manifested a constant recognition of the usefulness of disembodied data discourses in operating within the larger society. They also acknowledged the value of the "objective" perspective in facilitating intra-group communication. In order both to set up the contrast between embodied and disembodied discourses, and to characterize the situatedness of communities preferring embodied discourse within a larger society that still mostly understands data in a disembodied way, I will first discuss how the data advocates in this study employ the discourse of disembodiment.

Disembodied discourses

Donna Haraway suggests that the usefulness of disembodied data discourses comes from the authority and power that are achieved by pretending to maintain an objective distance from local knowledge and particularities – by playing the "god trick," as she puts it (Haraway, 1988). The god trick involves imagining the scientist (in this case perhaps the data scientist) as without a body, and who can observe, as a god would, everything from nowhere in particular, achieving perfectly objective knowledge.

For the purpose of this research, disembodied discourses include concepts that generally map onto the first- and second-generation discourses described in the introduction. These discourses are common and familiar: understanding data as "raw material produced by abstracting the world into categories, measures and other representational forms…that constitute the building blocks from which information and knowledge are created" (Kitchin, 2014, p. 1). These discourses characterize data as sitting above politics, values, and emotion. They have historically validated efforts at exercising state control and private sector activities in particular.

In interviews, data advocates often signaled awareness of the problems of understanding data in this disembodied way. Disembodied discourses of data can be attached to projects and policies that perpetuate social harms in the form of data colonialism (Thatcher et al., 2016; Fraser, 2019; Couldry & Mejias, 2019), perpetuating a dominant world view that maintains hierarchies that benefit those at the center (Foucault,1995; Scott, 1998; Brown, 2017).

But data advocates also understood the social power of disembodied discourses and sometimes used these in their own favor. Some interviewees understood disembodied discourses to lend authority and a common language necessary to group cooperation, particularly for the Measure group, where trust between police and community members had been damaged. These dynamics, which may seem somewhat counterintuitive given the importance of embodiment to many data advocates, are described below.

Disembodied voice: Authoritative and common

Data advocates from both sites used disembodied discourses, describing the utility of data in terms of its ability to speak a common language between dissimilar and contentious groups as well as an authoritative language, capable of communicating a story believed by both sides.

> I think often when community members express themselves, especially to government entities who may not understand the culture or how they're articulating their stories, [they] can often get dismissed. Data is really like a concrete language that everyone can understand. We know that 2 is bigger than 5, and we can see those disparities visually…[but] it kind of takes away those cultural differences. Those experiences and those concerns [can] be dismissed, and so that's how we see that data being a common language. (M5).

According to this data advocate, data serves as a bridge, or common language, between different groups with different cultures. This common language is particularly useful when those different cultures view each other with suspicion and a lack of understanding.

One kind of mark of this recognition is the use of phrases that imply that data speak for themselves. It is still common to use the phrase "the data say," as if data had its own voice apart from individuals who do the actual speaking. Furthermore, policy scholars often argue for policy makers to "let the data speak!" (Zachmann et al., 2015; Boyles & Meyer, 2016; Kettl, 2017). Another Measure advocate explained that, "Data can tell that story. It can show we all want the same thing. We all want to have a great life and to come home safely. We all want that, police officers, black men…so I think the data show that in a way that we can't tell, we can't verbalize. I think that data has been instrumental in changing these ordinances and these policies" (M5).

In describing why data was an effective tool in working with community, this Measure advocate responded, "It basically shows police departments how to meaningfully interact with community members and activists to use the power of data to speak that common language" and, "they (the police) know these systems oppress them [communities of color] but the data allows them [the communities]t o be able to push back in a way that can be measured…So I think data is actually power" (M4). To put this another way, disembodied data abstracts from bodies that society might deem untrustworthy or unbelievable. By abstracting from black bodies, the stories that data tell are given more authority. "The numbers really tell a story that sometimes it just really can't understand coming from a black face" (M4). In this way, disembodied data is, perhaps counterintuitively, a powerful tool used by those occupying less powerful positions in society.

Disembodied sight: a god's eye view

Advocates often seem to believe that the authority of voice and clarity of thought provided by disembodied discourses not only paint an accurate picture of social phenomena but can also rectify superstitions and correct long-held misconceptions. The "god trick" is also at work in disembodied "sight," where data "see" from a position sitting above politics, ostensibly unclouded by our own prejudices and emotions. I refer to disembodied data sight as having a "god's eye view" that is objective and complete because of its distance from particular bodies. As one HUBs advocate explained, "the human brain isn't capable of seeing patterns and we just have to do it this way" (N8). Not only is the human brain not sufficiently capable of seeing patterns, according to this perspective, it is unable to piece together a complete picture that accurately describes and explains complex social phenomena. "You have this problem that's really hard in the social area but now you have the data to show a better picture. You haven't fixed the problem yet anyway and now you have a different tool" (N6). Data provide a better view of complex social phenomena.

It should be noted that this faith in the corrective power of data may be a fallacy. D'Ignazio and Klein (2020) disagree that this form of objectivity is as robust as many think it to be, diagnosing it instead as naïve and potentially producing inaccurate information and knowledge. As will be discussed in the embodied discourses section, allowing for the diversity of data stories that may compete with each other for relevance and meaning potentially produces a more accurate picture of the complexity of the situation. Embodied discourses could have the effect of replacing naïve objectivity with strong objectivity. According to D'Ignazio and Klein (2020), strong objectivity is *situated* knowledge. It is a recognition that knowledge is produced by people in specific contexts of history and circumstance. Therefore, knowledge produced by only one group of people is necessarily partial and not simply objective. Strong objectivity, "works toward more inclusive knowledge production by centering the perspectives– or *standpoints*– of groups that are otherwise excluded from knowledge-making processes" (p. 83). The notion of strong objectivity predates feminism and CDS and is a core concept in the field of STS where the larger debate around objectivity, or what counts as valid knowledge and the reasons why, is central to debates around knowledge production (Hess, 1997), particularly in the sociology of science where questions around the claim of the unique status of scientific knowledge was central.

The concept of strong objectivity was taken up by feminist theorist Donna Haraway, who describes objectivity as, "feminist objectivity means quite simply situated knowledge" (Haraway, 1998, p.581). Strong objectivity comes out of feminist standpoint epistemology (a theory of knowledge) which argues that traditional "value neutral" epistemologies, such as positivism, are rife with political prejudices of dominant societal views and therefore knowledge from those differently situated in society produces more accurate scientific knowledge. In other words, real knowledge is socially situated.

Feminist theorists developed strong objectivity and standpoint epistemology out of an effort to explain and describe how to produce research that included the lives and perspectives of women, or research that was less partial than what had traditionally been produced from the value-neutral perspective of science (Harding, 1995). Strong objectivity is then an attempt to maximize objectivity and move the idea of the scientific self beyond a narrow view of what Sanra Harding calls, "Mr. Nowhere" (Harding, 2015). Avoiding the naive perspective of "Mr. Nowhere" is taken up in Data Feminism where traditional or disembodied views of data tend to take the perspective of naive objectivity.

Disembodied discourses align with naive objectivity and could perpetuate a valorization of the neutrality ideal. Nevertheless, the value of disembodied discourses was widely held among data advocates. In the words of one HUBs data advocate, "As a society, we exist in a series of illusion bubbles about things we assume [are] true that get popped when there's data about it. There are these illusions that dissipate as soon as you create this compelling data resource" (N7). This data advocate went on to describe an example of popping one such illusion bubble: that women don't travel alone, and so there is no need to market toward solitary women travelers. "There's another idea out there that [it] is unsafe for women to travel alone, the majority of people who travel alone…these are women and that's a business! And so you start collecting data, [and] the thing you realize is that the thing you thought all along qualitatively is just wrong!" Although this statement was focused on creating new markets and launching new businesses, it was also understood that these misconceptions were a source of harmful prejudices. In this example, popping illusion bubbles and exposing the work for what it really is with compelling data had the effect of normalizing women's mobility.

As with the discussion on voice, understanding data's sight as coming from a god's eye view carries with it a potential danger that the optimism and enthusiasm for a picture seen or the story told by data can veer into a form of data hubris (Read et al., 2016) where local knowledge and lived experience are dismissed in favor of a dominant narrative. This can perpetuate colonial data practices and create inaccurate descriptions and analyses, under the banner of objectivity.

Opening up to embodiment: the "Big Data" revolution and playing with categories

One may be skeptical that data can produce a social good for individuals living outside the power centers of our regime, public and private, given the colonial baggage data practices tend to drag along. Why would activists from civil society who are concerned with undoing data colonialism, like Measure Austin, turn to the mechanism of data? In my interviews, data advocates from both groups answered the "why data now" question consistently with some version of, "Because it's available! That's it. Easy answer. Technologists have provided a lot of tools that collect a ton of data and can mine it and give you an answer" (N6). Or, "For me, this lightbulb switched on that data can be different…that data can come from pdfs, that it can come from words,…and that I can make a dataset. That never even crossed my mind and now I make databases for everything" (M9).

According to the data advocates interviewed, the reason for more participation in creating data infrastructures is simply that more data are available. Data are one more tool in the tool box and it is abundant, cheap, and powerful. The sheer volume of data now available, particularly in the form of social data (data from social media), has made it easy to manipulate categories in order to create information and knowledge that is relevant to one's purposes (Bowker & Starr, 1999; Kitchin, 2014). Data are widely available and abundant because, "everything produces data and so there's an explosion of things you can analyze and monetize" (N8). Furthermore, "why wouldn't they use it if they have it! 20 years ago you didn't have lightweight laptops on your desk" (N2). Cash-strapped non-profits, activist groups, and government entities will use the tools available. Not only is data abundant and cheap but data also has the virtue of reusability. "It's free because there is a fundamental purpose for the data…if you make that data open, then there's all sorts of things you could use that data for. And that's the fundamental thing is the usability of data, and data is a resource that you can reuse over and over again" (N7). The reusability of data allows more groups to use data for different purposes.

In other words, perhaps inadvertently, in pursuing the goals of creating new markets and profit-sources, goals that may in themselves perpetuate colonial practices, big tech has also opened possibilities for data to be used for other purposes. Perhaps surprisingly, here may be a link between market mechanisms and the democratization of data. This democratization of data has brought more players (data advocates) to the table and allowed for a continuing renegotiation of its meaning and purpose, challenging traditional assumptions about objectivity and neutrality.

Embodied discourses

The concept of embodied data implies that data lives through, is produced by, and comes out of physical bodies instead of the persistent perception that data descends from on high. Bodies are located in a particular time at a particular place where senses like smell, touch, taste, and sound are the primary ways we experience and know the world including cultural expectations and prejudices about particular bodies. For the purpose of this research, embodiment is defined as, "…the idea that there is a constitutive relationship of the lived body to thought, to knowledge, and to ethics, taking leave of the modern idea that bodies can be left behind as the mind does its work" (Martin-Alcoff, 2013, p.1). Furthermore, embodied data will carry with it cultural expectations of different bodies, including the social hierarchies in which particular bodies are embedded.

D'ignazio and Klein (2020) identify embodiment as a key concept and principle in data feminism, or in using data for social good understood as co-liberation. However, the authors do not develop the concept in their book. However, some scholars in the field of critical data studies have applied the concept of embodiment to digital systems. For example, by creating three-dimensional materializations of personal data, people can, "feel their data" both physically and emotionally, or 'viscerally' (Lupton, 2017).

Embodiment is a matter of taking the contextuality of data to the level of individual lived experience. Data advocates from a variety of backgrounds, whether research, policymaking, or activism, recognize the importance of putting data in context. Advocates from both the Measure and HUBs site both agreed that data without context is not knowledge. A Measure data advocate from a policing background put this in terms of high crime dots on crime maps: “Data is just information. It doesn’t tell you the whole story. It’s gotta be put in context...it’s not just cops on dots” (M3). Another advocate emphasized the importance of context for meaningful analysis: “Just because data is available doesn’t mean that anyone is gonna see any insight” (N8). Without context, according to this data advocate, data will not provide a full picture or explanation. Simply being more available does not generate insight.

Many policy makers ensconced in earlier data discourses consider context to mean metadata, or data about the data. However, situated knowledge (Haraway, 1988), produced by embodied actors in particular social contexts, is important even to creating meaning from data as metadata. While metadata can be understood as an important piece of creating meaning in databases, contextualizing data meant something different to Measure data advocates. Context meant understanding the local conditions in which data were collected. One data advocate explained that putting data in context encourages analysts and policy makers to ask deeper “why” questions, potentially reducing the risk of reactionary policy. “How do you know what’s working? Let’s find creative ways to deter crime; let’s find ways to prevent crime; let’s find ways to improve our legitimacy in our community…You’re looking at the place the way an investigator looks at a person and really dive deep into why that is. You look at the why’s. Why is there theft?” (M2). Data for good projects, as well as policy and political action that use data, could benefit from a deeper conversation about what constitutes enough context for data to reach a threshold of the kind of “strong objectivity” I discuss next.

Embodied voice: multiple narratives

As with disembodied voice, the voice of embodied data is also understood to be authoritative and "can't lie" (M10), but here the source of authority is not imagined objectivity, but the authority of diverse, real, lived experiences. Where disembodied data is presumed to have one voice, embodied data have multiple voices, each speaking with authority. From this perspective, because people are embodied, so must the data about people be embodied. The third principle of data feminism states that, "Data Feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world" (D'Ignazio & Klein, 2020, p. 73). Instead of naïve objectivity, data feminism offers the concept of strong objectivity. The difference between naïve and strong objectivity is the location of the body. Attending to more of the world, through multiple voices, strong objectivity brings us closer to true observations than does artificially imposing one narrative (D'Ignazio & Klein, 2020).

For example, one advocate discussed how their city imposed a curfew on a high crime area in order to improve public safety by reducing crime. "It wasn't intentional, but it was a disparity, and it was largely minority children that were being targeted for citations, in concentrated parts of town. We couldn't really deny that" (M1). From a naively objective point of view, citations issued in enforcement of the curfew measured success. Yet, when the social location of those cited was taken into account, along with the voices of those directly impacted, the strong objective story was different.

Embodied sight: a drone's eye view

According to the "god's eye view" of first-generation data discourses, the authority of data comes precisely from its disconnection from the "ground," from bodies. Embodied data discourse envisions the initial location and trajectory of data from the ground moving up, instead of from the heavens peering down. This is a critical re-framing that impacts the possibilities for data producing a social good. But it is still the case that getting *some* distance is important. Distance powerfully assists our understanding, showing us trends, forecasts, predictions, and patterns otherwise unseen and unknown. Data sight that is still tied to an embodied location can be compared to a drone's eye view –this is my own metaphor and did not arise from interviews. Since a drone rises a little way above the earth, yet is still controlled by a human hand, this metaphor reduces the distance between bodies and data, maintaining a connection between the two. Framing data sight in this way avoids the utopian/dystopian dichotomy as well as avoiding the naïve objectivity and arrogance that comes from the "god trick". The metaphor is also useful in recognizing that drones are a technology that, while empowering, are also fraught with controversy, as the technology in the development of drones can be used as a military weapon or as child's toy (Latonero & Gold, 2015; Floreano & Wood, 2015).

For data advocates in the Measure site, data provided a drone's eye view by "seeing" patterns of racial disparity, historically left out of dominant narratives.

> Data does have to be a piece of this because [...] you can tell people day and night that there are disparities, [but] unless you can show them…. And when I say people, I'm talking about literally the white middle upper class people. Because they don't see a problem, because

> they've never been harassed by police, they've never had all of these issues that people in poverty and people in different minority groups have had to endure. You have to be able to show them that there is an issue to begin with" (M2).

When considered from an embodied perspective, knowledge produced by one group, "white middle class", will not see a problem because their perspective –and the data collected from that perspective– is partial.

Data discourses and strategies for cooperation: Boundary objects and data thermostats

The differences in use between disembodied and embodied discourses is also theoretically useful as we move deeper into a datafied society. It might be tempting to think that simply replacing disembodied discourses with embodied ones would bring us closer to notions of data justice. And yet, as shown above, Measure advocates used both discourses to achieve policy goals and efforts at building community and collaboration.

I will explore this phenomenon through the concepts of the boundary object (Star & Griesemer, 1989) and the data thermostat (my own term). The concept of the boundary object developed in response to a need to develop effective collaborative tools and techniques among diverse groups (Bowker et al., 2015). This type of problem is similar to that described in this research: data advocates from a diversity of backgrounds, industry, non-profit, academia, and government sectors, with different agendas, who are attempting to collaborate toward a shared goal such as community policing or mapping water usage in California.

Boundary objects allow groups to work together without consensus. They bring and hold groups together, in part due to the object's interpretive flexibility--its capacity for meaning different things to different people. For example, a map to a campground could be interpreted by campers as a tool for finding one's way to leisure, whereas that same map to a geologist could be interpreted as a guide to sites of potential data collection (Leigh Star, 2010).

As a common language, data are relatively stable or fixed. And yet, like a boundary object, they have interpretive flexibility. For example, a data visualization showing a high crime concentration in a certain part of a city or in a specific neighborhood could be seen as evidence that more policing efforts are needed to patrol that area. This is the response of many police departments who rely on what is called, "hot spot policing" or what Measure advocates refer to as "cops on dots". To someone else, this same visualization could be evidence of racism in the police department, particularly if the person looking at the visualization lives in a minoritized community. The definition and cause of the crime is interpreted differently depending on who is looking at the data. The visualization could act as a boundary object, standing between police officers and community members as they attempt to work together on solutions that will build trust between them. The point of the boundary object is not so much its "accuracy" or one-to-one relationship with reality, but its ability to facilitate negotiation over values and action.

This seemed to be the case when Measure advocates described the data used to shift the practices of police departments. Data advocates who are sensitive to this boundary object function of data are careful to present data without politically sensitive language, such as racism, so that each side of the negotiations is free to interpret the data from their own point of view. The data presented to police departments by Measure often uses the word bias rather than racism. This is intentional. Bias is a scientific and technical term, but can also point to prejudices without having the same triggering effects that calling people or practices “racist” might. Data showing a police department bias toward communities of color could mean that processes and structures of the police department need to change as an institution. For a cop, the police department shows bias in the structure; it does not show a police department that is racist. For a person of color, data showing bias can be interpreted as a larger social and structural problem of racism.

What the concept of a boundary object perhaps does not adequately address is the continuing role played by emotion in data negotiations, cooperation, and community building. When data are recognized to have come from bodies, we not only avoid the naïve assumption that data are neutral and apolitical, but we must also consider the role of affect and emotion in data negotiations. Striving to manage emotional disparities is one reason that advocates may wish to shift between disembodied and embodied discourses. I suggest the concept of a data thermostat as a tool for explaining the code switching (Krasas, 2018; McCluney et. al., 2021) or ‘“discourse switching”’ in response to emotion and affect. The “data thermostat” comes into play when the reality and authority of embodied emotion and affect are explicitly recognized by the data advocate; the advocate is then able to make decisions about how to use and to talk about data in order to raise or lower the emotional temperature of the room. The use of the data thermostat is itself a hallmark of social skills refined by interacting in a variety of contexts and with diverse actors.

For example, one challenge identified by Measure advocates was walking the fraught path between a police officer’s experience of danger and a community member’s experience of danger, especially when that community member is a person of color. For the police officer, danger is perceived to come from the community member, who may be posing a variety of threats. But for the community member, the uniform of a police officer carrying a deadly weapon may embody state-sanctioned violence against communities of color. In this example, issues of existential fear and perceived racism are rife with lived experience and deep emotion. If these groups are to collaborate in identification of the problem, its measurement and potential solutions may depend upon not inflaming emotions in the room. The concept of the data thermostat–where emotions can be turned up or down by invoking more and less embodied data discourses–reflects awareness of how different data discourses can affect the emotional temperature of the discussion. By using the disembodied discourse of distance, for example, data advocates measure the implicit bias of organizations, rather than racism of cops or organizations. As one interviewee put it, "One of the major challenges is that we have to be careful what we say – we don't say that a police department is racist – we say that we see evidence of implicit bias. This means that we're distancing the blame. Our primary customers are cops, so we don't want to point the finger and start calling them names" (M7). This form of disembodied discourse facilitates cooperation by turning down the emotional temperature of the room.

Data can also be employed to turn up the emotional temperature of the room. When working with members of a minoritized community rather than police departments, activist data advocates in the Measure site did not use emotional distance but rather emotional connection to real and lived experiences of racism to facilitate collaboration. Data was brought closer to the body, invoked in its most embodied sense. According to one Measure advocate, "People are so afraid of repeating the same thing as our predecessors went through, like Martin Luther King, or the different folks who came before us who did not necessarily have the opportunity to create public information requests, to really see, to really prove, how they felt emotionally" (M4). This data advocate further explained that, by bringing data closer to the body, empowered communities of color could be owners of their own narratives.

> That's the reason for Measure here in Austin, because we see the importance [of] black people being the owners of their own data. Because then that translates into them being the narrators of their own story [...]. One thing we do is go into neighborhoods and do community design surveys for community development. So we create our surveys, we determine our own standards for data collection and we do the analysis of it and then we are the ones who create the recommendations based on what we see and learn. That's why it's so super important to involve those impacted from the beginning. I also have a fear of perpetuation of racism through data and in some cases, some data systems need to be completely cleaned out. We start from square one and we need to make sure we're providing some context for equity in the collection, in how we're actually going about doing it. (M4)

The data thermostat is visible working in both directions, in this statement. The data advocate switches discourse modes according to the temperature perceived to be most useful in the context.

Scholars studying politics and collaboration, or deliberative democracy and policy-making, have studied the role of affect (Marcus et al., 2008; Coleman & Wu, 2010; Coleman & Banning, 2016; Sumartojo et al., 2016; Amrute, 2019). However, the relationship between data and affect has not been studied adequately. Data as thermostat could provide a useful concept for further research exploring the employment of affect and data in collaborative environments.

As a concluding remark I would like to note that data advocates from the Measure site used embodied data and invoked embodied data discourses more than advocates from the HUBs site. It is possible that because Measure advocates come from a variety of backgrounds the discourses used to describe the nature and purposes of data tend to stretch traditional boundaries. This very stretching is an incorporation of embodiment into data discourses. By using embodied discourses, Measure advocates bring a deeper meaning to the purpose of data, expanding what context means and how data are extensions of bodies, including the lived experiences and cultural expectations attached to bodies. Even the fact that Measure advocates used both embodied and disembodied discourses to describe how to effectively collaborate with each other reflects their greater attention to embodied context. Code switching in accordance with the data thermostat is yet another example of this attention to context.

CHAPTER SIX: THEORIZING THE SOCIAL GOOD

My research asks how data can be used to promote social good. A full analysis of this question requires paying attention to how practitioners in the data for social good space understand what data are and what constitutes a social good. In the previous chapter, I addressed how advocates from my two sites understood the meanings of data. In this chapter, I turn my attention to analyzing data advocates' articulation of social good. Findings suggest that HUBs data advocates articulate social good in open and vague terms, according to disembodied discourses of experimentation and social impact. Measure data advocates, on the other hand, used embodied discourses in their articulation of social good more than did HUBs data advocates.

In this chapter, after a brief discussion of social good terminology, I will take up a more detailed description of the different understandings of the social good as they manifest at my two sites. I will end this section with a discussion of code-switching. As was the case with data discourses, Measure data advocates are generally more sophisticated in their ability to participate in "code switching,", a term most often used to refer to the ability of people of color to switch between language/mores coded as white and the language/mores of their own racial groups (Krasas, 2018; McCluney et.al., 2021). I apply this term to how my participants are able to switch between disembodied generations discourses (first and second) and third generation embodied discourse in a strategic maneuver to accomplish their goals of advancing cooperation between traditionally contentious groups. Throughout the chapter I will consider how the different understandings of data for good correlate to effects on existing hierarchical power arrangements—that is, which are more likely to reproduce or exacerbate existing inequalities, and which are more likely to move toward greater equality.

In what follows, I compare two differing visions of social good that appear among the data advocates using the metaphors: ‘a rising tide lifts all boats’ and having ‘a seat at the table’. These metaphors map quite closely onto D’Ignazio and Klein’s framework of ‘social good’ versus ‘co-liberation’ in *Data Feminism* (2020), but I choose to use my own metaphorical terms here because they better capture how data advocates in the HUBs and Measure actually speak about their work in the social good space. Roughly, HUBs advocates understand the social good just as it is described in *Data Feminism*, while the goal of Measure advocates is ‘co-liberation’, yet Measure advocates also speak of this goal as the *social good*. Data advocates from both groups, in other words, understood themselves to be working toward the goal of social good. In other words, where *Data Feminism* speaks of the social good, I employ the ‘rising tide lifts all boats’ metaphor for the disembodied view of the social good most common at the HUBs site, and where *Data Feminism* distinguishes co-liberation from social good, I use the ‘seat at the table’ metaphor for the embodied view of the social good—co-liberation *as* social good—most common at Measure. I also note and consider where variations from the analysis of *Data Feminism* appear: for example, where Measure advocates note the [goodness of establishing and maintaining interpersonal ties—i.e., of community or friendship], which seems to be left only implicit in D’Ignazio and Klein’s conceptualization of co-liberation. It may be helpful to refer to table 6.1 (which is the same as table 2.2) in considering the following discussion.

Table 6.1 Data for social good versus co-liberation

_

	Data for Social Good	Data for Co-Liberation
Leadership by members of minoritized groups working in community		X
Money and resources managed by members of minoritized groups		X
Data owned and governed by community		X
Quantitative data analysis is "ground truthed" through a participatory, community-centered data analysis process		X
Data scientists are not rock stars and wizards but rather facilitators and guides		X
Data education and knowledge transfer are part of the project design		X
Building social infrastructure is part of project design		X

(D'Ignazio & Klein, 2020, p. 140)

The lack of clarity around the concept of social good that is manifested by these divergent users is significant, since it obscures p important value differences and commitments. The vague use of the agreeable sounding term social good can also easily be used to hide—sometimes even from those employing it—power dynamics that might harm members of marginalized communities through the reproduction, perpetuation, and exacerbation of injustices resulting from existing power hierarchies. This issue is particularly marked in the 'rising tide lifts all boats' understanding of social good.

Social good: A rising tide lifts all boats

Data Feminism's framework of social good versus co-liberation sharpens the concept of social good by centering power relations – who has it and who does not. Their model of 'social good' does not consider how power is distributed. I agree with this distinction and adopt it here, but choose, for the reasons given above, to utilize the metaphor of ‘a rising tide lifts all boats’ to describe a version of social good that is disembodied, and therefore able to ignore the context of how power is distributed. This understanding of social good matches up with Stone’s market model of society (Stone, 2012). Referring back to the introduction, the two most important parts to this metaphor are:

1. It is not concerned with increasing a person’s or community’s access to democratic decision-making.
2. It does not suggest the necessity for human agency, at least not any agency from those outside locations of political power; the tide will rise whether those outside locations of power cast a vote or not.

Decision-making is located in the market mechanism or the institution, rather than in embodied individuals. The view closely matches that of the neoliberal worldview that is dominant in the wider world. Wendy Brown (2017) connects the expansion of neoliberalism with a greater acceptance of social inequalities. “As liberty is relocated from political to economic life, it becomes subject to the inherent inequality of the latter and is part of what secures that inequality” (p. 41). Because of its greater attention to the embodiment of a variety of actors, the ‘seat at the table’ understanding of social good common among Measure advocates more successfully resists neoliberal logic.

With its greater employment of first/second generation disembodied data discourses, its focus on overall systems, and the relative inattention it pays to the variety of specific embodied locations of all members of society, this understanding of the social good may be relatively blind to disparate effects on differently located persons. A poignant example of how disembodied data discourse can perpetuate and exacerbate social inequalities comes from one of the pioneers of data driven decision making: The Simulmatics group in the 1950s sought to use data to predict human behavior. As documented by the historian Jill Lepore, two of their first major projects were to predict the Black vote and the spending habits of white suburban housewives. In both examples, engineers at Simulmatics never spoke with their populations of interest. Instead, they built computer models and ran simulations. The problem was not simply the exclusionary practices employed as they developed their predictive models, but also the reproduction of underlying and unchecked racial biases like constructing the "Black mind" as a mysterious phenomenon. Data systems "designed for good" in the 1950s and 1960s by white liberal and progressive engineers (who were well-intentioned) ended up writing white supremacy into their technocratic systems, in effect reproducing and amplifying the power hierarchy responsible for the very problems in the Black community these white liberals were trying to solve (Lepore, 2020).

As noted in the previous chapter, both Measure and HUBs data advocates adopted start-up language to describe their work. But HUBs advocates were much more likely to intertwine this language with the 'rising tide' framing of the social good. In the startup model of collaboration, where experimentation and failing fast is part of the game, social good definitions are more loosely defined. The strategy of the HUB is to attract a broad variety of participants to see what works and what doesn't. Because different participants may have different understandings of the social good, the start-up mentality is here associated with broad definitions of social good; the only vague requirement is that it include some kind of social impact. It is assumed that this impact is likely to serve the social good. Table 6.2 shows the salient components of data advocates' conceptualization of social good as 'a rising tide lifts all boats.'

Table 6.2 Social good: A rising tide lifts all boats

Name	Description
Experimental	Start-up model of social good with fail fast mantras to test which partnerships and collaborations work
Social Impacts	Limits or defines social good in terms of the overall (but uneven) impact on society and does not address power distribution
Disembodied trust in neutral third parties	Trust is abstracted from individual bodies to institutions that are perceived as neutral third parties

The 'rising tide' framing of social good is often only alluded to, rather than concretely defined; advocates seemed to assume that it was obvious or intuitive (N6).

One HUBs advocate referred to social good as a catch-all phrase capturing multiple policy domains: "Social good [paints with] a very broad brush. So, it can include projects related to environmental science and projects related to climate change. And this could include health data and health care, but that's also maybe an equity issue. But I would say social good is a catch all" (N3). Social good was taken to encompass most of what HUBs data advocates worked toward. This interviewee continued,

> I think pretty much all the activities we participate in are all in the social good space. And it's part of our mission that we're tackling areas that are for societal benefit and so I would argue pretty much everything we do is touching on that. Again, it depends on how you define social good. The projects we're taking on are at least related to society and have a societal impact. (N3)

HUBs data advocate N1 similarly described social good as appealing to a broad audience. "For us, social good is something where we found that it resonates with a broader group of stakeholders than 'justice' [does], and it may just be that word is sort of easier to grasp."

The appeal to a broad audience was understood to be necessitated by the purpose and scope of the HUBs organization. The following story shows how the HUBs was created on the model of an experimental start up, striving to attract a variety of participants. One advocate used the image of a lighthouse helping ships find their way in the night to describe how HUBs guides organizations toward collaboration and social good:

> We called them these lighthouse stories, the way that different kinds of organizations across different sectors could engage with each other and share resources. And share analy[tical] resources, all of it with the goal of doing some greater social good, right? If you look at the announcement (RFP) you'll see pharm[aceutical] companies that are working with federal agencies to create new banks of data for clinical trials, you'll see AMPLab make an announcement making some of their resources open source, there's also work shared with Google and Google Maps and some of their work on the Amazon. (N7)

This data advocate explained a little later on that, "The purpose of phase1 of the Hubs was actually to experiment with different ways that they could engage with different partners to discover where they could offer the most value. Phase1 was really open and allowed them to do whatever they wanted, to engage in different ways. They had a lot of free rein" (N7). The vagueness of the notion of the social good at the HUBs, and at the National Science Foundation generally, leaves room for different constituent domain experts–academic investigators, government officials, industry leaders–to define it differently. At the HUBs, the relevant domain expert is often the principal investigator on a grant.

> We're coming with a question and how can we use the data to answer that question and what tools might be relevant for that question. And there's a difference between what one might do in the social good space versus in other types of areas of data science. I think when you're doing data science in the social good space it is driven by the questions and the questions are defined by the people for whom these issues matter and who are going to be making the policy or drive the decisions as they're relevant to them. (N3)

According to this understanding, it is the domain expert's understanding of the social good that counts, and it is that understanding that drives research. HUBs advocates seem to have a sort of provisional or instrumental understanding of the social good as serving their constituent domain experts.

Another difference that shows up between these two fundamental framings of the social good is in how data advocates conceptualize where trust is located, whether in an institution, in the data, or in actual embodied human beings. The necessity of neutral institutions is part of the disembodied view of the social good—a rising tide lifts all boats. Data advocates from both sites recognized that trust is still located in institutions and organizations that are perceived as politically neutral. From the perspective of data advocates in this study, without at least the veneer of neutrality, data for good projects could not move forward. According to CDS scholars Reider and Simon (2016), trust operates as the cement of society and, "is essential to the construction and establishment of epistemic systems" (p.3), that is, the social processes that generate what we take to be knowledge and truth. In mapping how the concept of trust has changed over time, they argue that society has progressively transferred their trust from flesh and blood individuals to faceless institutions, and finally to disembodied data. In other words, that our epistemic system has come ever more to ground upon our trust in numbers. This account is in some tension with the data discourse generations heuristic employed in the present study. Data advocates from neither site went so far as to recognize data itself as sufficient to establish trust between groups. The trajectory of trust identified by Reider and Simon makes sense through the first and second generations, but problems emerge in the third generation, where neoliberal logic is challenged, as is naïve optimism about the goodness of all scientific discovery.

As we have seen above, data advocates from the HUBs and Measure sites manifested features of both earlier and later data discourse generations. To the extent that they partake of earlier generations, they do sometimes locate trust in institutions or "neutral third parties" that were perceived as disembodied and abstracted from political agendas and narrow self-interest. Yet, they did not seem to go so far as to simply trust disembodied data themselves. To the extent that they partake of the attitudes of third generation, embodied, data discourse, they located trust differently than Reider and Simon suggest. The only sense in which one could say that they trusted data, is that data advocates from both sites shared a kind of trust in the utility of data in general—data as a kind of institution—yet, they recognized that the degree of trust in any particular data was contingent upon attitudes toward the data's source. One data advocate from the Measure site explained, "People always nay say data" (M2). Furthermore, "You're not going to change things with science or with just science and with the data because people say, and you've heard this a million times, that, 'oh you can make data say anything'" (M2). Echoing this concern, a HUBs data advocate (N5) explained that data could be perceived as “window-dressing,” where decision makers had already made their decision and adjusted their use of data to support that decision, rather than letting their decision be led by the data. As another data advocate from the Measure pointed out, the term ‘evidence based’ is a buzz phrase, used by those without any training in science or data analysis to either advance an agenda or sell a product

Data advocates from both sites did manifest trust in a neutral third-party institution. HUBs data advocates adopted this view as a guiding principle for the entire purpose and activities of the HUBs. In other words, the HUBs organization strove to become the trusted neutral third party.

> All the hubs are in the service, very broadly speaking, of industry, academia, non-profits and government throughout their regions, but also collectively across the nation. I think that is a very powerful motivator because a lot of things that provide value to these extended communities are provided by neutral third parties. There’s a lot of power to that in terms of building collaborations that are based on trust, because we are a neutral third party that’s helping to build those frameworks that will help in collaboration so I think that’s part of the secret sauce. (N2)

For HUBs data advocates, neutrality was important because it distanced the organization’s use of data from particular (located) political agendas and narrow self-interest. “I think social good implies that it’s good for the overarching society. Not in the self-interest of a specific group and those can be in conflict” (N6). Something like this view of institutional neutrality was shared by data advocates from the Measure site as well, insofar as their mission of improving communication between police and communities of color required being seen as a neutral third party or honest broker (Pielke, 2007). “That’s why it’s so important to me that we’re not perceived as being in the advocacy field. There’s so much data out there now, so many people putting it out, and who puts out research even can be very confusing these days. That's why I care so much about transparency...so there's that trust" (M9). Perceived neutrality was seen as necessary for collaboration. "But when we work with a police department, we work with them at no cost to them, because we need to remain neutral and we don't want them or us accused of giving them what they want to hear. They know that we're going to be neutral in it and they will get what they get when they work with us" (M 2).

The rising tide lifts all boats version of social good produces demonstrable social impacts, such as creating neutral institutions and spaces where trust can be built and maintained. However, lacking from these considerations is a concern with, and analysis of context or locadedness of the individuals involved. In the next, I deal with the version of social good that incorporates embodiment by actively incorporating the voices of people who have traditionally been left out of decision making.

Social good: A seat at the table

Whereas the 'rising tide lifts all boats' metaphor is thought to promote a social good regardless of power dynamics, like *Data Feminism*'s co-liberation model, the 'seat at the table' metaphor pays attention to the power dynamics at work. However, where community is mostly left implicit in that model, I want to use this metaphor to include the importance of interpersonal ties and authentic community that emerged from interviews with my participants. This metaphor is meant to suggest that the hearing the voices of all involved people is an intrinsic part of social good. In this view, particular attention is paid to those whose voices have traditionally been silenced and those who do not generally have access to political power. In the discussion that follows, it should be noted that all examples of the seat at the table version of social good come from the Measure site. This is potentially a result of the HUBs structure and purpose, which is meant to appeal to a broader diversity of potential partners than the Measure site.

Let me first discuss that aspect of the seat at the table model that fits with co-liberation, and then highlight the importance of authentic community that is left implicit in co-liberation. This metaphor includes all or most of *Data Feminism*'s framework of co-liberation. According to this framework, a project qualifies as co-liberation if certain conditions are met:

> 1. Leadership by members of minoritized groups, working in community.
> 2. Money and resources managed by members of minoritized groups.
> 3. Data owned and governed by the community.
> 4. Quantitative data analysis is "ground truthed" through a participatory, community-centered data analysis process.
> 5. Data scientists are not seen as 'rock stars' and 'wizards', but rather facilitators and guides.
> 6. Data education and knowledge transfer are part of the project design.
> 7. Building social infrastructure is part of project design.
>
> (D'Ignazio & Klein, 2020, p. 140).

Not all the elements in *Data Feminism*'s list above showed up explicitly in interviews with Measure advocates. However, it is also the case that nothing inconsistent with them showed up. The list offers a characterization of the cultural attitude toward data within Measure that broadly represents that organization's orientation to data collection, analysis, and application. In what follows, I will discuss the three items from this list most explicitly embraced at Measure: Leadership by members of minoritized groups working in community; data owned and governed by the community; and quantitative data analysis "ground truthed" through a participatory, community-centered data analysis process.

Although the first element of leadership was not explicitly discussed in the interviews, Measure describes itself on its website as, “an organization that was founded and is led by Black women” Throughout interviews with data advocates from Measure, data ownership by the community [#3] and ground-truthing through a participatory process [#4] were consistently presented as aspects of social good. By shifting ownership of data to the community and ground truthing in participatory action research, a seat at the table version of social good considers the embodied context of power dynamics.

According to Measure interviewees, not only should community members be included in “ground truthing” data for social good projects, but the ownership of data is shifted from governments and private industry to the community members. “That’s the reason for Measure here in Austin because we see the importance that Black people be the owners of their own data. Because then that translates into them being the narrators of their own story” (M4). The importance of owning the data, for this advocate, came down to having control over narrative. It is the historical existence of a pervasive power hierarchy, grounded in large part on race, and Measure advocates’ intention of working toward undoing that hierarchy, that makes voice, narrative control, and ownership of one’s own data so crucial. According to CDS scholar Tiara Roxanne, local and embodied control over one’s data narrative is a means of resisting ‘data colonialism’ (Couldry & Mejias, 2019), which refers to a combination of the predatory extractive practices of historical colonization with the abstract quantification methods of computation (Roxanne, 2020, p. 154). Because data colonialism comes out of historical colonialism (including domestic colonialism), it is “built from structural racism.” The practice of including voices and ownership of one's own data for the sake of narrative control by members of a community of color is then an act of resistance against data colonialism.

Mirroring the fourth element of the co-liberation model of social good, data advocates understood social good as empowering community by actively involving community members in the data collection process.

> One thing we do is go into neighborhoods and do community design surveys for community development. So we create our surveys, we determine our own standards for data collection and we do the analysis of it, and then we are the ones who create the recommendations based on what we see and learn, and that's why it's so super important to involve those impacted from the beginning. M4

The kind of engagement described by this data advocate is robust and embodied. Often those most affected by data-driven policy decisions experience issues differently than those collecting data and crafting policy. The lived experience of community members is considered to be the necessary ground truthing of any quantitative analysis. In the words of another Measure advocate, "We want people who are experiencing an issue to be able to reach out and know how to use data and present their case and have something that represents their lived experiences and use that to advocate for whatever issue they need" (M5). This advocate went on to describe working with parents at a public school who wanted better lighting for their children to cross the street. "And so, just advocating on things like that, things that matter to them, they get to have that voice" M5.

A significant feature of the seat at the table model that is left implicit in the model of co-liberation found in *Data Feminism* was the value of interpersonal relationships, or authentic community. This feature transcends the boundaries of the individuals. In other words, not only were the voices of all individual participants valued, so also were the relationships among participants. One data advocate, for example, repeatedly emphasized how often they witnessed friendship ties and their importance among members of the organization.

> [At first] I was their police advisor and over time as I got to know them… I've gotten to the point where they're my friends. I've gone beyond the police advisor role and we're good friends and I care about what they do and they care about what I do as a person. Developing those relationships is also about being available for more than just those questions about police and such, so it's actually developing a real relationship, going to have lunch when you don't have to go and have lunch. (M1)

This data advocate described friendships that developed through difficult and shared tasks but also articulated the importance of these relationships for working past challenges, particularly in highly contentious areas.

> We can move on because we've built that foundation, a real foundation. There have been times where I've had to talk to them about some stuff I didn't agree with, or some publication they were about to go live with, and say this is the way I see it, and it might have changed their minds. Because they respect me and I respect them. I know their heart. (M1)

The importance of deep, enduring, mutual relations captured by 'authentic community' is certainly consistent with the attention to embodiment and context indicated by the 'seat at the table' understanding of social good. Yet, little attention is given in the literature to the importance of such profound emotional bonds. This suggests an interesting avenue for further research.

Both metaphors of a rising tide and having a seat at the table are useful conceptions of social good for groups operating under the data for social good banner.

However, like we saw in the data discourses chapter, Measure data advocates manifested a much subtler understanding of the contexts in which each conception was likely to prove most useful, and more skill in switching between them. This political savvy is one of the prominent findings of this research, as we will see in the following section.

Code Switching

As we saw in the previous chapter regarding understandings of data, Measure data advocates manifested the ability to switch between the generations of data discourse regarding the social good according to what they saw as more useful to their purposes in a given context. Because their notion of social good as 'a seat at the table' includes actors from all walks of life, it includes an emphasis on communication across any number of differences in social location. In what follows I will consider two particular manifestations of this sensitivity and practice: switching between social good terminologies and switching between police culture and Measure culture.

Code-switching shows up in Measure data advocates' careful deployment of terminology. For example, they are careful to use words such as 'bias', 'accountability', or 'disparity' instead of more controversial terms like 'racism' or 'social justice'. In *Data Feminism*, D'Ignazio and Klein suggest that terms such as bias and accountability, when addressing data systems and practices, have a tendency to secure power, rather than challenge power. This is due to the fact that these concepts locate the source of the problem in a technical system rather than in a structure of power. (See table 6.4). In this way, we could say that Measure data advocates are not fully embracing a data liberation ethos in the way data feminism prescribes.

Table 6.4 From data ethics to data justice

Concepts that secure power because they locate the source of the problem in individuals or technical systems	Concepts that challenge power because they acknowledge structural power differentials and work toward dismantling them
Ethics	Justice
Bias	Oppression
Fairness	Equity
Accountability	Co-liberation
Transparency	Reflexivity
Understanding the algorithm	Understanding history, context, and culture

(D'Ignazio & Klein, 2020, p. 60)

However, in other ways, Measure advocates all work toward social justice as understood by D'Ignazio and Klein, "Measure is more on the social justice side and we're also for social good" (M5). What is fascinating about the use of terms by Measure advocates is precisely their sensitivity to language, and their careful deployment of terms from both perspectives, in order not to maintain relationships that seem to be highly valued. While data advocates from Measure hold an embodied understanding of social good, and while they may use stronger terms consistent with that understanding when speaking with some audiences, they recognized that using terms like oppression and justice would not be helpful in building collaborative projects between communities of color and police departments. For example, data advocates explicitly employed the term of racial disparity rather than racism. "We elevate data to address what we call disparities," M4. Data advocates are aware of the circumstances and contexts that might necessitate using concepts such as bias, instead of more visceral language like justice. I believe this shows an awareness of the intended audience and the necessity for careful speech. One Measure advocate put it this way, "One of the major challenges is that we have to be careful what we say – we don't say that a police department is racist – we say that we see evidence of implicit bias – this means that we're distancing the blame" (M7).

While the basic purpose of Measure is to disrupt existing power dynamics by bringing more people to seats at the table, the strategic maximization of this goal actually requires sometimes *not* using the language usually associated with it. While D'Ignazio and Klein would say that the language used by data advocates in the Measure site could have an effect of maintaining power rather than challenging it, I argue that this is better understood as a strategy, a kind of code switching, used by data advocates to speak to different audiences depending on the situation. Instead of showing naivete to power, as D'Ignazio and Klein' analysis might be taken to suggest, the strategic deployment of terms according to code switching actually illustrates the superior political acumen of data advocates from this site. If the goal is to use data in a way that promotes collaboration between traditionally contentious groups, such as between citizen activist groups and police departments, then the use of "concepts that challenge power" like equity and oppression, could undermine those efforts.

The second example of code-switching concerns switching between traditional policing culture and Measure culture. Measure is composed of citizen activists from the local community as well as advocates with a background in professional policing. This second subgroup of advocates straddles the two quite different cultures of Measure and policing. For the sake of simplicity, I will refer to these as 'police Measure advocates'. Police Measure advocates share the fundamental commitment of Measure to using data to promote the social good, including the disruption of existing power arrangements, yet they must also function within, and so be sensitive to, the policing culture. (For more on the challenges of this situation, see the discussion in Chapter Four on 'Misfits'.) This group of advocates, too, engages in code switching from their very particular situation advocating a do no harm principle but using a research method that is exclusive to community members, randomized control trials.

Measure advocates in general, and Police Measure advocates in particular, are committed to 'evidence-based policing.' According to these data advocates, evidence-based policing was described with reference to randomized control trials and quasi experiments. They see evidence-based policing as the alternative to "bloodletting," (M1) as one Measure advocate characterized the frequent violent conclusion of so many police encounters with which we are all so familiar from press coverage. Measure advocates are particularly concerned with the lack of evidence-driven policy making in police departments, and the historical immunity of the police from more careful public scrutiny, due to their social location. One police Measure advocate described this immunity and its potential harms like this:

> We create all these interventions all day long. Policing is the last bastion of social science experimentation… The public has no clue. But because we have good intentions, nobody thinks about stopping us from doing it. Nobody makes sure that any of these interventions are effective and nobody makes sure we're not causing harm! (M3).

Policing policies and practices, according to this data advocate, somehow escape oversight and are shielded from public scrutiny due to their perceived good intentions.

A powerful example of the harm done to communities and individuals was told by M3 on the harm done to juveniles who were part of a Cambridge Summerville Youth Study in 1936. In this study, A physician, Dr Richard Cabot, commissioned a study on how mentorship might reduce recidivism in boys.

> What we always want to do with the juveniles is mentor them! So, I love to tell them the story of how all these normal kids and kids on the edges of the criminal justice system were put into a mentoring program…and l love that part of the story is that they had normal kids, kids that had nothing to do with criminal justice system, kids that were just walking around and living their lives like normal human beings that had nothing to do with policing. And guess what, we screwed up those kids, too! M3

The kind of research and evidence that police Measure advocates often suggested in interviews is 'quasi experimental randomized control trials' (RCT), a kind of common quantitative research that is in some ways at odds with the ideals of the social good understood as co-liberation. For example, it is not "ground truthed" through a participatory, community-centered data analysis process (see #4 above), there is not much in the way of oversight or involvement from community groups. It treats data as disembodied. This method is sometimes preferred because it is relatively quick and easy to carry out, and is widely perceived as valid in the broader community, and therefore useful for some purposes. "Scared straight, the DARE program…these things are harmful and we wouldn't have known it unless we had looked at the data, gone through these randomized control trials" (M10).

Police Measure advocates' more basic commitment to the 'seat at the table' understanding of social good appears in their commitment carrying out this kind of research according to a "do no harm" requirement borrowed from the Hippocratic oath.

"It's to do no harm. When you look at the medical field, nothing gets introduced to the public without going through some kind of oversight or governing board, or neutral third party" (M6). The harm, according to data advocates in the Measure site, is a result of a lack of standards or oversight that tests whether programs and policy interventions accomplish what they are supposed to accomplish.

> I did a 34-day randomized control trial, lights on, lights off. Small sample sizes so there are some limitations, but we have some signals so that we can push this. You would normally have to deal with this lengthy academic, IRB, hidden behind paywalls, bureaucracy, and so on. But the data gov model is that this can be easy peasy, one page snap shots, tell us what your trial was, tell us the outcomes and let's move on. M10

One Police Measure advocate described how sometimes even police officers involved in an RCT are unaware of their participation in the program. "And we did that (RCT), but if you really drill down on what we were doing I was turning off activations for cops, turning off the license plate readers for these cops, but no one noticed!" (M10).

This do no harm commitment also characterizes police Measure advocates' general attitude toward policing. As another of these advocates put it,

> How do we know the way we're doing things is actually working? This might be more harmful. There are a lot of studies out there that show this. Like scared straight, the DARE program…these things are harmful and we wouldn't have known it unless we had looked at the data, gone through these randomized control trials. M10

This means that without the do no harm principle guiding policing practices, there is no ground on which to challenge the ineffectiveness of a program or articulate the social harms. Among the social harms that were identified were the harm of perpetuating racial bias through the implementation of policing policies that can destroy communities, families and individuals. One police Measure advocate compared such harms to adverse side effects from medications.

This means that without the do no harm principle guiding policing practices, there is no ground on which to challenge the ineffectiveness of a program or articulate the social harms. Among the social harms that were identified were the harm of perpetuating racial bias through the implementation of policing policies that can destroy communities, families and individuals. One police Measure advocate compared such harms to adverse side effects from medications. “It [the policing policy] took black and brown males, and disproportionately locked them up. You look at how you destroy the families and kids growing up with one parent who is working two jobs and now these male blacks have convictions and records and can’t get a job, and you’re talking about a cycle that just perpetuates itself” M6.

This police Measure advocate clearly shares Measure’s basic concern for undoing toxic racial hierarchy. The statement also shows the concern for the police mission. These advocates, straddling two cultures, often feel like misfits in their home departments. They, too, must strategically code switch in accordance with a well-tuned sense of context. Recall the words of M6 quoted in Chapter four above, “I was kind of a threat to [someone in the police department] and he pooh-poohed a lot of the things I did. So I had to do things kind of quietly, and I always tell people, do things quietly[...]. Partner with one or two other people and just do it.” M10 recognized the dangers of manifesting the use of data to ensure equity in the police world, saying that police peers might see them as “part of some cult,” so that they were “putting my badge on the line” (M10).

Code-switching shows a profound understanding of the complexity faced by data advocates in the social good space. Advocates who share the ‘seat at the table’ understanding of the social good must recognize the vast variety of locations and associated interpretations that are active in the world. While this certainly complicates the mission of data for good advocacy, it also promises stronger objectivity that speaks to our concern for accurate knowledge, and it manifests a more robust respect for human equality.

CONCLUSION

This research asks how data can be used to promote social good. Through a qualitative comparison of two dissimilar sites, I have shown how the concept of embodiment in a third-generation environment challenges and informs disembodied notions of data and social good informed by first and second-generations. In this research, data for social good manifests as a space where data advocates negotiate embodied and disembodied meanings of data.

This research question is important as we continue to build data systems for social good, often doing so without much thought about what data for social good looks like and the structures needed to support that good. By affixing data to social good, we run the risk of focusing on data problems and solutions alone but we cannot automate social good with data and big data. Data is only a tool created and used within an existing context of culture, history, and politics. Because it is always embedded in context, data reflect the best and worst of us.

I argue in this research that by framing data and the social good within third generation embodiment, we might be less likely to reproduce some of the harms caused by data systems in the past. These harms play out in predictable ways within a neoliberal context, which seeks to reduce all lived experiences into economic terms and market logic. This dissertation has shown the ways the emergence of the concept of embodiment challenges that logic by reimagining who can use data, what data are, and by insisting that power dynamics matter when defining social good. Ultimately, both sites understand their work as using data to promote the social good. However, conceptions of data for social good coming out of the HUBs site do not adequately take into account existing issues of power distributions and how that particular context can reproduce harm to marginalized communities. Because of the purpose of Measure, and the data advocates' proximity to street level policing making (advocates from this site come from activist organizations, non-profits, street-level and more senior administrators in government agencies, and groups of citizens–essentially, what we might think of as those making up civil society), data advocates from this site more readily articulated embodied data discourses and a seat at the table version of social good, which takes into account power dynamics.

One important development that challenges and changes existing power dynamics is the broadened view of expertise. Both the HUBs and Measure have adopted a broadened view of who counts as an expert. This is a critical shift in the possibilities of including more voices in future digital infrastructures. However, the expert as a community member with lived experience, someone who has lived crime or who recognizes that how they feel about an issue like poverty (embodied) is not captured in current data sets and stories, is not actively pursued at the HUBs as it is in Measure. The HUBs structure was not created to accomplish this task and its current efforts to broaden focus more on contextualizing data within a substantive social science academic field than community members, especially communities of color and others who are marginalized.

By recognizing the expertise of those who experience power dynamics differently, third generation environments open the possibility for embodied data discourses to redefine what counts as data. From the perspective of embodiment, data are no longer abstracted from flesh and blood bodies. Instead of playing the god trick, data can only “see” as a drone would see, providing a higher, abstracted view that is still tethered to bodies on the ground. This framing of data allows for the challenging of dominant narratives based on first and second generation discourses. It replaces naive objectivity with strong objectivity.

Embodied data discourses offer an alternative to the traditional naïve objectivity It is also the case that embodiment can inform conceptions of social good that work to change the distribution of political and economic power in a way that levels or democratizes the practice of power. When data is applied to this ‘seat at the table’ version of social good, those who find themselves on the margins of political power are able to move into spaces where decisions are often made for them by those at the top. Embodied data discourses can help to facilitate this democratizing effect. Unfortunately, disembodied discourses in the form of neoliberalism often maintain the status quo power hierarchy. Those at the top stay at the top and those at the bottom stay at the bottom. Even if those at the bottom are somehow materially better off than several decades ago, their access to political power, relative to those at the top, remains the same.

Even though disembodied discourses tend to align with perpetuating social harms, data advocates from Measure practiced code switching between disembodied and embodied discourses. I argue that code switching is an active and intentional strategy employed by data advocates for effective collaboration. For example, by employing disembodied data discourses, advocates could lower the emotional temperature of the room. And data as authoritative provides a justification and credibility to marginalized experiences. And by using terms such as racial disparity instead of racism, work with police departments could continue. Even though data for social good projects should work to adopt embodied discourses, the phenomenon of code-switching shows that disembodied discourses should not be abandoned and have a positive role to play in data for social good projects.

Code switching shows the positive role played by disembodied discourses in data for good projects, however, concerns remain over the inheritance of neoliberal assumptions that can change how data for good efforts, like collaboration, manifest in third-generation environments. This is especially the case with the reliance from both sites on start-up language and the gendered role of the Data Yenta where the established hierarchy at the national science foundation reproduces patriarchal roles at the HUBs.

Embodiment in the data for good movement is an exciting phenomenon. This line of research explicitly shows that data systems and processes do not appear out of nowhere and that the path is not predetermined. By exploring the role of embodiment, we have a concept that allows us to recognize how power dynamics play out and gives us an alternative to challenge neoliberalism.

My research contributes to the conversations in CDS and data feminism on how to build more equitable data infrastructures by exploring the role played by embodiment in the data for good space. However, this study is limited to one specific instance of data for good. Further research is needed to apply the concept of embodiment to more cases of data for good and track the extent to which this concept is challenging and shaping data practices. This research on embodiment in data discourses also suggests that cultural literacy is a key part of data literacy. Policy studies, by incorporating embodiment and the practices of code switching, could deepen understanding of data’s role within the policy making process, for example, in policy implementation. Furthermore, policy studies could benefit from adopting frameworks from Science and Technology studies and Critical Data Studies which understand data and technology as situated in social and political contexts. This potentially changes some of the questions asked in the policy process about how data are used and deepens understandings of who ought to be included in the policy process and why.

REFERENCES

Abma, T. A., & Noordegraaf, M. (2003). Public Managers amidst Ambiguity. *Evaluation*, *9*(3), 285–306.

Amrute, S. (2019). Of Techno-Ethics and Techno-Affects. *Feminist Review*, *123*(1), 56–73.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 1461444816676645.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias *ProPublica*. Machine Bias — ProPublica

Arantes, J., & Buchanan, R. (2022). Educational data advocates: emerging forms of teacher agency in postdigital classrooms. *Learning, Media and Technology*, *0*(0), 1–21.

Arnold, G. (2021). Does Entrepreneurship Work? Understanding What Policy Entrepreneurs Do and Whether It Matters. *Policy Studies Journal*, *49*(4), 968–991.

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography*, *3*(3), 297–302.

Barnes, T. J., & Wilson, M. W. (2014). Big Data, social physics, and spatial analysis: The early years. *Big Data & Society*, *1*(1), 1-14.

Beer, D. (2009). Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society*, *11*(6), 985–1002.

Benjamin, R. (2020). Race After Technology: Abolitionist Tools for the New Jim Code.

Social Forces, *98*(4), 1–3.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, *50*(1), 3–44.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: classification and its consequences*. MIT Press.

Bowker, G. C., Timmermans, S., Clarke, A. E., & Balka, E. (2015). *Boundary objects and beyond: working with Leigh Star*. MIT Press.

boyd, d. (2019). Differential Privacy in the 2020 Decennial Census and the Implications for Available Data Products. Available at SSRN: https://ssrn.com/abstract=3416572.

boyd, d., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679.

boyd, d., & Golebiewski, M. (2019, October 29). *Data Voids*. Data & Society. https://datasociety.net/library/data-voids/.

Boyles, J. L., & Meyer, E. (2016). Letting the Data Speak. *Digital Journalism*, *4*(7), 944–954.

Brayne S. (2017). Big Data Surveillance: The Case of Policing. *American Sociological Review*, *82*(5), 977–1008.

Brown, W. (2017). *Undoing the Demos: Neoliberalism's Stealth Revolution* (Reprint edition). Zone Books.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1-12.

Burton, A., & Confino, P. (2023, January). *ChatGPT could make this HR function obsolete*. https://fortune.com/2023/01/26/chatgpt-hr-recruiter-talent-obsolete/

Catlett, C., & Ghani, R. (2015). Big Data for Social Good. *Big Data*, *3*(1), 1–2.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 1-14.

Cohen, N., & Horev, T. (2017). Policy entrepreneurship and policy networks in healthcare systems – the case of Israel's pediatric dentistry reform. *Israel Journal of Health Policy Research*, *6*(1), 24.

Coleman, R., & Banning, S. (2016). Network TV News' Affective Framing of the Presidential Candidates: Evidence for a Second-Level Agenda-Setting Effect through Visual Framing: *Journalism & Mass Communication Quarterly*.

Coleman, R., & Wu, H. D. (2010). Proposing Emotion as a Dimension of Affective Agenda Setting: Separating Affect into Two Components and Comparing Their Second-Level Effects. *Journalism & Mass Communication Quarterly*, *87*(2), 315–327.

Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, *20*(4), 336–349.

Dalton, C. M., & Thatcher, J. (2015). Inflated granularity: Spatial "Big Data" and geodemographics. *Big Data & Society*, *2*(2), 1-15.

Dalton, C. M., Taylor, L., & Thatcher (alphabetical), J. (2016). Critical Data Studies: A dialog on data and space. *Big Data & Society Big Data & Society*, *3*(1), 1-9.

D'Ignazio, C., & Klein, L. (2020). *Data Feminism*. MIT Press.

Döerr, B. (2017). Profile: An Interview with Michael Markie - an open science and open data advocate. *Medical Writing*, *26*, 68–69.

Egliston, B., & Carter, M. (2021). Critical questions for Facebook's virtual reality: data, power and the metaverse. *Internet Policy Review*, *10*(4).

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment. *Criminal Justice* *and Behavior*, *46*(2), 185–209.

El-Taliawi, O. G., Goyal, N., & Howlett, M. (2021). Holding out the promise of Lasswell's dream: Big data analytics in public policy research and teaching. *Review of Policy Research*, *38*(6), 640–660.

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*. 21, 403-418.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Ferguson, A. G. (2017). *The rise of big data policing: surveillance, race, and the future of law enforcement*. NYU Press.

Floreano, D., & Wood, R. J. (2015). Science, technology and the future of small autonomous drones. *Nature*, *521*(7553), 460–466.

Floridi, L. (2012). Big Data and Their Epistemological Challenge [SSRN Scholarly Paper]. Available at https://ssrn.com/abstract=3854411.

Fotopoulou, A. (2019). Embodiment and the ethics of care. In *Citizen Media and Practice*. Routledge.

Fotopoulou, A. (2021). Conceptualising critical data literacies for civil society organisations: agency, care, and social responsibility. *Information, Communication & Society*, *24*(11), 1640–1657.

Foucault, M. (1995). *Discipline & Punish: The Birth of the Prison* (A. Sheridan, Trans.). Vintage Books.

Fowler, L. (2022). Using the Multiple Streams Framework to Connect Policy Adoption to Implementation. *Policy Studies Journal*, *50*(3), 615–639.

Fraser, A. (2019). Curating digital geographies in an era of data colonialism. *Geoforum*, *104*, 193–200.

Friedman, B., & Nissenbaum, H. (1996) "Bias in Computer Systems." *ACM Transactions on Information Systems*, vol. 14, no. 3.

Gitelman, L. (2013). *Raw Data Is an Oxymoron*. MIT Press.

Goel, S., Perelman, M., Shroff, R., & Sklansky, D. A. (2016). *Combatting Police Discrimination in the Age of Big Data* (SSRN Scholarly Paper ID 2787101).

Guo, E., & Noori, H. (2021, August 30). *This is the real story of the Afghan biometric databases abandoned to the Taliban*. MIT Technology Review. https://www.technologyreview.com/2021/08/30/1033941/afghanistan-biometric-databases-us-military-40-data-points/.

Gutiérrez, M. (2018). *Data Activism and Social Change*. Springer.

Hallberg, M., & Kullenberg, C. (2019). Happiness Studies: Co-Production of Social Science and Social Order. *Nordic Journal of Science and Technology Studies*, *7*(1), 42–50.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575–599.

Harding, S. (1995). "Strong objectivity": A response to the new objectivity question. *Synthese*, *104*(3), 331–349. https://doi.org/10.1007/BF01064504

Harding, S. (2015). Objectivity and Diversity: Another Logic of Scientific Research. In *Objectivity and Diversity*. University of Chicago Press.

Head, B. W., & Alford, J. (2015). Wicked Problems: Implications for Public Policy and Management. *Administration & Society*, *47*(6), 711–739.

Heitmueller, A., Henderson, S., Warburton, W., Elmagarmid, A., Pentland, A. "Sandy," & Darzi, A. (2014). Developing Public Policy To Advance The Use Of Big Data In Health Care. *Health Affairs*, *33*(9), 1523–1530.

Hess, D. (1997). *Science Studies: An Advanced Introduction*. NYU Press.

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016a). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce; Mahwah*, *26*(1–2), 147–169.

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016b). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and*

Electronic Commerce, *26*(1–2), 147–169.

Hooker, S. (2018, July 22). *Why "data for good" lacks precision.* Medium. https://towardsdatascience.com/why-data-for-good-lacks-precision-87fb48e341f1

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society Big Data & Society*, *3*(2), 1-7.

Ingrams, A. (2019). Public Values in the Age of Big Data: A Public Information Perspective. *Policy & Internet*, *11*(2), 128–148.

Irvin, R. A., & Stansbury, J. (2004). Citizen Participation in Decision Making: Is It Worth the Effort? *PUAR Public Administration Review*, *64*(1), 55–65.

Jackson, L. (2019). The smiling philosopher: Emotional labor, gender, and harassment in conference spaces. *Educational Philosophy & Theory*, *51*(7), 693–701.

Jarmin, R. S., & O'Hara, A. B. (2016). Big data and the transformation of public policy analysis. *Journal of Policy Analysis and Management*, *35*(3), 715–721.

Jasanoff, S. (Ed.). (2006). *States of Knowledge: The Co-Production of Science and the Social Order* (1st edition). Routledge.

Kettl, D. F. (2000). Public Administration at the Millennium: The State of the Field. *Journal of Public Administration Research and Theory*, *10*(1), 7–34.

Kettl, D. F. (2017). *Little Bites of Big Data for Public Policy*. CQ Press.

Kingdon, J. W. (2011). *Agendas, Alternatives, and Public Policies, Update Edition, with an Epilogue on Health Care* (2 edition). Pearson.

Kitchin, R. (2014). *The data revolution: big data, open data, data infrastructures & their consequences*. SAGE Publications Ltd.

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society Information, Communication & Society*, *20*(1), 14–29.

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal;*

Dordrecht, *80*(4), 463–475.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1).

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). A governance framework for algorithmic accountability and transparency. Report for the European Parliamentary Research Service Panel for the Future of Science and Technology. https://www.cs.ox.ac.uk/publications/publication13748-abstract.html

Krasas, J. (2018). The Work of Code Switching: Implications for Gender and Racial Inequality in Employment. *Religion & Theology*, *25*(3/4), 190–207.

Latonero, M., & Gold, Z. (2015). Data, Human Rights & Human Security. [SSRN scholarly

article]. https://ssrn.com/abstract=2643728 or http://dx.doi.org/10.2139/ssrn.2643728

Leigh Star, S. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values*, *35*(5), 601–617.

Lepore, J. (2020). *If Then: How the Simulmatics Corporation Invented the Future* (Illustrated edition). Liveright.

Levy, K. E., & Johns, D. M. (2016). When open data is a Trojan Horse: The weaponization of transparency in science and governance. *Big Data & Society*, *3*(1).

Lindquist, E. (2016, August 26). *BD Spokes: PLANNING: WEST: Big Data and Criminal Justice in the Western United States.* Proposal to the National Science Foundation. Award Abstract # 1636962.

Lönngren, J., & van Poeck, K. (2021). Wicked problems: a mapping review of the literature. *International Journal of Sustainable Development & World Ecology*, *28*(6), 481–502.

Lowrey, A. (2023, January 20). *How ChatGPT Will Destabilize White-Collar Work*. The

Atlantic. https://www.theatlantic.com/ideas/archive/2023/01/chatgpt-ai-economy-automation-jobs/672767/

Luker, K. (2009). *Salsa Dancing into the Social Sciences*. Harvard University Press.

Lupton, D. (2017). Feeling your data: Touch and making sense of personal digital data. *New Media & Society*, *19*(10), 1599–1614.

Machiavelli, Niccolò., & Bondanella, P. (1984). *The prince*. Oxford University Press.

Marche, S. (2022, December 6). *The College Essay Is Dead*. The Atlantic. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/

Marcus, G. E., MacKuen, M., & Neuman, W. R. (2008). *Affective intelligence and political judgment*. Univ. of Chicago Press.

Marr, B. (2023, January). *How ChatGPT And Natural Language Technology Might Affect Your Job If You Are A Computer Programmer*. Forbes. https://www.forbes.com/sites/bernardmarr/2023/01/23/how-chatgpt-and-natural-language-technology-might-affect-your-job-if-you-are-a-computer-programmer/

Martin-Alcoff, L. (2013). Embodiment: Introduction. *Hypatia*, *Virtual Issues*. https://onlinelibrary.wiley.com/page/journal/15272001/homepage/VirtualIssuesPage.html#embodiment.

Mayer, J. (2016). *Dark Money: The Hidden History of the Billionaires Behind the Rise of the Radical Right*. Knopf Doubleday Publishing Group.

Mayer-Schönberger, V., & Cukier, K. (2014). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt..

McCluney, C. L., Durkee, M. I., Smith, R. E., Robotham, K. J., & Lee, S. S.-L. (2021). To be, or not to be…Black: The effects of racial codeswitching on perceived professionalism in the workplace. *Journal of Experimental Social Psychology*, *97*.

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction*, *14*(4).

Milan, S., & Gutiérrez, M. (2015). Citizens' media meets big data: the emergence of data activism. *MEDIACIONES*, *11*(14), 120–133.

Milan, S., & Treré, E. (2019). Big Data from the South(s): Beyond Data Universalism. *Television & New Media*, *20*(4), 319–335.

Milan, S., & Velden, L. van der. (2016). The Alternative Epistemologies of Data Activism. *Digital Culture & Society*, *2*(2), 57–74.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2).

Miller, C. A., & Wyborn, C. (2020). Co-production in global sustainability: Histories and theories. *Environmental Science & Policy*, *113*, 88–95.

Mintrom, M., & Norman, P. (2009). Policy Entrepreneurship and Policy Change. *Policy Studies Journal*, *37*(4), 649–667.

Mühlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., & Streit, M. (2014). Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1643–1652.

Narassimhan, E., Koester, S., & Gallagher, K. S. (2022). Carbon Pricing in the US: Examining State-Level Policy Support and Federal Resistance. *Politics and Governance*, *10*(1).

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press http://ebookcentral.proquest.com/lib/boisestate/detail.action?docID=4834260

O’Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown, NY.

Pangrazio, L., & Selwyn, N. (2021). Towards a school-based ‘critical data education.’ *Pedagogy, Culture & Society*, *29*(3), 431–448.

Pielke, R. A. (2007). *The honest broker: making sense of science in policy and politics*. Cambridge University Press.

Pirog M.A. (2014). Data will drive innovation in public policy and management research in the next decade. *J. Policy Anal. Manage. Journal of Policy Analysis and Management*, *33*(2), 537–543.

Pozen, D. E. (2018). Transparency’s Ideological Drift. *Yale Law Journal*, *128*(1), 100–165.

Ramírez, E. E., Castillo, M. F., & Sánchez, E. I. (2022). How Policy Entrepreneurs Encourage or Hinder Urban Growth Within a Political Market. *Urban Affairs Review*, p.1.

Read, R., Taithe, B., & Mac Ginty, R. (2016). Data hubris? Humanitarian information systems and the mirage of technology. *Third World Quarterly*, *37*(8), 1314–1331.

Ricaurte, P. (2019). Data Epistemologies, The Coloniality of Power, and Resistance. *Television & New Media*, *20*(4), 350–365.

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1).

Rosenberg, D. (2013). Data before the fact. In *"Raw data is an oxymoron* (pp. 15–40). The MIT Press.

Roxanne, T. (2020). Data Colonialism: Decolonial Gestures of Storytelling. *Donau*

Reader. https://www.academia.edu/42641770/Data_Colonialism_Decolonial_Gestures_of_Storytelling.

Ruijer, E., Grimmelikhuijsen, S., & Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly*, *34*(1), 45–52.

Sanchez, M. (2023, January 26). As ChatGPT flourishes, university debates its merits. The Tulane Hullabaloo. *The Tulane Hullabaloo*. https://tulanehullabaloo.com/62237/news/as-chat-gpt-flourishes-university-debates-its-merits/

Sander, I. (2020). What is critical big data literacy and how can it be implemented? *Internet Policy Review*, *9*(2).

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2016). Automation, Algorithms, and Politics | When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software. *International Journal of Communication*, *10*(0), 19.

Schäfer, M. T., & Es, K. van (Eds.). (2017). *The Datafied Society: Studying Culture Through Data*. Amsterdam University Press.

Scott, J. C. (1998). *Seeing like a state: how certain schemes to improve the human condition have failed*. Yale University Press.

Segura, M. S., & Waisbord, S. (2019). Between Data Capitalism and Data Citizenship. *Television & New Media*, *20*(4), 412–419.

Shah, K. (2018, November 10). “Textbook voter suppression”: Georgia’s bitter election a battle years in the making. *The Guardian*. https://www.theguardian.com/us-news/2018/nov/10/georgia-election-recount-stacey-abrams-brian-kemp

Shine, K. T., & Bartley, B. (2011). Whose Evidence Base? The Dynamic Effects of

Ownership, Receptivity and Values on Collaborative Evidence-Informed Policy Making. *Evidence & Policy: A Journal of Research, Debate and Practice*, *7*(4), 511–530.

Simon, H. A. (1976). *Administrative Behavior, 4th Edition*. Simon and Schuster.

Solon, O. (2016, December 12). 2016: the year Facebook became the bad guy. *The Guardian*. 2016: the year Facebook became the bad guy | Facebook | The Guardian

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, “translations”, and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*, 387–420.

Stone, D. (2012). *Policy paradox: the art of political decision making*. New York: W.W. Norton & Co.

Sumartojo, S., Pink, S., Lupton, D., & LaBond, C. H. (2016). The affective intensities of datafied space. *Emotion, Space and Society*, *21*, 33–40.

Thamkittikasem, J., Saunders, B., & Jin, K. (2019). *New York City Automated Decision Systems Task Force Report*.

Thatcher, J., O’Sullivan, D., & Mahmoudi, D. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space*, *34*(6), 990–1006.

Townsend, A. M. (2013). *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. W. W. Norton & Company.

Tracy, S. J. (2012). *Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact*. John Wiley & Sons.

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, *383*(9), 874–882.

Wagner-Pacifici, R., Mohr, J. W., & Breiger, R. L. (2015). Ontologies, methodologies, and new uses of Big Data in the social and cultural sciences. *Big Data & Society*, *2*(2).

Williams, S. (2020). *Data Action: Using Data for Public Good*. The MIT Press.

Wilson, W. (1887). The Study of Administration. *Political Science Quarterly Political Science Quarterly*, *2*(2), 197.

Wingard, D. J. (2023). *ChatGPT: A Threat To Higher Education?* Forbes. https://www.forbes.com/sites/jasonwingard/2023/01/10/chatgpt-a-threat-to-higher-education/

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136.

Young, I. M. (1980). Throwing like a Girl: A Phenomenology of Feminine Body Comportment Motility and Spatiality. *Human Studies*, *3*(2), 137–156.

Zachmann, G., Serwaah-Panin, A., & Peruzzi, M. (2015). When and How to Support Renewables? Letting the Data Speak. In A. Ansuategi, J. Delgado, & I. Galarraga (Eds.), *Green Energy and Efficiency: An Economic Perspective* (pp. 291–332). Springer International Publishing.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Public Affairs, NY.