

Introduction: Why Data Science Needs Feminism

Christine Mann Darden first passed through the gates of NASA's Langley Research Center in Hampton, Virginia, in the summer of 1967. Her newly minted master's degree in applied math had earned her a position as a data analyst there. In the city of Hampton, and across the United States, tensions were running high. In Los Angeles, a massive protest against the Vietnam War ended only when more than a thousand armed police officers attacked the peaceful protestors. One month later, even more violence engulfed the city of Detroit after a police raid spiraled out of control. The 1967 Detroit Riot, or the 1967 Detroit Rebellion, as it is increasingly known, resulted in over forty deaths and one thousand injuries.

The gates of Langley might have shielded Darden from those physical confrontations, but her work there was no less removed from the national stage. By 1967, the space race was well underway, and the United States was losing. The Soviet Union had already sent a man into space and a rocket to the moon. The only thing standing in the way of a Soviet victory was to put those two pieces together. Meanwhile, the United States had suffered a series of defeats—and, in January of that year, an outright disaster, when a sudden fire during a launch test of the *Apollo 1* spacecraft killed all three astronauts on board.

While the nation mourned, everyone at NASA threw themselves back into their work—including Darden, who began her data analyst job in the immediate aftermath of the *Apollo 1* disaster. Two years later, it would be her precise analysis of the physics of rocket reentry that would help to ensure the successful return of the *Apollo 11* mission from the moon, effectively winning the space race for the United States. But it would be Darden herself, as a Black woman with technical expertise, working at a federal agency in which sexism and racism openly prevailed, who demonstrated that the ideological mission of the United States—as a land based on the ideals of liberty, equality, and opportunity for all—was far from accomplished (figure 0.1).



Figure 0.1

Christine Darden in the control room of the Unitary Plan Wind Tunnel at NASA's Langley Research Center in 1975. Courtesy of NASA.

The 1960s, after all, were years of social protest and transformation as well as exploration into outer space. Darden herself had participated in several lunch-counter sit-ins at the Hampton Institute, the historically Black college that she attended for her undergraduate studies.¹ By the time that Darden joined NASA, its Virginia facility had been officially desegregated for several years. But it had yet to reckon with the unofficial segregation of both race and gender that remained in place—particularly among its women employees known as *computers*.

Darden's arrival at Langley coincided with the early days of digital computing. Although Langley could claim one of the most advanced computing systems of the time—an IBM 704, the first computer to support floating-point math—its resources were still limited. For most data analysis tasks, Langley's Advanced Computing Division

relied upon human computers like Darden herself. These computers were all women, trained in math or a related field, and tasked with performing the calculations that determined everything from the best wing shape for an airplane, to the best flight path to the moon. But despite the crucial roles they played in advancing this and other NASA research, they were treated like unskilled temporary workers. They were brought into research groups on a project-by-project basis, often without even being told anything about the source of the data they were asked to analyze. Most of the engineers, who were predominantly men, never even bothered to learn the computers' names.

These women computers have only recently begun to receive credit for their crucial work, thanks to scholars of the history of computing—and to journalists like Margot Lee Shetterly, whose book, *Hidden Figures: The American Dream and the Untold Story of the Black Women Who Helped Win the Space Race*, along with its film adaptation, is responsible for bringing Christine Darden's story into the public eye.² Her story, like those of her colleagues, is one of hard work under discriminatory conditions. Each of these women computers was required to advocate for herself—and some, like Darden, chose also to advocate for others. It is because of both her contributions to data science and her advocacy for women that we have chosen to begin our book, *Data Feminism*, with Darden's story. For *feminism* begins with a belief in the "political, social, and economic equality of the sexes," as the *Merriam-Webster Dictionary* defines the term—as does, for the record, Beyoncé.³ And any definition of feminism also necessarily includes the activist work that is required to turn that belief into reality. In *Data Feminism*, we bring these two aspects of feminism together, demonstrating a way of thinking about data, their analysis, and their display, that is informed by this tradition of feminist activism as well as the legacy of feminist critical thought.

As for Darden, she did not only apply her skills of data analysis to spaceflight trajectories; she also applied them to her own career path. After working at Langley for a number of years, she began to notice two distinct patterns in her workplace: men with math credentials were placed in engineering positions, where they could be promoted through the ranks of the civil service, while women with the same degrees were sent to the computing pools, where they languished until they retired or quit. She did not want to become one of those women, nor did she want others to experience the same fate. So she gathered up her courage and decided to approach the chief of her division to ask him why. As Darden, now seventy-five, told Shetterly in an interview for *Hidden Figures*, his response was sobering: "Well, nobody's ever complained," he told Darden. "The women seem to be happy doing that, so that's just what they do."

In today's world, Darden might have gotten her boss fired—or at least served with an Equal Employment Opportunity Commission complaint. But at the time that

Darden posed her question, stereotypical remarks about “what women do” were par for the course. In fact, challenging assumptions about what women could or couldn’t do—especially in the workplace—was the central subject of Betty Friedan’s best-selling book, *The Feminine Mystique*. Published in 1963, *The Feminine Mystique* is often credited with starting feminism’s so-called second wave.⁴ Fed up with the enforced return to domesticity following the end of World War II, and inspired by the national conversation about equality of opportunity prompted by the civil rights movement, women across the United States began to organize around a wide range of issues, including reproductive rights and domestic violence, as well as the workplace inequality and restrictive gender roles that Darden faced at Langley.

That said, Darden’s specific experience as a Black woman with a full-time job was quite different than that of a white suburban housewife—the central focus of *The Feminine Mystique*. And when critics rightly called out Friedan for failing to acknowledge the range of experiences of women in the United States (and abroad), it was women like Darden, among many others, whom they had in mind. In *Feminist Theory: From Margin to Center*, another landmark feminist book published in 1984, bell hooks puts it plainly: “[Friedan] did not discuss who would be called in to take care of the children and maintain the home if more women like herself were freed from their house labor and given equal access with white men to the professions. She did not speak of the needs of women without men, without children, without homes. She ignored the existence of all non-white women and poor white women. She did not tell readers whether it was more fulfilling to be a maid, a babysitter, a factory worker, a clerk, or a prostitute than to be a leisure-class housewife.”⁵

In other words, Friedan had failed to consider how those additional dimensions of individual and group identity—like race and class, not to mention sexuality, ability, age, religion, and geography, among many others—intersect with each other to determine one’s experience in the world. Although this concept—*intersectionality*—did not have a name when hooks described it, the idea that these dimensions cannot be examined in isolation from each other has a much longer intellectual history.⁶ Then, as now, key scholars and activists were deeply attuned to how the racism embedded in US culture, coupled with many other forms of oppression, made it impossible to claim a common experience—or a common movement—for all women everywhere. Instead, what was needed was “the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking.”⁷ These words are from the Combahee River Collective Statement, written in 1978 by the famed Black feminist activist group out of Boston. In this book, we draw heavily from intersectionality and other concepts developed through the work of *Black feminist*

scholars and activists because they offer some of the best ways for negotiating this multidimensional terrain.

Indeed, feminism must be intersectional if it seeks to address the challenges of the present moment. We write as two straight, white women based in the United States, with four advanced degrees and five kids between us. We identify as middle-class and *cisgender*—meaning that our gender identity matches the sex that we were assigned at birth. We have experienced sexism in various ways at different points of our lives—being women in tech and academia, birthing and breastfeeding babies, and trying to advocate for ourselves and our bodies in a male-dominated health care system. But we haven't experienced sexism in ways that other women certainly have or that nonbinary people have, for there are many dimensions of our shared identity, as the authors of this book, that align with dominant group positions. This fact makes it impossible for us to speak *from experience* about some oppressive forces—racism, for example. But it doesn't make it impossible for us to educate ourselves and then speak about racism and the role that white people play in upholding it. Or to challenge ableism and the role that abled people play in upholding it. Or to speak about class and wealth inequalities and the role that well-educated, well-off people play in maintaining those. Or to believe in the logic of *co-liberation*. Or to advocate for justice through equity. Indeed, a central aim of this book is to describe a form of intersectional feminism that takes the inequities of the present moment as its starting point and begins its own work by asking: How can we use data to remake the world?⁸

This is a complex and weighty task, and it will necessarily remain unfinished. But its size and scope need not stop us—or you, the readers of this book—from taking additional steps toward justice. Consider Christine Darden, who, after speaking up to her division chief, heard nothing from him but radio silence. But then, two weeks later, she was indeed promoted and transferred to a group focused on sonic boom research. In her new position, Darden was able to begin directing her own research projects and collaborate with colleagues of all genders as a peer. Her self-advocacy serves as a model: a sustained attention to how systems of oppression intersect with each other, informed by the knowledge that comes from direct experience. It offers a guide for challenging power and working toward justice.

What Is Data Feminism?

Christine Darden would go on to conduct groundbreaking research on sonic boom minimization techniques, author more than sixty scientific papers in the field of computational fluid dynamics, and earn her PhD in mechanical engineering—all while

"juggling the duties of Girl Scout mom, Sunday school teacher, trips to music lessons, and homemaker," Shetterly reports. But even as she ascended the professional ranks, she could tell that her scientific accomplishments were still not being recognized as readily as those of her male counterparts; the men, it seemed, received promotions far more quickly.

Darden consulted with Langley's Equal Opportunity Office, where a white woman by the name of Gloria Champine had been compiling a set of statistics about gender and rank. The data confirmed Darden's direct experience: that women and men—even those with identical academic credentials, publication records, and performance reviews—were promoted at vastly different rates. Champine recognized that her data could support Darden in her pursuit of a promotion and, furthermore, that these data could help communicate the systemic nature of the problem at hand. Champine visualized the data in the form of a bar chart, and presented the chart to the director of Darden's division.⁹ He was "shocked at the disparity," Shetterly reports, and Darden received the promotion she had long deserved.¹⁰ Darden would advance to the top rank in the federal civil service, the first Black woman at Langley to do so. By the time that she retired from NASA, in 2007, Darden was a director herself.¹¹

Although Darden's rise into the leadership ranks at NASA was largely the result of her own knowledge, experience, and grit, her story is one that we can only tell as a result of the past several decades of feminist activism and critical thought. It was a national feminist movement that brought women's issues to the forefront of US cultural politics, and the changes brought about by that movement were vast. They included both the shifting gender roles that pointed Darden in the direction of employment at NASA and the creation of reporting mechanisms like the one that enabled her to continue her professional rise. But Darden's success in the workplace was also, presumably, the result of many unnamed colleagues and friends who may or may not have considered themselves feminists. These were the people who provided her with community and support—and likely a not insignificant number of casserole dinners—as she ascended the government ranks. These types of collective efforts have been made increasingly legible, in turn, because of the feminist scholars and activists whose decades of work have enabled us to recognize that labor—emotional as much as physical—as such today.

As should already be apparent, feminism has been defined and used in many ways. Here and throughout the book, we employ the term *feminism* as a shorthand for the diverse and wide-ranging projects that name and challenge sexism and other forces of oppression, as well as those which seek to create more just, equitable, and livable futures. Because of this broadness, some scholars prefer to use the term *feminisms*,

which clearly signals the range of—and, at times, the incompatibilities among—these various strains of feminist activism and political thought. For reasons of readability, we choose to use the term *feminism* here, but our feminism is intended to be just as expansive. It includes the work of regular folks like Darden and Champine, public intellectuals like Betty Friedan and bell hooks, and organizing groups like the Combahee River Collective, which have taken direct action to achieve the equality of the sexes. It also includes the work of scholars and other cultural critics—like Kimberlé Crenshaw and Margot Lee Shetterly, among many more—who have used writing to explore the social, political, historical, and conceptual reasons behind the inequality of the sexes that we face today.

In the process, these writers and activists have given voice to the many ways in which today's status quo is unjust.¹² These injustices are often the result of historical and contemporary differentials of power, including those among men, women, and nonbinary people, as well as those among white women and Black women, academic researchers and Indigenous communities, and people in the Global North and the Global South. Feminists analyze these power differentials so that they can change them. Such a broad focus—one that incorporates race, class, ability, and more—would have sounded strange to Friedan or to the white women largely credited for leading the fight for women's suffrage in the nineteenth century.¹³ But the reality is that women of color have long insisted that any movement for gender equality must also consider the ways in which privilege and oppression are intersectional.

Because the concept of intersectionality is essential for this whole book, let's get a bit more specific. The term was coined by legal theorist Kimberlé Crenshaw in the late 1980s.¹⁴ In law school, Crenshaw had come across the antidiscrimination case of *DeGraffenreid v. General Motors*. Emma DeGraffenreid was a Black working mother who had sought a job at a General Motors factory in her town. She was not hired and sued GM for discrimination. The factory did have a history of hiring Black people: many Black men worked in industrial and maintenance jobs there. They also had a history of hiring women: many white women worked there as secretaries. These two pieces of evidence provided the rationale for the judge to throw out the case. Because the company did hire Black people and did hire women, it could not be discriminating based on race or gender. But, Crenshaw wanted to know, what about discrimination on the basis of race *and* gender together? This was something different, it was real, and it needed to be named. Crenshaw not only named the concept, but would go on to explain and elaborate the idea of intersectionality in award-winning books, papers, and talks.¹⁵

Key to the idea of *intersectionality* is that it does not only describe the intersecting aspects of any particular person's identity (or *positionalities*, as they are sometimes

termed).¹⁶ It also describes the intersecting forces of privilege and oppression at work in a given society. *Oppression* involves the systematic mistreatment of certain groups of people by other groups. It happens when power is not distributed equally—when one group controls the institutions of law, education, and culture, and uses its power to systematically exclude other groups while giving its own group unfair advantages (or simply maintaining the status quo).¹⁷ In the case of gender oppression, we can point to the sexism, cissexism, and patriarchy that is evident in everything from political representation to the wage gap to who speaks more often (or more loudly) in a meeting.¹⁸ In the case of racial oppression, this takes the form of racism and white supremacy. Other forms of oppression include ableism, colonialism, and classism. Each has its particular history and manifests differently in different cultures and contexts, but all involve a dominant group that accrues power and privilege at the expense of others. Moreover, these forces of power and privilege on the one hand and oppression on the other mesh together in ways that multiply their effects.

The effects of privilege and oppression are not distributed evenly across all individuals and groups, however. For some, they become an obvious and unavoidable part of daily life, particularly for women and people of color and queer people and immigrants: the list goes on. If you are a member of any or all of these (or other) minoritized groups, you experience their effects everywhere, shaping the choices you make (or don't get to make) each day. These systems of power are as real as rain. But forces of oppression can be difficult to detect when you benefit from them (we call this a *privilege hazard* later in the book). And this is where data come in: it was a set of intersecting systems of power and privilege that Darden was intent on exposing when she posed her initial question to her division chief. And it was that same set of intersecting systems of power and privilege that Darden sought to challenge when she approached Champine. Darden herself didn't need any more evidence of the problem she faced; she was already living it every day.¹⁹ But when her experience was recorded as data and aggregated with others' experiences, it could be used to challenge institutional systems of power and have far broader impact than on her career trajectory alone.

In this way, Darden models what we call *data feminism*: a way of thinking about data, both their uses and their limits, that is informed by direct experience, by a commitment to action, and by intersectional feminist thought. The starting point for data feminism is something that goes mostly unacknowledged in data science: power is not distributed equally in the world. Those who wield power are disproportionately elite, straight, white, able-bodied, cisgender men from the Global North.²⁰ The work of data feminism is first to tune into how standard practices in data science serve to reinforce these existing inequalities and second to use data science to challenge and change the

distribution of power.²¹ Underlying data feminism is a belief in and commitment to *co-liberation*: the idea that oppressive systems of power harm all of us, that they undermine the quality and validity of our work, and that they hinder us from creating true and lasting social impact with data science.

We wrote this book because we are data scientists and data feminists. Although we speak as a “we” in this book, and share certain identities, experiences, and skills, we have distinct life trajectories and motivations for our work on this project. If we were sitting with you right now, we would each introduce ourselves by answering the question: What brings you here today? Placing ourselves in that scenario, here is what we would have to say.

Catherine: I am a hacker mama. I spent fifteen years as a freelance software developer and experimental artist, now professor, working on projects ranging from serendipitous news-recommendation systems to countercartography to civic data literacy to making breast pumps not suck. I’m here writing this book because, for one, the hype around big data and AI is deafeningly male and white and technoheroic and the time is now to reframe that world with a feminist lens. The second reason I’m here is that my recent experience running a large, equity-focused hackathon taught me just how much people like me—basically, well-meaning liberal white people—are part of the problem in struggling for social justice. This book is one attempt to expose such workings of power, which are inside us as much as outside in the world.²²

Lauren: I often describe myself as a professional nerd. I worked in software development before going to grad school to study English, with a particular focus on early American literature and culture. (Early means very early—like, the eighteenth century.) As a professor at an engineering school, I now work on research projects that translate this history into contemporary contexts. For instance, I’m writing a book about the history of data visualization, employing machine-learning techniques to analyze abolitionist newspapers, and designing a haptic recreation of a hundred-year-old visualization scheme that looks like a quilt. Through projects like these, I show how the rise of the concept of “data” (which, as it turns out, really took off in the eighteenth century) is closely connected to the rise of our current concepts of gender and race. So one of my reasons for writing this book is to show how the issues of racism and sexism that we see in data science today are by no means new. The other reason is to help translate humanistic thinking into practice and, in so doing, create more opportunities for humanities scholars to engage with activists, organizers, and communities.²³

We both strongly believe that data can do good in the world. But for it to do so, we must explicitly acknowledge that a key way that power and privilege operate in the world today has to do with the word *data* itself. The word dates to the mid-seventeenth century, when it was introduced to supplement existing terms such as *evidence* and *fact*. Identifying information as data, rather than as either of those other two terms, served a rhetorical purpose.²⁴ It converted otherwise debatable information into the solid basis for subsequent claims. But what information needs to become data before it can be trusted? Or, more precisely, whose information needs to become data before it can be considered as fact and acted upon?²⁵ Data feminism must answer these questions, too.

The story that begins with Christine Darden entering the gates of Langley, passes through her sustained efforts to confront the structural oppression she encountered there, and concludes with her impressive array of life achievements, is a story about the power of data. Throughout her career, in ways large and small, Darden used data to make arguments and transform lives. But that's not all. Darden's feel-good biography is just as much a story about the larger systems of power that required data—rather than the belief in her lived experience—to perform that transformative work. An institutional mistrust of Darden's experiential knowledge was almost certainly a factor in Champine's decision to create her bar chart. Champine likely recognized, as did Darden herself, that she would need the bar chart to be believed.

In this way, the alliance between Darden and Champine, and their work together, underscores the flaws and compromises that are inherent in any data-driven project. The process of converting life experience into data always necessarily entails a reduction of that experience—along with the historical and conceptual burdens of the term. That Darden and Champine were able to view their work as a success despite these inherent constraints underscores even more the importance of listening to and learning from people whose lives and voices are behind the numbers. No dataset or analysis or visualization or model or algorithm is the result of one person working alone. Data feminism can help to remind us that before there are data, there are people—people who offer up their experience to be counted and analyzed, people who perform that counting and analysis, people who visualize the data and promote the findings of any particular project, and people who use the product in the end. There are also, always, people who go uncounted—for better or for worse. And there are problems that cannot be represented—or addressed—by data alone. And so data feminism, like justice, must remain both a goal and a process, one that guides our thoughts and our actions as we move forward toward our goal of remaking the world.

Data and Power

It took five state-of-the-art IBM System/360 Model 75 machines to guide the *Apollo 11* astronauts to the moon. Each was the size of a car and cost \$3.5 million dollars. Fast forward to the present. We now have computers in the form of phones that fit in our pockets and—in the case of the 2019 Apple iPhone XR—can perform more than 140 million more instructions per second than a standard IBM System/360.²⁶ That rate of change is astounding; it represents an exponential growth in computing capacity (figure 0.2a). We've witnessed an equally exponential growth in our ability to collect and record information in digital form—and in the ability to have information collected about us (figure 0.2b).

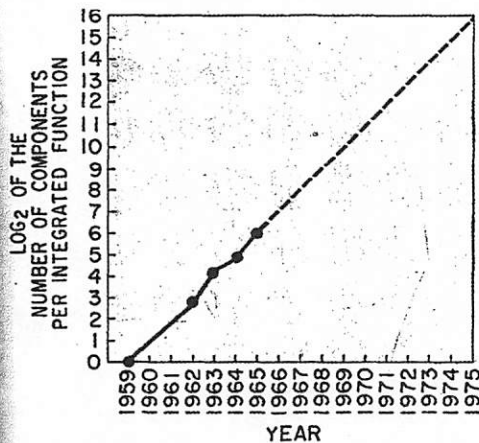


Figure 0.2

(a) The time-series chart included in the original paper on Moore's law, published in 1965, which posited that the number of transistors that could fit on an integrated circuit (and therefore contribute to computing capacity) would double every year. Courtesy of Gordon Moore. (b) Several years ago, researchers concluded that transistors were approaching their smallest size and that Moore's law would not hold. Nevertheless, today's computing power is what enabled Dr. Katie Bouman, a postdoctoral fellow at MIT, to contribute to a project that involved processing and compositing approximately five petabytes of data captured by the Event Horizon Telescope to create the first ever image of a black hole. After the publication of this photo in April 2019 showing her excitement—as one of the scientists on the large team that worked for years to capture the image—Bouman was subsequently trolled and harassed online. Courtesy of Tamy Emma Pepin/Twitter.

But the act of collecting and recording data about people is not new at all. From the registers of the dead that were published by church officials in the early modern era to the counts of Indigenous populations that appeared in colonial accounts of the Americas, data collection has long been employed as a technique of consolidating knowledge about the people whose data are collected, and therefore consolidating power over their lives.²⁷ The close relationship between data and power is perhaps most clearly visible in the historical arc that begins with the logs of people captured and placed aboard slave ships, reducing richly lived lives to numbers and names. It passes through the eugenics movement, in the late nineteenth and early twentieth centuries, which sought to employ data to quantify the superiority of white people over all others. It continues today in the proliferation of biometrics technologies that, as sociologist Simone Browne has shown, are disproportionately deployed to surveil Black bodies.²⁸

When Edward Snowden, the former US National Security Agency contractor, leaked his cache of classified documents to the press in 2013, he revealed the degree to which the federal government routinely collects data on its citizens—often with minimal regard to legality or ethics.²⁹ At the municipal level, too, governments are starting to collect data on everything from traffic movement to facial expressions in the interests of making cities “smarter.”³⁰ This often translates to reinscribing traditional urban patterns of power such as segregation, the overpolicing of communities of color, and the rationing of ever-scarcer city services.³¹

But the government is not alone in these data-collection efforts; corporations do it too—with profit as their guide. The words and phrases we search for on Google, the times of day we are most active on Facebook, and the number of items we add to our Amazon carts are all tracked and stored as data—data that are then converted into corporate financial gain. The most trivial of everyday actions—searching for a way around traffic, liking a friend’s cat video, or even stepping out of our front doors in the morning—are now hot commodities. This is not because any of these actions are exceptionally interesting (although we do make an exception for Catherine’s cats) but because these tiny actions can be combined with other tiny actions to generate targeted advertisements and personalized recommendations—in other words, to give us more things to click on, like, or buy.³²

This is the data economy, and corporations, often aided by academic researchers, are currently scrambling to see what behaviors—both online and off—remain to be turned into data and then monetized. Nothing is outside of *datafication*, as this process is sometimes termed—not your search history, or Catherine’s cats, or the butt that Lauren is currently using to sit in her seat. To wit: Shigeomi Koshimizu, a Tokyo-based professor of engineering, has been designing matrices of sensors that collect data at 360

different positions around a rear end while it is comfortably ensconced in a chair.³³ He proposes that people have unique butt signatures, as unique as their fingerprints. In the future, he suggests, our cars could be outfitted with butt-scanners instead of keys or car alarms to identify the driver.

Although datafication may occasionally verge into the realm of the absurd, it remains a very serious issue. Decisions of civic, economic, and individual importance are already and increasingly being made by automated systems sifting through large amounts of data. For example, PredPol, a so-called predictive policing company founded in 2012 by an anthropology professor at the University of California, Los Angeles, has been employed by the City of Los Angeles for nearly a decade to determine which neighborhoods to patrol more heavily, and which neighborhoods to (mostly) ignore. But because PredPol is based on historical crime data and US policing practices have always disproportionately surveilled and patrolled neighborhoods of color, the predictions of where crime will happen in the future look a lot like the racist practices of the past.³⁴ These systems create what mathematician and writer Cathy O'Neil, in *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, calls a "pernicious feedback loop," amplifying the effects of racial bias and of the criminalization of poverty that are already endemic to the United States.

O'Neil's solution is to open up the computational systems that produce these racist results. Only by knowing what goes in, she argues, can we understand what comes out. This is a key step in the project of mitigating the effects of biased data. Data feminism additionally requires that we trace those biased data back to their source. PredPol and the "three most objective data points" that it employs certainly amplify existing biases, but they are not the root cause.³⁵ The cause, rather, is the long history of the criminalization of Blackness in the United States, which produces biased policing practices, which produce biased historical data, which are then used to develop risk models for the future.³⁶ Tracing these links to historical and ongoing forces of oppression can help us answer the ethical question, Should this system exist?³⁷ In the case of PredPol, the answer is a resounding no.

Understanding this long and complicated chain reaction is what has motivated Yeshimabeit Milner, along with Boston-based activists, organizers, and mathematicians, to found Data for Black Lives, an organization dedicated to "using data science to create concrete and measurable change in the lives of Black communities."³⁸ Groups like the Stop LAPD Spying coalition are using explicitly feminist and antiracist methods to quantify and challenge invasive data collection by law enforcement.³⁹ Data journalists are reverse-engineering algorithms and collecting qualitative data at scale

about maternal harm.⁴⁰ Artists are inviting participants to perform ecological maps and using AI for making intergenerational family memoirs (figure 0.3a).⁴¹

All these projects are *data science*. Many people think of data as numbers alone, but data can also consist of words or stories, colors or sounds, or any type of information that is systematically collected, organized, and analyzed (figures 0.3b, 0.3c).⁴² The *science* in data science simply implies a commitment to systematic methods of observation and experiment. Throughout this book, we deliberately place diverse data science examples alongside each other. They come from individuals and small groups, and from across academic, artistic, nonprofit, journalistic, community-based, and for-profit organizations. This is due to our belief in a capacious definition of data science, one that seeks to include rather than exclude and does not erect barriers based on formal credentials, professional affiliation, size of data, complexity of technical methods, or other external markers of expertise. Such markers, after all, have long been used to prevent women from fully engaging in any number of professional fields, even as those fields—which include data science and computer science, among many others—were largely built on the knowledge that women were required to teach themselves.⁴³ An attempt to push back against this gendered history is foundational to data feminism, too.

Throughout its own history, feminism has consistently had to work to convince the world that it is relevant to people of all genders. We make the same argument: that data feminism is for everybody. (And here we borrow a line from bell hooks.)⁴⁴ You will notice that the examples we use are not only about women, nor are they created only by women. That's because *data feminism isn't only about women*. It takes more than one gender to have gender inequality and more than one gender to work toward justice. Likewise, *data feminism isn't only for women*. Men, nonbinary, and genderqueer people are proud to call themselves feminists and use feminist thought in their work. Moreover, *data feminism isn't only about gender*. Intersectional feminists have keyed us into how race, class, sexuality, ability, age, religion, geography, and more are factors that together influence each person's experience and opportunities in the world. Finally, *data feminism is about power—about who has it and who doesn't*. Intersectional feminism examines unequal power. And in our contemporary world, data is power too. Because the power of data is wielded unjustly, it must be challenged and changed.

Data Feminism in Action

Data is a double-edged sword. In a very real sense, data have been used as a weapon by those in power to consolidate their control—over places and things, as well as

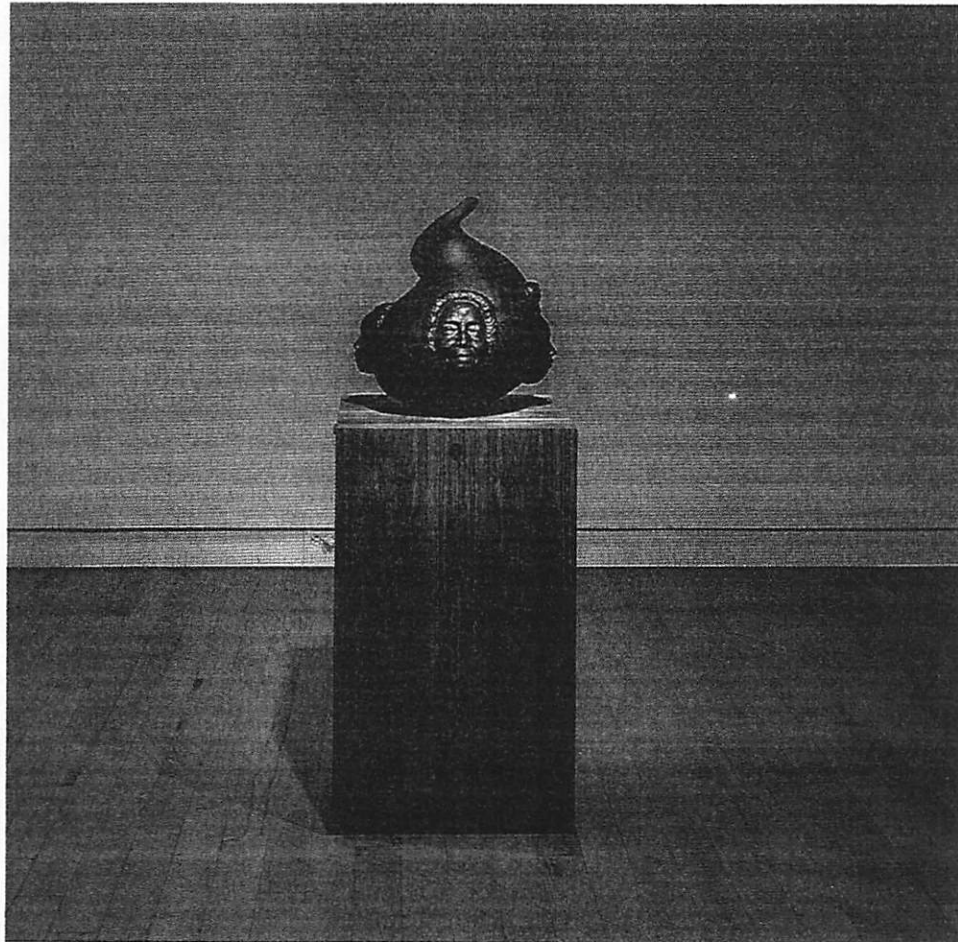


Figure 0.3

We define data science expansively in this book—here are three examples. (a) *Not the Only One* by Stephanie Dinkins (2017), is a sculpture that features a Black family through the use of artificial intelligence. The AI is trained and taught by the underrepresented voices of Black and brown individuals in the tech sector. (b) Researcher Margaret Mitchell and colleagues, in “Seeing through the Human Reporting Bias” (2016), have worked on systems to infer what is *not said* in human speech for the purposes of image classification. For example, people say “green bananas” but not “yellow bananas” because yellow is implied as the default color of the banana. Similarly, people say “woman doctor” but do not say “man doctor,” so it is the words that are *not spoken* that encode the bias. (c) A gender analysis of Hollywood film dialogue, “Film Dialogue from 2,000 Screenplays Broken Down by Gender and Age,” by Hanah Anderson and Matt Daniels, created for The Pudding, a data journalism start-up (2017).

(a) A woman standing next to a bicycle with basket.



	Human Label	Visual Label
Bicycle	✓	✓

(b) A city street filled with lots of people walking in the rain.



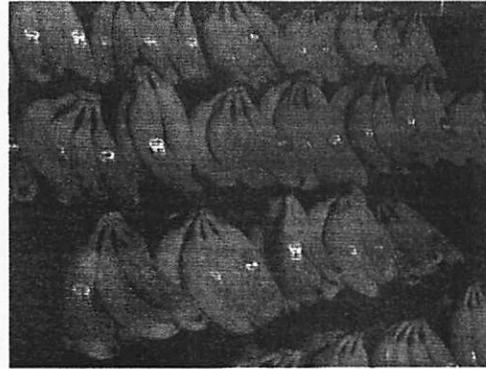
	Human Label	Visual Label
Bicycle	✗	✓

(c) A yellow Vespa parked in a lot with other cars.



	Human Label	Visual Label
Yellow	✓	✓

(d) A store display that has a lot of bananas on sale.



	Human Label	Visual Label
Yellow	✗	✓

Figure 0.3 (continued)

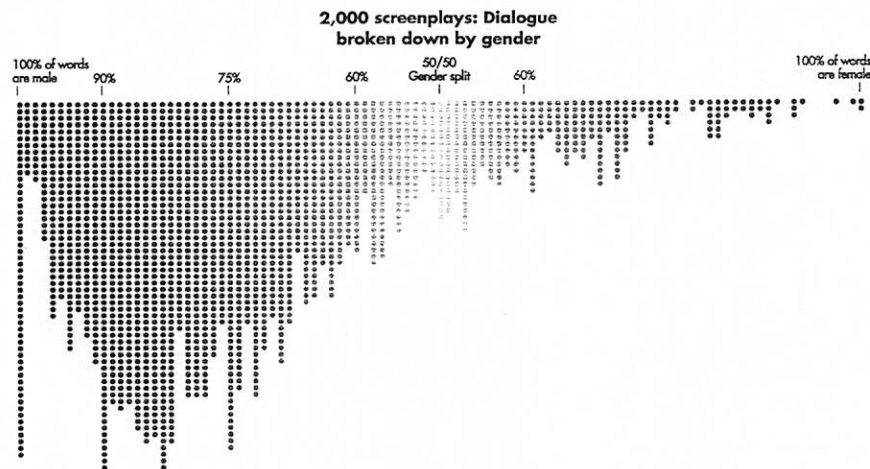


Figure 0.3 (continued)

people. Indeed, a central goal of this book is to show how governments and corporations have long employed data and statistics as management techniques to preserve an unequal status quo. Working with data from a feminist perspective requires knowing and acknowledging this history. To frame the trouble with data in another way: it's not a coincidence that the institution that employed Christine Darden and enabled her professional rise is the same that wielded the results of her data analysis to assert the technological superiority of the United States over its communist adversaries and to plant an American flag on the moon. But this flawed history does not mean ceding control of the future to the powers of the past. Data are part of the problem, to be sure. But they are also part of the solution. Another central goal of this book is to show how the power of data can be wielded back.

To guide us in this work, we have developed seven core principles. Individually and together, these principles emerge from the foundation of intersectional feminist thought. Each of the following chapters is structured around a single principle. The seven principles of data feminism are as follows:

1. **Examine power.** Data feminism begins by analyzing how power operates in the world.
2. **Challenge power.** Data feminism commits to challenging unequal power structures and working toward justice.

3. **Elevate emotion and embodiment.** Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.
4. **Rethink binaries and hierarchies.** Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.
5. **Embrace pluralism.** Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.
6. **Consider context.** Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.
7. **Make labor visible.** The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that it can be recognized and valued.

Each of the following chapters takes up one of these principles, drawing upon examples from the field of data science, expansively defined, to show how that principle can be put into action. Along the way, we introduce key feminist concepts like the matrix of domination (Patricia Hill Collins; see chapter 1), situated knowledge (Donna Haraway; see chapter 3), and emotional labor (Arlie Hochschild; see chapter 8), as well as some of our own ideas about what data feminism looks like in theory and practice. To this end, we introduce you to people at the cutting edge of data and justice. These include engineers and software developers, activists and community organizers, data journalists, artists, and scholars. This range of people, and the range of projects they have helped to create, is our way of answering the question: What makes a project feminist? As will become clear, a project may be feminist in *content*, in that it challenges power by choice of subject matter; in *form*, in that it challenges power by shifting the aesthetic and/or sensory registers of data communication; and/or in *process*, in that it challenges power by building participatory, inclusive processes of knowledge production. What unites this broad scope of data-based work is a commitment to action and a desire to remake the world.

Our overarching goal is to take a stand against the status quo—against a world that benefits us, two white college professors, at the expense of others. To work toward this goal, we have chosen to feature the voices of those who speak from the margins, whether because of their gender, sexuality, race, ability, class, geographic location, or any combination of those (and other) subject positions. We have done so, moreover,

because of our belief that those with direct experience of inequality know better than we do about what actions to take next. For this reason, we have attempted to prioritize the work of people in closer proximity to issues of inequality over those who study inequality from a distance. In this book, we pay particular attention to inequalities at the intersection of gender and race. This reflects our location in the United States, where the most entrenched issues of inequality have racism at their source. Our values statement, included as an appendix to this book, discusses the rationale for these authorial choices in more detail.

Any book involves making choices about whose voices and whose work to include and whose voices and work to omit. We ask that those who find their perspectives insufficiently addressed or their work insufficiently acknowledged view these gaps as additional openings for conversation. Our sincere hope is to contribute in a small way to a much larger conversation, one that began long before we embarked upon this writing process and that will continue long after these pages are through.

This book is intended to provide concrete steps to action for data scientists seeking to learn how feminism can help them work toward justice, and for feminists seeking to learn how their own work can carry over to the growing field of data science. It is also addressed to professionals in all fields in which data-driven decisions are being made, as well as to communities that want to resist or mobilize the data that surrounds them. It is written for everyone who seeks to better understand the charts and statistics that they encounter in their day-to-day lives, and for everyone who seeks to communicate the significance of such charts and statistics to others.

Our claim, once again, is that data feminism is for everyone. It's for people of all genders. It's by people of all genders. And most importantly: it's about much more than gender. Data feminism is about power, about who has it and who doesn't, and about how those differentials of power can be challenged and changed using data. We invite you, the readers of this book, to join us on this journey toward justice and toward remaking our data-driven world.

5 Unicorns, Janitors, Ninjas, Wizards, and Rock Stars

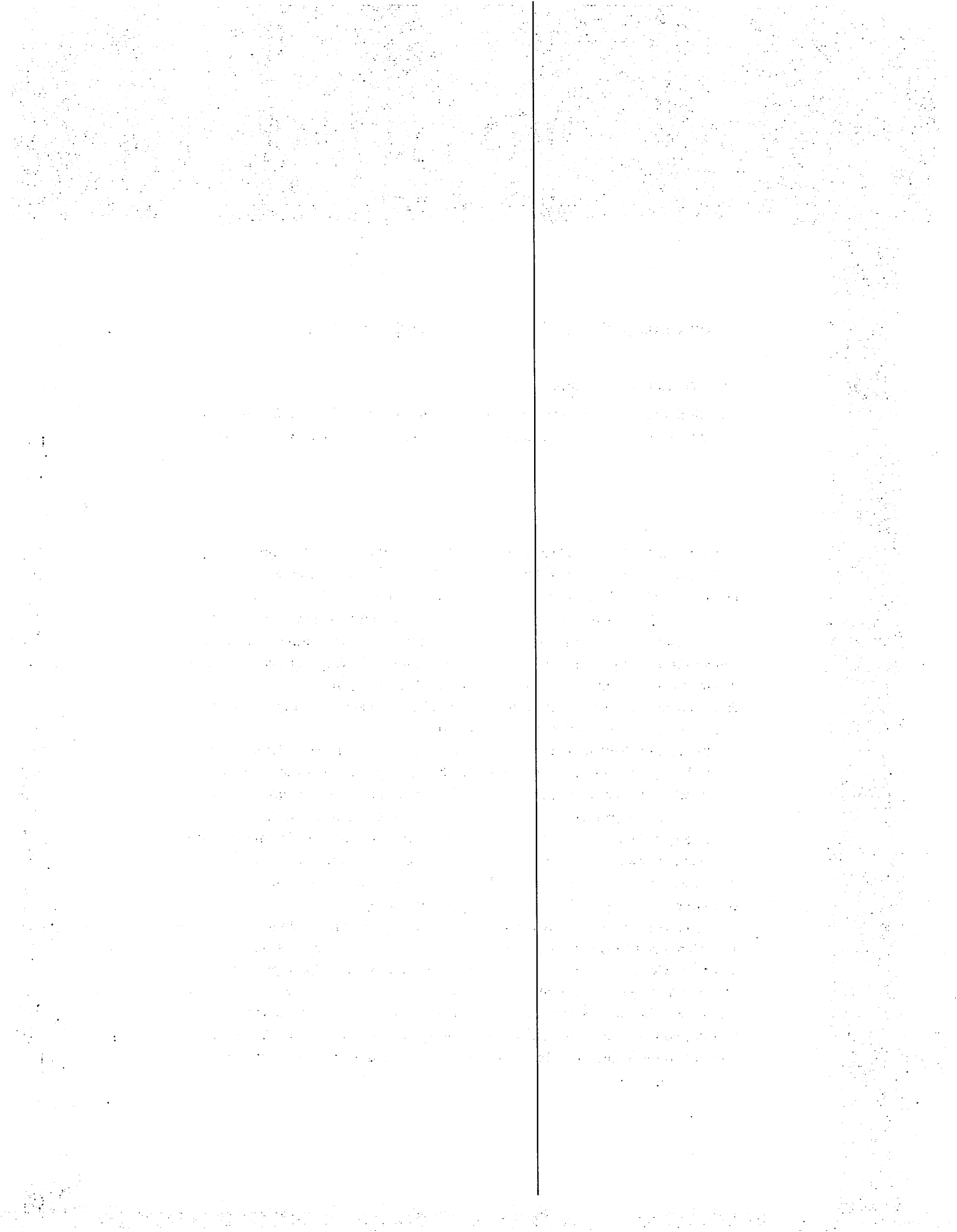
Principle: Embrace Pluralism

Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.

In Spring 2017, *Bloomberg News* ran an article with the provocative title “America’s Rich Get Richer and the Poor Get Replaced by Robots.”¹ Using census data, the authors reported that income inequality is widening across the nation. San Francisco is leading the pack, with an income gap of almost half a million dollars between the richest and the poorest twenty percent of residents. As in other places, the wealth gap has race and gender dimensions. In the Bay Area, people of color earn sixty-eight cents for every dollar earned by white people, and 59 percent of single mothers live in poverty.² San Francisco also has the lowest proportion of children to adults in any major US city, and—since 2003—an escalating rate of evictions.

Although the San Francisco Rent Board collects data on these evictions, it does not track where people go after they are evicted, how many of those people end up homeless, or which landlords are responsible for systematically evicting major blocks of the city. In 2013, the Anti-Eviction Mapping Project (AEMP) stepped in. The initiative is a self-described collective of “housing justice activists, researchers, data nerds, artists, and oral historians.” It is a multiracial group with significant, though not exclusive, project leadership by women. The AEMP is mapping eviction, and doing so through a collaborative, multimodal, and—yes—quite messy process.

If you visit antievictionmap.com, you won’t actually find a single map. There are seventy-eight distinct maps linked from the homepage: maps of displaced residents, of evictions, of tech buses, of property owners, of the Filipino diaspora, of the declining numbers of Black residents in the city, and more. The AEMP has a distinct way of working that is grounded in its stated commitment to antiracist, feminist, and decolonial methodologies.³ Most of the projects happen in collaboration with nonprofits and community-based organizations. Additional projects originate from



within the collective. For example, the group is working on producing an atlas of the Bay Area called *Counterpoints: Bay Area Data and Stories for Resisting Displacement*, which will cover topics such as migration and relocation, gentrification and the prison pipeline, Indigenous and colonial histories of the region, and speculation about the future.⁴

One of the AEMP's longest-standing collaborations is with the Eviction Defense Collaborative (EDC), a nonprofit that provides court representation for people who have been evicted. Although the city does not collect data on the race or income of evictees, the EDC does collect those demographics, and it works with 90 percent of the tenants whose eviction cases end up in San Francisco courts.⁵ In 2014, the EDC approached the AEMP to help produce its annual report and in return offered to share its demographic data with the organization.⁶ Since then, the two groups have continued to work together on EDC annual reports, as well as additional analyses of evictions with a focus on race. The AEMP has also gone on to produce reports with tenants' rights organizations, timelines of gentrification with Indigenous students, oral histories with grants from anthropology departments, and murals with arts organizations, as well as maps and more maps.

Some of the AEMP's maps are designed to leverage the ability of data visualization to make patterns visible at a glance. For example, the *Tech Bus Stop Eviction Map* produced in 2014 plots the locations of three years of Ellis Act evictions.⁷ This is a form of "no-fault" eviction in which landlords can claim that they are going out of the rental business. In many cases, this is so that they can convert the building to a condominium and sell the units at significant profit. San Francisco has seen almost five thousand invocations of the Ellis Act since 1994. On the map (figure 5.1), the AEMP plotted Ellis Act evictions in relationship to the location of technology company bus stops. Starting in the 2000s, tech companies with campuses in Silicon Valley began offering private luxury buses as a perk to attract employees who wanted to live in downtown San Francisco but didn't want the hassle of commuting. Colloquially known as "the Google buses" because Google was the most visible company to implement the practice, these vehicles use public bus stops—illegally at first—to shuttle employees to their offices in comfort (and also away from the public transportation system, which would otherwise reap the benefits of their fares).⁸ Because a new, wealthy clientele began to seek condos near the bus stops, property values soared—and so did the rate of evictions of long-time neighborhood residents. The AEMP analysis showed that, between 2011 and 2013, 69 percent of no-fault evictions occurred within four blocks of a tech bus stop. The map makes that finding plain.

Evictions Near Shuttle Stops 2011–2013

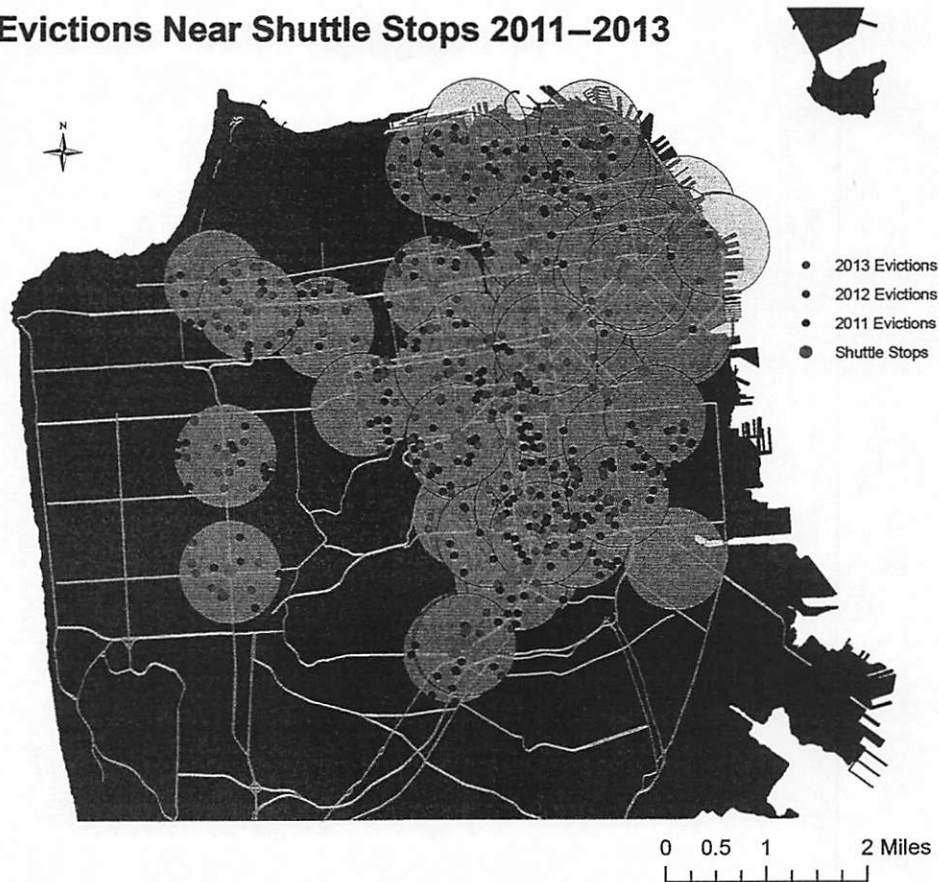


Figure 5.1

Tech Bus Stop Eviction Map by the Anti-Eviction Mapping Project (2014) plots evictions with "Google bus stops" in San Francisco. The group's analysis showed that 69 percent of no-fault evictions in the city occurred within four blocks of a tech bus stop. Courtesy of the Anti-Eviction Mapping Project.

But other AEMP maps are intentionally designed *not* to depict a clear correlation between evictions and place. In *Narratives of Displacement and Resistance* (figure 5.2a), five thousand evictions are each represented as a differently sized red bubble, so the base map of San Francisco is barely visible underneath.⁹ On top of this red sea, sky-blue bubbles dot the locations where the AEMP has conducted interviews with displaced residents, as well as activists, mediamakers, and local historians. Clicking on a blue bubble sets one of the dozens of oral histories in motion—for instance, the story of Phyllis Bowie (figure 5.2b), a resident facing eviction from her one-bedroom apartment. “I was born and raised in San Francisco proudly,” she begins. Bowie goes on to recall how she returned from the Air Force and worked like crazy for two years at her own small business, building up an income record that would make her eligible for a lease-to-own apartment in Midtown, the historically Black neighborhood where she had grown up. In 2015, however, the city broke the master lease and took away the rent control on her building. Now tenants like Bowie, who moved there with the promise of homeownership, are facing skyrocketing rents that none of them can afford. Bowie is leading rent strikes and organizing the building’s tenants, but their future is uncertain.

This uncertainty is carried over into the design of the map, which uses a phenomenon that is often discouraged in data visualization—*occlusion*—to drive home its point. Occlusion typically refers to the “problem” that occurs when some marks (like the eviction dots) obscure other important features (like the whole geography of the city). But here it underscores the point that there are very few patterns to detect when the entire city is covered in big red eviction bubbles and abundant blue story dots. Put another way, the whole city is a pattern, and that pattern is the problem—much more than a problem of information design.

In this way, the *Narratives* map enacts dissent from San Francisco city policies in the same way that it enacts dissent from the conventions of information design. It refuses both the clarity and cleanliness associated with the best practices of data visualization and the homogenizing and “cleanliness” associated with the forces of gentrification that lead to evictions in the first place.¹⁰ The visual point of the map is simple and exhortative: there are too many evictions. And there are too many eviction stories. The map does not efficiently reveal how evictions data may be correlated with Bay Area Rapid Transit (BART) stops, income, Google bus stops, or any other potential dimensions of the data. Even finding the *Narratives* map is difficult, given the sheer number of maps and visualizations on the AEMP website. There is no “master dashboard” that integrates all the information that the AEMP has collected into a single interface.¹¹

But all these design choices reinforce the main purpose of the AEMP: to document the effects of displacement and to resist it through critical and creative means.

The AEMP thus offers a rich set of examples of feminist counterdata collection and countervisualization strategies. Individually and together, the maps illustrate how Katie Lloyd Thomas, founding member of the feminist art architecture collective *taking place*, envisions “partiality and contestation” taking place in graphical design. “Rather than tell one ‘true’ story of consensus, [the drawing/graphic] might remember and acknowledge multiple, even contradictory versions of reality,” she explains.¹² The *Narratives* map populates its landscape with these contradictory realities. It demonstrates that behind each eviction is a person—a person like Bowie, with a unique voice and a unique story. The voices we hear are diverse, multiple, specific, divergent—and deliberately so.

In so doing, the *Narratives* map, and the AEMP more generally, exemplify the fourth principle of data feminism and the theme of this chapter: *embrace pluralism*. Embracing pluralism in data science means valuing many perspectives and voices and doing so at all stages of the process—from collection to cleaning to analysis to communication. It also means attending to the ways in which data science methods can inadvertently work to suppress those voices in the service of clarity, cleanliness, and control. Many of our received ideas about data science work against pluralistic meaning-making processes, and one goal of data feminism is to change that.

Strangers in the Dataset

Data cleaning holds a special kind of mythos in data science. “It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data,” writes Hadley Wickham in the first sentence of the abstract of “Tidy Data,” his widely cited paper from 2014.¹³ Wickham is also the author of the *tidyr* package for R, a popular programming language and statistical computing platform. (*Packages* are prewritten collections of functions, and other forms of code, that can be installed and used in any project.) Wickham’s package shares the sentiment expressed in his paper: that data are inherently messy and need to be tidied and tamed.

Wickham is not alone in this belief. Since the release of his *tidyr* package, Wickham has gone on to create the “tidyverse,” an expanded set of packages that formalize a clear workflow for processing data, which have been developed by a team of equally enthusiastic contributors. Articles in the popular and business press corroborate this insistence on tidiness, as well as its pressing need. In *Harvard Business Review*, the work of the data scientist is glorified in terms of this tidying function: “At ease in the digital

realm, they are able to bring structure to large quantities of formless data and make analysis possible."¹⁴ Here, the intrepid analyst wrangles orderly tables from unstructured chaos. According to the article, it's "the sexiest job of the 21st century."¹⁵ But for the *New York Times*, the data analyst's work is far less attractive. A 2014 article equated the task of cleaning data to the low-wage maintenance work of "janitors."¹⁶

Whether or not you think data scientists are sexy (they are), and whether or not you think janitors should be offended by this classist reference (we all should be), certain assumptions and anxieties remain consistent across these different articulations about the need for tidiness, cleanliness, and order, and the qualities of the people who should be doing this work. They must be able to tame the chaos of information overload. They must "scrub" and "cleanse" dirty data. And they must undertake deliberate action—either extraordinary or mundane—to put data back in their proper place.

But what might be lost in the process of dominating and disciplining data? Whose perspectives might be lost in that process? And, conversely, whose perspectives might be additionally imposed? The ideas expressed by Wickham, and by the press, carry the assumption that all data scientists, in all contexts, value cleanliness and control over messiness and complexity. But as the example of the AEMP demonstrates, these are not the requirements, nor the goals, of all data projects.

Before moving forward, we find it important to acknowledge that these ideas about cleanliness and control contain troubling traces of a movement from a prior era: eugenics, the upsetting, nineteenth-century source of much of modern statistics. As we have referenced in previous chapters via the work of Dean Spade and Rori Rohlf, many of the men often cited as the earliest statistical luminaries, such as Karl Pearson, Adolphe Quetelet, Francis Galton, and Ronald Fisher, were also leaders in the eugenics movement.¹⁷ In her book *Ghost Stories for Darwin: The Science of Variation and the Politics of Diversity*, postcolonial science studies scholar Banu Subramaniam details how over the course of the late nineteenth and early twentieth centuries, as statistics became the preferred language of communication between biologists and social scientists, certain ideas from the eugenics movement also carried over into this broader scientific conversation.¹⁸ While the most odious aspects of these ideas have been largely (and thankfully) stripped away, certain core principles—like a generalized belief in the benefit of control and cleanliness—remain.¹⁹ To be clear: the point here is *not* that anyone who cleans their data is perpetuating eugenics.²⁰ The point, rather, is that the ideas underlying the belief that data should always be clean and controlled have tainted historical roots. As data scientists, we cannot forget these roots, even as the ideas themselves have been tidied up over time.

This is the long history that library and information studies scholars Katie Rawson and Trevor Muñoz likely allude to in their assertion that “the cleaning paradigm assumes an underlying, ‘correct’ order.” Like Subramaniam, but in the context of cleaning data more specifically, Rawson and Muñoz caution that cleaning can function as a “diversity-hiding trick.”²¹ In the perceived messiness of data, there is actually rich information about the circumstances under which it was collected. Data studies scholar Yanni Loukissas concurs. Rather than talking about *datasets*, he advocates that we talk about *data settings*—his term to describe both the technical and the human processes that affect what information is captured in the data collection process and how the data are then structured.²²

As an example of how the data setting matters, Loukissas tells the story of being at a hackathon in Cambridge, Massachusetts, where he began to explore a dataset from the Clemson University Library, located in South Carolina. He stumbled across a puzzling record in which the librarian had noted the location of the item as “upstate.” Such a designation is, of course, relational to the place of collection. For South Carolinians, *upstate* is a very legible term that refers to the westernmost region of the state, where Clemson is located. But it does not hold the same meaning for a person from New York, where *upstate* refers to its own northern region, nor does it hold the same meaning for a person sitting at a hackathon in Massachusetts, which does not have an *upstate* part of the state. Had someone at the hackathon written the entry from where they sat, they might have chosen to list the ten or so counties that South Carolinians recognize as *upstate*, so as to be more clearly legible to a wider geographic audience. But there is meaning conveyed by the term that would not be conveyed by other, more general ways of indicating the same region. Only somebody already located in South Carolina would have referred to that region in that way. From that usage of the term, we can reason that the data were originally collected in South Carolina. This information is not included elsewhere in the library record.

It is because of records like this one that the process of cleaning and tidying data can be so complicated and, at times, can be a destructive rather than constructive act. One way to think of it is like chopping off the roots of a tree that connects it to the ground from which it grew. It’s an act that irreversibly separates the data from their context.

We might relate the growth of tools like *tidyr* that help to trim and tidy data to another human intervention into the environment: the proliferation of street names and signs that transformed the landscape of the nineteenth-century United States. Geographer Reuben Rose-Redwood describes how, for example, prior to the Revolutionary War, very few streets in Manhattan had signs posted at intersections.²³ Street names, such as they existed, were vernacular and related to the particularity of a

spot—for example, “Take a right at the red house.” But with the increased mobility of people and things—think of the postal system, the railroads, or the telegraph—street names needed to become systematized. Rose-Redwood calls this the production of “legible urban spaces.” Then, as now, there is high economic value to legible urban spaces, particularly for large corporations (we’re looking at you, Amazon) to deliver boxes of anything and everything directly to people’s front doors.²⁴

The point here is that *one does not need street names for navigation until one has strangers in the landscape*. Likewise, data do not need cleaning until there are *strangers in the dataset*. The Clemson University Library dataset was perfectly clear to Clemson’s own librarians. But once hackers in Cambridge get their hands on it, *upstate* started to make a lot less sense, and it was not at all helpful in producing, for instance, a map of all the library records in the United States. Put more generally, once the data scientists involved in a project are not from within the community, once the place of analysis changes, once the scale of the project shifts, or once a single dataset needs to be combined with others—then we have strangers in the dataset.

Who are these strangers? As we’ve already started to suggest, people who work with data are alternately called *unicorns* (because they are rare and have special skills), *wizards* (because they can do magic), *ninjas* (because they execute complicated, expert moves), *rock stars* (because they outperform others), and *janitors* (because they clean messy data) (figure 5.3). Amazon dropped the “janitor” part in a recent job ad, but it managed to work in a few of these metaphors: “Amazon needs a rockstar engineer ... You are passionate ... You succeed fearlessly ... You are a coding ninja.”²⁵

These rock stars and ninjas are strangers in the dataset because, like the hackers in Cambridge, they often sit at one, two, or many levels removed from the collection and maintenance process of the data that they work with. This is a *negative externality*—an inadvertent third-party consequence—that arises when working with open data, application programming interfaces (APIs), and the vast stores of training data available online. These data appear available and ready to mobilize, but what they represent is not always well-documented or easily understood by outsiders. Being a stranger in the dataset is not an inherently bad thing, but it carries significant risk of what renowned postcolonial scholar Gayatri Spivak calls *epistemic violence*—the harm that dominant groups like colonial powers wreak by privileging their ways of knowing over local and Indigenous ways.²⁶

This problem is compounded by the belief that data work is a solitary undertaking. This is reflected in the fact that unicorns are unique by definition, and wizards, ninjas, and rock stars are also all people who seem to work alone. This is a fallacy, of course; every rock star requires a backing band, and if we’ve learned anything from *Harry*

Data Scientists as Unicorns, Wizards, Ninjas, Rock Stars and Janitors Mentions in the Media, 2012–2018

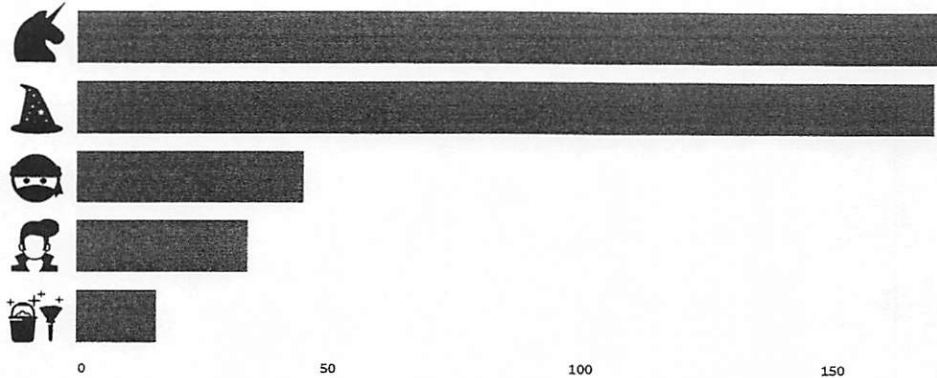


Figure 5.3

Searching Media Cloud between 2012 and 2018 shows that *unicorn* is the most commonly referenced metaphor in relation to data scientists, with *wizard* a close second. There are fewer than fifty articles about data *ninjas*, *rock stars*, and *janitors*, but they appear in high-profile venues like the *Washington Post* and *Forbes*. The Media Cloud platform at www.mediacloud.org was developed at the MIT Center for Civic Media and archives just under a million articles and blog posts every day. Graphic by Catherine D'Ignazio. Data from www.mediacloud.org.

Potter, it's that any particular tap of the wand is the culmination of years of education, training, and support. Wizards, ninjas, rock stars, and janitors each have something else in common: they are assumed to be men.²⁷ If you doubt this assertion, try doing a Google image search and count how many male-presenting wizards and janitors you see before you get to a single female-presenting one. Or consider why news articles about "data janitors" don't describe them as doing "cleaning lady" work? Like janitorial work, which is disproportionately undertaken by working-class people of color, the idea of the data ninja also carries racist connotations.²⁸ And there's even more: shared among the first four terms—unicorns, wizards, ninjas, rock stars—is a focus on the individual's extraordinary technical expertise and their ability to prevail when others cannot. We might have more accurately said "his ability to prevail" because these ideas about individual mastery and prevailing against steep odds are, of course, also associated with men.

There is a "genius" in the world of eviction data—it is Matthew Desmond, designated as such by the MacArthur Foundation for his work on poverty and eviction

in the United States. He is a professor and director of the Eviction Lab at Princeton University, which has been working for several years to compile a national database of evictions and make it available to the general public. Although the federal government collects national data on foreclosures, there is no national database of evictions, something that is desperately needed for the many communities where housing is in crisis.

Initially, the Eviction Lab had approached community organizations like the AEMP to request their data. The AEMP wanted to know more—about privacy protections and how the Eviction Lab would keep the data from falling into landlord hands. Instead of continuing the conversation, the Eviction Lab turned to a real estate data broker and purchased data of lower quality. But in an article written by people affiliated with the AEMP and other housing justice organizations, the authors state, “AEMP and Tenants Together have found three-times the amount of evictions in California as Desmond’s Eviction Lab show.”²⁹

As Desmond tells it, this decision was due to a change in the lab’s data collection strategy. “We’re a research lab, so one thing that’s important to us is the data cleaning process. If you want to know does Chicago evict more people than Boston, you’ve got to compare apples to apples.”³⁰ In Desmond’s view, it is more methodologically sound to compare datasets that have been already aggregated and standardized. And yet Desmond acknowledges that Eviction Lab’s data for the state of California is undercounting evictions; there is even a message on the site that makes that explicit. So here’s an ethical quandary: Does one choose cleaner data at a larger scale that is relatively easy and quick to purchase? Or more accurate data at a local scale for which one has to engage and build trust with community groups?

In this case, the priority was placed on speed at the expense of establishing trusted relationships with actors on the ground, and on broad national coverage at the expense of local accuracy. Though the Eviction Lab is doing important work, continued decisions that prioritize speed and comprehensiveness can’t help but maintain the cultural status of the solitary “genius,” effectively downplaying the work of coalitions, communities, and movements that are—not coincidentally—often led primarily by women and people of color.

What might be gained if we not only recognized but also valued the fact that data work involves multiple voices and multiple types of expertise? What if producing new social relationships—increasing community solidarity and enhancing social cohesion—was valued (and funded) as much as acquiring data? We think this would lead to a multiplication of projects like the AEMP: projects that do demonstrable good with data and do so together with the communities they seek to support.

On Power, Pluralism, and Process

Although the Anti-Eviction Mapping Project could have handed off its valuable data to a single mapmaking rock star-unicorn-ninja-wizard-janitor, the group made an intentional decision to include many designers in the process, including many nonexperts who experienced the power of making maps for the first time. In addition to the diverse array of data products that resulted, which reflected the diverse voices of the AEMP's various collaborators, this decision also had the (wholly intentional) result of building technical capacity. Slowly and surely, map by map, collaboration by collaboration, local residents strengthened their mapping skills and their relationships with each other. This latter result too was intentional; one of the stated goals of the AEMP is to "build solidarity and collectivity among the project's participants who could help one another in fighting evictions and collectively combat the alienation that eviction produces."³¹

This goal reflects a key tenet of feminist thinking, which is the recognition that a multiplicity of voices, rather than one single loud or technical or magical one, results in a more complete picture of the issue at hand. Feminist philosophers like Donna Haraway, who we introduced in chapter 3, prompted a wave of thinkers who have continued to develop the idea that all knowledge is partial, meaning no single person or group can claim an objective view of the capital-T Truth.³² But embracing *pluralism*, as this concept is often described today, does not mean that everything is relative, nor does it mean that all truth claims have equal weight. And it most certainly does not mean that feminists do not believe in science. It simply means that when people make knowledge, they do so from a particular standpoint: from a situated, embodied location in the world. More than that, by pooling our standpoints—or positionalities—together, we can arrive at a richer and more robust understanding of the world.³³

So, how do we begin down the path to this deeper understanding in data science? The first step in activating the value of multiple perspectives is to acknowledge the partiality of your own. This means disclosing your project's methods, your decisions, and—importantly for work that strives to address injustice—your own positionalities. This is called *reflexivity*, and we modeled this in the introduction to this book. You may have heard the phrase coined by David Weinberger, "transparency is the new objectivity."³⁴ We take this to mean that there is a way to build space for transparency plus reflexivity in data science, rather than undertaking projects that purport to be objective (but, as we've discussed, never really are). Transparency and reflexivity allow the people involved in any particular project to be explicit about the methods behind their project, as well as their own identities.

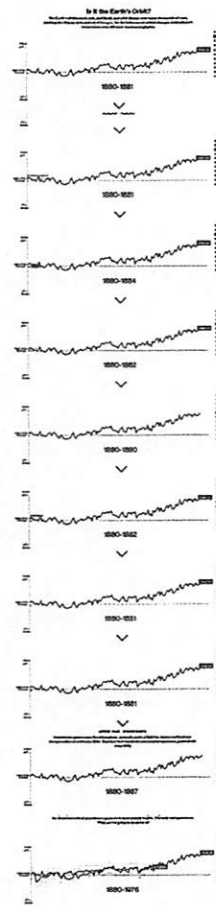
People in journalism have been doing this for some time—at least as it relates to their data and methods. For example, Bloomberg's interactive visualization "What's Really Warming the World?" (figure 5.4) walks the reader through a range of common arguments that try to explain away global warming with reasons that don't have to do with human industry or behavior.³⁵ It's a compelling piece in terms of content alone, but another interesting thing about it is that it devotes nearly a third of its screen real estate to describing its data and methods.

Providing access to data and describing the methods employed are becoming conventions in data journalism, just as they are in scientific fields. Although these accounts are presently focused on technical details—where the data were from, what statistical models were developed, and what analysis was performed—there is a seed of possibility for using this same space to reveal additional details: Who was on the team? What were points of tension and disagreement? When did the team need to talk to data stewards or domain experts or local communities? Which hypotheses were pursued but ultimately proved false? There is a story about how every evidence-based argument comes into being, and it is often a story that involves money and institutions, as well as humans and tools. Revealing this story through transparency and reflexivity can be a feminist act.

Reflexivity can sometimes be as simple as being explicit about or even visualizing who is doing the counting and mapping behind the scenes.³⁶ Take the example of the aerial-mapping image in figure 5.5.³⁷ The Public Laboratory for Open Technology and Science (Public Lab) is a citizen science group that got its start during the BP oil spill in 2010.³⁸ It makes high-resolution aerial maps by flying balloons and kites, which dangle cheap digital cameras over the environmental sites they seek to study. The technique is low cost, but the imagery produced is often higher resolution than existing satellite imagery because of the proximity to the ground. As in the image, the mappers themselves are often visible in the final product in the form of little bodies, gathered in boats or standing in clumps on a shoreline, looking up at the camera above them. The balloon string leads the eye back to their forms. Here the creators are not distanced or absent but represented in the final product. Literally.

Data for "Good" versus Data for Co-liberation

Embracing the value of multiple perspectives shouldn't stop with transparency and reflexivity. It also means actively and deliberately inviting other perspectives into the data analysis and storytelling process—more specifically, those of the people most marginalized in any given context. Intersectional feminist scholars have long insisted that



70% story

Methodology

Model

Global Temperature Project

Model Model Temperature

What the Data Shows

Conclusions

Data

Acknowledgments

30% method & "self-disclosure"

Figure 5.4

"What's Really Warming the World?," published in 2015, devotes a third of its real estate to describing the methods for how the authors worked with the data. Graphic by Catherine D'Ignazio, based on reporting by Eric Roston and Blacki Migliozi for *Bloomberg Businessweek*.



Figure 5.5

From a Public Lab research note by Eymund Diegel about mapping sewage flows in the Gowanus Canal in 2012 after Hurricane Sandy. Note the people on boats doing the mapping and the balloon tether that links the camera and image back to their bodies. Courtesy of Eymund Diegel for Public Lab.

we should be creating new knowledge and new designs from the margins. “Marginalized subjects have an epistemic advantage, a particular perspective that scholars should consider, if not adopt, when crafting a normative vision of a just society,” as Black feminist scholar Jennifer C. Nash explains.³⁹

What does this mean? From a gender perspective, it means beginning with the perspectives of women and nonbinary people. On a project that involves international development data, it means beginning not with institutional goals but with Indigenous standpoints. For the AEMP, it means centering the voices and experiences of those who have been evicted. Follow-up work about designing from the margins argues that designers and engineers shouldn’t only be engaging people at the margins but also actively working to dismantle the center/margins distinction in the first place.⁴⁰ More

recently, the Design Justice Network has transformed this key tenet of intersectional thinking into one of its design principles, stating: “We center the voices of those who are directly impacted by the outcomes of the design process.”⁴¹

How might this work? To begin, it requires a design process in which many actors can participate—people with technical expertise, as well as those with lived expertise, domain expertise, organizing expertise, and community history expertise. It also means shifting the overarching goal of such projects from “doing good with data” to designing for *co-liberation*; remember from chapter 2 that this is an end state in which people from dominant groups and minoritized groups work together to free themselves from oppressive systems. The key differences between data for good and data for co-liberation are highlighted in table 5.1.

Data for good is a frame that is increasingly employed to describe data science projects that are socially engaged and/or undertaken in the public interest. The Bloomberg corporation has been sponsoring Data for Good conferences since 2014. Nonprofit consulting groups like Delta Analytics have sprouted up to match volunteers with technical expertise with mission-driven organizations. In 2019, one such organization, DataKind, received a \$20 million gift from a funding collaborative called Data Science for Social Impact, with monies contributed by the Rockefeller Foundation

Table 5.1

Features of “data for good” versus data for co-liberation

	“Data for good”	Data for co-liberation
Leadership by members of minoritized groups working in community		✓
Money and resources managed by members of minoritized groups		✓
Data owned and governed by the community		✓
Quantitative data analysis “ground truthed” through a participatory, community-centered data analysis process		✓
Data scientists are not rock stars and wizards, but rather facilitators and guides		✓
Data education and knowledge transfer are part of the project design		✓
Building social infrastructure—community solidarity and shared understanding—is part of the project design		✓

and Mastercard.⁴² There are also educational experiments underway, like the Utrecht Data School and the University of Chicago's Data Science for Social Good summer fellowship program. In the latter, aspiring data scientists work with governments and nonprofit organizations to address problems in diverse domains such as education, health, public safety, and economic development.⁴³ Related efforts in artificial intelligence have sprung up, like the AI4Good Foundation, Project Impact sponsored by Intel Corporation, the women-centered AI Summer Lab at McGill University—the list goes on.

These efforts have had demonstrable social impact. And yet, there remains a nagging fuzziness with respect to what it means to “do good.” Whose good are we talking about? What are the terms? Who maintains the databases when the unicorn-wizards leave the community? And who pays for the cloud storage when the development portion of the project is complete?

These issues are beginning to be discussed within the data for good community. Sara Hooker, a deep-learning researcher at Google Brain and the founder of Delta Analytics, has observed that the idea of “data for good” lacks precision.⁴⁴ To contribute clarity to the phrase, Hooker proposes a rough taxonomy of this type of work, identifying four distinct flavors: (1) volunteering skilled labor, (2) donating tech, (3) working with nonprofits or governments as partners, and/or (4) running data-education programs in underserved communities.⁴⁵ In each of these areas, there remain some thorny issues: the fickleness of volunteer labor, the fact that even committed volunteers often lack local knowledge, the community's capacity (or lack thereof) to perform and/or pay for its own technical maintenance, the fact that work initiated by nonprofits and governments cannot always be assumed to be “good,” and so on. Hooker's point is that when the goal is a vague notion of “good,” there is no way to address such concerns.

By contrast, a model that positions *co-liberation* as the end goal leads to a very specific set of processes and practices, as well as criteria for success. Co-liberation is grounded in the belief that enduring and asymmetrical power relations among social groups serve as the root cause of many societal problems. Rather than framing acts of technical service as benevolence or charity, the goal of co-liberation requires that those technical workers acknowledge that they are engaged in a struggle for their own liberation as well, even and especially when they are members of dominant groups.

For these reasons, data projects designed with co-liberation in mind must be very specific about the power dynamics involved. They must take preemptive steps to counteract the privilege hazard that comes with work undertaken by members of dominant groups. For example, one such step is to deliberately concentrate strategic leadership,

and Mastercard.⁴² There are also educational experiments underway, like the Utrecht Data School and the University of Chicago's Data Science for Social Good summer fellowship program. In the latter, aspiring data scientists work with governments and nonprofit organizations to address problems in diverse domains such as education, health, public safety, and economic development.⁴³ Related efforts in artificial intelligence have sprung up, like the AI4Good Foundation, Project Impact sponsored by Intel Corporation, the women-centered AI Summer Lab at McGill University—the list goes on.

These efforts have had demonstrable social impact. And yet, there remains a nagging fuzziness with respect to what it means to “do good.” Whose good are we talking about? What are the terms? Who maintains the databases when the unicorn-wizards leave the community? And who pays for the cloud storage when the development portion of the project is complete?

These issues are beginning to be discussed within the data for good community. Sara Hooker, a deep-learning researcher at Google Brain and the founder of Delta Analytics, has observed that the idea of “data for good” lacks precision.⁴⁴ To contribute clarity to the phrase, Hooker proposes a rough taxonomy of this type of work, identifying four distinct flavors: (1) volunteering skilled labor, (2) donating tech, (3) working with nonprofits or governments as partners, and/or (4) running data-education programs in underserved communities.⁴⁵ In each of these areas, there remain some thorny issues: the fickleness of volunteer labor, the fact that even committed volunteers often lack local knowledge, the community's capacity (or lack thereof) to perform and/or pay for its own technical maintenance, the fact that work initiated by nonprofits and governments cannot always be assumed to be “good,” and so on. Hooker's point is that when the goal is a vague notion of “good,” there is no way to address such concerns.

By contrast, a model that positions *co-liberation* as the end goal leads to a very specific set of processes and practices, as well as criteria for success. Co-liberation is grounded in the belief that enduring and asymmetrical power relations among social groups serve as the root cause of many societal problems. Rather than framing acts of technical service as benevolence or charity, the goal of co-liberation requires that those technical workers acknowledge that they are engaged in a struggle for their own liberation as well, even and especially when they are members of dominant groups.

For these reasons, data projects designed with co-liberation in mind must be very specific about the power dynamics involved. They must take preemptive steps to counteract the privilege hazard that comes with work undertaken by members of dominant groups. For example, one such step is to deliberately concentrate strategic leadership,

financial resources, and data ownership in the hands of collaborators from minoritized groups.⁴⁶ It also recognizes that differential power has a silencing effect and that for a variety of reasons—as discussed in chapter 4—quantitative data can leave people out. To address these almost certain gaps, a model of data for co-liberation would deliberately pair quantitative analyses with inclusive civic processes, resulting in locally informed, ground-truthed insights that derive from many perspectives.

The scope of data for co-liberation is also broader. It explicitly includes two additional outcomes that data for good typically does not: (1) knowledge transfer and (2) building social infrastructure. The first involves a two-way exchange: in one direction, technical capacity building within the community so that any data products can be maintained and/or enhanced without requiring external expertise, and in the other, enhanced understanding of and respect for local knowledge for external collaborators, as well as chipping away at their own individual and institutional privilege hazards. Building social infrastructure involves an explicit focus on cultivating community solidarity through the project. This entails allocating financial and human resources to the community aspects of the project and not only to technical aspects like processing data or building apps. In the co-liberation model, data science projects become community science projects. They take place simultaneously in the database and in public space.

Data for Co-liberation in Action

What does data for co-liberation look like in action? Are there examples of feminist data science that value quantitative methods *and* pluralistic processes, data education *and* community solidarity?

Since 2012, Rahul and Emily Bhargava have partnered with community organizations from Belo Horizonte to Boston to create *data murals* in public spaces (figure 5.6). These are large-scale infographics that are both designed by and tell stories about the people who live and work in those spaces. In all cases, the people themselves sought out the collaboration, having recognized a need within their own community space. For example, in 2013, Groundwork Somerville, an urban agriculture nonprofit, approached the Bhargavas because it was in the process of establishing its first urban farm. As Emily recalls, “The site was disorderly—it was behind a used car parts building and hidden between other semi-industrial lots. They had built raised beds and planted for one growing season but passersby were stealing the vegetables.”⁴⁷ The organization was also running a high school employment program called the Green

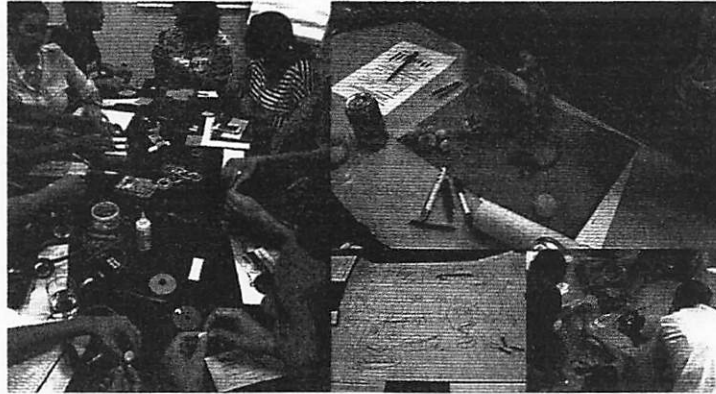


Figure 5.6

The process of making a data mural involves conversation, building prototypes with craft materials, workshops in data analysis, and actual painting. Courtesy of Data Therapy, Emily and Rahul Bhargava, 2018.

Team but struggling to fully involve the young people in its mission to create healthier communities.

Linking those two challenges together, Groundwork Somerville and the Bhargavas decided to enlist the young people in creating a mural that illustrated the purpose of the farm to the community (figure 5.7). First, they brought together data from several sources: demographic data from the city, geographic information system (GIS) data on unused lots, and internal data from Groundwork Somerville that included growing records, food donations, and attendance logs at community events. Then they worked with the learners over several after-school sessions to review and discuss the data and engage in storyfinding (aka data analysis). By the end of these sessions, the youth had sketched the overall outline and iconography of the resulting mural. Read left to right, the mural frames the problem: A man grasps for a basket of veggies, but it says “healthy food is hard to get.” The claim is backed up by data that show the cost of healthy food and the number of people with prediabetes. Additional data document the number of unused lots in the city and the amount of land that has been reclaimed for urban farming, pointing to future opportunity. In the next section, the mural depicts a Groundwork Somerville truck bringing affordable produce to the neighborhood and includes the fact that it employs over four hundred youth residents. The data mural ends with a vision for a unified and healthy community, “Together Livin’ Better.”

On July 30, 2013, the mayor of Somerville and other community leaders attended the ribbon-cutting to officially launch the renovated garden. Emily describes the visit:



Figure 5.7

The Groundwork Somerville data mural, painted by youth, staff and volunteers at Groundwork Somerville, and the Bhargavas in 2013. Courtesy of Data Therapy, Emily and Rahul Bhargava.

"The youth, having just spent weeks looking at the data, painting the mural together, and building relationships with staff and volunteers, were able to talk about the story in great detail to their elected officials."⁴⁸

The partnership represents some of the best aspects of data projects designed for co-liberation. But murals are just one possible output from this type of pluralistic, community-centered process.⁴⁹ For example, Digital Democracy works with Indigenous groups around the world to defend their rights through collecting data and making maps.⁵⁰ In the process, they have developed SMS services with domestic violence groups in Haiti and supported the Wapichana people in Guyana to make a data-driven case for land rights to the government. In another example, the Westside Atlanta Land Trust (WALT) has worked toward affordable housing in Atlanta, Georgia, through participatory data collection.⁵¹ Disappointed in what it termed "triple-incorrect" county-level data, the organization set out to collect its own spatial dataset about property abandonment and disinvestment and has used it to push municipal policymakers for change.⁵²

Each of these projects—the data mural, the land rights claim, and the abandoned property map—could be said to exemplify the data for co-liberation model. Each originates from a need articulated by the community, rather than projected onto it by more powerful institutions. Each relies upon methods of data collection and processing. And each also incorporates an explicit process of knowledge transfer from external collaborators (consultants, academics, nonprofit specialists) to the community itself. More importantly, none of those external collaborators envision themselves as unicorns, janitors, ninjas, wizards, or rock stars. "At Digital Democracy, we try to fight the superhero narrative," says director Emily Jacobi. "We are sidekicks rather than superheroes."⁵³

Along these lines, each of these projects is explicitly designed to build community solidarity and shared understanding around a civic issue. As Amanda Meng and Carl DiSalvo explain with respect to the WALT project, "Data collection became a way of generating not simply awareness of the conditions, but a sensitivity to the conditions and to [community members'] own experience of and situation of being blinded to these conditions."⁵⁴ Data collection and participatory analysis become a gathering technology—a kind of campfire—where information is exchanged, social cohesion is enhanced, and future actions are co-conspired.⁵⁵ The campfire model also has the effect of challenging the deficit narratives that so often surround underserved communities.⁵⁶ Instead, as social movement scholar Maribel Casas-Cortés has asserted, cultivating community solidarity contributes to a sense of collective agency and transformative possibility, what she describes as "an innovative sense of political participation and re-invigorating political imaginaries."⁵⁷

Does the Campfire Scale?

Data for co-liberation is simultaneously more specific than data for good and larger in scope. By advocating for community-based leadership and resources, qualitative investigations to complement quantitative analyses, and intentional project design in areas that exceed the purview of data for good projects, such as knowledge transfer and building community solidarity, data for co-liberation enables the participation of many actors and agents. But can this pluralistic model scale up? What would "big data for co-liberation" look like? Would it be possible at all?

Questions about big data versus little data or quantitative data versus qualitative data are far too often framed as false binaries. And as with all false binaries, there are paths through those distinctions that combine and multiply options, rather than reduce or divide the world into fake choices. The key question to keep in mind is how we can scale up data for co-liberation in ways that remain careful, community-based, and complex.⁵⁸

One real-world example of this intentional practice is the Global Atlas of Environmental Justice (EJ Atlas), a large-scale research project and open archive (figure 5.8) directed by scholars Leah Temper and Joan Martinez-Alier.⁵⁹ Initiated in 2012 by a team of researchers at the Universitat Autònoma de Barcelona, in Spain, the EJ Atlas represents a systematic collection of global ecological conflicts. It was created in part as a response to assertions that environmental justice scholars were focusing too much attention on local case studies at the expense of making global connections. For the

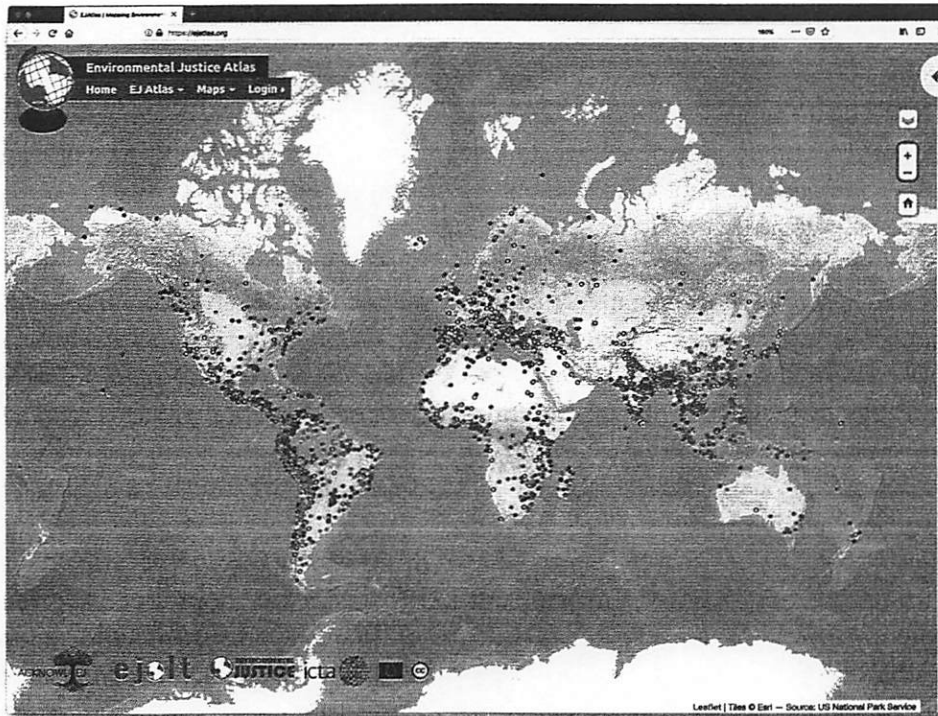


Figure 5.8

The Global Atlas of Environmental Justice (EJ Atlas; <https://ejatlas.org/>) works in partnership with activists, civil society organizations, and social movements to systematically document ecological conflicts around the globe. The scope and scale of the EJ Atlas enables activists to connect with others and facilitates researchers studying conflicts in a quantitative and comparative context, without sacrificing a commitment to a pluralistic process and the dignity of local and community knowledges. Courtesy of the Global Atlas of Environmental Justice, 2019.

most part, the global ecological conflicts that the EJ Atlas collects are cases straight out of the matrix of domination, of wealthy people overutilizing natural resources and displacing environmental risk and degradation to poorer people, who are often also minoritized because of their gender, indigeneity, race, and/or geography. For example, one of the cases in the EJ Atlas discusses the efforts of Berta Cáceres and the Lenca Indigenous people to oppose the construction of a dam and hydropower plant in Honduras.⁶⁰ Cáceres won international awards for her community organizing, only to be tragically assassinated by employees of the company building the dam. (The assassins were later convicted in a Honduran court of law.⁶¹) In the EJ Atlas, the entry on Cáceres's organizing efforts includes copious links, photos, and geographic data about the

dam, as well as information about where the incident falls in the atlas's typology of conflict.

The EJ Atlas demonstrates that scale is not incompatible with valuing local knowledges and relationships with local actors. Temper and her colleagues were able to draw on past relationships with partner organizations to assemble their archive, demonstrating how time invested in building relationships at the outset of a project, or of a career, continues to accrue benefits for all parties involved.⁶² At the time of writing, the EJ Atlas includes close to three thousand cases of ecological conflicts from around the world. Collaborators have created submaps from the atlas for countries including Colombia, Italy, and Turkey. Activists have used the atlas to draw media and policymaker attention to overlooked environmental conflicts, such as the construction of a ski resort in the middle of a nature park in Kazakhstan, which has disturbed hydrological systems.⁶³ The atlas has also enabled comparative empirical research, like that undertaken by economist Begüm Özkaynak and several colleagues.⁶⁴ Their work employs social network analysis to study the relationships among mining corporations, their financiers, and environmental groups challenging that practice to understand the geography of the actors and their connections to each other.⁶⁵

Scale is emphatically not incompatible with the feminist imperative to value multiple and local knowledges. Research like Özkaynak's and that of her colleagues would not be possible without a large-scale archive like the EJ Atlas. A pluralistic, participatory, iterative process like the EJ Atlas will take longer to scale than the extractive, quantity-at-all-costs approach of conventional big data. But ultimately the data—and the relationships, and the community capacity—will be of higher quality.

Embrace Pluralism

The fifth principle of data feminism is to *embrace pluralism* in the whole process of working with data, from collection to analysis to communication to decision-making. As we describe in this chapter, data work carries a high risk of enacting what Gayatri Spivak has termed *epistemic violence*, particularly when the people doing the work are strangers in the dataset, when they are one or more steps removed from the local context of the data, and when they view themselves (or are viewed by society) as unicorns, rock stars, and wizards.

Embracing pluralism is a feminist strategy for mitigating this risk. It allows both time and space for a range of participants to contribute their knowledge to a data project and to do so at all stages of that project. In contrast to an underspecified data for good model, embracing pluralism offers a way to work toward a model of data for

co-liberation. This means transferring knowledge from experts to communities and explicitly cultivating community solidarity in data work, as we see in the case of the Anti-Eviction Mapping Project. Moreover, embracing pluralism is not incompatible with “bigness” or scale; the EJ Atlas shows how pluralistic processes can be used to assemble a global archive and support empirical work in the service of justice. A single data scientist wizard will never defeat the matrix of domination alone, no matter how powerful their spells might be. But a well-designed, data-driven, participatory process, one that centers the standpoints of those most marginalized, empowers project participants, and builds new relationships across lines of social difference—well, that might just have a chance.

6 The Numbers Don't Speak for Themselves

Principle: Consider Context

Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.

In April 2014, 276 young women were kidnapped from their high school in the town of Chibok in northern Nigeria. Boko Haram, a militant terrorist group, claimed responsibility for the attacks. The press coverage, both in Nigeria and around the world, was fast and furious. SaharaReporters.com challenged the government's ability to keep its students safe. CNN covered parents' anguish. The *Japan Times* connected the kidnappings to the increasing unrest in Nigeria's northern states. And the BBC told the story of a girl who had managed to evade the kidnappers. Several weeks after this initial reporting, the popular blog *FiveThirtyEight* published its own data-driven story about the event, titled "Kidnapping of Girls in Nigeria Is Part of a Worsening Problem."¹ The story reported skyrocketing rates of kidnappings. It asserted that in 2013 alone there had been more than 3,608 kidnappings of young women. Charts and maps accompanied the story to visually make the case that abduction was at an all-time high (figure 6.1).

Shortly thereafter, the news website had to issue an apologetic retraction because its numbers were just plain wrong. The outlet had used the Global Database of Events, Language and Tone (GDELT) as its data source. GDELT is a big data project led by computational social scientist Kalev Leetaru. It collects news reports about events around the world and parses the news reports for actors, events, and geography with the aim of providing a comprehensive set of data for researchers, governments, and civil society. GDELT tries to focus on conflict—for example, whether conflict is likely between two countries or whether unrest is sparking a civil war—by analyzing media reports. However, as political scientist Erin Simpson pointed out to *FiveThirtyEight* in a widely cited Twitter thread, GDELT's primary data source is *media reports* (figure 6.2).² The project is not at a stage at which its data can be used to make reliable claims about *independent*

Abduction Trends

Daily kidnappings in Nigeria, 1982–2014

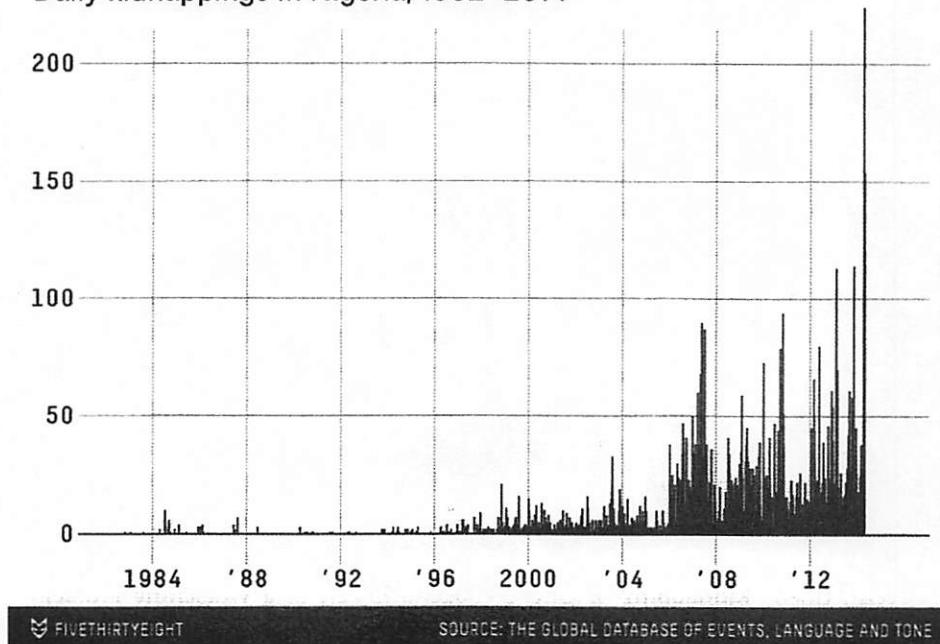


Figure 6.1

In 2014, *FiveThirtyEight* erroneously charted counts of “daily kidnappings” in Nigeria. The news site failed to recognize that the data source it was using was not counting *events*, but rather *media reports about events*. Or some events and some media reports. Or it was counting something, but we are still not sure what. Image by *FiveThirtyEight*.

cases of kidnapping. The kidnapping of schoolgirls in Nigeria was a single event. There were thousands of global media stories about it. Although GDELT de-duplicated some of those stories to a single event, it still logged, erroneously, that hundreds of kidnapping events had happened that day. The *FiveThirtyEight* report had counted each of those GDELT pseudoevents as a separate kidnapping incident.

The error was embarrassing for *FiveThirtyEight*, not to mention for the reporter, but it also helps to illustrate some of the larger problems related to data found “in the wild.” First, the hype around “big data” leads to projects like GDELT wildly overstating the completeness and accuracy of its data and algorithms. On the website and in publications, the project leads have stated that GDELT is “an initiative to construct a catalog



Figure 6.2

Two tweets by Erin Simpson in response to *FiveThirtyEight*'s erroneous interpretation of the GDELT dataset. Tweets by Erin Simpson on May 13, 2014.

of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day."³ That giant mouthful describes no small or impotent big data tool. It is clearly Big Dick Data.

Big Dick Data is a formal, academic term that we, the authors, have coined to denote big data projects that are characterized by masculinist, totalizing fantasies of world domination as enacted through data capture and analysis. Big Dick Data projects ignore context, fetishize size, and inflate their technical and scientific capabilities.⁴ In GDELT's case, the question is whether we should take its claims of big data at face value or whether the Big Dick Data is trying to trick funding organizations into giving the project massive amounts of research funding. (We have seen this trick work many times before.)

The GDELT technical documentation does not provide any more clarity as to whether it is counting media reports (as Simpson asserts) or single events. The database *FiveThirtyEight* used is called the GDELT Event Database, which certainly makes it sound like it's counting events. The GDELT documentation states that "if an event has been seen before it will not be included again," which also makes it sound like it's counting events. And a 2013 research paper related to the project confirms that GDELT is indeed counting events, but only events that are unique to specific publications. So it's counting events, but with an asterisk. Compounding the matter, the documentation offers no guidance as to what kinds of research questions are appropriate to ask the database or what the limitations might be. People like Simpson who are familiar with the area of research known as *event detection*, or members of the GDELT community, may know to not believe (1) the title of the database, (2) the documentation, and (3) the marketing hype. But how would outsiders, let alone newcomers to the platform, ever know that?

We've singled out GDELT, but the truth is that it's not very different from any number of other data repositories out there on the web. There are a proliferating number of portals, observatories, and websites that make it possible to download all manner of government, corporate, and scientific data. There are APIs that make it possible to write little programs to query massive datasets (like, for instance, all of Twitter) and download them in a structured way.⁵ There are test datasets for network analysis, machine learning, social media, and image recognition. There are fun datasets, curious datasets, and newsletters that inform readers of datasets to explore for journalism or analysis.⁶ In our current moment, we tend to think of this unfettered access to information as an inherent good. And in many ways, it is kind of amazing that one can just google and download data on, for instance, pigeon racing, the length of guinea pig teeth, or every single person accused of witchcraft in Scotland between 1562 and 1736—not to mention truckloads and truckloads of tweets.⁷

And though the schooling on data verification received by *FiveThirtyEight* was rightly deserved, there is a much larger issue that remains unaddressed: the issue of context. As we've discussed throughout this book, one of the central tenets of feminist thinking is that all knowledge is *situated*. A less academic way to put this is that *context matters*. When approaching any new source of knowledge, whether it be a dataset or dinner menu (or a dataset of dinner menus), it's essential to ask questions about the social, cultural, historical, institutional, and material conditions under which that knowledge was produced, as well as about the identities of the people who created it.⁸ Rather than seeing knowledge artifacts, like datasets, as raw input that can be simply fed into a statistical analysis or data visualization, a feminist approach insists on connecting data

back to the context in which they were produced. This context allows us, as data scientists, to better understand any functional limitations of the data and any associated ethical obligations, as well as how the power and privilege that contributed to their making may be obscuring the truth.

Situating Data on the Wild Wild Web

The major issue with much of the data that can be downloaded from web portals or through APIs is that they come without context or metadata. If you are lucky you *might* get a paragraph about where the data are from or a data dictionary that describes what each column in a particular spreadsheet means. But more often than not, you get something that looks like figure 6.3.

The data shown in the figure—open budget data about government procurement in São Paulo, Brazil—do not look very technically complicated. The complicated part is figuring out how the business process behind them works. How does the government run the bidding process? How does it decide who gets awarded a contract? Are all the bids published here, or just the ones that were awarded contracts? What do terms like *competition*, *cooperation agreement*, and *terms of collaboration* mean to the data publisher? Why is there such variation in the publication numbering scheme? These are only a few of the questions one might ask when first encountering this dataset. But without answers to even some of these questions—to say nothing of the local knowledge required to understand how power is operating in this particular ecosystem—it would be difficult to even begin a data exploration or analysis project.

This scenario is not uncommon. Most data arrive on our computational doorstep context-free. And this lack of context becomes even more of a liability when accompanied by the kind of marketing hype we see in GDELT and other Big Dick Data projects. In fact, the 1980s version of these claims is what led Donna Haraway to propose the concept of situated knowledge in the first place.⁹ Subsequent feminist work has drawn on the concept of situated knowledge to elaborate ideas about ethics and responsibility in relation to knowledge-making.¹⁰ Along this line of thinking, it becomes the responsibility of the person evaluating that knowledge, or building upon it, to ensure that its “situatedness” is taken into account. For example, information studies scholar Christine Borgman advocates for understanding data in relation to the “knowledge infrastructure” from which they originate. As Borgman defines it, a *knowledge infrastructure* is “an ecology of people, practices, technologies, institutions, material objects, and relationships.”¹¹ In short, it is the context that makes the data possible.

Nr. Publicação	Licitador	Modalidade	Dt. Abertura	Objeto
01-PREF/SECOM/2018	Secretaria do Governo Municipal - SGM	CONCORRÊNCIA	10/06/2019 14:00	Contratação de empresa para prestação de serviços de assessoria de imprensa e comunicação para a PREF/SECOM
03/SGM-2019	SGM - Administração de Compras e Contratos	CONCORRÊNCIA	03/06/2019 10:30	ALIENAÇÃO DO IMÓVEL MUNICIPAL, SITUADO NA AVENIDA PROFESSOR ALCEU MAYNARD ARAÚJO, NO DISTRITO DE SANTO AMARO.
01/SMPED/2019	Secretaria Municipal da Pessoa com Deficiência - SMPED	TOMADA DE PREÇOS	31/05/2019 10:30	Contratação de empresa especializada em produção e atualização de material didático orientador e informativo com produção de conteúdo em versão digital acessível, visando a subsidiar a capacitação do público alvo dos cursos e eventos oferecidos pela Secretaria Municipal da Pessoa com Deficiência - SMPED.
19/SME/2019	Secretaria Municipal de Educação - SME	PREGÃO ELETRÔNICO	30/05/2019 10:30	Registro de preços para aquisição de alimentos não perecíveis açúcar refinado.
007/2019	São Paulo Transporte S/A	PREGÃO ELETRÔNICO	27/05/2019 10:00	OBJETO: AQUISIÇÃO DE 6 (SEIS) EQUIPAMENTOS APPLIANCE DO TIPO UTM?, COM LICENÇAS DE SEGURANÇA, INSTALAÇÃO E SUPORTE TÉCNICO, PELO PERÍODO DE 24 (VINTE E QUATRO) MESES
109/SMADS/2019	Secretaria Municipal de Assistência e Desenvolvimento Social - SMADS	TERMO DE COLABORAÇÃO - EDITAL	24/05/2019 10:00	C J
108/SMADS/2019	Secretaria Municipal de Assistência e Desenvolvimento Social - SMADS	TERMO DE COLABORAÇÃO - EDITAL	24/05/2019 10:00	Centro de Acolhida com Inserção Produtiva para Adultos em Situação de Rua
001/2018/SEHAB	Secretaria Municipal de Habitação - SEHAB - GABINETE	CONCORRÊNCIA	24/05/2019 10:00	EXECUÇÃO DE OBRAS DE CONSTRUÇÃO DE EMPREENDIMENTO HABITACIONAL DE INTERESSE SOCIAL E DE USO MISTO, DENOMINADO COLISEU, NO ÂMBITO DA OPERAÇÃO URBANA CONSORCIADA FARIA LIMA
002/SVMA/2019	Secretaria Municipal do Verde e Meio Ambiente - SVMA	CONCORRÊNCIA	23/05/2019 10:30	CONTRATAÇÃO DE SERVIÇOS TÉCNICOS ESPECIALIZADOS PARA A ELABORAÇÃO DO PLANO DE MANEJO DA ÁREA DE PROTEÇÃO AMBIENTAL (APA) BORORÉ-COLÔNIA
070/18	São Paulo Turismo - SPTURIS	PREGÃO ELETRÔNICO	22/05/2019 10:00	Contratação de empresa, sob o regime de empreitada por preço unitário, para prestação de serviços de BOMBEIRO PROFISSIONAL CIVIL, por um período de 12 (doze) meses, prorrogáveis por iguais ou menores períodos, conforme bases, especificações e condições do Edital e seus Anexos.
093/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	22/05/2019 09:00	Registro de preços para o fornecimento de PAPEL CREPADO E SWAB, ALCOOL 70% PARA ANTI-SEPSIA.
121/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	21/05/2019 10:30	Registro de preços para o fornecimento de KIT PARA IDENTIFICAÇÃO QUALITATIVA PARA O COMPLEXO M. TUBERCULOSIS.
18/SME/2019	Secretaria Municipal de Educação - SME	PREGÃO ELETRÔNICO	21/05/2019 10:30	Registro de preço para aquisição de Item A: Sardinha em óleo comestível e Item B: Atum em pedaços em conserva.
166/2019	Autorarquia Hospitalar Municipal - AHM	PREGÃO ELETRÔNICO	21/05/2019 09:30	AQUISIÇÃO DE SULFAMETOXAZOL 80 MG/ML + TRIMETOPRIMA 16 MG/ML 5 ML, PARA AS UNIDADES DA AUTARQUIA HOSPITALAR MUNICIPAL.
119/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	20/05/2019 10:30	Registro de preços para o fornecimento de ETIQUETA TÉRMICA CONTINUA, AUTOADESIVA, PARA IMPRESSÃO TÉRMICA 7 62MM X 15M.
117/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	20/05/2019 09:30	Aquisição de MATERIAL ODONTOLÓGICO - FÓRCEPS PARA USO ODONTOLÓGICO.
047/2019-HMEC	Hospital Municipal Maternidade-Escola Dr. Mario de Moraes Altenfelder Silva	PREGÃO ELETRÔNICO	20/05/2019 09:00	BERACTANTO SUSPENSÃO INTRA-TRAQUEAL 25 MG/ML FAM 8,0 ML ? FAM
103/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	17/05/2019 10:30	Registro de preços para o fornecimento de MATERIAL DE LABORATÓRIO - COLETOR UNIVERSAL ESTÉRIL, PIPETA DE TRANSFERÊNCIA E SWAB DE RAYON.
055/2019-HMEC	Hospital Municipal Maternidade-Escola Dr. Mario de Moraes Altenfelder Silva	PREGÃO ELETRÔNICO	17/05/2019 10:00	PLACA DESCARTAVEL PARA ELETROCIURGIA
002/2019	São Paulo Obras - SP Obras	TOMADA DE PREÇOS	17/05/2019 09:30	Contratação de empresa especializada em engenharia e arquitetura para execução das obras de reforma para Implantação do DESCOMPLICA SP ? UNIDADE SÃO MATEUS .

Figure 6.3

Open budget data about procurement and expenses from the São Paulo prefecture in Brazil. Although Brazil has some of the most progressive transparency laws on the books, the data that are published aren't necessarily always accessible or usable by citizens and residents. In 2013, researcher Gisele Craveiro worked with civil society organizations to give this open budget data more context. Images from SIGRC for the Prefecture of São Paulo, Brazil.

Ironically, some of the most admirable aims and actions of the open data movement have worked against the ethical urgency of providing context, however inadvertently. *Open data* describes the idea that anyone can freely access, use, modify, and share data for any purpose. The open data movement is a loose network of organizations, governments, and individuals. It has been active in some form since the mid-2000s, when groups like the Open Knowledge Institute were founded and campaigns like Free Our Data from the *Guardian* originated to petition governments for free access to public records.¹² The goals are good ones in theory: economic development by building apps and services on open data; faster scientific progress when researchers share knowledge; and greater transparency for journalists, citizens, and residents to be able to use public information to hold governments accountable. This final goal was a major part of the framing of former US president Obama's well-known memorandum on transparency and open government.¹³ On his very first day in office, Obama signed a memorandum that directed government agencies to make all data open by default.¹⁴ Many more countries, states, and cities have followed suit by developing open data portals and writing open data into policy. As of 2019, seventeen countries and over fifty cities and states have adopted the International Open Data Charter, which outlines a set of six principles guiding the publication and accessibility of government data.¹⁵

In practice, however, limited public funding for technological infrastructure has meant that governments have prioritized the "opening up" part of open data—publishing spreadsheets of things like license applications, arrest records, and flood zones—but lack the capacity to provide any context about the data's provenance, let alone documentation that would allow the data to be made accessible and usable by the general public. As scholar Tim Davies notes, raw data dumps might be good for starting a conversation, but they cannot ensure engagement or accountability.¹⁶ The reality is that many published datasets sit idle on their portals, awaiting users to undertake the intensive work of deciphering the bureaucratic arcana that obscures their significance. This phenomenon has been called *zombie data*: datasets that have been published without any purpose or clear use case in mind.¹⁷

Zombies might be bad for brains, but is zombie data really a problem? *Wired* magazine editor Chris Anderson would say, emphatically, "No." In a 2008 *Wired* article, "The End of Theory," Anderson made the now-infamous claim that "the numbers speak for themselves."¹⁸ His main assertion was that the advent of big data would soon allow data scientists to conduct analyses at the scale of the entire human population, without needing to restrict their analysis to a smaller sample. To understand his claim, you need to understand one of the basic premises of statistics.

Statistical inference is based on the idea of sampling: that you can infer things about a population (or other large-scale phenomenon) by studying a random and/or representative sample and then mapping those findings back on the population (or phenomenon) as a whole. Say that you want to know who all of the 323 million people in the US will vote for in the coming presidential election. You couldn't contact all of them, of course, but you could call three thousand of them on the phone and then use those results to predict how the rest of the people would likely vote. There would also need to be some statistical modeling and theory involved, because how do you know that those three thousand people are an accurate representation of the whole population? This is where Anderson made his intervention: at the point at which we have data collected on the entire population, we no longer need modeling, or any other "theory" to first test and then prove. We can look directly at the data themselves.

Now, you can't write an article claiming that the basic structure of scientific inquiry is obsolete and not expect some pushback. Anderson wrote the piece to be provocative, and sure enough, it prompted numerous responses and debates, including those that challenge the idea that this argument is a "new" way of thinking in the first place (e.g., in the early seventeenth century, Francis Bacon argued for a form of inductive reasoning, in which the scientist gathers data, analyzes them, and only thereafter forms a hypothesis).¹⁹ One of Anderson's major examples is Google Search. Google's search algorithms don't need to have a hypothesis about *why* some websites have more incoming links—other pages that link to the site—than others; they just need a way to determine the number of links so they can use that number to determine the popularity and relevance of the site in search results. We no longer need causation, Anderson insists: "Correlation is enough."²⁰ But what happens when the number of links is also highly correlated with sexist, racist, and pornographic results?

The influence of racism, sexism, and colonialism is precisely what we see described in *Algorithms of Oppression*, information studies scholar Safiya Umoja Noble's study of the harmful stereotypes about Black and Latinx women perpetuated by search algorithms such as Google's. As discussed in chapter 1, Noble demonstrates that Google Search results do not simply correlate with our racist, sexist, and colonialist society; that society *causes* the racist and sexist results. More than that, Google Search reinforces these oppressive views by ranking results according to how many other sites link to them. The rank order, in turn, encourages users to continue to click on those same sites. Here, correlation without context is clearly not enough because it recirculates racism and sexism and perpetuates inequality.²¹

There's another reason that context is necessary for making sense of correlation, and it has to do with how racism, sexism, and other forces of oppression enter into the environments in which data are collected. The next example has to do with sexual assault and violence. If you do not want to read about these topics, you may want to skip ahead to the next section.

In April 1986, Jeanne Clery, a student at Lehigh University, was sexually assaulted and murdered in her dorm room. Her parents later found out that there had been thirty-eight violent crimes at Lehigh in the prior three years, but nobody had viewed that as important data that should be made available to parents or to the public. The Clerys mounted a campaign to improve data collection and communication efforts related to crimes on college campuses, and it was successful: the Jeanne Clery Act was passed in 1990, requiring all US colleges and universities to make on-campus crime statistics available to the public.²²

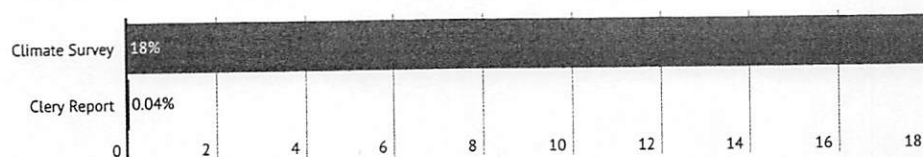
So we have an ostensibly comprehensive national dataset about an important public topic. In 2016, three students in Catherine's data journalism class at Emerson College—Patrick Torphy, Michaela Halnon, and Jillian Meehan—downloaded the Clery Act data and began to explore it, hoping to better understand the rape culture that has become pervasive on college campuses across the United States.²³ They soon became puzzled, however. Williams College, a small, wealthy liberal arts college in rural Massachusetts, seemed to have an epidemic of sexual assault, whereas Boston University (BU), a large research institution in the center of the city, seemed to have strikingly few cases relative to its size and population (not to mention that several high-profile sexual assault cases at BU had made the news in recent years).²⁴ The students were suspicious of these numbers, and investigated further. After comparing the Clery Act data with anonymous campus climate surveys (figure 6.4), consulting with experts, and interviewing survivors, they discovered, paradoxically, that the truth was closer to the *reverse* of the picture that the Clery Act data suggest. Many of the colleges with higher reported rates of sexual assault were actually places where more institutional resources were being devoted to support for survivors.²⁵

As for the colleges with lower numbers, this is also explained by context. The Clery Act requires colleges and universities to provide annual reports of sexual assault and other campus crimes, and there are stiff financial penalties for not reporting. But the numbers are self-reported, and there are also strong financial incentives for colleges *not* to report.²⁶ No college wants to tell the government—let alone parents of prospective students—that it has a high rate of sexual assault on campus. This is compounded by

Clery report data and anonymous survey results leave vastly different impressions of rape culture on college campuses.

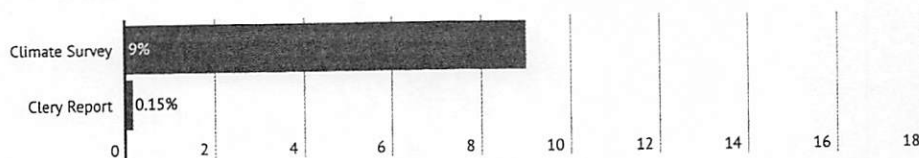


Boston University



Boston University surveyed its students in 2015, with a response rate of 22 percent. Nearly one in five respondents reported experiencing some type of sexual harassment or assault during their time at Boston University, compared to one in 2500 who reported assault in 2014.

Emerson College



Emerson College surveyed its students in 2015, with a 32 percent response rate. About one in 10 respondents said they experienced nonconsensual sexual contact on-campus during their time at Emerson, compared to one in 666 students that reported forcible sex offenses in 2014.

Figure 6.4

Data journalism students at Emerson College were skeptical of the self-reported Clery Act data and decided to compare the Clery Act results with anonymous campus climate survey results about nonconsensual sexual contact. Although there are data-quality issues with both data-sets, the students assert that if institutions are providing adequate support for survivors, then there will be less of a gap between the Clery-reported data and the proportion of students that report nonconsensual sexual conduct. Courtesy of Patrick Torphy, Michaela Halnon, and Jillian Meehan, 2016.

the fact that survivors of sexual assault often do not want to come forward—because of social stigma, the trauma of reliving their experience, or the resulting lack of social and psychological support. Mainstream culture has taught survivors that their experiences will not be treated with care and that they may in fact face more harm, blame, and trauma if they do come forward.²⁷

There are further power differentials reflected in the data when race and sexuality are taken into account. For example, in 2014, twenty-three students filed a complaint against Columbia University, alleging that Columbia was systematically mishandling cases of rape and sexual violence reported by LGBTQ students. Zoe Ridolfi-Starr, the lead student named in the complaint, told the *Daily Beast*, “We see complete lack of knowledge about the specific dynamics of sexual violence in the queer community, even from people who really should be trained in those issues.”²⁸

Simply stated, there are imbalances of power in the *data setting*—to use the phrase coined by Yanni Loukissas that we discussed in chapter 5—so we cannot take the numbers in the dataset at face value. Lacking this understanding of power in the collection environment and letting the numbers “speak for themselves” would tell a story that is not only patently false but could also be used to reward colleges that are systematically underreporting and creating hostile environments for survivors. Deliberately undercounting cases of sexual assault leads to being rewarded for underreporting. And the silence around sexual assault continues: the administration is silent, the campus culture is silent, the dataset is silent.²⁹

Raw Data, Cooked Data, Cooking

As demonstrated by the Emerson College students, one of the key analytical missteps of work that lets “the numbers speak for themselves” is the premise that data are a *raw input*. But as Lisa Gitelman and Virginia Jackson have memorably explained, data enter into research projects already fully cooked—the result of a complex set of social, political, and historical circumstances. “‘Raw data’ is an oxymoron,” they assert, just like “jumbo shrimp.”³⁰ But there is an emerging class of “data creatives” whose very existence is premised on their ability to *context-hop*—that is, their ability to creatively mine and combine data to produce new insights, as well as work across diverse domains. This group includes data scientists, data journalists, data artists and designers, researchers, and entrepreneurs—in short, pretty much everyone who works with data right now. They are the strangers in the dataset that we spoke of in chapter 5.

Data’s new creative class is highly rewarded for producing work that creates new value and insight from mining and combining conceptually unrelated datasets.

Examples include Google's now defunct Flu Trends project, which tried to geographically link people's web searches for flu symptoms to actual incidences of flu.³¹ Or a project of the *Sun Sentinel* newspaper, in Fort Lauderdale, Florida, which combined police license plate data with electronic toll records to prove that cops were systematically and dangerously speeding on Florida highways.³² Sometimes these acts of creative synthesis work out well; the *Sun Sentinel* won a Pulitzer for its reporting and a number of the speeding cops were fired. But sometimes the results are not quite as straightforward. Google Flu Trends worked well until it didn't, and subsequent research has shown that Google searches cannot be used as 1:1 signals for actual flu phenomena because they are susceptible to external factors, such as what the media is reporting about the flu.³³

Instead of taking data at face value and looking toward future insights, data scientists can first interrogate the context, limitations, and validity of the data under use. In other words, one feminist strategy for considering context is to consider the cooking process that produces "raw" data. As one example, computational social scientists Derek Ruths and Jürgen Pfeffer write about the limitations of using social media data for behavioral insights: Instagram data skews young because Instagram does; Reddit data contains far more comments by men than by women because Reddit's overall membership is majority men. They further show how research data acquired from those sources are shaped by sampling because companies like Reddit and Instagram employ proprietary methods to deliver their data to researchers, and those methods are never disclosed.³⁴ Related research by Devin Gaffney and J. Nathan Matias took on a popular corpus that claimed to contain "every publicly available Reddit comment."³⁵ Their work showed that the supposedly complete corpus is missing at least thirty-six million comments and twenty-eight million submissions.

Exploring and analyzing what is missing from a dataset is a powerful way to gain insight into the cooking process—of both the data and of the phenomenon it purports to represent. In some of Lauren's historical work, she looks at actual cooks as they are recorded (or not) in a corpus of thirty thousand letters written by Thomas Jefferson, as shown in figure 6.5.³⁶ Some may already know that Jefferson is considered the nation's "founding foodie."³⁷ But fewer know that he relied upon an enslaved kitchen staff to prepare his famous food.³⁸ In "The Image of Absence," Lauren used named-entity recognition, a natural language processing technique, to identify the places in Jefferson's personal correspondence where he named these people and then used social network analysis to approximate the extent of the relationships among them. The result is a visual representation of all of the work that Jefferson's enslaved staff put into preparing his meals but that he did not acknowledge—at least not directly—in the text of the letters themselves.

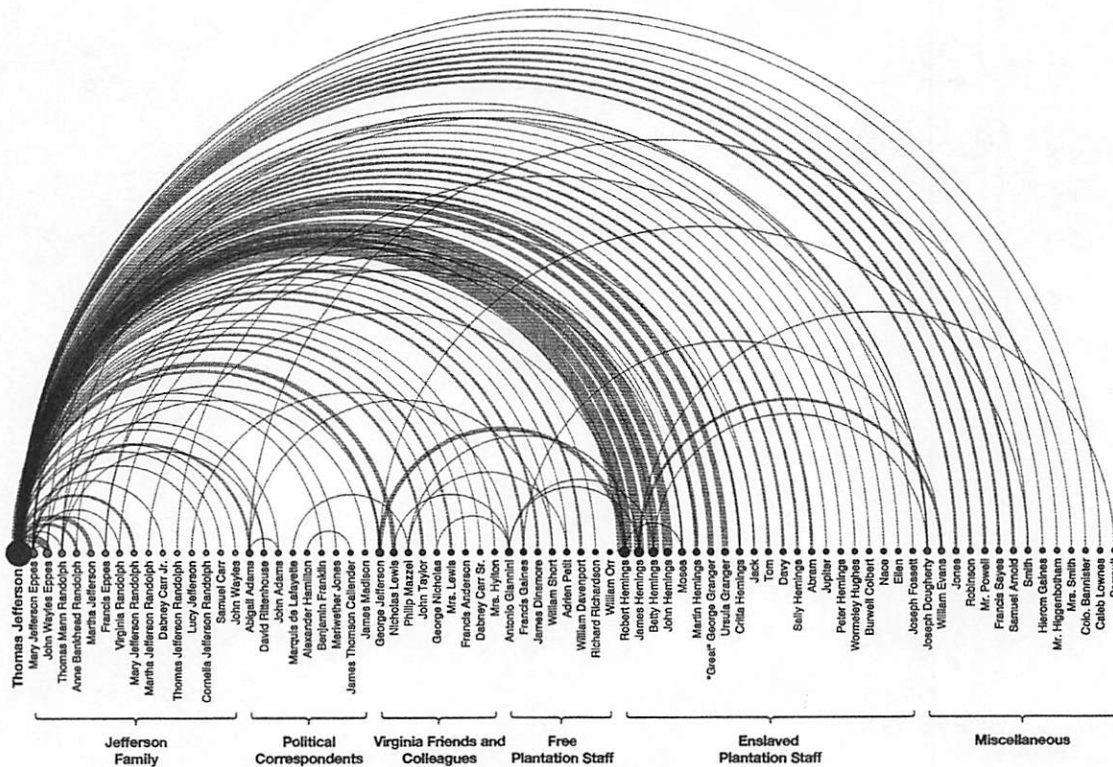


Figure 6.5

In "The Image of Absence" (2013), Lauren used machine learning techniques to identify the names of the people whom Thomas Jefferson mentioned in his personal correspondence and then visualized the relationships among them. The result demonstrates all of the work that his enslaved staff put into preparing Jefferson's meals but that was not directly acknowledged by Jefferson himself. Visualization by Lauren F. Klein.

On an even larger scale, computer scientists and historians at Stanford University used word embeddings—another machine learning technique—to explore gender and ethnic stereotypes across the span of the twentieth century.³⁹ Using several large datasets derived from sources such as the Google Books and the *New York Times*, the team showed how words like *intelligent*, *logical*, and *thoughtful* were strongly associated with men until the 1960s. Since that time, however, those words have steadily increased in association with women. The team attributed this phenomenon to the "women's movement in the 1960s and 1970s," making their work an interesting example of an attempt to quantify the impact of social movements. The paper is also notable for

openly acknowledging how their methods, which involved looking at the adjectives surrounding the words *man* and *woman*, limited the scope of their analysis to the gender binary. Furthermore, the researchers did not try to assert that the data represent how women and men “are,” nor did they try to “remove the bias” so that they could develop “unbiased” applications in other domains. They saw the data as what they are—cultural indicators of the changing face of patriarchy and racism—and interrogated them as such.

So, how do we produce more work like this—work that understands data as already “cooked” and then uses that data to expose structural bias? Unfortunately for Chris Anderson, the answer is that we need more theory, not less. Without theory, survey designers and data analysts must rely on their intuition, supported by “common sense” ideas about the things they are measuring and modeling. This reliance on “common sense” leads directly down the path to bias. Take the case of GDELT. Decades of research has demonstrated that events covered by the media are selected, framed, and shaped by what are called “news values”: values that confirm existing images and ideologies.⁴⁰ So what is it really that GDELT is measuring? What events are happening in the world, or what the major international news organizations are focusing their attention on? The latter might be the most powerful story embedded in the GDELT database. But it requires deep context and framing to draw it out.

Refusing to acknowledge context is a power play to avoid power. It’s a way to assert authoritativeness and mastery without being required to address the complexity of what the data actually represent: the political economy of the news in the case of GDELT, entrenched gender hierarchies and flawed reporting environments in the case of the Clery data, and so on. But deep context and computation are not incompatible. For example, SAFElab, a research lab at Columbia run by scholar and social worker Desmond Patton, uses artificial intelligence to examine the ways that youth of color navigate violence on and offline. He and a team of social work students use Twitter data to understand and prevent gang violence in Chicago. Their data are big, and they’re also complicated in ways that are both technical and social. The team is acutely aware of the history of law enforcement agencies using technology to surveil Black people, for example, and acknowledges that law enforcement continues to do so using Twitter itself. What’s more, when Patton started his research, he ran into an even more basic problem: “I didn’t know what young people were saying, period.”⁴¹ This was true even though Patton himself is Black, grew up in Chicago, and worked for years in many of these same neighborhoods. “It became really clear to me that we needed to take a deeper approach to social media data in particular, so that we could really grasp culture,

context and nuance, for the primary reason of not misinterpreting what's being said," he explains.⁴²

Patton's approach to incorporating culture, context, and nuance took the form of direct contact with and centering the perspectives of the youth whose behaviors his group sought to study. Patton and doctoral student William Frey hired formerly gang-involved youth to work on the project as domain experts. These experts coded and categorized a subset of the millions of tweets, then trained a team of social work students to take over the coding. The process was long and not without challenges. It required that Patton and Frey create a new "deep listening" method they call the *contextual analysis of social media* to help the student coders mitigate their own bias and get closer to the intended meaning of each tweet.⁴³ The step after that was to train a machine learning classifier to automatically label the tweets, so that the project could categorize all of the millions of tweets in the dataset. Says Patton, "We trained the algorithm to think like a young African American man on the south side of Chicago."⁴⁴

This approach illustrates how context can be integrated into an artificial intelligence project, and can be done with an attention to *subjugated knowledge*. This term describes the forms of knowledge that have been pushed out of mainstream institutions and the conversations they encourage. To explain this phenomenon, Patricia Hill Collins gives the example of how Black women have historically turned to "music, literature, daily conversations, and everyday behavior" as a result of being excluded from "white male-controlled social institutions."⁴⁵ These institutions include academia, or—for a recent example raised by sociologist Tressie McMillan Cottom—the op-ed section of the *New York Times*.⁴⁶ And because they circulate their knowledge in places outside of those mainstream institutions, that knowledge is not seen or recognized by those institutions: it becomes *subjugated*.

The idea of subjugated knowledge applies to other minoritized groups as well, including the Black men from Chicago whom Patton sought to understand. An approach that did not attend to this context would have resulted in significant errors. For example, a tweet like "aint kill yo mans & ion kno ya homie" would likely have been classified as aggressive or violent, reflecting its use of the word "kill." But drawing on the knowledge provided by the young Black men they hired for the project, Frey and Patton were able to show that many tweets like this one were references to song lyrics, in this case the Chicago rapper Lil Durk. In other words, these tweets are about sharing culture, not communicating threats.⁴⁷

In the case of SAFElab, as with all research projects that seek to make use of subjugated knowledge, there is also significant human, relational infrastructure required.

Frey and Patton have built long-term relationships with individuals and organizations in the community they study. Indeed, Frey lives and works in the community. In addition, both Frey and Patton are trained as social workers. This is reflected in their computational work, which remains guided by the social worker's code of ethics.⁴⁸ They are using AI to broker new forms of human understanding across power differentials, rather than using computation to replace human relationships. This kind of social innovation often goes underappreciated in the unicorn-wizard-genius model of data science. (For more on unicorns, see chapter 5.) As Patton says, "We had a lot of challenges with publishing papers in data science communities about this work, because it is very clear to me that they're slow to care about context. Not that they don't care, but they don't see the innovation or the social justice impact that the work can have."⁴⁹ Hopefully that will change in the future, as the work of SAFElab and others demonstrates the tremendous potential of combining social work and data science.

Communicating Context

It's not just in the stages of data acquisition or data analysis that context matters. Context also comes into play in the framing and communication of results. Let's imagine a scenario. In this case, you are a data journalist, and your editor has assigned you to create a graphic and short story about a recent research study: "Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service."⁵⁰ This study looks at the medical records of more than forty-five thousand first-time incarcerated people and finds that some groups are more likely to receive treatment, while others are more likely to receive punishment. More specifically, white people are more likely to receive a mental health diagnosis, while Black and Latinx people are more likely to be placed in solitary confinement. The researchers attribute some of this divergence to the differing diagnosis rates experienced by these groups before becoming incarcerated, but they also attribute some of the divergence to discrimination within the jail system. Either way, the racial and ethnic disparities are a product of structural racism.

Consider the difference between the two graphics shown in figure 6.6. The only variation is the title and framing of the chart.

Which one of these graphics would you create? Which one should you create? The first—Mental Health in Jail—represents the typical way that the results of a data analysis are communicated. The title *appears* to be neutral and free of bias. This is a graphic about rates of mental illness diagnosis of incarcerated people broken down by race and ethnicity. The people are referred to as *inmates*, the language that the study used. The

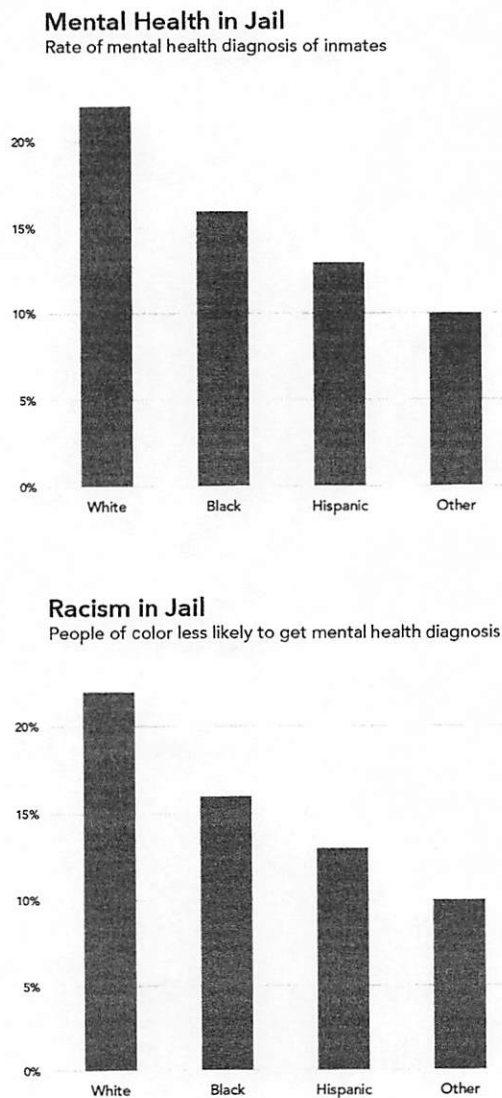


Figure 6.6

Two portrayals of the same data analysis. The data are from a study of people incarcerated for the first time in NYC jails between 2011 and 2013. Graphics by Catherine D'Ignazio. Data from Fatos Kaba et al., "Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service."

title does not mention race or ethnicity, or racism or health inequities, nor does the title point to what the data mean. But this is where additional questions about context come in. Are you representing only the four numbers that we see in the chart? Or are you representing the context from which they emerged?

The study that produced these numbers contains convincing evidence that we should distrust diagnosis numbers due to racial and ethnic discrimination. The first chart does not simply fail to communicate that but also actively undermines that main finding of the research. Moreover, the language used to refer to people in jail as *inmates* is dehumanizing, particularly in the context of the epidemic of mass incarceration in the United States.⁵¹ So, consider the second chart: Racism in Jail: People of Color Less Likely to Get Mental Health Diagnosis. This title offers a frame for how to interpret the numbers along the lines of the study from which they emerged. The research study was about racial disparities, so the title and content of this chart are about racial disparities. The people behind the numbers are *people*, not *inmates*. In addition, and crucially, the second chart names the forces of oppression that are at work: racism in prison.

Although naming racism may sound easy and obvious to some readers of this book, it is important to acknowledge that fields like journalism still adhere to conventions that resist such naming on the grounds that it is “bias” or “opinion.” John Daniszewski, an editor at the Associated Press, epitomizes this view: “In general our policy is to try to be neutral and precise and as accurate as we possibly can be for the given situation. We’re very cautious about throwing around accusations of our own that characterize something as being racist. We would try to say what was done, and allow the reader to make their own judgement.”⁵²

Daniszewski’s statement may sound democratic (“power to the reader!”), but it’s important to think about whose interests are served by making racism a matter of individual opinion. For many people, racism exists as a matter of fact, as we have discussed throughout this book. Its existence is supported by the overwhelming empirical evidence that documents instances of structural racism, including wealth gaps, wage gaps, and school segregation, as well as health inequities, as we have also discussed. Naming these structural forces may be the most effective way to communicate broad context. Moreover, as the data journalist in this scenario, it is your responsibility to connect the research question to the results and to the audience’s interpretation of the results. Letting the numbers speak for themselves is emphatically not more ethical or more democratic because it often leads to those numbers being misinterpreted or the results of the study being lost. Placing numbers in context and naming racism or sexism when

it is present in those numbers should be a requirement—not only for feminist data communication, but for data communication full stop.

This counsel—to name racism, sexism, or other forces of oppression when they are clearly present in the numbers—particularly applies to designers and data scientists from the dominant group with respect to the issue at hand. White people, including ourselves, the authors of this book, have a hard time naming and talking about racism. Men have a hard time naming and talking about sexism and patriarchy. Straight people have a hard time seeing and talking about homophobia and heteronormativity. If you are concerned with justice in data communication, or data science more generally, we suggest that you practice recognizing, naming, and talking about these structural forces of oppression.⁵³

But our work as hypothetical anti-oppression visualization designers is not over yet. We might have named racism as a structural force in our visualization, but there are still two problems with the “good” visualization, and they hinge on the wording of the subtitle: *People of Color Less Likely to Get Mental Health Diagnosis*. The first problem is that this is starting to look like a deficit narrative, which we discuss in chapter 2—a narrative that reduces a social group to negative stereotypes and fails to portray them with creativity and agency. The second issue is that by naming racism and then talking about people of color in the title, the graphic reinforces the idea that race is an issue for people of color only. If we care about righting the balance of power, the choice of words matters as much as the data under analysis. In an op-ed about the language used to describe low-income communities, health journalist Kimberly Seals Allers affirms this point: “We almost always use a language of deficiency, calling them disadvantaged, under-resourced and under-everything else. ... It ignores all the richness those communities and their young people possess: the wealth of resiliency, tenacity and grit that can turn into greatness if properly cultivated.”⁵⁴

So let's give it a third try, with the image in figure 6.7.

In this third version, we have retained the same title as the previous chart. But instead of focusing the subtitle on what minoritized groups lack, it focuses on the unfair advantages that are given to the dominant group. The subtitle now reads, *White People Get More Mental Health Services*. This avoids propagating a deficit narrative that reinforces negative associations and clichés. It also asserts that white people have a race, and that they derive an unfair advantage from that race in this case.⁵⁵ Finally, the title is proposing an interpretation of the numbers that is grounded in the context of the researchers' conclusions on health disparities.

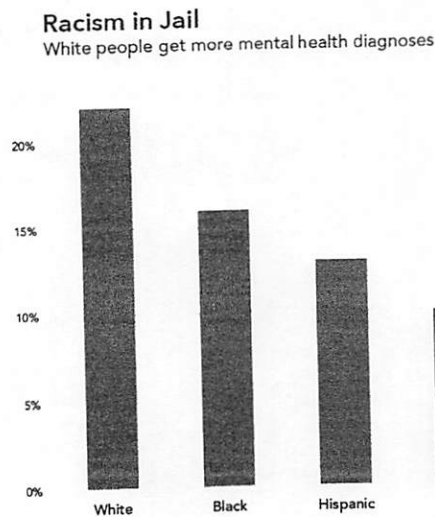


Figure 6.7

A third portrayal of the same data, with only the framing title and subtitle changed. *Source:* Data from Kaba et al., "Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service." Graphic by Catherine D'Ignazio. Data from Fatos Kaba et al., "Disparities in Mental Health."

Restoring Context

Three iterations on a single chart title might feel excessive, but it also helps to underscore the larger point that considering context always involves some combination of interest and time. Fortunately, there is a lot of energy around issues of context right now, and educators, journalists, librarians, computer scientists, and civic data publishers are starting to develop more robust tools and methods for keeping context attached to data so that it's easier to include in the end result.

For example, remember figure 6.3, that confusing chart of government procurements in São Paulo that we discussed earlier in this chapter? Gisele Craveiro, a professor at the University of São Paulo, has created a tool called *Cuidando do Meu Bairro* (Caring for My Neighborhood) to make that spending data more accessible to citizens by adding additional local context to the presentation of the information.⁵⁶ In the classroom, Heather Krause, a data scientist and educator, has developed the concept of the "data biography."⁵⁷ Prior to beginning the analysis process, Krause asks people working with

data, particularly journalists, to write a short history of a particular dataset and answer five basic questions: Where did it come from? Who collected it? When? How was it collected? Why was it collected? A related but slightly more technical proposal advocated by researchers at Microsoft is being called *datasheets for datasets*.⁵⁸ Inspired by the datasheets that accompany hardware components, computer scientist Timnit Gebru and colleagues advocate for data publishers to create short, three- to five-page documents that accompany datasets and outline how they were created and collected, what data might be missing, whether preprocessing was done, and how the dataset will be maintained, as well as a discussion of legal and ethical considerations such as whether the data collection process complies with privacy laws in the European Union.⁵⁹

Another emerging practice that attempts to better situate data in context is the development of *data user guides*.⁶⁰ Bob Gradeck, manager of the Western Pennsylvania Regional Data Center, started writing data user guides because he got the same questions over and over again about popular datasets he was managing, like property data and 311 resident reports in Pittsburgh. Reports Gradeck, "It took us some time to learn tips and tricks. ... I wanted to take the stuff that was in my head and put it out there with additional context, so other data users didn't have to do it from scratch."⁶¹ Data user guides are simple, written documents that each contain a narrative portrait of a dataset. They describe, among other things, the purpose and application of the data; the history, format, and standards; the organizational context; other analyses and stories that have used the dataset; and the limitations and ethical implications of the dataset. This is similar to the work that data journalists are doing to compile datasets and then make them available for reuse. For example, the Associated Press makes comprehensive national statistics about school segregation in the United States available for purchase.⁶² The spreadsheets are accompanied by a twenty-page narrative explainer about the data that includes limitations and sample story ideas.

These developments are exciting, but there is further to go with respect to issues of power and inequality that affect data collection environments. For example, professor of political science Valerie Hudson has worked for decades to trace the links between state security and the status of women. "I was interested in whether forms of oppression or subordination or violence against women were related to national, and perhaps international, instability and conflict," she explains. She and geographer Chad Emmett started the project WomanStats as a modest Excel spreadsheet in 2001. It has since grown to a large-scale web database with over a quarter of a million data points, including over 350 variables ranging from access to health care to the prevalence of rape to the division of domestic labor.⁶³

Notably, their sources are qualitative as well as quantitative. Says Hudson, "If you want to do research on women, you have to embrace qualitative data. There's no two ways about it, because the reality of women's lives is simply not captured in quantitative statistics. Absolutely not."⁶⁴ At the present, WomanStats includes two types of qualitative variables: practice variables are composed from women's reports of their lived experiences, and law variables are coded from the legal frameworks in a particular country. Indeed, the WomanStats codebook is a context nerd's dream that outlines measurement issues and warns about the incompleteness of its own data, especially with respect to difficult topics.⁶⁵ In regard to the data that records reports of rape, for example—a topic upsetting enough to even consider, let alone contemplate its scale and scope in an entire country—the codebook states: "CAVEAT EMPTOR! Users are warned that this scale only reflects reported rape rates, and for many, if not most, countries, this is a completely unreliable indicator of the actual prevalence of rape within a society!"⁶⁶ Instead of focusing on a single variable, users are directed to WomanStats's composite scales, like the Comprehensive Rape Scale, which look at reported prevalence in the context of laws, whether laws are enforced, reports from lived experience, strength of taboos in that environment, and so on.

So tools and methods for providing context are being developed and piloted. And WomanStats models how context can also include an analysis of unequal social power. But if we zoom out of project-level experiments, what remains murky is this: Which actors in the data ecosystem are responsible for providing context?

Is it the end users? In the case of the missing Reddit comments, we see how even the most highly educated among us fail to verify the basic claims of their data source. And datasheets for datasets and data user guides are great, but can we expect individual people and small teams to conduct an in-depth background research project while on a deadline and with a limited budget? This places unreasonable expectations and responsibility on newcomers and is likely to lead to further high-profile cases of errors and ethical breaches.

So is it the data publishers? In the case of GDELT, we saw how data publishers, in their quest for research funding, overstated their capabilities and didn't document the limitations of their data. The Reddit comments were a little different: the dataset was provided by an individual acting in good faith, but he did not verify—and probably did not have the resources to verify—his claim to completeness. In the case of the campus sexual assault data, it's the universities who are responsible for self-reporting, and they are governed by their own bottom line.⁶⁷ The government is under-resourced to verify and document all the limitations of the data.

Is it the data intermediaries? Intermediaries, who have also been called *infomediaries*, might include librarians, journalists, nonprofits, educators, and other public information professionals.⁶⁸ There are strong traditions of data curation and management in library science, and librarians are often the human face of databases for citizens and residents. But as media scholar Shannon Mattern points out, librarians are often left out of conversations about smart cities and civic technology.⁶⁹ Examples of well-curated, verified and contextualized data from journalism, like the Associated Press database on school segregation or other datasets available in ProPublica's data store, are also promising.⁷⁰ The nonprofit Measures for Justice provides comprehensive and contextualized data on criminal justice and incarceration rates in the United States.⁷¹ Some data intermediaries, like Civic Switchboard in Pittsburgh, are building their own local data ecosystems as a way of working toward sustainability and resilience.⁷² These intermediaries who clean and contextualize the data for public use have potential (and have fewer conflicts of interest), but sustained funding, significant capacity-building, and professional norms-setting would need to take place to do this at scale.

Houston, we have a public information problem. Until we invest as much in providing (and maintaining) context as we do in publishing data, we will end up with public information resources that are subpar at best and dangerous at worst. This ends up getting even more thorny as the sheer quantity of digital data complicates the verification, provenance, and contextualization work that archivists have traditionally undertaken. Context, and the informational infrastructure that it requires, should be a significant focus for open data advocates, philanthropic foundations, librarians, researchers, news organizations, and regulators in the future. Our data-driven lives depend on it.

Consider Context

The sixth principle of data feminism is to *consider context*. The bottom line for numbers is that they cannot speak for themselves. In fact, those of us who work with data must actively prevent numbers from speaking for themselves because when those numbers derive from a data setting influenced by differentials of power, or by misaligned collection incentives (read: pretty much all data settings), and especially when the numbers have to do with human beings or their behavior, then they run the risk not only of being arrogantly grandiose and empirically wrong, but also of doing real harm in their reinforcement of an unjust status quo.

The way through this predicament is by considering context, a process that includes understanding the provenance and environment from which the data was collected,

as well as working hard to frame context in data communication (i.e., the numbers should not speak for themselves in charts any more than they should in spreadsheets). It also includes analyzing social power in relation to the data setting. Which power imbalances have led to silences in the dataset or data that is missing altogether? Who has conflicts of interest that prevent them from being fully transparent about their data? Whose knowledge about an issue has been subjugated, and how might we begin to recuperate it? The energy around context, metadata, and provenance is impressive, but until we fund context, then excellent contextual work will remain the exception rather than the norm.