Data Feminism

Catherine D'Ignazio Lauren Klein

2019

Contents

Introduction

It was 1967 when Christine Mann Darden first passed through the gates of NASA's Langley Research Center, in Hampton, Virginia. In the city of Hampton, and across the United States, tensions were running high. In Los Angeles that June, a massive protest against the Vietnam War ended in violence when over a thousand armed police officers attacked the peaceful protestors, sparking national outrage. One month later, in July of that year, even more violence engulfed the city of Detroit after a police raid gone awry. The 1967 Detroit Riot, or the 1967 Detroit Rebellion, as it is increasingly known, resulted in over 40 deaths and 1000 injuries--most of which were sustained by the city's predominantly Black residents. The rebellion ended only after the governor, with the support of the President of the United States, called in the Army's Airborne Division and the Michigan National Guard.

The gates of Langley might have shielded Darden from those physical confrontations, but her work there was no less politically engaged. 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 U.S. had suffered a series of defeats--and, in January of that year, an honest-to-goodness tragedy, when a fire during a launch test of the *Apollo 1* spacecraft killed all three astronauts on board.

While the nation was in mourning, everyone at NASA threw themselves back into their work– including Darden, who held a master's degree in applied math and was employed as a data analyst at the time. Two years later, it would be Darden's point-perfect 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 was far from accomplished.

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 Hampton Institute, the historically Black college that

she attended for her undergraduate studies. The Hampton sit-ins were the first in the state of Virginia, and contributed significantly to the dismantling of the Jim Crow era policies of segregation that were still in place at the time. By the time that Darden joined NASA, its Virginia facility had been fully desegregated for several years. But it had yet to reckon with another issue of equality of opportunity: the status of its human computers.

Darden's arrival at Langley coincided with the early days of digital computing. While 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 instead. These computers were all women, trained in math but treated like secretaries. 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 male engineers never even bothered to learn the female computers' names.

But Darden had a sense of justice in addition to her advanced degree. So after several years of working as a computer, she decided to ask her boss why men with her credentials were placed in engineering positions, where they could be promoted through the ranks of the civil service, while women like herself were sent to the computing pools, where they languished until they retired or quit. As Darden, now 75, told Margot Lee Shetterly, who interviewed Darden for her book, *Hidden Figures The American Dream and the Untold Story of the Black Women Who Helped Win the Space Race*, her boss's response was sobering: "Well, nobody's ever complained," he told her. "The women seem to be happy doing that, so that's just what they do."[1]

Today, a response like that would get a boss fired (or, at the least, served with a Title IX complaint). But at the time, stereotypical remarks about "what women do" were par for the course. In fact, 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."[2] 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 being said, Darden's experience as a Black woman with a fulltime job was quite different than that of the white suburban housewife--the presumed

[1] The concepts taught address specific mathematical content and skills outlined by the Common Core State Standards in New York.

[2] CS109 at Harvard is taught jointly by Computer Science and Statistics. As of this writing, there are 37 male faculty (69%) and 17 female faculty (31%).

audience of *The Feminine Mystique*. And when critics called out Friedan, rightly, for failing to acknowledge how no single person could claim to speak on behalf of all women, everywhere, it was women like Darden, among all sorts of different folks, whom they had in mind. In *Feminist Theory: From Margin to Center*, a pioneering Black feminist text, 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."

In other words, Friedan had failed to consider how additional factors like race and class, not to mention sexuality, ability, age, religion, and geography, among many others, intersect with each other in order to determine any particular woman's personal experience in the world. Although this concept did not have a name when hooks described it--the term *intersectionality* would be coined by legal theorist Kimberlé Crenshaw in the late 1980s--its necessity was already clearly borne out by, well, everything. In the face of the racism embedded into U.S. culture, coupled with the many other forms of oppression experienced by minoritized groups, it would be impossible to claim a common experience--or a common movement--for all women, everywhere. Instead, what was needed was what the Combahee River Collective, the famed Black lesbian activist group out of Boston, advocated for back in 1977: "the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking."

We'll have more to say about the importance of intersectionality in a few pages, but first let's find out what happened with Christine Darden and her goal of becoming an aerospace engineer. As Shetterly tells it, Darden heard nothing from her boss but radio silence. But two weeks later, she was indeed promoted and transferred to an aerospace engineering group. Darden would go on to conduct ground-breaking 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," she recalls.

For over a decade, this research required that Darden work overtime in the office as well as at home. But she could tell that her scientific accomplishments were still not being recognized at the same level as her male colleagues. Once again, data analysis opened doors for Darden. But this time, Darden wasn't responsible for the math. Instead, her technical expertise provided a key datapoint for a larger advocacy project.

Over in Langley's Equal Opportunity office, a white woman by the name of Gloria Champine had been compiling a set of internal statistics about gender

and rank. The data showed that men and women with identical academic credentials, publication records, and performance reviews, were still promoted at vastly different rates. Champine then visualized the data—in the form of a bar chart—and presented her findings to the head of her Directorate. He was "shocked at the disparity," Shetterly reports, and Darden received the promotion she had long deserved.

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 the head of a Directorate herself.

Figure 1: Christine Darden in the control room of the Unitary Plan Wind Tunnel at NASA's Langley Research Center in 1975. ¶ Credit: NASA. ¶ Source: Wikipedia, https://en.wikipedia.org/wiki/Christine_Darden#/media/File:Christine_Darden.jpg

---

Christine Darden's rise into the leadership ranks at NASA was largely the result of her own knowledge, experience, and grit. But Darden's story is one we can only tell as a result of the past several decades of feminist activism and critical thought. For it was a national feminist movement that brought issues of women in the workplace to the forefront of U.S. cultural politics, and it was a local feminist advocate in the form of Gloria Champine who enabled Darden

to continue her own professional rise. It was also, presumably, the work of many unnamed colleagues and friends, who may or may not have considered themselves feminists, who provided Darden with community and support--and likely a significant number of casseroles--as she ascended the ranks of NASA. And it was the work of feminist scholars and activists that allows us to recognize that labor, emotional as much as physical, as such today.

Now it's probably time to clarify the relationship between the dictionary definition of a feminist and what is described, more generally, as *feminist thought*. The Merriam-Webster dictionary (and also, for the record, Beyoncé) define a feminist as "a person who believes in the political, social, and economic equality of the sexes."[3] Feminist thought has its basis in this theory of the equality of the sexes, but it is much more expansive. It includes the work of activists like Champine, or bell hooks, or--however problematically-- Betty Friedan, who have taken direct action to achieve the equality of the sexes. It also includes the work of scholars and cultural critics--again like hooks, or like Kimberlé Crenshaw, or like Margot Lee Shetterly--who have explored the social, political, historical, and conceptual reasons behind the inequality of the sexes that we face today.

In the process, these scholars and activists have given voice to many of the additional ways in which the status quo is unjust. These include the power differentials between men and women, as well as those between--for instance--white women and Black women, academic researchers and indigenous communities, and people in the Global North and the Global South. These feminist thinkers arrived at their emphasis on power, rather than gender alone, because of their insistence on intersectionality, the concept we started to get at just a few pages ago. So let's get a little more specific in our explanation of intersectionality. The concept doesn't simply describe the intersecting aspects of any particular person's identity that shape their experience in the world. Rather, it describes the intersecting systems of power--the systems of privilege, on the one hand; and systems of oppression, on the other--that determine that particular person's experiences. When you stop to think about it, many people experience at least a little of both.

In fact, it was an example of how oppression and privilege themselves intersect that prompted Crenshaw to name the concept that she'd seen play out over the course of her legal career. In law school, Crenshaw came across the anti-discrimination case of *DeGraffenreid v. General Motors*. Emma DeGraffenreid was a Black working mom who had sought a job at a General Motors factory in her town. She was not hired. The factory did have a history of hiring Black people: many Black men worked in industrial and maintenance jobs there. They also 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

[3]This does not mean there are no data ethics courses, only that it is not the norm to address these concerns in introductory coursework. Indeed, there is a long list compiled by Dr. Casey Fiesler of technical courses that specifically address ethics and what is being called "fairness, accountability and transparency" in technical fields: http://bit.ly/tech-ethics-syllabi

throw out the case. The company did hire Black people and did hire women, so it could not be discriminating on the basis of race or gender. But what about discrimination on the basis of race and gender together, Crenshaw wanted to know? This was something different, it was real, and it needed to be named.

Intersectionality helps us name and recognize the interaction between categories of social difference, such as race and gender. It helps us see when people who embody two or more of those characteristics fall through the cracks, because they are doubly or triply marginalized. It also helps us unmask the privilege that comes with embodying the dominant dimensions of identity, and helps us understand how oppression and privilege can co-exist in the same body. For example, a white, gay, disabled, cisgendered man might reap the benefits of privilege for his race and gender, but experience oppression for his sexual orientation and disability. A straight, college-educated, cisgendered Muslim woman might experience certain privilege on account of her sexual orientation and level of education, but experience oppression on account of her gender and religion. The intersection of categories of social difference, and of the forces of privilege and oppression that are bound up in them, are what Crenshaw's term names.

In the case of Christine Darden and her promotion, Gloria Champine was primarily concerned with the issue of gender. But she was also, like Crenshaw, intent on exposing a larger system of power and privilege. She knew that unless she confronted the systematic nature of the discrimination faced by women at NASA, she would continue to hear from individuals like Darden for the rest of her career. Her goal was to implement changes that would improve the lives of all women at NASA, and to achieve that goal, she required a complete picture--in the form of her bar chart-- of the problem at hand.

Of course, when Champine created her bar chart, she also recognized in the 1980s what many of us are only now beginning to understand: that data visualizations, and the data science that underlies them, hold tremendous rhetorical force. Now, as then, a single data visualization can dazzle, inform, and persuade. Champine aligned her goal of challenging the systemic nature of the gender discrimination that plagued NASA at the time with the rhetorical power of a bar chart. She asked herself: "What is the source of the problem that my colleague is facing? What information do I need in order to bring this problem to light? And what is the format through which I can best advocate on her behalf, and effect structural change?" In doing so, Champine joined with Darden to model a key aspect of what we call in this book *data feminism*: a way of thinking about data and its communication that is informed by direct experience, by a commitment to action, and by the ideas associated with intersectional feminist thought.

Data feminism can show us how images like Champine's bar chart might seem neutral and objective, but are in fact the result of very human and necessarily imperfect design processes. Data feminism can also show us how the categories of data collection matter deeply, especially when dividing people into groups. Because Champine was able to collect data on both gender and rank, she was

able to show the extent of the gender discrimination that was taking place at NASA at the time. But because she did not include any additional demographic information in her report--whether by circumstance, or by design--she was unable to show the effects of any additional forms discrimination that might also have been then taking place. The alliance between Champine and Darden, and the chart they together produced, also helps underscore the importance of listening to and learning from the data's stewards: the people who serve as the source of knowledge about the issues the data purports to represent. Champine knew to crunch the numbers only because Darden shared her personal experience of gender discrimination with her. Without Darden's first-hand knowledge of the problem, Champine might never have known that action was necessary.

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 and we now have computers in the form of our phones that fit in our pockets, and-- in the case of the iPhone 6-- can operate 120 million times faster than a standard IBM System/360. We've also witnessed an equally remarkable growth in our capacity to collect information in digital form--and in the capacity to have data collected about us.

As it turns out, the IBM System/360 line was viewed as a major milestone in the history of data processing as well as rocket science. It was the first family of machines that could be scaled up to be used for aerospace-level operations--and also scaled down to be used by a single data analyst in, for instance, an insurance company or a bank. But those banks and insurance companies today? They now collect data on our purchase histories and online behaviors, the times of day we're most active on Facebook and the number of items we add to our Amazon cart. Our most trivial everyday actions – taking a single step, searching for a way around traffic, or liking a friend's cat video – are now hot commodities. Not because our friends' cats are exceptionally cute (they are cute, of course, but not exceptional), or because our step counts are exceptionally exciting (they are emphatically not), but because those tiny actions can be combined with other tiny actions in order to determine, for instance, whether we're in a liking kind of mood, whether we tend to click on links when we're also liking videos, whether we might also happen to be frustrated with our daily commute, and whether today might be the day for a well-placed ad for new sneakers, or a coupon for 20% off. An alternative to the daily grind and a way to increase our step counts in one fell (Nike) swoop.

This is the data economy. And corporations and governments, often aided by scholars and researchers, are scrambling to see what consumer behaviors remain untapped and unrefined. Nothing is safe from datafication, the process of turning phenomena in the world into digital information. Not your cat, or your aspirational exercise goals, or-- more realistically-- the butt you are currently using to sit in your seat. Shigeomi Koshimizu, a Tokyo-based professor of engineering, has been designing matrices of sensors that collect data at 360

different positions around your rear end while it's smushed in a chair. Those data are then analyzed by custom software that detects micropatterns in weight and pressure. The result is a data profile of your butt that is, according to Koshimizo's research, as unique as your fingerprints. In the future, he suggests, our cars could be outfitted with butt-scanners instead of keys or car alarms. If Catherine sits down in her car of the future--self-driving, of course--this technology would scan her butt, crunch the data, and welcome her with a warm hello. But if Lauren sat in Catherine's car, it would refuse to move--or it might even be programmed to call the police.

While this redefinition of butt-dialing may still be a few years away, the datafication of our everyday lives is already a reality--and not only when we're actively clicking. Decisions of social and civic importance, ranging from which products to stock in the grocery store before a hurricane, to which city buildings to inspect for the risk of fire, to which citizens to tag as pre-trial flight risks, are increasingly being made by automated systems sifting through large amounts of data. For example, Walmart's predictive analytics team has long combined consumer purchasing patterns with weather data in order to address the first of these scenarios. If a hurricane is coming, what goods should Walmart get on the shelves quickly? It turns out that obvious items like flashlights and generators are in high demand. But Walmart also always sends truckloads of strawberry Pop-Tarts to areas where there are hurricane warnings, because their data analysis detected spikes in Pop-Tart sales during episodes of severe weather. As VP of Information Systems Dan Phillips told *Fortune* magazine, "They are preserved until you open them, the whole family can eat them, and they taste good."

There are similar examples of data-driven decision making in place in the government sector. In a widely-cited example, a stats team employed by the City of New York helped to integrate data analysis into the building inspection process. In New York, there are upwards of 25,000 complaints per year about buildings that have been illegally converted into apartments. But there are only around two hundred inspectors on city payroll who can handle the complaints. These illegal apartments often pose fire hazards, and in fact, many firefighters have lost their lives trying to save the residents who live there. So the stats team brought together several datasets relating to property tax delinquency (an indicator of neglect), rat complaints (ditto), arrest locations (a proxy for poverty), and more, in order to rank the 25,000 complaints by fire risk. The inspectors began with the buildings that were determined to hold the highest risk, and postponed their inspections of lower-risk sites. To their surprise, the inspectors issued five times more "vacate orders" than they had without the data-assisted ranking system. They might have been responsible for causing short-term inconvenience on the part of the people who were required to move, but they were able to reduce the longer-term risk of fire and potential loss of life. Here, the power of data was wielded for civic good: to allocate scarce resources to address issues of public health and safety.

But data-driven decision-making can be just as easily used to amplify the inequities already entrenched in public life. In *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Cathy O'Neill has drawn attention to how predictive policing models use data about property tax delinquency and arrest locations, the very same datasets used for good by the City of New York's stats team, in order to determine which neighborhoods to patrol more heavily, and which neighborhoods to leave alone. Like city building inspectors, it turns out that police are also in short supply. But in the case of policing, data analysis does not always save lives. Because more police are sent into neighborhoods that have had more crimes already reported, the police are more likely to be there when a new crime is committed (or, as in most cases, simply suspected of taking place). And because the police are already there, the people involved--usually poor, and usually people of color-- are more likely to get ticketed, arrested, or even killed. This creates what O'Neill calls a "pernicious feedback loop," amplifying the effects of the already pernicious criminalization of poverty that takes place in the United States. Meanwhile, in more affluent neighborhoods, the very same petty crimes--like jaywalking or littering, for instance-- are far less likely to be prosecuted because the police simply aren't there to see those crimes take place. This disparity in law enforcement is what led to the creation of White Collar Crime Risk Zones, a satirical map of all of the white collar crime that goes uninvestigated because neighborhoods of color are so overpoliced.

The double-edged sword of data shows just how important it is to understand how structures of power and privilege operate in the world. The questions we might ask about these structures can relate to issues of gender in the workplace, as in the case of Christine Darden and her wrongly delayed promotion. Or they can relate to issues of broader social inequality, as in the case of predictive policing described just above. So one thing you will notice throughout this book is that not all of our examples are about women--and deliberately so. This is because data feminism is about more than women. It's is about more than gender. Put simply: *Data Feminism* is a book about power in data science. Because feminism, ultimately, is about power too. It is about who has power and who doesn't, about the consequences of those power differentials, and how those power differentials can be challenged and changed.

This paragraph deserves re-stating, because we want you to remember these points as you read this book:

- **Data feminism isn't only about women.**

  It takes more than one gender to have gender inequality; and more than one gender to work towards justice.

- **Data feminism isn't only for women.**

  Men, non-binary, and genderqueer people are proud to call themselves feminists and use feminist thought in their work.

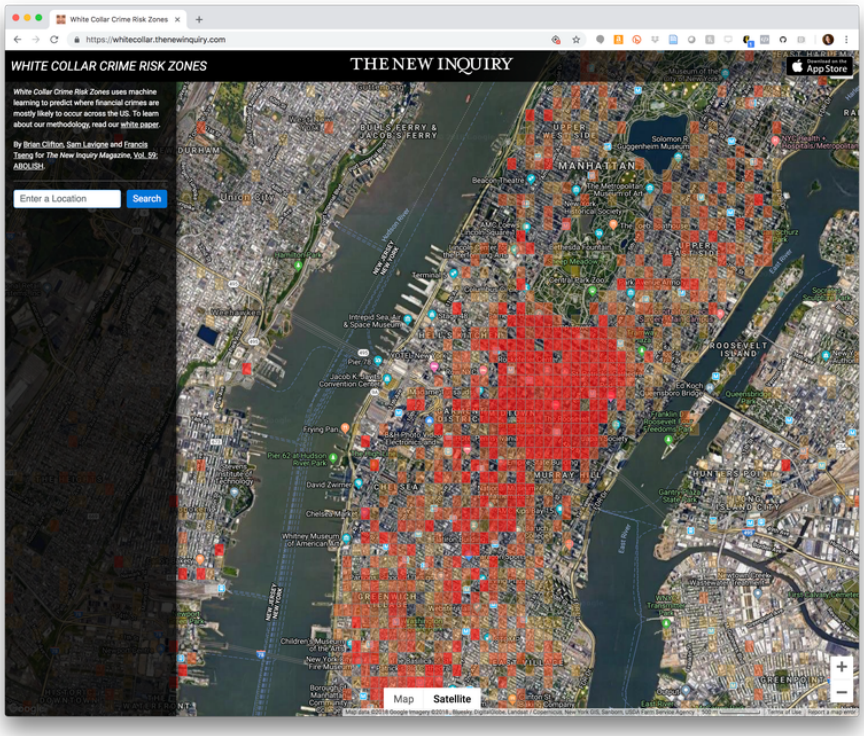


Figure 2: White Collar Crime Risk Zones uses machine learning to predict where financial crimes by white people may occur. ¶ Credit: Brian Clifton, Sam Lavigne, and Francis Tseng for *The New Inquiry Magazine*, Vol. 59: ABOLISH. ¶ Source: (https://whitecollar.thenewinquiry.com/

- **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. In this book, we choose to foreground many examples where racism and patriarchy intersect. This reflects our location in the United States, where the most entrenched issues of injustice have racism at their source.

- **Feminism *is* about power - who has it and who doesn't.**

```
And in our contemporary world, data *is* *power*. Which is why we
wrote this book.
```

While the feminisms of the 19th and 20th centuries accomplished a great deal, feminism remains an unfinished project, and an urgent one. Consider that, while Darden's story ended well, her accomplishments are just beginning to be recognized--largely due to the work of a single person, Margot Lee Shetterly, and her efforts to shine a light on the "hidden figures" of the history of computing. Or, consider that, in 1996, the field of computing held a celebration for the fiftieth anniversary of the first electronic computer, the ENIAC. But the organizers of the event did not think to invite most of the computer's original programmers to the party! These programmers were all women, of course. So is it a surprise that the number of female graduates of information sciences programs continues to decline? In 2013, the numbers were bleak: the field was only 26% women, the same percentage as in 1974.

On a larger cultural scale, Kimberlé Crenshaw and her colleagues have started a campaign called #SayHerName to call attention to police brutality against Black women, whose stories of racialized and gendered violence are so often left out of public conversations. And yet, the US federal government still has no comprehensive database on people killed by police officers. The #MeToo movement has demonstrated the pervasiveness of sexual assault as well as the bravery of women from all backgrounds to come forward. And yet, the U.S. has placed accused sexual predators in the White House and nominated them to the Supreme Court, all while their elite white male colleagues rally angrily behind their innocence. Feminism is unfinished and urgent work, in data and technology as well as in our most powerful political institutions.

In the chapters that follow, we draw from a wide range of examples of feminist data science in order to show how we can take steps towards a more just and equal world. In the examples we discuss, we are guided by our values around intersectionality, equity, and proximity, which we outline in Appendix 1. We do so in order to challenge a set of widely held assumptions, like the idea that "the numbers speak for themselves," and to explore projects that expand our ideas about what constitutes "data" in the first place. We call attention to the people and their bodies who are typically included in the data collection process, as well as to the people and their bodies who are typically left out. We question whose work gets recognized, and whose research questions should matter most.

Along the way, we introduce you to some of the analysts, designers, journalists, scholars, and teachers who are already doing the transformative work we hope to see more of in the world. There are a lot of data feminists out there! You just might not know about them until you read this book.

Chapter 1

Bring Back the Bodies

When Serena Williams disappeared from Instagram in early September, 2017, her six million followers thought they knew what had happened. Several months earlier, in March of that year, Williams had accidentally announced her pregnancy to the world via a bathing suit selfie and a caption that was hard to misinterpret: "20 weeks." Now, they assumed, her baby had finally arrived.

But then they waited, and waited some more. Two weeks later, Williams finally re-appeared on Instagram, announcing the birth of her daughter and inviting her followers to watch a video that welcomed Alexis Olympia Ohanian Jr. to the world. A montage of baby bump pics interspersed with clips of a pregnant Williams playing tennis and cute conversations with her husband, Reddit cofounder Alexis Ohanian, segued into the shot that her fans had been waiting for: the first of baby Olympia. Williams was narrating: "So we're leaving the hospital," she explains. "It's been a long time. We had a lot of complications. But look who we got!" The scene fades to white, and ends with a set of stats: Olympia's date of birth, birth weight, and number of grand slam titles: 1. (Williams, as it turned out, was already eight weeks pregnant when she won the Australian Open earlier that year).

Williams's Instagram followers were, for the most part, enchanted. But a fair number of her followers – many of them Black women like Williams herself– fixated on the comment she'd made as she was heading home from the hospital with her baby girl. Those "complications" that Williams mentioned – they'd had them too.

On Williams's Instagram feed, the evidence was anecdotal –women posting about their own experience of childbirth gone horribly wrong. But a few months later, Williams returned to social media –Facebook, this time –armed with data. Citing a 2017 study from the US Centers for Disease Control and Prevention (CDC), Williams wrote that "Black women are over 3 times more likely than white women to die from pregnancy- or childbirth-related causes."

While these disparities were well known to Black women-led reproductive justice groups like Sister Song, the Black Mamas Matter Alliance, and Raising Our Sisters Everywhere, Williams helped to shine a national spotlight on them. And she wasn't the only one. A few months earlier, Nina Martin of the investigative journalism outfit ProPublica, working with Renee Montagne of NPR, had reported on the same phenomenon. "Nothing Protects Black Women From Dying in Pregnancy and Childbirth," the headline read. In addition to the study also cited by Williams, Martin and Montagne cited a second study from 2016 which showed that neither education nor income level – the factors usually invoked when attempting to account for healthcare outcomes that diverge along racial lines– impacted the fates of Black women giving birth. On the contrary, the data showed that Black women with college degrees suffered more severe complications of pregnancy and childbirth than white women who had never even graduated from high school.

But what were these complications, more precisely? And how many women had actually died as a result? ProPublica couldn't find out, and neither could *USA Today*, which took up the issue a year later to see what, after a year of increased attention and advocacy, had changed. What they found was that there was still no national system for tracking complications sustained in pregnancy and childbirth, even as similar systems have long been in place for tracking things like, for instance, teen pregnancy, hip replacements, and heart attacks. They also found that there is also still no reporting mechanism for ensuring that hospitals follow national childbirth safety standards, as is required for both hip surgery and cardiac care. "Our maternal data is embarrassing," stated Stacie Geller, a professor obstetrics and gynecology at the University of Illinois, when asked for comment. The Chief of the CDC's Maternal and Infant Health Branch, William

Callaghan, makes the significance of this "embarrassing" data more clear: "What we choose to measure is a statement of what we value in health," he explains. We might edit his statement to add: it's a measure of *who* we value in health, too

---

The lack of data about maternal health outcomes, and its impact on matters of life and death, underscores how it is *people* who end up affected by the choices we make in our practices of data collection, analysis, and communication. More than that, it's almost always the bodies of those who have been disempowered by forces they cannot control, such as sexism, racism, or classism –or, more likely, some combination of all three –who experience the most severe consequences of these choices. Serena Williams acknowledged this exact phenomenon when asked by *Glamour* magazine about the statistics she cited in her Facebook post. "If I wasn't who I am, it could have been me—" she said, referring to the fact that she had to demand that her medical team perform additional tests in order to diagnose her own postnatal complications, and because she was Serena Williams, 23-time grand slam champion, they listened. But, she told *Glamour*, "that's not fair."

It is absolutely not fair. But without a significant intervention into our current data practices, this unfairness –and many other inequities with issues of power and privilege at their core – will continue to get worse. Stopping that downward spiral is the real reason we wrote this book. We wrote this book because we are data scientists and data feminists. We think that data science and the fields that rely upon it stand a lot to learn from feminist writing, thinking, scholarship, and action[1]. As we explain in *Why Data Science Needs Feminism*, feminism isn't only about women. It isn't even only about issues of gender. Feminism is about power –about who has it, and who doesn't. In a world in which data is power, and that power is wielded unequally, feminism can help us better understand how it operates and how it can be challenged. As data feminists –a group that includes women, men, non-binary and genderqueer people, and everyone else –we can take steps, together, towards a more just and equal world.

A good starting point is to understand how power operates on bodies and through them. "But!" you might say. "Data science is premised on things like objectivity and neutrality ! And those things have nothing to do with bodies!" But that is precisely the point. Data science, as it is generally understood in the world today, has very little to do with bodies. But that is a fundamental misconception about the field, and about data more generally. Because even though we don't see the bodies that data science is reliant upon, it most certainly relies upon them. It relies upon them as the sources of data, and it relies upon them to make decisions about data. As we discuss more in depth in a couple of pages, it even relies on them to decide what concepts like "objective" and "neutral" really

[1]Feminism is one key conceptual orientation that can help mitigate inequality and work towards justice, but it is not the only one. We talk about some others in *Now Let's Multiply.*

mean. And when not all bodies are represented in those decisions – as in the case of the federal and state legislatures which might fund data collection on maternal mortality –well, that's when problems enter in.

What kind of problems? Structural ones. Structural problems refer to problems that are systemic in nature, rather than due to a specific point (or person) of origin. It might be counterintuitive to think that individual bodies can help expose structural problems, but that's precisely what the past several decades –centuries, even –of feminist activism and critical thought has allowed us to see. Because many of the problems that individual people face are often the result of larger systems of power, but they remain invisible until those people bring them to light. In a contemporary context, we might easily cite the #MeToo movement as an example of how individual experience, taken together, reveals a larger structural problem of sexual harassment and assault. We might also cite the fact that the movement's founder was a Black woman, T arana Burke, whose contributions have largely been overshadowed by the more famous white women who joined in only after the initial –and therefore most dangerous –work had already taken place.

Burke's erasure from the #MeToo movement is only one datapoint in a long line of Black women who have stood on the vanguard of feminist advocacy work, only to have their contributions subsumed by white feminists after the fact. This is a structural problem too. It's the result of several intersecting differentials of power –differentials of power that must be made visible and acknowledged before they can be challenged and changed.

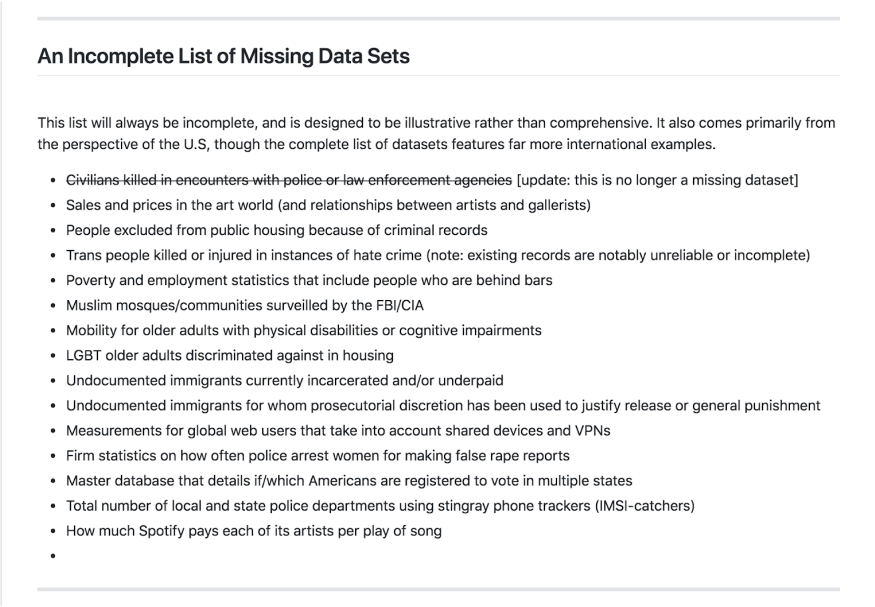
To be clear, there are already a significant number of data scientists, designers, policymakers, educators, and journalists, among others, who share our goal of using data to challenge inequality and help change the world. These include the educators who are introducing data science students to real-world problems in health, economic development, the environment, and more, as part of the Data Science for Social Good initiative; the growing number of organizations like DataKind, Tactical Tech, and the Engine Room, that are working to strengthen the capacity of the civil sector to work with data; newsrooms like ProPublica and the Markup that use data to hold Big Tech accountable; and public information startups like MuckRock, which streamlines public records requests into reusable databases Even a commercial design firm, Periscopic, has chosen the tagline, "Do Good With Data." We agree that data can do good in the world. But we can do only do good with data if we acknowledge the inequalities that are embedded in the data practices that we ourselves rely upon. And this is where the bodies come back in.

In the rest of this chapter, we explain how it's people and their bodies who are missing from our current data practices. Bodies are missing from the data we collect; bodies are extracted into corporate databases; and bodies are absent from the field of data science. Even more, it's the bodies with the most power that are ever present, albeit invisibly, in the products of data science. Each of these is a problem, because without these bodies present in the field of data

science, the power differentials currently embedded in the field will continue to spread. It's by bringing back these bodies –into discussions about data collection, about the goals of our work, and about the decisions we make along the way –that a new approach to data science, one we call *data feminism*, begins to come into view.

Bodies uncounted, undercounted, silenced

One person already attuned to certain things missing from data science, and to the power differentials responsible for those gaps, is artist, designer, and educator Mimi Onuoha. Her project, *Missing Data Sets*, is a list of precisely that: descriptions of data sets that you would expect to already exist in the world, because they describe urgent social issues and unmet social needs, but in reality, do not. These include "People excluded from public housing because of criminal records," "Mobility for older adults with physical disabilities or cognitive impairments," and "Measurements for global web users that take into account shared devices and VPNs." These data sets are missing for a number of reasons, Onuoha explains in her artist statement, many relating to issues of power. By compiling a list of the data that are missing from our "otherwise data-saturated" world, she states, we can "reveal our hidden social biases and indifferences."



An Incomplete List of Missing Data Sets

This list will always be incomplete, and is designed to be illustrative rather than comprehensive. It also comes primarily from the perspective of the U.S, though the complete list of datasets features far more international examples.

- ~~Civilians killed in encounters with police or law enforcement agencies~~ [update: this is no longer a missing dataset]
- Sales and prices in the art world (and relationships between artists and gallerists)
- People excluded from public housing because of criminal records
- Trans people killed or injured in instances of hate crime (note: existing records are notably unreliable or incomplete)
- Poverty and employment statistics that include people who are behind bars
- Muslim mosques/communities surveilled by the FBI/CIA
- Mobility for older adults with physical disabilities or cognitive impairments
- LGBT older adults discriminated against in housing
- Undocumented immigrants currently incarcerated and/or underpaid
- Undocumented immigrants for whom prosecutorial discretion has been used to justify release or general punishment
- Measurements for global web users that take into account shared devices and VPNs
- Firm statistics on how often police arrest women for making false rape reports
- Master database that details if/which Americans are registered to vote in multiple states
- Total number of local and state police departments using stingray phone trackers (IMSI-catchers)
- How much Spotify pays each of its artists per play of song
-

The lack of data about women who die in childbirth makes Onuoha's point plain. In the absence of U.S. government-mandated action or federal funding ProPublica had to resort to crowdsourcing to find out the names of the estimated 700 to 900 U.S. women who died in childbirth in 2016. So far, they've identified only 134. Or, for another example: In 1998, youth of color in Roxbury, Boston,

were sick and tired of inhaling polluted air. They led a march demanding clean air and better data collection, which led to the creation of the AirBeat community monitoring project. Just south of the U.S. border, in Mexico, a single anonymous woman is compiling the most comprehensive dataset on femicides – gender-related killings. The woman, who goes by the name "Princesa," has logged 3,920 cases of femicide since 2016. Her work provides the most up-to-date information on the subject for Mexican journalists and legislators –information that, in turn, has inspired those journalists to report on the subject, and has compelled those legislators to act.

Princesa has undertaken this important data collection effort because women's deaths are being neglected and going uncounted by the local, regional, and federal governments of Mexico. But it's not better anywhere else. *The Washington Post* and *The Guardian UK* currently compile the most comprehensive national count of police killings of citizens in the United States, and not the U.S. federal government. But it's powerful institutions like the federal government that, more often than not, control the terms of data collection –for several reasons that Onuoha's *Missing Data Sets* points us towards. In the present moment, in which the most powerful form of evidence is data –a fact we may find troubling, but is increasingly true –the things that we do not or cannot collect data about are very often perceived to be things that do not exist at all.

Even when the data are collected, however, they still may not be disaggregated or analyzed in terms of the categories that make issues of inequality apparent. This is, in part, what is responsible for the lack of data on maternal mortality in the United States. While there is (as of 2003) a box to check on the official U.S. death certificate that indicates whether the person who died, if female, was pregnant at the time or within a year of death, it would require a researcher who was already interested in racial disparities in healthcare to combine those data with the data collected on race for the "three times more likely" stat that Serena Williams cited in her Facebook post to be revealed.

As feminist geographer Joni Seager states, "If data are not available on a topic, no informed policy will be formulated; if a topic is not evident in standardized databases, then, in a self-fulfilling cycle, it is assumed to be unimportant." Princesa's femicide map is an outlier, a case when a private citizen stood up and took action on behalf of the bodies that were going uncounted. ProPublica solicited stories and trawled Facebook groups and private crowdfunding sites in order to compile their list of the women who would otherwise go uncounted and unnamed. But this work is precarious in that it relies upon the will of individuals or the sustained attention of news organizations in order to take place. In the case of Princesa, this work is even more precarious in that it places herself and her family at risk of physical harm.

Sometimes, however, it's the subjects of data collection who can find themselves in harm's way. When power in the collection environment is not distributed equally, those who fear reprisal have strong reasons not to come forward. Collecting data on the locations of undocumented immigrants in the United States, for

example, could on the one hand be used to direct additional resources to them; but on the other hand, it could send ICE officials to their doors. A similar paradox of exposure is evident among transgender people. Journalist Mona Chalabi has written about the challenges of collecting reliable data on the size of the transgender population in the U.S. Among other reasons, this is because transgender people are afraid to come forward for fear of violence or other harms. And so many choose to stay silent, leading to a set of statistics that does not accurately reflect the populations they seek to represent.

There is no universal solution to the problem of uncounted, undercounted, and silenced bodies. But that's precisely why it's so important to listen to, and take our cues from, the communities that we as data scientists, and data feminists, seek to support. Because these communities are disproportionately those of women, people of color, and other marginalized groups, it's also of crucial importance to recognize how data and power, far too often, easily and insidiously align. Bringing the bodies back into our discussions and decisions about what data gets collected, by whom, and why, is one crucial way in which data science can benefit from feminist thought. It's people and their bodies who can tell us what data will help improve lives, and what data will harm them*[2].

Bodies extracted for science, surveillance, and selling

Far too often, the problem is not that bodies go uncounted or undercounted, or that their existence or their interests go unacknowledged, but the reverse: that their information is enthusiastically scooped up for the narrow purposes of our

[2]There is a growing body of work dedicated the difficulties of uncounted and undercounted populations, and related phenomena. The emerging field of Critical Data Studies advocates for using frameworks from cartography and GIS which "have long been concerned with the nature of missing data", including theorizing their origins in power imbalances as well as determining ethical courses of action for mappers in diverse situations. Jonathan Gray, Danny Lämmerhirt, and Liliana Bounegru wrote a report, Changing What Counts, which includes case studies of citizen involvement in collecting data on drones, police killings, water supplies and pollution. Environmental health and justice represents an area where communities are out front collecting data when agencies refuse or neglect to do so. For example, Sara Wylie, co-founder of Public Lab, works with communities impacted by fracking to measure hydrogen sulfide using low-cost DIY sensors. The lack of data on women impacted by police violence in the U.S. led Kimberlé Crenshaw and the African American Policy Forum to develop the Black Women Police Violence database, designed to challenge the narrative that policy violence only affects males of color. Erin McElroy's work on community-collected eviction data in San Francisco, as part of the Anti-Eviction Mapping Project, demonstrates how data that originates in communities can be more complete and grounded than outside data collection efforts. Indigenous cartographers Margaret Pearce and Renee Pualani Louis describe cartographic techniques for recuperating indigenous perspectives and epistemologies (often absent or misrepresented) into GIS maps. And through methods like crowdsourcing or sensor journalism, the data journalism community is not just reporting with existing data, but increasingly undertaking projects that involve compiling their own databases in the absence of official data sources. That said, participatory data collection efforts have their own silences, as Heather Ford and Judy Wajcman show in their study of the 'missing women' of Wikipedia.

data-collecting institutions. For example, in 2012, *The New York Times* published an explosive article by Charles Duhigg, "How Companies Learn Your Secrets," which soon became the stuff of legend in data and privacy circles. Duhigg describes how Andrew Pole, a data scientist working at Target, synthesized customers' purchasing histories with the timeline of those purchases in order to infer if a customer might be pregnant. (Evidently, pregnancy is the second major life event, after leaving for college, that determines whether a casual shopper will become a customer for life). Pole's algorithm was so accurate that he could not only identify the pregnant customers, but also predict their due dates.

But then Target turned around and put this algorithm into action by sending discount coupons to pregnant customers. Win-win. Or so they thought, until a Minneapolis teenager's dad saw the coupons for maternity clothes that she was getting in the mail, and marched into his local Target to read the manager the riot act. Why was his daughter getting coupons for pregnant women when she was only a teen?!

It turned out that the young woman was, indeed, pregnant. Pole's algorithm informed Target before the teenager informed her father. Evidently, there are approximately twenty-five common products, including unscented lotion and large bags of cotton balls, that, when analyzed together, can predict whether or not a customer is pregnant, and if so, when they are due to give birth. But in the case of the Minneapolis teen, the win-win quickly became a lose-lose, as Target lost a potential customer and the pregnant teenager lost far worse: her privacy over information related to her own body and her health. In this way, Target's pregnancy prediction algorithm helps to illustrate another reason why bodies must be brought back to the data science table: without the ability of individuals and communities to shape the terms of their own data collection, their bodies can be mined and their data can be extracted far too easily –and done so by powerful institutions who rarely have their best interests at heart.

At root, this is another question of power, along with a question of priorities and resources – financial ones. Data collection and analysis can be prohibitively expensive. At Facebook's newest data center in New Mexico, the electrical cost alone is estimated at $31 million annually. Only corporations like Target, along with well-resourced governments and elite research universities, have the resources to collect, store, maintain, and analyze data at the highest levels. It's the flip side of the lack of data on maternal health outcomes. Put crudely, there is no profit to be made collecting data on the women who are dying, but there is significant profit in knowing whether women are pregnant.

Data has been called "the new oil"for, among other things, its untapped potential for profit and its value once it's processed and refined. But just as the original oil barons were able to use that profit to wield outsized power in the world –think of John D. Rockefeller, J. Paul Getty, or, more recently, the Koch brothers – so too do the Targets of the world use their data capital to consolidate control over their customers. But it's not petroleum that's extracted in this case; it's data that's extracted from people and communities with minimal consent. This

basic fact creates a profound asymmetry between who is collecting, storing, analyzing and visualizing data, and whose information is collected, stored, analyzed, and visualized. The values that drive this extraction of data represent the interests and priorities of the universities, governments, and corporations that are dominated by elite, white men. We name these values the three S's: science (universities), surveillance (governments) and selling (corporations).[3]

In the case of Target and the pregnant teen, the originating charge from the marketing department to Andrew Pole was: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?" But did the teenager have access to her purchasing data? No. Did she or her parents have a hand in formulating any of the questions that Target might wish to ask of its millions of records of consumer purchases? No. Did they even know that their family's purchasing data was being analyzed and recorded? No no no. They were not invited to the design table, even though it was one on which their personal data was put out on (corporate) display. Instead, it was Target –a company currently valued at $32 billion dollars – that determined what data to collect, and what questions to ask of it.

The harms inflicted by this asymmetry don't only have to do with personal exposure and embarrassment, but also with the systematic monitoring, control, and punishment of the people and groups who hold less power in society. For example, Paola Villareal's data analysis for the ACLU reveals clear racial disparities in the City of Boston's approach to policing marijuana-related offenses. (Additional analyses have found this phenomenon to be true in cities across the United States). In *Automating Inequality*, Virginia Eubanks provides another example of how the asymmetrical relationship between data-collecting institutions and the people about which they collect data plays out. The Allegheny County Office of Children, Youth, and Families, in Pennsylvania, employs an algorithmic model to predict the risk of child abuse. Additional methods of detecting child abuse would seem to be a good thing. But the problem with this particular model, as with most predictive algorithms in use in the world today, is that it has been

[3]In their widely cited paper Critical Questions for Big Data, danah boyd and Kate Crawford outlined the challenges of unequal access to big data, noting that the current configuration (in which corporations own and control massive stores of data about people) creates an imbalance of power in which there are "Big Data rich" and "Big Data poor." Media scholar Seeta Peña Gangadharan has detailed how cont emporary data profiling disproportionately impacts poor, communities of color, migrants and indigenous groups. Social scientist Zeynep Tufecki warns that corporations have emerged as "power brokers" with outsized potential to influence politics and publics precisely because of their exclusive data ownership. Building on this, Mark Andrejevic has outlined a "big data divide" in which only elite institutions have abilities to capture, mine and utilize data whereas individuals do not, privileging "a form of knowledge available only to those with access to costly resources and technologies." Jeff Warren describes how this gives "data shepherds" (technologists) disproportionate power over knowledge production and discourse, circumscribing the kinds of questions that can be asked in a democracy. And in advancing the idea of "Black data" to refer to the intersection of informatics and Black queer life, Shaka McGlotten states, "How can citizens challenge state and corporate power when those powers demand we accede to total surveillance, while also criminalizing dissent?"

designed unreflexively. In this case, the problem is rooted in the fact that it takes into account every single data source that it can get. For wealthier parents, who can more easily access private health care and mental health services, there is simply not that much data. But for poor parents, who primarily access public resources, the model scoops up records from child welfare services, drug and alcohol treatment programs, mental health services, jail records, Medicaid histories, and so on. Because there is far more data about poor parents, they are oversampled in the model, and disproportionately targeted for intervention. The model "confuse[s] parenting while poor with poor parenting," Eubanks explains – with the most profound of results.

Ensuring that bodies are not simply viewed as a resource, like oil, that can be "extracted" and "refined," is another way that data feminism can intervene in our current data practices. Like the process of data collection, this process of extracting bodies is one that disproportionately impacts women, people of color, and others who are more often subject to power rather than in possession of it. And it's another place where bringing the bodies back into discussions about data collection, and its consequences, can begin to challenge and transform the unequal systems that we presently face.

Bodies absent from data work

One place where these conversations need to be happening is in the field of data science itself. It's no surprise to observe that women and people of color are underrepresented in data science, just as they are in STEM fields as a whole. The surprising thing is that the problem is getting worse. According to a research report published by the American Association of University Women in 2015, women comprised 35% of computing and mathematical occupations in 1990, but this percentage dropped to 26% in 2013.[4] They are being pushed out as "data analysts" have become rebranded as "data scientists," in order to make room for more highly valued and more highly compensated men.[5] We identify this later in the book as what we call a "privilege hazard," one in which discrimination

[4]For comparison, this is the same percentage of female information science graduates in 1974. And in subfields like machine learning, the proportion of women is even less. As per the points made in this chapter, even knowing the exact extent of the disparity is challenging. According to a 2014 Mother Jones report about diversity in Silicon Valley, tech firms convinced the U.S. Labor Department to treat their demographics as a trade secret, and didn't divulge any data until after they were sued by Mike Swift of the *San Jose Mercury News*. There are analyses that have obtained the data in other ways. For example, a gender analysis by data scientists at LinkedIn has shown that tech teams at tech companies have far *less* gender parity than tech teams in other industries including healthcare, education, and government.

[5]This phenomenon, while new to data science, is unfortunately as old as time. Scholars such as Marie Hicks and Nathan Ensmenger have shown how the push to professionalize computer science resulted in the pushing out of the women who had previously performed those same roles. Historians of medicine often point to the history of obstetrics, in which female midwives were replaced by male obstetricians after the advent of formal medical schools. The same phenomenon can be found in the kitchen, with women performing most home cooking, unpaid altogether, while men attend culinary school to become celebrity chefs.

becomes hard-coded into so-called "intelligent systems," because the people doing the coding are the most privileged – and therefore least well-equipped – to acknowledge and account for inequity.[6]

This privilege hazard is a risk that can rear its head in harmful ways. For example, in 2016, MIT Media Lab graduate student Joy Buolamwini, founder of the Algorithmic Justice League, was experimenting with software libraries for the Aspire Mirror project. This project used computer vision software to overlay inspirational images (like a favored animal or an admired celebrity) onto a reflection of the user's face. She would open up her computer and run some code that she'd written, built on a free JavaScript library that used her computer's built-in camera to detect the contours of her face. Buolamwini's code was bug-free, but she couldn't get the software to work for a more basic reason: it had a really hard time detecting her face in front of the camera. Buolamwini has dark skin. While her computer's camera picked up her lighter-skinned colleague's face immediately, it took much longer for the camera to pick up Buolamwini's face, when it did at all. Even then, sometimes, her nose was identified as her mouth. What was going on?

What was going on was this: facial analysis technology, which uses machine learning approaches, learns how to detect faces based on existing collections of data that are used to train, validate, and test the algorithms for use in future models. These datasets are constructed in advance, in order to present any particular learning algorithm with a representative sample of the kinds of things it might encounter in the real world. But problems arise very quickly when the biases that already exist in the world are replicated in these datasets. Upon digging into the benchmarking data for facial analysis algorithms, Buolamwini learned that they consisted of 78% male faces and 84% pale faces, sharply at odds with a global population that is majority female and non-pale.[7] How could such an oversight have happened? Easily, when most engineering teams have 1) few women or people of color; and 2) no training to think about #1 as a problem.

Oversights like this happen more often than you might think, and with a wide

[6]Social scientist Kate Crawford has advanced the idea that the biggest threat from artificial intelligence systems is not that they will become smarter than humans, but rather that they will hardcode sexism, racism and discrimination into the digital infrastructure of our societies. This is evident not only in data products and systems themselves but also in the divisions of labor in the data economy. The book Ghost Work by anthropologist Mary Gray and computer scientist Siddharth Suri details the existence of a "global underclass" performing work like content moderation, transcription, and captioning. While Silicon Valley tech workers remain steadily young, white and male, these "ghost workers" are often older, often female and minority, and always precarious.

[7]Specifically, the breakdown for the Labeled Faces in the Wild (LFW) dataset was 77.5% male faces and 83.5% white faces. And Buolamwini and Timnit Gebru showed that the breakdown for the IARPA Janus Benchmark A (IJB-A) dataset published by the US government was 75% male and 80% pale faces (as determined by the Fitzpatrick skin type). But Buolamwini makes the additional point that population parity in the test data is not always the answer, because small populations like Native Americans might not have enough test cases to determine whether the model was working.

Figure 1.1: Joy Buolamwini had to resort to "white face" to get a computer vision algorithm to detect her face. Many facial detection algorithms have only been trained on pale and male faces. ¶ Credit: Joy Buolamwini ¶ Source: https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148 ¶ Permissions: Pending

range of consequences. Consider a craze that (briefly) swept the Internet in Spring 2018. In order to promote awareness of its growing number of digitized museum collections, Google released a new feature for its Arts and Culture app. You could take a selfie, upload the image, and the app would find the face from among its millions of digitized artworks that looked the most like you. All over Facebook, Twitter, and Instagram, people were posting side-by-side shots of themselves and – for instance, the *Mona Lisa*, *American Gothic,* or a Vermeer.

Well, white people were. Because most of the museums with collections that Google had helped to digitize came from the U.S. and Europe, most featured artworks from the Western canon. And because most artworks from the Western canon feature white people, the white users of the Arts and Culture app found really good matches for their faces. But some Asian users of the app, for example, found themselves matched with one of only the handful of portraits of Asian people included in those collections.

On Twitter, the response to this inadequacy was tellingly resigned. One user, @pitchaya, whose Tweet was quoted in a *digg.com* article on the subject, tweeted sarcastically: "If you do that whole Google Arts & Culture app portrait comparison as an Asian male, it gives you one of 5-6 portraits that hardly resembles you but, hey, looks Asian enough." Another user, @rgan0, also quoted in the piece, called out Google directly: "The Google Arts and Culture app thinks I look like a"Beautiful [Japanese] Woman"! :p get more Asian faces in your art

database, Google."

And if the disparities of representation in Western art museums weren't enough of a problem, some Art and Culture App users worried about something more insidious taking place. For app users to upload their images for analysis, they had to agree to allow Google to access those images. Were their images also being stored for future internal research? Was Google secretly using crowdsourcing to improve its training data for its own facial recognition software, or for the NSA? A short-lived internet uproar ensued, ending only when Google updated the user agreement to say: "Google won't use data from your photo for any other purpose and will only store your photo for the time it takes to search for matches."

But what if they had been? The art selfie conspiracy theorists weren't actually too far from reality, given that earlier that year, Amazon had briefly contracted with the Orlando Police Department to use its own proprietary facial recognition software, trained on its own proprietary data, to help the police automatically identify suspects in real time. How representative was Amazon's training, benchmarking, or validation data? Was it more or less representative than the data that Buolamwini explored in her research? There was no way to know. And while a best match of 44% between Asian Art and Culture App users and Terashima Shimei's *Beautiful Woman* (which is the painting @rgon0 matched with) might earn RTs of solidarity on Twitter, a best match of 44% between a suspected criminal and a random person identified through traffic camera footage –the image source for the Amazon project –could send an innocent person to jail. Who any particular system is designed for, and who that system is designed by, are both issues that matter deeply. They matter because the biases they encode, and often unintentionally amplify, remain unseen and unaddressed –that is, until someone like Buolamwini literally has to face them. What's more, without women and people of color more involved in the coding and design process, the new research questions that might yield groundbreaking results don't even get asked –because they're not around to ask them. As the example of facial analysis technology, or the Google Arts and Culture app help to show, there is a much higher likelihood that biases will be designed into data systems if the bodies of the system's designers themselves only represent the dominant group.

Bodies invisible: The view from nowhere is always a view from somewhere

So far, we've shown how bringing the bodies back into data science can help expose the inequities in the scope and contents of our data sets, as in the example of the hundreds of unnamed U.S. women who die in childbirth each year. We've also shown how bringing back the bodies can help avoid their data being mined without their consent, as in the example of the Minneapolis teenager who Target identified as pregnant. And we've also shown how bringing bodies that are more representative of the population into the field of data science can help avert the

increasing number of racist, sexist data products that are inadvertently released into the world, as in the example of the Google Arts and Culture app, or of the facial recognition software that is the focus of Joy Buolamwini's research. (We'll have more to say about some of its worst applications, like state surveillance, in the chapters to come).

But there are other bodies that need to be brought back into the field of data science not because they're not yet represented, but because of the exact opposite reason: they are overrepresented in the field. They are so overrepresented that their identities and their actions are simply assumed to be the default. An example that Yanni Loukissas includes in his book, *All Data are Local*, makes this point crystal clear: Marya McQuirter, a historian at the Smithsonian Institution's National Museum of African American History and Culture, recalls searching the Smithsonian's internal catalog for the terms "black" and "white." Searching the millions of catalog entries for "black" yielded a rich array of objects related to Black people, Black culture, and Black history in the US : the civil rights movement, the jazz era, the history of enslavement, and so on. But searching for "white" yielded only white-colored visual art. Almost nothing showed up relating to the history of white people in the United States.

McQuirter, who is Black, knew the reason why: in the United States, it's white people and their bodies who occupy the "default" position. Their existence seems so normal that they go unremarked upon. They need not be categorized, because – it is, again, assumed – most people are like them. This is how the perspective of only one group of bodies –the most dominant and powerful group –becomes invisibly embedded in a larger system, whether it's a system of classification, as in the case of McQuirter's catalog search; a system of surveillance, as in the case of Amazon and the Orlando police; or a system of knowledge, as reflected in a data visualization, as we'll now explain –

Whose perspective are we seeing when we see a visualization like this one of global shipping routes?

We are not seeing any particular person's perspective when we look at this map (unless you are an astronaut on the space station and you have weird blue glasses on that make all the continents blue). In terms of visualization design, this is for good reason - it is precisely this impossible, totalizing view which makes any particular visualization so dazzling and seductive, so rhetorically powerful, and so persuasive.[8] This image appears to show us the "big picture" of the entire world. Because we do not see the designers of this image, nor can we detect any visual indicators of human involvement, the image appears truthful, accurate, and free of bias.

This is what feminist philosopher Donna Haraway describes as "the god trick." By the "god" part, Haraway refers to how data is often presented as though

[8]Sociologist Helen Kennedy and her colleagues have shown how visual conventions such as two-dimensional layouts, and geometric shapes, contribute to the pervasive view of data visualization as a neutral and scientific method of display.

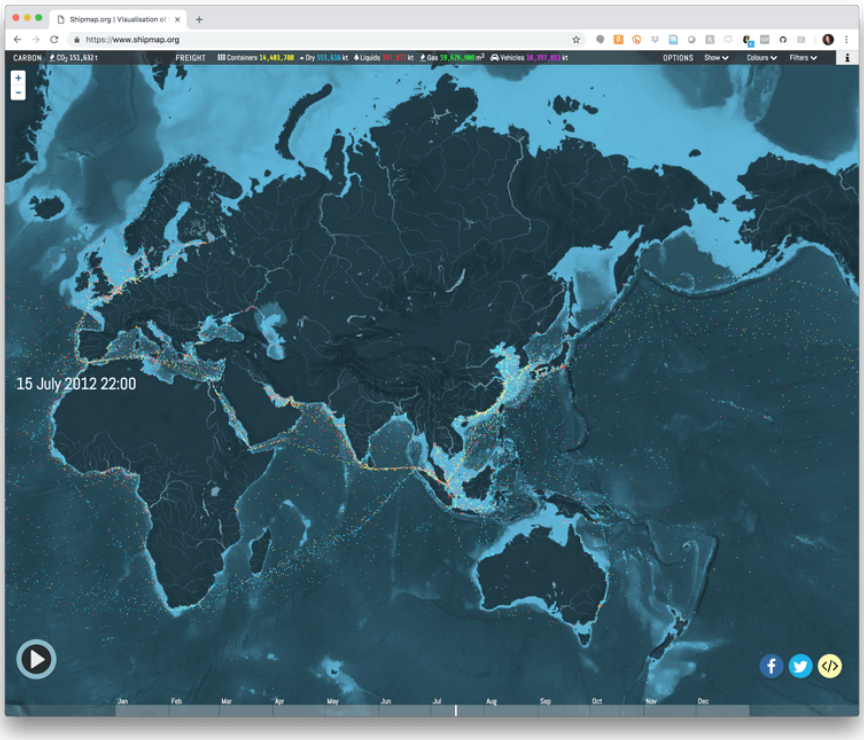


Figure 1.2: Time-based visualization of global shipping routes designed by Kiln based on data from the UCL Energy Institute. ¶ Credit: Website created by Duncan Clark & Robin Houston from Kiln. Data compiled by Julia Schaumeier & Tristan Smith from the UCL EI. The website also includes a soundtrack: Bach's Goldberg Variations played by Kimiko Ishizaka. ¶ Source: https://www.shipmap.org/

it inhabits an omniscient, godlike perspective. But the "trick" is that the bodies who helped to create the visualization – whether through providing the underlying data, collecting it, processing it, or designing the image that you see it –have themselves been rendered invisible. There are no bodies in the image anymore.

Haraway terms this "the view from nowhere." But the view from nowhere is always a view from somewhere: the view from the default. Sometimes this view comes into focus when considering what isn't revealed, as in the case of McQuirter's search query. But when we do not remind ourselves to ask what we are *not* seeing, and about *who* we are not seeing –well, that is the most serious body issue of all. It's serious because all images and interactions, the data they are based on, and the knowledge they produce, comes from bodies. As a result, this knowledge is necessarily incomplete. It's also necessarily culturally, politically, and historically circumscribed. Pretending otherwise entails a belief in what sociologist Ruha Benjamin, in *Race After Technology: Abolitionist Tools for the New Jim Code*, describes as the "imagined objectivity of data and technology," because it's not objectivity at all.

To be clear: this does *not* mean that there is no value in data or technology. What this means for data science is this: if we truly care about objectivity in our work, we must pay close attention to whose perspective is assumed to be the default. Almost always, this perspective is the one of elite white men, since they occupy the most privileged position in the field, as they do in our society overall. Because they occupy this position, they rarely find their dominance challenged, their neutrality called into question, or their perspectives open to debate. Their privilege renders their bodies invisible– in datasets, in algorithms, and in visualizations, as in their everyday lives.

Ever heard of the phrase, "History is written by the victors"? It's the same sort of idea. Both in the writing of history and in our work with data, we can learn so much more – and we can get closer to some sort of truth – if we bring together as many bodies and perspectives as we can. And when it comes to bringing these bodies back into data science, feminism becomes increasingly instructive, as the rest of the chapters in this book explain.

In *On Rational, Scientific, Objective Viewpoints from Mythical, Imaginary, Impossible Standpoints. we build on Haraway's notion of the god trick, exploring some reasons why emotion has been kept out of data science as a field, and what we think emotion can, in fact, contribute. We talk about emotional data, among data of many other forms, in* What Gets Counted Counts* –a chapter that emphasizes the importance of thinking though each and every one of the choices we make when collecting and classifying data. The next chapter, *Unicorns, Janitors, Ninjas, Wizards, and Rock Stars*, challenges the assumption that data scientists are lone rangers who wrangle meaning from mess. Instead, we show how working with communities and embracing multiple perspectives can lead to a more detailed picture of the problem at hand. This argument is continued in *The Numbers Don't Speak for Themselves*, in which we show how much of

today's work involving "Big Data" prioritizes size over context. In contrast, feminist projects connect data back to their sources, pointing out the biases and power differentials in their collection environments that may be obscuring their meaning. We turn to the contexts and communities that ensure that the work of data science can take place in *Show Your Work*, a chapter that centers on issues of labor. In *The Power Chapter*, it's, well, power, privilege, and structural inequality that we take up and explore. *Teach Data Like an Intersectional Feminist* provides a series of examples of how to implement the lessons of the previous chapters in classrooms, workshops, and offices, so that we can train the next generations of data feminists. And in *Now Let's Multiply*, we speculate about other approaches that might enrich a conversation about data science, its uses, and its limits.

There is growing discussion about the uses and limits of data science, especially when it comes to questions of ethics and values. But so far, feminist thinking hasn't directed the conversation as it might. As a starting point, let's take the language that is increasingly employed to discuss questions of ethics in data and the algorithms that they support, such as the computer vision and predictive policing algorithms we've described just above. The emerging best practices in the field of data ethics involve orienting algorithmic work around concepts like "bias," and values like "fairness, accountability, and transparency." This is a promising development, especially as conversations about data and ethics enter the mainstream, and funding mechanisms for research on the topic proliferate. But there is an additional opportunity to reframe the discussion before it gathers too much speed, so that its orienting concepts do not inadvertently perpetuate an unjust status quo.

Consider this chart, which uses Benjamin's prompt to reconsider the "imagined objectivity of data and technology" in order to develop an alternative set of orienting concepts for the field. These concepts have legacies in intersectional feminist activism, collective organizing, and critical thought, and they are unabashedly explicit in how they work towards justice:

Concepts Which Uphold "Imagined Objectivity" *Because they locate the source of the problem in individuals or technical systems*	**Intersectional Feminist Concepts Which Strengthen Real Objectivity** *Because they acknowledge structural power differentials and work towards dismantling them*
Ethics	Justice
Bias	Oppression
Fairness	Equity
Accountability	Co-liberation
Transparency	Reflexivity

Concepts Which Uphold "Imagined Objectivity" *Because they locate the source of the problem in individuals or technical systems*	**Intersectional Feminist Concepts Which Strengthen Real Objectivity** *Because they acknowledge structural power differentials and work towards dismantling them*
Understanding algorithms	Understanding history, culture, and context

The concept of "bias," for example, locates the source of inequity in the behavior of individuals (i.e. a prejudiced person) or in the outcomes of a technical system (i.e. a system that favors white people or men). Under this conceptual model, a technical goal might be to create an "unbiased" system. First we would design a system, use data to tune its parameters and then we would test for any biases that result. We could even define what might be more "fair," and then we could optimize for that.

But this entire approach is flawed, like the imagined objectivity that shaped it. Just as Benjamin cautions against imagining that data and technology are objective, we must caution ourselves against locating the problems associated with "biased" data and algorithms in technical systems alone. This is a danger that computer scientists have noted in relation to high-stakes domains like criminal justice, where hundreds of years of history, politics, and economics, not to mention the complexities of contemporary culture, are distilled into black-boxed algorithms that determine the course of people's lives. In this context, computer scientist Ben Green warns about the narrowness of computationally conceived fairness, writing that "computer scientists who support criminal justice reform ought to proceed thoughtfully, ensuring that their efforts are driven by clear alignment with the goals of justice rather than a zeitgeist of technological solutionism." And in keynoting the Data Justice Conference in 2018, Sasha Costanza-Chock challenged the audience to expand their concept of ethics to justice, in particular *restorative justice* which recognizes and accounts for the harms of the past. We do not all arrive in the present moment with equal power and privilege. When "fairness" is a value that does not acknowledge context or history, it fails to acknowledge the systematic nature of the "unfairness" perpetrated by certain groups on other groups for centuries.

Does this make fairness political? Emphatically yes, because all systems are political. In fact, the appeal to avoid politics is a very familiar move for those in power to continue to uphold the status quo. The ability to do so is also a privilege, one held only by those whose existence does not challenge that same status quo. Rather than designing algorithms that are "color blind," Costanza-Chock says, we should be designing algorithms that are *just*. This means shifting from ahistorical notions of fairness to a model of *equity*. This model would take time, history, and differential power into account. Researcher Seeta Peña Gangadharan, co-lead of the Our Data Bodies project, states, "The question is

not 'How do we make automated systems fairer?' but rather to think about how we got here. How might we recover that ability to collectively self determine?"

This is why *bias* (in individuals, in data sets, or in algorithms) is not a strong enough concept in which to anchor ideas about equity and justice. In writing about the creation of New York's Welfare Management System in the early 1970s, for example, Virginia Eubanks describes: "These early big data systems were built on a specific understanding of what constitutes discrimination: personal bias." The solution at the time was to remove the humans from the loop, and it remains so today: without potentially bad –in this case, racist – apples, there would be less discrimination. But this line of thinking illustrates what Robin DiAngelo would call the " 'new' racism": the belief that racism is due to individual bad actors, rather than structures or systems. In relation to welfare management, this often means replacing the women of color social workers, who have empathy and flexibility and listening skills, with an automated system that applies a set of rigid criteria, no matter what the circumstances.

Bias is not a problem that can be fixed after the fact. Instead, we must look to understand and design systems that address *oppression* at the structural level. Oppression, as defined by the comic artist Robot Hugs, is what happens "when prejudice and discrimination is supported and encouraged by the world around you. It is when you are harmed or not helped by government, community or society at large because of your identity," they explain. And while the research and energy emerging around algorithmic *accountability* is promising, why should we settle for retroactive audits of potentially flawed systems if we could design for *co-liberation* from the start? Here co-liberation doesn't mean "free the data," but rather "free the people." And the people in question are not only those with less power, but also those with relative privilege (like data scientists, designers, researchers, educators; like ourselves) who play a role in upholding oppressive systems. Poet and community organizer Tawana Petty defines what co-liberation means in relation to anti-racism in the U.S.: "We need whites to firmly believe that their liberation, their humanity is also dependent upon the destruction of racism and the dismantling of white supremacy." The same goes for gender – men are often not even thought to have a gender, let alone prompted to think about how unequal gender relations seep into our institutions and artifacts and harm all of us. In these situations, it is not enough to do audits after-the-fact. We should be able to dream of data-driven systems that position co-liberation as their primary design goal.

Designing data sets and data systems that dismantle oppression and work towards justice, equity, and co-liberation requires new tools in our collective toolbox. We have some good starting points – building more *understandable algorithms* is a laudable, worthy research goal. And yet, what we need to explain and account for are not only the inner workings of machine learning, but also the *history, culture, and context* that lead to discriminatory outputs in the first place. Did you know, for example, that the concept of homophily which provides the rationale for most contemporary network clustering algorithms in fact derives from 1950s-era

models of housing segregation? (If not, we recommend you read Wendy Chun). Or, for another example, did you know that the "Lena" image used to test most image processing algorithms is the centerfold from the November 1972 issue of *Playboy*, cropped demurely at the shoulders? (If not, Jacob Gaboury is the one to consult on the subject). These are not merely bits of trivia to be pulled out to impress dinner party guests. On the contrary, they have very real implications for the design of algorithms, and for their use.

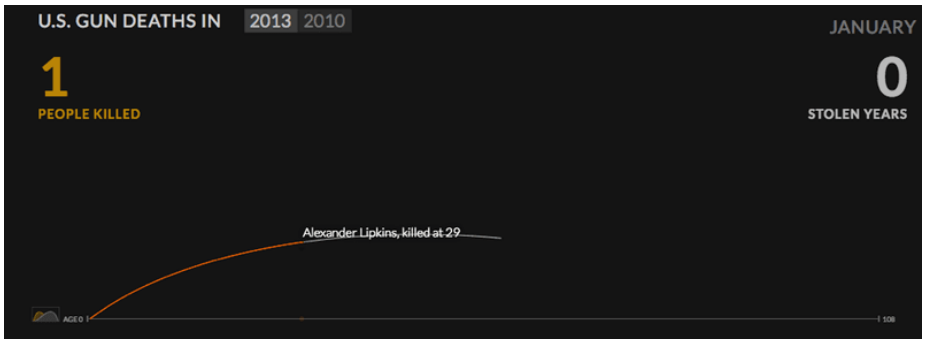
How might we design a network clustering algorithm that does not perpetuate segregation, but actively strives to bring communities together? (This is a question that Chun is pursuing in her current research). How might we ensure that the selection of test data isn't ever relegated to happenstance? (This is how the "Lena" image, which encoded sexism into the field of image processing, is explained away). The first step requires *transparency* in our methods as well as the *reflexivity* to understand how our own identities, our communities, and our domains of expertise are part of the problem. But they can also be part of the solution.

When we start to ask questions like: "Whose bodies are benefiting from data science?" "Whose bodies are harmed?" "How can we use data science to design for a more just and equitable future?" and "By whose values will we re-make the world?" we are drawing from *data feminism*. It's data feminism that we describe in the rest of this book. It's what can help us understand how power and privilege operate in the present moment, and how they might be rebalanced in the future.

Chapter 2

On Rational, Scientific, Objective Viewpoints from Mythical, Imaginary, Impossible Standpoints

In 2012, twenty kindergarten children and six adults were shot and killed at an elementary school in Sandy Hook, CT. In the wake of this tragedy, and the weight of others like it, the design firm Periscopic started a new project – to visualize gun deaths in the United States. While there is no shortage of prior work in the form of bar charts or line graphs of deaths per year, Periscopic, a company whose tagline is "do good with data", took a different approach. When you load the webpage, you see a single, arcing line that reaches out over time. Then, the color abruptly shifts from orange to white. A small dot drops down, and you see the phrase, "Alexander Lipkins, killed at 29". The arc continues to stretch across the screen, coming to rest on the x-axis, where you see a second phrase, "could have lived to be 93." Then, a second arc appears, displaying another arcing life. The animation speeds up over time, and the arcing lines increase, along with a counter that displays how many years of life have been "stolen" from these gun victims. After a couple of (long) minutes, the visualization moves through all of the year 2013, arriving at 11,419 people killed and 502,025 stolen years.



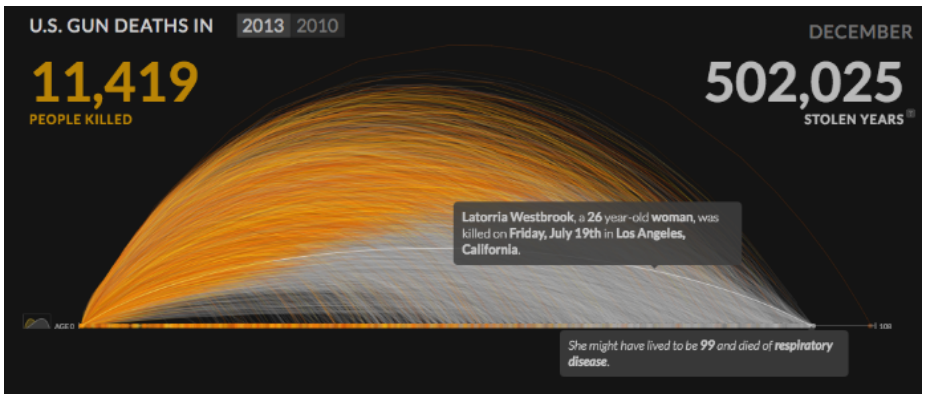
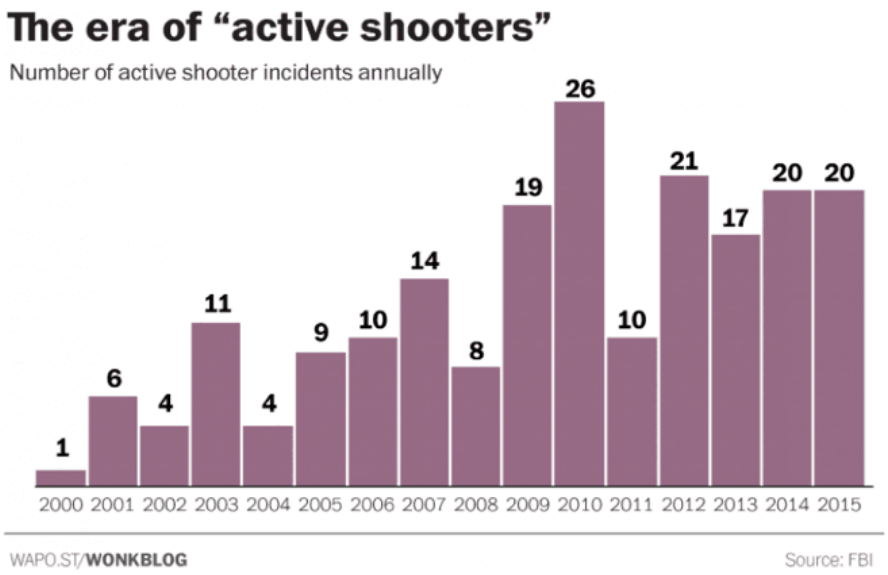


Figure 2.1: An animated visualization of the "stolen years" of people killed by guns in the United States in 2013. Credit: Periscopic Source: https://guns.periscopic.com/?year=2013



:::

What is different about Periscopic's visualization than a more conventional bar chart of similar information such as "The era of 'active shooters'" from the Washington Post?

The Post's graphic has a proposition - that active shooter incidents are on the rise - and demonstrates visual evidence to that effect. But Periscopic's work is framed around a singular emotion: loss. People are dying, their remaining time on earth has been stolen from them. These people have names and ages. We presume they have parents and partners and children who also suffer from that loss. The data scientists who worked on the project used rigorous statistical methods and demographic information in order to infer how long that person would have lived, which are documented in their notes. But in spite of the statistical rigor, and its undeniable emotional impact, “U.S. Gun Deaths” drew mixed responses from the visualization community. They couldn't decide: *should a visualization evoke emotion?*

The received wisdom in technical communication circles is, emphatically, "NO." In the recent book, "A Unified Theory of Information Design," the authors state: "The plain style normally recommended for technical visuals is directed toward a deliberately neutral emotional field, a blank page in effect, upon which viewers are more free to choose their own response to the information." Here, plainness is equated with the absence of design, and thus greater freedom on the part of the viewer to interpret the results for themselves. Things like colors and icons, it is implied, work only to stir up emotions and cloud the viewer's rational mind. In fact, in the field of data visualization, any kind of ornament has historically been viewed as suspect. Why? Well, as historian of science Theodore Porter puts it, "quantification is a technology of distance" and distance was historically imagined

to serve objectivity by producing knowledge independently of the people that make it. This echoes nineteenth-century statistician Karl Pearson's exhortation for people to set aside their own feelings and emotions when it came to statistics. The more seemingly neutral, the more rational, the more true, the better. At a data visualization master class in 2013, workshop leaders from the Guardian newspaper called spreadsheet data– those endless columns and rows– "clarity without persuasion."

Back in the olden days of visualization, before the rise of the web elevated the visual display of data into a prized (and increasingly pervasive) artform, Edward Tufte, statistician and statistical graphics expert, invented a metric for measuring the superfluous information included in a chart– what he called the "data-ink" ratio. In his view, a visualization designer should strive to use ink only to display the data. Any ink devoted to something other than the data itself – such as background color, iconography, or embellishment – should be immediately erased and, he all but says, the designer spat upon. Visual minimalism, according to this logic, appeals to reason first. ("Just the facts, ma'am" says Joe Friday to every female character on Dragnet). Decorative elements, on the other hand, are associated with messy feelings– or, worse, represent stealthy (and, of course, unscientific) attempts at emotional persuasion. Data visualization has even be classified as the "unempathetic" art, in the words of designer Mushon Zer-Aviv, because of its emphatic rejection of emotion.

The gendered dimension of this thinking should be as clear as when President Richard Nixon, in a now-infamous statement, declared that no "woman should be in any government job... because they are erratic. And emotional." (He subsequently admitted that "men are erratic and emotional, too," but insisted that "a woman is more likely to be.") We saw this same sentiment play out in the presidential debates of 2016, between Hillary Clinton and Donald Trump, but it's not limited to politics alone. The truth is that most Anglo-Western cultures have long prized reason over emotion for its supposedly greater neutrality and universality. And the belief that women are more emotional than men (and, by contrast, that men are more reasoned than women) is one of the most persistent stereotypes in the world today. Indeed, psychologists have called it the "master stereotype," and puzzled over how it endures even when certain emotions– even extreme ones, like anger and pride– are simultaneously coded as male. One need only compare any number of "Hulk smash!" GIFs to the equal number of crazy-lady PMS memes floating around the Internet to prove, first of all, that everyone is crazy; and second, that the stereotype of the unstable, irrational woman persists.

But what happens if we let go of the binary logic for a minute and posit two questions to challenge this master stereotype. First, is visual minimalism really more neutral? And second, how might activating emotion** – leveraging it, rather than resisting emotion in data visualization – help us learn, remember, and communicate with data?

Until recently, data visualization was a rather specialized form of communication,

more common in scientific papers and earnings reports than on the front page of *The New York Times*. Because of this history, the field's theorists and practitioners have come from technical disciplines aligned with engineering and computer science, and have not been trained in the most fundamental of all Western communication theories: *rhetoric*. In the ancient Greek treatise of the same name, Aristotle defines rhetoric as "the faculty of observing in any given case the available means of persuasion." Rhetoric does not (only) consist of political speeches made by men in robes on ancient stages. Any communicating object that makes choices about the selection and representation of reality is a rhetorical object. Whether or not it is rhetorical (it is) has nothing to do with whether or not it is "true" (it may or may not be).

Why does the question of rhetoric matter? Well, because "a rhetorical dimension is present in every design," as Jessica Hullman, a researcher at the University of Washington, says of data visualization. This includes visualizations that do not deliberately intend to persuade people of a certain message. We would say that it *especially and definitively* includes those so-called "neutral" visualizations that do not appear to have an editorial hand. In fact, those might even be the most perniciously persuasive visualizations of all!

Hullman and co-author Nicholas Diakopoulous wrote an influential paper in 2011 introducing concepts of rhetoric to the information visualization community. Their main argument is that visualizing data involves editorial choices – some things are necessarily highlighted, while others are necessarily obscured. When designers make these choices, they carry along with them "framing effects," which is to say they have an impact on how people interpret the graphics and what they take away from them. For example, it is standard practice to cite the source of one's data. This functions on a practical level – so that a reader may go out and download the data themselves. But this choice also functions as what Hullman and Diakopoulous call *provenance rhetoric* designed to signal the transparency and trustworthiness of the presentation source to end-users. This trust between the designers and their audience, in turn, increases the likelihood that viewers will believe what they see.

So if plain, "unemotional" visualizations are not neutral, but are actually extremely persuasive, then what does this mean for the concept of neutrality in general? Scientists and journalists are just some of the people that get nervous and defensive when questions about neutrality and objectivity come up. Auditors and accountants get nervous, too. They often assume that the only alternative to objectivity is a retreat into complete relativism, and a world in which everyone gets a medal for having an opinion. But feminists would beg to differ. (Feminists, generally speaking, do not like alternative facts any more than scientists do). Rather than valorizing the "neutrality ideal," feminist thinkers like Sandra Harding would posit a different kind of objectivity that strives for truth *at the same time* that it considers– and discloses– the standpoint of the designer. This has come to be called "standpoint theory." It is defined by what Harding calls "strong objectivity" which acknowledges that regular-grade, vanilla

objectivity is mainly made by mostly rich white guys in power and does not include the experiences of women and other marginalized groups.

This myopia inherent in traditional "objectivity" is what provoked renowned cardiologist Dr. Nieca Goldberg to title her book "Women Are Not Small Men," because she found that heart disease in women unfolds in a fundamentally different way than in men. The vast majority of scientific studies– not just of heart disease, but of most medical conditions– are conducted on male subjects, with women viewed as varying from this "norm" only by their smaller size. Harding and her followers would say that the key to fixing this issue is to acknowledge that all science, and indeed all work in the world, is undertaken by individuals, each with a particular standpoint – gender, race, culture, heritage, life experience, and so on. Rather than viewing these standpoints as threats that might *bias* our work – for, after all, even the standpoint of a rich white guy in power is a standpoint – we should embrace each of our standpoints as valuable perspectives that can *frame* our work. Our diverse standpoints can generate creative and wholly new research questions. We discuss standpoint theory further in *Unicorns, Janitors, Ninjas, Wizards and Rock Stars*, where we assert that a participatory process that centers the standpoints of those most marginalized is what makes for strong objectivity.

Along with this embrace of our various standpoints goes the rebalancing of the false binary between reason and emotion.[1] Since the early 2000s, there has been an explosion of research about "affect"– the term that academics use to refer to emotions and other subjective feelings– from fields as diverse as neuroscience, geography, and philosophy. This work challenges the thinking, inherited from Descartes, which casts emotion as irrational and illegitimate, even as it undeniably influences all of the social and political processes of our world. Feminist thinkers have long believed that emotion, and other forms of subjective experiences, are legitimate ways of knowing and producing knowledge about the world. Evelyn Fox Keller, a physicist-turned-philosopher, famously employed the Nobel-prize-winning research of geneticist Barbara McClintock, in order to show how even the most profound of scientific discoveries are generated from a combination of experiment and insight, reason and emotion. And sociologist Patricia Hill Collins describes an ideal knowledge situation as one in which "neither ethics nor emotions are subordinated to reason."

Once we embrace the idea of leveraging emotion in data visualization, we can truly appreciate what sets Periscopic's Gun Deaths apart from the *Washington Post* graphic, or any number of other gun death charts that have appeared in newspapers and policy documents. The *Washington Post* graphic represents death counts as blue ticks on a generic bar chart. If we didn't read the caption, we wouldn't know whether we were counting gun deaths in the U.S., or haystacks in Kansas, or exports from Malaysia, or any other semi-remote statistics of passing interest. But the *Periscopic* visualization leads with loss, grief, and

[1]The concepts taught address specific mathematical content and skills outlined by the Common Core State Standards in New York.

mourning. It provides a visual language for representing the years that could have been– numbers that are accurate, but not technically facts. It uses pacing and animation to help us appreciate the scale of one life, and then compounds that scale 11,419-fold. The magnitude of the loss, especially when viewed in aggregate, is a staggering and profound truth– and the visualization helps recognize it as such through our own emotions. The generic WaPo bar chart cannot do the work of communicating this truth.

Skilled data artists and designers know these things already, and are pushing the boundaries for what affective and embodied data visualization could look like. In 2010, Kelly Dobson founded the Data Visceralization research group at the Rhode Island School of Design (RISD) Digital + Media Graduate program. The goal for this group was not to visualize data but to *visceralize* it. Visual things are for the eyes, but visceralizations are data that the whole body can experience– emotionally, as well as physically.

The reasons for doing visceralizing data have to do with more than simply creative experimentation. How do visually impaired people access charts and dashboards? According to the World Health Organization, 253 million people globally live with some form of visual impairment. This might include cataracts, glaucoma and complete blindness. Creators in the visceralization mode have crafted haptic data visualizations,[2] data walks, data quilts, musical scores from scientific data, wearable objects that capture your breaths and play them back later, and data performances. These types of objects and events are more likely to be found in the context of galleries and museums and research labs, but there are many lessons to be learned from them for those of us who make visualizations in more everyday settings.

For example, in the project "A Sort of Joy (Thousands of Exhausted Things)", a theater troupe joined with a data visualization firm to craft a live performance based on metadata about the artworks held by New York's Museum of Modern Art. With 123,951 works in its collection, MoMA's metadata consists of the names of artists, the titles of artworks, their media formats, and their time periods. But how does an artwork make it into the museum collection to begin with? Major art museums and their collection policies have long been the focus of feminist critique because the question of whose work gets collected translates into the question of whose work is counted in the annals of history– and, as you might guess, this history has mostly consisted of a parade of white male "masters."

In 1989, for example, the Guerrilla Girls, an anonymous collective of female artists, published what we would today call an infographic: Do women have to be naked to get into the Met. Museum? The graphic was designed to be displayed on a billboard. However, it was rejected by the sign company because it "wasn't clear enough." (If you ask us, it's pretty clear). The Guerrilla Girls then paid

[2]CS109 at Harvard is taught jointly by Computer Science and Statistics. As of this writing, there are 37 male faculty (69%) and 17 female faculty (31%).

Figure 2.2: An infographic (of a sort) created by the Guerrilla Girls, intended to be displayed on a billboard. ¶ Credit: The Guerrilla Girls ¶ Source: https://www.nga.gov/collection/art-object-page.139856.html ¶ Permissions: PENDING

for it to be printed on posters which were displayed throughout the New York City bus system, until the bus company cancelled their contract, stating that the figure "seemed to have more than a fan in her hand." (It is definitely more than a fan). The figure is certainly provocative, but the poster also makes a data-driven argument by tabulating gender statistics for artists included in the Met collection, and comparing them to the gender stats for the *subjects* of art included in the collection. As per the poster, the Met readily collects paintings in which women are the subjects, but not those in which women are the artists themselves.

A Sort of Joy tackles a similar subject, but with wholly different tactics. The performance starts with a group of white men standing in a circle in the center of the room. They face out towards the audience, which stands around them. The men are dressed like stereotypical museum visitors: collared shirts, slacks, etc. Each wears headphones and holds an iPad on which the names of artists in the collection scroll by. "John," the men say together. We see the iPads scrolling through all of the names of artists in the MOMA collection whose first name is John: John Baldessari, John Cage, John Lennon, John Waters, and so on. Three female performers, also wearing headphones and carrying iPads with scrolling names, pace around the circle of men,. "Robert," the men say together, and the names scroll through the Roberts alphabetically. The women are silent and keep walking. "David," the men say together. It soon becomes apparent that the artists are sorted by first name, and then ordered by which first name has the most works in the collection. Thus, the Johns and Roberts and Davids come first, because they have the most works in the collection. But Marys have fewer works, and Mohameds and Camilas are barely in the register. Several minutes later, after the men say "Michael", "James", "George", "Jean", "Hans", "Thomas", "Walter", "Edward", "Yan", "Joseph", "Martin", "Mark", "José", "Louis", "Frank",

"Otto", "Max", "Steven", "Jack", "Henry", "Henri", "Alfred", "Alexander", "Carl", "Andre", "Harry", "Roger" and "Pierre", "Mary" finally gets her due. It's spoken by the female performers; the first sound they've made.

For audience members, the experience starts as one of slight confusion. Why are there men in a circle? Why do they randomly speak someone's name? And what are those women walking around so intently? But "Mary" becomes a kind of a-ha moment– the same that researcher Robert Kosara says that data visualization is so good at producing – when the highly gendered nature of the collection is revealed. From that point on, audience members start to listen differently, eagerly awaiting the next female name. It takes more than three minutes for "Mary" to be spoken, and the next female name, "Joan," doesn't come for a full minute longer. "Barbara" follows immediately after that, and then the men return to reading, "Werner", "Tony", "Marcel", "Jonathan".

From a data analysis perspective, "A Sort of Joy" consists of simple operations: only counting and grouping. The results could easily have been represented by a bar chart or a tree map of first names. But rendering the dataset as a time-based experience makes the audience wait and listen. It also contradicts long-held wisdom in visualization design, as expressed by Ben Shneiderman in the mid-1990s: "Overview first, zoom and filter, then details-on-demand." Instead, in this data performance, we do not see "the whole picture". We hear and see and experience each datapoint one at a time. The different gender expressions, body movements, and verbal tones of the performers draw our collective attention to the issue of gender in the MoMA collection. We start to anticipate when the next female name will arise, and begin to speculate on whether it will be a Rosa, a Rhonda, or another name entirely. We *feel* the gender differential, rather than *see* it. This feeling is affect. It comprises the emotions that arise when experiencing the performance and the physiological reactions to to the sounds and movements made by the performers, as well as the desires and drives that result– even if that drive is to walk into another room because the performance is disconcerting or just plain long.

Designing data visceralizations requires a much more holistic conception of the viewer. The viewer isn't just a pair of eyes attached to a brain. They are a whole body– a complex, feeling one. Theirs is a body located in space, with a history and a future. This notion of the visceral viewer, and of the additional knowledge that can be conveyed when designing with visceralization in mind, can be found in many current projects, even if they don't always describe their work in those terms. For example, Catherine (one of the authors of this book) and artist Andi Sutton led walking tours of the future coastline of Boston based on sea level rise. And Lauren (the other author) and her team of Georgia Tech undergrads recreated Elizabeth Peabody's living-room-rug-sized charts from the 19th century using touch sensors and individually addressable LEDs. Mikhail Mansion made a leaning, bobbing chair that animatronically shifts based on real-time shifts in river currents. Teri Rueb staged "sound encounters" between the geologic layers of a landscape and the human body that is affected by them.

Simon Elvins drew a giant paper map of silence in London that you can actually listen to. While these projects may seem to be speaking to another part of brain than your standard Sankey diagrams or network maps, there is something to be learned from the opportunities opened up by visceralizing data. In fact, scientists are now proving by experiment what designers and artists have long known through practice: activating emotion, leveraging embodiment, and creating novel presentation forms help people learn more from data-driven arguments, and remember them more fully.

It turns out that visceralizing data may also help designers solve a pernicious problem in the visualization community: how to represent uncertainty in a medium that's become rhetorically synonymous with the truth. Its "truthiness" is both a feature and a bug. One of the best things about data visualization is that it *does* look so certain, so factual, and so authoritative. But why? After doing a sociological analysis, Helen Kennedy determined that four conventions of data visualization reinforce people's perceptions of its factual basis: 1) two-dimensional viewpoints, 2) clean layouts, 3) geometric shapes and lines, and 4) the inclusion of data sources at the bottom. These conventions contribute to the perception of data visualization as objective, scientific and neutral.

But even if you use these conventions with the best and most pure intentions, it's something of a problem for feminist design, because feminist theory maintains that there is no such thing as a purely objective view of the world. Knowledge is always partial, as Sandra Harding has shown us, and these conventions would seem to contradict that basic philosophical tenet. Haraway's "God trick," which we discuss in *Bring Back the Bodies*, is exactly that: a trick to make you believe that you can see everything, all at once, from an imaginary and impossible standpoint.

This is the argument from philosophy. But representing uncertainty is also a known problem in data journalism and visualization research. In these realms, people may care a tiny bit less about feminist epistemology but do care deeply about end users' ability to accurately interpret graphic depictions of data and use them to make decisions. To this end, designers have created a huge array of charts and techniques for quantifying and representing uncertainty. These include box-plots, violin plots, gradient plots, and confidence intervals. Unfortunately, however, people are terrible at recognizing uncertainty in data visualizations, even when they're explicitly told that something is uncertain. According to work by Geoff Cumming and colleagues, researchers themselves have a hard time understanding confidence intervals. And even everyday weather forecasts such as "There's a 30% chance of rain tomorrow" are generally interpreted by the public to mean "It will rain 30% of the time" or "It will rain in 30% of my area."[3]

Emotion is not only useful for communicating uncertainty. Let's return to

[3]This does not mean there are no data ethics courses, only that it is not the norm to address these concerns in introductory coursework. Indeed, there is a long list compiled by Dr. Casey Fiesler of technical courses that specifically address ethics and what is being called "fairness, accountability and transparency" in technical fields: http://bit.ly/tech-ethics-syllabi

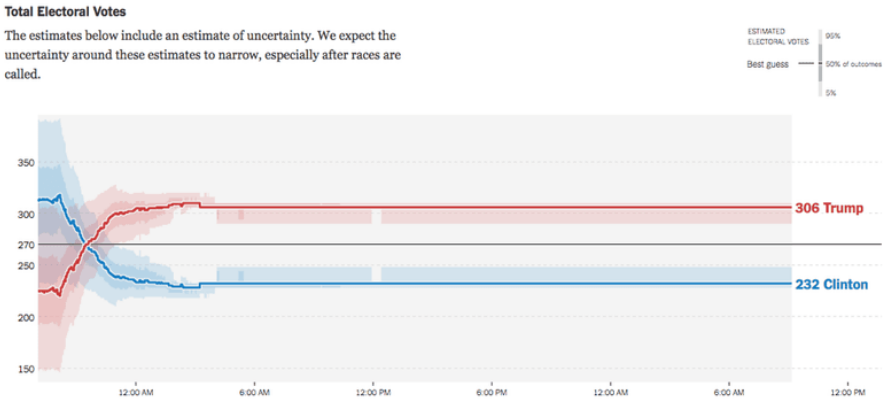


Figure 2.3: < p>A chart from *The New York Times* that uses opacity in order to indicate uncertainty. ¶ Credit: Gregor Aisch, Nate Cohn, Amanda Cox, Josh Katz, Adam Pearce, and Kevin Quealy for *The New York Times* ¶ Source: https://www.nytimes.com/elections/2016/forecast/president

the question around visual minimalism: Are minimal charts really "better," as Edward Tufte has claimed? Charts like "Monstrous Costs," depicted above, have been historically dismissed as "chart junk." In this case, the data has to do with presidential campaign spending from 1972 - 1982. You will notice that there is a very low data to ink ratio– recall Tufte's other major pet peeve. For example, the monster's eyes and body would be deemed as "not data" and thus embellishment. Likewise, there is gratuitous spittle emanating from the monster's mouth, and the tail forms an S-curve that is then turned into a dollar sign with two utterly superfluous grey lines. The monster is also wearing a VOTE badge, which, again, does not represent any aspect of the data at hand. Plus, monsters do not vote– so how irrational is that?

There are many popular blogs devoted to ridiculing this so-called chart junk. But recently, researchers have challenged the notion that chart junk is junk at all. A 2010 study by Scott Bateman and colleagues in Computer Science at the University of Saskatchewan that found that "embellished" charts do not hinder people's ability to accurately read them, and in fact, they are actually better for memorability. When polled 2 to 3 weeks later, people were much more likely to recall the message of the embellished chart over a minimalist chart that displayed the same data. People also thought the junk charts were more attractive, and enjoyed them more (Duh. Who doesn't like monsters better than bar charts?!) Likewise, in 2016, Michelle Borkin and colleagues showed that visualizations that make use of novel presentation styles are more memorable, and therefore more effective. Relating the visual form to the topical content of a chart *works*. So, as data journalist Mona Chalabi says, "If it's about farts, draw a butt for god's sakes."

After all this, you might be wondering, "So why is it that people like junk so much?" But that is not really the right question. The question we should all be asking is, "How has the view from an imaginary and impossible standpoint– the 'whole picture,' the 'overview,' – come to be seen as rational and objective at all?" The rational, scientific, objective viewpoint actually comes from a mythical, imaginary, impossible standpoint. The view from no body. "The god trick of seeing everything from nowhere," as Donna Haraway says it.

But let's say it more simply, "How did we arrive at conventions in data visualization that prioritize rationality, devalue emotion, and completely ignore the human body except for two eye stalks attached to a brain?" Any knowledge community inevitably places certain things at the center and casts others out, in the same way that male bodies have been taken as the norm in scientific study and female bodies imagined as deviation from the norm, or that rationality has been valued as an authoritative mode of communication and emotion cast out. But, following feminist theorist Elizabeth Grosz, what is regarded as "excess" in any given system might possibly be the most interesting thing to explore because it tells us the most about what *and who* the system is trying to exclude.

In the case of data visualization, this excess is emotion and affect, embodiment and expression, embellishment and decoration. These are the aspects of human experience coded "female," and thus devalued by the logic of our master stereotype. But how might we intentionally flip this? All design fields, including visualization, are fields of possibility. We must actively strive to question what (and who) is at the center of the discourse of our field, what (and who) is at the periphery. And then we must work to center the things and people that have been cast off. The first step in this re-centering process is to legitimize affect and embodiment in data visualization. This means to moving emotion to the center of visualization design, and to start to imagine data experiences for whole human bodies in all of their glorious, situated, uncontainable excesses.

Chapter 3

Chapter Three: "What Gets Counted Counts"

"Sign in or create an account to continue." These may be the most unwelcome words on the internet. For most who encounter them, these words elicit a groan– and the inevitability of yet another password that will soon be forgotten. But for people like Maria Munir, the British college student who famously came out as gender non-binary to (then) President Barack Obama on live TV, the prospect of creating a new user account is more than an annoyance. "I wince as I'm forced to choose female over male every single time, because that's what my passport says, and... being non-binary is still not legally recognised in the UK," Munir explains.

For the estimated 9 to 12 million gender non-binary people in the world-- that is, people who are not *either* male *or* female-- the seemingly simple request to "select gender" can be difficult to answer, if it can be answered at all. Yet when creating an online user account, not to mention applying for a passport, the choice between "male" or "female," and only "male" or "female," are almost always the only options. These options (or the lack thereof) have consequences, as Munir clearly states: "If you refuse to register non-binary people like me with birth certificates, and exclude us in everything from creating bank accounts to signing up for mailing lists, you do not have the right to turn around and say that there are not enough of us to warrant change.

"What gets counted counts," as feminist geographer Joni Seager has asserted, and Munir is one person who understands that. Without the right categories, the right data can't be collected. And increasingly, without the right data, there can be no social change. We live in a world in which "data-driven" decisions are prioritized over anecdotal ones, and "evidence"--Fox News notwithstanding--is taken to mean "backed up by numbers and facts." Now, any self-respecting feminist would be the first to tell you that personal accounts should matter

as much as any meta-study, and "evidence" can take a range of qualitative and quantitative forms. To disagree with those statements would undo the work of the many feminist activists and scholars of the 1980s and early 1990s who struggled to get qualitative methods, such as interviews and participant observations, accepted as legitimate evidence in the first place. But there is, undeniably, what feminist demographers Christina Hughes and Rachel Lara Cohen call a "pragmatic politics" of using quantitative methods for feminist aims. If the goal is to work towards justice, then by all means use whatever form of evidence is most convincing. It would be an injustice not to!

That being said, there is a second argument in favor of quantitative methods that has less to do with pragmatism, and more to do with the nature of the problem at hand. So many issues of structural inequality are problems of scale, and can seem anecdotal until they are seen as a whole. For instance, when Natalie Wreyford and Shelley Cobb set out to count the women involved in the film industry in the UK, they encountered a female screenwriter who had never considered the fact that, in the UK, male screenwriters outnumber women at a rate of four to one. "Isn't that a funny old thing?" she said. "I didn't even know that because screenwriters never get to meet each other."

But it's far less funny when the subject is a matter of life or death, as in ProPublica's reporting on the racial divide in maternal mortality in the United States, which we discuss in *Bring Back the Bodies*. While they interviewed the families of many Black women who had died while giving birth, few were aware that the phenomenon extended beyond their own family. But the racial disparity in maternal health outcomes is indeed a structural problem, and it's why feminist sociologists like Ann Oakley have long advocated for the use of quantitative methods alongside qualitative ones. Without big data, Oakley explains--although she just used the term "quantitative research," since she was writing in 1999--"it is difficult to distinguish between personal experience and collective oppression."

But before issues like the racial divide in maternal mortality, or the structural racism that underlies it, can be identified through large-scale analyses like the one that ProPublica conducted, the data must exist in the first place. Which brings us back to Maria Munir and the importance of collecting data that reflects the population it claims to represent. On this issue, Facebook of all corporations was ahead of the curve when, in 2014, it expanded its gender options from the standard two to over fifty choices, ranging from "genderqueer" to "neither"--a move that was widely praised by a range of gender non-binary groups. One year later, when the company abandoned its select-from-options model altogether, replacing the "Gender" dropdown menu with a blank text field, the decision was touted as even more progressive. Because Facebook users could input any word or phrase in order to indicate their gender, they were at last unconstrained by the assumptions imposed by any preset choice.

But research by Rena Bivens, a scholar of social media, has revealed that, below the surface, Facebook continues to resolve users' genders into one of either male

Figure 3.1: Facebook's initial attempt to allow users to indicate additional genders, ca. 2014. ¶ Credit: Slate ¶ Source: http://www.slate.com/blogs/future_tense/2014/02/13/facebook_custom_gender_options_here_are_all_56_custom_options.html

Figure 3.2: Facebook's updated gender field, ca. 2018. ¶ Credit: Facebook. Screenshot by Lauren F. Klein ¶ Source: http://www.facebook.com

or female. Evidently, this decision was made so that Facebook could allow its primary clients-- advertisers-- to more easily market to one gender or the other. Put another way, even if you can choose the gender that you show to your Facebook friends, you can't change the gender that Facebook's advertisers ultimately see. And this discrepancy leads right back to the body issues we discussed in Chapter One: it's corporations like Facebook, and not individuals like Maria Munir, who control the terms of data collection--even if it's folks like Munir, who have personally (and often painfully) run up against the limits of our current classification systems, who are best positioned to improve them.

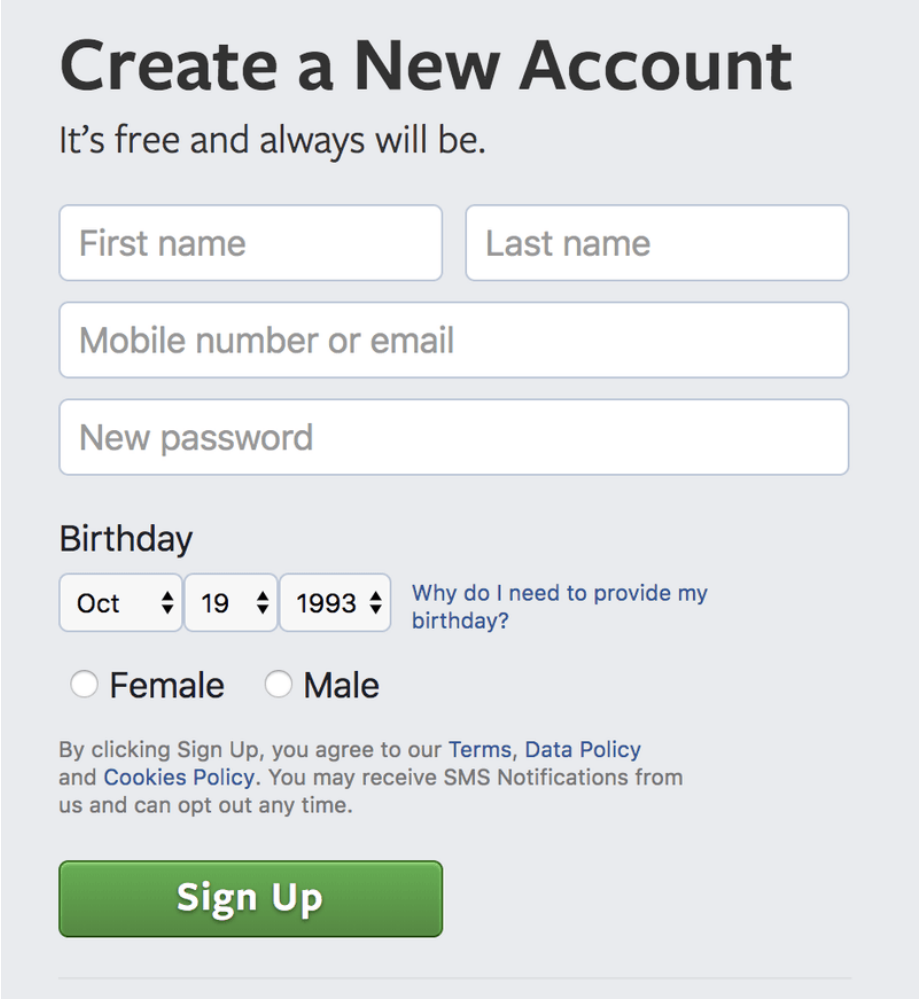


Figure 3.3: Detail of the Facebook new account creation page, ca. 2018. ¶ Credit: Facebook. Screenshot by Lauren Klein. ¶ Source: http://www.facebook.com/

Feminists have also spent a lot of time thinking about classification systems, as it turns out, since the criteria by which people are divided into the categories of "male" and "female" is exactly that: a classification system. And while the gender binary is one of the most universal classification systems in the world today, it is no less constructed than the Facebook advertising platform or, say, the Golden Gate Bridge. The Golden Gate Bridge is a physical structure; Facebook Ads is a virtual structure; and the gender binary is a conceptual one. But each of these structures was created by people: people living in a particular place, at a particular time, and who were influenced--as we all are-- by the world around them.

So this starts to get at the meaning behind the phrase, "gender is a social construct." Our current ideas about the gender binary can be traced to a place (Europe) and a time (the Enlightenment) when new theories about democracy and what philosophers called "natural rights" began to emerge. Before then, there was definitely a gender *hierarchy*, with men on the top and women on the bottom. (Thanks, Aristotle!) But there wasn't a *binary* distinction between those two genders. In fact, according to the historian of gender, Thomas Laqueur, most people believed that women were just inferior men, with penises located inside instead of outside of their bodies, and that-- for reals!-- could descend at any time in life.

For the gender binary to emerge, it would take figures like Thomas Jefferson declaring that all men were created equal, and entire countries (like the U.S.) founded on that principle, before those same figures began to worry what, exactly, they had declared-- and, even more worrisome, to whom it actually applied. All sorts of systems for classifying people date to that era-- not only gender but also, crucially, race. Before the eighteenth century, Western societies understood "race" as a concept tied to religious affiliation, geographic origin, or some combination of both. Although it's hard to believe, race had nothing to do with skin color until the rise of the transatlantic slave trade, in the seventeenth century. Even then, race was still a hazy concept. It would take the so-called "scientific racism" of the mid-eighteenth century for race to begin to be defined in terms of black and white.

Ever heard of Carl Linnaeus? Think back to middle school, when you likely learned about the binomial classification system that he is credited with creating. Well, Linnaeus's revolutionary system didn't just include the category of *homo sapiens*; it also, lamentably--but as historians would tell you, unsurprisingly--included five subcategories of humans separated by race. (One of these five was set aside for mythological humans who didn't exist in real life, in case you're still ready to get behind his science). But Linnaeus's classification system wasn't even the worst of the lot. Over the course of the eighteenth century, increasingly racist systems of classification began to emerge. These were systems that were designed to exclude, and in instances as far-ranging as the maternal health outcomes we've already discussed, to Google search results for "black girls" vs. "white girls," as information studies scholar Safiya Noble has shown, we can detect the

effects of those racist systems every day.

A simple solution would be to say, "Fine, then. Let's just not classify anything, or certainly anyone!" But the flaw in that plan is that data must also be classified in some way in order to be put to use. Data, after all, is information made *tractable*, to borrow a term from computer science (and from another essay that Lauren wrote with a colleague in information studies, Miriam Posner). "What distinguishes data from other forms of information is that it can be processed by a computer, or by computer-like operations," we write there. And in order to enable those operations, which range from counting to sorting, and from modeling to visualizing, the data must be placed into some kind of category--if not always into a conceptual category like "gender," then at the least into a computational category like "integer" (a type of number) or "string" (a sequence of letters or words).

It's been argued that classification systems are essential to any working infrastructure-- and not only to computational infrastructures or even conceptual ones, but also to physical infrastructures like the checkout line at the grocery store. Think about how angry you get when you're stuck in the express line behind someone with more than fifteen items. Or, if that's not something that gets you going, just think of the system you use to sort your clothes for the wash. It's not that we should reject these classification systems out of hand, or even that we could if we wanted to. (We're pretty sure that no one wants all of their socks to turn pink). It's just that we rarely question how classification systems are constructed, or ask why they might have been thought up in the first place. In fact-- and this is a point also made by the influential information theorists Geoffrey Bowker and Susan Leigh Star-- we tend not to even think to ask these questions until our systems break.

Classification systems can break for any number of reasons. They can break when an object-- or, more profoundly, a person-- can't be placed in the appropriate category. They can break when that object or person doesn't want to be placed in an appropriate category. And they can break when that object or person shouldn't even be placed in a category to begin with. In each of these cases, it's important to ask whether it's the categories that are broken, or whether-- and this is a key feminist move-- it's the system of classification itself. Whether it's the gender binary, or the patriarchy, or-- to get a little heady-- the distinction between nature and culture, or reason and emotion, or public and private, or body and world, decades of feminist thinking would tell us to question why these distinctions might have come about; what social, cultural, or political values they reflect; and, crucially, whether they should exist in the first place.

But let's spend some time with at an actual person who has done this kind of thinking: one Michael Hicks, an eight-year-old Cub Scout from New Jersey. Why has this kid started to question the broken systems of classification that surround him? Well, Mikey, as he's more commonly known, shares his first and last name with someone who has been placed on a terrorist watch list by the U.S. federal government. As a result, Mikey is subjected to the highest level of airport

security screening each time that he travels. “A terrorist can blow his underwear up and they don’t catch him. But my 8-year-old can’t walk through security without being frisked,” his mother lamented to Lizette Alvarez, a reporter for *The New York Times*, who covered the issue in 2010.

Of course in some ways, Mikey is lucky. His is white, so he does not run the risk of racial profiling. (Any number of Black women can tell you how many times they’ve received a pat-down only because of their hair). Moreover, Mikey’s name is not Muslim-sounding, so he does not need to worry about religious or ethnic profiling either. (Anyone in the U.S. named Muhammad can tell you how many times they’ve been pulled over by the police). But Mikey the Cub Scout still helps to expose the brokenness of the categories that structure the TSA’s terrorist classification system; the combination of first and last name is simply insufficient to classify someone as a terrorist or not.

Or, consider another person with a history of bad experiences at the (literal) hands of the TSA: Sasha Costanza-Chock. Costanza-Chock is, like Maria Munir, gender non-binary. They are also a design professor at MIT, so they have a lot of experience not only living with, but also thinking through broken classification systems. In a recent essay, they describe how the seemingly simple system employed by the operators of those hand-in-the-air millimeter-wave-scanning machines is in fact quite complex-- and also fundamentally flawed.

No one but a gender non-conforming person would know that, before you step into a scanning machine, the TSA agent operating the machine looks you up and down, decides whether you are male or female, and then pushes a button to select the appropriate gender on the scanner’s touch-screen interface. That decision loads the algorithmic profile for either male bodies or female ones, against which your measurements are compared. If your body’s measurements diverge from the statistical norm of that gender’s body-- whether the discrepancy is because you’re concealing a deadly weapon, or because the TSA agent just made the wrong choice-- you trigger a “risk alert,” and are subjected to the same full-body pat-down as a potential terrorist. So here it’s not that the scanning machines rely upon an insufficient number of categories, as in the case of Mikey the Cub Scout; or even that they employ the wrong ones, as Mikey’s mom would likely say. It’s that the the TSA scanners shouldn’t rely on the category of gender to classify air-travelers to begin with.

So when we say that what gets counted counts, it’s folks like Costanza-Chock, or Mikey, or Maria Munir, that we’re thinking about. Because broken classification systems like the one that underlies the airport scanner’s risk detection algorithm, or the one that determines which names end up on terrorist watch lists, or simply (simply!) the gender binary, are often the result of larger systems that are themselves broken, but that most people don’t often have the opportunity to see. These invisible systems are what philosopher Michel Foucault would call systems of power. Systems of power don’t simply determine the categories into which individual objects or people are sorted; they over-determine how those groups of objects or people experience the world.

What does it mean for a system to over-determine how people experience the world? Many feminists would point to the example of the patriarchy--a word that describes the combination of legal frameworks, social structures, and cultural values that contribute to the continued male domination of society. But for a more concrete example, we could return to Facebook. It's not only that anyone who types in a gender that is not "male" or "female" is reduced, in the eyes of advertisers, to the single category of " unknown." It's also that, at the level of code, these three remaining categories are further reduced to numerical values: 1, 2, and 3, respectively. So when an app developer requests a list of users sorted by gender for any reason-- whether it's to sell them useless diet pills, 50% off retail (which no one ever wants); or to offer them a free financial consultation, first come first served (which many people do)-- they receive a list in which male Facebook users are hard-coded to be always first in line.

Now, the software engineers who wrote the word-to-number code were almost certainly not intending to discriminate. They were probably only thinking, "How can we make our gender data easier to sort and manage?" And when it comes to computational data, it's almost always easier and more efficient to deal with numbers than it is to deal with words. But it's also not a surprise that in a group of engineers which is a reported 87% male, no one thought to point out (or maybe just that no one felt comfortable saying out loud) that a data classification system in which men are always ranked first might lead to problems for those who ranked second or third-- not to mention those excluded from the list altogether. In fact, if you were to ask a feminist theorist like Judith Butler to weigh in, she'd tell you that the inadvertent and invisible way in which systems of power reproduce themselves is exactly how the gender binary consolidates its force.

It's not only Facebook that's to blame. Gender data is almost always collected in the binary categories of male and female, and visually represented by some form of binary as well. This remains true even as a recent Stanford study found that, when given the choice among seven points on a gender spectrum, more than two-thirds of the subjects polled placed themselves somewhere in the middle. It's also important to remember that there have always been more variations in gender identity than Anglo-Western societies have cared to outwardly acknowledge or collectively remember. These third, fourth and n-th genders go by different names in the different historical and cultural circumstances in which they originate, including *female husbands*, *indigenous berdaches*, *Hijras*, *two-spirits*, *pansy performers*, and *sworn virgins,* along with the category of *transgender* that we most commonly use today.

Now, as data analysts and visualization designers, we can't always control the collection process for the data we use in our research. Like the Facebook engineers, we're often working with data that we've obtained from someplace else. But even in those cases--and, arguably, especially in those cases--it's important to ask how and why the categories of the dataset we're using were constructed, and what systems of power they might represent. Because when it comes to

classification systems, there's power up and down, side to side, and everywhere in between. And it's on us, as data feminists,to ensure that any differentials of power that are encoded in our datasets don't continue to spread.

Whether we like it or not, we're all already swayed by these systems of power, as well as by the heuristic techniques that reinforce them. Before you say, "Wait! No one taught *me* those techniques!" consider that "heuristic techniques" is just a fancy term for the use of mental shortcuts to make judgements--in other words, common sense. The tendency of people to adhere to common sense offers a great evolutionary advantage, in that it's enabled humanity to survive over many millennia. (What tells you to run away from a bear? Common sense! What tells you not to eat rancid meat? Also common sense (and your gag reflex)). But as the renowned work of cognitive psychologists Daniel Kahnemann and Amos Tversky has showed, this reliance on heuristics eventually leads to an accumulation of *cognitive biases*--what might be otherwise understood as a snowball of mistaken assumptions that, in a world more challenged by structural inequalities than by grizzly bears, leads to profoundly flawed decision-making, and equally profoundly flawed results.

Buster Benson, a product manager at the crowd-funding platform Patreon, has made a hobby of classifying these cognitive biases, and with John Manoogian, has visualized them in the chart you see above. If you look at the lower half of the image, you see can see the two quadrants-- "Need to Act Fast" and "Not Enough Meaning" --that include some of the key cognitive biases that come into play when collecting and classifying data.

Now imagine, for a moment, that you are designing a new survey for an analysis of gender and cell phone usage, but you have not yet finished reading this book. Gender is something you are pretty familiar with, you might say to yourself, since you have a gender, and everyone else you know has a gender too. But this is called the *overconfidence effect*, found on the lower left of the chart in lime green. Still, you go on: in your experience there are two genders, male and female, and everyone else you know would say so, too. (This is called the *false-consensus effect*, also on the lower left). Men and women should clearly be placed in separate categories, since they are different kinds of people. (This is called *essentialism*; file under "Not Enough Meaning"). Also, everyone knows that women like talking—*stereotyping* alert!—so in addition to gender data, how about collecting cell phone minutes data too. (You've just committed a *fundamental attribution error*, in blue on the right).

Fast forward past the data collection phase to the analysis portion of the project. You note that you were right in your initial assessment of the situation: women did talk on their cell phones more than men. This forms the basis of your subsequent analysis. (This is called *confirmation bias*). In addition, in your zeal to confirm your essentialist beliefs, you entirely missed an important phenomenon: millennial-aged people of all genders have extremely large social networks. Your *expectation bias* prevented you from discovering some important insights that might have informed the design of a new product. You receive a negative

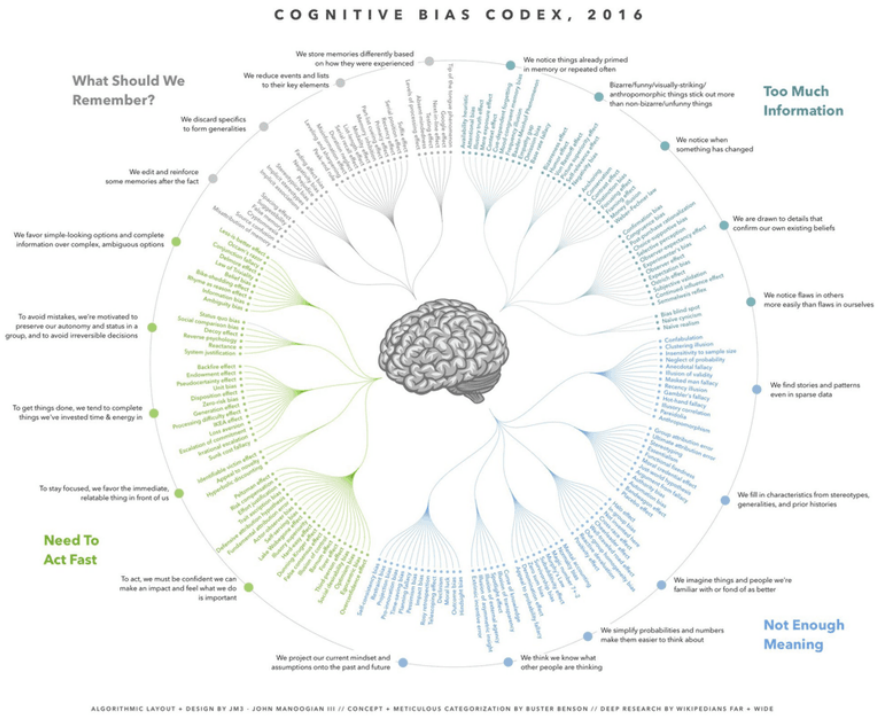


Figure 3.4: The Cognitive Bias Codex groups known cognitive biases into four different categories. ¶ Credit: Design by John Manoogian III based on grouping by Buster Benson. ¶ Source: https://commons.wikimedia.org/wiki/File:The_Cognitive_Bias_Codex_-_180%2B_biases ,_designed_by_John_Manoogian_III_(jm3).png

performance review, and you are fired.

What interrupts this series of bad decisions? Recognizing that common sense is often sexist, racist, and harmful for entire groups of people--especially those groups, like women, who find themselves at the bottom end of a hierarchical classification system; or like non-binary folks, who are excluded from the system altogether.

As should now be clear, a feminist critique of classification systems is not limited to data about women, or to the category of gender alone. This point can't be overstated, as it forms the basis for the theories of intersectional feminism that inspire this book. Feminist scholars Brittney Cooper and Margaret Rhee address this issue directly in their call to use feminist thinking to "hack" the binary logic that simultaneously underlies the racism experienced by Black people in the United States, and erases the other forms of racism experienced by Latinx, Asian American, and Indigenous groups. "Binary racial discourses elide our struggles for justice," they state clearly, and we agree. By hacking the binary distinctions that erase the experiences of certain groups, as well as the systems of power that position those groups against each other, we can work towards a more just and equitable future.

Even though the stakes of this project are high, it's possible for anyone, including you, our readers, to contribute. One of the best visualizations of the concept of intersectionality that we've found, for instance, comes from a series of posts on anonymously-authored WordPress blog. "Intersectionality, Illustrated" offers a series of visualizations that employ color gradients to represent the multiple axes of privilege (or the lack thereof) that a person might encounter in the world. At the center of each visualization is a solid circle, which represents that person's goals and dreams for their life. Colorful lenses spiral out from the center, each representing an aspect of that person's identity: ethnicity, age, sexual orientation, and so on. In this visualization, opacity is employed to show whether a particular identity trait contributes to an enhanced capacity to achieve one's personal goals, or a diminished one. A directional gradient underscores how that trait alternately supports the person's goals, or distances them from them. In this way, the viewer begins to literally see how an intersection of privileged positions-- a term used to describe the advantages offered only particular groups, such as those that come along with being white, male, able-bodied, or college-educated--can lead to an array of colorful options for the future. An intersection of disadvantaged positions, on the other hand, such as being gay, or transgender, or disabled, or poor, reduces– and, at times, eliminates altogether– that person's ability to pursue a particular life path. It's a simple visualization, which relies only upon the creative use of color, opacity, gradient, and form, and yet it illustrates a powerful point: that one's identity, and therefore one's privilege, is determined by multiple factors that all intersect.

In addition to the intersection of the various aspects of a person's identity, each individual aspect can be quite complex. Again, an anonymous person on the internet offers among the most inspiring examples for considering how we

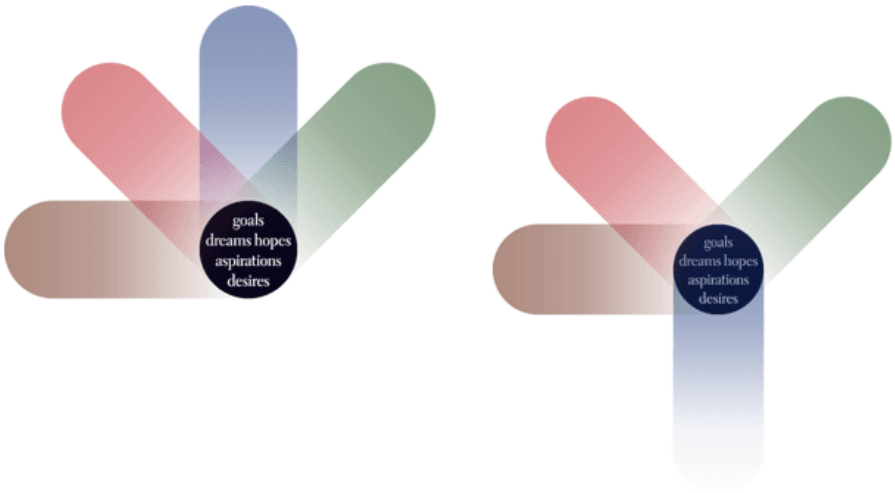


Figure 3.5: Left: "Four intersections, with four intersectional privileges" | Right" Four intersections, with one intersectional barrier." ¶ Credit: Ententa's Magic ¶ Source: https://ententasmagic.wordpress.com/2013/04/14/intersectionality-illustrated/

might visualize gender, for example, if we weren't limited to to the male/female split. The creator of the Non-Binary Safe Space Tumblr shows how gender might be visualized as a spectrum, or as a branching tree. They sketch out how non-binary genders might be placed around a circle, in order to emphasize shared sensibilities rather than differences; or plotted on a Cartesian plane, in which "male" and "female" serve as the axes, with infinite points in between. They even wonder about designing a series of interactive sliders, with "female" and "not female," "male" and "not male," and "other" and "not other," serving as the respective poles; or even a 3D cube, with a vector charting a person's changing course through their evolving sense of self. These are designs that, like "Intersectionality, Illustrated," come from personal experience, and they offer a powerful point of departure for thinking through new classification systems and visualization schemes.

When we went to track down the permissions for the Non Binary Safe Space Tumblr, we discovered that the site had been taken over by spammers. But maybe it's a sign of the times (along with the inevitable descent into spam) that some of these ideas have already begun to enter major publications. For example, when Amanda Montañez, a designer for *Scientific American,* was tasked with creating an infographic to accompany an article on the evolving science of sex and gender, she envisioned a spectrum not unlike the one pictured above. But she soon found confirmation of what feminist theorists have been saying for decades (and what we've been saying so far in this book): that sex and gender

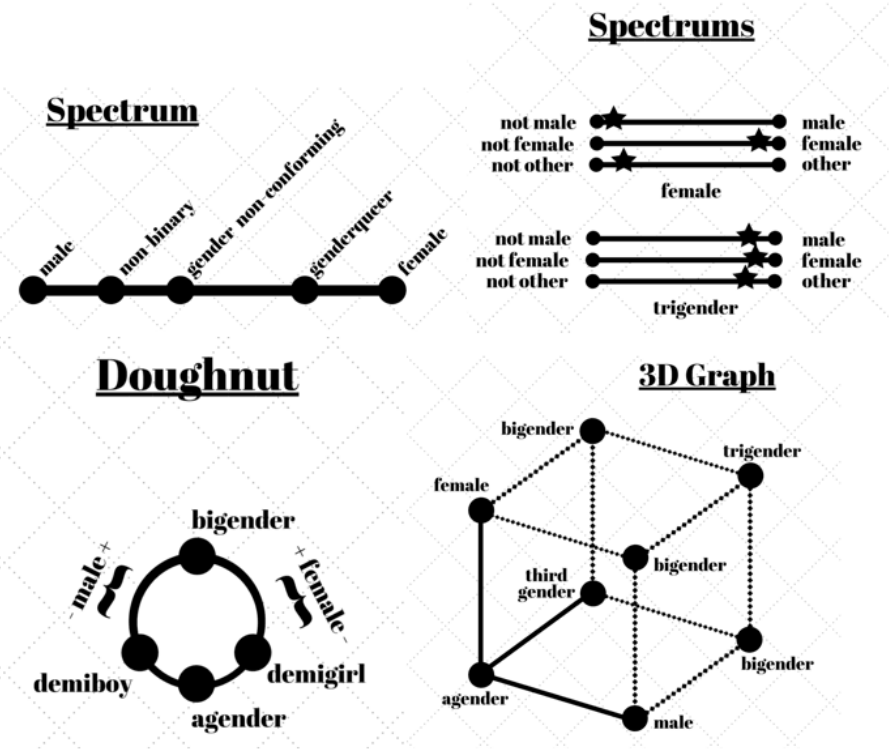


Figure 3.6: Caption: Four different ways to visualize gender: spectrum, spectrums, donut, cube. ¶ Credit: Non Binary Safe Space Tumblr ¶ Source: https://nonbinarysafespace.tumblr.com/

are not exactly the same thing. More than that, what we might think of as the easier concept to explain--the biological category of sex--is just as fluid and complicated as the social category of gender.

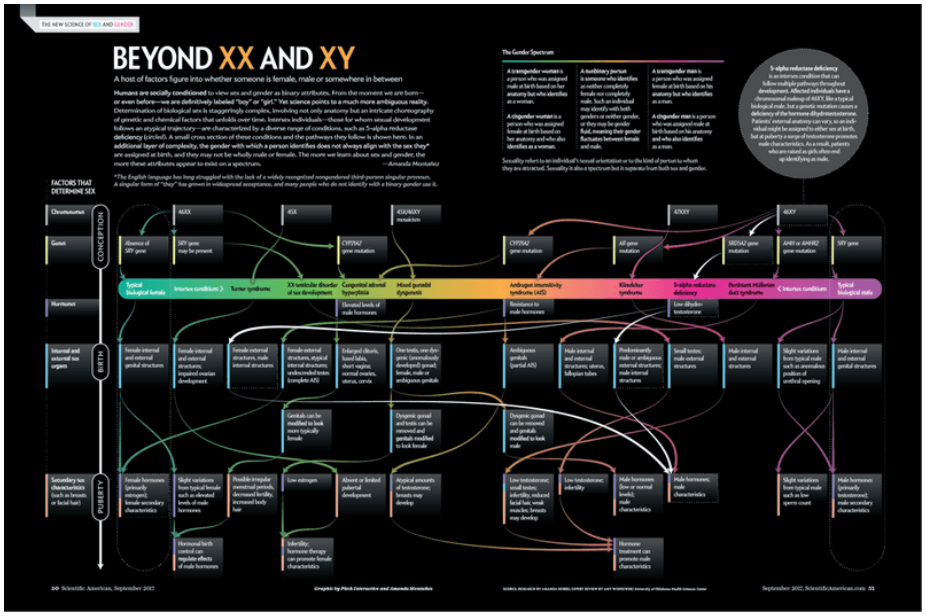


Figure 3.7: Visualizing Sex as a Spectrum ¶ Credit: Credit: Pitch Interactive and Amanda Montaez; Source: Research by Amanda Hobbs; Expert review by Amy Wisniewski *University of Oklahoma Health Sciences Center* ¶ Source: https://blogs.scientificamerican.com/sa-visual/visualizing-sex-as-a-spectrum/ ¶ Permissions: PENDING

The result, "Beyond XX and YY," a collaboration between Montañez and the design firm Pitch Interactive, is a complex diagram, which employs a color spectrum to represent the sex spectrum, a vertical axis to represent change over time, and branching arrows to connect to text blocks that provide additional information. Montañez hopes that visualization, with its careful adherence to terminology, and inclusion of only properly categorized data, will help "raise public awareness" about intersex as well as transgender and non-binary people, and "help align policies more closely with scientific reality, and by extension, social justice." In other words, Montañez made what was already counted count.

Even when working with binary gender data, designers can still make those limited categories count. For example, in March 2018, when the reporters on the Lifestyle Desk of *The Telegraph*, a British newspaper, were considering how to honor International Women's Day, they were struck by the significant gender gap in the UK in terms of education, politics, business, and culture. They didn't have the time or the expertise to collect their own data, and even if they had,

there's no telling as to whether they would have collected non-binary gender data. But they wanted to ensure that they didn't further reinforce any gender stereotypes. They paid particular attention to color, with the awareness that even as many designers are moving away from using pink for girls and blue for boys, most still adhere to the logic that associates warm colors with women and girls, and cool colors with men and boys. Because the stereotype that women are warmer and more caring, while men are cooler and more aloof, is still firmly entrenched in many cultures, the associated colors are easier to interpret—or so this argument goes.

This stereotype is, of course, another hierarchy, and the goal of the *Telegraph* team was to mitigate inequality, not reinforce it, and so they took a different source for inspiration: the "Votes for Women" campaign of early 20th century England, in which purple was employed to represent freedom and dignity, and green to represent hope. When thinking about which of these colors to assign to each gender, they took a design principle as their guide: "Against white, purple registers with far greater contrast and so should attract more attention when putting alongside the green, not by much but just enough to tip the scales. In a lot of the visualizations men largely outnumber women, so it was a fairly simple method of bringing them back into focus," Fraser Lyness, the Telegraph's Director of Graphic Journalism told Lisa Charlotte Rost, herself a visualization designer who interviewed Lyness for her blog. Here, one hierarchy, the hierarchy in which colors are perceived by the eye—was employed to challenge another one—the hierarchy of gender. Lyness was right. It was a "fairly simple method" to employ. But when put into practice, it had profound results.

There are all sorts of instances of designers, as well as journalists, artists, activists, and scholars, using data and design to bring issues of gender into view. P. Gabrielle Foreman and her team at the University of Delaware are creating a historical dataset of women who would otherwise go uncounted, and therefore unrecognized for their work. The team's focus is on the women who attended but were not named as participants in the nineteenth-century Colored Conventions, organizing meetings in which Black Americans, fugitive and free, met to strategize about how to achieve educational, economic, and legal justice. Because these women often worked behind the scenes, packing the lunches and watching the children so that their husbands could attend; running the boarding-houses where out-of-town delegates stayed during the conventions; or even, as research has shown, standing in the back of the meeting hall in order to make their presence known, their contributions were not considered as participation in the events. But as continues to be true today--think back to the issue of maternal mortality mentioned at the beginning of this chapter, or to the issue of sexual assault, as we discuss more in *The Numbers Don't Speak for Themselves*—the systems of power that place women below men in patriarchal societies such as ours are the same that ensure that the types of contributions that women make to those societies are valued less, and therefore less likely to be counted.

But counting is not always an unmitigated good. Sometimes counting can have

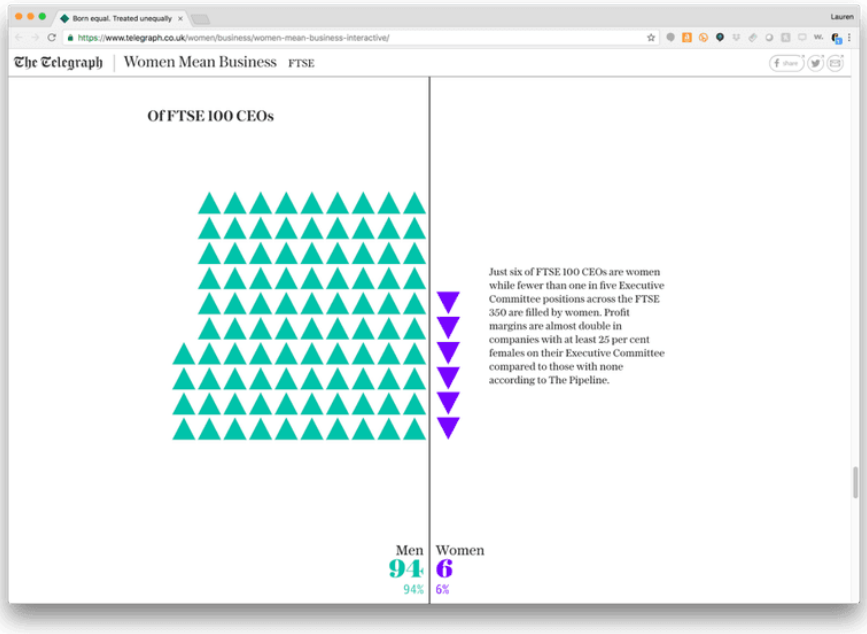


Figure 3.8: The gender divide among CEOs in the UK's FTSE 100 list. ¶ Credits: Claire Cohen, Patrick Scott, Ellie Kempster, Richard Moynihan, Oliver Edgington, Dario Verrengia, Fraser Lyness,George Ioakeimidis, Jamie Johnson ¶ Source: https://www.telegraph.co.uk/women/business/women-mean-business-interactive/

unintended consequences, really bad ones, especially for marginalized groups. Some transgender people, for example, prefer not to disclose the sex they were assigned at birth, keeping their identity as a trans person private. Even for those who generally choose to make their trans identity public, being visibly identified as trans on a map, or in a database, for example, could expose them to violence. Even in a big dataset, there is no additional strength in numbers. Compared to cisgendered people (folks whose genders match the sex they were assigned at birth, trans people are so small a group that they are more exposed, and therefore more vulnerable.

A similar paradox of exposure is evident among undocumented immigrants; visualizing the precise locations of undocumented immigrants may, on the one hand, help make an argument for directing additional resources to a particular area, but on the other, may alert ICE officials of the locations of their homes or schools, making the threat of deportation more likely. In cases where lives are at stake, and the security of the data can't be guaranteed, not collecting statistical outliers can be the best way to go, as Catherine has argued in some of her other work. In other cases, however, the decision to exclude outliers can be viewed as "demographic malpractice," since it completely erases the record of those whose experiences are already marginalized in their everyday lives, and forecloses any future analysis for good or ill.

Is there any way out of this paradox? Feminist geographer Joni Seager has studied this issue for decades, and in 2004, experienced its effects firsthand when she began what she thought would be an easy project: making a map of female doctors for her monumental *Altas of Women in the World*. But she hit a wall when she discovered that the World Health Organization data on medical professionals did not include a field for gender. Seager had to abandon the map, and as a result, she could not include any information about female doctors in her *Atlas*. Ever since, her approach has been to always collect gender data according to the most precise possible categories, and also to always ask-- before the analysis phase-- whether the data should be aggregated or otherwise anonymized in order to mask any potential adverse effects.

Seager's research is focused on the collection practices associated with global and nation-wide data, where she has found that gender data is often collected but rarely made available or analyzed in disaggregated form. For example, in 2015, the Pew Research Center published a report about cellphone use in Africa. "Cell Phone Ownership Surges in Africa," was the title of the report; and the first chart showed the growth in cell phone ownership in the United States compared with several African countries. But buried in the text of report was a surprising finding: "Men are more likely than women to own a cell phone in six of the seven countries surveyed." Now, this would seem like an important distinction-- and perhaps one tied to other inequities-- but because gender was not treated as a primary category of analysis, those who didn't read the fine print might not come way with one of its most important findings. In the case of the Cell Phone study, it wasn't a question of what got counted that turned out to matter, but

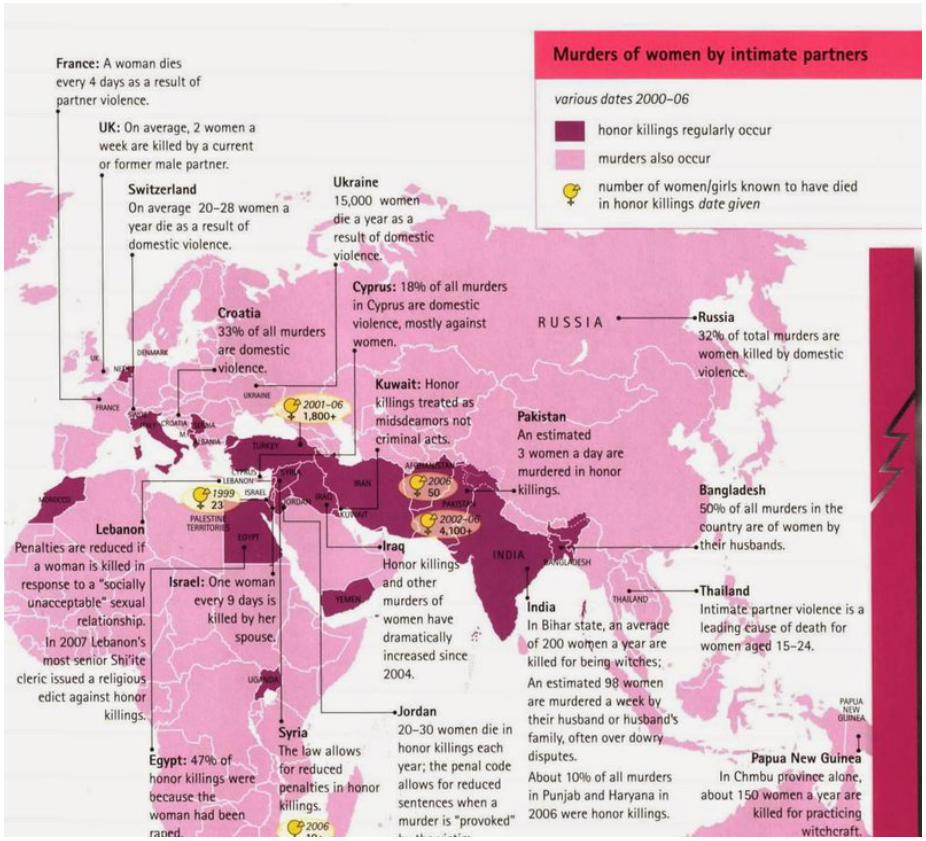


Figure 3.9: A map from Joni Seager's *State of Women in the World* which depicts countries where honor killings occur. ¶ Credit: Joni Seager, *State of Women in the World*, 2nd ed. (1997) ¶ Source: http://bsumaps.blogspot.com/2014/05/maps-in-news.html

how that counting was put to use.

Sometimes, however, questions about counting shouldn't be answered by the survey designer, or by the data analyst, or even by the most careful reader of this book. As a final example helps to show, questions about counting often go hand-in-hand with questions of consent. Flash back to another era -- 2006 -- when another debate about a border wall was underway. Its source was the Secure Fence Act, a bill signed into law by then President George W. Bush, which authorized the construction of 700-mile-fence along the US-Mexico border. But for the fence to be completed, it would have to pass through the Tohono O'odham Nation, which straddles both countries. Recognizing that they would have to build around several sacred burial sites, the U.S. Government requested that the O'odham nation provide them with the locations of those remains.

Figure 3.10: Ofelia Rivas is a Tohono O'odham elder who fought against the US government erecting a fence that cut her nation in half. ¶ Credit: Catherine D'Iganzio ¶ Source: Catherine D'Ignazio

In O'odham tradition, however, the locations of burial sites constitute sacred knowledge, and cannot be shared with outsiders under any circumstances. The O'odham Nation refused to violate its own laws by divulging information about its burial sites to the U.S. government, but it could not oppose the legal or political power of the United States. The United States built the fence, unearthing many O'odham remains in the process, and the tribe spent months attempting to get

the US to return them.

But why should it be assumed that the O'odham Nation, which has existed for thousands of years, weigh its own laws less heavily than those of the United States, which-- after all-- has existed for less than two hundred fifty? Who has the right to demand that information be made public, and who has the right to protect it? And what are the cultural assumptions-- and not just the logistical considerations-- that go along with making knowledge visible and information known.

We've all heard the phrase "knowledge is power," and the example of the border wall shows how this is undeniably true. But the range of examples in this chapter, we hope, also help to show how knowledge can be used to contest power, and to begin transform it. By paying attention to the politics of data collection, and to the systems of power that influence how that data is collected, we can work to rebalance some of the relationships that would otherwise contribute to their force. We might look to large institutions like the National Library of New Zealand, which began the Ngā Upoko Tukutuku Reo Māori Working Group to develop new subject headings for the Māori materials in its collections, ensuring that those materials would be classified in terms of subjects that make sense within a Māori worldview. We might look to small research groups like Mobilized Humanities, which aggregated and visualized dozens of public data sets relating to the U.S.'s "Zero Tolerance" policy, in order to call attention to the humanitarian crisis that unfolded along the U.S./Mexico border in Summer 2018. We might look to individual artists like Caroline Sinders, who is developing a data set of intersectional feminist content that can be used to train the next generation of feminist AI. Or we might look to distributed movements like #SayHerName, which employed that Twitter hashtag to create a digital record of the police violence against Black women that would otherwise go unrecorded.

These are each projects that recognized that what gets counted counts, and how the act of counting, and how we decide to show our results, profoundly influences the ideas we're able to take away. An intersectional feminist approach to counting, like the one we've demonstrated here, insists that you always ask questions about the categories that structure your data, and the systems of power that might, in turn, have structured them.

Chapter 4

Unicorns, Janitors, Ninjas, Wizards, and Rock Stars

In Spring 2017, Bloomberg News ran an article with the provocative title "America's Rich Get Richer and the Poor Get Replaced By Robots". Using census data, the authors report 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. It has the lowest proportion of children for any major US city and a growing rate of evictions since 2003.

While the San Francisco Rent Board collects data on these evictions, it does not track where people go, how many end up homeless or which landlords and developers are systematically evicting major blocks of the city. This is where the Anti-Eviction Mapping Project (AEMP) stepped in, starting in 2013. Led by two women—Erin McElroy and Terra Graziani—the project is a self-described collective of "housing justice activists, researchers, data nerds, artists, and oral historians." They are mapping eviction, but they are doing so through a collaborative, multimodal, and yes, quite messy, process.

If you visit antievictionmap.com, there isn't one single eviction map. There are a total of 78 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 group has a distinctly fluid, collaborative, and community-based way of working. Some projects originate from within the group. 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 has chapters on such topics as Migration/Relocation, Gentrification and the Prison Pipeline, Indigenous and Colonial Histories, and Speculation. But the majority of the projects happen in collaboration with nonprofits, students, and community-based organizations. For

example, the Eviction Defense Collaborative (EDC) is a nonprofit that represents people who have been evicted in housing court. While the City does not collect data on the race or income of evictees, EDC does collect those demographics, and they work with 90% of evicted tenants in the city. In 2014, they approached AEMP to help produce their annual report [1] and in return offered to share their demographic data with the organization. Since then, the two groups have been working together on data sharing, annual reports and spatial analysis of evictions based on race. And the AEMP has 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. They all have the singular goal of documenting displacement and creative resistance, from the standpoint of the residents and community members. Once you dive into the seventy-eight maps, the charts and stories and voices multiply further. It's not a simple story.

The Anti-Eviction Mapping Project's process and products would seem to be antithetical to the received wisdom in data science and visualization circles. Business writers tout the ability of data visualization to reduce complexity and create new insight, quickly and clearly. "Nothing going on in the field of business intelligence today can bring us closer to fulfilling its promise of intelligence in the workplace than data visualization," wrote Stephen Few in an early 2007 white paper on the promise of data visualization. The story – told by prominent figures in the field such as Few and Nadieh Bremer and David McCandless and Ben Schneiderman – goes something like this: We are living in the age of Big Data in which humans cannot process and make sense of the vast stores of information they are collecting. While our capacity for processing text and numbers is limited, our eyes are uniquely suited to detect patterns from a sea of visual information. As leading researcher Colin Ware has written, "Visualization provides an ability to comprehend huge amounts of data. The important information from more than a million measurements is immediately available." Thus, visualizing large datastores in sensible, user-friendly ways is our ticket to making sense, making decisions, and making money.

This is not an untrue story. It really is measurably easier and faster to see patterns in a table of numbers if they are presented in graphic form. And some of AEMP's seventy-eight maps perform this widely acknowledged function of data visualization by making patterns in the data visually apparent, at a glance.

For example, the *Tech Bus Stop Eviction Map* produced by the collective in 2014 plots the location of three years of Ellis Act evictions. This is a form of "no-fault" eviction in which landlords claim that they are going out of the rental business, in many cases to convert the building to a condominium and sell units at significant profit. San Francisco has seen almost five thousand uses of the Ellis Act to evict residents since 1994. In this case, AEMP plotted these evictions in relationship to the location of technology company bus stops. Starting in the

[1] The concepts taught address specific mathematical content and skills outlined by the Common Core State Standards in New York.

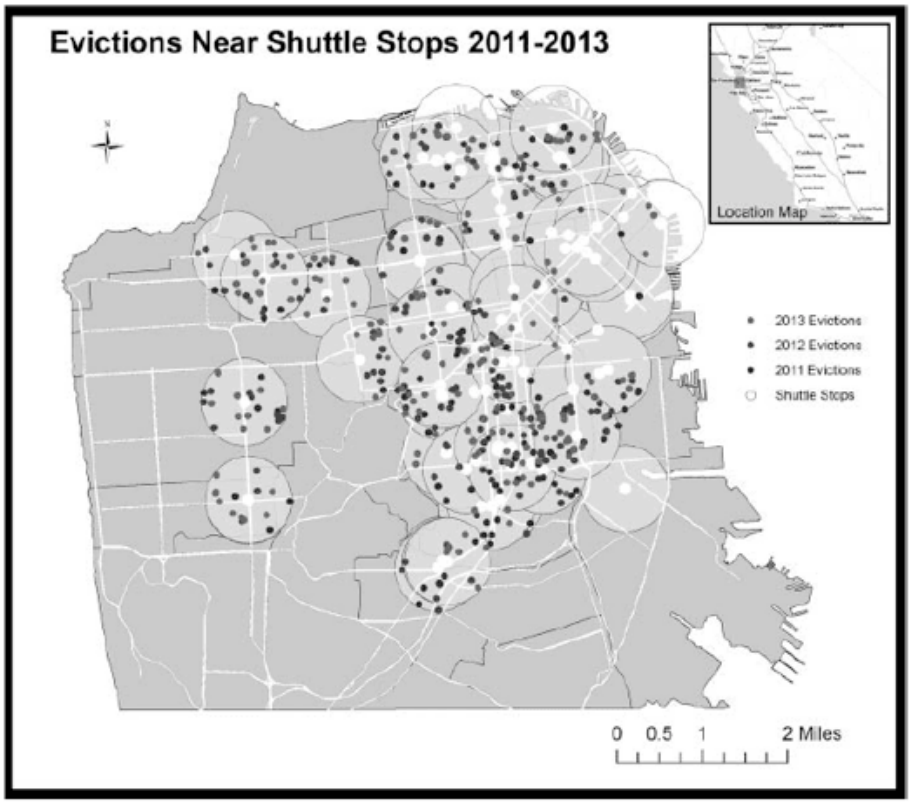


Figure 4.1: Tech Bus Stop Eviction Map by the Anti-Eviction Mapping Project, 2014. Plots evictions with "Google Bus Stops" in San Francisco. Their analysis showed that 69% of no-fault evictions in the city occurred within four blocks of a tech bus stop. ¶ Credit: Anti-Eviction Mapping Project ¶ Source: http://www.antievictionmappingproject.net/techbusevictions.html ¶ Permissions: PENDING

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", these vehicles used public bus stops – illegally at first – to shuttle their riders in comfort. The location of the bus stops also meant that there was a new, wealthy clientele for condos in the area and so property values around the bus stops soared. Here the AEMP makes the case that so, too, did evictions of long-standing residents. Their analysis, shown in the image above, demonstrates that 69% of no-fault evictions between 2011-13 occurred within four blocks of a tech bus stop.

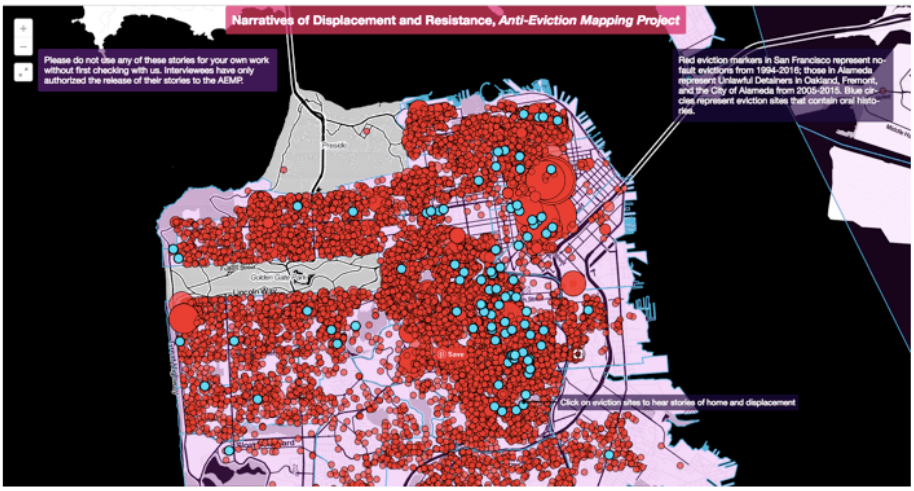


Figure 4.2: "Narratives of Displacement and Resistance" ¶ Credit: The Anti-Eviction Mapping Project ¶ Source: http://www.antievictionmappingproject.net/narratives.html ¶ Permissions: PENDING

But other AEMP maps, like *Narratives of Displacement and Resistance*, pictured above, do not have an efficient analytical function. All 5000 evictions are represented as sized red bubbles, so the basemap of San Francisco is barely visible. On top of the sea of red bubbles, sky blue bubbles dot the locations where the AEMP conducted audio and video interviews with displaced residents, activists, mediamakers, and local historians. Clicking on a sky blue bubble sets the story in motion: "I was born and raised in San Francisco proudly," begins Phyllis Bowie, a resident facing eviction in her Midtown apartment. She goes on to tell the story of returning from serving in the Air Force and working like crazy for two years to build up her income record at her small business to be eligible for a one bedroom lease-to-own apartment in Midtown, a historically Black neighborhood where she had grown up. In 2015, the city broke the master lease and took away the rent control on their building. Now, the tenants, who moved

there on the promise of a future of property ownership, are facing skyrocketing rents that none of them can afford. Bowie is leading rent strikes and organizing tenants but their future is unclear.

The point of this map is not for the eyes to efficiently detect a correlation between space and evictions. There are very few patterns to detect when the entire city is covered in big red eviction bubbles and abundant blue story dots. Rather, the visual point is simple and exhortative: "There are too many evictions". Behind each eviction is a person, with a unique voice and a story like Bowie's.

The *Narratives* map would appear to be messy. It does not efficiently reveal how evictions data may be correlated with BART stops, income, Google bus stops or any other potential dimensions of the data. Moreover, even finding the *Tech Bus Stop Eviction Map* or *Narratives* map is complex given the sheer number of maps and visualizations on the AEMP website. There is no "master map" that integrates all of the information that AEMP has collected into a single interface. So AEMP's efforts would seem to fail at what proponents indicate is the basic value of data visualization: taming information overload, integrating large amounts of information and detecting visual patterns efficiently.

But perhaps cleanliness, efficiency and control are not the only criteria by which to judge data visualizations.

A related story that falls into the *Data Science Bin of Received Ideas That We Might Want To Think About More* is that data always needs to be tamed—it is messy and in need of cleaning. "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 for his widely read and cited 2014 paper called "Tidy Data". Wickham is the author of the *tidyr* package for the R statistical computing platform which logs around 177,000 downloads per month. Articles in the popular press and business corroborate this need for tidiness. The Harvard Business Review calls the Data Scientist "the sexiest job of the 21st century" and talks about this new special form of human: "At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible." Here, the intrepid analyst wrangles an orderly table from unstructured chaos. For the New York Times in 2014 ("For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights"), the analyst's work is less sexy and equated to the low-wage, maintenance labor of janitors.

Whether or not you think data scientists are sexy (they are) or whether you think janitors should be offended by this classist reference (we all should be), there are some interesting assumptions and anxieties surfacing in both of these sets of received ideas. In one story, humans are able to tame the chaos of information overload visually—visual organization helps us go from data to intelligence. In the second story, data are dirty and it is actions of cleaning and tidying that put them back in their proper places.

But what might be lost in the process? Or, more specifically, whose perspectives are lost in the process of dominating and disciplining data and whose perspec-

tives are imposed on the results? Both sets of received ideas make normative assumptions—namely, that we all value cleanliness, efficiency and control over messiness, inefficiency and complexity.

Scholars Katie Rawson and Trevor Muñoz have advanced the idea that "the cleaning paradigm assumes an underlying, 'correct' order," and warn that cleaning can be a "diversity-hiding trick." In the perceived "messiness" of data there is actually rich information about the circumstances under which it was collected. Yanni Loukissas concurs. Rather than talking about *data sets*, he advocates that we talk about *data settings*—both the technical and the human processes that affect what information is captured and how it is structured. Loukissas likes to quote the anthropologist Mary Douglas, who said, famously, "What is dirt but matter out of place?" He employs the example of data in actual dirt: the soil of the Arnold Arboretum, in Boston, MA. In order to determine the origin of a single cherry tree, he explains, you need to know to look at multiple database fields, including one that indicates who incorporated the plant into the collection, another that records the native country of the species, and a third that documents the way it came to the collection. The "messiness" of storing related data in three different fields is actually a *signal*—i.e. meaningful information—that points back to the complex history of recordkeeping practices at the institution.

Loukissas' assertion is that all data are "local", by which he means they are connected, sometimes inextricably, to the human and technical conditions under which they are collected and maintained. For example, he tells the story of exploring data from the Clemson University library in South Carolina while he was at a hackathon in Cambridge, MA. He stumbled across a puzzling record where the librarian had noted the place it described as "upstate". Such a classification is relational to the place of collection—"Upstate" is a term immediately understandable to South Carolinians and refers to the westernmost region of the state, where Clemson is located. But it has no relevance to a person sitting at a hackathon in New England, who might prefer a more generalized way of denoting the ten or so counties that count as Upstate. But note that even though the outsider may be frustrated with the fact that the record doesn't use latitude and longitude, there is meaningful and precise geographic information contained in the "upstate" reference. Not only that, but there is meaningful metadata provided by this cultural insider reference: Only somebody collecting the data in South Carolina would have referred to that region as "upstate", so we can reason that the data was collected there. It is because of records like this that taming and cleaning data is such a chore—it is like chopping off the roots of a tree that connects it to the ground from which it grew. It is painstaking and cumbersome. Plus, uprooting a tree might not always make sense.

We might relate the growth of tools to tidy, tame and discipline data to the proliferation of street names and signs in the landscape. Geographer Reuben Rose-Redwood describes how, for example, prior to the Revolutionary War, very

few streets in Manhattan had signs posted at intersections.[2] Street names, such as they existed, were vernacular and related to the particularity of a spot, e.g. "Take a right at the red house". With the increased mobility of people and things—think of the postal system, the railroads, the telegraph—street names became systematized in the 19th century in the United States and institutionalized by the early 20th century. Rose-Redwood calls this the production of "legible urban spaces." There is high economic value to legible spaces, particularly for large, international corporations to deliver boxes of anything and everything directly to our front door.

The point here is that *one does not need street names for navigation until one has strangers* [3] in the landscape. Likewise, data does not need cleaning until there are *strangers in the dataset.*

Who are those strangers in the dataset? 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).

These operators are "strangers" in data sets because 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 of open data, APIs and the vast stores of training data sets available online: the data appear available and ready to mobilize, but what they represent is not always well-documented or easily understood by outsiders. This problem—that data do a very poor job of speaking for themselves, especially when the listener is a stranger—is something we will return to at length in the next chapter.

Unicorns, wizards, ninjas, rock stars and janitors all have something in common. Apart from the unicorn, a mythical creature that is not usually depicted with a readily apparent gender, they are mostly stereotyped as male undertakings. [4] And unicorns, wizards, ninjas, rock stars and janitors work alone. Their work is solitary and singular. When applied to data science, the focus is on an individual's extraordinary technical expertise and ability to determine meaning where others cannot. Solo superheroes and informational geniuses that weave meaning from chaos.

There is a "genius" in the world of eviction data – it is Matthew Desmond,

[2] CS109 at Harvard is taught jointly by Computer Science and Statistics. As of this writing, there are 37 male faculty (69%) and 17 female faculty (31%).

[3] This does not mean there are no data ethics courses, only that it is not the norm to address these concerns in introductory coursework. Indeed, there is a long list compiled by Dr. Casey Fiesler of technical courses that specifically address ethics and what is being called "fairness, accountability and transparency" in technical fields: http://bit.ly/tech-ethics-syllabi

[4] Stingrays are devices that mimic cell phone towers and trick cell phones nearby into connecting with them so that they can gather personal data. They are used primarily by law enforcement and their use is contested by the ACLU and other organizations concerned with privacy and civil liberties.

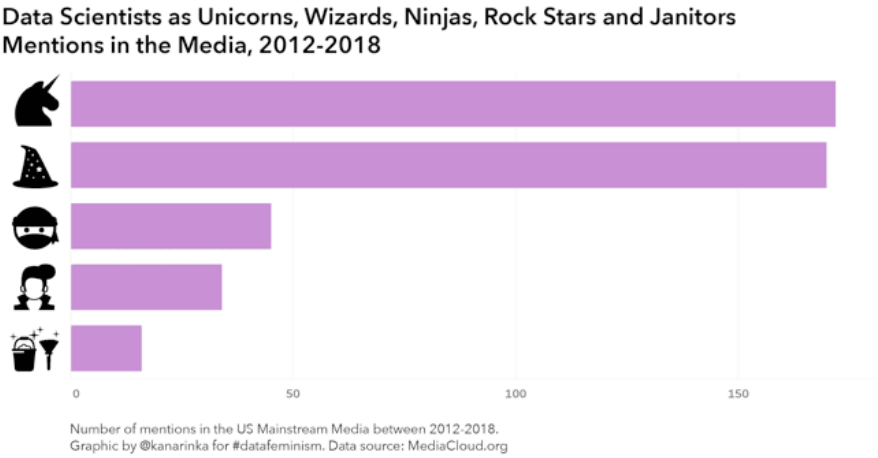


Figure 4.3: Searching MediaCloud [4], between 2012 - 2018 shows that "unicorn" is the most commonly referenced metaphor in relation to data scientists, with "wizard" a close second. There are fewer than fifty articles about data ninjas, rock stars and janitors but they appear in high-profile venues like the Washington Post and Forbes. ¶ Credit: Image by Catherine D'Ignazio ¶ Source: Catherine D'Iganzio

officially designated as such by the MacArthur Foundation for his work on poverty and eviction in the US. He is a professor and director of the Eviction Lab at Princeton University. In an article for Shelterforce, organizers from the AEMP, the Workers' Center of Central New York, the Community Alliance of Tenants in Oregon, and the Housing Justice League in Atlanta detail how Desmond and the Eviction Lab were pursuing a big data agenda for acquiring national evictions data at the expense of understanding local context and providing adequate protections for the communities represented by the data.]{.highlight .h3pblzt9e}Initially, the Eviction Lab approached community organizations like 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, Eviction Lab turned to a real estate data broker and purchased data of lower quality. The authors write, "AEMP and Tenants Together have found three-times the amount of evictions in California as Desmond's Eviction Lab show." Unfortunately, due to the fact that Desmond is a "genius," his social capital (combined with Princeton's) mean that the numbers that many policymakers see and use to make decisions are inaccurate. In this case, the priority was on speed – at the expense of establishing trusted relationships with actors on the ground – and national coverage – at the expense of accuracy. Note that speed and perceived comprehensiveness help to maintain and secure the status of the white male genius and his institution, while strategically downplaying the work of coalitions, communities and movements

that are led primarily by women and people of color. This is a classic case of *Big Dick Data*, a phenomenon you can read more about in the next chapter of this book. "We're unpacking America's eviction crisis," proclaims the Eviction Lab home page, "Find out how many evictions happen in your community."

But what might be gained if we understood data work not as a genius-like wizardly undertaking, but rather one that embraced multiple voices and valued different types of expertise at all stages of the process?

While the Anti-Eviction Mapping Project could have handed off the data it collects to a single mapmaking rock star-unicorn-ninja-wizard-janitor, they made an intentional decision to include many designers in the process, including many non-experts who experienced the power of making maps for the first time. The resultant proliferation of maps, oral histories, events, murals and reports reflects the diverse voices of many collaborators who are working together to document the scope and the scale of San Francisco's housing crisis. And this has the (wholly intentional) consequences of building the technical capacity of residents as well as relationships between community members – slowly and surely, map by map, collaboration by collaboration. In fact one of the explicit goals of 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." [5] In addition to translating evictions into insights, the AEMP wants to use the process of making maps to produce new human relationships.

A key contribution of feminist thinking has been to recognize how a multiplicity of voices, rather than one single loud or magical or technical voice, often results in a more complete picture of the issue at hand. Feminist philosophers like Donna Haraway are part of a postmodern wave of thinkers that maintain that all knowledge is partial, meaning no one person or group has the privilege of a distant, objective view of The Truth. Even if they self-identify as a unicorn, janitor, ninja, wizard or rockstar. But embracing *pluralism* – as this concept is sometimes described – does *not* mean that everything is relative, nor that all truth claims have equal weight, nor that feminists don't believe in science. It simply means that when people make knowledge, they do so from a particular standpoint – a situated, embodied location in the world.

This is called standpoint theory in feminist thinking. And the easiest way to start to understand standpoint theory is to think of it like perspectives that you have from identities you are born into as well as your experiences. For example, we—the co-authors of this book speaking to you at this moment—are two white, cisgendered women, not Latino transmen. We live in Boston and Atlanta, not Bangalore or São Paulo. We've been trained as designers, software developers, and scholars, and not as bank tellers or biomedical engineers. These perspectives

[5] You probably know what WTF stands for. But csv stands for "comma separated values" and is a text-based spreadsheet file format. Each column break is denoted by a comma and each row break is denoted by a carriage return. You can open csv files in spreadsheet programs and most data software packages.

matter. They will shape the questions that we ask of the world, the data we collect, the results that we see, and the meaning that we make. The idea behind standpoint theory is that pooling our standpoints makes for a richer and more robust objectivity.

But there are forces beyond the individual operating in standpoint theory. Standpoints, writes sociologist Patricia Collins Hill, are group-based experiences, "Groups have a degree of permanence over time such that group realities transcend individual experiences." She gives the example of being African American, a stigmatized racial group in the US. "While my individual experiences with institutionalized racism will be unique, the types of opportunities and constraints that I encounter on a daily basis will resemble those confronting African Americans as a group." Hill calls on us to use standpoint theory to acknowledge (and address) social inequality based on existing unequal power relations between social groups. Note how this is different than a call for simple diversity in individual perspectives—what people in the tech industry characterize as "thought diversity." [6] This means explicitly acknowledging and taking steps to address the unjust structural forces at play in our work, including racism, sexism and more.

Indeed, beyond simply "embracing different perspectives", feminist standpoint theory asserts that the best way to strengthen objectivity and address injustice in the system is to begin with the lives, experiences and interpretations of the people most marginalized in a particular context. Applying this to computational systems design, Shaowen Bardzell calls for starting first and foremost with the perspective of the "marginal user." From a gender perspective, that would mean beginning with female and non-binary perspectives. On a project that involves international development data, that might mean beginning not with institutional goals but with indigenous standpoints. For the AEMP, that means centering the voices and experiences of those who have been evicted. Privileging marginal perspectives helps to expose aspects of the world that appear to be neutral and objective, but are actually distorted and one-sided accounts of the world. And centering marginalized standpoints can generate new and critical questions that would otherwise go un-asked because the system is set up to suppress those voices. As Kim Tallbear says, "If we promiscuously account for standpoints, objectivity will be strengthened." So, how do we begin to embrace this kind of plurality of voices and perspectives in data science?

The first step in activating the value of multiple perspectives is to acknowledge the partiality of your own. But how? It can be particularly hard to see just how partial your own perspective is if you are a member of a dominant group whose way of operating in the world stands in for the "default" or "normal" way. Think back to the example of Marya McQuirter's catalog search, described in *Bring Back the Bodies*, in which white people were not labeled as such, because being white in the United States is simply so normal that it goes without saying

[6] "Civic data guides" is the name of the collaboration undertaken by Catherine, Yanni Loukissas and Bob Gradeck around the production of data user guides by students and learners.

or labeling. Feminist sociologist Michael Kimmel illustrates this concept in the following way: while sitting in a small study group in graduate school, his African American colleague said, "When I look in the mirror I see a *Black woman*. When a white woman looks in the mirror she sees a *woman*," to which Kimmel rejoins, half-joking and half-serious, "And when I look in the mirror, I see a *human being*."

Kimmel says that he sees "a human being" rather than a white man because throughout his life his race and gender have granted him privileges. One of the most notable of these privileges is, paradoxically, to not need to see his race and gender as markers of difference. So, as Kimmel articulates it: "privilege is invisible to those that have it." You know you are a member of the dominant group when your gender or race or religion or sexuality is invisible – the fact that you do not have to consider and negotiate it every day. As whiteness scholar Robin DiAngelo says, "a significant aspect of white identity is to see oneself as an individual outside of or innocent of race, 'just human'."

This basic insight is one of the main reasons standpoint theory is so important: it helps us understand the power dynamics at play in knowledge production and unmask partial perspectives (created by dominant social groups) masquerading as universal truths. If the world's data science is created by mostly men, then can we consider it objective? No, it represents the standpoint of the dominant group. If machine vision programs are trained on majority pale faces then can we consider them accurate? No, they represent the standpoint of the dominant group. If 3% of the people that work on criminal justice algorithms in the US are African American [7] then can we consider that software fair and unbiased? No, criminal justice algorithms represent the standpoint of the dominant group. These are localized standpoints that, because of the accumulated power and privilege of the dominant group, "pass" as objective truths.

These are large-scale, structural inequities, which we address further in *The Power Chapter*. But we can take individual and institutional steps towards addressing them right here and right now by incorporating more and different standpoints into data-oriented work. How?

For one, you can disclose your own project's methods – rather than sweeping them under the proverbial rug. This is called *self-disclosure*. You may have even heard the phrase coined by David Weinberger, "transparency is the new objectivity." So rather than attempting to create visualizations and data science products that purport to be objective, you might build a space for transparency and self-disclosure into your design. People in journalism and science have been doing this for some time, at least as it relates to the technical methods employed in their analysis process. For example, Bloomberg's interactive visualization What's Really Warming the World? walks the reader through common climate denier arguments that try to explain away the warming planet with reasons that

[7] This is not to say tools and individual skills are not important (they are), or that your co-authors have never led tool-focused workshops (we have). Rather, the problem when this is the only model of learning that is ever undertaken in a workshop or course.

Figure 4.4: What's Really Warming the World published on Bloomberg.com in 2015 devotes a third of its real estate to describing the methods for how they worked with the data. ¶ Credit: Catherine D'Ignazio, based off reportage by Eric Roston and Blacki Migliozzi for Bloomberg Businessweek. ¶ Source: Catherine D'Ignazio. Original article: https://www.bloomberg.com/graphics/2015-whats-warming-the-world/

don't have to do with human activity. Is it volcanos? Is it deforestation? Is it ozone pollution? The piece systematically demonstrates that these factors have little to do with global warming whereas greenhouse gas emissions from human industries are the clear factor in rising global temperatures.

One of the most interesting things about the piece is that it devotes almost a third of its real estate to describing the data it draws from and the methods the authors used for analysis. Providing access to the data as well as describing the methods used to analyze it is increasingly a convention in data journalism and aligned with the growing trend towards openness and reproducibility in scientific research. While these methodological accounts are presently focused on technical details – where is the data from, what program was used to analyze it, what statistical models were developed – there is a seed of possibility for revealing other details about the human process of making decisions about data storytelling. Who was on the team? Which hypotheses were pursued but ultimately proved false? What were points of tension and disagreement? When did data need some ground-truthing by talking to data owners or domain experts? There is a story for how every evidence-based case came into being and it is a story that involves money, institutions, humans and tools. Revealing this story through reflection and self-disclosure can be a feminist act.

But data doesn't lend itself naturally to self-disclosure. Dietmar Offenhuber, head of the Information Visualization program at Northeastern University, has advanced the idea that data appears so neutral because it is unclear who is the author. In written text (even the driest and most academic tome), there is always an author, a tone, and a connection back to a human speaker through language and attribution. But, as we described in *On Rational, Scientific, Objective Viewpoints from Mythical, Imaginary, Impossible Standpoints,* data and its visualizations carry tremendous rhetorical power, particularly for newcomers. Databases and charts are often so sophisticated at obscuring the perspectives of their human speakers. How can we connect spreadsheets back to speakers, and visuals back to voices?

Self-disclosure could be as simple as being explicit, transparent and possibly even visualizing who is doing the counting and mapping behind the scenes. Take the example of the aerial mapping image in the photo below. The Public Laboratory for Open Technology and Science (PLOTS) is a citizen science group that got its start during the BP oil spill in 2010 [^[9]](#_ftn9). They make high-resolution aerial maps by flying balloons and kites, which dangle cheap digital cameras, over the environmental sites they seek to study. While the technique is low-cost, the imagery produced is often higher resolution than existing satellite imagery because of the proximity to the ground.

As you can see 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 bodies are not missing but represented in the final product. Literally.

Figure 4.5: A photo from a Public Lab research note by Eymund Diegal about mapping sewage flows in the Gowanus Canal. Note the people on boats doing the mapping and the balloon tether that links the camera and image back to their bodies. ¶ Credit: Public Lab ¶ Source: https://publiclab.org/notes/eymund-diegel/12-18-2012/mapping-sewage-flows-gowanu s-canal-after-sandy-flood-damages-sequel

Self-disclosure illustrates the feminist method of reflexivity—rigorous interrogation and transparency about one's own position in the world. Not just one's technical methods, but one's social position, one's institutional position, one's racial position. Reflexivity is a meticulous tactic for addressing that aforementioned conundrum that "privilege is blind to those who have it.

Embracing the value of multiple perspectives shouldn't stop with transparency and self-disclosure. It also means actively and deliberately inviting other standpoints—specifically, those most marginalized by current power relations—into the analysis and storytelling process. As we have seen, the Anti-Eviction Mapping Project does this through producing maps with—not *for*—low-income residents and community organizations like the Eviction Defense Collective and Tenants Together. And there are projects that embrace pluralism and privilege marginal standpoints that we can look to as models. They come from diverse sectors such as university research labs, private consulting groups and government.

Since 2012, Rahul and Emily Bhargava have worked with community organizations from Belo Horizonte to Boston to create "data murals". These are large-scale infographics designed to be displayed in public spaces, that tell data-driven stories about the people who occupy those spaces. For example, Groundwork Somerville, an urban agriculture nonprofit, approached the Bhargavas in 2013. Emily recalls that the organization was in the process of establishing its first urban farm, "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." The organization was also running a high school employment program called "The Green Team" but struggling to fully involve the youth in their mission to create healthier communities.

The Bhargavas and the Director of Groundwork Somerville collected demographic data from the city, GIS data on unused lots, and internal data such as growing records, food donations, and attendance logs at community events. They worked with the youth over several after- school sessions to review and discuss the data, as well as engage in storyfinding (a.k.a. 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". They back that claim up with numbers depicting the cost of healthy food and the number of people with prediabetes. It then transitions into showing opportunity: the number of unused lots in the city and how much land has been reclaimed for urban farming. Finally, the mural shows how the Groundwork Somerville truck brings many pounds of affordable produce to low income neighborhoods and employs over 400 youth residents. It ends with a vision of a unified and healthy community.

Youth, staff and volunteers worked together over the course of several weeks to paint the twenty-meter-long mural on the corrugated steel wall of the garden.

Figure 4.6: The process of making a data mural involves conversation, building prototypes with craft materials, workshops in data analysis, and actual painting. ¶ Credit: Emily and Rahul Bhargava ¶ Source: Emily and Rahul Bhargava. ¶ Permissions: PENDING

Figure 4.7: The Groundwork Somerville data mural, painted by youth, staff and volunteers at Groundwork Somerville, and the Bhargavas in 2013. ¶ Credit: Emily and Rahul Bhargava ¶ Source: Emily and Rahul Bhargava ¶ Permissions: PENDING

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

Data murals like the one in Somerville are becoming a more common practice—Communities from Detroit [8] to Dar Es Salaam have undertaken data murals to tell a public story about an important issue. In Dar Es Salaam, the Data Zetu project ("our data" in Swahili) ran a listening campaign in four low-income districts. They compiled the residents' concerns, as well as statistical data, into a data mural about teenage pregnancy and sexual health. In the image, a young woman is pregnant and wants to grow up and be a doctor. Her peers tell her that she can still dream big, but that she should seek counseling at the youth health clinic to do the best by her new family.

And murals are just one kind of output from a pluralistic, community-centered data process. There are many examples of participatory mapping that combine data collection and community storytelling. For example, in the project Map Kibera, the GroundTruth Initiative worked with residents to map the largest and most well-known slum in Nairobi. While Kibera was not unmapped prior to 2009 when the project began, [9] the maps that were made were used by the government, NGOs and researchers to drive policy, but not made available to residents.

Do you remember the example of the teenager who Target identified as pregnant before her parents did? The issue in that case, as is true of many instances of corporate and government-sponsored data collection efforts, is that the people who collect, store, and derive insight from the data often wield outsize power over those about whom the data are collected. Embracing pluralism is a way to rectify this power imbalance. Map Kibera seeks to redress that asymmetry by teaching residents to collect their own data, make their own maps and tell their own stories through community radio and video journalism. Likewise, the Digital Democracy project works with indigenous groups around the world to defend their rights through collecting data and making maps. In the process, they have developed SMS services with domestic violence groups in Haiti and

[8] The concept of a discotech was created by the Detroit Digital Justice Coalition in 2009 based loosely on the idea of a potluck event for technology. The goal of a discotech is to create "a genuine collaborative, collective learning environment that is accessible to all skill levels, ages, and learning styles." The first Data DiscoTech was run in 2015 as a response to a Detroit open data ordinance. The organizers were concerned about some of the harms that might arise from open data. "Public data impacts different people differently," as one organizer stated, so the goals of the discotechs have included capacity building as well as consciousness building. The Coalition has published a free guide to running your own discotechs here: https://www.alliedmedia.org/files/ddjc_zine_4.pdf

[9] KPIs are Key Performance Indicators – measures that help to evaluate the performance of an organization in regards to a particular activity that it has deemed important. For example, a nonprofit might track its fundraising efforts with a KPI like "Cost per dollar raised" – how much did they spend on fundraising for every dollar that they brought in.

Figure 4.8: Data mural about teenage pregnancy and sexual health in Tanzania by Data Zetu ("our data" in Swahili). ¶ Credit: Data Zetu ¶ Source: https://datazetu.or.tz/photos/ ¶ Permissions: PENDING

helped the Wapichana people in Guyana make a data-driven case for land rights to the government.

Neither Abella Bateyunga from the Data Zetu project, nor Erica Hagen and Mikel Maron from the Map Kibera project, nor Emily Jacobi from Digital Democracy, nor the Bhargavas envision themselves as unicorns, or janitors, or ninjas, or wizards, or rock stars. "At Digital Democracy, we try to fight the superhero narrative", says Jacobi. "We are sidekicks rather than superheroes." Through a series of workshops and trainings, these groups enhance the capacity of an entire community to engage in data analysis and storytelling. Emily Bhargava reflects, "Painting a mural is great for building community relationships. But the time when people actually become empowered is during the storyfinding process when they learn to translate the data and own its meaning. And doing that collectively helps to even the power differential." The facilitators in these cases act more like cheerleaders, or guides, or advocates, or even therapists. Perhaps appropriately, the Bhargavas' website is located at DataTherapy.org.

Facilitating data-informed community conversations, mapping forests in Guyana and painting data murals may seem foreign for those who are accustomed to being indoors with their data, but the ideas about participatory meaning-making are transferable to more conventional contexts like municipal government. In 2015, the City of Boston was in the process of developing its first master plan in fifty years. A master plan is a document that guides future growth and development across a range of city systems like transportation, the built environment and social services and settings. The transportation wing of this effort was called Go Boston 2030 and, with the help of a grant from the Barr Foundation, they assembled a team that attempted to do something highly inventive in community engagement and participatory data analysis.

The way that transportation master plans typically work is that the city planners set up a framework as well as metrics for success. They then hire external consultants to help undertake a visioning process, involve various stakeholders and members of the public, collect data, analyze, and synthesize that data, and produce a report. The City of Boston did all of these steps, which resulted in the Go Boston 2030 Vision and Action Plan in 2017.

But they also did something different along the way. In addition to holding community input meetings, the City dispatched colorful food trucks and decked-out "Idea Bikes" with trailers to all of the neighborhoods in the city. These mobile units served hot chocolate, and staff offered post-it notes and friendly conversation about commuting in Boston now and in the future. Their goal was to collect data in the form of residents' questions and ideas about the future of transportation in Boston. And collect they did. Over a period of eight months, the City collaborated with the nonprofit Interaction Institute for Social Change to collect 8,700 questions and ideas, public engagement data at an unprecedented scale in relation to prior efforts.

So here is the critical juncture point in the story – what did they do with the

Figure 4.9: One of the Idea Bikes from the Go Boston 2030 project. Residents submitted ideas for improving transportation in the city. ¶ Credit: Livable Streets ¶ Source: From http://www.livablestreets.info/go_boston_2030_idea_bike_stop_fenway

data? At this stage in the process, the typical thing to do would be to fork over the citizen ideas to a data wizard-ninja employed by the transportation consultants and await "the answers." Instead, Go Boston 2030 decided to use the data analysis process as an opportunity for participatory meaning-making and consensus building amongst a multitude of stakeholders. They organized a large meeting of policy makers, public servants, and community leaders (including Catherine), in which the participants–organized into thematic groups–reviewed each of the ideas that had been collected, making note of the ideas that warranted further discussion. One idea in Catherine's group provoked controversy: "More housing density around transit." A state transportation official kicked off the conversation by making an impassioned case for changing the city's zoning codes. Then, a city planner offered a short history lesson on the dangers of prioritizing high-density (but also high-cost) housing. The nonprofit representatives followed up by advocating for a requirement for affordable housing should any zoning codes be changed. While not everyone agreed on the details details, the group did agree to add that idea to a list of priority recommendations for the City to pursue.

Figure 4.10: In early 2016, Go Boston 2030 invited 75 community stakeholders into a participatory data analysis process. ¶ Credit: Go Boston 2030 ¶ Source: ht tps://www.flickr.com/photos/135496995@N05/23154181990/in/dateposted/

In early 2016, Go Boston 2030 invited 75 community stakeholders into a participatory data analysis process. - This is actually one of the Idea Roundtables, but placeholder until we get an image from the Idea Review session.

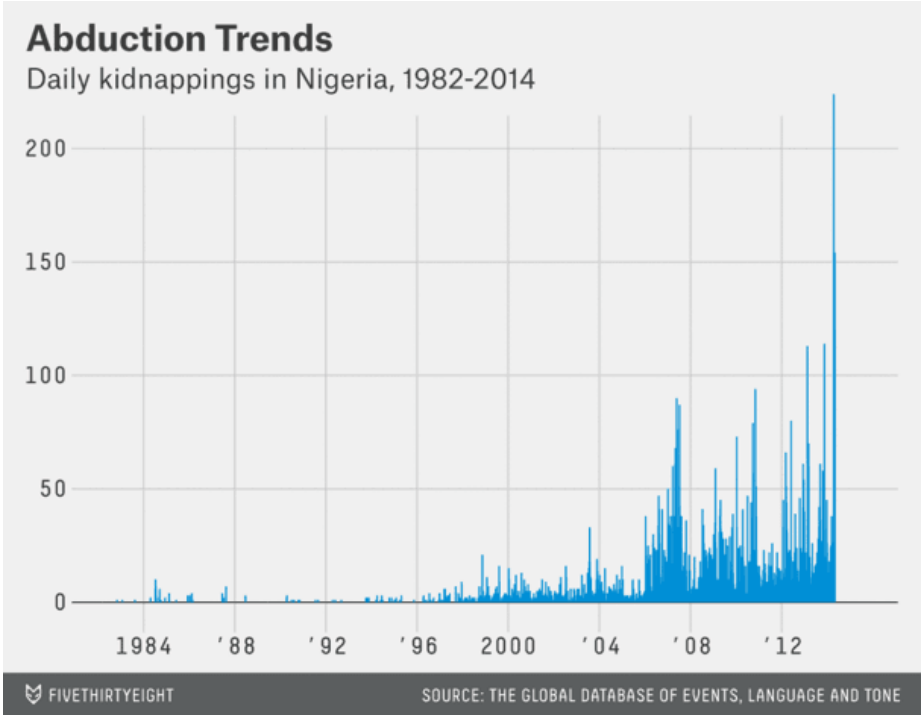
The event was instructive in that the process of data analysis was not imagined as a purely technical problem – something to be handled by a natural language processing expert or a statistician. The designers of this process – the consulting firm Interaction Institute for Social Change – understood that the choice to prioritize one idea over another would carry real weight and material consequences for the people of Boston, consequences that a natural language processing expert or a statistician could not understand simply by looking at word frequencies in the data. Determining which ideas to prioritize among the thousands that were collected could only be achieved by bringing many forms of expertise to an actual table.

The GoBoston 2030 process, the Bhargavas' Data Murals, and the Anti-Eviction Mapping Project are examples of the feminist values in action. Note how they represent different institutional starting points. The first was a government endeavor, the second is a small private consulting group, and the third is a fluid grassroots collective. All embrace pluralism, value knowledge from distinct standpoints, disclose their own standpoint, and center the perspectives of marginalized groups. While we do not necessarily need a seventy-five person community meeting to compute average daily temperatures from instrument readings, we may want to consider that meeting the second we try to make meaning from those readings – defining questions to ask the data, interpreting the data or thinking about how to allocate resources across a vast geography based on the data. Which is to say that as soon as data start to become information that can be operationalized for decision-making, they leave the technical domain and cannot be black-boxed into the server closets of even the most crackerjack unicorn-rock-stars. A data scientist is not going to save democracy, but a well-designed, data-driven, participatory process that centers the standpoints of those most marginalized, empowers participants and builds new relationships across lines of social difference? Well, that might just have a chance.

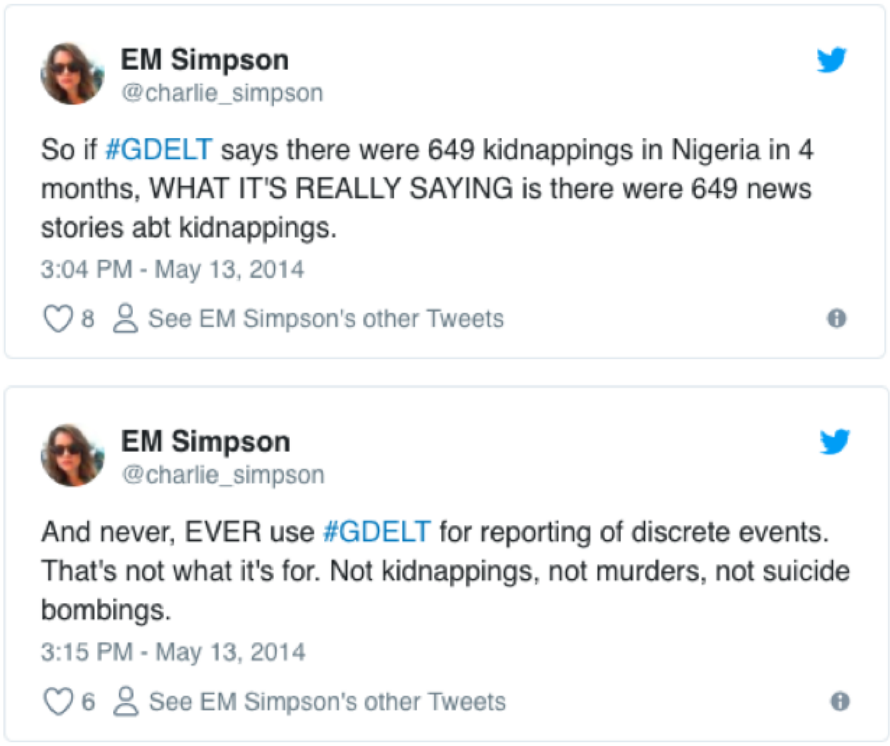
Chapter 5

The Numbers Don't Speak for Themselves

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 keep their 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 their own data-driven story about the event, titled "Kidnapping of Girls in Nigeria Is Part of a Worsening Problem." They reported skyrocketing rates of kidnappings. In 2013 alone, the story asserted that 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.



Shortly thereafter, they had to issue an apologetic retraction because their numbers were just plain wrong. The outlet had used the Global Database of Events, Language and Tone (GDELT) as their data source. GDELT is a big data project by computational social scientist Kalev Leetaru, with previous development by Philip Schrodt and Patrick Brandt. 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, government and civil society. GDELT particularly tries to focus on conflict – whether conflict is likely between two countries, whether unrest is sparking a civil war – all by analyzing media reports. However, as political scientist Erin Simpson pointed out to FiveThirtyEight in a widely cited Twitter rant GDELT's primary data source is media reports and it's not at a stage where you can use it to make reliable claims about separate cases of kidnapping. The kidnapping of schoolgirls in Nigeria was a single event. There were thousands of global media stories about it. GDELT deduplicated some of those to a single event but still logged, erroneously, that hundreds of events happened that day. And the FiveThiryEight report was counting each GDELT pseudo-event as a separate kidnapping incident.



One of Simpson's final admonishments in her long thread to FiveThirtyEight is to "never, ever use #GDELT for reporting of discrete events," because "that's not what it's for." This, combined with other commentary and critique from political scientists, statisticians and bloggers, was embarrassing for FiveThirtyEight– not to mention the reporter– but it also illustrates some larger problems about using data found "in the wild." First of all, the hype around "Big Data" leads projects like GDELT to wildly overstate the completeness and accuracy of their data and algorithms. On their website and in publications, the project leads have stated that GDELT is "an initiative to construct a catalog 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." That giant mouthful is no small dick of Big Data. It is clearly Big Dick Data[1] The question is whether we should take the Big Dick Data at face value or if the Big Dick Data is designed to trick us into giving it massive amounts of research funding.

Yet even when you get past the marketing hype aimed at funders, the GDELT

[1]The concepts taught address specific mathematical content and skills outlined by the Common Core State Standards in New York.

technical documentation is not quite forthright when it comes to whether it is counting media reports (as Simpson asserts) or events. The database FiveThirtyEight used is called the "GDELT Event Database", which makes it sound like it's counting events and not media reports. And the documentation for it states that "if an event has been seen before it will not be included again". This also makes it sound like it's counting events. And a research paper from 2013 states that it's counting events but only specific to publications. So still events, but with an asterisk. Both in documentation and in practice, it is unclear how the system actually works, since there are multiple assertions that events (derived from media reports) are what is being counted, but from a practical standpoint when you use the system the results of counting include many duplicate events. Moreover, there is no guidance in its documentation about what kinds of research questions are appropriate to ask the database and what the limitations are. [2] People who, like Simpson, are insiders to the field of Natural Language Processing event detection[3] and/or the GDELT community, may know to not believe 1) the hype, 2) the title of the database, and 3) the documentation, but how would outsiders or newcomers ever know that?

The stakes are high for context and data. GDELT is not that different from many other data repositories out there. There are a growing number of portals, observatories, and websites where one can download all manner of open government, corporate and scientific data. There are APIs[4] where you can write little programs to query large data sets (like, say, tweets on Twitter) and download them in a structured way. There are test data sets – for network analysis, machine learning, social media, and image recognition. There are fun data sets,[5] curious data sets,[6] proliferating newsletters of numbers and data sets[7] and so on. This would all appear to be a "good thing" in the sense that we, in our contemporary moment, tend to think of unfettered access to information as a kind of inherent good. And, in general, it is kind of amazing that one can just go download pigeon racing statistics, social networks of jazz musicians, the

[2] CS109 at Harvard is taught jointly by Computer Science and Statistics. As of this writing, there are 37 male faculty (69%) and 17 female faculty (31%).

[3] This does not mean there are no data ethics courses, only that it is not the norm to address these concerns in introductory coursework. Indeed, there is a long list compiled by Dr. Casey Fiesler of technical courses that specifically address ethics and what is being called "fairness, accountability and transparency" in technical fields: http://bit.ly/tech-ethics-syllabi

[4] Stingrays are devices that mimic cell phone towers and trick cell phones nearby into connecting with them so that they can gather personal data. They are used primarily by law enforcement and their use is contested by the ACLU and other organizations concerned with privacy and civil liberties.

[5] You probably know what WTF stands for. But csv stands for "comma separated values" and is a text-based spreadsheet file format. Each column break is denoted by a comma and each row break is denoted by a carriage return. You can open csv files in spreadsheet programs and most data software packages.

[6] "Civic data guides" is the name of the collaboration undertaken by Catherine, Yanni Loukissas and Bob Gradeck around the production of data user guides by students and learners.

[7] This is not to say tools and individual skills are not important (they are), or that your co-authors have never led tool-focused workshops (we have). Rather, the problem when this is the only model of learning that is ever undertaken in a workshop or course.

lengths of guinea pig teeth and truckloads of tweets for various hashtags.

While FiveThirtyEight did need a schooling on verifying their data, there's a larger problem at work here that has to do with context. One of the central tenets of feminist thinking, outlined by Donna Haraway, is that all knowledge is "situated." What this means is that context matters – What are the social, cultural, historical and material conditions in which knowledge is produced? What are the identities of the humans making the knowledge? Rather than seeing knowledge artifacts – like datasets – as neutral and objective fodder to use for more knowledge making, a feminist perspective advocates for connecting them back to their context, to better understand their limitations and ethical obligations. And, ultimately, to better understand the ways in which power and privilege may be obscuring the truth.

The issue is that much of the data downloaded from web portals and APIs comes 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 spreadsheet means. But more than likely you get something like this.

Nr. Publication	Bidder	Modality	Dt. Opening	Object
6016.2017 / 0054930-1	Regional Board of Education - Penha	CONVENTION	10/22/2018 10:00 AM	CELEBRATION OF PARTNERSHIP BY TERM OF COLLABORATION
001 / SP-G / 2015	Prefeitura Regional Guaianases - PRG	ELECTRONIC WORK	03/09/2018 11:00	PROVISION OF CLEANING SERVICES AND ASSISTANCE AND PREDIAL CONSERVATION FOR SP-G.
129 / SMADS / 2018	Municipal Office of Assistance and Social Development - SMADS	TERM OF COLLABORATION	08/23/2018 09:00	MSE MA
CRI 001/2018 COHAB	Metropolitan Housing Company of São Paulo - COHAB	COMPETITION	7/31/2018 10:31 AM	INTERNATIONAL COMPETITION No. COHAB-SP 001/2018? ADMINISTRATIVE PROCEDURE No. 2017-0.185.313-9 - PUBLIC-PRIVATE PARTNERSHIP FOR ADMINISTRATIVE AWARD FOR THE IMPLANTATION OF ROOMS OF SOCIAL INTEREST AND POPULAR MARKET IN THE CITY OF SÃO PAULO, ACCOMPANIED BY URBAN INFRASTRUCTURE AND PUBLIC EQUIPMENT, PROVIDING SERVICES THAT SPECIFY, in accordance with the specifications included in the Invitation to Bid and its Annexes.
002/18	Traffic Engineering Company - CET	COOPERATION AGREEMENT	06/01/2018 17:00	Conclusion of Technical Cooperation Terms, which aims to evaluate the automatic systems for controlling the use of parking spaces on public roads developed by private companies for a Pilot Project, to be carried out in the areas of BLUE ZONE of the Municipality of São Paulo designated by the CET for this Project.
015189160	São Paulo Works - SP Works	COMPETITION	05/15/2018 10:00	CONCESSION OF PUBLIC UTILITY SERVICE, WITH PUBLIC USE, WITH ONEROSA PARTY, UNDERSTANDING THE MANUFACTURE, INSTALLATION, MAINTENANCE AND HYGIENIZATION OF FIXED AND MOVABLE PUBLIC SANITARY WARE, WITH ADVERTISING EXPLOITATION.
002/18 / SMSO	Municipal Secretary of Services and Works - SMSO	COMPETITION	05/09/2018 14:30	TECHNICAL SERVICES SPECIALIZED PROFESSIONALS OF CONSULTATIVE ENGINEERING FOR THE MANAGEMENT AND TECHNICAL ADVICE FOR IMPLEMENTATION OF URBAN INFRASTRUCTURE AND PUBLIC BUILDING PROGRAMS IN THE CITY UNDER THE RESPONSIBILITY OF THE MUNICIPAL SECRETARIAT OF SERVICES AND WORKS OF THE SÃO PAULO MUNICIPALITY.

The data shown here – open budget data about government procurement in São Paulo – do not look very technically complicated. Rather, the complicated part is figuring out how the business process behind them works – how does the government run open bids? How do they decide who gets awarded a contract? Are all bids published here or just the ones that got the contract? 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? Without answers to some of these basic questions, it's hard to even begin a data exploration or analysis process.

This is not an uncommon situation. Most data arrives on our computational doorstep context-free and ripe for misinterpretation. And context becomes extra-complicated when poor data documentation is accompanied by the kind of marketing hype we see from GDELT or other Big Dick Data projects. Claims such as GDELT made to totality and universal objectivity are exactly what led Haraway to propose situated knowledge as a feminist antidote to these kind of "unlocatable, and so irresponsible, knowledge claims." The goal of feminist objectivity, then, becomes to connect knowledge back to the bodies of its producers and institutions, with their particular histories, values, limitations, and oversights. In short, to consider context in relation to data.

Ironically, some of the the admirable aims and actions of the Open Data Movement have worked against the urgency of providing context (often inadvertently). Open data is the idea that anyone can freely access, use, modify, and share data for any purpose. And the Open Data Movement – a loose network of organizations, governments, and individuals – has been active since the mid-2000's when groups the like the Open Knowledge Institute were founded and campaigns like the Guardian's "Free Our Data"[8] originated to petition governments for free access to public records. The ideals are great: Economic development by building apps and services on open data; Faster scientific progress when researchers share knowledge; More transparency for citizens to use public information to hold government accountable. The last ideals – transparency and accountability – were key in the framing of President Obama's well-known 'Open Government Directive' in 2009 which directed agencies to make government data open by default. And many more countries, states and cities have followed suit to host open data portals and write open data into policy. 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.

But in practice, limited public funds for technological infrastructure mean that governments prioritize the "opening up" part of open data – publishing spreadsheets of things like licenses granted, arrest data, or flood zones – but cannot go further on developing context or making data usable in order to ensure

[8]The concept of a discotech was created by the Detroit Digital Justice Coalition in 2009 based loosely on the idea of a potluck event for technology. The goal of a discotech is to create "a genuine collaborative, collective learning environment that is accessible to all skill levels, ages, and learning styles." The first Data DiscoTech was run in 2015 as a response to a Detroit open data ordinance. The organizers were concerned about some of the harms that might arise from open data. "Public data impacts different people differently," as one organizer stated, so the goals of the discotechs have included capacity building as well as consciousness building. The Coalition has published a free guide to running your own discotechs here: https://www.alliedmedia.org/files/ddjc_zine_4.pdf

access and use by broad groups of the public. Raw data dumps might be good for starting a conversation, notes scholar Tim Davies, but they cannot actually ensure engagement or accountability. And in fact, many published data sets sit idle on their portals, awaiting users to undertake the labor of deciphering their bureaucratic arcana. There is even a neologism coined by Daniel Kaufmann, an economist with the Revenue Watch Institute, that has been coined to describe this phenomenon: "Zombie data" is data that has been published without any purpose or clear use case in mind.

So, open data has a context problem. Or, a better way to say this is that governments and data providers have not invested as much time and resources in providing context to end users as they have in providing data

But do we need to invest in context? Wired magazine editor Chris Anderson would say "No". In 2008, in a now famous Wired article titled "The End of Theory," Anderson made the claim that "with enough data, the numbers speak for themselves." His assertion was that the age of Big Data will soon permit data scientists to do analysis at the scale of the population. Statistics is based on the idea that you can infer things about a population by taking a random and representative sample. For example, say you want to know which candidate all 323 million people in the US will vote for in a presidential election. You can't contact all of them, but you can call 3,000 of them and use those results to predict what the rest of the people will do. Of course, there's some statistical modeling and uncertainty calculations that need to take place here. And this is the point where Anderson is saying that the theory happens – bridging data collected from a sample with calculations to infer things about a population. At the point when we have data collected about an entire population, theory is no longer necessary. We also, he says, don't need models and theories to understand why something is happening, just to be able to see that one thing is correlated with another: "Correlation is enough." Anderson's main example is Google search. Google's systems don't need to understand why some pages are more linked to than others, only that it's happening, and they will then use that as an indicator of relevance in search.

Now, you can't write an article claiming that the scientific method and all theory are obsolete and not expect some pushback. Anderson wrote the piece to be provocative, and there have been numerous responses and debates, including those that challenge the idea that this argument is a "new" way of thinking in the first place (for example, Francis Bacon argued for inductive reasoning in which the scientist gathers data, analyzes it, and only thereafter she forms a hypothesis. What has unfolded since 2008 in feminist thinking is also a more sophisticated understanding of the ways in which data-driven systems like Google Search do not just reflect back the racism and sexism embedded throughout society but also participate in reinforcing it. This is the central argument of Algorithms of Oppression, Safiya Noble's study of the harmful stereotypes about Black and Latina women perpetuated by search algorithms. In direct opposition to Anderson, Noble asserts that it is the corporation's responsibility to understand

racism in page-linking. Correlation, without context, is not enough when it means that Google recirculates racism.

But there's another reason that the numbers don't speak for themselves when it comes to data about women and marginalized people: not all standpoints are valued equally by society. In writing about a Black woman's standpoint, sociologist Patricia Hill Collins explains that when a group's standpoint is consistently devalued, it becomes subjugated knowledge: "Traditionally, the suppression of Black women's ideas within White male-controlled social institutions led African-American women to use music, literature, daily conversations, and everyday behavior as important locations for constructing a Black feminist consciousness." When groups of people are systematically taught that mainstream culture excludes their experience, stigmatizes their experience or completely neglects their experience, then their knowledge and cultural practices either go underground or are completely silenced. When mainstream institutions try to collect data in the context of subjugated knowledge, the results are uneven because the data setting has major imbalances of power. Nowhere is this more evident than in the case of violence against women and the data that tries (though, in most cases, does not try very hard) to capture the reality of this phenomenon.

In April 1986, Jeanne Clery was sexually assaulted and brutally murdered in her dorm room by an acquaintance at Lehigh University. Clery's parents were devastated. "Most Americans saw the [the space shuttle] Challenger splinter into a billion pieces. That's what happened to our hearts," Connie Clery told People Magazine. The Clerys mounted a campaign to improve data collection about crimes on college campuses and it was successful – the Jeanne Clery Act was passed in 1990 and requires all US colleges and universities to make on-campus crime statistics available to the public. This now includes separate and specific numbers on sexual violence such as sexual assault, dating violence, domestic violence and stalking.

So we have an ostensibly comprehensive national data set about an important public topic. In 2016, three senior students – Patrick Torphy, Michaela Halnon, and Jillian Meehan – in Catherine's data journalism class at Emerson College decided this was a good starting point for their final project. Could the Clery Act data tell them something important about the rape culture[9] that is currently under scrutiny on college campuses?

However, upon downloading and exploring the data for colleges in Massachusetts the students were puzzled. Williams College, a small, wealthy liberal arts college in rural Massachusetts, seemed to have an epidemic of sexual assault, while Boston University, a large research institution in downtown Boston, seemed to have strikingly few cases for its size and population. Not to mention that several high-profile sexual assault cases at BU had made the news in recent years so BU

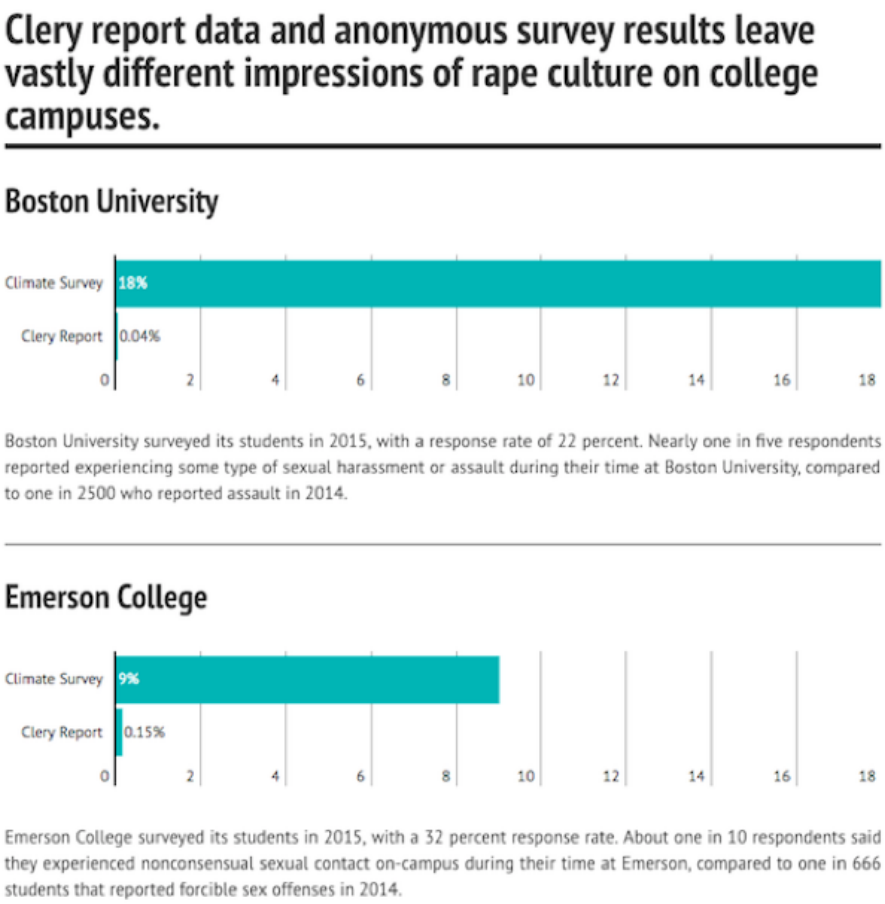
[9]KPIs are Key Performance Indicators – measures that help to evaluate the performance of an organization in regards to a particular activity that it has deemed important. For example, a nonprofit might track its fundraising efforts with a KPI like "Cost per dollar raised" – how much did they spend on fundraising for every dollar that they brought in.

did not have a great reputation around Boston. The students were suspicious – and with good reason. Their further investigation revealed that the truth is likely closer to the reverse of the picture that the Clery Act numbers paint. But you cannot know that without understanding the context of the data.

Colleges and universities are required to report sexual assault data and other campus crimes annually per the Clery Act, and there are stiff financial penalties for not reporting. But it's important to note that the numbers are self-reported. There are only sixteen staff members at the US Department of Education devoted to monitoring the more than 7,000 higher education institutions in the country so it is unlikely that underreporting by an institution would be discovered except in very high-profile cases like the Sandusky Case at Penn State. Moreover, there are strong incentives not to file a Clery report with high numbers. First of all, no college wants to tell the government– let alone parents of prospective students– that it has a high rate of sexual assault on campus. High numbers of sexual assault are bad for the bottom line so universities are actually financially incentivized to not encourage survivors to come forward. And survivors of sexual assault themselves often do not come forward because of social stigma, the trauma of reliving their experience, and the resulting lack of social and psychological support. This is subjugated knowledge - by normalizing male sexual violence, mainstream culture has taught survivors that their experience will not be treated with care and, in fact, they may face more harm, blame and trauma if they do come forward. The result is silence, and the effect on the data is that vast rows of survivors go unaccounted for.

As the students consulted with experts, compared Clery Act data with anonymous campus climate surveys, and interviewed survivors, they found that, paradoxically, 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.[10] They quoted Sylvia Spears, Vice President for Diversity and Inclusion at Emerson College, who stated it like this, "Institutions that have high numbers—it's not always just that high incidents are happening. It's that you've created a culture where people feel they can report and will be supported in that process." The places with higher numbers had convinced their administrators not to turn a blind eye and were actively working on shifting the campus climate so that survivors could come forward without fear of shame, blame or retribution.

[10]If you want a viscerally enlightening reading about privilege, check out White Privilege: Unpacking the Invisible Knapsack by Peggy McIntosh from 1989. Written in the first-person perspective of a white person in the US, it lists fifty ways that white privilege manifests in everyday life, including "My culture gives me little fear about ignoring the perspectives and powers of people of other races."



Do you remember the body issues we described in Chapter One? One of the key reasons that data science needs feminism is that bodies go uncounted, particularly bodies of women and people of color. This is true in the case of data about maternal health, human migration, police killings, health impacts of pollution, and more. And this is certainly the case with sexual assault data, where society systematically neglects and devalues the standpoint of survivors. Their experiences become subjugated knowledge - stigmatized and silenced. Thus, the collection environment has social, political and cultural incentives around reporting that are misaligned and work against collecting reliable, accurate data. Simply stated, there are imbalances of power in the data setting, so we cannot take the numbers in the data set at face value. Here one needs a sophisticated understanding of the context of the data and the actors in the data collection system in order to be able to work with it ethically and truthfully. Lacking this understanding of context 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. Cathy O'Neil, the author of Weapons of Math Destruction, has a term for this: A "pernicious feedback loop" helps to reinforce the unfair environment from which it spawned. 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 data set is silent.

One of the key analytical missteps of work that "lets the numbers speak for themselves" is the assumption that data are a raw input rather than seeing them as artifacts that have emerged fully cooked into the world, birthed out of a complex set of social and political circumstances already existing in the data setting. It's important to note that there is an emerging class of "data creatives" whose very existence is premised on context-hopping by combining disparate data. 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. Data's new creative class is highly rewarded for producing work that creates new value and insight from mining and combining conceptually unrelated data sets.

But data is an output first. After that, it can become an input into a new process, but only with understanding of what the limitations of the collection environment were. "Raw Data" is an Oxymoron is the lovely title and primary assertion of a book edited by literary and information studies scholar Lisa Gitelman that traces the history of data and its connections to today's data culture. Many data-driven projects aiming towards producing new, future insights forget to interrogate how the data got collected and cooked in the first place. FiveThirtyEight got the GDELT events data and jumped into the analysis without looking backwards into how the data was acquired and processed. Clery counts emerge out of a data setting that has an imbalance of power, subjugated knowledge and misaligned incentives and so does not measure what it appears to on first glimpse.

This kind of "data-is-raw-input" mentality happens in scholarly research as well. An academic paper about "the Baumgartner Reddit Corpus" authored by Devin Gaffney and J. Nathan Matias made waves in spring 2018. Three years prior, software developer Jason Baumgartner published a dataset that he claimed contained "every publicly available Reddit comment". Computational social scientists were thrilled. To date, at least fifteen peer- reviewed studies have used the dataset for research studies on topics like politics, online behavior, breaking news, and hate speech. But Gaffney and Matias found a big problem with this big data set: The supposedly complete corpus is missing at least 36 million comments and 28 million submissions. Depending on what the researchers used the corpus for, the missing data may affect the validity of their results. Some researchers have re-run their experiments and found no changes in their findings when they included the missing data[11] and others have not given any public

[11] We bring this up so that you might question the wisdom of uncritically pulling 18th century philosophers into 21st century ethics conversations, i.e. so you can drop some knowledge on the

statement.

Gaffney and Matias' work represents an emerging feminist methodological approach to big data research. Instead of using large data sets as raw inputs to create other meaning, they are interrogating the context, limitations and validity of the data itself. Which is to say, they are examining the data to understand the cooking process. In a similar vein, computer scientists and historians at Stanford undertook a study called "Word embeddings quantify 100 years of gender and ethnic stereotypes." Using machine learning and a data set of 200 million words taken from US books, media and census data from the 20th century, they sought to analyze gender and ethnic stereotypes over time. Word embeddings are numbers that quantify associations between words. So, for example if "woman" appears more frequently near "emotional", that bias is quantifiable and we can see whether that relationship shifts or changes over time. Their results show that words like "intelligent", "logical", and "thoughtful" had masculine associations before the 1960s and have increasingly been associated with women since then. But other words, like those associated with physical appearance did not show such comparative "progress." In the paper, the researchers assert that shifts in word embeddings can quantify the effects of social activism. They write, "The women's movement in the 1960s and 1970s especially had a systemic and drastic effect in women's portrayals in literature and culture."

What makes this project feminist in both topic and method is its use of computation to situate gender and ethnic bias in a social and temporal context. Note that the researchers did not try to assert that the data represent "how women and men are." They also did not try to "remove the bias" so that they could study gender differences. 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 collectively produce more work that situates data, interrogates bias and sensitively treats subjugated standpoints and knowledges?

Unfortunately for Chris Anderson, the answer is that we need more theory, context, and scientific method, not less. Why? Because, quite simply, the humans are always in the loop. Even when the algorithms are doing the heavy lifting. As we showed in What Gets Counted Counts, without theory, survey designers and data analysts are relying on their intuition and "common sense" theories of the things they are measuring and modeling and this leads directly down the path towards cognitive bias.

Deep context, subjugated standpoints and computation are not incompatible. Desmond Patton has a unique background – trained as a social worker, he now runs SAFElab, a research lab at Columbia that 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 social media, specifically Twitter data, to

next person who sings the praises of the categorical imperative in a machine learning ethics discussion.

understand and prevent gang violence in Chicago. But when he started doing this work five years ago, he ran into a problem. Even though he is African American, grew up in Chicago, and worked in many of these neighborhoods for years in violence prevention, "I didn't know what young people were saying, period." At the same time, Patton and his team are acutely aware of the fact that many groups, such as law enforcement and corporate platforms, are already surveilling youth of color online. He continues, "it became really clear to me that we needed to 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."

The solution to context, in this case, came through direct contact with and centering the perspectives of the youth whom they sought to understand. Patton and doctoral student William Frey hired formerly gang-involved youth to work on the project as domain experts. These experts coded a subset of the millions of tweets, then a team of social work students was trained to code them. The process was long, and not without challenges. Patton and Frey actually created a new "deep listening" method, called Contextual Analysis of Social Media, in order to help human coders mitigate their own bias in the coding process and get closer to the intended meaning of a single tweet. Finally, they trained a machine learning algorithm to classify the youths' tweets. Says Patton, "We trained the algorithm to think like young African American man on the south side of Chicago."

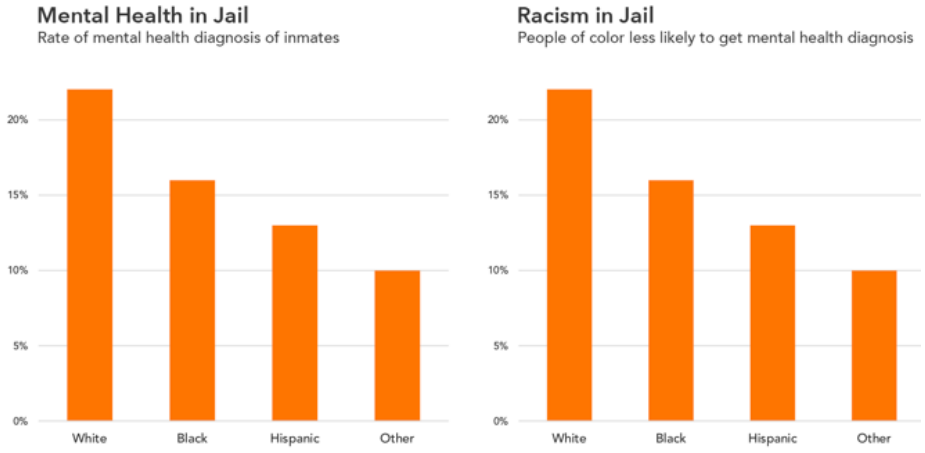
Here is feminist standpoint theory in action in artificial intelligence. The dominant culture approach would have done something naive or misinformed, like counting violent words in tweets. So a tweet like "aint kill yo mans & ion kno ya homie" would have been classified as aggressive or violent, fulfilling the dominant culture's stereotype of Black youth. Taking a situated view, from the standpoint of Black youth themselves, Frey and Patton were able to show that many tweets like this one are actually youth quoting music lyrics of local rap stars, in this case Chicago rapper Lil Durk. These tweets are about sharing culture, not communicating threats.

Note that in order to train an algorithm to understand the context of subjugated standpoints, significant human infrastructure and ethical navigation is required. Frey and Patton have built long-term, ongoing relationships with individuals and organizations in the community. Indeed, Frye lives and works in the community. Both are trained social workers, with years of experience working in violence prevention. According to Patton, they lead with the social worker's code of ethics, one of whose principles is "Social workers recognize the central importance of human relationships." Rather than using computation to replace human relations, SAFELab is using AI to broker new forms of understanding across power differentials. This kind of social innovation often goes underappreciated in technical disciplines. 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."

Note that it's not just in the stages of data acquisition and analysis that context matters. Context also comes into play in the framing and communication of numbers. 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". This study looked at the medical records of 45,189 first-time inmates and found that some groups are likely to receive a treatment response and others are more likely to receive a punishment response. Older people and white people were more likely to receive a mental health diagnosis. Black and Hispanic inmates were more likely to enter solitary confinement. While the researchers explain some of this variation from differential diagnostic rates outside of jail, they also attribute some of the variation to discrimination within the jail system. Either way, the racial and ethnic disparities are a product of structural racism.

Consider the difference between these two graphics. The only variation is the title and framing of the chart.



Which one of these graphics would you (should you) choose to use? The first – "Mental Health in Jail" – represents the most typical way that data is communicated. The title appears to be neutral and free of bias. This is a graphic about rates of mental illness diagnosis of inmates broken down by race and ethnicity. The title does not mention race, ethnicity, racism, or health inequities. The title also does not point to what the data means. And remember from What Gets Counted Counts what happens in the case of "not enough meaning"? Our helpful, heuristic-loving brains will start to fill in the gaps. The particular subject matter of this data – combining mental illness and race and incarceration – contains three charged issues that are particularly prone to stigma and stereotypes. In chart #1, your viewers' brains will likely start to draw inferences based on stereotypes that use essentialist ideas about race/ethnicity

(the main category of analysis depicted) to explain the variation in the data, as in, "Oh, this is because white people are x and Black people are y."

So, here is where an important context question comes in. Are you representing only the four numbers that we see in the chart? Or are you representing the context from which they emerged? Because the research study that produced these numbers presents convincing evidence that *we should distrust the diagnosis numbers* due to racial and ethnic disparities. So, if you publish a chart of those same numbers without questioning them in the title, you are actually undermining the main claim of the research. The scientists' results showed that white prisoners disproportionately receive treatment measures like mental health services and people of color disproportionately receive punitive measures like solitary confinement. So, chart #1 is not only not providing enough meaning (and letting stereotypes flow in) but it is also not giving enough information about the main claim of the research.

Enter chart #2: "Racism in Jail: People of color less likely to get mental health diagnosis." There are a couple of important things this title is doing. First, the title is framing a way to interpret the numbers that is in line with the context and claims of the research study from which they emerged. The research study was about racial disparities, so the title and content of this chart are about racial disparities. Additionally, and crucially, this chart names the forces of oppression that are at work: "Racism in Prison." Rather than leave the door open for stereotypes and essentialist views of race and ethnicity and mental illness to flood your viewer's minds, this chart names the force at work that produces this inequality: it is racism.

"But", you may say (and our students say this a lot), "I don't want to tell people what to think. I want to let them interpret the numbers for themselves." This is an ostensibly noble sentiment, but it fails to acknowledge the power relationship between the author and the audience. As the data journalist in this scenario, you are in a position of power to be able to communicate something to your readers. Presumably, you have researched the topic and know more about it than your audience. Because of that, your audience is in a position of listening and paying attention. What this means is that you have a responsibility – precisely because of your position of privilege – to communicate both the data and the most accurate interpretation of the data. If you let the numbers speak for themselves, this is emphatically not more ethical and more democratic. Why? Because, stereotypes and heuristics love a vacuum. Your audience is highly unlikely to go read all the research you read, do the calculations you did, and interview the people you interviewed. They are reading your story or analysis precisely because they do not have the time for that. So if your work fails to provide meaning and context to numbers, their minds will fill in the gaps with the path of least resistance – and that will probably include stereotypes in the case of stigmatized topics like race, gender, mental illness, incarceration, or a host of other topics. The feminist imperative to consider context and situate numbers in relation to their social and political context is not just a recommendation but

a responsibility of ethical data communication.

This counsel – to name forces of oppression when they are clearly present in the numbers – particularly applies to data scientists and designers from the dominant group. White people, including ourselves, 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 heteronormativity. If you are concerned with truth and justice in data, we suggest that you practice recognizing, naming and talking about these structural forces of oppression because it is in aggregated data that they are most evident. We go into further detail about these forces in *The Power Chapter*.

Part of considering context is understanding that data collection always involves an investment of some combination of interest, money, and time. Like we said in Bring Back the Bodies, counting is power, and that power is not distributed equally across all social groups. There are many important issues, often related to women and other marginalized groups, about which we have little-to-no data. As artist Mimi Onuoha points out in relation to her project *Missing Data Sets*, this is primarily because institutional incentives do not exist to collect it. And the groups who are most affected by the problem often do not have the resources of either time or money or expertise to do it on their own.

These structural issues can feel overwhelming. But honoring context responsibly in the course of one's work with data is also not that complicated. It merely involves a reconception of the role of the data scientist from a "raw data" massager to a "cooked data" investigative biographer. It involves looking backwards at the data setting – and reflecting on your own identity in relation to the data – before you look forwards to create new insights produced with the data set.

Educators, journalists, and civic data publishers are starting to develop more robust tools and methods for context, and we'll take you on a quick tour of several. The first and most important is simply to take an "equity pause" at the beginning of the project, and later at key strategic moments. An equity pause is a process step in EquityXDesign, a justice-focused design framework developed by Christine Ortiz, Caroline Hill and Michelle Molitor. Their framework asserts that research and design can proceed hand-in-hand with racial equity, but only with an additional set of checks around power and privilege. As applied to data science, an equity pause would involve questioning your research questions, questioning your categories and questioning your expectations, particularly as they relate to data about people. As Confucious said, "Real knowledge is to know the extent of one's ignorance". But this is really difficult – especially for people who are members of the dominant social group (and thus more susceptible to *the overconfidence bias* and *the illusory superiority bias* and the *status quo bias*, among others). And the smaller and less diverse your team is, the more likely you are to fall prey to habits of thinking like the self-serving bias and *the egoecentric bias*.

Remember the scenario in What *Gets Counted Counts* in which you were

designing a data project about cell phone data usage? Planning time for an equity pause in the research and discovery phase of a project might lead you to entirely different design decisions as the survey designer. It would allow you the time and space to research contemporary ideas about gender and mobile technology and incorporate them into your work. For example, you might collect more than simply "Women" and "Men" cell phone usage. Or you might collect gender on a spectrum, as a continuous variable. And an equity pause – especially together with a team – may lead you to make some of your implicit assumptions explicit like "women talk more so let's collect minutes." To which your female colleagues might respond, diplomatically, "that's an over-generalizing essentialist assumption." Finally, an equity pause informed by deeper research, may lead to highly innovative methods. For example, demographer Jill Williams suggests that quantitative work informed by feminist theory may need to treat gender as a dependent variable. Meaning, this work would look at gender as an outcome of other intersecting aspects of identity such as age, race, class, sexuality and ethnicity. Doing that kind of intersectional analysis might lead you to discover the connection between millennials and expanded social networks that you had previously missed.

A related but slightly more technical proposal advocated by researchers at Microsoft is being called *datasheets for datasets*. Inspired by the datasheets that accompany hardware components, Timnit Gebru and colleagues advocate for data publishers to create a short, 3-5 page document that accompanies data sets and outlines how they were created and collected, what data is missing, whether preprocessing was done, how the dataset will be maintained, and legal and ethical considerations such as whether the data collection process complies with privacy laws in the EU.

Providing more context is in line with a feminist approach to data and it also helps move towards some of the unrealized ideals of the Open Data Movement around participation, transparency and civic empowerment. For example, Gisele Craveiro, a professor at the University of São Paulo, researches the dissemination and reuse of open government data. Brazil has a transparency law on the books that requires the government to publish data about every expenditure in 24 hours or less. Most of this gets published in impenetrable tables with little metadata or documentation, as shown earlier in the chapter with the example of the procurement table. In the project "Cuidando do Meu Bairro"(Caring for My Neighborhood), Craveiro and her team created a tool to make this spending data more accessible to citizens by *adding context* to the presentation of the information. Their results showed that people could engage with the data better once they could see expenditures that occurred in their neighborhood and what their funding status was (planned, committed or paid). Not only that, the research team was also able to communicate accessibility struggles around lack of context back to government officials and influence how the data was published in the first place.

So tools and methods for providing context are being developed and piloted,

and there is still hope (we hope!) for the future of open data. But 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 reddit comments, we have seen how even the most highly educated among us failed to verify the basic claims of their data source. And datasheets for data sets are great, but can we expect individual people and small teams to conduct an in-depth background research project while on a deadline and a 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 have seen how data publishers, in their quest for research funding, overstate their capabilities and don't document the limitations of the data. In the case of the reddit comments, the data was provided by an individual acting in good faith, but who did not verify – and probably did not have the resources to verify – his claim to completeness. In the case of the sexual assault data, the universities self-reporting cases are incentivized to underreport and government is under-resourced to verify and document all the limitations of the data. And if one of the goals is transparency and accountability, the institutions in power often have strong incentives to not provide context,[12] so the data setting is rife with conflicts of interest. Indeed, Gebru and colleagues foresee challenges to publishers specifying ethical considerations on their datasheets because they may perceive it as exposing themselves to legal and public relations risks.

So, is it data intermediaries? Intermediaries might include librarians, journalists, nonprofits, educators and other public information professionals. These folks are doing context-building work in some piecemeal but important ways. For example, ProPublica, the nonprofit news organization, has compiled the largest US database on school segregation from public data sources. They provide a 21-page document to give context on where the data comes from, the time period it covers and what kinds of questions are appropriate to ask of the data. The nonprofit Measuring Justice provides comprehensive and contextualized data on criminal justice and incarceration rates in the US. So, intermediaries who clean and contextualize the data for public use have potential (and have fewer conflicts of interest), but there would have to be a funding mechanism, significant capacity building and professional norms-setting that 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

[12] A breast pump is a device used to extract milk when a breastfeeding mom is separated from her baby or cannot/does not want to nurse them at the breast. Despite the fact that the medical establishment sees breastfeeding as a public health issue, it is socially stigmatized and has faced neglect as a space of innovation. Lack of paid family leave policy in the US, means that nursing mothers (and trans dads) often end up back at work secretly pumping in closets, bathrooms and cars, if they are able to pump at all.

up with public information resources that are subpar at best and dangerous at worst. 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 come from a data setting with a power imbalance or misaligned collection incentives (read: pretty much all data settings!), and especially when the numbers have to do with human beings, then they run the risk of being not only discriminatory, not only empirically wrong, but actually dangerous in their reinforcement of an unjust status quo. Considering context should be a frontier for open data advocates, philanthropic foundations, researchers, news organizations, and – perhaps most importantly – regulators.

Chapter 6

Show Your Work

If you work in software development, chances are that you have a GitHub account. As of June 2018, the online code management platform had over 28 million users worldwide. By allowing users to create web-based repositories of source code (among other forms of content) to which project teams of any size can then contribute, GitHub makes collaborating on a single piece of software, or a website, or even a book, much easier than it's ever been before.

Well, easier if you're a man. A 2016 study found that female GitHub users were less likely to have their contributions accepted if they identified themselves in their user profiles as women. Critics of GitHub's commitment to inclusivity (or lack thereof) also point to the company's internal politics. In 2014, GitHub's co-founder was forced to resign after allegations of sexual harassment were brought to light. But problematic gender politics do not necessarily preclude other feminist interventions. And here, GitHub makes an important one: the platform helps *show the work* of writing collaborative code. In addition to basic project management tools, like bug tracking and feature requests, the Github website also generates visualizations of each team member's contributions to a project's codebase. Area charts, arranged in small multiples, allow viewers to compare the quantity, frequency, and duration of any particular member's contributions. A virtual "punch-card" reveals patterns in the time of day when those contributions took place. And a flowchart-like diagram of the relationships between various branches of the project's code helps to acknowledge any sources for the project that might otherwise go uncredited, as well as any additional projects that might build upon the project's initial work.

Coding is work, as anyone who's ever programmed anything knows well. But it's not always work that is easy to see. The same is true for collecting, analyzing, and visualizing data. We tend to marvel at the scale and complexity of an interactive visualization like the Ship Map, which we first discussed in *Bring Back the Bodies* as an example of the view from nowhere. That view, as it turns

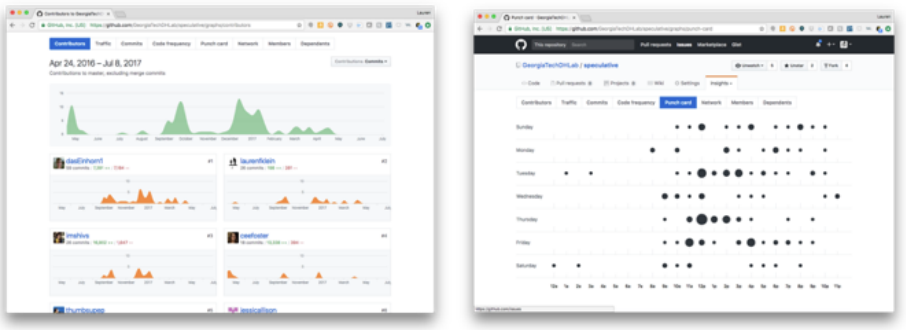


Figure 6.1: Caption: Two visualizations of the code commits associated with a project from Lauren's research group, the Digital Humanities Lab, showing the significant contributions of her student researchers. ¶ Credit: Screenshot by Lauren Klein of GItHub data ¶ Source: https://github.com/GeorgiaTechDHLab/speculative/graphs/contributors and https://github.com/GeorgiaTechDHLab/speculative/graphs/punch-card

out, presents the path of every ship in the global merchant fleet over the course of the 2012 calendar year. By plotting every single trip, the Ship Map exposes the network of waterways that constitute our global product supply chain. But we are less often exposed to the network of processes and people that help constitute the visualization itself. From the seventy-five corporate researchers at Clarksons Research UK who assembled and validated the underlying dataset, to the academic research team at University College London's Energy Institute that developed the data model, to the design team at Kiln that transformed the data model into the visualization that we see– and that is to say nothing of the tens of thousands of commercial ships that served as the source of data in the first place–visualizations like the Ship Map involve the work of many hands.

Unfortunately, though, when releasing a visualization to the public, we tend not to credit the many hands who perform this work. We often cite the source of a dataset, and the names of the people who designed and implemented the visualization. But we rarely dig deeper to discover who collected our data, who processed it for use, and who else might have labored to made our visualizations possible. Admittedly, this information is sometimes hard to find. At other times, it can't be found at all. But the difficulty we encounter when trying to acknowledge this work reflects a larger problem in our data supply chain, as Miriam Posner explains. Like the contents of the ships visualized on the Ship Map, about which we only know vague details– the map can tell us *if* a shipping container was loaded onto the boat, but not *what* the shipping container contains– the invisible labor involved in data work is something that, Posner argues, we willfully see with "partial sight."

To put it more simply, it's not a coincidence that much of the work that goes into

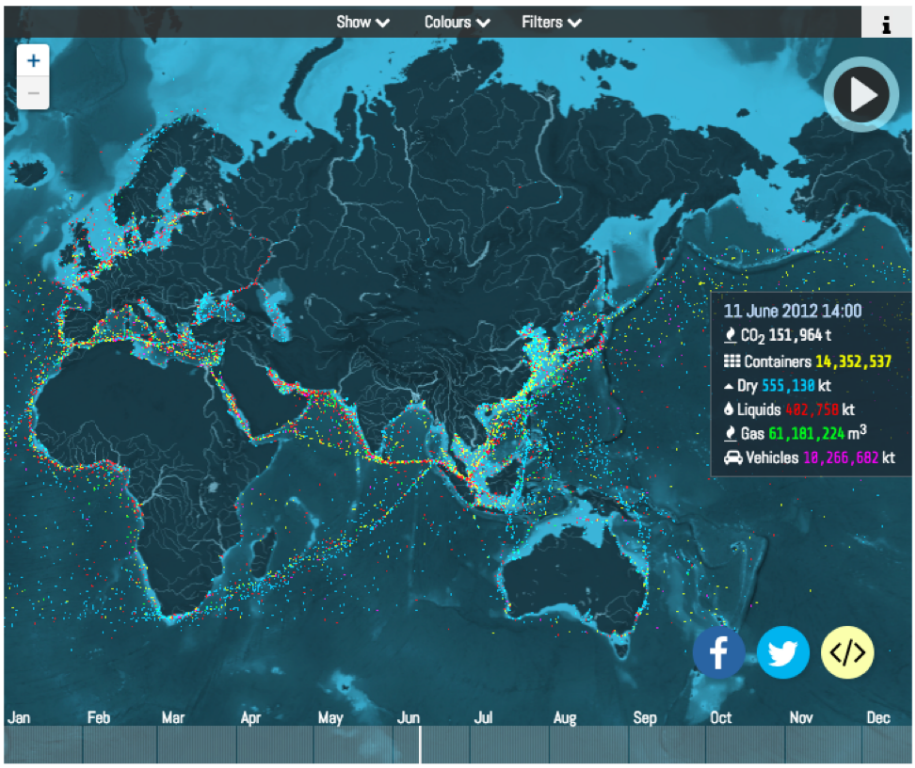


Figure 6.2: Time-based visualization of global shipping routes designed by Kiln based on data from the UCL Energy Institute. ¶ Credit: Website created by Duncan Clark & Robin Houston from Kiln. Data compiled by Julia Schaumeier & Tristan Smith from the UCL EI. The website also includes a soundtrack: Bach's Goldberg Variations played by Kimiko Ishizaka. ¶ Source: https://www.shipmap.org/

designing a data visualization remains invisible and uncredited. In our capitalist society, we tend to value labor that we can see. When, in the early 1970s, the International Feminist Collective launched the Wages for Housework campaign, it was this phenomenon of *invisible labor* that they were trying bring to light. By demanding wages for housework, the group was attempting to erase the distinction between the paid labor of traditional jobs, like office or factory work, and the unpaid labor of household tasks, like cooking or cleaning or child-rearing. Housework might be invisible, these women insisted, performed out of sight and away from the marketplace, but it's certainly not without value. On the contrary, the invisible labor performed inside the home is precisely what enables those who work outside the home to continue to do so.

Figure 6.3: A Wages for Housework march, 1977. ¶ Source: https://hollisarchives.lib.harvard.edu/repositories/8/archival_objects/1438878 ¶ Credit: Schlesinger Library, Radcliffe Institute / Bettye Lane ¶ Permissions: Pending

Unlike washing dishes, however, data work doesn't require that you get your hands wet. (Unless, of course, you're a citizen scientist associated with Public Lab and you're actually collecting data on water). But invisible labor is what sustains the world of data science as well. When was the last time you saw an analysis of census data list the names of any Federal Census Workers, those people outfitted in orange safety vests who knock on your door to remind you to fill out your census form? Or what about the pool of typists who hand-keyed the text of the historical newspapers that you used to train your neural network? Or the metadata librarian who created the fields for the collections database that you visualized? Or the archivists (or, more likely, student employees) who

entered all of the actual records into those fields? This work is not always performed for free, nor is it always performed by women. But we can still view it as invisible labor for the way that it remains invisible to the public eye, and uncredited in the end result.

When looking at the various forms of invisible labor that characterize our present moment, information studies scholars tend to focus on the forms of labor that are not only uncredited, but also unpaid. Visit WagesforFacebook.com and you'll find a version of the Wages for Housework argument, updated for the present. "They call it sharing. We call it stealing," is one of the lines that scrolls down the screen in large black type. The "it" refers to a form of invisible labor that most of us perform every day, in the form of our Facebook likes, Instagram posts, and Twitter tweets. We might do it because it's fun, and we might not expect to be paid for it, but the point made by Laurel Ptak, the artist behind Wages for Facebook, which is the same made by theorists of digital labor, most notably by Tiziana Terranova, is that the invisible unpaid labor of our likes and tweets is precisely what enables the Facebooks and Twitters of the world to profit and thrive.

The world of data science is able to profit and thrive because of unpaid invisible labor as well. How did Netflix improve their movie recommendation algorithm? They crowdsourced it. How did the Guardian, the British newspaper, determine which among two million leaked documents might contain incriminating information about government misspending? They crowdsourced it. The error correction performed on the dataset of early modern books that you downloaded for your text analysis project? That was crowdsourced, too.

"But crowdsourcing is fun," its proponents might say. "People wouldn't do it otherwise!" (And in the case of Netflix, they'd be quick to point out that the winning team was paid a million dollar prize). But someone like Ashe Dryden, the software developer behind Programming Diversity, would point out that people can only help crowdsource if they have the inclination and the time. Think back to the example of GitHub. If you were a woman, and you knew your contributions to a programming project were less likely to be accepted than if you were a man, would that motivate you to contribute the project? Or, for another example, Wikipedia. While the exact demographics of Wikipedia contributors are unknown, numerous surveys have indicated that those who contribute content to the crowdsourced encyclopedia are between 84% and 91.5% male. Why? It could be that there, too, edits are less likely to be accepted if they come from female editors. It could also go back to the housework argument. A 2011 study showed that women spend more than twice as much time on household tasks than men do, even when controlling for women who hold full-time jobs. Women simply don't have as much time.

No one would argue with the fact that time is money, but it's important to remember to ask whose time is being spent, and whose money is being saved. The premise behind Amazon's Mechanical Turk, or MTurk, as the crowd-sourcing platform is more commonly known, is that data scientists want to save their

own time, and their own bottom line. The MTurk website touts its access to a "global marketplace" of "on-demand Workers," who are advertised as being more "scalable and cost-effective" than the "time consuming [and] expensive" process of hiring actual employees. But the data entry and data processing tasks performed by these workers earn them less than minimum wage, even as a recent study by the Pew Research Center showed that 51% of U.S.-based Turkers, as they are known, hold college degrees; and 88% are below the age of 50, among other metrics that would otherwise rank them among the most desired demographic for salaried employees.

This *underwaged work*, as feminist labor theorists would call it, is also increasingly outsourced to countries with fewer (or worse) labor laws, and fewer (or worse) opportunities for economic advancement. A 2010 University of California-Irvine study measured a 20% drop in the number of U.S.-based Turkers over the eighteen months that it monitored. This trend has continued, the real-time MTurk Tracker shows, with workers from India alone now comprising roughly 20% of the total MTurk workforce. (The gender split, interestingly, has evened out over time).

But even in the United States– and even at companies like Amazon and Google– the work of data entry is profoundly undervalued in proportion to the knowledge it helps to create. Andrew Norman Wilson's 2011 documentary, *Workers Leaving the Googleplex,* exposes how the workers tasked with scanning the books for the Google Books database are hired as a separate but unequal class of employee, with ID cards that restrict their access to most of the Google campus, and that prevent them from enjoying the company's famed employee perks. (Evidently, working overtime to preserve the world's cultural heritage still does not entitle you to a free lunch, let alone a free class on how to cook Pad Kee Mao.)

Wilson also observes that Google's book-scanning workers are disproportionately women and people of color– a fact that would not surprise the long line of women of color scholar-activists, including Angela Davis, Patricia Hill Collins, and Evelyn Nakano Glenn, who have insisted that economic oppression be recognized as a vector that cuts across the matrix of domination as a whole. (See *The Power Chapter* where we talk more about this idea). Information studies scholar Lilly Irani confirms that "today's hierarchy of data labor echoes older gendered, classed, and raced technology hierarchies." Here, Irani is referring to the underwaged contributions of the first generation of female computers like Christine Darden, who we met in this book's introduction, who had to resort to NASA's Equal Opportunity Office in order to receive her long overdue raise; or, for another example, the below-minimum-wage pay of the Navajo women who, in the early days of digital computing, were tapped to assemble integrated circuits for the largest electronics supplier in the country, Fairchild Semiconductor–a story that Lisa Nakamura has recently exposed.

Irani's own research focuses on Mechanical Turk, the people it employs, and the people it exploits. As part of this work, Irani built a web tool, the Turkopticon, which enables Turkers to anonymously report unfair labor conditions, as well as

Figure 6.4: Andrew Norman Wilson's "Workers Leaving the Googleplex" (2011) documents the hidden inequities at Google's Mountain View headquarters. ¶ Credit: Andrew Norman Wilson ¶ Source: http://www.andrewnormanwilson.com/WorkersGoogleplex.html

any additional information that might help them decide whether to accept any future task.

But the people who perform this "cultural data work," as Irani terms it, are not only found at Amazon; they're increasingly the people on whom the entire information economy depends. Cultural data workers are responsible for everything from transcribing audio clips to fine-tuning search algorithms. Even Google relies upon people to confirm the quality of its search results, as official job postings for "Ads Quality Raters" confirm. Among more specific skills like a college degree and "excellent written communications skills," the job ad specifies that "a deep understanding of the culture is required."

Cultural data workers are also responsible for the invisible labor involved in moderating the veritable deluge of content produced online every day, ensuring that your Facebook feed is free of dick pics–and, much more disturbingly, videos of beheadings. When a recent exposé in *Wired Magazine* documented the emotional costs of this labor, performed by some of the least empowered of these workers–women in the Global South–it was met with an outpouring of shock and outrage. But those who study global capitalism for a living would be quick to point out that this exploitation of *racialized labor*, as they'd term it, has a long and sordid history, one that has its roots in the original form of human exploitation: slavery.

There is an infamous story that is often told in order to illustrate the close

connection between capitalism and slavery: in 1781, the British slave ship, *Zong*, made a series of navigational errors while crossing the Atlantic, resulting in a shortage of drinking water for the 17 crew members and 133 captives on board. After performing a cost-benefit analysis, the crew decided to throw their enslaved human "cargo" overboard, calculating that they could collect enough insurance money on that loss of life to come out ahead. For scholars such as Ian Baucom and, more recently, Fred Moten and Stefano Harney, there is no clearer example of the all-too-easy exchange between people and profit that capitalism enabled then, and still enables today.

Not by coincidence does our present global technological infrastructure follow this same pattern of exploitation, as Miriam Posner, along with Robert Meija and Safiya Noble, among others, have shown. The cobalt required to produce the lithium-ion batteries that power our cell phones and laptops, for example, may no longer be mined by people in physical shackles, but its extraction is still associated with significant human rights violations, including coercing labor from Congolese children as young as seven. The unregulated disposal of this and other minerals, as well as the electronics that house them, have resulted in entire cities along the west coast of Africa, as well as in China, becoming toxic "e-waste" sites. The humanitarian and ecological stakes couldn't be higher, nor could their source be any more clear: the capitalist forces that encourage the exploitation of Black bodies so that white bodies can thrive.

The weight of these forces can seem overwhelming. But, as any activist would remind you, any global resistance must begin at home. It follows, then, that we can each start by working harder to acknowledge the range of people who have contributed to our own projects, as well as those whose labor we might inadvertently exploit.

How, more specifically, can we go about this? For one model, we might look to an existing subfield of information visualization, known as data provenance. Data provenance typically refers to the practice of visualizing the history of the changes to a dataset that take place over the course of a project. Because the practice derives, ironically, from supply chain management, most extant data provenance visualizations focus solely on the data itself. They document which fields from which databases were combined with which data from which API, and which technical processes were performed on those data. But just as the ships on the Ship Map are piloted by people–people making real-world decisions like whether to steer around pirates in Somalia; or whether to wait on line for the Suez Canal–we might improve upon existing data provenance diagrams to include human processes as well. Who first collected the data, or processed it for use? What was the workflow employed by the design team, and how could it be plotted in relation to the data provenance chart? Was there a point at which the data analysis phase shifted from exploration to confirmation; or were there other significant conceptual shifts that could be rendered visible? These are only some of the questions that might be answered in such a diagram so that everyone involved in the data analysis process could receive credit for their work, and any

dependencies–both intellectual and interpersonal–could be acknowledged.

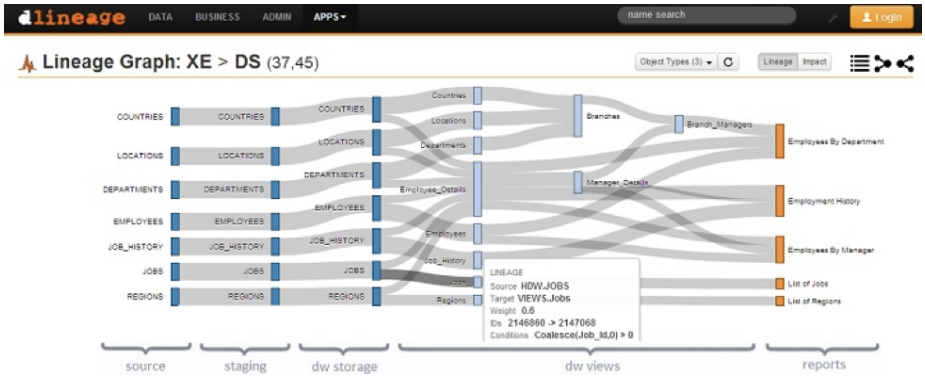


Figure 6.5: A data provenance chart. ¶ Credit: Dlineage. Screenshot by Catherine D'Ignazio. ¶ Source: dlineage.com ¶

One level up, at the level of labor itself, we might take an additional cue from the Next System Project, a research group aimed at documenting and visualizing alternative economic systems. In one report, the group compiled information on the diversity of community economies operating in locations as far-ranging as Negros Island, in the Philippines, Quebec province, in Canada, and the state of Kerala, in India. Their report employs the visual metaphor of an iceberg, in which "wage labor" is positioned at the tip of the iceberg, floating above the water, while dozens of other forms of labor– "informal lending," "consumer cooperatives," and work "within families," among others– are positioned below the water, providing essential economic ballast, but remaining out of sight.

With the idea of underwater labor in mind, we might return to the example of GitHub, which begins this chapter, in order to ask what additional forms of labor might contribute to the production of code, but that cannot be represented by the visualization scheme that GitHub currently employs. We might also think of the work of the project manager, which is not directly expressed in a particular number, or size, or frequency, of contributions, but nevertheless ensures the quality and consistency of all project code. We might wonder about the work of the designer on a project, or of the technical writer– both of whom might have helped to shape the project in its initial phases, but who have likely moved on to other tasks. We might additionally consider the contributions of the user experience specialist, or the quality assurance tester, who might enter the development process at a later phase of the project, but whose work is no less essential to the project's ultimate success. In the case of a consumer-facing project, we might also consider the contributions of the sales or customer support teams. These forms of labor, both productive and reproductive, are of course essential to the success of the project, but are not currently rendered visible, nor could they ever be easily visualized, by a scheme that considers project contributions to consist of code alone.

Figure 6.6: The Next System Project, "Cultivating Community Economies" ¶ Credit: J.K. Gibson-Graham, Jenny Cameron, Kelly Dombrowski, Stephen Healy, and Ethan Miller for the Next System Project ¶ Source: https://thenextsystem.org/cultivating-community-economies ¶ Permissions: Pending

When designing data products from feminist perspectives, we must aspire to show the work involved in the entire lifecycle of the project, even if it can be difficult to do. Whether it be a team of software developers working on GitHub, a team of visualization designers at Kiln, or a group of sugarcane farmers in the Philippines, a feminist approach would insist on recognizing the range of communities that produce the data. It would include the people who then collect, digitize, and transform it into a dataset; those who are subsequently enlisted to process the dataset; those then work to analyze and/or visualize the dataset; and finally, those who interpret the images or interactions that are produced, or otherwise experience their effects. Each of these roles is essential to the process of producing knowledge, but relies upon a variety of forms of labor– some visible, some not– in order to take place.

Showing all of this work is a tall order, and as designers and data analysts ourselves, we'll be the first to admit that it's not always one that can be fully achieved. But showing the work, as this chapter is named, begins with a commitment to acknowledging the range of forms of work that have been performed, even if they can't be ascribed to a specific person or credited by a single name.

In more instances than you might think, this work can be surfaced from the data themselves. For instance, Benjamin Schmidt, whose research centers on the role of government agencies in shaping public knowledge, decided to visualize the metadata associated with the digital catalog of U.S. Library of Congress, the largest library of the world. Schmidt's initial goal was to understand the collection and the classification system that structured the catalog. But in the process of visualizing the catalog records, he discovered something else: a record of the labor of the cataloguers themselves. When he plotted the year that each book's record was created against the year that the book was published, he saw some unusual patterns in the image: shaded vertical lines, step-like structures, and dark vertical bands that didn't match up with what one might otherwise assume would be a basic two-step process of 1) acquire a book; and 2) enter it in.

The shaded vertical lines, Schmidt soon realized, showed the point at which the cataloguers began to turn back to the books that had been published before the library went digital, filling in the online catalogue with older books. The step-like patterns indicated the periods of time, later in the process, when the cataloguers returned to specific subcollections of the library, entering in the data for the entire set of books in a short period of time. And the horizontal lines? Well, given that they appear only in the years 1800 and 1900, Schmidt inferred that they indicated missing publication information, as best practices for library cataloguing dictate that the first year of the century be entered when the exact publication date is unknown.

With an emphasis on showing the work, these visual artifacts should also prompt us to consider just how much physical work was involved in converting the library's paper records to digital form. The darker areas of the chart don't

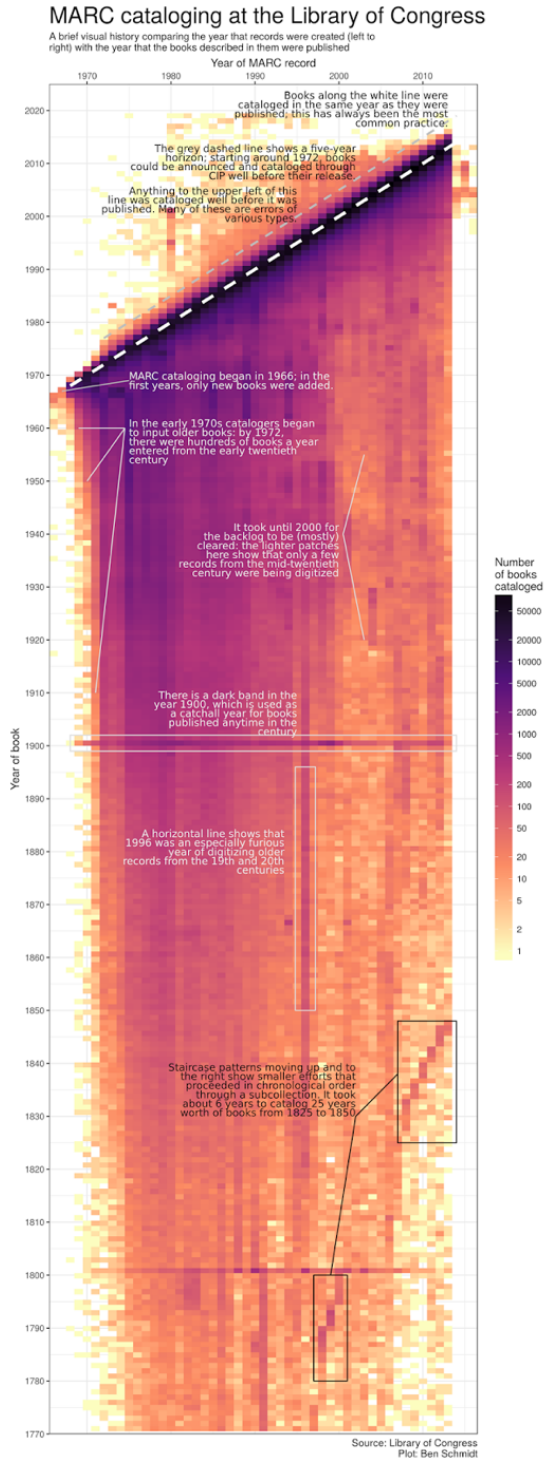


Figure 6.7: A visualization of when books at the Library of Congress entered their digital catalog. ¶ Credit: Benjamin M. Schmidt ¶ Source: http://sappingattention.blogspot.com/2017/05/a-brief-visual-history-of-marc.html ¶ Permissions: pending

just indicate a larger number of books entered into the catalog, after all. They also indicate the people who typed in all of those records–millions and millions of them. (Schmidt estimates the total number of records at ten million and growing). Similarly, the step-like formations don't just indicate a higher volume of data entry. They indicate strategic decisions made by library staff to return to specific parts of the collection, and reflect those staff members' prior knowledge of the gaps that needed to be filled. In other words, their intellectual labor as well.

There is also a political dimension of this work. For instance, in 1996, we see a dark vertical line that Schmidt tell us indicates "an especially furious year of digitizing older records." Could it possibly be that, in the lead-up to the presidential election that would result in Bill Clinton's second term, the federal government granted additional funding to the Library of Congress that enabled them to hire additional staff? Or did it indicate the fear that the Republican candidate, Bob Dole, would win the election and reduce the amount of funding for federal agencies, leading to the existing cataloguers to redouble their efforts? We can't know the answer without additional research, but these questions help to show how the dataset always points back to the data setting–a term coined by Yanni Loukissas, which we introduce in Chapter Four–and to the people who labored in that setting in order to produce the data that we see.

The people who labor in office buildings, and on top of them, are the subject of *Builders of the Vision,* by Daniel Cardoso Llach. That project employs data collection, as well as visualization, in order to analyze the hidden social hierarchies at work in international construction projects. For *Builders of the Vision*, Cardoso Llach wrote a script to register the design conflicts between different versions of the plans for the Thomas Wynne Mall, which recently opened along a stretch of otherwise desolate road in Abu Dhabi. In any large building project, design conflicts are common; they emerge when different subcontracting teams–say, the architects and the mechanical engineers–employ different software to draw up their plans for the project. It then becomes the task of a project coordinator to translate those plans into a single file format, identifying and resolving any inconsistencies along the way. Cardoso Llach's code sat on top of the project management software, recording the date and time of each conflict, the subcontracting teams involved, and the team whose design was accepted. By visualizing this conflict history, he was able to expose the evolving work patterns and shifting power dynamics within the project as it reached completion. In so doing, Cardoso Llach also exposes the hidden complexity of computational labor today.

Of course, there is also labor that remains hidden because we are not trained to think of it as labor at all. This is what's known as *emotional labor*, and it's (yet) another form of work that feminist theory has helped to bring to light. As described by feminist sociologist Arlie Hochschild, emotional labor describes the work involved in managing one's feelings, or someone else's, in response to the demands of society or a particular job. Hochschild coined the term in the

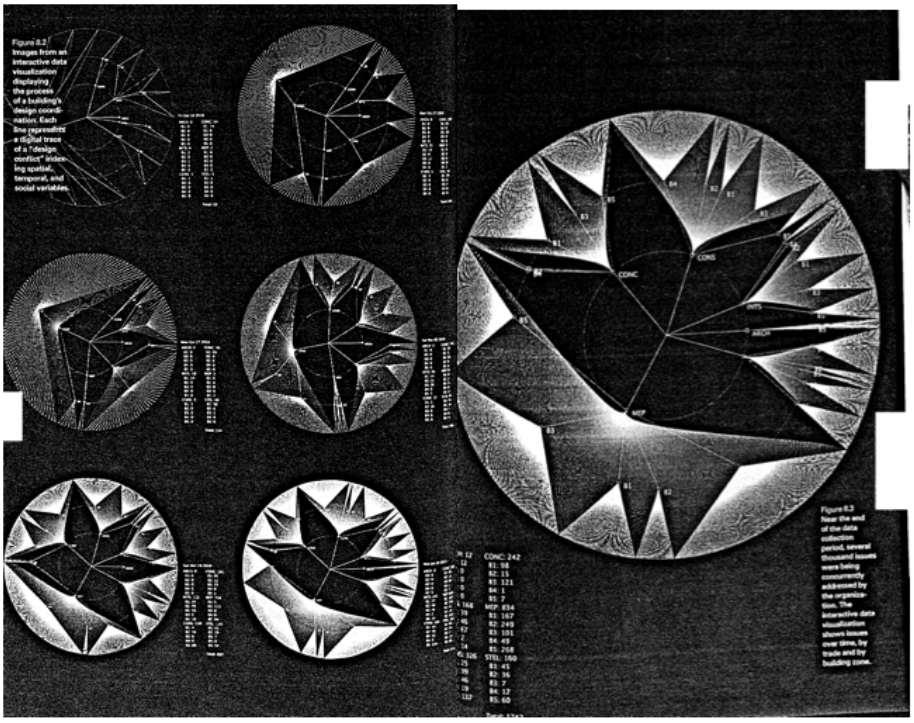


Figure 6.8: A series of visualizations documenting design conflicts within an architectural project's Building Information Management (BIM) software. ¶ Credit: Daniel Cardoso Llach ¶ Source: Builders of the Vision (Routledge, 2015), p. 132 and 133. ¶ Permissions: Pending

early 1980s to describe the labor required of service industry workers, such as flight attendants, who are required to manage their own fear while also calming passengers, during adverse flight conditions. In the decades that followed, the notion of emotional labor was supplemented by a related concept, *affective labor,* so that the work of projecting a feeling (the definition of emotion) could be distinguished from the work of experiencing that feeling (the definition of affect).

We can see both emotional and affective labor at work all across the technology industry today. Consider, for instance, how call center workers and other technical support specialists must exert a combination of affective and emotional labor, as well as technical expertise, in order to absorb the rage of irate customers (affective labor), reflect back their sympathy (emotional labor), and then help them with–for instance–the configuration of their wireless router (technical expertise). In corporate headquarters, we might also consider the affective labor required by women and underrepresented groups of all kinds, in all situations, who must take steps to disprove (or simply ignore) the sexist, racists, or otherist assumptions they face– about their technical ability, or about anything else. And they must do so while also performing the emotional labor that ensures that they do not threaten those who hold those assumptions, who often also hold positions of power over them. Are there ways to visualize these forms of labor, giving visual presence–and therefore acknowledgement and credit–to these outlays of work?

One example to prompt our thinking can be found in the *Atlas of Caregiving*, an ongoing project aimed at documenting the work involved in caring for a chronically ill family member. The project's name plays on the concept of the anatomy atlas, a compendium of illustrations of the human body that doctors can consult for information and reference. In this case, the goal was to illustrate the sometimes physical, and sometimes emotional or affective work of care. The research team outfitted its participants with a variety of biometric sensors, including accelerometers and heart-rate monitors, as well as with body cameras programmed to take a picture every fifteen minutes. They then visualized these data alongside excerpts from personal interviews, as well as from the activity logs they asked the caregivers in the study to complete. The result is a complex picture of caregiving, one that marshalls data in the interest of creating a comprehensive view of the range of labor involved in caregiving work.

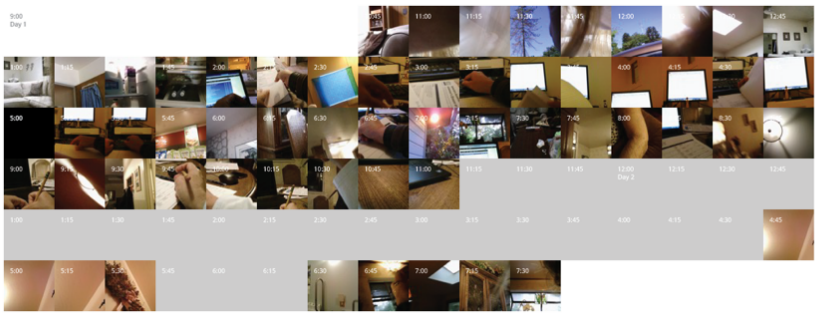


Figure 6.9: Caption: Clockwise from top left: A 36 hour log of caregiving activities; caregiving activities separate by type; photo log during that same time. ¶ Credit: The Atlas of Caregiving ¶ Source: https://atlasofcaregiving.com/studies/chantals-household/chantal/24-hour/ ¶ Permissions: Pending

Of course, a measure like heart rate or skin temperature is only a proxy for human feelings, and this is a common critique of the Quantified Self movement overall. This understanding served as the genesis for "Bruises: The Data We Don't See." This artful visualization, created by visualization designer Giogia Lupi, and accompanied by a musical score composed by Kaki King, attempts to make visible the emotional toll of parenting a child with chronic illness, as King herself was required to do when her own child was diagnosed with a rare autoimmune disease. Her daughter's illness, Idiopathic Thrombocytopenic Purpura, or ITP, is described as a "very visual disease," and presents as bruises and burst blood vessels all over the body. For this reason, King was instructed to watch her daughter's skin and record any significant changes. She also thought to record her own feelings in terms of hope, stress, and fear, creating subjective data to complement the hard numbers she received from the blood tests her daughter was required to endure.

When Lupi, who knew King from previous collaborations, set out to design her visualization, her goal was to "evoke empathy," and make her audience "feel a part of a story of a human's life." In contrast to the Atlas of Caregiving, which relies upon standard visualization techniques like radial timelines and Gantt-style charts in order to legitimate the work of care, Lupi sought alternative visualization strategies that would evoke the emotions she sought. She employed a fluid timeline to reflect the subjective nature of what feminist disability studies scholar Alison Kafer calls "crip time." Days become white aspen-shaped leaves, segmented not by weeks or years but by hospital visits. Red dots indicate platelet counts, with color deployed mimetically in order to convey the intensity

Figure 6.10: Detail from "Bruises: The Data We Don't See" ¶ Credit: Giorgia Lupi and Kaki King ¶ Source: https://medium.com/@giorgialupi/bruises-the-data-we-dont-see-1fdec00d0036 ¶ Permissions: Pending

of the bruises, as well as the visuality of the "data" recorded by King. Lupi also employed color to represent King's record of her feelings, with black corresponding to stress and fear; and yellow to signify hope. King's fear and hope are also visualized by hand-drawn lines that reflect each on a scale of one to ten. The result is both visually and aurally affecting composition of the affective labor of mothering and care.

Lupi and King, or the Atlas of Caregivers, are not the first to want to identify and make visible the work of care. Since the mid-1990s, when Nancy Folbre introduced the term, *care work* has been a significant topic of interest for feminist scholars. Folbre's primary model of care work was the everyday work of caring for a child. But we might also think of additional burden of caring for a sick child, as documented in "Bruises," or for a family member, as in the "Atlas of Caregivers." Care work isn't necessarily performed for free. It can also include the underwaged work performed by daycare workers or home health aids, as well as the waged work of doctors, nurses, physical therapists, mental health professionals, and so on. What binds these forms of work together across economic lines is their motivation: as theorized by Folbre, care work is undertaken out of a sense of compassion with, or responsibility for others, rather than with a goal of monetary gain. But when it comes to the market, altruism is a double-edged sword. These same professional care-workers–who are predominantly women and people of color–are often paid less than they would be in other fields. Why? Because they care.

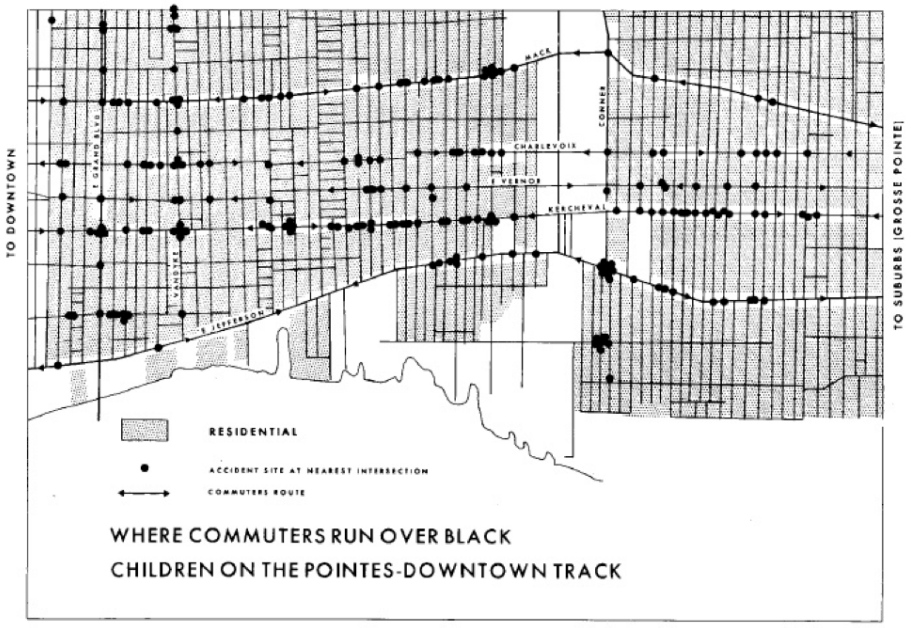
A similar commitment to others–and, increasingly, an awareness of how care

can also oppress–is what has prompted groups like The Maintainers to take up theories and practices of care in relation to data work. Through a series of workshops, conferences, and publications, The Maintainers are trying to counter the current tendency to celebrate technological innovation and discovery. The work that should be celebrated, they argue, is the work that sustains and maintains the world we live in today; and not work that passes over the problems of the present in order to look ahead. But there is yet more work to be done. Data work, after all, is part of a larger ecology of knowledge. Like the network of ships visualized on the Ship Map, or the network of source code stored on GitHub, the network of people who contribute to data projects is vast and complex. We may never be able to acknowledge all of the work that goes into these projects, at least not explicitly. But thinking through the many of forms of labor involved in data science, by making them visible, and by giving them credit, might be the most important data work of all.

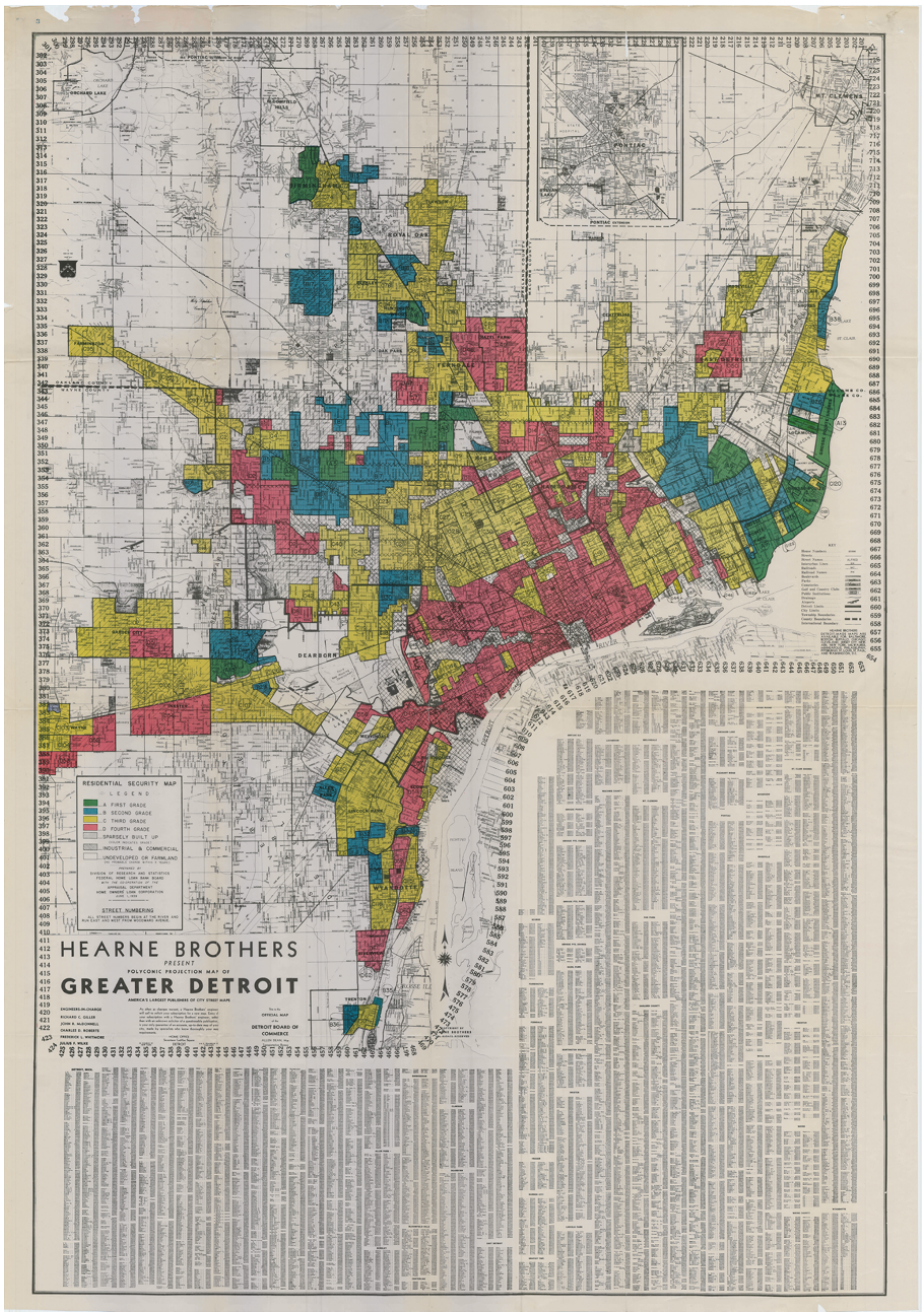
Chapter 7

The Power Chapter

In 1970, the Detroit Geographic Expedition and Institute released a provocative map, titled "Where Commuters Run Over Black Children on the Pointes-Downtown Track". The map starkly shows where many Black children were killed. On one single corner alone, there were six children killed by white drivers over the course of six months. Just gathering the data that the community already knew to be true posed a difficult problem. No one was keeping detailed records of these deaths, nor making them publicly available. The only reason it ended up being collected and published was because of an unlikely collaboration formed between low-income, urban, Black youth led by Gwendolyn Warren and white male academic geographers.



Contrast this map with a map made thirty years prior by the (all white and male) Detroit Chamber of Commerce and the (all white and male) Federal Home Loan Bank Board. This map set the stage for "redlining", a discriminatory practice of rating the risk of home loans in particular neighborhoods based on residents' demographics (their race, not their creditworthiness). Redlining began as a visual technique of red shading for all the neighborhoods in a city that were deemed "undesirable" for granting loans. All of Detroit's Black neighborhoods in 1940 fall in red areas on this map. Denying loans to Black residents set the stage for decades of structural racism and blight that was to follow.



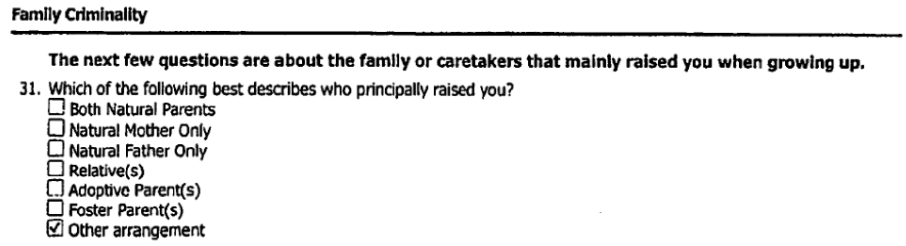
Both of these maps use straightforward cartographic techniques: aerial view, legends and keys, color and shading, to indicate different characteristics. But what is starkly, undeniably different about the two maps are the worldviews of the makers and their communities. In the second map you have the racist,

male-dominated city and federal institutions seeking to further institutionalize segregation and secure white wealth. Black neighborhoods were deemed to pose a "high risk" to the financial solvency of white institutions, so redlining maps became a way to systematically and "scientifically" protect white resources. These institutions succeeded, in no small part because of maps like this one. In contrast, in the first map you have a community who had recently learned the cutting-edge geographic techniques of their era who decided to take action against those same structures of power that created the first map. One is a map of securing power and the other is a map contesting power.

Who makes maps and who gets mapped? The DGEI map is, unfortunately, a rare instance in which communities of color, led by a young Black woman, determined what they wanted to map. It is more frequently the case that communities of color are mapped by institutions in power, whose worldviews and value systems may differ vastly from those of the community. One of the most dangerous outcomes of this imbalance of power – in evidence in this example of harm that was inflicted on people systematically for decades using maps and data – is when those institutions in power obscure their political agendas behind a veil of objectivity and technology.

This veil is not just a historical phenomenon. One can make a direct comparison between yesterday's redlining maps and today's risk assessment algorithms. The latter are used in many locales to inform whether a person who has been detained should be considered at low or high risk of committing a future crime. Risk assessment scores can affect whether a person is let out on bail and what kind of sentence they receive – they have the power to set you free or lighten your sentence.

The issue is that different bodies are differently weighted by the risk assessment algorithm. For example, in 2016 Julia Angwin led a team at ProPublica to investigate one of the most widely used risk assessment algorithms created by the company Northpointe (now Equivant). Her team found that white defendants were more often mislabeled as low risk than Black defendants, and conversely, that Black defendants were mislabeled as high risk more often than white defendants. Digging further into the details, the journalists uncovered a 137-question worksheet that detainees fill out. Their answers feed into the software and are compared with other data in order to spit out the risk assessment score for the individual. While the questionnaire does not ask directly about race, it asks questions that are direct proxies for race, like whether the you were raised by a single mother, whether you have friends or family that have been arrested, and whether you have ever been suspended from school. In the US context, each of those data points has been demonstrated to have disproportionate occurrences for Black people – 67% of Black kids grow up in single parent households, for example, whereas the rate is only 25% for white kids. So, while the algorithm creators claim that it isn't considering race, it is considering race by proxy and using that information to systematically disadvantage Black and brown people.



Family Criminality

The next few questions are about the family or caretakers that mainly raised you when growing up.

31. Which of the following best describes who principally raised you?
- ☐ Both Natural Parents
- ☐ Natural Mother Only
- ☐ Natural Father Only
- ☐ Relative(s)
- ☐ Adoptive Parent(s)
- ☐ Foster Parent(s)
- ☑ Other arrangement

The redlining map and the Northpointe risk assessment algorithm have a lot of similarities. Both use the cutting-edge technologies and aggregated data about social groups for institutions to make decisions about individuals – Should we grant a loan to this person? What's the risk that this person will re-offend? Both use past data to predict and constrain future individual behaviors. Note that the past data in question (like segregated housing patterns or single parentage) are a product of structurally unequal conditions amongst social groups, and yet the technology uses those data as a causal element that will influence an individual's future behavior. Effectively this constitutes a demographic penalty that tracks an individual through their lives and limits their future potential – Live in a Black neighborhood? Then you don't get a loan. Raised by a single mom? Then you can't be freed on bail because you are a flight risk. And the kicker is that because of their use of tech and data, both of these racist data products have the appearance of neutrality. Scholar Ruha Benjamin has a term for this – "the New Jim Code" – a situation which combines software code and imagined objectivity to contain and control Black and brown people.

What's the alternative? Let us for a moment imagine a completely different set of values to encode in our data products. The values in evidence in redlining maps and risk assessment algorithms are about preserving a race- and class-based status quo. White, wealthy men working in powerful institutions adopt a focus on risk – a single loan in default threatens to decrease the wealth of their institution and the data and computational systems are mobilized to avoid this possibility at all costs. But instead of penalizing people for their statistical affiliation with specific race, gender and class demographics, we could imagine an alternative approach grounded in equity and demographic healing. A system could mobilize the same data – say, zip code and neighborhood demographics – to determine where more strategic investment was needed to counteract the toxic effects of structural inequality. And when people applied for loans, the red color in certain neighborhoods would indicate their higher need and place them higher up in the priority line for individual loans.[1] The values in our alternate world are not about preserving the dominance of certain institutions and elite people but about equalizing the effects of structural inequality. Sharing power and wealth could easily be hardcoded into the computational systems of the future. The data and technology would remain almost the same but the values driving their

[1]The concepts taught address specific mathematical content and skills outlined by the Common Core State Standards in New York.

use (and the people who derive benefit from their use) would be almost exactly opposite. But this alternate world won't happen of its own accord. As Frederick Douglass stated in 1857, and as Yeshimebeit Milner recently reminded Data for Black Lives members: "Power concedes nothing without a demand."

What is the demand? Which demands and on behalf of whom? In order to formulate those demands it is important to do two things: examine how power is currently wielded with and through data and, in parallel, imagine and model how things could be different. Examining intersecting dimensions of power has long been part of a feminist toolkit. Back in 1977, the Combahee River Collective, the famed Black lesbian activist group out of Boston, urgently advocated for "the development of integrated analysis and practice based upon the fact that the major systems of oppression are interlocking."

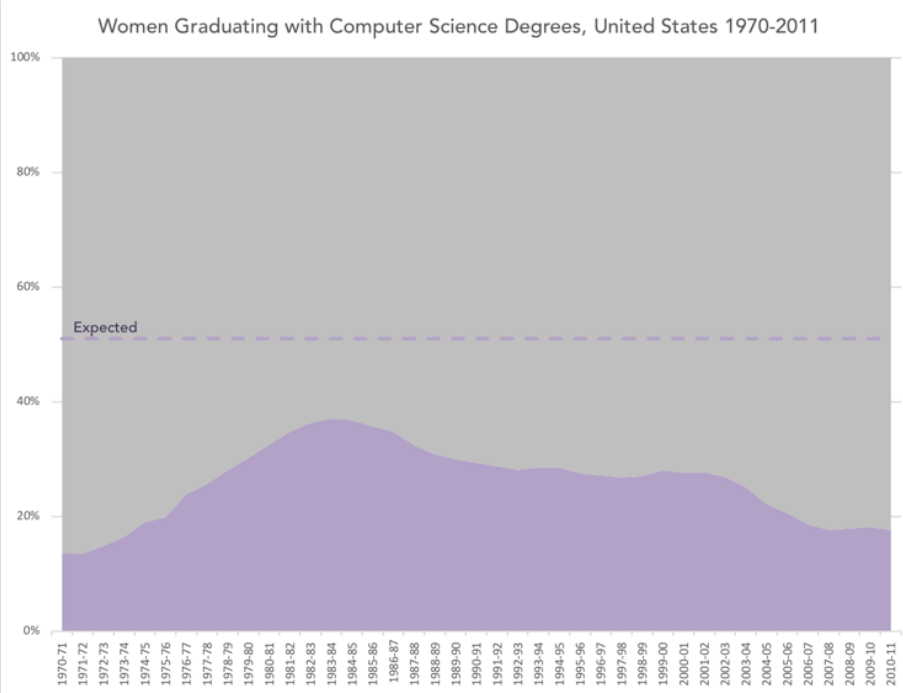
Examining how power is wielded through data means doing projects that wield it back like Warren's map and ProPublica's Machine Bias story – These deal openly and explicitly with who has power and who doesn't, as well as naming the structural conditions like racism and sexism that underlie those facts. It involves lifting the veil of what Benjamin calls the "imagined objectivity" of code and exposing the differential harms and benefits resulting from the deployment of data science. Good work in this vein is emerging from spaces like activism, journalism,[2] machine learning, [3] and law.

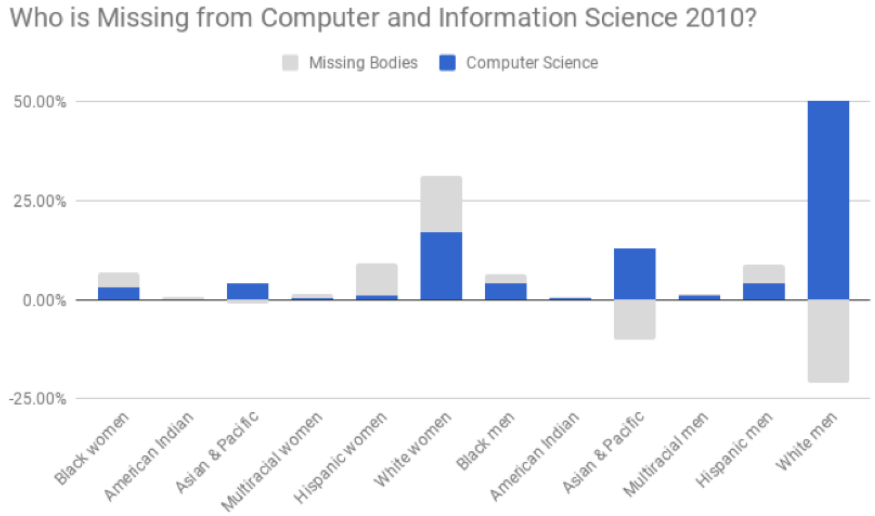
But data science and visualization work that examines power still mostly happens around the margins of the field, for three reasons. First, unless you work in an accountability field (such as journalism or law), there typically isn't funding or other professional incentives for such work. Corporations typically want to visualize their supply chain, not their sexism.

Second, the people that have access to data and to the technical skills to work with it are those that have the most stake in reproducing the status quo. The elephant in the server room, only very occasionally acknowledged, goes back to one of the issues that we raised in Bring Back the Bodies: that women and people of color are not well-represented in the fields of data science and visualization, and the problem is getting worse. In the graphic below, you can see that female graduates in Computer/Information Science in the US peaked in the mid 1980's at 37%. We have seen a slow decline in the years since then. The rate of female graduates in 2010-11 fell below the rate of female graduates in 1974-5. What this means is that the most highly-touted methods of producing knowledge and deriving insight in the age of big data and artificial intelligence are being designed and deployed primarily by the people with the most privilege.

[2] CS109 at Harvard is taught jointly by Computer Science and Statistics. As of this writing, there are 37 male faculty (69%) and 17 female faculty (31%).

[3] This does not mean there are no data ethics courses, only that it is not the norm to address these concerns in introductory coursework. Indeed, there is a long list compiled by Dr. Casey Fiesler of technical courses that specifically address ethics and what is being called "fairness, accountability and transparency" in technical fields: http://bit.ly/tech-ethics-syllabi





Relatedly, the third and final reason that examining power is the exception rather than the norm is that, as feminist sociologist Michael Kimmel says, "privilege is blind to those who have it."

What does this mean? If you remember Kimmel's colleague's powerful statement

from Chapter X, it went like this. His African-American colleague said, "When I look in the mirror I see a Black woman. When a white woman looks in the mirror she sees a woman." And Kimmel, a white man, rejoins, "And when I look in the mirror, I see a human being." For people in the dominant group, their gender, race, sexuality or class is so normal that it is invisible. It is not seen as a marker of difference, but rather simply "the way things are". Take enough of those privileged individuals and put them together collectively at the helm of data science and algorithm development and you have a major structural deficiency. This basic imbalance of power remains mostly unacknowledged – except when it reveals itself in surprising and uncomfortable ways.

For example, Joy Buolamwini, a Ghanaian-American graduate student at MIT, was working on a class project using facial analysis technology. These are software packages that will detect a face in an image, similar to when your phone camera will create outlines around the people's faces that it "sees" in the picture. But there was a problem – the software couldn't "see" Buolamwini's dark-skinned face. It had no problem seeing her lighter skinned collaborators. When she drew a face on her hand and put it in front of the camera, it detected that. And then when Buolamwini put on a white mask, essentially going in "white face," the system detected the mask's facial features perfectly. Digging deeper into the code and benchmarking data behind these systems, Buolamwini discovered that the data set on which many of the facial recognition algorithms are tested contains 77.5% male faces and 83.5% white faces. When she did an intersectional breakdown of a separate test dataset – looking at gender and skin type together – only 4.4% of the faces in that data set were female and dark skinned. In their evaluation of three commercial systems, Buolamwini and Timnit Gebru showed that darker-skinned females were up to forty-four times more likely to be misclassified than lighter skinned males. No wonder the software was failing on faces like Buolamwini's if both the training data and the benchmarking data relegate women of color to a tiny fraction of the overall data set

As she tells it, “I didn’t start out on a mission for social justice,” but after seeing the need for more fairness and accountability, Buolamwini has now gone on to launch the Algorithmic Justice League (AJL) – an organization that works to highlight and address algorithmic bias. Buolamwini and the AJL have done art projects, written research papers, taken to the media to call for a moratorium on facial analysis and policing, and they are even advising on legislation and professional standards for the field of computer vision.

But imagine, for a moment, a world where female, Black and brown engineers are the ones designing the computer vision training data sets and algorithms in the first place. A world where people like Joy Buolamwini and Gwendolyn Warren are the norm, not the exception. Would it have worked from the start? Not necessarily, says Buolamwini. “No technologist works in isolation – we rely on the libraries and datasets developed by the community over time.” And these datasets have what she has termed “power shadows” – they reflect the structural inequality of the world they draw from. So when it is easiest to collect faces of powerful public figures for your benchmarking data, those datasets will contain power shadows - disproportionate male and white representation.

So what does “working” mean if you want to make data products that are anti-racist and anti-sexist? On the one hand, the software did “work”. It was pretty good at detecting faces for the white men who comprised 78% of the data set. But Buolamwini likes to remind her audiences that Europeans are less than 10% of the world’s population, so it didn’t work for the majority of the global population. And even so, “it’s not just about creating accurate algorithms but creating equitable systems,” she says. We can’t just be building more precise surveillance apparatuses but also look at the deployment, governance, use and impacts of these technologies – “Communities not companies should determine whether and how this technology is used by law enforcement.”

Where we might say that the technology did "work" is that it accurately reflected back to Buolamwini the biases of the people in power towards Black women. In that sense, it faithfully reinforced the racist messages Black people receive all the time that their lives as well as their voices, bodies and representations do not matter. bell hooks referred to this phenomena as "representational harms". Specifically writing about data, artist Mimi Onuoha has called this phenomenon "algorithmic violence" and data ethicist Anna Lauren Hoffmann has used the term "data violence" for the way in which it participates in (and legitimates) the circulation of damaging narratives and ideas about particular groups of people. This is the harm that occurs with imagined objectivity – when software engineers wield data "neutrally" (in an attempt to wiggle out of having to deal with squishy things like values) they build things that support the existing status quo. And that status quo is ugly – it is racist, patriarchal, heteronormative and more.

In fact, one of the structural forces that software engineers and data scientists need to contend with is that data is by and large a tool of management, wielded by those institutions in power, like the Detroit Chamber of Commerce in the 1940s, who have a vested interest in maintaining the ugly status quo because they benefit from it. Joseph Weizenbaum, artificial intelligence trailblazer and creator of the famous ELIZA experiment in the 1970s, looked back on the history of computing and said it like this: "What the coming of the computer did, 'just in time,' was to make it unnecessary to create social inventions, to change the system in any way. So in that sense, the computer has acted as fundamentally a conservative force, a force which kept power or even solidified power where is already existed."

The first step to pushing back against this fundamentally conservative force is to understand that the single most damaging thing one can do to uphold the racist, sexist order of the world is to claim that they have no values, no politics, and that their work with data is neutral. This is Haraway's god trick and Benjamin's imagined objectivity - the veil at work to obscure power differentials. This neutrality narrative would be item #1 in the BuzzFeed listicle "Things Men Tell Themselves to Stay on Top".

The second step is to begin to understand the ways that privilege – and oppression, its counterpoint – manifest themselves in data science. Privilege and oppression are complicated and there are "few pure victims and oppressors," as sociologist Patricia Hill Collins notes. A helpful way to start to grasp these functionings is through Collins' concept of the matrix of domination. As we described at the outset of this book, a core distinguishing feature of contemporary feminism is its insistence on intersectionality – the idea that we must take into account not only gender but also race, class, sexuality and other aspects of identity in order to fully understand and resist how power operates to maintain an unjust status quo. Collins' matrix of domination describes the overall social organization of those intersecting oppressions. She outlines four major domains in which the matrix of domination operates: the structural domain, the disciplinary domain,

the hegemonic domain and the interpersonal domain. "Each domain serves a particular purpose," writes Collins.

The *structural domain* is that of laws and policies and schools and institutions – it organizes and codifies oppression. If we take the example of voting, most US states prohibited women from voting in elections until the 1910s. Even after the passage of the Nineteenth amendment in 1920, many state voting laws included literacy tests and other ways to specifically exclude women of color,[4] so it wasn't until the Voting Rights Act in 1965 that all Black and brown women were enfranchised. The *disciplinary domain* administers and manages oppression through bureaucracy and hierarchy (rather than explicit laws). In our voting example, this might take the shape of a company prohibiting factory workers from leaving early to vote or penalizing workers who distribute information about voting.

Neither of these domains are possible without the hegemonic domain which deals with culture, media, and ideas. Discriminatory policies and practices in voting can only happen in a world that widely circulates oppressive ideas about who "counts" as a citizen. For example, an anti-suffrage pamphlet from the 1910s proclaimed that "You do not need a ballot to clean out your sink spout." This and other such memes of the era reinforced pre- existing societal notions that a woman's place is in the domestic arena, outside of public life. And the final part of the matrix of domination is the interpersonal domain which influences the everyday lived experience of individuals. For example, what would it feel like to be the butt of jokes made by males in your family as they read that pamphlet? How did it feel like to wait in line for twelve hours to cast your vote, knowing that the system was deliberately trying to screw you out of a voice?

If you are a Black woman in the US, you are intimately familiar with the matrix of domination because you brush up against it in everyday encounters. Writes Collins, "Oppression is not simply understood in the mind—it is felt in the body in myriad ways. Moreover, because oppression is constantly changing, different aspects of an individual U.S. Black woman's self-definitions intermingle and become more salient: Her gender may be more prominent when she becomes a mother, her race when she searches for housing, her social class when she applies for credit, her sexual orientation when she is walking with her lover, and her citizenship status when she applies for a job. In all of these contexts, her position in relation to and within intersecting oppressions shifts." In each of these cases, the woman is made aware of her differences and her subjugated position in relation to a dominant norm. This experience is an essential form of data – lived experience as primary source knowledge.

But let's imagine for a moment you are an middle-class, straight, white male US citizen. Your body doesn't change in childbirth and breastfeeding so you don't

[4]Stingrays are devices that mimic cell phone towers and trick cell phones nearby into connecting with them so that they can gather personal data. They are used primarily by law enforcement and their use is contested by the ACLU and other organizations concerned with privacy and civil liberties.

think about workplace accommodations. You look for a home or apply for a credit card and people are eager for your business. People smile or don't look twice when you hold your girlfriend's hand in public. You present your social security number in jobs as a formality, but it never hinders an application from being processed or brings unwanted attention. The ease with which you traverse the world is invisible to you because it is quite simply the way things are and you imagine they are the same for everyone else. This is what it means to be blind to your own privilege – despite having the best education, the most elite among us are pathetically deficient when it comes to recognizing injustice, across all of the domains in the matrix of domination. They lack the lived experience – the undeniable data of lived experience – that reminds them everyday that their bodies, their sexuality, and/or their race depart from a desired norm.

Projects that reveal those norms often focus on the absences and silences – those who are purposefully omitted or simply forgotten because of who has consolidated privilege and power. We've already introduced you to the work of artist and designer Mimi Onuoha in Chapter One. Her project, Missing Data Sets, if you recall, is a list she maintains of issues and events that go uncounted. Her missing data sets name important phenomena that you would expect institutions to collect systematic information about – topics such as police killings, hate crimes, sexual harassment, and caucasian children adopted by people of color.[5]

Onuoha exhibits Missing Data Sets as an empty set of tabbed file folders in art

[5] You probably know what WTF stands for. But csv stands for "comma separated values" and is a text-based spreadsheet file format. Each column break is denoted by a comma and each row break is denoted by a carriage return. You can open csv files in spreadsheet programs and most data software packages.

exhibitions. The viewer can browse the files and open the folders to reveal that there are no papers inside. What should be there, in the form of paper records, is "missing" – absent not because the topics are unimportant, but because of bias, social and political will, and structural disregard. As Onuoha says, "That which we ignore reveals more than what we give our attention to. It's in these things that we find cultural and colloquial hints of what is deemed important. Spots that we've left blank reveal our hidden social biases and indifferences."

What is to be done about missing data sets? Taking a feminist perspective in this unequal ecosystem can mean pointing at their absence, as in the case of Onuoha. Or, sometimes, it means walking right straight ahead into the unequal playing field and collecting the missing data yourself, because somebody has to do it.

This is exactly what pioneering data journalist and civil rights advocate Ida B. Wells did as early as 1895, when she assembled a set of statistics on the epidemic of lynching that was sweeping the United States at the time; or what Princesa, the anonymous Mexican woman who we introduced in Bring Back the Bodies, has been doing for the past three years. She has logged 2,355 cases of femicide since 2016,[6] and her work provides the most accessible information on the subject for journalists, activists and victims' families seeking justice.

Femicide is a term first used publicly by feminist writer and activist Diana Russell in 1976 while testifying before the first International Tribunal on Crimes Against Women. Her goal was to situate the murders of cis and trans women in a context of unequal gender relations. In this context, men use violence to systematically dominate and exert power over women. Indeed, the research bears this out. While male victims of homicide are more likely to have been killed by strangers, a 2008 report notes a "universal finding in all regions" that for women and femmes are far more likely to have been murdered by someone they know. Femicide includes a range of gender-related crimes, including intimate and interpersonal violence, political violence, so-called "honor" crimes, gang activity, and female infanticide.While such deaths are often depicted as isolated incidents, and treated as such by authorities, those who study femicides characterize them as a pattern of underrecognized and under-addressed systemic violence.

Femicides in Mexico rose to global visibility in the mid-2000's with widespread media coverage about the deaths of poor and working-class women in Ciudad Juárez. A border town, located across the Río Grande from El Paso, Juárez is a home to more than 300 maquiladoras – factories that employ many women to assemble goods and electronics, often for low wages and in substandard working conditions. Between 1993 - 2005, nearly four hundred women were murdered in the city, with around a third in brutal or sexual form. A conviction was made in only three of those deaths. When alleged perpetrators were arrested, they were often tortured into confessions by police, casting doubt on the investigations.

[6] "Civic data guides" is the name of the collaboration undertaken by Catherine, Yanni Loukissas and Bob Gradeck around the production of data user guides by students and learners.

Activist groups like Ni Una Más (Not One More) and Nuestras Hijas de Regreso a Casa (Our Daughters Back Home) were formed in large part by mothers who demanded justice for their daughters, often at great personal risk to themselves.[7] These groups succeeded in gaining the attention of the Mexican State who established a Special Commission on Femicide chaired by politician Marcela Lagarde. After three years of investigating, the Commission found in 2006 that femicide was indeed occurring and that the Mexican State was systematically failing to protect women and girls from being killed. Moreover, Lagarde suggested that femicide be considered, "a crime of the state which tolerates the murders of women and neither vigorously investigates the crimes nor holds the killers accountable."

Despite the Commission's work and the fourteen volumes of detailed accounts and statistics about femicide – As well as a 2009 ruling against the Mexican state by the Inter-American Human Rights Court; As well as a United Nations Symposium on Femicide in 2012; As well as the fact that sixteen Latin American countries have now passed laws defining femicide – despite all this, deaths in Juárez have continued to rise and the toll is now more than 1500. Three hundred women were killed in Juárez in 2011 alone, and only a tiny fraction of those cases have been investigated. The problem extends beyond Ciudad Juárez in the state of Chihuahua to other states in the nation such as Chiapas and Veracruz.

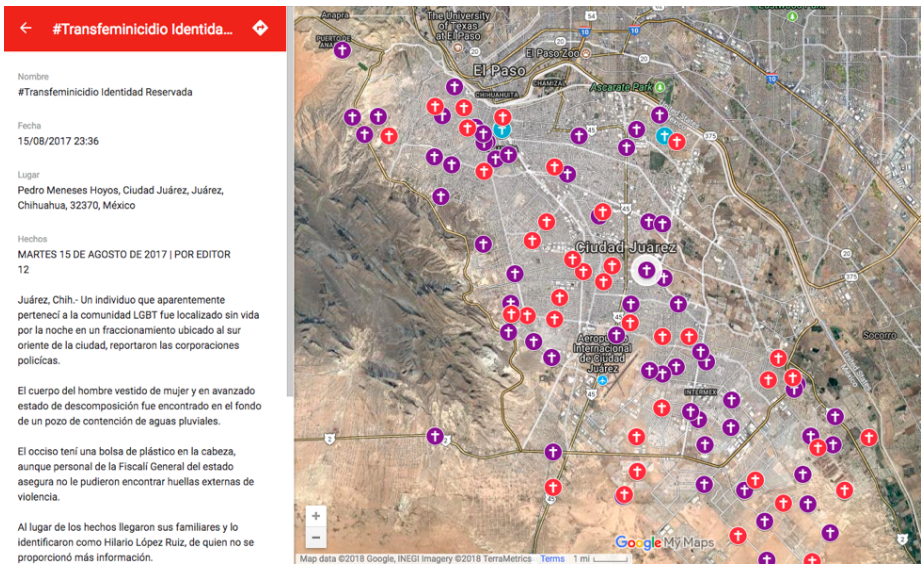
While there is increasingly a legal and analytical basis for characterizing deaths as femicides, there is still a great deal of missing data. In a report titled Strengthening Understanding of Femicide, the authors state that "instances of missing, incorrect, or incomplete data mean that femicide is significantly underreported in every region." In the case of femicides, as in so many cases of data collected (or not) about women and marginalized groups, the *collection environment is compromised*. Largarde's very definition of femicide includes the fact that the State – comprised mainly of privileged men who have a vested interest in maintaining a gendered order – is complicit through indifference and impunity, so how could data be reliably collected?

This circles us back to a point we first made in Chapter One, and elaborated in Unicorns, Ninjas, Janitors, and Rock Stars, about how collecting large amounts of data is costly and resource-intensive. Only government states, corporations, and some elite institutions have those resources, so data collection efforts tend to be driven by their values and priorities. Not surprisingly, those institutional actors can be compromised by their own privilege, and their interest in maintaining the status quo. In the case where a government state is itself the bad actor, there can be no other authority with enough resources or channels of influence to shift collection practices. This is especially true in the case of femicides, in which collecting high-quality data would rely on shifting policy for local law enforcement and medical examiners, the entities that log homicide information.

[7] This is not to say tools and individual skills are not important (they are), or that your co-authors have never led tool-focused workshops (we have). Rather, the problem when this is the only model of learning that is ever undertaken in a workshop or course.

But as data journalist Jonathan Stray asserts, "Quantification is representation." Looking at U.S. census data prior to 1970, he explains, you might come to the conclusion that there were no Latinx people living in the United States. This is not true, of course. There were actually already millions of Latinx people living in the U.S. But 1970 was the first year that "Hispanic" was included as an ethnic category on the census. Prior to that, it would have been hard to know anything about Latinx people as a group because the federal government was simply not collecting any information about them. So when the category was added to the census, most Latinx people were pleased to see it. It meant that they mattered.

But the inverse of the "quantification is representation" equation is also true: if data is not collected on a particular group, or on a particular issue, then institutions in power can pretend that the issue doesn't exist. Similar to the case of universities and sexual assault statistics, as discussed in *The Numbers Don't Speak for Themselves*, no Mexican state wants to have high rates of femicide. It is into this lack of government will that Princesa, who recently has spoken out in public under her given name María Salguero, has inserted her map of femicides. Salguero studied Geophysical Engineering in Mexico's Instituto Politécnico Nacional. She learned her mapping and journalism skills from attending trainings with Chicas Poderosas, a Latin American feminist group that focuses on training cis and trans women in data storytelling. The femicides map takes two forms – one, depicted in Figure 7.08a, is a point map where Salguero manually plots a pin for every femicide that she collects through media reports or through crowdsourced contributions. The other visualization, seen in figure 7.08b, consists of the same data in a dashboard format, with gender-related killings grouped as smaller or larger bubbles for different geographies depending on their incidence. One her goals is to "show that these victims had a name and that they had a life. They weren't statistics," so Salguero logs as many details as she can about each death. These include name, age, relationship with the perpetrator, mode and place of death, whether the victim identified as transgender, as well as the full content of the news report which served as the source. It can take her three to four hours a day to do this unpaid work (see Show Your Work for a further discussion on labor, gender, and data). She takes breaks for preserving her mental health, and she typically has a backlog of of a month's worth of femicides to add to the map.



While media reports and crowdsourcing are imperfect ways of collecting information, this map – created and maintained by an individual – fills a vacuum created by the government's deflection of responsibility. Mexico's National Health Information System (SINAIS) logs national homicide data, but only records the

name, location and how the person died, and to count a death as a femicide you must know the circumstances of the death as well as the relationship between the perpetrator and the victim. Various federal agencies point fingers in different directions regarding femicide data collection. In 2017, the Federal Institute for Access to Public Information and Data Protection (INAI) – led by Commissioner Ximena Puente de la Mora – ordered the National Commission for Human Rights (CNDH) to turn over statistics about femicides for 2015 and 2016. The CNDH declared itself incompetent to know such information and referred the request to two other federal agencies, neither of whom collect data about femicides.

In the meantime, Salguero's femicides map provides the most authoritative source of data on femicides at the national level. It has been featured in national Mexican media outlets and used to help find missing people. Salguero herself has testified before the Mexican Senate. Though Salguero is not affiliated with a specific group, she makes the data available to activist groups for their efforts. And parents of victims have called her to give their thanks for making their daughters visible. The urgency of the problem makes the labor worthwhile. Princesa affirms, "this map seeks to make visible the sites where they are killing us, to find patterns, to bolster arguments about the problem, to georeference aid, to promote prevention and try to avoid femicides."

How might we explain the missing data around femicides in relation to the four domains of power that constitute Collins' matrix of domination? The most grave and urgent manifestation is in the interpersonal domain, where women are victims of extreme violence and murder at the hands of men. And although the structural domain – law and policy – has recognized femicide, there are no specific policies implemented in order to ensure adequate information collection, either by federal agencies or local authorities. Thus, the disciplinary domain, where law and policy are enacted, is characterized by deferral of responsibility, failure to investigate, and victim blaming, precisely because there are no consequences in the structural domain.

And none of this would be possible without the hegemonic domain - the realm of media and culture – that presents men as dominant and women and women as subservient; men as public, women as private; with any challenge to this gendered order of operations perceived as a grave transgression, deserving of punishment. Indeed, government agencies have used their position to publicly blame victims. Following the femicide of 22- year-old Mexican student Lesvy Osorio in 2017, as Maria Rodriguez-Dominguez reports, the Public Prosecutor's Office of Mexico City shared on social media that the victim was an alcoholic and drug user who had been living out of wedlock with her boyfriend. Here was the office that was supposed to be investigating the murder, and instead of doing their job they turned to social media to imply that Osorio was a degenerate. This led to public backlash and the hashtag "#SiMeMatan (If they kill me)" and tweets such as "#SiMeMatan it's because I liked to go out at night and drink a lot of beer."

This is the data collection environment for femicide information and it is char-

acterized by extremely asymmetrical power relations, where those with power and privilege are the only ones who can actually collect the data but they have overwhelming incentives to ignore the problem, precisely because addressing it poses a threat to their dominance. Here it is important to note that femicides data is not an isolated case. It is an expected outcome and regular feature of an unequal society, in which a gendered, racialized order is maintained through willful disregard, deferral of responsibility and organized neglect for data and statistics about those bodies who do not hold power. For example, doctoral student Annita Lucchesi has created "The Missing and Murdered Indigenous Women Database" which tracks indigenous women who are killed or disappear under suspicious circumstances in the US and Canada. She thinks approximately 300 indigenous women per year are killed but the exact number is unknown because nobody (other than Lucchesi) is actually counting. Other examples in the US context include police killings of unarmed Black and brown people,[8] maternal mortality statistics, and people killed by US drones.

What is to be done? It's important to remember that asymmetrical power relations don't mean absolute power. And it's also important to remember that States and entities with power are not monolithic. There are plenty of public servants – women and men and others – in Mexico advocating internally for better data collection around femicides, like Ximena Puente de la Mora from INAI who initiated the femicides data request.

Crowdsourced data collection efforts that count and measure the extent of structural oppression can be a first step towards demanding public accountability. This is an important, urgent role for data journalism in the 21st century.[9] As we discussed in Bring Back the Bodies, ProPublica has an on-going investigative series about "Lost Mothers" – mothers in the US who lose their lives in childbirth due to poor care and preventable causes. One of their findings was that there was no comprehensive federal data on maternal mortality, so ProPublica began crowdsourcing stories of individuals to attempt to count the phenomenon. Their database and their reporting has spurred the creation of more than 35 state level review committees who are investigating maternal mortality in their state, as well as a proposed bill in Congress to allocate $12.5 million to the Centers for Disease Control and Prevention to undertake better data collection.

[8]The concept of a discotech was created by the Detroit Digital Justice Coalition in 2009 based loosely on the idea of a potluck event for technology. The goal of a discotech is to create "a genuine collaborative, collective learning environment that is accessible to all skill levels, ages, and learning styles." The first Data DiscoTech was run in 2015 as a response to a Detroit open data ordinance. The organizers were concerned about some of the harms that might arise from open data. "Public data impacts different people differently," as one organizer stated, so the goals of the discotechs have included capacity building as well as consciousness building. The Coalition has published a free guide to running your own discotechs here: https://www.alliedmedia.org/files/ddjc_zine_4.pdf

[9]KPIs are Key Performance Indicators – measures that help to evaluate the performance of an organization in regards to a particular activity that it has deemed important. For example, a nonprofit might track its fundraising efforts with a KPI like "Cost per dollar raised" – how much did they spend on fundraising for every dollar that they brought in.

But, at the same time, we also have to work on dismantling the consolidated power and privilege that organize the matrix of domination.

Could we statistically model oppression? It's a provocative question and one that Google researcher Margaret Mitchell has been investigating at the level of collective human speech. She describes how, in speech patterns, people use unqualified nouns for the "default" case of something. For example, bananas that are green are modified with "green bananas" or "unripe bananas" to indicate that they depart from the ready-to-eat yellow banana. But nobody needs to say "yellow banana" because it is implied by our shared concept of banana. This is called "reporting bias" in artificial intelligence research. So, studying the adjectives that modify "banana" in large data sets can actually tell us a lot about what people's default idea of bananas is in a particular culture. And when applied to humans, the "default case" reveals a lot about our collective norms and biases. For example, a doctor who is female is more typically qualified as a "female doctor" in human speech because it represents a departure from a perceived norm of doctors being male. So if "female doctor" is used in speech patterns for a particular culture, we might be able to infer that the social norms for that culture are patriarchal and thus pay special attention to the ways in which women are oppressed. Of course, this only works with those ideas that make it into human speech. As we have already outlined, there are many important issues related to cis and trans women, such as sexual assault, about which people are almost completely silent.

Or perhaps we need to start looking at privilege as an ethical and legal liability and start quantifying it. Anti-racist feminists have long opposed quantifying privilege at the scale of the individual body (which can lead to something Roxane Gay calls "the oppression olympics" - competition for who is most oppressed). However, building off of recent calls for monitoring Big Tech with things like Sasha Costanza-Chock's "Intersectional Media Equity Index," one could fairly easily quantify the collective privilege of an organization and then create a prediction score for just how likely that institution is to create racist, sexist data products that lead to harmful impacts for users as well as legal and public relations disasters for the firm. Such a score could incorporate demographic information for firm ownership, leadership, employees (with a special focus on the demographics of those who are producers of data products for the company) and users. It could consist of a grade from 0-100, where 0 signifies perfect alignment between the firm and its users and 100 signifies a high risk of discrimination because of misalignment between the firm and its users. This privilege hazard score would measure just how much or how little the firm was influenced by those who already have the most privilege and power, and conversely, just how likely it would be to produce discriminatory "mistakes" and oversights. Consequently, the media might be less surprised when Google, whose board consists of 82% white men, creates image classification algorithms that only show white men in image searches for "CEO". Or when the Mexican State, comprised of X% rich men, is complicit in the murders of its working class women and girls. Such discriminatory outputs would have been entirely expected based on their

privilege hazard score. As discussed in What Gets Counted Counts, there is an explicit politics of being counted here. Quantification can operate as a kind of sousveillance - "watching from below" – where the Great Quantifiers like Google and Amazon and even whole nation-states are quantified and predicted right back.

But let us return to the Frederick Douglass quote, "Power concedes nothing without demand." So far, we have discussed the feminist project of *examining power* – interrogating how power works through data to prioritize some bodies over other bodies and to secure the wealth and status of the dominant group. Buolamwini's algorithmic auditing quantified exactly how much facial recognition software was failing women of color. Princesa's map exposes the fact that gender-based killings are rampant and going untracked by the powers that be. Many important efforts to redress data discrimination and algorithmic bias are working in this mode of examining power. But in order to truly formulate demands, a feminist approach additionally requires *imagining and modeling power differently to achieve equity.* Equity is justice of a specific flavor, and it is slightly different than equality. Fighting for a world which treats everyone equally means that those who start out with more privilege will get further, achieve more and stay on top. Fighting for a world which treats everyone equitably means taking into account present power differentials and distributing resources accordingly. More simply said, equality upholds patriarchy and white supremacy. Equity dismantles them.

So how might data be used not only to examine power but also to transform gendered power relations? To support self-determination of marginalized groups? What does a society that values data and equity look like and feel like?

Let's circle back to present day Detroit. At the end of 2017, the Detroit Digital Justice Coalition and the Detroit Community Technology Project published a collaborative report entitled *Recommendations for Equitable Open Data.* It was the result of two years of research, conversations and explorations about the city's Open Data Portal. The report is specific about what equitable open data is and who it benefits: "[W]e mean accountable, ethical uses of public information for social good that actively resist the criminalization and surveillance of low income communities, people of color and other targeted communities." Note here how the authors named and made explicit whose perspectives they were centering and why – these communities have been historical targets of discriminatory institutional practices. We saw this targeting explicitly in the redlining map introduced earlier in the chapter. The report goes on to outline seven recommendations for the City of Detroit to adopt to make their open data practices more equitable and more likely to benefit people of color and low-income communities. These include "Protect the people represented by the numbers","Engage residents offline about open data" and "Prioritize the release of new datasets based on community interest." These are concrete demands, offered to improve the use and benefits of open data for the people who are most often left out of open data conversations.

So, following Collins, there is a matrix of domination with four different domains

of power. Examining that power using data-driven methods is an important step towards challenging that matrix, particularly in egregious cases like femicides where there is a violent, unjust status quo. Additionally, we have a responsibility to create space for women, people of color, queer and trans folks and others to imagine and dream power differently – to model better and beautiful futures where all can thrive – something which we will address further in *Teach Data Like an Intersectional Feminist!* But it's hard to see the contours of the matrix of domination, let alone empower others or imagine things differently, when you are the recipient of a lot of benefits from it. When the system works for you, you are able to set racism and sexism and other oppressive forces aside and you will experience little penalty for such ignorance.[10]

So what is to be done when you are in a position of power and privilege? Most people working in data science, visualization, machine learning and statistics have significant privilege and power accumulated through their education and their institutional connections, as well their race, gender and ability. Can you use your power and privilege for "good", even though we have explored how much of a hazard it is for your ability to accurately apprehend the injustice of the world? Emphatically, unequivocally "Yes!", with some caveats and elaborations.

The feminist grounding for navigating this quandary is called an "ethics of care", which we introduced in *Show Your Work*. While there are many contemporary discussions about data ethics, most derive from a version of moral reasoning introduced by Immanuel Kant in the 18th century which prioritizes abstract dilemmas, rules and obligations, and universal application. In these conceptions, the focus is on an individual, independent human actor and their relationships with others are conceived as contractual, business-like negotiations among equals. It is important to note that Kant based morality on reasoning, believed women to be incapable of reason, and thus concluded that women could never be full moral persons, i.e. were not fully human.[11] This relates back to the "master narrative" we described in *On Rational, Scientific, Objective Viewpoints from Mythical, Impossible, Imaginary Standpoints*, which valorizes reason and (supposed) impartiality over all other ways of knowing and asserts the superiority of males in that capacity. More recently, technical folk are digging this approach because this kind of blanket ethical logic is easy to code into large systems. [But it's important to note that this approach was explicitly designed to exclude half of humanity.

On the other hand, a feminist ethics of care prioritizes responsibilities, issues in

[10] If you want a viscerally enlightening reading about privilege, check out White Privilege: Unpacking the Invisible Knapsack by Peggy McIntosh from 1989. Written in the first-person perspective of a white person in the US, it lists fifty ways that white privilege manifests in everyday life, including "My culture gives me little fear about ignoring the perspectives and powers of people of other races."

[11] We bring this up so that you might question the wisdom of uncritically pulling 18th century philosophers into 21st century ethics conversations, i.e. so you can drop some knowledge on the next person who sings the praises of the categorical imperative in a machine learning ethics discussion.

context, and, above all else, relationships. Feminist philosopher Alison Jaggar has detailed numerous ways that traditional ethics has failed women. Masculine ethical approaches have systematically showed less concern for women's issues as opposed to men's issues, have devalued ethical quandaries in the "private" realm (the realm of housework, family and children), and have valued traditionally masculine traits like independence, autonomy, universality and impartiality, over traditionally feminine traits like interdependence, community, relationships, and responsibility. While the central unit of analysis in Kantian ethics is the individual human, an ethics of care focuses on the relationship between two or more things (possibly human, possibly not), and the ways that they are bound together by that relationship. Which is to be said, rather than valuing impartiality, an ethics of care prioritizes intimacy and honors the deep, emotional, personal investment that comes with being responsible for the well-being of another, whether that is a child or the environment. And vice versa – the ways that your own well-being is tied up in how a child or the environment cares back towards you and nurtures you. This kind of situated ethics is not as easy to encode into large computational systems, but we shouldn't rule it out as impossible until someone has actually tried.

What does a feminist ethics of care mean for those of us who work everyday with data science, journalism or visualization and enjoy some relatively high degree of privilege? First, accept that your privilege and power are not just an asset, but also a liability. They structure what you and your institutions see in the world and also what (and who) you and your institutions disregard about the world. The antidote to your privilege deficiency is to establish meaningful, authentic, on-going relationships across power differentials (whether based on gender, race, class, technical knowledge, ability, etc) – and to listen deeply to those new friends. This sounds simple, but it is hard, both at the individual level and at the institutional level, because it involves a reorganization of priorities and revaluation of the metrics of success.

Relationships in an ethics of care are a two-way street. For this reason, it's also important to reframe "doing good" with data as something more akin to "doing equity" or "doing co-liberation" with data to remove some of its paternalistic overtones. All too often, well meaning "help" is conceived as saving unfortunate victims from their own technological ignorance. In presenting the origin story of the *Detroit Geographic Expedition and Institute*, Gwendolyn Warren reflected on the ignorance of the white male academics her community worked with, "We had this group of geographers, one of whom lived in the neighborhood, who decided that they were going to 'discover us'. They were going to go and explore the 'hood and discover us. And show us how to make change[. . .] There was no way in hell they were going to save us, but they didn't know it."

Whereas an act of data service performed by a technical organization for a community-based group is often framed as charity, an ethics of care would frame it as one step in deeper relationship building and broader demographic healing. There is a famous saying from aboriginal activists that goes like this,

> If you have come here to help me, you are wasting your time. But if you have come because your liberation is bound up with mine, then let us work together.

Following a logic of co-liberation leads to different metrics of success. The success of a single project would not only rest on whether the database was organized according to spec or whether the algorithm was able to classify things properly, but also on how much trust was built between institutions and communities, how effectively those with power and resources shared their power and resources, how much learning happened in both directions, how much the people and organizations were transformed in the process, and how much inspiration for future work, together, was co-conspired.

Likewise, data projects undertaken by technical folk with an ethics of care would openly acknowledge and account for power differentials by explicitly prioritizing whose voices matter most in the process as input. We saw this in the Detroit Equitable Open Data Report – the authors prioritized the needs of communities that are targeted for surveillance – those who stand to experience the least benefits and the most harm from open data. By prioritizing the needs of those at the margins, we create a system that works for everyone. In some situations, this means working your absolute hardest to establish authentic relationships that did not previously exist. For example, for the past five years, Catherine has been co-leading a feminist hackathon project called *Make the Breast Pump Not Suck*. The first version of the hackathon took place in 2014 and focused primarily on the product design and experience of using [a breast pump.[12]

But after a couple years, it was clear that the innovations emerging in the breast pump space were primarily for white knowledge workers – the smart pumps were coming in at $400, $500 and $1000, not covered by insurance and thus only accessible to those with disposable income. So, in organizing the second *Make the Breast Pump Not Suck Hackathon* in 2018, our leadership team decided to center the voices of mothers of color, low wage workers and queer parents because those are the groups that face the most barriers to breastfeeding in the US context. We invited members of those groups as hackers – and we also put into place an Advisory Board composed primarily of high-profile advocates of color that work directly with community organizations. This Board caught multiple oversights of the majority white leadership team, and shifted the project in significant ways. Everyone was paid for their time. In *On Rational, Scientific, Objective Viewpoints from Mythical, Impossibly, Imaginary Standpoints*, we discussed "design from the margins" as an underlying principle of feminist human computer interaction. This additional layer might be characterized as "governance from the margins." It functioned as an accountability mechanism to simultaneously

[12] A breast pump is a device used to extract milk when a breastfeeding mom is separated from her baby or cannot/does not want to nurse them at the breast. Despite the fact that the medical establishment sees breastfeeding as a public health issue, it is socially stigmatized and has faced neglect as a space of innovation. Lack of paid family leave policy in the US, means that nursing mothers (and trans dads) often end up back at work secretly pumping in closets, bathrooms and cars, if they are able to pump at all.

check the leadership team's privilege and prevent us from doing harm, and also to deepen emerging relationships across race, social capital and technical knowledge.

But for this to work, those that are doing co-liberation with data have to trust that the people who experience the most harms from a social issue have the best ideas for reimagining it. As Kimberly Seals Allers, one of our Advisory Board members said at her keynote, "whatever the question, the answer is in the community." And while the emphasis of data projects is often to develop a time-bounded thing – a database, an algorithm, a model, a visualization – it's important to remember that the longer-term goal is to build meaningful, authentic, on- going relationships across differences in power and privilege in order to transform yourself and your institution and the world.

---

breaks: false intro-text: | Much of current data science education functions as a "Man Factory", focused on reproducing data work that is abstract, individual, & led by elite men. But what if we imagined teaching data as a place to start creating the connected, collective, caring world that we want to see? —

Chapter 8

Teach Data Like an Intersectional Feminist!

If you were on the city streets of Brooklyn or the Bronx in the past five years, you may have inadvertently crossed paths with a data science class. You probably didn't realize it because the classes looked nothing like a traditional classroom. Teenagers from the neighborhood wandered around in small groups. They were outfitted with tablets, pens, paper, cameras and maps. They periodically took pictures on the street, entered bodegas, chatted with passersby in Spanish or English, and entered information on their tablets.

These young people were attending their regularly scheduled school classes as well as participating in a project called "Local Lotto" created by math educator Laurie Rubel, the Civic Data Design Lab at MIT, and the Center for Urban Pedagogy in New York City. Local Lotto was designed to teach place-based statistics and data analysis to high school seniors and community college students. The learning goals for this curriculum are tied to standard mathematical concepts.[1] For example, teachers want to introduce ratios and probability as well as expose students to more advanced concepts like spatial data analysis, combinatorics and mathematical modeling.

The regular way this is done is to line students up in chairs facing a teacher, and talk directly about the mathematical and numerical concepts, and then give students some problem sets. For example, Harvard's Introduction to Data Science classroom, pictured below, looks a lot more like the data science class we might all be expecting. It takes place in a standard large lecture hall with auditorium seating, students listen to lectures, and then they turn in problem sets individually in Jupyter Notebooks.

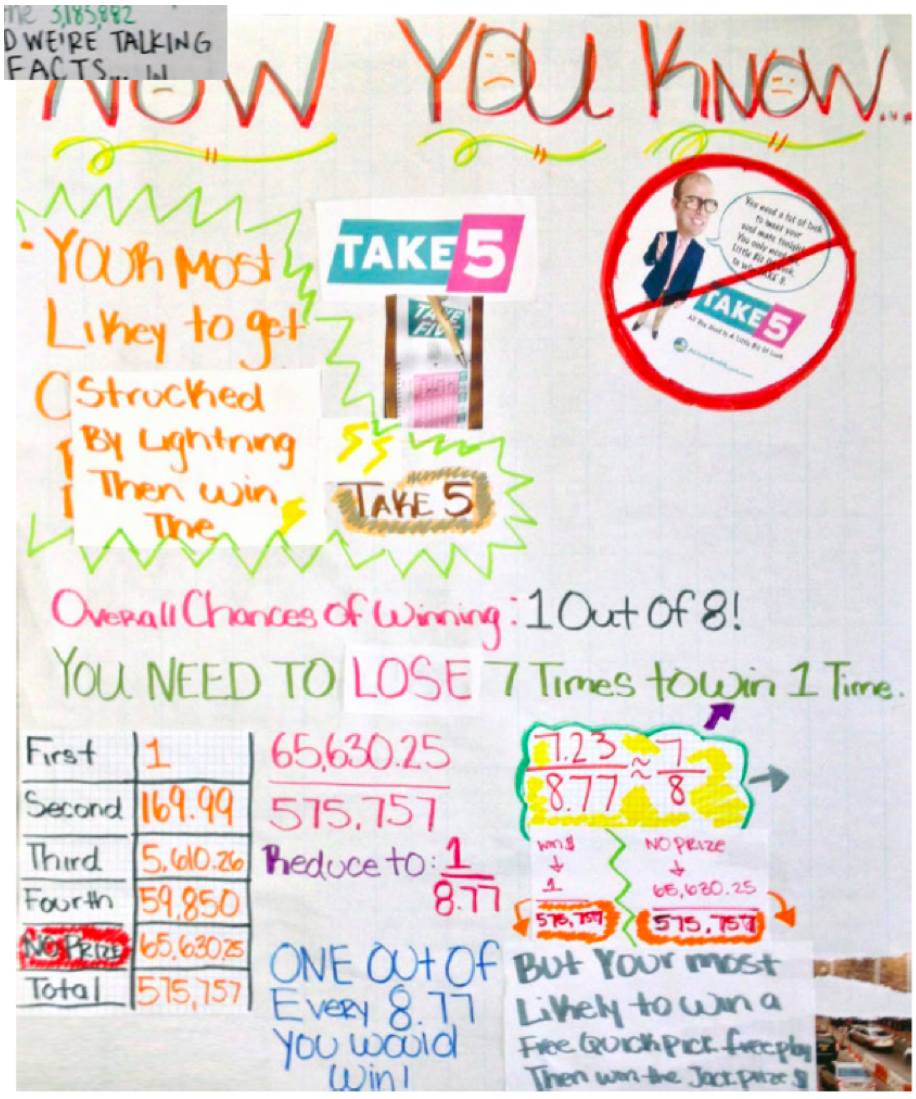
[1]The concepts taught address specific mathematical content and skills outlined by the Common Core State Standards in New York.

But Local Lotto had different aspirations – Laurie Rubel, the math educator who helped design it, is a leader in something called "mathematics for spatial justice". She and the other organizers wanted students to learn probability in relationship to issues in their everyday lives, in their everyday neighborhoods, in dialogue with people in their community. Specifically, they wanted young learners to come up with a data-driven argument to the question: "Is the lottery good or bad for your neighborhood?"

In New York State, like other US states that operate lotteries, the proceeds from sales of lottery tickets are directed towards state revenues. Lottery ticket purchases correlate with socioeconomic status – low-wage workers and people of color buy more tickets than their highly educated, white counterparts. Ticket sales are not allocated geographically to the districts where sales were made. Because of this, scholars have argued for decades that the lottery system is a form of unfair, regressive taxation – essentially a "poverty tax" – whereby low-income, Black and Latinx neighborhoods are taxed more because they play more, but do not receive a proportionate share of the public benefits that follow.

The Local Lotto curriculum is organized around 10-15 standard class sessions. Learners start talking about the lottery and probability by playing chance-based games and creating large, colored charts of ratios. Then they really dive into probabilities by computing and considering the jackpot games, like Sweet Millions. The State of New York runs an advertising campaign for these games with the slogan, "Hey, you never know…" And they advertise Sweet Millions as "your

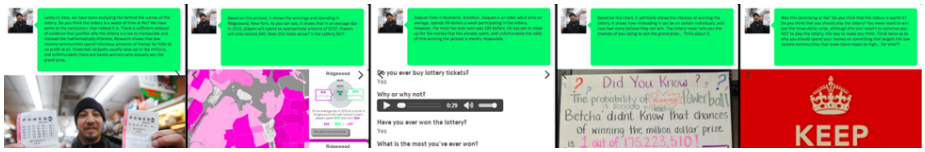
best chance from the New York Lottery to win a million for just a buck." The best chance turns out to be about 1 in 4 million, and a whole class session is devoted to learners coming up with creative interpretations of 4 million that relate to their lives. For example, you would have to drink 15,444 Arizona ice teas to consume 4 million calories; or wait 61 years for 4 million M subway trains to pass you on the platform. After these sessions, learners create infographic posters that respond to the lottery's "you never know" slogan.



At this point in the module, students leave the classroom with the goal of collecting data about how people in their neighborhood experience the lottery, which is how you might have encountered them on the streets of New York. They

map stores that sell lottery tickets. They record interviews with shopkeepers and ticket-buyers on their tablets and geolocate them on their maps. They take pictures of lottery advertising. Afterwards, students present their results to the classroom – including themes and quotes that emerge from different perspectives such as consumers, people who choose not to buy lottery tickets, and shopkeepers.

With this introduction to participatory mapping and qualitative data analysis, the module now turns entirely to mapping the impacts of the lottery spatially. They examine choropleth maps of median income in New York City, they make ratio tables, and they examine how the State's profits from the lottery are distributed spatially to see whether they correlate with spending or with median family income (there is no correlation). Finally, learners prepare their culminating data-driven argument – an opinion story with evidence from their mathematical analysis, their interviews and their spatial analysis. Is the lottery good for their neighborhood?



The Local Lotto approach worked – in the sense that it led to greater understanding of the mathematical concepts with students who had previously been struggling in math. For example, before one iteration of Local Lotto, only 2 of 47 learners were able to use mathematics to find the correct number of possible combinations in a state lottery game where they choose six numbers from a set of forty. Later, almost half (21 of 47) were successfully able to calculate the number of combinations. The Local Lotto approach also worked in the sense that it produced deep engagement with the subject matter. One student shared that what he learned was "something new that could help me in my local environment, in my house actually," and that he now tries to convince his mother to spend less money on the lottery by "showing her my math book and all the work." The female native Spanish speakers in the class, who didn't often participate in classroom discussion, became essential translators during the participatory mapping module in order to interview neighbors. And students went on to teach other teachers about the curriculum, both locally and nationally. As Rubel describes, "We brought them to Brooklyn College to present to faculty there. We had a group of New York City school kids, and each kid led a small group of faculty from lots of disciplines, showing them how to read the maps and how to interpret some of the data. That was neat." The students subsequently presented at Math for America and a national conference on math and social justice in San Francisco.

What's different about the Local Lotto approach to teaching data analysis and statistical concepts? And why should people outside of math education be paying very, very close attention to learning experiments like this that disrupt the status

quo of data science education?

Teaching and technology have a couple things in common. To start with, both are associated with utopian social aspirations. Horace Mann stated famously in 1848, "Education, then, beyond all other divides of human origin, is a great equalizer of conditions of men—the balance wheel of the social machinery." Mann is credited as one of the first advocates of universal public education and was quoted as recently as the Obama administration to illustrate its commitment to poverty reduction through public education. Likewise, information technology has been the subject of utopian speculation. If only everyone had access to the new tools of information, the thinking goes, then perhaps our social ills would be fixed. We can see this in the discourse around the digital divide, or the Open Data Movement, or in specific examples like the One Laptop per Child program, where it was imagined that one way to alleviate poverty and inequality was through giving every child in the Global South a laptop.

The democratic aspirations around both teaching and technology are fascinating to explore in historical relief. While it is hard with those of us with privilege to see our own blind spots right now, it is from a historical distance that we can begin to perceive the ways in which people's attempts to "do good" with the innovations of their day ended up reproducing a status quo in which the people in power are still on top. We know in retrospect, for example, that Mann was serious about the "man" part of his quote. That is to say, education was to be an equalizer of the conditions of men, but not women. In public speeches, Mann argued that men and women could not be treated as equals in the education system because their anatomy is different, "... there is not one single organ in structure, position and function alike in man and woman, and therefore there can be no equality between the sexes." So, the radical part of Mann's social imaginary for the time was that he imagined an education system which treated all white, Anglo-Saxon, Christian men, regardless of class background, as worthy of education. But women, people of color, immigrants, disabled people and others remained excluded from the equalizing.

Likewise, consider the One Laptop per Child project (OLPC), conceived by Nicholas Negroponte, an elite white man based at MIT. The bold idea is to transform global education by distributing low-cost laptops to kids. At the Techonomy conference in 2010, Negroponte responded to some of his critics, "One the things people told me about technology, particularly about laptops, in the beginning was 'Nicholas, you can't give a kid a laptop that's connected and walk away.' Well, you know what, you can. You actually can. And we have found that kids in the remotest parts of the world, when given that connected laptop, like some of the kids in these pictures, not only teach themselves how to read and write, but most importantly, and this we found in Peru first, they teach their parents how to read and write." Unfortunately, evidence from deployments of OLPC in various countries has not borne out the theory of change that laptops lead to literacy. But what is crystal clear in OLPC is the way in which "doing good" reinscribes existing power relationships. The white men at MIT are cast

in the part of the magnanimous benefactors and the children in Africa, Asia, and Latin America in the role of victims that need their help (in the form of cheap computers).

So, teaching and technology have utopian social aspirations, but those are inextricably linked to the people who are doing the imagining, including their gender, class, colonial status and race position in society. Note how neither Mann nor Negroponte envisioned a mutual exchange – they did not imagine that those they were helping might have something to teach them. The utopian imagining done by those in power almost always stops short of mutual transformation or sharing power. Why?

The other thing that teaching and technology have in common is that they are both what feminist computer scientist Lynette Kvasny would call "sites of social reproduction." Feminist theories of social reproduction demonstrate how supposedly neutral places and practices are actually ways to maintain and secure an existing, unequal social order, based in patriarchy, racism and other forms of exclusion. When we frame marginalized people as in need of help from the dominant group, she writes, then "[h]umanity is stolen from historically disadvantaged people as they come to be seen as have-nots, the unemployed, and the urban poor." Although the goal of "doing good" or "helping" or "democratizing" might be well-intentioned, that first act of imagining others in a deficit position in relation to yourself repeats the cycle of domination and oppression. Access to education and technology becomes a way to socialize "those people" into a given social order, without challenging the very basis of the order's existence.

What does this mean practically in the data science classroom? Imagine teaching as a way to model the world. The world that is modeled in the Harvard data science classroom is threefold: Elite men lead. Female faculty comprise less than a third of computer science and statistics faculty.[2] In all the publicly available syllabi for CS109, no female faculty has ever been the lead professor for the course. Second, data science is abstract and technical. Steps like cleaning and wrangling data are depicted as solely technical conundrums and there is little to no discussion of the social context, ethics, values, or politics of data. Third, the goal of learning data science is modeled as individual mastery of technical concepts and skills. The teachers impart the technical knowledge via lecture and students complete assignments and quizzes individually.

Beginning courses such as CS109 play a key role in introducing learners to the concerns of a field (and by extension, what the field is not concerned with). Becoming socialized into the CS109 model of the world means that one sets aside any concern with the social and political, with justice and fairness, with values and motivations. As such, it is no wonder that people who have been socialized into this world order are terrible at creating inclusive visions of how

[2]CS109 at Harvard is taught jointly by Computer Science and Statistics. As of this writing, there are 37 male faculty (69%) and 17 female faculty (31%).

technology might actually be used for collective benefit. They have never been taught a vocabulary for seeing and addressing how gaps in power, privilege and justice relate to technical and informatic fields. This model of teaching might be called "The Horace Man-Factory Model of Data Science" or "Let A Thousand One Laptops Per Child Bloom."

While we have been picking on Harvard, the issue is really not this single class at this single institution. Rather it is that the Man-Factory model of teaching data science is the dominant model across the hundreds of courses now running in high schools, institutions of higher education and on MOOCs like Coursera and Udacity.[3] Data science courses turn a healthy profit, so institutions have been racing to set up online Masters' and certificate programs in order to capitalize on the demand. What gets lost in the process, and in academia's prioritization of research over teaching, are larger questions about power, ethics and values and how those are modeled in teaching practices.

Paulo Freire, the Brazilian educator who wrote the now-classic *Pedagogy of the Oppressed* in 1969, stated it this way, "Education either functions as an instrument which is used to facilitate integration of the younger generation into the logic of the present system and bring about conformity or it becomes the practice of freedom, the means by which men and women deal critically and creatively with reality and discover how to participate in the transformation of their world." Freire called status quo education "the banking model" in which teachers deposit knowledge into students heads, which are conceived of as empty vessels. In contrast, the feminist pedagogy of bell hooks draws from Freire to assert that if learning is to be a practice of freedom then it must be a two-way street – a process of mutual transformation. Indeed, there is much that those in power need to unlearn if they seek to challenge the status quo. bell hooks commented that, "In my books I try to show how much my work is influenced by what students say in the classroom, what they do, what they express to me... This is one of the primary differences between education as a practice of freedom and the conservative banking system which encourages professors to believe deep clown in the core of their being that they have nothing to learn from their students."

So, while there is an emerging focus on data ethics and accountability in research, we need to turn that same focus now to teaching. Teaching is particularly high stakes as a site of social reproduction right now because of the sheer number of people who are working their way through these newly spawned courses and programs in data science. Additionally, due to the rapid pace of change in database technologies, machine learning libraries, and visualization packages, even the most accomplished professionals have to be teaching and learning at all times.

[3]This does not mean there are no data ethics courses, only that it is not the norm to address these concerns in introductory coursework. Indeed, there is a long list compiled by Dr. Casey Fiesler of technical courses that specifically address ethics and what is being called "fairness, accountability and transparency" in technical fields: http://bit.ly/tech-ethics-syllabi

For the remainder of this chapter, we take you on a tour of what it looks like to teach data science with an intersectional feminist lens. Walking back through the prior chapters in this book, we outline how the feminist principles that we discussed – like valuing multiple voices or resisting binary thinking or embracing emotion – apply to teaching. Luckily, none of us have to imagine new things from scratch because projects like Local Lotto are already disrupting the status quo and modeling more emancipatory alternatives to the Man-Factory.

How might we teach a data science that is grounded in values of equity and co-liberation?

In *The Power Chapter*, we detailed how data and its products (like maps and algorithms), can be used to secure power or to contest power. Our examples showed how it matters deeply who is doing the mapping – Gwendolyn Warren mapped the deaths of Black children by white commuters so her community could demand justice whereas the Federal Home Owners Loan Corporation mapped Black and brown residents so that they could systematically deny them bank loans. In general, those who wield their data from a position of power tend to use that technology to preserve a status quo in which they are on top. This is true even when the people in power think of themselves as being anti-racist and anti-sexist because "privilege is blind to those who have it." We named this as a "privilege hazard" and argued for it being a key consideration in the data ethics toolbox. Perhaps unsurprisingly, the people with the most privilege are also the people who argue most stridently against embedding specifically named and detailed values in technology. For example, according to Safiya Noble it took Google until 2013 to start suppressing derogatory and pornographic characterizations of Black women in autocomplete search suggestions. Here we might remind you again that their board is 82% white men.

How might we teach a data science that is grounded in values of equity and co- liberation? As data ethicist Anna Lauren Hoffman has written, "Most important, engineers and data scientists need to listen to and engage with affected communities." Listening and engaging is the first step towards co-liberation. And the only way to work respectfully with those most affected by a problem is to develop a sophisticated understanding of structural oppression and how your own identity factors into that. While computer science offers no help in navigating these waters, emerging technical design frameworks do. For example, the equityXdesign framework that we discussed in *The Numbers Don't Speak for Themselves*, retools IDEO's human-centered design process with an explicit focus on oppression and deliberately centers equity as core value. Importantly, the framework was developed by three African American women who have a powerful vision for ending racism: "Racism and inequity are products of design. They can be redesigned."

Coming out of the field of interaction design, Jill Dimond & Thomas Smyth have

developed a concept called "anti-oppressive design", which focuses on creating interactive systems that strive to end one of more forms of oppression. The framework consists of a series of questions and practices that teams can use in order to prioritize which work to take on, make governance decisions for the organization, and assess whether the work that was done met their core values around ending oppression. Dimond herself used the anti- oppressive design framework to create *Hollaback*, a platform and movement for change that supports survivors of gender-based sexual harassment on the street. At a larger scale, Una Lee, Wesley Taylor, Victoria Barnett, Ebony Dumas, Carlos (L05) Garcia, and Sasha Costanza-Chock are organizing a movement for "design justice" that advocates for design with ten explicitly stated values. The first of these is: "We use design to sustain, heal, and empower our communities, as well as to seek liberation from exploitative and oppressive systems."

How might these emerging frameworks be made manifest in teaching about data science? The simplest answer is to start making contact and building relationships with the people most affected. Let's look back to Local Lotto, where the high school students were trying to answer a question with data: Is the lottery good or bad for your neighborhood? The group could have remained in the classroom, learning about probability concepts and working with data sets about family income and lottery winners, and made their estimation. But the organizers made an explicit choice to value the voices and experiences of neighborhood residents, as data. These interactions sparked new potential lines of inquiry for both the learners and the teachers in the project. As Laurie Rubel describes it, "by talking to people on the street, multiple groups discovered that some people traveled to other parts of the city to buy their lottery tickets because they felt like more winners are going to happen in these other spots. Like, they go to the Upper East Side to buy their tickets. That's an interesting idea. Are there more winners on the Upper East Side?" This was not originally a question that either students or teachers would have asked.

The choice to venture outside the classroom reveals two important things about data science that are not typically emphasized in standard curricula: 1) Quantitative data requires context - The data may appear to say one thing, but what do the people say? What are their attitudes and beliefs about the lottery? What new lines of inquiry might be discovered by young people in the neighborhood talking to older people? And 2) Building social cohesion - Kubel and her associates repeatedly describe the importance of cultivating sociopolitical consciousness in the learners. We can think about the relationships initiated in the interviewing process as contributing to the social fabric of the neighborhood – knitting residents together through conversations about shared interests. These relationships may be deepened and cultivated for later mobilization, say in the form of a concrete demand or community action.

There have also been experiments in situated, partnered ethics in higher education classrooms. For example, in 2014, Sasha Costanza-Chock ran a semester of their Co- Design Studio course at MIT about the topic of data and surveillance.

In this course, MIT undergrad and grad students work in partnership with an outside organization to co-design technologies around a critical issue. This is MIT, so Costanza-Chock could have easily partnered with the Department of Defense or Google or AT&T for the course. But instead of brand name partners, "we were interested in how to amplify the already existing work against the harmful impacts of surveillance, so we partnered with organizations that are leading that fight based on the lived experience of their communities." Costanza-Chock prioritized working with organizations led by people from the community. Partners included the Detention Watch Network, which works to challenge injustices in the challenge the U.S. immigration detention and deportation system, and Transition House, a domestic violence prevention organization that was interested in developing technologies to protect survivors from interpersonal surveillance from their partners. Students initiated relationships with and learned about the perspectives of groups who they otherwise never would have come into contact with. They built technologies like *SpideyApp* – an Android-based Stingray[4] detector – and graphics and media like the Surveillance *Self-Defense Guide* as products of these new relationships.

Racial justice educator Chris A. Miller emphasizes the importance of contact as the first step in fighting oppression. What this means is that relationship-building between people of different backgrounds is the first step in any quest to use technology for co-liberation, whether that is across differences in race, gender, technology literacy, age, profession, ability status or other. If we take that to its logical conclusion, then there can be no "data for good" and no "ethical AI" without contact, relationship-building and trust-building between systems designers and the people with the least power in the system.

How might we teach a data science that names and values the labor of all those involved?

In *Show Your Work*, we explored the many forms of labor involved in data work, from the individuals and groups that serve as the source of the data, to those who collect and process it, to those who analyze it and put it on display. We explained why naming all of these forms of labor, especially those we cannot see, is a feminist act. We also outlined how the current landscape of work, with its work from any place, at any time attitude, leads itself to additional forms of *immaterial labor*. These perpetuate the exploitation and inequality that we already see in other aspects of twenty-first century global life. Identifying these myriad forms of invisible labor is the first step in acknowledging how the visualizations and other data analyses that we see rely upon the work of many hands.

[4]Stingrays are devices that mimic cell phone towers and trick cell phones nearby into connecting with them so that they can gather personal data. They are used primarily by law enforcement and their use is contested by the ACLU and other organizations concerned with privacy and civil liberties.

Thinking about invisible labor can also help point to the forms of work that are harder to quantify, and therefore visualize, because they involve emotional outlays rather than physical or financial ones. The work of caring for a sick relative, as documented in the Atlas of Caregiving, for example, involves not only keeping track of medication schedules and doctor visits, but also helping to bear the burden of the illness, and projecting calm in moments of medical crisis. There is a version of care work involved in data science as well–for instance, the work of Te Whakakaokao, the Ngā Upoko Tukutuku Reo Māori Working Group, which is responsible for designing the library subject headings that enable members of New Zealand's Māori community to locate information about their history and culture in the National Library. Like care-giving, this work is performed for benefit of others, to enable the creation of future knowledge.

How might we teach a data science that names and values the labor of all those involved in the process? And how might we become more attuned to the invisible forms of labor, like care work, that would otherwise be overlooked? We can start by naming all of the people involved in our own projects, and in the courses about data that we teach. Where did we find the datasets that we work with? Is there information on those websites, or in those communities, about who contributed to the dataset? How was the dataset processed, and by whom? Have those people since left the project? If so, can we record their contributions and their names? Similar questions can be asked about the texts and activities that we assign in our classes. By whom were they authored? What topics do they engage? If there are activities that we "borrowed" from other courses, have we acknowledge the teachers whose work we employ?

This approach to crediting intellectual labor derives from feminist practices of citation, but it is not limited to academic contexts alone. Think of it as resistance to what Sarah Ahmed calls *screening techniques*, a concept that describes how "certain bodies take up spaces by screening out the existence of others." When bodies are screened out, they don't appear–let alone have their contributions recognized by others. This is not always intentional, but it is unfortunately self-perpetuating. To borrow another example from Ahmed, it's like sinking into a leather armchair that is comfortable because it's molded to the shape of your body over time. You probably wouldn't notice how the chair would be uncomfortable for those who haven't spent time sitting in it– those with different bodies, and with demands on their time.

Even in classrooms outfitted with the most rigid of plastic seats, we can still work to create a more comfortable intellectual space in the room. We can include more women and people of color among the scholars whose work we assign; and we can include more projects relating to women and people of color among the examples we discuss. Brian Croxall suggests that we should think about how we might "fork" our classes–a metaphor he borrows from version control software–so that we can acknowledge the intellectual labor of the scholars and teachers that our own classes rest upon. We can also acknowledge the intellectual labor (and other forms of work) that our projects rest upon by naming the people

who performed that work, and by working hard, ourselves, to ensure that any invisible labor is better accounted for.

In data science, the names of these people, and the work they perform, are not always easy to locate–if they can be located at all. But taking steps to document all of the people who worked on a particular project at the time that it is taking place can help to ensure that a record of that work remains after the project has been completed. In fact, this is among the four core principles that comprise the Collaborators' Bill of Rights, a document developed by an interdisciplinary team of librarians, staff technologists, scholars, and postdoctoral fellows in 2011, in response to the proliferation of types of positions, at widely divergent ranks, that were being enlisted on scholarly digital projects. More recently, at UCLA, a team of eleven students and faculty members worked together to author the Student Collaborators' Bill of Rights. Supplementing the original document with ten additional principles, the student version emphasizes the importance of empowering students to "make critical decisions about the intellectual design of a project or a portion of a project," and credit them accordingly.

In distinguishing between the intellectual opportunities offered by collaborative digital projects, and the mechanical work that is also required, the Student Collaborator's Bill of Rights draws attention to the importance of affording students the space to grow as project leads. But at Northeastern University, in Boston, literature professor Elizabeth Maddock Dillon takes another approach. She includes assignments involving complex data processing tasks in her courses, so that students can also appreciate the intellectual labor involved in tasks that would otherwise seem purely mechanical. For instance, in her Literature and Digital Diversity course, a text encoding exercise– a required part of the process for converting unstructured text into structured data– presents students with the lines from Shakespeare's The Tempest spoken by Caliban, the only native inhabitant of the island where the play takes place, and asks students to think about how they might make note of any colonialist language in their markup scheme.

The Colored Conventions Project(CCP), directed by P. Gabrielle Foreman, at the University of Delaware, and run by a large team of students, staff, and faculty, seeks to address issues of labor at the level of data entry and content creation, as well as in the actual dataset that the CCP is working to create: a corpus of meeting minutes from the nineteenth-century Colored Conventions, events in which Black Americans, fugitive and free, gathered to strategize about how to achieve legal, social, economic, and educational justice. Justifiably wary of the free labor of crowd-sourcing, the CCP asks its teaching partners to sign a Memo of Understanding (MoU) before contributing to the project. The MoU makes explicit the importance of keeping track of and crediting any student contributions to the project. (In fact, there is a second MoU that students complete, which asks them to share their contact information, should they feel comfortable doing so, so that they can be named on the CCP site).

In addition, the MoU asks that teaching partners address issues of labor in the

dataset itself. Because the dataset is derived from the conventions' meeting minutes, which tended to record the official convention participants and the discussions they initiated, it does not sufficiently acknowledge the contributions of the women who were often in attendance, albeit in an unofficial capacity; or those who worked in the boarding-houses where the male delegates stayed during the conventions, enabling their participation; or those who stayed home altogether, taking care of children and housework, ensuring that their husbands and sons could attend. To address this disparity in the dataset, the MoU asks that all instructors introduce a woman involved in the conventions, such as a wife, daughter sister, or fellow church member, alongside every male delegate who is named. (A growing body of information about these women is housed on the CCP website). As the MoU explains, "This is our shared commitment to recovering a convention movement that includes women's activism and presence—even though it's largely written out of the minutes themselves."

The issue of invisible labor in data science is significant, but it can seem difficult to address only because it has gone unacknowledged for so long. Taking simple steps to keep track of the participants who contribute to a project, and crediting them in the end result, can contribute to a visible record of the work that data analyses and visualizations rest upon. That information is, after all, data. And data, as we know, can be a powerful tool for combating the inequities we encounter in the world–and in our own workplaces, labs and classrooms as well.

How might we teach a data science that honors context?

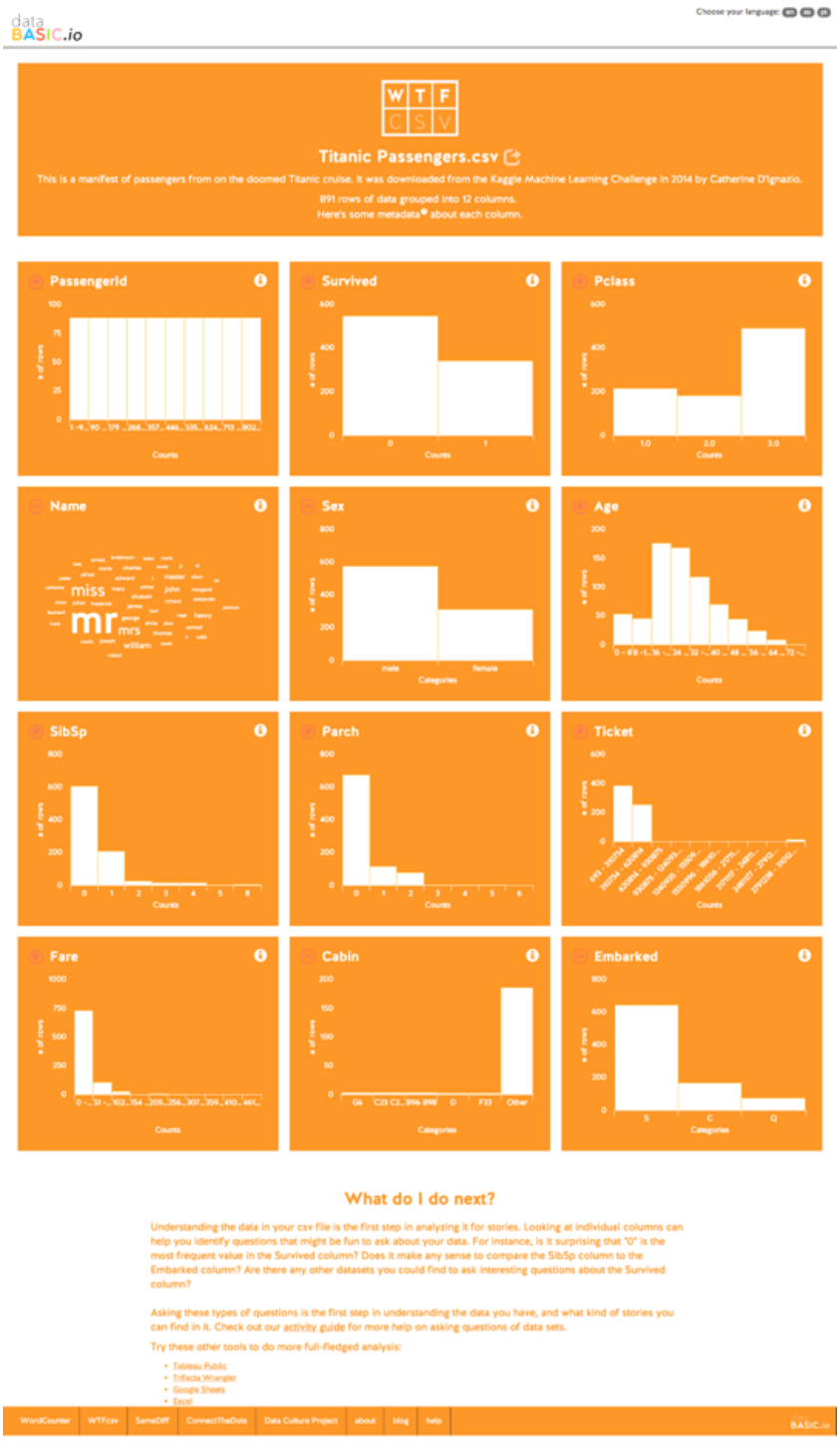
In *The Numbers Don't Speak for Themselves*, we outlined how data do not always represent what they appear to, particularly when it comes to data about about women and marginalized groups. Following Donna Haraway, knowledge is never absolute, but always situated in a social, cultural, historical and material context. Untangling and investigating how it is that datasets are products of those contexts can help us understand the ways in which power and privilege may be obscuring the truth. The collection environment – *or data setting*, as Yanni Loukissas suggests we call it – may have power imbalances, marketing hype, social stigma or incentives at cross purposes that complicate how and whether data are complete and representative. Sexual assault information on college campuses, for example, is self-reported by higher ed institutions whose bottom line is directly threatened by reporting high rates. Likewise, data published online in spreadsheets and APIs often lack robust metadata, including the reason and purpose they are collected as well as the limitations (ethical, social and technical) of what they can and should be used for. Lacking this context for orientation, strangers in the data set run the risk of getting things entirely wrong or actually doing harm by filling in the missing information with their own biases and assumptions.

How might we teach a data science that honors context? Situating data for learners is a particular challenge, since the conventions of both spreadsheets and precise graphics make them appear objective. If data is never "raw", but always cooked through a process that involves people, institutions, politics and processes, then we must give newcomers tools and vocabulary to examine the cooking process. This has not been the norm in statistics education, which has tended to teach data sets as illustrations of theoretical methods. As Chris Wild and Maxine Pfannkuch write, "The arid, context- free landscape on which so many examples used in statistics teaching are built ensures that large numbers of students never even see, let alone engage in, statistical thinking." And even in data science learning, where there is more value placed on "real-world", "messy" examples, deep interrogation of the business purpose and political factors of a data set is not the norm. A data librarian at a Boston-based research university recently grumbled to Catherine that he has to support hundreds of students in developing a basic understanding of what data are and where they come from. For the first assignment in their introduction to data science class, the professor gives them a spreadsheet of several hundred thousand rows of Boston health inspections and tells them to "find something interesting." At a loss for how to get started, the students show up on the data librarian's doorstep.

Honoring context and situated knowledge would proceed in the exact opposite direction – newcomers would be taught that you cannot look forwards towards new insights from data until you look backwards at the data setting. One example of a learning activity that tries to do this is called "Asking Questions". Learners use WTFcsv (part of the Databasic.io suite of tools, co-developed by Catherine and Rahul Bhargava) to learn how to get started with basic spreadsheet analysis. True to its name, WTFcsv is a simple online tool designed to help people understand WTF is going on with thei .csv[5] file. It takes each column from a spreadsheet and characterizes patterns in the data across that column.

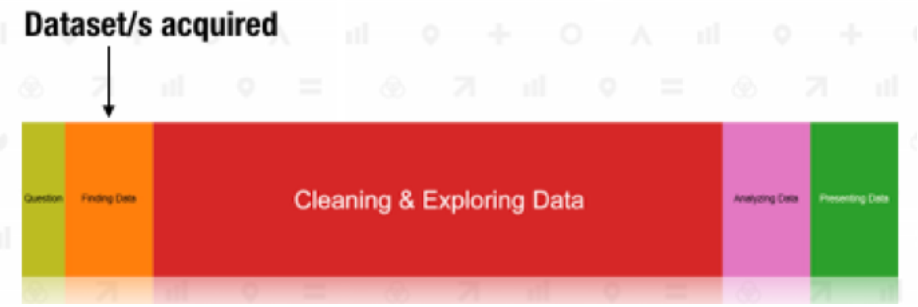
Learners are charged with "asking questions" about the patterns in their spreadsheet data, rather than finding immediate stories or insights from analysis. Above is the WTFcsv results screen showing column summaries from a spreadsheet of passengers on the Titanic. From Databasic.io. [SOURCE: databasic.io and we have permission bc Catherine can grant it]

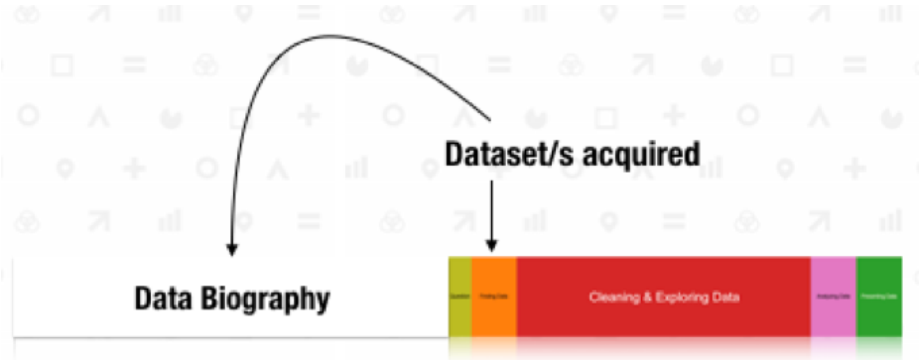
[5]You probably know what WTF stands for. But csv stands for "comma separated values" and is a text-based spreadsheet file format. Each column break is denoted by a comma and each row break is denoted by a carriage return. You can open csv files in spreadsheet programs and most data software packages.


data BASIC.io
Choose your language:
WTF CSV
Titanic Passengers.csv
This is a manifest of passengers from on the doomed Titanic cruise. It was downloaded from the Kaggle Machine Learning Challenge in 2014 by Catherine D'Ignazio.
891 rows of data grouped into 12 columns.
Here's some metadata about each column.
PassengerId
Survived
Pclass
Name
Sex
Age
SibSp
Parch
Ticket
Fare
Cabin
Embarked
# of rows
Counts
Categories
miss
mr
mrs
william
male
female
What do I do next?
Understanding the data in your csv file is the first step in analyzing it for stories. Looking at individual columns can help you identify questions that might be fun to ask about your data. For instance, is it surprising that "0" is the most frequent value in the Survived column? Does it make any sense to compare the SibSp column to the Embarked column? Are there any other datasets you could find to ask interesting questions about the Survived column?
Asking these types of questions is the first step in understanding the data you have, and what kind of stories you can find in it. Check out our activity guide for more help on asking questions of data sets.
Try these other tools to do more full-fledged analysis:
Tableau Public
Trifacta Wrangler
Google Sheets
Excel
WordCounter
WTFcsv
SameDiff
ConnectTheDots
Data Culture Project
about
blog
help
BASIC.io

Rather than being instructed to "find a story" immediately in their data, learners are charged with asking good questions to pursue as lines of inquiry. They break into small groups, choose one of the sample data sets to work with, examine WTFcsv's summary visualizations and brainstorm questions that they want to ask the data. Facilitators encourage learners to use the visualizations to generate many types of questions, including context questions ("What's the source of this data? Why did they collect this information? Who uses it?"), ethical questions ("Is it ok to publish people's full names? How did sex end up as a binary variable?"), quality questions ("Is this data complete? How was it acquired?"), data formatting questions ("What does the 'Parch' column mean?"), as well as data analysis questions ("Did women survive at a higher rate than men?") After 10 minutes of brainstorming questions and sources for related data, learners are asked to select one as the most interesting one to share back to the group. The debrief conversation focuses on how developing rich questions usually means that you need a lot more information than what is contained in the dataset itself – you may need to do background research, consult with domain experts or seek other data sources in order to take the next step with a dataset. Encouraging many types of questions, including questions about trust in the source, missing data, and data formatting, helps learners start to connect the data back to the institutional and historical context where it was collected, emphasizing that those things also matter deeply to any meaning that comes from patterns observed in the data.

Beyond developing learners' muscles to connect datasets back to their context, this activity also models a process that consists of iterative question generation in dialogue with others about factors outside of the dataset. Think about this in contrast to the "find something interesting" assignment where students are individually charged with coming up with creative insights from a large- ish data set with no tools, background information, or other scaffolding. The latter models a process in which data science starts with data, domain knowledge is not necessary, it is imagined that all of the answers are contained within the single dataset itself and it is one individual's job to find them.





Other practices are being developed for learners who are at more advanced stages of their learning process. Miriam Posner requires that learners in her digital humanities class at UCLA interview domain experts about their data set before they do any analysis. The *data biography*, used by Heather Krause, a data scientist and educator, is another emerging context and data verification tool. Prior to beginning the analysis process, Krause asks people, particularly journalists, working with data to write a short history of a particular data set and answer four basic questions: Where did it come from? Who collected it? When? How was it collected? Why was it collected? Krause advocates using data biographies as a first step in understanding the origin story of the data set because, as she says, "you need to treat data with as much care as you would treat any source in any journalism project." In her online tutorial about data biographies, she describes how data about violence against women in Malawi appears to improve dramatically from one year to the next, but in fact these variations are due to the data collection being undertaken by two different organizations employing two different methodologies. Undertaking a data biography can reveal these inconsistencies, places in the data pipeline where disparate data were combined, and power imbalances.

Another emerging practice that attempts to better situate data in context is the development of *data user guides*. 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 data sets 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." Data user guides are simple, written documents that contain a narrative portrait of a data set. 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 data set; and the limitations and ethical implications of the data set. While it is more of a commitment than a data biography and takes up more classroom time, writing a data user guide impresses upon learners just how much background context and complexity there is to uncover about even

the most seemingly simple data. To date, groups in several different learning situations have been assigned to write data user guides with promising results: graduate students in Digital Media at Georgia Tech, undergraduates in a data visualization course at Emerson College, and fifty librarians participating in an online course called "Civic Data Ambassadors."[6]

These are promising experiments and practices, but if we aspire to honoring context in a more systemic way, it is important to pose the question, "Who is good at context?" While most data science education has tended to be situated in departments of Computer Science, nobody would think of computer scientists as having great conceptual, theoretical or practical tools for understanding the social and political environment. It's simply not what the field has been concerned with. Social scientists, humanists, ethnographers, psychologists and designers all have more robust ways of navigating and understanding context. What this points to is modeling a transdisciplinary approach to data science education where it isn't "owned" by a single discipline, but rather taught in studio-form and grounded in the topical subject matter of the data.

How might we teach a data science that is not about individual mastery but about arriving at shared meaning?

Modeling a data collection and analysis process that embraces many voices and perspectives leads us back to the chapter *Unicorns*, *Janitors*, *Ninjas*, *Wizards* and *Rock Stars*. While many metaphors that are used in the popular media for data scientists promote an image of a lone wizard (man) who dominates and tames unruly data to extract "intelligence" (for his corporate employer), there are powerful counter narratives embodied in feminist-led projects like the Anti-Eviction Mapping Project and GoBoston 2030. What the latter get right is that they start with the idea that traditional methods of institutional data collection and analysis are not working for everyone equally. So, they organize a process to center the perspectives of marginalized groups and value knowledge from distinct standpoints. Which is to say that process matters. A lot. And participatory processes are inherently messy and multivocal, hard to tame and dominate like rows and columns. And that's ok.

How might we teach a data science that is not about individual mastery but about arriving at shared meaning? Unfortunately, many data science courses and workshops plant individual learners in front of computer screens and walk them through technical trainings in R, D3, Excel, or Tableau. What this models to learners is a world in which data science is primarily a technical endeavor isolated from social circumstances, and success is defined by one's individual

[6] "Civic data guides" is the name of the collaboration undertaken by Catherine, Yanni Loukissas and Bob Gradeck around the production of data user guides by students and learners.

mastery of tools.[7] This does wonders to perpetuate the newcomer perception of data science as highly exclusionary male wizardry.

What are some alternatives to this learning model? In Detroit, three community technology groups have banded together to produce something called *Data DiscoTechs*. These drop-in, peer-to-peer learning events create a space where people can discover technology together.[8] Each *Data DiscoTech* features a variety of stations including topics like data scraping, data visualization and an introduction to social media data. The Detroit Community Technology Project emphasizes that it's important that stations are staffed by volunteers from the community, "Participants learn at their own pace and from people who understand the context of their neighborhoods and communities." *Data DiscoTechs* connect open government data to issues that the community cares about and that have immediate relevance for people's everyday lives. For example, the Transit Justice station taught people how to make heatmaps that combined transit and survey data in order to analyze bus routes in Detroit in terms of quality and availability, and then propose alternatives.

While there are plenty of computers and software programs in view at a *Data DiscoTech*, there is also a lot of engaged conversation, collaborative problem-solving and laughing going on. There are kids, teenagers, and older adults asking questions and teaching things to each other. As one of the organizers states, "A *Data DiscoTech* not only creates an opportunity to demystify technology and data but it also creates an opportunity to build relationships." It's worth a pause here to reflect on that last bit: an opportunity to build relationships. Most data science educators imagine that they are educating individuals and would likely not put "building relationships" as a top learning goal for their work. But perhaps they should. Because when is a data-driven project wholly conceived, executed and implemented by a single person? Like, never.

Is data literacy a set of individual skills where learners graduate to being technical ninjas or could it be a collective muscle – learned and practiced in context and in community? Could it be both? Learning experiments like *Data DiscoTechs* simultaneously build individual skills and build social cohesion – the relational infrastructure of the community to address its own data challenges. They model to newcomers that you don't have to be a wizard to work with data, and you also don't have to hire an expensive outside wizard who totally doesn't get your needs

[7] This is not to say tools and individual skills are not important (they are), or that your co-authors have never led tool-focused workshops (we have). Rather, the problem when this is the only model of learning that is ever undertaken in a workshop or course.

[8] The concept of a discotech was created by the Detroit Digital Justice Coalition in 2009 based loosely on the idea of a potluck event for technology. The goal of a discotech is to create "a genuine collaborative, collective learning environment that is accessible to all skill levels, ages, and learning styles." The first Data DiscoTech was run in 2015 as a response to a Detroit open data ordinance. The organizers were concerned about some of the harms that might arise from open data. "Public data impacts different people differently," as one organizer stated, so the goals of the discotechs have included capacity building as well as consciousness building. The Coalition has published a free guide to running your own discotechs here: https://www.alliedmedia.org/files/ddjc_zine_4.pdf

and concerns. Rather, you build trust and relationships with guides in your community which can be mobilized when needed for deeper collaborations in the future. Another advantage of prioritizing relationship-building relates back to the feminist concept of design from the margins that we discussed in *Unicorns*, *Janitors*, *Ninjas*, *Wizards* and *Rock Stars*. As Kimberly Seals Allers, women's health advocate, says, "Whatever the question, the answer is in the community." People in a community know its problems, intimately, and they know which phenomena go uncounted, underreported or neglected by institutions in power – such as the deaths of Black children killed by white commuters in Detroit mapped by Gwendolyn Warren. They also know what the harmful impacts of data are for their people. Building trust across power differentials increases the chances that people at the margins could engage authentically and help educate those in power, particularly the ones that aspire to "doing good" with data.

Building the capacity of organizations and communities to work with data is what led Rahul Bhargava to pen a blog post titled "You Don't Need a Data Scientist, You Need a Data Cultur*." In it, he describes the* Data Culture Project*, co-developed with Catherine, to scale data literacy across an organization. Many of the challenges nonprofit and community-based organizations face in making effective use of data have less to do with tech skills and more to do with organizational process and culture.

For example, many nonprofits collect lots of data but don't actually use it. Or data gets silo'ed in the IT department and seen as "the tech people's job". Or departments engage in separate data analysis efforts where they could be creatively pooling efforts. Or staff are alienated by KPIs[9] and dashboards and don't see the relevance for their everyday work.

The answer to these challenges is not that everyone needs more spreadsheet training but rather that there needs to be a more effective participatory process around how data is used in the organization. The *Data Culture Project* addresses this through a free, self-service curriculum with monthly activities. For example, after his organization led its staff in the Sketch a Story activity, Michael Smith from the Telluride Foundation reported that, "One of my colleagues came back after the session to share a 'breakthrough' on how to use the tool to analyze our program content. I also overheard our Communications/Marketing team discussing how they could use it to analyze interview and social media data." In the process of learning a new concept, like quantitative text analysis, staff also build relationships across silos.

There's that *building relationships* thing again. Prioritizing relational infrastructure and multiple voices in the data learning process helps learners implicitly understand feminist standpoint theory, even if they don't address it in those terms. Meaning, there is never one singular possible interpretation of a set of

[9]KPIs are Key Performance Indicators – measures that help to evaluate the performance of an organization in regards to a particular activity that it has deemed important. For example, a nonprofit might track its fundraising efforts with a KPI like "Cost per dollar raised" – how much did they spend on fundraising for every dollar that they brought in.

data (that one discovers by sitting at a computer for long enough) but rather a better or worse collective process informed by data to arrive at shared meaning.

How might we teach a data science that addresses the politics of and the absences in counting and measuring?

In *What Gets Counted Counts*, we discussed how the data we collect, and the categories we place them into, matter deeply for the analyses that can then be performed. Once categories have been established, it can be nearly impossible to go back and look for information that has been left out. Because standard data collection categories, like gender, are often derived from existing social categories, it is absolutely essential that we question those categories and the assumptions that underlie them, before we translate them into the categories we use to collect our data. As the example of the gender binary makes clear, binary distinctions are also often secretly hierarchies, with one category on top and the other on the bottom–to say nothing of the people, like Maria Munir, who are excluded from the classification system altogether.

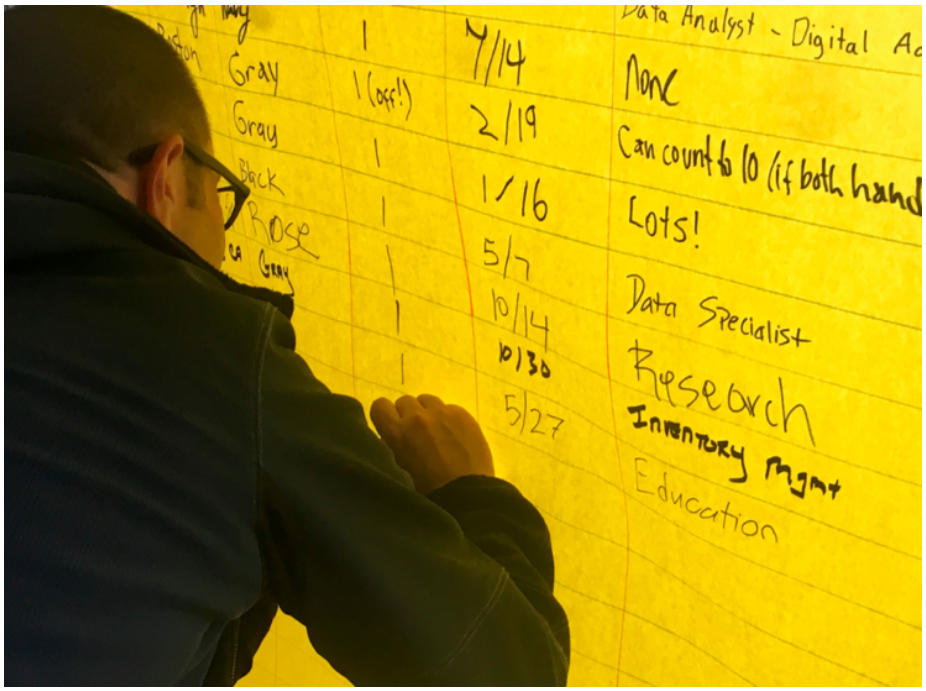
Counting quickly gets complicated, as Munir's story shows, but most data science courses present datasets as if they just dropped from the sky. In these cases, the teaching environment is modeling a world in which the role of the data scientist is of a pure technician. Somebody else asks the research questions, somebody else convinces the institution that the project is worthy, somebody else allocates the resources, somebody else designs the data entry and somebody else does the data collection. Then the data scientist enters the picture. If this is the way it works, then we argue that the "data scientist" shouldn't be elevated with the "scientist" bit at the end, but rather revert back to the less sexy "analyst" who doesn't ask hard questions and dutifully does the computing handed down by management.

What this prevents the teacher and learners from doing is having a productive discussion about what actually gets counted and measured, and what does not. It prevents them from discussing the thorny problems involved in collecting data about hard-to-measure phenomena that matter deeply for the well-being of women and people of color – like sexual harassment, domestic violence, discrimination in the healthcare system, police killings, hate crimes, indigenous land use, food deserts. The list goes on. It prevents the learners from discussing institutional ethics and responsibility – whose job is it to measure maternal mortality? Whose job is it to mobilize that data to do something? It prevents the class from discussing values and consent – when is it unethical to count and measure something? When do you walk away from the institution that hired you?

How might we teach a data science that addresses the politics of and the absences

in counting and measuring? Educators can start by refusing to model "Plop Data Science" where the learner is repeatedly plopped into the middle of the process with a data set, with the implied takeaway that "someone else" will always figure out that hard stuff that comes before the data set is collected. At Georgia Tech, for example, Lauren teaches a data visualization class that relies upon data from the U.S. Census. But before examining the dataset, students learn about the history of the census, its political origins, and its change over time. She explains how, in the census, even just counting the number of people in the country is political; before emancipation, white residents of the slave-holding South wanted as many representatives as possible in Congress, but they didn't want to acknowledge that Black southerners counted in equal measure of themselves. Hence the notorious "three-fifths compromise," which counted each enslaved person as three-fifths of a citizen. This allowed slaveholder interests to dominate the US government until the Civil War.

Another way to show how counting matters is to have the students themselves learn basic methods of counting and measuring and collecting. This does not have to exhaustively cover all possible methods but can focus on tuning learners into the right questions to ask. For example, Catherine and Rahul Bhargava have a learning activity called *Paper Spreadsheet* designed to introduce newcomers to basic ideas of data collection as well as the limitations of what data can and cannot represent about the world. In *Paper Spreadsheet*, learners fill out a row of information about themselves on a large, colored piece of paper, including their name, hometown, number of siblings and color of their shirt. In the ensuing discussion, the facilitator asks questions about what the data does represent about the people in the class and what it does not represent, leading the learners towards questions about representation and ethics. For example, the column "color of your shirt" leads learners to enter in a single color, but often their shirt is multicolored or patterned. The facilitator can point out that data collection is reductive, by definition and by design, but that it's important to be attentive to when more complexity is necessary to answer the questions at hand. What do we miss out on if we only have one column about shirt color? That resolution might be fine for a basic characterization of people in the room, but completely insufficient for a fashion designer to create new shirt designs. Likewise, the facilitator can ask the learners questions about ethics, consent and privacy, such as "When would you refuse to input your data into this spreadsheet? What columns would be invasive to collect, such as your sexuality or mental health status? Does it matter who is doing the measuring?"



Case study discussions can allow learners to go further and unpack some of the ways that structural oppression shows up in counting and measuring. For example, Anna Lauren Hoffmann taught an undergraduate course called *Data & Ethics* at Berkeley for two years that used a case-based approach to introduce learners to ethical issues across the whole pipeline of data processing. Students read and discussed cases based on high-profile examples featured in news reports and op-eds. In one example around data collection, learners read about the launch of Apple's Health app in 2014. The app initially enabled the tracking of health data ranging from calories consumed to heart rate to blood alcohol content. A notable absence, quickly seized upon by many female tech writers , was the lack of ability to track menstrual cycles, something that women have been doing for centuries longer than people have taken chromium supplements. In Hoffmann's class, she used this case as a way to discuss the power of default choices made by designers and engineers, and whose bodies were included and excluded based on those defaults. She also invited students into the design process – "we would use white boards to draw alternative fitness tracking dashboards for different populations." Students imagined what a fitness dashboard for people with scoliosis might look like, for example, or for migrant agricultural laborers. In the case of the latter, the student designers included information about hours of sunlight, alerts for when to take water breaks and connected the laborers to ways to document workplace abuse and take political action.

Beyond discussions of case studies, learners might be introduced to participating in open source communities, or other volunteer groups, whose explicit purpose

is to collect data – like Open Street Maps or New York's Homeless Outreach Population Estimate.

Counting is complicated, but it's easy to forget when you're handed a dataset and told "ready, set, go." A data science that calls attention to the decisions made when counting, and shows how those decisions impact the questions that can then be asked, leads to more truthful or accurate answers with respect to the dataset at hand. In addition, it can help to shine a light on new questions worth asking–questions that students who are trained to think hard about the source of their sources, will be well-equipped to begin to explore.

How might we teach a data science that equally values ethics, emotions and reason?

In the *chapter On Rational, Scientific, Objective Viewpoints from Mythical, Imaginary, Impossible Standpoints*, we contrasted the Periscopic gun deaths visualization with a graphic by the Washington Post about active shooters to initiate a conversation about emotion in data science and visualization. We outlined how contemporary Western thinking about data has evolved from a flawed model of a "master stereotype" where what is perceived as rational and objective is valued more than that which is perceived as emotional and subjective. The master stereotype would say that emotions cloud judgement and distance amplifies objectivity. But a feminist perspective challenges everything about that master stereotype. Emotions don't cloud judgement – they produce deep engagement and incentive to learn. Patricia Hill Collins, for example, describes an ideal knowledge situation as one in which "neither ethics nor emotions are subordinated to reason."

So, how might we teach a data science that equally values ethics, emotions and reason? The question sounds abstract, but it might be as simple as reconsidering the subject matter of the data that you teach with.

Data analysis techniques are often discussed as though the subject matter of the data is interchangeable and neutral. Many teaching examples use so-called "classic" data sets like mtcars – a dataset about different features of cars from a 1974 edition of Motor Trends Magazine. Each row of data contains measures for horsepower, miles per gallon, # of cylinders, and so on. But who cares about measuring car efficiency? Mechanics, salespeople, car companies, and people that like cars – disproportionately dudes. The authors of this book – not so much. Valuing emotion, in this case, would mean ensuring that the choice of teaching data has some cultural and emotional proximity to the teacher and to the learners – i.e. they have some reason to care about the subject matter, they have some ground truth experiential knowledge of the data, and they have some emotional or ethical investment in asking questions of that data. In the case of *Local Lotto*, that proximity is literal and geographic - the learners are from the neighborhood where they are collecting data about lottery usage.

In another example, Tahir Hemphill is an artist who founded the Rap Research Lab (RRL), can after school program for Black, hispanic, immigrant, female, and transgender youth. RRL teaches data analysis using a vast database of hip hop lyrics which Hemphill assembled called the Rap Almanac. RRL learners produce sophisticated research questions, analyses and visualizations, which are informed by their deep knowledge of hip hop culture. Rather than seeing the youth's emotional connection to hip hop lyrics as something that makes them "biased", Hemphill's project uses that intimate experiential knowledge to bridge their investment in learning data analysis techniques. The youth are already experts in the experience and context of hip hop music, so they can mobilize that existing expertise to ask interesting questions of the data. Emotional and ethical proximity to the subject matter is an asset – they are a form of insider knowledge which makes for better and different questions than those asked by strangers in the data set. Indeed, Helen Kennedy and her research team at the University of Sheffield found in 2016 that the subject matter of a visualization, and its relation to the viewer's everyday life, was a key factor in people's engagement with them. People invested more time and energy in looking at visualizations whose subject matter mattered to them.

Valuing emotion and ethics as equal to reason may also mean that educators could expand the output forms that they show as examples and that they structure into assignments. Valuing emotions leads us to ask – why stop at a web visualization or a statistical model as "proof" of learning? How about sculptures, murals, quilts, VR experiences, sonifications, and other "visceralizations" of data? And valuing ethics leads us to ask - what is the most appropriate form of output given an analyst's obligation to communicate results to a given community, most especially the community who is represented in the data? If you think back to the case of the Groundwork Somerville data analysis discussed in *Unicorns, Janitors, Ninjas, Wizards, and Rock Stars*, the output took the form of a data mural. This large-scale analog painting, situated in a working garden, communicated the youths' analysis of food security data in an accessible, novel format and helped the sponsoring organization build awareness and solidarity around its mission in the community.

Valuing emotions and ethics may represent a shift for more technical data science educators, who may not be used to thinking about data as creative communication, nor have the social-relational muscles to think about multiple audiences and stakeholders. In these cases, educators could benefit from partnerships with media, art and design educators, whose fields are built on experimenting with these questions.

Justice is a Journey

We opened this chapter with the example of the *Local Lotto* project. The organizers did many innovative things in this project – they situated data analysis as an issue of cultural relevance for the learners, they built relationships

between learners and neighbors, they valued the voices and experiences of the community, they gave multiple opportunities for creative communication with data – all while teaching important concepts around spatial data analysis and probability. But *Local Lotto* also has its failures and open questions when it comes to achieving its stated ethical goals around spatial justice. In this, it resembles every other project in the entire world, including those undertaken by you, the readers, and us, the authors.

The organizers of Local Lotto wrote a paper for the Harvard Educational Review in 2016 in which they reflected on the successes and failures of the project. While there is much evidence to point to success, including the test scores and learners' engagement with the material, they note several oversights and issues to expand on in the next version of the curriculum. First, and importantly, they noted that the collaborating teachers and course designers on the project are white and Asian whereas the youth in the classes were predominantly Latinx and Black. While the narrative of the course designers focused primarily on income inequality and did not engage race, they write that "the students consistently surfaced race." Since race and ethnicity were not part of the teaching material, the teachers felt that they did not have the experience or background to discuss it explicitly, and would deflect those conversations. The organizers are now taking steps to explicitly integrate discussions about race. They also plan to include race, ethnicity and age data in the next version of the curriculum, because "youth, and in this case youth of color, have different understandings about racial boundaries; theirs are differently nuanced and scaled than affluent, White, or adult perspectives."

Another question the course designers pose to themselves has again to do with the identity of who is in the classroom and who is leading the classroom (primarily white or Asian outsiders to the neighborhood and the community). The organizers describe "limited but recurring instances of resistance from students" to the project's central thrust around investigating income inequality in the lottery. This resistance from youth learners might be summarized like this, "you have no business coming into my world and telling my people that we are doing things wrong and that we should use your tools to do things right." While this was not the course designers' purpose, it is easy to see how a curriculum taught by outsiders, focused on income inequality, could be seen as passing judgement on people in the neighborhood and perpetuating a deficit view of low-income people. As in, "If only they knew what was good for them, they would not buy lottery tickets. We will use our tools of privilege to teach them." In reflecting on these unintended and possibly harmful results, the course designers determined that their steps for the next iteration would include revised maps and visualizations that did not paint such a simple narrative about income inequality as well as connecting students with people in their community who are actively working on issues of income inequality. Which is to say, they will work harder next time to build relationships between the youth and their community.

So, is the *Local Lotto* project a feminist failure? Or an admirable achievement

for equity-focused learning? It is both of these things. What is important is to understand justice as a journey and intersectional feminism as a set of tools to apply along the way. While one might master commands in R or write a badass Python scraper, there is no such mastery of feminism because you are (hopefully!) always in a state of unlearning your own privilege and encountering new social and political differences that challenge your prior worldview and make you extremely uncomfortable. Discomfort, shame and failure are par for the course. Python scrapers might be frustrating but will never make you confront your own sexism. What becomes important in working towards justice is what Donna Haraway calls "staying with the trouble" – having the courage to keep going when the work is difficult and fuzzy and you and your people and your institutions are a major part of the problem. One of the biggest strengths of *Local Lotto* is the courage of the project creators to publicly and reflexively interrogate themselves and their process, to detail their failures as well as their commitments to doing better on the next iteration. Justice is a journey and the most important part of it is that you stay with the trouble (and hopefully cause some trouble, yourself)

Conclusion: Now Let's Multiply

Less than 24 hours after the 2016 US Presidential Election, and the unexpected defeat of Hillary Clinton, a retired grandmother from Hawaii, Teresa Shook, took to Facebook to declare her intention to march on Washington in protest of the president-elect. Less than two days after that, Shook's idea became a reality when a group of professional activists and organizers offered to help her plan an official event–what they decided to call, as Shook had, the "Million Women March."

When Shook picked the name, she was thinking about the Million Man March, the 1995 event which brought a reported 1.5 to 2 million African American men to the National Mall in order to call attention to issues of civil rights and racial injustice. It's unclear as to whether Shook knew that there had already been a protest called the Million Women March, which took place in Philadelphia, in 1997, on behalf of African American women. But in either case, the name of the 2016 event drew critics– who pointed out, rightly, that because the initial group of organizers consisted only of white women, they lacked the crucial perspectives that would allow them to develop a truly inclusive social and political platform.

The critics were right. And after adding additional women of color to the organizing committee– many of whom were professional activists and organizers– and also changing the name of the event, it expanded into a more powerful and inclusive movement, if not ever controversy-free. With more than 600 distinct locations across the US and an estimated three to five million participants, the Women's March on Washington became the largest single protest in U.S. history.

The story of the Women's March serves as a reminder that while feminism can serve as a valuable starting point for identifying issues of inequity and injustice, it's not the only position that a person might start from. As we hope we've made clear, a feminist approach– to data science, to visualization, or to anything else in the world– should *always* be accompanied by an awareness of the perspectives that it does not, or cannot, account for.

We are two white women with four advanced degrees and five kids between

us, who work in the privileged world of higher education. While we have learned immensely from, for example, the Black feminist activists and organizers whose work we describe in this book, we can never speak directly from the life experiences that motivate their work. But imagine a book on data that takes Black feminism as its focus. What concepts would it introduce, what principles would it propose, and what examples would it cite?

Some of these possibilities are hinted at in the mission of the group Data for Black Lives, which we discuss in The Power Chapter. The group's emphasis on liberation, rather than a generic form of social good, leads them to projects that actively work to overturn the discrimination and injustice experienced in Black communities as a result of data-driven systems like predictive policing, predatory lending, and risk-based prison sentencing. Or, for another example, consider the principles that guide the corpus creation work of the Colored Conventions Project, which we discuss in Teach Data Like an Intersectional Feminist. Like the nineteenth century organizing meetings that the project seeks to document, the CCP promotes the work of collectives over individuals, and insists on acknowledging the humanity of any person or group represented in their data set.

Or, as another starting point, consider what a queer approach to data science might entail. Queer data science might build off the concept of failure, as described by Maria Munir with respect to the lack of non-binary gender categories in What Gets Counted Counts. Amplifying the moments in the data processing pipeline when our work leads not to new knowledge, but to something we can't ever know, a queer data science could help to call out the otherwise invisible assumptions embedded in our technical and social systems. This might take the form of visualizing the gaps in a data set–a sort of inverse of Daniel Cardozo Llach's visualization of architectural data traces that we discuss in Show Your Work. Or it could lead to the design of an entire visualization that, rather than employing interaction to lead to increasing insight, instead becomes increasingly opaque over time–a sort of data science version of the "refusal of legibility," to borrow Jack Halberstam's term, that characterizes much of queer life.

Or how would a disability studies perspective, which shifts the focus from the individual body to the social structures that enable the capacities of certain bodies, while disabling the capacities of others, push us to rethink our approach to interface design, and of the larger frames in which we display the results of our data analyses? Or a postcolonial approach, which would challenge us to connect issues of power, politics, and geography, in the context of data? Or an explicitly indigenous approach that values cultural knowledge as sacred and has rigorous accountability structures for working ethically with outsiders? These are only a few of the possibilities that approaches to data science, informed by additional perspectives, might present.

Our goal with this book has been to provide a model of how feminist thinking might be applied to data science, and to plant the seeds for exploring how other modes of thinking that intersect with social and political concerns can help

advance the field. In the examples in this chapter, we sketch out some of these possibilities. This conversation – about data, design, and justice – is one that's only just begun.

About Us

Catherine D'Ignazio is a scholar, artist/designer and hacker mama who focuses on feminist technology, data literacy and civic engagement. She has run breast pump hackathons, designed global news recommendation systems, created talking and tweeting water quality sculptures, and led walking data visualizations to envision the future of sea level rise. Her research at the intersection of technology, design & the humanities has been published in the *Journal of Peer Production*, the *Journal of Community Informatics*, and the proceedings of *Human Factors in Computing Systems (ACM SIGCHI)*. D'Ignazio is an Assistant Professor of Civic Media and Data Visualization in the Journalism Department at Emerson College, a Senior Fellow at the Emerson Engagement Lab and a research affiliate at the MIT Center for Civic Media & MIT Media Lab. Learn more: www.kanarinka.com.

Lauren F. Klein is a scholar and teacher whose work crosses the fields of data visualization, digital humanities, and media history, among others. She has designed platforms for exploring the contents of historical newspapers, recreated forgotten visualization schemes with fabric and addressable LEDs, and, with her students, cooked meals from early American recipes—and then visualized the results. Her writing has appeared in *American Literature, Digital Scholarship in the Humanities,* and *Feminist Media Studies,* among other venues. With Matthew K. Gold, she edits *Debates in the Digital Humanities*, a hybrid print-digital publication stream that explores debates in the field as they emerge. Klein is an Associate Professor in the School of Literature, Media, and Communication at Georgia Tech, where she also directs the Digital Humanities Lab. Learn more: www.lklein.com.

Acknowledgments

We are indebted to many people for getting *Data Feminism* to this point. At the MIT Press, Gita Manaktala has provided crucial editorial guidance. Catherine Ahearn has devoted countless hours to creating the draft that appears on this site. Nhora Lucia Serrano and Kyle Gipson have also provided invaluable editorial assistance.

We are grateful to our our student research assistants, Izii Carter at Emerson College, who compiled the metrics that appear in our values statement, and Zoe Wangstrom at Georgia Tech, who compiled information about the images that appear in this draft.

We would also like to thank David Weinberger, who first solicited the project, and who reviewed multiple drafts of our book proposal; Patsy Baudoin, who has offered feedback on many phases of the project; Alison Booth and Liz Losh, whose endorsement of the project has enabled it to advance; our colleagues and friends who have read and commented on portions of this manuscript; and the institutions which have extended us invitations to talk about our work, and offered their own thoughts and ideas.

We are grateful to our partners, our children, and their grandparents and caregivers, who have given us the time to write.

We would also like to thank the activists, journalists, artists, scholars, and teachers whose work we describe in this book. It remains an inspiration.

Data Feminism is supported by a 2019-2020 ACLS Collaborative Fellowship.

Image: Emily & Rahul Bhargava, youth, staff and volunteers at Groundwork Somerville, who all participated in creating a data mural in their community garden.

Code of Conduct

Data Feminism is a book that aspires to take the challenges of our present moment head on. It deals directly and explicitly with issues of social, political, and economic inequality, including sexism and racism, among many others. In so doing, it also aims to elevate the voices of those with experience closest to those issues, as we describe in our values statement. These voices must be respected, and offensive or harassing comments will not be tolerated. Comments of this nature include sexualized, racialized, and/or otherwise derogatory language, as well as deliberately intimidating or bullying language. These comments will not be tolerated, and will be removed by PubPub administrators. Users contributing comments of this nature will be blocked from future commenting on PubPub sites.

This code of conduct is intended to enhance conversation, not restrict it. To this end, we also ask that you remain attentive to and supportive of the commenters themselves. These include undergraduate students and senior academics, data justice advocates and outside observers, journalists, librarians, artists, activists, and many more. This range is deliberate. We believe that each commenter brings a valuable perspective, and we aim to create an environment in which each of those who contributes comments feels free to express their thoughts, ideas, concerns, and critiques.

Our goal is not that this online review will always feel comfortable. Discussions of difficult topics always involve discomfort. Rather, our goal is that the review process remain civil and professional, in the interest of working towards a final version of this book in which the topics and issues that we engage, as well as the projects that we describe and the voices we include, are represented as accurately as possible, in language that is respectful, inclusive, and clear.

Please feel free to email Catherine and/or Lauren if you encounter any comments that violate this code of conduct, or if you have any comments that you would rather not publicly disclose.

This code of conduct draws from the Digital Frontiers Code of Conduct, the Society for Medieval Feminist Scholarship's Ground Rules for Civil Discussions of Difficult Issues, and the African American History, Culture, and Digital Humanities (AADHum) Initiative's Statement of Our Values.

Our Values and Our Metrics for Achieving Them

We insist on intersectionality

Feminism has always been multi-vocal and multi-racial, but the movements' diverse voices have not always been valued equally. The women's suffrage movement largely excluded Black women and the abolition of slavery from its agenda. In the 1970s, lesbian feminists were called "the purple menace" by straight feminists. But feminism fails altogether if it is only for elite, white, straight, Christian, Anglo women. The work of activists and scholars, particularly Black feminists, over the past forty years insists on a feminism that is intersectional, meaning it looks at issues of social power related not just to gender, but also to race, class, ability, sexuality, immigrant status, and more. It does so, moreover, by looking to collectives as well as individuals, structural issues as well as specific instances of injustice.

We advocate for equity

Equity is both an outcome and a process. Future justice must account for an unjust past in which some groups' knowledges have been valued and others have been "subjugated," as Patricia Hill Collins teaches us. In the process of achieving equity, those of us in positions of relative power must learn to listen deeper and listen differently – with the ultimate goal of taking action against the status quo that benefits us at the expense of others. For this reason, we listen and give priority in the text to voices who speak from marginalized perspectives, whether because of their gender, ability, race, class, colonial status, or other aspects of their identity.

We prioritize proximity

As Kimberly Seals Allers, women's health advocate, says, "Whatever the question, the answer is in the community." People in a community know its problems intimately, and they know which phenomena go uncounted, underreported, or neglected by institutions in power (or, conversely, who is overly surveilled by institutions in power). They also know what interventions will work to solve those problems. In this book, we try to prioritize voices with closer and more

direct experience of issues of injustice over those that study a data injustice from a distance.

We acknowledge the humanity of data

We recognize that the transformation of human experience into data often entails a reduction in complexity and context. We further acknowledge that there is a long history of data being "all too often wielded as an instrument of oppression, reinforcing inequality and perpetuating injustice," as the group Data for Black Lives explains. We keep these inherent constraints in mind as we write, attempting to introduce context and complexity whenever possible, and acknowledge the limits of the methods we discuss as well as their strengths.

We are reflexive, transparent and accountable

Acknowledging that our knowledge is shaped by our own perspectives and limitations, we strive to be reflexive, transparent, and accountable for our work. We are on a journey towards justice and that inevitably involves making mistakes. We are grateful to those who have shown us generosity in letting us learn up to this point. And we respectfully say to our future teachers that you will find in us open listeners – we recognize direct and critical words as a generous offer and a vote of confidence in our ability to hear and be transformed by you.

To that end, we have an evolving table of explicit metrics that will guide us in auditing our citations and the examples that we elevate in the book. We note, here, that our foregrounding of race and racism reflects our location in the United States, where the most entrenched issues of inequality and injustice have racism at their source.

NB: The metrics for this draft (see "Draft Metrics" below) were compiled by Izii Carter, a graduate student of journalism and research assistant for the *Data Feminism* project. We plan to take these metrics into account as we revise, and will release the final metrics upon the publication of the book.

Structural Problem	**Aspirational Metrics to Live Our Values For This Book**	**Draft Metrics**	**Final Metrics**

Racism	• 75% of citations of feminist scholarship from people of color • 75% of examples of feminist data projects discussed led by people of color	Scholarship: 36% from people of color Projects: 49% led by people of color
Patriarchy	• 75% of all citations and examples from women and nonbinary people	67% of citations and examples from women and nonbinary people

Classism	• Acknowledge that data science, as a field, is premised on economic, educational, and technological privilege • 50% of feminist projects discussed come from outside the academy • Example or theorist in every chapter that demonstrates how the ideas can be applied without expensive technology and/or formal training	Projects: 88% from outside academy 10 of 10 chapters feature non-academic example and/or theorist

Colonialism	• 30% of projects discussed come from the Global South • Example or theorist in every chapter about indigenous knowledges and/or activism	Projects: 8.5% from the Global South 5 of 10 chapters feature indigenous example and/or theorist
Transgender Oppression	• Center trans perspectives in discussions of the gender binary • Use transinclusive language throughout the book • Example or theorist in every chapter from a transgender perspective	3 of 10 chapters feature transgender example and/or theorist

Heteronormativity	• Resist assumptions about family structure and gender roles • Example or theorist in every chapter that illustrates the power of communal (vs. family) support networks	10 of 10 chapters feature communal example and/or theorist
Ableism	• Challenge the dominance of visualization in the presentation of data • Example or theorist in every chapter that employs non-visual methods of presenting data	9 of 10 chapters feature non-visual example and/or theorist

Proximity	• 50% of feminist projects discussed feature and quote people directly impacted by an issue (versus those who study or report on the phenomena from a distance)	Projects: 49% feature people directly impacted