demonstrates how data science can be enlisted on behalf of individuals and communities that need more power on their side.[66]

Data Science with Whose Interests and Goals?

Far too often, the problem is not that data about minoritized groups are missing but the reverse: the databases and data systems of powerful institutions are built on the excessive surveillance of minoritized groups. This results in women, people of color, and poor people, among others, being overrepresented in the data that these systems are premised upon. In *Automating Inequality*, for example, Virginia Eubanks tells the story of the Allegheny County Office of Children, Youth, and Families in western Pennsylvania, which employs an algorithmic model to predict the risk of child abuse in any particular home.[67] The goal of the model is to remove children from potentially abusive households before it happens; this would appear to be a very worthy goal. As Eubanks shows, however, inequities result. For wealthier parents, who can more easily access private health care and mental health services, there is simply not that much data to pull into the model. For poor parents, who more often rely on public resources, the system scoops up records from child welfare services, drug and alcohol treatment programs, mental health services, Medicaid histories, and more. Because there are far more data about poor parents, they are oversampled in the model, and so their children are overtargeted as being at risk for child abuse—a risk that results in children being removed from their families and homes. Eubanks argues that the model "confuse[s] parenting while poor with poor parenting."

This model, like many, was designed under two flawed assumptions: (1) that more data is always better and (2) that the data are a neutral input. In practice, however, the reality is quite different. The higher proportion of poor parents in the database, with more complete data profiles, the more likely the model will be to find fault with poor parents. And data are never neutral; they are always the biased output of unequal social, historical, and economic conditions: this is the matrix of domination once again.[68] Governments can and do use biased data to marshal the power of the matrix of domination in ways that amplify its effects on the least powerful in society. In this case, the model becomes a way to administer and manage classism in the disciplinary domain—with the consequence that poor parents' attempts to access resources and improve their lives, when compiled as data, become the same data that remove their children from their care.

So this raises our next *who question*: Whose goals are prioritized in data science (and whose are not)? In this case, the state of Pennsylvania prioritized its bureaucratic goal of efficiency, which is an oft-cited reason for coming up with a technical solution to a social and political dilemma. Viewed from the perspective of the state, there were simply not enough employees to handle all of the potential child abuse cases, so it needed a mechanism for efficiently deploying limited staff—or so the reasoning goes. This is what Eubanks has described as a *scarcity bias*: the idea that there are not enough resources for everyone so we should think small and allow technology to fill the gaps. Such thinking, and the technological "solutions" that result, often meet the goals of their creators—in this case, the Allegheny County Office of Children, Youth, and Families—but not the goals of the children and families that it purports to serve.

Corporations also place their own goals ahead of those of the people their products purport to serve, supported by their outsize wealth and the power that comes with it. For example, in 2012, the *New York Times* published an explosive article by Charles Duhigg, "How Companies Learn Your Secrets,"[69] which soon became the stuff of legend in data and privacy circles. Duhigg describes how Andrew Pole, a data scientist working at Target, was approached by men from the marketing department who asked, "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?"[70] He proceeded to synthesize customers' purchasing histories with the timeline of those purchases to give each customer a so-called pregnancy prediction score (figure 1.6).[71] Evidently, pregnancy is the second major life event, after leaving for college, that determines whether a casual shopper will become a customer for life.

Target turned around and put Pole's pregnancy detection model into action in an automated system that sent discount coupons to possibly pregnant customers. Win-win—or so the company thought, until a Minneapolis teenager's dad saw the coupons for baby 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 model informed Target before the teenager informed her family. By analyzing the purchase dates of approximately twenty-five common products, such as unscented lotion and large bags of cotton balls, the model found a set of purchase patterns that were highly correlated with pregnancy status and expected due date. But the win-win quickly became a lose-lose, as Target lost the trust of its customers in a PR disaster and the Minneapolis teenager lost far worse: her control over information related to her own body and her health.

Figure 1.6: Screenshot from a video of statistician Andrew Pole’s presentation at Predictive Analytics World about Target’s pregnancy detection model in October 2010, titled “How Target Gets the Most out of Its Guest Data to Improve Marketing ROI.” He discusses the model at 47:50. Image by Andrew Pole for Predictive Analytics World. *Source:* Andrew Pole, “How Target Gets the Most out of Its Guest Data to Improve Marketing ROI,” filmed October 2010 at Predictive Analytics World, video, 47:50, https://www.predictiveanalyticsworld.com/patimes/how-target-gets-the-most-out-of-its-guest-data-to-improve-marketing-roi/6815/.

This story has been told many times: first by Pole, the statistician; then by Duhigg, the *New York Times* journalist; then by many other commentators on personal privacy and corporate overreach. But it is not only a story about privacy: it is also a story about gender injustice—about how corporations approach data relating to women’s bodies and lives, and about how corporations approach data relating to minoritized populations more generally. Whose goals are prioritized in this case? The corporation’s, of course. For Target, the primary motivation was maximizing profit, and quarterly financial reports to the board are the measurement of success. Whose goals are *not* prioritized? The teenager’s and those of every other pregnant woman out there.

How did we get to the point where data science is used almost exclusively in the service of profit (for a few), surveillance (of the minoritized), and efficiency (amidst scarcity)? It’s worth stepping back to make an observation about the organization of the data economy: data are expensive and resource-intensive, so only already powerful institutions—corporations, governments, and elite research universities—have the means to work with them at scale. These resource requirements result in data science that serves the primary goals of the institutions themselves. We can think of these goals as the *three Ss*: science (universities), surveillance (governments), and selling (corporations). This is not a normative judgment (e.g., “all science is bad”) but rather an observation about the organization of resources. If science, surveillance, and selling are the main goals that data are serving, because that’s who has the money, then what other goals and purposes are going underserved?

Let’s take “the cloud” as an example. As server farms have taken the place of paper archives, storing data has come to require large physical spaces. A project by the Center for Land Use Interpretation (CLUI) makes this last point plain (figure 1.7). In 2014, CLUI set out to map and photograph data centers around the United States, often in those seemingly empty in-between areas we now call *exurbs*. In so doing, it called attention to “a new kind of physical information architecture” sprawling across the United States: “windowless boxes, often with distinct design features such as an appliqué of surface graphics or a functional brutalism, surrounded by cooling systems.” The environmental impacts of the cloud—in the form of electricity and air conditioning—are enormous. A 2017 Greenpeace report estimated that the global IT sector, which is largely US-based, accounted for around 7 percent of the world’s energy use. This is more than some of largest countries in the world, including Russia, Brazil, and Japan.[72] Unless that energy comes from renewable sources (which the Greenpeace report shows that it does not), the cloud has a significant accelerating impact on global climate change.

So the cloud is not light and it is not airy. And the cloud is not cheap. The cost of constructing Facebook’s newest data center in Los Lunas, New Mexico, is expected to reach $1 billion.[73] The electrical cost of that center alone is estimated at $31 million per year.[74] These numbers return us to the question about financial resources: Who has the money to invest in centers like these? Only powerful corporations like Facebook and Target, along with wealthy governments and elite universities, have the resources to collect, store, maintain, analyze, and mobilize the largest amounts of data. Next, who is in charge of these well-resourced institutions? Disproportionately men, even more disproportionately white men, and even more than that, disproportionately rich white men. Want the data on that? Google’s Board of Directors is comprised of 82 percent white men. Facebook’s board is 78 percent male and 89 percent white. The 2018 US Congress was 79 percent male—actually a better percentage than in previous years—and with a median net worth of five times more than the average American household.[75] These are the people who experience the most privilege within the matrix of domination, and they are also the people who benefit the most from the current status quo.[76]

Figure 1.7: Photographs from Networked Nation: The Landscape of the Internet in America, an exhibition by the Center for Land Use Interpretation staged in 2013. The photos show four data centers located in North Bergen, NJ; Dalles, OR; Ashburn, VA; and Lockport, NY (counterclockwise from top right). They show how the “cloud” is housed in remote locations and office parks around the country. Images by the Center for Land Use Interpretation. *Source: Networked Nation: The Landscape of the Internet in America*, exhibit, 2013, Center for Land Use Interpretation. *Credit:* Images by the Center for Land Use Interpretation.

In the past decade or so, many of these men at the top have described data as “the new oil.”[77] It’s a metaphor that resonates uncannily well—even more than they likely intended. The idea of data as some sort of untapped natural resource clearly points to the potential of data for power and profit once they are processed and refined, but it also helps highlight the exploitative dimensions of extracting data from their source—people—as well as their ecological cost. Just as the original oil barons were able to use their riches 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 corporate gain to consolidate control over their customers. But unlike crude oil, which is extracted from the earth and then sold to people, data are both extracted from people and sold back to them—in the form of coupons like the one the Minneapolis teen received in the mail, or far worse.[78]

This extractive system creates a profound asymmetry between who is collecting, storing, and analyzing data, and whose data are collected, stored, and analyzed.[79] The goals that drive this process are those of the corporations, governments, and well-resourced universities that are dominated by elite white men. And those goals are neither neutral nor democratic—in the sense of having undergone any kind of participatory, public process. On the contrary, focusing on those *three Ss*—science, surveillance, and selling—to the exclusion of other possible objectives results in significant oversights with life-altering consequences. Consider the Target example as the flip side of the missing data on maternal health outcomes. Put crudely, there is no profit to be made collecting data on the women who are dying in childbirth, but there is significant profit in knowing whether women are pregnant.

How might we prioritize different goals and different people in data science? How might data scientists undertake a feminist analysis of power in order to tackle bias at its source? Kimberly Seals Allers, a birth justice advocate and author, is on a mission to do exactly that in relation to maternal and infant care in the United States. She followed the Serena Williams story with great interest and watched as Congress passed the Preventing Maternal Deaths Act of 2018. This bill funded the creation of maternal health review committees in every state and, for the first time, uniform and comprehensive data collection at the federal level. But even as more data have begun to be collected about maternal mortality, Seals Allers has remained frustrated by the public conversation: “The statistics that are rightfully creating awareness around the Black maternal mortality crisis are also contributing to this gloom and doom deficit narrative. White people are like, ‘how can we save Black women?’ And that’s not the solution that we need the data to produce.”[80]

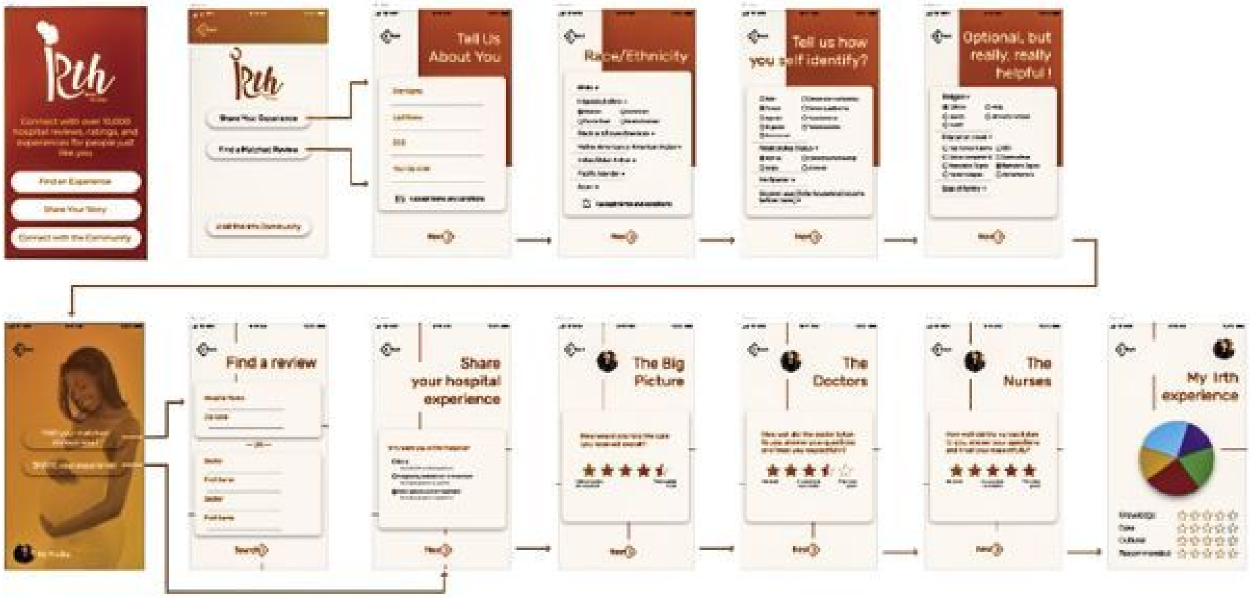

Figure 1.8: Irth is a mobile app and web platform focused on removing bias from birth (including prenatal, birth, and postpartum health care). Users post intersectional reviews of the care they received from individual nurses and doctors, as well as whole practices and hospitals. When parents to be are searching for providers, they can consult Irth to see what kind of care people like them received in the hands of specific caregivers. Wireframes from Irth's first prototype are shown here. Images by Kimberly Seals Allers and the Irth team, 2019. *Credit:* Kimberly Seals Allers and the Irth team.

Seals Allers—and her fifteen-year-old son, Michael—are working on their own data-driven contribution to the maternal and infant health conversation: a platform and app called Irth—from *birth*, but with the *b* for *bias* removed (figure 1.8). One of the major contributing factors to poor birth outcomes, as well as maternal and infant mortality, is biased care. Hospitals, clinics, and caregivers routinely disregard Black women's expressions of pain and wishes for treatment.[81] As we saw, Serena Williams's own story almost ended in this way, despite the fact that she is an international tennis star. To combat this, Irth operates like an intersectional Yelp for birth experiences. Users post ratings and reviews of their prenatal, postpartum, and birth experiences at specific hospitals and in the hands of specific caregivers. Their reviews include important details like their race, religion, sexuality, and gender identity, as well as whether they felt that those identities were respected in the care that they received. The app also has a taxonomy of bias and asks users to tick boxes to indicate whether and how they may have experienced different types of bias. Irth allows parents who are seeking care to search for a review from someone like them—from a racial, ethnic, socioeconomic, and/or gender perspective—to see how they experienced a certain doctor or hospital.

Seals Allers's vision is that Irth will be both a public information platform, for individuals to find better care, and an accountability tool, to hold hospitals and providers responsible for systemic bias. Ultimately, she would like to present aggregated stories and data analyses from the platform to hospital networks to push for change grounded in women's and parents' lived experiences. "We keep telling the story of maternal mortality from the grave," she says. "We have to start preventing those deaths by sharing the stories of people who actually lived."[82]

Irth illustrates the fact that "doing good with data" requires being deeply attuned to the things that fall outside the dataset—and in particular to how datasets, and the data science they enable, too often reflect the structures of power of the world they draw from. In a world defined by unequal power relations, which shape both social norms and laws about how data are used and how data science is applied, it remains imperative to consider who gets to do the "good" and who, conversely, gets someone else's "good" done to them.

Examine Power

Data feminism begins by examining how power operates in the world today. This consists of asking *who questions* about data science: Who does the work (and who is pushed out)? Who benefits (and who is neglected or harmed)? Whose priorities get turned into products (and whose are overlooked)? These questions are relevant at the level of individuals and organizations, and are absolutely essential at the level of society. The current answer to most of these questions is "people from dominant groups," which has resulted in a *privilege hazard* so acute that it explains the near-daily revelations about another sexist or racist data product or algorithm. The *matrix of domination* helps us to understand how the privilege hazard—the result of unequal distributions of power—plays out in different domains. Ultimately, the goal of examining power is not only to understand it, but also to be able to challenge and change it. In the next chapter, we explore several approaches for challenging power with data science.

Footnotes

1. Serena Williams, "Meet Alexis Olympia Ohanian Jr. You have to check out link in bio for her amazing journey. Also check out my IG stories 😍😍❤❤," September 13, 2017, https://www.instagram.com/p/BY-7H9zhQD7/. ↩
2. See Serena Williams, Facebook, January 15, 2018, https://www.facebook.com/SerenaWilliams/videos/10156086135726834/. ↩