Data Feminism

Catherine D'Ignazio and Lauren F. Klein

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in ITC Stone Serif Std and ITC Stone Sans Std by Toppan Best-set Premedia Limited. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Kanarinka, author. | Klein, Lauren F., author.
Title: Data feminism / Catherine D'Ignazio and Lauren F. Klein.
Description: Cambridge, Massachusetts : The MIT Press, [2020] | Series: <Strong> ideas series | Includes bibliographical references and index.
Identifiers: LCCN 2019036137 | ISBN 9780262044004 (hardcover)
Subjects: LCSH: Feminism. | Feminism and science. | Big data—Social aspects. | Quantitative research—Methodology—Social aspects. | Power (Social sciences)
Classification: LCC HQ1190 .K375 2020 | DDC 305.42—dc23
LC record available at https://lccn.loc.gov/2019036137

10 9 8 7 6 5 4 3 2 1

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

Principle: Embrace Pluralism

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

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

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

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

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

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

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

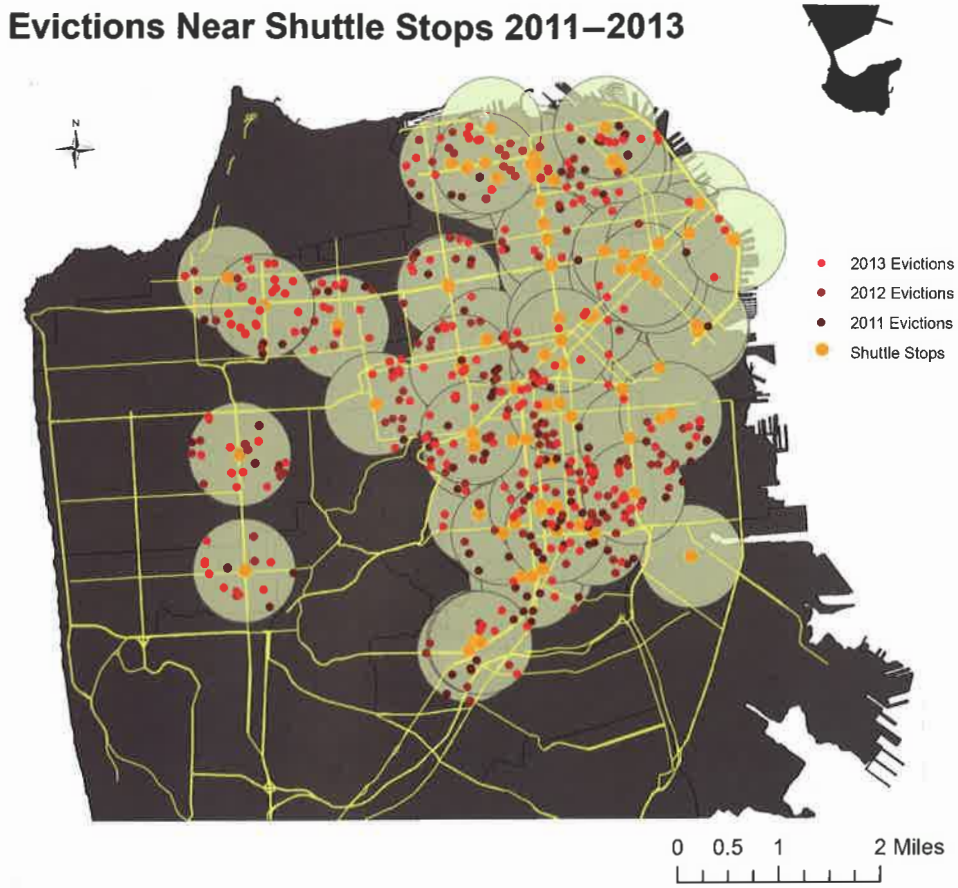

Figure 5.1
Tech Bus Stop Eviction Map by the Anti-Eviction Mapping Project (2014) plots evictions with "Google bus stops" in San Francisco. The group's analysis showed that 69 percent of no-fault evictions in the city occurred within four blocks of a tech bus stop. Courtesy of the Anti-Eviction Mapping Project.

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

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

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

Figure 5.2

Narratives of Displacement and Resistance (2018) by the Anti-Eviction Mapping Project (a), including a detail view from Phyllis Bowie's story (b). Courtesy of the Anti-Eviction Mapping Project. Made in collaboration with the San Francisco Ruth Assawa School of the Arts. The interview was shot by Marianne Maeckelbergh and Brandon Jourdan and edited by students Shilo Arkinson and Avidan Novogrodsky-Godt, and facilitated by Alexandra Lacey and Jin Zhu.

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

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

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

Strangers in the Dataset

Data cleaning holds a special kind of mythos in data science. "It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data," writes Hadley Wickham in the first sentence of the abstract of "Tidy Data," his widely cited paper from 2014.[13] Wickham is also the author of the *tidyr* package for R, a popular programming language and statistical computing platform. (*Packages* are prewritten collections of functions, and other forms of code, that can be installed and used in any project.) Wickham's package shares the sentiment expressed in his paper: that data are inherently messy and need to be tidied and tamed.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

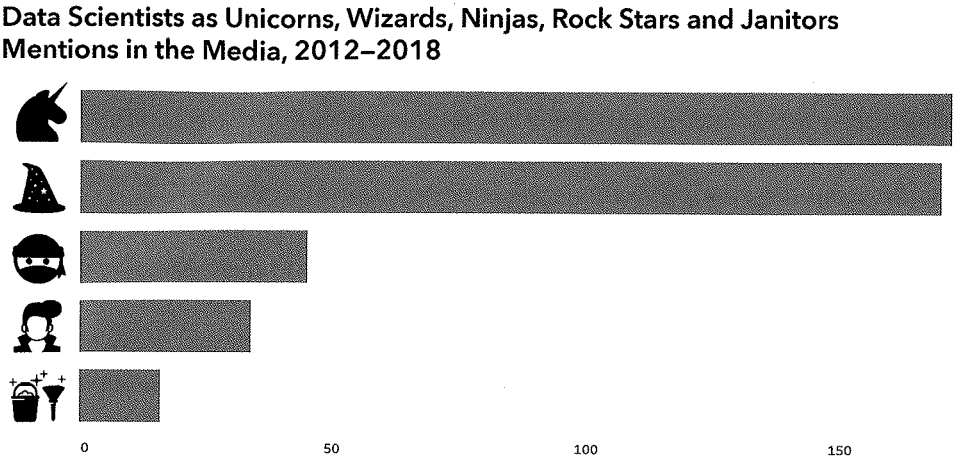

Figure 5.3
Searching Media Cloud between 2012 and 2018 shows that *unicorn* is the most commonly referenced metaphor in relation to data scientists, with *wizard* a close second. There are fewer than fifty articles about data *ninjas*, *rock stars*, and *janitors*, but they appear in high-profile venues like the *Washington Post* and *Forbes*. The Media Cloud platform at www.mediacloud.org was developed at the MIT Center for Civic Media and archives just under a million articles and blog posts every day. Graphic by Catherine D'Ignazio. Data from www.mediacloud.org.

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

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

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

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

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

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

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

On Power, Pluralism, and Process

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

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

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

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

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

Reflexivity can sometimes be as simple as being explicit about or even visualizing who is doing the counting and mapping behind the scenes.[36] Take the example of the aerial-mapping image in figure 5.5.[37] The Public Laboratory for Open Technology and Science (Public Lab) is a citizen science group that got its start during the BP oil spill in 2010.[38] It makes high-resolution aerial maps by flying balloons and kites, which dangle cheap digital cameras over the environmental sites they seek to study. The technique is low cost, but the imagery produced is often higher resolution than existing satellite imagery because of the proximity to the ground. As in the image, the mappers themselves are often visible in the final product in the form of little bodies, gathered in boats or standing in clumps on a shoreline, looking up at the camera above them. The balloon string leads the eye back to their forms. Here the creators are not distanced or absent but represented in the final product. Literally.

Data for "Good" versus Data for Co-liberation

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

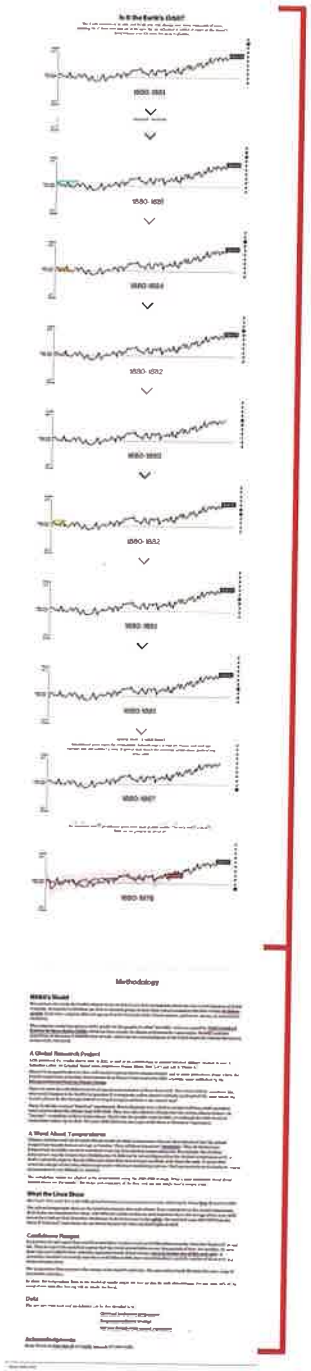

Figure 5.4

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

Figure 5.5
From a Public Lab research note by Eymund Diegel about mapping sewage flows in the Gowanus Canal in 2012 after Hurricane Sandy. Note the people on boats doing the mapping and the balloon tether that links the camera and image back to their bodies. Courtesy of Eymund Diegel for Public Lab.

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

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

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

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

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

Table 5.1
Features of "data for good" versus data for co-liberation

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

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

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

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

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

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

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

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

Data for Co-liberation in Action

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

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

Figure 5.6
The process of making a data mural involves conversation, building prototypes with craft materials, workshops in data analysis, and actual painting. Courtesy of Data Therapy, Emily and Rahul Bhargava, 2018.

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

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

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

Figure 5.7
The Groundwork Somerville data mural, painted by youth, staff and volunteers at Groundwork Somerville, and the Bhargavas in 2013. Courtesy of Data Therapy, Emily and Rahul Bhargava.

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

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

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

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

Does the Campfire Scale?

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

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

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

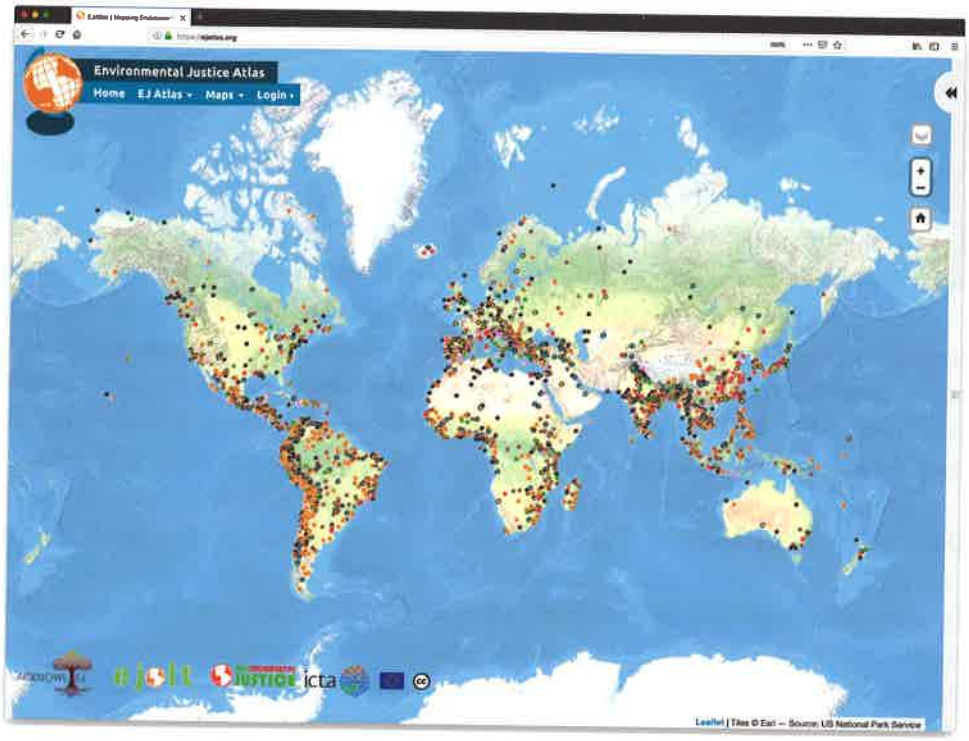

Figure 5.8
The Global Atlas of Environmental Justice (EJ Atlas; https://ejatlas.org/) works in partnership with activists, civil society organizations, and social movements to systematically document ecological conflicts around the globe. The scope and scale of the EJ Atlas enables activists to connect with others and facilitates researchers studying conflicts in a quantitative and comparative context, without sacrificing a commitment to a pluralistic process and the dignity of local and community knowledges. Courtesy of the Global Atlas of Environmental Justice, 2019.

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

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

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

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

Embrace Pluralism

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

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

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