Data Science for Whom?

One of the downstream effects of the privilege hazard—the risks incurred when people from dominant groups create most of our data products—is not only that datasets are biased or unrepresentative, but that they never get collected at all. Mimi Onuoha—an artist, designer, and educator—has long been asking *who questions* about data science. Her project, *The Library of Missing Datasets* (figure 1.4), is a list of datasets that one might expect to already exist in the world, because they help to address pressing social issues, but that in reality have never been created. The project exists as a website and as an art object. The latter consists of a file cabinet filled with folders labeled with phrases like: "People excluded from public housing because of criminal records," "Mobility for older adults with physical disabilities or cognitive impairments," and "Total number of local and state police departments using stingray phone trackers (IMSI-catchers)." Visitors can tab through the folders and remove any particular folder of interest, only to reveal that it is empty. They all are. The datasets that should be there are "missing."

Figure 1.4: The Library of Missing Datasets, by Mimi Onuoha (2016) is a list of datasets that are not collected because of bias, lack of social and political will, and structural disregard. Courtesy of Mimi Onuoha. Photo by Brandon Schulman. *Credit:* Photo by Brandon Schulman

By compiling a list of the datasets that are missing from our "otherwise data-saturated" world, Onuoha explains, "we find cultural and colloquial hints of what is deemed important" and what is not. "Spots that we've left blank reveal our hidden social biases and indifferences," she continues. And by calling attention to these datasets as "missing," she also calls attention to how the matrix of domination encodes these "social biases and indifferences" across all levels of society.[49] Along similar lines, foundations like Data2X and books like *Invisible Women* have advanced the idea of a systematic "gender data gap" due to the fact that the majority of research data in scientific studies is based around men's bodies. The downstream effects of the gender data gap range from annoying—cell phones slightly too large for women's hands, for example—to fatal. Until recently, crash test dummies were designed in the size and shape of men, an oversight that meant that women had a 47 percent higher chance of car injury than men.[50]

The *who question* in this case is: Who benefits from data science and who is overlooked? Examining those gaps can sometimes mean calling out missing datasets, as Onuoha does; characterizing them, as *Invisible Women* does; and advocating for filling them, as Data2X does. At other times, it can mean collecting the missing data yourself. Lacking comprehensive data about women who die in childbirth, for example, ProPublica decided to resort to crowdsourcing to learn the names of the estimated seven hundred to nine hundred US women who died in 2016.[51] As of 2019, they've identified only 140. Or, for another example: in 1998, youth living in Roxbury—a neighborhood known as "the heart of Black culture in Boston"[52]—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.[53]

Scholars have proposed various names for these instances of ground-up data collection, including *counterdata* or *agonistic data* collection, *data activism*, *statactivism*, and *citizen science* (when in the service of environmental justice).[54] Whatever it's called, it's been going on for a long time. In 1895, civil rights activist and pioneering data journalist Ida B. Wells assembled a set of statistics on the epidemic of lynching that was sweeping the United States.[55] She accompanied her data with a meticulous exposé of the fraudulent claims made by white people—typically, that a rape, theft, or assault of some kind had occurred (which it hadn't in most cases) and that lynching was a justified response. Today, an organization named after Wells—the Ida B. Wells Society for Investigative Reporting—continues her mission by training up a new generation of journalists of color in the skills of data collection and analysis.[56]

A counterdata initiative in the spirit of Wells is taking place just south of the US border, in Mexico, where a single woman is compiling a comprehensive dataset on femicides—gender-related killings of women and girls.[57] María Salguero, who also goes by the name Princesa, has logged more than five thousand cases of femicide since 2016.[58] Her work provides the most accessible information on the subject for journalists, activists, and victims' families seeking justice.

The issue of femicide in Mexico rose to global visibility in the mid-2000s with widespread media coverage about the deaths of poor and working-class women in Ciudad Juárez. A border town, Juárez is the site of more than three hundred *maquiladoras*: factories that employ women to assemble goods and electronics, often for low wages and in substandard working conditions. Between 1993 and 2005, nearly four hundred of these women were murdered, with around a third of those murders exhibiting signs of exceptional brutality or sexual violence. Convictions were made in

only three of those deaths. In response, a number of activist groups like Ni Una Más (Not One More) and Nuestras Hijas de Regreso a Casa (Our Daughters Back Home) were formed, largely motivated by mothers demanding justice for their daughters, often at great personal risk to themselves.[59]

These groups succeeded in gaining the attention of the Mexican government, which established a Special Commission on Femicide. But despite the commission and the fourteen volumes of information about femicide that it produced, and despite a 2009 ruling against the Mexican state by the Inter-American Human Rights Court, and despite a United Nations Symposium on Femicide in 2012, and despite the fact that sixteen Latin American countries have now passed laws defining femicide—despite all of this, deaths in Juárez have continued to rise.[60] In 2009 a report pointed out that one of the reasons that the issue had yet to be sufficiently addressed was the lack of data.[61] Needless to say, the problem remains.

How might we explain the missing data around femicides in relation to the four domains of power that constitute Collins's matrix of domination? As is true in so many cases of data collected (or not) about women and other minoritized groups, the collection environment is compromised by imbalances of power.

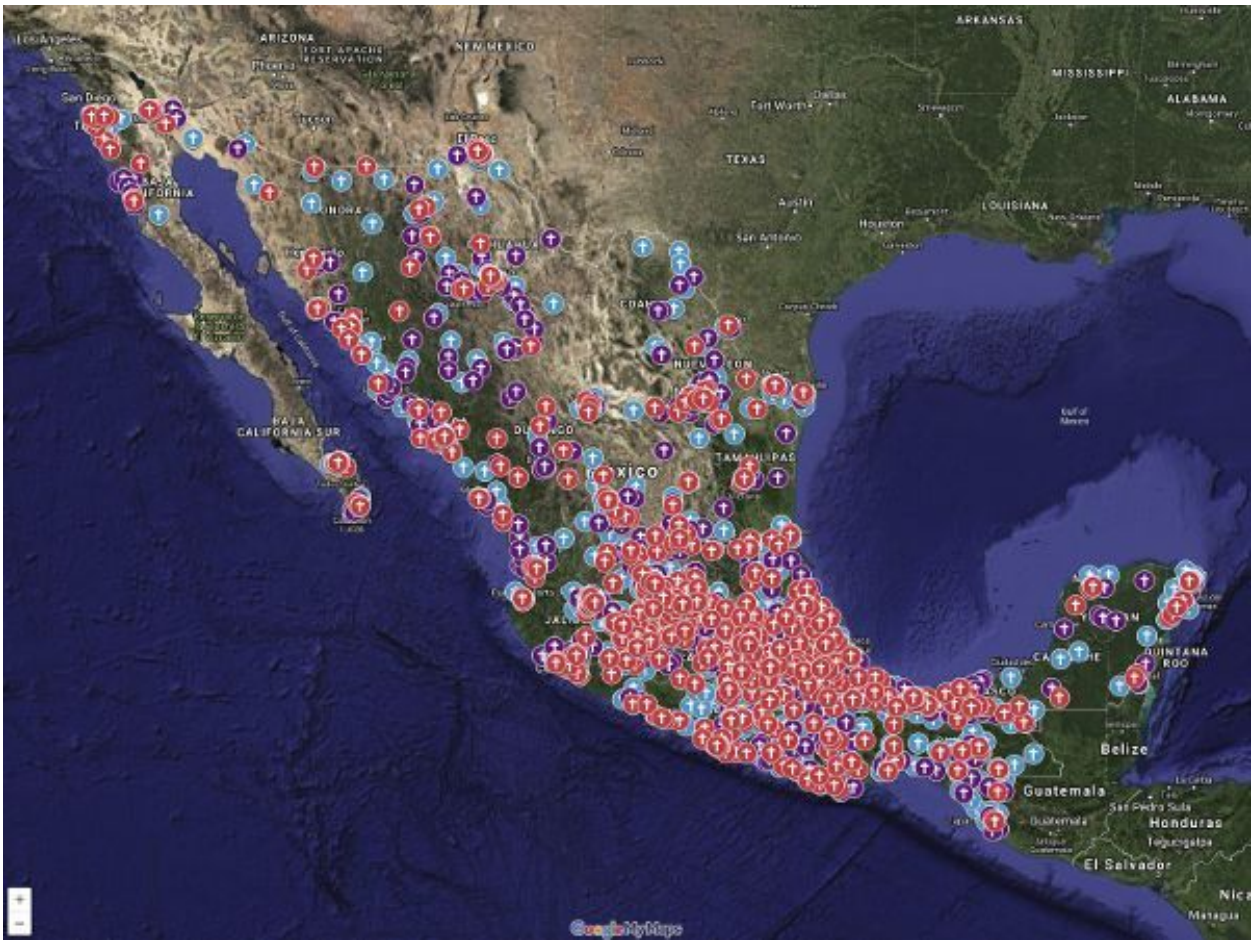
The most grave and urgent manifestation of the matrix of domination is within the interpersonal domain, in which cis and trans women become the victims of violence and murder at the hands of men. Although law and policy (the structural domain) have recognized the crime of femicide, no specific policies have been implemented to ensure adequate information collection, either by federal agencies or local authorities. Thus the disciplinary domain, in which law and policy are enacted, is characterized by a deferral of responsibility, a failure to investigate, and victim blaming. This persists in a somewhat recursive fashion because there are no consequences imposed within the structural domain. For example, the Special Commission's definition of femicide as a "crime of the state" speaks volumes to how the government of Mexico is deeply complicit through inattention and indifference.[62]

Of course, this inaction would not have been tolerated without the assistance of the hegemonic domain—the realm of media and culture—which presents men as strong and women as subservient, men as public and women as private, trans people as deviating from "essential" norms, and nonbinary people as nonexistent altogether. Indeed, government agencies have used their public platforms to blame victims. Following the femicide of twenty-two-year-old Mexican student Lesvy Osorio in 2017, researcher Maria Rodriguez-Dominguez documented how 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.[63] This led to justified public backlash, and to the hashtag #SiMeMatan (If they kill me), which prompted sarcastic tweets such as "#SiMeMatan it's because I liked to go out at night and drink a lot of beer."[64]

It is into this data collection environment, characterized by extremely asymmetrical power relations, that María Salguero has inserted her femicides map. Salguero manually plots a pin on the map for every femicide that she collects through media reports or through crowdsourced contributions (figure 1.5a). One of her goals is to "show that these victims [each] had a name and that they had a life," and 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, and whether the victim was transgender, as well as the full content of the news report that served as the source. Figure 1.5b shows a detailed view for a single report from an unidentified transfemicide, including the date, time, location, and media article about the killing. It can take Salguero three to four hours a day to do this unpaid work. She takes occasional breaks to preserve her mental health, and she typically has a backlog of a month's worth of femicides to add to the map.

Although media reportage and crowdsourcing are imperfect ways of collecting data, this particular map, created and maintained by a single person, fills a vacuum created by her national government. The map has been used to help find missing women, and Salguero herself has testified before Mexico's Congress about the scope of the problem. Salguero is not affiliated with an activist group, but she makes her data available to activist groups for their efforts. Parents of victims have called her to give their thanks for making their daughters visible, and Salguero affirms this function as well: "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."


ARKANSAS
ARIZONA
NEW MEXICO
MISSISSIPPI
ALABAMA
TEXAS
LOUISIANA
Belize
Guatemala
Honduras
El Salvador

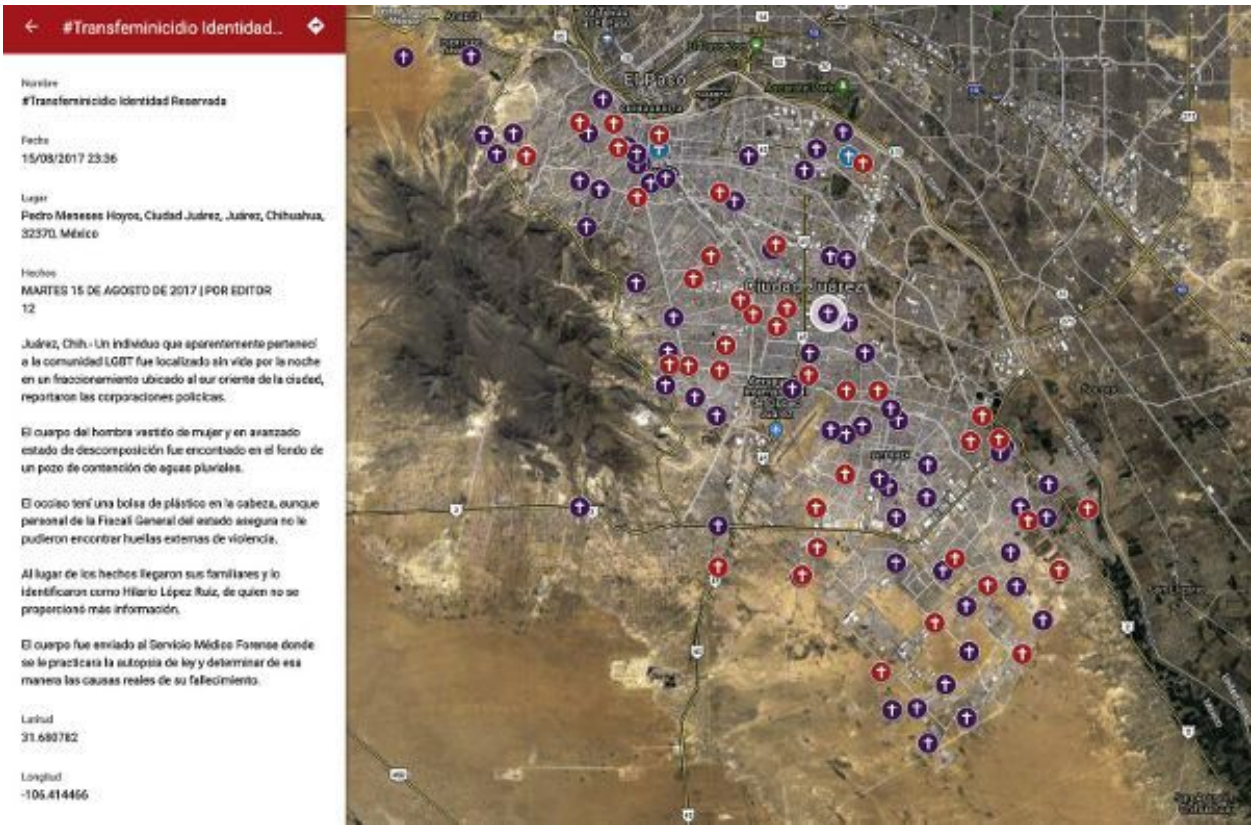


Figure 1.5: María Salguero's map of femicides in Mexico (2016–present) can be found at https://feminicidiosmx.crowdmap.com/. (a) Map extent showing the whole country. (b) A detailed view of Ciudad Juárez with a focus on a single report of an anonymous transfemicide. Salguero crowdsources points on the map based on reports in the press and reports from citizens to her. Courtesy of María Salguero. (a) *Source:* https://feminicidiosmx.crowdmap.com/. (b) *Source:* https://www.google.com/maps/d/u/0/viewer?mid=174IjBzP-fl_6wpRHg5pkGSj2egE&ll=21.347609098250942%2C-102.05467709375&z=5. *Credit:* María Salguero.

It is important to make clear that the example of missing data about femicides in Mexico is not an isolated case, either in terms of subject matter or geographic location. The phenomenon of missing data is a regular and expected outcome in all societies characterized by unequal power relations, in which a gendered, racialized order is maintained through willful disregard, deferral of responsibility, and organized neglect for data and statistics about those minoritized bodies who do not hold power. So too are examples of individuals and communities using strategies like Salguero's to fill in the gaps left by these missing datasets—in the United States as around the world.[65] If "quantification is representation," as data journalist Jonathan Stray asserts, then this offers one way to hold those in power accountable. Collecting counterdata

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