On Missing Data Sets

This repo will be periodically updated with more information, links, and topics.

Overview

What is a Missing Data Set?

"Missing data sets" are my term for the blank spots that exist in spaces that are otherwise data-saturated. My interest in them stems from the observation that within many spaces where large amounts of data are collected, there are often empty spaces where no data live. Unsurprisingly, this lack of data typically correlates with issues affecting those who are most vulnerable in that context.

The word "missing" is inherently normative. It implies both a lack and an ought: something does not exist, but it should. That which should be somewhere is not in its expected place; an established system is disrupted by distinct absence. Just because some type of data doesn't exist doesn't mean it's missing, and the idea of missing data sets is inextricably tied to a more expansive climate of inevitable and routine data collection.

Why Do They Matter?

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.

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
-

Why Are They Missing?

There are a number of reasons why a data set that seems like it *should* exist might not, and they are all tied to the quiet complications inherent in data collection. Below are four reasons, with accompanying real-world examples.

1. **Those who have the resources to collect data lack the incentive to (corollary: often those who have access to a dataset are the same ones who have the ability to remove, hide, or obscure it)**.

   Police brutality towards civilians provides a powerful example. Though policing and crime are among the most data-driven areas of public policy, traditionally there has been little history of standardized and rigorous data collected about police brutality.

   Nowadays we have a political and cultural climate where this issue has become one of public discussion. Public interest campaigns like Fatal Encounters and the Guardian's The Counted have helped fill that void. But even for these individuals/organizations, the work is difficult and time-consuming. The group who would make the most sense to monitor this issue—the law enforcement agents who *create* the data set in the first place—have no incentive to actually gather such data, which could prove incriminating.

2. **The data to be collected resist simple quantification (corollary: we prioritize collecting things that fit our modes of collection).**

The defining tension of data collection is the struggle of taking a messy, organic world and defining it in formats that are neat, clean, and structured.

Some things are difficult to collect and quantify by nature of their structure. We don't know how much US currency is outside of our borders. There's no incentive for other countries to monitor US currency within their countries, and the very nature of cash and the anonymity it affords makes it difficult to track.

But then there are other subjects that resist quantification entirely. Things like emotions are hard to quantify (at this time, at least). Institutional racism is subtle and deniable; it reveals itself more in effects than acts. Not all things are easily quantifiable, and at times the very desire to render the world more abstract, trackable, and machine-readable is an idea that itself deserves questioning.

3. **The act of collection involves more work than the benefit the presence of the data is perceived to give.**

Sexual assault and harassment are woefully underreported. And while there are many reasons why this is, one major one is that in many cases the very act of reporting sexual assault is a very intensive, painful, and difficult process. For some, the benefit of reporting isn't perceived to be equal or greater than the cost of the process.

4. **There are advantages to nonexistence.**

Every missing dataset is a testament to this fact. Just as the presence of data benefits someone, so too does the absence. This is important to keep in mind.

However, there's an even more specific angle to this point. To collect, record, and archive aspects of the world is an intentional act, one that typically benefits those who have the power to decide what should be collected. Often, remaining outside of the bounds of collection can be a form of response for a situationally-disadvantaged group. In short, sometimes a missing dataset can function as a form of protection.