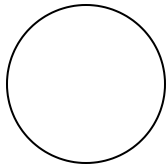


POINTS (/POINTS)

## THE POINT OF COLLECTION



**Mimi Onuoha**  
Advisor

(<https://datasociety.net/people/onuoha-mimi/>)

---

February 10, 2016

The conceptual, practical, and ethical issues surrounding “big data” and data in general begin at the very moment of data collection. Particularly when the data concern people, not enough attention is paid to the realities entangled within that significant moment and spreading out from it.

I try to do some disentangling here, through five theses around data collection — points that are worth remembering, communicating, thinking about, dwelling on, and keeping in mind, if you have anything to do with data on a daily basis (read: *all of us*) and want to do data responsibly.



CC BY-SA 2.0-licensed photo by Lynn Dombrowski

#### 1. Data sets are the results of their means of collection.

It's easy to forget that the people collecting a data set, and how they choose to do it, directly determines the data set.

An illustrative example can be found in the statistics for how many hate crimes were committed in the United States in 2012. According to the FBI Uniform Crime Reporting Program (UCR), the number was 5,796. However, the Department of Justice's Bureau of Statistics reported 293,800 hate crimes.

The reason for the variation was simple. The URC gathers data that is voluntarily reported by law enforcement agencies across the country. The Bureau of Statistics, on the other hand, distributes the National Crime Victimization Survey, which collects data from the victims of hate crimes. The result is a more transparent and inclusive surveying.

Same data set, two different means of collection, two wildly different results. What they show is an important fact we must keep in mind: There's no pure objectivity encoded into data sets. Each one is the result of a number of human processes and decisions that affect, in a variety of ways, the data that they aim to report. In this sense, the moment of data collection starts before any data is actually produced.

#### 2. As we collect more data, we prioritize things that fit patterns of collection.

Or as Rob Kitchin and Martin Dodge say in Code/Space, “The effect of abstracting the world is that the world starts to structure itself in the image of the capta and the code.” Data emerges from a world that is increasingly software-mediated, and software thrives on abstraction. It flattens out individual variations in favor of types and models.

As we abstract the world, we prioritize abstractions of the world. The more we look to data to answer our big questions (in areas like policing, safety, and security), the more incentives we have to shape the world into an input that fits into an algorithm. Our need to generate things that feed a model rings true even in cases where the messy bounds of experiences can’t be neatly categorized into bits and bytes, or easily retrieved from tables through queries.

Biometric data is a great example of this. Fingerprint authentication technologies and iris scanners point to a system where individuals are uniquely identified through metrics and data. In order for this to work, people themselves have to be conceptualized more and more as machine-readable.

### 3. Data sets outlive the rationale for their collection.

Spotify can come up with a list of reasons why having access to users’ photos, locations, microphones, and contact lists can improve the music streaming experience. But the reasons why they decide these forms of data might be useful can be less important than the fact that they have the data itself. This is because the needs or desires influencing the decisions to collect some type of data often eventually disappear, while the data produced as a result of those decisions have the potential to live for much longer. The data are capable of shifting and changing according to specific cultural contexts and to play different roles than what they might have initially been intended for.

Ultimately, the question of intention behind the collection or generation of a data set can be rendered irrelevant. Thinking through the moment of collection can reveal the distance between it and the data’s use. And it’s often far more critical to think about the potentials and possibilities surrounding what can be done with collected data.

### 4. Corollary: Especially combined, data sets reveal far more than intended.

We sometimes fail to realize that data sets, both on their own and combined with others, can be used to do far more than what they were originally intended for. You can make inferences from one data set that result in conclusions in completely different realms. Facebook, by having huge amounts of data on people and their networks, could make reasonable hypotheses regarding people’s sexual orientations.

People who work with data know this intimately, but it can often be difficult to see the connections between the collection of one thing and the inference of something else. Unfortunately, the effects of these connections can become very strongly felt. As Bruce Schneier puts it, “data we’re willing to share can imply conclusions that we don’t want to share.”

5. Data collection is a transaction that is the result of an invisible relationship.

This is a frame — connected to my first point — useful for understanding how to think about data collection on the whole:

Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings. I point this out not for any fluffy reasons related to humanism or human-centered design, but because power arises out of hierarchies, interactions, and dynamics. The below-the-surface work of a particular data set is joined to the reasons and means that created it and the relationships running through those reasons and means. If we can keep that in mind, we’re better positioned to see data as an intermediate result, one piece in a larger process, something that is as much human-oriented as it is systematic. The challenge is for us to keep in mind both aspects of data collection, to see systematic as well as human tensions and biases.

The point of data collection is a unique site for unpacking change, abuse, unfairness, bias, and potential. We can’t talk about responsible data without talking about the moment when data becomes data.

*Points: “The Point of Collection” takes off from Mimi Onuoha’s recent Machine Eatable talk: Data as Process. Thinking through the implications of the moment of data collection, she offers a compact set of reminders for those who work with and think about data. Tape it to your monitor. — Ed.*

Reporters and Media  
[press@datasociety.net](mailto:press@datasociety.net) (<mailto:press@datasociety.net>)

General Inquiries  
[info@datasociety.net](mailto:info@datasociety.net) (<mailto:info@datasociety.net>)

---

Subscribe to the Data & Society newsletter

Enter your email address



Unless otherwise noted, this site and its contents are licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

Website by [Jake Dow-Smith Studio](#)