Harvard
Business
Review

Analytics And Data Science

The Hidden Biases in Big Data

by Kate Crawford

April 01, 2013

This looks to be the year that we reach peak big data hype. From wildly popular big data conferences to columns in major newspapers, the business and science worlds are focused on how large datasets can give insight on previously intractable challenges. The hype becomes problematic when it leads to what I call "data fundamentalism," the notion that correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth. Former *Wired* editor-in-chief Chris Anderson embraced this idea in his comment, "with enough data, the numbers speak for themselves." But can big data really deliver on that promise? Can numbers actually speak for themselves?

Sadly, they can't. Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves.

For example, consider the Twitter data generated by Hurricane Sandy, more than 20 million tweets between October 27 and November 1. A fascinating study combining Sandy-related Twitter and Foursquare data produced some expected findings (grocery shopping peaks the night before the storm) and some surprising ones (nightlife picked up the day after — presumably when cabin fever strikes). But these data don't represent the whole picture. The greatest number of tweets about Sandy came from Manhattan. This makes sense given the city's high level of smartphone ownership and Twitter use, but it creates the illusion that Manhattan was the hub of the disaster. Very few messages originated from more severely affected locations, such as Breezy Point, Coney Island and Rockaway. As extended power blackouts drained batteries and limited cellular access, even fewer tweets came from the worst hit areas. In fact, there was much more going on outside the privileged, urban experience of Sandy that Twitter data failed to convey, especially in aggregate. We can think of this as a "signal problem": Data are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities.

While massive datasets may feel very abstract, they are intricately linked to physical place and human culture. And places, like people, have their own individual character and grain. For example, Boston has a problem with potholes, patching approximately 20,000 every year. To help allocate its resources efficiently, the City of Boston released the excellent StreetBump smartphone app, which draws on accelerometer and GPS data to help passively detect potholes, instantly reporting them to the city. While certainly a clever approach, StreetBump has a signal problem. People in lower income groups in the US are less likely to have smartphones, and this is particularly true of older residents, where smartphone penetration can be as low as 16%. For cities like Boston, this means that smartphone data sets are missing inputs from significant parts of the population — often those who have the fewest resources.

Fortunately Boston's Office of New Urban Mechanics is aware of this problem, and works with a range of academics to take into account issues of equitable access and digital divides. But as we increasingly rely on big data's numbers to speak for themselves, we risk misunderstanding the results and in turn misallocating important public resources. This could well have been the case had public health officials relied exclusively on Google Flu Trends, which mistakenly estimated that peak flu levels reached 11% of the US public this flu season, almost double the CDC's estimate of about 6%. While Google will not comment on the reason for the overestimation, it seems likely that it was caused by the extensive media coverage of the flu season, creating a spike in search queries. Similarly, we can imagine the substantial problems if FEMA had relied solely upon tweets about Sandy to allocate disaster relief aid.

Big data's signal problems won't disappear as the use of smartphones and other digital technologies increases. As the geographers Michael Crutcher and Matthew Zook noted after Hurricane Katrina, technologies are always differentially adopted, and "any divide in accessing digital technology is not a one-time event but a constantly moving target as new devices, software and cultural practices emerge." As we move into an era in which personal devices are seen as proxies for public needs, we run the risk that already existing inequities will be further entrenched. Thus, with every big data set, we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big data sets?

This points to the next frontier: how to address these weaknesses in big data science. In the near term, data scientists should take a page from social scientists, who have a long history of asking where the data they're working with comes from, what methods were used to gather and analyze it, and what cognitive biases they might bring to its interpretation (for more, see "Raw Data is an Oxymoron"). Longer term, we must ask how we can bring together big data approaches with small data studies — computational social science with traditional qualitative methods. We know that data insights can be found at multiple levels of granularity, and by combining methods such as ethnography with analytics, or conducting semi-structured interviews paired with information retrieval techniques, we can add depth to the data we collect. We get a much richer sense of the world when we ask people the why and the how not just the “how many”. This goes beyond merely conducting focus groups to confirm what you already want to see in a big data set. It means complementing data sources with rigorous qualitative research. Social science methodologies may make the challenge of understanding big data more complex, but they also bring context-awareness to our research to address serious signal problems. Then we can move from the focus on merely “big” data towards something more three-dimensional: data with depth.

KC

Kate Crawford is a principal researcher at Microsoft Research and a visiting professor at the MIT Center for Civic Media. Follow her on Twitter @katecrawford. This post is drawn from *a keynote* given at the Strata Conference in Santa Clara, Feb 28, 2013.