# Big data's 'streetlight effect'—where and how we look affects what we see

May 17 2016, by Mark Moritz, The Ohio State University

Don’t just look where the streetlight shines. Credit: darwinbell/flickr, CC BY

Big data offers us a window on the world. But large and easily available datasets may not show us the world we live in. For instance, epidemiological models of the recent Ebola epidemic in West Africa using big data consistently overestimated the risk of the disease's spread and underestimated the local initiatives that played a critical role in controlling the outbreak.

Researchers are rightly excited about the possibilities offered by the availability of enormous amounts of computerized data. But there's reason to stand back for a minute to consider what exactly this treasure trove of information really offers. Ethnographers like me use a cross-cultural approach when we collect our data because family, marriage and household mean different things in different contexts. This approach informs how I think about big data.

We've all heard the joke about the drunk who is asked why he is searching for his lost wallet under the streetlight, rather than where he thinks he dropped it. "Because the light is better here," he said.

This "streetlight effect" is the tendency of researchers to study what is easy to study. I use this story in my course on Research Design and Ethnographic Methods to explain why so much research on disparities in educational outcomes is done in classrooms and not in students' homes. Children are much easier to study at school than in their homes, even though many studies show that knowing what happens outside the classroom is important. Nevertheless, schools will continue to be the focus of most research because they generate big data and homes don't.

The streetlight effect is one factor that prevents big data studies from being useful in the real world – especially studies analyzing easily available user-generated data from the Internet. Researchers assume that this data offers a window into reality. It doesn't necessarily.

## Looking at WEIRDOs

Based on the number of tweets following Hurricane Sandy, for example, it might seem as if the storm hit Manhattan the hardest, not the New Jersey shore. Another example: the since-retired Google Flu Trends, which in 2013 tracked online searches relating to flu symptoms to predict doctor visits, but gave estimates twice as high as reports from the

Centers for Disease Control and Prevention. Without checking facts on the ground, researchers may fool themselves into thinking that their big data models accurately represent the world they aim to study.

The problem is similar to the "WEIRD" issue in many research studies. Harvard professor Joseph Henrich and colleagues have shown that findings based on research conducted with undergraduates at American universities – whom they describe as "some of the most psychologically unusual people on Earth" – apply only to that population and cannot be used to make any claims about other human populations, including other Americans. Unlike the typical research subject in psychology studies, they argue, most people in the world are not from Western, Educated, Industrialized, Rich and Democratic societies, i.e., WEIRD.

Twitter users are also atypical compared with the rest of humanity, giving rise to what our postdoctoral researcher Sarah Laborde has dubbed the "WEIRDO" problem of data analytics: most people are not Western, Educated, Industrialized, Rich, Democratic and Online.

## Context is critical

Understanding the differences between the vast majority of humanity and that small subset of people whose activities are captured in big data sets is critical to correct analysis of the data. Considering the context and meaning of data – not just the data itself – is a key feature of ethnographic research, argues Michael Agar, who has written extensively about how ethnographers come to understand the world.

What makes research ethnographic? It is not just the methods. It starts with fundamental assumptions about the world, the first and most important of which is that people see and experience the world in different ways, giving them different points of view. Second, these differences result from growing up and living in different social and

cultural contexts. This is why WEIRD people are not like any other people on Earth.

The task of the ethnographer, then, is to translate the point of view of the people they study into the point of view of their audience. Discovering other points of view requires ethnographers go through multiple rounds of data collection and analysis and incorporate concepts from the people they study in the development of their theoretical models. The results are models that are good representations of the world – something analyses of big data frequently struggle to achieve.

Here is an example from my own research with mobile pastoralists. When I tried to make a map of my study area in the Logone Floodplain of Cameroon, I assumed that places had boundaries, as the one separating Ohio from Michigan. Only later, after multiple interviews and observations, did I learn that it is better to think of places in the floodplain as points in an open system, like Columbus and Ann Arbor, without any boundary between them. Imagine that!

Don't get me wrong: I think big data is great. In our interdisciplinary research projects studying the ecology of infectious diseases and regime shifts in coupled human and natural systems, we are building our own big data sets. Of course, they are not as big as those generated by Twitter or Google users, but big enough that the analytical tools of complexity theory are useful to make sense of the data because the systems we study are more than the sum of their parts.

Moreover, we know what the data represents, how it was collected and what its limitations are. Understanding the context and meaning of the data allows us to check our findings against our knowledge of the world and validate our models. For example, we have collected data on livestock movements using a combination of surveys and GPS technology in Cameroon to build computer models and examine its

impact on the spread of foot-and-mouth disease. Because we know the pastoralists and the region in which they move, we can detect the errors and explain the patterns in the data.

For data analytics to be useful, it needs to be theory- or problem-driven, not simply driven by data that is easily available. It should be more like ethnographic research, with data analysts getting out of their labs and engaging with the world they aim to understand.

*This article was originally published on* The Conversation. *Read the* original article.

Source: The Conversation