Commentary

Big Data & Society
July–December 2015: 1–3
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053951715604334
bds.sagepub.com

Surfeit and surface

Monica Lee[1] **and John Levi Martin**[2]

Abstract

"Would you like another EXTRA BIG ASS FRIES?"
(Carl's Jr computer, *Idiocracy*)

Keywords
Big Data, surfaces, correlation, big-ass, cartography, social theory

Big Data, more than anything else, is...more. Lots of muchness or, as we might say in sociology, "large *N*s." Large *N*s are a challenge for us not because we have physical difficulties in handling the data, but because they come with a conceptual problem. While the problem is rooted in a statistical issue, this approach to statistics has so infected our minds, as we have built a conceptual world to justify what is technically tractable, that it has become a fundamental problem of our imagination.

Sociologists quickly gave up on the idea that they were going to find numerical laws of behavior, where the precise values of numbers really mattered (such with $F = gm_1m_2/d^2$, the gravitational force, where $g = -32\,\text{ft/s}^2$). As historians of science have shown, we re-thought the shocking imprecision of our results and decided to declare a victory not if we actually knew what our estimates were, but if we were reasonably sure that they weren't *zero*. The problem is, that as we get more and more data, the chance of *anything* being *precisely* zero goes to...precisely zero.

To put it simply, we are used to using tests of statistical significance to determine whether some effect is "significant." But with Big Data, or "big-ass data," as we prefer to call it, *everything* is significant because "*p* values" become disappearingly small. For this reason, as sociologists got their hands on more and more data sets with *N*s in the five-plus figures, there has been a slow defection to Bayesian statistics, but that has not solved the problems with our imagination. As is generally the case with innovations, we just use new tools to accomplish old tasks. In a word, we have been trying to *make things insignificant*. Like a neurologically impaired subject with dilated pupils, we are putting our hands over our eyes and hoping to peep through our fingers. But we believe that this is wrong. Because everything *is* significant...to someone.

We can call this imagination problem the "population problem." So far, we've assumed that the muchness of big-ass data (henceforward, BAD) comes from merely increasing the *size* of a sample. But we think that it goes further in that in many cases, big-ass data eats away at the very idea of the "population." Modern social statistics starts with the creative (if somewhat insane) idea that all members of a population are actually replications of a single underlying ideal type, the average man—just with a little bit of random error here and there that we can cleverly "correct for" with a wave of the statistical wand. But we are, in most cases, literally unable to distinguish *variation* from *error*, which means that we are, in effect, *forcing* people to hide their

[1] Facebook Inc., Menlo Park, CA, USA
[2] Department of Sociology, University of Chicago, Chicago, IL, USA

Corresponding author:
John Levi Martin, Department of Sociology, University of Chicago, 1126 E 59th St, Chicago, IL 60637, USA.
Email: jlmartin@uchicago.edu

heterogeneity, though we now tolerate a limited amount in, say, growth curve models. But if people are "doing their own thing," instead of doing like a specimen of the population, we don't want to hear about it. TMI.

Think, if you will, about the apparent obsession that we have with isolating the "average treatment effect," which has to be one of the least enlightening statistics available. If there is a treatment x whose effect on y is conditional upon type of person w (denote this conditional effect b_w), then the average treatment effect is $\int b_w dw$. There is nothing theoretically compelling about this average. Unless we are sure that there will be no change in the population composition of w, this is not even a good practical guide to the results of an intervention. To bring it home: it's thinking in terms of an "average treatment effect" that allows pundits to claim that the US economy is doing well while most workers get poorer (a mistake Adam Smith never made).

Out of sheer laziness, we are projecting a population, which induces a *surface* of effects in a hyper-dimensional space, onto a single point. We started in small N statistics with the liquidation of "outliers" (outsiders?) for excessive deviance. Only without their annoying refusal to take their place in a "normal" distribution could we theorize the population as a mean-plus-error. In so doing, we liquefied *everyone* into a homogenous soup despite how much they varied individually. But when you look carefully enough, we are *all* deviants. "It's not weird to be weird—in fact, it's absolutely normal," as Harrison White wrote.

That everyone is a package of idiosyncracies and incomparabilities need not undermine the possibility of large scale statistical analysis, just like each molecule might well have its unique characteristics without invalidating the gas law. In fact, with BAD, we have new tools to overcome the population problem. BAD analysts now expect a power law distribution for pretty much any cultural behavior, the way Queteletians once expected a normal distribution. That's because as we leave the worlds of biology and organizational regulation (call it "nature") for that of culture and independent action (call it "freedom") we are finding that people tend to do their own thing. And so we see many different power law distributions, with individuals who are at the left of one finding themselves on the right of others. Time spent bidding on *Lost in Space* memorabilia, looking at cat pictures on the internet, or going to psychobilly concerts, take on such a form because a very few people care a great deal about this, some people care a bit about it, and most don't care at all. Those weirdly attached to *Lost in Space* may be quite normally indifferent to cats in sinks.

Of course, we may want to go beyond simply graphing univariate distributions, and move towards relationships among variables, where the population problem is consequential. Suppose you want to know whether people are more likely to attend psychobilly concerts if they also consume lots of Pabst Blue Ribbon (PBR) beer. And we have data from 500,000,000 people to investigate that. Adhering to traditional techniques, we would log and correlate these variables. And we will almost certainly find a statistically significant but miniscule (say, $r = .0003$) effect. So yes, by normal standards, we have a finding (***!) that people who drink PBR are more likely to attend psychobilly concerts than people who drink other beverages.

But we should suspect that something's wrong. What does taking the average effect for this whole population even mean when most people in the sample neither attend psychobilly concerts nor drink PBR at all? The "average man" here certainly does not summarize the relationship between music show and beer choice for everyone in the sample; he's just a half-hearted attempt to find the lowest common denominator among a diverse set of people. And the larger a population is, the more likely it is to be too heterogeneous to characterize through that single average.

Now, this is not to say that BAD introduces us to the population problem—it's the same thing we encounter when we get a bi-modal distribution or see outliers. But BAD gives us a special opportunity to not *eliminate* these outliers, but to analyze them in their own right and to see how they relate to the rest of the population. That's because in BAD, a group of outliers, though a very small percentage of the total population, may still consist of thousands of people—many times more than the total respondents to the GSS. They should not be expunged as outliers but understood as a significant population in its own right. In so doing, we can analyze multiple trends in the same total population, moving from one average man that poorly represents one big population to multiple average men that represent segments of the total population more accurately—not analyze a given population, but discover populations inductively. Then we can finally begin searching for something a bit more like the gas law. Such lawfulness will not be found in derivative by-product statistics like the compositional artifacts of averages, but rather, in the topology of surfaces of effects. In other words, we think that the task for the next decade or so is shaking off the last commitments to causal explanation and shifting towards cartography—the construction of question-independent, though theoretically organized, reductions of information to make possible the answering of many questions.

Imagine arranging nearly everyone alive in a Blau Space—a multidimensional coordinate system, in which socio-demographic variables like age, income, education, race, and geographical location are treated as dimensions, and individuals who are close to each other are more similar to those further apart. Consider any variable corresponding to a taste, preference, or action *conditional* on position in this space to be a "surface." For example, one may represent the number of psychobilly concerts each person has attended. We can envision this number being represented as the intensity of a color (say, red), such that opacity is for those who attend concerts nightly, while the color is invisible for those who have never attended. A second (green) surface on top of the first would represent the number of cans of PBR people drink in a month. Say we find that people who go to concerts often go even more often as they increase their consumption of PBR—they go to socialize with their friends, so for them, cheap beer and psychobilly shows are highly correlated. Thus, opaque green is on top of opaque red. Perhaps then we find that listening to the Bazooka Baltimore Psychobilly radio station—blue—contributes to higher attendance for those who attend concerts only occasionally, but has little effect on those who attend concerts often. So opaque blue would be on top of translucent red, but there is no visible blue over opaque red.

Why the elaborate set up for what might seem a reproduction of the notion of correlation? It is because we do not necessarily want to collapse all areas with the same numerical values. Instead, we want to inspect the contours to understand the social logic of the distribution. The map allows us to scan our eyes up, down, left, and right, to draw both horizontal and vertical comparisons—how people in the population relate to each other in terms of demographics or any single surface (e.g. psychobilly concert attendance), as well as which factors contribute concert attendance for each subpopulation. We realize, for example, that there are several snake-like shapes of red moving through Blauspace, suggesting separable if not also independent "psychobillies" that are mapped onto the same action-space. We find a clumping of blue that suggests the importance of region. And we find, let us say, a set of widely spaced, horizontal planes of green, suggesting that there are different PBR drinkers organized by their position as the dominated fraction of the class they most identify with. Thus the population can be disaggregated and flexibly explored to answer a number of different questions instead of mean-averaged out to answer one poorly posed and unchanging question. Further, rather than *assuming* that all effects or relationships are global, we instead assume that effects are local until proven otherwise.

Choosing the variables to be included in any one map then emerges as the primary challenge. In BAD analysis, we often find ourselves in a world in which there is bountiful possibility but not always a natural stopping place. We may be brought towards a system level view, where we are examining open systems. It is rarely obvious from the start where to close the investigation—as Latour would say, which things to follow, and which not. Less and less often can we say with complacent regret, "we couldn't do that, because it's not in our data set." Like a real scientist, our problem isn't running out of information, but choosing which path to follow.

It is not that one needs such a jar to begin to think in terms of open systems of mutually interacting elements; theorists have been proposing this for years. It is a source of satisfaction to us that computational techniques now make this tractable for many cases, although sociological implementations are still crude and tend to focus on overly convenient cases. We may find ourselves in a position somewhat like meteorology—to determine the most feasible approximations to systems that we recognize as inherently open and overly complex, and to employ sets of models with known biases and blindspots. And while such models are not directly deducible from lower level theories, it is impossible to construct models that substantially outperform common sense without both huge quantities of accurate data and an understanding both of the representation of vector fields and their relation to fundamental dynamics of social interaction.

In sum, since Durkheim, in sociology, we have used convenient fictions—not the least of which is "society"—to justify otherwise bizarre conventions whereby we link aggregate data to claims about classes or persons. These fictions may have been methodologically necessary, but they are no longer, while they retain their falsity. Big-ass data allows us new ways of finding meaningful patterns in human experience, while we continue to pursue our fundamental theoretical interest in reciprocal patterns on alignment, conflict, competition, affiliation, and influence—precisely the theoretical approach of Durkheim's arch-enemy Tarde. Perhaps his day has dawned at last.

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

This article is part of a special theme on *Colloquium: Assumptions of Sociality*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/colloquium-assumptions-sociality.