IJoC | Big Data, Big Questions

The Theory/Data Thing

Commentary

GEOFFREY C. BOWKER
University of California at Irvine, USA

First we take Manhattan, then we take Berlin.
~ Leonard Cohen[1]

First we take Chris Anderson, then we take Latour . . .

The end of theory is being proclaimed on multiple fronts, and big data has a lot to do with it. Chris Anderson proclaims: Theory is dead, long live data! Away with every theory "of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology."[2] We can model the world and behavior well enough that we don't need to fit data into theory in order to create opportunities for more data gathering. The model's the thing. All science is subject to Anderson's new rules.

And these rules can be highly effective. In the sciences, this approach arguably works for much of climate science, which is less about why things occur than about whether we can accurately retrodict, portray, and predict (Edwards, 2010). For those of us brought up learning that correlation is not causation, there's a certain reluctance to examine the possibility that correlation is basically good enough. It is surely the case that we are moving from the knowledge/power nexus portrayed by Foucault to a data/action nexus that does not need to move through theory: All it needs is data together with preferred outcomes.

If science is about acting in the world, then there is no doubt much virtue to this position. It is Skinnerian psychology writ large—if all we care about is what goes in (stimulus) and what comes out (response), then to be effective we do not need to know what happens inside the mind/brain of the

---

[1] See http://www.sing365.com/music/lyric.nsf/First-We-Take-Manhattan-lyrics-Leonard-Cohen/926CCB64249F308848256AF00028CB85

[2] See http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Geoffrey C. Bowker: gbowker@uci.edu

individual. The death of Freud and the rise of neuropharmacology have engrained this within academia. *Data sunt potestas.* This leads to our intelligence being that of the ant colony, an arguably sad apotheosis. Ants act as if they are intelligent, in terms of organizing their colonies, farming fungi, and so forth, but they do not need to pass through ratiocination in order to achieve these goals. It is a stripped-down version of Teilhard de Chardin's numinous noosphere: global consciousness as glorified instinct rather than spiritual insight.

A strong virtue to correlationalism is that it avoids funneling our findings through vapid stereotypes. Thus, in molecular biology, most scientists do not believe in the categories of ethnicity (Reardon, 2001)—and are content to assign genetic clusters to diseases without passing through ethnicity (e.g., Karposi's sarcoma as initially a Jewish disease). Similarly, from the commercial end, many recommender systems work through correlation of purchases without passing through the vapid categories of the marketers—you don't need to know whether someone is male or female, queer or straight, you just need to know his or her patterns of purchases and find similar clusters.

But there is a series of problems with this movement, which we can start to adumbrate if we look to Bruno Latour. Latour (2002) argues for Gabriel Tarde *contra* Emile Durkheim. The latter reified society and explained constant correlations (e.g., suicide rates) as social facts. Social conditions cause social effects. The Tardean position, for Latour, involves replacing statistics (etymologically, facts about the State) with aggregating clusters on the fly through large-scale data analysis. There is no need to go "outside" of events for their explanation—we do not need to assume that there are categories like society, class, ethnicity, and so forth: Everything depends on describing a specific correlation at a specific time. Thus for Latour, as for the molecular biologists and the marketers, there is no need to appeal to analytic categories in order to study and write about events. (I am deliberately not using "understand," since understanding is precisely what is at stake.)

Latour here is retrojecting onto Tarde his own prior views that actor-network theory is not a theory but a way of flattening all categories and replacing theory with method. His is the *nec plus ultra of* Margaret Thatcher's infamous proclamation: "And, you know, there is no such thing as society. There are individual men and women, and there are families."[3] Latour would just add in that there aren't families or individuals either (the latter being the more interesting ontological point).

So a two-part question—do we need theories, and do theories need categories? In *The Fragile Absolute: Or, Why Is the Christian Legacy Worth Fighting For?* Žižek (2009) provides one way in to these questions. Take the social dimension first. If we accept the underlying ontology that we are all individuals (atoms) who aggregate in unnamed clusters rather than categories, then Žižek argues that we certainly lose the ability to recognize constant and meaningful forces in "society" (which I'll put in scare quotes for the nonce). It does not just happen that there is a net protein, natural resource drain from the Third World to the First, nor that women in the United States are consistently paid less for the same quality of work as men. These categories represent a reality. Certainly, they should not be essentialized. The Third World/First World divide overlooks regions of intense underdevelopment in, say, the United States and

---

[3] See http://briandeer.com/social/thatcher-society.htm

regions of vast wealth in, say, India. Similarly, "woman" is a category that can and should be questioned. And yet . . . the rough, aggregate truth is that there is not a level playing field for either, broadly construed. No data deluge will explain these truths—at best, it can help direct policies to mitigate the injustice; at worst (and most commonly), it can deny that there are indeed broad social forces. Willy-nilly, our social world is one in which categories have deep meaning. This is not just about the social truths: The same can be argued for truths in the natural sciences. A category system like the species concept is indeed highly problematic (Wilkins, 2011); however, the aggregate behavior of most entities can be described along certain dimensions as if this categorization were real. In both cases, the world is structured in such a way as to make the categories have real consequences.

So in some ways, categories are central to being in the world. Big data does not do away with categories at all. As I have argued elsewhere, the term "raw data" is itself an oxymoron. Antonia Walford (2012) writes about the work it takes to turn data from sensors in the Amazon rain forest into manipulable data within databases. There is a plenum of data: For her, the art of the scientific database is to take this undifferentiated onslaught and conjure it into models (structured data fields, metadata) that allow Amazon data to circulate scientifically. As Derrida (1998) argues in *Archive Fever* and Cory Knobel (2010) so beautifully develops with his concept of ontic occlusion, every act of admitting data into the archive is simultaneously an act of occluding other ways of being, other realities. The archive cannot in principle contain the world in small; its very finitude means that most slices of reality are not represented. The question for theory is what the forms of exclusion are and how we can generalize about them. Take the other Amazon as an illustration. If I am defined by my clicks and purchases and so forth, I get represented largely as a person with no qualities other than "consumer with tastes." However, creating a system that locks me into my tastes reduces me significantly. Individuals are not stable categories—things and people are not identical with themselves over time. (This is argued in formal logic in the discipline of mereology and in psychiatry by, say, ethnopsychiatry.) The unexamined term the "individual" is what structures the database and significantly excludes temporality.

Two things, then. Just because we have big data does not mean that the world acts as if there are no categories. And just because we have big (or very big, or massive) data does not mean that our databases are not theoretically structured in ways that enable certain perspectives and disable others.

There is, however, also the overarching problem with both Anderson and Latour. Sure, with the above caveats, I can imagine living in a world where science and social science are about manipulating the world—effective action is after all a good thing. However, this is a massive reduction of what it means to "know." I have already witnessed in the unhallowed halls of the National Science Foundation a line of argument that says we don't really need ethnography any more. After all, ethnographers just reason from an *n* of, say, 20, where other methods deploy an *n* of 200,000. In John King's immortal words, "numbers beats no numbers every time." The hyping of big data leads to the withering away of interpretation—not through the actions of a cabal, but through a sociologic of excluding from the archive all data which is not big. This unconsidered exclusion is occurring in small across the sciences ("first they came for ethnography, but I did not speak out because I was not an ethnographer. . ." and so forth). It demands a systematic response.

The theory/data thing is very much about "things," in the sense in which Pelle Ehn uses the term—for him, a designed object (a thing) contains within it a host of contradictory discourses, never finally resolved—as in the Icelandic Thing (the original parliament) (Binder, T., De Michelis, G., Ehn, P., & Jacucci, G., 2011). Any "thing" that we create (object, way of looking at the world) irreducibly embodies theory and data. And that is a good thing.

References

Binder, T., De Michelis, G., Ehn, P., & Jacucci, G. (2011). *Design things (design thinking, design theory).* Cambridge, MA: MIT Press.

Derrida, J. (1998). *Archive fever: A Freudian impression.* Chicago, IL: University of Chicago Press.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming.* Cambridge, MA: MIT Press.

Knobel, C. (2010). Ontic occlusion and exposure in sociotechnical systems (Doctoral dissertation, University of Michigan). Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/78763 (UMI Number AAI3441199)

Latour, B. (2002). Gabriel Tarde and the end of the social. In P. Joyce (Ed.), *The social in question: New bearings in history and the social sciences* (pp. 117–132). London, UK, Routledge.

Reardon, J. (2001). The human genome diversity project: A case study in coproduction. *Social Studies of Science 31*, 357–388.

Walford, A. (2012). Data moves: Taking Amazonian climate science seriously. *Cambridge Anthropology, 30* (2), 101–117.

Wilkins, J. S. (2011). *Species: A history of the idea.* Berkeley: University of California Press.

Žižek, S. (2009). The fragile absolute: Or, why is the Christian legacy worth fighting for? London, UK: Verso.