Search: Go

blog.castac.org

From the Committee on the Anthropology of Science, Technology, and Computing (CASTAC)

Dealing with Big Data: David Hakken Weighs In

January 15th, 2013, by Patricia G. Lange Comments Off

Although anthropologists have been working with large-scale data sets for quite some time, the term "big data" is currently being used to refer to large, complex sets of data combined from different sources and media that are difficult to wrangle using standard coding schemes or desktop database software. Last year saw a rise in STS approaches that try to grapple with questions of scale in research, and the trend toward data accumulation seems to be continuing unabated. According to IBM, we generate 2.5 quintillion bytes of data each day. This means that 90% of the data in the world was created during the last 2 years.

Big data are often drawn and aggregated from a very large variety of sources, both personal and public, and include everything from social media participation to surveillance footage to consumer buying patterns. Big data sets exhibit complex relationships and yield information to entities who may mine highly personal information in a variety of unpredictable and even potentially violative ways.

The rise of such data sets yields many questions for anthropologists and other researchers interested both in using such data and investigating the techno-cultural implications and ethics of how such data is collected, disseminated, and used by unknown others for public and private purposes. Researchers have called this phenomenon the "politics of the algorithm," and have called for ways to collect and share big data sets as well as to discuss the implications of their existence.

I asked David Hakken to respond to this issue by answering questions about the direction that big data and associated research frameworks are headed. David is currently directing a Social Informatics (SI) Group in the School of Informatics and Computing (SoIC) at Indiana University Bloomington. Explicitly oriented to the field of Science, Technology, and Society studies, David and his group are developing a notion of social robustness, which calls for developers and designers to take responsibility for the creation and implications of techno-cultural objects, devices, software, and

systems. The CASTAC Blog is interested in providing a forum to exchange ideas on the subject of Big Data, in an era in which it seems impossible to return to data innocence.

Patricia: How do you define "big data"?

David: I would add three, essentially epistemological, points to your discussion above. The first is to make explicit how "Big Data" are intimately associated with computing; indeed, the notion that they are a separate species of data is connected to the idea that they are generated more or less "automatically," as traces normally a part of mediation by computing. Such data are "big" in the sense that they are generated at a much higher rate than are those large-scale, purpose-collected sets that you refer to initially.

The second point is the existence of a parallel phenomenon, "Data Science," which is a term used in computing circles to refer to a preferred response to "Big Data." Just as we have had large data sets before Big Data, so we have had formal procedures for dealing with any data. The new claim is that Big Data has such unique properties that it demands its own new Data Science. Also part of the claim is that new procedures, interestingly often referred to as "data mining." will be the ones characteristic of Data Science. (What are interesting to me are the rank empiricist implications of "data mining.") Every computing school of which I know is in the process of figuring out how to deal with/"capitalize" on the Data Science opportunity.

The third point is the frequently-made claim that the two together, Big Data and Data Science, provide unique opportunities to study human behavior. Such claims become more than annoying for me when it is asserted that the Big Data/Data Sciences uniquenesses are such that those pursuing them need not pay any attention to any previous attempt to understand human behavior, that only they and they alone are capable of placing the study of human behavior on truly "scientific" footing, again because of their unique scale.

Patricia: Do you think that anthropologists and other researchers should use big data, for instance, using large-scale, global information mined from Twitter or Facebook? Do you view this as "covert research"?

David: We should have the same basic concern about these as we would any other sources of data: Were they gathered with the informed consent of those whose activities created the traces in the first place? Many of the social media sites, game hosts, etc., include permission to gather data as one of their terms of service, to which users agree when they access the site. This situation makes it hard to argue that collection of such data are "covert." Of course, when such agreement has not been given, any gathered data in my view should not be used.

In the experience of my colleagues, the research problem is not so much the ethical one to which you refer so much as its opposite—that the commercial holders of the Big Data will not allow independent researchers access to it. This situation has led some colleagues to "creative" approaches to gathering big data that have caused some serious problems for my University's Institutional Review Board.

In sum, I would say that there are ethical issues here that I don't feel I understand well enough to take a firm position. I would in any particular case begin with whether it makes any sense to use these data to answer the research questions being asked.

Patricia: Who "owns" big data, and how can its owners be held accountable for its integrity and ethical use?

David: I would say that the working assumption of the researchers with whom I am familiar is either the business whose software gathers the traces or the researcher who is able to get users to use their data gathering tool, rather than the users themselves. I take it as a fair point that such data are different from, say, the personal demographic or credit card data that are arguably owned by the individual with whom they are associated. The dangers of selling or similar commercial use of these latter data are legion and clear; of the former, less clear to me, mostly because I don't know enough about them.

Patricia: What new insights are yielded by the ability to collect and manipulate multi-terrabyte data sets?

David: This is where I am most skeptical. I can see how data on the moves typically made by players in a massive, multiplayer, online game (MMOG) like World of Warcraft ™ would be of interest to an organization that wants to make money building games, and I can see how an argument could be made that analysis of such data could lead to better games and thus be arguably in the interest of the gamers. When it comes to broader implications, say about typical human behavior in general, however, what can be inferred is much more difficult to say. There remain serious sampling issues however big the data set, since the behaviors whose traces are gathered are in no sense that I can see likely to be randomly representative of the population at large. Equally important is a point made repeatedly by my colleague John Paolillo, that the traces gathered are very difficult to use directly in any meaningful sense; that they have to be substantially "cleaned," and that the principles of such cleaning are difficult to articulate. Paolillo works on Open Source games, where issues of ownership are less salient that they would be in the proprietary games and other software of more general interest.

Equally important: These behavioral traces are generated by activities executed in response to particular stimulations designed into the software. Such stimuli are most likely not typical of those to which humans respond; this is the essence of a technology. How they can be used to make inferences about human behavior in general is beyond my ken.

Let me illustrate in terms of some of my current research on MMOGs. Via game play ethnography, my co-authors (Shad Gross, Nic True) and I arrived at a tripartite basic typology of game moves: those essentially compelled by the physics engine which rendered the game space/time, those responsive to the specific features designed into the game by its developers, and those likely to be based on some analogy with "real life" imported by the player into the game. As the first two are clearly not "normal," while the third is, we argue that games could be ranked in terms of the ratio between the third and the first two, such ratio constituting an initial indicator of the extent of familiarity with "real life" that could conceivably be inferred from game behavior. Perhaps more important, the kinds of traces to be gathered from play could be changed to help make measures like this easier to develop.

Patricia: What are the epistemological ramifications of big data? Does its existence change what we mean by "knowledge" about behavior and experience in the social sciences?

David: I have already had a stab at the first question. To be explicit about the second: I don't think so.

There are no fundamental knowledge alterations regarding those computer mediations of common human activity, and we don't know what kind of knowledge is contained in manipulations of data traces generated in response to abnormal, technology-mediated stimuli.

Patricia: boyd and Crawford (2011) argue that asymmetrical access to data creates a new digital divide. What happens when researchers employed for Facebook or Google obtain access to data that is not available to researchers worldwide?

David: I find their argument technically correct, but, as above, I'm not sure how important its implications are. I am reminded of a to-remain-unnamed NSF program officer who once pointed out to a panel on which I served that NSF was unlikely to be asked to fund the really cutting edge research, as this was likely to be done as a closely guarded, corporate secrete.

Patricia: What new skills will researchers need to collect, parse, and analyze big data?

David: This is interesting. When TAing the PhD data analysis course way back in the 1970s, I argued that to take random strolls through data sets in hopes of stumbling on a statistically significant correlation was bad practice, yet this is in my understanding, the approach in "data mining." We argue in our game research that ethnography can be used to identify the kinds of questions worth asking and thus give a focus, even foster hypothesis testing, as an alternative to such rampant empiricism. Only when such questions are taken seriously will it be possible to articulate what new skills of data analysis are likely to be needed.

Patricia: How can researchers insure data integrity across such mind-boggling large and diverse sets of information?

David: Difficult question if dealing with proprietary software; as with election software, "trust me" is not enough. This is why I have where possible encouraged study of Open Source Projects, like that of Giacomo Poderi in Trento, Italy. Here, at least, the goals of designers and researchers should be aligned.

Patricia: To some extent, anthropologists and other qualitative researchers have always struggled to have their findings respected among colleagues who work with quantitative samples of large-scale data sets. Qualitative approaches seem especially under fire in an era of Big Data. As we move forward, what is/will be the role and importance of qualitative studies in these areas?

David: As I suggested above, in my experience, much of the Data Science research is epistemologically blind. Ethnography can be used to give it some sight. By and large, however, my Data Science colleagues have not found it necessary to respond positively to my offers of collaboration, nor do I think it likely that either their research communities of funders like the NSF, a big pusher for Data Science, will push them toward collaboration with us any time soon.

Patricia: What does the future hold for dealing with "big data," and where do we go from here?

David: I think we keep asking our questions and turn to Big Data when we can find reason to think that they can help us answer them. I see no reason to jump on the BD/DS bandwagon any time soon.

On behalf of The CASTAC Blog, please join me in thanking David Hakken for contributing his insights into a challenging new area of social science research!

Patricia G. Lange
Editor-in-Chief
The CASTAC Blog

Tagged: big data, ethics, methods

Comments are closed.

« Call for Papers: "Big Data, Big Questions, or, Accounting for Big Data" [Abstracts DUE October 1, 2012]

Looking Ahead to 2013: A Question of Scale »

There remain serious sampling issues however big the data set, since the behaviors whose traces are gathered are in no sense that I can see likely to be randomly representative of the population at large.

What's this?

You are currently reading **Dealing with Big Data: David Hakken Weighs In** at blog.castac.org.

meta

- **Author:** Patricia G. Lange
- **Comments:** Comments Off
- **Categories:** Research, Tools & Techniques