Routledge
Taylor & Francis Group

Information, Communication & Society

ISSN: 1369-118X (Print) 1468-4462 (Online) Journal homepage: http://www.tandfonline.com/loi/rics20

Bigger sociological imaginations: framing big social data theory and methods

Alexander Halavais

To cite this article: Alexander Halavais (2015) Bigger sociological imaginations: framing big social data theory and methods, Information, Communication & Society, 18:5, 583-594, DOI: 10.1080/1369118X.2015.1008543

To link to this article: http://dx.doi.org/10.1080/1369118X.2015.1008543

Information, Communication & Society, 2015
Vol. 18, No. 5, 583–594, http://dx.doi.org/10.1080/1369118X.2015.1008543

Routledge
Taylor & Francis Group

Bigger sociological imaginations: framing big social data theory and methods

Alexander Halavais*

School of Social & Behavioral Sciences, Arizona State University, P.O. Box 37100, Phoenix, AZ 85069-7100, USA

(*Received 4 November 2014; accepted 13 January 2015*)

Making effective use of big social data requires us to frame that work in useful ways, ways that draw connections between new methods and a long history of social methods and theories. In particular, the key questions of big social data – those of relating observations of features at scale to practical outcomes for individuals and groups – are core sociological questions. We need to develop a new, bigger sociological imagination that allows us to incorporate big social data rather than reinventing the wheel. That requires careful mining of our methodological and theoretical history, along with a reexamination of the ways in which we collect and use our data.

Keywords: big data; data science; social theory; sociology

Over the last several years, the term 'big data' has found its way into the discourse of a number of fields, including the social sciences. There seem to be more people with opinions about big data than there are studies utilizing large social data sets. Those opinions, at least those that agree that big data is anything other than hyperbolic marketing, see the advent of methods of collecting and analyzing large sets of trace observations as marked by promise or peril, and often by both at once. And many, wisely, approach the question of the degree to which big social data marks a departure from traditional social science with some trepidation: prognostication is a perilous art.

Nonetheless, the real danger is allowing some combination of availability, methods, marketing, and scholarly fashion to bend and shape social research rather than being guided by a deeper sense of inquiry. There are significant ethical issues surrounding the massive collection and analysis of social data, but perhaps more dangerous is the possibility that we undertake or avoid big social studies without a broader frame of engagement. What follows is an initial attempt to define the boundaries of big social data, to navigate the place of theory, to trace a deeper history of big data within the social sciences and map current debates to a broader agenda, to place some of the ethical challenges we face within this agenda, and finally, to suggest some ways in which this shift might affect the role of the social scientist and how new social scientists are trained.

Big data does provide a challenge to the social sciences, but not a particularly new one. It is, in fact, the core challenge of sociology: connecting the micro-connections between individuals to

*Email: theprof@asu.edu

the vast social structures that shape us (and are shaped by us) as a society. Mills (2000) famously suggested that this ability to both connect and disconnect the personal with the social was at the core of what he called the 'sociological imagination'. To incorporate new sources of large-scale traces and ways of mapping social flows and transactions requires a larger sociological imagination, but it also informs the struggle that has always been at the core of social inquiry: how it is that human relationships relate to social structure.

Big social data

We encounter the idea of big data at a particular historical moment and it appears to mean different things to different audiences. The term and cognates (including 'data science' and 'computational social science') appear in a range of popular and scholarly contexts, yielding special issues of journals, new funding lines, new degree programs, and job titles. At one extreme end of this, there is a suggestion that the social scientist is properly relegated to the dustbin of history, superseded by the data scientist, who, unhindered by theoretical baggage, is able to finally perfect the ideal of 'social physics' and discover truth in massive collections of trace data that map out human relations. At the same time, we see criticisms of this new thrust, many of them arguing that we miss opportunities for observation or for theorization by focusing on big data, others suggesting that an emphasis on big data places the social scientist in the service of the technologies, platforms, institutions, and economic structures that produce, collect, and concentrate this information. This has yielded a growing counter-vocabulary: 'small data' (boyd & Crawford, 2012), 'long data' (Arbesman, 2013), 'thick data' (Boellstorff, 2013), and 'slow data' (Barns, 2014), among others.

Although there are some who suggest that we have moved beyond the hype surrounding big data, few could ignore the degree to which the term has been used as a faddish buzzword, and sometimes little more. Google's search trends suggest that searches for the term began growing exponentially from the early part of 2011. By 2012, *Harvard Business Review* crowned data science as 'the sexiest job of the twenty-first century' (Davenport & Patil, 2012). Some of the more recent headlines suggest a slightly more critical stance than in the past (*Forbes*: 'Taxi Stockholm shows us that big data needs a big idea'; *Slate*: 'The big data paradox'; *New York Times*: 'Is big data spreading inequality?'). Few of the laudatory or critical articles either in the popular or scholarly press offer a consistent definition of the term. 'Big Data', like 'Web 2.0' and 'social media', has come to represent an amorphous set of practices and technologies that are only very loosely related.

Perhaps the most widespread use of the term is by companies that provide hardware and software solutions to address the 'problem' of big data. Ward and Barker (2013) note that the term has been used in a wide range of fields and this 'shared provenance has led to multiple, ambiguous and often contradictory definitions', but they go on to survey definitions offered mainly by information technology companies – Oracle, Intel, Microsoft, and IBM – all of which have used the term to suggest unmet needs. The Gartner Group's adopted definition of big data as consisting of 'three Vs' – Volume, Velocity, and Variety – is often cited (Laney, 2001), and sometimes IBM's added V, 'Veracity', is included. Unfortunately, rather than clarifying the issue, this provides further dimensions of confusion. The implicit fifth 'V' is Vexatious vagueness.

Ward and Barker (2013) draw on definitions from standards-producing organizations as well to suggest that all of these seem to touch on size, complexity, and technology as defining elements, but only very rarely quantify or specify how big, fast, or unstructured is enough. Big data is a moving target, and is often defined by failure: it is any collection of data that exceeds our ability to effectively collect, manage, transmit, and analyze it (Jacobs, 2009). Such a definition moves big data into the realm of the ever unknowable, which may be useful for encouraging sales,

but not particularly helpful in scholarly pursuits. Given this, it is tempting to dismiss the term as marketing hyperbole and merely avoid its use.

There are two reasons that big data remains worthy of interest. The first is that it has become entrenched in the formal institutions that support scholarly work in the social sciences. A surprising array of public funding agencies and private foundations has taken up the banner of big social data. Dozens of universities have promoted programs or tracks in 'big data' or 'data science'. In 2014, we saw the launch of a new scholarly journal entitled *Big Data & Society*, which joins slightly older scholarly outlets like the *Journal of Big Data*, *Big Data Research*, the *Journal of Data Science*, and – perhaps most succinctly – *Big Data*. One might expect that this is in response to a groundswell of research and theory in the area, but if these institutional changes are the result of a latent community, it is as invisible a college as you could hope to imagine. Nonetheless, the rapid rise of material and institutional support for data-oriented social research would alone spur continued interest.

But second, and more importantly, to ignore big data is to ignore a set of questions that are core to the research of society, and the ongoing evolution of social methods and theories. The greatest promise of big data is the opportunity to connect the very large scale of social interactions with the microinteractions of everyday relationships. Rather than an analytical definition, then, we might draw on some of the existing work on big social data as indicative of a kind of definition through example.

Many of these studies draw on the microblogging platform Twitter (boyd & Crawford, 2012). There are a number of reasons for this, not least that it is seen as a largely public space for interaction, and that despite some technical difficulties in acquiring large collections of tweets, it remains relatively open to researchers. Although the tweets themselves are small, relatively large collections represent a challenge. The most-cited articles draw on large Twitter collections to provide descriptive content analysis (Krishnamurthy, Gill, & Arlitt, 2008; Kwak, Lee, Park, & Moon, 2010; Vieweg, Hughes, Starbird, & Palen, 2010), sentiment analysis (Bollen, Mao, & Zeng, 2011; Jansen, Zhang, Sobel, & Chowdury, 2009; Pak & Paroubek, 2010; Tumasjan, Sprenger, Sandner, & Welpe, 2010), analyze influence (Bakshy, Hofman, Mason, & Watts, 2011; Cha, Haddadi, Benevenuto, & Gummadi, 2010), or detect communities and relationships (boyd, Golder, & Lotan, 2010; Huberman, Romero, & Wu, 2009; Java, Song, Finin, & Tseng, 2007), among other topics. Although we could consider all the microcontent of Twitter itself as the 'big data', with large numbers of users and tweets accumulating rapidly over time, these studies have generally drawn on subsets of the data available. Of those cited above, the largest corpus studied, Bakshy et al., consisted of over a billion individual tweets, and most of the studies used collections of over a million tweets. Although the United States Library of Congress holds an archive of over 170 billion tweets, difficulties of access have thus far left this resource untapped.

A number of event-driven studies of Twitter (e.g. Earl, McKee Hurwitz, Mejia Mesinas, Tolan, & Arlotti, 2013; Gaffney, 2010; Halavais & Garrido, 2014; Hughes & Palen, 2009; Poell & Borra, 2011) use relatively more modest collections of tens or hundreds of thousands of tweets. Other constrained approaches, like attempts to capture regional twitterspheres (e.g. Bruns, 2014), likewise might collect very large subsets of the totality of tweets. Are there a particular number of items that leads to data being considered 'big'? In these cases, there was an effort to capture the universe of relevant tweets. While it seems in some sense that analyzing any data that are available makes for more informative or compelling research, we already know that beyond a certain level, sample sizes in inferential statistics are irrelevant. If the aim is to measure central tendency, samples of a billion provide little benefit.

Astronomical data, medical imaging, and large-scale physics measurements can also lead to enormous data sets. These often seek out either small anomalies in large collections, or

relationships that are only revealed at scale. The same is true of social research – it is difficult to sample when you are seeking out anomalies: the needle in the haystack. Likewise, it is often difficult to sample networks: the more you know about every node and the relationships between these, the better measure of the totality can be reached. The size of the collected data should match the intended analysis.

While Twitter may in some ways be emblematic of big social science, there are, of course, other sources of big data of interest to the social scientist. Many of these remain in the realm of social media, and each platform – or a combination of platforms – provides an opportunity to capture traces of social interaction. Moreover, as more social processes that have traditionally occurred offline leave online traces (Lazer et al., 2009), there arise new opportunities to obtain and make use of large-scale, non-reactive data. Collecting these trace data may lead to surprisingly accurate indications of aggregate phenomena. The most widely noted example of this is the use of search engine data to indicate early trends (Goel, Hofman, Lahaie, Pennock, & Watts, 2010). Google Flu Trends predicted influenza outbreaks over time based on particular query terms on the Google search engine. This approach was, at least until recently, as effective at predicting influenza outbreaks as more costly traditional public health monitoring.

In practice, then, 'big social data' may have as much to do with how the data is recorded, transferred, and manipulated as it does with its size. Because storage is inexpensive and grows ever more inexpensive, accumulations of everyday interactions are kept as part of an ongoing transactional record. These unobtrusive traces of the lives of large numbers of connected people represent a source of data that is, if not exactly new, less commonly utilized than the usual toolkit of qualitative and quantitative methods in the social sciences today. That these data can grow to be large is less important than the ways in which they can (and cannot) be obtained and analyzed.

The death and resurrection of theory

The rise of big data in the natural sciences has led to suggestions that the hypothetico-deductive model of science is over, and with it, the need for theory. This idea was distilled to its most extreme and popular form in an article that appeared in *Wired Magazine*, in which Anderson (2008) suggests that 'faced with massive data, this approach to science – hypothesize, model, test – is becoming obsolete'. Particularly for those for whom 'theory' is seen as the opposite of 'practice', this death knell was celebrated. The idea that data might speak to us unassisted has deep roots. In 1962, an article by Orrin Clotworthy appeared in *Studies in Intelligence*, an in-house publication for the Central Intelligence Agency. In it, he suggests that the future of intelligence will be found in collecting data from a wide variety of disparate sources (e.g. 'the size of the next coffee crop, bullfight attendance figures, local newspaper coverage of UN matters'), and drawing out correlations that can predict future behavior:

> To learn just what the factors are, how to measure them, how to weight them, and how to keep them flowing into a computing center for continual analysis will someday become a matter of great concern to all of us in the intelligence community.

Of course, in some ways, this prediction was already present in Asimov's (1951) fictional 'psychohistory' in the *Foundation* series, an approach to collecting data and mathematically modeling (and shaping) the evolution of society. With enough data, and enough dimensions, answers can be found without questions being necessary, the reasoning goes, and without the need for questions, theory is superfluous. This view is wrong in a number of ways.

First, although theory provides an important heuristic function, identifying testable hypotheses, as well as the ability to predict outcomes, these are hardly its only functions. More important than both is the ability to explain social structure and change. To the patient who is in need of life-saving medication, the fact that particular compounds can alleviate the symptoms is of primary importance. Likewise, an approximated solution to a logistics problem for a retail chain can save money. But from early on, the practice of the social sciences has aimed to understand the ways in which social structures emerge and change and to ground observations of apparent relationships to the causal network that explains them (Durkheim, 1982, p. 94).

As noted earlier, many point to Google Flu Trends as an exemplar of the promise of big data. Relationships in the data – between elements, over time, or in overall structure – can provide answers to questions you did not know you had. In 2013, the prediction was flawed, and since the cause of the correlation was never fully understood, the cause of the failure was likewise opaque. At least some have suggested that the failure of the Google Flu Trends predictions in 2013 should be seen as an indictment of 'big data hubris' more broadly (Lazer, Kennedy, King, & Vespignani, 2014), and not just a failure of analysis in one case. Not only is there a danger that confounding variables remain unidentified in anything other than a perfect collection of factors, spurious relationships are almost guaranteed to be found in any large enough collection of data. Some reasoning around relationships, some theoretical interpretation, is a practical necessity.

But even if we were able to accurately create models that predicted the spread of infectious diseases, the popularity of a candidate, or future commodities prices, such an understanding would be only partially satisfactory without fully understanding the reasons for those changes. Models are important, but represent an intermediate step in the process of social science. Part of what makes the social sciences sciences is a desire to explain social change and structure – indeed, such explanations are important even in the very likely event that they fail to predict future social change.

Finally, any claim to atheoreticality is difficult to support. Instead, those relying entirely on correlation and mapping are engaging in a naive form of reasoning; theory by assumption. As Henri Poincaré famously noted, the scientist is charged with creating order: a collection of facts is no more a science than a heap of stones is a house. While large-scale mapping of variables can act as an inductive tool for discovery, it represents one piece of a process that leads to understanding – to social theory. Methods of big data are not an end unto themselves, but a process for arriving at explanatory theory; fortunately, sociology has some experience with this process.

One of the central questions of social theory is how society shapes, and is shaped by, individual actions. Or, to put it another way, how is it that the aggregation of the infinite microinteractions we engage in each day and throughout our lives, of various kinds and with different sets of people and artifacts, leads at larger scales to rules, expectations, values, desires, and structures that do not seem to be easily observed at the individual or group level? This question is, fundamentally, a question of big social data, of understanding how the dynamics of large-scale structure evolve and are related to our everyday existence.

Big social data requires us to think about how the abstract is related to the particular, and recognize that this relationship is complex, tenuous, and difficult to discern. It aims to ground grand social theory in everyday experience. While, say, structuration provides a way of understanding individual agency in the face of (or in support of) broader social rules, it is difficult to ground this in empirical work. Any attempt to examine the particular, often through ethnographic methods, risks only a cursory view of the general, social fabric. And attempts to understand the overarching structures of society, while they may provide some internally consistent explanation of social change or control, are often too abstract to be of practical importance or usefulness.

Mills (2000) warns against these twin traps: on the one hand, grand theory, which is necessarily abstract, and when analyzed, yields only further abstractions; on the other, what he calls the pitfall of abstract empiricism, where the 'content swallows the idea' (p. 124). But as Manovich (2012) suggests, big social data present the possibility of empirical observation of interactions at a microlevel all the way up to the largest accumulations, of collections that are both broad and deep. Of course, there is the danger that such observations are large in scale, but lacking in context. But the more dangerous outcome is being swallowed by big data and assuming that it can do the job of the researcher.

Big data before it was cool

If big data is defined by the 'petabyte age' of inexpensive digital storage or by the rise of social media as an object of study, then it is perhaps fair to suggest that it is a kind of a revolution. But a closer examination finds more continuities with a long history of social inquiry than it does disconnect. Indeed, the roots of sociology are found in explorations of big data, and attempts to make sense of how to connect phenomena observed at scale and individually. The question at the core of early social theories remains as pressing today: how is individual action related to social structure?

This was the question embedded in Emile Durkheim's work on suicide (2006), and despite the lack of computing power, it would be difficult to miss the place of what we might call big data in his work. Indeed, if it is the case that Durkheim chose the topic of suicide because statistics were available, his work might also demonstrate one of the pitfalls of big data – allowing the available data to choose the research topic. Nonetheless, his work required him to reuse trace data collected for a different purpose to help support not just an ecological theory of suicide, but the larger proposal that society represents a separate field of inquiry. Finding, assembling, standardizing, and analyzing the collected statistics bared more than a passing resemblance to what is now considered 'big social data', and certainly appeared different from other methods of studying collective behavior at the time.

And from these initial steps into sociology, the need to arrive at new methods of collection and analysis was acute. Many of these remain important to us today, and are finding new applications as more data is available. This includes the continued development of statistical techniques that can be applied to complex social data. The development of multilevel analysis (Goldstein & McDonald, 1988), for example, was needed to address issues of grouping and hierarchy in aggregated data, and continues to be part of a set of tools used in interpreting large-scale data. And, of course, social scientists have long made use of statistical computing packages to analyze large data sets even before those data were 'born digital' and the job title of 'data scientist' existed.

Likewise, despite the seeming recent rediscovery of social network analysis, 80 years ago, Moreno (1934) was using sociographs to describe networked social relationships, and Simmel (2010) was thinking about social networks decades earlier than that. That work set in motion decades of applying networked approaches to social systems, culminating in recent work on the 'new' network sciences by scholars like Granovetter (1973), Wellman et al. (1996), and Watts & Strogatz (1998). Social scientists also drew from the various traditions of cybernetics, including new forms of systems approaches (Luhmann, 1989), complex adaptive systems (Miller & Page, 2009), and agent-based modeling (Epstein, 2006). Particularly in the latter case, the work has tended to be more theoretical than empirical, but now, some of the models that might have been difficult to test because of lack of empirical data can draw on new streams of digital information to inform these social simulations. Among the earliest scholars of online community and interaction were those who studied communication and media, and drew together data from online social systems (Hiltz & Turoff, 1993) and later the Internet and

mobile networks. The idea that communications can help to map social structure (e.g. Deutsch, 1963) likewise provides a decades-long record of approaches to understanding big social data (González-Bailón, 2013).

Despite this long history of thinking about tools that can be used to understand large social systems, often, when those interested in big data from the perspective of computing technologies or algorithms take on questions of social dynamics, social theory appears as a quick citation, if at all, and methodological questions are likewise given short shrift. Those who are more familiar with computing than with social science can be forgiven for thinking that they have encountered a whole new world of questions and new tools for answering them. This is all the more true when social scientists fail to take on some of the technical challenges of acquiring and manipulating big social data. This results in more than a missed opportunity; the lack of voices from sociologists and others with an understanding and interest in social theory is an abrogation of responsibility. If social scientists wish to see these new sources of data used appropriately, they need to demonstrate how to do so.

Data ethics

It is at the human scale where most sociologists work, and perhaps because of that are drawn to qualitative and ethnographic approaches. This human-scale interaction comes with its own ethical challenges, but makes it easier to remember that our subjects are people. There are at least two significant ethical challenges for social scientists who wish to engage with sources of big data, and neither of these is easily solved. There are, however, ways of reducing the potential harm of engaging in this kind of research and those ways are closely tied to the purpose and practice of the research itself.

Big data often calls to mind a vision of extreme technological surveillance portrayed in dystopic science fiction, and with good reason. News that the NSA's PRISM project was collecting huge amounts of electronic traffic from open and closed sources has been one of the most notable revelations of the last several years, and this followed publicity around the Total Information Awareness project, which aimed to create the kind of clearinghouse of personal information envisioned in the Clotworthy article 50 years earlier. The idea that the government is collecting and storing information about many people's everyday social interactions is troubling, not least because the targets are unaware of what is being collected, who is able to view it, and what decisions are being made based on that information.

The analysis of data by companies who are able to collect social data from their own platforms is equally problematic. When researchers at Facebook collaborated with others to study the effect of timeline organization on users' emotional state (Kramer, Guillory, & Hancock, 2014), it led to a public uproar; this despite the fact that Facebook naturally collected this kind of data as part of its ongoing operations. As Fiske and Hauser (2014) suggest, because the research often serves a business purpose, and because consent is often made as part of an unread, click-through user agreement, the public can be ill-served by work done within a corporate setting. Other software and online platform companies (Microsoft, Yahoo!, and Google, among others) have created research centers that focus on how people use their products and platforms, and how this relates to the broader online environment. Of course, companies have always had research centers, and have always done market research, but the combination of these two tasks, along with unprecedented access to our personal and our public communications, leads to research that fails to support the researched.

Once publicity around the Facebook study brought it to public scrutiny, OkCupid, an online dating site, posted some of its own research to its blog, and titled the update 'We Experiment on Human Beings!' (Rudder, 2014). In the posting, the author presents some of their work and

suggests that experimentation is simply part of how websites make their offerings better. This usually happens behind the scenes and users remain unaware of it. Generalizing this research and making it more broadly available, while it could potentially serve the subjects of the research better than A-B testing or market testing might, nonetheless moves these transactions into a context that the users neither expected nor explicitly consented to.

It is worth noting that access to these data – and control of that access by platform owners – often places researchers not working for the companies themselves both at a disadvantage and potentially in an ethically compromising position. There are companies that provide free access to transactional data (StackExchange, 2014) as well as the longstanding Internet Archive's web archives (http://web.archive.org), for example. But more often, as in the case of Twitter, companies have sought to control access to data related to their site, often by including in their end user licensing agreement (EULA) language that prohibits users from collecting or disseminating information from the platform. They then sell these data to researchers and marketers who are able to afford it, or release the data to researchers on their own terms. Researchers who are unable to pay for these streams of data may either have to cobble together their own collections or reach a compromise with the businesses providing the platform. Recently, there have been other efforts by foundations, governments, and sometimes companies to provide access to such resources, including the Twitter archive at the United States Library of Congress. But more often, access requires substantial financial resources or a willingness to collaborate with businesses. The result may be a rift between data-rich and data-poor researchers (boyd & Crawford, 2012), and often this leaves relatively underfunded social scientists in the latter category. While questions of open access and financial support are hardly unique to big social data, at present, even those who have the requisite skills and training to do work in the area may find themselves unable to pay the price of admission.

Aside from the problem – for both researcher and user – of private or government control of large-scale transactional data, there is the overarching issue of consent and of the use of trace data. These data are unobtrusive, which for the social scientist interested in understanding how social interactions actually take place represents an opportunity to avoid some of the common observer effects and priming effects that more intrusive forms of research require. It also means that the user remains unaware of the collection process and the shift in context of these uses. Even, as in the case of Twitter, when interactions are relatively public, a user may not recognize that they are being observed or recorded for purposes other than that public conversation.

Smith, Milberg, and Burke (1996) lay out four dimensions of disclosure that can lead to greater degrees of privacy violation: the amount of private information collected, whether access is granted to those without permission, whether the data is used for purposes other than its original intent, and whether the data is accurate and free from errors. It would be difficult to find a large social data set that does not represent threats on each of these four dimensions. As Zimmer (2010) has shown, at least one archive of Facebook data clearly represents a substantial intrusion on users' privacy. Even if permission to disclosure users' data is granted in the EULA, as Good et al. (2005) have shown, users rarely read and understand such licenses and what they have agreed to. They rely instead on a sense of contextual integrity (Nissenbaum, 2004) and an expectation that they understand the space in which they are participating.

These two issues are certainly not independent. It is worth noting that social scientists working in a university setting generally have a human subjects board that provides an external review of their research plans and provides some level of protection for the privacy of the subjects. As part of their review, they consider not only the potential for privacy violations, but how this is balanced against the potential benefits of the research to society and to research subjects themselves. Most studies place the subjects at some risk; the question remains whether the research will benefit them significantly enough to offset this risk. Researchers with close ties to industry or government are

less likely to prioritize outcomes for the research subjects or social good. Sociology is grounded in the practice of addressing social issues. While this balance may not always be perfectly struck, research done outside of industry and government is more likely to consider ethical implications. And, to the degree that it is theoretically motivated, such research is less likely to consist merely of a 'fishing expedition' hoping to find useful relationships at the expense of subjects' privacy.

The data imaginary and craft of collection

Since sociologists and other social scientists have the methodological background to analyze big social data, an ethical grounding to help ensure that participants are protected, and a rich theoretical history to draw on, it is perhaps surprising that we do not see more work that applies these to big social data. There are two things that seem likely to restrain such research. The first is a needed shift in perspective and the second is a set of practical skills. Again, we turn to Mills' *Sociological imagination* (2000) to address these.

The 'sociological imagination' that Mills strives for is one that allows the scholar to step away from her everyday experience and recognize experiences within other milieux. But it also requires that the sociologist be able to move across scales, to recognize how people's 'troubles' may or may not relate to broader social 'issues'. For Mills, this largely meant engaging in ethnographic methods, which put him at odds with much of the postwar research establishment, and he encouraged students to unite their research and their lived existence. But he was in no way a methodological purist, and made substantial use of quantitative methods and even what might be considered 'administrative research' resources, to use Paul Lazarsfeld's term (Sterne, 2005). He recognized the need to discern social issues and then employ the tools necessary to understand these structures more clearly.

This perspective places explanation at the forefront, and recognizes that critical social science can and should draw on the broadest possible array of tools. We should not let the availability of large-scale trace data draw us away from our investigations, but neither should we shy away from these sources if they help us. This means accepting that 'data' are not the objects of our study, but a means of understanding social structure, and that they are the result of processes of observation and recording.

There was a time when natural scientists would build their own apparatus, grind their own lenses. Even if they eventually would use shared equipment, the experience of understanding and building the instruments with which they observed the world led them to see data as something shaped by the process by which it is gathered. Recent years have seen increasing interest in software studies and critical efforts to understand the biases of algorithms (Gillespie, 2010). When we collect data from these new platforms (just as when we collected data in traditional spaces), context matters. That means that an archive produced at a particular time, using particular software tools, is affected in important ways by that process (Burgess & Bruns, 2012). Unfortunately, researchers often do a poor job of communicating the processes by which we make data (Vis, 2013), but this is an important step in making clear the relationship between our tools and our observations. Just as understanding statistical methods is vital to being able to apply and interpret them, a basic grounding in programing and online systems is essential to working on big social data.

Especially during a period when many undergraduate social science programs are reducing instruction in quantitative methods, the idea of introducing networking and programming coursework may represent a challenge. After all, sociologists have successfully partnered with those with technical abilities to carry out their work. Ford (2014) describes such a coming together, with each collaborator drawing on specific knowledge and skills. Such work can succeed, but only in the rare case that the collaborators have enough training and common ground to be

able to do theoretically oriented work. That is not to say that social scientists should not make good use of skilled computer programmers to help them to achieve their ends. It is becoming increasingly essential to be able to draw on such expertise. However, just as a sociologist needs a good grounding in statistics to be able to effectively consult a statistician, she must have a foundational understanding of programming and networking to work effectively with programmers on tools to acquire data. The key to a successful bigger sociological imagination lies in training new social scientists (Lazer et al., 2009), and while the theoretical and methodological traditions of the social sciences are well suited to this new environment, the institutions and existing social science disciplines may be too ossified to rise to the challenge.

The data imaginary and the craft of making data exist on opposite ends of the spectrum, the highly conceptual in tension with the abundantly practical. Big social data requires us to expand our own perceptual apparatus, and build our own tools for manipulating and making sense of the streams of data being created in the world. Of course, we have a role in critically analyzing others' work, but unless our claim is that these sources should never be used, it is important that we both do work that can stand up as exemplary and help to assess and collect examples of excellent work as models.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Alexander Halavais is an Associate Professor of Sociology in the School of Social and Behavioral Sciences at Arizona State University, where he teaches in the graduate Social Technologies program. His research addresses issues of social change and social media, including issues of learning communities and activism. He tweets at @halavais and blogs at http://alex.halavais.net. [email: theprof@asu.edu]

ORCID

Alexander Halavais http://orcid.org/0000-0002-7164-9208

References

Anderson, C. (2008). The end of theory. *Wired Magazine 16*(7).

Arbesman, S. (2013, January 29). Stop hyping big data and start paying attention to 'long data'. *Wired*. Retrieved from http://www.wired.com/2013/01/forget-big-data-think-long-data/

Asimov, I. (1951). *Foundation: The 1,000 year plan*. New York: Ace Books.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). *Everyone's an influencer: Quantifying influence on twitter*. Proceedings of the fourth ACM international conference on Web search and data mining, Hong Kong (pp. 65–74).

Barns, S. (2014). Plus ça change? Remaking the city, 'one site, one app, one click at a time'. *City*, *18*(2), 226–229.

Boellstorff, T. (2013). Making big data, in theory. *First Monday*, *18*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4869

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, and Society*, *15*(5), 662–679.

boyd, d., Golder, S., & Lotan, G. (2010). *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter*. 43rd Hawaii International Conference on Systems Sciences, Poipu, Kauai, Hawaii.

Bruns, A. (2014). First steps in exploring the Australian twitter sphere. *Mapping Online Publics*. Retrieved from http://mappingonlinepublics.net/2014/08/04/first-steps-in-exploring-the-australian-twittersphere/

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of 'Big Social Data' for media and communication research. *M/C Journal*, *15*(5). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. *ICWSM*, *10*, 10–17.

Clotworthy, O. (1962). Some far-out thoughts on computers. Studies in Intelligence, *6*(4). Retrieved from: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol-56-no-4/pdfs/Clotworthy-Imaginative-Use-of-Computers.pdf

Davenport, T. S., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*, 70–76.

Deutsch, K. W. (1963). *The nerves of government: Models of political communication and control*. New York: Free Press of Glencoe.

Durkheim, E. (1982). *The rules of the sociological method*. New York: The Free Press.

Durkheim, Emile (2006). *On suicide*. New York: Penguin.

Earl, J., McKee Hurwitz, H., Mejia Mesinas, A., Tolan, M., & Arlotti, A. (2013). This protest will be tweeted: Twitter and protest policing during the Pittsburgh G20. *Information, Communication & Society*, *16*(4), 459–478.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton, NJ: Princeton University Press.

Fiske, S. T., & Hauser, R. M. (2014). Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences*, *111*(38), 13675–13676.

Ford, H. (2014). Big data and small: Collaborations between ethnographers and data scientists. *Big Data & Society*, *1*, 1–3.

Gaffney, D. (2010, April). #iranElection: Quantifying online activism. *Proceedings of the WebSci 10: Extending the frontiers of society* online, Raleigh, NC.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, *107*(41), 17486–17490.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrica*, *53*(4), 455–467.

González-Bailón, S. (2013). Social science in the era of big data. *Policy & Internet*, *5*(2), 147–160.

Good, N., Dhamija, R., Grossklags, J., Thaw, D., Aronowitz, S., Mulligan, D., & Konstan, J. (2005). *Stopping spyware at the gate: A user study of privacy, notice and spyware*. Proceedings of the 2005 symposium on usable privacy and security, Pittsburgh, PA (pp. 43–52).

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Halavais, A., & Garrido, M. (2014). Twitter as the people's microphone: Emergence of authorities during protest tweeting. In M. McCaughey (Ed.), *Cyberactivism on the participatory web* (pp. 117–139). London: Routledge.

Hiltz, S. R., & Turoff, M. (1993). *The network nation: Human communication via computer*. Cambridge: MIT Press.

Huberman, B., Romero, D. M., & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, *14*(1). Retrieved from http://firstmonday.org/article/view/2317/2063

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, *6*(3), 248–260.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*(8), 36–44.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). *Why we twitter: Understanding microblogging usage and communities*. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, San Jose, CA (pp. 56–65).

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788–8790.

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). *A few chirps about twitter*. Proceedings of the first workshop on online social networks, Seattle, WA (pp. 19–24).

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media?* Proceedings of the 19th international conference on world wide web, Raleigh, NC (pp. 591–600).

Laney, D. (2001, February 6). 3D data management: Controlling data volume, velocity, and variety. META Group. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949–3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science*, *323*(5915), 721.

Luhmann, N. (1989). *Ecological communication*. Chicago, IL: University of Chicago Press.

Manovich, L. (2012). Trending: The promises and challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis: University of Minnesota Press.

Miller, J. H., & Page, S. E. (2009). *Complex adaptive systems: An introduction to computational models of social life*. Princeton, NJ: Princeton University Press.

Mills, C. W. (2000). *The sociological imagination*. New York: Oxford University Press.

Moreno, J. L. (1934). *Who shall survive? A new approach to the problem of human interrelations*. Washington, DC: Nervous and Mental Disease.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, *79*(1), 119–158.

Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining*. Proceedings of the international conference on language resources and evaluation, Valletta, Malta.

Poell, T., & Borra, E. (2011). Twitter, YouTube, and Flickr as platforms for alternative journalism: The social media account of the 2010 Toronto G20 protests. *Journalism*, *13*(6), 695–713.

Rudder, J. (2014, July 28). We experiment on human beings! *oktrends.* Retrieved from http://blog.okcupid.com/index.php/we-experiment-on-human-beings/

Simmel, G. (2010). *Conflict and the web of group affiliations*. New York: Simon & Schuster.

Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, *20*(2), 167–196.

StackExchange. (2014). Database schema documentation for the public data dump and SEDE. *StackExchange Meta*. Retrieved from http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

Sterne, J. (2005). C. Wright Mills, the bureau for applied social research, and the meaning of critical scholarship. *Cultural Studies-Critical Methodologies*, *5*(1), 65–94.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*, 178–185.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). *Microblogging during two natural hazards events: What Twitter may contribute to situational awareness*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, GA (pp. 1079–1088).

Vis, F. (2013). A critical reflection on big data: Considering APIs, researchers and tools as data makers. *FirstMonday*, *18*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4878

Ward, J. S., & Barker, A. (2013). Undefined by data: A survey of big data definitions. arXiv preprint arXiv:1309.5821.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, *22*, 213–238.

Zimmer, M. (2010). 'But the data is already public': On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325.