danah boyd & Kate Crawford

CRITICAL QUESTIONS FOR BIG DATA

Provocations for a cultural, technological, and scholarly phenomenon

The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing genetic sequences, social media interactions, health records, phone logs, government records, and other digital traces left by people. Significant questions emerge. Will large-scale search data help us create better tools, services, and public goods? Or will it usher in a new wave of privacy incursions and invasive marketing? Will data analytics help us understand online communities and political movements? Or will it be used to track protesters and suppress speech? Will it transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means? Given the rise of Big Data as a socio-technical phenomenon, we argue that it is necessary to critically interrogate its assumptions and biases. In this article, we offer six provocations to spark conversations about the issues of Big Data: a cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis, and mythology that provokes extensive utopian and dystopian rhetoric.

Keywords Big Data; analytics; social media; communication studies; social network sites; philosophy of science; epistemology; ethics; Twitter

(*Received 10 December 2011; final version received 20 March 2012*)

Technology is neither good nor bad; nor is it neutral . . . technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves. (Kranzberg 1986, p. 545)

Routledge
Taylor & Francis Group

Information, Communication & Society Vol. 15, No. 5, June 2012, pp. 662–679
ISSN 1369-118X print/ISSN 1468-4462 online
http://www.tandfonline.com http://dx.doi.org/10.1080/1369118X.2012.678878

> We need to open a discourse – where there is no effective discourse now – about the varying temporalities, spatialities and materialities that we might represent in our databases, with a view to designing for maximum flexibility and allowing as possible for an emergent polyphony and polychrony. Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care. (Bowker 2005, pp. 183–184)

The era of Big Data is underway. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing genetic sequences, social media interactions, health records, phone logs, government records, and other digital traces left by people. Significant questions emerge. Will large-scale search data help us create better tools, services, and public goods? Or will it usher in a new wave of privacy incursions and invasive marketing? Will data analytics help us understand online communities and political movements? Or will analytics be used to track protesters and suppress speech? Will large quantities of data transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means?

Big Data is, in many ways, a poor term. As Manovich (2011) observes, it has been used in the sciences to refer to data sets large enough to require supercomputers, but what once required such machines can now be analyzed on desktop computers with standard software. There is little doubt that the quantities of data now available are often quite large, but that is not the defining characteristic of this new data ecosystem. In fact, some of the data encompassed by Big Data (e.g. all Twitter messages about a particular topic) are not nearly as large as earlier data sets that were not considered Big Data (e.g. census data). Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets.

We define Big Data[1] as a cultural, technological, and scholarly phenomenon that rests on the interplay of:

(1) *Technology*: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
(2) *Analysis*: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
(3) *Mythology*: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

Like other socio-technical phenomena, Big Data triggers both utopian and dystopian rhetoric. On one hand, Big Data is seen as a powerful tool to address

various societal ills, offering the potential of new insights into areas as diverse as cancer research, terrorism, and climate change. On the other, Big Data is seen as a troubling manifestation of Big Brother, enabling invasions of privacy, decreased civil freedoms, and increased state and corporate control. As with all socio-technical phenomena, the currents of hope and fear often obscure the more nuanced and subtle shifts that are underway.

Computerized databases are not new. The US Bureau of the Census deployed the world's first automated processing equipment in 1890 – the punch-card machine (Anderson 1988). Relational databases emerged in the 1960s (Fry & Sibley 1974). Personal computing and the Internet have made it possible for a wider range of people – including scholars, marketers, governmental agencies, educational institutions, and motivated individuals – to produce, share, interact with, and organize data. This has resulted in what Savage and Burrows (2007) describe as a crisis in empirical sociology. Data sets that were once obscure and difficult to manage – and, thus, only of interest to social scientists – are now being aggregated and made easily accessible to anyone who is curious, regardless of their training.

How we handle the emergence of an era of Big Data is critical. While the phenomenon is taking place in an environment of uncertainty and rapid change, current decisions will shape the future. With the increased automation of data collection and analysis – as well as algorithms that can extract and illustrate large-scale patterns in human behavior – it is necessary to ask which systems are driving these practices and which are regulating them. Lessig (1999) argues that social systems are regulated by four forces: market, law, social norms, and architecture – or, in the case of technology, code. When it comes to Big Data, these four forces are frequently at odds. The market sees Big Data as pure opportunity: marketers use it to target advertising, insurance providers use it to optimize their offerings, and Wall Street bankers use it to read the market. Legislation has already been proposed to curb the collection and retention of data, usually over concerns about privacy (e.g. the US Do Not Track Online Act of 2011). Features like personalization allow rapid access to more relevant information, but they present difficult ethical questions and fragment the public in troubling ways (Pariser 2011).

There are some significant and insightful studies currently being done that involve Big Data, but it is still necessary to ask critical questions about what all this data means, who gets access to what data, how data analysis is deployed, and to what ends. In this article, we offer six provocations to spark conversations about the issues of Big Data. We are social scientists and media studies scholars who are in regular conversation with computer scientists and informatics experts. The questions that we ask are hard ones without easy answers, although we also describe different pitfalls that may seem obvious to social scientists but are often surprising to those from different disciplines. Due to our interest in and experience with social media, our focus here is mainly on Big Data in social

media context. That said, we believe that the questions we are asking are also important to those in other fields. We also recognize that the questions we are asking are just the beginning and we hope that this article will spark others to question the assumptions embedded in Big Data. Researchers in all areas – including computer science, business, and medicine – have a stake in the computational culture of Big Data precisely because of its extended reach of influence and potential within multiple disciplines. We believe that it is time to start critically interrogating this phenomenon, its assumptions, and its biases.

1. Big Data changes the definition of knowledge

In the early decades of the twentieth century, Henry Ford devised a manufacturing system of mass production, using specialized machinery and standardized products. It quickly became the dominant vision of technological progress. 'Fordism' meant automation and assembly lines; for decades onward, this became the orthodoxy of manufacturing: out with skilled craftspeople and slow work, in with a new machine-made era (Baca 2004). But it was more than just a new set of tools. The twentieth century was marked by Fordism at a cellular level: it produced a new understanding of labor, the human relationship to work, and society at large.

Big Data not only refers to very large data sets and the tools and procedures used to manipulate and analyze them, but also to a computational turn in thought and research (Burkholder 1992). Just as Ford changed the way we made cars – and then transformed work itself – Big Data has emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community. 'Change the instruments, and you will change the entire social theory that goes with them', Latour (2009) reminds us (p. 9).

Big Data creates a radical shift in how we think about research. Commenting on computational social science, Lazer *et al*. (2009) argue that it offers 'the capacity to collect and analyze data with an unprecedented breadth and depth and scale' (p. 722). It is neither just a matter of scale nor is it enough to consider it in terms of proximity, or what Moretti (2007) refers to as distant or close analysis of texts. Rather, it is a profound change at the levels of epistemology and ethics. Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality. Just as Du Gay and Pryke (2002) note that 'accounting tools . . . do not simply aid the measurement of economic activity, they shape the reality they measure' (pp. 12–13), so Big Data stakes out new terrains of objects, methods of knowing, and definitions of social life.

Speaking in praise of what he terms 'The Petabyte Age', Anderson, Editor-in-Chief of Wired, writes:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (2008)

Do numbers speak for themselves? We believe the answer is 'no'. Significantly, Anderson's sweeping dismissal of all other theories and disciplines is a tell: it reveals an arrogant undercurrent in many Big Data debates where other forms of analysis are too easily sidelined. Other methods for ascertaining why people do things, write things, or make things are lost in the sheer volume of numbers. This is not a space that has been welcoming to older forms of intellectual craft. As Berry (2011, p. 8) writes, Big Data provides 'destablising amounts of knowledge and information that lack the regulating force of philosophy'. Instead of philosophy – which Kant saw as the rational basis for all institutions – 'computationality might then be understood as an ontotheology, creating a new ontological "epoch" as a new historical constellation of intelligibility' (Berry 2011, p. 12).

We must ask difficult questions of Big Data's models of intelligibility before they crystallize into new orthodoxies. If we return to Ford, his innovation was using the assembly line to break down interconnected, holistic tasks into simple, atomized, mechanistic ones. He did this by designing specialized tools that strongly predetermined and limited the action of the worker. Similarly, the specialized tools of Big Data also have their own inbuilt limitations and restrictions. For example, Twitter and Facebook are examples of Big Data sources that offer very poor archiving and search functions. Consequently, researchers are much more likely to focus on something in the present or immediate past – tracking reactions to an election, TV finale, or natural disaster – because of the sheer difficulty or impossibility of accessing older data.

If we are observing the automation of particular kinds of research functions, then we must consider the inbuilt flaws of the machine tools. It is not enough to simply ask, as Anderson has suggested 'what can science learn from Google?', but to ask how the harvesters of Big Data might change the meaning of learning, and what new possibilities and new limitations may come with these systems of knowing.

2. Claims to objectivity and accuracy are misleading

'Numbers, numbers, numbers', writes Latour (2009). 'Sociology has been obsessed by the goal of becoming a quantitative science'. Sociology has never reached this goal, in Latour's view, because of where it draws the line between what is and is not quantifiable knowledge in the social domain.

Big Data offers the humanistic disciplines a new way to claim the status of quantitative science and objective method. It makes many more social spaces quantifiable. In reality, working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth – particularly when considering messages from social media sites. But there remains a mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing facts. In this way, Big Data risks re-inscribing established divisions in the long running debates about scientific method and the legitimacy of social science and humanistic inquiry.

The notion of objectivity has been a central question for the philosophy of science and early debates about the scientific method (Durkheim 1895). Claims to objectivity suggest an adherence to the sphere of objects, to things as they exist in and for themselves. Subjectivity, on the other hand, is viewed with suspicion, colored as it is with various forms of individual and social conditioning. The scientific method attempts to remove itself from the subjective domain through the application of a dispassionate process whereby hypotheses are proposed and tested, eventually resulting in improvements in knowledge. Nonetheless, claims to objectivity are necessarily made by subjects and are based on subjective observations and choices.

All researchers are interpreters of data. As Gitelman (2011) observes, data need to be imagined as data in the first instance, and this process of the imagination of data entails an interpretative base: 'every discipline and disciplinary institution has its own norms and standards for the imagination of data'. As computational scientists have started engaging in acts of social science, there is a tendency to claim their work as the business of facts and not interpretation. A model may be mathematically sound, an experiment may seem valid, but as soon as a researcher seeks to understand what it means, the process of interpretation has begun. This is not to say that all interpretations are created equal, but rather that not all numbers are neutral.

The design decisions that determine what will be measured also stem from interpretation. For example, in the case of social media data, there is a 'data cleaning' process: making decisions about what attributes and variables will be counted, and which will be ignored. This process is inherently subjective. As Bollier explains,

> As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an 'objective truth' or is any interpretation necessarily biased by some subjective filter or the way that data is 'cleaned?'. (2010, p. 13)

In addition to this question, there is the issue of data errors. Large data sets from Internet sources are often unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are used together. Social scientists have a long history of asking critical questions about the collection of data and trying to account for any biases in their data (Cain & Finch 1981; Clifford & Marcus 1986). This requires understanding the properties and limits of a data set, regardless of its size. A data set may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a data set, we need to know where data is coming from; it is similarly important to know and account for the weaknesses in that data. Furthermore, researchers must be able to account for the biases in their interpretation of the data. To do so requires recognizing that one's identity and perspective informs one's analysis (Behar & Gordon 1996).

Too often, Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions. In one notable example, Leinweber (2007) demonstrated that data mining techniques could show a strong but spurious correlation between the changes in the S&P 500 stock index and butter production in Bangladesh.

Interpretation is at the center of data analysis. Regardless of the size of a data, it is subject to limitation and bias. Without those biases and limitations being understood and outlined, misinterpretation is the result. Data analysis is most effective when researchers take account of the complex methodological processes that underlie the analysis of that data.

3. Bigger data are not always better data

Social scientists have long argued that what makes their work rigorous is rooted in their systematic approach to data collection and analysis (McCloskey 1985). Ethnographers focus on reflexively accounting for bias in their interpretations. Experimentalists control and standardize the design of their experiment. Survey researchers drill down on sampling mechanisms and question bias. Quantitative researchers weigh up statistical significance. These are but a few of the ways in which social scientists try to assess the validity of each other's work. Just because Big Data presents us with large quantities of data does not mean that methodological issues are no longer relevant. Understanding sample, for example, is more important now than ever.

Twitter provides an example in the context of a statistical analysis. Because it is easy to obtain – or scrape – Twitter data, scholars have used Twitter to examine a wide variety of patterns (e.g. mood rhythms (Golder & Macy 2011), media event engagement (Shamma *et al*. 2010), political uprisings (Lotan *et al*. 2011), and conversational interactions (Wu *et al*. 2011)). While

many scholars are conscientious about discussing the limitations of Twitter data in their publications, the public discourse around such research tends to focus on the raw number of tweets available. Even news coverage of scholarship tends to focus on how many millions of 'people' were studied (Wang 2011).

Twitter does not represent 'all people', and it is an error to assume 'people' and 'Twitter users' are synonymous: they are a very particular sub-set. Neither is the population using Twitter representative of the global population. Nor can we assume that accounts and users are equivalent. Some users have multiple accounts, while some accounts are used by multiple people. Some people never establish an account, and simply access Twitter via the web. Some accounts are 'bots' that produce automated content without directly involving a person. Furthermore, the notion of an 'active' account is problematic. While some users post content frequently through Twitter, others participate as 'listeners' (Crawford 2009, p. 532). Twitter Inc. has revealed that 40 percent of active users sign in just to listen (Twitter 2011). The very meanings of 'user' and 'participation' and 'active' need to be critically examined.

Big Data and whole data are also not the same. Without taking into account the sample of a data set, the size of the data set is meaningless. For example, a researcher may seek to understand the topical frequency of tweets, yet if Twitter removes all tweets that contain problematic words or content – such as references to pornography or spam – from the stream, the topical frequency would be inaccurate. Regardless of the number of tweets, it is not a representative sample as the data is skewed from the beginning.

It is also hard to understand the sample when the source is uncertain. Twitter Inc. makes a fraction of its material available to the public through its APIs.[2] The 'firehose' theoretically contains all public tweets ever posted and explicitly excludes any tweet that a user chose to make private or 'protected'. Yet, some publicly accessible tweets are also missing from the firehose. Although a handful of companies have access to the firehose, very few researchers have this level of access. Most either have access to a 'gardenhose' (roughly 10 percent of public tweets), a 'spritzer' (roughly one percent of public tweets), or have used 'white-listed' accounts where they could use the APIs to get access to different subsets of content from the public stream.[3] It is not clear what tweets are included in these different data streams or sampling them represents. It could be that the API pulls a random sample of tweets or that it pulls the first few thousand tweets per hour or that it only pulls tweets from a particular segment of the network graph. Without knowing, it is difficult for researchers to make claims about the quality of the data that they are analyzing. Are the data representative of all tweets? No, because they exclude tweets from protected accounts.[4] But are the data representative of all public tweets? Perhaps, but not necessarily.

Twitter has become a popular source for mining Big Data, but working with Twitter data has serious methodological challenges that are rarely addressed by those who embrace it. When researchers approach a data set, they need to

understand – and publicly account for – not only the limits of the data set, but also the limits of which questions they can ask of a data set and what interpretations are appropriate.

This is especially true when researchers combine multiple large data sets. This does not mean that combining data does not offer valuable insights – studies like those by Acquisti and Gross (2009) are powerful, as they reveal how public databases can be combined to produce serious privacy violations, such as revealing an individual's Social Security number. Yet, as Jesper Anderson, co-founder of open financial data store FreeRisk, explains: combining data from multiple sources creates unique challenges. 'Every one of those sources is error-prone . . . I think we are just magnifying that problem [when we combine multiple data sets]' (Bollier 2010, p. 13).

Finally, during this computational turn, it is increasingly important to recognize the value of 'small data'. Research insights can be found at any level, including at very modest scales. In some cases, focusing just on a single individual can be extraordinarily valuable. Take, for example, the work of Veinot (2007), who followed one worker – a vault inspector at a hydroelectric utility company – in order to understand the information practices of a blue-collar worker. In doing this unusual study, Veinot reframed the definition of 'information practices' away from the usual focus on early-adopter, white-collar workers, to spaces outside of the offices and urban context. Her work tells a story that could not be discovered by farming millions of Facebook or Twitter accounts, and contributes to the research field in a significant way, despite the smallest possible participant count. The size of data should fit the research question being asked; in some cases, small is best.

4. Taken out of context, Big Data loses its meaning

Because large data sets can be modeled, data are often reduced to what can fit into a mathematical model. Yet, taken out of context, data lose meaning and value. The rise of social network sites prompted an industry-driven obsession with the 'social graph'. Thousands of researchers have flocked to Twitter and Facebook and other social media to analyze connections between messages and accounts, making claims about social networks. Yet, the relations displayed through social media are not necessarily equivalent to the sociograms and kinship networks that sociologists and anthropologists have been investigating since the 1930s (Radcliffe-Brown 1940; Freeman 2006). The ability to represent relationships between people as a graph does not mean that they convey equivalent information.

Historically, sociologists and anthropologists collected data about people's relationships through surveys, interviews, observations, and experiments. Using this data, they focused on describing people's 'personal networks' – the

set of relationships that individuals develop and maintain (Fischer 1982). These connections were evaluated based on a series of measures developed over time to identify personal connections. Big Data introduces two new popular types of social networks derived from data traces: 'articulated networks' and 'behavioral networks'.

Articulated networks are those that result from people specifying their contacts through technical mechanisms like email or cell phone address books, instant messaging buddy lists, 'Friends' lists on social network sites, and 'Follower' lists on other social media genres. The motivations that people have for adding someone to each of these lists vary widely, but the result is that these lists can include friends, colleagues, acquaintances, celebrities, friends-of-friends, public figures, and interesting strangers.

Behavioral networks are derived from communication patterns, cell coordinates, and social media interactions (Onnela *et al*. 2007; Meiss *et al*. 2008). These might include people who text message one another, those who are tagged in photos together on Facebook, people who email one another, and people who are physically in the same space, at least according to their cell phone.

Both behavioral and articulated networks have great value to researchers, but they are not equivalent to personal networks. For example, although contested, the concept of 'tie strength' is understood to indicate the importance of individual relationships (Granovetter 1973). When mobile phone data suggest that workers spend more time with colleagues than their spouse, this does not necessarily imply that colleagues are more important than spouses. Measuring tie strength through frequency or public articulation is a common mistake: tie strength – and many of the theories built around it – is a subtle reckoning in how people understand and value their relationships with other people. Not every connection is equivalent to every other connection, and neither does frequency of contact indicate strength of relationship. Further, the absence of a connection does not necessarily indicate that a relationship should be made.

Data are not generic. There is value to analyzing data abstractions, yet retaining context remains critical, particularly for certain lines of inquiry. Context is hard to interpret at scale and even harder to maintain when data are reduced to fit into a model. Managing context in light of Big Data will be an ongoing challenge.

5. Just because it is accessible does not make it ethical

In 2006, a Harvard-based research group started gathering the profiles of 1,700 college-based Facebook users to study how their interests and friendships changed over time (Lewis *et al*. 2008). These supposedly anonymous data were released to the world, allowing other researchers to explore and analyze

them. What other researchers quickly discovered was that it was possible to de-anonymize parts of the data set: compromising the privacy of students, none of whom were aware their data were being collected (Zimmer 2008).

The case made headlines and raised difficult issues for scholars: what is the status of so-called 'public' data on social media sites? Can it simply be used, without requesting permission? What constitutes best ethical practice for researchers? Privacy campaigners already see this as a key battleground where better privacy protections are needed. The difficulty is that privacy breaches are hard to make specific – is there damage done at the time? What about 20 years hence? 'Any data on human subjects inevitably raise privacy issues, and the real risks of abuse of such data are difficult to quantify' (Nature, cited in Berry 2011).

Institutional Review Boards (IRBs) – and other research ethics committees – emerged in the 1970s to oversee research on human subjects. While unquestionably problematic in implementation (Schrag 2010), the goal of IRBs is to provide a framework for evaluating the ethics of a particular line of research inquiry and to make certain that checks and balances are put into place to protect subjects. Practices like 'informed consent' and protecting the privacy of informants are intended to empower participants in light of earlier abuses in the medical and social sciences (Blass 2004; Reverby 2009). Although IRBs cannot always predict the harm of a particular study – and, all too often, prevent researchers from doing research on grounds other than ethics – their value is in prompting researchers to think critically about the ethics of their project.

Very little is understood about the ethical implications underpinning the Big Data phenomenon. Should someone be included as a part of a large aggregate of data? What if someone's 'public' blog post is taken out of context and analyzed in a way that the author never imagined? What does it mean for someone to be spotlighted or to be analyzed without knowing it? Who is responsible for making certain that individuals and communities are not hurt by the research process? What does informed consent look like?

It may be unreasonable to ask researchers to obtain consent from every person who posts a tweet, but it is problematic for researchers to justify their actions as ethical simply because the data are accessible. Just because content is publicly accessible does not mean that it was meant to be consumed by just anyone. There are serious issues involved in the ethics of online data collection and analysis (Ess 2002). The process of evaluating the research ethics cannot be ignored simply because the data are seemingly public. Researchers must keep asking themselves – and their colleagues – about the ethics of their data collection, analysis, and publication.

In order to act ethically, it is important that researchers reflect on the importance of accountability: both to the field of research and to the research subjects. Accountability here is used as a broader concept than privacy, as Troshynski *et al.*

(2008) have outlined, where the concept of accountability can apply even when conventional expectations of privacy are not in question. Instead, accountability is a multi-directional relationship: there may be accountability to superiors, to colleagues, to participants, and to the public (Dourish & Bell 2011). Academic scholars are held to specific professional standards when working with human participants in order to protect informants' rights and well-being. However, many ethics boards do not understand the processes of mining and anonymizing Big Data, let alone the errors that can cause data to become personally identifiable. Accountability requires rigorous thinking about the ramifications of Big Data, rather than assuming that ethics boards will necessarily do the work of ensuring that people are protected.

There are also significant questions of truth, control, and power in Big Data studies: researchers have the tools and the access, while social media users as a whole do not. Their data were created in highly context-sensitive spaces, and it is entirely possible that some users would not give permission for their data to be used elsewhere. Many are not aware of the multiplicity of agents and algorithms currently gathering and storing their data for future use. Researchers are rarely in a user's imagined audience. Users are not necessarily aware of all the multiple uses, profits, and other gains that come from information they have posted. Data may be public (or semi-public) but this does not simplistically equate with full permission being given for all uses. Big Data researchers rarely acknowledge that there is a considerable difference between being in public (i.e. sitting in a park) and being public (i.e. actively courting attention) (boyd & Marwick 2011).

6. Limited access to Big Data creates new digital divides

In an essay on Big Data, Golder (2010) quotes sociologist Homans (1974): 'The methods of social science are dear in time and money and getting dearer every day'. Historically speaking, collecting data has been hard, time consuming, and resource intensive. Much of the enthusiasm surrounding Big Data stems from the perception that it offers easy access to massive amounts of data.

But who gets access? For what purposes? In what contexts? And with what constraints? While the explosion of research using data sets from social media sources would suggest that access is straightforward, it is anything but. As Manovich (2011) points out, 'only social media companies have access to really large social data – especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not'. Some companies restrict access to their data entirely; others sell the privilege of access for a fee; and others offer small data sets to university-based researchers. This produces considerable unevenness in the system: those with money – or those inside the company – can produce

a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access.

It is also important to recognize that the class of the Big Data rich is reinforced through the university system: top-tier, well-resourced universities will be able to buy access to data, and students from the top universities are the ones most likely to be invited to work within large social media companies. Those from the periphery are less likely to get those invitations and develop their skills. The result is that the divisions between scholars will widen significantly.

In addition to questions of access, there are questions of skills. Wrangling APIs, scraping, and analyzing big swathes of data is a skill set generally restricted to those with a computational background. When computational skills are positioned as the most valuable, questions emerge over who is advantaged and who is disadvantaged in such a context. This, in its own way, sets up new hierarchies around 'who can read the numbers', rather than recognizing that computer scientists and social scientists both have valuable perspectives to offer. Significantly, this is also a gendered division. Most researchers who have computational skills at the present moment are male and, as feminist historians and philosophers of science have demonstrated, who is asking the questions determines which questions are asked (Harding 2010; Forsythe 2001). There are complex questions about what kinds of research skills are valued in the future and how those skills are taught. How can students be educated so that they are equally comfortable with algorithms and data analysis as well as with social analysis and theory?

Finally, the difficulty and expense of gaining access to Big Data produce a restricted culture of research findings. Large data companies have no responsibility to make their data available, and they have total control over who gets to see them. Big Data researchers with access to proprietary data sets are less likely to choose questions that are contentious to a social media company if they think it may result in their access being cut. The chilling effects on the kinds of research questions that can be asked – in public or private – are something we all need to consider when assessing the future of Big Data.

The current ecosystem around Big Data creates a new kind of digital divide: the Big Data rich and the Big Data poor. Some company researchers have even gone so far as to suggest that academics should not bother studying social media data sets – Jimmy Lin, a professor on industrial sabbatical at Twitter argued that academics should not engage in research that industry 'can do better' (Conover 2011). Such explicit efforts to demarcate research 'insiders' and 'outsiders' – while by no means new – undermine the research community. 'Effective democratisation can always be measured by this essential criterion', Derrida (1996) claimed, 'the participation in and access to the archive, its constitution, and its interpretation' (p. 4).

Whenever inequalities are explicitly written into the system, they produce class-based structures. Manovich (2011) writes of three classes of people in the realm of Big Data: 'those who create data (both consciously and by leaving digital footprints), those who have the means to collect it, and those who have expertise to analyze it'. We know that the last group is the smallest, and the most privileged: they are also the ones who get to determine the rules about how Big Data will be used, and who gets to participate. While institutional inequalities may be a forgone conclusion in academia, they should nevertheless be examined and questioned. They produce a bias in the data and the types of research that emerge.

By arguing that the Big Data phenomenon is implicated in some broad historical and philosophical shifts is not to suggest it is solely accountable; the academy is by no means the sole driver behind the computational turn. There is a deep government and industrial drive toward gathering and extracting maximal value from data, be it information that will lead to more targeted advertising, product design, traffic planning, or criminal policing. But we do think there are serious and wide-ranging implications for the operationalization of Big Data, and what it will mean for future research agendas. As Suchman (2011) observes, via Levi Strauss, 'we are our tools'. We should consider how the tools participate in shaping the world with us as we use them. The era of Big Data has only just begun, but it is already important that we start questioning the assumptions, values, and biases of this new wave of research. As scholars who are invested in the production of knowledge, such interrogations are an essential component of what we do.

Acknowledgements

We wish to thank Heather Casteel for her help in preparing this article. We are also deeply grateful to Eytan Adar, Tarleton Gillespie, Bernie Hogan, Mor Naaman, Jussi Parikka, Christian Sandvig, and all the members of the Microsoft Research Social Media Collective for inspiring conversations, suggestions, and feedback. We are indebted to all who provided feedback at the Oxford Internet Institute's 10th Anniversary. Finally, we appreciate the anonymous reviewers' helpful comments.

Notes

1 We have chosen to capitalize the term 'Big Data' throughout this article to make it clear that it is the phenomenon we are discussing.

2 API stands for application programming interface; this refers to a set of tools that developers can use to access structured data.

3 Details of what Twitter provides can be found at https://dev.Twitter.com/docs/streaming-api/methods White-listed accounts were commonly used by researchers, but they are no longer available.

4 The percentage of protected accounts is unknown, although attempts to identify protected accounts suggest that under 10 percent of accounts are protected (Meeder *et al*. 2010).

References

Acquisti, A. & Gross, R. (2009) 'Predicting social security numbers from public data', *Proceedings of the National Academy of Science*, vol. 106, no. 27, pp. 10975–10980.

Anderson, C. (2008) 'The end of theory, will the data deluge makes the scientific method obsolete?', *Edge*, [Online] Available at: http://www.edge.org/3rd_culture/anderson08/anderson08_index.html (25 July 2011).

Anderson, M. (1988) *The American Census: A Social History*, Yale University Press, New Haven, CT.

Baca, G. (2004) 'Legends of Fordism: between myth, history, and foregone conclusions', *Social Analysis*, vol. 48, no. 3, pp. 169–178.

Behar, R. & Gordon, D. A. (eds) (1996) *Women Writing Culture*, University of California Press, Berkeley, CA.

Berry, D. (2011) 'The computational turn: thinking about the digital humanities', *Culture Machine*, vol. 12, [Online] Available at: http://www.culturemachine.net/index.php/cm/article/view/440/470 (11 July 2011).

Blass, T. (2004) *The Man Who Shocked the World: The Life and Legacy of Stanley Milgram*, Basic Books, New York.

Bollier, D. (2010) 'The promise and peril of big data', [Online] Available at: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf (11 July 2011).

Bowker, G. C. (2005) *Memory Practices in the Sciences*, MIT Press, Cambridge, MA.

Boyd, D. & Marwick, A. (2011) 'Social privacy in networked publics: teens' attitudes, practices, and strategies', paper given at *Oxford Internet Institute*, [Online] Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1925128 (28 September 2011).

Burkholder, L. (ed.) (1992) *Philosophy and the Computer*, Westview Press, Boulder, San Francisco, and Oxford.

Cain, M. & Finch, J. (1981) 'Towards a rehabilitation of data', in *Practice and Progress: British Sociology 1950–1980*, eds P. Abrams, R. Deem, J. Finch & P. Rock, George Allen and Unwin, London, pp. 105–119.

Clifford, J. & Marcus, G. E. (eds) (1986) *Writing Culture: The Poetics and Politics of Ethnography*, University of California Press, Berkeley, CA.

Conover, M. (2011) 'Jimmy Lin', *Complexity and Social Networks Blog*, [Online] Available at: http://www.iq.harvard.edu/blog/netgov/2011/07/the_international_conference_o.html (9 December 2011).

Crawford, K. (2009) 'Following you: disciplines of listening in social media', *Continuum: Journal of Media & Cultural Studies*, vol. 23, no. 4, pp. 532–533.

Derrida, J. (1996) *Archive Fever: A Freudian Impression*, trans. Eric Prenowitz, University of Chicago Press, Chicago.

Dourish, P. & Bell, G. (2011) *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing*, MIT Press, Cambridge, MA.

Du Gay, P. & Pryke, M. (2002) *Cultural Economy: Cultural Analysis and Commercial Life*, Sage, London.

Durkheim, E. (1895/1982) *Rules of Sociological Method*, The Free Press, New York, NY.

Ess, C. (2002) 'Ethical decision-making and Internet research: recommendations from the aoir ethics working committee', *Association of Internet Researchers*, [Online] Available at: http://aoir.org/reports/ethics.pdf (12 September 2011).

Fischer, C. (1982) *To Dwell Among Friends: Personal Networks in Town and City*, University of Chicago, Chicago.

Forsythe, D. (2001) *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*, Stanford University Press, Stanford.

Freeman, L. (2006) *The Development of Social Network Analysis*, Empirical Press, Vancouver.

Fry, J. P. & Sibley, E. H. (1996) [1974] 'Evolution of database management systems', *Computing Surveys*, vol. 8, no. 1.1, pp. 7–42. Reprinted in (1996) *Great Papers in Computer Science*, ed. L. Laplante, IEEE Press, New York.

Gitelman, L. (2011) *Notes for the Upcoming Collection 'Raw Data' is an Oxymoron*, [Online] Available at: https://files.nyu.edu/lg91/public/ (23 July 2011).

Golder, S. (2010) 'Scaling social science with hadoop', *Cloudera Blog*, [Online] Available at: http://www.cloudera.com/blog/2010/04/scaling-social-science-with-hadoop/ (18 June 2011).

Golder, S. & Macy, M. W. (2011) 'Diurnal and seasonal mood vary with work, sleep and daylength across diverse cultures', *Science*, vol. 333, no. 6051, pp. 1878–1881, [Online] Available at: http://www.sciencemag.org/content/333/6051/1878.

Granovetter, M. S. (1973) 'The strength of weak ties', *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380.

Harding, S. (2010) 'Feminism, science and the anti-Enlightenment critiques', in *Women, Knowledge and Reality: Explorations in Feminist Philosophy*, eds A. Garry & M. Pearsall, Unwin Hyman, Boston, MA, pp. 298–320.

Homans, G. C. (1974) *Social Behavior: Its Elementary Forms*, Harvard University Press, Cambridge, MA.

Kranzberg, M. (1986) 'Technology and history: kranzberg's laws', *Technology and Culture*, vol. 27, no. 3, pp. 544–560.

Latour, B. (2009) 'Tarde's idea of quantification', in *The Social after Gabriel Tarde: Debates and Assessments*, ed. M. Candea, Routledge, London, pp. 145–162,

[Online] Available at: http://www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf (19 June 2011).

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. & Van Alstyne, M. (2009) 'Computational social science', *Science*, vol. 323, no. 5915, pp. 721–723.

Leinweber, D. (2007) 'Stupid data miner tricks: overfitting the S&P 500', *The Journal of Investing*, vol. 16, no. 1, pp. 15–22.

Lessig, L. (1999) *Code: and Other Laws of Cyberspace*, Basic Books, New York, NY.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. & Christakis, N. (2008) 'Tastes, ties, and time: a new social network dataset using Facebook.com', *Social Networks*, vol. 30, no. 4, pp. 330–342.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I. & boyd, D. (2011) 'The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions', *International Journal of Communications*, vol. 5, pp. 1375–1405, [Online] Available at: http://ijoc.org/ojs/index.php/ijoc/article/view/1246.

Manovich, L. (2011) 'Trending: the promises and the challenges of big social data', in *Debates in the Digital Humanities*, ed. M. K. Gold, The University of Minnesota Press, Minneapolis, MN, [Online] Available at: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (15 July 2011).

McCloskey, D. N. (ed.) (1985) 'From methodology to rhetoric', *The Rhetoric of Economics*, University of Wisconsin Press, Madison, pp. 20–35.

Meeder, B., Tam, J., Gage Kelley, P. & Faith Cranor, L. (2010) 'RT @IWantPrivacy: widespread violation of privacy settings in the Twitter social network', paper presented at Web 2.0 Security and Privacy, W2SP 2011, Oakland, CA.

Meiss, M. R., Menczer, F. & Vespignani, A. (2008) 'Structural analysis of behavioral networks from the Internet', *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, pp. 220–224.

Moretti, F. (2007) *Graphs, Maps, Trees: Abstract Models for a Literary History*, Verso, London.

Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. & Barabási, A. L. (2007) 'Structure and tie strengths in mobile communication networks', *Proceedings from the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336.

Pariser, E. (2011) *The Filter Bubble: What the Internet is Hiding from You*, Penguin Press, New York.

Radcliffe-Brown, A. R. (1940) 'On social structure', *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, vol. 70, no. 1, pp. 1–12.

Reverby, S. M. (2009) *Examining Tuskegee: The Infamous Syphilis Study and Its Legacy*, University of North Carolina Press, Chapel Hill, NC.

Savage, M. & Burrows, R. (2007) 'The coming crisis of empirical sociology', *Sociology*, vol. 41, no. 5, pp. 885–899.

Schrag, Z. M. (2010) *Ethical Imperialism: Institutional Review Boards and the Social Sciences, 1965–2009*, Johns Hopkins University Press, Baltimore, MD.

Shamma, D. A., Kennedy, L., and Churchill, E. F. (2010) 'Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events?,' *Paper presented at the Computer-Supported Cooperative Work-2010, Association for Computing Machinery*, February 6–10, Savannah, Georgia USA. Available at: http://research.yahoo.com/pub/3041.

Suchman, L. (2011) 'Consuming anthropology', in *Interdisciplinarity: Reconfigurations of the Social and Natural Sciences*, eds A. Barry & G. Born, Routledge, London, [Online] Available at: http://www.lancs.ac.uk/fass/doc_library/sociology/Suchman_consuming_anthroploogy.pdf.

Troshynski, E., Lee, C. & Dourish, P. (2008) 'Accountabilities of presence: reframing location-based systems,' *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, April 5–10, Florence, Italy.

Twitter (2011) 'One hundred million voices', *Twitter Blog*, [Online] Available at: http://blog.Twitter.com/2011/09/one-hundred-million-voices.html (12 September 2011).

Veinot, T. (2007) 'The eyes of the power company: workplace information practices of a vault inspector', *The Library Quarterly*, vol. 77, no. 2, pp. 157–180.

Wang, X. (2011) 'Twitter posts show workers worldwide are stressed out on the job', *Bloomberg Businessweek*, [Online] Available at: http://www.businessweek.com/news/2011-09-29/Twitter-posts-show-workers-worldwide-are-stressed-out-on-the-job.html (12 March 2012).

Wu, S., Hofman, J. M., Mason, W. A. & Watts, D. J. (2011) 'Who says what to whom on Twitter', *Proceedings of the International World Wide Web Conference (WWW 2011)*, March 28-April 1, Hyderabad, India, pp. 705–714.

Zimmer, M. (2008) 'More on the "Anonymity" of the Facebook dataset – it's Harvard College', *MichaelZimmer.org Blog*, [Online] Available at: http://www.michaelzimmer.org/2008/01/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/ (20 June 2011).

---

danah boyd is Senior Researcher at Microsoft Research, Research Assistant Professor at New York University, and Fellow at Harvard's Berkman Center for Internet & Society. Her work focuses on how people integrate social media into their everyday practices, with a particular eye towards youth's socio-technical practices. Her next book is called *It's Complicated: The Social Lives of Networked Teens* (Yale University Press). *Address*: Microsoft Research, One Memorial Drive, Cambridge, 02142 MA, USA. [email: danah-ics@danah.org]

Kate Crawford is Associate Professor at the University of New South Wales, Sydney, and Principal Researcher at Microsoft Research New England. She has conducted large scale studies of mobile and social media use, and has been published widely on the cultural and political contexts of social media. Her next book is the coauthored *Internet Adaptations: Language, Technology, Media, Power* (Palgrave Macmillan). *Address*: Microsoft Research, One Memorial Drive, Cambridge, 02142 MA, USA. [email: kate@katecrawford.net]