Paper to be presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011.

Six Provocations for Big Data

danah boyd
Microsoft Research
dmb@microsoft.com

Kate Crawford
University of New South Wales
k.crawford@unsw.edu.au

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.

Melvin Kranzberg (1986, p. 545)

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.

Geoffrey Bowker (2005, p. 183-184)

The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and many others 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 information from Twitter, Google, Verizon, 23andMe, Facebook, Wikipedia, and every space where large groups of people leave digital traces and deposit data. Significant questions emerge. Will large-scale analysis of DNA help cure diseases? Or will it usher in a new wave of medical inequality? Will data analytics help make people's access to information more efficient and effective? Or will it be used to track protesters in the streets of major cities? Will it transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means? Some or all of the above?

Big Data is, in many ways, a poor term. As Lev Manovich (2011) observes, it has been used in the sciences to refer to data sets large enough to require supercomputers, although now vast sets of data can be analyzed on desktop computers with standard software. There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem. Big Data is notable not because of its size, but because of its relationality to other data. Due to efforts to mine

and aggregate data, Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself.

Furthermore, Big Data is important because it refers to an analytic phenomenon playing out in academia and industry. Rather than suggesting a new term, we are using Big Data here because of its popular salience and because it is the phenomenon around Big Data that we want to address. Big Data tempts some researchers to believe that they can see everything at a 30,000-foot view. It is the kind of data that encourages the practice of apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions. Due to this, it is crucial to begin asking questions about the analytic assumptions, methodological frameworks, and underlying biases embedded in the Big Data phenomenon.

While databases have been aggregating data for over a century, Big Data is no longer just the domain of actuaries and scientists. New technologies have made it possible for a wide range of people – including humanities and social science academics, marketers, governmental organizations, educational institutions, and motivated individuals – to produce, share, interact with, and organize data. Massive data sets that were once obscure and distinct are being aggregated and made easily accessible. Data is increasingly digital air: the oxygen we breathe and the carbon dioxide that we exhale. It can be a source of both sustenance and pollution.

How we handle the emergence of an era of Big Data is critical: while it is taking place in an environment of uncertainty and rapid change, current decisions will have considerable impact in the future. With the increased automation of data collection and analysis – as well as algorithms that can extract and inform us of massive patterns in human behavior – it is necessary to ask which systems are driving these practices, and which are regulating them. In *Code*, Lawrence Lessig (1999) argues that systems are regulated by four forces: the market, the law, social norms, and architecture – or, in the case of technology, code. When it comes to Big Data, these four forces are at work and, frequently, at odds. The market sees Big Data as pure opportunity: marketers use it to target advertising, insurance providers want to optimize their offerings, and Wall Street bankers use it to read better readings on market temperament. Legislation has already been proposed to curb the collection and retention of data, usually over concerns about privacy (for example, the Do Not Track Online Act of 2011 in the United States). Features like personalization allow rapid access to more relevant information, but they present difficult ethical questions and fragment the public in problematic ways (Pariser 2011).

There are some significant and insightful studies currently being done that draw on Big Data methodologies, particularly studies of practices in social network sites like Facebook and Twitter. Yet, it is imperative that we begin asking critical questions about what all this data means, who gets access to it, how it is deployed, and to what ends. With Big Data come big responsibilities. In this essay, we are offering six provocations that we hope can spark conversations about the issues of Big Data. Social and cultural researchers

have a stake in the computational culture of Big Data precisely because many of its central questions are fundamental to our disciplines. Thus, we believe that it is time to start critically interrogating this phenomenon, its assumptions, and its biases.

1. Automating Research Changes the Definition of Knowledge.

In the early decades of the 20th century, Henry Ford devised a manufacturing system of mass production, using specialized machinery and standardized products. Simultaneously, it became the dominant vision of technological progress. Fordism meant automation and assembly lines, and 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 20th 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 reminds us (2009, p. 9).

We would argue that Bit Data creates a radical shift in how we think about research. Commenting on computational social science, Lazer *et al* argue that it offers 'the capacity to collect and analyze data with an unprecedented breadth and depth and scale' (2009, p. 722). But it is not just a matter of scale. Neither is 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. It 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 note that 'accounting tools...do not simply aid the measurement of economic activity, they shape the reality they measure' (2002, 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', Chris 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? The answer, we think, is a resounding '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 all other forms of analysis can be sidelined by production lines of numbers, privileged as having a direct line to raw knowledge. Why people do things, write things, or make things is erased by the sheer volume of numerical repetition and large patterns. This is not a space for reflection or the older forms of intellectual craft. As David 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. One is the issue of time. 'Big Data is about exactly right now, with no historical context that is predictive,' observes Joi Ito, the director of the MIT Media Lab (Bollier 2010, p. 19). For example, Twitter and Facebook are examples of Big Data sources that offer very poor archiving and search functions, where 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 suggests 'what can science learn from Google?', but to ask how Google and the other 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 (2010). 'Sociology has been obsessed by the goal of becoming a quantitative science.' Yet 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 reinscribing established divisions in the long running debates about scientific method.

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 Lisa Gitelman (2011) observes, data needs 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. 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 dataset, regardless of its size. A dataset may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a dataset, 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).

Spectacular errors can emerge when researchers try to build social science findings into technological systems. A classic example arose when Friendster chose to implement Robin Dunbar's (1998) work. Analyzing gossip practices in humans and grooming habits in monkeys, Dunbar found that people could only actively maintain 150 relationships at any time and argued that this number represented the maximum size of a person's personal network. Unfortunately, Friendster believed that people were replicating their pre-existing personal networks on the site, so they inferred that no one should have a friend list greater than 150. Thus, they capped the number of 'Friends' people could have on the system (boyd, 2006).

Interpretation is at the center of data analysis. Regardless of the size of a data set, it is subject to limitation and bias. Without those biases and limitations being understood and outlined, misinterpretation is the result. Big Data is at its most effective when researchers take account of the complex methodological processes that underlie the analysis of social 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 (McClosky, 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. Unfortunately, some who are embracing Big Data presume the core methodological issues in the social sciences are no longer relevant. There is a problematic underlying ethos that bigger is better, that quantity necessarily means quality.

Twitter provides an example in the context of a statistical analysis. First, Twitter does not represent 'all people', although many journalists and researchers refer to 'people' and 'Twitter users' as synonymous. 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. 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 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.

Due to uncertainties about what an account represents and what engagement looks like, it is standing on precarious ground to sample Twitter accounts and make claims about people and users. Twitter Inc. can make claims about all accounts or all tweets or a random sample thereof as they have access to the central database. Even so, they cannot

easily account for lurkers, people who have multiple accounts or groups of people who all access one account. Additionally, the central database is also prone to outages, and tweets are frequently lost and deleted.

Twitter Inc. makes a fraction of its material available to the public through its APIs[1]. 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 and startups have access to the firehose, very few researchers have this level of access. Most either have access to a 'gardenhose' (roughly 10% of public tweets), a 'spritzer' (roughly 1% 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.[2] 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. Given uncertainty, it is difficult for researchers to make claims about the quality of the data that they are analyzing. Is the data representative of all tweets? No, because it excludes tweets from protected accounts.[3] Is the data representative of all public tweets? Perhaps, but not necessarily.

These are just a few of the unknowns that researchers face when they work with Twitter data, yet these limitations are rarely acknowledged. Even those who provide a mechanism for how they sample from the firehose or the gardenhose rarely reveal what might be missing or how their algorithms or the architecture of Twitter's system introduces biases into the dataset. Some scholars simply focus on the raw number of tweets: but big data and whole data are not the same. Without taking into account the sample of a dataset, the size of the dataset 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 – from the stream, the topical frequency would be wholly inaccurate. Regardless of the number of tweets, it is not a representative sample as the data is skewed from the beginning.

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

---

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

[2] Details of what Twitter provides can be found at https://dev.twitter.com/docs/streaming-api/methods White-listed accounts were a common mechanism of acquiring access early on, but they are no longer available.

[3] The percentage of protected accounts is unknown. In a study of Twitter where they attempted to locate both protected and public Twitter accounts, Meeder et al (2010) found that 8.4% of the accounts they identified were protected.

This is especially true when researchers combine multiple large datasets. Jesper Anderson, co-founder of open financial data store FreeRisk, explains that 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). This does not mean that combining data doesn't have value – studies like those by Alessandro Acquisti and Ralph Gross (2009), which reveal how databases can be combined to reveal serious privacy violations are crucial. Yet, it is imperative that such combinations are not without methodological rigor and transparency.

Finally, in the era of the 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 Tiffany Veinot (2007), who followed one worker - a vault inspector at a hydroelectric utility company - in order to understand the information practices of 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 being sampled should fit the research question being asked: in some cases, small is best.

4. Not All Data Are Equivalent

Some researchers assume that analyses done with small data can be done better with Big Data. This argument also presumes that data is interchangeable. Yet, taken out of context, data lose meaning and value. Context matters. When two datasets can be modeled in a similar way, this does not mean that they are equivalent or can be analyzed in the same way. Consider, for example, the rise of interest in social network analysis that has emerged alongside the rise of social network sites (boyd & Ellison 2007) and the industry-driven obsession with the 'social graph'. Countless researchers have flocked to Twitter and Facebook and other social media to analyze the resultant social graphs, making claims about social networks.

The study of social networks dates back to early sociology and anthropology (e.g., Radcliffe-Brown 1940), with the notion of a 'social network' emerging in 1954 (Barnes) and the field of 'social network analysis' emerging shortly thereafter (Freeman 2006). Since then, scholars from diverse disciplines have been trying to understand people's relationships to one another using diverse methodological and analytical approaches. As researchers began interrogating the connections between people on public social media, there was a surge of interest in social network analysis. Now, network analysts are turning to study networks produced through mediated communication, geographical movement, and other data traces.

However, the networks produced through social media and resulting from communication traces are not necessarily interchangeable with other social network data. Just because two people are physically co-present – which may be made visible to cell towers or captured through photographs – does not mean that they know one another. Furthermore, rather than indicating the presence of predictable objective patterns, social network sites facilitate connectedness across structural boundaries and act as a dynamic source of change: taking a snapshot, or even witnessing a set of traces over time does not capture the complexity of all social relations. As Kilduff and Tsai (2003, p. 117) note, 'network research tends to proceed from a naive ontology that takes as unproblematic the objective existence and persistence of patterns, elementary parts and social systems.' This approach can yield a particular kind of result when analysis is conducted only at a fixed point in time, but quickly unravels as soon as broader questions are asked (Meyer et al. 2005).

Historically speaking, when sociologists and anthropologists were the primary scholars interested in social networks, data about people's relationships was collected through surveys, interviews, observations, and experiments. Using this data, social scientists focused on describing one'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 a mediating technology (boyd 2004). There are three common reasons in which people articulate their connections: to have a list of contacts for personal use; to publicly display their connections to others; and to filter content on social media. These articulated networks take the form of 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 (Meiss *et al.* 2008; Onnela *et al*. 2007). 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 often contested, the concept of 'tie strength' is understood to indicate the importance of individual relationships (Granovetter, 1973). When a person chooses to list someone as their 'Top Friend' on MySpace, this may or may not be their closest friend; there are all sorts of social reasons to not list one's most intimate connections first (boyd, 2006). Likewise, when mobile phones recognize that a worker spends more time with colleagues than their spouse, this

does not necessarily mean that they have stronger ties with their colleagues than their spouse. 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.

Fascinating network analysis can be done with behavioral and articulated networks. But there is a risk in an era of Big Data of treating every connection as equivalent to every other connection, of assuming frequency of contact is equivalent to strength of relationship, and of believing that an absence of connection indicates a relationship should be made. Data is not generic. There is value to analyzing data abstractions, yet the context remains critical.

5. Just Because it is Accessible Doesn't Make it Ethical

In 2006, a Harvard-based research project 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). This supposedly anonymous data was released to the world, allowing other researchers to explore and analyze it. What other researchers quickly discovered was that it was possible to de-anonymize parts of the dataset: compromising the privacy of students, none of whom were aware their data was being collected (Zimmer 2008).

The case made headlines, and raised a difficult issue 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 twenty 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 2010).

Even when researchers try to be cautious about their procedures, they are not always aware of the harm they might be causing in their research. For example, a group of researchers noticed that there was a correlation between self-injury ('cutting') and suicide. They prepared an educational intervention seeking to discourage people from engaging in acts of self-injury, only to learn that their intervention prompted an increase in suicide attempts. For some, self-injury was a safety valve that kept the desire to attempt suicide at bay. They immediately ceased their intervention (Emmens & Phippen 2010).

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 scholars to think critically about the ethics of their research.

With Big Data emerging as a research field, little is understood about the ethical implications of the research being done. 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 consent look like?

It may be unreasonable to ask researchers to obtain consent from every person who posts a tweet, but it is unethical for researchers to justify their actions as ethical simply because the data is accessible. Just because content is publicly accessible doesn't mean that it was meant to be consumed by just anyone (boyd & Marwick, 2011). 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 is seemingly accessible. Researchers must keep asking themselves – and their colleagues – about the ethics of their data collection, analysis, and publication.

In order to act in an ethical manner, it is important that scholars reflect on the importance of accountability. In the case of Big Data, this means both accountability to the field of research, and accountability to the research subjects. Academic researchers are held to specific professional standards when working with human participants in order to protect their 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 to the field and to human subjects required rigorous thinking about the ramifications of Big Data, rather than assuming that ethics boards will necessarily do the work of ensuring people are protected. Accountability here is used as a broader concept that privacy, as Troshynski *et al.* (2008) have outlined, where the concept of accountability can apply even when conventional expectations of privacy aren't 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).

There are 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 was created in highly context-sensitive spaces, and it is entirely possible that some social media 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, neither are users 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. There is a

considerable difference between being in public and being public, which is rarely acknowledged by Big Data researchers.

6. Limited Access to Big Data Creates New Digital Divides

In an essay on Big Data, Scott Golder (2010) quotes sociologist George 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 Lev 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; other sell the privilege of access for a high 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 those who went to the top universities and the rest 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 (Forsythe 2001; Harding 1989). 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 produces 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 it. Big Data researchers with access to proprietary data sets are less likely to choose questions that are contentious to a social media company, for example, 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 shouldn't bother studying social media - as in-house people can do it so much better.[4] Such explicit efforts to demarcate research 'insiders' and 'outsiders' – while by no means new – undermine the utopian rhetoric of those who evangelize about the values of Big Data. 'Effective democratisation can always be measured by this essential criterion,' Derrida claimed, 'the participation in and access to the archive, its constitution, and its interpretation' (1996, p. 4). Whenever inequalities are explicitly written into the system, they produce class-based structures. Manovich 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' (2011). 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 much broader 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 Lucy Suchman (2011) observes, via Levi Strauss, 'we are our tools.' We should consider how they 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.

[4] During his keynote talk at the International Conference on Weblogs and Social Media (ICWSM) in Barcelona on July 19, 2011, Jimmy Lin – a researcher at Twitter – discouraged researchers from pursuing lines of inquiry that internal Twitter researchers could do better given their preferential access to Twitter data.

Acknowledgements

We wish to thank Heather Casteel for her help in preparing this article. We are also deeply grateful to Eytan Adar, Tarleton Gillespie, and Christian Sandvig for inspiring conversations, suggestions, and feedback.

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, <http://www.edge.org/3rd_culture/anderson08/ anderson08_index.html>. [25 July 2011]

Baca, G. (2004) 'Legends of Fordism: Between Myth, History, and Foregone Conclusions', Social Analysis, vol. 48, no.3, pp. 169-178.

Barnes, J. A. (1954) 'Class and Committees in a Norwegian Island Parish', Human Relations, vol. 7, no. 1, pp. 39–58.

Barry, A. and Born, G. (2012) Interdisciplinarity: reconfigurations of the Social and Natural Sciences. Taylor and Francis, London.

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

Berry, D. (2011) 'The Computational Turn: Thinking About the Digital Humanities', Culture Machine. vol 12. <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, New York.

Bollier, D. (2010) 'The Promise and Peril of Big Data', <http:// www.aspeninstitute.org/sites/default/files/content/docs/pubs/ The_Promise_and_Peril_of_Big_Data.pdf>. [11 July 2011].

boyd, d. (2004) 'Friendster and Publicly Articulated Social Networks', Conference on Human Factors and Computing Systems (CHI 2004). ACM, April 24-2, Vienna.

boyd, d. (2006) 'Friends, Friendsters, and Top 8: Writing community into being on social network sites', First Monday vol. 11, no. 12, article 2.

boyd, d. and Ellison, N. (2007) 'Social Network Sites: Definition, History, and Scholarship', Journal of Computer-Mediated Communication, vol. 13, no.1, article 11.

boyd, d. and Marwick, A. (2011) 'Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies,' paper given at Oxford Internet Institute Decade in Time Conference. Oxford, England.

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

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

Cain, M. and Finch, J. (1981) Towards a Rehabilitation of Data. In: P. Abrams, R. Deem, J. Finch, & P. Rock (eds.), Practice and Progress: British Sociology 1950-1980, George Allen and Unwin, London.

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

Crawford, K. (2009) 'Following you: Disciplines of listening in social media', Continuum: Journal of Media & Cultural Studies vol. 23, no. 4, 532-33.

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

Dunbar, R. (1998) Grooming, Gossip, and the Evolution of Language, Harvard University Press, Cambridge.

Derrida, J. (1996) Archive Fever: A Freudian Impression. Trans. Eric Prenowitz, University of Chicago Press, Chicago & London.

Emmens, T. and Phippen, A. (2010) 'Evaluating Online Safety Programs', Harvard Berkman Center for Internet and Society, <http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Emmens_Phippen_Evaluating-Online-Safety-Programs_2010.pdf>. [23 July 2011].

Ess, C. (2002) 'Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee,' Association of Internet Researchers, <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.

Gitelman, L. (2011) Notes for the upcoming collection 'Raw Data' is an Oxymoron, <https://files.nyu.edu/lg91/public/>. [23 July 2011].

Golder, S. (2010) 'Scaling Social Science with Hadoop', Cloudera Blog, <http://www.cloudera.com/blog/2010/04/scaling-social-science-with-hadoop/>. [June 18 2011].

Granovetter, M. S. (1973) 'The Strength of Weak Ties,' *American Journal of Sociology* vol. 78, issue 6, pp. 1360-80.

Harding, S. (2010) 'Feminism, science and the anti-Enlightenment critiques', in Women, knowledge and reality: explorations in feminist philosophy, eds A. Garry and M. Pearsall, Boston: Unwin Hyman, 298–320.

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

Isbell, C., Kearns, M., Kormann, D., Singh, S. & Stone, P. (2000) 'Cobot in LambdaMOO: A Social Statistics Agent', paper given at the 17th National Conference on Artificial Intelligence (AAAI-00). Austin, Texas.

Kilduff, M. and Tsai, W. (2003) Social Networks and Organizations, Sage, London.

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, London: Routledge, pp. 145-162.< 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, pp. 721-3.

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, pp. 330-342.

Manovich, L. (2011) 'Trending: The Promises and the Challenges of Big Social Data', Debates in the Digital Humanities, ed M.K.Gold. The University of Minnesota Press, Minneapolis, MN <http://www.manovich.net/DOCS/Manovich_trending_paper.pdf>.[15 July 2011].

McCloskey, D. N. (1985) 'From Methodology to Rhetoric', In The Rhetoric of Economics au D. N. McCloskey, 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., and A. Vespignani. (2008) 'Structural analysis of behavioral networks from the Internet', Journal of Physics A: Mathematical and Theoretical, vol. 41, no. 22, pp. 220-224.

Meyer D, Gaba, V., Colwell, K.A., (2005) 'Organizing Far from Equilibrium: Nonlinear Change in Organizational Fields', Organization Science, vol. 16, no. 5, pp.456-473.

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, NY.

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.

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

Suchman, L. (2011) 'Consuming Anthropology', in Interdisicpinarity: Reconfigurations of the social and natural sciences, eds Andrew Barry and Georgina Born, Routledge, London and New York.

Twitter. (2011) 'One hundred million voices', Twitter blog, <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.

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