International Journal of Communication 8 (2014), 1663–1672 1932–8036/20140005

IJoC Big Data, Big Questions

# Critiquing Big Data: Politics, Ethics, Epistemology

## *Special Section Introduction*

KATE CRAWFORD
Microsoft Research, New England, USA
MIT Center for Civic Media, USA
NYU Information Law Institute, USA

KATE MILTNER
Microsoft Research, USA

MARY L. GRAY
Microsoft Research New England, USA
Indiana University, USA

*Why now?* This is the first question we might ask of the big data phenomenon. Why has it gained such remarkable purchase in a range of industries and across academia, at this point in the 21st century? *Big data* as a term has spread like kudzu in a few short years, ranging across a vast terrain that spans health care, astronomy, policing, city planning, and advertising. From the RNA bacteriophages in our bodies to the Kepler Space Telescope, searching for terrorists or predicting cereal preferences, big data is deployed as the term of art to encompass all the techniques used to analyze data at scale. But why has the concept gained such traction now?

### The Technical Terrain

A common answer is that big data approaches have been produced by our current technological capacities: that the steady drumbeat of Moore's Law, the doubling of integrated circuits every two years, has brought us to a point where massive amounts of data can be easily gathered, stored, analyzed and interlinked. Apache's Hadoop, a common big data platform that utilizes distributed nodes to act as processing and analysis clusters, was launched in 2005—reinforcing the idea of big data's newness. Yet

Kate Crawford: kate@katecrawford.net
Kate Miltner: a-kamilt@microsoft.com
Mary Gray: mlg@microsoft.com
Date submitted: 2013–04–10

this isn't satisfying; it overemphasizes the technological element of our historical conjuncture and fails to account for the economic, political, and cultural forces at work. And, in fact, the term *big data* has been around for almost two decades, and yet it has only been in the last five years that it has acquired such popular resonance. In its earliest incarnations in the archive of the Association for Computing Machinery, the concept of big data simply referred to data sets that were too large for any single computer: "Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data" (Cox & Ellsworth, 1997, p. 235).

This definition of the 'problem' of big data as being about storage and compression was well established in the mid-20th century, as RAND was developing its Relational Data File, a system designed for "the logical analysis of large collections of factual data." In November 1967, two computer scientists were struggling with the difficulty of working with large data sets, and they discovered "a wide variety of problems—logical and linguistic, hardware and software, practical and theoretical" that plagued their endeavors (Levien & Maron, 1967, p. 715).

And these difficulties persist. As cloud computing platforms distribute storage, the pipes to move data remain expensive and limited and storage remains problematic. Jonathan Sterne notes that "further increases in computing power and bandwidth may be used for higher definition, but they will also be used for more elaborate compression schemes . . . there will be no post-compression age" (2012, p. 231). In other words, big data is neither new nor free of the technical challenges raised since the emergence of supercomputers. Just as the Large Hadron Collider, the poster child for truly large data projects, has turned to magnetic tape as its best data storage solution, big data contains the techniques, artifacts, and challenges of older computational forms. Older concerns—technical, epistemological, and ethical—haunt the domains of big data.

## Big Data as Political, Economic and Cultural

If we are to reject the claim that the "big data moment" has been precipitated by technology alone, then we need to look more widely at the cultural, political, and economic making of big data: as a science, a business, and—importantly—as a mythology (boyd & Crawford, 2012). This cultural mythology can be seen in city billboards promoting "big data solutions," at highly profitable big data conferences, and in the many newspaper and magazine columns covering the advances brought about by big data science. The very term *big data science* is itself a kind of mythological artifact: implying that the precepts and methods of scientific research change as the data sets increase in size. Some big data fundamentalists argue that at sufficient scale, data is enough; "statistical algorithms find patterns where science cannot" (Anderson, 2008, para. 14), and thus big data represents "the end of theory" (Graham, 2012). But we argue that big data *is* theory. It is an emerging *Weltanschauung* grounded across multiple domains in the public and private sectors, one that is need of deeper critical engagement.

The mythic power of big data is part of what unifies it as a concept and informs its legibility as a set of tools. And this is, of course, not a novel claim. As Donna J. Haraway first wrote in 1983:

> The boundary is very permeable between tool and myth, instrument and concept, historical systems of social relations and historical anatomies of possible bodies, including objects of knowledge. Indeed, myth and tool mutually constitute each other. (1983, para. 7)

Haraway writes of this "common move" in the technological sciences, of translating the world into a problem of coding, where differences in the world are "submitted to disassembly, reassembly, investment, and exchange" in what she describes as "the informatics of domination" (1991, pp. 302–303). We can see how the process of coding the world has progressed in the multiple prehistories of big data, from the first instantiations of the U.S. census tracking human populations (Driscoll, 2012), the quantification of climate shifts (Edwards, 2010), the rapid analysis and projection of financial data (MacKenzie, 2006), and the complete capture of communications systems by intelligence agencies (Landau, 2010).

This special section of the *International Journal of Communication* brings together critical accounts of big data as theory, practice, archive, myth, and rhetorical move. The essays collected here interrogate rather than accept the realities conjured through our political, economic, and cultural imaginings of big data. From neoliberal economic logics shaping the deployment of big data to the cultural precursors that underlie data mining techniques, the issue covers both the macro and micro contexts. We have drawn together researchers from communication, anthropology, geography, information science, sociology, and critical media studies to, among other things, examine the political and epistemological ramifications of big data for a range of audiences. Articles in this collection also interrogate the ethics of big data use and critically consider who gets access to big data and how access (or lack of it) matters to the issues of class, race, gender, sexuality, and geography.

But, importantly, this is not a wholesale rejection of big data: Several of the authors presented here use big data as tools and techniques in their everyday work. By analyzing big data's applications, methods, and assumptions, they aim to improve the way social and cultural research is done. The already tired binary of big data—is it good or bad?—neglects a far more complex reality that is developing. There is a multitude of different—sometimes completely opposed—disciplinary settings, techniques, and practices that still assemble (albeit uncomfortably) under the banner of big data. Fields engaged in media and communication research that draw on big data to address dilemmas or raise new questions push us to carefully consider the ways in which the term and techniques are deployed. This is particularly necessary given that the big data nomenclature has generated nationally funded multi-billion-dollar grant programs and tenure-track jobs across academe; it is the megafauna of the academic landscape. The rapid and widespread ascendancy of the concept attests to its significance and "stickiness" across multiple fields—it has become a "thing," despite the ways in which the term is often at odds with itself semantically and industrially. As Tom Boellstorff (2013) suggests, "there is no unitary phenomenon 'big data' . . . yet the impact of big data is real and worthy of sustained attention" (para. 2).

## Ethical Challenges

Another way to consider the "why now?" question is to ask *who* and *what* is subjected to analysis. For decades, the "informatics of domination" have been focused and tested on historically marginalized groups. As Virginia Eubanks (2014) has shown, drawing on her ethnographic research of Electronic Benefit Transfer card and food stamp use in the United States, "poor and working-class Americans already live in the surveillance future" (para. 3). Thanks to the archive released by Edward Snowden, it is now public knowledge that consentless big data gathering is out of its testing phase and has been widely franchised to the mass populace. This has also demonstrated the extent of the erosion of civil liberties and privacy. However, in the case of scientific inquiry, it also raises the question of how big data tools should be used. How can data be gathered without people's knowledge or consent and still meet the ethical obligation to treat people with "justice, beneficence, and respect," as the Belmont Report on human subjects research first outlined in 1978? Scientific research that involves drawing on what is euphemistically known as "passively collected" big data must face difficult questions and develop new ethical frameworks. This is particularly urgent given the leading professional bodies for computing and engineering, the ACM and IEEE, both have ethical guidelines that are almost two decades old.

Snowden's trove of documents exposed to the public that the e-mails, phone calls, text messages, and social media activity of millions of people around the worlds are collected and stored, that enormous cloud servers have been breached, and both data and metadata have been fair game. But it also revealed the driving economic imperatives: Big data's promise is economic efficiency, more observation at less cost. Bankston and Soltani (2014) have shown in detail just how cheap mass surveillance is compared to hiring police officers: just 6.5 cents per hour to monitor a person electronically rather than $275 for a covert pursuit. The excitement about harnessing the promise of big data through the widespread collection of disparate online transactions and interactions coincides with its cost efficiency in targeting niche markets and providing oversight of populations.

In an "informatics of domination" that gathers all the data it can to unlock some presumed or as-yet-unknown value down the road, data generation and collection are equated with innovation and scientific breakthroughs. As such, participation in the big data project—offering up the data we generate through the social interactions that shape our everyday lives—becomes the "responsibility" of all good citizens. To contribute one's data to the pool is to contribute to the advancement of science, innovation, and learning. This rhetoric can be seen most clearly with regard to health data. To be concerned about individual risk is equated with hindering progress; why be concerned about releasing data if it could help others, in the aggregate? Of course, this fails to acknowledge the ways in which our data can reveal much about us that we cannot know or intend, and can be used to discriminate against individuals and groups. And how much trust should we have in the custodianship of data? The repositories of data are characteristically unstable; data is leaky, and it escapes in unexpected ways, be it through errors, hacks, or whistleblowing.

## Big Data Literatures

There is a strong celebratory thread in the literature on big data: that more data will bring better

science, safer cities, and rapid innovation. One such example is *The Human Face of Big Data* (Smolan & Erwitt, 2012), a collection of essays about the potential of large-scale data gathering to design personalized drugs, predict divorce, and research Parkinson's and retinal disease. Even India's controversial biometric ID card program, the Aadhaar card, is described in glowing terms, with no mention of privacy or ethics concerns, and only a brief mention of the information security risks. Likewise, in *Social Physics*, Pentland (2014) outlines his goal to gather "the digital bread crumbs we all leave behind as we move through the world—call records, credit card transactions, and GPS location fixes" (2014, 16) to predict who is more likely to get diabetes and "whether someone is the sort of person who will pay back loans" (p. 7). In his view, big data brings us closer to a probabilistic universe where human behavior can be predicted from metadata, to "build a society that is better at avoiding market crashes, ethnic and religious violence, political stalemates, widespread corruption, and dangerous concentrations of power" (p. 16). Similarly, Mayer-Schönberger and Cukier (2013) suggest that large data sets have the potential to replace the exactitude of causality for the "good enough" of correlation.

What these arguments fail to fully consider, however, is that data sets—including predictive data—may lead to new concentrations of power, and they are never methodologically removed from human design and bias (Crawford, 2013). Big data continues to present blind spots and problems of representativeness, precisely because it cannot account for those who participate in the social world in ways that do not register as digital signals. It is big data's opacity to outsiders and subsequent claims to veracity through volume that discursively neutralizes the tendency to make errors, fail to account for certain people and communities, or discriminate. The rhetoric of objectivity can be very seductive to public policy makers traversing the complex world of social phenomena. In Will Davies' (2013) terms, "data is being icily naturalized, with its institutional and methodological preconditions being marginalized from discussion" (para. 7). Indeed, the celebratory promises of big data as "good enough" to produce predictors of social behavior fundamentally ignore a key insight of social theory: Aggregated, individual actions cannot, in and of themselves, illustrate the complicated dynamics that produce social interaction—the whole of society is greater than the sum of its parts.

More critical and historical investigations are emerging that address how big data is being understood, operationalized, and resisted across the fields of media, computer science, law, and economics. Two collections in particular have addressed the issue of how big data is made. *Raw Data Is an Oxymoron*, a collection edited by Lisa Gitelman (2013), takes up this question by examining "the imagination of data" across various disciplines, eras, and media. These essays observe how data is generated and shaped, with the very definition of data changing across time and media, from newspaper clippings in the 1860s to the computational cloud. A *First Monday* special issue likewise argues that "data creation is a process that is extended in time and across spatial and institutional settings" (Helles & Jensen, 2013). This special section contributes to this growing critical conversation. These articles bring a nuanced and grounded analysis to engage big data practices, tools, and rhetorics directly and ask how they function, how they build interpretations, and how they could be different, more ethical, and more historically aware. This collection identifies three threads of inquiry.

■■■

**Methodological Interventions and Ethical Critiques**

To claim that the dynamics of human interaction and the complexity of the social world can be reduced to a self-explanatory set of nodes and edges defies important insights from fields as diverse as machine learning, sociology, and economics. Data sets are not, and can never be, neutral and theory-free repositories of information waiting to give up their secrets. They require the active interpretation of researchers, all of whom have their own ways of seeing. And in the words of historian James Clifford, "no one reads from a neutral or final position. This rather obvious caution is violated in new accounts that purport to set the record straight or to fill a gap in 'our' knowledge" (Clifford & Marcus, 1986, p. 18).

Three articles in this collection directly address methodological issues of sampling and bias, and they address the difficulty of producing representativeness from social media data. Lawrence Busch assesses how big data analyses fall prey to a range of problems, including distortion, errors, bias, and misinterpretation. He argues that, because these data sets are used to make policy decisions, the manner of their construction is particularly important, and he notes that there are "complex trade-offs between the size and apparent precision of the data set." Kevin Driscoll and Shawn Walker demonstrate that even the most robust and careful social data collection methods can produce skewed research results. They illustrate the importance of understanding how data access and technological infrastructure impact experimental outcomes, and they show that, even within privileged institutions that have "full" access to Twitter's database, small differences in timing or network connectivity can result in vastly different results for the same experiment. Driscoll and Walker note the danger in assuming that these types of studies are representative of larger-scale phenomena, particularly since Twitter has a relatively small user base.

Jim Thatcher contends with issues of representation with the "data fumes" produced by geolocative apps. He notes that only a small, mostly urban, and comparatively privileged percentage of the population regularly engages with platforms such as Foursquare. Even within the already-limited framework of these apps and their users, the data available for collection have been shaped by the affordances of the apps in question; these, in turn, have been determined by a very small, often homogenous group of developers. Thatcher warns that, for researchers using this data, "the very limits of knowledge are set through the data infrastructure of private corporations."

These articles make the key point that data sets available from social media platforms are inherently exclusionary, both because of the populations represented as well as the methodologies used to harness them. The authors challenge us to consider the ethical implications of using these types of data sets, particularly if they are used to make decisions—be it policy, planning, or resource allocation—that impact entire populations based on the data of the few.

**Social and Political Implications**

Questions of who is represented, and how, necessarily lead to the political and social implications of big data research. Who benefits most from big data techniques? How is the idea of the social reconstructed by big data logics? Mark Andrejevic approaches these questions by examining the once-heralded potential of the "daily me," and arguing that it is corporations, not us, that benefit from our data,

Andrejevic sees a "big data divide" between us and our data; not only are we rarely granted access our own data, we lack the capability to analyze it and make sense of it, particularly in the context of other users. Andrejevic argues that it is not simply access to data sets but the technologies, infrastructure, and expertise to analyze these data that reinforce power differentials between those who have the capacity to make use of big data and those who are simply part of the sorting process. Andrejevic observes that the users who opt in to data systems often feel powerless in regard to their participation. When users click "I agree," it is often with a sense of coercion; to access essential technologies, relinquishing control over their personal data is the price they must pay. In contrast to the purportedly democratizing effect of widespread Internet access and use, Andrejevic argues that the big data paradigm offers a re-entrenchment of societal power differentials, with big data being used by an elite few to make decisions of wide-ranging impact for the many.

Whereas Andrejevic argues that the promise of personalization has failed us individually, Nick Couldry and Joseph Turow argue that personalization has failed us collectively, albeit in a different way. In their analysis of the "deep personalization" of online content enabled by big data, Couldry and Turow argue that big data practices have the potential to undermine the public sphere. When content—journalistic or otherwise—is tailored to individuals based upon the needs of advertisers and content providers, it fractures the reference points necessary for a shared political and social life and "risks eliminating the connective media necessary for an effective democracy." Couldry and Turow remind us that the unexpected negative externalities that result from "successful" applications of big data analysis have the potential to undermine fundamental societal structures more than poorly implemented big data initiatives.

## Epistemological Challenges and Research Provocations

If researchers are to intervene in the debates outlined above, we must collectively invest in an explicit epistemological pluralism. This would mean scholars from a range of disciplines engaging in dialogue about how data shapes understanding and productively questioning the rhetoric of objectivity and claims to knowledge.

Cornelius Puschmann and Jean Burgess examine the metaphors used to describe big data across various publications reporting on the business and management, technology, news, and telecommunications sectors. They find that big data is explicated in two key ways: as a resource to be consumed and as a natural force to be controlled. Both of these metaphors position big data as reliable, value-neutral sources of information. However, far from innocuous framings to help explain a technical and complicated concept, the authors argue that by positioning big data in this way, these metaphors obscure the many ways that data—big or otherwise—are socially constructed, consequently reifying the notion that big data is somehow a source of objective truth.

Dawn Nafus and Jamie Sherman's essay on the data practices of the Quantified Self (QS) community encourages us to think about the role that individual agency plays when it comes to resisting dominant data logics. Several articles in this special section note how the process of data generation and collection affects big data analyses: the conditions for what we can know are shaped by what data is

recognized, how that data is collected, and by whom. QSers engage in self-monitoring, but they collect data to serve their own needs, often generating data and practices that confound the wishes of institutional data collectors. The QS community provides examples of a subtle subversion of the dominant data logic. Nafus and Sherman argue that even when data is being generated on the most intimate levels, individuals are not necessarily willing participants in the big data project and they complicate naturalistic epistemologies grounded in a straightforward empiricism.

Finally, Geoffrey Bowker's epilogue brings the collection full circle, questioning how an archive's size could ever serve as sufficient justification for certain beliefs or as a self-evident conveyor of truth claims. He argues that different levels of interpretation and scope are necessary, because data of any size do not operate in a social vacuum. Even though some databases are bigger than ever, they are still structured in ways that privilege certain ontologies and obscure others.

Above all, we need new critical approaches to big data that begin with deep skepticism of its a priori validity as a naturalized representation of the social world. We can make big data sets productive archives for theory building if we reimagine what big data offers us. Combining separate, often disparate, multiterabyte sets of information can reframe our understandings of people, institutions, and things. Rather than invest in big data as an all-knowing prognosticator or a shortcut to ground truth, we need to recognize and make plain its complexities and dimensionality as an emerging theory of knowledge.

**References**

Anderson, C. (2008, June 23). The end of theory. *Wired*, *16*. Retrieved from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Bankston, K. S., & Soltani, A. (2014, January 9). Tiny constables and the cost of surveillance: Making cents out of United States v. Jones. *Yale Law Journal, 123*. Retrieved from http://yalelawjournal.org/forum/tiny-constables-and-the-cost-of-surveillance-making-cents-out-of-united-states-v-jones

Boellstorff, T. (2013). Making big data, in theory. *First Monday*, *18*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4869/3750

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

Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture: The poetics and politics of ethnography*. Berkeley, CA: University of California Press.

Cox, M., & Ellsworth, D. (1997, October). Application-controlled demand paging for out-of-core visualization. In *Proceedings of the 8th conference on visualization '97* (pp. 235–244). Los Alamitos, CA: IEEE.

Crawford, K. (2013, April 1). The hidden biases in big data. *Harvard Business Review*. Retrieved from http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data

Davies, W. (2013, June 12). Data's double life. *Potlatch.* Retrieved from http://potlatch.typepad.com/weblog/2013/06/datas-double-life.html

Driscoll, K. (2012). From punched cards to "big data": A social history of database populism. *communication+ 1, 1*(1), 4.

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

Eubanks, V. (2014, January 15). Want to predict the future of surveillance? Ask poor communities. *The American Prospect*. Retrieved from http://prospect.org/article/want-predict-future-surveillance-ask-poor-communities

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron*. Cambridge, MA: MIT Press.

Graham, M. (2012, March 9). Big data and the end of theory? *The Guardian*. Retrieved from http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory

Haraway, D. J. (1983). The ironic dream of a common language for women in the integrated circuit: Science, technology, and socialist feminism in the 1980s or a socialist feminist manifesto for cyborgs. History of Consciousness Board, University of California at Santa Cruz. Retrieved from http://homepages.herts.ac.uk/~comqcln/HarawayCyborg.html

Haraway, D. J. (1991). A cyborg manifesto: Science, technology and socialist feminism in the late twentieth century. In D. J. Haraway, *Simians, cyborgs, and women: The reinvention of nature*. London, UK: Free Association Books

Helles, R., & Jensen, K. B. (2013). Introduction to the special issue "making data—big data and beyond." *First Monday*, *18*(10). Retrieved from http://firstmonday.org/article/view/4860/3748

Landau, S. E. (2010). *Surveillance or security? The risks posed by new wiretapping technologies*. Cambridge, MA: MIT Press.

Levien, R. E., & Maron, M. E. (1967). A computer system for inference execution and data retrieval. *Communications of the ACM, 10*(11) (November 1967), 715–721.

MacKenzie, D. A. (2006). *An engine, not a camera: How financial models shape markets*. Cambridge, MA: MIT Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Pentland, A. S. (2014). *Social physics: How good ideas spread—The lessons from a new science*. New York, NY: Penguin Books.

Smolan, R., & Erwitt, J. (2012). *The human face of big data*. Chicago, IL: Against All Odds Productions.

Sterne, J. (2012). *MP3: The meaning of a format*. Durham, NC: Duke University Press.

Tape rescues big data. (2013, September 26.) *The Economist.* Retrieved from http://www.economist.com/blogs/babbage/2013/09/information