2013

The Role of Big Data in the Social Sciences

Steven Ovadia
CUNY La Guardia Community College

Behavioral & Social Sciences Librarian, 32:130–134, 2013
Published with license by Taylor & Francis
ISSN: 0163-9269 print / 1544-4546 online
DOI: 10.1080/01639269.2013.787274

Routledge
Taylor & Francis Group

INTERNET CONNECTION

The Role of Big Data in the Social Sciences

STEVEN OVADIA
LaGuardia Community College, Long Island City, New York

Big Data is an increasingly popular term across scholarly and popular literature but lacks a formal definition (Lohr 2012). This is beneficial in that it keeps the term flexible. But it can be frustrating in that the term can have different meanings depending on the context. A McKinsey report defined it as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al. 2011, 1). Looking at Big Data from an academic perspective, boyd and Crawford (2012) define it as the interplay of technology, analysis, and mythology, with the mythology coming from "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" (663).

For librarians, Big Data represents a few important ideas. One idea is the idea of balancing accessibility with privacy. Librarians tend to want information to be as open and available as possible, but they also understand the importance of maintaining the privacy of the individual. Big Data also has tremendous implications for the social sciences. So while it is not a given that Big Data will impact everyone's research, it is safe to say it's a concept many academics are interested in learning about.

If Big Data has an air of magic or "mythology" to it, especially in the social science context, it's because of the impact it's had in the natural sciences: "The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven 'computational social science' has been much slower" (Lazer et al. 2009, 721). Which is not to say that social scientists have avoided the challenge of Big Data. There is quite a lot of sociology-based activity around Big Data sets, some of the activity from within the academy and some from outside of it.

Address correspondence to Steven Ovadia, LaGuardia Community College, Library Media Resources Center, 31–10 Thomson Avenue, Long Island City, NY 11101. E-mail: sovadia@lagcc.cuny.edu

Andreas Weigend, Amazon.com's former chief scientist, straddles both worlds, teaching and running a social data lab at Stanford, while also acting as a corporate consultant. Weigend expands the idea of social data beyond the analysis of services like Facebook and Twitter: "Social media is just a small part of the social data universe—one of the many data sources that represent the front end of the process. The back end is when you bring together the data from different sources" (Krivda 2011). The idea of combining data sets to create new insights is echoed by Prabhakar Raghavan, who, like Weigend, works in both the technology world, as a vice president at Google, and in the academic world, as a consulting professor at Stanford: "There is a huge opportunity to take the 'big data' that the computer scientists have access to and then combine it with the big problems that the social scientists contemplate. The opportunity here is the confluences of the two, to get to really interesting social insights that are statistically robust" (Mann 2012).

The National Science Foundation (NSF) also sees a need for social scientists to work with Big Data, issuing a report predicting "future research will be interdisciplinary, data-intensive, and collaborative," and then going on to mention four specific areas where data-intensive social science work seems to be congregating: population change; source of disparities; communication, language, and linguistics; and technology, new media and social networks (Gutmann and Friedlander 2011, 5). NSF support for big data in the social sciences expands to a program called "Building Community and Capacity for Data-Intensive Research in the Social, Behavioral, and Economic Sciences and in Education and Human Resources." One of the goals of the program is to develop data resources and methods of Big Data analysis (NSF).

Big Data is becoming increasingly important in the social sciences, but where does that leave librarians? Little (2012) foresees a role where, as has occurred historically, librarians serve as guides into the world of Big Data, helping to connect interested patrons with tools, and data sets, as well as assisting with archiving.

Librarians are also free to pursue their own work with Big Data sets, not just finding and archiving them, but also using these sets for their own research.

Another potential role for librarians is helping to facilitate the sharing of Big Data sets. Because Big Data often intersects between business and scholarship, there have been some challenges in terms of the public availability of certain data sets. In May 2012, *The New York Times* reported on the outcry at a conference where presenters from Google and the University of Cambridge kept their data set private (Markoff 2012). The practice of keeping data private has raised some concerns within the academic community: "Many of the emerging 'big data' come from private sources that are inaccessible to other researchers. The data source may be hidden, compounding

problems of verification as well as concerns about the generality of the results. These results are meaningful only if many other data sets reveal the same behavior" (Huberman 2012). Ravetz (2012) echoes Huberman, calling for "open science" and open data as a way to monitor and maintain research quality.

The issues of sharing data are complex, but there seem to be two main issues. One issue is businesses wanting to protect proprietary information that could potentially aid competitors. However, a larger and, for academics, more pressing issue is the question of privacy. As data sets become larger and more detailed, it becomes easier for individuals in those sets to be identified. One of the more memorable examples of this occurred when two University of Texas researchers were able to identify certain individuals in a massive data set from the Netflix film rental/streaming service by comparing the anonymous Netflix data to reviews posted on another site (Singel 2009). This revelation eventually led to a lawsuit against Netflix.

Researchers were also able to identify individuals with information contained within a genetic data set, using not only the genetic data, but also genealogy websites, as well as publicly available data (Kolata 2013). boyd and Crawford (2012) raised concerns about Big Data based upon public data: "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 available" (672).

This privacy concern has led to discussions on how academics can work with large data sets while still protecting the privacy of individuals within those data sets. Lane (2012) has proposed a cross-discipline community of practice that might develop best practices to share with the larger academic community, as well as training to help academics understand the privacy vulnerabilities of large data sets—vulnerabilities that might require a certain level of statistical expertise to understand and identify.

The privacy implications and concerns of Big Data also present another opportunity for librarians within the Big Data movement. Patron privacy is an important professional ethic for librarians, meaning most librarians not only understand why privacy is important, but can also articulately explain it, and demonstrate techniques to help protect the privacy of users. One could make the case the integrated library system is, depending upon the size of system, a large data set where the privacy of each individual must be protected.

Users interested in exploring public Big Data sets have a number of options. Integrated Public Use Microdata Series (IPUMS; www.ipums.org), which has been discussed in this space before (Ovadia 2010), includes detailed U.S. Census information going back to 1850, as well as international data sets. The project sponsor, the Minnesota Population Center, also has data sets from the American Time Use Survey-X and Integrated Health Interview Series.

Amazon Web Services, a cloud hosting service provided by the retailer, also hosts some public data sets, including Census information, genome data sets, and the content of the Freebase.com data project, a wide-ranging collection of data culled from public sources like Wikipedia and the Securities and Exchange Commission. The data sets can be accessed via https://aws.amazon.com/publicdatasets/.

Users might also have access to large data sets from their own institution. Many institutional repositories now include data sets and tools like The Dataverse Network (thedata.org/DVN-software), providing a free and open-source interface allowing institutions to host and make available their own data sets. As many institutions using Dataverse also use the name in the web address for their archive, a quick and easy way to survey Dataverse repositories is to do a Google search for inurl:dataverse.

For librarians, Big Data represents an opportunity to connect users to data sets. Just as many librarians perform a readers' advisor role, connecting users to titles they might enjoy, it is not out of the realm of possibility to imagine a day where libraries have a data advisor who connects users to data sets that will aid patrons in their research.

A final role for librarians to consider is as users and consumers of big data. These data sets that have captured the imagination of so many academics also have the ability to be tremendously useful to librarians, helping them to understand user behavior and choices. Librarians don't need to just preserve and find data sets. There's also an opportunity to use these data sets, too, learning anything and everything from patron interface preferences to preferred book formats.

Big Data may be a buzz word that could sound dated in a few years, but the movement and opportunity it represents to social scientists, librarians, and librarians who are also social scientists are much too important to dismiss. Many users are interested in working with large data sets, and the more librarians understand these sets, as well as their challenges, the better librarians can connect users to the data—big and small—that they need.

REFERENCES

boyd, danah, and Kate Crawford. 2012. Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society 15* (5): 662–79.

Gutmann, Myron P. and Amy Friedlander. 2011. *Rebuilding the mosaic: Fostering research in the social, behavioral, and economic sciences at the National Science Foundation in the next decade*. Arlington, VA: National Science Foundation, Directorate for Social, Behavioral, and Economic Sciences. http://www.nsf.gov/pubs/2011/nsf11086/nsf11086.pdf.

Huberman, Bernardo A. 2012. Correspondence: Big data deserve a bigger audience. *Nature 482* (7385): 308.

Kolata, Gina. 2013. Web hunt for DNA sequences leaves privacy compromised. *The New York Times*, January 17. http://www.nytimes.com/2013/01/18/health/search-of-dna-sequences-reveals-full-identities.html

Krivda, Cheryl D. 2011. Socialization of data. *Teradata Magazine 11* (2): 38–41.

Lane, Julia. 2012. O privacy, where art thou?: Protecting privacy and confidentiality in an era of big data access. *Chance 25* (4): 39–41.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, et al. 2009. Computational social science. *Science 323* (5915): 721–22.

Little, Geoffrey. 2012. Managing the data deluge. *Journal of Academic Librarianship 38* (5):263–64.

Lohr, Steve. 2012. How big data became so big. *The New York Times*, August 11. http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html.

Mann, Rebecca. 2012. Five minutes with Prabhakar Raghavan: Big data and social science at Google. *Impact of Social Sciences* (blog)*, London School of Economics and Political Science*, September 19. http://blogs.lse.ac.uk/impactofsocialsciences/2012/09/19/five-minutes-with-prabhakar-raghavan.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2012. *Big data: The next frontier for innovation, competition, and productivity*. New York, NY: McKinsey Global Institute. http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx.

Markoff, John. 2012. Troves of personal data, forbidden to researchers. *The New York Times*, May 21. http://www.nytimes.com/2012/05/22/science/big-data-troves-stay-forbidden-to-social-scientists.html.

National Science Foundation. 2012. Building community and capacity for data-intensive research in the social, behavioral, and economic sciences and in education and human resources (BCC-SBE/HER). *Directorate for Social, Behavioral & Economic Sciences*. Accessed January 23, 2012. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504747.

Ovadia, Steven. 2010. Finding data sets online. *Behavioral & Social Sciences Librarian 29* (1): 81–85.

Ravetz, Jerome. 2012. Keep standards high. *Nature 481* (7379): 25.

Singel, Ryan. 2009. Netflix spilled your Brokeback Mountain secret, lawsuit claims. *Threat Level (blog), Wired*, December 17, http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit.