The AAG Review of Books

ISSN: (Print) 2325-548X (Online) Journal homepage: www.tandfonline.com/journals/rrob20

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Ian M. Dunham

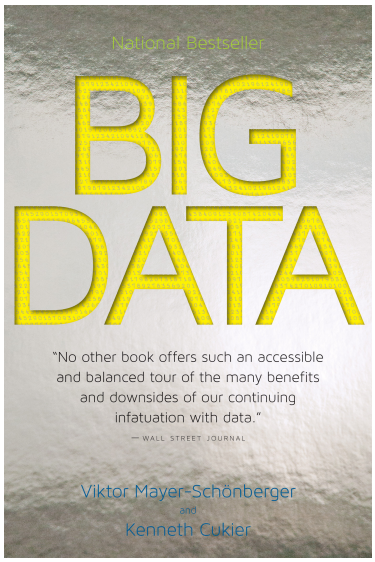
To cite this article: Ian M. Dunham (2015) Big Data: A Revolution That Will Transform How We Live, Work, and Think, The AAG Review of Books, 3:1, 19-21, DOI: 10.1080/2325548X.2015.985533

To link to this article: https://doi.org/10.1080/2325548X.2015.985533

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Viktor Mayer-Schönberger and Kenneth Cukier. Boston: Houghton Mifflin Harcourt, 2013. 242 pp., notes, biblography, index. $27.00 cloth (ISBN 978-0-544-00269-2); $14.95 electronic (ISBN 978-0-544-00293-7).

Reviewed by Ian M. Dunham, Department of Geography and Urban Studies, Temple University, Philadelphia, PA.



This volume by Mayer-Schönberger and Cukier weaves together real-world examples, the future potential, and the caveats of big data in an easily digestible format. A common catchphrase today, big data is an all-encompassing concept that refers to the ability to analyze huge quantities of information and make valuable conclusions that otherwise could not be revealed using smaller data and traditional methods. The book is approachable from the onset and maintains reader interest while espousing the scientific, societal, and economic potential of big data.

So what is all the hype about? Recent advances in the ability to quantify, store, and analyze data have culminated to make big data possible. The first advancement is "datification." The digitization of information gathered from communications devices and digital sensors allows for the quantification of many phenomena that have never before been rendered into data. The authors briefly describe how technological innovations—from Eratosthenes, Ptolemy, and Mercator, to GPS and Google—have allowed geographic location to become georeferenced and thus become data. The second advancement is the exponential growth of data storage capacity and an accompanying reduction in cost. In the year 2000 only about a quarter of the stored information in the world was digital and the rest was analog (paper, books, film, photographs, cassette tapes, etc.). In 2007 only about 7 percent of data were analog. In 2013 it is estimated that over 98 percent of all data are digital. Lastly, increasingly powerful computer processing capability and sophisticated algorithms now allow information to be analyzed in new and interesting ways.

In 2009 researchers at Google published a paper in *Nature* that showed that they were able to track influenza outbreaks using search records and historical data from the Centers for Disease Control and Prevention (CDC). The CDC gathers information from hospitals and clinics on individual patient flu cases, but there is a lag time of one to two weeks in their reporting. The Google team created an algorithm that compared search queries with CDC data and ran nearly half a billion calculations to identify forty-five search terms that have a strong correlation with flu outbreaks. Google now has the ability to track influenza outbreaks with surprising accuracy in near real time, augmenting traditional CDC reporting.

Big data is already a major source of competitive advantage for existing companies, and new startups are emerging to use big data in innovative ways. Whether or not users are aware of it, machine-generated content is present in a number of predictive e-commence examples: an alternative spelling for your web search, the next book you might be interested in reading, what advertisements are of interest to you, or the ideal time to purchase a low-price plane ticket. The authors remain eruditely grounded by offering examples of existing and potential uses of big data to substantively address issues such as climate change, conflict prevention, decision making, disease

The AAG Review of Books 3(1) 2015, pp. 19–21. doi: 10.1080/2325548X.2015.985533.

eradication, economic development, good governance, and the management of cities.

Three chapters are dedicated to addressing big data's potential to fundamentally change the very nature of how statistical analysis is carried out through three interlinked and self-reinforcing shifts. The first shift is the ability to analyze vast amounts of data rather than relying on traditional statistical sampling. For the better part of the twentieth century, statisticians adhered to the fundamental conviction that statistical inferences improve most dramatically, not with increased sampling size, but with increased randomness of the sample. After a certain point relatively early on, as the sample size increases, the benefit to accuracy of increasing the sample size becomes less. Big data allows a fundamental move away from random sampling to using all of the data, "N=all," which might reveal new details and connections.

The second shift is a willingness to embrace "real-world messiness" and a loosening of the desire for exactitude. As a data set increases in size, the likelihood of errors increases and analysts are less likely to clean every datum point. As the sample size increases, however, shortcomings in accuracy, exactitude, and cleanliness are overcome. A sample that is large and messy might yield more accurate results than a sample that is few and exact. Contrary to attempting to confirm the richest findings by using the smallest amount of exact data, a fundamental tenet of big data is that, as more data are analyzed, predictive performance is improved.

The third shift facilitated by big data is a growing respect for correlation rather than causality, or a move away from attempting to understand why something is happening to simply leaning what it is happening. One tenet of behavioral economics is that humans have a desire to establish causes where they might not exist. Kahneman's (2013) landmark book *Thinking, Fast and Slow* delineates fast and slow modes of thinking to which humans are prone. When thinking fast, imaginary causal relationships, "cognitive shortcuts," are made that lead to misunderstandings about the world. Mayer-Schönberger and Cukier argue that big data presents an opportunity to prevent cognitive biases that could lead to incorrect causal intuitions.

In 2008 the editor-in-chief of *Wired* magazine made the bold statement that big data has the potential to render the scientific method obsolete (Anderson 2008). His argument is that, whereas the traditional process of scientific discovery seeks to establish causality through the scientific method, big data allows research to be driven by the abundance of data rather than by theory. Big data presents a fundamentally different approach that begins with data collection, then analysis, and then drawing conclusions from the patterns that appear. Establishing causality is, of course, desirable, but it requires expert knowledge, theory, and the testing of a hypothesis to prove results. This is time consuming, and if causality is established, the results might reflect the biases of the researcher, omit other causes, and remain limited to only being able to answer the question being asked. Although Mayer-Schönberger and Cukier do not believe that correlation will completely replace causality, by "letting the data speak," new correlations might be revealed that point the way toward future hypothesis testing that examines causal relationships.

Notably the three reinforcing shifts surrounding big data might represent a detraction from reliance on expert authoritative knowledge, as subject-area experts will need to make way for generalists who know how to utilize big data. Although latent knowledge and the application of theory by experts will always have a rightful place, generalists are more likely to remain impartial to the theories of a specified field, thus possessing the ability to reach findings without prejudgments. In today's early stages of big data, the ideas and skills of those able to use big data hold the greatest value; however, as data become a commodity, and data holders see it as an asset to be monetized, data ownership and democratization of data will become more significant issues.

The idealism surrounding big data is tempered by concerns about potential negative implications. One downside is that these new technologies present opportunities to undermine privacy. Big data introduces challenges to the three core strategies—individual notice and consent, opting out, and anonymization—currently employed to ensure privacy. The authors also raise concerns about the potential for overreliance on data when making important decisions. As the Scott (1998) argues in his book *Seeing Like a State*, quantification serves the powerful, and government's desire for data collection and quantifiable order could have adverse impacts on citizens. Problems arise when decision makers rely solely on data and omit vital knowledge about the subtleties of citizens and processes. A more extreme example of the danger of overreliance in the predictive power of big data is that if big data could predict that an individual will commit a crime before it happens, people could be punished for their propensity to commit a crime. This raises concerns about the ethical considerations of the role of free will versus the dictatorship of data.

Given the aforementioned concerns, new principles are required to govern the use of big data and protect individual rights and freedoms. Johannes Gutenburg's invention of movable type around 1450 catalyzed the spread of new ideas, and the technology ultimately raised new legal issues including censorship, ownership of the printed word, and freedom of speech. Similarly, big data will require paradigmatic changes in governance. The authors call for accountable usage of data, and a transfer of the responsibility of privacy away from individuals and toward data users.

Overall, this book provides a lively and accessible introduction to big data. The breadth of examples of how those in academia, governance, and the private sector might benefit from the increased ability to analyze vast amounts of data are widely appealing. A thorough discussion of the potential negative implications makes for a balanced debate. Given the short hype cycle of catchphrases in Silicon Valley, the long-term sticking power of big data is yet to be determined; however, Mayer-Schönberger and Cukier are enthusiastic that the most transformative benefits are yet to come.

References

Anderson, C. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired* 16 (July).

Kahneman, D. 2013. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Scott, J. C. 1998. *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven, CT: Yale University Press.