EDITORS' INTRODUCTION

Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum

Massive amounts of data on human beings can now be accessed and analyzed. And the new 'big data'[1] are much more likely to be harvested from a wide variety of different sources. Much has been made of the many uses of such data for pragmatic purposes, including selling goods and services, winning political campaigns, and identifying possible terrorists. Yet big data can also be harnessed to serve the public good in other ways: scientists can use new forms of data to do research that improves the lives of human beings; federal, state, and local governments can use data to improve services and reduce taxpayer costs; and public organizations can use information to advocate for public causes, for example.

Much has also been made of the privacy and confidentiality issues associated with access. Statisticians are not alone in thinking that consumers should worry about privacy issues, and that an ethical framework should be in place to guide data scientists;[2] the European Commission and the U.S. government have begun to address the problem. Yet there are many unanswered questions. What are the ethical and legal requirements for scientists and government officials seeking to use big data to serve the public good without harming individual citizens? What are the rules of engagement with these new data sources? What are the best ways to provide access while protecting confidentiality? Are there reasonable mechanisms to compensate citizens for privacy loss?

The goal of this book is to answer some of these questions. The book's authors paint an intellectual landscape that includes the legal, economic, and statistical context necessary to frame the many privacy issues, including the value to the public of data access, clarifying personal data ownership questions, and raising issues of agency in personal data. The authors also identify core practical approaches that use new technologies to simultaneously maximize the utility of data access while minimizing information risk. As is appropriate for such a new and evolving field, each chapter also identifies important questions that require future research.

The work in this book is also intended to be accessible to an audience broader than those in the research and policy spheres. In addition

to informing the public, we hope that the book will be useful to people trying to provide data access within confidentiality constraints in their roles as data custodians for federal, state, and local agencies, or decision makers on Institutional Review Boards.

Historical and Future Use of Data for the Public Good

Good data are critically important for good public decisions. For example, national and international government policies depend on GDP estimates – indeed, international crises have been exacerbated when statistical agencies have cooked the data books.[3] Good data are also important for good science – as Daniel Kahneman famously pointed out, the first big breakthrough in our understanding of the mechanism of association was an improvement in a method of measurement.[4]

Historically the leading producers of high-quality data were statistical government agencies engaged in collecting data through large-scale statistically representative surveys. There are several reasons for this. One was the sheer scale of the necessary activity: generating representative samples required an expensive, constantly updated population frame and extensive investments in survey methodology and data storage, cleaning, and dissemination. The second was that the public trusted the government to protect confidentiality, and statistical agencies invested heavily in the appropriate statistical disclosure limitation methodologies. The third was that the statistical agencies were seen to be objective, and not trying to sell a product. The U.S. Census Bureau's mission statement reflects all three of these reasons:

> The Census Bureau's mission is to serve as the **leading source of quality** data about the nation's people and economy. We honor **privacy, protect confidentiality**, share our expertise globally, and conduct our work openly. We are guided on this mission by **scientific objectivity**, our strong and capable workforce, our devotion to research-based innovation, and our abiding commitment to our customers.[5] (*Emphases added*)

The public good has clearly been served by the creation and careful dissemination of data by both the government and the research community. Of course, the nature of the data, as well as the dissemination modality, has evolved over time. The development and release of large-scale public use datasets like the Current Population Survey and, later, the National Longitudinal Surveys of Youth and the Panel Study of Income Dynamics and the German Socio-Economic Panel have transformed our

understanding of labor markets while protecting respondent confidentiality. The development of large-scale administrative datasets and their access through secure data enclaves have lowered costs, increased sample size, and reduced respondent burden,[6] as well as created completely new classes of information.[7]

Big data, by which we mean the data necessary to support new types of data-intensive research,[8] hold the promise of even more profound contributions to the public good. Knowledge derived from big data is likely to become one of the foundational elements in the functioning of society by, for example, generating real-time information about economic and social activity or generating new insights into human behavior.[9] Yet the pathway to developing this foundation is not clear, because experience provides little guidance. The data currently used to inform decisions – survey and administrative data – have benefited from decades of statistical research, as well as clear rules defining ownership and responsibility. Statistical agencies, the primary custodians, have developed clear ways to both protect and access the data. By contrast, the value of big data in informing evidence-based policy is still being established, and the ownership of big data, typically distributed across multiple entities, is not well defined. Big data have many elements of a natural resource, and sensible rules must be developed in order to avoid a tragedy of the commons, and to create a commonly pooled resource for improving scientific understanding for the public good.[10]

Privacy, Big Data, and the Public Good: The Contributions of This Book

The vast changes in the data world have brought with them changes in the nature of data collectors and producers. Data on human beings, though now more likely to be collected, are much *less* likely to be purposively collected by researchers and government agencies, and are thus less often held by organizations with traditional knowledge about how to protect privacy. There are serious consequences: the lack of dissemination experience of non-governmental collectors can lead to massive privacy breaches (the 2006 AOL data release is but one famous example).[11] Even worse, if no dissemination is allowed, the quality of privately held data is largely unknown[12] absent detailed researcher inspection and validation. Similarly, Institutional Review Boards with few reference guidelines are likely to slow or prevent research on human subjects with complex data.[13]

https://doi.org/10.1017/CBO9781107590205.001 Published online by Cambridge University Press

Because of the importance of the topic, there is a rich and vibrant literature; the contributors to this book have provided, for the first time in one place, an accessible summary of existing research on many of the important aspects of balancing access to data with protection of privacy. They have also identified practical suggestions – to help guide practitioners and Institutional Review Boards – and identified important areas for future research.

Opening Part I, on the conceptual framework, *Strandburg* argues that the acquisition, transfer, and aggregation of data on a massive scale for data mining and predictive analysis raises questions that simply are not answered by the paradigms that have dominated privacy law to date. She develops a taxonomy of current U.S. privacy law and uses that taxonomy to elucidate the mismatch between current law and big data privacy concerns. *Barocas and Nissenbaum* argue that big data involves practices that have radically disrupted entrenched information flows. From modes of acquisition to aggregation, analysis, and application, these disruptions affect actors, information types, and transmission principles. Privacy and big data are simply incompatible, and the time has come to reconfigure choices that we made decades ago to enforce certain constraints. They argue that it is time for the background of rights, obligations, and legitimate expectations to be explored and enriched so that notice and consent can do the work for which it is best suited. *Acquisti* discusses how the economics and behavioral economics of privacy can be applied to investigate the implications of consumer data mining and business analytics. An important insight is that personal information, when shared, can become a public good whose analysis can reduce inefficiencies and increase economic welfare; when abused, it can lead to transfer of economic wealth from data subjects to data holders. The interesting economic question then becomes who will bear the costs if privacy-enhancing technologies become more popular in the age of big data: data subjects (whose benefits from business analytics and big data may shrink as they restrict the amount of information they share), data holders (who may face increasing costs associated with collecting and handling consumer data), or both?

There are practical implications. *Ohm* provides an overview of how information privacy laws regulate the use of big data techniques, if at all. He discusses whether these laws strike an appropriate balance between allowing the benefits of big data and protecting individual privacy and, if not, how the laws might be extended and amended to better strike this balance. He notes that most information privacy law focuses on collection or disclosure and not use. Once data has been legitimately obtained, few

laws dictate what may be done with the information. The chapter offers five general proposals for change. *Stodden* sets out the scientific rationale for access to data and computational methods, to enable the verification and validation of published research findings. She describes the legal landscape in the context of big data research and suggests two guiding principles to facilitate reproducibility and reuse of research data and code within and beyond the scientific context.

Koonin and Holland open Part II, on the practical framework, by addressing the motivations of the new urban science and the value for cities in big data – particularly with respect to analysis of the infrastructure, the environment, and the people. They discuss the key technical issues in building a data infrastructure for curation, analytics, visualization, machine learning, and data mining, as well as modeling and simulation to keep up with the volume and speed of data. *Goerge* describes the creation of a data warehouse that links data on multiple services provided by the public sector to individuals and families as a way to highlight both the opportunities and the challenges in a city's use of data. He identifies the key issues that need to be addressed – what data to develop and access from counties, states, the federal government, and private sources; how to develop the capacity to use data; how to present data and be transparent; and how best to keep data secure so that individuals and organizations are protected – as well as the key barriers. *Elias* provides a broader perspective than simply the United States by noting that many of the legal and ethical issues associated with big data have wider relevance. Much can be learned from examining the progress made in Europe toward developing a harmonized approach to legislation designed to provide individuals and organizations with what has become known as the 'right to privacy'. The legislative developments have had and are continuing to have substantial impact on cross-border access to microdata for research purposes; that impact is also examined.

Greenwood, Stopczynski, Sweatt, Hardjono and Pentland explore the emergence of the Big Data society, arguing that the 'personal data sector' of the economy needs productive collaboration between the government, the private sector, and the citizen to create new markets – just as the automobile and oil industries did in prior centuries. They sketch a model of data access that is governed by 'living informed consent', whereby the user is entitled to know what data are being collected about her by which entities, empowered to understand the implications of data sharing, and finally is put in charge of any data-sharing authorizations. They envision the establishment of a New Deal on Data, grounded in principles such as the

opt-in nature of data provision, the framing of data usage boundaries, and the credentialing of parties authorized to access data.

Landwehr takes a very pragmatic approach. He notes that, regardless of what data policies have been agreed, access must be allowed through controls engineered into the data infrastructure. Without sound technical enforcement, incidents of abuse, misuse, theft of data, and even invalid scientific conclusions based on undetectably altered data can be expected. He discusses what features access controls might have – delineating the characteristics of subjects, objects, and access modes – and notes that advances in practical cryptographic solutions to computing on encrypted data could change the picture in the future by reducing the need to trust hardware and system software. Advances in methods for building systems in which information flow, rather than access control, is the basis for policy enforcement could also open the door to better enforcement of comprehensible policies.

Wilbanks is similarly practical. He provides an overview of frameworks that are available to permit data reuse and discusses how legal and technical systems can be structured to allow people to donate their data to science. He argues that traditional frameworks to permit data reuse have been left behind by the mix of advanced techniques for re-identification and cheap technologies for the creation of data about individuals. He surveys the approaches developed in technological and organizational systems to 'create' privacy where it has been eroded while allowing data reuse, but also discusses a new approach of 'radical honesty' toward data contribution and the development of 'portable' approaches to informed consent that could potentially support a broad range of research without the unintended fragmentation of data created by traditional consent systems

Kreuter and Peng open Part III, on the statistical framework, with a discussion of the new statistical challenges associated with inference in the context of big data. They begin by noting that reliable statistical inference requires an understanding of the data-generating process. That process is not well understood in the case of big data, so it is important that researchers be given access to the source data so that coverage and quality issues can be identified and addressed. Standard statistical disclosure limitations are unlikely to work, because an important feature of big data is the ability to examine different, targeted populations, which often have unique and easily re-identifiable characteristics. *Karr and Reiter* explore the interactions between data dissemination, big data, and statistical inference. They identify a set of lessons that stewards of big data can learn from statistical

agencies' experiences with the measurement of disclosure risk and data utility. Their conclusion is that the sheer scale and potential uses of big data will require that analysis be taken to the data rather than the data to the analyst or the analyst to the data. They argue that a viable way forward for big data access is an integrated system including (i) unrestricted access to highly redacted data, most likely some version of synthetic data, followed with (ii) means for approved researchers to access the confidential data via remote access solutions, glued together by (iii) verification servers that allow users to assess the quality of their inferences with the redacted data so as to be more efficient with their use (if necessary) of the remote access to the confidential data. *Dwork* concludes with a vision for the future. She shows how differential privacy provides a mathematically rigorous theory of privacy, a theory amenable to measuring (and minimizing) cumulative privacy loss, as data are analyzed and re-analyzed, shared, and linked. There are trade-offs – differential privacy requires a new way of interacting with data, in which the analyst accesses data only through a privacy mechanism, and in which accuracy and privacy are improved by minimizing the viewing of intermediate results. But the approach provides a measure that captures cumulative privacy loss over multiple releases; it offers the possibility that data usage and release could be accompanied by publication of privacy loss.

Thanks As with any book, we have benefited enormously from the support and help of many people. Our editor, Diana Gillooly, has worked tirelessly and efficiently at all phases – going well beyond the call of duty. Our referees took time out of their busy schedules to give thoughtful, constructive guidance to the authors. They include Micah Altman, Mike Batty, Jason Bobe, Aleksandra Bujnowska, Fred Conrad, Josep Domingo-Ferrer, Stephanie Eckman, Mark Elliot, Martin Feldkircher, Simson Garfinkel, Bob Goerge, Eric Grosse, Patricia Hammar, David J. Hand, Dan Harnesk, Kumar Jayasuriya, Gary King, Frauke Kreuter, Tom Kvan, Bethany Letalien, William Lowrance, Lars Lyberg, Tim Mulcahy, Kobbi Nissim, Onora O'Neill, Kathleen Perez Lopez, Carlo Reggiani, Jerome H. Reichman, Guy Rothblum, Subu R. Sangameswar, Fred Schneider, Aleksandra Slavkovic, Tom Snijders, Omer Tene, Vincenc Torra, Paul Uhlir, Richard Valliant, and Felix Wu.

The book was sponsored by New York University, through the Center for Urban Science and Progress. Konstantin Baetz, Veronika Zakrocki, Felicitas Mittereder, Reinhard Sauckel, Dominik Braun, Shaylee Nielson,

and Christian Rafidi provided superb administrative support, Mark Righter in NYU General Counsel's office provided legal assistance, and Ardis Kadiu of Spark451 gave assistance with our project website. The book was also supported by the Privacy and Confidentiality Subcommittee of the American Statistical Association and the Institute for Employment Research of the German Federal Employment Agency.

We are also very grateful to Steve Pierson, American Statistical Association; Kim Alfred, CUSP; and Kelly Berschauer, Microsoft Research, for their help with outreach to key stakeholders.

NOTES

1. 'Big data' is given many definitions. One fairly representative version says that the term "refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future" (NSF BIGDATA solicitation http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767&org=CISE). Another uses the velocity, volume, and variety rubric: "data is now available faster, has greater coverage and scope, and includes new types of observations and measurements that previously were not available" (L. Einav and J. D. Levin, "The Data Revolution and Economic Analysis," National Bureau of Economic Research Working Paper, No. 19035, Cambridge, MA, 2013, retrieved from http://www.nber.org/papers/w19035).
2. For a survey of statisticians' opinions on the privacy and ethics related to data collections, see http://blog.revolutionanalytics.com/2013/09/statistician-survey-results.html.
3. For a good review of the role of statistics in the Greek financial crisis, see http://www.ft.com/cms/s/0/82b15932-18fe-11e1-92d8-00144feabdc0.html#axzz2g7W3pWOJ.
4. D. Kahneman, *Thinking Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).
5. See http://www.census.gov/aboutus/#.
6. J. Groen, "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures," *Journal of Official Statistics* 28, no. 2 (2012): 173–198.
7. J. M. Abowd and L. Vilhuber, "National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail," *Journal of Econometrics* 161, no. 1 (2011): 82–99, retrieved from http://ideas.repec.org/a/eee/econom/v161y2011i1p82-99.html.
8. T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data Intensive Scientific Discovery* (Redmond, WA: Microsoft Research, 2009).
9. Einav and Levin, "The Data Revolution and Economic Analysis."
10. E. Ostrom, Governing the Commons: The Evolution of Institutions for Collective Action (Cambridge, UK: Cambridge University Press, 1990).
11. See http://en.wikipedia.org/wiki/AOL_search_data_leak.

12. See; e.g., http://www.theguardian.com/commentisfree/2013/may/04/adp-forecasting-monthly-bls-jobs-reports.
13. Many of the issues associated with the role of Institutional Review Boards are highlighted in a recent National Academies' workshop http://sites.nationalacademies.org/DBASSE/BBCSS/CurrentProjects/DBASSE_080452.