Legal Considerations in Data Science

Victoria Stodden
School of Information Sciences and the College of Law
University of Illinois at Urbana-Champaign

LAW AND DATA SCIENCE SUMMIT
Session on “Law Governing Data”
College of Law
University of Illinois at Urbana-Champaign
October 5, 2018
On May 1st, 2014, the White House released two landmark reports on Privacy and Big Data. Laws are going to follow. What should they be?

On May 13th, 2014, the European Union's court of justice upheld a 'Right to be Forgotten'. Why are so many thoughtful people worried by this?

Big Data can make us safer, richer, smarter. It can also strip us of our privacy without our consent.

This book is your intellectual starter kit for engaging with one of today's most important issues.

Download a book flyer for the special promotional price

Also available at Amazon
Ideas Motivating the Book

the book’s focus is on the legal, economic, and statistical context necessary to frame the many issues regarding big data, including privacy, the value to the public of data access, clarifying personal data ownership questions, and raising issues of agency in personal data.

1. *Conceptual Framework*: What data can be legally collected? What can be legally used? What’s the role of informed consent?

2. *Practical Framework*: How can data be used in novel research areas including the new urban science, infrastructure for a city’s use of data, and how does this change in the international context?

Chapter 1: Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context

Katherine J. Strandburg, NYU

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. This chapter develops a taxonomy of current U.S. privacy law. It then uses that taxonomy to elucidate the mismatch between current law and big data privacy concerns and makes five suggestions to reduce the mismatch. Of these, probably the most important is to recognize that there is no functioning market for assessing citizens' preferences, and that there is a critical need for measuring both the privacy impact of data acquisition and the potential benefit of data use; much can be learned from the experience of environmental regulation.

Related video: 2 minutes with Katherine Strandburg: on her recent work
"We're not living just in real space anymore. We're living out there online."

Chapter 2: Big Data's End Run around Anonymity and Consent

Solon Barocas and Helen Nissenbaum, NYU

Big data involves practices that have radically disrupted entrenched information flows. From modes of acquiring 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. 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.
This chapter surveys the growing body of theoretical and empirical research on the economics and behavioral economics of privacy, and discusses how these streams of research can be applied to the investigation of 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 reduces inefficiencies and increases 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 would shrink with the amount of information they share), data holders (who may face increasing costs associated with collecting and handling consumer data), or both?

Related video: 15 minutes with Alessandro Acquisti: why privacy matters

"Privacy is not about having something negative to hide."

Chapter 4: Changing the Rules: General Principles for Data Use and Analysis

Paul Ohm, Colorado

How do information privacy laws regulate the use of big data techniques, if at all? Do these laws strike an appropriate balance between allowing the benefits of big data and protecting individual privacy? If not, how might we amend or extend laws to better strike this balance? 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 proposes five general approaches for change.
Chapter 5: Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency

Victoria Stodden, Columbia

This chapter begins by motivating the scientific rationale for access to data and computational methods to enable the verification and validation of published research findings. It 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.

II. Practical Framework

Chapter 6: The Value of Big Data for Urban Science

Steven E. Koonin and Michael J. Holland, CUSP

This chapter addresses the motivations for the new urban science, and the value for cities – particularly with respect to analysis of the infrastructure, the environment, and the people. It discusses the key technical issues necessary to build a data infrastructure for curation, analytics, visualization, machine learning, data mining, as well as modeling and simulation to keep up with the volume and speed of data.
Chapter 7: Data for the Public Good: Challenges and Barriers in the Context of Cities

Robert M. Goerge, University of Chicago

This chapter uses an example of the creation of a data warehouse which links data on multiple services provided by the public sector to individuals and families as a way to highlight both the barriers to and opportunities for cities to use data. It 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.

Chapter 8: A European Perspective on Research and Big Data Access

Peter Elias, University of Warwick

Many of the legal and ethical issues associated with big data have wider relevance; this chapter discusses them from a European perspective. The first part gives an historical overview of the progress that has been made across Europe to develop a harmonised approach to legislation designed to provide individuals and organisations with what has become known as the ‘right to privacy’. The second part examines the impact that these legislative developments have had and are continuing to have on cross-border access to microdata for research purposes.
Chapter 9: The New Deal on Data: A Framework for Institutional Controls

Daniel Greenwood, MIT, Arkadiusz Stopczynski, DTU, Brian Sweatt, MIT, Thomas Hardjono, MIT, Alex Pentland, MIT

This chapter explores 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. It envisions data access to be governed by ‘living informed consent’, where the user is entitled to know what data is being collected about her by which entities, empowered to understand the implications of data sharing, and finally put in charge of the sharing authorizations. It discusses the establishment of a New Deal on Data, grounded in principles, such as the opt-in nature of data provision, the boundaries of the data usage, and parties accessing the data.

Chapter 10: Engineered Controls for Dealing with Big Data

Carl Landwehr, George Washington University

Regardless of what data policies have been agreed to, 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. This chapter discusses what features those access controls might have – delineating the characteristics of subjects, objects, and access modes. Although fundamental computing concepts for engineered controls on access to data and on information flows are reasonably well developed, they are perhaps not so widely deployed as they might be. Areas of research that could change the picture in the future include advances in practical cryptographic solutions to computing on encrypted data, which could reduce the need to trust hardware and system software. Advances in methods for building systems in which information flow, rather than access control, as the basis for policy enforcement could also open the door for better enforcement of comprehensible policies.
Open Licensing for Data-Enabled Research

- Background: Open Licensing

  ➡ Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

- Hundreds of open source software licenses:
  - GNU Public License (GPL)
  - (Modified) BSD License
  - MIT License
  - Apache 2.0 License
  - ... see http://www.opensource.org/licenses/alphabetical
The Reproducible Research Standard

The Reproducible Research Standard (RRS) (Stodden, 2009)

• A suite of license recommendations for computational science:

• Release media components (text, figures) under CC BY,

• Release code components under Modified BSD or similar,

• Release data to public domain or attach attribution license.

➡ Remove copyright’s barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.