Privacy, Big Data, and the Public Good: Frameworks for Engagement

Victoria Stodden
Department of Statistics
Columbia University

Invited Special Presentation: Privacy and Big Data
Joint Statistical Meetings
Boston, MA
August 6, 2014

Part I. Conceptual Framework

1. Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context *Katherine J. Strandburg*

2. Big Data's End Run around Anonymity and Consent *Solon Barocas and Helen Nissenbaum*

3. The Economics and Behavioral Economics of Privacy *Alessandro Acquisti*

4. Changing the Rules: General Principles for Data Use and Analysis *Paul Ohm*

5. Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency *Victoria Stodden*

Part I. Conceptual Framework

- What are the legal constraints stemming from privacy on the collection and use of big data?
- What are the gaps in the current legal framework?
- How can we improve on this framework to access the benefits of big data while protecting privacy?

Part I. Conceptual Framework

3 themes emerge:

1. that the concepts used in the larger discussion of privacy and big data require updating;
2. that how we understand and assess harms from privacy violations needs updating;
3. and that we must rethink established approaches to managing privacy in the big data context.

Some emerging points..

- The term ‘big data’ is interpreted as a change in paradigm, rather than solely as a change in technology.
- Assessing harm from privacy breaches is complicated by big data, extending harm from an individual concept to that of groups or classes, and even society as a whole.
- Anonymity and informed consent are not panaceas, and do not solve these problems, even if they were possible in all cases.

Some emerging points..

- The concept of notice is complicated by big data (notice to whom? for what?),
- The concept of risk is complicated as individuals appear in various different datasets,
- Verification of inferences from big data is complicated by privacy: consider maximizing access within regulatory and ethical constraints to maximize research reliability.

More transparency?

- Legal considerations
- Ethical considerations
- Data owner agency considerations
- Propriety data
- Data linking and future exposure considerations..

Assertion: Traditional restrictions on data access need to be revisited in the big data context, to ensure replicable reliable research results.

Between Open and Closed

- Example 1: “Walled Gardens”

  For protected data, ie. subject to HIPAA, limit access to authorized researchers from independent groups to enable the verification of scientific findings, within a walled garden.

- Example 2: “Data Lakes”

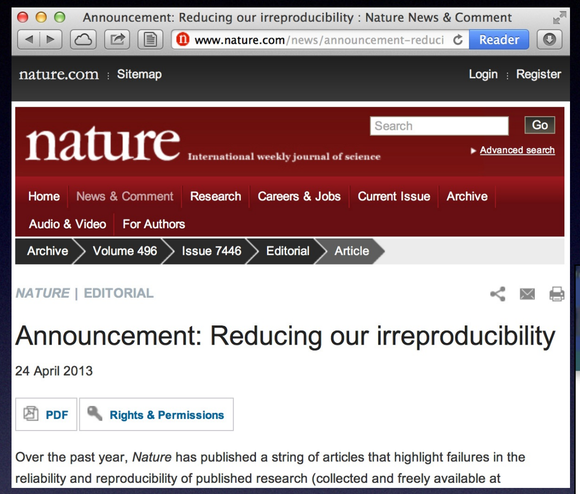
  Department of Homeland Security approach: proactively tag permission levels for each dataset in the “lake” e.g. core biographical data, extended biographical data, DHS encounter data. (Neptune and Cerberos pilots)

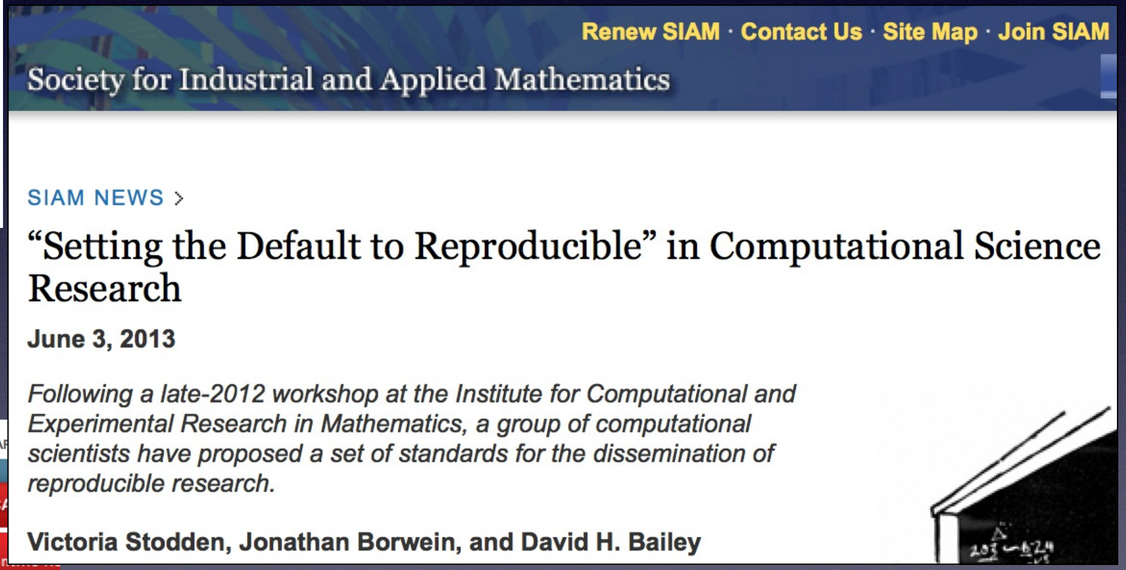
Parsing Reproducibility

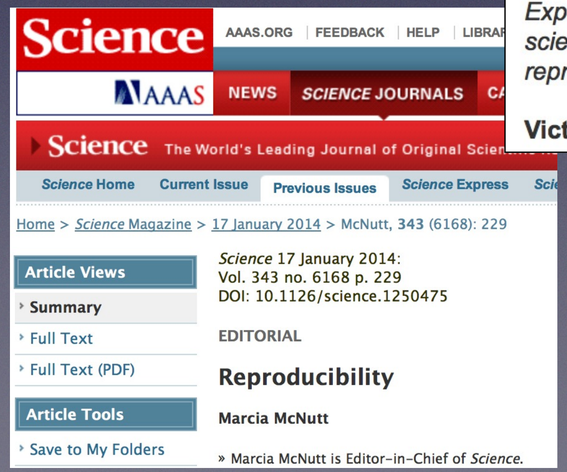
"Empirical Reproducibility"

"Computational Reproducibility"

"Statistical Reproducibility"







V. Stodden, IMS Bulletin (2013)

Supporting Computational Science

- Dissemination Platforms:

ResearchCompendia.org	IPOL	Madagascar
MLOSS.org	thedatahub.org	nanoHUB.org
Open Science Framework		RunMyCode.org

- Workflow Tracking and Research Environments:

VisTrails	Kepler	CDE	IPython Notebook
Galaxy	GenePattern	Paper Mâché	
Sumatra	Taverna	Pegasus	

- Embedded Publishing:

Verifiable Computational Research	SOLE	knitR
Collage Authoring Environment	SHARE	Sweave