Relying on Data Science: Reproducible Research and the Role of Policy

Victoria Stodden
Department of Statistics
Columbia University

DataEDGE Conference
UC Berkeley
May 9, 2014

Big Data and Reproducibility

1. What's Reproducibility?
2. *Big Data:* Scientific Research and Industry Engagement
3. *Policy Impact:* The OSTP Executive Memorandum and Executive Order

Mongo like “Reproducibility”!

- Reproducibility good. But what is it?
- Reproducibility good. But why?
- Reproducibility hard. But why?
- Reproducibility hard to sell. But why?
- Hard to teach an old dog new tricks.
- Solution: Work with puppies.

Why do we like reproducibility?

- Provides (a way to generate) evidence of correctness
- Enables re-use, modification, extension, . . .
- Exposes methods, which might be interesting and instructive
- Gut feeling that transparency and openness are good

Claim: Reproducibility is a tool, not a primary goal.

Might accomplish some of those goals without it, but it's a Very Powerful Tool.

Open Data Crucial to Science Today

- not a new concept, rooted in *skepticism*
- Transactions of the Royal Society 1660's
- Transparency, knowledge transfer -> goal to perfect the *scholarly record*. Nothing else.
- Technology has changed the nature of experimentation, data, and communication.

Computation is Becoming Central to Scientific Research

1. enormous, and increasing, amounts of data collection:
    - CMS project at LHC: 300 "events" per second, 5.2M seconds of runtime per year, .5MB per event = 780TB/yr => several PB when data processed,
    - Sloan Digital Sky Survey: 9th data release (SDSS-III 2012), 60TB,
    - quantitative revolution in social science due to abundance of social network data (Lazier et al, *Science*, 2009)
    - *Science* survey of peer reviewers: 340 researchers regularly work with datasets >100GB; 119 regularly work with datasets >1TB (N=1700, Feb 11, 2011, p. 692)
2. massive simulations of the complete evolution of a physical system, systematically varying parameters,
3. deep intellectual contributions now encoded in software.

Credibility Crisis

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

Ioannidis, 2011: 9% of authors studied made data available

Generally, data and code not available at the time of publication, insufficient information in the publication for verification of results.

A Credibility Crisis

Scientific Perspective

- “Really Reproducible Research” inspired by Stanford Professor Jon Claerbout:

  “The idea is: An article about computational science in a scientific publication is *not* the scholarship itself, it is merely *advertising* of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.” David Donoho, 1998.

Updating the Scientific Method

Argument: computation presents only a *potential* third branch of the scientific method (Stodden et al 2009):

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments,
- Branch 3,4? (computational): large scale simulations / data driven computational science.

The Ubiquity of Error

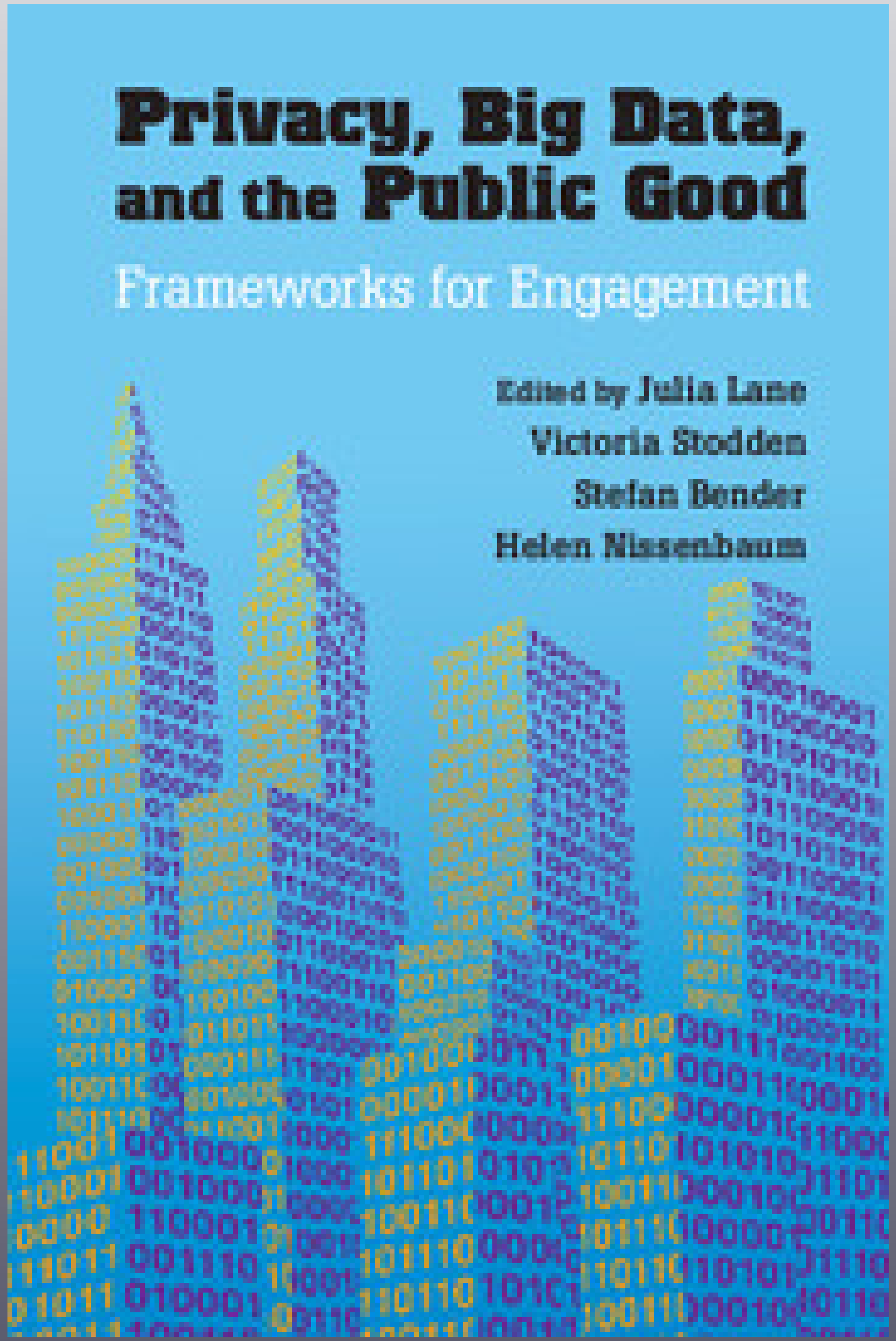
The central motivation for the scientific method is to root out error:

1. Deductive branch: the well-defined concept of the proof,
2. Empirical branch: the machinery of hypothesis testing, structured communication of methods and protocols.

Claim: We must develop comparable communication standards to elevate the 3rd and 4th branches to the scientific method

Big Data: Scientific Research and Industry Engagement

Barriers to Transparency



Privacy, Big Data, and the Public Good

Frameworks for Engagement

Edited by

Julia Lane
American Institutes for Research, Washington DC

Victoria Stodden
Columbia University

Stefan Bender
Institute for Employment Research of the German Federal Employment Agency

Helen Nissenbaum
New York University

OSTP Executive Memorandum and Executive Order

2013: Open Science in DC

- Feb 22: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.
- May 9: Executive Order directing federal agencies to make their data publicly available.

Data in the Memorandum

1. Digitally formatted data arising from federal grants should be stored and publicly accessible to search, retrieve, and analyze.
2. "[D]ata is defined... as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications"

Each Public Access Plan Shall...

a) Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while [respecting privacy, proprietary interests and IP, need for long-term preservation],

b) Ensure that all ... researchers receiving Federal grants ... develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why longterm preservation and access cannot be justified,

c) Allow the inclusion of appropriate costs for data management and access in proposals for Federal funding for scientific research,

d) Ensure appropriate evaluation of the merits of submitted data management plans,

e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies,

f) Promote the deposit of data in publicly accessible databases, where appropriate and available,

g) Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations,

h) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan,

i) In coordination with other agencies and the private sector, support ... workforce development related to scientific data management, analysis, storage, preservation, and stewardship, and

j) Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.