Law, Ethics, & Economics of Big Data

Victoria Stodden
Department of Statistics
Columbia University

Book Preview: Privacy, Big Data, And The Public Good: Frameworks For Engagement
NY Academies of Science
New York, NY
June 16, 2014
Credibility Crisis

Los Angeles Times | BUSINESS

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren’t even true.

Science 17 January 2014:
Vol. 343 no. 6188 p. 229
DOI: 10.1126/science.1250475

Reproducibility
Marcia McNutt

Science advances on a foundation of trusted approach that scientists use to gain confidence. Community was shaken by reports that a result not reproducible. Because confidence in results community, we are announcing new initiatives Science. For preclinical studies (one of the targets) recommendations of the U.S. National Institute increasing transparency. Authors will indicate handling (such as how to deal with outliers), will ensure a sufficient signal-to-noise ratio, whether experimenter was blind to the conduct of the guidelines.

NIH Tackles Irreproducibility

The federal agency speaks out about how to improve the quality of scientific research.

By Jef Akst | January 28, 2014

The Scientist

Must try harder

The Scientist 483, 509 (29 March 2012) | doi:10.1088/08928367/483/509/a
Published online 28 March 2012

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.
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)

• “Principle 1. Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims.”
Parsing Reproducibility

“Empirical Reproducibility”

“Computational Reproducibility”

“Statistical Reproducibility”

V. Stodden, IMS Bulletin (2013)
Supporting Computational Science

• Dissemination Platforms:
  - ResearchCompendia.org
  - MLOSS.org
  - Open Science Framework
  - IPOL
  - thedatahub.org
  - Madagascar
  - nanoHUB.org
  - RunMyCode.org

• Workflow Tracking and Research Environments:
  - VisTrails
  - Galaxy
  - Sumatra
  - Kepler
  - GenePattern
  - Taverna
  - CDE
  - Paper Mâché
  - Pegasus
  - IPython Notebook

• Embedded Publishing:
  - Verifiable Computational Research
  - Collage Authoring Environment
  - SOLE
  - knitR
  - SHARE
  - Sweave
Research Compendia

Pilot project: improve understanding of reproducible computational science, trace sources of error.

• link data/code to published claims, re-use,
• research produces a guide to empirical researchers, certifies results,
• large scale validation of findings,
• stability, sensitivity checks.
Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items

Mark Dingemanse, Francisco Torreira, N. J. Enfield, Johan J. Bolhuis

Code and Data Abstract

A word like Huh?—used as a repair initiator when, for example, one has not clearly heard what someone just said—is found in roughly the same form and function in spoken languages across the globe. We investigate it in naturally occurring conversations in ten languages and present evidence and arguments for two distinct claims: that Huh? is universal, and that it is a word. In support of the first, we show that the similarities in form and function of this interjection across languages are much greater than expected by chance. In support of the second claim we show that it is a lexical, conventionalised form that has to be learnt, unlike grunts or emotional cries. We discuss possible reasons for the cross-linguistic similarity and propose an account in terms of convergent evolution. Huh? is a universal word not because it is innate but because it is shaped by selective pressures in an interactional environment that all languages share: that of other-initiated repair. Our proposal enhances evolutionary models of language change by suggesting that conversational infrastructure can drive the convergent cultural evolution of linguistic items.
Reproducibility in Computational and Experimental Mathematics (December 10-14, 2012)

Description

In addition to advancing research and discovery in pure and applied mathematics, computation is pervasive across the sciences and now computational research results are more crucial than ever for public policy, risk management, and national security. Reproducibility of carefully documented experiments is a cornerstone of the scientific method, and yet is often lacking in computational mathematics, science, and engineering. Setting and achieving appropriate standards for reproducibility in computation poses a number of interesting technological and social challenges. The purpose of this workshop is to discuss aspects of reproducibility most relevant to the mathematical sciences among researchers from pure and applied mathematics from academics and other settings, together with interested parties from funding agencies, national laboratories, professional societies, and publishers. This will be a working workshop, with relatively few talks and dedicated time for breakout group discussions on the current state of the art and the tools, policies, and infrastructure that are needed to improve the situation. The groups will be charged with developing guides to current best practices and/or white papers on desirable advances.

Organizing Committee

- David H. Bailey
  (Lawrence Berkeley National Laboratory)
- Jon Borwein
  (Centre for Computer Assisted Research Mathematics and its Applications)
- Randall J. LeVeque
  (University of Washington)
- Bill Rider
  (Sandia National Laboratory)
- William Stein
  (University of Washington)
- Victoria Stodden
  (Columbia University)
Setting the Default to Reproducible

Reproducibility in Computational and Experimental Mathematics

Developed collaboratively by the ICERM workshop participants

Compiled and edited by the Organizers
V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein

Abstract

Science is built upon foundations of theory and experiment validated and improved through open, transparent communication. With the increasingly central role of computation in scientific discovery this means communicating all details of the computations needed for others to replicate the experiment, i.e. making available to others the associated data and code. The “reproducible research” movement recognizes that traditional scientific research and publication practices now fall short of this ideal, and encourages all those involved in the production of computational science – scientists who use computational methods and the institutions that employ them, journals and dissemination mechanisms, and funding agencies – to facilitate and practice really reproducible research.