Big Data and the Credibility Crisis

Victoria Stodden
School of Graduate and Library Science
University of Illinois at Urbana-Champaign

Advisory Committee for the Education and Human Resources Directorate of the NSF
Washington, D. C. (remotely)
Nov 6, 2014
Framing: The Scientific Method

Traditionally two branches to the scientific method:

• Branch 1 (deductive): mathematics, formal logic,

• Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

• Branch 3,4? (computational): large scale simulations / data driven computational science.

Argument: computation presents only a potential third/fourth branch of the scientific method (Donoho et al 2009).
New Paradigms for Discovery?

Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

Gaithersburg, MD
September 21, 1995

Workshop Overview

The workshop consisted of an introduction; five talks, each followed by a discussion period; and an open discussion session. Capsule versions follow immediately; more substantial summaries follow later.

Jim Blue opened the workshop with brief introductory remarks. He emphasized that the purpose of doing modeling and simulation is to gain understanding and insight. The three benefits are that modeling and simulation can be cheaper, quicker, and better than experimentation alone. It is common now to consider computation as a third branch of science, besides theory and experiment.

“This book is about a new, fourth paradigm for science based on data-intensive computing.”
The Impact of Technology

1. **Big Data / Data Driven Discovery**: high dimensional data, $p >> n$,

2. **Computational Power**: simulation of the complete evolution of a physical system, systematically varying parameters,

3. Deep intellectual contributions now encoded only in software.

The software contains “ideas that enable biology...”

*Stories from the Supplement, 2013.*
The Ubiquity of Error

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

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

**Claim:** Computation presents only a potential third/fourth branch of the scientific method (Donoho, Stodden, et al. 2009), until the development of comparable standards.
Parsing Reproducibility

“Empirical Reproducibility”

“Computational Reproducibility”

“Statistical Reproducibility”

V. Stodden, IMS Bulletin (2013)
“Really Reproducible Research” pioneered 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.”

Reproducibility is a Statistical Issue

- False discovery, chasing significance, p-hacking (Simonsohn 2012), overuse and mis-use of p-values,
- Multiple testing, file drawer problem, sensitivity analysis, poor reporting/tracking practices,
- Data preparation, treatment of outliers,
- Poor statistical methods (nonrandom sampling, inappropriate methods,..)
- Model robustness to parameter changes and data perturbations,
- Investigator bias toward previous findings; conflicts of interest.
Experimental Bias

Experimental biases:

Figure 2: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).
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.
Statistical Issues in Software

The challenge of reproducible computational science:

• shareable encoding of good statistical practices,
• permitting independent verification and comparison,
• extending statistical notions of integrity to statistical software practices,

Foundational research needed..
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
  - Kepler
  - Galaxy
  - GenePattern
  - Sumatra
  - Taverna
  - CDE
  - Paper Mâché
  - IPython Notebook
  - Pegasus

• Embedded Publishing:
  - Verifiable Computational Research
  - SOLE
  - Collage Authoring Environment
  - knitR
  - SHARE
  - Sweave
Open Science from the Whitehouse

- Feb 22, 2013: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.
- May 9, 2013: Executive Order directing federal agencies to make their data publicly available.
Request for Input: “Strategy for American Innovation”

• “to guide the Administration's efforts to promote lasting economic growth and competitiveness through policies that support transformative American innovation in products, processes, and services and spur new fundamental discoveries that in the long run lead to growing economic prosperity and rising living standards.”

• “(11) Given recent evidence of the irreproducibility of a surprising number of published scientific findings, how can the Federal Government leverage its role as a significant funder of scientific research to most effectively address the problem?”
Science Policy in Congress

• America COMPETES due to be reauthorized, drafting underway.


• Hearing on Research Integrity and Transparency by the House Science, Space, and Technology Committee (March 5, 2013).

• Reproducibility cannot be an unfunded mandate.

• “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.”

We need:

Standards for reproducibility of computational findings:

1. data access, software access, persistent linking to publications.
2. innovation around data and code access for privacy protection and scale.
3. robust methods, producing stable results, emphasis on reliability and reproducibility.

References


available at http://www.stodden.net
Credibility Crisis

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.

Reproducibility

Marcia McNutt

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

NIH Tackles Irreproducibility

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

By Jef Akst | January 28, 2014

Must try harder

Nature 483, 509 (29 March 2012) | doi:10.1038/483509a
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.