Reproducibility in Science

Why all the fuss?

Victoria Stodden
Department of Statistics
Columbia University

Publish or Perish? The Future of Scholarly Publishing and Careers
UC Davis
Feb 13, 2014

Advances in Technology

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),
    - NIH Associate Director for Data Science, 2014.
2. *computational power*: massive simulations of the complete evolution of a physical system, systematically varying parameters,
3. deep intellectual contributions now encoded in *software*.

The Rationale

- *Skepticism* requires that the claim can be independently verified,
- requiring transparency in the communication of the research process.
- Instantiated by Robert Boyle and the Transactions of the Royal Society in the 1660's.
- Advances in the technology used for scientific discovery have changed how scientists effect reproducibility.

The Scientific Method

Traditionally two branches to the scientific method:

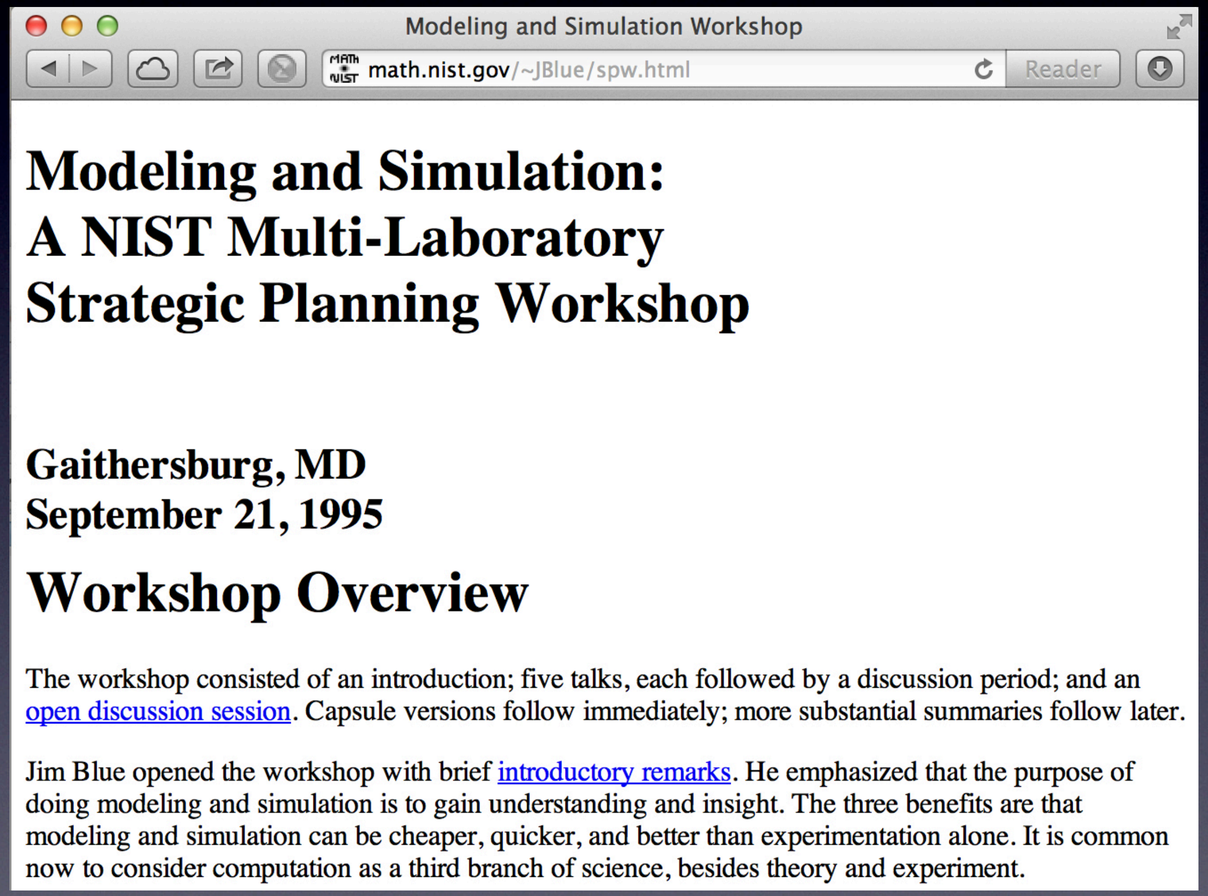
- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

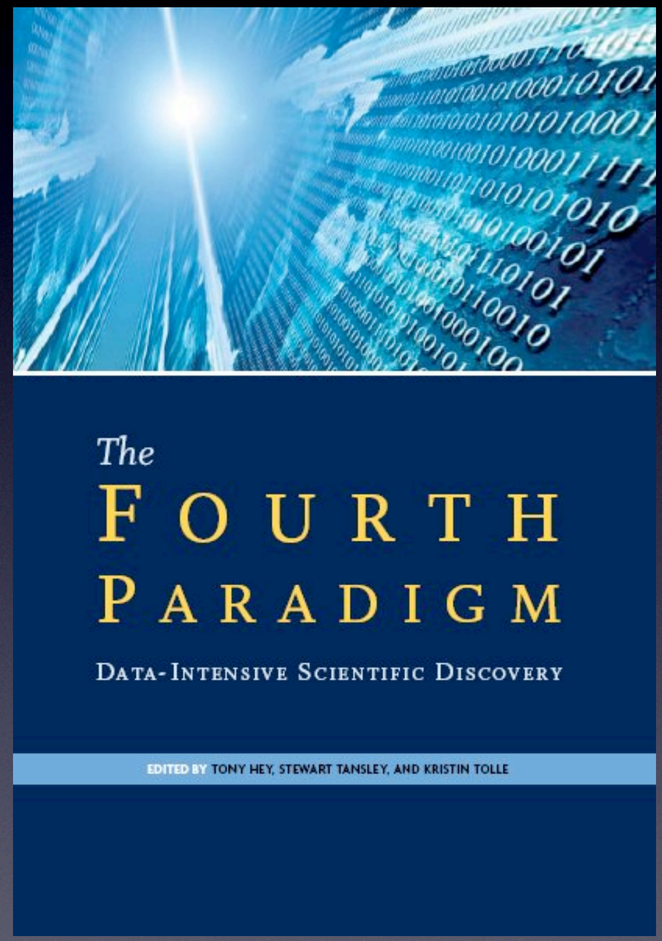
- Branch 3,4? (computational): large scale simulations / data driven computational science.

Argument: computation presents only a *potential* third/fourth branch of the scientific method (Stodden et al 2009). Appropriate standards of communication and review largely absent.

New Paradigms for Discovery?



"It is common now to consider computation as a third branch of science, besides theory and experiment."



2009

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, structured communication of methods and protocols.

Computational science as practiced today does not generate reliable knowledge. Instead, "breezy demos" of results, characterized by the inability to reproduce findings.

Access to code and data typically essential for scientific integrity.

Credibility Crisis

Los Angeles Times | BUSINESS

LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH STYLE TRAV

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.

Announcement: Reducing our irreproducibility : Nature News & Comment

www.nature.com/news/announcement-reduci Reader

nature.com : Sitemap Login : Register

nature International weekly journal of science

Search Go Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 496 > Issue 7446 > Editorial > Article

NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

PDF | Rights & Permissions

Over the past year, *Nature* has published a string of articles tha
reliability and reproducibility of published research (collected an

Science AAAS.ORG | FEEDBACK | HELP | LIBRARIANS

All Science Journals | Enter Search Ter

GUEST ALERTS | ACCESS RIGH

AAAS NEWS *SCIENCE* JOURNALS CAREERS MULTIMEDIA COLLECTIONS

Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.

Science Home | Current Issue | Previous Issues | *Science* Express | *Science* Products | My *Science* | About the Journal

Home > *Science* Magazine > 17 January 2014 > McNutt, 343 (6168): 229

Article Views
- Summary
- Full Text
- Full Text (PDF)

Article Tools
- Save to My Folders
- Download Citation
- Alert Me When Article is Cited
- Post to CiteULike
- E-mail This Page
- Rights & Permissions
- Commercial Reprints and E-Prints
- View PubMed Citation

Related Content

Science 17 January 2014:
Vol. 343 no. 6168 p. 229
DOI: 10.1126/science.1250475

‹ Prev | Table of Contents | Next ›

Read Full Text to Comment (8)

EDITORIAL

Reproducibility

Marcia McNutt

» Marcia McNutt is Editor-in-Chief of *Science*.

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

TheScientist
EXPLORING LIFE. INSPIRING INNOVATION

NIH Tackles Irreproducibility

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

By Jef Akst | January 28, 2014

nature International weekly journal of science

Menu | Advanced search | Search | Go

archive ▸ volume 483 ▸ issue 7391 ▸ editorials ▸ article

NATURE | EDITORIAL

Must try harder

Nature **483**, 509 (29 March 2012) | doi:10.1038/483509a
Published online 28 March 2012

PDF | Citation | Reprints | Rights & permissions | Article metrics

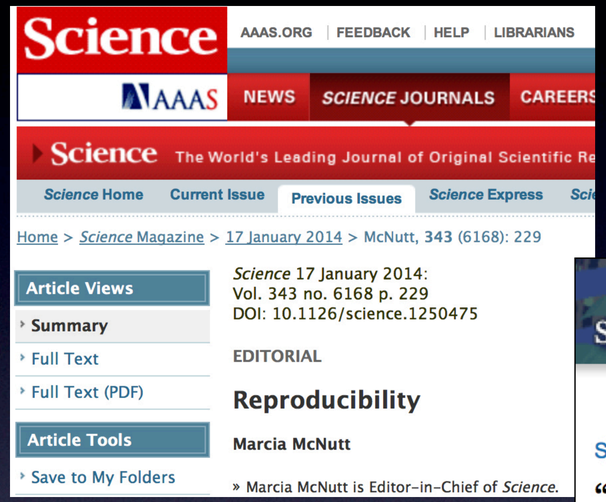
Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

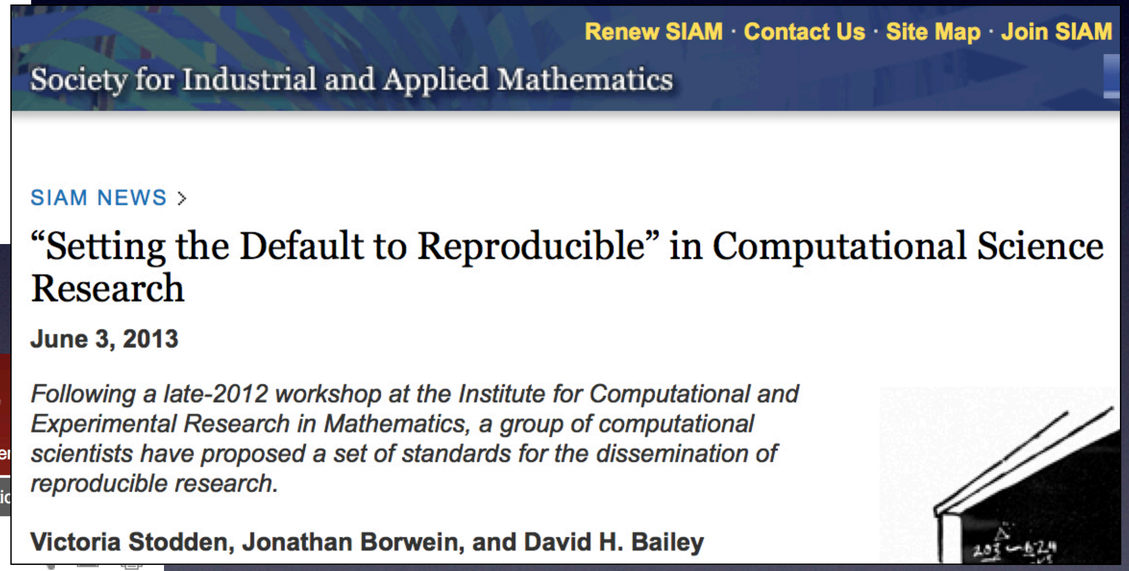
Thinking about Reproducibility

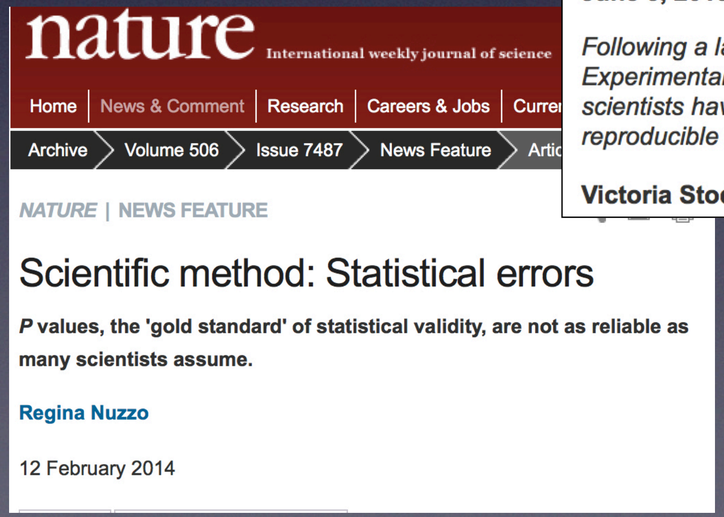
"Empirical Reproducibility"

"Computational Reproducibility"

"Statistical Reproducibility"







“Reproducible Research” is Grassroots

- reproducibility@XSEDE: An XSEDE14 Workshop
- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- Open Science Framework / Reproducibility Project in Psychology
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: “The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer”
- SIAM CSE 2011: “Verifiable, Reproducible Computational Science”
- Yale Law School 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials”
- ...

Integrity in Publishing

- Who reviews computational findings? (right now, unreviewed)
- Is computational replication a post-publication event?
- How to inform the community of the results of your replication?
- Who reviews code?
- Standards for release and re-use of code? Plagiarism is desirable, so is citation.
- Who is responsible for data integrity? Fraud?
- Statistical issues: file drawer problem, antiquated experiment reporting standards (ie. data preparation), inappropriate methods.
- How different are the needs of different fields?

Efforts and Solutions

Best Practices

PLOS | BIOLOGY
TENTH ANNIVERSARY

Browse | For Authors | About Us | Search

OPEN ACCESS
COMMUNITY PAGE

26,543 VIEWS | 25 SAVES

Best Practices for Scientific Computing

Greg Wilson, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

Published: January 07, 2014 • DOI: 10.1371/journal.pbio.1001745

siam Society for Industrial and Applied Mathematics

Renew SIAM · Contact Us · Site Map · Join SIAM · My Account

SIAM Home
About SIAM
Activity Groups
Advertising
Books
Careers & Jobs
Conferences
Customer Service
Digital Library
Fellows Program
History Project
Journals
Membership
Prizes & Recognitions
Proceedings
Public Awareness
Reports
Sections
SIAM Connect
SIAM News
Students

SIAM NEWS >

"Setting the Default to Reproducible" in Computational Science Research

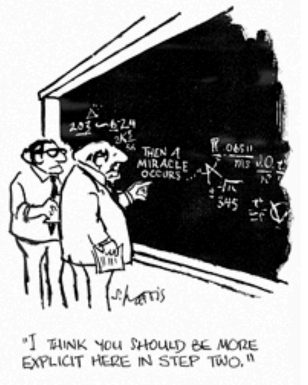
June 3, 2013

Following a late-2012 workshop at the Institute for Computational and Experimental Research in Mathematics, a group of computational scientists have proposed a set of standards for the dissemination of reproducible research.

Victoria Stodden, Jonathan Borwein, and David H. Bailey

Computation is now central to the scientific enterprise, and the emergence of powerful computational hardware, combined with a vast array of computational software, presents novel opportunities for researchers. Unfortunately, the scientific culture surrounding computational work has evolved in ways that make it difficult to verify findings, efficiently build on past research, or even apply the basic tenets of the scientific method to computational procedures.

As a result, computational science is facing a credibility crisis [1,2,4,5]. The enormous scale of state-of-the-art



Courtesy of S. Harris, ScienceCartoonsPlus.com.

Abstract
http://ssrn.com/abstract=2322276

Download This Paper | Share | Email | Add to MyBriefcase

Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research

Victoria Stodden
Columbia University - Department of Statistics

Sheila Miguez
Columbia University

September 6, 2013

Abstract:
Scholarly dissemination and communication standards are changing to reflect the increasingly computational nature of scholarly research, primarily to include the sharing of the data and code associated with published results. This paper presents a formalized set of best practice recommendations for computational scientists wishing to disseminate reproducible research, facilitate innovation by enabling data and code re-use, and enable broader communication of the output of digital scientific research. We distinguish two forms of collaboration to motivate choices of software environment for computational scientific research. We also present these Best Practices as a living, evolving, and changing document on wiki.

siam Society for Industrial and Applied Mathematics

Renew SIAM · Contact Us · Site Map · Join SIAM · My Account

SIAM Home
About SIAM
Activity Groups
Advertising
Books
Careers & Jobs
Conferences
Customer Service
Digital Library
Fellows Program
History Project
Journals
Membership

SIAM NEWS >

Top Ten Reasons To Not Share Your Code (and why you should anyway)

April 1, 2013

Randall J. LeVeque

There is no . . . mathematician so expert in his science, as to place entire confidence in any truth immediately upon his discovery of it. . . . Every time he runs over his proofs, his confidence encreases; but still more by the approbation of his friends; and is raised to its utmost perfection by the universal assent and applauses of the learned world.---David Hume, 1739*

New Tools for Computational Reproducibility

- Dissemination Platforms:

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

- Workflow Tracking and Research Environments:

VisTrails	Kepler	CDE
Galaxy	GenePattern	Synapse
Sumatra	Taverna	Pegasus

- Embedded Publishing:

Verifiable Computational Research	Sweave	knitR
Collage Authoring Environment	SHARE	

Journal Data Sharing Policy

	2011	2012
Required as condition of publication, barring exceptions	10.6%	11.2%
Required but may not affect editorial decisions	1.7%	5.9%
Encouraged/addressed, may be reviewed and/or hosted	20.6%	17.6%
Implied	0%	2.9%
No mention	67.1%	62.4%

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

Journal Code Sharing Policy

	2011	2012
Required as condition of publication, barring exceptions	3.5%	3.5%
Required but may not affect editorial decisions	3.5%	3.5%
Encouraged/addressed, may be reviewed and/or hosted	10%	12.4%
Implied	0%	1.8%
No mention	82.9%	78.8%

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

2013: Open Science in DC

- Feb 22: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications:

  "data 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..."

- May 9: Executive Order directing federal agencies to make their data publicly available.

Sharing: Funding Agency Policy

- NSF grant guidelines: "NSF ... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable." (2005 and earlier)
- NSF peer-reviewed Data Management Plan (DMP), January 2011.
- NIH (2003): "The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers." (>$500,000, include data sharing plan)

NAS Data Sharing Report

- Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences, (2003)
- "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."

Selected References

- Reproducible Research, Guest editor for Computing in Science and Engineering, July/August 2012.
- "Enabling Reproducible Research: Open Licensing for Scientific Innovation," International Journal of Communications Law and Policy, 2009.
- "Reproducible Research: Tools and Strategies for Scientific Computing," July 2011

available at http://www.stodden.net

Journal Publication Requirements

- Journal Policy snapshots June 2011 and June 2012:
- Select all journals from ISI classifications "Statistics & Probability," "Mathematical & Computational Biology," and "Multidisciplinary Sciences" (this includes Science and Nature).
- N = 170, after deleting journals that have ceased publication.

Software in Scientific Discovery

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%

Reproducible Research Movement:
Data and code are made conveniently available at the time of publication

Conflict between incentives to patent academic code and the scientific method?

Barriers to Sharing

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Survey of the Machine Learning Community, NIPS (Stodden 2010)