# 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:

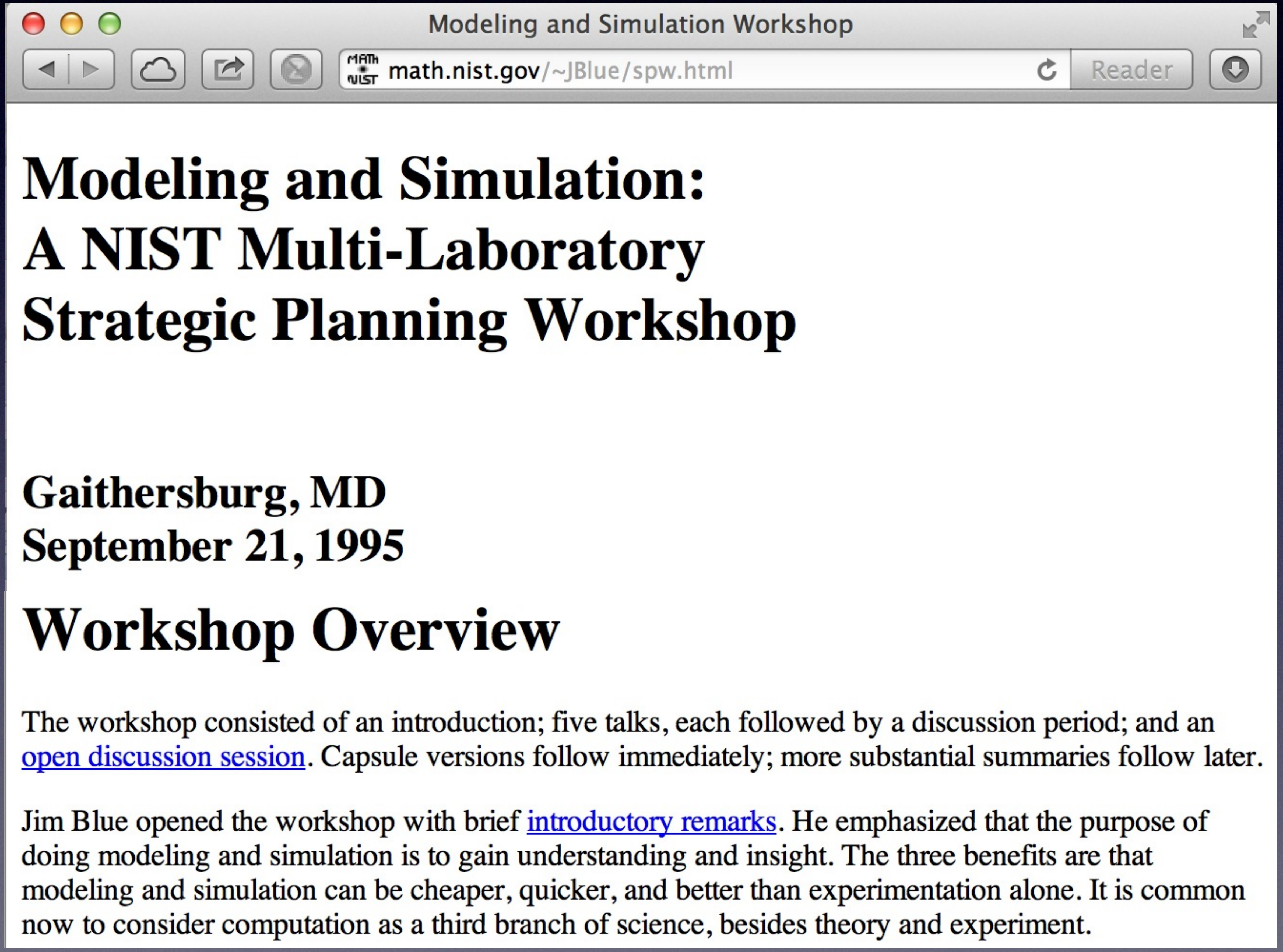
- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments.

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 Workshop
math.nist.gov/~JBlue/spw.html
Reader

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

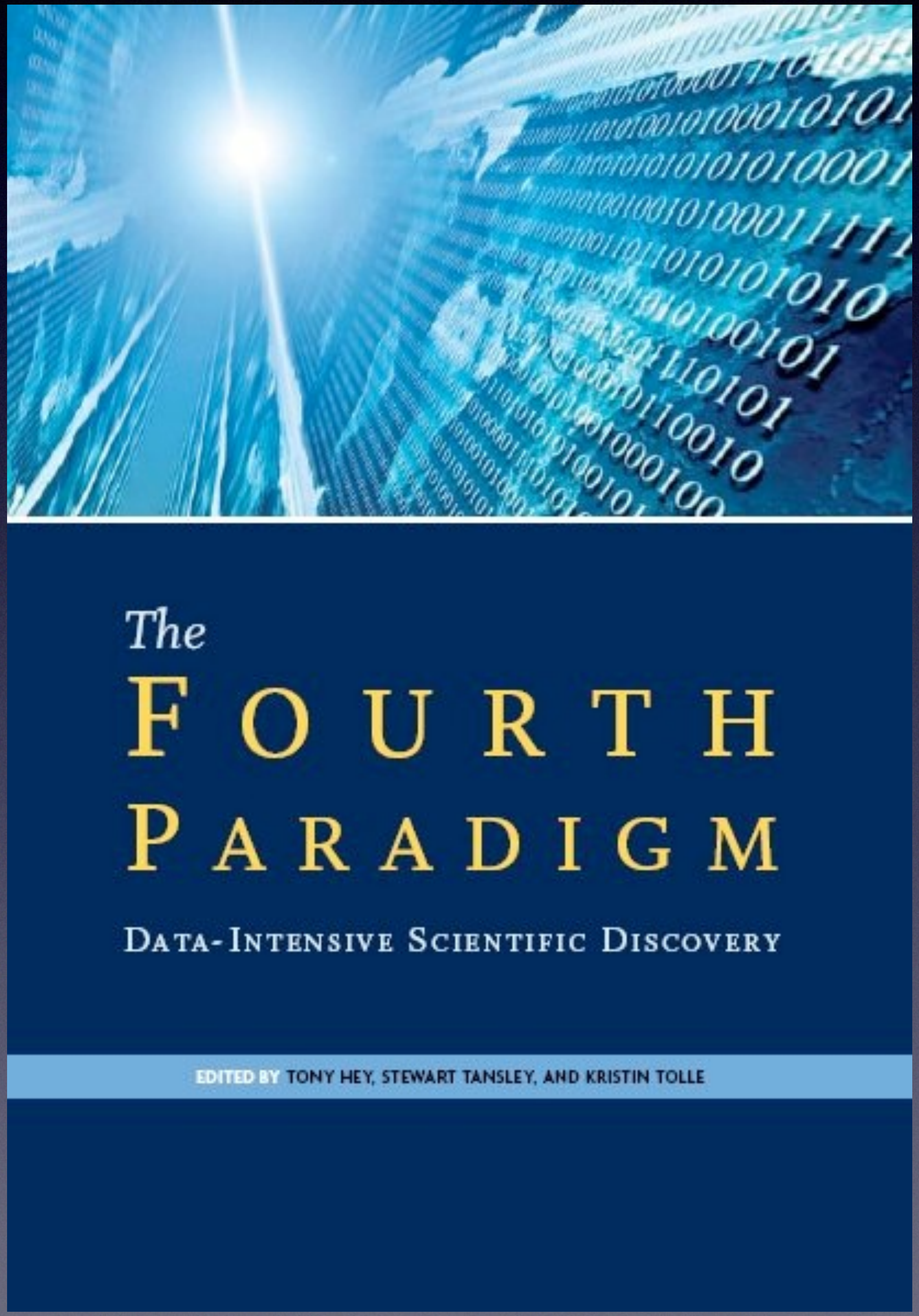
**Gaithersburg, MD**
**September 21, 1995**

## 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.

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



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

“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*.

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

The software contains “ideas that enable biology...”
*Stories from the Supplement*, 2013.

# The Ubiquity of Error

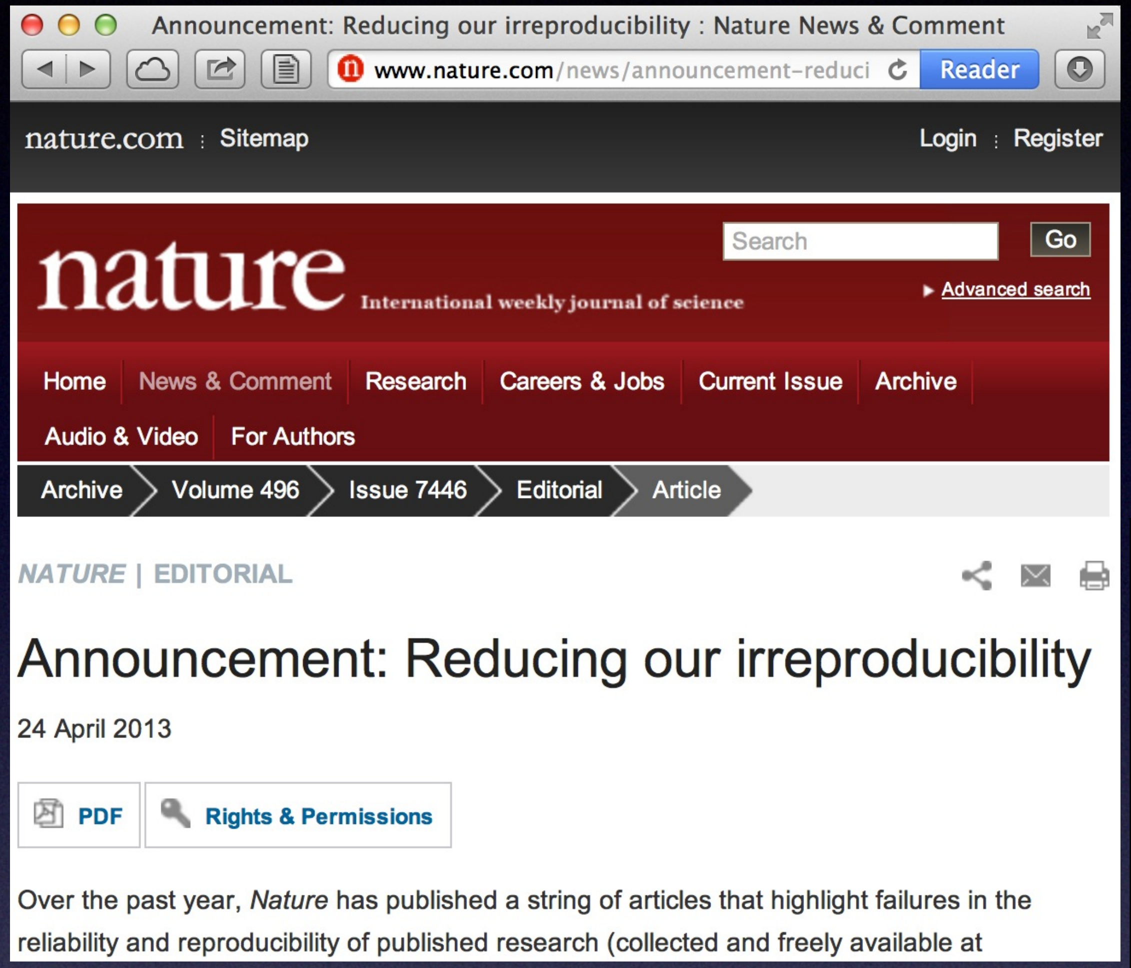
The central motivation for the scientific method is to root out error:

- 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"



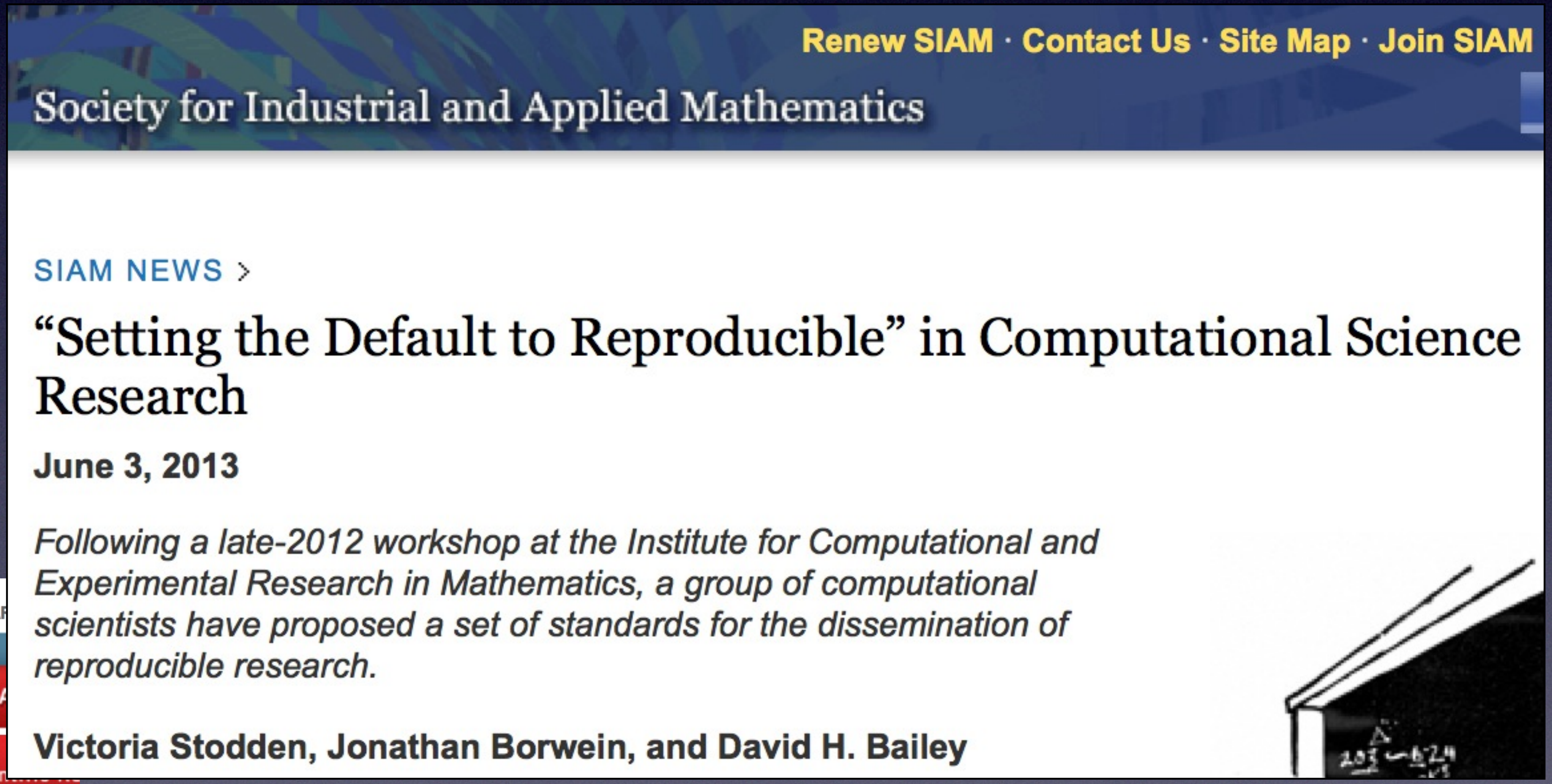
*NATURE* | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at

"Computational Reproducibility"

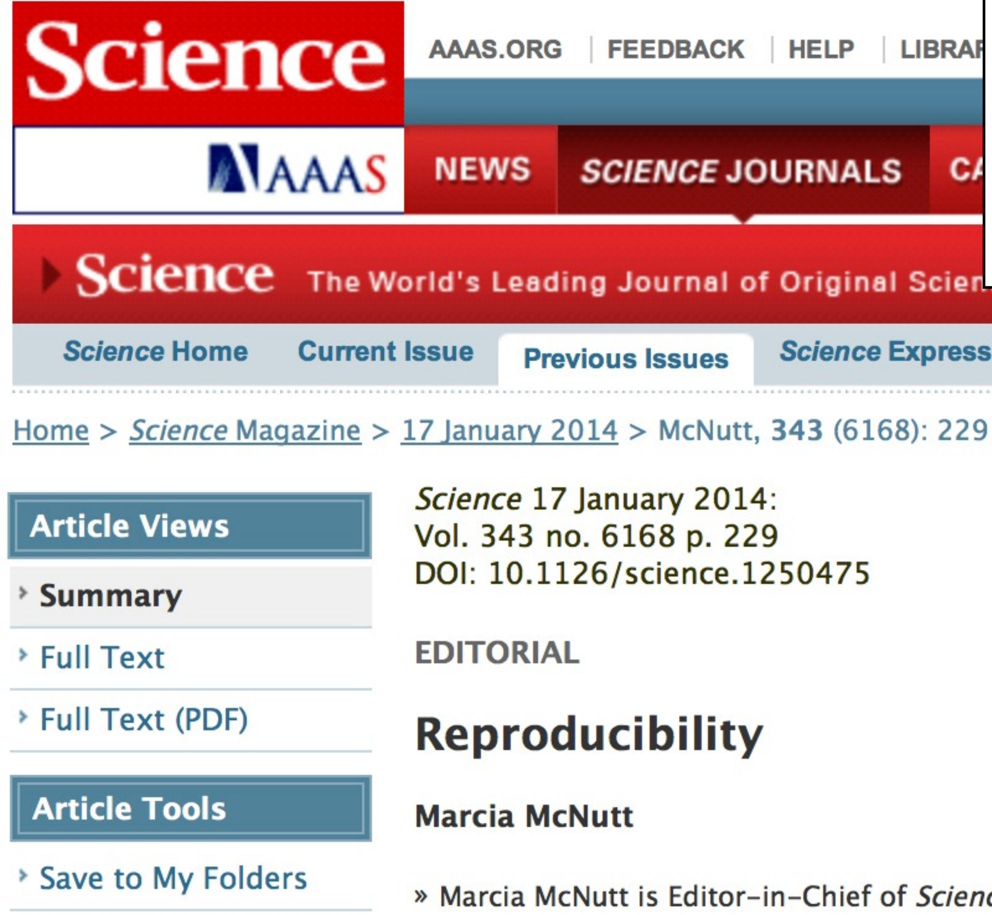

Renew SIAM · Contact Us · Site Map · Join SIAM
Society for Industrial and Applied Mathematics
SIAM NEWS >

"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**



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

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

EDITORIAL

**Reproducibility**

**Marcia McNutt**

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

"Statistical Reproducibility"

V. Stodden, IMS Bulletin (2013)

# Computational Reproducibility

“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.”

paraphrased by David Donoho, 1998.

# 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,

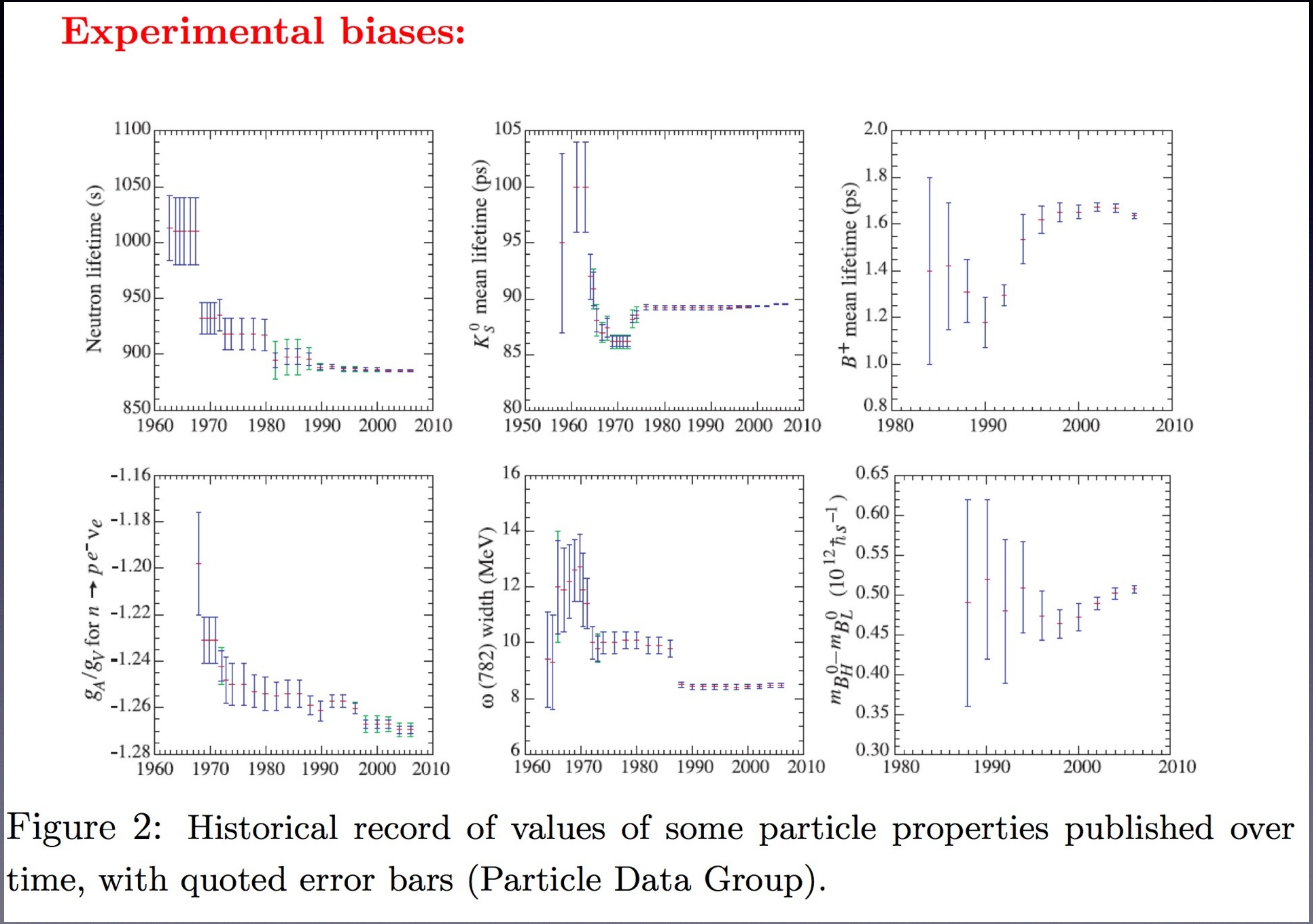
- Data preparation, treatment of outliers,

- 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

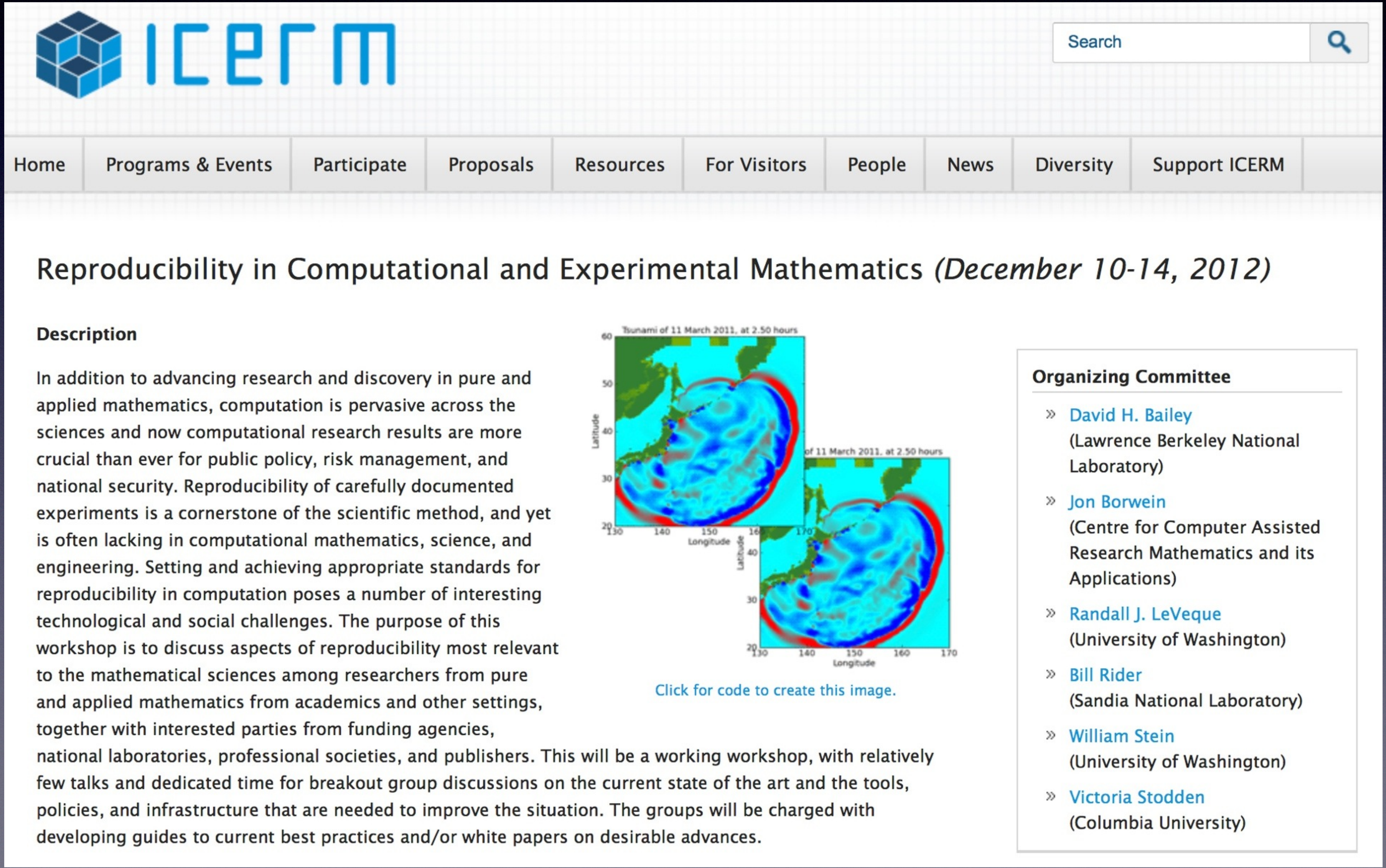


**Experimental biases:**

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

Figure courtesy of James Berger

# ICERM Workshop



## Reproducibility in Computational and Experimental Mathematics *(December 10-14, 2012)*

**Description**

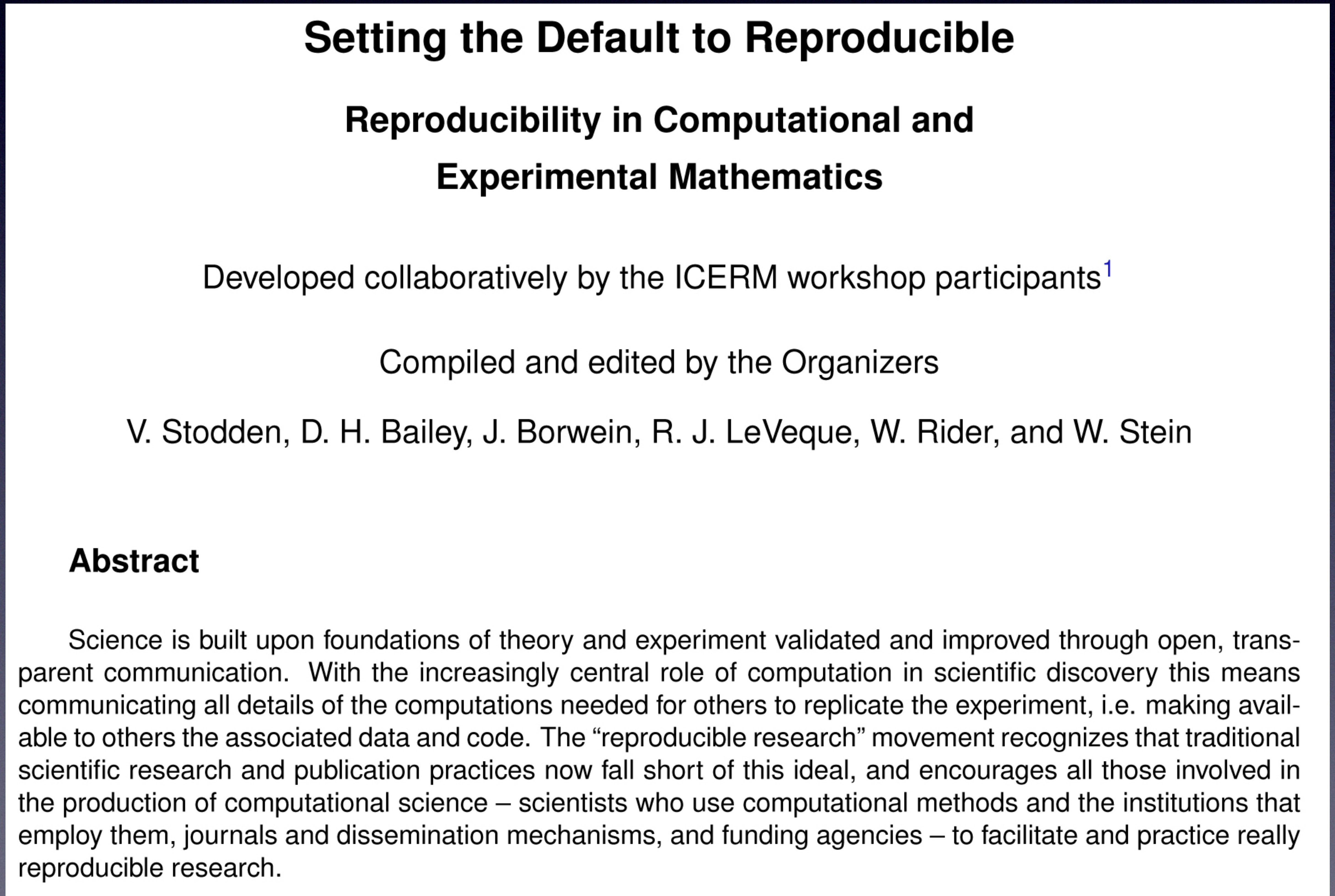
In addition to advancing research and discovery in pure and applied mathematics, computation is pervasive across the sciences and now computational research results are more crucial than ever for public policy, risk management, and national security. Reproducibility of carefully documented experiments 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.

Click for code to create this image.

**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)

# ICERM Workshop Report



## Setting the Default to Reproducible

### Reproducibility in Computational and Experimental Mathematics

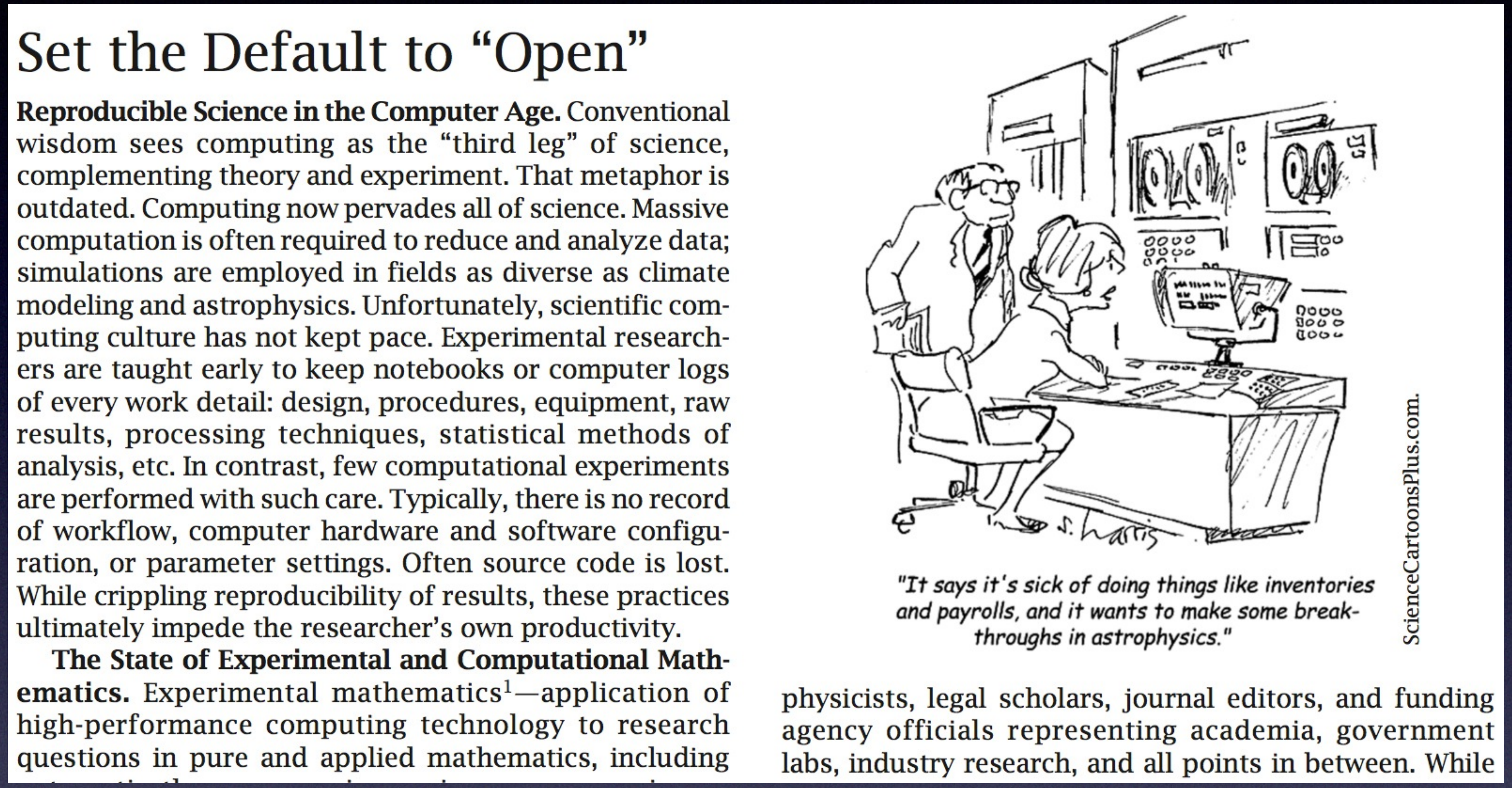
Developed collaboratively by the ICERM workshop participants[1]

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.



## Set the Default to “Open”

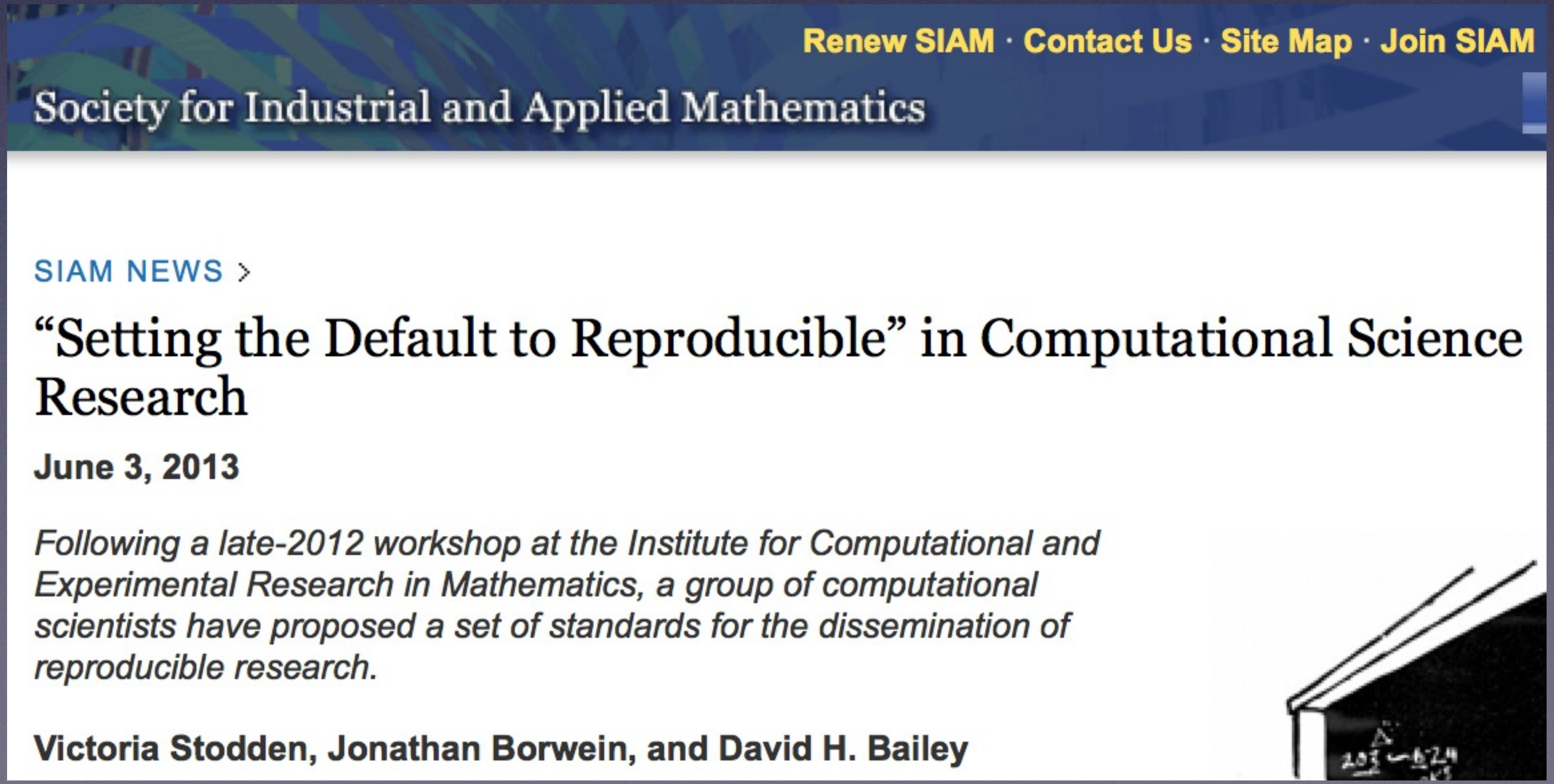
**Reproducible Science in the Computer Age.** Conventional wisdom sees computing as the “third leg” of science, complementing theory and experiment. That metaphor is outdated. Computing now pervades all of science. Massive computation is often required to reduce and analyze data; simulations are employed in fields as diverse as climate modeling and astrophysics. Unfortunately, scientific computing culture has not kept pace. Experimental researchers are taught early to keep notebooks or computer logs of every work detail: design, procedures, equipment, raw results, processing techniques, statistical methods of analysis, etc. In contrast, few computational experiments are performed with such care. Typically, there is no record of workflow, computer hardware and software configuration, or parameter settings. Often source code is lost. While crippling reproducibility of results, these practices ultimately impede the researcher’s own productivity.

**The State of Experimental and Computational Mathematics.** Experimental mathematics[1]—application of high-performance computing technology to research questions in pure and applied mathematics, including

"It says it's sick of doing things like inventories and payrolls, and it wants to make some breakthroughs in astrophysics."

ScienceCartoonsPlus.com.

physicists, legal scholars, journal editors, and funding agency officials representing academia, government labs, industry research, and all points in between. While



Renew SIAM · Contact Us · Site Map · Join SIAM
Society for Industrial and Applied Mathematics
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**

# 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
  - Galaxy
  - Sumatra
  - Kepler
  - GenePattern
  - Taverna
  - CDE
  - Paper Mâché
  - Pegasus
  - IPython Notebook
- Embedded Publishing:
  - Verifiable Computational Research
  - Collage Authoring Environment
  - SOLE
  - SHARE
  - knitR
  - 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.
- July 29, 2014: Notice of Request for Information "Strategy for American Innovation"

# 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.
- Sensenbrenner introduced "Public Access to Science," Sept 19, 2013.
- Hearing on Research Integrity and Transparency by the House Science, Space, and Technology Committee (March 5, 2013).
- *Reproducibility cannot be an unfunded mandate.*

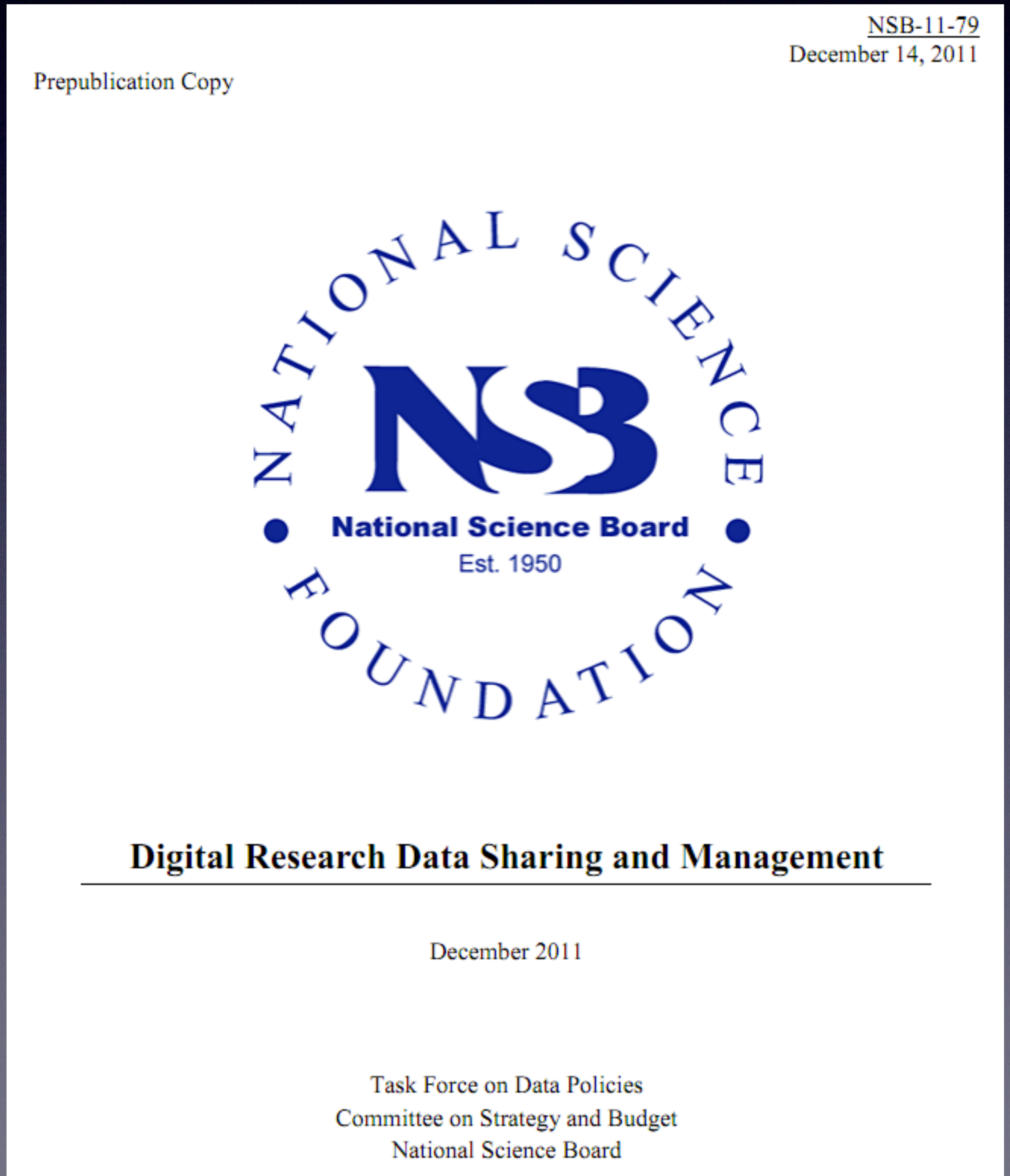
# NAS Data Sharing Report

SHARING PUBLICATION-RELATED DATA AND MATERIALS

Responsibilities of Authorship in the Life Sciences

- 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.”

# National Science Board Report



NSB-11-79
December 14, 2011

Prepublication Copy

NATIONAL SCIENCE FOUNDATION

NSB

National Science Board
Est. 1950

**Digital Research Data Sharing and Management**

December 2011

Task Force on Data Policies
Committee on Strategy and Budget
National Science Board

"Digital Research Data Sharing and Management," December 2011.

http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

# 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.

Example: Google Flu Trends results: worked at first, but what happened? (Lazer et al. "The Parable of Google Flu: Traps in Big Data Analysis" *Science*, 2014)

# References

"Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," PLoS ONE, June 2013

"Reproducible Research," guest editor for Computing in Science and Engineering, July/August 2012.

"Reproducible Research: Tools and Strategies for Scientific Computing," July 2011.

"Enabling Reproducible Research: Open Licensing for Scientific Innovation," 2009.

available at http://www.stodden.net

# 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.

Science AAAS
AAAS.ORG | FEEDBACK | HELP | LIBRARIANS
All Science Journals
Enter Search Ter
GUEST ALERTS ACCESS RIGH
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
‹ Prev | Table of Contents | Next ›
Read Full Text to Comment (8)

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

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.

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
PDF
Rights & Permissions

*NATURE* | EDITORIAL

### Announcement: Reducing our irreproducibility

24 April 2013

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

The Economist
OCTOBER 19TH–25TH 2013
Economist.com
Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar
HOW SCIENCE GOES WRONG
99
Einsteinium

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
PDF
Citation
Reprints
Rights & permissions
Article metrics

*NATURE* | EDITORIAL

### 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.**