Reliability and Reproducibility in Cyber Security Research

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

Workshop on the Science of Cyber-Security
67th Meeting of IFIP Working Group 10.4 on Dependable Computing and Fault Tolerance
Tortworth, Bristol, UK
January 23, 2015

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.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 result
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

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

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

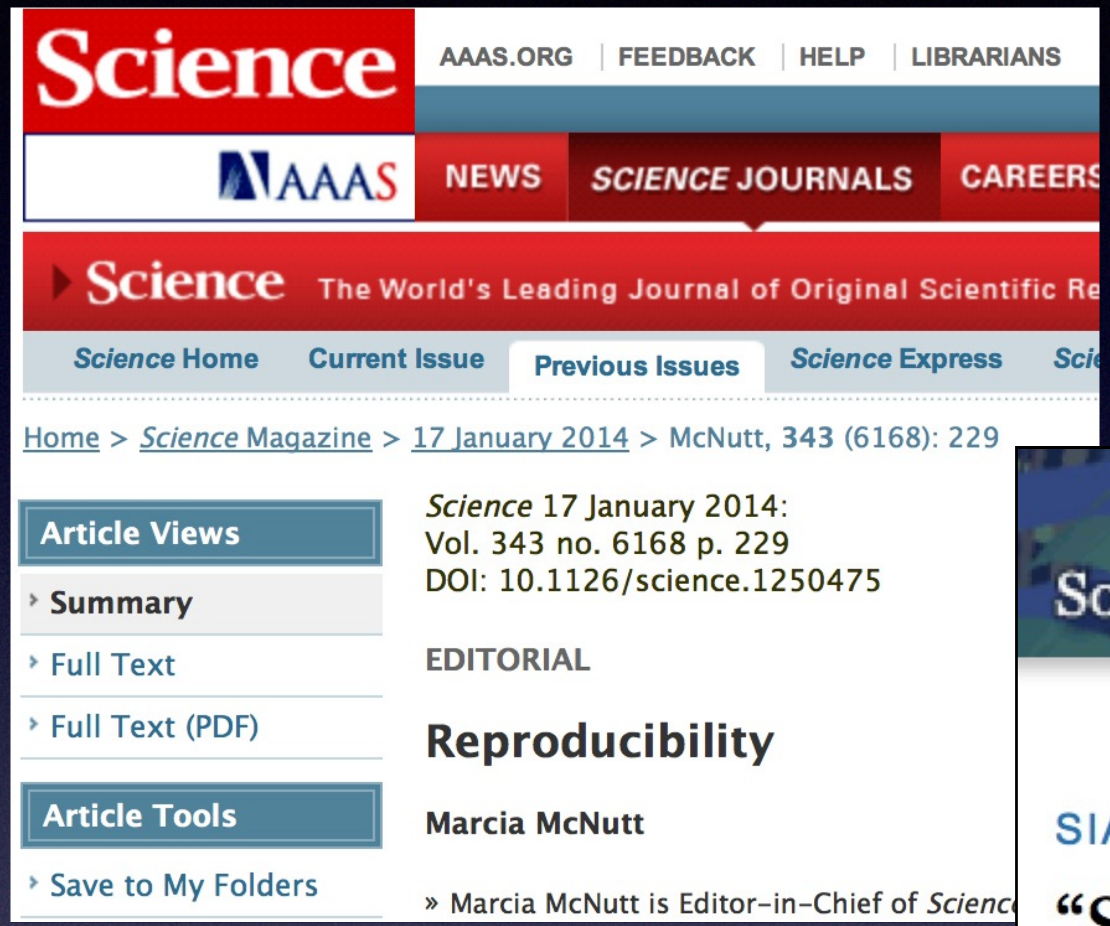
Nature **483**, 509 (29 March 2012) | doi:10.1038/483509a
Published online 28 March 2012

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.

Unpacking "Reproducibility"

"Empirical Reproducibility"



Science | AAAS.ORG | FEEDBACK | HELP | LIBRARIANS

AAAS | NEWS | SCIENCE JOURNALS | CAREERS

Science The World's Leading Journal of Original Scientific Re

Science Home | Current Issue | Previous Issues | Science Express | Sci

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

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

Article Tools
- Save to My Folders

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

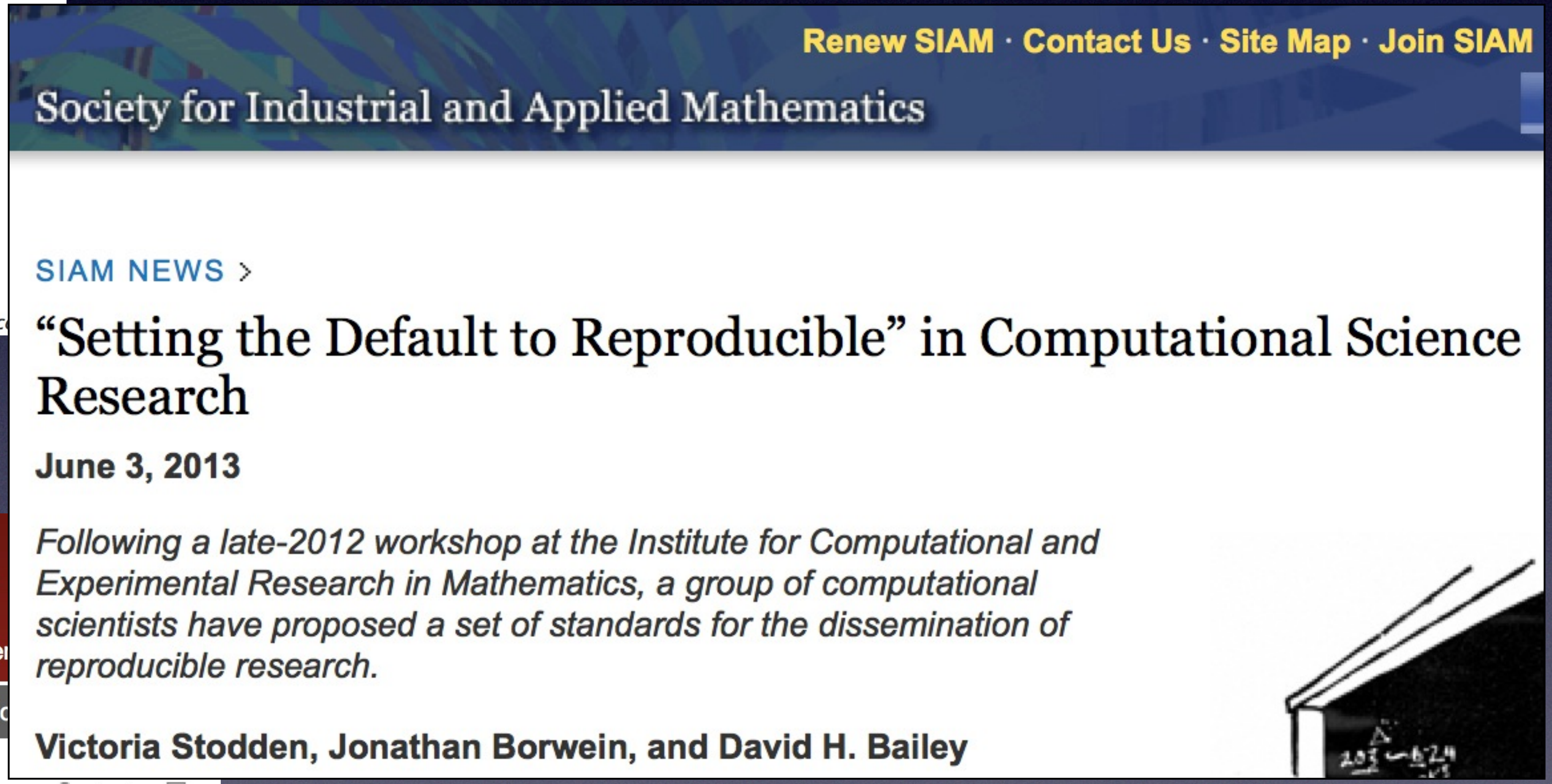
EDITORIAL

Reproducibility

Marcia McNutt

» Marcia McNutt is Editor-in-Chief of *Scienc*

"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

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Curre

Archive > Volume 506 > Issue 7487 > News Feature > Artic

NATURE | NEWS FEATURE

Scientific method: Statistical errors

P **values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.**

Regina Nuzzo

12 February 2014

"Statistical Reproducibility"

V. Stodden, IMS Bulletin (2013)

1. Empirical Reproducibility

Skepticism requires that the claim can be independently verified,

This in turn requires 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.

Examples

Cell Reports
Commentary

Cell PRESS
Open ACCESS

Sorting Out the FACS: A Devil in the Details

William C. Hines,[1,5,*] Ying Su,[2,3,4,5,*] Irene Kuhn,[1] Kornelia Polyak,[2,3,4,5] and Mina J. Bissell[1,5]
[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[3]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[4]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[5]These authors contributed equally to this work
*Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.celrep.2014.02.021

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and specimens, our two laboratories quite reproducibly were unable to replicate each other's fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of measuring up to—and beyond—an astonishing 50 individual markers per cell (Cheung and Utz, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.'s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.'s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for

Reproducibility Issues in Research with Animals and Animal Models

The missing "*R*": Reproducibility in a Changing Research Landscape

A workshop of the Roundtable on Science and Welfare in Laboratory Animal Use

National Academy of Sciences, NAS 125
2100 C Street NW, Washington DC
June 4-5, 2014

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues.

Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.

2. Computational Reproducibility

Traditionally two branches of the scientific method:

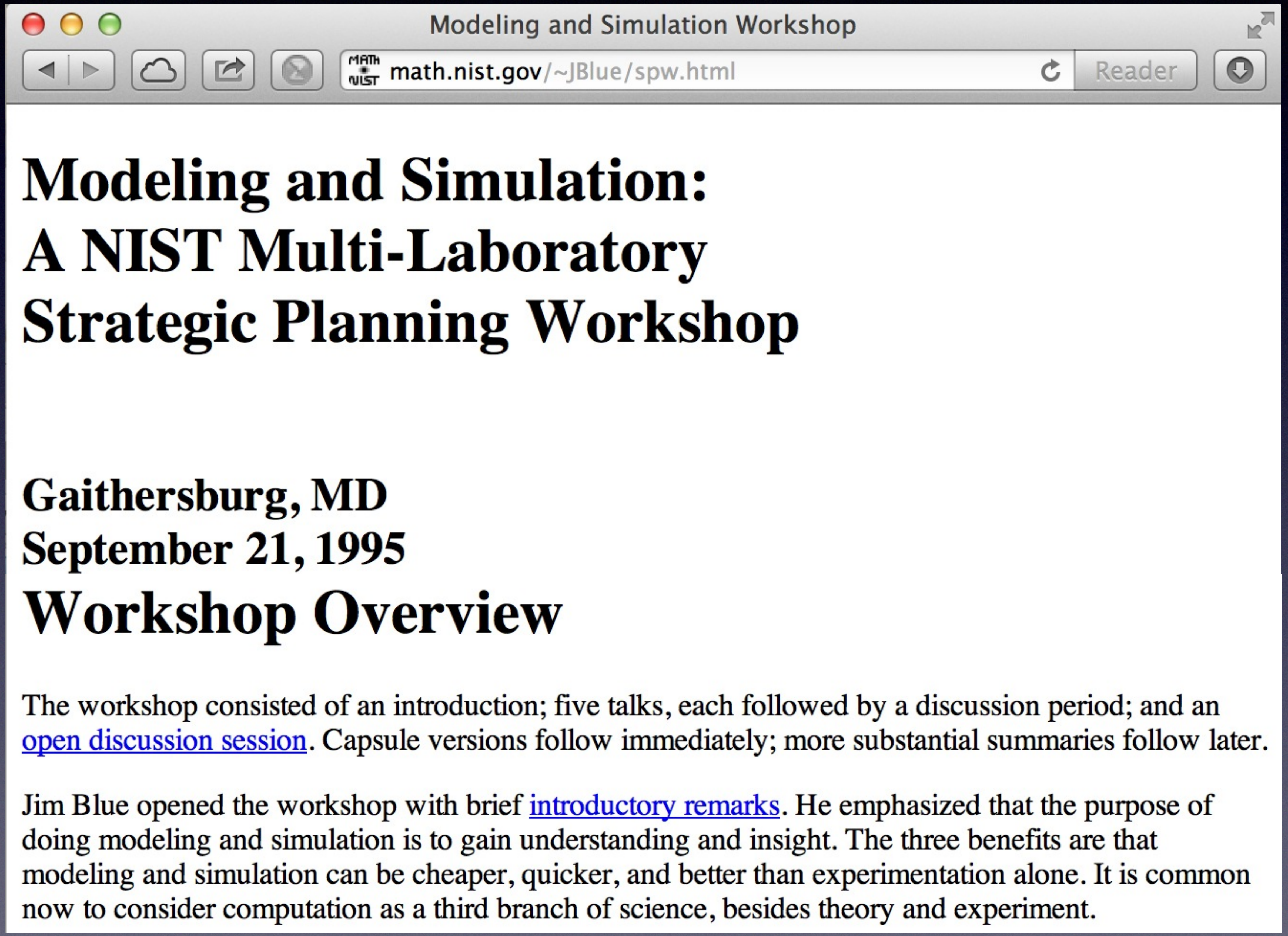
Branch 1 (deductive): mathematics, formal logic,

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

Many claim the emergence of new branches:

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

Commonly believed...


Modeling and Simulation Workshop

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

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

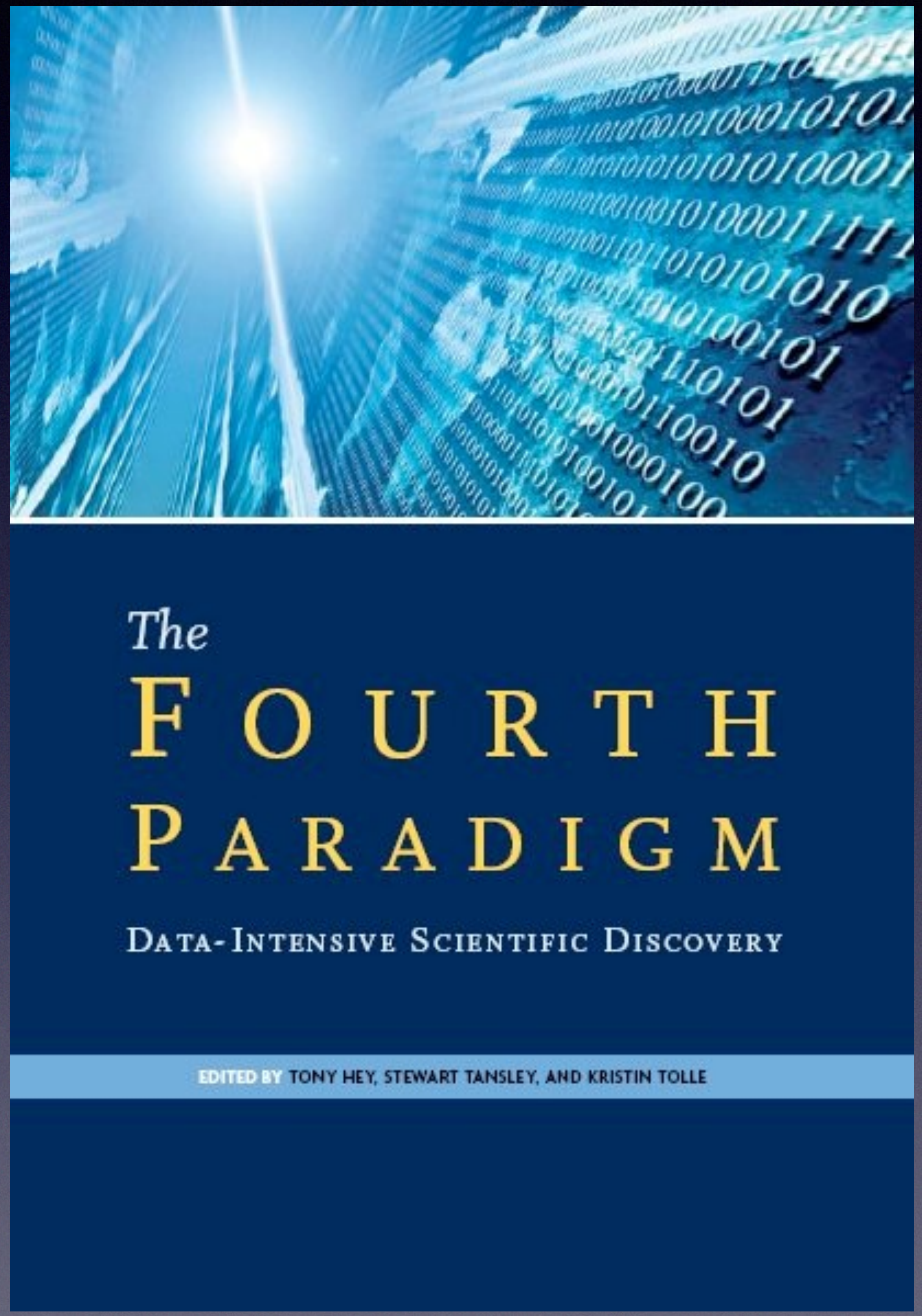
Gaithersburg, MD
September 21, 1995

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."



"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.

Scoping the Issue

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): of 500 papers studied, 9% had full primary raw data deposited.

Stodden (to come): estimates that the computations in 27% of scientific articles published in *Science* today are reproducible.

Examples: Voodoo Correlation in fMRI, Nevins/Potti Duke case, …

ICERM Workshop

ICERM

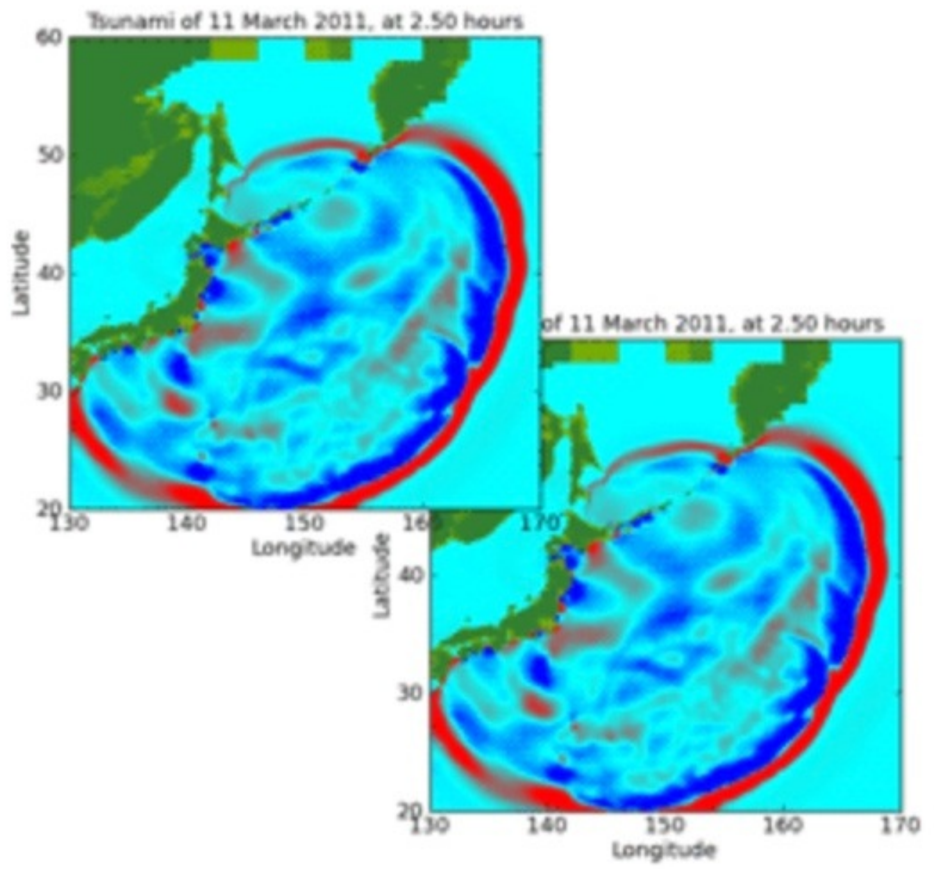
Search

Home | Programs & Events | Participate | Proposals | Resources | For Visitors | People | News | Diversity | Support ICERM

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

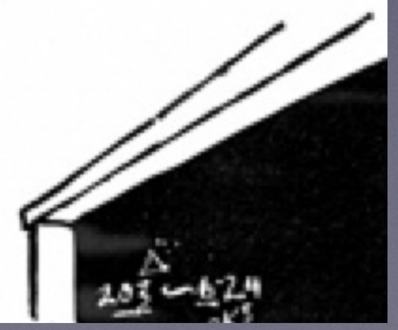
Society for Industrial and Applied Mathematics

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

Criterion	Definition
Assertions (#1)	A precise statement of assertions to be made in the paper.
Comp. Approach (#2)	A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.
Software Cited (#3 & 4)	Complete statements of, or references to, every algorithm employed, and salient details of auxiliary software (both research and commercial software) used in the computation.
Hardware Discussed (#5)	Salient details of the test environment, including hardware, system software and the number of processors utilized.
Analysis (#6)	Salient details of data reduction and statistical analysis methods.
Parameter Discussed (#7)	Discussion of the adequacy of parameters such as precision level and grid resolution.
Parameters Given (#7)	Were necessary run parameters given?
Results (#8)	Full statement (or at least a valid summary) of experimental results.
Available Code (#10)	Availability of computer code, input data and output data, with some reasonable level of documentation.
Functions Calls	Which precise functions were called, with what settings?
Comp. Instructions (#12)	Instructions for repeating computational experiments described in the paper.
Alternate Avenues (#14)	Avenues of exploration examined throughout development, including information about negative findings.
Citation (#15)	Proper citation of all code and data used, including that generated by the authors.

Supporting Computational Science

- Dissemination Platforms:

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

- Workflow Tracking and Research Environments:

  VisTrails   Kepler   CDE
  Galaxy   GenePattern   Jupyter / IPython Notebook
  Sumatra   Taverna   Pegasus

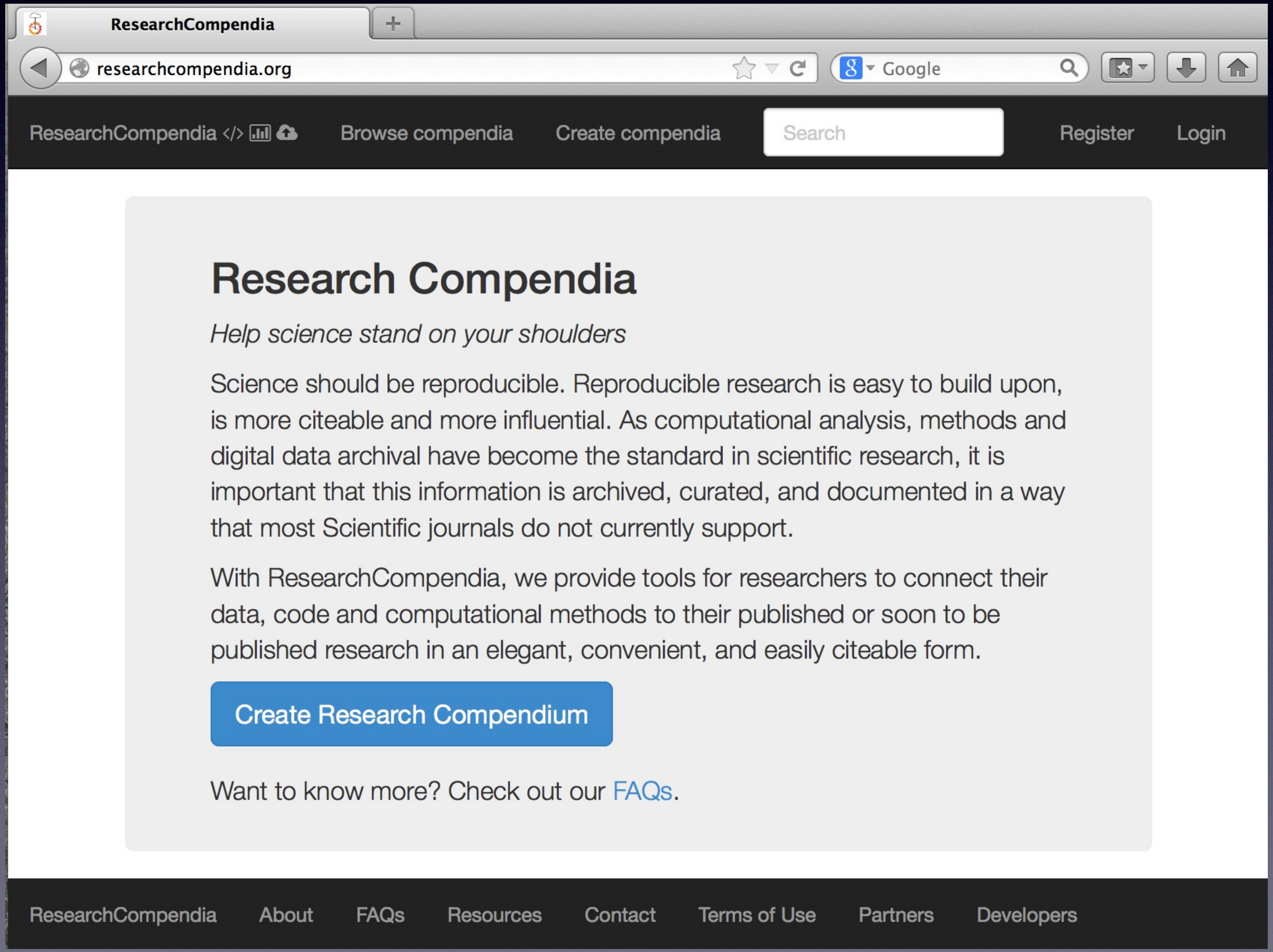
- Embedded Publishing:

  Verifiable Computational Research   SOLE   knitR
  Collage Authoring Environment   SHARE   Sweave

Research Compendia

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

- link data/code to published claims,
- enable re-use,
- sharing guide for researchers,
- certification of 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.

</> code | data | article

Compendium Type:: article
Content License:: CC0
Code License:: MIT

ResearchCompendia

labs.researchcompendia.org/compendia/2014.8/

Random survival forests for high-dimensional data

Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, Andy J. Minn

Code and Data Abstract

Minimal depth is a dimensionless order statistic that measures the predictiveness of a variable in a survival tree. It can be used to select variables in high-dimensional problems using Random Survival Forests (RSF), a new extension of Breiman's Random Forests (RF) to survival settings. We review this methodology and demonstrate its use in high-dimensional survival problems using a public domain R-language package randomSurvivalForest. We discuss effective ways to regularize forests and discuss how to properly tune the RF parameters 'nodesize' and 'mtry'. We also introduce new graphical ways of using minimal depth for exploring variable relationships.

</> code | data | Article | Verify

Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, Andy J. Minn, et al. 2011. "Random survival forests for high-dimensional data." *Statistical Analysis and Data Mining*. 4 (1) 115–132. doi:10.1002/sam.10103. Retrieved 12/04/2014 from labs.researchcompendia.org/compendia/2014.8/

Code DOI: doi:10.7938/M1H41PBB.
Data DOI: doi:10.7938/M1CC0XMM.

Compendium Type: Journal or Magazine Articles
Primary Research Field: Computer and Information Sciences
Secondary Research Field: Mathematics
Content License: Public Domain Mark
Code License: MIT License

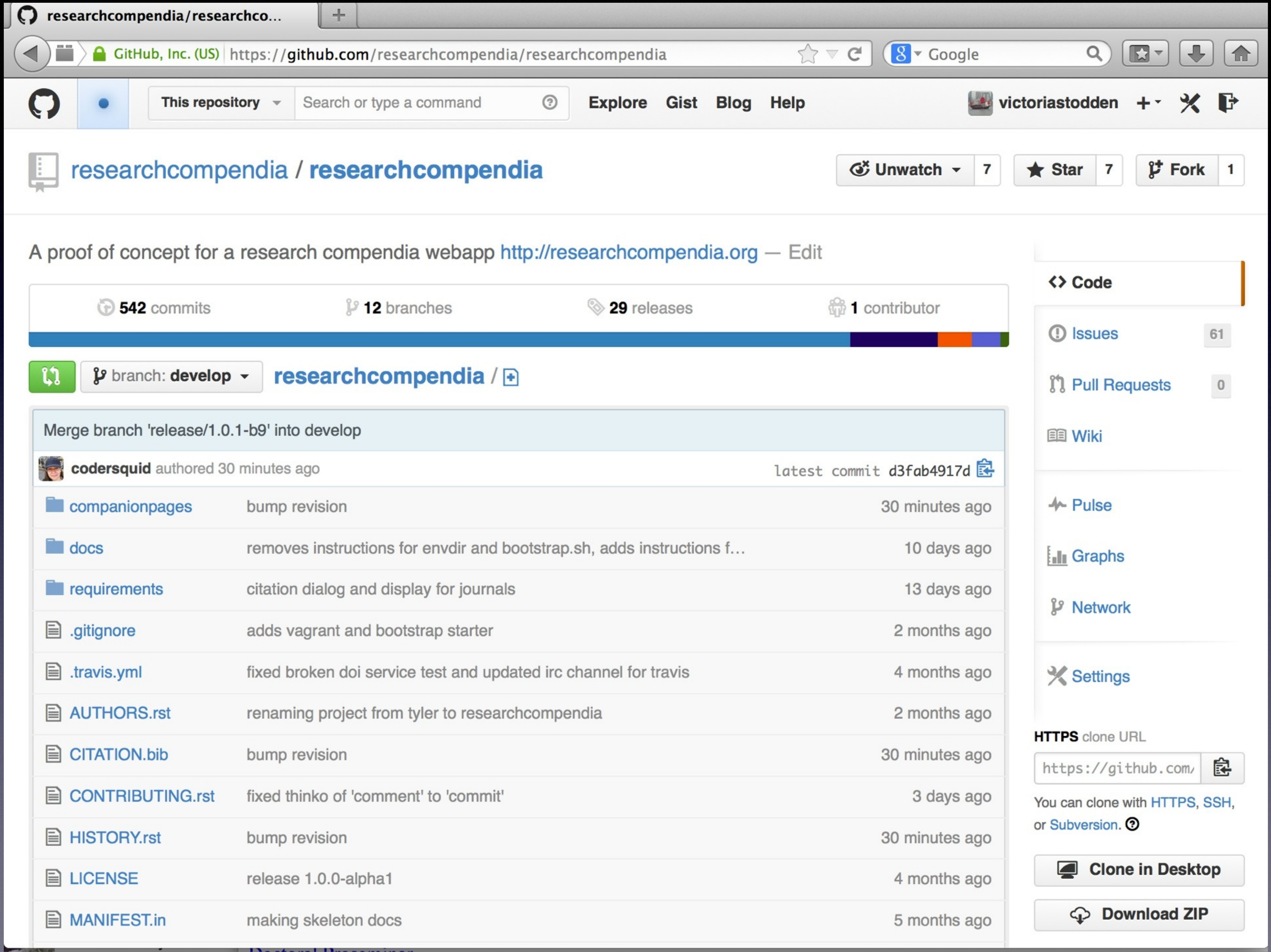
Verification

</> verification code

Recent Verifications

Verification run 41 created March 21, 2014, 9:43 a.m.

Verification run 40 created March 21, 2014, 4:19 a.m.


researchcompendia/researchco...
GitHub, Inc. (US) https://github.com/researchcompendia/researchcompendia
Google
This repository
Search or type a command
Explore
Gist
Blog
Help
victoriastodden
researchcompendia / researchcompendia
Unwatch 7
Star 7
Fork 1
A proof of concept for a research compendia webapp http://researchcompendia.org — Edit
542 commits
12 branches
29 releases
1 contributor
branch: develop
researchcompendia / +
Merge branch 'release/1.0.1-b9' into develop
codersquid authored 30 minutes ago
latest commit d3fab4917d
companionpages bump revision 30 minutes ago
docs removes instructions for envdir and bootstrap.sh, adds instructions f… 10 days ago
requirements citation dialog and display for journals 13 days ago
.gitignore adds vagrant and bootstrap starter 2 months ago
.travis.yml fixed broken doi service test and updated irc channel for travis 4 months ago
AUTHORS.rst renaming project from tyler to researchcompendia 2 months ago
CITATION.bib bump revision 30 minutes ago
CONTRIBUTING.rst fixed thinko of 'comment' to 'commit' 3 days ago
HISTORY.rst bump revision 30 minutes ago
LICENSE release 1.0.0-alpha1 4 months ago
MANIFEST.in making skeleton docs 5 months ago
Code
Issues 61
Pull Requests 0
Wiki
Pulse
Graphs
Network
Settings
HTTPS clone URL
https://github.com/
You can clone with HTTPS, SSH, or Subversion.
Clone in Desktop
Download ZIP

Journal Policy?

- Journal Policy setting study design:
- 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.
- Create dataset with ISI information (impact factor, citations, publisher) and supplement with publication policies as listed on journal websites, in June 2011 and June 2012.

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)

Findings

- Changemakers are journals with high impact factors.
- Progressive policies are not widespread, but being adopted rapidly.
- Close relationship between the existence of a supplemental materials policy and a data policy.
- No statistically significant relationship between data and code policies and open access policy.
- Data and supplemental material policies appear to lead software policy.

Three sections:

1. Tools
2. Practices and Guidelines
3. Platforms

Chapters available for download: https://osf.io/s9tya/wiki/

3. Statistical Reproducibility

- False discovery, chasing significance, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.
- Low power, poor experimental design,
- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,
- Poor statistical methods (nonrandom sampling, inappropriate tests or models, model misspecification..)
- Model robustness to parameter changes and data perturbations,
- Investigator bias toward previous findings; conflicts of interest.

Really Reproducible Research

- “Really Reproducible Research” (1992) 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

- NB. Reproducing the *computational steps* vs replicating the experiments *independently* including data collection and software implementation.

Journal Requirements

In January 2014 Science enacted new policies. The will check for:

1. a "data-handling plan" i.e. how outliers will be dealt with,
2. sample size estimation for effect size,
3. whether samples are treated randomly,
4. whether experimenter blind to the conduct of the experiment.

Statisticians added to the Board of Reviewing Editors.

Updating Reproducibility

- Failings of traditional reporting methods vs adaptation of standards to accommodate changes in the research process.
- Interaction of computational systems in the process of scientific discovery viz. reliability of parameter estimates and models.
- Benchmarking and testing: either nonexistent or over reliance on inappropriate benchmarks (see e.g. http://www.in-cites.com/scientists/DrDavidDonoho.html)
- Collective action problem: coordination of researcher incentives, universities, funding agencies, journals, scientific societies, legal and policy environment,

References

Implementing Reproducible Research, CRC Press 2014.

"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.

"Reproducible Research in Computational Harmonic Analysis", IEEE Computing in Science and Engineering, 11(1), January 2009

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

available at http://www.stodden.net

Data / Code Sharing Practices

Survey of the NIPS community:

- 1,758 NIPS registrants up to and including 2008,
- 1,008 registrants when restricted to .edu registration emails,
- After piloting, the final survey was sent to 638 registrants,
- 37 bounces, 5 away, and 3 in industry, gave a final response rate was 134 of 593 or 23%.
- Queried about reasons for sharing or not sharing data/code associated with their NIPS paper.

Sharing Incentives

Code		Data
91%	Encourage scientific advancement	81%
90%	Encourage sharing in others	79%
86%	Be a good community member	79%
82%	Set a standard for the field	76%
85%	Improve the calibre of research	74%
81%	Get others to work on the problem	79%
85%	Increase in publicity	73%
78%	Opportunity for feedback	71%
71%	Finding collaborators	71%

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

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)

Legal Barriers: Copyright

“To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.” (U.S. Const. art. I, §8, cl. 8)

- Original expression of ideas falls under copyright *by default* (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
    - reproduce the work
    - prepare derivative works based upon the original

Exceptions and Limitations: Fair Use.

Responses Outside the Sciences 1: Open Source Software

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

- Hundreds of open source software licenses:
  - GNU Public License (GPL)
  - (Modified) BSD License
  - MIT License
  - Apache 2.0 License
  - ... see http://www.opensource.org/licenses/alphabetical

Responses Outside the Sciences 2: Creative Commons

- Founded in 2001, by Stanford Law Professor Larry Lessig, MIT EECS Professor Hal Abelson, and advocate Eric Eldred.
- Adapts the Open Source Software approach to artistic and creative digital works.

Response from Within the Sciences

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

- A suite of license recommendations for computational science:
    - Release media components (text, figures) under CC BY,
    - Release code components under Modified BSD or similar,
    - Release data to public domain or attach attribution license.

➡ Remove copyright's barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.

Winner of the Access to Knowledge Kaltura Award 2008

Copyright and Data

- Copyright adheres to raw facts in Europe.
- In the US raw facts are not copyrightable, but the original "selection and arrangement" of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
- the possibility of a residual copyright in data (attribution licensing or public domain certification).
- Law doesn't match reality on the ground: What constitutes a "raw" fact anyway?