Enabling Reproducibility in Computational and Data-enabled Science

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

Workshop II: HPC and Data Science for Scientific Discovery
Part of the Long Program Science at Extreme Scales: Where Big Data Meets Large-Scale Computing

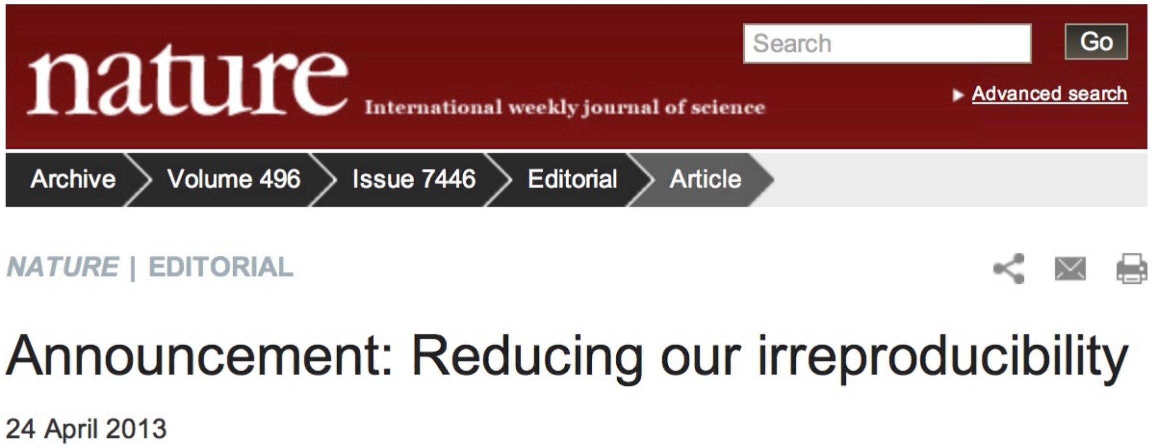
Institute for Pure and Applied Mathematics, UCLA
October 19, 2018

Agenda

1. Framing Reproducibility in Data-enabled Scientific Discovery
2. A (Very) Brief History of Recent Community Efforts
3. Infrastructure Solutions: “WholeTale” and “ezDMP”
4. How much of a Problem is Computational Reproducibility?

Parsing Reproducibility

“Empirical Reproducibility”

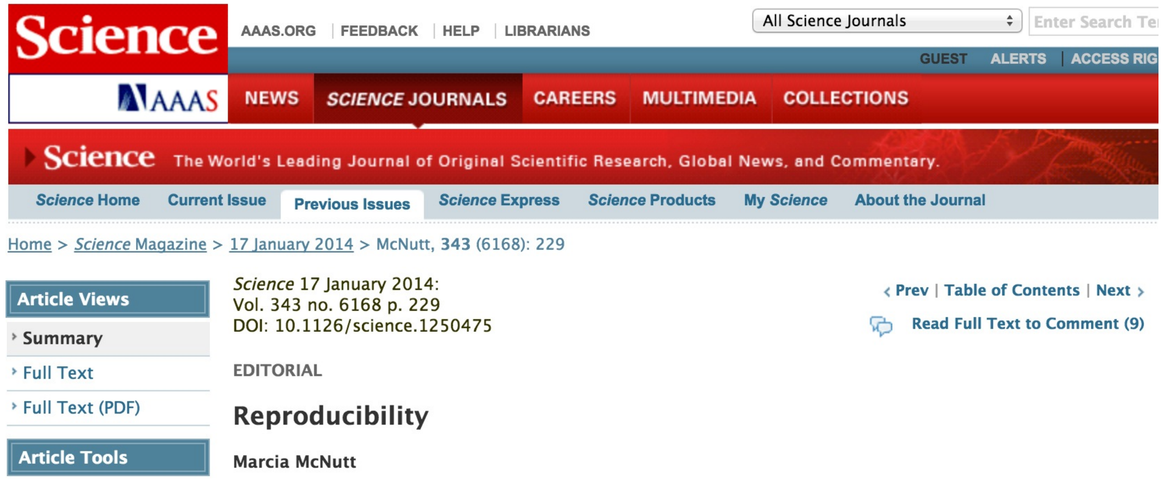


NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

“Statistical Reproducibility”



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

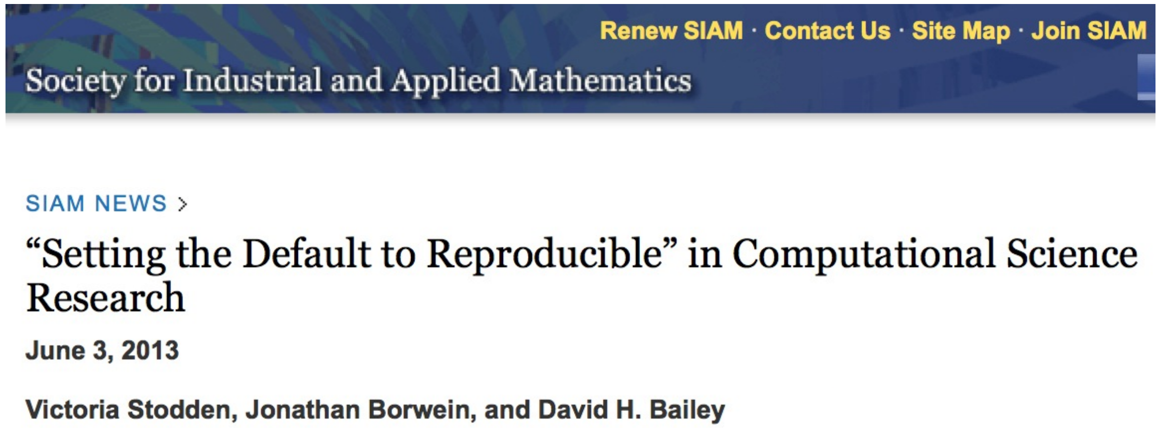
Science 17 January 2014:
Vol. 343 no. 6168 p. 229
DOI: 10.1126/science.1250475

EDITORIAL

Reproducibility

Marcia McNutt

“Computational Reproducibility”


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

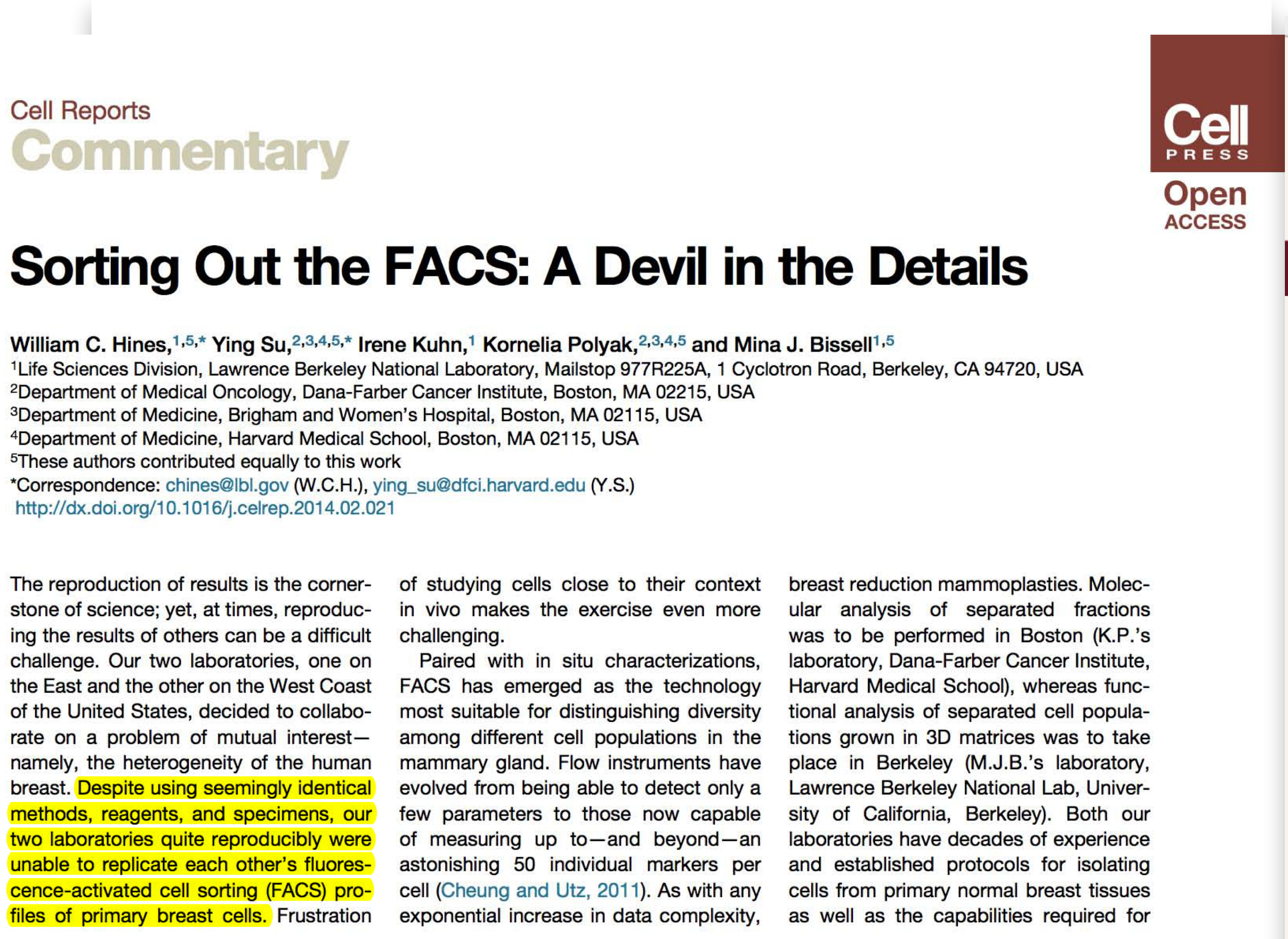
“Setting the Default to Reproducible” in Computational Science Research

June 3, 2013

Victoria Stodden, Jonathan Borwein, and David H. Bailey

V. Stodden, IMS Bulletin (2013)

Empirical Reproducibility



Cell Reports

Commentary

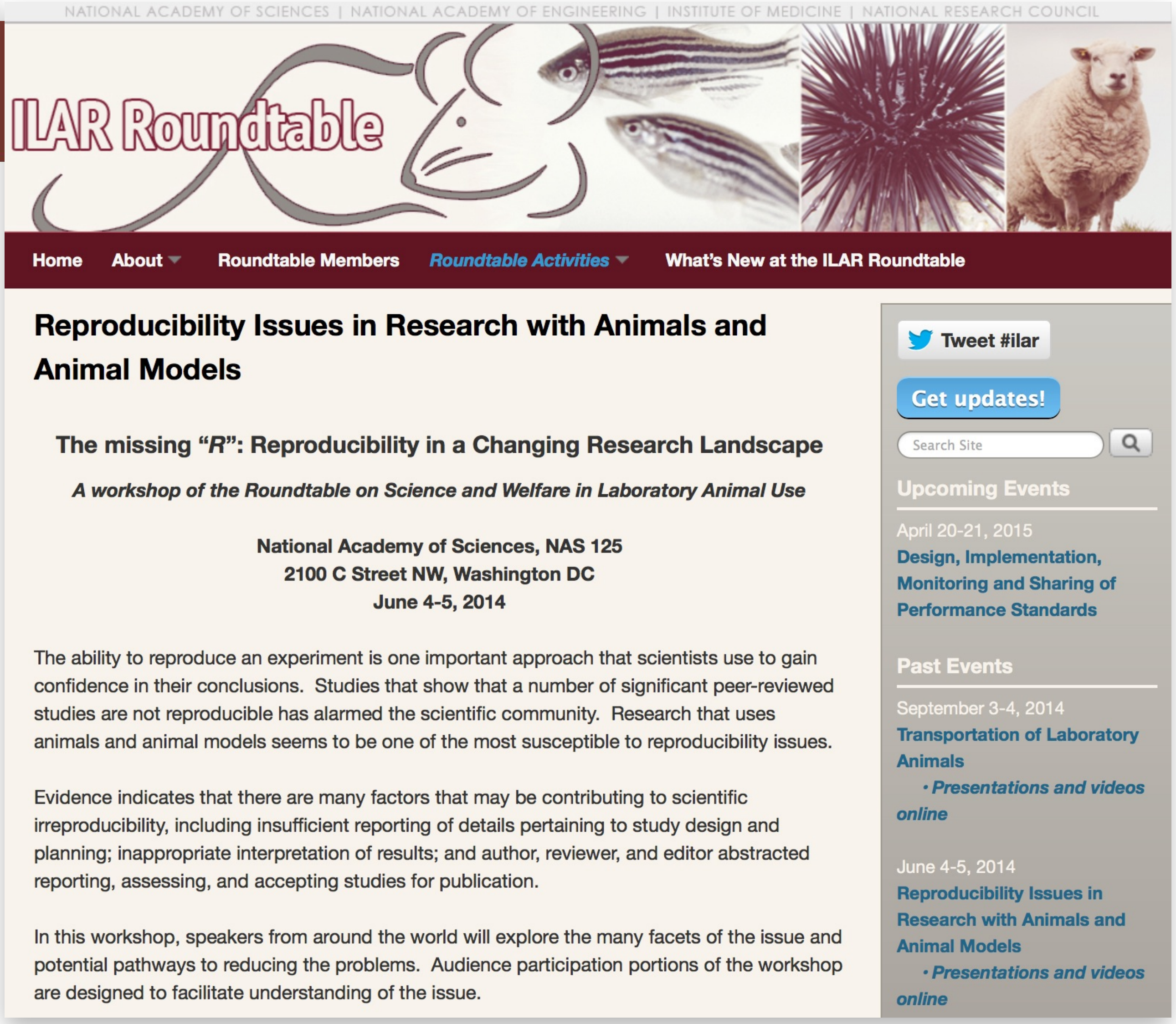
Sorting Out the FACS: A Devil in the Details

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



NATIONAL ACADEMY OF SCIENCES | NATIONAL ACADEMY OF ENGINEERING | INSTITUTE OF MEDICINE | NATIONAL RESEARCH COUNCIL

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.

Upcoming Events

April 20-21, 2015
Design, Implementation, Monitoring and Sharing of Performance Standards

Past Events

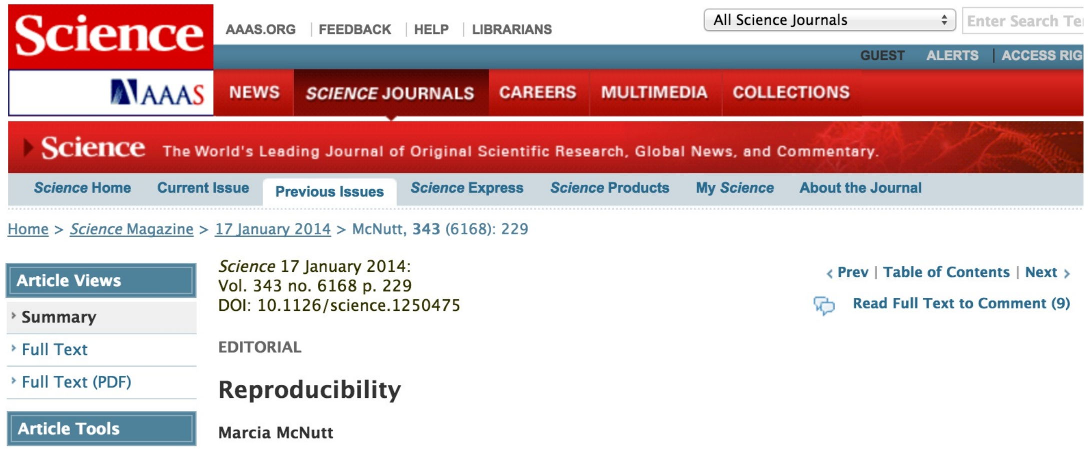
September 3-4, 2014
Transportation of Laboratory Animals
• *Presentations and videos online*

June 4-5, 2014
Reproducibility Issues in Research with Animals and Animal Models
• *Presentations and videos online*

Statistical Reproducibility

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

Statistical Reproducibility



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

EDITORIAL

Reproducibility

Marcia McNutt

In January 2014 Science enacted new manuscript submission requirements:

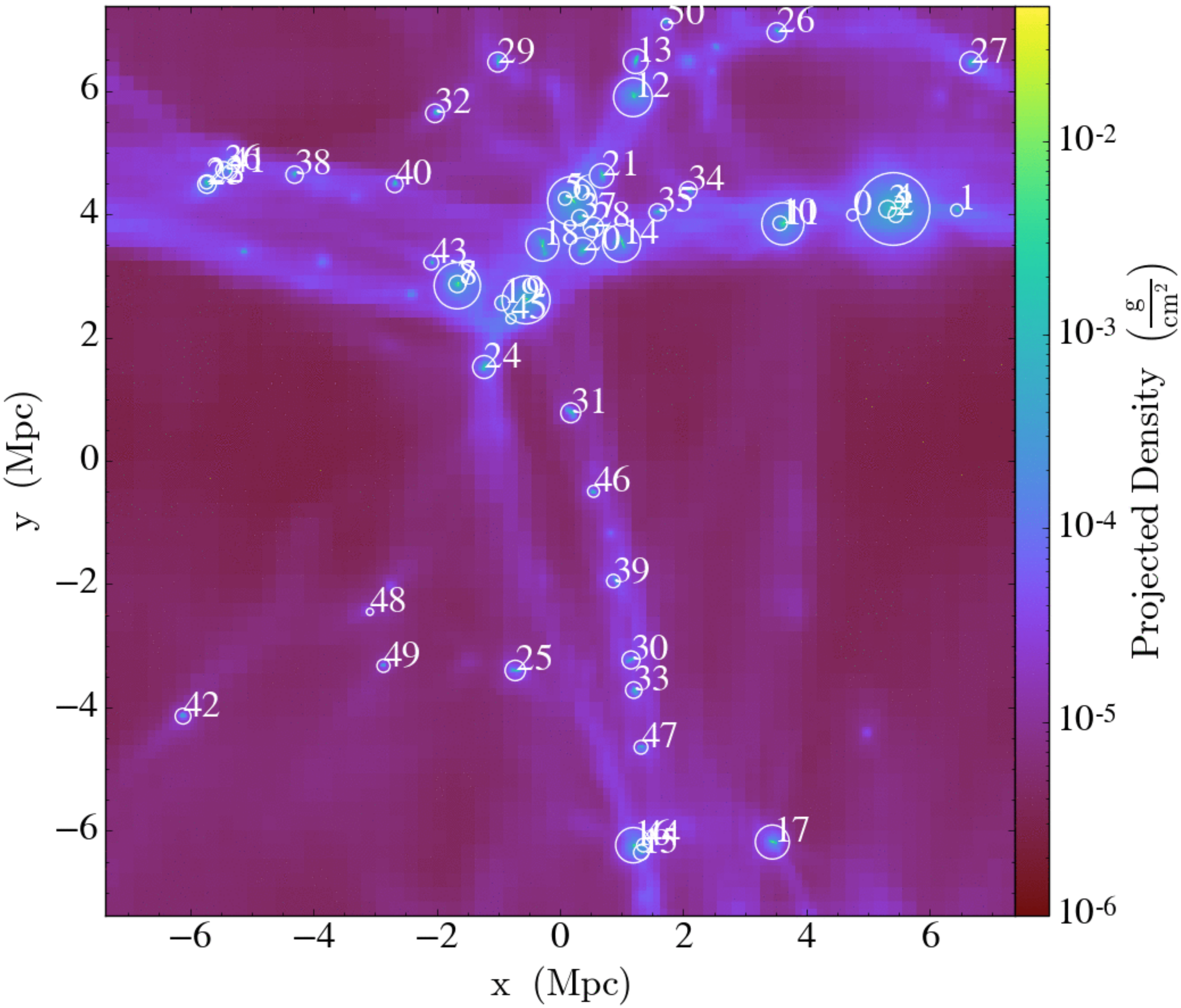
- a “data-handling plan” i.e. how outliers will be dealt with,
- sample size estimation for effect size,
- whether samples are treated randomly,
- whether experimenter blind to the conduct of the experiment.

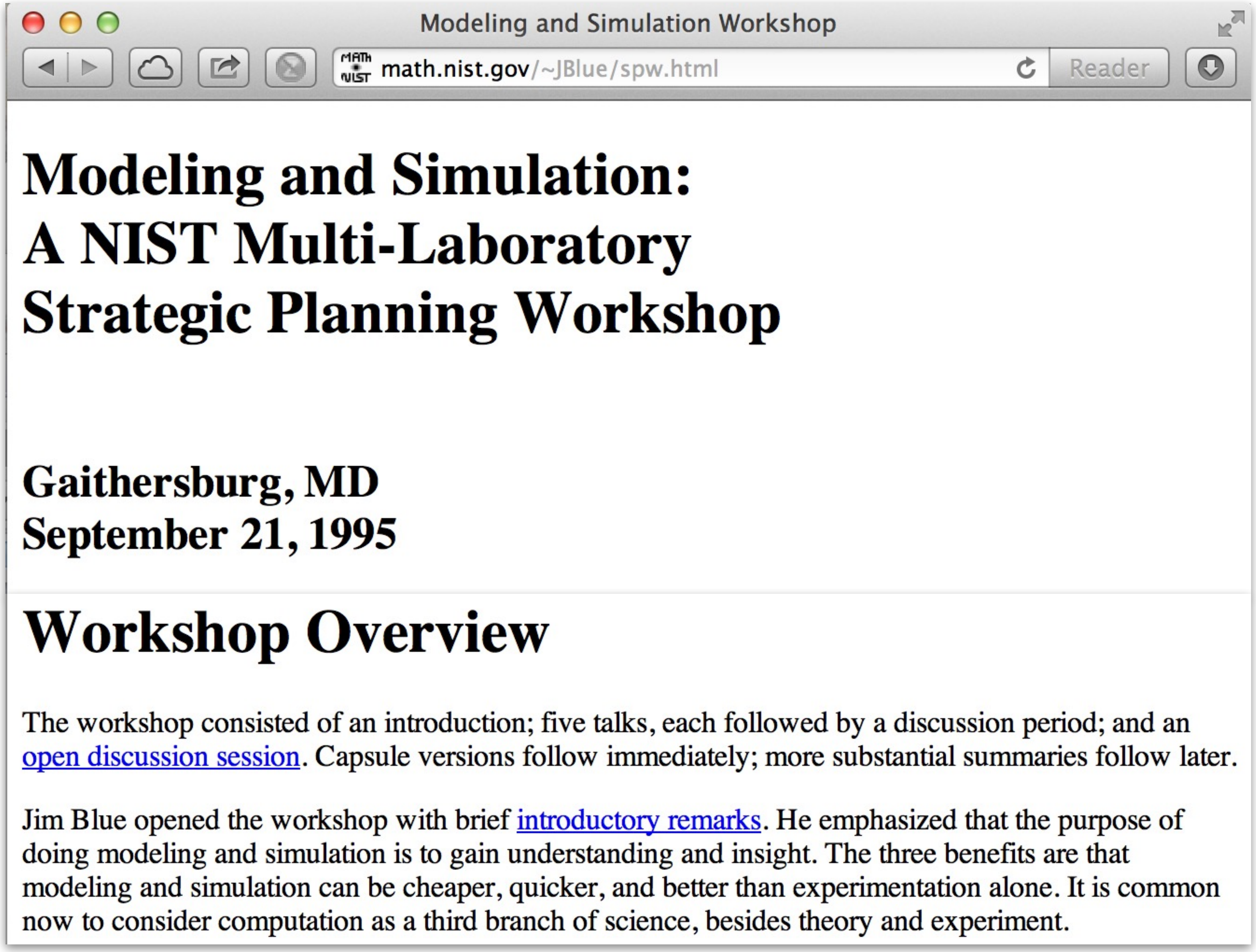
Also added statisticians to the Board of Reviewing Editors.

Computational Reproducibility

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
http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf


50
26
29
13
27
12
32
36
41
38
40
21
34
5
6
35
37
28
10
0
3
2
1
18
20
14
43
8
19
45
24
31
46
39
48
49
25
30
33
42
47
16
44
15
17
y (Mpc)
x (Mpc)
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
Projected Density $\left(\frac{\mathrm{g}}{\mathrm{cm}^2}\right)$
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}


Modeling and Simulation Workshop
math.nist.gov/~JBlue/spw.html
Reader

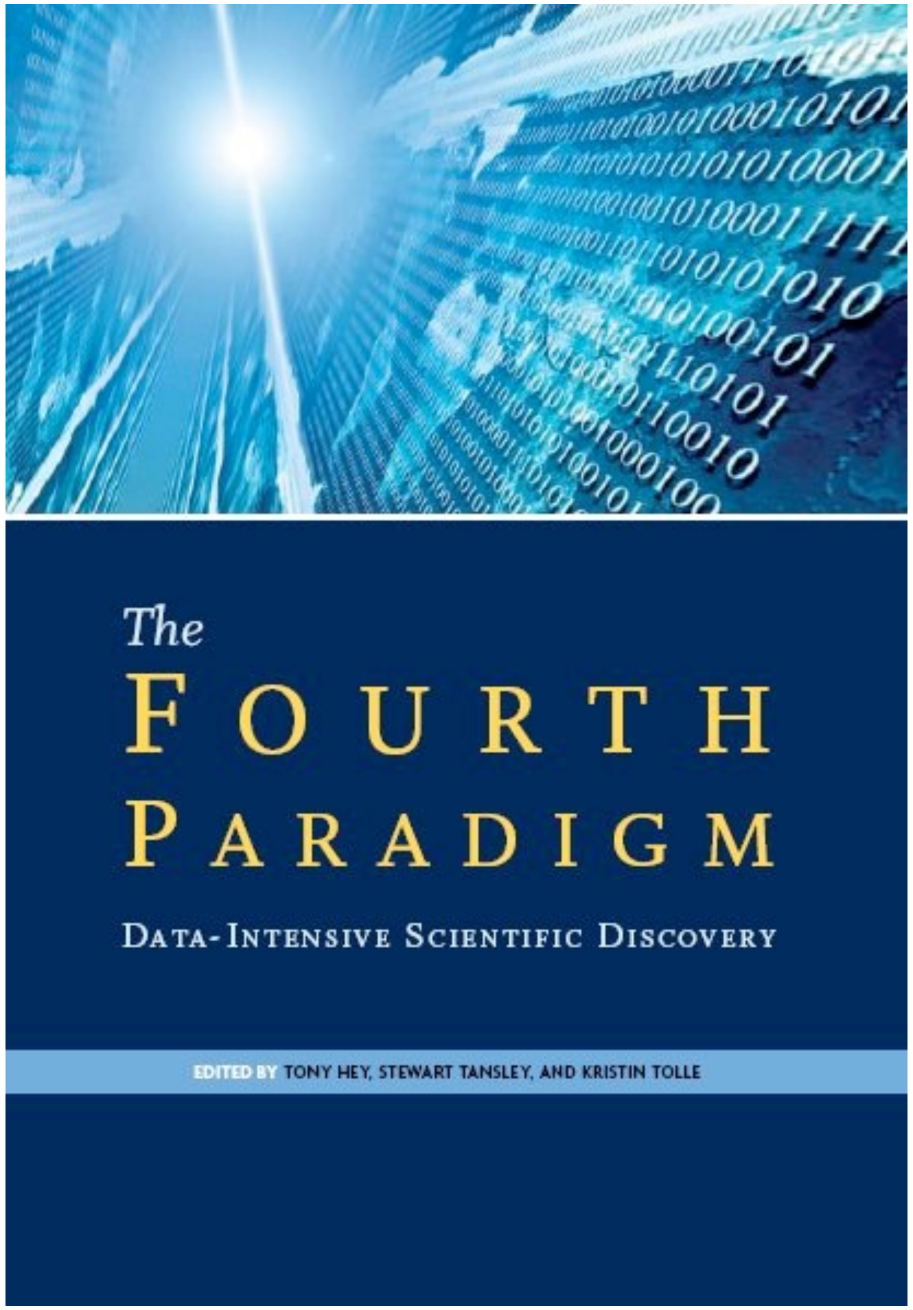
Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

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.



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

"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 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 et al. 2009), until the development of comparable standards.

The digital age in science

Claim 1:

Virtually all published discoveries today have a computational component.

Claim 2:

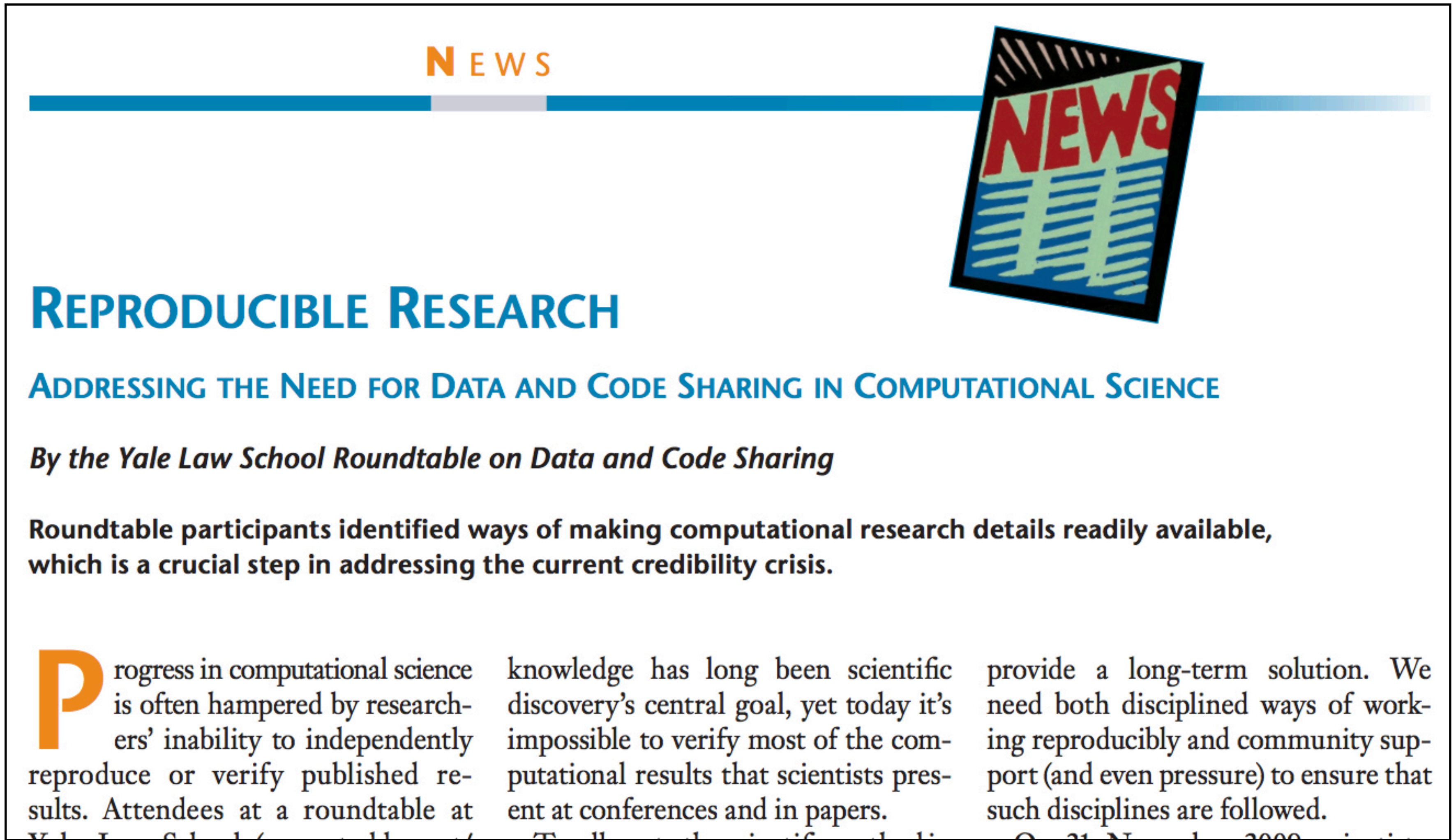
There is a mismatch between the traditional scientific process and computation, leading to reproducibility concerns.

A (Very) Brief History of Recent Community Efforts..

Yale 2009

Inspired by the Bermuda Principles, “Data and Code Sharing Roundtable” on November 21, 2009. See http://stodden.net/RoundtableNov212009

We collectively produced the Data and Code Sharing Declaration including a description of the problem, proposed solutions, and dream goals we’d like to see.



NEWS

REPRODUCIBLE RESEARCH

ADDRESSING THE NEED FOR DATA AND CODE SHARING IN COMPUTATIONAL SCIENCE

By the Yale Law School Roundtable on Data and Code Sharing

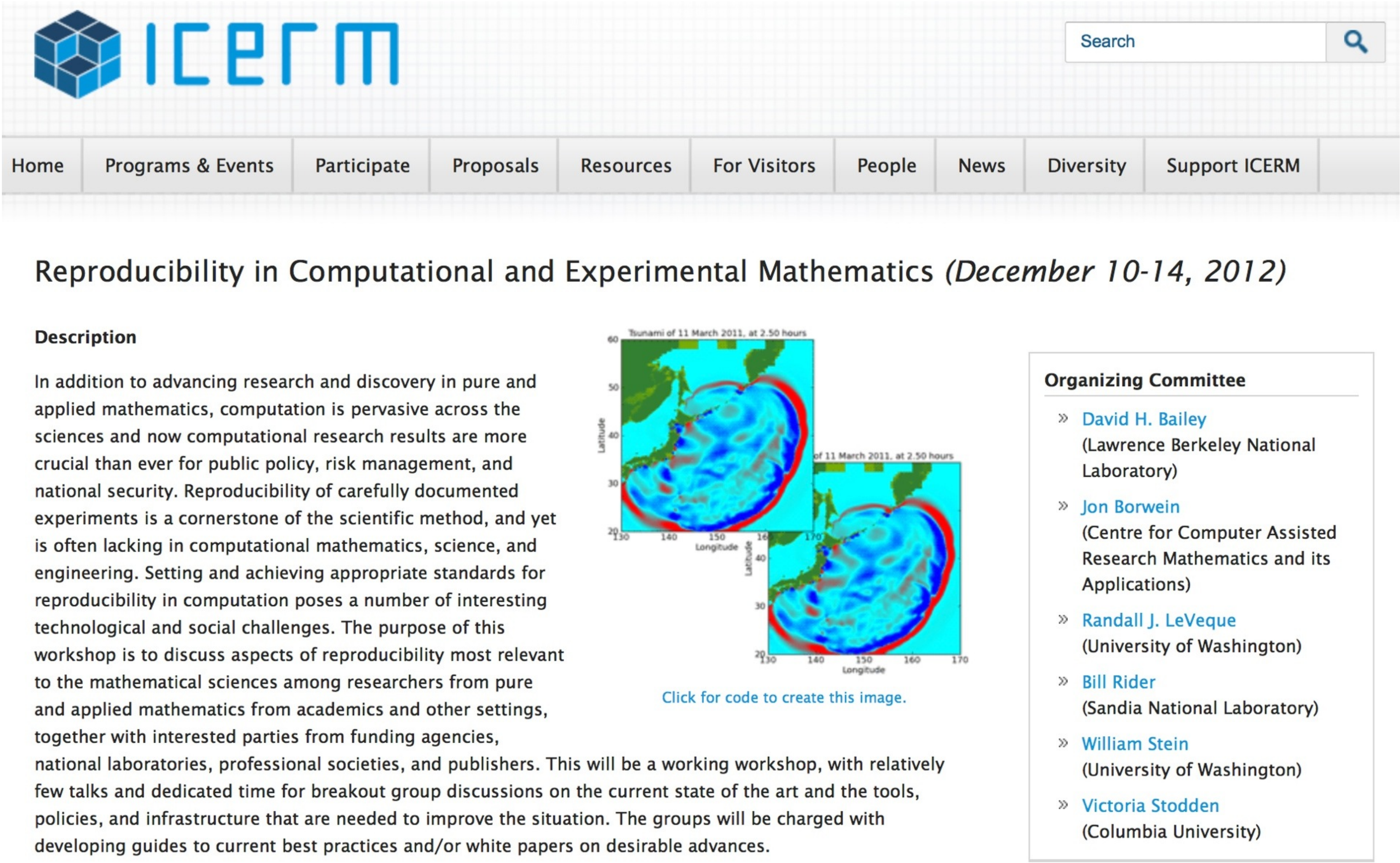
Roundtable participants identified ways of making computational research details readily available, which is a crucial step in addressing the current credibility crisis.

Progress in computational science is often hampered by researchers’ inability to independently reproduce or verify published results. Attendees at a roundtable at

knowledge has long been scientific discovery’s central goal, yet today it’s impossible to verify most of the computational results that scientists present at conferences and in papers.

provide a long-term solution. We need both disciplined ways of working reproducibly and community support (and even pressure) to ensure that such disciplines are followed.

ICERM 2012



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

Description

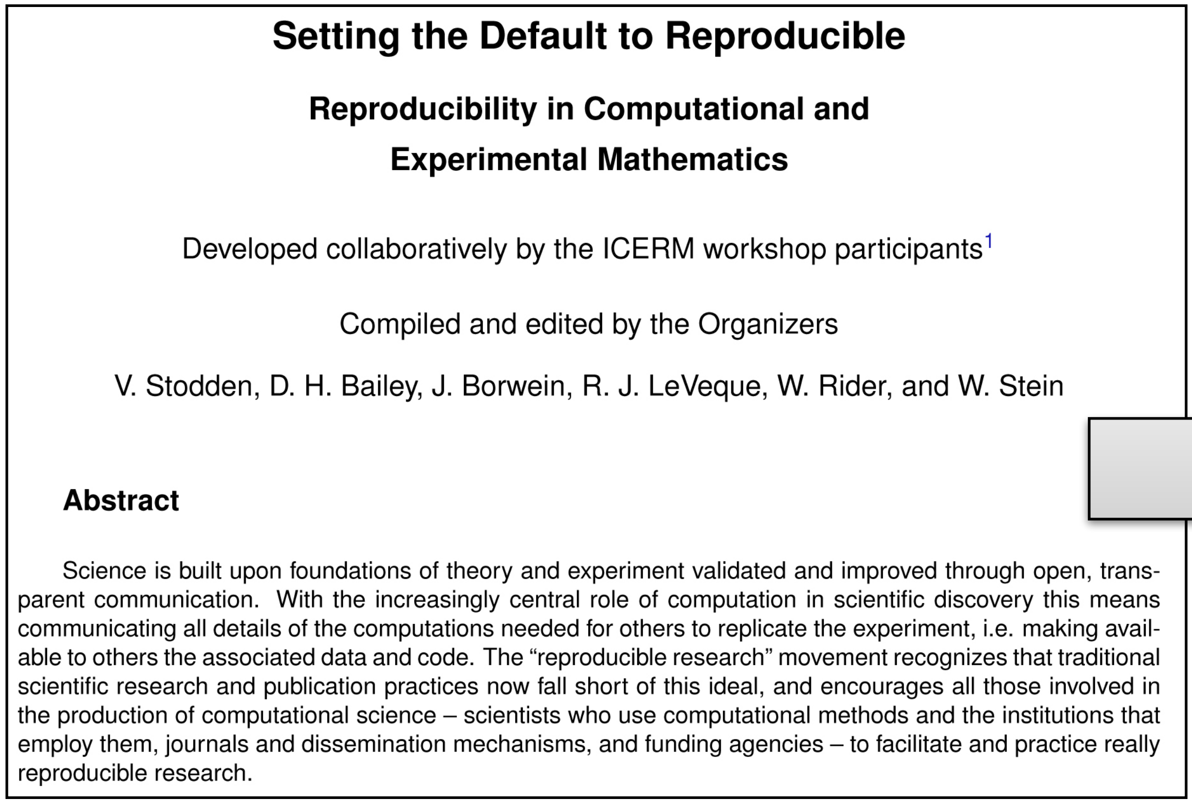
In addition to advancing research and discovery in pure and applied mathematics, computation is pervasive across the sciences and now computational research results are more crucial than ever for public policy, risk management, and national security. Reproducibility of carefully documented experiments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

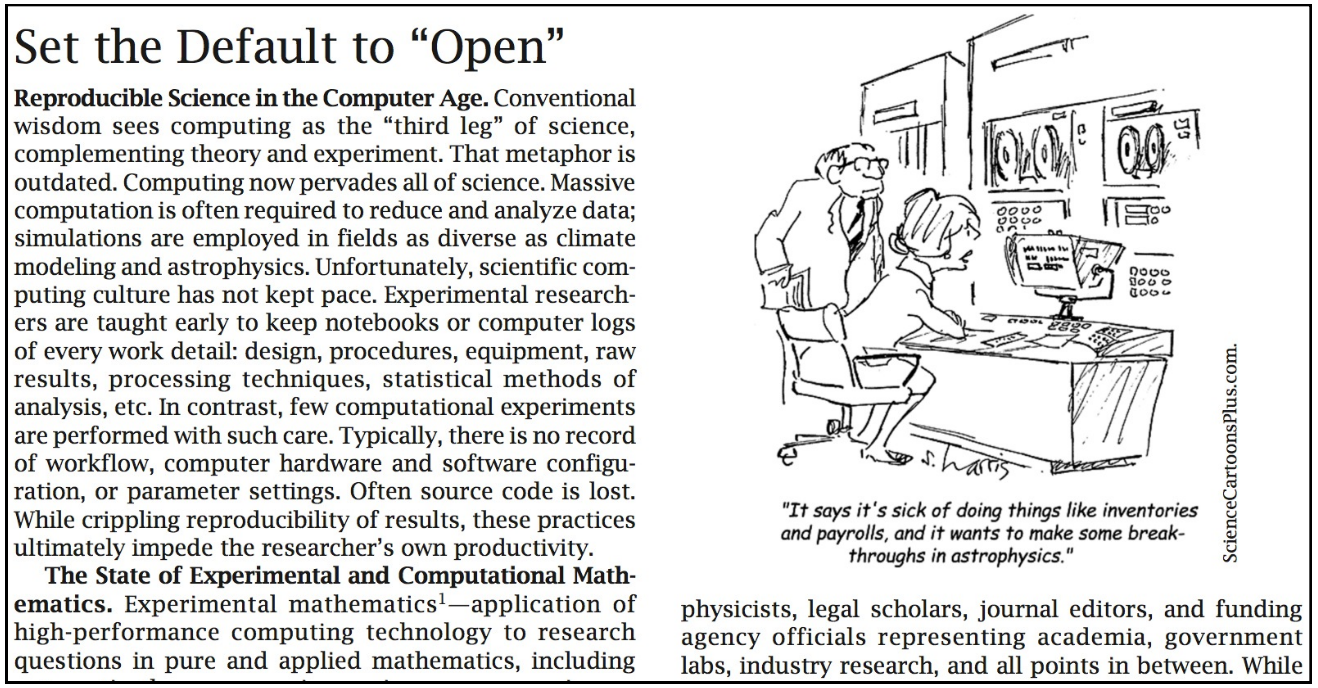
Developed collaboratively by the ICERM workshop participants[1]

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"

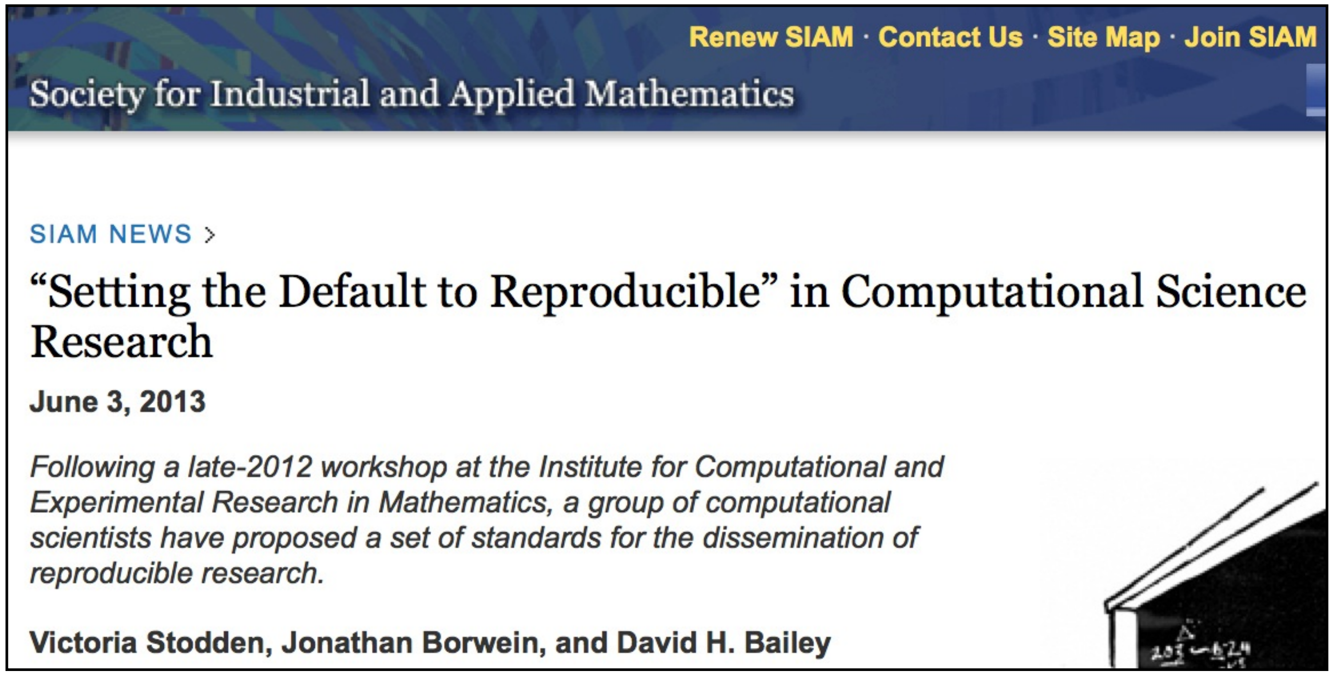
Reproducible Science in the Computer Age. Conventional wisdom sees computing as the "third leg" of science, complementing theory and experiment. That metaphor is outdated. Computing now pervades all of science. Massive computation is often required to reduce and analyze data; simulations are employed 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

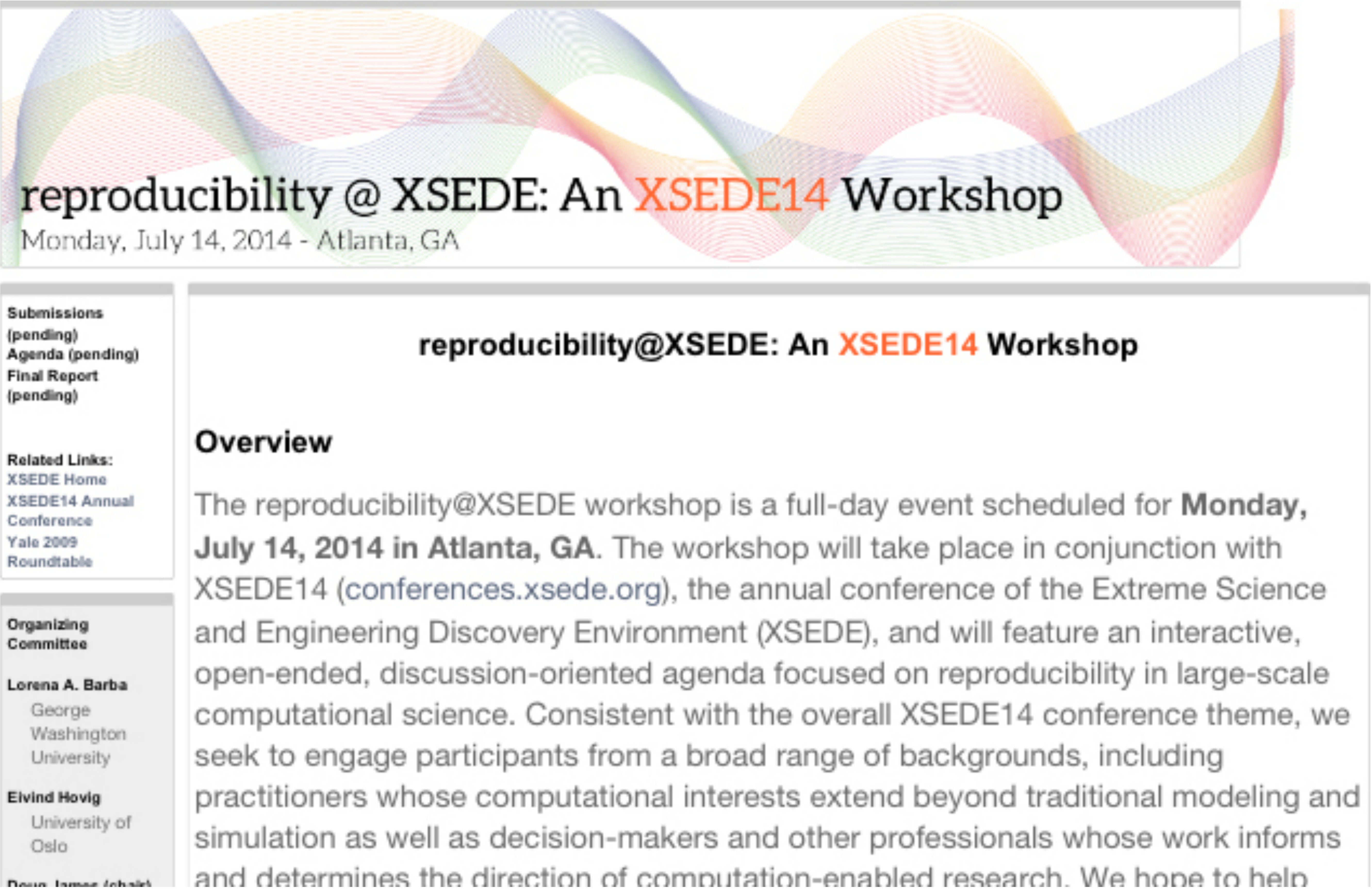
June 3, 2013

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

Issues from ICERM

- The need to carefully document the full context of computational experiments including system environment, input data, code used, computed results, etc.
- The need to save the code and data in a permanent repository, with version control and appropriate meta-data.
- The need for reviewers, research institutions, and funding agencies to recognize the importance of computing and computing professionals, and to allocate funding for after-the-grant support and repositories.
- The increasing importance of numerical reproducibility, and the need for tools to ensure and enhance numerical reliability.
- The need to encourage publication of negative results as other researchers can often learn from them.
- The re-emergence of the need to ensure responsible reporting of performance.



reproducibility @ XSEDE: An XSEDE14 Workshop

Monday, July 14, 2014 - Atlanta, GA

Submissions (pending)
Agenda (pending)
Final Report (pending)

Related Links:
XSEDE Home
XSEDE14 Annual Conference
Yale 2009 Roundtable

Organizing Committee

Lorena A. Barba
George Washington University

Eivind Hovig
University of Oslo

Doug James (chair)

reproducibility@XSEDE: An XSEDE14 Workshop

Overview

The reproducibility@XSEDE workshop is a full-day event scheduled for **Monday, July 14, 2014 in Atlanta, GA**. The workshop will take place in conjunction with XSEDE14 (conferences.xsede.org), the annual conference of the Extreme Science and Engineering Discovery Environment (XSEDE), and will feature an interactive, open-ended, discussion-oriented agenda focused on reproducibility in large-scale computational science. Consistent with the overall XSEDE14 conference theme, we seek to engage participants from a broad range of backgrounds, including practitioners whose computational interests extend beyond traditional modeling and simulation as well as decision-makers and other professionals whose work informs and determines the direction of computation-enabled research. We hope to help

Standing Together
for
Reproducibility in Large-Scale Computing

Report on reproducibility@XSEDE

An XSEDE14 Workshop
July 14, 2014
Atlanta, GA

Developed collaboratively by the reproducibility@XSEDE workshop participants[1]

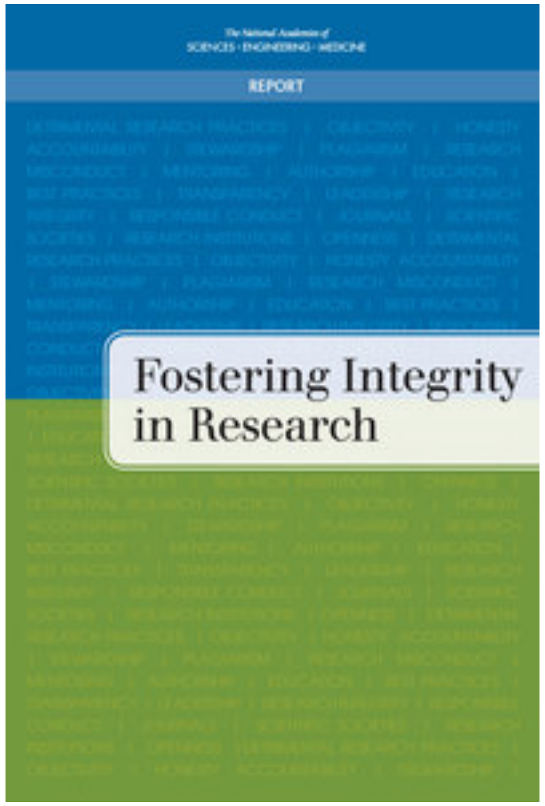
Principal Editors:
Doug James, Nancy Wilkins-Diehr, Victoria Stodden, Dirk Colbry, and Carlos Rosales

Finalized 17 Dec 2014

Abstract. *This is the final report on reproducibility@xsede, a one-day workshop held in conjunction with XSEDE14, the annual conference of the Extreme Science and Engineering Discovery Environment (XSEDE). The workshop's discussion-oriented agenda focused on reproducibility in large-scale computational research. Two important themes capture the spirit of the workshop submissions and discussions: (1) organizational stakeholders, especially supercomputer centers, are in a unique position to promote, enable, and support reproducible research; and (2) individual researchers should conduct each experiment as though someone will replicate that experiment. Participants documented numerous issues, questions, technologies, practices, and potentially promising initiatives emerging from the discussion, but also highlighted four areas of particular interest to XSEDE: (1) documentation and training that promotes reproducible research; (2) system-level tools that provide build- and run-time information at the level of the individual job; (3) the need to model best practices in research collaborations involving XSEDE staff; and (4) continued work on gateways and related technologies. In addition, an intriguing question emerged from the day's interactions: would there be value in establishing an annual award for excellence in reproducible research?*

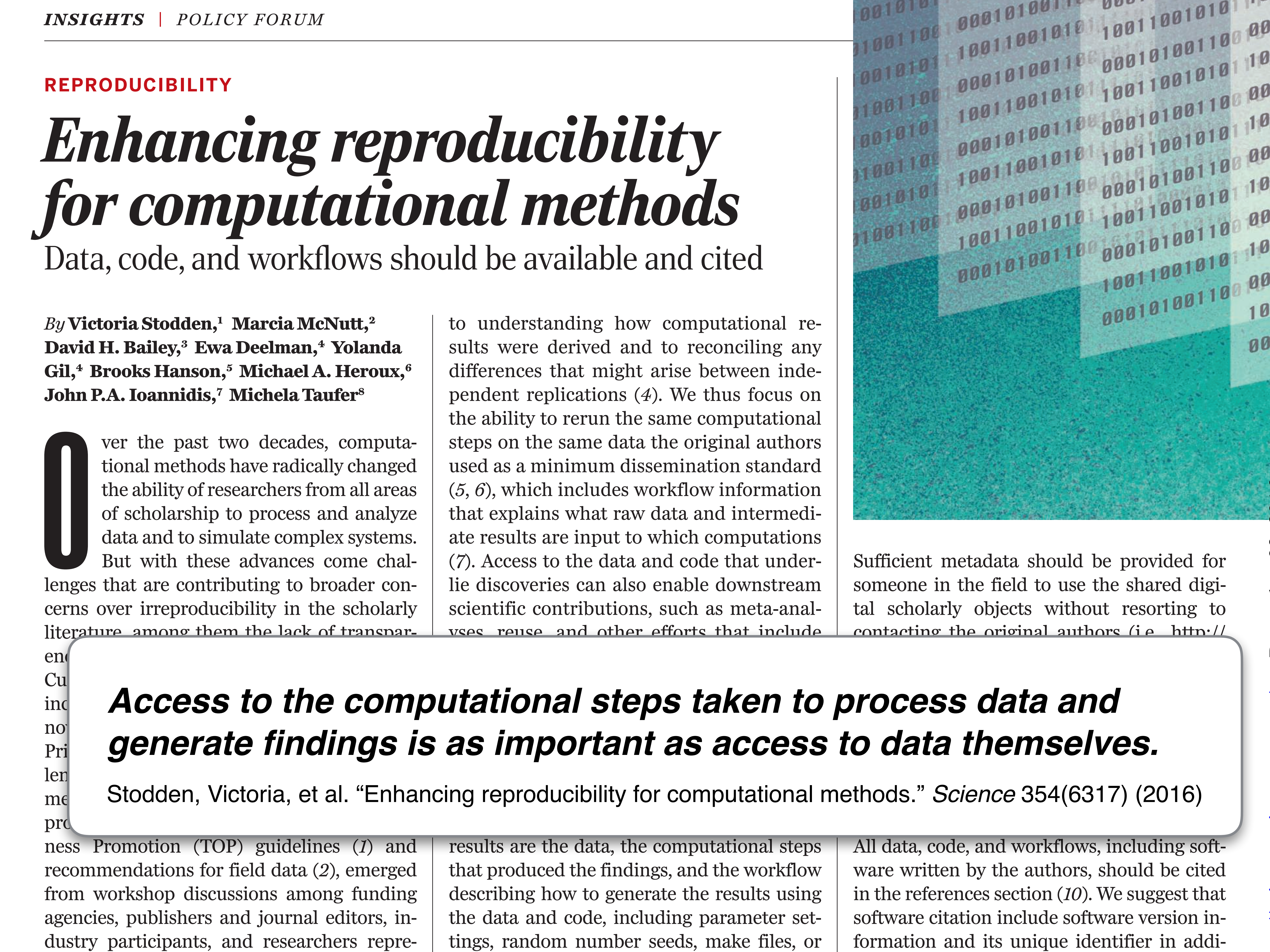
“Fostering Integrity in Research”

6: Through their policies and through the development of supporting infrastructure, research sponsors and science, engineering, technology, and medical journal and book publishers should ensure that **information sufficient** for a person knowledgeable about the field and its techniques **to reproduce reported results is made available at the time of publication** or as soon as possible after publication.



7: Federal funding agencies and other research sponsors should allocate sufficient funds to **enable the long-term storage, archiving, and access of datasets and code necessary for the replication of published findings**.

Fostering Integrity in Research, National Academies of Sciences, Engineering, and Medicine, 2017



REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By **Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[4] Brooks Hanson,[5] Michael A. Heroux,[6] John P.A. Ioannidis,[7] Michela Taufer[8]**

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transpar-
en
Cu
inc
no
Pri
len
me
pro
ness Promotion (TOP) guidelines (*1*) and recommendations for field data (*2*), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers repre-

to understanding how computational results were derived and to reconciling any differences that might arise between independent replications (*4*). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (*5*, *6*), which includes workflow information that explains what raw data and intermediate results are input to which computations (*7*). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include

results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://

All data, code, and workflows, including software written by the authors, should be cited in the references section (*10*). We suggest that software citation include software version information and its unique identifier in addi-

Access to the computational steps taken to process data and generate findings is as important as access to data themselves.

Stodden, Victoria, et al. "Enhancing reproducibility for computational methods." *Science* 354(6317) (2016)

Reproducibility Enhancement Principles

1: To facilitate reproducibility, **share the data, software, workflows**, and details of the computational environment in open repositories.

2: To enable discoverability, **persistent links** should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

3: To enable credit for shared digital scholarly objects, **citation** should be standard practice.

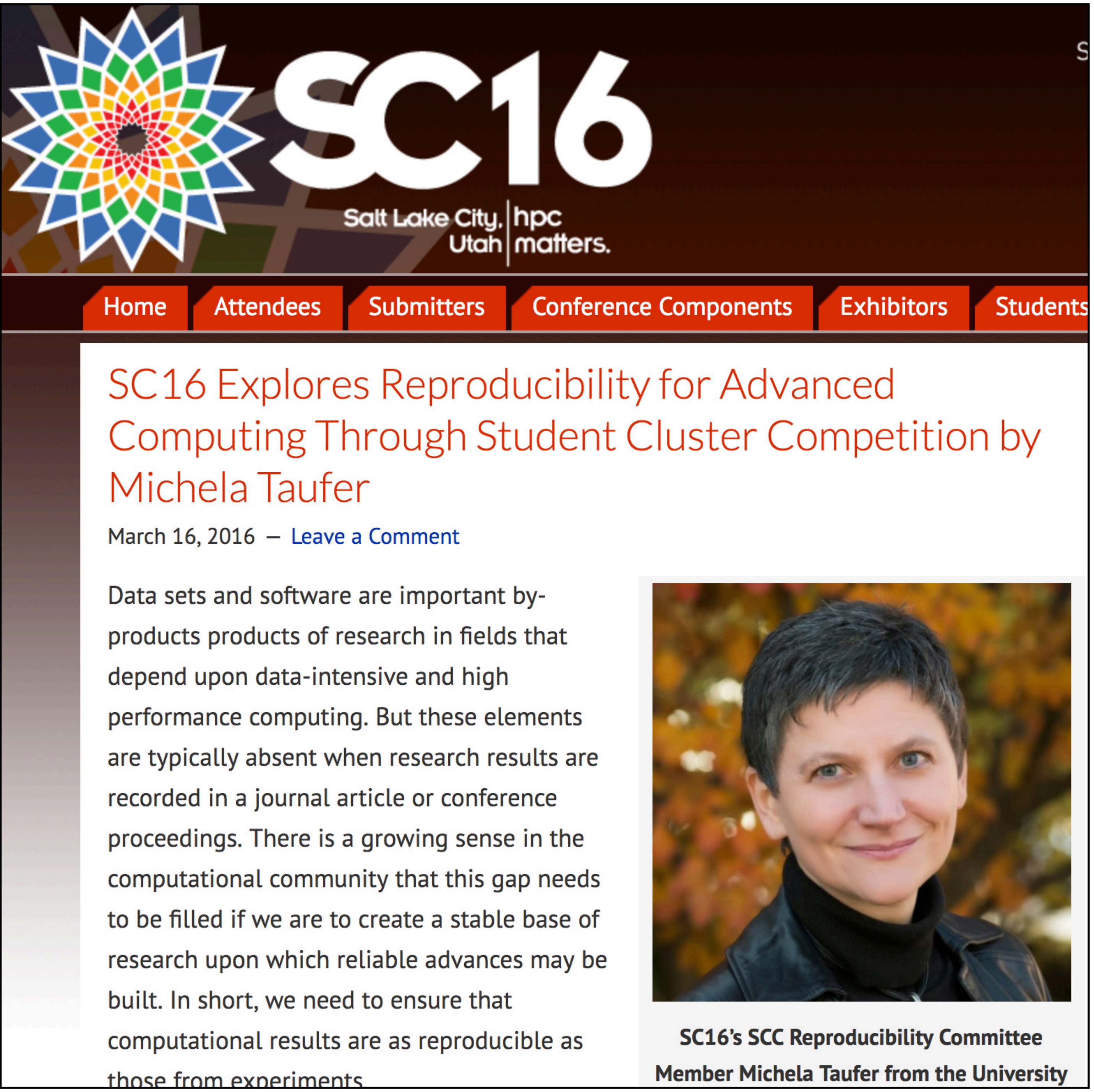
4: To facilitate reuse, adequately **document** digital scholarly artifacts.

5: Journals should conduct a **Reproducibility Check** as part of the publication process and enact the TOP Standards at level 2 or 3.

6: Use **Open Licensing** when publishing digital scholarly objects.

7: Funding agencies should instigate **new research** programs and pilot studies.

Supercomputing

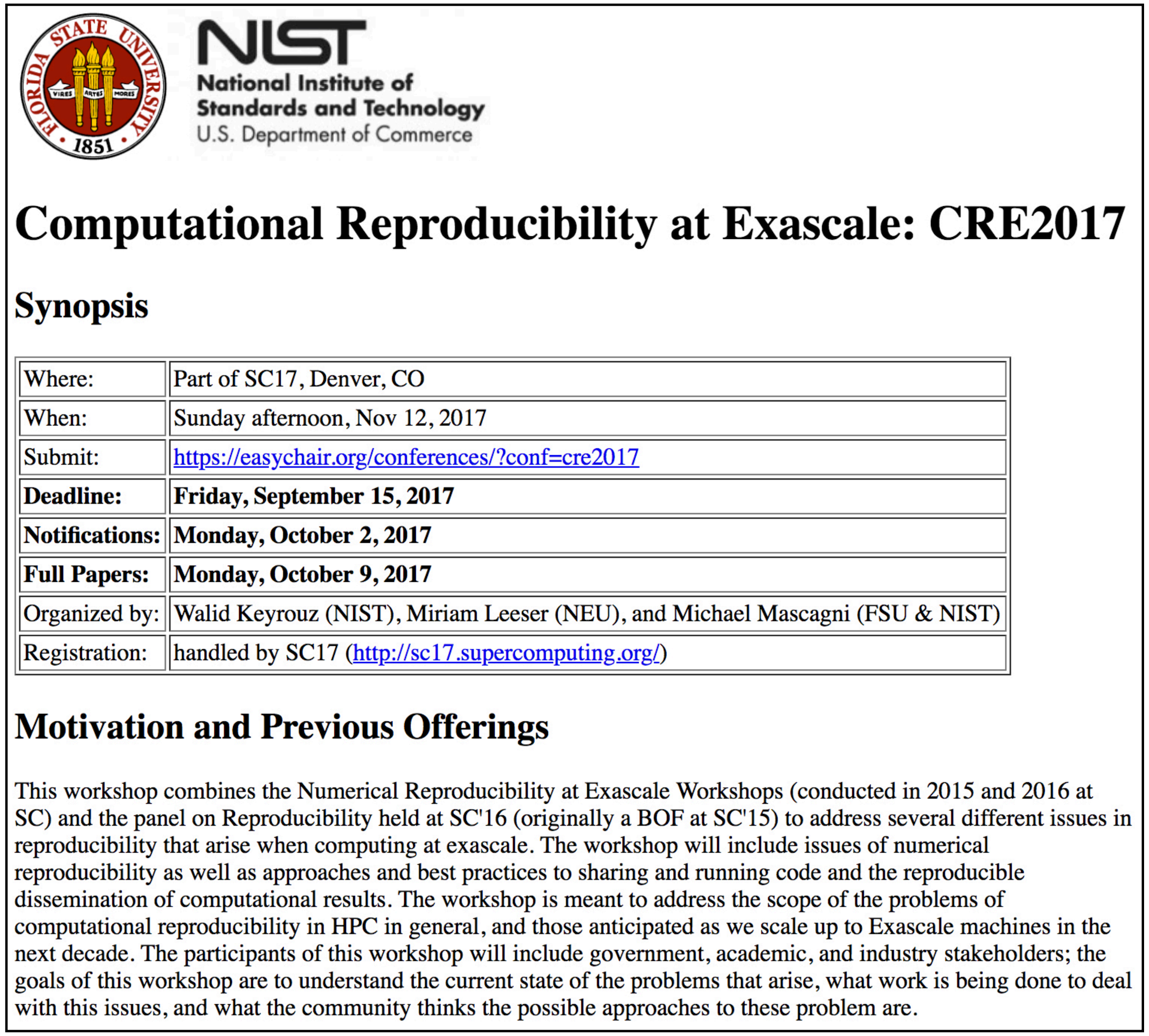


SC16 Explores Reproducibility for Advanced Computing Through Student Cluster Competition by Michela Taufer

March 16, 2016 – Leave a Comment

Data sets and software are important by-products products of research in fields that depend upon data-intensive and high performance computing. But these elements are typically absent when research results are recorded in a journal article or conference proceedings. There is a growing sense in the computational community that this gap needs to be filled if we are to create a stable base of research upon which reliable advances may be built. In short, we need to ensure that computational results are as reproducible as those from experiments

SC16's SCC Reproducibility Committee Member Michela Taufer from the University


FLORIDA STATE UNIVERSITY 1851
NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Computational Reproducibility at Exascale: CRE2017

Synopsis

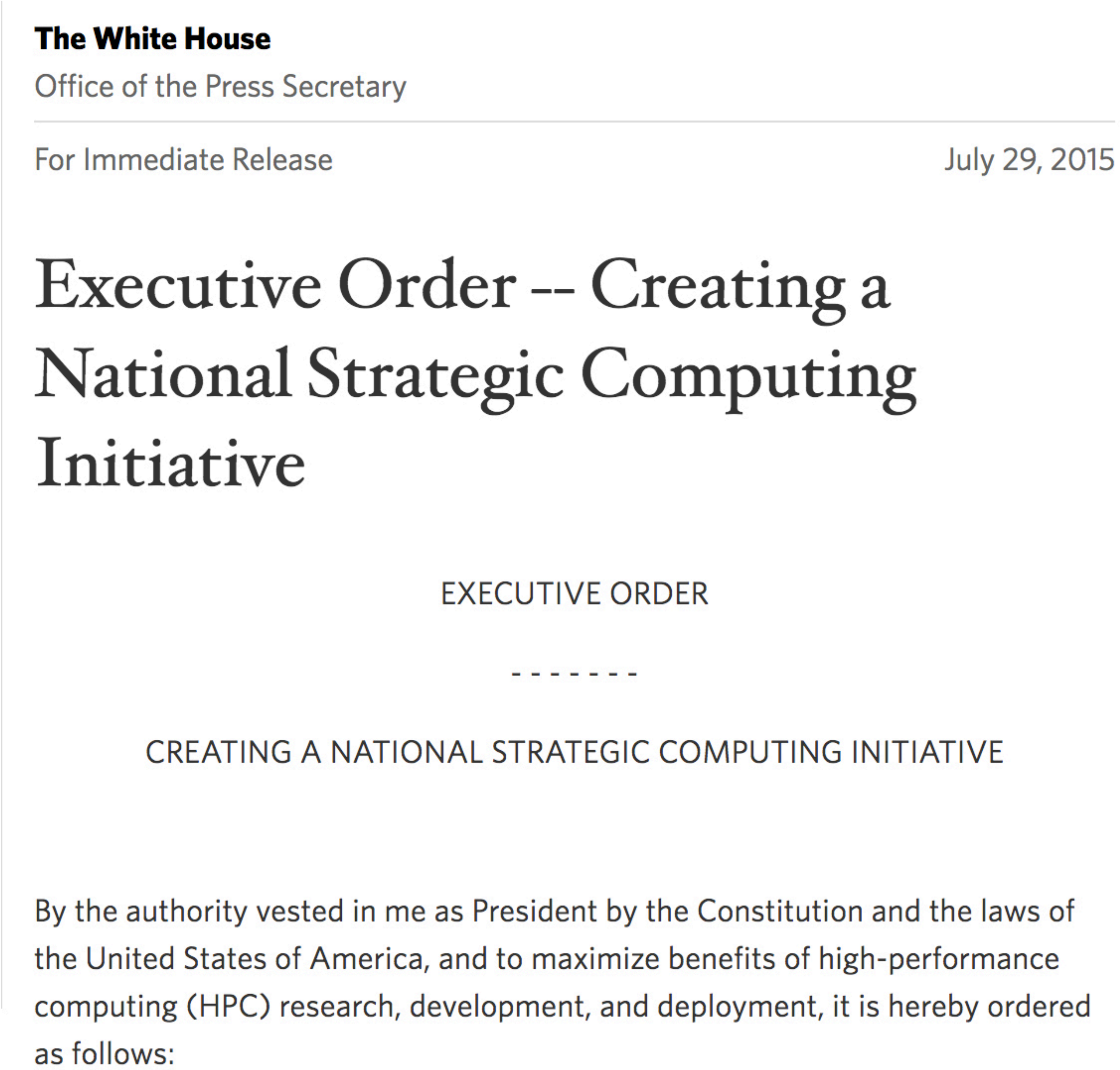
Where:	Part of SC17, Denver, CO
When:	Sunday afternoon, Nov 12, 2017
Submit:	https://easychair.org/conferences/?conf=cre2017
Deadline:	**Friday, September 15, 2017**
Notifications:	**Monday, October 2, 2017**
Full Papers:	**Monday, October 9, 2017**
Organized by:	Walid Keyrouz (NIST), Miriam Leeser (NEU), and Michael Mascagni (FSU & NIST)
Registration:	handled by SC17 (http://sc17.supercomputing.org/)

Motivation and Previous Offerings

This workshop combines the Numerical Reproducibility at Exascale Workshops (conducted in 2015 and 2016 at SC) and the panel on Reproducibility held at SC'16 (originally a BOF at SC'15) to address several different issues in reproducibility that arise when computing at exascale. The workshop will include issues of numerical reproducibility as well as approaches and best practices to sharing and running code and the reproducible dissemination of computational results. The workshop is meant to address the scope of the problems of computational reproducibility in HPC in general, and those anticipated as we scale up to Exascale machines in the next decade. The participants of this workshop will include government, academic, and industry stakeholders; the goals of this workshop are to understand the current state of the problems that arise, what work is being done to deal with this issues, and what the community thinks the possible approaches to these problem are.

Efforts by SIGHPC, SIGMOD, SIGCOMM…

National Strategic Computing Initiative 2015



The White House

Office of the Press Secretary

For Immediate Release July 29, 2015

Executive Order -- Creating a National Strategic Computing Initiative

EXECUTIVE ORDER

- - - - - - -

CREATING A NATIONAL STRATEGIC COMPUTING INITIATIVE

By the authority vested in me as President by the Constitution and the laws of the United States of America, and to maximize benefits of high-performance computing (HPC) research, development, and deployment, it is hereby ordered as follows:

NSCI Sec. 2. Objectives.

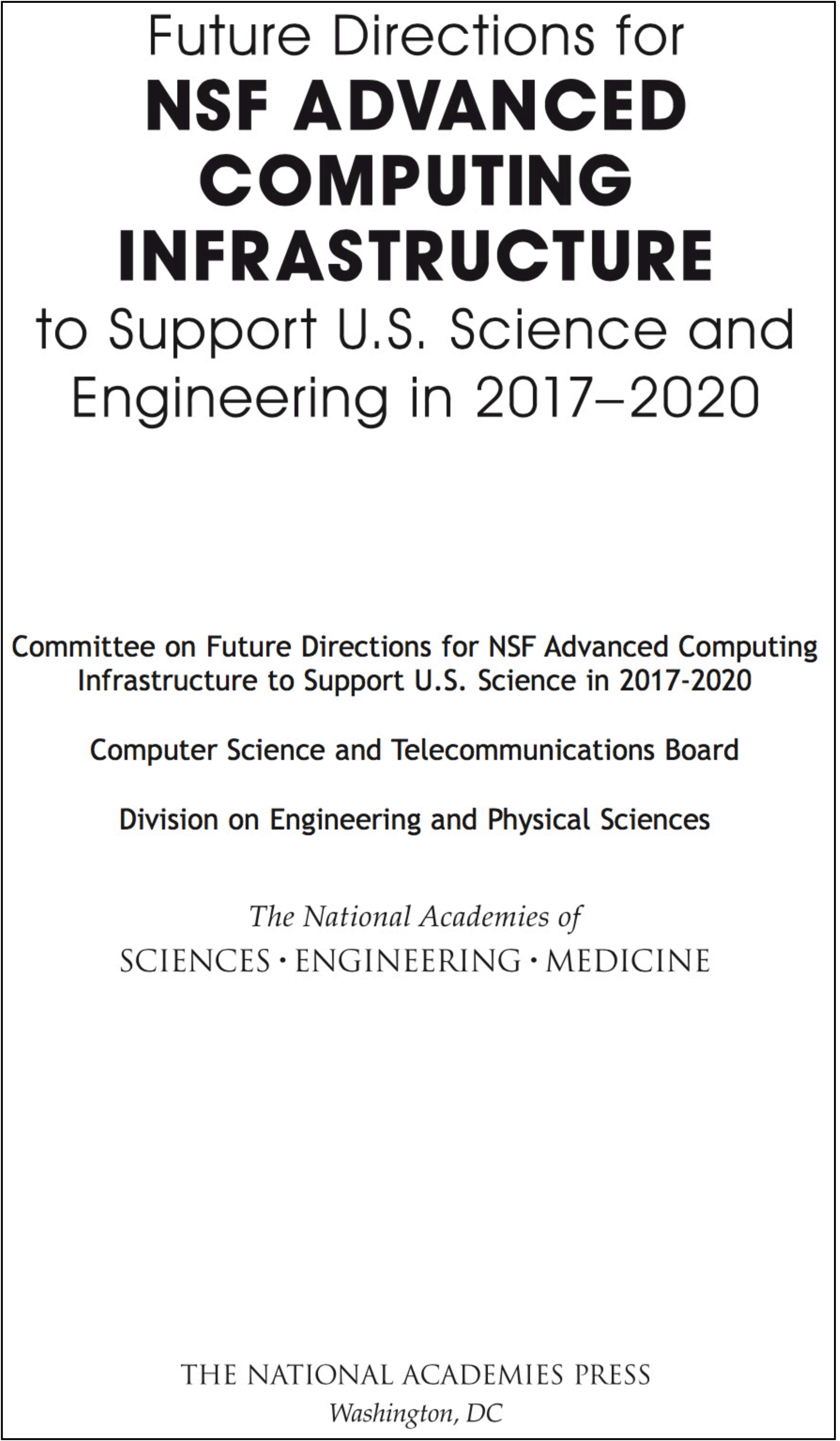
1. Accelerating delivery of a capable exascale computing system that integrates hardware and software capability to deliver approximately 100 times the performance of current 10 petaflop systems across a range of applications representing government needs.

2. Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing.

3. Establishing, over the next 15 years, a viable path forward for future HPC systems even after the limits of current semiconductor technology are reached (the "post-Moore's Law era").

4. **Increasing the capacity and capability of an enduring national HPC ecosystem by employing a holistic approach that addresses relevant factors such as networking technology, workflow, downward scaling, foundational algorithms and software, accessibility, and workforce development.**

5. Developing an enduring public-private collaboration to ensure that the benefits of the research and development advances are, to the greatest extent, shared between the United States Government and industrial and academic sectors.



Future Directions for
NSF ADVANCED COMPUTING INFRASTRUCTURE
to Support U.S. Science and Engineering in 2017–2020

Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC

- From a technical requirements perspective, infrastructure for data- intensive science needs to consider data acquisition, storage and archiving, search and retrieval, analytics, and collaboration (including publish/sub- scribe services). Recent NSF requirements to submit data management plans as part of proposals signal recognition that **access to data is increasingly important for interdisciplinary science and for research reproducibility.** Although the focus is sometimes on the hardware infrastructure (amount of storage, bandwidth, etc.), the human and software infrastructure is also important. Understanding the software frameworks that are enabled within the various cloud services and then mapping scientific workflows onto them requires a high level of both technical and scientific insight. Moreover, these new services enable a deeper level of collaboration and software reuse that are critical for data-intensive science.
- changing scientific workflows extend to the human side of scientific computing as well. Especially in regards to data-intensive science, reproducibility will be challenging. **These requirements will often be as important as the traditional technical requirements of CPU performance, latency, storage, and bandwidth.**
- deciding how much data to save is a trade-off between the cost of saving and the cost of reproducing, and this is **potentially more significant than the trade-off between disks and processors.**

Infrastructure Solutions

Research Environments and Document Enhancement Tools

- StatTag.org
- Verifiable Computational Research
- knitR
- Collage Authoring Environment
- Sumatra
- SHARE
- Sweave
- SOLE
- GenePattern
- torch.ch
- Code Ocean
- Cyverse
- Open Science Framework
- IPOL
- Whole Tale
- Jupyter
- NanoHUB
- Vistrails
- Popper
- flywheel.io

Workflow Systems

- Taverna
- Kurator
- Wings
- Kepler
- Pegasus
- Everware
- CDE
- Reprozip
- binder.org
- Galaxy

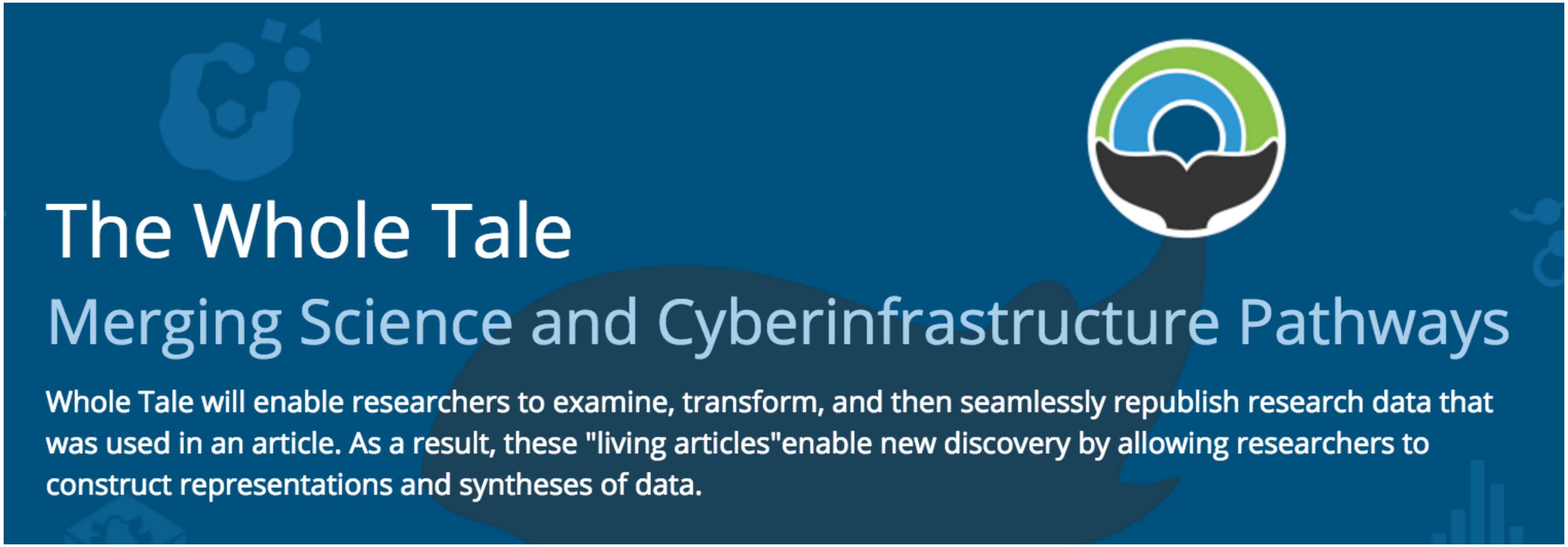
Dissemination Platforms

- ResearchCompendia.org
- Occam
- Wavelab
- DataCenterHub
- RCloud
- Sparselab
- RunMyCode.org
- TheDataHub.org
- ChameleonCloud
- Madagascar

“Whole Tale” Project

The Whole Tale project seeks to leverage & contribute to **existing cyberinfrastructure and tools** to support the **whole research story**, and provide access to data and computing power.

➡ ***Integrate** tools to **simplify usage** and promote **best practices***



The Whole Tale

Merging Science and Cyberinfrastructure Pathways

Whole Tale will enable researchers to examine, transform, and then seamlessly republish research data that was used in an article. As a result, these "living articles"enable new discovery by allowing researchers to construct representations and syntheses of data.

B. Ludaescher, K. Chard, N. Gaffney, M. B. Jones, J. Nabrzyski, V. Stodden, M. Turk
NSF CC*DNI DIBBS awarded 2016: 5 Institutions for 5 Years ($5M total)

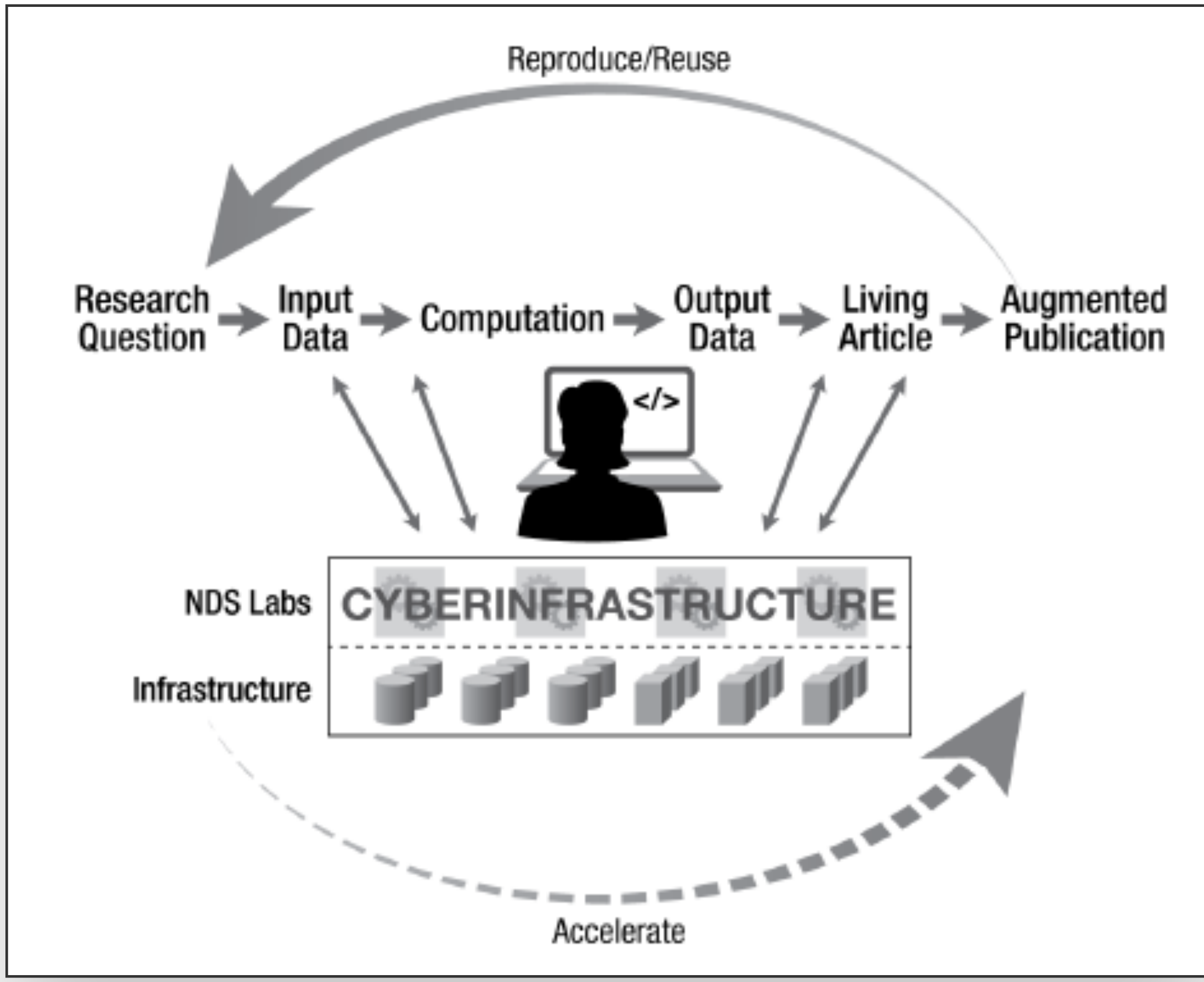
Whole Tale: What’s in a Name?

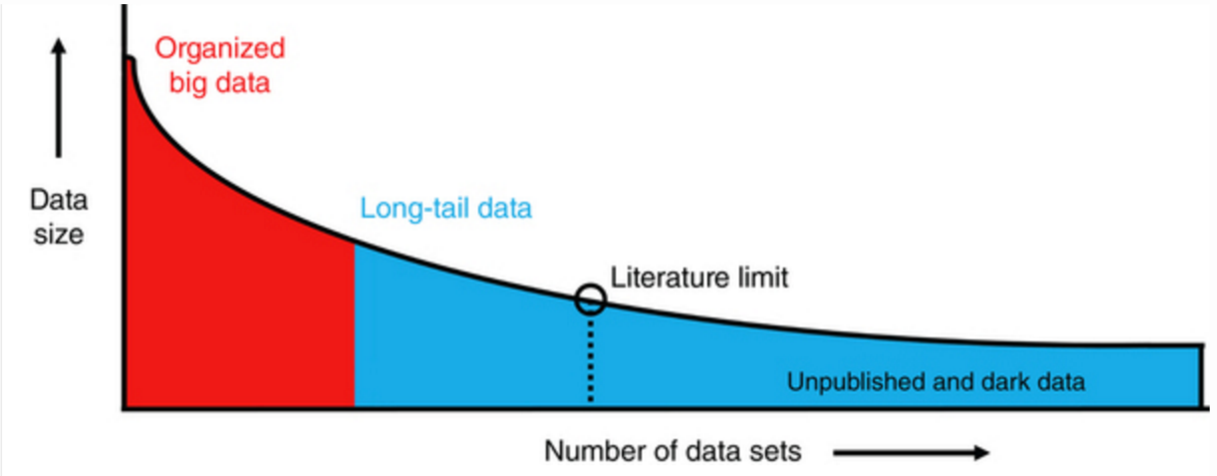
(1) Whole Tale ⇔ Whole **Story**:

Support (computational & data) **scientists** along the **complete research lifecycle** from **experiment** to **publication** and back!

(2) Whole Tale ⇔ **Long Tail of Science:**

Engage researchers of all project scales





Studies that have plotted data set size against the number of data sources reliably uncover a skewed distribution. Well-organized big science efforts featuring homogenous, well-organized data represent only a small proportion of the total data collected by scientists. A very large proportion of scientific data falls in the long-tail of the distribution, with numerous small independent research efforts yielding a rich

image from Ferguson et al. 2014 doi:10.1038/nn.3838

“Tales”

“Tales” are the final published research output from a project, capturing the complete provenance of a particular activity/analysis within the system:

- easily sharable with others,
- publishable in repositories,
- associated with persistent identifiers,
- linked to publications,
- execute in the same state as it was when first published,
- acts as a starting point for research.

Try it!

The first Whole Tale platform was released in July!

http://wholetale.readthedocs.io/users_guide/

Feedback is very welcome at feedback@wholetale.org and/or at https://github.com/whole-tale/whole-tale/issues

“ezDMP”

NSF funded project to provide structured guidance for a second generation data management plan.

EAGER: Collaborative Proposal: Supporting Public Access to Supplemental Scholarly Products Generated from Grant Funded Research (2016).

Helen M. Berman (Rutgers)
Kerstin Lehnert (Columbia)
Vicki Ferrini (Columbia)
Victoria Stodden (UIUC)
Maggie Gabanyi (Rutgers)

ezDMP Released!

Research progression:

- Examined selected data management plans to understand gaps, successes, and patterns of use in IEDA DMP Tool.
- Reviewed the patterns exhibited by DMP creators using the IEDA DMP Tool.
- Implemented into IEDA DMP Tool (“ezDMP”)

Try our prototype! http://dev.ezdmp.org and we have a feedback rubric here https://goo.gl/forms/CaEB3ddJ3iuUmpxS2

How Much of a Problem is Computational Reproducibility?

Does artifact access on demand work?

February 11, 2011:

> "***All data*** *necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science.* ***All computer codes*** *involved in the creation or analysis of data* ***must also be available to any reader of Science****. After publication,* ***all reasonable requests for data and materials must be fulfilled****....*"

- Survey of publications in Science Magazine from Feb 11, 2011 to June 29, 2012 inclusive.
- Obtained a random sample of 204 scientific articles with computational findings. Asked for the data and code!

Stodden et al., "Journal Policy for Computational Reproducibility," PNAS, March 2018

Response	**% of Total**
No response	26%
Email bounced	2%
Impossible to share	2%
Refusal to share	7%
Contact to another person	11%
Asks for reasons	11%
Unfulfilled promise to follow up	3%
Direct back to SOM	3%
Shared data and code	36%
Total	100%

24 articles provided direct access to code/data.

Replicating Computational Findings

- We deemed 56 of the 89 articles for which we had data and code potentially reproducible

- We chose a random sample of 22 from these 56 to replicate

Computational Replication Rates

We were able to obtain data and code from the authors of 89 articles in our sample of 204,

- ➡ overall **artifact recovery rate** estimate: **44%** with 95% confidence interval [0.36, 0.50]

Of the 56 potentially reproducible articles, we randomly choose 22 to attempt replication, and all but one provided enough information that we were able to reproduce their computational findings.

- ➡ overall **computational reproducibility** estimate: **26%** with 95% confidence interval [0.20, 0.32]

When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

The data files remains our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.

We do not typically share our internal data or code with people outside our collaboration.

The code we wrote is the accumulated product of years of effort by [redacted] and myself. Also, the data we processed was collected painstakingly over a long period by collaborators, and so we will need to ask permission from them too.

Normally we do not provide this kind of information to people we do not know. It might be that you want to check the data analysis, and that might be of some use to us, but only if you publish your findings while properly referring to us.

Thank you for your interest in our paper. For the [redacted] calculations I used my own code, and there is no public version of this code, which could be downloaded. Since this code is not very user-friendly and is under constant development I prefer not to share this code.

I'm sorry, but our computer code was not written with an eye toward distributing for other people to use. The codes are not documented and we don't have the time or resources to document them. If you have a particular calculation you would like done and it is not a major extension of what we are presently set up to do, we might be able to run the codes for you.

R is a free software package available at www.r-project.org/ I used R for the [redacted] models. As you probably know, [redacted] and [redacted] are quite complicated. But I don't have to tell you that given that you are a statistics student! I used Matlab for the geometry.

Our program [redacted] is available here [URL redacted] (documentation and tutorials were included)

If you go to [URL redacted], under the publications, I have a link to the gitHub repository. I don't know if I have all of the raw simulated data, but I certainly have the processed data used to make the plots. What do you need? All of the simulated data could of course be regenerated from the code.

Please find attached a .zip file called [redacted].zip that has the custom MATLAB [redacted] analysis code. If you run Masterrunfigureone.m this will generate several panels from the paper.

In the next email I will enclose the custom image analysis software. This can also be accessed from [URL redacted] where there is a manual and tutorial.

Please let me know if you have any troubles, or if there is anything else I can help with.

Converging Trends

Two (competing?) conjectures:

1. Scientific research will become massively more computational,
2. Scientific computing will become dramatically more transparent.

These trends need to be addressed simultaneously:

Better transparency will **allow people to run much more** ambitious computational experiments.

And **better** computational experiment **infrastructure** will allow **researchers** to be **more transparent**.

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

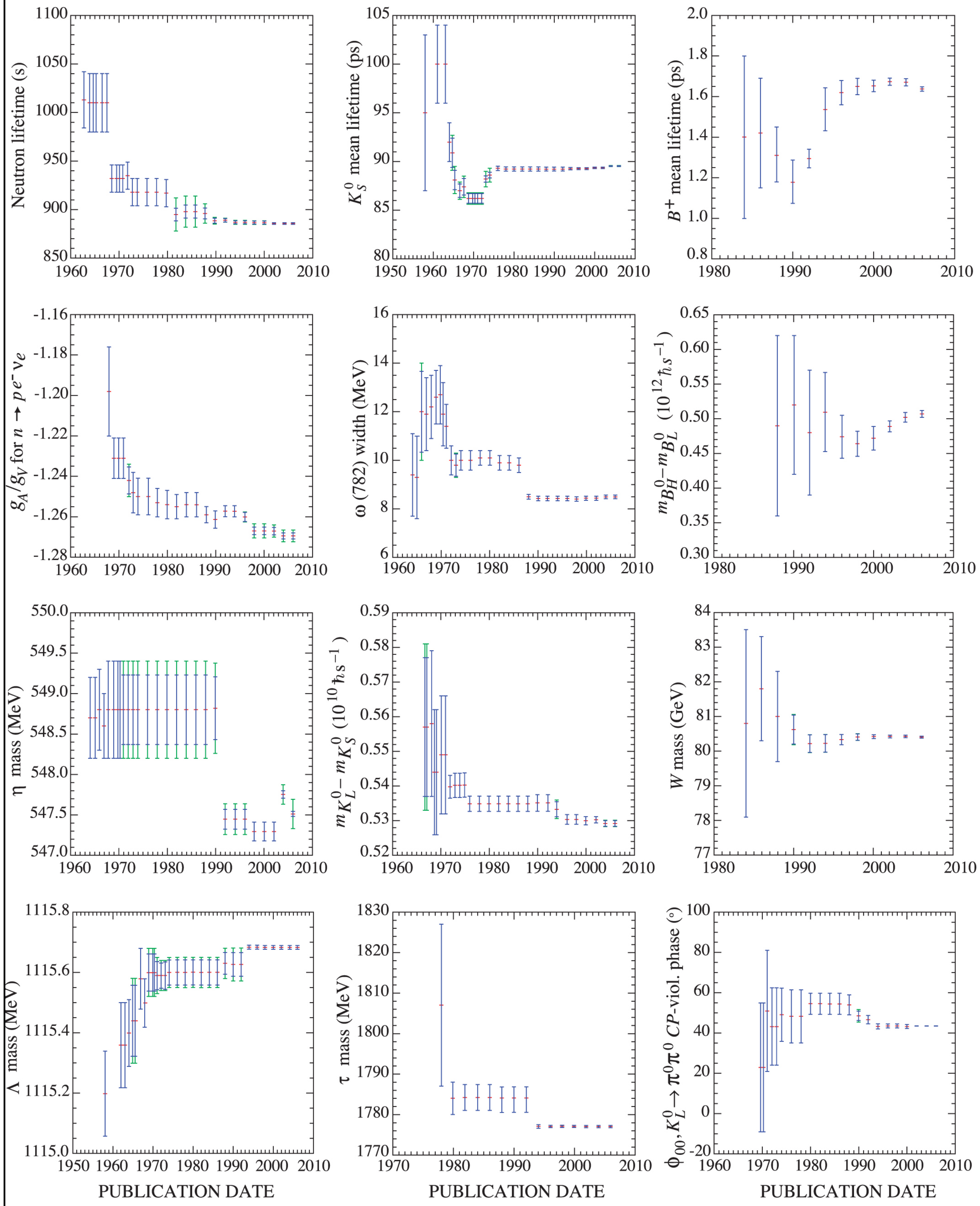


Figure 2: A historical perspective of values of a few particle properties tabulated in this *Review* as a function of date of publication of the *Review*. A full error bar indicates the quoted error; a thick-lined portion indicates the same but without the "scale factor."