Reproducibility at Exascale

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

SOS 22 Workshop
HPC and Data Science
Hilton Waikoloa Village, HI
March 28, 2018

Agenda

1. Framing Reproducibility in Computational Science

2. A (Very) Brief History of Recent Efforts

3. Steps Forward

Merton's Scientific Norms (1942)

Communalism: scientific results are the common property of the community.

Universalism: all scientists can contribute to science regardless of race, nationality, culture, or gender.

Disinterestedness: act for the benefit of a common scientific enterprise, rather than for personal gain.

Skepticism: scientific claims must be exposed to critical scrutiny before being accepted.

Skepticism and Boyle's Idea for Scientific Communication

Skepticism interpreted to mean claims can be **independently verified**, which requires **transparency** of the research process in publications.

Standards established by Transactions of the Royal Society in the 1660's (Robert Boyle).

ROBERT BOYLE.

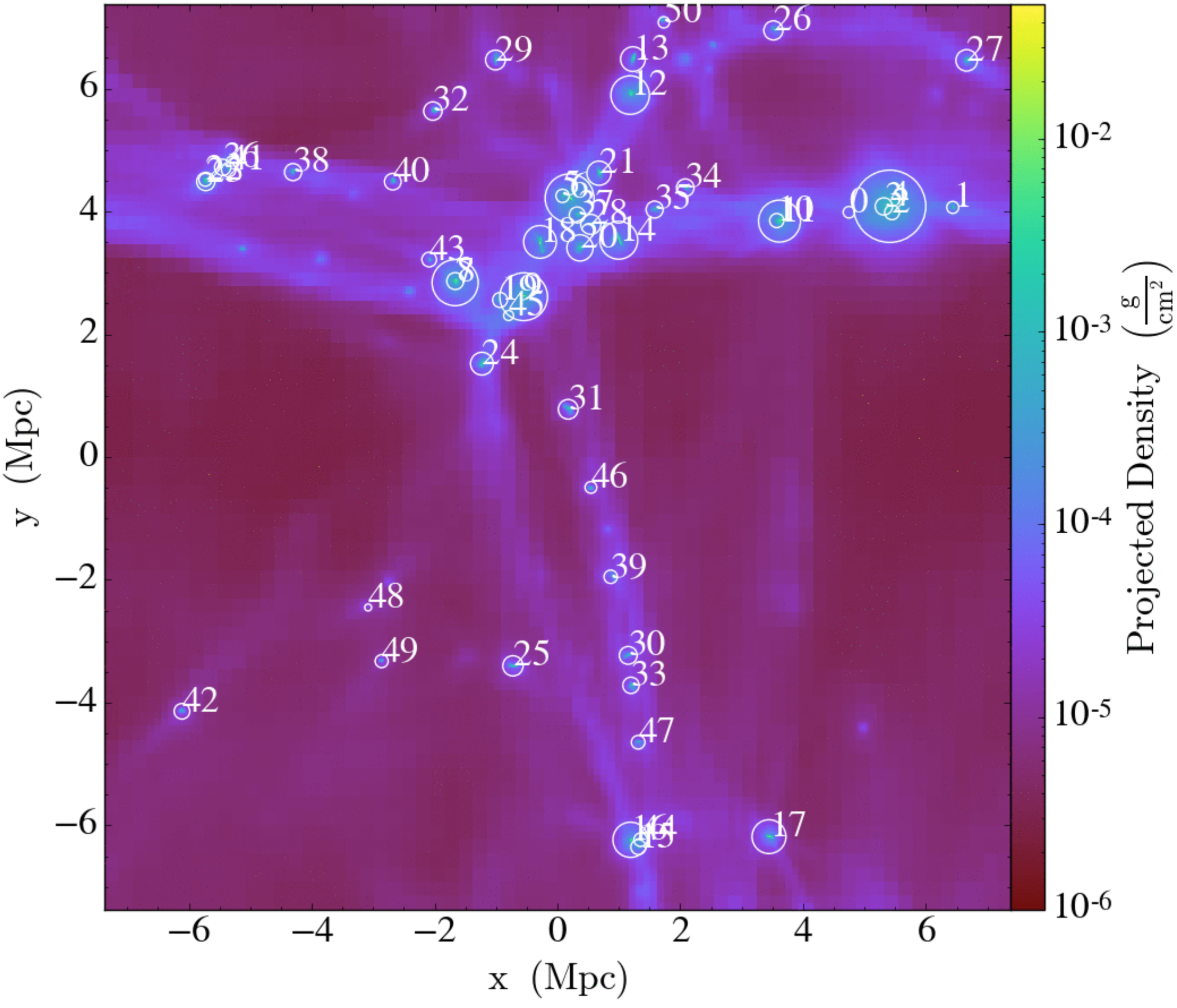
Today: Technology drives a re-assessment of transparency

- Big Data / Data Driven Discovery: e.g. high dimensional data,
- Computational Power: simulation of the complete evolution of a physical system, systematically varying parameters,
- Deep intellectual contributions now encoded only in software.

The software contains "ideas that enable biology…"

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

"Stories from the Supplement" from the Genome Informatics meeting 11/1/2013
https://youtu.be/5NiFibnbE8o


y (Mpc)
x (Mpc)
Projected Density $\left(\frac{g}{cm^2}\right)$
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

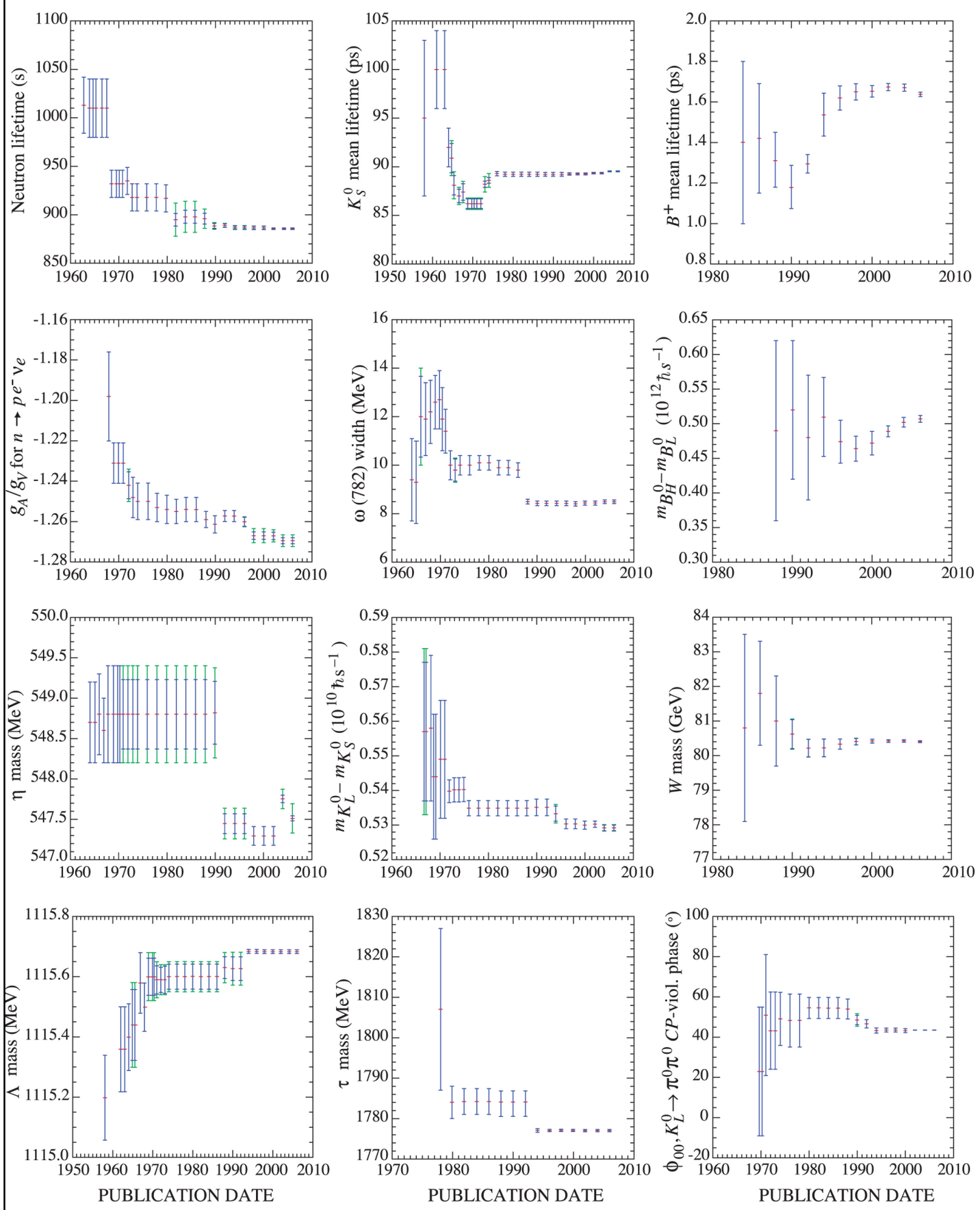
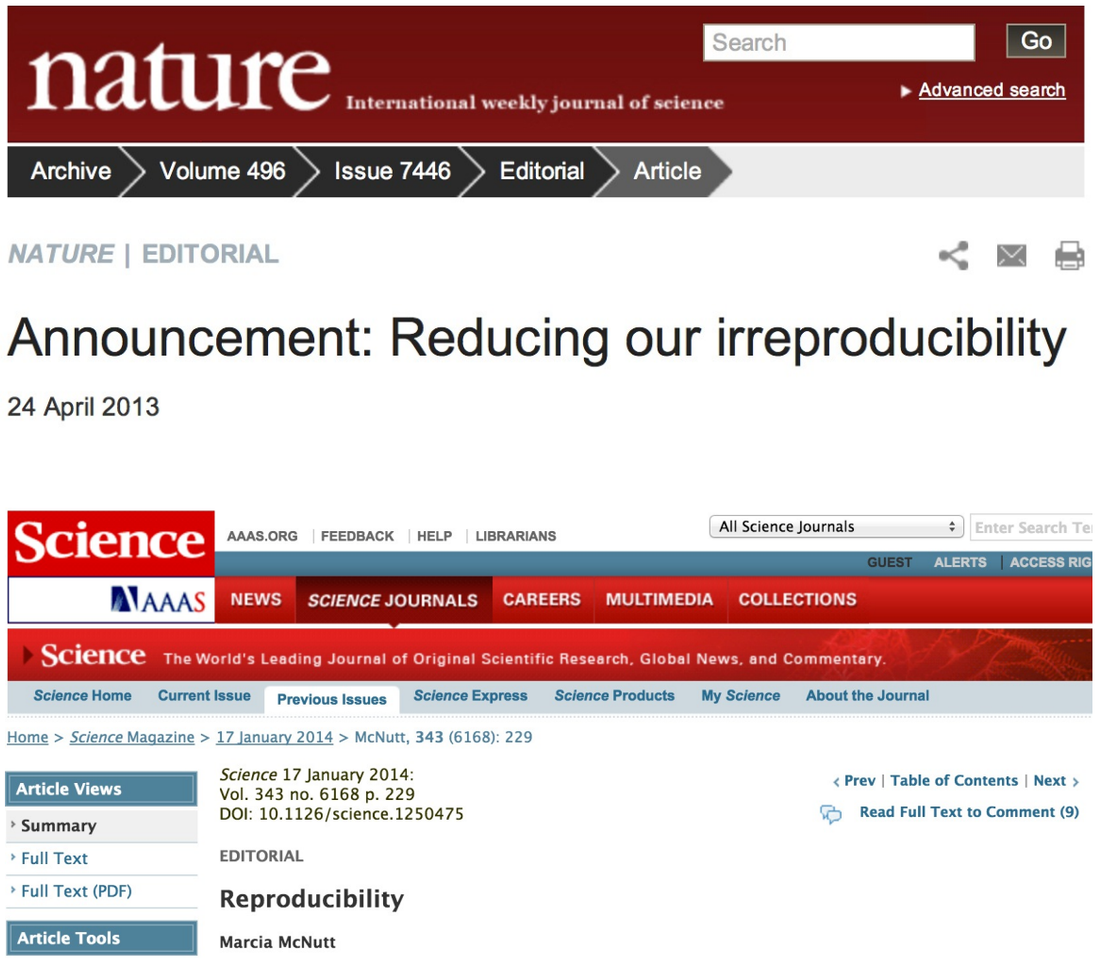


Figure 2: A historical perspective of values of a few particle properties tabulated in this *Review* as a function of date of publication of the *Review*. A full error bar indicates the quoted error; a thick-lined portion indicates the same but without the "scale factor."

Parsing Reproducibility

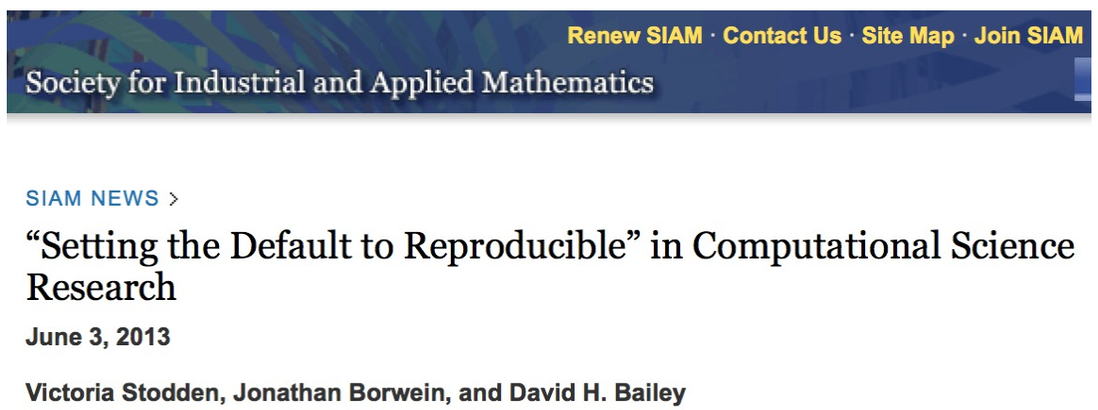
“Empirical Reproducibility”



“Statistical Reproducibility”

“Computational Reproducibility”

V. Stodden, IMS Bulletin (2013)



A Credibility Crisis

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)

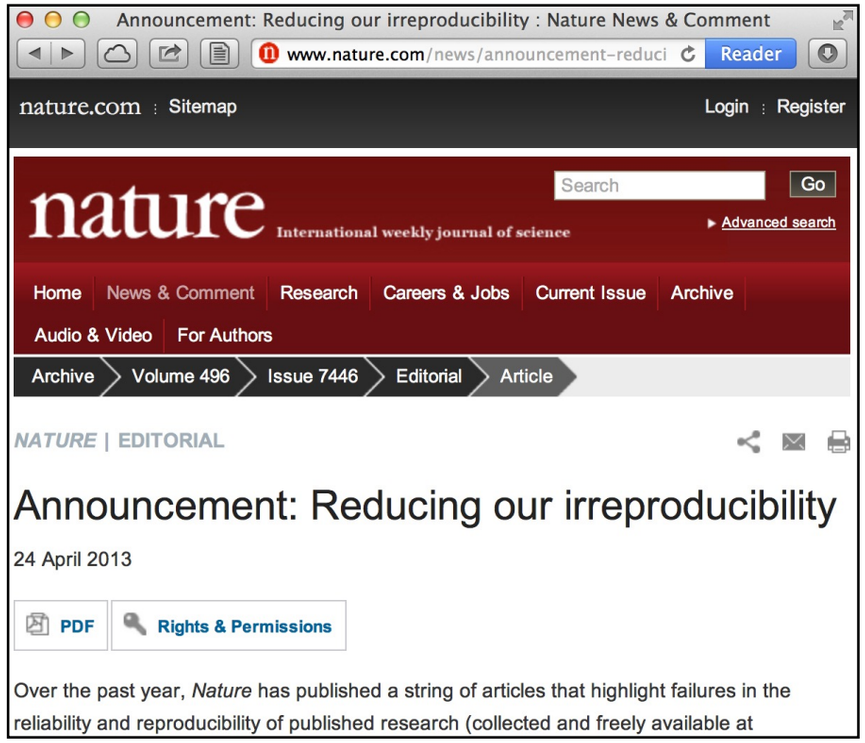
EDITORIAL

Reproducibility

Marcia McNutt

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

Science advances on a foundation of trusted discoveries. Reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author 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

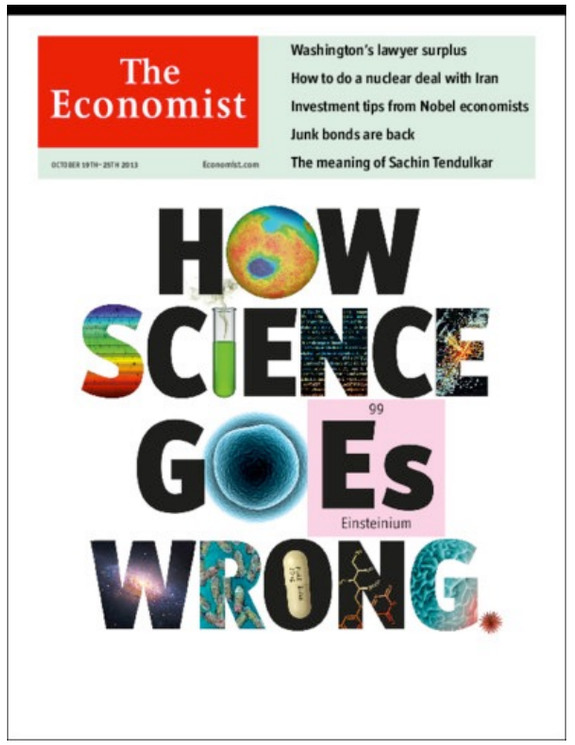
NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

PDF | Rights & Permissions

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



The Economist
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

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

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.

Email | Share 9K | Tweet 1,076 | Like 7.5k | LinkedIn 85 | +1 299

By Michael Hiltzik
October 27, 2013

In today's world, brimful as it is with opinion and falsehoods masquerading as facts, you'd think the one place you can depend on for verifiable facts is science.

You'd be wrong. Many billions of dollars' worth of wrong.

NIH Tackles Irreproducibility

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

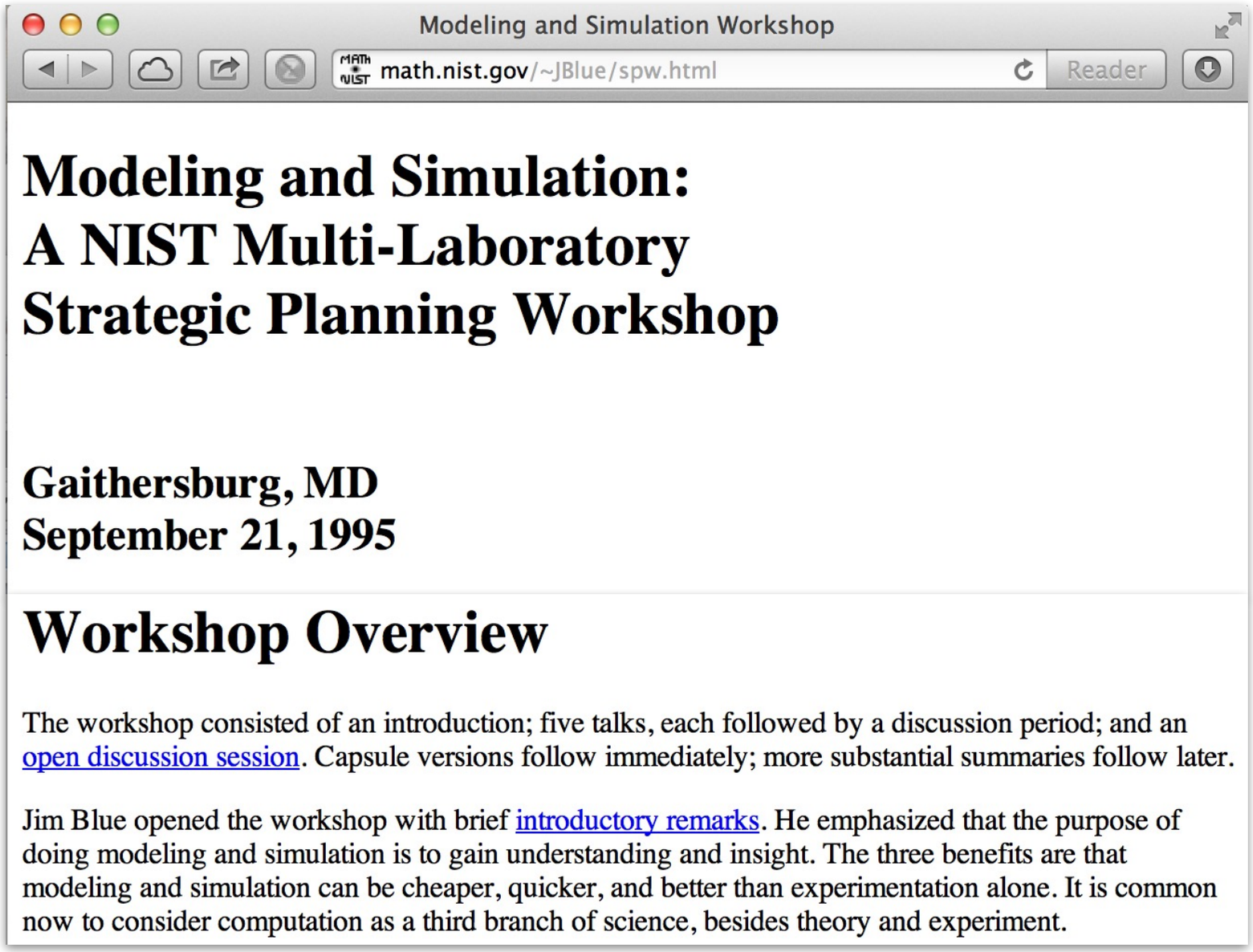
By Jef Akst | January 28, 2014

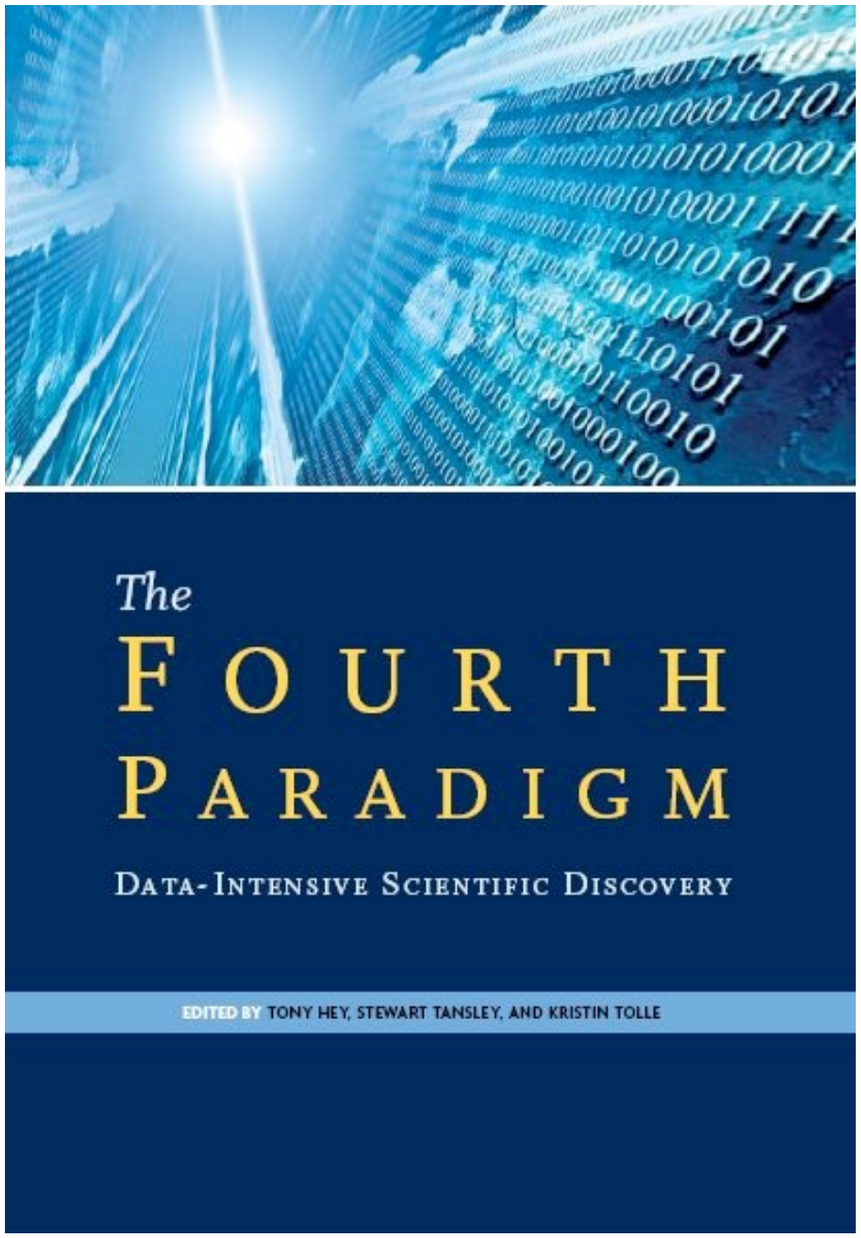
The Ubiquity of Error

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.





"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."

Really Reproducible Research

“Really Reproducible Research” (1992) inspired by Stanford Professor Jon Claerbout:

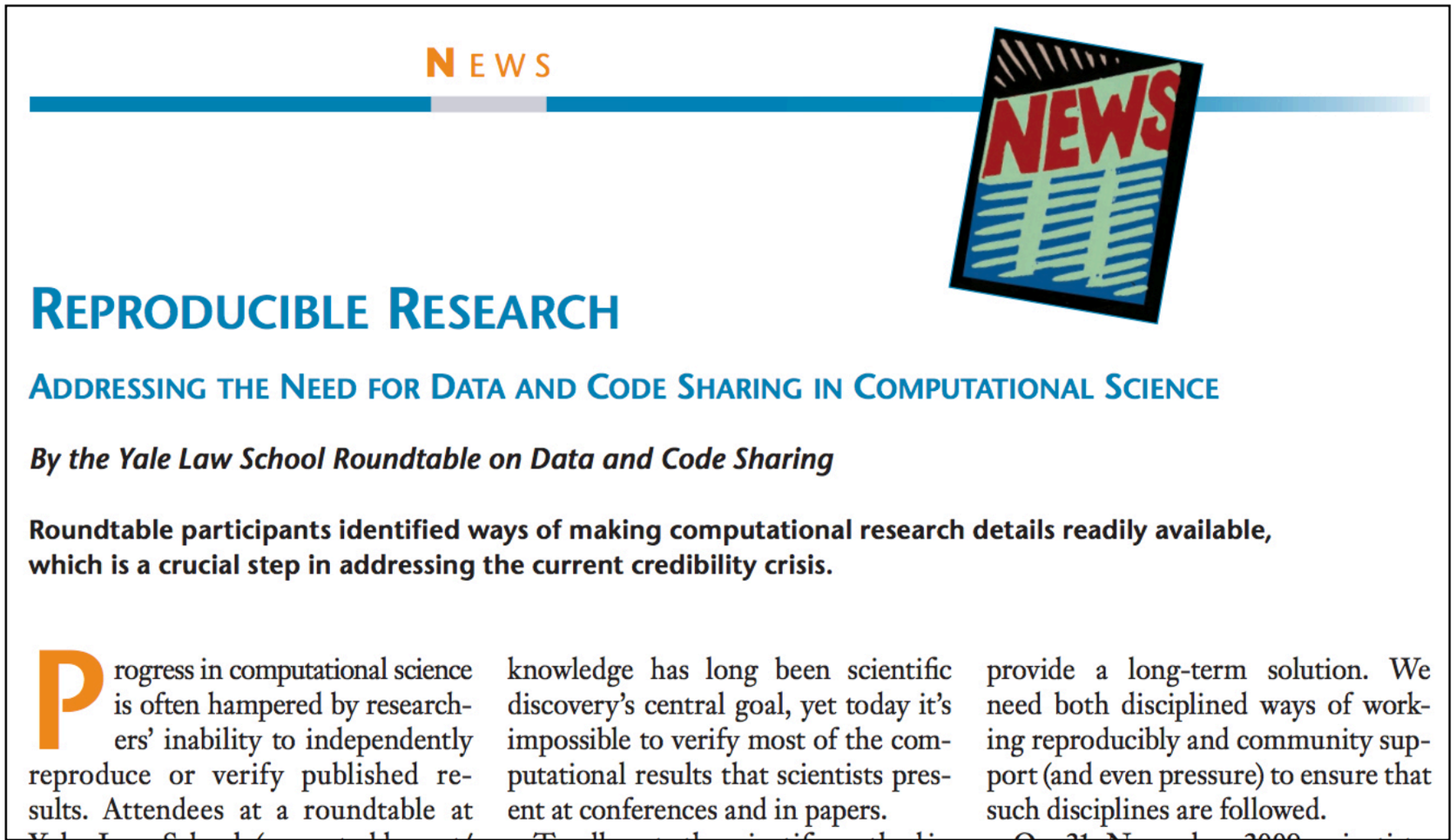
“The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.” David Donoho, 1998

A (Very) Brief History..

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

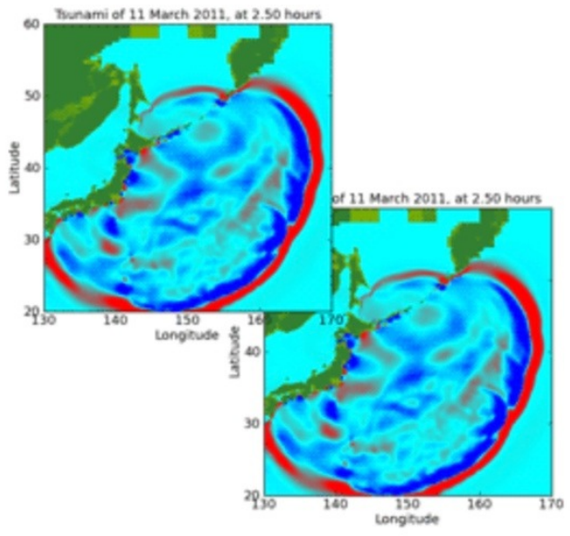
knowledge has long been scientific discovery’s central goal, yet today it’s impossible to verify most of the computational results that scientists present at conferences and in papers.

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



Click for code to create this image.

In addition to advancing research and discovery in pure and applied mathematics, computation is pervasive across the sciences and now computational research results are more crucial than ever for public policy, risk management, and national security. Reproducibility of carefully documented experiments is a cornerstone of the scientific method, and yet is often lacking in computational mathematics, science, and engineering. Setting and achieving appropriate standards for reproducibility in computation poses a number of interesting technological and social challenges. The purpose of this workshop is to discuss aspects of reproducibility most relevant to the mathematical sciences among researchers from pure and applied mathematics from academics and other settings, together with interested parties from funding agencies, national laboratories, professional societies, and publishers. This will be a working workshop, with relatively few talks and dedicated time for breakout group discussions on the current state of the art and the tools, policies, and infrastructure that are needed to improve the situation. The groups will be charged with developing guides to current best practices and/or white papers on desirable advances.

Organizing Committee

- David H. Bailey (Lawrence Berkeley National Laboratory)
- Jon Borwein (Centre for Computer Assisted Research Mathematics and its Applications)
- Randall J. LeVeque (University of Washington)
- Bill Rider (Sandia National Laboratory)
- William Stein (University of Washington)
- Victoria Stodden (Columbia University)

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

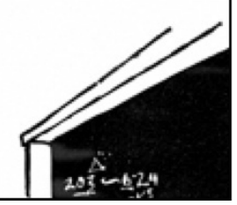
SIAM NEWS >

"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

Elvind 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?*

Supercomputing

SC16 Salt Lake City, Utah | hpc matters.

Home | Attendees | Submitters | Conference Components | Exhibitors | Students

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

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/)

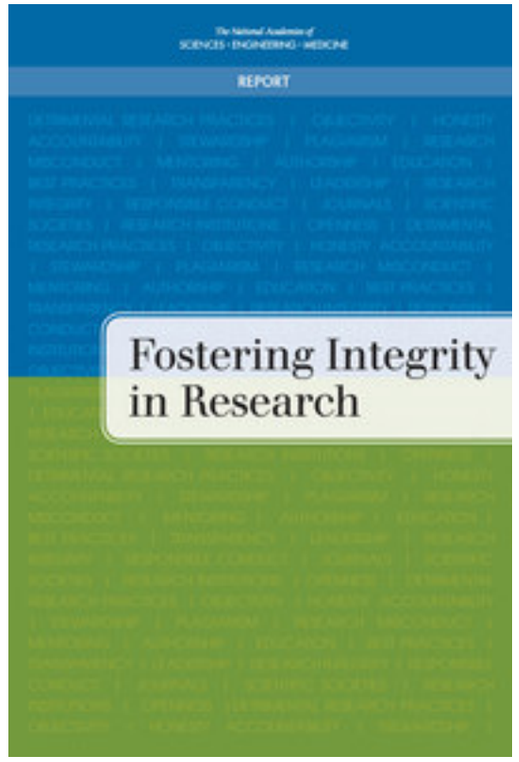
Motivation and Previous Offerings

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…

“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-

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://

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)

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

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-

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.

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.

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

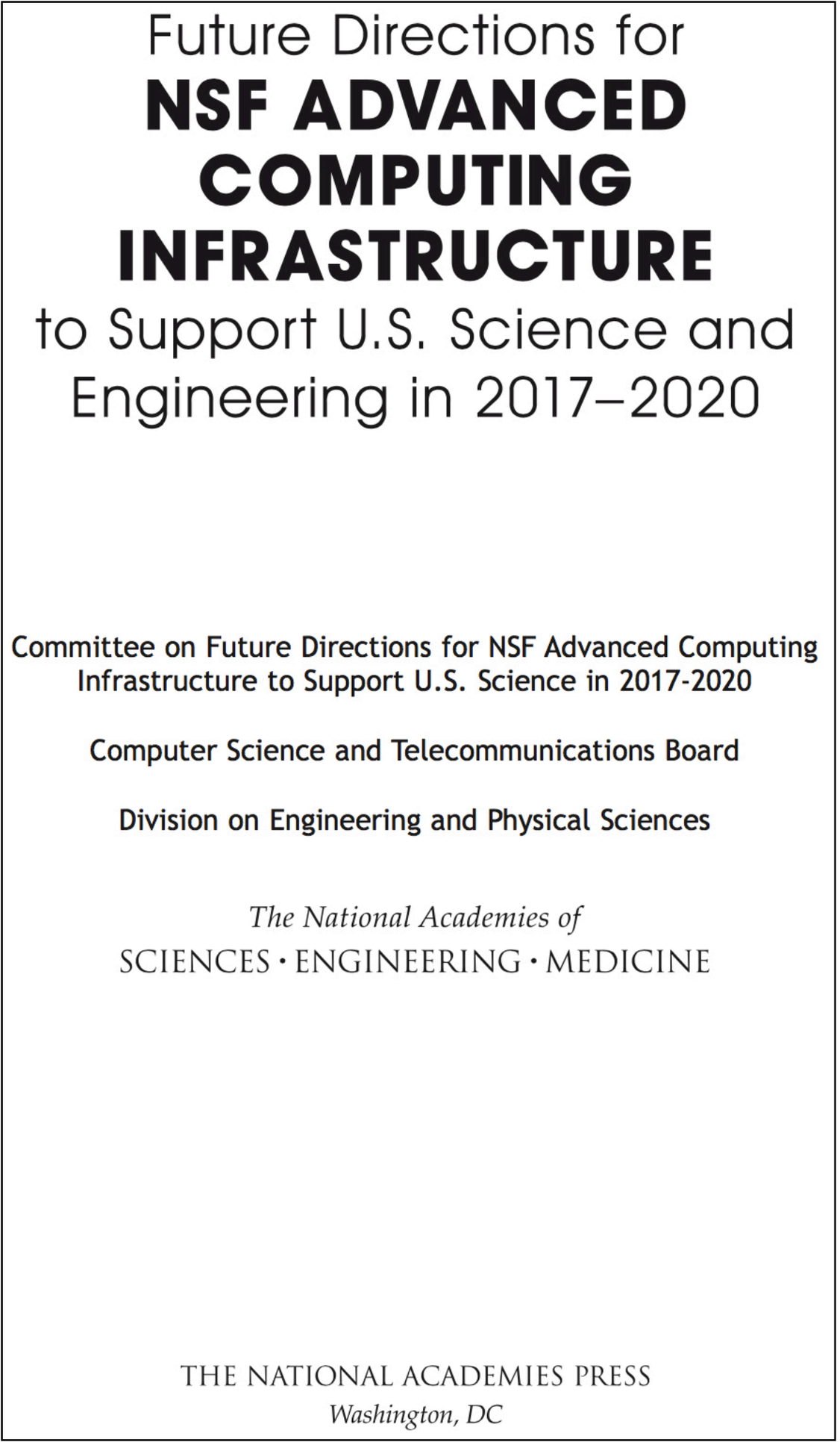
- - - - - - -

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.



- 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.**

Community Infrastructure Innovations

Research Environments

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

Workflow Systems

Taverna	Wings	Pegasus	CDE	binder.org
Kurator	Kepler	Everware	Reprozip	

Dissemination Platforms

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

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**.

Looking ahead…

We imagine a major effort to develop infrastructure that promotes good scientific practice downstream like transparency and reproducibility.

But plan for people to use it not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work.

This infrastructure is used because it enables **efficiency** and **productivity**, and **discovery**.

Life Cycle of Data Science

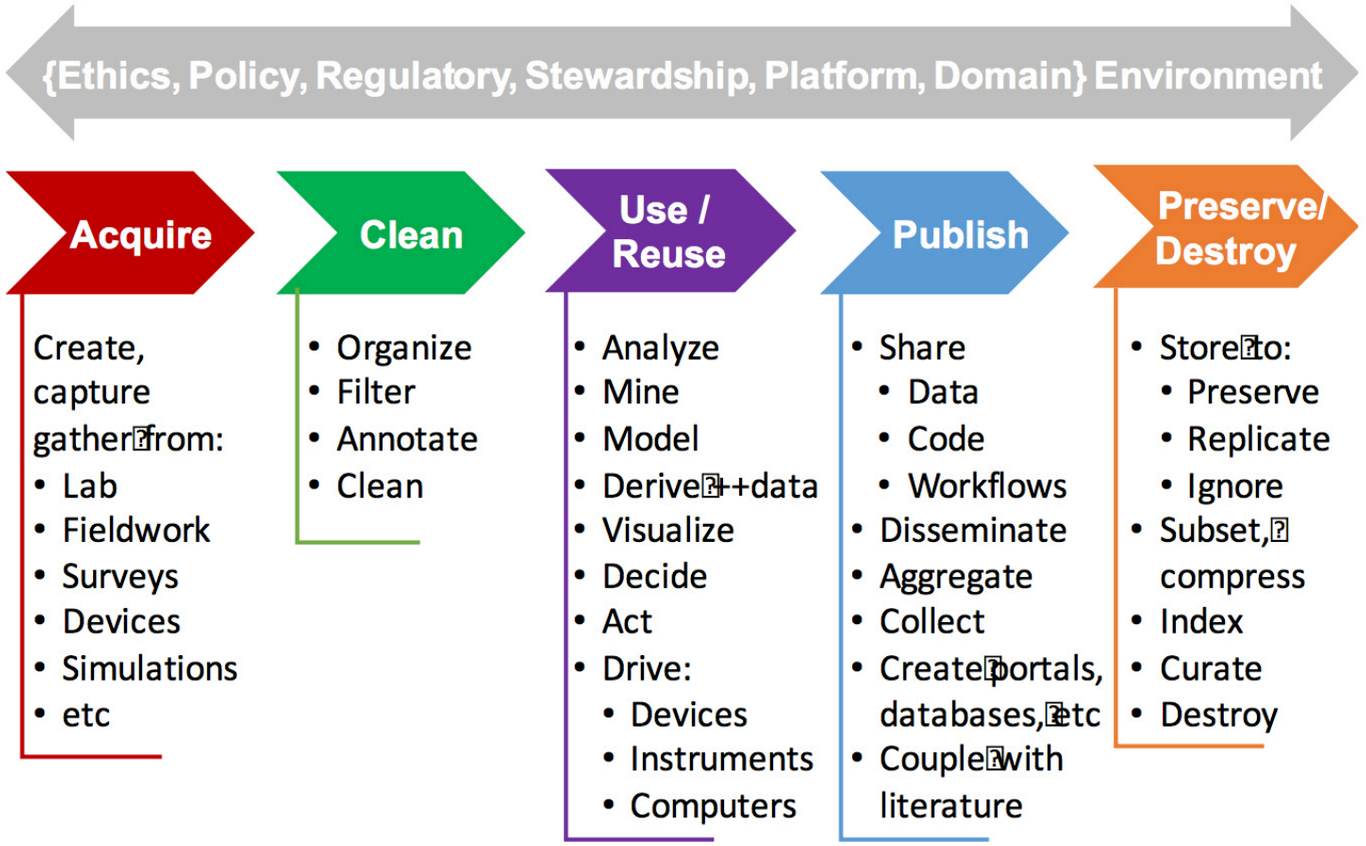


FIGURE 1: The Data Life Cycle and Surrounding Data Ecosystem

The *Data Life Cycle* is critical to understanding the opportunities and challenges of making the most of digital data. **Figure 1** shows a simplified cartoon with essential components of the data life cycle. Data is *acquired* from some source (measured, observed, generated), *cleaned* and edited to remove the outliers inevitable in real-world measurement scenarios and render it suitable for subsequent analysis; *used* (or reused) via some analysis leading to insight, action, or decision; *published* or disseminated in some way so the community at large is made aware of the data and its outcome(s); *preserved* (or not) so that others can revisit and reuse this data now or in the future. Surrounding this overall pipeline is a broader *environment* of concerns: *stewardship* to maximize the quality of the data and promote effective use, *ethics* issues that touch on proper or improper actions with these data; *policy* and *regulatory* constraints that impose legal limitations on these data; *platform* and infrastructure issues that affect technically how we can work with data; and *domain* and disciplinary needs specific to the application communities that create, operate, and use the data from these pipelines.