Reproducibility Across Research Disciplines and Stakeholder Communities

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

American Association for the Advancement of Science Annual Meeting
The Reproducibility Revolution: Impacts on Science, Journalism, and Society
Seattle, WA
February 15, 2020

Agenda

1. Three Types of Reproducibility Discussions (in Parallel)
2. AAAS 2016 Workshop Report Recommendations
3. National Academies of Science, Engineering, and Medicine 2019 Consensus Report Recommendations

This material is based upon work supported by the National Science Foundation under Grant No. 1941443. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

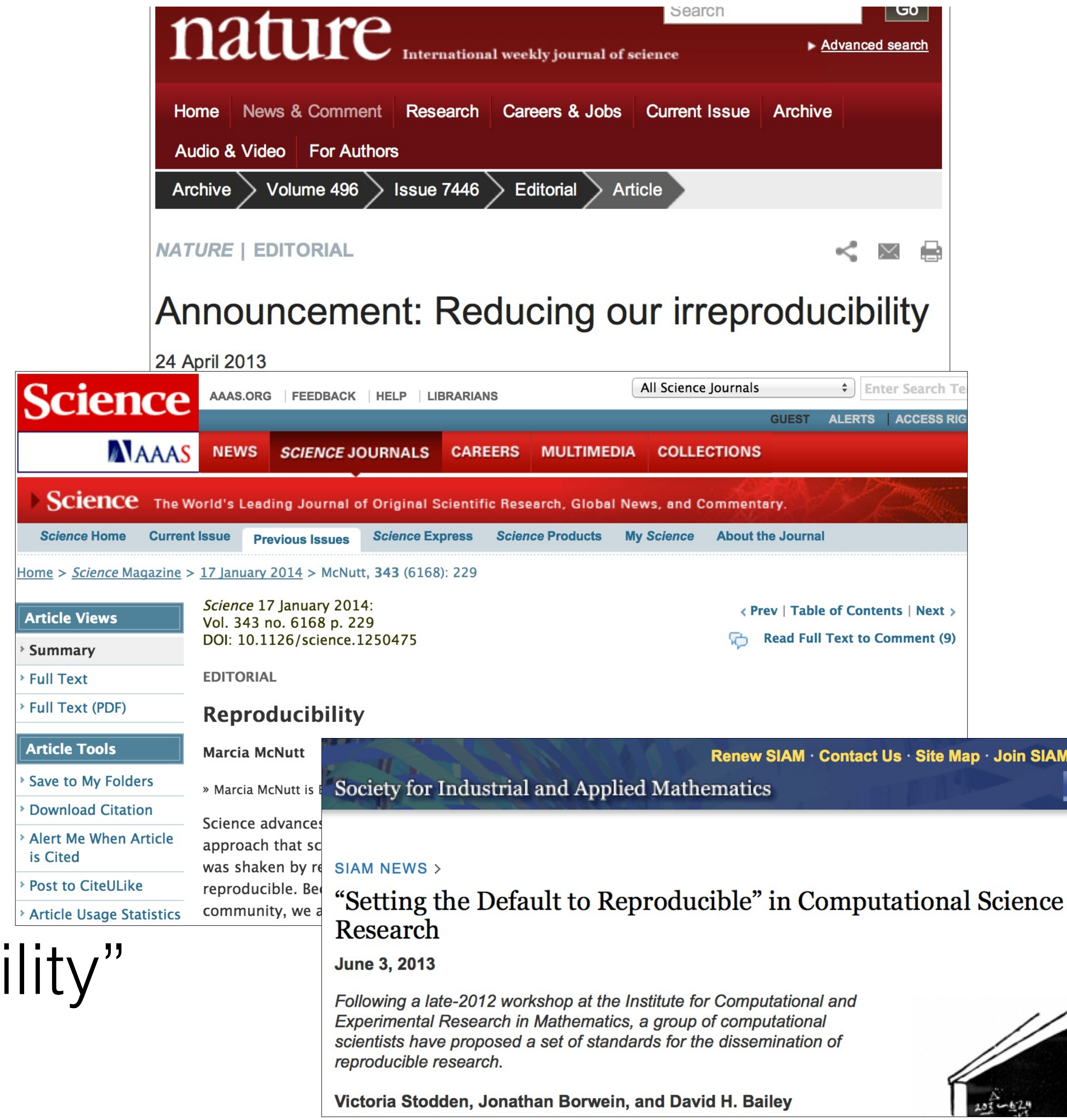
Parsing Reproducibility

"Empirical Reproducibility"

"Statistical Reproducibility"

"Computational Reproducibility"

V. Stodden, IMS Bulletin (2013)



Empirical Reproducibility

Cell Reports
Commentary

Cell PRESS
Open ACCESS

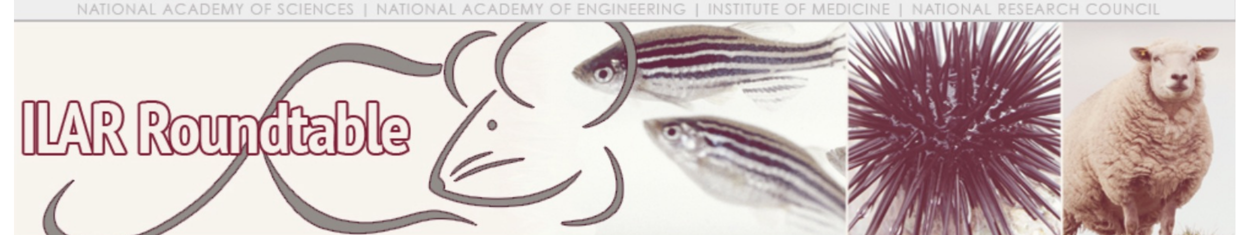
Sorting Out the FACS: A Devil in the Details

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

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



Home | About | Roundtable Members | Roundtable Activities | What's New at the ILAR Roundtable

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.

Tweet #ilar

Get updates!

Search Site

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*

Technological Sources of Impact

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

Claim: *Virtually all published discoveries today have a computational component.*

Corollary: *There is a mismatch between traditional scientific dissemination practices and modern computational research processes, leading to reproducibility concerns.*

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, insufficient sample size,
- Data preparation, treatment of outliers and missing values, re-combination of datasets,
- Inappropriate tests or models, model misspecification, poor parameter estimation techniques,
- Model robustness to parameter changes and data perturbations,
- …

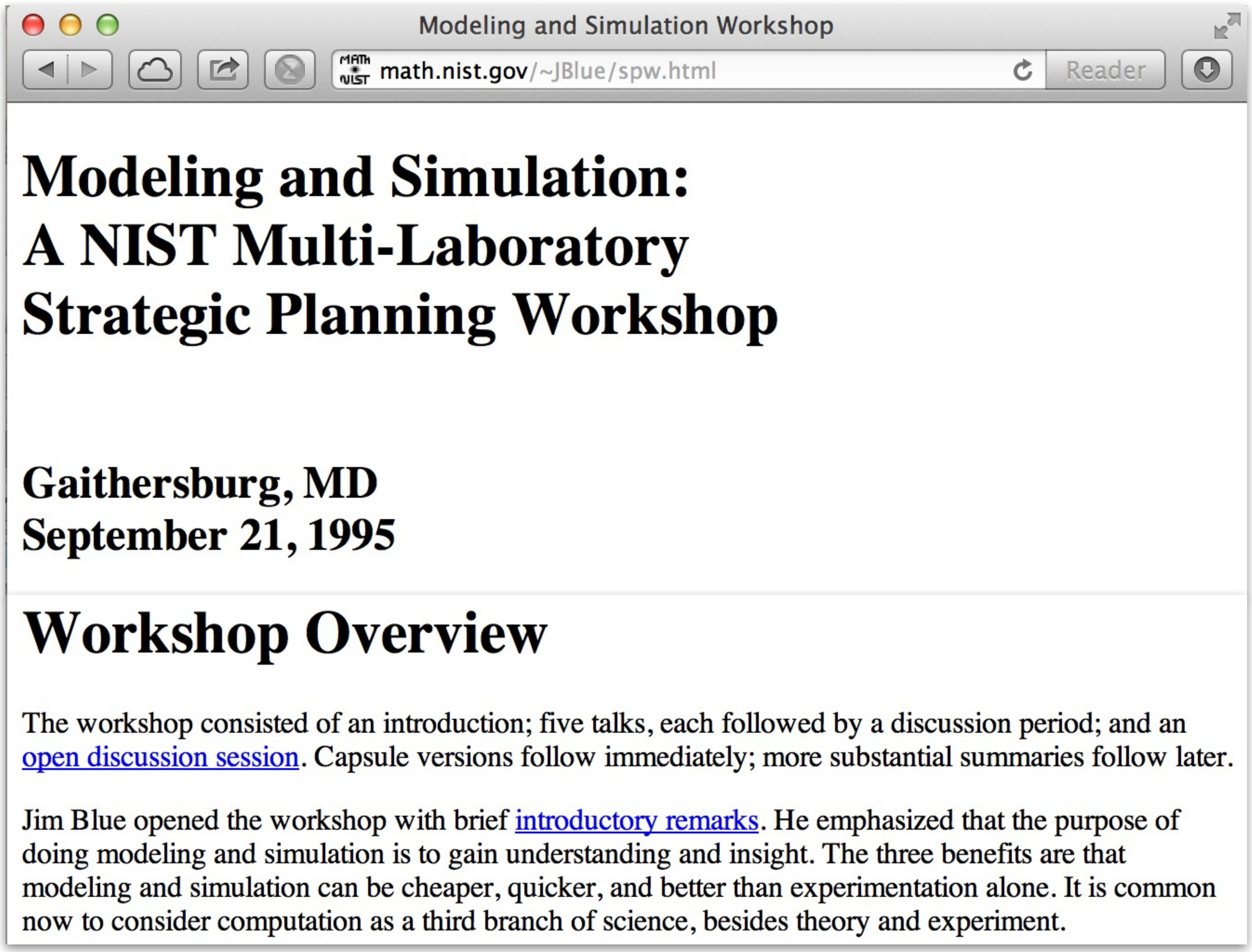
Computational Reproducibility

Traditionally two branches to the scientific method:

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

Now, new branches due to technological changes?

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



Modeling and Simulation Workshop

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

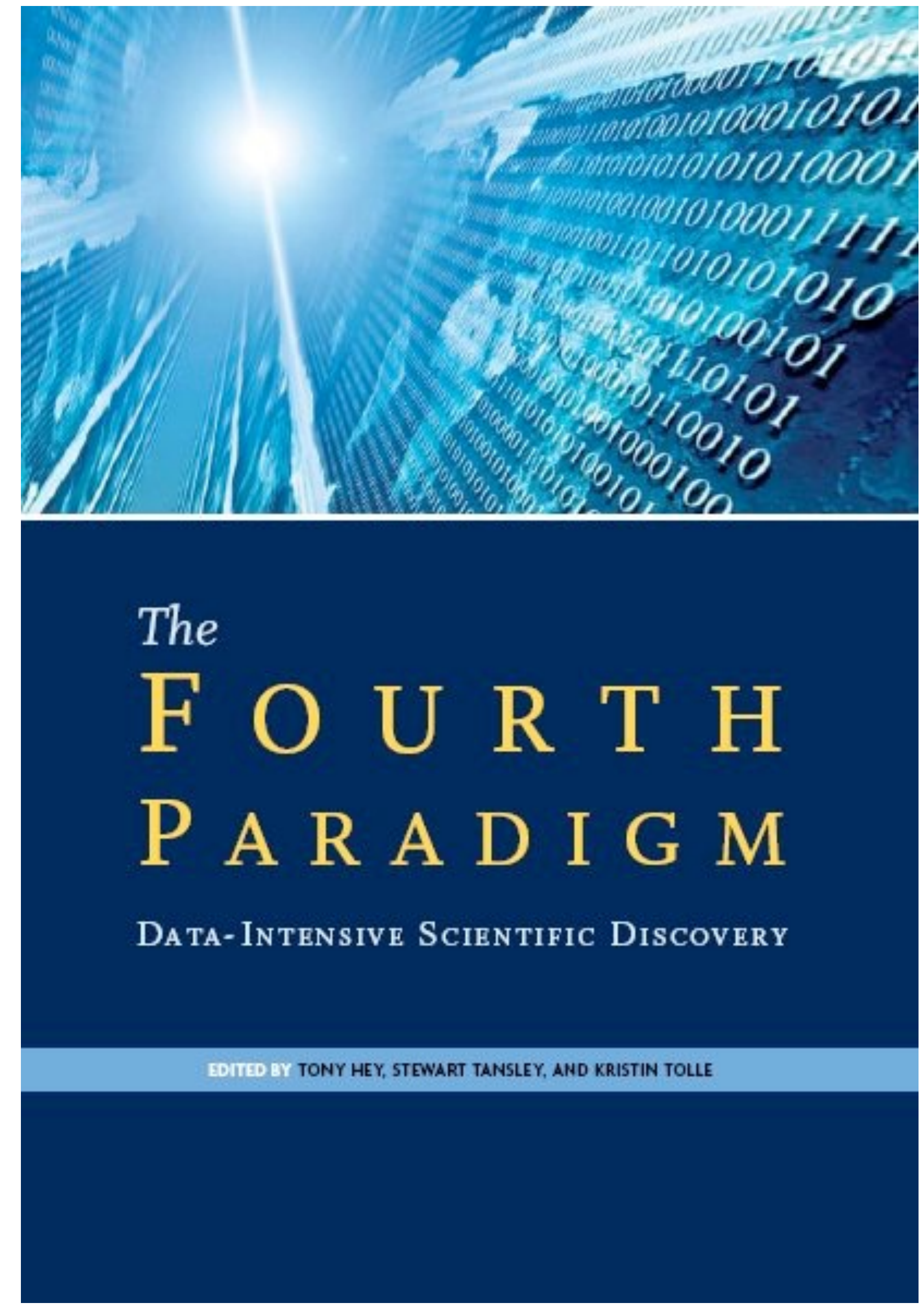
Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

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 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 and Data Science present only *potential* third/fourth branches of the scientific method (Donoho et al. 2009), until the development of comparable standards.

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 transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general proposals from the Transparency and Openness 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 from multiple studies.

RECOMMENDATIONS

Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable independent regeneration of computational 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://bit.ly/2fVwjPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly objects, citation should be standard practice. 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-

Workshop Recommendations: “Reproducibility Enhancement Principles”

1. Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.

2. 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.

Workshop Recommendations: "Reproducibility Enhancement Principles"

5. Use Open Licensing when publishing digital scholarly objects.

6. Journals should conduct a reproducibility check as part of the publication process and should enact the TOP standards at level 2 or 3.

7. To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.

“Reproducibility and Replication in Science”
Consensus Report, April 2019

National Academies of Science, Engineering, and Medicine

Definitions

- The terms, “reproducibility” and “replicability” have different meanings and uses across science and engineering, which has led to confusion in collectively understanding problems in reproducibility and replicability. The committee adopted specific definitions for the purpose of this report to clearly differentiate between the terms, which are otherwise interchangeable in everyday discourse.

- **Reproducibility** is obtaining *consistent results using the same input data, computational steps, methods, and code, and conditions of analysis.* This definition is synonymous with “computational reproducibility,” and the terms are used interchangeably in this report.

- **Replicability** is obtaining *consistent results across studies aimed at answering the same scientific question*, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.

Key Recommendation 1

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, *researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results* in order to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment:

- *the input data* used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;

- *a detailed description of the study methods (ideally in executable form)* together with its computational steps and associated parameters; and

- *information about the computational environment* where the study was originally executed, such as operating system, hardware architecture, and library dependencies (which are relationships described in and managed by a software dependency manager tool to mitigate problems that occur when installed software packages have dependencies on specific versions of other software packages).

Key Recommendation 2

RECOMMENDATION 6-3: *Funding agencies and organizations should consider investing in research and development of open-source, usable tools and infrastructure that support reproducibility* for a broad range of studies across different domains in a seamless fashion. Concurrently, investments would be helpful in outreach to inform and train researchers on best practices and how to use these tools.

Key Recommendation 3

RECOMMENDATION 6-5: In order to facilitate the transparent sharing and availability of digital artifacts, such as data and code, for its studies, the National Science Foundation (NSF) should:

- Develop a set of *criteria for trusted open repositories* to be used by the scientific community for objects of the scholarly record.

- Seek to *harmonize with other funding agencies* the repository criteria and data-management plans for scholarly objects.

- *Endorse or consider creating code and data repositories for long-term archiving and preservation of digital artifacts that support claims made in the scholarly record based on NSF-funded research*. These archives could be based at the institutional level or be part of, and harmonized with, the NSF-funded Public Access Repository.

- *Consider extending NSF's current data-management plan to include other digital artifacts, such as software.*

- Work with communities reliant on non-public data or code to *develop alternative mechanisms* for demonstrating reproducibility. Through these repository criteria, NSF would enable discoverability and standards for digital scholarly objects and discourage an undue proliferation of repositories, perhaps through endorsing or providing one go-to website that could access NSF-approved repositories.

Key Recommendation 4

RECOMMENDATION 6-6: *Many stakeholders have a role to play* in improving computational reproducibility, including educational institutions, professional societies, researchers, and funders.

- Educational institutions should educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.

- Professional societies should take responsibility for educating the public and their professional members about the importance and limitations of computational research. Societies have an important role in educating the public about the evolving nature of science and the tools and methods that are used.

- Researchers should collaborate with expert colleagues when their education and training are not adequate to meet the computational requirements of their research.

- In line with its priority for "harnessing the data revolution," the National Science Foundation (and other funders) should consider funding of activities to promote computational reproducibility.

Key Recommendation 5

RECOMMENDATION 6-9: Funders should require a thoughtful discussion in *grant applications of how uncertainties will be evaluated, along with any relevant issues regarding replicability and computational reproducibility. Funders should introduce review of reproducibility and replicability guidelines and activities into their merit-review criteria*, as a low-cost way to enhance both.

Conclusions

We see the convergence of two (ordinarily antagonistic) trends:

1. Scientific projects will become massively more computing intensive.

2. Research computing will become dramatically more transparent.

These are (in fact) reinforcing trends, which supports the convergence toward verifying and comparing findings.

Response: *Science* 2014

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.

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

Note: reproducing the computational steps vs re-implementing the experiment independently (both types needed).

Infrastructure Solutions

Research Environments and Document Enhancement Tools

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

Workflow Systems

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

Dissemination Platforms

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

Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) e.g the U.S. Constitution and subsequent Acts:

> "*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)

Copyright

- 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
- limited time: generally life of the author +70 years
- Exceptions and Limitations: e.g. Fair Use.

Patents

Patentable subject matter: "*new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof*" (35 U.S.C. §101) that is

1. *Novel*, in at least one aspect,

2. *Non-obvious*,

3. *Useful*.

USPTO Final Computer Related Examination Guidelines (1996) "A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena" (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.

Legal Issues in Data

- 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)).

- Copyright adheres to raw facts in Europe.

- the possibility of a residual copyright in data (attribution licensing or public domain certification).

- Legal mismatch: What constitutes a “raw” fact anyway?

The Reproducible Research Standard

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 **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

➡ *Remove copyright's barrier to reproducible research and,*

➡ *Realign the IP framework with longstanding scientific norms.*

A Convergence of Trends

- ➡ Scientific projects will become massively more computing intensive, and
- ➡ Scientific computing will become dramatically more transparent

Simultaneity: better transparency allows much more ambitious computational experiments. *And* better computational experiment infrastructure allows greater transparency.

Such a system is used not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work, enabling *efficiency* and *productivity*, and *discovery*.

“Quantitative Programming Environments”

- Define and create “Quantitative Programming Environments” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis

- The two 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.

Whole Tale: What's in a name… wholetale.org

A Double Entendre:

- Whole tale: captures the end-to-end scientific discovery story, including computational aspects
- Long tail: includes all computational research, e.g. bespoke or small scale research

Addresses Problems scientists face:

- Reproducibility (and reuse) challenges in computational & data-enabled research (e.g. data+code access, dependencies, …)

Whole Tale Approach:

- directly respond to community needs and requirements

Simplifying Computational Reproducibility in Whole Tale

Researchers can easily package and share tales:

- Data, Code, and Compute Environment
  - .. including narrative and workflow information including inputs, outputs, and intermediates
- to re-create the computational results from a study
- achieving computational reproducibility
- thus “setting the default to reproducible.”

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein. (2013). *Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics*, ICERM workshop (2013)

Empowers users to verify and extend results with different data, methods, and environments.