

Advancing Data Science as a Science: Paradigms, Practices, and Infrastructure

Victoria Stodden

School of Information Sciences
University of Illinois at Urbana-Champaign

DATAPALOOZA 2019

University of Virginia

November 1, 2019

Agenda

1. *Technology* is Disrupting Research (in good ways!)
2. *Reproducibility* in a Digital World
3. *Recommendations: AAAS 2016 Workshop Report and National Academies of Science, Engineering, and Medicine 2019 Consensus Report*
4. Data Science as a Science

Technological Sources of Impact

1. Big Data / Data Driven Discovery: high dimensional data, $p \gg n$,
2. Computational Power: simulation of the complete evolution of a physical system, systematically varying parameters,
3. Deep intellectual contributions now encoded only in software.



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

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

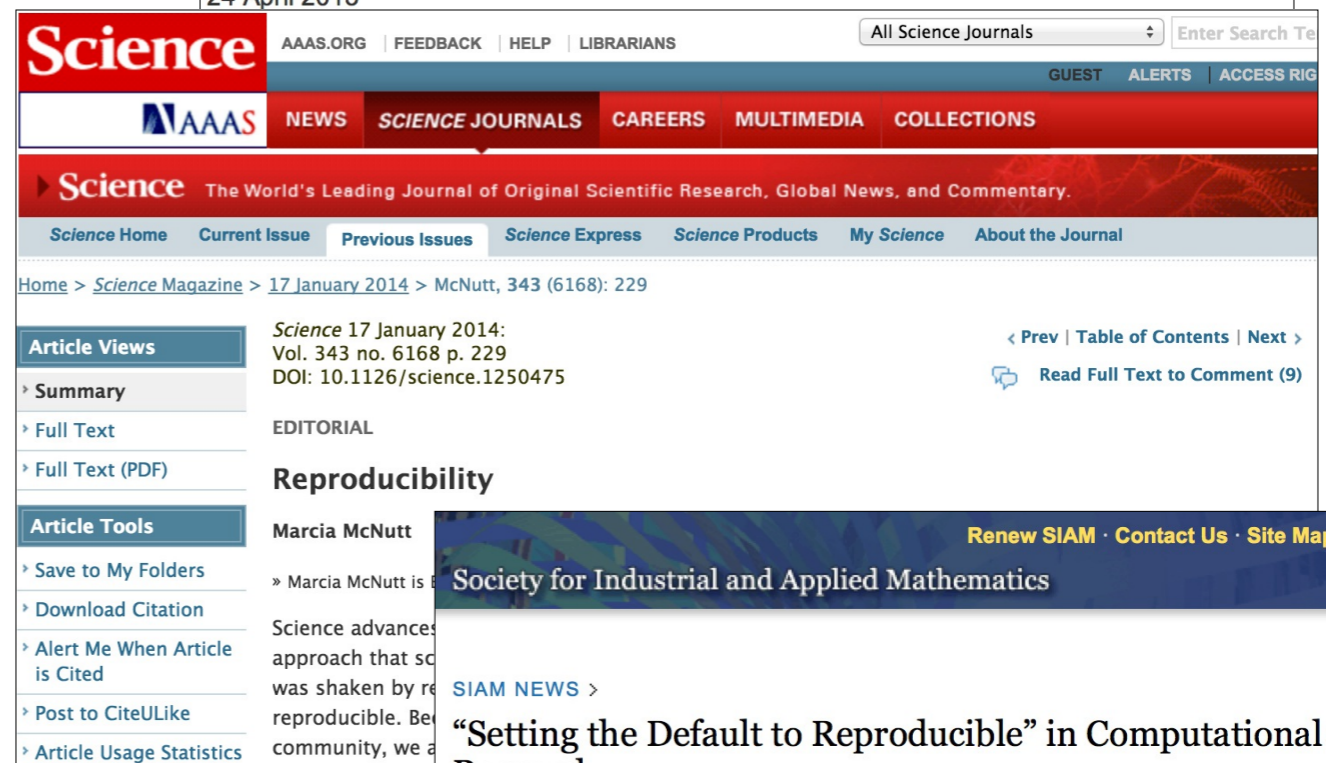
Corollary: *There is a mismatch between traditional scientific dissemination practices and modern computational research processes, leading to reproducibility concerns.*

Parsing Reproducibility

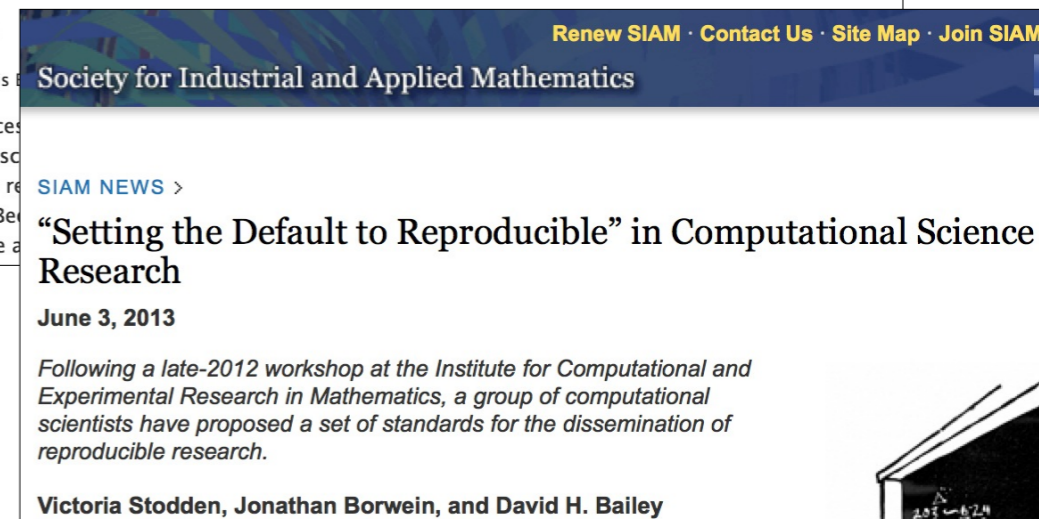
“Empirical Reproducibility”



“Statistical Reproducibility”



“Computational Reproducibility”



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}

¹Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA

²Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

³Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

⁴Department of Medicine, Harvard Medical School, Boston, MA 02115, USA

⁵These 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

ILAR Roundtable

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](#)

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,
- ...

Response to Statistical Reproducibility: *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.

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

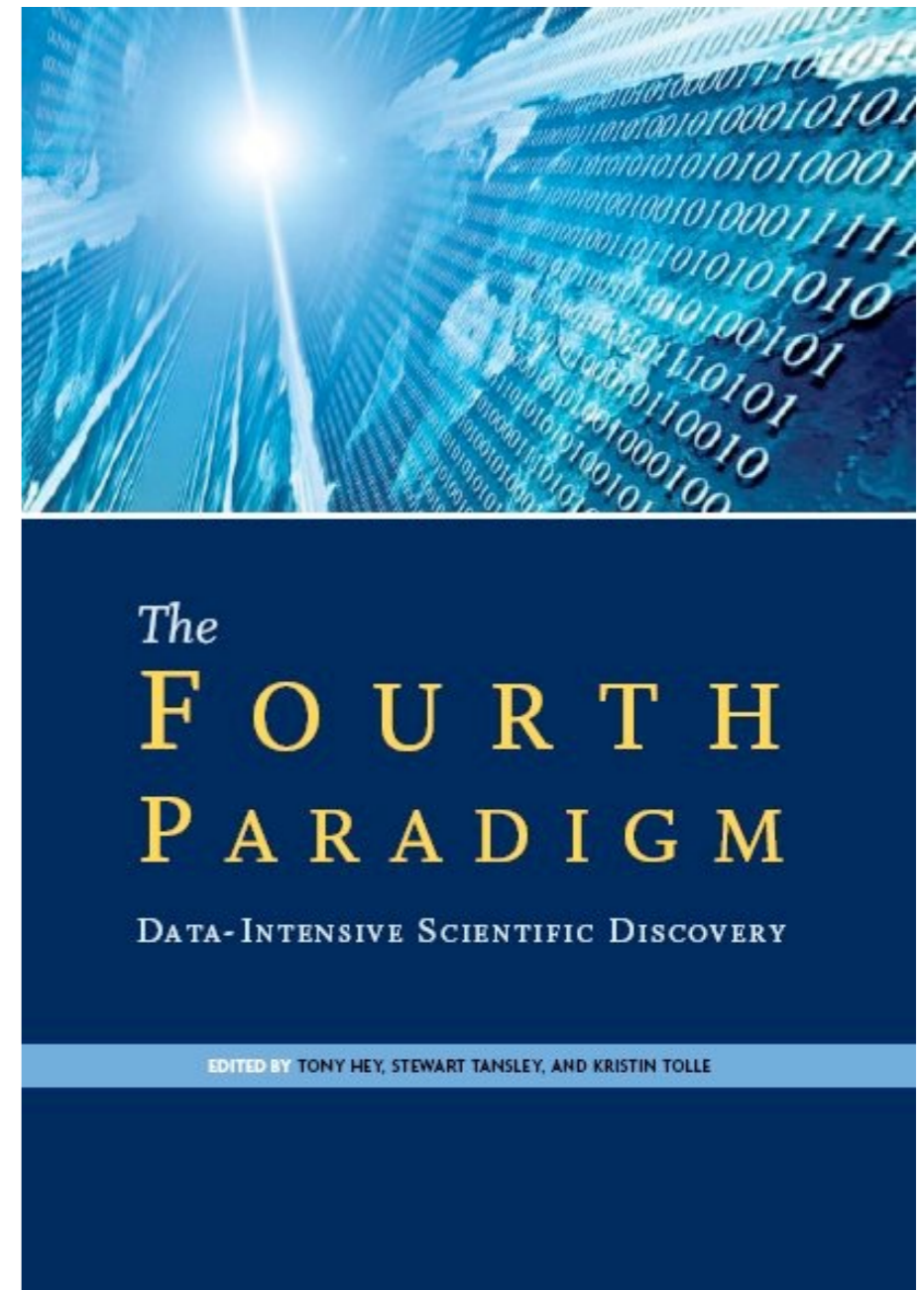
Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

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.

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

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

Workflow Systems

<u>Taverna</u>	<u>Wings</u>	<u>Pegasus</u>	<u>CDE</u>	<u>binder.org</u>
<u>Kurator</u>	<u>Kepler</u>	<u>Everware</u>	<u>Reprozip</u>	<u>Galaxy</u>

Dissemination Platforms

<u>ResearchCompendia.org</u>	<u>DataCenterHub</u>	<u>RunMyCode.org</u>	<u>ChameleonCloud</u>
<u>Occam</u>	<u>RCloud</u>	<u>TheDataHub.org</u>	<u>Madagascar</u>
<u>Wavelab</u>	<u>Sparselab</u>		

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 Taufer⁸

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-

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



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-

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-

Workshop Recommendations: “Reproducibility Enhancement Principles”

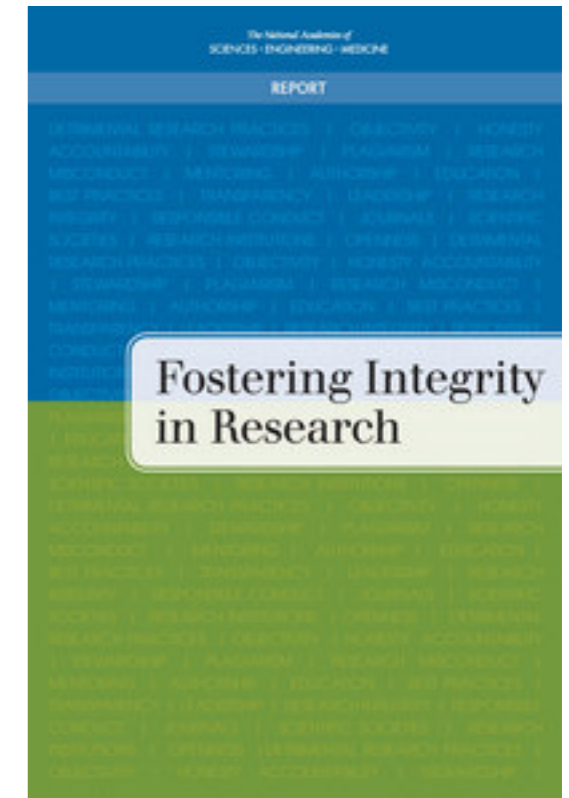
1. Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.
2. Persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.
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.

“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.**

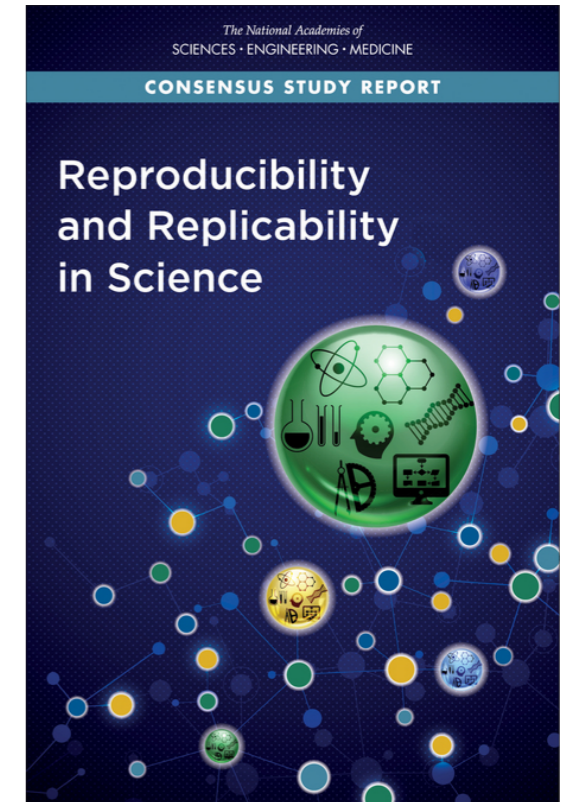
"Fostering Integrity in Research," National Academies of Sciences, Engineering, and Medicine, 2017

“Reproducibility and Replication in Science”

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.



“Reproducibility and Replication in Science” National Academies of Science, Engineering, and Medicine, May 2019

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 6-9: 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.

Key Recommendation 5

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.

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!

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 Supplement	3%
Shared data and code	36%
<hr/>	
Total	100%

24 articles provided direct access to code/data.

Replicating Computational Findings

Computational Replication Rates

- 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

We obtained 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 Masterrunfigure-one.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. Research will become massively more computational,
2. Research 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**.

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

Imagine: Querying the Scholarly Record

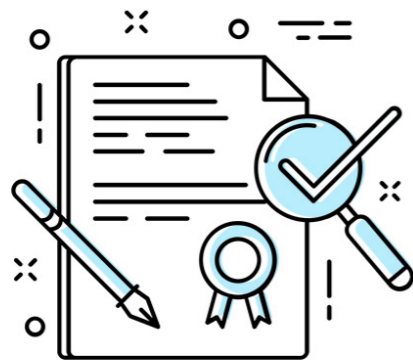
- Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;
- Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;
- List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;
- Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;
- Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the 2003 and list the trial name and histogram side by side.

Conclusion

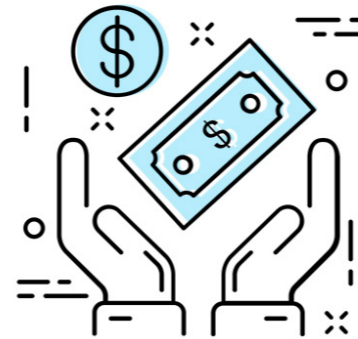
Change is enabled by standards and community emphasis from a variety of stakeholders.



**Universities/
institutions**
(hiring/promotion;
programmatic change)



Publishers
(TOP guidelines)



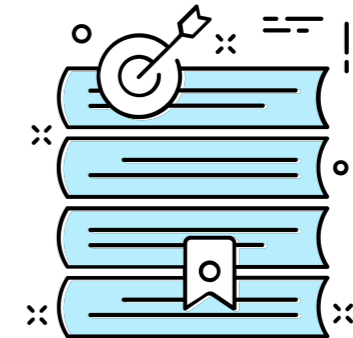
Funders
(policy)



Researchers
(processes)



Scientific Societies



**Universities/
libraries**
(empowering w/tools)



Regulatory Bodies
(OSTP)

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)

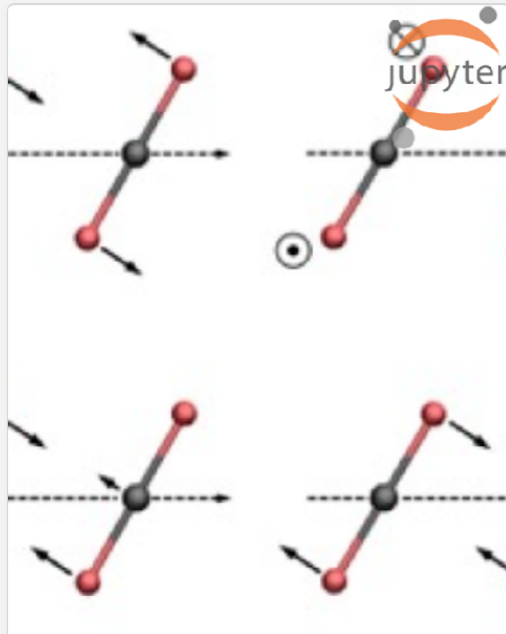
Empowers users to verify and extend results with different data, methods, and environments.

Browse Tales Launch to add to Launched Tales list



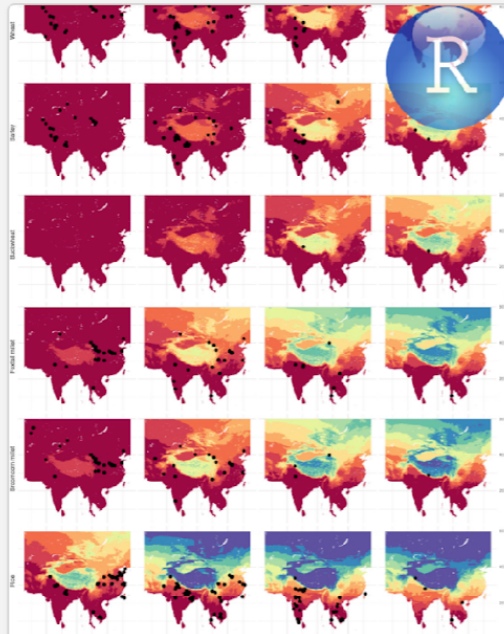
Search tales...

All [Switch to list view](#)



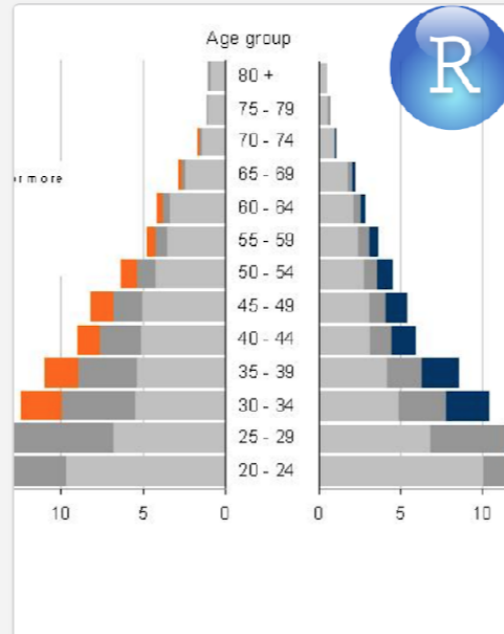
COMPUTATIONAL CHEMISTRY Anharmonic vibrational structure of...

This project produces all of the data from the Anharmonic vibrational structure of the carbon dioxide dimer with a many-body potential energy surface journal article. The project solves the vibrational Schrodinger equation for the CO2 monomer and dimer



ARCHAEOLOGY Climate change stimulated agricultu...

Ancient farmers experienced climate change at the local level through variations in the yields of their staple crops. However, archaeologists have had difficulty in determining where, when, and how changes in climate affected ancient farmers. We



ECONOMICS L2-Boosting for Economic Applicatio...

Replication package for: L₂-Boosting for Economic Applications
The authors present the L₂--Boosting algorithm and two variants, namely post-Boosting and orthogonal Boosting. Building on results in Ye and Spindler (2016), they

Launched Tales

L2-Boosting for Economic Applicatio...

Browse Existing Tales ...

 **Compose** *Create a new Tale by pairing a compute environment with a dataset* 

Tale name:

L2-Boosting for Economic Applications

Compute environment:

 RStudio (rocker/geospatial)









Input data:

[Launch New Tale](#)

... *Compose New Tales* ...

 **Environments**

Search compute environments... 

-  RStudio (rocker/geospatial) 
-  Jupyter Classic 
-  RStudio 
-  Jupyter Lab 

L2-Boosting for Economic Applicatio...
Ye Luo and Martin Spindler

Interact Files Metadata

File Edit Code View Plots Session Build Debug Profile Tools Help

```

1 | #####
2 | # L2-Boosting for Economic Applications
3 | #####
4 | # Parameter for simulation study
5 | rm(list=ls())
6 | source("DGP.R")
7 | source("helper.R")
8 | R <- 500 # number of repitions
9 | set.seed(12345)
10 | library(MASS)
11 | library(mvtnorm)
12 | library(hdm)
13 | library(newboost) # can be downloaded from R-Forge or requested by the
14 | #####
15 | # IV Estimation
    
```

Console Terminal x

```

/WholeTale/workspace/

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> load("/WholeTale/workspace/Sim_AER.RData")
> |
    
```

Environment History Connections Jobs

Global Environment

Data

data	List of 3
ds	num [1:90, 1] -1.24 -0.974 1.33 -0.154 -0...
ED	List of 6
ED1	List of 6
EDB	List of 6

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

WholeTale > workspace

Name	Size	Modified
..		
apt.txt	5 B	Mar 6, 2019, 1:43 PM
DGP.R	1.5 KB	Mar 5, 2019, 3:36 PM
helper.R	9.2 KB	Mar 5, 2019, 3:36 PM
install.R	148 B	Mar 5, 2019, 3:36 PM
Readme.pdf	60.7 KB	Mar 5, 2019, 3:36 PM
runtime.txt	13 B	Mar 5, 2019, 3:36 PM
Sim_AER.RData	6.6 MB	Mar 5, 2019, 4:14 PM
Sim_AER_V3.R	5.3 KB	Mar 5, 2019, 3:46 PM

Launched Tales

L2-Boosting for Economic Applicatio...

...

Run & Interact with Tales ...

L2-Boosting for Economic Applicatio... Ye Luo and Martin Spindler

Interact Files Metadata

Title L2-Boosting for Economic Applications

Authors Ye Luo and Martin Spindler

Category Economics

Environment RStudio (rocker/geospatial)

Date Created Tue Mar 05 2019 15:36:05 GMT-0600 (Central Standard Time)

Last Updated Wed Mar 06 2019 13:18:07 GMT-0600 (Central Standard Time)

Edit Preview

Replication package for: L₂-Boosting for Economic Applications

Description

The authors present the L₂-Boosting algorithm and two variants, namely post-Boosting and orthogonal Boosting. Building on results in Ye and Spindler (2016), they demonstrate how boosting can be used for estimation and inference of low-dimensional treatment effects. In particular, we consider estimation of a treatment effect in a setting with very many controls and in a setting with very many instruments. We provide simulations and analyze two real applications.

Based on <https://www.aeaweb.org/articles?id=10.1257/aer.p20171040>

Illustration

<https://raw.githubusercontent.com/whole-tale/dashboard/master/public/images/demo-graph2.jpg>

or Generate Illustration

Launched Tales

L2-Boosting for Economic Applicatio...

...
**Use
Tale
Metada
ta ...**

Setting the Stage: Data *Science*

sci·ence

/ˈsiːəns/

The intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation and experiment.

sci·en·tif·ic meth·od

A method of procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses.

re·pro·du·ci·bil·i·ty

/ˌriːprəˌdjuːsəˈbɪlɪti/

The extent to which consistent results are obtained when an experiment is repeated.
‘the experiments were conducted numerous times to test the reproducibility of the results’

Trans·par·en·cy

/ˌtrænsˈpærənsi/

The transparency of a process, situation, or statement is its quality of being easily understood or recognized, for example because there are no secrets connected with it, or because it is expressed in a clear way.

How are we doing?

re·pro·du·ci·bil·i·ty
/ˌriːprəˌdjuːsəˈbɪlɪti/

The extent to which consistent results are obtained when an experiment is repeated.
'the experiments were conducted numerous times to test the reproducibility of the results'

Trans·par·en·cy
/ˌtrænsˈpærənsi/

The transparency of a process, situation, or statement is its quality of being easily understood or recognized, for example because there are no secrets connected with it, or because it is expressed in a clear way.

There is an ongoing convergence of two (ordinarily antagonistic) trends that will resolve with transparency and reproducibility:

- 1. Scientific projects will continue to become massively more computing intensive*
- 2. Research computing will become dramatically more transparent*

These are reinforcing trends, whose resolution essential for verifying and comparing findings.

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**.