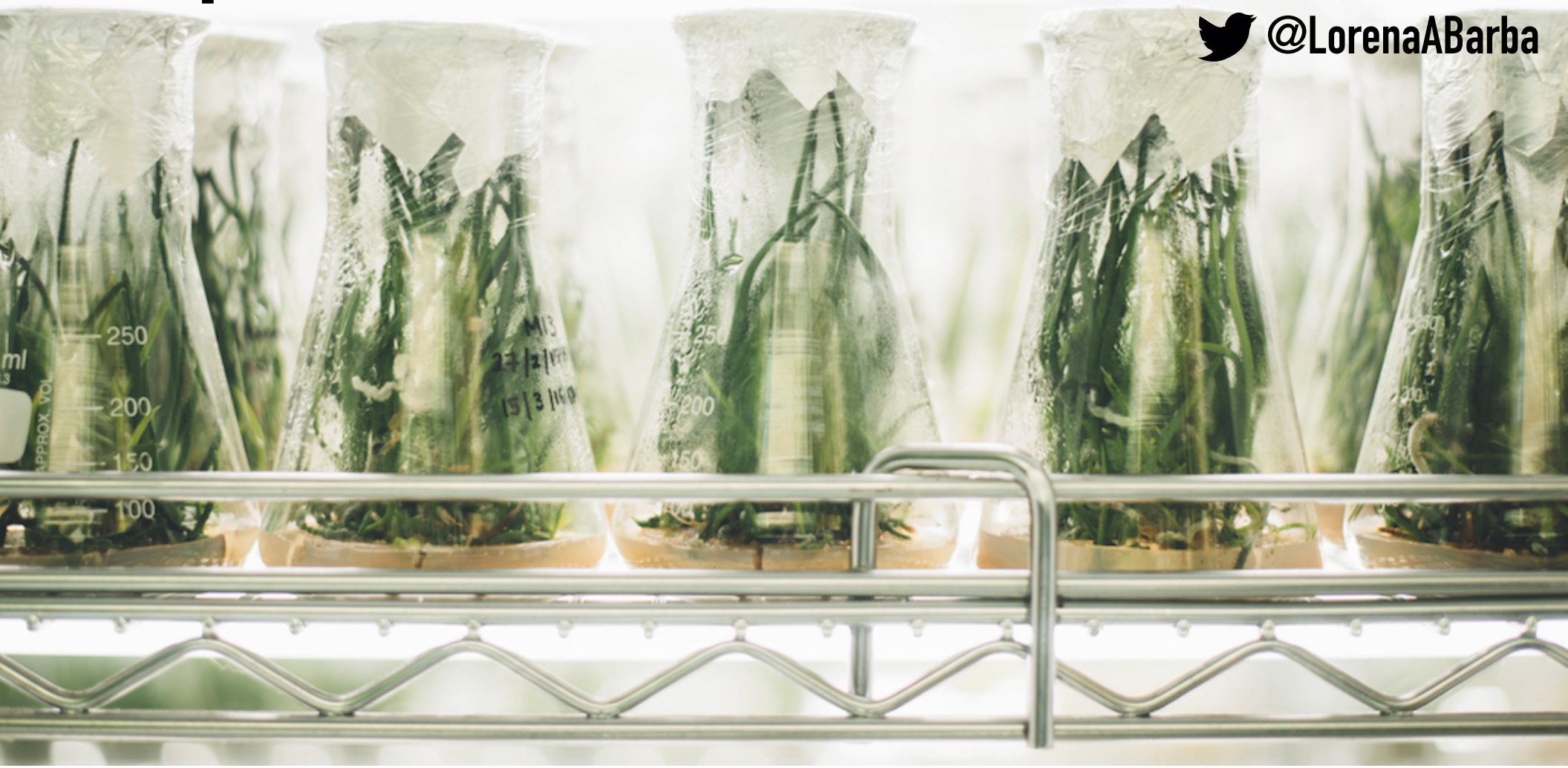
IACS Stony Brook University

The Why and How of Reproducible Computational Research

@LorenaABarba



Reproducibility hit the mainstream. . .

NSF 17-022
Dear Colleague Letter: Encouraging Reproducibility in Computing and Communications Research

CISE, October 21, 2016

https://www.nsf.gov/pubs/2017/nsf17022/nsf17022.jsp

NSF SBE subcommittee on replicability in science:

“reproducibility refers to the ability of a researcher to duplicate results of a prior study using the same materials as were used by the original investigator."
“… new evidence is provided by new experimentation, defined in the NSF report as ‘replicability’ “

SBE, May 2015

National Institutes of Health
Turning Discovery Into Health

Health Information | Grants & Funding | News & Events | Research & Training

Home » Research & Training

RIGOR AND REPRODUCIBILITY

Rigor and Reproducibility

- Principles and Guidelines
- Expanded Guidelines
- Application Instructions
- Training
- Funding Opportunities
- Meetings and Workshops
- Announcements
- Publications

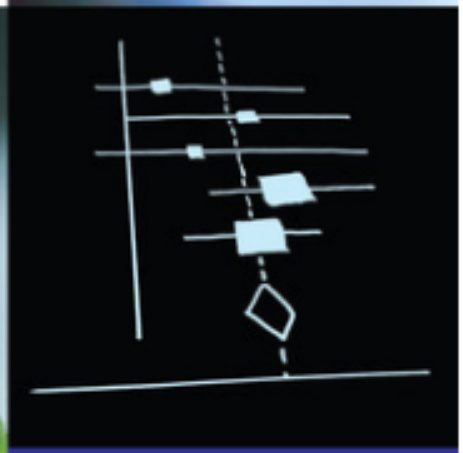
When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research.

students in a
University

https://www.nih.gov/research-training/rigor-reproducibility

Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results

Summary of a Workshop



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

The National Academies of
SCIENCES
ENGINEERING
MEDICINE

https://doi.org/10.17226/21915

http://sc16.supercomputing.org/2016/03/16/sc16-explores-reproducibility-advanced-computing-student-competition-michela-taufer/

SC16 Panel: "Different Architectures, Different Times: Reproducibility and Repeatability in High Performance Computing"

26 JUNE 2015
VOL 348, ISSUE 6242

SCIENTIFIC STANDARDS

Promoting an open research culture

Author guidelines for journals could help to promote transparency, openness, and reproducibility

By **B. A. Nosek,* G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni**

8 MODULAR STANDARDS

Citation Standards Describes citation of data	**Data Transparency** Describes availability and sharing of data
Analytical Methods Transparency Describes analytical code accessibility	**Research Materials Transparency** Describes research materials accessibility
Design and Analysis Transparency Sets standards for research design disclosures	**Preregistration of Studies** Specification of study details before data collection
Preregistration of Analysis Plans Specification of analytical details before data collection	**Replication** Encourages publication of replication studies

ACROSS 3 TIERS

1 DISCLOSURE: the final research output must disclose if the work satisfies the standard

2 REQUIREMENT: the final research output must satisfy the standard

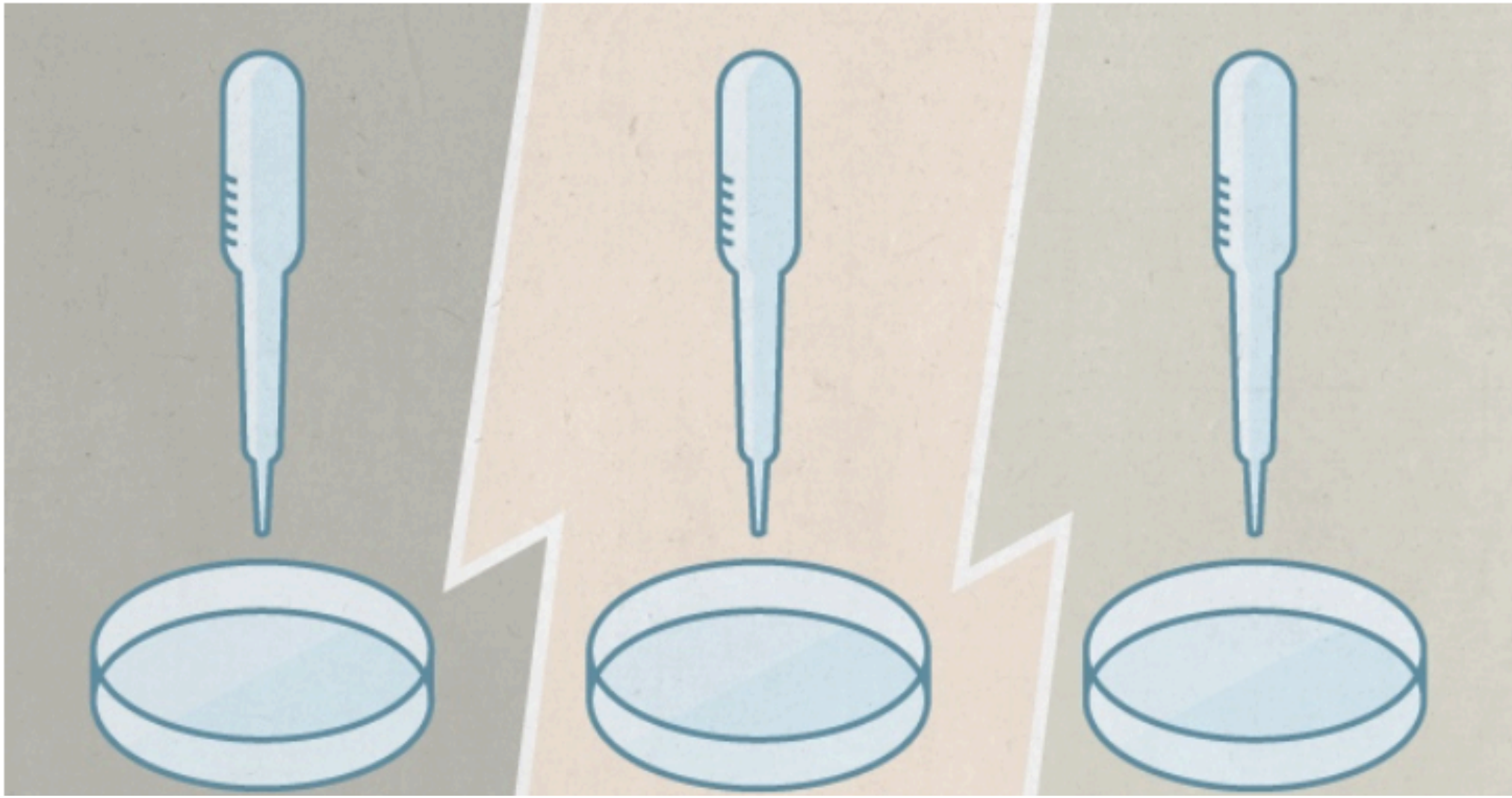
3 VERIFICATION: third party must verify that the standard is being met

https://cos.io/our-services/top-guidelines/

nature
International weekly journal of science

SPECIAL

See all specials



nature International weekly journal of science

Nature's survey of 1,576 researchers

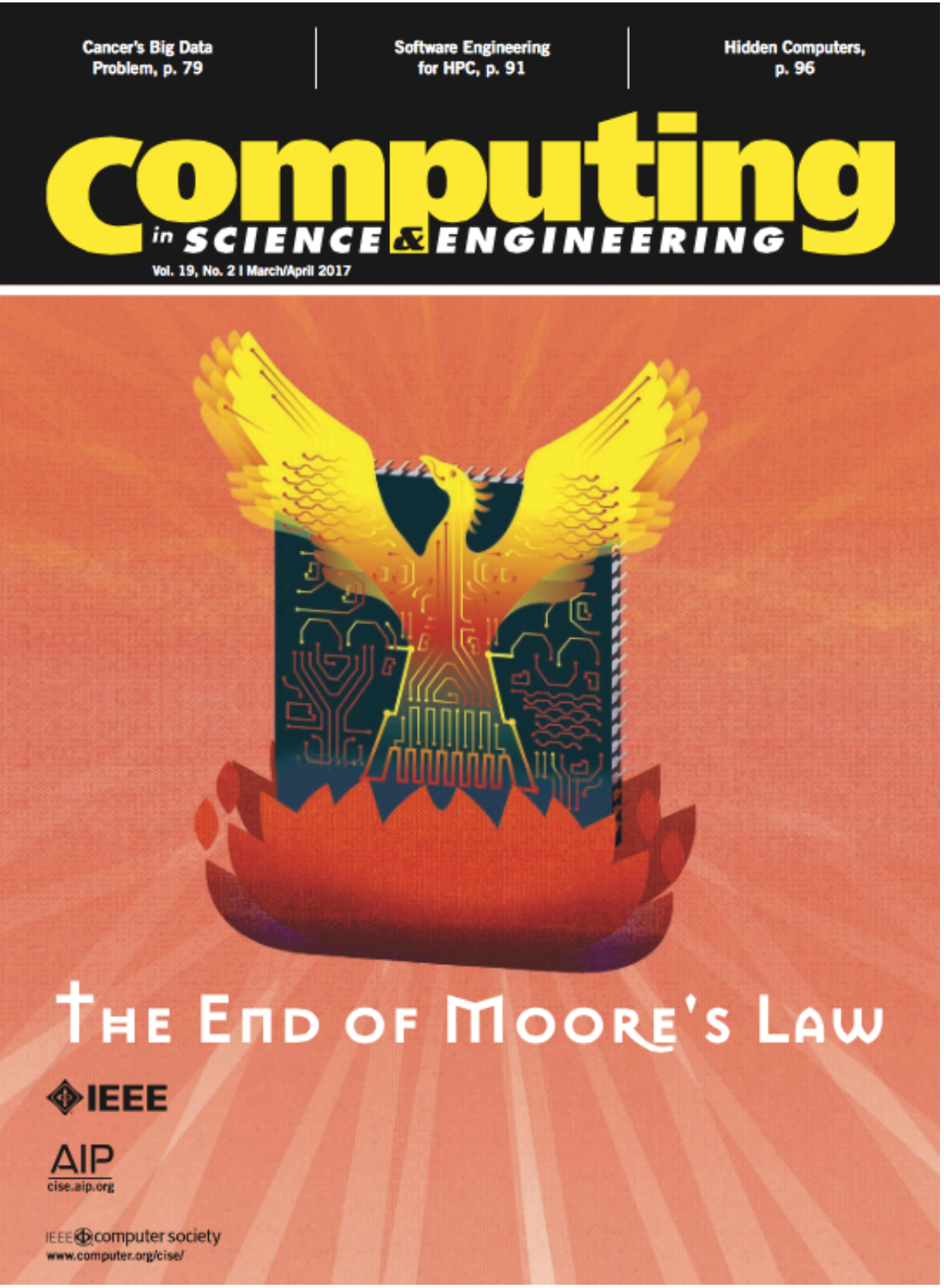
- >70% of researchers have tried and failed to reproduce another scientist's experiments
- >50% have failed to reproduce their own experiments
- ~52% agree that there is a significant 'crisis' of reproducibility

http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Technical Consortium on High Performance Computing

New initiative on Reproducibility, led by Barba.

`https://www.computer.org/web/tchpc`



Reproducible Research Track (peer reviewed)

Lorena A. Barba
George Washington University
labarba@gwu.edu

George K. Thiruvathukal
Loyola University Chicago
gkt@cs.luc.edu

https://www.computer.org/cise/

Def.— Reproducible research

Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results.

Schwab, M., Karrenbach, N., Claerbout, J. (2000) "Making scientific computations reproducible," *Computing in Science and Engineering* Vol. 2(6):61–67

Jon F. Claerbout
Professor Emeritus of Geophysics
Stanford University

… pioneered the use of computers in processing and filtering seismic exploration data [Wikipedia]

… from 1991, he required theses to conform to a standard of reproducibility.

“In 1990, we set this sequence of goals:

1. Learn how to merge a publication with its underlying computational analysis.
2. Teach researchers how to prepare a document in a form where they themselves can reproduce their own research results a year or more later by “pressing a single button”.
3. Learn how to leave finished work in a condition where coworkers can reproduce the calculation including the final illustration by pressing a button in its caption.
4. Prepare a complete copy of our local software environment so that graduating students can take their work away with them to other sites, press a button, and reproduce their Stanford work.
5. Merge electronic documents written by multiple authors (SEP reports).
6. Export electronic documents to numerous other sites (sponsors) so they can readily reproduce a substantial portion of our Stanford research.

http://library.seg.org/doi/abs/10.1190/1.1822162

Think about your latest paper or report . . .

Def.— Replication

Arriving at the same scientific findings as another study, collecting new data (possibly with different methods) and completing new analyses.

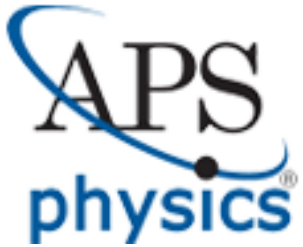
Roger D. Peng (2011), "Reproducible Research in Computational Science" *Science*, Vol. 334, Issue 6060, pp. 1226-1227

Why?

We use computing to create scientific knowledge.

What is Science?

- American Physical Society:



- Ethics and Values, 1999

  "The success and credibility of science are anchored in the willingness of scientists to [...] Expose their ideas and results to **independent testing and replication by others**. This requires the open exchange of data, procedures and materials."

https://www.aps.org/policy/statements/99_6.cfm

Published online 13 October 2010 | *Nature* **467**, 775-777 (2010) |
doi:10.1038/467775a

News Feature

Computational science: ...Error

...why scientific programming does not compute.

Zeeya Merali

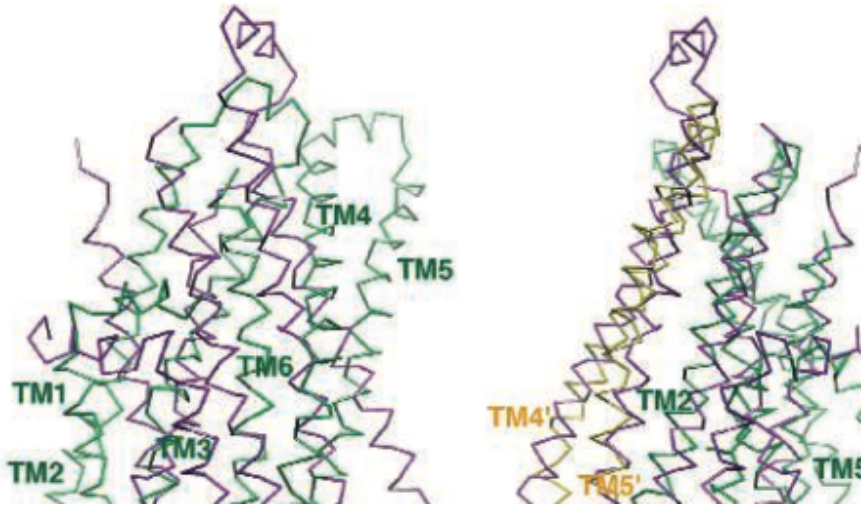
A dolphin's demise
1860

SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was on a trajectory most young scientists only dream about. In 1999, at the age of 28, the protein crystallographer landed a faculty position at the prestigious Scripps Research Institute in San Diego, California. The next year, in a ceremony at the White House, Chang received a Presidential Early Career Award for Scientists and Engineers, the country's highest honor for young researchers. His lab generated a stream of high-profile papers detailing the molecular structures of important proteins embedded in cell membranes.

Then the dream turned into a nightmare. In September, Swiss researchers published a paper in *Nature* that cast serious doubt on a 2001 *Science* paper, which described the structure of a protein called MsbA, isolated from the bacterium *Escherichia coli*. MsbA belongs to a huge and ancient family of molecules that use energy from adenosine triphosphate to transport molecules across cell membranes. These so-called ABC transporters perform many



Retraction Watch

Error in one line of code sinks cancer study

without comments

Authors of a 2016 cancer paper have retracted it after finding an error in one line of code in the program used to calculate some of the results.

Sarah Darby, last author of the now-retracted paper from the University of Oxford, UK, told *Retraction Watch* that the mistake was made by a doctoral student. When the error was realized, Darby said, she contacted the *Journal of Clinical Oncology* (*JCO*), explained the issue, and asked whether they would prefer a retraction or a correction. *JCO* wanted a retraction, and she complied.

The journal allowed the authors to publish a correspondence article outlining their new results.

The New York Times

The Opinion Pages | OP-ED COLUMNIST

The Excel Depression

Paul Krugman APRIL 18, 2013

The story so far: At the beginning of 2010, two Harvard economists, Carmen Reinhart and Kenneth Rogoff, circulated a paper, "Growth in a Time of Debt," that purported to identify a critical "threshold," a tipping point, for government indebtedness. Once debt exceeds 90 percent of gross domestic product, they claimed, economic growth drops off sharply.

BUSINESS INSIDER

MARKETS

Shocking Paper Claims That Microsoft Excel Coding Error Is Behind The Reinhart-Rogoff Study On Debt

Mike Konczal, NewDeal2.0

Apr. 16, 2013, 12:40 PM 92,101

THE WALL STREET JOURNAL.

REAL TIME ECONOMICS

Reinhart, Rogoff Admit Excel Mistake, Rebut Other Critiques

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

How?

Transparency & Open Science

Donoho, D. et al. (2009) “Reproducible research in computational harmonic analysis,” *Computing in Science and Engineering* Vol. 11(1):8–18.

Data and Code Sharing **Recommendations**

- assign a unique identifier to every version of the data and code
- describe in each publication the computing environment used
- use open licenses and non-proprietary formats
- publish under open-access conditions (and/or post pre-prints)

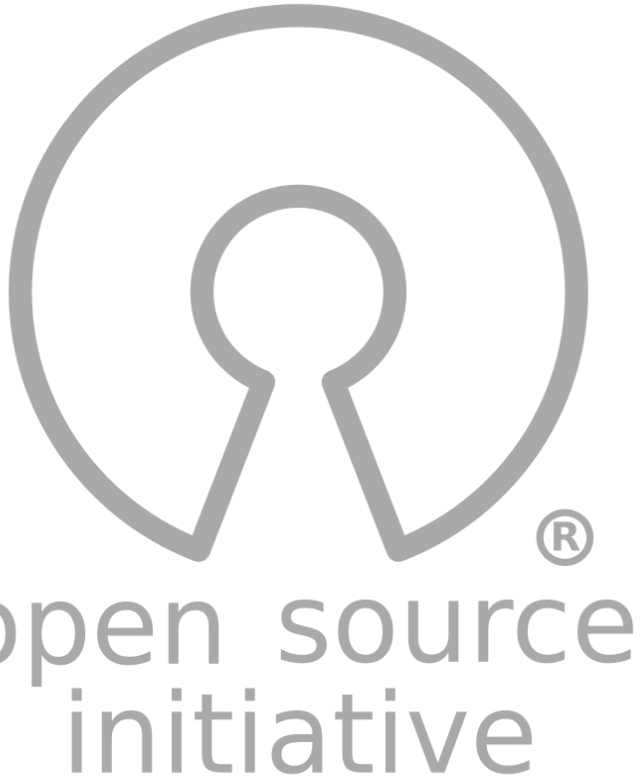
NEWS

REPRODUCIBLE RESEARCH

ADDRESSING THE NEED FOR DATA AND CODE SHARING IN COMPUTATIONAL SCIENCE

By the Yale Law School Roundtable on Data and Code Sharing

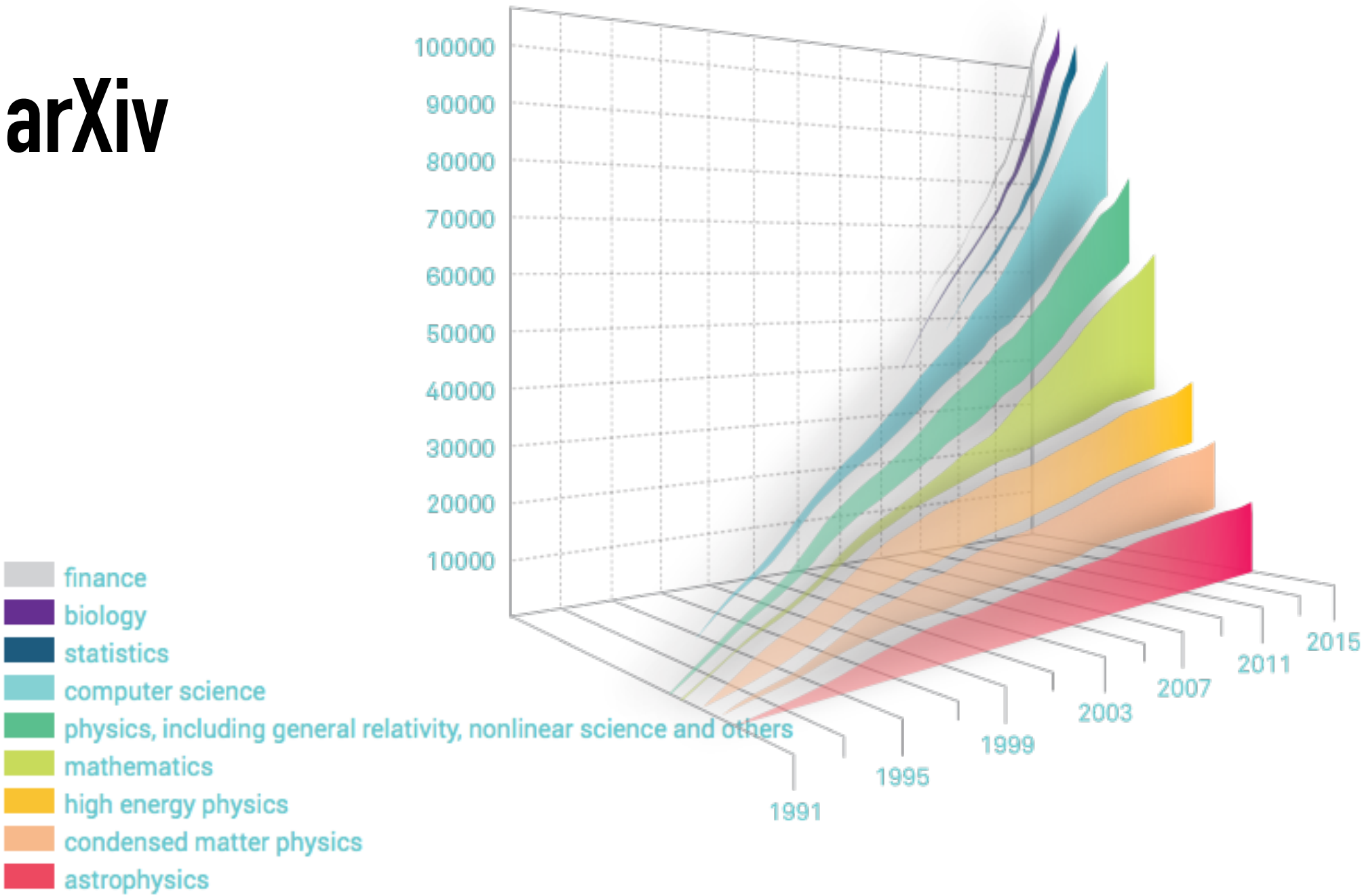
Roundtable participants identified ways of making computational research details readily available,



Open-source licenses:

People can coordinate their work freely, within the confines of copyright law, while making access and wide distribution a priority.

arXiv



https://www.simonsfoundation.org/report2015/stories/arxiv.html


engrχiv


bioRχiv
beta
THE PREPRINT SERVER FOR BIOLOGY


SOC
ARXIV

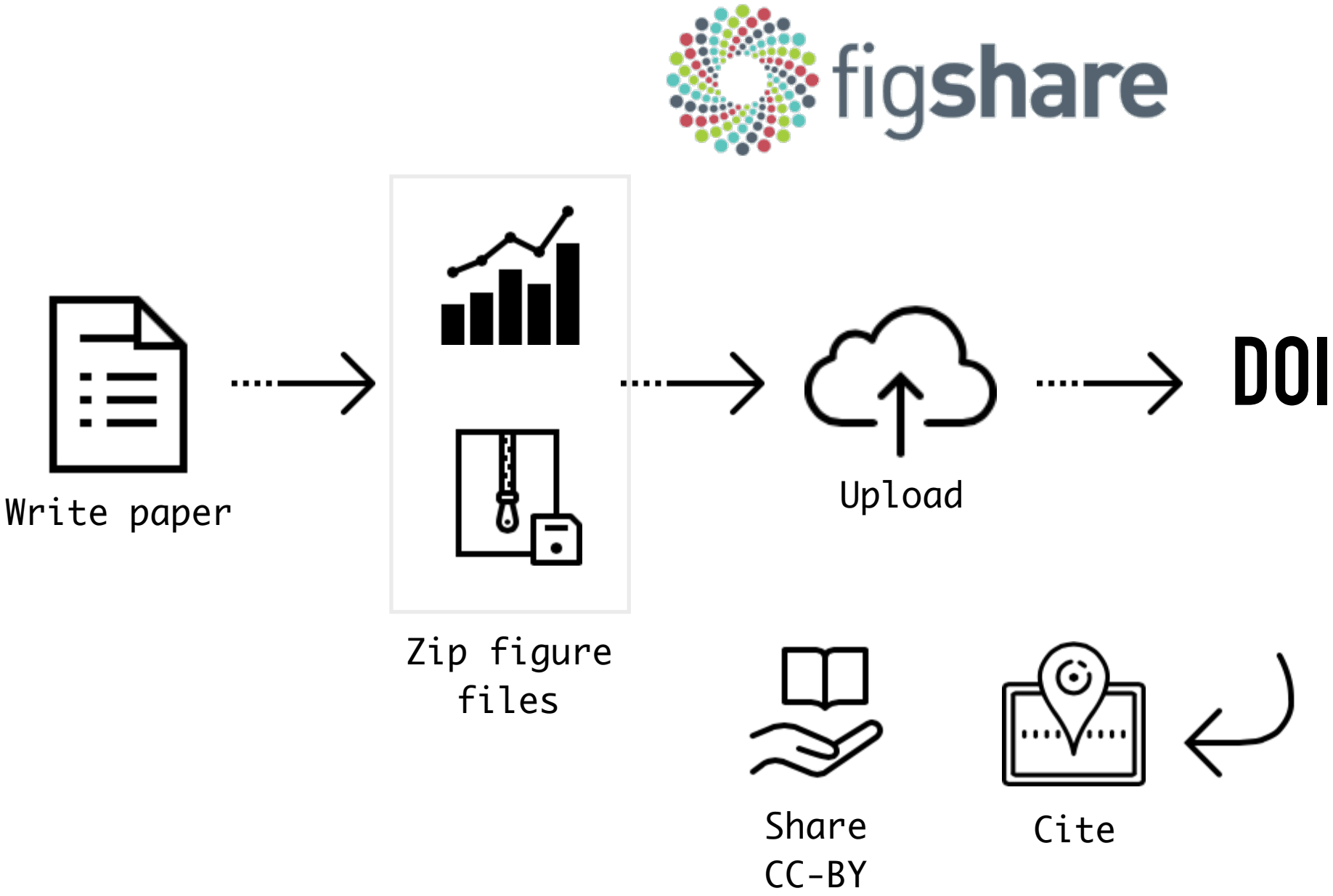

OSFPREPRINTS

ReproPacks

For main results in a paper, we share data, plotting script & figure under CC-BY.

File bundle with input data, running scripts, plotting scripts, and figure.

We cite our own figure in the caption!


figshare
Write paper
Zip figure files
Upload
DOI
Share CC-BY
Cite

The Journal of Open Source Software

Submit Papers About Sign In

The Journal of Open Source Software

A **developer friendly** journal for research software packages.

open source initiative

NUMFOCUS

OPEN CODE = BETTER SCIENCE

http://joss.theoj.org

The right tools

Two points of contention:
— scripted figures (vs. GUI-based tools)
— version control

“I’ve learned that interactive programs are slavery (unless they include the ability to arrive in any previous state by means of a script).”
— Jon Claerbout

Automation:
Turn protocols into code

Version control

10 août 1887

9 septembre 1887

8 octobre 1887

10 novembre 1887

14 décembre 1887

15 mars 1888

10 avril 1888

10 mai 1888

juin 1888

juillet 1888

14 août 1888

14 septembre 1888

14 octobre 1888

14 novembre 1888

26 décembre 1888

20 janvier 1889

12 février 1889

12 mars 1889

NATUREJOBS | NATUREJOBS BLOG

TechBlog: My digital toolbox: Lorena Barba

17 Apr 2017 | 12:00 BST | Posted by Jeffrey Perkel | Category: Blog, Technology

http://blogs.nature.com/naturejobs/2017/04/17/techblog-my-digital-toolbox-lorena-barba/

Lorena A. **Barba group**

Reproducibility PI Manifesto

2012

http://lorenabarba.com/gallery/reproducibility-pi-manifesto/

Reproducibility **PI Manifesto** (2012)

- I teach my graduate students about reproducibility
- All our research code (and writing) is under version control
- We always carry out verification & validation (and make them public)
- For main results, we share data, plotting script & figure under CC-BY
- We upload preprint to arXiv at the time of submission to a journal
- We release code at the time of submission of a paper to a journal
- We add a "Reproducibility" declaration at the end of each paper
- I develop a consistent open-science policy & keep an up-to-date web presence

Onboarding

OPINION

Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] **and Roger D. Peng**[b]

[a]*Associate Professor of Biostatistics and Oncology and* [b]*Associate Professor of Biostatistics, Johns Hopkins University, Baltimore, MD*

PNAS | **February 10, 2015** | vol. 112 | no. 6 | **1645–1646**

http://dx.doi.org/10.1073/pnas.1421412111

"The key is prevention via the training of more people on techniques for data analysis and reproducible research."

Leek & Peng, PNAS 2015

Lorena A Barba
Engineering professor, computational scientist, jazz buff, techie, mac fan, academic writer and...
Oct 30, 2016 · 10 min read

Lockheed P-80A airplane (1946). Credit: NASA Commons. — A reminder to test your code.

Barba-group reproducibility syllabus

A syllabus for research computing

1. command line utilities in Unix/Linux
2. an open-source scientific software ecosystem (our favorite is Python's)
3. software version control (we like the distributed kind: our favorite is git / GitHub)
4. good practices for scientific software development: code hygiene and testing
5. knowledge of licensing options for sharing software

`https://barbagroup.github.io/essential_skills_RRC/`

“private reproducibility”

…we can rebuild our own past research results from the precise version of the code that was used to create them.

WORKING LIFE

By **Lorena A. Barba**

The hard road to reproducibility

Early in my Ph.D. studies, my supervisor assigned me the task of running computer code written by a previous student who was graduated and gone. It was hell.

"My students and I continuously discuss and perfect our standards."

http://science.sciencemag.org/content/354/6308/142

Lorena A Barba
Engineering professor, computational scientist, jazz buff, techie, mac fan, academic writer and font …
Feb 15 · 6 min read

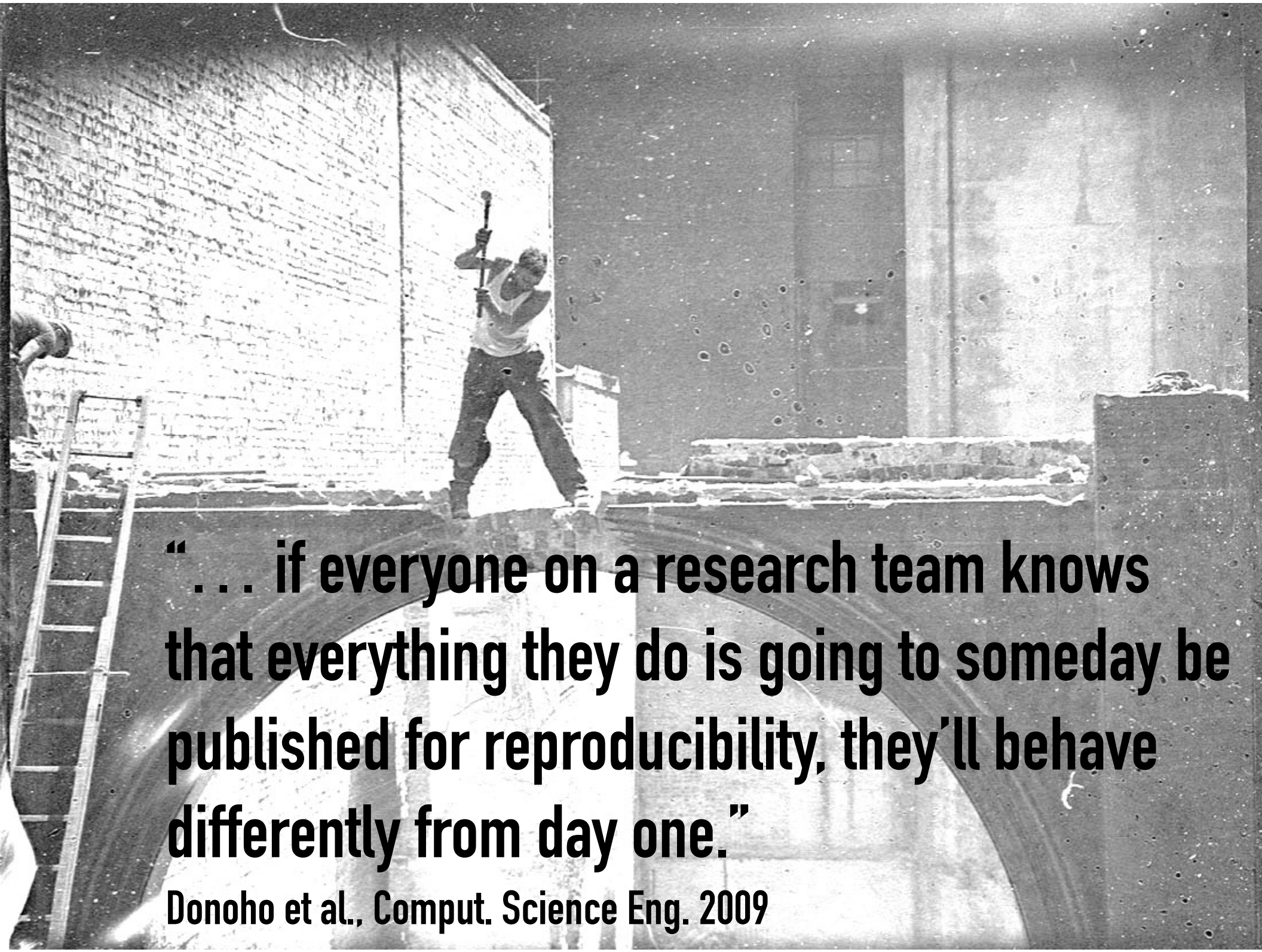
Blue Mountain supercomputer at Los Alamos National Laboratory, decommissioned 2004. (Public domain)

Why should I believe your supercomputing research?

In parallel, even two runs with identical input data can differ!

Different versions of your code, external libraries, even compilers may change results.

In HPC, peers may not be able to reproduce, but they will trust the results more if built over a consistent *practice* of reproducible research.


“. . . if everyone on a research team knows that everything they do is going to someday be published for reproducibility, they’ll behave differently from day one.”
Donoho et al., Comput. Science Eng. 2009