Reproducibility in Computational Science

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

Stanford University | Stats285 Guest Lecture | October 16, 2017

Agenda

1. Unpacking Reproducibility
2. The Research Ecosystem
3. Policy and Progress
4. Intellectual Property and Openness

Unpacking Reproducibility

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.

Originality: scientific claims contribute something new

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

Skepticism: Boyle's ideas

Skepticism requires that the claim can be independently verified,

This in turn requires transparency in the communication of the research process.

Instantiated by Robert Boyle and the Transactions of the Royal Society in the 1660's.

ROBERT BOYLE,

Today: Technology is driving a re-assessment of transparency

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.

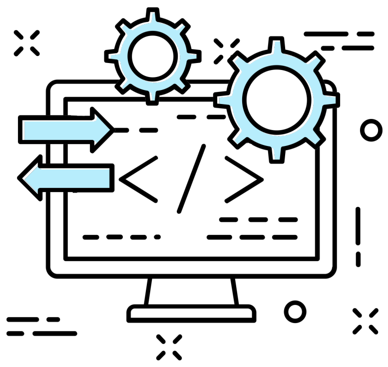
The software contains "ideas that enable biology…"

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

"Stories from the Supplement" from the Genome Informatics meeting 11/1/2013

https://youtu.be/5NiFibnbE8o

The digital age in science



Claim 1:

Virtually all published discoveries today have a computational component.

Claim 2:

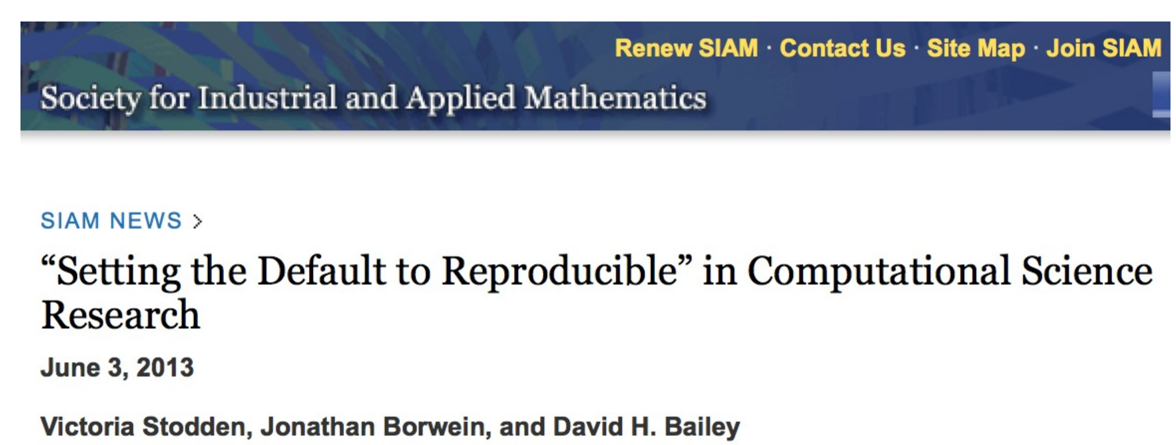
There is a mismatch between the traditional scientific process and computation, leading to reproducibility concerns.

Parsing Reproducibility

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”



V. Stodden, IMS Bulletin (2013)

Empirical Reproducibility

Cell Reports
Commentary

Sorting Out the FACS: A Devil in the Details

William C. Hines,[1,5,*] Ying Su,[2,3,4,5,*] Irene Kuhn,[1] Kornelia Polyak,[2,3,4,5] and Mina J. Bissell[1,5]
[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[3]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[4]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[5]These authors contributed equally to this work
*Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.celrep.2014.02.021

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and specimens, our two laboratories quite reproducibly were unable to replicate each other's fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of measuring up to—and beyond—an astonishing 50 individual markers per cell (Cheung and Utz, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.'s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.'s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for

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

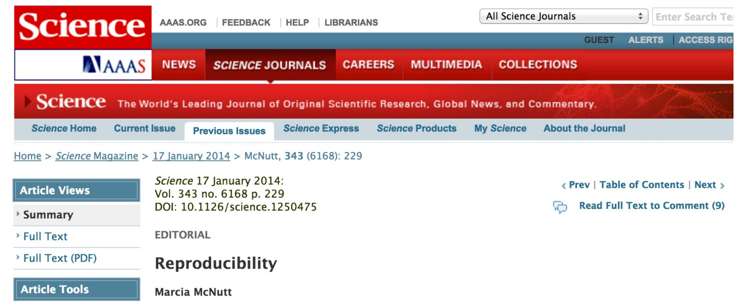
Past Events
September 3-4, 2014
Transportation of Laboratory Animals
• *Presentations and videos online*

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,
- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,
- inappropriate tests or models, model misspecification,
- Model robustness to parameter changes and data perturbations,
- Investigator bias toward previous findings; conflicts of interest.
- …

Statistical Reproducibility



In January 2014 Science enacted new manuscript submission requirements:

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

Also added statisticians to the Board of Reviewing Editors.

Computational Reproducibility

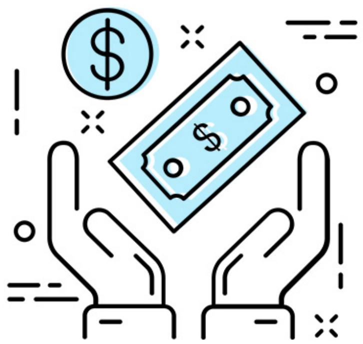
An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.

David Donoho, 1998 http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf

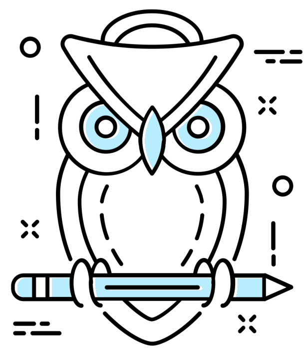
“Really Reproducible Research” (1992) inspired by Stanford Professor Jon Claerbout

2. The Research Ecosystem

Ecosystem



Funders
(policy)

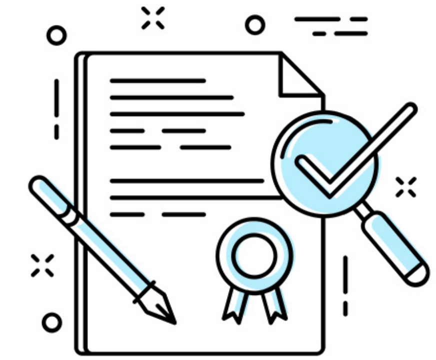


Scientific Societies

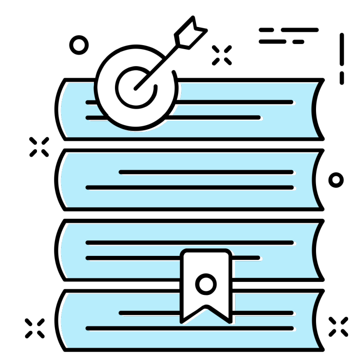
Regulatory Bodies
(OSTP Memos)

Researchers
(processes)

Universities/ institutions
(hiring/promotion)



Publishers
(TOP guidelines)



Universities/ libraries
(empowering w/tools)

Researchers

Data / Code Sharing Practices

Survey of the NIPS community:

- 1,758 NIPS registrants up to and including 2008,
- 1,008 registrants when restricted to .edu registration emails,
- After piloting, the final survey was sent to 638 registrants,

Sharing Incentives

Code		Data
91%	Encourage scientific advancement	81%
90%	Encourage sharing in others	79%
86%	Be a good community member	79%
82%	Set a standard for the field	76%
85%	Improve the calibre of research	74%
81%	Get others to work on the problem	79%
85%	Increase in publicity	73%
78%	Opportunity for feedback	71%
71%	Finding collaborators	71%

Survey of the Machine Learning Community, NIPS (Stodden 2010)

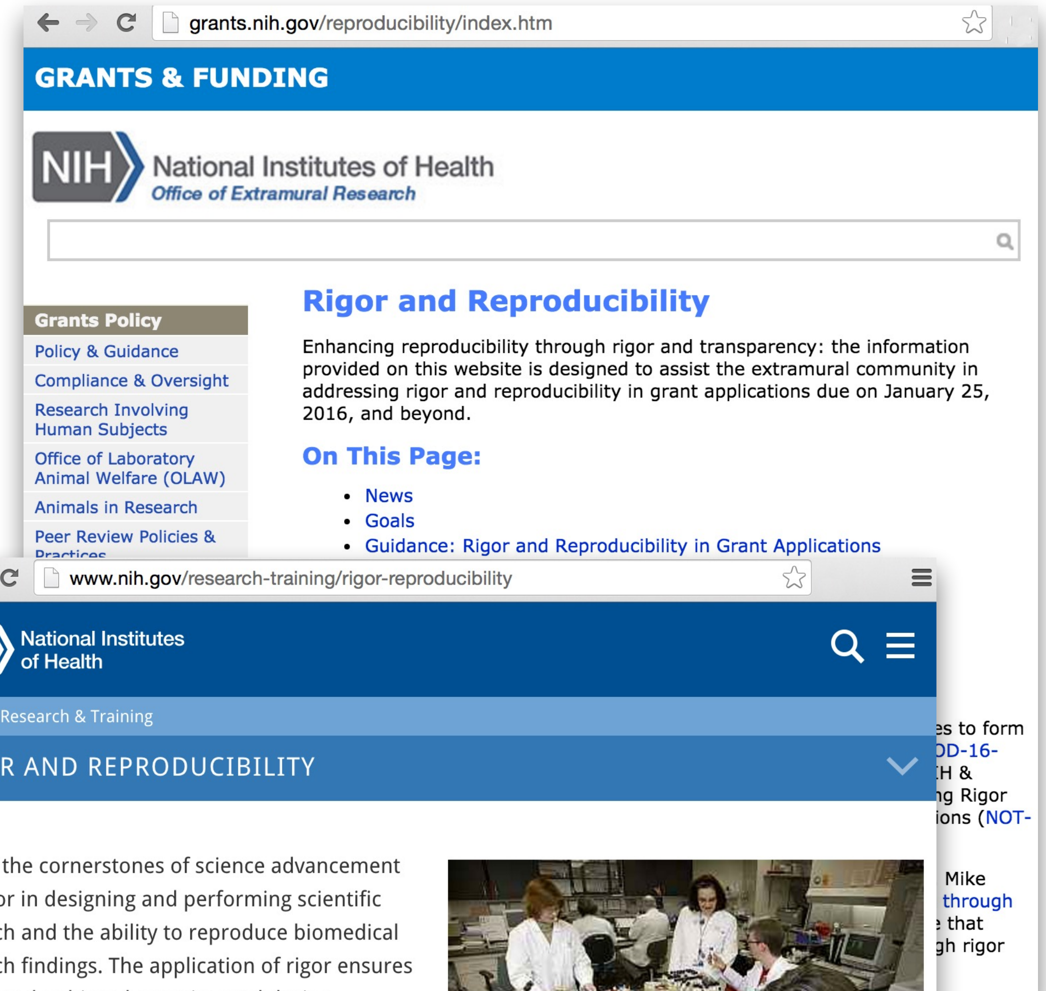
Barriers to Sharing

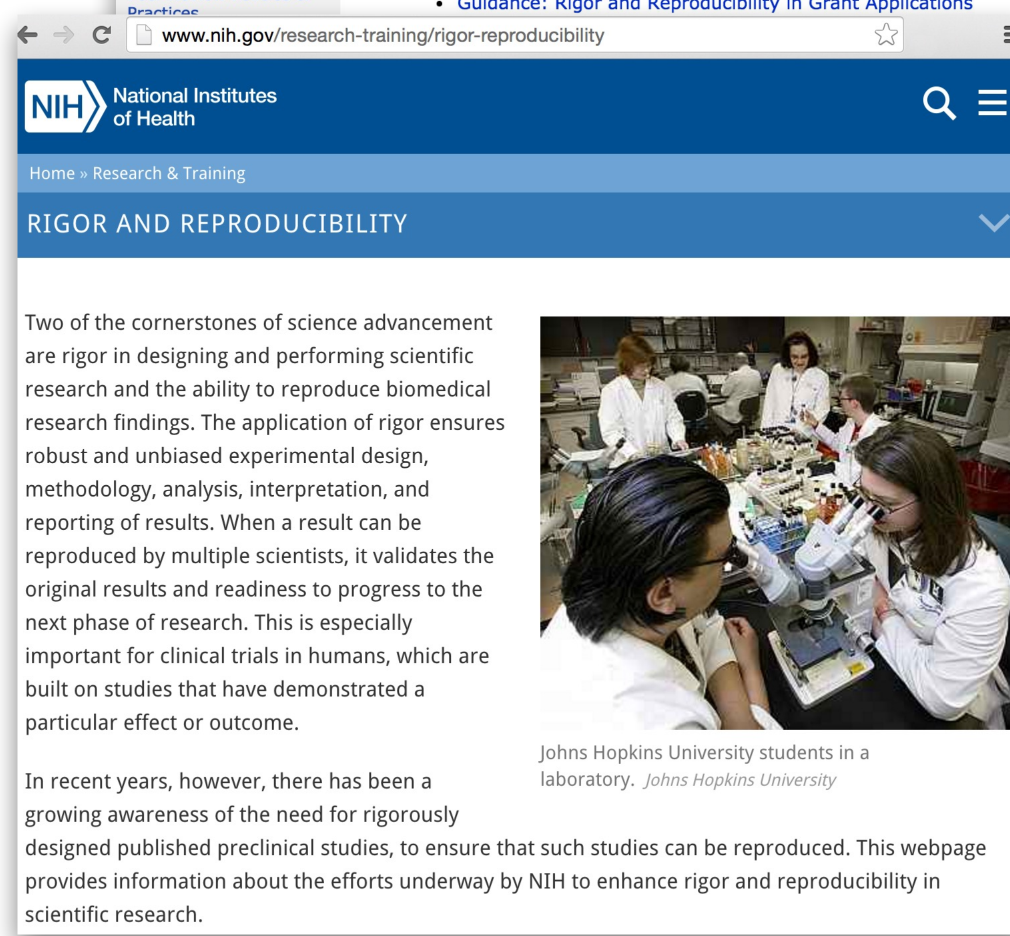
Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Survey of the Machine Learning Community, NIPS (Stodden 2010)

Funding Bodies

Federal Agencies





www.mrsec.harvard.edu/2017NSFReliability/

Charge

Schedule

Attendees

Venue

Travel

Lodging

Restaurants

Resources

Lexicon

Report

Support

Contact Us

NSF Workshop
Systematic Approach to Robustness, Reliability, and Reproducibility in Scientific Research

February 25 - 26, 2017

Beckman Center of the National Academies of Sciences & Engineering
University of California at Irvine
100 Academy Way
Irvine, CA 92617
(949) 721-2200

The federal investment in scientific and engineering research drives innovation across our society; it also provides a foundation for national competitiveness, prosperity, and sound public policy. Recently, several prominent studies have highlighted a significant proportion of research reports, in certain fields, that are not reproducible. There is growing concern within the scientific enterprise and a loss of public trust in the reliability of science, especially the results of basic research funded by the taxpayer, is a serious issue.

The Administration, through OMB and OSTP, has directed that funding agencies, including the NSF, address these problems of irreproducibility, which includes cases where the data generated by publicly-funded research is not accessible. As part of its response to this mandate, the NSF is supporting the scientific community in efforts to find the root causes of these problems, and through extensive discussions identify ways in which they can best be solved.

Principal Investigator
David A. Weitz (Harvard University)

Workshop Leaders
Andrea Liu (University of Pennsylvania)
Wallace Marshall (UC San Francisco)
Roger D. Peng (Johns Hopkins University)
Victoria Stodden (University of Illinois)

Workshop Participants
Keith Baggerly (UTexas/MD Anderson)
Paul Chaikin (New York University)
George Fuller (UC San Diego)
Carol Hall (North Carolina State University)
Robert Hanisch (ODI, NIST)
Leslie Hatton (University of Kingston)
Amy E. Herr (UC Berkeley)
Mike Hildreth (Notre Dame)
Daniel S. Katz (University of Illinois)
Gareth H. McKinley (MIT)
Peter J. Mohr (NIST)
Jose Onuchic (Rice University)
Manish Pararashar (Rutgers University)
Steven Vigdor (Indiana University)
George Whitesides (Harvard University)
William Allen Zajc (Columbia University)

Agency Contacts
Bogdan Mihaila (NSF, Mathematical and Physical Sciences)
Gregory W. Warr (NSF, Molecular and Cellular Biosciences)

Journals and Publishers

Journal Data and Code Sharing Policies

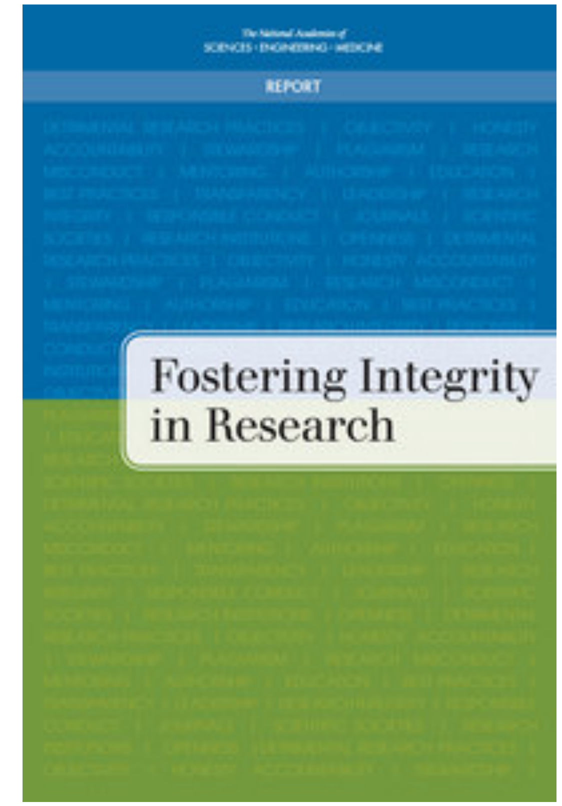
	Data 2011	Data 2012	Code 2011	Code 2012
Required as condition of publication, barring exceptions	10.6%	11.2%	3.5%	3.5%
Required but may not affect editorial decisions	1.7%	5.9%	3.5%	3.5%
Encouraged/addressed, may be reviewed and/or hosted	20.6%	17.6%	10%	12.4%
Implied	0%	2.9%	0%	1.8%
No mention	67.1%	62.4%	82.9%	78.8%

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)

3. Policy and Progress

“Fostering Integrity in Research”

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



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

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

REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

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

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transpar-

en

Cu

inc

no

Pri

len

me

pro

ness Promotion (TOP) guidelines (*1*) and recommendations for field data (*2*), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers repre-

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

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

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

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

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

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

Reproducibility Enhancement Principles

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

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

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

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

Reproducibility Enhancement Principles

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

4. Intellectual Property

Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) via the 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)

Argument: both types of intellectual property are an imperfect fit with scholarly norms, and require action from the research community to enable re-use, verification, reproducibility, and support the acceleration of scientific discovery.

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.

Licensing in Research
Background: Open Source Software

Innovation: Open Licensing

- ➡ Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

Hundreds of open source software licenses:

- GNU Public License (GPL)
- (Modified) BSD License
- MIT License
- Apache 2.0 License
- … see http://www.opensource.org/licenses/alphabetical

Creative Commons

- Founded in 2001, by Stanford Law Professor Larry Lessig, MIT EECS Professor Hal Abelson, and advocate Eric Eldred.
- Adapts the Open Source Software approach to artistic and creative digital works.

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 Modified BSD or similar,
- Release data to public domain or attach attribution license.
    - ➡ Remove copyright's barrier to reproducible research and,
    - ➡ Realign the IP framework with longstanding scientific norms.

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 PubIns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
- Copyright adheres to raw facts in Europe.
- Residual copyright in data is possible (attribution licensing or public domain certification).
- Legal mismatch: What constitutes a "raw" fact anyway?

Privacy and Data

- HIPAA, FERPA, IRB mandates create legally binding restrictions on the sharing human subjects data
  (see e.g. http://www.dataprivacybook.org/)
- Potential privacy/proprietary implications for industry generated data.
- Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..

Data and Code Ownership: What Defines Contribution?

- Issue for producers: credit and citation.
- What is the role of peer-review?
- Repositories adding meta-data and discoverability make a contribution.
- Velocity of data and code contributions/updates
- Future coders may contribute in part to new software, other software components may already be in the scholarly record. Attribution vs sharealike.
  - ➡ (at least) 2 aspects: legal ownership vs scholarly credit.
- Rethinking plagiarism for software contributions.

In a World of Radical Transparency..

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 year 2003 and list the trial name and histogram side by side.

Courtesy Donoho and Gavish 2012