Reproducibility in Computational Science

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

METRICS Conference
Stanford University
Nov 19, 2015
Agenda:
Sources of Irreproducibility and Solution Steps

1. Empirical Reproducibility

2. Computational Reproducibility:
   - Tools and Cyberinfrastructure
   - Community and Policy Responses

3. Statistical Reproducibility
Credibility Crisis

Los Angeles Times

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren’t even true.

Science

Reproducibility

Marcia McNutt

Science advances on a foundation of trusted approach that scientists use to gain confidence community was shaken by reports that a result not reproducible. Because confidence in results community, we are announcing new initiatives Science. For preclinical studies (one of the rare recommendations of the U.S. National Institute increasing transparency. Authors will indicate handling (such as how to deal with outliers), will ensure a sufficient signal-to-noise ratio, whether experimenter was blind to the conduct of the guidelines.

The Scientist

NIH Tackles Irreproducibility

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

By Jef Akst | January 28, 2014

The Economist

HOW SCIENCE GOES WRONG

The meaning of Sachin Tendulkar

nature

Announcement: Reducing our irreproducibility

Over the past year, Nature has published a string of articles that highlight the reliability and reproducibility of published research (collected at https://www.nature.com/reproducibility). This is an unprecedented commitment to transparency and accountability in our field.

Nature 483, 609 (29 March 2012) | doi:10.1038/483609a
Published online 28 March 2012

Nature 483, 609 (29 March 2012) | doi:10.1038/483609a
Published online 28 March 2012

Nature 483, 609 (29 March 2012) | doi:10.1038/483609a
Published online 28 March 2012
Unpacking “Reproducibility”

“Empirical Reproducibility”

“Computational Reproducibility”

“Statistical Reproducibility”

V. Stodden, IMS Bulletin (2013)
1. Empirical Reproducibility

Sorting Out the FACS: A Devil in the Details

William C. Hines,1,3,5* Ying Su,3,4,5* Irene Kuhn,1 Kornelia Polyak,2,3,4,5* and Mina J. Bissell1,5

1Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977225A, 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.cellrep.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 characterization, 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 Uz, 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 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.
Traditionally two branches of the scientific method:

Branch 1 (deductive): mathematics, formal logic,

Branch 2 (empirical): statistical analysis of controlled experiments.

Many claim the emergence of new branches:

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

“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.”

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.
The Impact of Technology

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...”
Stories from the Supplement, 2013.
The Ubiquity of Error

The central motivation for the scientific method is to root out error:

• Deductive branch: the well-defined concept of the proof,

• Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation presents only a potential third/fourth branch of the scientific method (Donoho, Stodden, 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

• Reproducing the computational steps vs replicating the experiments independently including data collection and software implementation.
Reproducibility in Computational and Experimental Mathematics (December 10-14, 2012)

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

Setting the Default to Reproducible

Reproducibility in Computational and Experimental Mathematics

Developed collaboratively by the ICERM workshop participants

Compiled and edited by the Organizers

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein

Abstract

Science is built upon foundations of theory and experiment validated and improved through open, transparent communication. With the increasingly central role of computation in scientific discovery this means communicating all details of the computations needed for others to replicate the experiment, i.e., making available to others the associated data and code. The "reproducible research" movement recognizes that traditional scientific research and publication practices now fall short of this ideal, and encourages all those involved in the production of computational science — scientists who use computational methods and the institutions that employ them, journals and dissemination mechanisms, and funding agencies — to facilitate and practice really reproducible research.
reproducibility@XSEDE: An XSEDE14 Workshop

Overview

The reproducibility@XSEDE workshop is a full-day event scheduled for Monday, July 14, 2014 in Atlanta, GA. The workshop will take place in conjunction with XSEDE14 (conferences.xsede.org), the annual conference of the Extreme Science and Engineering Discovery Environment (XSEDE), and will feature an interactive, open-ended, discussion-oriented agenda focused on reproducibility in large-scale computational science. Consistent with the overall XSEDE14 conference theme, we seek to engage participants from a broad range of backgrounds, including practitioners whose computational interests extend beyond traditional modeling and simulation as well as decision-makers and other professionals whose work informs and determines the direction of computation-enabled research. We hope to help
Supporting Computational Science

• Dissemination Platforms:
  - ResearchCompendia.org
  - IPOL
  - Madagascar
  - MLOSS.org
  - thedatahub.org
  - nanoHUB.org
  - Open Science Framework
  - RunMyCode.org

• Workflow Tracking and Research Environments:
  - VisTrails
  - Kepler
  - CDE
  - Chameleon
  - Galaxy
  - GenePattern
  - Jupyter / IPython Notebook
  - Flywheel
  - Sumatra
  - Taverna
  - Pegasus

• Embedded Publishing:
  - Verifiable Computational Research
  - SOLE
  - knitR
  - Collage Authoring Environment
  - SHARE
  - Sweave
Research Compendia

Goal: improve understanding of reproducible computational science, trace sources of error.

- link data/code to published claims,
- enable re-use,
- sharing guide for researchers,
- certification of results,
- large scale validation of findings,
- stability, sensitivity checks.
Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items

Mark Dingemanse, Francisco Torreira, N. J. Enfield, Johan J. Bolhuis

Code and Data Abstract

A word like Huh?—used as a repair initiator when, for example, one has not clearly heard what someone just said—is found in roughly the same form and function in spoken languages across the globe. We investigate it in naturally occurring conversations in ten languages and present evidence and arguments for two distinct claims: that Huh? is universal, and that it is a word. In support of the first, we show that the similarities in form and function of this interjection across languages are much greater than expected by chance. In support of the second claim we show that it is a lexical, conventionalised form that has to be learnt, unlike grunts or emotional cries. We discuss possible reasons for the cross-linguistic similarity and propose an account in terms of convergent evolution. Huh? is a universal word not because it is innate but because it is shaped by selective pressures in an interactional environment that all languages share: that of other-initiated repair. Our proposal enhances evolutionary models of language change by suggesting that conversational infrastructure can drive the convergent cultural evolution of linguistic items.

Compendium Type:: article
Content License:: CC0
Code License:: MIT
Random survival forests for high-dimensional data
Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, Andy J. Minn

Code and Data Abstract
Minimal depth is a dimensionless order statistic that measures the predictiveness of a variable in a survival tree. It can be used to select variables in high-dimensional problems using Random Survival Forests (RSF), a new extension of Breiman's Random Forests (RF) to survival settings. We review this methodology and demonstrate its use in high-dimensional survival problems using a public domain R-language package randomSurvivalForest. We discuss effective ways to regularize forests and discuss how to properly tune the RF parameters 'nodesize' and 'mtry'. We also introduce new graphical ways of using minimal depth for exploring variable relationships.


Code DOI: doi:10.7938/M1H41PBB.
Data DOI: doi:10.7938/M1CC0XMM.

Compendium Type: Journal or Magazine Articles
Primary Research Field: Computer and Information Sciences
Secondary Research Field: Mathematics
Content License: Public Domain Mark
Code License: MIT License

Verification

Recent Verifications
Verification run 41 created March 21, 2014, 9:43 a.m.
Verification run 40 created March 21, 2014, 4:19 a.m.
A proof of concept for a research compendia webapp http://researchcompendia.org — Edit

542 commits  12 branches  29 releases  1 contributor

branch: develop

Merge branch 'Release/1.0.1-b9' into develop

codersquid authored 30 minutes ago

- companionpages  bump revision  30 minutes ago
- docs  removes instructions for envdir and bootstrap.sh, adds instructions f...  10 days ago
- requirements  citation dialog and display for journals  13 days ago
- .gitignore  adds vagrant and bootstrap starter  2 months ago
- .travis.yml  fixed broken doi service test and updated irc channel for travis  4 months ago
- AUTHORS.rst  renaming project from tyler to researchcompendia  2 months ago
- CITATION.bib  bump revision  30 minutes ago
- CONTRIBUTING.rst  fixed thinko of 'comment' to 'commit'  3 days ago
- HISTORY.rst  bump revision  30 minutes ago
- LICENSE  release 1.0.0-alpha1  4 months ago
- MANIFEST.in  making skeleton docs  5 months ago
The MIT License (MIT)

Copyright (c) 2013 Sheila Miguez, Victoria Stodden, Jennifer Seiler

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.
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,
- 37 bounces, 5 away, and 3 in industry, gave a final response rate was 134 of 593 or 23%.
- Queried about reasons for sharing or not sharing data/code associated with their NIPS paper.
# Sharing Incentives

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>91%</td>
<td>Encourage scientific advancement</td>
</tr>
<tr>
<td>90%</td>
<td>Encourage sharing in others</td>
</tr>
<tr>
<td>86%</td>
<td>Be a good community member</td>
</tr>
<tr>
<td>82%</td>
<td>Set a standard for the field</td>
</tr>
<tr>
<td>85%</td>
<td>Improve the calibre of research</td>
</tr>
<tr>
<td>81%</td>
<td>Get others to work on the problem</td>
</tr>
<tr>
<td>85%</td>
<td>Increase in publicity</td>
</tr>
<tr>
<td>78%</td>
<td>Opportunity for feedback</td>
</tr>
<tr>
<td>71%</td>
<td>Finding collaborators</td>
</tr>
</tbody>
</table>

Survey of the Machine Learning Community, NIPS (Stodden 2010)
## Barriers to Sharing

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>Time to document and clean up</td>
</tr>
<tr>
<td>52%</td>
<td>Dealing with questions from users</td>
</tr>
<tr>
<td>44%</td>
<td>Not receiving attribution</td>
</tr>
<tr>
<td>40%</td>
<td>Possibility of patents</td>
</tr>
<tr>
<td>34%</td>
<td>Legal Barriers (ie. copyright)</td>
</tr>
<tr>
<td>-</td>
<td>Time to verify release with admin</td>
</tr>
<tr>
<td>30%</td>
<td>Potential loss of future publications</td>
</tr>
<tr>
<td>30%</td>
<td>Competitors may get an advantage</td>
</tr>
<tr>
<td>20%</td>
<td>Web/disk space limitations</td>
</tr>
</tbody>
</table>

Survey of the Machine Learning Community, NIPS (Stodden 2010)
Editorial: ACM TOMS Replicated Computational Results Initiative

MICHAEL A. HEROUX, Sandia National Laboratories

The scientific community relies on the peer review process for assuring the quality of published material, the goal of which is to build a body of work we can trust. Computational journals such as the ACM Transactions on Mathematical Software (TOMS) use this process for rigorously promoting the clarity and completeness of content, and citation of prior work. At the same time, it is unusual to independently confirm computational results.

ACM TOMS has established a Replicated Computational Results (RCR) review process as part of the manuscript peer review process. The purpose is to provide independent confirmation that results contained in a manuscript are replicable. Successful completion of the RCR process awards a manuscript with the Replicated Computational Results Designation.

This issue of ACM TOMS contains the first [Van Zee and van de Geijn 2015] of what we anticipate to be a growing number of articles to receive the RCR designation, and the related RCR reviewer report [Willenbring 2015]. We hope that the TOMS RCR process will serve as a model for other publications and increase the confidence in and value of computational results in TOMS articles.

- provide independent confirmation that results contained in a manuscript are correct and replicated.

- Successful completion of the RCR process gives the manuscript a Replicated Computational Results designation on the first page of the published article.
3. Statistical Reproducibility

- False discovery, chasing significance, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.

- Low power, poor experimental design,

- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,

- Poor statistical methods (nonrandom sampling, inappropriate tests or models, model misspecification..)

- Model robustness to parameter changes and data perturbations,

- Investigator bias toward previous findings; conflicts of interest.
Journal Requirements

In January 2014 Science enacted new policies that will check for:

1. a “data-handling plan” i.e. how outliers will be dealt with,

2. sample size estimation for effect size,

3. whether samples are treated randomly,

4. whether experimenter blind to the conduct of the experiment.

Statisticians added to the Board of Reviewing Editors.
Responses to Irreproducibility

- **Failure of traditional reporting methods** vs adaptation of standards to accommodate changes in the research process.

- Conjecture: requiring power calculations and negative results disclosure would resolve the majority of irreproducibility in empirical data-driven research (with code/data access).

- **Benchmarking and testing**: either nonexistent or over reliance on inappropriate benchmarks (see e.g. [http://www.in-cites.com/scientists/DrDavidDonoho.html](http://www.in-cites.com/scientists/DrDavidDonoho.html))

- **Collective action problem**: coordination of researcher incentives, universities, funding agencies, journals, scientific societies, legal and policy, libraries, the public.
Some Dream Applications

• 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 of Donoho and Gavish 2012
Self-correction in science at work

Jay Bruce Alberts, Ralph C. Cicerone,* Stephen E. Fienberg, Alexander Kamban, Muzella McKnight, Robert M. Nemenyi, Randy Schekman, Richard Shiffrin, Victoria Stolinski, Steven Suarez, Mark T. Zuber, Barbara Klimes-Popa, Kathleen Hall Jamieson*8

Weckerly quoted in work, newsmakers carry word of new scientific discover,ies, but the media sometimes give impetus to the idea that science is equal to play with substantive discoveries. Careful of what is known are lost in categorical headlines. Rare instances of misconduct or instances of irresponsibility are transmuted into concerns that science is broken. The October 2013 Economist headlines proclaimed "Trouble at the lab: Scientists like to think of science as self-correcting. To an alarming degree, it is not." Yet, that article is also rich with instances of science both policing itself, which is how the problems came to The Economist's attention in the first place, and addressing discovered lapses and irreproducibility concerns. In light of such issues and efforts, the U.S. National Academy of Sciences (NAS) and the Ammerberg Retreat at Hungary convened our group to examine ways to remove some of the current disincentives to high standards of integrity in science.

Like all human endeavors, science is imperfect. However, if Robert Morss wrote more than half a century ago "the activities of scientists are subject to rigorous policing, to a degree perhaps unparalleled in any other field of activity," 2 as a result, as Popper argued, "Science is one of the very few human activities—which perhaps the only one—of which every error is systematically criticized and fairly often, in time, corrected." 3 Instances in which scientists detect and address flaws in work constitute evidence of success, not failure, because they demonstrate the underlying protective mechanisms of science at work. Still, as in any human venture, science with large does not always live up to its ideals. Although attempts to replicate the 1998 Wakefield study alleging an association between autism and the MMR vaccine,
One-year postdoctoral fellowship at UIUC in reproducibility

One Postdoctoral Researcher Position within the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign, full-time, 100% (1-year duration, with possibility of extension, contingent upon performance and funding)

Summary: This postdoctoral appointment provides an opportunity to help develop a new area of transdisciplinary research in reproducibility in computational and data-intensive research. Researchers, journals, scientific societies and funding agencies are requiring the reproducibility of computational findings, which introduces new aspects to research including understanding factors that drive or prevent reproducibility, and enabling factors such as software and data access, interoperability, provenance, workflow capture, discoverability, and policy development. The postdoc will develop research on all these fronts including a focus on high end computing and cyberinfrastructure.

The position is funded by the Computing and Data Sciences Theme at NCSA, and the successful applicant will be working primarily with Dr. Victoria Stodden. The successful candidate will use one of the most advanced supercomputers for open scientific research, Blue Waters, and also work with a team of leading computational and data scientists at NCSA.
References

Implementing Reproducible Research, CRC Press 2014.


available at http://www.stodden.net
Three sections:

1. Tools
2. Practices and Guidelines
3. Platforms

Chapters available for download: https://osf.io/s9tya/wiki/
Data and Privacy

Question: can we find methods that permit access to data with confidentiality concerns?

0 = no access; 1 = complete access

What does .5 look like?

“Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency” (chapter 5)
Journal Policy?

• Journal Policy setting study design:
  
  • Select all journals from ISI classifications “Statistics & Probability,” “Mathematical & Computational Biology,” and “Multidisciplinary Sciences” (this includes Science and Nature).
  
  • N = 170, after deleting journals that have ceased publication.
  
  • Create dataset with ISI information (impact factor, citations, publisher) and supplement with publication policies as listed on journal websites, in June 2011 and June 2012.
## Journal Data Sharing Policy

<table>
<thead>
<tr>
<th>Policy Description</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required as condition of publication, barring exceptions</td>
<td>10.6%</td>
<td>11.2%</td>
</tr>
<tr>
<td>Required but may not affect editorial decisions</td>
<td>1.7%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Encouraged/addressed, may be reviewed and/or hosted</td>
<td>20.6%</td>
<td>17.6%</td>
</tr>
<tr>
<td>Implied</td>
<td>0%</td>
<td>2.9%</td>
</tr>
<tr>
<td>No mention</td>
<td>67.1%</td>
<td>62.4%</td>
</tr>
</tbody>
</table>

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)
## Journal Code Sharing Policy

<table>
<thead>
<tr>
<th>Requirement</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required as condition of publication, barring exceptions</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Required but may not affect editorial decisions</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Encouraged/addressed, may be reviewed and/or hosted</td>
<td>10%</td>
<td>12.4%</td>
</tr>
<tr>
<td>Implied</td>
<td>0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>No mention</td>
<td>82.9%</td>
<td>78.8%</td>
</tr>
</tbody>
</table>

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

- Changemakers are journals with high impact factors.
- Progressive policies are not widespread, but being adopted rapidly.
- Close relationship between the existence of a supplemental materials policy and a data policy.
- No statistically significant relationship between data and code policies and open access policy.
- Data and supplemental material policies appear to lead software policy.
Legal Barriers: Copyright

“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)

- 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

Exceptions and Limitations: Fair Use.
Responses Outside the Sciences 1: Open Source Software

- 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
Responses Outside the Sciences 2: 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.
Response from Within the Sciences

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.

Winner of the Access to Knowledge Kaltura Award 2008
Copyright and Data

• Copyright adheres to raw facts in Europe.

• 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)).

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

• Law doesn’t match reality on the ground: What constitutes a “raw” fact anyway?
Bayh-Dole Act (1980)

Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices),

Bayh-Dole Act 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.

Greatest impact in biomedical research collaborations and drug discovery. Now, software patents also impact science.
Ownership of Research Codes

Patent and Copyright Agreement for Personnel at Stanford - SU18

I understand that, consistent with applicable laws and regulations, Stanford University is governed in the handling of intellectual property by its official policies titled Inventions, Patents and Licensing and Copyright Policy (both published in the Research Policy Handbook), and I agree to abide by the terms and conditions of those policies, as they may be amended from time to time.

Pursuant to those policies, and in consideration of my employment by Stanford, the receipt of remuneration from Stanford, participation in projects administered by Stanford, access to or use of facilities or resources provided by Stanford and/or other valuable consideration, I hereby agree as follows:

1. I will disclose to Stanford all potentially patentable inventions conceived or first reduced to practice in whole or in part in the course of my University responsibilities or with more than incidental use of University resources. I hereby assign to Stanford all my right, title and interest in such patentable inventions and to execute and deliver all documents and do any and all things necessary and proper on my part to effect such assignment. (See Inventions, Patents and Licensing for further clarification and discussion related to this paragraph.)

2. I am free to place my inventions in the public domain as long as in so doing neither I nor Stanford violates the terms of any agreements that governed the work done.

3. Stanford policy states that all rights in copyright shall remain with the creator unless the work:
   a. is a work-for-hire (and copyright therefore vests in the University under copyright law),
   b. is supported by a direct allocation of funds through the University for the pursuit of a specific project,
   c. is commissioned by the University,
   d. makes significant use of University resources or personnel, or
   e. is otherwise subject to contractual obligations.

I hereby assign or confirm in writing to Stanford all my right, title and interest, including associated copyright, in and to copyrightable materials falling under a) through e), above.

4. I am now under no consulting or other obligations to any third person, organization or corporation in respect to rights in inventions or copyrightable materials which are, or could be reasonably construed to be, in conflict with this agreement.

NOTE: An alternative to this agreement may be appropriate for personnel with a prior existing and conflicting employment agreement that establishes a right to intellectual property in conflict with Stanford policies. Personnel in this situation should contact the Office of the Vice Provost and Dean of Research.

5. I will not enter into any agreement creating copyright or patent obligations in conflict with this agreement.

6. This agreement is effective on the later of July 1, 2011 (on the one hand) or my date of hire, enrollment, or participation in projects administered by Stanford (on the other hand), and is binding on me, my estate, heirs and assigns.

Electronic Signature in AXESS
http://axess.stanford.edu

The signer should make a copy of this agreement for his or her own records, and hereby waives any objection to Stanford’s use of an electronic version of this agreement as a substitute for the original for any legally recognized purpose.

July 2011

Provider: Office of the Vice Provost and Dean of Research, Stanford University
Contact: Assistant Dean of Research
Last updated: July 2011
Disclosure of Research Codes

Claim: Codes would (eventually) be fully open in the absence of Bayh-Dole:

- Grassroots “Reproducible Research” movement in computational science (policy development, best practices, tool development),
- Changes in funding agency and journal publication requirements.

Other legal barriers:

- HIPAA (Health Information Portability and Accountability Act) and privacy regulations,
- Collaboration agreements with industry,
- Hiring agreements, institutional rules,
- National security.
Best Practices for Scientific Computing


Published: January 07, 2014 • DOI: 10.1371/journal.pbio.1001745

Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research

Victoria Stodden
Columbia University - Department of Statistics

Sheila Miguez
Columbia University

September 6, 2013

Abstract:
Scholarly dissemination and communication standards are changing to reflect the increasingly computational nature of scholarly research, primarily to include the sharing of the data and code associated with published results. This paper presents a formalized set of best practice recommendations for computational scientists wishing to disseminate reproducible research, facilitate innovation by enabling data and code re-use, and enable broader communication of the output of digital scientific research. We distinguish two forms of collaboration to motivate choices of software environment for computational scientific research. We also present these Best Practices as a living, evolving, and changing document on wiki.