Reproducibility and Reliability in Statistical and Data Driven Research

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

Department of Statistics Seminar
Carnegie Mellon University
Nov 9, 2015
Agenda

1. Motivating Reproducibility: What’s New?
   • Big Data and Big Computation
   • My Thesis Approach
2. Sources of Error: Statistical, Computational, Empirical
3. Modeling Policy Responses
4. Software as Method: A Statistics Issue
5. Future Directions
Scoping the Issue

<table>
<thead>
<tr>
<th>JASA June</th>
<th>Computational Articles</th>
<th>Code Publicly Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>9 of 20</td>
<td>0%</td>
</tr>
<tr>
<td>2006</td>
<td>33 of 35</td>
<td>9%</td>
</tr>
<tr>
<td>2009</td>
<td>32 of 32</td>
<td>16%</td>
</tr>
<tr>
<td>2011</td>
<td>29 of 29</td>
<td>21%</td>
</tr>
</tbody>
</table>

Ioannidis (2011): of 500 papers studied, 9% had full primary raw data deposited.

Stodden (to come): estimates that the computations in 27% of scientific articles published in *Science* today are reproducible.
The Scientific Method

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.”
The Impact of Technology

1. *Big Data / Data Driven Discovery*: high dimensional data, $p \gg n$; divorce of data generation from data analysis

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.
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.

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

Nature

Announcement: Reducing our irreproducibility

Over the past year, Nature has published a string of articles that have highlighted the reliability and reproducibility of published research (collected as a special theme guide on the Nature website).

The Economist

How science goes wrong

Washington’s lawyer surplus

How to do a nuclear deal with Iran

Investment dips from Nobel economists

Hard lessons at last

The meaning of Sachin Tendulkar
Parsing Reproducibility

“Empirical Reproducibility”

“Computational Reproducibility”

“Statistical Reproducibility”

V. Stodden, IMS Bulletin (2013)
Really Reproducible Research

• “Really Reproducible Research” 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.”
SparseLab (circa 2005)

SparseLab

SEEKING SPARSE SOLUTIONS TO LINEAR SYSTEMS OF EQUATIONS

SparseLab is provided free of charge, but we do request you register your use of the software by clicking on this link: REGISTER

Please see the Documentation tab on the left to find helpful materials for the installation and use of SparseLab. SparseLab 2.1 is now available! See the documentation folder in SparseLab 2.1 for changes and updates.

The SparseLab package is downloadable in three components: a "core" package containing the code (including Demos, Examples, Papers, etc), and two "Data Supplements". Some of the Demo figures use large datasets and we've made these into separate downloads for those interested in SparseLab, but not necessarily interested in reproducing these figures.

To download the core package click here: DOWNLOAD SPARSELAB 2.1 (~33MB)

Older versions:
- DOWNLOAD SPARSELAB 2.0 (~26MB)
- DOWNLOAD SPARSELAB 1.0 (~12MB)

To download the data supplements:
- DOWNLOAD "Extensions of Compressed Sensing" DATA SUPPLEMENT (~23MB)
- DOWNLOAD "Sparse Solution to Underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit" DATA SUPPLEMENT (~11MB)
- DOWNLOAD "Fast Solution of 1-norm Minimization Problems When the Solution May be Sparse" DATA SUPPLEMENT (~23MB)
In the case of $p >> n$ (microarray data, image processing, data mining applications), traditional linear regression fails.

George Box defined Factor Sparsity: the vast majority of factors have zero effect ($\beta$ is sparse).

If $k < n$, $y = X\beta + z$ can still be modeled effectively.

How to select $k$?
**\( l_0/l_1 \) Equivalence**

The Phase Diagram: Organizes simulation results to show when model selection algorithms begin to fail when model complexity exceeds a certain threshold - the *phase transition*.

1. We would like: \( \min_\beta \|\beta\|_0 \) s.t. \( y = X\beta \)

2. We can compute: \( \min_\beta \|\beta\|_1 \) s.t. \( y = X\beta \)

Result (Donoho et al): Under certain broad circumstances, the solutions to 2. also solves 1.
The Problem Suite

A collection of problems with sparse solutions, \( S\{k,n,p\} \):

Each problem has: \( n \times p \) model matrix \( X \), \( k \)-sparse \( p \)-vector of coefficients \( \beta \),

For each \( k, n, p \) combination we run the model selection algorithm multiple times, and measure its success according to \( \| \beta^\hat - \beta \|_2 / \| \beta \|_2 \).

Code, data released in Sparselab in an interactive MATLAB environment.
Two Examples

Normalized $L_2$ Error, $\sqrt{2\log(p)}$ threshold, $z \sim N(0,4^2)$

Forward Stepwise

Stepwise with FDR threshold, $z \sim N(0,16)$, Normalized $L_2$ error, $p=200$

Forward Stepwise with False Discovery Rate Thresholding
Reproducibility is a Statistical Issue

• False discovery, multiple testing, file drawer problem, p-hacking (Simonsohn 2012), overuse and mis-use of p-values,

• Insufficient or inappropriate reporting/tracking practices,

• Data preparation, treatment of outliers,

• Poor statistical methods (nonrandom sampling, inappropriate methods,..)

• Model robustness to parameter changes and data perturbations,

• Investigator bias toward previous findings; conflicts of interest.
Experimental Bias

Experimental biases:

Figure 2: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).

Figure courtesy of James Berger
Measuring Advances

• 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>Category</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.
Journal Requirements

In January 2014 Science enacted new policies. 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.
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)
## Supporting Computational Science

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

### Workflow Tracking and Research Environments:
- VisTrails
- Galaxy
- Sumatra
- Kepler
- GenePattern
- Taverna
- CDE
- Paper Mâché
- Pegasus
- IPython Notebook

### Embedded Publishing:
- Verifiable Computational Research
- Collage Authoring Environment
- SOLE
- knitR
- SHARE
- Sweave
Statistical Issues in Software

The challenge of reproducible computational science:

- encoding good statistical practices, i.e. capturing multiple comparisons,
- permitting independent verification and comparison,
- software testing: e.g. reconstructing figures,
- extending statistical notions of integrity to statistical software practices.
Research Compendia

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

- link data/code to published claims, re-use,
- research produces a guide to empirical researchers, certifies 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 interational 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
A proof of concept for a research compendia webapp http://researchcompendia.org — Edit

- 542 commits
- 12 branches
- 29 releases
- 1 contributor

Merge branch 'release/1.0.1-b9' into develop

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

- improvements in mentoring
- labeling retractions as “voluntary withdrawal” or “withdrawal for cause”
- neutral language such as “disclosure of relevant relationships” rather than “conflict of interest”
- universal and improved ethics education
- an independent Scientific Integrity Advisory Board to provide leadership in addressing ethical issues in research conduct
- avoiding hype in publicizing discoveries

Self-correction in science at work
Improve incentives to support research integrity


Week after week, news headlines carry word of new scientific discoveries, but the media sometimes give suspect science equal play with substantive discoveries. Careful qualifications about what is known are lost in categorical headlines. Rare instances of misconduct or instances of irreproducibility are transmuted into concerns that science is broken. The October 2013 Economist headlines proclaimed “Trouble at the Lab: Scientists lose confidence in the scientific method.” In the opening sentence, the Economist noted, “In an alarming degree, it is now the case that the results of experiments are not reproducible.” An image of a scientist holding his head in his hands under the headline, “The scientific method has hit a crisis.”

This article explores the causes of distrust in science and how it can be improved. The authors argue that the scientific community must work to correct its own mistakes and that such self-correction is essential for the health of science. They propose several strategies to improve research integrity, including improvements in mentoring, labeling retractions, using neutral language, universal ethics education, creating an independent Scientific Integrity Advisory Board, and avoiding hype in publicizing discoveries.
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.

**Organizing Committee**

- David H. Bailey  
  (Lawrence Berkeley National Laboratory)
- Jon Borwein  
  (Centre for Computer Assisted Research Mathematics and its Applications)
- Randall J. LeVeque  
  (University of Washington)
- Bill Rider  
  (Sandia National Laboratory)
- William Stein  
  (University of Washington)
- Victoria Stodden  
  (Columbia University)
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.
<table>
<thead>
<tr>
<th>Criterion</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assertions (#1)</td>
<td>A precise statement of assertions to be made in the paper.</td>
</tr>
<tr>
<td>Comp. Approach (#2)</td>
<td>A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.</td>
</tr>
<tr>
<td>Software Cited (#3 &amp; 4)</td>
<td>Complete statements of, or references to, every algorithm employed, and salient details of auxiliary software (both research and commercial software) used in the computation.</td>
</tr>
<tr>
<td>Hardware Discussed (#5)</td>
<td>Salient details of the test environment, including hardware, system software and the number of processors utilized.</td>
</tr>
<tr>
<td>Analysis (#6)</td>
<td>Salient details of data reduction and statistical analysis methods.</td>
</tr>
<tr>
<td>Parameter Discussed (#7)</td>
<td>Discussion of the adequacy of parameters such as precision level and grid resolution.</td>
</tr>
<tr>
<td>Parameters Given (#7)</td>
<td>Were necessary run parameters given?</td>
</tr>
<tr>
<td>Results (#8)</td>
<td>Full statement (or at least a valid summary) of experimental results.</td>
</tr>
<tr>
<td>Available Code (#10)</td>
<td>Availability of computer code, input data and output data, with some reasonable level of documentation.</td>
</tr>
<tr>
<td>Functions Calls</td>
<td>Which precise functions were called, with what settings?</td>
</tr>
<tr>
<td>Comp. Instructions (#12)</td>
<td>Instructions for repeating computational experiments described in the paper.</td>
</tr>
<tr>
<td>Alternate Avenues (#14)</td>
<td>Avenues of exploration examined throughout development, including information about negative findings.</td>
</tr>
<tr>
<td>Citation (#15)</td>
<td>Proper citation of all code and data used, including that generated by the authors.</td>
</tr>
</tbody>
</table>
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)
We need:

Standards for reproducibility of big data findings:

1. data access, software access, persistent linking to publications.

2. innovation around data and code access for privacy protection and scale.

3. robust methods, producing stable results, emphasis on reliability and reproducibility.

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

- 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

Abstract

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.
Open Science from the Whitehouse

- Feb 22, 2013: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.

- May 9, 2013: Executive Order directing federal agencies to make their data publicly available.

Executive Memorandum: “Expanding Public Access to the Results of Federally Funded Research”

• “Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries.”

• “digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze.”

• “digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings”

• “Each agency shall submit its draft plan to OSTP within six months of publication of this memorandum.”
Executive Order: “Making Open and Machine Readable the New Default for Government Information”

• “The Director … shall issue an Open Data Policy to advance the management of Government information as an asset”

• “Agencies shall implement the requirements of the Open Data Policy”

• “Within 30 days of the issuance of the Open Data Policy, the CIO and CTO shall publish an open online repository of tools and best practices”
Request for Input: “Strategy for American Innovation”

• “to guide the Administration's efforts to promote lasting economic growth and competitiveness through policies that support transformative American innovation in products, processes, and services and spur new fundamental discoveries that in the long run lead to growing economic prosperity and rising living standards.”

• “(11) Given recent evidence of the irreproducibility of a surprising number of published scientific findings, how can the Federal Government leverage its role as a significant funder of scientific research to most effectively address the problem?”
• NSF grant guidelines: “NSF ... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.” (2005 and earlier)

• NSF peer-reviewed Data Management Plan (DMP), January 2011.

• NIH (2003): “The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.” (>500,000, include data sharing plan)
September 8, 2015

These days, much discussion about the reproducibility of scientific results seems driven by critiques of research in biomedicine and psychology. Most recently, an article in Science concluded that 60 percent of a collection of studies were not replicable. This result along with similar analyses of cancer research results have stimulated strong commentary. For example, the New York Times print edition headline about the Science article was “Psychology’s Fears Confirmed: Repeated Studies Don’t Hold Up,” coverage that prompted a strong op-ed rebuttal titled, “Psychology Is Not in Crisis.”

Issues that arise with human subjects or with other complex systems do not plague physical science to the same degree. However, the notion of measuring the same value of a physical quantity or the same behavior of a physical system in different laboratories at different times is central to our concept of a valid scientific result. Often the approach is not simple to replicate an experiment, but rather to get at the same quantity via different paths. For example, we can measure the gravitational constant, G, with approaches ranging from a torsional pendulum to atom interferometry.

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

In recent years, however, there has been a growing awareness of the need for rigorously designed published preclinical studies, to ensure that such studies can be reproduced. This webpage provides information about the efforts underway by NIH to enhance rigor and reproducibility in scientific research.
National Science Board Report


• **Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences, (2003)**

• “Principle 1. Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims.”
Future Directions

Computational reproducibility addresses whether fixed codes/data can replicate findings, permitting verification, validation, and the reconciliation of differences in independent efforts.

• does not directly address whether these findings improve our understanding of the world.

• we might expect that for such findings, repeated independent replications yield results that are “close.” Possible sources of variation (Yu, 2013):
  
  • Stability: “reasonable” perturbations in the underlying data.
  
  • Robustness: perturbations in methods (due to changes in the parametrization, model or model assumptions).

Relationship to VV&EQ in scientific computing.
Caution: Roadwork Ahead..
Accessing code/data

Science Magazine policy as of Feb 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. Any restrictions on the availability of data, codes, or materials, including fees and original data obtained from other sources... must be disclosed to the editors upon submission”
Metrics for Empirical Evaluation


• Obtained a random sample of 204 scientific articles with computational findings.

• Posed three questions:

  1. How effectively were code/data procured?

  2. Could the published results be reproduced? (why not?)

  3. How effective are the ICERM standards in addressing the failures?
<table>
<thead>
<tr>
<th>Criterion</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assertions (#1)</td>
<td>A precise statement of assertions to be made in the paper.</td>
</tr>
<tr>
<td>Comp. Approach (#2)</td>
<td>A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.</td>
</tr>
<tr>
<td>Software Cited (#3 &amp; 4)</td>
<td>Complete statements of, or references to, every algorithm employed, and salient details of auxiliary software (both research and commercial software) used in the computation.</td>
</tr>
<tr>
<td>Hardware Discussed (#5)</td>
<td>Salient details of the test environment, including hardware, system software and the number of processors utilized.</td>
</tr>
<tr>
<td>Analysis (#6)</td>
<td>Salient details of data reduction and statistical analysis methods.</td>
</tr>
<tr>
<td>Parameter Discussed (#7)</td>
<td>Discussion of the adequacy of parameters such as precision level and grid resolution.</td>
</tr>
<tr>
<td>Parameters Given (#7)</td>
<td>Were necessary run parameters given?</td>
</tr>
<tr>
<td>Results (#8)</td>
<td>Full statement (or at least a valid summary) of experimental results.</td>
</tr>
<tr>
<td>Available Code (#10)</td>
<td>Availability of computer code, input data and output data, with some reasonable level of documentation.</td>
</tr>
<tr>
<td>Functions Calls</td>
<td>Which precise functions were called, with what settings?</td>
</tr>
<tr>
<td>Comp. Instructions (#12)</td>
<td>Instructions for repeating computational experiments described in the paper.</td>
</tr>
<tr>
<td>Alternate Avenues (#14)</td>
<td>Avenues of exploration examined throughout development, including information about negative findings.</td>
</tr>
<tr>
<td>Citation (#15)</td>
<td>Proper citation of all code and data used, including that generated by the authors.</td>
</tr>
</tbody>
</table>
Obtaining code/data

- Of the random sample of 204, 24 papers provided direct access to code/data.
- For the remaining 180 articles, the corresponding author was sent an email request:

<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>46</td>
<td>26%</td>
</tr>
<tr>
<td>Email bounced</td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Impossible to share</td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Refusal to share</td>
<td>12</td>
<td>7%</td>
</tr>
<tr>
<td>Contact to another person</td>
<td>20</td>
<td>11%</td>
</tr>
<tr>
<td>Asks for reasons</td>
<td>20</td>
<td>11%</td>
</tr>
<tr>
<td>Unfulfilled promise to follow up</td>
<td>5</td>
<td>3%</td>
</tr>
<tr>
<td>Direct back to SOM</td>
<td>6</td>
<td>3%</td>
</tr>
<tr>
<td>Shared data and code</td>
<td>65</td>
<td>36%</td>
</tr>
</tbody>
</table>

Total 180 100%

51% compliance rate
Upon inspection we deemed 56 of the 89 articles potentially reproducible (not including the 3 papers who could not share), and chose a random sample of 22 from the 56 to implement:

<table>
<thead>
<tr>
<th>ICERM Criterion</th>
<th>#Papers</th>
<th>Percent</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewable</td>
<td>22</td>
<td>100%</td>
<td>The descriptions permit the research methods to be independently assessed and the results judged credible.</td>
</tr>
<tr>
<td>Replicable</td>
<td>19</td>
<td>87%</td>
<td>Tools are made available that would allow one to duplicate the results of the research.</td>
</tr>
<tr>
<td>Confirmable</td>
<td>20</td>
<td>91%</td>
<td>The main conclusions of the research can be attained independently without the use of software provided by the author. (But using the complete description of algorithms and methodology provided.)</td>
</tr>
<tr>
<td>Auditable</td>
<td>17</td>
<td>77%</td>
<td>Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks.</td>
</tr>
<tr>
<td>Reproducible (by us)</td>
<td>21</td>
<td>95%</td>
<td>Auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.</td>
</tr>
<tr>
<td>Criterion</td>
<td>Papers</td>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>--------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Assertions (#1)</td>
<td>56</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Comp. Approach (#2)</td>
<td>46</td>
<td>82.14%</td>
<td></td>
</tr>
<tr>
<td>Software Cited (#3 &amp; 4)</td>
<td>45</td>
<td>80.36%</td>
<td></td>
</tr>
<tr>
<td>Hardware Discussed (#5)</td>
<td>7</td>
<td>12.5%</td>
<td></td>
</tr>
<tr>
<td>Analysis (#6)</td>
<td>51</td>
<td>91.07%</td>
<td></td>
</tr>
<tr>
<td>Parameter Discuss (#7)</td>
<td>44</td>
<td>78.57%</td>
<td></td>
</tr>
<tr>
<td>Parameters given (#7)</td>
<td>48</td>
<td>85.71%</td>
<td></td>
</tr>
<tr>
<td>Results (#8)</td>
<td>56</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Available Code (#10)</td>
<td>43</td>
<td>76.79%</td>
<td></td>
</tr>
<tr>
<td>Functions Calls</td>
<td>23</td>
<td>41.07%</td>
<td></td>
</tr>
<tr>
<td>Comp. Instructions (#12)</td>
<td>35</td>
<td>62.5%</td>
<td></td>
</tr>
<tr>
<td>Alternate Avenues (#14)</td>
<td>38</td>
<td>67.86%</td>
<td></td>
</tr>
<tr>
<td>Citation (#15)</td>
<td>44</td>
<td>78.57%</td>
<td></td>
</tr>
</tbody>
</table>
**Verification and Validation**

For the 56 potentially replicable papers, we evaluated:

<table>
<thead>
<tr>
<th>ICERM Criterion #9</th>
<th>#Papers</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifiable</td>
<td>34</td>
<td>Check that the computer code correctly solves the mathematical problem it claims to solve.</td>
</tr>
<tr>
<td>Validatable</td>
<td>51</td>
<td>Is it the right mathematical formulation? Check that the results agree with experiments or observations of the phenomenon being studied.</td>
</tr>
<tr>
<td>Verified (by author)</td>
<td>36</td>
<td>Verification tests performed by the author(s).</td>
</tr>
<tr>
<td>Validated (by author)</td>
<td>51</td>
<td>Validation tests performed by the author(s).</td>
</tr>
</tbody>
</table>
Preliminary Conclusions

Where did the papers fail?

1. 49% did not make sufficient code/data available, 2% could not.

2. Of the remaining 49%, we replicated 19 of the 22 randomly chosen from 56 possibly reproducible articles; we estimate 48 of the 56 may replicate.

3. In our random sample of 22, model details were missing for 2 papers, code was missing for 3 papers, and code changes since publication made replication impossible for one.

We estimate 27% (48/177) of the computational articles published in Science since Feb 11, 2011 will replicate.
Future Directions

• Quantify the impact of data and model perturbations to develop metrics for assessing empirical evidence from replication
  
  • Implement on sparse problem set $S\{k,n,p\}$ with $p>>n$.
  
  • Implement on examples (i.e. microarray data).

• Identify sources of error from reproducibility issues: uncertainty quantification adapted to include data and model errors, and software encoding issues.
## Defining Reproducibility

<table>
<thead>
<tr>
<th>ICERM Criterion</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewable</td>
<td>The descriptions permit the research methods to be independently assessed and the results judged credible.</td>
</tr>
<tr>
<td>Replicable</td>
<td>Tools are made available that would allow one to duplicate the results of the research.</td>
</tr>
<tr>
<td>Confirmable</td>
<td>The main conclusions of the research can be attained independently without the use of software provided by the author. (But using the complete description of algorithms and methodology provided.)</td>
</tr>
<tr>
<td>Auditable</td>
<td>Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks.</td>
</tr>
<tr>
<td>Reproducible</td>
<td>Auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.</td>
</tr>
</tbody>
</table>