Open Data and Reproducibility in Research

Victoria Stodden
School of Graduate and Library Science
University of Illinois at Urbana-Champaign

OpenCon 2014
American University Washington College of Law
Washington, D.C.
Nov 15, 2014
Agenda

1. What makes scientific research scientific?
2. How has computation interacted with the research process?
3. Treading carefully: conflicts with scientific goals
4. Implications for public discourse
5. Some solutions and ideas
Advances in Technology

1. enormous, and increasing, amounts of data collection:
   - CMS project at LHC: 300 “events” per second, 5.2M seconds of runtime per year, 5MB per event = 780TB/yr => several PB when data processed,
   - Sloan Digital Sky Survey: 9th data release (SDSS-III 2012), 60TB,
   - quantitative revolution in social science due to abundance of social network data (Lazier et al, Science, 2009),
   - NIH Associate Director for Data Science, 2014.

2. computational power: massive simulations of the complete evolution of a physical system, systematically varying parameters,

3. deep intellectual contributions now encoded in software.
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
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.
Implementation of Norms

*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.

Advances in the technology used for scientific discovery have changed how scientists effect reproducibility.
The Scientific Method

Traditionally two branches to the scientific method:

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

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

Argument: computation presents only a potential third/fourth branch of the scientific method (Donoho et al 2009).
New Paradigms for Discovery?

“New Paradigms for Discovery?”

“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 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, structured communication of methods and protocols.

Conjecture: Computational science as practiced today does not generate routinely verifiable knowledge.
Credibility Crisis

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.

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 was not reproducible. Because confidence in result, community, we are announcing new initiatives.

NIH Tackles Irreproducibility

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

By Jef Akst | January 28, 2014

Must try harder

Natur 483, 509 (29 March 2012) | doi:10.1038/483509a
Published online 28 March 2012

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.
Parsing Reproducibility

“Empirical Reproducibility”

“Computational Reproducibility”

“Statistical Reproducibility”

V. Stodden, IMS Bulletin (2013)
Computational Reproducibility

“Really Reproducible Research” pioneered 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.”

“Reproducible Research” is Grassroots

- **reproducibility@XSEDE**: An XSEDE14 Workshop
- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- Open Science Framework / Reproducibility Project in Psychology
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: “The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer”
- SIAM CSE 2011: “Verifiable, Reproducible Computational Science”
- Yale Law School 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- **ACM SIGMOD** conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials”
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.
Supporting Computational Science

- **Dissemination Platforms:**
  - ResearchCompendia.org
  - MLOSS.org
  - Open Science Framework

- **Workflow Tracking and Research Environments:**
  - VisTrails
  - Kepler
  - Galaxy
  - GenePattern
  - Sumatra
  - Taverna

- **Embedded Publishing:**
  - Verifiable Computational Research
  - Collage Authoring Environment

- **Verifiable Computational Research SOLE knitR**
- **Collage Authoring Environment SHARE Sweave**
- **MLOSS.org thedatahub.org nanoHUB.org RunMyCode.org**
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 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 an 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
Joint work with Sheila Miguez and Jennifer Seiler
Funded by the Alfred P. Sloan Foundation
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.

• July 29, 2014: Notice of Request for Information “Strategy for American Innovation”
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?”
Science Policy in Congress

• America COMPETES due to be reauthorized, drafting underway.


• Hearing on Research Integrity and Transparency by the House Science, Space, and Technology Committee (March 5, 2013).

• Reproducibility cannot be an unfunded mandate.

Sharing: Funding Agency Policy

• 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)
NSF Data Management Plan

“Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled ‘Data Management Plan.’ This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.” (http://www.nsf.gov/bfa/dias/policy/dmp.jsp)

Software management plans appearing.. (BigData joint NSF/NIH solicitation)
DOE Data Management Plan

“The Department is taking a phased approach to the implementation of requirements set forth by the OSTP memo. In particular, the Office of Science, which supports roughly two-thirds of the total R&D for the Department, plans to pilot a data management policy with the requirements described below by July 28, 2014. Other DOE Offices and elements with over $100 million in annual conduct of research and development expenditures will implement data management plan requirements that satisfy the requirements of the OSTP memo no later than October 1, 2015 in such a way that there is a single DOE policy for data management planning.” (DOE Public Access Plan 2014)

“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.”
The Larger Community

1. Production: Crowdsourcing and public engagement in science
   • primarily data collection/donation today, but open up pipeline:
     - access to “coherent” digital scholarly objects,
     - mechanism for ingesting/evaluating new findings,
     - addressing legal issues (use, re-use, privacy,…).

2. Use: “Evidence-based”- {policy, medicine, …}, decision making.
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?
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.

References


available at http://www.stodden.net
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:
   - is a work-for-hire (and copyright therefore vests in the University under copyright law),
   - is supported by a direct allocation of funds through the University for the pursuit of a specific project,
   - is commissioned by the University,
   - makes significant use of University resources or personnel, or
   - is otherwise subject to contractual obligations.

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.

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.

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 requirements

• Changes in journal publication requirements
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.
Journal Publication Requirements

• Journal Policy snapshots June 2011 and June 2012:

• 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.
## Journal Data 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>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)
Software in Scientific Discovery

<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>

Reproducible Research Movement:
Data and code are made conveniently available at the time of publication

Conflict between incentives to patent academic code and the scientific method?
## 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><strong>Possibility of patents</strong></td>
</tr>
<tr>
<td>34%</td>
<td>Legal Barriers (ie. copyright)</td>
</tr>
<tr>
<td>30%</td>
<td>Time to verify release with admin</td>
</tr>
<tr>
<td>30%</td>
<td>Potential loss of future publications</td>
</tr>
<tr>
<td>20%</td>
<td>Competitors may get an advantage</td>
</tr>
<tr>
<td></td>
<td>Web/disk space limitations</td>
</tr>
</tbody>
</table>

*Survey of the Machine Learning Community, NIPS (Stodden 2010)*
Potential implications for science as a disruptor of openness norms:

• delay in revealing code, or closed code,

• potential obfuscation of methods submitted for patents (*Bilski v. Kappos*),

• alteration of a scientist's incentives toward commercial ends, instead of the production of science as a public good.