Open Licensing and Science Policy

Victoria Stodden
Department of Statistics
Columbia University

Guest Lecture
Columbia University
April 16, 2014
Agenda

1. Creating Reliable Computational Science: Updating Research Paradigms
2. Statistical Inference in “Big Data” Settings
3. Rethinking Inference and Scientific Knowledge Production
4. Some Efforts to Address these Challenges
I. Creating Reliable Computational Science: Updating Research Paradigms
Reproducibility in Science

• not a new concept, rooted in skepticism
• Transactions of the Royal Society 1660’s
• Transparency, knowledge transfer -> goal to perfect the scholarly record.
• Technology has changed the nature of experimentation, data, and communication.
Computation is Becoming Central to Scientific Research

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 = \(780\text{TB/yr}\) => 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)
   - *Science survey* of peer reviewers: 340 researchers regularly work with datasets >100GB; 119 regularly work with datasets >1TB (N=1700, Feb 11, 2011, p. 692)

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

3. deep intellectual contributions now encoded in software.
## Credibility Crisis

<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): 9% of authors studied made data available.

Generally, data and code not made available at the time of publication, insufficient information in the publication for verification, replication of results. A Credibility Crisis
“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.”

Updating the Scientific Method

Argument: computation presents only a potential third branch of the scientific method (Stodden et al 2009):

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments,
- Branch 3,4? (computational): large scale simulations / data driven computational science.
Commonly asserted...

“"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.”
Credibility Crisis

Los Angeles Times | BUSINESS

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.

The Economist

HOW SCIENCE GOES WRONG.

Nature

Announcement: Reducing our irreproducibility

Over the past year, Nature has published a string of articles that highlighted the reliability and reproducibility of published research (collected as a special collection on the Progress in Science website: http://scim.ag/reproducibility).

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

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)
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.
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.
Data and Software

1. Digital databases require software and computing power for curation, filtering, analysis, visualization, and storage.

2. Computation requires software, and scientific investigation either produces data (e.g. simulations), or uses data as an input (e.g. data science), or both.

→ Data and software should be considered together as different but inextricably linked components of the computational science research process.
Intellectual Property Barriers

• Software is both copyrighted (by default) and patentable.

• Copyright: author sets terms of use using an open license:
  • Attribution only (ie. Modified BSD, MIT license, LGPL)
  • Reproducible Research Standard (Stodden 2009)

• Patents: Bayh-Dole (1980) vs reproducible research (Stodden 2012)
  • delays, barriers to software access
  • Bilski v Kappos (2011)
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?
GPL and Copyleft

“sharealike” See e.g. [https://www.gnu.org/copyleft/](https://www.gnu.org/copyleft/)
Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (i.e. 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.
Impact of Computation

In the computational sciences, disclosure of data and code are considered essential for reproducibility. Software can be patent-eligible (Bilski v. Kappos 130 S. Ct. 3218 2010), increasing the reach of Bayh-Dole in the sciences.

Universities can claim ownership over software developed in the course of research on this basis and potentially patent then license access to the code.

*Hypothesis:* The Bayh-Dole Act inhibits reproducibility in the computational sciences, and is a barrier to access to research inventions.
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)
“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)
2013: Open Science in DC

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

• May 9: Executive Order directing federal agencies to make their data publicly available.
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).

• Reproducibility cannot be an unfunded mandate.
National Science Board Report


NAS Data Sharing Report


• “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.”
March 23 2012, Institute of Medicine releases report,

Recommends new standards for omics-based tests, including a fixed version of the software, expressly for verification purposes.
“The fully specified computational procedures are locked down in the discovery phase and should remain unchanged in all subsequent development steps.”
2. Statistical Inference in “Big Data” Settings
Traditional Sources of Error

Statistics has a long history of identifying and controlling for error in empirical experiments, e.g.:

- sampling errors,
- selection bias,
- model specification errors, functional form misspecification,
- omitted variable bias,
- ...

ICERM Reporting Approach

• Omitted information regarding the implementation of the experiment (e.g. parameters, parameter estimation) as source of irreproducibility.

• Introduces uncertainty when interpreting or re-estimating results.

• Addressed in ICERM workshop report.
Reproducibility at Scale

Scale Issues: both for large datasets and compute time.

- data produced by code - making the code available permit data regeneration, but may involve prohibitive runtimes.

- partial data precomputation - what standards? partial results checking and testing. data synopses, simulated data.

- do some research areas pose extra difficulties?
reproducibility @ XSEDE: An XSEDE14 Workshop
Monday, July 14, 2014 - Atlanta, GA

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...
Stability and Robustness of Results

• Computational reproducibility addresses whether fixed codes/data can replicate findings, permitting 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 (B.Yu, 2013):

  • Stability: “reasonable” perturbations in the underlying data.

  • Robustness: perturbations in methods (due to changes in the parametrization, model, or model assumptions).
Some “New” Sources of Error

- Statistical: frequentist paradigm, p-values, multiplicity, power (dataset size improves the reliability results!), outlier classification, hypothesis testing.

- Computational: traversing data, tools, scaling algorithms (dataset size is a challenge!).

- Reconciling conflicting results,

- Duplication of efforts,

- Standing on the shoulders of giants: software perspective.
3. Rethinking Inference and Scientific Knowledge Production
We need:

• Standards for reproducibility of big data findings:

  1. data access, software access (what form? what documentation? collaboration with proprietary software holders).

  2. data generation vs data analysis.

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

Example: Google Flu Trends results: worked at first, but what happened?
4. Some Efforts to Address these Challenges
Openness and Science Policy

• Science Policy must support scientific ends: Reliability and accuracy of the scientific record.

• Facilitate Reproducibility - the ability to regenerate published results (data and code availability, alongside results).

• Need infrastructure to facilitate (1):
  1. deposit/curation of versioned data and code,
  2. link to published article,
  3. permanence of link.
Constructing Policy

- “Open Data” is not well-defined. Scope: Share data and code that permit others in the field to replicate published results. (traditionally done by the publication alone). Corollary: maximizes data reuse.

- Data and code availability at the time of publication.

- Public access. “With many eyeballs, all bugs are shallow.”

- Need infrastructure/software tools to facilitate (2):
  1. data/code suitable for sharing, created during the research process.
2013: Open Science in DC

- Feb 22: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.
- May 9: Executive Order directing federal agencies to make their data publicly available.
Tools for Computational Science

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

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

• Embedded Publishing:
  Verifiable Computational Research  Sweave  knitR
  Collage Authoring Environment  SHARE
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 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.
A proof of concept for a research compendia webapp [http://researchcompendia.org](http://researchcompendia.org) — Edit

<table>
<thead>
<tr>
<th>branch</th>
<th>description</th>
<th>latest commit</th>
<th>committer</th>
<th>authored</th>
<th>ago</th>
</tr>
</thead>
<tbody>
<tr>
<td>develop</td>
<td>merge branch 'release/1.0.1-b9' into develop</td>
<td>d3fb4917d</td>
<td>codersquid</td>
<td>30 minutes ago</td>
<td></td>
</tr>
<tr>
<td>companionpages</td>
<td>bump revision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>docs</td>
<td>removes instructions for envdir and bootstrap.sh, adds instructions for...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>requirements</td>
<td>citation dialog and display for journals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.gitignore</td>
<td>adds vagrant and bootstrap starter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.travis.yml</td>
<td>fixed broken doi service test and updated irc channel for travis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUTHORS.rst</td>
<td>renaming project from tyler to researchcompendia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CITATION.bib</td>
<td>bump revision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONTRIBUTING.rst</td>
<td>fixed thinko of 'comment' to 'commit'</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HISTORY.rst</td>
<td>bump revision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LICENSE</td>
<td>release 1.0.0-alpha1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MANIFEST.in</td>
<td>making skeleton docs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HTTPS clone URL

[https://github.com/](https://github.com/)
“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”
References


available at http://www.stodden.net