Rethinking the Scholarly Record: Facilitating Reproducibility of Computational Results

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

Bringing Rigor and Reproducibility to Research
University of Pittsburgh
September 22, 2016
Agenda

1. Conceptualizing Technological Changes
   i. data collection and storage,
   ii. computational power,
   iii. software,
   iv. communication.

2. Grounding in Scientific Norms

3. Impact on the Scholarly Record
1. Conceptualizing Technological Change
The Impact of Technology I

1. Big Data / Data Driven Discovery: high dimensional data, \( p >> n \),

2. Computational Power: simulation of the complete evolution of a physical system, systematically varying parameters,

3. Deep intellectual contributions now encoded only in software.

The software contains “ideas that enable biology...”
Stories from the Supplement, 2013
The Impact of Technology II

1. Communication: nearly all aspects of research becoming *digitized* and *accessible* due to the Internet.
   - myriad examples.. including the Open Access movement.

2. Intellectual Property Law: digitally shared objects often have more and more easily enforceable IP rights associated.
   - Reproducible Research Standard (Stodden 2009).
2. Grounding Changes in Scientific Norms
Parsing Reproducibility

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”

V. Stodden, IMS Bulletin (2013)
Empirical Reproducibility

Sorting Out the FACS: A Devil in the Details

William C. Hines,1,8,5* Ying Su,2,3,4,5* Irene Kuhn,1 Kornelia Polyak,2,3,4,5* and Mina J. Bissell1,5

1Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 9771255A, 1 Cyclotron Road, Berkeley, CA 94720, USA
2Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
3Department of Medicine, Brigham and Women’s Hospital, Boston, MA 02115, USA
4Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
5These authors contributed equally to this work
6Correspondence: china@lfnl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.celrep.2014.02.021

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

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

Reproducibility Issues in Research with Animals and Animal Models

The missing “R”: Reproducibility in a Changing Research Landscape

A workshop of the Roundtable on Science and Welfare in Laboratory Animal Use

National Academy of Sciences, NAS 125
2100 C Street NW, Washington DC
June 4-5, 2014

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible have alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues. Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.
Statistical Reproducibility

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

• Low power, poor experimental design, nonrandom sampling,

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

• inappropriate tests or models, model misspecification,

• Model robustness to parameter changes and data perturbations,

• ...
Computational Reproducibility

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.
The Ubiquity of Error

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

- Deductive branch: the well-defined concept of the proof,
- Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation presents only a potential third/fourth branch of the scientific method (Donoho, Stodden, et al. 2009), until the development of comparable standards.
Really Reproducible Research

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

“The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.” David Donoho, 1998

Note the difference between: reproducing the computational steps and, replicating the experiments independently including data collection and software implementation. (Both required)
Merton’s Scientific Norms (1942)

**Communalism**: scientific results are the common property of the community.

**Universalism**: all scientists can contribute to science regardless of race, nationality, culture, or gender.

**Disinterestedness**: act for the benefit of a common scientific enterprise, rather than for personal gain.

**Originality**: scientific claims contribute something new

**Skepticism**: scientific claims must be exposed to critical scrutiny before being accepted.
Skepticism -> Reproducibility

- 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.
3. The Impact on the Scholarly Record
Rethinking the Notion of the Scholarly Record

**Idea**: The Scholarly Record comprises *access to* and/or the ability to regenerate:

1. items relied on in the generation of results AND/OR
2. items required for independent replication and reproducibility.

The difference is that *unreported research paths* are included in 1.
“Items”

- digital scholarly objects such as articles, texts, code, software, data, workflow information, research environment details, ...

- material objects such as reagents, lab equipment, instruments (telescopes, hadron colliders..), texts, historical artifacts, ...

Note: versioning and identification is crucial.
Infrastructure Responses

Tools and software to enhance reproducibility and disseminate the scholarly record:

Dissemination Platforms

- ResearchCompendia.org
- MLOSS.org
- Open Science Framework
- IPOL
- thedatahub.org
- nanoHUB.org
- Madagascar
- RunMyCode.org

Workflow Tracking and Research Environments

- Vistrails
- Kepler
- GenePattern
- CDE
- Jupyter
- torch.ch
- Galaxy
- GenePattern
- Sumatra
- Taverna
- DataCenterHub
- Pegasus
- Kurator
- RCloud

Embedded Publishing

- Verifiable Computational Research
- SHARE
- knitR
- Collage Authoring Environment
- SOLE
- Sweave
Research Compendia

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

• link data/code to published claims, re-use,
• 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>Commit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>d3fab4917d</td>
<td>latest commit</td>
</tr>
<tr>
<td>542</td>
<td>commits</td>
</tr>
<tr>
<td>12</td>
<td>branches</td>
</tr>
<tr>
<td>29</td>
<td>releases</td>
</tr>
<tr>
<td>1</td>
<td>contributor</td>
</tr>
</tbody>
</table>

Merge branch 'release/1.0.1-b9' into develop

- **codersquid** authored 30 minutes ago
  - Latest commit: d3fab4917d

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>companionpages</td>
<td>bump revision</td>
</tr>
<tr>
<td>docs</td>
<td>removes instructions for envdir and bootstrap.sh, adds instructions for...</td>
</tr>
<tr>
<td>requirements</td>
<td>citation dialog and display for journals</td>
</tr>
<tr>
<td>.gitignore</td>
<td>adds vagrant and bootstrap starter</td>
</tr>
<tr>
<td>.travis.yml</td>
<td>fixed broken doi service test and updated irc channel for travis</td>
</tr>
<tr>
<td>AUTHORS.rst</td>
<td>renaming project from tyler to researchcompendia</td>
</tr>
<tr>
<td>CITATION.bib</td>
<td>bump revision</td>
</tr>
<tr>
<td>CONTRIBUTING.rst</td>
<td>fixed thinko of 'comment' to 'commit'</td>
</tr>
<tr>
<td>HISTORY.rst</td>
<td>bump revision</td>
</tr>
<tr>
<td>LICENSE</td>
<td>release 1.0.0-alpha1</td>
</tr>
<tr>
<td>MANIFEST.in</td>
<td>making skeleton docs</td>
</tr>
</tbody>
</table>

[HTTPS clone URL](https://github.com/researchcompendia/researchcompendia)
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.
Background: Open Source Software

- Innovation: Open Licensing
  → Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

- Hundreds of open source software licenses:
  - GNU Public License (GPL)
  - (Modified) BSD License
  - MIT License
  - Apache 2.0 License
  - ... see http://www.opensource.org/licenses/alphabetical
The Reproducible Research Standard

The *Reproducible Research Standard (RRS)* (Stodden, 2009): A suite of license recommendations for computational science

- Release *media* components (text, figures) under CC BY,
- Release *code* components under Modified BSD or similar,
- Release *data* to public domain or attach attribution license.

➡ Remove copyright’s barrier to reproducible research and,
➡ Realign the IP framework with longstanding scientific norms.
Community Responses

Declarations and Documents:

- Yale Declaration 2009
- ICERM 2012
- XSEDE 2014
Government Mandates

- OSTP 2013 Open Data and Open Access Executive Memorandum; Executive Order.
- “Public Access to Results of NSF-Funded Research”
- NOAA Data Management Plan, Data Sharing Plan
- NIST “Common Access Platform”
- ...
Federal Agencies

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.
Journal Requirements

• Science: code/data sharing since 2011.
• Nature: data sharing.
• JASA-ACS Associate Editor for Reproducibility
• ACM TOMS Replicated Computational Results Initiative
• ...

Journal Policy

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.
# Data Sharing Policies

<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)
## Code Sharing Policies

<table>
<thead>
<tr>
<th>Policy</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.
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.
CI Wishes

- data access, software access, persistent linking to publications.
- linked DOI assignment on article, data, code, workflows, compendia.
- innovation around data and code access for privacy protection and scale.
- robust methods, producing stable results, emphasis on reliability and reproducibility,
- open source.

Three Principles for CI

1. *Supporting scientific norms*—not only should CI enable new discoveries, but it should also permit others to reproduce the computational findings, reuse and combine digital outputs such as datasets and code, and facilitate validation and comparisons to previous findings.

2. *Supporting best practices in science*—CI in support of science should embed and encourage best practices in scientific research and discovery.

3. *Taking a holistic approach to CI*—the complete end-to-end research pipeline should be considered to ensure interoperability and the effective implementation of 1 and 2.

Social and political environment..

Exceptions: privacy, HIPAA, FERPA, other constraints on sharing.

Querying the Scholarly Record

- 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
Open Questions

• Incentivizing changes toward the production and dissemination of reproducible research.

• Who funds and supports cyberinfrastructure? Who controls access and gateways?

• Who owns data, code, and research outputs? Working around and within blocks such as privacy, legal barriers, ..

• What are community standards around documentation, citation standards, best practices? Who enforces?
New 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.