Implementing Reproducible Computational Research: Policy and Practice

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

Summer@CENSUS
June 5, 2017
Agenda

1. Framing the Reproducibility


3. Intellectual Property and Legal Considerations

4. Big Data, Privacy and the Public Good

5. Infrastructure and Looking Ahead
Remember Google Flu Trends?

In 2008 Google Flu Trends claimed it can tell you whether “the number of influenza cases is increasing in areas around the U.S., earlier than many existing methods”.

In 2013 Google Flu Trends was predicting more than double the proportion of doctor visits for flu than the CDC.

Today:

Thank you for stopping by.
Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue – we’re excited to see what comes next. Academic research groups interested in working with us should fill out this form.

Sincerely,
The Google Flu and Dengue Trends Team.
What Happened?

• How did Google Flu Trends work? What was the data collection process? What was the algorithm?

• Why should we believe Google Flu Trends output? Many people did in 2008..

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani

In February 2013, Google Flu Trends (GFT) made headlines. Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.
The Impact of Technology

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

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

3. Deep intellectual contributions now encoded only in software.

**Claim 1:** Virtually all published discoveries today have a computational component. *(is Data Science all science?)*

**Claim 2:** There is a mismatch between the traditional scientific process and computation, leading to reproducibility concerns.

---

"The actual scholarship is the full software environment, code and data, that produced the result.”

*Stories from the Supplement, 2013*

"The software contains “ideas that enable biology…”*

*Buckheit & Donoho, 1995*
2. Defining 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,1,2,* Ying Su,3,4,5,6,7 Irene Kuhn,1 Kornelia Polyak,3,4,5,6 and Mina J. Bissell7,8

1Life Sciences Division, Lawrence Berkeley National Laboratory, Malatiop 977R255A, 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 02215, USA
4Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
5These authors contributed equally to this work
6Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)

http://dx.doi.org/10.1016/j.cellrep.2014.02.021

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

Paired with in situ 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 Utz, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.’s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.’s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for

The 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 has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues. Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.
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, insufficient sample size,
- Data preparation, treatment of outliers and missing values, recombination of datasets,
- Inappropriate tests or models, model misspecification, poor parameter estimation techniques,
- Model robustness to parameter changes and data perturbations,
Example: Science 2014

In January 2014 Science enacted new manuscript submission requirements:

- a “data-handling plan” i.e. how outliers will be dealt with,
- sample size estimation for effect size,
- whether samples are treated randomly,
- whether experimenter blind to the conduct of the experiment.

Also added statisticians to the Board of Reviewing Editors.
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.
“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 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 and Data Science present only potential third/fourth branches of the scientific method (Donoho 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: reproducing the computational steps vs re-implementing the experiment independently (both types needed).
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
This workshop will consider ways to make code and modeling information more readily available, and include a variety of stakeholders.

The computational steps that produce scientific findings are increasingly considered a crucial part of the scholarly record, permitting transparency, reproducibility, and re-use. Important information about data preparation and model implementation, such as parameter settings or the treatment of outliers and missing values, is often expressed only in code. Such decisions can have substantial impacts on research outcomes, yet such details are rarely available with scientific findings.

http://www.aaas.org/event/iii-arnold-workshop-modeling-and-code
Feb 16-17, 2016
REPRODUCIBILITY

Enhancing reproducibility for computational methods
Data, code, and workflows should be available and cited

By Victoria Stodden,1 Marcia McNutt,2 David H. Bailey,3 Ewa Deelman,4 Yolanda Gil,4 Brooks Hanson,5 Michael A. Heroux,6 John P.A. Ioannidis,7 Michela Taufer8

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general proposals from the Transparency and Openness Promotion (TOP) guidelines (1) and recommendations for field data (2), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers representing a broad range of disciplines.

To understanding how computational results were derived and to reconciling any differences that might arise between independent replications (4). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (5, 6), which includes workflow information that explains what raw data and intermediate results are input to which computations (7). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.

RECOMMENDATIONS
Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable independent regeneration of computational results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fWwjPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly objects, citation should be standard practice. All data, code, and workflows, including software written by the authors, should be cited in the references section (10). We suggest that software citation include software version information and its unique identifier in addition to the author(s) name(s).

Workshop Recommendations: “Reproducibility Enhancement Principles”

RECOMMENDATION 1: To facilitate reproducibility, share the data, software, workflows, and details of the computational environment in open repositories.

RECOMMENDATION 2: To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

RECOMMENDATION 3: To enable credit for shared digital scholarly objects, citation should be standard practice.

RECOMMENDATION 4: To facilitate reuse, adequately document digital scholarly artifacts.
Workshop Recommendations: “Reproducibility Enhancement Principles”

RECOMMENDATION 5: Journals should conduct a Reproducibility Check as part of the publication process and enact the TOP Standards at level 2 or 3.

RECOMMENDATION 6: Use Open Licensing when publishing digital scholarly objects.

RECOMMENDATION 7: To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.
Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) via the Constitution and subsequent Acts:

“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)

Argument: both types of intellectual property are an imperfect fit with scholarly norms, and require action from the research community to enable re-use, verification, reproducibility, and support the acceleration of scientific discovery.
Copyright

• 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

• limited time: generally life of the author +70 years

• Exceptions and Limitations: e.g. Fair Use.
Licensing in Research

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 MIT License or similar,
- Release data to public domain (CC0) or attach attribution license.

- Remove copyright’s barrier to reproducible research and,
- Realign the IP framework with longstanding scientific norms.
Patents

Patentable subject matter: “new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof” (35 U.S.C. §101) that is

1. Novel, in at least one aspect,

2. Non-obvious,

3. Useful.

USPTO Final Computer Related Examination Guidelines (1996) “A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena” (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).
Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (i.e., Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole 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.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.
Legal Issues in Data

- 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)).

- Copyright adheres to raw facts in Europe.

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

- Legal mismatch: What constitutes a “raw” fact anyway?
Privacy and Data

- HIPAA, FERPA, IRB mandates create legally binding restrictions on the sharing human subjects data (see e.g. http://www.dataprivacybook.org/)

- Potential privacy implications for industry generated data.

- Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..
Privacy, Big Data, and the Public Good

Julia Lane,
Victoria Stodden,
Stefan Bender,
Helen Nissenbaum (eds)
Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency

Between Open and Closed:

Example 1: “Walled Gardens”

For protected data, i.e. subject to HIPAA, limit access to authorized researchers from independent groups to enable the verification of scientific findings, within a walled garden.

Example 2: “Data Lakes”

Department of Homeland Security approach: proactively tag permission levels for each dataset in the “lake” e.g. core biographical data, extended biographical data, DHS encounter data. (i.e. Neptune and Cerberos pilots)
Ownership: What Defines Contribution?

• Issue for producers: credit and citation.

• What is the role of peer-review?

• Repositories adding meta-data and discoverability make a contribution.

• Data repositories may be inadequate: velocity of contributions

• Future coders may contribute in part to new software, other software components may already be in the scholarly record. Attribution vs sharealike.

  ➔ (at least) 2 aspects: legal ownership vs scholarly credit.

• Redefining plagiarism for software contributions.
# Infrastructure Solutions

## Research Environments

<table>
<thead>
<tr>
<th>Verifiable Computational Research</th>
<th>SHARE</th>
<th>Code Ocean</th>
<th>Jupyter</th>
</tr>
</thead>
<tbody>
<tr>
<td>knitR</td>
<td>Sweave</td>
<td>Cyverse</td>
<td>NanoHUB</td>
</tr>
<tr>
<td>Collage Authoring Environment</td>
<td>SOLE</td>
<td>Open Science Framework</td>
<td>Vistrails</td>
</tr>
<tr>
<td>Sumatra</td>
<td>GenePattern</td>
<td>IPOL</td>
<td>Popper</td>
</tr>
<tr>
<td>Galaxy</td>
<td>torch.ch</td>
<td>Whole Tale</td>
<td>flywheel.io</td>
</tr>
</tbody>
</table>

## Workflow Systems

<table>
<thead>
<tr>
<th>Taverna</th>
<th>Wings</th>
<th>Pegasus</th>
<th>CDE</th>
<th>binder.org</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurator</td>
<td>Kepler</td>
<td>Everware</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Dissemination Platforms

<table>
<thead>
<tr>
<th>ResearchCompendia.org</th>
<th>DataCenterHub</th>
<th>RunMyCode.org</th>
<th>ChameleonCloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occam</td>
<td>RCloud</td>
<td>TheDataHub.org</td>
<td>Madagascar</td>
</tr>
<tr>
<td>Wavelab</td>
<td>Sparselab</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Experiment Definition Systems”

- Define and create “Experiment Definition Systems” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis.

- The two trends need to be addressed simultaneously: better transparency will allow people to run much more ambitious computational experiments. And better computational experiment infrastructure will allow researchers to be more transparent.
Looking Ahead

We see the convergence of two (ordinarily antagonistic) trends:

- Scientific projects will become massively more computing intensive
- Research computing will become dramatically more transparent

Reinforcing trends, resolution essential for verifying and comparing findings.
Proposition

• We propose a major effort to develop a new infrastructure that promotes good scientific practice downstream like transparency and reproducibility.

• But plan for people to use it not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work. Enables efficiency and productivity, and discovery.
Inducing a Reproducibility Industry by Grant Set-asides

• Previously, NIH required that clinical trials hire Biostatistician PhD's to design and analyze experiments. This set-aside requirement more or less directly transformed clinical trials practice and resulted in much more good science being done. It also spawned the modern field of Biostatistics, by creating a demand for a specific set of services and trained people who could conduct them.

• Why not try a similar idea for reproducibility?