Reproducibility in Computational Science: A Computable Scholarly Record

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

Center for Research Computing Seminar
Notredame University
South Bend, IN
October 24, 2016
1. Framing the Issues

2. Defining Reproducibility


4. Solutions, Tools, and Future Work
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:

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?


**BIG DATA**

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.
A Credibility Crisis

Science advances on a foundation of trusted discoveries. Reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community has shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in Science. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency. Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

Over the past year, Nature has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at

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.

By Michael Hilzik
October 27, 2013

NIH Tackles Irreproducibility

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

By Jef Akst | January 28, 2014

In today’s world, brimming as it is with opinion and falsehoods masquerading as facts, you’d think the one place you can depend on for verifiable facts is science.

You’d be wrong. Many billions of dollars’ worth of wrong.
The Impact of Technology

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.

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.”
Buckheit & Donoho, 1995

CSHL Keynote; Dr. Lior Pachter, UC Berkeley
The software contains “ideas that enable biology...”
Stories from the Supplement, 2013
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,a* Ying Su,2,b,a,a1 Irene Kuhn,1 Kornelia Polyak,2,a,a2,a3,a4 and Mina J. Bissell1,a
1Life Sciences Division, Lawrence Berkeley National Laboratory, Malitop 977RR255A, 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
*Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.cell.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...
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,

• …
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” (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 needed).
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
Agenda

Panel 1: Integration with the scholarly record: Case Studies and Lessons Learned (Michela Taufer)

Panel 2: Interoperability standards, proprietary codes, and verification/testing (Michael Heroux)

Panel 3: Licensing and facilitating re-use (Victoria Stodden)

Panel 4: Credit and citation standards, persistence (i.e. DOIs, repositories, embargo periods) (Kate Keahey)

Panel 5: Edge cases, specialized hardware, large or exceptionally complex code bases (Lorena Barba)

Panel 6: Minimal sharing requirements, workflows (Ewa Deelman)
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.
## Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation standards</td>
<td>Journal encourages citation of data, code, and materials—or says nothing.</td>
<td>Journal describes citation of data in guidelines to authors with clear rules and examples.</td>
<td>Article provides appropriate citation for data and materials used, consistent with journal’s author guidelines.</td>
<td>Article is not published until appropriate citation for data and materials is provided that follows journal’s author guidelines.</td>
</tr>
<tr>
<td>Data transparency</td>
<td>Journal encourages data sharing—or says nothing.</td>
<td>Article states whether data are available and, if so, where to access them.</td>
<td>Data must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Analytic methods (code) transparency</td>
<td>Journal encourages code sharing—or says nothing.</td>
<td>Article states whether code is available and, if so, where to access them.</td>
<td>Code must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Research materials transparency</td>
<td>Journal encourages materials sharing—or says nothing.</td>
<td>Article states whether materials are available and, if so, where to access them.</td>
<td>Materials must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Design and analysis transparency</td>
<td>Journal encourages design and analysis transparency or says nothing.</td>
<td>Journal articulates design transparency standards.</td>
<td>Journal requires adherence to design transparency standards for review and publication.</td>
<td>Journal requires and enforces adherence to design transparency standards for review and publication.</td>
</tr>
<tr>
<td>Preregistration of studies</td>
<td>Journal says nothing.</td>
<td>Journal encourages preregistration of studies and provides link in article to preregistration if it exists.</td>
<td>Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.</td>
<td>Journal requires preregistration of studies and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td>Preregistration of analysis plans</td>
<td>Journal says nothing.</td>
<td>Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.</td>
<td>Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.</td>
<td>Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td>Replication</td>
<td>Journal discourages submission of replication studies—or says nothing.</td>
<td>Journal encourages submission of replication studies.</td>
<td>Journal encourages submission of replication studies and conducts blind review of results.</td>
<td>Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.</td>
</tr>
</tbody>
</table>
Result and Artifact Review and Badging

An experimental result is not fully established unless it can be independently reproduced. A variety of recent studies, primarily in the biomedical field, have revealed that an uncomfortably large number of research results found in the literature fail this test, because of sloppy experimental methods, flawed statistical analyses, or failures to report results in a way that permits independent verification. The problem is not limited to the biomedical sciences, and the lack of reproducibility is a major cause of the long-standing reputation for low reliability in empirical research in many fields. The need for independent reproduction of results is, of course, not limited to the biomedical sciences; it is an essential part of the scientific method. And the need for independent verification is not limited to the research process; it is also essential in the process of preparing publications (and grants, fellowship awards, etc.).
**Terminology.**

A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms. The following are inspired by the International Vocabulary for Metrology (VIM); see the Appendix for details.

- **Repeatability** (Same team, same experimental setup)
  
  The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.

- **Replicability** (Different team, same experimental setup)
  
  The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author’s own artifacts.

- **Reproducibility** (Different team, different experimental setup)
  
  The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.
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.
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 (ie. 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.
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.
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 (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.
Computational Barriers

Barriers to Replication in Computational Science:

- rerunning same code, same parameter settings, same system can produce different results (?),

- same code (Reprozip, containerization/Docker), but updated libraries, compiler, operating system..

- software customization to underlying architectures; portability, modularity, re-usability,

- numerical stability of the underlying software architecture,

- unique hardware, scarce allocations, long runtimes..
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
- Madagascar
- nanoHUB.org
- RunMyCode.org

Workflow Tracking and Research Environments

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

Embedded Publishing

- Verifiable Computational Research
- Collage Authoring Environment
- SOLE
- SHARE
- knitR
- Sweave
Encouraging Reproducibility While Expanding Access to Massive Computation

We are at the convergence of two (ordinarily antagonistic) trends:

1. Scientific projects will become massively more computing intensive,

2. Scientific computing dramatically more transparent.

These two trends can reinforce each other: better transparency will allow people to run much more ambitious computational experiments. And better computational experiment infrastructure will allow researchers to be more transparent.
Encouraging reproducibility while expanding access to massive computation: leverage & contribute to existing cyberinfrastructure and tools to support the whole discovery story (= run-to-pub-cycle).

Organization through Working Groups. Examples needed..

CC*DNI DIBBS:
- 5 Institutions, 5 Years ($5M total).
- Cooperative Agreement.
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
Conclusions

1. Computation is near-ubiquitous in modern research.

2. Reproducibility issues travel with all computational research.

3. Cyberinfrastructure is underdeveloped and could help resolve irreproducibility if done carefully.
“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.
Proposition 1

• 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?
A Credibility Crisis

Science advances on a foundation of trusted discoveries. Reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in Science. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency. Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.

Over the past year, Nature has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at)

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.

By Michael Hiltzik

In today’s world, brimful as it is with opinion and falsehoods masquerading as facts, you’d think the one place you can depend on for verifiable facts is science.

You’d be wrong. Many billions of dollars’ worth of wrong.