Stakeholder Roles in Enabling Reproducibility in Computational Research

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

IEEE Panel of Editors Meeting: Research Reproducibility
Hollywood, CA
April 13, 2018
Conclusion: Progress on computational reproducibility is enabled through coordination by a variety of stakeholders.
Multiple group influence

Stodden 2018 (submitted)
Gridlock?

Community members need to:

1) act on their own initiative, but be mindful of the impact on other stakeholders (for their own success, as well as that of the community),

2) act in consort with other stakeholders: workshops; coordinating mechanisms such as RDA; curriculum development discussions; multidisciplinary bodies; new mechanisms.

• Community-wide solutions are likely beyond the reach of any single scientific society.
• Rulemakers can enforce 2).
Technology is driving a re-assessment of transparency irrespective of discipline

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…”

CSHL Keynote; Dr. Lior Pachter, Caltech
“Stories from the Supplement” from the Genome Informatics meeting 11/1/2013
https://youtu.be/5NiFibnbE8o
The digital age in science

Claim 1:
Virtually all published discoveries today have a computational component.

Claim 2:
There is a mismatch between the traditional scientific process and computation, leading to reproducibility concerns.
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, Ying Su, Irene Kuhn, Kornelia Polyak, and Mina J. Bissell

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 Uzt, 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 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,
- 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,
Statistical Reproducibility

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.
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

http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf
Examples and Steps
Example: LIGO

**SIGNAL PROCESSING WITH GW150914 OPEN DATA**

Welcome! This ipython notebook (or associated python script GW150914_tutorial.py) will go through some typical signal processing tasks on strain time-series data associated with the LIGO GW150914 data release from the LIGO Open Science Center (LOSC):

- [https://losc.ligo.org/events/GW150914/](https://losc.ligo.org/events/GW150914/)
- View the tutorial as a web page - [https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html/](https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html/)
- Download the tutorial as a python script - [https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.py/](https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.py/)
- Download the tutorial as iPython Notebook - [https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.ipynb/](https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.ipynb/)

To begin, download the ipython notebook, readligo.py, and the data files listed below, into a directory / folder, then run it. Or you can run the python script GW150914_tutorial.py. You will need the python packages: numpy, scipy, matplotlib, h5py.

On Windows, or if you prefer, you can use a python development environment such as Anaconda (https://www.continuum.io/why-anaconda) or Enthought Canopy (https://www.enthought.com/products/canopy/).

Questions, comments, suggestions, corrections, etc: email losc@ligo.caltech.edu

v20160208b
**BINARY BLACK HOLE SIGNALS IN LIGO OPEN DATA**

Version 1.63, 2017 Sept 11

Welcome! This IPython notebook (or associated python script LOSC_Event_tutorial.py) will go through some typical signal processing tasks on strain time-series data associated with the LIGO Event data releases from the LIGO Open Science Center (LOSC):

- Find events at [https://losc.ligo.org/events/](https://losc.ligo.org/events/).
- View the tutorial as a [web page](https://losc.ligo.org/software/) for GW150914.
- Run this tutorial with Binder using the link on the [tutorials](https://losc.ligo.org/software/) page.
- If you are running this tutorial on your own computer, see the [Download](https://losc.ligo.org/software/) section below.
- This notebook works with nbformat version 4. If you are running version 3, pick it up from the [tutorials](https://losc.ligo.org/software/) page.
- After setting the desired “eventname” below, you can just run the full notebook.

Questions, comments, suggestions, corrections, etc: email losc@ligo.caltech.edu

---

This tutorial is intended for educational purposes. The code shown here is not used to produce results papers published by the LIGO Scientific Collaboration, which instead rely on special purpose analysis software packages.

For publicly available, gravitational-wave software analysis packages that are used to produce LSC and Virgo Collaboration results papers, see [https://losc.ligo.org/software/](https://losc.ligo.org/software/).
Tutorials

Each tutorial will lead you step-by-step through some common data analysis tasks. While LIGO data can be analyzed using libraries in many software languages (C, C++, Matlab, etc.), most of these tutorials use Python. See also the software page for more examples.

See the tutorial setup page for help installing software to run these tutorials.

Tutorials shown here are not used to produce published results. For gravitational-wave software analysis packages that are used to produce LSC and Virgo Collaboration publications, see https://losc.ligo.org/software/.

Binary Black Hole Events

Use matched filtering to find signals hidden in noise.

Run: Azure | mybinder (Beta)

View: GW150914 | LVT151012 | GW151226 | GW170104

Download: zip file with data | Jupyter notebook | python script

Quickview Notebook

Make summary plots for any short segment of LIGO data.

Run: Azure | mybinder (Beta)

Download: IPython 4
Societal impact through scientific advances is predicated on discovery and new knowledge that is reliable and robust and provides a solid foundation on which further advances can be built. Unfortunately, there is evidence many published scientific results will not stand the test of time, in part due to the lack of good scientific practices for reproducibility.

Our statistical profession has a responsibility to establish publication standards that improve the transparency and robustness of what we publish and to promote awareness within the scientific community of the need for rigor in our statistical research to ensure reproducibility of our scientific results. JASA is committed to helping lead the effort by presenting solutions that can help improve research quality and reproducibility.

Starting September 1, JASA ACS will require code and data as a minimum standard for reproducibility of statistical scientific research.

New infrastructure is being established to support this initiative. Each manuscript will go through the current review process managed by an associate editor (AE), who will assign to one of the reviewers the broad evaluation of the code. A new editorial role—associate editor for reproducibility (AER)—will be added to ensure we meet a standard of reproducibility.

Reproducibility of scientific research is our ultimate goal, and the code and data requirement is a first step in that direction.
A (Very) Brief History..
Yale 2009


We collectively produced the Data and Code Sharing Declaration including a description of the problem, proposed solutions, and dream goals we’d like to see.

---

**REPRODUCIBLE RESEARCH**

**ADDRESSING THE NEED FOR DATA AND CODE SHARING IN COMPUTATIONAL SCIENCE**

By the Yale Law School Roundtable on Data and Code Sharing

Roundtable participants identified ways of making computational research details readily available, which is a crucial step in addressing the current credibility crisis.

Progress in computational science is often hampered by researchers’ inability to independently reproduce or verify published results. Attendees at a roundtable at Yale Law School (co-organized by ...) knowledge has long been scientific discovery’s central goal, yet today it’s impossible to verify most of the computational results that scientists present at conferences and in papers. To address this issue, scientists need to provide a long-term solution. We need both disciplined ways of working reproducibly and community support (and even pressure) to ensure that such disciplines are followed. On 21 November 2009, I"
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.

Set the Default to “Open”

Reproducible Science in the Computer Age. Conventional wisdom sees computing as the “third leg” of science, complementing theory and experiment. That metaphor is outdated. Computing now pervades all science. Massive computation is often required to reduce and analyze data; simulations are employed in fields as diverse as climate modeling and astrophysics. Unfortunately, scientific computing culture has not kept pace. Experimental researchers are taught early to keep notebooks or computer logs of every work detail: design, procedures, equipment, raw results, processing techniques, statistical methods of analysis, etc. In contrast, few computational experiments are performed with such care. Typically, there is no record of workflow, computer hardware and software configuration, or parameter settings. Often source code is lost. While crippling reproducibility of results, these practices ultimately impede the researcher’s own productivity.

The State of Experimental and Computational Mathematics. Experimental mathematics — application of high-performance computing technology to research questions in pure and applied mathematics, including physicists, legal scholars, journal editors, and funding agency officials representing academia, government labs, industry research, and all points in between.

Society for Industrial and Applied Mathematics

SIAM NEWS >

“Setting the Default to Reproducible” in Computational Science Research

June 3, 2013

Following a late-2012 workshop at the Institute for Computational and Experimental Research in Mathematics, a group of computational scientists have proposed a set of standards for the dissemination of reproducible research.

Victoria Stodden, Jonathan Borwein, and David H. Bailey
Issues from ICERM

• The need to carefully document the full context of computational experiments including system environment, input data, code used, computed results, etc.

• The need to save the code and data in a permanent repository, with version control and appropriate meta-data.

• The need for reviewers, research institutions, and funding agencies to recognize the importance of computing and computing professionals, and to allocate funding for after-the-grant support and repositories.

• The increasing importance of numerical reproducibility, and the need for tools to ensure and enhance numerical reliability.

• The need to encourage publication of negative results as other researchers can often learn from them.

• The re-emergence of the need to ensure responsible reporting of performance.
6: Through their policies and through the development of supporting infrastructure, research sponsors and science, engineering, technology, and medical journal and book publishers should ensure that information sufficient for a person knowledgeable about the field and its techniques to reproduce reported results is made available at the time of publication or as soon as possible after publication.

7: Federal funding agencies and other research sponsors should allocate sufficient funds to enable the long-term storage, archiving, and access of datasets and code necessary for the replication of published findings.
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 the computational steps and a failure to understand 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

Access to the computational steps taken to process data and generate findings is as important as access to data themselves.


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://

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

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

1 Center for Information Technology Research in the Interest of Society, Santa Clara, CA 95054, USA;
2 U.S. Department of Energy, Washington, DC 20585, USA;
3 Lawrence Livermore National Laboratory, Livermore, CA 94550, USA;
4 University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA;
5 National Institute of Standards and Technology, Gaithersburg, MD 20899, USA;
6 Los Alamos National Laboratory, Los Alamos, NM 87545, USA;
7 Stanford University, Stanford, CA 94305, USA;
8 University of California, Berkeley, CA 94720, USA.

The authors declare no competing financial interests.
Reproducibility Enhancement Principles

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

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.

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

4: To facilitate reuse, adequately document digital scholarly artifacts.

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

6: Use Open Licensing when publishing digital scholarly objects.

7: 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></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>
Community Infrastructure Innovations

Research Environments

- Verifiable Computational Research: knitR, Sumatra, Galaxy
- Collage Authoring Environment: Taverna, Kurator, Wavelab
- SHARE: Sweave, SOLE, GenePattern, torch.ch
- Code Ocean: Cyverse, Open Science Framework, IPOL, Whole Tale
- Jupyter: NanoHUB, Vistrails, Popper, flywheel.io

Workflow Systems

- Taverna, Wings, Pegasus
- Kurator, Kepler, Everware

Dissemination Platforms

- ResearchCompendia.org: Occam, Wavelab
- DataCenterHub: RCloud, Sparselab
- RunMyCode.org: TheDataHub.org, Reprozip
- ChameleonCloud: Madagascar
How Much of a Problem is Computational Reproducibility?
Does artifact access on demand work?

February 11, 2011:

“All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data and materials must be fulfilled….”

- Obtained a random sample of 204 scientific articles with computational findings. Asked for the data and code!

Stodden et al., “Journal Policy for Computational Reproducibility,” PNAS, March 2018
<table>
<thead>
<tr>
<th>Response</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>26%</td>
</tr>
<tr>
<td>Email bounced</td>
<td>2%</td>
</tr>
<tr>
<td>Impossible to share</td>
<td>2%</td>
</tr>
<tr>
<td>Refusal to share</td>
<td>7%</td>
</tr>
<tr>
<td>Contact to another person</td>
<td>11%</td>
</tr>
<tr>
<td>Asks for reasons</td>
<td>11%</td>
</tr>
<tr>
<td>Unfulfilled promise to follow up</td>
<td>3%</td>
</tr>
<tr>
<td>Direct back to SOM</td>
<td>3%</td>
</tr>
<tr>
<td>Shared data and code</td>
<td>36%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

24 articles provided direct access to code/data.
Replicating Computational Findings

• We deemed 56 of the 89 articles for which we had data and code potentially reproducible

• We chose a random sample of 22 from these 56 to replicate
Computational Replication Rates

We were able to obtain data and code from the authors of 89 articles in our sample of 204,

- overall **artifact recovery rate** estimate: 44% with 95% confidence interval [0.36, 0.50]

Of the 56 potentially reproducible articles, we randomly choose 22 to attempt replication, and all but one provided enough information that we were able to reproduce their computational findings.

- overall **computational reproducibility** estimate: 26% with 95% confidence interval [0.20, 0.32]
When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

The data files remains our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.

We do not typically share our internal data or code with people outside our collaboration.

The code we wrote is the accumulated product of years of effort by [redacted] and myself. Also, the data we processed was collected painstakingly over a long period by collaborators, and so we will need to ask permission from them too.

Normally we do not provide this kind of information to people we do not know. It might be that you want to check the data analysis, and that might be of some use to us, but only if you publish your findings while properly referring to us.
Thank you for your interest in our paper. For the [redacted] calculations I used my own code, and there is no public version of this code, which could be downloaded. Since this code is not very user-friendly and is under constant development I prefer not to share this code.

I’m sorry, but our computer code was not written with an eye toward distributing for other people to use. The codes are not documented and we don’t have the time or resources to document them. If you have a particular calculation you would like done and it is not a major extension of what we are presently set up to do, we might be able to run the codes for you.

R is a free software package available at www.r-project.org/ I used R for the [redacted] models. As you probably know, [redacted] and [redacted] are quite complicated. But I don’t have to tell you that given that you are a statistics student! I used Matlab for the geometry.
Our program [redacted] is available here [URL redacted] (documentation and tutorials were included)

If you go to [URL redacted], under the publications, I have a link to the GitHub repository. I don’t know if I have all of the raw simulated data, but I certainly have the processed data used to make the plots. What do you need? All of the simulated data could of course be regenerated from the code.

Please find attached a .zip file called [redacted].zip that has the custom MATLAB [redacted] analysis code. If you run MastroRunfigure-one.m this will generate several panels from the paper.

In the next email I will enclose the custom image analysis software. This can also be accessed from [URL redacted] where there is a manual and tutorial.

Please let me know if you have any troubles, or if there is anything else I can help with.
Converging Trends

Two (competing?) conjectures:
1. Scientific research will become massively more computational,
2. Scientific computing will become dramatically more transparent.

These 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.
We imagine a major effort to develop 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.

This infrastructure is used because it enables efficiency and productivity, and discovery.
Imagine: 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 2003 and list the trial name and histogram side by side.

Donoho & Gavish, “Three Dream Applications of Verifiable Computational Results,” CiSE, 2012