Computational Knowledge and the Scholarly Record

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

Barnard College
February 8, 2018
Reproducibility of Scientific Findings

- 1992 Claerbout and Karrenbach coined the phrase “reproducible research”
- My research is inspired by this notion:
  - SparseLab software (2005)
  - Reproducible Research Standard (2009)
  - User-led innovation in science; the scientific method (2010)
  - Funding Agency Policy Changes (2011)
  - Tools and Strategies for Changing the Culture (2012)
  - Testified for the House Committee on Science, Space and Technology hearing on Scientific Integrity & Transparency (2013)
  - Cyberinfrastructure and the Digital Research Environment; Statistics and Reproducibility; Self-correction in Science; ResearchCompendia software (2015)
  - Reproducibility in Computational Methods; “Whole Tale” Cyberinfrastructure (2016)

In this talk I focus on one specific result that illustrates an approach to improving machine learning research and the scholarly record.
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.

**Skepticism**: scientific claims must be exposed to critical scrutiny before being accepted.
Skepticism: Boyle’s ideas

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.
Today: Technology is driving a reassessment of transparency

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, UC Berkeley

“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.
“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”

V. Stodden, IMS Bulletin (2013)
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 empirically reproducing the experiments.
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.

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
An Abstraction to Improve Machine learning (AIM)
Query 3: The Acute Leukemia Dataset

- Introduced in Golub et al. “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring” (1999):

  “cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias [to] discover the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)”

- In joint work with Xiaomian Wu and April Tang, we carried out this scholarly record query.
Querying the Literature

We wanted:

• A **list of all classifiers** applied to the Golub dataset;
• A comparison of **their misclassification rates**.

A literature search produced 30 articles, but they did not give comparable misclassification rates.

Our next step was to create a table of comparable misclassification rates. We identified 5 articles for which this seemed possible.
Our (Naive) Expectation

We hoped to apply the machine learning algorithms from the literature to the Golub dataset, in the 5 cases we identified. However, we found that the articles implemented (at least) three steps, each varying from one article to the next:

1. data preprocessing,
2. feature selection,
3. application of machine learning algorithm.
Computational Steps in the 5 Articles
<table>
<thead>
<tr>
<th>Paper</th>
<th>Dataset Size</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72 x 6817</td>
<td>Golub Classifier: informative genes+weighted vote</td>
</tr>
<tr>
<td>2</td>
<td>72 x 6817</td>
<td>Golub Classifier: informative genes+weighted vote</td>
</tr>
<tr>
<td>3</td>
<td>72 x 7129</td>
<td>Nearest Neighbor; SVM(linear kernel, quadratic kernel); Boosting (100, 1000, 10000 iterations)</td>
</tr>
<tr>
<td>4</td>
<td>72 x 7129</td>
<td>SVM (top 25, 250, 500, 1000 features)</td>
</tr>
<tr>
<td>5</td>
<td>72 x 7070</td>
<td>MVR(median vote relevance); NBGR(naive bayes global relevance); MAR(Golub relevance)+SVM</td>
</tr>
<tr>
<td>6</td>
<td>72 x 6817</td>
<td>Logistic and Quadratic discriminant analysis</td>
</tr>
<tr>
<td>7</td>
<td>72 x 7129</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>9</td>
<td>72 x 6817</td>
<td>Linear &amp; Quadratic discriminant analysis analysis; Classification Trees; Nearest Neighbors</td>
</tr>
<tr>
<td>10</td>
<td>72 x 7129</td>
<td>Decision Trees; AdaBoost</td>
</tr>
<tr>
<td>11</td>
<td>72 x 7129</td>
<td>MAVE-LD, DLDA, DQDA, MAVE-NPLD</td>
</tr>
<tr>
<td>12</td>
<td>72 x 7129</td>
<td>SIMCA classification</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Paper/Classifier</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>---------------------</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>WeightedVote</td>
<td>.91</td>
<td>.94</td>
</tr>
<tr>
<td>NearestN</td>
<td>.97</td>
<td>.94</td>
</tr>
<tr>
<td>SVM Linear</td>
<td>.97</td>
<td>.97</td>
</tr>
<tr>
<td>SVM Quadratic</td>
<td>.97</td>
<td>.88</td>
</tr>
<tr>
<td>Adaboost</td>
<td>.91</td>
<td>.91</td>
</tr>
<tr>
<td>Logit</td>
<td>.97</td>
<td>.97</td>
</tr>
<tr>
<td>QDA</td>
<td>.94</td>
<td>.91</td>
</tr>
<tr>
<td>NearestN</td>
<td>.97</td>
<td>.91</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>.91</td>
<td>.91</td>
</tr>
<tr>
<td>Bagging</td>
<td>.94</td>
<td>.91</td>
</tr>
<tr>
<td>Bagging (CPD)</td>
<td>.74</td>
<td>.85</td>
</tr>
<tr>
<td>FFLDA</td>
<td>.88</td>
<td>.88</td>
</tr>
<tr>
<td>DLDA</td>
<td>.97</td>
<td>.94</td>
</tr>
<tr>
<td>DQDA</td>
<td>.97</td>
<td>.94</td>
</tr>
<tr>
<td>Bayes Network</td>
<td>.74</td>
<td>.88</td>
</tr>
<tr>
<td>mean</td>
<td>.92</td>
<td>.92</td>
</tr>
</tbody>
</table>
Query Conclusions

- Lengthy to obtain comparable estimates (200+ student hours)
- Many points of variability: starting dataset; preprocessing steps; feature selection methods; algorithm choice; parameter tuning...
- Details not well-captured in the traditional article, making comparisons difficult or impossible.
- Would be easier if:
  - there was prior agreement on the dataset,
  - prior agreement on hold-out data for testing,
  - full disclosure of preprocessing and feature selection steps,
  - full disclosure of algorithm application and parameter tuning.
Abstraction for Improving Machine learning (AIM)

- Agreement on datasets prior to analysis, conferences around those datasets,
- Hold-out data held by a neutral third party (e.g. NIST), not seen by researchers,
- Researchers distinguish and specify feature selection and preprocessing vs learning algorithm application,
- Send code to the third party who returns your misclassification rate on the test data.

Side effect: training data and code/algorithm shared.
How Should the Scholarly Record Look for Computational Research?

Journal Policy Effectiveness
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.**

Fostering Integrity in Research, National Academies of Sciences, Engineering, and Medicine, 2017
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 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

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
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.
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, forthcoming 2018
<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>46</td>
<td>26%</td>
</tr>
<tr>
<td>Email bounced</td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Impossible to share</td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Refusal to share</td>
<td>12</td>
<td>7%</td>
</tr>
<tr>
<td>Contact to another person</td>
<td>20</td>
<td>11%</td>
</tr>
<tr>
<td>Asks for reasons</td>
<td>20</td>
<td>11%</td>
</tr>
<tr>
<td>Unfulfilled promise to follow up</td>
<td>5</td>
<td>3%</td>
</tr>
<tr>
<td>Direct back to SOM</td>
<td>6</td>
<td>3%</td>
</tr>
<tr>
<td>Shared data and code</td>
<td>65</td>
<td>36%</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>100%</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]
Conclusion

We can see the scholarly record as a body of numerical data, and we find **publications are unstructured for analysis**

Why?

- Research is enabled by new technological forces
- Overuse of underpowered studies
- Editorial preference for positive results
- Exploitation of researcher degrees of freedom

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 of 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. While

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
<table>
<thead>
<tr>
<th>Difficulty Level</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impossible to reproduce (missing essential code, data or methodology)</td>
<td>5%</td>
</tr>
<tr>
<td>Nearly impossible to reproduce (specialized hardware, intense computation</td>
<td>14%</td>
</tr>
<tr>
<td>requirements, sensitive data, human study, or other unavoidable reasons)</td>
<td></td>
</tr>
<tr>
<td>Difficult to reproduce because of unavoidable inherent complexity (e.g. requiring</td>
<td>14%</td>
</tr>
<tr>
<td>300 million MCMC steps on each data set, or months to do runs)</td>
<td></td>
</tr>
<tr>
<td>Reproducible with substantial tedious effort (e.g. individual download of a large</td>
<td>5%</td>
</tr>
<tr>
<td>number of datasets, hand coding of data into a new format i.e. from an image,</td>
<td></td>
</tr>
<tr>
<td>many archiving steps required)</td>
<td></td>
</tr>
<tr>
<td>Reproducible with substantial intellectual effort (e.g. methods well defined but</td>
<td>5%</td>
</tr>
<tr>
<td>required some knowledge of jargon or understanding of the field; or down the</td>
<td></td>
</tr>
<tr>
<td>rabbit hole references to past articles required to reproduce; etc.)</td>
<td></td>
</tr>
<tr>
<td>Could reproduce with fairly substantial skill and knowledge (e.g. required GPU</td>
<td>23%</td>
</tr>
<tr>
<td>programming abilities to run code that wasn't given; translating complex models</td>
<td></td>
</tr>
<tr>
<td>into MATLAB code; pseudo code with functions not detailed described in text into</td>
<td></td>
</tr>
<tr>
<td>code; missing scripts)</td>
<td></td>
</tr>
<tr>
<td>Reproducible after tweaking (e.g. missing parameters required fiddling to find,</td>
<td>5%</td>
</tr>
<tr>
<td>missing modified code lines, missing arguments required for differing architecture;</td>
<td></td>
</tr>
<tr>
<td>missing minor method step)</td>
<td></td>
</tr>
<tr>
<td>Minor difficulty in reproducing (e.g. installing a specialized library, converting</td>
<td>18%</td>
</tr>
<tr>
<td>to a different computational system)</td>
<td></td>
</tr>
<tr>
<td>Straightforward to reproduce with minimal effort</td>
<td>14%</td>
</tr>
</tbody>
</table>
Meta-Analysis

- Elsevier publishes ~1,000 medical journals with ~1 million articles a year, mostly clinical findings
- Typically single-center studies with a small number of patients (e.g. n = 20)
- Meta Analysis: *aggregate across many studies*

Meta-analysis of the association between TP53 status and the risk of death at 2 years

Kyzas et al., “Selective Reporting Biases in Cancer Prognostic Factor Studies,” *JNCI*, 97(14), 2005
What Does Meta-Analysis Tell Us?

- Most published findings do not replicate
- Most published effects are inflated
- Incorrect findings have more impact than true ones e.g. negative results

Suggests an important approach: *study the scholarly record as a body of evidence*
Example from Genomics

- Late 1990’s: microarray and sequencing technology provided gene expression data for statistical analysis
- Goal was to find “candidate genes” that were related to a phenomena of interest:
  - small n studies
  - risk factors chosen from “diverse considerations”
  - use of conventional statistical tests and thresholding ($p < 0.05$)
  - studies subject to confounding and selective reporting
- Entirely replaced by Genome-Wide Association Studies (GWAS)
Recall: False Positives and False Negatives

<table>
<thead>
<tr>
<th>True Underlying Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
</tr>
<tr>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
</tr>
<tr>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>+</th>
<th>False Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Positive</td>
<td></td>
</tr>
<tr>
<td>N₁</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
<tr>
<td>N₂</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Efforts to Replicate “Candidate Gene” Association Studies Fail

Table 1 shows “at least 20 false-positive findings for every one true-positive result.”

“approximately 1000 early gene loci-phenotype associations for the conditions listed in Table 1 were false positives from the candidate-gene approach.”

“There are no documented false-negative results arising from candidate-gene studies. Therefore, for the phenotypes listed in Table 1, the numerator of the FP:FN ratio is over 1000, while the denominator is apparently 0.”

Ioannidis et al. “The false-positive to false-negative ratio in epidemiologic studies,” Epidemiology, 22(4), 2011

<table>
<thead>
<tr>
<th>First Author</th>
<th>Disease/Phenotype</th>
<th>Gene Loci Tested</th>
<th>Sample Size (Design)</th>
<th>Replicated Gene Loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bosker et al</td>
<td>Major depressive disorder</td>
<td>57</td>
<td>3540 (case-control)</td>
<td>1</td>
</tr>
<tr>
<td>Caporaso et al</td>
<td>Smoking (7 phenotypes)</td>
<td>359</td>
<td>4611 (cohort)</td>
<td>1</td>
</tr>
<tr>
<td>Morgan et al</td>
<td>Acute coronary syndrome</td>
<td>70</td>
<td>1461 (case-control)</td>
<td>0</td>
</tr>
<tr>
<td>Richards et al</td>
<td>Osteoporosis (2 phenotypes)</td>
<td>150</td>
<td>19,195 (cohort)</td>
<td>3; 9</td>
</tr>
<tr>
<td>Samani et al</td>
<td>Coronary artery disease</td>
<td>55</td>
<td>4864; 2519 (case-control)</td>
<td>1</td>
</tr>
<tr>
<td>Scuteri et al</td>
<td>Obesity (3 phenotypes)</td>
<td>74</td>
<td>6148 (cohort)</td>
<td>0</td>
</tr>
<tr>
<td>Söber et al</td>
<td>Blood pressure</td>
<td>149</td>
<td>1644; 8023 (cohort)</td>
<td>0</td>
</tr>
<tr>
<td>Wu et al</td>
<td>Childhood asthma</td>
<td>237</td>
<td>1476 (triads)</td>
<td>1</td>
</tr>
</tbody>
</table>
# Infrastructure Innovations

## Research Environments

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

## Workflow Systems

<table>
<thead>
<tr>
<th>Taverna</th>
<th>Wings</th>
<th>Pegasus</th>
<th>CDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurator</td>
<td>Kepler</td>
<td>Everware</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>
“the paper of the birch”

Researchers today are perhaps like Lewis and Clark.

Jefferson’s instructions (1803):

“The object of your mission is to explore the Missouri river, & such principal stream of it, as, by it's course & communication with the water of the Pacific ocean may offer the most direct & practicable water communication across this continent, for the purposes of commerce.”

Thanks to Dave Culler for the analogy
“Beginning at the mouth of the Missouri, you will take observations of latitude and longitude at all remarkable points on the river”

“The courses of the river between these points of observation may be supplied by the compass, the log-line & by time, corrected by the observations themselves. The variations of the compass too, in different places should be noticed”
“Your observations are to be taken with great pains & accuracy to be entered distinctly, & intelligibly for others as well as yourself, to comprehend all the elements necessary, with the aid of the usual tables to fix the latitude & longitude of the places at which they were taken, & are to be rendered to the war office, for the purpose of having the calculations made concurrently by proper persons within the U.S. Several copies of these as well as of your other notes, should be made at leisure times, & put into the care of the most trustworthy of your attendants, to guard by multiplying them against the accidental losses to which they will be exposed. A further guard would be that one of these copies be written on the paper of the birch, as less liable to injury from damp than common paper.”
“To provide, on the accident of your death, against anarchy, dispersion & the consequent danger to your party, and total failure of the enterprise, you are hereby authorised, by any instrument signed & written in your own hand, to name the person among them who shall succeed to the command”
Intellectual Property
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
Creative Commons

- Adapts the Open Source Software approach to artistic and creative digital works.
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.
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 PublIns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).

• Copyright adheres to raw facts in Europe.

• Residual copyright in data is possible (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/proprietary 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.
The Research Ecosystem
Ecosystem

- Funders (policy)
- Publishers (TOP guidelines)
- Researchers (processes)
- Scientific Societies
- Regulatory Bodies (OSTP Memos)
- Universities/institutions (hiring/promotion)
- Universities/libraries (empowering w/tools)
# 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><strong>Citation standards</strong></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><strong>Data transparency</strong></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><strong>Analytic methods (code) transparency</strong></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><strong>Research materials transparency</strong></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><strong>Design and analysis transparency</strong></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><strong>Preregistration of studies</strong></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><strong>Preregistration of analysis plans</strong></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><strong>Replication</strong></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>
Conclusion

Meta-queries on the scholarly record are hard, but they can be realized:

- Facilitated by open data/code, including meta-data, associated with publications in the scholarly record.

- Frameworks necessary for comparing answers to the same scientific questions, e.g. CompareML, Common Task Framework.

- Side effect: open data and code; transparency, verifiability
Data / Code Sharing Practices

Survey of the NIPS community:

• 1,758 NIPS registrants up to and including 2008,

• 1,008 registrants when restricted to .edu registration emails,

• After piloting, the final survey was sent to 638 registrants,
## Sharing Incentives

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>91%</td>
<td>Encourage scientific advancement</td>
</tr>
<tr>
<td>90%</td>
<td>Encourage sharing in others</td>
</tr>
<tr>
<td>86%</td>
<td>Be a good community member</td>
</tr>
<tr>
<td>82%</td>
<td>Set a standard for the field</td>
</tr>
<tr>
<td>85%</td>
<td>Improve the calibre of research</td>
</tr>
<tr>
<td>81%</td>
<td>Get others to work on the problem</td>
</tr>
<tr>
<td>85%</td>
<td>Increase in publicity</td>
</tr>
<tr>
<td>78%</td>
<td>Opportunity for feedback</td>
</tr>
<tr>
<td>71%</td>
<td>Finding collaborators</td>
</tr>
</tbody>
</table>

Survey of the Machine Learning Community, NIPS (Stodden 2010)
## Barriers to Sharing

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>Time to document and clean up</td>
<td>54%</td>
</tr>
<tr>
<td>52%</td>
<td>Dealing with questions from users</td>
<td>34%</td>
</tr>
<tr>
<td>44%</td>
<td>Not receiving attribution</td>
<td>42%</td>
</tr>
<tr>
<td>40%</td>
<td>Possibility of patents</td>
<td>-</td>
</tr>
<tr>
<td>34%</td>
<td>Legal Barriers (ie. copyright)</td>
<td>41%</td>
</tr>
<tr>
<td>-</td>
<td>Time to verify release with admin</td>
<td>38%</td>
</tr>
<tr>
<td>30%</td>
<td>Potential loss of future publications</td>
<td>35%</td>
</tr>
<tr>
<td>30%</td>
<td>Competitors may get an advantage</td>
<td>33%</td>
</tr>
<tr>
<td>20%</td>
<td>Web/disk space limitations</td>
<td>29%</td>
</tr>
</tbody>
</table>

Survey of the Machine Learning Community, NIPS (Stodden 2010)
Federal Agencies

Reliable Science: The Path to Robust Research Results

September 8, 2015

These days, much discussion about the reproducibility of scientific results seems driven by critiques of research in biomedicine and psychology. Most recently, an article in Science concluded that 60 percent of a collection of studies were not replicable. This result along with similar analyses of cancer research results has stimulated strong commentary. For example, the New York Times print edition headline about the Science article was “Psychology’s Fears Confirmed: Rechecked Studies Don’t Hold Up,” coverage that prompted a strong op-ed rebuttal tried, “Psychology Is Not in Crisis.”

Issues that arise with human subjects or with other complex living systems do not plague physical science to the same degree. However, the notion of measuring the same value of a physical quantity or the same behavior of a physical system in different laboratories at different times is central to our concept of a valid scientific result. Often the approach is not simply to replicate an experiment, but rather to get the same quantity via different paths. For example, we can measure the gravitational constant, G, with

Rigor and Reproducibility

Enhancing reproducibility through rigor and transparency: the information provided on this website is designed to assist the extramural community in addressing rigor and reproducibility in grant applications due on January 25, 2016, and beyond.

On This Page:
- News
- Goals
- Guidance: Rigor and Reproducibility in Grant Applications

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.
NSF Workshop
Systematic Approach to Robustness, Reliability, and Reproducibility in Scientific Research

February 25 - 26, 2017
Beckman Center of the National Academies of Sciences & Engineering
University of California at Irvine
100 Academy Way
Irvine, CA 92617
(949) 721-2200

The federal investment in scientific and engineering research drives innovation across our society; it also provides a foundation for national competitiveness, prosperity, and sound public policy. Recently, several prominent studies have highlighted a significant proportion of research reports, in certain fields, that are not reproducible. There is growing concern within the scientific enterprise and a loss of public trust in the reliability of science, especially the results of basic research funded by the taxpayer, is a serious issue.

The Administration, through OMB and OSTP, has directed that funding agencies, including the NSF, address these problems of irreproducibility, which includes cases where the data generated by publicly-funded research is not accessible. As part of its response to this mandate, the NSF is supporting the scientific community in efforts to find the root causes of these problems, and through extensive discussions identify ways in which they can best be solved.

Principal Investigator
David A. Weitz (Harvard University)

Workshop Leaders
Andrea Liu (University of Pennsylvania)
Wallace Marshall (UC San Francisco)
Roger D. Peng (Johns Hopkins University)
Victoria Stodden (University of Illinois)

Workshop Participants
Keith Baggerly (UTexas/MD Anderson)
Paul Chaikin (New York University)
George Fuller (UC San Diego)
Carol Hall (North Carolina State University)
Robert Hanisch (ODI, NIST)
Leslie Hatton (University of Kingston)
Amy E. Herr (UC Berkeley)
Mike Hildreth (Notre Dame)
Daniel S. Katz (University of Illinois)
Gareth H. McKinley (MIT)
Peter J. Mohr (NIST)
Jose Onuchic (Rice University)
Manish Pararashar (Rutgers University)
Steven Vigdor (Indiana University)
George Whitesides (Harvard University)
William Allen Zajc (Columbia University)

Agency Contacts
Bogdan Mihaila (NSF, Mathematical and Physical Sciences)
Gregory W. Warr (NSF, Molecular and Cellular Biosciences)
## Journal Data and Code Sharing Policies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Required as condition of publication, barring exceptions</td>
<td>10.6%</td>
<td>11.2%</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Required but may not affect editorial decisions</td>
<td>1.7%</td>
<td>5.9%</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Encouraged/addressed, may be reviewed and/or hosted</td>
<td>20.6%</td>
<td>17.6%</td>
<td>10%</td>
<td>12.4%</td>
</tr>
<tr>
<td>Implied</td>
<td>0%</td>
<td>2.9%</td>
<td>0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>No mention</td>
<td>67.1%</td>
<td>62.4%</td>
<td>82.9%</td>
<td>78.8%</td>
</tr>
</tbody>
</table>

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)