Structuring Machine Learning Research in Data Driven Science

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

Statistics Seminar
Department of Statistical Science at Cornell University
November 1, 2017
Agenda

1. Unpacking Reproducibility

2. The Research Ecosystem

3. CompareML: Reconciling Machine Learning Results

4. (if time) Intellectual Property and Openness
Unpacking Reproducibility
Merton’s Scientific Norms (1942)

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

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

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

**Originality**: scientific claims contribute something new

**Skepticism**: scientific claims must be exposed to critical scrutiny before being accepted.
Skepticism: 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 re-assessment 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)
Empirical Reproducibility

Sorting Out the FACS: A Devil in the Details

William C. Hines,1,4* Ying Su,2,3,5,4* Irene Kuhn,1 Kornelia Polyak,2,3,4,5 and Mina J. Bissell1,4
1Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 9778R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA
2Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
3Department of Medicine, Brigham and Women’s Hospital, Boston, MA 02215, USA
4Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
5These authors contributed equally to this work
*Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.cellrep.2014.02.021

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

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

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

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

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

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues.

Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

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

- False discovery, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.
- Low power, poor experimental design, nonrandom sampling,
- 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,
- ...
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

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.


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

- **Funders** (policy)
- **Publishers** (TOP guidelines)
- **Researchers** (processes)
- **Scientific Societies**
- **Regulatory Bodies** (OSTP Memos)
- **Universities/libraries** (empowering w/tools)
- **Universities/institutions** (hiring/promotion)
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 methods that were used. To address these shortcomings, researchers have proposed various initiatives to improve reproducibility. Notable among these efforts 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 any other software configuration details. 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.

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://...}

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

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.
Reproducibility Enhancement Principles

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: 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>
In a World of Radical Transparency.

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 Donoho and Gavish 2012
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 various machine learning algorithms from the literature to the Golub dataset, in the 5 cases we identified.

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

1. Train Data
   - Preprocessing
   - Feature Selection
   - Model parameter estimation

2. Test Data
   - Classification Result
   - Test Misclassification Rate

3. Data Dimensionality
   - LOOCV Used
   - LOOCV not Used

4. Train Data
   - Preprocessing
   - Feature Selection
   - Model Building

5. Test Data
   - Classification Result
   - Test Accuracy

6. Data Dimensionality
   - Repeat 200 times and using accuracy upper quartile to compare the classifiers.
   - Repeated 200 times
## Learning Algorithms (47 ALL, 25 AML)

<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>
</tbody>
</table>
## Misclassification Rates

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Feature Selection Method</th>
<th>1</th>
<th>3</th>
<th>6PCA</th>
<th>6PLS</th>
<th>9</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper1 Classifier</td>
<td></td>
<td>0.912</td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
<td>0.958</td>
<td>0.706</td>
</tr>
<tr>
<td>Paper3 NN</td>
<td></td>
<td>0.971</td>
<td>0.941</td>
<td>0.912</td>
<td>0.941</td>
<td>1</td>
<td>0.912</td>
</tr>
<tr>
<td>Paper3 SVM Linear</td>
<td></td>
<td>0.971</td>
<td>0.971</td>
<td>0.941</td>
<td>0.971</td>
<td>1</td>
<td>0.765</td>
</tr>
<tr>
<td>Paper3 SVM Quadratic</td>
<td></td>
<td>0.912</td>
<td>0.882</td>
<td>0.971</td>
<td>0.971</td>
<td>1</td>
<td>0.912</td>
</tr>
<tr>
<td>Paper3 Adaboost</td>
<td></td>
<td>0.971</td>
<td>0.912</td>
<td>0.912</td>
<td>0.971</td>
<td>0.958</td>
<td>0.941</td>
</tr>
<tr>
<td>Paper6 PCA Logit</td>
<td></td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
<td>1</td>
<td>1</td>
<td>0.853</td>
</tr>
<tr>
<td>Paper6 PCA QDA</td>
<td></td>
<td>0.971</td>
<td>0.912</td>
<td>0.941</td>
<td>1</td>
<td>1</td>
<td>0.853</td>
</tr>
<tr>
<td>Paper6 PLS Logit</td>
<td></td>
<td>0.971</td>
<td>0.882</td>
<td>0.971</td>
<td>1</td>
<td>1</td>
<td>0.853</td>
</tr>
<tr>
<td>Paper6 PLS QDA</td>
<td></td>
<td>0.971</td>
<td>0.882</td>
<td>0.971</td>
<td>1</td>
<td>1</td>
<td>0.853</td>
</tr>
<tr>
<td>Paper9 NN</td>
<td></td>
<td>0.971</td>
<td>0.912</td>
<td>0.853</td>
<td>0.971</td>
<td>0.958</td>
<td>0.971</td>
</tr>
<tr>
<td>Paper9 Decision Tree</td>
<td></td>
<td>0.912</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.917</td>
<td>0.735</td>
</tr>
<tr>
<td>Paper9 Bagging</td>
<td></td>
<td>0.971</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.958</td>
<td>0.735</td>
</tr>
<tr>
<td>Paper9 Bagging (CPD)</td>
<td></td>
<td>0.941</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.917</td>
<td>0.794</td>
</tr>
<tr>
<td>Paper9 LDA</td>
<td></td>
<td>0.912</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.958</td>
<td>0.794</td>
</tr>
<tr>
<td>Paper9 Diagonal LDA</td>
<td></td>
<td>0.941</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.958</td>
<td>0.765</td>
</tr>
<tr>
<td>Paper9 Diagonal QDA</td>
<td></td>
<td>0.912</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.958</td>
<td>0.735</td>
</tr>
<tr>
<td>Paper29 Bayesian Network</td>
<td></td>
<td>0.735</td>
<td>0.882</td>
<td>0.971</td>
<td>0.971</td>
<td>1</td>
<td>0.647</td>
</tr>
</tbody>
</table>
Query Conclusions

• Hard to respond to Query 3! (200+ student hours)

• Many points of variability: starting dataset; preprocessing steps; feature selection methods; algorithm choice; tuning of algorithm and parameters...

• 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 feature selection steps,
  ➡ full disclosure of algorithm application and parameter tuning.
The “CompareML Framework”

Adapt the Common Task Framework from Natural Language Processing: “CompareML Framework”

- 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.
Conclusion

Meta-queries on the scholarly record 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
Thank you
4. 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

• Founded in 2001, by Stanford Law Professor Larry Lessig, MIT EECS Professor Hal Abelson, and advocate Eric Eldred.

• 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 Publns 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.
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 81%</td>
</tr>
<tr>
<td>90%</td>
<td>Encourage sharing in others 79%</td>
</tr>
<tr>
<td>86%</td>
<td>Be a good community member 79%</td>
</tr>
<tr>
<td>82%</td>
<td>Set a standard for the field 76%</td>
</tr>
<tr>
<td>85%</td>
<td>Improve the calibre of research 74%</td>
</tr>
<tr>
<td>81%</td>
<td>Get others to work on the problem 79%</td>
</tr>
<tr>
<td>85%</td>
<td>Increase in publicity 73%</td>
</tr>
<tr>
<td>78%</td>
<td>Opportunity for feedback 71%</td>
</tr>
<tr>
<td>71%</td>
<td>Finding collaborators 71%</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 titled, “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 at the same quantity via different paths. For example, we can measure the gravitational constant, G, with...
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 Parashar (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)