An Overview of Reproducible Research, Uncertainty Quantification and Verification & Validation

Victoria Stodden
Department of Statistics
Columbia University

2014 SIAM Conference on Uncertainty Quantification
Savannah, Georgia
April 1, 2014
Data Science and Simulation (CSE)

- Traditional observational science increasingly computational and large scale.
- Traditional computational science and engineering addressing an increasingly large range of problems.

Increasing crossover between traditionally separate disciplines.
Data and Software

1. Digital databases require software and computing power for curation, filtering, analysis, visualization, and storage.

2. Computation requires software, and scientific investigation either produces data (e.g. simulations), or uses data as an input (e.g. data science), or both.

Data and software should be considered together as different but inextricably linked components of the computational science research process.
“Borrowing Strength”

- Lack of access to software/data implies irreproducibility.
- Reproducibility addressed in observational science. Can we learn anything to apply to computational science?
- Scientific computing assesses simulations. Can we learn anything to apply to data driven empirical science?
Reproducibility

• Reproducibility rooted in **skepticism**

• Robert Boyle and the Transactions of the Royal Society, 1660’s

• Transparency, knowledge transfer → goal to perfect the **scholarly record**.

• Advances in the technology used for scientific discovery are changing communication standards for reproducibility.
The Scientific Method

Traditionally two branches of the scientific method:

Branch 1 (deductive): mathematics, formal logic,

Branch 2 (empirical): statistical analysis of controlled experiments.

Many claim the emergence of new branches:

Branch 3,4? (computational): large scale simulations / data driven computational science.
## Observational Science

<table>
<thead>
<tr>
<th>JASA June</th>
<th>Computational Articles</th>
<th>Code Publicly Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>9 of 20</td>
<td>0%</td>
</tr>
<tr>
<td>2006</td>
<td>33 of 35</td>
<td>9%</td>
</tr>
<tr>
<td>2009</td>
<td>32 of 32</td>
<td>16%</td>
</tr>
<tr>
<td>2011</td>
<td>29 of 29</td>
<td>21%</td>
</tr>
</tbody>
</table>

Ioannidis (2011): of 500 papers studied, 9% had full primary raw data deposited.

Stodden and Seiler (to come): estimates that the computations in 27% of scientific articles published in *Science* today are reproducible.
Commonly asserted...

“...common now to consider computation as a third branch of science, besides theory and experiment.”

“We are working on a new, fourth paradigm for science based on data-intensive computing.”
The Ubiquity of Error

The central motivation for the scientific method is to root out error:

- Deductive branch: the well-defined concept of the proof,
- Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation presents only a potential third/fourth branch of the scientific method (Donoho, Stodden, et al. 2009), until the development of comparable standards.
“Really Reproducible Research” inspired by Stanford Prof. Jon Claerbout:

“The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.” David Donoho, 1998.
Credibility Crisis

Los Angeles Times

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren’t even true.

Science

Reproducibility

Marcia McNutt

Science advances on a foundation of trusted approaches that scientists use to gain confidence. Community was shaken by reports that a result was not reproducible. Because confidence in results is key, we are announcing new initiatives to ensure a sufficient signal-to-noise ratio.

The Scientist

NIH Tackles Irreproducibility

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

By Jef Akst | January 28, 2014

Nature

Announcement: Reducing our irreproducibility

Over the past year, Nature has published a string of articles that highlighted the unreliability and reproducibility of published research (collected in the special issue). This has been a wake-up call for the scientific community, which is now taking steps to improve the quality of scientific research.
Parsing Reproducibility

“Empirical Reproducibility”

“Computational Reproducibility”

“Statistical Reproducibility”

V. Stodden, IMS Bulletin (2013)
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 policy making, 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)
ICERM Workshop Report

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.
# Defining Reproducibility

<table>
<thead>
<tr>
<th>ICERM Criterion</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reviewable</strong></td>
<td>The descriptions permit the research methods to be independently assessed and the results judged credible.</td>
</tr>
<tr>
<td><strong>Confirmable</strong></td>
<td>The main conclusions of the research can be attained independently without the use of software provided by the author (using the complete description of algorithms and methodology provided).</td>
</tr>
<tr>
<td><strong>Replicable</strong></td>
<td>Tools are made available that would allow one to duplicate the results of the research.</td>
</tr>
<tr>
<td><strong>Auditable</strong></td>
<td>Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks.</td>
</tr>
<tr>
<td><strong>Reproducible</strong></td>
<td>Auditable research made openly available. This comprises well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.</td>
</tr>
</tbody>
</table>
Stability and Robustness of Results

• Computational reproducibility addresses whether fixed codes/data can replicate findings, permitting the reconciliation of differences in independent efforts.

• does not directly address whether these findings improve our understanding of the world.

• we might expect that for such findings, repeated independent replications yield results that are “close.” Possible sources of variation (B. Yu, 2013):
  
  • *Stability*: “reasonable” perturbations in the underlying data.
  
  • *Robustness*: perturbations in methods (due to changes in the parametrization, model, or model assumptions).
“Borrowing Strength” 2

• MATHEMATICAL MODEL: A collection of mathematical constructs that describe a system: a mathematical representation of the essential aspects of a system which presents the knowledge of the system in a usable form, thus a mathematical representation of a theory and of observations.

• COMPUTATIONAL MODEL: A discretization (or corruption) of a mathematical model designed to render it to a form that can be processed by computing devices.

(much of this section is due to J.T. Oden)
• **VERIFICATION:** The process of determining the accuracy with which a computational model can produce results deliverable by the mathematical model on which it is based:

  ➜ Code verification

  ➜ Solution verification

• **VALIDATION:** The process of determining the accuracy with which a model can predict observed physical events (or the important features of a physical reality).

  P. Roache (2009): “The process of determining the degree to which a model (and its associated data) is an accurate representation of the real world from the perspective of the intended uses of the model”.
Uncertainty Quantification

Establish quantitative confidence in computational results.

- identify sources of error: e.g. parameter choices, boundary conditions, edge effects, sampling uncertainty, initial conditions, includes discretization error and model uncertainty.

- create a probabilistic resolution of sources of uncertainty to understand their impact on system behavior: e.g. maximum likelihood to estimate error.
ICERM Reporting Approach

• Omitted information regarding the implementation of the experiment (e.g. parameters, parameter estimation) as source of irreproducibility.

• Introduces uncertainty when interpreting or re-estimating results.

• Addressed in ICERM workshop report.
<table>
<thead>
<tr>
<th><strong>Criterion</strong></th>
<th><strong>Definition</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assertions (#1)</td>
<td>A precise statement of assertions to be made in the paper.</td>
</tr>
<tr>
<td>Comp. Approach (#2)</td>
<td>A statement of the computational approach, and why it constitutes a rigorous test of the hypothesized assertions.</td>
</tr>
<tr>
<td>Software Cited (#3 &amp; 4)</td>
<td>Complete statements of, or references to, every algorithm employed, and salient details of auxiliary software (both research and commercial software) used in the computation.</td>
</tr>
<tr>
<td>Hardware Discussed (#5)</td>
<td>Salient details of the test environment, including hardware, system software and the number of processors utilized.</td>
</tr>
<tr>
<td>Analysis (#6)</td>
<td>Salient details of data reduction and statistical analysis methods.</td>
</tr>
<tr>
<td>Parameters Given (#7)</td>
<td>Were necessary run parameters given?</td>
</tr>
<tr>
<td>Parameters Discussed (#7)</td>
<td>Discussion of the adequacy of parameters such as precision level and grid resolution.</td>
</tr>
<tr>
<td>Results (#8)</td>
<td>Full statement (or at least a valid summary) of experimental results.</td>
</tr>
<tr>
<td>Available Code (#10)</td>
<td>Availability of computer code, input data and output data, with some reasonable level of documentation.</td>
</tr>
<tr>
<td>Functions Calls</td>
<td>Which precise functions were called, with what settings?</td>
</tr>
<tr>
<td>Comp. Instructions (#12)</td>
<td>Instructions for repeating computational experiments described in the paper.</td>
</tr>
<tr>
<td>Alternate Avenues (#14)</td>
<td>Avenues of exploration examined throughout development, including information about negative findings.</td>
</tr>
<tr>
<td>Citation (#15)</td>
<td>Proper citation of all code and data used, including that generated by the authors.</td>
</tr>
</tbody>
</table>
Sources of Error

Statistics has a long history of identifying and controlling for error in empirical experiments, e.g.:

• sampling errors,
• selection bias,
• model specification errors, functional form misspecification,
• omitted variable bias.
Nature Checklist 2013

In April 2013 Nature enacted the following checklist:

• Each figure legend should contain, for each panel where they are relevant:

1. the exact sample size (n);

2. a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates;

3. a statement of how many times the experiment was replicated in the laboratory;
4. definitions of statistical methods and measures:

- statistical test results, e.g., P values;

- very common tests, such as t-test, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;

- are tests one-sided or two-sided?

- are there adjustments for multiple comparisons?

- definition of ‘center values’ as median or average; definition of error bars as s.d. or s.e.m.
• Statistics and General Methods:
  
  • How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?
  
  • Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?
  
  • If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it.
  
  • If the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, state the extent of blinding.
  
  • For every figure, are statistical tests justified as appropriate?
• Data Deposition:

• 17. Provide accession codes for deposited data.

• Data deposition in a public repository is mandatory for: Protein, DNA and RNA sequences, Macromolecular structures, Crystallographic data for small molecules, Microarray data;

• Deposition is strongly recommended for many other datasets for which structured public repositories exist; more details on our data policy are available here. We encourage the provision of other source data in supplementary information or in unstructured repositories.

• 18. Is computer source code provided with the paper or deposited in a public repository? If so, indicate how it can be obtained.
Science Checklist 2014

In January 2014 Science enacted new policies. Check for:

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

Statisticians added to the Board of Reviewing Editors.
Reproducibility at Scale

Scale Issues: both for large datasets and compute time.

• data produced by code - making the code available permit data regeneration, but may involve prohibitive runtimes.

• partial data precomputation - what standards? partial results checking and testing. data synopses, simulated data.

• do some research areas pose extra difficulties?
The reproducibility@XSEDE workshop is a full-day event scheduled for Monday, July 14, 2014 in Atlanta, GA. The workshop will take place in conjunction with XSEDE14 (conferences.xsede.org), the annual conference of the Extreme Science and Engineering Discovery Environment (XSEDE), and will feature an interactive, open-ended, discussion-oriented agenda focused on reproducibility in large-scale computational science. Consistent with the overall XSEDE14 conference theme, we seek to engage participants from a broad range of backgrounds, including practitioners whose computational interests extend beyond traditional modeling and simulation as well as decision-makers and other professionals whose work informs and determines the direction of computation-enabled research. We hope to help...
Synthesizing Approaches..

- Proposal 1: Augment UQ with ICERM criteria.
- Proposal 2: Augment UQ with statistical criteria.
- Proposal 3: Augment reproducibility with UQ.
- Proposal 4: Operationalize augmented UQ as a guiding FAQ for computational science.
References


National Science Board Report


NAS Data Sharing Report

• **Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences, (2003)**

• “Principle 1. Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims.”
IOM “Evolution of Translational Omics: Lessons Learned and the Path Forward”

- March 23 2012, Institute of Medicine releases report,
- Recommends new standards for omics-based tests, including a fixed version of the software, expressly for verification purposes.
The fully specified computational procedures are locked down in the discovery phase and should remain unchanged in all subsequent development steps.
Intellectual Property Barriers

- Software is both copyrighted (by default) and patentable.

- Copyright: author sets terms of use using an open license:
  - Attribution only (ie. Modified BSD, MIT license, LGPL)
  - Reproducible Research Standard (Stodden 2009)

- Patents: Bayh-Dole (1980) vs reproducible research (Stodden 2012)
  - delays, barriers to software access
  - Bilski v Kappos (2011)
Legal Barriers: Copyright

“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)

- 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

Exceptions and Limitations: Fair Use.
Responses Outside the Sciences 1: Open Source Software

- 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
Responses Outside the Sciences 2: 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.
Response from Within the Sciences

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.

Winner of the Access to Knowledge Kaltura Award 2008
Copyright and Data

• Copyright adheres to raw facts in Europe.

• 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)).

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

• Law doesn’t match reality on the ground: What constitutes a “raw” fact anyway?
Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices),
- Bayh-Dole Act 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.
- greatest impact in biomedical research collaborations and drug discovery.
Impact of Computation

In the computational sciences, disclosure of data and code are considered essential for reproducibility. Software can be patent-eligible (Bilski v. Kappos 130 S. Ct. 3218 2010), increasing the reach of Bayh-Dole in the sciences.

Universities can claim ownership over software developed in the course of research on this basis and potentially patent then license access to the code.

Hypothesis: The Bayh-Dole Act inhibits reproducibility in the computational sciences, and is a barrier to access to research inventions.
Ownership of Research Codes

Patent and Copyright Agreement for Personnel at Stanford - SU18

I understand that, consistent with applicable laws and regulations, Stanford University is governed in the handling of intellectual property by its official policies titled Inventions, Patents and Licensing and Copyright Policy (both published in the Research Policy Handbook), and I agree to abide by the terms and conditions of those policies, as they may be amended from time to time.

Pursuant to those policies, and in consideration of my employment by Stanford, the receipt of remuneration from Stanford, participation in projects administered by Stanford, access to or use of facilities or resources provided by Stanford and/or other valuable consideration, I hereby agree as follows:

1. I will disclose to Stanford all potentially patentable inventions conceived or first reduced to practice in whole or in part in the course of my University responsibilities or with more than incidental use of University resources. I hereby assign to Stanford all my right, title and interest in such patentable inventions and to execute and deliver all documents and do any and all things necessary and proper on my part to effect such assignment. (See Inventions, Patents and Licensing for further clarification and discussion related to this paragraph.)

2. I am free to place my inventions in the public domain as long as in so doing neither I nor Stanford violates the terms of any agreements that governed the work done.

3. Stanford policy states that all rights in copyright shall remain with the creator unless the work:
   a. is a work-for-hire (and copyright therefore vests in the University under copyright law),
   b. is supported by a direct allocation of funds through the University for the pursuit of a specific project,
   c. is commissioned by the University,
   d. makes significant use of University resources or personnel, or
   e. is otherwise subject to contractual obligations.

I hereby assign or confirm in writing to Stanford all my right, title and interest, including associated copyright, in and to copyrightable materials falling under a) through e), above.

4. I am now under no consulting or other obligations to any third person, organization or corporation in respect to rights in inventions or copyrightable materials which are, or could be reasonably construed to be, in conflict with this agreement.

NOTE: An alternative to this agreement may be appropriate for personnel with a prior existing and conflicting employment agreement that establishes a right to intellectual property in conflict with Stanford policies. Personnel in this situation should contact the Office of the Vice Provost and Dean of Research.

5. I will not enter into any agreement creating copyright or patent obligations in conflict with this agreement.

6. This agreement is effective on the later of July 1, 2011 (on the one hand) or my date of hire, enrollment, or participation in projects administered by Stanford (on the other hand), and is binding on me, my estate, heirs and assigns.

Electronic Signature in AXESS
http://axess.stanford.edu

The signer should make a copy of this agreement for his or her own records, and hereby waives any objection to Stanford’s use of an electronic version of this agreement as a substitute for the original for any legally recognized purpose.

July 2011

Provider: Office of the Vice Provost and Dean of Research, Stanford University
Contact: Assistant Dean of Research
Last updated: July 2011
Disclosure of Research Codes

Claim: Codes would (eventually) be fully open in the absence of Bayh-Dole:

- Grassroots “Reproducible Research” movement in computational science (policy development, best practices, tool development),
- Changes in funding agency requirements
- Changes in journal publication requirements
“Reproducible Research” is Grassroots

- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- Open Science Framework / Reproducibility Project in Psychology
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: “The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer”
- SIAM CSE 2011: “Verifiable, Reproducible Computational Science”
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials”
- ...

...
Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research

Victoria Stodden
Columbia University - Department of Statistics

Sheila Miguez
Columbia University

Abstract:
Scholarly dissemination and communication standards are changing to reflect the increasingly computational nature of scholarly research, primarily to include the sharing of the data and code associated with published results. This paper presents a formalized set of best practice recommendations for computational scientists wishing to disseminate reproducible research, facilitate innovation by enabling data and code re-use, and enable broader communication of the output of digital scientific research. We distinguish two forms of collaboration to motivate choices of software environment for computational scientific research. We also present these Best Practices as a living, evolving, and changing document on wiki.
FAQ 3

- FAQ 2 + IP issues (RRS + book)
References

- J.T. Oden
- Barth
- RRS
Funding Agency Policy
Sharing: Funding Agency Policy

- NSF grant guidelines: “NSF ... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.” (2005 and earlier)

- NSF peer-reviewed Data Management Plan (DMP), January 2011.

- NIH (2003): “The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.” (>$500,000, include data sharing plan)
NSF Data Management Plan

“Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled ‘Data Management Plan.’ This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.” (http://www.nsf.gov/bfa/dias/policy/dmp.jsp)

Software management plans appearing.. (BigData joint NSF/NIH solicitation)
Inclusiveness For Research Products

- As of January 2013, the NSF permits the listing of “products” in submitted biosketches, not just publications.
- “products are…including but not limited to publications, data sets, software, patents, and copyrights.”
- Accessible and citable.
2013: Open Science in DC

- Feb 22: Executive Memorandum directing federal funding agencies to develop plans for public access to data and publications.

- May 9: Executive Order directing federal agencies to make their data publicly available.
Science Policy in U.S. Congress

- America COMPETES due to be reauthorized, drafting underway.
- Hearing on Research Integrity and Transparency by the House Science, Space, and Technology Committee (March 5).
- Reproducibility cannot be an unfunded mandate.
Journal Policy
Measuring Advances

- Journal Policy setting study design:

- Select all journals from ISI classifications “Statistics & Probability,” “Mathematical & Computational Biology,” and “Multidisciplinary Sciences” (this includes Science and Nature).

- N = 170, after deleting journals that have ceased publication.

- Create dataset with ISI information (impact factor, citations, publisher) and supplement with publication policies as listed on journal websites, in June 2011 and June 2012.
## Journal Data Sharing Policy

<table>
<thead>
<tr>
<th>Requirement</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required as condition of publication, barring exceptions</td>
<td>10.6%</td>
<td>11.2%</td>
</tr>
<tr>
<td>Required but may not affect editorial decisions</td>
<td>1.7%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Encouraged/addressed, may be reviewed and/or hosted</td>
<td>20.6%</td>
<td>17.6%</td>
</tr>
<tr>
<td>Implied</td>
<td>0%</td>
<td>2.9%</td>
</tr>
<tr>
<td>No mention</td>
<td>67.1%</td>
<td>62.4%</td>
</tr>
</tbody>
</table>

Source: Stodden, Guo, Ma (2013) PLoS ONE, 8(6)
## Journal Code Sharing Policy

<table>
<thead>
<tr>
<th>Requirement</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Required as condition of publication, barring exceptions</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Required but may not affect editorial decisions</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Encouraged/addressed, may be reviewed and/or hosted</td>
<td>10%</td>
<td>12.4%</td>
</tr>
<tr>
<td>Implied</td>
<td>0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>No mention</td>
<td>82.9%</td>
<td>78.8%</td>
</tr>
</tbody>
</table>

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

• Changemakers are journals with high impact factors.

• Progressive policies are not widespread, but being adopted rapidly.

• Close relationship between the existence of a supplemental materials policy and a data policy.

• No statistically significant relationship between data and code policies and open access policy.

• Data and supplemental material policies appear to lead software policy.
Barriers to Journal Policy Making

- Standards for code and data sharing,
- Meta-data, archiving, re-use, documentation, sharing platforms, citation standards,
- Review, who checks replication pre-publication, if anyone,
- Burdens on authors, especially less technical authors,
- Evolving, early research; affects decisions on when to publish,
- Business concerns, attracting the best papers.
Researcher Incentives
Data / Code Sharing Practices

Survey of the NIPS (Machine Learning) 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,
- 37 bounces, 5 away, and 3 in industry, gave a final response rate was 134 of 593 or 23%.
- Queried about reasons for sharing or not sharing data/code associated with their NIPS paper.
## Sharing Incentives

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>91%</td>
<td>81%</td>
</tr>
<tr>
<td>90%</td>
<td>79%</td>
</tr>
<tr>
<td>86%</td>
<td>79%</td>
</tr>
<tr>
<td>82%</td>
<td>76%</td>
</tr>
<tr>
<td>85%</td>
<td>74%</td>
</tr>
<tr>
<td>81%</td>
<td>79%</td>
</tr>
<tr>
<td>85%</td>
<td>73%</td>
</tr>
<tr>
<td>78%</td>
<td>71%</td>
</tr>
<tr>
<td>71%</td>
<td>71%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Barriers to Sharing</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)
Tools and CyberInfrastructure
Openness in Science

• Science Policy must support scientific ends: Reliability and accuracy of the scientific record.

• Facilitate Reproducibility - the ability to regenerate published results (data and code availability, alongside results).

• Need infrastructure to facilitate (1):
  1. deposit/curation of versioned data and code,
  2. link to published article,
  3. permanence of link.
Constructing Policy

• “Open Data” is not well-defined. Scope: Share data and code that permit others in the field to replicate published results. (traditionally done by the publication alone). Corollary: maximizes data reuse.

• Data and code availability at the time of publication.

• Public access. “With many eyeballs, all bugs are shallow.”

• Need infrastructure/software tools to facilitate (2):
  1. data/code suitable for sharing, created during the research process.
Tools for Computational Science

• Dissemination Platforms:
  - ResearchCompendia.org
  - MLOSS.org
  - Open Science Framework
  - IPOL
  - thedatahub.org
  - Madagascar
  - nanoHUB.org
  - RunMyCode.org

• Workflow Tracking and Research Environments:
  - VisTrails
  - Kepler
  - CDE
  - IPython Notebook
  - Galaxy
  - GenePattern
  - Paper Mâché
  - Taverna
  - Pegasus

• Embedded Publishing:
  - Verifiable Computational Research
  - Sweave
  - knitR
  - Collage Authoring Environment
  - SHARE
Research Compendia

Pilot project: improve understanding of reproducible computational science, trace sources of error.

- link data/code to published claims, re-use,
- research produces a guide to empirical researchers, certifies results,
- large scale validation of findings,
- stability, sensitivity checks.
Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items

Mark Dingemanse, Francisco Torreira, N. J. Enfield, Johan J. Bolhuis

Code and Data Abstract

A word like Huh?–used as a repair initiator when, for example, one has not clearly heard what someone just said–is found in roughly the same form and function in spoken languages across the globe. We investigate it in naturally occurring conversations in ten languages and present evidence and arguments for two distinct claims: that Huh? is universal, and that it is a word. In support of the first, we show that the similarities in form and function of this interjection across languages are much greater than expected by chance. In support of the second claim we show that it is a lexical, conventionalised form that has to be learnt, unlike grunts or emotional cries. We discuss possible reasons for the cross-linguistic similarity and propose an account in terms of convergent evolution. Huh? is a universal word not because it is innate but because it is shaped by selective pressures in an interactional environment that all languages share: that of other-initiated repair. Our proposal enhances evolutionary models of language change by suggesting that conversational infrastructure can drive the convergent cultural evolution of linguistic items.
A proof of concept for a research compendia webapp http://researchcompendia.org — Edit
A Grassroots Movement

- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- Open Science Framework / Reproducibility Project in Psychology
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: “The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer”
- SIAM CSE 2011: “Verifiable, Reproducible Computational Science”
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials”
- ...
References


• “Reproducible Research,” guest editor for Computing in Science and Engineering, July/August 2012.


available at http://www.stodden.net