Enhancing Reproducibility for Computational Methods

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

“Toward an Open Science Enterprise”
National Academies of Science, Engineering, and Medicine'
July 20, 2017

Take Home Message

1. For any policy, clarify the goal (e.g. reproducibility),

2. Stay in scope: sharing artifacts necessary for computational reproducibility (e.g. reusable code, data, workflows),

3. Coordinate with stakeholders (institutions, journals, funding bodies, regulatory bodies and agencies, libraries, societies, researchers, the public),

4. Enforce, react, change, Enforce, react, change, …

Take Home Message 2

Ideas to move toward computational reproducibility:

1. Transparency policy e.g. TOP Guidelines for journals, Stodden V, Guo P, Ma Z (2013) *Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals*. PLoS ONE 8(6),

2. Grant set asides to support an ecosystem,

3. Compare *workflows*, not results.

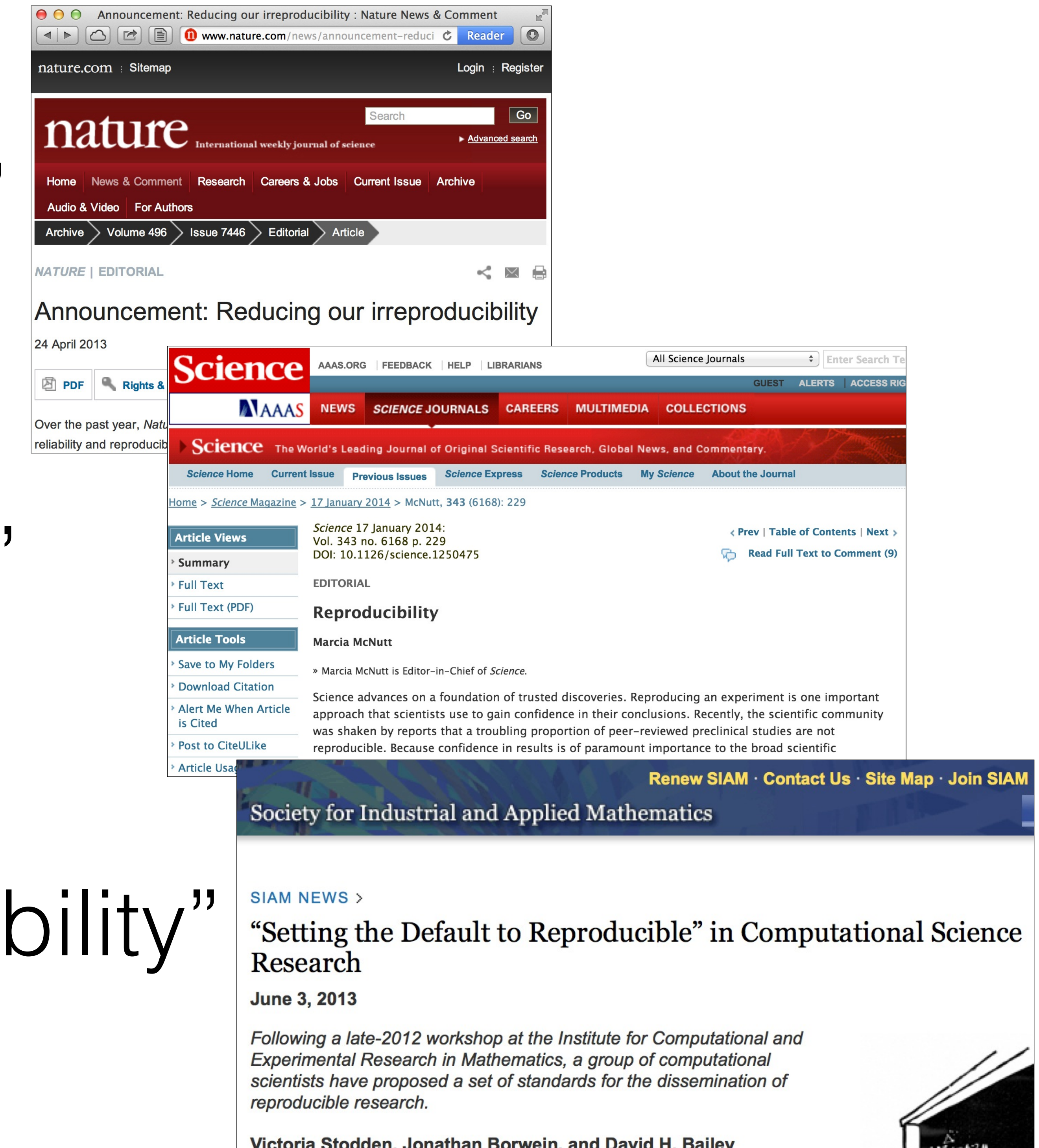
4. Leadership in Intellectual Property policy.

1. The Goal: Reproducibility

"Empirical Reproducibility"

"Statistical Reproducibility"

"Computational Reproducibility"



V. Stodden, IMS Bulletin (2013)

Computational Reproducibility

Traditionally two branches to the scientific method:

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

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

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 et al. 2009), until the development of comparable standards.

Really Reproducible Research

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

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

Note the difference between: reproducing the computational steps and, replicating the experiments independently including data collection and software implementation. (Both required)

Evidence: Requesting Artifacts

	Science	**JCP**
Shared Data and Code	36%	18%
Contact Another Person	11%	3%
Asked for Reasons	11%	2%
Refusal to Share	7%	31%
Directed back to Supplement	3%	1%
Unfulfilled promise to follow up	3%	3%
Impossible to share	2%	9%
Email bounced	2%	0%
No response	26%	32%
	n=170	n=147

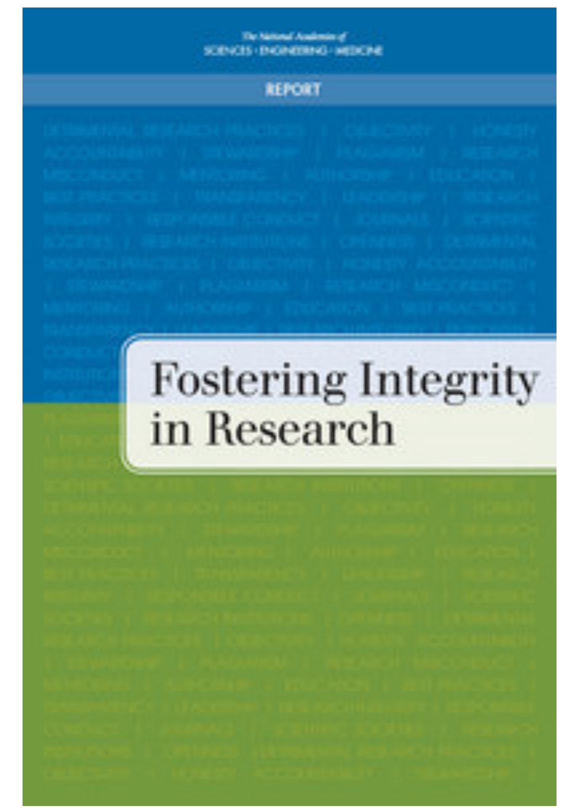
Evidence: Reusing Artifacts

- For Science 56 articles were deemed “potentially reproducible.”
- We attempted replication for a random sample of 22 of the 56. In all but one the computations replicated. Estimate 53 of 170 would replicate (~31%).
- Work on JCP in progress this summer (lower replication rates so far).

2. Stay in Scope

Fostering Integrity in Research

RECOMMENDATION SIX: 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.



RECOMMENDATION SEVEN: 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

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 Taufer[8]**

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 disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general proposals from the Transparency and Openness Promotion (TOP) guidelines (*1*) and recommendations for field data (*2*), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers repre- 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 results from multiple studies.

RECOMMENDATIONS

Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable independent regeneration of computational 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

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://bit.ly/2fVwjPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly objects, citation should be standard practice. 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 addi-

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 e.g. Reproducible Research Standard (Stodden 2009).

7: To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.

Reporting Templates

AMSTATNEWS
The Membership Magazine of the American Statistical Association

HOME | ABOUT | EDITORIAL CALENDAR | PDF ARCHIVES | ADVERTISE

Home » Additional Features, Featured, News and Announcements

Reproducible Research in *JASA*

1 JULY 2016 910 VIEWS 3 COMMENTS

Montse Fuentes, Coordinating Editor of JASA and Editor of JASA ACS

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

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

Starting September 1, *JASA ACS* will require code and data as a minimum standard for reproducibility of statistical scientific research. New infrastructure is being established to support this initiative. Each manuscript will go through the current review process managed by an associate editor (AE), who will assign to one of the reviewers the broad evaluation of the code. A new editorial role—associate editor for reproducibility (AER)—will be added to ensure we meet a standard of reproducibility.

Reproducibility of scientific research is our ultimate goal, and the code and data requirement is a first step in that direction.

A. Artifact description

A.1 Abstract

*The artifact contains all the executables of the current existing graph primitives in Gunrock's latest version on github, as well as the shell scripts of running them. It can support the runtime and/or edge throughput results in Table 3 of our PPoPP'2016 paper **Gunrock: A High-Performance Graph Processing Library on the GPU**. To validate the results, run the test scripts and check the results piped in the according text output files.*

A.2 Description

A.2.1 Check-list (artifact meta information)

- **Algorithm: breadth-first search, single-source shortest path, betweenness centrality, Pagerank, connected component**
- **Program: CUDA and C/C++ code**
- **Compilation: Host code: gcc 4.8.4 with the -O3 flag; device code: nvcc 7.0.27 with the -O3 flag**
- **Binary: CUDA executables**
- **Data set: Publicly available matrix market files**
- **Run-time environment: Ubuntu 12.04 with CUDA and GPU Computing SDK installed**
- **Hardware: Any GPU with compute capability $\geq$ 3.0 (Recommended GPU: NVIDIA K40c GPU)**
- **Output: Runtime and/or edge throughput**
- **Experiment workflow: Git clone project; download the datasets; run the test scripts; observe the results**
- **Publicly available?: Yes**

A.2.2 How delivered

Gunrock is an open source library under Apache 2.0 license and is hosted with code, API specifications, build instructions, and design documentations on Github.

A.2.3 Hardware dependencies

Gunrock requires NVIDIA GPU with the compute capability of no less than 3.0.

A.2.4 Software dependencies

Gunrock requires Boost (for CPU reference) and CUDA with version no less than 5.5. Gunrock has been tested on Ubuntu 12.04/14.04, and is expected to run correctly under other Linux distributions.

A.2.5 Datasets

All datasets are either publicly available or generated using standard graph generation software. Users will be able to run script to get these datasets once they built Gunrock code. The rgg graph is generated by Gunrock team. The download link is provided here: `https://drive.google.com/uc?export=download&id=0Bw6LwCuER0a3VWNrVUV6eTZyeFU`. Please located the unzipped `rgg_n_2_24_s0.mtx` file under `gunrock_dir/datasets/large/rgg_n_2_24_s0/`. Users are welcom to try other datasets or generate rgg/R-MAT graphs using the command line option during the test. We currently only support matrix market format files as input.

A.3 Installation

Follow the build instruction on Gunrock's github page (`http://gunrock.github.io/`), users can build Gunrock and generate the necessary executables for the experiments.

A.4 Experiment workflow

For the convenience of the artifact evaluation, we provide a series of shell scripts which run the graph primitives we have described in the paper and store the results in the output text files. Below are the steps to download Gunrock code, build, run the experiments, and observe the results.

- Clone Gunrock code to the local machine:

```
$ git clone https://github.com/gunrock/gunrock.git
$ cd gunrock
$ git submodule init && git submodule update
```

- Use CMake to build Gunrock. Make sure that boost and CUDA is correctly installed before this step:

```
$ cd /path/to/gunrock/../
$ mkdir gunrock_build && cd gunrock_build
$ cmake ../gunrock/
$ make -j16
```

The last comand will build Gunrock's executables under `gunrock_build/bin` and shared library under `gunrock_build/lib`.

- Prepare the dataset. First step into Gunrock directory:

```
$ cd /path/to/gunrock/
$ cd dataset/large/ && make
```

This will download and extract all the large datasets, including the 6 datasets in the paper.

- Step into the test script directory and run scripts for five graph primitives:

```
$ cd ../test-scripts
$ sh ppopp16-test.sh
```

- Observe the results for each dataset under five directories: BFS, SSSP, BC, PR, and CC.

A.5 Evaluation and expected result

For BFS and SSSP, the expected results include both runtime and edge throughput. For BC, Pagerank, and CC, the expected results contain runtime only.

A.6 Notes

To know more about our library, send feedback, or file issues, please visit our github page (`http://gunrock.github.io/`).

Example from https://arxiv.org/abs/1501.05387

3. Coordinate with Stakeholders

Actual question..

“[My Federal agency] is struggling with a lot of the same questions as the broader community around reproducible science. Are there particular groups that you would recommend we follow to keep track of progress that is being made?”

Infrastructure Solutions

Research Environments

Verifiable Computational Research	SHARE	Code Ocean	Jupyter
knitR	Sweave	Cyverse	NanoHUB
Collage Authoring Environment	SOLE	Open Science Framework	Vistrails
Sumatra	GenePattern	IPOL	Popper
Galaxy	torch.ch	Whole Tale	

Workflow Systems

Taverna	Wings	Pegasus	CDE	binder.org
Kurator	Kepler	Everware	Reprozip	

Dissemination Platforms

ResearchCompendia.org	DataCenterHub	RunMyCode.org	ChameleonCloud
Occam	RCloud	TheDataHub.org	Madagascar

The Convergence of Two Trends

Two (ordinarily antagonistic) trends are converging:

- ➡ Across all disciplines scientific projects will become massively more computing intensive,
- ➡ Research computing will become dramatically more transparent.

These are reinforcing trends, whose resolution is essential for verifying and comparing findings.

We will compare at the workflow level.

“Experiment Definition Systems”

- Define and create “Experiment Definition Systems” to (easily):
  - manage the conduct of massive computational experiments and
  - expose the resulting data for analysis and structure the subsequent data analysis

- The two trends need to be addressed simultaneously:
  - better transparency will allow people to run much more ambitious computational experiments,
  - *and* better computational experiment infrastructure will allow researchers to be more transparent.

Proposition

- Develop a new infrastructure that promotes good scientific practice downstream like transparency and reproducibility.
- But plan for people to use it not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work enabling *efficiency* and *productivity*, and *discovery*.

In Support of Reproducibility

- Enable (automated) capture of crucial computational information,
- Licensing for re-use and reproducibility,
- Persistence and archiving of artifacts,
- Discoverability and linking to specific scientific claims (not releasing software packages).

Problem: Two Paths

Currently there is a distribution of largely unconnected scholarly objects in various repositories, with different ownership structures.

Some repositories are institutional or federally funded, some are owned by publishers e.g. figshare, Mendeley.

Inducing a Reproducibility Industry by Grant Set-asides

- Previously, NIH required that clinical trials hire Biostatistician PhD's to design and analyze experiments. This set-aside requirement directly transformed clinical trials practice and resulted in much more good science being done. It also spawned the modern field of Biostatistics, by creating a demand for a specific set of services and trained people who could conduct them.

- Why not try a similar idea for reproducibility?

Reproducible Research Standard

Legal Issues in Software

Intellectual property is associated with 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: Intellectual property law is a poor fit with scholarly norms, and require action from the research community to enable re-use, verification, reproducibility, and support the acceleration of scientific discovery.

Copyright

- Original expression of ideas falls under copyright by default (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original
- limited time: generally life of the author +70 years
- Exceptions and Limitations: e.g. Fair Use.

Patents

Patentable subject matter: "*new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof*" (35 U.S.C. §101) that is

1. *Novel*, in at least one aspect,
2. *Non-obvious*,
3. *Useful*.

USPTO Final Computer Related Examination Guidelines (1996) "A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena" (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.

Legal Issues in Data

- In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).

- Copyright adheres to raw facts in Europe.

- Legal mismatch: What constitutes a “raw” fact anyway?

Privacy and Data

- HIPAA, FERPA, IRB mandates create legally binding restrictions on the sharing human subjects data (see e.g. http://www.dataprivacybook.org/)
- Potential privacy implications for industry generated data.
- Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..

Licensing in Research
Background: Open Source Software

Innovation: Open Licensing

➡ Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

Hundreds of open source software licenses:

- GNU Public License (GPL)
- (Modified) BSD License
- MIT License
- Apache 2.0 License
- ... see http://www.opensource.org/licenses/alphabetical

The Reproducible Research Standard

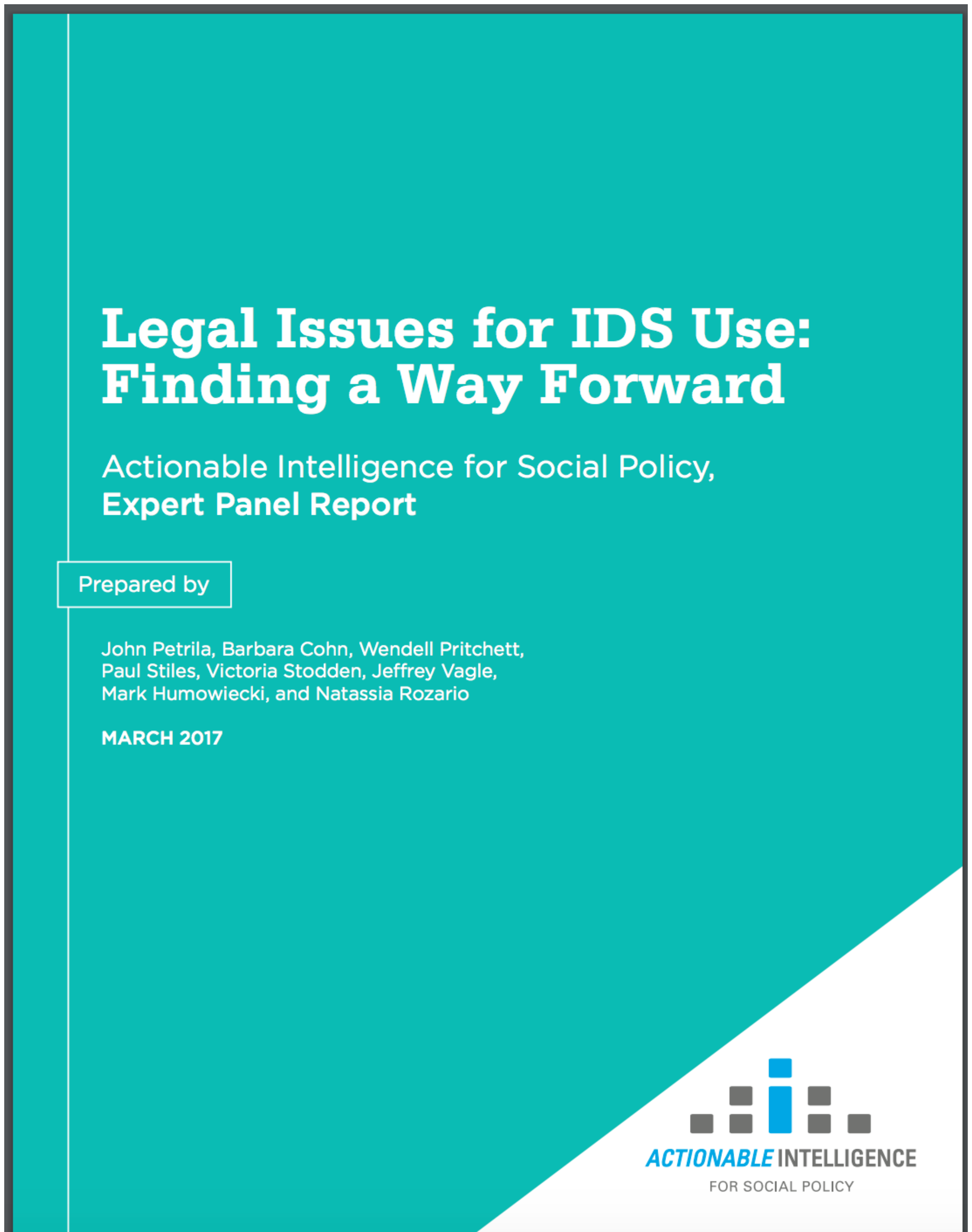
The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

A suite of license recommendations for computational science:

- Release media components (text, figures) under **CC BY**,
- Release code components under **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

  - ➡ Remove copyright's barrier to reproducible research and,
  - ➡ Realign the IP framework with longstanding scientific norms.

Legal Templates


Legal Issues for IDS Use:
Finding a Way Forward

Actionable Intelligence for Social Policy,
Expert Panel Report

Prepared by

John Petrila, Barbara Cohn, Wendell Pritchett,
Paul Stiles, Victoria Stodden, Jeffrey Vagle,
Mark Humowiecki, and Natassia Rozario

MARCH 2017

ACTIONABLE INTELLIGENCE
FOR SOCIAL POLICY

- Actionable Intelligence for Social Policy, 2017
- Four reports presenting Integrated Data Systems (IDS) technology for state and local government data sharing.

Effects of Computational Reproducibility: Fantasy Searches

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 effect sizes. Do this for all clinical trials published in 2003 and list the trial name and histogram side by side.

Courtesy of Donoho and Gavish 2012

www.mrsec.harvard.edu/2017NSFReliability/index.html

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 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)

ACM Badges

https://www.acm.org/publications/policies/artifact-review-badging

Digital Library CACM

Association for Computing Machinery

Advancing Computing as a Science & Profession

ABOUT ACM MEMBERSHIP PUBLICATIONS SPECIAL INTEREST GROUPS CONFERENCES CHAPTERS AWARDS

Publications Home | About Publications Digital Library ACM Books ICPS Submit Propose Polici

Home > Publications > Policies > Result And Artifact Review And Badging

Result and Artifact Review and Badging

An experimental result is not fully established unless it can be independently reproduced. A variety of recent studies, primarily in the biomedical field, have revealed that an uncomfortably large number of research results found in the literature fail this test, because of sloppy experimental methods, flawed https://www.acm.org/publications/policies/artifact-review-badging

Terminology.

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

- Repeatability (Same team, same experimental setup)
  - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Replicability (Different team, same experimental setup)
  - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Reproducibility (Different team, different experimental setup)
  - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.