The Role of Cyberinfrastructure in Reproducible Science

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

Chameleon Cloud User Meeting
Argonne National Labs
September 13, 2017

Agenda

1. Why Reproducibility?
2. Definitions and Framing
3. Tools and Cyberinfrastructure Solutions
4. Imagining a Reproducible Scholarly Record

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.

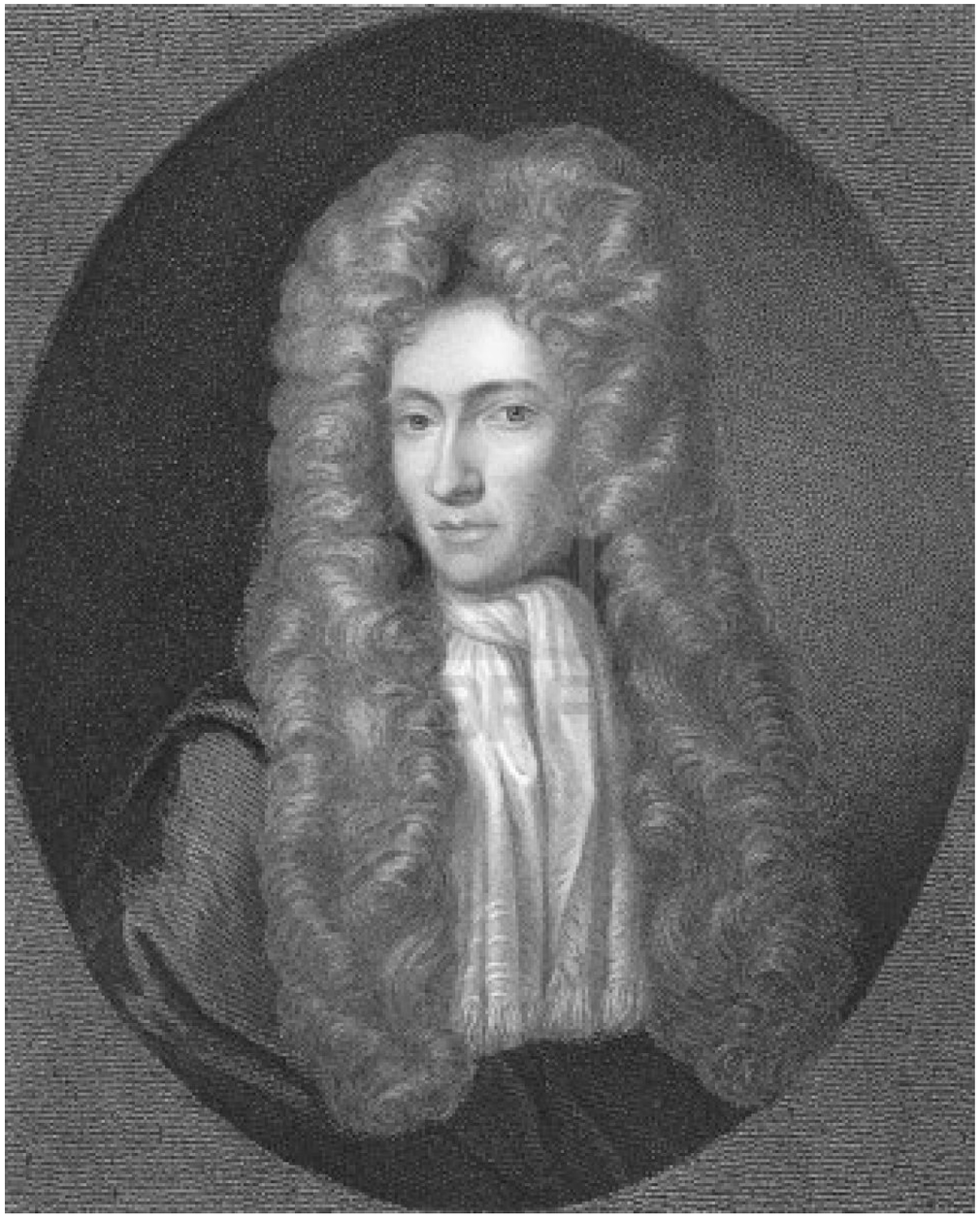
Disinterestedness: act for the benefit of a common scientific enterprise, rather than for personal gain.

Originality: scientific claims contribute something new

Skepticism: scientific claims must be exposed to critical scrutiny before being accepted.

Skepticism -> Reproducibility

- 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.



The Impact of Technology

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.

Claim 1: *Virtually all published discoveries today have a computational component. (is Data Science all science?)*

Claim 2: *There is a mismatch between the traditional scientific process and computation, leading to reproducibility concerns.*

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

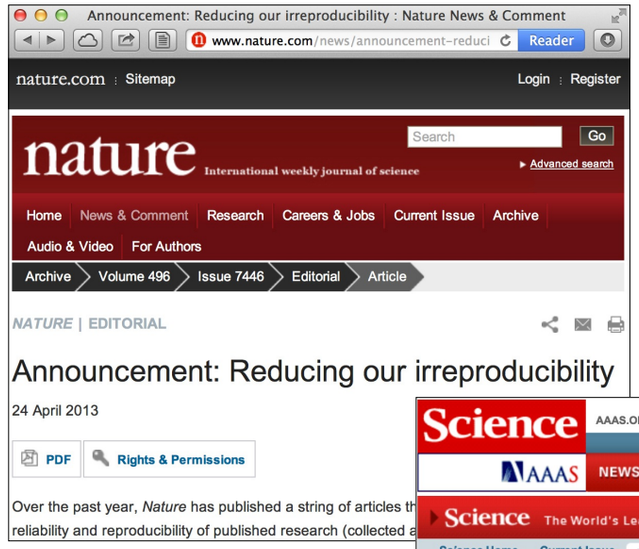
The software contains "ideas that enable biology..."
Stories from the Supplement, 2013

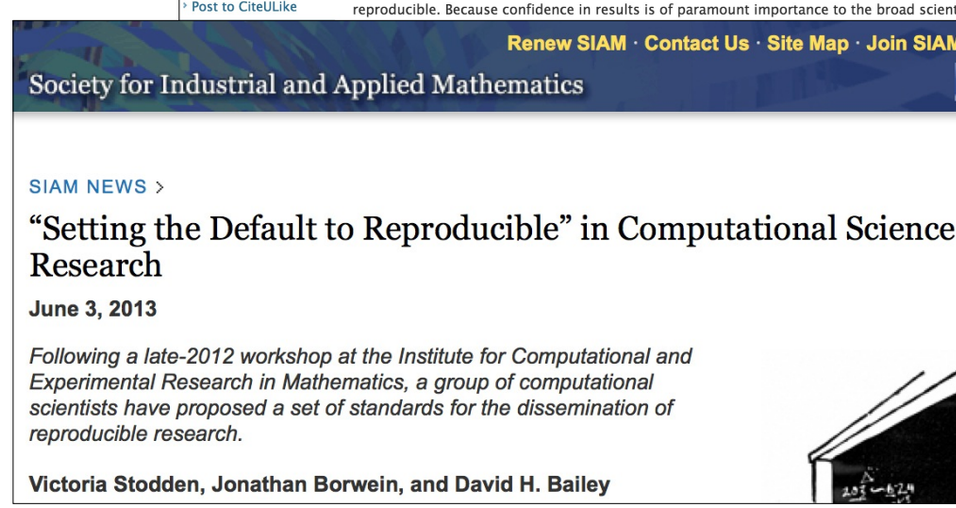
Parsing Reproducibility

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”





V. Stodden, IMS Bulletin (2013)

Empirical Reproducibility

Cell Reports
Commentary

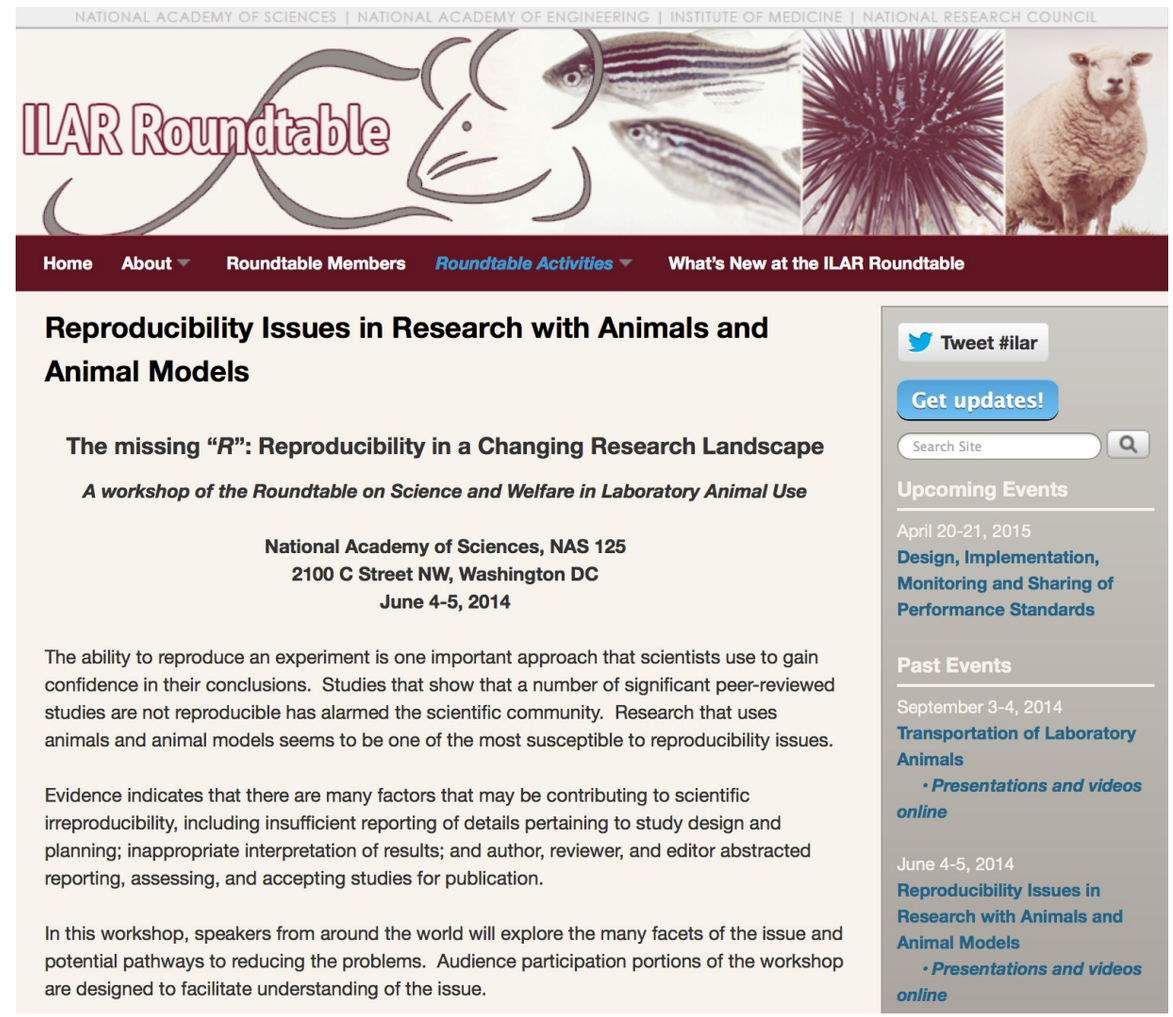
Sorting Out the FACS: A Devil in the Details

William C. Hines,[1,5,*] Ying Su,[2,3,4,5,*] Irene Kuhn,[1] Kornelia Polyak,[2,3,4,5] and Mina J. Bissell[1,5]
[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[3]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[4]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[5]These 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.celrep.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

Cell PRESS
Open ACCESS



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,
- Investigator bias toward previous findings; conflicts of interest.
- …

Response: Science 2014

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

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.

Argument: computation presents only a potential third/fourth branch of the scientific method (Donoho et al 2009).

Modeling and Simulation Workshop

math.nist.gov/~JBlue/spw.html

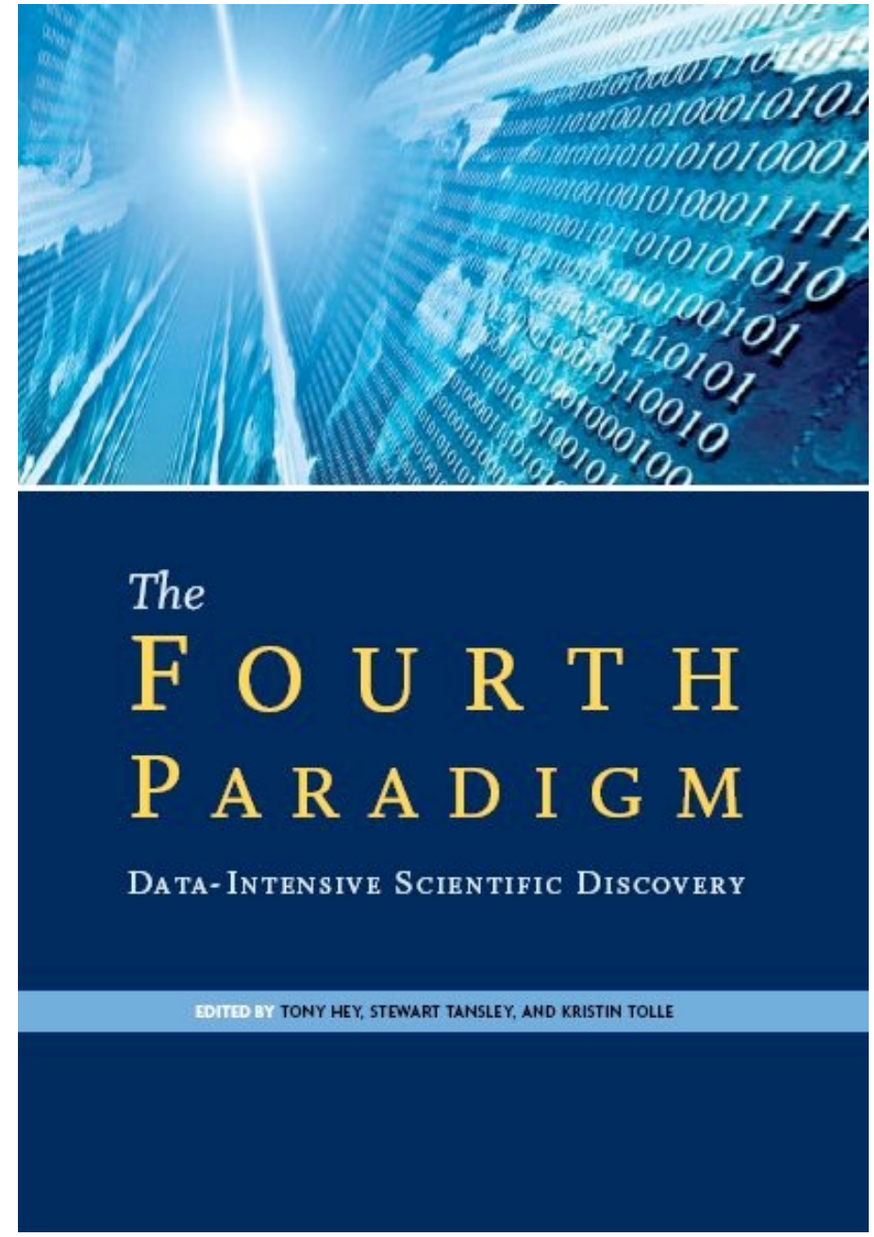
Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

Gaithersburg, MD
September 21, 1995

Workshop Overview

The workshop consisted of an introduction; five talks, each followed by a discussion period; and an open discussion session. Capsule versions follow immediately; more substantial summaries follow later.

Jim Blue opened the workshop with brief introductory remarks. He emphasized that the purpose of doing modeling and simulation is to gain understanding and insight. The three benefits are that modeling and simulation can be cheaper, quicker, and better than experimentation alone. It is common now to consider computation as a third branch of science, besides theory and experiment.



"It is common now to consider computation as a third branch of science, besides theory and experiment."

"This book is about 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

"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)

Why Reproducibility?

1. Scientific verifiability, but also:

2. The concept of Reproducibility naturally integrates data, code, workflows, and other scholarly artifacts into the scholarly record.

3. Researchers understand the idea of making their results reproducible.

4. Opportunity: Convergence of two (ordinarily antagonistic) trends.

   ➡ Scientific projects will become massively more computing intensive

   ➡ Scientific computing dramatically more transparent

Converging Trends

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.

Looking ahead..

We imagine a major effort to 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.

Enables **efficiency** and **productivity**, and **discovery**.

Where Could We Be?

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.

Donoho and Gavish 2012

Example: Cancer Research

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)”

With Xiaomian Wu and April Tang, we carried out the third scholarly record query.

ALL/AML Query

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.

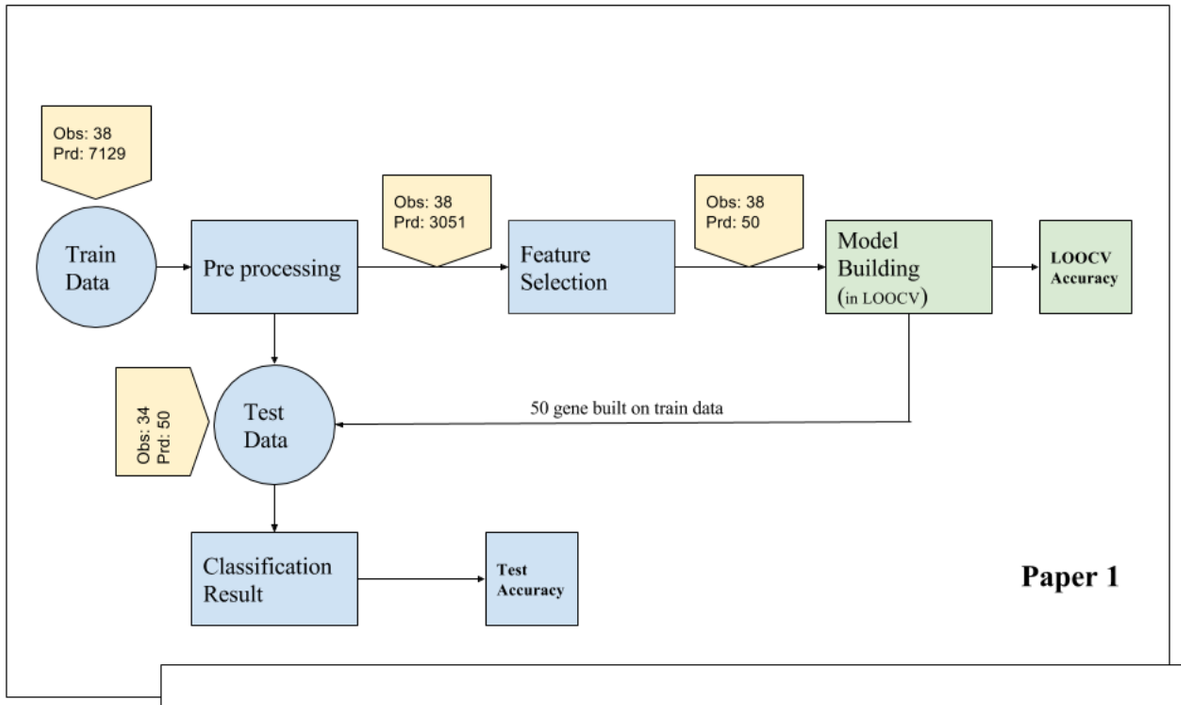
Our next step was to create the table of misclassification rates. We identified 5 articles for which this seemed possible.

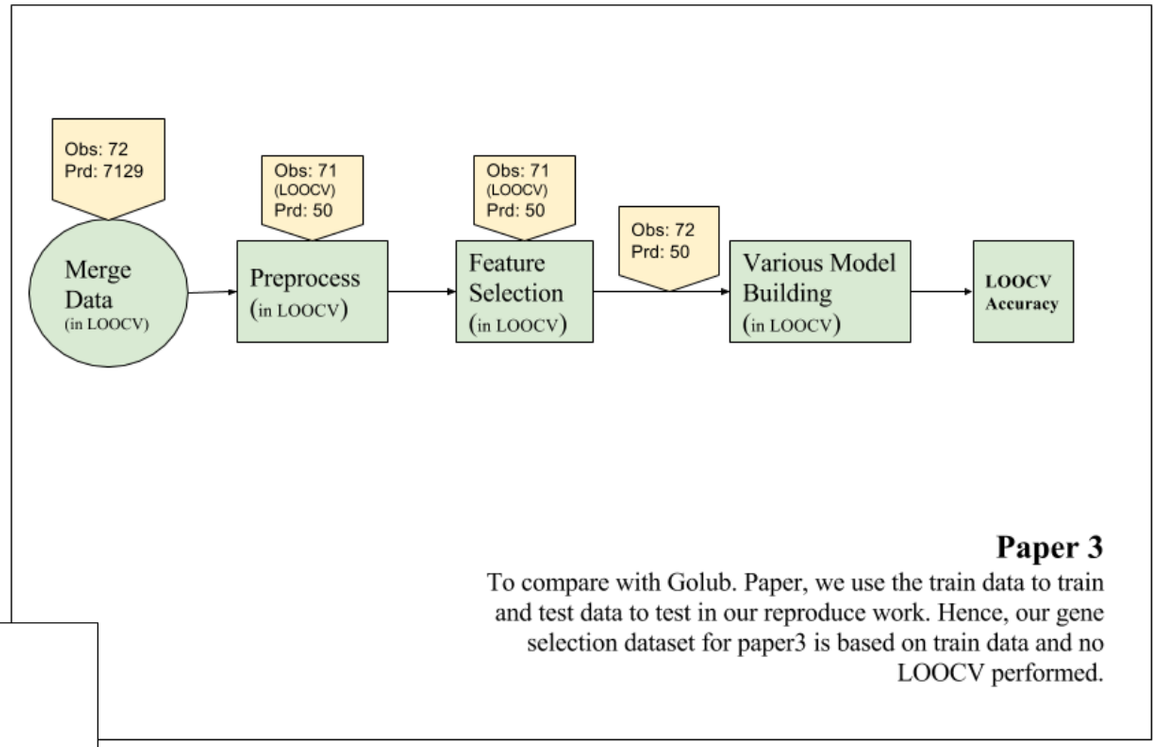
Lack of Comparability

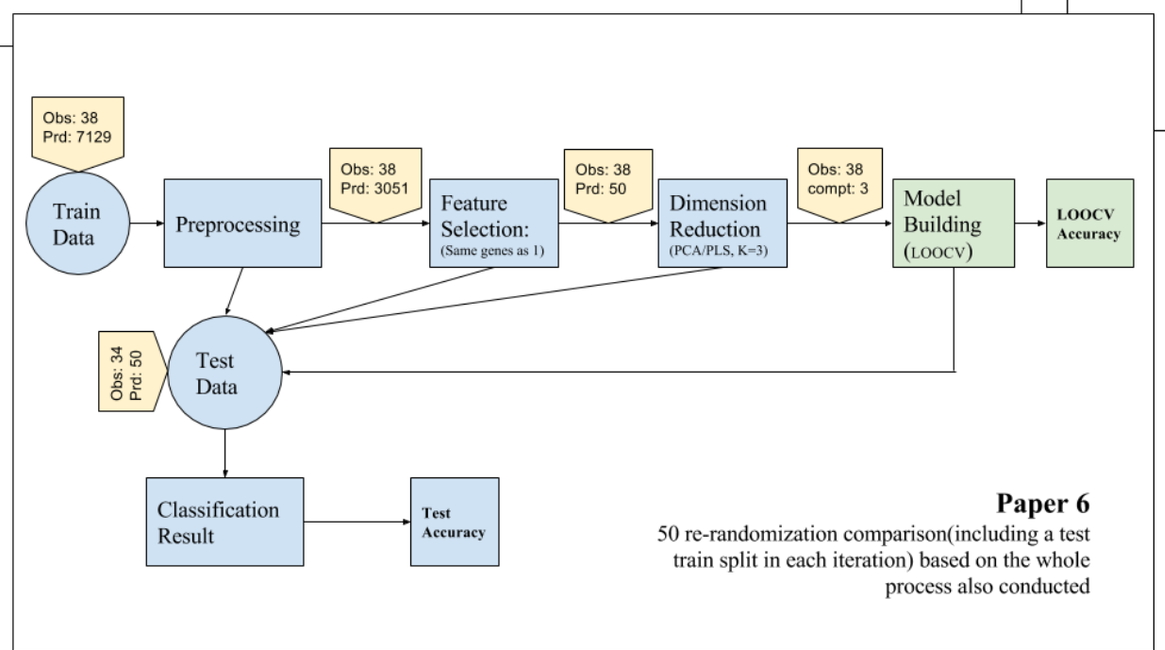
We obtained the original Golub data. We hoped to apply the various machine learning algorithms from the literature, in the 5 cases we identified.

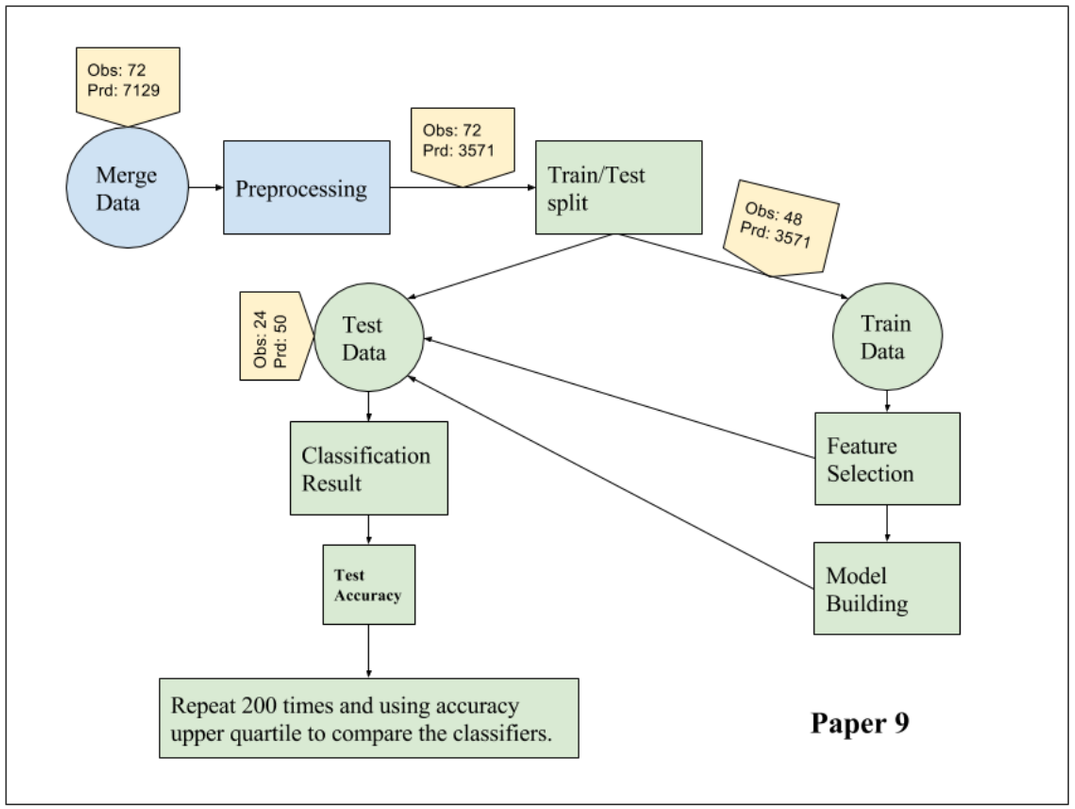
We found that the articles implemented (at least) three steps, each varying from one article to the next:

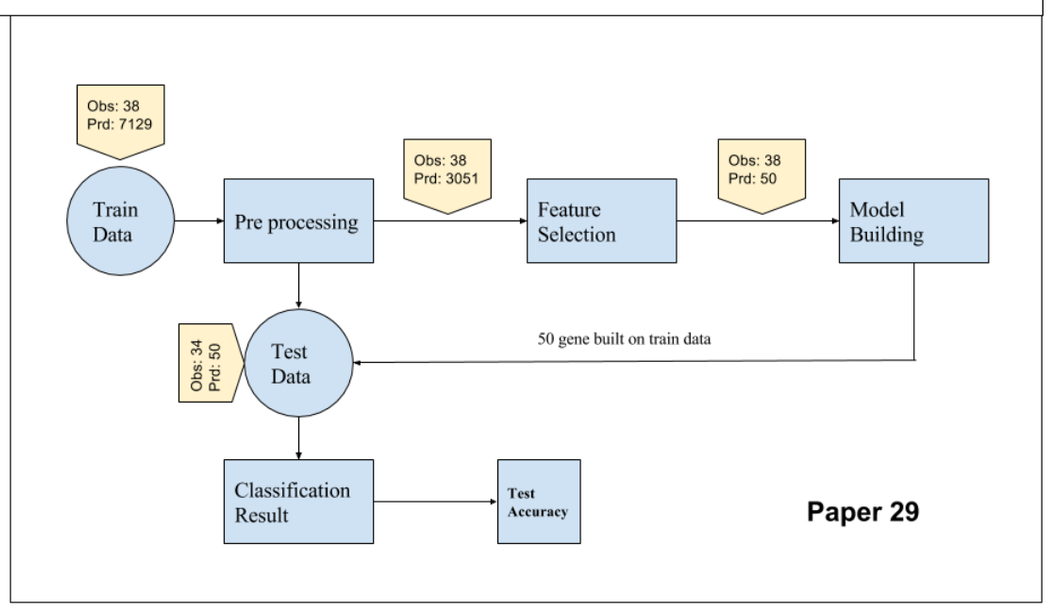
1. data preprocessing,
2. feature selection,
3. application of machine learning algorithm.


Obs: 38
Prd: 7129
Train Data
Pre processing
Obs: 38
Prd: 3051
Feature Selection
Obs: 38
Prd: 50
Model Building
(in LOOCV)
LOOCV Accuracy
Obs: 34
Prd: 50
Test Data
50 gene built on train data
Classification Result
Test Accuracy
Paper 1


Obs: 72
Prd: 7129
Merge Data
(in LOOCV)
Obs: 71
(LOOCV)
Prd: 50
Preprocess
(in LOOCV)
Obs: 71
(LOOCV)
Prd: 50
Feature Selection
(in LOOCV)
Obs: 72
Prd: 50
Various Model Building
(in LOOCV)
LOOCV Accuracy
Paper 3
To compare with Golub. Paper, we use the train data to train and test data to test in our reproduce work. Hence, our gene selection dataset for paper3 is based on train data and no LOOCV performed.


Obs: 38
Prd: 7129
Train Data
Preprocessing
Obs: 38
Prd: 3051
Feature Selection:
(Same genes as 1)
Obs: 38
Prd: 50
Dimension Reduction
(PCA/PLS, K=3)
Obs: 38
compt: 3
Model Building
(LOOCV)
LOOCV Accuracy
Obs: 34
Prd: 50
Test Data
Classification Result
Test Accuracy
Paper 6
50 re-randomization comparison(including a test train split in each iteration) based on the whole process also conducted


Obs: 72
Prd: 7129
Merge Data
Preprocessing
Obs: 72
Prd: 3571
Train/Test split
Obs: 48
Prd: 3571
Obs: 24
Prd: 50
Test Data
Train Data
Classification Result
Feature Selection
Test Accuracy
Model Building
Repeat 200 times and using accuracy upper quartile to compare the classifiers.
Paper 9


Obs: 38
Prd: 7129
Train Data
Pre processing
Obs: 38
Prd: 3051
Feature Selection
Obs: 38
Prd: 50
Model Building
Obs: 34
Prd: 50
Test Data
50 gene built on train data
Classification Result
Test Accuracy
Paper 29

Comparisons Difficult..

Hard to synthesize (200+ student hours)

Many points of variability: starting dataset; preprocessing steps; feature selection methods; algorithm choice; tuning of algorithm, model validation steps…

Details not well-captured in the traditional article…

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.

Infrastructure as Solution

Idea: Adapt the “Common Task Framework”:

1. Agreement on datasets prior to analysis, conferences around those datasets,

2. Hold-out data held by a neutral third party, not seen by researchers,

3. Researchers distinguish and specify feature selection and preprocessing vs learning algorithm application,

4. Send code to the third party who returns your misclassification rate on the test data.

Note: training data and code/algorithm shared.

Infrastructure that supports the computational experiments can facilitate the comparison not only of the results, but the workflows that generated the results.

Interlude: A Brief History of Computational Reproducibility

Community Responses

Declarations and Documents:

- Yale Declaration 2009
- ICERM 2012
- XSEDE 2014

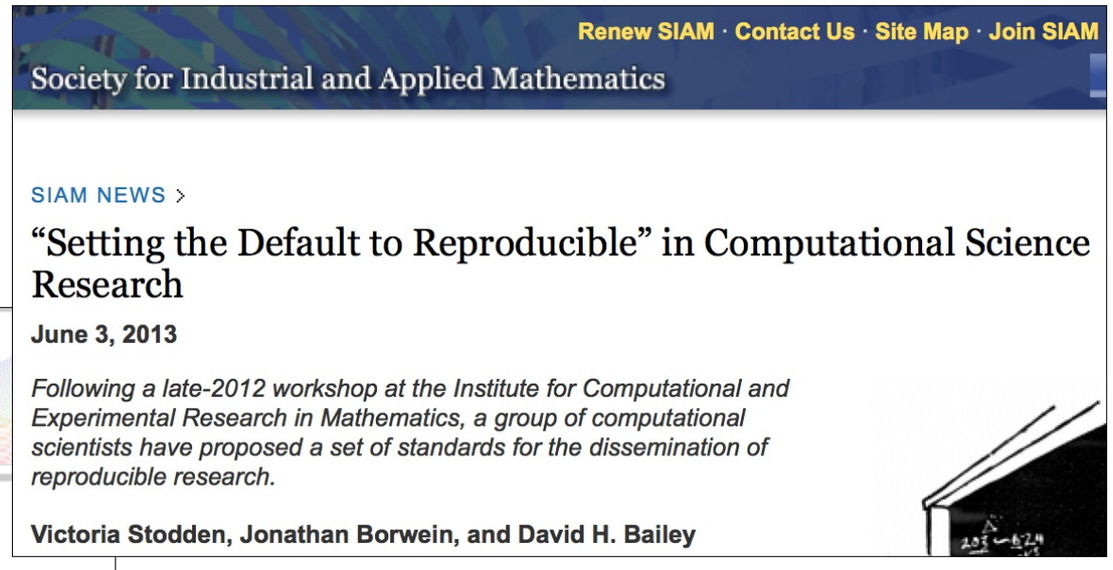
NEWS

Reproducible Research

Addressing the Need for Data and Code Sharing in Computational Science

By the Yale Law School Roundtable on Data and Code Sharing

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



Renew SIAM · Contact Us · Site Map · Join SIAM

Society for Industrial and Applied Mathematics

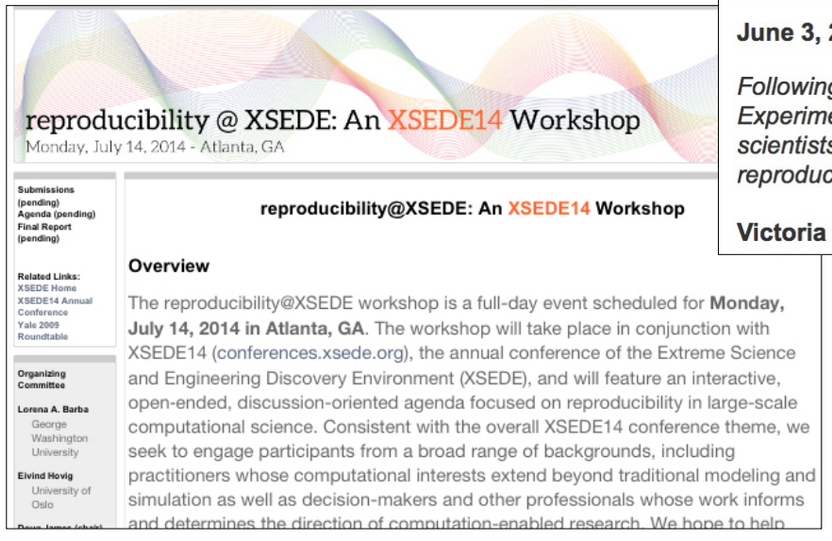
SIAM NEWS >

"Setting the Default to Reproducible" in Computational Science Research

June 3, 2013

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

Victoria Stodden, Jonathan Borwein, and David H. Bailey



reproducibility @ XSEDE: An XSEDE14 Workshop

Monday, July 14, 2014 - Atlanta, GA

Submissions (pending)
Agenda (pending)
Final Report (pending)

Related Links:
XSEDE Home
XSEDE14 Annual Conference
Yale 2009 Roundtable

Organizing Committee

Lorena A. Barba
George Washington University

Eivind Hovig
University of Oslo

reproducibility@XSEDE: An XSEDE14 Workshop

Overview

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

Setting the Default to Open

Setting the Default to Reproducible

Reproducibility in Computational and Experimental Mathematics

Developed collaboratively by the ICERM workshop participants[1]

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 "repro
scientific research and publication practices now fall sh
the production of computational science – scientists wh
employ them, journals and dissemination mechanisms,
reproducible research.

siam news

Click here to sign up for

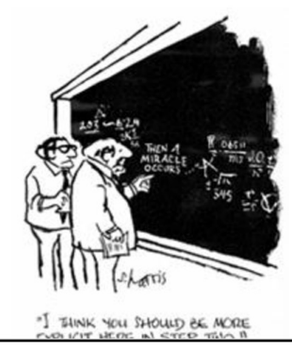
CURRENT ISSUE | ALL ISSUES | ABOUT SIAM NEWS | SIAM NEWS ARCHIVE

Printable version

June 03, 2013

"Setting the Default to Reproducible" in Computational Science Research

By Victoria Stodden, Jonathan M. Borwein, David H. Bailey



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

Computation is now central to the scientific enterprise, and the emergence of powerful computational hardware, combined with a vast array of computational software, presents novel opportunities for researchers. Unfortunately, the scientific culture surrounding computational work has evolved in ways that make it difficult to verify findings, efficiently build on past research, or even apply the basic tenets of the scientific method to computational procedures.

As a result, computational science is facing a credibility crisis [1,2,4,5]. The enormous scale of state-of-the-art scientific computations, using tens or hundreds of thousands of processors, presents unprecedented challenges. Numerical reproducibility is a major issue, as is hardware reliability. For some ubatomic particles matter.

Set the Default to "Open"

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

The State of Experimental and Computational Mathematics. Experimental mathematics[1]—application of high-performance computing technology to research questions in pure and applied mathematics, including

"It says it's sick of doing things like inventories and payrolls, and it wants to make some breakthroughs in astrophysics."

ScienceCartoonsPlus.com.

physicists, legal scholars, journal editors, and funding agency officials representing academia, government labs, industry research, and all points in between. While

Government Mandates

- OSTP 2013 Open Data and Open Access Executive Memorandum; Executive Order.
- "Public Access to Results of NSF-Funded Research"
- NOAA Data Management Plan, Data Sharing Plan
- NIST "Common Access Platform"
- …

Federal Agencies

www.nsf.gov/mps/perspectives/reliable_science_sep2015.jsp

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT NSF | FASTLANE

National Science Foundation
Directorate for Mathematical & Physical Sciences (MPS)

QUICK LINKS

SEARCH

MPS HOME | MPS FUNDING | MPS AWARDS | MPS DISCOVERIES | MPS NEWS | ABOUT MPS

Math & Physical Sciences

- MPS Home
- About MPS
- Funding Opportunities
- Awards
- News
- Events
- Discoveries
- Publications
- Advisory Committee
- Career Opportunities
- 2013-2014 Distinguished Lecture Series
- View MPS Staff

Search MPS Staff

MPS Organizations

- Astronomical Sciences (AST)
- Chemistry (CHE)
- Materials Research (DMR)
- Mathematical Sciences (DMS)
- Physics (PHY)
- Office of Multidisciplinary

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 have 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 approaches ranging from a torsional pendulum to atom interferometry.

grants.nih.gov/reproducibility/index.htm

GRANTS & FUNDING

NIH National Institutes of Health
Office of Extramural Research

Grants Policy

- Policy & Guidance
- Compliance & Oversight
- Research Involving Human Subjects
- Office of Laboratory Animal Welfare (OLAW)
- Animals in Research

Rigor and Reproducibility

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

On This Page:

- News
- Goals

lications

utlining updates to form
verview (NOT-OD-16-
sparency in NIH &
d Implementing Rigor
Award Applications (NOT-

l Research Dr. Mike
rust in Science through
raining module that
lucibility through rigor

www.nih.gov/research-training/rigor-reproducibility

NIH National Institutes of Health

Home » Research & Training

RIGOR AND REPRODUCIBILITY

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

Johns Hopkins University students in a laboratory. *Johns Hopkins University*

In recent years, however, there has been a growing awareness of the need for rigorously designed published preclinical studies, to ensure that such studies can be reproduced. This webpage provides information about the efforts underway by NIH to enhance rigor and reproducibility in scientific research.

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

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

RECOMMENDATION 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.

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

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

Reproducibility Enhancement Principles

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

RECOMMENDATION 6: Use Open Licensing when publishing digital scholarly objects.

RECOMMENDATION 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.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

www.mrsec.harvard.edu/2017NSFReliability/

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

February 25 - 26, 2017

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

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

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

Principal Investigator
David A. Weitz (Harvard University)

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

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

Agency Contacts
Bogdan Mihaila (NSF, Mathematical and Physical Sciences)
Gregory W. Warr (NSF, Molecular and Cellular Biosciences)

Journal Data and Code Sharing Policies

	Data 2011	Data 2012	Code 2011	Code 2012
Required as condition of publication, barring exceptions	10.6%	11.2%	3.5%	3.5%
Required but may not affect editorial decisions	1.7%	5.9%	3.5%	3.5%
Encouraged/addressed, may be reviewed and/or hosted	20.6%	17.6%	10%	12.4%
Implied	0%	2.9%	0%	1.8%
No mention	**67.1%**	**62.4%**	**82.9%**	**78.8%**

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

“ezDMP”

NSF funded project to provide structured guidance for a second generation data management plan.

EAGER: Collaborative Proposal: Supporting Public Access to Supplemental Scholarly Products Generated from Grant Funded Research (2016).

Helen M. Berman, Rutgers
Kerstin Lehnert, Columbia
Victoria Stodden, UIUC
Maggie Gabanyi, Rutgers
Vicki Ferrini, Columbia

exDMP Progress

- Examined selected data management plans to understand gaps, successes, and patterns of use in IEDA DMP Tool
- Reviewed the patterns exhibited by DMP creators using the IEDA DMP Tool
- GEO (under the lead of K. Lehnert & V. Ferrini)
- BIO (under the lead of H. Berman)
- SBE, MPS (under the lead of V. Stodden)
- Implement into IEDA (“ezDMP”)

SC16 Explores Reproducibility for Advanced Computing Through Student Cluster Competition by Michela Taufer

March 16, 2016 — Leave a Comment

Data sets and software are important by-products products of research in fields that depend upon data-intensive and high performance computing. But these elements are typically absent when research results are recorded in a journal article or conference proceedings. There is a growing sense in the computational community that this gap needs to be filled if we are to create a stable base of research upon which reliable advances may be built. In short, we need to ensure that computational results are as reproducible as those from experiments.

SC16's SCC Reproducibility Committee Member Michela Taufer from the University of Delaware.

Blog Archive

November 2016 (37)
October 2016 (15)
September 2016 (17)
August 2016 (20)
July 2016 (10)
June 2016 (11)
May 2016 (9)
April 2016 (11)
March 2016 (8)
February 2016 (3)
December 2015 (2)
November 2015 (22)
October 2015 (20)

Infrastructure Innovations

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 | flywheel.io

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
Wavelab | Sparselab

Conclusions: Three Principles for Computational Infrastructure

1. *Supporting scientific norms*: Not only should CI enable new discoveries, but it should also permit others to reproduce the computational findings, reuse and combine digital outputs such as datasets and code, and facilitate validation and comparisons to previous findings.

2. *Supporting best practices in science*: CI in support of science should embed and encourage best practices in scientific research and discovery.

3. *Taking a holistic approach to CI*: The complete end-to-end research pipeline should be considered to ensure interoperability and the effective implementation of 1 and 2.

See Stodden, Miguez, Seiler, “ResearchCompendia.org: Cyberinfrastructure for Reproducibility and Collaboration in Computational Science” CiSE 2015

Cyberinfrastructure

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 | flywheel.io

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
Wavelab | Sparselab

“Whole Tale” Project

The Whole Tale project seeks to leverage & contribute to **existing cyberinfrastructure and tools** to support the **whole research story**, and provide access to data and computing power.

➡ ***Integrate*** *tools to* ***simplify usage*** *and promote* ***best practices***

B. Ludaescher, K. Chard, N. Gaffney, M. B. Jones, J. Nabrzyski, V. Stodden, M. Turk
NSF CC*DNI DIBBS awarded 2016: 5 Institutions for 5 Years ($5M total)

Whole Tale Project Goals

Expose existing digital resources to researchers

… through popular frontends (Jupyter, RStudio, ..)

Develop necessary “software glue”

… for seamless access to different CI-backend capabilities

Enhance conceptualization-to-publication lifecycle

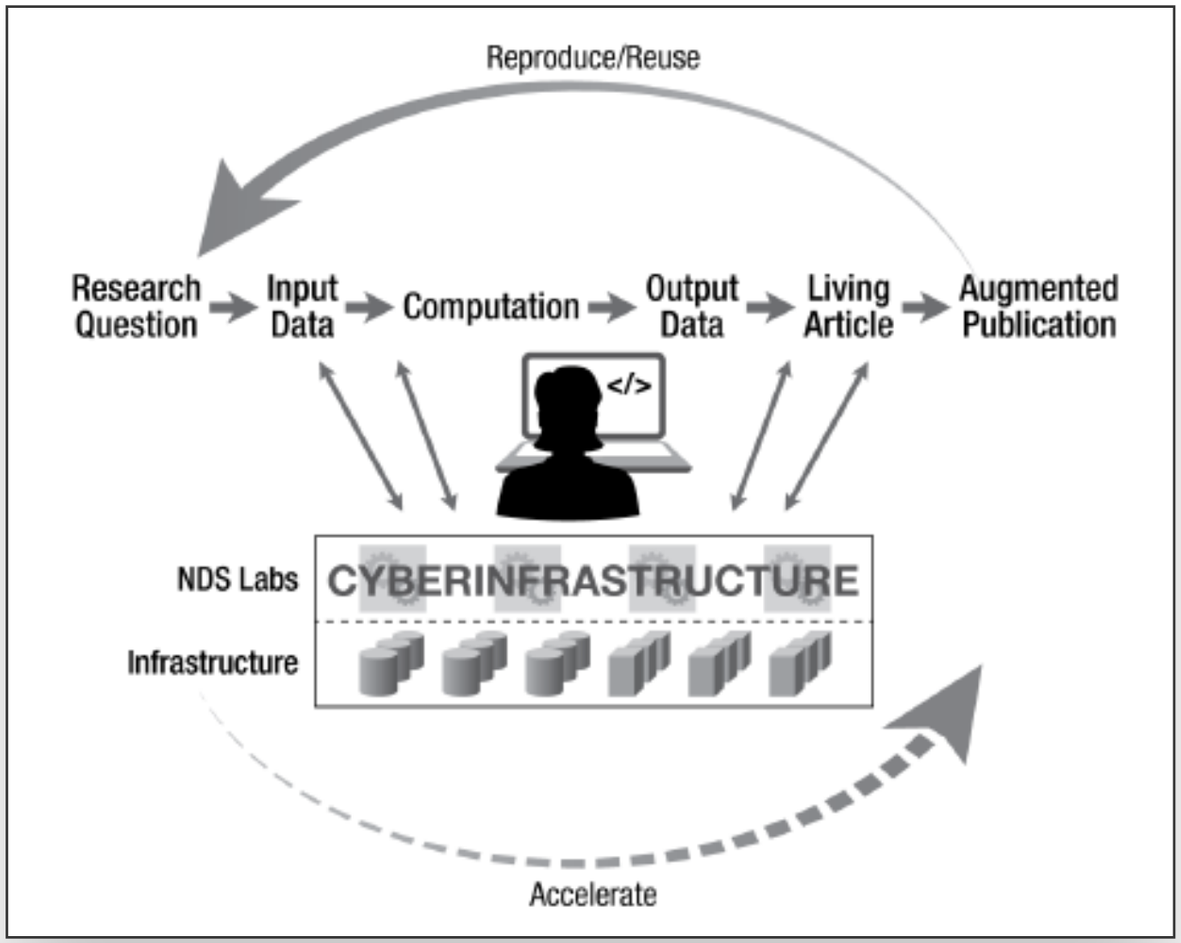
… by empowering scientists to create computational narratives in their usual programming environments

Embed reproducibility and best/better practices in the digital research environment

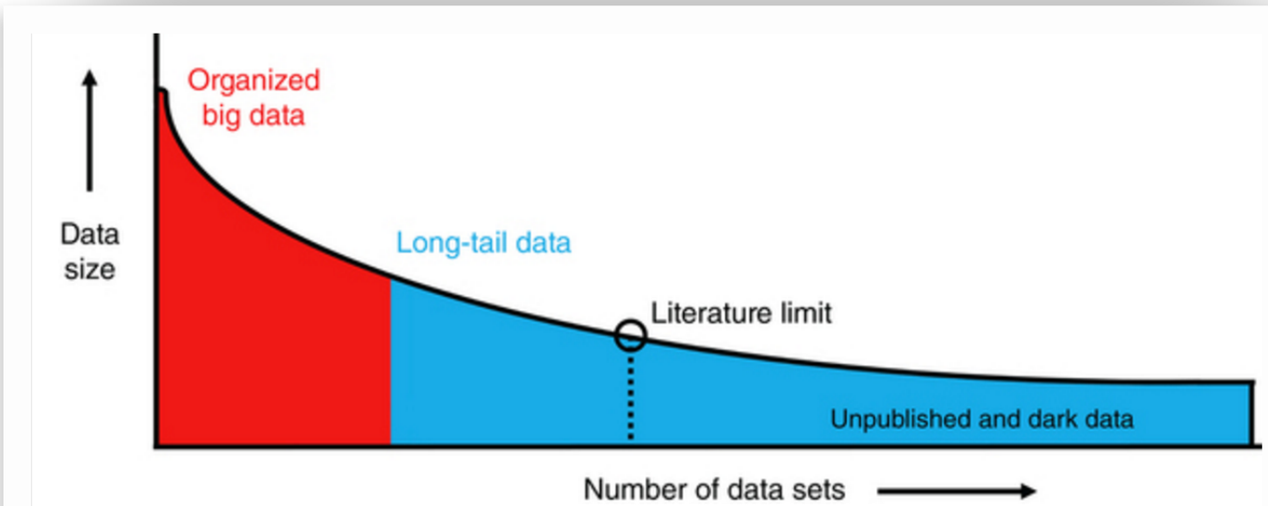
Whole Tale: What's in a Name?

(1) Whole Tale ⇔ Whole **Story**:

Support (computational & data) **scientists** along the **complete research lifecycle** from **experiment** to **publication** and back!



(2) Whole Tale ⇔ **Long Tail of Science**



Studies that have plotted data set size against the number of data sources reliably uncover a skewed distribution. Well-organized big science efforts featuring homogenous, well-organized data represent only a small proportion of the total data collected by scientists. A very large proportion of scientific data

image from Ferguson et al. 2014 doi:10.1038/nn.3838

“Tales”

“Tales” are the final research output from a project, capturing the complete provenance of a particular activity within the system:

- capture full provenance of an analysis recorded transparently,
- easily shared with others,
- publishable in repositories,
- associated with persistent identifiers,
- linked to publications,
- execute in the same state as it was when first published,
- acts as a starting point for research.

Danger: A Distributed Scholarly Record

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.

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 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.

Querying the Scholarly Record

- Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;
- Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;
- List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;
- Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;
- Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the year 2003 and list the trial name and histogram side by side.

Courtesy of Donoho and Gavish 2012

Cyberinfrastructure Goals

- minimize time commitment by the user for both learning and using the CI,
- automate as much of the discovery and dissemination process as possible,
- facilitate queries across the scholarly record,
- capture all information needed to assess the findings.

Open Questions

- Who funds and supports CI?
- Who owns data, code, and research outputs?
- Who controls access and gateways?
- What are community standards around documentation, citation standards, best practices? Who enforces?
- Citation of CI? What are the incentives? What should they be?