Trust in Research Findings

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

Learning From The Science Of Cognition And Perception For Decision-Making:
A Workshop

Decadal Survey Of Social And Behavioral Sciences For Applications To National Security
Agenda

Washington, DC
January 24, 2018

Los Angeles Times | BUSINESS

LOCAL U.S. WORLD BUSINESS SPORTS ENTERTAINMENT HEALTH STYLE TR

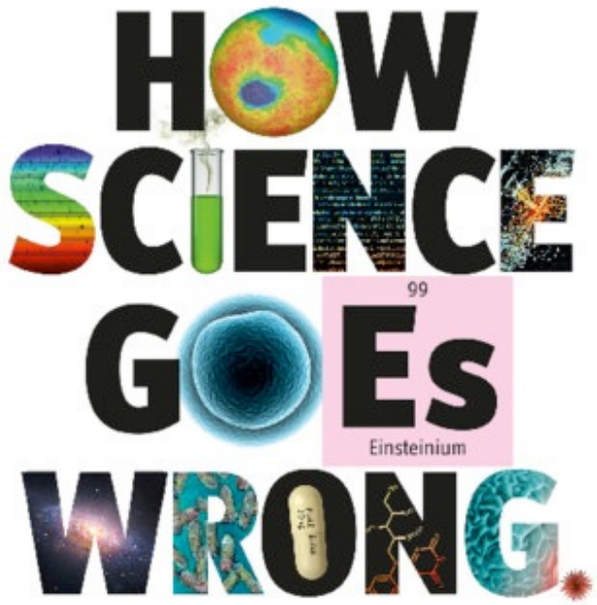
Science has lost its way, at a big cost to humanity

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.

The Economist

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

OCTOBER 19TH–25TH 2013 Economist.com



TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

NIH Tackles Irreproducibility

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

Science AAAS.ORG | FEEDBACK | HELP | LIBRARIANS All Science Journals
GUEST ALERT
AAAS NEWS SCIENCE JOURNALS CAREERS MULTIMEDIA COLLECTIONS
Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.
Science Home Current Issue Previous Issues Science Express Science Products My Science About the Journal

Home > Science Magazine > 17 January 2014 > McNutt, 343 (6168): 229

Article Views
- Summary
- Full Text
- Full Text (PDF)

Article Tools
- Save to My Folders
- Download Citation
- Alert Me When Article is Cited
- Post to CiteULike
- E-mail This Page
- Rights & Permissions
- Commercial Reprints and E-Prints
- View PubMed Citation

Related Content

Science 17 January 2014:
Vol. 343 no. 6168 p. 229
DOI: 10.1126/science.1250475

‹ Prev | Table of Co
Read Full Text t

EDITORIAL

Reproducibility

Marcia McNutt

» Marcia McNutt is Editor-in-Chief of *Science*.

Science advances on a foundation of trusted discoveries. Reproducing an experiment is on approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical not reproducible. Because confidence in results is of paramount importance to the broad s community, we are announcing new initiatives to increase confidence in the studies publis *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NIND increasing transparency.* Authors will indicate whether there was a pre-experimental plan handling (such as how to deal with outliers), whether they conducted a sample size estimat ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whe experimenter was blind to the conduct of the experiment. These criteria will be included in guidelines.

Announcement: Reducing our irreproducibility : Nature News & Comment
www.nature.com/news/announcement-reduci Reader

nature.com : Sitemap Login : Register

nature International weekly journal of science
Search Go
Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive
Audio & Video | For Authors

Archive > Volume 496 > Issue 7446 > Editorial > Article

NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

PDF | Rights & Permissions

nature International weekly journal of sc

Menu Advanced searc

archive ▸ volume 483 ▸ issue 7391 ▸ editorials ▸ article

NATURE | EDITORIAL

Must try harder

Nature **483**, 509 (29 March 2012) | doi:10.1038/48350
Published online 28 March 2012

PDF | Citation | Reprints | Rights & permiss

Too many sloppy mistakes are creeping into scientif look more rigorously at the data — and at themselve

The Reproducibility Discussion

- Most of the discussion of reproducibility of scientific findings has been *inward-facing* in orientation,
- However, public trust in science can be impacted by our internal discussions:
  - use of jargon, scientific phrases
  - reporting of changing “answers”
  - University Press offices (Alberts et al. 2015)

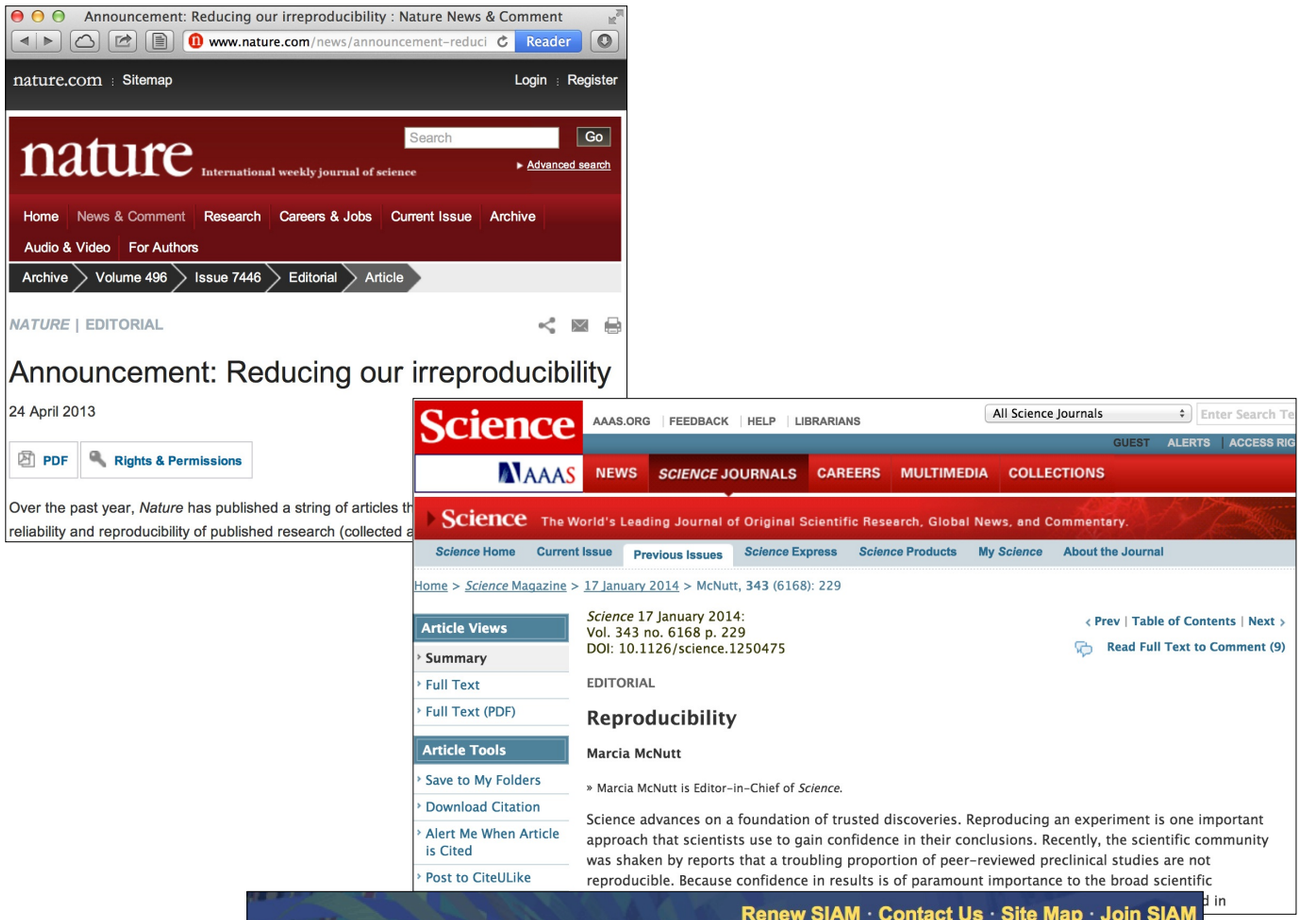
- Morale: Communication is very important.

Parsing Reproducibility I

"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

Reproducibility Issues in Research with Animals and Animal Models

The missing "*R*": Reproducibility in a Changing Research Landscape

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

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

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

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

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

Statistical Reproducibility

- False discovery, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments.

- Low power, poor experimental design, nonrandom sampling,

- Data preparation, treatment of outliers, re-combination of datasets, insufficient reporting/tracking practices,

- inappropriate tests or models, model misspecification,

- Model robustness to parameter changes and data perturbations,

- Investigator bias toward previous findings; conflicts of interest.

- …

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).

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.

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 transpar-

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://

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

Stodden, Victoria, et al. "Enhancing reproducibility for computational methods." *Science* 354(6317) (2016)

ness 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 are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or

All data, code, and workflows, including software written by the authors, should be cited in the references section (*10*). We suggest that software citation include software version information and its unique identifier in 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.

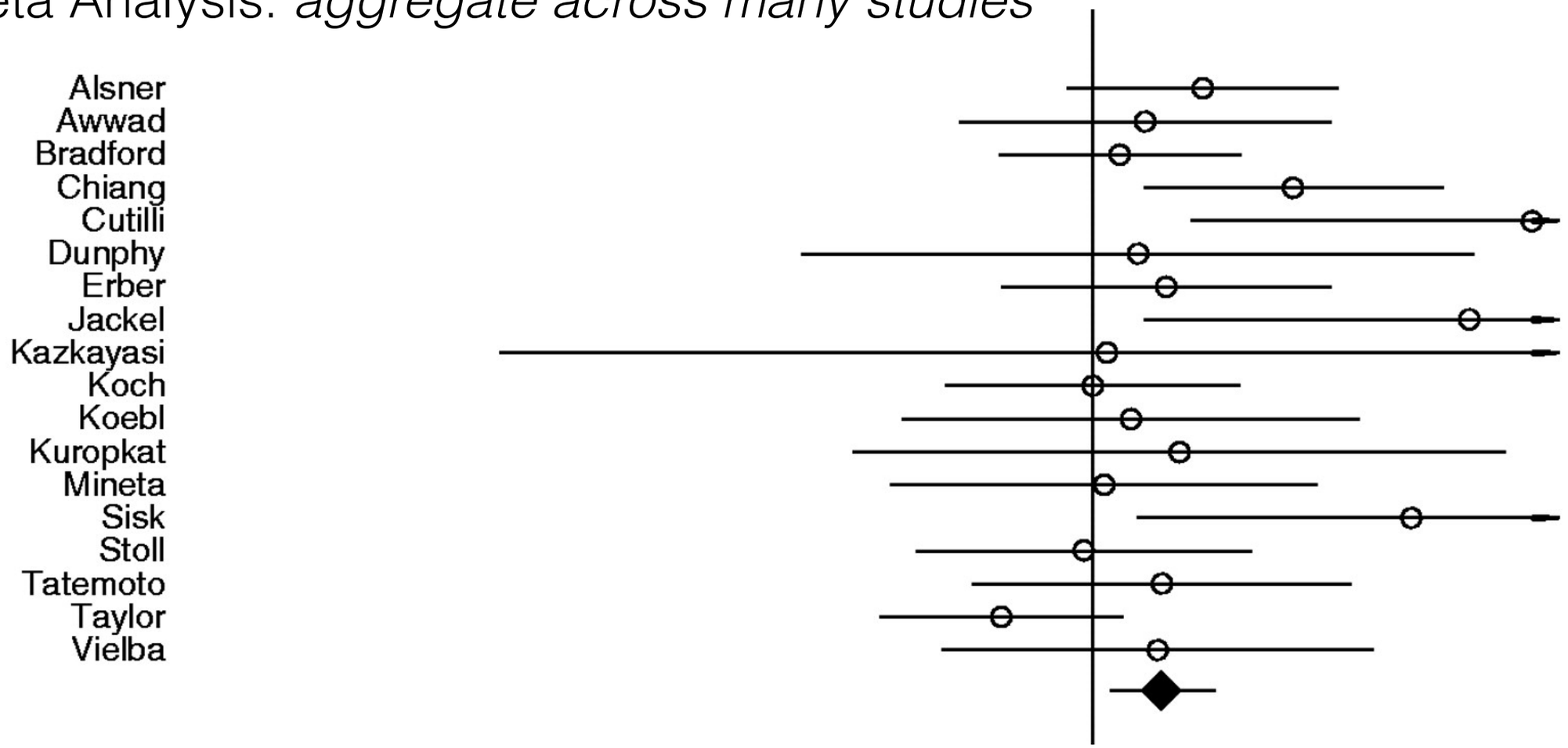
5: Journals should conduct a Reproducibility Check as part of the publication process and enact the TOP Standards at level 2 or 3.

6: Use Open Licensing when publishing digital scholarly objects.

7: Funding agencies should instigate new research programs and pilot studies.

Meta-Analysis

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



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

Kyzas et al., "Selective Reporting Biases in Cancer Prognostic Factor Studies," *JNCI*, 97(14), 2005

What Does Meta-Analysis Tell Us?

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

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

Example from Genomics

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

Efforts to Replicate "Candidate Gene" Association Studies Fail

TABLE 1. Large-scale Efforts to Massively Replicate Reported Candidate-gene Associations[a]

First Author	Disease/Phenotype	Gene Loci Tested	Sample Size (Design)	Replicated Gene Loci[b]
Bosker et al[15]	Major depressive disorder	57	3540 (case-control)	1
Caporaso et al[16]	Smoking (7 phenotypes)	359	4611 (cohort[c])	1
Morgan et al[17]	Acute coronary syndrome	70	1461 (case-control)	0
Richards et al[18]	Osteoporosis (2 phenotypes)	150	19,195 (cohort[d])	3[e]:9[f]
Samani et al[19]	Coronary artery disease	55	4864; 2519 (case-control)	1[g]
Scuteri et al[20]	Obesity (3 phenotypes)	74	6148 (cohort)	0
Sõber et al[21]	Blood pressure	149	1644; 8023 (cohort[h])	0
Wu et al[22]	Childhood asthma	237	1476 (triads[i])	1

- Table 1 shows "at least 20 false-positive findings for every one true-positive result"
- "approximately 1000 early gene loci-phenotype associations for the conditions listed in Table 1 were false positives from the candidate-gene approach."
- "There are no documented false-negative results arising from candidate-gene studies. Therefore, for the phenotypes listed in Table 1, the numerator of the FP:FN ratio is over 1000, while the denominator is apparently 0"

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

Recall: False Positives and False Negatives

Statistical Inference	True Underlying Relationship: +	True Underlying Relationship: −
+	True Positive	False Positive
−	False Negative	True Negative
	N_1	N_2

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.

Donoho & Gavish, “Three Dream Applications of Verifiable Computational Results,” CiSE, 2012

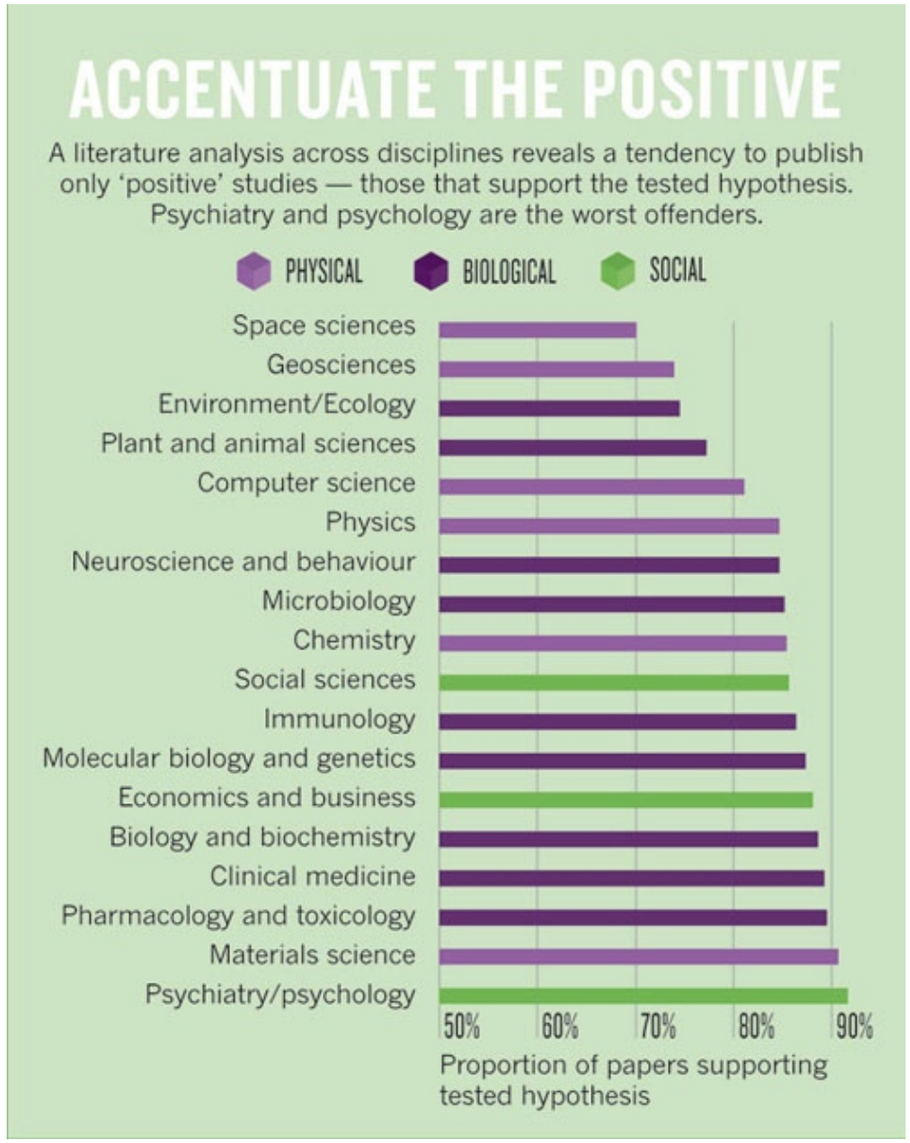
Summary

We now see the scholarly record as a body of numerical data, and we find:

- ➡ False Positives can overwhelm fields
- ➡ Entire fields are systemically failing
- ➡ Publications unstructured for analysis

Why?

- ➡ Overuse of underpowered studies
- ➡ Editorial preference for positive results
- ➡ Exploitation of researcher degrees of freedom



Yong, "Replication studies: Bad copy," Nature, 2012

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		

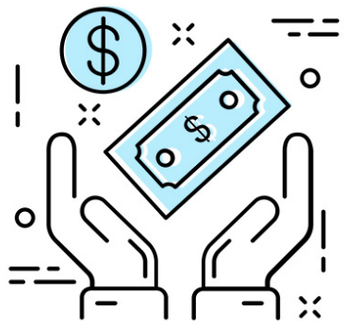
Computational Reproducibility

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

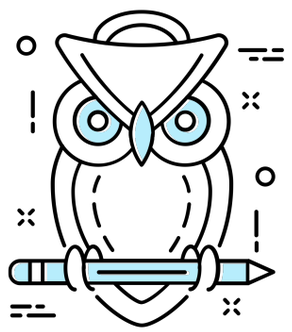
David Donoho, 1998 http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf

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

Ecosystem



Funders
(policy)

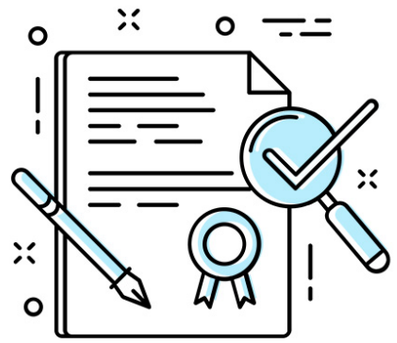


Scientific Societies

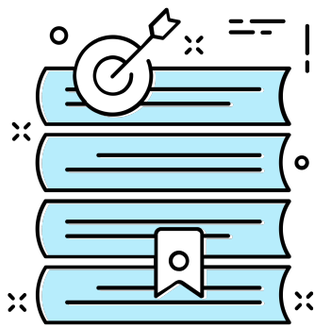
Regulatory Bodies
(OSTP Memos)

Researchers
(processes)

Universities/ institutions
(hiring/promotion)



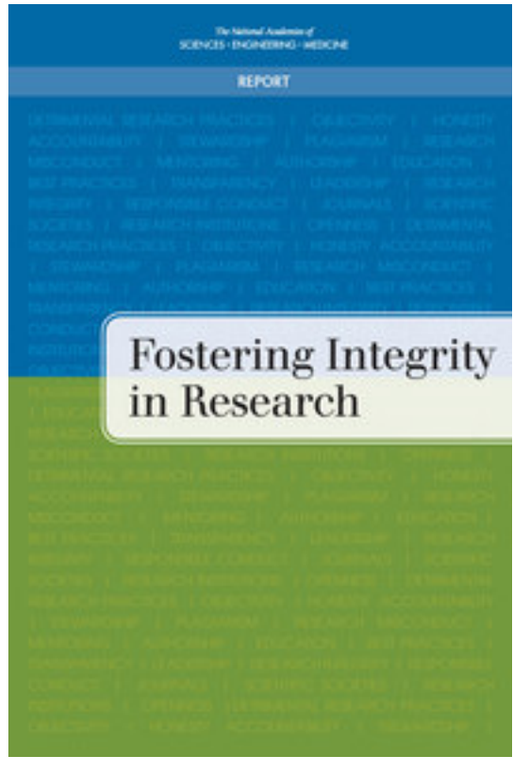
Publishers
(TOP guidelines)



Universities/ libraries
(empowering w/tools)

“Fostering Integrity in Research”

6: Through their policies and through the development of supporting infrastructure, research sponsors and science, engineering, technology, and medical journal and book publishers should ensure that **information sufficient** for a person knowledgeable about the field and its techniques **to reproduce reported results is made available at the time of publication** or as soon as possible after publication.



7: Federal funding agencies and other research sponsors should allocate sufficient funds to **enable the long-term storage, archiving, and access of datasets and code necessary for the replication of published findings**.

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

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.