Reproducibility and Replicability in Science

Presentation to the Council of Governmental Relations (COGR)

David B. Allison, Dean, Distinguished Professor, and Provost Professor, Indiana University
allison@iu.edu
7 June 2019

The National Academies of SCIENCES ENGINEERING MEDICINE

Committee on Reproducibility and Replicability in Science

Harvey V. Fineberg, Chair, Gordon and Betty Moore Foundation

David B. Allison, Indiana University

Lorena A. Barba, The George Washington University

Dianne Chong, Boeing Research and Technology (Retired)

David L. Donoho,* Stanford University

Juliana Freire, New York University

Gerald Gabrielse, Northwestern University

Constantine Gatsonis, Brown University

*Resigned from committee July 2018

Edward (Ned) Hall, Harvard University

Thomas H. Jordan, University of Southern California

Dietram A. Scheufele, University of Wisconsin-Madison

Victoria Stodden, University of Illinois at Urbana-Champaign

Simine Vazire,** University of California, Davis

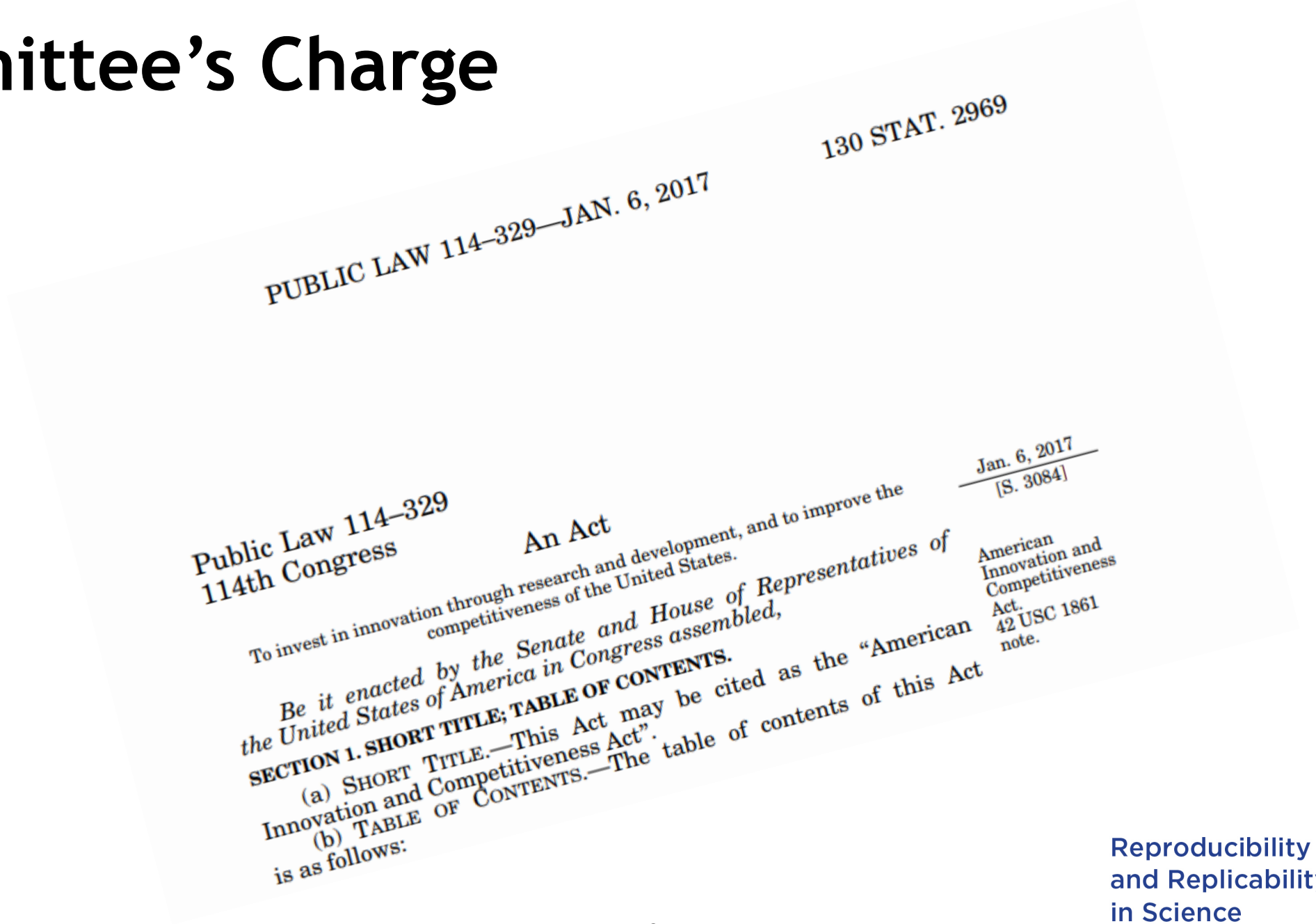
Timothy Wilson, University of Virginia

Wendy Wood, University of Southern California

**Resigned from committee October 2018

Reproducibility and Replicability in Science

Committee's Charge



PUBLIC LAW 114–329—JAN. 6, 2017 130 STAT. 2969

Public Law 114–329
114th Congress

An Act

To invest in innovation through research and development, and to improve the competitiveness of the United States.

Jan. 6, 2017
[S. 3084]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) SHORT TITLE.—This Act may be cited as the "American Innovation and Competitiveness Act".

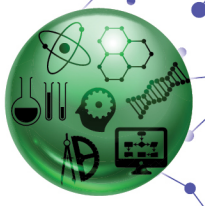
American Innovation and Competitiveness Act.
42 USC 1861 note.

(b) TABLE OF CONTENTS.—The table of contents of this Act is as follows:

Reproducibility and Replicability in Science

Committee's Charge

- Define reproducibility and replicability accounting for the diversity of fields in science and engineering.
- Examine the extent of non-reproducibility and non-replicability.
- Review current activities to improve reproducibility and replicability.
- Determine if the lack of replicability and reproducibility impacts the overall health of science and engineering as well as the public's perception of these fields.



Reproducibility and Replicability in Science

No crisis . . . No complacency.

- Improvements are needed.
- Reproducibility is important but not currently easy to attain.
- Aspects of replicability of individual studies are a serious concern.

Neither are the main or most effective way to ensure reliability of scientific knowledge.

Confusion Reigns in Defining the Terms

reproducibility = replicability

reproducibility = replicability = repeatability

reproducibility ≠ replicability

"One big problem keeps coming up among those seeking to tackle the issue: different groups are using terminologies in utter contradiction with each other."

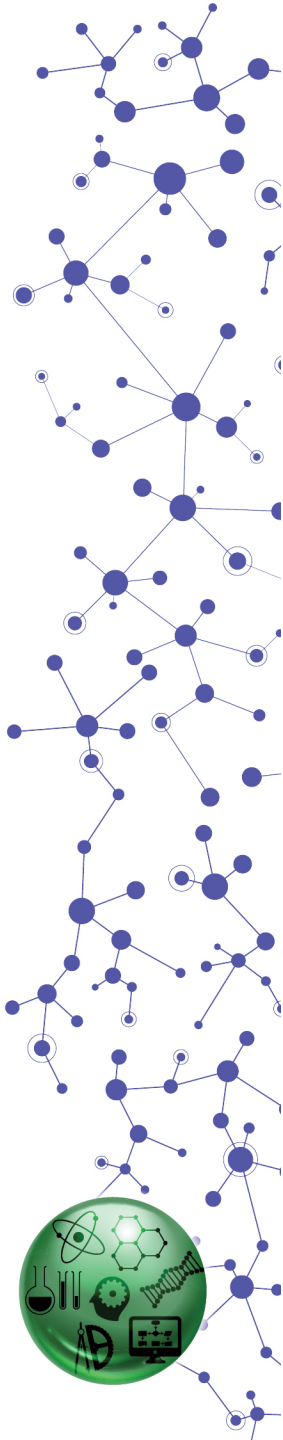
Barba, 2018

Reproducibility and Replicability in Science

Definitions

Reproducibility is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis.

Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.



Reproducibility and Replicability in Science

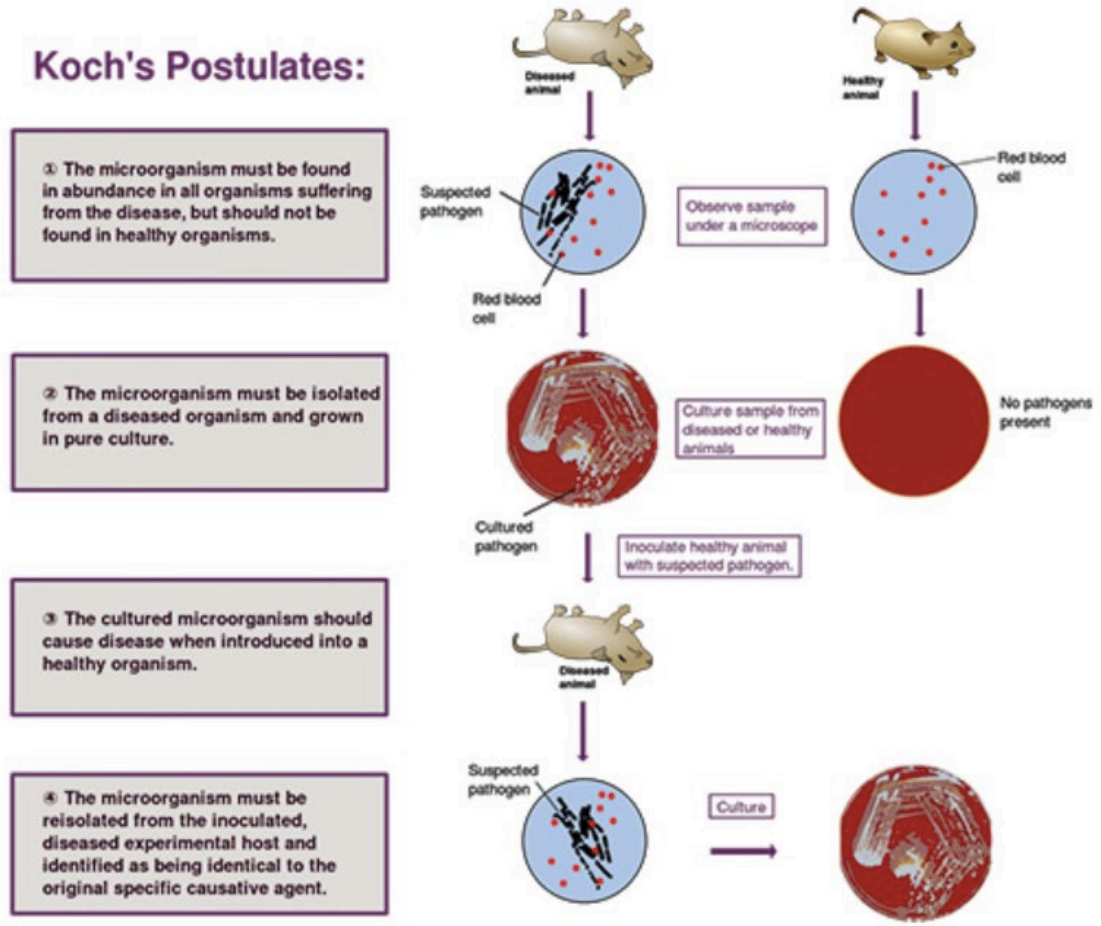
Gaining Confidence in Scientific Results

- Replicability and reproducibility focus on individual studies
- Research synthesis and meta-analysis provide broader review
- Multiple channels of evidence from a variety of studies provide a robust means for gaining confidence in scientific knowledge over time.

The goal of science is to understand the overall effect or inference from a set of scientific studies, not to strictly determine whether any one study has replicated any other.

Reproducibility and Replicability in Science

Example: Affirming the Causes of Infectious Diseases



Source: Aryal, 2019.

Widespread Use of Computation and Data across Science

Here's the moment when the first black hole image was processed, from the eyes of researcher Katie Bouman. #EHTBlackHole #BlackHoleDay #BlackHole (v/@dfbarajas)

LIGO control room

Credit: David Ryder/Bloomberg via Getty Images

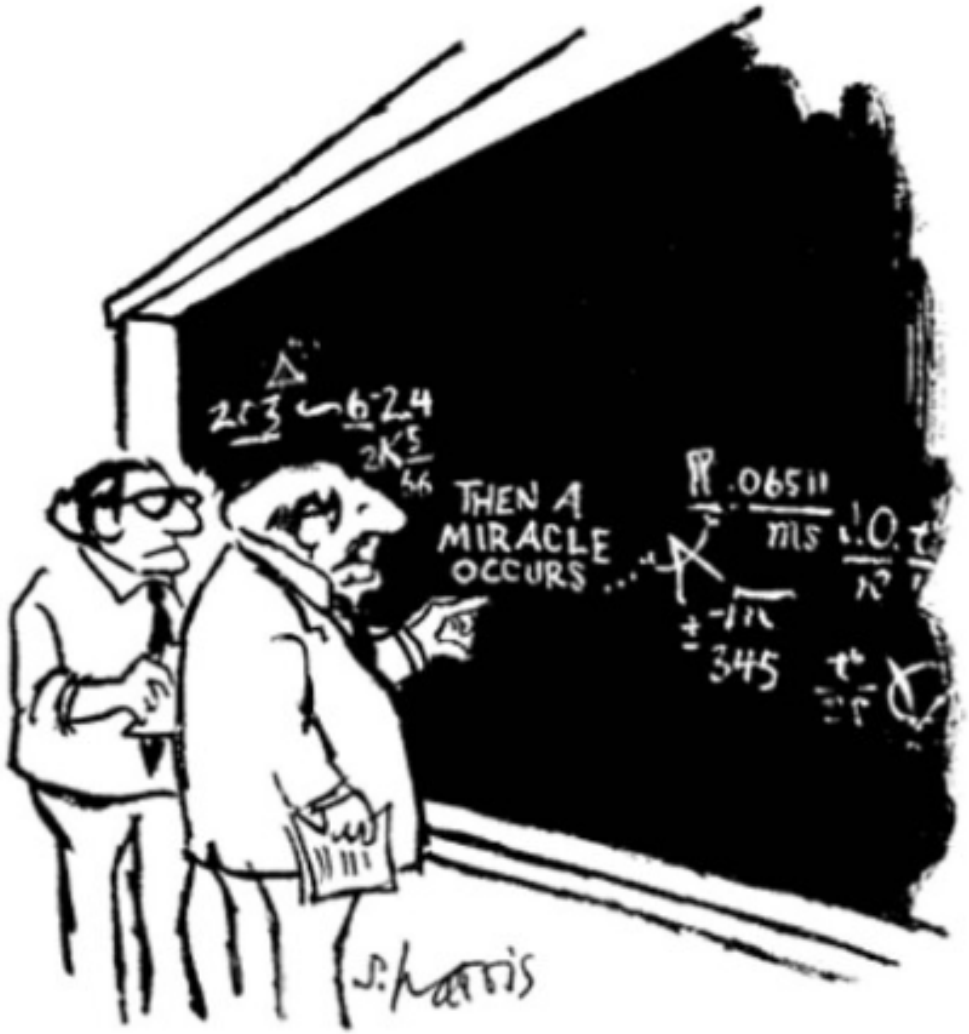
https://twitter.com/MIT_CSAIL/status/1116020858282180609?s=20
https://i1.wp.com/images.firstpost.com/wp-content/uploads/2019/04/Katie-Bowman-1.jpg?w=640&ssl=1

Reproducibility Is Not Always Straightforward

TABLE 4-1 Examples of Reproducibility-Related Studies

Author	Field	Scope of Study	Reported Concerns
Prinz et al. (2011)	Biology (oncology, women's health, cardiovascular health)	Data from 67 projects within Bayer Healthcare	Published data in line with in-house results: ~20 to 25 percent of total projects
Iqbal et al. (2016)	Biomedical	An examination of 441 biomedical studies published between 2000 and 2014	Of 268 papers with empirical data, 267 did not include a link to a full study protocol, and none provided access to all of the raw data used in the study.
Stodden et al. (2018a)	Computational physics	An examination of the availability of artifacts for 307 articles published in the *Journal of Computational Physics*	Over half (50.9 %) of the articles were impossible to reproduce. About 6 percent of the articles (17) made artifacts available in the publication itself, and about 36 percent discussed the artifacts (e.g., mentioned code) in the article.

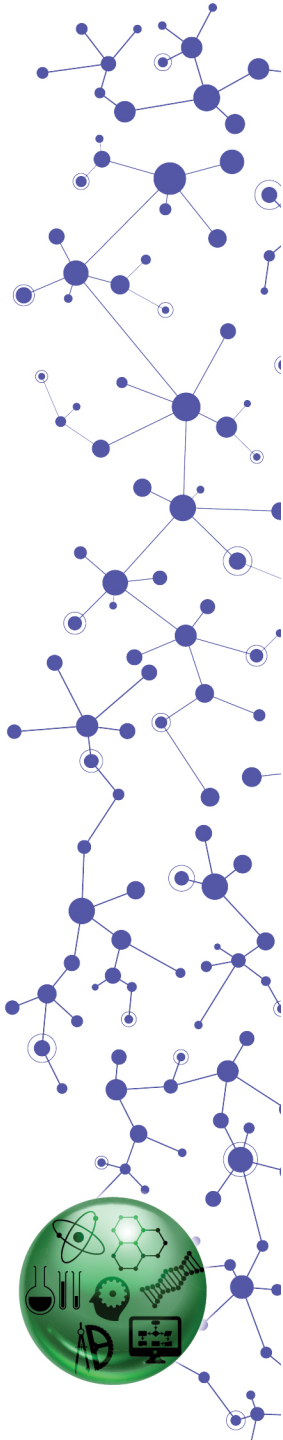
Table 4-1: National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*.



"I think you should be more explicit here in step two."

Sources of Non-Reproducibility

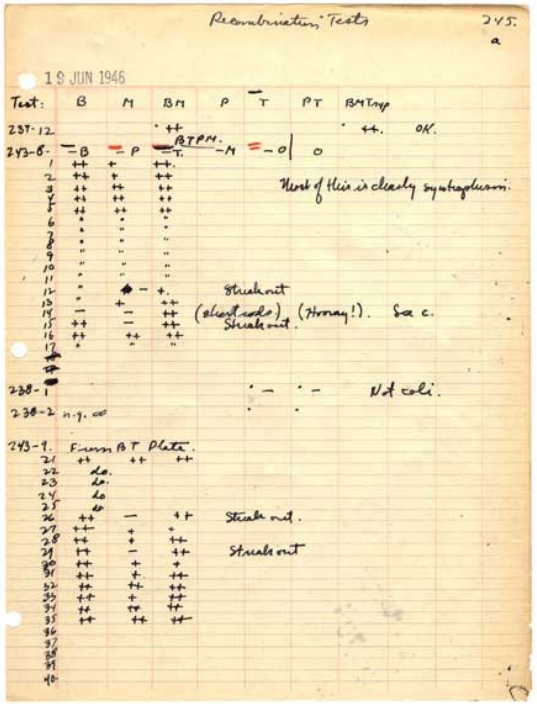
- Inadequate record keeping
- Non-transparent reporting
- Obsolescence of the digital artifacts
- Flawed attempts to reproduce other's results
- Barriers in culture



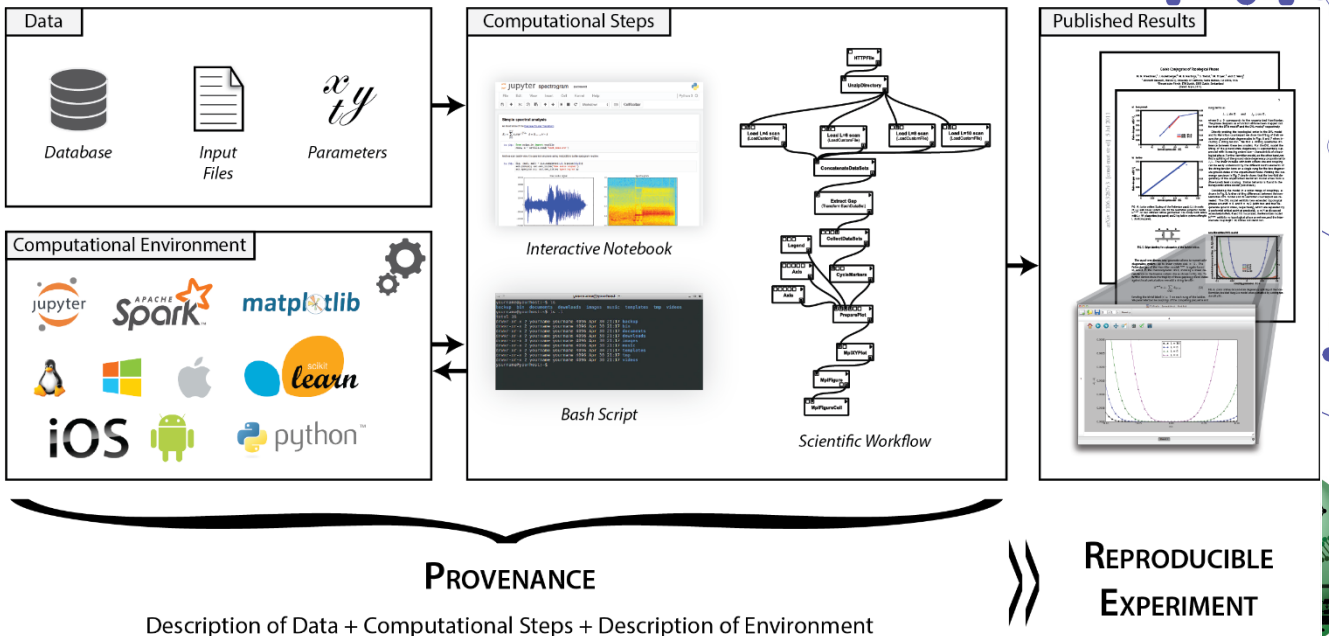
Reproducibility and Replicability in Science

Reproducibility: Challenge

- Experiments are complex and involve many steps: need to systematically capture and report detailed provenance: data, code, computational environment
- Full reproducibility is not always possible: proprietary and non-public data, code and hardware
- Transparency contributes to the confidence in results



DNA recombination
By Lederberg

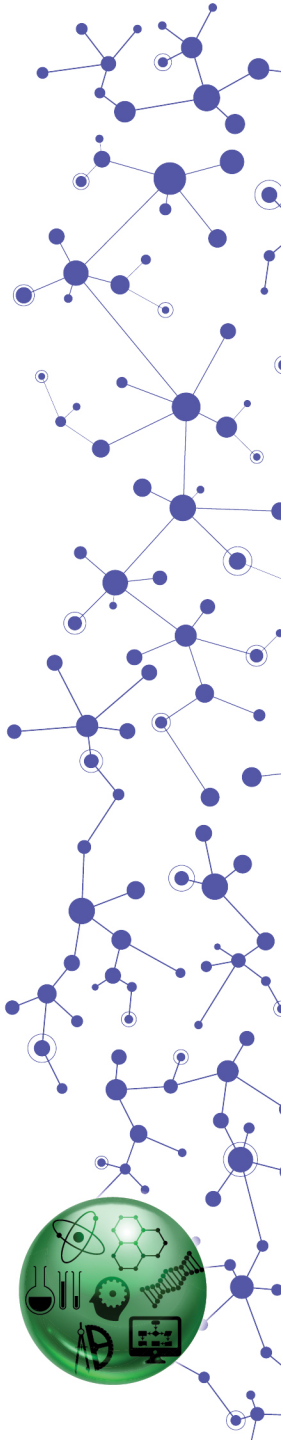


Replicability Is Nuanced

- One can expect bitwise reproducibility, but one does not expect exact replicability
- Some important studies are not amenable to direct replication: Ephemeral phenomena, long-term epidemiological studies
- Many *de facto* replications go unreported as such

Replicability Is Nuanced

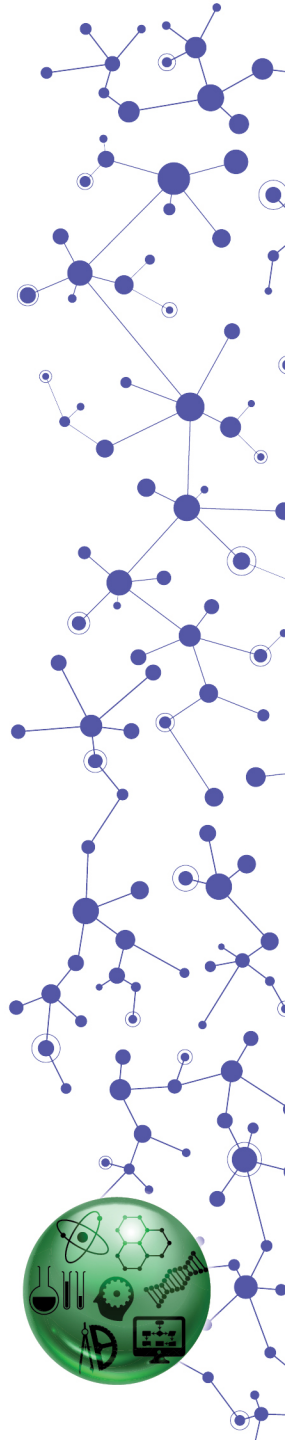
- Non-replicability in any scientific discipline is related to key attributes of the scientific system under study
  - -- Complexity
  - -- Intrinsic variability
  - -- Controllability
  - -- Precision of measurement
- Assess and report uncertainty along with clear, specific and complete reporting of methods
- In tests of replicability, criteria for replication should take account of both the central tendency and variability in results



Reproducibility and Replicability in Science

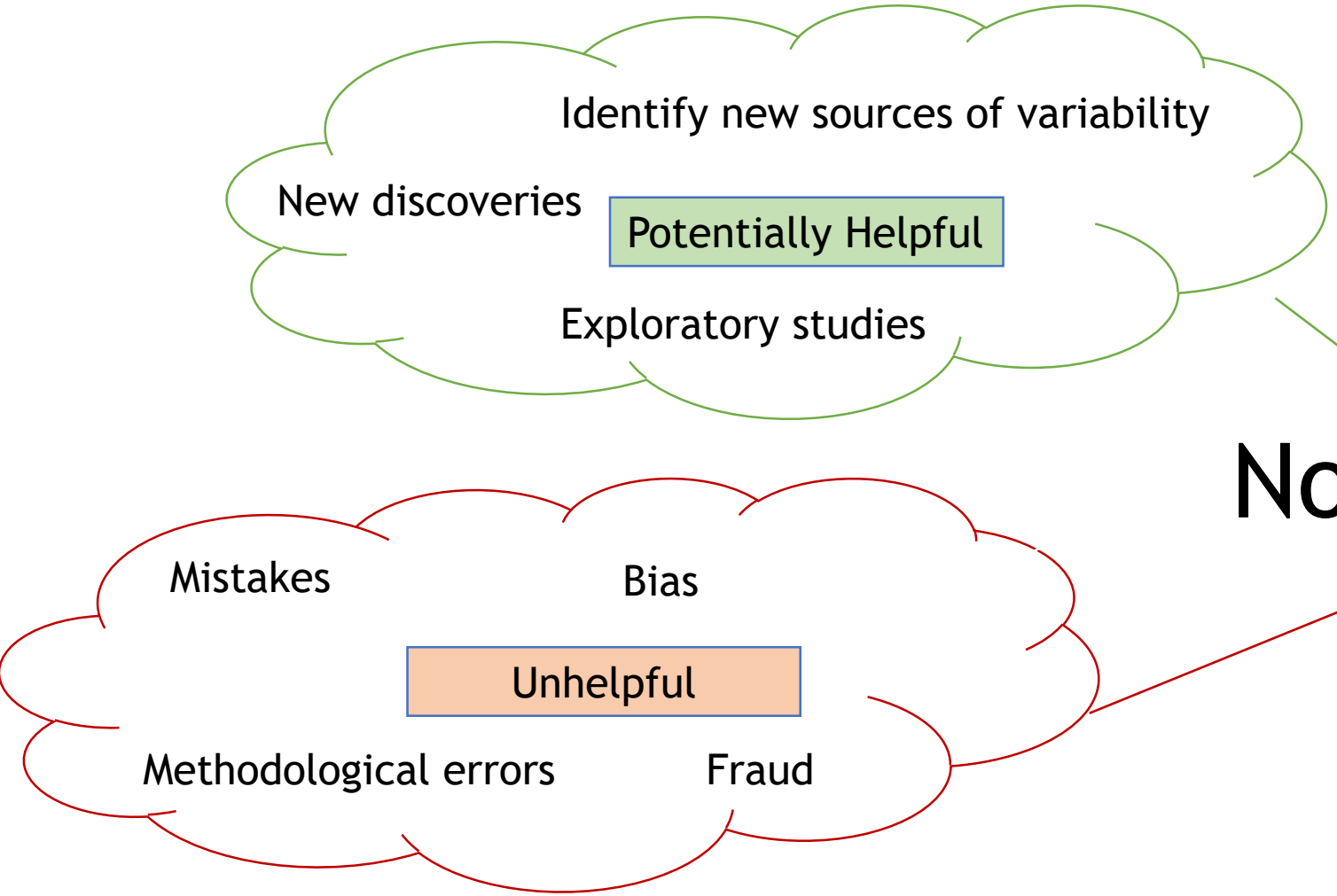
Criteria for Undertaking Replicability Studies

- Importance of the results for policy, decision making, and science
- Unexpected or controversial results, or potential bias
- Recognized weaknesses or flaws in the design, methods, or analysis of the original study
- Costs offset by potential benefits for science and society



Reproducibility and Replicability in Science

Sources of Non-Replicability:
"Potentially Helpful" and "Unhelpful" to the Advancement of Science

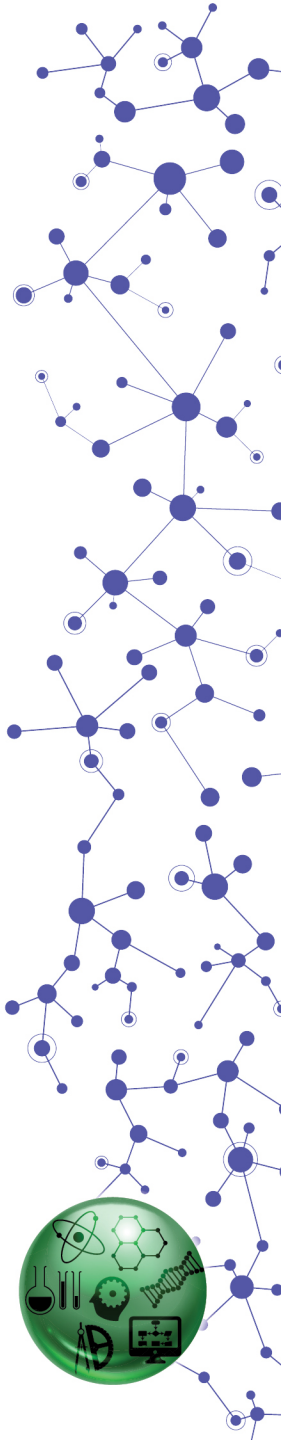


Non-Replicability

Reproducibility and Replicability in Science

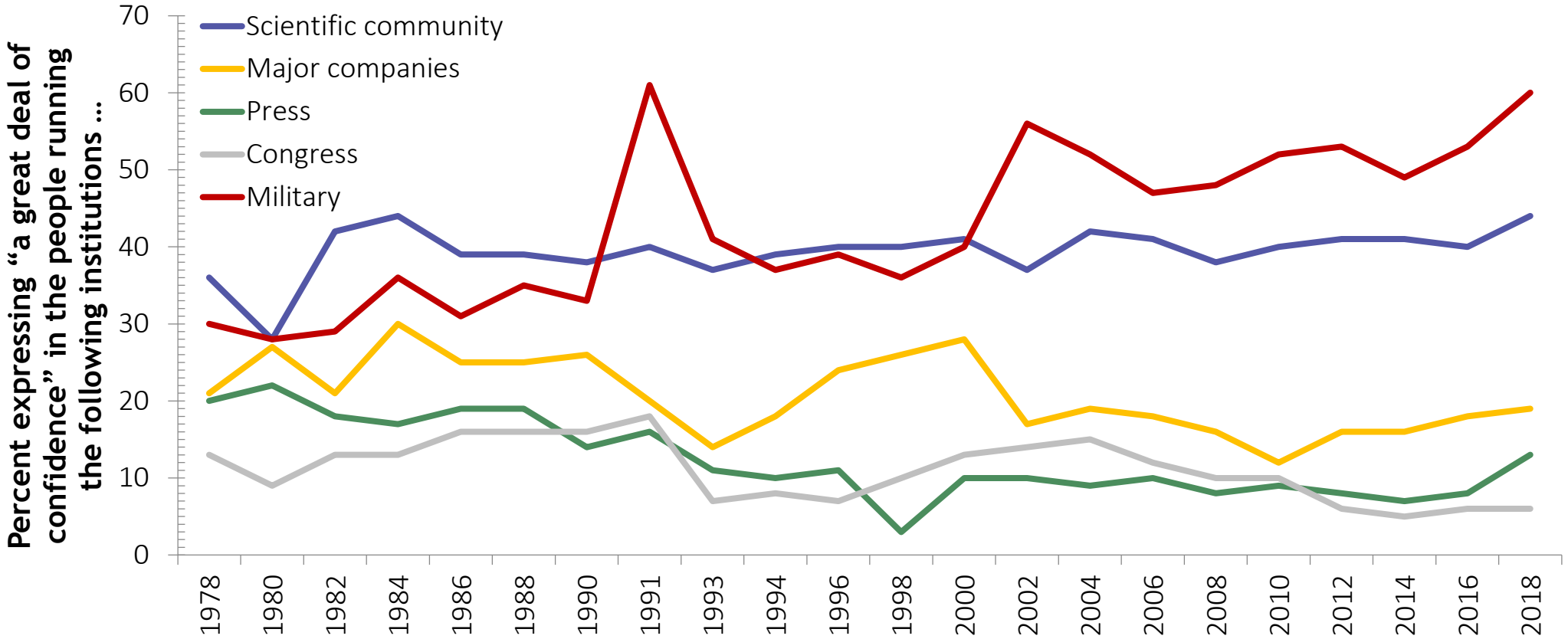
Statistical Inference and Replicability

- Outsized role in the replicability debate
- Misunderstanding and misuse of *p*-values
  - Erroneous calculations
  - Confusion about meaning
  - Excess reliance on arbitrary thresholds of “statistical significance”
  - Bias in reporting
- Meta-analysis and research synthesis



Reproducibility and Replicability in Science

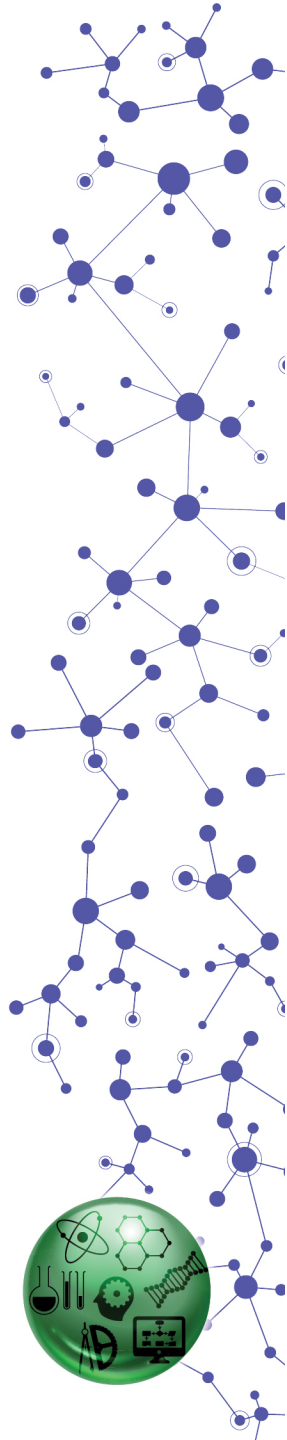
Public Trust



SOURCE: National Science Foundation (2018e, Figure 7-16) and General Social Survey (2018 data from http://gss.norc.org/Get-The-Data).

Key Recommendations for:

- Educational Institutions
- Researchers
- NSF and other funders
- Professional societies
- Journal editors and conference organizers
- Journalists
- Policy makers



Reproducibility and Replicability in Science

Key Recommendations for Educational Institutions

- Educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.

- Include training in the proper use of statistical analysis and inference. Researchers who use statistical inference analyses should learn to use them properly.

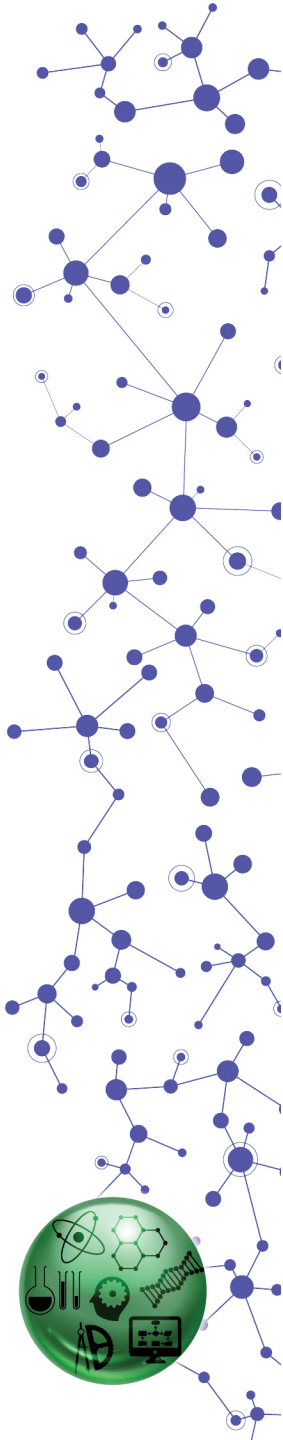
Key Recommendations for Researchers

- Convey clear, specific and complete information about:

  any computational methods, computational environment and data products,

  how the reported result was reached characterization of uncertainties relevant to the study.

- Properly use statistical analysis and inference and in computational methods; adhere to sound methodological practices.
- Collaborate with expert colleagues to meet computational or statistical requirements.
- Avoid overstating the implications of research



Reproducibility and Replicability in Science

Key Recommendations for NSF and Other Funders (1 of 2):

Investments to consider:

- Explore the limits of computational reproducibility
- Promote computational reproducibility
- Support reproducibility tools and infrastructure
- Support training of researchers in best practices and use of these tools.

Key Recommendations for NSF and Other Funders (2 of 2):

- Improve archives and repositories for data, code, and other digital artifacts
- Consider criteria developed to guide investment in replication studies
- Require evaluation of uncertainties as part of grant applications and review of reproducibility and replicability into merit-review criteria

Frege's Letter to Russell

Jena
22 June 1902

Dear Colleague,

Many thanks for your interesting letter of 16 June. I am glad that you agree with me in many things and that you intend to discuss my work in detail. . . .

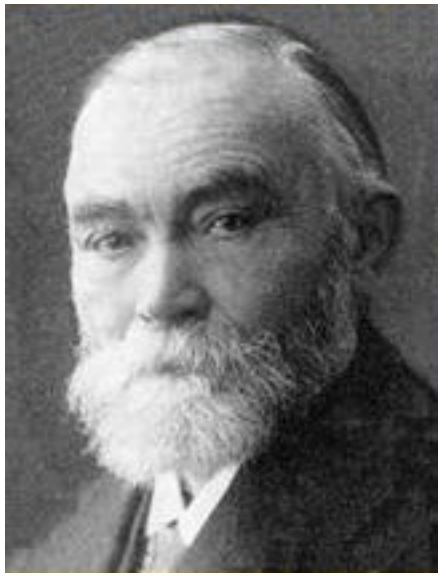
Your discovery of the contradiction has surprised me beyond words and, I should almost like to say, left me thunderstruck, because it has rocked the ground on which I meant to build arithmetic. It seems accordingly that the transformation of the generality of an identity into an identity of ranges of values (sect. 9 of my *Basic Laws*) is not always permissible, that my law V (sect. 20, p. 36) is false, and that my explanations in sect. 31 do not suffice to secure a meaning for my combinations of signs in all cases. I must give some further thought to the matter. It is all the more serious as the collapse of my law V seems to undermine not only the foundations of my arithmetic but the only possible foundations of arithmetic as such. And yet, I should think, it must be possible to set up conditions for the transformation of the generality of an identity into an identity of ranges of values so as to retain the essentials of my proofs. Your discovery is at any rate a very remarkable one, and it may perhaps lead to a great advance in logic, undesirable as it may seem at first sight.

Incidentally, the expression 'A predicate is predicated of itself' does not seem exact to me. A predicate is as a rule a first-level function which requires an object as argument and which cannot therefore have itself as argument (subject). Therefore I would rather say: 'A concept is predicated of its own extension.' If the function $\Phi(\xi)$ is a concept, I designate its extension (or the pertinent class) by '$\dot{\varepsilon}\Phi(\varepsilon)$' (though I now have some doubts about the justification for this). '$\Phi(\dot{\varepsilon}\Phi(\varepsilon))$' or '$\dot{\varepsilon}\Phi(\varepsilon) \cap \dot{\varepsilon}\Phi(\varepsilon)$' is then the predication of the concept $\Phi(\xi)$ of its own extension.

The second volume of my *Basic Laws* is to appear shortly. I shall have to give it an appendix where I will do justice to your discovery. If only I could find the right way of looking at it!

Yours sincerely,
G. Frege

Source: Marcus and McEvoy, 2016

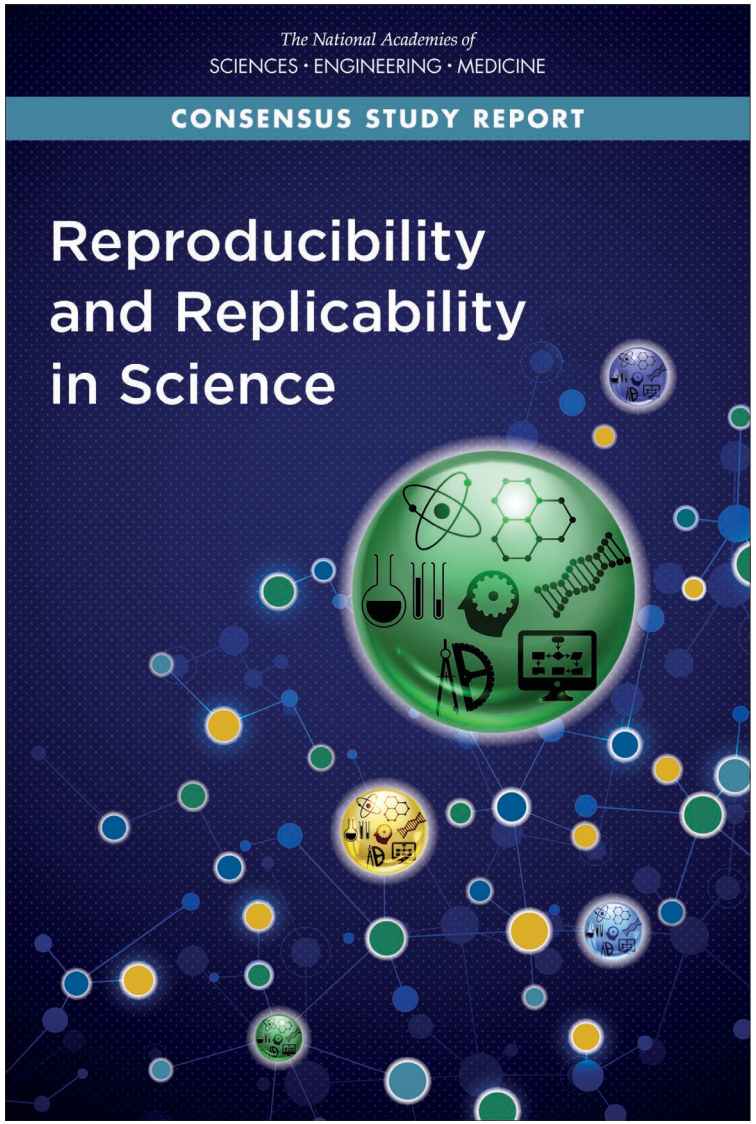


Friedrich Ludwig Gottlob Frege
1848-1925

Science does not aim at establishing immutable truths and eternal dogmas; its aim is to approach the truth by successive approximations, without claiming that at any stage final and complete accuracy has been achieved.

– Bertrand Russell

Reproducibility and Replicability in Science



www.nationalacademies.org/ReproducibilityinScience

Thank you to the sponsors of this study:

National Science Foundation
Alfred P. Sloan Foundation

Reproducibility and Replicability in Science

References

Aryal, S. (2019). Robert Koch and Koch's Postulates. Available: https://microbenotes.com/robert-koch-and-kochs-postulates/ [June 2019].

Barba, L.A. (2018). Terminologies for Reproducible Research. arXiv, 1802.03311. Available: https://arxiv.org/pdf/1802.03311 [December 2018].

Marcus, R., and McEvoy, M. (Eds.). (2016). *An historical introduction to the philosophy of mathematics: A reader*. New York, NY: Bloomsbury.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. Washington, DC: The National Academies Press.

Key Recommendations for Journalists

- Report on scientific results with as much context and nuance as the medium allows.
- Be cautious about scientific reports on complex, hard-to-control systems; when a result is particularly surprising or at odds with existing bodies of research; when a study deals with an emerging area of science with substantial disagreement; when there may be conflicts of interest.

Key Recommendations for Policy Makers

- Be wary of making a serious decision or policy based on results of a single study; be similarly wary of allowing a single contrary study to refute scientific conclusions supported by multiple lines of previous evidence.

Reproducibility and Replicability in Science

Key Recommendations for Professional Societies

- Educate the public and professional members
- Develop and disclose policies
- Require that all research reports include a discussion of uncertainty in measurements and conclusions as a review criterion.

Key Recommendations for Journals and Conference Organizers

- Consider ways to ensure computational reproducibility for publications, to the extent ethically and legally possible. Make and enforce transparency requirements.
- Reserve stronger claims to studies meeting higher levels of reproducibility and replicability.

Key Recommendations for Educational Institutions

- Train students and faculty in computational reproducibility and in proper use of statistical methods.

Key Recommendations for Journalists:

- Report on scientific results with as much context and nuance as the medium allows.
- Be especially cautious about scientific reports when:
  - complex, hard-to-control systems are the subject of study;
  - result is particularly surprising or at odds with existing bodies of research;
  - study deals with an emerging area of science with substantial disagreement; and
  - conflicts of interest may be present.

Key Recommendations for Policy Makers:

- Seek convergent evidence when contemplating a serious decision or policy based on results of a single study;
- Be wary of allowing a single contrary study to refute scientific conclusions supported by multiple lines of previous evidence.

Role of Uncertainty in Replicability

Most scientific inquiries encounter irreducible uncertainties, which can be due to:

- random processes in the system under study
- limits to our understanding or ability to control that system
- limitations in the precision of measurement

Uncertainties or confidence levels should be included in research results so other researchers and stakeholders can correctly interpret the results.

Growing Adoption of Reproducible Science

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Formerly the National Climatic Data Center (NCDC)... more about NCEI »

Home | Climate Information | Data Access | Customer Support

Datasets | Search | Contribute | Products | Perspectives | Outreach | About

Home > Paleoclimatology Data > Paleo Data Search > Study

If you would like to help us understand our user community bet

Global and Regional 500 Year Temperature Reconstructions

Originator:

Abram, N.J.; McGregor, H.V.; Tierney, J.E.; Evans, M.N.; McKay, N.P.; Kaufman, D.S.; Thirumalai, K.

Citation Information:

Abram et al. 2016 Code	Compresse
Abram et al. 2016 Data	Compresse

2016. Early onset of industrial-era warming across the oceans and continents. Nature, 536(7617), 411-418. doi: 10.1038/nature19082

NOAA Study Page:

https://www.ncdc.noaa.gov/paleo/study/20083

JSON Metadata:

https://www.ncdc.noaa.gov/paleo-search/study/search.json?xmlId=17895

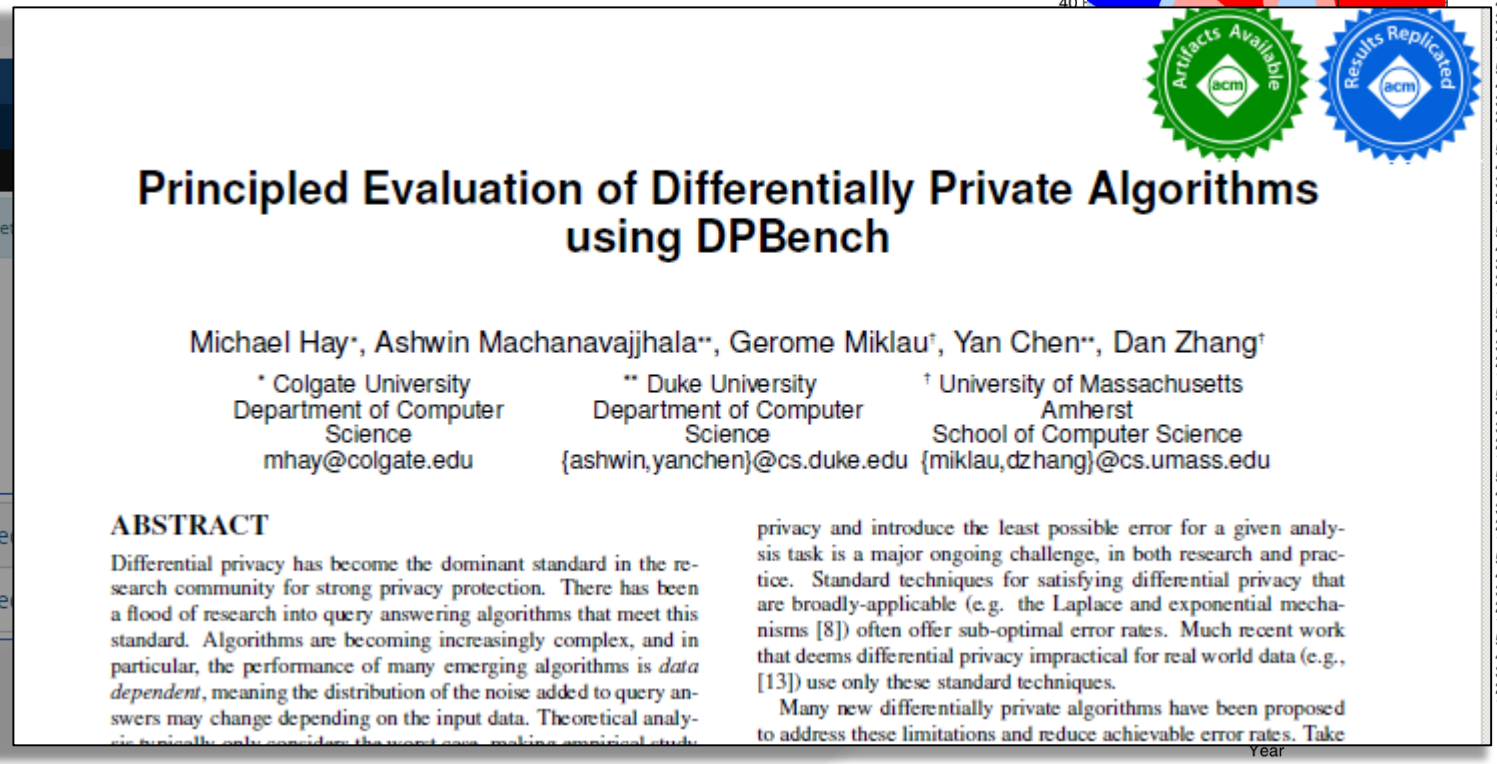
DIF Metadata:

http://www1.ncdc.noaa.gov/pub/data/metadata/published/paleo/dif/xml/noaa-recon-20083.xml

Download Data:

Show Data File Variables

Abram et al. 2016 Code	Compressed ZIP File containing Abram et al. 2016 Code
Abram et al. 2016 Data	Compressed ZIP File containing Abram et al. 2016 Input Data



Artifacts Available acm

Results Replicated acm

Principled Evaluation of Differentially Private Algorithms using DPBench

Michael Hay*, Ashwin Machanavajjhala**, Gerome Miklau†, Yan Chen**, Dan Zhang†

* Colgate University
Department of Computer Science
mhay@colgate.edu

** Duke University
Department of Computer Science
{ashwin,yanchen}@cs.duke.edu

† University of Massachusetts Amherst
School of Computer Science
{miklau,dzhang}@cs.umass.edu

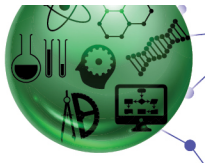
ABSTRACT

Differential privacy has become the dominant standard in the research community for strong privacy protection. There has been a flood of research into query answering algorithms that meet this standard. Algorithms are becoming increasingly complex, and in particular, the performance of many emerging algorithms is *data dependent*, meaning the distribution of the noise added to query answers may change depending on the input data. Theoretical analysis typically only considers the worst case, making empirical study

privacy and introduce the least possible error for a given analysis task is a major ongoing challenge, in both research and practice. Standard techniques for satisfying differential privacy that are broadly-applicable (e.g. the Laplace and exponential mechanisms [8]) often offer sub-optimal error rates. Much recent work that deems differential privacy impractical for real world data (e.g., [13]) use only these standard techniques.

Many new differentially private algorithms have been proposed to address these limitations and reduce achievable error rates. Take

Reproducibility and Replicability in Science



Editorial: Note About Inaccurate Results Published in the *American Journal of Epidemiology* and the *American Journal of Public Health*

This article was jointly published in the *American Journal of Epidemiology* (*Am J Epidemiol.* 2017;185(6): 407–408) and the *AJPH* (*Am J Pub Health.* 2017;107(4):502).

In 2013, Masters et al. published articles in the *American Journal of Public Health* (*AJPH*)[1] and the *American Journal of Epidemiology* (*AJE*)[2] in which they reported results of analyses of data from the National Health Interview Survey that were linked to individual National Death Index mortality records from 1986 to 2006. The two papers, which were related, were about age variation in the association between obesity status and mortality risk in US adults. In the *AJE* article, Masters et al. reported that contrary to

BIASED RATES

letters, the authors concluded that the bias was such that instead of increasing with age, as reported by Masters et al.,[1,2] the proportion of mortality attributable to obesity should have decreased with older age. Masters et al. responded to Wang in the *AJE*[5]; however, the letter by Wang and Liu in the *AJPH*[4] was published alone because Masters et al. declined the invitation to respond.

TO WITHDRAW OR NOT

results of Masters et al. could be used by insurance companies to justify increasing the life insurance premiums of obese persons as they grow older. The results could also lead insurance companies to reduce rates for obese persons younger than 50 years of age because they spuriously suggested that younger obese persons are less likely to die than are nonobese persons.

Reproducibility and Replicability in Science