The Foundations of Rigorous, Reproducible Research in Statistical Research

Victoria Stodden
Department of Industrial and Systems Engineering
University of Southern California

Seminar 1
Professor Maurice H. Belz Lectures
School of Mathematics and Statistics, Faculty of Science
The University of Melbourne
August 12, 2025

USC

Agenda

1. My Background / Introductions
2. A Brief History

<Break>

3. Defining the Problem: Reproducibility Differs in Different Research Settings
4. The Data Science Lifecycle
5. Challenges and Ongoing Work

1. My Background

Educational Experience

Ph.D. Sept 2006. Statistics, Stanford University. Advisor: David Donoho,

Committee Chair: Michael Saunders (Management Science and Engineering)

Committee: Michael Saunders, David Donoho, Brad Efron, Jerry Friedman, Trevor Hastie, and Rob Tibshirani

M.L.S. Dec 2007. Stanford Law School

M.S. June 2000. Statistics, Stanford University

M.S. July 1996. Economics, University of British Columbia

B.Soc.Sci. Dec 1994. Economics (magna cum laude), University of Ottawa

Work Experience

- Associate Professor, Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California
- Associate Professor, School of Information Sciences, University of Illinois at Urbana Champaign
- Assistant Professor, Department of Statistics, Columbia University
- Kauffman Fellow in Law and Innovation at Yale Law School
- Berkman Klein fellow at Harvard Law School

Introductions

2. Reproducibility in Statistical Inference and Data Science: A Brief History

A Brief History

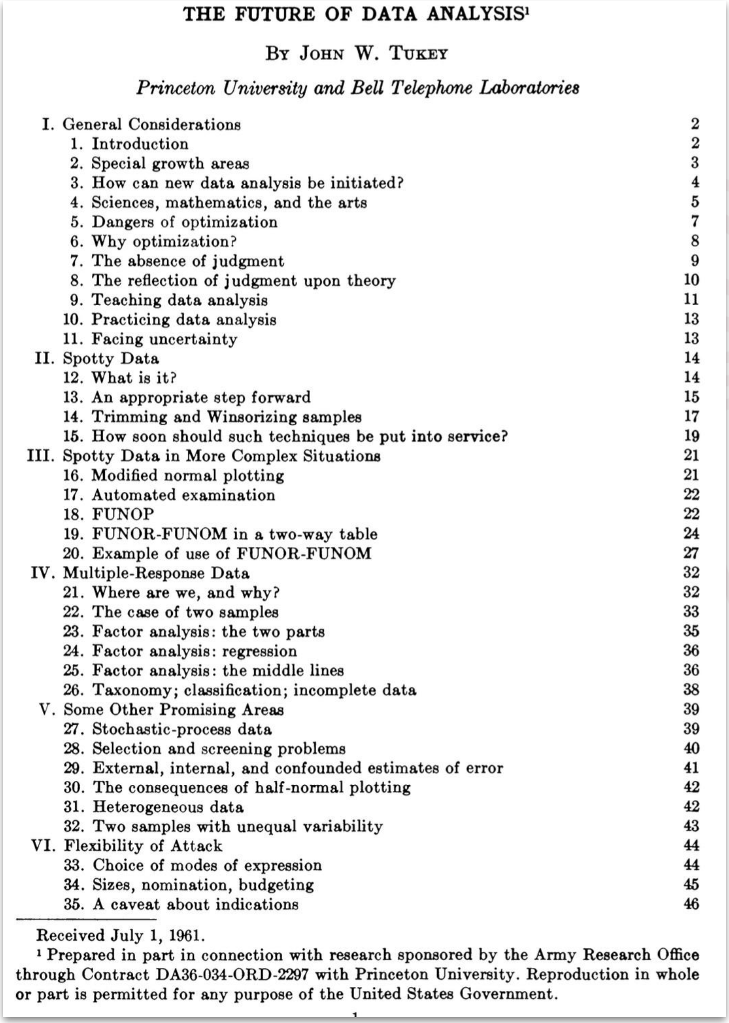
- Researchers leveraged the rise of computation and data collection to advance discovery.

- Researchers could only include English language descriptions of their computational steps and data, since publications afforded no other options.

➢ First sign of computational reproducibility problem

First Steps In Statistics

- John Tukey 1962: “The Future of Data Analysis” Annals of Mathematical Statistics.
  - Advocated for the use of computational methods in data inference
- John Tukey contributed groundbreaking discoveries to statistics, broadened recognition of the field, coined the work ‘bit’ (1947) and developed the Fast Fourier Transform, among others. Worked on computational methods at Bell Labs.



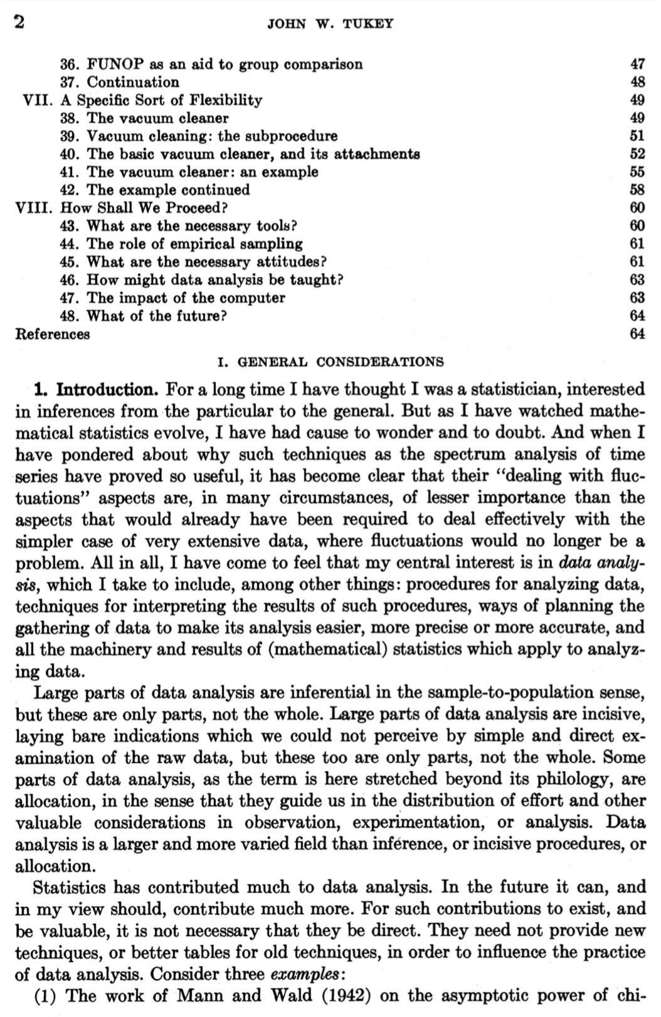
THE FUTURE OF DATA ANALYSIS[1]

By John W. Tukey

Princeton University and Bell Telephone Laboratories

Received July 1, 1961.

[1] Prepared in part in connection with research sponsored by the Army Research Office through Contract DA36-034-ORD-2297 with Princeton University. Reproduction in whole or part is permitted for any purpose of the United States Government.



I. GENERAL CONSIDERATIONS

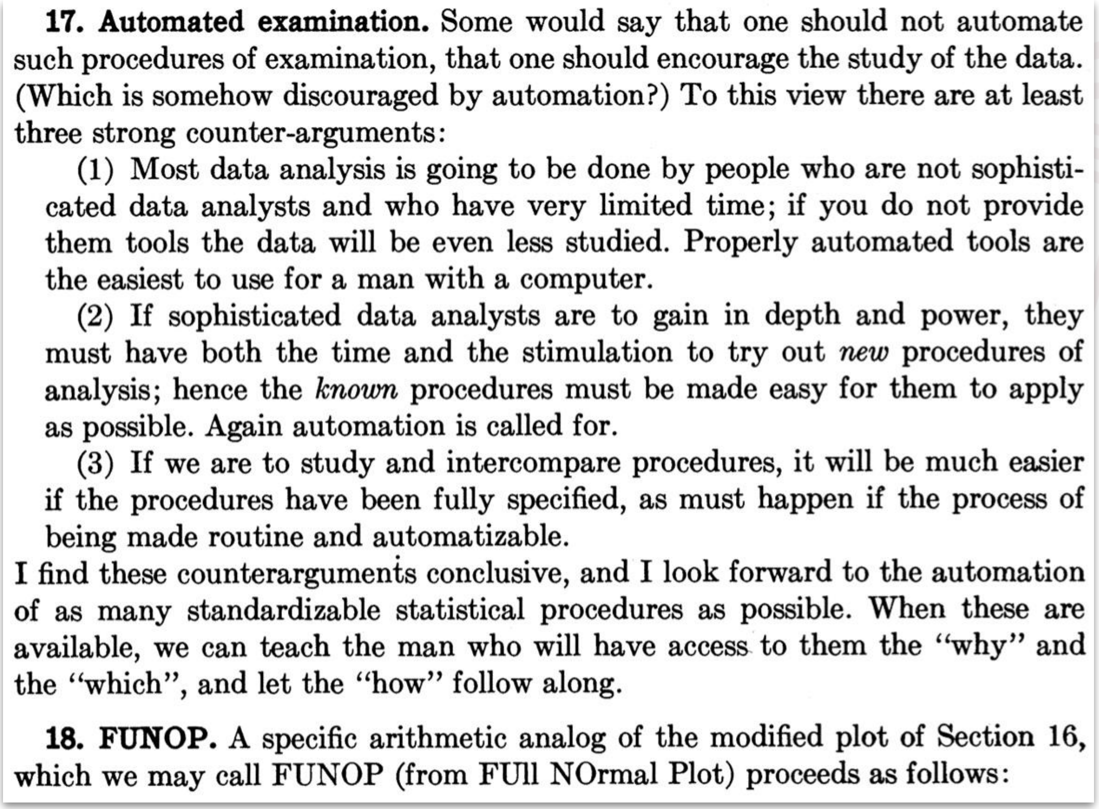
1. Introduction. For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their "dealing with fluctuations" aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data, where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole. Large parts of data analysis are incisive, laying bare indications which we could not perceive by simple and direct examination of the raw data, but these too are only parts, not the whole. Some parts of data analysis, as the term is here stretched beyond its philology, are allocation, in the sense that they guide us in the distribution of effort and other valuable considerations in observation, experimentation, or analysis. Data analysis is a larger and more varied field than inference, or incisive procedures, or allocation.

Statistics has contributed much to data analysis. In the future it can, and in my view should, contribute much more. For such contributions to exist, and be valuable, it is not necessary that they be direct. They need not provide new techniques, or better tables for old techniques, in order to influence the practice of data analysis. Consider three *examples*:

(1) The work of Mann and Wald (1942) on the asymptotic power of chi-

“Automation” is a new idea..



17. Automated examination. Some would say that one should not automate such procedures of examination, that one should encourage the study of the data. (Which is somehow discouraged by automation?) To this view there are at least three strong counter-arguments:

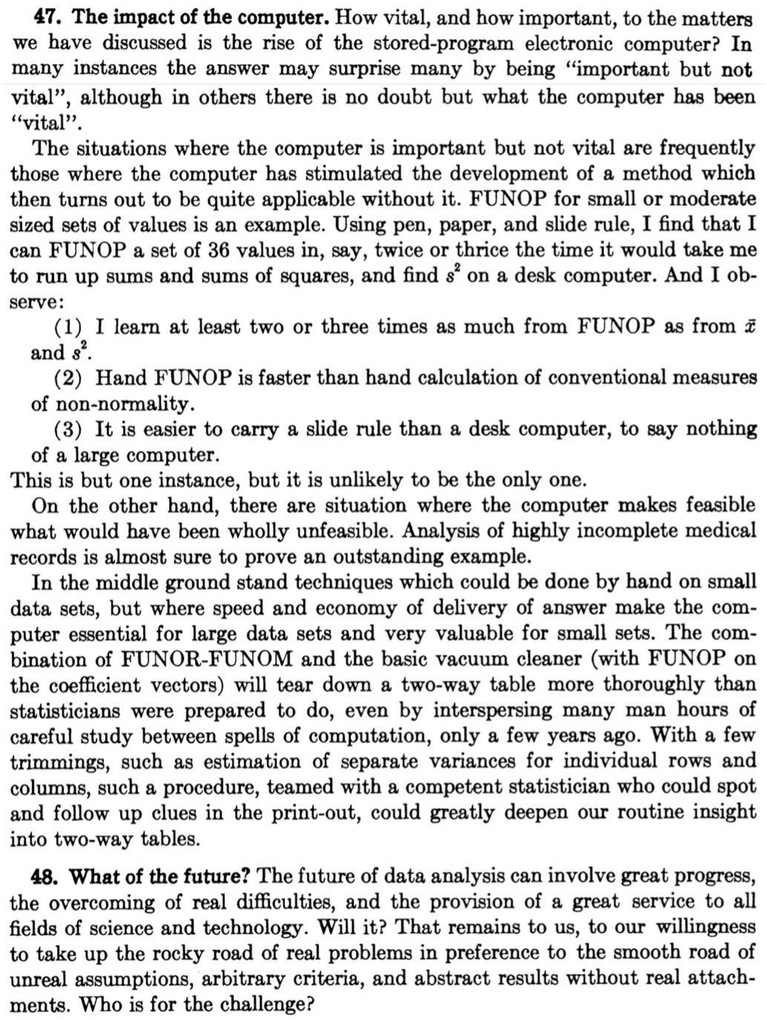
(1) Most data analysis is going to be done by people who are not sophisticated data analysts and who have very limited time; if you do not provide them tools the data will be even less studied. Properly automated tools are the easiest to use for a man with a computer.

(2) If sophisticated data analysts are to gain in depth and power, they must have both the time and the stimulation to try out *new* procedures of analysis; hence the *known* procedures must be made easy for them to apply as possible. Again automation is called for.

(3) If we are to study and intercompare procedures, it will be much easier if the procedures have been fully specified, as must happen if the process of being made routine and automatizable.

I find these counterarguments conclusive, and I look forward to the automation of as many standardizable statistical procedures as possible. When these are available, we can teach the man who will have access to them the “why” and the “which”, and let the “how” follow along.

18. FUNOP. A specific arithmetic analog of the modified plot of Section 16, which we may call FUNOP (from FUll NOrmal Plot) proceeds as follows:



47. The impact of the computer. How vital, and how important, to the matters we have discussed is the rise of the stored-program electronic computer? In many instances the answer may surprise many by being "important but not vital", although in others there is no doubt but what the computer has been "vital".

The situations where the computer is important but not vital are frequently those where the computer has stimulated the development of a method which then turns out to be quite applicable without it. FUNOP for small or moderate sized sets of values is an example. Using pen, paper, and slide rule, I find that I can FUNOP a set of 36 values in, say, twice or thrice the time it would take me to run up sums and sums of squares, and find s^2 on a desk computer. And I observe:

(1) I learn at least two or three times as much from FUNOP as from $\bar{x}$ and s^2.

(2) Hand FUNOP is faster than hand calculation of conventional measures of non-normality.

(3) It is easier to carry a slide rule than a desk computer, to say nothing of a large computer.

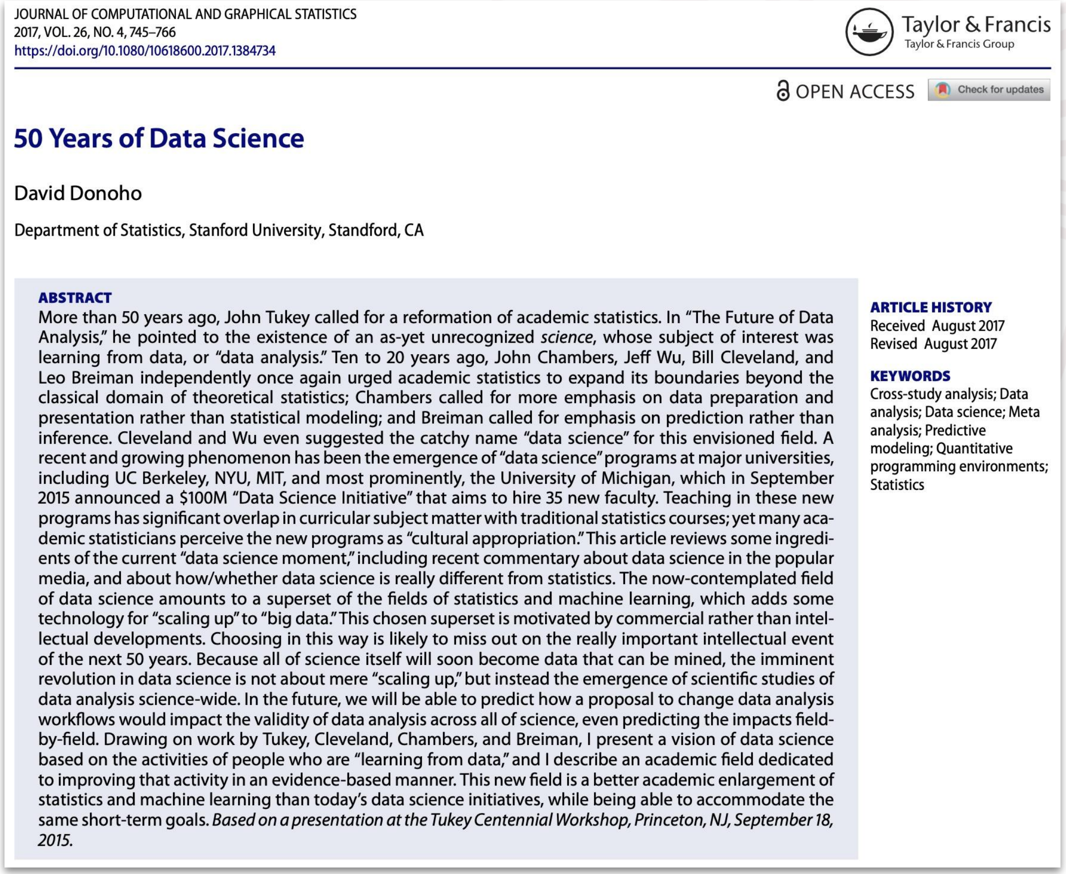
This is but one instance, but it is unlikely to be the only one.

On the other hand, there are situation where the computer makes feasible what would have been wholly unfeasible. Analysis of highly incomplete medical records is almost sure to prove an outstanding example.

In the middle ground stand techniques which could be done by hand on small data sets, but where speed and economy of delivery of answer make the computer essential for large data sets and very valuable for small sets. The combination of FUNOR-FUNOM and the basic vacuum cleaner (with FUNOP on the coefficient vectors) will tear down a two-way table more thoroughly than statisticians were prepared to do, even by interspersing many man hours of careful study between spells of computation, only a few years ago. With a few trimmings, such as estimation of separate variances for individual rows and columns, such a procedure, teamed with a competent statistician who could spot and follow up clues in the print-out, could greatly deepen our routine insight into two-way tables.

48. What of the future? The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it? That remains to us, to our willingness to take up the rocky road of real problems in preference to the smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments. Who is for the challenge?

“50 Years of Data Science” 2017



JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 4, 745–766
https://doi.org/10.1080/10618600.2017.1384734

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Standford, CA

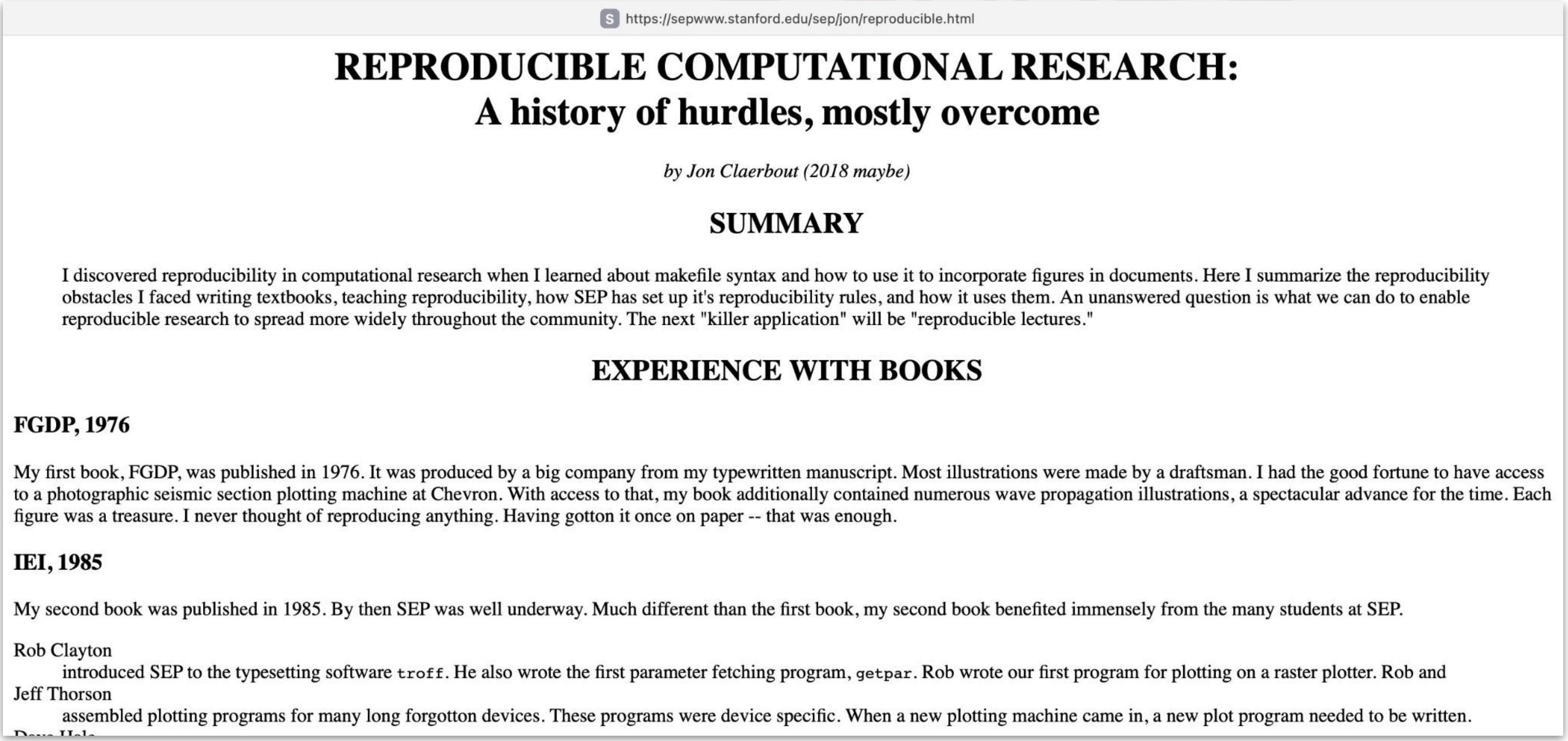
ABSTRACT

More than 50 years ago, John Tukey called for a reformation of academic statistics. In “The Future of Data Analysis,” he pointed to the existence of an as-yet unrecognized *science*, whose subject of interest was learning from data, or “data analysis.” Ten to 20 years ago, John Chambers, Jeff Wu, Bill Cleveland, and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland and Wu even suggested the catchy name “data science” for this envisioned field. A recent and growing phenomenon has been the emergence of “data science” programs at major universities, including UC Berkeley, NYU, MIT, and most prominently, the University of Michigan, which in September 2015 announced a $100M “Data Science Initiative” that aims to hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; yet many academic statisticians perceive the new programs as “cultural appropriation.” This article reviews some ingredients of the current “data science moment,” including recent commentary about data science in the popular media, and about how/whether data science is really different from statistics. The now-contemplated field of data science amounts to a superset of the fields of statistics and machine learning, which adds some technology for “scaling up” to “big data.” This chosen superset is motivated by commercial rather than intellectual developments. Choosing in this way is likely to miss out on the really important intellectual event of the next 50 years. Because all of science itself will soon become data that can be mined, the imminent revolution in data science is not about mere “scaling up,” but instead the emergence of scientific studies of data analysis science-wide. In the future, we will be able to predict how a proposal to change data analysis workflows would impact the validity of data analysis across all of science, even predicting the impacts field-by-field. Drawing on work by Tukey, Cleveland, Chambers, and Breiman, I present a vision of data science based on the activities of people who are “learning from data,” and I describe an academic field dedicated to improving that activity in an evidence-based manner. This new field is a better academic enlargement of statistics and machine learning than today’s data science initiatives, while being able to accommodate the same short-term goals. *Based on a presentation at the Tukey Centennial Workshop, Princeton, NJ, September 18, 2015.*

ARTICLE HISTORY
Received August 2017
Revised August 2017

KEYWORDS
Cross-study analysis; Data analysis; Data science; Meta analysis; Predictive modeling; Quantitative programming environments; Statistics

First Steps in Geophysics



https://sepwww.stanford.edu/sep/jon/reproducible.html

REPRODUCIBLE COMPUTATIONAL RESEARCH:
A history of hurdles, mostly overcome

by Jon Claerbout (2018 maybe)

SUMMARY

I discovered reproducibility in computational research when I learned about makefile syntax and how to use it to incorporate figures in documents. Here I summarize the reproducibility obstacles I faced writing textbooks, teaching reproducibility, how SEP has set up it's reproducibility rules, and how it uses them. An unanswered question is what we can do to enable reproducible research to spread more widely throughout the community. The next "killer application" will be "reproducible lectures."

EXPERIENCE WITH BOOKS

FGDP, 1976

My first book, FGDP, was published in 1976. It was produced by a big company from my typewritten manuscript. Most illustrations were made by a draftsman. I had the good fortune to have access to a photographic seismic section plotting machine at Chevron. With access to that, my book additionally contained numerous wave propagation illustrations, a spectacular advance for the time. Each figure was a treasure. I never thought of reproducing anything. Having gotton it once on paper -- that was enough.

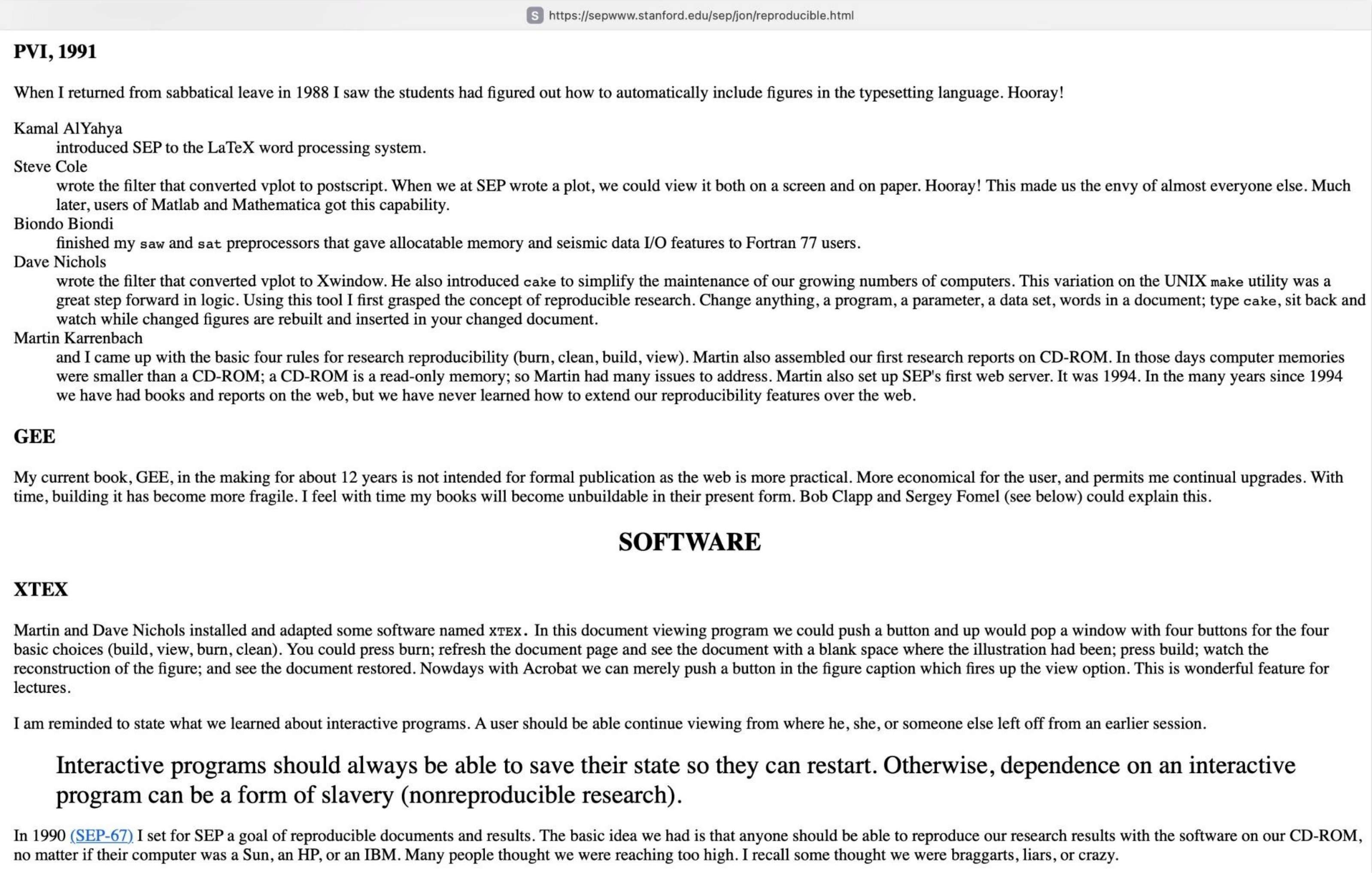
IEI, 1985

My second book was published in 1985. By then SEP was well underway. Much different than the first book, my second book benefited immensely from the many students at SEP.

Rob Clayton
introduced SEP to the typesetting software `troff`. He also wrote the first parameter fetching program, `getpar`. Rob wrote our first program for plotting on a raster plotter. Rob and

Jeff Thorson
assembled plotting programs for many long forgotton devices. These programs were device specific. When a new plotting machine came in, a new plot program needed to be written.

Dave Hale



PVI, 1991

When I returned from sabbatical leave in 1988 I saw the students had figured out how to automatically include figures in the typesetting language. Hooray!

Kamal AlYahya
: introduced SEP to the LaTeX word processing system.

Steve Cole
: wrote the filter that converted vplot to postscript. When we at SEP wrote a plot, we could view it both on a screen and on paper. Hooray! This made us the envy of almost everyone else. Much later, users of Matlab and Mathematica got this capability.

Biondo Biondi
: finished my `saw` and `sat` preprocessors that gave allocatable memory and seismic data I/O features to Fortran 77 users.

Dave Nichols
: wrote the filter that converted vplot to Xwindow. He also introduced `cake` to simplify the maintenance of our growing numbers of computers. This variation on the UNIX `make` utility was a great step forward in logic. Using this tool I first grasped the concept of reproducible research. Change anything, a program, a parameter, a data set, words in a document; type `cake`, sit back and watch while changed figures are rebuilt and inserted in your changed document.

Martin Karrenbach
: and I came up with the basic four rules for research reproducibility (burn, clean, build, view). Martin also assembled our first research reports on CD-ROM. In those days computer memories were smaller than a CD-ROM; a CD-ROM is a read-only memory; so Martin had many issues to address. Martin also set up SEP's first web server. It was 1994. In the many years since 1994 we have had books and reports on the web, but we have never learned how to extend our reproducibility features over the web.

GEE

My current book, GEE, in the making for about 12 years is not intended for formal publication as the web is more practical. More economical for the user, and permits me continual upgrades. With time, building it has become more fragile. I feel with time my books will become unbuildable in their present form. Bob Clapp and Sergey Fomel (see below) could explain this.

SOFTWARE

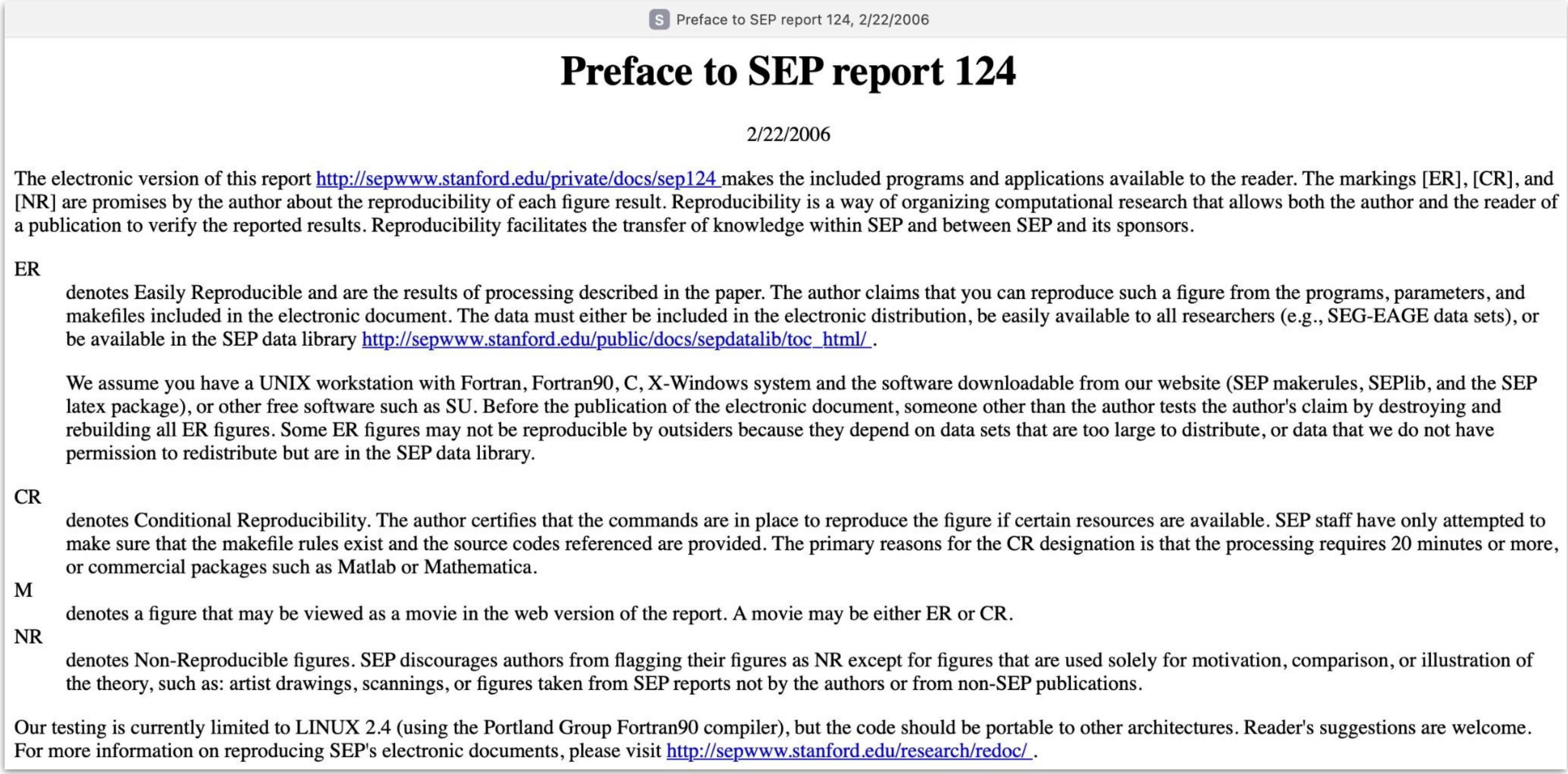
XTEX

Martin and Dave Nichols installed and adapted some software named `XTEX`. In this document viewing program we could push a button and up would pop a window with four buttons for the four basic choices (build, view, burn, clean). You could press burn; refresh the document page and see the document with a blank space where the illustration had been; press build; watch the reconstruction of the figure; and see the document restored. Nowdays with Acrobat we can merely push a button in the figure caption which fires up the view option. This is wonderful feature for lectures.

I am reminded to state what we learned about interactive programs. A user should be able continue viewing from where he, she, or someone else left off from an earlier session.

> Interactive programs should always be able to save their state so they can restart. Otherwise, dependence on an interactive program can be a form of slavery (nonreproducible research).

In 1990 (SEP-67) I set for SEP a goal of reproducible documents and results. The basic idea we had is that anyone should be able to reproduce our research results with the software on our CD-ROM, no matter if their computer was a Sun, an HP, or an IBM. Many people thought we were reaching too high. I recall some thought we were braggarts, liars, or crazy.



Preface to SEP report 124

2/22/2006

The electronic version of this report http://sepwww.stanford.edu/private/docs/sep124 makes the included programs and applications available to the reader. The markings [ER], [CR], and [NR] are promises by the author about the reproducibility of each figure result. Reproducibility is a way of organizing computational research that allows both the author and the reader of a publication to verify the reported results. Reproducibility facilitates the transfer of knowledge within SEP and between SEP and its sponsors.

ER

denotes Easily Reproducible and are the results of processing described in the paper. The author claims that you can reproduce such a figure from the programs, parameters, and makefiles included in the electronic document. The data must either be included in the electronic distribution, be easily available to all researchers (e.g., SEG-EAGE data sets), or be available in the SEP data library http://sepwww.stanford.edu/public/docs/sepdatalib/toc_html/ .

We assume you have a UNIX workstation with Fortran, Fortran90, C, X-Windows system and the software downloadable from our website (SEP makerules, SEPlib, and the SEP latex package), or other free software such as SU. Before the publication of the electronic document, someone other than the author tests the author's claim by destroying and rebuilding all ER figures. Some ER figures may not be reproducible by outsiders because they depend on data sets that are too large to distribute, or data that we do not have permission to redistribute but are in the SEP data library.

CR

denotes Conditional Reproducibility. The author certifies that the commands are in place to reproduce the figure if certain resources are available. SEP staff have only attempted to make sure that the makefile rules exist and the source codes referenced are provided. The primary reasons for the CR designation is that the processing requires 20 minutes or more, or commercial packages such as Matlab or Mathematica.

M

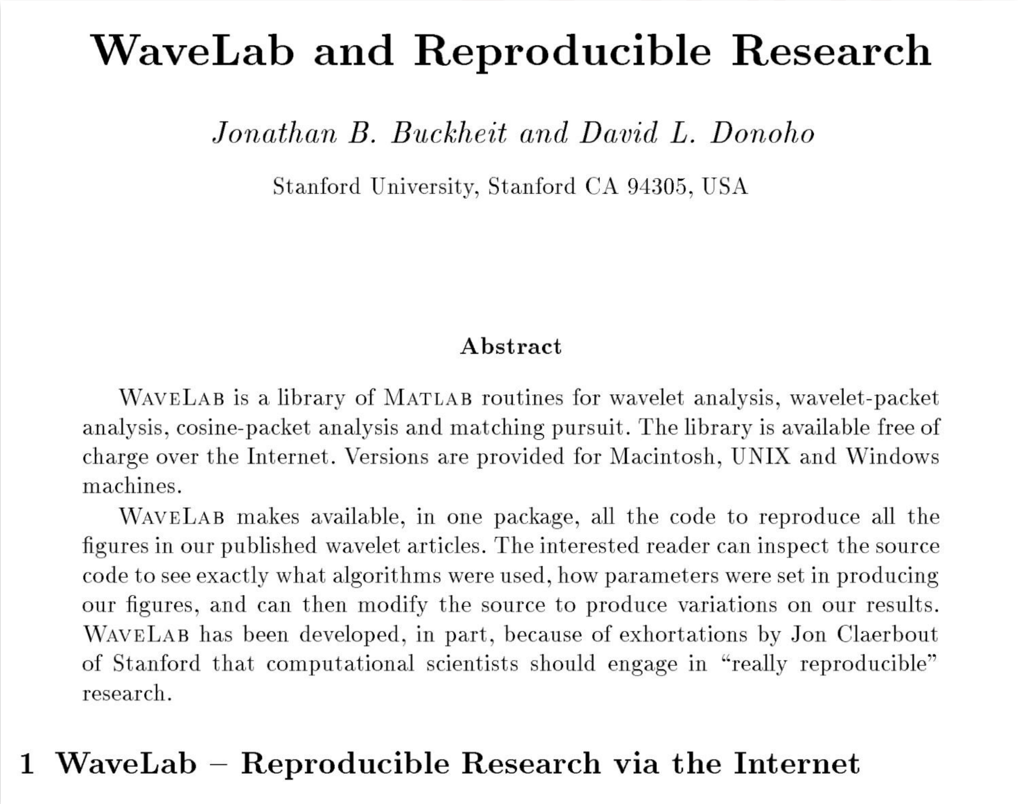
denotes a figure that may be viewed as a movie in the web version of the report. A movie may be either ER or CR.

NR

denotes Non-Reproducible figures. SEP discourages authors from flagging their figures as NR except for figures that are used solely for motivation, comparison, or illustration of the theory, such as: artist drawings, scannings, or figures taken from SEP reports not by the authors or from non-SEP publications.

Our testing is currently limited to LINUX 2.4 (using the Portland Group Fortran90 compiler), but the code should be portable to other architectures. Reader's suggestions are welcome. For more information on reproducing SEP's electronic documents, please visit http://sepwww.stanford.edu/research/redoc/ .

Statistics, 1998: In Part Inspired by Claerbout



WaveLab and Reproducible Research

Jonathan B. Buckheit and David L. Donoho

Stanford University, Stanford CA 94305, USA

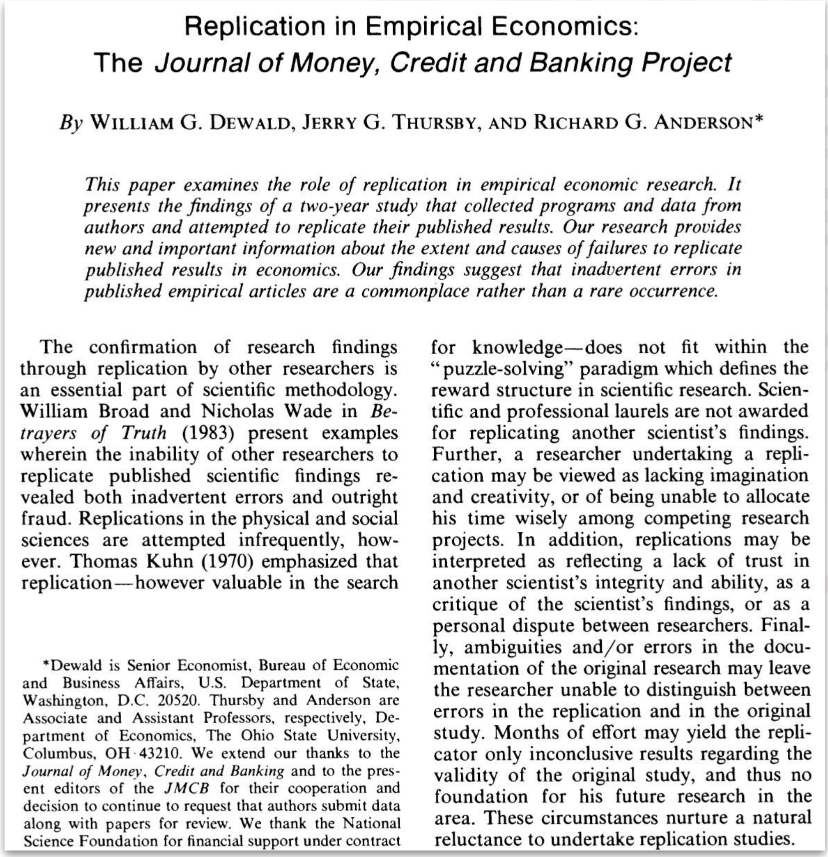
Abstract

WaveLab is a library of Matlab routines for wavelet analysis, wavelet-packet analysis, cosine-packet analysis and matching pursuit. The library is available free of charge over the Internet. Versions are provided for Macintosh, UNIX and Windows machines.

WaveLab makes available, in one package, all the code to reproduce all the figures in our published wavelet articles. The interested reader can inspect the source code to see exactly what algorithms were used, how parameters were set in producing our figures, and can then modify the source to produce variations on our results. WaveLab has been developed, in part, because of exhortations by Jon Claerbout of Stanford that computational scientists should engage in "really reproducible" research.

1 WaveLab – Reproducible Research via the Internet

First Steps in Economics (1986)



Replication in Empirical Economics: The *Journal of Money, Credit and Banking Project*

By William G. Dewald, Jerry G. Thursby, and Richard G. Anderson*

This paper examines the role of replication in empirical economic research. It presents the findings of a two-year study that collected programs and data from authors and attempted to replicate their published results. Our research provides new and important information about the extent and causes of failures to replicate published results in economics. Our findings suggest that inadvertent errors in published empirical articles are a commonplace rather than a rare occurrence.

The confirmation of research findings through replication by other researchers is an essential part of scientific methodology. William Broad and Nicholas Wade in *Betrayers of Truth* (1983) present examples wherein the inability of other researchers to replicate published scientific findings revealed both inadvertent errors and outright fraud. Replications in the physical and social sciences are attempted infrequently, however. Thomas Kuhn (1970) emphasized that replication—however valuable in the search for knowledge—does not fit within the "puzzle-solving" paradigm which defines the reward structure in scientific research. Scientific and professional laurels are not awarded for replicating another scientist's findings. Further, a researcher undertaking a replication may be viewed as lacking imagination and creativity, or of being unable to allocate his time wisely among competing research projects. In addition, replications may be interpreted as reflecting a lack of trust in another scientist's integrity and ability, as a critique of the scientist's findings, or as a personal dispute between researchers. Finally, ambiguities and/or errors in the documentation of the original research may leave the researcher unable to distinguish between errors in the replication and in the original study. Months of effort may yield the replicator only inconclusive results regarding the validity of the original study, and thus no foundation for his future research in the area. These circumstances nurture a natural reluctance to undertake replication studies.

*Dewald is Senior Economist, Bureau of Economic and Business Affairs, U.S. Department of State, Washington, D.C. 20520. Thursby and Anderson are Associate and Assistant Professors, respectively, Department of Economics, The Ohio State University, Columbus, OH 43210. We extend our thanks to the *Journal of Money, Credit and Banking* and to the present editors of the *JMCB* for their cooperation and decision to continue to request that authors submit data along with papers for review. We thank the National Science Foundation for financial support under contract

Other Efforts Emerged..

- Baggerly, Keith A., and Kevin R. Coombes. “Deriving Chemosensitivity From Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology.” The Annals of Applied Statistics, 3(4), 1309–34, (2009).

- Baker, M. “1,500 Scientists Lift the Lid on Reproducibility.” Nature 533, 452–454 (2016).

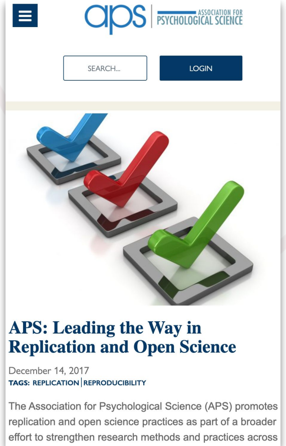
- National Academies of Sciences, Engineering, and Medicine, “Reproducibility and Replicability in Science.” The National Academies Press, (2019).

Reproducibility in Social Psychology

In 2012 an email by Daniel Kahneman was published in Nature revealing reproducibility concerns of “priming” studies in social psychology. A constellation of questions had arisen regarding such studies, and several highly visible cases of fraud

Since then several initiatives in psychology have arisen to take on these challenges

CENTER FOR OPEN SCIENCE

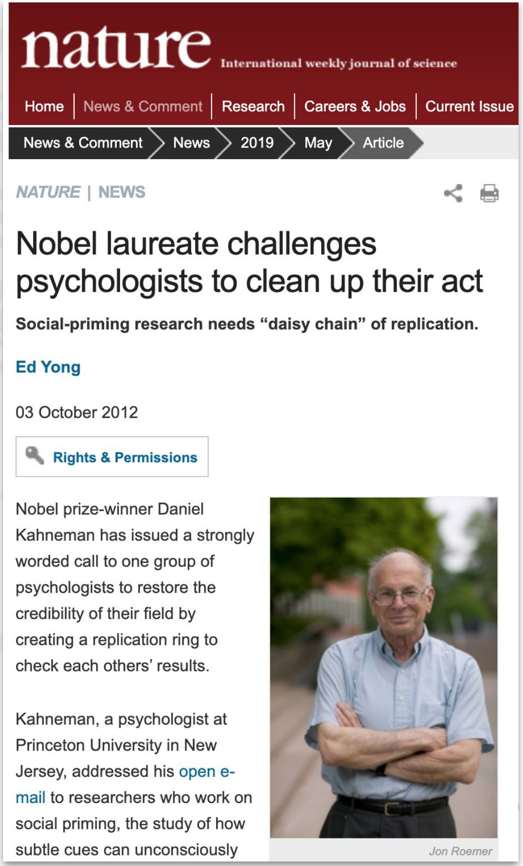


APS: Leading the Way in Replication and Open Science

December 14, 2017

TAGS: REPLICATION | REPRODUCIBILITY

The Association for Psychological Science (APS) promotes replication and open science practices as part of a broader effort to strengthen research methods and practices across



Nobel laureate challenges psychologists to clean up their act

Social-priming research needs “daisy chain” of replication.

Ed Yong

03 October 2012

Nobel prize-winner Daniel Kahneman has issued a strongly worded call to one group of psychologists to restore the credibility of their field by creating a replication ring to check each others’ results.

Kahneman, a psychologist at Princeton University in New Jersey, addressed his open e-mail to researchers who work on social priming, the study of how subtle cues can unconsciously

USC

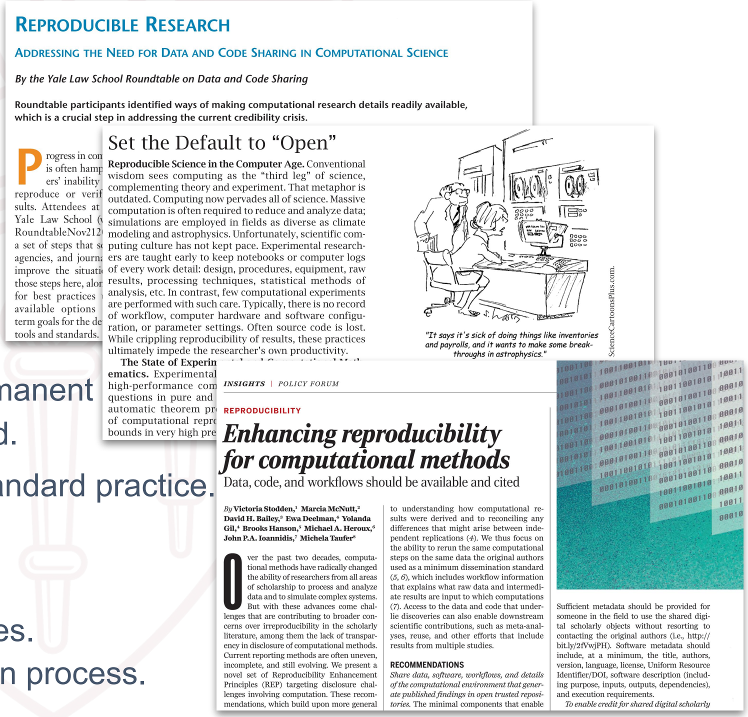
Break

3. Defining the Problem

Reproducibility Standards Development

Community Efforts: AAAS 2016 Workshop on Code and Modeling Reproducibility recommended:

- **Share** data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.
- **Persistent links** should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.
- To enable credit for shared digital scholarly objects, **citation** should be standard practice.
- To facilitate reuse, adequately **document** digital scholarly artifacts.
- Use **Open Licensing** when publishing digital scholarly objects.
- Funding agencies should instigate new research programs and pilot studies.
- Journals should conduct a **reproducibility check** as part of the publication process.



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.

Progress in com
is often hamp
ers' inability
reproduce or verif
sults. Attendees at
Yale Law School (
RoundtableNov212
a set of steps that s
agencies, and journa
improve the situati
those steps here, alon
for best practices
available options
term goals for the de
tools and standards.

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 Experim
ematics. Experimental
high-performance com
questions in pure and
automatic theorem pr
of computational repro
bounds in very high pre

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

ScienceCartoonsPlus.com.

INSIGHTS | *POLICY FORUM*

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 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

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

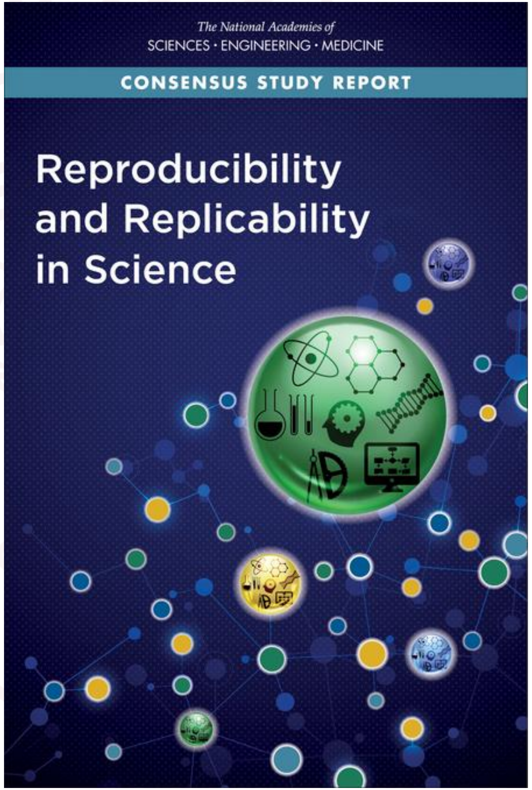
Stodden, McNutt, Bailey, Deelman, Gil, Hanson, Heroux, Ioannidis, Taufer (2016). Enhancing Reproducibility for Computational Methods. Science.

National Academies Consensus Report 2019

“Reproducibility and Replication in Science”

- 15 distinguished members (I was a committee member)
- Chair: Harvey Fineberg, President of Gordon and Betty Moore Foundation
- Stakeholder input: over 50 individuals representing a range of disciplines

⇨ Produced key definitions and several recommendations.



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science

Report and white papers available at https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

Committee Charge

- Define reproducibility and replicability accounting for the diversity of fields in science and engineering.

- Examine state of contemporary science with regard to reproducibility and replication.

- Determine if lack of replication and reproducibility impacts the overall health of science and engineering as well as the public’s perception of these fields.

- Make recommendations for improving rigor and transparency in scientific and engineering research.

Reproducibility Definitions

- ***Reproducibility*** is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with "**computational reproducibility**."

- ***Replicability*** is obtaining **consistent results across studies** aimed at answering the same scientific question, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.

Recommendation 4-1(Transparency)

To help ensure the reproducibility of computational results, **researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results** in order to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment:

- **the input data** used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;

- **a detailed description of the study methods (ideally in executable form)** together with its computational steps and associated parameters; and

- **information about the computational environment** where the study was originally executed, such as operating system, hardware architecture, and library dependencies (which are relationships described in and managed by a software dependency manager tool to mitigate problems that occur when installed software packages have dependencies on specific versions of other software packages).

Recommendation 6-6 (Coordination)

Many stakeholders have a role to play in improving computational reproducibility, including educational institutions, professional societies, researchers, and funders.

- Educational institutions should educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.
- Professional societies should take responsibility for educating the public and their professional members about the importance and limitations of computational research. Societies have an important role in educating the public about the evolving nature of science and the tools and methods that are used.
- Researchers should collaborate with expert colleagues when their education and training are not adequate to meet the computational requirements of their research.
- In line with its priority for "harnessing the data revolution," the National Science Foundation (and other funders) should consider funding of activities to promote computational reproducibility.

Recommendation 6-7 (Publishers)

Journals and scientific societies requesting submissions for conferences should **disclose** their policies relevant to achieving reproducibility and replicability. The strength of the claims made in a journal article or conference submission should reflect the reproducibility and replicability standards to which an article is held, with stronger claims reserved for higher expected levels of reproducibility and replicability.

Journals and conference organizers are encouraged to:

- set and implement desired standards of reproducibility and replicability
- adopt policies to reduce the likelihood of non-replicability
- require as a review criterion that all research reports include a thoughtful discussion of the uncertainty in measurements and conclusions.

Recommendation 6-8 (Funding Initiatives)

Many considerations enter into decisions about what types of scientific studies to fund, including striking a balance between exploratory and confirmatory research. If private or public funders choose to invest in initiatives on reproducibility and replication, two areas may benefit from additional funding:

- *education and training initiatives* to ensure that researchers have the knowledge, skills, and tools needed to conduct research in ways that adhere to the highest scientific standards; that describe methods clearly, specifically, and completely; and that express accurately and appropriately the uncertainty involved in the research;

- *reviews of published work*, such as testing the reproducibility of published research, conducting rigorous replication studies, and publishing sound critical commentaries.

Recommendation 6-3 (Tools and Training)

Funding agencies and organizations should consider investing in research and development of **open-source, usable tools and infrastructure that support reproducibility** for a broad range of studies across different domains in a seamless fashion.

Concurrently, investments would be helpful in outreach to inform and **train researchers** on best practices and how to use these tools.

Recommendation 6-5 (Repositories)

In order to facilitate the transparent sharing and availability of digital artifacts, such as data and code, for its studies, the National Science Foundation (NSF) should:

- Develop a set of **criteria for trusted open repositories** to be used by the scientific community for objects of the scholarly record.
- Seek to **harmonize with other funding agencies** the repository criteria and data-management plans for scholarly objects.
- Endorse or consider creating code and data repositories for **long-term archiving** and preservation of digital artifacts that support claims made in the scholarly record based on NSF-funded research. These archives could be based at the institutional level or be part of, and harmonized with, the NSF-funded Public Access Repository.
- Consider extending NSF’s current **data-management plan** to include other digital artifacts, such as software.
- Work with communities reliant on non-public data or code to develop **alternative mechanisms** for demonstrating reproducibility.

Recommendation 6-9 (Proposal Review)

Funders should require a thoughtful discussion in grant applications of how uncertainties will be evaluated, along with any relevant issues regarding replicability and computational reproducibility.

Funders should introduce review of reproducibility and replicability guidelines and activities into their merit-review criteria, as a low-cost way to enhance both.

Recommendation 6-10 (Funding Replication)

When funders, researchers, and other stakeholders are considering whether and where to direct resources for replication studies, they should consider the following criteria:

- The scientific results are important for individual decision-making or for policy decisions.
- The results have the potential to make a large contribution to basic scientific knowledge.
- The original result is particularly surprising, that is, it is unexpected in light of previous evidence.
- There is controversy about the topic.
- There was potential bias in the original investigation, due, for example, to the source of funding.
- There was a weakness or flaw in the design, methods, or analysis of the original study.
- The cost of a replication is offset by the potential value in reaffirming the original results.
- Future expensive and important studies will build on the original scientific results.

Recommendation 7-1 & 7-2 (Communication)

RECOMMENDATION 7-1: Scientists should take care to **avoid overstating** the implications of their research and also exercise caution in their review of press releases, especially when the results bear directly on matters of keen public interest and possible action.

RECOMMENDATION 7-2: Journalists should report on scientific results with as much **context and nuance** as the medium allows. In covering issues related to replicability and reproducibility, journalists should help their audiences understand the differences between non-reproducibility and non- replicability due to fraudulent conduct of science and instances in which the failure to reproduce or replicate may be due to evolving best practices in methods or inherent uncertainty in science. Particular care in reporting on scientific results is warranted when:

- the scientific system under study is complex and with limited control over alternative explanations or confounding influences;
- a result is particularly surprising or at odds with existing bodies of research;
- the study deals with an emerging area of science that is characterized by significant disagreement or contradictory results within the scientific community; and
- research involves potential conflicts of interest, such as work funded by advocacy groups, affected industry, or others with a stake in the outcomes.

Recommendation 7-3 (Context)

Anyone making personal or policy decisions based on scientific evidence should be wary of making a serious decision based on the results, no matter how promising, of a single study.

Similarly, no one should take a new, single contrary study as refutation of scientific conclusions supported by multiple lines of previous evidence.

4. The Data Science LifeCycle

Developing Frameworks for Policy

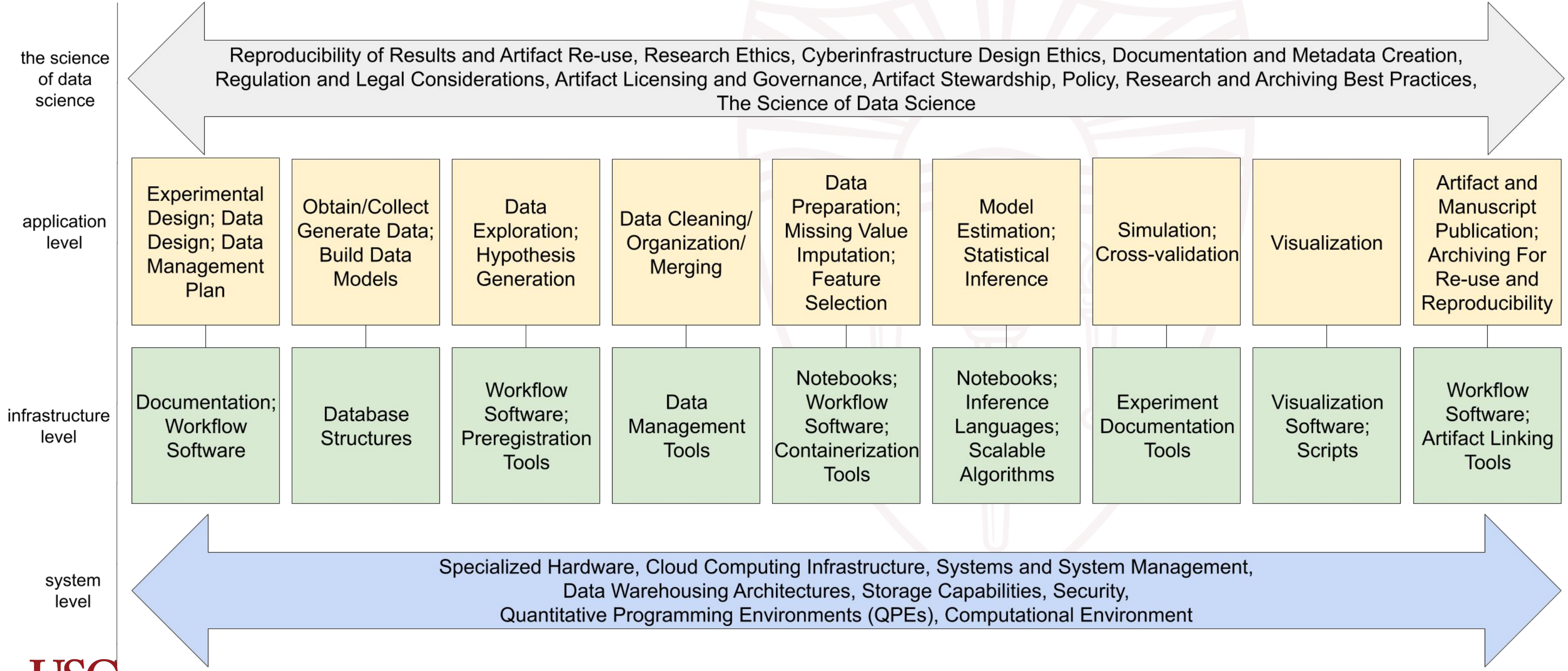
“Lifecycle of Data” is an abstraction from the Information Sciences

- Describes and relates actors in the ecosystem of data use and re-use.

What if we applied this idea to data-enabled science?

- **Clarify steps** in research projects: people/skills involved, tools and infrastructure, and reproducibility through the cycle.
- **Holistically guide implementations**: infrastructure, ethics, reproducibility and sources of uncertainty, curricula, training, and other programmatic initiatives.
- **Develop and reward contributing areas**.

A Proposal: Lifecycle of Data Science



the science of data science

Reproducibility of Results and Artifact Re-use, Research Ethics, Cyberinfrastructure Design Ethics, Documentation and Metadata Creation, Regulation and Legal Considerations, Artifact Licensing and Governance, Artifact Stewardship, Policy, Research and Archiving Best Practices, The Science of Data Science

application level

- Experimental Design; Data Design; Data Management Plan
- Obtain/Collect Generate Data; Build Data Models
- Data Exploration; Hypothesis Generation
- Data Cleaning/ Organization/ Merging
- Data Preparation; Missing Value Imputation; Feature Selection
- Model Estimation; Statistical Inference
- Simulation; Cross-validation
- Visualization
- Artifact and Manuscript Publication; Archiving For Re-use and Reproducibility

infrastructure level

- Documentation; Workflow Software
- Database Structures
- Workflow Software; Preregistration Tools
- Data Management Tools
- Notebooks; Workflow Software; Containerization Tools
- Notebooks; Inference Languages; Scalable Algorithms
- Experiment Documentation Tools
- Visualization Software; Scripts
- Workflow Software; Artifact Linking Tools

system level

Specialized Hardware, Cloud Computing Infrastructure, Systems and System Management, Data Warehousing Architectures, Storage Capabilities, Security, Quantitative Programming Environments (QPEs), Computational Environment

V. Stodden (2020). The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science. CACM.

A Proposed Formalism: The “Tale”

What information do we need to reproduce and verify computational findings?

- Manuscript
  - source or reference
- Documentation
  - README, codebook, install instructions, user guide, etc.
  - License, copyright, permissions
- Code
  - Preprocessing, analysis, workflow
- Data
  - By copy, by reference, data access protocol
- Results
  - Output, figures, tables
- Environment
  - Hardware, OS, compilers, dependent software
  - Runtime, image, container
- Provenance
  - Computational, archival
- Metadata
  - Identifiers, related artifacts, Domain metadata
  - Badges
- Version

Chard et al. (2019) Implementing Computational Reproducibility in the Whole Tale Environment. P-RECS '19: Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems

Challenges Across the Community

- Incentives and culture change: rewards for reproducible research; potentially enabling bad behaviors e.g. data and software capture, minimal value add, ignoring or quashing disruption.
- Relating data and software e.g. LLMs.
- Upskilling in the era of Data Science / Data Inference / Data Collection / Data Visualization / Data Policy / Data Ethics / Data CI / AI.
- Cost/benefit/risk analysis.
- Public perception of science.
- Funding long term curation and archiving.

Challenge: IP and Transparency

Researchers generally don't resolve IP issues regarding their research products.

- ⇨ Funding agency policy setting (in cooperation with institutions and other stakeholders).

Public access to research artifacts and scholarly information data, support of scholarly norms. "Giving back."

- ⇨ "Reproducible Research Standard" (Stodden 2008)

Legal Issues in Data

- In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).
- Copyright adheres to raw facts in Europe.
- The possibility of a residual copyright in data (attribution licensing or public domain certification).
- Legal mismatch: What constitutes a “raw” fact anyway?

Copyright in Australia

Not too different than the US:

“Copyright protects the original forms or way an idea or information is expressed, not the idea or information itself. The most common form of copyright are writing, visual images, music and moving images.”

“Copyright provides an owner of a material with exclusive economic rights to do certain acts with that material. These rights include the right to copy and the right to communicate the material to the public.”

“Copyright also provides authors and performers with non-economic rights, known as moral rights. Moral rights recognised in Australia are the right of integrity, the right of attribution and the right against false attribution.”

“Copyright subsists in a dataset or database where the work of an author, in reducing that compilation to material form (including digital form) involves some intellectual activity that is directed not at collecting or inputting the data, but in expressing the work. Accordingly, a given dataset may or may not be subject to copyright.”

“As it is difficult to determine whether copyright subsists in a dataset, it is recommended that agencies apply a copyright licence. The licence should state that the data or dataset is subject to the terms of the licence to the extent that the data or dataset is protected by copyright.”

See https://www.ag.gov.au/rights-and-protections/copyright/copyright-basics and https://www.data.vic.gov.au/datavic-access-policy-guidelines/licensing-datasets

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 Creative Commons Attribution License **CC BY**,
- Release code components under **MIT License** or similar,
- Release data to public domain (Creative Common Public Domain **CC0**) or attach attribution license.

➡ *Remove copyright's barrier to reproducible research and,*

➡ *Realign the IP framework with longstanding scientific norms.*

Long Term Goals for Reproducibility?

An **integrated computable scholarly record** that is queryable e.g.:

- Show a table of effect sizes and p-values in all vaccination/autism studies published after 1997;
- 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; and
- Randomly re-assign 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.

M. Gavish, D. Donoho, and A. Onn. (2013) Dream applications of verifiable computational results. XRDS, 19, 3.

Reproducibility Issues?

- What have you encountered in:
  - Your readings
  - Your research
  - The classroom
  - Funding
  - Publishing
  - Collaborators?

Continued on Thursday…

- Publisher and Funding Agency Requirements
- Foundational and emerging frameworks for social, legal, and ethical issues e.g. Australian Code for the Responsible Conduct of Research
- Australian Research Council (ARC) data management requirements, and writing data management plans (DMPs).
- Interpreting the FAIR principles (Findability, Accessibility, Interoperability, and Reusability).
- Leveraging repositories and tools to create and openly share reproducible research.
- Privacy and confidentiality
- Leveraging an open code / open data future for research
- Code Requirements; Repository Funding and Integration; Teaching and Curriculum; AI-enabled Data Science.