Reproducible Research: a Perspective

Arnaud Legrand
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP

TILECS Workshop, July 2019

Common beliefs

- *RR mainly allows to fight scientific misconduct (e.g., fraud). That's nice but I'm honnest so just let me do my work!*

Common beliefs

- *RR mainly allows to fight scientific misconduct (e.g., fraud). That's nice but I'm honnest so just let me do my work!*
- *RR is all about re-executing the same code, even if the code is stupid and makes wrong computation. It's pointless!*

Common beliefs

- *RR mainly allows to fight scientific misconduct (e.g., fraud). That's nice but I'm honnest so just let me do my work!*
- *RR is all about re-executing the same code, even if the code is stupid and makes wrong computation. It's pointless!*
- *My student has done everything with org-jupyter-studio-mode. Now he's gone and I can't reuse what he did. See, what's the point? RR does not help!*

Common beliefs

- *RR mainly allows to fight scientific misconduct (e.g., fraud). That's nice but I'm honnest so just let me do my work!*
- *RR is all about re-executing the same code, even if the code is stupid and makes wrong computation. It's pointless!*
- *My student has done everything with org-jupyter-studio-mode. Now he's gone and I can't reuse what he did. See, what's the point? RR does not help!*
- *RR is about controling and checking everything, which slows down the scientific discovery process. Changing the way we work and publish may be harmful!*

Past

First Appearance

Claerbout & Karrenbach, meeting of the Society of Exploration Geophysics, 1992

Electronic Documents Give Reproducible Research a New Meaning **RE1.3**

Jon F. Claerbout and Martin Karrenbach, Stanford Univ.

SUMMARY

A revolution in education and technology transfer follows from the marriage of word processing and software command scripts. In this marriage an author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters, and programs. This provid[illegible] of reproducibility in compute[illegible]

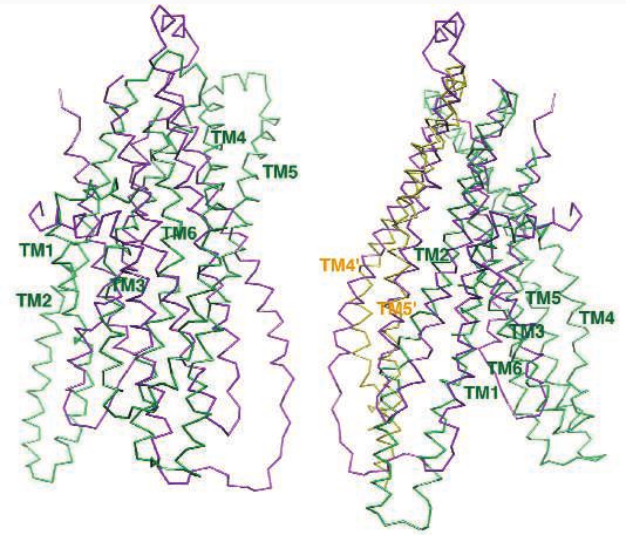
In 1990, we set this sequence of goals:

- Learn how to merge a publication with its underlying computational analysis.
- Teach researchers how to prepare a document in a form where they themselves can reproduce their own research results a year or more later by "pressing a single button".
- Learn how to leave finished work in a condition where coworkers can reproduce the calculation including the final illustration by pressing a button in its caption.
- Prepare a complete copy of our local software environment so that graduating students can take their work away with them to other sites, press a button, and reproduce their Stanford work.
- Merge electronic documents written by multiple authors (SEP reports).
- make incremental improvements in electronic-document software
- seek partners for broadening standards (and making incremental improvements).

Our basic goal is reproducible research. The electronic document is our means to this end. In principle, reproducibility in research can be achieved without electronic documents and that is how we started. Our first nonelectronic reproducible document was a textbook in which the paper document contained the name of a program script in every figure caption. The program scripts were organized by book chapter and section so they could be correlated to an accompanying magnetic tape dump of the file system. The magnetic tape also contained all the necessary data to feed the program script.

Now that we have begun using CD-ROM publication, we can go much further. Every figure caption contains a pushbutton that jumps to the appropriate science directory (folder) and initiates a figure rebuild command and then displays the figure, possibly as a movie or interactive program. We normally display seismic images of the earth's interior. but to reach wider audiences, Figure 1 shows a satellite weather picture which the pushbutton will animate as seen on commerical television. We include all our plot software as well as freely available software from many sources, including compilers and the LaTeX word processing system. Naturally

Unfortunate mistakes



Geoffrey Chang (Scripps, UCSD) works on crystalography and studies the structure of cell membrane proteins.
He specialized in structures of multidrug resistant transporter proteins in bacteria:
MsbA de Escheria Choli (Science, 2001),
Vibrio cholera (Mol. Biology, 2003),
Salmonella typhimurium (Science, 2005)

2006: Inconsistencies reveal a programming mistake

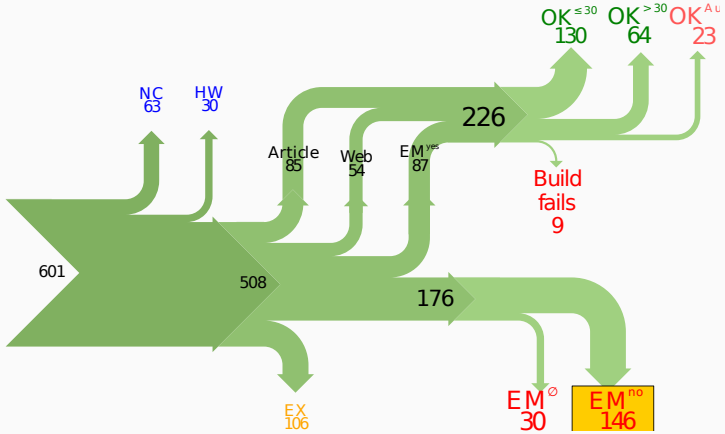
> *a homemade data-analysis program had flipped two columns of data, inverting the electron-density map from which his team had derived the protein structure.*

5 retractations that motivate improved software engineering practices in computational biology

Whistle Blowing

”clinical trials in oncology have the highest failure rate [only 5% are licensed] compared with other therapeutic areas [..] difficulty to repeat experiments even in the laboratory of the original investigator”

Begley and Ellis, Nature, 2012. *Raise standards for preclinical cancer research*



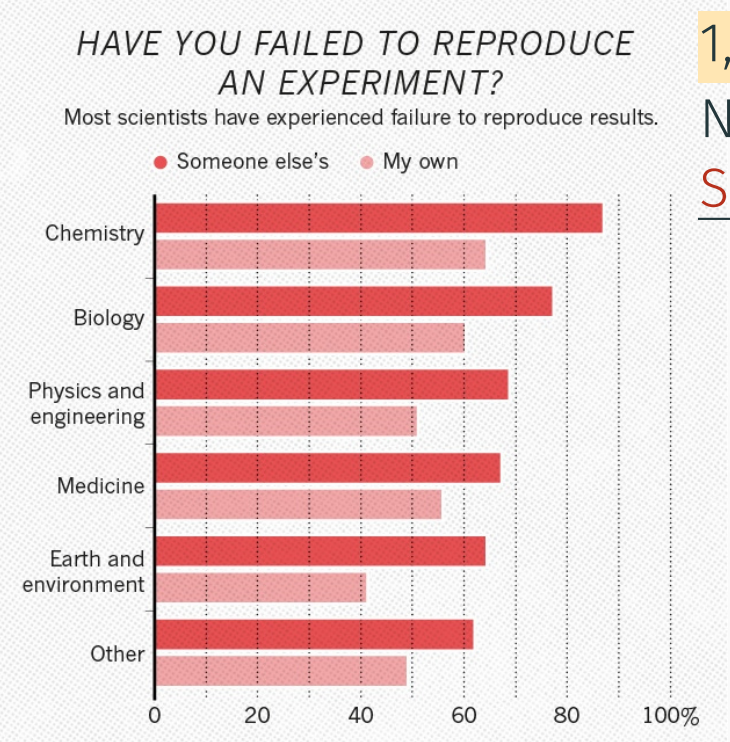
8 ACM conferences and 5 journals

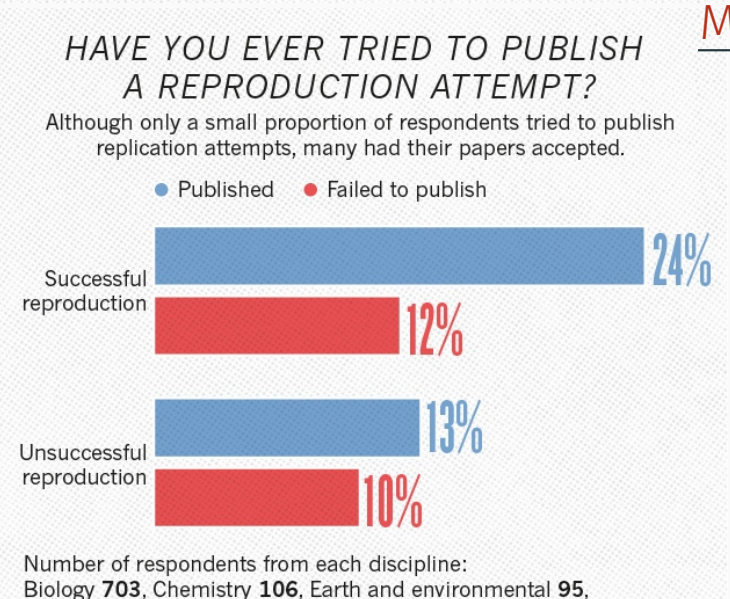
EM^{no} = the code cannot be provided

Collberg, Christian *et al.*, *Measuring Reproducibility in Computer Systems Research*, 2015

The scientific process demands the highest standards of **quality**, **ethics** and **rigor**.

Why are scientific studies so difficult to reproduce?





1,500 scientists lift the lid on reproducibility, Nature, May 2016

Social causes

- Fraud, conflict of interest (pharmaceutic, ...)
- No incentive to reproduce/check our own work (afap), nor the work of others (big results!), nor to allow others to check (competition)
- Peer review does not scale: 1+ million articles per year!

Methodological or technical causes

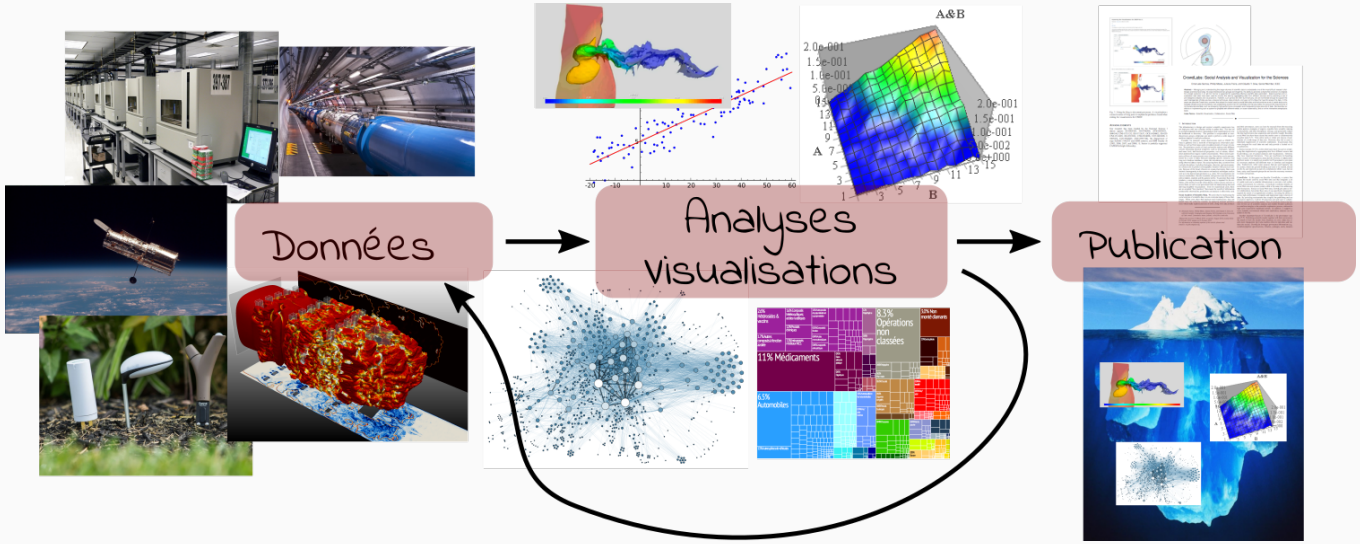
- The many biases (apophenia, confirmation, hindsight, experimenter, ...): bad designs
- Selective reporting, weak analysis (statistics, data manipulation mistakes, computational errors)
- Lack of information, code/raw data unavailable

Different concerns

Social Sciences, Oncology, ... methodology, statistics

Genomics software engineering, computational reproducibility, provenance, ...

Computational fluid dynamics numerical issues



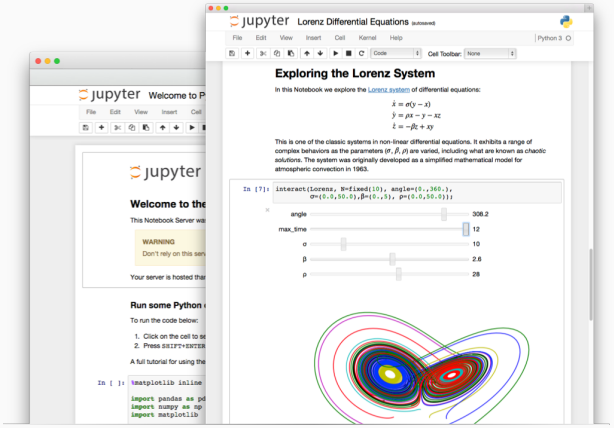
Distinguish between:

- experimental science
- observational science
- computational science (simulation)
- (big) data analysis

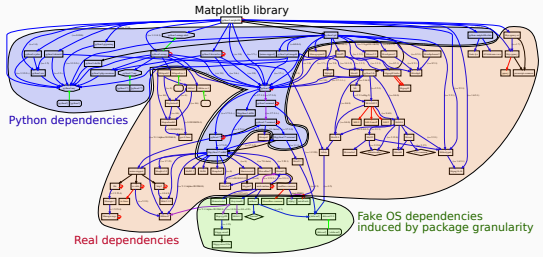
Present

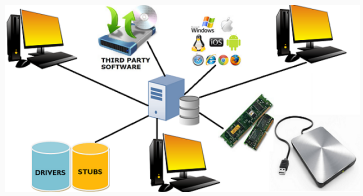
Existing Tools, Emerging Standards

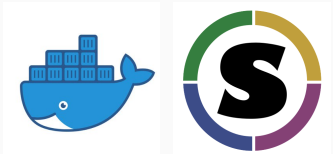
Notebooks and workflows



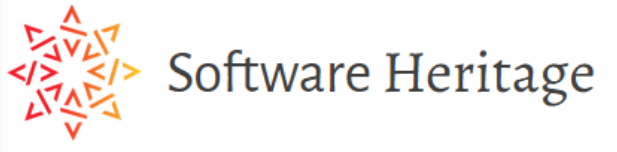
Software environments







Sharing platforms



Changing Practices

Manifesto: *"I solemnly pledge"* (WSSSPE, Lorena Barba, FAIR)

1. I will teach my graduate students about reproducibility
2. All our research code (and writing) is under version control
3. We will always carry out verification and validation
4. We will share data, plotting script & figure under CC-BY
5. We will upload the preprint to arXiv at the time of submission of a paper
6. We will release code at the time of submission of a paper
7. We will add a "Reproducibility" declaration at the end of each paper
8. I will keep an up-to-date web presence

Software Engineering and Reproducible Research in the curricula

- Webinars on RR 2016-2017
- **MOOC on RR** (3rd edition planned for January 2020)
- Book on RR in June 2019

Publishing

Artifact evaluation and ACM badges

Major conferences

- Supercomputing: Artifact Description (AD) mandatory, Artifact Evaluation (AE) still optional, Double blind vs. RR
- NeurIPS, ICLR: open reviews, reproducibility challenge

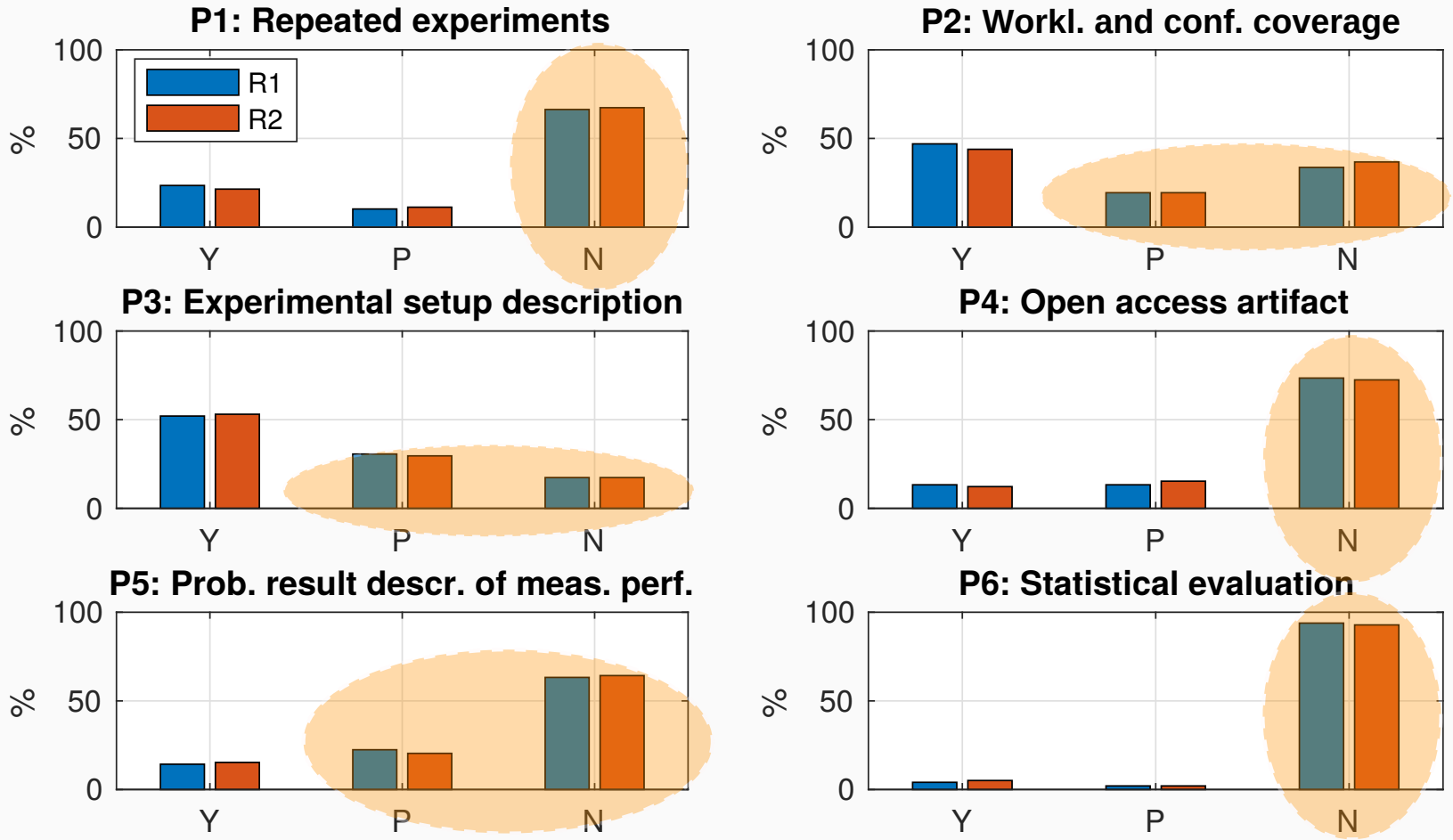
Joelle Pineau @ NeurIPS'18

- ACM SIGMOD 2015-2019, ...

Mentality is evolving: people care and make stuff available

Key Concerns for our Community (Room for Improvement)

- **Awareness of Experiments and Statistics** How are cloud performance currently obtained and reported?, March 2019

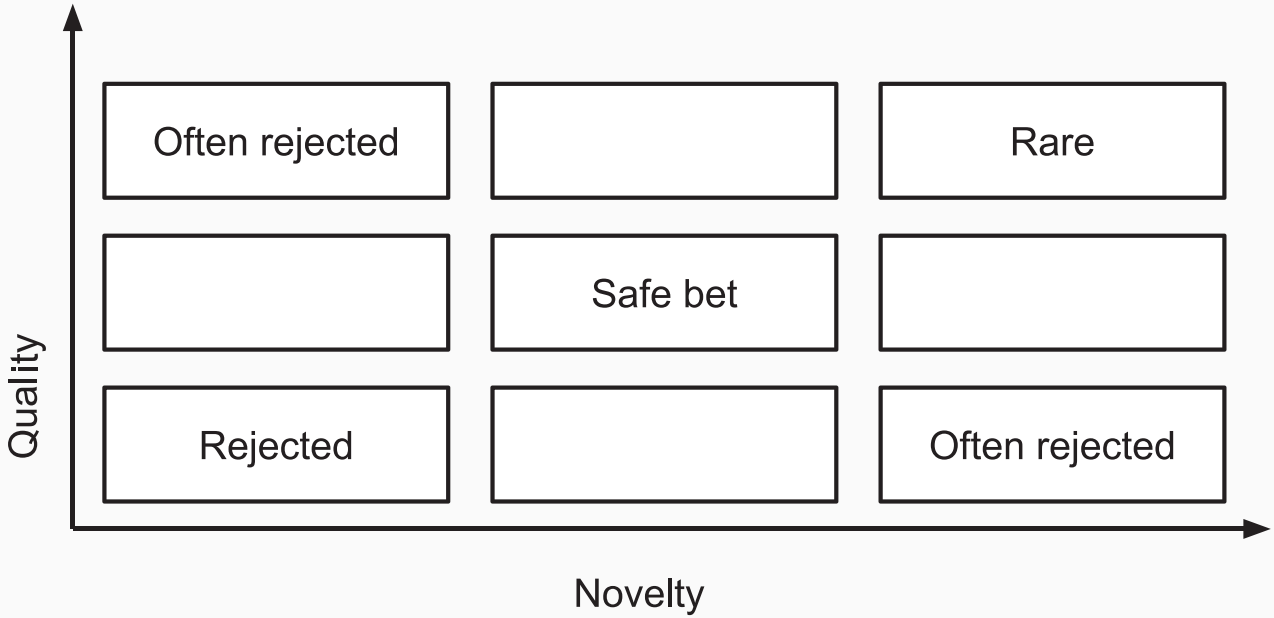


- **Shared testbeds and Experimental control** TILECS

Future

Publish or Perish (ok, this is past and present)

- Goodhart's Law: Are Academic Metrics Being Gamed?, M. Fire 2019
  - AI: over 1,000 ranked journals (×10 in 15 years)
  - Shorter papers with increasing self references
  - More and more papers without any citation
  - Sharp increase in the number of new authors publishing at a much faster rate given their career age
- The Truth, The Whole Truth, and Nothing But the Truth: A Pragmatic Guide to Assessing Empirical Evaluations, *TOPLAS* 2016



Reproducible Research = Transparency

To err is human.

Good research requires time and resources

1. **Train yourself and your students on RR**, statistics, experiments
   - Beware of checklists and norms
   - Understand what's at stake
2. **Make publication practices evolve**
3. **Prepare the Future: Toward literate experimentation?**
   - Reuse, reuse, reuse
   - How to share Experiments