Data Infrastructure, Cyberinfrastructure and Reproducibility

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

Building Bridges Across the S&T Enterprise
A White House National Science and Technology Council (NSTC) Conference
NIH Natcher Conference Center, Bethesda, MD
June 13, 2019

Agenda

1. Highlights from the National Academies Report on Reproducibility and Replication (April 2019)

2. Reproducibility Definitions and Data Infrastructure

3. Moving to Panel Discussion

The NASEM Reproducibility & Replication Report (April 2019)

- The 2017 "**American Innovation and Competitiveness Act**" contained a section called '*Research Reproducibility and Replication*'

- This section allocated funding for a report to be submitted to the Director of the National Science Foundation

- The report was to assess "*research and data reproducibility and replicability issues in interdisciplinary research*" and make "*recommendations for improving rigor and transparency in scientific research*."

Definitions

The terms “reproducibility” and “replicability” have different meanings and uses across science and engineering. For our report:

- **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.*

- **Generalizability** refers to the extent that results of a study apply in other contexts or populations that differ from the original one.

- In short, reproducibility involves the original data and code; replicability involves new data collection to test for consistency with previous results of a similar study.

These two processes also differ in the type of results that should be expected. In general, when a researcher transparently reports a study and makes available the underlying digital artifacts, such as data and code, the results should be computationally reproducible. In contrast, even when a study was rigorously conducted according to best practices, correctly analyzed, and transparently reported, it may fail to be replicated.

Key Recommendation 4-1

RECOMMENDATION 4-1: 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).

Key Recommendation 6-3

RECOMMENDATION 6-3: 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.

Key Recommendation 6-5

RECOMMENDATION 6-5: 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.

Through these repository criteria, NSF would **enable discoverability and standards for digital scholarly objects** and discourage an undue proliferation of repositories, perhaps through endorsing or providing one go-to website that could access NSF-approved repositories.

Key Recommendation 6-6

RECOMMENDATION 6-6: **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**.

Key Recommendation 6-9

RECOMMENDATION 6-9: 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.

Kickstarting a “Reproducibility Industry” by Grant Set-asides

- Previously, NIH required that clinical trials hire Biostatistician PhD's to design and analyze experiments. This set-aside requirement transformed clinical trials practice and resulted in better science, and spawned the field of Biostatistics by creating a demand for a specific set of services and trained people to perform them.

- Try a similar idea for reproducibility? Set asides for funded research to support reproducibility -> infrastructure development.

- Reproducibility startups: CodeOcean.com & Flywheel.io

Recap of Discussion

- Joanna:
- Joni:
  - Data science & open science promise to accelerate innovation & progress in biomedical research
  - NIH has extensive plans to realize this promise, relying on FAIR principles
  - Implementation of these plans is underway
- Antony:
  - FAIR and Open Data is critical to building scientific data hubs for the community
  - Transparency – in data and predictive models is the new approach to science and should be embraced
  - Data QUALITY is key and community collaboration and crowdsourcing is critical to success
  - Interoperability is enabled by the adoption of open standards – especially ontologies and taxonomies
- Victoria
  - NASEM R&R Report Recommendation clarify the role of Reproducibility and Replication in the Open Science and Data Ecosystem.

Discussion Questions

- How closely do **agencies work together** to ensure interoperability of data systems / information? How closely should they? What are challenges? Models of successful practices? Opportunities?

- What are challenges with identifying **common infrastructure models and standards**, including artifact standards?

- What are your specific challenges faced with making data and code **transparent and accessible**?

- How have increased concerns for **privacy, security, and intellectual property** impacted ability to make information publicly available? What are strategies to assess risks for any given dataset / information? How can data and software infrastructure evolve to address concerns?

- How has movement to **cloud-based platforms** enabled progress and what challenges need to be addressed?

- What are agencies doing, specifically, to support the **FAIR** (Findability, Accessibility, Interoperability, and Reusability) principles in their public access implementations?

Technological Sources of Impact

1. Big Data / Data Driven Discovery: high dimensional data, p >> n,

2. Computational Power: simulation of the complete evolution of a physical system, systematically varying parameters,

3. Deep intellectual contributions now encoded only in software.

Claim: *Virtually all published discoveries today have a computational component. (Isn't Data Science all science?)*

Corollary: *There is a mismatch between traditional scientific dissemination practices and modern computational research processes, leading to reproducibility concerns.*

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

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

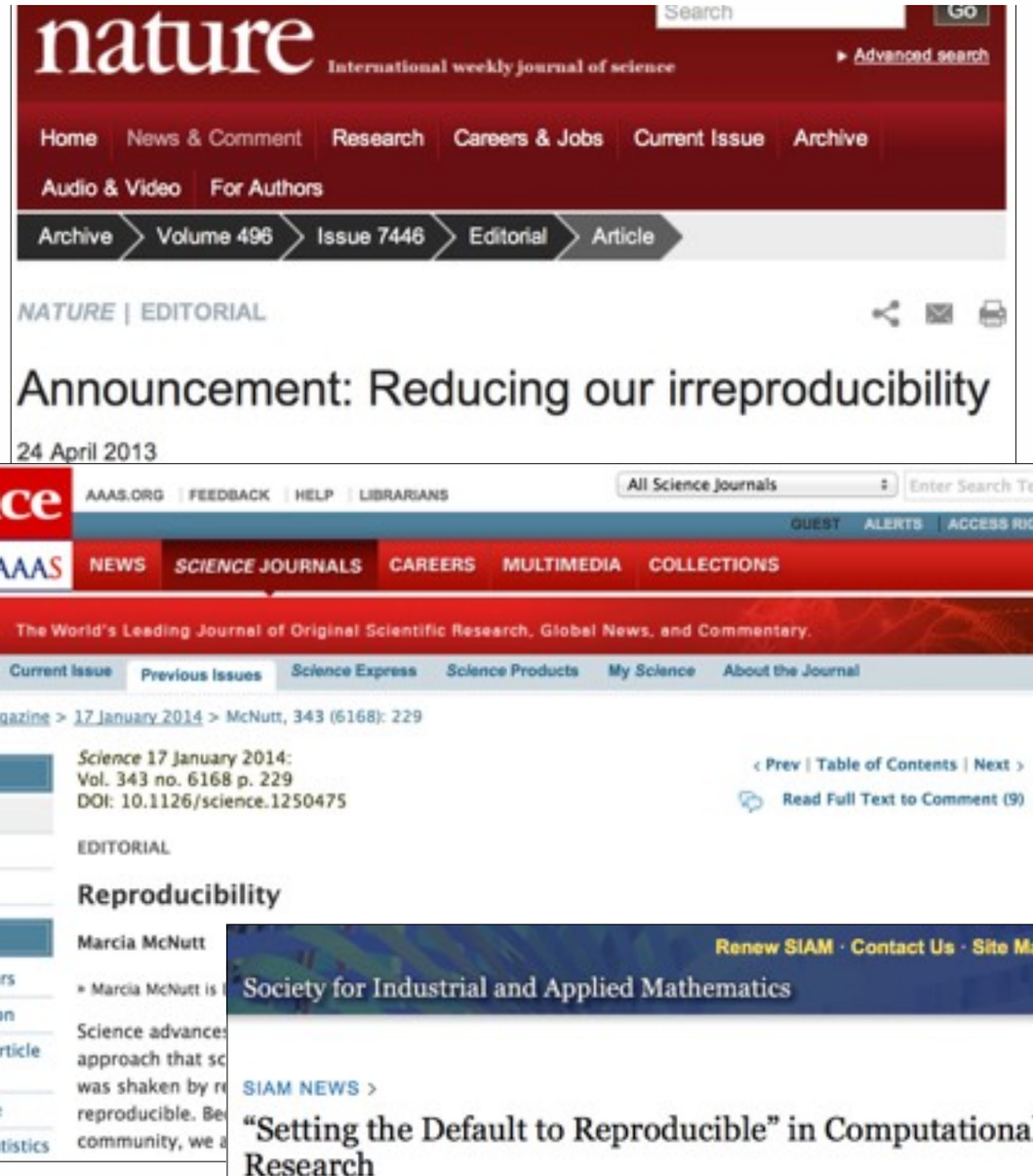
Parsing Reproducibility

"Empirical Reproducibility"

"Statistical Reproducibility"

"Computational Reproducibility"

V. Stodden, IMS Bulletin (2013)



Workshop Recommendations: "Reproducibility Enhancement Principles"

1. Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.

2. 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.

Workshop Recommendations: "Reproducibility Enhancement Principles"

5. Use Open Licensing when publishing digital scholarly objects.

6. Journals should conduct a reproducibility check as part of the publication process and should enact the TOP standards at level 2 or 3.

7. To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.

Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) e.g the U.S. Constitution and subsequent Acts:

"*To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries*." (U.S. Const. art. I, §8, cl. 8)

Copyright

- Original expression of ideas falls under copyright by default (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original
- limited time: generally life of the author +70 years
- Exceptions and Limitations: e.g. Fair Use.

Patents

Patentable subject matter: "*new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof*" (35 U.S.C. §101) that is

1. *Novel*, in at least one aspect,

2. *Non-obvious*,

3. *Useful*.

USPTO Final Computer Related Examination Guidelines (1996) "A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena" (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.

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?

The Reproducible Research Standard

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

A suite of license recommendations for computational science:

- Release media components (text, figures) under **CC BY**,
- Release code components under **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

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

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

A Convergence of Trends

- Scientific projects will become massively more computing intensive, and

- Scientific computing will become dramatically more transparent

Simultaneity: better transparency allows much more ambitious computational experiments. *And* better computational experiment infrastructure allows greater transparency.

Such a system is used not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work, enabling *efficiency* and *productivity*, and *discovery*.

“Quantitative Programming Environments”

- Define and create “Quantitative Programming Environments” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis

- The two trends need to be addressed simultaneously: better transparency will allow people to run much more ambitious computational experiments. *And* better computational experiment infrastructure will allow researchers to be more transparent.

https://www.acm.org/publications/policies/artifact-review-badging

Terminology.

A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms. The following are inspired by the International Vocabulary for Metrology(VIM); see the Appendix for details.

- Repeatability (Same team, same experimental setup)
    - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Replicability (Different team, same experimental setup)
    - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Reproducibility (Different team, different experimental setup)
    - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Privacy and Data

- (U.S.) HIPAA, FERPA, Institutional Review Boards create legally binding restrictions on the sharing human subjects data (see e.g. http://www.dataprivacybook.org/)

- Potential privacy implications for industry generated data.

- Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..

Ownership: What Defines Contribution?

- Issue for producers: credit and citation.
- What is the role of peer-review?
- Repositories adding meta-data and discoverability make a contribution.
- Data repositories may be inadequate: velocity of contributions
- Future coders may contribute in part to new software, other software components may already be in the scholarly record. Attribution vs sharealike.

  ➡ (at least) 2 aspects: legal ownership vs scholarly credit.

- Redefining plagiarism for software contributions.

Licensing in Research
Background: Open Source Software

Innovation: Open Licensing

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

Hundreds of open source software licenses:

- GNU Public License (GPL)
- (Modified) BSD License
- MIT License
- Apache 2.0 License
- ... see http://www.opensource.org/licenses/alphabetical

The Reproducible Research Standard

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

A suite of license recommendations for computational science:

- Release media components (text, figures) under **CC BY**,
- Release code components under **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

➡ Remove copyright's barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.

Computational Barriers

Barriers to Replication in Computational Science:

- rerunning same code, same parameter settings, same system can produce different results (?),
- same code (Reprozip, containerization/Docker), but updated libraries, compiler, operating system..
- software customization to underlying architectures; portability, modularity, re-usability,
- numerical stability of the underlying software architecture,
- unique hardware, scarce allocations, long runtimes..

Encouraging Reproducibility While Expanding Access to Massive Computation

We are at the convergence of two (ordinarily antagonistic) trends:

1. Scientific projects will become massively more computing intensive,

2. Scientific computing dramatically more transparent.

These two trends can reinforce each other: better transparency will allow people to run much more ambitious computational experiments. *And* better computational experiment infrastructure will allow researchers to be more transparent.