Advancing Data-enabled Research via Reproducibility and Transparency

Victoria Stodden
<br><vcs@stodden.net>

Cognitive Systems Institute Group Speaker Series

September 26, 2019
Who am I?

1. PhD in Statistics; Law Degree (Stanford University)

2. IBM T. J. Watson Research Center (Summer 2000)

3. Currently Associate Professor in the iSchool, University of Illinois at Urbana-Champaign, faculty affiliations in CS, Stats, Law School, NCSA, and CSL.

4. Research on Reproducibility: Computational Transparency; Statistical Generalization; Policy.
Agenda

1. Three Types of Reproducibility


3. Proposal: A Computable Scholarly Record
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.

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

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”

Traditionally two branches to the scientific method:

- Branch 1 (deductive): mathematics, formal logic.
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

- Branch 3,4? (computational): large scale simulations / data driven computational science.
```
Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

Gaithersburg, MD
September 21, 1995

Workshop Overview

The workshop consisted of an introduction; five talks, each followed by a discussion period; and an open discussion session. Capsule versions follow immediately; more substantial summaries follow later.

Jim Blue opened the workshop with brief introductory remarks. He emphasized that the purpose of doing modeling and simulation is to gain understanding and insight. The three benefits are that modeling and simulation can be cheaper, quicker, and better than experimentation alone. It is common now to consider computation as a third branch of science, besides theory and experiment.

“It is common now to consider computation as a third branch of science, besides theory and experiment.”
```

```
“This book is about a new, fourth paradigm for science based on data-intensive computing.”
```
The Ubiquity of Error

The central motivation for the scientific method is to root out error:

• Deductive branch: the well-defined concept of the proof,

• Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation and Data Science present only potential third/fourth branches of the scientific method (Donoho et al. 2009), until the development of comparable standards.
Really Reproducible Research

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

“The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.” David Donoho, 1998

Note: reproducing computational steps vs re-implementing the experiment independently (both needed).
“Reproducibility and Replication in Science”
Consensus Report, May 2019

National Academies of Science, Engineering, and Medicine
Definitions

The terms, “reproducibility” and “replicability” have different meanings and uses across science and engineering, which has led to confusion in collectively understanding problems in reproducibility and replicability. The committee adopted specific definitions for the purpose of this report to clearly differentiate between the terms, which are otherwise interchangeable in everyday discourse.

**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,” and the terms are used interchangeably in this report.

**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: 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 6-3: 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 3

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 4

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.
Proposal: A Computable Scholarly Record

Exposure of computational steps? A dream:

- potential executability
- code re-use
- methods application in new contexts
- pooling data and improved experimental power
- improved validation of findings
- organization of discovery pipeline information
- structured dissemination of reproducible findings

A More Modest Proposal: The Knowledge Integrator

- Development of dissemination standards around results (stack agnostic)

- Central deposition of computationally reproducible results; open access, open deposit to grow the computable scholarly record

- Integration of results to extend knowledge

- Identification of computational components: results, methods and method types, scientific questions, experiment types

- The scholarly record as a data source for calculations such as: overall false discovery rate; key questions in different fields (and current understanding); meta-science and assessment; benchmarking and algorithm performance..
Conclusions

Two (ordinarily antagonistic) trends are converging:

1. Scientific projects will become massively more compute intensive

2. Research computing will become dramatically more transparent

• These are in fact reinforcing trends, whose convergence admits a computable scholarly record, for example the Knowledge Integrator; essential for verifying and comparing findings.