Advancing Computational Scientific Discovery by Enabling Reproducibility and Transparency: Policies and Practice

Victoria Stodden

University of Southern California

http://stodden.net

SIAM Conference on Computational Science and Engineering (CSE21)

March 1 - 5, 2021

This material is based upon work supported by the National Science Foundation under Grant Nos 2028881 and 21941443

Agenda

1. *Really* Reproducible Research

2. Key AAAS / Arnold Foundation Workshop Recommendations

3. Key National Academies *Reproducibility and Replication in Science* Recommendations

Really Reproducible Research

"Really Reproducible Research" introduced by Jon Claerbout (1992), summarized by David Donoho (1998):

"**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 software development environment and the complete set of instructions [and data] which generated the figures.**"

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 proposals from the Transparency and Openness Promotion (TOP) guidelines (*1*) and recommendations for field data (*2*), emerged from workshop discussions among funding

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 independent regeneration of computational results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using

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 objects, citation should be standard practice. All data, code, and workflows, including software written by the authors, should be cited in the references section (*10*). We suggest that

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.

“Reproducibility and Replication in Science” Consensus Report, April 2019

National Academies of Science, Engineering, and Medicine

Definitions: “reproducibility” and “replicability”

- **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.

Key Recommendation 1

RECOMMENDATION 4-1: ...*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 2

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, (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 … 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.

Key Recommendation 5

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.

Conclusions

We see the convergence of two (ordinarily antagonistic) trends:

1. Scientific projects will become massively more computing intensive

2. Research computing will become dramatically more transparent

These are reinforcing trends, whose resolution is essential for verifying and comparing findings.

Join our next Virtual World Cafés

- Developers and testers working on hardware and software systems
  - Wed Apr 21, 2021 noon-3pm EST / 9am-noon PST
  - Register at: **https://qrgo.page.link/ZpiLV**

Café 2

- Trainers and educators working with workforce development
  - Wed Aug , 2021 noon-3pm EST / 9am-noon PST
  - Register at: **https://qrgo.page.link/xbrRu**

Café 3

Visit the project website at: **robustscience.org**