Transparency and Reproducibility: Case Studies, Formalisms, and Structured Guidance in Computational Social Science Applications

Victoria Stodden
University of Illinois at Urbana-Champaign

SIAM Conference on Parallel Processing for Scientific Computing (PP20)
Session: "Transparency, Reproducibility, Sustainability, and Security: The Four Pillars of the Next Generation Scientific Software Stack"

Seattle, WA
February 13, 2020

Agenda

1. **Setting the Stage: Research Reproducibility**
   - National Academies of Science, Engineering, and Medicine report

2. **A Tour of Three Examples**
   - Container-based Reproducible Data Science with the Whole Tale project
   - The “Time/Value Tradeoff” for Reproducibility: Execution in the Long Run
   - Reproducibility Journal Policy: Who Re-executes the Research? Where?

3. **A “Lifecycle of Data Science” Approach Includes Security**

This material is based upon work supported by the National Science Foundation under Grant No. 1941443. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

1. Setting the Stage: Research Reproducibility

Reproducibility Definitions: National Academies

In 2019 the "Reproducibility and Replication in Science" committee published consensus report (I was a committee member).

Produced key definitions and several recommendations.

- *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.

Some Reproducibility Efforts

FROM THE EDITOR'S DESK

On Reproducible Research

By Bram Vanderborght

During the IEEE Panel of Editors meeting held this past April in Los Angeles, California, I was invited, as editor-in-chief of *IEEE Robotics and Automation Magazine*, to participate in a panel discussing reproducible research under the lead of the leading role of our associate editor, Fabio Bonsignorio and colleagues, who began work on this topic ten years ago and has since organized several related workshops. Remember, we had a special issue in

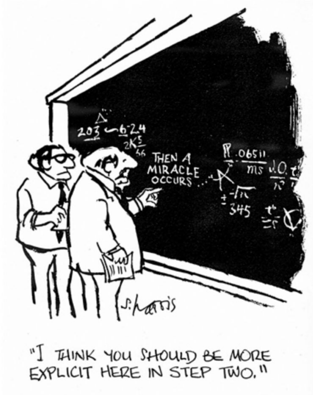
Again, I offer a warm call to submit articles with reproducible research. Together with the authors, we need to improve the procedure over time, but we can only learn from experience.

With full awareness that it requires time and effort that may be in opposition to current publish-or-perish pressure, we need to increase efforts to reward the positive behavior of authors contributing to reproducible research. One initiative suggests highlighting paper with reproducible research content, perhaps even granting awards dedicated to such papers. Discussions are in progress with the Association for

"Setting the Default to Reproducible" in Computational Science Research

By Victoria Stodden, Jonathan M. Borwein and David H. Bailey

Following a late-2012 workshop at the Institute for Computational and Experimental Research in Mathematics, a group of computational scientists have proposed a set of standards for the dissemination of reproducible research.



Courtesy of S. Harris, ScienceCartoonsPlus.com.

Computation is now central to the scientific enterprise, and the emergence of powerful computational hardware, combined with a vast array of computational software, presents novel opportunities for researchers. Unfortunately, the scientific culture surrounding computational work has evolved in ways that make it difficult to verify findings, efficiently build on past research, or even apply the basic tenets of the scientific method to computational procedures.

As a result, computational science is facing a credibility crisis [1,2,4,5]. The enormous scale of state-of-the-art scientific computations, using tens or hundreds of thousands of processors, presents unprecedented challenges. Numerical reproducibility is a major issue, as is hardware reliability. For some applications, even rare interactions of circuitry with stray subatomic particles matter.

In December 2012, more than 70 computational scientists and stakeholders, such as journal editors and funding agency officials, gathered at Brown University for the ICERM Workshop on Reproducibility in Computational and Experimental Mathematics. This workshop gave a broad cross section of computational scientists their first opportunity to discuss these issues and brainstorm ways to improve on current practices; the result was a series of recommendations for establishing really reproducible computational science as a standard [13]. Three main recommendations emerged from the workshop discussions:

New NISO Project: Badging Scheme for Reproducibility in the Computational and Computing Sciences

January 2019

Call for Participation

NISO voting members have approved a new project, Recommended Practice: Toward a Compatible Taxonomy, Definitions, and Recognition Badging Scheme for Reproducibility in the Computational and Computing Sciences. As publishers and researchers are placing greater emphasis on the practice of reproducibility as an essential ingredient of the scientific research process, it is critical to make compatible the taxonomies used to define the various levels of reproducibility

SPOTLIGHT ON TRANSACTIONS

EDITOR RON VETTER
University of North Carolina Wilmington; vetter@uncw.edu

The Reproducibility Initiative

Manish Parashar, Rutgers University

This installment of Computer's series highlighting the work published in IEEE Computer Society journals comes from IEEE Transactions on Parallel and Distributed Systems.

Reproducibility is a foundation of solid scientific and technical research. The ability to repeat research is key to confirming the validity of a

to upload the code to Code Ocean, which generates a "compute capsule" that includes the code, data, results, and computational environment specifications. Code Ocean sends the EiC a review copy of the compute capsule, which is passed on to the assigned reproducibility associate editor for the article.

SIAM News 2013

- Editorial Policies and Badging
- Pilot Partnerships: Code Ocean

2. A Tour of Three Examples

1. Data Science in the Whole Tale Project

- Building an **open platform for computational reproducibility**
  - Create and publish **executable research objects** ("*Tales*")
- Simplify process of creating & verifying reproducible computational artifacts for scientific discovery

Easy-to-access **cloud-based** computing **environments**

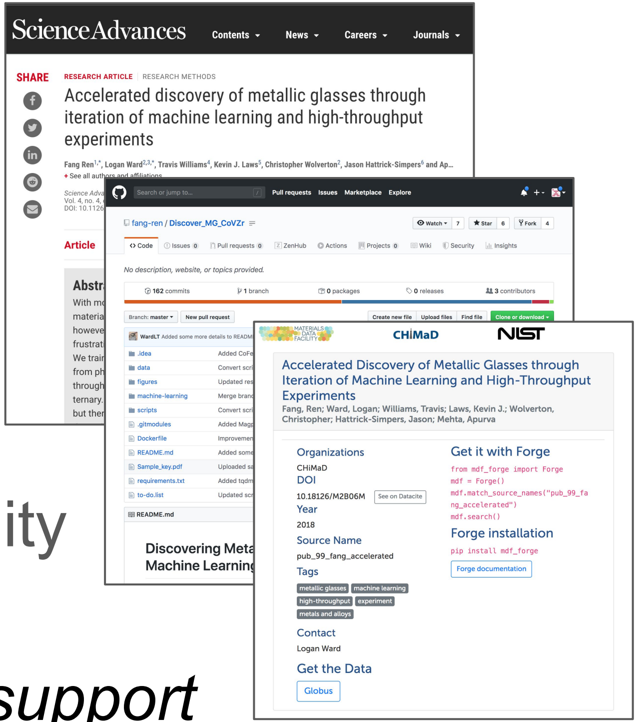
Transparent access to research **data**

Export and **publish executable research objects**

This material is based upon work supported by the National Science Foundation under Grant No. OAC-1541450

Use case: Ren et al. (2018)

- ML experiments in materials science
- Published in *Science Advances*
- Code in Github
- Data published to Materials Data Facility



How can we publish the code and data to support computational reproducibility and reuse/exploration?

- Reproducibility implemented in Whole Tale

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

Tale Packaging for Sharing, Dissemination, Archiving

- Research Object
  - Beyond PDFs and datasets -- include code, workflows
  - Distributed elements
- Interoperability between systems
  - Archives/repositories
  - Active compute platforms
- BagIt serialized "Research Object" bundle
  - Zip archive + metadata + JSON-LD
  - https://github.com/ResearchObject/bagit-ro (=> ro-crate)

researchobject.org

Chard et al. (2019) Application of BagIt-Serialized Research Object Bundles for Packaging and Re-execution of Computational Analyses. RO-5 at Workshop on Research Objects (RO 2019)

2. Reproducibility Journal Policy: Who Re-executes the Research? Where?

2. Reproducibility Standards Development

Reproducibility requires community adoption and standards development.
Example: 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 r**eproducibility check** as part of the publication process.

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

➔ NASEM 2019 "Reproducibility and Replication in Science" report recommendations.

Transparency and Openness Promotion (TOP) and Open Problems

- Responsibility for verification; 3rd party re-execution of codes?
- JASA-ACS Reproducibility Editors? Cloud infrastructure (Whole Tale?)? Automation?
- Documentation and meta-data for data and code: transparency and liability

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

1424 26 JUNE 2015 • VOL 348 ISSUE 6242 sciencemag.org SCIENCE

3. A “Lifecycle of Data Science” Includes Security

The Lifecycle of Data Science

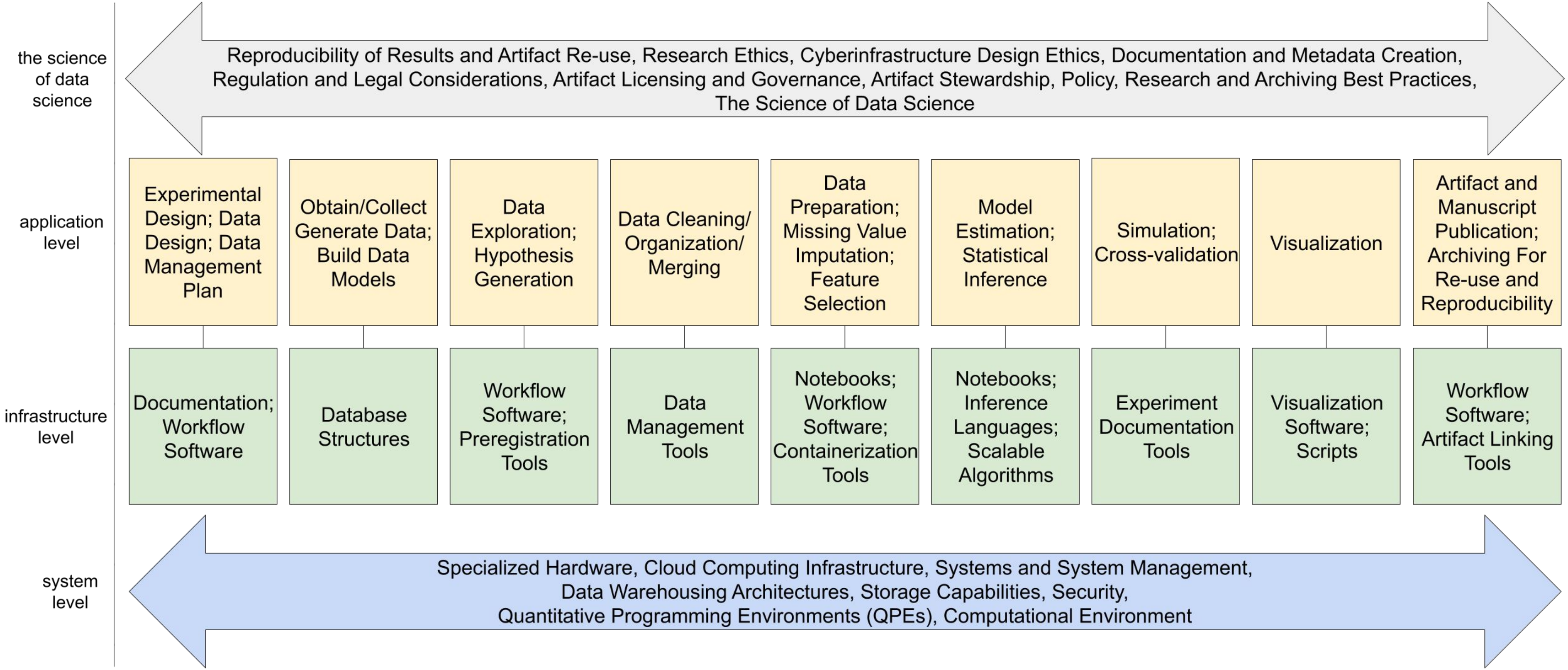
"Lifecycle of Data" is an abstraction from the Information Sciences

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

What if we applied this idea to Data Science?

- **Clarify steps** in data science projects: people/skills involved, tools and infrastructure, and reproducibility through the cycle.
- **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



Leveraging the Lifecycle of Data Science

An abstraction that organizes the computational pipeline.. and so recognizes different contributions including from e.g.:

- Ethicists
- Knowledge and data managers
- Compute resources and cyberinfrastructure

Goals:

- Improve understanding of Data Science advancement.
- Permit the comparison of results.
- Improve research output and social impact.

V. Stodden (2020). The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science. forthcoming Communications of the ACM.

Conclusion

Two (ordinarily antagonistic) trends are converging:

Research will become **massively more compute and data intensive**, and
Research computing will become **dramatically more transparent**.

These are reinforcing trends, which can admit exciting new opportunities:

- greater understanding of norms and social structures for discovery,
- enabling **efficiency**, **productivity**, and **discovery**.

Security issues pervasive and of ongoing importance with cyberinfrastructure development.