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)
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

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

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.

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

---

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

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](https://github.com/ResearchObject/bagit-ro) ( => ro-crate)
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 **reproducibility check** as part of the publication process.

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

<table>
<thead>
<tr>
<th>LEVEL 0</th>
<th>LEVEL 1</th>
<th>LEVEL 2</th>
<th>LEVEL 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Citation standards</strong></td>
<td>Journal encourages citation of data, code, and materials—or says nothing.</td>
<td>Journal describes citation of data in guidelines to authors with clear rules and examples.</td>
<td>Article provides appropriate citation for data and materials used, consistent with journal’s author guidelines.</td>
</tr>
<tr>
<td><strong>Data transparency</strong></td>
<td>Journal encourages data sharing—or says nothing.</td>
<td>Article states whether data are available and, if so, where to access them.</td>
<td>Data must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
</tr>
<tr>
<td><strong>Analytic methods (code) transparency</strong></td>
<td>Journal encourages code sharing—or says nothing.</td>
<td>Article states whether code is available and, if so, where to access them.</td>
<td>Code must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
</tr>
<tr>
<td><strong>Research materials transparency</strong></td>
<td>Journal encourages materials sharing—or says nothing.</td>
<td>Article states whether materials are available and, if so, where to access them.</td>
<td>Materials must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
</tr>
<tr>
<td><strong>Design and analysis transparency</strong></td>
<td>Journal encourages design and analysis transparency or says nothing.</td>
<td>Journal articulates design transparency standards.</td>
<td>Journal requires adherence to design transparency standards for review and publication.</td>
</tr>
<tr>
<td><strong>Preregistration of studies</strong></td>
<td>Journal says nothing.</td>
<td>Journal encourages preregistration of studies and provides link in article to preregistration if it exists.</td>
<td>Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.</td>
</tr>
<tr>
<td><strong>Preregistration of analysis plans</strong></td>
<td>Journal says nothing.</td>
<td>Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.</td>
<td>Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.</td>
</tr>
<tr>
<td><strong>Replication</strong></td>
<td>Journal discourages submission of replication studies—or says nothing.</td>
<td>Journal encourages submission of replication studies.</td>
<td>Journal encourages submission of replication studies and conducts blind review of results.</td>
</tr>
</tbody>
</table>
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

Reproducibility of Results and Artifact Re-use, Research Ethics, Cyberinfrastructure Design Ethics, Documentation and Metadata Creation, Regulation and Legal Considerations, Artifact Licensing and Governance, Artifact Stewardship, Policy, Research and Archiving Best Practices, The Science of Data Science

Experimental Design; Data Design; Data Management Plan
Obtain/Collect Generate Data; Build Data Models
Data Exploration; Hypothesis Generation
Data Cleaning/ Organization/ Merging
Data Preparation; Missing Value Imputation; Feature Selection
Model Estimation; Statistical Inference
Simulation; Cross-validation
Visualization
Artifact and Manuscript Publication; Archiving For Re-use and Reproducibility

Documentation; Workflow Software
Database Structures
Workflow Software; Preregistration Tools
Data Management Tools
Notebooks; Workflow Software; Containerization Tools
Notebooks; Inference Languages; Scalable Algorithms
Experiment Documentation Tools
Visualization Software; Scripts
Workflow Software; Artifact Linking Tools

Specialized Hardware, Cloud Computing Infrastructure, Systems and System Management, Data Warehousing Architectures, Storage Capabilities, Security, Quantitative Programming Environments (QPEs), Computational Environment
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.