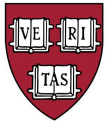


Computational Reproducibility: Expectations, Challenges and Recommendations

JEDI 2022 Workshop – *May the force be with you: Resources to help journal editors advance their fields*
May 20, 2022

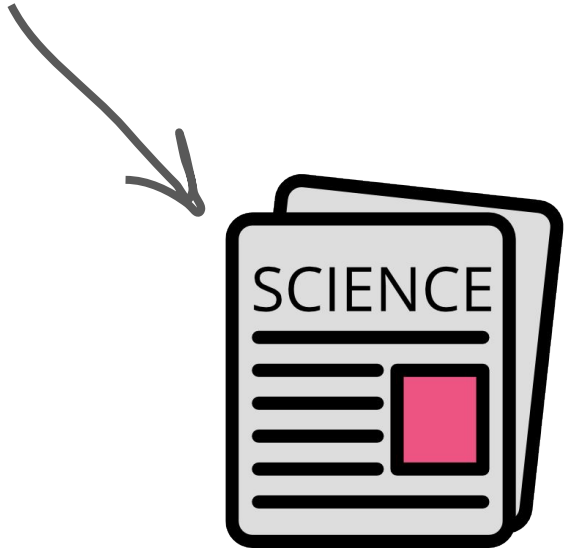
Ana Trisovic, Harvard Biostatistics
Review board member of the Journal of Systems Research (JSys)
Early-career board member of Harvard Data Science Review (HDSR)

“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 software, [data] … and set of instructions which generated the figures. ”

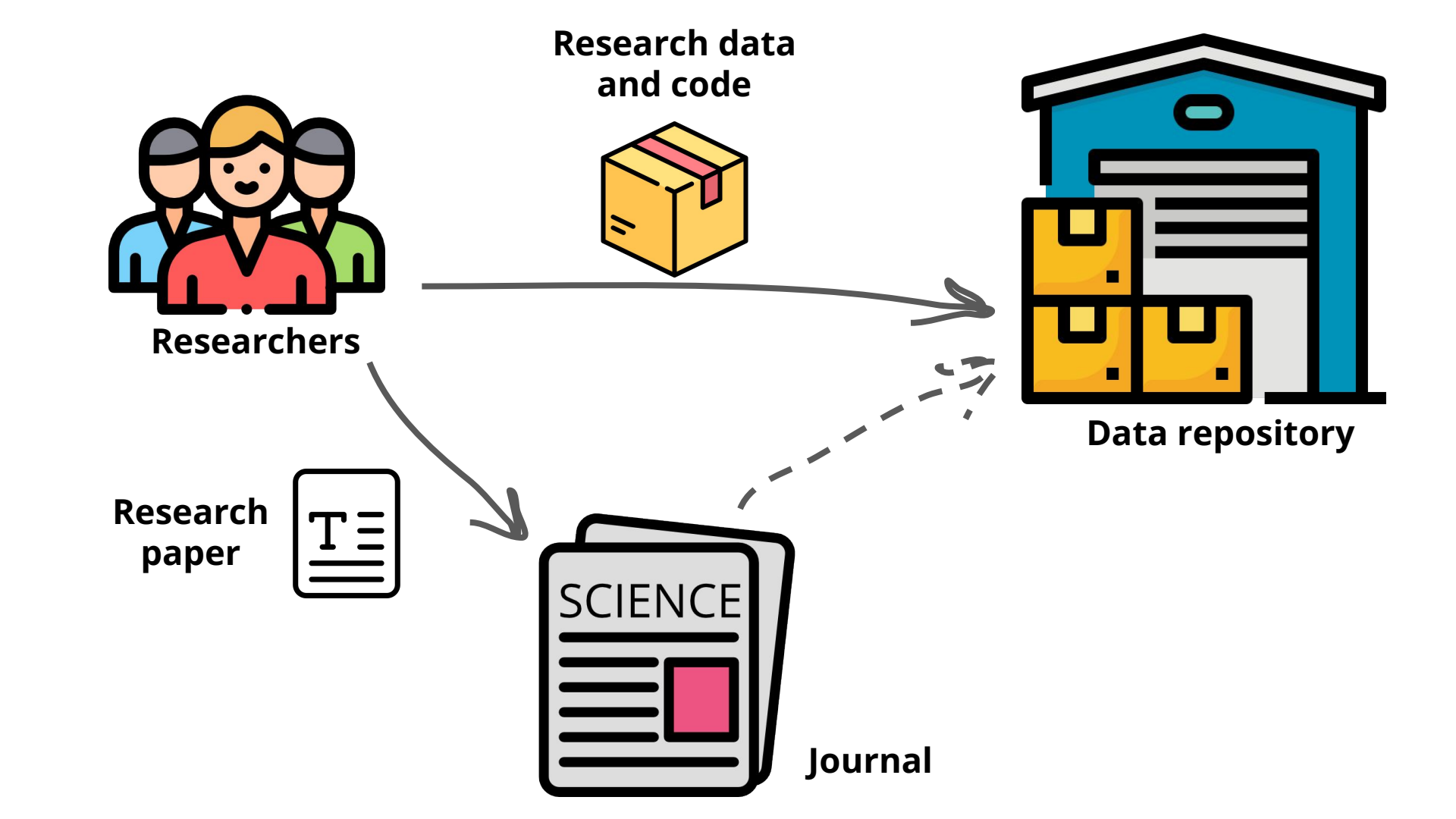
~ Prof Claerbout

Buckheit. 1995. Wavelab and reproducible research.

Researchers



Journal


Research data
and code
Researchers
Data repository
Research
paper
SCIENCE
Journal

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

ANALYSIS

A large-scale study on research code quality and execution

Ana Trisovic[1]✉, Matthew K. Lau[2], Thomas Pasquier[3] & Mercè Crosas[1]

This article presents a study on the quality and execution of research code from publicly-available replication datasets at the Harvard Dataverse repository. Research code is typically created by a group of scientists and published together with academic papers to facilitate research transparency and reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility and reuse. First, we retrieve and analyze more than 2000 replication datasets with over 9000 unique R files published from 2010 to 2020. Second, we execute the code in a clean runtime environment to assess its ease of reuse. Common coding errors were identified, and some of them were solved with automatic code cleaning to aid code execution. We find that 74% of R files failed to complete without error in the initial execution, while 56% failed when code cleaning was applied, showing that many errors can be prevented with good coding practices. We also analyze the replication datasets from journals' collections and discuss the impact of the journal policy strictness on the code re-execution rate. Finally, based on our results, we propose a set of recommendations for code dissemination aimed at researchers, journals, and repositories.

Introduction

Researchers increasingly publish their data and code to enable scientific transparency, reproducibility, reuse, or compliance with funding bodies, journals, and academic institutions[1]. Reusing data and code should propel new research and save researchers' time, but in practice, it is often easier to write new code than reuse old. Even attempting to reproduce previously published results using the same input data, computational steps, methods, and code has shown to be troublesome. Studies have reported a lack of research reproducibility[2,3] often caused by inadequate documentation, errors in the code, or missing files.

- Dataverse = a free and open-source software platform to archive, share, and cite research data
    - Focus on making research data and code available
- Retrieved 2109 publicly available replication datasets containing 9078 R files
- Over 94% of the datasets belonged to social sciences

The Dataverse® Project

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

ANALYSIS

A large-scale study on research code quality and execution

Ana Trisovic[1✉], Matthew K. Lau[2], Thomas Pasquier[3] & Mercè Crosas[1]

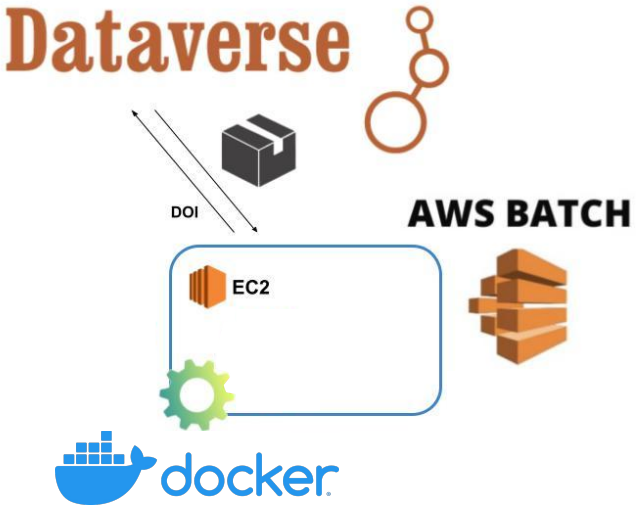
This article presents a study on the quality and execution of research code from publicly-available replication datasets at the Harvard Dataverse repository. Research code is typically created by a group of scientists and published together with academic papers to facilitate research transparency and reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility and reuse. First, we retrieve and analyze more than 2000 replication datasets with over 9000 unique R files published from 2010 to 2020. Second, we execute the code in a clean runtime environment to assess its ease of reuse. Common coding errors were identified, and some of them were solved with automatic code cleaning to aid code execution. We find that 74% of R files failed to complete without error in the initial execution, while 56% failed when code cleaning was applied, showing that many errors can be prevented with good coding practices. We also analyze the replication datasets from journals' collections and discuss the impact of the journal policy strictness on the code re-execution rate. Finally, based on our results, we propose a set of recommendations for code dissemination aimed at researchers, journals, and repositories.

Introduction

Researchers increasingly publish their data and code to enable scientific transparency, reproducibility, reuse, or compliance with funding bodies, journals, and academic institutions[1]. Reusing data and code should propel new research and save researchers' time, but in practice, it is often easier to write new code than reuse old. Even attempting to reproduce previously published results using the same input data, computational steps, methods, and code has shown to be troublesome. Studies have reported a lack of research reproducibility[2,3] often caused by inadequate documentation, errors in the code, or missing files.

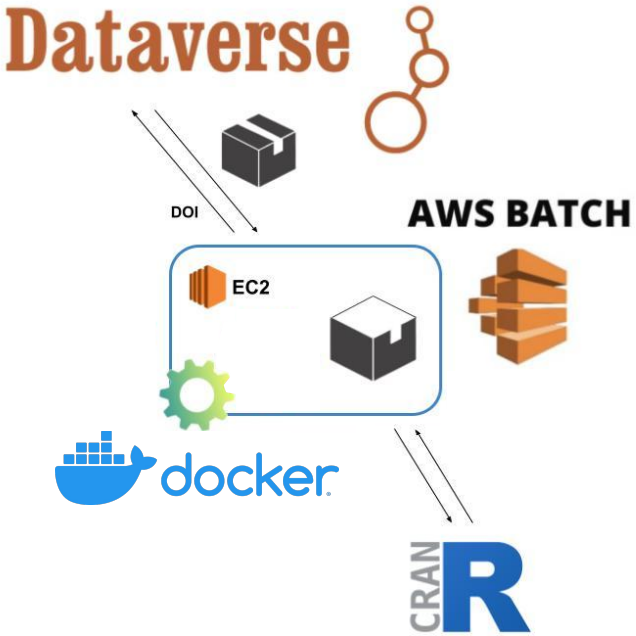
Our data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse to AWS



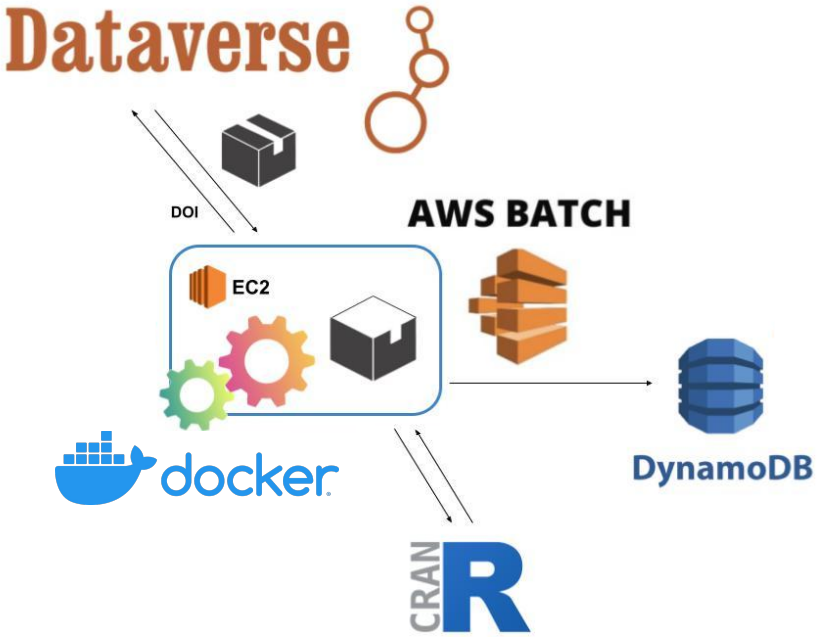
Our data collection workflow

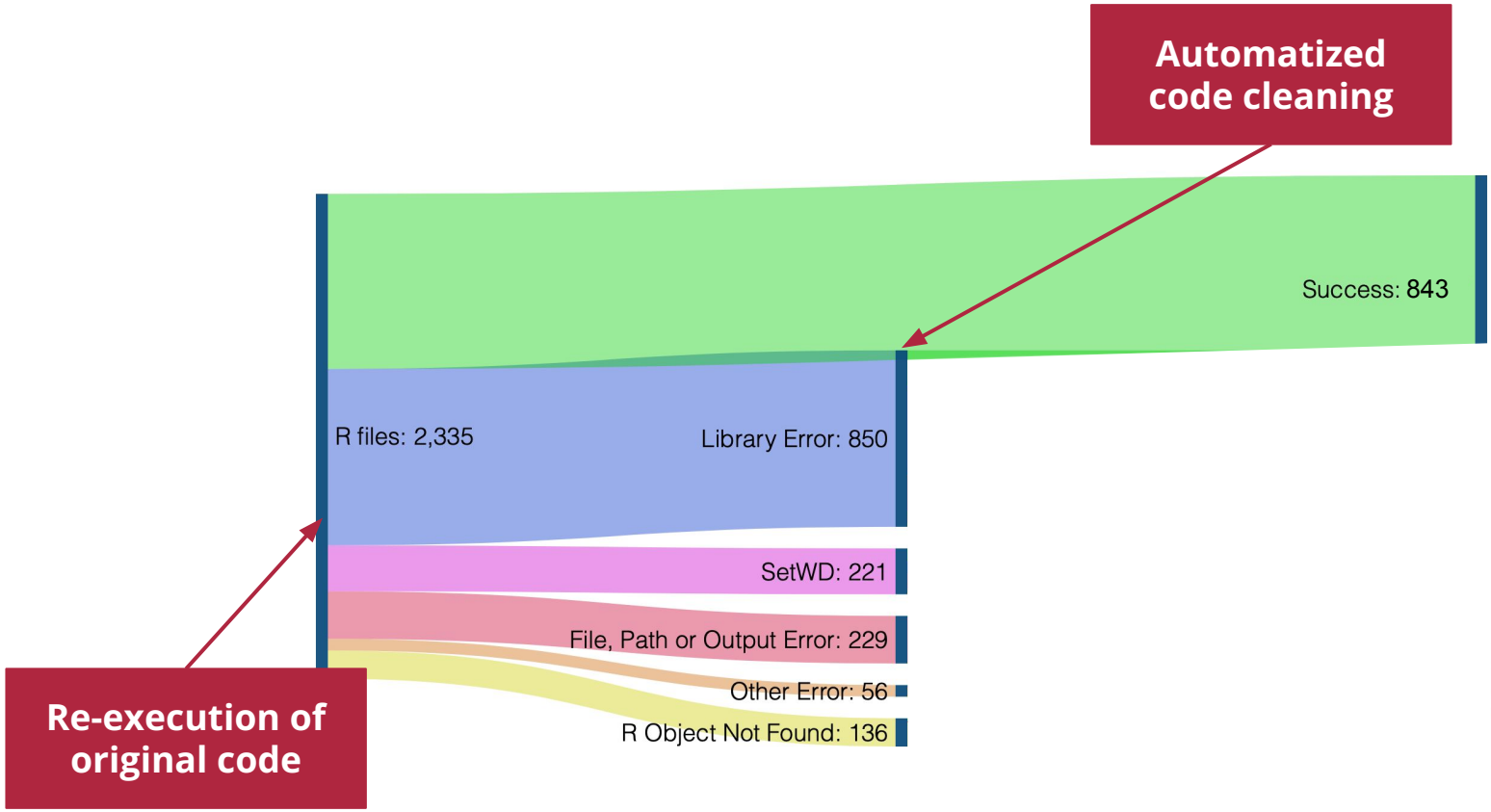
1. Replication dataset is retrieved from Harvard Dataverse to AWS
2. We collect data on the content, install used libraries and attempt automatic code re-execution

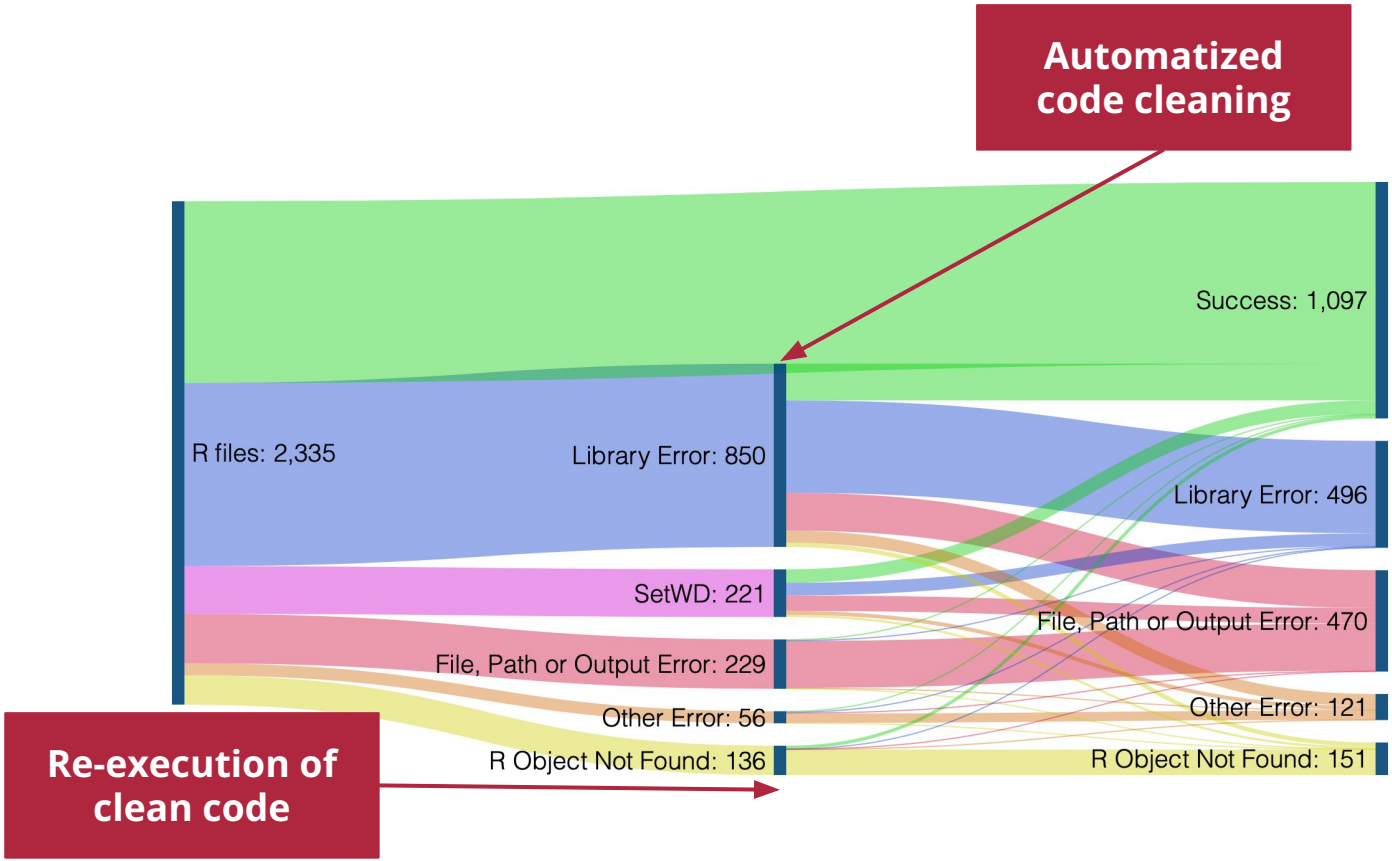


Our data collection workflow

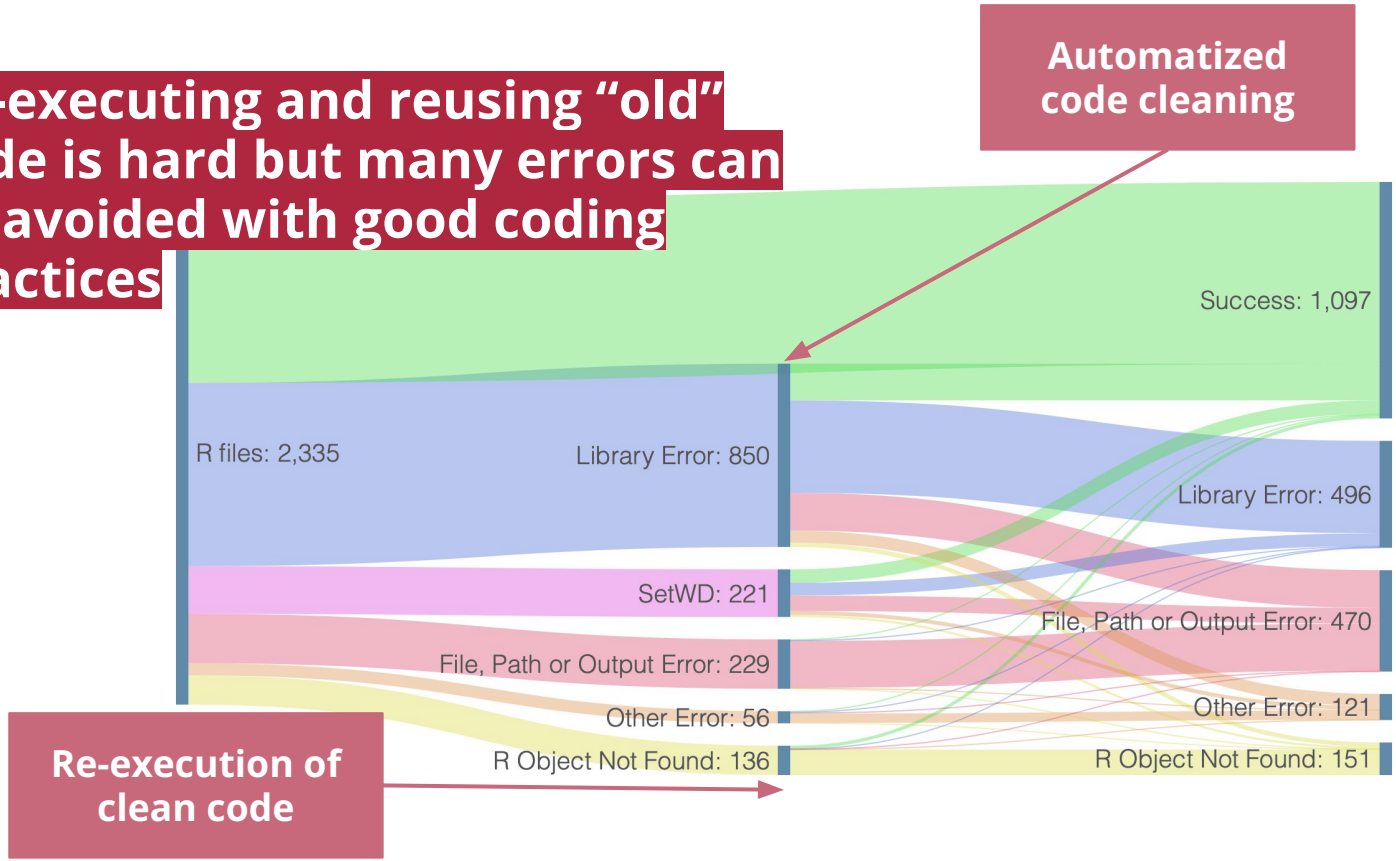
1. Replication dataset is retrieved from Harvard Dataverse to AWS
2. We collect data on the content, install used libraries and attempt automatic code re-execution
3. The re-execution result and other collected data are passed to the backend database for analysis




Automatized code cleaning
Success: 843
R files: 2,335
Library Error: 850
SetWD: 221
File, Path or Output Error: 229
Other Error: 56
R Object Not Found: 136
Re-execution of original code


Automatized code cleaning
R files: 2,335
Library Error: 850
SetWD: 221
File, Path or Output Error: 229
Other Error: 56
R Object Not Found: 136
Success: 1,097
Library Error: 496
File, Path or Output Error: 470
Other Error: 121
R Object Not Found: 151
Re-execution of clean code

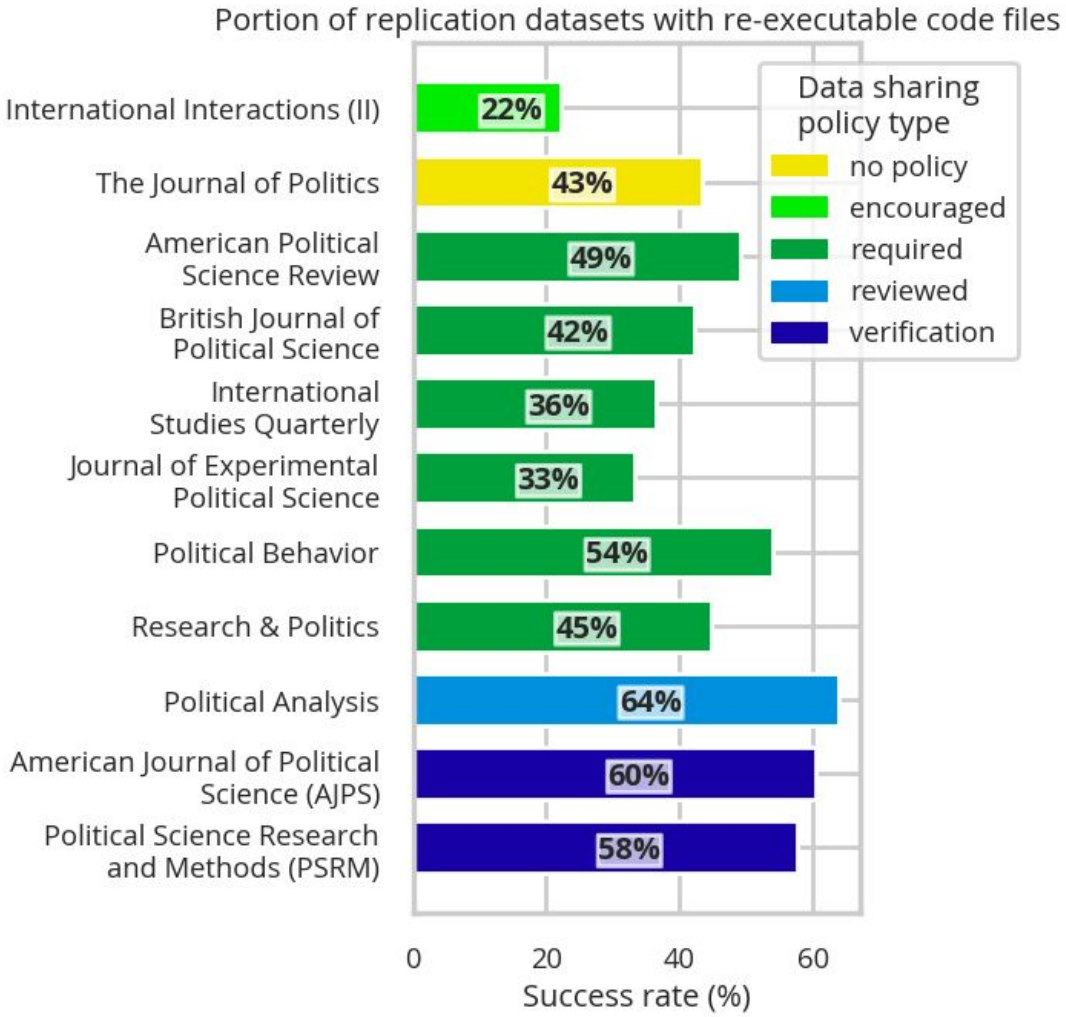
Re-executing and reusing “old” code is hard but many errors can be avoided with good coding practices




Portion of replication datasets with re-executable code files
Data sharing policy type
no policy
encouraged
required
reviewed
verification
International Interactions (II)
22%
The Journal of Politics
43%
American Political Science Review
49%
British Journal of Political Science
42%
International Studies Quarterly
36%
Journal of Experimental Political Science
33%
Political Behavior
54%
Research & Politics
45%
Political Analysis
64%
American Journal of Political Science (AJPS)
60%
Political Science Research and Methods (PSRM)
58%
0
20
40
60
Success rate (%)

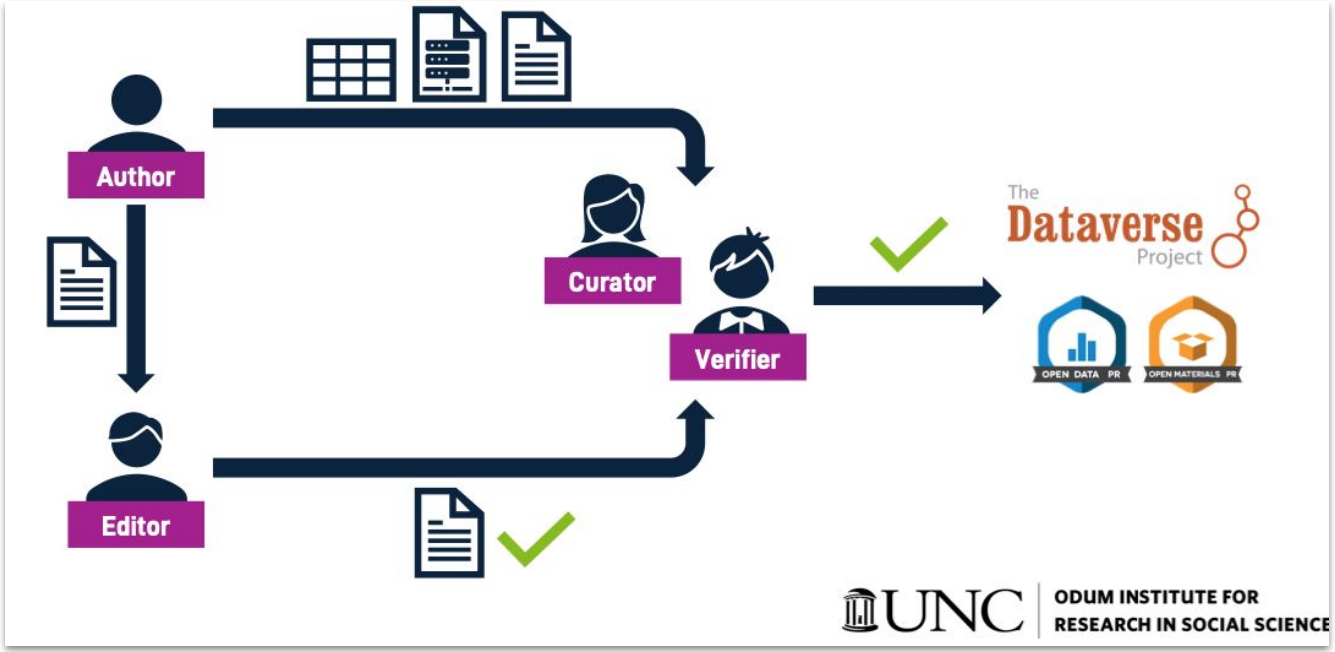
Journals with stricter data policies have higher rate of executable code

Journal average: 47%
Total average: 45%



What can journals do?

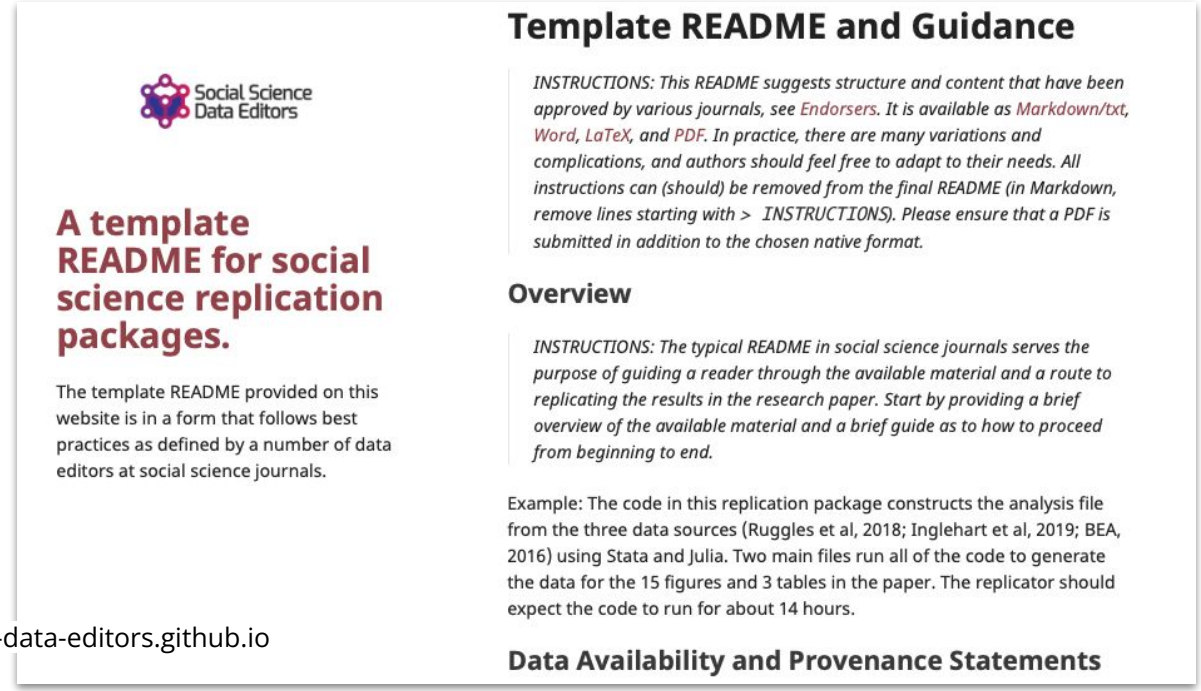
1. Verification or review of deposited materials



@amagooch

1. Verification or review of deposited data and code
2. Reproducibility checklist or README templates for authors

1. Verification or review of deposited data and code
2. Reproducibility checklist or README templates for authors

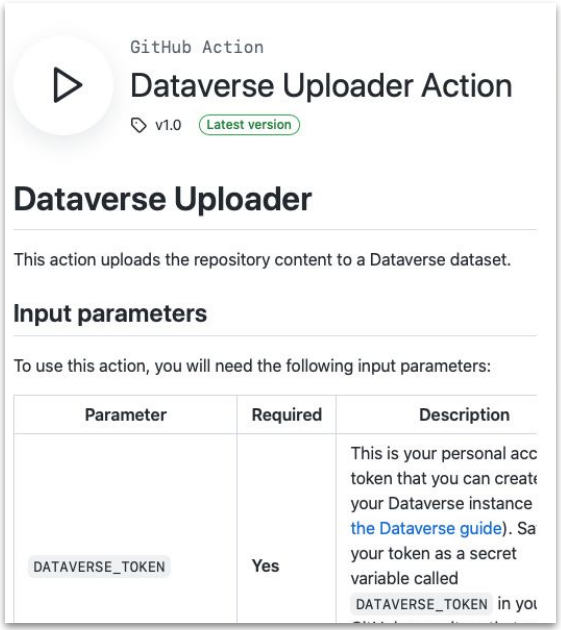


https://social-science-data-editors.github.io

@larsvil

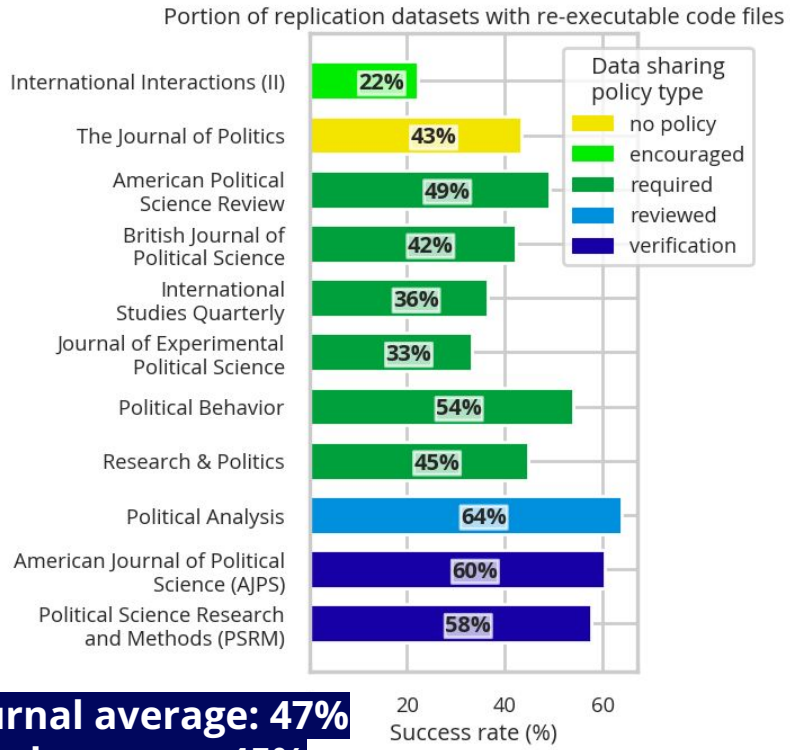
1. Verification or review of deposited data and code
2. Reproducibility checklist or README templates for authors
3. Integration with data and software repositories

1. Verification or review of deposited data and code
2. Reproducibility checklist or README templates for authors
3. Integration with data and software repositories



Email: anatrisovic@g.harvard.edu
Twitter & GitHub: atrisovic

ALFRED P. SLOAN FOUNDATION



Journal average: 47%
Total average: 45%

1. **Verification or review** of deposited materials
2. Reproducibility **checklist** or README **templates** for authors
3. **Integration** with data and software repositories