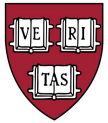
ALFRED P. SLOAN
FOUNDATION



The Landscape of Data Sharing and Computational Reproducibility for Social Research

Pew Research Center
Feb 25, 2022

Ana Trisovic, Harvard University

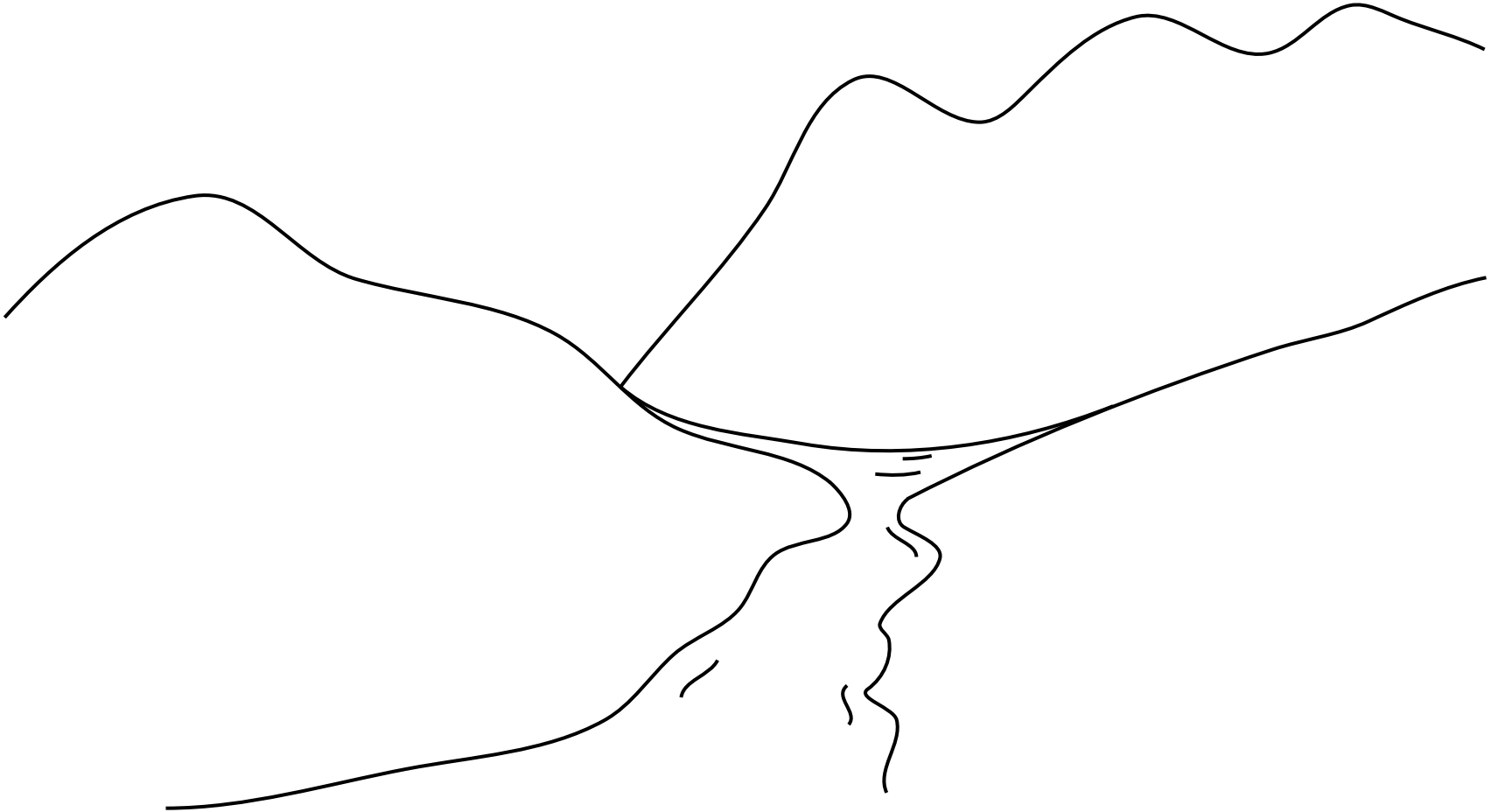
“An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the software, [data] … and set of instructions which generated the figures. ”

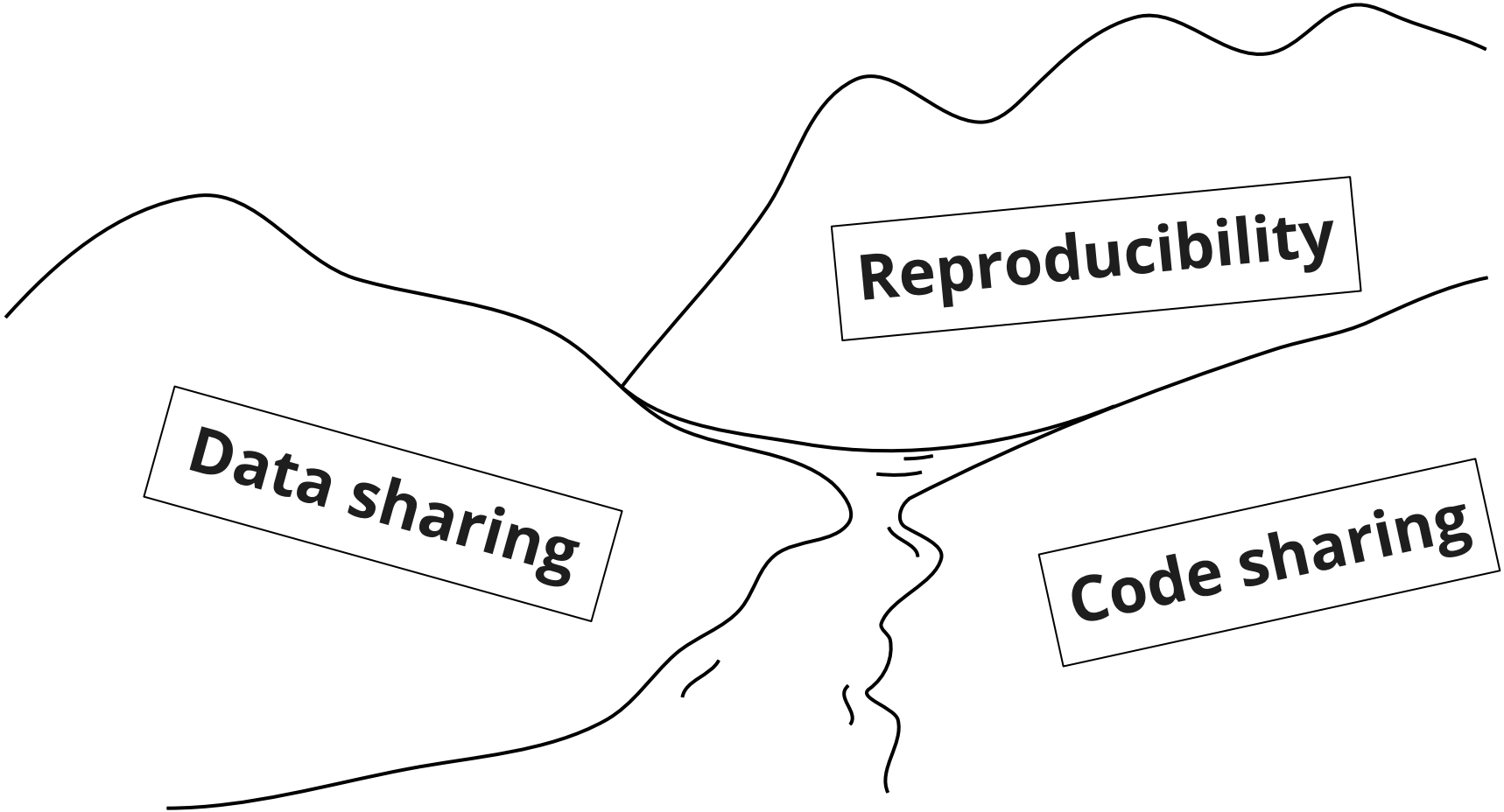
~ Prof Claerbout

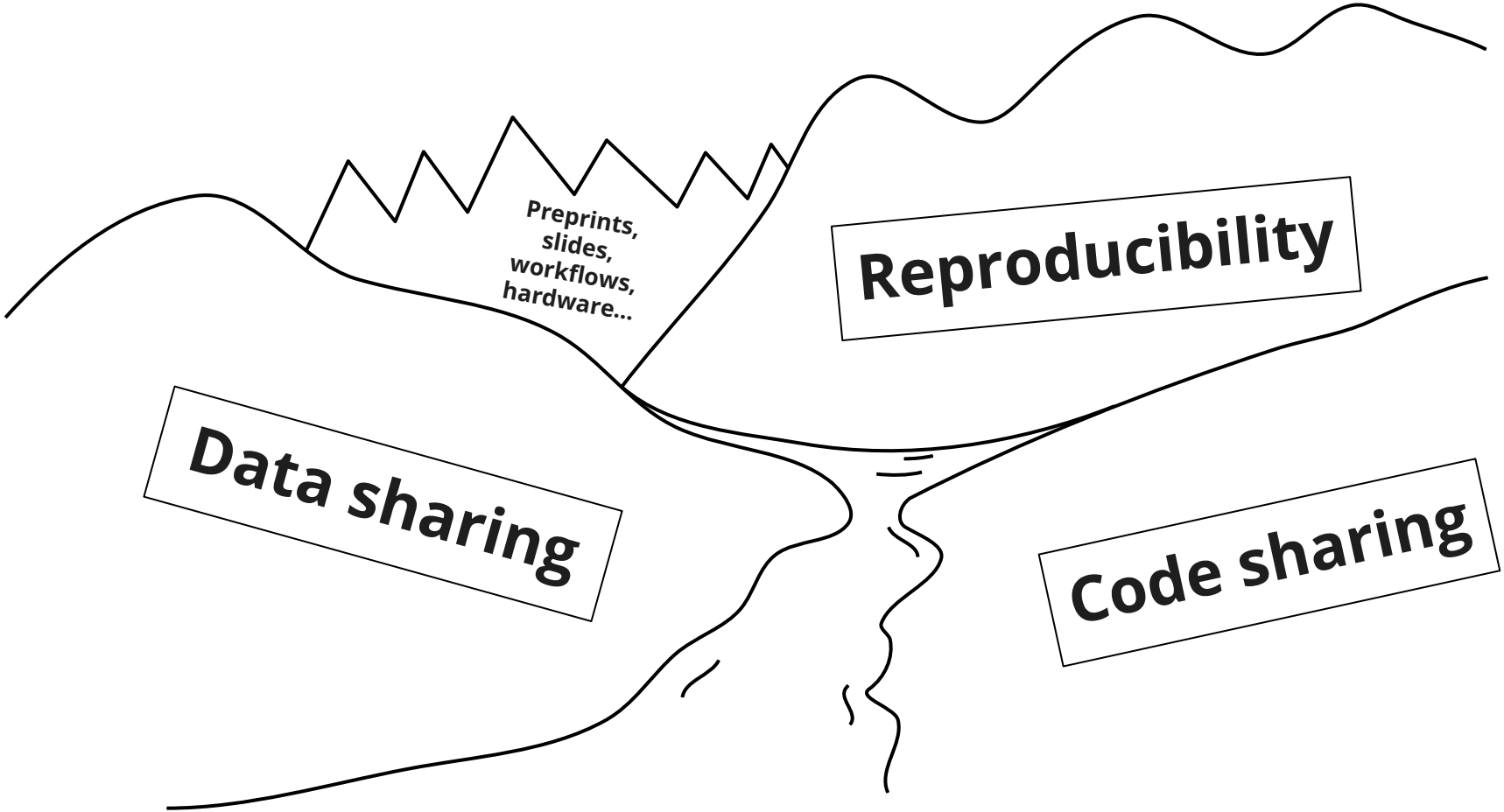
Buckheit. 1995. Wavelab and reproducible research.

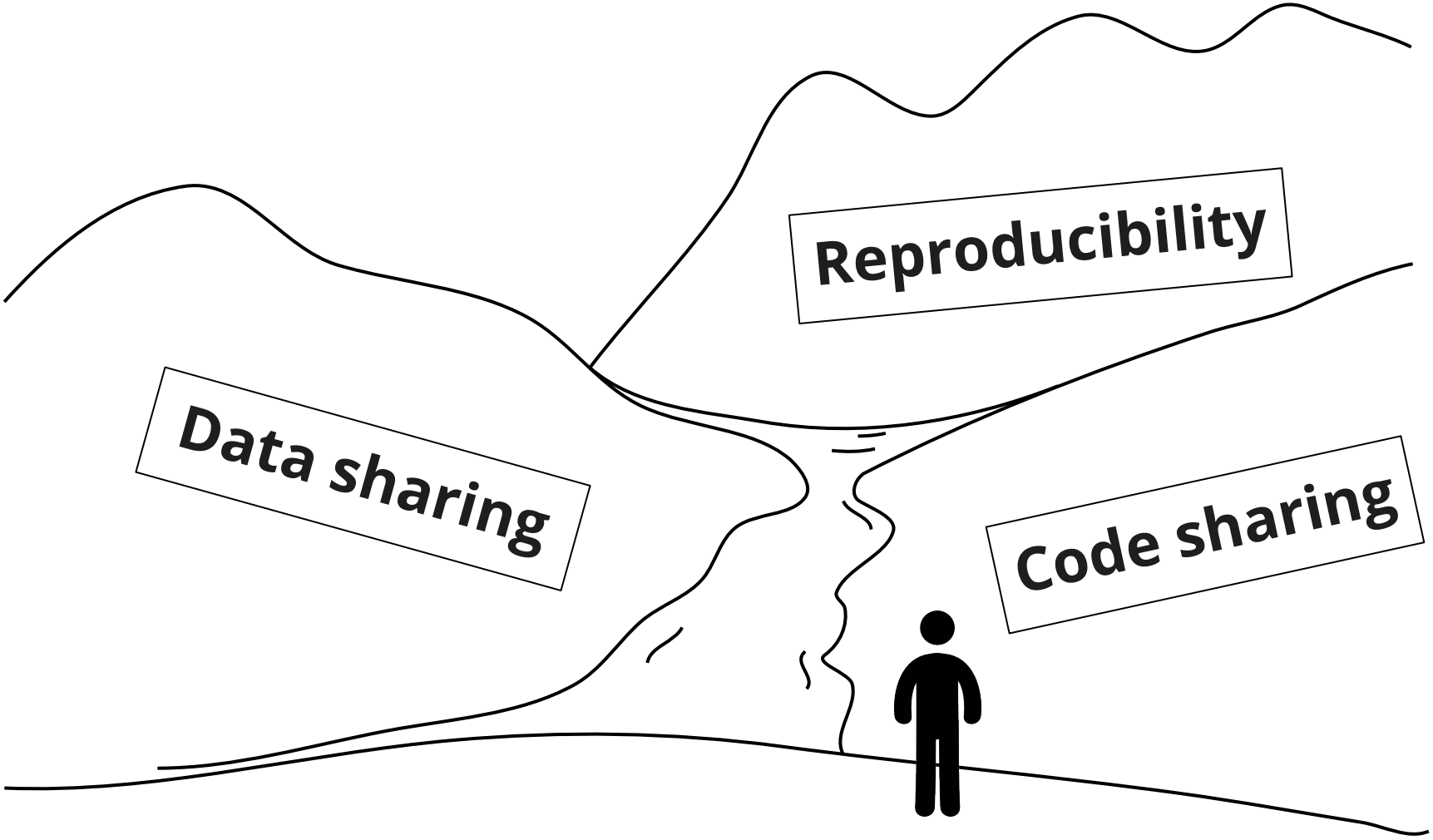
Reproducibility: “obtaining consistent computational results using the same input data, steps, code, and conditions of analysis”

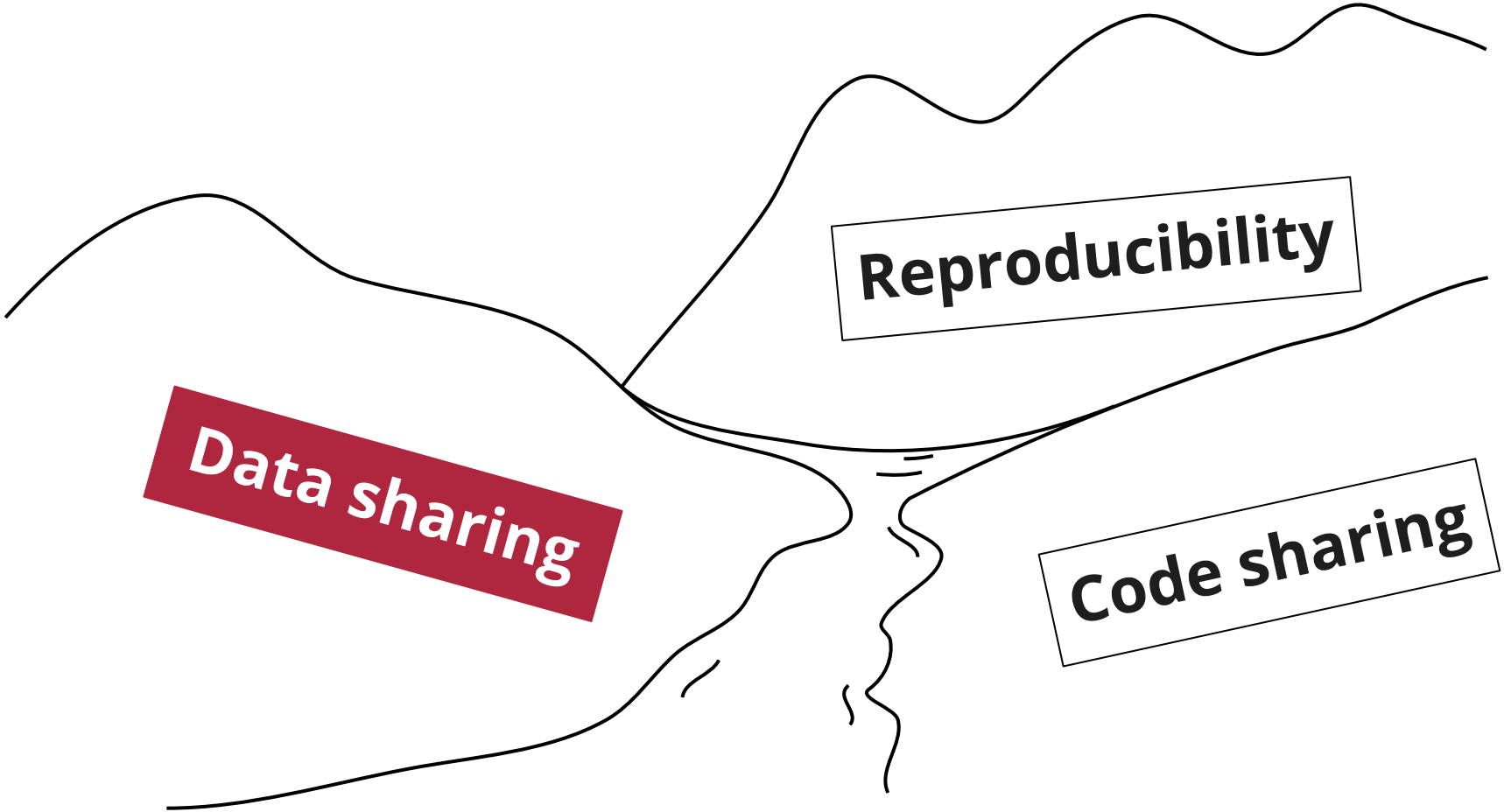
National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science.




Reproducibility
Data sharing
Code sharing

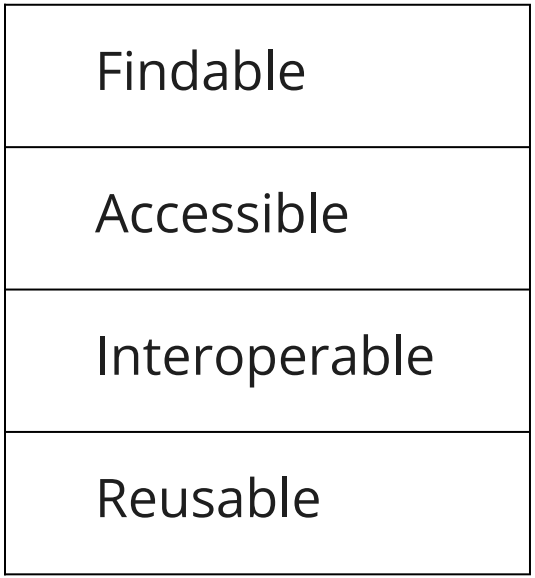

Preprints,
slides,
workflows,
hardware...
Reproducibility
Data sharing
Code sharing


Reproducibility
Data sharing
Code sharing


Reproducibility
Data sharing
Code sharing

- A free and open-source software platform to archive, share, and cite research data
    - Focus on data sharing and making data available

76 institutions around the globe run Dataverse installations as their official data repository

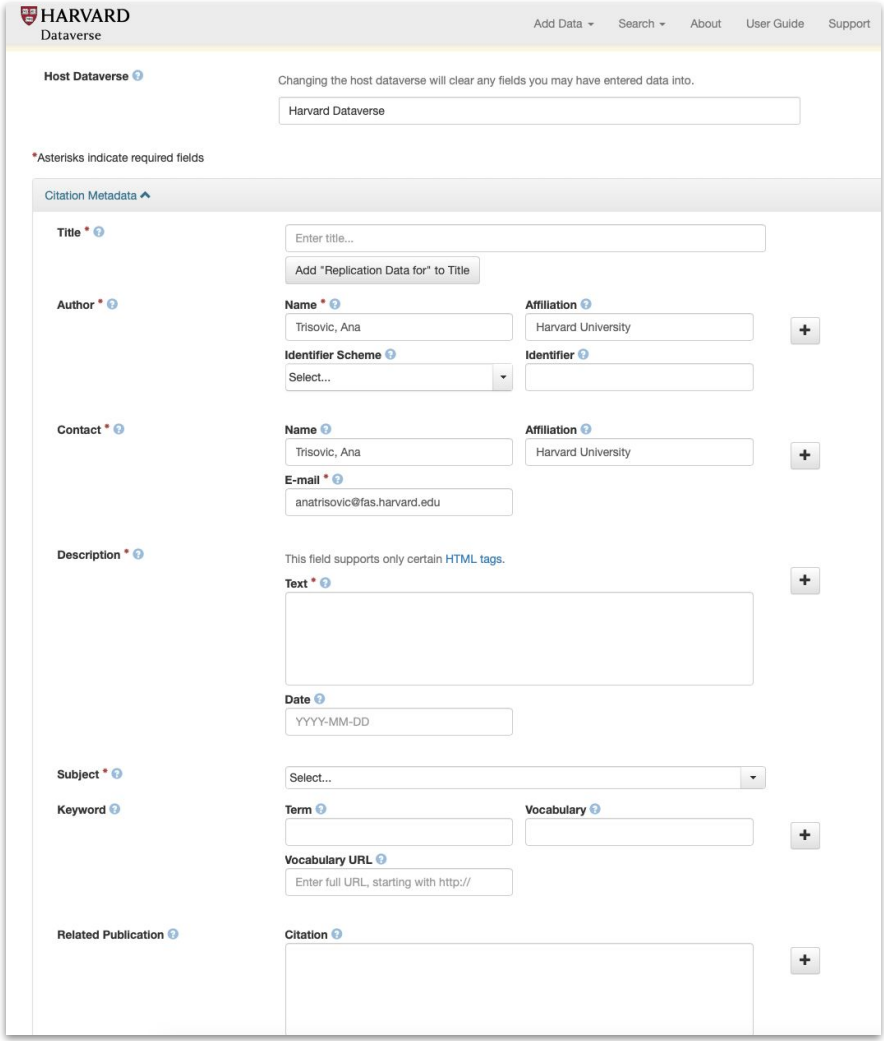

Findable
Accessible
Interoperable
Reusable

Findable	Describe data in metadata, assign DOI Metadata record is shared in data repository
Accessible	Accessible but not necessarily open Standard access protocol
Interoperable	File format open or proprietary Formal knowledge representation
Reusable	License and usage rights Data provenance

Wilkinson, M. D., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship.

Two approaches for data sharing:

- Via UI in the web browser
- Via API (command line or Dataverse Software clients)



Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0

Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", https://doi.org/10.7910/DVN/Z5BTCQ, Harvard Dataverse, V2, UNF:6:YTyX+kJtxsSZUNEND/3GGg== [fileUNF]

Cite Dataset Learn about Data Citation Standards.

Access Dataset | Contact Owner | Share

Dataset Metrics

1,092 Downloads

Description How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinions changed immediately and substantially when their party switched its policy position---even when the new position went against citizens' previously held views. These findings advance the current, largely experimental literature on partisan elite influence. (2020-03-26)

Subject Social Sciences

Keyword Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

Related Publication Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. http://ajps.org/

Notes This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the Center for Open Science.

Files | Metadata | Terms | Versions

Search this dataset...

Filter by
File Type: All Access: All

Sort

1 to 10 of 25 Files

Download

build_data.R
R Syntax - 12.1 KB
Published Jun 29, 2020
56 Downloads
MD5: a94...597

codebook ess.pdf
Adobe PDF - 508.8 KB
Published Jun 29, 2020
46 Downloads

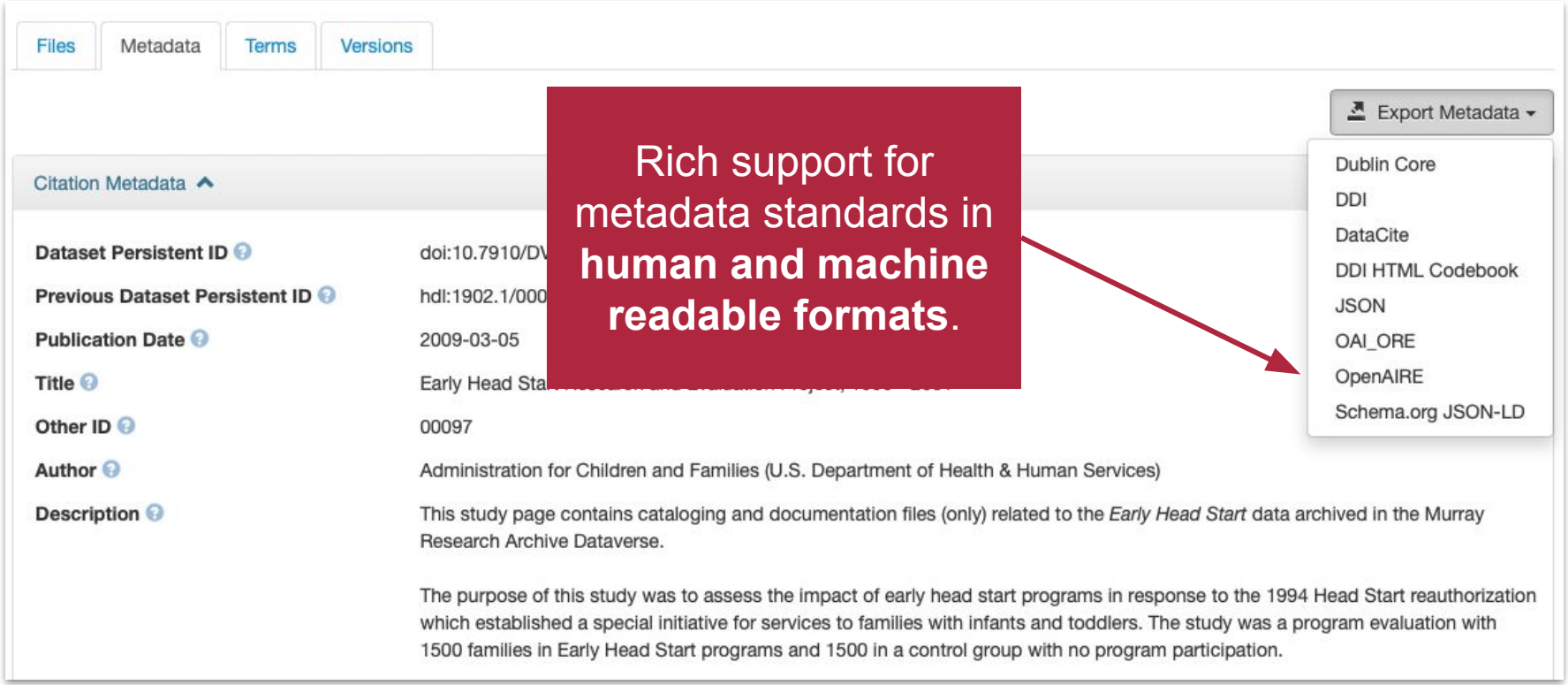
Dataset version

Unique DOI

Mandatory citation-level metadata

Center for Open Science Badges

Code, documentation and other files


Files
Metadata
Terms
Versions
Export Metadata
Dublin Core
DDI
DataCite
DDI HTML Codebook
JSON
OAI_ORE
OpenAIRE
Schema.org JSON-LD
Rich support for metadata standards in **human and machine readable formats**.
Citation Metadata
Dataset Persistent ID
doi:10.7910/DV
Previous Dataset Persistent ID
hdl:1902.1/000
Publication Date
2009-03-05
Title
Early Head Sta
Other ID
00097
Author
Administration for Children and Families (U.S. Department of Health & Human Services)
Description
This study page contains cataloging and documentation files (only) related to the *Early Head Start* data archived in the Murray Research Archive Dataverse.
The purpose of this study was to assess the impact of early head start programs in response to the 1994 Head Start reauthorization which established a special initiative for services to families with infants and toddlers. The study was a program evaluation with 1500 families in Early Head Start programs and 1500 in a control group with no program participation.


zenodo


Dataverse®


figshare


DRYAD

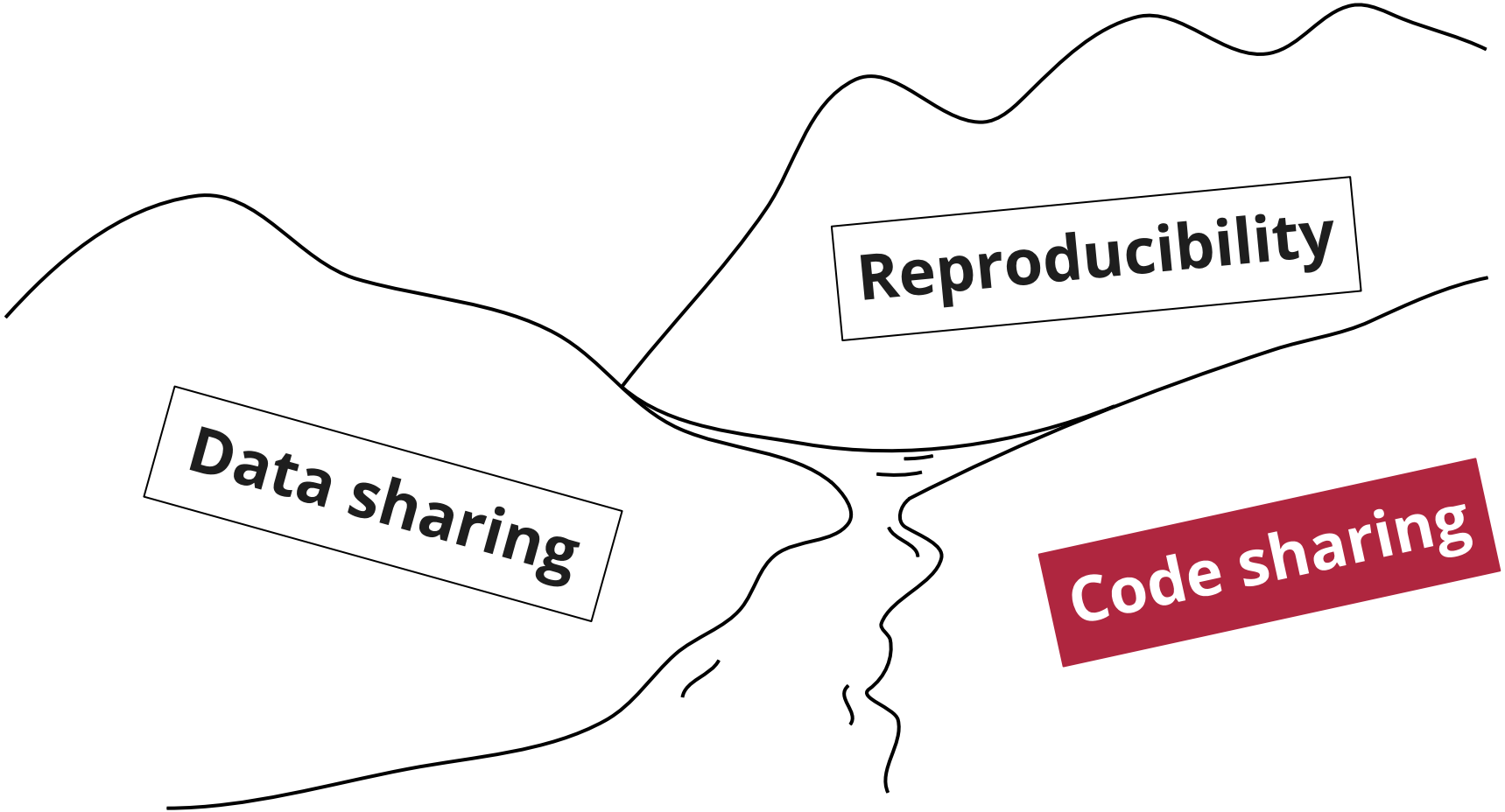

ckan


re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Data sharing

Summary

- FAIR principles are a standard for sharing research datasets
- Data repositories implement FAIR with standard (and domain-specific) metadata, identifiers, data curation, etc.

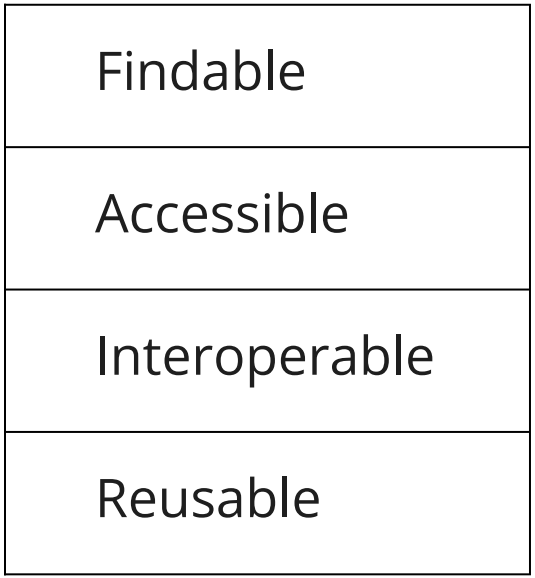

Reproducibility
Data sharing
Code sharing

Research software includes code files, algorithms, scripts, workflows and executables created during the research process or for a research purpose.

Gruenpeter, M et al. 2021. Defining Research Software.

- Research software is a critical component of research
  - Software intensive projects are a majority of publications
  - Most-cited papers are methods and software
  - Funders encourage open-source software
- Effort to recognize software are a valuable research output

Katz, D. S., et al. 2022. FAIR Research Software as a Step Towards Better Research.


Findable
Accessible
Interoperable
Reusable

Findable	Describe software with metadata, assign id Metadata are FAIR and searchable/indexable
Accessible	Retrievable by its identifier via standard protocol Metadata are accessible even when sw is not
Interoperable	Software interoperates with other software by exchanging data and/or metadata
Reusable	Software is usable (executable) and reusable (documented, licensed, extendable)

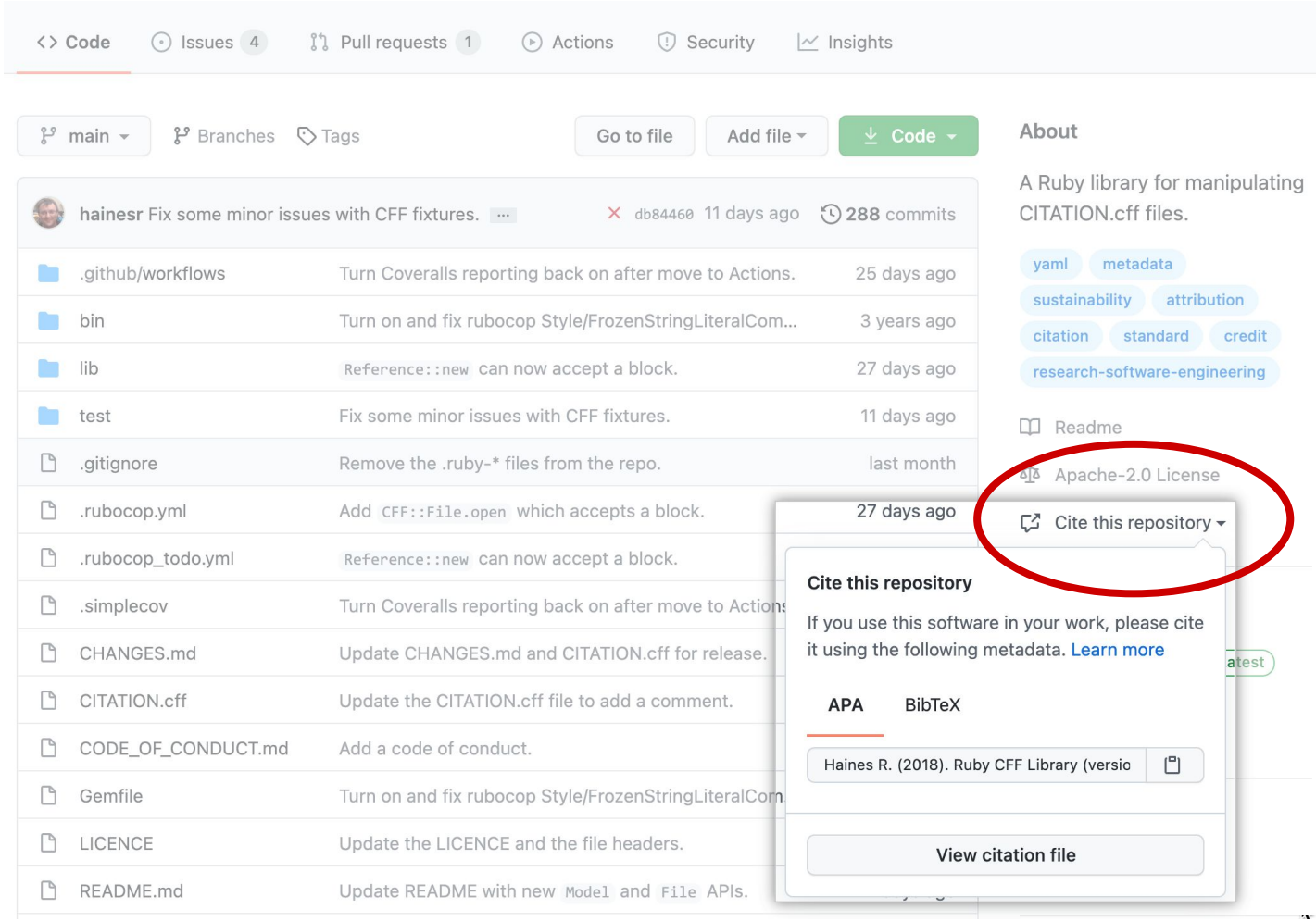
Hong, N.P.C., et al. 2021. FAIR Principles for Research Software.

GitLab

- Long-term solution for sharing software?
- Where to store metadata?
- Software citation?

Citation File Format and `CITATION.cff`

```
cff-version: 1.2.0
message: "If you use this software, please cite it as below."
authors:
- family-names: "Lisa"
  given-names: "Mona"
  orcid: "https://orcid.org/0000-0000-0000-0000"
- family-names: "Bot"
  given-names: "Hew"
  orcid: "https://orcid.org/0000-0000-0000-0000"
title: "My Research Software"
version: 2.0.4
doi: 10.5281/zenodo.1234
date-released: 2017-12-18
url: "https://github.com/github/linguist"
```


Code
Issues 4
Pull requests 1
Actions
Security
Insights
main
Branches
Tags
Go to file
Add file
Code
hainesr Fix some minor issues with CFF fixtures.
db84460 11 days ago
288 commits
.github/workflows
Turn Coveralls reporting back on after move to Actions.
25 days ago
bin
Turn on and fix rubocop Style/FrozenStringLiteralCom...
3 years ago
lib
Reference::new can now accept a block.
27 days ago
test
Fix some minor issues with CFF fixtures.
11 days ago
.gitignore
Remove the .ruby-* files from the repo.
last month
.rubocop.yml
Add CFF::File.open which accepts a block.
27 days ago
.rubocop_todo.yml
Reference::new can now accept a block.
.simplecov
Turn Coveralls reporting back on after move to Actions
CHANGES.md
Update CHANGES.md and CITATION.cff for release.
CITATION.cff
Update the CITATION.cff file to add a comment.
CODE_OF_CONDUCT.md
Add a code of conduct.
Gemfile
Turn on and fix rubocop Style/FrozenStringLiteralCom
LICENCE
Update the LICENCE and the file headers.
README.md
Update README with new Model and File APIs.
About
A Ruby library for manipulating CITATION.cff files.
yaml
metadata
sustainability
attribution
citation
standard
credit
research-software-engineering
Readme
Apache-2.0 License
Cite this repository
Cite this repository
If you use this software in your work, please cite it using the following metadata. Learn more
APA
BibTeX
Haines R. (2018). Ruby CFF Library (versio
View citation file

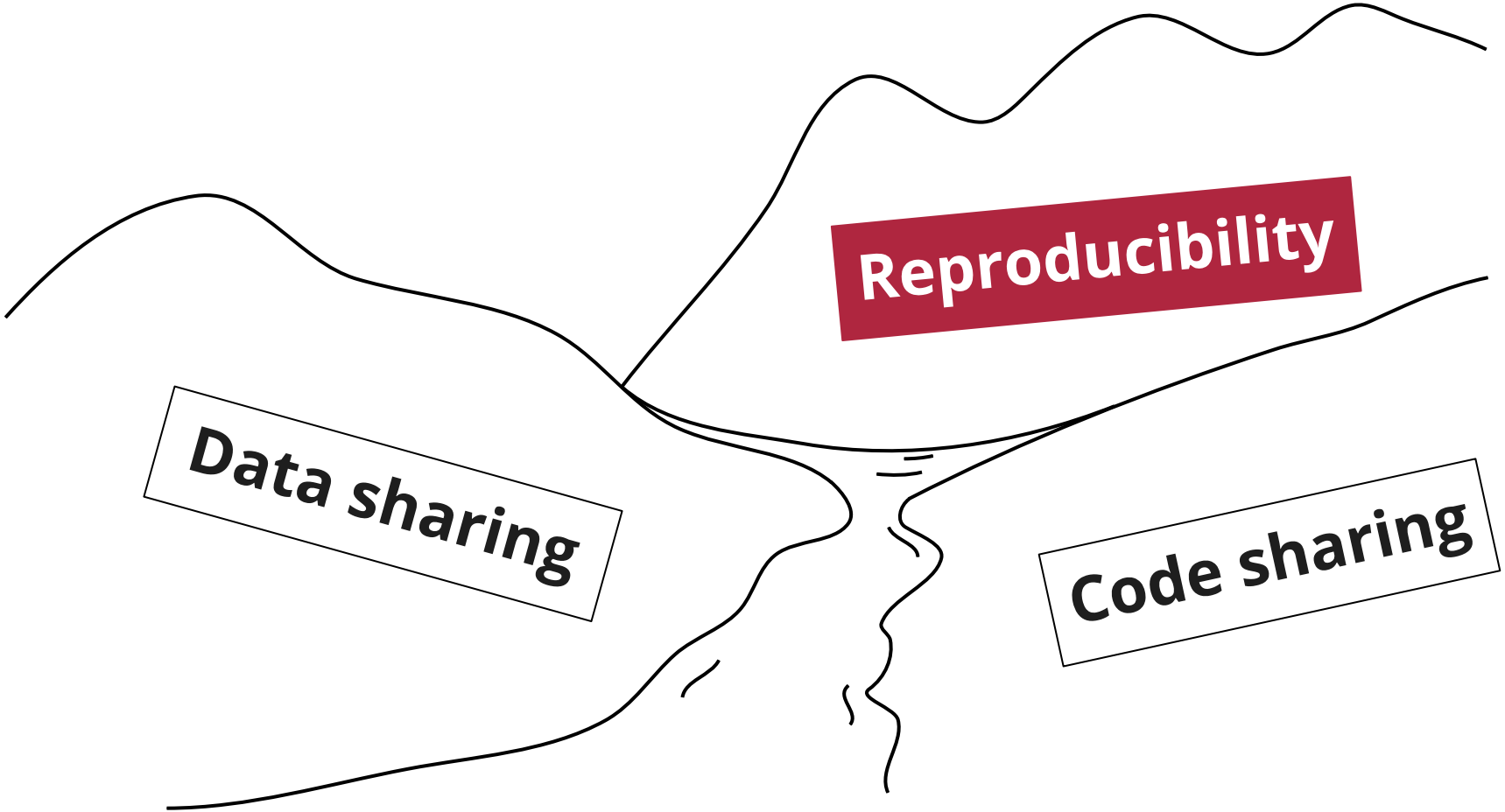
Software archival

- SoftwareHeritage.org
  - Initiative to collect, preserve and share software source code for the long-term future
- Zenodo
  - General purpose repository that support software deposits

Code sharing

Summary

- Research software as a recognized scientific output
- FAIR principles for software help discoverability, licencing, attribution, citation, documentation, etc.


Reproducibility
Data sharing
Code sharing

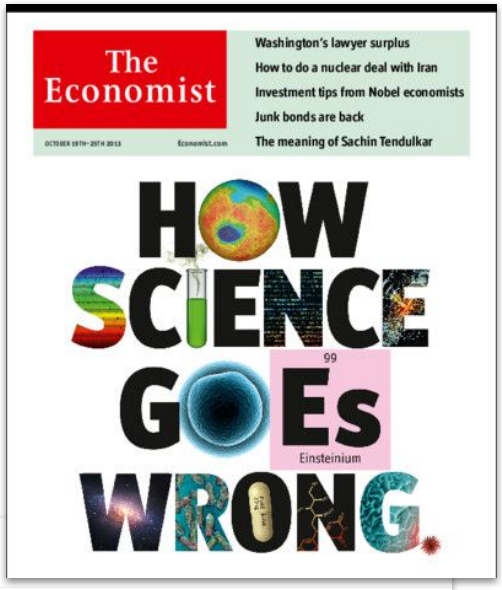
Vox

UKRAINE CORONAVIRUS RECODE THE GOODS FUTURE PERFECT THE HIGHLIGHT CROSSWORD MORE

Science has been in a "replication crisis" for a decade. Have we learned anything?

Bad papers are still published. But some other things might be getting better.

By Kelsey Piper | Oct 14, 2020, 12:20pm EDT



TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

NEWS & OPINION PUBLICATIONS CATEGORIES

Home / News & Opinion

Potential Causes of Irreproducibility Revealed

Five independent groups got different results in a drug-response experiment, despite sharing protocols, reagents, and cell lines. The researchers identify technical variables could be to blame.

nature

Explore content About the journal Publish with us

nature > news feature > article

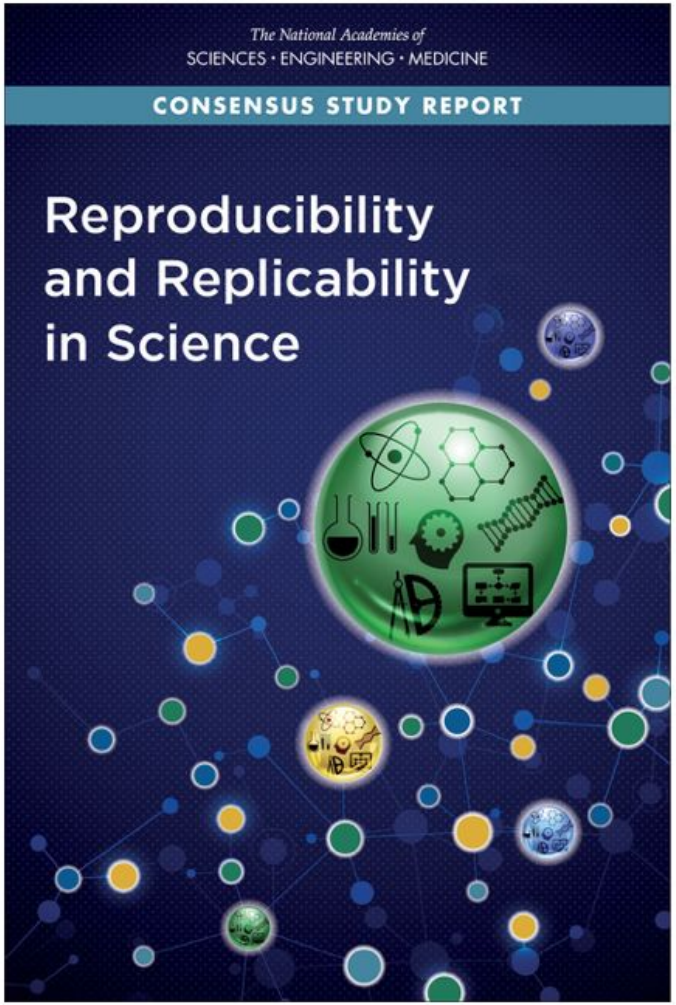
Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

Monya Baker

Nature **533**, 452–454 (2016) | Cite this article

34k Accesses | **1489** Citations | **3920** Altmetric | Metrics



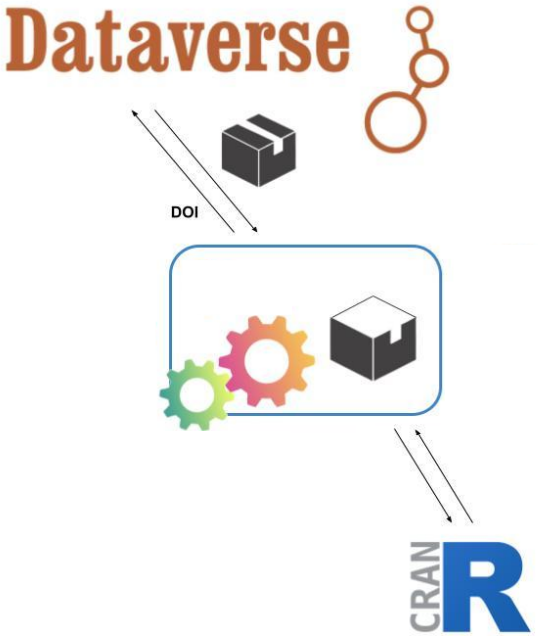
National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science.

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

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

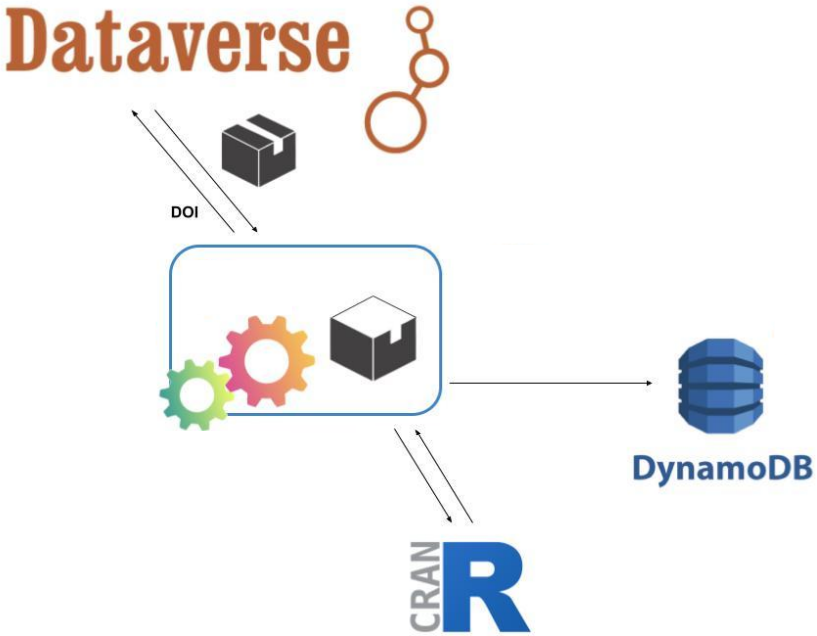
Our data collection workflow

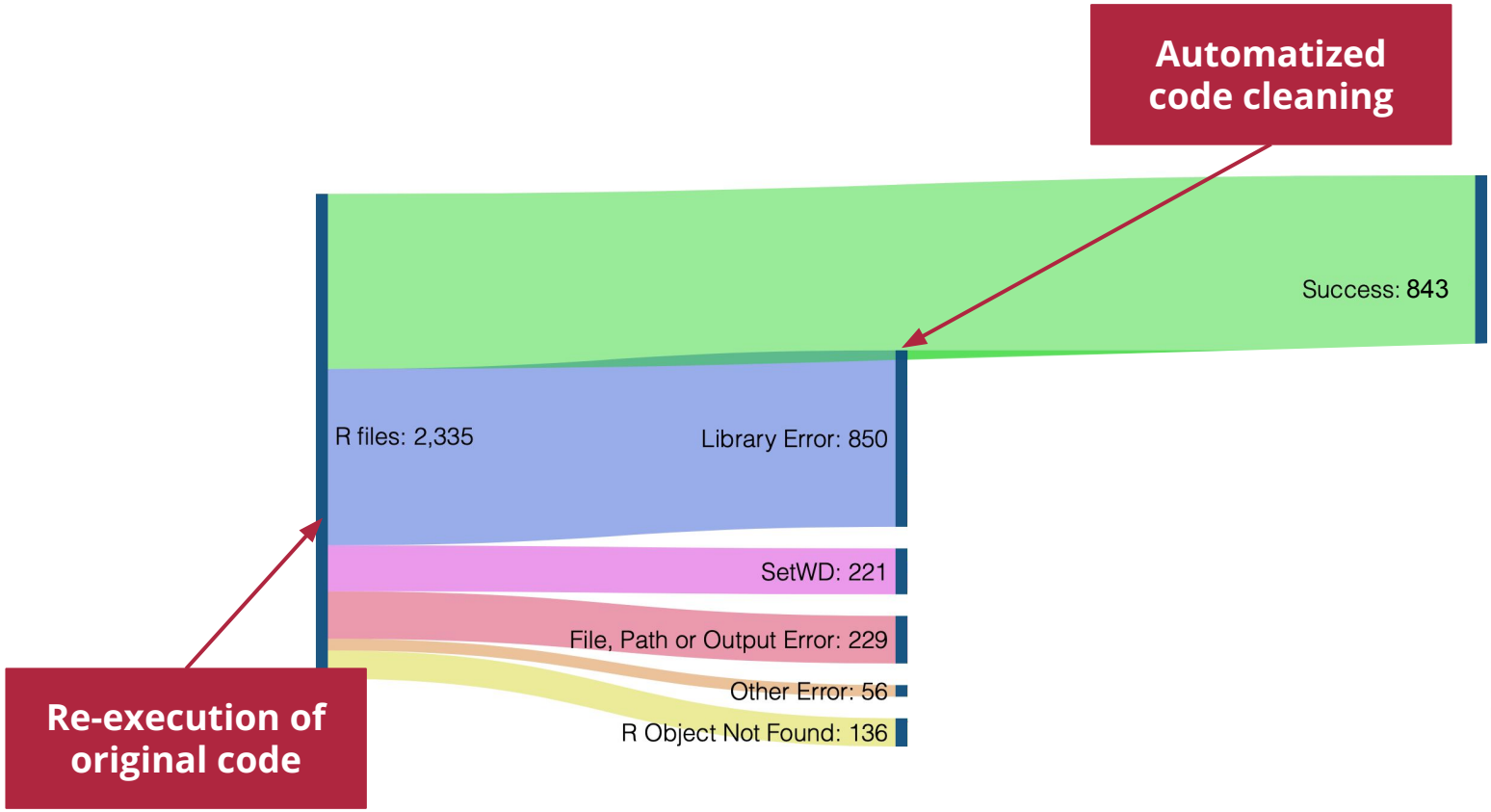
1. Replication dataset is retrieved from Harvard Dataverse
2. We collect data on the content, install used R libraries and attempt automatic code re-execution

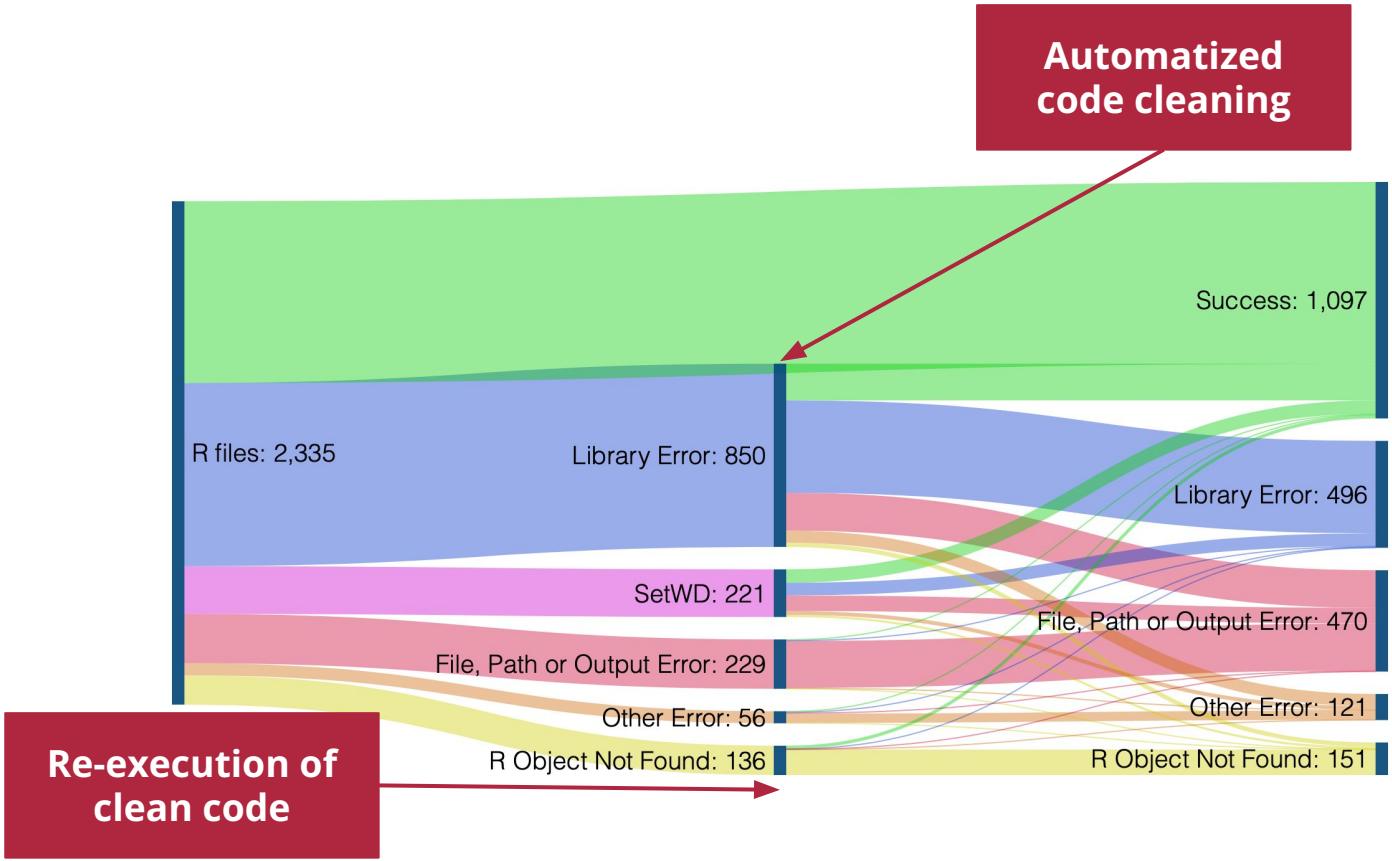


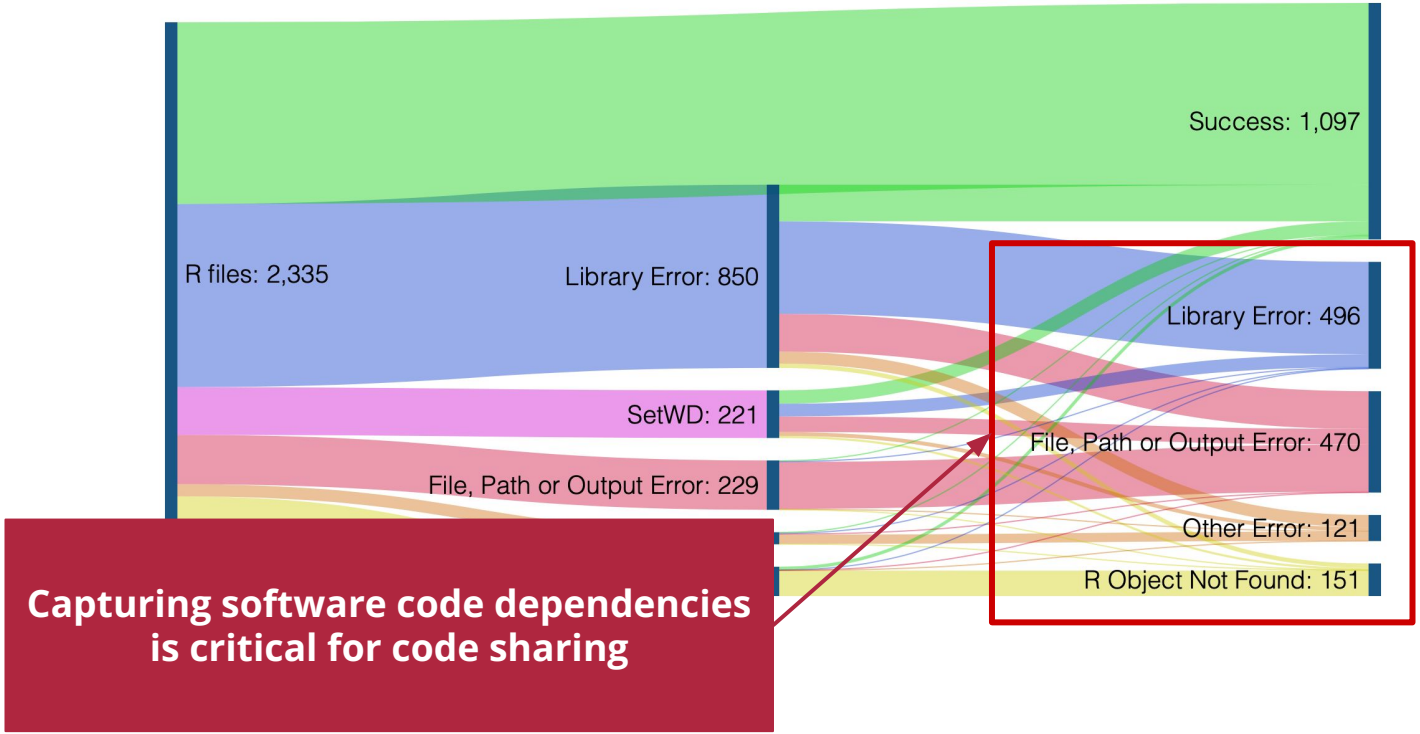
Our data collection workflow

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

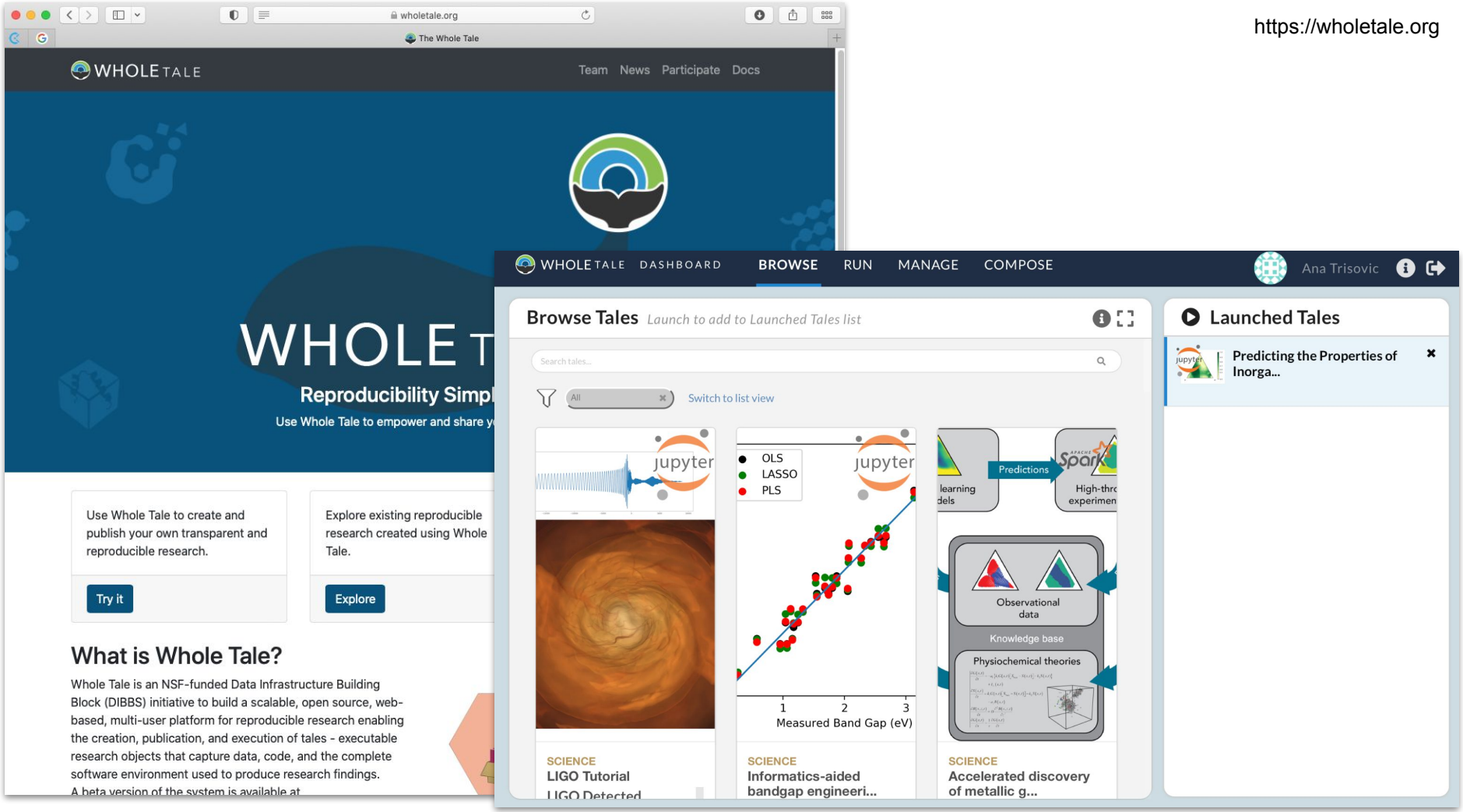



Automatized code cleaning
Success: 843
R files: 2,335
Library Error: 850
SetWD: 221
File, Path or Output Error: 229
Other Error: 56
R Object Not Found: 136
Re-execution of original code

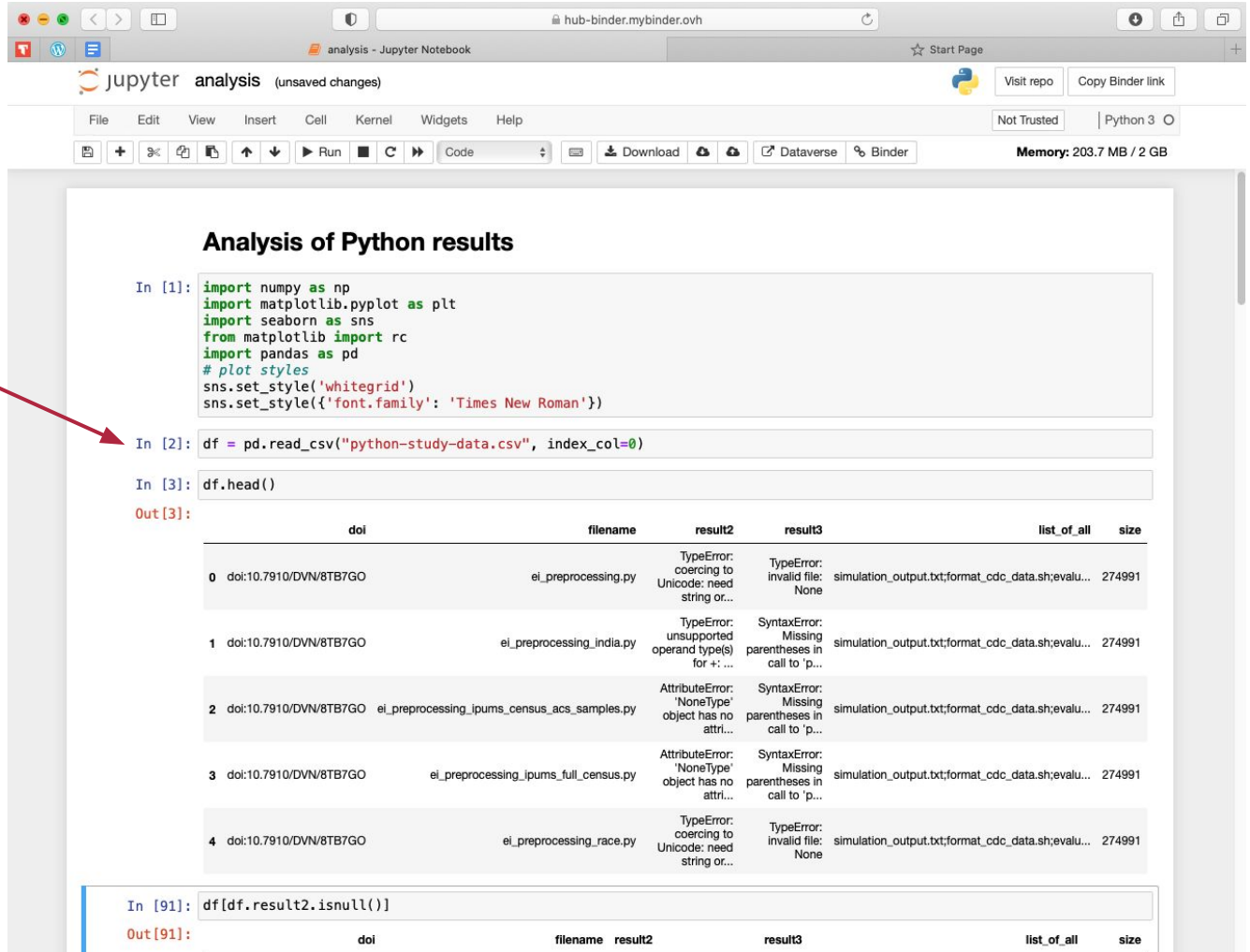

Automatized code cleaning
R files: 2,335
Library Error: 850
SetWD: 221
File, Path or Output Error: 229
Other Error: 56
R Object Not Found: 136
Success: 1,097
Library Error: 496
File, Path or Output Error: 470
Other Error: 121
R Object Not Found: 151
Re-execution of clean code


R files: 2,335
Library Error: 850
SetWD: 221
File, Path or Output Error: 229
Success: 1,097
Library Error: 496
File, Path or Output Error: 470
Other Error: 121
R Object Not Found: 151
Capturing software code dependencies is critical for code sharing

https://wholetale.org



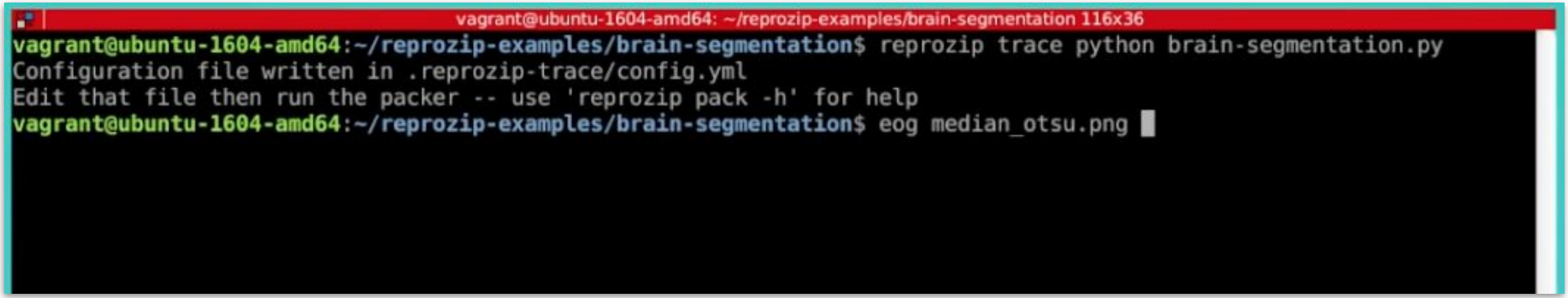
Anyone can explore data, re-run code or modify it



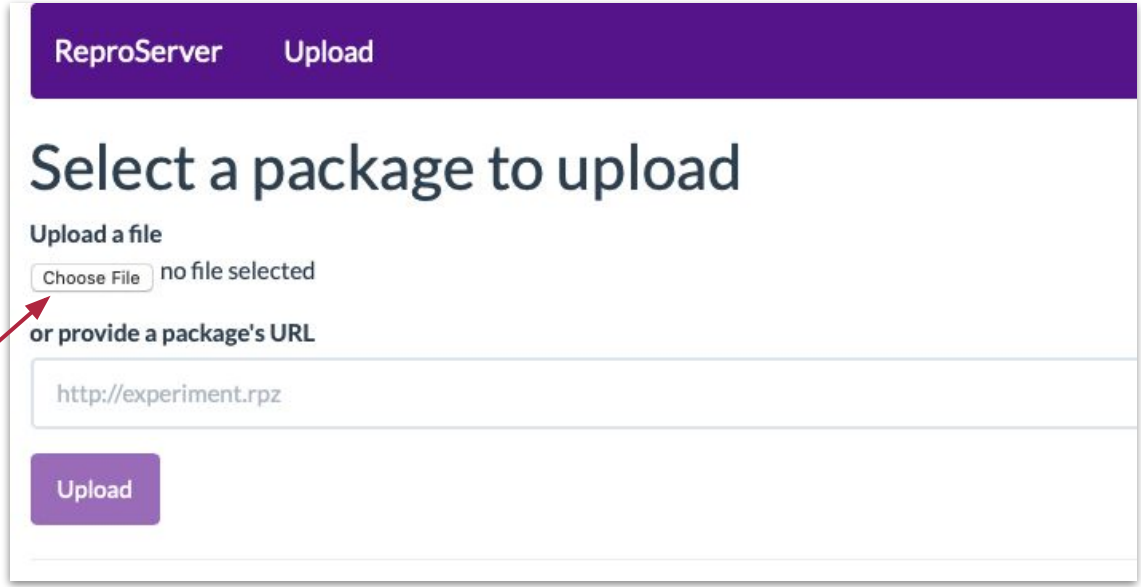
https://mybinder.org

ReproZip

https://www.reprozip.org



ReproZip



https://www.reprozip.org

```
vagrant@ubuntu-1604-amd64: ~/reprozip-examples/brain-segmentation 116x36
vagrant@ubuntu-1604-amd64:~/reprozip-examples/brain-segmentation$ reprozip trace python brain-segmentation.py
Configuration file written in .reprozip-trace/config.yml
Edit that file then run the packer -- use 'reprozip pack -h' for help
vagrant@ubuntu-1604-amd64:~/reprozip-examples/brain-segmentation$ eog median_otsu.png
```


CODE OCEAN


WHOLE
TALE

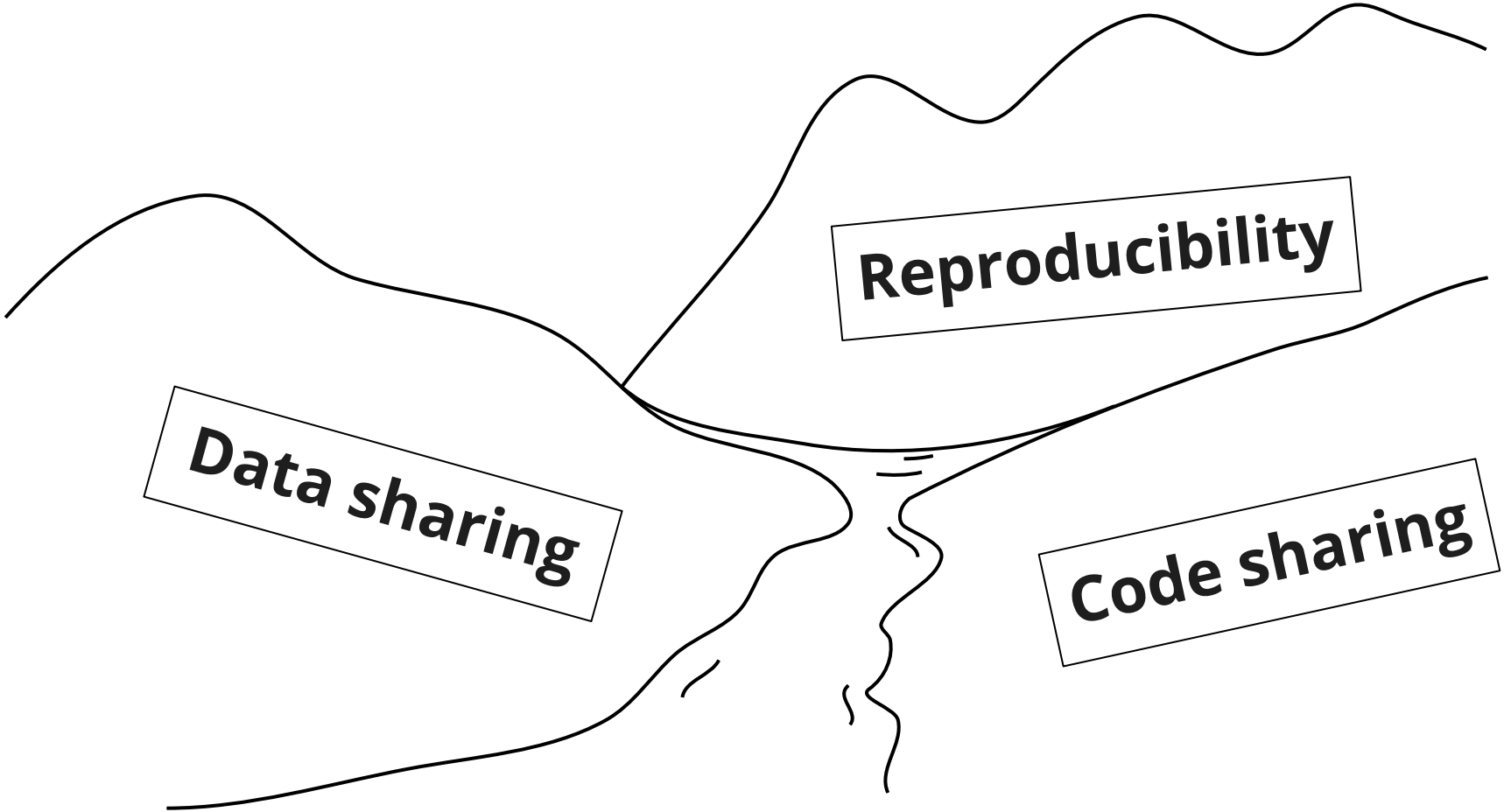

binder

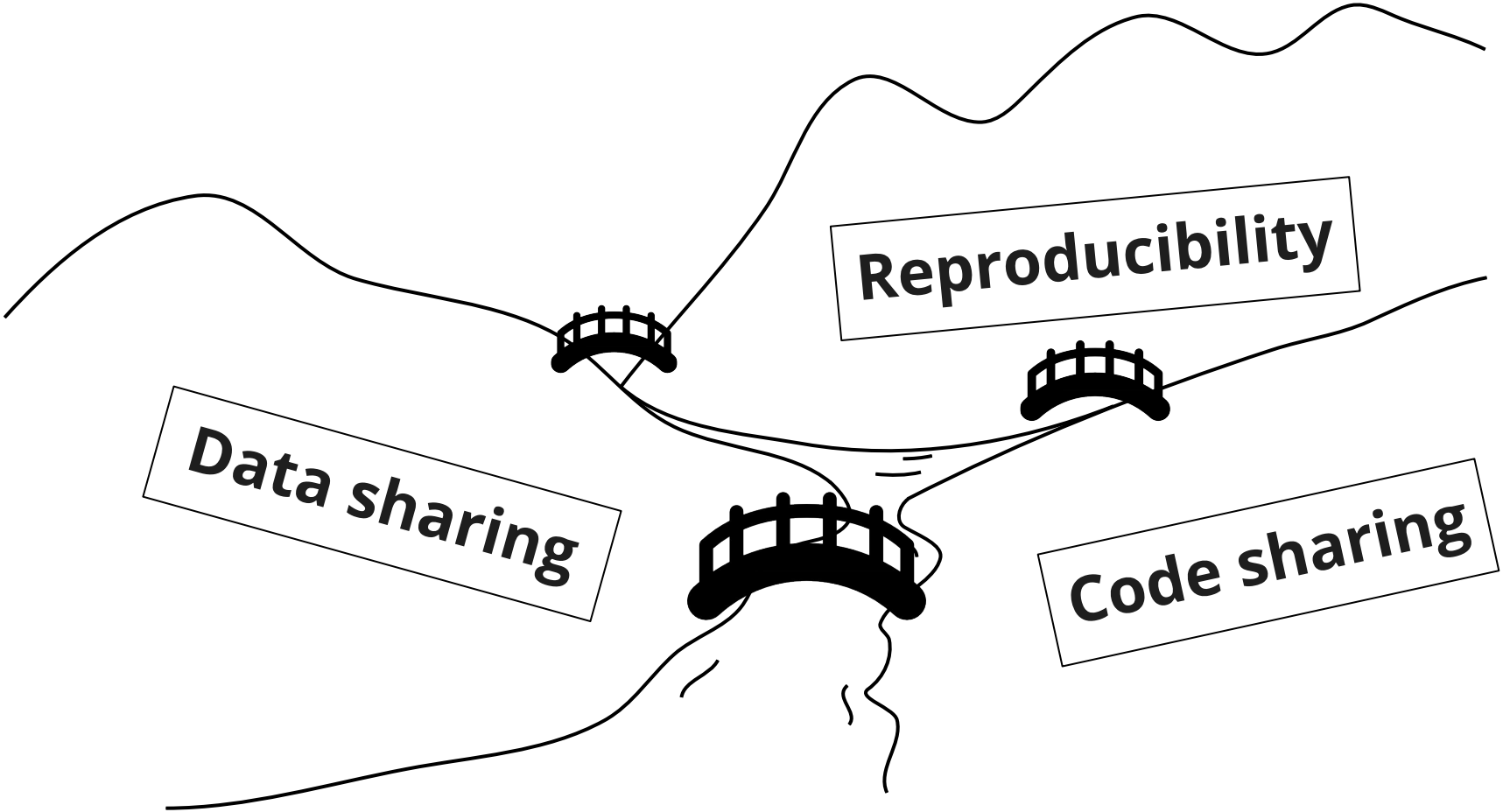

RENKU

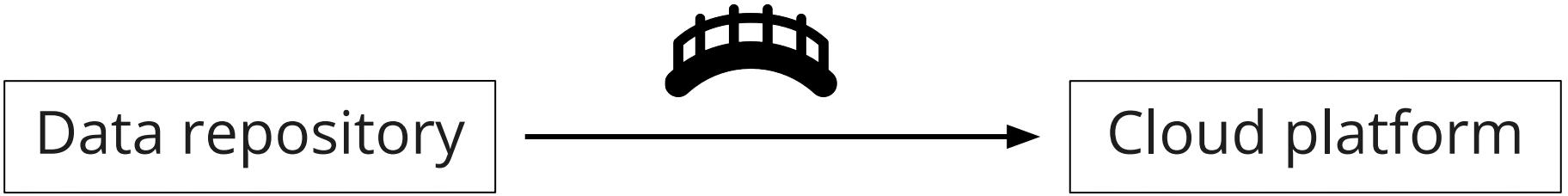

Stencila


ReproZip


colab

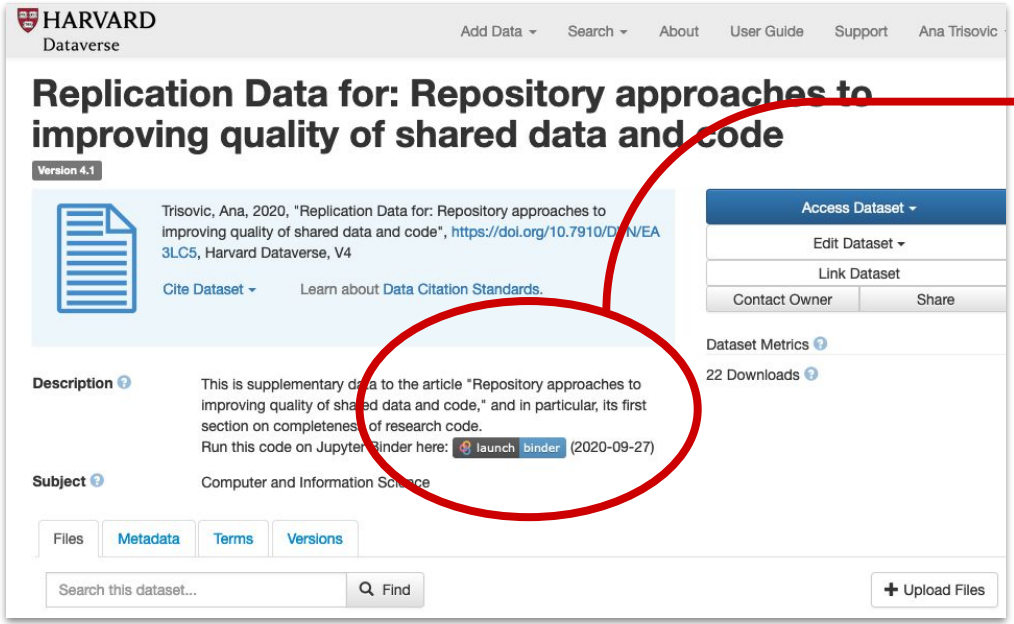

Reproducibility
Data sharing
Code sharing


Reproducibility
Data sharing
Code sharing

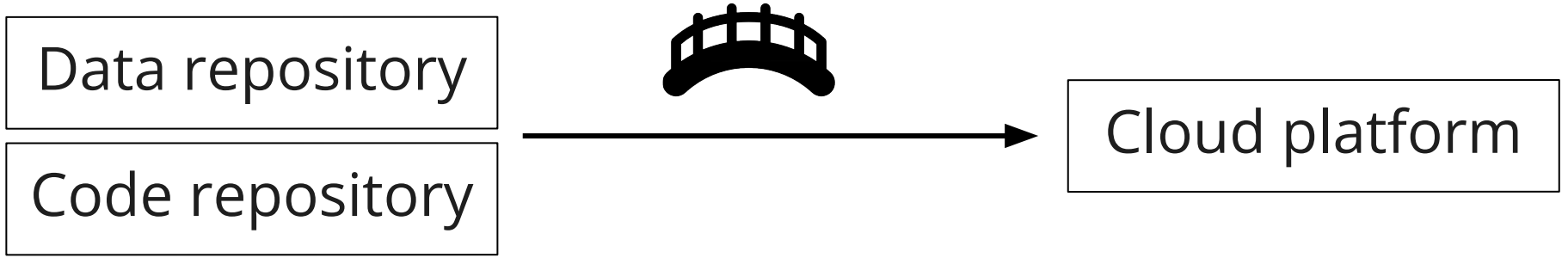

Data repository
Cloud platform

Data repository

Cloud platform



Trisovic, Ana, et al. "Advancing computational reproducibility in the Dataverse data repository platform." P-RECS'20.


Data repository
Code repository
Cloud platform

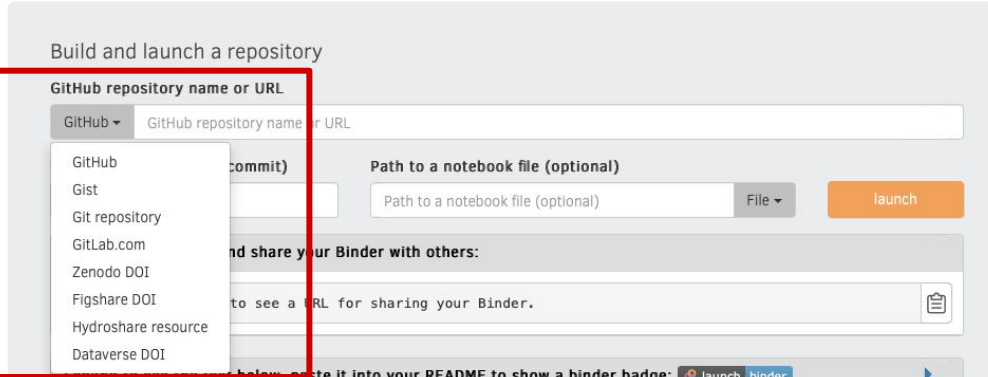
Data repository

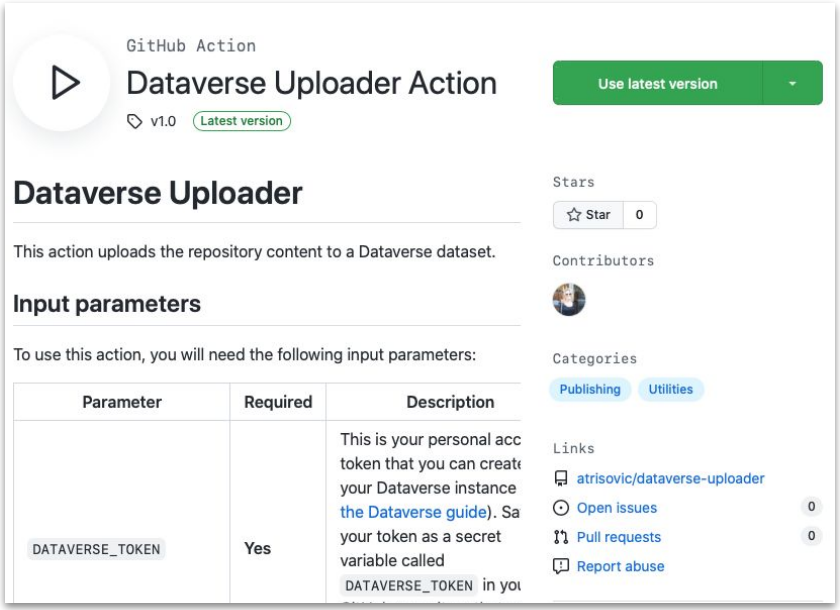
Code repository

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

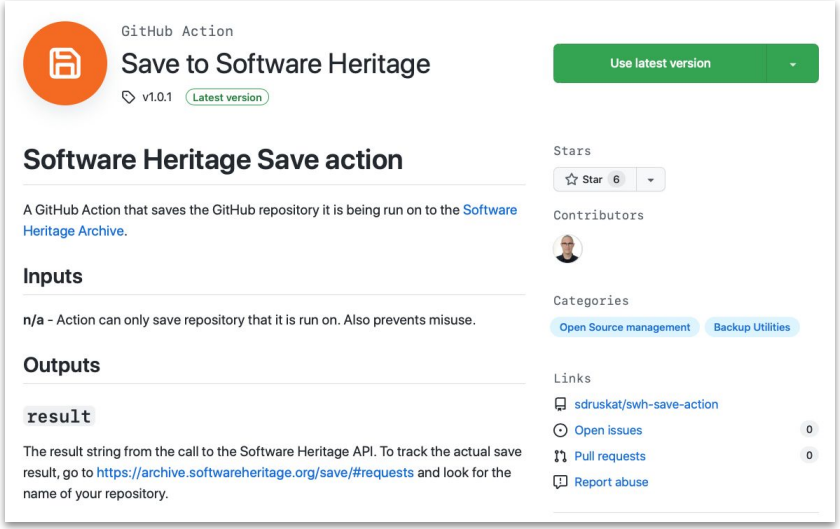
New to Binder? Get started with a Zero-to-Binder tutorial in Julia, Python or R.

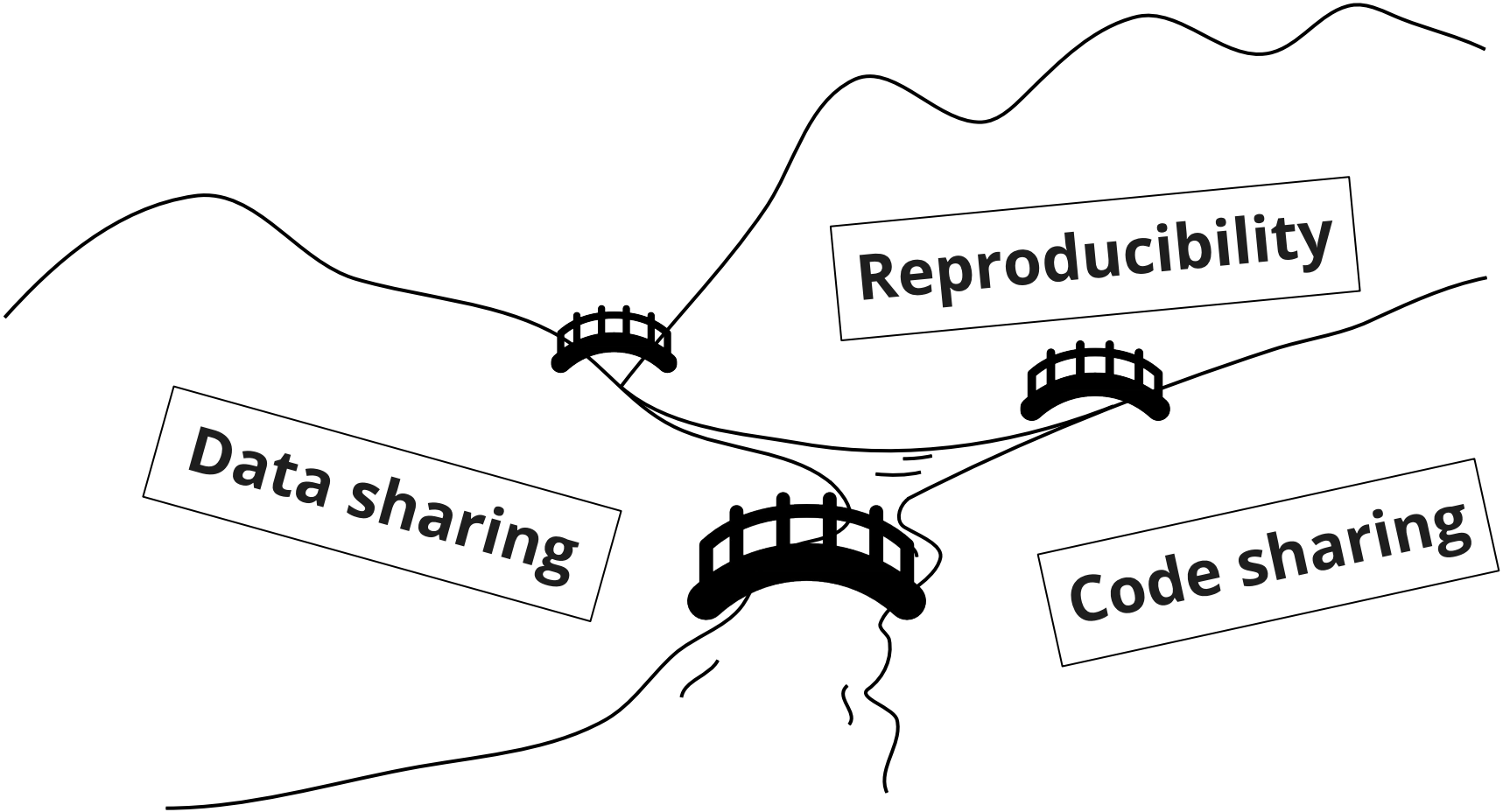


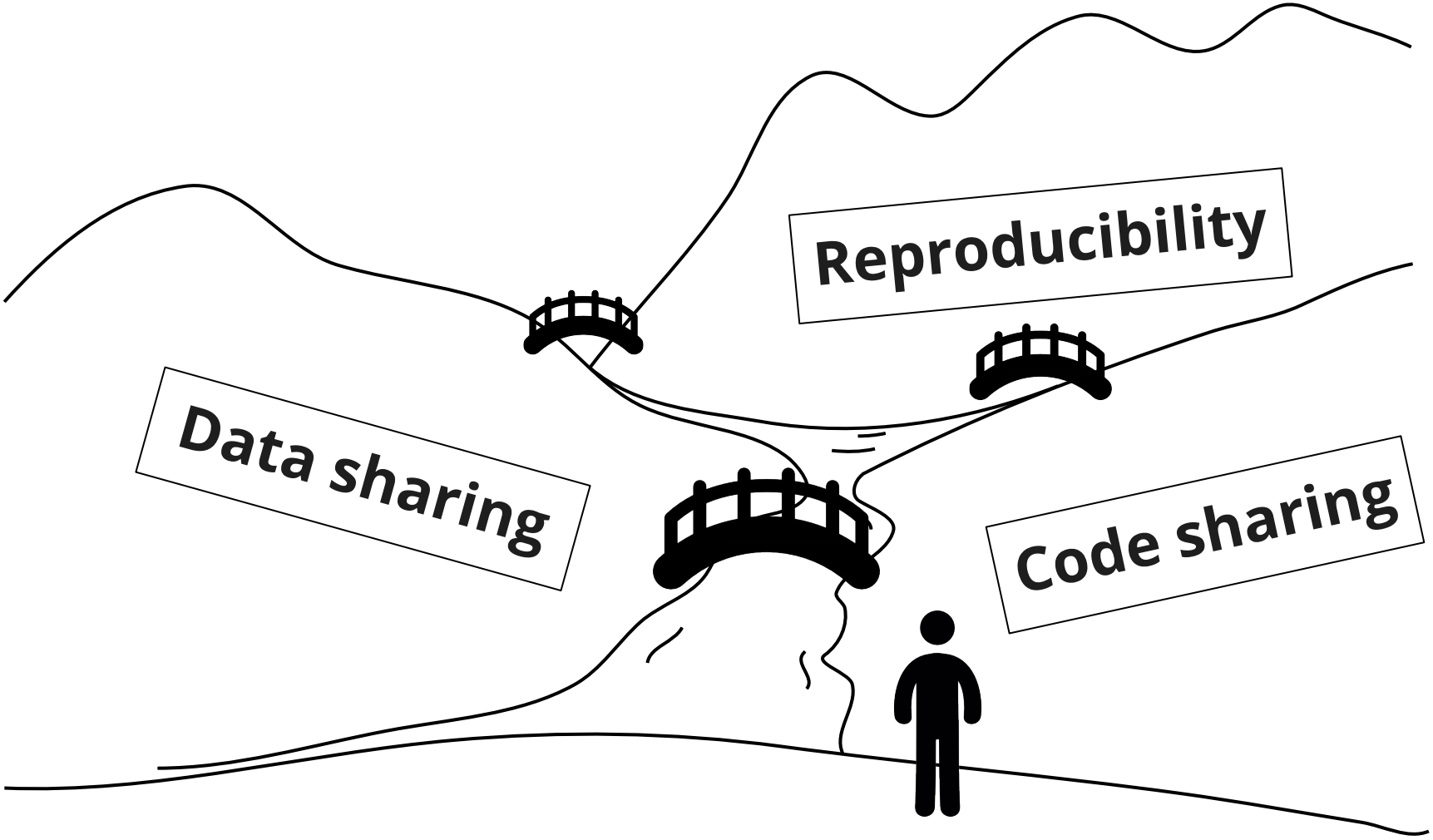

GitHub Action
Dataverse Uploader Action
v1.0
Latest version
Use latest version
Dataverse Uploader
This action uploads the repository content to a Dataverse dataset.
Input parameters
To use this action, you will need the following input parameters:
Parameter
Required
Description
DATAVERSE_TOKEN
Yes
This is your personal acc
token that you can creat
your Dataverse instance
the Dataverse guide). Sa
your token as a secret
variable called
DATAVERSE_TOKEN in yo
Stars
Star
0
Contributors
Categories
Publishing
Utilities
Links
atrisovic/dataverse-uploader
Open issues
0
Pull requests
0
Report abuse

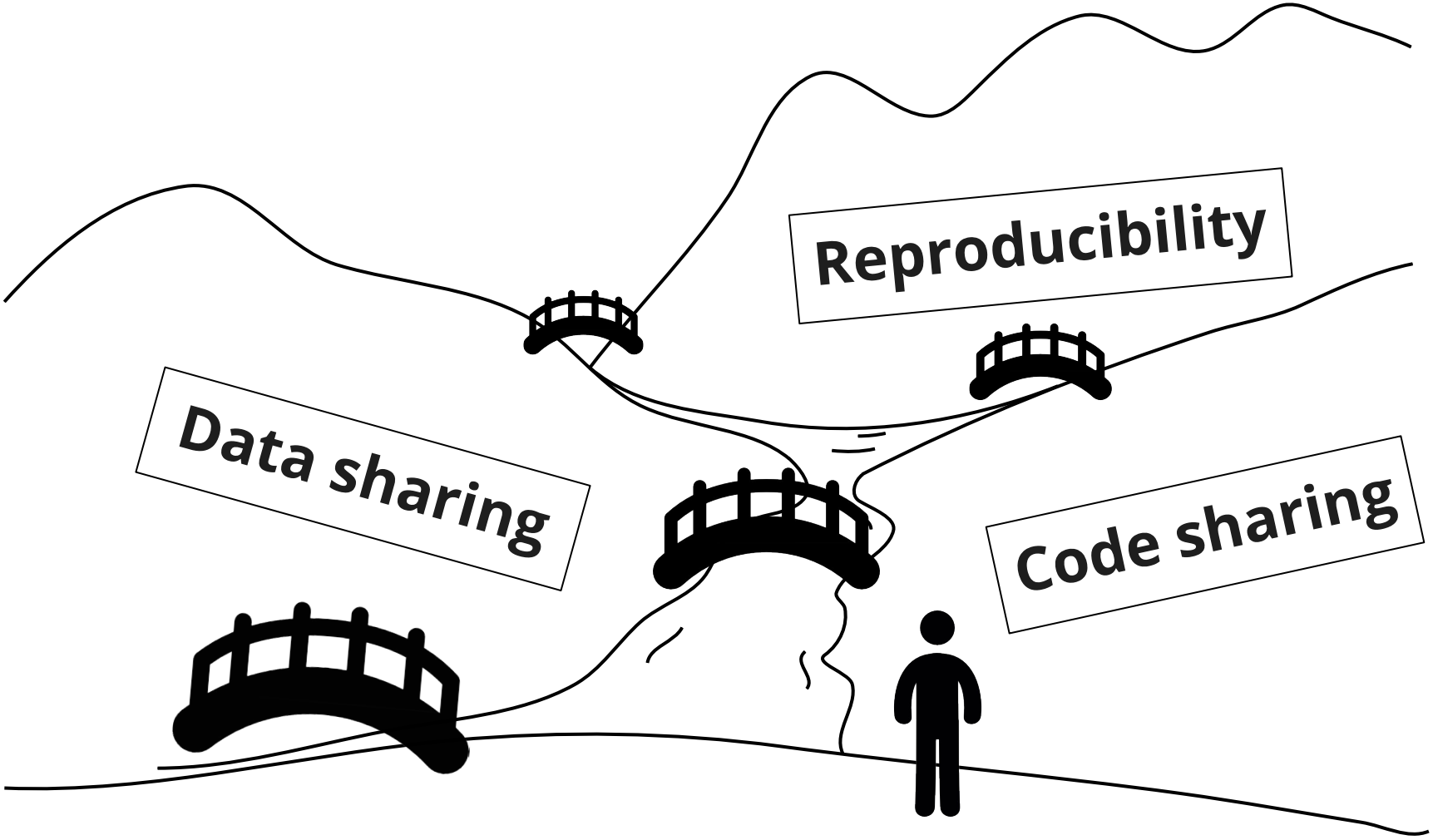
GitHub

Software Heritage



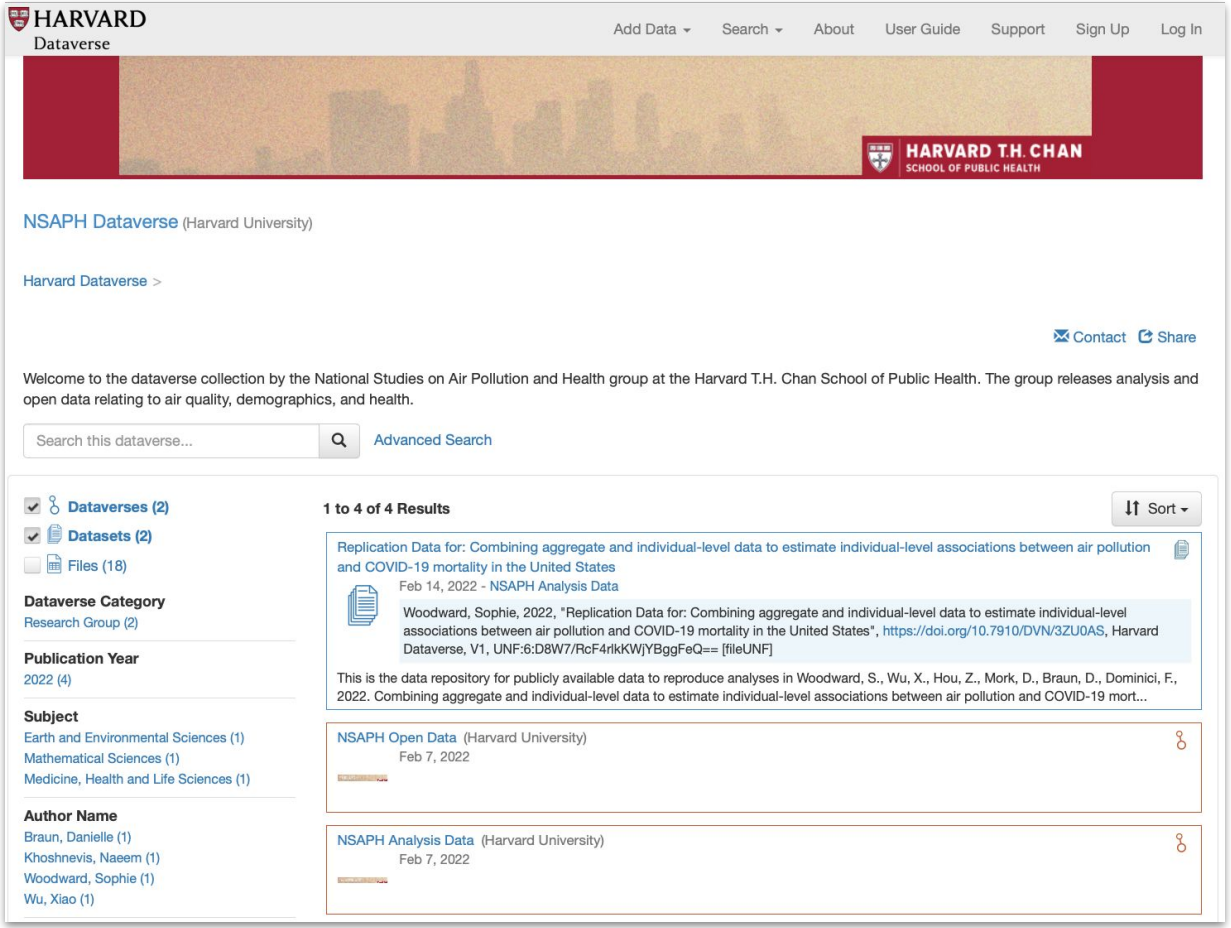

Reproducibility
Data sharing
Code sharing


Reproducibility
Data sharing
Code sharing


Reproducibility
Data sharing
Code sharing

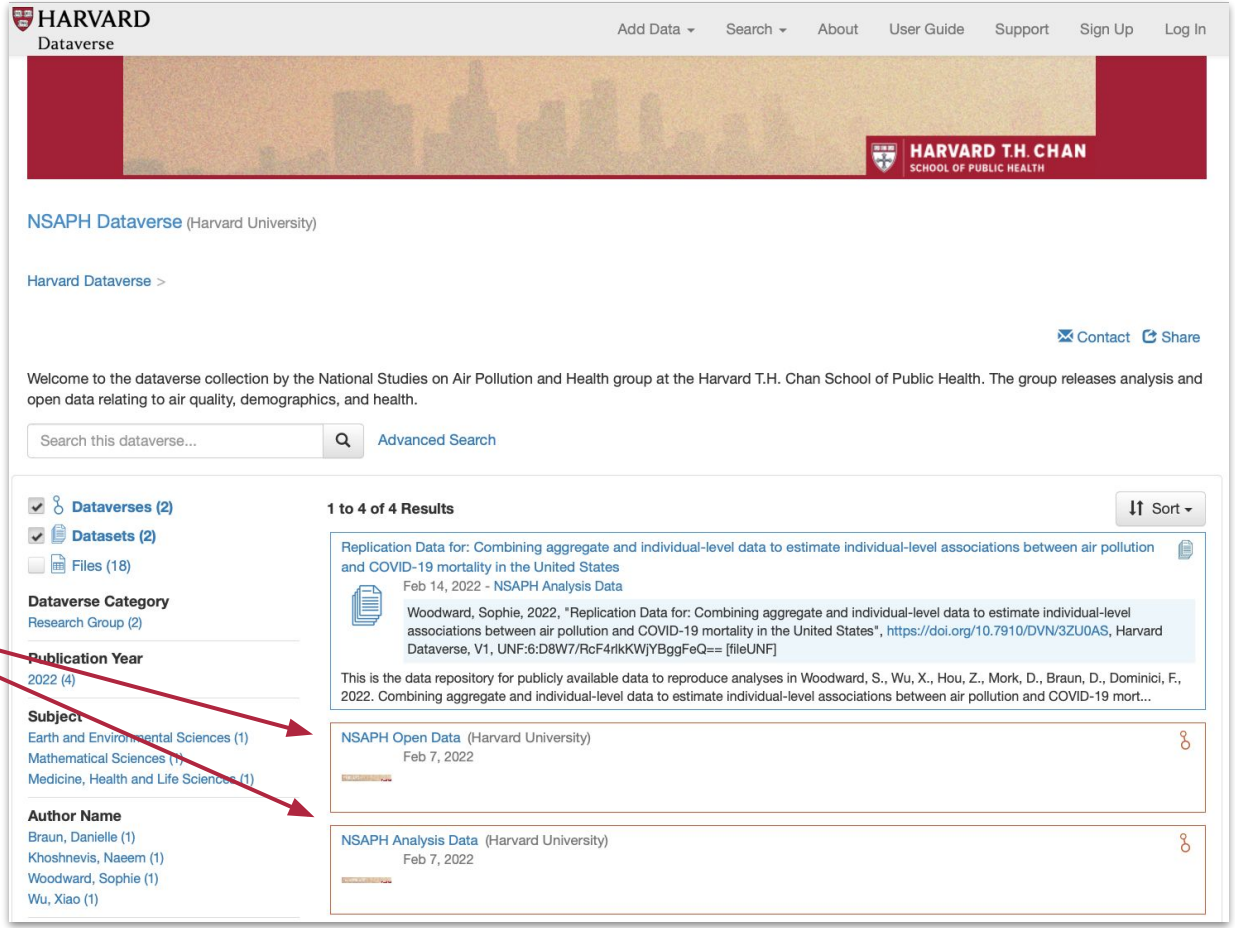
1. Data sharing

1. Data sharing



1. Data sharing

Data collections



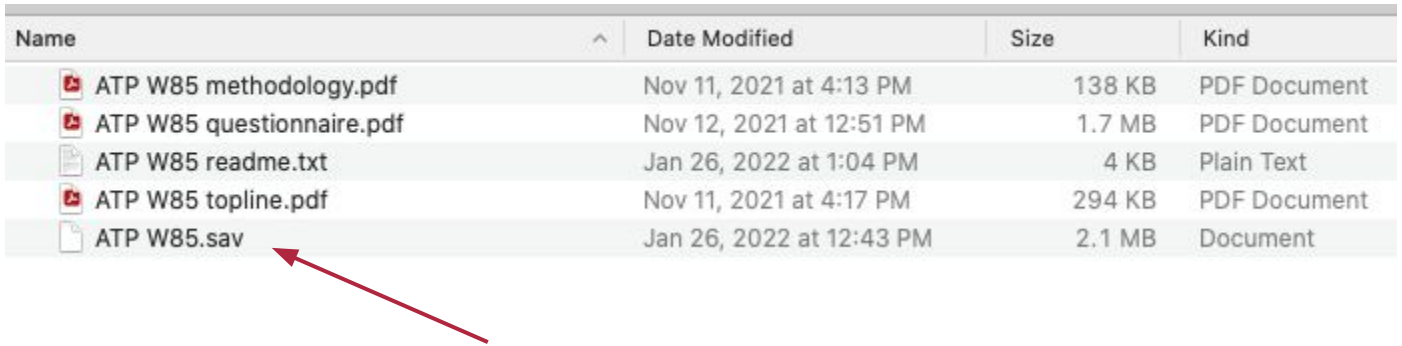
Data that does not explicitly have an open license is not open data.

1. Data sharing

- Data should be licensed
- Metadata
- It should be complete
- It should be shared in a (open) machine-readable format

1. Data sharing

- Data should be licensed
- Metadata
- It should be complete
- It should be shared in a (open) machine-readable format



Name	Date Modified	Size	Kind
ATP W85 methodology.pdf	Nov 11, 2021 at 4:13 PM	138 KB	PDF Document
ATP W85 questionnaire.pdf	Nov 12, 2021 at 12:51 PM	1.7 MB	PDF Document
ATP W85 readme.txt	Jan 26, 2022 at 1:04 PM	4 KB	Plain Text
ATP W85 topline.pdf	Nov 11, 2021 at 4:17 PM	294 KB	PDF Document
ATP W85.sav	Jan 26, 2022 at 12:43 PM	2.1 MB	Document

2. Data documentation

Data Dictionary

Fieldname	Source	Description	Role
QID	Medicare	Person's ID	ID
ADATE	Medicare	Admission date	Outcome
DDATE	Medicare	Discharge date	Outcome
zipcode_R	Medicare	Zipcode	Location
DIAG 1-10	Medicare	Billing codes as ICD codes	Outcome
AGE	Medicare	Age	Confounder
PROV_NUM	Medicare	Hospital ID	Location
ADM_SOURCE	Medicare	Admission source	Confounder
ADM_TYPE	Medicare	Admission type (1 - Emergency, 2 - Urgent, 3 - Elective)	Confounder
Dual	Medicare	Eligible for both Medicare and Medicaid (1 - yes, 0 - otherwise)	Confounder
year	R	Year of hospital admission (from 2000 to 2016)	Outcome
AD_primary	R	Does the ICD code of Alzheimer's disease appear in the first billing code? (T/F)	Outcome

2. Data documentation

Data Dictionary

Fieldname	Source	Description	Role
QID	Medicare	Person's ID	ID
ADATE	Medicare	Admission date	Outcome
DDATE	Medicare	Discharge date	Outcome
zipcode_R	Medicare	Zipcode	Location
DIAG 1-10	Medicare	Billing codes as ICD codes	Outcome
AGE	Medicare	Age	Confounder
PROV_NUM	Medicare	Hospital ID	Location
ADM_SOURCE	Medicare	Admission source	Confounder
ADM_TYPE	Medicare	Admission type (1 - Emergency, 2 - Urgent, 3 - Elective)	Confounder
Dual	Medicare	Eligible for both Medicare and Medicaid (1 - yes, 0 - otherwise)	Confounder
year	R	Year of hospital admission (from 2000 to 2016)	Outcome
AD_primary	R	Does the ICD code of Alzheimer's disease appear in the first billing code? (T/F)	Outcome

Variable Description

	ADM_SOURCE - Admission source
1	Physician referral
2	Clinic referral
3	HMO referral
4	Transfer from hospital
5	Transfer from a SNF
6	another health care facility
7	Emergency room

3. Survey data documentation

- What is the topic of your survey?
- What are the questions? What are the allowable answers?
- Special variables - such as a score?
- Survey administration
  - Who enters the data?
  - Paper survey (errors handling, interviews)
  - Online survey (skip patterns)

4. Use version control for code

5. Runtime environment capture

Pin dependencies

When you install a dependency, include its version number (depending on the language you use, the exact syntax may vary). E.g., don't just specify `numpy`, specify `numpy==1.12.0`.

`pip freeze` is a handy tool to export the exact version of every Python package in your environment in a format that can be used in `requirements.txt`.

`conda env export -n <env-name>` is the equivalent for anaconda's environment.yml file.

When exporting a conda environment, you can add the conda-forge broken channel (`conda-forge/label/broken`) as a low-priority channel in your exported `environment.yml` file in order to maximize durability. Thus, if a package is marked broken after you froze the environment, said package will still install during the Binder image build process. Only do this when you intend to truly freeze the environment.

For example (in `environment.yml`):

```
channels:
  - conda-forge
  - defaults
  - conda-forge/label/broken
```

Using Dockerfiles

Ensuring reproducibility with Dockerfiles comes with its own set of challenges. For more

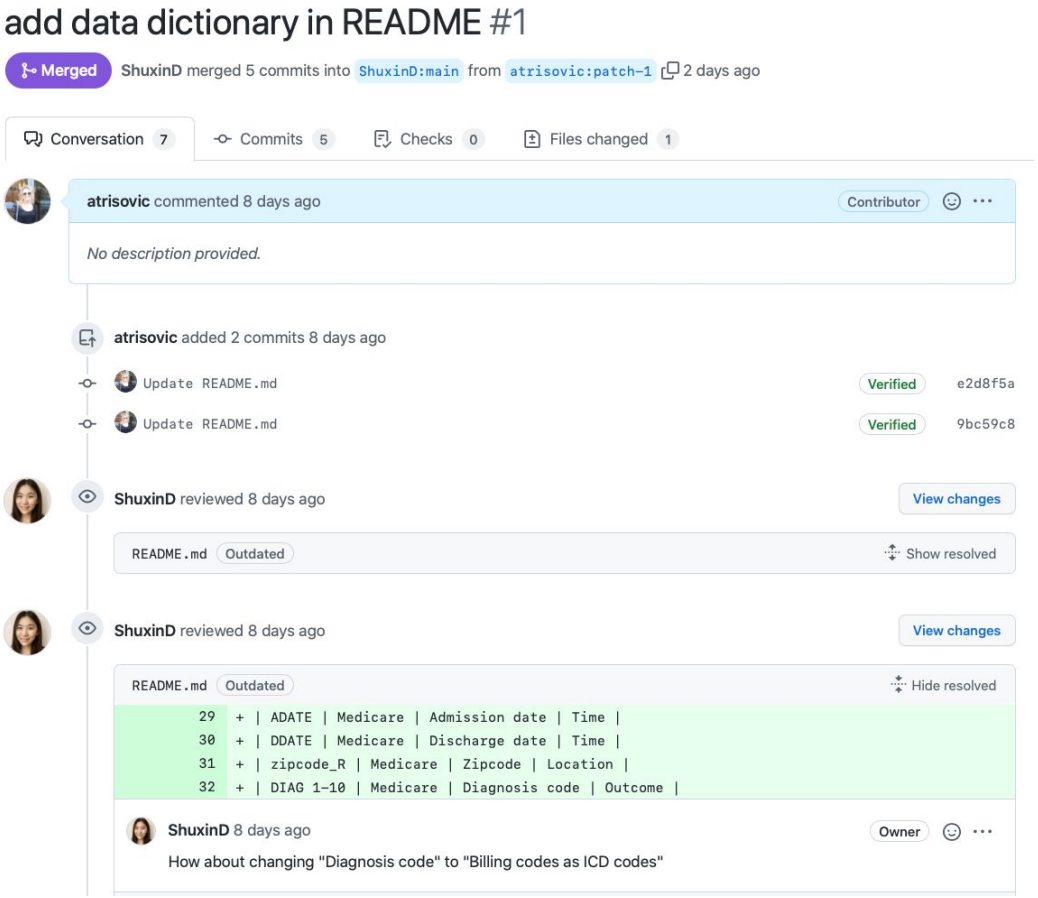
https://mybinder.readthedocs.io/en/latest/tutorials/reproducibility.html

6. Automation

Galaxy PROJECT

COMMON
WORKFLOW
LANGUAGE

7. Internal review



Recommendations

Summary

- Use data repositories for data sharing
- Document your data
- Capture dependencies of your code
- Establish internal review workflow

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