

# Responding to reproducibility challenges: from physics to social sciences

---

*Ana Trisovic*  
*IQSS, Harvard University*

# Agenda

- Reproducibility and replicability
- Preliminary results reveal that more than 75% of deposited research code (in R) is not easily re-executable
- Possible reproducibility solutions for
  - social science research
  - particle physics research



# Reproducibility and replicability

- "Reproducibility (computational) is obtaining consistent results using the same input data, computational steps, methods and code"
- "Replicability is obtaining consistent results across studies aimed at answering the same scientific questions, each of which has obtained its own data"



# Necessary conditions for reproducibility

- Research code
- Code dependencies (packages, system dependencies etc.)
- Input Data
- Research workflow (i.e., sequence of analysis steps)
- Other (computer infrastructure dependencies, OS, contextual information etc.)






# Current practice of research preservation for reproducibility







- Typically code and data are preserved, together with instructions and metadata


[Harvard Dataverse](#) > [American Journal of Political Science \(AJPS\) Dataverse](#) > **Replication Data for: Do Inheritance Customs Affect Political and Social Inequality**

[Contact](#) [Share](#)

 **Replication Data for: Do Inheritance Customs Affect Political and Social Inequality**

**Version 2.0**

<input type="checkbox"/>	 <b>Data_Main.tab</b> Tabular Data - 3.2 MB - May 6, 2019 - 22 Downloads 40 Variables, 8670 Observations - UNF:6:HvOhg+eZddd6KDnYhWBF3A==	<a href="#">Explore</a> <a href="#">Download</a>
<input type="checkbox"/>	 <b>GESIS_Councils.tab</b> Tabular Data - 2.7 MB - May 6, 2019 - 16 Downloads 23 Variables, 13312 Observations - UNF:6:WfiaoEC/BiTEBQzL41pJPg==	<a href="#">Explore</a> <a href="#">Download</a>
<input type="checkbox"/>	 <b>Instrument_Analysis.R</b> R Syntax - 18.3 KB - May 6, 2019 - 10 Downloads MD5: 47a4655ef08cdb25a7cee290bf77e543	<a href="#">Download</a>
<input type="checkbox"/>	 <b>IPEHD_1892.tab</b> Tabular Data - 25.2 KB - May 6, 2019 - 15 Downloads 14 Variables, 272 Observations - UNF:6:PilUC/9fQQlhNgdP5vewPw==	<a href="#">Explore</a> <a href="#">Download</a>
<input type="checkbox"/>	 <b>IPEHD_1901.tab</b> Tabular Data - 25.8 KB - May 6, 2019 - 14 Downloads 14 Variables, 281 Observations - UNF:6:jgPS5ou2I2PfGbWyCI6tnA==	<a href="#">Explore</a> <a href="#">Download</a>
<input type="checkbox"/>	 <b>LandGini.tab</b> Tabular Data - 27.2 KB - May 6, 2019 - 15 Downloads 14 Variables, 234 Observations - UNF:6:OeLIGqIU4vjO0Man6efFsg==	<a href="#">Explore</a> <a href="#">Download</a>



# A reproducibility study

- **Goal:** find success rate of easily re-executable research code that uses R programming language and is publicly available
- Mimics the situation in which a third-party researcher downloads a dataset, installs the used packages and tries to rerun the code
- R programming language
  - free software environment for statistical computing and graphics
  - widely used among statisticians, data scientists and academics



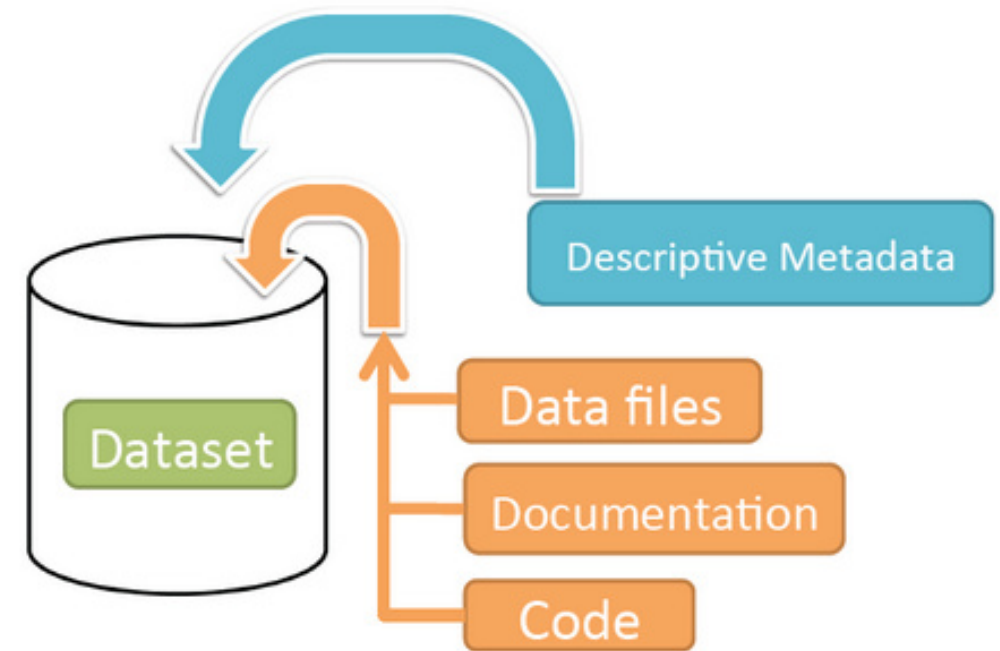


**a free and open-source web application to  
preserve, share, explore, analyse research data.  
48 institutions around the globe run Dataverse  
instances as their official data repository.**



# Implementation

- One job executes one Dataverse dataset
  - A dataset contains:  
code, data, documentation
- Allocated time to run one job is **1 hour**
  - Some jobs ran for 17 hours or longer
- The study is implemented with **AWS Batch**
- **3685 R files** from **1568 Dataverse datasets** were studied



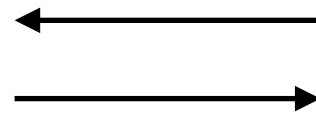
# **DOI**

**(Digital Object Identifier) is a globally unique persistent identifier that references digital objects.**



# Implementation

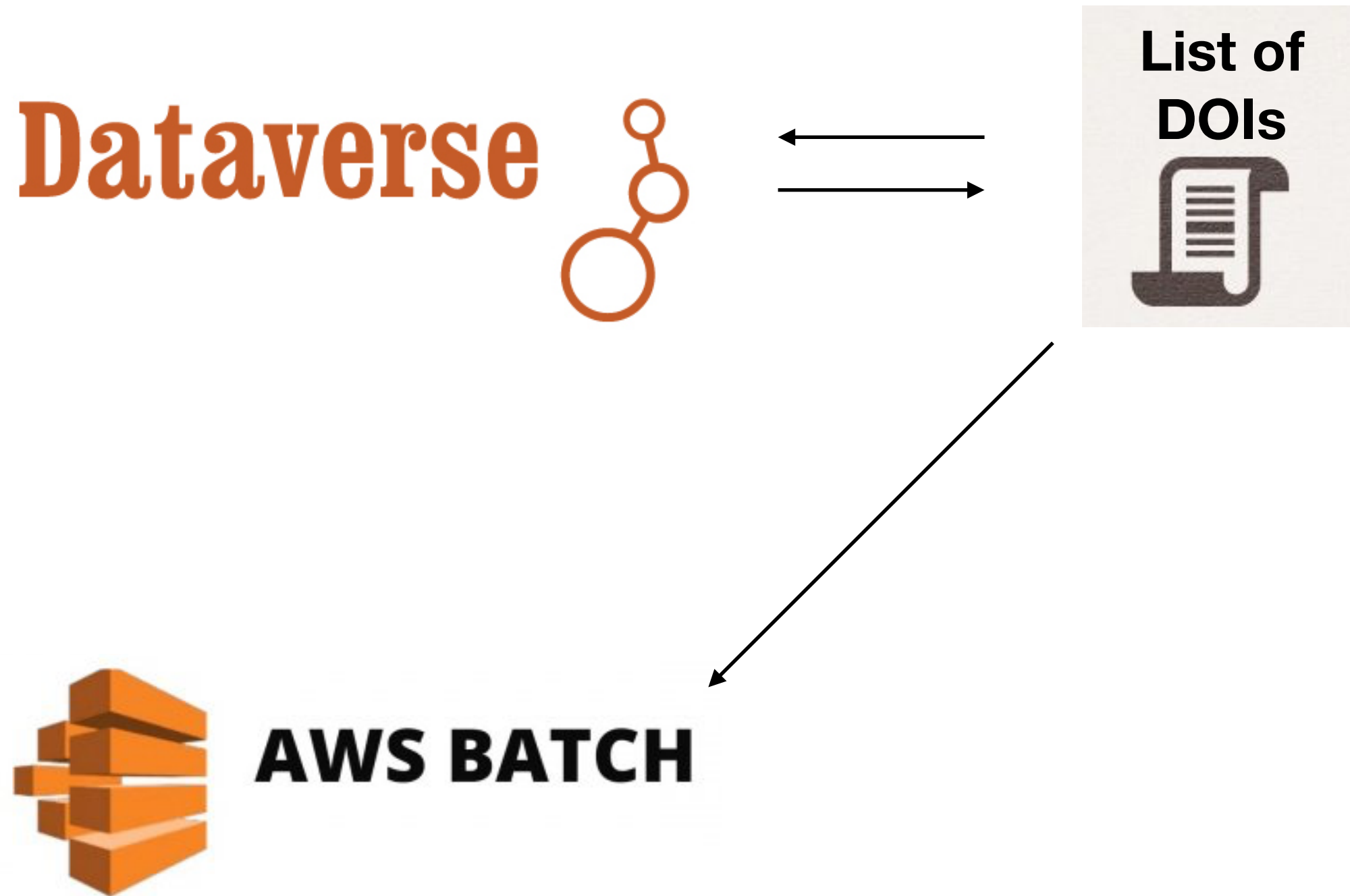
**Dataverse**



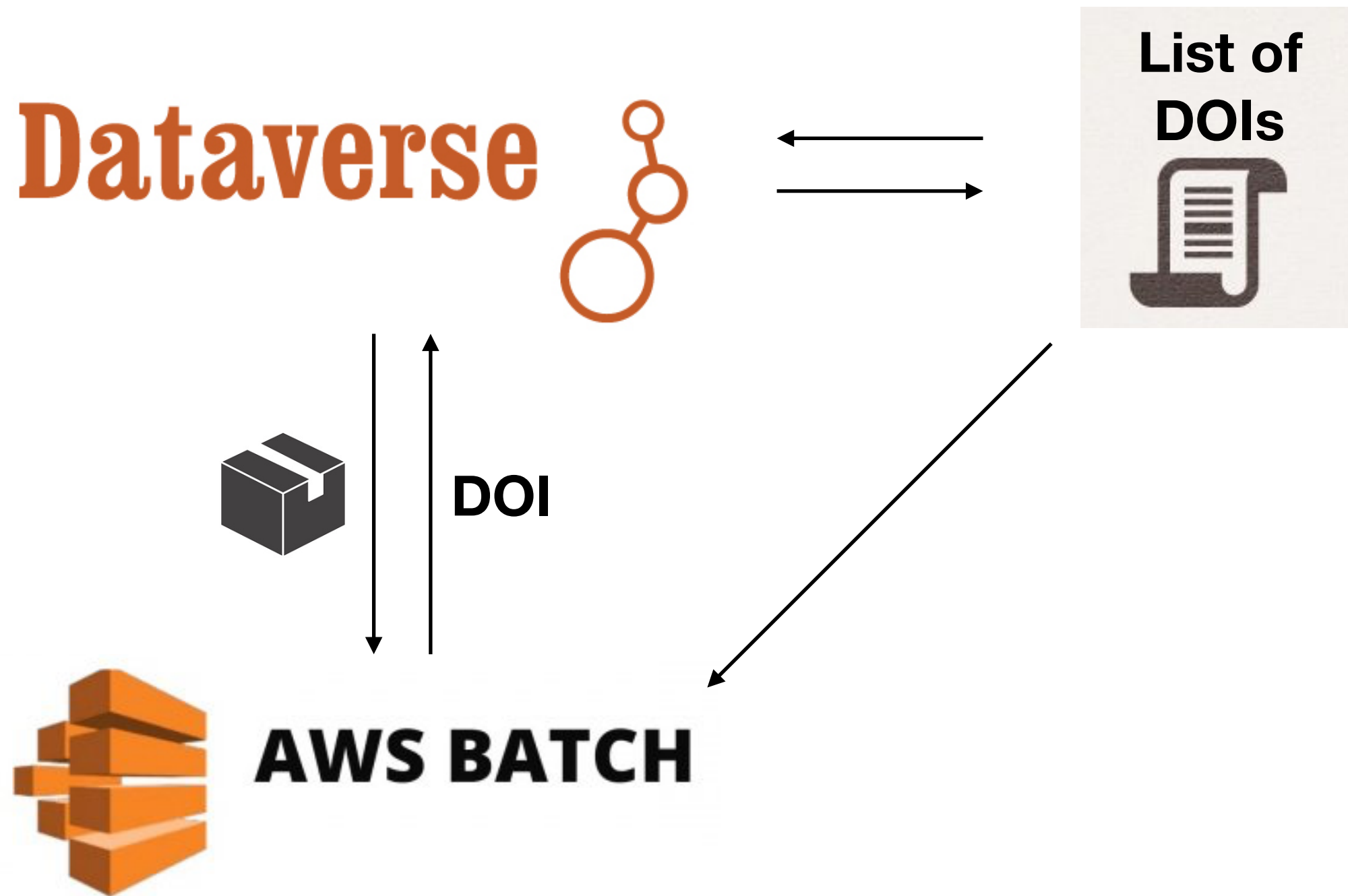
**List of  
DOIs**



# Implementation

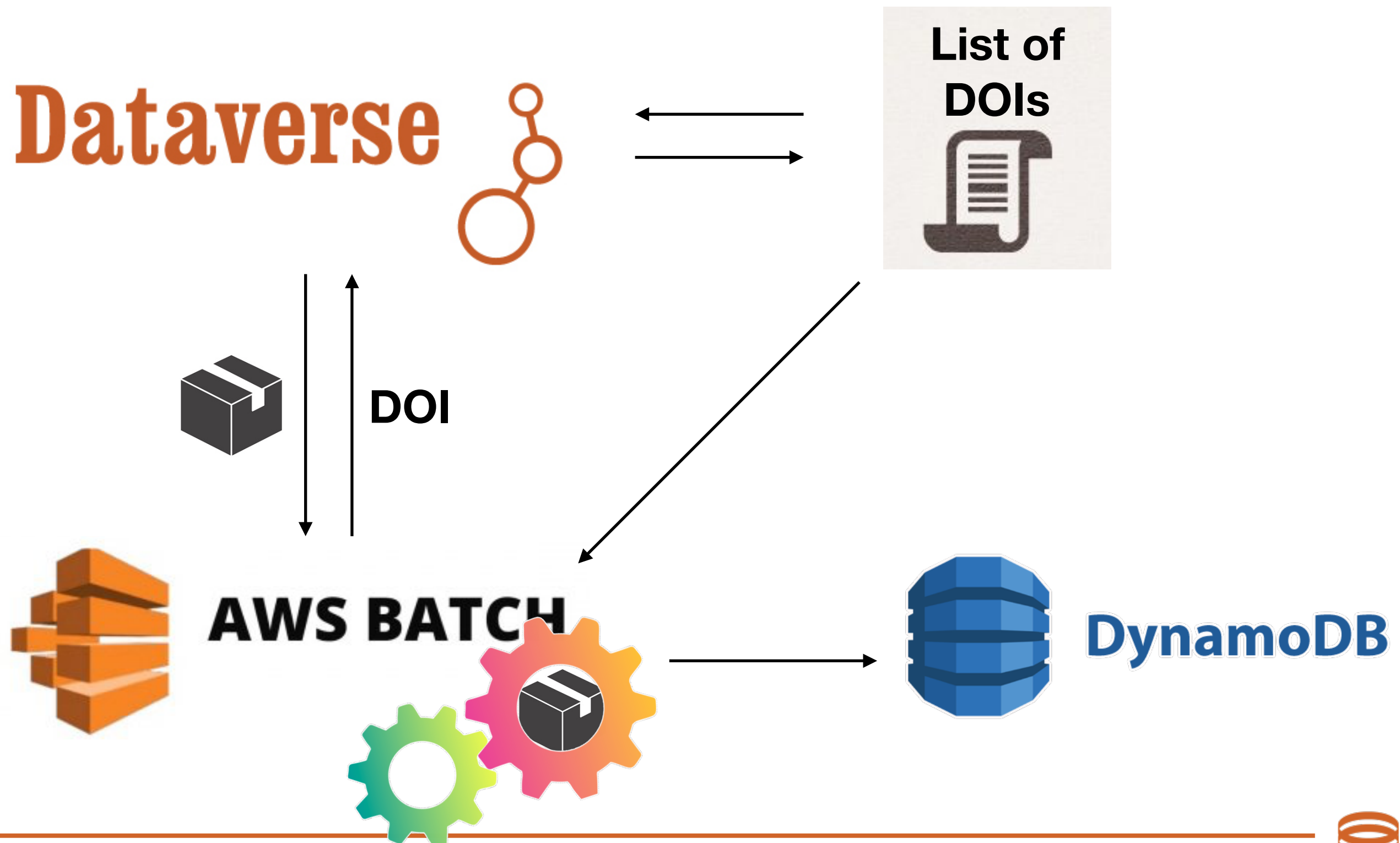


# Implementation





# Implementation



# Results

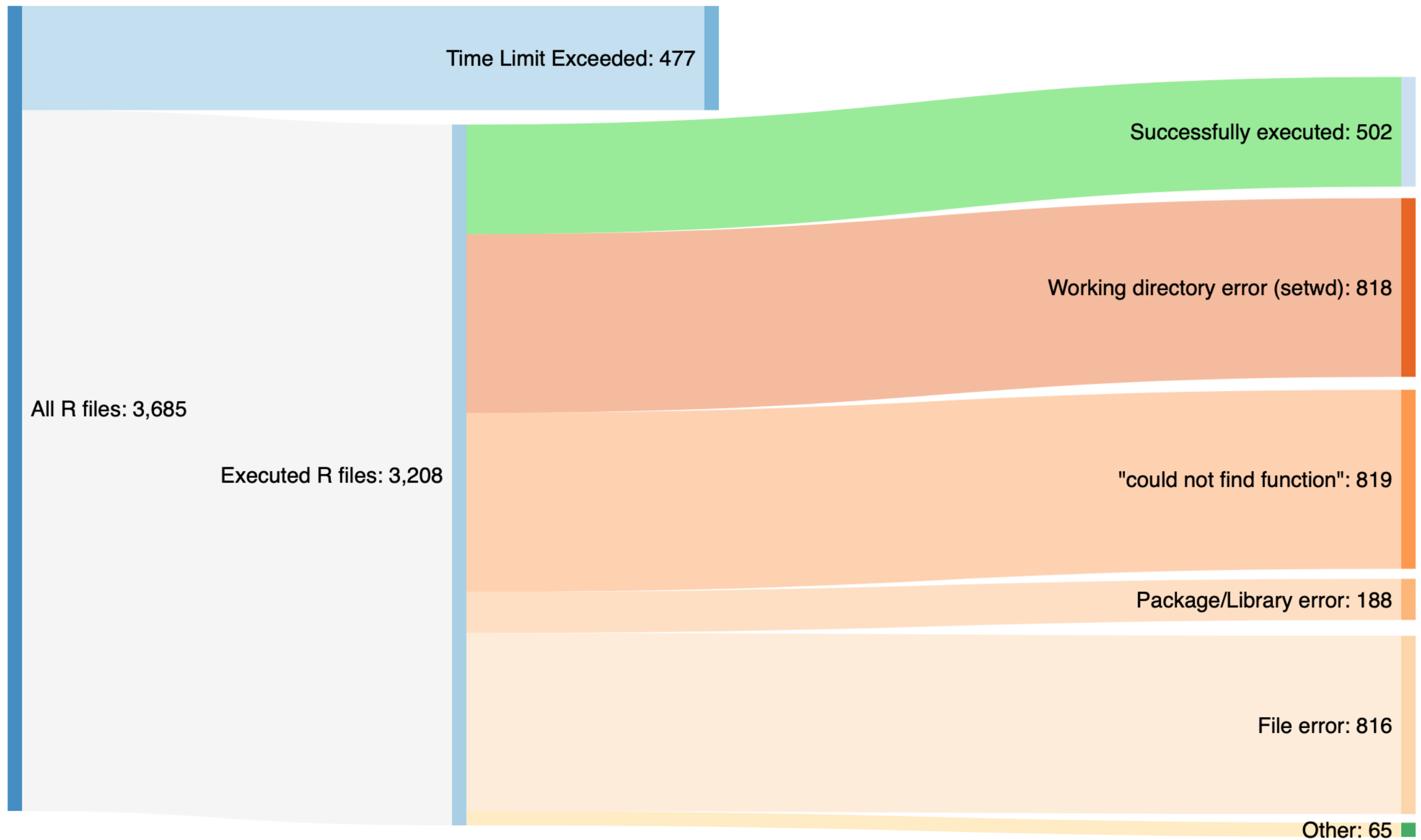
Time Limit Exceeded: 477

All R files: 3,685

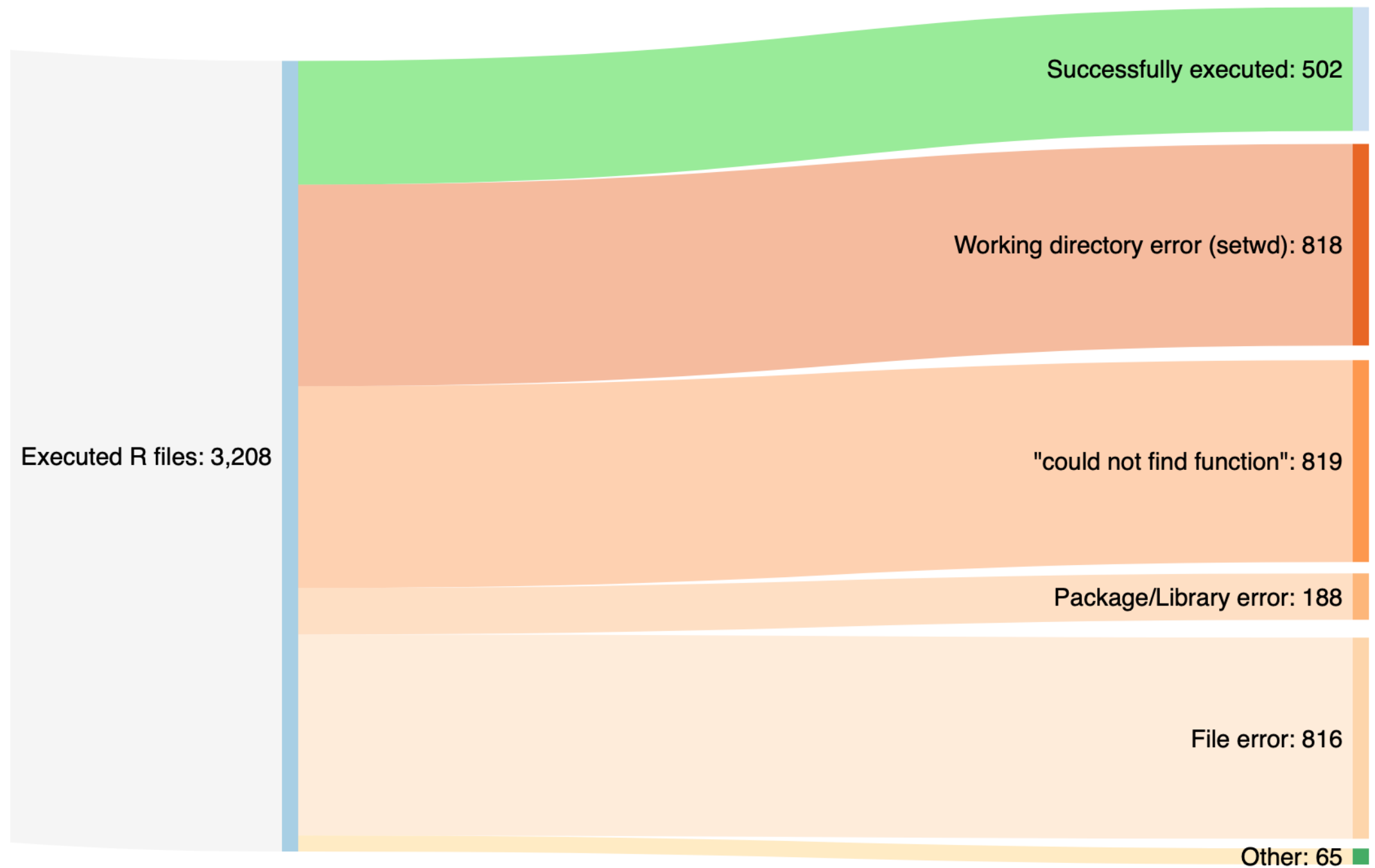
Executed R files: 3,208



# Results

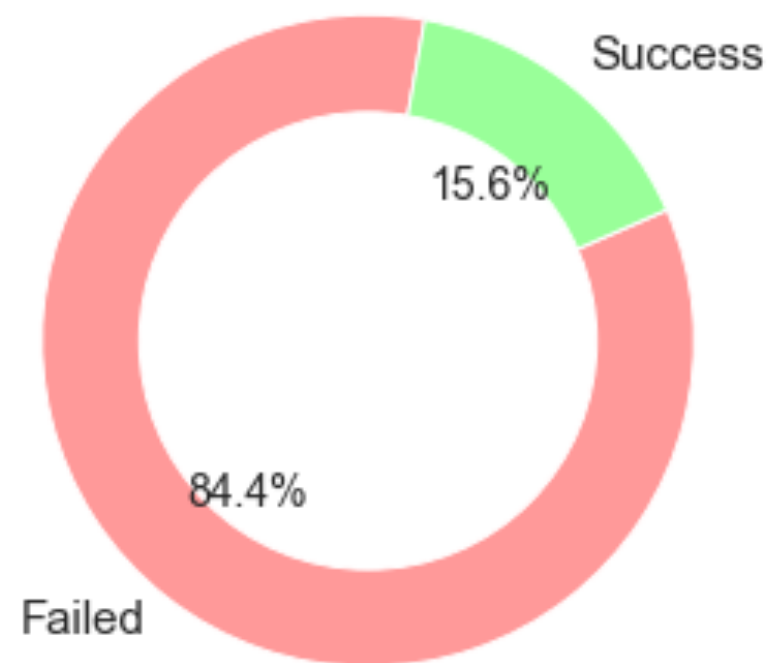


# Results



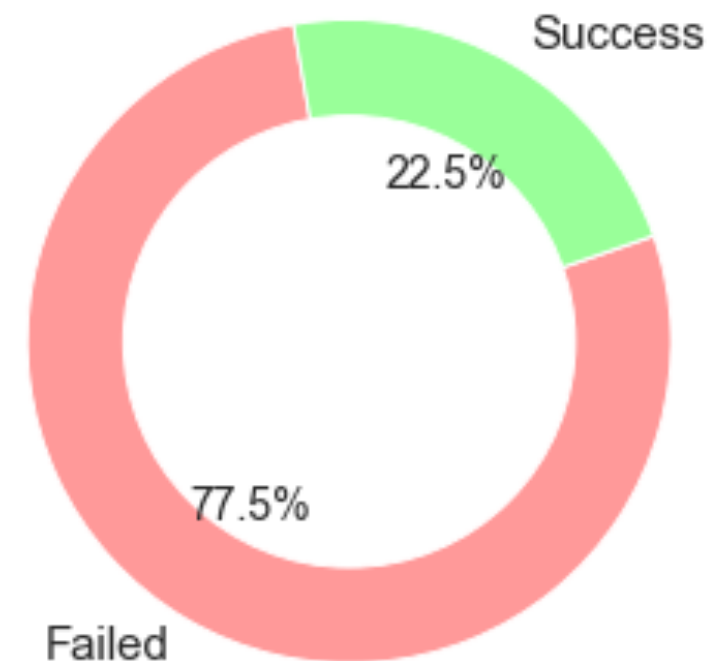
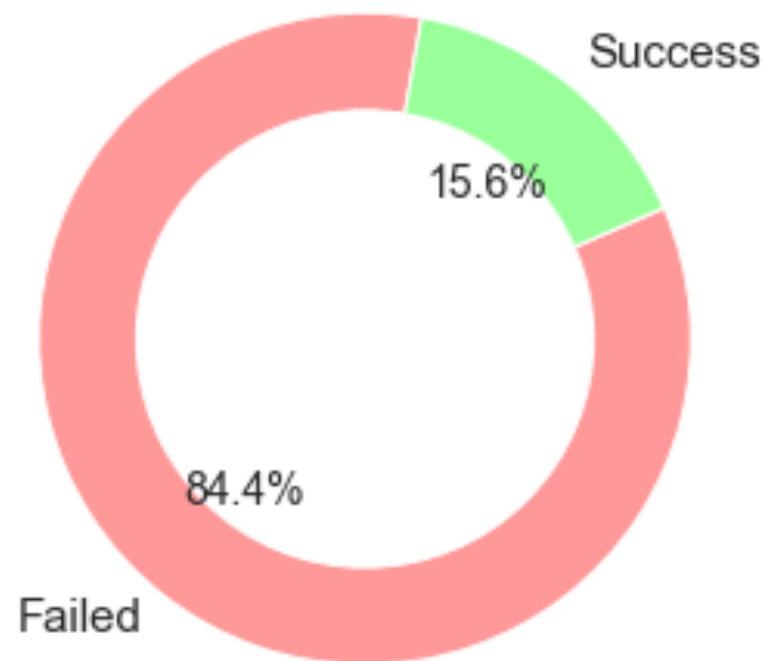
# Re-execution rate

- Re-execution rate of 3208 R files

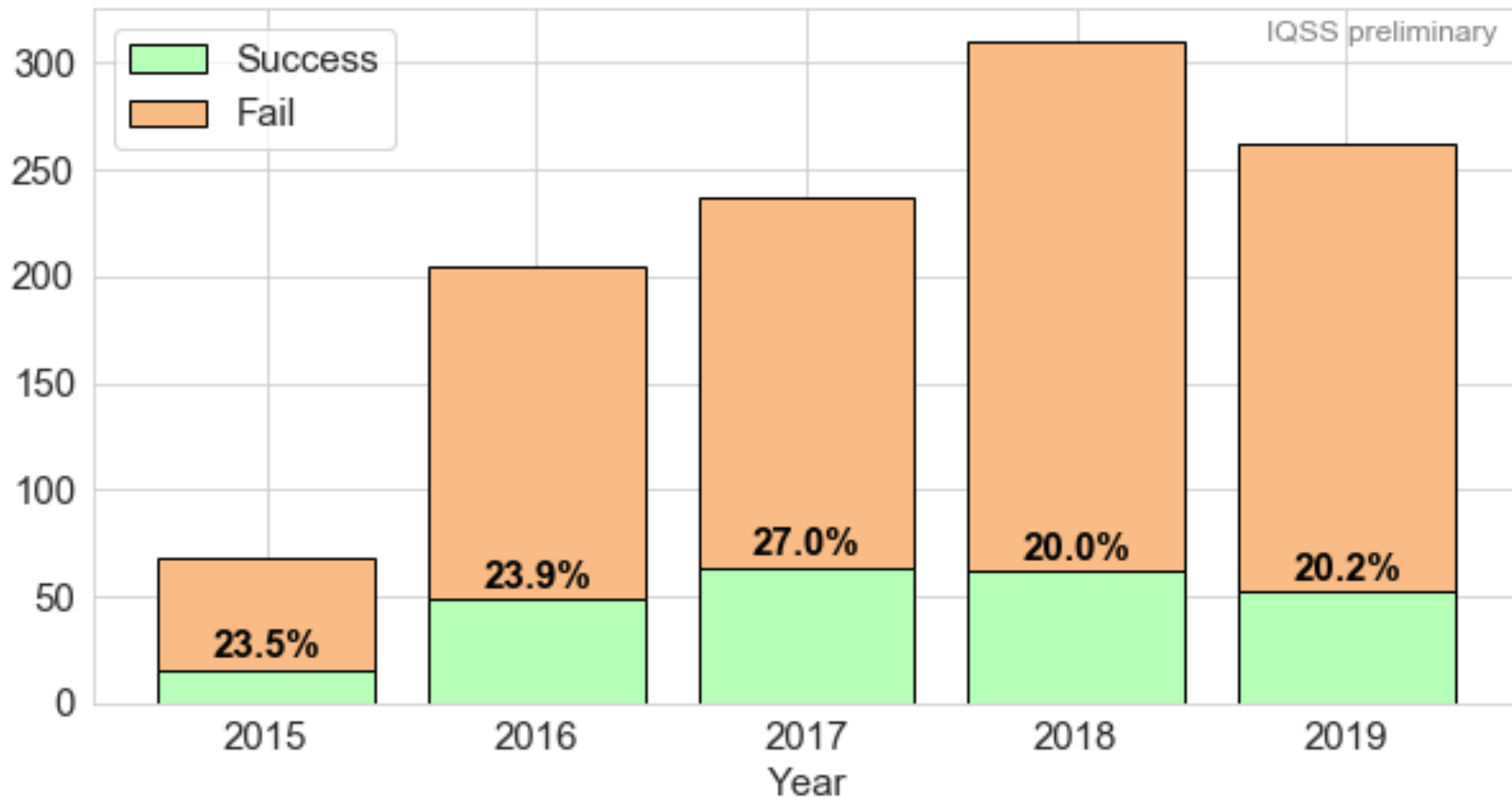


# Re-execution rate

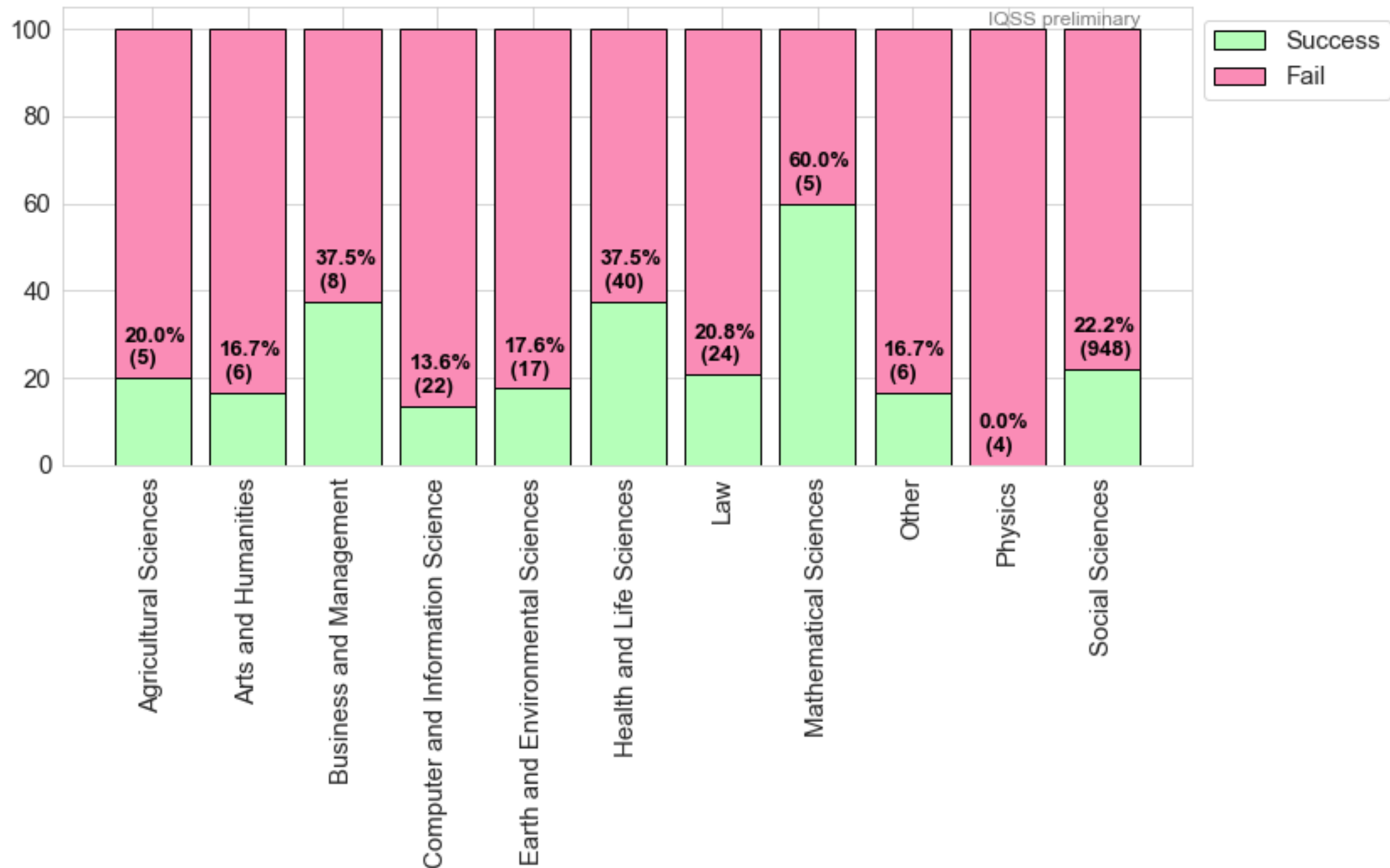
- Re-execution rate of 3208 R files
- Re-execution rate of 1091 Dataverse datasets



# Re-execution rate per year of publishing

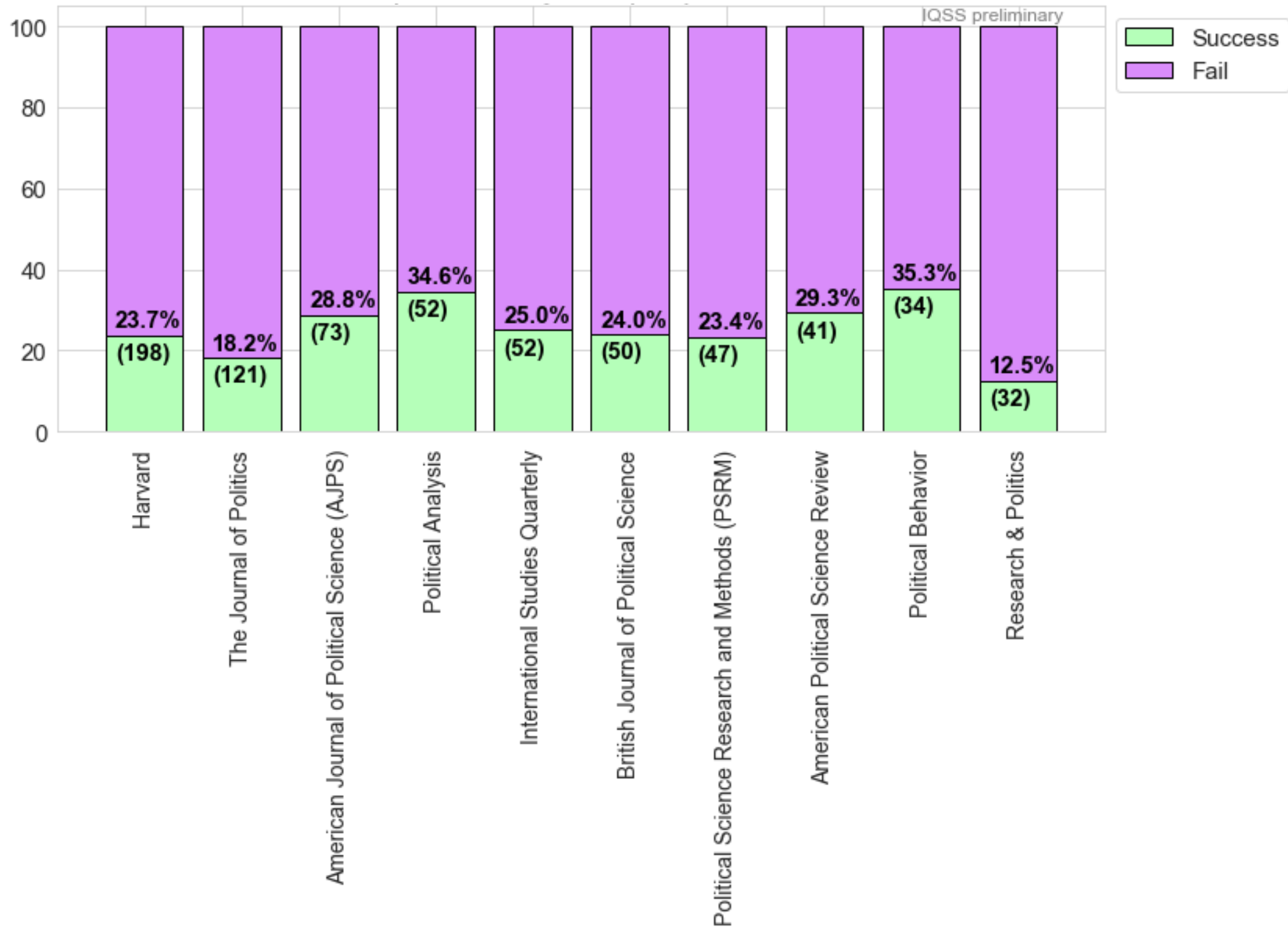


# Re-execution rate per field of study





# Re-execution rate per publisher



# Discussion

- Limitations of the language
  - R is not backward compatible

```
## requires JAGS version 2.1.0, rjags version 2.1.0-2, and R2jags version 0.02-07  
## it is incompatible with newer versions of JAGS e.g. 2.2 and 3.x
```

```
Error in setwd(~/.Dropbox/Pollution and Public Perceptions in China/Data and Analysis/Data  
Error in setwd(~/.Dropbox/Current projects/1953/Replication archive) : [newline] cannot c  
Error in setwd(C:/Users/████████/Desktop/PA replic) : [newline] cannot change working dire  
Error in setwd(/Users/████████/Documents/Papers/Representation of Party Preferences  
Error in setwd(~replication) : cannot change working directory  
...
```



# Discussion

- Limitations of the language
  - R is not backward compatible

```
## requires JAGS version 2.1.0, rjags version 2.1.0-2, and R2jags version 0.02-07  
## it is incompatible with newer versions of JAGS e.g. 2.2 and 3.x
```

- `setwd` considered a good practice in R

```
Error in setwd(~/.Dropbox/Pollution and Public Perceptions in China/Data and Analysis/Data  
Error in setwd(~/.Dropbox/Current projects/1953/Replication archive) : [newline] cannot c  
Error in setwd(C:/Users/████████/Desktop/PA replic) : [newline] cannot change working dire  
Error in setwd(/Users/████████/Documents/Papers/Representation of Party Preferences  
Error in setwd(~replication) : cannot change working directory  
...
```

- More sophisticated analysis



## Mini summary:

The way in which we currently deposit research code in data repositories is not adequate to guarantee its re-execution. On average, only about 15% of the R code files can be easily re-executed.



# Characterising social science research

- Typically small research groups
- Use both *survey data* and *big data*
- Often use sensitive data and proprietary software
- Workflows include the use of:
  - Dropbox
  - git
  - sometimes continuous integration



# What went wrong?

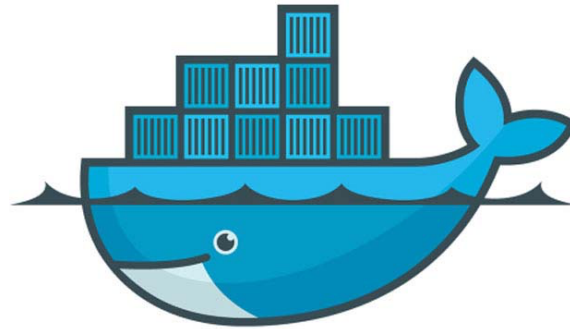
- There is a culture for sharing research materials
  - Often imposed by top journals and funding bodies
- Computational and system dependencies are often overlooked



# Possible solution

- Container-based reproducibility platforms
  - Allow seamless use of the container technology
  - Containers capture necessary system dependencies and can vastly help with reproducibility
- Examples:
  - Whole Tale
  - Code Ocean (proprietary software)
  - Binder (Jupyter)





**Docker containers provide a OS-level virtualisation that bundle software, libraries and configuration files in an isolated environment. They are typically more lightweight than virtual machines.**





# The Whole Tale project



- A representation of the entire research activity: data, software, workflows, provenance into a single executable package
- External data sources: Globus, DataONE, Dataverse
- <http://wholetale.org>





## Browse Tales

Launch to add to Launched Tales list



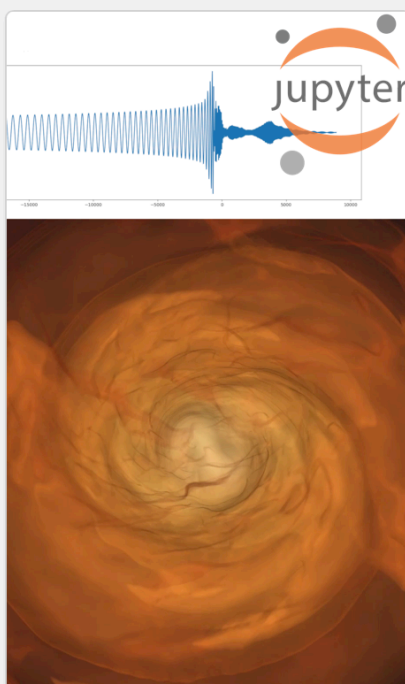
Search tales...



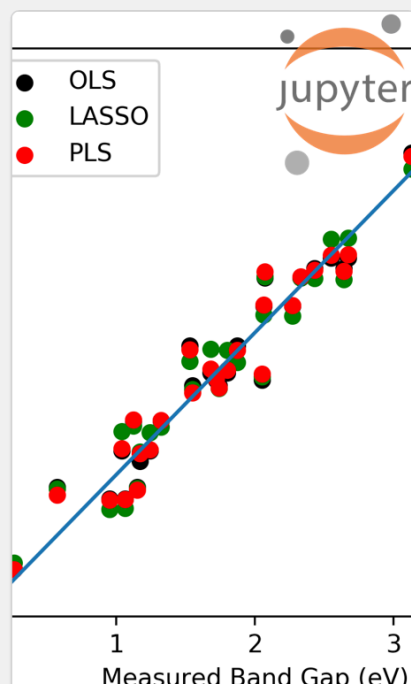
All



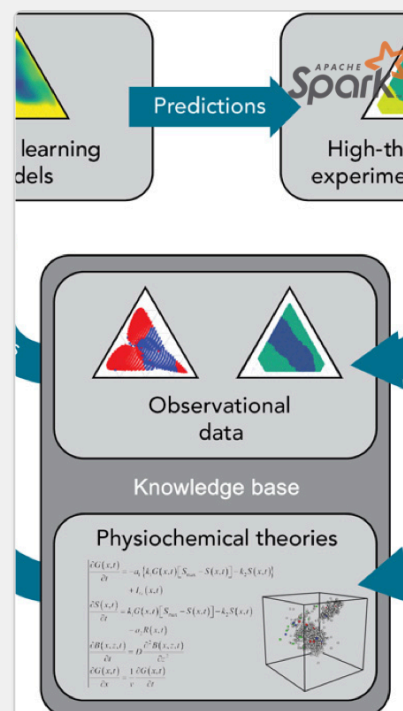
Switch to list view



SCIENCE  
LIGO Tutorial  
LIGO Detected



SCIENCE  
Informatics-aided  
bandgap engineeri...



SCIENCE  
Accelerated discovery  
of metallic g...

## Launched Tales



Predicting the Properties of Inorga...

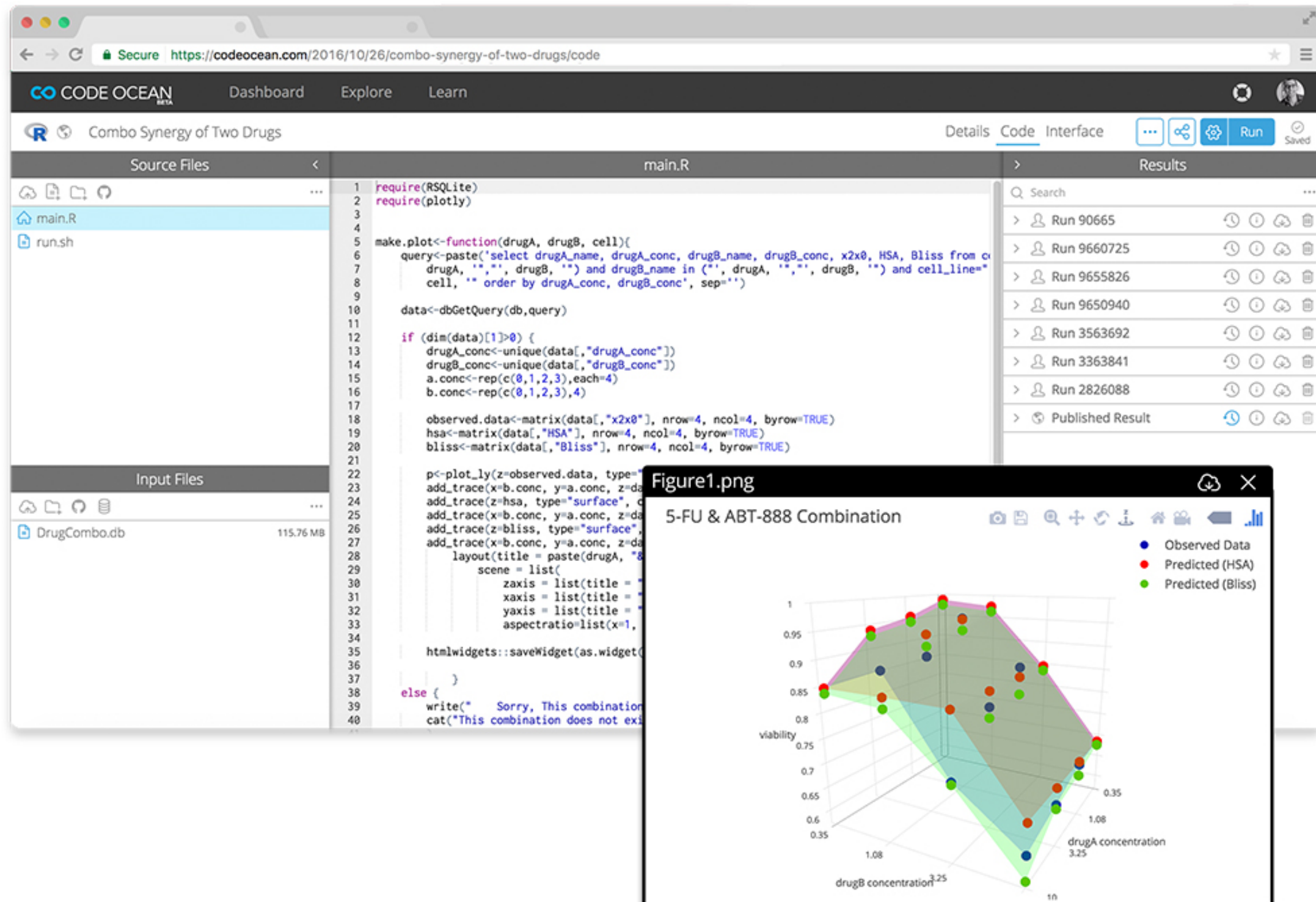


# Code Ocean

- Research collaboration platform integrated into a web browser
- Encourages users to define research workflow
- Support for a number of programming languages:
  - python
  - R
  - stata
  - MATLAB



# Code Ocean



# Binder



Turn a GitHub 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.

Build and launch a repository

GitHub repo or URL

Git branch, tag, or commit

Path to a notebook file (optional)

File ▼

launch



Already built!

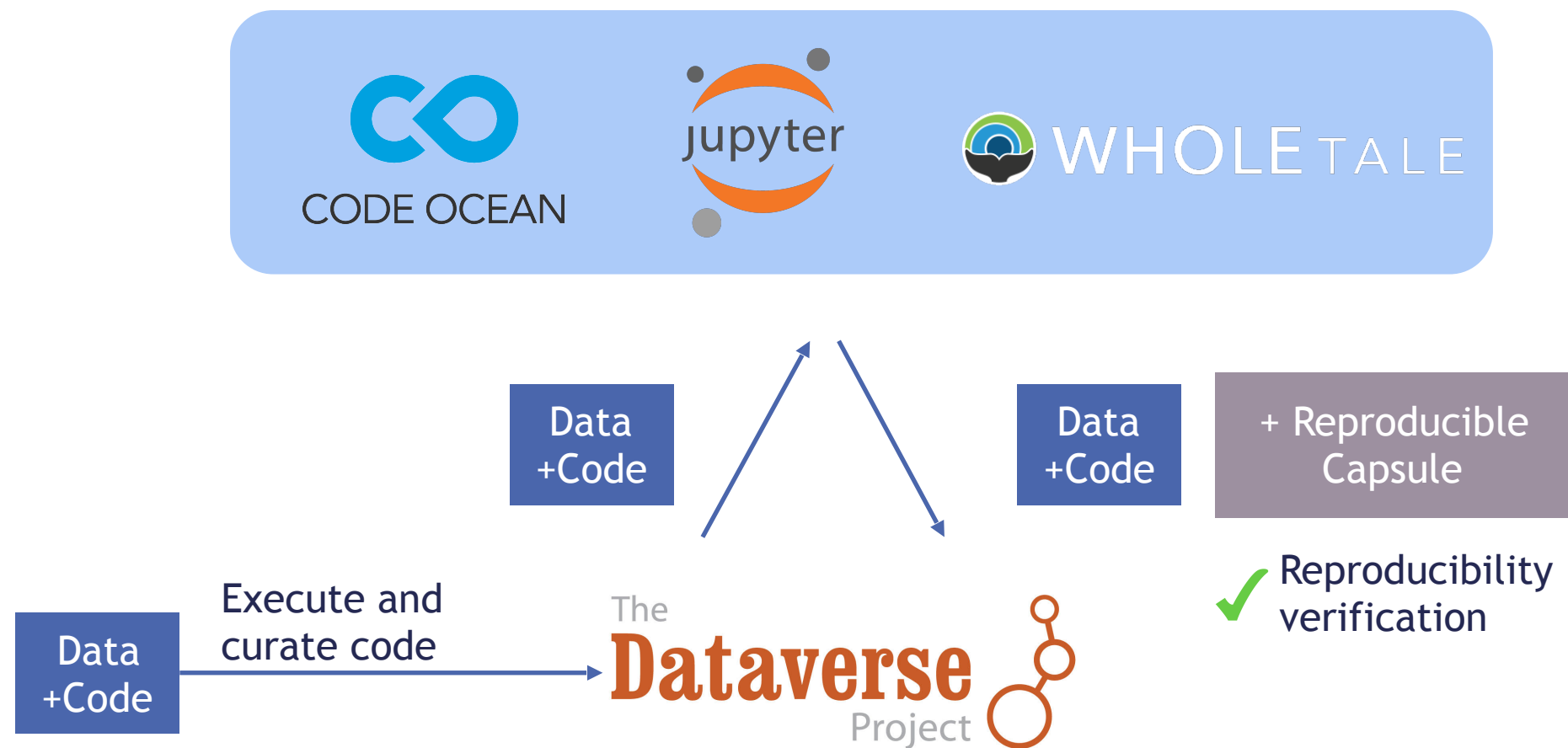
Launching



# Functionality of reproducibility platforms

- Interactive environments
- Provide versioning and provenance tools
- Support for verification, review, licensing
- Assigning permissions (for accessing proprietary data)
- Support cloning and republishing

# Integration with Dataverse



## **Mini summary:**

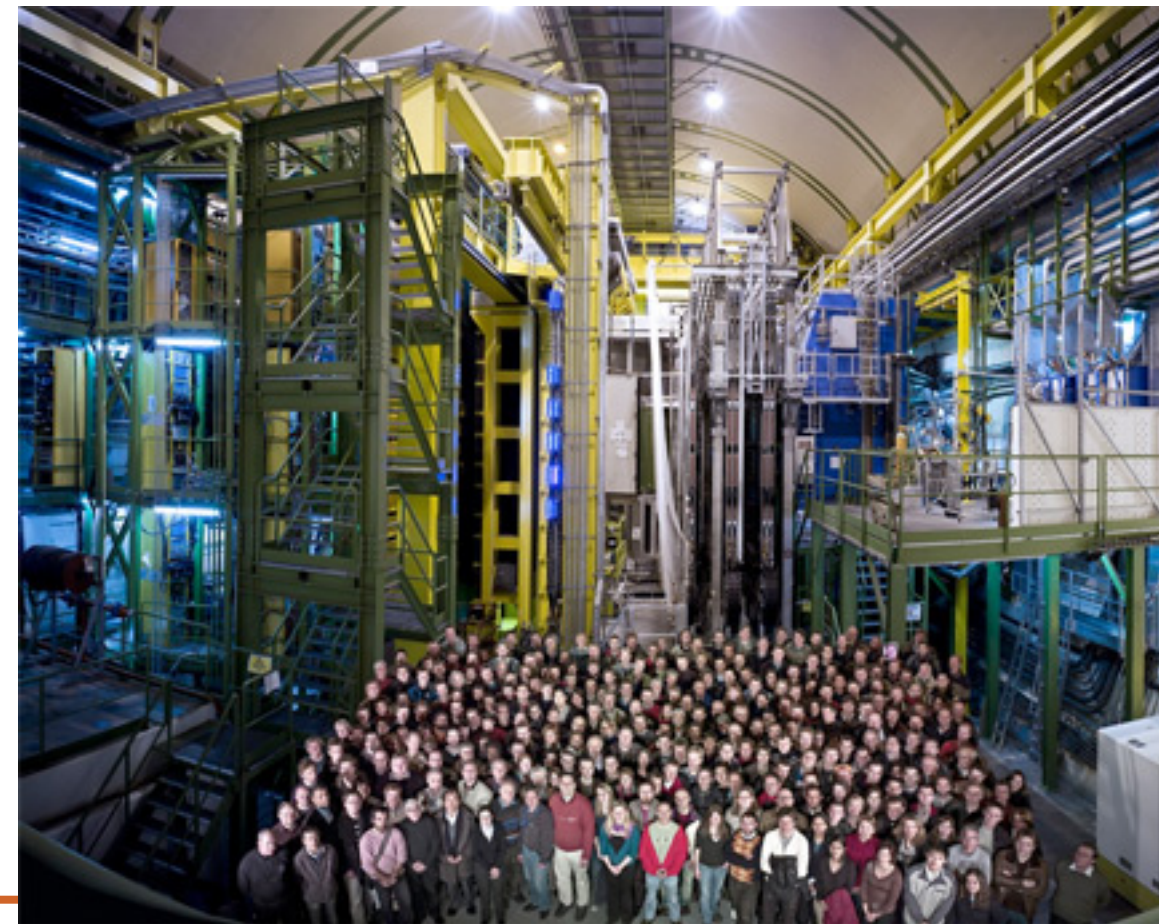
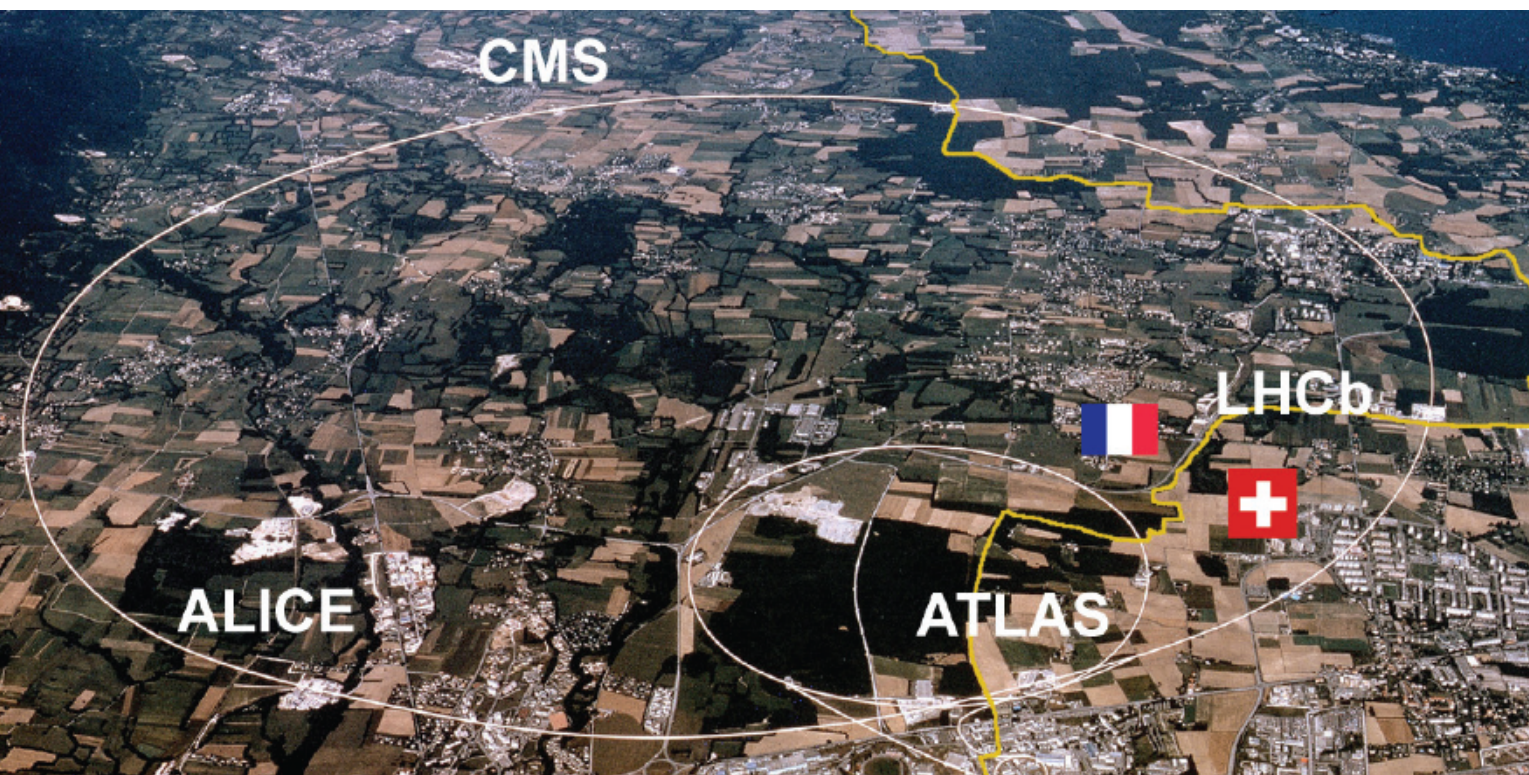
**The use of container- and cloud-based reproducibility platforms is a promising solution that can adequately preserve a wide variety of research studies across sciences.**





# Characterising particle physics research

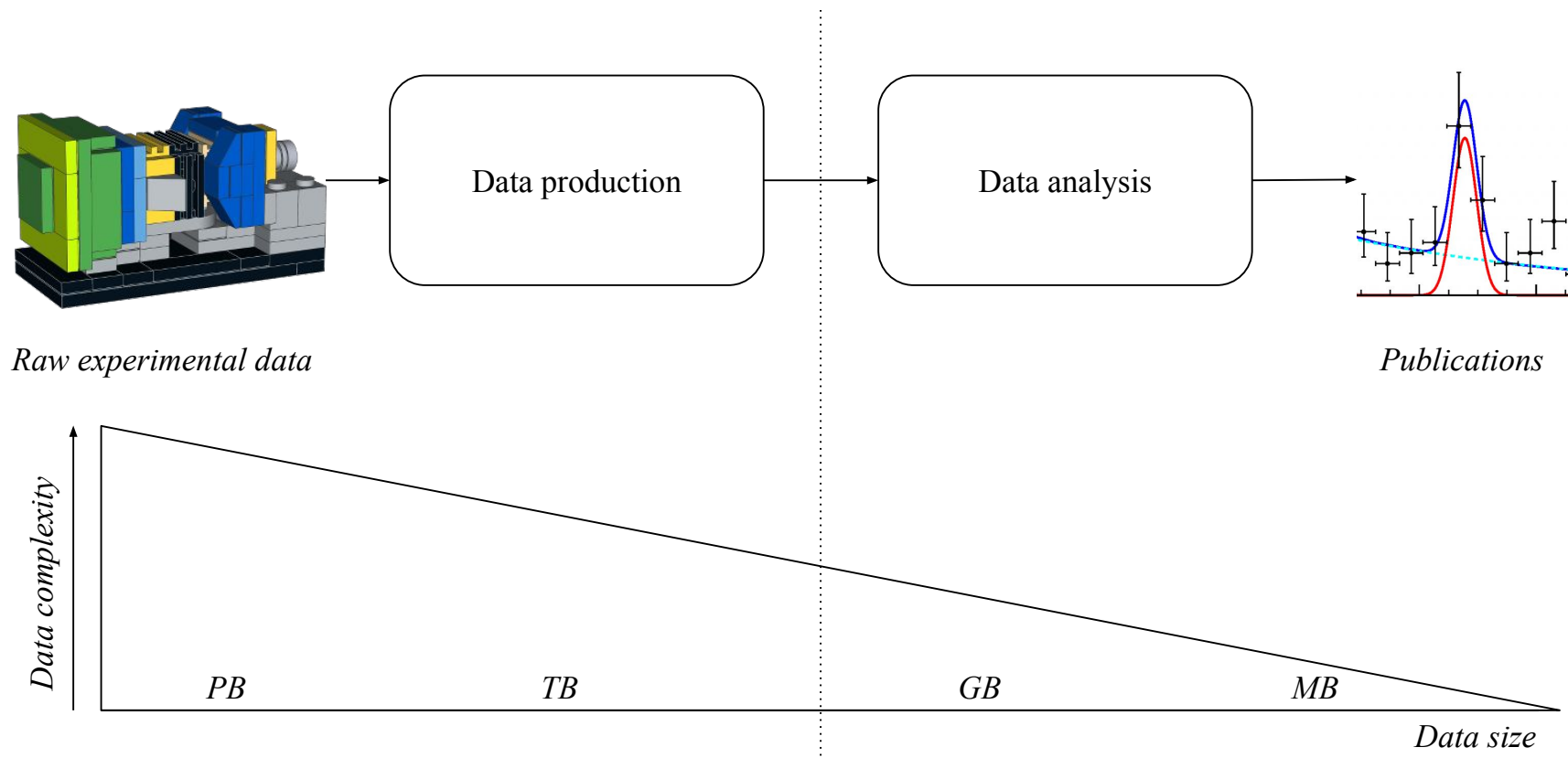
- Large experiments producing petabytes of data
- Large collaborations





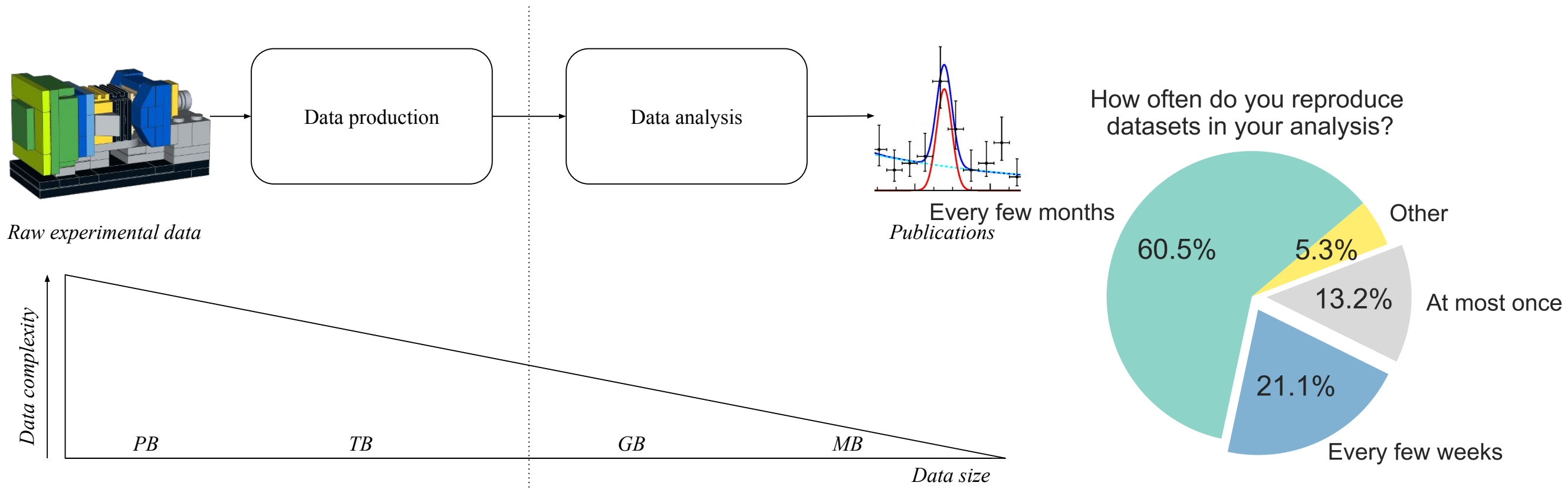
# Characterising particle physics research

- Typically use free and open source software
- Data to be released open access due to funding-body policies



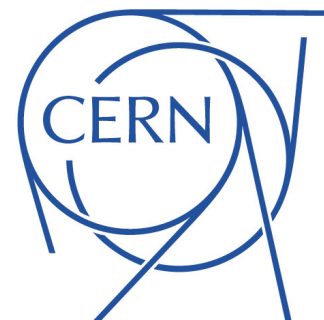
# Characterising particle physics research

- Typically use free and open source software
- Data to be released open access due to funding-body policies



# CERN Analysis Preservation

- Digital repository for describing and capturing physics analyses
- Supports integration with databases of the LHC collaborations (analysis databases)
- Supports integration with GitLab, GitHub, DockerHub



**No files have been attached.** Upload your analysis files here (n-tuples, macros, publication, output, etc). 10 GB of storage are available for each analysis

### Basic Information

Please provide some information relevant for all parts of the Analysis here



### Stripping/Turbo Selections *[0 items]*



### ntuple/userDST-production *[0 items]*

Please provide information about the steps of the analysis



### User Analysis



### Additional Resources

Please provide information about the additional resources of the analysis



:: Untitled document

 SAVE & CONTINUE

Files | Data | Source Code

Submission Form



**No files have been attached.** Upload your analysis files here (n-tuples, macros, publication, output, etc). 10 GB of storage are available for each analysis

### Basic Information

Please provide some information relevant for all parts of the Analysis here



Analysis Name

Measurement

Proponents



Reviewers



Review eGroup

Status

Choose from list





Reproducible research data analysis platform

### Flexible

Run many computational workflow engines.



### Scalable

Support for remote compute clouds.



### Reusable

Containerise once, reuse elsewhere. Cloud-native.



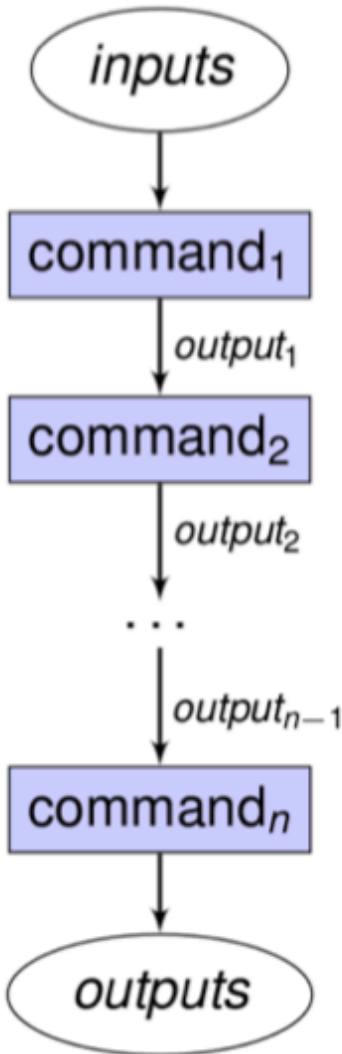
### Free

Free Software. MIT licence.  
Made with ❤️ at CERN.



- The project REANA allows analysis re-execution on a kubernetes cluster
- [reana.io](https://reana.io)

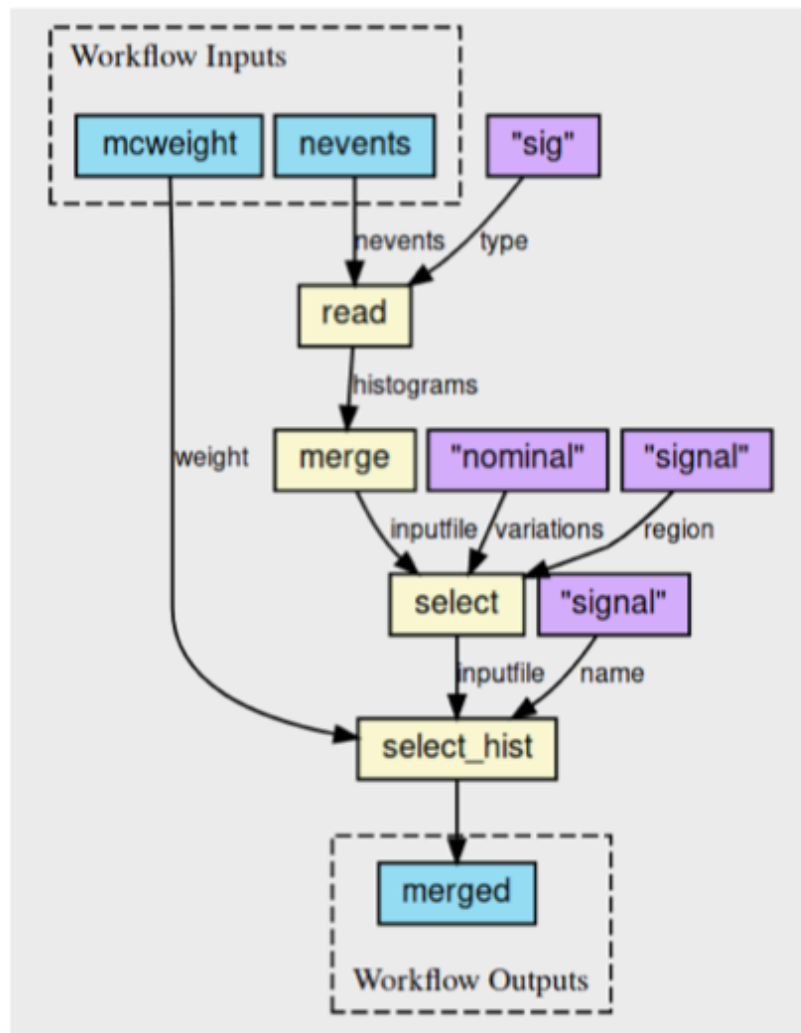
# Reana supports tools for workflow design and execution



## Serial



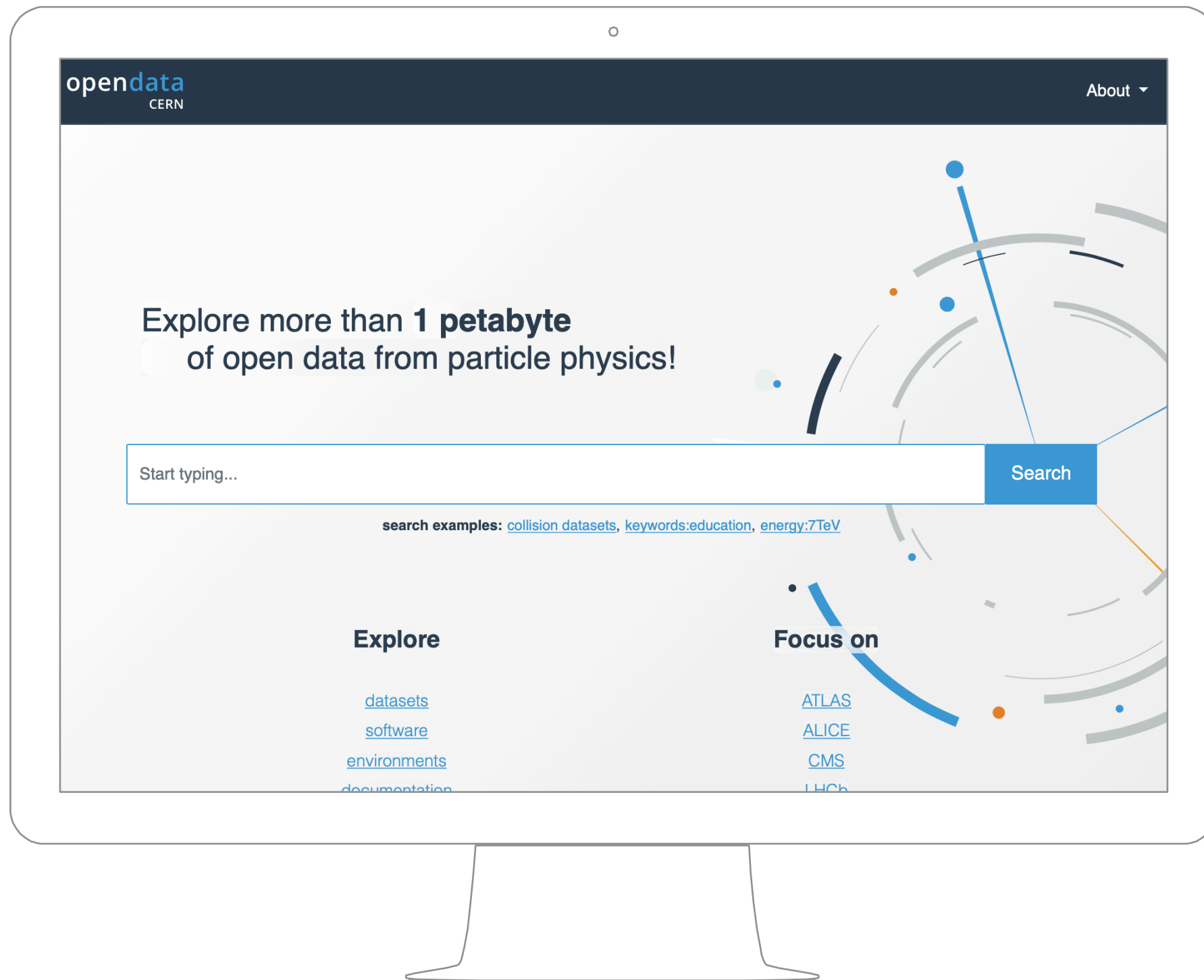
## Yadage



CWL



# CERN Open Data



masterclass ×

Filter by type

Dataset58

Derived58

Documentation4

Activities4

Environment1

VM1

Software3

Analysis2

Tool1

Filter by experiment

ALICE4

ATLAS57

CMS1

LHCb4

Filter by year

201118

201237

Filter by file type

jpg1

root3

Filter by keywords

education13

external resource9

masterclass66

teaching7

Sort by: Best match ▾ asc. ▾

Display: detailed ▾ 20 results ▾

Found 66 results.

ATLAS ZPath 2015 Masterclass dataset

A dataset of 1000 event display files accessible  
events were recorded in 2012 by the ATLAS det

Dataset Derived ATLAS

ATLAS WPath 2015 Masterclass dataset

A dataset of 1000 event display files accessible  
events were recorded in 2011 by the ATLAS det

Dataset Derived ATLAS

ATLAS WPath 2015 Masterclass dataset

A dataset of 1000 event display files accessible  
events were recorded in 2011 by the ATLAS det

Dataset Derived ATLAS

ATLAS WPath 2014 Masterclass dataset

A dataset of 1000 events taken in 2011 by the A  
Path...

Dataset Derived ATLAS

Event Selection

Physics Objects

Item #1

Objectbjeta ▾

Jet typeAK5Calo ▾

Jet CorrectionsJetCorrections ▾

Number<, >, =, <= ▾  
<=, >=

NumberNumber, e.g. 1

Selection CriteriaLoose Medium Other Tight

DiscriminatorTag Select Tag ▾  
Value1

pT CutsItem #1

<, >, = > ▾

GeV

|η| Cuts

+ Add New Item

# Conclusions

- Computational reproducibility is a hard problem that is likely becoming worse
- Container-based reproducibility platforms as a possible solution
- Some disciplines (like particle physics) require tailored solutions

