CERN openlab

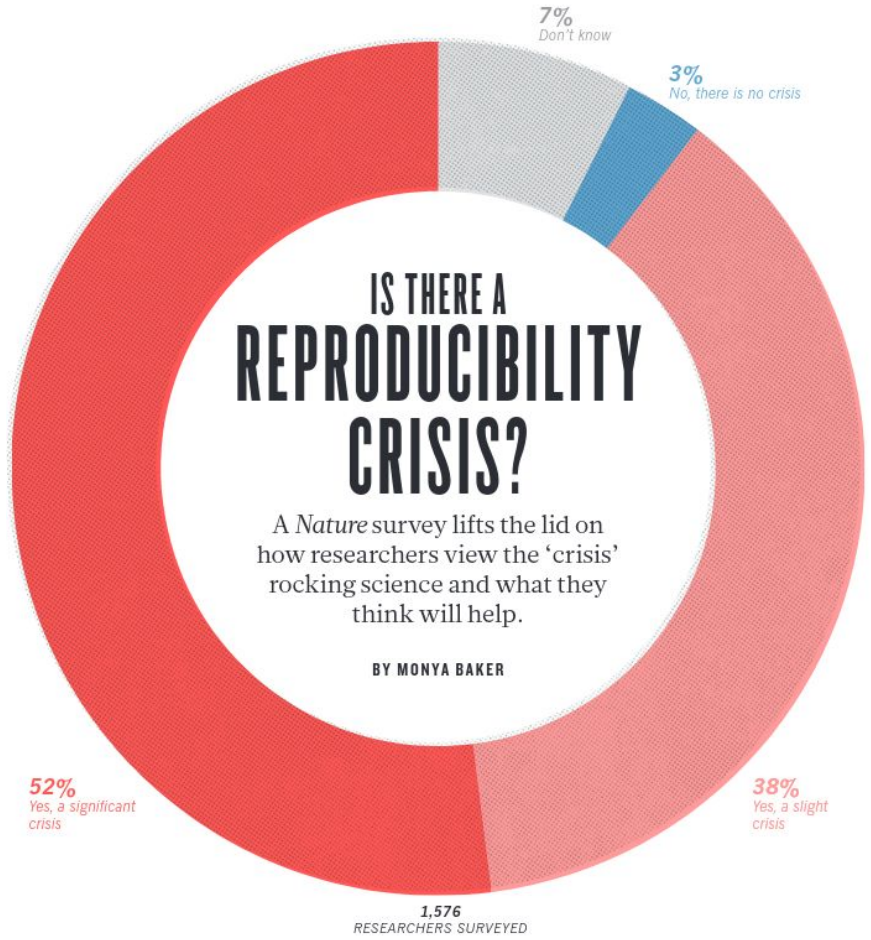
Abstraction of user storage mechanisms for heterogeneous REANA scientific pipelines

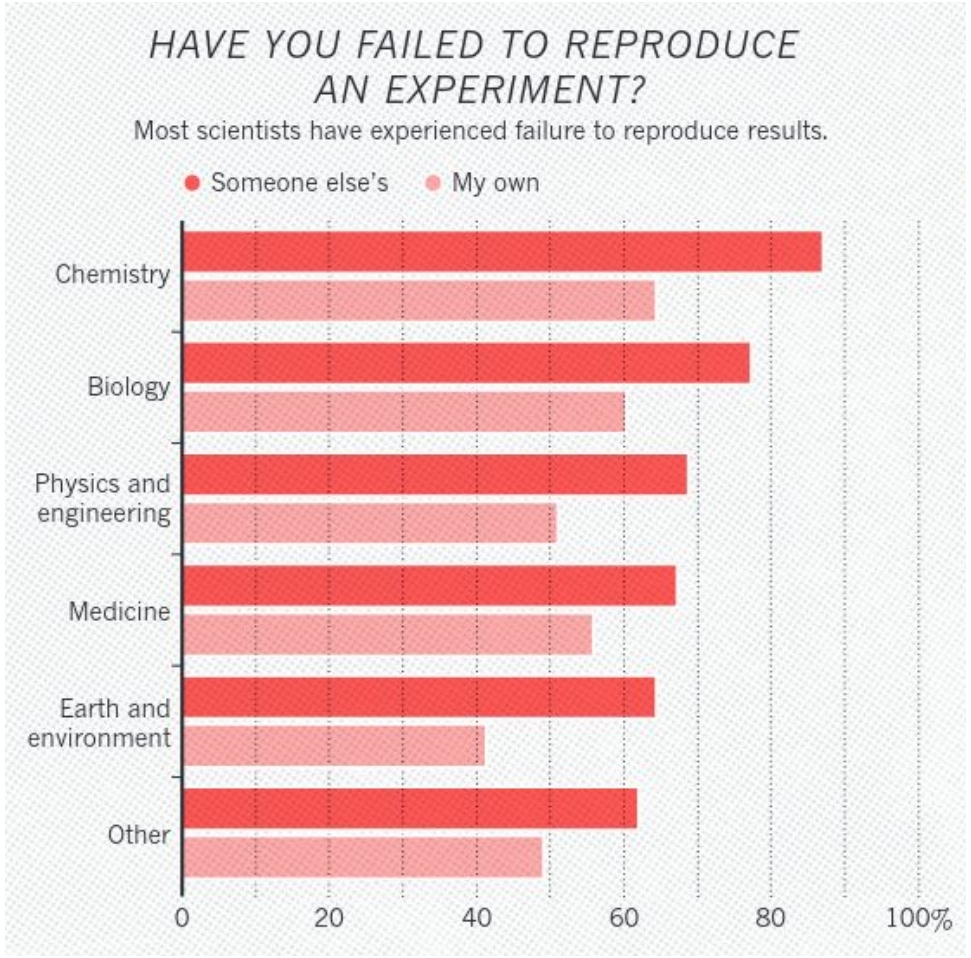
Maria Camila Diaz Sanchez

Supervisors: Tibor Šimko, Audrius Mečionis

December 1st, 2021

Reproducibility Crisis

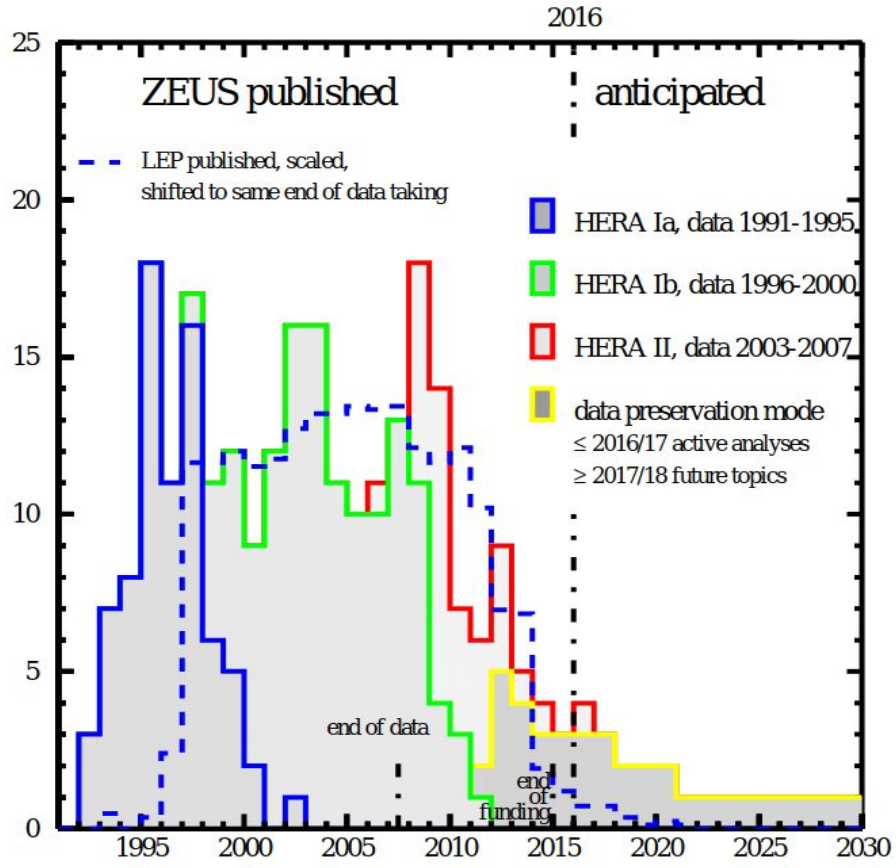




[Nature: 1,500 scientists lift the lid on reproducibility]

HEP

The complexity and cost of HEP research makes it difficult to reproduce the experiments.



Verbytskyi, A (2016)

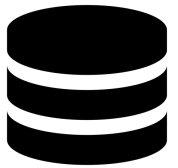
With the emergence of new theories, scientists analyse and reuse the data many years later

```
1 import pandas as pd
2 pd.__version__
```

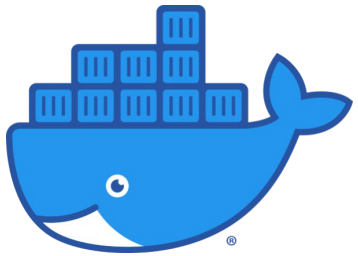
```
Out[1]: '0.24.2'
```

```
1 import pandas as pd
2 pd.__version__
```

```
Out[1]: '1.0.1'
```

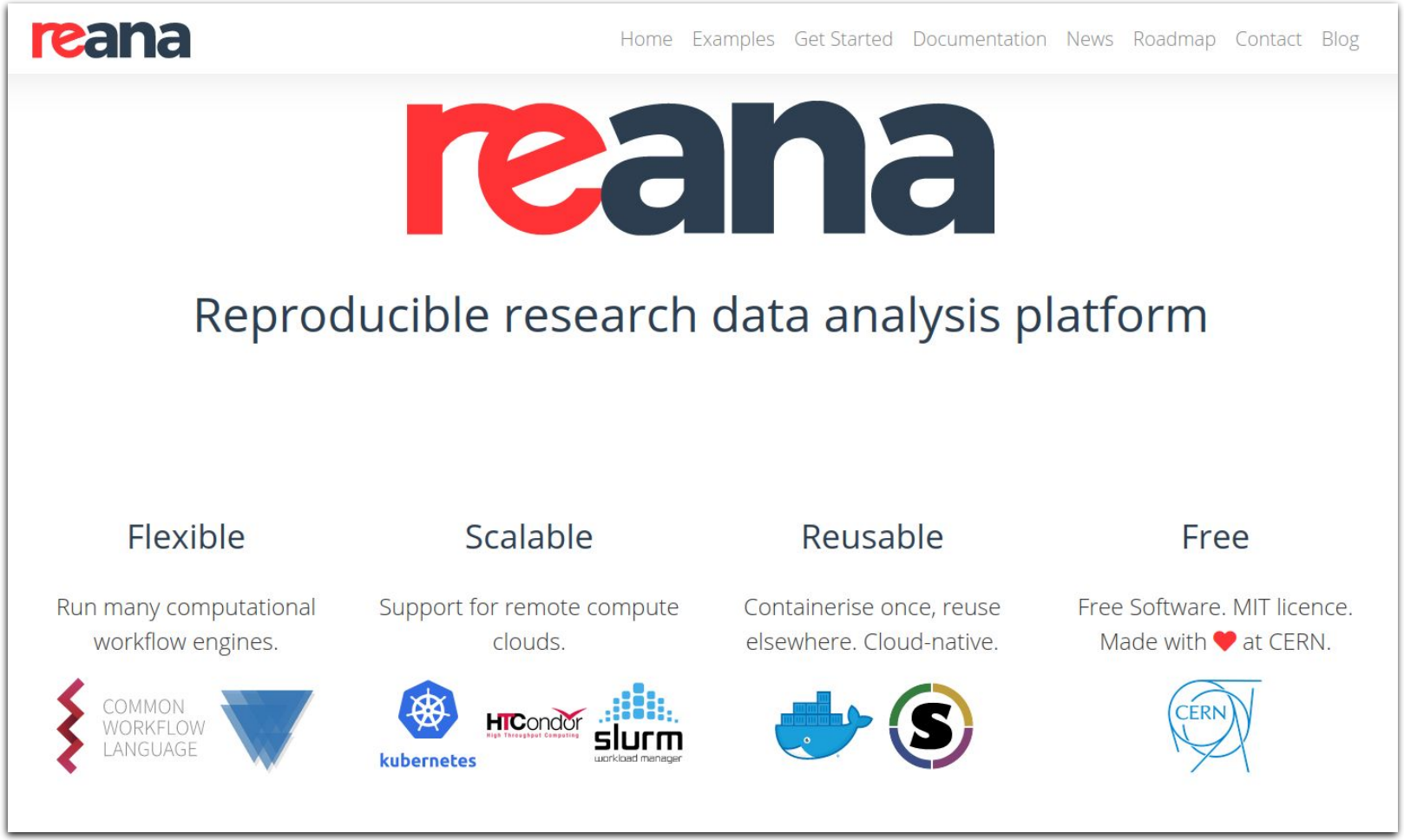


Can you easily run you analyses under new computing environments?



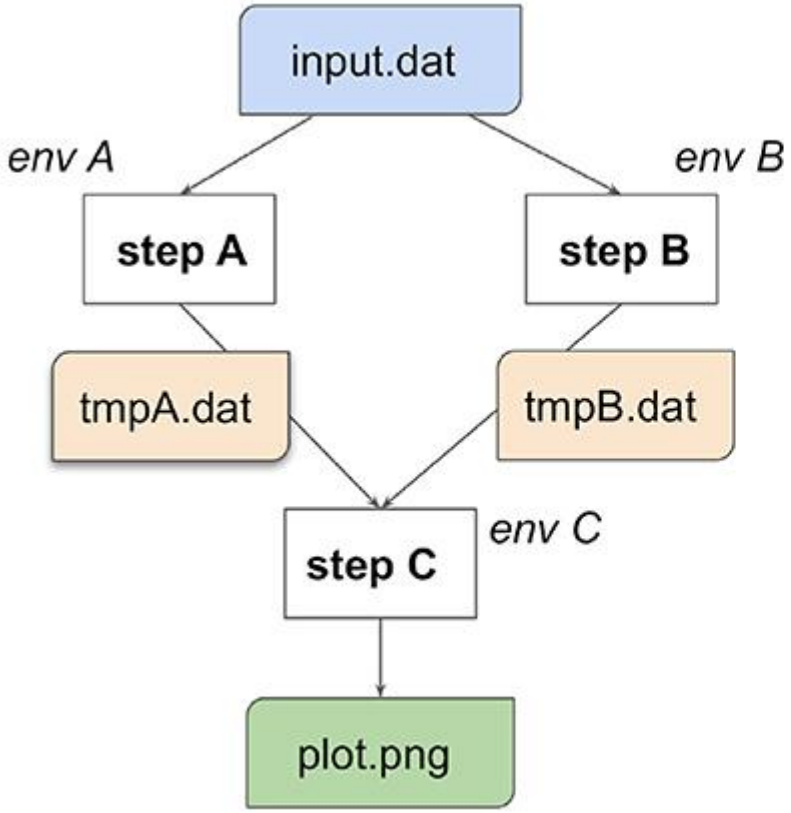
Non-reproducible research is problematic for advancing science

Some technologies are heading to make this changes less problematic


reana
Home Examples Get Started Documentation News Roadmap Contact Blog
reana
Reproducible research data analysis platform
Flexible
Run many computational workflow engines.
COMMON WORKFLOW LANGUAGE
Scalable
Support for remote compute clouds.
kubernetes
HTCondor
slurm
workload manager
Reusable
Containerise once, reuse elsewhere. Cloud-native.
Free
Free Software. MIT licence.
Made with ❤ at CERN.
CERN

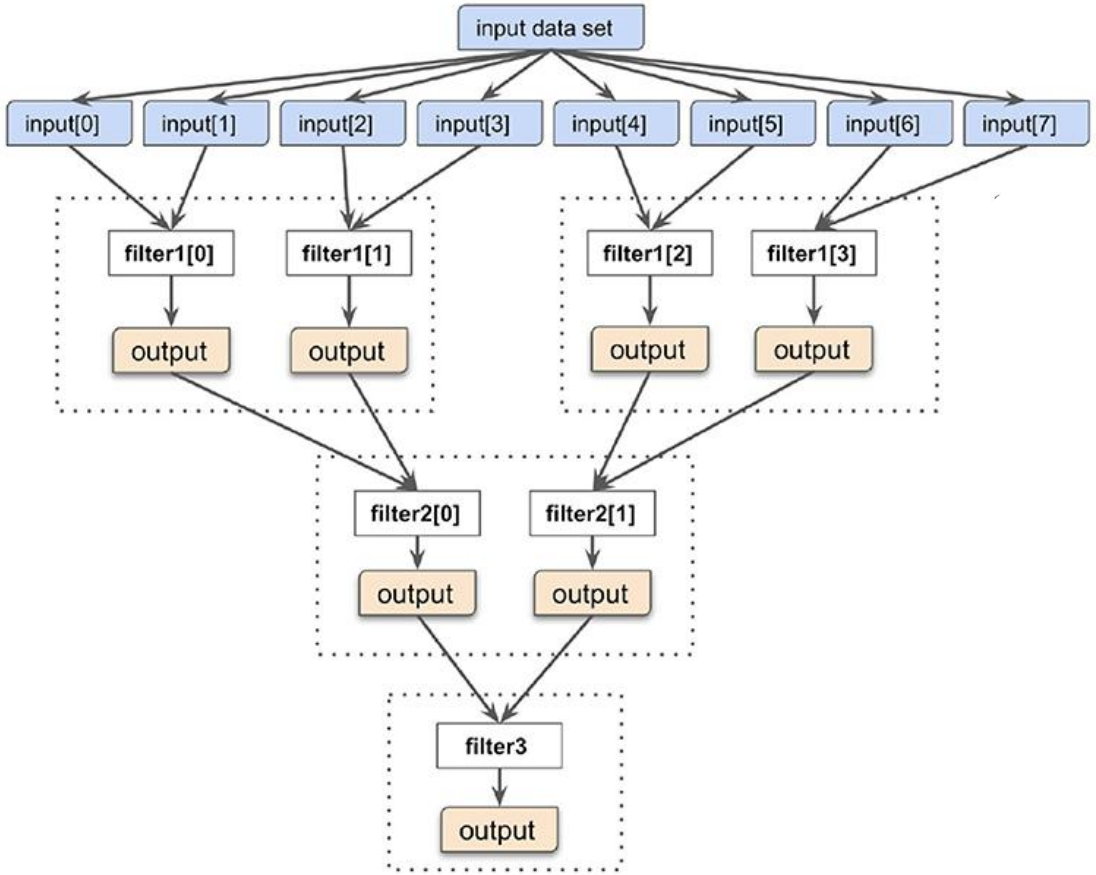

Input Data
Code
reana
Environment
COMMON
WORKFLOW
LANGUAGE
Workflow

What is a declarative workflow?



The focus is on expressing **"what"** step should do, separating from and without focusing on **"how"***

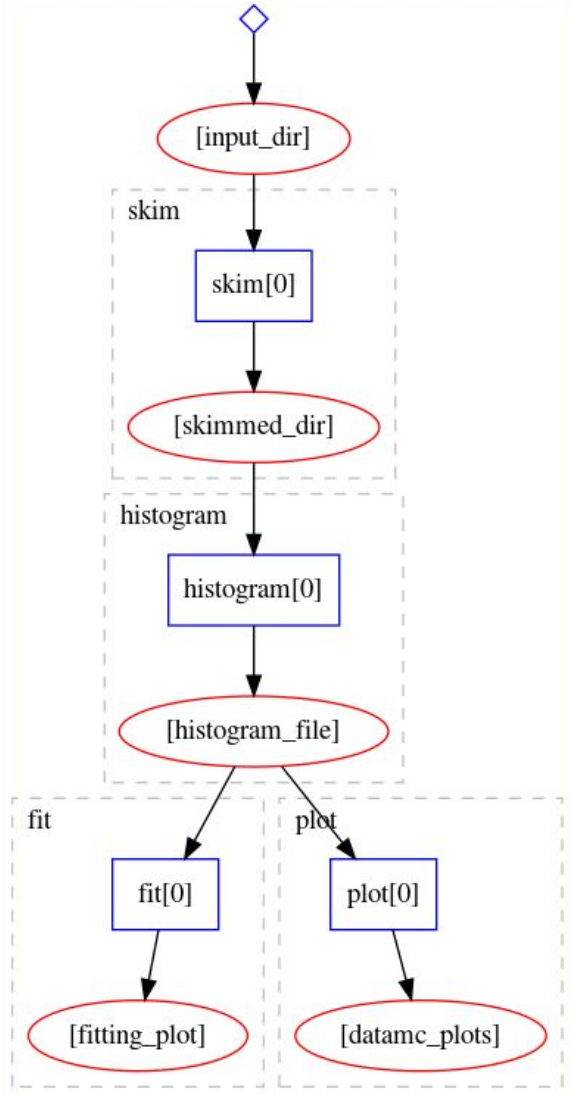
*Until is needed, declared and used

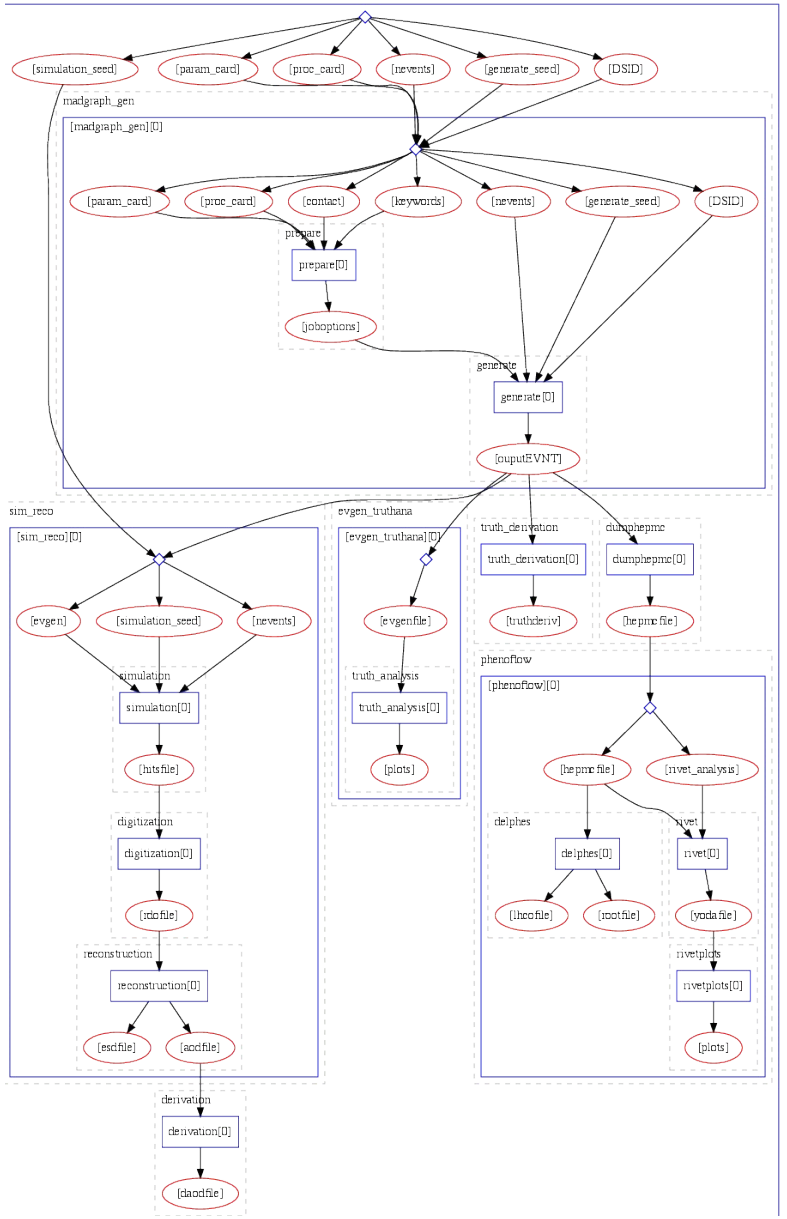


Can be orchestrated

Each step is isolated and independent

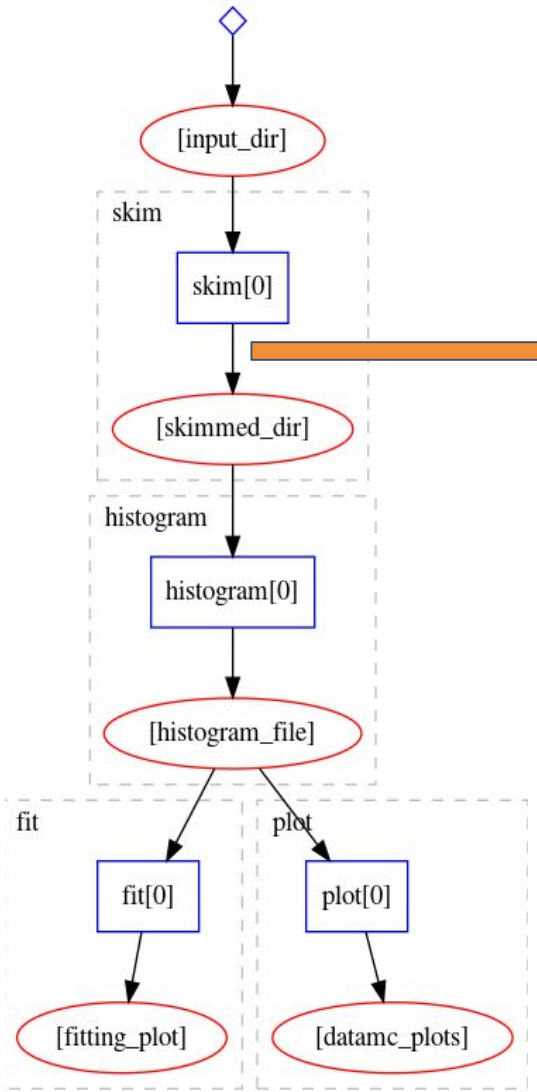
Can be parallelized and scalable





Hildreth, M; et al (2018)

My task



Where are all the files?
Where is the workspace?

Currently: Predefined folder in shared storage

How to generalize it, to plug other type of storage?

Abstract the workspace from the workflow

models: extract the workspace concept from within the Workflow model #94 Edit New issue

Closed **audrium** opened this issue on Oct 9, 2020 · 0 comments

audrium commented on Oct 9, 2020 Member

Addressing the following line in #89:

> (2) opening interactive sessions without necessarily having any workflow yet.

There is now no notion of `workspace` for `InteractiveSession`'s, however, the `InteractiveSession`'s implementation depends on it. So far, we have been depending on the workflow model to retrieve it, which was ok because the interactive session was also within the workflow model.

As we are evolving to this new idea of separate sessions, I see a need to extract the workspace concept from within the `Workflow` model to a new entity. This idea gains more importance once we start thinking about using different workspaces than CephFS.

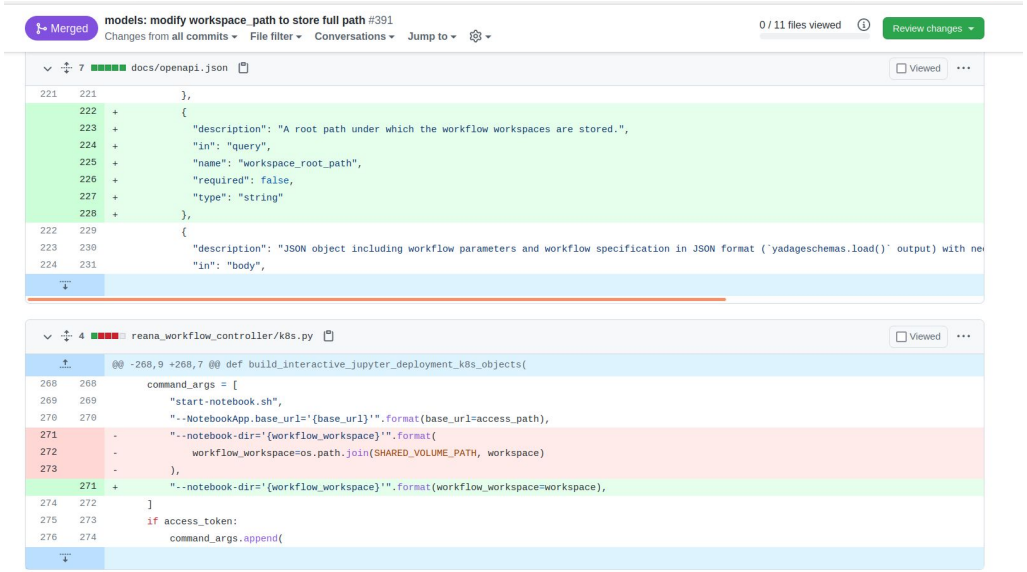
*Originally posted by **@diegodelemos** in #90 (comment)*

Assignees
marcdiazsan

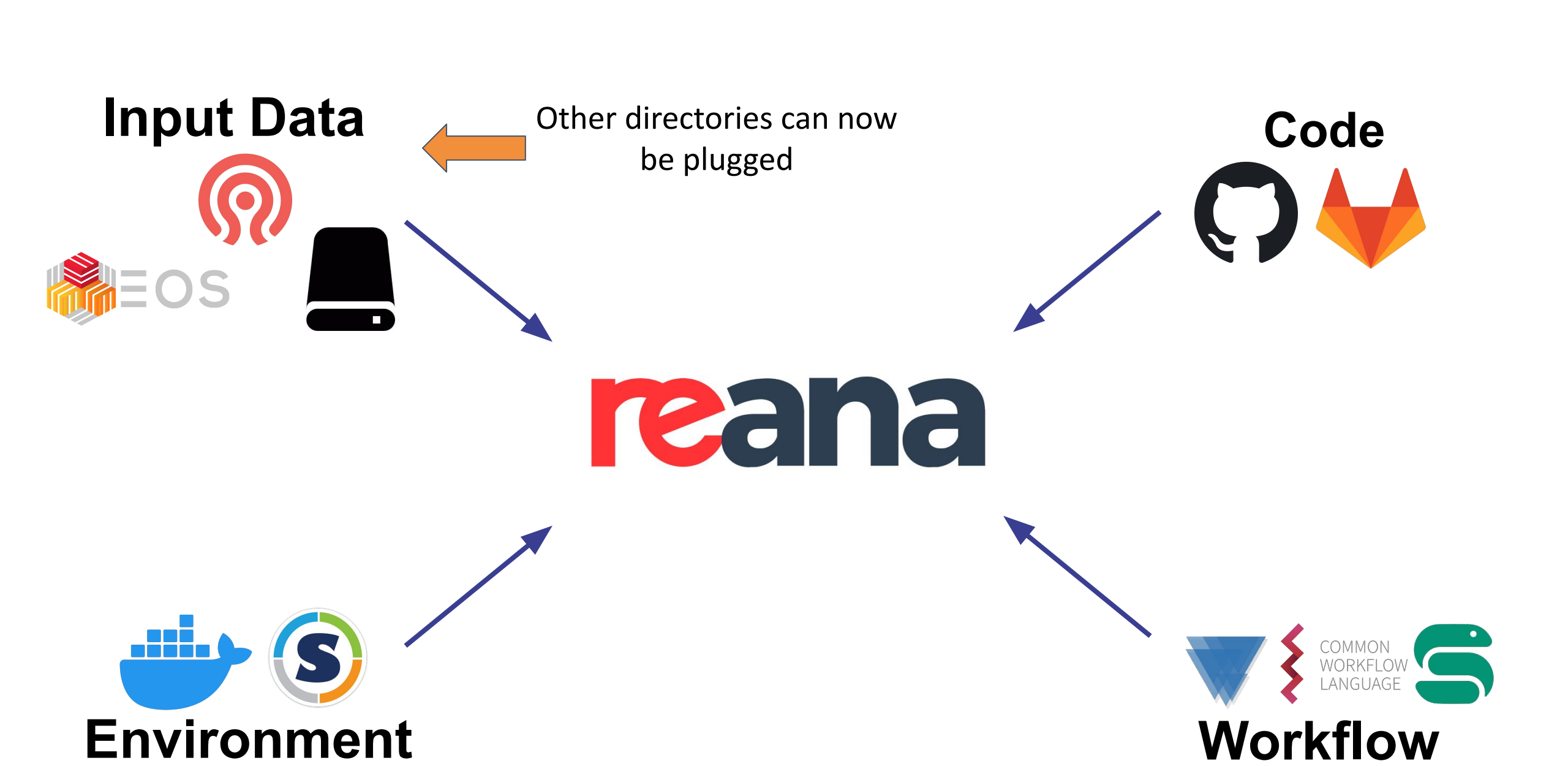
Labels
None yet

Projects
Workspace-Abstraction
Done

Milestone



Refactoring the code to allow different storage strategies


Input Data
Other directories can now be plugged
Code
EOS
reana
Environment
COMMON WORKFLOW LANGUAGE
Workflow

EXTRA RESOURCES


www.reanahub.io


docs.reanahub.io


@reanahub


info@reanahub.io


@reanahub

Contact reana team!

QUESTIONS?

marcdiazsan@unal.edu.co
@marcdiazsan