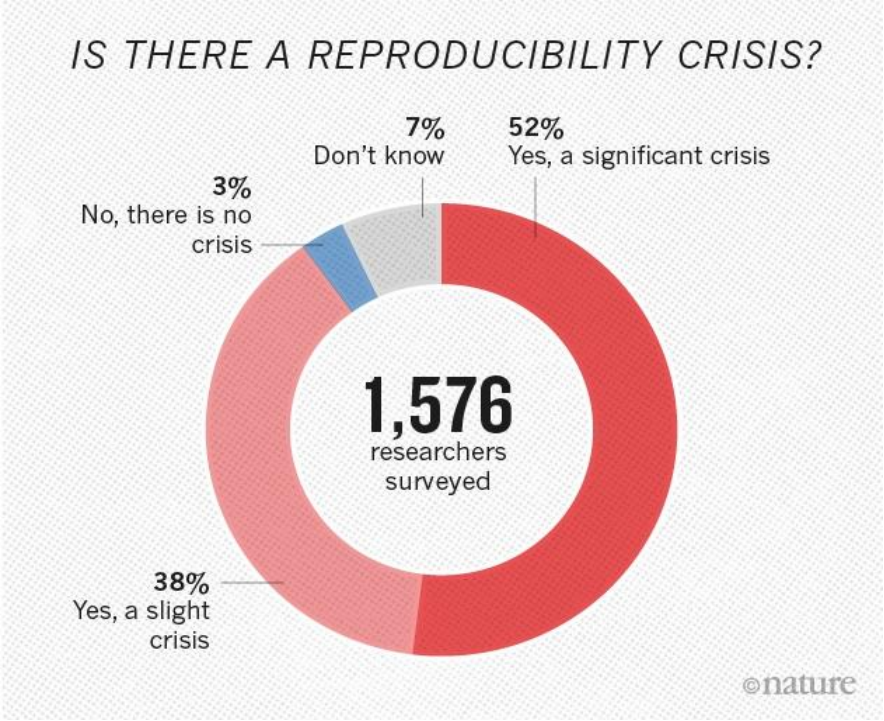
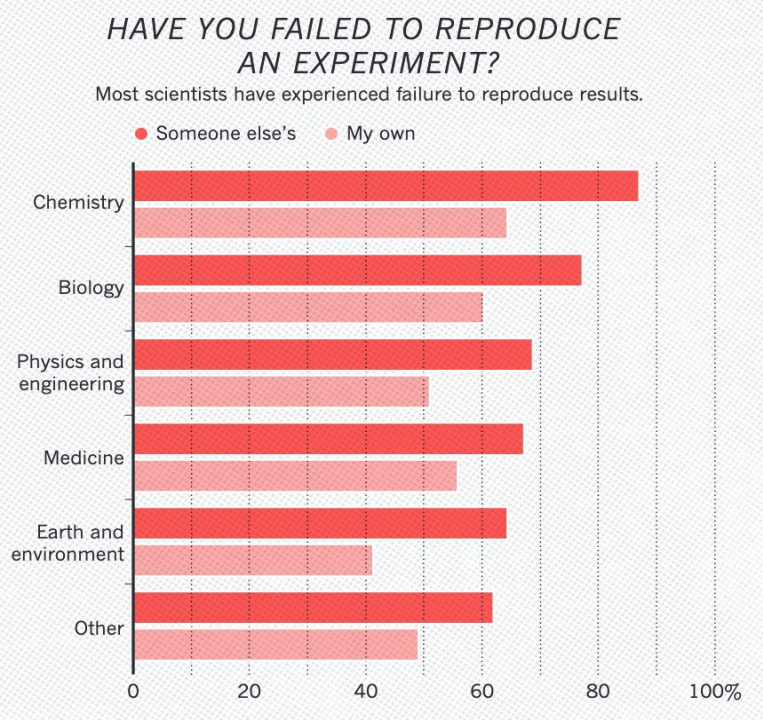
Containerised data analysis workflows using reana

Marco Donadoni - CERN
CSC 2022, Krakow

Reproducibility crisis

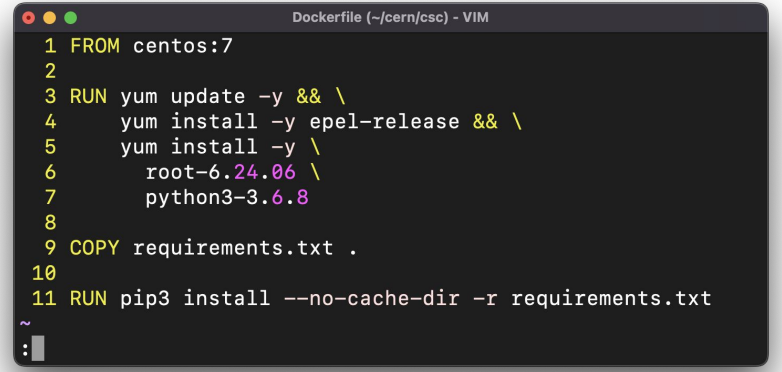




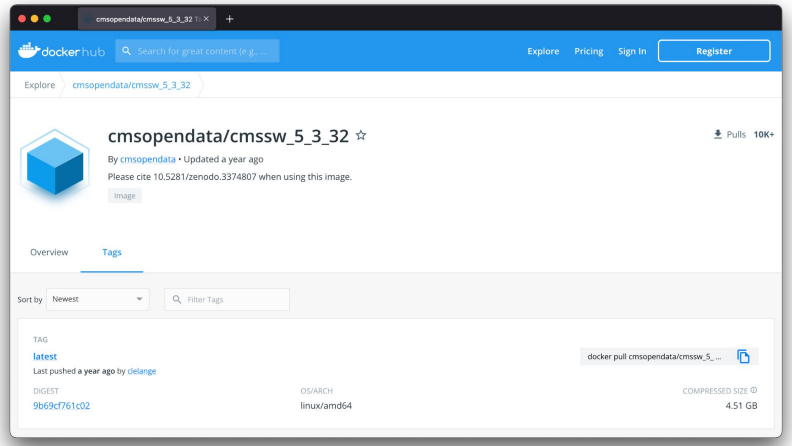
Baker, M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454 (2016). https://doi.org/10.1038/533452a

Containers

- Container images encapsulate computational environments that are independent from the hosting system
- Container images have everything needed by the data analysis
- Data analyses can be run in an isolated environment
  - … on your machine
  - … on a remote cloud
  - … by someone else



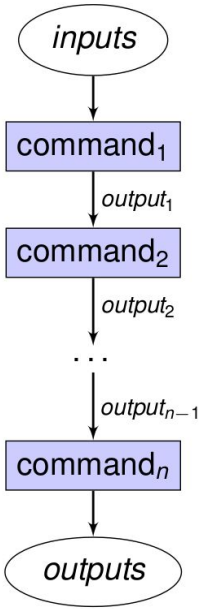
Dockerfile example

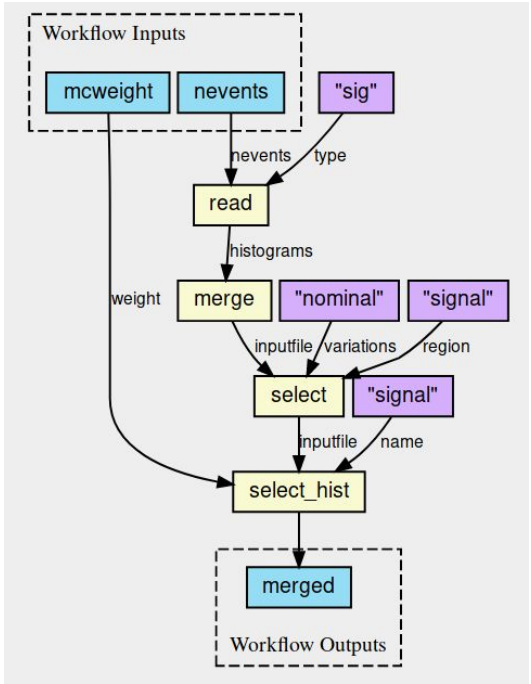


Containerised CMS software framework

Workflows

- A workflow is a recipe explaining how to produce results from input data
- What is needed?
  - Input data
  - Code
  - Computing environment
  - Workflow steps
- Steps can be described by means of declarative workflow languages (CWL, Snakemake, Yadage, …)
- Each step can use (possibly) different container images

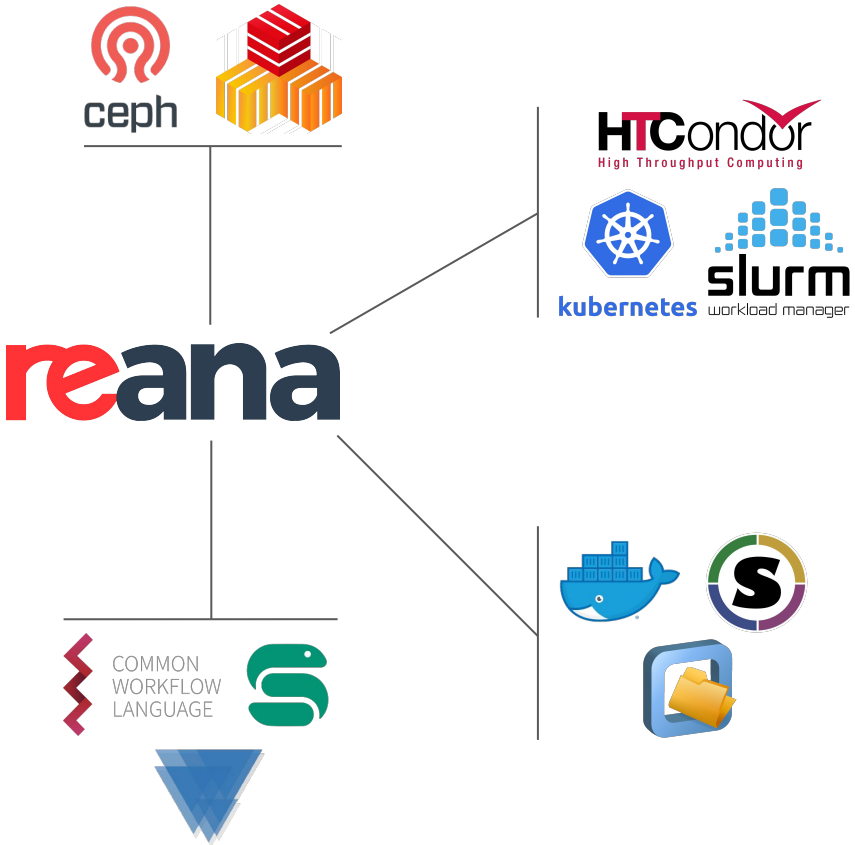




What is REANA?

REANA helps researchers run declarative workflows in containerised clouds

- multiple container technologies (Docker, Singularity)
- multiple workflow systems (CWL, Snakemake, Yadage)
- multiple compute backends (HTCondor, Kubernetes, Slurm)
- multiple shared storage platforms (Ceph, EOS, NFS)



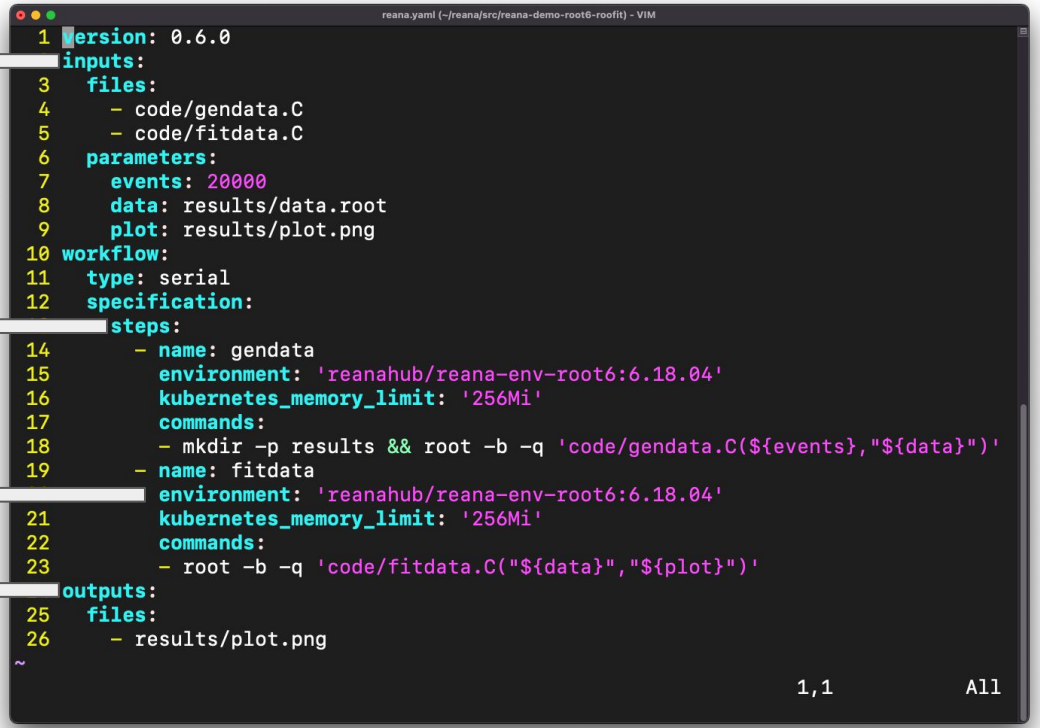
Analysis Example

Input data/params & Code

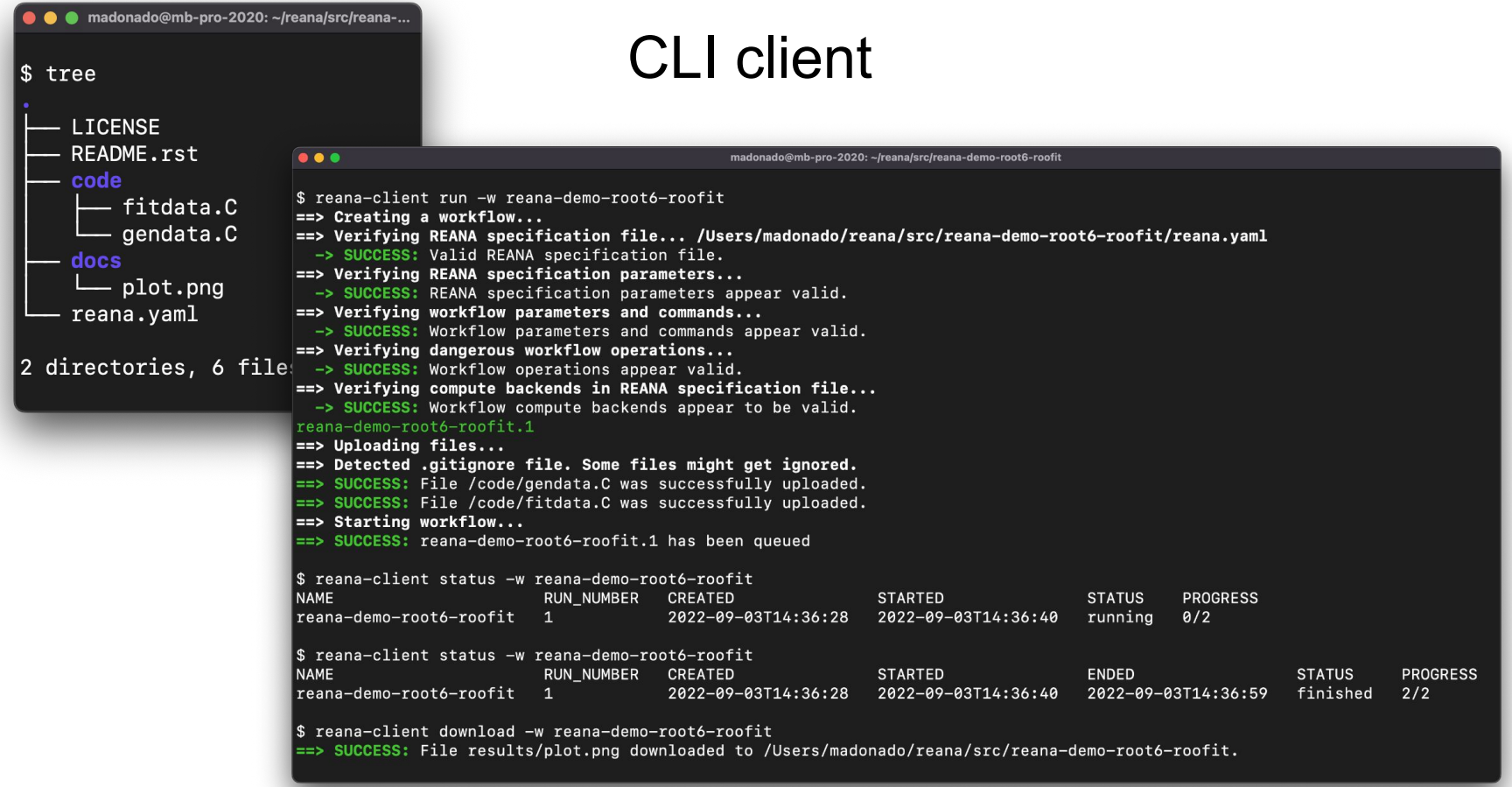
Workflow steps

Environment

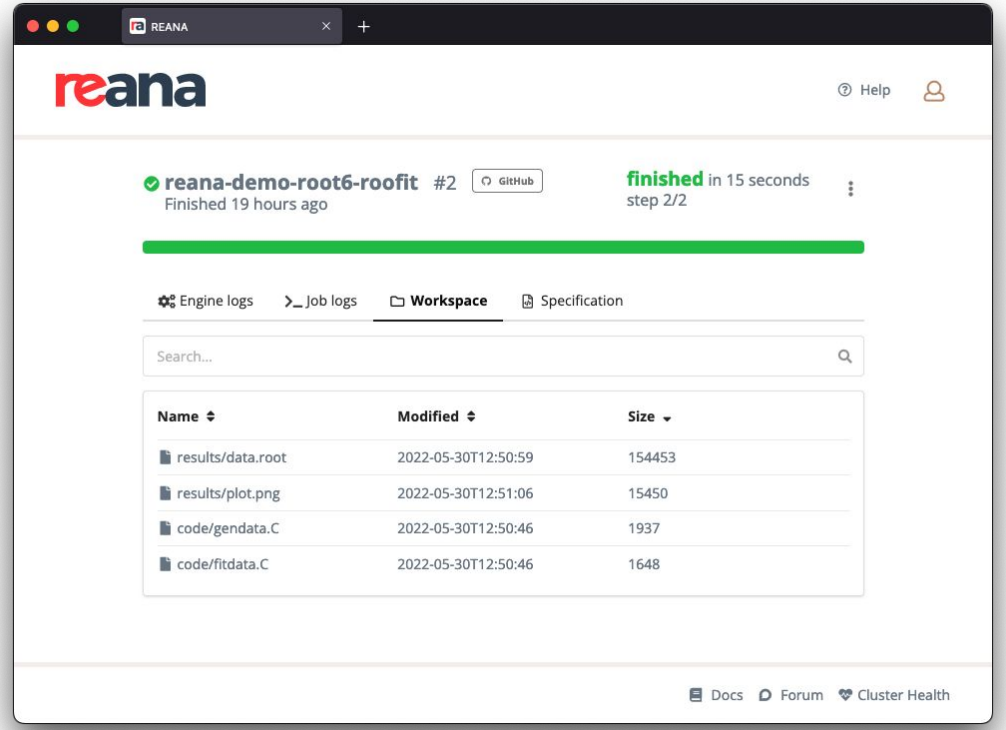
Results



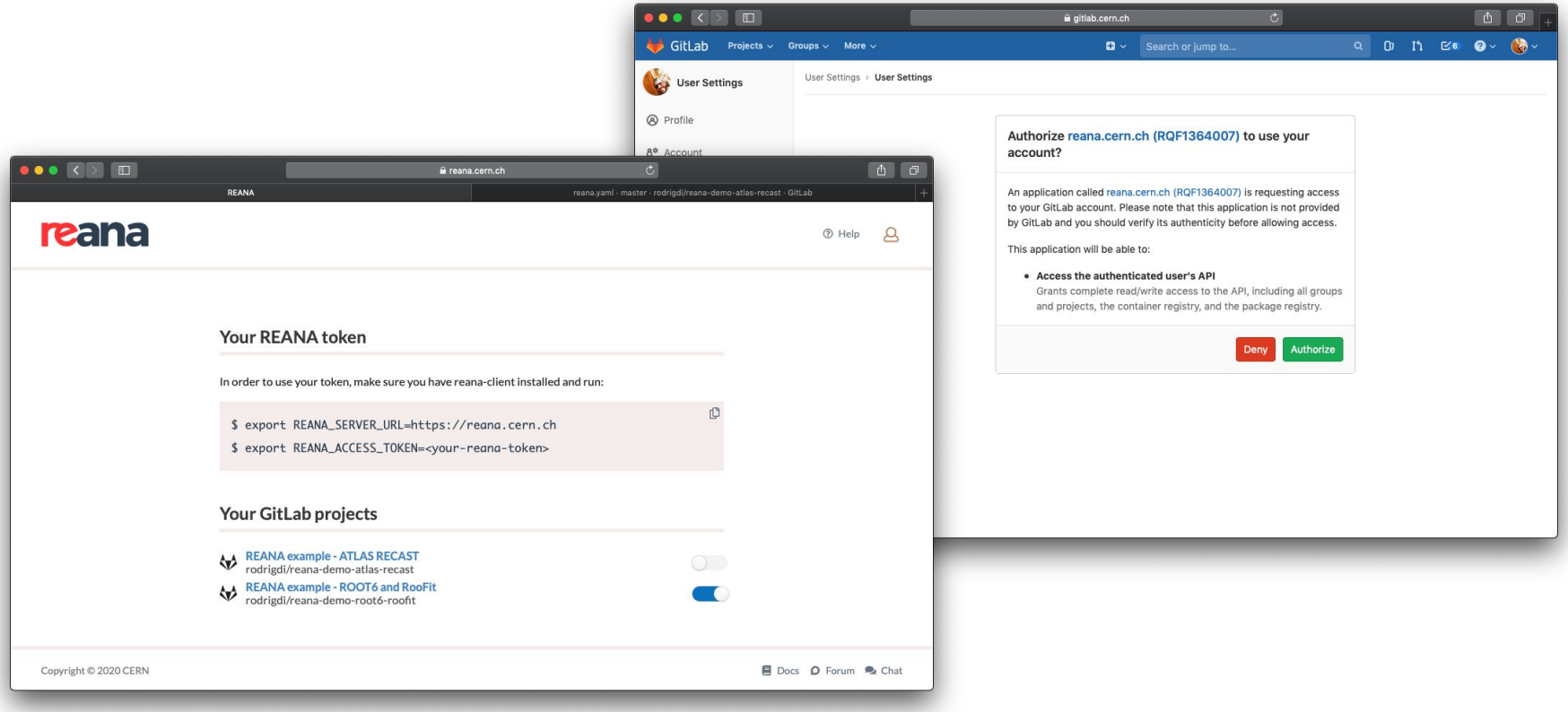
CLI client


madonado@mb-pro-2020: ~/reana/src/reana-...
$ tree
.
├── LICENSE
├── README.rst
├── code
│   ├── fitdata.C
│   └── gendata.C
├── docs
│   └── plot.png
└── reana.yaml
2 directories, 6 file
madonado@mb-pro-2020: ~/reana/src/reana-demo-root6-roofit
$ reana-client run -w reana-demo-root6-roofit
==> Creating a workflow...
==> Verifying REANA specification file... /Users/madonado/reana/src/reana-demo-root6-roofit/reana.yaml
-> SUCCESS: Valid REANA specification file.
==> Verifying REANA specification parameters...
-> SUCCESS: REANA specification parameters appear valid.
==> Verifying workflow parameters and commands...
-> SUCCESS: Workflow parameters and commands appear valid.
==> Verifying dangerous workflow operations...
-> SUCCESS: Workflow operations appear valid.
==> Verifying compute backends in REANA specification file...
-> SUCCESS: Workflow compute backends appear to be valid.
reana-demo-root6-roofit.1
==> Uploading files...
==> Detected .gitignore file. Some files might get ignored.
==> SUCCESS: File /code/gendata.C was successfully uploaded.
==> SUCCESS: File /code/fitdata.C was successfully uploaded.
==> Starting workflow...
==> SUCCESS: reana-demo-root6-roofit.1 has been queued
$ reana-client status -w reana-demo-root6-roofit
NAME RUN_NUMBER CREATED STARTED STATUS PROGRESS
reana-demo-root6-roofit 1 2022-09-03T14:36:28 2022-09-03T14:36:40 running 0/2
$ reana-client status -w reana-demo-root6-roofit
NAME RUN_NUMBER CREATED STARTED ENDED STATUS PROGRESS
reana-demo-root6-roofit 1 2022-09-03T14:36:28 2022-09-03T14:36:40 2022-09-03T14:36:59 finished 2/2
$ reana-client download -w reana-demo-root6-roofit
==> SUCCESS: File results/plot.png downloaded to /Users/madonado/reana/src/reana-demo-root6-roofit.

Web interface

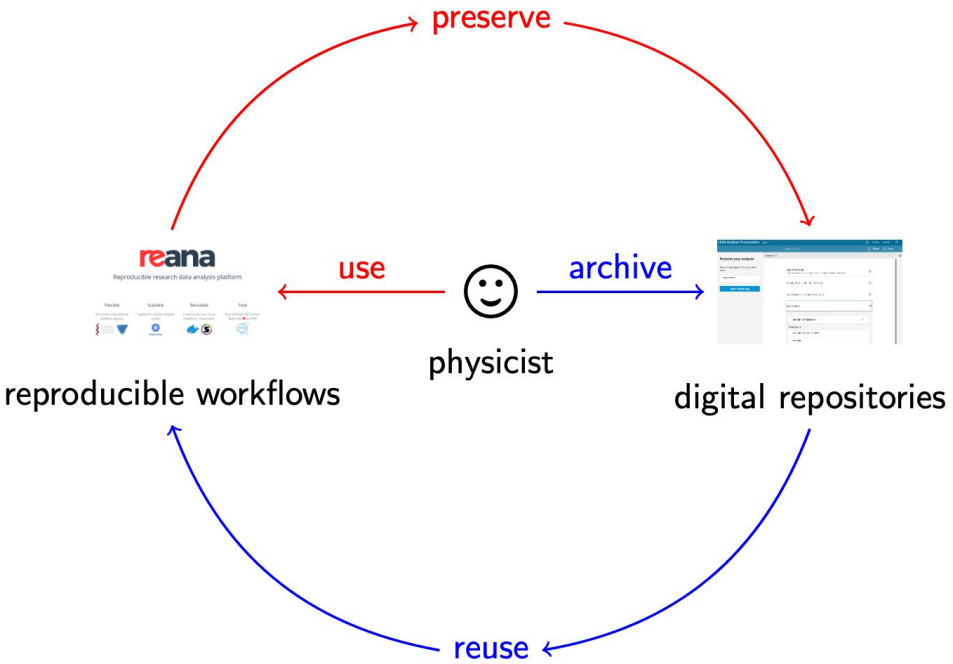


… and many other features such as GitLab integration



Conclusions

- Containerization helps in encapsulating computational environments
- Declarative workflow languages describe the analysis computation logic
- data + code + environment + workflow → reproducible analyses!
- REANA aims at helping researchers
  - execute complex data analyses on remote compute clouds
  - organise data analyses to facilitate preservation and reuse



https://www.reana.io @reanahub @reanahub