

# Curating Data Flows

Leveraging REANA for Reproducible Analyses  
of Dimensionality Reduction Workflows

# Data in the digital space of physics

- Fundamental physics is an early adopter of working with data in the digital space
- Consequently, a plethora of solutions has been developed within the different communities for:
  - Data Storage formats
  - Descriptions of data = Metadata and their schemas
  - Default tools for data handling and analysis
- Simulations and simulated data became a means for testing theories, where experiments can't serve
- The scale of experiments and other material research infrastructures drove a different division of labour and a culture of sharing
- The scale of the data collections require more than 'making data FAIR' for modern scientific research

# PUNCH4NFDI approach to FAIR data

- F(indable)
  - do have working approaches for managing and working with their own data
  - don't have working implementations for access by the other communities
  - **But:** there are good points to start from (various Registries for data publication sites, common protocols)
- A(ccessible)
  - Sharing data is well developed within the respective communities
  - Sharing data cross-community is tightly bound to making the software tools available and cooperative
  - Although there is no huge problem with **data privacy protection**,
    - data curation processes,
    - embargo periods,
    - and (computational) resources for working with the data

are serious challenges, require rethinking of well worn community (silo) solutions

  - **But:** bringing data and compute resources together is a hard problem for huge data sizes

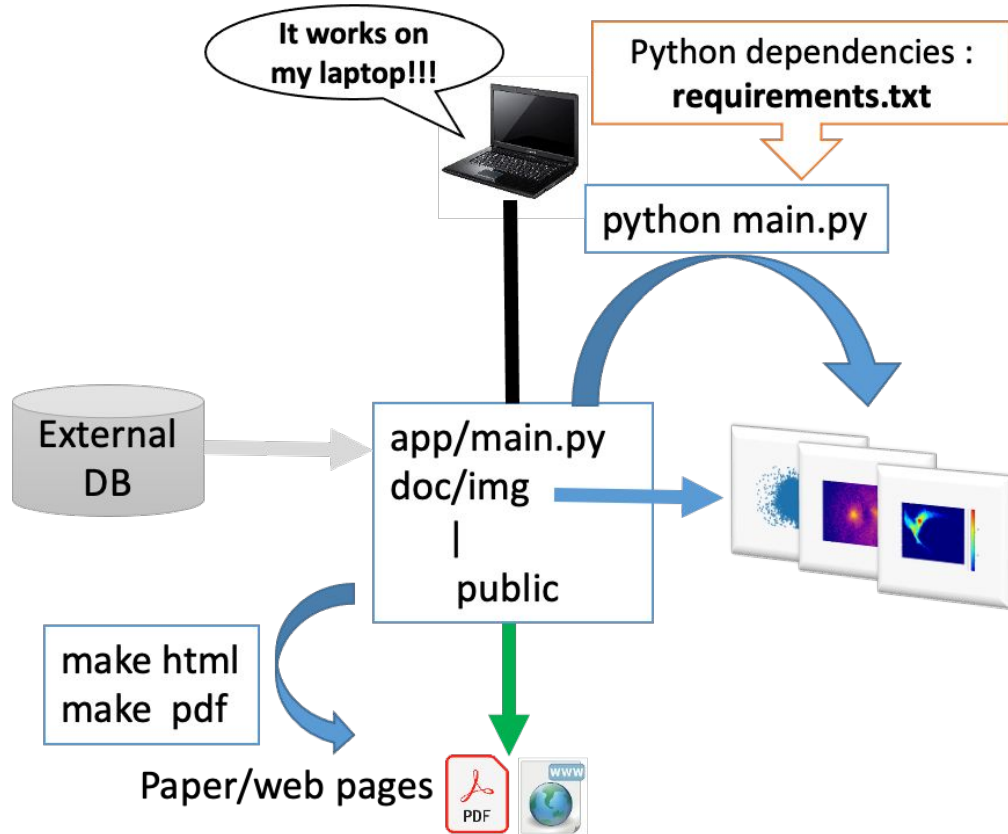
# PUNCH4NFDI approach to FAIR data

- I(nteroperable)
  - Data structures are very different
  - Mapping data structures via 'classical' metadata concepts are not sufficient
  - **But:** the mapping in itself is no goal, it's about the consistent availability for digital analysis
- R(eusable)
  - Reusable data for digital analysis implies the challenge of repeating analyses and achieving consistent results
  - Many elements of a digital analysis need to be stable and consistent:
    - Data
    - Software
    - Computing environment
    - Workflows
  - Reusability implies also reproducibility
    - Some modern instruments require real time decisions (for discarding lots of observed data) while preserving a degree of reproducibility of subsequent analysis results
  - **But:** with modern Cloud environment and tools the means are available to address these challenges

# PUNCH4NFDI Tasks:

- Making Storage and Compute facilities available across institutional and disciplinary domains
- Enabling efficient Authentication and Authorisation methods for resource providers of the communities
- Providing a registry for results of research products in this environment
- Hiding the complexities of the underlying digital landscape
- **Defining and Capturing workflows and execution environments**

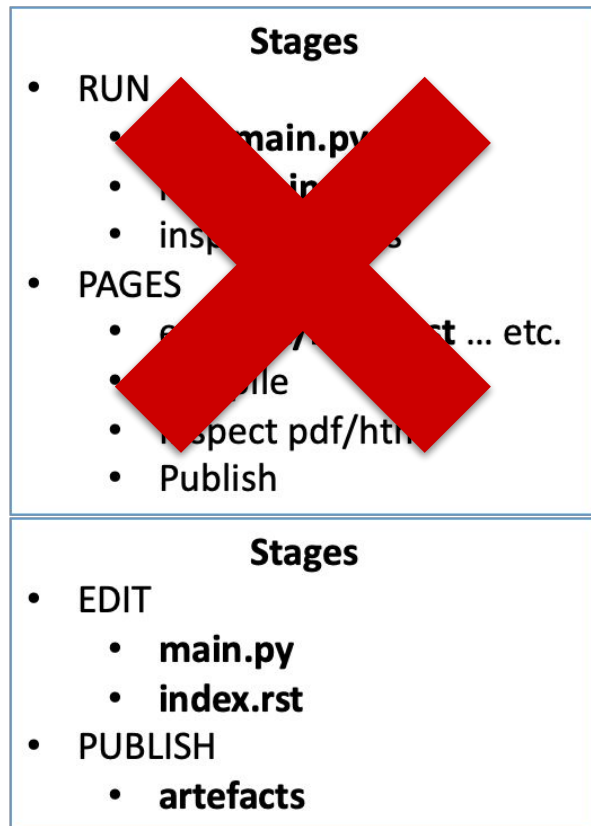
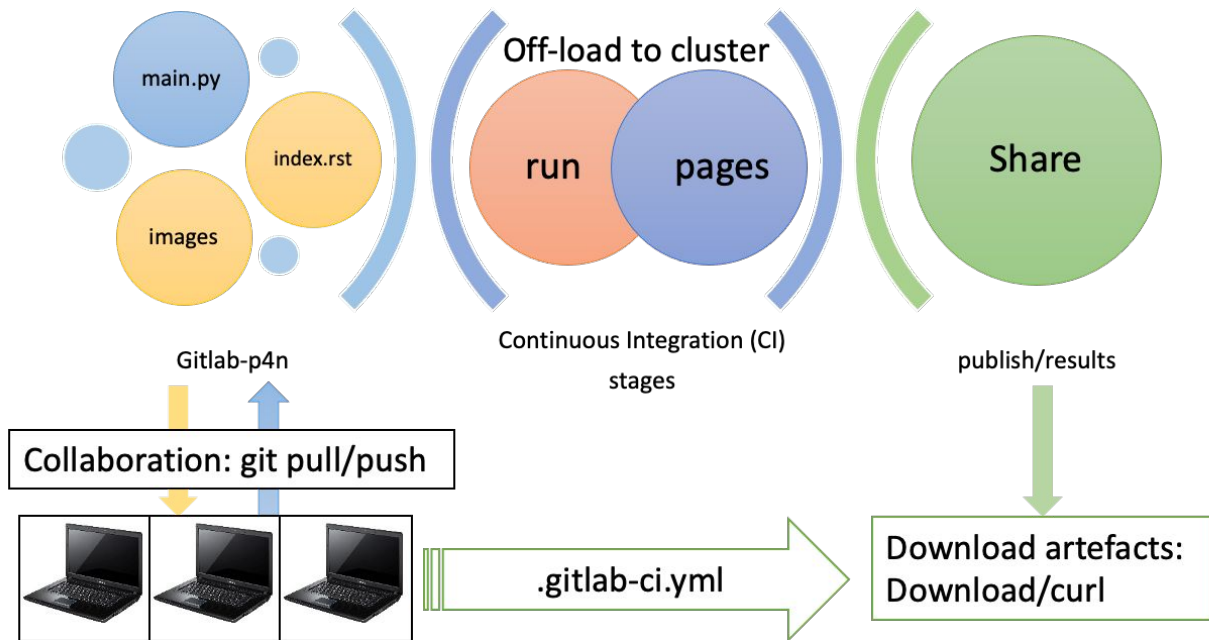
# From local Workflows...



## Stages

- RUN
  - edit **main.py**
  - run: **main.py**
  - inspect images
- PAGES
  - edit **doc/index.rst** ... etc.
  - compile
  - Inspect pdf/html
  - Publish

# ...to reproducible Workflows!



# reana

Reproducible research data analysis platform

## Flexible

Run many computational workflow engines.



## Scalable

Support for remote compute clouds.



## Reusable

Containerise once, reuse elsewhere. Cloud-native.



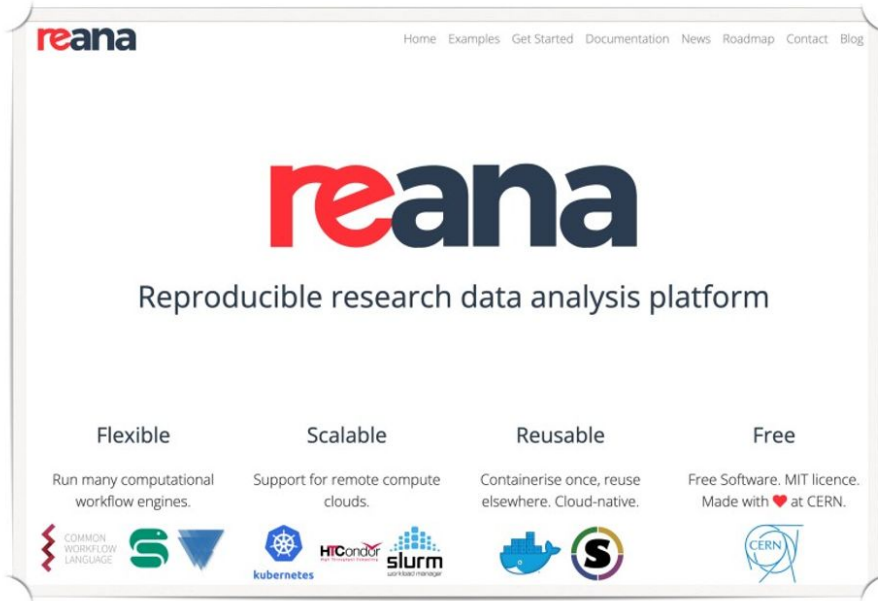
## Free

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





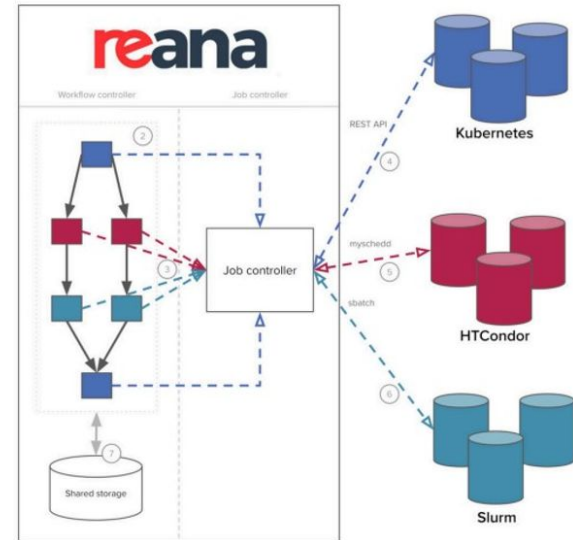
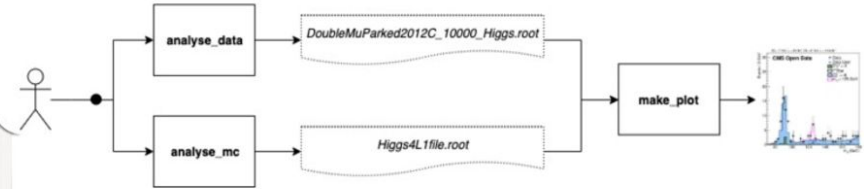
# What is REANA?



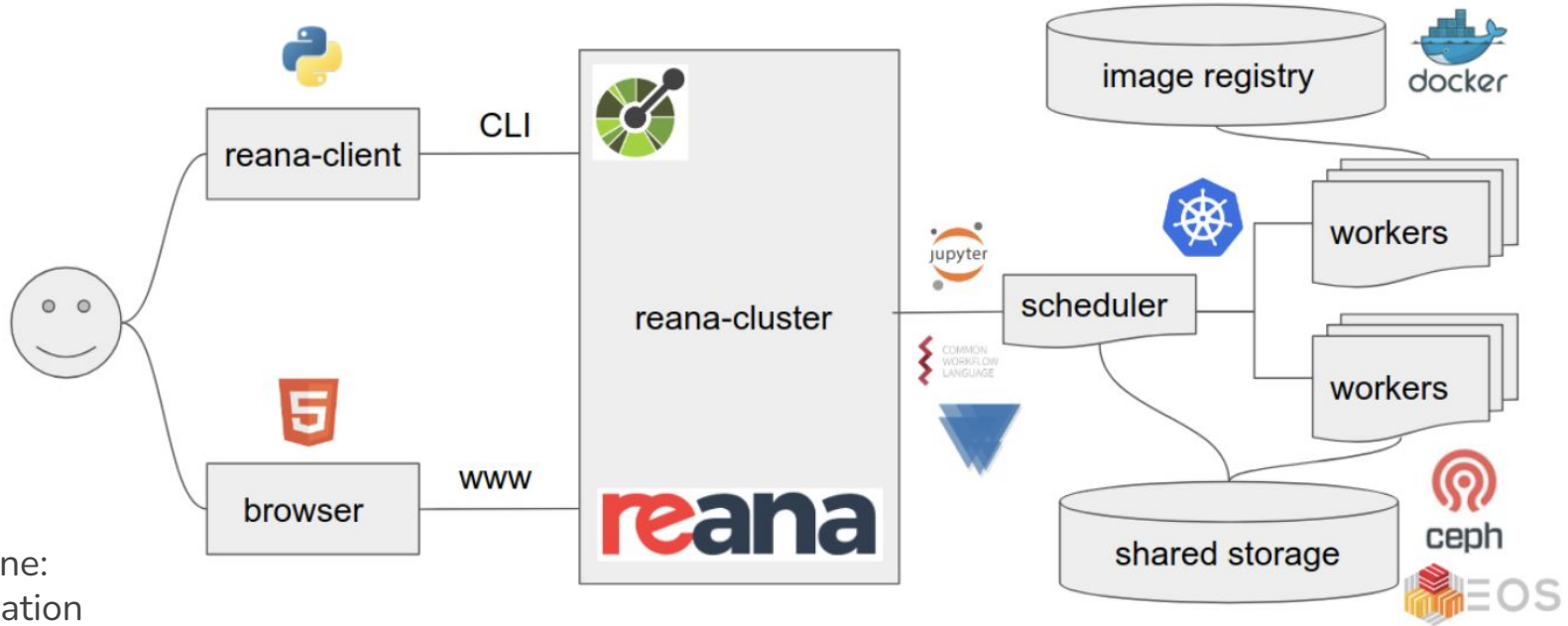
The screenshot shows the REANA website homepage. At the top, there is a navigation bar with links: Home, Examples, Get Started, Documentation, News, Roadmap, Contact, and Blog. The main heading is "reana" in a large, bold font, with "re" in red and "ana" in blue. Below the heading is the tagline "Reproducible research data analysis platform".

Four key features are listed in a row:

- Flexible**: Run many computational workflow engines. Logos for Common Workflow Language, S, and V are shown.
- Scalable**: Support for remote compute clouds. Logos for Kubernetes, HTCondor, and Slurm are shown.
- Reusable**: Containerise once, reuse elsewhere. Cloud-native. Logos for Docker and Singularity are shown.
- Free**: Free Software. MIT licence. Made with ❤ at CERN. The CERN logo is shown.



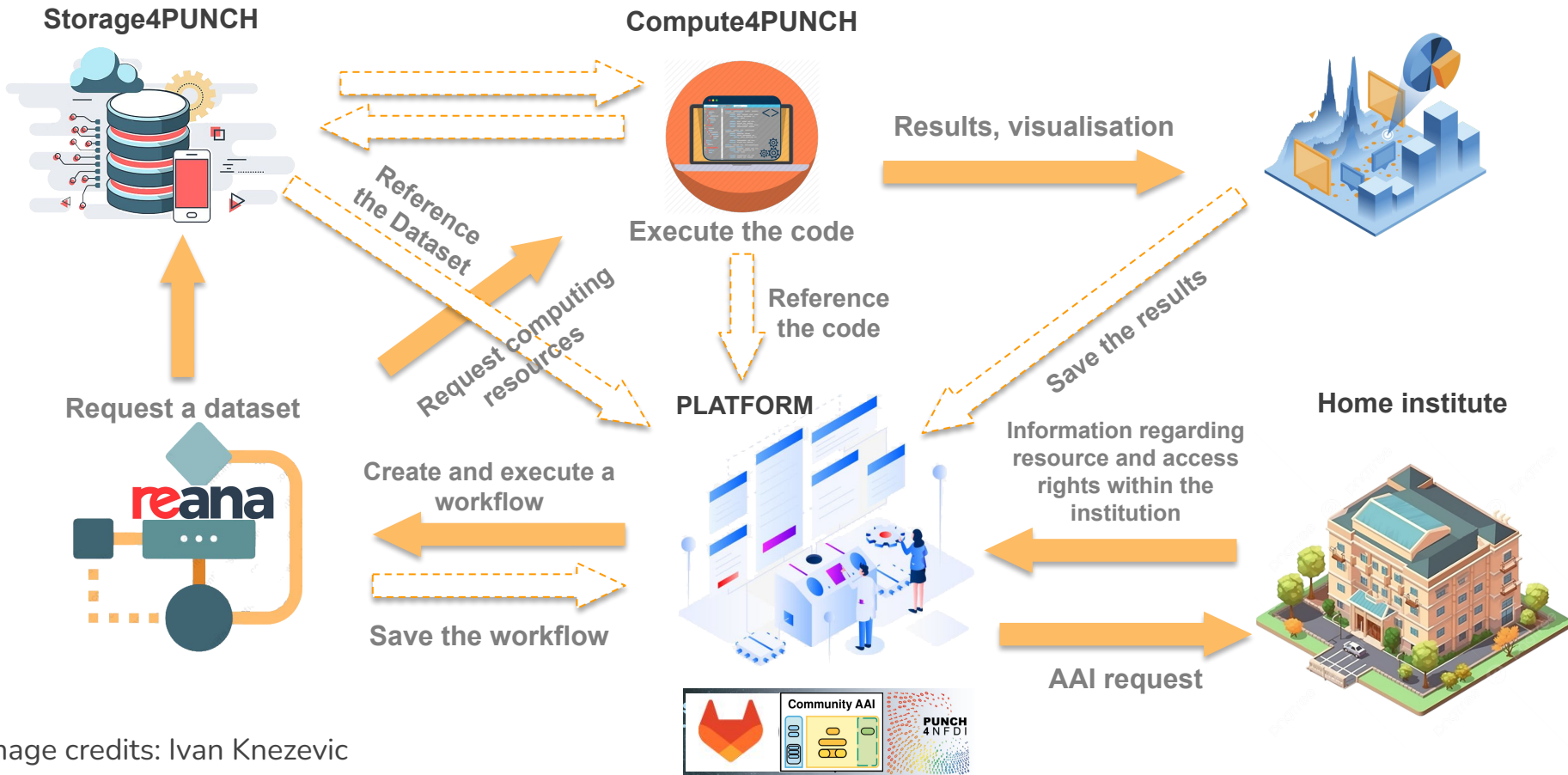
# What is REANA?

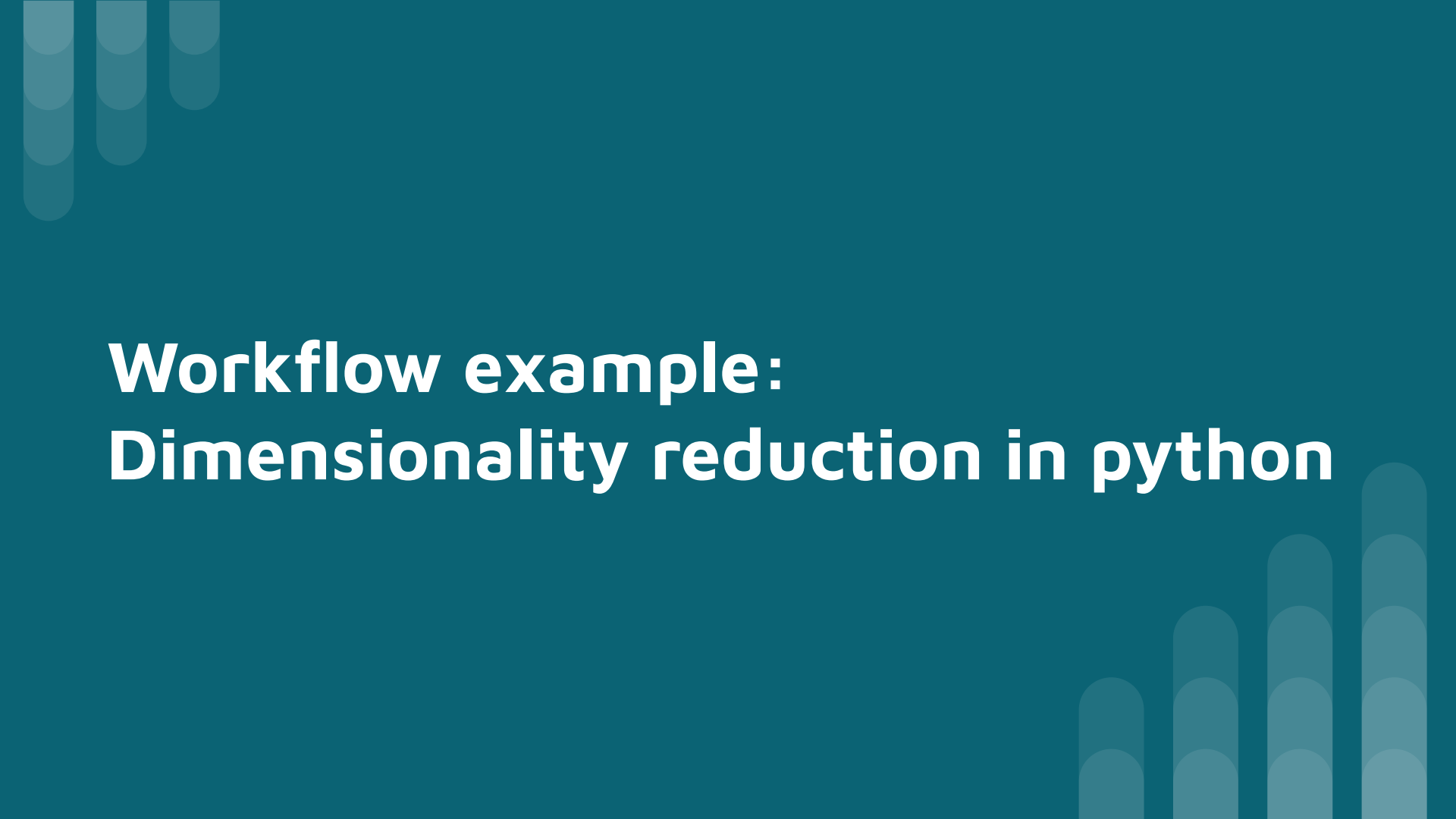


## Workflow Engine:

- authentication
- verification
- execution
  - environment
  - algorithm(s)
  - data I/O

# REANA integration in the PUNCH infrastructure





# **Workflow example: Dimensionality reduction in python**

# Dimensionality reduction – structure

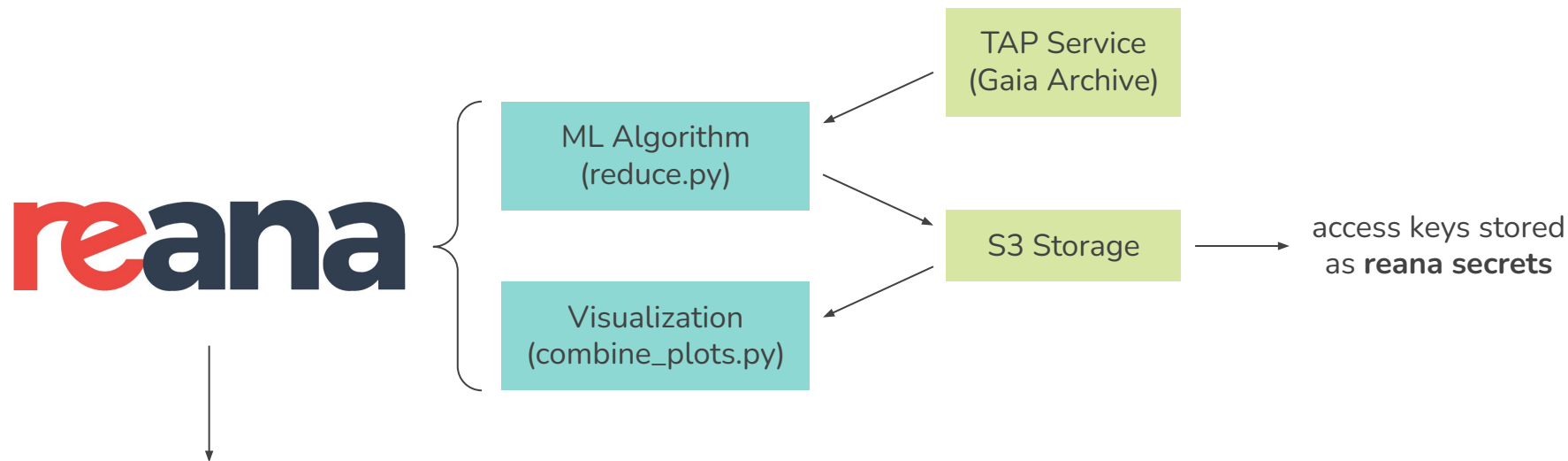
We show an example of 3 different algorithms to perform dimensionality reduction:

- UMAP (Uniform Manifold Approximation and Projection)
- PCA (Principal Component Analysis)
- t-SNE (t-distributed Stochastic Neighbor Embedding)

The data are managed between different pipelines using S3 private storage:

- download data from remote **TAP service**
- analyze and plot the results
- save and **upload to S3**
- **download** plots **from S3** if they exist
- combine them and save the result

# Dimensionality reduction – structure



- **yaml file** to organize the workflow (I/O, parameters, steps)
- **computational power**
- **memory**

# Dimensionality reduction – REANA secrets

To access private S3 storage, we use **reana-secrets**, a way to store tokens in REANA environment. For S3, we need two keys, that we can add with:

```
reana-client secrets-add --env access_key=XXX
```

```
reana-client secrets-add --env secret_key=XXX
```

Now we can call them within the python script with:

```
os.environ[ 'access_key' ]
```

```
os.environ[ 'secret_key' ]
```

# Dimensionality reduction – yaml file

The inputs are the 2 python scripts and some useful parameters:

```
inputs:  
  files:  
    - reduce.py  
    - combine_plots.py  
  parameters:  
    user_folder: new_user  
    n_test: 5
```



# Dimensionality reduction – yaml file

The inputs are the 2 python scripts and some useful parameters:

```
inputs:  
  files:  
    - reduce.py  
    - combine_plots.py  
  parameters:  
    user_folder: new_user  
    n_test: 5
```



**Customizable parameters**

# Dimensionality reduction – yaml file

The inputs are the 2 python scripts and some useful parameters:

```
inputs:  
  files:  
    - reduce.py  
    - combine_plots.py  
  parameters:  
    user_folder: new_user  
    n_test: 5
```

The final output is the combined pdf with all plots from the **n\_test** iterations:

```
outputs:  
  files:  
    - results/merged_plots.pdf
```

# Dimensionality reduction – yaml file

The data round trip is performed in 2 steps:

```
workflow:  
  type: serial  
  specification:  
    steps:  
      - name: make-projections  
        environment: 'gitlab-p4n.aip.de:5005/p4nreana/reana-env:py311-astro-ml.10134'  
        commands:  
          - mkdir -p results  
          - python reduce.py -d ${user_folder} -n ${n_test}  
      - name: combine-plots  
        environment: 'gitlab-p4n.aip.de:5005/p4nreana/reana-env:py311-astro-ml.10134'  
        commands:  
          - python combine_plots.py -d ${user_folder} -n ${n_test}
```

This runs the 3 algorithms  
and uploads the plots on S3

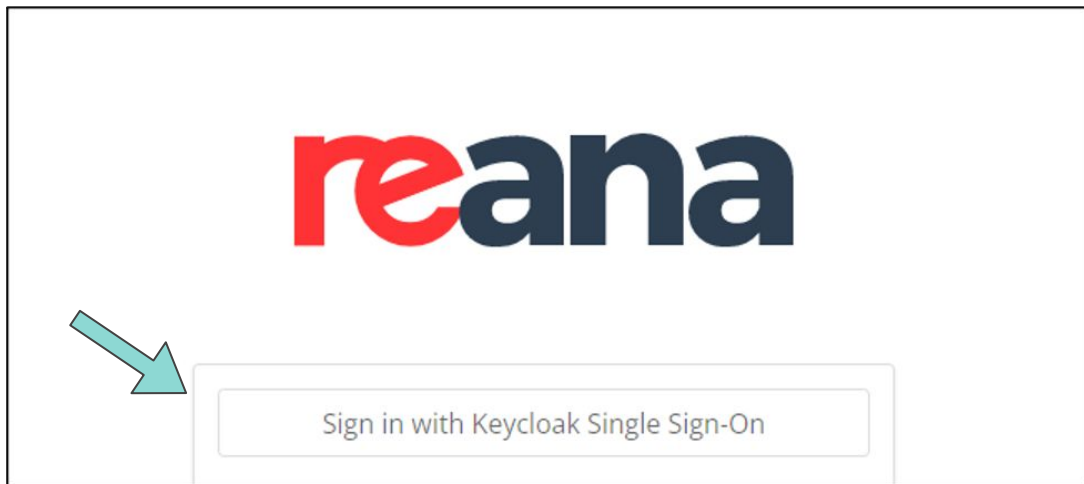
# Dimensionality reduction – yaml file

The data round trip is performed in 2 steps:

```
workflow:
  type: serial
  specification:
    steps:
      - name: make-projections
        environment: 'gitlab-p4n.aip.de:5005/p4nreana/reana-env:py311-astro-ml.10134'
        commands:
          - mkdir -p results
          - python reduce.py -d ${user_folder} -n ${n_test}
      - name: combine-plots
        environment: 'gitlab-p4n.aip.de:5005/p4nreana/reana-env:py311-astro-ml.10134'
        commands:
          - python combine_plots.py -d ${user_folder} -n ${n_test}
```

This grabs the plots from S3 and combines them

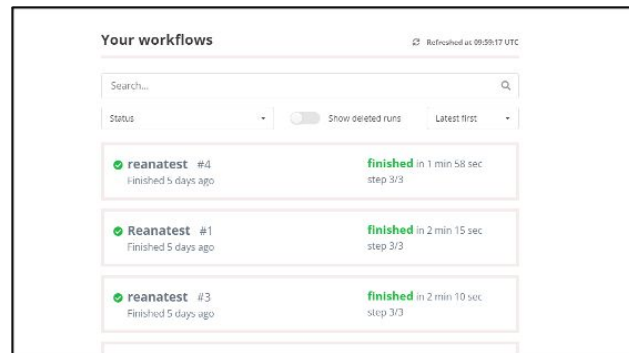
# REANA in practice



Launch from the terminal:

```
reana-client run -w <workflow-name>
```

```
reana-client download results
```



Launch from a URL:

## "Hello World!"

Launch on REANA



# REANA in practice

## Your workflows

Refreshed at 09:02:16 UTC

Search...

Status



Show deleted runs

Latest first

hello #8

35.5 KiB

Finished 3 days ago

finished in 28 seconds

step 2/2

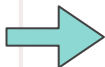
s3\_flow #2

1.93 MiB

Finished 15 days ago

finished in 1 min 22 sec

step 3/3



s3\_flow #2  
Finished 15 days ago

finished in 1 min 22 sec  
step 3/3

Engine logs

Job logs

Workspace

Specification

```
2024-04-10 09:29:03,523 | root | MainThread | INFO | Publishing step:0, cmd: mkdir -p results, total steps 2 to MQ
2024-04-10 09:29:12,665 | root | MainThread | INFO | Publishing step:0, cmd: python reduce.py -d esacchi -n 5, total steps 2 to MQ
2024-04-10 09:30:15,959 | root | MainThread | INFO | Publishing step:1, cmd: python combine_plots.py -d esacchi -n 5, total steps 2 to MQ
2024-04-10 09:30:24,989 | root | MainThread | INFO | Workflow 053c6fb5-ba57-4a66-9357-5fab558421b8 finished. Files available at /var/reana/users/dfabb887-0dbd-4bdd-a1a2-8a234ec5552a/workflows/053c6fb5-ba57-4a66-9357-5fab558421b8.
```

# REANA in practice

## Your workflows

Refreshed at 09:02:16 UTC

Search...

Status



Show deleted runs

Latest first

hello #8

35.5 KiB

Finished 3 days ago

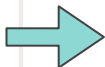
finished in 28 seconds  
step 2/2

s3\_flow #2

1.93 MiB

Finished 15 days ago

finished in 1 min 22 sec  
step 3/3



s3\_flow #2

Finished 15 days ago

finished in 1 min 22 sec  
step 3/3

Engine logs

Job logs

Workspace

Specification

Step make-projections

finished in 1 min 3 sec

Kubernetes

gitlab-p4n.aip.de:5005/p4nreana/re...

\$ python reduce.py -d esacchi -n 5

```
t-SNE done
Successfully uploaded projections_comparison_3.png to S3!

Random seed = 635
UMAP done
PCA done
t-SNE done
Successfully uploaded projections_comparison_4.png to S3!

Random seed = 22
UMAP done
PCA done
t-SNE done
Successfully uploaded projections_comparison_5.png to S3!

Completed
```

# REANA in practice

## Your workflows

Refreshed at 09:02:16 UTC

Search...

Status  Show deleted runs Latest first

hello #8

35.5 KiB

Finished 3 days ago

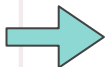
finished in 28 seconds  
step 2/2

s3\_flow #2

1.93 MiB

Finished 15 days ago

finished in 1 min 22 sec  
step 3/3



s3\_flow #2  
Finished 15 days ago

finished in 1 min 22 sec  
step 3/3

Engine logs Job logs Workspace Specification

Search...

Name	Modified	Size
reduce.py	2024-04-10T09:28:51	3.24 KiB
combine_plots.py	2024-04-10T09:28:51	1.03 KiB
reana.yaml	2024-04-10T09:28:50	634 Bytes
results/projections_comparison_4.png	2024-04-10T09:30:17	220.39 KiB
results/projections_comparison_3.png	2024-04-10T09:30:17	215.62 KiB
results/merged_plots.pdf	2024-04-10T09:30:18	513.46 KiB
results/projections_comparison_2.png	2024-04-10T09:30:17	216.01 KiB
results/projections_comparison_5.png	2024-04-10T09:30:18	214.25 KiB
results/projections_comparison_1.png	2024-04-10T09:30:17	216.89 KiB



# REANA in practice

**finished** in 1 min 22 sec  
step 3/3

Open Jupyter Notebook



Delete workflow



 jupyter

Files

Running

Clusters

Select items to perform actions on them.

0

▼ /

results

combine\_plots.py

reana.yaml

reduce.py

To open Jupyter Notebooks from command line and custom image:

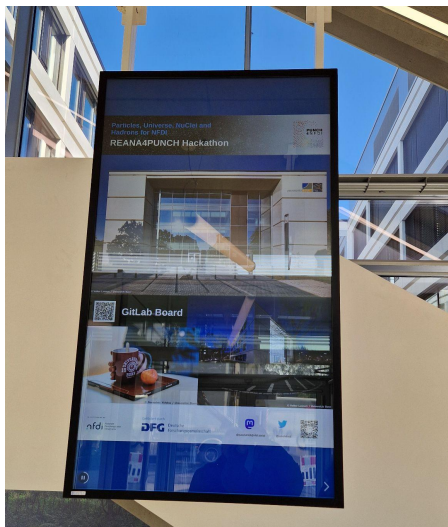
```
reana-client open -w <WF-name> -i <image-name> jupyter
```



**Further developments...**



# REANA Hackathon in Bonn, April 22 & 23, 2024



Bonn University  
Physics Institute

Hackthon Work & Discussion Items Search Show labels Edit board Create list

- Discussion Items** (6 items)
  - PUNCH: Resources sharing between HPC-REANA-Storage** (#16)
  - Quota policy-CPU,GPU,RAM,Disk: How to implement projects , groups and users? How much?** (#15)
  - Selecting HW: GPU or CPU** (#14)
  - Allow users to select Compute-backends: Slurm,HTCondor, kubernetes(default)** (#13)
  - How to currate data between SLURM clusters and REANA** (#12)
  - Data curation and communication: S3(AIP or external) and S4P** (#11)
- Issues** (0 items)
- Closed** (7 items)
  - Re-run REANA tutorial reading from and writing to S4P** (Hackathon Work Items) (#9)
  - Deploy REANA pre-production instance at AIP** (Hackathon Work Items) (#1)
  - Adjust REANA tutorial to stage-out results to S4P** (Hackathon Work Items) (#8)
  - Re-run REANA tutorial reading from S4P** (Hackathon Work Items) (#7)
  - Adjust REANA tutorial to read data from S4P Storage** (Hackathon Work Items) (#6)
  - Current issues with AAI+AIP+Keycloak: Does not work** (Issues) (#10)

# REANA Hackathon in Bonn, April 22 & 23, 2024

- REANA pre-production instance

- Keycloak authentication

prepared on AIP Kubernetes Cluster  
running reana-server 0.9.2a

- Integration with Compute4PUNCH

- HTCondor backend to run jobs on C4P

HTCondor interface (KIT):  
reana-job-controller:0.9.2a

- Storage4PUNCH

- Access management
- Data flow from/to S4P

OIDC token management integrated into  
the reana workflow

- Testing ongoing...

Integration of reana fork into  
reana main branch ongoing

# REANA Hackathon in Bonn, April 22 & 23, 2024



✓ reana-cern-open-data-tutorial #8

Finished a few seconds ago

finished in 12 min 18 sec

step 3/3

🔧 Engine logs > Job logs Workspace Specification

Step stage\_out

finished in 1 min 54 sec

compute4punch

wlwg:wn:latest

\$ chmod +x code/stage\_out.sh && code...

```
Copying 16177 bytes file:///srv/results/higgs_2el2mu.pdf => https://dcache-desy-webdav.desy.de:2880//pnfs/desy.de/punch/tutorials/reana-cern-
open-data-tutorial/9b761fb1-9ab5-4d46-88e5-86195aaca4cc/higgs_2el2mu.pdf
event: [1713865981742] BOTH GFAL2:CORE:COPY LIST:ENTER
event: [1713865981743] BOTH GFAL2:CORE:COPY LIST:ITEM file:///srv/results/higgs_2el2mu.pdf => https://dcache-desy-
webdav.desy.de:2880//pnfs/desy.de/punch/tutorials/reana-cern-open-data-tutorial/9b761fb1-9ab5-4d46-88e5-86195aaca4cc/higgs_2el2mu.pdf
event: [1713865981743] BOTH GFAL2:CORE:COPY LIST:EXIT
event: [1713865981743] BOTH http_plugin PREPARE:ENTER file:///srv/results/higgs_2el2mu.pdf => https://dcache-desy-
webdav.desy.de:2880//pnfs/desy.de/punch/tutorials/reana-cern-open-data-tutorial/9b761fb1-9ab5-4d46-88e5-86195aaca4cc/higgs_2el2mu.pdf
event: [1713865981808] BOTH http_plugin PREPARE:EXIT file:///srv/results/higgs_2el2mu.pdf => https://dcache-desy-
webdav.desy.de:2880//pnfs/desy.de/punch/tutorials/reana-cern-open-data-tutorial/9b761fb1-9ab5-4d46-88e5-86195aaca4cc/higgs_2el2mu.pdf
event: [1713865981808] BOTH http_plugin TRANSFER:ENTER file:///srv/results/higgs_2el2mu.pdf => https://dcache-desy-
webdav.desy.de:2880//pnfs/desy.de/punch/tutorials/reana-cern-open-data-tutorial/9b761fb1-9ab5-4d46-88e5-86195aaca4cc/higgs_2el2mu.pdf
event: [1713865981808] BOTH http_plugin TRANSFER:TYPE streamed
event: [1713865982054] BOTH http_plugin TRANSFER:EXIT file:///srv/results/higgs_2el2mu.pdf => https://dcache-desy-
webdav.desy.de:2880//pnfs/desy.de/punch/tutorials/reana-cern-open-data-tutorial/9b761fb1-9ab5-4d46-88e5-86195aaca4cc/higgs_2el2mu.pdf
Copying 9764 bytes file:///srv/results/higgs_2el2mu.root => https://dcache-desy-webdav.desy.de:2880//pnfs/desy.de/punch/tutorials/reana-cern-
open-data-tutorial/9b761fb1-9ab5-4d46-88e5-86195aaca4cc/higgs_2el2mu.root
```

# REANA Hackathon in Bonn, April 22 & 23, 2024

- REANA pre-production instance
  - Keycloak authentication
- Integration with Compute4PUNCH
  - HTCondor backend to run jobs on C4P
- Storage4PUNCH
  - Access management
  - Data flow from/to S4P
- Testing ongoing...



The Hackers assembly, away from the keyboard!