V irtual
R esearch
E nvironment

A collaborative platform to serve astro-particle communities

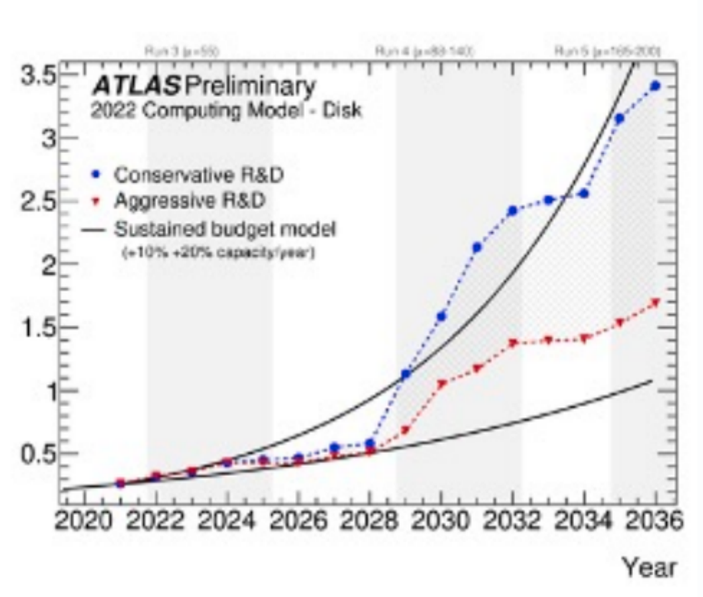
Elena Gazzarrini, Alba Vendrell Moya, Enrique Garcia, Agisilaos Kounelis, Domenic Gosein, Xavier Espinal

Challenges of today

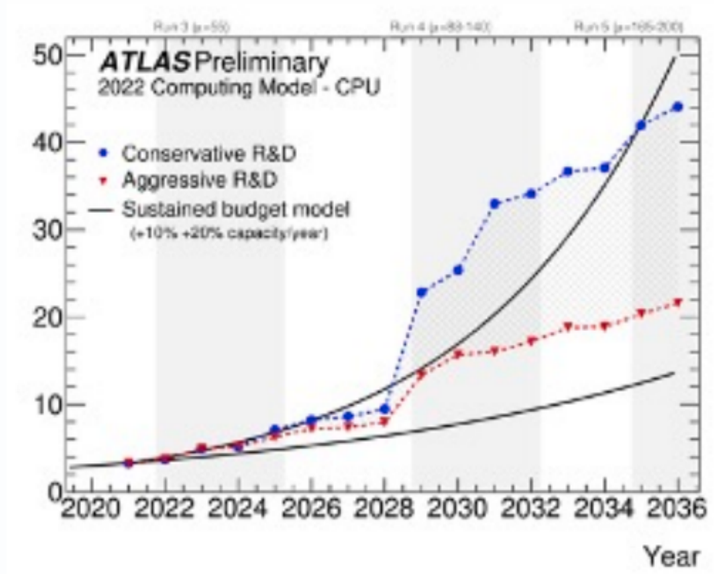
Astro-particle physics communities around the world are quickly increasing **storage** and **CPU** needs, and scientific collaborations are growing.

TECHNOLOGY	SCIENCE
exabyte scale needs	more interest in collaboration
unified way of sharing data and software	cross-disciplinary research

DISK space (EB)



Annual CPU consumption (MHS06years)



The **Virtual Research Environment** is a collaborative analysis platform where researchers from various scientific communities are able to develop and share end-to-end workflows, having access to all the digital content needed to produce a scientific result in compliance with FAIR principles.

An analysis platform?

Yes! A place to run an end-to-end workflow.

Generation and simulation of events

Experimental data

An analysis platform?

Yes! A place to run an end-to-end workflow.

Generation and simulation of events

Experimental data

Data processing

Prepare for analysis

An analysis platform?

Yes! A place to run an end-to-end workflow.

Generation and simulation of events

Experimental data

Data processing

Prepare for analysis

Analysis

Preserved software

An analysis platform?

Yes! A place to run an end-to-end workflow.

Generation and simulation of events

Experimental data

Data processing

Prepare for analysis

Analysis

Preserved software

Interpretation of results

Combination of results and comparison across workflows

An analysis platform?

Yes! A place to run an end-to-end workflow.

Generation and simulation of events

Experimental data

Data processing

Prepare for analysis

Analysis

Preserved software

Interpretation of results

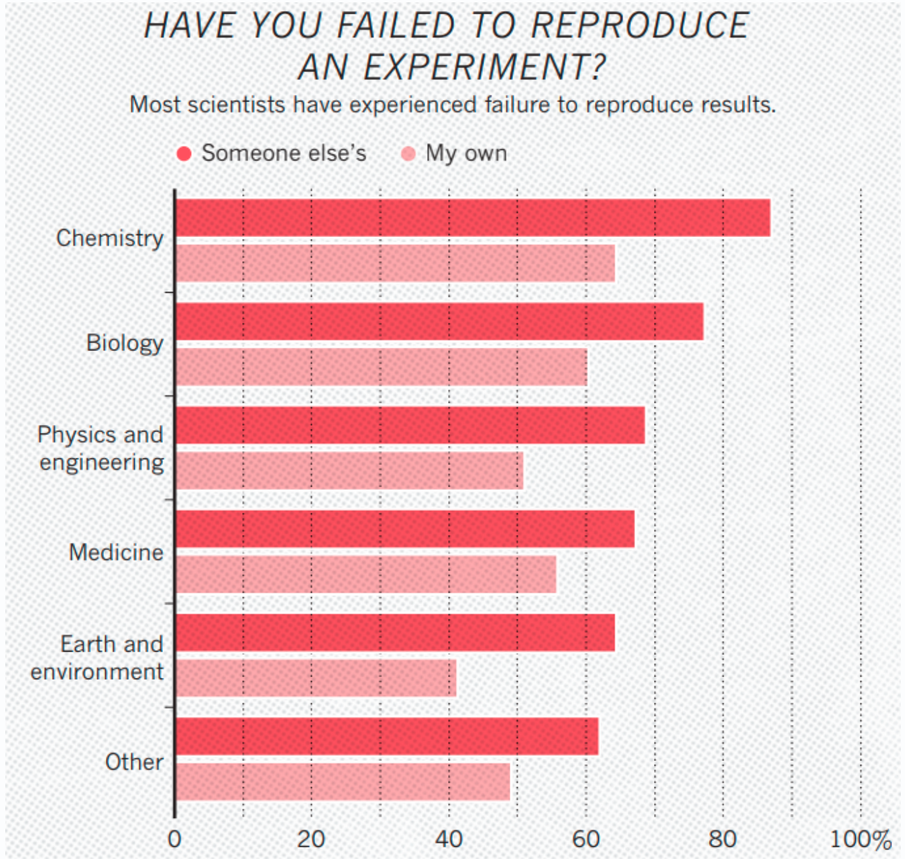
Combination of results and comparison across workflows

Preservation of analysis

For later reuse or reanalysis

An analysis platform?

Yes! A place to run an end-to-end workflow.



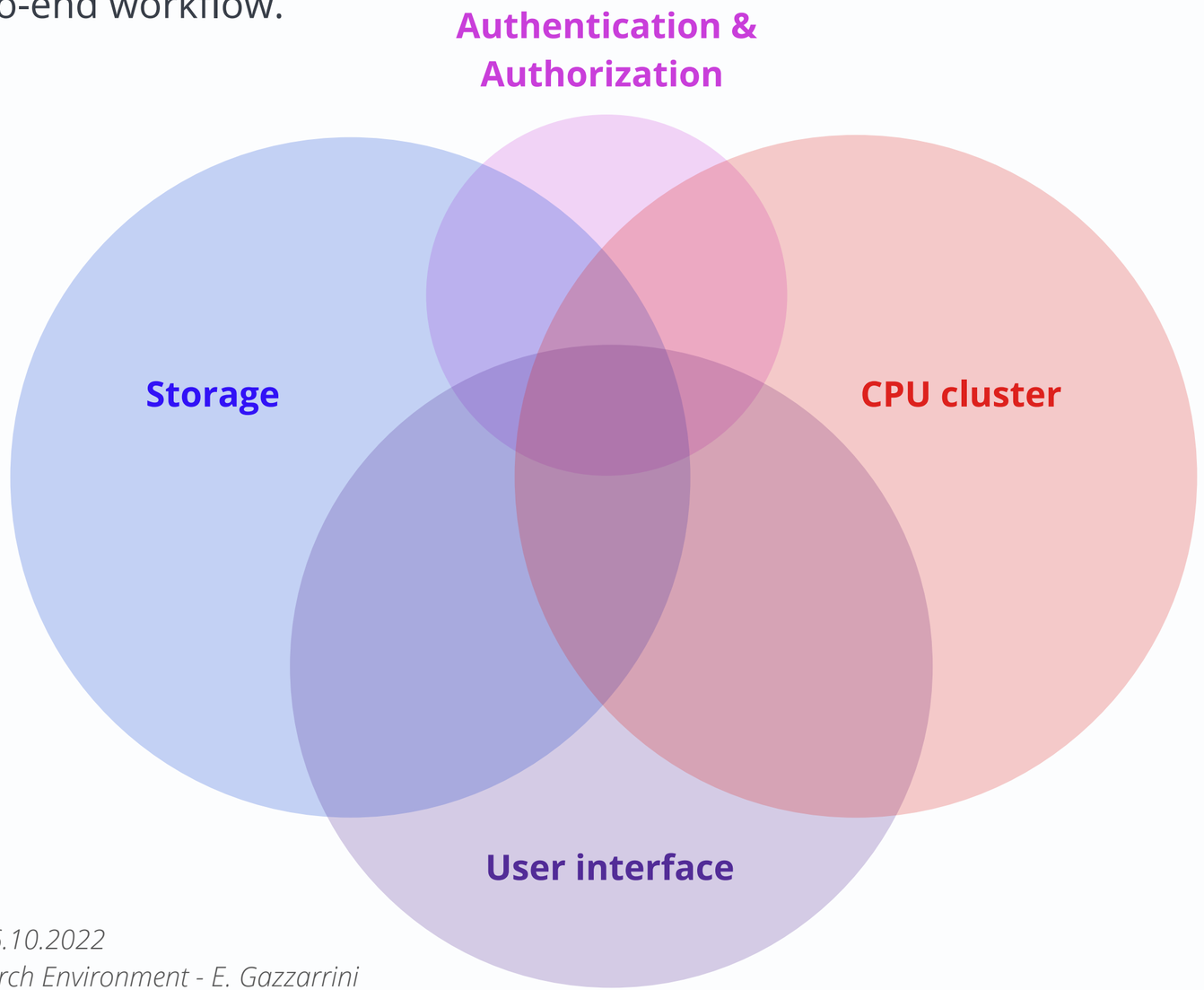
Nature 533, 452–454 (2016). https://doi.org/10.1038/533452a

Preservation of analysis

For later reuse or reanalysis

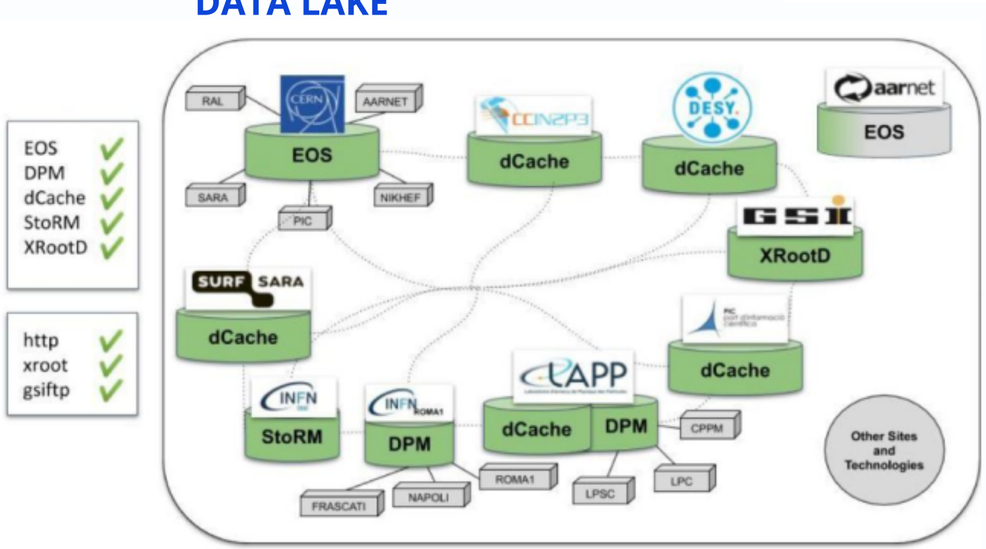
An analysis platform?

Yes! A place to run an end-to-end workflow.



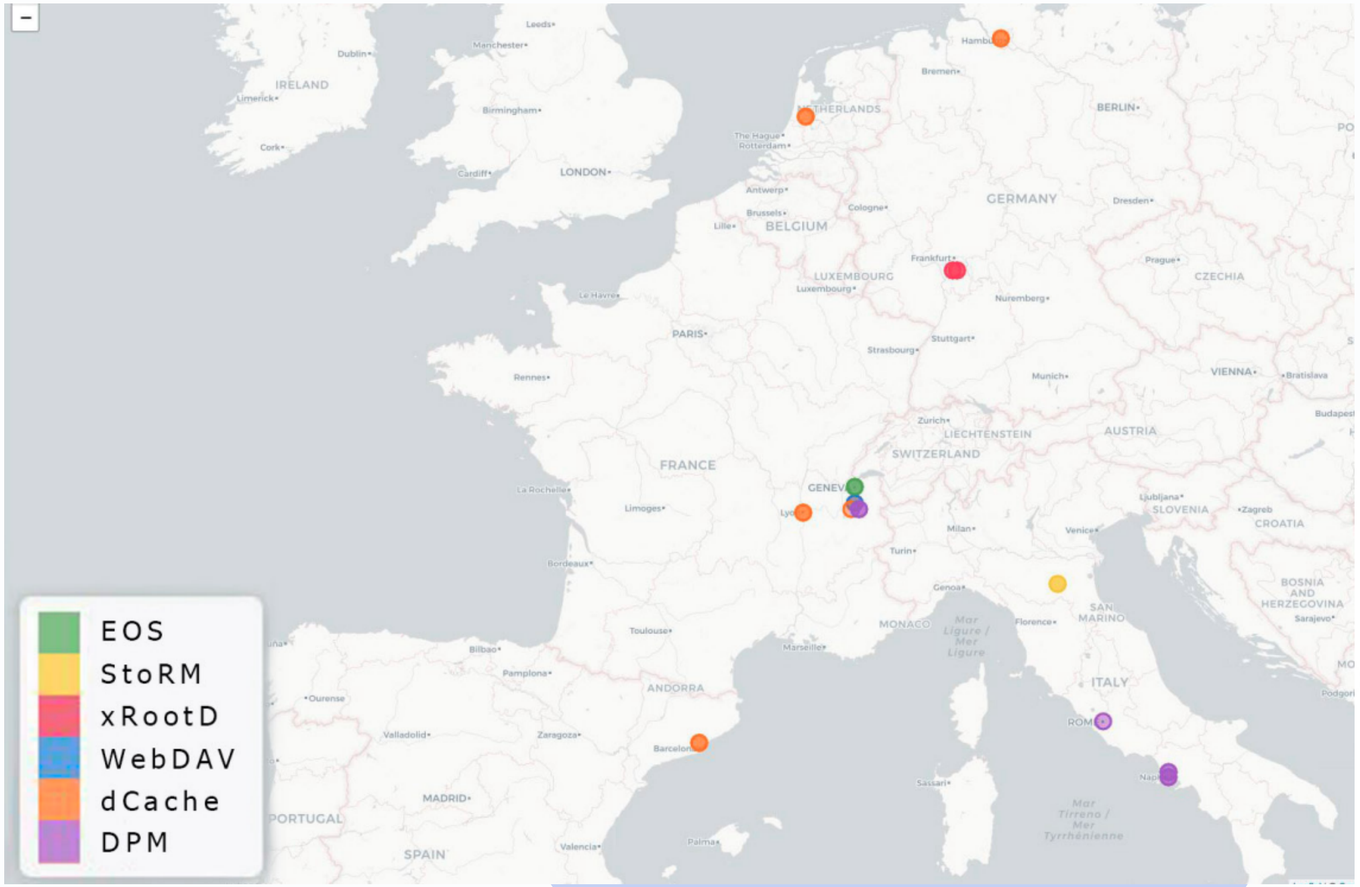
Storage

DATA LAKE



- distributed storage on WLCG model
- support for all data types
- ensures data security, quality, access
- experiments *dump* data... scientists *fish* data
- data access and transfer with gsiftp, http(s), and xrootd protocols

ESCAPE Data Lake -10.1051/epjconf/202125102056

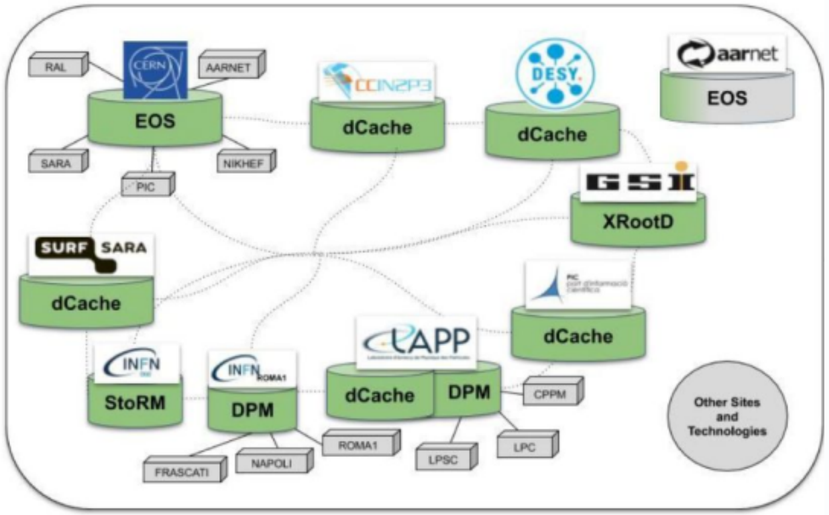


Storage

EOS ✓
DPM ✓
dCache ✓
StoRM ✓
XRootD ✓

http ✓
xroot ✓
gsiftp ✓

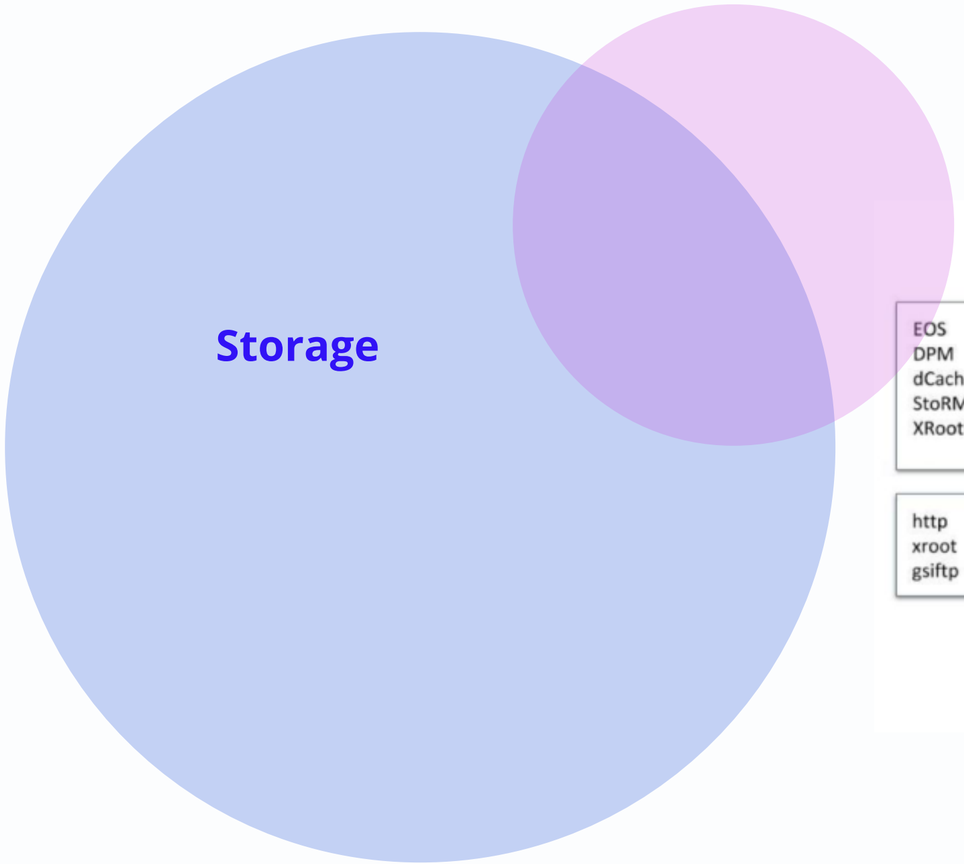
DATA LAKE



- distributed storage on WLCG model
- support for all data types
- ensures data security, quality, access
- experiments *dump* data... scientists *fish* data
- data access and transfer with gsiftp, http(s), and xrootd protocols

ESCAPE Data Lake -10.1051/epjconf/202125102056

Authentication & Authorization



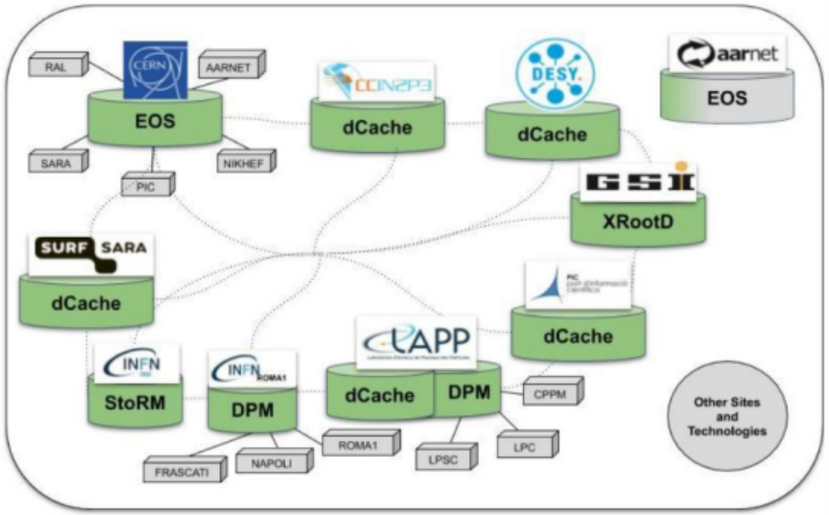
EOS ✓
DPM ✓
dCache ✓
StoRM ✓
XRootD ✓

http ✓
xroot ✓
gsiftp ✓

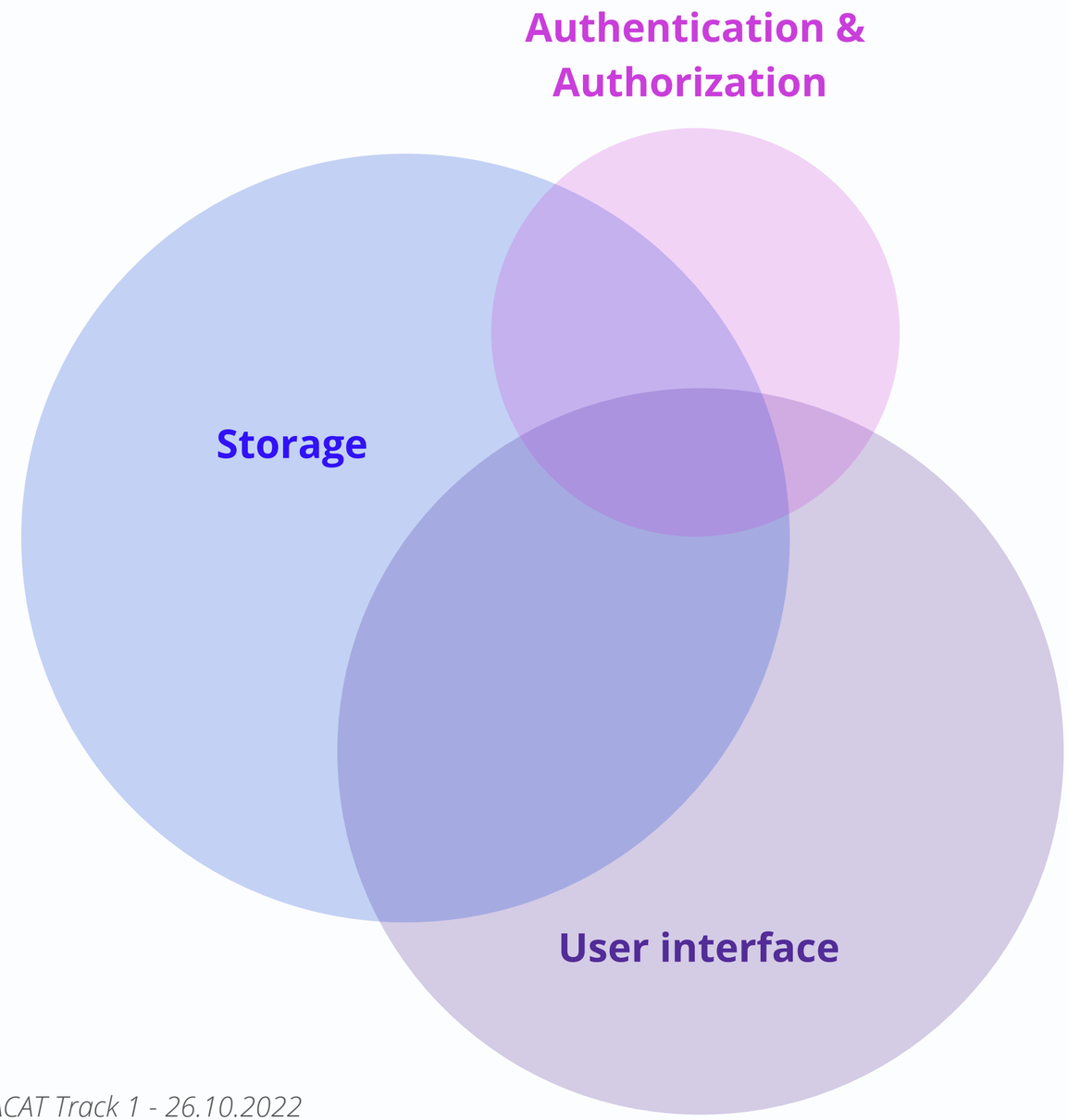
Identity provider via:

- x509 certificates
- OIDC tokens

DATA LAKE

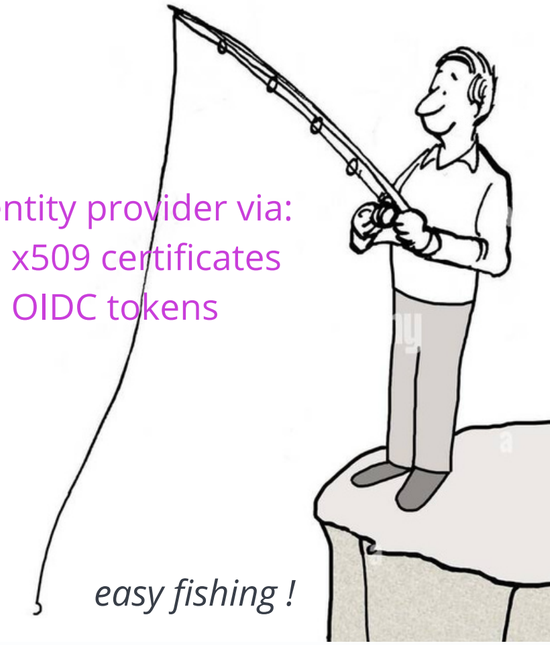


- distributed storage on WLCG model
- support for all data types
- ensures data security, quality, access
- experiments *dump* data... scientists *fish* data
- data access and transfer with gridFTP, HTTP(S), and Xrootd protocols



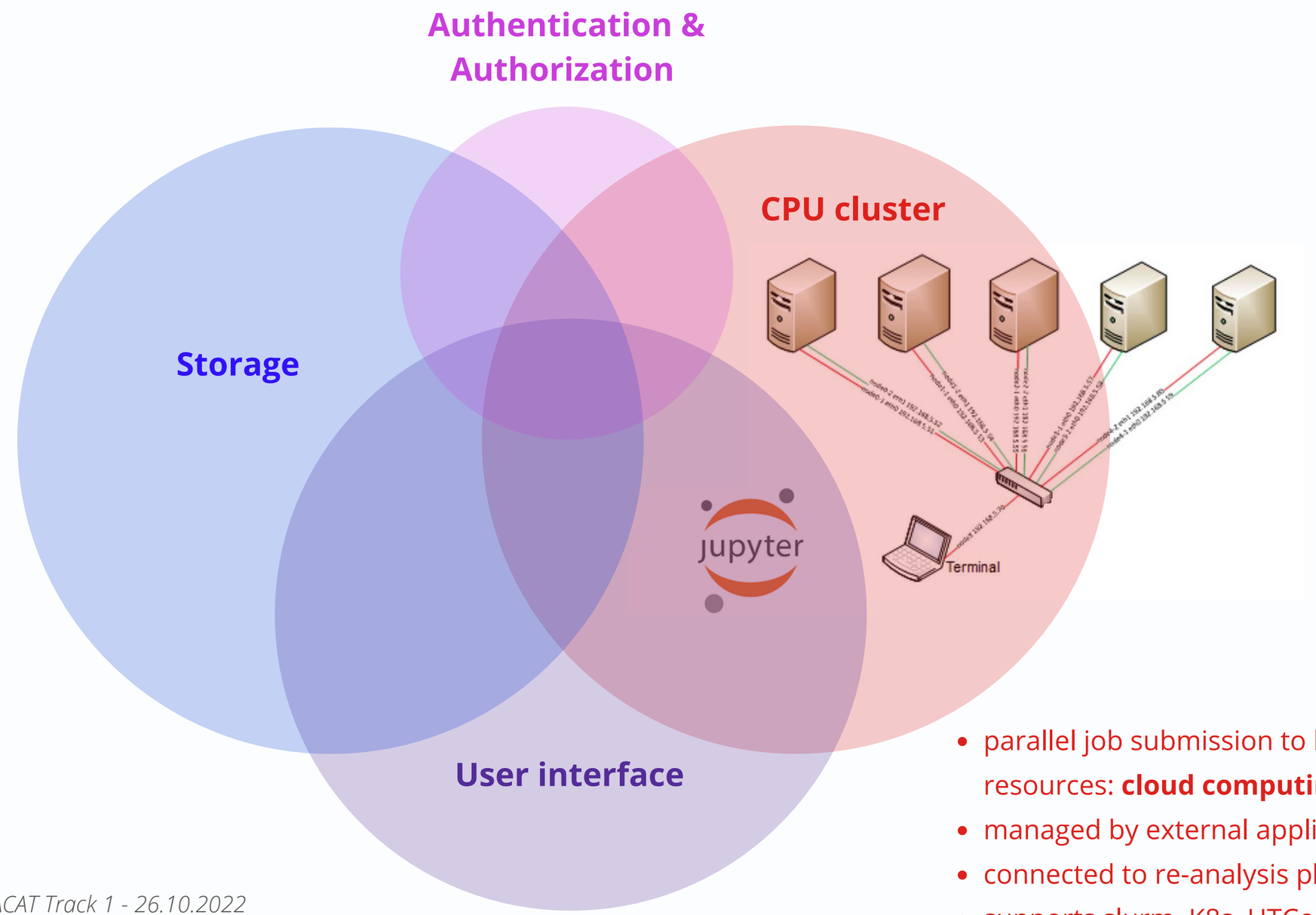
Identity provider via:

- x509 certificates
- OIDC tokens



DATA LAKE AS A SERVICE

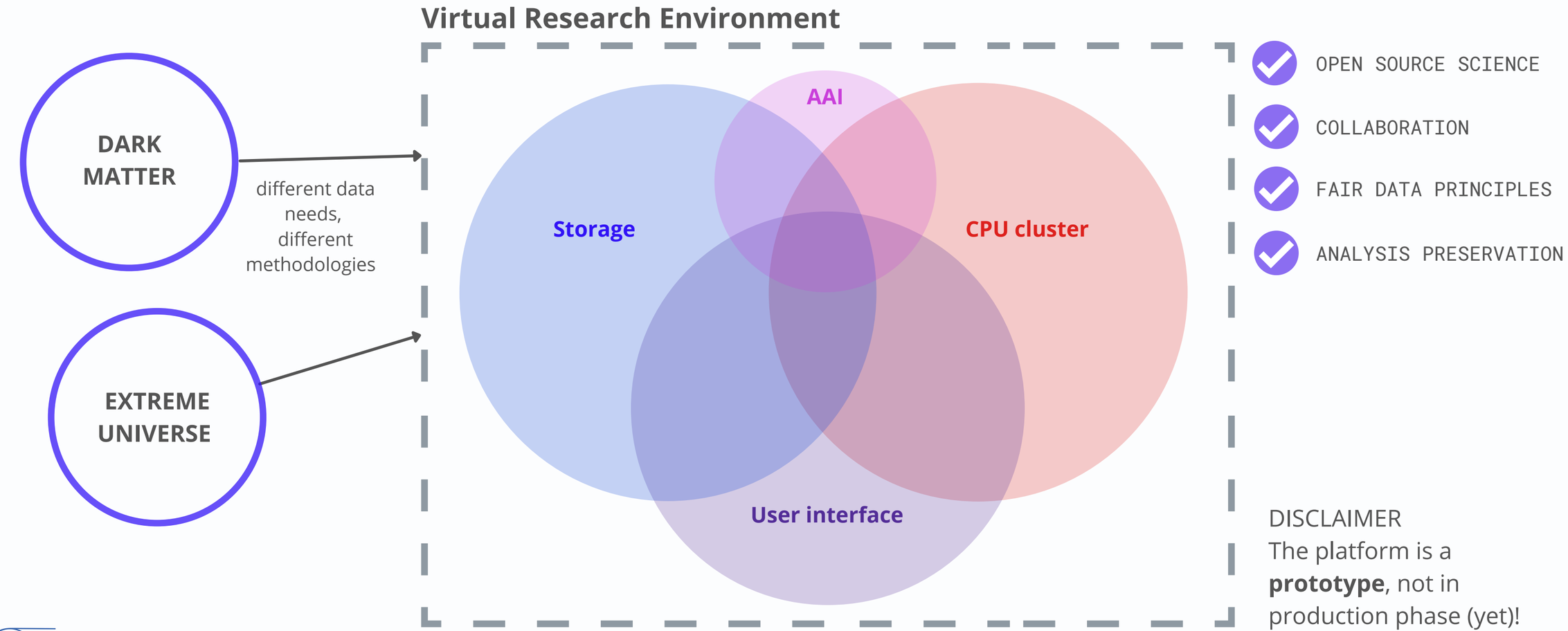
- hides complexity of Data Lake - scientists browse data
- pre-installed software
- run preliminary analysis
- send batch/parallel jobs to compute cluster (no need of local file copy)



- parallel job submission to heterogeneous resources: **cloud computing** and **HPCs**
- managed by external applications
- connected to re-analysis platform
- supports slurm, K8s, HTCondor

Context: EOSC-Future

European Open Science Cloud project to demonstrate the use of collaborative tools within astro-particle research (2 Science Projects). Science driver for other communities.



DISCLAIMER
The platform is a **prototype**, not in production phase (yet)!

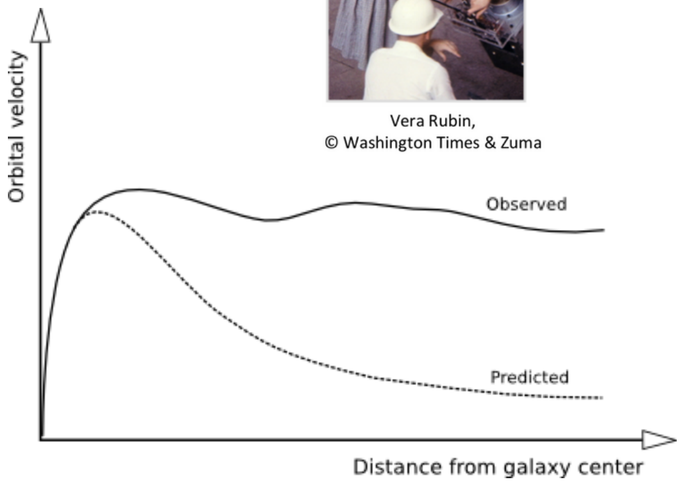
SCIENTIFIC MOTIVATION

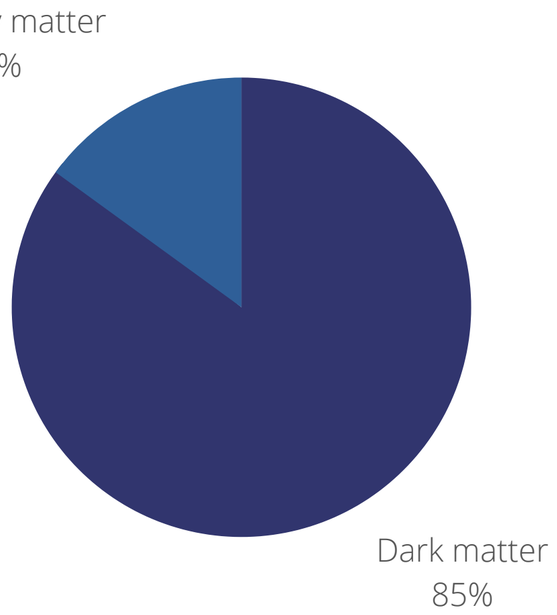
Why is the VRE useful to the astro-particle physics community?

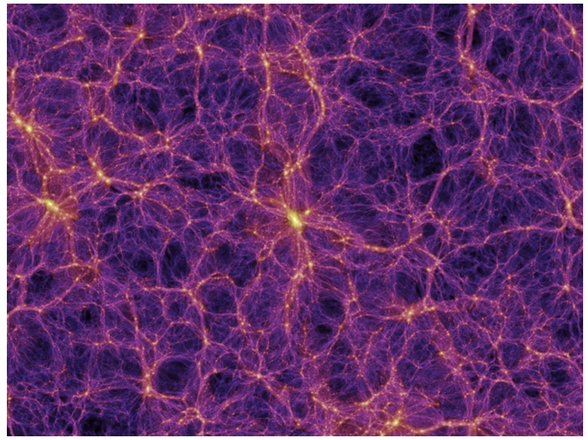
Science Project 1: Dark Matter

→ *Galaxy rotation curves* --> a larger amount of gravitational mass is expected to exist in the universe.
→ It does not interact with the electromagnetic filed and *cannot therefore be seen.*
→ Many DM candidates. Many experiments target the problem. Many different research approaches.

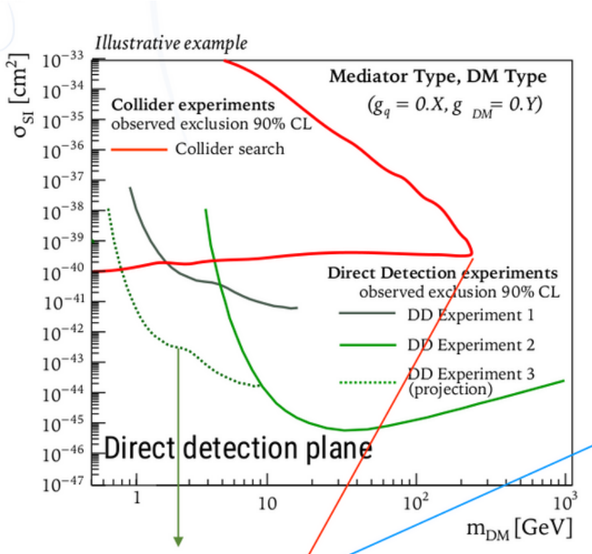
Vera Rubin,
© Washington Times & Zuma

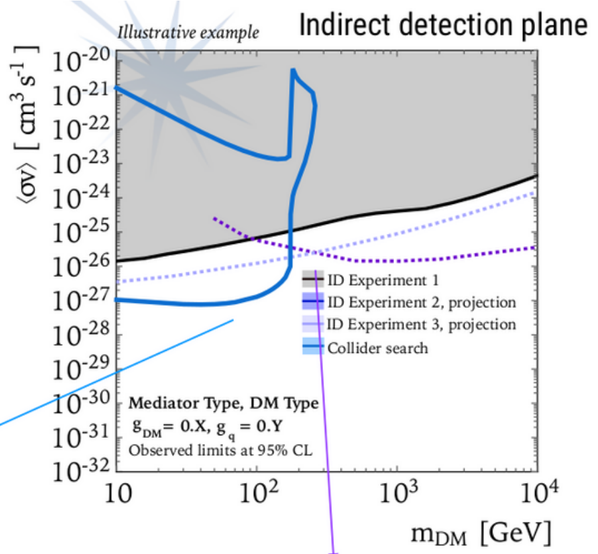




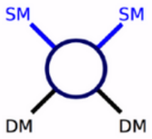
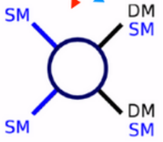
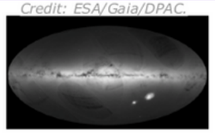
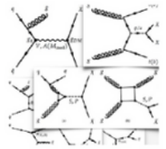
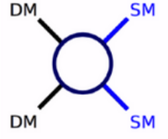


Expected outcomes





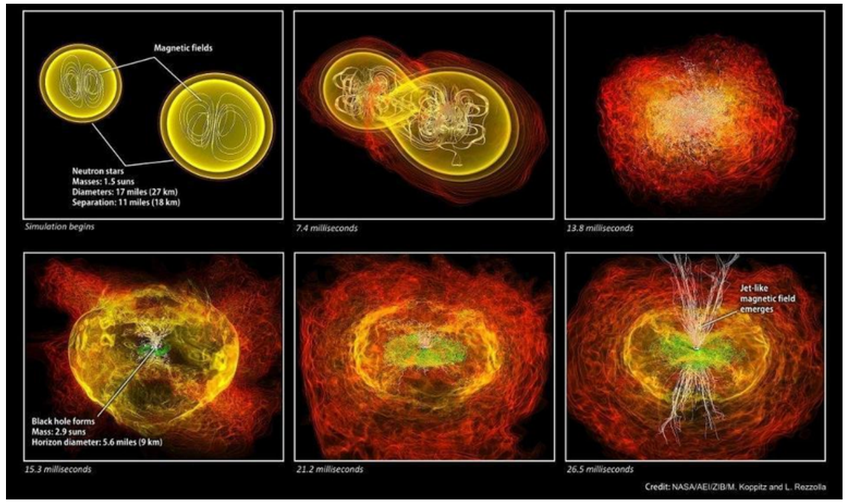
*Objective is to collect all the **digital objects** + workflows in a cohesive way, output **combined plots** and provide an interdisciplinary open science example from **bottom-up effort***

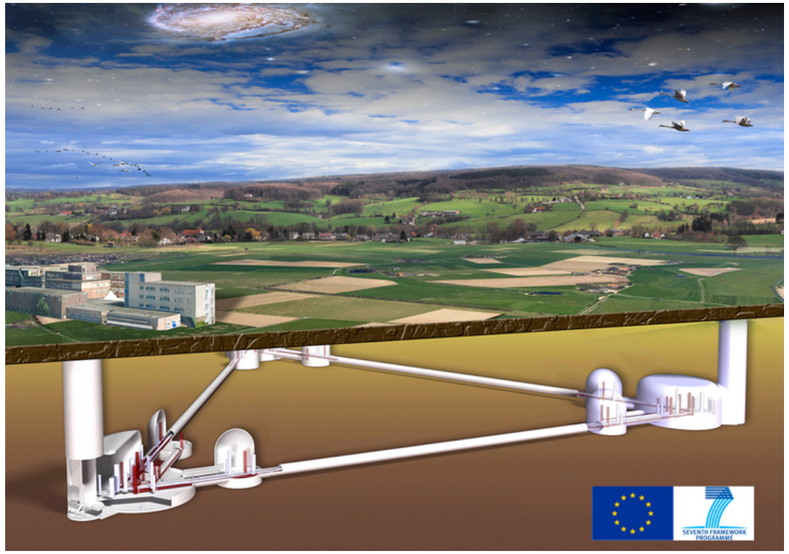
METHODS	 Direct Detection	 Colliders	 Astrophysics	 Theory	 Indirect Detection
RELEVANCE	DM that interacts inside the detector (WIMPs, axions)	produce DM and probe the dark interaction	necessary for all	necessary for all	detect annihilating/ decaying DM through its decays (i.e. neutrino searches, gamma rays)
EXPERIMENTS INVOLVED	DARKSIDE	ATLAS EXPERIMENT			KM3NeT, cta, FermiLAT

Science Project 2: Extreme Universe

➔ **Multi-messenger astronomy** (i.e. EM radiation, GW, neutrinos, cosmic rays are created by different astrophysical processes, and thus reveal different information about their sources)

INPUT DATA	Binary Neutron Star Merger	Active Galactic Nuclei	Core-Collapse Supernovae
STUDY	• GW • Fast Radio Bursts • Broadband follow ups	multi wavelength observations	• Neutrinos • GW
EXPERIMENT	VIRGO LOFAR	 FermiLAT	VIRGO KM3NeT





WHAT IS NEEDED?

To create a collaborative platform, useful for the sciences?

Components

FROM SCIENTISTS / EXPERIMENTS

FROM IT DEVELOPERS

DATA

uploaded & downloaded →

Flexible data management system to preserve data quality and integrity, INDEPENDENT from local storage

=

Data Lake
a cluster of virtual resources to manage storage elements, user subscriptions and monitoring of all data operations

SCIENTIFIC DATA MANAGEMENT
RUCIO

Components

FROM SCIENTISTS / EXPERIMENTS

FROM IT DEVELOPERS

DATA — *uploaded & downloaded* → Flexible data management system to preserve data quality and integrity, INDEPENDENT from local storage

SOFTWARE — *uploaded & synchronised* → Open source sofware repository

=

Data Lake
a cluster of virtual resources to manage storage elements, user subscriptions and monitoring of all data operations

SCIENTIFIC DATA MANAGEMENT
RUCIO

Components

FROM SCIENTISTS / EXPERIMENTS

FROM IT DEVELOPERS

DATA	*uploaded & downloaded* →	Flexible data management system to preserve data quality and integrity, INDEPENDENT from local storage
SOFTWARE	*uploaded & synchronised* →	Open source sofware repository
ENVIRONMENT	*uploaded & synchronised* →	Docker images spawned when user starts notebook

=

Data Lake
a cluster of virtual resources to manage storage elements, user subscriptions and monitoring of all data operations

=

CernVM File system, GitLab, GitHub, zenodo, docker

Components

FROM SCIENTISTS / EXPERIMENTS

FROM IT DEVELOPERS

DATA	*uploaded & downloaded* →	Flexible data management system to preserve data quality and integrity, INDEPENDENT from local storage
SOFTWARE	*uploaded & synchronised* →	Open source sofware repository
ENVIRONMENT	*uploaded & synchronised* →	Docker images spawned when user starts notebook
STEPS	*uploaded & synchronised* →	Declarative format file to keep track of the workflow (.yaml)

=

Data Lake
a cluster of virtual resources to manage storage elements, user subscriptions and monitoring of all data operations

=

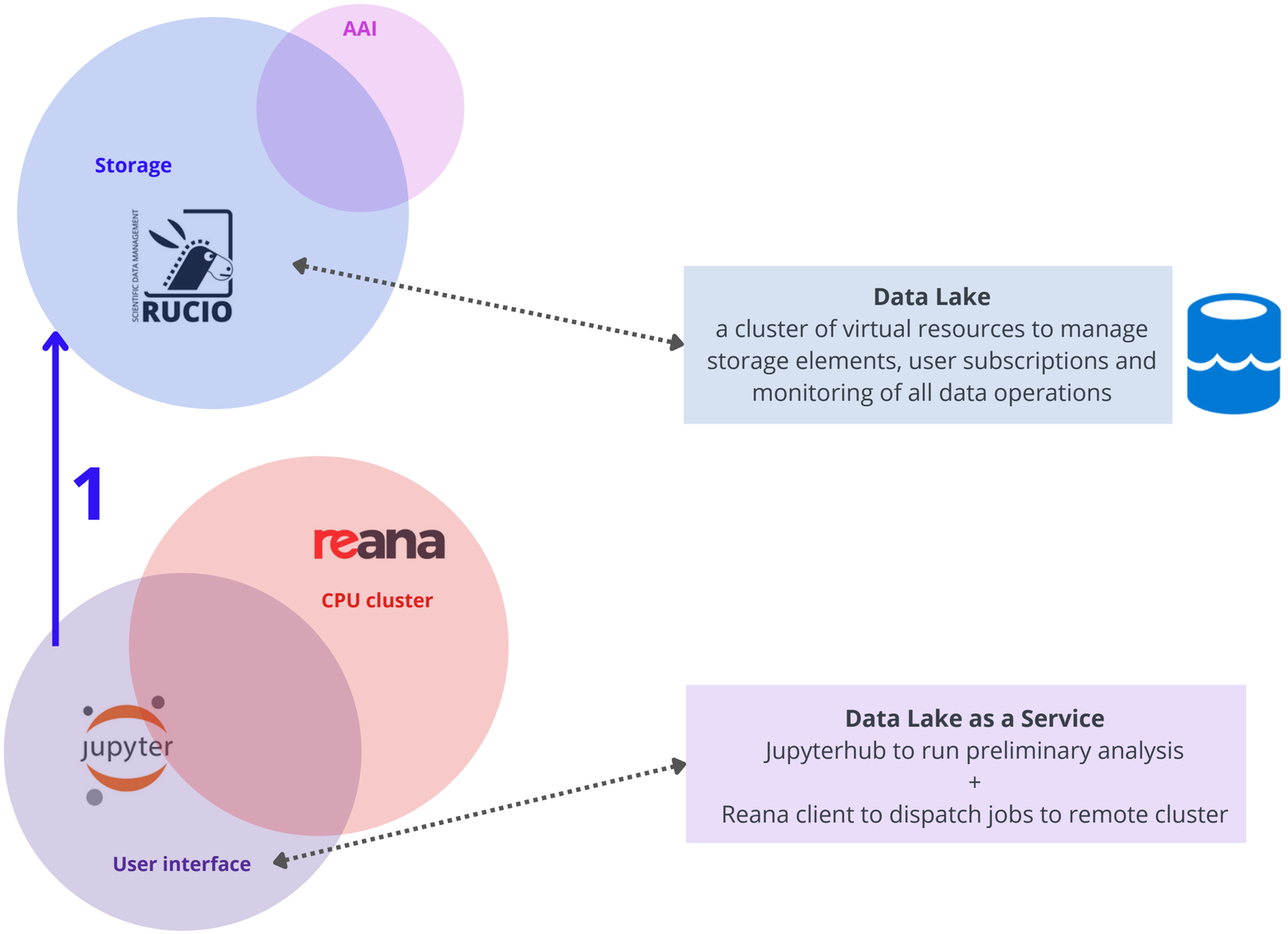
CernVM File system, GitLab, GitHub, zenodo, docker

=

Data Lake as a Service
Jupyterhub to run preliminary analysis
+
Reana client to dispatch jobs to remote cluster

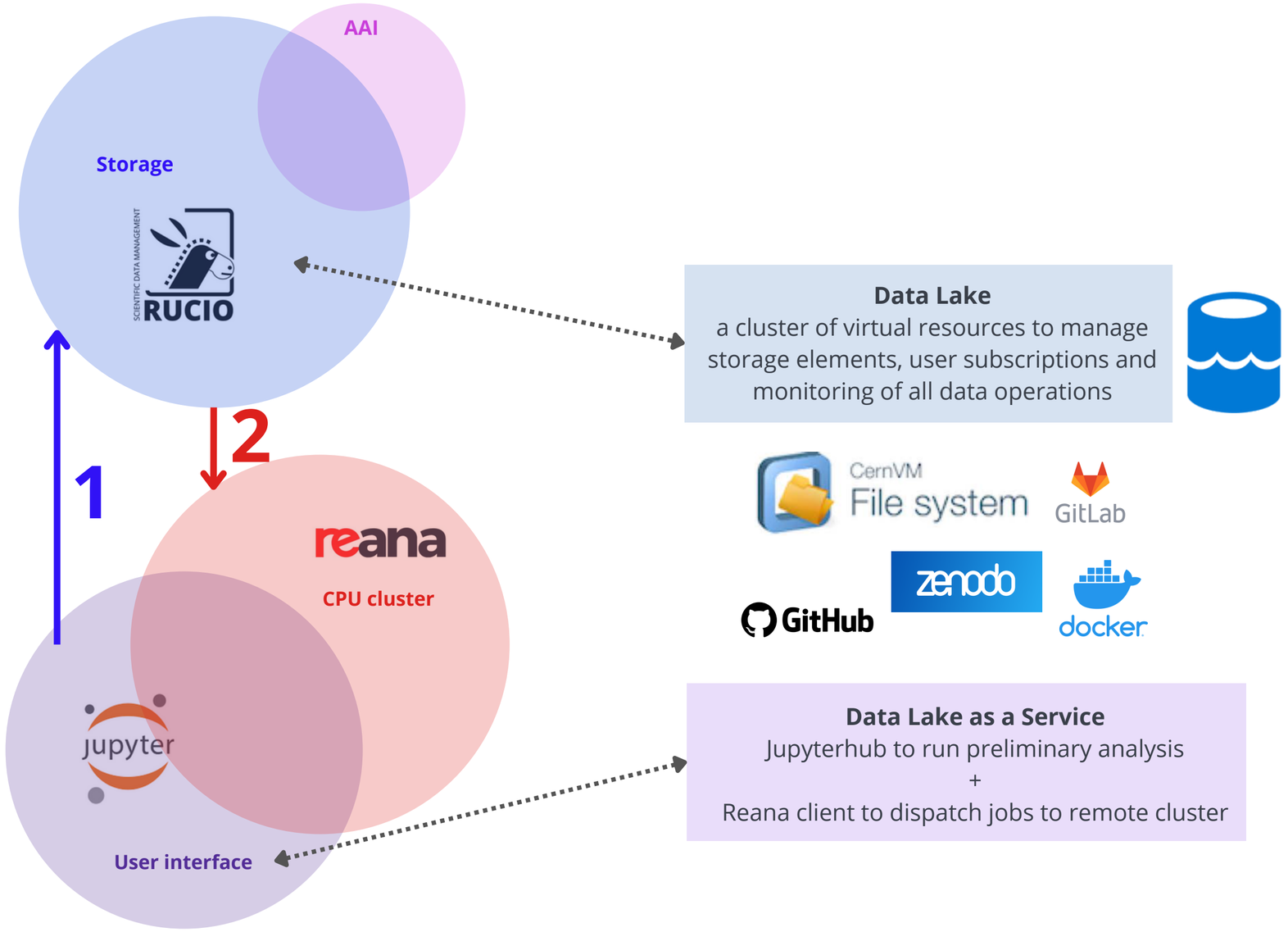
REANA: A System for Reusable Research Data Analyses - 10.1051/epjconf/201921406034

Components



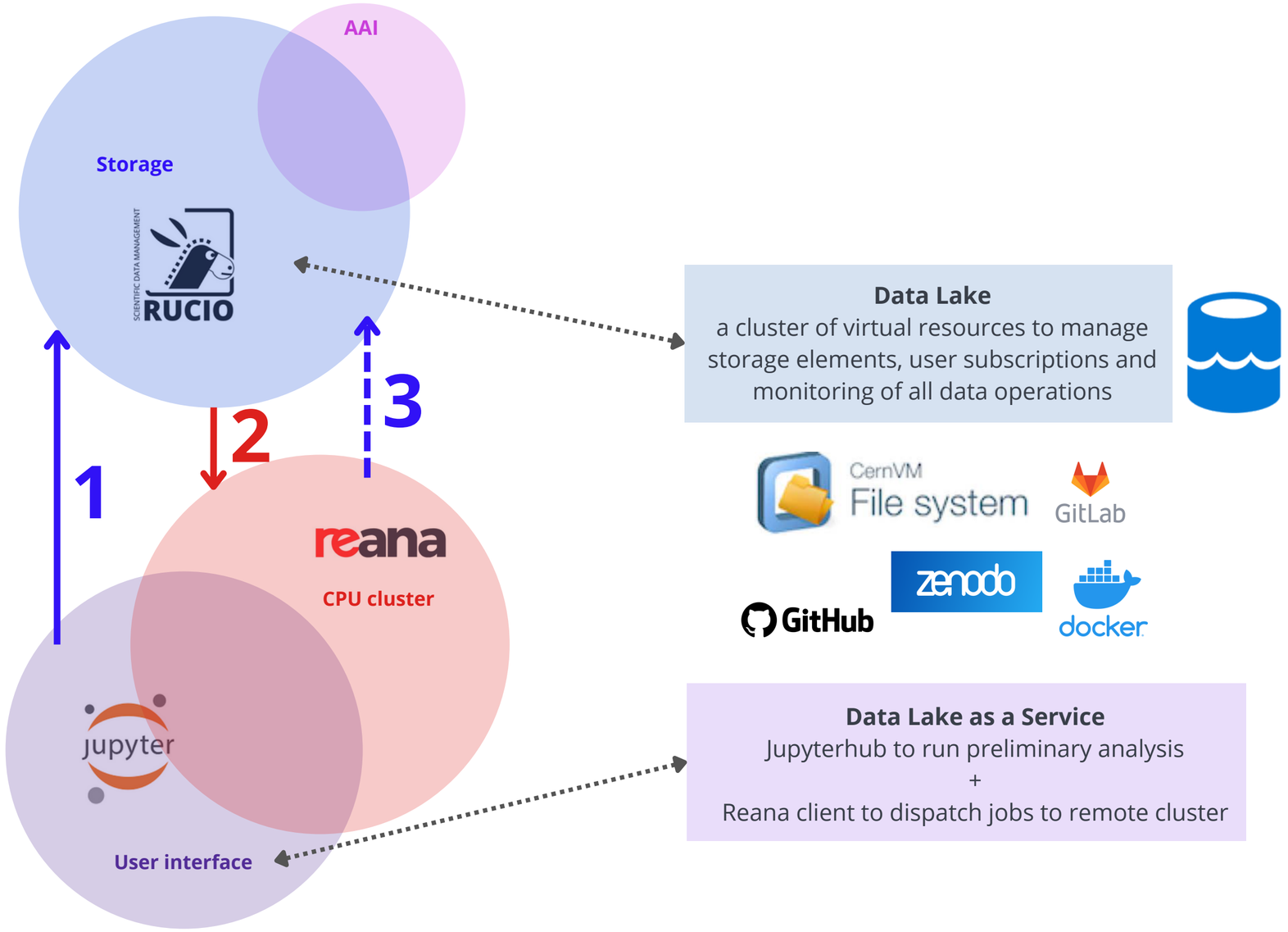
REANA: A System for Reusable Research Data Analyses - 10.1051/epjconf/201921406034

Components



REANA: A System for Reusable Research Data Analyses - 10.1051/epjconf/201921406034

Components



REANA: A System for Reusable Research Data Analyses - 10.1051/epjconf/201921406034

Components

FROM SCIENTISTS / EXPERIMENTS

FROM IT DEVELOPERS

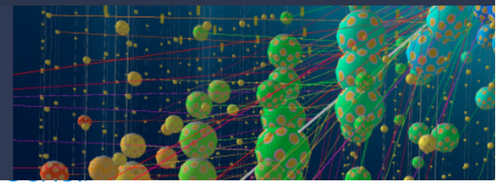
DATA	*uploaded & downloaded* →	Flexible data management system to preserve data quality and integrity, INDEPENDENT from local storage	= **Data Lake** a cluster of virtual resources to manage storage elements, user subscriptions and monitoring of all data operations
SOFTWARE	*uploaded & synchronised* →	Open source sofware repository	= CernVM File system, GitLab, GitHub, zenodo, docker
ENVIRONMENT	*uploaded & synchronised* →	Docker images spawned when user starts notebook	
STEPS	*uploaded & synchronised* →	Declarative format file to keep track of the workflow (.yaml)	= **Data Lake as a Service** Jupyterhub to run preliminary analysis + Reana client to dispatch jobs to remote cluster
NARRATIVE	→	Outreach activity to link the dots	= HUGO webpage

Dark Matter

ESCAPE AND EOSC FUTURE SCIENCE PROJECTS

Dark Matter studies at LHC

ne Standard Model (SM). If such interactions exist, DM particles could be produced
e detector, so we would have to observe them in association with visible SM



Virtual Research Enviroment

A COLLABORATIVE ONLINE PLATFORM WHERE SCIENCE PROJECTS ARE ABLE TO DEVELOP AND SHARE END-TO-END ANALYSIS WORKFLOWS, HAVING ACCESS TO ALL THE DIGITAL CONTENT NEEDED TO PRODUCE A SCIENTIFIC RESULT IN COMPLIANCE WITH FAIR PRINCIPLES.

ACCESS THE SCIENCE PROJECTS →

uploaded & synchronised

ESCAPE

Dark Matter

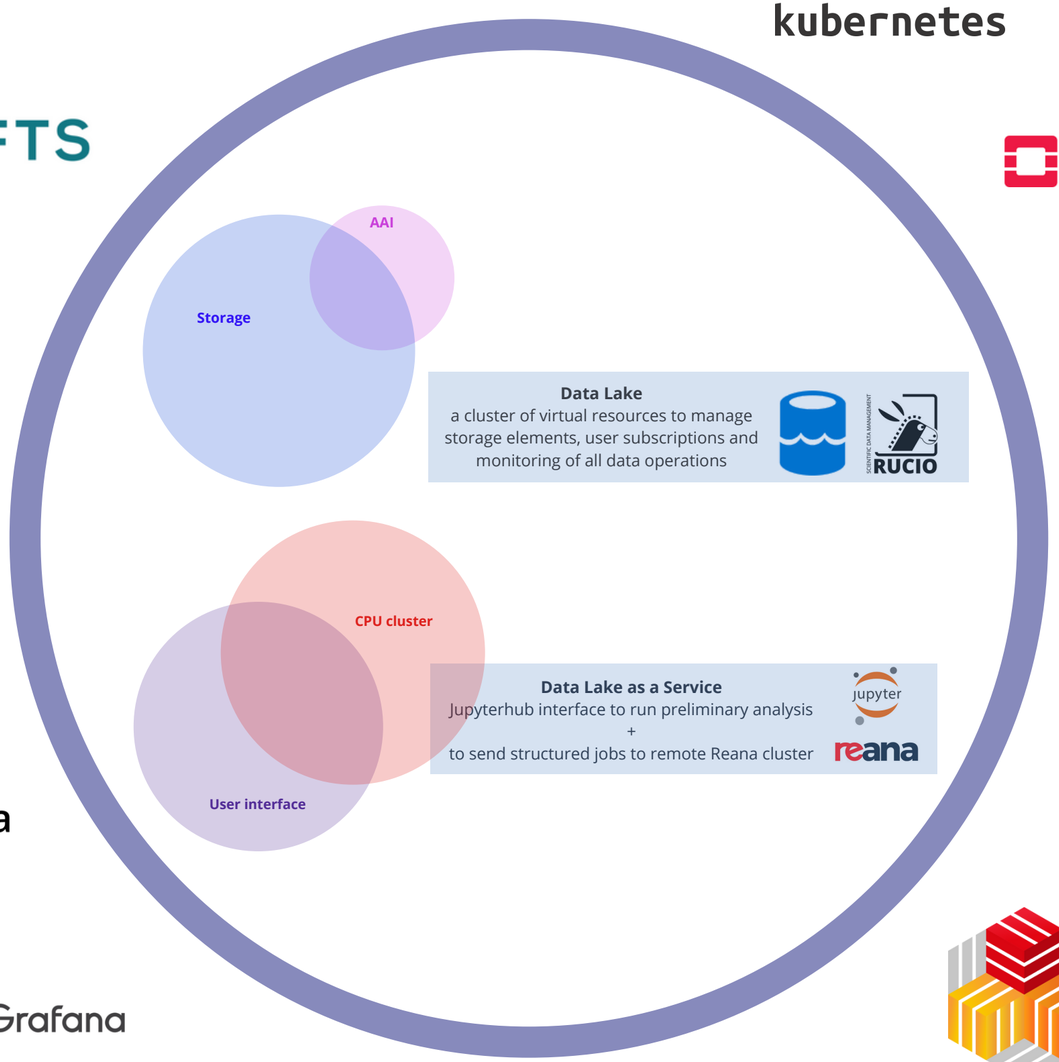
ESCAPE & EOSC FUTURE SCIENCE PROJECT

Extreme Universe

ESCAPE & EOSC FUTURE SCIENCE PROJECT

HUGO webpage

Technologies



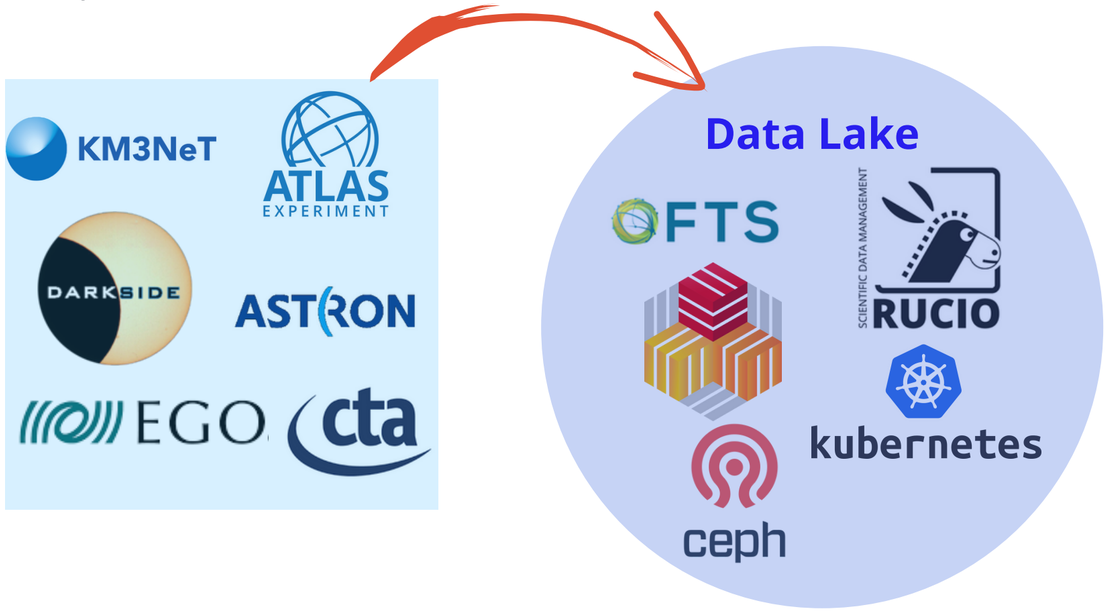
VRE supporting workflows

Generation and simulation of events

Experimental data

Data processing

Prepare for analysis



VRE supporting workflows

Generation and simulation of events

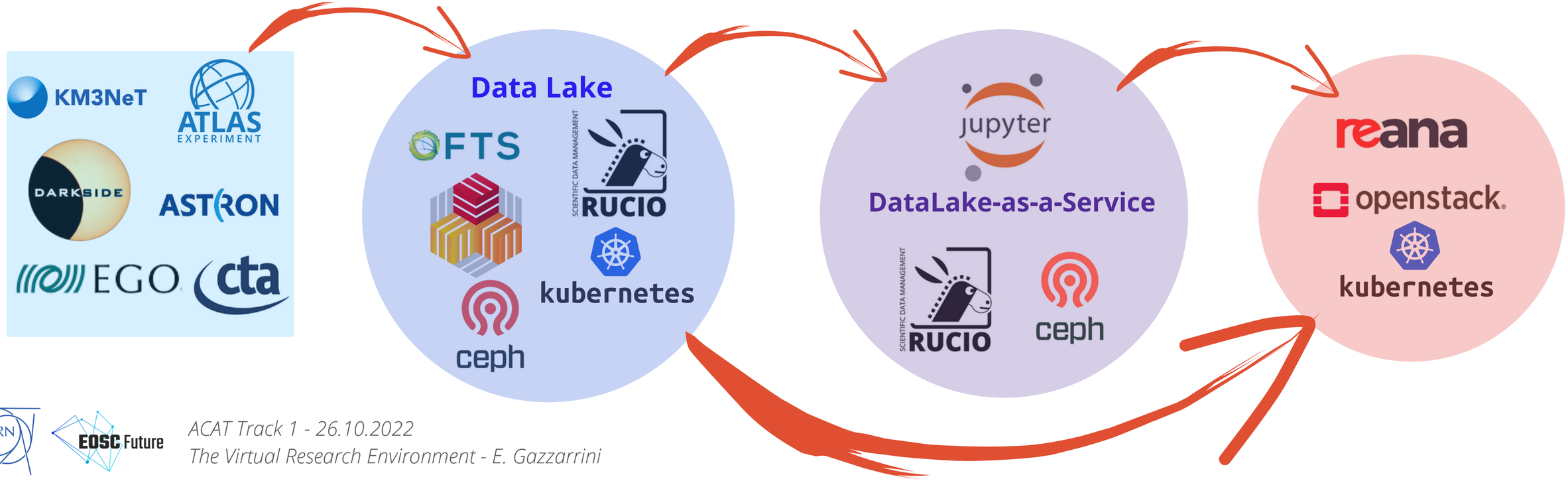
Experimental data

Data processing

Prepare for analysis

Analysis

Preserved software



VRE supporting workflows

Generation and simulation of events

Experimental data

Data processing

Prepare for analysis

Analysis

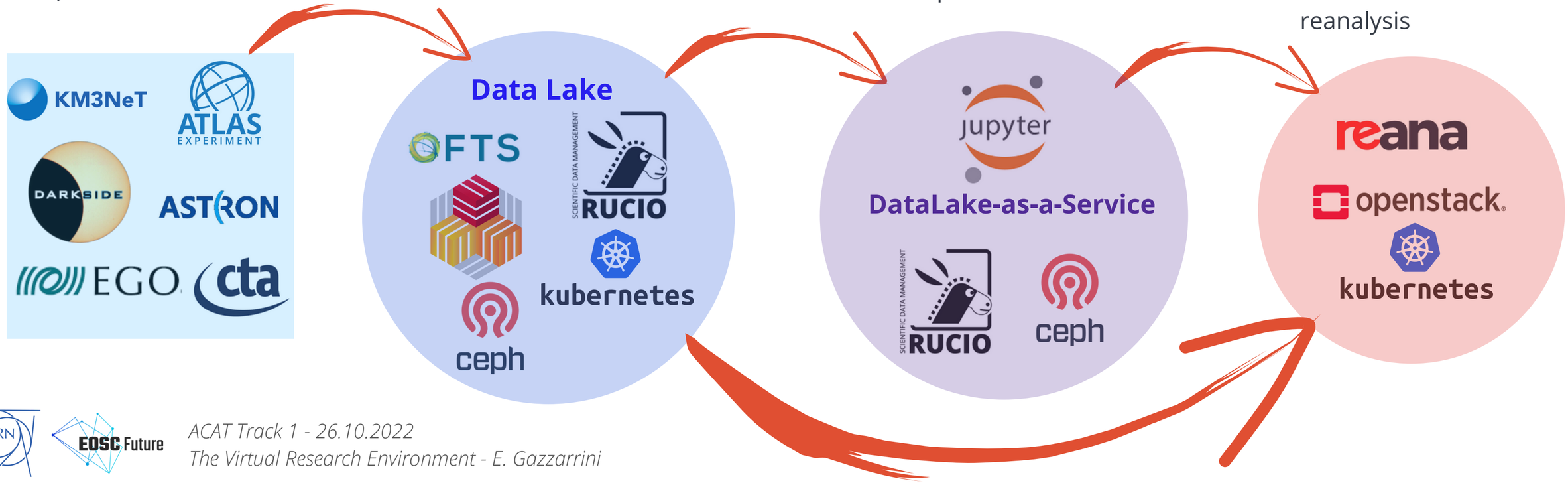
Preserved software

Interpretation of results

Combination of results and comparison across workflows

Preservation of analysis

For later reuse or reanalysis

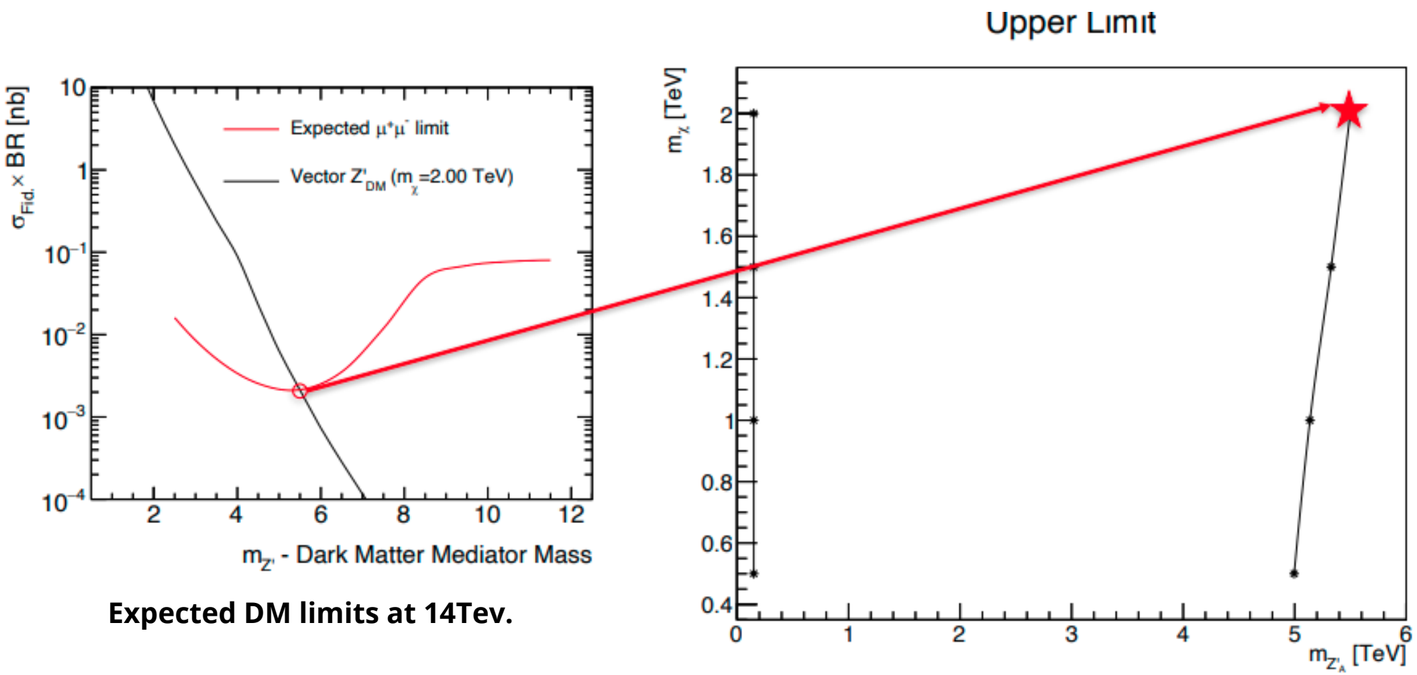


AN EXAMPLE USE CASE

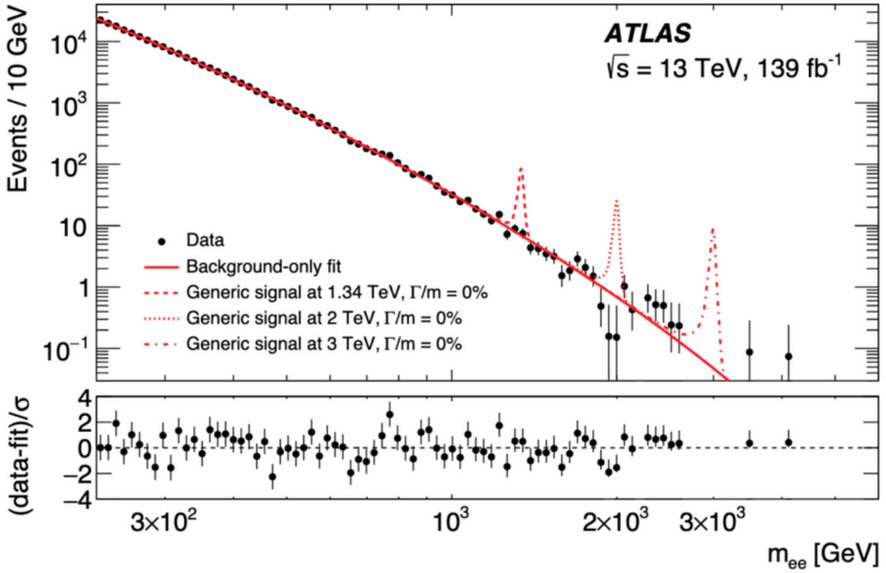
ATLAS Dark Matter reinterpretation analysis

DM@LHC with ATLAS

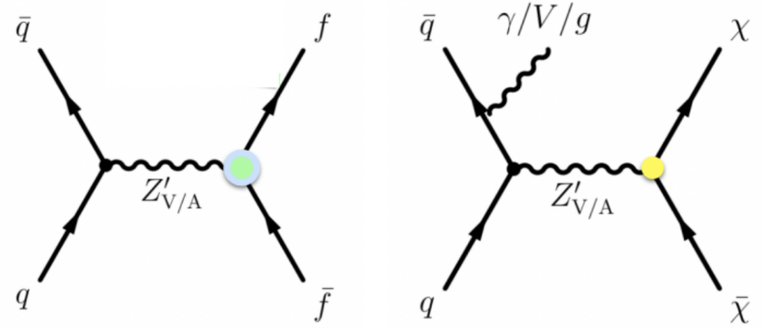
1. **Dark Matter Reinterpretation**: setting limits on High-Luminosity LHC contraints on $Z' \rightarrow \chi\chi$ (Z' mediated Dark Matter models).
2. The **dilepton** inclusive search (right) concluded in 2019
   a. objective: projecting limits to 14 TeV and computing the fiducial cross-sections in **lower mass regions.**



Expected DM limits at 14Tev.



Dilepton Inclusive Search. Results of this analysis demonstrate good agreement with SM predictions.

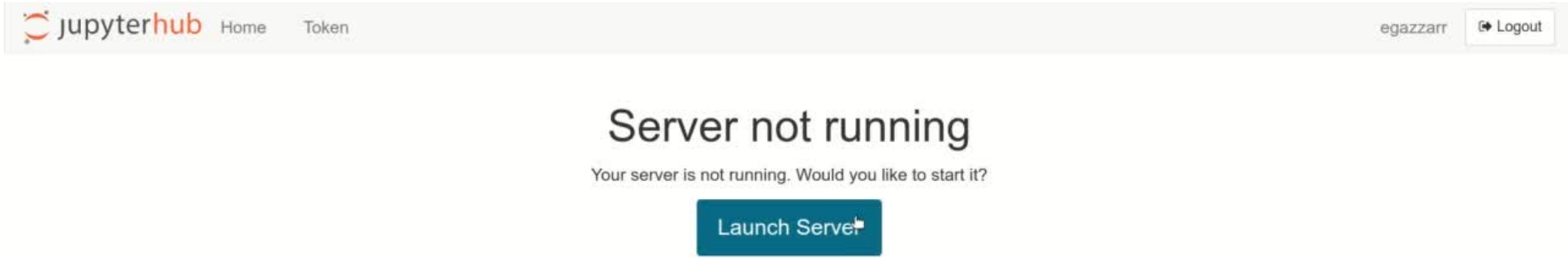


Fom Jared Little's presentation.

Demo workflow

1. **Docker environment** in DataLake-as-a-service (ROOT installed)
2. **Rucio extension** to browse data
3. Preliminary analysis on notebooks
4. Job submission to **Reana cluster** (data directly from Rucio to Reana)
5. Generating **plots**

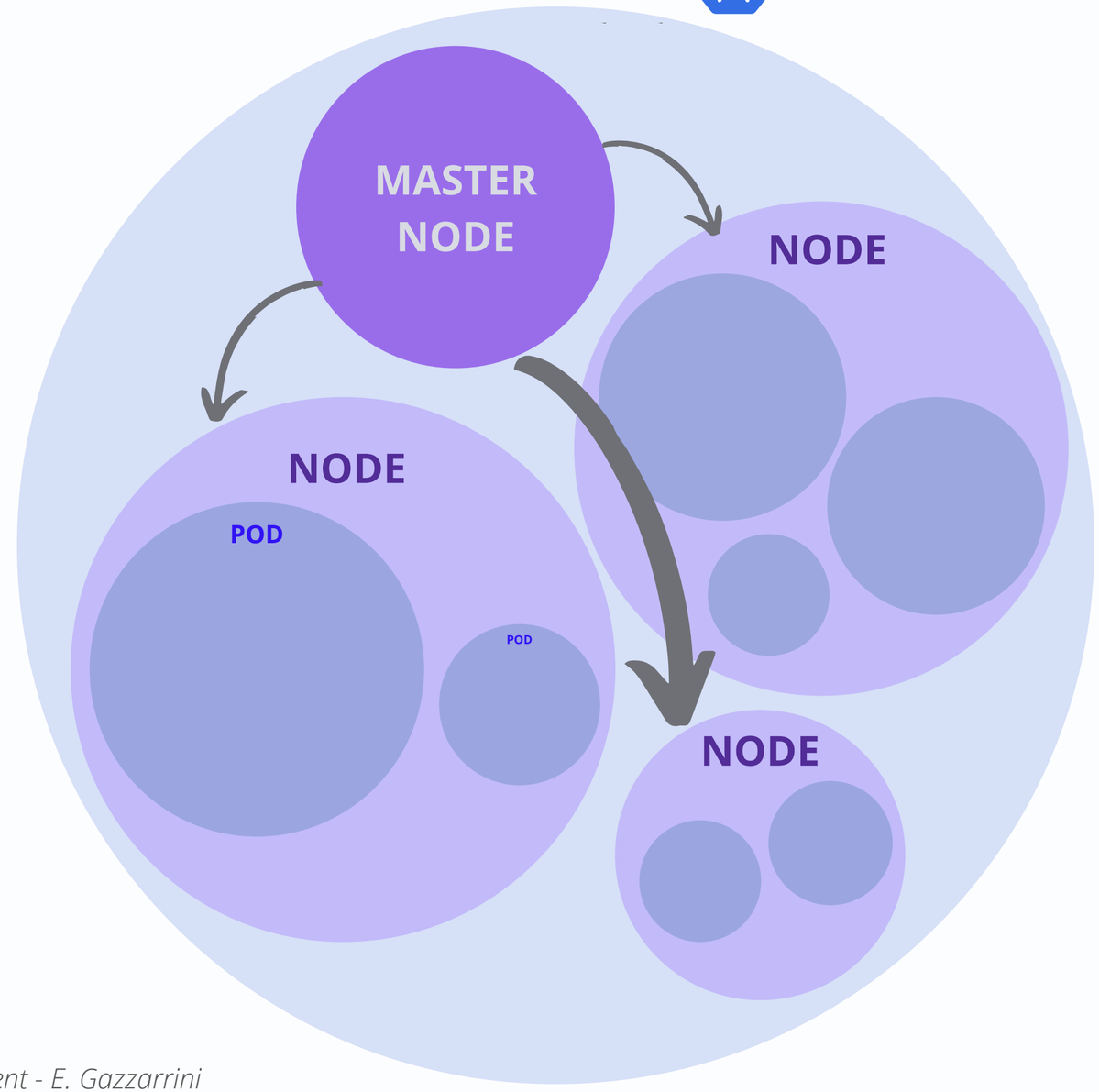
LINK: https://youtu.be/hvlJLo_7xXc



CERN EOSC Future

Behind the scenes...

Rucio Data Lake



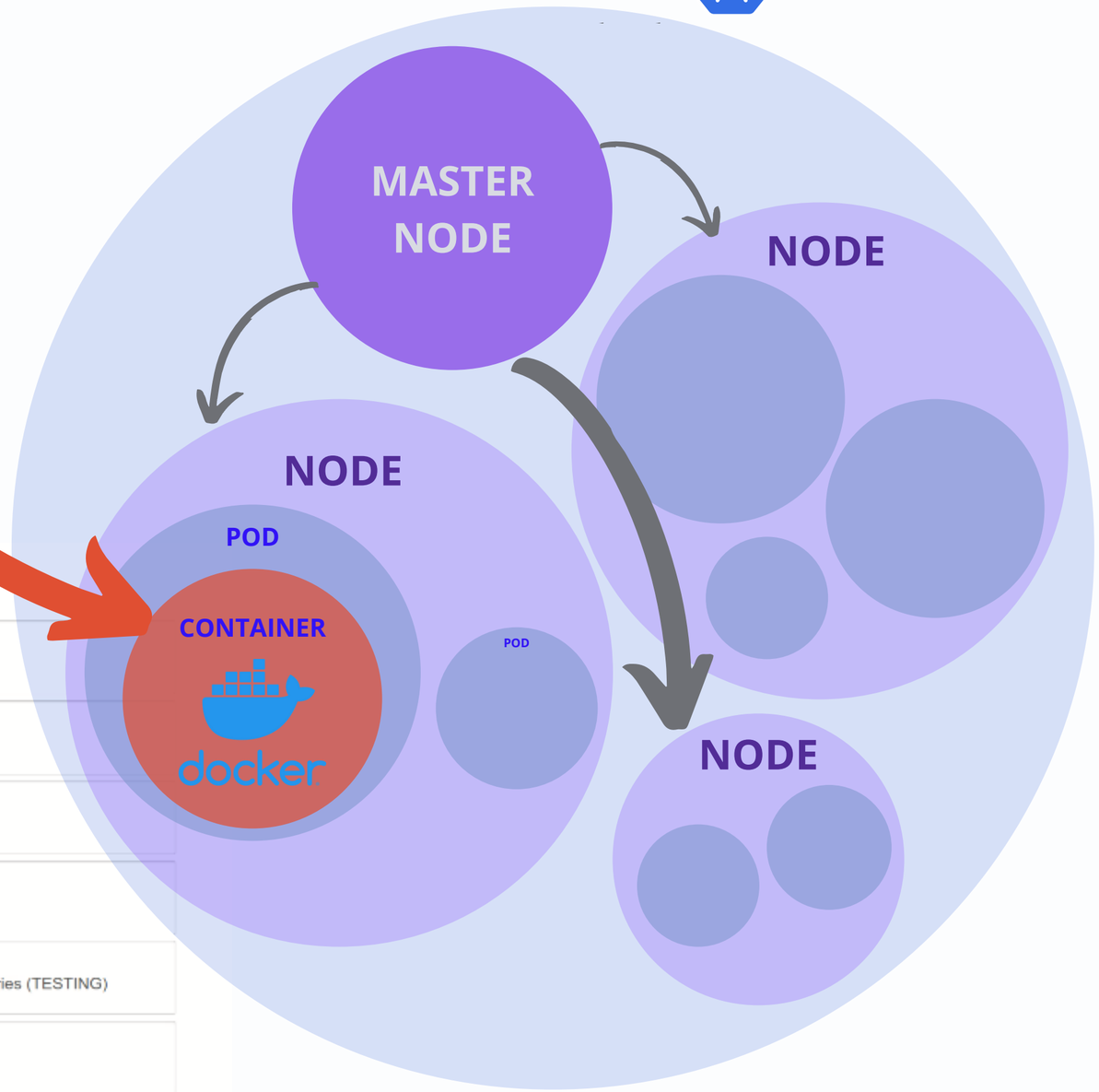
Behind the scenes...

Server Options

- **Minimal environment**
  Based on jupyter/scipy-notebook (active reana-client)
- **ROOT environment**
  Start your ROOT analysis. A C++ kernel is implemented too
- **ROOT environment (Xcache testing)**
  Run the extension in Download mode
- **SKA SDC1**
  SKA environment profile for SDC
- **Indirect Dark Matter Detection Environment**
  Contains a GCC compiler and the MLFermiLATDwarfs, fermipy, fermitools libraries (TESTING)
- **Common gamma analysis tools**
  Contains a GCC compiler and astropy, sherpa, agnpy, gammapy libraries

docker

Rucio Data Lake



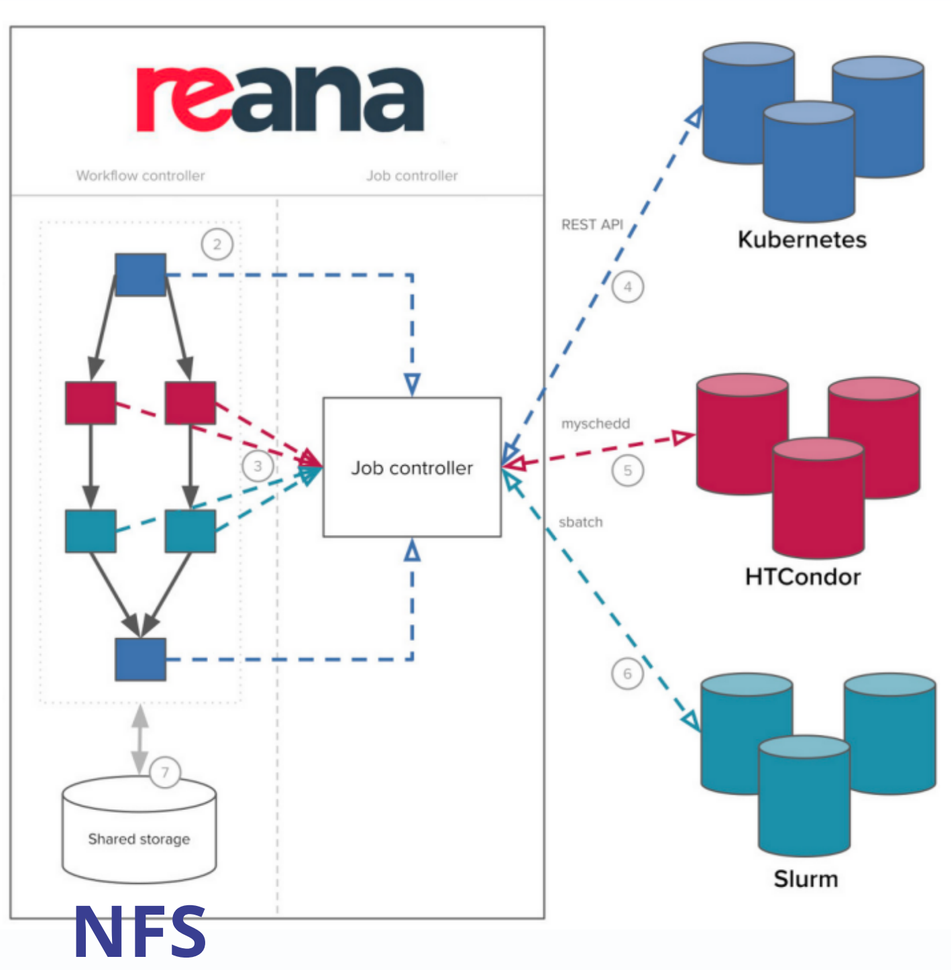
Behind the scenes...

DataLake-as-a-Service

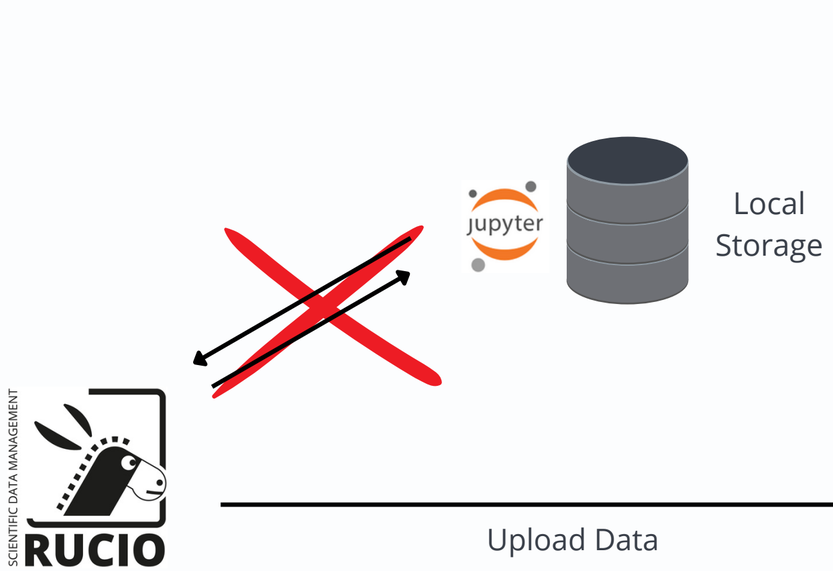
Behind the scenes...

Reana file

```yaml
version: 0.8.1
inputs:
  directories:
    - python/
workflow:
  type: serial
  specification:
    steps:
      - name: fetchdata-rucio
        voms_proxy: true
        rucio: true
        environment: 'projectescape/rucio-client'
        commands:
          - rucio whoami
          - rucio get ATLAS_LAPP_SP:DMs[illegible]mary.dileptonReinterpretat
      - name: SetLimits
        environment: 'reanahub/reana-env-r[illegible]t6:6.18.04'
        compute_backend: kubernetes
        kubernetes_memory_limit: '9Gi'
        commands:
          - mkdir plots/
          - python python/MakeLimit.py
outputs:
  directories:
    - plots/
```



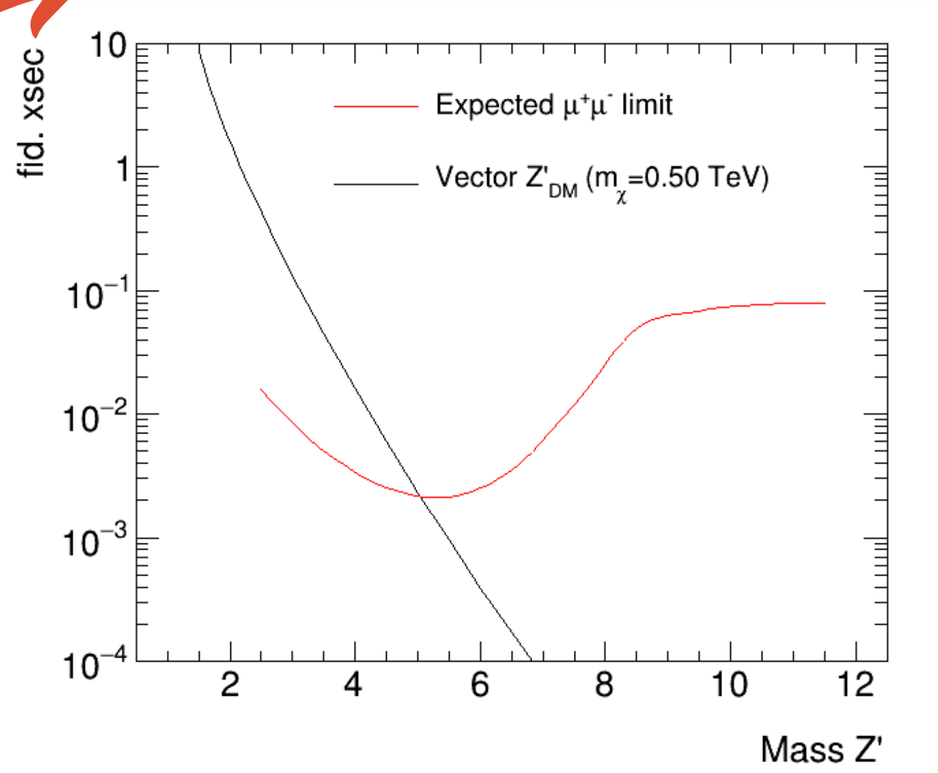
Behind the scenes...



reana

WORKSPACE
=
StorageClass
pointing to NFS
deployed in cluster

OUTPUT PLOT



Improvements

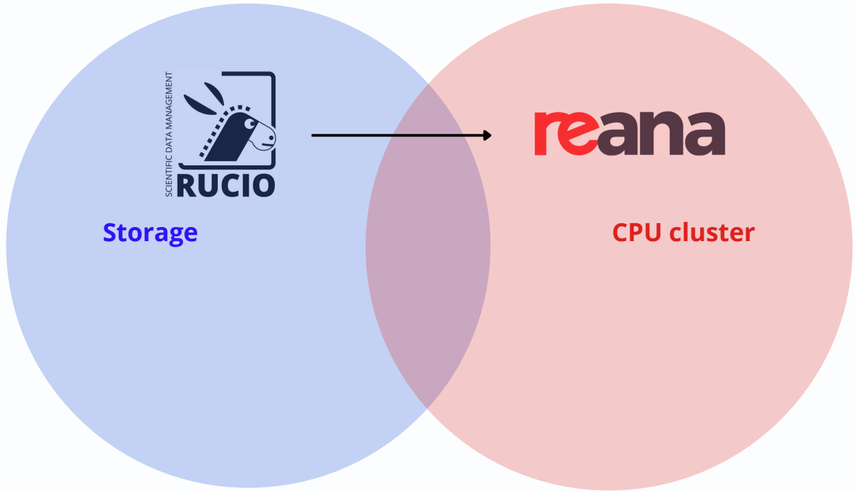
- How to evaluate **performance**?
- Better **resource usage** monitoring!
- **Network** overhead for data transfer
- **Plugins** on jupyterhub
- Authentication through **tokens** not yet fully supported

The VRE is ...

- **Useful!**
  - around 10 scientists from ATLAS, Virgo, CTA, Km3Net, FermiLat, Darkside, LOFAR have tested workflows successfully and have provided positive feedback
  - jupyter interface for exploratory analysis

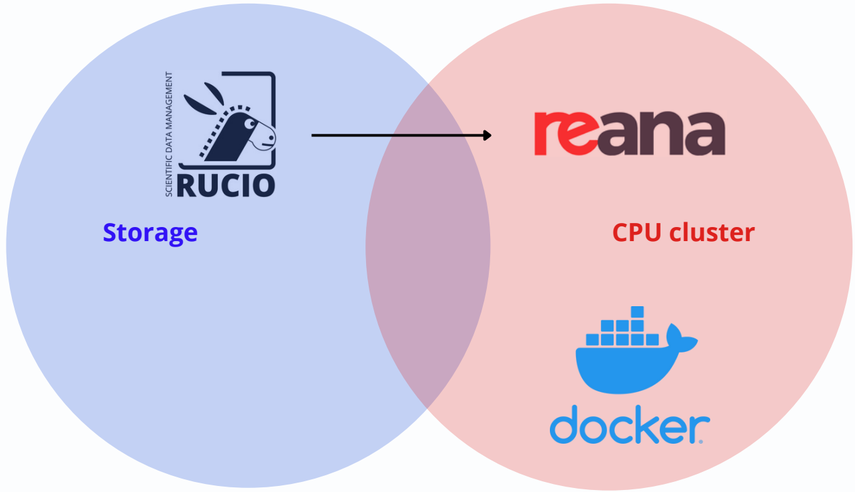
The VRE is ...

- **Useful!**
  - around 10 scientists from ATLAS, Virgo, CTA, Km3Net, FermiLat, Darkside, LOFAR have tested workflows successfully and have provided positive feedback
  - jupyter interface for exploratory analysis
- **Independent**
  - of local storage through Reana-Rucio connection
  - of restrictions on CERN resources



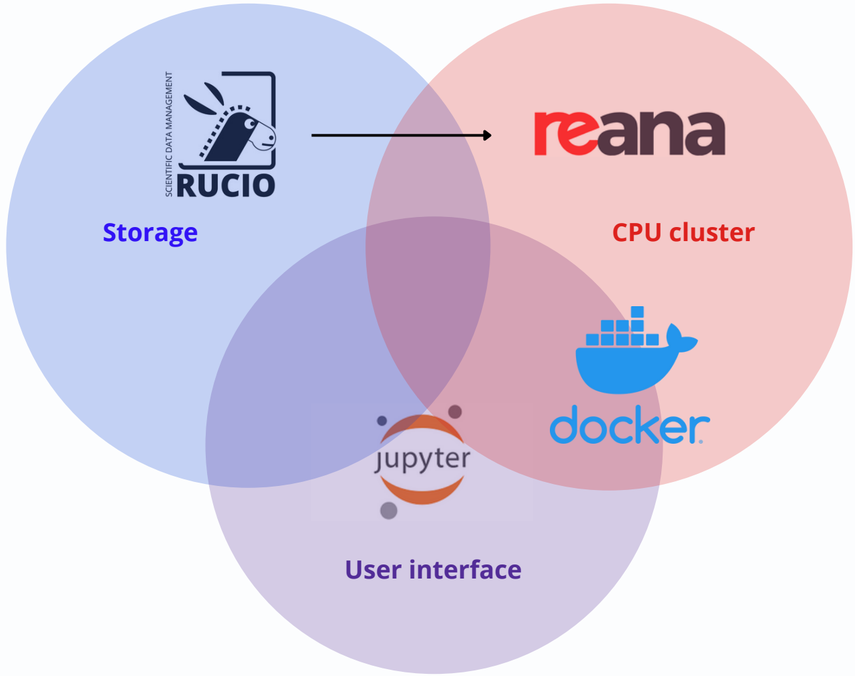
The VRE is ...

- **Useful!**
  - around 10 scientists from ATLAS, Virgo, CTA, Km3Net, FermiLat, Darkside, LOFAR have tested workflows successfully and have provided positive feedback
  - jupyter interface for exploratory analysis
- **Independent**
  - of local storage through Reana-Rucio connection
  - of restrictions on CERN resources
- **Abstraction**
  - Docker



The VRE is ...

- **Useful!**
  - around 10 scientists from ATLAS, Virgo, CTA, Km3Net, FermiLat, Darkside, LOFAR have tested workflows successfully and have provided positive feedback
  - jupyter interface for exploratory analysis
- **Independent**
  - of local storage through Reana-Rucio connection
  - of restrictions on CERN resources
- **Abstraction**
  - Docker
- **Heterogeneous**
  - Various resources as workflow back-end
- **Flexible**
  - ad-hoc workflows can be created via easily editable declarative files



Thank you! Questions? Advice?

E-mail

elena.gazzarrini@cern.ch

VRE website

https://escape2020.pages.in2p3.fr/virtual-environment/home/

Where to find me

CERN Meyrin site, room 513-1-014

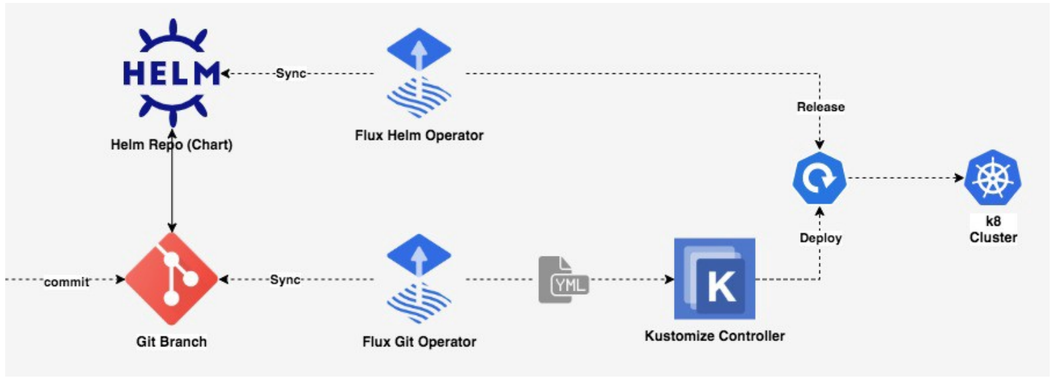
Back-up slides

References

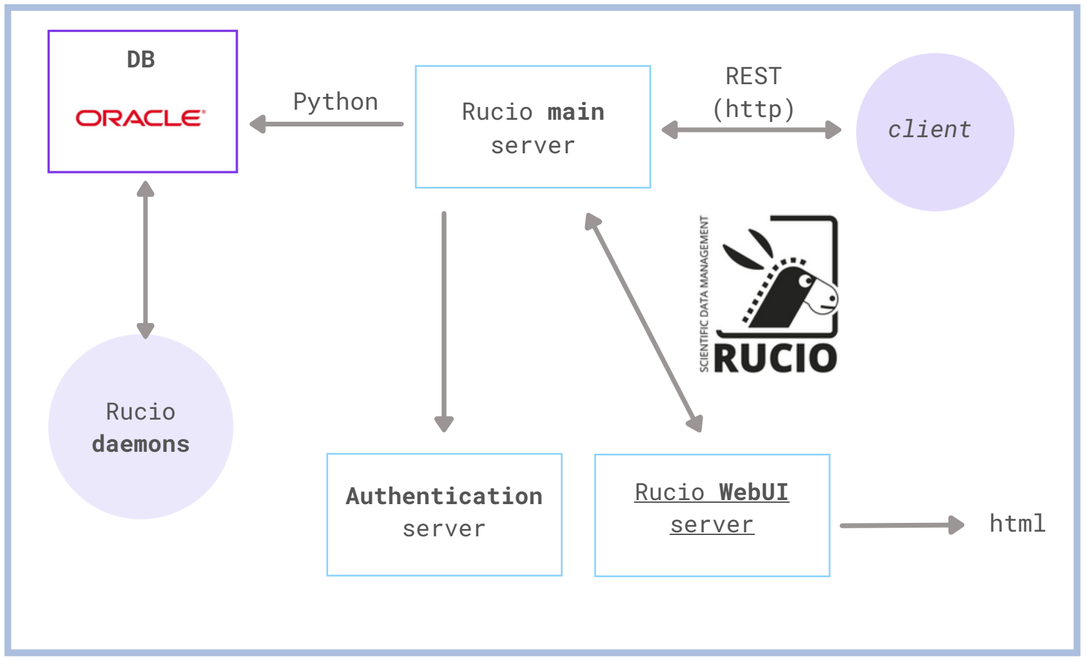
- Rucio
- Rucio Escape WebUI
- K8s cluster CI/CD Helm + Flux configurations
- Cluster testing
- Grafana monitoring
- VRE webpage (in progress)
- VRE documentation
- VRE onboarding
- VRE scientific analyses + docker images for notebooks

Rucio infrastructure

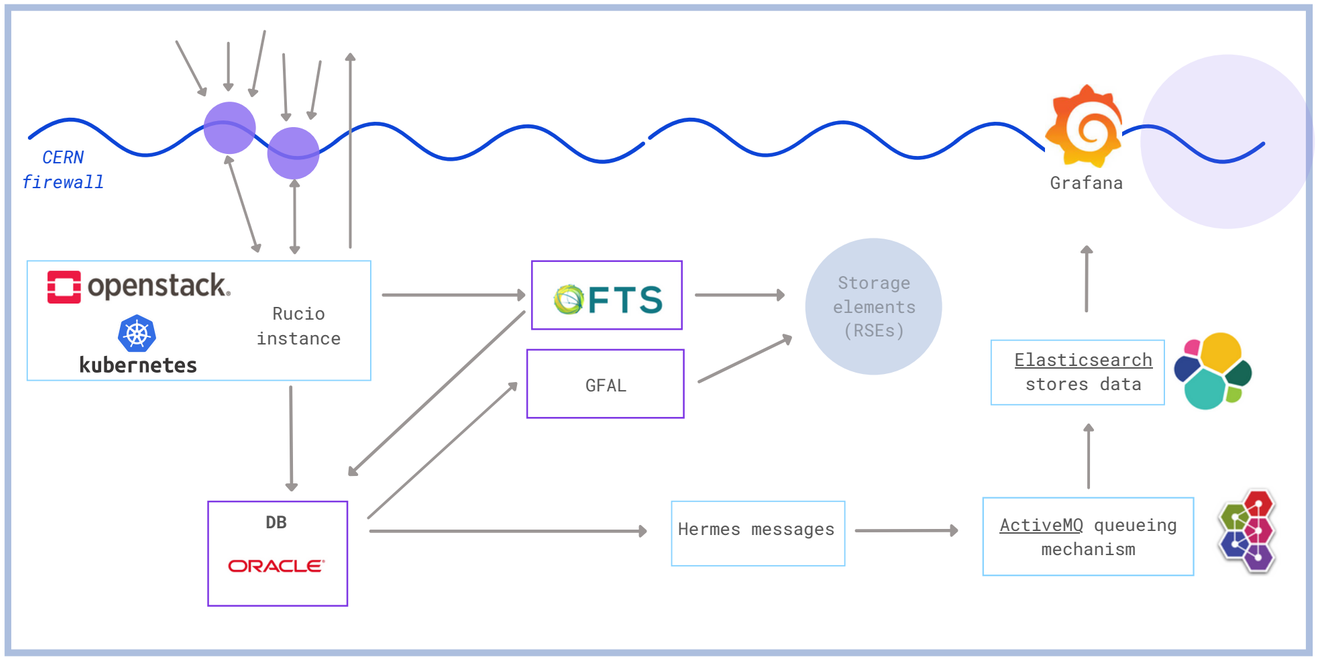
CI/CD cycle



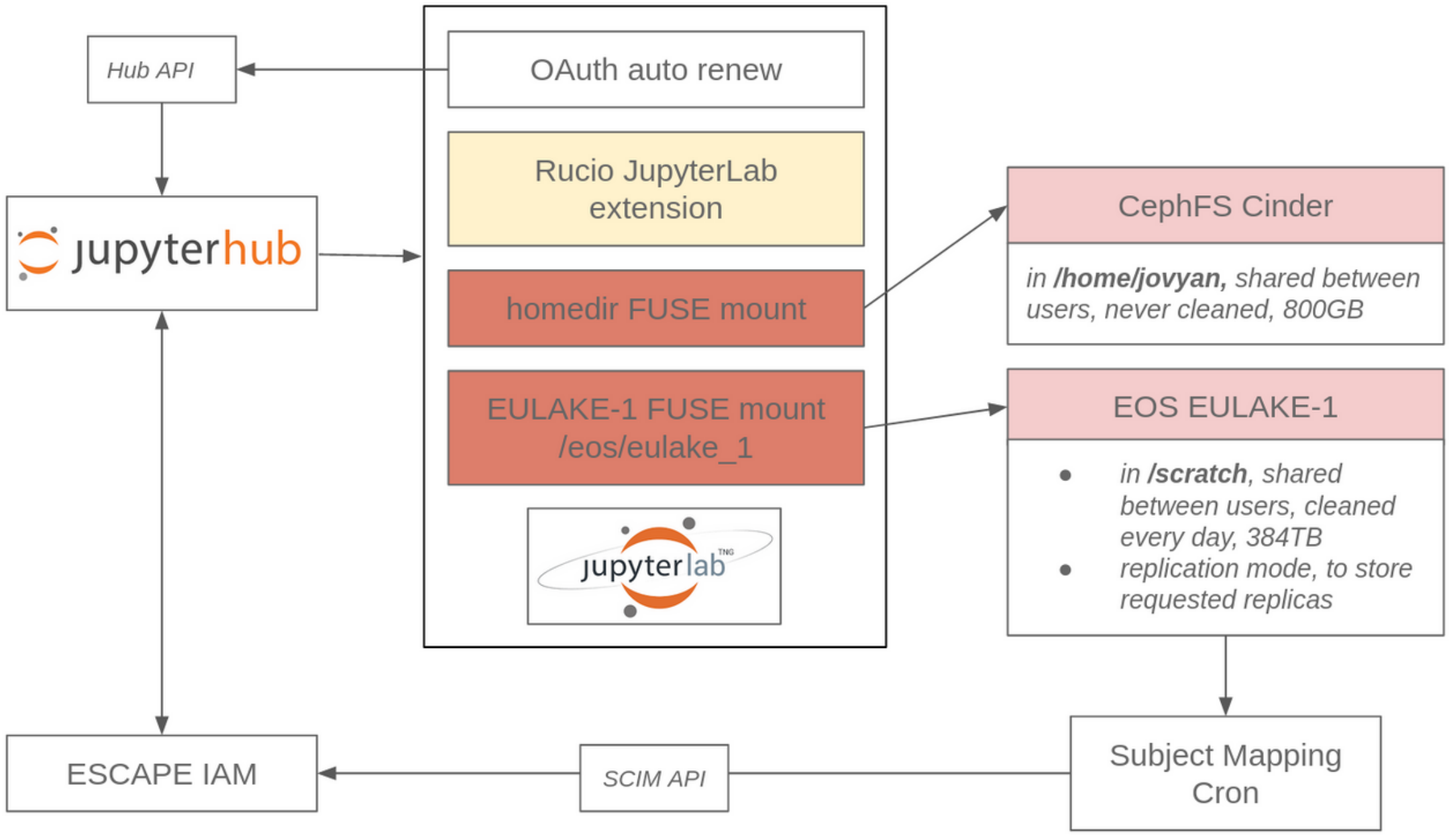
Server overview



Networking overview



The DataLake-as-a-Service (DLaaS)



How to use it

1. **Add secrets through terminal command line**

```
$ reana-client secrets-add
--file userkey.pem --file usercert.pem
--env VOMSPROXY_PASS=xxx
--env VONAME=escape
--env RUCIO_USERNAME=xxx
```

Sorry, you still need your X509 certificate..
But no need to execute the 'voms-proxy-init' command!

How to use it

2. Set voms_proxy & rucio to TRUE in .yaml file

```
steps:
  - name: fetchdata
    voms_proxy: true
    rucio: true
```

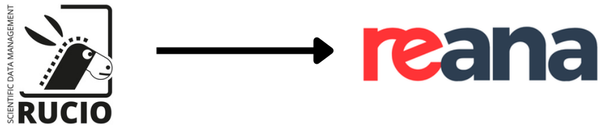
How to use it

3. Docker environment and execute Rucio commands

```
steps:
  - name: fetchdata
    voms_proxy: true
    rucio: true
    environment: 'reanahub/reana-env-rucioclient:latest'
    commands:
      - rucio get DID_name && rucio upload DID_name
```

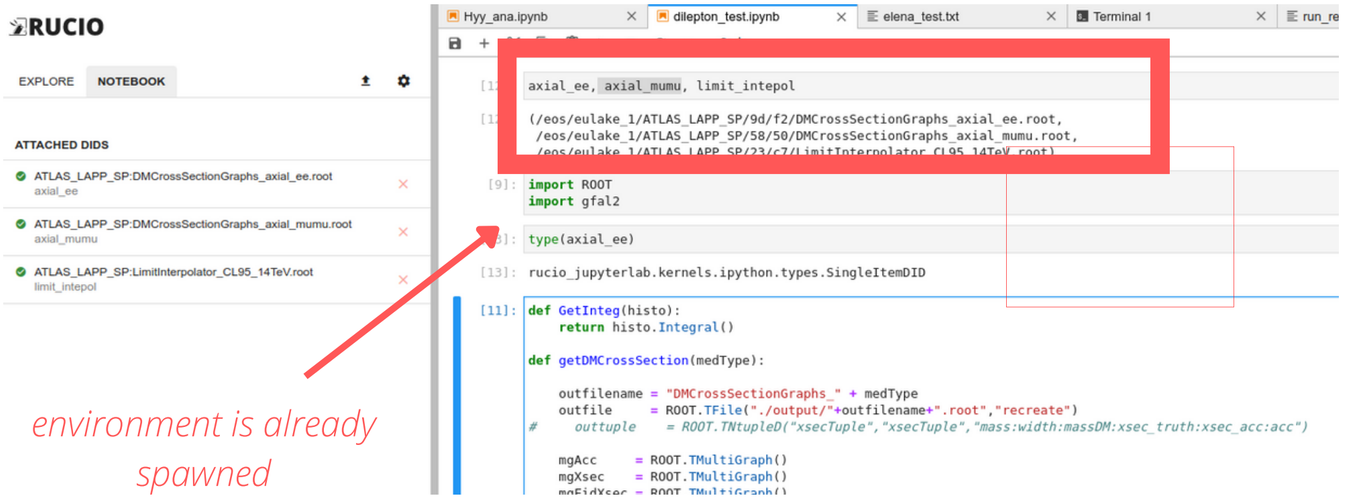
DEMO time!

```yaml
workflow:
  type: serial
  specification:
    steps:
      - name: fetchdata
        voms_proxy: true
        rucio: true
        environment: 'projectescape/rucio-client'
        commands:
        - rucio get agis_test:fitdata.C agis_test:gendata.C
      - name: gendata
        environment: 'reanahub/reana-env-root6:latest'
        kubernetes_memory_limit: '256Mi'
        commands:
        - mkdir -p results && root -b -q 'agis_test/gendata.C(${events},"${data}")'
      - name: fitdata
        environment: 'reanahub/reana-env-root6:latest'
        kubernetes_memory_limit: '256Mi'
        commands:
        - root -b -q 'agis_test/fitdata.C("${data}","${plot}")'
      - name: uploaddata
        voms_proxy: true
        rucio: true
        environment: 'projectescape/rucio-client'
        kubernetes_memory_limit: '256Mi'
        commands:
        - rucio upload --scope agis_test results/plot.png --rse EULAKE-1
```



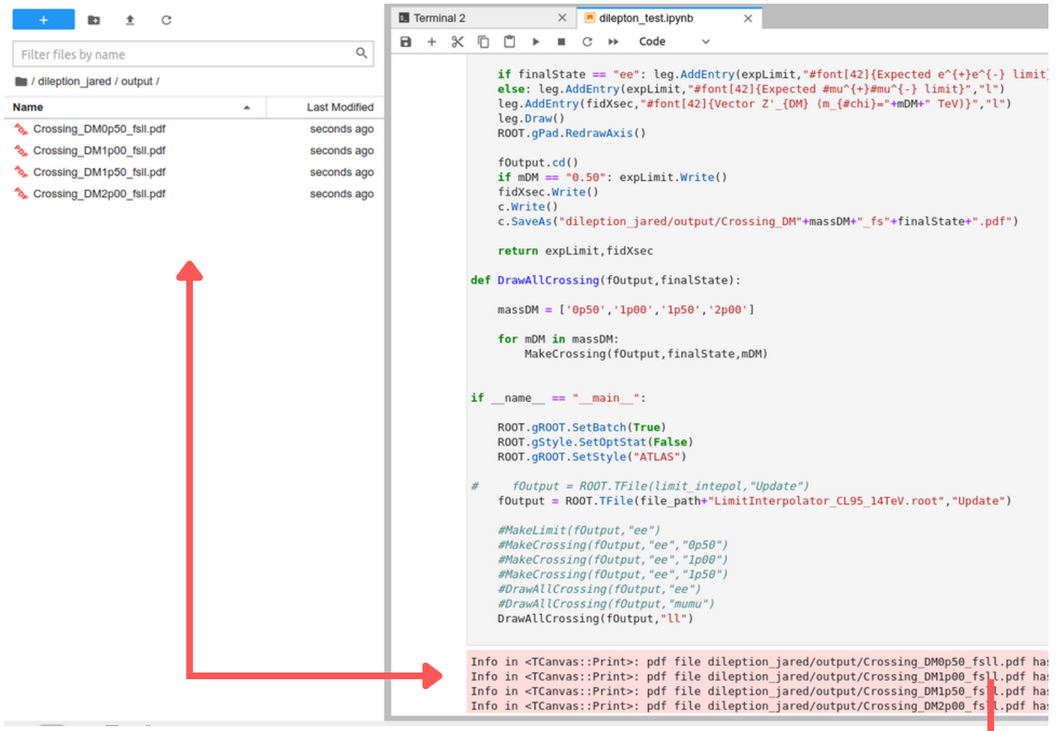
ATLAS Dark Matter Reinterpretation - Dilepton Resonance

1. Import files as variables into notebook

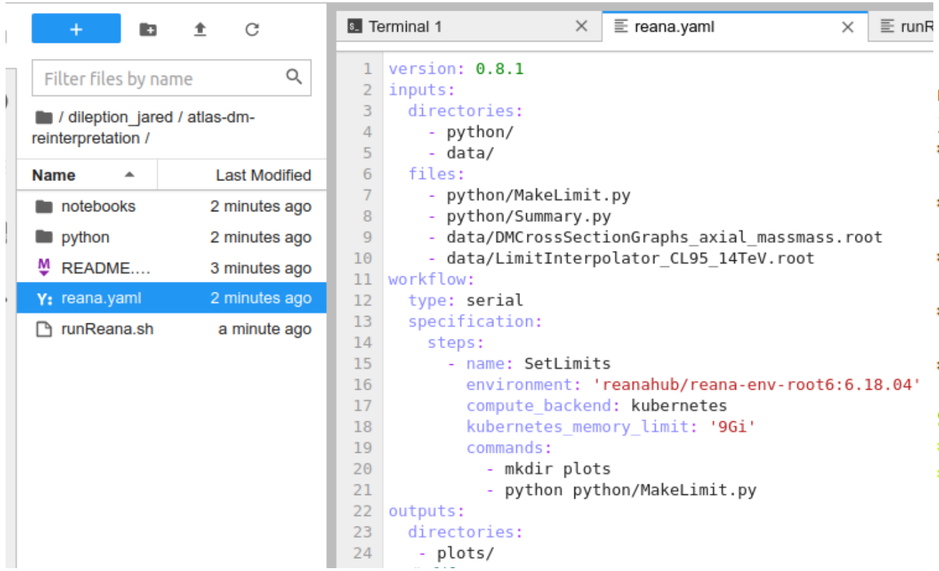


environment is already spawned

2a. Output generation



2b. REANA automatises workflow execution




```
notebooks  python  README.md  reana.yaml  runReana.sh
jovyan@jupyter-egazzarr:~/dileption_jared/atlas-dm-reinterpretatio
==> Verifying REANA specification file... /home/jovyan/dileption_j
  -> 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.
SettingLimits.1
==> SUCCESS: File /python/MakeLimit.py was successfully uploaded.
==> SUCCESS: File /python/Summary.py was successfully uploaded.
```