Virtual Research Environment

Towards a comprehensive analysis platform

Elena Gazzarrini, Enrique Garcia

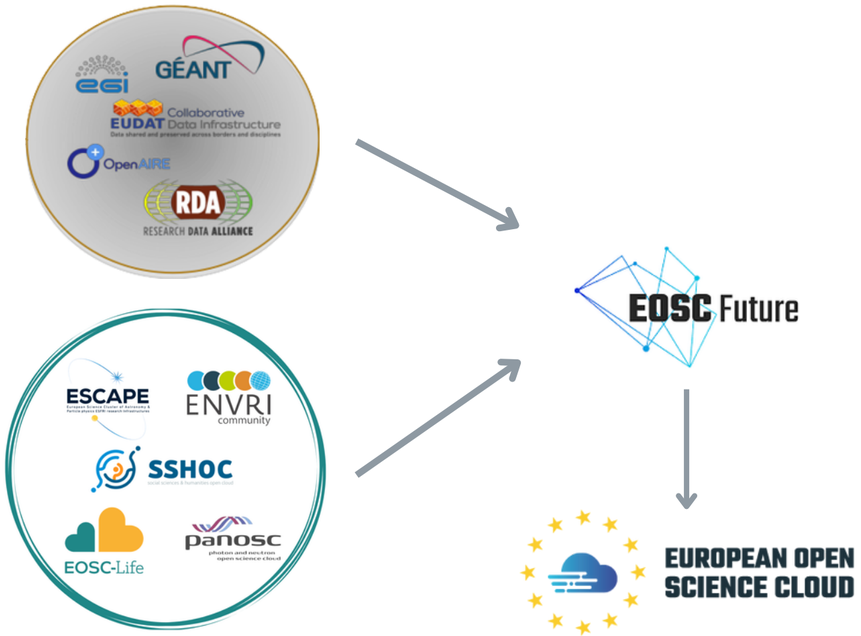
European Union's Horizon 2020 programme Grant Agreement 824064 and 101017536

The Virtual Research Environment

The VRE is an **open source** analysis platform where researchers have access to all the digital content needed to **develop, share and reproduce an end-to-end scientific result** in compliance with **FAIR** (findable, accessible, interoperable, reproducible) principles.

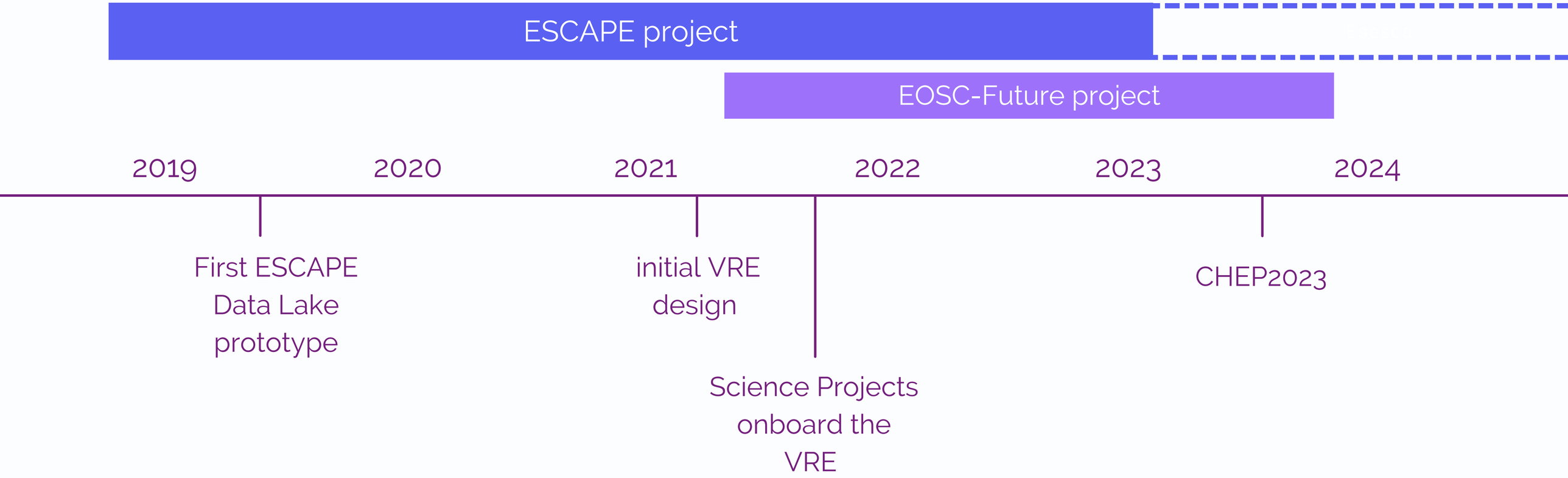
(documentation: https://vre-hub.github.io/)

- Researchers
- Community Software & Analysis development
- Data & Workflow Management
- IT administrators



CERN EOSC Future ESCAPE European Commission

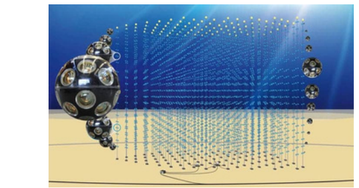
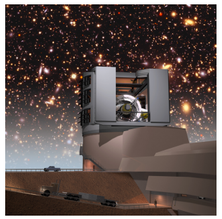
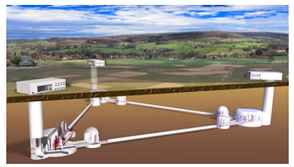
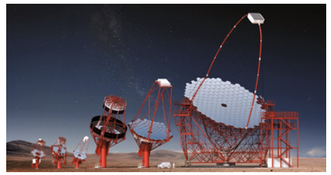
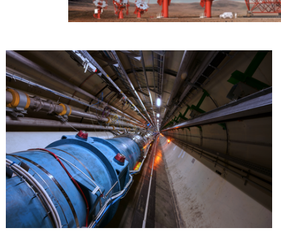
Timeline


ESCAPE project
EOSC-Future project
2019
2020
2021
2022
2023
2024
First ESCAPE Data Lake prototype
initial VRE design
Science Projects onboard the VRE
CHEP2023

Context: EOSC-Future

EOSC-Future Science Projects demonstrate

- multi-domain science integration across the **ESCAPE** project
- unification of services under **one Proof of Concept (PoC) analysis platform,** the **VRE**
- **interdisciplinary open science** example from bottom-up effort as a science driver for other communities

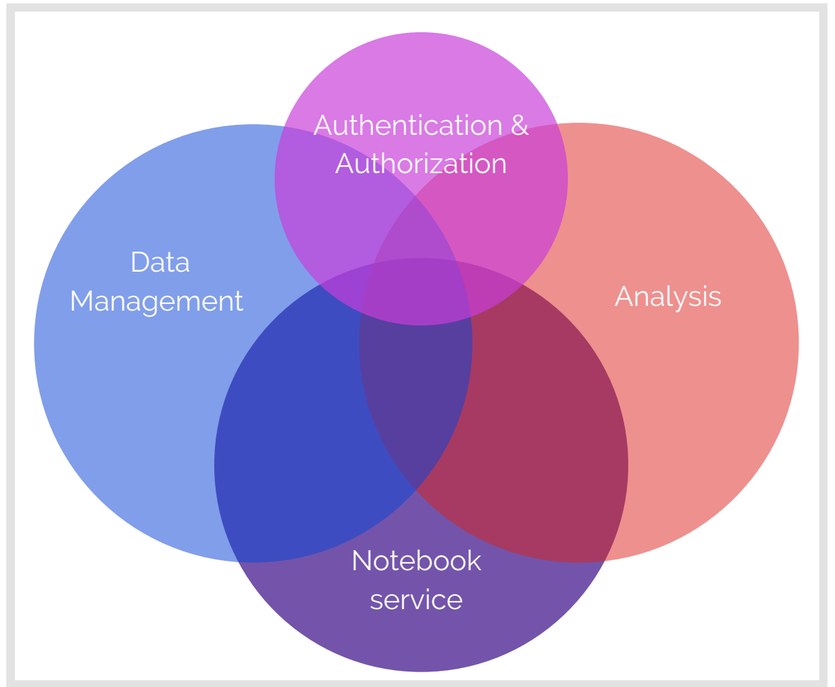






- Neutrino Observations
- High Energy Astronomy
- Low Frequency Telescopes
- Particle Physics
- Gravitational Waves

DARK MATTER

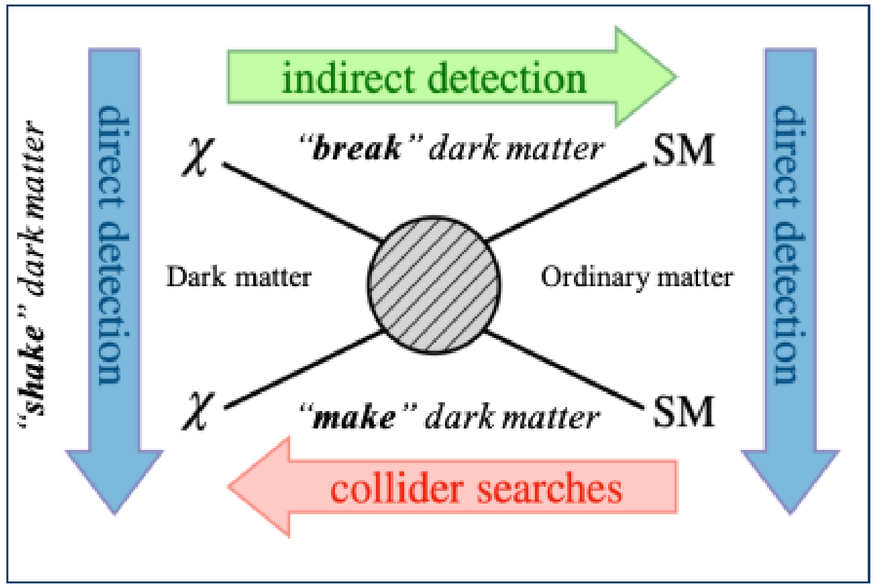
EXTREME UNIVERSE

VRE



Dark Matter: Complementary Approach

Focus: Looking for **W**eakly **I**nteracting **M**assive **P**articles (**WIMP**s)



A joint discovery of the nature of dark matter requires different experiments and inputs

Experiments have **different** data sizes, workflows, data, and result sharing policies

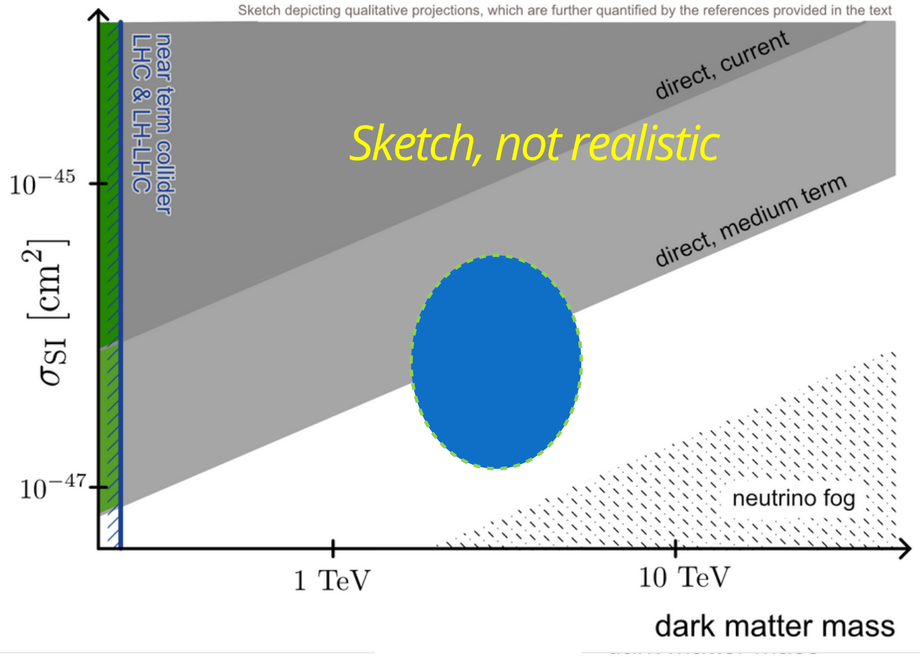
Example of a **discovery scenario**

Late 2020s

Direct detection experiment sees a hint of a signal, with characteristics compatible with WIMP DM

Mid 2030s

2040s



Inspired by:
Dark Matter Complementarity (Snowmass report), arXiv:2210.01770
T. Slatyer's "Paths to discovery" talk at Snowmass 2022

Example of a **discovery scenario**

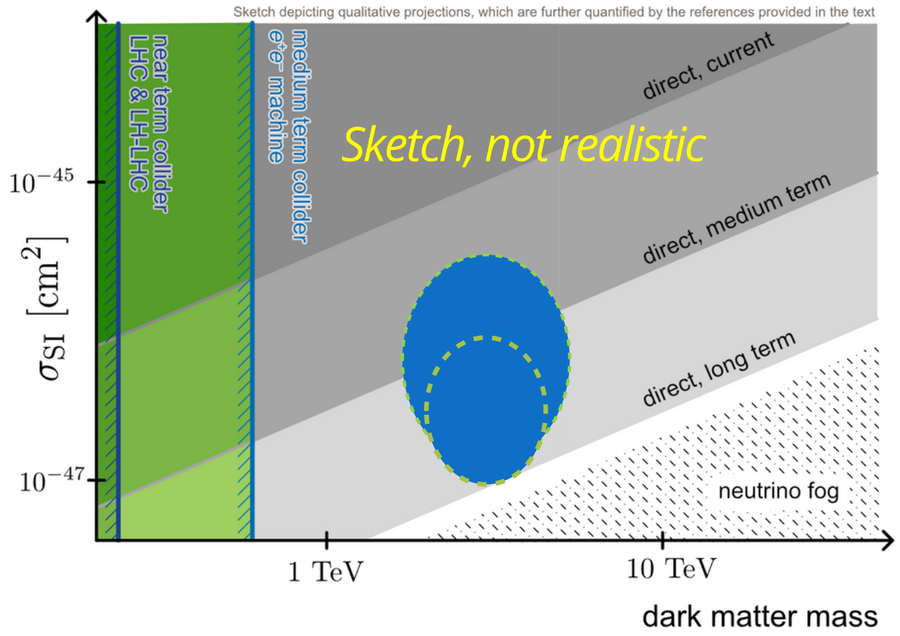
Late 2020s

Direct detection experiment sees a hint of a signal, with characteristics compatible with WIMP DM

Mid 2030s

Direct detection experiment (using another technique) confirms these hints

2040s



Inspired by:
Dark Matter Complementarity (Snowmass report), arXiv:2210.01770
T. Slatyer's "Paths to discovery" talk at Snowmass 2022

Example of a **discovery scenario**

Late 2020s

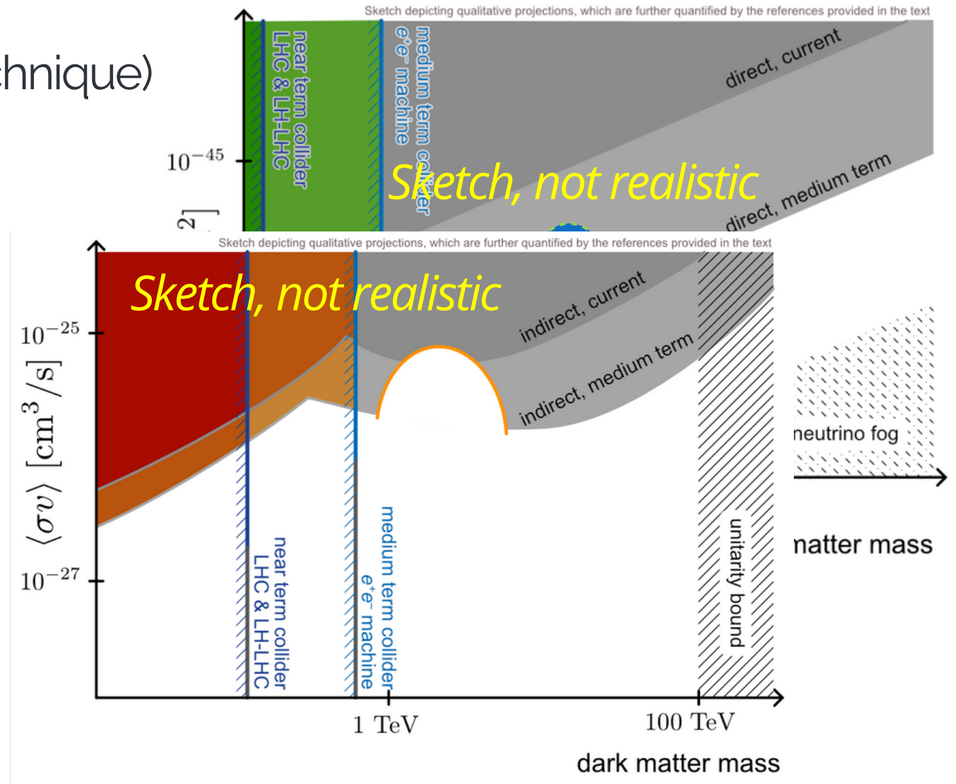
Direct detection experiment sees a hint of a signal, with characteristics compatible with WIMP DM

Mid 2030s

Direct detection experiment (using another technique) confirms these hints

2040s

Indirect detection experiment observes signals of DM annihilation



Inspired by:
Dark Matter Complementarity (Snowmass report), arXiv:2210.01770
T. Slatyer's "Paths to discovery" talk at Snowmass 2022

Example of a **discovery scenario**

Late 2020s

Direct detection experiment sees a hint of a signal, with characteristics compatible with WIMP DM

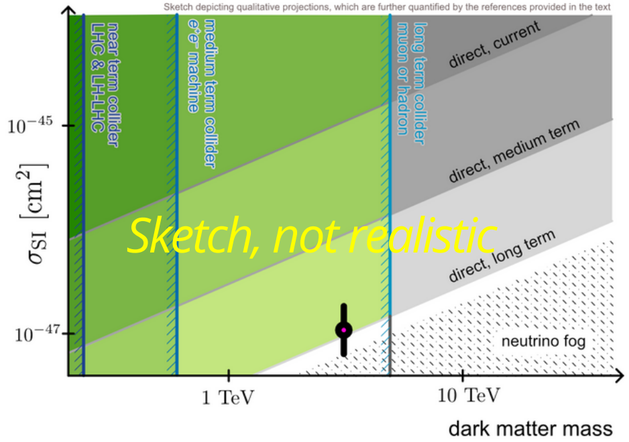
Direct detection experiment (using another technique) confirms these hints

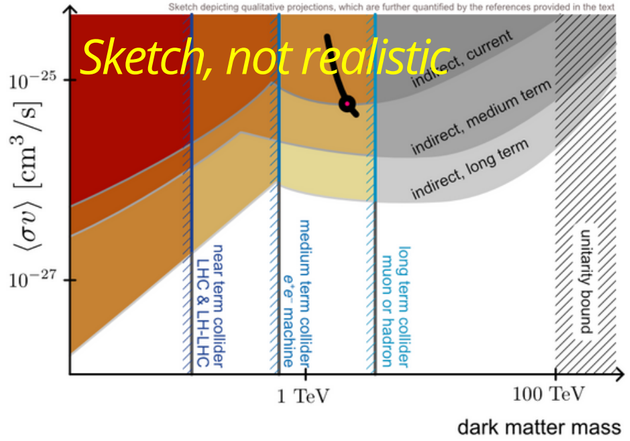
Mid 2030s

Indirect detection experiment observes signals of DM annihilation

2040s

Future collider, built to target particles with the mass of the putative DM candidate, sheds light on interactions between DM and ordinary matter





Inspired by:
Dark Matter Complementarity (Snowmass report), arXiv:2210.01770
T. Slatyer's "Paths to discovery" talk at Snowmass 2022

Example of a **discovery scenario**

Late 2020s

Direct detection experiment sees a hint of a signal, with characteristics compatible with WIMP DM

Direct detection experiment (using another technique) confirms these hints

Mid 2030s

Indirect detection experiment observes signals of DM annihilation

2040s

Future collider, built to target particles with the mass of the putative DM candidate, sheds light on interactions between DM and ordinary matter

Inspired by:
Dark Matter Complementarity (Snowmass report), arXiv:2210.01770
T. Slatyer's "Paths to discovery" talk at Snowmass 2022

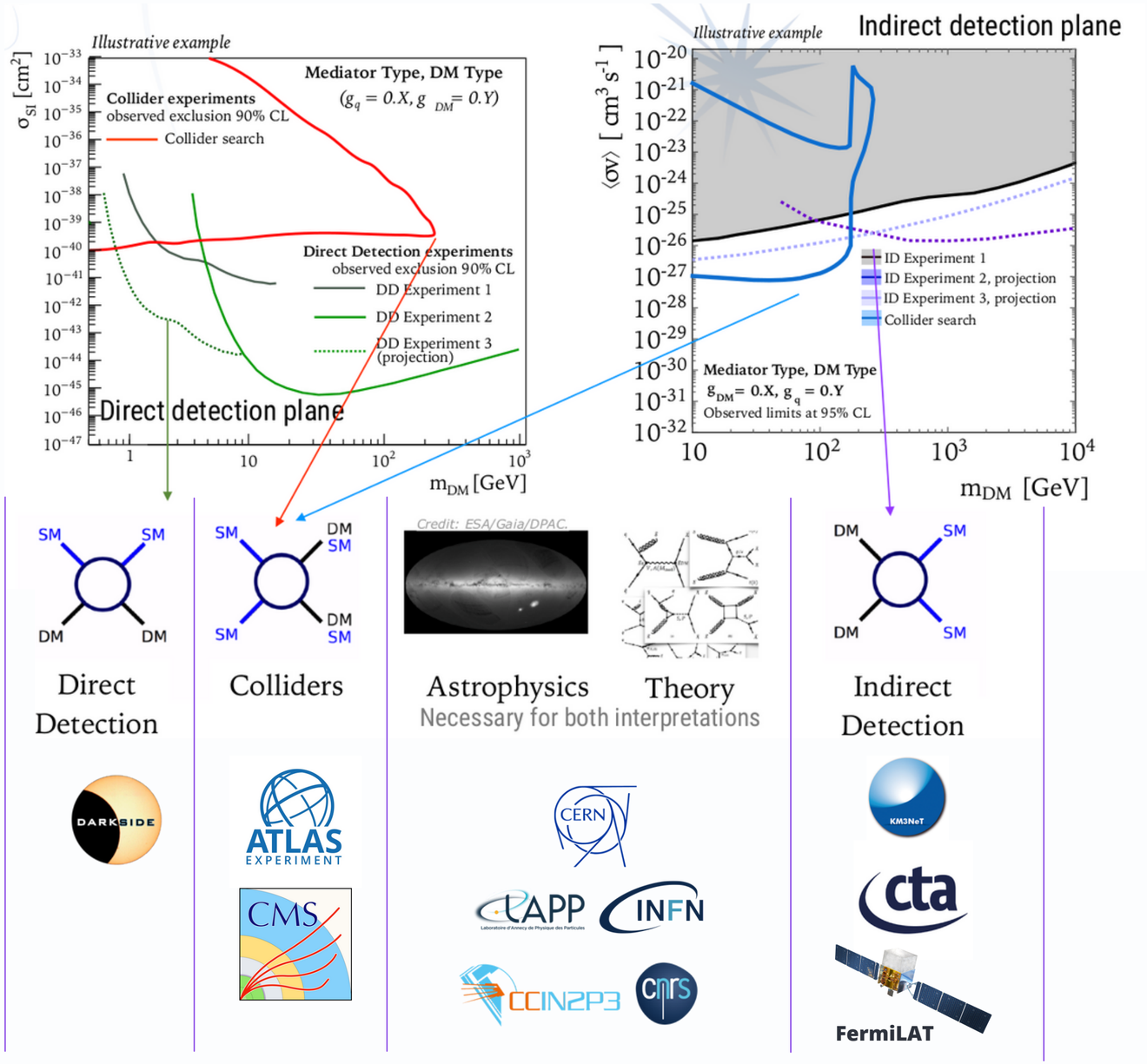
Such a scenario requires **interoperable** and **reproducible** analyses

- **comparison** and **combination** of results from different experiments
- **end-to-end workflows** available for cross-checks

with the **Dark Matter Science Project**, we build a **prototype** that fulfils these requirements

Science outputs of the Dark Matter SP

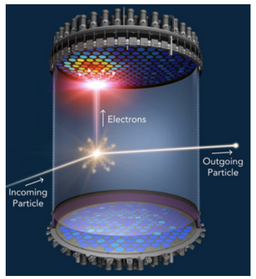
- Individual results and publications
- Plots highlighting complementarity of different experimental efforts
- Data and software objects + pipelines
- Data on the Data Lake, and software on the ESCAPE Software Catalogue
- Pipelines accessible via VRE
- **Combination of experimental results**

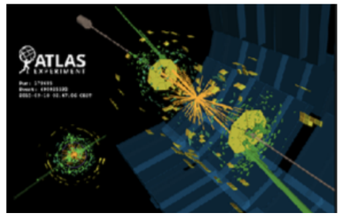


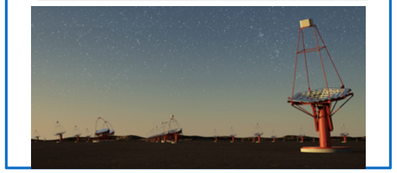
Analysis Workflows for the DM Science Project

Generation & simulation

Expt. data







Credit: SLAC/LZ/ATLAS/CTA

Data **processing**

JupyterHub
(smaller workflows)

REANA
(larger workflows)

Analysis of events/distributions

Virtual Research Environment

Interpretation of results

Combination of results with other searches/experiments

Comparison of results with other searches / experiments

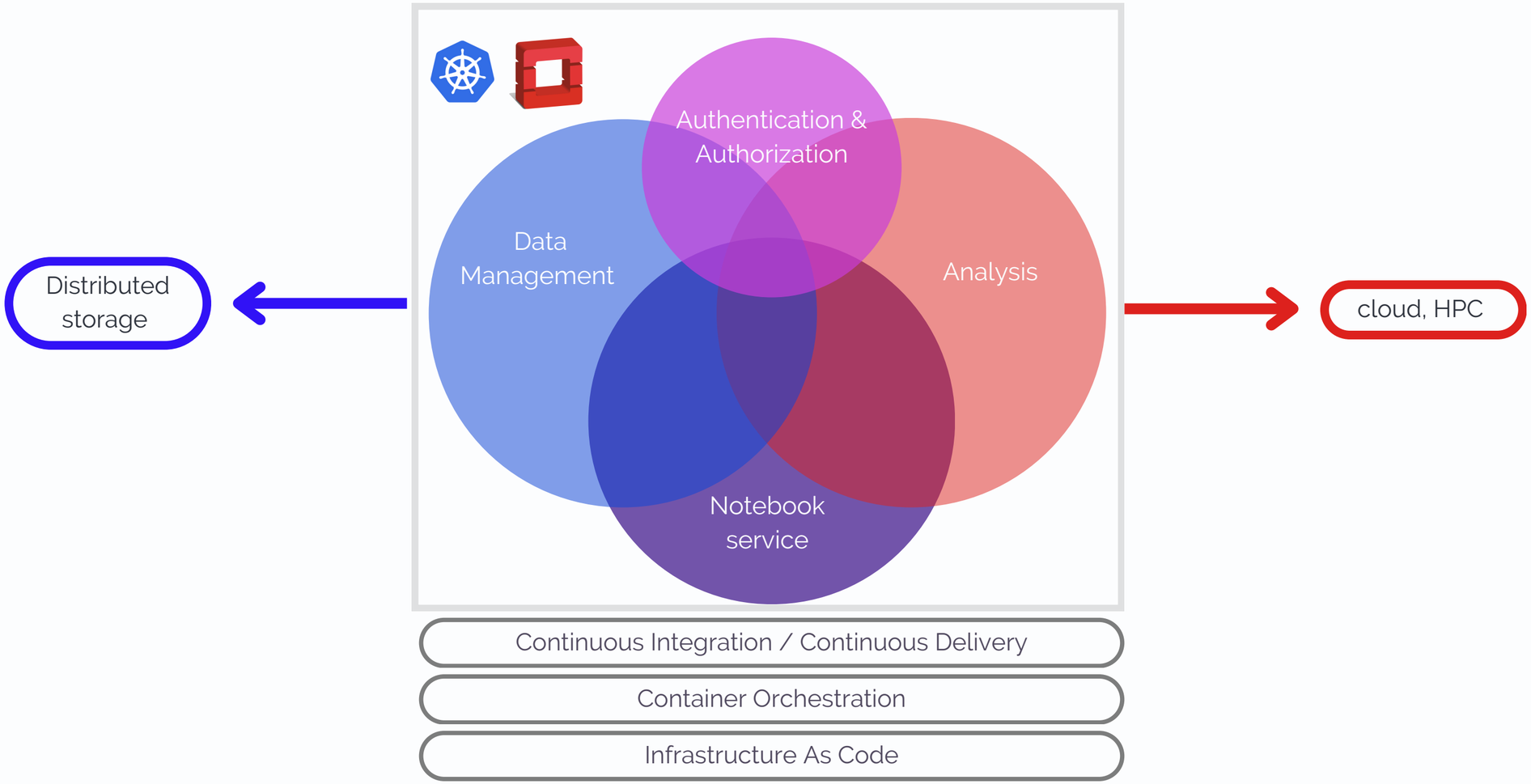
ESCAPE Data Lake

ESCAPE Software Catalogue

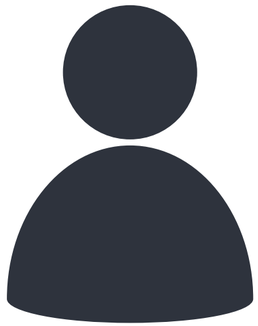
Data sharing and data processing

Data analysis, preservation and interpretation

The VRE building blocks



Authentication & Authorisation



INDIGO Identity and Access Management (IAM) - adopted by WLCG for token

- OIDC tokens
- X.509 certificates / one VO for all the experiments

@ INFN CNAF

Authentication & Authorization

subject mapping cronjob

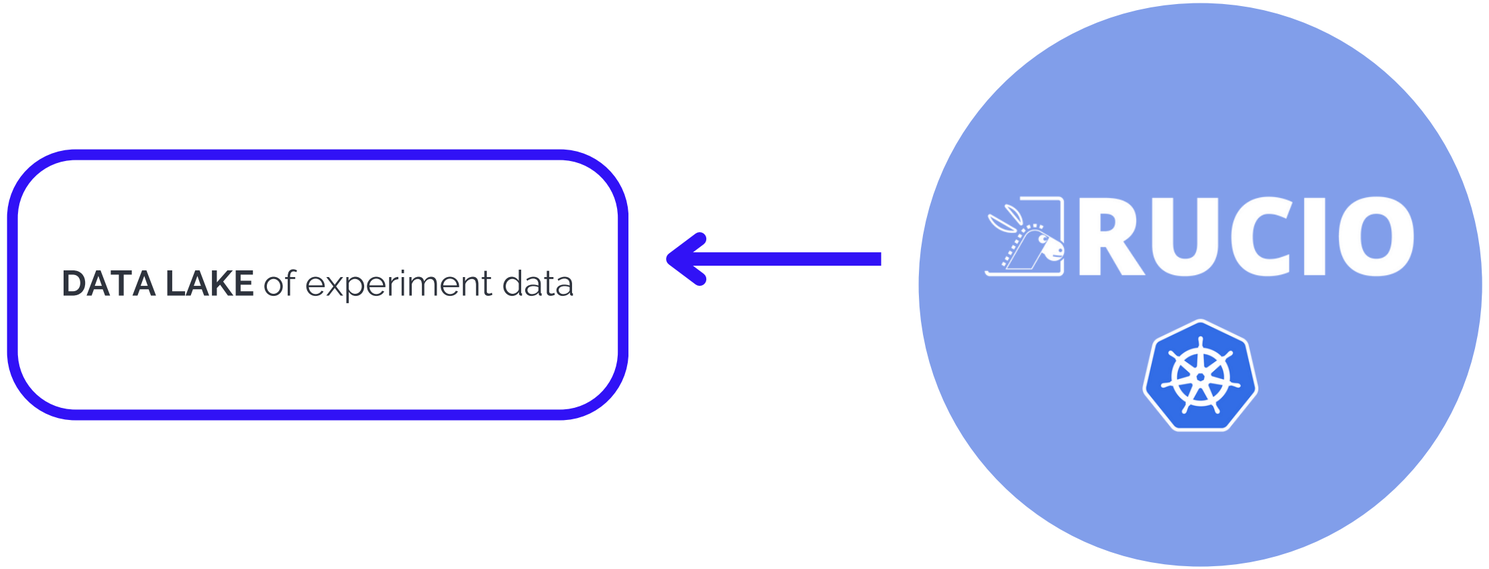
Data Management

Notebook service

Analysis

CERN EOSC Future ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures European Commission

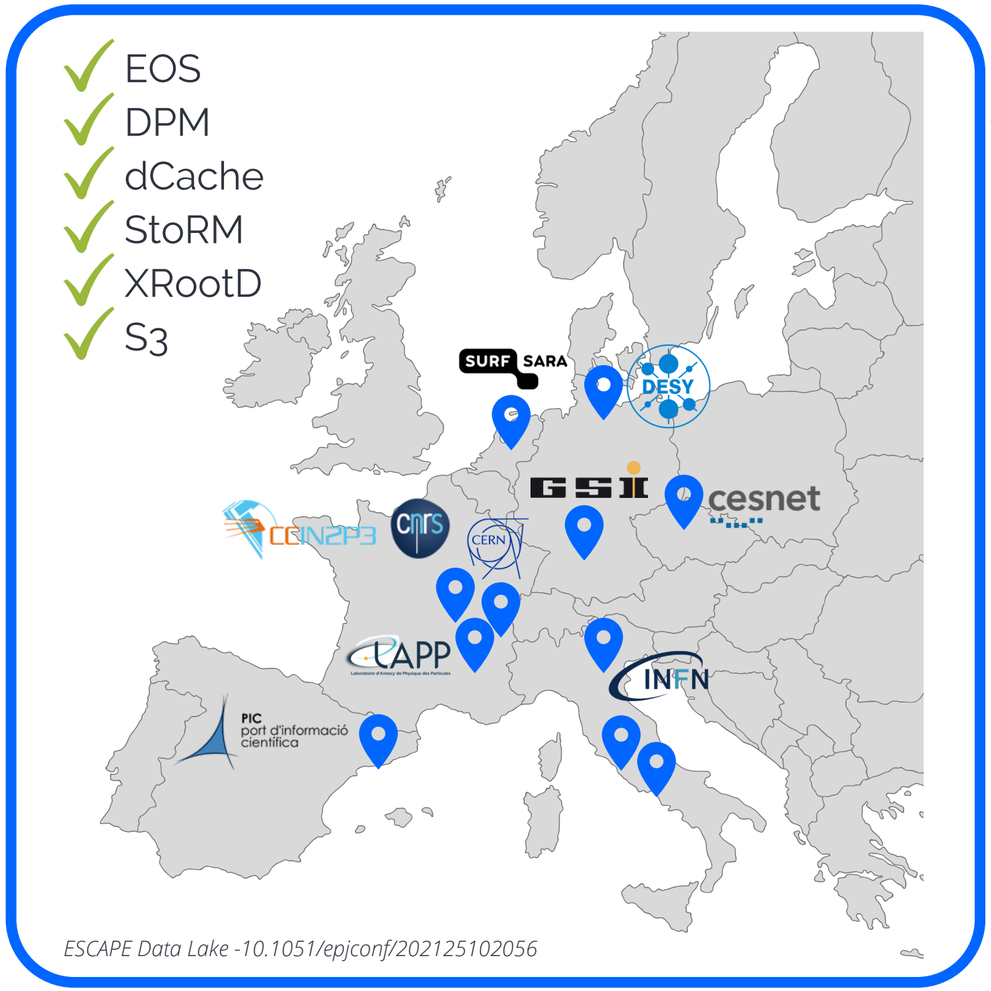
Data Management



Inherited from the ESCAPE WP2 data management group at CERN -> evolved into a GitHub hosted, k8s managed Openstack cluster maintained to comply with latest Rucio software developments.

Rucio instance

Data Lake



Central Relational Database

FTS3and GFAL2

Daemons for data access & replications through gridFTP, HTTP(S)/webDAV, XRoot protocols

Main & authorisation servers

CERN EOSC Future ESCAPE European Commission

Notebook Service

https://jhub-vre.cern.ch/

interface to run preliminary analysis

containerised environments on public repositories

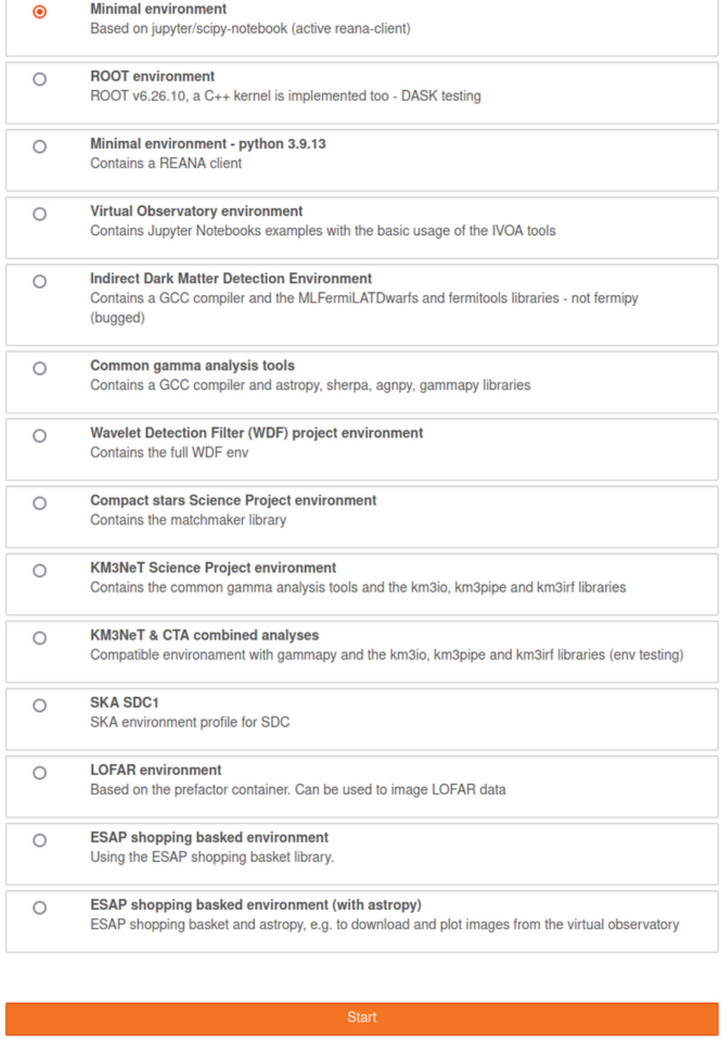
EOS accessible from all images jhub nodes

CERN Virtual Machine FS (CVMFS) installed and accessible from all images

client libraries and software installed to interact with underlying services

CephFS volumes provided as shared, temporary storage solution

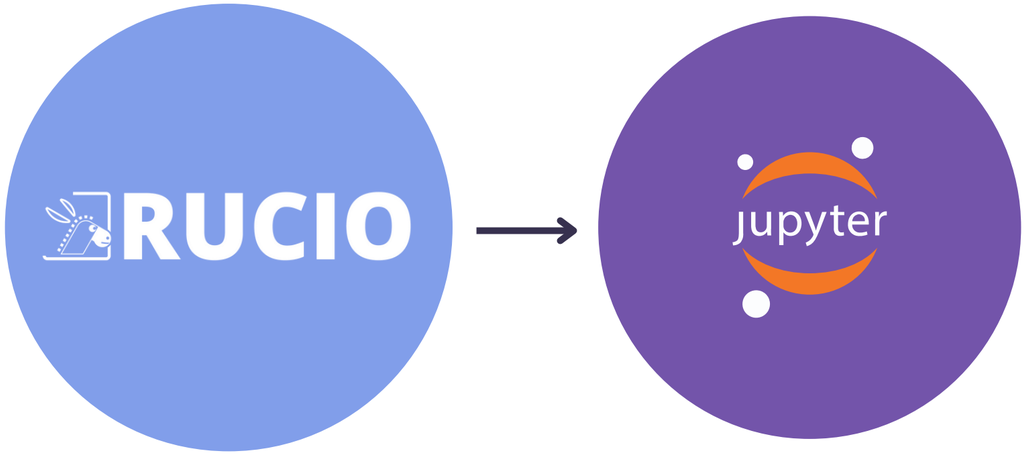
Rucio VRE instance access

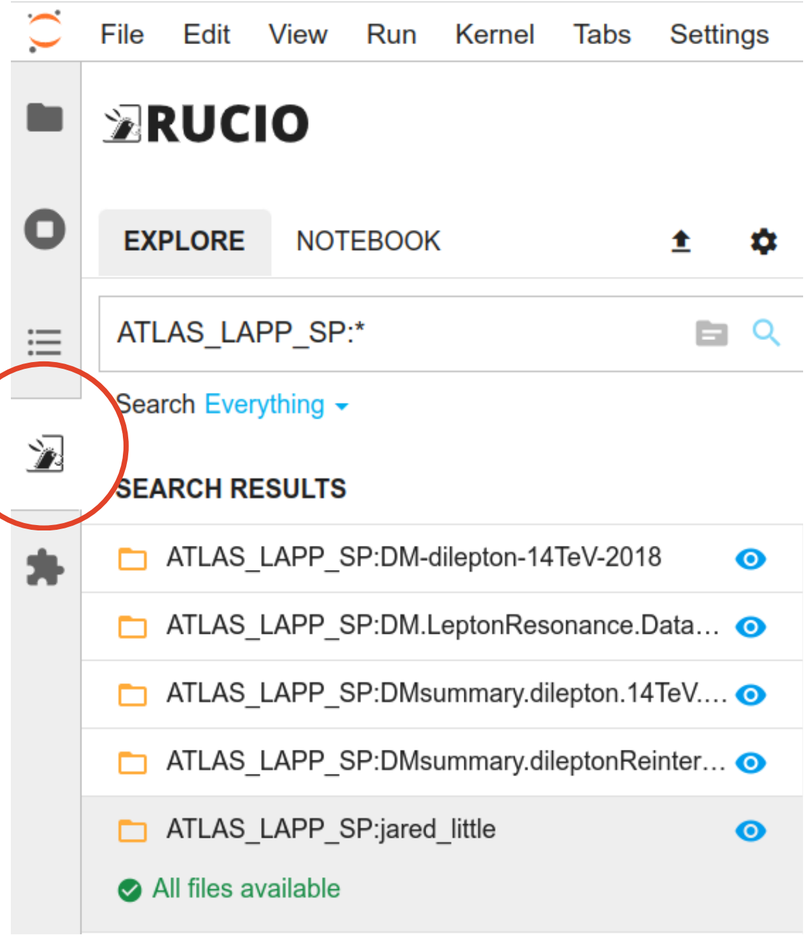


Data into the notebook

The **Jupyterhub Rucio extension** hides the complexity of the Data Lake and allows users to

- browse experiments' data catalogue
- authenticate with OIDC tokens to the Rucio infrastructure
- replicate data into the notebook
- import the data into the notebook by assigning a parameter to it
- run preliminary analysis to prototype code





CERN EOSC Future ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures European Commission

Data into the notebook

Data gets replicated through Rucio daemons from any storage element to an EOS storage element of half a Petabyte FUSE mounted on the Jupyterhub node.

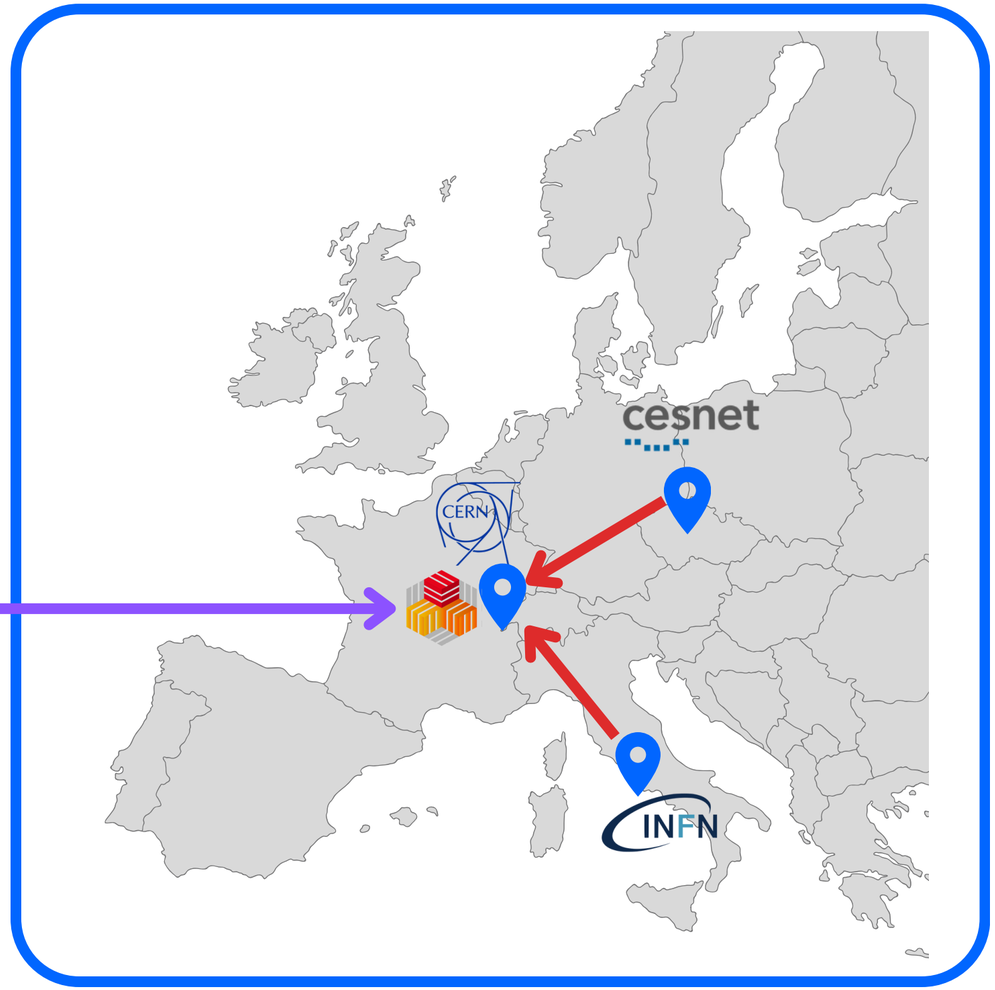
The computation is limited to the CPU capacity of the node.

How do we SCALE OUT?

Openstack
Jupyterhub node

FUSE mount

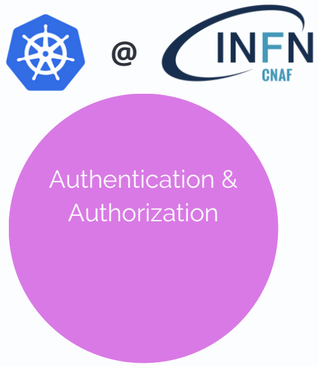
Data Lake

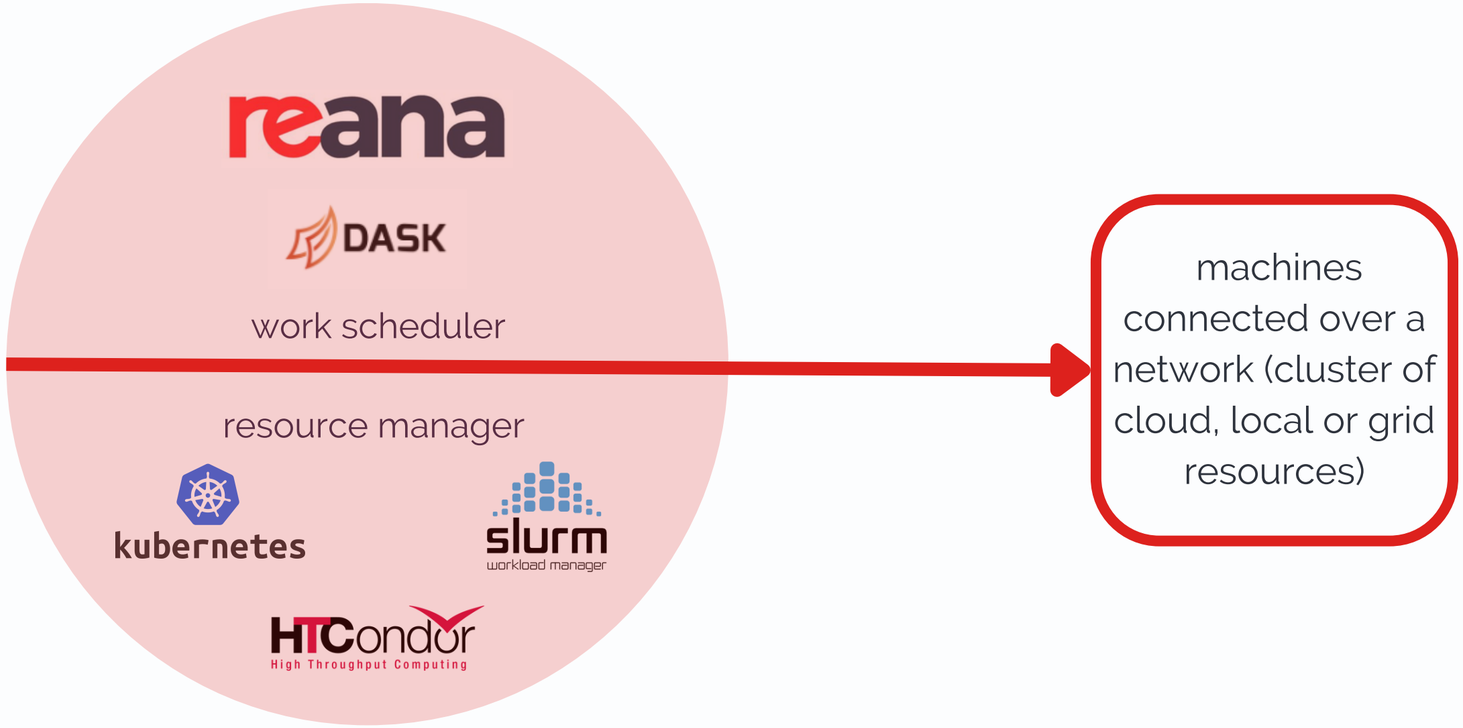


Computing

- **Distribute** the analysis
  - **resource managers** (Kubernetes, HTCondor (High Throughput Computing (HTC)) and Slurm (High Performance Computing (HPC))
  - **work schedulers** (Dask, Reana, Spark)

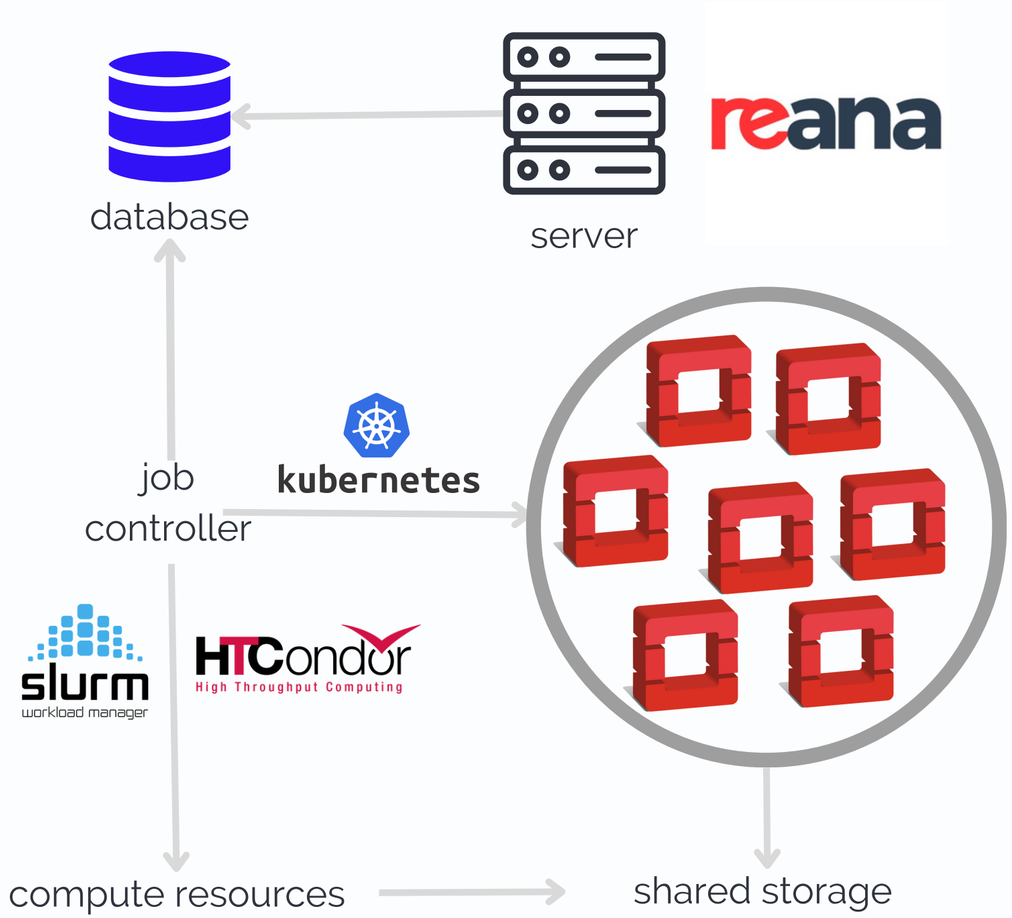
- **Preserve** the analysis for reuse
  - work schedulers (Reana)





CERN EOSC Future ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures European Commission

Analysis preservation and distribution

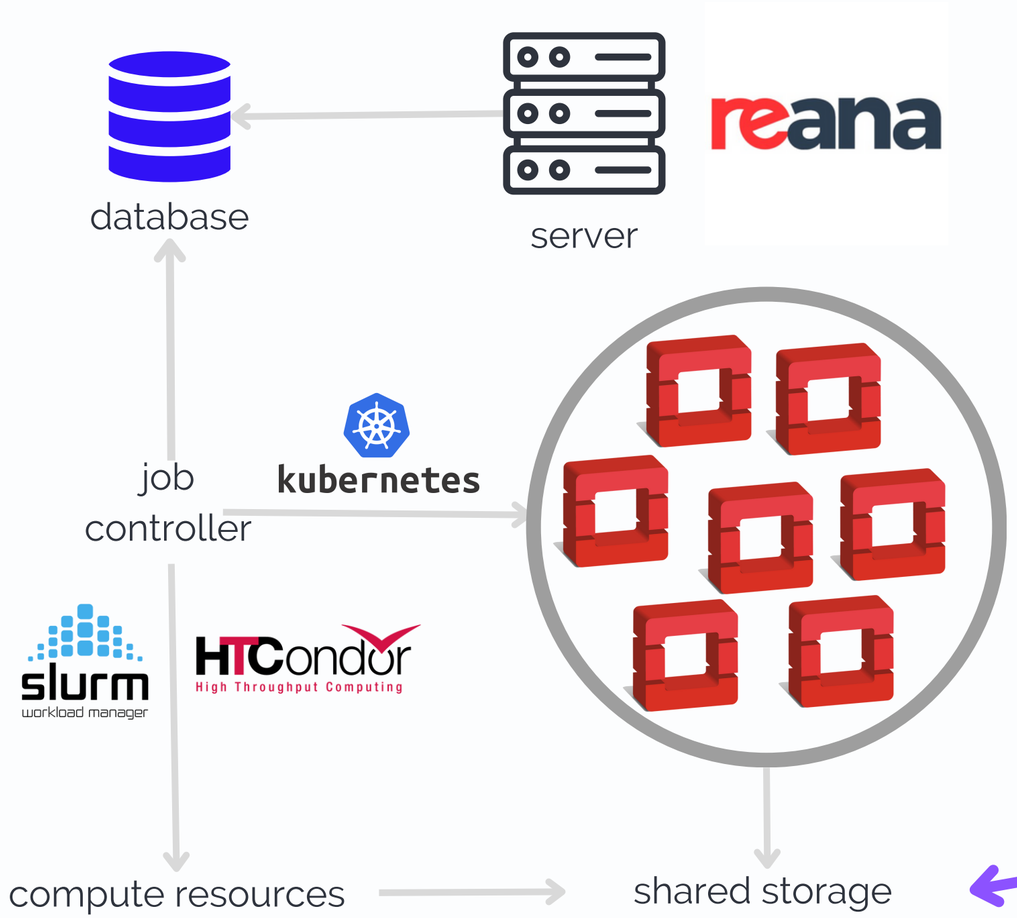


Reana is a reproducible analysis project developed at CERN, to make the preservation of heavier analyses seamless.

- Easily installed via Helm
- Intuitive declarative programming approach (reana.yaml file) with:
  - input data
  - environment
  - code
  - computational steps
- Isolates each step with different containers
- Supports workflow engines
  - CWL
  - Snakemake
  - Yadage --> workflow concatenation (output becomes input)

https://reana-vre.cern.ch/

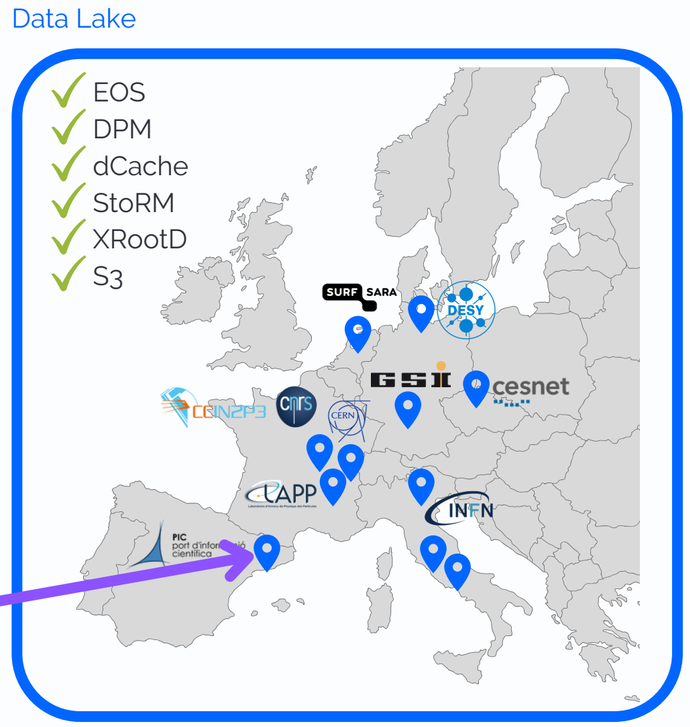
Non-local analysis preservation



From the Reana client:

- authenticate via IAM to Rucio with a side-car container
- get data from distributed storage

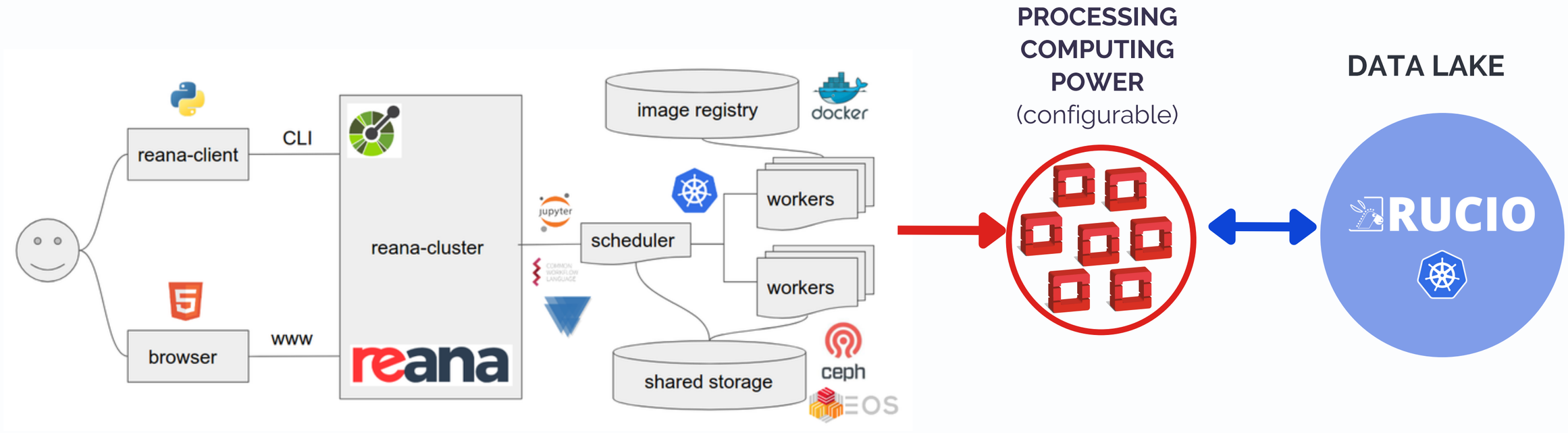
→ the analysis can be reproduced fully and **independently from local storage**



CERN EOSC Future ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures European Commission

Analysis preservation and distribution

Reana makes preservation of heavier analyses seamless

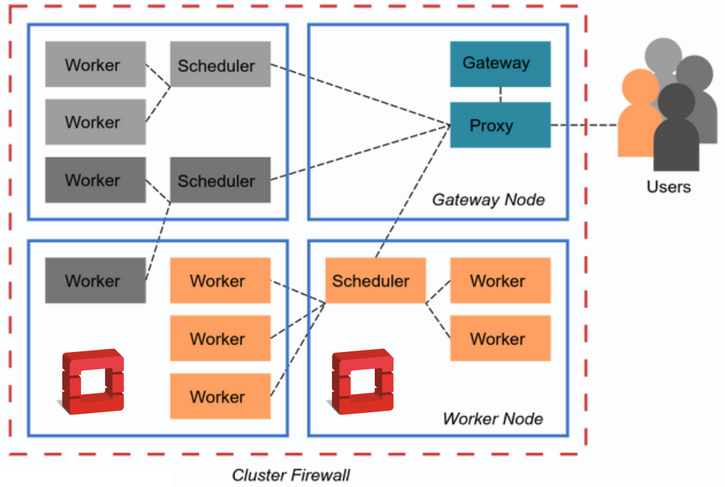


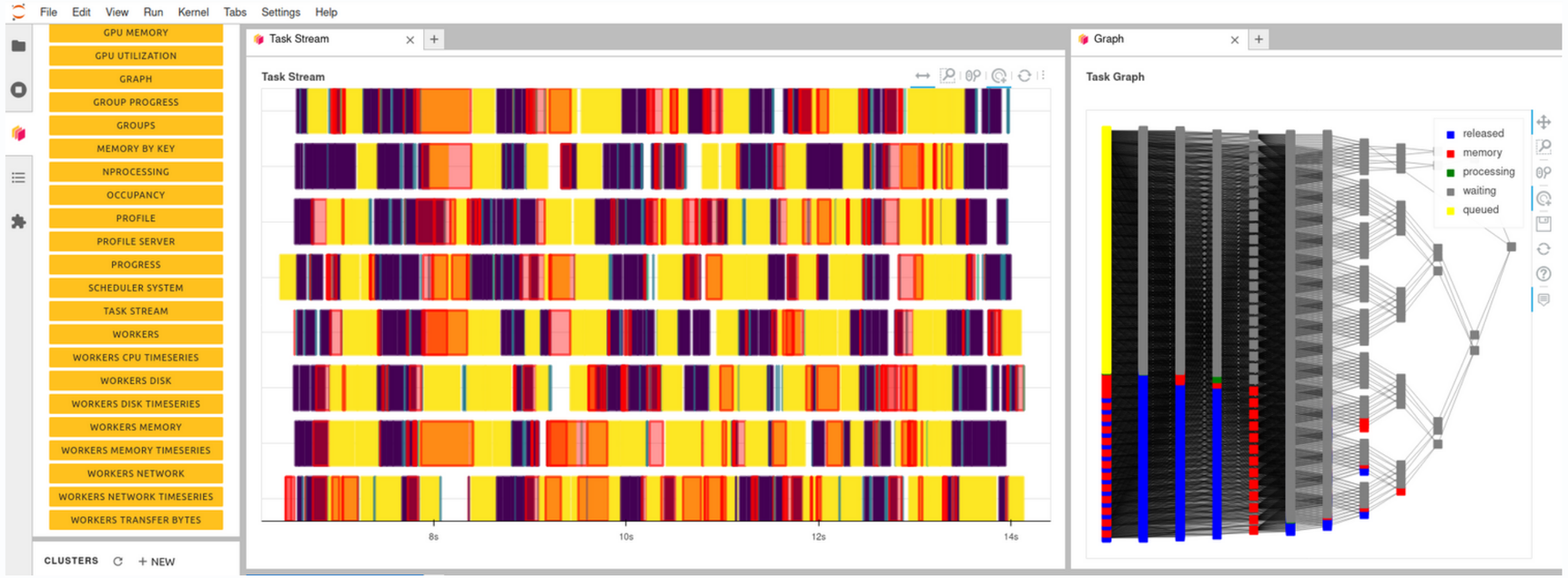
CERN EOSC Future ESCAPE European Commission

Workflow distribution with Dask

Daskhub helm chart: Dask Gateway + Jupyterhub

- multi-user, configurable usage profiles
- gateway to distribute access to all cloud nodes of the VRE
- code needs to be adapted
- dashboards of work progress

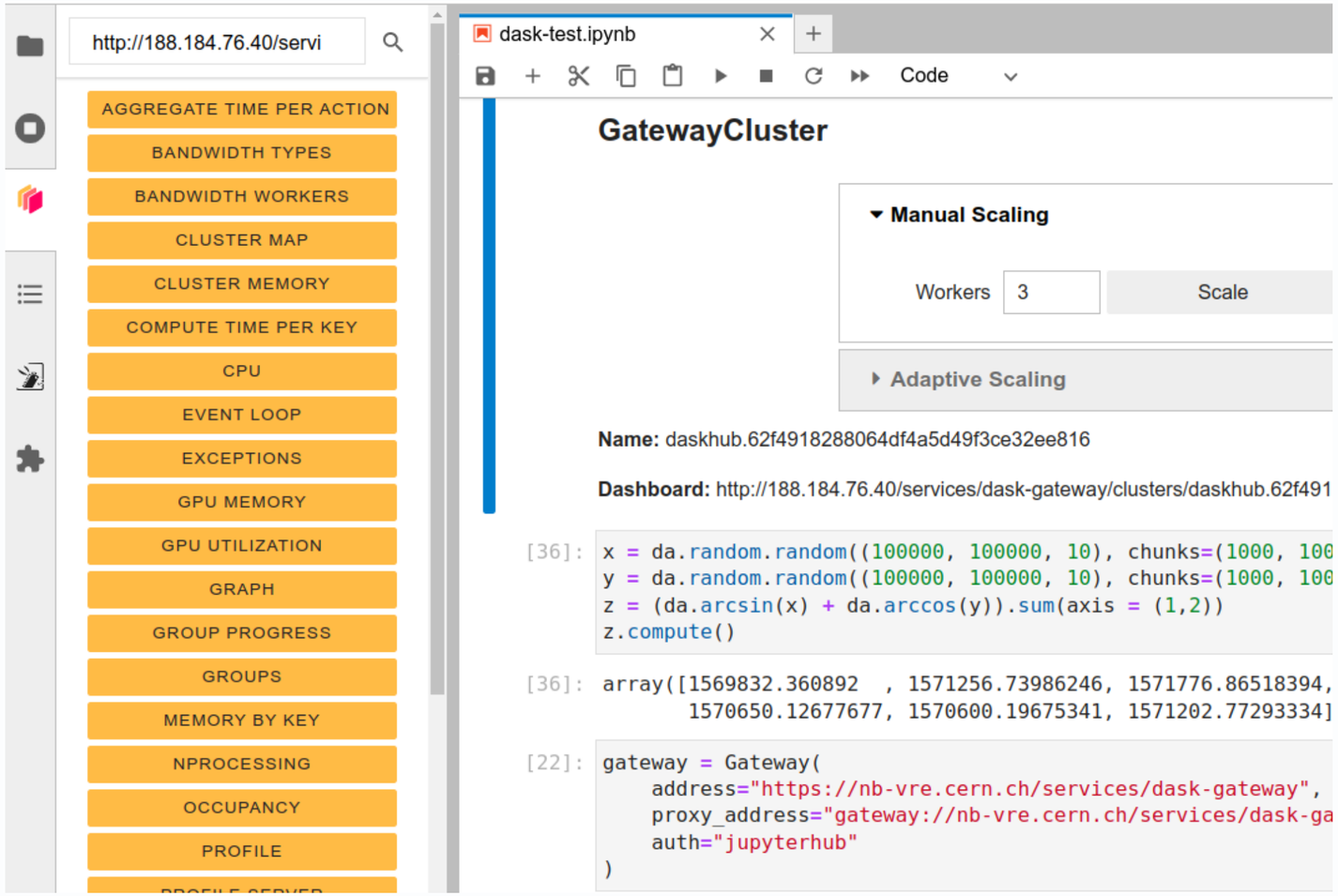




jupyter + DASK

Workflow distribution with Dask

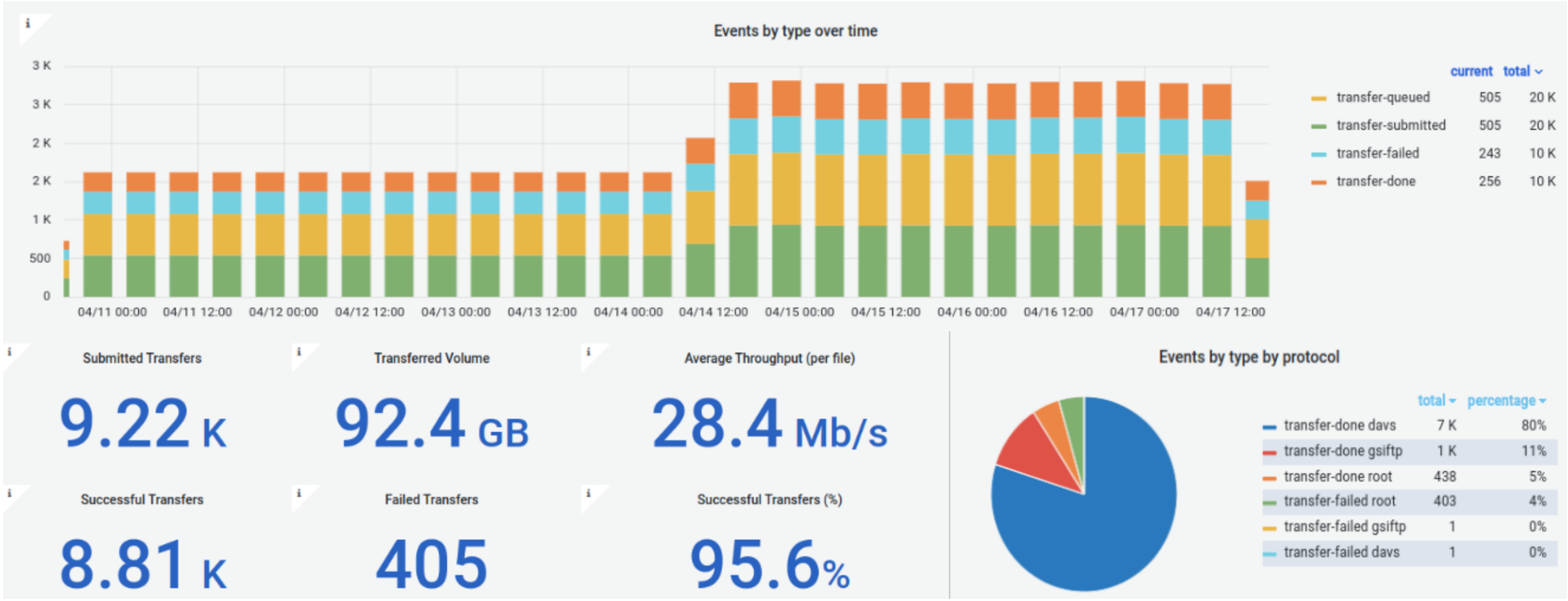
Work in progress:
Docker image with Rucio Jupyterlab extension and Dask, giving the possibility to spawn a cluster to each user



Monitoring, testing, dashboards, on-boarding

- Continuous **monitoring and testing** of transfers between Rucio Storage Elements (RSEs) is in place on Grafana dashboards hosted at CERN.

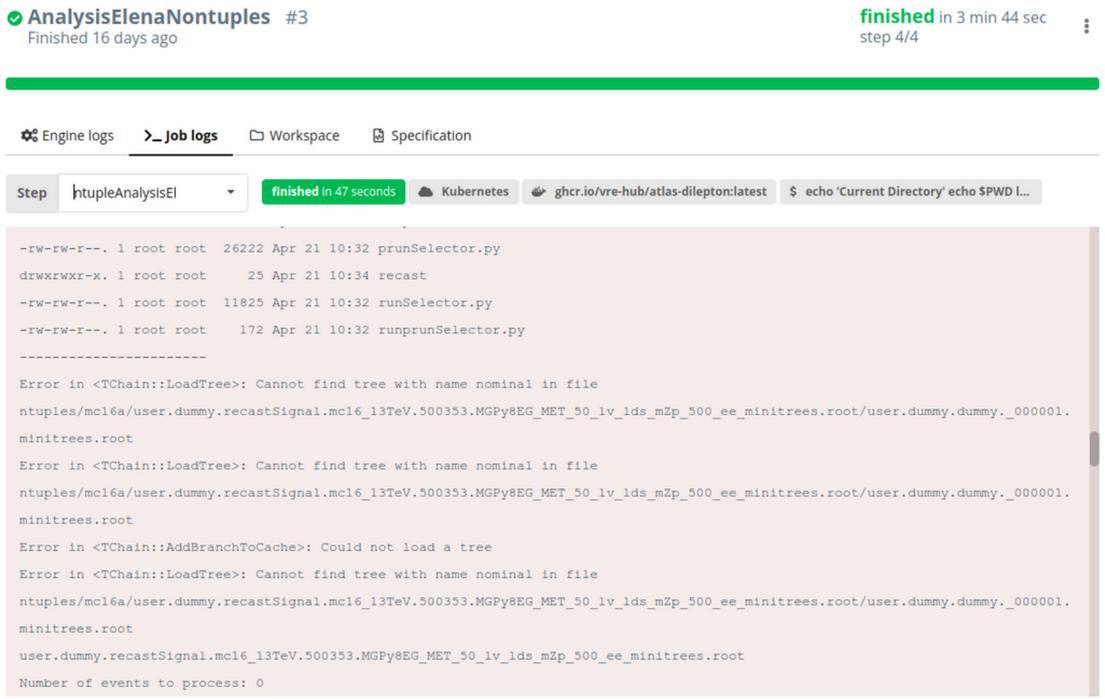
https://monit-grafana-open.cern.ch/d/PJ65OqBVz/vre-rucio-events?orgId=16&from=now-7d&to=now



Monitoring, testing, dashboards, on-boarding

- **Rucio and Reana UI** interfaces deployed with K8s allow to explore and debug failed transfers and workflows.

https://reana-vre.cern.ch/



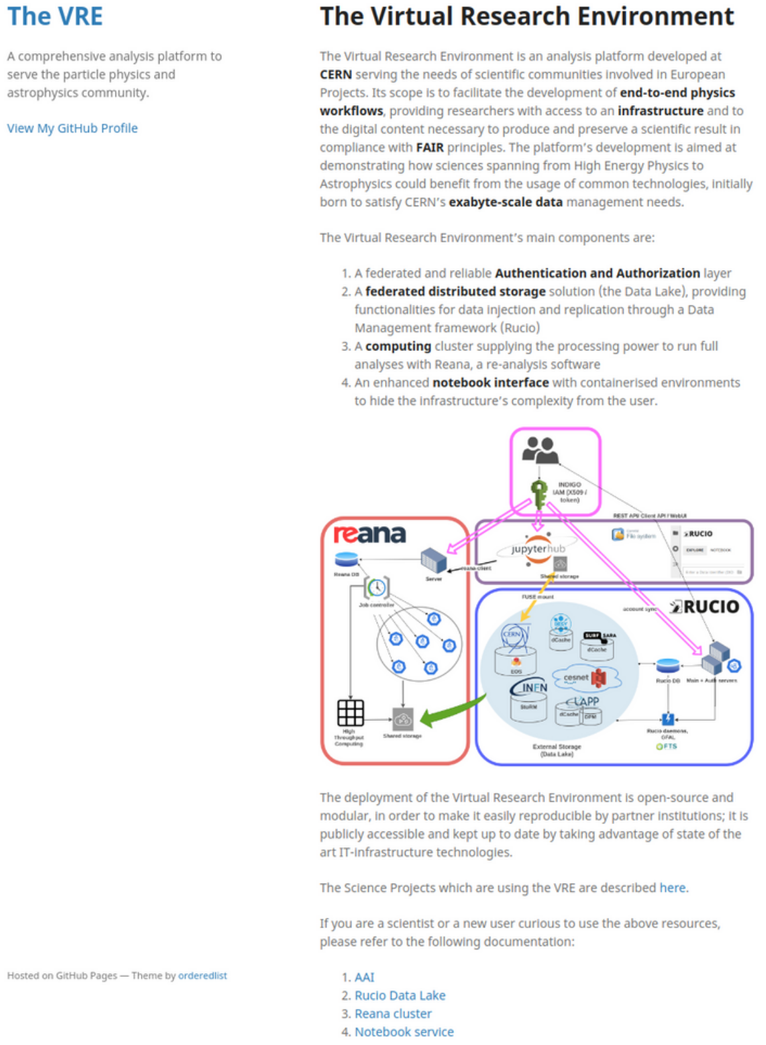
https://vre-rucio-ui.cern.ch

Name	Account	RSE Expression	Creation Date	Remaining Lifetime	State
elena_test:2023.03.16-11.19.03.txt	egazzarr	EULAKE-1	2023-05-07T13:22:23.000Z	7d	STUCK
user.ron:test_from_CERN-030523_1643.txt	garcia	SURF-IOP-EXP	2023-05-04T10:35:14.000Z	-	STUCK
user.ron:test_from_CERN-030523_1643.txt	garcia	EULAKE-1	2023-05-03T14:43:27.000Z	-	OK
user.ron:mytestfile_2	garcia	DESY-DCACHE	2023-05-03T14:35:27.000Z	-	OK
elena_test:test-file-rucio-2023-04-24-01.txt	egazzarr	PIC-DCACHE	2023-04-24T14:13:33.000Z	-	OK
elena_test:test-file-rucio-2023-04-24-02.txt	egazzarr	PIC-DCACHE	2023-04-24T14:12:45.000Z	-	REPLICATING
elena_test:test-file-rucio-2023-04-24-01.txt	egazzarr	EULAKE-1	2023-04-24T14:12:12.000Z	-	OK
elena_test:test-file-rucio-2023-04-20-04.txt	egazzarr	IN2P3-CC-DCACHE	2023-04-20T15:08:51.000Z	-	REPLICATING
elena_test:test-file-rucio-2023-04-20-03.txt	egazzarr	DESY-DCACHE	2023-04-20T15:06:00.000Z	-	REPLICATING
elena_test:test-file-rucio-2023-04-19-01.txt	egazzarr	SURF-IOP-EXP	2023-04-19T15:53:19.000Z	-	STUCK
elena_test:test-file-rucio-2023-04-19-01.txt	egazzarr	IN2P3-CC-DCACHE	2023-04-19T15:42:32.000Z	-	OK
elena_test:test-file-rucio-2023-04-19-01.txt	egazzarr	EULAKE-1	2023-04-19T15:35:53.000Z	-	OK
elena_test:test-file-rucio-2023-04-19-01.txt	egazzarr	DESY-DCACHE	2023-04-19T15:33:53.000Z	-	OK
elena_test:test-file-rucio-2023-04-19-01.txt	egazzarr	CESNET-S3	2023-04-19T15:33:34.000Z	-	OK

Monitoring, testing, dashboards, on-boarding

- **Documentation** is hosted on Github pages and is made easy for both users and system administrators who would like to get inspired by the VRE model.

https://vre-hub.github.io/



The VRE

A comprehensive analysis platform to serve the particle physics and astrophysics community.

View My GitHub Profile

Hosted on GitHub Pages — Theme by orderedlist

The Virtual Research Environment

The Virtual Research Environment is an analysis platform developed at **CERN** serving the needs of scientific communities involved in European Projects. Its scope is to facilitate the development of **end-to-end physics workflows**, providing researchers with access to an **infrastructure** and to the digital content necessary to produce and preserve a scientific result in compliance with **FAIR** principles. The platform's development is aimed at demonstrating how sciences spanning from High Energy Physics to Astrophysics could benefit from the usage of common technologies, initially born to satisfy CERN's **exabyte-scale data** management needs.

The Virtual Research Environment's main components are:

1. A federated and reliable **Authentication and Authorization** layer
2. A **federated distributed storage** solution (the Data Lake), providing functionalities for data injection and replication through a Data Management framework (Rucio)
3. A **computing** cluster supplying the processing power to run full analyses with Reana, a re-analysis software
4. An enhanced **notebook interface** with containerised environments to hide the infrastructure's complexity from the user.

The deployment of the Virtual Research Environment is open-source and modular, in order to make it easily reproducible by partner institutions; it is publicly accessible and kept up to date by taking advantage of state of the art IT-infrastructure technologies.

The Science Projects which are using the VRE are described here.

If you are a scientist or a new user curious to use the above resources, please refer to the following documentation:

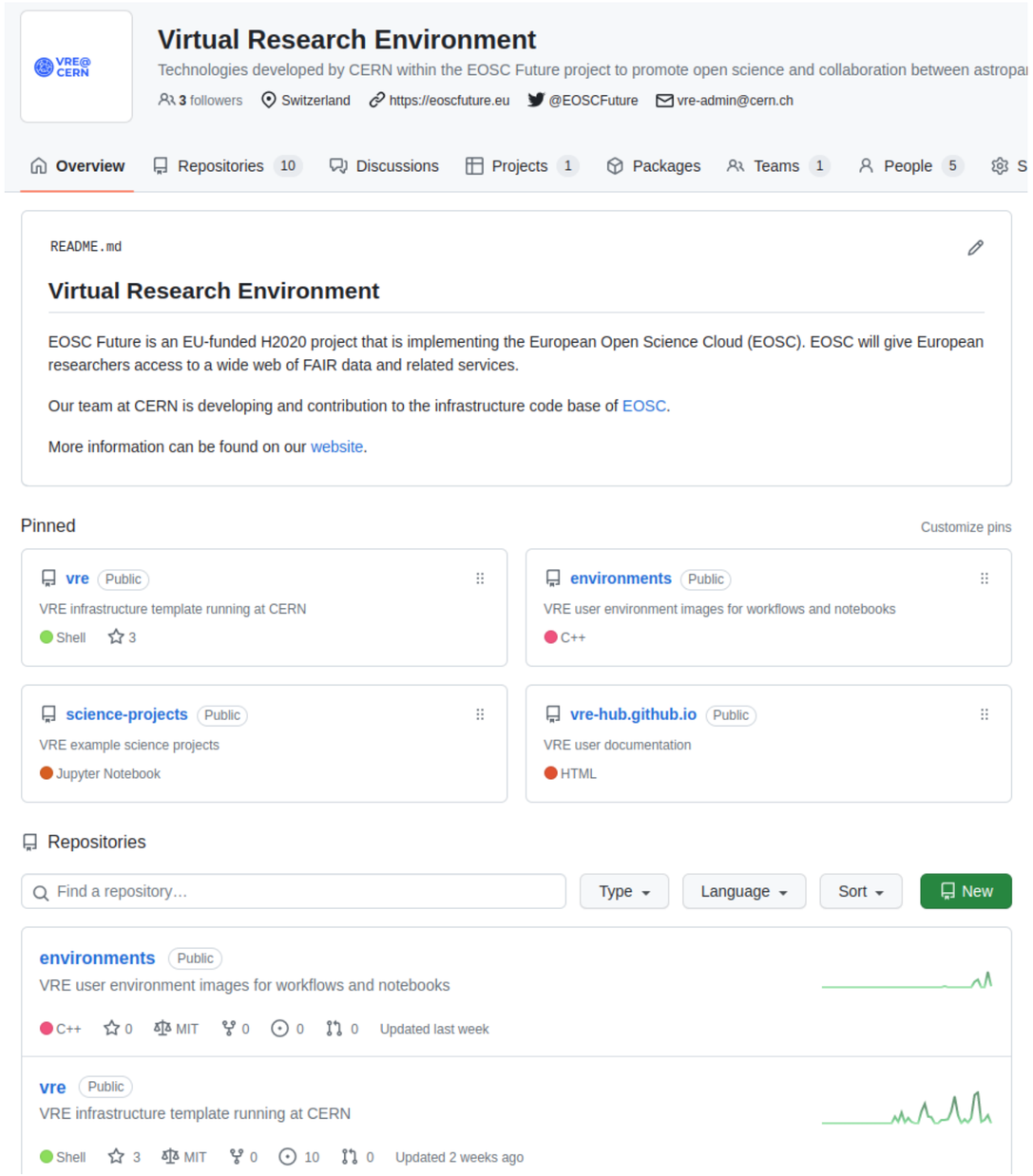
1. AAI
2. Rucio Data Lake
3. Reana cluster
4. Notebook service

Deployment

VRE public Github repository hosts

- cloud deployment of the infrastructure components with Helm, Flux, Terraform and K8s
- Science Projects software to produce the environments for the Jupyterhub instance
- scientific code to be shared
- reana.yaml files to reproduce the analysis
- forums and discussions

https://github.com/vre-hub

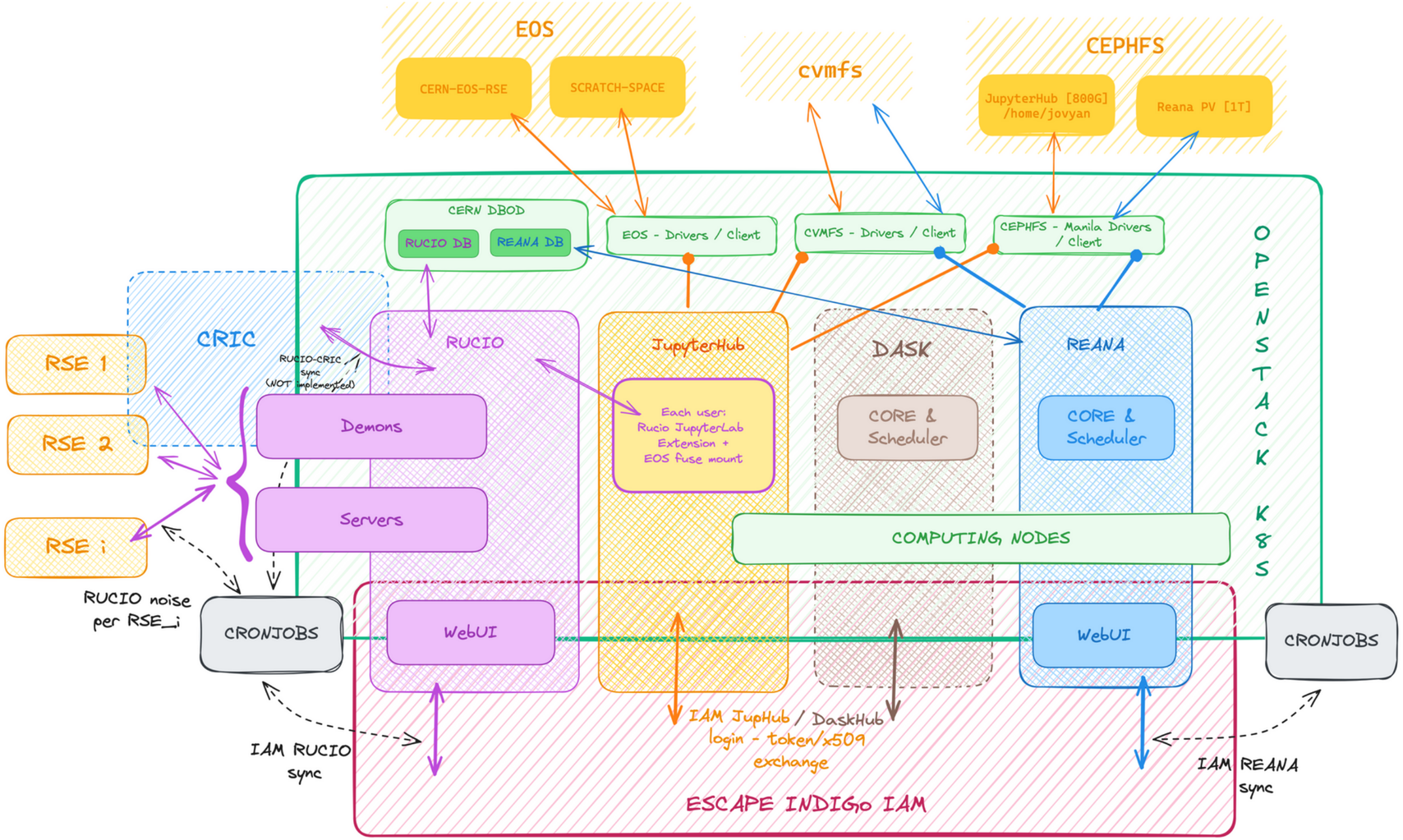


Support and meetings

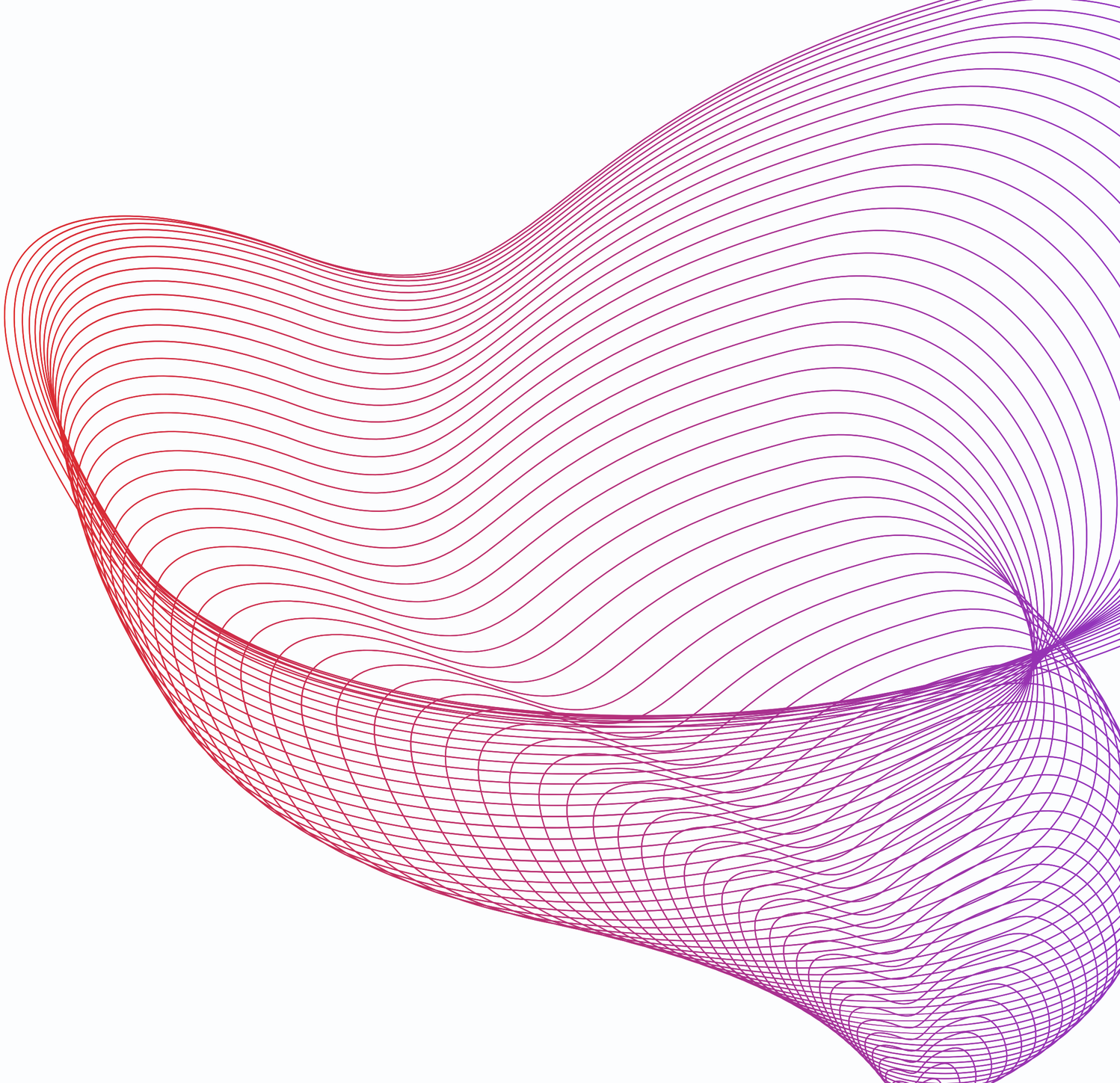
- VRE users support meeting every 2 Tuesdays @14

- VRE working group meeting 1st Thursday of the month @11
  - for developers and system administrators who are building and deploying scientific infrastructures within the broad field of scientific analysis platforms

CERN VRE architecture: a closer look

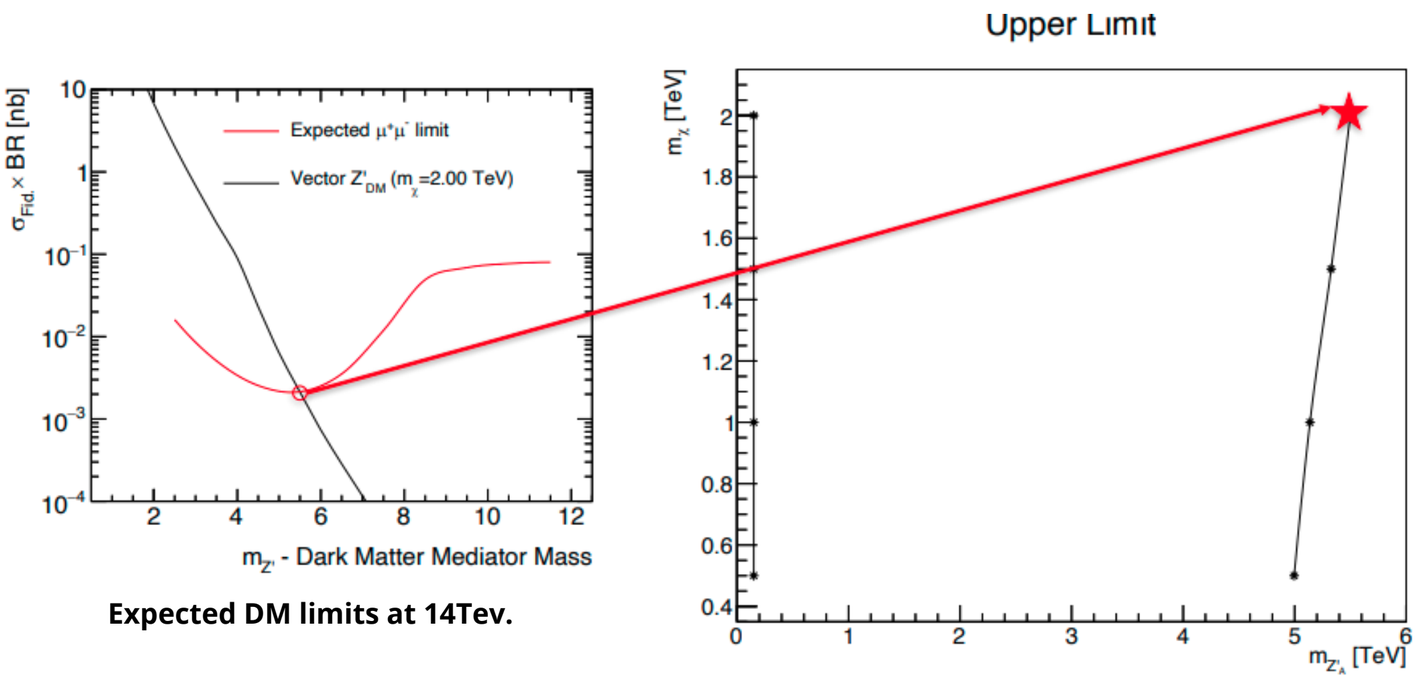


Demo

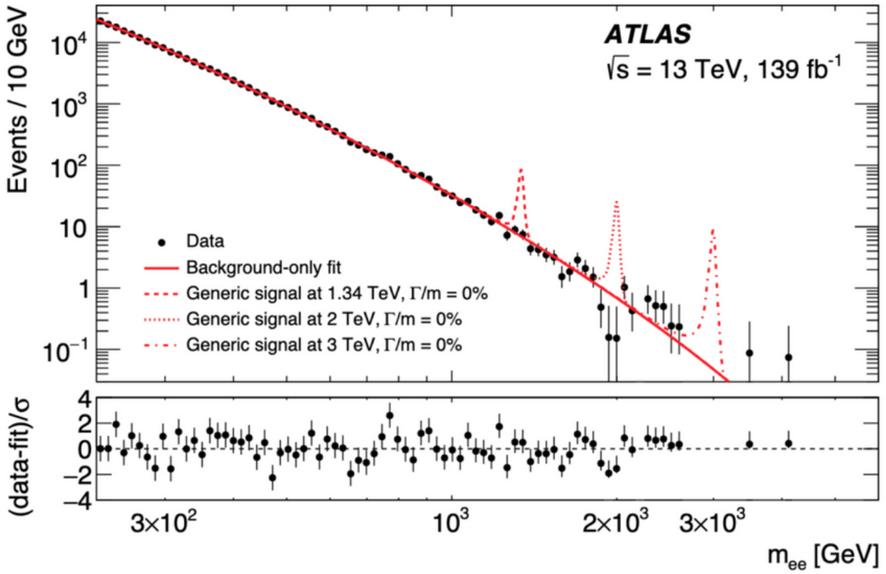


DM@LHC with ATLAS

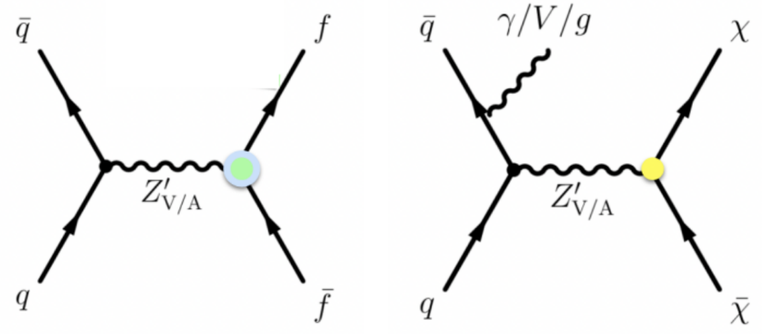
1. **Dark Matter Reinterpretation**: setting limits on High-Luminosity LHC contraints on $Z' \rightarrow \chi\chi$ (Z' mediated Dark Matter models).
2. The **dilepton** inclusive search (right) concluded in 2019
   a. objective: projecting limits to 14 TeV and computing the fiducial cross-sections in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



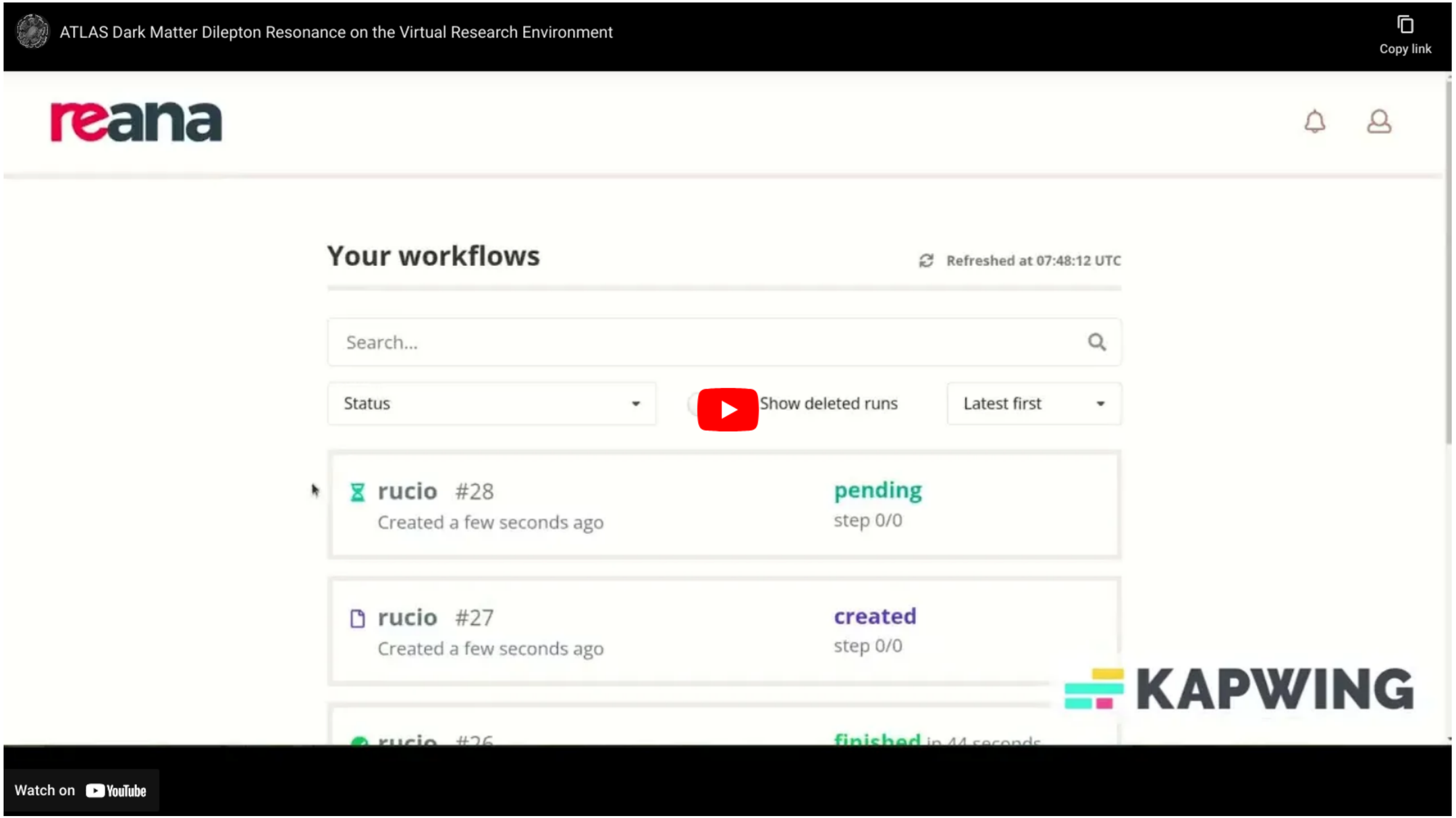
Expected DM limits at 14Tev.



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



See yesterday's talk by J. Little



https://www.youtube.com/watch?v=hvlJLo_7xXc&ab_channel=ElenaGazzarrini

Success stories

- Escape Open Collaboration Agreement ensures the collaboration and joint common activities across scientific communities in the development of VREs
- VRE awoke interest from scientific domains who are in early-stage prototype phase (i.e. Einstein telescope)
- Interest from new digital models (i.e. digital twins) developed within European projects

Future outlook

VRE future steps:

- connection with HPCs, commercial clouds and other external computing resources
- Caching data on distributed storage
- Engage more scientific communities

- The pursuit of commonalities in Scientific Computing is fundamental to pursue economies of scale at all levels.
- Agreed joint multi-year roadmaps is mandatory if we pursue common approaches in the areas of **Data Management, Data Analysis, Security, Networking and Software.**
- Promoting alignment of efforts and ensuring positive return is challenging but is the way to go
  - --> **VRE + ESAP could align their interests and join their efforts**

CERN EOSC Future ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures European Commission

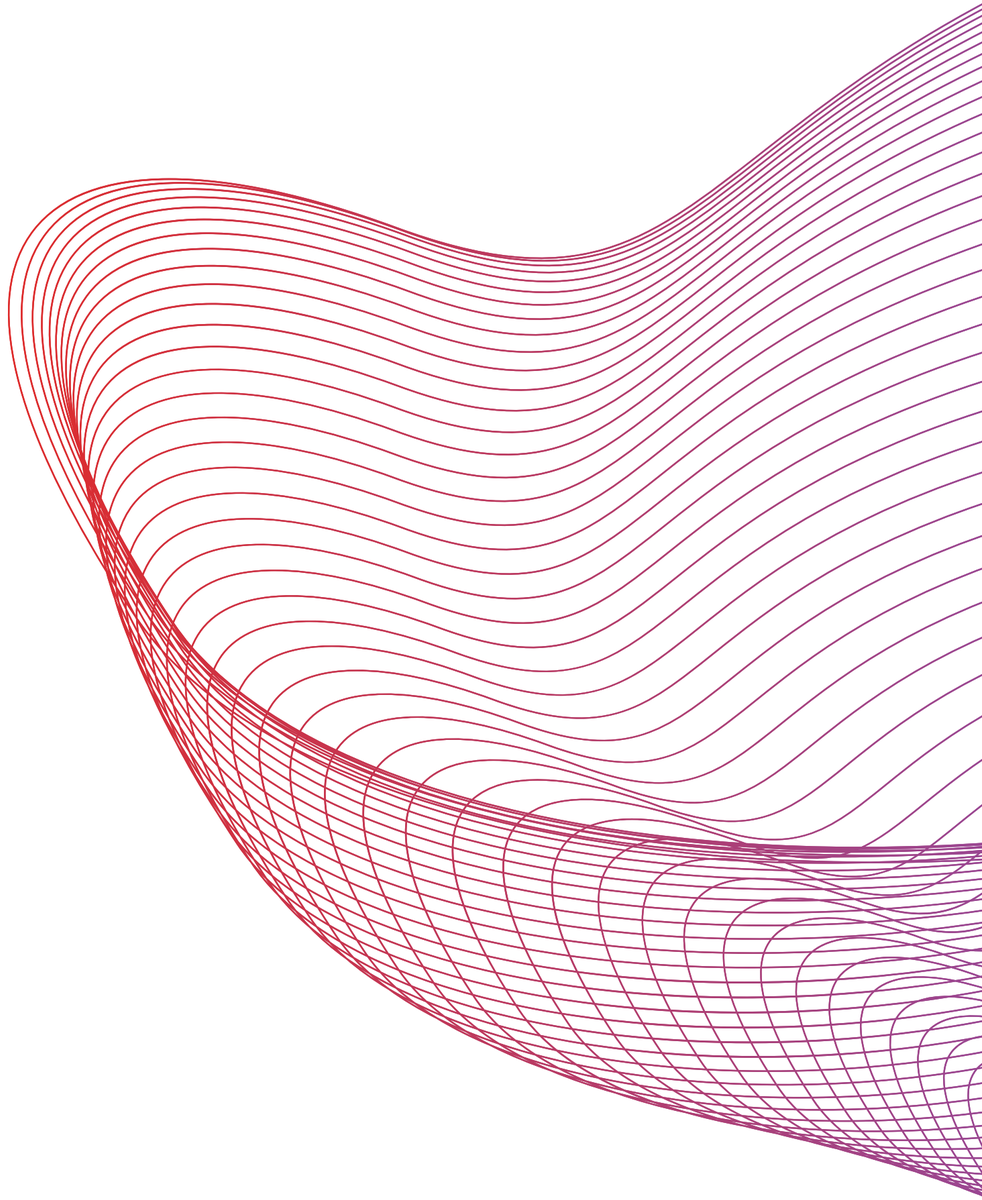
Thank you

special thanks to

Tibor Simko & Reana team, Martin Barisits & Rucio team, Xavier Espinal, Ian Bird, CNAF IAM team and all the Science Projects researchers (Jared Little, Caterina Doglioni, Christopher Eckner, Alexander Ekman, Axel Gallen, Mikhail Smirnov, Francesca Calore, Pooja Bhattacharjee, Valerio Ippolito, Estelle Pons, Elena Cuoco, Alberto Iess, Alessandro Parisi, Dany Vohl)

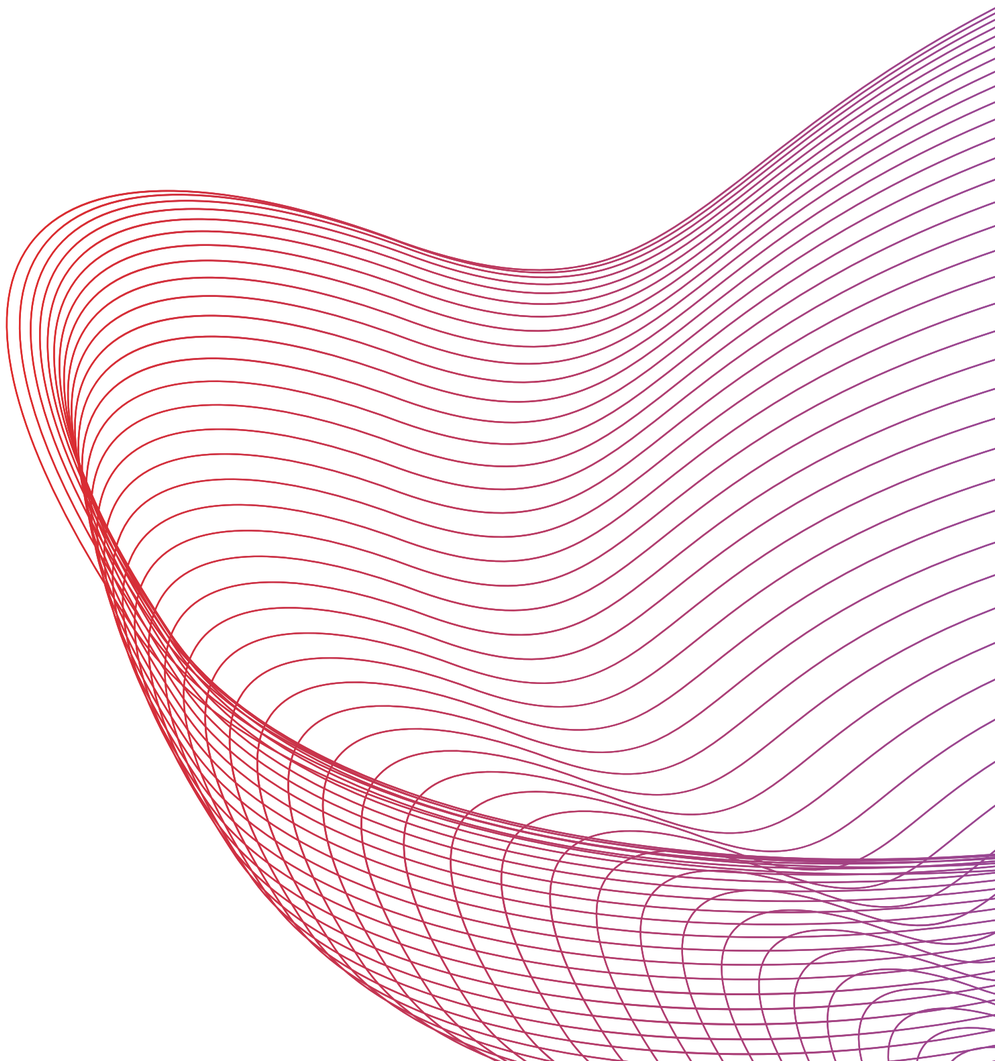
e-mail

elena.gazzarrini@cern.ch



CERN EOSC Future ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures European Commission

Back up



Status

The VRE is an R&D project and it is not a production system. As such, the platform is maintained by a team of 3 people.

For the moment, ~ 230 users subscribed on the IAM platform and have therefore access to the resources.

VRE documentation and links to resources at: https://vre-hub.github.io/.

Links to useful related works are provided by clicking on the underlined text in the slides.

vCPUs	RAM (GB)	Masters	Nodes	Remote Storage (TB)	CephFS (TB)
184	335.8	3	23	646	1.8

25 Openstack machines

- 14.6GB RAM
- 8 VCPU
- 80GB Disk
- Fedora CoreOS 35
- LINUX