Virtual Research Environment: an end-user perspective

Sukanya Sinha

WLCG/HSF Workshop, Paris, May 2025

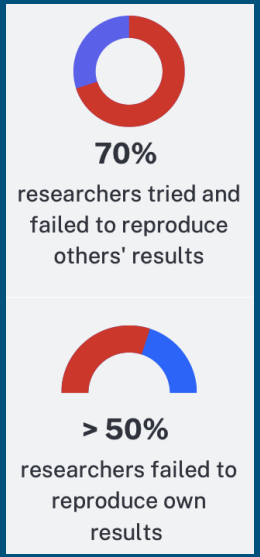
06/05/2025

University of Manchester

…with inputs from C. Doglioni, E. Garcia and G. Guerrieri!

The big picture...

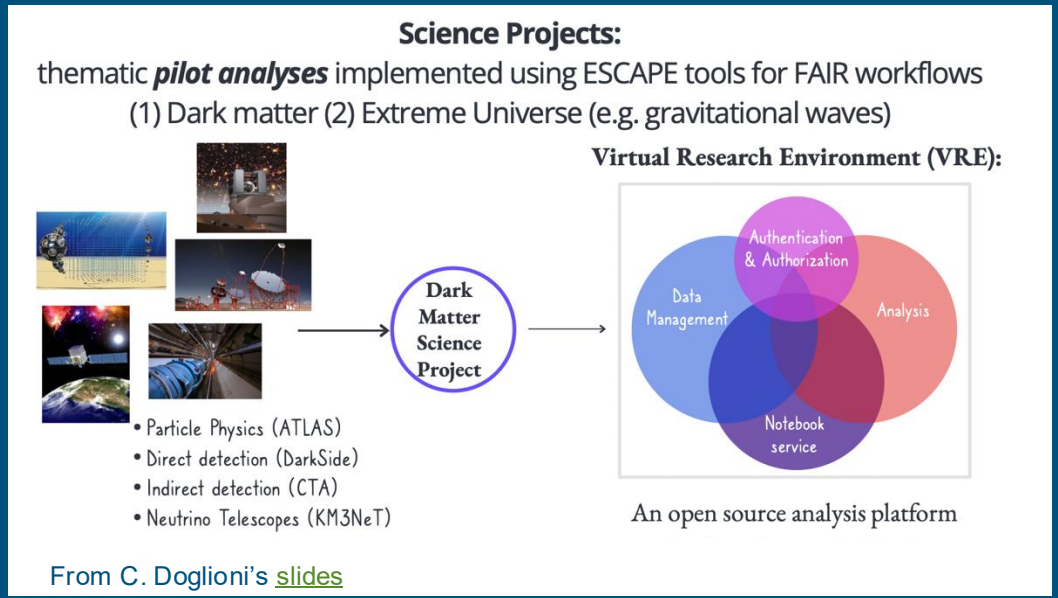
Experimental analyses are not always easily reproducible → not all information or workflow steps publicly available and/or difficult to provide without an experimental framework-like structure



[source]

European Open Science Cloud-FUTURE was part of ESCAPE [EU funded project aims to bring together different research infrastructures: KM3NET, Einstein telescope, LHC, DarkSide, etc] (more details in G. Guerrieri's talk)

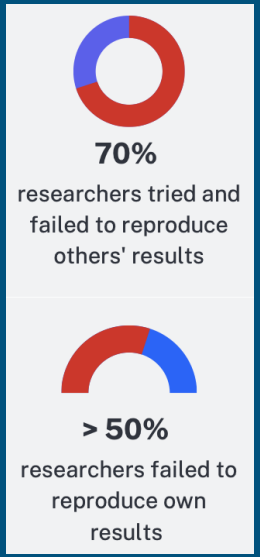
→ Attempts to have a coordinated effort to open science and enable FAIR Workflows.



From C. Doglioni's slides

The big picture...

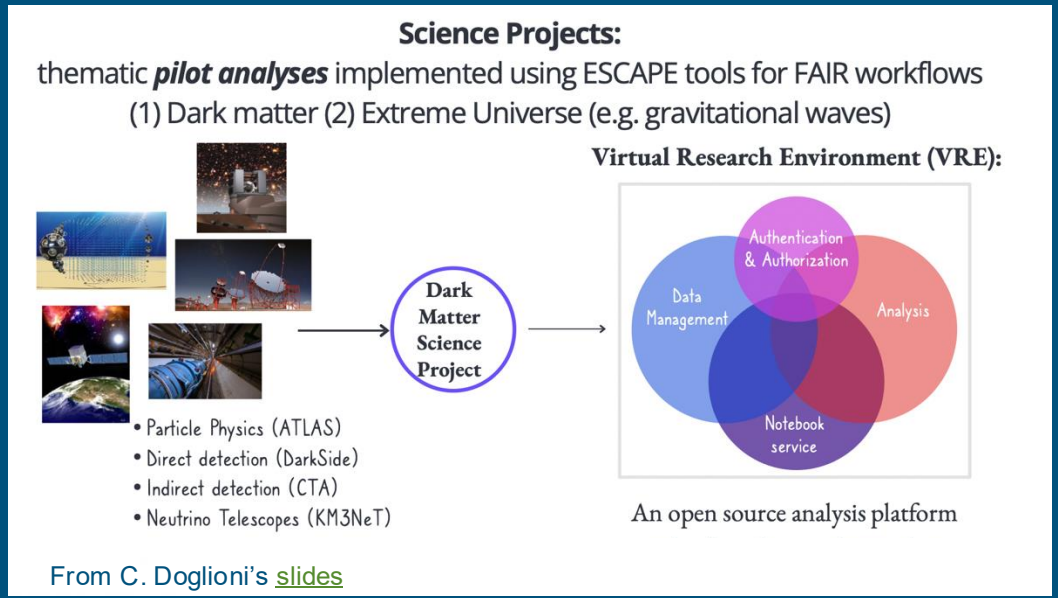
Experimental analyses are not always easily reproducible → not all information or workflow steps publicly available and/or difficult to provide without an experimental framework-like structure



[source]

European Open Science Cloud-FUTURE was part of ESCAPE [EU funded project aims to bring together different research infrastructures: KM3NET, Einstein telescope, LHC, DarkSide, etc] (more details in G. Guerrieri's talk)

→ Attempts to have a coordinated effort to open science and enable FAIR Workflows.



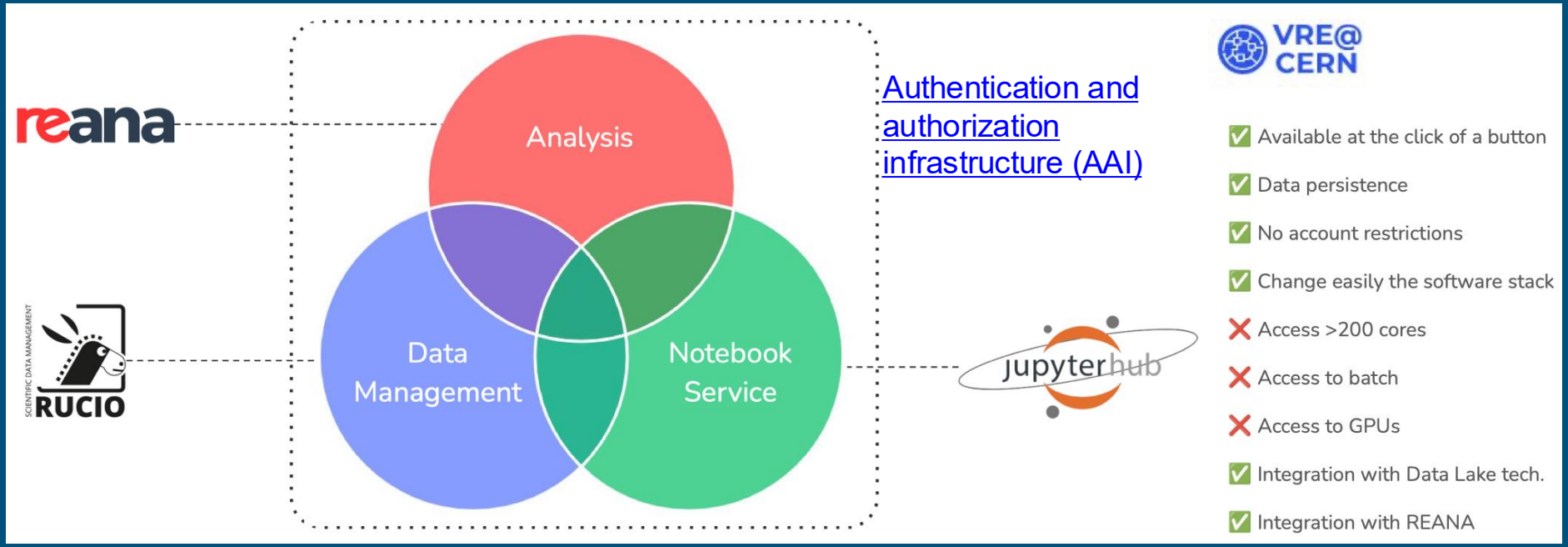
From C. Doglioni's slides

VRE@CERN

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

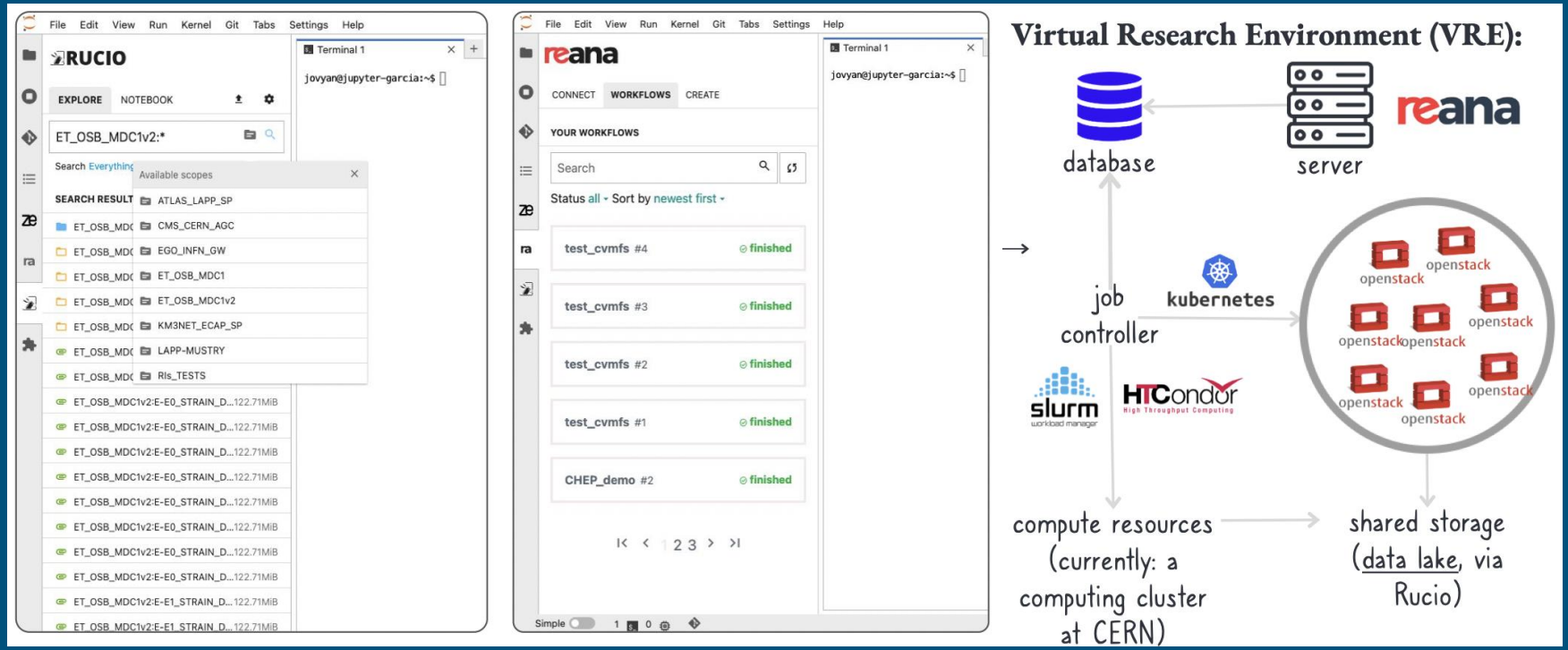
Open-source analysis platform developed @ CERN
→ access to all the digital content needed to develop, share and reproduce analyses using REANA, in compliance with FAIR principles.

Virtual Research Environment basics [for a user]



VRE uses AAI and shares different experimental data in a reliable distributed storage infrastructure via ESCAPE Data Lake.
Entry point (for an end-user) is a jupyterhub instance deployed on top of Kubernetes infrastructure. Analysis run on REANA-VRE instance.
VRE provides an interactive graphical UI. The data access and browsability enabled through API calls to high level data management and storage orchestration software (Rucio).

UI for Virtual Research Environment



How does it help experimental analysers and how could we use VRE to contribute to the global effort of analysis preservation?

Example case study:

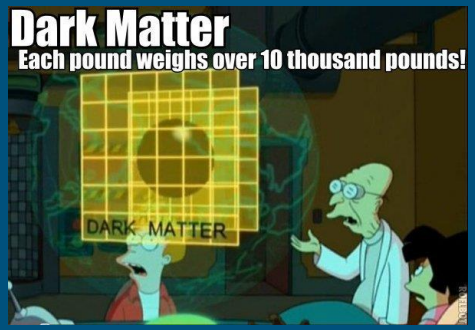
ATLAS dark QCD analysis implementation and reproducibility test

Hidden Sectors

SM Sector — Connectors / Portals — Hidden Sectors

Z', SUSY particles, Higgs, Extra Dim, Leptoquarks, CP-odd...

can be strongly or weakly coupled i.e., dark Higgs, dark photon, dark SU(N), Asym DM...



When a Hidden Sector particle is (quasi-)stable, a dark matter candidate can potentially exist

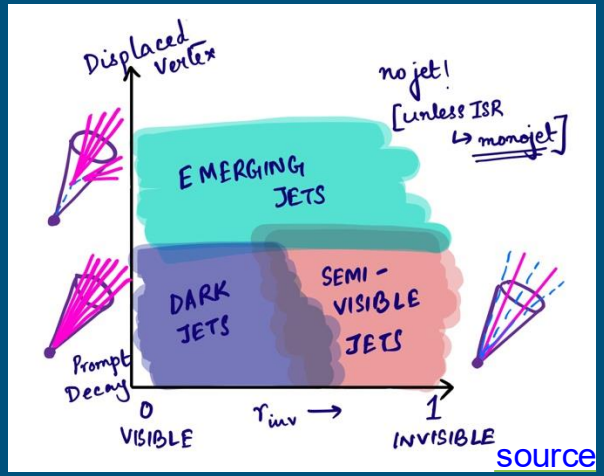
We have not found any concrete sign of new physics … yet!
Looking at unusual topologies and hidden corners of the phase space
→ **signature based searches, using benchmark models.**

The big picture...

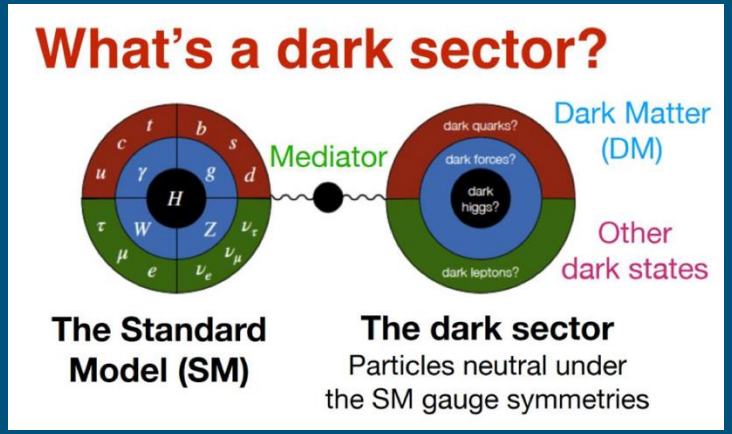
Dark hadrons decaying promptly in a QCD-like fashion,
- fully, or
- partially back to the visible sector

Dark hadrons undergoing displaced decays in a QCD-like fashion

Each signature has a distinctive feature:
-> Dark jets: unique substructure
-> Emerging jets: displaced objects
-> Semi-visible jets: substantial missing energy

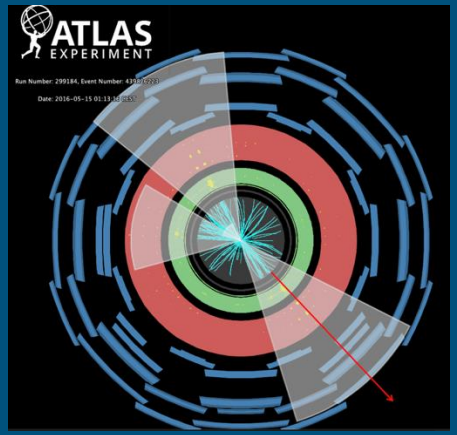


source

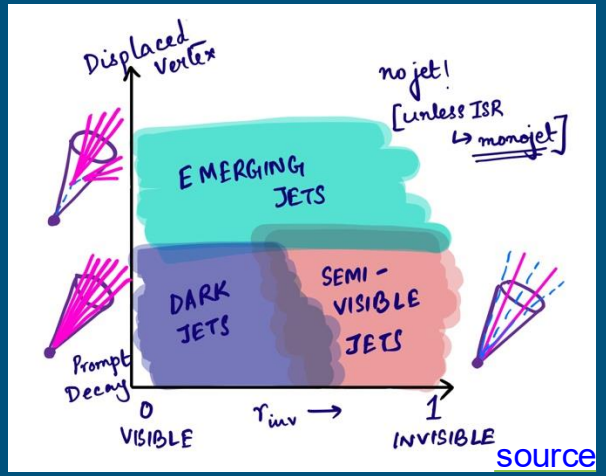


More details in N. Hemme's MA5 reinterpretation talk

Semi-visible jets

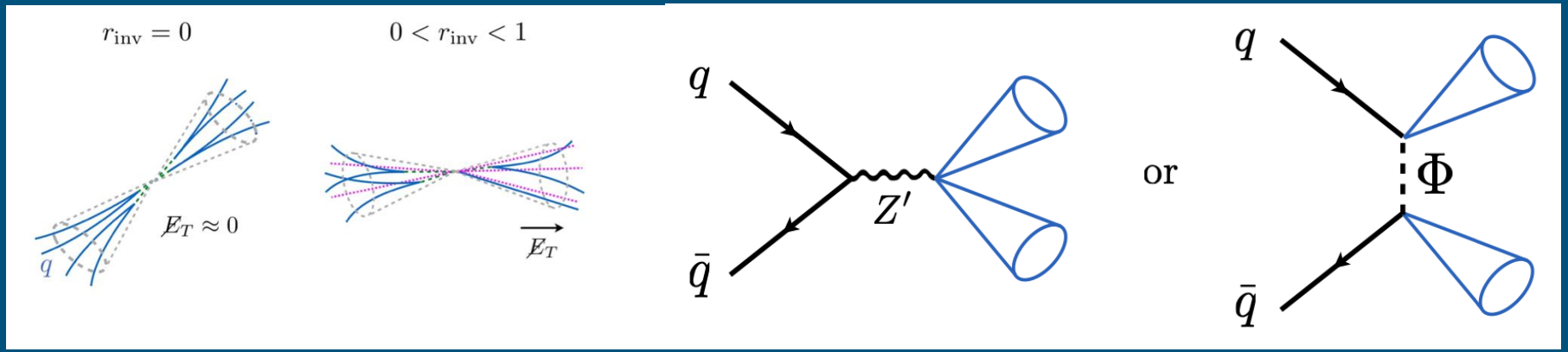


Assume dark sector mimicing QCD
→ Some dark hadrons decay back to SM
→ Some dark hadrons are stable and invisible



source

Jets with missing energy aligned along one of the jet direction!



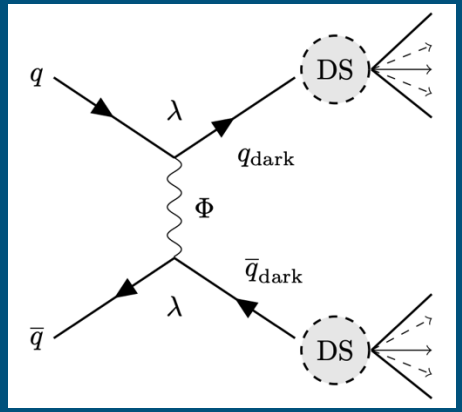
R_{inv} = #stable dark hadrons / #all dark hadrons

t-channel production mode

ATLAS Full Run-2 search for semi-visible jets (t-channel)

Signal model parameters:

1. M_ϕ = Mass of scalar bi-fundamental mediator
2. R_{inv}
3. M_d = Mass of dark hadrons

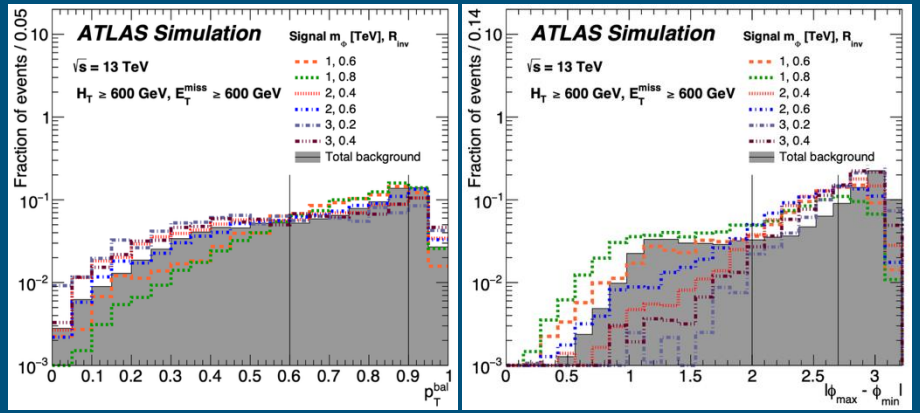


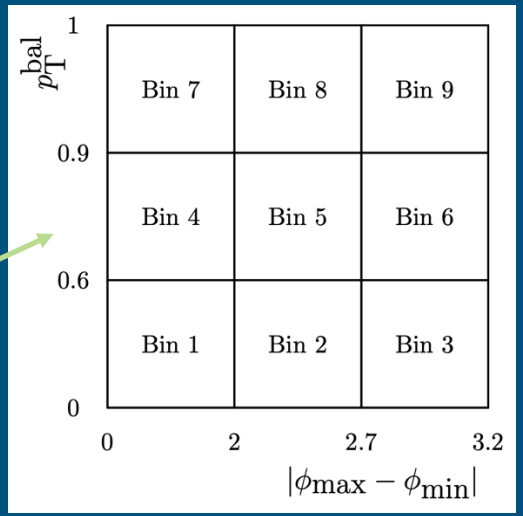
Background samples: W/Z+jets, ttbar, singletop, multi-jet, diboson

The signal events typically have high missing energy

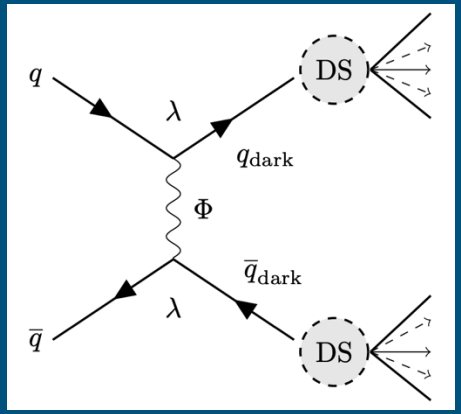
Partially data-driven method, simultaneously fit signal region and three control regions to obtain scale factors for each bkg process

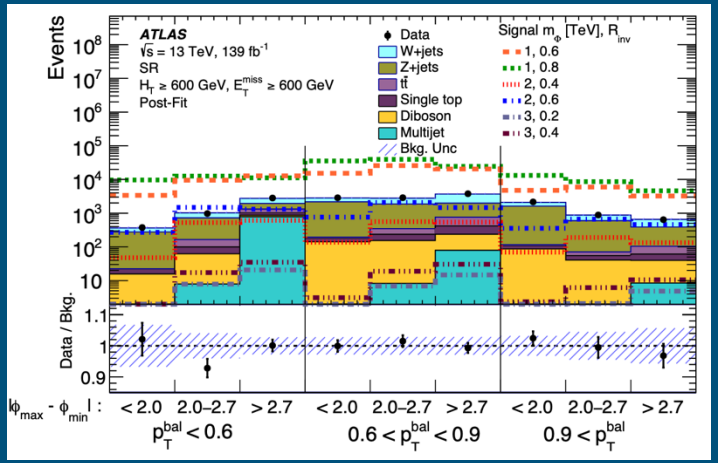
Two key observables used to design a 9-bin grid: Yield in each bin treated as an observable

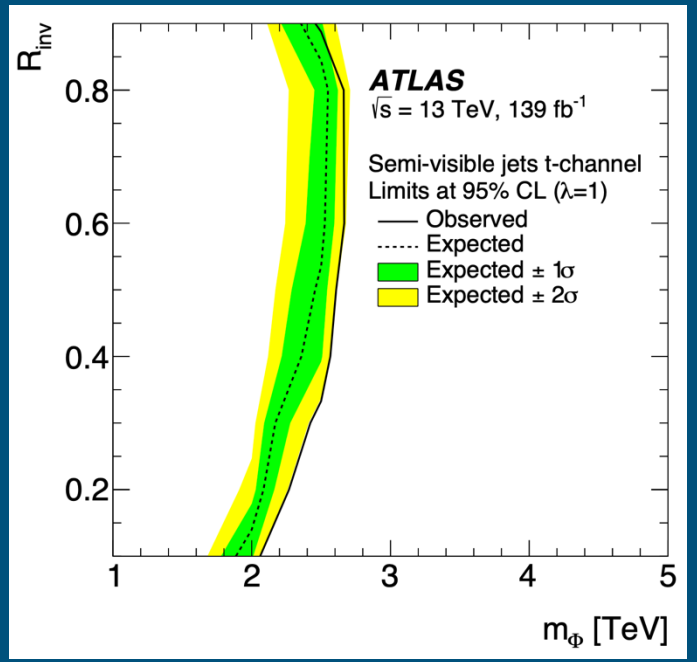




ATLAS Full Run-2 search for semi-visible jets (t-channel)







Phys. Lett. B 848 (2024) 138324

Search for non-resonant production of semi-visible jets using Run 2 data in ATLAS

The ATLAS Collaboration

Semi-visible jets, with a significant contribution to the event's missing transverse momentum, can arise in strongly interacting dark sectors. This results in an event topology where one of the jets can be aligned with the direction of the missing transverse momentum. A search for semi-visible jets produced via a *t*-channel mediator exchange is presented. The analysis uses pp collisions with an integrated luminosity of 139 fb^{-1} and a centre-of-mass energy of 13 TeV, collected with the ATLAS detector during Run 2 of the LHC. No excess over Standard Model predictions is observed. Assuming a coupling strength of unity between the mediator, a Standard Model quark and a dark quark, mediator masses up to 2.7 TeV can be excluded at the 95% confidence level. Upper limits on the coupling strength are also derived.

First limits on t-channel production mode from ATLAS!

Potential for preservation

Semi-visible jet t-channel search has fairly inclusive selections (so does several other experimental searches)
→ ALLOWS TO PROBE OTHER MODELS WITH SIMILAR FINAL STATES

BUT several times it becomes difficult for even the ATLAS analysers to reproduce the analysis results after a few months → either code not publicly available or previous analyser left the field…

Code preservation of experimental analyses is essential!

Science 2061

- Many members of the generation who may wish to re-analyze the LHC Run 3 data in 2061 have yet to be born.
- That generation may wish to re-analyze these data simultaneously with old LSST and LIGO data using BSM models that they invent from which (with the help of their AI assistants) they are able to compute testable predictions for LHC, LSST, and LIGO data.
- Now ask yourself: What are we doing wrong, or what are we forgetting, that risks thwarting such heroic efforts forty year hence?

Harrison Prosper, RiF 2021

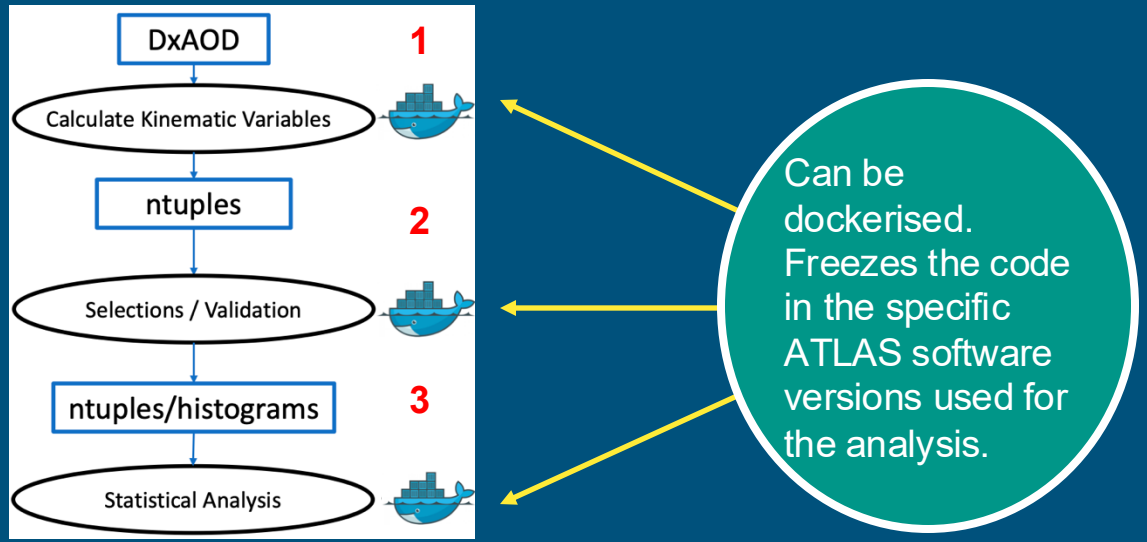
Make analyses live beyond the paper publication !

From S. Kraml's talk

Workflow for an analysis implementation and usage within VRE

An ATLAS experimental search (cut & count) can be split into 2-3 different/distinct components, depending on the tools involved in the workflow (starting from reconstructed object level):

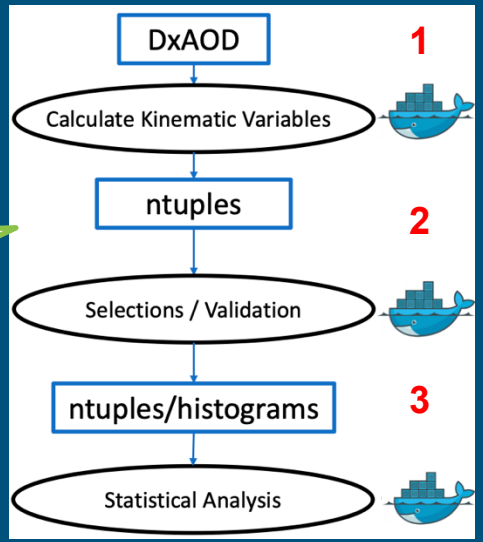
1. Object and trigger selections
2. Ntuple analyser
3. Statistical analysis



Workflow for an analysis implementation and usage within VRE

An experimental search (cut & count) can be split into 2-3 different/distinct components, depending on the tools involved in the workflow (starting from reconstructed object level):

1. Object and trigger selections
2. Ntuple analyser
3. Statistical analysis



This part of workflow is not new! This is essentially what ATLAS internal recast has been doing for several years.

Can be dockerised. Freezes the code in the specific ATLAS software versions used for the analysis.

Running semi-visible jet analysis on VRE

Steps TLDR:-

- Follow the authentication and authorisation infrastructure (AAI) setup instructions on https://vre-hub.github.io/,
- Clone the REANA implementation of semi-visible jets t-channel analysis: GitHub link to repository,
- source runReana.sh within the VRE jupyter hub instance #script containing all the REANA workflow setup and running instructions

Running semi-visible jet analysis on VRE

Steps TLDR:-

- Follow the authentication and authorisation infrastructure (AAI) setup instructions on https://vre-hub.github.io/,
- Clone the REANA implementation of semi-visible jets t-channel analysis: GitHub link to repository,
- source runReana.sh within the VRE jupyter hub instance #script containing all the REANA workflow setup and running instructions

Non-TLDR version:-

The REANA analysis workflow is split into 3 steps (== 3 docker images, that run one after the other in the REANA-VRE instance):

1. Object and trigger selections using AnalysisTop package (part of ATLAS AnalysisBase public docker image)

2. Flat ntuple analyser (this part is up to the collaboration to make public → might require code approval, subject to other collaboration software dependencies)

3. Fitting mechanism via TRExFitter package (part of StatAnalysis public docker image)

The workflow is run within the reana-vre(.cern.ch) platform

svj-tchannelreana / specs / workflow.yml

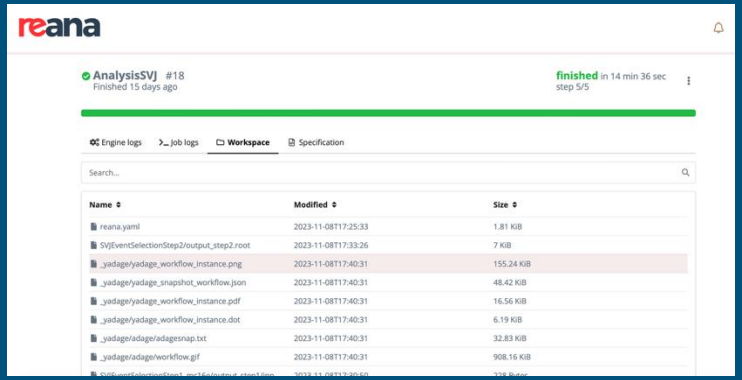
susinha svj reana scripts and files

Code | Blame | Executable File · 74 lines (67 loc) · 2.81 KB | Code 55%

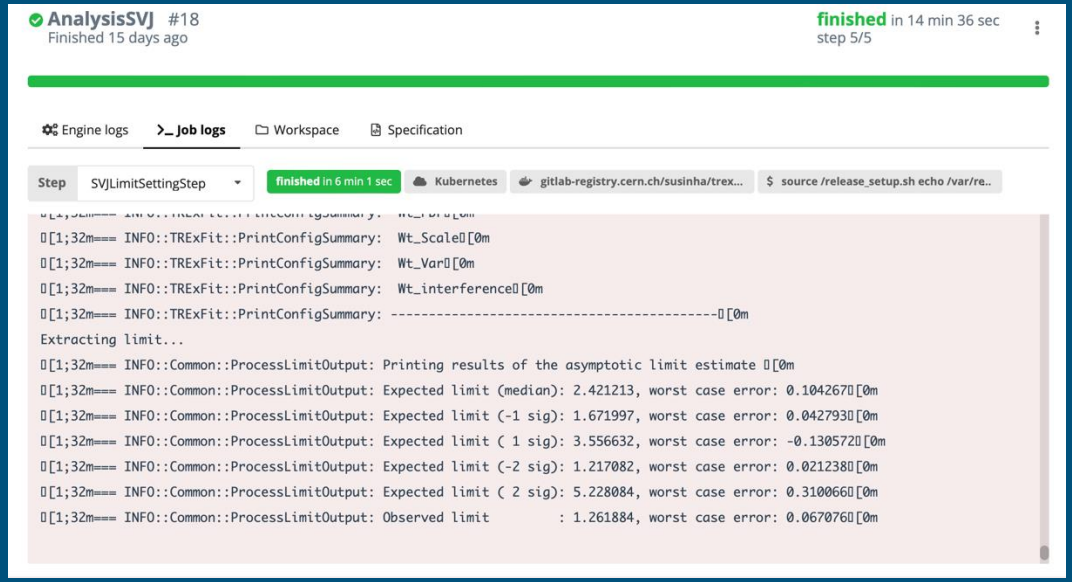
```
stages:
- name: SVJEventSelectionStep1_mc16a
  dependencies: [init]
  scheduler:
    scheduler_type: singlestep-stage
    parameters:
      AT_config_file: 'config_files/config_mc16a.txt'
      dxaod_file_mc16a: {step: init, output: dxaod_file_mc16a}
      prw_file_mc16a: {step: init, output: prw_file_mc16a}
      tdp_file_path: {step: init, output: tdp_file_path}
      output_step1: '{workdir}/output_step1'
    step: {$ref: 'specs/steps.yml#/SVJEventSelectionStep1_mc16a'}

- name: SVJEventSelectionStep1_mc16d
  dependencies: [init]
  scheduler:
    scheduler_type: singlestep-stage
    parameters:
      AT_config_file: 'config_files/config_mc16d.txt'
      dxaod_file_mc16d: {step: init, output: dxaod_file_mc16d}
      prw_file_mc16d: {step: init, output: prw_file_mc16d}
      tdp_file_path: {step: init, output: tdp_file_path}
      output_step1: '{workdir}/output_step1'
    step: {$ref: 'specs/steps.yml#/SVJEventSelectionStep1_mc16d'}

- name: SVJEventSelectionStep1_mc16e
  dependencies: [init]
```

Implementation similar to ATLAS Z' analysis also included in VRE [talk by J. Little.]

```
(reana) [susinha@lxplus788 reana]$ source runReana.sh
==> Verifying REANA specification file... /afs/cern.ch/work/s/susinha/reana/reana.yaml
  -> SUCCESS: Valid REANA specification file.
==> Verifying REANA specification parameters...
  -> SUCCESS: REANA specification parameters appear valid.
==> Verifying workflow parameters and commands...
==> SUCCESS: File /reana.yaml was successfully uploaded.
==> SUCCESS: File /specs/steps.yml~ was successfully uploaded.
==> SUCCESS: File /specs/steps.yml was successfully uploaded.
==> SUCCESS: File /specs/workflow.yml was successfully uploaded.
==> SUCCESS: File /specs/workflow.yml~ was successfully uploaded.
==> SUCCESS: AnalysisSVJ has been queued
```



Files uploaded on rucio-vre and can be accessed via same authentication. Run successful and sensible limit extraction using reana-vre instance



Elephant in the room – piping new physics models to existing REANA analyses

The REANA + VRE workflow currently enables users to run the present analyses with existing samples and datasets [raw analysis code preservation in the broadest sense of the term]

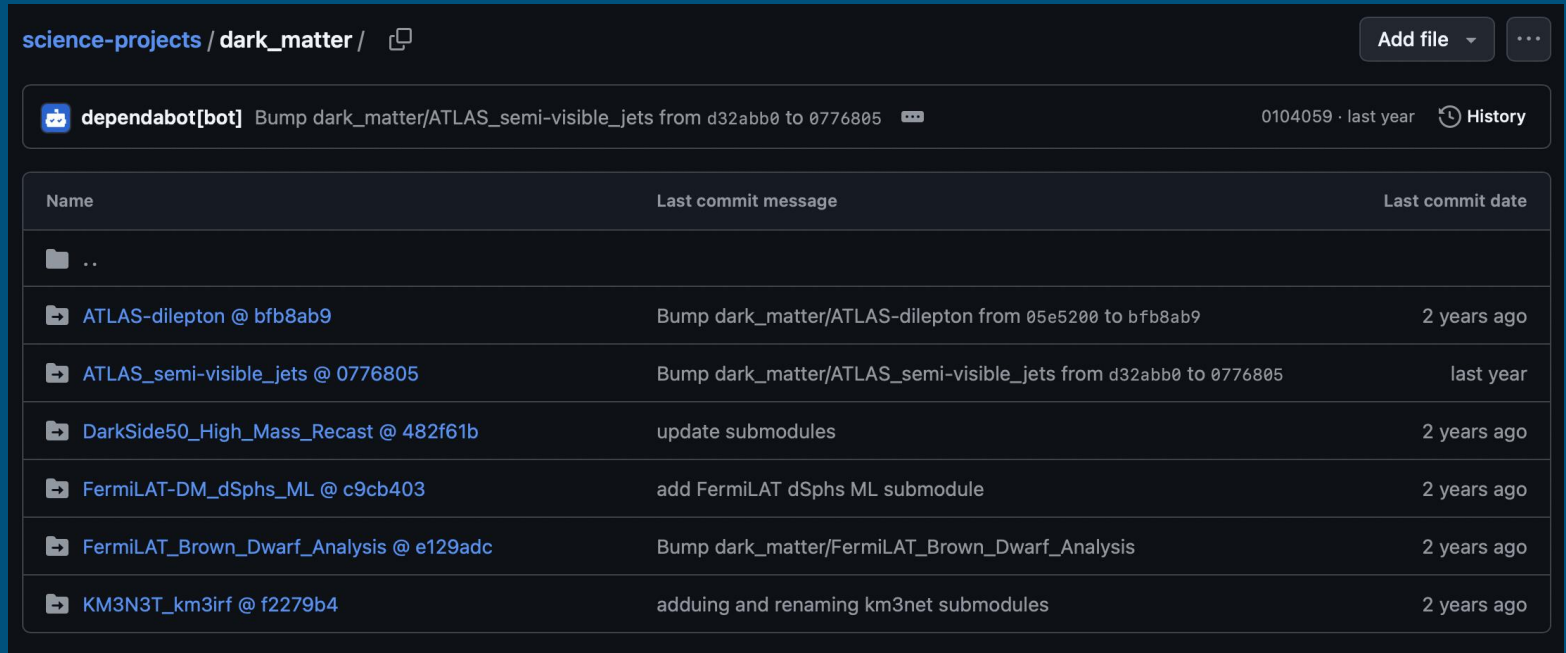
→ How to go to the next step of reusability and reinterpretation of the analyses for other new physics signals?

* requires the availability of full ATLAS simulation software (Event Generation -> detector simulation -> reconstructed objects) in a dockerised format → difficult, requires sufficient motivation, collaboration management level discussions needed

Elephant in the room – piping new physics models to existing REANA analyses

→ What can be done in the short-term?

* Add more ATLAS analyses to the existing platform, make enough of a physics and reinterpretation motivated case for collaborations to see the value of the effort (collaboration-internal discussion initiated).



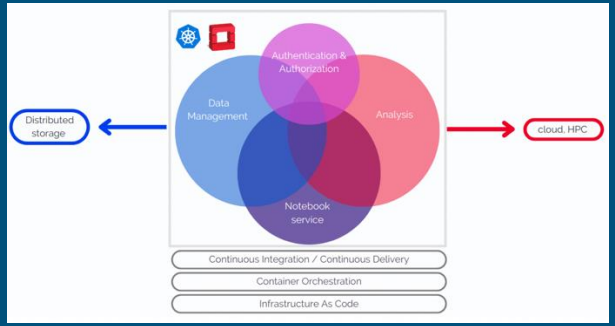
User perspective of the Analysis Facility goals and VRE

Analysis facility: infrastructure and services that provide integrated data, software and computational resources to execute one or more elements of an analysis workflow. (Whitepaper)

-- Perform fast analysis iterations on large datasets interactively.
-- Convert interactive to batch-schedulable workloads.
-- Interact with the WLCG and scale outside of the facility on occasion.

-- Collaborate in a multi-organisational team.
-- Execute desired analysis workflows reproducibly.

-- Efficiently train machine learning models for HEP.

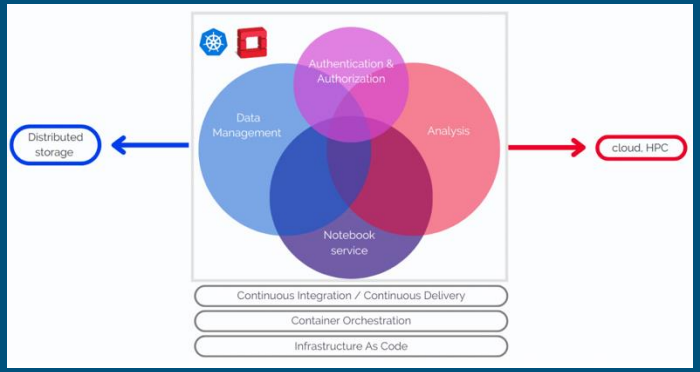


Looking forward...

- Analysis code preservation of experimental analyses is essential → it's a useful starting point for reinterpretations.

- Analyser perspective: VRE gives a platform to experimental/non-experimental physicists to dig through the official analysis code (currently available in the database) from different experiments and reproduce the workflow, **independent from local storage**

- Discussions ongoing for having code approval within ATLAS before storing in VRE. Piping new signals through the whole ATLAS sample processing step (Event Generation -> detector simulation -> reconstructed objects) still needs to be figured out



PRESERVATION and REINTERPRETABILITY!