REANA reproducible analyses: status update

M. Donadoni, T. Šimko

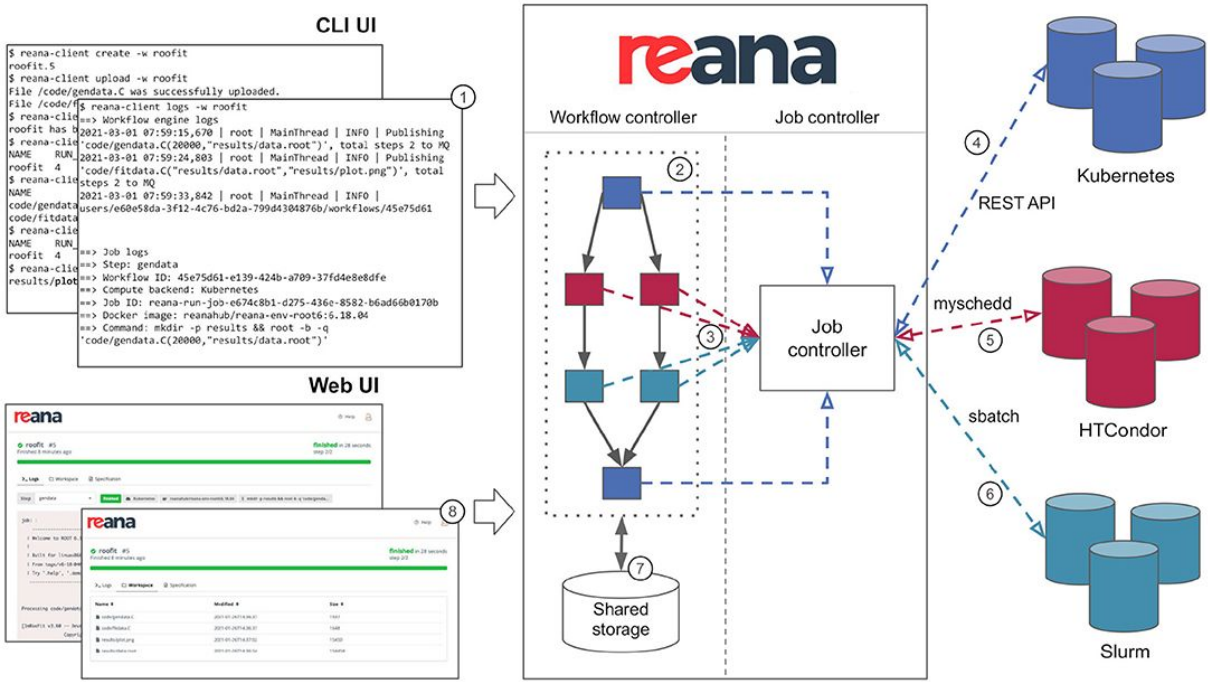
CERN

4th DPHEP Collaboration Workshop, October 2nd-3rd 2024

https://indico.cern.ch/event/1432766/

What is REANA?

Running containerised analysis workflows on the cloud



Multiple **compute backends**:

- Kubernetes
- HTCondor
- Slurm

Multiple **workflow languages**:

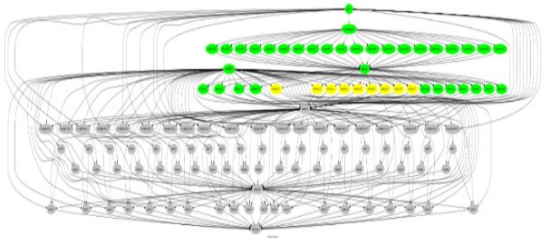
- CWL
- Serial
- Snakemake
- Yadage

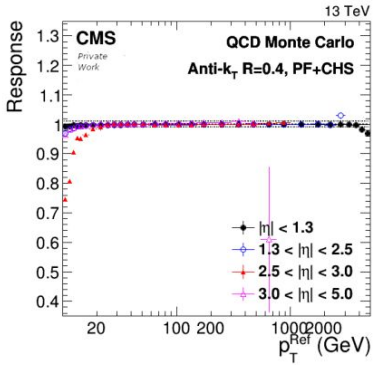
Multiple **means of use**:

- Command-line client
- Web UI

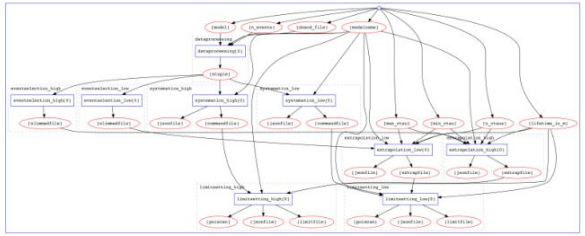
https://www.reana.io

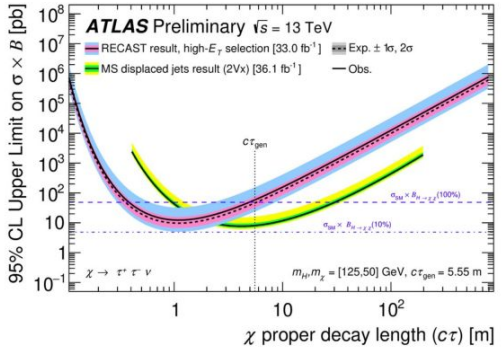
Use cases: data production and data analyses





Data production example: CMS jet energy resolutions and corrections
https://github.com/alintulu/reana-demo-JetMETAnalysis





Data analysis example: ATLAS displaced jet reinterpretations
https://cds.cern.ch/record/2714064

ATLAS pMSSM searches

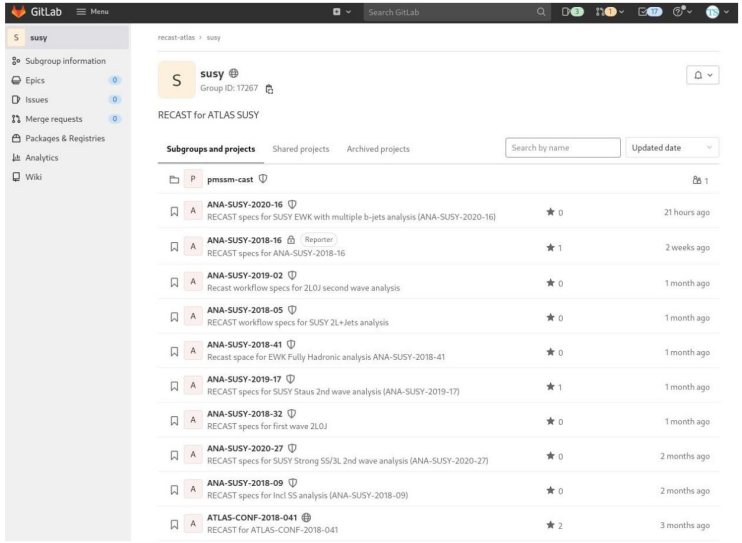


Figure 1. A screenshot of the ATLAS SUSY group analyses preserved on GitLab. Each repository is labeled with the internal ATLAS analysis identifier and contains both workflow files and additional data files needed for the computational processing.

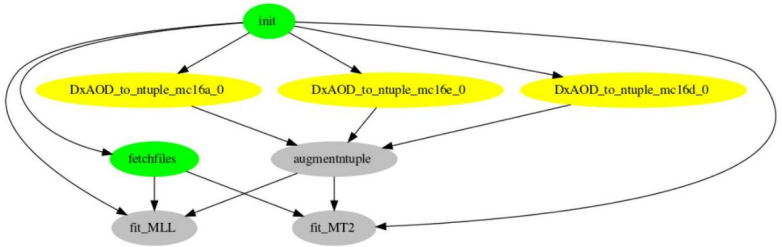


Figure 2. A typical pMSSM workflow. The computational runtime is about 10 minutes without systematics (test payload) and about 10 hours with all systematics (real payload).

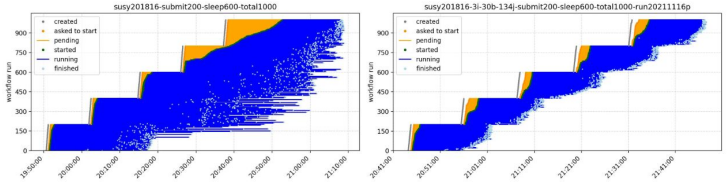
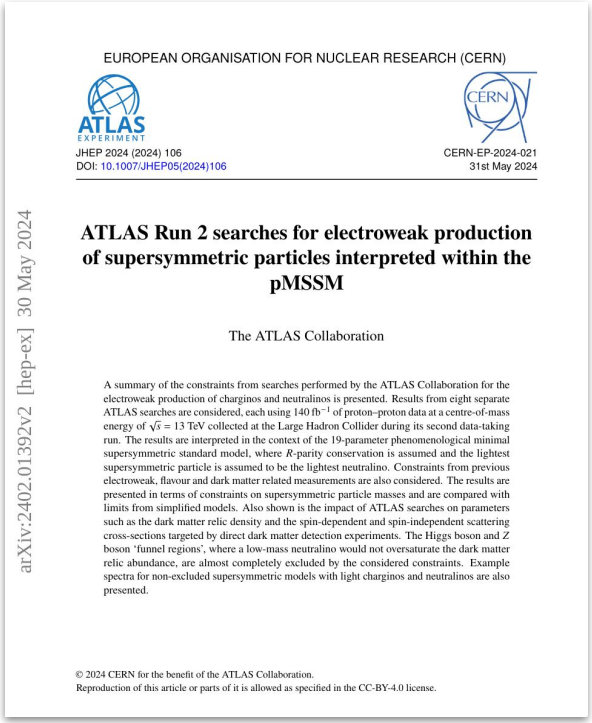


Figure 8. A scalability test submitting 200 workflows every 10 minutes. A cluster with 448 cores (left) cannot keep up with the load. A cluster with 1072 cores (right) can comfortably hold the incoming workload.

https://arxiv.org/abs/2403.03494

Streamlining the execution of thousands of reinterpretation workflows at scale

First ATLAS pMSSM Run-2 searches published



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

ATLAS EXPERIMENT

CERN

JHEP 2024 (2024) 106
DOI: 10.1007/JHEP05(2024)106

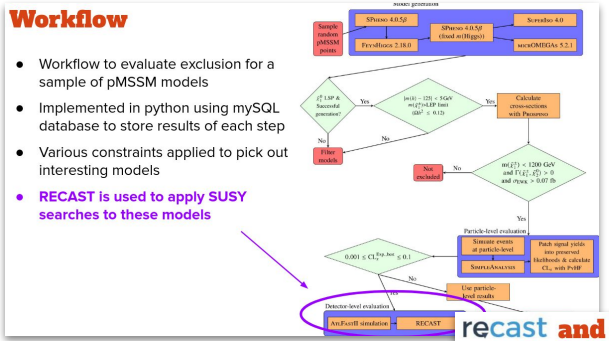
CERN-EP-2024-021
31st May 2024

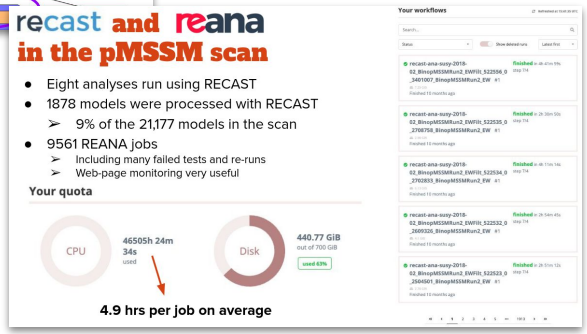
ATLAS Run 2 searches for electroweak production of supersymmetric particles interpreted within the pMSSM

The ATLAS Collaboration

A summary of the constraints from searches performed by the ATLAS Collaboration for the electroweak production of charginos and neutralinos is presented. Results from eight separate ATLAS searches are considered, each using 140 fb^{-1} of proton–proton data at a centre-of-mass energy of $\sqrt{s} = 13$ TeV collected at the Large Hadron Collider during its second data-taking run. The results are interpreted in the context of the 19-parameter phenomenological minimal supersymmetric standard model, where R-parity conservation is assumed and the lightest supersymmetric particle is assumed to be the lightest neutralino. Constraints from previous electroweak, flavour and dark matter related measurements are also considered. The results are presented in terms of constraints on supersymmetric particle masses and are compared with limits from simplified models. Also shown is the impact of ATLAS searches on parameters such as the dark matter relic density and the spin-dependent and spin-independent scattering cross-sections targeted by direct dark matter detection experiments. The Higgs boson and Z boson 'funnel regions', where a low-mass neutralino would not oversaturate the dark matter relic abundance, are almost completely excluded by the considered constraints. Example spectra for non-excluded supersymmetric models with light charginos and neutralinos are also presented.

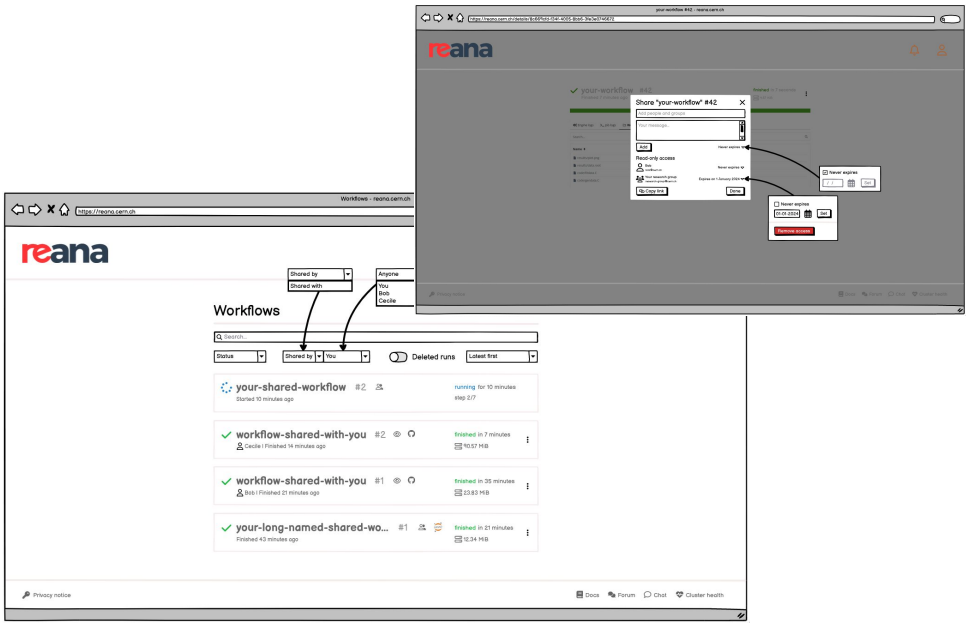
https://arxiv.org/abs/2402.01392



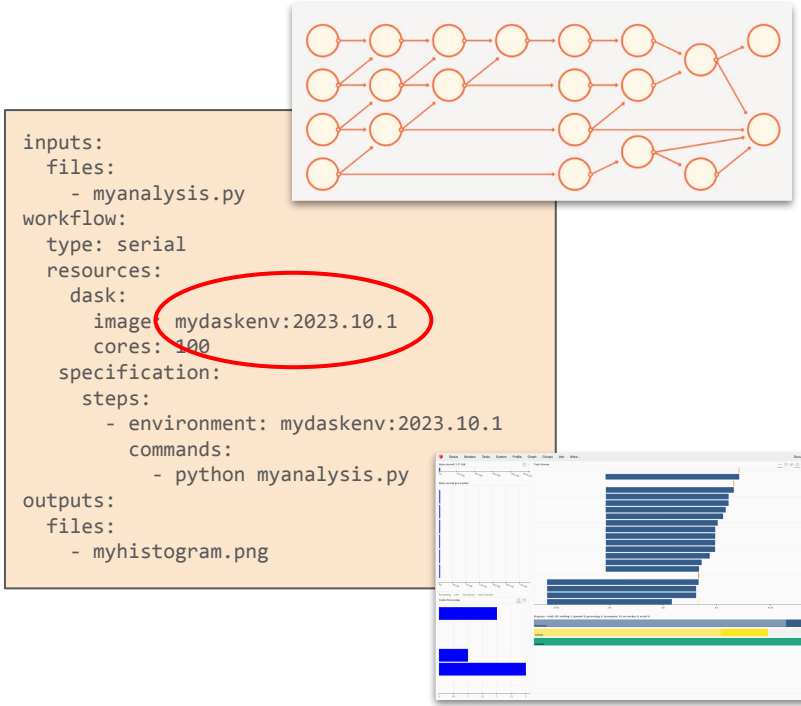


Ben Hodkinson
https://indico.cern.ch/event/1380367/

A glimpse on forthcoming REANA features

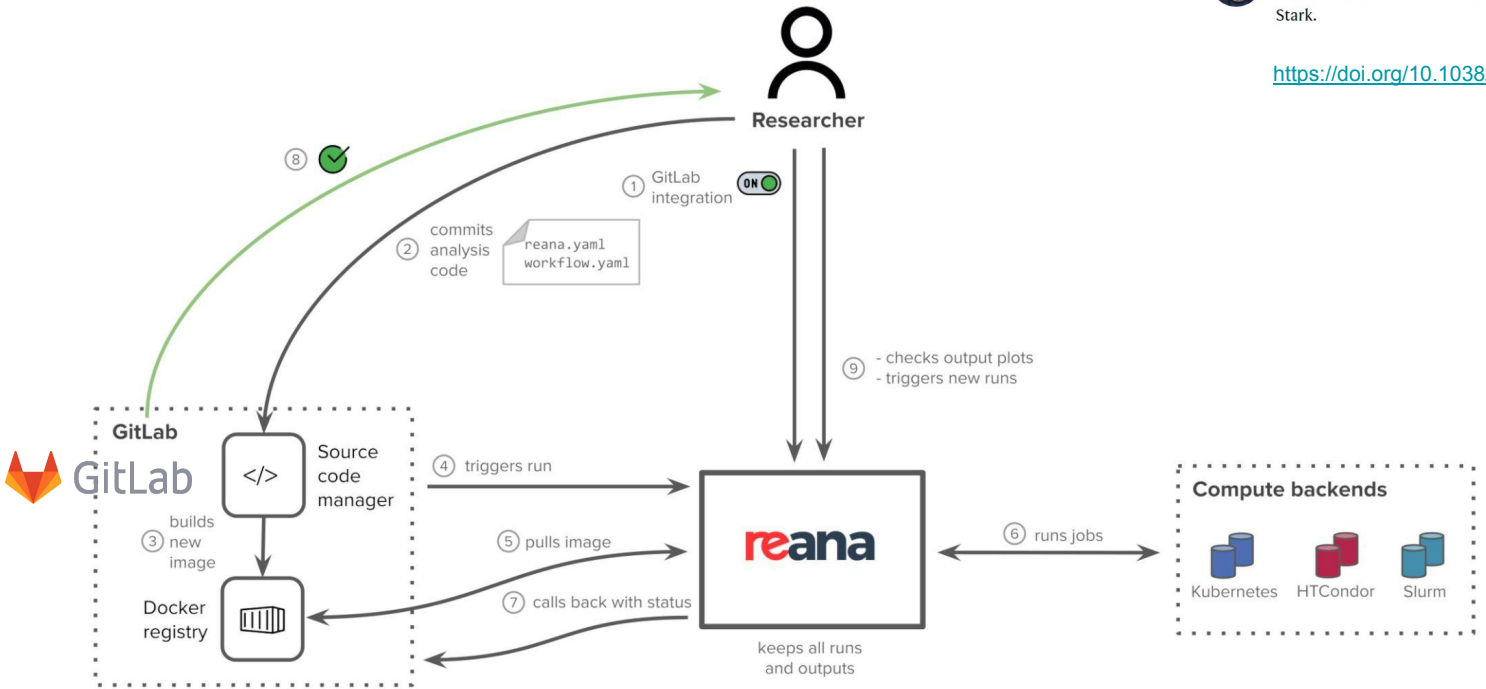


Share workflows with colleagues



Support for Dask workflows

Driving future reproducibility



WORLD VIEW · 24 MAY 2018

Before reproducibility must come preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says Philip Stark.

https://doi.org/10.1038/d41586-018-05256-0

REANA as a continuous integration engine for source code management systems

Community

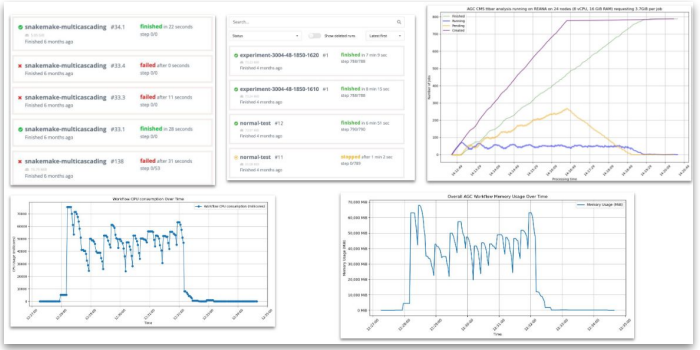
IRIS-HEP HSF Analysis preservation training (Valencia, October 2023)
https://hsf-training.github.io/hsf-training-reana-webpage/

HSF trainings on reproducibility

Workshop on workflow languages for HEP (May 2024)
https://indico.cern.ch/event/1380367/

Seeking synergies across experiments

HSF Analysis Facilities White Paper



THE HEP SOFTWARE FOUNDATION (HSF)

HSF-TN-2024-01
April 2024

Analysis Facilities White Paper

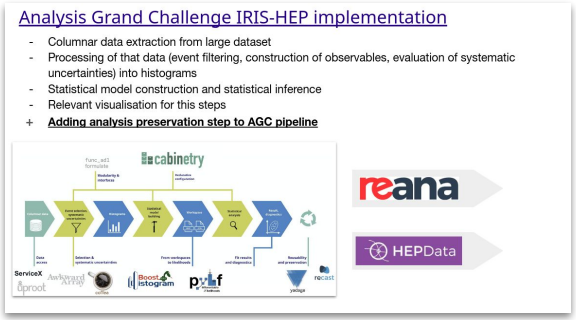
arXiv:2404.02100v2 [hep-ex] 15 Apr 2024

To support complicated workflow pipelines, as expressed in the user requirements, workflow languages such as Snakemake [22], Yadage [23], Common Workflow Language (CWL) [24] and others may be suitable. Also, some workload schedulers have features that allow users to express dependencies between workloads (e.g. HTCondor's tool DAGMan). Workflow services such as REANA [25], capable of processing such languages, should be integrated into the existing batch infrastructure at the facility in order to have access to the same data and compute resources that would be available if the pipeline were processed manually. Note that REANA also works with a kubernetes back end.

The large number of new technologies may trigger the question: should users be exposed to all of these or should there be a common entry point interface? Reaching agreement on a common interface will be challenging, the introduction of Jupyter Notebooks/JupyterLab is an attempt at having a common interface but this remains an open question.

Keywords: High energy physics, analysis facilities, data analysis, scientific computing, data access, grid computing, federated identity management, analysis preservation, resource provisioning, HL-LHC

HSF
HEP Software Foundation



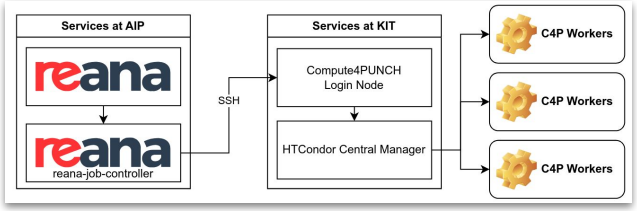
https://indico.cern.ch/event/1446410/

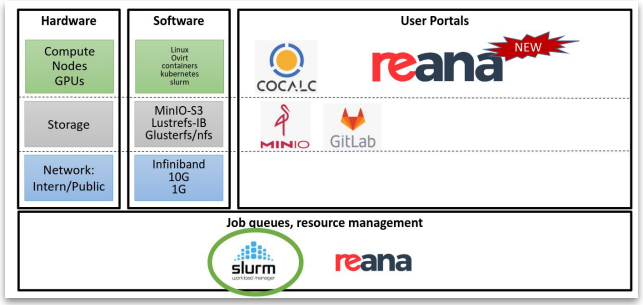
iris hep

https://arxiv.org/abs/2404.02100

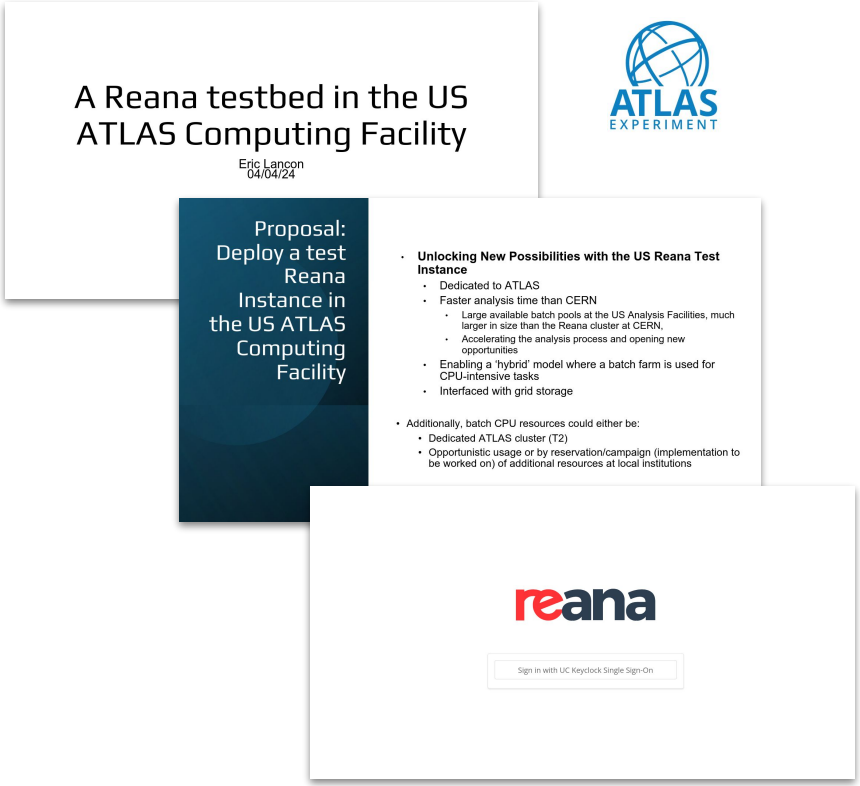
Analysis Grand Challenge CMS ttbar analysis on REANA
(see forthcoming CHEP 2024 talk)

Recent new deployments





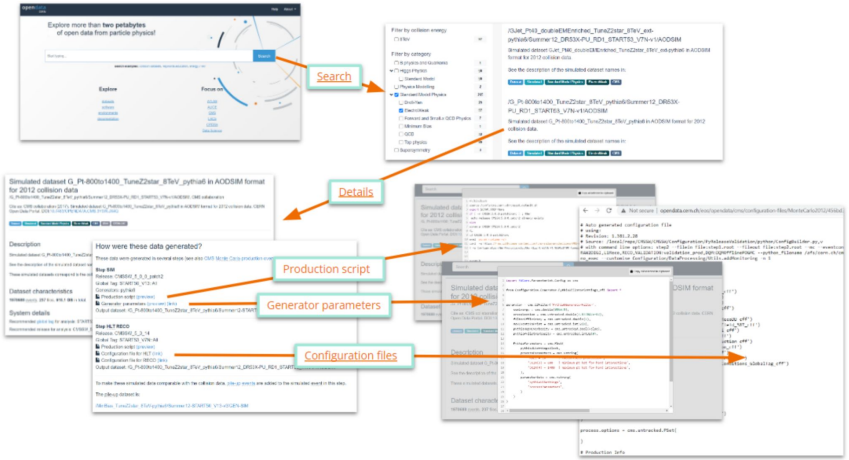
REANA @ AIP (astronomy)
https://reana-p4n.aip.de/
https://indico.desy.de/event/44722/

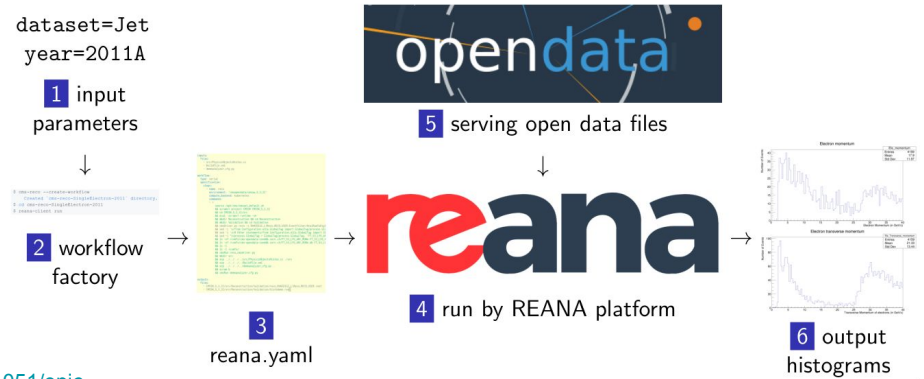


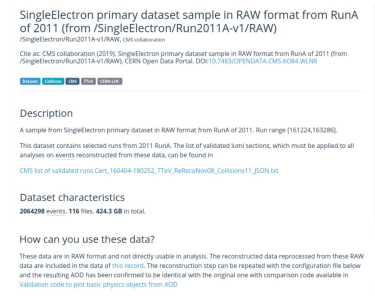
REANA@UChicago (ATLAS Analysis Facility)
https://reana.af.uchicago.edu/
https://indico.cern.ch/event/1386696/

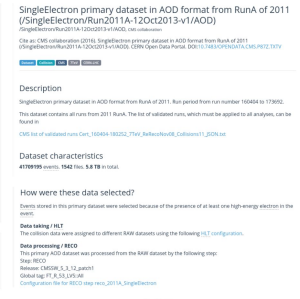
Use case 1: Is preserved data correct?

Verifying data provenance









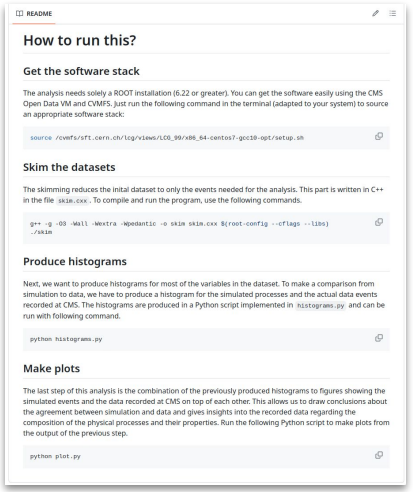
CMS open data coming with detailed provenance information

https://doi.org/10.1051/epjconf/202024508014

Reprocessing AOD from RAW samples

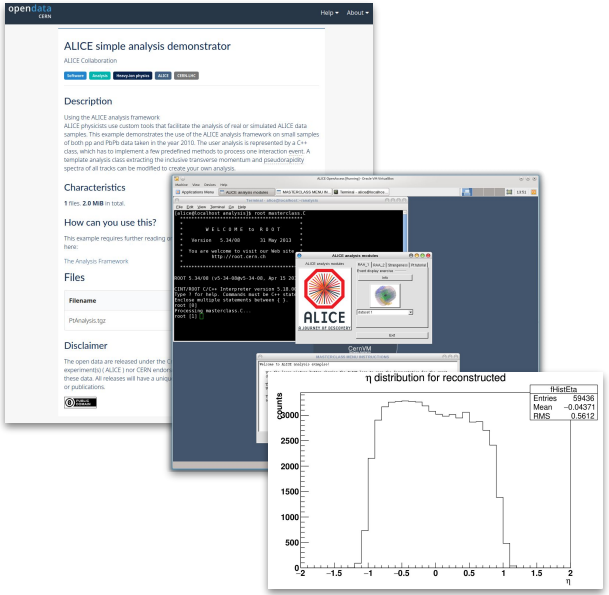
Use case 2: Is preserved data usable?

Verifying usage patterns



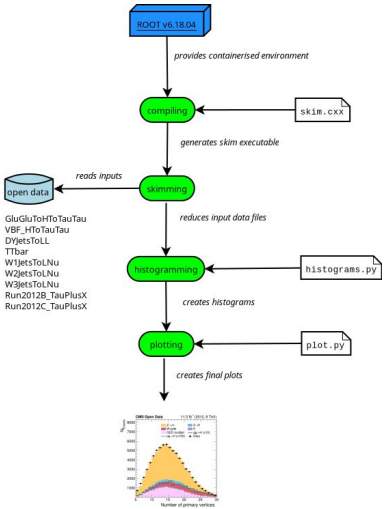
How-to-run recipes in README files is a good start; but they are not actionable

“Out of 863,878 attempted executions of valid notebooks (...) only 24.11% executed without errors and only 4.03% produced the same results”



ALICE pt analysis example in the VM stopped working due to microCernVM format compatibility issues

“Continuous reuse”

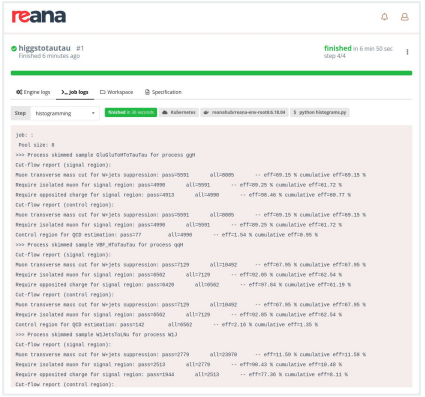


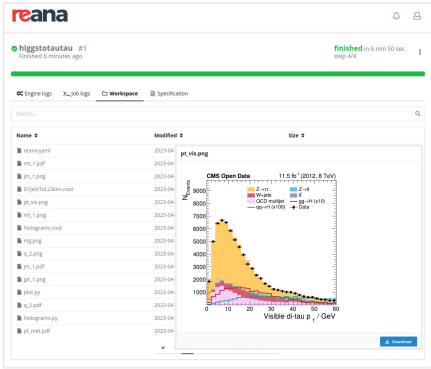
An example studying
H → ττ lepton decays
uses nine published
CMS open datasets

```
Feature: cms-htautau-nanoaod

  Scenario: Workspace content
    When the workflow is finished
    Then the workspace should contain "njets.png"
    And the workspace should contain "phi_1.png"

  Scenario: Workspace size
    When the workflow is finished
    Then the workspace size should be less than 75 MiB

  Scenario: Log content
    When the workflow is finished
    Then the job logs of the step "skimming" should contain \
        "Event has good muons: pass=36921"
    And the job logs of the step "histogramming" should contain \
       "Muon transverse mass cut for W+jets suppression: pass=5063"

  Scenario: Run duration
    When the workflow is finished
    Then the workflow run duration should be less than 25 minutes
    And the duration of the step "skimming" should be less than 20 minutes
```

Define expected outcomes in natural language thanks to Gherkin behavioural test language

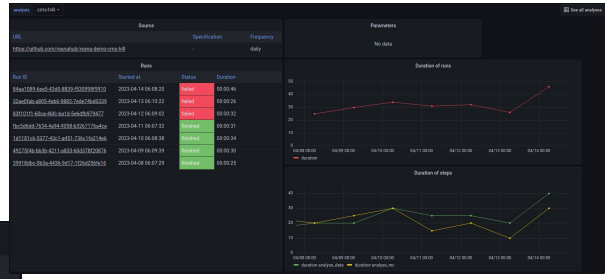
- check workspace content and output files
- check produced log messages
- check execution duration of steps





Test outcomes periodically on REANA

Making sure data remains usable



Analyses

Name ↑	Last success	Last failure	Last duration	R1	R2	R3	R4	R5
alice-lego-train-test-run	5 hours ago		00:01:35					
alice-pt-analysis	5 hours ago		00:01:10					
atlas-recast	5 hours ago		00:01:10					
cms-dimuon-mass-spectrum	8 days ago		00:01:03					
cms-dimuon-spectrum	5 hours ago		00:01:22					
cms-dimuon-spectrum-nanoaod	5 hours ago		00:01:52					
cms-h4l	5 hours ago	2 days ago	00:02:02					
cms-h4l-nanoaod	5 hours ago		00:03:46					
cms-htautau-nanoaod	5 hours ago		00:07:08					

- displays a history of various reuse examples and their statistics
- allows to quickly check the last success and failure timestamps
- shows the results of last five runs
- displays duration of individual steps

Dashboard monitoring continuous reuse of periodically re-executed open data analysis examples

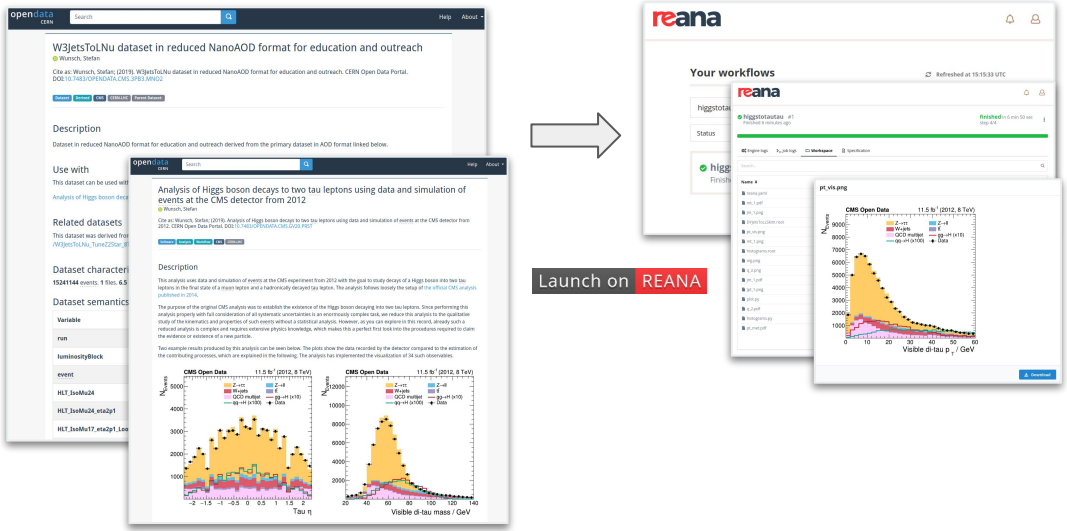
Conclusions

REANA as an “analysis engine” complementing your data preservation repository activities.

Ultimate goal: facilitate future reuse of scientific data.

- Use cases for “preproducible” analyses
- Use cases for data provenance verification
- Use cases for analysis reinterpretations
- Use cases for data usage pattern validations

"adaptable software examples [are] the most efficient way to pass on the knowledge needed for research-level studies on [the] data" — CMS



Data + Code + Environment + Services + Workflow = Reusable Analyses