

REANA reproducible analyses: status update

M. Donadoni, T. Šimko

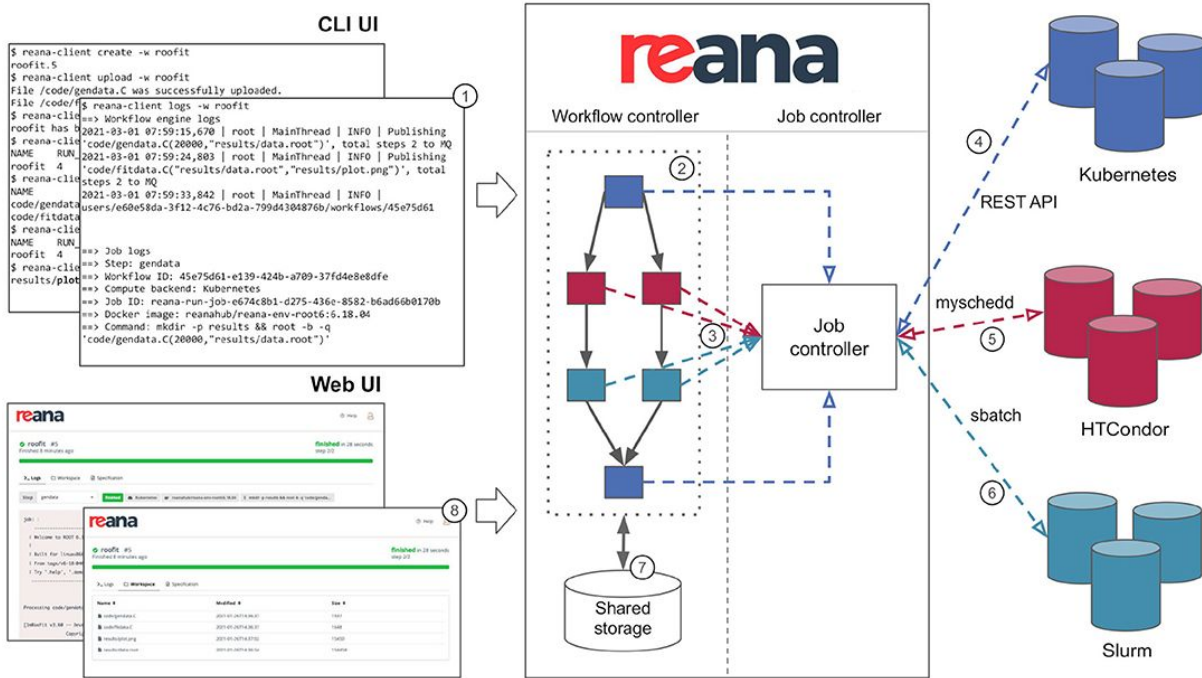
CERN

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**:

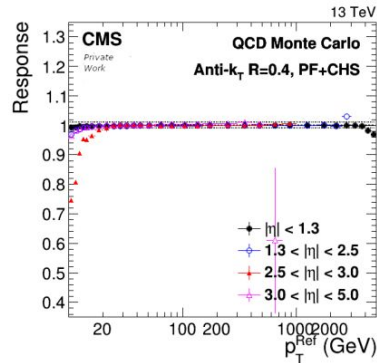
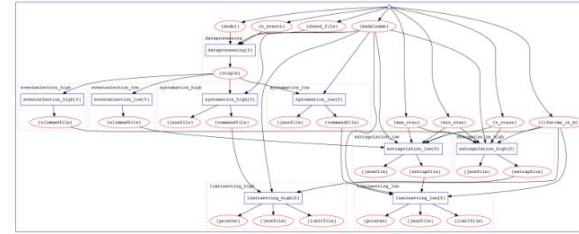
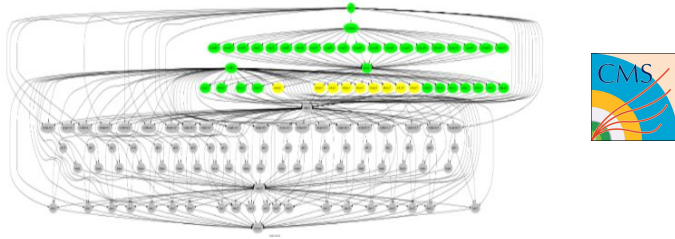
- CWL
- Serial
- Snakemake
- Yadage

Multiple **means of use**:

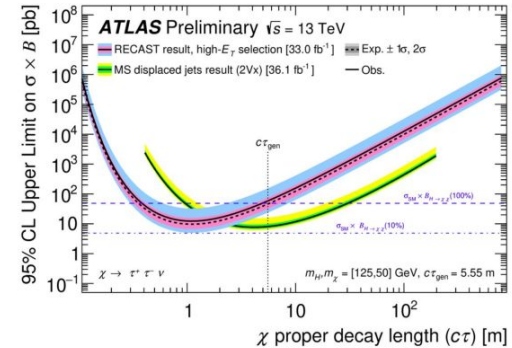
- Command-line client
- Web UI

<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

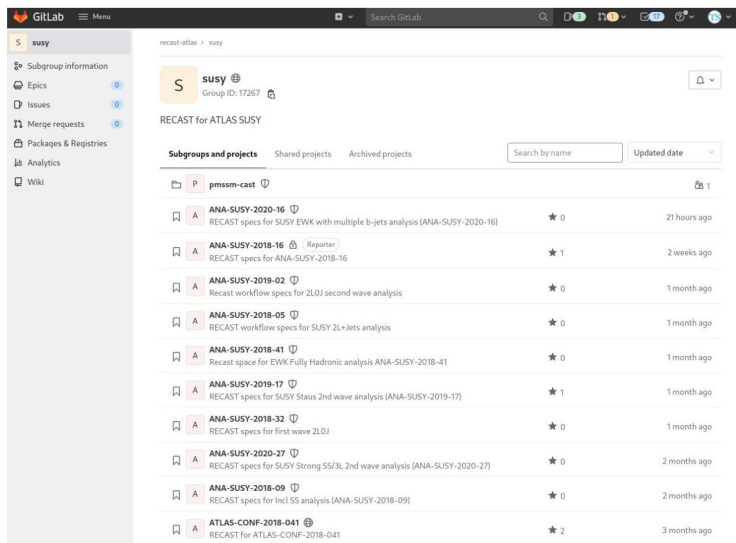


Figure 1. A screenshot of the ATLAS SUSY group analyses preserved on GitLab. Each repository is labeled with the internal ATLAS analysis identifier and contains both workflow files and additional data files needed for the computational processing.

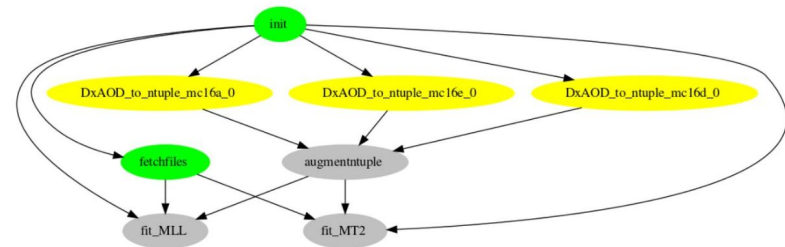


Figure 2. A typical pMSSM workflow. The computational runtime is about 10 minutes without systematics (test payload) and about 10 hours with all systematics (real payload).

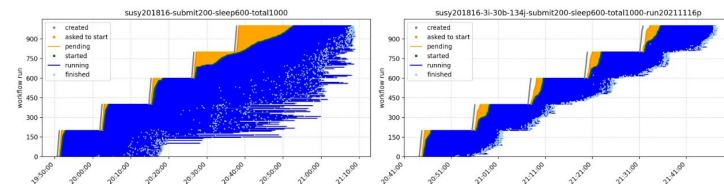



Figure 8. A scalability test submitting 200 workflows every 10 minutes. A cluster with 448 cores (left) cannot keep up with the load. A cluster with 1072 cores (right) can comfortably hold the incoming workload.

<https://arxiv.org/abs/2403.03494>

Streamlining the execution of thousands of reinterpretation workflows at scale


First ATLAS pMSSM Run-2 searches published

arXiv:2402.01392v2 [hep-ex] 30 May 2024



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (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

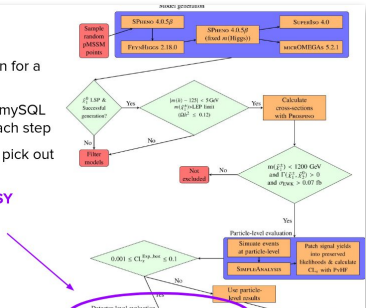
A summary of the constraints from searches performed by the ATLAS Collaboration for the electroweak production of charginos and neutralinos is presented. Results from eight separate ATLAS searches are considered, each using 140 fb^{-1} of proton–proton data at a centre-of-mass energy of $\sqrt{s} = 13 \text{ TeV}$ 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 “tunnel 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.

© 2024 CERN for the benefit of the ATLAS Collaboration.
Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.

<https://arxiv.org/abs/2402.01392>

Workflow

- Workflow to evaluate exclusion for a sample of pMSSM models
- Implemented in python using MySQL database to store results of each step
- Various constraints applied to pick out interesting models
- **RECAST is used to apply SUSY searches to these models**



recast and reana in the pMSSM scan

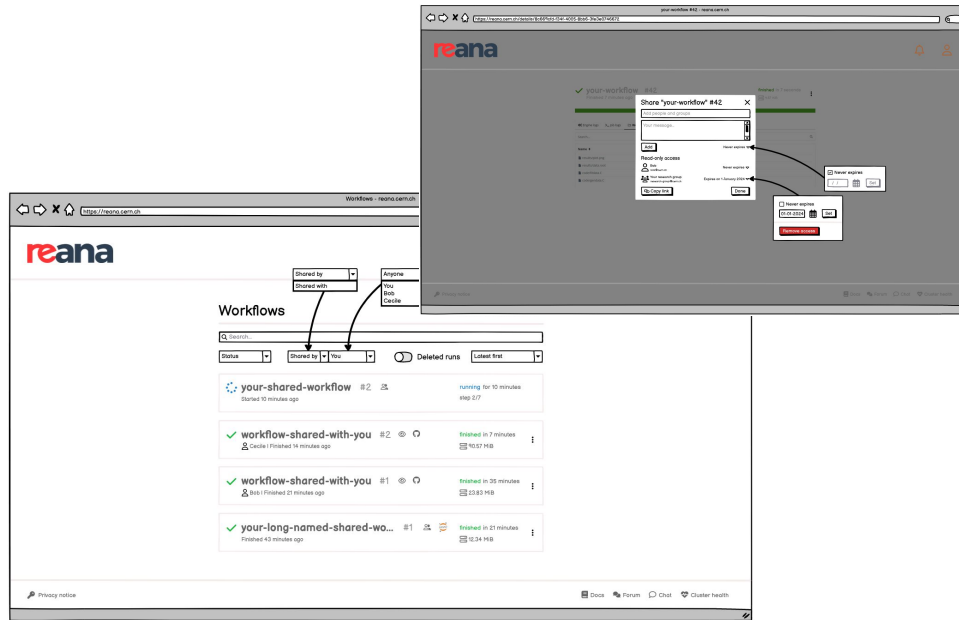
- Eight analyses run using RECAST
- 1878 models were processed with RECAST
- $\geq 9\%$ of the 21,177 models in the scan
- 9561 REANA jobs
 - including many failed tests and re-runs
 - Web-page monitoring very useful



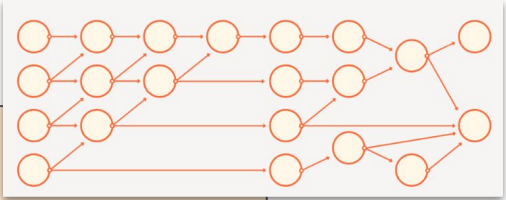
Ben Hodkinson

<https://indico.cern.ch/event/1380367/>

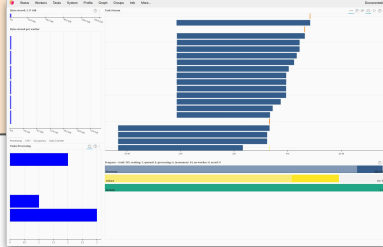
A glimpse on forthcoming REANA features



Share workflows with colleagues



```
inputs:  
  files:  
    - myanalysis.py  
workflow:  
  type: serial  
  resources:  
    dask:  
      image: mydaskenv:2023.10.1  
      cores: 100  
  specification:  
    steps:  
      - environment: mydaskenv:2023.10.1  
        commands:  
          - python myanalysis.py  
outputs:  
  files:  
    - myhistogram.png
```



The image displays a workflow configuration snippet for a Dask workflow. The configuration includes inputs (files), workflow type (serial), resources (dask), and specification (steps). The 'dask' resource is configured with 'image: mydaskenv:2023.10.1' and 'cores: 100'. The 'steps' section includes an environment configuration and a command to run 'python myanalysis.py'. The outputs section lists 'myhistogram.png'. To the right, a diagram shows a workflow graph with nodes and arrows. Below the code, a screenshot of the Dask dashboard shows a histogram and other workflow metrics.

Support for Dask workflows

Driving future reproducibility

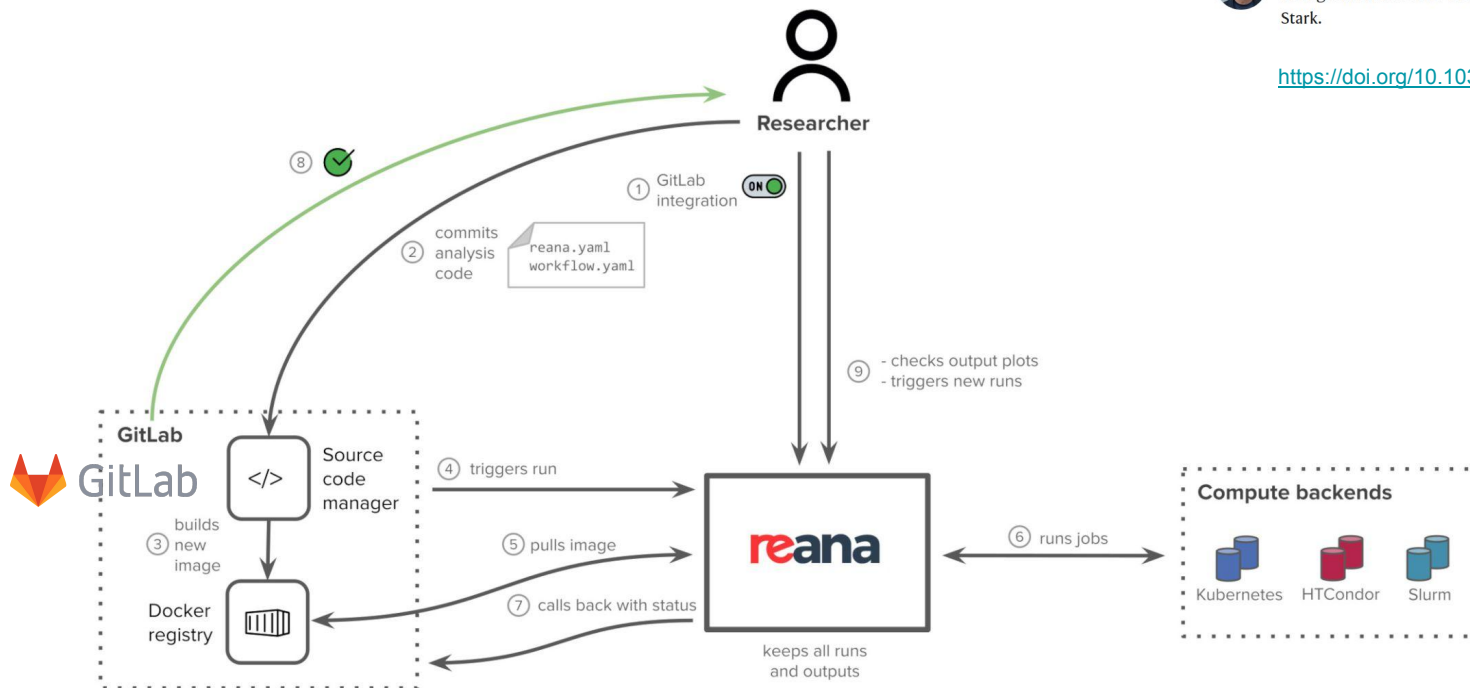
WORLD VIEW · 24 MAY 2018

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

Analysis Preservation Training
CI/CD + Containerization

October 16 - 18, 2023

Instructors: <ul style="list-style-type: none">• CI / CD: Mason Proffitt• Docker: Marco Donadoni• REANA: Tibor Simko	Mentors and local organizers: <ul style="list-style-type: none">• Carlos Escobar• Miguell Vilaplana• Emma Torro
---	--

 GOBIERNO DE ESPAÑA
 MINISTERIO DE CIENCIA E INNOVACIÓN
PID2021-124912NB-I00
PID2019-104301RB-C21
CIPROM/2022/70

 IFIC
 CSIC
UNIVERSITAT ID VALÈNCIA

 ATLASVLC
Instituto de Física Corpuscular - Valencia

 GENERALITAT VALENCIANA
Conselleria d'Educació, Cultura i Esport

 GenT

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

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

THE HEP SOFTWARE FOUNDATION (HSF)

HSF-TN-2024-01
April 2024

Analysis Facilities White Paper



HEP Software Foundation

D. Ciangottini^{1,3}, A. Forti^{2,3}, L. Heinrich^{3,4}, N. Skidmore^{5,6},
 C. Alpigiani³, M. Aly², D. Benjamin⁶, B. Bockelman⁷, L. Bryant⁸, J. Catmore⁹, M. D'Alfonso¹⁰, A.
 Delgado Peris¹¹, C. Dogliani⁶, G. Duckeck¹², P. Elmer¹³, J. Eschler¹², M. Fairclert¹³, J. Frust¹⁴, R.
 Gordinos¹⁵, V. Gammone¹⁶, M. Giffels¹⁶, J. Gooding¹⁶, E. Gramstad¹⁷, L. Gray¹⁸, B. Hegner¹⁹, A.
 Held¹³, J. Hernández¹⁶, B. Holzman¹⁶, F. Hu²⁰, B. K. Jasha^{18,19}, D. Kondratyev²⁰, E. Koufalis³, L.
 Kreczko²¹, I. Krommydas²², T. Kuhn¹⁰, E. Laucou⁶, C. Lange²³, D. Lange¹¹, J. Lange¹¹, P. Lenzi¹,
 T. Lindner²⁴, V. Martino Onofre²⁵, S. McKee²⁶, J. F. Molina²⁷, M. Nouhouar²⁸, A. Novak²⁹, I.
 Osherson³¹, F. Ould-Saada³, A. P. Pages²⁸, K. Pedro³⁰, A. Perez-Calero Yaguirero¹⁷, S. Piperno²⁰, J.
 Pivarski³¹, E. Rodrigues²⁹, N. Sahoo³⁰, A. Sciala³¹, M. Schulz³¹, L. Sexton-Kennedy³¹, O.
 Shadura³², T. Simko³³, N. Smith³⁰, D. Spiga³, G. Stark³³, G. Stewart³¹, I. Vukotic⁶, G. Watts⁶,

¹Editor, ³INFN, ²University of Manchester, ³Technische Universität München, ⁴University of Warwick ⁵University of Washington ⁶Brockhaven National Laboratory ⁷Margridge Institute for Research ⁸University of Chicago ⁹University of Duke ¹⁰Ladislav Mates ¹¹Wisconsin-Madison ¹²Urbain Le Verrier ¹³Wisconsin-Madison ¹⁴Urbain Le Verrier ¹⁵Purdue University ¹⁶Urbain Le Verrier ¹⁷Purdue University ¹⁸Urbain Le Verrier ¹⁹Urbain Le Verrier ²⁰Urbain Le Verrier ²¹Urbain Le Verrier ²²Urbain Le Verrier ²³Urbain Le Verrier ²⁴Urbain Le Verrier ²⁵Urbain Le Verrier ²⁶Urbain Le Verrier ²⁷Urbain Le Verrier ²⁸Urbain Le Verrier ²⁹Urbain Le Verrier ³⁰Urbain Le Verrier ³¹Urbain Le Verrier ³²Urbain Le Verrier ³³Urbain Le Verrier

This white paper attempts to summarize collected through if forms [1], established May 2022, and the attempts to cover a

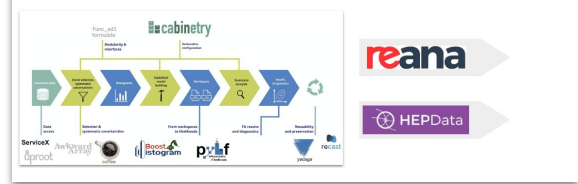
© Licence CC-BY-4.0

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



Analysis Grand Challenge IRIS-HEP implementation

- Columnar data extraction from large dataset
 - Processing of that data (event filtering, construction of observables, evaluation of systematic uncertainties) into histograms
 - Statistical model construction and statistical inference
 - Relevant visualisation for this steps
- + **Adding analysis preservation step to AGC pipeline**



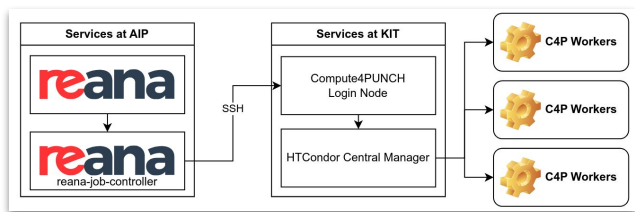
<https://indico.cern.ch/event/1446410/>



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

<https://arxiv.org/abs/2404.02100>

Recent new deployments



Hardware	Software	User Portals
Compute Nodes GPUs	Linux Docker containers Kubernetes Slurm	COCALC reana ^{NEW}
Storage	MinIO-S3 LustreFS-iB GlusterFS/mfs	MINIO GitLab
Network: Intern/Public	Infiniband 10G 1G	
Job queues, resource management		

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

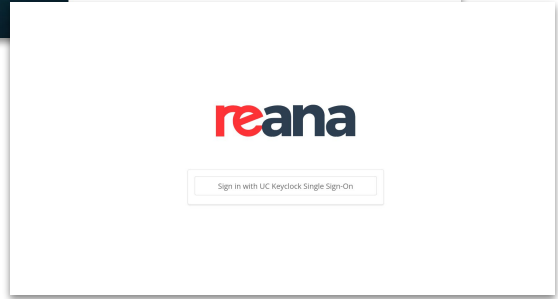
A Reana testbed in the US ATLAS Computing Facility

Eric Lancon
04/04/24



Proposal:
Deploy a test Reana Instance in the US ATLAS Computing Facility

- **Unlocking New Possibilities with the US Reana Test Instance**
 - Dedicated to ATLAS
 - Faster analysis time than CERN
 - Large available batch pools at the US Analysis Facilities, much larger in size than the Reana cluster at CERN.
 - Accelerating the analysis process and opening new opportunities
 - Enabling a 'hybrid' model where a batch farm is used for CPU-intensive tasks
 - Interfaced with grid storage
- Additionally, batch CPU resources could either be:
 - Dedicated ATLAS cluster (T2)
 - Opportunistic usage or by reservation/campaign (implementation to be worked on) of additional resources at local institutions



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



The screenshot shows the OpenData portal interface with several key sections highlighted by orange arrows:

- Search:** A search bar and filter options at the top.
- Details:** A section showing dataset characteristics and system details.
- Production script:** A section showing the workflow used to generate the data.
- Generator parameters:** A section showing the parameters used in the simulation.
- Configuration files:** A section showing the configuration files used in the simulation.

SingleElectron primary dataset sample in RAW format from RunA of 2011 (from /SingleElectron/Run2011A-v1/RAW)

File list for this dataset

Description

A sample from SingleElectron primary dataset in RAW format from RunA of 2011. Run range [141224,163285]. This dataset contains selected runs from 2011 RunA. The list of validated lumi sections, which must be applied to all analyses on events reconstructed from these data, can be found in CMS list of validated runs CERN Open Data Portal. DOI:10.7802/OPENDATA.CMS.OPEN.WLNR

Dataset characteristics

294428 events, 316 files, 424.3 GB in total.

How can you use these data?

These data are in RAW format and not directly usable in analysis. The reconstructed data reprocessed from these RAW data are included in the data of this record. The reconstruction step can be repeated with the configuration file below and the resulting AOD has been confirmed to be identical with the original one with comparison code available in Validation code to plot basic physics objects from AOD

SingleElectron primary dataset in AOD format from RunA of 2011 (SingleElectron/Run2011A-12002013-v1/AOD)

File list for this dataset

Description

SingleElectron primary dataset in AOD format from RunA of 2011. Run range [141224,163285]. This dataset contains all runs from 2011 RunA. The list of validated lumi sections, which must be applied to all analyses, can be found in CMS list of validated runs CERN Open Data Portal. DOI:10.7802/OPENDATA.CMS.OPEN.WLNR

Dataset characteristics

470999 events, 1542 files, 5.4 TB in total.

How were these data selected?

Events stored in this dataset were selected because of the presence of at least one high-energy electron in the event.

Data linking / FET

The collision data were assigned to different RAW datasets using the following configuration file.

Data processing / RECO

This primary AOD dataset was processed from the RAW dataset by the following step:

Step: RECO

Release: CMSSW_5_3_12_patch1

Global tag: FT-3-12-10-48

Configuration file for RECO step: rec_2011A_SimReco

dataset=Jet
year=2011A
1 input parameters

2 workflow factory

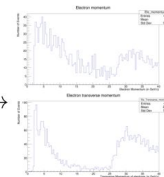
3 reana.yaml



5 serving open data files



4 run by REANA platform



6 output histograms

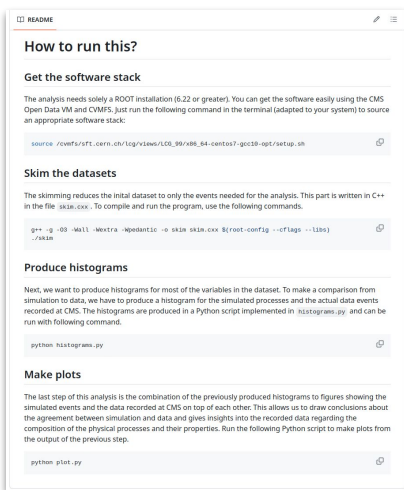
CMS open data coming with detailed provenance information

<https://doi.org/10.1051/epic/onf/202024508014>

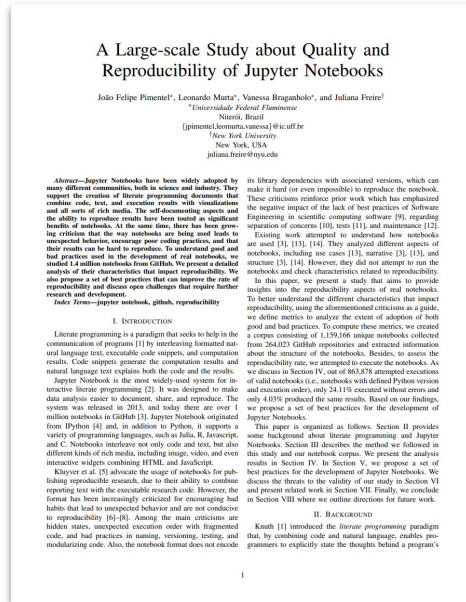
Reprocessing AOD from RAW samples

Use case 2: Is preserved data usable?

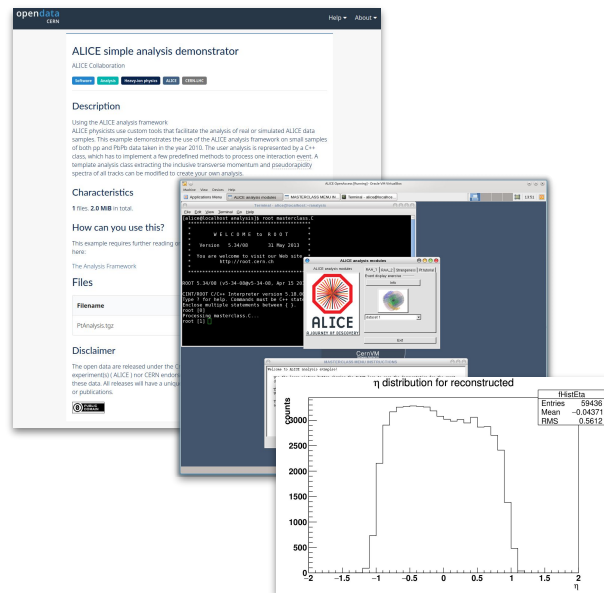
Verifying usage patterns



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”

Feature: cms-htautau-nanoaoad

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

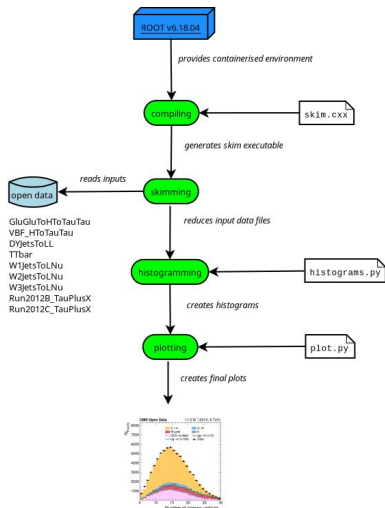
When the workflow is finished
Then the workspace size should be less than 75 MiB

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

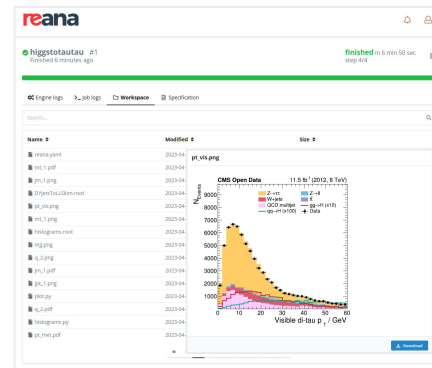
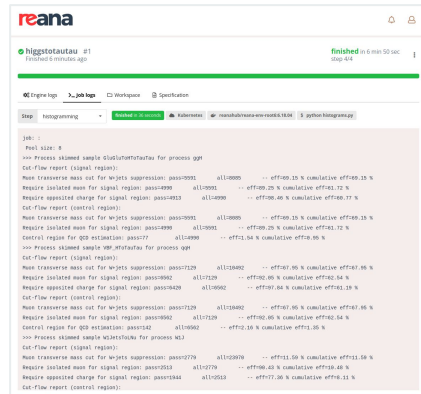
When the workflow is finished
Then the workflow run duration should be less than 25 minutes
And the duration of the step "skimming" should be less than 20 minutes



An example studying
 $H \rightarrow \tau\tau$ lepton decays
uses nine published
CMS open datasets

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

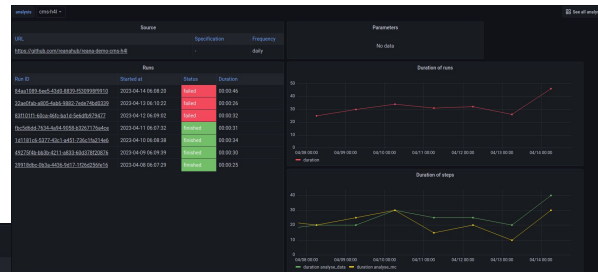
- check workspace content and output files
- check produced log messages
- check execution duration of steps



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	Success	Success	Success	Success	Success
alice-pt-analysis	5 hours ago		00:01:10	Success	Success	Success	Success	Success
atlas-recast	5 hours ago		00:01:10	Success	Success	Success	Success	Success
cms-dimuon-mass-spectrum	8 days ago		00:01:03	Success	Success	Success	Success	Success
cms-dimuon-spectrum	5 hours ago		00:01:22	Success	Success	Success	Success	Success
cms-dimuon-spectrum-nanoaod	5 hours ago		00:01:52	Success	Success	Success	Success	Success
cms-h4l	5 hours ago	2 days ago	00:02:02	Success	Success	Failure	Failure	Failure
cms-h4l-nanoaod	5 hours ago		00:03:46	Success	Success	Success	Success	Success
cms-htautau-nanoaod	5 hours ago		00:07:08	Success	Success	Success	Success	Success



- 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

The diagram illustrates a workflow for data analysis. It begins with an 'open data' page showing a dataset titled 'WJsetsToLNU dataset in reduced NanoAOD format for education and outreach'. This dataset is used for an 'analysis' of Higgs boson decays, which includes a plot of 'Visible d-tau mass / GeV' versus 'Tau η '. The analysis is then launched on the 'REANA' platform, which provides a 'Your workflows' dashboard. The dashboard shows a workflow named 'higgs' with a status of 'Finished'. A 'Launch on REANA' button is visible, and a 'plot_v0.png' plot is shown, which is a histogram of 'Visible d-tau mass / GeV' versus 'Tau η '.

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