The vision for the PUNCH Science Data Platform

PUNCH LUNCH, 14 January 2021

Particles, Universe, NuClei and Hadrons for the NFDI

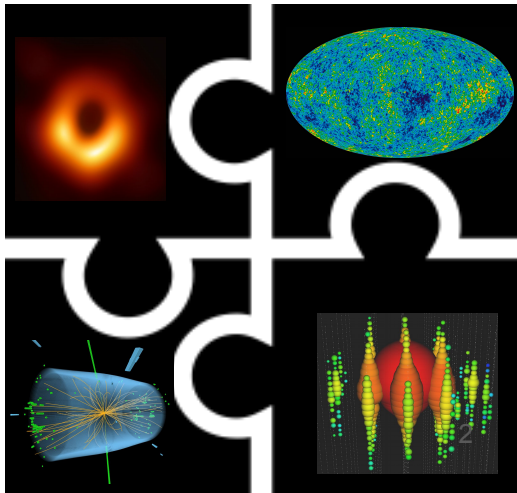
S. Neubert, P. Bechtle (U Bonn)

Contents

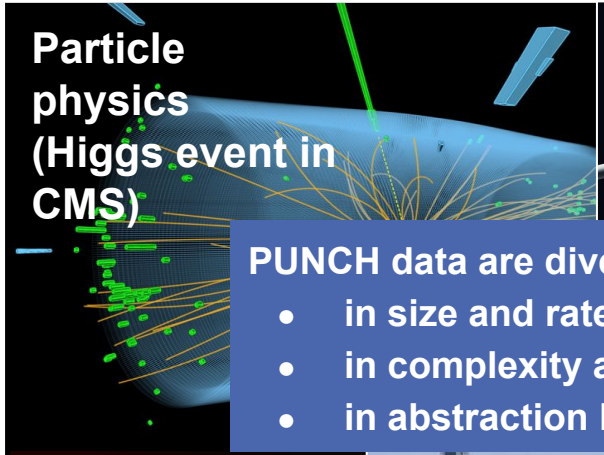
- **Introduction**
  - PUNCH4NFDI & Motivation
  - The environment of the PUNCH SDP: Other platform projects
  - The specific goals of the PUNCH SDP: Open, FAIR and *knowledge centered*

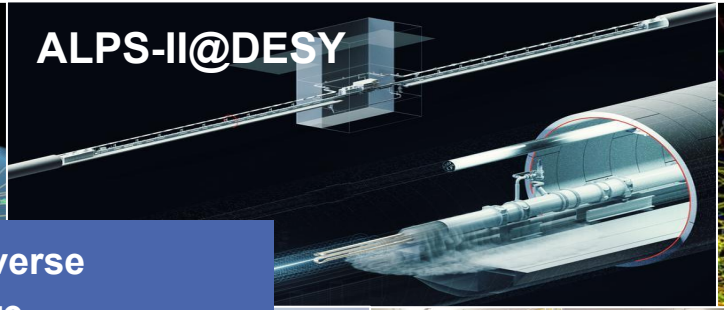
- **Why PUNCH is the right community to engage in a challenging new idea**
  - Examples for complex integrated analysis workflows in our fields
  - Open data in PUNCH
  - The interplay of open workflows and analysis preservation

- **Some components and capabilities of the PUNCH SDP**
  - The idea of the (Dynamic) Research Product, and the catalogue behind it
  - The use cases and a few exemplifications



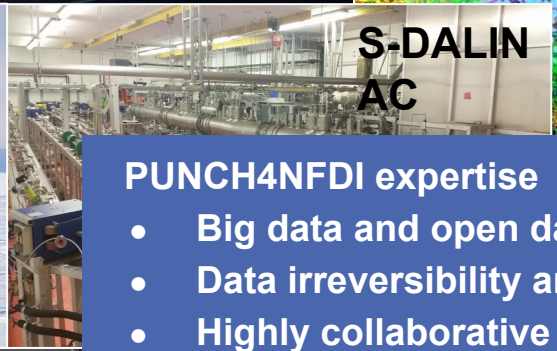
PUNCH4NFDI: Particles, Universe, NuClei, Hadrons for the **NFDI**

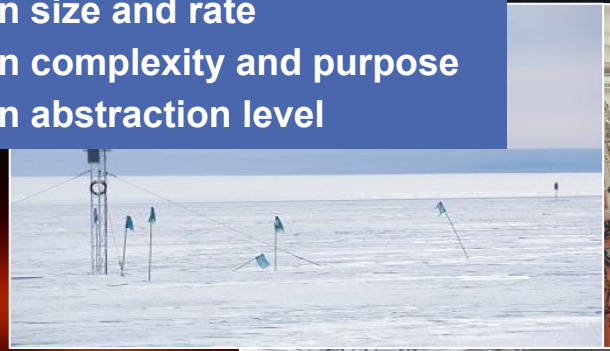




PUNCH data are diverse

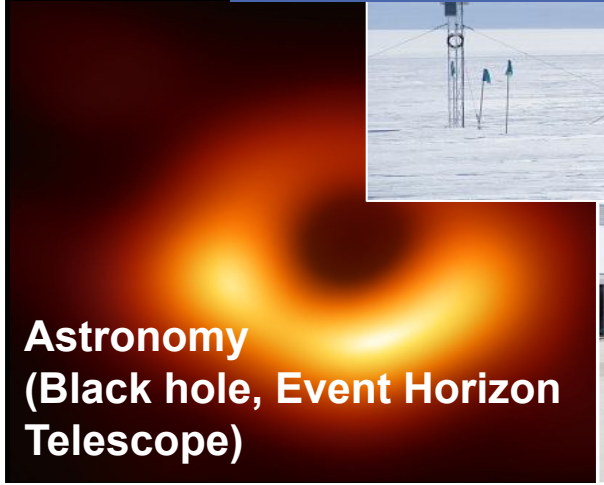
- in size and rate
- in complexity and purpose
- in abstraction level

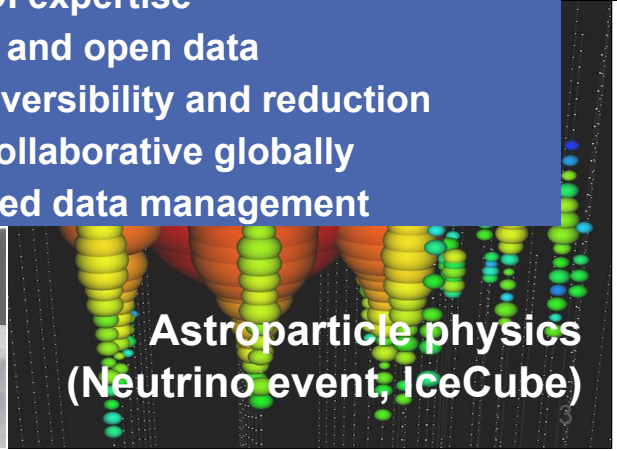




PUNCH4NFDI expertise

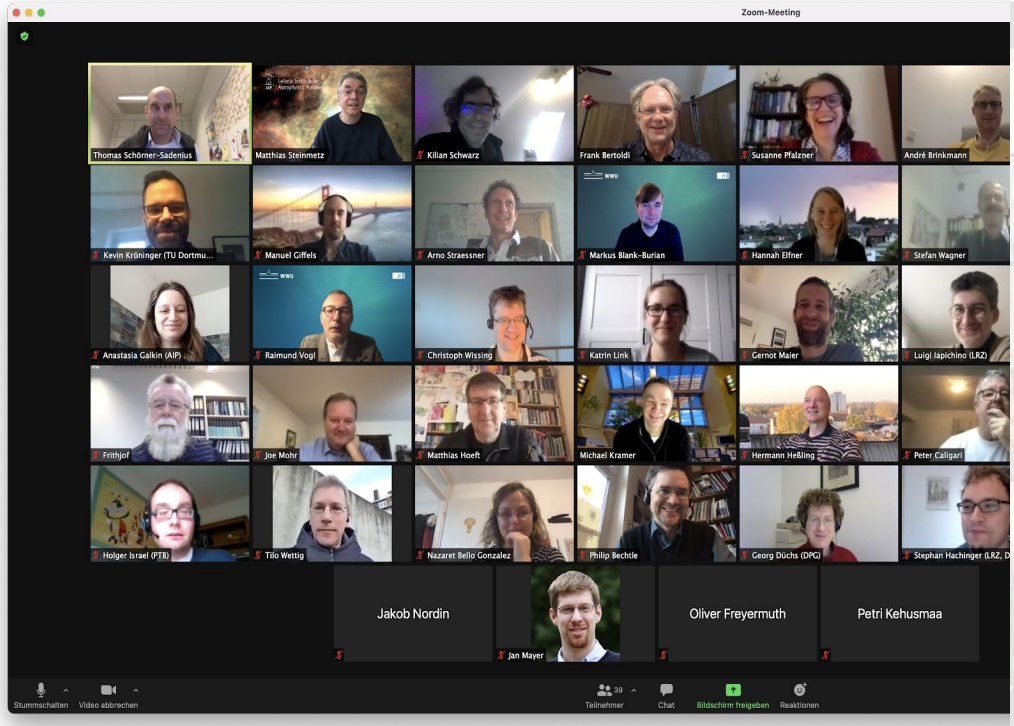
- Big data and open data
- Data irreversibility and reduction
- Highly collaborative globally distributed data management





Disclaimer:

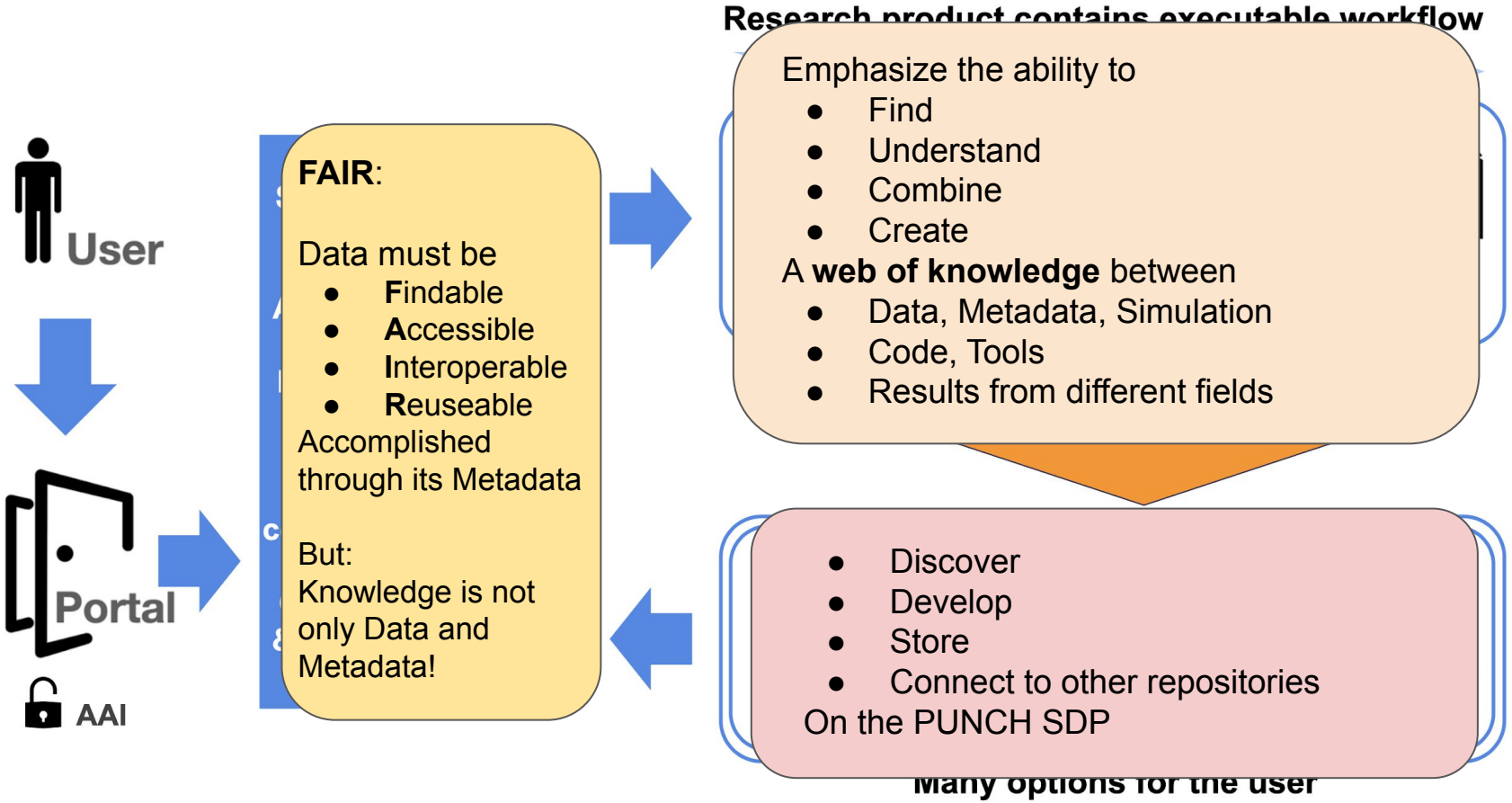
Both authors of this talk are particle/hadron physicists.
Almost all examples are borrowed from our own respective areas, but the **design of the tools** designed here is inspired by and will serve a **wide variety of sciences, even beyond PUNCH**



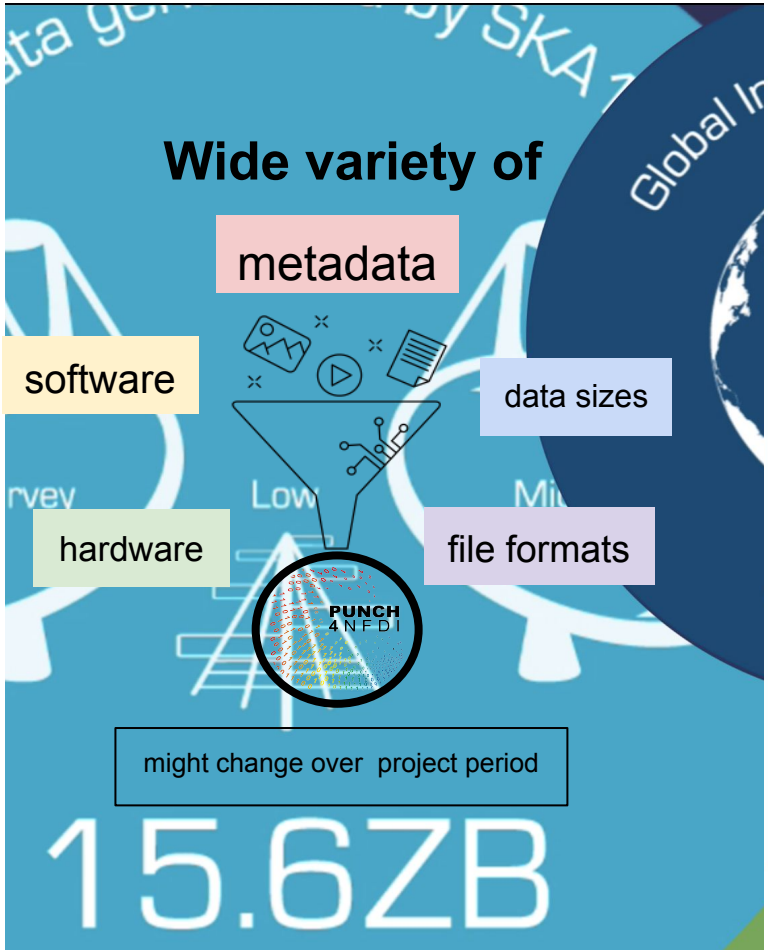
- **Merger of particle, astro(particle), hadron & nuclear physics – basic research at large infrastructures**
- **>4000 scientists (Ph.D.), >9000 DPG members, and deeply rooted in our community**
- **Users are developers are providers.**

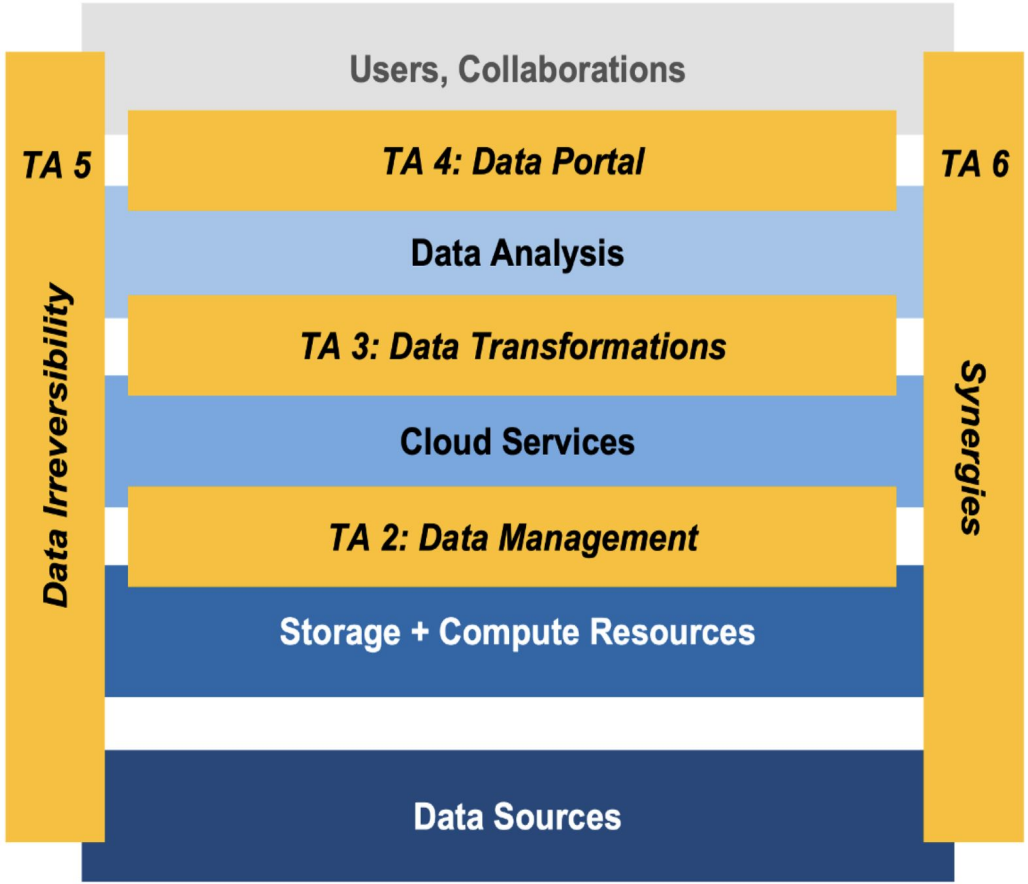
PUNCH4NFDI – a consortium by the community for the community – and beyond

The PUNCH Science Data Platform SDP



The Genesis of the PUNCH SDP





The Layer model of data abstractions

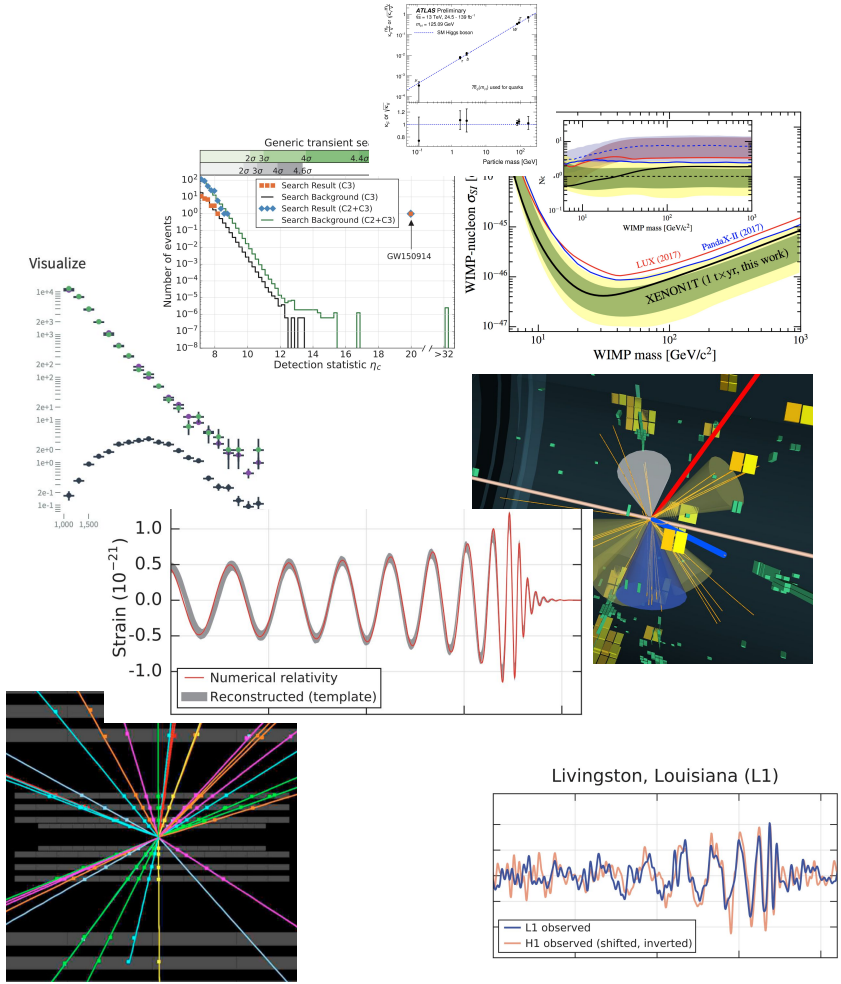
Find, Access, Interact, Reuse (TA4+all)

Transformations (TA3)

Interpreted Results (Properties of physics models, particle masses, ...)	+ Metadata on Theory Calculations, Statistics and Interpretation
Derived Quantities (Event or Astronomical Object Classification, Distributions)	+ Metadata on Analysis and Simulation
Reconstructed Physics Objects (Jets, Particles, Spectra, ..)	+ Metadata on Higher Level Reconstruction and Calibration
Reconstructed Constituents (Clusters, Tracks, ...)	+ Metadata on Calibration and Low Level Reconstruction
Raw Hits	Metadata associated to data taking

Processing and Storage (TA2)

Data taking and reduction (TA5)



The planned PUNCH SDP is not alone: ESCAPE and EOSC

MANAGEMENT INNOVATION NETWORKING DISSEMINATION

Contacts:
Mark Allen

Added benefits of the PUNCH SDP:

1.) In close collaboration with developments for ESCAPE and the EOSC, bring the technology behind individual parts of ESCAPE and the EOSC to all of PUNCH, and then to more science
2.) Add important functionality development: RP and their catalogue

What is it?

The Virtual Observatory (VO) for astronomy integrates distributed infrastructures into one single virtual astronomy facility. ESCAPE supports the integration of the various multi-messenger ESFRI facilities and other research infrastructures into the EOSC through the VO framework, to ensure that their high-level scientific products and big data sets are openly accessible.

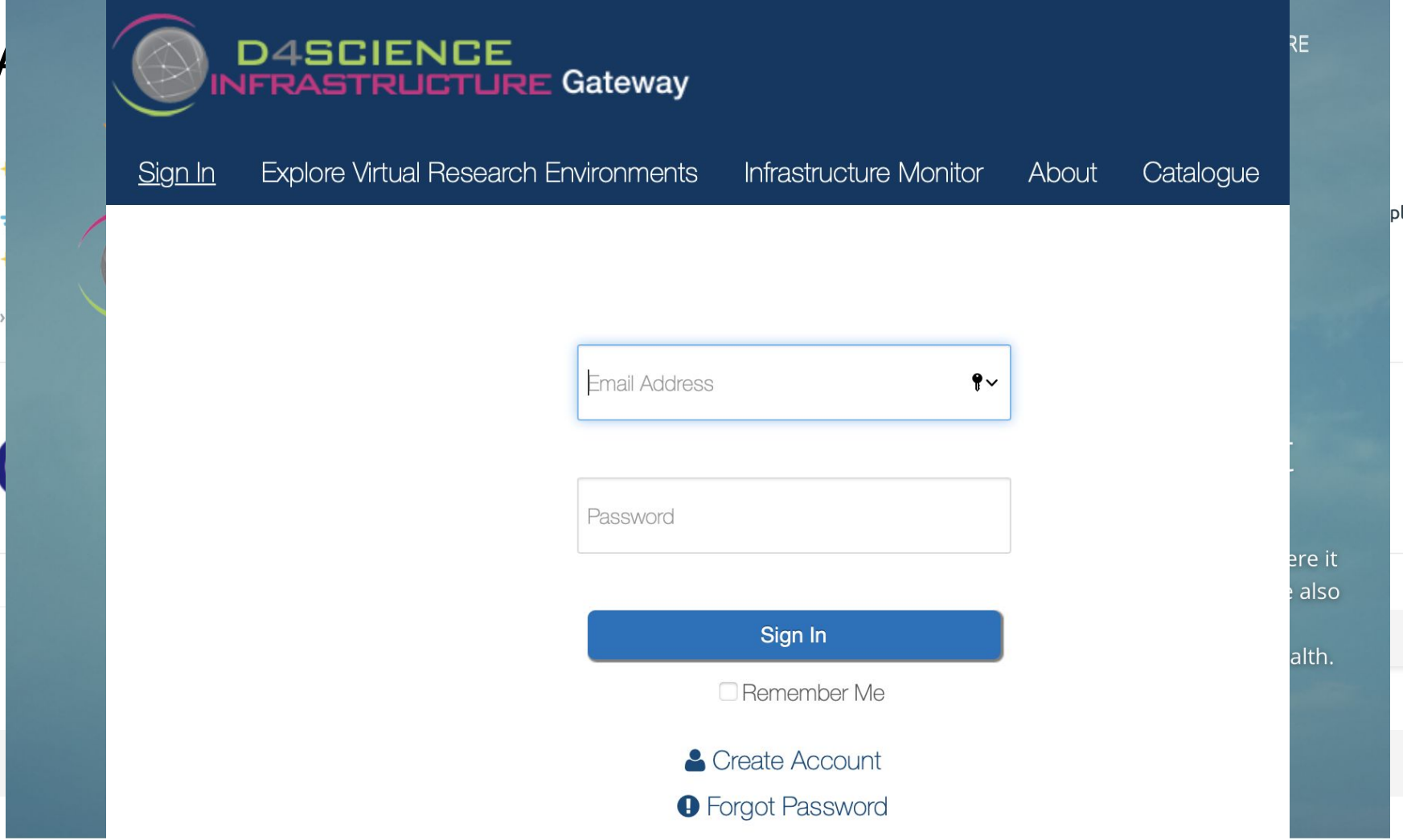
The VO is an essential component for the astronomy data landscape, it facilitates interoperability and re-use of data. ESCAPE will ensures that VO data are accessible to the European and international communities through EOSC. It sets the path for a new era of cross-disciplinary interoperability and connections to the necessary computing resources, as well as enabling the use of VO data in scientific analysis platforms. The ESCAPE enabled VO services aim to be a key for discovery and reuse, using common disciplinary IVOA standards and fostering good practices for data access, deposition and sharing of data, as well as for data management curation and preservation.

Who are the Target Users?

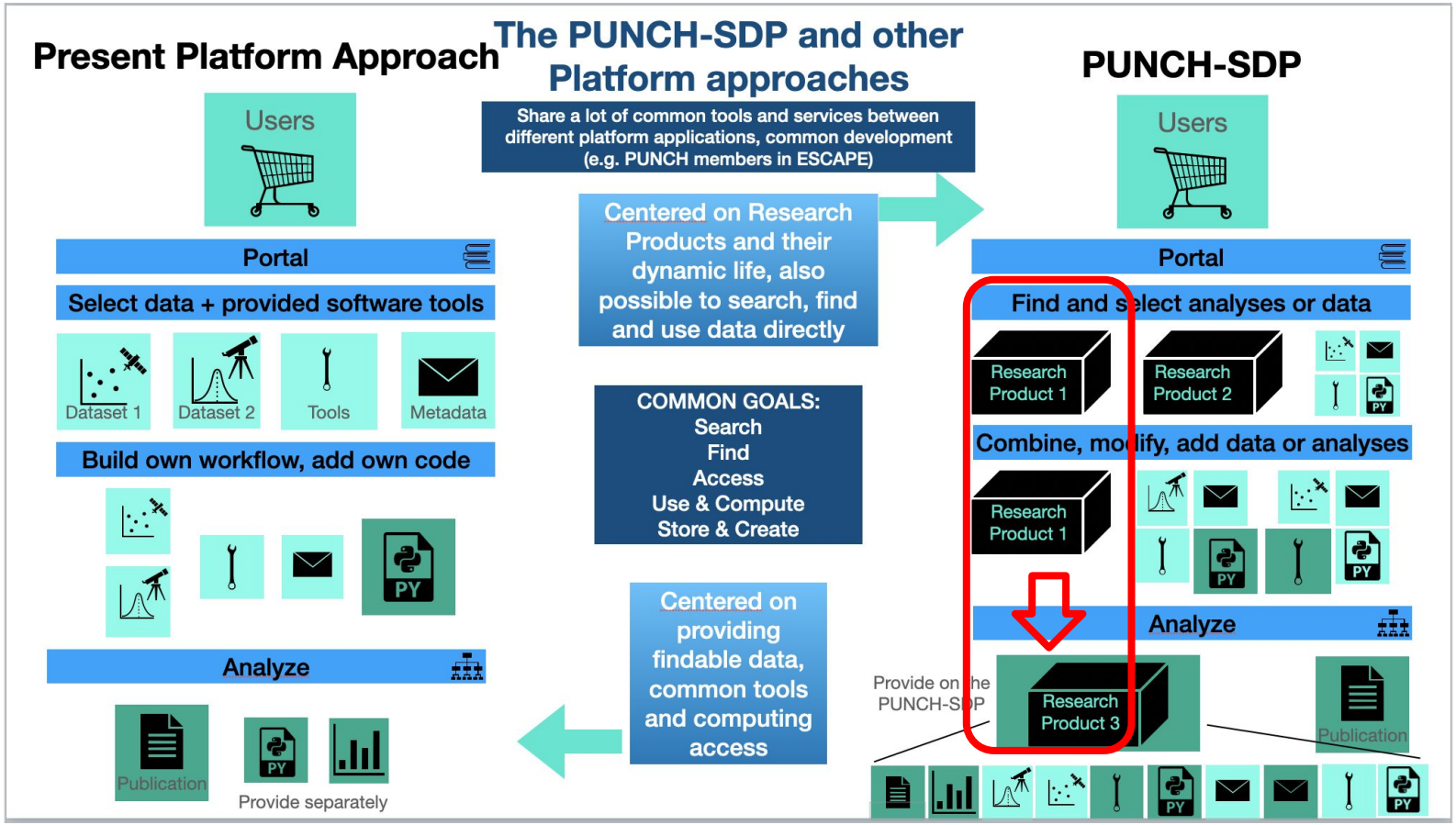
ESFRI, Research Infrastructures and other data providers, Research scientists, and Virtual Observatory builders

What is the value added / impact?

The VO provides interoperable access to European and international astronomical data and services. It is a pioneer of data sharing with a well established alliance of international partners who cooperate to build the necessary interoperability standards.


D4SCIENCE
INFRASTRUCTURE Gateway
Sign In
Explore Virtual Research Environments
Infrastructure Monitor
About
Catalogue
Email Address
Password
Sign In
Remember Me
Create Account
Forgot Password

The planned PUNCH SDP is different



The RP:
Research Product

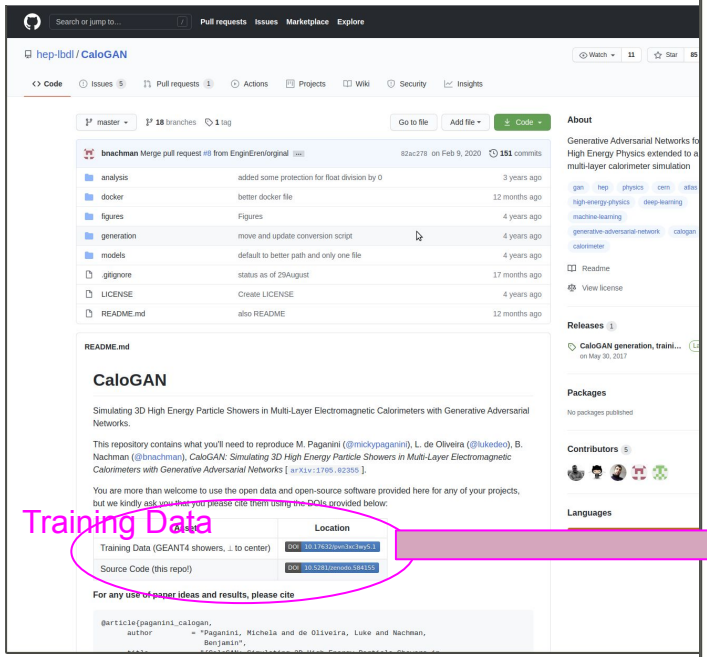
A representation of the data transformation and analysis workflow, and of the knowledge which lies in the combination of

- Data, metadata
- Simulation
- Code
- Tools
- Workflow

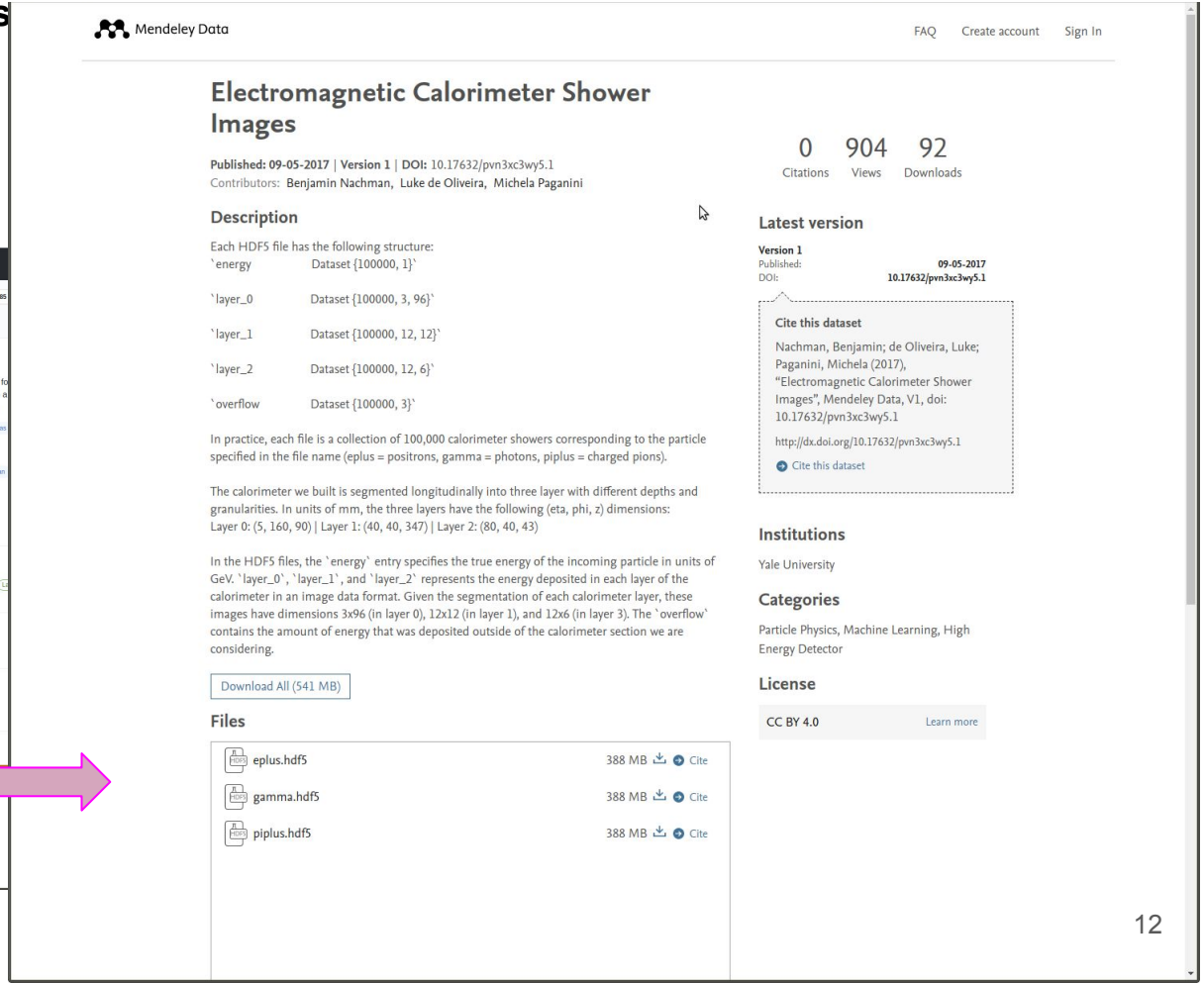
The foreseen components of the PUNCH SDP

- Access Control, Accounting: **A**uthentication and **A**uthorization **I**nfrastructure
- Searchable **catalogue** of **RP**s ((combination of) data, metadata, simulation, code, tools, example workflows, …)
- **Metadata** representation of the role of individual components in the RPs, and of metadata conversions
- Links to **other repositories** and catalogues (arxiv, inspirehep, VO, HepData, DOIs,...)
- Different **user interfaces** to choose from -- web, work in a terminal on your computer, ...
- Interface to **job execution**
- Data, (parts of the) Metadata and Simulation can be **stored** in the **data lakes** of the platform or on external repositories

Finding open data / code from papers
(current status)
https://arxiv.org/abs/1701.05927v2



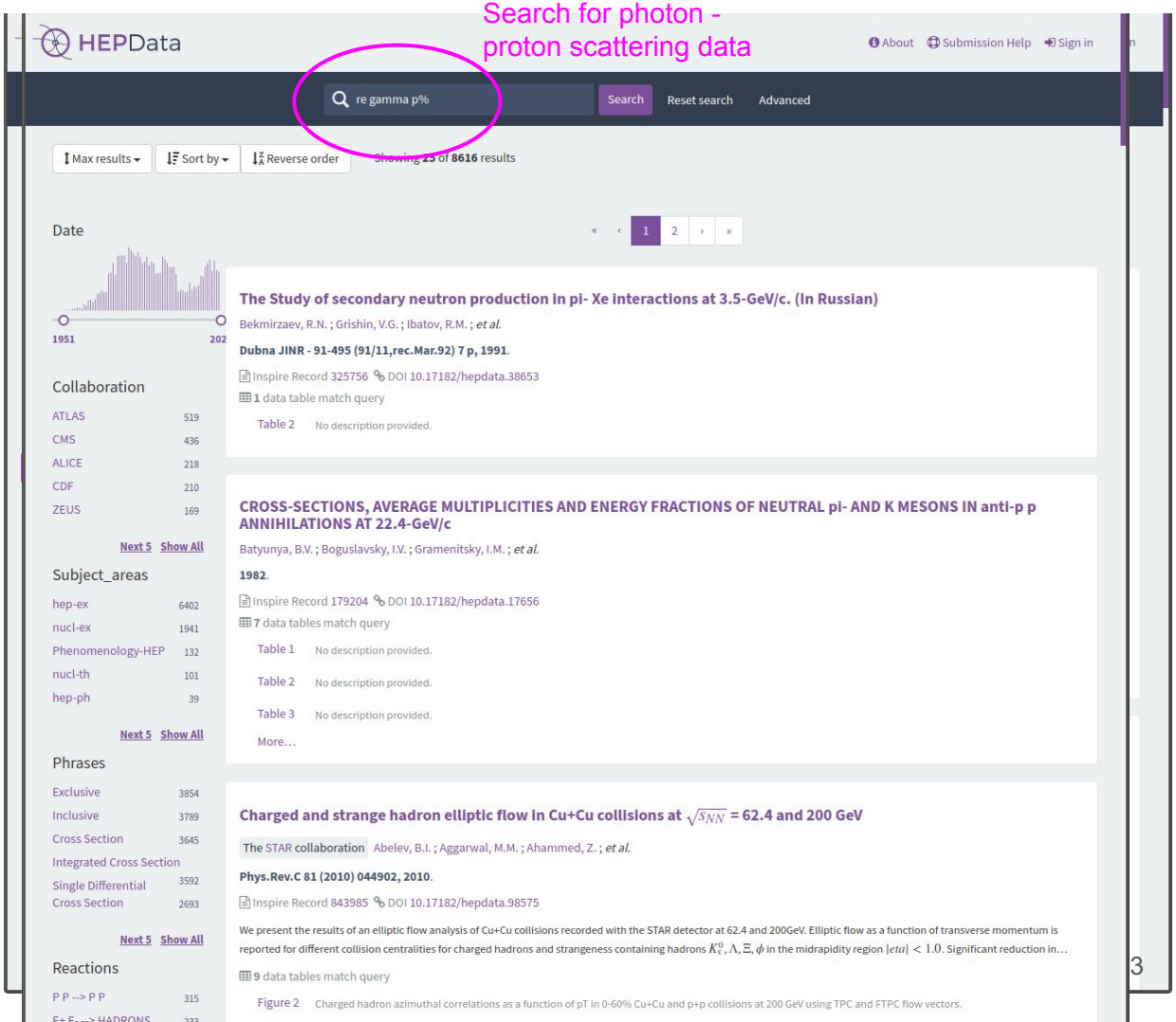
Training Data



HEPData
https://www.hepdata.net/

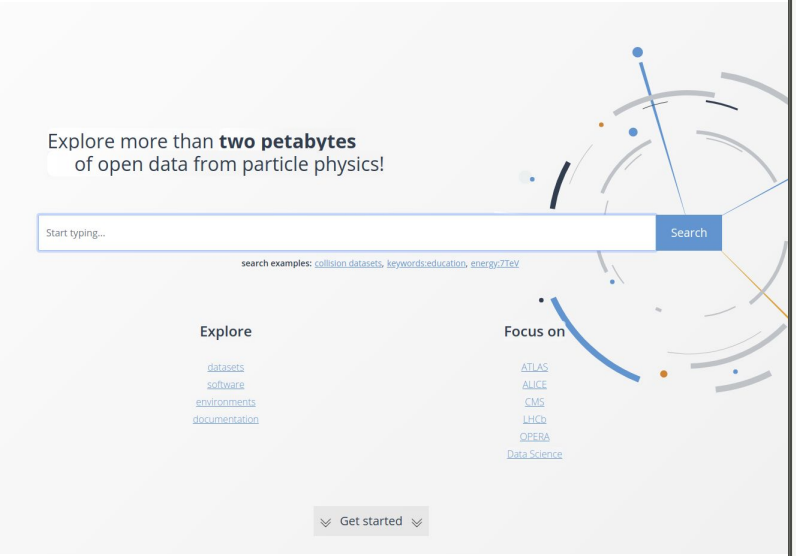
- mainly used to publish **high level observables** / analysis results
  - cross sections
  - particle properties
  - SM parameters
  - limits and confidence regions

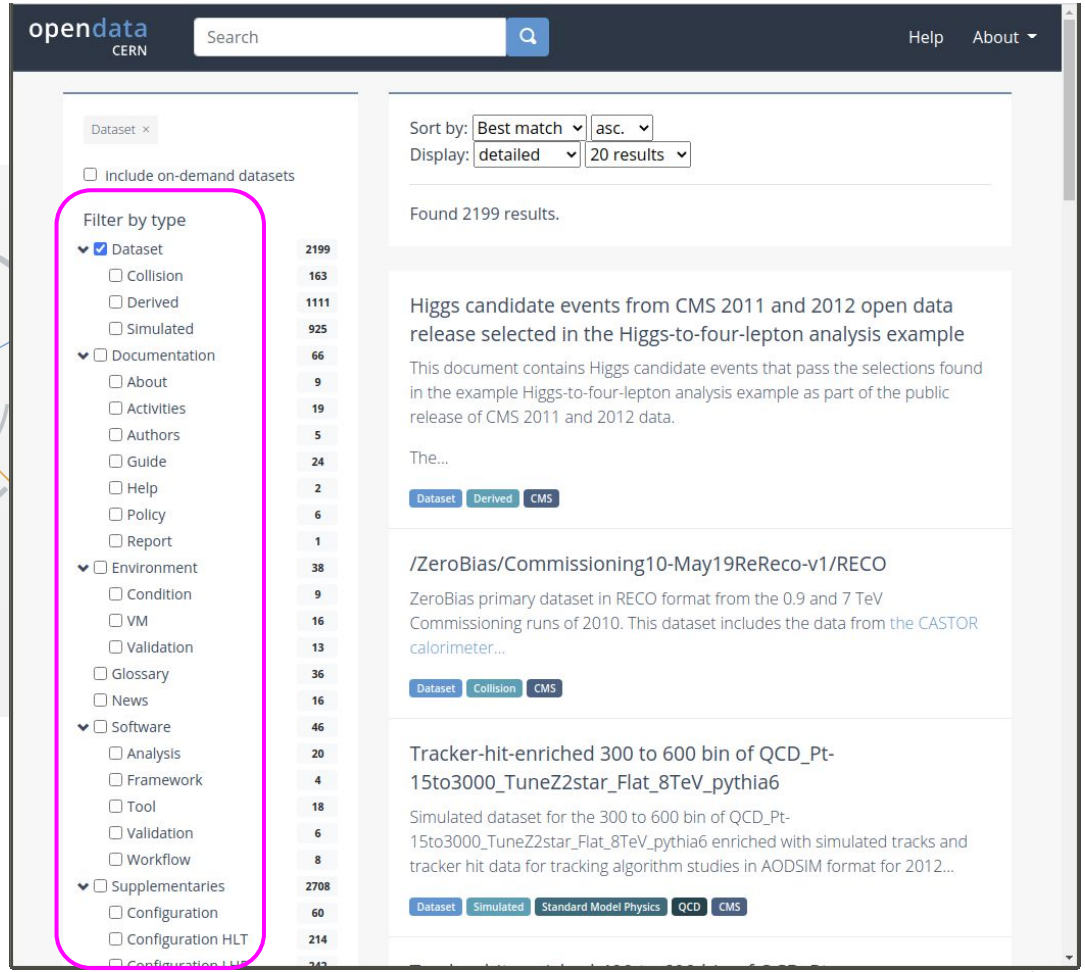
- machine readable format for observables and uncertainties



CERN Open Data

opendata.cern.ch

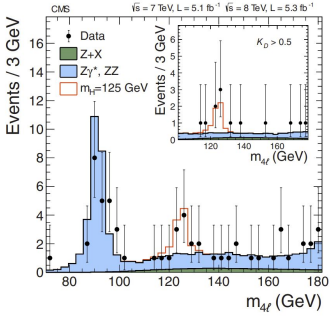


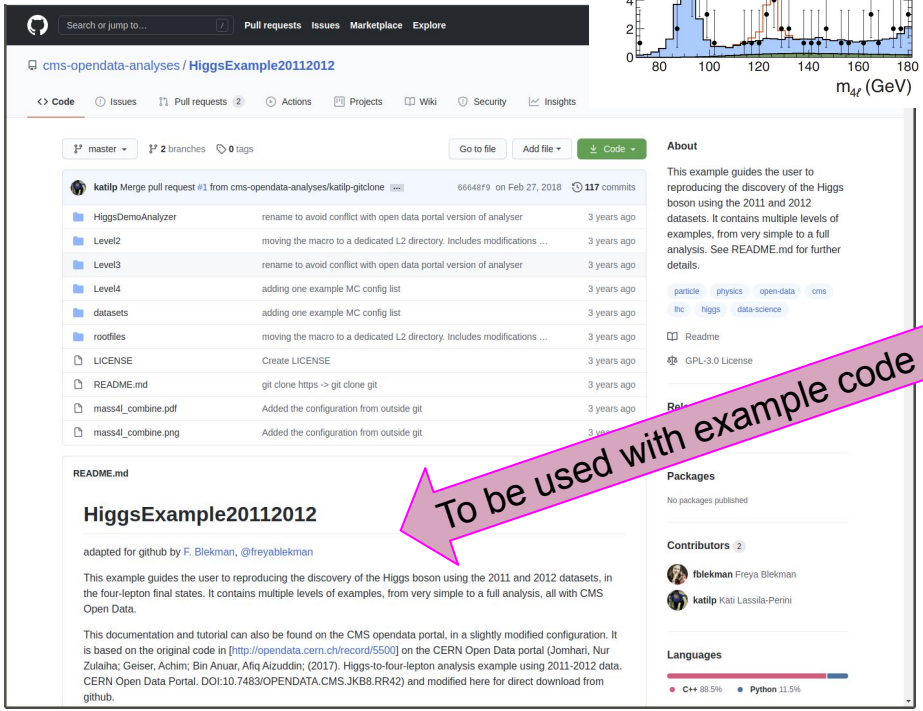


CERN Open Data example entry

opendata.cern.ch

allows to reproduce the Higgs discovery





To be used with example code

Root files for Higgs-to-four-lepton analysis example using 2011-2012 data

Jomhari, Nur Zulaiha ; Geiser, Achim ; Bin Anuar, Afiq Aizuddin

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin; (2017). Root files for Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JKB8.D634

Dataset | Derived | CMS | CERN-LHC

Description

This record contains root files that were processed for the Higgs analysis example on the CMS 2011-2012 Open Data.

Dataset characteristics

24 files. **1.6 MB** in total.

How can you use these data?

These datasets are provided as part of the Higgs-to-four-lepton analysis example using 2011-2012 data. You can download all root files needed for the Level 3 exercise to your working area by downloading first the file list `rootfilelist.txt` and then the root files with `wget -i rootfilelist.txt`.

Please note that the following files are not needed to produce the plot but are only the products of intermittent processing steps:

- DY101Jets12.root
- DY50Mag12.root
- DYTo2mu12.root
- TTJets11.root
- TTJets12.root

Higgs-to-four-lepton analysis example using 2011-2012 data

Files

Filename	Size	
rootfilelist.txt	1.1 kB	Download
DY1011.root	64.7 kB	Download
DY1012.root	66.1 kB	Download
DY101Jets12.root	63.9 kB	Download
DY50Mag12.root	69.7 kB	Download

First | Previous | 1 | 2 | 3 | 4 | 5 | Next | Last

Disclaimer

GNU General Public License (GPL) version 3

New CERN Open Data Policy published December 2020
https://cds.cern.ch/record/2745133
endorsed by all LHC collaborations

Lays out policy for 4 levels of complexity [arXiv:1205.4667]

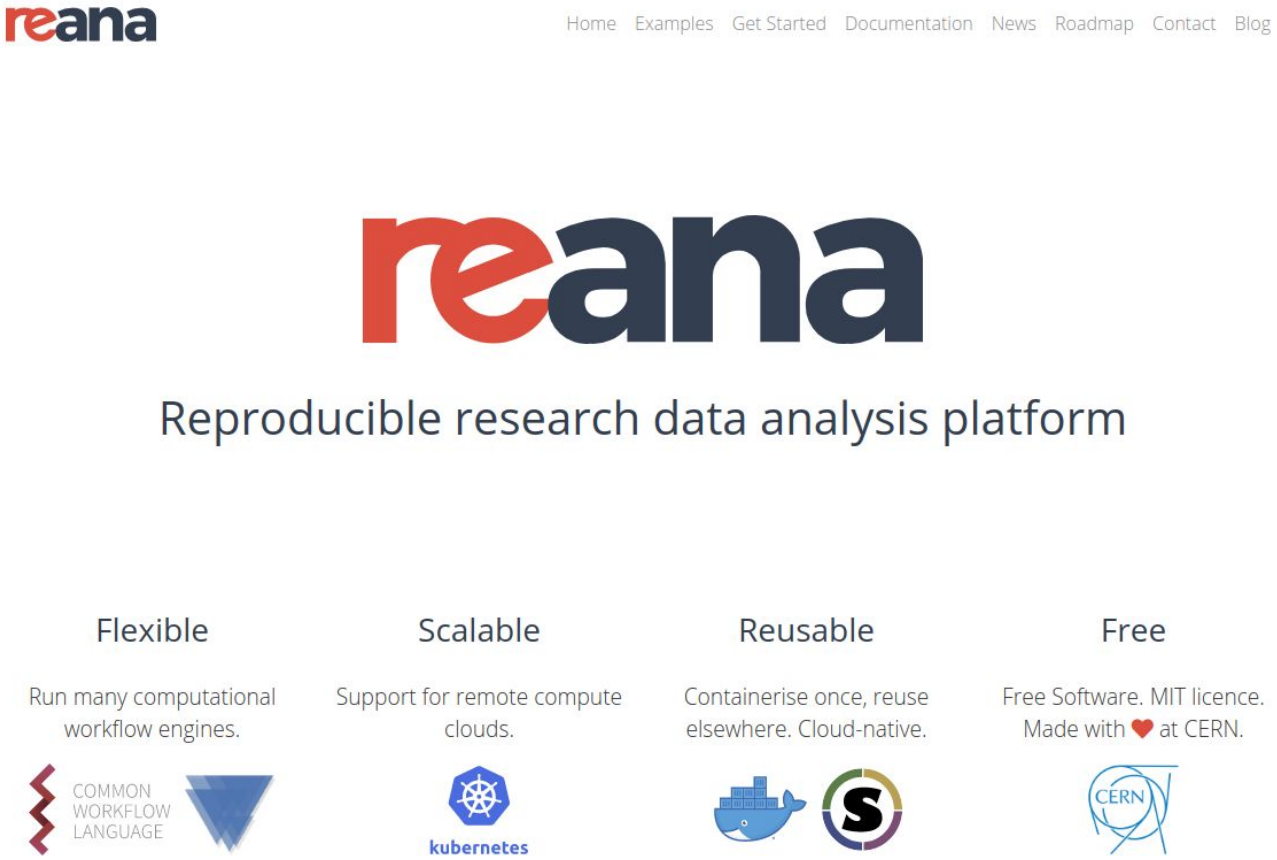
1. **Public results**
2. Outreach and education
3. **Reconstructed data**
4. Raw data
   (will not be made public)

"To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData, with ***selection routines stored in specialised tools****. The data made available may include* ***simplified or full binned likelihoods, as well as unbinned likelihoods*** *based on datasets of event-level observables extracted by the analyses.* ***Reinterpretation of published results is also made possible*** *through analysis preservation and direct collaboration with external researchers."*

"The LHC experiments will release ***calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies****. The release of these data will be accompanied by* ***provenance metadata****, and by a concurrent release of appropriate simulated data samples, software,* ***reproducible example analysis workflows****, and documentation.* ***Virtual computing environments*** *that are compatible with the data and software will be made available. The information provided will be* ***sufficient to allow high-quality analysis*** *[..]"*

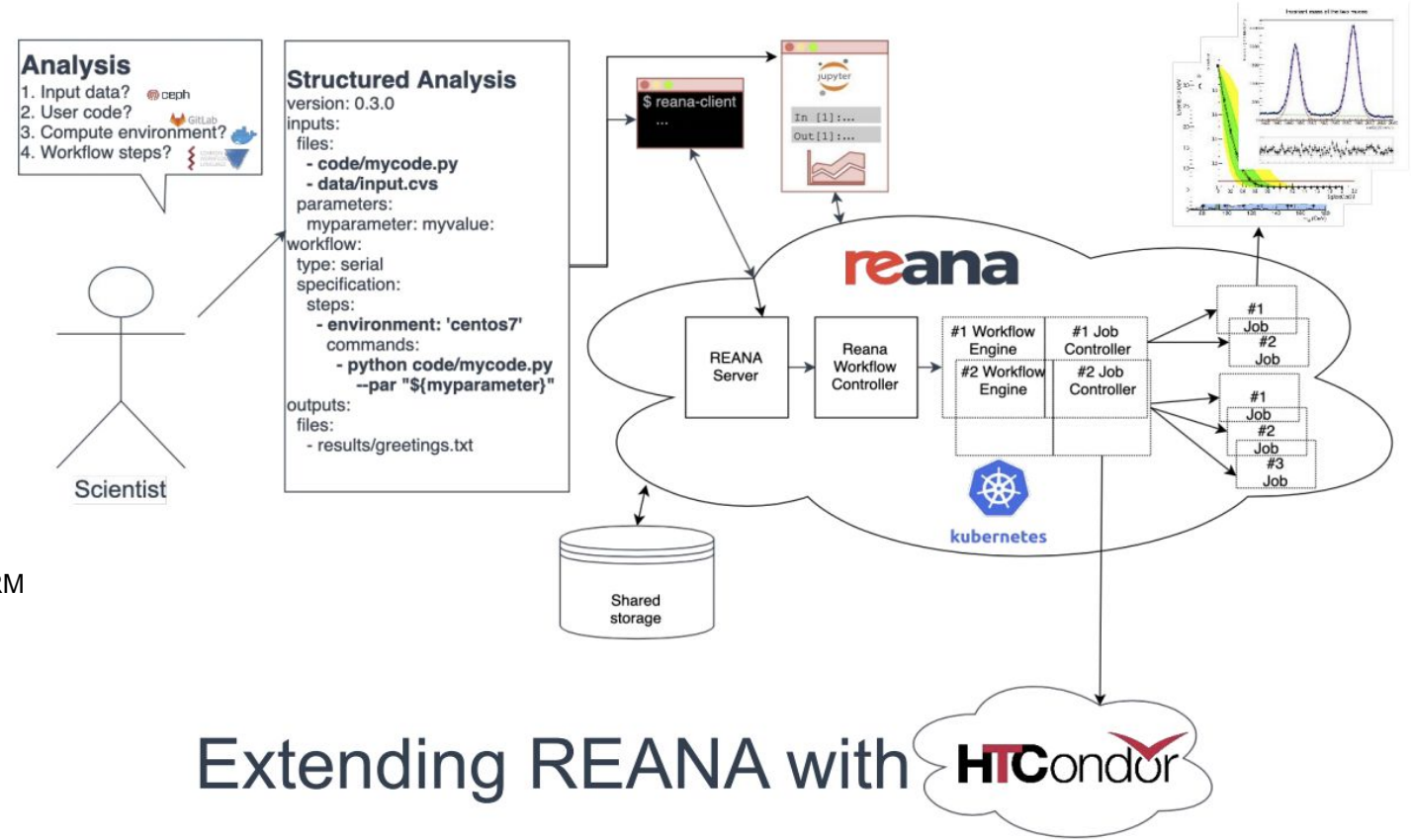
The REANA Project
https://reanahub.io/

Life Demo
ATLAS BSM Search



Access to CERN reana cloud currently limited to CERN network

REANA architecture
https://reanahub.io/



Supported workflow engines:

- CWL
  Common Workflow Language
  https://www.commonwl.org/
- Yadage
  https://github.com/yadage
- Serial (simple generic)
- Snakemake under discussion

Supported compute backends:

- HTCondor, Kubernetes, SLURM

Supported storage services:

- GitLab, CVMFS, EOS, NFS

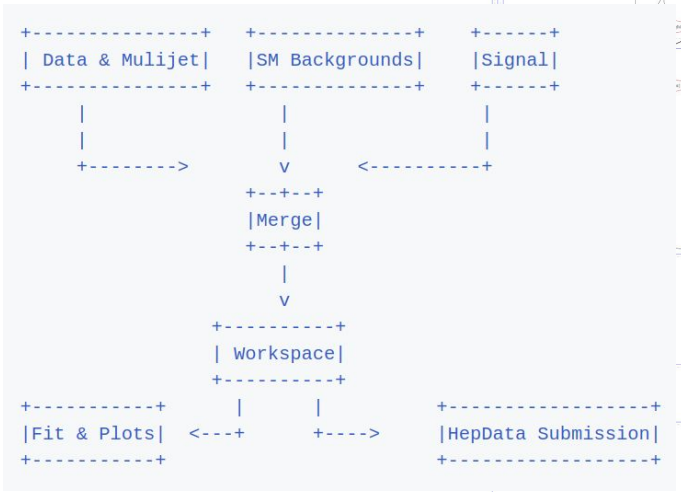
Tibor Simko, poster presented at eScience IEEE 2019

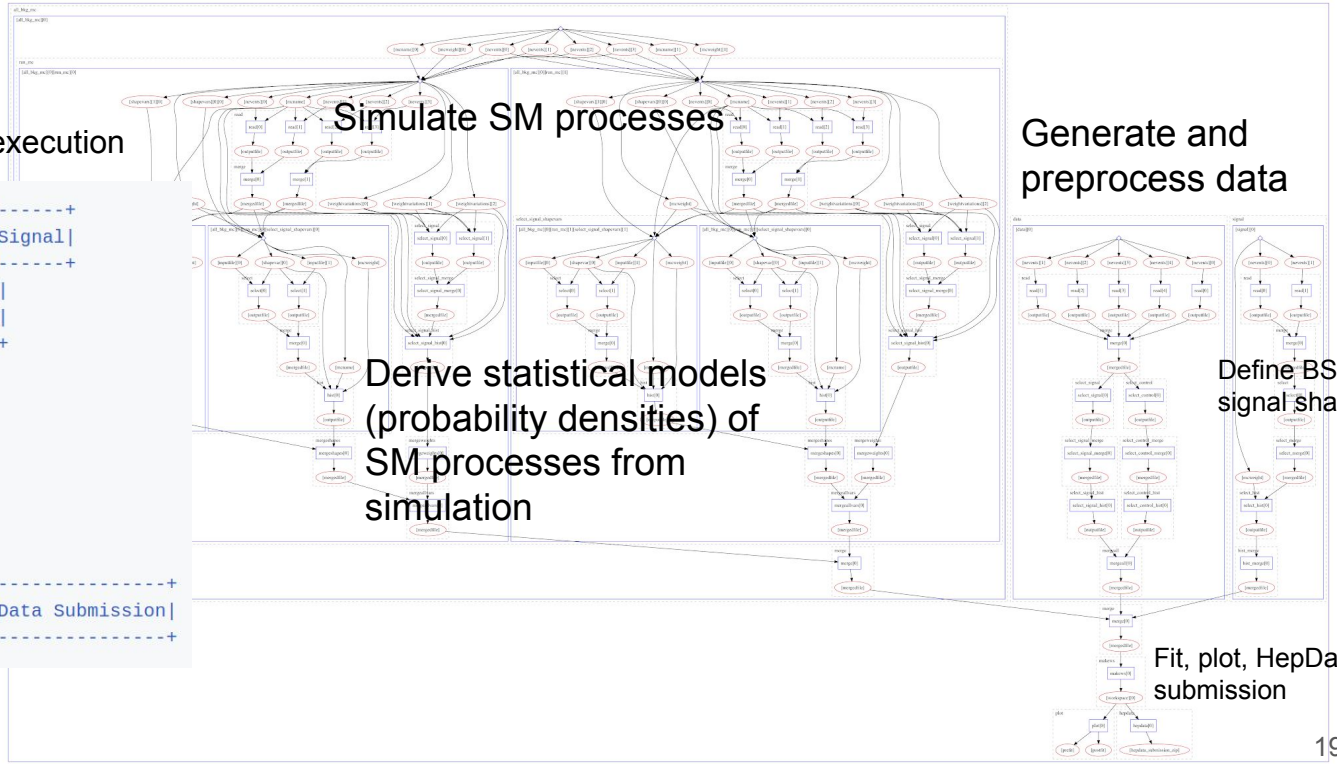
Example Analysis
Search for physics beyond the standard model at ATLAS
https://github.com/reanahub/reana-demo-bsm-search
Input data is simulated from statistical models
as part of workflow here

Workflow described by a directed acyclic graph (DAG) using the **yadage** engine

- automatic parallelisation of workflow graph

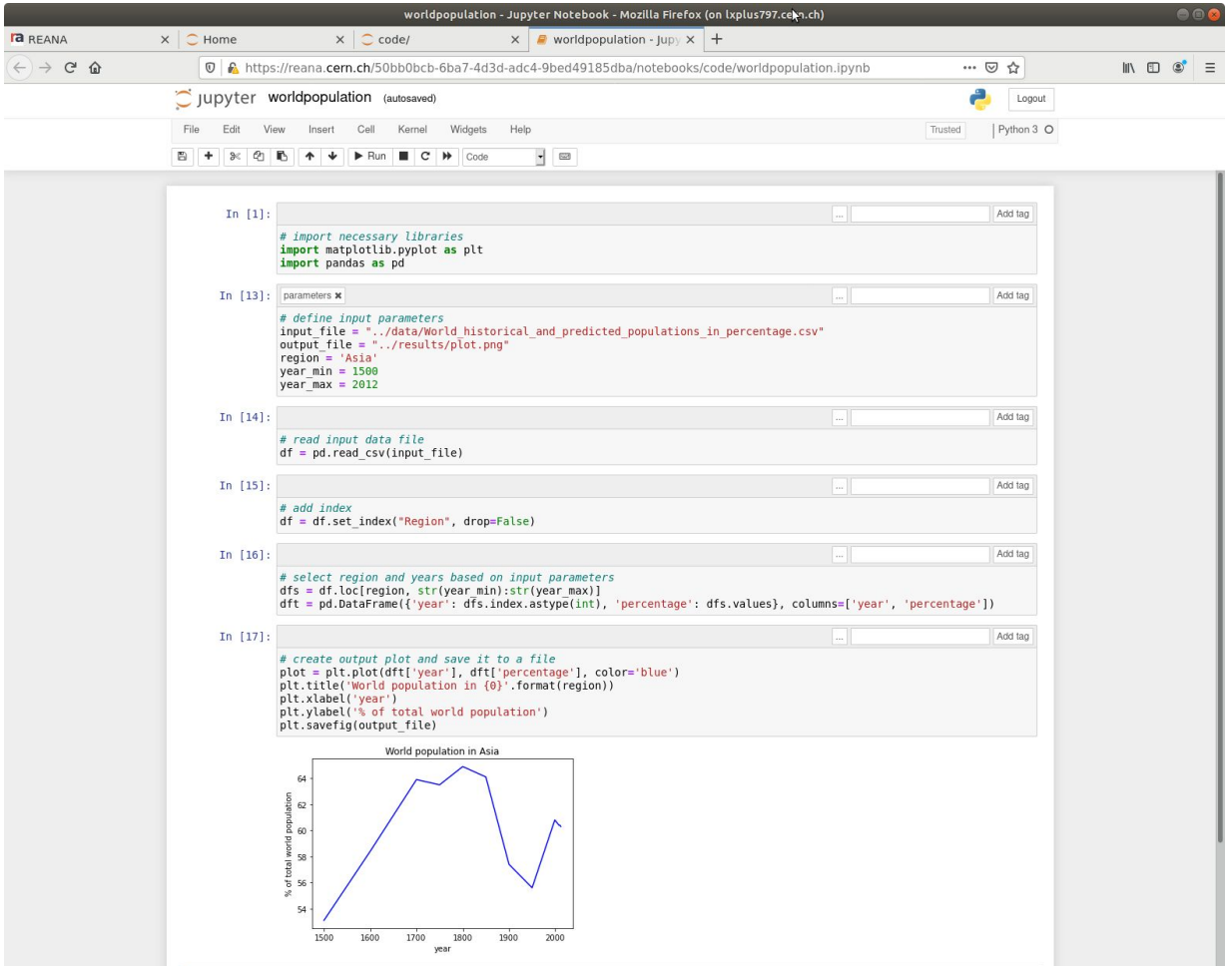
High level workflow
broken down into a large number
of intermediary steps => parallel execution





REANA interactive session
https://reanahub.io/

- REANA manages projects in so called “workspaces” (basically a directory)
- Data and code are uploaded into the workspace
- Workspaces can be examined interactively using Jupyter Notebooks



Current Paradigm

Publications
may be supported by

- open data sets (various repositories)
- code, workflows (github)

High level results available
in machine readable form

PUNCH4NFDI development

Data catalogues (e.g VO)

Items loosely linked

- hyperlinks in project descriptions / abstracts
- “papers with code” links (arXiv)
- Inspired by how we reference papers
- no standard way of linking / interfacing

Executing computational elements left to users

DSP and the role of the Research Product (RP)

RPs provide a common concept to talk about the output of research. An RP can be a

- paper
- (meta)data set
- scientific code
- workflow
- computational element

Building a catalogue of RPs establishes a common standard to link RPs

- **papers, data and code as first class citizens**

RPs facilitate composition => reuse previous work to create new scientific value

RPs can be active computational elements (Dynamic RPs)

- Workflows + execution environment
- Statistical analysis, combination with other results, ...

PUNCH Science Catalogue

- Collect available material from across PUNCH community in a common, searchable catalogue
  - publications, code, datasets/sources, existing catalogues
  - Common interface to **establish links**

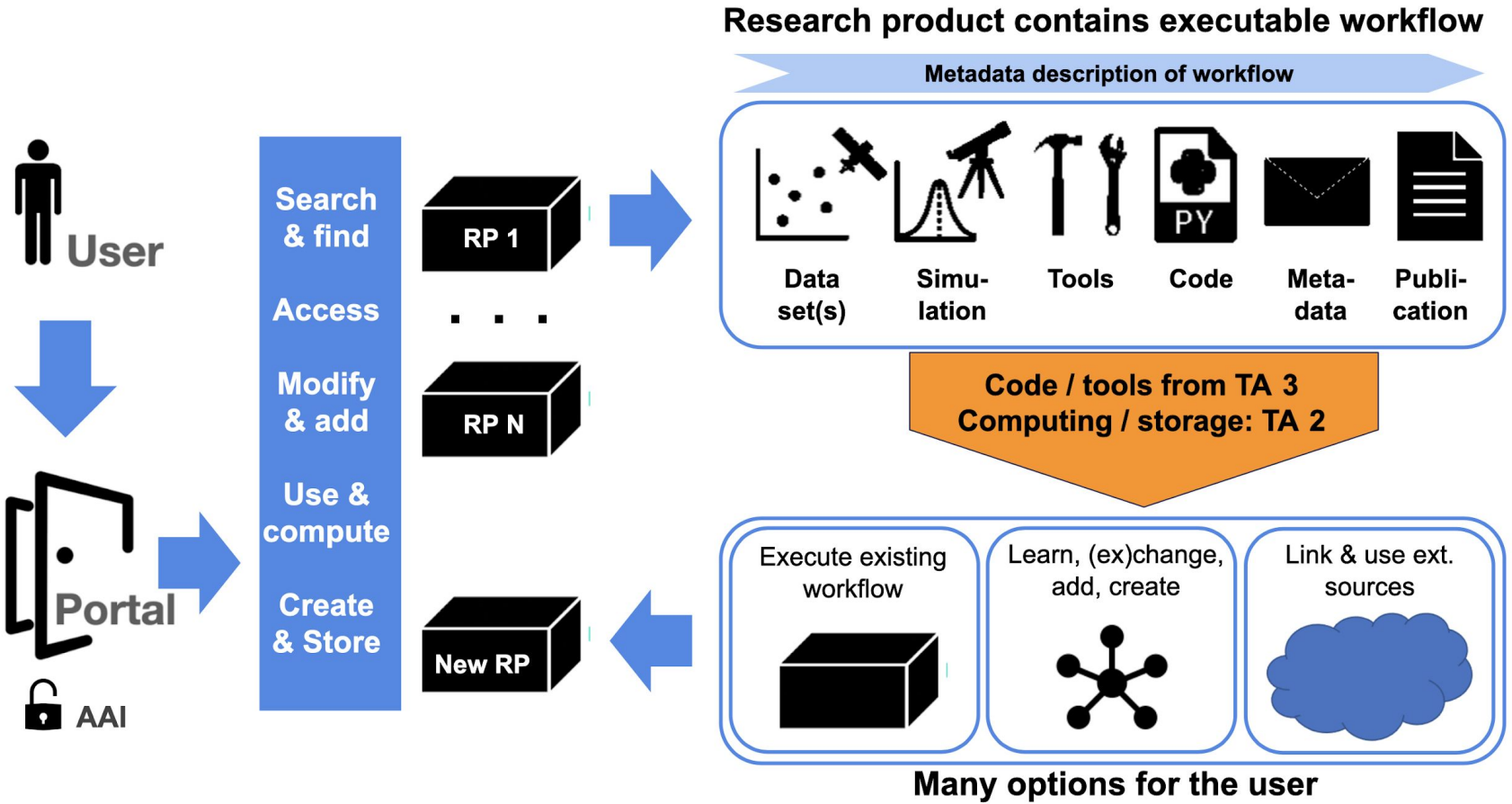
- Catalogue entries are the **research products** (RPs)

- Use these collections to develop meta-data schemes, describing the RPs

  - building on existing developments in particular in astrophysics
  - establish schemata to **allow composition of RPs**

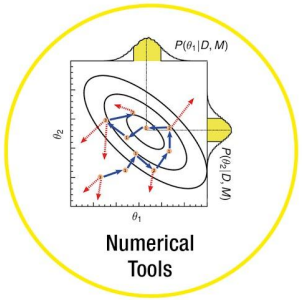
- **Portal Entry point** for user searches

Platform SDP



The Use Cases: How and where our work might improve

https://www.punch4nfdi.de



PUNCH 4NFDI

if we ontologically know
what this is: easily combine
with a separate workflow
→ HUGE benefit from

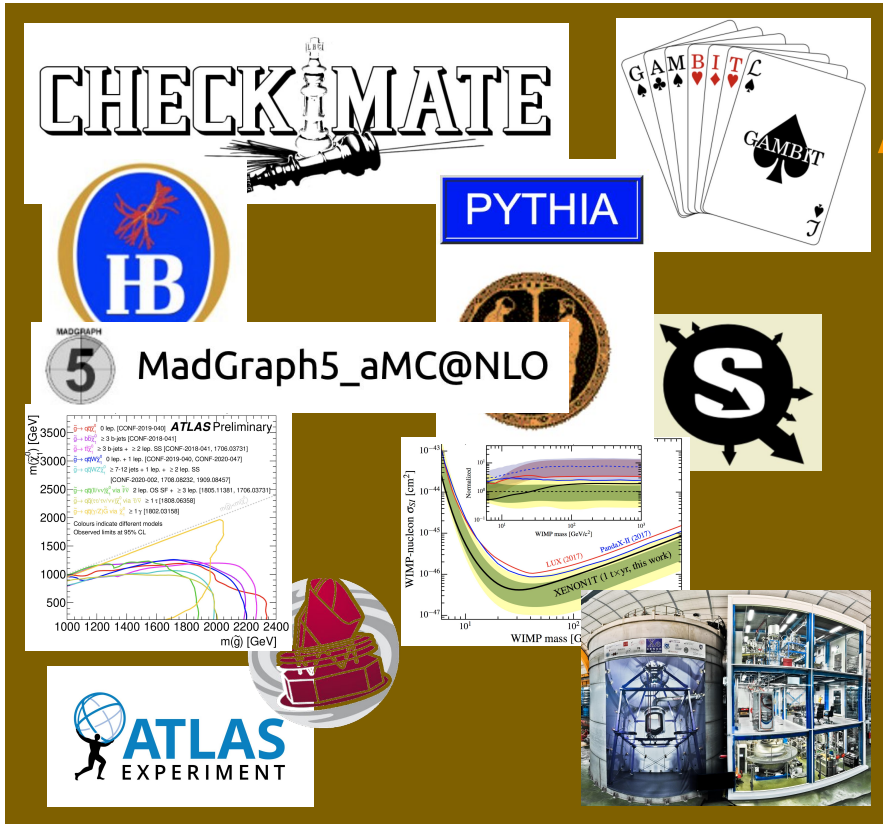
store as new DRP

result
from e.g.
cosmological
analysis of Ω_k^2

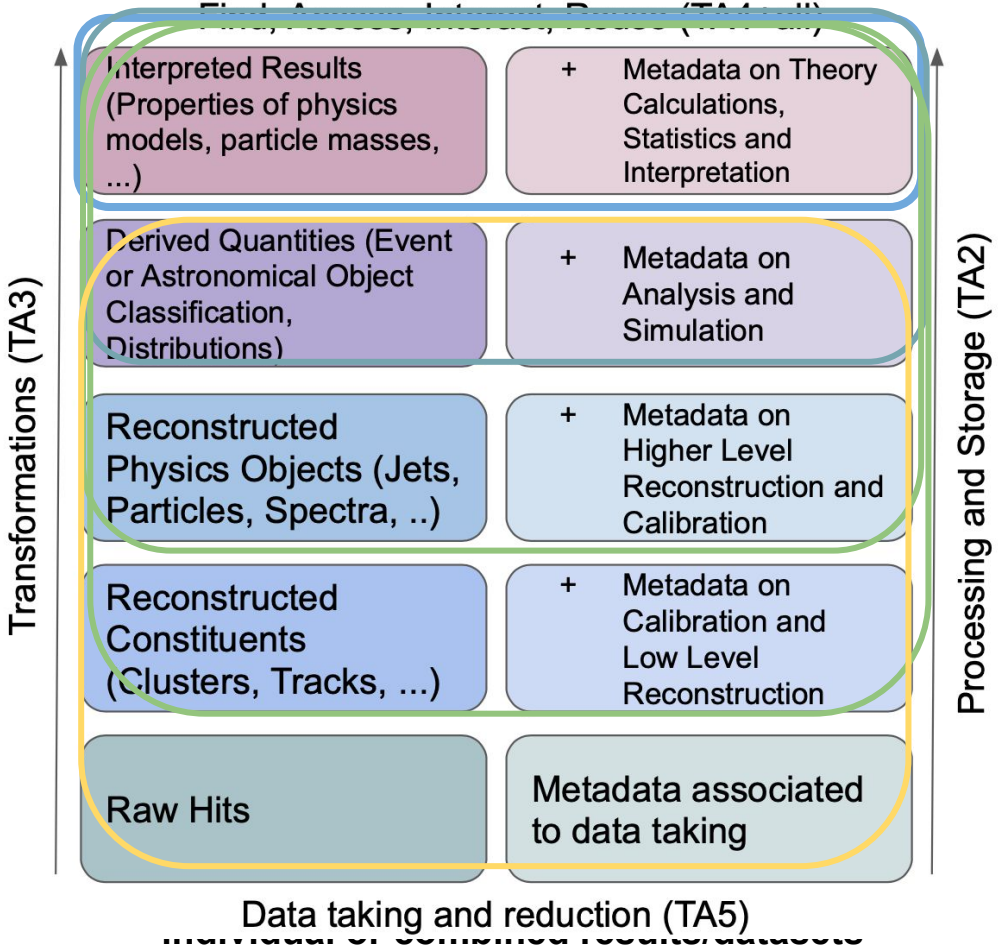
new statistical
analysis
(tools to test and combine
likelihoods easily available)

result
from e.g.
LUX

The Use Cases: An example from a global analysis



Now: Every analyst must download, compile, interface every code, read results from tables, ..

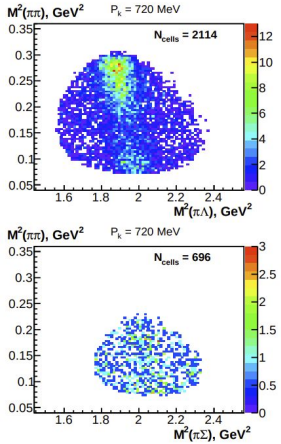


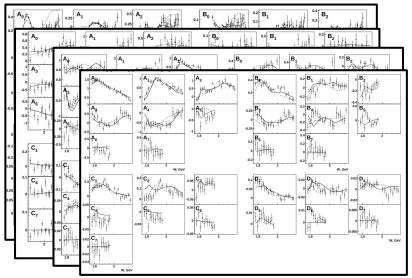
The Use Cases: An example from a hadron spectroscopy analysis

Hyperons coupled channel analysis [EPJ A55(2019)180] combination of data from several experiments JLab, CERN,... "by hand"

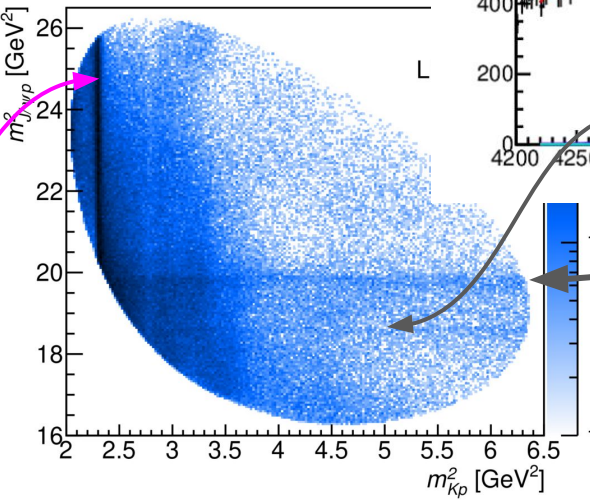
Pentaquarks @ LHCb [PRL122(2019)222001]

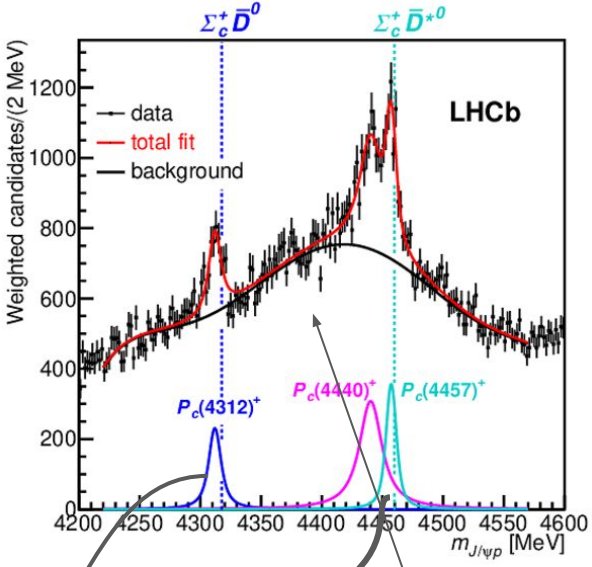
Combined analysis scattering data + B-decays could constrain background from hyperons. Need infrastructure to combine datasets and fitters





$$K^- p \to \pi^0 \Lambda(1520),$$
$$K^- p \to K^- \Delta^+(1232),$$
$$K^- p \to \pi^\mp \Sigma^\pm(1385),$$
$$K^- p \to \bar{K}^* N.$$



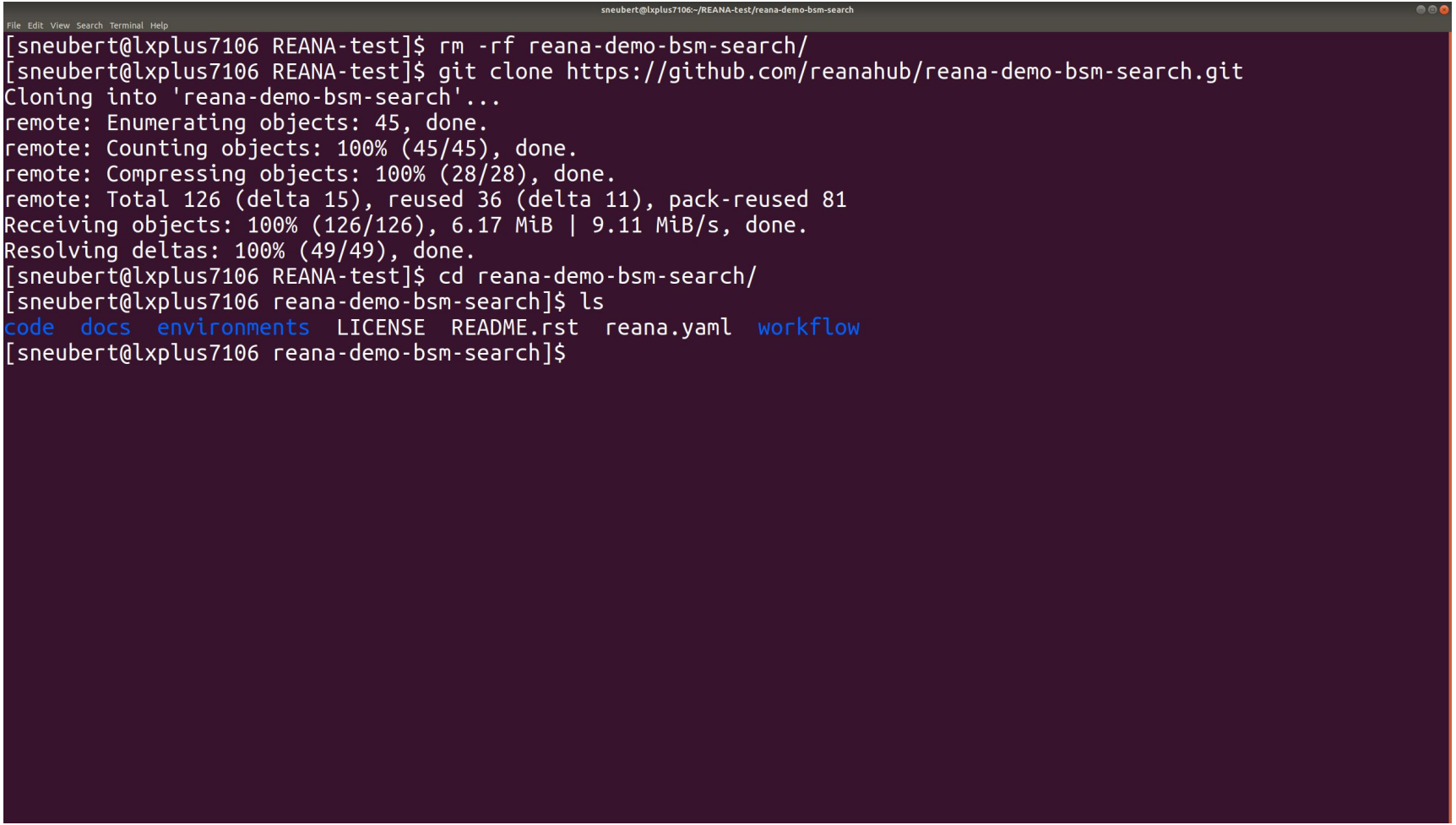


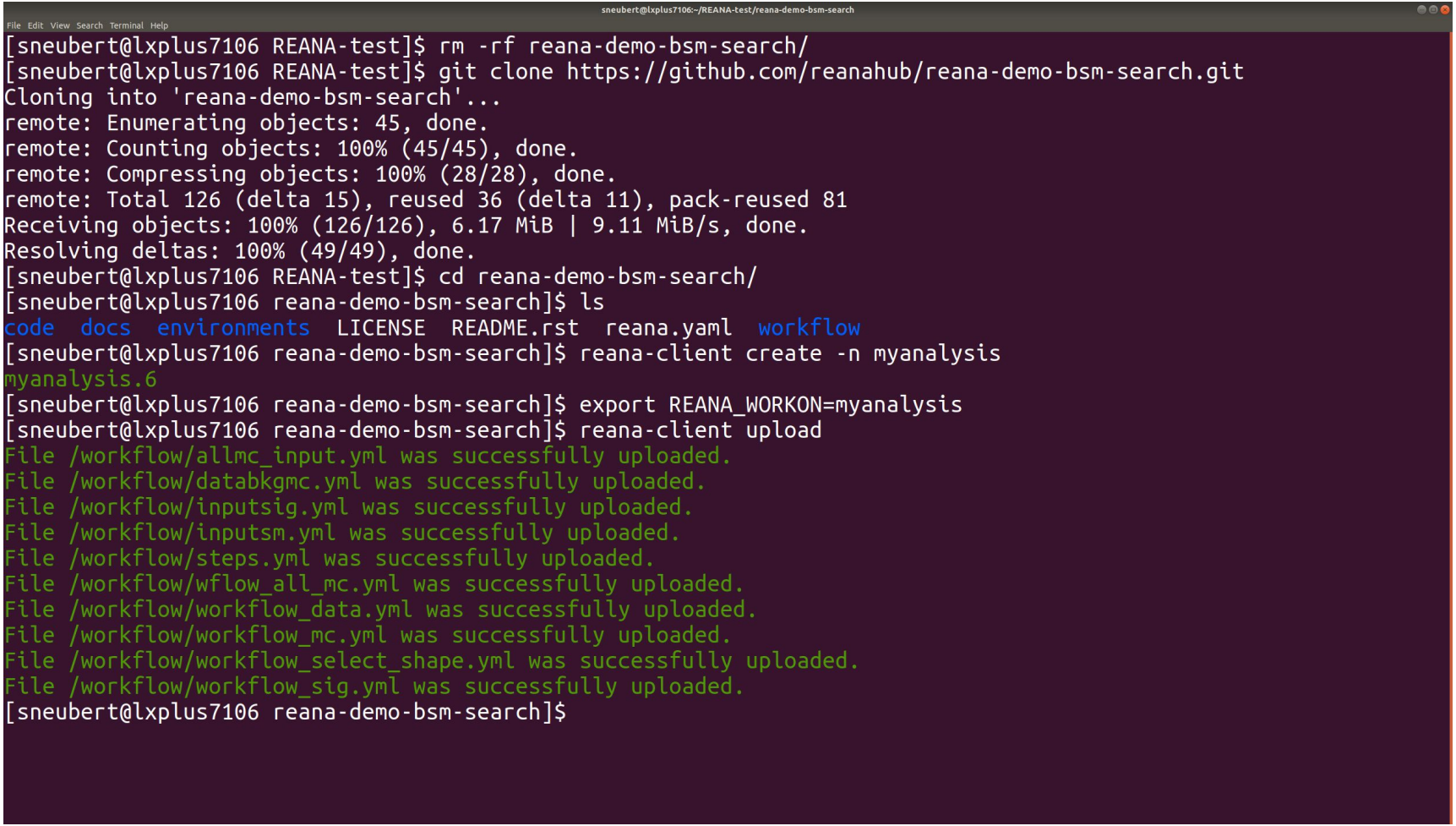
Is there another broad resonance here?

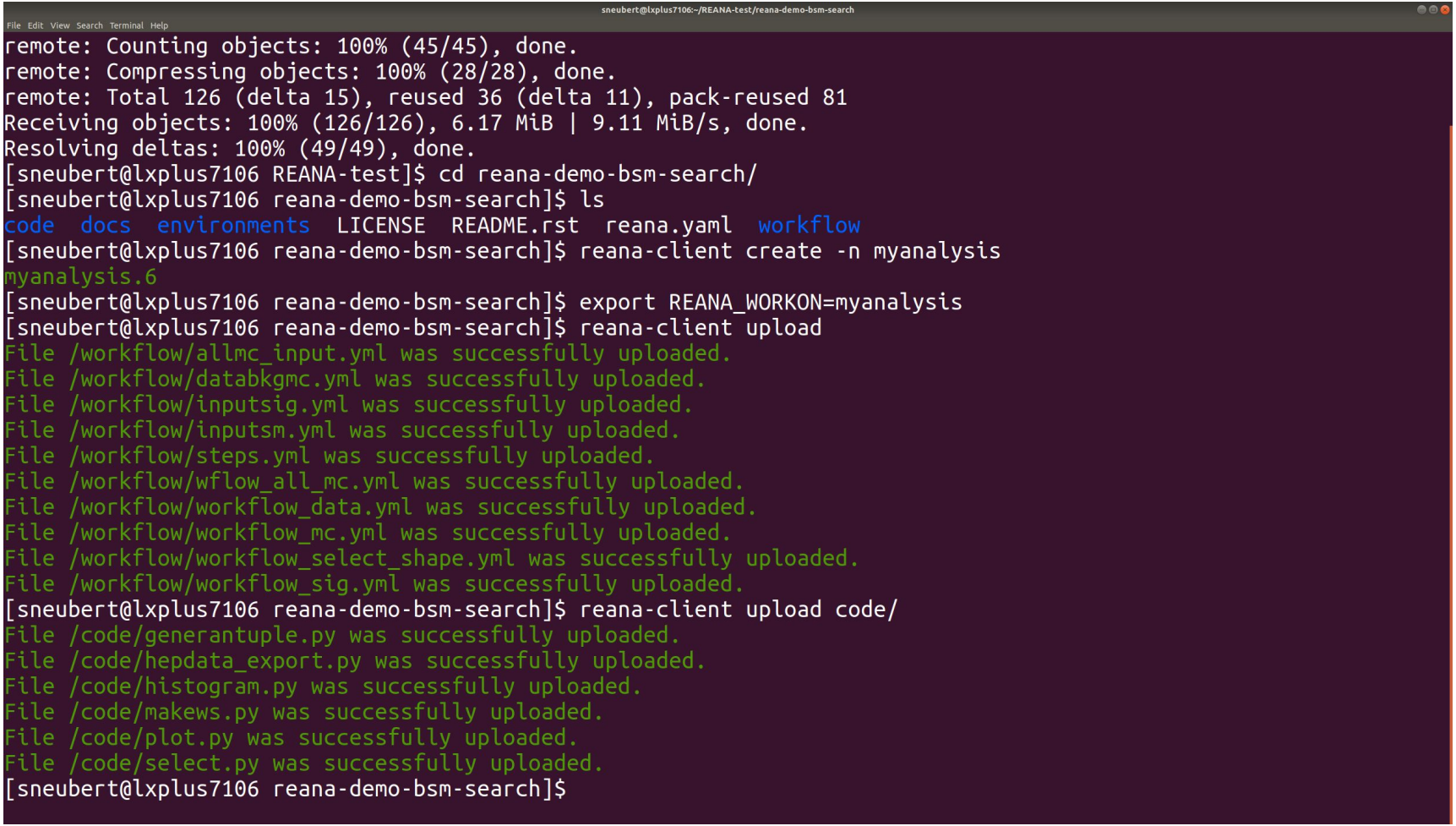
Summary

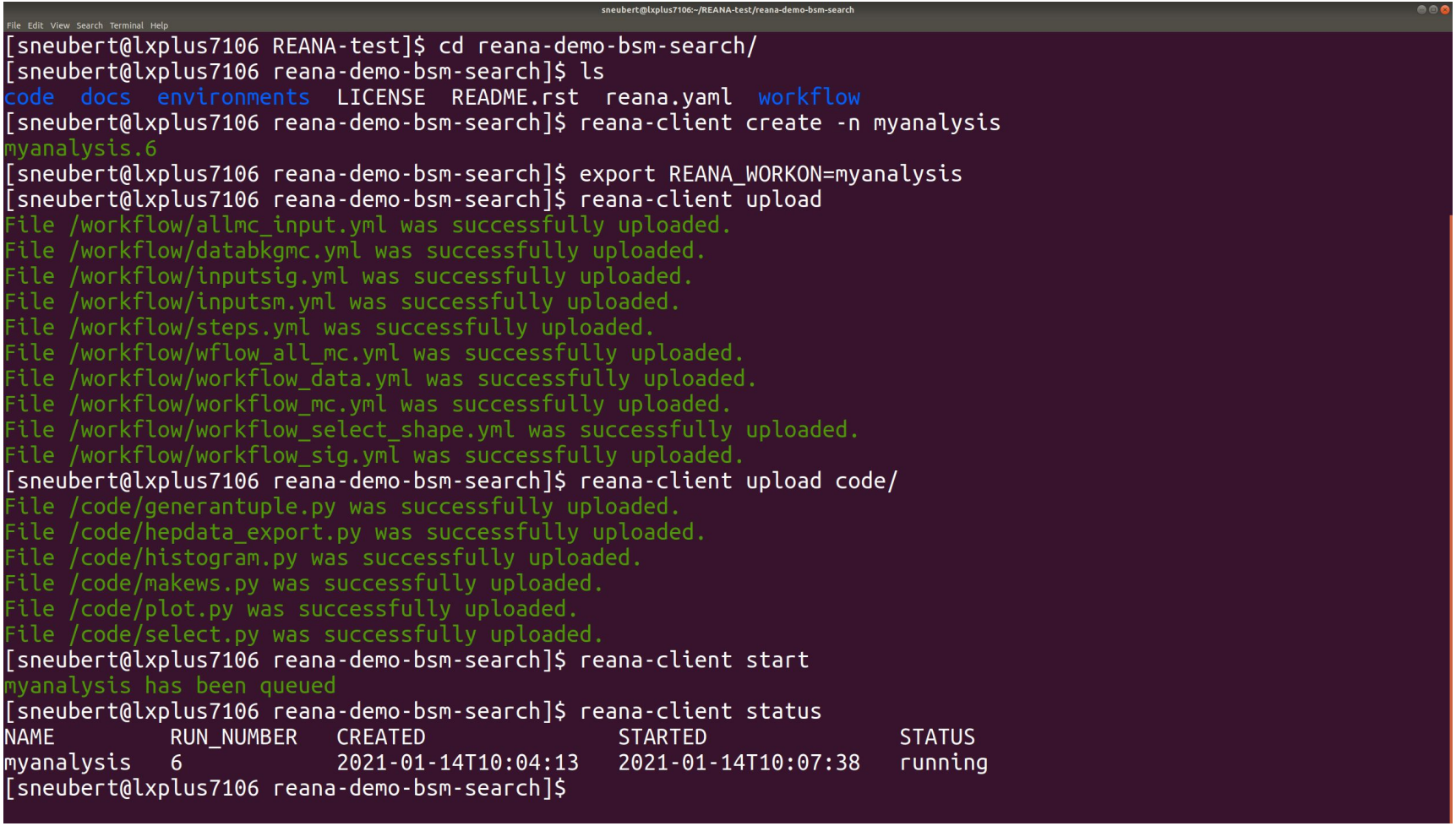
- The **PUNCH Science Data Platform SDP** is an important part of the PUNCH4NFDI consortium's plans for advancing the computing of the PUNCH sciences and the whole NFDI
- It builds on the collaboration with other platform developments, and on the improvement of platform use through a continuation of the developments of **techniques available in our communities**
- The **Research Product** is a central part of the PUNCH-SDP: It can add an important representation of *knowledge in the interaction and integration* of *data, metadata, simulation, tools and code*, which is especially strong in our communities due to the one-off nature of our experiments: Big Data and unique one-off experiments place common challenges for (Astro)Particle-, Hadron-Physics and Astronomy

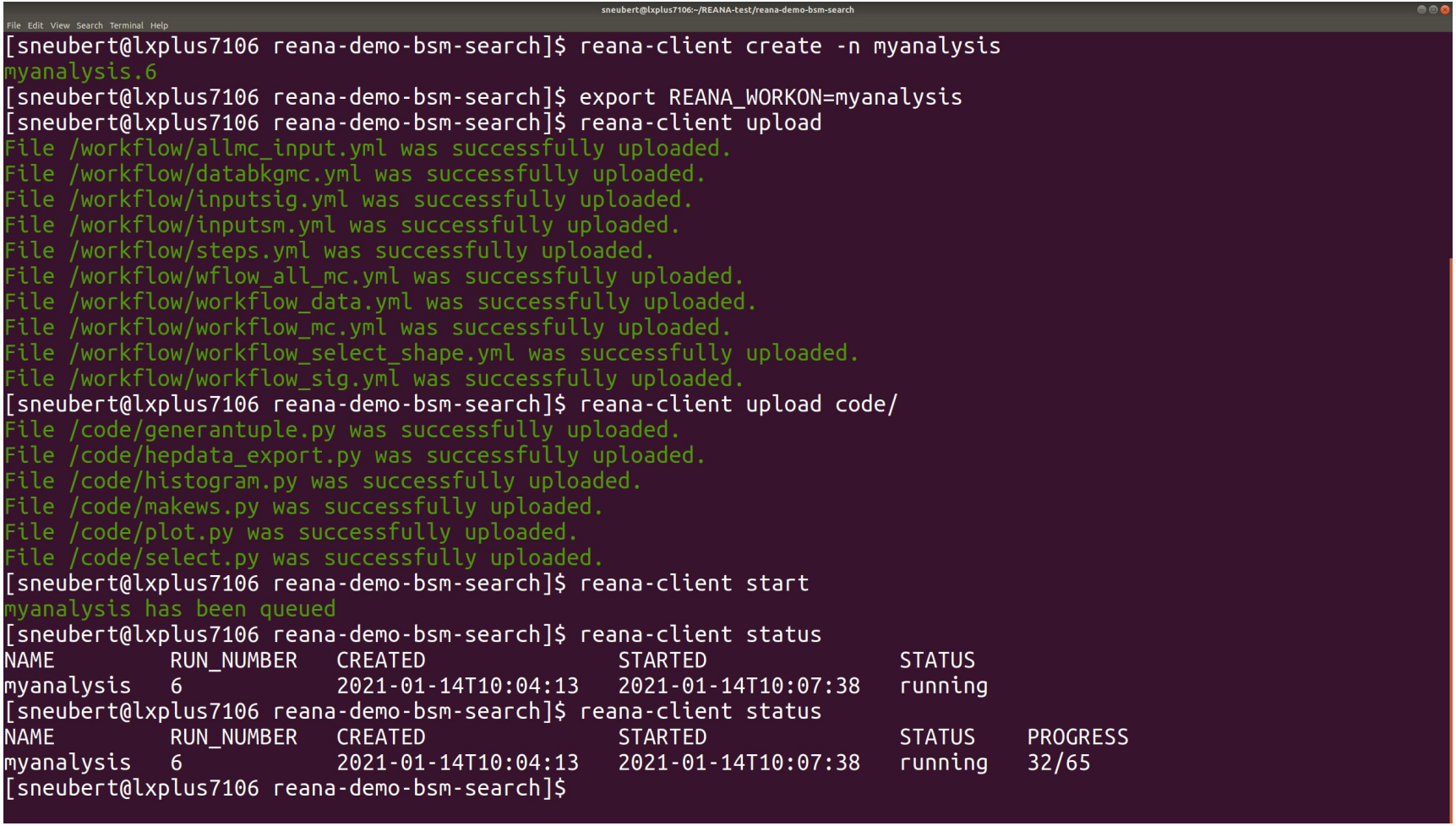
Backup

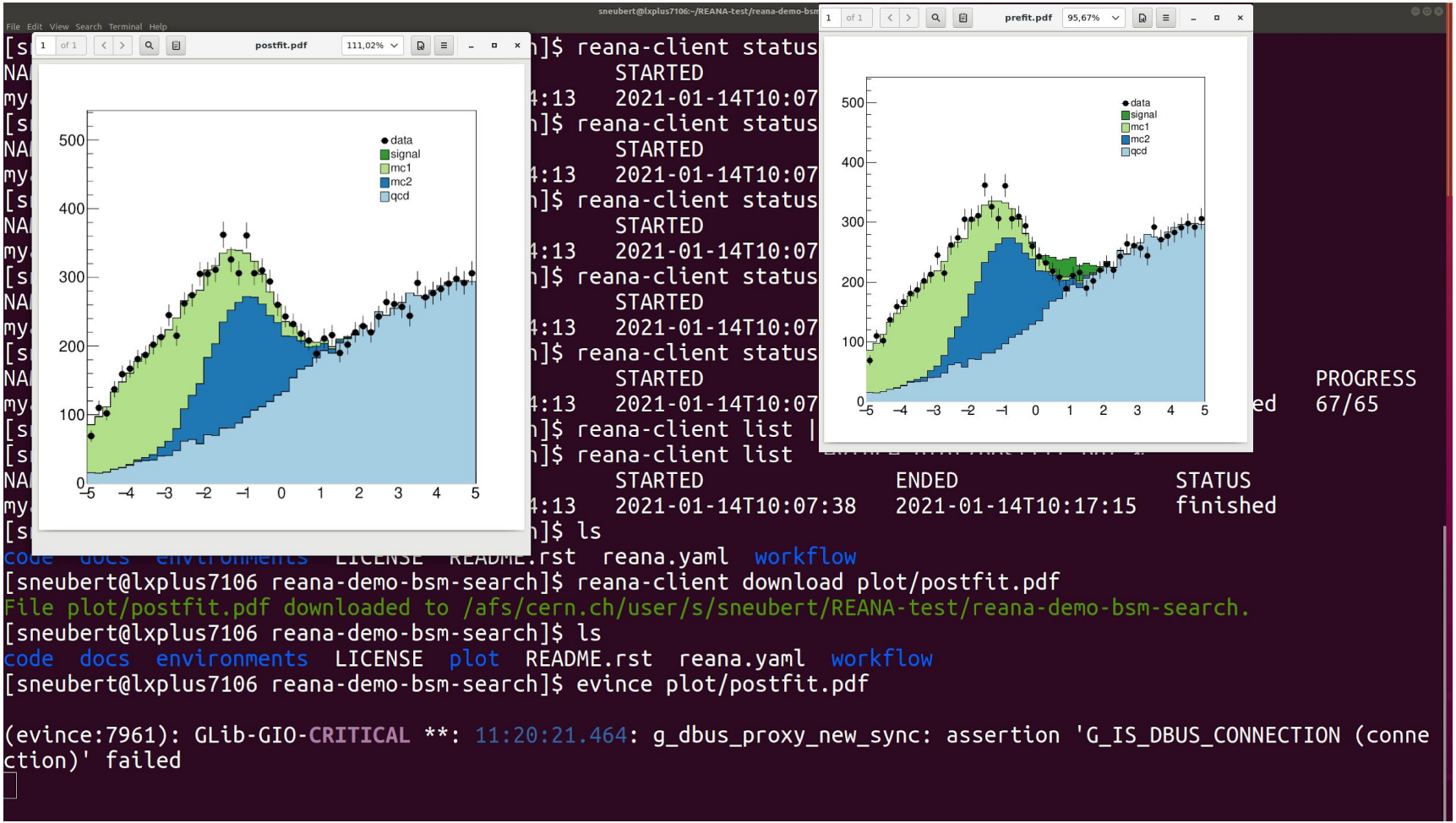

sneubert@lxplus7106:~/REANA-test/reana-demo-bsm-search
File Edit View Search Terminal Help
[sneubert@lxplus7106 REANA-test]$ rm -rf reana-demo-bsm-search/
[sneubert@lxplus7106 REANA-test]$ git clone https://github.com/reanahub/reana-demo-bsm-search.git
Cloning into 'reana-demo-bsm-search'...
remote: Enumerating objects: 45, done.
remote: Counting objects: 100% (45/45), done.
remote: Compressing objects: 100% (28/28), done.
remote: Total 126 (delta 15), reused 36 (delta 11), pack-reused 81
Receiving objects: 100% (126/126), 6.17 MiB | 9.11 MiB/s, done.
Resolving deltas: 100% (49/49), done.
[sneubert@lxplus7106 REANA-test]$ cd reana-demo-bsm-search/
[sneubert@lxplus7106 reana-demo-bsm-search]$ ls
code docs environments LICENSE README.rst reana.yaml workflow
[sneubert@lxplus7106 reana-demo-bsm-search]$


sneubert@lxplus7106:~/REANA-test/reana-demo-bsm-search
File Edit View Search Terminal Help
[sneubert@lxplus7106 REANA-test]$ rm -rf reana-demo-bsm-search/
[sneubert@lxplus7106 REANA-test]$ git clone https://github.com/reanahub/reana-demo-bsm-search.git
Cloning into 'reana-demo-bsm-search'...
remote: Enumerating objects: 45, done.
remote: Counting objects: 100% (45/45), done.
remote: Compressing objects: 100% (28/28), done.
remote: Total 126 (delta 15), reused 36 (delta 11), pack-reused 81
Receiving objects: 100% (126/126), 6.17 MiB | 9.11 MiB/s, done.
Resolving deltas: 100% (49/49), done.
[sneubert@lxplus7106 REANA-test]$ cd reana-demo-bsm-search/
[sneubert@lxplus7106 reana-demo-bsm-search]$ ls
code docs environments LICENSE README.rst reana.yaml workflow
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client create -n myanalysis
myanalysis.6
[sneubert@lxplus7106 reana-demo-bsm-search]$ export REANA_WORKON=myanalysis
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client upload
File /workflow/allmc_input.yml was successfully uploaded.
File /workflow/databkgmc.yml was successfully uploaded.
File /workflow/inputsig.yml was successfully uploaded.
File /workflow/inputsm.yml was successfully uploaded.
File /workflow/steps.yml was successfully uploaded.
File /workflow/wflow_all_mc.yml was successfully uploaded.
File /workflow/workflow_data.yml was successfully uploaded.
File /workflow/workflow_mc.yml was successfully uploaded.
File /workflow/workflow_select_shape.yml was successfully uploaded.
File /workflow/workflow_sig.yml was successfully uploaded.
[sneubert@lxplus7106 reana-demo-bsm-search]$


sneubert@lxplus7106:~/REANA-test/reana-demo-bsm-search
File Edit View Search Terminal Help
remote: Counting objects: 100% (45/45), done.
remote: Compressing objects: 100% (28/28), done.
remote: Total 126 (delta 15), reused 36 (delta 11), pack-reused 81
Receiving objects: 100% (126/126), 6.17 MiB | 9.11 MiB/s, done.
Resolving deltas: 100% (49/49), done.
[sneubert@lxplus7106 REANA-test]$ cd reana-demo-bsm-search/
[sneubert@lxplus7106 reana-demo-bsm-search]$ ls
code docs environments LICENSE README.rst reana.yaml workflow
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client create -n myanalysis
myanalysis.6
[sneubert@lxplus7106 reana-demo-bsm-search]$ export REANA_WORKON=myanalysis
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client upload
File /workflow/allmc_input.yml was successfully uploaded.
File /workflow/databkgmc.yml was successfully uploaded.
File /workflow/inputsig.yml was successfully uploaded.
File /workflow/inputsm.yml was successfully uploaded.
File /workflow/steps.yml was successfully uploaded.
File /workflow/wflow_all_mc.yml was successfully uploaded.
File /workflow/workflow_data.yml was successfully uploaded.
File /workflow/workflow_mc.yml was successfully uploaded.
File /workflow/workflow_select_shape.yml was successfully uploaded.
File /workflow/workflow_sig.yml was successfully uploaded.
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client upload code/
File /code/generantuple.py was successfully uploaded.
File /code/hepdata_export.py was successfully uploaded.
File /code/histogram.py was successfully uploaded.
File /code/makews.py was successfully uploaded.
File /code/plot.py was successfully uploaded.
File /code/select.py was successfully uploaded.
[sneubert@lxplus7106 reana-demo-bsm-search]$


sneubert@lxplus7106:~/REANA-test/reana-demo-bsm-search
File Edit View Search Terminal Help
[sneubert@lxplus7106 REANA-test]$ cd reana-demo-bsm-search/
[sneubert@lxplus7106 reana-demo-bsm-search]$ ls
code docs environments LICENSE README.rst reana.yaml workflow
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client create -n myanalysis
myanalysis.6
[sneubert@lxplus7106 reana-demo-bsm-search]$ export REANA_WORKON=myanalysis
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client upload
File /workflow/allmc_input.yml was successfully uploaded.
File /workflow/databkgmc.yml was successfully uploaded.
File /workflow/inputsig.yml was successfully uploaded.
File /workflow/inputsm.yml was successfully uploaded.
File /workflow/steps.yml was successfully uploaded.
File /workflow/wflow_all_mc.yml was successfully uploaded.
File /workflow/workflow_data.yml was successfully uploaded.
File /workflow/workflow_mc.yml was successfully uploaded.
File /workflow/workflow_select_shape.yml was successfully uploaded.
File /workflow/workflow_sig.yml was successfully uploaded.
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client upload code/
File /code/generantuple.py was successfully uploaded.
File /code/hepdata_export.py was successfully uploaded.
File /code/histogram.py was successfully uploaded.
File /code/makews.py was successfully uploaded.
File /code/plot.py was successfully uploaded.
File /code/select.py was successfully uploaded.
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client start
myanalysis has been queued
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client status
NAME RUN_NUMBER CREATED STARTED STATUS
myanalysis 6 2021-01-14T10:04:13 2021-01-14T10:07:38 running
[sneubert@lxplus7106 reana-demo-bsm-search]$


sneubert@lxplus7106:~/REANA-test/reana-demo-bsm-search
File Edit View Search Terminal Help
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client create -n myanalysis
myanalysis.6
[sneubert@lxplus7106 reana-demo-bsm-search]$ export REANA_WORKON=myanalysis
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client upload
File /workflow/allmc_input.yml was successfully uploaded.
File /workflow/databkgmc.yml was successfully uploaded.
File /workflow/inputsig.yml was successfully uploaded.
File /workflow/inputsm.yml was successfully uploaded.
File /workflow/steps.yml was successfully uploaded.
File /workflow/wflow_all_mc.yml was successfully uploaded.
File /workflow/workflow_data.yml was successfully uploaded.
File /workflow/workflow_mc.yml was successfully uploaded.
File /workflow/workflow_select_shape.yml was successfully uploaded.
File /workflow/workflow_sig.yml was successfully uploaded.
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client upload code/
File /code/generantuple.py was successfully uploaded.
File /code/hepdata_export.py was successfully uploaded.
File /code/histogram.py was successfully uploaded.
File /code/makews.py was successfully uploaded.
File /code/plot.py was successfully uploaded.
File /code/select.py was successfully uploaded.
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client start
myanalysis has been queued
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client status
NAME RUN_NUMBER CREATED STARTED STATUS
myanalysis 6 2021-01-14T10:04:13 2021-01-14T10:07:38 running
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client status
NAME RUN_NUMBER CREATED STARTED STATUS PROGRESS
myanalysis 6 2021-01-14T10:04:13 2021-01-14T10:07:38 running 32/65
[sneubert@lxplus7106 reana-demo-bsm-search]$


postfit.pdf
prefit.pdf
data
signal
mc1
mc2
qcd
[sneubert@lxplus7106 reana-demo-bsm-search]$ reana-client download plot/postfit.pdf
File plot/postfit.pdf downloaded to /afs/cern.ch/user/s/sneubert/REANA-test/reana-demo-bsm-search.
[sneubert@lxplus7106 reana-demo-bsm-search]$ ls
code docs environments LICENSE plot README.rst reana.yaml workflow
[sneubert@lxplus7106 reana-demo-bsm-search]$ evince plot/postfit.pdf
(evince:7961): GLib-GIO-CRITICAL **: 11:20:21.464: g_dbus_proxy_new_sync: assertion 'G_IS_DBUS_CONNECTION (connection)' failed