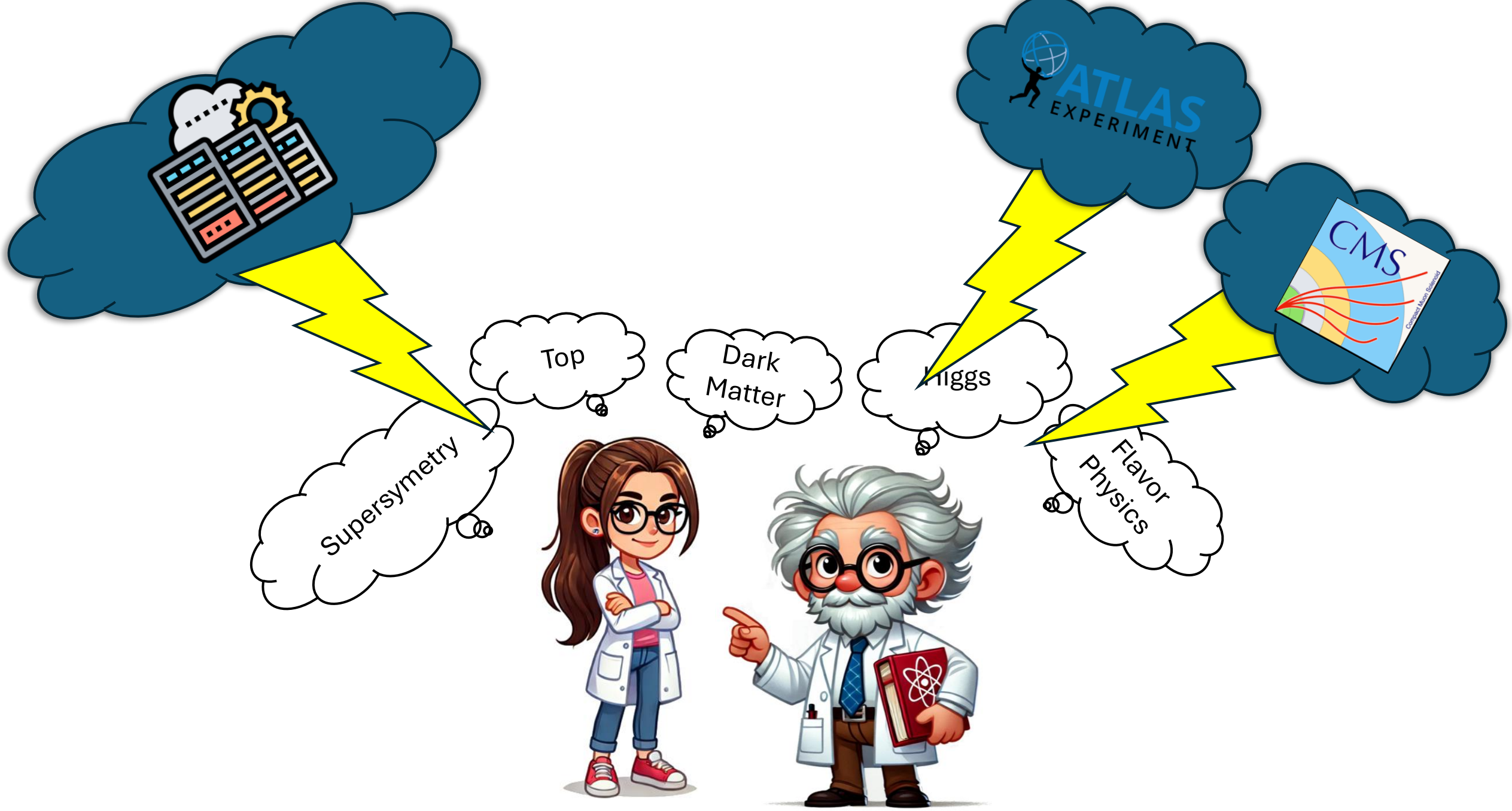
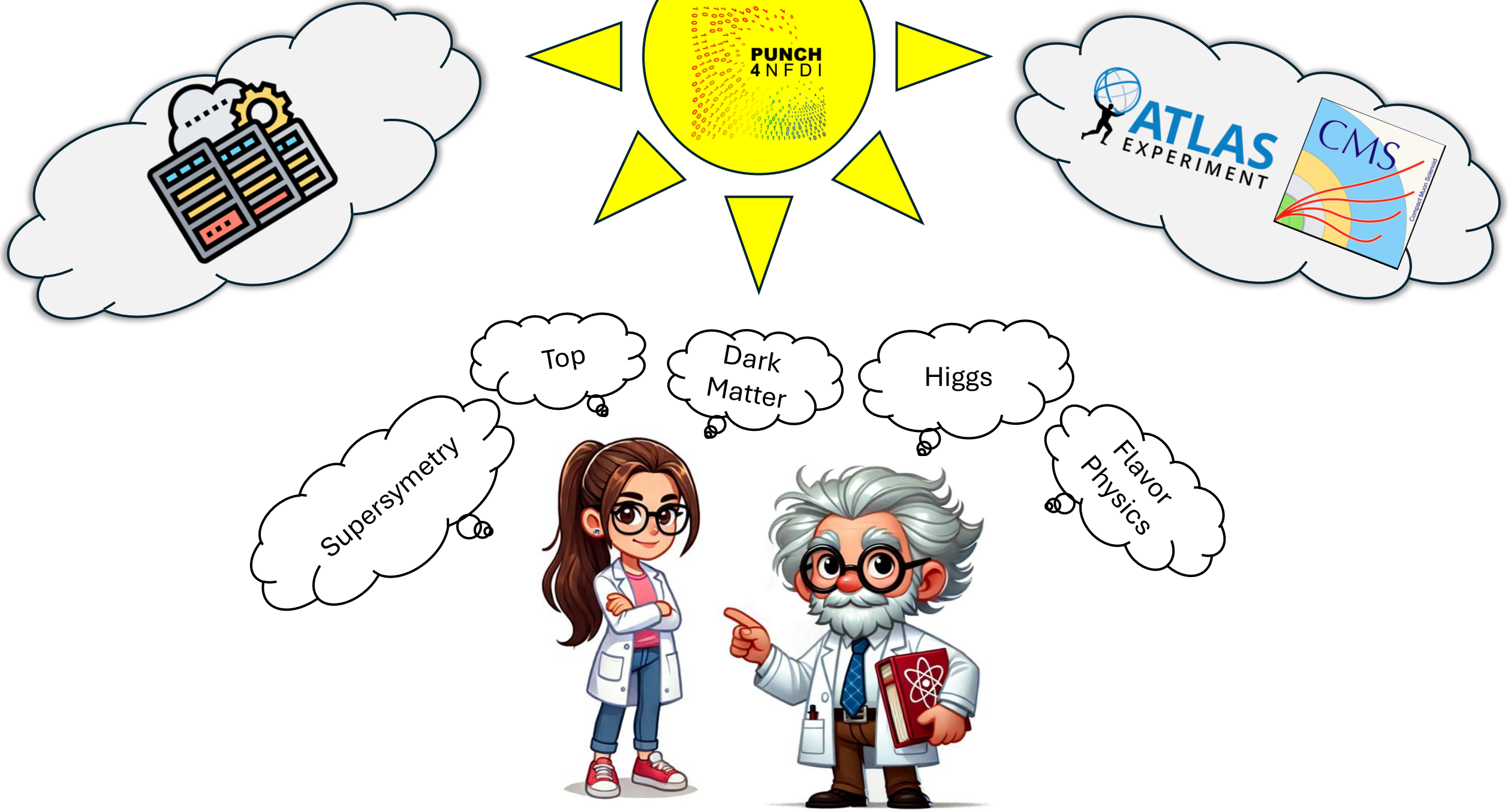
Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Public Multi-Experiment Higgs Analysis Demonstrator

Lucas Karwatzki, Achim Geiser



Created by DALL-E



Created by DALL-E

The Idea

Make science using LHC Open data FAIR

What is the target group?

You:

- Want to do research using LHC data
- Are a researcher or an advanced Student

What do we want to offer?

Transformed, unified data

- Unified approach to ATLAS & CMS data analysis
  - Handle data the same way on an event basis
- Reduced storage requirements
- Minimized runtime for analyses

Computation resources

- Storage
- Processing

Analysis demonstrator

- Showcase feasibility
- Provide entry point

PUNCH4NFDI

What is PUNCH4NFDI?

Particle, **U**niverse, **NuC**lei and **H**adrons
for
National research data infrastructure

The main Goal of PUNCH4NFDI

Essence

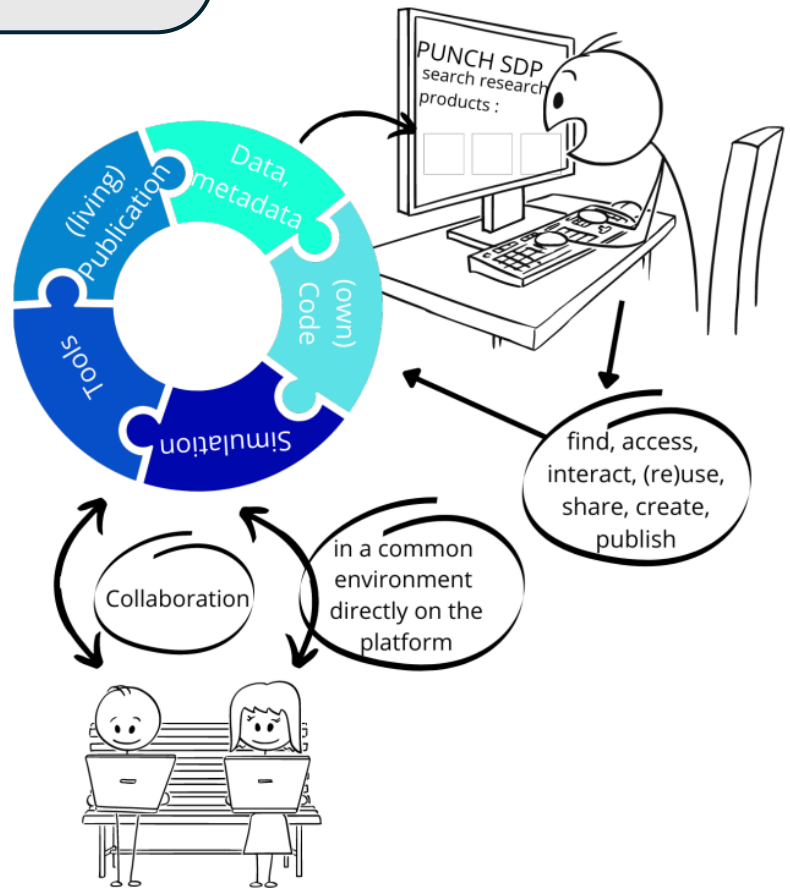
Create a "FAIR" science data platform

- **F**indability
- **A**ccessibility
- **I**nteroperability
- **R**eusable

Offer the infrastructures

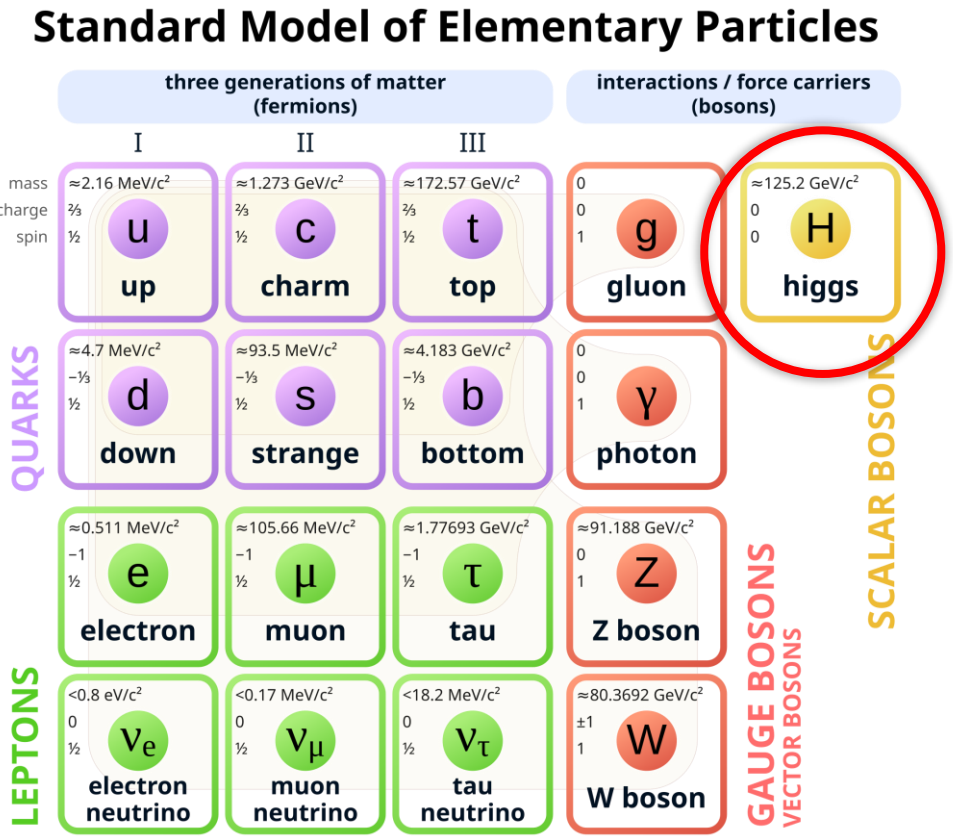
- Computing
- Storage

From and **For** the involved communities and beyond

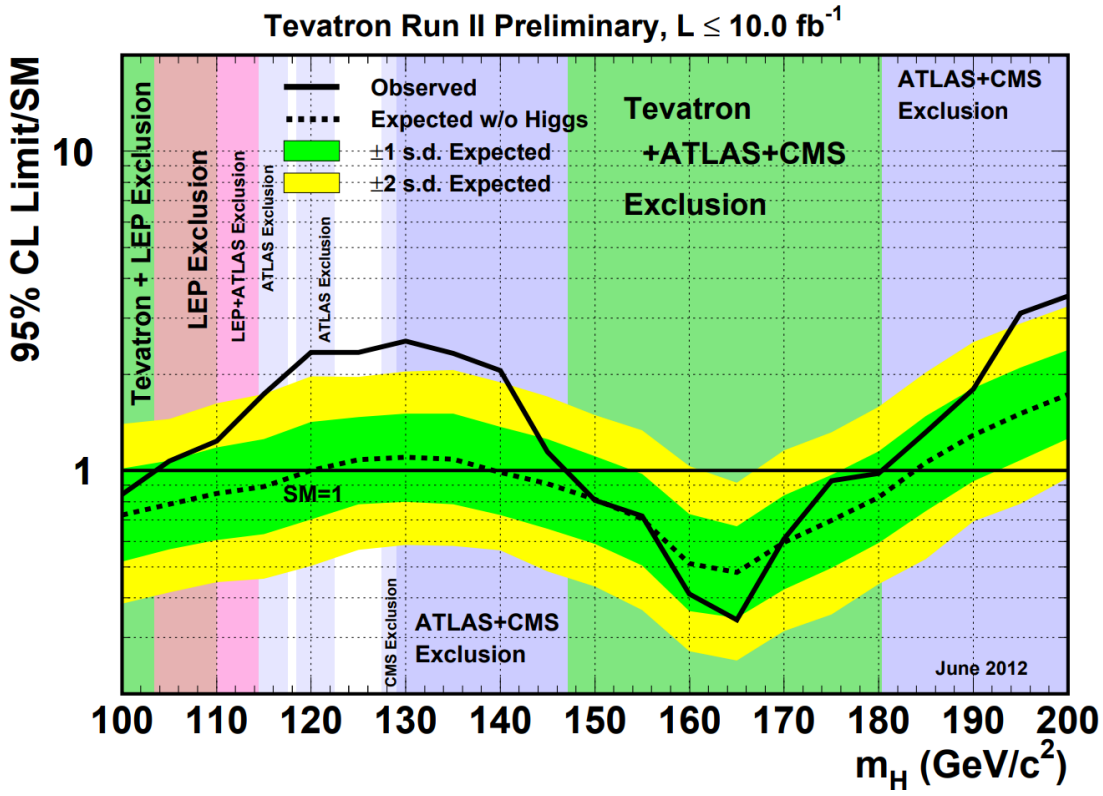


Analysis Demonstrator

What Analysis would we like to showcase?



PBS NOVA



1207.0449

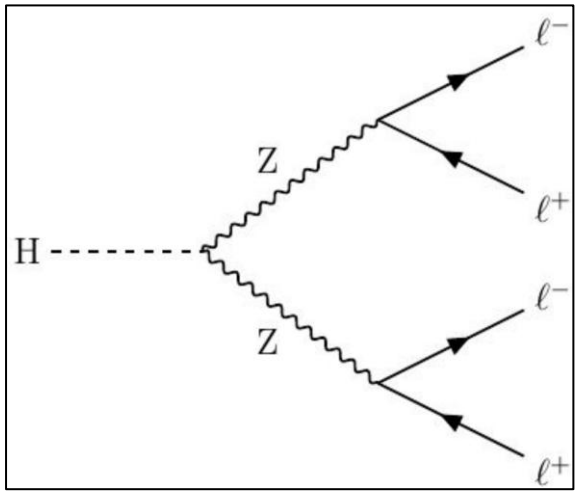
The 4 Lepton Channel

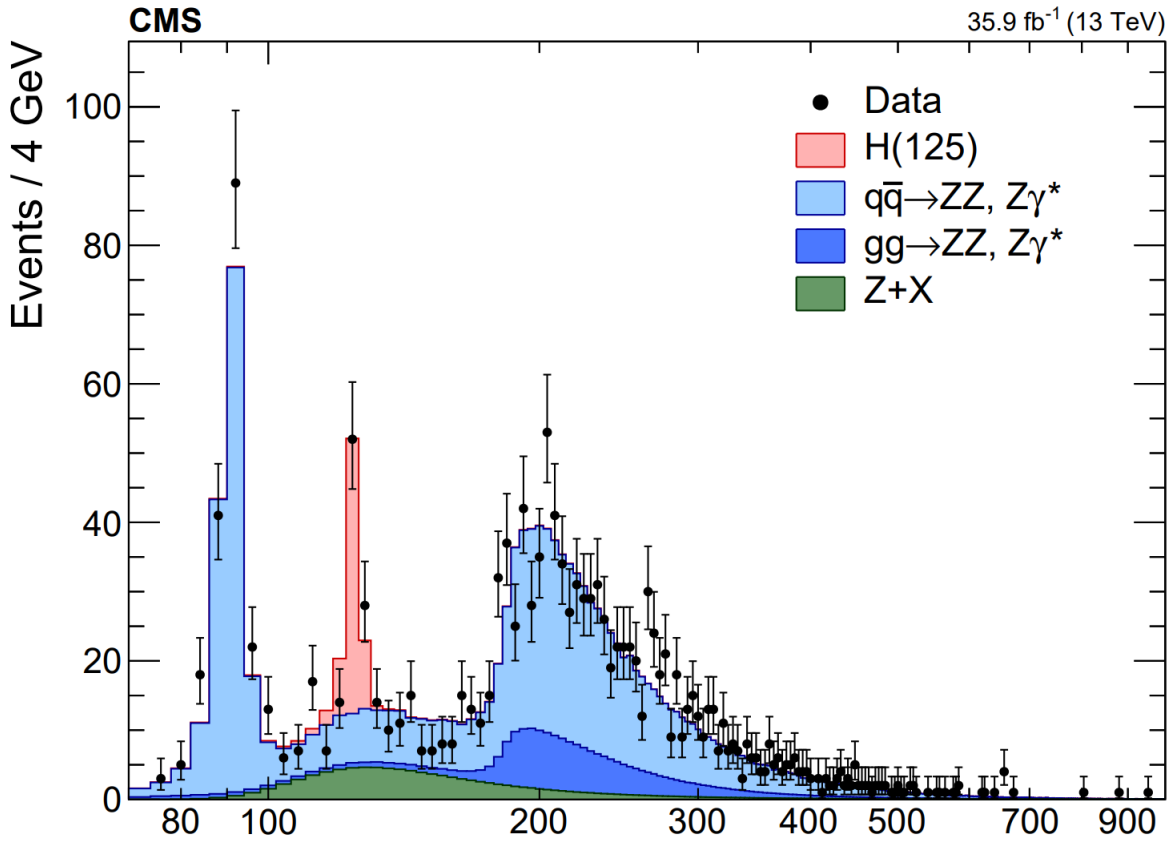
One of the main discovery channel

Precise reconstruction

High signal to background ratio

- Irreducible ZZ background
- Reducible Background “Z+X”
  - Z
  - $t\bar{t}$
  - $Z\gamma$
  - WZ
  - WW





arxiv.org/pdf/1706.09936

Workflow
Unified data

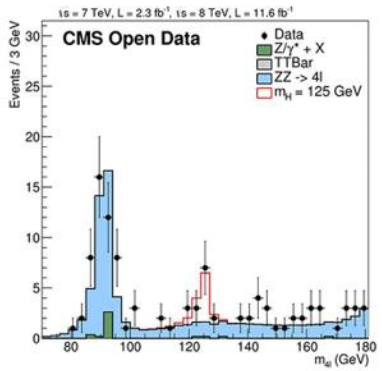
CMS

CMS legacy research data

opendata.cern.ch

Original CMS research data
(2010 100%, 2011/12 70%)

CMS legacy software
(Many CPU Month)

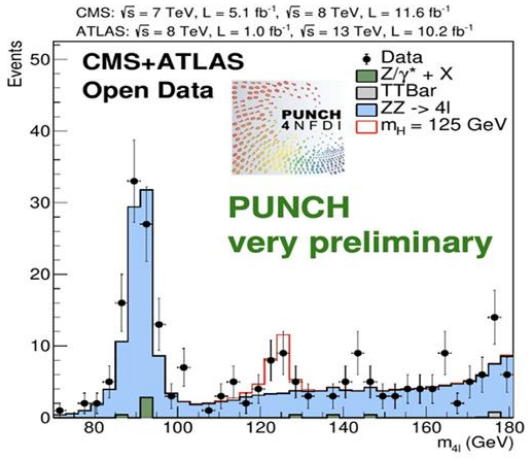


Transform once

Transformed Files in S4P / EOSPUBLIC (NanoAOD)

Transform once

1 script,
~2 hours,



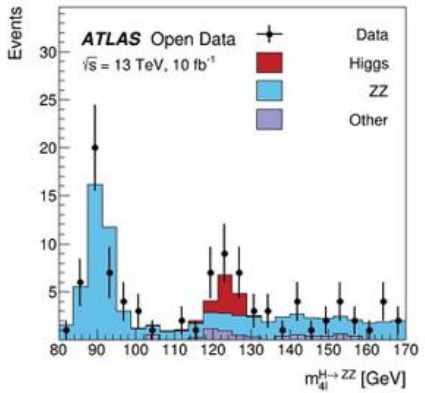
ATLAS

Non-public Atlas legacy data

opendata.cern.ch

Simplified educational data set of 2012 and 2016

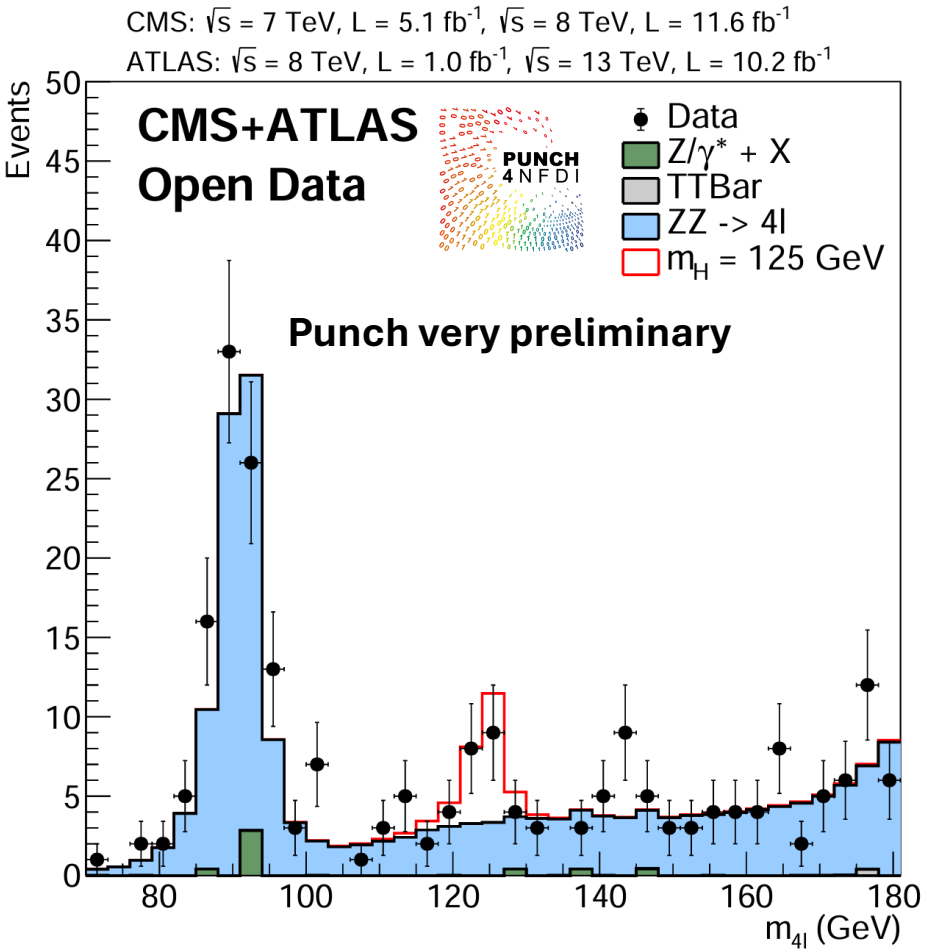
VM with dedicated software package or Jupyter notebook



Multi Experiment Higgs ➔ 4Lepton

Create the 4Lepton mass spectrum:

Using 2010, 2011, 2012 CMS and 2012, 2016 ATLAS open data.



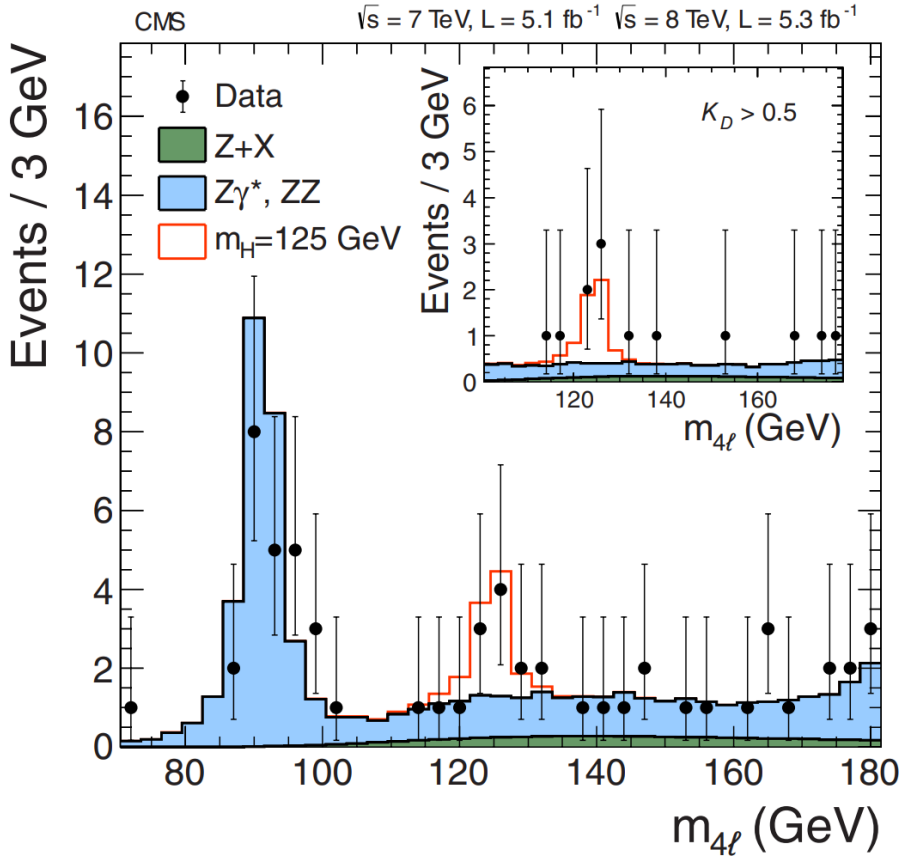
Four complexity levels are implemented:

1. **Observe** and compare the final output plot
2. **Reproduce** the final plot from predefined histograms
3. **Produce** a root data input from parts of the full analysis
4. **Produce** the full analysis

➔ Ready to be run, by PUNCH user!

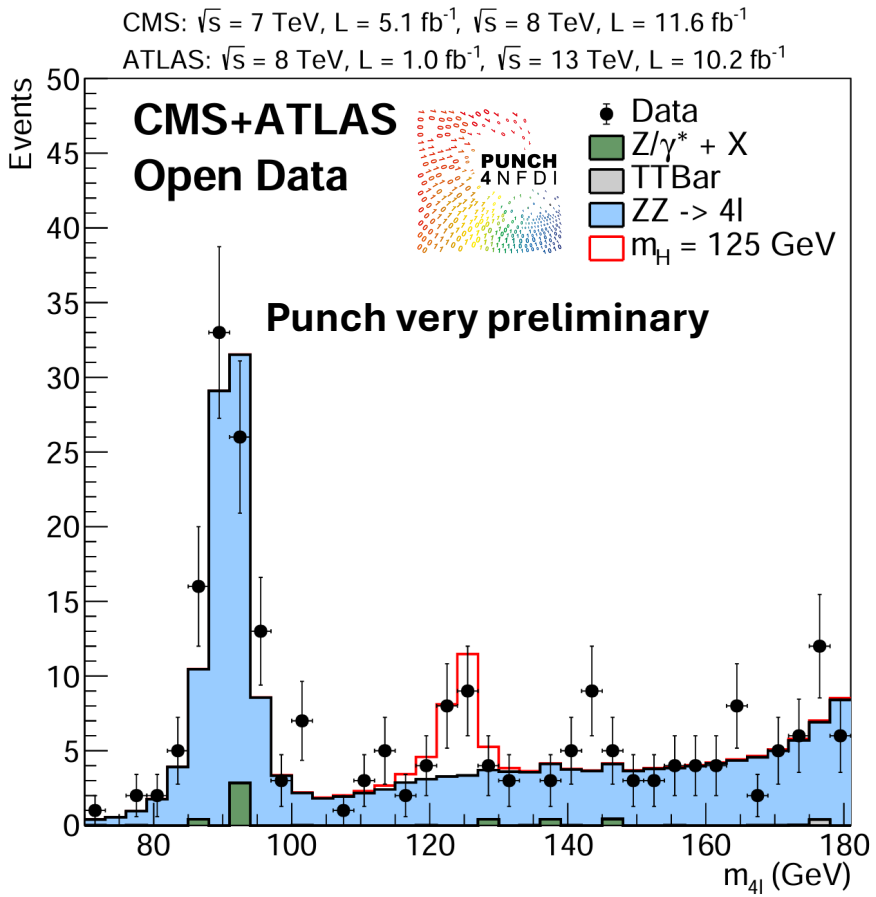
Level 1: Compare

CMS: Higgs discovery



arXiv:1207.7235

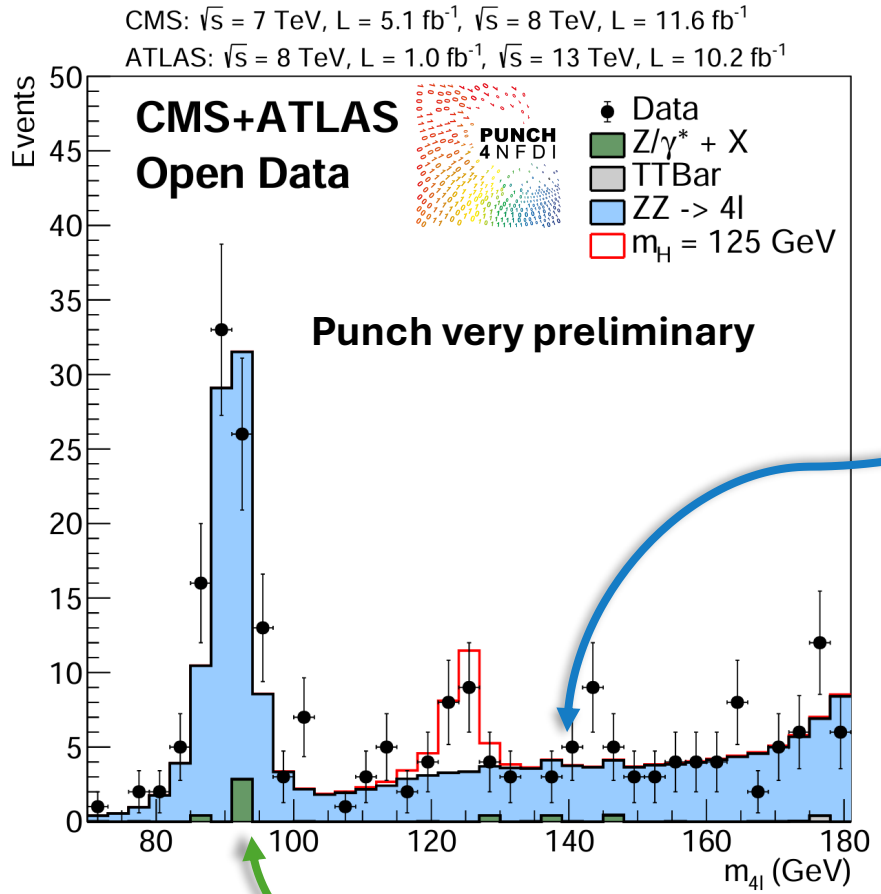
PUNCH4NFDI: Open Data

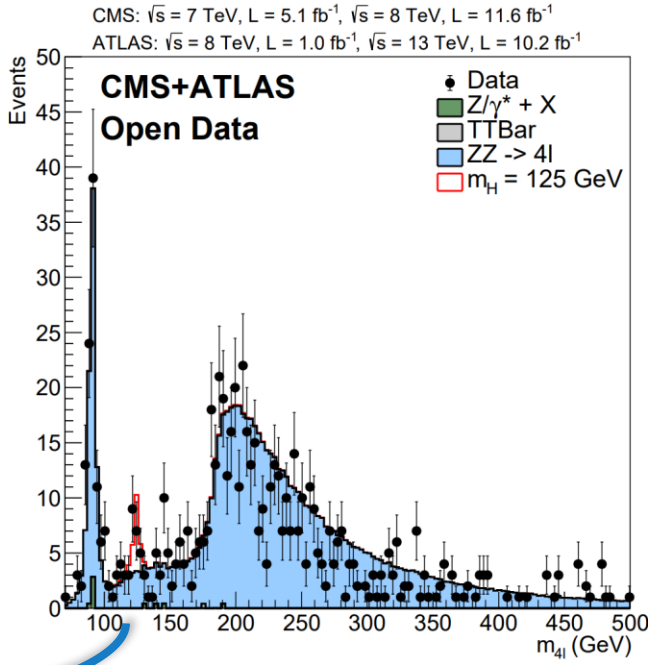


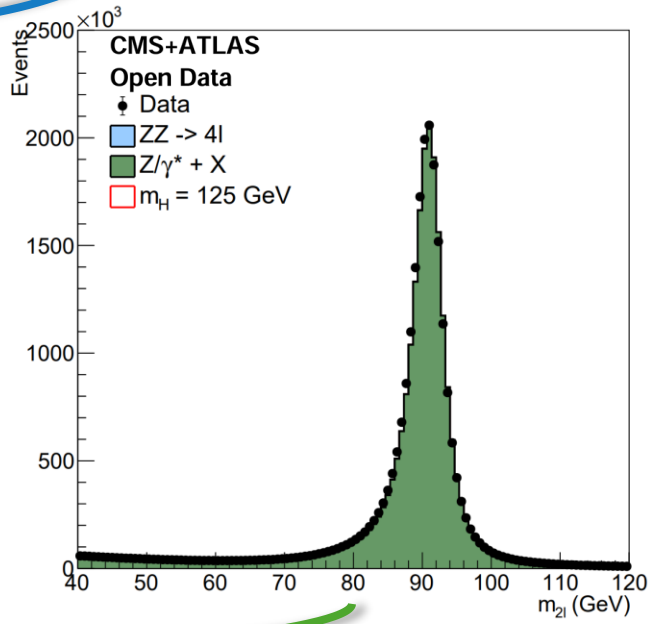
Data driven background normalization

Normalise the MC to the Data:

- The Drell Yan background to the 2l Z-peak
- The ZZ background to fit the whole mass spectrum
- The Higgs sample to the signal region







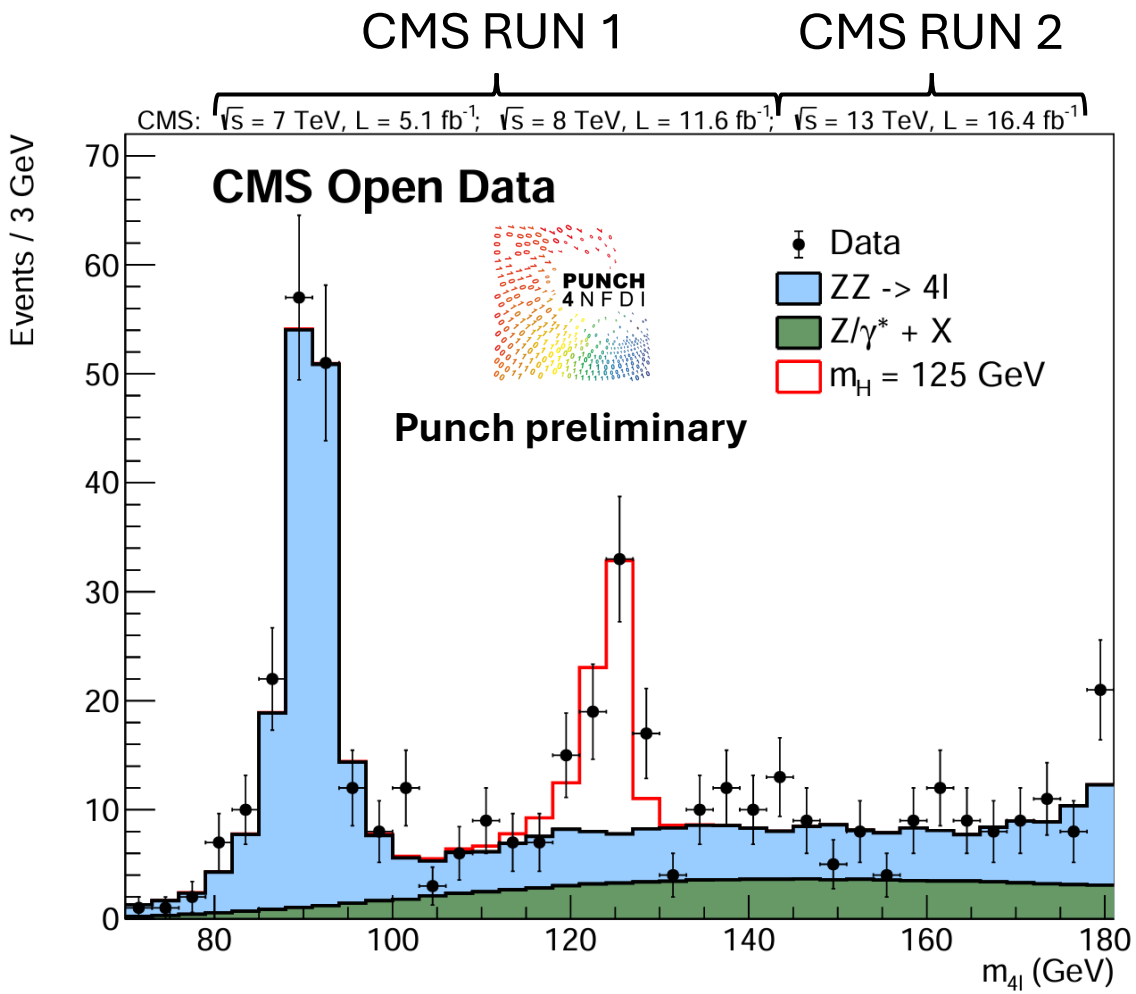
What about you?

Access to the transformed files on S4P is limited to PUNCH users

Multi run analysis with data from EOSPUBLIC

You can reproduce this analysis within hours!

- No registration
- No token required
- No affiliation required
- No access restriction
- You only need a working root (CVMFS) installation



Ready to be run, by **YOU**

https://gitlab-p4n.aip.de/lucas.karwatzki/DPG_HEP_HiggsTo4L_PUNCH

Data driven background normalization 2.0

Many issues with the background:

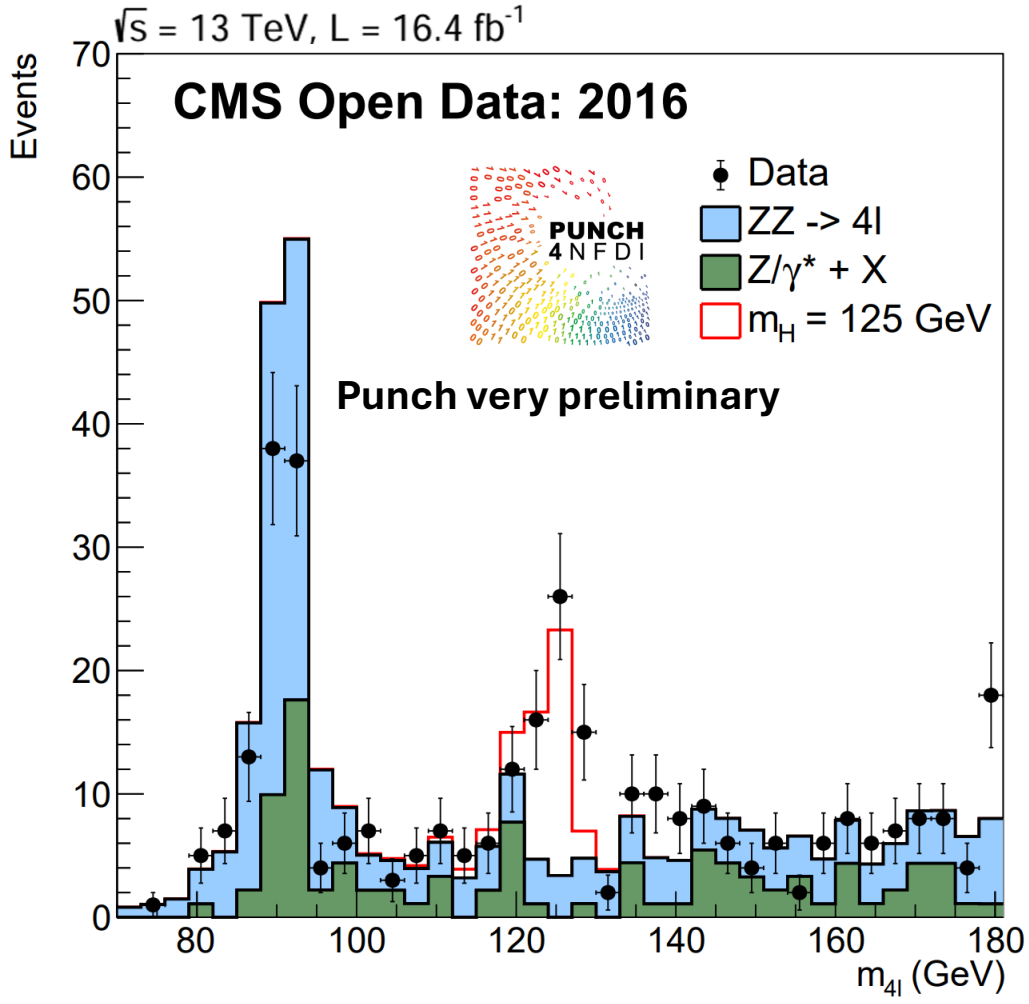
- Background samples missing on EOS public
- Double counting in the Z peak
- Non uniform Background

➢ **Improve our background implementation**

Perform a Landau fit

Adjust empirically to fit the whole irreducibly background

Adjust ZZ background based on the whole distribution



Data driven background normalization 2.0

Many issues with the background:

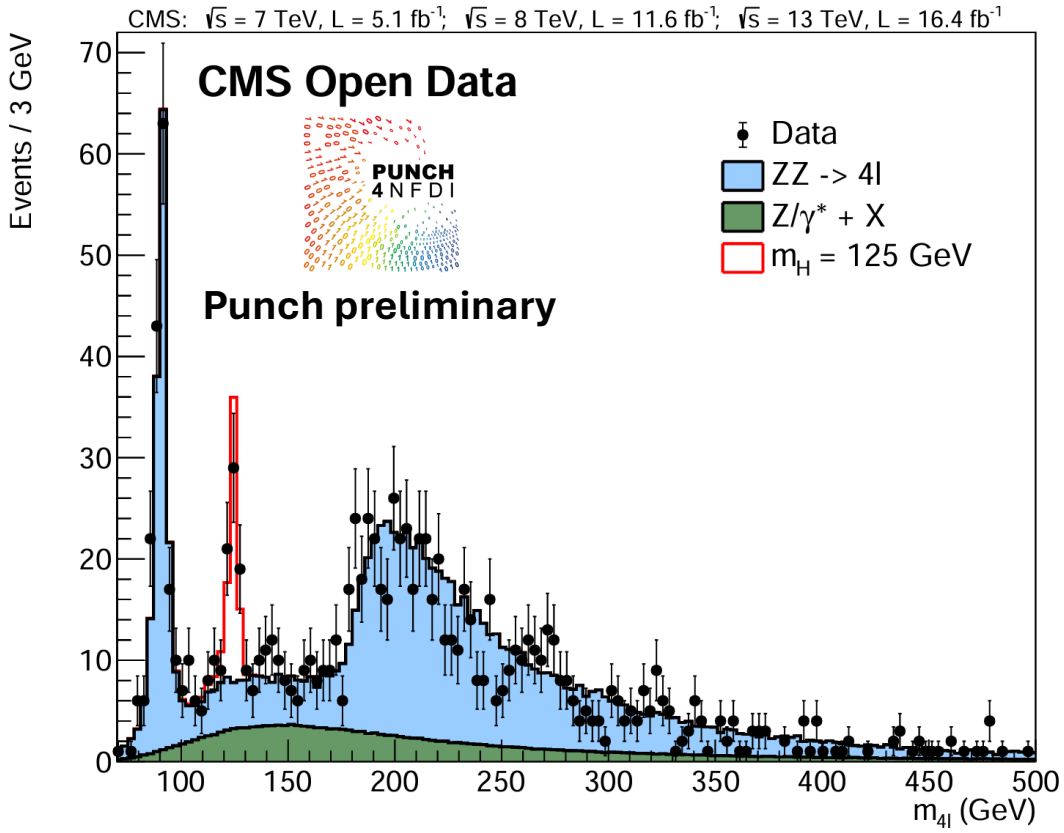
- Background samples missing on EOS public
- Double counting in the Z peak
- Non uniform Background

➢ **Improve our background implementation**

Perform a Landau fit

Adjust empirically to fit the whole irreducibly background

Adjust ZZ background based on the whole distribution



Conclusion

We aim to provide:

- A unified framework to LHC open data
- Fully implemented and well-documented use case examples
- A *FAIR* way for everybody to perform analysis

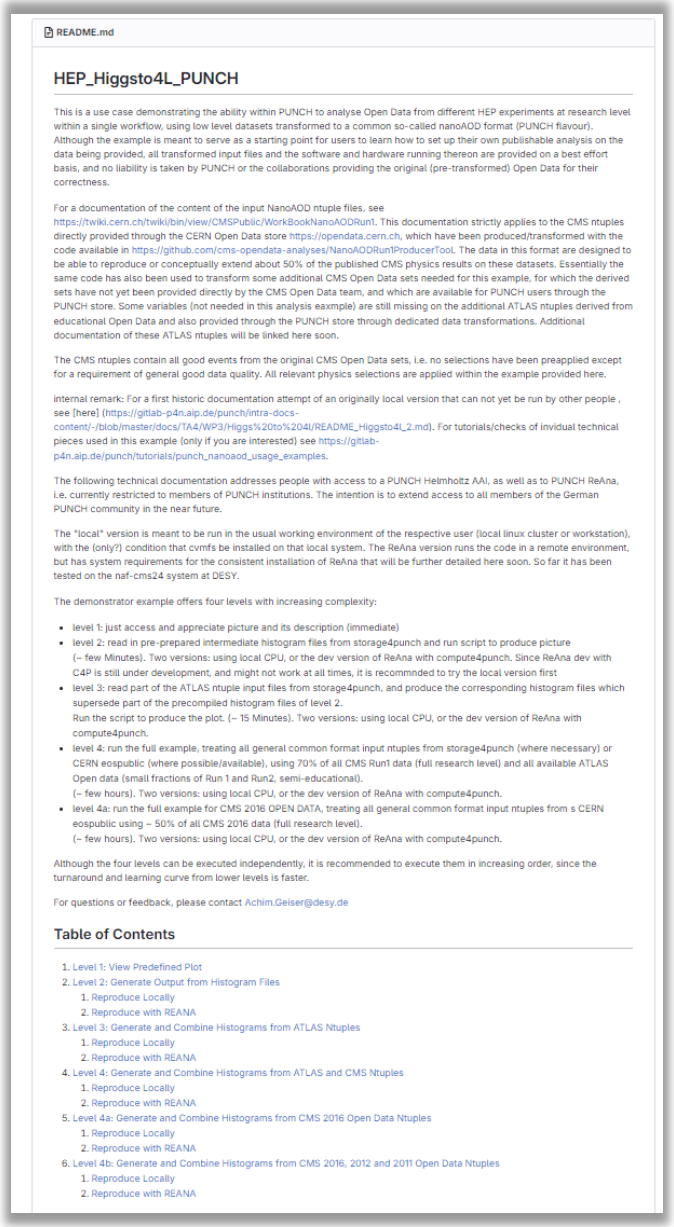
To scientist, who want to

- Conduct research in particle physics
- Combine data from multiple experiments

Have a look at the examples and any feedback is very welcome!

https://gitlab-p4n.aip.de/lucas.karwatzki/DPG_HEP_HiggsTo4L_PUNCH

Thank you for your Attention!



README.md

HEP_Higgsto4L_PUNCH

This is a use case demonstrating the ability within PUNCH to analyse Open Data from different HEP experiments at research level within a single workflow, using low level datasets transformed to a common so-called nanoAOD format (PUNCH flavour). Although the example is meant to serve as a starting point for users to learn how to set up their own publishable analysis on the data being provided, all transformed input files and the software and hardware running thereon are provided on a best effort basis, and no liability is taken by PUNCH or the collaborations providing the original (pre-transformed) Open Data for their correctness.

For a documentation of the content of the input NanoAOD ntuple files, see https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAODRun1. This documentation strictly applies to the CMS ntuples directly provided through the CERN Open Data store https://opendata.cern.ch, which have been produced/transformed with the code available in https://github.com/cms-opendata-analyses/NanoAODRun1ProducerTool. The data in this format are designed to be able to reproduce or conceptually extend about 50% of the published CMS physics results on these datasets. Essentially the same code has also been used to transform some additional CMS Open Data sets needed for this example, for which the derived sets have not yet been provided directly by the CMS Open Data team, and which are available for PUNCH users through the PUNCH store. Some variables (not needed in this analysis eaxmple) are still missing on the additional ATLAS ntuples derived from educational Open Data and also provided through the PUNCH store through dedicated data transformations. Additional documentation of these ATLAS ntuples will be linked here soon.

The CMS ntuples contain all good events from the original CMS Open Data sets, i.e. no selections have been preapplied except for a requirement of general good data quality. All relevant physics selections are applied within the example provided here.

internal remark: For a first historic documentation attempt of an originally local version that can not yet be run by other people , see [here] (https://gitlab-p4n.aip.de/punch/intra-docs-content/-/blob/master/docs/TA4/WP3/Higgs%20to%204l/README_Higgsto4L_2.md). For tutorials/checks of invidual technical pieces used in this example (only if you are interested) see https://gitlab-p4n.aip.de/punch/tutorials/punch_nanoaod_usage_examples.

The following technical documentation addresses people with access to a PUNCH Helmholtz AAI, as well as to PUNCH ReAna, i.e. currently restricted to members of PUNCH institutions. The intention is to extend access to all members of the German PUNCH community in the near future.

The "local" version is meant to be run in the usual working environment of the respective user (local linux cluster or workstation), with the (only?) condition that cvmfs be installed on that local system. The ReAna version runs the code in a remote environment, but has system requirements for the consistent installation of ReAna that will be further detailed here soon. So far it has been tested on the naf-cms24 system at DESY.

The demonstrator example offers four levels with increasing complexity:

- level 1: just access and appreciate picture and its description (immediate)
- level 2: read in pre-prepared intermediate histogram files from storage4punch and run script to produce picture (~ few Minutes). Two versions: using local CPU, or the dev version of ReAna with compute4punch. Since ReAna dev with C4P is still under development, and might not work at all times, it is recommnded to try the local version first
- level 3: read part of the ATLAS ntuple input files from storage4punch, and produce the corresponding histogram files which supersede part of the precompiled histogram files of level 2.
  Run the script to produce the plot. (~ 15 Minutes). Two versions: using local CPU, or the dev version of ReAna with compute4punch.
- level 4: run the full example, treating all general common format input ntuples from storage4punch (where necessary) or CERN eospublic (where possible/available), using 70% of all CMS Run1 data (full research level) and all available ATLAS Open data (small fractions of Run 1 and Run2, semi-educational).
  (~ few hours). Two versions: using local CPU, or the dev version of ReAna with compute4punch.
- level 4a: run the full example for CMS 2016 OPEN DATA, treating all general common format input ntuples from s CERN eospublic using ~ 50% of all CMS 2016 data (full research level).
  (~ few hours). Two versions: using local CPU, or the dev version of ReAna with compute4punch.

Although the four levels can be executed independently, it is recommended to execute them in increasing order, since the turnaround and learning curve from lower levels is faster.

For questions or feedback, please contact Achim.Geiser@desy.de

Table of Contents

BACKUP

Resource usage: Higgs ➔ 4Lepton

All tests are run on DESY naf-cms24.

Level	Storage4punch (Read only)	Eospublic (Read only)	Local			ReAna		
			Real time	CPU time	Storage	Real time	CPU time	Storage
1	/	/	/	/	/	/	/	/
2 - Local	1.9 MB	/	6 min*	30 sec*	3.2 MB	/	/	/
2 - ReAna	1.9 MB	/	>10 sec*	>10 sec*	3.2 MB	19 min*	12 sec*	120 KB
3 - Local	1.3 GB	/	7 min*	1 min 15 sec*	3.4 MB	/	/	/
3 - ReAna	1.3 GB	/	>10 sec*	>10 sec*	3.2 MB	20 min*	55 sec*	340 KB
4 - Local	206 GB	668 GB	2h 35 min*	1 h 30 min*	7.1 MB	/	/	/
4 - ReAna	206 GB	668 GB	>10 sec*	>10 sec*	3.2 MB	1 h 37 min*	53 min*	5.9 MB

*: Local times are only approximate and will vary from system to system.
*: ReAna times are approximate only and will vary depending on cluster occupancy.

Cuts for H→4L analysis

Quality Cuts:
- Muon_isGlobal = true
- Muon_isPFcand = true
- Electron_isPFcand = true

Event Selection:
- QL1 + QL2 + QL3 + QL4 = 0

Selection and Kinematic Cuts:
- p_T > 7 GeV and |η|<2.5 for electrons
- p_T > 5 GeV and |η|<2.4 for muons
- | D_{xy} | < 0.5 and | D_z |<1.0 for electrons+muons
- Rel. Iso < 0.4 for electrons and muons
- Lost Hits <= 1 for electrons

Electron/muon pairs must arise from same primary vertex: | SIP_{3D} |<4.0

For the final event selection two pairs of same flavour, opposite charge leptons are formed.
Possible combination: 4μ and 4e and 2μ2e

For the pair (Z_a → 2L) with invariant mass closest to Z boson mass:
- $Q_{L_i} + Q_j = 0$
- p_{T,L_i} > 20.0 GeV and p_{T,L_j} > 10.0 GeV
- 40.0 GeV < M_{Z_a} < 120.0 GeV

For the other pair (Z_b → 2L), we require:
- $Q_{L_i} + Q_j = 0$
- p_{T,L_i} > 20.0 GeV and p_{T,L_j} > 10.0 GeV
- 12.0 GeV < M_{Z_b} < 120.0 GeV