

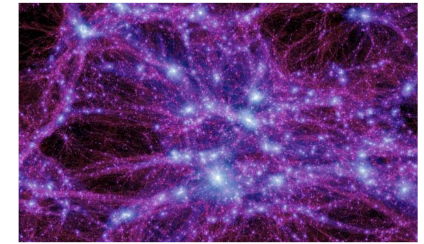
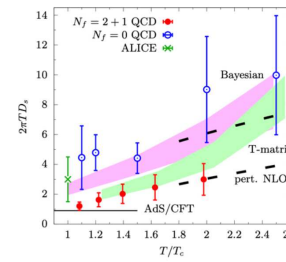
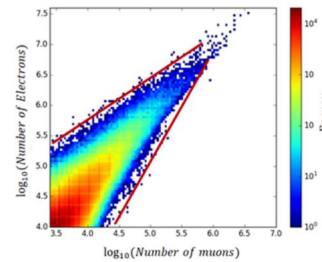
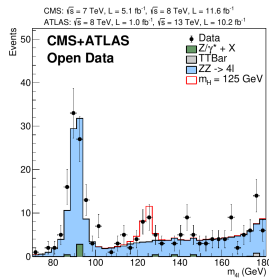
Multi-experiment and multi-era research level LHC Open Data analysis with PUNCH4NFDI



A. Geiser, DESY, 03.10.24, DPHEP, CERN, for **PUNCH4NFDI**:

Particles, Universe, NuClei and Hadrons for the NFDI –

the consortium of particle, astroparticle, hadron&nuclear physics and astronomy

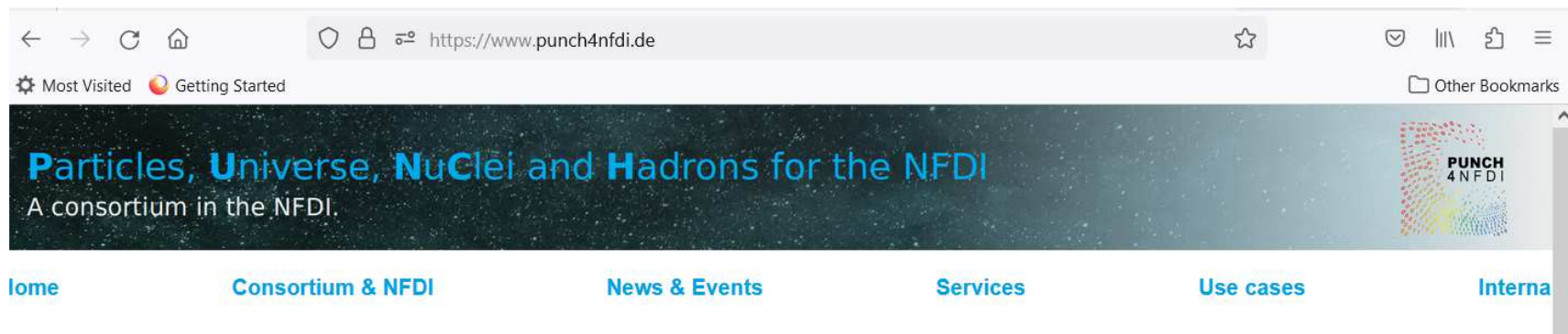


- Scope of the NFDI, PUNCH and the **PUNCH4NFDI** science data platform

Use case/work flow examples:

- Higgs Boson “Rediscovery” using **CERN Open Data**
- Summary and outlook

What is NFDI?

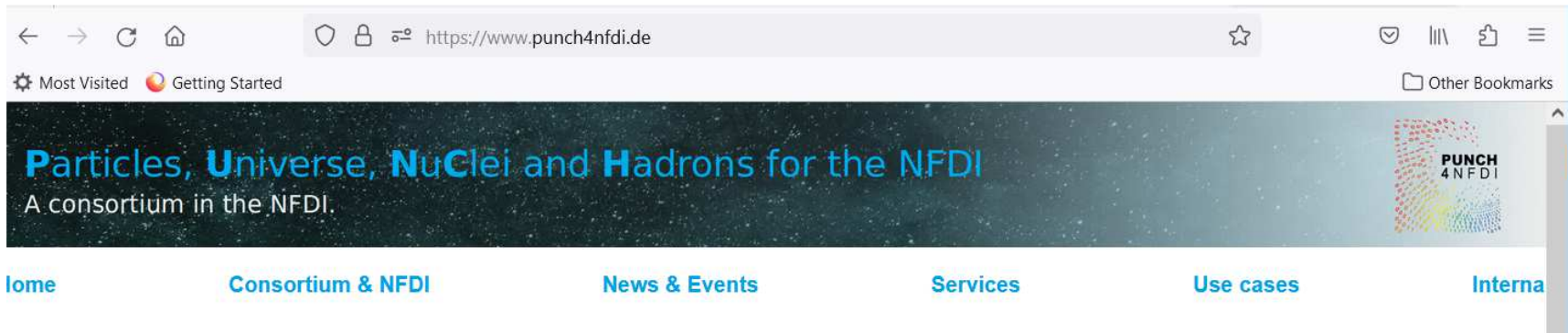


- The (German) National Research Data Infrastructure (in original German: **N**ationale **F**orschungs-**D**aten **I**nfrastruktur [NFDI](#)) has the objective to systematically index, edit, interconnect and make available the valuable stock of data from science and research..

What is PUNCH?

- **PUNCH4NFDI** is the **NFDI consortium of particle, astro-, astroparticle, hadron and nuclear physics**, representing about 9.000 scientists with a Ph.D. in Germany, from universities, the Max Planck society, the Leibniz Association, and the Helmholtz Association. **PUNCH physics** addresses the **fundamental constituents of matter and their interactions**, as well as their role for the development of the largest structures in the universe - stars and galaxies. The achievements of PUNCH science range from the **discovery of the Higgs boson** over the installation of a 1 cubic kilometre particle detector for neutrino detection in the antarctic ice to the detection of the quark-gluon plasma in heavy-ion collisions and the first picture ever of the black hole at the heart of the Milky Way.
- The prime goal of PUNCH4NFDI is the **setup of a federated and "FAIR" science data platform**, offering the infrastructures and interfaces necessary for the access to and use of data and computing resources of the involved communities and beyond. The High-Level Milestones of the PUNCH4NFDI consortium can be seen here: [\(PDF, 42KB\)](#)
- PUNCH4NFDI will also **offer services for the efficient scientific exploitation of research data**. In doing so, PUNCH4NFDI collaborates closely with its European and international partners at the forefront of research in scientific data management. **CERN** is one of these participant partner institutions.

Scope of the PUNCH4NFDI Science Data Platform



- via PUNCH4NFDI portal and related AAI, the **PUNCH Science Data Platform** aims at providing **coherent access to the rich data collections** of the PUNCH communities.
- uses **federated infrastructures and resources** like **Compute4PUNCH (C4P)** and **Storage4PUNCH (S4P)**, to enable **complex workflows**. **Make them work together coherently also for users without special expertise.**

Technicalities (plenty!) not discussed here.

- special focus on eventual **cross-community uses** of the **heterogeneous data and software** from particle, astroparticle and hadron&nuclear physics and astronomy.

Development of public demonstrator use cases

- **Final goal: Cross-community use cases** (not yet advanced enough to go public)
- **Currently: Community-specific research-level use cases. Here:**
 - **Cross-Experiment HEP analysis.**
Challenge: **Experiments have different detectors and different data formats, analysis software and analysis workflows.**
 - **Cross-era HEP analysis.**
Challenge: **Even within a given experiment, data from different eras** (e.g. LHC Run 1, Run 2, or even different suberas) **reflect slightly different detectors and often different mutually incompatible versions of data formats and software.**
- For first examples from **Astrophysics and lattice QCD**, see e.g.
https://indico.scc.kit.edu/event/3476/contributions/14171/attachments/6681/10513/MUdaysSept23_2.pdf

Combined CMS + ATLAS Higgs “Rediscovery”



First example of analysis of low level (open) data from two different HEP experiments within the same analysis framework.

Main purpose: **demonstrate practical feasibility of a PUNCH use case on the PUNCH Science Data Platform going significantly beyond what is already available outside PUNCH** (i.e. not just an import of things already available elsewhere), **using PUNCH resources** already now wherever possible.

Relevant **PUNCH-transformed** (or CMS-OD transformed) **data sets**

(mimicking official CMS nanoAOD format)

already produced and **stored on S4P** (Storage4PUNCH)

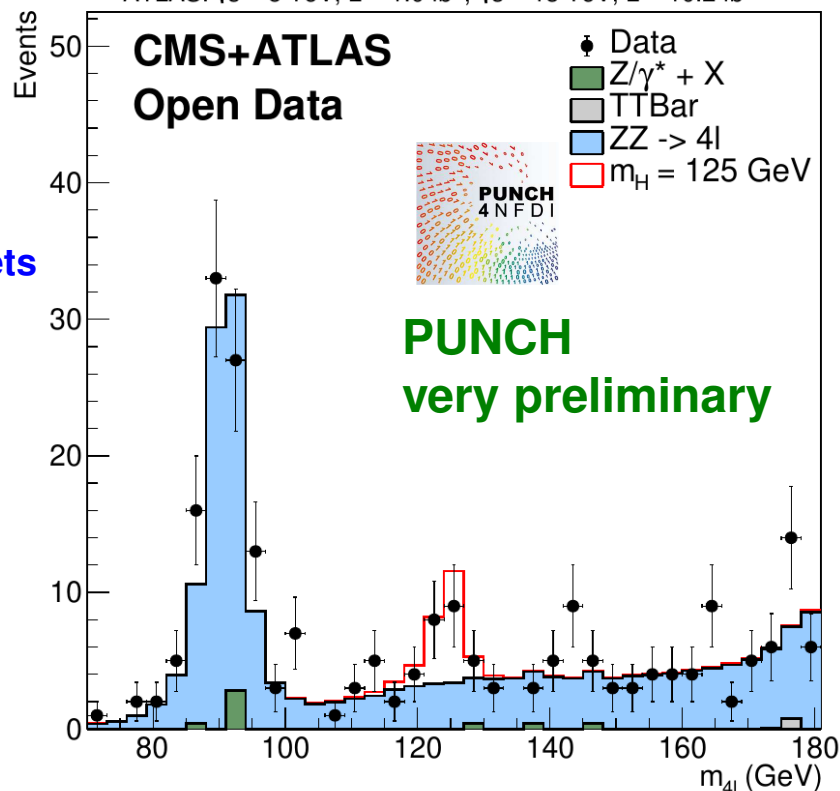
or **accessed from CERN eospublic** when available,

Using Xrootd streaming.

Full PUNCH workflow (series of ROOT scripts)

implemented on **PUNCH instance of REANA** workflow manager

CMS: $\sqrt{s} = 7 \text{ TeV}$, $L = 5.1 \text{ fb}^{-1}$, $\sqrt{s} = 8 \text{ TeV}$, $L = 11.6 \text{ fb}^{-1}$
ATLAS: $\sqrt{s} = 8 \text{ TeV}$, $L = 1.0 \text{ fb}^{-1}$, $\sqrt{s} = 13 \text{ TeV}$, $L = 10.2 \text{ fb}^{-1}$



Combined CMS + ATLAS Higgs "Rediscovery"



LHC Open Data

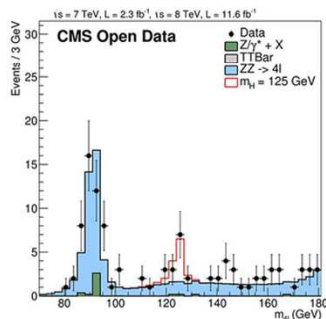
Original CMS legacy research data
(2 PB via CERN Open Data portal)
(2010 data 100%, 2011/12 data 70%)



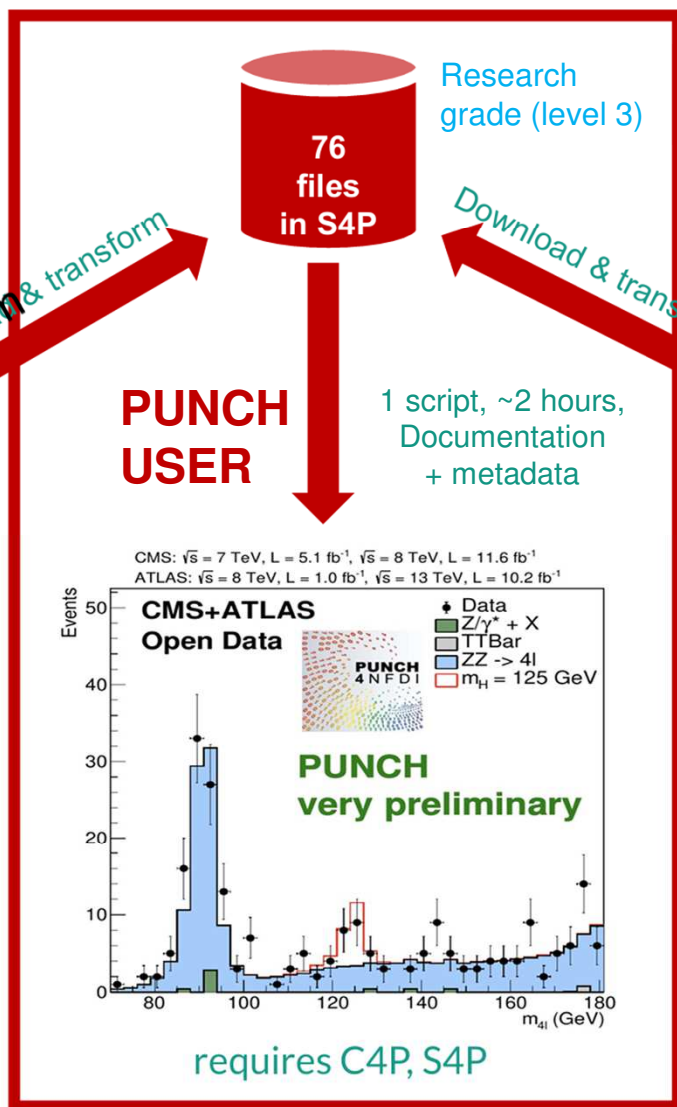
CMS legacy software from public github via CERN Open Data portal



Produce histograms (many CPU months)



Alternative: educational (level 2) version by ROOT team running within few hours



REANA for workflow management and resource organisation

can alternatively also be run outside ReAna on local computing

Non-public ATLAS legacy data

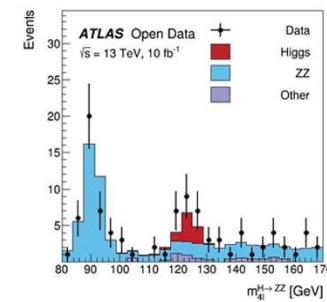


Simplified educational data sets 2012 / 2016 via CERN Open Data portal or ATLAS Open Data portal

(to be replaced by recent ATLAS research grade PHYSLite data release)



VM with dedicated software package or Jupyter notebook



CMS + ATLAS Higgs->4Lepton demonstrator project for PUNCH platform



Will hopefully soon become fully public, so far accessible to registered PUNCH users

https://gitlab-p4n.aip.de/punch/usecase_demonstrators/hep_higgsto4l_punch

ting (29-Marc...

punch / UseCase_Demonstrators / HEP_Higgsto4L_PUNCH

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 example) 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.



PUNCH dev version

Your workflows

Refreshed at 21:59

Search...	Status	Show deleted runs	Latest first
	✔		finished in 1h 43m 59s step 3/3
hep_higgsto4l_punch_level4 #2	11 MiB	Finished 6 days ago	
✔		finished in 2h 42m 11s step 3/3	
hep_higgsto4l_punch_level4 #1	10.99 MiB	Finished 6 days ago	
✔		finished in 12 min 19 sec step 3/3	
hep_higgsto4l_punch_level3 #4	392 KiB	Finished 13 days ago	
✘		failed after 0 seconds step 0/3	
hep_higgsto4l_punch_level3 #3	100 KiB	Finished 13 days ago	

03.10.24

A. Geiser, DPFIET meeting.

Higgs to four-lepton spectrum, CMS only version including CMS Run 2 Open Data

Part of DESY/PUNCH summer student project 2024, Lucas Karwatzki

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

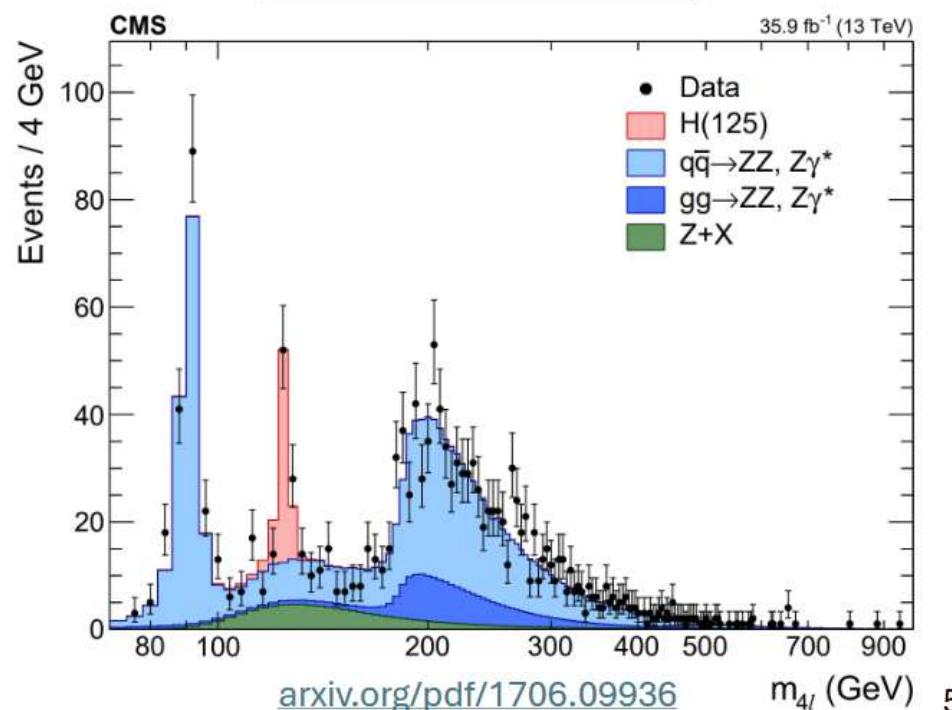
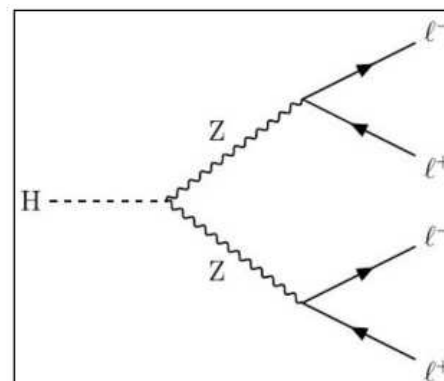
The 4 Lepton Channel

One of the main discovery channel

Precise reconstruction

High signal to background ratio

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



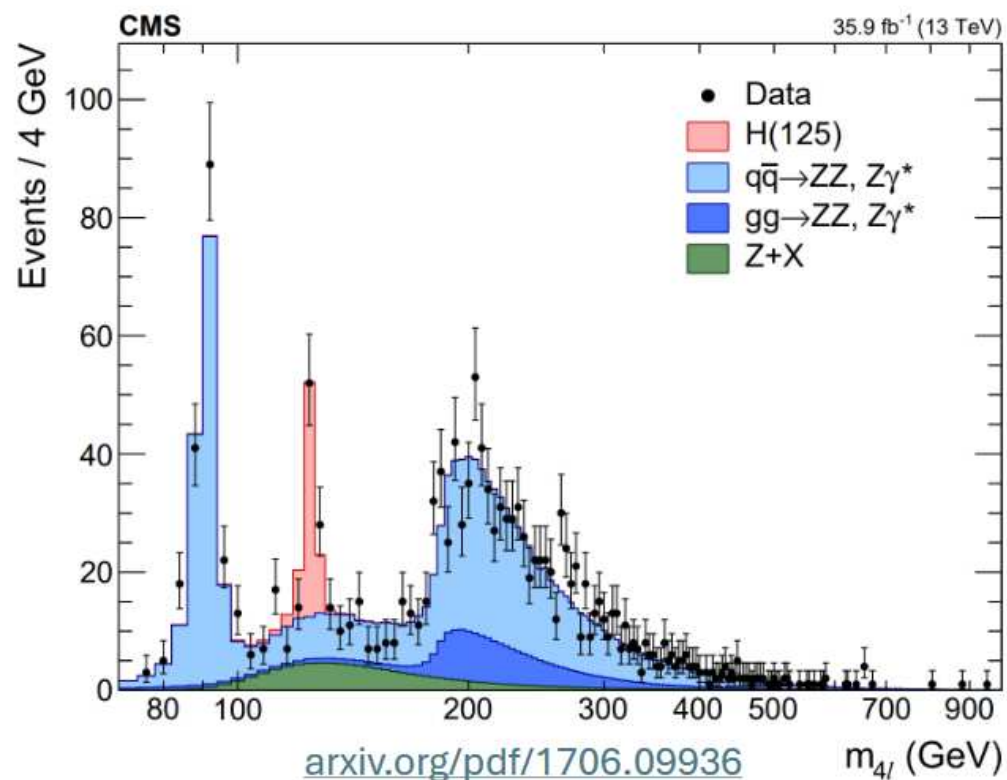
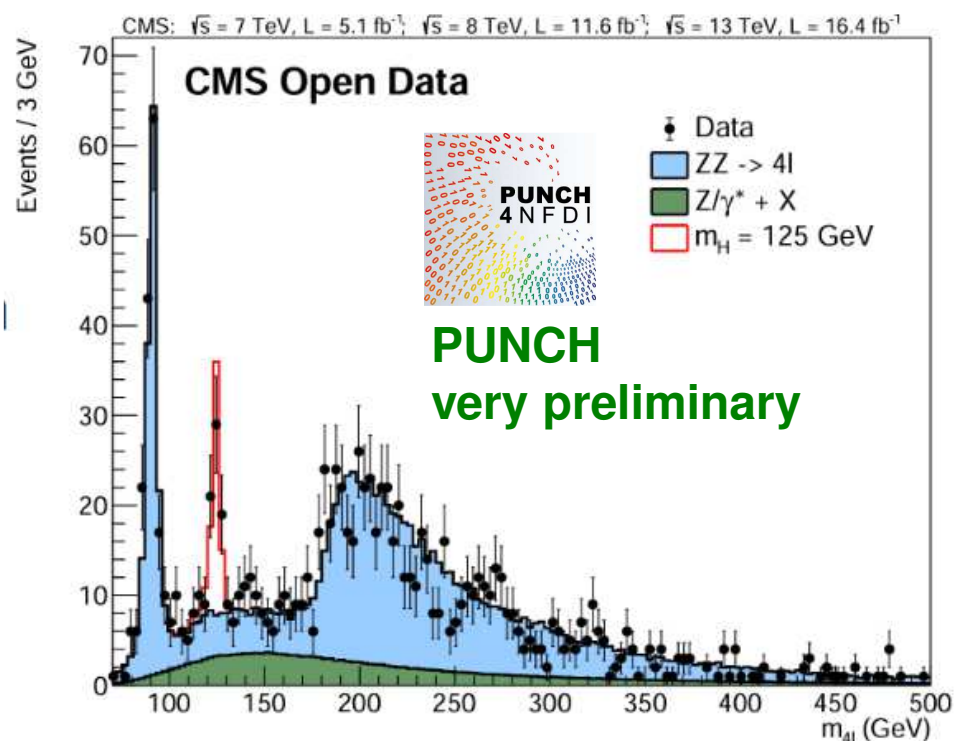
Higgs to four-lepton spectrum, CMS only version including Run 2 Open Data

Part of DESY/PUNCH summer student project 2024, Lucas Karwatzki

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

Public example, 70% Run1, 50% 2016

CMS published reference, 100% 2016



Summary and Outlook



- Via PUNCH4NFDI portal and related AAI, the **PUNCH Science Data Platform** aims at providing **coherent access to the rich data collections** of the PUNCH communities, including **cross-community applications** on **federated infrastructures**.
- Some **first example usecases/workflows** are **already being implemented, and starting to be made public in parallel to the gradual implementation of the data management infrastructures**. Many more, and more ambitious ones, to come.
- These demonstrator use-cases are intended as **examples and inspirations for users of the platform to design and implement their own applications on the platform** in a greatly simplified/unified environment which is so far not available anywhere else. Test users welcome. In practice, **some `under development` restrictions currently still apply**. More data are going to be integrated.
- The data and workflow setup for the **H to 4L use case example** outlined here is in principle suited for doing **comprehensive analyses** like e.g. a determination of alphas or the W mass, **combining data from different origins** (experiments or eras), although in practice the level of systematic sophistication reached by proprietary experiment analyses might be hard/impossible to reach without additional internal information. -> **Rather for the implementation of new ideas**.