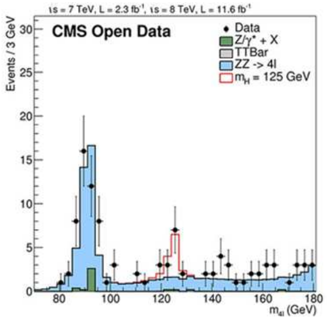
Combined CMS + ATLAS Higgs “Rediscovery”

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

CMS OD & PUNCH

stream

Download & transform

76 files in S4P

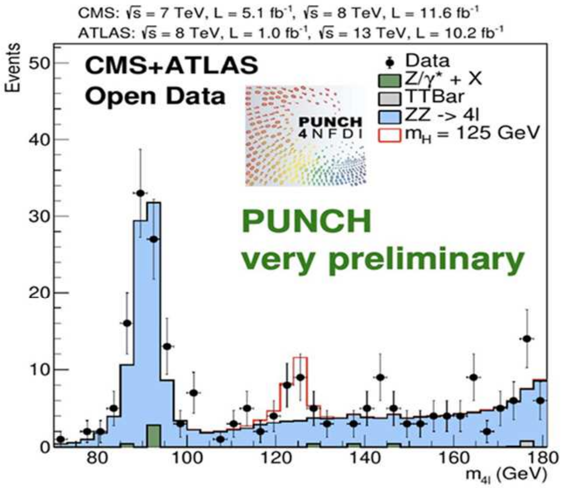
Research grade (level 3)

Download & transform

PUNCH

PUNCH USER

1 script, ~2 hours, Documentation + metadata



requires C4P, S4P

REANA for workflow management and resource organisation

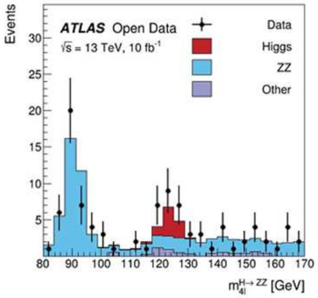
can alternatively also be run outside ReAna on local computing

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



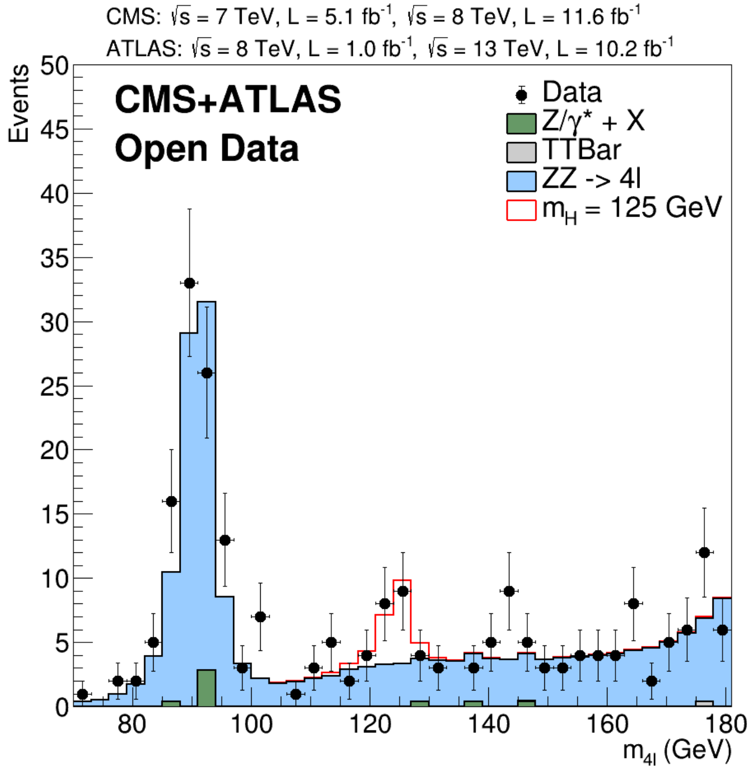
8

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

Higgs → 4Lepton

Create the 4Lepton mass spectrum:

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



Four complexity levels are implemented:

1. **Observe** 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 (WIP)

(sometimes) ready to be run, by (most) other PUNCH users [here]

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

H->4L example, RECAP of some history

* Original CMS H->4L example running on CMS Open Data with VM on CMS-specific format implemented within three months (by PhD student) at the end of 2017, and has continued to work publicly since then without major interruption.

Documented on CERN Open Data Portal. Needs local CPU and standard VM installation (working on any platform), but almost no disk resources needed, and no AAI.

4 levels (as for PUNCH example), level 4 needs 2 months of CPU – actually run by ATLAS person within 10 minutes on 3000 Google-CLOUD CPUs at CHEP conference. Nice, but not so realistic for non-power users.

* Much faster example set up by CERN ROOT team as a ROOT example. Portable format in principle, but educational, thus not suited for change of analysis topic or cross-experiment analysis. Tested successfully within PUNCH early-on by KIT team on C4P.

* Most of current PUNCH H->4L example set up before PUNCH started:
Content similar to original H->4L example, but extended to ATLAS data.
Reduction of disk space and CPU by ~ 2 orders of magnitude through transformation to simplified portable research-level format.
Including transformation of ATLAS data -> interoperable across experiments!

Running locally on separate threads already in summer 2021 (summer student project).
Bottleneck: place to store transformed ATLAS ntuples (and part of CMS transformed ones), since no cross-experiment storage forseen at CERN
--> plan: move the example to PUNCH and use PUNCH resources.

PUNCH H->4L: RECAP of some history

Initial great progress: worked almost immediately on DESY hifis/dcache-demonstrator store w/o AAI from fall 2021. Content streamlining to single thread and documentation ready by summer 2022 (summer student project).
Start thinking about public release (as planned, `within first year').

Then AAI requirement was added on S4P. Took until early summer 2024 (~ 2 years!) to resolve the issues for xrootd with token AAI on S4P. Essentially solved by CERN.
Now working on 4 different S4P stores (3 partial, 1 full: KIT) ☺, but AAI still partially an inconvenient hassle for users.

Running locally at DESY and other HEP sites with S4P since end of summer 2024.
Bottleneck: availability of cvmfs outside HEP to pick up relevant root version.
--> plan: make available on ReAna, with interface to C4P/S4P.

Work started in summer 2023 with 'local' ReAna dev version kindly provided by AIP (summer student project). Stopped working (for us) in fall 2023.

Work resumed with new much better remote ReAna dev version kindly provided by AIP in summer 2024, including documentation (summer student project).
-> essentially ready for public release, advertised in mid term report.
Stopped working (for us) in fall 2024 (update of ReAna-dev, no AAI access to C4P/S4P).

Resolved by integration into standard ReAna (via CERN) in early December 2024.
Stopped working again one week later (update of ReAna, version/installation issues).
Resolved by mid-January 2025, but still AAI issues for some users (e.g. Christiane).

Summary/conclusions

At some point, remaining issues will hopefully be resolved, but

Main recurring practical bottleneck: AAI issues for xrootd access to storage.

Has already twice delayed public release of the example by up to two years, and is still delaying it.

AAI necessary for some applications (CPU resources), but for many others not:

As already proposed several times:
Propose to create read-only S4P instance w/o AAI requirement, like CERN eospublic store for Open Data, to store PUNCH transformed input data that can not be stored at CERN or elsewhere.

Not all S4P sites would need to do this. Just one would be sufficient.

Initially few Tb would do (later maybe ~100 Tb?).

If we would not have stopped having such access after initial PUNCH S4P tests, the pre-ReAna version of the PUNCH H->4L example would have been out publicly and working since more than two years.

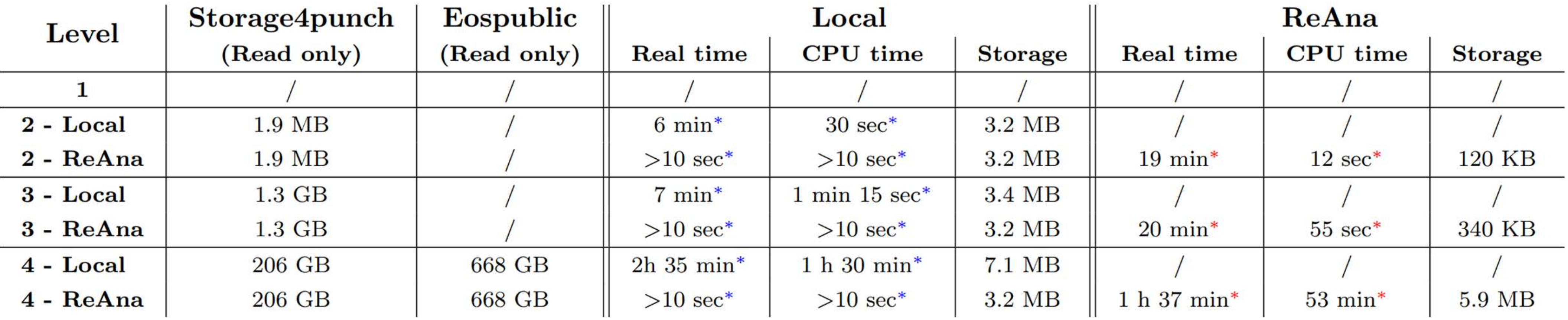
Also, even if it does work, avoiding AAI for input data reading makes life a lot simpler for users.

Backup

L. Karwatzki, PUNCH4NFDI WP3 meeting | 19.08.2024

Resource usage: Higgs ➔ 4Lepton

All tests are run on DESY naf-cms24.



Level	Storage4punch	Eospublic		Local			ReAna	
	(Read only)	(Read only)	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.