



The First Release of ATLAS Open Data for Research

CHEP 2024

21 October 2024

Zach Marshall (LBNL) on behalf of the ATLAS Collaboration

ATLAS Open Data for Research — CHEP 2024 — 21 Oct 2024 — Zach Marshall



Open Data for Research

Open Data for Research

Public, licensed,
documented, [FAIR](#)

Understandable, light,
useful format

Can be used for
scientific publications

Open Data for Research

Public, licensed,
documented, [FAIR](#)

Understandable, light,
useful format

Can be used for
scientific publications

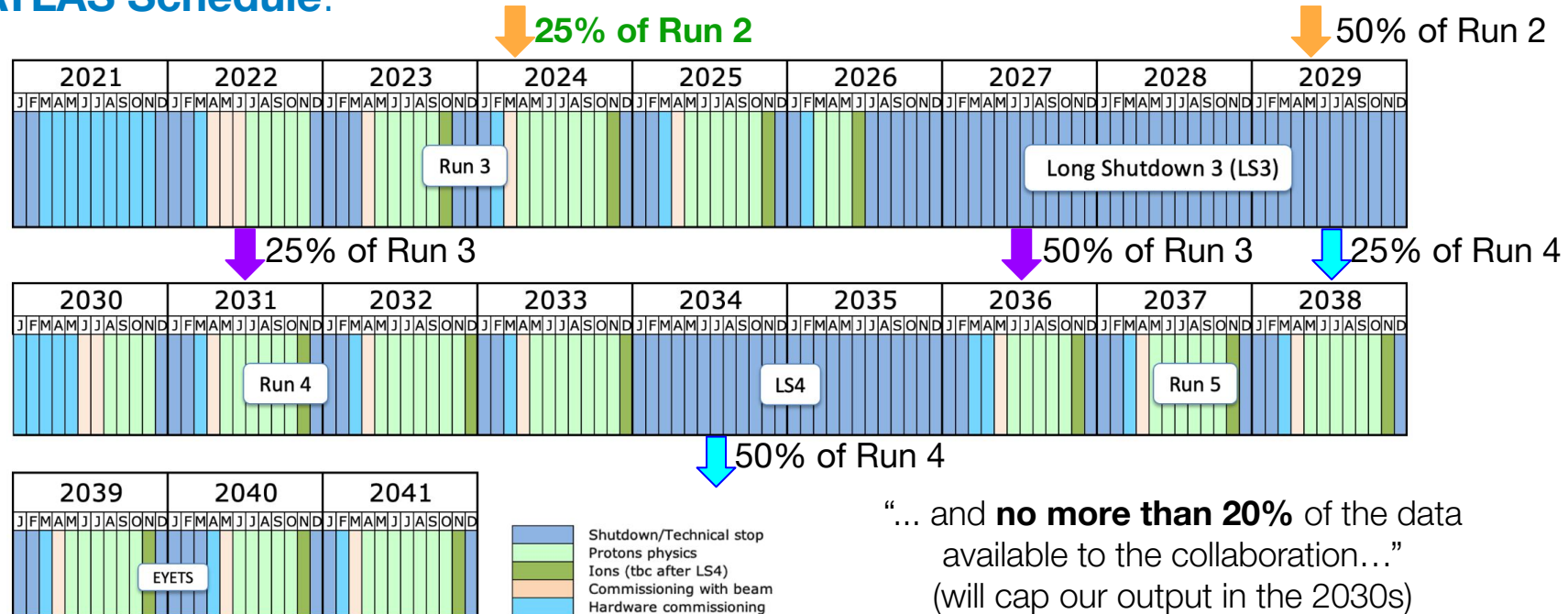
This talk: what we're doing, when, why, how, and a bunch of **fun facts**

- The Data Preservation for HEP (DPHEP) Collaboration [defined four levels](#) of open data

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

1. Yep, we do that all the time! [Plot records](#), [HepData](#), [Rivet analyses](#)...
2. We've been doing that for years as well! Previous open data ([8 TeV](#), [13 TeV](#)) [widely used for education and outreach](#): **more in Giovanni's talk**
3. **This is what we're talking about today**
4. Being preserved, but will not be released. 50 PB/year in Run 3 in a complex format, huge resources required for processing — not useful to the public (even expert public).

- We have lots of **bespoke datasets** for **targeted research applications**
 - [Top jet tagging](#), [Fast Simulation training](#), [BSM BDT training](#), ...
 - Kaggle Challenges: [Higgs boson ML Challenge](#), [TrackML Challenge](#)
- Four large LHC experiments agreed to release data for **general research use**
- **ATLAS Schedule:**



Last update: September 24

- All the goals of Outreach and Education open data
 - Democratic access, education and inspiration of future scientists
- These Open Data meet research standards — can be used for scientific papers
- Would love for excited researchers to use this as a step towards collaboration
 - Particularly if they want access to the **full dataset**
- This is also a step towards **data preservation**
 - Open data represents the **minimum** level of long-term support for data
 - If in 2040 we can no longer reconstruct Run 2 data, **at least we have open data**

Open Data

Definition

“ Open research data refers to the publishing of the data underpinning scientific research results so that they have no restrictions on their access and usage. Openly sharing data opens it up to inspection and re-use, forms the basis for research verification and reproducibility, and opens up a path to broader collaboration.

”

[UNESCO Open research data definition](#)

More on the web

What's the Goal?

- All the goals of Outreach and Education open data
 - Democratic access, education and inspiration of future scientists
- These Open Data meet research standards — can be used for scientific papers
- Would love for excited researchers to use this as a step towards collaboration
 - Particularly if they want access to the **full dataset**
- This is also a step towards **data preservation**
 - Open data represents the **minimum** level of long-term support for data
 - If in 2040 we can no longer reconstruct Run 2 data, **at least we have open data**

Open Data

More on the web

Definition

“ Open research data refers to the publishing of the data underpinning scientific research results so that they have no restrictions on their access and usage. Openly sharing data opens it up to inspection and re-use, forms the basis for **research verification and reproducibility**, and opens up a path to broader collaboration.

Open question: is this realistic? ”

UNESCO Open research data definition

What did we Release?

ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data

ATLAS collaboration

Cite as: ATLAS collaboration (2024). ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data. CERN Open Access Library. DOI:10.7483/OPENDATA.ATLAS.4ZES.DJHA

Dataset characteristics

538348881 events. 45571 files. 35.4 TiB in total.

More on the web

More on the web

Documentation on PHYSLITE Variables for ATLAS Open Data

Page generated from sample: mc20_13TeV_410471.PHPy8EG_A14_ttbar_hdamp258p75_allhad.deriv.DAOD_PHYSLITE.e6337_s3681_r13167_p5631

List of Containers:

- [AnalysisElectrons](#) | [AnalysisJets](#) | [AnalysisLargeRJets](#) | [AnalysisMuons](#) | [AnalysisPhotons](#) | [AnalysisTauJets](#) | [AnalysisTrigMatch](#) | [AntiK10TruthSoftDropBeta100Zcut10Jets](#) | [AntiK4TruthDressedWZJets](#) | [BTagging_AntiK4EMPFlow](#) | [CombinedMuonTrackParticles](#) | [egammaClusters](#) | [EventInfo](#) | [ExtrapolatedMuonTrackParticles](#) | [GSFConversionVertices](#) | [GSFTTrackParticles](#) | [HardScatterParticles](#) | [HardScatterVertices](#) | [InDetTrackParticles](#) | [K4EMPFlowEventShape](#) | [MET_Core_AnalysisMET](#) | [MET_Truth](#) | [MuonSpectrometerTrackParticles](#) | [PrimaryVertices](#) | [TauTracks](#) | [TruthBoson](#) | [TruthBosonsWithDecayParticles](#) | [TruthBosonsWithDecayVertices](#) | [TruthBottom](#) | [TruthElectrons](#) | [TruthEvents](#) | [TruthForwardProtons](#) | [TruthMuons](#) | [TruthNeutrinos](#) | [TruthPhotons](#) | [TruthPrimaryVertices](#) | [TruthTaus](#) | [TruthTop](#)

AnalysisElectrons [\(back to top\)](#)

Variable Name	Type	Description
ambiguityLink	vector<ElementLink<DataVector<xAOD::Egamma_v1>>>	Links Photon<-> Electron when ambiguous
ambiguityType	vector<unsigned char>	Ambiguity (almost surely electron 0 or photon 1) rel22/re121 or ambiguous 1-6/5, (>= rel22)
author	vector<unsigned short>	Electron, Photon, Ambiguous, Forward
caloClusterLinks	vector<vector<ElementLink<DataVector<xAOD::CaloCluster_v1>>>>	Photon/electron -> Cluster
charge	vector<float>	Electron charge
DFCommonElectronsECIDS	vector<char>	Charge sign (0: no decision)
DFCommonElectronsECIDSResult	vector<double>	Charge sign (0: no decision)
DFCommonElectronsLH.Loose	vector<char>	Like
DFCommonElectronsLH.LooseBL	vector<char>	Like

More on the web

- 2015+2016 Run 2 pp collision data
 - 45 TB of data, 6.3 kB/event, 7.1B events, 55k files in ~300 runs
 - 20 TB of MC, ~10 kB/event, 2B events, 16k files in ~300 MC datasets
- Explanation of [our nomenclature](#)
- Giant [tables of metadata](#)
 - Cross sections, k-factors, filters / efficiencies, processes, how to combine samples, configurations, ...
- PHYSLITE (ROOT-based) format
 - Already columnar — Uproot friendly
 - [Used for our own papers too](#)
- Pre-calibrated (first for ATLAS)
 - Just draw a plot!
- Extensive effort to document variables
- **Useful documentation for us as well!**

What did we Release?

ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data

ATLAS collaboration

Cite as: ATLAS collaboration (2024). ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data. CERN Open Access Library. DOI:10.7483/OPENDATA.ATLAS.4ZES.DJHA

Dataset characteristics

5383448881 events, **45571** files, **35.4 TIB** in total.

More on the web

More on the web

Documentation on PHYSLITE Variables for ATLAS Open Data

Page generated from sample: mc20_13TeV_410471.PHPy8EG_A14_ttbar_hdamp258p75_allhad.deriv.DAOD_PHYSLITE.e6337_s3681_r13167_p5631

List of Containers:

- [AnalysisElectrons](#) | [AnalysisJets](#) | [AnalysisLargeRJets](#) | [AnalysisMuons](#) | [AnalysisPhotons](#) | [AnalysisTauJets](#) | [AnalysisTrigMatch](#) | [AntiK10TruthSoftDropBeta100Zcut10Jets](#) | [AntiK14TruthDressedWZJets](#) | [BTagging_AntiK4EMPFlow](#) | [CombinedMuonTrackParticles](#) | [EgammaClusters](#) | [EventInfo](#) | [ExtrapolatedMuonTrackParticles](#) | [GSFConversionVertices](#) | [GSFTTrackParticles](#) | [HardScatterParticles](#) | [HardScatterVertices](#) | [InDetTrackParticles](#) | [K14EMPFlowEventShape](#) | [MET_Core_AnalysisMET](#) | [MET_Truth](#) | [MuonSpectrometerTrackParticles](#) | [PrimaryVertices](#) | [TauTracks](#) | [TruthBoson](#) | [TruthBosonsWithDecayParticles](#) | [TruthBosonsWithDecayVertices](#) | [TruthBottom](#) | [TruthElectrons](#) | [TruthEvents](#) | [TruthForwardProtons](#) | [TruthMuons](#) | [TruthNeutrinos](#) | [TruthPhotons](#) | [TruthPrimaryVertices](#) | [TruthTaus](#) | [TruthTop](#)

AnalysisElectrons [\(back to top\)](#)

Variable Name	Type	Description
ambiguityLink	vector<ElementLink<DataVector<xAOD::Egamma_v1>>>	Links Photon<-> Electron when ambiguous
ambiguityType	vector<unsigned char>	Ambiguity (almost surely electron 0 or photon 1) rel22/re121 or ambiguous 1-6/5, (>= rel22)
author	vector<unsigned short>	Electron, Photon, Ambiguous, Forward
caloClusterLinks	vector<vector<ElementLink<DataVector<xAOD::CaloCluSter_v1>>>>	Photon/electron -> Cluster
charge	vector<float>	Electron charge
DFCommonElectronsECIDS	vector<char>	Charge sign (0: no decision, 1: assignment)
DFCommonElectronsECIDSResult	vector<double>	Charge sign for the charge
DFCommonElectronsLHLoose	vector<char>	Like/Unlike decision
DFCommonElectronsLHLooseBL	vector<char>	Like/Unlike decision

More on the web

- 2015+2016 Run 2 pp collision data
 - 45 TB of data, 6.3 kB/event, 7.1B events, 55k files in ~300 runs
 - 20 TB of MC, ~10 kB/event, 2B events, 16k files in ~300 MC datasets
- Explanation of [our nomenclature](#)
- Giant [tables of metadata](#)
 - Cross sections, k-factors, filters / efficiencies, processes, how to combine samples, configurations, ...
- PHYSLITE (ROOT-based) format
 - Already columnar — Uproot friendly
 - [Used for our own papers too](#)
- Pre-calibrated (first for ATLAS)
 - Just draw a plot!
- Extensive effort to document variables
- **Useful documentation for us as well!**

What is the Right License?

- We want our Open Data to be usable for everyone
- We want people to cite us when they use our Open Data
- What's the right License to choose?

CC 0 CC0 1.0


CC0 1.0 UNIVERSAL

Deed

Canonical URL : <https://creativecommons.org/publicdomain/zero/1.0/>

[See the legal code](#)

No Copyright

 The person who associated a work with this deed has **dedicated** the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.

You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. See **Other Information** below.

CC BY CC BY 4.0

ATTRIBUTION 4.0 INTERNATIONAL

Deed

Canonical URL : <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

What is the Right License?

- We want our Open Data to be usable for everyone
- We want people to cite us when they use our Open Data
- What's the right License to choose?

CC BY CC0 1.0

CC0 1.0 UNIVERSAL

Deed

Canonical URL : <https://creativecommons.org/publicdomain/zero/1.0/>

[See the legal code](#)

No Copyright

- The person who associated a work with this deed has **dedicated** the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.

You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. See **Other Information** below.

Citing ATLAS

The public datasets are accessible on the [CERN Open Data portal](#), under [Creative Commons CC0 license](#).

Any paper published using these data should cite the corresponding DOI of the datasets. The citation should be similar to this:

ATLAS Collaboration (2020). *ATLAS simulated samples collection for jet reconstruction training, as part of the 2020 Open Data release*. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.L806.5CKU

A few additional useful papers for citation are provided below. Please ensure that the ATLAS Collaboration is acknowledged as well. Our preferred acknowledgement is:

"We acknowledge the work of the ATLAS Collaboration to record or simulate, reconstruct, and distribute the Open Data used in this paper, and to develop and support the software with which it was analysed."

For citing the ATLAS Detector:

ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider." JINST 3 (2008) S08003. DOI:10.1088/1748-0221/3/08/S08003.

⚠ Disclaimer

Neither ATLAS nor CERN endorse any works, scientific or otherwise, produced using ATLAS Open Data.



- [Athena](#) is already open-source
- cvmfs distributions available
 - Great for folks who know what this is
- [Containers are available](#)
 - Great for tutorials, already in use
 - For some applications these are **huge**
 - Through cvmfs we currently distribute **138 GB of PDFs** for event generation
- For **notebooks**, [lots of resources](#)
 - Also [binder](#) and so on to run on
- Making public an **ntuple maker**
 - Based on our analysis tutorial
 - Exactly the example used to create the education and outreach open data
- **Analysis code** is ~never public (!)
 - All examples are custom — no real code has been shared to date
 - (It is preserved, [sometimes well](#))

ATLAS releases new open software

20 November 2020 | By [Mariana Velho](#), [Katarina Anthony](#)

The ATLAS Collaboration has just released a collection of 200 software packages that make up the Trigger and Data Acquisition System (TDAQ). With this new release, most ATLAS software is now open – reinforcing the Collaboration's ongoing commitment to open science.

ATLAS' first major step into open software was the release of Athena in November 2018. Athena (available [here](#)) is a collision-event processing software for the ATLAS experiment for event reconstruction, detector simulation, and other key tasks. It is licensed for data

More on the web

Jupyter Notebooks

Uproot

[Higgs to ZZ](#)

This notebook uses ATLAS Open Data to show you the steps to rediscover the Higgs boson yourself! You will discover the Higgs boson decaying into a pair of Z bosons, which are in turn decaying into a lepton-antilepton pair each.

Physics: ★
Coding: ★★
Time: ★★★
[launch binder](#)

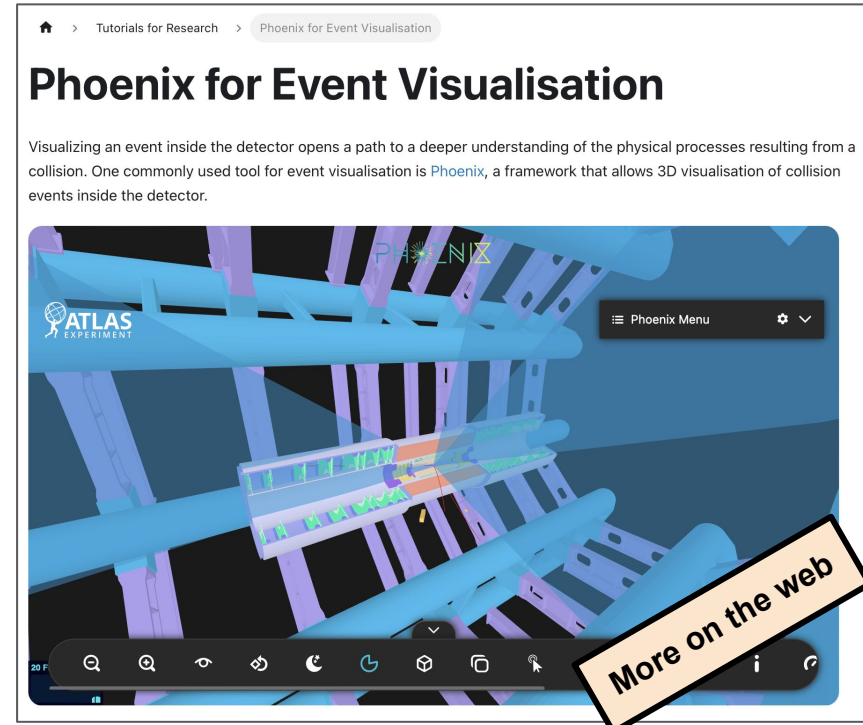
[Higgs to ZZ with Boosted Decision Tree](#)

This notebook uses ATLAS Open Data to show you the steps to apply a Machine Learning approach to discover the Higgs boson yourself! You will discover the Higgs boson decaying into a pair of Z bosons, which are in turn decaying into a lepton-antilepton pair each, and you will learn how to use a boosted decision tree (BDT) like a professional.





Physics: ★
Coding: ★★
Time: ★★
[launch binder](#)

More on the web

- With Phoenix we have an in-browser interactive event display
 - Detector geometry exploration tool!
- Providing lots of examples for users to play with (SM and BSM, MC and real)
- Light python script to go from PHYSLITE or education ntuple to Phoenix input
 - Users can make event displays out of any event they wish!
 - Minimal infrastructure — don't need any ATLAS software or containers

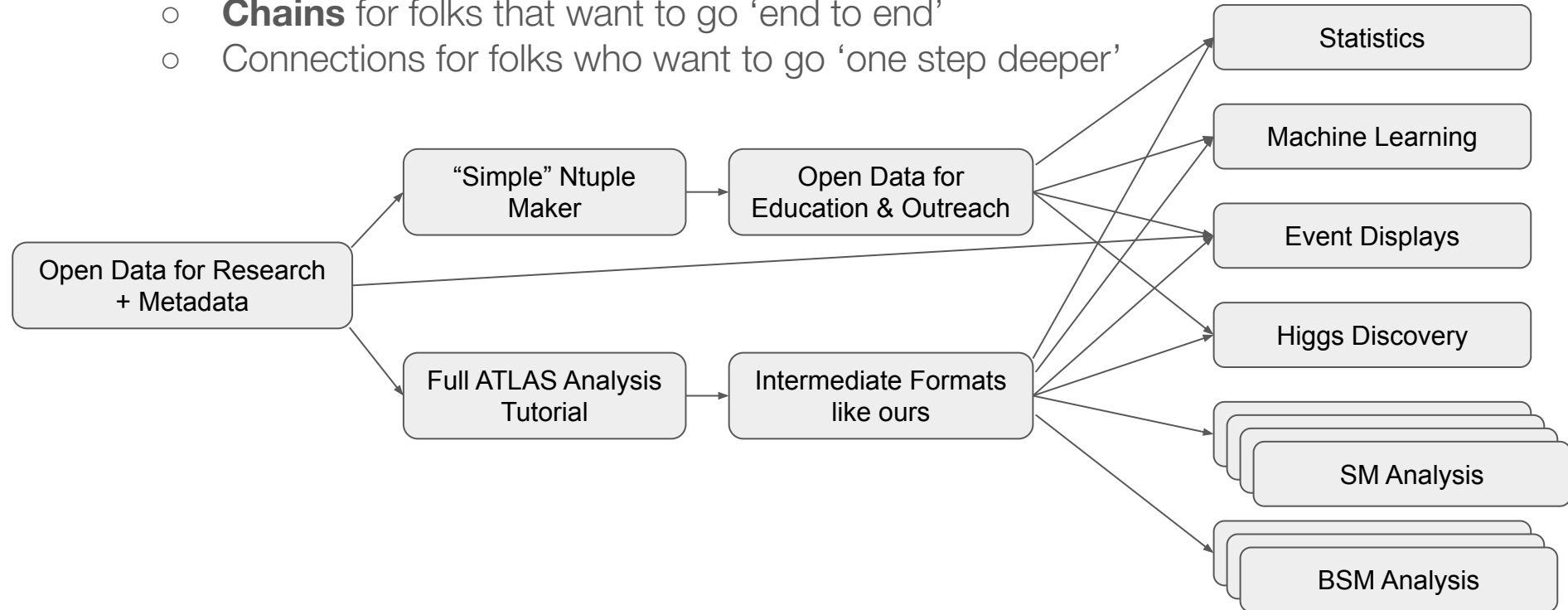


- We've constructed “paths” through the Open Data for different kinds of users

 Quick start The quickest way to start learning with ATLAS Open Data.	 Deep Dive For extended use. Let's dive into what ATLAS has to offer!
 Researchers Toolkit Detailed information and resources for researchers	 Online Data Analyser Explore ATLAS Open Data at a glance!

More on the web

- We've constructed "paths" through the Open Data for different kinds of users
- The eventual goal is to have a **web** of paths that satisfy many users and interests
 - Independent **modules** for specific learning objectives (hours – months)
 - **Chains** for folks that want to go 'end to end'
 - Connections for folks who want to go 'one step deeper'



Documentation: How to Write an Analysis

- Jupyter Notebooks are **fantastic** for fast-uptake documentation and examples
- Once they want something sufficiently complex, there's no (easy) option:
 - We send users to our **real analysis software tutorial**
- Have not yet found a great way to bridge these two worlds
 - Perhaps this will happen naturally if our analyses evolve to be more notebook-like?

How to use a PHYSLITE file

For the ATLAS Open Data for Research, we have released hundreds of datasets in PHYSLITE format. In this notebook we are going to show you how to access the event and variables, and how to do a basic analysis using `uproot` and `awkward` arrays.

```
In [1]: # Import basic libraries
import copy # copy variables
import os # manage paths

import uproot # use of root files
import awkward as ak # nested, variable sized data
import vector # lorentz vectors
vector.register_awkward()
import matplotlib.pyplot as plt # plotting
import tqdm # progress bars
```

In this notebook we are using a dataset from `mc20_13TeV.410470.PhPy8EG_A14_ttbar_hdamp258p75_nonallhad.deriv.DA0D_PHYSLITE.e6337_s3681_r13167_p5855`. To run this notebook locally, please download a dataset from this container and substitute the file and path in the `filename` variable below.

```
In [2]: # mc20_13TeV.410470.PhPy8EG_A14_ttbar_hdamp258p75_nonallhad.deriv.DA0D_PHYSLITE.e6337_s3681_r13167_p5855
filename = "DA0D_PHYSLITE.34865537_000312.pool.root.1"
```

Read PHYSLITE with uproot

We can open a `TFile` using `uproot.open`. To check the `TTree` objects in the file we use the `.keys()` method.

```
In [3]: print('TTree objects inside the ROOT file:')
for ii in uproot.open(filename).keys():
    print('-',ii)
```

```
TTree objects inside the ROOT file:
- ##Params;3
```

More on the web

ATLAS Analysis Software Tutorial



Welcome to the **ATLAS Analysis Software Tutorial** pages.

This is the portal to the ATLAS analysis software tutorial held multiple times throughout the year.

This tutorial is aimed at (new and old) members of the ATLAS collaboration interested in learning the basics of ATLAS software and the latest physics analysis tools. There are several introductory lectures aimed at introducing the topics.

Tutorial Week

Follow a week-long analysis software tutorial with lectures and hands-on exercises. The new tutorial format (from November 2022) is designed to be done in small groups and to follow an example ATLAS analysis from start to finish. The introductory material and hands-on exercises are kept available online, so if you cannot attend the tutorial week, you can still work through the material.

The curriculum, recorded lectures (for self-guided/asynchronous study), and links to the material are available at:

Tutorial material

More on the web

Documentation: The Really Hard Part

Setting Uncertainties

One of the most important parts of any data analysis is the inclusion of proper uncertainties. Uncertainties help quantify the reliability and precision of a conclusion obtained from data.

When comparing detector data to simulations, you can see a difference that might seem significant. However, whether that difference is interesting or important requires understanding uncertainties. Agreement within uncertainties implies that the observed and predicted values are consistent. If a number is measured to be 1000 and it was predicted to be 2000 ± 1000 , then the measurement and prediction agree. Despite the measurement appearing far from the prediction, the large uncertainty range indicates that the prediction is not very precise, allowing for agreement.

Similarly, it is important not to misinterpret agreement that is better than the uncertainty suggests. If a number is measured to be 1000 and the prediction was 1000 ± 500 , that does not mean that the true value will be 1000. A more precise model might give a prediction of 600 ± 100 , which would be in consistent with the original prediction, but would no longer agree with the measurement.

A key part of scientific training is understanding when a difference between a prediction and an observation is meaningful and significant, and that comes down to understanding uncertainties.

Why Consider Uncertainties?

In ATLAS analyses we consider uncertainties for several reasons:

- Accurate Parameter Estimation:** To get reliable estimates of the parameters of interest (POIs) – boson couplings or the top quark mass, we need to account for all sources of uncertainty. Ignoring uncertainties can lead to biased estimates and incorrect conclusions.
- Robust Hypothesis Testing:** In testing theoretical models against experimental data, uncertainties ensure that discrepancies between the observed data and theoretical predictions are not mistakenly attributed to new physics or phenomena, instead that they are correctly identified as arising from uncertainties in the experimental or theoretical setup.

More on the web



- The **hardest part** of an analysis is understanding and calculating systematic uncertainties
- Explaining how to do this in an approachable way is **extraordinarily difficult** and **important**
- **Evergreen documentation** of concepts
 - Useful for our students as well!
 - Can be integrated into our tutorials
- **Technical documentation** for code
 - **Momentarily** matches internal documentation until we move on (except CERN-specifics, Grid use...)
 - Needs to be **fool proof** to avoid science problems
 - Good examples are a **huge** help here
- Related documentation we haven't written yet: what you **cannot** do
 - Things our systematics don't cover
 - Things our datasets / simplification don't permit

Resources

Downloading all the available Open Data requires significant resources. For those who wish to tinker, but might not have the computing resources to hand, there are a few options.

- There is a [cluster at the University of Nebraska-Lincoln](#) on which an account can be requested. It's possible to authenticate with Google, for example (no institutional affiliation is required).
- Google offers [cloud resources](#), with free credit for new users.

On these resources, we recommend installing dependencies via a terminal:

```
pip3 install --user jupyterlab matplotlib tqdm xrootd zstandard uproot==5.1.2 awkward==2.5.0 vector-datasource-client[xrootd]
```

More on the web

Opendata Analysis Facility @ T2_US_Nebraska

Useful Links

[Coffea-Casa Support Page](#) [Coffea-Casa Docs](#)

News

Watch here for announcements!

[Register for access](#)

Authorized Users Only:
[Sign in with OAuth 2.0](#)

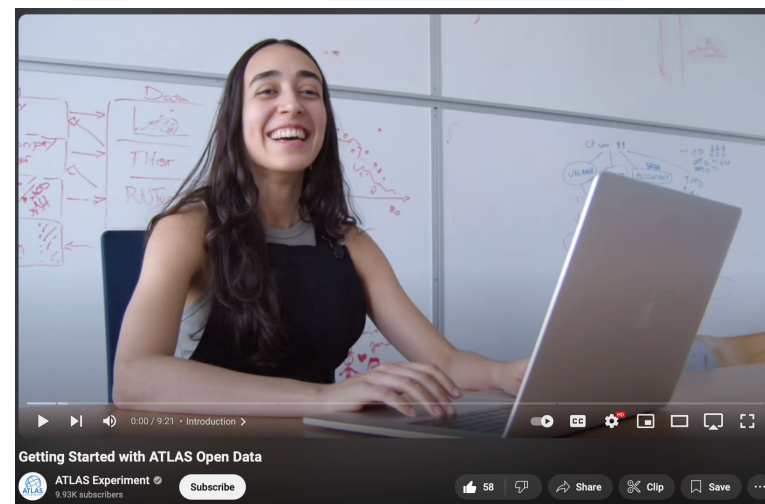
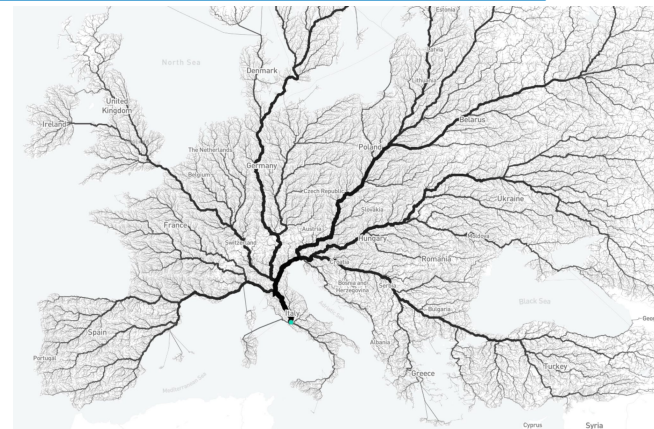
More on the web



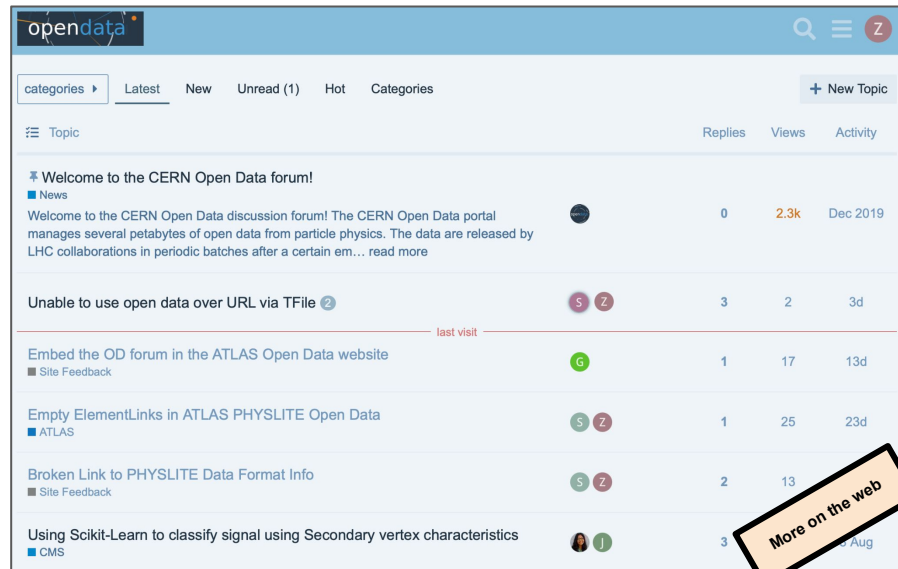

- **Who pays** for resources?
- CERN's Open Data Portal, storage, support provided by CERN IT
 - Is this a 'host lab responsibility'? Long-term?
- CERN only providing open *storage*
 - CERN users have access to standard resources (notebooks, etc)
 - Security and abuse concerns with providing CPU to 'any' user
- CPU is *critical* to **equitable** open data
 - We cannot expect people to have network, local storage, local CPU, ...
 - Notebooks with a web interface might work?
- A few sites are trying to provide CPU
 - [UNL](#) has done a great job
- Trying to provide instructions for commercial cloud resources (second copy of data too?)

How are Users Finding Us?

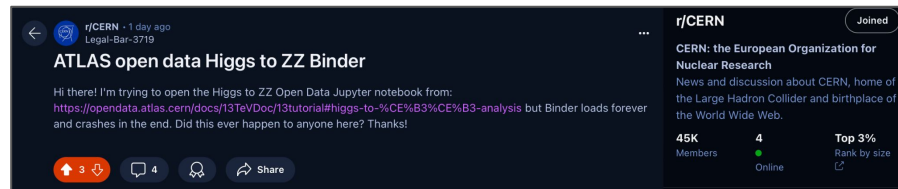
- How many entry points do we have?
 - [ATLAS news articles](#)
 - [CERN news articles](#)
 - [YouTube videos](#), Tweets, Reels, ...
 - [ATLAS Open Data website](#)
 - [Individual Open Data records](#)
 - ...
- Documentation that **can't be found** is wasted effort!
 - Duplicate documentation is painful and risks inconsistencies and other major problems...
- **Single** [Open Data Portal 'entry' record](#)
- Everything redirecting to the [ATLAS Open Data Website](#)



- Open Data Forum at CERN
 - Social media login allowed
 - Harder to track who's there to help (who is tracking the relevant issues)
- Issues / Pull Requests on GitHub
 - GitLab requires CERN account
- Several egroups dedicated to help
 - Within ATLAS and within CERN
 - Harder to track answers — Did we cc an expert? Did something get answered off-list?
- Asking expert users to come to our meetings and give feedback
 - IRIS-HEP summer students, e.g.
- Have to meet the users where they are!
 - If someone asks for help on Reddit, we should find it and answer

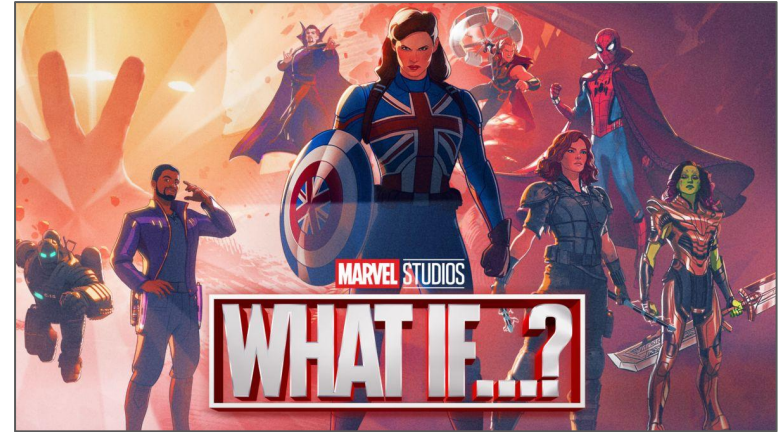


The screenshot shows the 'open data' forum interface. It features a navigation bar with 'categories', 'Latest', 'New', 'Unread (1)', 'Hot', and 'Categories'. Below the navigation bar, there is a table of forum topics. The first topic is 'Welcome to the CERN Open Data forum!' with 0 replies, 2.3k views, and a date of Dec 2019. The second topic is 'Unable to use open data over URL via TFile' with 3 replies, 2 views, and a date of 3d. The third topic is 'Embed the OD forum in the ATLAS Open Data website' with 1 reply, 17 views, and a date of 13d. The fourth topic is 'Empty ElementLinks in ATLAS PHYSLITE Open Data' with 1 reply, 25 views, and a date of 23d. The fifth topic is 'Broken Link to PHYSLITE Data Format Info' with 2 replies, 13 views, and a date of 13d. The sixth topic is 'Using Scikit-Learn to classify signal using Secondary vertex characteristics' with 3 replies and a date of 1 Aug. A yellow callout box with the text 'More on the web' is overlaid on the bottom right of the forum screenshot.




The screenshot shows a Reddit post from the r/CERN subreddit. The post title is 'ATLAS open data Higgs to ZZ Binder'. The post content reads: 'Hi there! I'm trying to open the Higgs to ZZ Open Data Jupyter notebook from: <https://opendata.atlas.cern/docs/13TeVDoc/13tutorial#higgs-to-%CE%B3%CE%B3-analysis> but Binder loads forever and crashes in the end. Did this ever happen to anyone here? Thanks!'. The post has 3 upvotes, 4 comments, and 1 share. The subreddit information shows 45K members, 4 online, and a top 3% rank by size.

- Significant discussion in the collaboration about effort for support — “What ifs”
 - What if someone requests an SM sample?
 - What if someone wants to run their own (BSM?) MC?
 - What if someone needs help with a tool?
 - What if someone publishes a paper that requires a response?
- Agreed on ‘best effort’ volunteer support
- Point of concern to be watched
 - Need to ensure that we don’t need Sherpas to guide all Open Data users
 - Shouldn’t *waste* significant effort, but it’s ok to *spend* effort that helps us too
- Hoping to connect with expert users via Short Term Associations
 - Become ‘insiders’ for specific projects



The ATLAS Mountain Range



opendata
CERN

ATLAS top tagging open data set with systematic uncertainties

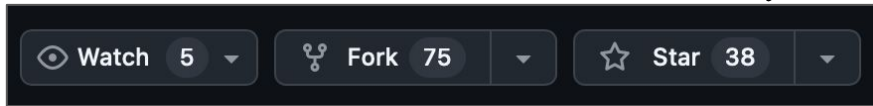
ATLAS collaboration

Cite as: ATLAS collaboration (2024). ATLAS top tagging open data set with systematic uncertainties. CERN Open Data. DOI:10.7483/OPENDATA.ATLAS.SOAY.LABE

There is one publication referring to these data

Dataset Derived Datascience ATLAS CERN-LHC

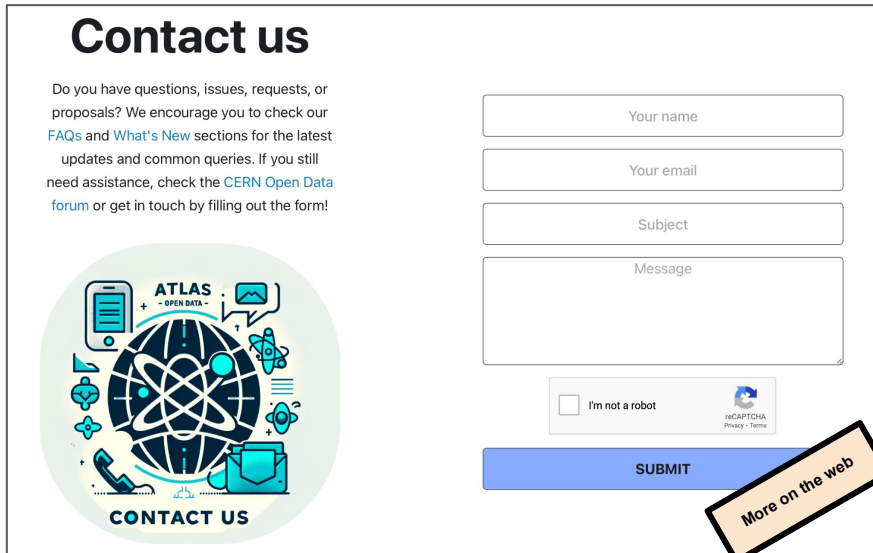
More on the web



Watch 5

Fork 75

Star 38



Contact us

Do you have questions, issues, requests, or proposals? We encourage you to check our [FAQs](#) and [What's New](#) sections for the latest updates and common queries. If you still need assistance, check the [CERN Open Data forum](#) or get in touch by filling out the form!

Your name

Your email

Subject

Message

I'm not a robot

reCAPTCHA
Privacy - Terms

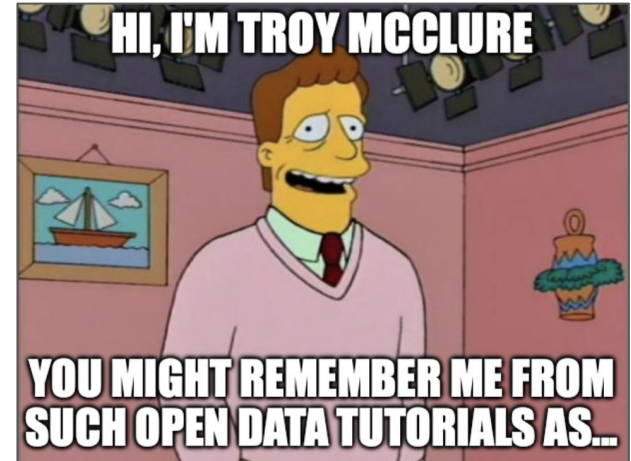
SUBMIT

CONTACT US

More on the web

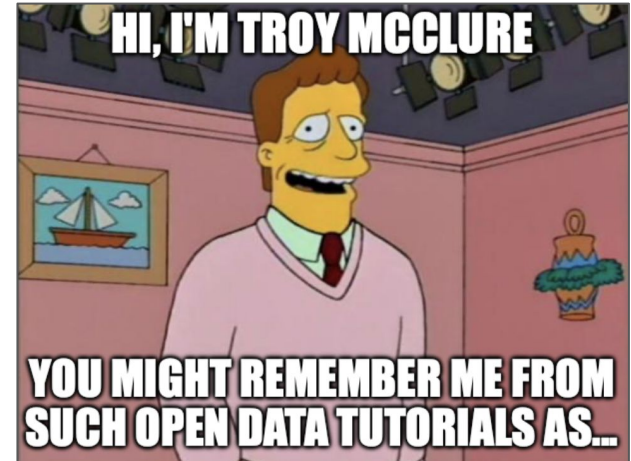
- **Research** output KPIs
 - Now tracking citations via DOI and URL
 - Indications this is not well reported ([tools?](#))
- **Repository / code** tracking KPIs
 - Watches / forks / stars
- **Website** KPIs
 - Clicks, website search results
- **Direct Data** KPIs
 - Downloads, remote reads, reads from eos
 - These require pretty good infrastructure monitoring
 - Useful to know how many people are using the [opendata client](#) software
- **'Experience'** KPIs
 - Surveys for usage and satisfaction, user feedback forms, attendance at events
- **Some day: CPU** KPIs
 - Integrating CPU monitoring from 'friendly' sites to understand CPU used for open data
- **Must track educational** usage as well!

- Who will store the data in 2050?
 - Does this scale for the lifetime of the experiment?
 - What about the software? Websites? Documentation? Examples?
- Where physically are they stored?
 - **All** of them? Even the little example mini-formats?
- Who has rights to the repos?
 - Ownership? At CERN? Outside?
 - What about **after** the experiment?
- Can we reproduce them? Add to them?
 - If someone wants “just one more” MC dataset? 15 years from now?
- How do we release **more** data?
 - New format... new release of everything? Keep what’s there?
 - Update examples? Abandon them? Delete them?
- Are old tutorials / recordings useful? Watched?
 - Just the most recent?



A Playground for Future-Proofing

- Who will store the data in 2050?
 - Does this scale for the lifetime of the experiment?
 - What about the software? Websites? Documentation? Examples?
- Where physically are they stored?
 - **All** of them? Even the little example mini-formats?
- Who has rights to the repos?
 - Ownership? At CERN? Outside?
 - What about **after** the experiment?
- Can we reproduce them? Add to them?
 - If someone wants “just one more” MC dataset? 15 years from now?
- How do we release **more** data?
 - New format... new release of everything? Keep what's there?
 - Update examples? Abandon them? Delete them?
- Are old tutorials / recordings useful? Watched?
 - Just the most recent?



What's Next

- Monitoring and watching
 - Users will let us know what they think of our open data in comparison to other offerings
- Heavy Ion open data coming soon
 - Required by / following the Open Data Policy
 - Different format that supports different analyses
- Filling in more parts of the documentation chain
- More examples, integrating **your** examples
 - We would love to build a library in the style of Rivet with examples developed in projects around the world
 - Maybe also **our own** analysis? (TBD)
- Beginning to plan a workshop / hackathon

🏠 > Tutorials for Research > Community Contributions

Community Contributions

Here we gather various projects and analyses created using our open data for research. We believe in the power of collaboration and the insights that can emerge from diverse perspectives. If you've used our open data for something cool, we would love to hear about it! Please share your work with us through the [contact us](#) form. Your contributions can inspire others and help to show the potential of open data.

A full list of uses of ATLAS Open Data can be found on [INSPIRE-HEP](#).

Notebooks

HOW THE SCIENTIFIC PYTHON ECOSYSTEM HELPS ANSWERING FUNDAMENTAL QUESTIONS OF THE UNIVERSE

By Vangelis Kourlitis

[launch](#) [binder](#)

More on the web

- Rivet home
 - Contour
 - Professor
 - YODA
 - MCPests
 - AGILE
- Downloads
- Analysis
 - Standard analyses
 - Analysis changinglog
 - Writing an analysis
- Analysis coverage & wishlists
 - General
 - No searches/HL
 - Searches
 - Heavy ion
- Documentation
 - Manual & talk links
 - Getting started / tutorials
 - Rivet via Docker
 - Changinglog
 - Example code/API docs
- Source code
- Contact

[Follow @rivetio](#)

Rivet analysis coverage

Rivet analyses exist for 1838/6446 papers = 29%. 261 priority analyses required.

Total number of Inspire papers scanned = 10889, at 2024-08-08

Breakdown by identified experiment (in development):

Key	ALICE	ATLAS	CMS	LHCb	Forward	HERA	e ⁺ e ⁻ (>12 GeV)	e ⁺ e ⁻ (≤12 GeV)	Tevatron	RHIC	SPS	Other
Rivet wanted (total):	380	477	582	205	19	478	647	54	1116	523	59	90
Rivet REALLY wanted:	54	62	98	15	0	15	1	0	9	2	5	0
Rivet provided:	44/124 = 10%	212/689 = 31%	138/897 = 19%	71/776 = 26%	19/25 = 34%	37/513 = 7%	243/390 = 27%	996/1050 = 95%	81/1177 = 5%	11/534 = 2%	15/74 = 20%	21/111 = 19%

[Show graph](#) [Show tooltip](#) [Hide list](#)

ALICE
ATLAS
CMS
LHCb
Forward
HERA
e⁺e⁻ (>12 GeV)
e⁺e⁻ (≤12 GeV)
Tevatron
RHIC
SPS
Other

ATLAS: Search for heavy Majorana neutrinos in *ee* and *etp* final states via WW scattering in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector

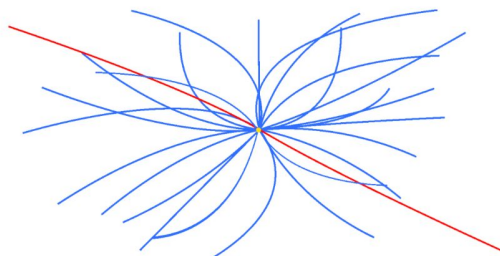
Inspire: 2776861 · arXiv: 2403.15016 (pdf) · DOI/journal: 10.1016/j.physletb.2024.138865 · CDS: 2892200

Report ID: CERN-EP-2024-033

ATLAS Open Data for Research — CHEP 2024 — 21 Oct 2024 — Zach Marshall

26

ATLAS Open Data



High Energy Physics data for everyone.

For Education

To provide data and tools to high school, undergraduate and graduate students, as well as teachers and lecturers, to help educate them and exercise in physics analysis techniques used in experimental particle physics.

For Research

To provide researchers with high-quality data recorded by the ATLAS detector, enabling them to conduct state of the art analyses in particle physics.



GET STARTED

Thank you!



- Publications will not be reviewed by, endorsed by, or marked as ATLAS
 - We will, however, provide instructions for how to cite and acknowledge our work
- “ATLAS expects its members to publish as part of the collaboration but does not forbid them from publishing on open data”
- Composed an “appropriate set of MC samples” in PHYSLITE format
 - Including both baseline and variation samples for systematics, and some signals
 - Luminosity of $\sim 2x$ the data to start (rounding to the nearest file)