The First Release of ATLAS Open Data for Research

CHEP 2024

21 October 2024

Zach Marshall (LBNL) on behalf of the ATLAS Collaboration

Open Data for Research

Let’s unpack this

Open Data for Research

Public, licensed, documented, FAIR

Understandable, light, useful format

Can be used for scientific publications

Let's unpack this

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**

A Brief Introduction to Open Data

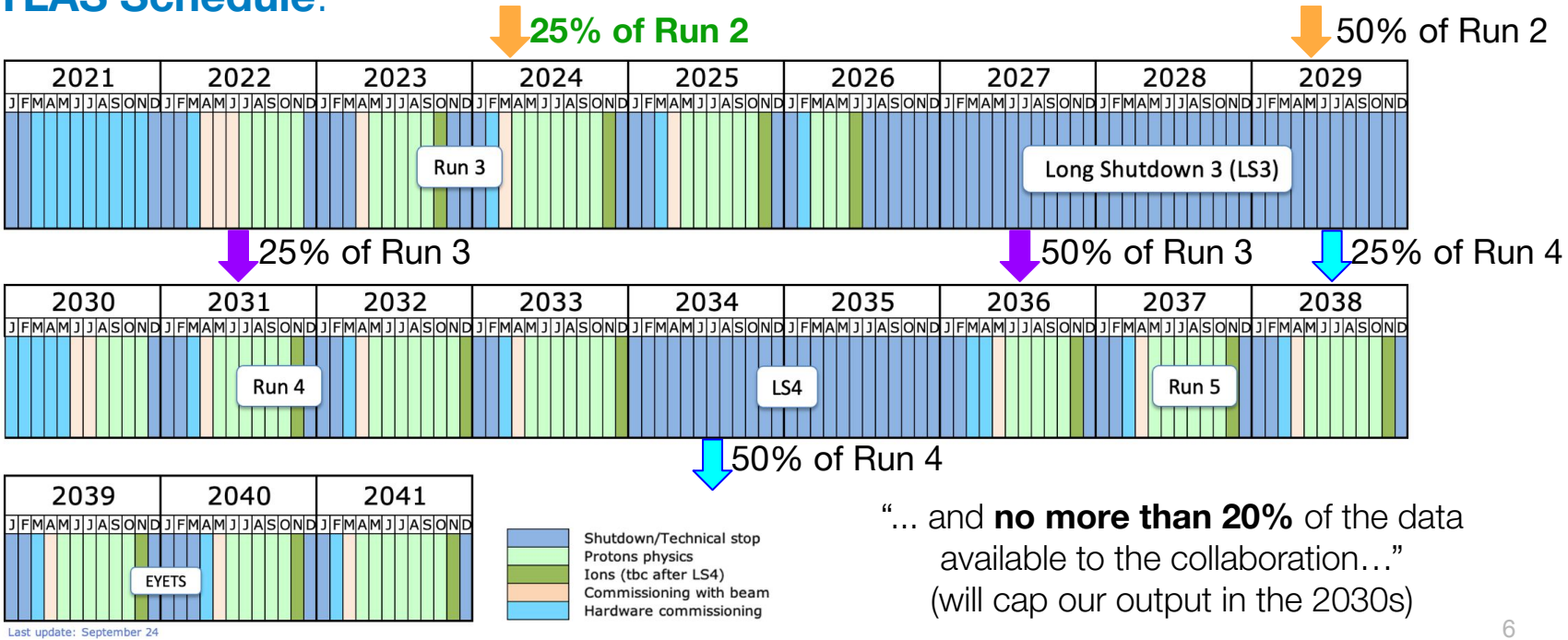
- The Data Preservation for HEP (DPHEP) Collaboration defined four levels of open data

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).

Open Data for Research

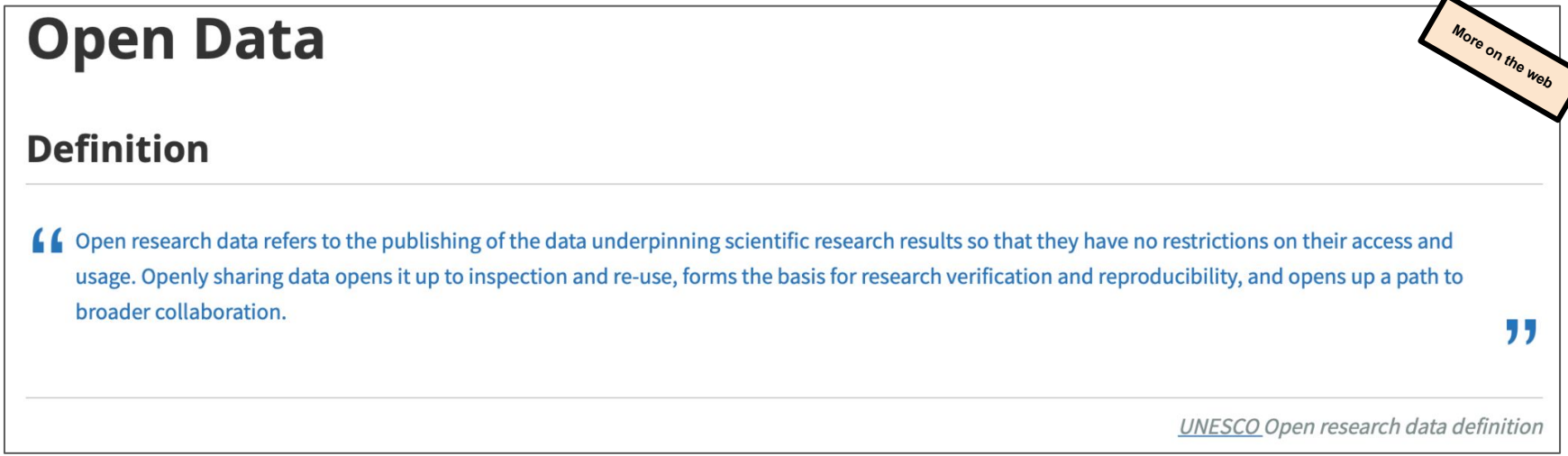
- We have lots of **bespoke datasets** for **targeted research applications**
  - Top jet tagging, Fast Simulation training, BSM BDT training, …
  - Kaggle Challenges: Higgs boson ML Challenge, TrackML Challenge
- Four large LHC experiments agreed to release data for **general research use**
- **ATLAS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



"... and **no more than 20%** of the data available to the collaboration…"
(will cap our output in the 2030s)

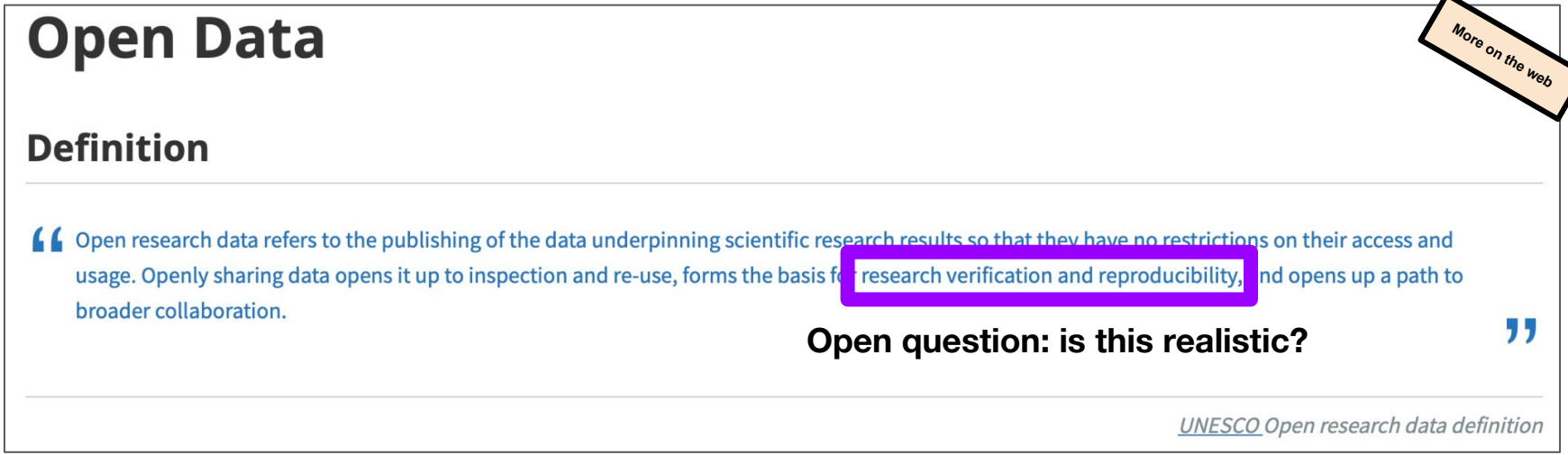
What's the Goal?

- 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**

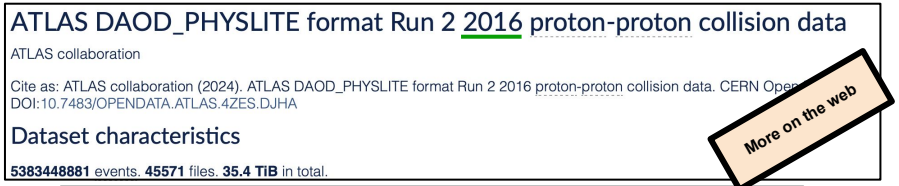


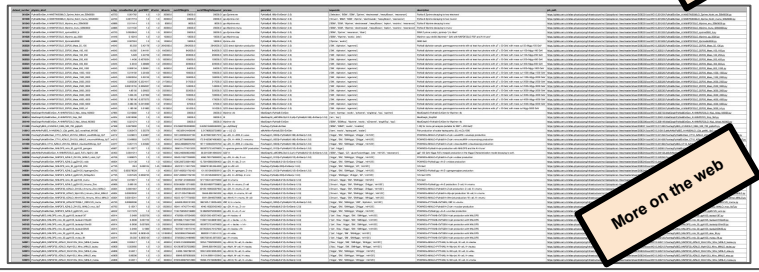
What's the Goal?

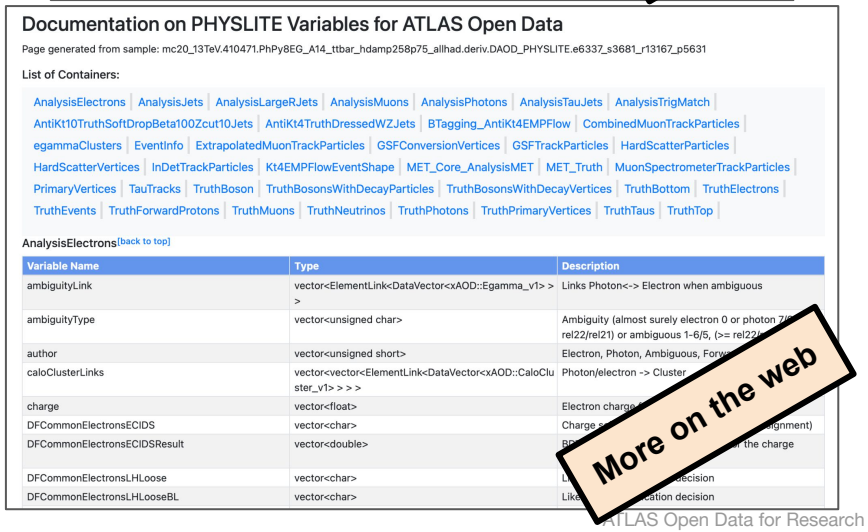
- All the goals of Outreach and Education open data
  - Democratic access, education and inspiration of future scientists
- These Open Data meet research standards — can be used for scientific papers
- Would love for excited researchers to use this as a step towards collaboration
  - Particularly if 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**



What did we Release?







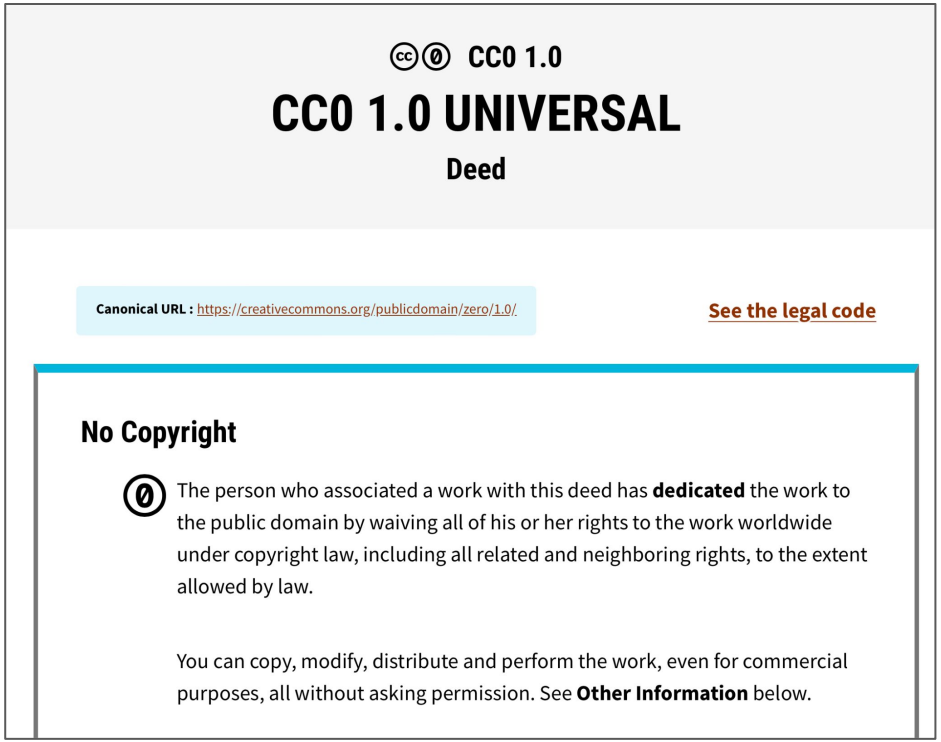
- 2015+2016 Run 2 pp collision data
  - 45 TB of data, 6.3 kB/event, 7.1B events, 55k files in ~300 runs
  - 20 TB of MC, ~10 kB/event, 2B events, 16k files in ~300 MC datasets
- Explanation of our nomenclature
- Giant tables of **metadata**
  - Cross sections, k-factors, filters / efficiencies, processes, 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?

- 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?



CC0 1.0

CC0 1.0 UNIVERSAL

Deed

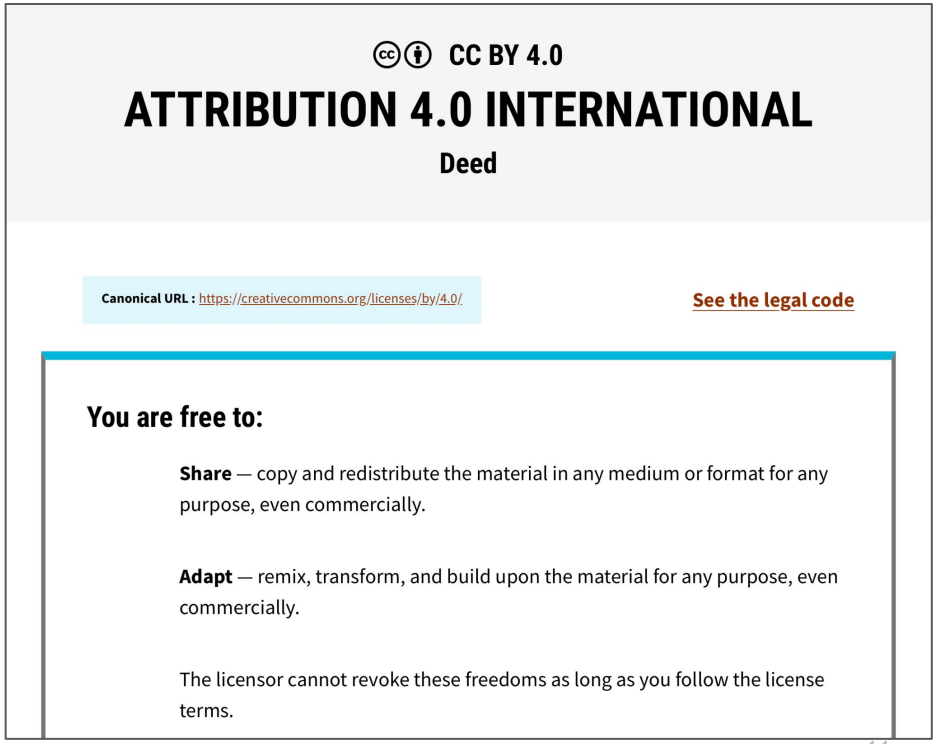
Canonical URL : https://creativecommons.org/publicdomain/zero/1.0/

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 4.0

ATTRIBUTION 4.0 INTERNATIONAL

Deed

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

See the legal code

You are free to:

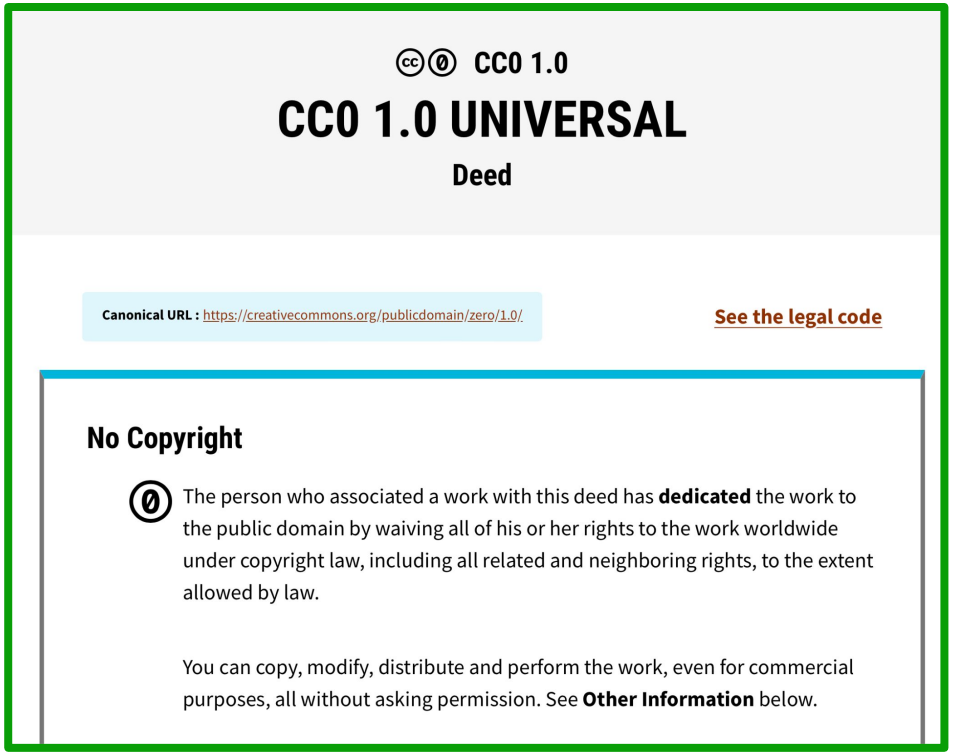
Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

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

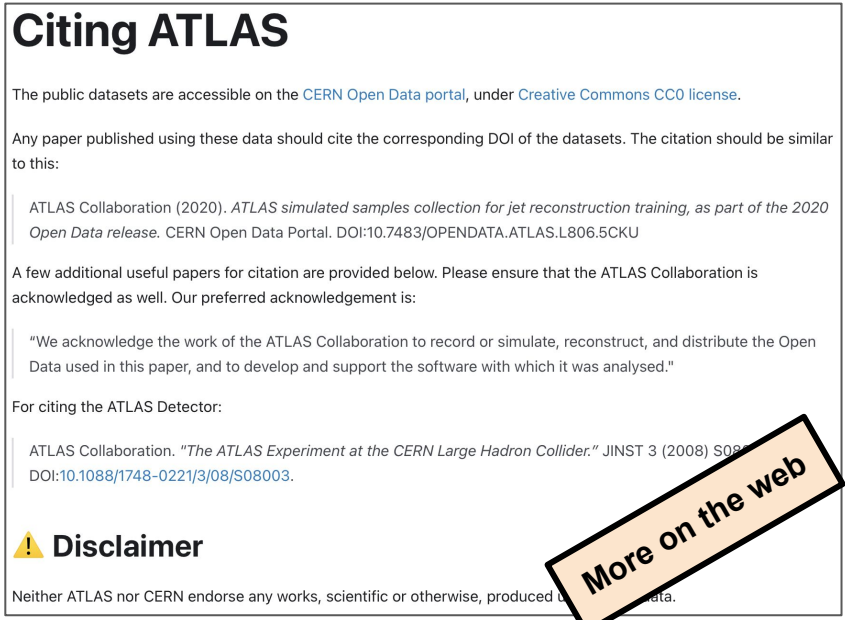
The licensor cannot revoke these freedoms as long as you follow the license terms.

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?

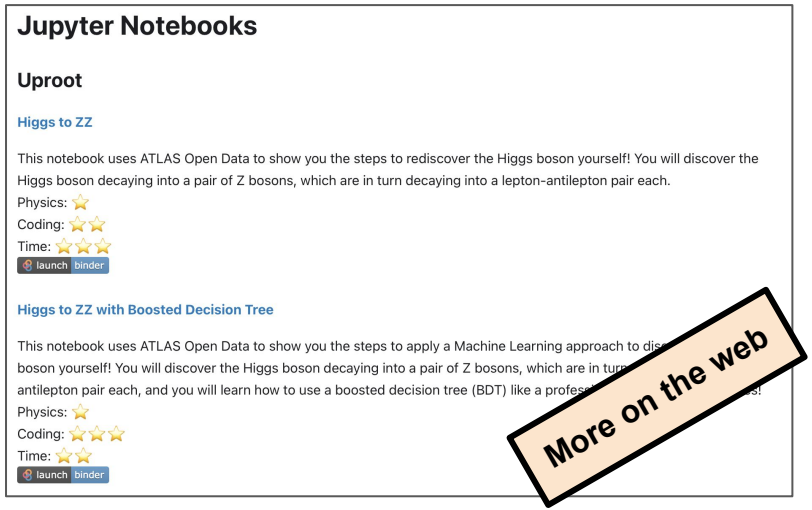


+



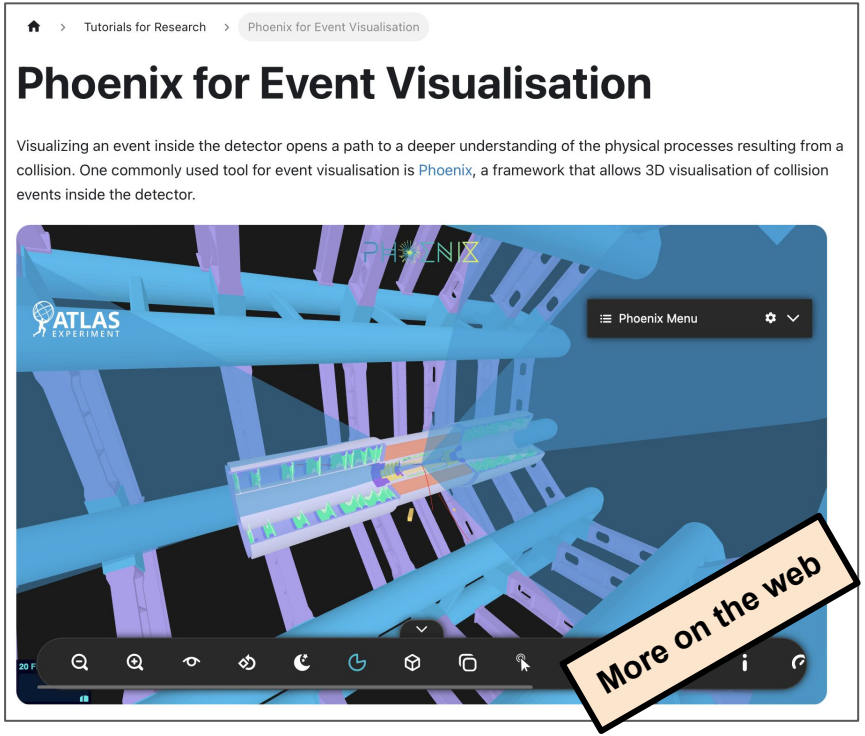
Software to Use Them

- 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)



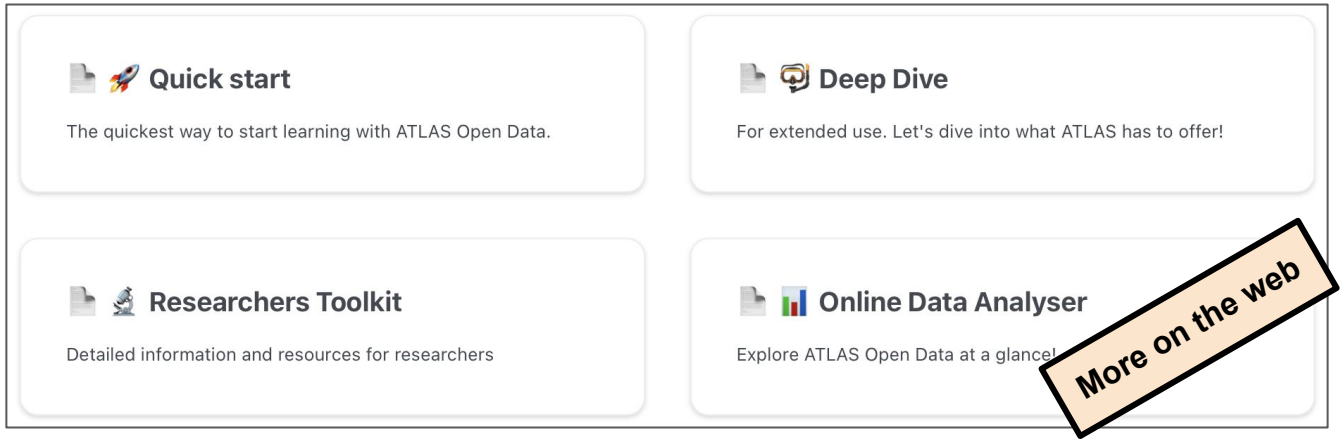
Event Displays

- With Phoenix we have an in-browser interactive event display
  - Detector geometry exploration too!
- 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



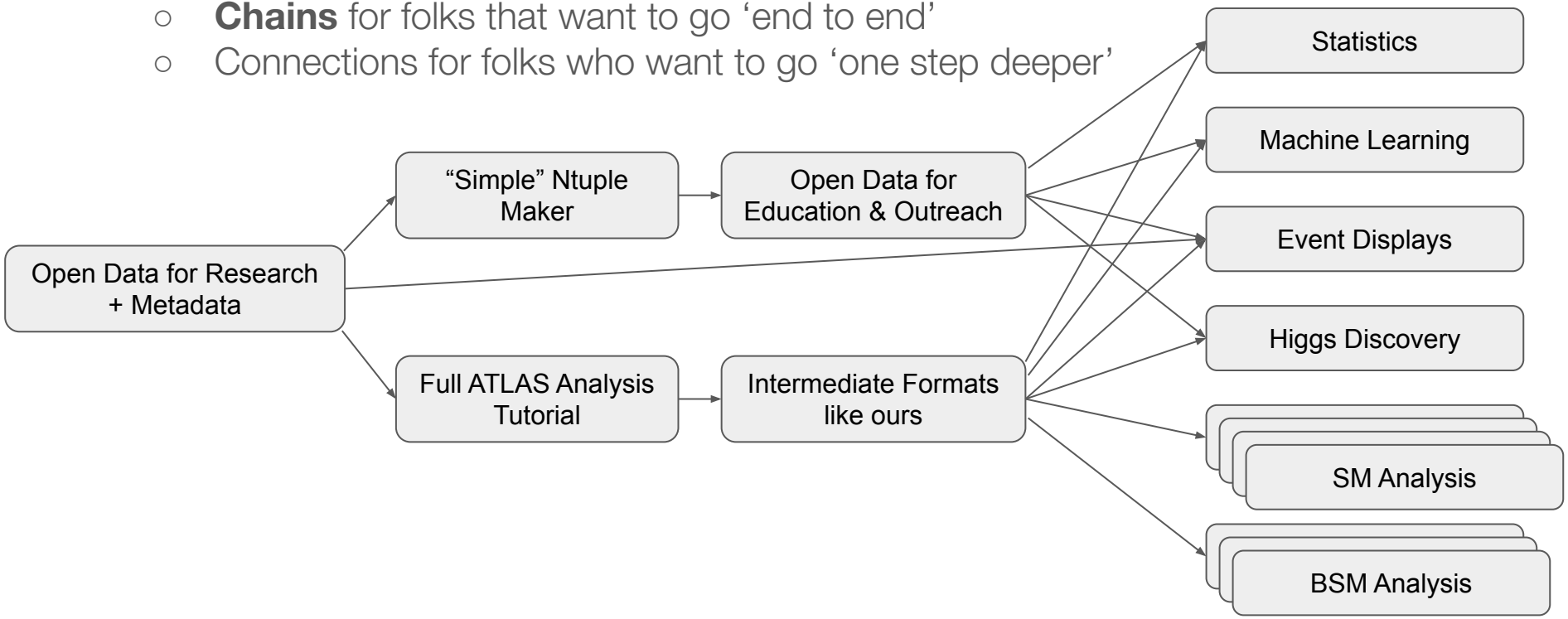
Documentation!

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



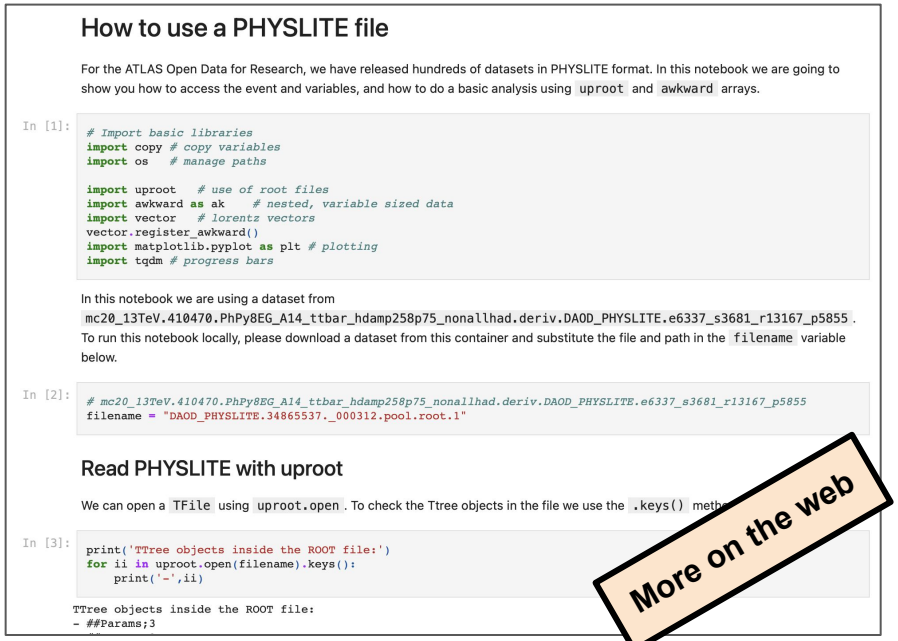
Documentation!

- 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.DAOD_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.DAOD_PHYSLITE.e6337_s3681_r13167_p5855
filename = "DAOD_PHYSLITE.34865537._000312.pool.root.1"
```

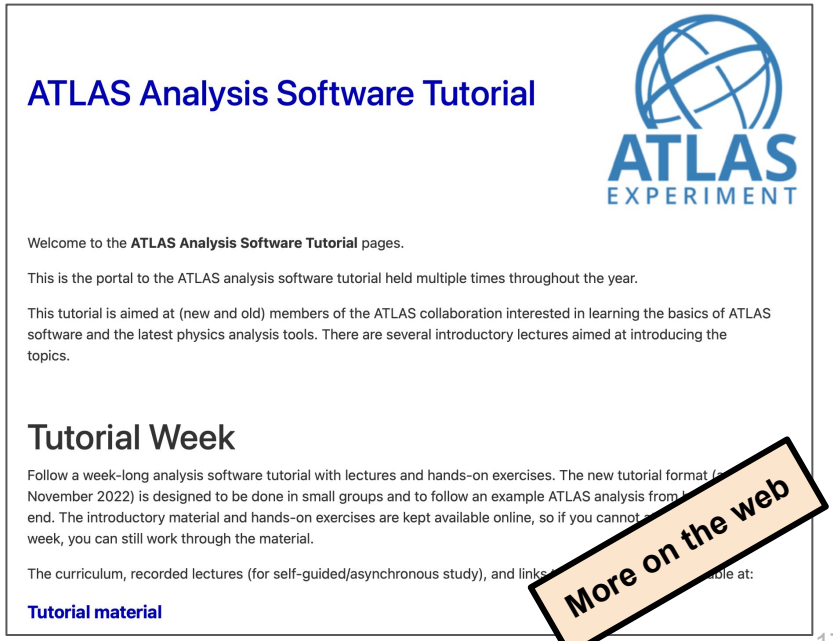
Read PHYSLITE with uproot

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

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

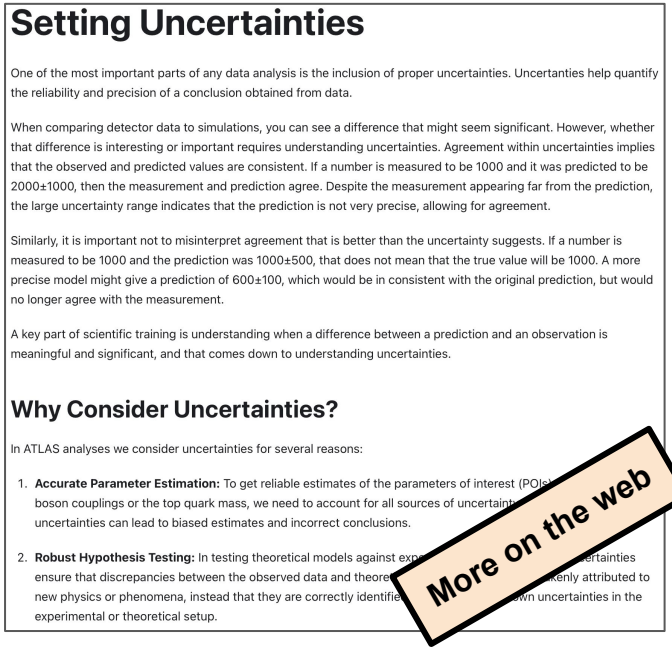
Follow a week-long analysis software tutorial with lectures and hands-on exercises. The new tutorial format (
November 2022) is designed to be done in small groups and to follow an example ATLAS analysis from
end. The introductory material and hands-on exercises are kept available online, so if you cannot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

The curriculum, recorded lectures (for self-guided/asynchronous study), and links
ble 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. Uncertanties 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:

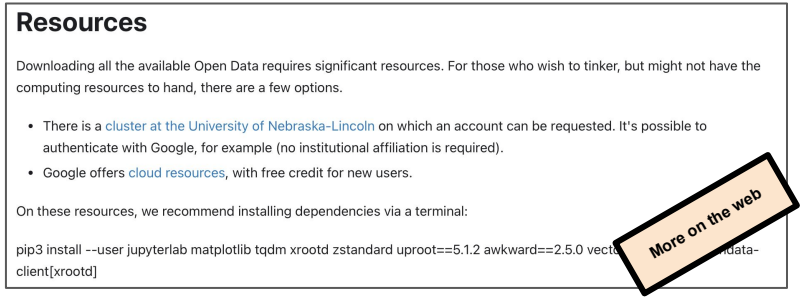
1. **Accurate Parameter Estimation:** To get reliable estimates of the parameters of interest (POI
   boson couplings or the top quark mass, we need to account for all sources of uncertainty
   uncertainties can lead to biased estimates and incorrect conclusions.
2. **Robust Hypothesis Testing:** In testing theoretical models against exp
   ensure that discrepancies between the observed data and theore
   new physics or phenomena, instead that they are correctly identifie
   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 for Open Data



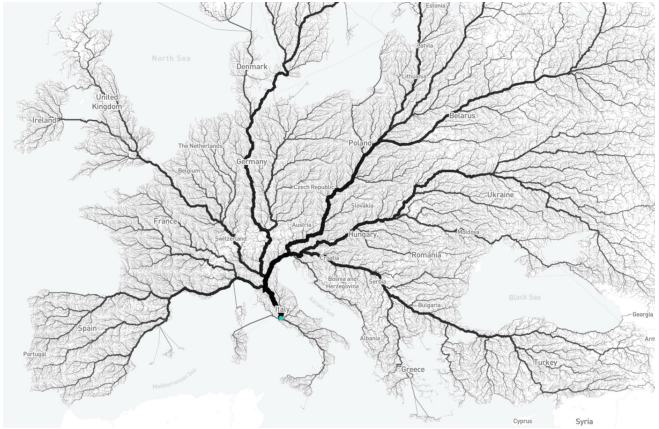
Google Cloud

aws

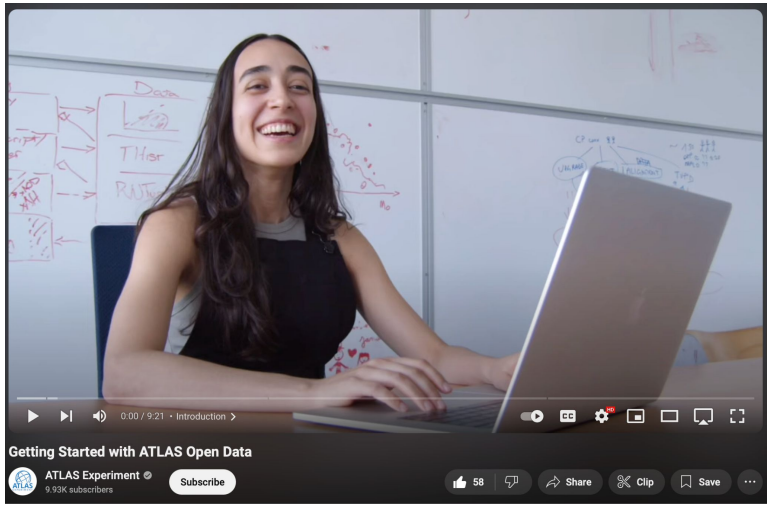
- **Who pays** for resources?
- CERN's Open Data Portal, storage, support provided by CERN IT
  - Is this a 'host lab responsibility'? Long-term?
- CERN only providing open *storage*
  - CERN users have access to standard resources (notebooks, etc)
  - Security and abuse concerns with providing CPU to '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…

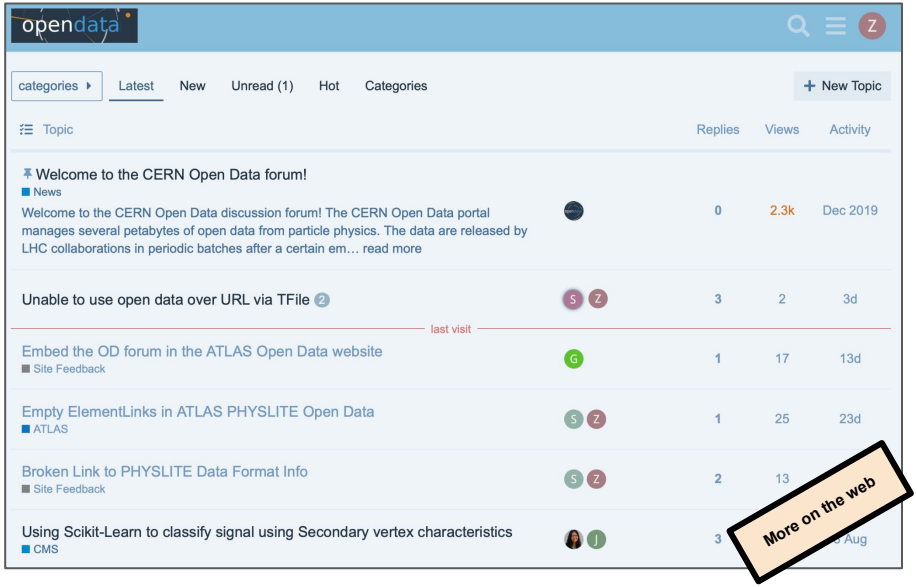


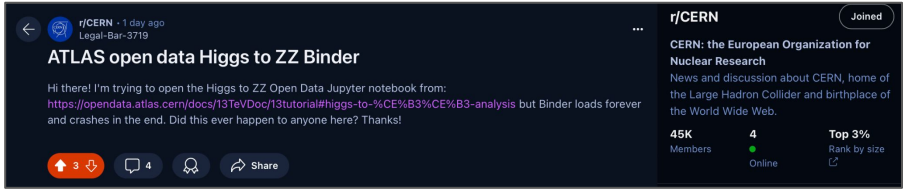
- **Single** Open Data Portal 'entry' record

- Everything redirecting to the ATLAS Open Data Website

User Support

- 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



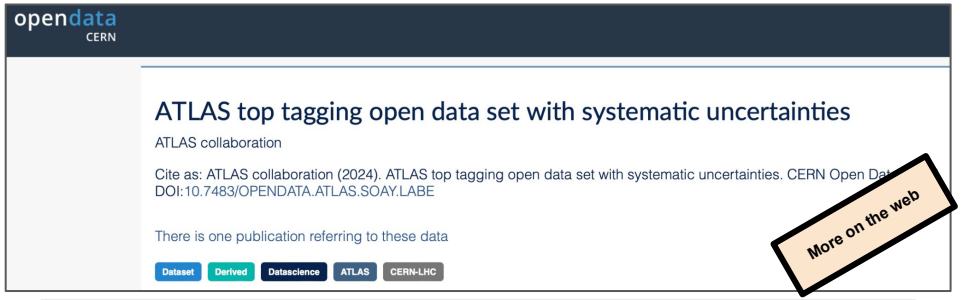


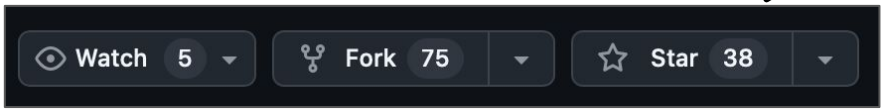
Advanced User Support

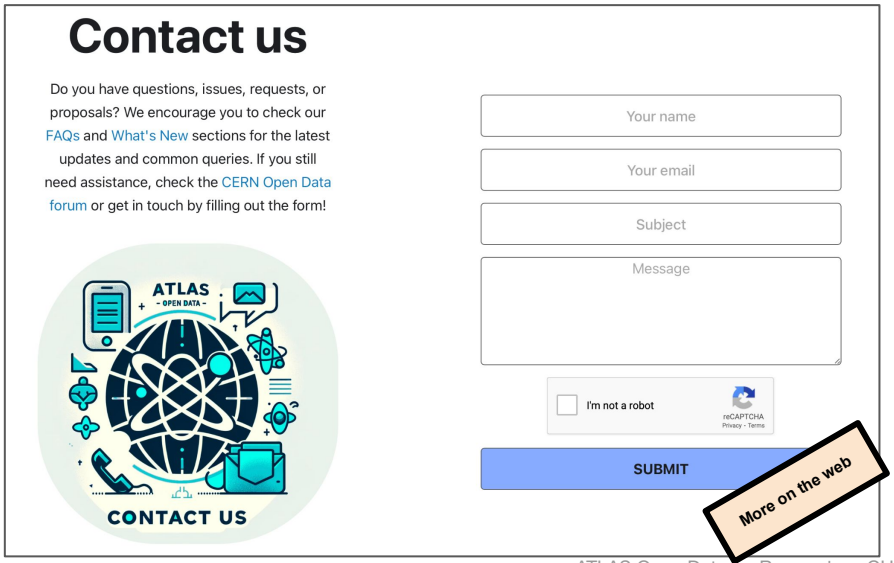
- Significant discussion in the collaboration about effort for support — “What Ifs”
  - What if someone requests an SM sample?
  - What if someone wants to run their own (BSM?) MC?
  - What if someone needs help with a tool?
  - What if someone publishes a paper that requires a response?
- Agreed on ‘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

Tracking Open Data Usage: Key Performance Indicators



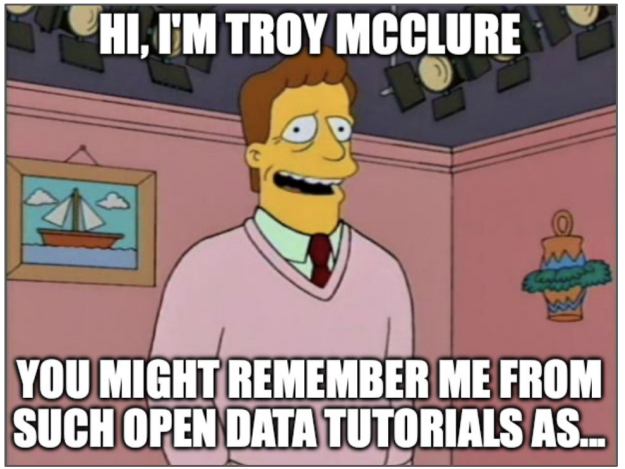




- **Research** output KPIs
  - Now tracking citations via DOI and URL
  - Indications this is not well reported (tools?)
- Repository / code tracking KPIs
  - Watches / forks / stars
- Website KPIs
  - Clicks, website search results
- Direct Data KPIs
  - Downloads, remote reads, reads from eos
  - These 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!

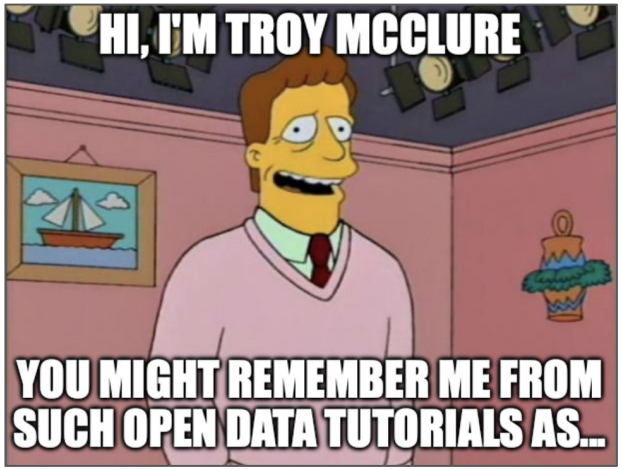
A Playground for Future-Proofing

- 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?



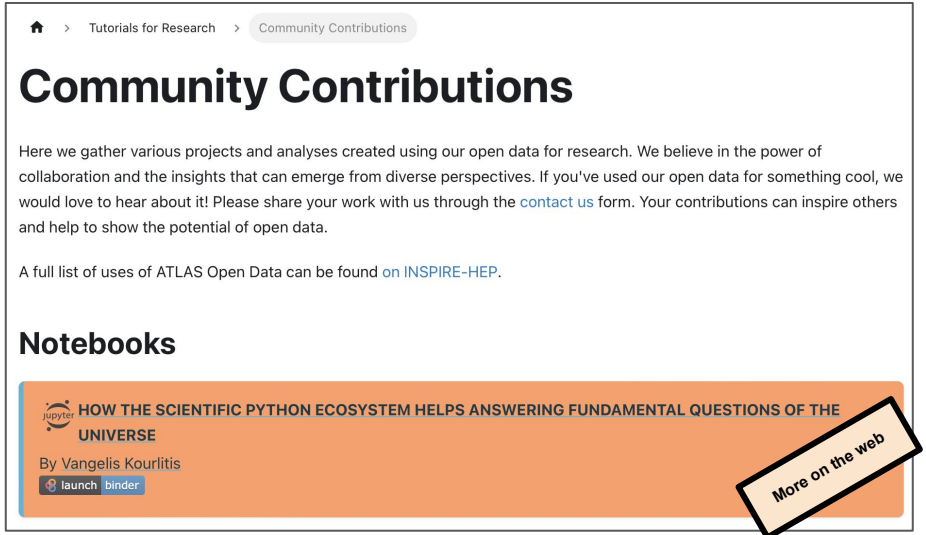
A Playground for Future-Proofing

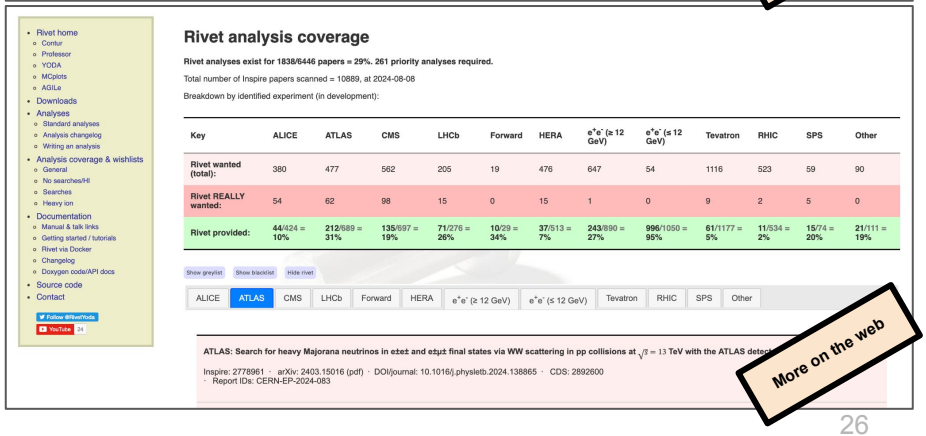
- Who will store the data in 2050?
  - Does this scale for the lifetime of the experiment?
  - What about the software? Websites? Documentation? Examples?
- Where physically are they stored?
  - **All** of them? Even the little example mini-formats?
- Who has rights to the repos?
  - Ownership? At CERN? 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?



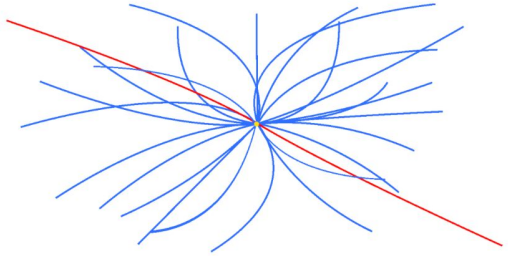
What's Next

- 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





ATLAS Open Data



High Energy Physics data for everyone.

NEW

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!

Extra Notes on Open Data Release

- 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 ~2x the data to start (rounding to the nearest file)