



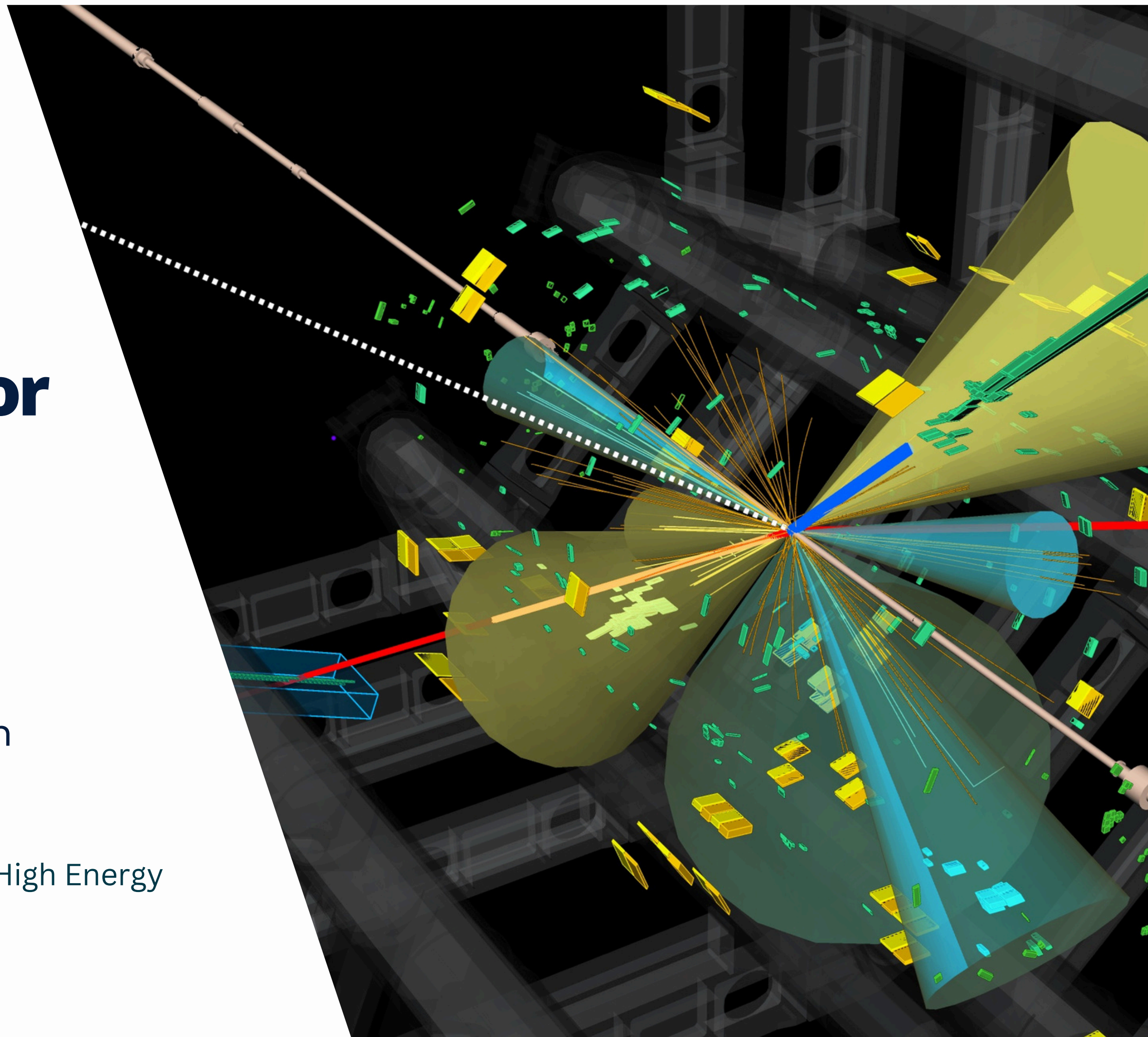
# Latest Releases of ATLAS Open Data for Education and Research

KATE SHAW

UNIVERSITY OF SUSSEX

On behalf of the ATLAS Collaboration

European Physical Society Conference on High Energy  
Physics, Marseille, France, July 2025





# Open Science Movement



Open science is an **accelerator** for the **Sustainable Development Goals (SDGs) 2030** and a powerful tool to bridge the science divide between and within countries

**Open science** aims at making scientific knowledge openly **available, accessible** and **reusable**.

The key elements include open access to scientific publications, **data**, educational resources, software and hardware, and open infrastructures

# CERN OPEN DATA POLICY

## Level 1: Published Results

- Available with Open Access
- HEPData: Repository for publication-related HEP data
- Rivet toolkit: Robust Independent Validation of Experiment and Theory

## Level 2: Outreach and Education

- Dedicated subsets of data selected and formatted to provide rich samples to maximise their educational impact, and to facilitate the easy use of the data.

## Level 3: Reconstructed Data

- Experiments release calibrated reconstructed data useful for algorithmic, performance and physics studies

## Level 4: Raw Data – Not feasible

## Open Science at CERN website

CERN Open Data Policy for the LHC Experiments.”  
<https://cds.cern.ch/record/2745133> , November 2020

### CERN Open Data Policy for the LHC Experiments November, 2020

The CERN Open Data Policy reflects values that have been enshrined in the CERN Convention for more than sixty years that were reaffirmed in the European Strategy for Particle Physics (2020)<sup>1</sup>, and aims to empower the LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Making data available responsibly (applying FAIR standards<sup>2</sup>), at different levels of abstraction and at different points in time, allows the maximum realisation of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community. CERN understands that in order to optimise reuse opportunities, immediate and continued resources are needed. The level of support that CERN and the experiments will be able to provide to external users will depend on available resources.

This policy relates to the data collected by the LHC experiments, for the main physics programme of the LHC — high-energy proton–proton and heavy-ion collision data. The foreseen use cases of the Open Data include reinterpretation and reanalysis of physics results, education and outreach, data analysis for technical and algorithmic developments and physics research. The Open Data will be released through the CERN Open Data Portal which will be supported by CERN for the lifetime of the data. The data will be tailored to the different uses, and will be made available in formats defined by each experiment that afford a range of opportunities for long-term use, reuse and preservation. In general, four levels of complexity of HEP data have been identified by the Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group<sup>3</sup>, which serve varying audiences and imply a diversity of openness solutions and practices.

**Published Results (Level 1) Policy:** Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,<sup>4</sup> with selection routines stored in specialised tools. The data made available may include simplified or full binned likelihoods, as well as unbinned likelihoods based on datasets of event-level observables extracted by the analyses. Reinterpretation of published results is also made possible through analysis preservation and direct collaboration with external researchers.

**Outreach and Education (Level 2) Policy:** For the purposes of education and outreach, dedicated subsets of data are used, selected and formatted to provide rich samples to maximise their educational impact, and to facilitate the easy use of the data. These data are released with a schedule and scope determined by each experiment. The data are provided in simplified, portable and self-contained formats suitable for educational and public understanding purposes; but are not intended nor adequate for the publication of scientific results. Lightweight environments to allow the easy exploration of these

<sup>1</sup> European Strategy Group (2020), ‘2020 Update of the European Strategy for Particle Physics’.

<sup>2</sup> FAIR Guiding Principles for scientific data management and stewardship. Available at: <https://www.go-fair.org/fair-principles/>.

<sup>3</sup> Data management plans are defined by the LHC experiments to address the long-term preservation of internal data products. See: Akopov et al., Status report of the DPHEP Study Group: Towards a global effort for sustainable data preservation in high energy physics. arXiv preprint arXiv:1205.4667 (2012).

<sup>4</sup> Repository for publication-related High-Energy Physics data: <http://www.hepdata.net>.

# CERN OPEN DATA PORTAL

opendata  
CERN

Help ▾ About ▾

Explore more than **five petabytes**  
of open data from particle physics!

Search

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

## Explore

[datasets](#)  
[software](#)  
[environments](#)  
[documentation](#)

## Focus on

[ALICE](#)  
[ATLAS](#)  
[CMS](#)  
[DELPHI](#)  
[LHCb](#)  
[OPERA](#)  
[PHENIX](#)  
[TOTEM](#)  
[Data Science](#)

### ATLAS $\sqrt{s}$ simulation for ML-based jet flavour tagging (JetSet)

Flavour-tagging — the task of identifying the flavour of jets — is essential for many physics analyses at the ATLAS experiment.

[Dataset](#) [Derived](#) [Simulated](#) [ATLAS](#)

### ATLAS releases first open data from heavy-ion collisions

The ATLAS Collaboration has released its first open data of heavy-ion collisions for research purposes. This data includes a nucleon pair, recorded in 2015 as part of the Large Hadron Collider's second operation period (LHC Run 2).

[News](#) [ATLAS](#)

### ATLAS releases 65 TB of open data for research

Explore over 75 billion LHC collision events — from home

[News](#) [ATLAS](#)

### ATLAS DAOD\_HION14 format Run 2 2015 Pb-Pb MC simulation

Run 2 2015 Pb-Pb MC simulation from the ATLAS experiment

[Dataset](#) [Simulated](#) [Heavy-Ion Physics](#) [ATLAS](#)

### ATLAS DAOD\_HION14 format Run 2 2015 Pb-Pb collision data

Run 2 2015 Pb-Pb collision data from the ATLAS experiment

[Dataset](#) [Collision](#) [ATLAS](#)

### DAOD\_HION14 format 2015 Pb-Pb Open Data for Research from the ATLAS experiment

2015 Pb-Pb Open Data for Research from the ATLAS experiment

[Dataset](#) [Simulated](#) [Collision](#) [Heavy-Ion Physics](#) [ATLAS](#)

### ATLAS top tagging open data set with systematic uncertainties

Boosted top tagging is an essential binary classification task for experiments at the Large Hadron Collider (LHC). The Open Data Set is...

[Dataset](#) [Derived](#) [Simulated](#) [ATLAS](#)

### DAOD\_PHYSLITE format 2015-2016 Open Data for Research from the ATLAS experiment

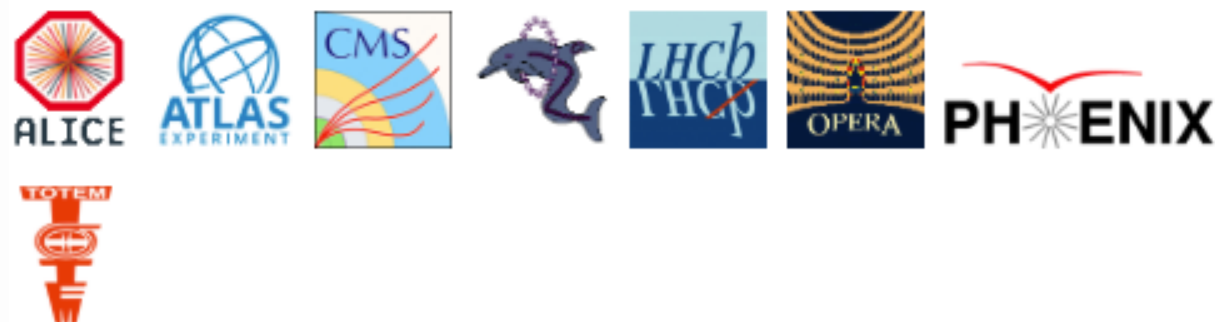
2015-2016 Open Data for Research from the ATLAS experiment

[Dataset](#) [Simulated](#) [Collision](#) [ATLAS](#)

### ATLAS DAOD\_PHYSLITE format MC simulation top systematic variation samples

MC simulation top systematic variation samples from the ATLAS experiment

[Dataset](#) [Simulated](#) [Standard Model Physics](#) [Top physics](#) [ATLAS](#)



<https://opendata.cern.ch>



# ATLAS Open Data For Education & Research

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

## Our values

The collaboration shares the data gathered by the ATLAS detector committing to three fundamental principles:

### Accessibility

Make the data and the tools openly available for everyone to use, without technology, region, or knowledge restrictions.

### Transferable expertise

Along with particle physics analysis and ATLAS learning objectives, provide skills in programming, software and machine learning.

### Usability

Different target audiences, with different backgrounds and skills must be able to use the data and tools for a wide range of learning objectives.

Purpose made website:

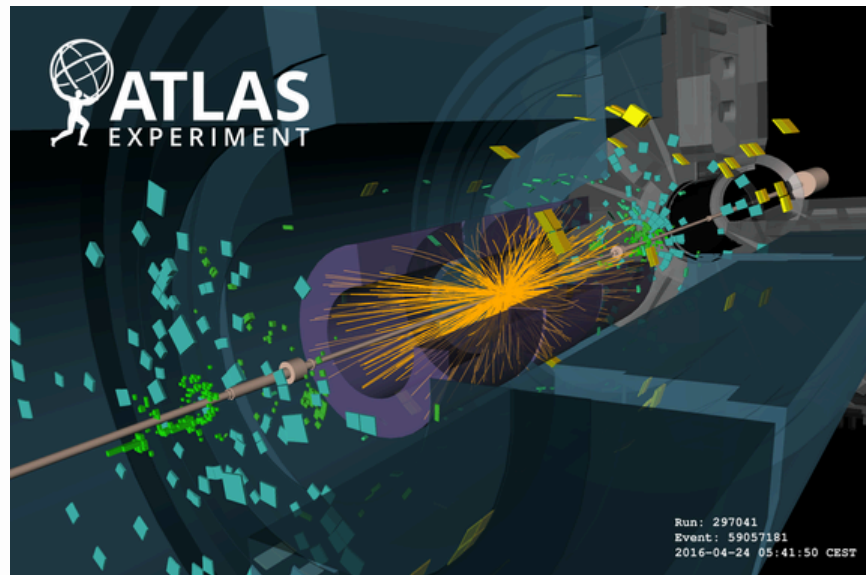
- Datasets
- Resources
- Tutorials
- Data visualisation
- Website framework

<https://opendata.atlas.cern/>

# ATLAS OPEN DATA for Research

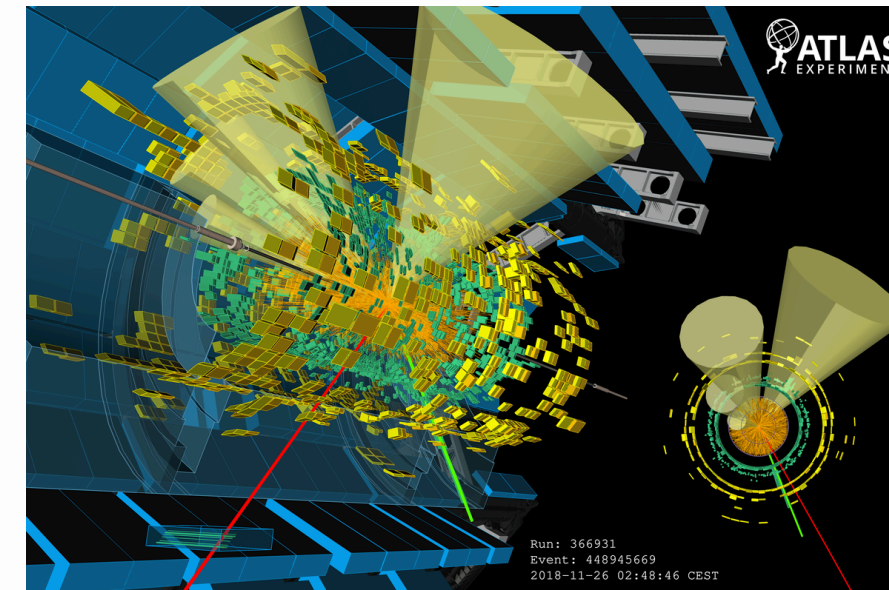
## Webpage

### Proton-Proton Collisions



**13 TeV Proton-Proton collision datasets**, 36 fb<sup>-1</sup>, 2015–2016, 65 TB in PHYSLITE files, with 2 billion events of simulated data

### Lead-Lead Collisions



**5 TeV Lead-Lead collision datasets**, 486 μb<sup>-1</sup>, 2015, 4 TB in DAOD\_HION14 files, with corresponding simulations

### Coming soon

- **Event generation data** in HEPMC format
  - provided in 10,000 event text files, tarred and gzipped to save space, for both 13 TeV and 13.6 TeV configurations
- **Heavy ion data** from the hard probes stream with corresponding simulations



# ATLAS OPEN DATA for Education

## Webpage



Datasets on  
CERN Open  
Data Portal

Proton-Proton Collisions

Open Data for  
Research release  
65 TB, 36 fb<sup>-1</sup>  
PHYSLITE

Open Data for  
Education release  
2 TB, 36 fb<sup>-1</sup>  
ROOT NTuples

Skimmed samples  
selecting dedicated  
final states  
1.5 GB to ~350 GB  
ROOT NTuples

- **PhysLiteToOpenData Framework** based on Combined Performance algorithms using the latest ATLAS analysis release
- Can be run inside a **docker container image**
- Repository will be registered in Zenodo with a DOI so that it can be cited with a persistent identifier

# ATLAS OPEN DATA for Education



Datasets on  
CERN Open  
Data Portal

Website hosted 8TeV 1 fb<sup>-1</sup> datasets, and 13 TeV 10 fb<sup>-1</sup> educational datasets, used by 10,000 students worldwide!

## Whats New with the third release?

Almost **four times** more data (36fb<sup>-1</sup> vs. 10fb<sup>-1</sup>)

Uses data from **recent** ATLAS Athena **releases** (Rel. 25)

**Latest recommendations** from the ATLAS combined performance groups

New features:

- Truth collections for electrons, taus, jets, met, muons and photons
- An example of jet energy scale resolution systematics
- Multi-lepton triggers, MET triggers, tau triggers (and jet triggers)
- Lepton ID and isolation working points



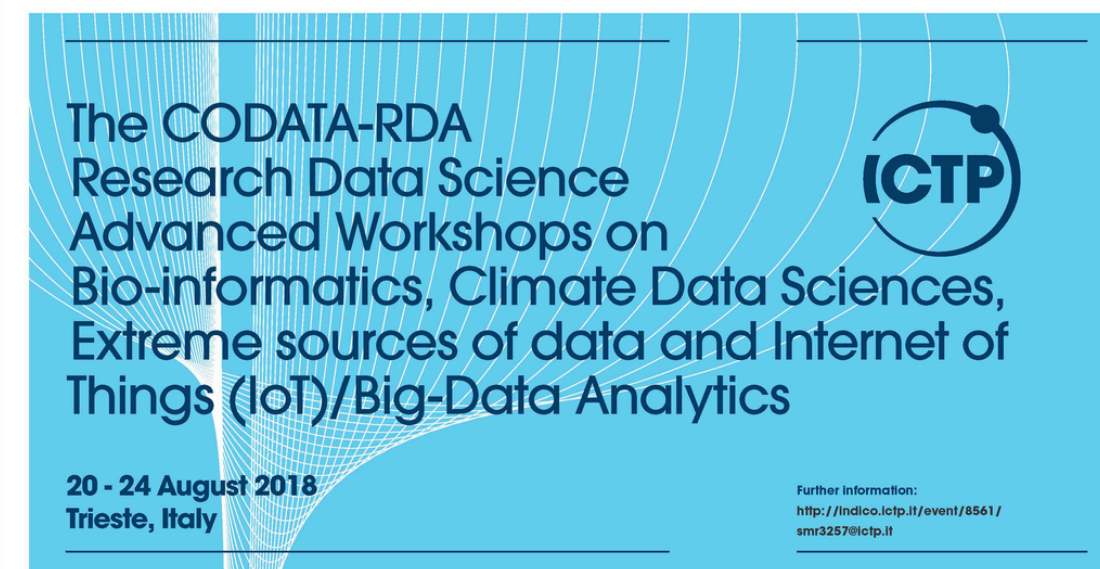
# ATLAS OPEN DATA for Education

## How is it used

Many members of the ATLAS collaboration use the simplified Ntuples, and our ready to use Notebooks and software for:

- Undergraduate courses and BSc and MSc projects
- Students can dive right into the learning objectives immediately (physics, statistics, analysis skills such as fitting and machine learning);
- Training and outreach activities such as hackathons and workshops, with PhD students, university students and 16–18 year olds.
- Been used by 10,000's students worldwide!

**Users also include non scientists, teachers and motivated students**



The CODATA-RDA Research Data Science Advanced Workshops on Bio-informatics, Climate Data Sciences, Extreme sources of data and Internet of Things (IoT)/Big-Data Analytics

20 - 24 August 2018  
Trieste, Italy

Further information:  
<http://indico.ictp.it/event/8561/>  
smr3257@ictp.it

During this activity, four applied/thematic workshops on Research Data Science would run in parallel.

### Description:

Workshop on Extreme sources of data: Introduction to CERN LHC and ATLAS Experiment. Hands-on sessions will include python coding and tutorials on using the ATLAS Open Data Platforms/Tools.

Workshop on Bioinformatics: Advanced hands-on-tutorials on computational methods for the management and analysis of genomic and sequencing data.

Workshop on IoT/Big Data Analytics: Topics will include Big Data tools and technology; real time event processing; low latency query; analyzing social media and customer sentiment. Hands-on sessions will include deploying and using Big-Data Analytic tools and platforms including Hadoop, Apache Kafka and HDF Workshop on Climate Data Science: Cloud computing platform/tools for Climate Data Sciences including integration and visualization of on-line and local datasets. Hands-on sessions will focus on using on-line high performance platforms and tools for Climate Data Science.

Participation in any of these applied workshops requires some knowledge of Research Data Science, which may be obtained by applying separately for the "Research Data Science Summer School" (SMR3231) which takes place August 6-17 2018.

### How to apply:

Online application:  
<http://indico.ictp.it/event/8561/>

Female scientists are encouraged to apply.

### Grants:

A limited number of grants are available to support the attendance of selected participants, with priority given to participants from developing countries. There is no registration fee.

### Directors:

A. HARRISON (Department of Mathematical Sciences, University of Essex)  
S. HUDSON (CODATA)  
H. SHANAHAN (Department of Computer Science, Royal Holloway University of London, UK)  
C. VAN GELDER (Dutch Techcentre for Life Sciences (DTL), Netherlands)  
R. MURENZI (TWAS)  
T.K. ATTWOOD (University of Manchester, UK)  
R. QUICK (Indiana University, U.S.A)  
S. JONES (University of Glasgow, UK)  
N. MULDER (University of Cape Town, South Africa)  
U. SINGE (ICTP)  
M. ZENNARO (ICTP)  
A. TOMPKINS (ICTP)

### Local Organizer:

C. ONIME (ICTP)

### Speakers:

ELIXIR  
University of Trieste  
European Open Science Cloud  
CERN  
Green Climate Fund

### Deadline:

21 May 2018



# ATLAS OPEN DATA for Education

## How is it used



## Final Project ZBoson and Search for New Resonances with ATLAS

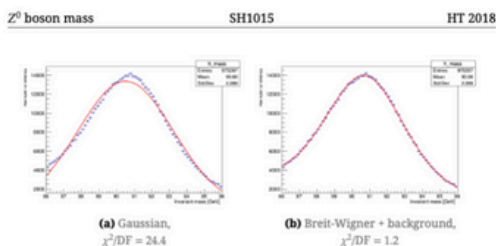


Figure 5: Fitting different distributions to the results. DF signifies the number of degrees of freedom

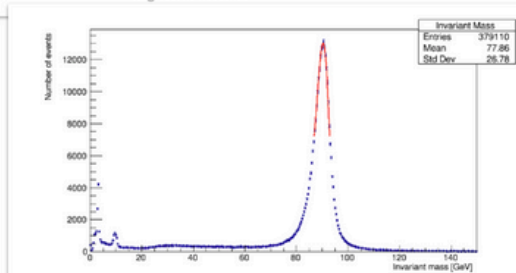
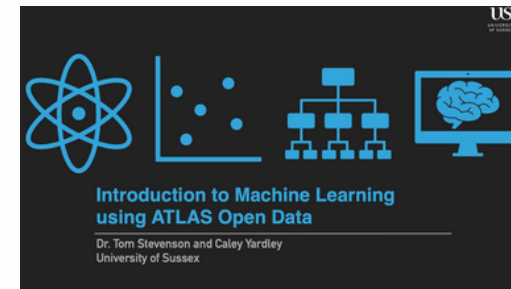
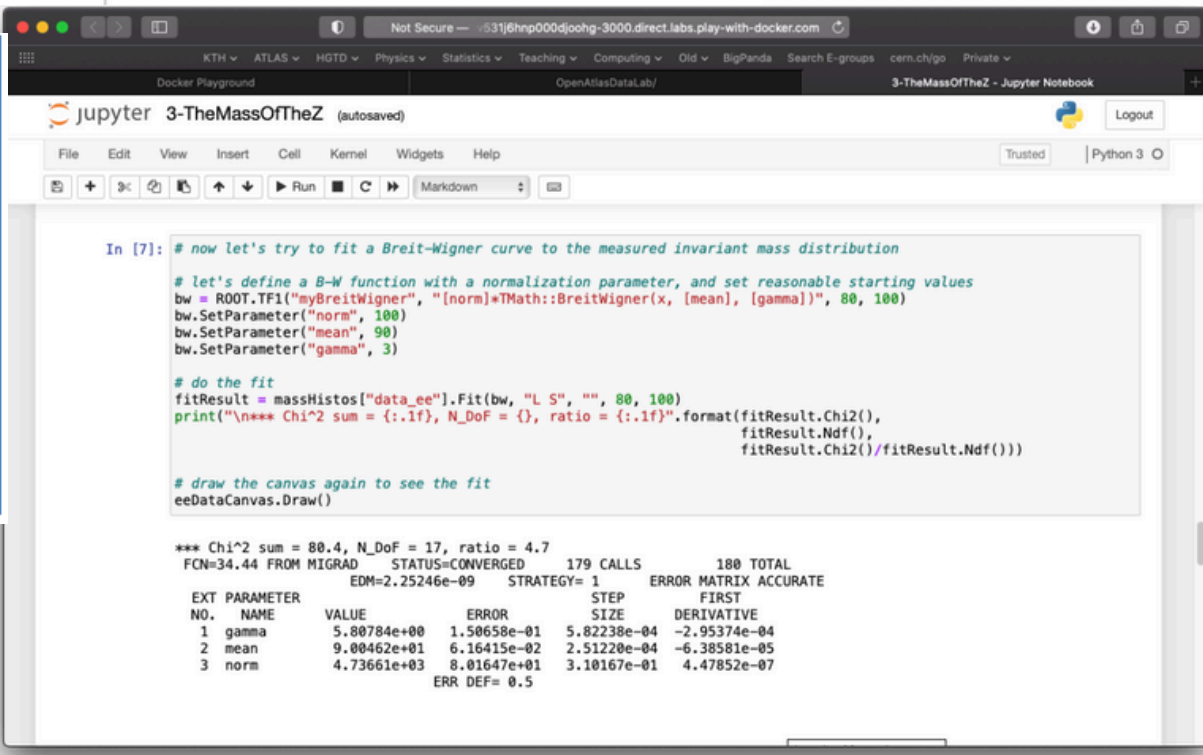


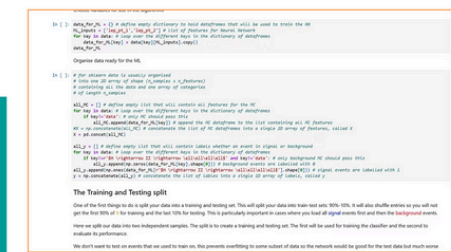
Figure 3: Invariant mass of the  $Z^0$  boson, approximated as the function  $f(x)$ , on  $p_T=171$



- Lots to learn:
- Different Z shapes for  $e^+e^-$  and  $\mu^+\mu^-$  - bremsstrahlung, resolution, etc
- Width-lifetime connection
- Background processes, small but still important to model



## Machine Learning Workbook

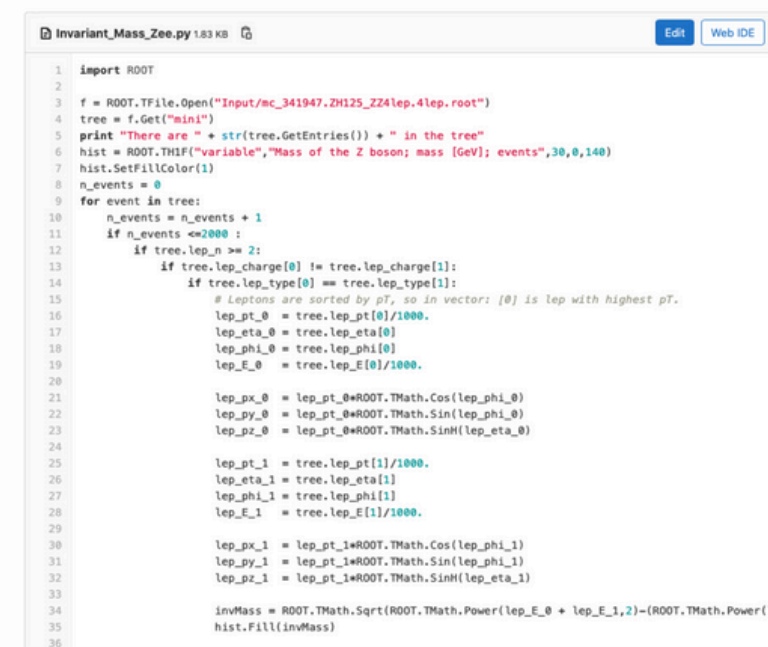
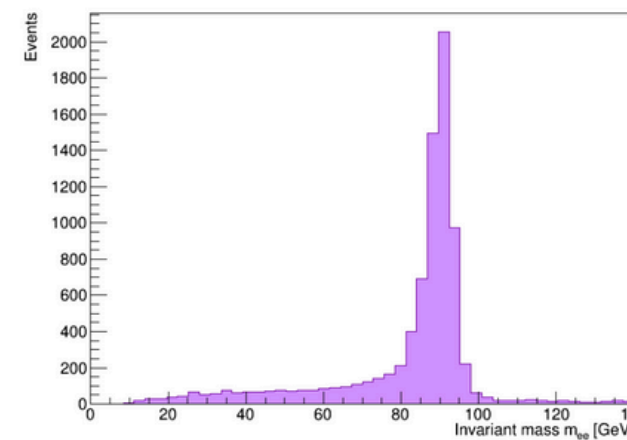


ACCESS WORKBOOK HERE

This online workbook, created by Thomas Stevenson, will introduce the basics of Machine Learning and how it can be exploited in an experimental physics context. Experience with python is required, however no knowledge of ML is necessary. Warning: it may take a long time to load the workbook and associated repositories! A preview is available by clicking on the image.

The notebook uses real proton-proton collision data from the ATLAS experiment at CERN to explore some hands-on examples, implementing some of the most commonly used Machine Learning techniques and exploring the potential uses of Machine Learning.

**Their task:** inv. mass of H from two Z, add selections, scale according to XS, add backgrounds, stack histograms, add data etc.



UNIVERSITY OF AMSTERDAM



# Getting Started

We use [Matomo](#) to internally analyze traffic and help us improve your experience.; check [our privacy policy!](#)

ATLAS Open Data

Get Started

## Get Started

In this section you will find different suggested paths to get involved with ATLAS Open Data. This are just suggestions on what we think it will be more usefull to check in each case. However, feel free to check the website freely.

 Quick start

The quickest way to start learning with ATLAS Open Data.

- No technical knowledge required
- Full introduction provided
- Step-by-step tutorials
- Instructions how to access data
- Histogram Analyser

[Get Started Webpage](#)

## Histogram Analyser

**Tool** shows how physicists differentiate between physics processes applying cuts using just your mouse.

### How to Separate Signals: Higgs to WW

Let's look at the simulated data.  
Using the Histogram Analyser we can look at each sample separately and understand a little more about their characteristics.  
This will help us separate our signal from the background later.

Select the sample by clicking on the bar in the Expected Number of Events histogram.  
The rest of the histograms now just display the characteristics of your chosen sample.

$$H \rightarrow W^+ W^-$$

$$H \rightarrow W^+ W^- \rightarrow \ell^+ \ell^- \nu \bar{\nu}$$

( $\ell$  = electron, muon)

### Curriculum Learning Objective: Particle decay balancing with charge

[see e.g. OCR A-level physics 6.4.2]

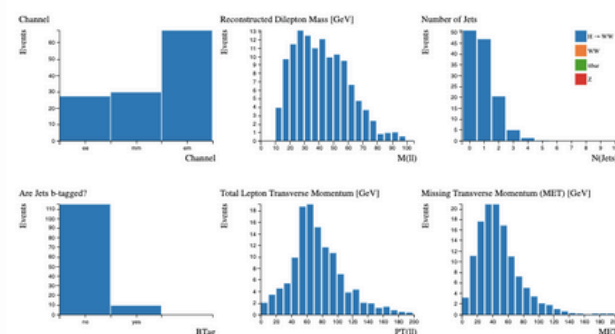
Why does the higgs boson decay into particles whose charges sum to zero?

### Curriculum Learning Objective: Classification of leptons

[see e.g. OCR A-level physics 6.4.2(d)]

Electron and muon channels are shown separately in the histograms.

Our signal is the Higgs boson which decays into two  $W$  bosons which subsequently decay into leptons and neutrinos.

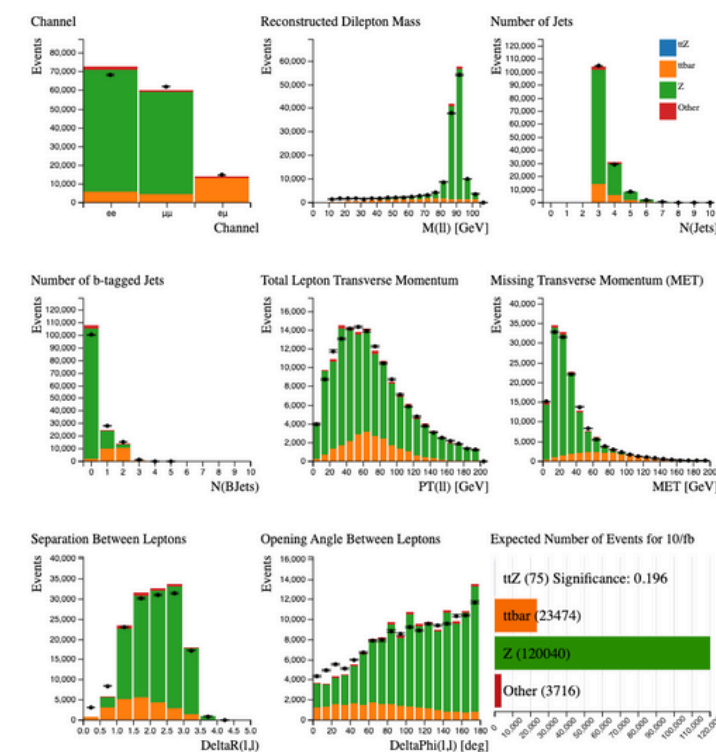


### The histograms explained

Histogram Analyser displays nine histograms. The description of each follows.

The histograms can take about 30 seconds to load. Whilst loading you'll only see the histogram titles. Once loaded you'll see the histograms appear under their titles.

We think it really helps to be able to see all nine histograms on your screen at the same time. So if this isn't the case to start with, we suggest decreasing the zoom in your web browser until you can see all nine (e.g 67%).

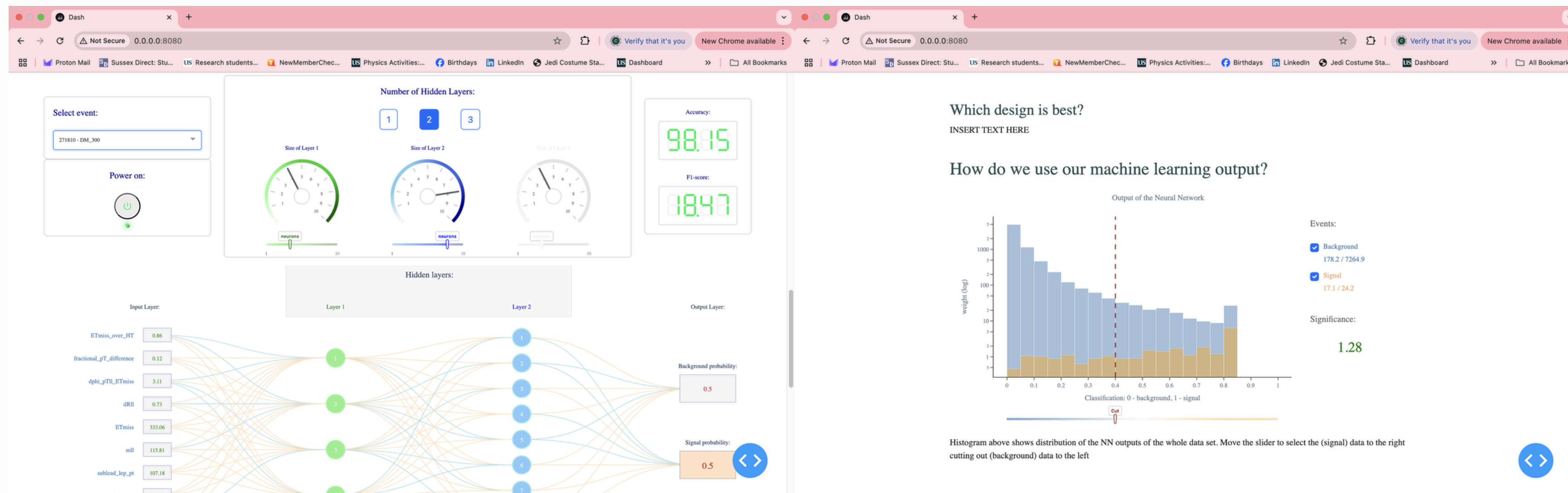


You can select just the  $t\bar{t}Z$  events by clicking on the  $t\bar{t}Z$  in the 'Expected Number of Events' histogram.

# Machine Learning Online Interactive Application

- Machine Learning tutorial using only your mouse!!
- <https://ml-visual-dashboard-atlas-open-data.app.cern.ch/>
- Run a Neural Network to discover Dark Matter!

**NEW**





# Teacher workshop

## Welcome to ATLAS Open Data in the Classroom



[Webpage](#)

An application to facilitate the learning of useful experimental particle physics techniques

Available in languages:  
English, Spanish, Italian

Google Chrome

### ATLAS Open Data in the Classroom

#### ▶ Getting Started

▶ Foundations of Particle Physics

▶ Experimental Particle Physics

▶ Analyze ATLAS Open Data

▶ Intro to Python

▶ Classroom Toolkit

# Advanced Tutorials

## Jupyter Notebooks

### Uproot

Higgs to ZZ **NEW**

This notebook uses the 2025 release of the ATLAS Open Data to show you to rediscover the Higgs boson yourself! You will discover the Higgs boson into a pair of Z bosons, which are in turn decaying into a lepton-antilepton

Physics: ★

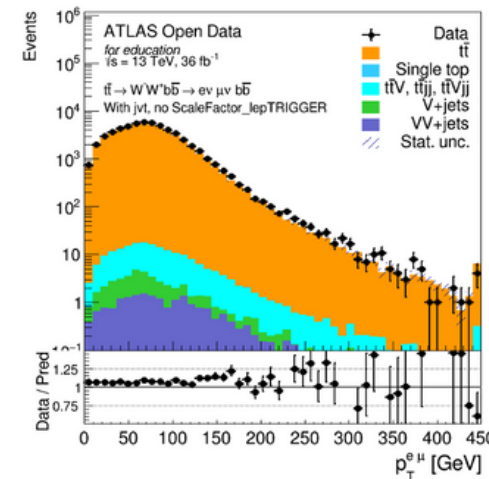
Coding: ★★

Time: ★★★

 launch  binder  Open in Colab

Higgs to  $\gamma\gamma$  analysis **NEW**

This notebook uses the 2025 release of ATLAS Open Data, with  $36.1 \text{ fb}^{-1}$ ,



### Developing of online tutorials:

Jupyter notebooks analyses (with & without ROOT framework), in C++, pyROOT 8 TeV and 13 TeV & uproot.

Different frameworks available to suit different learning objectives and use cases:


- Python
- C++
- RDataFrame
- Uproot / Coffea
- 

All Notebooks are available on the GitHub repository.

Jupyter Notebooks



# ATLAS Open Data Videos on YouTube










## ATLAS Open Data Tutorials

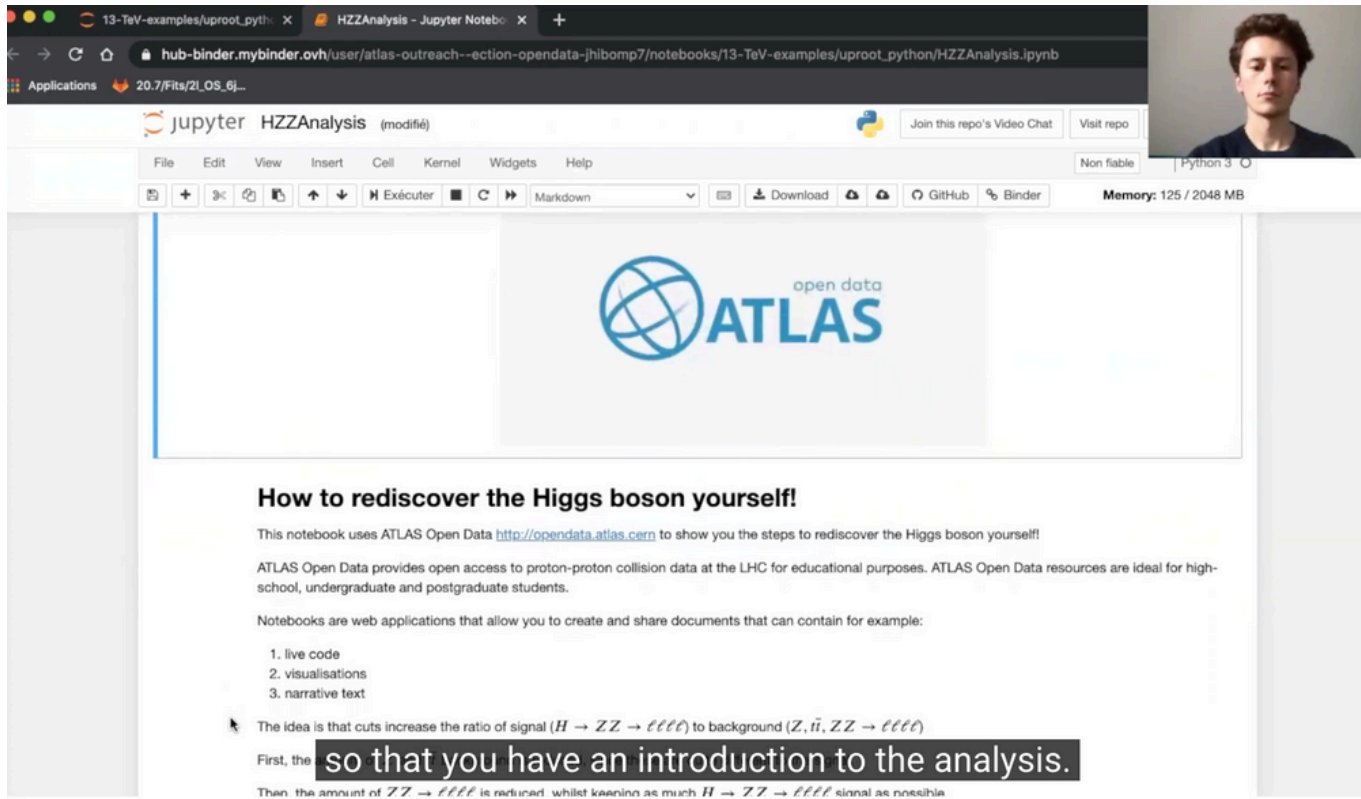
by ATLAS Experiment

Playlist • 8 videos • 4,280 views

Video tutorials explaining different ways to use Open Data from the ATLAS Experiment at CERN.

▶ Play all

1		<b>Find the Higgs boson with your mouse – ATLAS Open Data Tutorial</b> ATLAS Experiment • 3.3K views • 4 years ago
2		<b>Data analysis in a web browser – ATLAS Open Data Tutorial</b> ATLAS Experiment • 1.9K views • 4 years ago
3		<b>How to rediscover the Higgs boson – ATLAS Open Data Tutorial</b> ATLAS Experiment • 3.2K views • 4 years ago
4		<b>Installing a VirtualBox - ATLAS Open Data Tutorial</b> ATLAS Experiment • 700 views • 4 years ago
5		<b>Installing a Virtual Machine - ATLAS Open Data Tutorial</b> ATLAS Experiment • 1.5K views • 4 years ago
6		<b>Making Selection Cuts with PyROOT – ATLAS Open Data Tutorial</b> ATLAS Experiment • 1.8K views • 3 years ago
7		<b>Create Histograms with PyROOT – ATLAS Open Data Tutorial</b> ATLAS Experiment • 2.5K views • 3 years ago



**How to rediscover the Higgs boson yourself!**

This notebook uses ATLAS Open Data <http://opendata.atlas.cern> to show you the steps to rediscover the Higgs boson yourself!

ATLAS Open Data provides open access to proton-proton collision data at the LHC for educational purposes. ATLAS Open Data resources are ideal for high-school, undergraduate and postgraduate students.

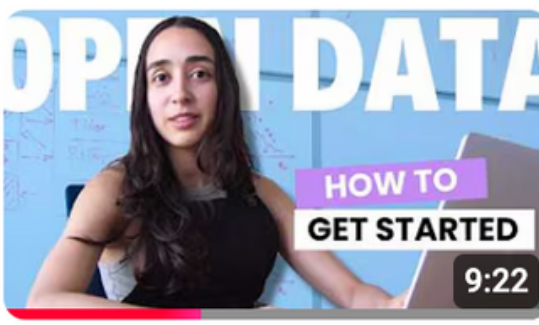
Notebooks are web applications that allow you to create and share documents that can contain for example:

- live code
- visualisations
- narrative text

The idea is that cuts increase the ratio of signal ( $H \rightarrow ZZ \rightarrow \ell\ell\ell\ell$ ) to background ( $Z, \bar{t}\bar{t}, ZZ \rightarrow \ell\ell\ell\ell$ )

First, the **so that you have an introduction to the analysis.**

Then the amount of  $ZZ \rightarrow \ell\ell\ell\ell$  is reduced whilst keeping as much  $H \rightarrow ZZ \rightarrow \ell\ell\ell\ell$  signal as possible



**Getting Started with ATLAS Open Data**

ATLAS Experiment • 2.1K views • 1 year ago

**Find the videos here**

# Computing Environments and Analysis Facilities



SWAN at CERN: Lots of resources, but bound to CERN authorisation



ESCAPE Virtual Research Environment: Anyone can get an account, useful for small workshops



GoeGRID at Goettingen: A batch service, does not support interactive analysis



Binder & Colab: Great for running workshops, only up to so much memory



Docker: Robust Platform for developing, running and sharing applications within containers



# Summary

- ATLAS has released 36 fb<sup>-1</sup> proton-proton collision data For Research, along with Event Generation data, and Heavy Ion Data
- ATLAS has released – just last week!! – 36 fb<sup>-1</sup> proton-proton collision data for Education, its third dataset release
- Our team has been developing a dedicated website with step by step tutorials, tools, software, data visualisation tools, to allow users to access and use the data depending on their learning objectives, level and goals.
- Please contact us with feedback & suggestions

