

# ATLAS Open Data

Bringing TeV collisions to the World

CHEP 2024

21 October 2024

Giovanni Guerrieri (CERN) on behalf of the ATLAS Collaboration



ATLAS Open data for

Research

Education

## ATLAS Open data for

Research

See [Zach's talk](#)

Education

*We talk about this!*

## Accessibility

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

## 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

## Transferable expertise

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

ATLAS Open Data releases for education are being used by several schools, universities, interested individuals, as well as in public events, masterclasses and international workshops

The datasets are used for an ***educational purpose only***

## The FAIR principles

**F**indable

*Data are assigned a globally unique and **persistent identifier***

**A**ccessible

*Data are **retrievable** by their identifier using a standardized communications protocol*

**I**nteroperable

*Data or tools from non-cooperating resources are able to **integrate or work together***

**R**eusable

*Meta(data) are **richly described** with a plurality of accurate and relevant attributes*

**F**indable

*Where do I **find** ATLAS Open Data?*

**A**ccessible

*How do I use ATLAS Open Data?*

**I**nteroperable

*Where do I use ATLAS Open Data?*

**R**eusable

*When can I use ATLAS Open Data?*


*One question missing: **what is ATLAS Open Data for education?***

# What is ATLAS Open Data for education?

## A collection of data

Gathered by the ATLAS detector in its data acquisition runs, together with the associated [Monte Carlo simulations](#), including [systematic uncertainties](#)

## Three releases:

- [8 TeV](#) (2016):  $1\text{fb}^{-1}$  of data ( ~4.5% of 2012 data , ~6GB)
- [13 TeV](#) (2020):  $10\text{fb}^{-1}$  of data ( ~30% of 2016 data , ~150GB)
- 13 TeV (2024):  $36\text{fb}^{-1}$  of data  Coming soon!

## The datasets

Includes [calibrated and simplified information](#) about the reconstructed physics objects

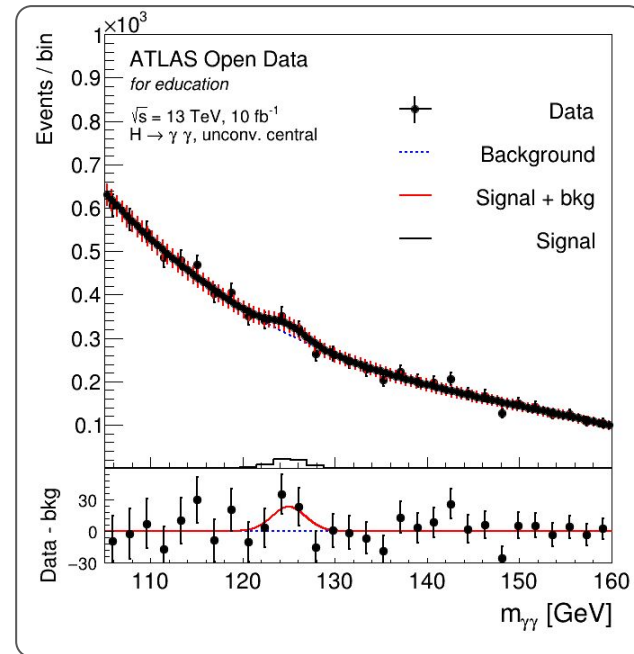
## Labels matter

Notice the “for education” label in the plot title

Come with [extensive documentation](#), tutorials and resources to make data usable

[Enable users](#) to experience the analysis of particle-physics data in educational environments

Example of analysis performed with the 2020 release for education



# Where do I find ATLAS Open Data?



# Where do I find ATLAS Open Data?

The ATLAS Open Data website

<https://opendata.atlas.cern>

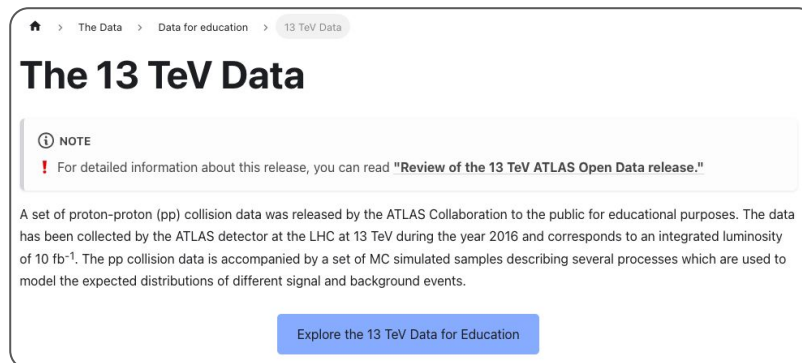
Provides information, tutorials and instructions on the data

The CERN Open Data Portal

<https://opendata.cern.ch>

Hosts the samples, together with basic information

- [8 TeV](#) (2016): ~O(10GB)
- [13 TeV](#) (2020)
- 13 TeV (2024) *Coming soon!*



Home > The Data > Data for education > 13 TeV Data

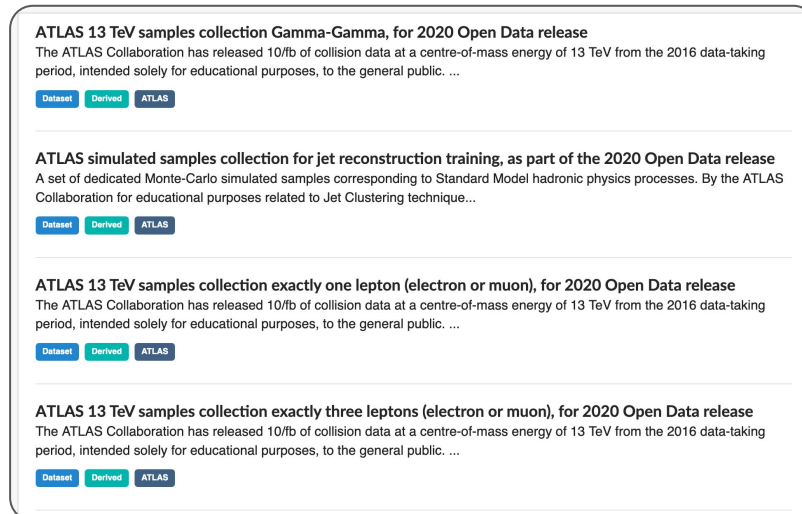
## The 13 TeV Data

**NOTE**

For detailed information about this release, you can read "[Review of the 13 TeV ATLAS Open Data release.](#)"

A set of proton-proton (pp) collision data was released by the ATLAS Collaboration to the public for educational purposes. The data has been collected by the ATLAS detector at the LHC at 13 TeV during the year 2016 and corresponds to an integrated luminosity of 10 fb<sup>-1</sup>. The pp collision data is accompanied by a set of MC simulated samples describing several processes which are used to model the expected distributions of different signal and background events.

Explore the 13 TeV Data for Education



**ATLAS 13 TeV samples collection Gamma-Gamma, for 2020 Open Data release**  
The ATLAS Collaboration has released 10/fb of collision data at a centre-of-mass energy of 13 TeV from the 2016 data-taking period, intended solely for educational purposes, to the general public. ...

**ATLAS simulated samples collection for jet reconstruction training, as part of the 2020 Open Data release**  
A set of dedicated Monte-Carlo simulated samples corresponding to Standard Model hadronic physics processes. By the ATLAS Collaboration for educational purposes related to Jet Clustering technique...

**ATLAS 13 TeV samples collection exactly one lepton (electron or muon), for 2020 Open Data release**  
The ATLAS Collaboration has released 10/fb of collision data at a centre-of-mass energy of 13 TeV from the 2016 data-taking period, intended solely for educational purposes, to the general public. ...

**ATLAS 13 TeV samples collection exactly three leptons (electron or muon), for 2020 Open Data release**  
The ATLAS Collaboration has released 10/fb of collision data at a centre-of-mass energy of 13 TeV from the 2016 data-taking period, intended solely for educational purposes, to the general public. ...

# How do I use ATLAS Open Data?

*How do I access and analyse data?*

# How do I use ATLAS Open Data?

## Qualitative exploration of data

[Histogram analysers](#) constitute an interactive and intuitive web based tool for fast, cut-based analysis of data. Visualise the data using online histograms

### Currently available:

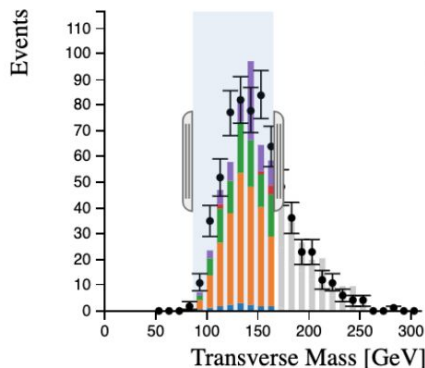
- Higgs boson decaying into two W bosons
- Associated production of a top-quark-pair and a Z boson

### Coming soon:

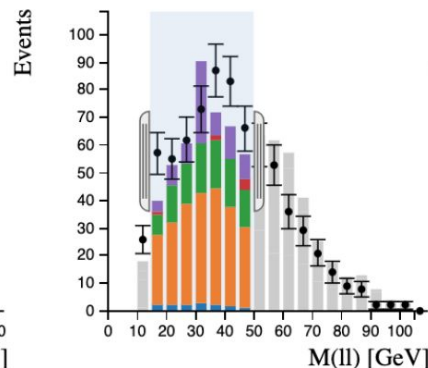
- Higgs boson decaying into two Z bosons
- Dark matter searches in the dilepton channel

- No technical knowledge required
- Introduction to the studied process
- Step-by-step explanation
- Advanced contents

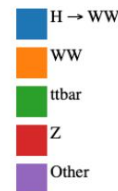
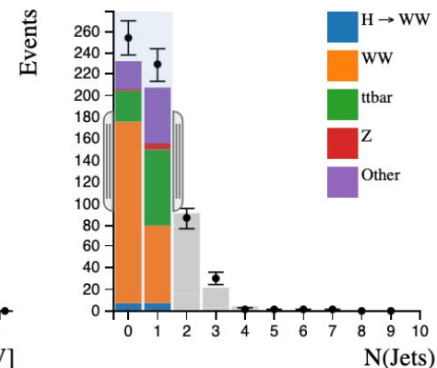
Transverse Mass



Reconstructed Dilepton Mass



Number of Jets



# How do I use ATLAS Open Data?

## Interactive analysis

We've built a set of [Jupyter notebooks](#) that allow data analysis to be performed directly in a [web browser](#)

- We list and summarise the tutorials in [our website](#)
- The notebooks are available in our [GitHub repository](#)
- Several analysis examples targeting [different users](#), with [different expertise](#) and interests
- Different frameworks, to adapt to everyone's interest:
  - C++
  - Python
  - RDataFrame
  - Uproot/Coffea

[Video tutorials also available](#)

### 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 data analyst in Physics!

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

##### [Higgs to ZZ with a neural network](#)

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 simple neural network like a professional data analyst in Physics!

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

##### [Higgs to ZZ with the Coffea framework](#)

This notebook uses ATLAS Open Data to show you the steps to rediscover the Higgs boson yourself, with the Coffea framework!

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

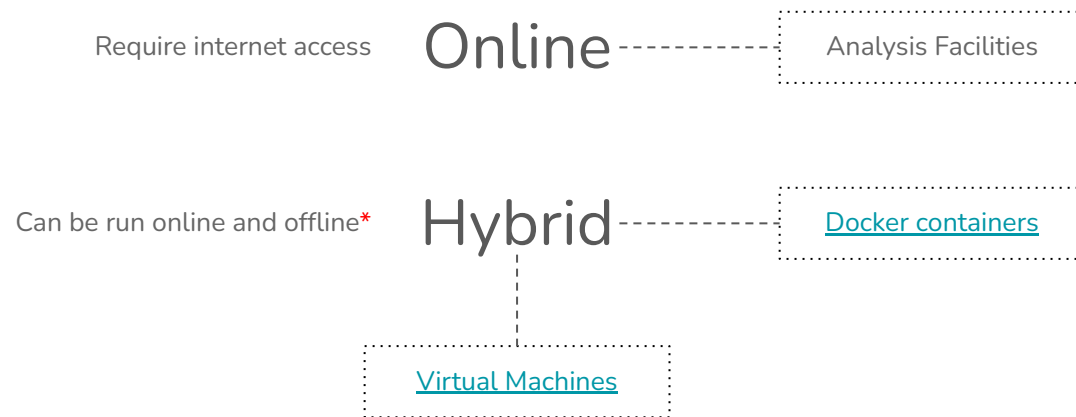
##### [Higgs to \$\gamma\gamma\$ analysis](#) **NEW**

This notebook uses the 2024 release of ATLAS Open Data, with  $36.1 \text{ fb}^{-1}$ , to show you the steps to rediscover the Higgs boson yourself! You will discover the Higgs boson decaying into two photons.

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

# Where do I use ATLAS Open Data?

*Which platforms / frameworks integrates with ATLAS Open Data?*



\*Internet connection is required in order to download material at the beginning

# Where do I use ATLAS Open Data?

## Online platforms

Swan/Binder platforms: very useful for setting up a **quick** and **individual** workspace

Documentation and tutorials on [our website](#)

Available at the click of a button

Data persistence

No account restrictions

No timeout time for sessions

Spawn time <1min

Change easily the software stack

I have a CERN account

### SWAN



- Requires CERN credentials

I don't have a CERN account

### Binder



Available at the click of a button

Data persistence

No account restrictions

No timeout time for sessions (1 CPU-h max)

Spawn time ~O(min)

Change easily the software stack  
Need to re-build the underlying image

# Where do I use ATLAS Open Data?

## Online platforms (ctd.)

Other analysis facilities: the [ESCAPE Virtual Research Environment](#)

Platform similar to SWAN integrated with additional resources:

- Data management: [Rucio](#)
- Reproducibility/Re-analysis: [Reana](#)
- Results/publications repository: [Zenodo](#)

Available at the click of a button

Data persistence

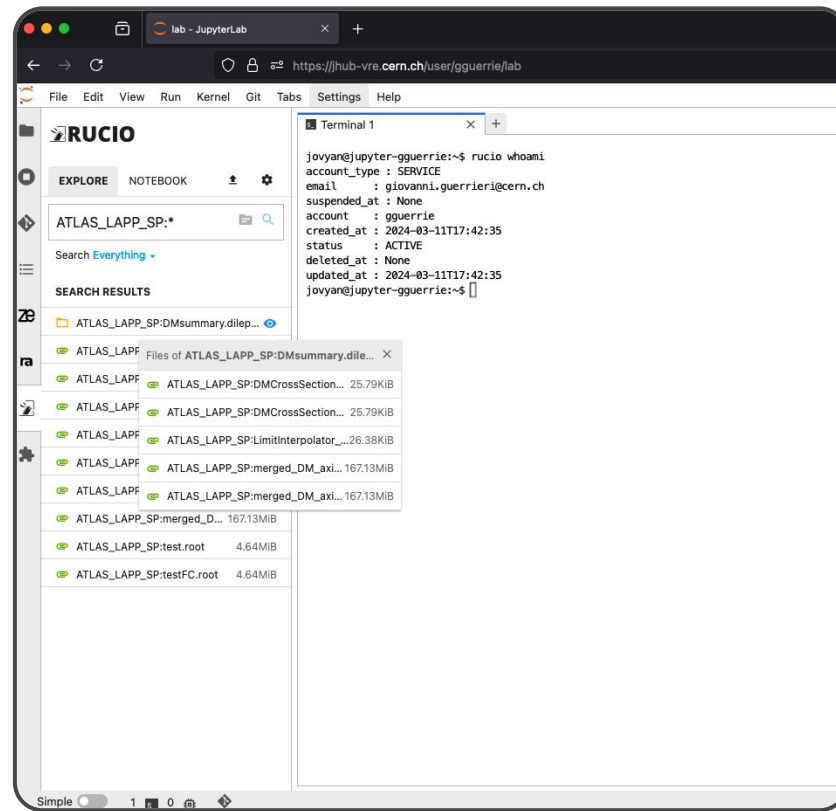
No account restrictions

**Anyone can create an account!** → No timeout time for sessions

Spawn time <1min

Change easily the software stack

See [Enrique's talk](#)





# When can I use ATLAS Open Data?

# When can I use ATLAS Open Data?

Whenever you like.

Data is not going anywhere

## What could go wrong then?

Several factors can affect usability, e.g.:

- Data changes location or gets corrupted
- Analysis tools are outdated or use deprecated dependencies
- Documentation is not available/up to date
- Users experience access restrictions
- Team does not have enough personpower to maintain everything

We need **all** the resources to be concurrently available and functioning



ATLAS-outreach-Data-Tools / MYATLAS-113  
**8 TeV open data gone? Breaks university lab!**



ATLAS-outreach-Data-Tools / MYATLAS-130  
**Missing images in data visualisation/ATLAS events**



ATLAS-outreach-Data-Tools / MYATLAS-156  
**The URLs to atlassoftwaredocs in open data website are broken**

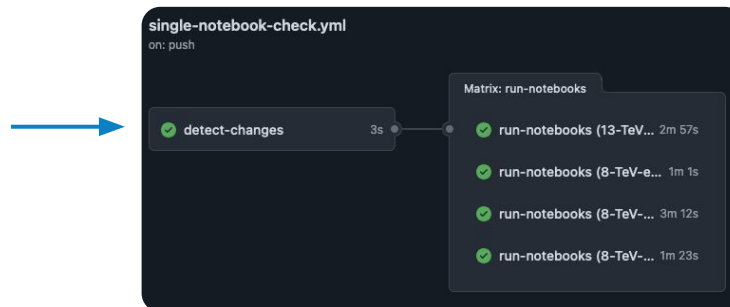
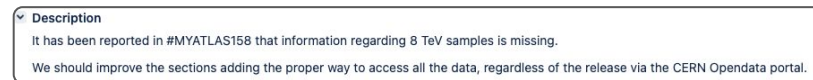
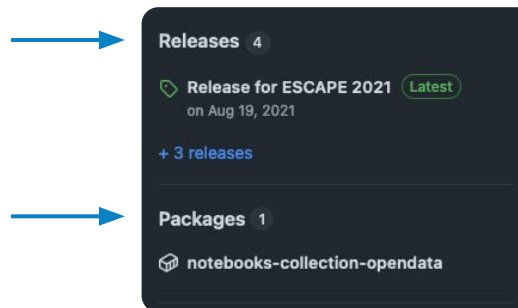


ATLAS-outreach-Data-Tools / MYATLAS-158  
**8 TeV data for educational purposes gone - again?**

# Best practices to foster usability

## How do we ensure usability through time?

- Manage code in [versioned repositories](#)
- Package the analysis environment in software [containers](#)
- [Document](#) everything from the start
- Define easily reusable [workflows](#)
- Use continuous, automated [testing](#)



## Serving Open Data for Education since 2016

- Beyond data, many resources, tutorials and examples are available.
- Widely used by several institutions for trainings, workshops, masterclasses

## New 13 TeV open data release coming soon

- What to do with the 2020 release?

## New possibilities with Open Data for Research

- In the process of finding synergies and complementarities
- Beginning to plan a workshop / hackathon

## Monitoring and watching

- Continue collecting usage statistics and conducting user surveys

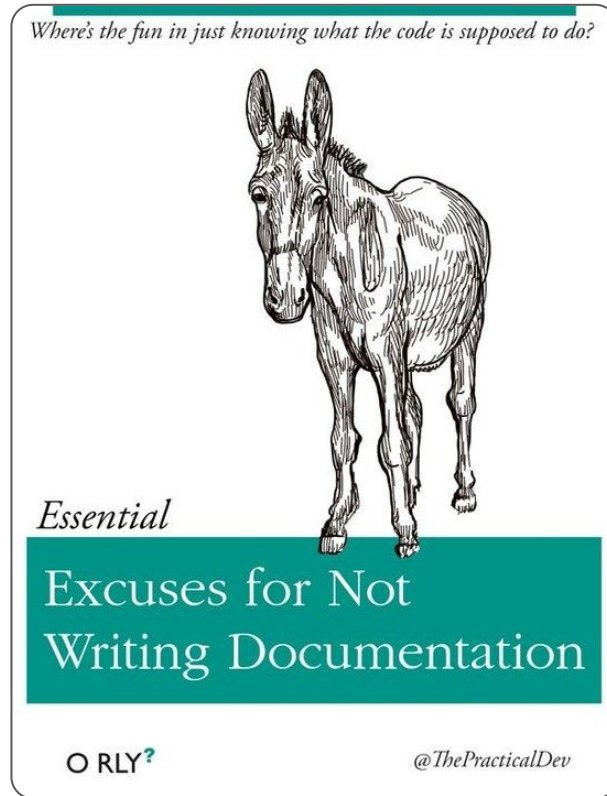
## Keep improving the documentation

- Data is nothing without knowing how to use it

## New examples and community contributions

- We are not only creating our own material, but also collecting examples developed in projects around the world.

Do you have a [notebook/project](#) that you want to share? [Contact us!](#)



# Thank you!

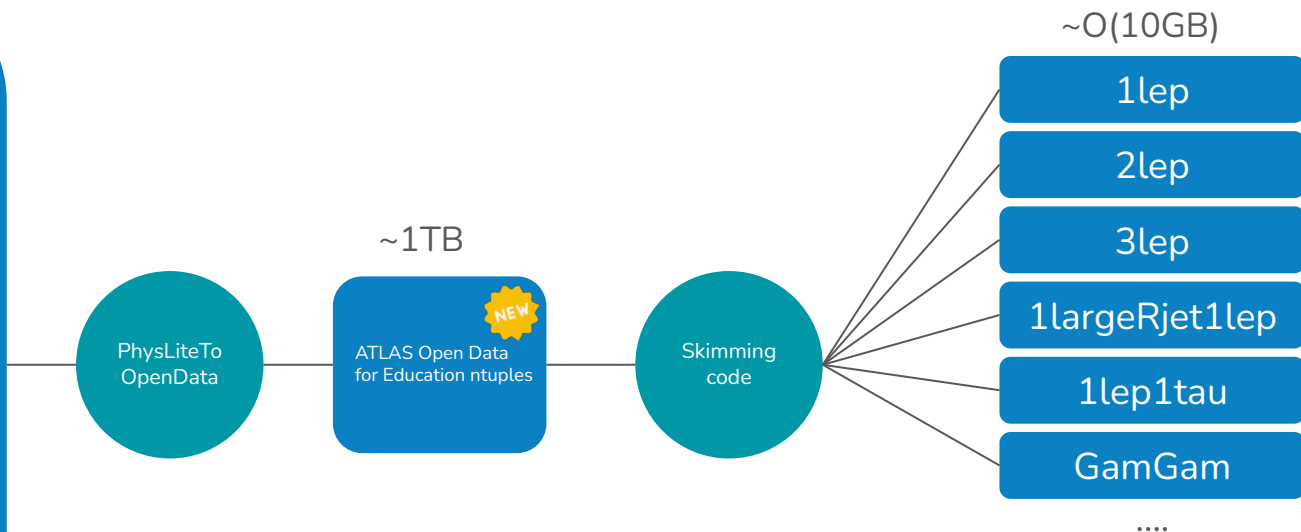
# Backup

# How do we produce ATLAS Open Data? (2024 edition)

~65TB

DISCLAIMER: This diagram is related to the 2024 release, coming soon.

ATLAS Open Data for research



## A new reduced common data format for ATLAS

- **Reduced File Size:** [PHYSLITE](#) targets a file size of **10 kB per event** for data and 12 kB for MC, a 60%-80% reduction compared to previous formats (DAOD\_PHYS are ~30-50 kB in size)
- **CPU Efficiency:** **25% reduction in CPU usage** compared to previous models
- **Unskimmed and Monolithic:** **one-size-fits-all** solution, fitting various use cases with no need for multiple versions
- **Direct Analysis Capability:** PHYSLITE can be **analysed directly**, no need for flat n-tuples and further reducing storage demands

Format	Run 2 MC $\bar{t}\bar{t}$	Run 3 MC $\bar{t}\bar{t}$	Data 16	Data 22
PHYS (kB/event)	33.8	40.9	18.2	20.5
PHYSLITE (kB/event)	13.0	16.1	6.2	6.2

The current file sizes (in kB per event) for PHYS and PHYSLITE for various data and MC campaigns  
Work is ongoing to reduce the size of PHYSLITE further.



## Hybrid platforms

Docker containers: robust, replicable environment

Image available on the [github registry](#)

Documentation and tutorials on [our website](#)

No internet required (after pulling the container and the data) ✓

Data persistence ✓

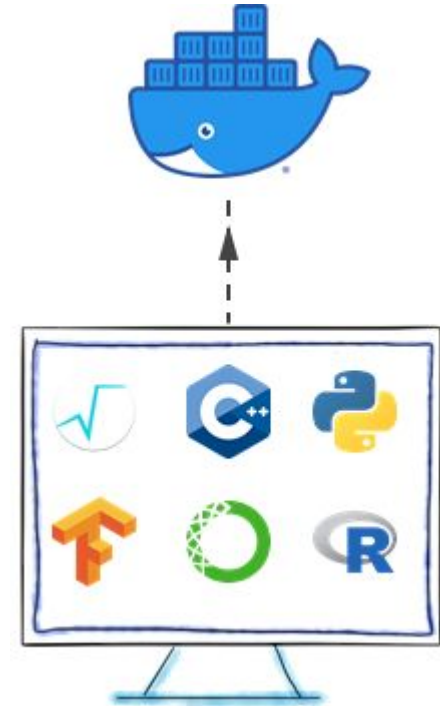
Do not need prerequisites ✗

No timeout time for sessions ✓

Spawn time <1min ✓

Software stack choice ✗

Relies on local computational resources ⚠



# Where do I use ATLAS Open Data?

## “Offline” platforms

Virtual Machines: download it and use it or put it in a USB key and take it where you want

e.g. Image available on the [opendata portal](#)

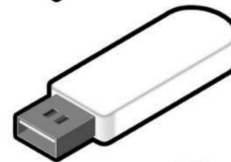
## How to plug in a USB key

- Plug 'n play ✓
- Data persistence ✓
- Do not need prerequisites ✗
- Larger overhead ☠
- No timeout time for sessions ✓
- Spawn time ~O(min) ✓
- Software stack available ✗

**Wrong**



**Wrong**



**Right**

