HEP Data for Everyone:

CERN open data and the ATLAS and CMS experiments

Thomas McCauley

@tpmccauley

University of Notre Dame, USA

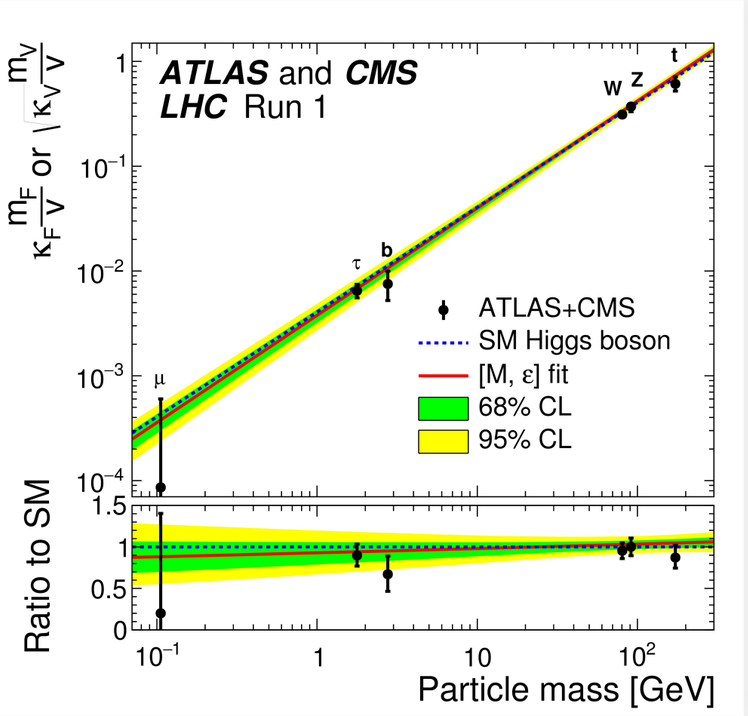
for the ATLAS and CMS Collaborations

https://tpmccauley.github.io/opendata-ichep2016


Who?

ATLAS and CMS

- Two general-purpose experiments at the Large Hadron Collider, CERN
- Broad physics programs
- Currently taking proton-proton collision data at $\sqrt{s}$ = 13 TeV
- http://cern.ch/cms
- http://atlas.cern
- If you want to find out more, ICHEP is the right place




What?

ATLAS and CMS have adopted open-access policies, both of which use common notions of levels of access to data:

- Level 1: data directly related to publications
- Level 2: simplified data formats suitable for education and outreach
- Level 3: “analysis-level” reconstructed data, simulation, and software
- Level 4: raw data and associated software

CMS Policy, ATLAS Policy

- Level 1: data directly related to publications
- **Level 2: simplified data formats suitable for education and outreach**
- **Level 3: “analysis-level” reconstructed data, simulation, and software**
- Level 4: raw data and associated software


Why?

CERN makes public first data of LHC experiments

20 Nov 2014

Geneva, 20 November 2014. CERN[1] launched today its Open Data Portal where data from real collision events, produced by the LHC experiments will for the first time be made openly available to all. It is expected that these data will be of high value for the research community, and also be used for education purposes.

"Launching the CERN Open Data Portal is an important step for our Organization. Data from the LHC programme are among the most precious assets of the LHC experiments, that today we start sharing openly with the world. We hope these open data will support and inspire the global research community, including students and citizen scientists," said CERN Director General Rolf Heuer.

The principle of openness is enshrined in CERN's founding Convention, and all LHC publications have been published Open Access, free for all to read and re-use. Widening the scope, the LHC collaborations recently approved Open Data policies and will release collision data over the coming years.

Open data benefits **the public,**

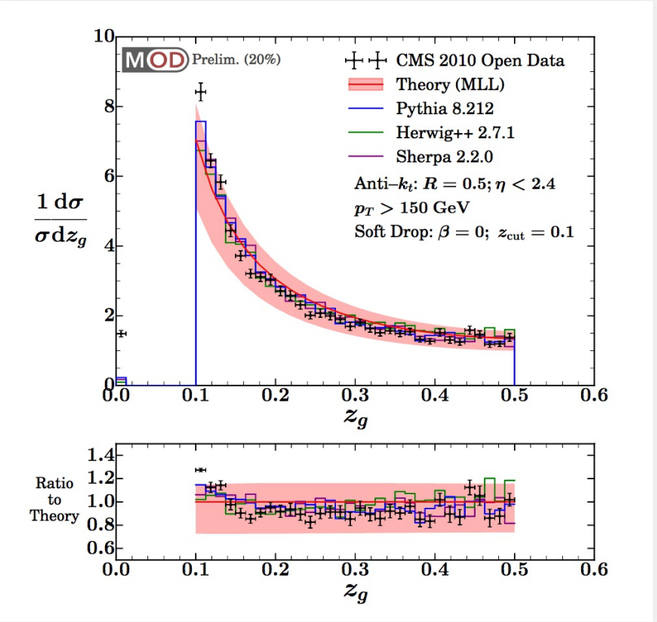
- Education
- Public engagement

Open data benefits the public, **the experiments,**

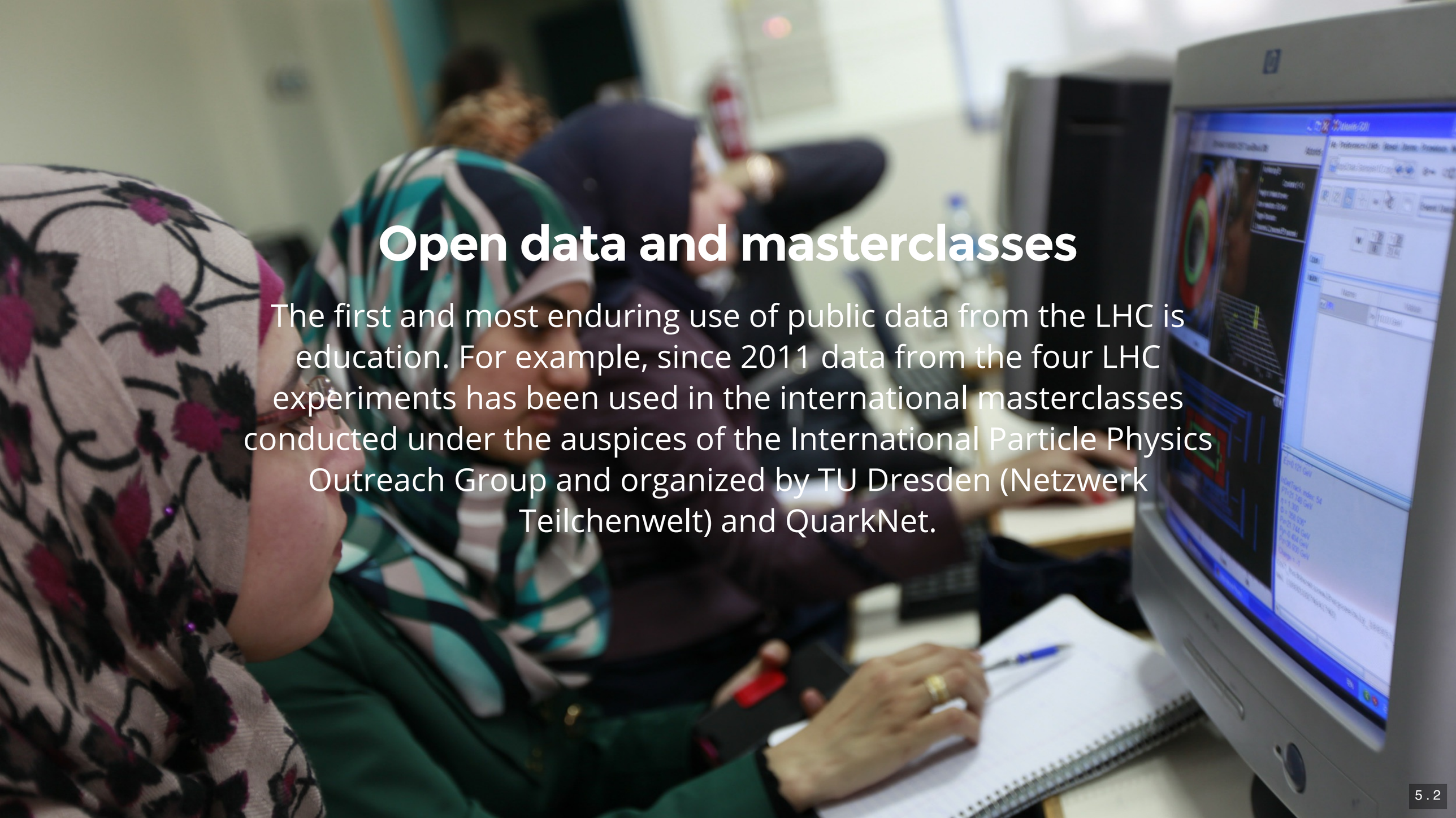
- Education
- Public engagement
- Data and analysis preservation

Open data benefits the public, the experiments, **and science**

- Education
- Public engagement
- Data and analysis preservation
- Physicists
- Data scientists
- Citizen scientists

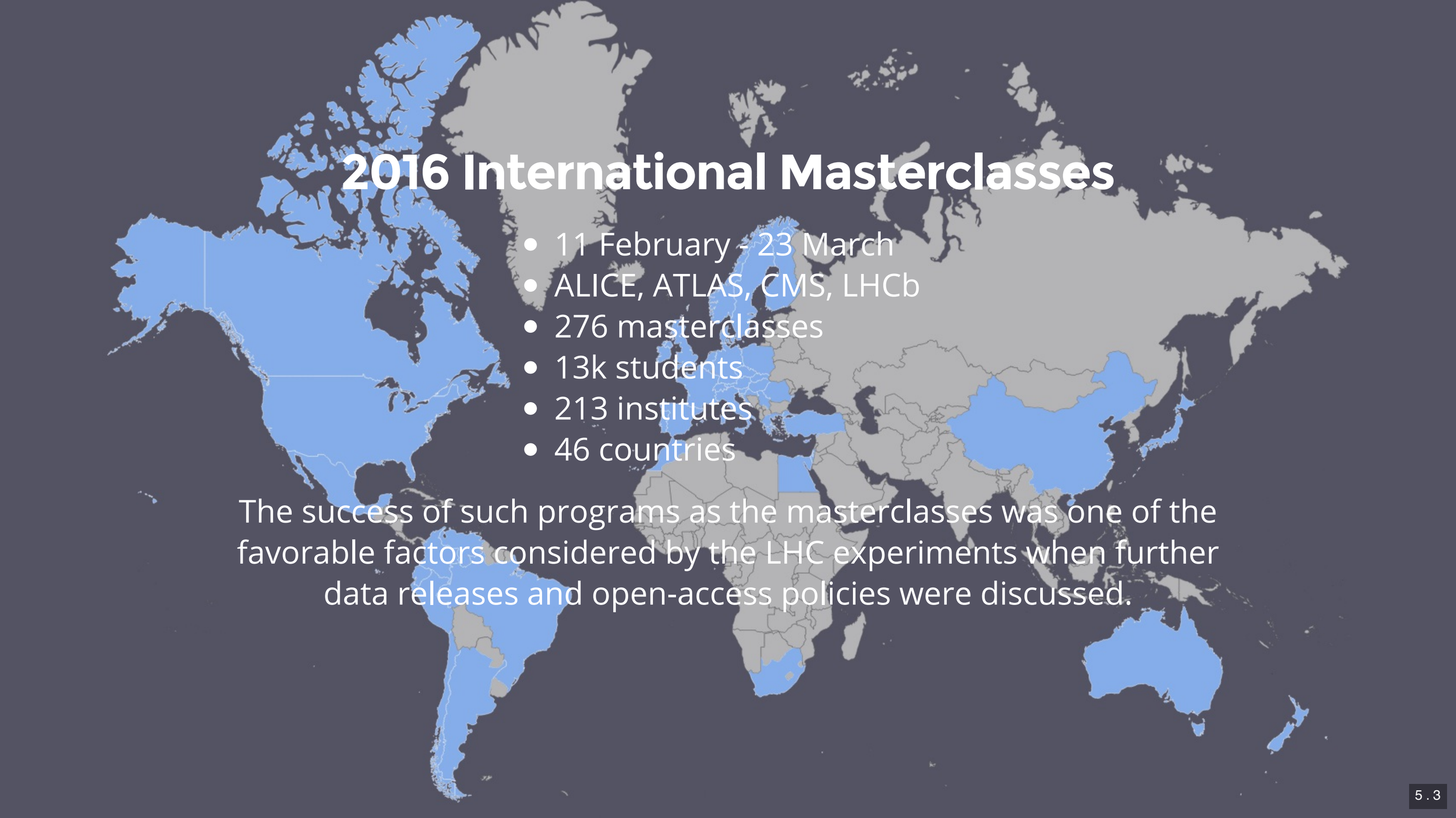


A bit of history...



Open data and masterclasses

The first and most enduring use of public data from the LHC is education. For example, since 2011 data from the four LHC experiments has been used in the international masterclasses conducted under the auspices of the International Particle Physics Outreach Group and organized by TU Dresden (Netzwerk Teilchenwelt) and QuarkNet.



2016 International Masterclasses

- 11 February - 23 March
- ALICE, ATLAS, CMS, LHCb
- 276 masterclasses
- 13k students
- 213 institutes
- 46 countries

The success of such programs as the masterclasses was one of the favorable factors considered by the LHC experiments when further data releases and open-access policies were discussed.


How?

Challenges

"Making the data public does not make them any simpler".

In order to make it useful and useable for the public several challenges have to be overcome:

- Data volume: datasets of up to hundreds of terabytes is a lot to handle and releases will only get larger
- Data complexity: reflects the complexity of the experiments themselves
- Software environment: large, custom-made software frameworks are required to read and analyze data; not-inconsiderable software skill is needed as well
- Physics knowledge needed: in an experiment those analyzing data either have or are working towards a PhD in physics

The details of how ATLAS and CMS responded to these challenges can and do differ but in general the following common approaches were used:

- Provide access and analyses at different levels of knowledge and expertise
- Where possible and applicable provide simplified datasets
- Provide good documentation
- Make example analysis code available
- Allow for analysis and visualization via the browser: *e.g.* histograms, event displays
- Provide virtual machines with software environment for more advanced users


CMS
JLG LIFTLUX 153-12

CMS data and tools

- Intended audience: all levels of the public
- There is therefore a spectrum of levels of access and complexity
- Also attempt to address issues of data cataloging, validation, and preservation
- Data and tools available via CERN Open Data Portal

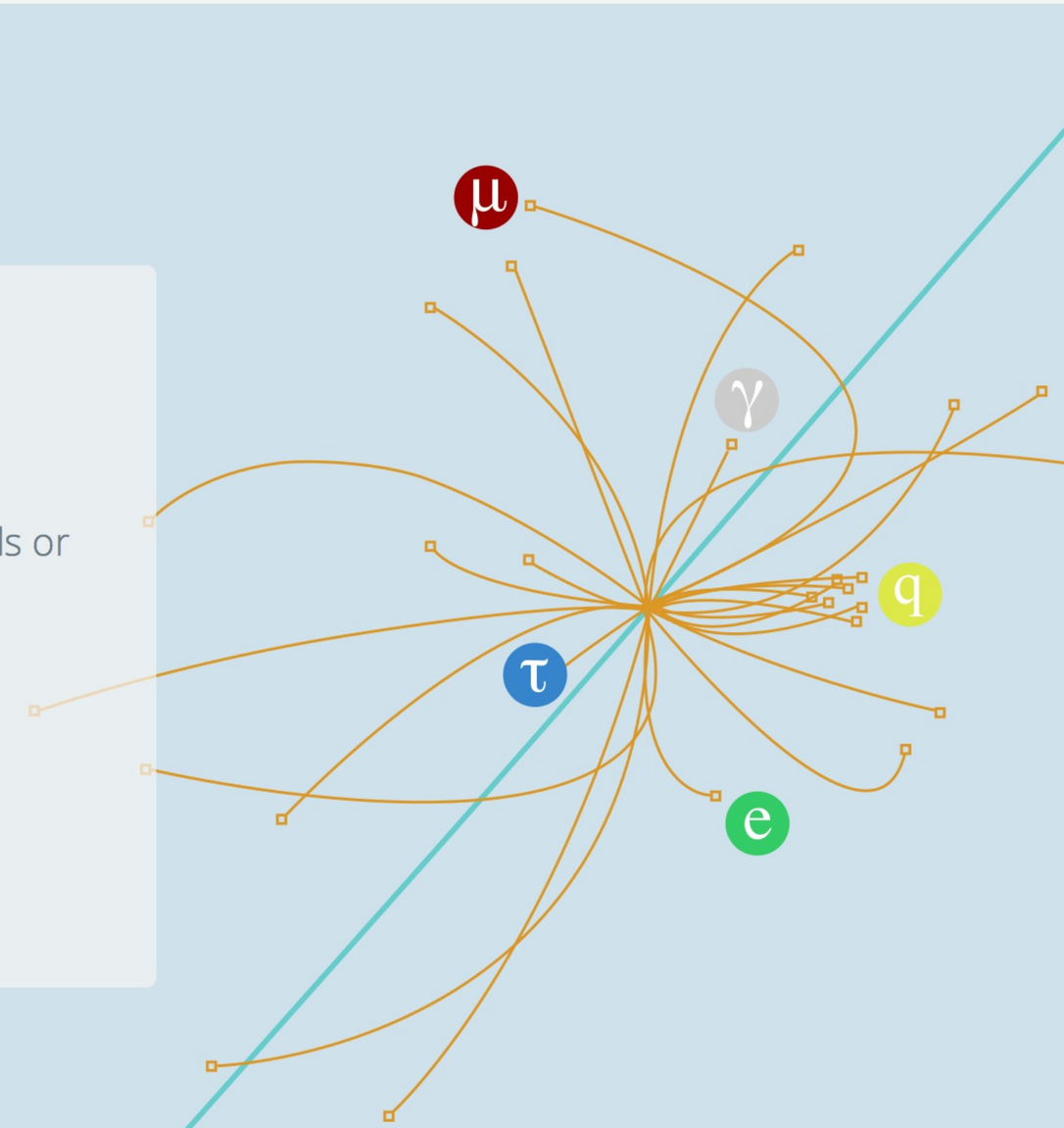
CMS data and tools (the details)

- Nov 2014: half of 2010 pp collision data at $\sqrt{s}$ = 7 TeV released, 27 TB in size, equivalent to tens of pb^{-1}
- April 2016: half of 2011 pp collision data at $\sqrt{s}$ = 7 TeV released, 100 TB in size, equivalent to ~ 2.5 fb^{-1}, along with 200 TB of Monte Carlo samples
- "Level 3" primary datasets are released in CMS AOD (Analysis Object Data) ROOT format
- Derived (*i.e.* reduced) datasets in csv, JSON, and CMS PAT (Physics Analysis Tool) ROOT format
- Extensive documentation includes data validation, trigger, and detector condition information as well as analysis code examples
- Virtual machines with CMS software environments needed for AOD analysis
- Data is stored in EOS and can be downloaded directly or accessed via XRootD

CERN Open Data Portal

http://opendata.cern.ch

- Portal is divided into two main areas: “Education” and “Research”
- Datasets are distinguished as either “primary" or “derived” (roughly falling into Level 3 and Level 2 categories, respectively)
- Philosophy: include and build upon the previous and current success of public data in education and outreach but also include the possibility for more in-depth, complex analysis
- Web applications for immediate use are available: event display and histogram tool
- Built with Invenio digital library software
- All four LHC experiments use the portal to various extents; this section focuses on CMS



Education

Visualise events, check reconstructed data, run tools or build your own!

Start learning

Research

Get the genuine working environments, virtual machines and datasets to start your research

Start analysing

Education

The CMS (Compact Muon Solenoid) experiment is one of two large general-purpose detectors built on the Large Hadron Collider (LHC). Its goal is to investigate a wide range of physics such as the characteristics of the Higgs boson, extra dimensions or dark matter.

Explore CMS >

ALICE (A Large Ion Collider Experiment) is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms. More than 1000 scientists are part of the collaboration.

Explore ALICE >

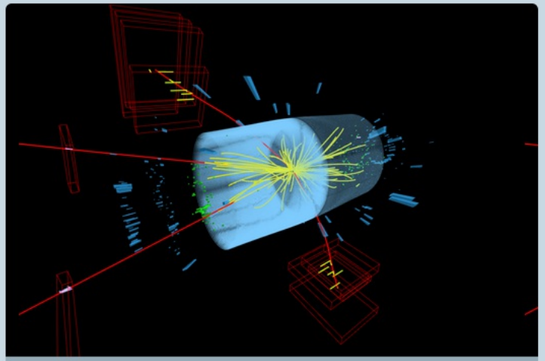
The ATLAS (A Toroidal LHC ApparatuS) experiment is a general-purpose detector exploring topics like the properties of the Higgs-like particle, extra dimensions of space, unification of fundamental forces and evidence for dark matter candidates in the Universe.

Explore ATLAS >

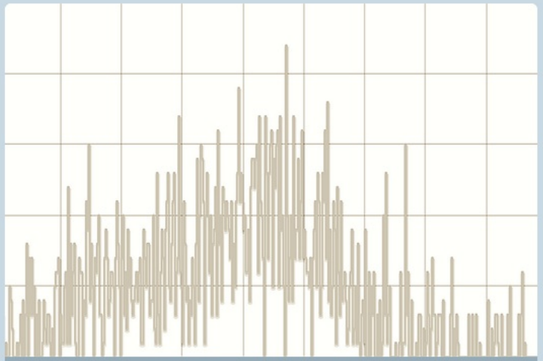
The LHCb (Large Hadron Collider beauty) experiment aims to record the decay of particles containing b and anti-b quarks, known as B mesons. The detector is designed to gather information about the identity, trajectory, momentum and energy of each particle.

Explore LHCb >

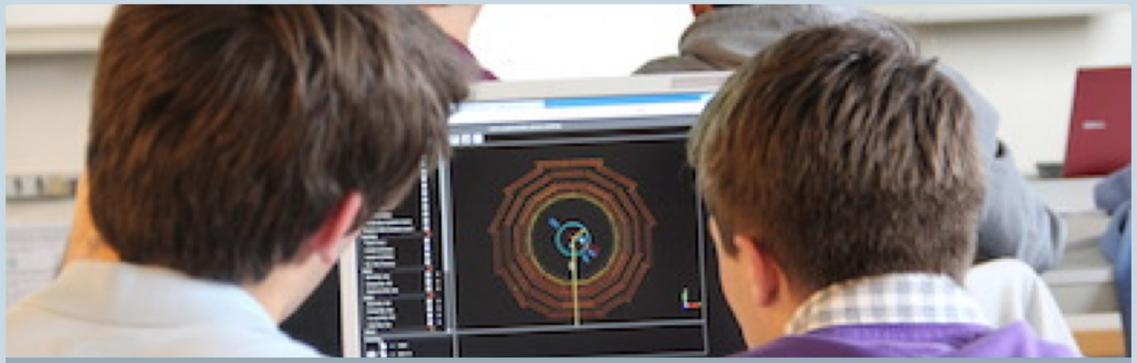
For education purposes, the complex primary data need to be processed into a format (examples below) that is good for simple applications. Get in touch if you wish to build your own applications similar to those shown here



Visualise events >

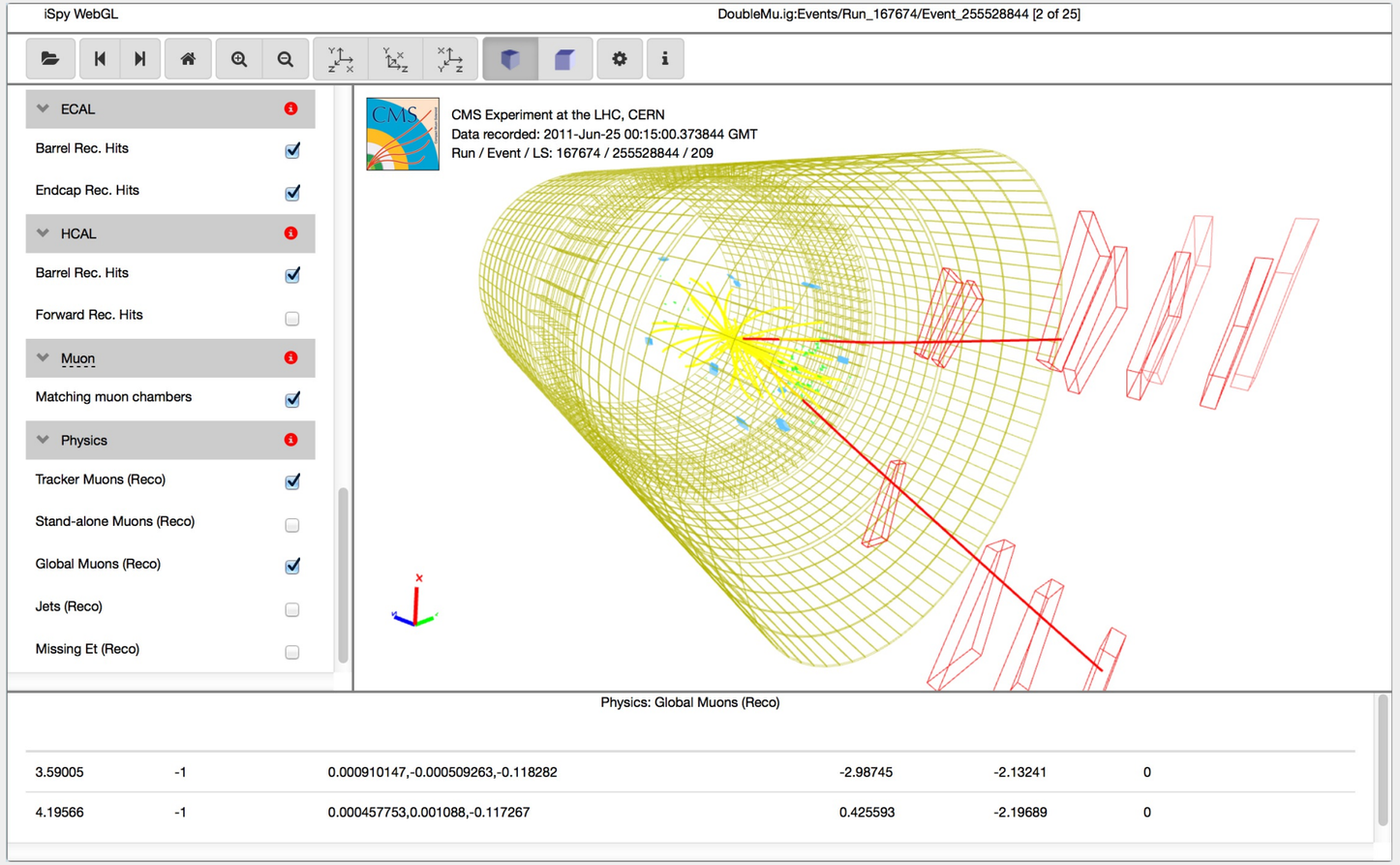


Visualise histograms >



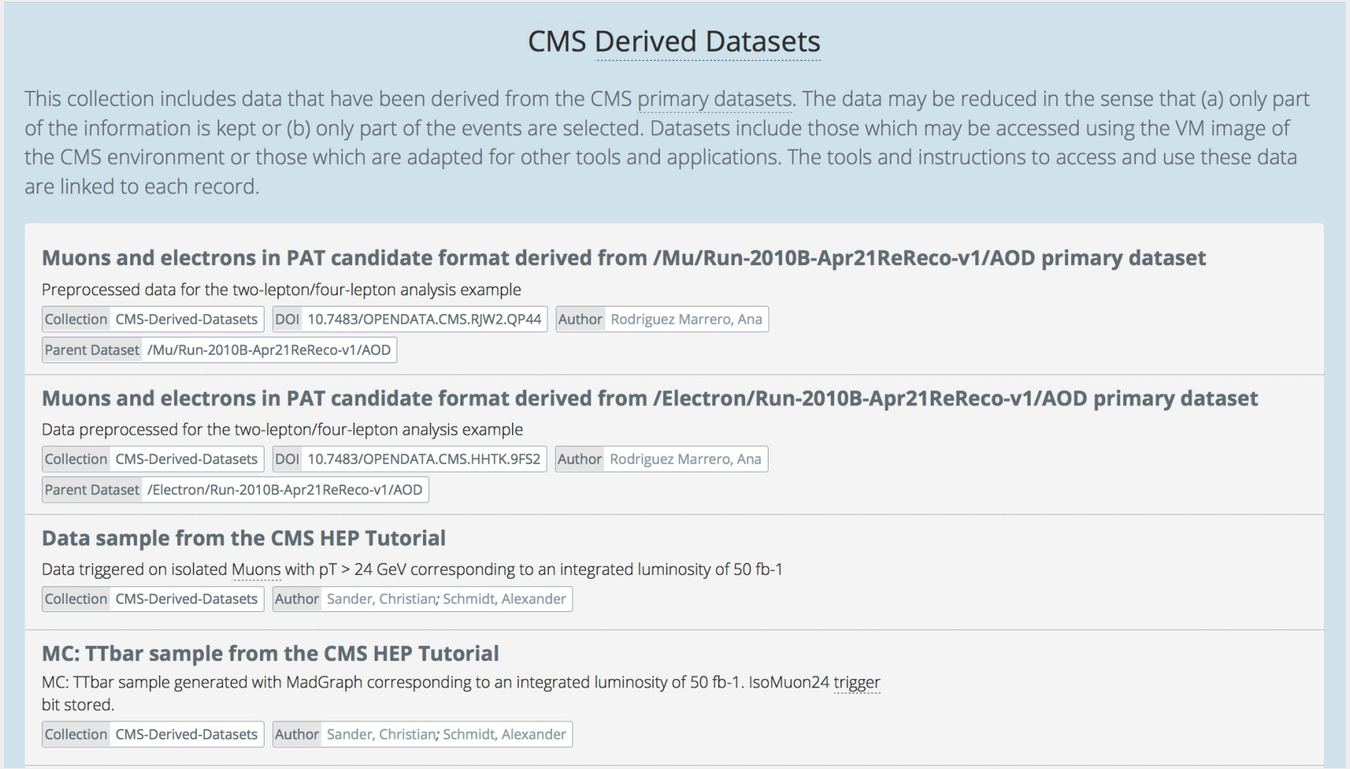
Learning Resources >

Education



Visualize and explore events from the datasets with an interactive event display

Education



CMS Derived Datasets

This collection includes data that have been derived from the CMS primary datasets. The data may be reduced in the sense that (a) only part of the information is kept or (b) only part of the events are selected. Datasets include those which may be accessed using the VM image of the CMS environment or those which are adapted for other tools and applications. The tools and instructions to access and use these data are linked to each record.

Muons and electrons in PAT candidate format derived from /Mu/Run-2010B-Apr21ReReco-v1/AOD primary dataset

Preprocessed data for the two-lepton/four-lepton analysis example

Collection CMS-Derived-Datasets | DOI 10.7483/OPENDATA.CMS.RJW2.QP44 | Author Rodriguez Marrero, Ana

Parent Dataset /Mu/Run-2010B-Apr21ReReco-v1/AOD

Muons and electrons in PAT candidate format derived from /Electron/Run-2010B-Apr21ReReco-v1/AOD primary dataset

Data preprocessed for the two-lepton/four-lepton analysis example

Collection CMS-Derived-Datasets | DOI 10.7483/OPENDATA.CMS.HHTK.9FS2 | Author Rodriguez Marrero, Ana

Parent Dataset /Electron/Run-2010B-Apr21ReReco-v1/AOD

Data sample from the CMS HEP Tutorial

Data triggered on isolated Muons with pT > 24 GeV corresponding to an integrated luminosity of 50 fb-1

Collection CMS-Derived-Datasets | Author Sander, Christian; Schmidt, Alexander

MC: TTbar sample from the CMS HEP Tutorial

MC: TTbar sample generated with MadGraph corresponding to an integrated luminosity of 50 fb-1. IsoMuon24 trigger bit stored.

Collection CMS-Derived-Datasets | Author Sander, Christian; Schmidt, Alexander

Recently high school teachers as part of the CERN High School Teacher program successfully used CMS derived data to come up with exercises for the high-school level

Research

opendata CERN

ABOUT SEARCH EDUCATION RESEARCH

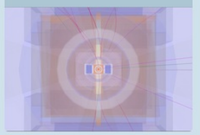
Search

Research › CMS

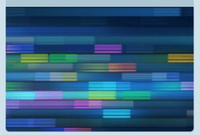
CMS Open Data are available in the same format as used in analysis by CMS physicists. A CMS-specific analysis framework is needed, and it is provided as a Virtual Machine image with the CMS analysis environment. The data can be accessed directly through the VM image. Basic information of the data contents is provided in About CMS and in About CMS Physics Objects. The original data are in primary datasets, i.e. no selection nor identification criteria have been applied (apart from the trigger decision), and these have to be applied in the subsequent analysis step. The 2011 data release includes simulated Monte Carlo datasets, but no simulated datasets are provided for the 2010 release.

VMs

Getting started!



Software and tools



CMS Primary Datasets

CMS primary datasets are AOD (Analysis Object Data) files, which contain the information that is needed for analysis

Years: **2010**, **2011**

Total records:

33

CMS Simulated Datasets

This collection contains CMS Simulated Datasets.

Years: **2010**, **2011**

Total records:

381

CMS Derived Datasets

This collection includes data that have been derived from the CMS primary datasets

Years: **2010**, **2011**

Total records:

59

CMS Tools

This collection includes tools with which the CMS open data can be accessed and used

Years: **2010**, **2011**

Total records:

17

CMS Validation Utilities

This collection contains CMS Validation Utilities.

Years: **2010**, **2011**

Total records:

5

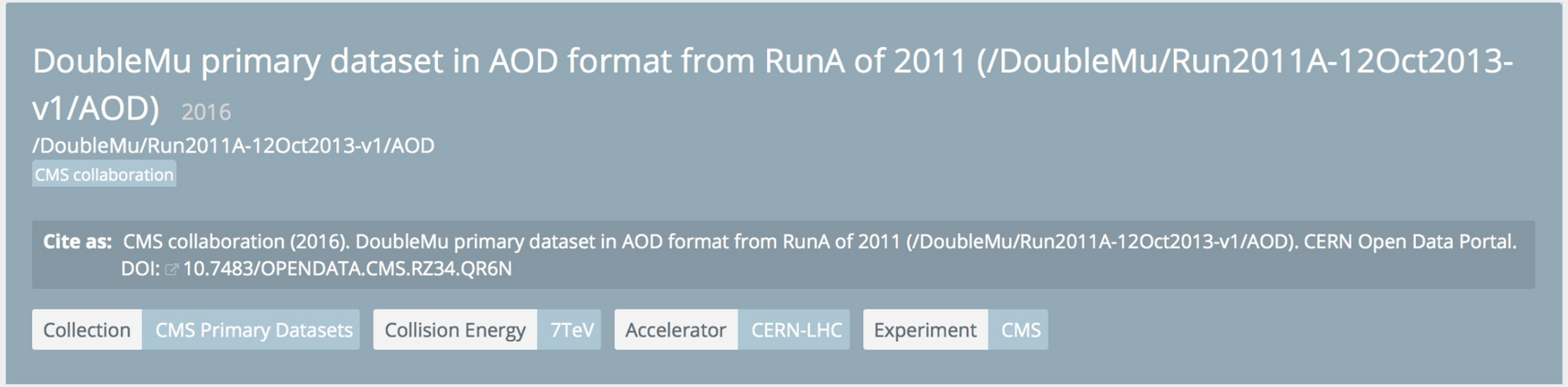
CMS Learning Resources

This collection includes learning resources that use CMS public data

Total records:

6

Research



Data are identified with persistent, citable digital object identifiers (DOI) and are released under the Creative Commons CC0 waiver, essentially releasing it into the public domain.

Research

Example code to produce the di-muon spectrum from a CMS 2010 primary dataset

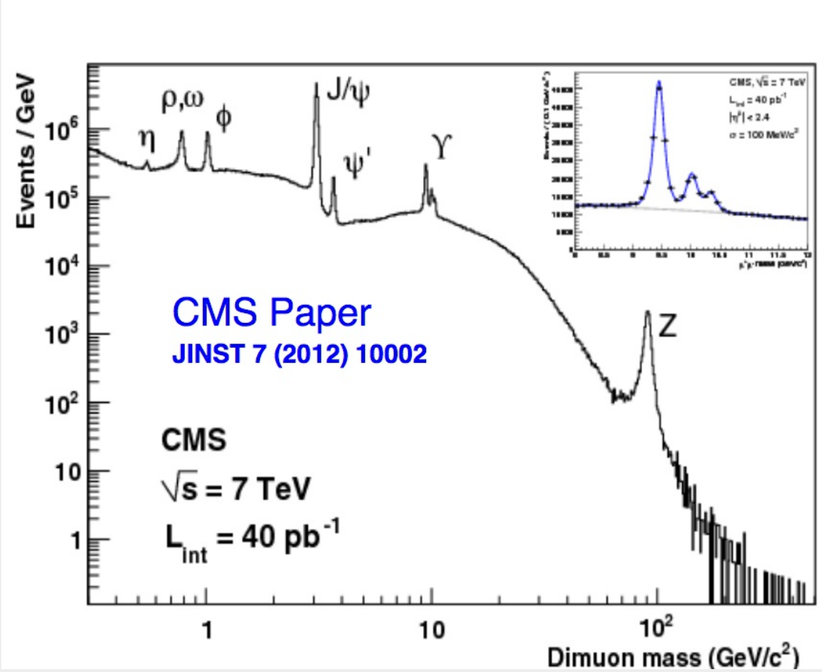
Geiser, Achim; Dutta, Irene; Hirvonsalo, Harri; Sheeran, Bridget

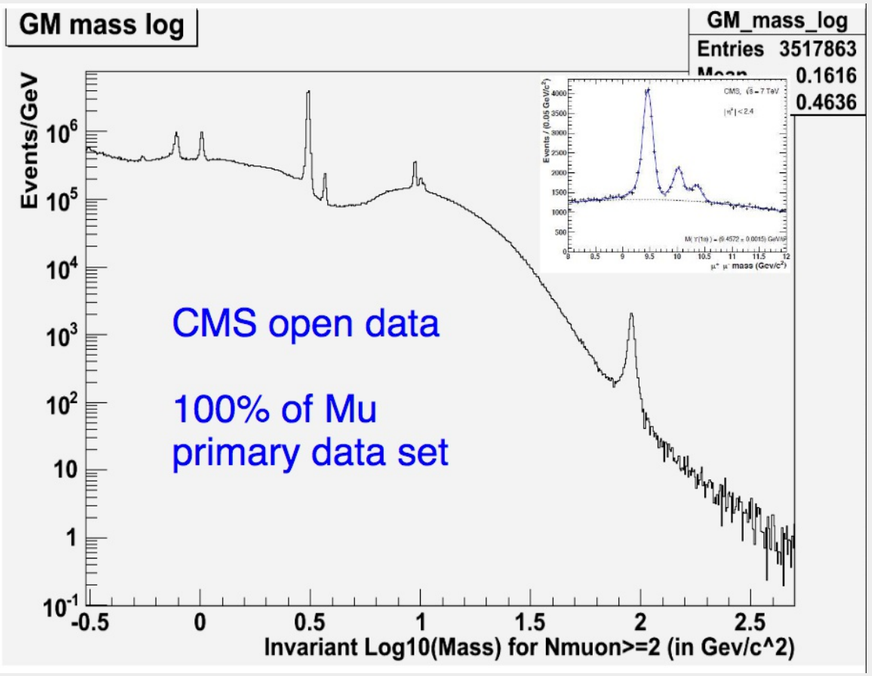
Cite as: Geiser, A. , Dutta, I. , Hirvonsalo, H. & Sheeran, B. (2016). Example code to produce the di-muon spectrum from a CMS 2010 primary dataset. CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.TF26.KG2D

Collection: CMS Tools | Accelerator: CERN-LHC | Experiment: CMS

```
// WHAT: Fill histograms for the following attributes from the current
//       globalMuon-Track:
// - p (momentum vector magnitude)
// - pt (track transverse momentum)
// - eta (pseudorapidity of momentum vector)
// - chi-square
// - ndof (number of degrees of freedom of the fit)
// - normalizedChi2 (normalized chi-square == chi-squared divided by ndof
//                   OR chi-squared * 1e6 if ndof is zero)
    h1->Fill(it->p());
    h2->Fill(it->pt());
    h3->Fill(it->eta());
    h4->Fill(it->phi());
    h53->Fill(it->chi2());
    h54->Fill(it->ndof());
    h55->Fill(it->normalizedChi2());
```

Research


Events / GeV
ρ,ω
η
φ
J/ψ
ψ'
ϒ
Z
CMS Paper
JINST 7 (2012) 10002
CMS
√s = 7 TeV
L_int = 40 pb^-1
Dimuon mass (GeV/c^2)


GM mass log
GM_mass_log
Entries 3517863
0.1616
0.4636
Events/GeV
CMS open data
100% of Mu
primary data set
Invariant Log10(Mass) for Nmuon>=2 (in Gev/c^2)


ATLAS

ATLAS data and tools

- Initial focus: undergraduate and postgraduate students (but eventually to expand target audience)
- Within this scope, provide access at 3 levels: from visualizations, to web analysis, to more complex analysis
- Data and tools available via the ATLAS Open Data Page: released on 29 July, just last week!

ATLAS data and tools (the details)

- 1 fb^{-1} of 2012 pp collision data at $\sqrt{s}$ = 8 TeV and Monte Carlo
- Datasets: Electron/gamma and muon
- ROOT TTree format
- Interactive visualization and analysis tools via browser
- ROOT + Jupyter notebooks
- python-based analysis framework code
- Virtual machines with software available
- Data and tools (including VM): ~14 GB; all can fit on a USB stick
- Data and VMs also made available via CERN Open Data Portal

ATLAS Open Data Page

http://atlasopendata.web.cern.ch

ATLAS Experiment

DOWNLOAD COMMUNITY

open data ATLAS

Get Started
Documentation, Histogram Analyser, ROOTbrowser

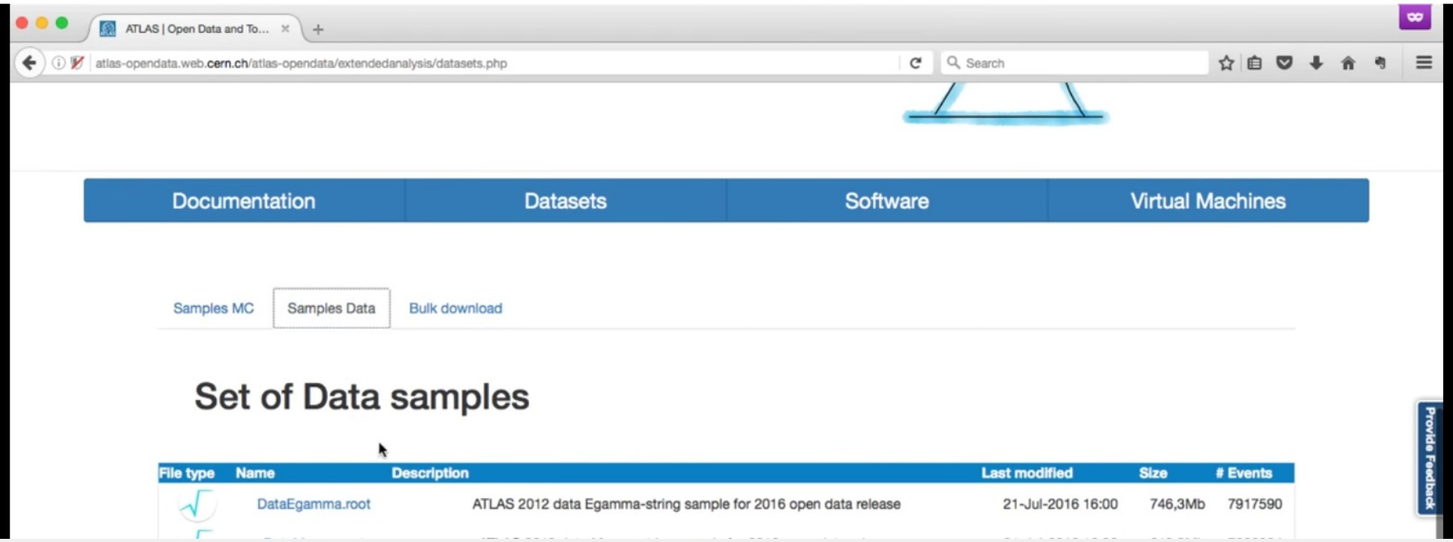
Web Analysis
Documentation, Online ROOTbooks

Data & Tools
Documentation, Datasets, Software, Virtual Machines

Access Open Data from the ATLAS Experiment at CERN

The ATLAS data from 100 trillion proton collisions is now public! This marks the world's first open release of 8 TeV data, gathered from the Large Hadron Collider in 2012.

ATLAS Open Data guides you through how to visualise the data, how to download and use the data, and even provides open-source software for you to make your own discoveries. **Check the introductory video and get started now!**



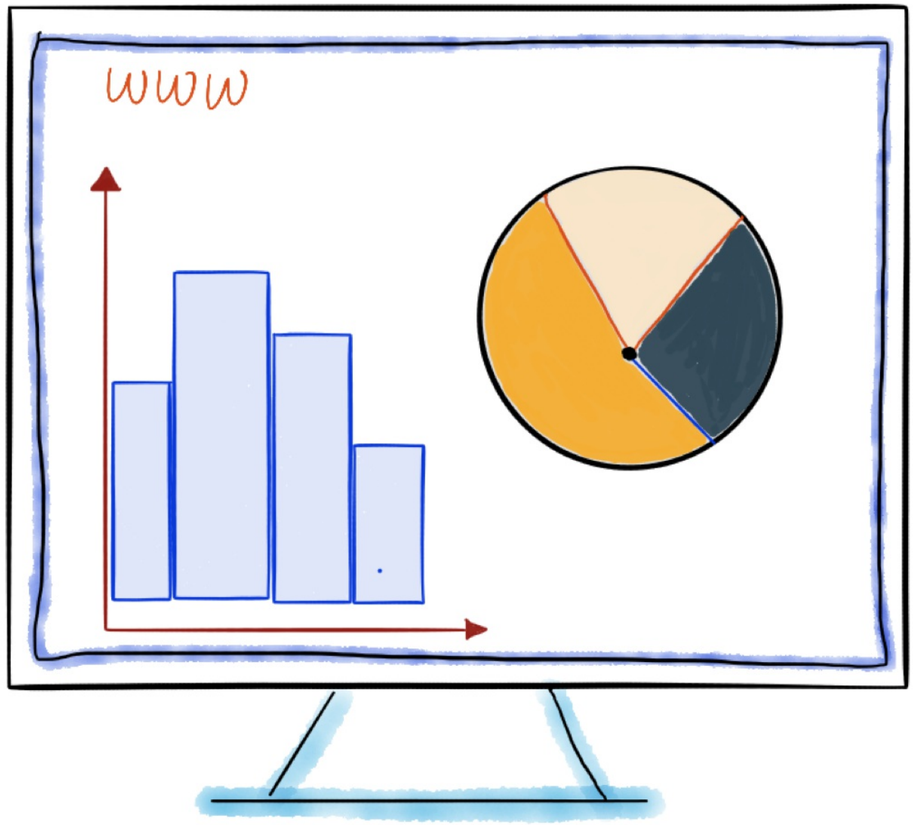
Provide Feedback

Level 1: **Get Started**

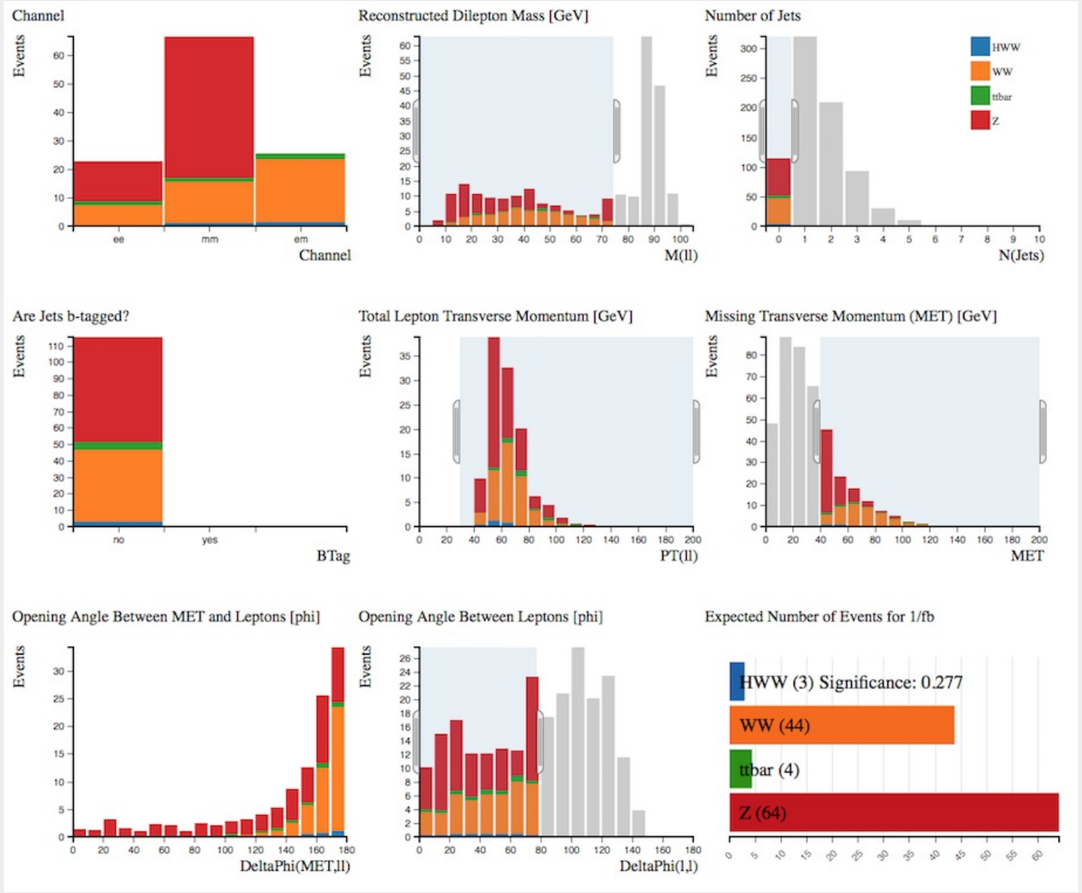
Physicists at the ATLAS Experiment visualise collision data with histograms. They are used in every publication, from simple analyses to headline-making discoveries. *In this section, you will learn how the data is visualised.*

Explore:

- **Documentation:** a step-by-step guide to using Histogram Analyser and ROOTbrowser
- **Histogram Analyser:** a web based tool for fast, cut-based analysis of data. Visualise data using online histograms
- **ROOTbrowser:** a web based tool for displaying and analysing data. Visualise data online
- **Live events:** see live events from the ATLAS experiment



Get Started



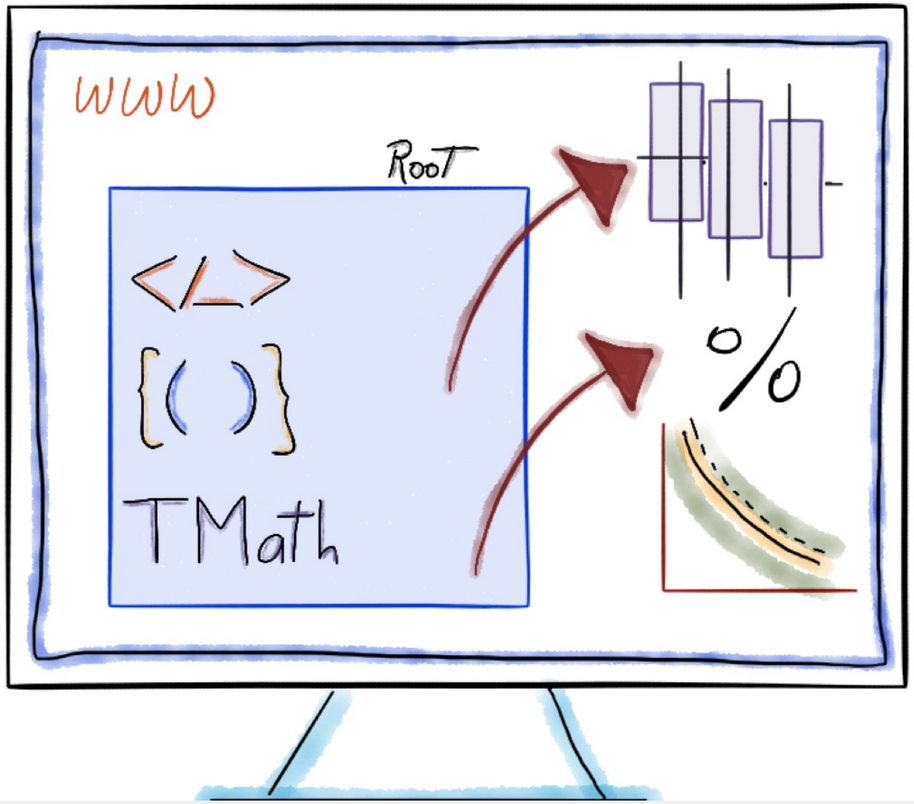
Explore correlations between variables in the datasets with histogram tool

Level 2: **Web Analysis**

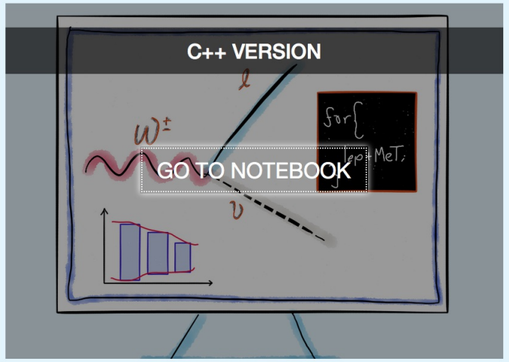
The ATLAS Experiment has made 7 analyses available to help you get started with your own research! These analyses range from measuring **Standard Model** particles such as the Higgs boson to searching for a Beyond the Standard Model particle. Avoid local installations by using notebooks in a Software as a Service environment on your computer or in the Cloud.

Explore:

- **Documentation:** a step-by-step guide to using, creating and executing ROOT notebooks
- **ROOT notebooks (ROOTbooks):** use Jupyter technology and the power of ROOT to review, execute and develop your own analysis directly in your browser
- **Executable ROOTbooks:** execute, edit and save ROOTbooks using our datasets and examples. You can create your own notebooks as well!

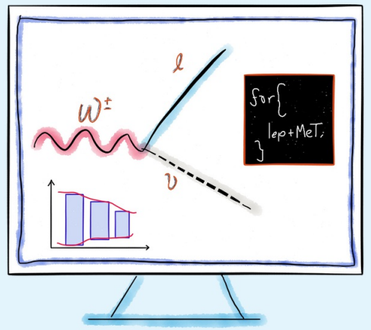


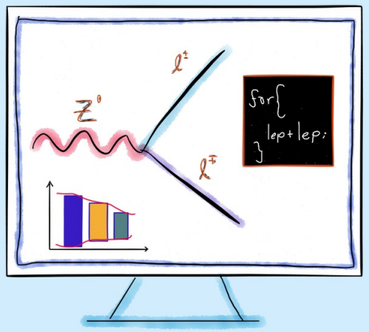
Web Analysis



The W Analysis ROOTbook

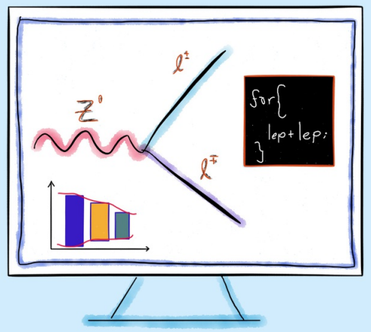
The W boson analysis is intended to provide an example for a high statistics analysis using the ATLAS open data dataset. Furthermore it tests the description of the real data by the simulated W boson data which represents the most extensive dataset in terms of luminosity.





The Z Analysis ROOTbook

Many analyses selecting leptons suffer from Z + jets as a contributing background due to its large production cross section. It is therefore vital to check the correct modelling of this process by the Monte-Carlo simulated data. It is important to measure well known Standard Model particles, to confirm that we understand properly the detector and software. We are then ready to search for new physics.



nbviewer.jupyter.org/github/artfisica/rootbinder/tree/master/notebooks/SummerStudents/

Provide Feedback

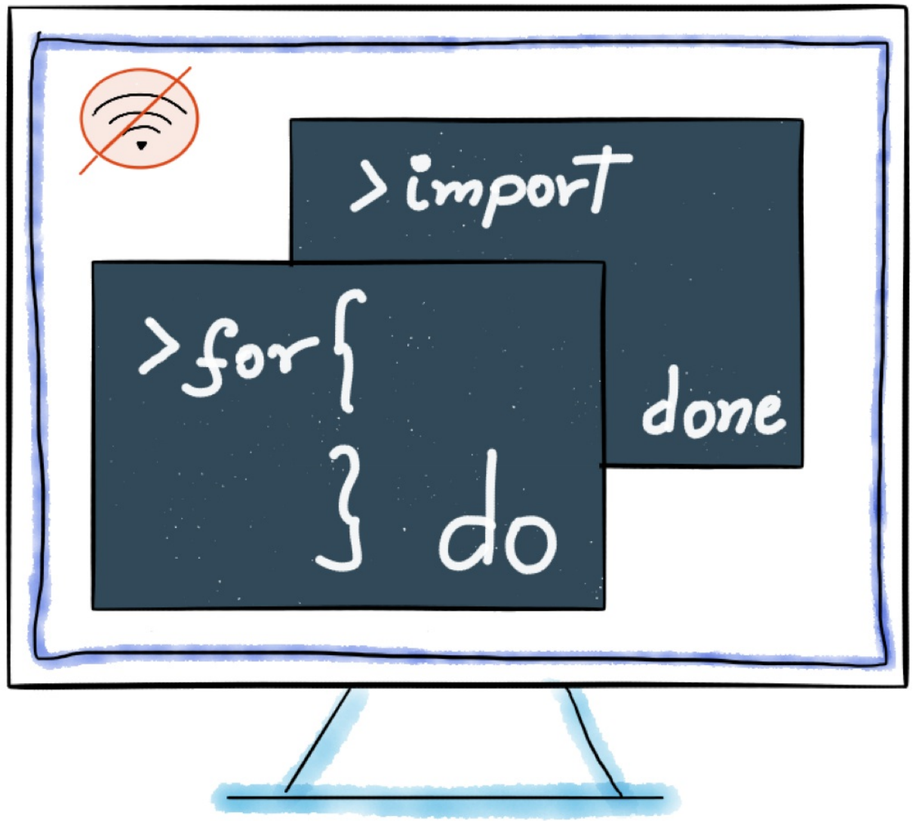
C++ and python notebooks available for various analyses: W, Z, $t\bar{t}$, WZ, ZZ, $H \rightarrow WW$, Z'

Level 3: **Data & Tools**

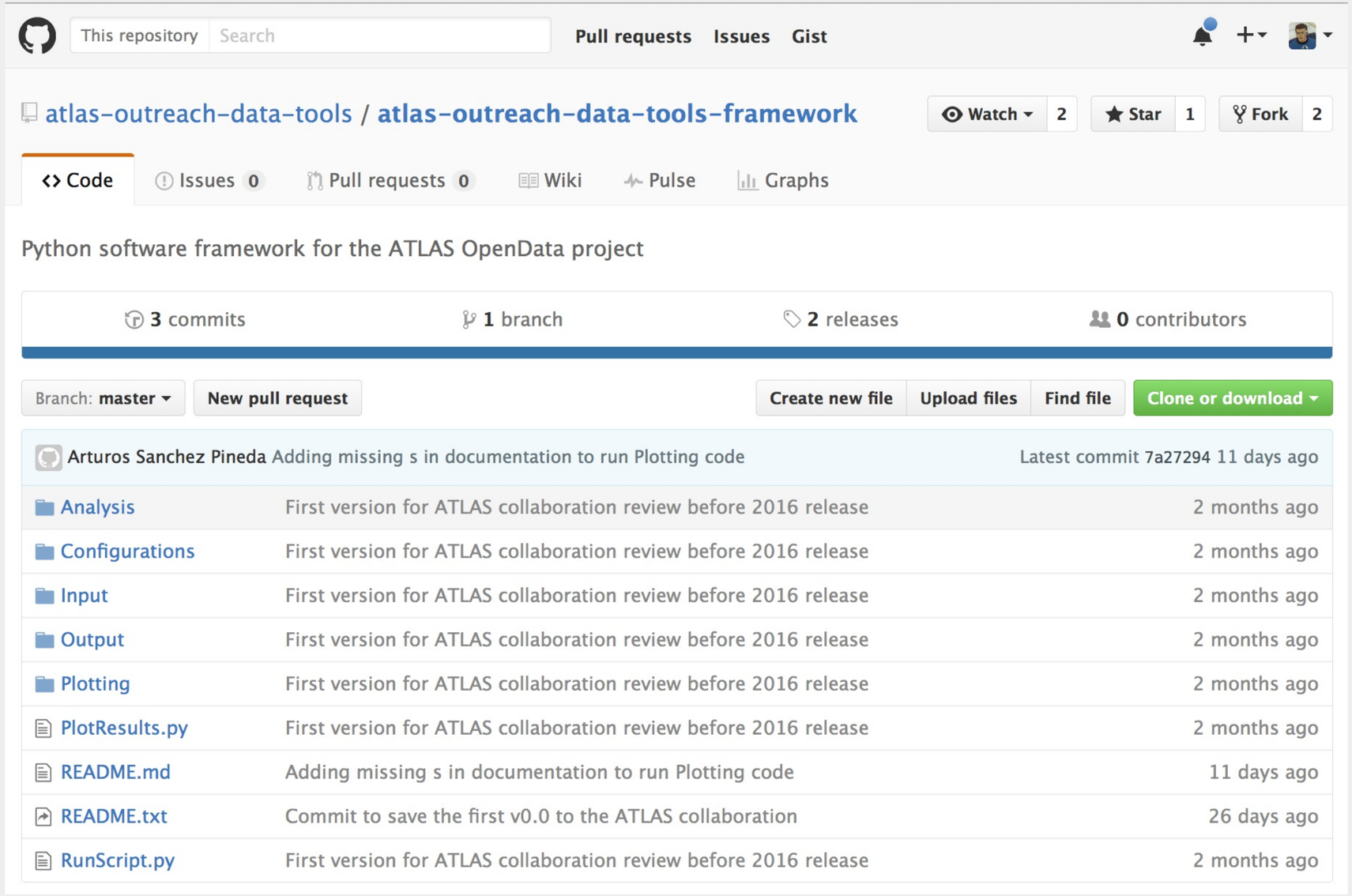
Now that you have learned to visualise data and use code for analysis, you are ready to take an in-depth look at ATLAS data. **Start your analysis now!** In this section, you can download the full datasets, install a virtual machine and learn how to execute analysis software.

Explore:

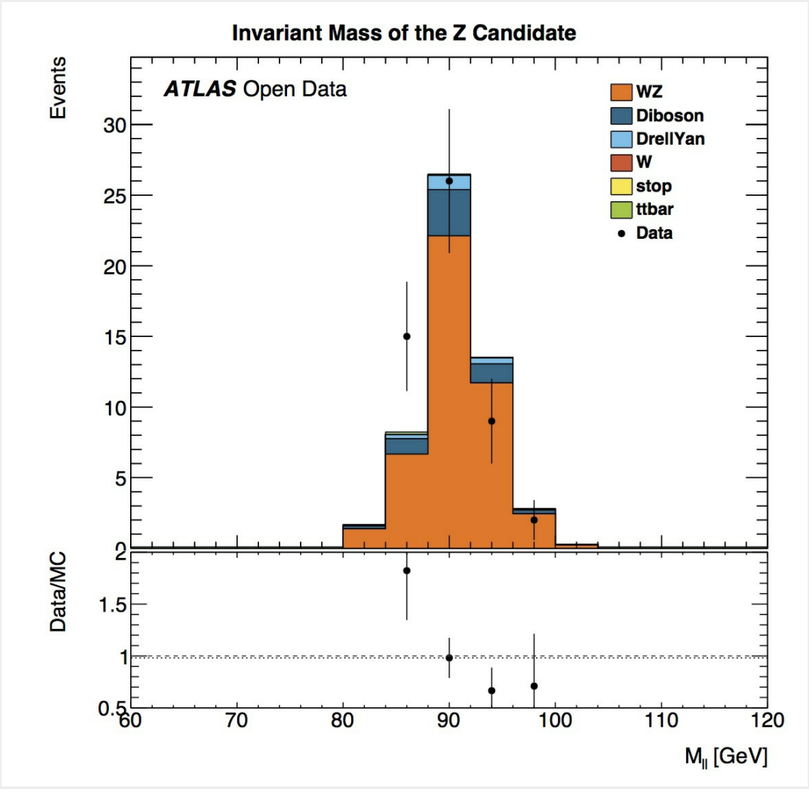
- **Documentation:** a step-by-step guide to downloading datasets, software and virtual machines
- **Datasets:** download the ATLAS datasets
- **Software:** download and run analysis software
- **Virtual Machines:** download and prepare a virtual machine to run on your computer

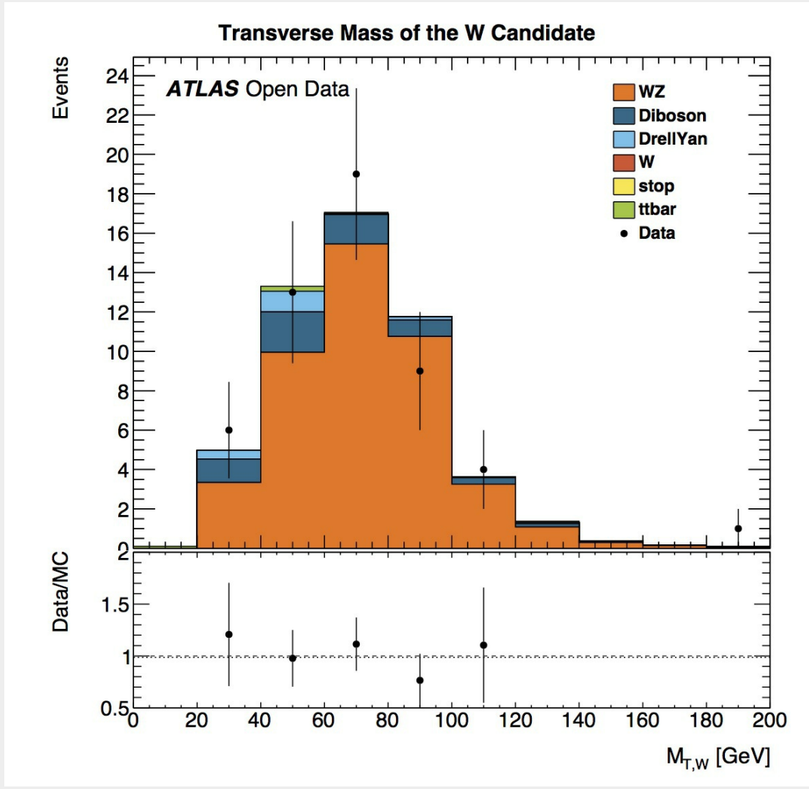


Data and Tools



Data and Tools





ATLAS open data **Community**

Join the ATLAS open data community!

Explore:

- **Forum:** Join the forum and share your experiences and successes with fellow ATLAS open data users. Join or start a thread to ask for tips, suggest changes, report bugs...
- **Frequently Asked Questions (FAQs):** Find answers to common questions
- **Contact:** Use our contact form to get in touch with the ATLAS open data team

Future plans

- For CMS: to continue with regular data releases
- For ATLAS: to work towards a second release of data (13 TeV)
- Improve the data analysis tools available to the public
- Develop accompanying educational material data
- Overall, enable as many as possible to use and enjoy the data

Acknowledgements

- ATLAS Collaboration
- CMS Collaboration
- CERN Scientific Information Services
- CERN Invenio team


Thank you