ATLAS Open Data

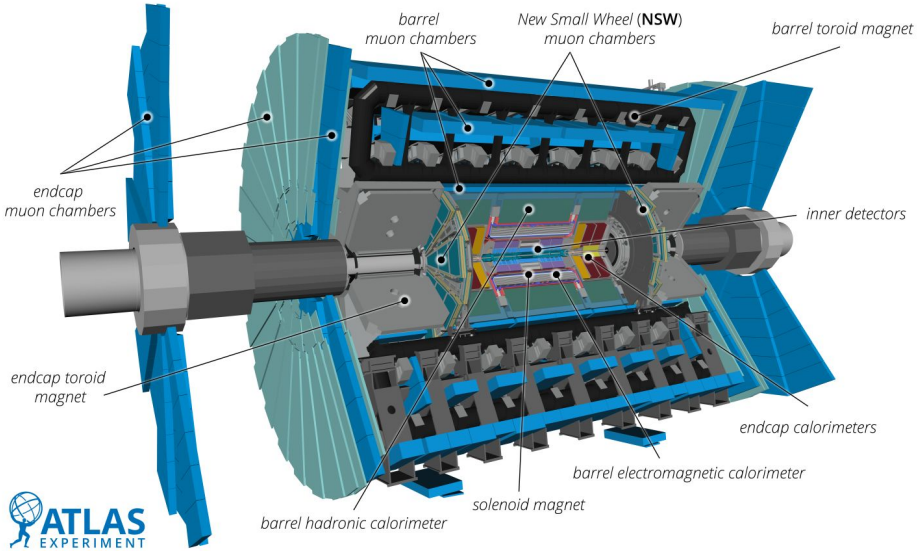
Developing Education and Outreach Resources From Research Data

Attila Krasznahorkay

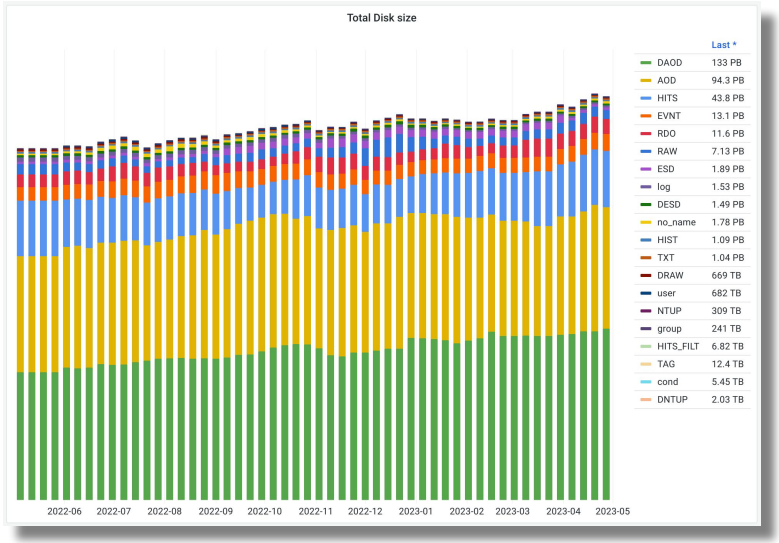
on behalf of the ATLAS Outreach Team

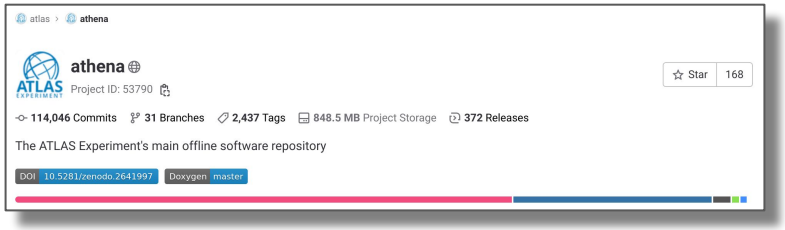
The ATLAS Experiment

- ATLAS is one of the two general purpose experiments at the Large Hadron Collider
  - Looking for all kinds of physics, with a special emphasis on Higgs and Beyond Standard Model physics
- Physics analyses require hundreds of petabytes of data storage and hundreds of thousands of CPUs for running millions of lines of C++ code
- The ATLAS OpenData Project attempts to bring a glimpse into this for students and non-physicists in general



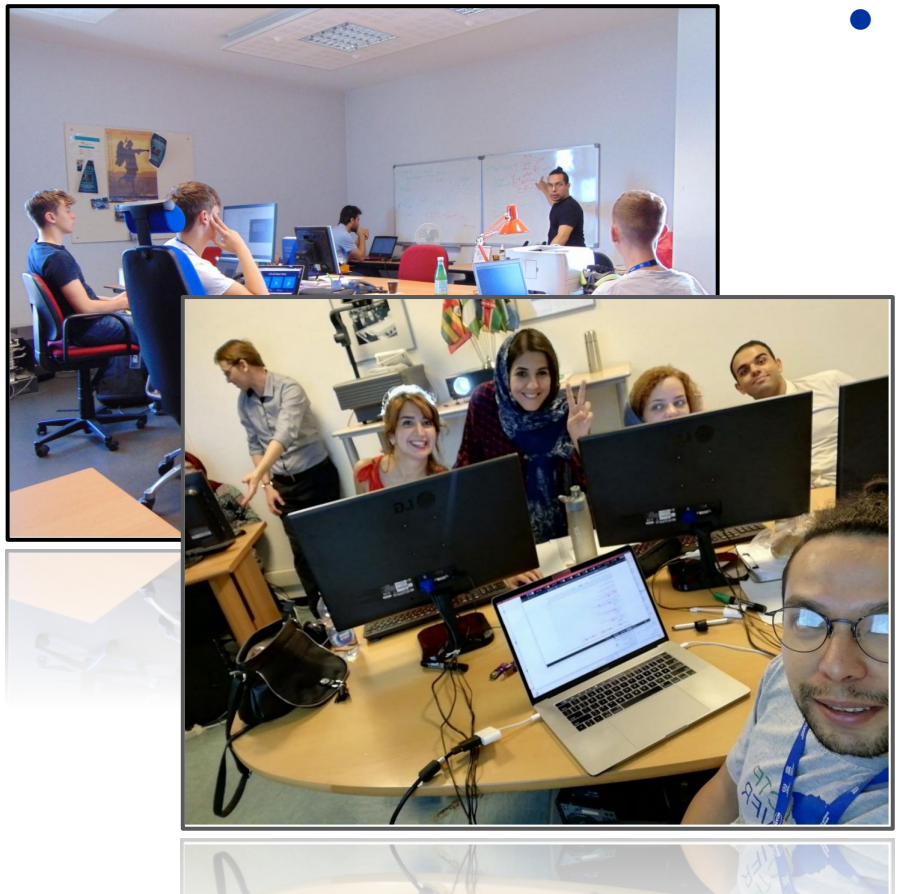
ATLAS Data / Software in Numbers





- ATLAS collected ~4 PB of RAW data so far
  - Plot includes data replication 😉
- The reconstructed / pre-processed analysis data takes up >100 PB of space
- The ATLAS reconstruction / simulation / common analysis software is ~4 million lines of C++
  - The installed size of the software, with all of its “dependencies”, is O(10) GB
- Neither of this is trivial to communicate to the public…

ATLAS Data in Education / Outreach

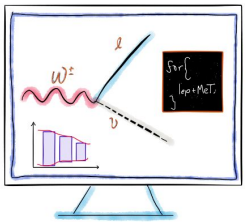


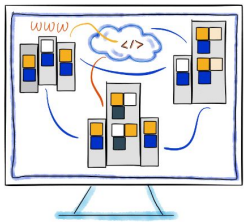
- Making (some of) ATLAS's data public has multiple benefits
  - Exposing students to this data, and the method of data analysis used in the experiment, helps with raising interest in our field
  - It provides resources to teach transferable skills in programming and analysis techniques
  - The data and software can help teachers when talking about modern high-energy physics
  - In the long term it should improve the scientific literacy of the public
  - Our governments and scientific funders are extremely pleased with the effort made by our experiment to make data and educational resources available to the public

What is ATLAS OpenData?

- Consists of 2 “campaigns”
  - 1 fb^{-1} of 8 TeV data
  - 10 fb^{-1} of 13 TeV data
  - Both with accompanying Monte Carlo datasets
- Data and MC pre-selected with different filters for different use-cases
  - To allow using a much smaller dataset in certain situations, for simpler analyses / demonstrations

The 8 TeV samples | **The 13 TeV samples**





Learn more about the 2016 datasets | Explore the 10x more data in 2020 datasets

Description	Name	link to ZIP file
events selected with at least one lepton (electron or muon) and exactly one large-Radius jet (R = 1.0)	1largeRjet1lep	5.5 Gb
events selected with exactly one lepton (electron or muon). This is a very large collection, so, it was divided into three ZIP files	1lep	17 Gb, 20 Gb, 21 Gb
events selected with exactly one lepton (electron or muon) and exactly one hadronic-reconstructed tau	1lep1tau	1.3 Gb
events selected with at least two leptons (electron or muon)	2lep	24 Gb
events selected with exactly three leptons (electron or muon)	3lep	1.0 Gb
events selected with at least four leptons (electron or muon)	4lep	427 Mb
events selected with at least two photons	GamGam	1.5 Gb

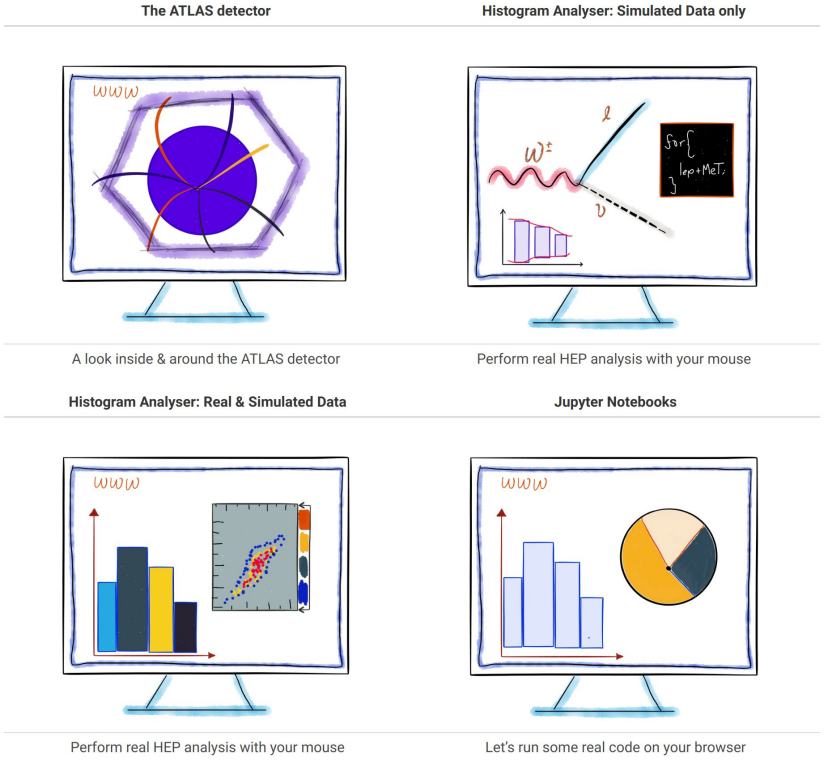
What is ATLAS OpenData?

Tuple branch name	C++ type	Variable description
runNumber	int	number uniquely identifying ATLAS data-taking run
eventNumber	int	event number and run number combined uniquely identifies event
channelNumber	int	number uniquely identifying ATLAS simulated dataset
mcWeight	float	weight of a simulated event
XSection	float	total cross-section, including filter efficiency and higher-order correction factor
SumWeights	float	generated sum of weights for MC process
scaleFactor_PILEUP	float	scale-factor for pileup reweighting
scaleFactor_ELE	float	scale-factor for electron efficiency
scaleFactor_MUON	float	scale-factor for muon efficiency
scaleFactor_PHOTON	float	scale-factor for photon efficiency
scaleFactor_TAU	float	scale-factor for tau efficiency
scaleFactor_BTAG	float	scale-factor for b-tagging algorithm @70% efficiency
scaleFactor_LepTRIGGER	float	scale-factor for lepton triggers
scaleFactor_PhotonTRIGGER	float	scale-factor for photon triggers
trigE	bool	boolean whether event passes a single-electron trigger
trigM	bool	boolean whether event passes a single-muon trigger
trigP	bool	boolean whether event passes a diphoton trigger
lep_n	int	number of pre-selected leptons
lep_truthMatched	vector<bool>	boolean indicating whether the lepton is matched to a simulated lepton
lep_trigMatched	vector<bool>	boolean indicating whether the lepton is the one triggering the event
lep_pt	vector<float>	transverse momentum of the lepton
lep_eta	vector<float>	pseudo-rapidity, η, of the lepton
lep_phi	vector<float>	azimuthal angle, ϕ, of the lepton
lep_E	vector<float>	energy of the lepton
lep_z0	vector<float>	z-coordinate of the track associated to the lepton wrt. primary vertex
lep_charge	vector<int>	charge of the lepton
lep_type	vector<int>	number signifying the lepton type (e or μ)
lep_isTightID	vector<bool>	boolean indicating whether lepton satisfies tight ID reconstruction criteria
lep_ptcone30	vector<float>	scalar sum of track p_{T} in a cone of R=0.3 around lepton, used for tracking isolation
lep_etcone20	vector<float>	scalar sum of track E_{T} in a cone of R=0.2 around lepton, used for calorimeter isolation
lep_trackd0pvunbiased	vector<float>	d_0 of track associated to lepton at point of closest approach (p.c.a.)
lep_tracksigd0pvunbiased	vector<float>	d_0 significance of the track associated to lepton at the p.c.a.
met_et	float	transverse energy of the missing momentum vector
met_phi	float	azimuthal angle of the missing momentum vector
jet_n	int	number of pre-selected jets
jet_pt	vector<float>	transverse momentum of the jet
jet_eta	vector<float>	pseudo-rapidity, η, of the jet
jet_phi	vector<float>	azimuthal angle, ϕ, of the jet
jet_E	vector<float>	energy of the jet
jet_jvt	vector<float>	jet vertex tagger discriminant [21] of the jet
jet_trueflav	vector<int>	flavour of the simulated jet
jet_truthMatched	vector<bool>	boolean indicating whether the jet is matched to a simulated jet
jet_MV2c10	vector<float>	output from the multivariate b-tagging algorithm [22] of the jet

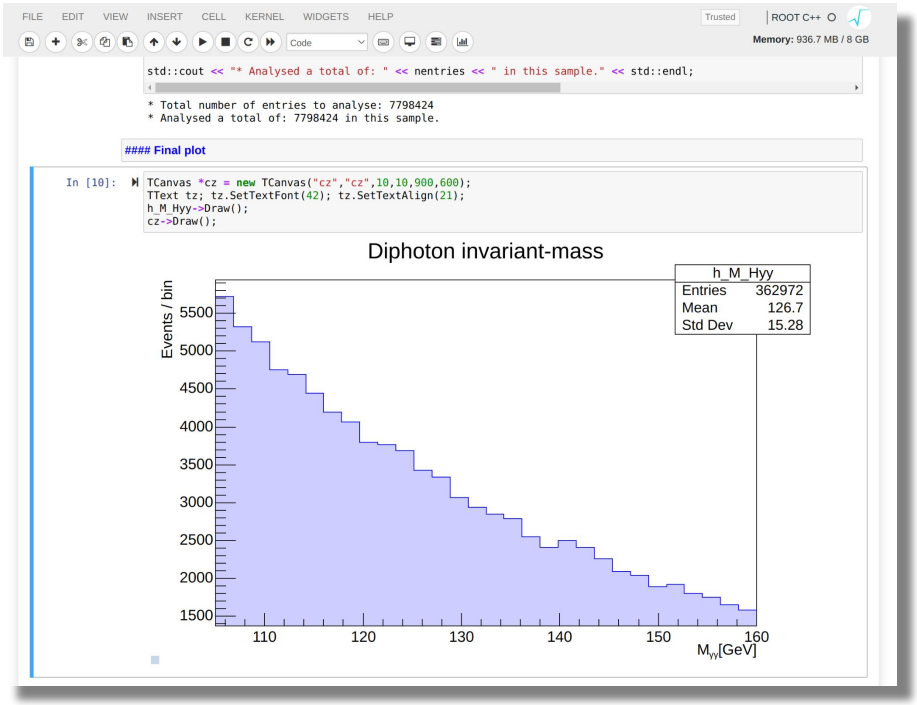
- Stored in a “flat ROOT ntuple format”, calibrated and simplified information about the reconstructed high level objects
  - With simplified information for taking systematic uncertainties into account on the properties of the objects
- Generated with the help of one of ATLAS’s “analysis frameworks”
  - Since ATLAS physicists are expected to perform some calibrations themselves
  - Different frameworks actually used for the 8 TeV and 13 TeV datasets

ATLAS OpenData Software

- Analysis code is provided with the data to help with its usage
  - Both in online Jupyter Notebooks;
  - and in downloadable VMs preloaded with all software needed for the analysis
- Different methods provided to support online/offline clients with different level of “analysis complexity”
- Also working on providing the examples in the form of Docker containers



ATLAS OpenData Software



- With a solid internet connection using Jupyter notebooks, without downloading anything, can be a good option
- Local VMs with locally downloaded data can be used in a completely offline environment as well
- Docker containers can combine the two, presenting a notebook interface on top of a locally running web service

Experience in Outreach

- Many ATLAS collaborators use the Open Data in their BSc and MSc programmes, in courses such as particle physics, programming courses, and advanced labs.
- We have excellent feedback from many institutes!
- It has also been used by trainers in countries that are not part of the ATLAS Collaboration

Training in Venezuela, Argentina, Uruguay, Honduras, Peru and many others in Latin America

Online course in particle physics and machine learning at the Royal University of Bhutan

Co-Creation

- We worked with many secondary school students, interns, CERN summer students, and BSc students to co-create many of the resources.
- Our methods included content creation, testing, peer review between students, and feedback.
- This helped us to check the level was appropriate for different age groups, and we validated the tools and software, and created new analyses frameworks.
- We did this in various geographic locations where high speed internet is not common-place, or powerful computers, to understand the limitations and make our data and resources accessible to all.
- Great way for students to learn about particle physics, analysis techniques, and genuinely contribute to high level efforts.

Future OpenData

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)[1], 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[2]), 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[3], 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,[4] 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

CERN-OPEN-2020-013
20/11/2020

- ATLAS is in the process of changing its analysis data formats to some level
  - See Caterina's presentation on Thursday for some of the details
- The DAOD_PHYSLITE format, meant for HL-LHC ATLAS analyses, is being investigated as the starting point for the next ATLAS OpenData release
  - It will make the preparation of the public data a lot easier this time around
  - Which will help us to implement the policy agreed on in 2020
- Data formats other than ROOT ntuples are being considered
  - But it is unlikely that a different format would be used in the end

Summary

- Both open releases of ATLAS's data proved very successful in outreach
  - Many schools used it over the years to engage high-school and university students
    - Was even used in some BSc / MSc theses!
  - Was a great help in involving students from all over the world
- The software associated with the datasets continues to be improved
  - With the involvement of the students themselves in the process
  - Embracing new technologies with the software (Jupyter notebooks, Docker containers), which are not even used by all ATLAS analyzers at this point
- A future data release is in planning
  - The infrastructure developed for it will hopefully be usable for data releases afterwards as well

http://home.cern