PROCEEDINGS OF SCIENCE

The ATLAS Open Data Project

Caterina Doglioni, on behalf of the ATLAS Collaboration[*][†]
Lund University
E-mail: caterina.doglioni@hep.lu.se

The ATLAS Open Data project is a collection of high-level data from the Large Hadron Collider and tools to analyze this data. The ATLAS Open Data can be used for education, training, outreach and citizen science. These resources are meant to be platform-independent, and are available on the web and on storage devices such as USB drives.

Sixth Annual Conference on Large Hadron Collider Physics (LHCP2018)
4-9 June 2018
Bologna, Italy

[*] Speaker.

[†] With input from Else Lytken and Arturo Sanchez Pineda.

The ATLAS Open Data project [1, 2, 3] comprises software tools and a subset of proton–proton collision data from the Large Hadron Collider (LHC) collected by the ATLAS detector and reconstructed into final physics objects such as leptons, photons and hadronic jets. The ATLAS Open Data Project allows newcomers to the experiments, young students, and the general public to understand how a high energy physics analysis is conducted, by analyzing the data using both interactive analysis tools and software that they can modify on their own local machines.

In this contribution, we describe the ATLAS datasets that are made available, the analyses provided as a starting point in the software suite, and conclude with a description of use cases of the ATLAS Open Data project worldwide.

1. ATLAS Open datasets

The ATLAS data available within the Open Data project correspond to approximately 1/fb of LHC proton–proton data recorded in 2012, called the ATLAS Open Data 2016 dataset. It corresponds to approximately 100 trillion proton–proton collisions. This is part of the data that allowed ATLAS to discover the Higgs boson [4]. For this reason, this fraction of the 2012 data has an important scientific, educational and historic value. Simulated data are also made available for various Standard Model and new physics signal models, for comparison with LHC data.

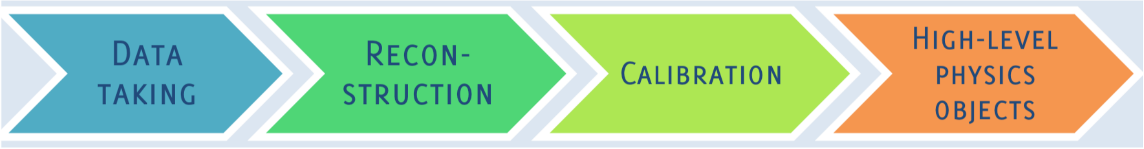

Figure 1: Schema of the reconstruction chain that leads from data taking to physics objects for a particle physics experiment such as ATLAS.

One of the goals of the ATLAS Open Data project is to allow users to easily gain insight into the process of doing a physics analysis. For this reason, data are provided in a simplified format containing high-level physics objects (e.g. electrons, muons, jets...), reconstructed and calibrated with algorithms employed by the ATLAS detector at the time of the data release. A schematic representation of this process is shown in Fig. 1. The data events are encoded in a ROOT ntuple [5], containing a collection of each of the physics objects. A loose preselection on the quality of the events and of the objects is applied to decrease the processing time on events that would not pass the analysis selection. Wherever necessary, simplicity is favored over precision. This means for example that not all the information available to an actual ATLAS physics analysis is available in the ATLAS Open Data format, in order to maintain a simpler layout and documentation.

2. Analyses provided in the Open Data suite

A number of simplified analyses are provided alongside the ATLAS Open Data dataset:

- High-statistics Standard Model analyses, allowing users to measure properties of the Standard Model particles (e.g. the mass of the Z boson), or to confirm agreement between data and simulation. These analyses select the processes $W \to l\nu$, $Z \to ll$, $tt \to l\nu jjbb$.

- Low-statistics Standard Model analyses, showcasing the difficulties encountered by searches for known, rare processes over larger backgrounds, such as the Higgs boson or diboson production. These analyses select WZ, ZZ and Higgs events. An event display from the $H \to ZZ* \to 4\mu$ analysis is shown in the left panel of Figure 2
- A search for a new physics process, allowing users to search for a hypothetical Z' signal beyond the Standard Model, decaying into top-antitop pairs.

These analyses produce comparison plots of data and simulation, as shown in the right panel of Figure 2 for the $H \to WW$ analysis.

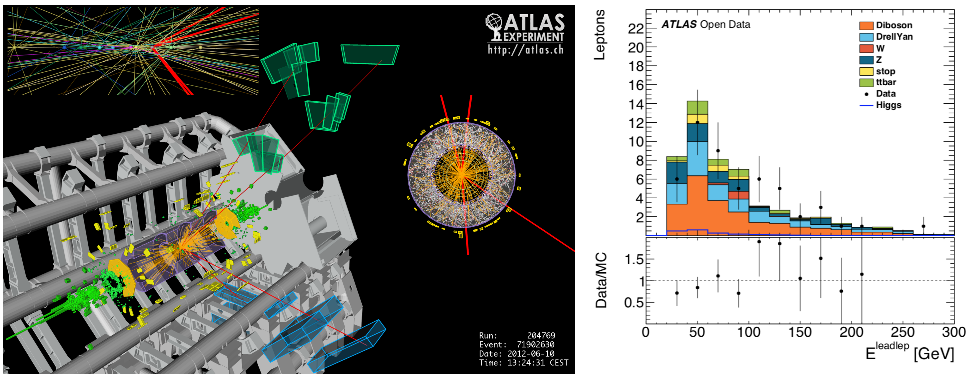

Figure 2: Left: Event display from the $H \to ZZ* \to 4\mu$ analysis [6]. Right: Example plot produced by the $H \to WW$ analysis, showing the energy of the leading lepton in the event [3], where MC stands for Monte Carlo simulation and "stop" stands for single top.

3. Software tools

The ATLAS Open Data dataset is complemented by analysis tools that can be run on either stand-alone virtual machines (VM) or ROOTBooks [7].

The virtual machines are prepared using the CERN VM software application [8]. They contain both data and analysis tools that can be installed either from the web or from a USB drive. ROOTBooks use the Jupyter notebook technology [9] using the ROOT software framework [5], in order to execute analyses on the Cloud or in the provided VMs on the Open Data website with histogram visualization tools.

The documentation for the software tools is provided via GitBooks[1]. The documentation comprises a general introduction and specific instructions for each one of the analyses. Video guides are also available in a YouTube playlist [10].

4. Users of the ATLAS Open Data

Universities from all over the world use the ATLAS Open Data for teaching students at the basic and advanced level the essentials of data analysis in particle physics. Among the universities and institutes using open data are Maastricht (Belgium), Montreal (Canada), UIS (Colombia),

[1] https://www.gitbook.com/about

Athens (Greece), TU Dresden (Germany), KTH and Lund (Sweden, as described in more detail below), Oslo (Norway), LIP (Portugal), CERN, Birmingham (UK), U of Michigan and California State (US), UCV and USB (Venezuela). The ATLAS Open Data is also used for the IPPOG Masterclasses [11], day-long events in which high school students interact with researchers for a day, do hands-on data analysis, and share their results with other institutes in a CERN-hosted videoconference. ATLAS Open Data is also used for e-courses and student theses by CEVALE2VE (Centro Virtual de Altos Estudios de Altas Energias) [12], which builds collaborative networks with and between Latin American institutions.

5. Use of the ATLAS Open Data at Lund University

At Lund University (Sweden), we use the ATLAS Open Data for hands-on data analysis in the *Modern Experimental Particle Physics* course. Modern experimental particle physics at Lund University is a course for students that have already a basic knowledge of particle physics, usually taken by students who are currently or will soon be doing their Master's or Bachelor's project at the Particle Physics Division. Most of the students are already familiar with basic programming (e.g. Python).The aim of the course is to bring the students up-to-date with contemporary particle physics and the status of the Standard Model from the experimentalist's point of view. It also includes a basic introduction to the statistical methods used. An important part of the course is the reconstruction and identification of particles as well as understanding the analysis strategies, and the ATLAS Open Data plays an important role in introducing the students to all those concepts.

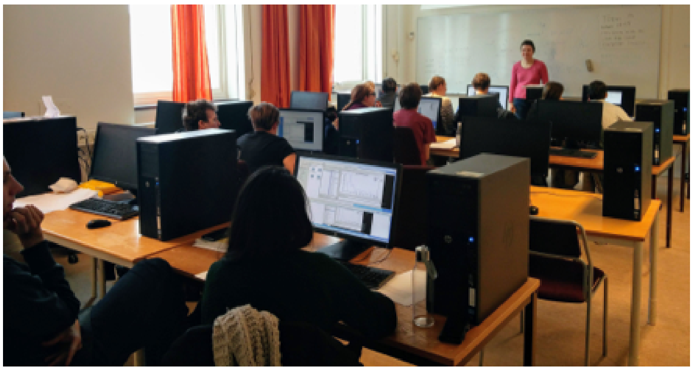

Figure 3: Picture from a laboratory session from the Modern Experimental Particle Physics course at Lund University.

During the ATLAS Open Data exercise, the students are divided in groups of two or three. Each group chooses one of the Open Data analyses and reads additional material to prepare a preliminary presentation of about 15 minutes. The presentation is rehearsed in presence of ATLAS physicists, including a question and answer session. This prepares the students to present scientific material in public, to review published references, to make plans for the analysis that they will undertake during the practical exercise, and to answer questions testing their understanding of the material. During the second lecture, each group goes through the introductory GitBook and receives a guide-sheet to guide them towards specific tasks of each analysis, e.g. superimpose signal and background for the Beyond the Standard Model analysis. The results are discussed among the students and the teachers. A picture of the laboratory session is shown in Fig. 3. This practical exercise received positive reviews in the students evaluations.

6. Conclusions

This contribution presented the ATLAS Open Data tools, a combination of public software and data released by the ATLAS Collaboration. The ATLAS Open Data tools are used to exemplify how a LHC physics analysis takes place, for newcomers to the experiment, students and visitors of the Open Data website. Examples of the educational use of the ATLAS Open Data are presented, with an emphasis on the recent experience of its use in the Master's degree level Modern Experimental Particle Physics course at Lund University in Sweden.

References

[1] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, JINST **3** (2008) S08003. doi:10.1088/1748-0221/3/08/S08003

[2] ATLAS Collaboration, ATLAS Open Data Portal website http://opendata.atlas.cern

[3] ATLAS Collaboration, Review Studies for the ATLAS Open Data Dataset, Technical Report ATL-OREACH-PUB-2016-001, CERN, Geneva, Aug 2016.

[4] ATLAS Collaboration, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," Phys. Lett. B **716**, 1 (2012) doi:10.1016/j.physletb.2012.08.020 [arXiv:1207.7214 [hep-ex]].

[5] Rene Brun and Fons Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. http://root.cern.ch/

[6] ATLAS Collaboration, Observation of an excess of events in the search for the Standard Model Higgs boson in the $H \to ZZ^{(*)} \to 4\ell$ channel with the ATLAS detector. Technical Report ATLAS-CONF-2012-092, CERN, Geneva, Jul 2012.

[7] Arturo Sanchez Pineda, on behalf of the ATLAS Collaboration, Integration of ROOT notebook as an ATLAS analysis web-based tool in outreach and public data release projects. Technical Report ATL-SOFT-PROC-2016-011, CERN, Geneva, Nov 2016.

[8] P. Buncic et al., "CernVM: A virtual software appliance for LHC applications," J. Phys. Conf. Ser. **219**, 042003 (2010). doi:10.1088/1742-6596/219/4/042003

[9] Thomas Kluyver et al, Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

[10] A. Sanchez Pineda, Youtube playlist of Open Data videos. link

[11] K. Cecire, "Developments in International Masterclasses", Proceedings of the Meeting of the APS Division of Particles and Fields (DPF 2017), arXiv:1710.00927 [physics.ed-ph].

[12] Arturo Sanchez Pineda, on behalf of the ATLAS Collaboration The CEVALE2VE case. Technical Report ATL-OREACH-PROC-2017-001, CERN, Geneva, Jan 2017.