PROCEEDINGS OF SCIENCE

Open data as seen by the 4 LHC experiments

Lara Lloret Iglesias[1]
Instituto de Fisica de Cantabria
Av. de los Castros s/n 39005 Santander, Spain
E-mail: lara@cern.ch

Open Data and, in general, reproducible research, has become one of the main goals of scientific communities. An adequate policy on Open Data together with the right sharing tools enables academic and scientific reuse of the data, boosting future discoveries, improving the learning curve of the future researchers and helping science to move forward faster. These proceedings summarize the Open Data strategy from the point of view of the four main experiments at the LHC: ATLAS, CMS, LHCb and ALICE.

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

[1]Speaker

PoS(LHCP2018)300

1. Introduction

The four main experiments at the LHC operate independently of each other without sharing their data nor the analysis tools to exploit them. This is very important in order to avoid biases that could alter the scientific conclusions extracted from the data, especially when searching for new physics. Once the data have been extensively exploited, and all the relevant scientific results published by the corresponding experiment, an adequate policy to openly release the data can be a great improvement for the scientific communities. Open Data can boost new discoveries, accelerating the learning curve of the future researchers and thus helping science to move forward faster. These proceedings summarize the Open Data strategy from the point of view of the four main experiments at the LHC: ATLAS, CMS, LHCb and ALICE.

2. Why releasing data?

The main motivation for releasing Open Data from the CERN experiments is the consideration that science should be as open, collaborative, inclusive and transparent as possible. Open Data allows research groups to explore and reuse the data from other communities in order to verify their scientific claims and further exploit the data aiming to extract new knowledge from them.

When the Open Data is realeased with the adequate tools, it can allow the general public to access and *play* with data, helping to engage them with scientific research. This is extremely important both to promote science and to give something back to society since they are the ones funding the public research.

In terms of education, organising events for students with meaningful activities using Open Data, such as the Master Classes [1], is a very effective way of teaching them about particle physics and data analysis, providing the motivation to become scientists in the future.

For all these reasons, the Open Data release at CERN targets a broad public going from university and high school students to the general public and scientific research communities.

3. Cern Open Data Portal

With the growing demand for open research data, CERN has released an online Open Data Portal [2]. This portal is an access point to a growing range of data produced in the main particle detectors at CERN, including accompanying software and documentation needed to understand and analyze the data being shared. The published data adhere to

the established global standards in data preservation and Open Science, being shared under open licenses and issued with a digital object identifier (DOI) to make them citable.

This portal provides content for both education and research, aiming to support high school students and teachers as well as university students and professors. This platform also caters to the broader research community that may be interested in these materials. It includes detailed guidelines to introduce the user to physics analyses that make use of the datasets available. Event displays can also be found in the portal. They allow to visualise collision data and examine what happens in a particle interaction in a graphic way.

The success of CERN Open Data has been made possible through working closely with various CERN-affiliated research collaborations and with external partners. It provides access to data and resources from the four big LHC collaborations and also from Opera.

4.Data access policy

All four experiments at the LHC have released a data policy following the different levels of open access to data described in the DPHEP [3] model. These levels correspond to various preservation models, listed in order of increasing complexity:

Level 1 - Provide additional documentation. Publication-related information search.

Level 2 - Preserve the data in a simplified format. Outreach, simple training analyses

Level 3 - Preserve the analysis level software and data format. Full scientific analysis based on existing reconstruction.

Level 4 - Preserve the reconstruction and simulation software as well as the basic level data. Full potential of the experimental data

The four experiments have different approaches regarding how many data they are willing to release and when. LHCb [3] will grant access to portions of the data five years after data is taken. The portion of the data which LHCb would normally make available is 50% after 5 years, rising to 100% after 10 years. All requests will be considered by the Collaboration Board and the period and proportion may be varied for specific requests. ALICE [4] data will be conditionally made publicly available after an embargo period of 5 years after publication for 10% of the data and 10 years for 100% of the data. In the last version of the CMS policy data [5], updated in April 2018, the experiment commits to make available 50% after 3 years from data taking, rising to 100% within 10 years, but the Collaboration Board can, in exceptional circumstances, decide to release some particular data sets either earlier or later. ATLAS [6] has not set any schedule.

Despite some differences, all of the experiments share some common goals. They all provide access and analyses at different levels of knowledge and expertise to cover the whole range of interested users. Also, where it is possible, the experiments have released simplified datasets and made example analysis code available. Another useful tool that is common to all the experiments is allowing for analysis and visualization via the browser (e.g. histograms, event displays…). In addition, all of the four experiments provide virtual machines with software environment for more advanced users.

CMS is currently the only experiment that has actually released research level data.

5.Research level data

5.1Training

Using research data for training purposes is an extremely useful tool to allow undergraduate students to obtain and understand physics results supervised by experienced physicists One of the last research level examples consists on a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis published in [7].

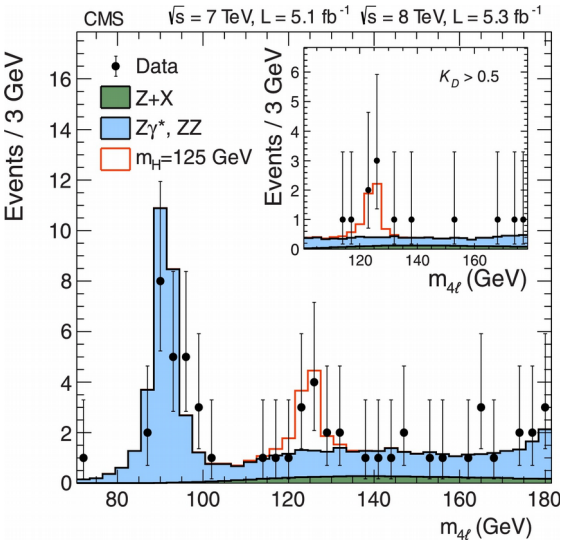

Figure 1: Distribution of the four-lepton invariant mass for the ZZ → 4 analysis. The points represent the data, the filled histograms represent the background, and the open histogram shows the signal expectation for a Higgs boson of mass m_H = 125 GeV, added to the background expectation

The published reference plot which is being approximated in this example can be found in Figure 1. Other Higgs final states (e.g. Higgs to two photons), which were also part of

the same CMS paper and strongly contributed to the Higgs boson discovery, were not covered by this example.

The example consists of different levels of complexity. The highest level of this example addresses users who feel they have at least some minimal understanding of the content of this paper and of the meaning of this reference plot, which can be reached via separate educational exercises. The lower levels might also be interesting for educational applications. The example requires a minimal acquaintance with the linux operating system and the ROOT analysis tool.

The resulting plot obtained with Open Data within this example can be seen in Figure 2 and compared with the original one in Figure 1.

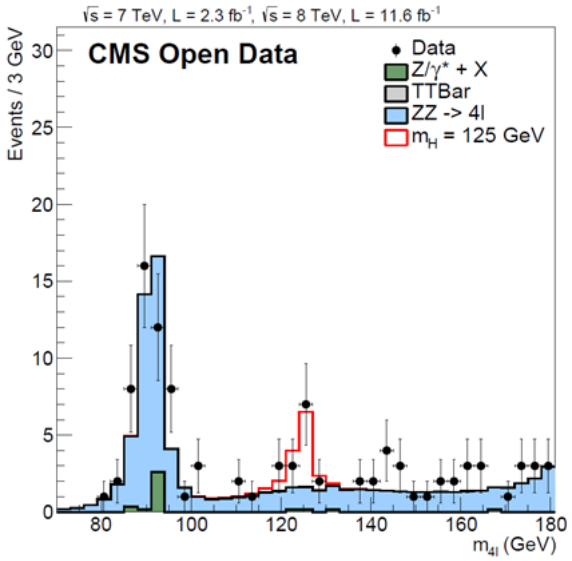

Figure 2: Approximation using Open Data of one of the main results in the Higgs to four leptons analysis

5.2 Scientific results

There are currently two new scientific results using CMS Open Data

- **Jet substructure studies with CMS open data** [8]: Open Data from the CMS experiment has been used to study the two-prong substructure of jets. A good agreement has been found between results obtained from the CMS Open Data and those obtained from parton shower generators. Although the 2010 CMS Open Data do not include simulated data to help estimate systematic

uncertainties, track-only observables has been used to validate these substructure studies.

- **Exposing the QCD Splitting Function with CMS Open Data** [9]: The splitting function is a universal property of quantum chromodynamics (QCD) which describes how energy is shared between partons. It cannot be measured directly since it always appears multiplied by a collinear singularity factor. Recently, however, a new jet substructure observable was introduced which approaches to the splitting function for sufficiently high jet energies. This provides a way to expose the splitting function through jet substructure measurements at the LHC. In this publication, public data released by the CMS experiment are used to study the two-prong substructure of jets and test the $1 \rightarrow 2$ splitting function of QCD.

6. Other activities

6.1 International Master Classes

LHCb, ALICE, ATLAS and CMS participate actively in the international Master Classes. Each year more than 13000 high school students in 52 countries come to one of about 215 nearby universities or research centres for one day in order to unravel the mysteries of particle physics. Lectures from active scientists give insight in topics and methods of basic research at the fundaments of matter and forces, enabling the students to perform measurements on real data from particle physics experiments themselves. At the end of each day, like in an international research collaboration, the participants join in a video conference for discussion and combination of their results.

6.2 Kaggle Challenges

Several datasets have been released for machine learning challenges online in the Kaggle Platfom:

- **Higgs boson Machine Learning Challenge [10]:** The goal of the Higgs Boson Machine Learning Challenge is to explore the potential of advanced machine learning methods to improve the discovery significance of the experiment using simulated data with features characterizing events detected by ATLAS. The task is to classify events into *tau tau decay of a Higgs boson* versus *background.*
- **Flavour of Physics: Finding $\tau \rightarrow \mu\mu\mu$ [11]**: This challenge uses real data from the LHCb experiment, mixed with simulated datasets of the rare decay. The metric used in this challenge includes checks that physicists perform in their analysis to make sure the results are unbiased. These checks have been built into

the competition design to help ensure that the results will be useful for physicists in future studies

- **TrackML: Particle Tracking Challenge [12]**: The challenge in this competition is to build an algorithm that quickly reconstructs particle tracks from 3D points left in the silicon detectors.

7. Challenges and conclusions

High Energy Physics is doing well with immediate metadata such as beam conditions, event and run numbers, provenance information (processing and reconstruction chain, software versions) recorded together with data at time of dataset creation. Context metadata (i.e how to pick up the right objects in the data and their documentation, how to know if there are additional selections, corrections…) are to be improved. This information is readily available and even obvious at time of immediate data analysis, but then easily forgotten. The Open Data group will work on progressing towards more extensive metadata.

In summary, all four experiments are doing as much as they can to make the Open Data goals advance, always limited by experiment resources. Open Data enable High Energy physicists to engage with the outside world on a more meaningful level allowing us to show how we work. Many tools and datasets have already been released. This material can be used to develop lab courses, full visualisations, create teaching materials and also research results. Master Classes and Kaggle challenges using Open Data are being succesfully organised yearly using data from the four experiments.

References

[1] http://physicsmasterclasses.org/

[2] Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics, arXiv:1205.4667 [hep-ex]

[3] P. Clarke. LHCb collaboration (2013). LHCb External Data Access Policy. CERN Open Data Portal. DOI:10.7483/OPENDATA.LHCb.HKJW.TWSZata.cern.ch/record/410

[4] ALICE collaboration (2013). ALICE data preservation strategy. CERN Open Data Portal. DOI:10.7483/OPENDATA.ALICE.54NE.X2EA

[5] https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/RetrieveFile?docid=6032&filename=CMSDataPolicyV1.2.pdf&version=2

[6] ATLAS collaboration (2014). ATLAS Data Access Policy. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.T9YR.Y7MZ

[7] CMS collaboration (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC Phys.Lett. B716 (2012) 30-61, arXiv:1207.7235.

[8] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, and J. Thaler. (2017) Jet substructure studies with CMS open data. Phys. Rev. D 96, 074003

[9] A. Larkoski, S. Marzani, J. Thaler, A. Tripathee, and W. Xue . (2017) Exposing the QCD Splitting Function with CMS Open Data. Phys. Rev. Lett. 119, 132003

[10] https://www.kaggle.com/c/higgs-boson

[11] https://www.kaggle.com/c/flavours-of-physics

[12] https://www.kaggle.com/c/trackml-particle-identification