Using CMS Open Data for education, outreach and software development

Stefan Wunsch[1,2,*]

[1]CERN
[2]Karlsruhe Institute of Technology

Abstract. The CMS collaboration at the CERN LHC has made more than one petabyte of open data available to the public, including large parts of the data which formed the basis for the discovery of the Higgs boson in 2012. Apart from their scientific value, these data can be used not only for education and outreach, but also for software development. However, in their original format, the data cannot be accessed easily without experiment-specific knowledge and skills. Work is presented that allows to set up open analyses that are performed close to the published ones, but which meet minimum requirements for experiment-specific knowledge and software. The suitability of this approach for education and outreach is demonstrated with analyses that have been made fully accessible to the public via the CERN Open Data portal. Further, the value of these data for software development and as basis for benchmarks of analysis software under realistic conditions of a high-energy physics experiment is discussed.

1 Introduction

The CMS collaboration has recently published on the CERN Open Data portal [1] a new batch of open data to the public [2]. The release increases the volume of the open data to more than two petabyte including large parts of the data used for the discovery of the Higgs boson in 2012. In the future, we can expect a continuous growth of these resources due to the CMS Open Data policy [3], which states that the collaboration commits to releasing 100 % of its analysable data within ten years of collecting them, making CMS Open Data an invaluable resource for open science. All other large LHC experiments have published similar statements [4–6]. The CMS Open Data releases are already today basis of scientific publications, naturally in the field of particle physics [7, 8] but as well in studies related to data science and machine learning [9, 10].

Besides the purely scientific use-cases, open data is also valuable for education and outreach being already actively used around the world [11, 12], especially for attracting young people to CMS and high-energy physics. This paper presents an approach to ease the usage of open data for education and outreach and points out the additional value of such resources for software development in high-energy physics.

Section 2 shows in which data-format most of the CMS Open Data is currently distributed and discusses which aspects are most important for facilitating its use. Section 3 presents

[*] e-mail: stefan.wunsch@cern.ch

examples for education and outreach using these accessible resources and section 4 puts emphasis on the impact and importance of such accessible data on software development.

2 Data-format of CMS Open Data

In the current release of the CMS Open Data, most of the collision data and simulated data are published in a data-format known as AOD (Analysis Object Data). This data-format consists of serialized C++ objects, which requires experiment-specific software (CMSSW [13]) and ROOT [14] to be read. Further, each event holds about 500 kB/event of information resulting in general in large files. AOD is a powerful but complex data-format, which does not fit the needs of the here discussed use-cases of the data.

The CMS collaboration has developed derived data-formats called MiniAOD [15] and its successor NanoAOD [16]. Each format reduces the information content so that the size per event is reduced by an order of magnitude each, resulting in about 2 kB/event for NanoAOD files. Whereas MiniAOD is still similar to AOD and stores serialized C++ objects, NanoAOD is based on the storage of basic types such as floats, integers and arrays thereof. Table 1 shows as example the layout of a muon collection in the NanoAOD data-format.

Table 1. Data-format of a muon collection in the NanoAOD format

Variable	Type	Description
nMuon	unsigned int	Number of muons in this event
Muon_pt	float[nMuon]	Transverse momentum of the muons
Muon_eta	float[nMuon]	Pseudorapidity of the muons
Muon_phi	float[nMuon]	Azimuth of the muons
Muon_mass	float[nMuon]	Mass of the muons
Muons_charge	int[nMuon]	Charge of the muons (either 1 or -1)

Therefore, NanoAOD is readable independent from experiment-specific software with any library capable to read ROOT files. The data-format of NanoAOD satisfies all requirements for the emphasized use-cases, which are:

- Simple data-format
- Readable without experiment-specific software
- Small files

Moreover, the data-format is in use by actual CMS analyses, which facilitates the usage for education and outreach by being close to actual research and provides a perfect playground for studying analysis workflows.

For these reasons, a conversion tool for the AOD data-format to a reduced NanoAOD format was developed in order to make the CMS Open Data release more accessible [17]. The conversion tool targets the conversion of AOD files taken in 2012 with a total recorded integrated luminosity of 21.8 fb^{-1} [18]. At the time of the development only half of the data taken in 2012 was published (Run B and C) so that the examples in sections 3 and 4 are based on an integrated luminosity of about 11.5 fb^{-1}. It should be noted that the content of the reduced CMS Open Data NanoAOD files is not validated and therefore must not be used for any physics measurements.

3 Usage for education and outreach

Physics analysis examples for education and outreach can be separated roughly in two categories:

- High-school level: Students or the general public who have the first contact with high-energy particle physics and little knowledge about programming and data analysis
- University level: Students or individuals studying advanced particle physics with existent knowledge about programming and data analysis

In the following, two examples are presented, each serving one of the categories above.

3.1 Analysis of the di-muon spectrum

This high-school level example consists in the analysis of the di-muon spectrum. The input file is in the reduced NanoAOD format as described above, generated from the validated runs of the double muon primary dataset and contains only the muon collection such as shown in table 1. The full information about the data is transparently documented in the record on the portal [19]. The conversion and filtering reduces the AOD files with a size of 19.4 TB to a dataset with only 2.2 GB while keeping about 61.5 million events.

The analysis selects events with exactly two muons, checks for opposite-charged pairs and eventually computes the invariant mass of the di-muon system and makes a histogram. The result can be seen in figure 1. A reference implementation is provided in C++ and Python using only ROOT with in total less than 100 lines of code. It has an approximate runtime of one minute on a consumer laptop with an SSD running on a single thread. It should be noted that the analysis is trivially parallelizable, which is also supported by the reference implementation powered by the `RDataFrame` facility from ROOT [20].

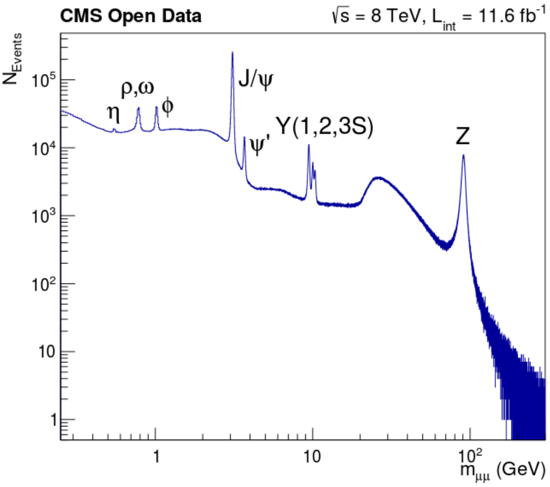

Figure 1. Analysis of the di-muon spectrum using data from the CMS detector taken in 2012 [19].

The analysis illustrates the concept of particle resonances and an invariant mass and finally enables the user to rediscover particle resonances in a wide energy range from the η meson at about 548 MeV up to the Z boson at about 91 GeV.

3.2 Analysis of Higgs boson decays to two tau leptons

This example uses data and simulation of events analyzing 11.5 fb^{-1} with the goal to study decays of a Higgs boson into two tau leptons in the final state of a muon lepton and a hadronically decayed tau lepton. The analysis follows loosely the setup of the official CMS analysis published in 2014 [21]. The purpose of the original CMS analysis was to establish the existence of the Higgs boson decaying into two tau leptons. Since performing this analysis properly with full consideration of all systematic uncertainties is an highly complex task, we reduce this analysis to the qualitative study of the kinematics and properties of such events without a statistical analysis. However, as you can find in the following, already such a reduced analysis is complex and requires to understand an extensive physics context, which makes this a perfect first look into the procedures required to claim the evidence or existence of new particles.

The analysis processes data, two signal samples representing a Higgs boson produced by gluon-fusion and vector-boson fusion, and the four most dominant background processes as simulated events. In addition, the QCD multijet contribution is estimated with a data-driven technique via an extrapolation from a control region with tau candidates of the same charge. Full details can be found in the record on the portal [22]. The analysis processes nine different datasets for a total of 69.3 GB consisting of 87 million recorded events and 114 million simulated events. The workflow follows a typical analysis at CMS:

1. The NanoAOD files are preprocessed and reduced to the event information needed for this analysis. In addition, a pair selection is performed to find from the muon and tau collections the pair which originates most likely from a Higgs boson. The result is a flat analysis dataset for further processing. The runtime on a consumer laptop and locally read files is around twenty minutes.

2. The flat analysis dataset is processed to compute histograms of 34 observables and all contributing processes. Further, similar histograms have to be produced in a control region for the data-driven QCD estimation, which sums up in total to multiple hundreds of histograms representing all processes and observables. The runtime for the processing is in the order of one minute.

3. The histograms are combined to physically meaningful plots. For example, figure 2 shows the visible mass of the di-tau system. The runtime of this step with less than a few seconds is negligible.

The analysis allows to study the contribution of the different physics processes to the data taken with the CMS detector. The complexity level is close to an actual CMS analysis with a rich physics content including complex objects such as jets and taus.

4 Usage for software development

Software development in high-energy physics is currently very active in studying analysis workflows of the future, see for example the efforts in the HEP Software Foundation [23]. However, since the developments are driven from inside the experiment collaborations, any analysis examples using new developed tools take typically experiment datasets as input. The datasets are usually subject to the collaboration policies, which prevents free sharing of the datasets outside of the collaboration. The lack of realistic and openly accessible analyses beyond over-simplified examples for education and outreach impedes the progress in this effort since freely shareable one-to-one comparison of different analysis tools are complicated

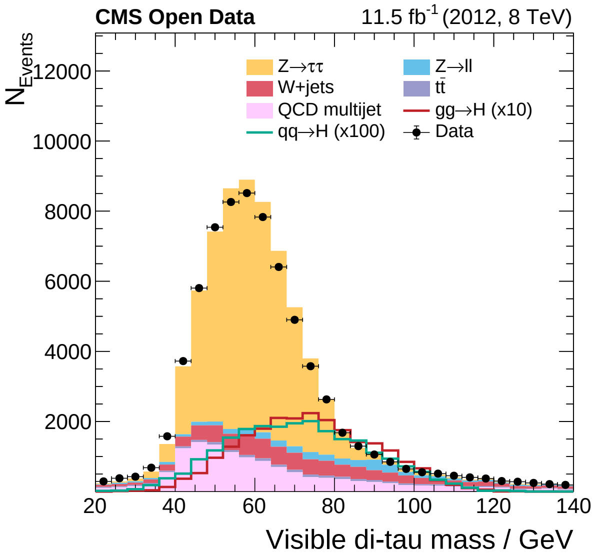

Figure 2. Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012 [22].

to put in place. Further, software in high-energy physics is usually open source. However, the lack of open analysis examples do not allow to publish comprehensive analysis examples with the software since the data is protected.

For these use-cases, the resources presented in section 3 are well suited for several reasons:

- High accessibility of the resources minimizes the initial effort to run the examples.
- The datasets are open and free from experiment-specific software.
- The NanoAOD format is a lightweight and easily readable data format.
- Software examples and analyses for education and outreach gain from a realistic setting but usually do not require the same precision as for a physics measurement.

The examples presented in section 3 have been used successfully for software tutorials, hands-on teaching and for software benchmarking of the ROOT framework [14, 24]. All educational material and benchmark studies are fully open and reproducible, which facilitates learning the usage of the software and promotes collaborative work for future developments.

5 Summary

The paper introduces new resources that allow to set up analyses with open data from the CMS experiment, which require a minimum of experiment-specific knowledge and software. Two example analyses are presented that are well suited for education and outreach on high-school and university level but also have the properties to facilitate software development in high-energy physics. The similarities between the requirements of education and outreach and software development are demonstrated. Because the datasets are free from experiment-specific software and any restrictions imposed by the collaborations, these examples are well suited to facilitate open discussions and inter-experimental collaborative work most notably in the field of software development and future analysis workflows.

References

[1] CERN, *The CERN Open Data portal*, http://opendata.cern.ch

[2] The CMS collaboration, *CMS releases open data for Machine Learning*, http://opendata.cern.ch/docs/cms-releases-open-data-for-machine-learning

[3] The CMS collaboration, *CMS preservation policy*, 10.7483/OPENDATA.CMS.7347.JDWH

[4] The ATLAS collaboration, *ATLAS preservation policy*, 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ

[5] The LHCb collaboration, *LHCb preservation policy*, 10.7483/OPENDATA.LHCb.HKJW.TWSZ

[6] The ALICE collaboration, *ALICE preservation policy*, 10.7483/OPENDATA.ALICE.54NE.X2EA

[7] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, J. Thaler, *Jet substructure studies with CMS open data* (2017), http://dx.doi.org/10.1103/PhysRevD.96.074003

[8] C. Cesarotti, Y. Soreq, M.J. Strassler, J. Thaler, W. Xue, *Searching in CMS open data for dimuon resonances with substantial transverse momentum* (2019), http://dx.doi.org/10.1103/PhysRevD.100.015021

[9] P. Musella, F. Pandolfi, *Fast and Accurate Simulation of Particle Detectors Using Generative Adversarial Networks* (2018), http://dx.doi.org/10.1007/s41781-018-0015-y

[10] A. Di Florio, F. Pantaleo, M. Pierini, *Sample with tracker hit information for tracking algorithm ML studies*, 10.7483/OPENDATA.CMS.N1IN.TQHD

[11] University of Central Florida, *Physics Hosts Teaching & Data Workshops*, https://sciences.ucf.edu/news/physics-hosts-teaching-data-workshops/

[12] Helsinki Insitute of Physics, *Education and Open Data*, http://www.hip.fi/?page_id=2985

[13] The CMS collaboration, *CMS Offline Software*, https://github.com/cms-sw/cmssw

[14] R. Brun, F. Rademakers, *ROOT - An object oriented data analysis framework* (1997)

[15] G. Petrucciani, A. Rizzi, C. Vuosalo, *Mini-AOD: A New Analysis Data Format for CMS* (2015), http://dx.doi.org/10.1088/1742-6596/664/7/072052

[16] A. Rizzi, G. Petrucciani, M. Peruzzi (CMS Collaboration), EPJ Web Conf. **214**, 06021. 6 p (2019)

[17] S. Wunsch, *Tool for conversion of CMS AOD files to reduced NanoAOD format for the purpose of education and outreach* (2019), 10.7483/OPENDATA.CMS.944Q.PN2X

[18] The CMS collaboration, *Public CMS Luminosity Information*, https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults

[19] S. Wunsch, *Analysis of the di-muon spectrum using data from the CMS detector taken in 2012* (2019), 10.7483/OPENDATA.CMS.AAR1.4NZQ

[20] E. Guiraud, A. Naumann, D. Piparo, *TDataFrame: functional chains for ROOT data analyses* (2017), https://doi.org/10.5281/zenodo.260230

[21] The CMS collaboration, *Evidence for the 125 GeV Higgs boson decaying to a pair of tau leptons* (2014), http://dx.doi.org/10.1007/JHEP05(2014)104

[22] S. Wunsch, *Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012* (2019), `10.7483/OPENDATA.CMS.GV20.PR5T`

[23] *The HEP Software Foundation*, `https://hepsoftwarefoundation.org/`

[24] V.E. Padulano, *Blurring High Energy Physics Data Analysis Techniques and Data Science Approaches* (2019), presented 25 Oct 2019, `http://cds.cern.ch/record/2693575`