The Ecuadorian experience with the CMS open data initiative and future projects

A. Chicaiza[1*], D. Merizalde[2], P. Llerena[1], J. Ochoa[2], X. Tintin[1], E. Carrera[2], E. Ayala[1]

1. Escuela Politécnica Nacional, Physics Department, Quito, Ecuador
2. Universidad San Francisco de Quito, Physics Department, Quito, Ecuador

*andres.chicaiza@cern.ch

Introduction

The CMS (compact muon solenoid) is a detector that stores information about the particles created after colliding proton beams. We cannot directly observe all the particles created after the collision because of the decay rate or the lack of interaction with the CMS. Nevertheless, those decays can lead to stable particles able to interact with the detector. Therefore, it is possible to use the signals obtained from the decay products to infer data about the original decay. The decay products are called physics objects.

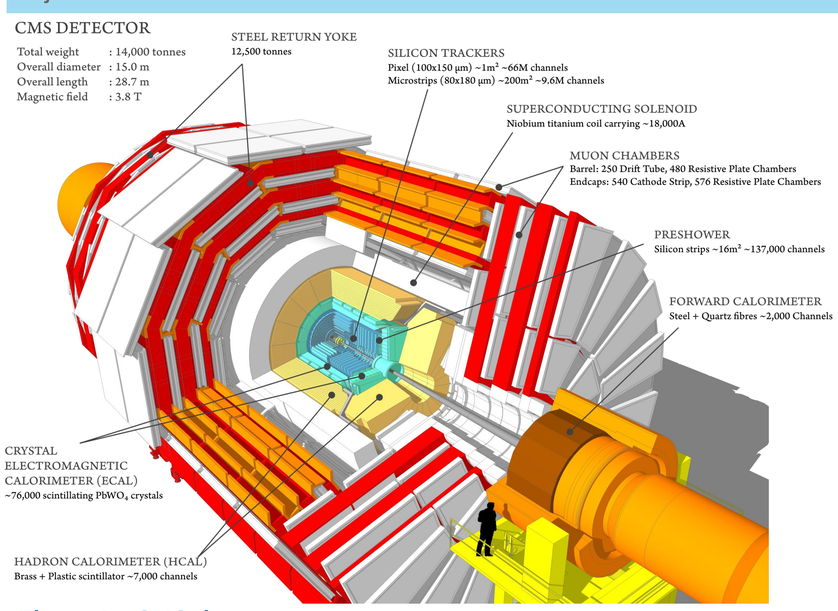


Figure A: CMS detector.

What is Open Data?

- ✓ Information from real collisions and simulations that can be found in the CERN Open Data Portal.
- ✓ It is expected that open data will support and inspire the global research community, including students and citizen scientists to start research beyond the actual knowledge.

- **How to find the data?**

The starting point is the Open Data Portal

Below the research bar and "Explore" is a link to datasets.

The filters are related to the research needs.

- On filter by type, select Data set.
- On filter by experiment, select CMS. The name of CMS datasets shows the conditions for classifying a specific event called triggers, the run period, and the data format.

For example, look at the following title:

/DoubleElectron/Run2012B-v1/RAW

The trigger is **DoubleElectron**, which is related to the minimum energy threshold two electrons must have to be considered as data. **Run2012B – v1** is the run period. **RAW** is the data format.

Which type of information can be obtained?

- ✓ **Collision:** refers to the real data that came off from the CMS detector.
- ✓ **Derived:** refers to datasets that have been further processed for some specific purpose, such as outreach and education.
- ✓ **Simulated:** refers to Monte Carlo datasets.

This information can be used to create Physics objects which could be:

- •Muons
- •Electrons
- •Jets
- •Photons
- •Taus
- •Missing transverse momentum

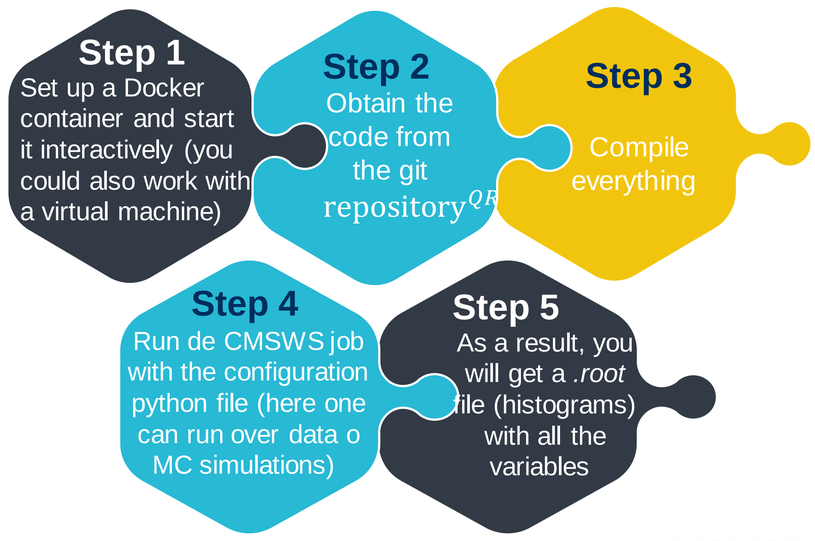
Then one can extract variables related such that momentum, energy, number of primary vertices and so on.

How to extract this information?

- The open data files are ROOT files following the EDM data model in the so-called miniAOD format. Configure CMSSW software to read and extract information from these files.
- The POET repository^{QR} contains packages that verse instructions and examples on how to extract physics object information from Run 2 (MiniAOD format) CMS open/legacy data and methods or tools needed for processing them.

QR

SCAN ME



CMS Open Data Workshop 2022

Since 2020, the CMS Collaboration releases the Open Data workshop that aims to bridge the technical gap that usually exists between the scientific creativity of an external analyst and the fundamental details of a full analysis with CMS open data. Ecuadorian students both from EPN and USFQ have actively contributed to the development of the workshops.

Primarily aimed at students and scientists with prior knowledge of collider physics and a deep interest in learning the works and arts of conducting experimental analysis using CMS Open Data.

Content:

- Pre-exercises: Git, Docker containers, ROOT, Intro to CMSSW.
- Physics Objects.
- Trigger and Luminosity.
- Simplified Run 2 analysis.
- Cloud computing.

Indico's site

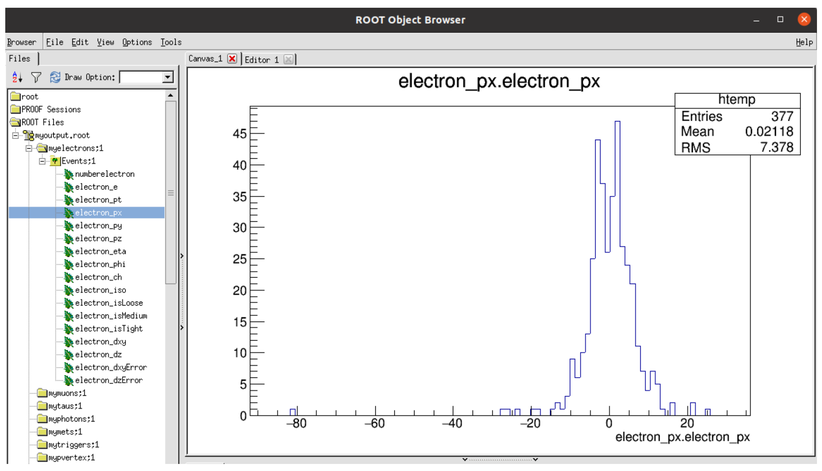


Figure B: Root Object Browser interface. Histogram of electron's x momentum. There also are other objects like muons, taus, etc.

Contact: cms-opendata-workshop-organizers@cern.ch.

- Analysing the dimuon samples

We were running over some larger ROOT files, and we took in account that memory issues may cause some errors or crashes of the code.

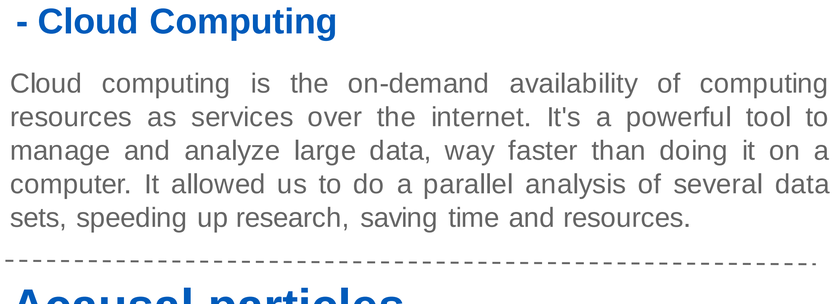


Figure C: Invariant mass of a selected sample of oppositely charged dimuon pairs, obtained from 2011 data.

- Cloud Computing

Cloud computing is the on-demand availability of computing resources as services over the internet. It's a powerful tool to manage and analyze large data, way faster than doing it on a computer. It allowed us to do a parallel analysis of several data sets, speeding up research, saving time and resources.

Acausal particles

The objective of this investigation was to take a theoretical approach that propose the existence of acausal particles.
This work was done by Jonathan Sánchez as part of the USFQ's graduates program.

Future Plans

Machine learning methods for data analysis. Particularly in the application of Convolutional Neural Networks for image for the analysis of collision.

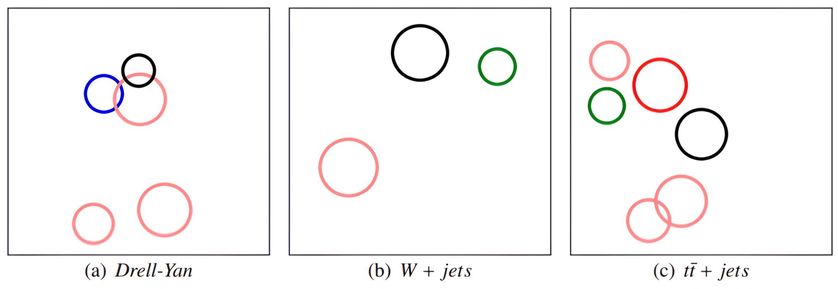


Figure D: Examples of images corresponding to the three different classes of collisions being classified. [1]

Conclusions

- Several students have successfully experimented with CMS open data. There exists enough information about how to access, filter and extract information.
- Ecuadorian contribution into the development of the 2022 workshop was a success.
- We have started to use this results formally in order to make original investigation.
- In the short term it is planned to improve AI algorithms focused on the analysis of particle physics.

Acknowledgements: CMS experiment and CERN Open Data Portal

References

[1] Application of a Convolutional Neural Network for image classification for the analysis of collisions in High Energy Physics. Celia Fernández Madrazo, Ignacio Heredia, Lara Lloret and Jesús Marco de Lucas. EPJ Web Conf., 214 (2019) 06017.
DOI: https://doi.org/10.1051/epjconf/201921406017