BETHEL UNIVERSITY

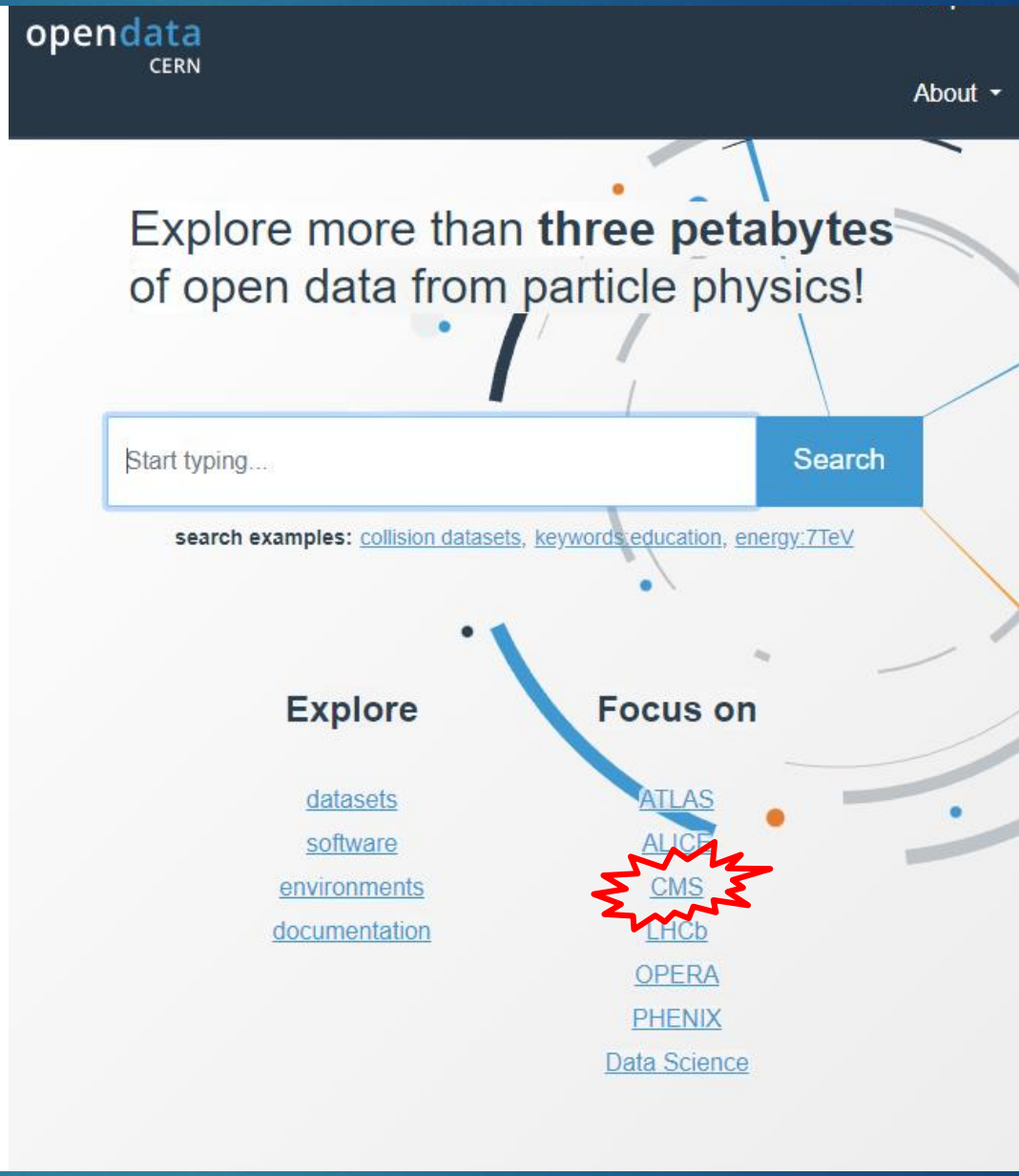
CMS Open Data

Julie Hogan, on behalf of the CMS Collaboration

2/8/2023

Outline

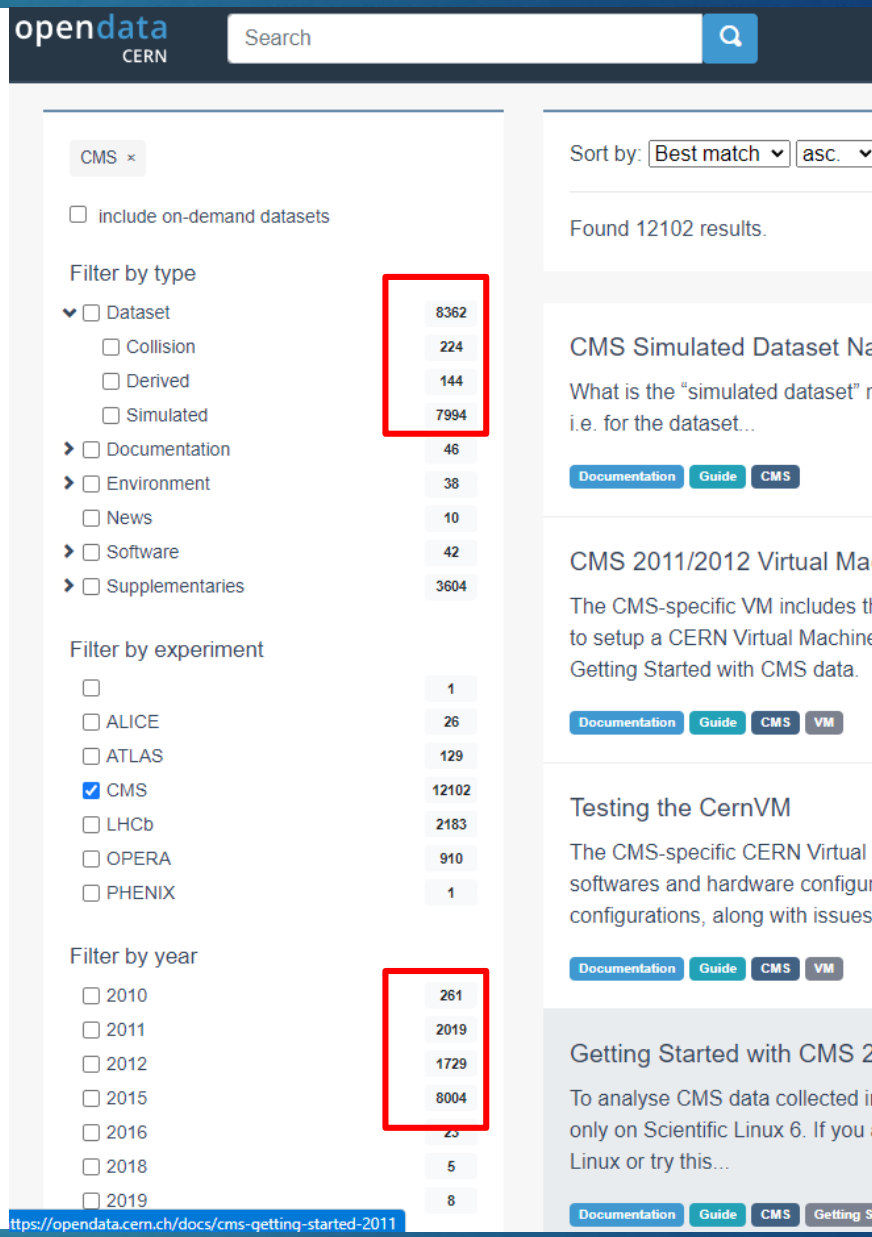
- CMS on the CERN Portal
- CMS Open Data policies
- Educational resources
- Research resources
- Plans for the future



CMS on the CERN Portal

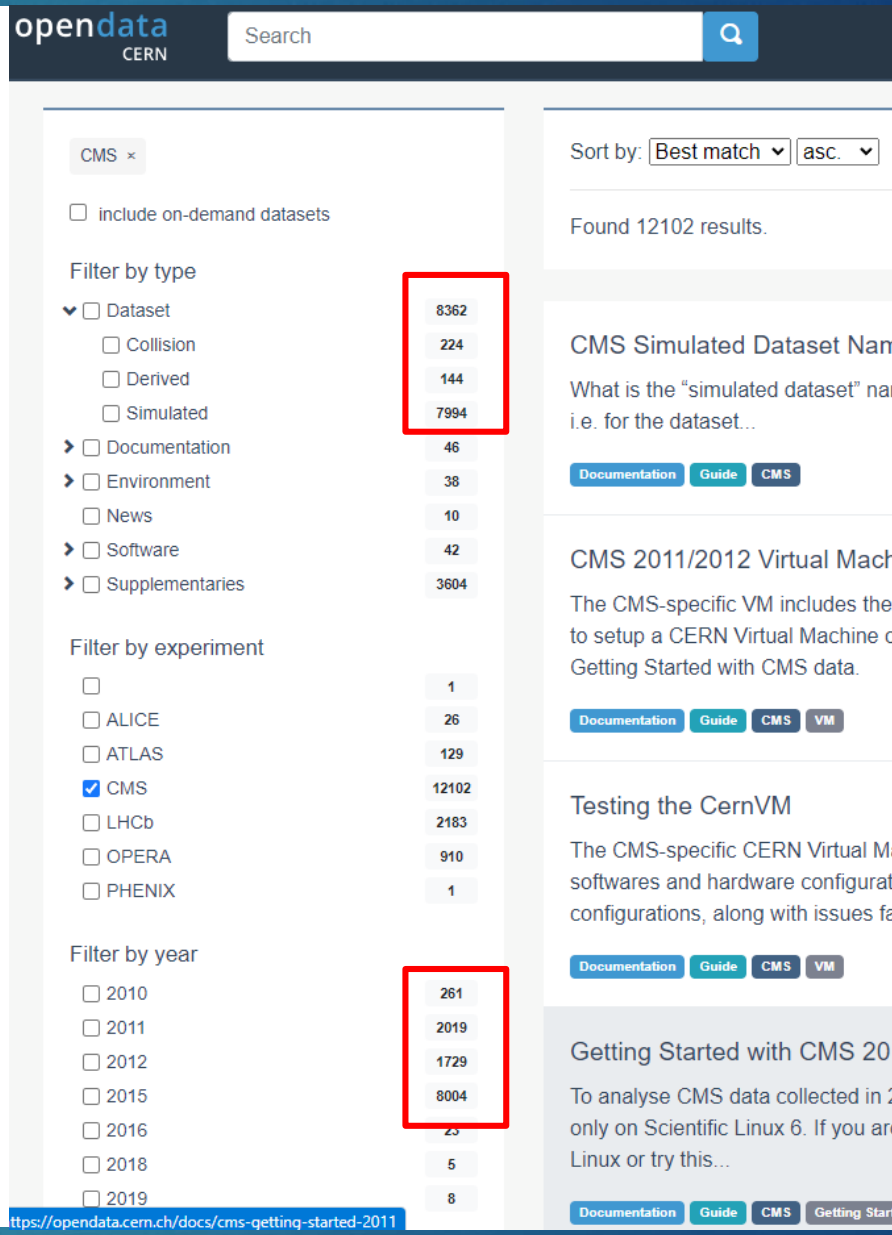
- Full Run 1 data!
  - Analysis Object Data (AOD) format
- First Run 2 data! (most of 2015)
  - MiniAOD format
- Derived datasets in NanoAOD format for simple analysis reproduction

Data Tier	Event size
Reconstructed data	~3 MB
Analysis Object Data (AOD)	~500 kB
MiniAOD	~50 kB
NanoAOD (flat ROOT)	1-2 kB



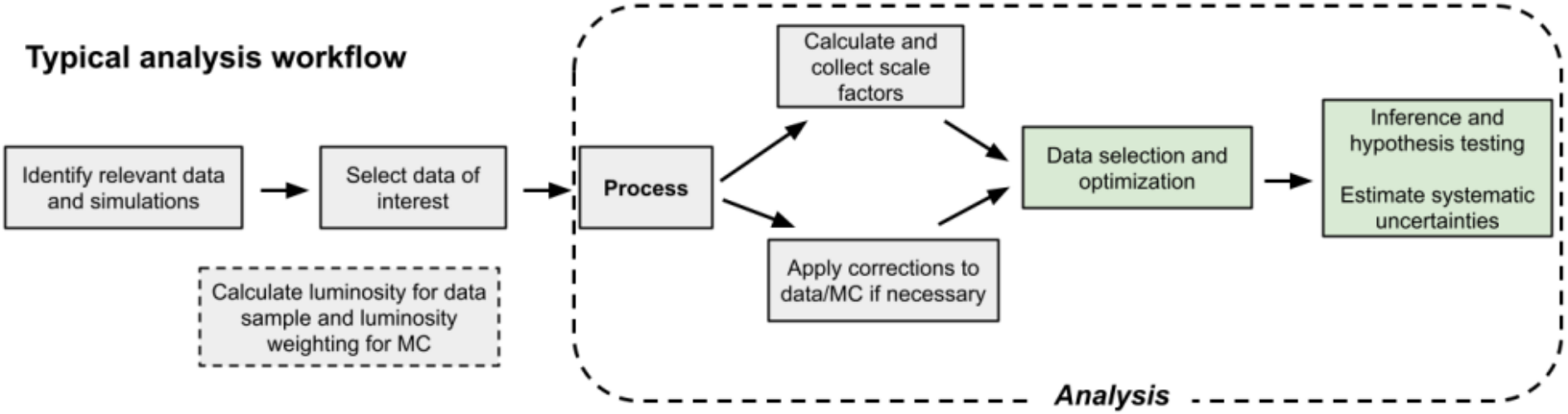
CMS on the CERN Portal

- Full Run 1 data!
  - Analysis Object Data (AOD) format
- First Run 2 data! (most of 2015)
  - MiniAOD format
- Derived datasets in NanoAOD format for simple analysis reproduction
- Simulations at each energy
- Software!
  - CMSSW via docker containers or VM
  - Validated data json files
  - Conditions database access
- Instruction guides and example code
- Support forum



CMS on the CERN Portal

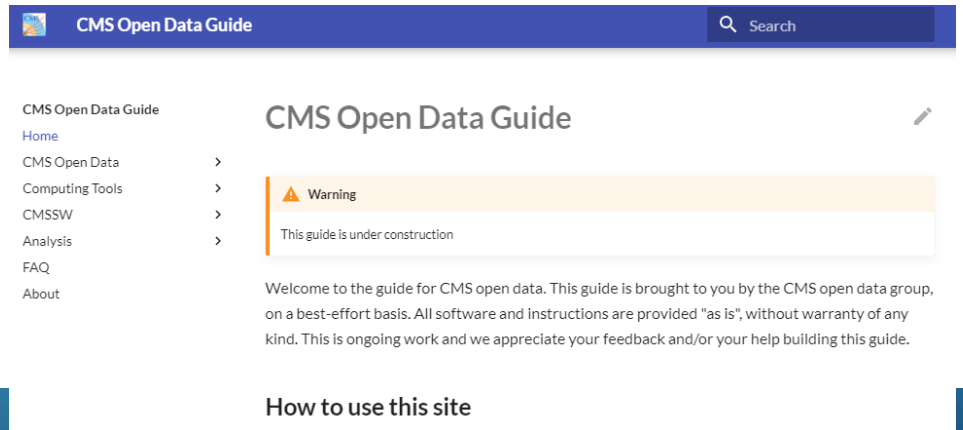
Data	Access data from CERN storage	Store reduced data locally
Processing	CMSSW	User-developed software
Software	CMS open data virtual machines/containers	User's computing environment



Documentation	Community documentation and training	
	CMS-provided documentation	

K. L.-P. et al, CHEP 2021

https://cms-opendata-guide.web.cern.ch/



CMS Open Data policies

2020 Policy

- Publications now required to release info on HEPData
- Full and simplified data released ~annually since 2014
  - Embargo of 6 years, then release 50% of a year's luminosity
  - Remainder released w/n 10 years
  - At ongoing collision energies, Open Data will have max 20% of luminosity
  - Anyone can request to "affiliate" with CMS for more access!
- Managed by Offline Software & Computing + Collab. Board
  - Data Preservation & Open Access
  - Kati Lassila-Perini, Julie Hogan coordinators
  - cms-dpoa-coordinators@cern.ch

Educational Resources

BETHEL UNIVERSITY

opendata CERN | Search | Help | About

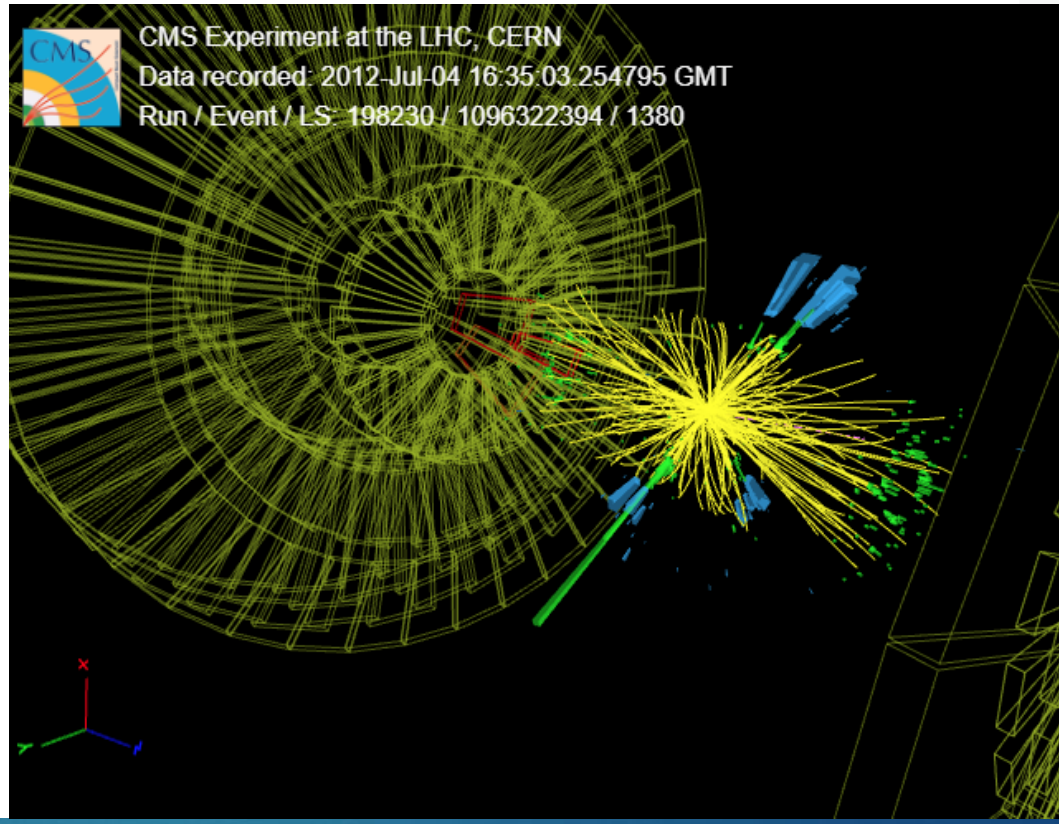
CMS Guide to education use of CMS Open Data

Documentation Guide

https://opendata.cern.ch/docs/cms-guide-for-education

This page will guide you through contents of the CMS Open Data collections that are meant for educational use (or for physics enthusiasts!). It is roughly broken down into three levels of difficulty:

- Beginner: *Visualise collisions*
- Intermediate: *Make histograms with collision data*
- Advanced: *Dive deeper into the data*



- iSpy Event Displays
  - + worksheets!
- Virtual Reality displays
- International Masterclass
- Dimuon analyses for schools
- University-level course tools

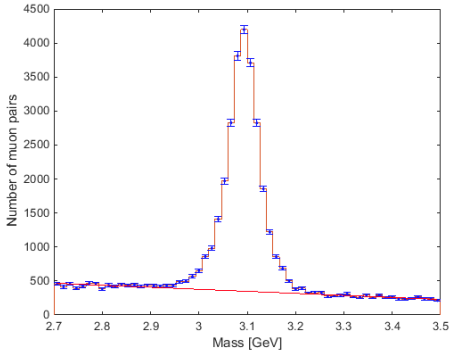
Educational Resources

launch binder

Particle Discovery lab (for students!)

The particle discovery lab uses CMS dimuon data from 2012 published via the CERN Open Data Portal. We have developed an undergraduate intermediate-level lab exercise to complement the many high school-level exercises available via the Open Data Portal. Solutions and student code are available in both MATLAB and Python, but do not require ROOT or Open Data Virtual Machines for students or instructors.

The goal of this exercise is for students to reconstruct decays of unknown particle X (initial state) to 2 muons (final state). They will use histograms to display their calculated mass for particle X, and learn about fitting and subtracting background contributions from data. Uncertainty propagation concepts are included through each step of the analysis. After isolating the signal distribution they will determine which particle they have discovered and compare their observed properties (mass and width) to the known properties.



ALPhA's Laboratory Immersions

Bethel University, St Paul, MN

High energy physics data analysis in a BFY lab

Research Resources

- Early feedback from authors that documentation of analysis procedures needed to be improved
- Three major thrusts to improve user experience:

1. Grow "NanoAOD"-like formats with corrections applied, identification algorithms computed, etc
   - 2011, 2012, 2015: Physics Object Extractor Tool (POET)
   - 2015: NanoAOD under preparation (and work ongoing to backport)

2. Build the CMS Open Data Guide

3. Offer annual CMS Open Data Workshops

Data Tier	Event size
Reconstructed data	~3 MB
Analysis Object Data (AOD)	~500 kB
MiniAOD	~50 kB
NanoAOD (flat ROOT)	1-2 kB

Research Resources

1. Grow "NanoAOD"-like formats with corrections applied, identification algorithms computed, etc
   - Physics Object Extractor Tool (POET) 2011, 2012, 2015
   - NanoAOD under preparation for 2015 (and work ongoing to backport)
     - NanoAOD-like files available for select 2012 datasets, without full corrections

Physics object extractor tool for the CMS 2012 data

Physics Object Extractor Tool (POET) is an example C++ code (and its configuration in Python) to extract the physics object information from CMS Open data. These instructions are valid to work with...

Software Tool Workflow CMS

Physics object extractor tool for the CMS 2015 data

Physics Object Extractor Tool (POET) is an example C++ code (and its configuration in Python) to extract the physics object information from CMS Open data. These instructions are valid to work with...

Software Tool Workflow CMS

Tool for conversion of CMS AOD files to reduced NanoAOD format for the purpose of education and outreach

The tool can be used to read events from CMS AOD files and convert them to a reduced NanoAOD data format for the purpose of education and outreach. Note that the tool is published for the documenta...

Software Tool Workflow CMS

Research Resources

BETHEL UNIVERSITY

2. CMS Open Data Guide: links, lessons, POET examples

Jet Uncertainty

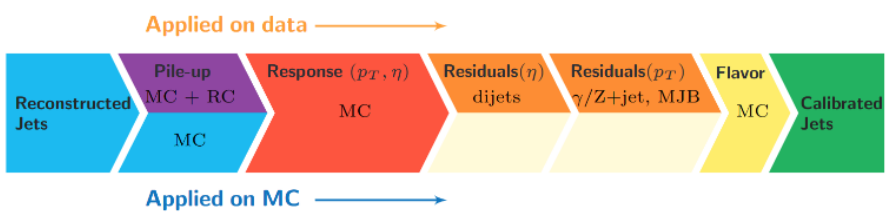
Unsurprisingly, the CMS detector does not measure jet energies perfectly, nor do simulation and data agree perfectly! The measured energy of jet must be corrected so that it can be related to the true energy of its parent particle. These corrections account for several effects and are factorized so that each effect can be studied independently.

Jet Energy Corrections (JEC)

What is JEC?

JEC is the first set of corrections applied on jets that adjust the mean of the response distribution in a series of correction levels.

Correction Levels



Particles from additional interactions in nearby bunch crossings of the LHC contribute energy in the calorimeters that must somehow be distinguished from the energy deposits of the main interaction. Extra energy in a jet's cone can make its measured momentum larger than the momentum of the parent particle. The first layer ("L1") of jet energy corrections accounts for pileup

Table of contents

CMS Open Data Guide

Research Resources

BETHEL UNIVERSITY

2. CMS Open Data Guide: links, lessons, POET examples

Tau 4-vector information

Run 1 Data | Run 2 Data

An example of an EDAnalyzer tau information is available in the TauAnalyzer of the Physics Object Extractor Tool (POET). The following header files needed for accessing tau information are included:

```
//classes to extract tau information
#include "DataFormats/TauReco/interface/PFTau.h"
#include "DataFormats/TauReco/interface/PFTauFwd.h"
#include "DataFormats/TauReco/interface/PFTauDiscriminator.h"
```

In TauAnalyzer.cc, the tau four-vector elements are accessed as shown below.

```
Handle<reco::PFTauCollection> mytaus;
iEvent.getByLabel(tauInput, mytaus);

[...]

for (reco::PFTauCollection::const_iterator itTau=mytaus->begin(); itTau!=mytaus->e
    if (itTau->pt() > tau_min_pt) {
        tau_e.push_back(itTau->energy());
        tau_pt.push_back(itTau->pt());
        tau_px.push_back(itTau->px());
        tau_py.push_back(itTau->py());
        tau_pz.push_back(itTau->pz());
        tau_eta.push_back(itTau->eta());
        tau_phi.push_back(itTau->phi());

[...]
}
```

Research Resources

3. Offer annual CMS Open Data Workshops – began in 2020

CMS Open Data Workshop for Theorists

Fermilab LHC Physics Center (LPC)

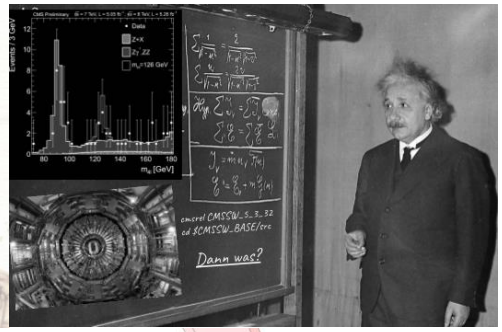
Online (link)

Sep 30 - Oct 2, 2020

8:30 am - 5:00 pm (US Central Time Zone)

Instructors: Matt Bellis, Edgar Carrera, Thomas Gaehtgens, Allan Da Silva Jales, Julie Hogan, Clemens Lange, Kati Lassila-Perini, Santeri Laurila, Adelina Lintuluoto, Tom McCauley, Sezen Sekmen, Jesse Thaler

Helpers: Asdrubal Cruz, Nada Mohamed, Nikolas Pervan, Farrah Simpson, Stefan Wunsch



CMS Open Data Workshop 2021

CERN (Virtual)

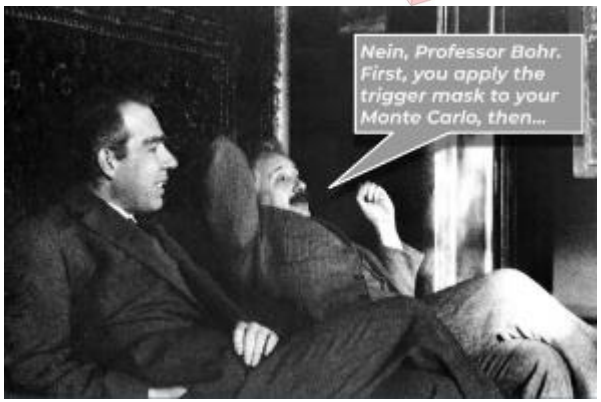
Online

Jul 19-22, 2021

2:30 pm - 6:40 pm (CET)

Instructors: Matt Bellis, Edgar Carrera Jarrin, Julie Hogan, Clemens Lange, Kati Lassila-Perini

Helpers: Marcelo Anda, Sid Boros, Anniina Kinnunen, Sarah Markham, Nick Pervan, Andro Petkovic, Farrah Simpson, Julian Westerlund



CMS Open Data Workshop 2022

CERN

Aug 1-4, 2022

2:30 pm - 6:30 pm (CET)

Instructors: M. Bellis, E. Carrera, A. Geiser, J. Hogan, C. Lange, K. Lassila-Perini, T. McCauley, S. Sekmen, X. Tintin, J. Yoo

Helpers: A. Chicaiza, K. Chicaiza, N. Dhingra, E. Jimenez, K. Johnson, D. Liko, P. LLerena, S. Markham, D. Mena, D. Merizalde, J. Ochoa, E. Piedra,

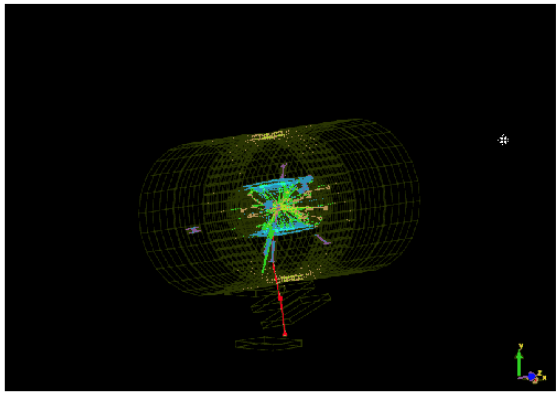
Research Resources

3. Offer annual CMS Open Data Workshops – began in 2020

- Pre-exercises on Docker, ROOT, CMSSW, etc
- Detailed lessons on using CMS physics objects
- Understanding corrections, uncertainties, scale factors
- Trigger usage, luminosity calculations, etc
- Hands-on analysis examples
- Scaling up analyses with cloud computing

Home | Code of Conduct | Setup | Episodes | Extras | License | Improve this page

Physics Objects



Prerequisites

- For this lesson we will be using the `CMSSW` Docker container you already used during the pre-exercises.
- After you completed the pre-exercises, you should have now a good understanding about `ROOT` and the basic aspects of `CMSSW`.

Schedule

	Setup	Before we begin
00:00	1. General Physics Objects and POET	What do we call physics objects in CMS? How are physics objects reconstructed? What is POET? How do I run POET?
01:00	2. Electrons	What are electromagnetic objects How are electrons treated in CMS? What variable are available using POET? How can one add a new electron variable?

Plans for the future

- Continue the release of Run 2 data!
  - Increased use of NanoAOD
  - Open Data will inherit “full Run 2” analysis coherency for 2016 – 2018 data
- Continue improving automation of tools
- Continue documentation of 2015 data usage
- CMS Open Data Workshop 2023!
  - July 11-14, 2023
  - Fermilab
  - USA mornings each day
  - Watch for more…

DO NOT LOOK