

# CMS Open Data

**Tom McCauley on behalf of the CMS Collaboration**

**[thomas.mccauley@cern.ch](mailto:thomas.mccauley@cern.ch)**

**University of Notre Dame, USA**

**[1st COMETA Workshop on Artificial Intelligence for Multi-boson Physics, 1 Oct 2024](#)**



# Outline

- CMS Open Data policy
- CMS Open Data releases
- Research resources
- Summary and future plans
- Hands-on

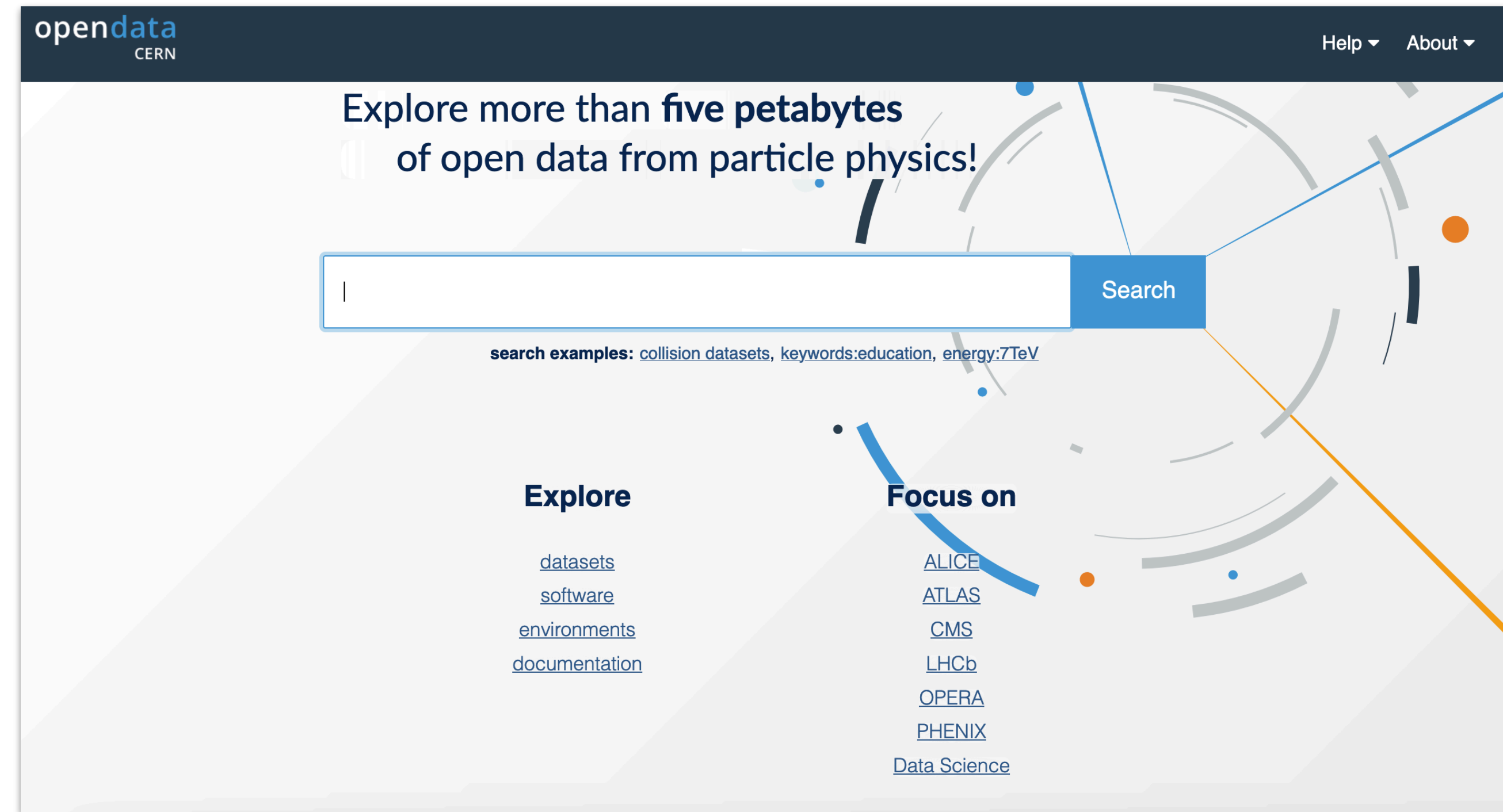
# CMS Open Data Policy

- CMS data preservation, re-use, and open access policy DOI:[10.7483/OPENDATA.CMS.1BNU.8V1W](https://doi.org/10.7483/OPENDATA.CMS.1BNU.8V1W)
- Data releases since 2014
- Publish 50% of luminosity after 6 years, remainder released within 10 years
- “Amount of open data will be limited to 20% of data with the similar centre-of-mass energy and collision type while such data are still planned to be taken”
- Releases are made under the open license Creative Commons [CC0](https://creativecommons.org/licenses/by/4.0/) waiver, essentially releasing into the public domain
- CMS Open Data Policy is coordinated by the CMS DPOA (Data Preservation and Open Access) group
- Motivation: Data preservation and open access are interdependent. Data can't be used and (re-used) unless it and the conditions for its use are preserved; data used are data preserved

# CMS Open Data releases

## CERN Open Data Portal

- CMS makes use of the [CERN Open Data Portal](#)
- Datasets are categorised, searchable, and citable (with each assigned a DOI)
- CLI via the `cernopendata-client` is available



[DoubleMu primary dataset in AOD format from RunA of 2011 \(/DoubleMu/Run2011A-12Oct2013-v1/AOD\)](#)

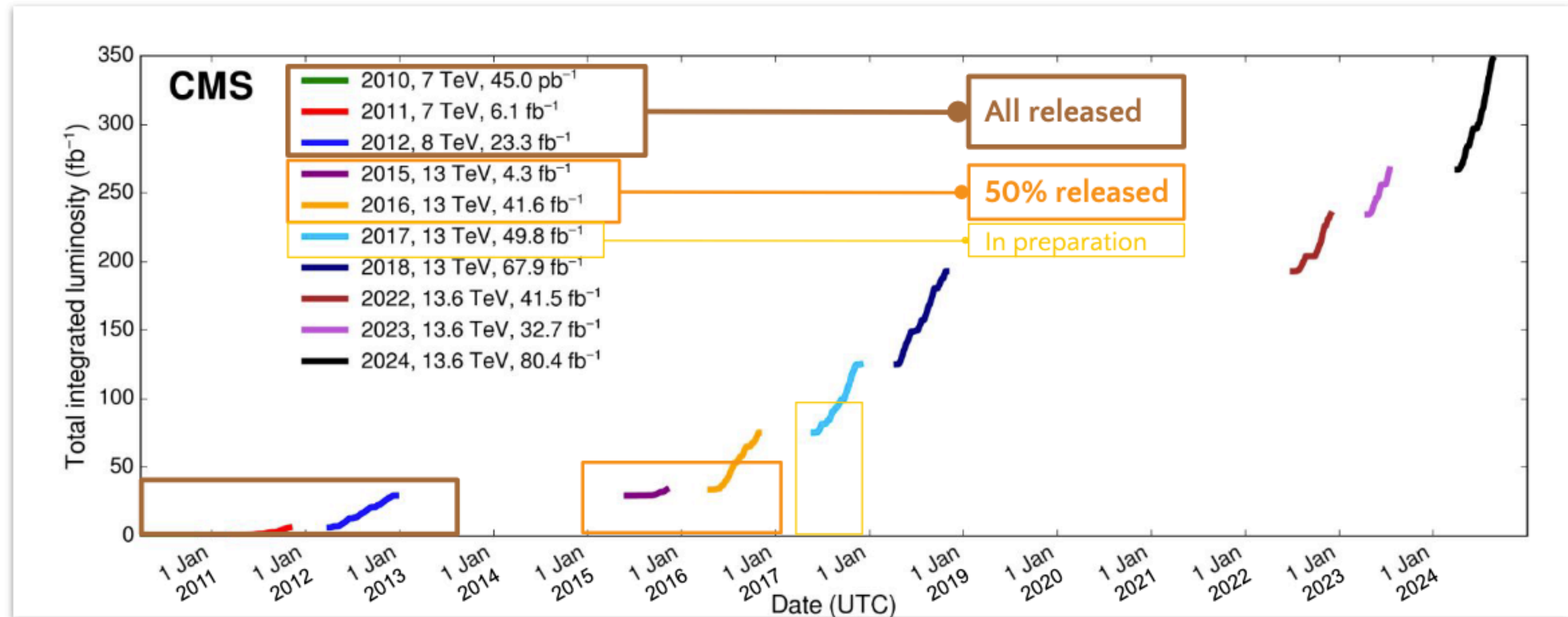
[/DoubleMu/Run2011A-12Oct2013-v1/AOD](#), CMS collaboration

Cite as: CMS collaboration (2016). DoubleMu primary dataset in AOD format from RunA of 2011 (/DoubleMu/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.RZ34.QR6N

[Dataset](#) [Collision](#) [CMS](#) [7TeV](#) [pp](#) [CERN-LHC](#)

# CMS Open Data releases

## Run eras



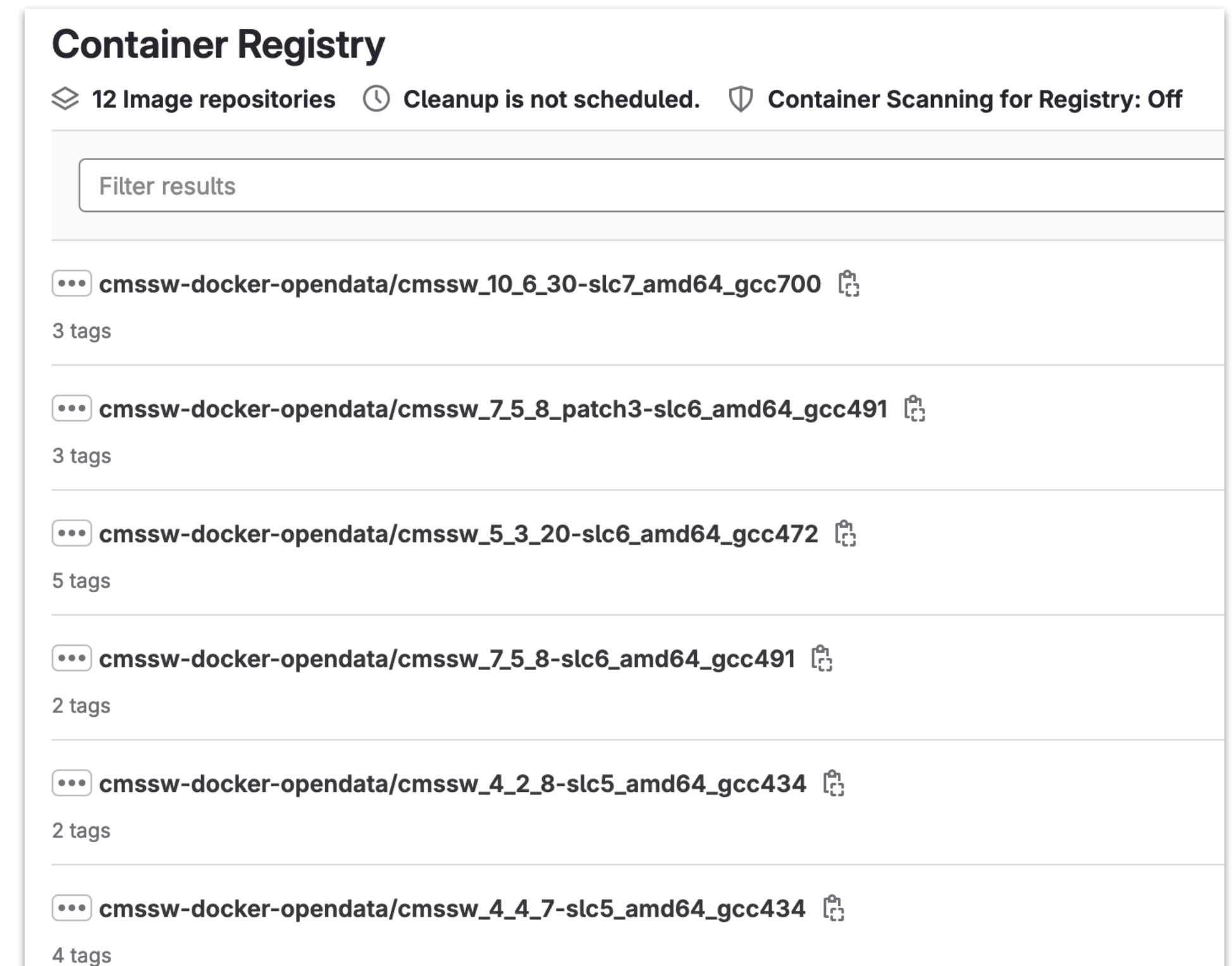
Credit: K. Lassila-Perini

# CMS Open Data releases

## Content

Providing the datasets isn't enough. A data release also includes:

- Accompanying simulation datasets
- Software environments via Docker containers and/or virtual machines
- Analysis software: CMSSW, example analyses, validated runs JSON files, conditions database access, ...
- Documentation (such as the [CMS Open Data Guide](#))
- Continued support via e.g. a [support forum](#)



The screenshot displays the Container Registry interface. At the top, it shows '12 Image repositories', 'Cleanup is not scheduled.', and 'Container Scanning for Registry: Off'. Below this is a search bar labeled 'Filter results'. The main content area lists several image repositories, each with a three-dot menu icon, the repository name, and a copy icon. The listed repositories are:

- cmssw-docker-opendata/cmssw\_10\_6\_30-slc7\_amd64\_gcc700 (3 tags)
- cmssw-docker-opendata/cmssw\_7\_5\_8\_patch3-slc6\_amd64\_gcc491 (3 tags)
- cmssw-docker-opendata/cmssw\_5\_3\_20-slc6\_amd64\_gcc472 (5 tags)
- cmssw-docker-opendata/cmssw\_7\_5\_8-slc6\_amd64\_gcc491 (2 tags)
- cmssw-docker-opendata/cmssw\_4\_2\_8-slc5\_amd64\_gcc434 (2 tags)
- cmssw-docker-opendata/cmssw\_4\_4\_7-slc5\_amd64\_gcc434 (4 tags)

# CMS Open Data releases

## Dataset record walkthrough

SingleMuon primary dataset in MINIAOD format from RunH of 2016 (/SingleMuon/Run2016H-UL2016\_MiniAODv2-v2/MINIAOD)

← Dataset name

/SingleMuon/Run2016H-UL2016\_MiniAODv2-v2/MINIAOD, CMS Collaboration

Cite as: CMS Collaboration (2024). SingleMuon primary dataset in MINIAOD format from RunH of 2016 (/SingleMuon/Run2016H-UL2016\_MiniAODv2-v2/MINIAOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.18K5.94CA

← DOI

Dataset Collision CMS 13TeV pp CERN-LHC Parent Dataset:

### Description

SingleMuon primary dataset in MINIAOD format from RunH of 2016. Run period from run number 281613 to 284044.

The list of validated runs, which must be applied to all analyses, either with the full validation or for an analysis requiring only muons, can be found in:

Validated runs, full validation

Validated runs, muons only

← List of validated (i.e. “good”) runs

### Related datasets

The corresponding NANOAO dataset:

/SingleMuon/Run2016H-UL2016\_MiniAODv2\_NanoAODv9-v1/NANOAO

### Dataset characteristics

174035164 events. 1356 files. 4.6 TiB in total.

← Datasets that are either the children or parent of this dataset

Which Docker container image is needed and where to get it

### System details

Recommended global tag for analysis: 106X\_dataRun2\_v37

Recommended release for analysis: CMSSW\_10\_6\_30

Recommended container image for analyses is available in the following locations (see guide):

- [docker.io/cmsopendata/cmssw\\_10\\_6\\_30-slc7\\_amd64\\_gcc700:latest](https://docker.io/cmsopendata/cmssw_10_6_30-slc7_amd64_gcc700:latest)
- [gitlab-registry.cern.ch/cms-cloud/cmssw-docker-opendata/cmssw\\_10\\_6\\_30-slc7\\_amd64\\_gcc700:latest](https://gitlab-registry.cern.ch/cms-cloud/cmssw-docker-opendata/cmssw_10_6_30-slc7_amd64_gcc700:latest)

# CMS Open Data releases

## Dataset record walkthrough

### How were these data selected?

Events stored in this [primary dataset](#) were selected because of the presence of at least one energetic [muon](#), or at least one [muon](#) and one or more jets, [tau](#) or high missing transverse momentum.

#### Data taking / HLT

The collision data were assigned to different RAW datasets using the following [HLT configuration](#).

#### Data processing

This MINIAOD dataset was processed from the RAW dataset by the following steps:

#### Step PAT

Release: [CMSSW\\_10\\_6\\_25](#)

Global tag: [106X\\_dataRun2\\_v35](#)

Configuration file for PAT step [ReReco-Run2016H-SingleMuon-UL2016\\_MiniAODv2](#)

Output dataset: [/SingleMuon/Run2016H-UL2016\\_MiniAODv2-v2/MINIAOD](#)

#### Step RECO

Release: [CMSSW\\_10\\_6\\_8\\_patch1](#)

Global tag: [106X\\_dataRun2\\_v27](#)

Configuration file for RECO step [recoSim\\_Run2016H\\_SingleMuon](#)

Output dataset: [/SingleMuon/Run2016H-21Feb2020\\_UL2016-v1/AOD](#)

#### HLT trigger paths

The possible HLT trigger paths in this dataset are:

[HLT\\_DoubleIsoMu17\\_eta2p1\\_noDzCut](#)

[HLT\\_DoubleIsoMu17\\_eta2p1](#)

[HLT\\_IsoMu16\\_eta2p1\\_MET30\\_LooseIsoPFTau50\\_Trk30\\_eta2p1](#)

[HLT\\_IsoMu16\\_eta2p1\\_MET30](#)

[HLT\\_IsoMu17\\_eta2p1\\_LooseIsoPFTau20\\_SingleL1](#)

[HLT\\_IsoMu17\\_eta2p1\\_LooseIsoPFTau20](#)

Data provenance and trigger paths

### How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, [trigger](#), [lumi](#) and physics objects ([tracking](#), [electron](#), [muon](#), [photon](#), [jet](#) and [MET](#)) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

[The Data Quality Monitoring Software for the CMS experiment at the LHC: past, present and future](#)

### How can you use these data?

You can access these data through the CMS Open Data container or the CMS Virtual Machine. See the instructions for setting up one of the two alternative environments and getting started in

[Running CMS analysis code using Docker](#)

[How to install the CMS Virtual Machine](#)

[Getting started with CMS open data](#)

How-tos: docker containers / VM and analysis



# Research data formats

## Tiers

- AOD: largest data format, requires CMS software for analysis, only available for Run 1
- miniAOD: smaller data format derived from AOD, requires CMS software for analysis, available for Run 2
- nanoAOD (*i.e.* ROOT-based ntuples) formats (with *e.g.* corrections applied and ID used) are produced and used more and more by CMS and have several advantages over larger formats beyond purely size: *e.g.* flatter physics object structure, no need for large C++ frameworks for analysis, possible use of frameworks and tools such as Coffea, RDataFrame, uproot, awkward, ...
- Note: there are other data tiers and formats available but the above cover most data records

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

# CMS ML datasets

- Several datasets have been generated from CMS Open Data specifically for ML applications
- The generator code has been provided as well as example code
- The datasets have been derived from miniAOD into TTrees
- Note: [Additional datasets for ML studies](#) have been released as well but are in other CMS formats

The screenshot displays the CMS ML datasets interface. On the left, there are filter panels for 'Current parameters', 'Availability', 'Type', and 'Experiment'. The 'Current parameters' panel shows 'CMS' and 'datascience' tags. The 'Availability' panel has a toggle for 'include on-demand datasets'. The 'Type' panel has 'Dataset (4)' selected. The 'Experiment' panel has 'CMS (4)' selected. The main content area shows three dataset entries, each with a title, description, and tags.

**Current parameters** [Clear all](#)

CMS x datascience x  
Dataset x

**Availability**

include on-demand datasets

**Type**

Dataset (4)  
 Derived (4)  
 Software (4)  
 Tool (4)

**Experiment**

ATLAS (7)  
 CMS (4)  
 LHCb (1)

---

**Sample with tracker hit information for tracking algorithm ML studies TTbar\_13TeV\_PU50\_PixelSeeds**  
This dataset consists of a collection of pixel doublet seeds, i.e. the hit pairs that could belong to the same particle flying through the CMS Silicon Pixel Detector. These can be used in ML studie...

Dataset Derived CMS

---

**Samples with full event information including tracker hits for tracking, ML, and top quark tagging studies**  
Samples in this record are in a custom root ntuple format and contain the position of the hits and information from the generator-level objects associated to the tracker hits. The samples can be us...

Dataset Derived CMS

---

**Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple\_HiggsToBB\_QCD\_RunII\_13TeV\_MC**  
The dataset consists of particle jets extracted from simulated proton-proton collision events at a center-of-mass energy of 13 TeV generated with Pythia 8. It has been produced for developing machi...

Dataset Derived CMS

---

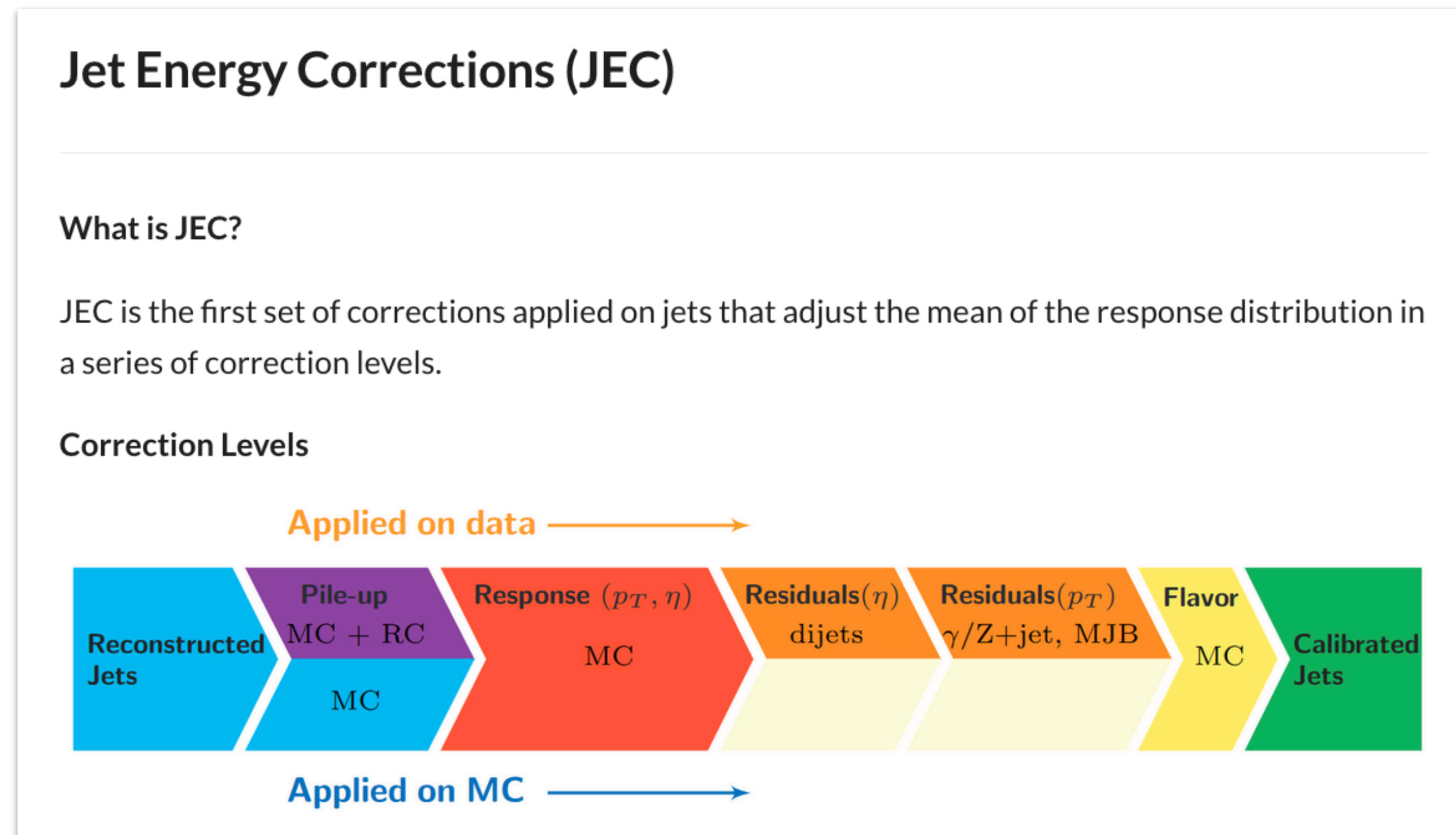
**Sample with jet properties for jet-flavor and other jet-related ML studies JetNTuple\_QCD\_RunII\_13TeV\_MC**  
The dataset consists of particle jets extracted from simulated proton-proton collision events at a center-of-mass energy of 13 TeV generated with Pythia 8. The particles emerging from the collision...

Dataset Derived CMS

# Research resources

## CMS Open Data Guide

The [CMS Open Data Guide](#) provides the information needed for analysis in one place

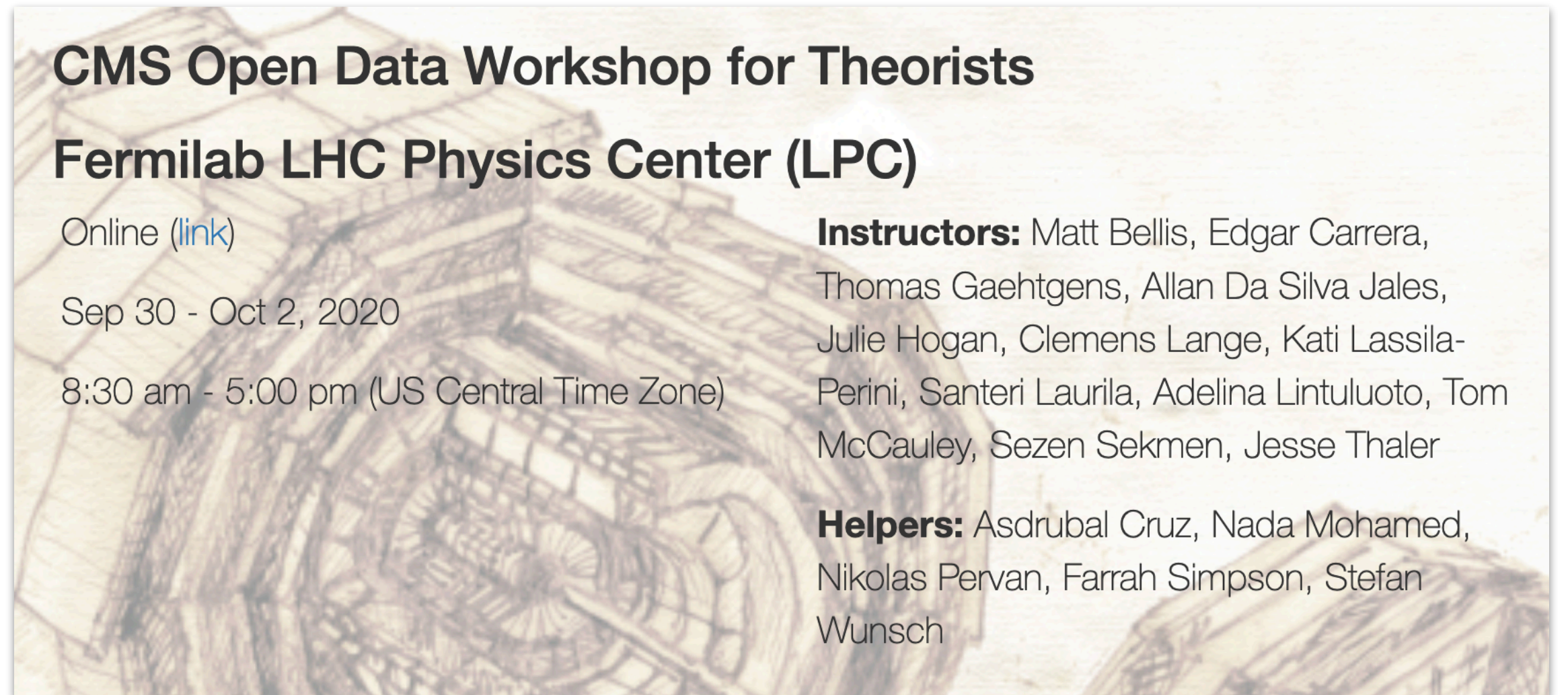


- CMS Open Data Guide
  - Home
  - CMS Open Data
    - CERN Open Data Portal
    - CMS Open Data
    - Finding Data
    - Workshops
  - Computing Tools
    - UNIX
    - ROOT
    - C++ and Python
    - Git
    - Docker
    - Virtual Machines
  - CMSSW
    - Overview
    - Data Model
    - Analyzers
    - Configuration
    - Conditions Data
  - Analysis
    - Data and Simulation
    - Selection
    - Luminosity
    - Background Modelling
    - Systematics
    - Statistical Interpretation
  - FAQ
  - About

# Research resources

## Open Data Workshops

- Since 2020 CMS have been offering [Open Data Workshops](#)
- The goal: to lower the threshold to access and use open data for theorists, phenomenologists, MLs, ...
- Perhaps come to the next one



**CMS Open Data Workshop for Theorists**  
**Fermilab LHC Physics Center (LPC)**

Online ([link](#))

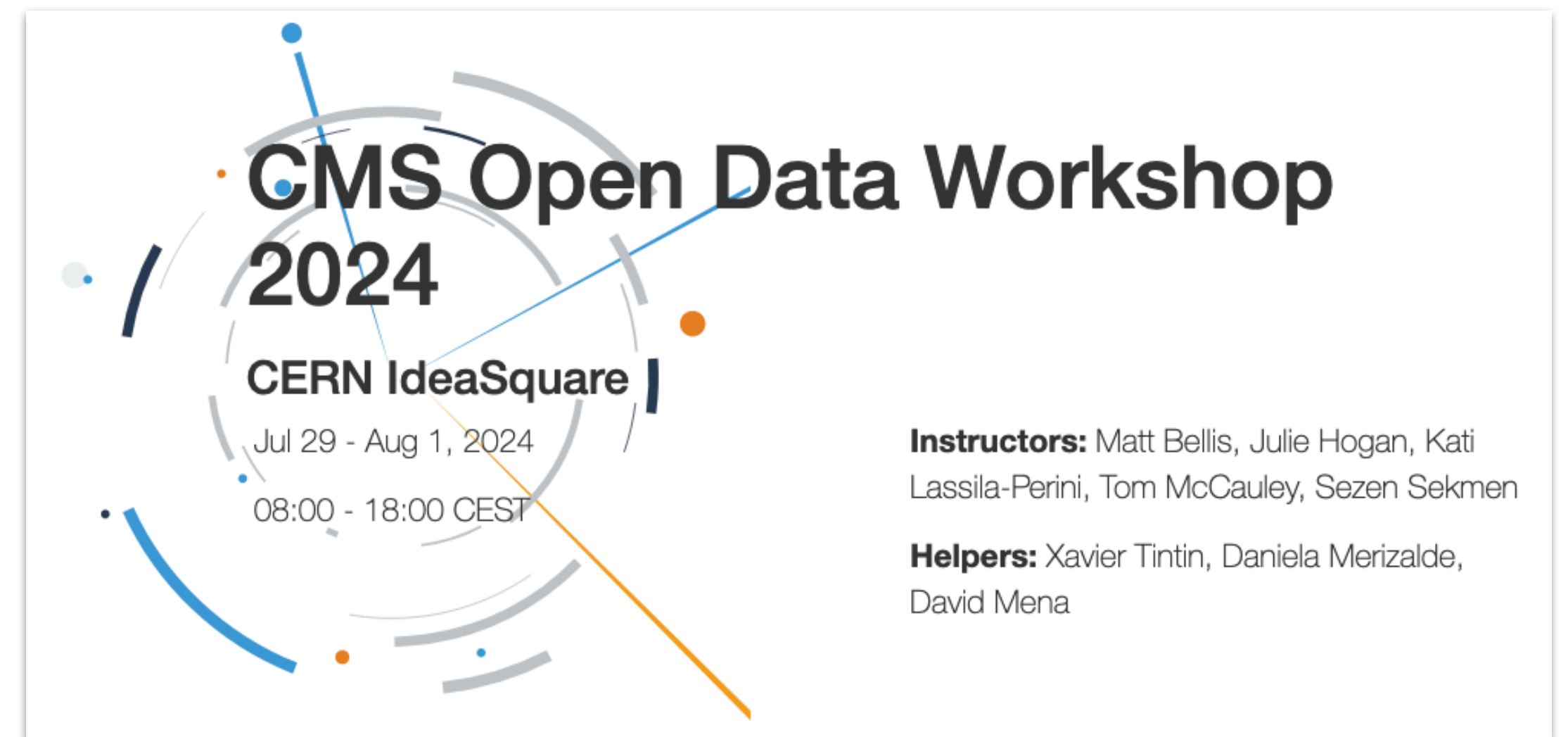
Sep 30 - Oct 2, 2020

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**  
**2024**

**CERN IdeaSquare |**

Jul 29 - Aug 1, 2024

08:00 - 18:00 CEST

**Instructors:** Matt Bellis, Julie Hogan, Kati Lassila-Perini, Tom McCauley, Sezen Sekmen

**Helpers:** Xavier Tintin, Daniela Merizalde, David Mena

# Research resources

## Open Data Workshops

- Uses [Carpentries](#) training template
- Pre-exercises on version control, containers, ROOT, C++, python, CMS software
- Hands-on lessons on CMS physics object usage, corrections, scale factors, uncertainties, trigger, luminosity calculation, etc.
- Hands-on analysis examples

Monday July 29			Tuesday July 30		
14:00-14:30	Welcome to the IdeaSquare	IdeaSquare Team	9:00-9:30	Hackathon Introduction	Julie Hogan Xavier Tintin
14:30-15:15	Introduction to CMS Open Data	Kati Lassila-Perini	9:30-10:30	Hackathon project division	Julie Hogan Xavier Tintin
15:15-16:00	Open Discussion: Your hopes for Open Data		10:30-11:00	Break	
16:00-16:30	Break		11:00-12:30	Hackathon working period	Julie Hogan Xavier Tintin
16:30-17:30	Exploring CMS NanoAOD (lesson)	Kati Lassila-Perini Tom McCauley	12:30-14:00	Lunch	
17:30-18:00	Exploring CMS NanoAOD (activity)	Kati Lassila-Perini Tom McCauley	14:00-14:30	Inspiration talk: BSM physics via anomaly detection	Julie Hogan
			14:30-15:30	Triggers & Luminosity (lesson)	Julie Hogan
			15:30-16:00	Triggers & Luminosity (activity)	Julie Hogan
			16:00-16:30	Break	
			16:30-17:30	Event Selection (lesson)	Matt Bellis
			17:30-18:00	Event Selection (activity)	Matt Bellis
Wednesday July 31			Thursday Aug 1		
9:00-10:30	Hackathon working period	Julie Hogan Xavier Tintin	9:30-10:30	Hackathon working period	Julie Hogan Xavier Tintin
10:30-11:00	Break		10:30-11:00	Break	
11:00-12:30	Hackathon working period	Julie Hogan Xavier Tintin	11:00-12:30	Hackathon working period	Julie Hogan Xavier Tintin
12:30-14:00	Lunch		12:30-14:00	Lunch	
14:00-15:00	Background modeling (lesson)	Matt Bellis	14:00-14:30	Hackathon Progress Report	Xavier Tintin (TBC)
15:00-15:45	Background modeling (activity)	Matt Bellis	14:30-15:30	Statistical Inference (lesson)	Sezen Sekmen
15:45-16:15	Break		15:30-16:00	Statistical Inference (activity)	Sezen Sekmen
16:15-17:15	Experimental uncertainties (lesson)	Julie Hogan	16:00-16:30	Break	
17:15-18:00	Experimental uncertainties (activity)	Julie Hogan	16:30-17:30	Resources & tools for CMS Open Data	Julie Hogan
			17:30-18:00	Closing discussion & survey	Julie Hogan



# Resources

## Summary

- Over 4 PB of [research-level collision data and simulation](#)
- CERN Open Data Portal: <https://opendata.cern.ch/>
- CMS Open Data Guide: <https://cms-opendata-guide.web.cern.ch/>
- CMS Open Data Forum: <https://opendata-forum.cern.ch/c/cms/6>
- CMS Open Data Workshops: <https://cms-opendata-guide.web.cern.ch/cmsOpenData/workshops/>

# Summary and future plans

- CMS continues to implement its open data policy with regular data releases (including documentation, code, software environments, ...)
- There are currently over 4 PB of level 3 (“research level”) collision data and simulation available as open data via the CODP
- Release of Run 2 data from 2017 has been recently approved and is in the midst of preparation
- We’re always working to improve and expand documentation (including training materials and workshops)



# Acknowledgements

- DPOA coordinators present and past including Julie Hogan and Kati Lassila-Perini
- DPOA team: a small but dedicated group
- CERN IT and SIS
- Thanks to COMETA/NIKHEF for the invitation

# Hands-on

GitHub page for the hands-on tutorial:

<https://github.com/cms-dpoa/cms-nikhef-tutorial>

**Backup**

# CMS Open Data Policy

## Levels

- Level 1: data directly related to publications
- **Level 2: simplified data formats suitable for education and outreach**
- **Level 3: “analysis level” reconstructed data and simulation and software**
- Level 4: raw data and associated software

# Level 3 collision data and MC releases

- 2010 p-p collision data at 7 TeV
- 2010 Pb-Pb collision data at 2.76 TeV
- 2011 p-p collision data at 7 TeV + MC
- 2011 p-p collision data at 2.76 TeV and p-Pb collision data at 5.02 TeV
- 2012 p-p collision data at 8 TeV + MC
- 2013 p-p collision data at 2.76 TeV and p-Pb collision data at 5.02 TeV + MC
- 2015 p-p collision data at 5.02 TeV and at 13 TeV + MC
- 2018 p-p collision MC at 13 TeV for ML studies
- 2016 p-p collision data at 13 TeV + MC