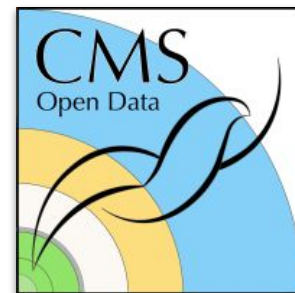


# CMS Open Data

**Matt Bellis (Siena University)** and Tom McCauley (Notre Dame)  
for the CMS DPOA effort

[TREASURE: Tokenizing HEP Collider Data for AI discovery](#)

April 27, 2026



# CERN 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



CMS

[Data Preservation in High Energy Physics \(DPHEP\) paper](#)

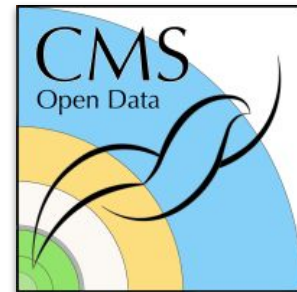
# CMS Open Data Policy

Data releases since 2014

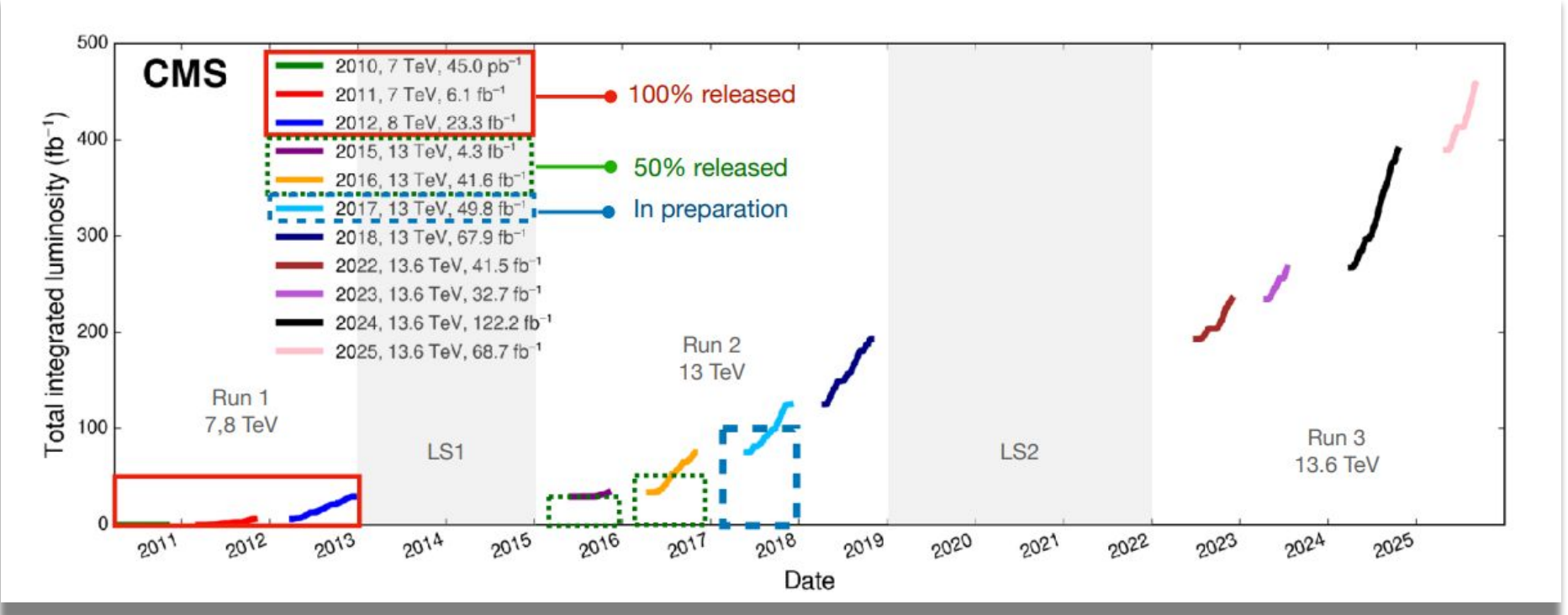
CMS data preservation, re-use, and open access policy

[DOI:10.7483/OPENDATA.CMS.1BNU.8V1W](https://doi.org/10.7483/OPENDATA.CMS.1BNU.8V1W)

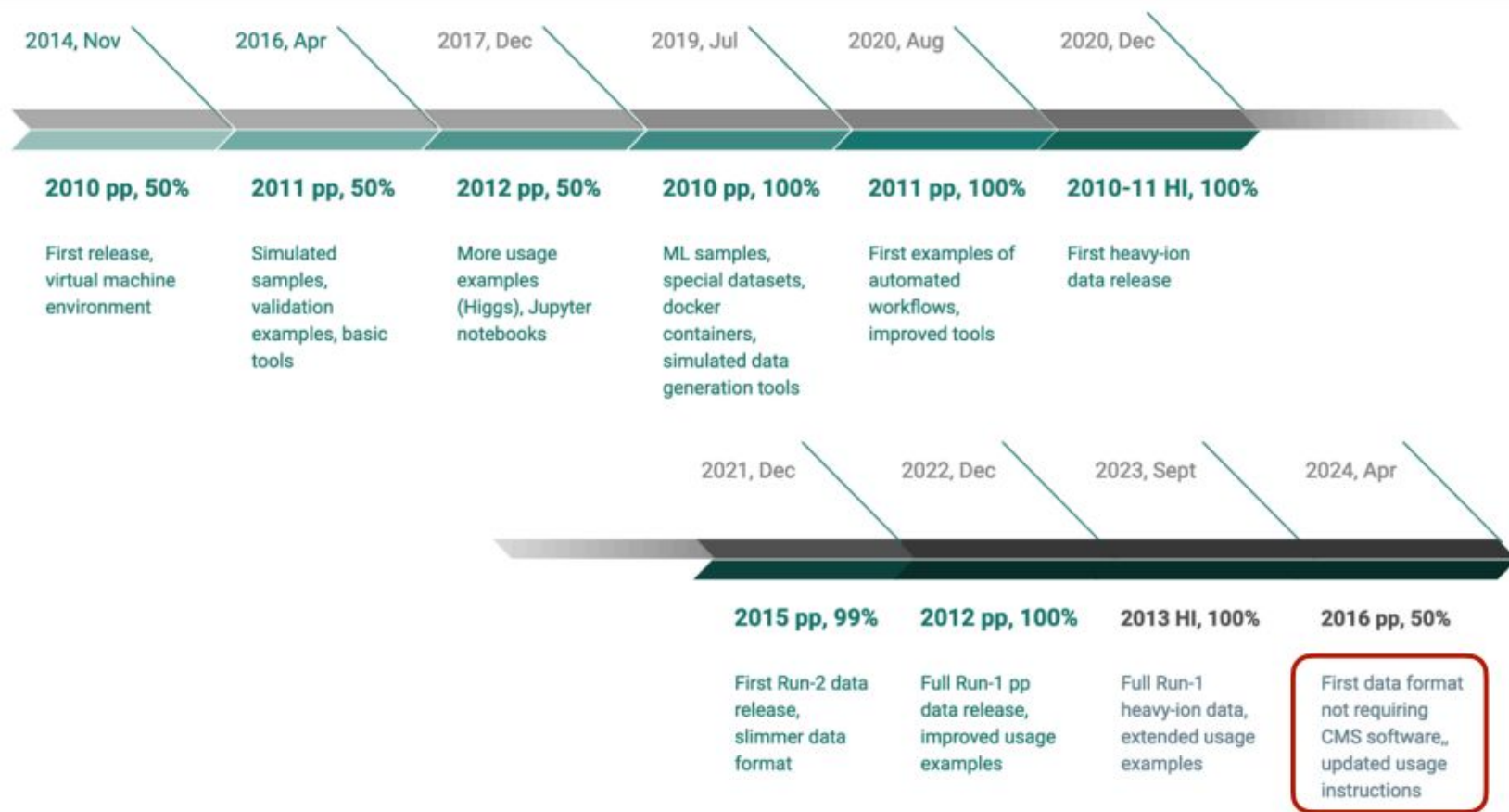
- Publish 50% of luminosity after 6 years
  - Remainder released within 10 years
- *“Amount of open data will be limited to 20% of data with the similar centre-of mass energy and collision type while such data are still planned to be taken”*
- Releases are made under the open license [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/)
  - Essentially releasing into the public domain
- Principles aligned with the [CERN open data policy for the LHC experiments](https://cern.ch/osp/open-access-policy)



# CMS Open Data Releases - *run eras and released data*



# CMS Open Data Releases - *timeline*



# CMS Open Data Releases - *CERN Open Data Portal*

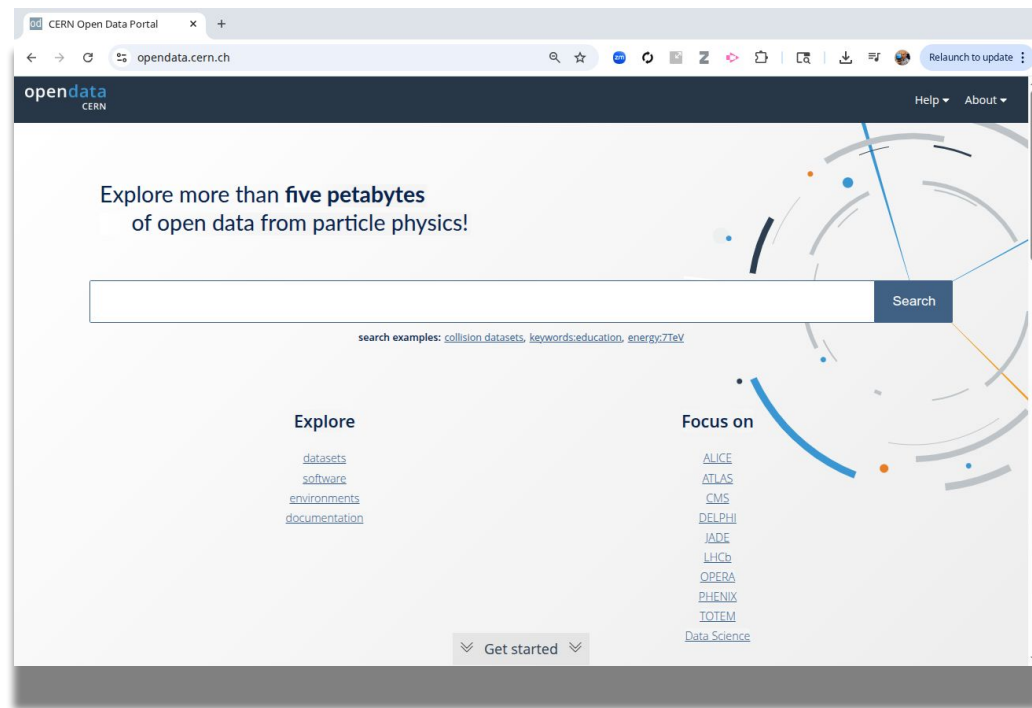
CMS data are available via the [CERN Open Data Portal](https://opendata.cern.ch)

Datasets are

- *Categorised*
- *Searchable*
- *Citable*

Lots of data!

- 300+ collision datasets
- 50k+ simulated datasets
- 4+ PB (disk and tape)

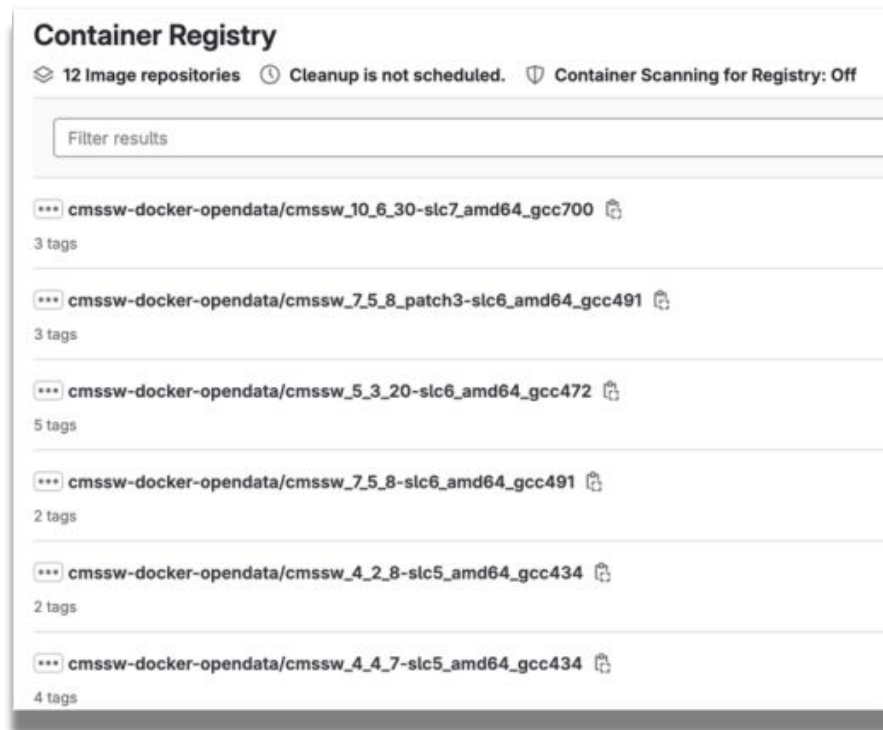


# CMS Open Data Releases - *content*

Access to datasets isn't enough.

A data release also includes:

- **Software environments**
  - Docker containers
  - Virtual machines
- **Analysis software:** CMS software, example analyses, validated run information, conditions database access, ...
- **Documentation** (such as the CMS Open Data Guide)
- **Continued support**
  - [CMS support forum](#)



# Data formats - *tiers*

- **AOD**: largest data format,
  - Requires CMS software for analysis
  - Only available for Run 1
  - ([link](#), [link](#), [link](#))
- **miniAOD**: smaller data format derived from AOD
  - Requires CMS software for analysis
  - Available for Run 2
  - ([link](#), [link](#))
- **nanoAOD**: flat, ROOT-based ntuples
  - Corrections applied, selectors included
  - Are produced and used more and more by CMS
  - No need for large C++ frameworks for analysis
  - Analysts can use standard python frameworks and tools such as Coffea, RDataFrame, uproot, awkward, ...
  - ([link](#))

Data Tier	Event size (kB)
<i>Reconstructed data</i>	~3000
<i>AOD</i>	~500
<i>miniAOD</i>	~50
<i>NanoAOD</i>	~2

# Data formats - *PFNanoAOD*

While the NanoAOD format has reduced information in comparison to MiniAOD (e.g. it does not have jet constituents) it is possible to create enhanced, extended versions of NanoAOD

**PFNano** is such an enhanced version derived from MiniAOD which included information on *Particle Flow Candidates* (PFCands)

*Helpful for jet reclustering, for example*

Software for production of PFNano is available [here](#)

One usage: [Aspen Open Jets](#)

Current parameters Clear all

CMS (17) 17

Type

Dataset (17)

Derived (17)

Experiment

CMS (17)

Availability

online (17)

Year

DoubleEG dataset in NanoAOD format enhanced with Particle Flow candidates from RunG of 2016  
DoubleEG dataset in NanoAOD format enhanced with Particle Flow candidates, readable with bare ROOT or other ROOT-compatible software. In addition to the default NanoAOD content, it contains candid...

JetHT dataset in NanoAOD format enhanced with Particle Flow candidates from RunG of 2016  
JetHT dataset in NanoAOD format enhanced with Particle Flow candidates, readable with bare ROOT or other ROOT-compatible software. In addition to the default NanoAOD content, it contains candidates...

Charmonium dataset in NanoAOD format enhanced with Particle Flow candidates from RunG of 2016  
Charmonium dataset in NanoAOD format enhanced with Particle Flow candidates, readable with bare ROOT or other ROOT-compatible software. In addition to the default NanoAOD content, it contains candi...

BTagCSV dataset in NanoAOD format enhanced with Particle Flow candidates from RunG of 2016  
BTagCSV dataset in NanoAOD format enhanced with Particle Flow candidates, readable with bare ROOT or other ROOT-compatible software. In addition to the default NanoAOD content, it contains candid...

PFCands (back to top)

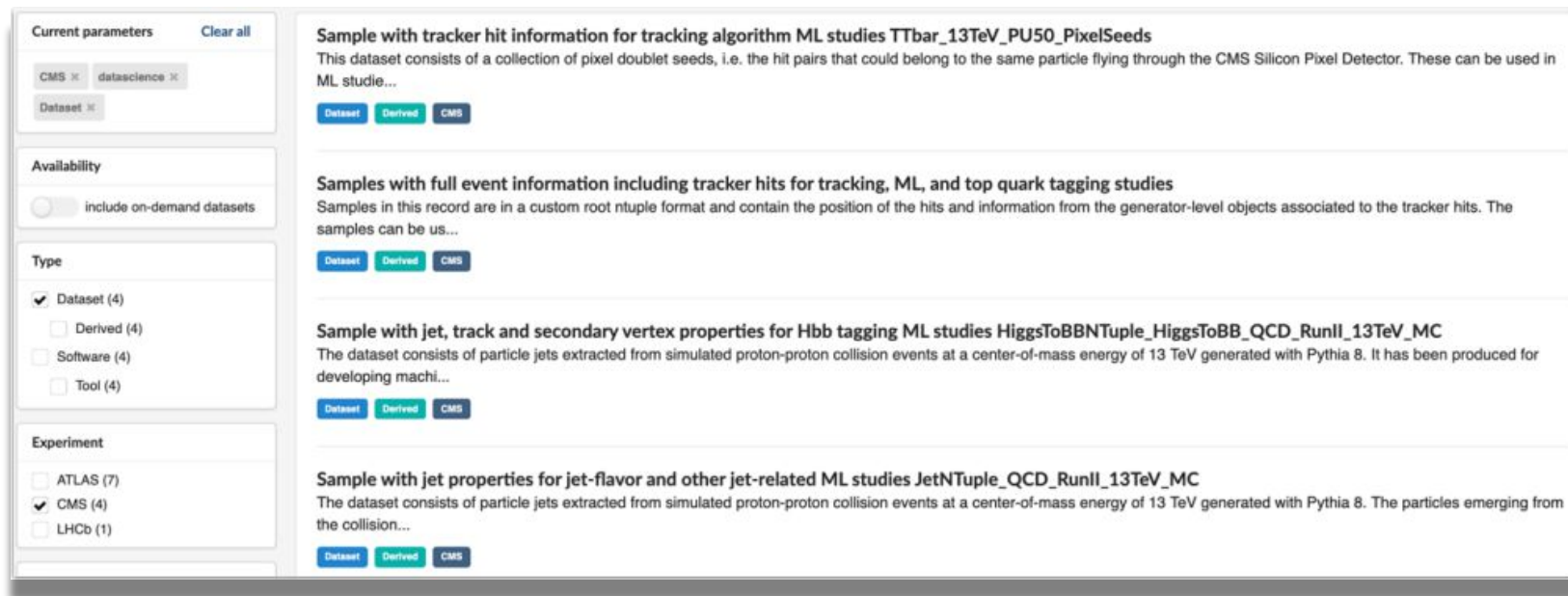
Object property	Type	Description
PFCands_charge	Int_t	electric charge
PFCands_d0	Float_t	pf d0
PFCands_d0Err	Float_t	pf d0 err
PFCands_dx	Float_t	pf dx
PFCands_dxErr	Float_t	pf dx err
PFCands_eta	Float_t	eta
PFCands_lostInnerHits	Int_t	lost inner hits
PFCands_mass	Float_t	mass
PFCands_pdgId	Int_t	PDG code assigned by the event reconstruction (not by MC truth)
PFCands_phi	Float_t	phi
PFCands_pt	Float_t	pt
PFCands_puppiWeight	Float_t	Puppi weight
PFCands_puppiWeightNoLep	Float_t	Puppi weight removing leptons
PFCands_pvAssocQuality	Int_t	primary vertex association quality
PFCands_trkChi2	Float_t	normalized trk chi2
PFCands_trkQuality	Int_t	track quality mask
PFCands_vtxChi2	Float_t	vertex chi2
nPFCands	UInt_t	interesting particles from AK4 and AK8 jets

# Datasets - *ML applications*

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



The screenshot displays a web interface for searching datasets. On the left, there are filter panels for 'Current parameters', 'Availability', 'Type', and 'Experiment'. The 'Current parameters' panel shows 'CMS' and 'datascience' selected. 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 search results, each with a title, description, and buttons for 'Dataset', 'Derived', and 'CMS'.

**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](#)

# Datasets - *storage and access going forward*

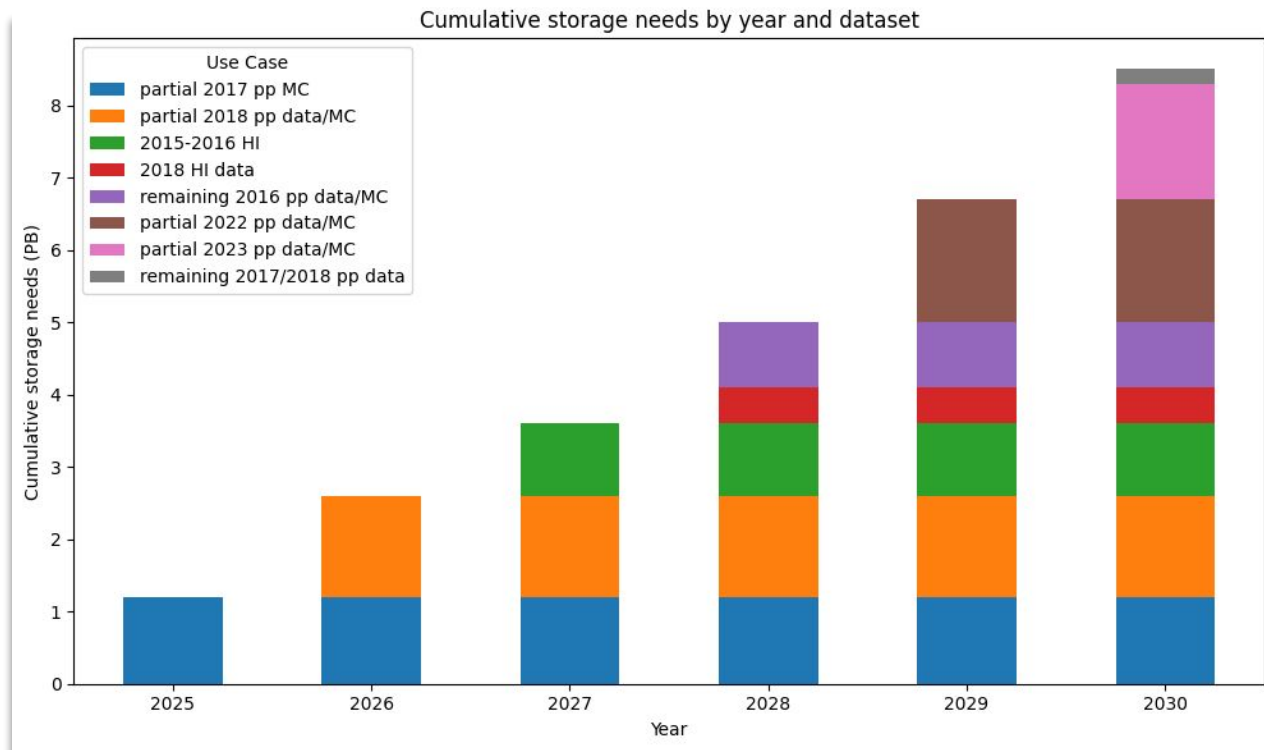
**Planned CMS releases going forward** (*on top of current datasets*)

An ambitious plan!

Working with CERN IT...

Some datasets will be on *tape* (cold storage) rather than *disk*

Access could be days or weeks for *tape storage*



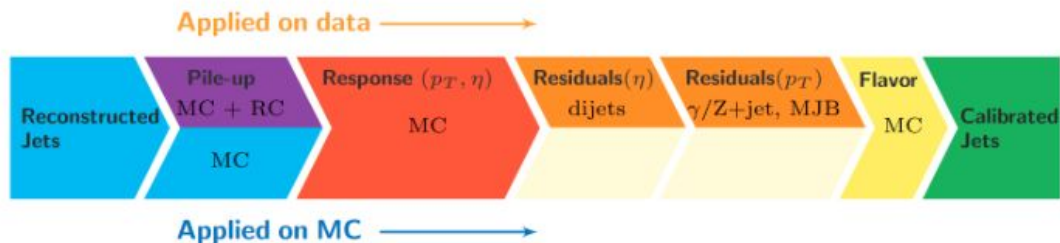
# Resources - *CMS Open Data Guide*

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

## 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



## 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](#)

CMS SW ▼

[Overview](#)

[Data Model](#)

[Analyzers](#)

[Configuration](#)

[Conditions Data](#)

Analysis ▼

[Data and Simulation](#) >

[Selection](#) >

# Resources - workshops

## [CMS Open Data Guide](#)

Aimed at theorists/phenomenologists

Attracted experimentalists and some teachers

- 2020 workshop
- 2021 workshop
- 2022 workshop
- 2023 workshop
- 2024 WHEPP India
- 2024 workshop

Paused in 2025

**Next one - July 28-July 30 - Notre Dame**  
**Pedagogical focus**

Hosted by University of Notre Dame & QuarkNet

Calling all high school and higher ed teachers!

Who wants to use particle physics data in their classroom?

**CMS Open Data Workshop**  
**Education and Pedagogy Hackathon!**

July 28<sup>th</sup> - July 30<sup>th</sup>, 2026

Location: University of Notre Dame & virtual participation

Join us for the 6th workshop, the first focused on education and pedagogy using the CMS experiment's open data. Over three days, we'll host tutorials and hacking sessions to create new, usable K-12 or college classroom products.

Since 2014, the CMS experiment at the Large Hadron Collider has released data publicly via the CERN Open Data Portal. This workshop will cover these datasets and their underlying physics. We want your input on making this data classroom-useful. We want you to help create resources, during the workshop, to share with other teachers.

<p><b>Organizing Committee</b></p> <p>Matt Bellis (Siena University)          Julie Hogan (Bethel University)          Clemens Lange (Paul Scherrer Institute)          Kati Laszlo-Pivari (Helsinki Institute of Physics)          Thomas McCauley (Notre Dame)</p> <p><b>Facilitators</b></p> <p>Emily Renssch (Siena University)          Pablo Sarz (CERN)          Kotte Salvatore (Siena University)</p>	<p><b>Local Organizing Committee</b></p> <p>Ken Cecire (Notre Dame &amp; QuarkNet)          Abhishek Bhatta (Notre Dame)          Michael Hirschi (Notre Dame)          Colin Joseph (Notre Dame)          Kevin Lannon (Notre Dame)          Marc Osherson (Notre Dame)          Mitchell Wayne (Notre Dame)          Shane Wood (Notre Dame &amp; QuarkNet)</p>
--	---

<https://indico.cern.ch/e/cmsdws2026>

Hosted by the Fermilab LHC Physics Center

CMS Open Data Workshop for Theorists  
 Virtual!  
 Sept 30-Oct 2, 2020

So much data! I can learn to analyze them!

CMS Open Data Workshop  
 July 19 - 22, 2021 Virtual

Hosted by the Fermilab LHC Physics Center

CMS Open Data Workshop  
 Aug 1<sup>st</sup> - 4<sup>th</sup>, 2022 CERN, Geneva, CH

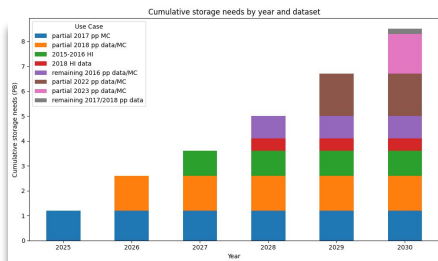
Start your day with a cup of CMS open data

CMS Open Data Workshop  
 July 11<sup>th</sup> - 14<sup>th</sup>, 2023 Fermilab, LHC, Batavia, IL USA

Hosted by CMS @ CERN's IdeaSquare

CMS Open Data Workshop & Hackathon  
 July 29<sup>th</sup> - Aug 1<sup>st</sup>, 2024 CERN IdeaSquare

# Summary and plans



CMS continues to implement its open data policy

- Regular data releases
- Updating documentation, code, software environments, etc

Currently over **4 PB research level** collision data and simulation available as open data via the CODP

Release of Run 2 data from 2017 is in the midst of preparation

**In 2014, we weren't preserving data because we foresaw AI and the need for training data**

**We preserved it because...who knows what's coming next?**

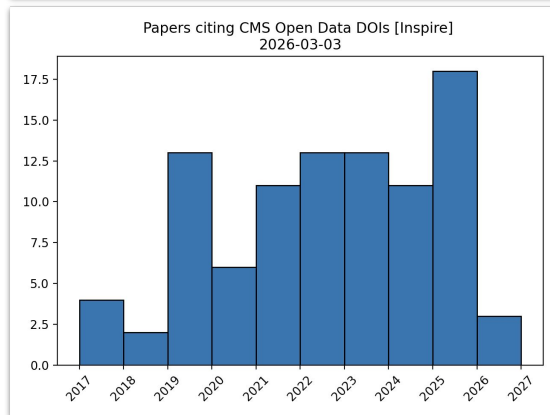
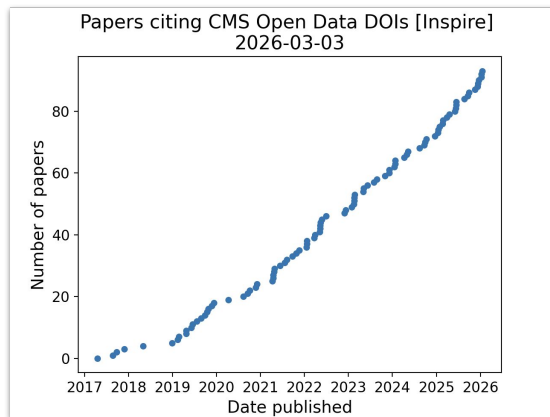
Thank you for your time!



# Backup slides

# CMS Open Data: Research

- CMS Open Data has found use out in the high-energy physics community, producing published research results: see for example the [citations of CMS Open Data DOIs in Inspire](#)
- CMS and CERN open data team authors are excluded in the plots at right



## Analyzing $N$ -Point Energy Correlators inside Jets with CMS Open Data

Patrick T. Komiske (MIT, Cambridge, CTP), Ian Mould (Yale U.), Jesse Thaler (MIT, Cambridge, CTP), Hua Xing Zhu (Hangzhou, Zhejiang U.)

Jan 19, 2022

13 pages

Published in: *Phys.Rev.Lett.* 130 (2023) 5, 051901

Published: Feb 1, 2023

e-Print: 2201.07800 [hep-ph]

DOI: 10.1103/PhysRevLett.130.051901 (publication)

Report number: MIT-CTP 5389

View in: [ADS Abstract Service](#)

[pdf](#) [cite](#) [claim](#)

[reference search](#) [159 citations](#)

## Conformal collider physics meets LHC data

Kyle Lee (LBL, Berkeley and LBNL, NSD), Bianka Meçaj (Yale U. and Yale U., Math. Dept.), Ian Mould (Yale U. and Yale U., Math. Dept.)

May 6, 2022

8 pages

Published in: *Phys.Rev.D* 111 (2025) 1, L011502

Published: Jan 1, 2025

e-Print: 2205.03414 [hep-ph]

DOI: 10.1103/PhysRevD.111.L011502 (publication)

View in: [ADS Abstract Service](#)

[pdf](#) [cite](#) [claim](#)

[reference search](#) [118 citations](#)

## Fast and Accurate Simulation of Particle Detectors Using Generative Adversarial Networks

Pasquale Musella (ETH, Zurich (main)), Francesco Pandolfi (INFN, Rome)

May 2, 2018

8 pages

Published in: *Comput.Softw.Big Sci.* 2 (2018) 1, 8

Published: Nov 2, 2018

e-Print: 1805.00850 [hep-ex]

DOI: 10.1007/s41781-018-0015-y

View in: [ADS Abstract Service](#)

[pdf](#) [cite](#) [claim](#)

[reference search](#) [104 citations](#)

# CERN 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

# Education and Outreach

- Education and outreach was the first and is the most enduring use-case for CMS open data
- Open data from CMS have formed the raw material of the CMS masterclasses since the beginning of the [International Masterclasses](#) (in collaboration with QuarkNet); the masterclasses which just finished used collision data from 2016
- There are over 200 “derived” (i.e. Level 2) datasets available on the CODP for use in education and outreach
- [Example analysis code](#) is also available (e.g. 4-lepton analysis)
- There are also other great resources like the [Particle Physics Playground](#)

