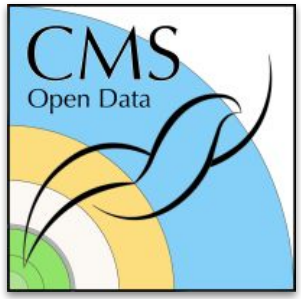
CMS Open Data

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

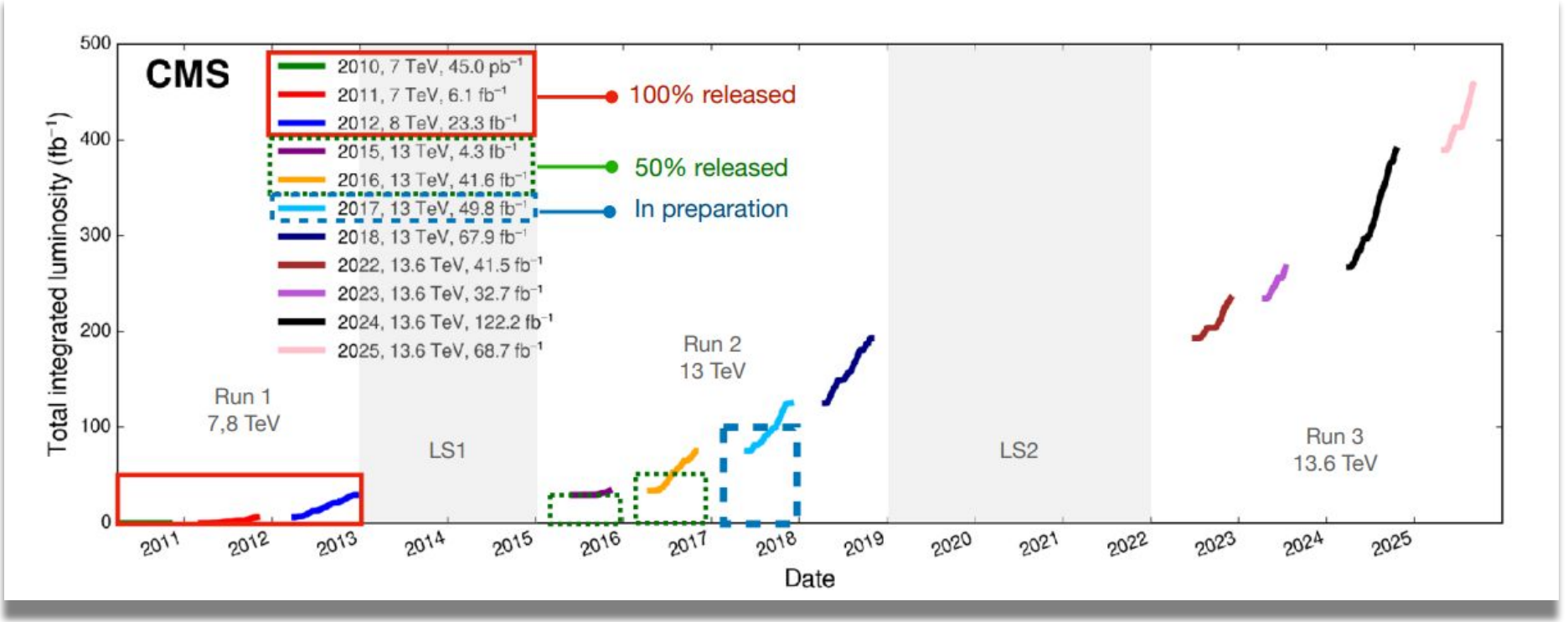
CMS Open Data Policy

Data releases since 2014

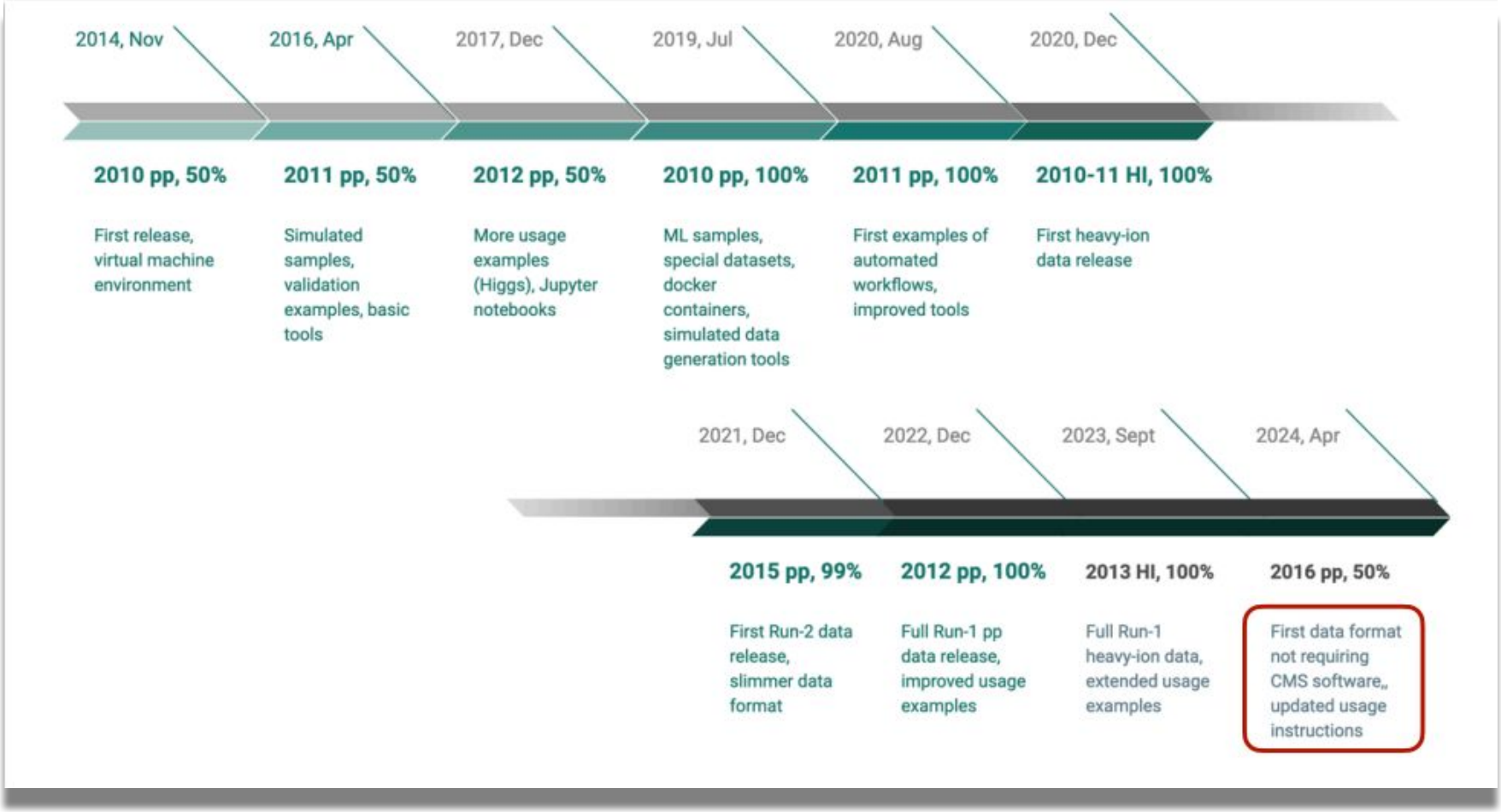
CMS data preservation, re-use, and open access policy
DOI:10.7483/ OPENDATA.CMS.1BNU.8V1W

- 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
  - Essentially releasing into the public domain
- Principles aligned with the CERN open data policy for the LHC experiments

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



CMS Open Data Releases - *timeline*



CMS Open Data Releases - *CERN Open Data Portal*

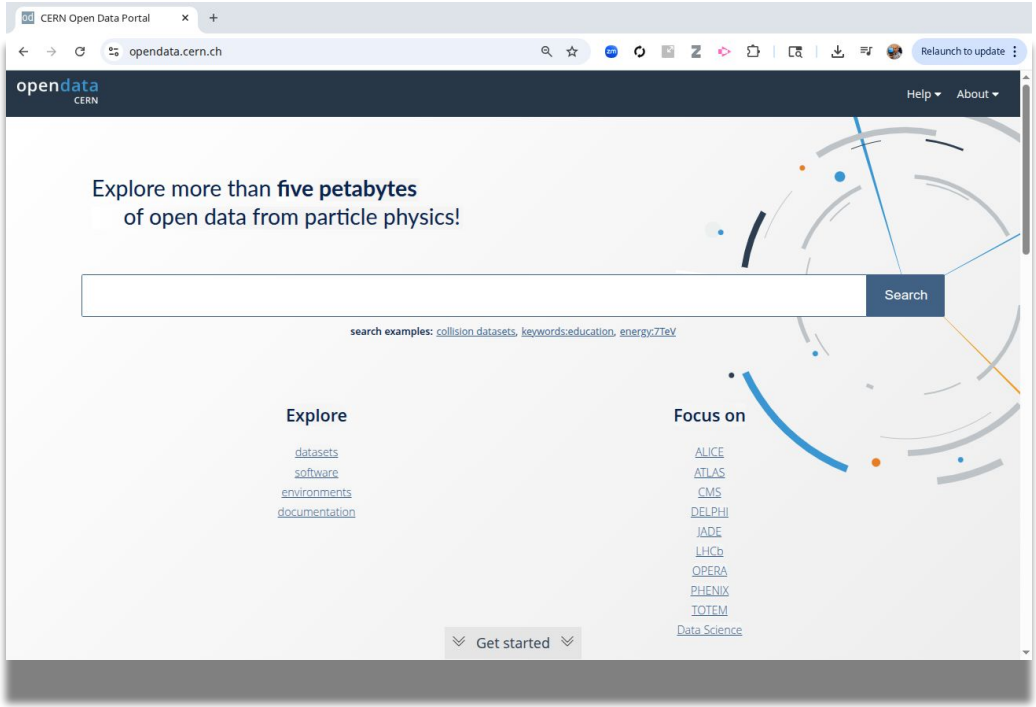
CMS data are available via the [CERN Open Data Portal]

Datasets are

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

Lots of data!

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

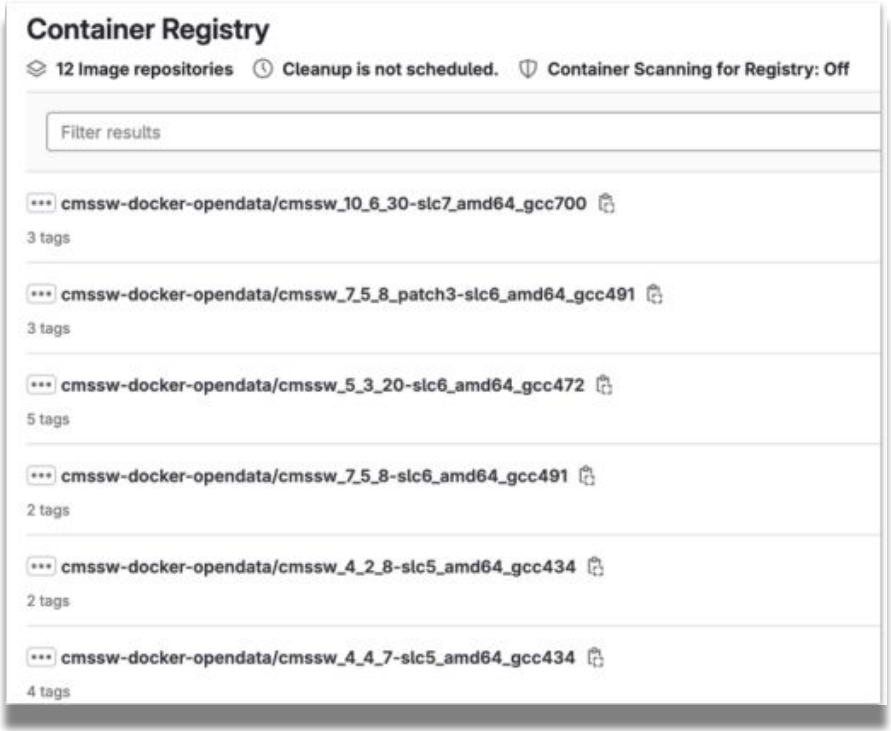


CMS Open Data Releases - *content*

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)
Reconstructed data	~3000
AOD	~500
miniAOD	~50
NanoAOD	~2

Data formats - *PFNanoAOD*

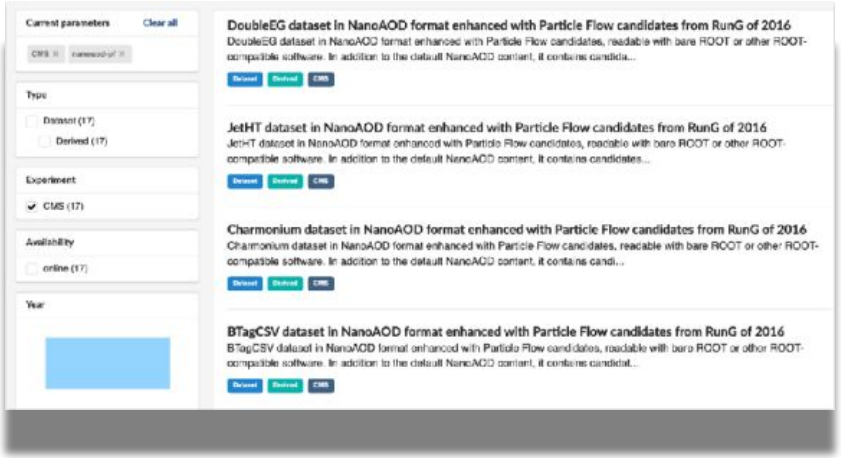
While the NanoAOD format has reduced information in comparison to MiniAOD (e.g. it does not have jet constituents) it is possible to created enhanced, extended versions of NanoAOD

PFNano is such 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



PFCands [back to top]

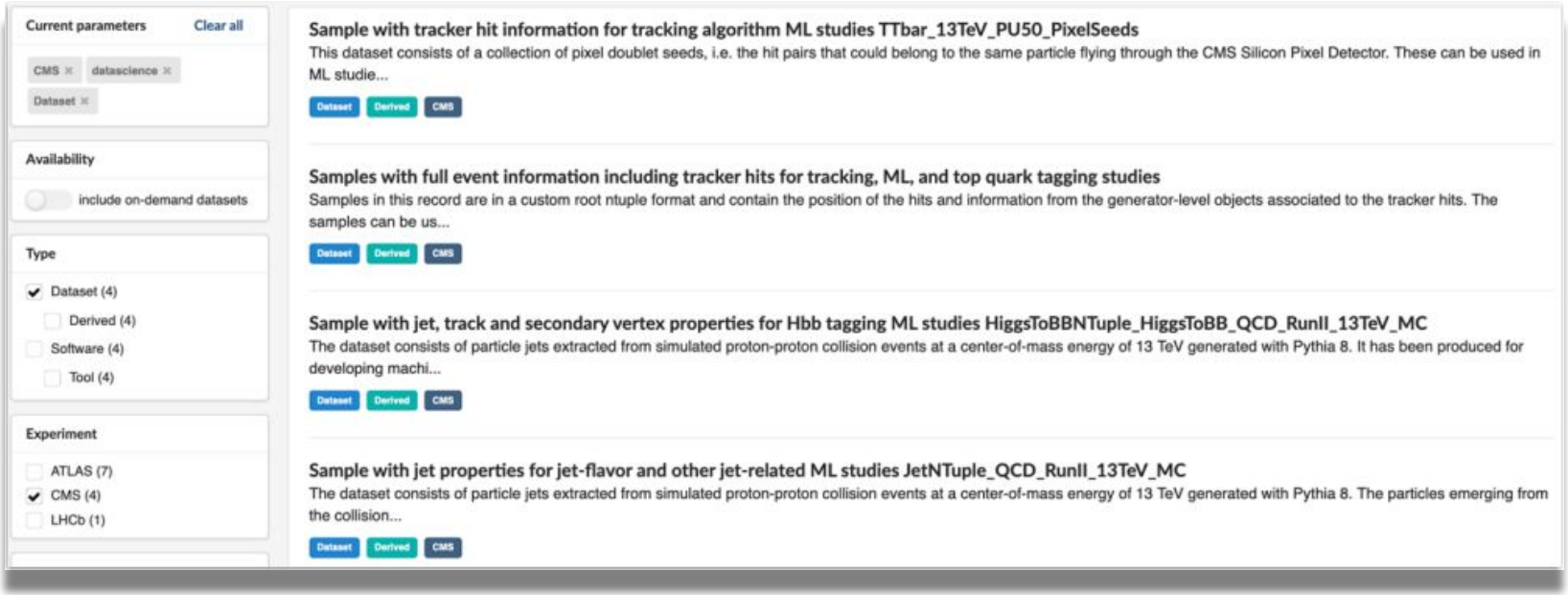
Object property	Type	Description
PFCands_charge	Int_t	electric charge
PFCands_d0	Float_t	pf d0
PFCands_d0Err	Float_t	pf d0 err
PFCands_dz	Float_t	pf dz
PFCands_dzErr	Float_t	pf dz err
PFCands_eta	Float_t	eta
PFCands_lostInnerHits	Int_t	lost inner hits
PFCands_mass	Float_t	mass
PFCands_pdgId	Int_t	PDG 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



Datasets - *storage and access going forward*

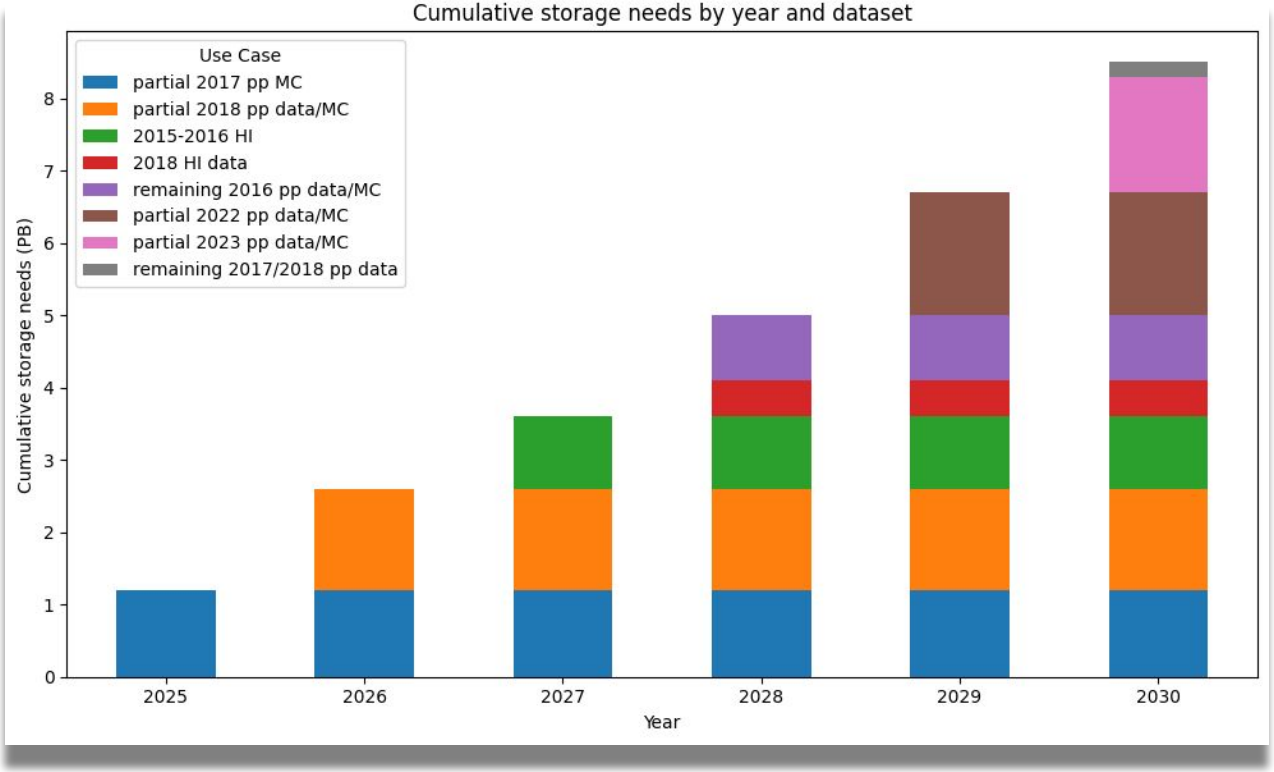
Planned CMS releases going forward *(on top of current datasets)*

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*



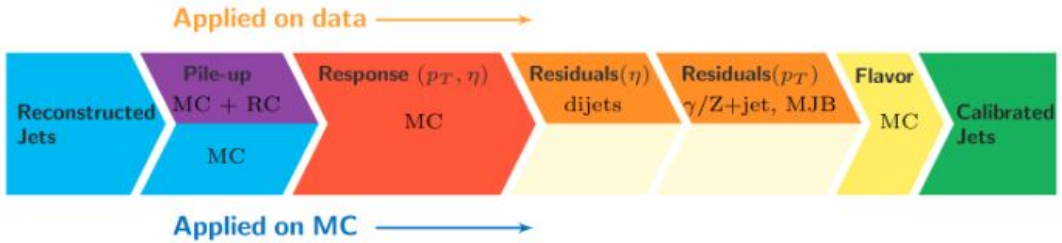
Resources - *CMS Open Data Guide*

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
- CMSSW
  - Overview
  - Data Model
  - Analyzers
  - Configuration
  - Conditions Data
- Analysis
  - Data and Simulation
  - Selection

Resources - *workshops*

[CMS Open Data Guide]

Aimed at theorists/phenomenolgists

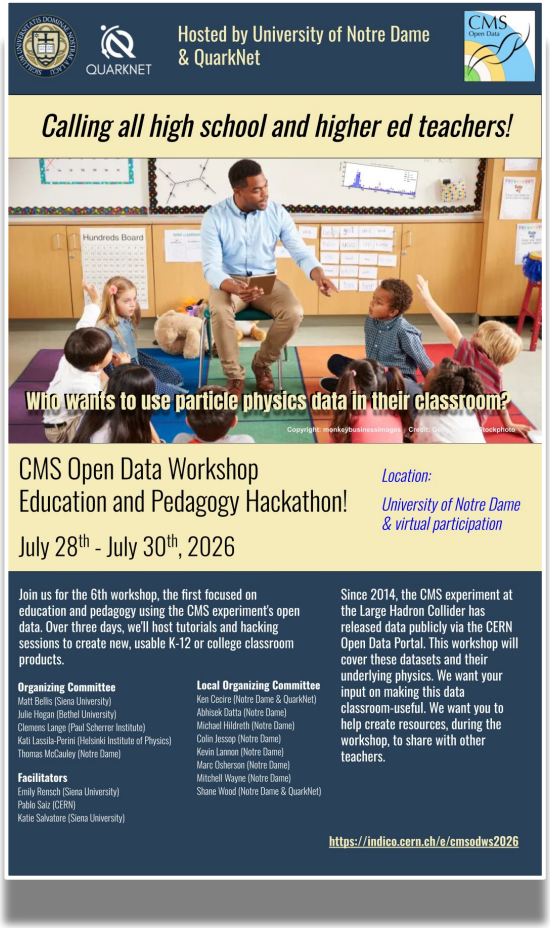
Attracted experimentalists and some teachers

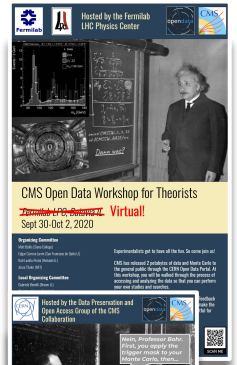
- 2020 workshop
- 2021 workshop
- 2022 workshop
- 2023 workshop
- 2024 WHEPP India
- 2024 workshop

Paused in 2025

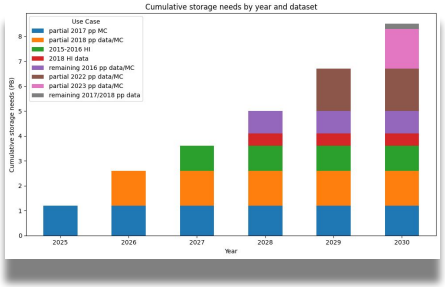
Next one - July 28-July 30 - Notre Dame

Pedagogical focus





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

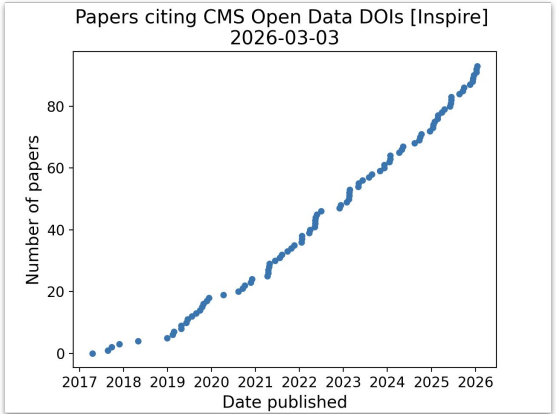
We preserved it because…who knows what's coming next?

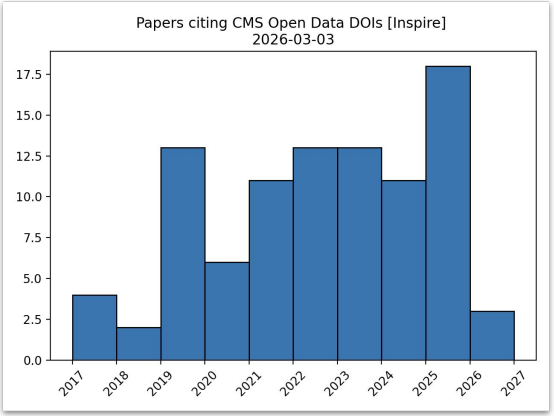
Thank you for your time!

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 Moult (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 Moult (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