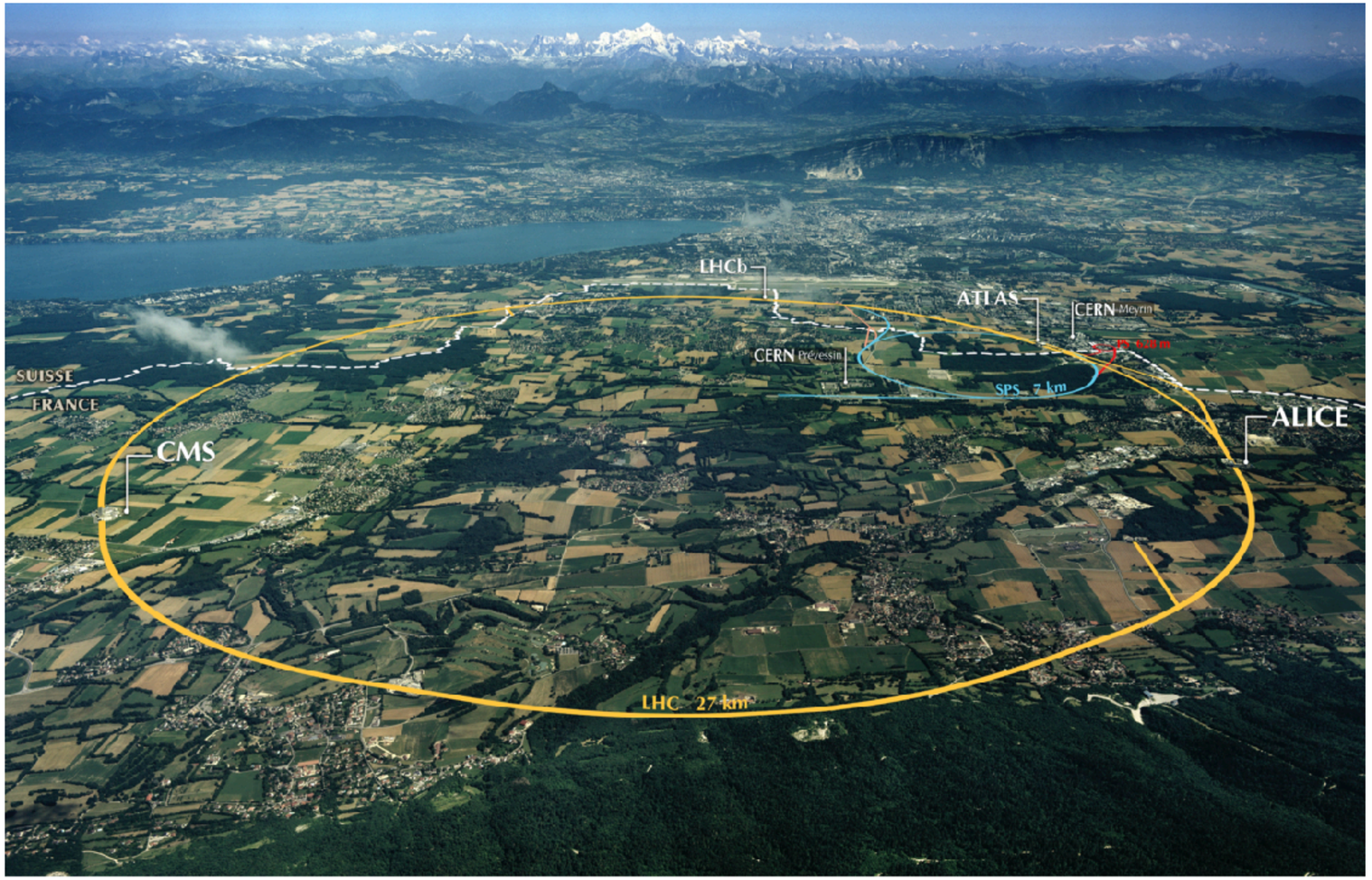
Open data from the CMS experiment at the LHC

Tom McCauley
for the CMS Collaboration
University of Notre Dame, USA
thomas.mccauley@cern.ch

IOP Open Data Workshop, 14 May 2026, University of Sussex

CERN, the LHC, and CMS

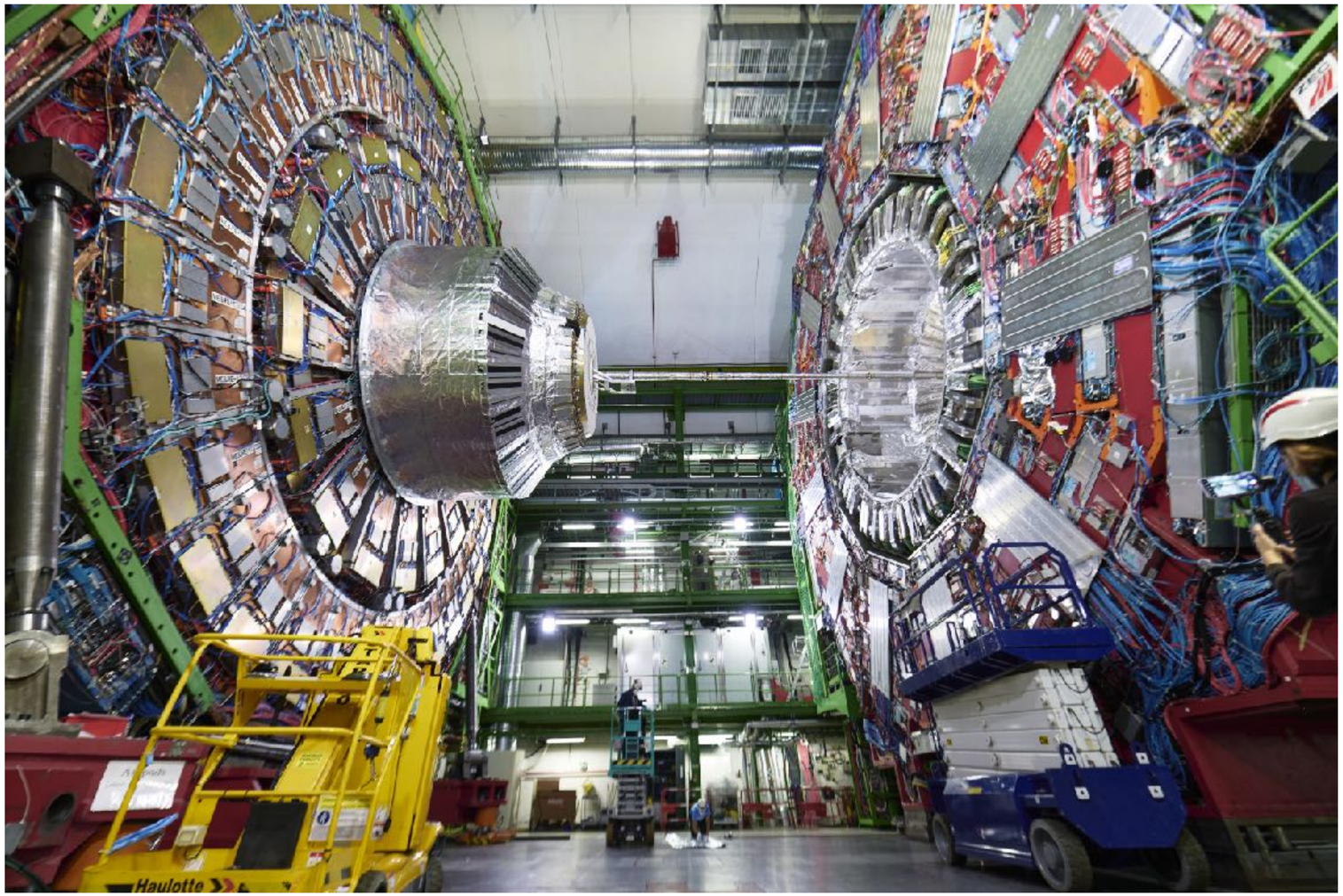
- Since 2010 the Large Hadron Collider (LHC) at CERN has been delivering stable proton-proton collisions to the 4 large LHC experiments at centre-of-mass energies from 7-13.6 TeV in addition to heavy-ion collisions.
- CMS (Compact Muon Solenoid) is a general-purpose detector which is used to conduct a broad physics program of Higgs, heavy-flavour, top quark, electroweak, Standard Model, and heavy-ion physics as well as searches for physics beyond the SM



CERN, the LHC, and CMS

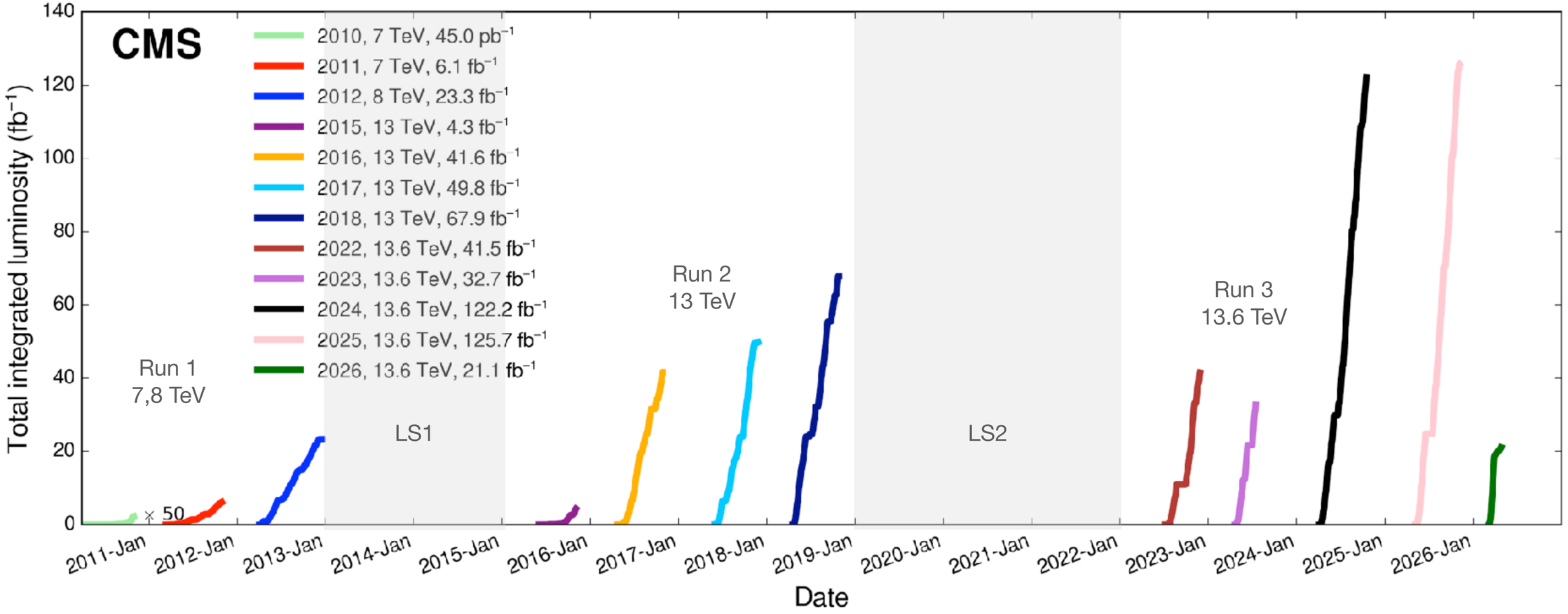
- Since 2010 the Large Hadron Collider (LHC) at CERN has been delivering stable proton-proton collisions to the 4 large LHC experiments at centre-of-mass energies from 7-13.6 TeV in addition to heavy-ion collisions.

- CMS (Compact Muon Solenoid) is a general-purpose detector which is used to conduct a broad physics program of Higgs, heavy-flavour, top quark, electroweak, Standard Model, and heavy-ion physics as well as searches for physics beyond the SM.



LHC luminosity

Run eras and data delivered



Why Open Data?
CMS Open Data Policy

“CMS data are unique and are the result of vast and long-term moral, human and financial investment by the international community. There is unique scientific opportunity in re-using these data, at different levels of abstraction and at different points in time . This opportunity calls for our collective responsibility, and poses unprecedented challenges as no data sample of this complexity and value has ever been preserved or made available for later re-use.

The CMS collaboration is committed to preserve its data, at different levels of complexity, and **to allow their re-use by a wide community including: collaboration members long after the data are taken, experimental and theoretical HEP scientists who were not members of the collaboration, educational and outreach initiatives, and citizen scientists in the general public**.

CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential.” - *CMS Data Preservation and Open Access (DPOA) Policy (first drafted and adopted in 2012 and updated in 2020)* DOI:10.7483/OPENDATA.CMS.1BNU.8V1W

CMS Open Data Policy

- 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 waiver, essentially releasing into the public domain
- Principles aligned with the CERN open data policy for the LHC experiments
- There are also CMS policies (restrictions) on use of open data by CMS collaborators
- Recommendations for Best Practices for Data Preservation and Open Science in HEP

Data preservation, re-use, and open access are interrelated, as used data are preserved data.

The best time to prepare for data preservation and open data is before data are even taken; the second best time is now.

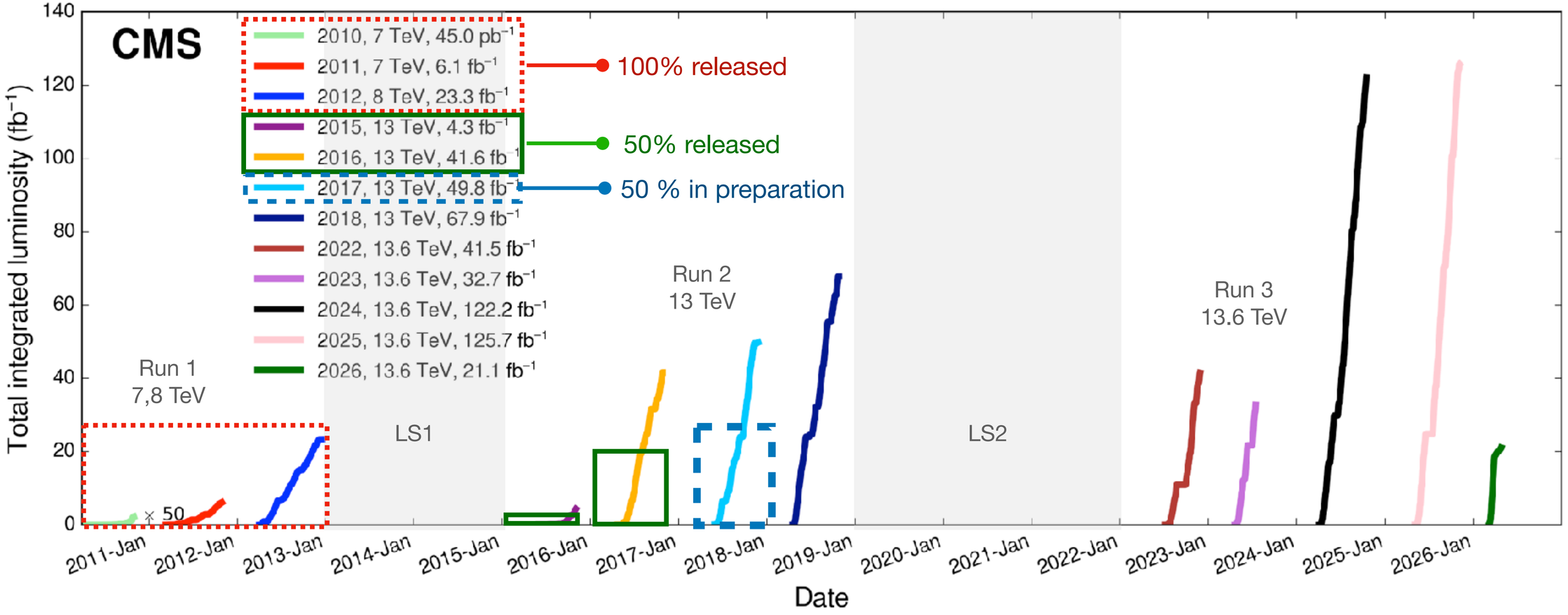
CERN Open Data Policy

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

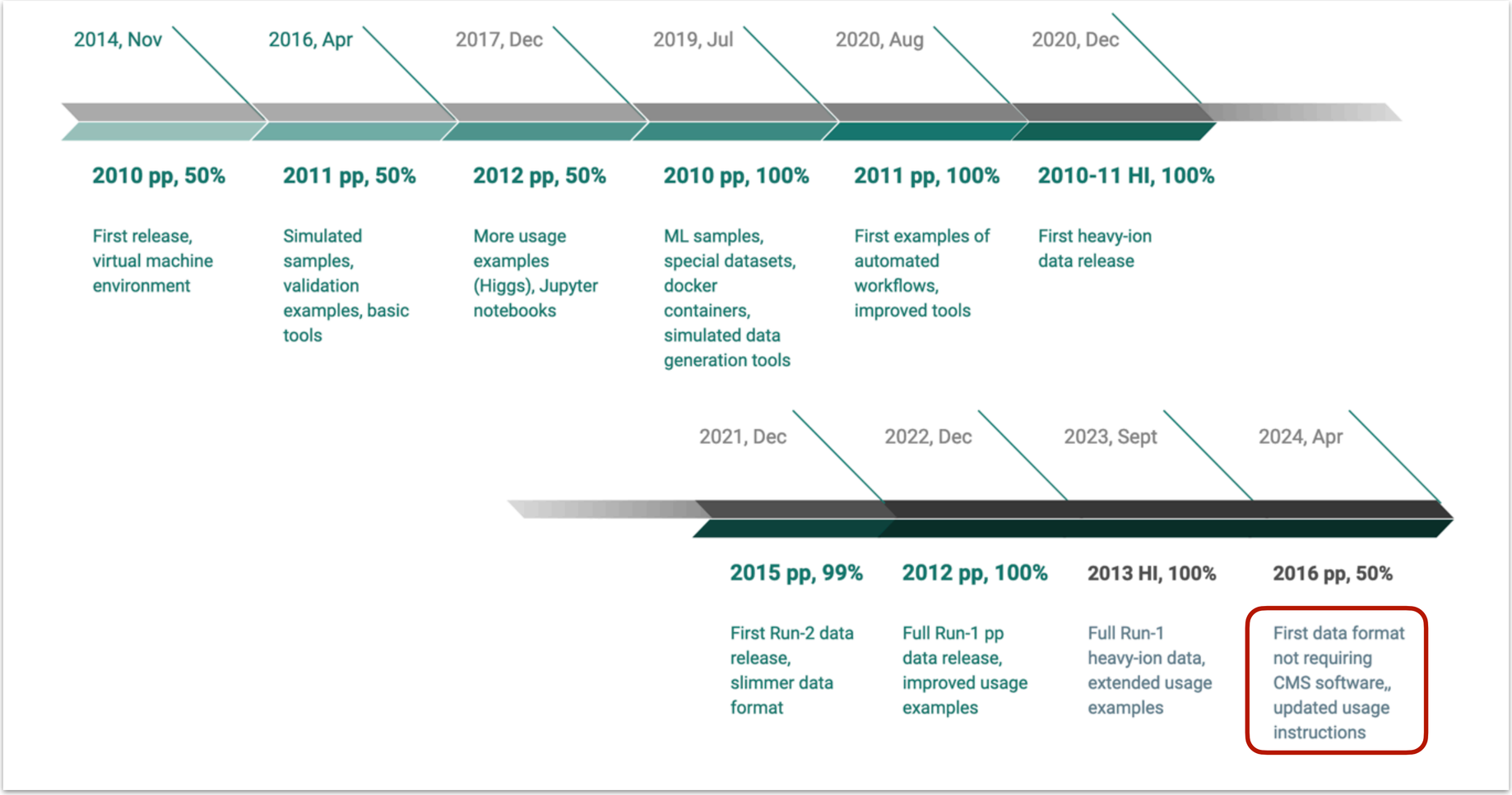
LHC luminosity

Run eras and data delivered



CMS Open Data releases

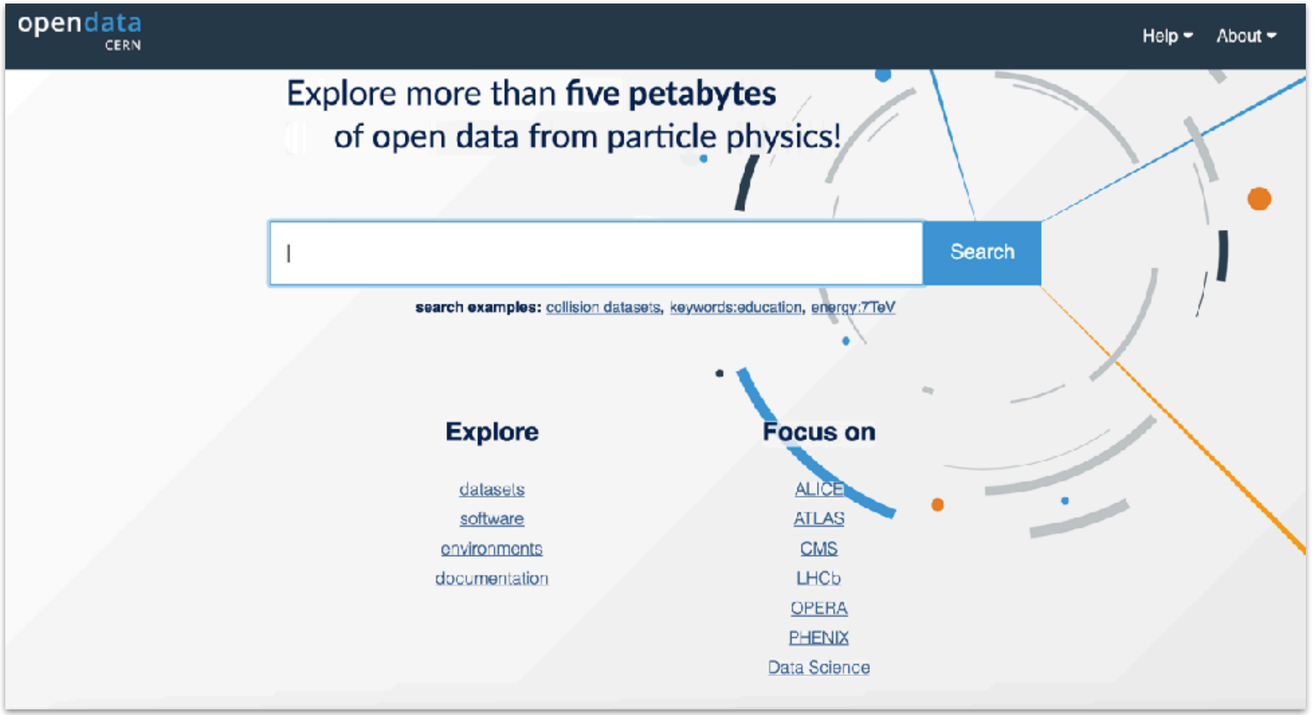
Timeline



CMS Open Data releases

CERN Open Data Portal

- CMS data are available via the CERN Open Data Portal
- Datasets are categorised, searchable, and citable (with a DOI attached to each dataset)
- There are **300+ collision and 50k+ simulated datasets** available, over 4 PB (disk and tape)
- Over 16 billion real collision events

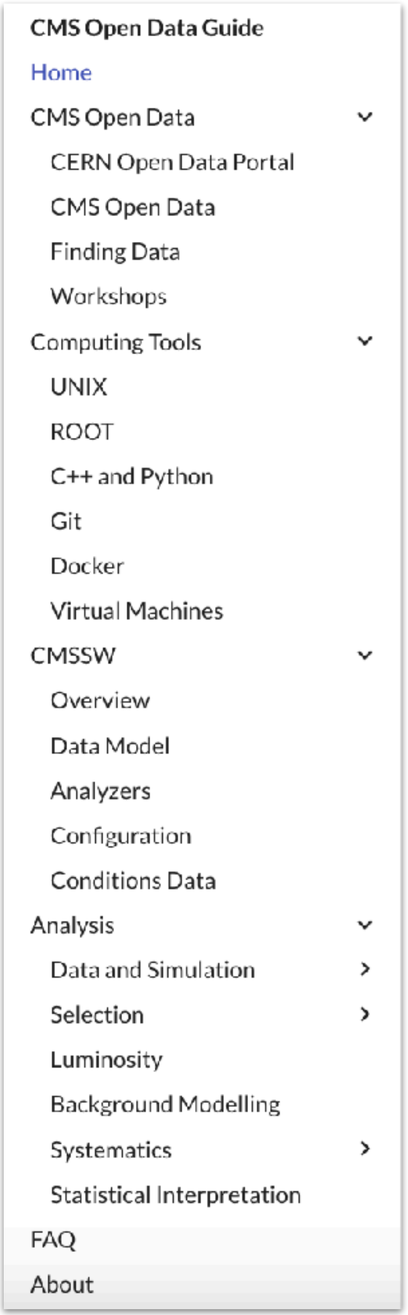


CMS Open Data releases

Content

Providing just the datasets isn't enough. If the data aren't usable, then what's the point? Therefore a data release also includes:

- Software environments via Docker containers and/or virtual machines
- Analysis software: CMS software, example analyses, validated run information, conditions database access, …
- Documentation (such as the CMS Open Data Guide)
- Continued support via *e.g.* a support forum



CMS Open Data usage

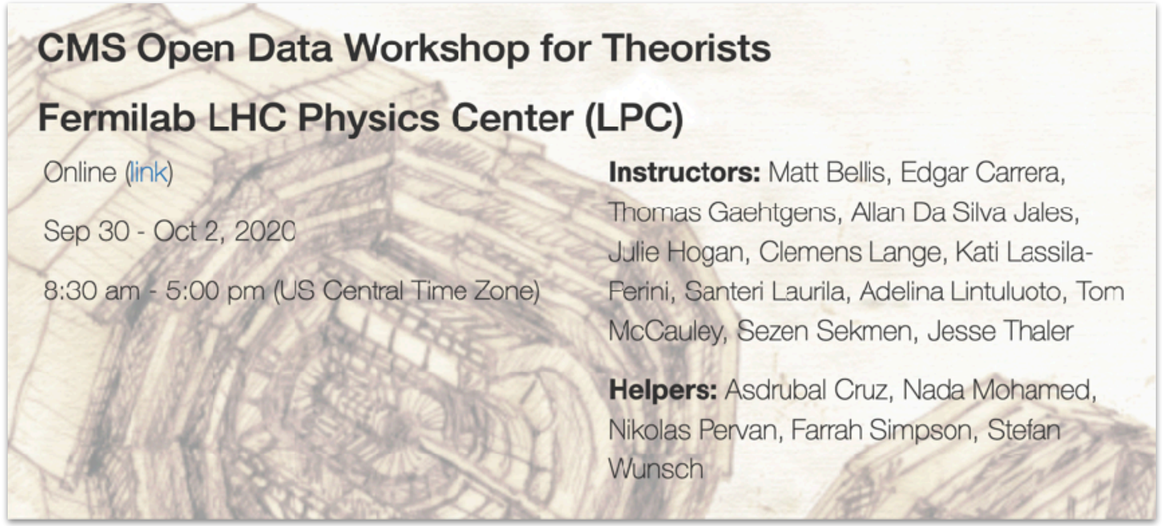
General challenges

- A certain amount of knowledge about physics, detectors, software, and data analysis is needed in order to use the open data “like a physicist”
- The datasets available are large and cover a broad range of physics, what data should one use and once one knows that, how does one find them?
- Some data formats are complex and require use and knowledge of CMS-specific software
- ***On the experiment end:*** person-power is needed to prepare the releases and to prepare and maintain the open data infrastructure like documentation, software environments, examples, etc.

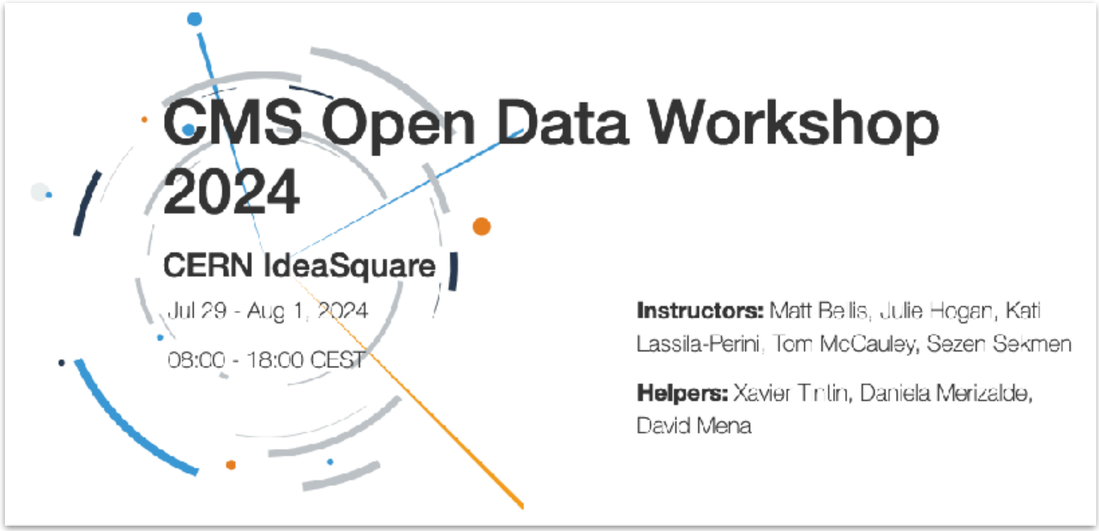
Resources

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 students, educators, theorists, phenomenologists, MLs, …**
- Open Data Workshops are in-person and hybrid
- Even after the workshops, the material prepared remain a persistent and invaluable resource for learning
- **The next workshop, 28-30 July at ND, will focus on educational uses, targeting teachers and lecturers**



[…]



CMS Open Data usage

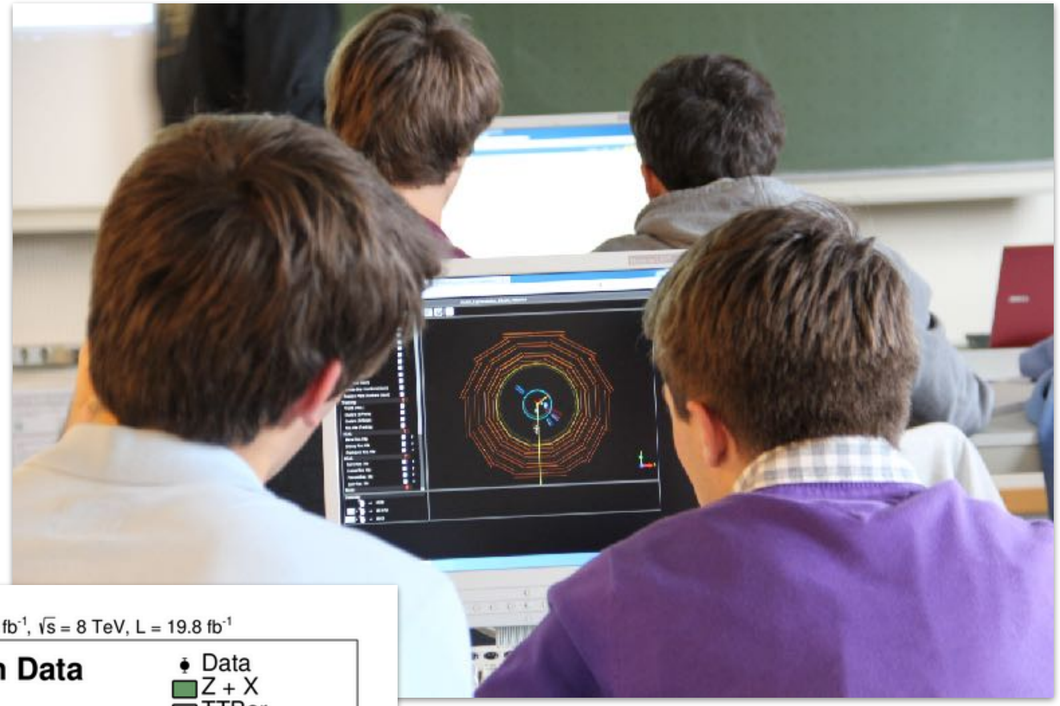
Education

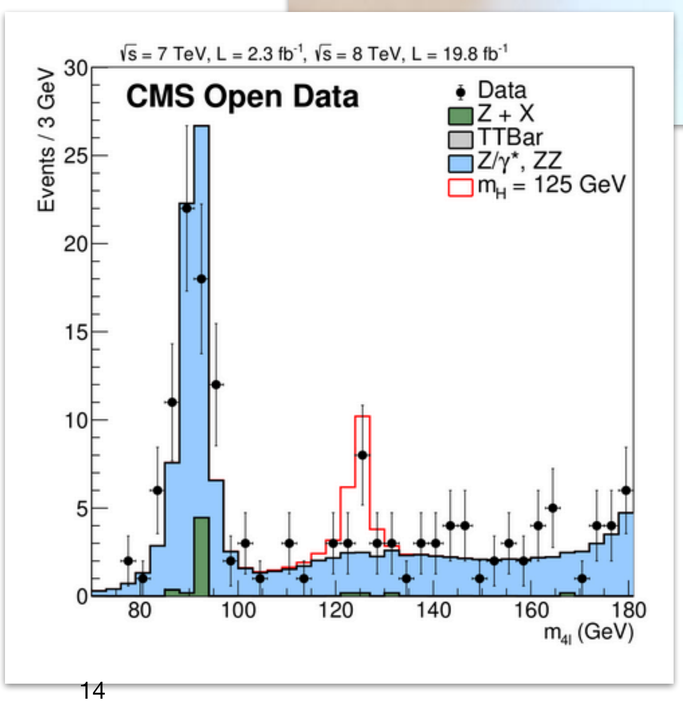
- Education (and outreach) was the first and is the most enduring use-case for CMS open data. In fact, **the successful use of small datasets released especially for education helped to encourage the collaboration to release more and to adopt an open data policy**.

- Open data from CMS have formed the raw material of the CMS masterclasses since the beginning of the International Masterclasses (in collaboration with QuarkNet). These data have been used by tens of thousands of students all over the world.

- 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, *i.e.* discovery of the Higgs)





Educational usage

How do (did) we get there?

- The bulk of CMS open data are research-level (data preservation is part of our mandate as well), *i.e.* at a level used by the physicists on the experiment
- Collaboration between educational experts and physicists to help bridge the gap of expertise and knowledge: the collaboration between QuarkNet (an NSF-funded physics teacher training program) and CMS has been crucial to success
- Access at different levels appropriate to the intended audience needs to be provided: *e.g.* example analyses in Jupyter notebooks, simplified datasets and formats, open access applications and tools such as event displays and learning portals
- As with everything, person-power is required: ***we need sustained collaboration between experiments, universities, and funding bodies to create shared teaching resources, maintain tools, and recognise the contribution of those who build and support open data infrastructure***

Resources

Summary

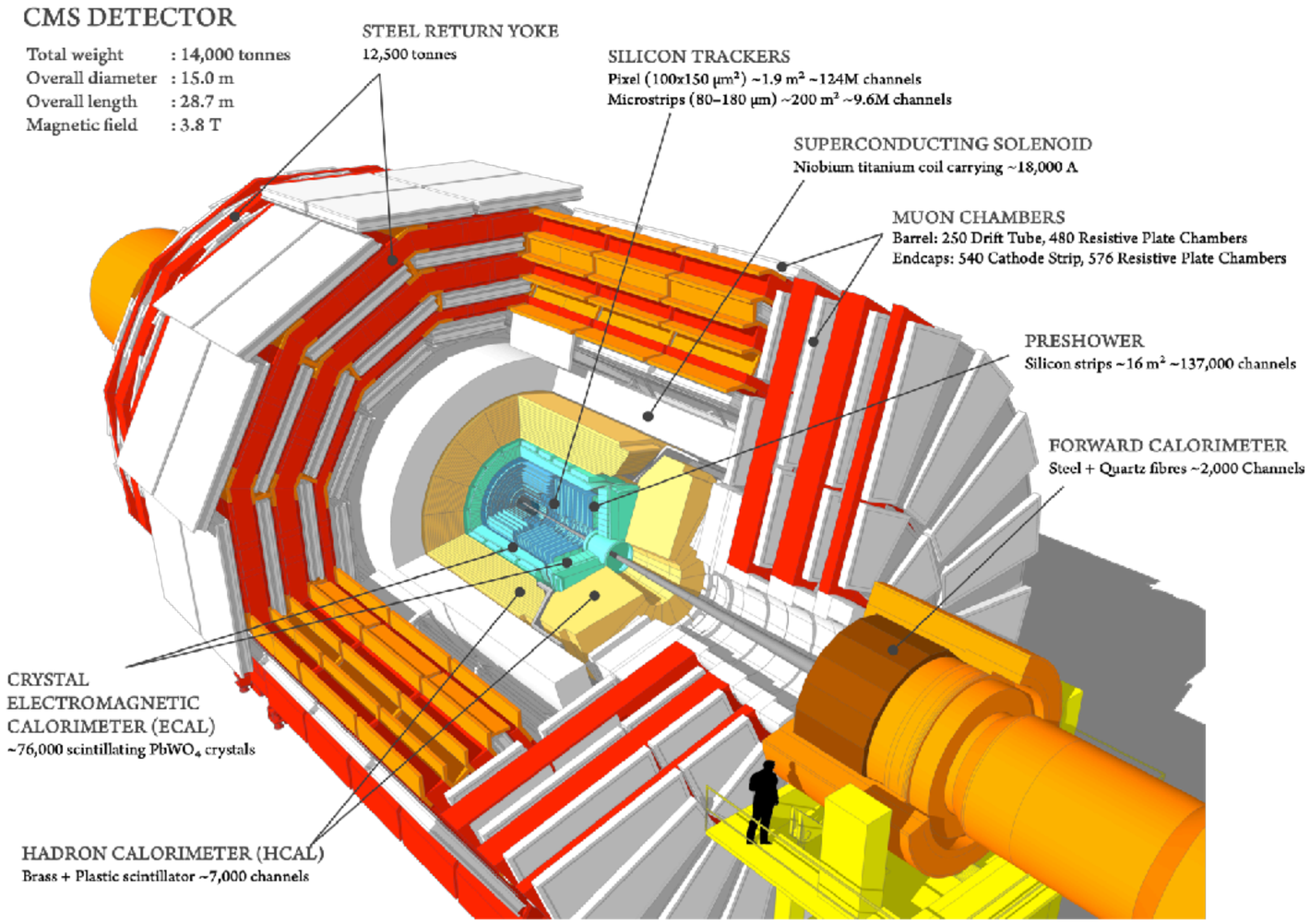
- Over 4 PB of collision data and simulation
- CERN Open Data Portal: https://opendata.cern.ch/
- CMS Open Data Guide: https://cms-opendata-guide.web.cern.ch/
- CMS Open Data Forum: https://opendata-forum.cern.ch/c/cms/6
- CMS Open Data Workshops: https://cms-opendata-guide.web.cern.ch/cmsOpenData/workshops/

Thank you

Backup

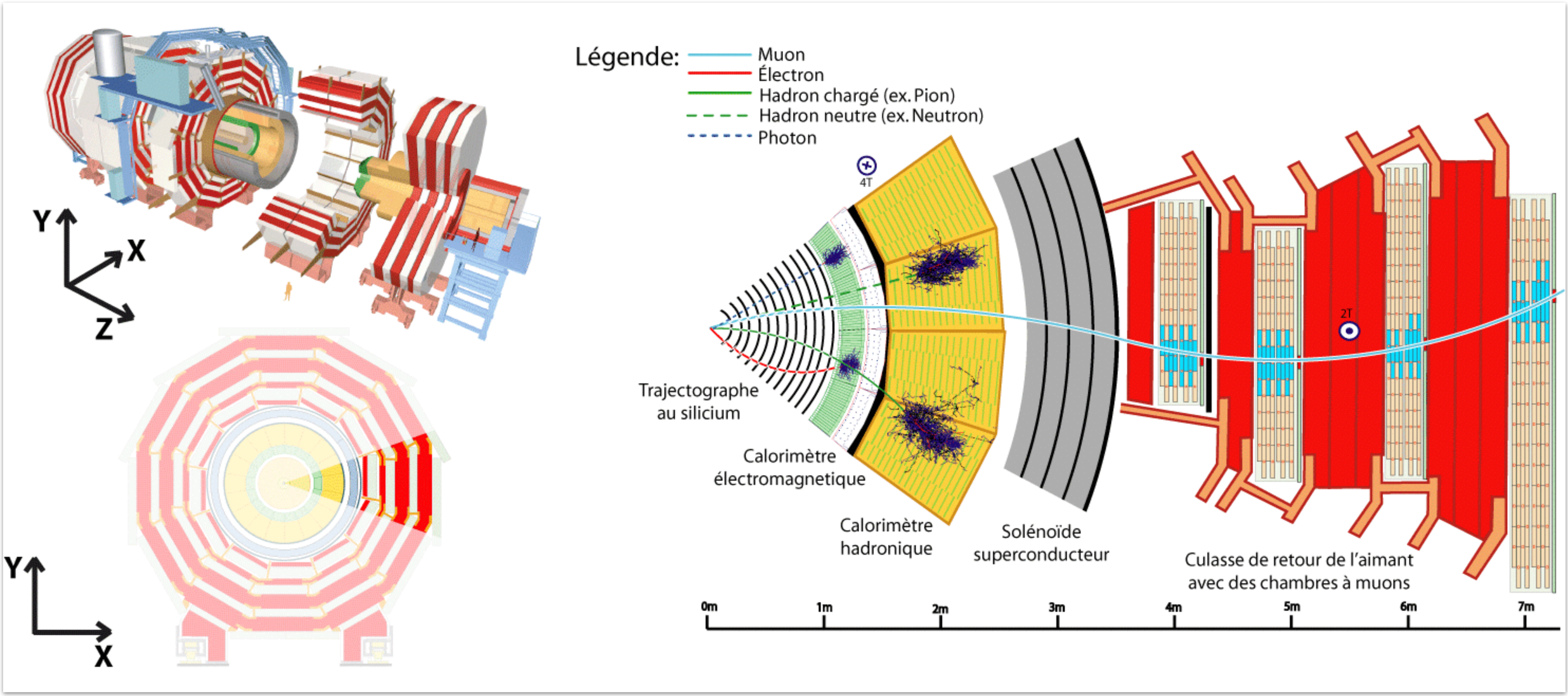
CMS detector

Requirements and features



- A high-quality central tracking system to give accurate momentum measurements
- A high-resolution electromagnetic calorimeter to detect and measure electrons and photons
- A “hermetic” hadron calorimeter, designed to entirely surround the collisions and prevent any particles from escaping undetected
- A high performance system to detect and measure muons
- A solenoidal magnet to provide a large magnetic field

CMS detector



CMS collaboration

CMS collaboration

- 246 institutes from 58 regions
- ~ 2300 authors
- Thousands of engineers, students, …

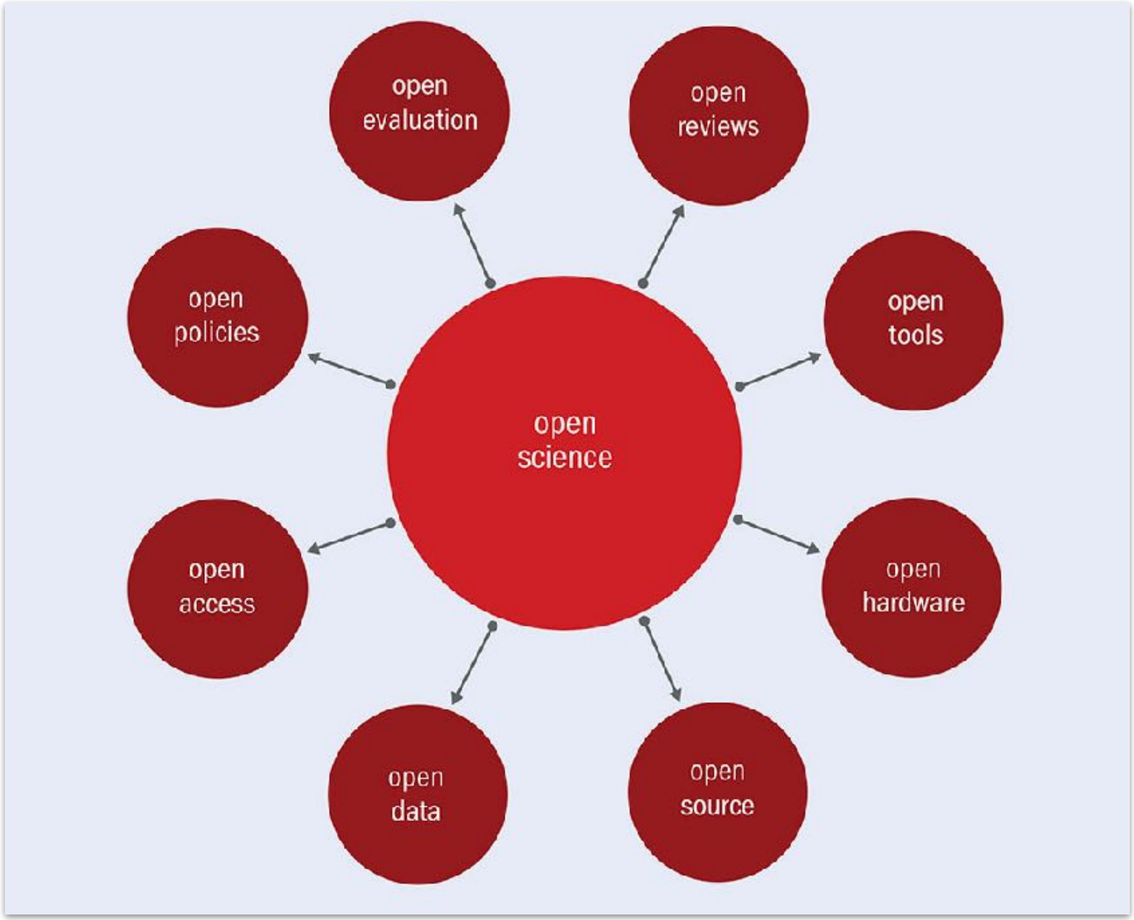
(As of end-of-summer 2025)

What is Open Data?

ChatGPT 5 >

What is open data?

Open data is data that anyone can access, use, and share freely—usually with minimal restrictions such as attribution. The core idea is that data should be available to the public in a usable format so that it can be analyzed, reused, and redistributed for purposes like research, innovation, education, and transparency.



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, and the possible use of python frameworks and tools such as Coffea, RDataFrame, uproot, awkward, …

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

Why Open Data?

But steady publication of LHC data has multiple benefits. First, it encourages prompt archiving, before collective memory fades and knowledge is lost. Second, other scientists can analyse the data while the LHC is still running, testing unconventional strategies and potentially leading to unexpected discoveries, new approaches and fruitful discussions. And third, as a by-product, these scientists can stress test the archiving methods; any deficiencies found are easier to fix now than later. In this way, public collider data can complement the overall LHC research effort. We, therefore, favour a slow but steady approach to full publication of the LHC experiments' data; it is in the best interest of particle physics.

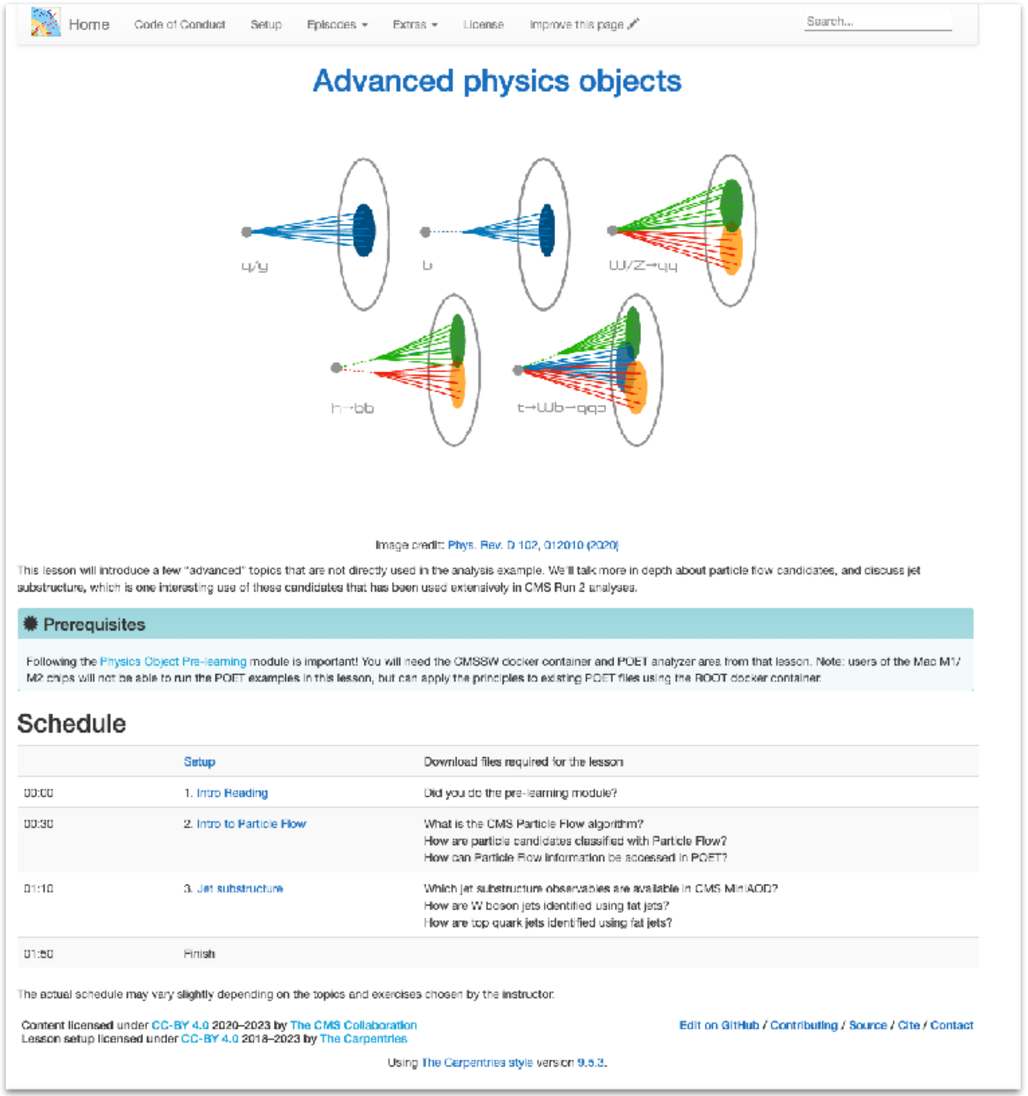
Matthew Strassler ✉ & Jesse Thaler ✉

Nature Physics **15**, 725 (2019) | Cite this article

Open Data Workshops

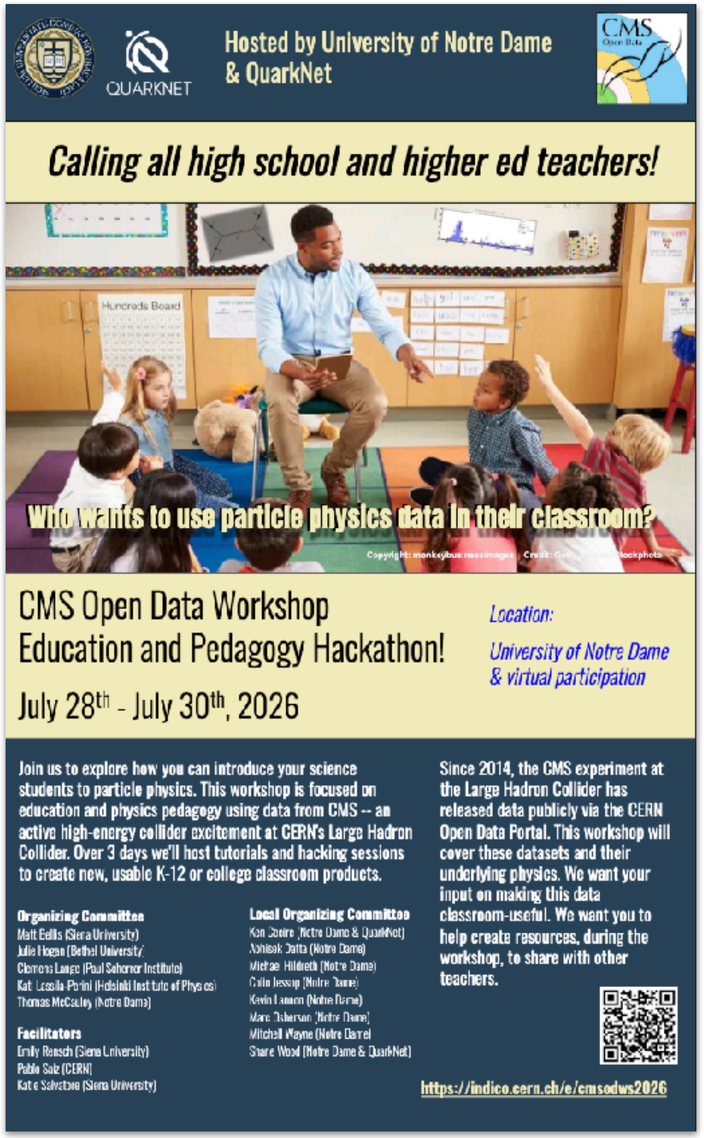
Structure and Schedule

- The workshops run over several days
- The pre-learning and pre-exercises are pretty common across workshops (Unix, python, version control, containers, particle physics, CMS detector,...)
- We use the Software Carpentries framework for all workshops
- The advanced topics (e.g. cloud computing, statistical inference, specific analyses,...) may change over workshops
- “Hackathon” sessions are sometimes included
- Programming languages include C++ and python but the direction of travel in more recent workshops is towards python-only



Next workshop

- The next open data workshop: 28-30 July at University of Notre Dame
- The focus of this workshop will be educational use-cases: hopefully empowering secondary school teachers and university lecturers to use CMS open data in their lessons
- Format will include user-initiated hackathon sessions as well as tutorials



Level 3 collision data and MC releases

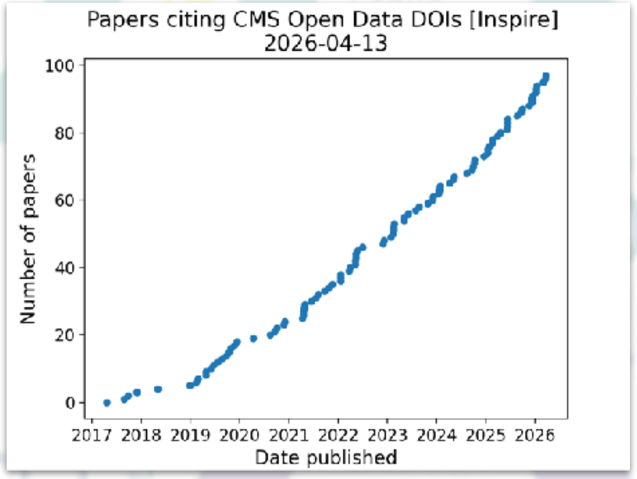
- 2010 p-p collision data at 7 TeV
- 2010 Pb-Pb collision data at 2.76 TeV
- 2011 p-p collision data at 7 TeV + MC
- 2011 p-p collision data at 2.76 TeV and p-Pb collision data at 5.02 TeV
- 2012 p-p collision data at 8 TeV + MC
- 2013 p-p collision data at 2.76 TeV and p-Pb collision data at 5.02 TeV + 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

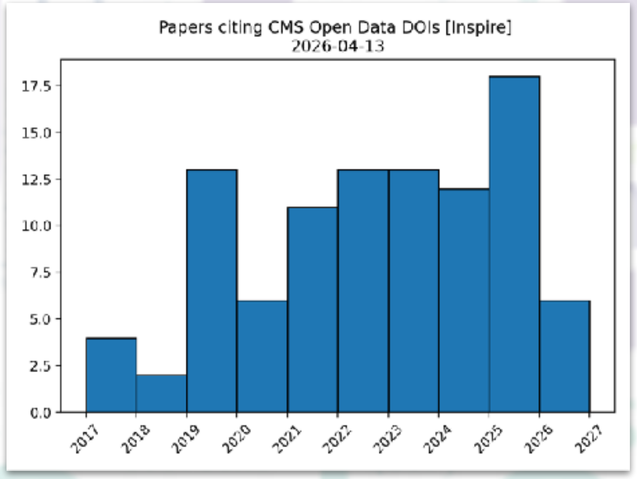
CMS Open Data usage

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]

- Topics: searches, jet/QCD studies, machine learning





Papers with CMS and CERN open data authors are excluded

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), Blanka 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