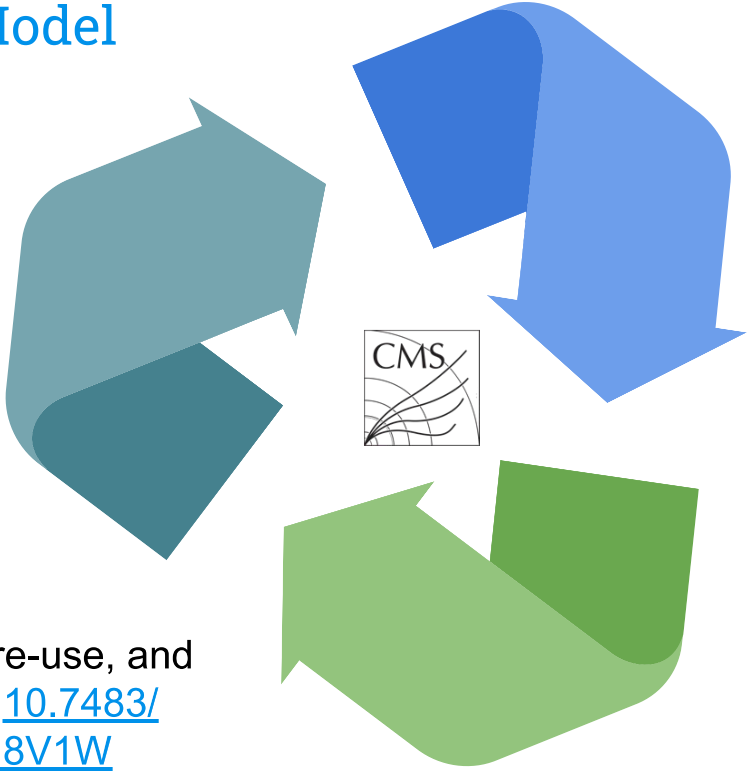
New Opportunities with CMS Open Data

Julie Hogan (Bethel U), Tom McCauley (Notre Dame U)

DPHEP Workshop

Oct 3, 2024

CMS Open Data Model



Data:

- collision data
- simulations
- additional data for analysis

Tools:

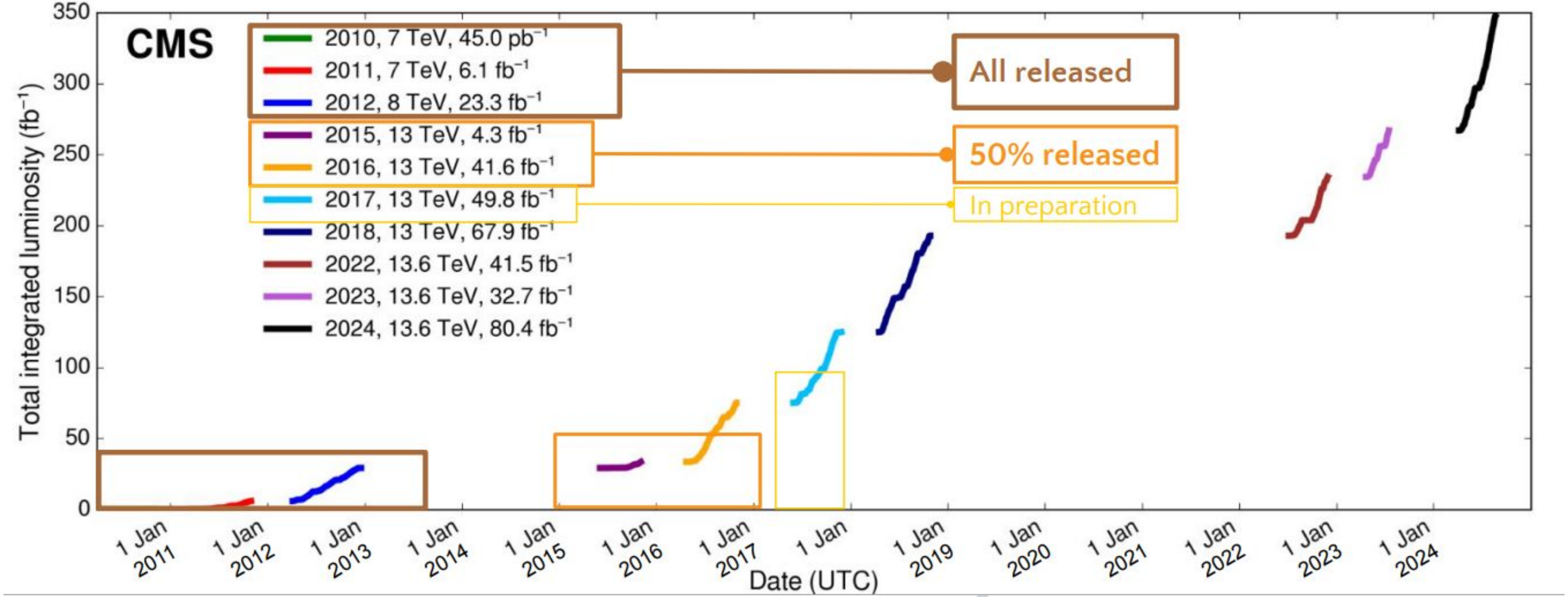
- software
- environments
- interfaces

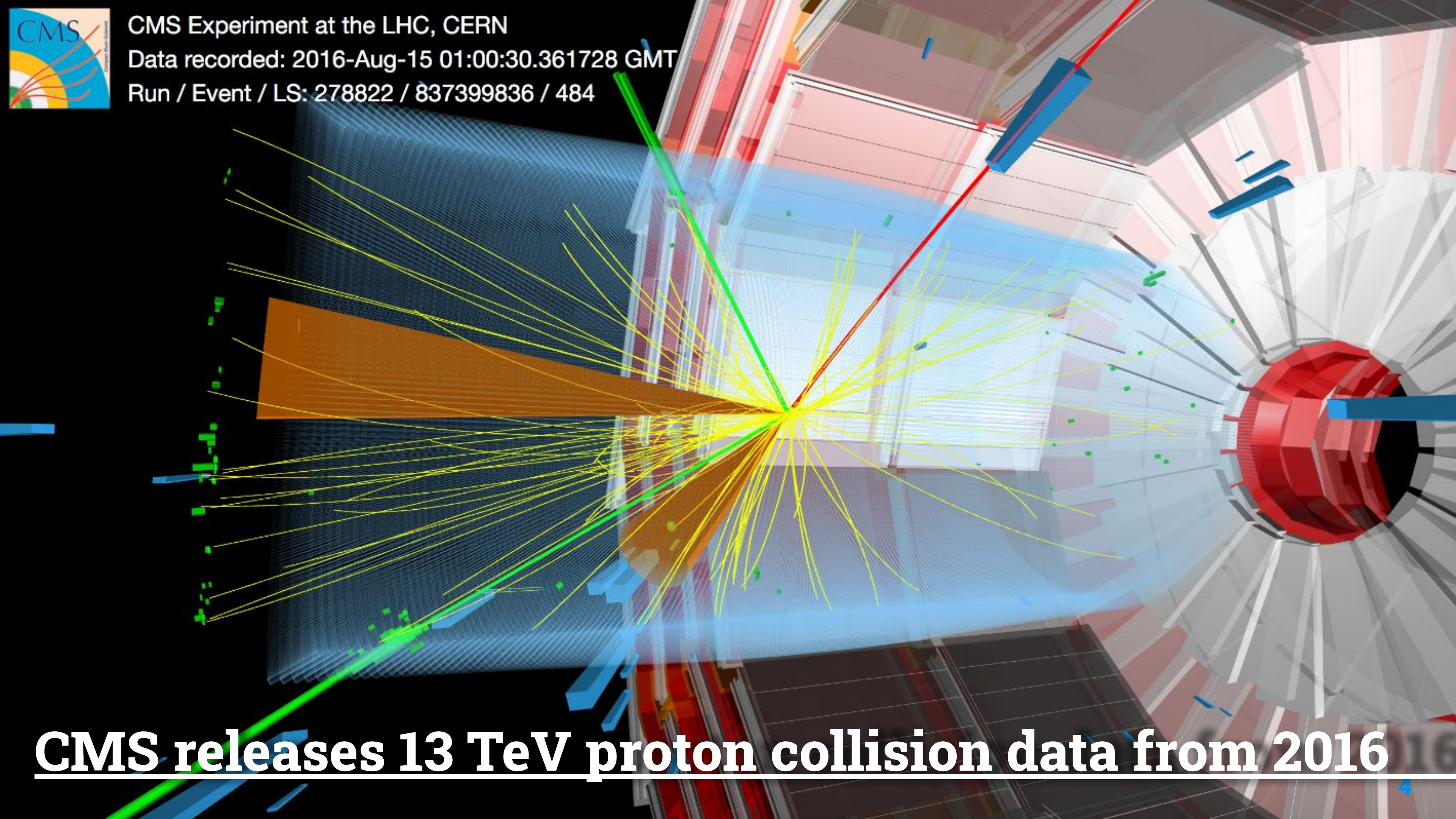
Knowledge:

- instructions
- actionable examples
- understanding of experimental data

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)

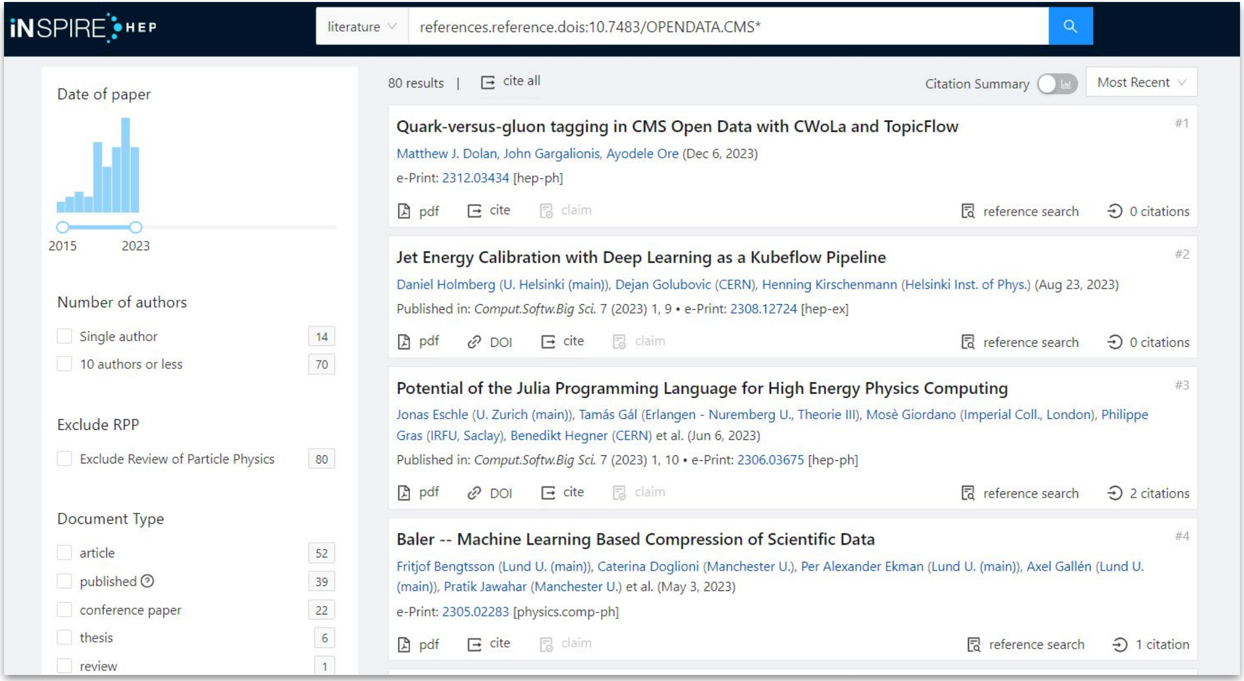
CMS Open Data releases




CMS
CMS Experiment at the LHC, CERN
Data recorded: 2016-Aug-15 01:00:30.361728 GMT
Run / Event / LS: 278822 / 837399836 / 484

CMS releases 13 TeV proton collision data from 2016

CMS Open data in use



references.reference.dois:10.7483/OPENDATA.CMS*

Date of paper

2015 2023

Number of authors

☐ Single author	14
☐ 10 authors or less	70

Exclude RPP

☐ Exclude Review of Particle Physics	80

Document Type

☐ article	52
☐ published ⓘ	39
☐ conference paper	22
☐ thesis	6
☐ review	1

80 results | cite all

Citation Summary Most Recent

Quark-versus-gluon tagging in CMS Open Data with CWoLa and TopicFlow #1

Matthew J. Dolan, John Gargalionis, Ayodele Ore (Dec 6, 2023)

e-Print: 2312.03434 [hep-ph]

pdf cite claim reference search 0 citations

Jet Energy Calibration with Deep Learning as a Kubeflow Pipeline #2

Daniel Holmberg (U. Helsinki (main)), Dejan Golubovic (CERN), Henning Kirschenmann (Helsinki Inst. of Phys.) (Aug 23, 2023)

Published in: *Comput.Softw.Big Sci.* 7 (2023) 1, 9 • e-Print: 2308.12724 [hep-ex]

pdf DOI cite claim reference search 0 citations

Potential of the Julia Programming Language for High Energy Physics Computing #3

Jonas Eschle (U. Zurich (main)), Tamás Gál (Erlangen - Nuremberg U., Theorie III), Mosè Giordano (Imperial Coll., London), Philippe Gras (IRFU, Saclay), Benedikt Hegner (CERN) et al. (Jun 6, 2023)

Published in: *Comput.Softw.Big Sci.* 7 (2023) 1, 10 • e-Print: 2306.03675 [hep-ph]

pdf DOI cite claim reference search 2 citations

Baler -- Machine Learning Based Compression of Scientific Data #4

Fritjof Bengtsson (Lund U. (main)), Caterina Doglioni (Manchester U.), Per Alexander Ekman (Lund U. (main)), Axel Gallén (Lund U. (main)), Pratik Jawahar (Manchester U.) et al. (May 3, 2023)

e-Print: 2305.02283 [physics.comp-ph]

pdf cite claim reference search 1 citation

Records have unique DOIs: 10.7483/OPENDATA.CMS*

2016 data release features

Collision Data

- 16 fb^{-1} of 13 TeV proton collision data from 2016
- Ultra-Legacy processing!
- MiniAOD and NanoAOD data formats

Simulation

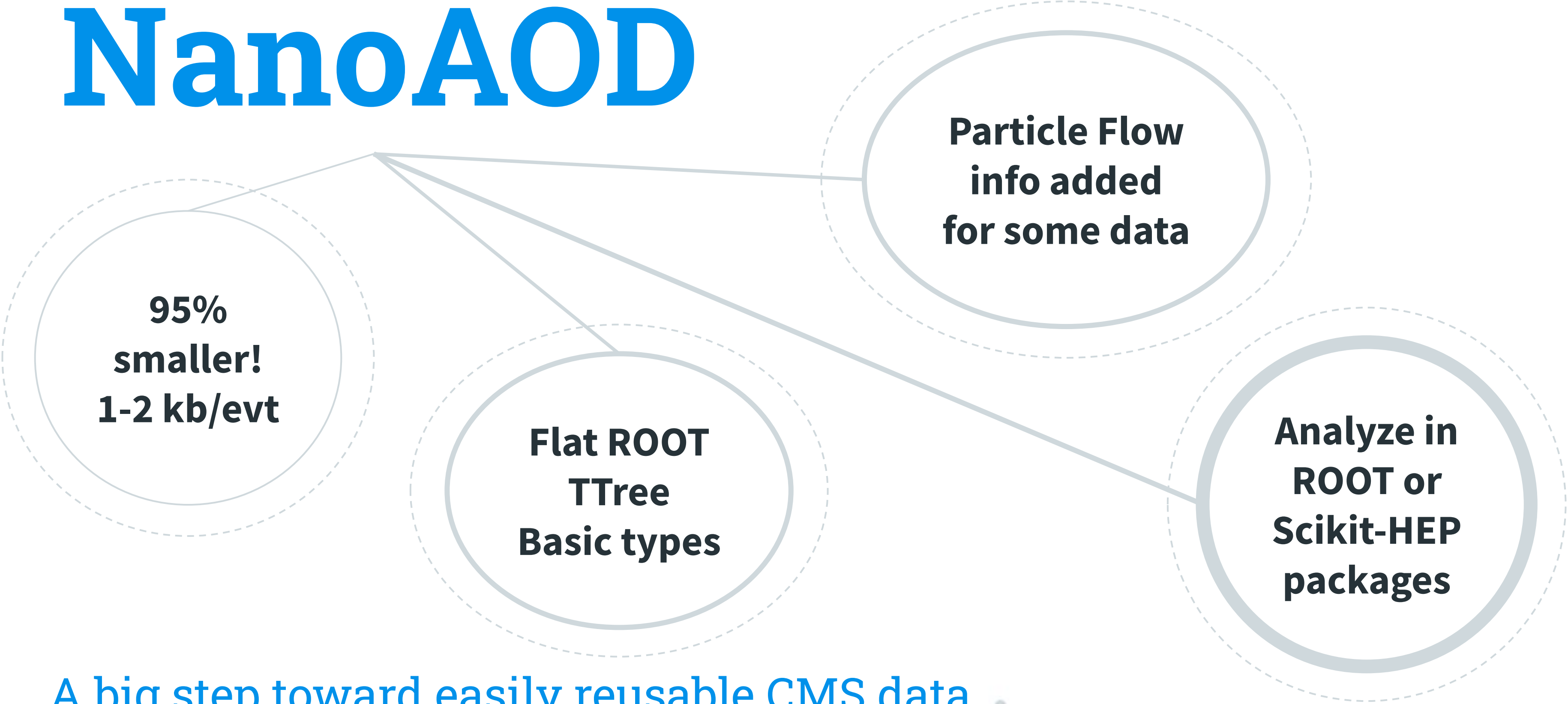
- Over 830 TB of simulation going to the portal
- Over 20,000 unique processes!
- MiniAOD and NanoAOD formats

Software

- Container & VM for CMSSW 10
- Containers for ROOT & python
- New guides
- New analysis tools

First significant luminosity from 13 TeV collisions

NanoAOD



95%
smaller!
1-2 kb/evt

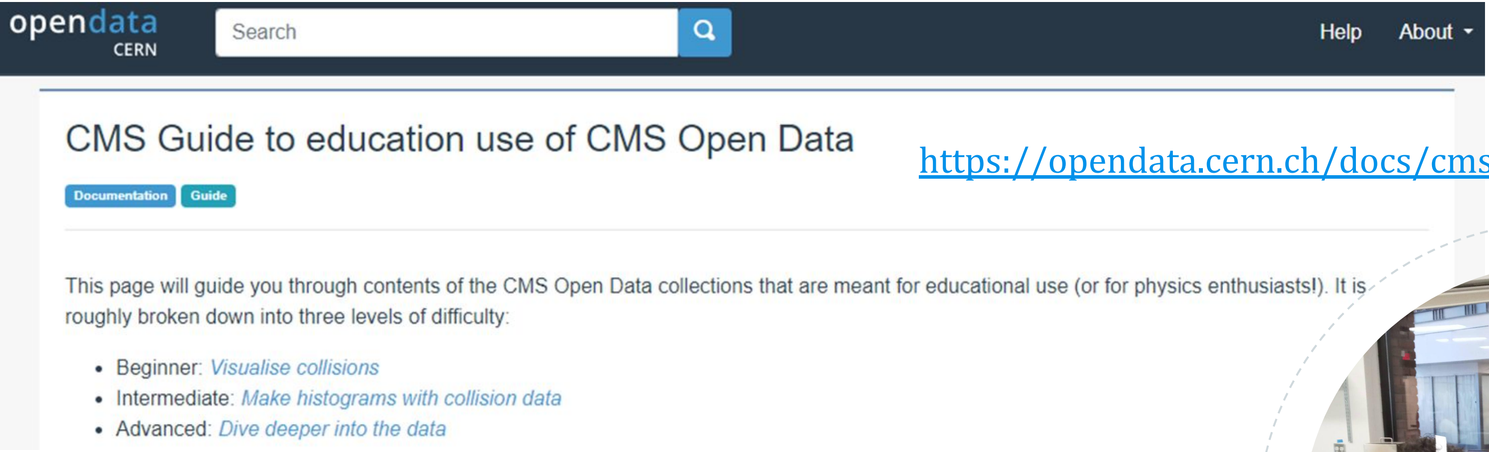
Flat ROOT
TTree
Basic types

Particle Flow
info added
for some data

Analyze in
ROOT or
Scikit-HEP
packages

A big step toward easily reusable CMS data

Resources: Education & Outreach



CMS Guide to education use of CMS Open Data

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*

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

◎ Interactive & VR event displays

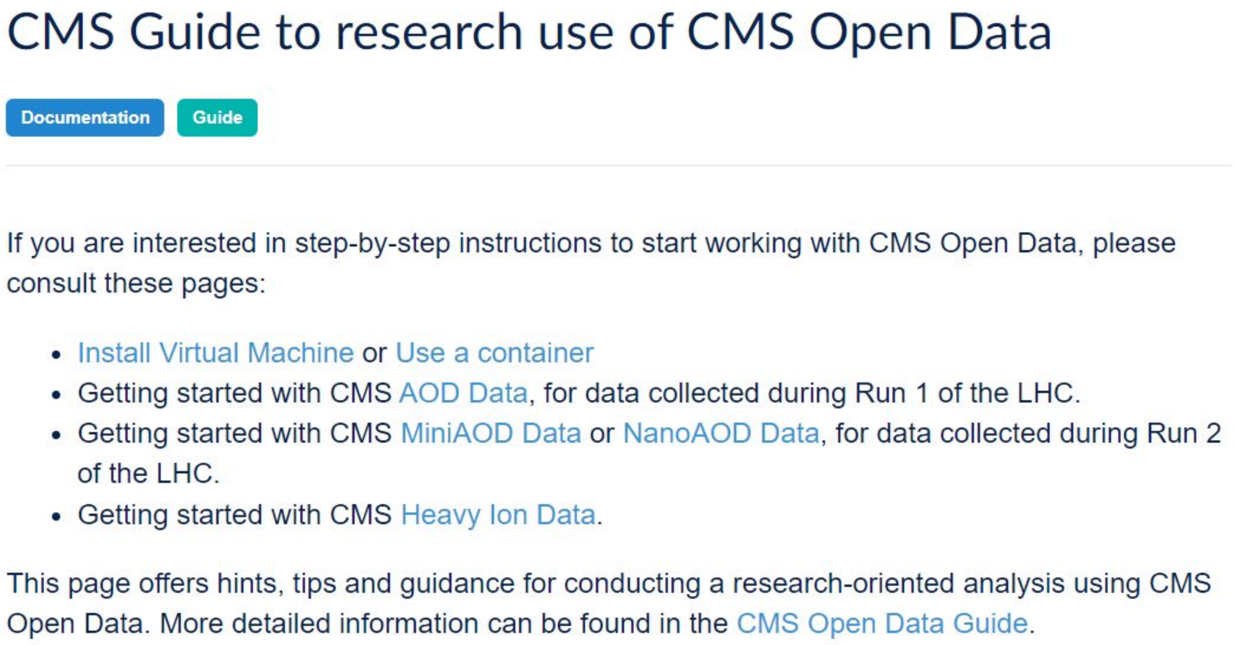
◎ International Masterclass

◎ Dimuon analyses for schools

◎ University-level course tools

Resources: Research Use

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



CMS Guide to research use of CMS Open Data

If you are interested in step-by-step instructions to start working with CMS Open Data, please consult these pages:

- Install Virtual Machine or Use a container
- Getting started with CMS AOD Data, for data collected during Run 1 of the LHC.
- Getting started with CMS MiniAOD Data or NanoAOD Data, for data collected during Run 2 of the LHC.
- Getting started with CMS Heavy Ion Data.

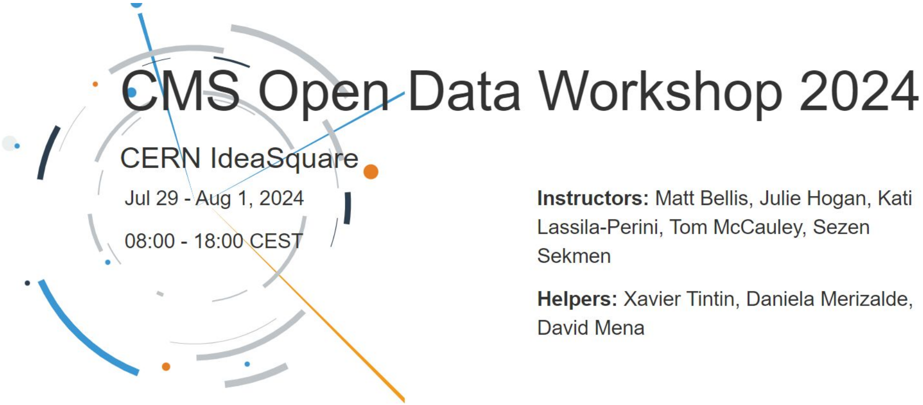
This page offers hints, tips and guidance for conducting a research-oriented analysis using CMS Open Data. More detailed information can be found in the CMS Open Data Guide.

- Logistics & Software
  "Getting Started" Guides
  Containers & VMs
  Metadata & Artifacts
- Physics analysis
  CMS Open Data Guide
  Example analysis records
- Hands-on help
  Annual Workshops!

Workshop series

Summer workshops teach:

- Finding data
- Data format structure
- Software environments
- Trigger system
- Physics object algorithms
- Event selection techniques
- Histogram creation
- Statistical analysis
- Scale-up techniques



CMS Open Data Workshop 2024

CERN IdeaSquare

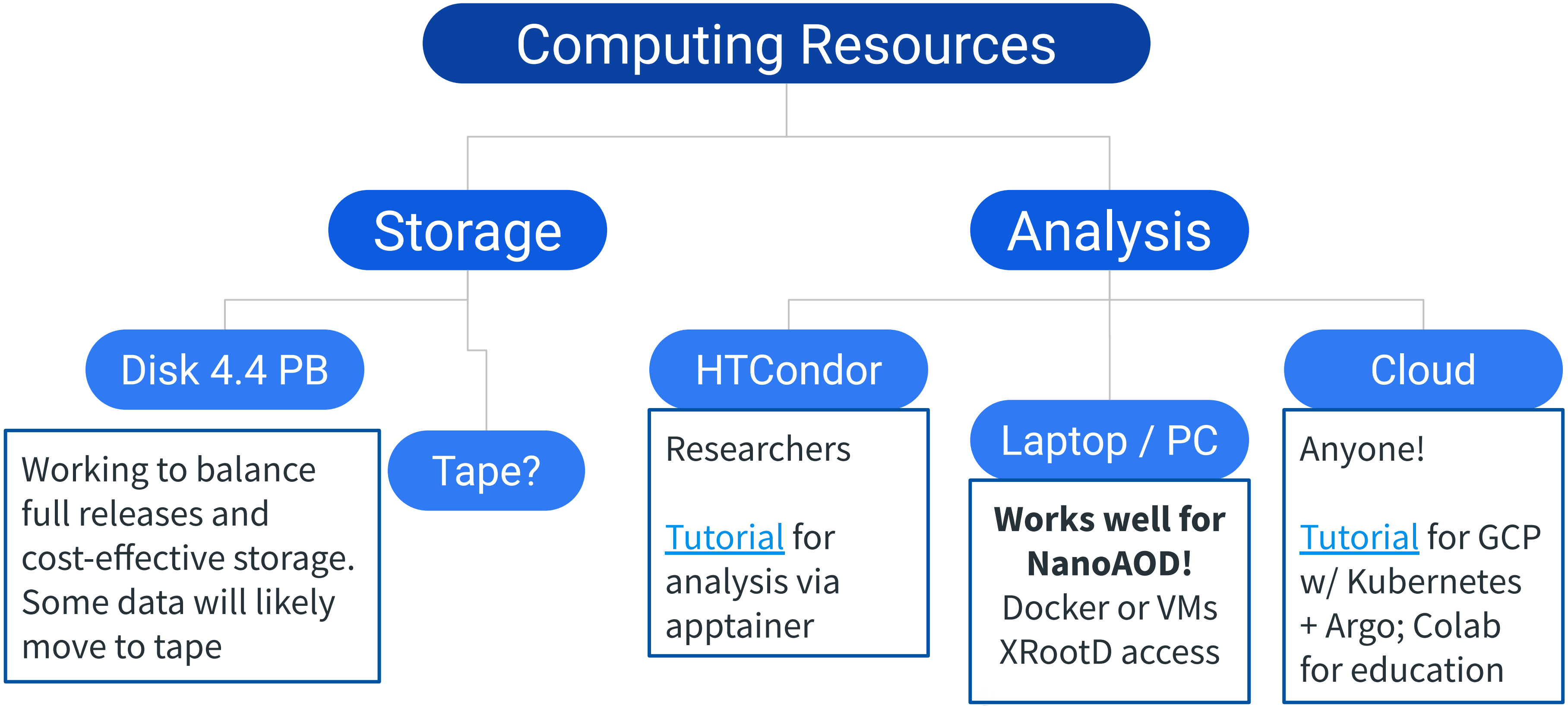
Jul 29 - Aug 1, 2024

08:00 - 18:00 CEST

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

Helpers: Xavier Tintin, Daniela Merizalde, David Mena

https://cms-opendata-workshop.github.io/2024-07-29-CERN/



Computing Resources

Storage

Disk 4.4 PB

Working to balance full releases and cost-effective storage. Some data will likely move to tape

Tape?

Analysis

HTCondor

Researchers

Tutorial for analysis via apptainer

Laptop / PC

Works well for NanoAOD!

Docker or VMs

XRootD access

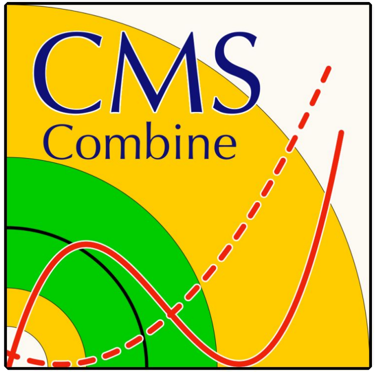
Cloud

Anyone!

Tutorial for GCP w/ Kubernetes + Argo; Colab for education

Resources: Statistical analysis software

- Paper on the COMBINE software, used for Higgs discovery
- Model that can be used in COMBINE to replicate the CMS Higgs discovery
- New analysis preservation practice in CMS to release likelihoods via CDS upon publication
- Supplements HEPdata, Open Data, etc
- Enabled a COMBINE limit setting lesson at our July workshop!



CMS Commitment to Open Science Takes the New Step

Positive experience,
model for the CERN policy

Continuous interest,
steady publication rate

Pioneering work for archiving
and serving data through CERN
Open data portal

Credits

Thanks to our colleagues:

- in the DPOA group in CMS
  - all organizers and contributors
- in the CERN Data preservation services
  - CERN Open data portal team, and many other services we rely on

And great thanks to all CMS open data users!

And thanks to SlidesCarnival for this free presentation template