Research use of event-level open data - how to get there?

(Re)interpretation of the LHC results for new physics, CERN, 26 February 2025

Kati Lassila-Perini
Helsinki Institute of Physics - Finland

Hello!

I am Kati Lassila-Perini

experimental particle physicist

CMS data preservation and open access (DPOA) coordinator (2012 - 2024)

ICFA Data Lifecycle panel chair (2024 -) - with a focus on Open Science

Find me at: kati.lassila-perini @ cern.ch

@katilp.bsky.social

HELSINKI INSTITUTE OF PHYSICS

Glossary for this talk

Data:

- reconstructed collision data **and** corresponding simulations in the same format (with the MC truth)

Event-level:

- generic primary datasets, no publication-specific selections other than online selection or MC process definition*

Research-quality:

- data in sufficient detail and accompanied with the necessary information so that they can be used in publication-level research

1

Your expectations

What would you want to find out about event-level open data before using it?

- ...
- ...
- ...
- ...
- ...
- ...

2 Providing open data

ATLAS releases first open data from heavy-ion collisions

Ten years of CERN Open Data portal

TOTEM releases first set of open data

DELPHI Collaboration releases its entire data collection

ATLAS releases 65 TB of open data for research

CMS releases 13 TeV proton collision data from 2016

Upgrade of the CERN Open Data portal

LHCb releases the entire Run I dataset

CMS completes Run-1 heavy ion open data collection

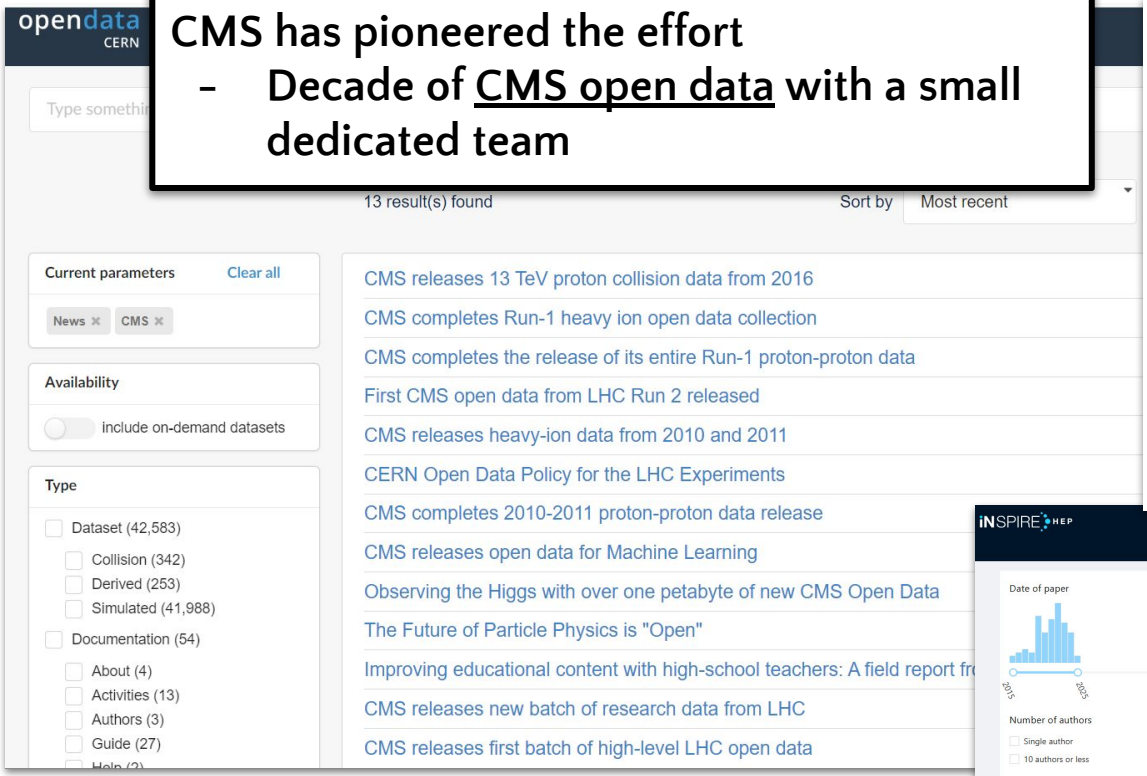
LHCb releases first set of data to the public

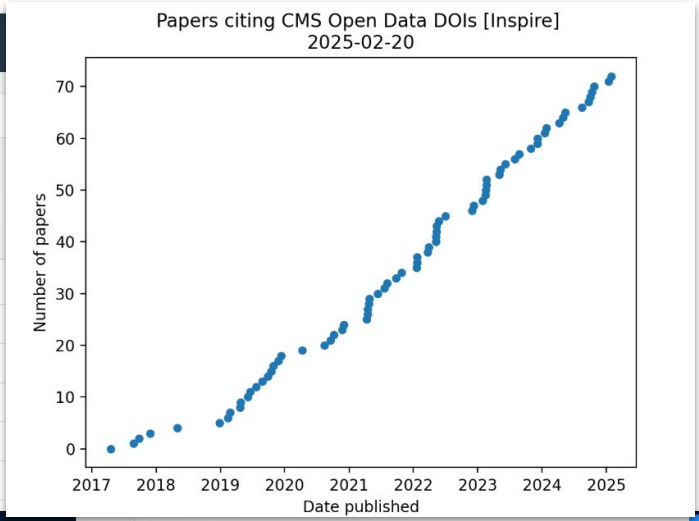
All LHC experiments committed releasing open data for research through the CERN Open data portal

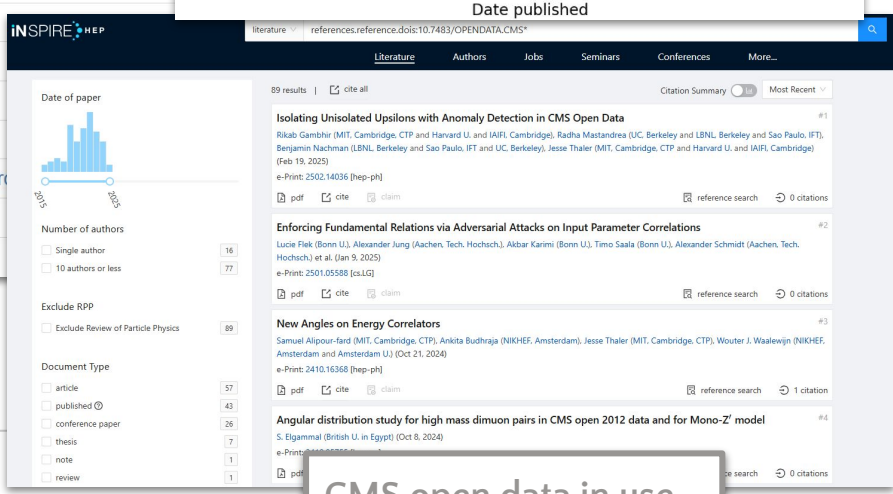
- **and not only LHC!**

CMS has pioneered the effort

- Decade of CMS open data with a small dedicated team







CMS open data in use

Isolating Unisolated Upsilons with Anomaly Detection in CMS Open Data

Rikab Gambhir (MIT, Cambridge, CTP and Harvard U. and IAIFI, Cambridge), Radha Mastandrea (UC, Berkeley and LBNL, Berkeley and Sao Paulo, IFT), Benjamin Nachman (LBNL, Berkeley and Sao Paulo, IFT and UC, Berkeley), Jesse Thaler (MIT, Cambridge, CTP and Harvard U. and IAIFI, Cambridge) (Feb 19, 2025)

e-Print: 2502.14036 [hep-ph]

Now that we have "rediscovered" the Υ, we plan to build upon this study in the future and perform a full scan over the dimuon mass range to search for resonances beyond the Standard Model. We encourage the anomaly detection community to use this study and this dataset for development and testing, since the Υ peak serves as a "standard candle" whose properties are well known. This dataset is complementary to previously publicized benchmark datasets, since it corresponds to real, noisy, detector-level experimental data. It also presents new challenges; for example, we found that not all auxiliary feature sets avoided sculpting, a hurdle which we had not encountered in previous studies on synthetic data. We hope this Υ analysis inspires further anomaly detection studies in real collider data.

End Matter: Code and Data. All of the code to reproduce the analyses and figures in this paper can be found at `https://github.com/hep-lbdl/dimuonAD`. The code consists of a series of numbered Python files and Jupyter notebooks, that, when run in order, will completely reproduce our analyses from scratch. The BDTs are trained using the XGBoost package [47] and the CATHODE normalizing flow is built and trained with PyTorch [48]. The architectures and hyperparameters used in this analysis can be found in the GitHub repository in the `configs` folder. The datasets used in this paper are available on Zenodo at `https://zenodo.org/records/14618719`.

Open data have value only when in use

From open data to open science

Tools:

- common software
- environments
- interfaces

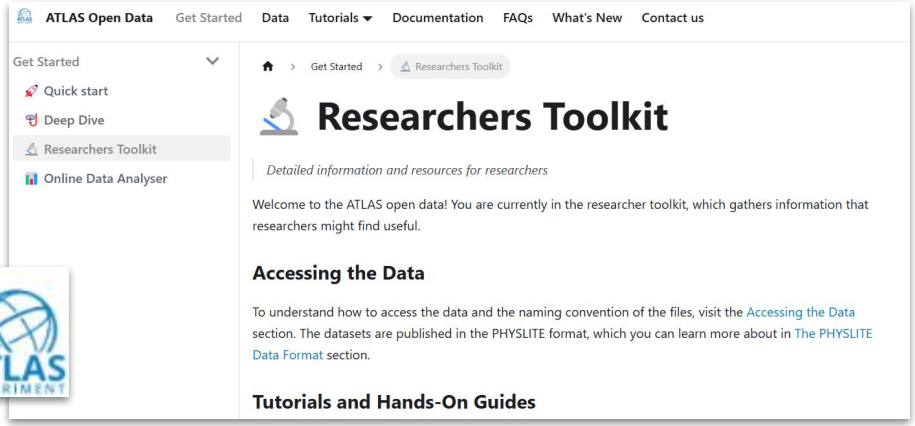
Data:

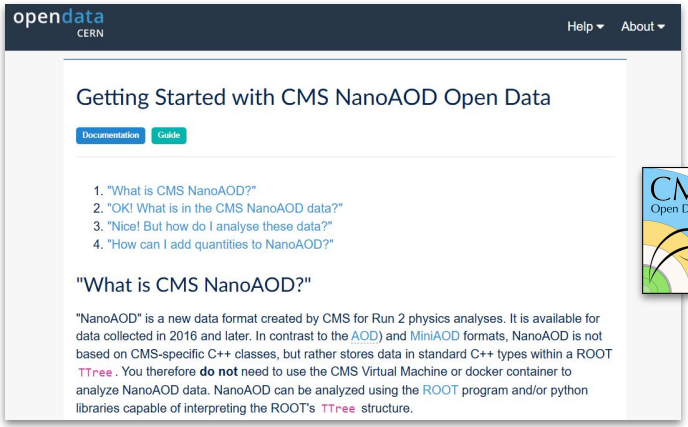
- collision data
- simulations
- auxiliary data for analysis

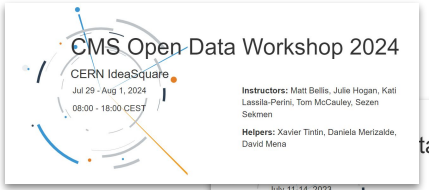
Knowledge:

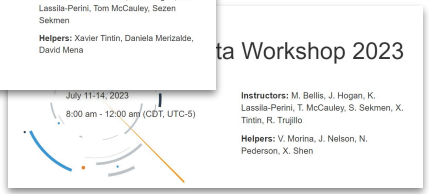
- instructions
- example code
- actionable examples

3 Using open data



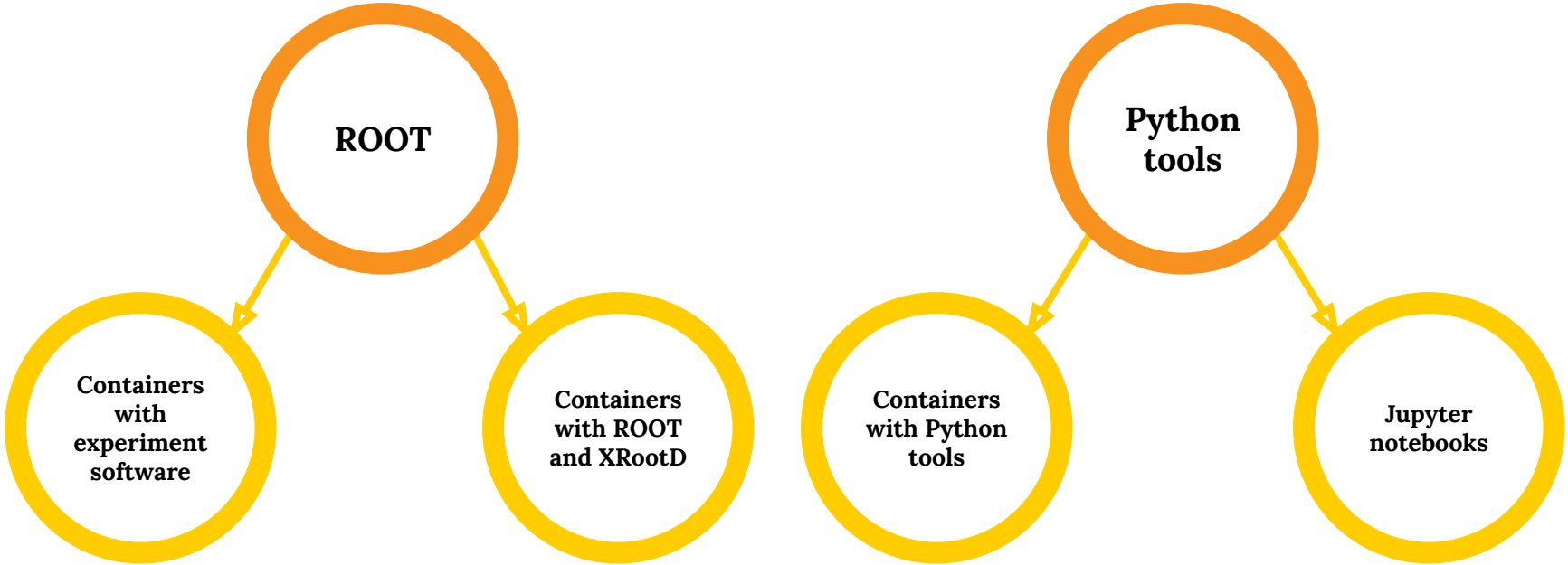




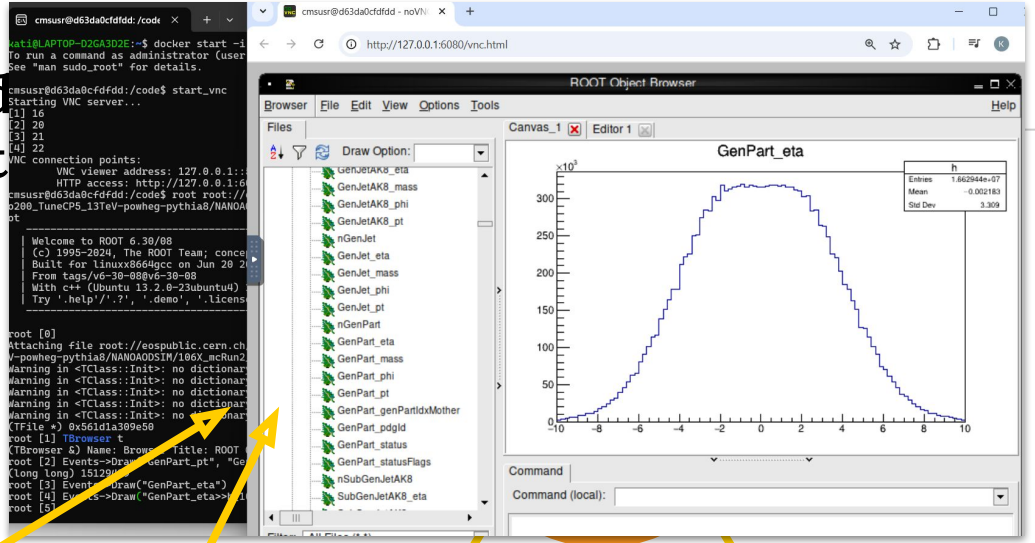


Getting started: easy 🚀

Choose your tool and get it running on your laptop



Choose your tool and
running on your lapt



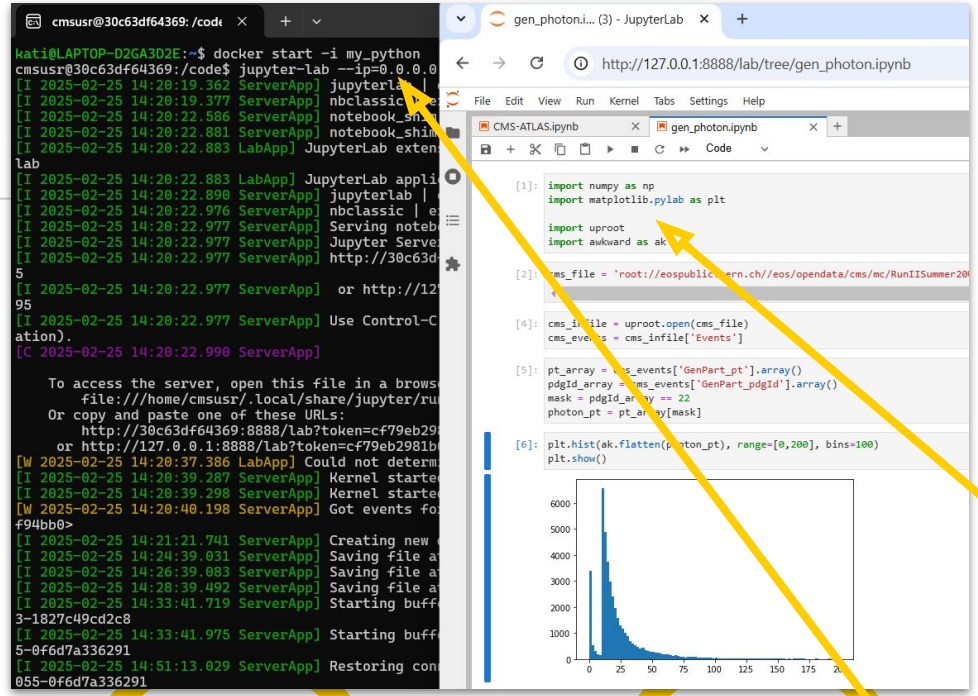
ROOT

Containers with experiment software

Containers with ROOT and XRootD

Containers with Python tools

Jupyter notebooks



Python tools

Containers with experiment software

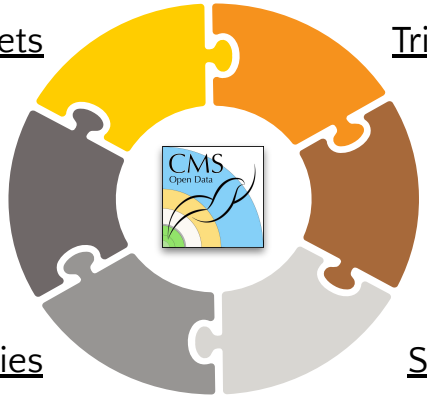
Containers with ROOT and XRootD

Containers with Python tools

Jupyter notebooks

Find and understand datasets

Triggers and luminosity



Event selection

Background modeling

Experimental uncertainties

Statistical inference

CMS Open data guide
CMS Open data workshops

Research level work is more complex but we have it covered

3.1 A quick example

MC generator-level information in open data

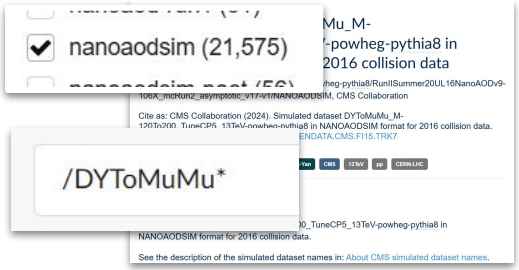
Did you know that:

- there are > 20000 MC datasets available with (reduced) generator particle information in flat ROOT format (NanoAODSIM)?
- there are many datasets with > 50M events?
- the description of variables for NanoAOD/NanoAODSIM is available in the dataset record?
- CMS organizes regular tutorials and the material is available for self study?

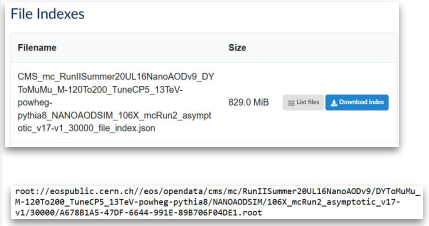
Now you know!

Let CMS know if these datasets are useful for you!

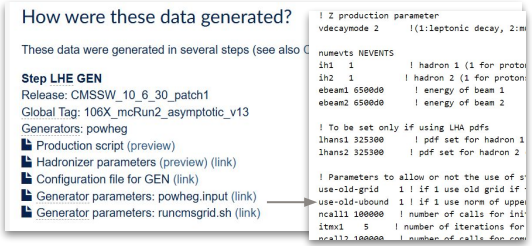
1. Find a dataset



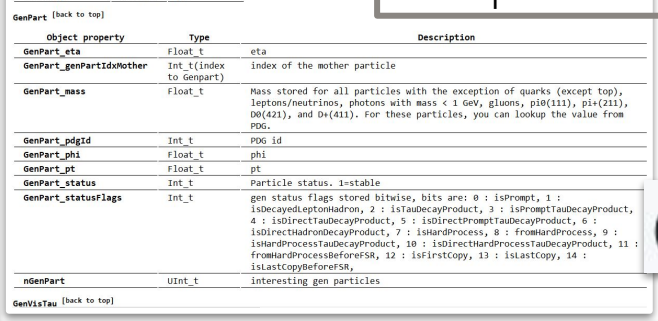
2. Find a file



3. Check generator parameters



4. Explore variables



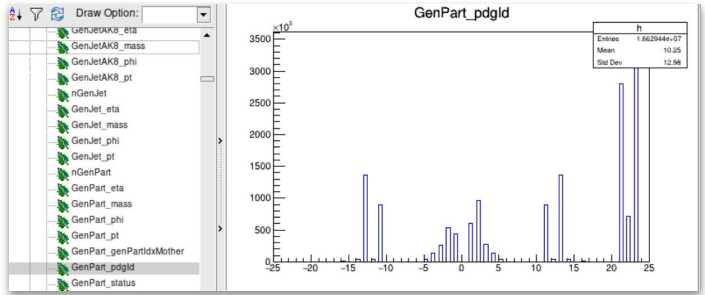
GenMET_pt	Float_t	pt

GenPart [back to top]

Object property	Type	Description
GenPart_eta	Float_t	eta
GenPart_genPartIdxMother	Int_t(index to Genpart)	index of the mother particle
GenPart_mass	Float_t	Mass stored for all particles with the exception of quarks (except top), leptons/neutrinos, photons with mass < 1 GeV, gluons, pi0(111), pi+(211), D0(421), and D+(411). For these particles, you can lookup the value from PDG.
GenPart_pdgId	Int_t	PDG id
GenPart_phi	Float_t	phi
GenPart_pt	Float_t	pt
GenPart_status	Int_t	Particle status. 1=stable
GenPart_statusFlags	Int_t	gen status flags stored bitwise, bits are: 0 : isPrompt, 1 : isDecayedLeptonHadron, 2 : isTauDecayProduct, 3 : isPromptTauDecayProduct, 4 : isDirectTauDecayProduct, 5 : isDirectPromptTauDecayProduct, 6 : isDirectHadronDecayProduct, 7 : isHardProcess, 8 : fromHardProcess, 9 : isHardProcessTauDecayProduct, 10 : isDirectHardProcessTauDecayProduct, 11 : fromHardProcessBeforeFSR, 12 : isFirstCopy, 13 : isLastCopy, 14 : isLastCopyBeforeFSR,
nGenPart	UInt_t	interesting gen particles

GenVisTau [back to top]

5. Explore contents with ROOT or Python tools

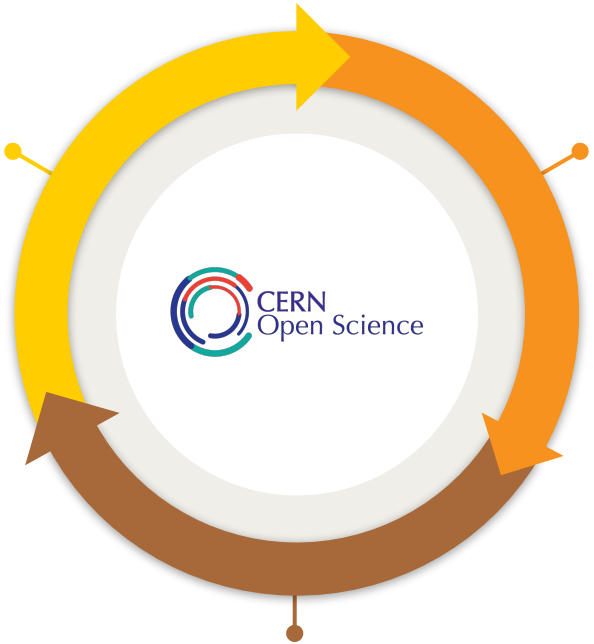


5 Open science

“Open science is one of our core missions.”

F. Gianotti: Keynote speech at
ALLEA General Assembly 2024

Open data are accessible to you through CERN Open data/science services 🩷



Open data resources do not yet have a well-defined budget line at CERN.

If you value open data, speak out!

Open science through open data

6 Take home messages

Homework for you:

- Make friends with software containers📦🧡
  - you do not need to, but they make working in different environments easier
- Explore open data🔎
  - getting started is easy: have a look, even without having immediate projects in mind
- Find out the obstacles🚧
  - we can maybe fix them
- Set expectations for preserved analysis workflows ✅✅✅
  - See next talks about analysis preservation
- Get in touch with us✍️💬
  - we need your feedback to improve
- Speak out📢
  - we cannot take the support to open data granted!

Thank you!

Questions ?

The presentation template SlidesCarnival

Icons by HAJICON, Iconjam, Darius Dan from Flaticon p 3