Open Science at CERN

... using the CMS experiment as leading example

opensciencecern

Clemens Lange (Paul Scherrer Institute PSI)

EuroLabs Advanced Training: Open Science and Data Management

26th November 2024

Introducing the CMS Experiment at CERN

Video: https://www.youtube.com/watch?v=EB5eZIR3AoM

Hi, I’m Clemens :-)

I’m a particle physicist at Paul Scherrer Institute, working on the CMS experiment at the Large Hadron Collider (LHC) at CERN

I roughly share my work time as follows

> Analysing the particle collisions provided by the LHC, recorded by the CMS detector (30-50%)

> Building and testing new pixel detectors for the upgrade of the CMS experiment (40-60%)

> Detector operations (5%)

> Open data/computing (~10%)

???

Open Science Governance at CERN

CERN | CERN Open Science

Director General | Director of Research and Computing

Open Science Steering Board

Open Science Practitioners Forum (OSPF)

Chair: Clemens Lange

How did he end up there?

… Office

Open Data Working Group

Implementation Working Group

…ic Information Policy Board

… Working Group

… Interest Group

Open Source Programme Office (OSPO)

High-Energy Physics at the CERN LHC

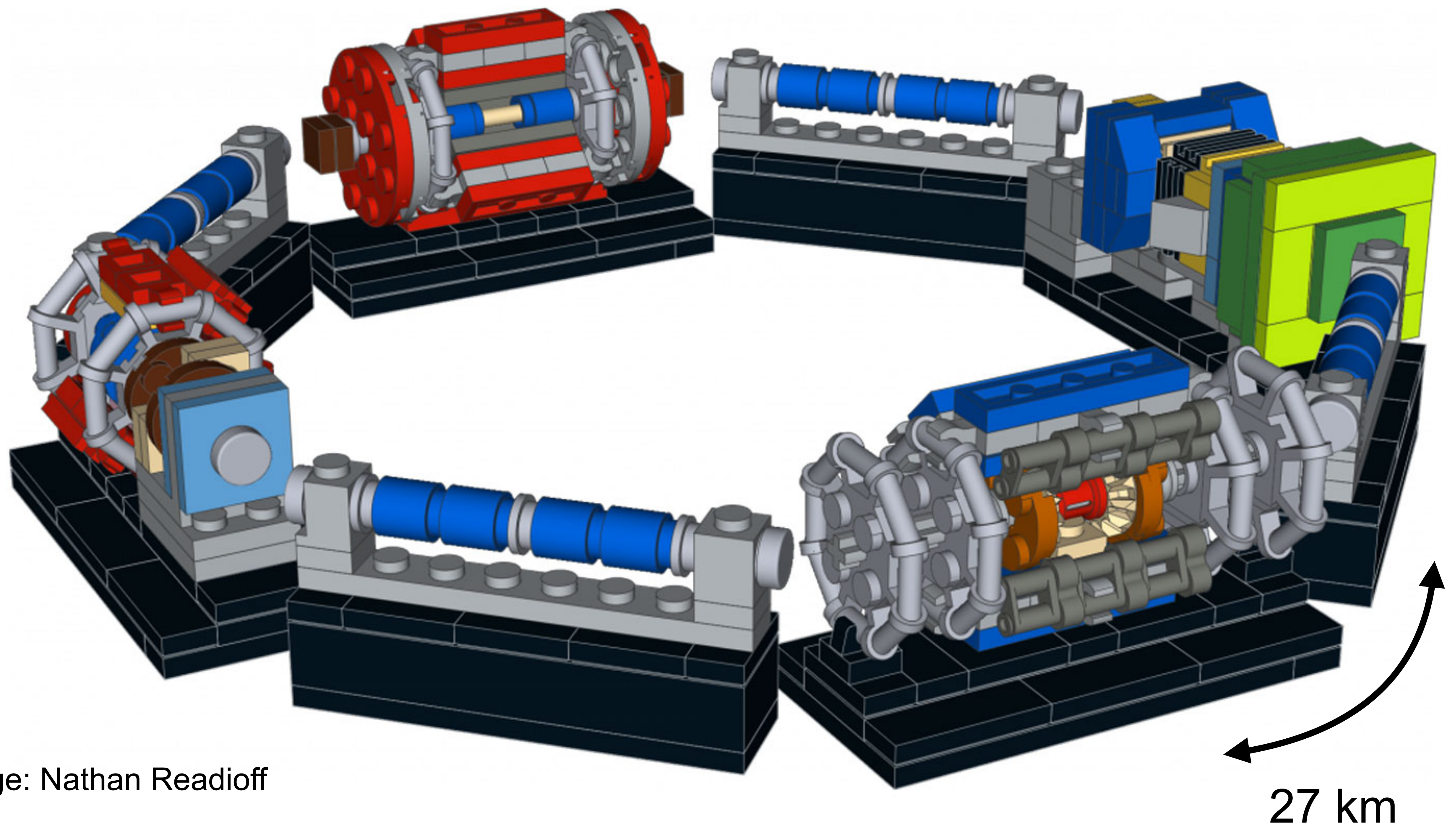


Image: Nathan Readioff

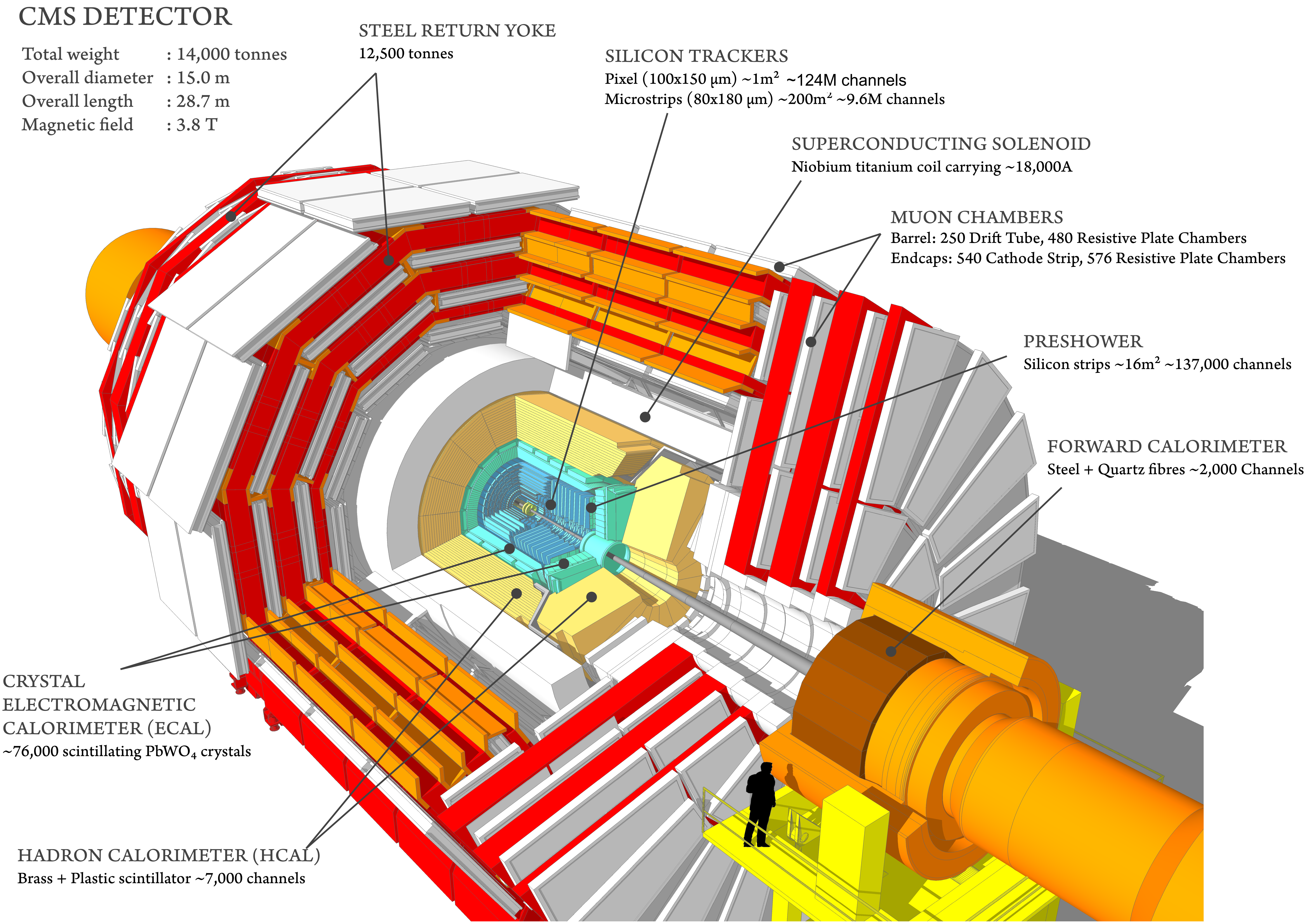
Four large experiments:

- ATLAS (5500 members of which almost 3000 scientific authors)
- ALICE (almost 2000 members)
- CMS (4000 particle physicists, engineers, computer scientists, technicians and students)
- LHCb (about 1700 scientists, engineers and technicians)

... plus several smaller ones

Today: more than 13,000 people involved in the experiments

The CMS Experiment



Record up to **40,000,000 events** of the LHC collisions **per second**, 24/7 (almost) all year long

Goal: understand the smallest building blocks of matter

~134 million readout channels — extraordinary levels of technical sophistication

These data are unique, e.g. the Higgs boson can only be measured at the LHC

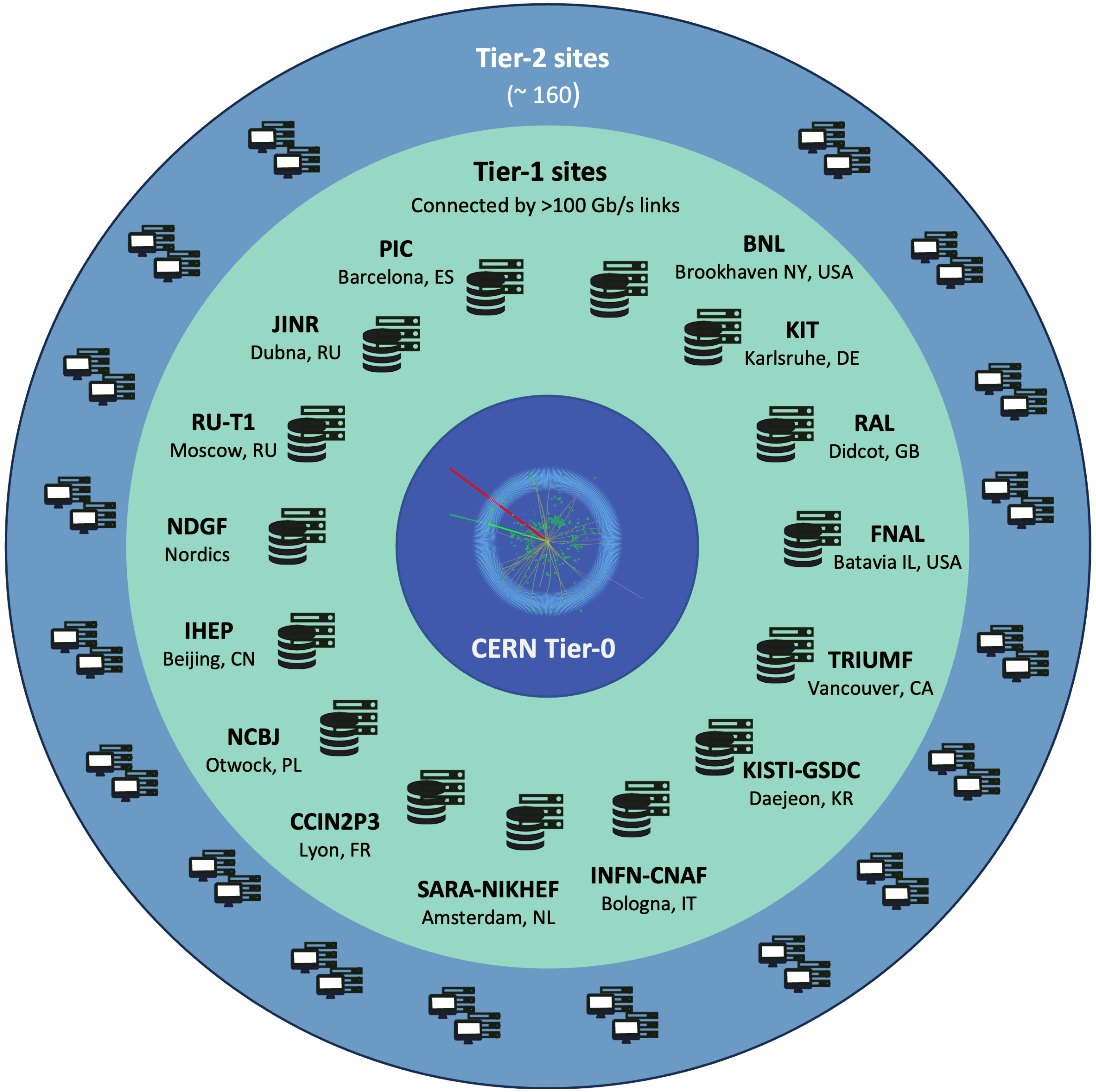
Data Processing using the Worldwide LHC Computing Grid

Centralised processing takes place in an **automated** way on the Worldwide LHC Computing Grid (> 170 computer centres, > 1 million computer cores, 2 exabytes of storage)

Each experiment has its own computing model

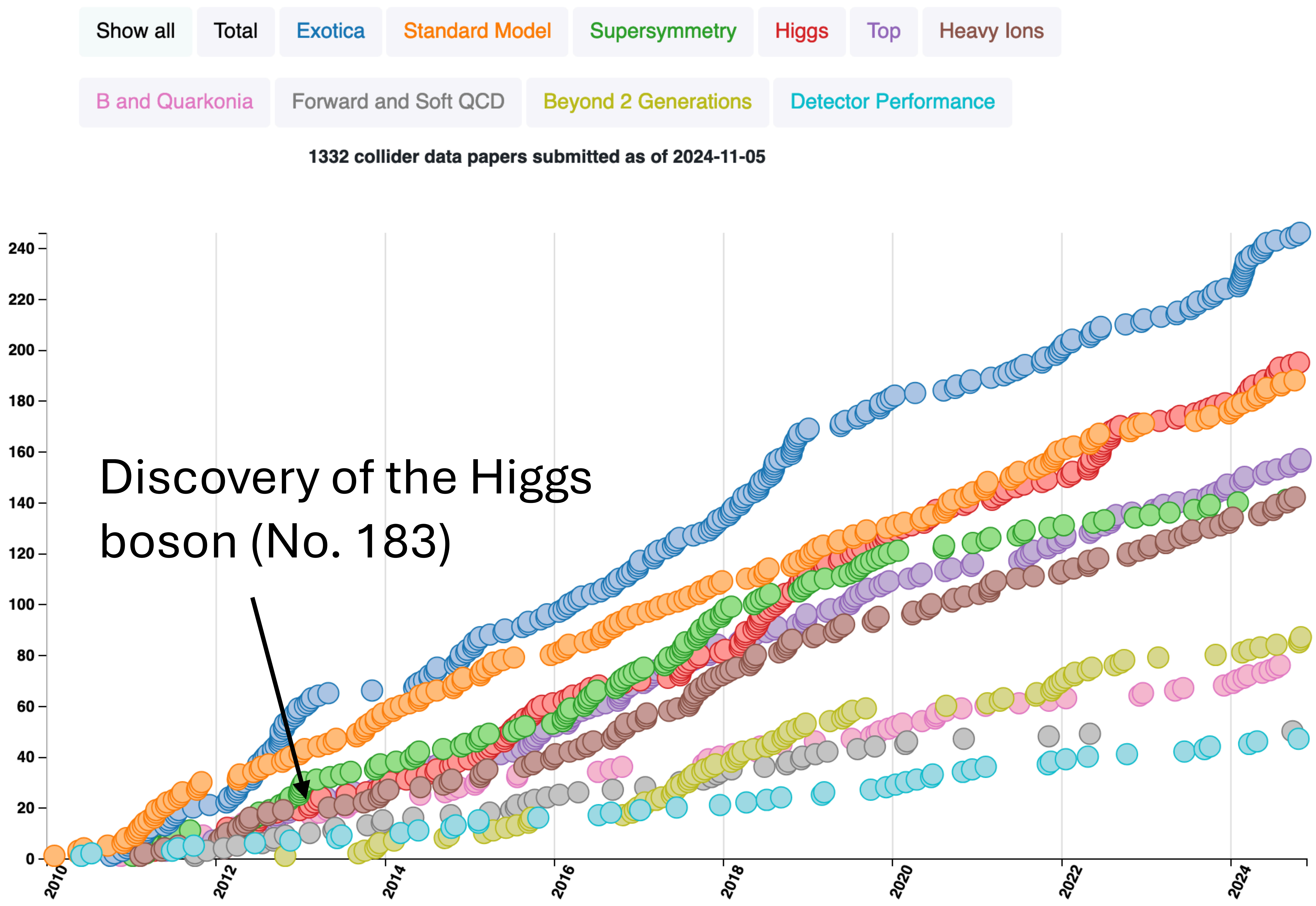
CMS: container images containing base Linux operating system with experiment software served through CernVM file system

This is what makes things easy



Source: https://wlcg-public.web.cern.ch/tiers

CMS Publications

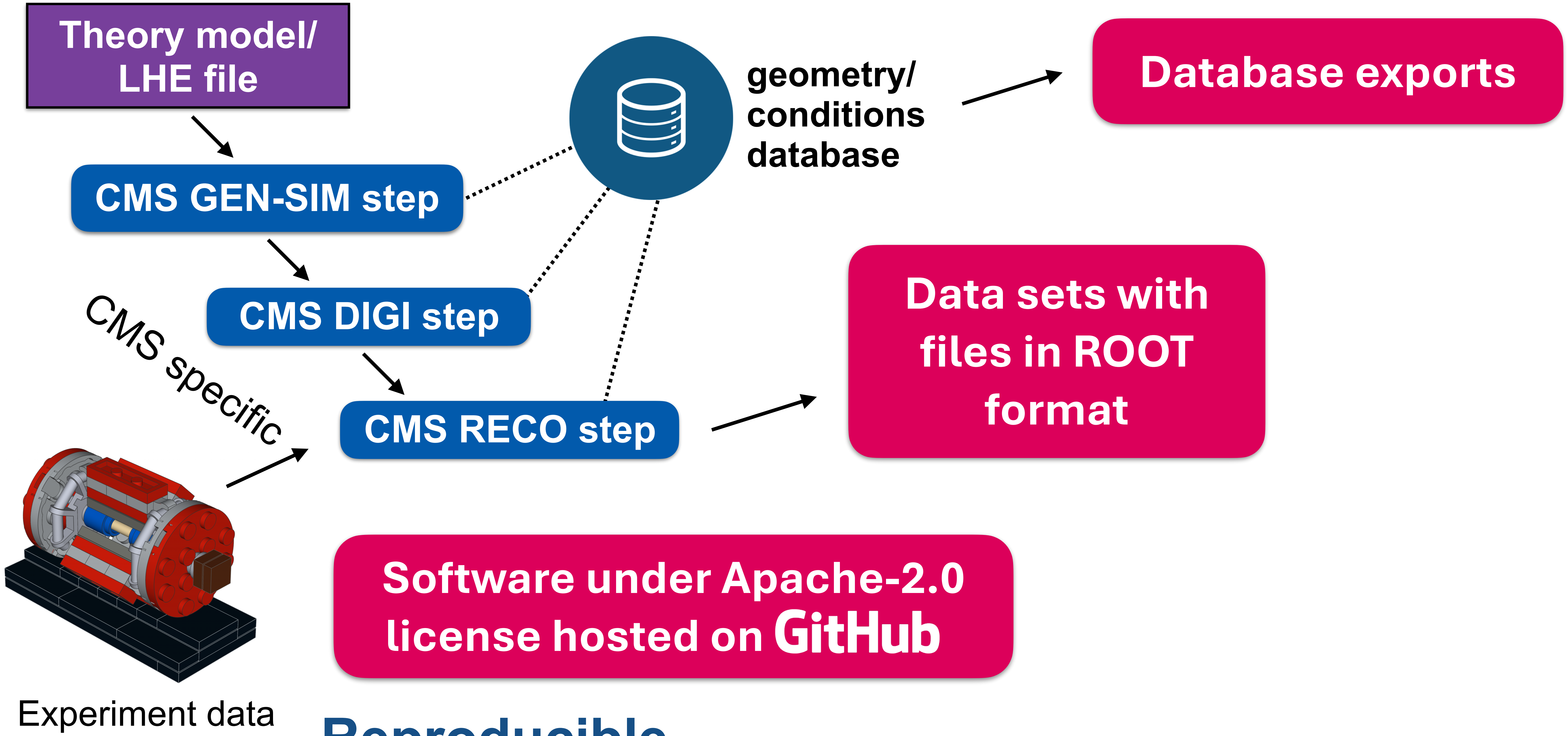


Interactive version at http://cms-results.web.cern.ch/cms-results/public-results/publications-vs-time/

Current State of Analysis Preservation in CMS (1)

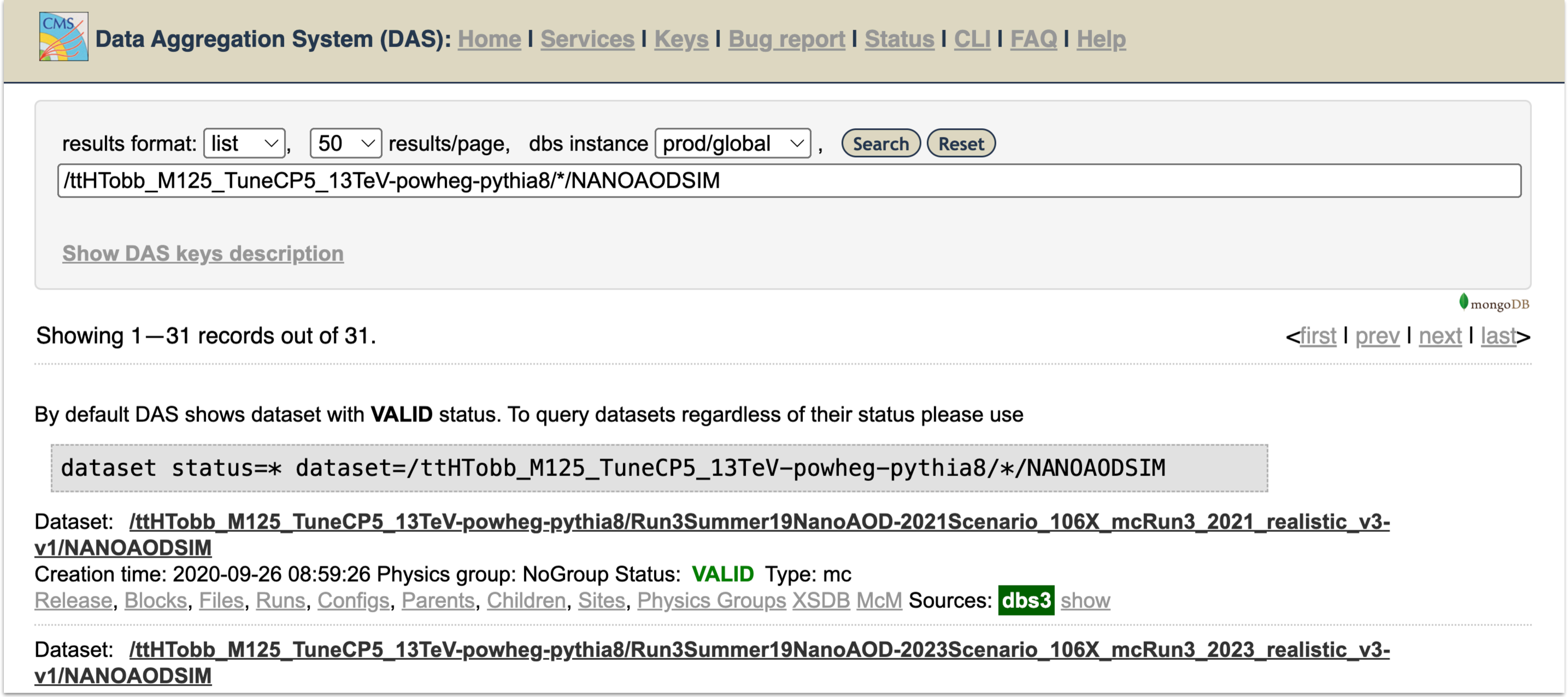
Event generation+simulation as well as reconstruction (both data and MC) centralised

- Software and database tags preserved and archived



Reproducible

CMS Data Aggregation Service (internal)



All data sets created using grid resources are registered

> Naming is important!

From Internal CMS Data Sets to Open Data

A lot of information encoded in a data set:

m_{Higgs} | centre-of-mass energy | process | underlying event tune | event generator(s) | Processing campaign

/ttHTobb_M125_TuneCP5_13TeV-powheg-pythia8/

RunIISummer20UL16NanoAODv9-106X_mcRun2_asymptotic_v17-v2/

NANOAODSIM

Data tier

This is the hard part

There is **a lot more metadata that is not visible** though, added to public records such as https://opendata.cern.ch/record/67645 — **can trace/reproduce entire chain** — often used by CMS collaborators now!

Open Data are eventually transferred to public EOS instance at CERN

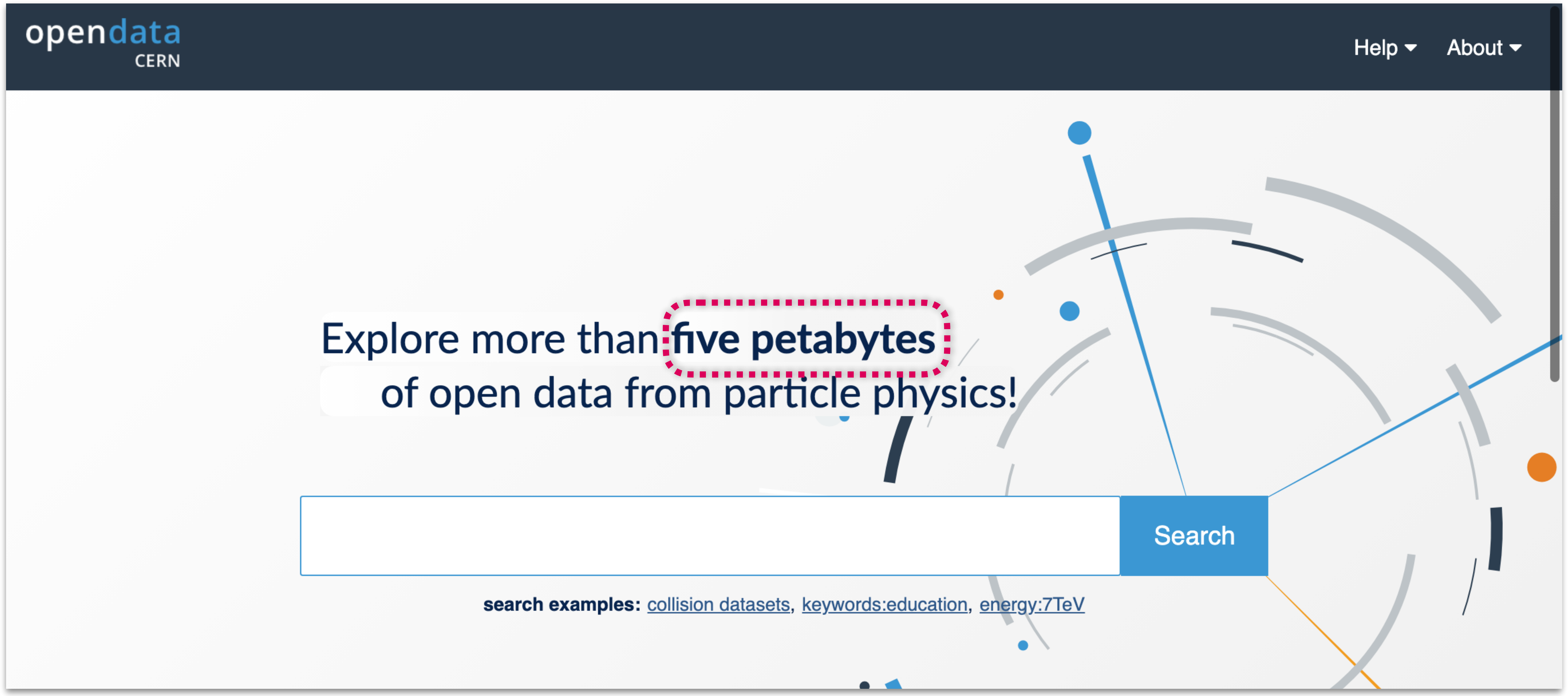
Corresponding records published on CERN Open Data Portal (with DOIs)

10 Years of CMS Open Data and the CERN Open Data Portal

Since 2014, CMS has released ~4.5 petabytes of open data available on the CERN Open Data Portal

- Both collision and simulation data sets
- Entire Run-1 (2010-2012) + 2015 data sets, fraction of 2016

More information: https://cms.cern/news/cms-celebrates-decade-open-data



From CMS to CERN Open Data

At the end of 2020, all large LHC experimental collaborations have endorsed a new open data policy

- Following existing CMS policy

Commit to publicly **releasing data required to ma**

Data and simula released approximately five years after collection (50%)

- Released und Commons CC
- Full dataset by the close of the experiment

ATLAS and LHCb (0.8 PB) recently released their first research-level Open Data

higher comp ... rt

- Level 1: Open access publication and additional
- ...utreach and Education
- ... and the software to analyse them
- Level 4: Raw data, and the software to reconstruct and analyse them

In 2022, developed CERN Open Science Policy and its Implementation Plan

Challenges of Providing Open Data

Data: available ≠ usable

Open Data needs to be FAIR:

Findable ➜ CERN Open Data Portal records

Accessible ➜ reliable storage and access technology

Interoperable ➜ provide good documentation, avoid jargon

Reusable ➜ **preserve software** (and hardware to run it if needed), data provenance, workflows

Building CMSSW container (Docker) images got me involved in CMS Open Data

Getting Others to Use Your Open Data

Beyond the data sets available on the CERN Open Data Portal, we provide:

Analysis examples with different levels of complexity (scientific and education)

The required software

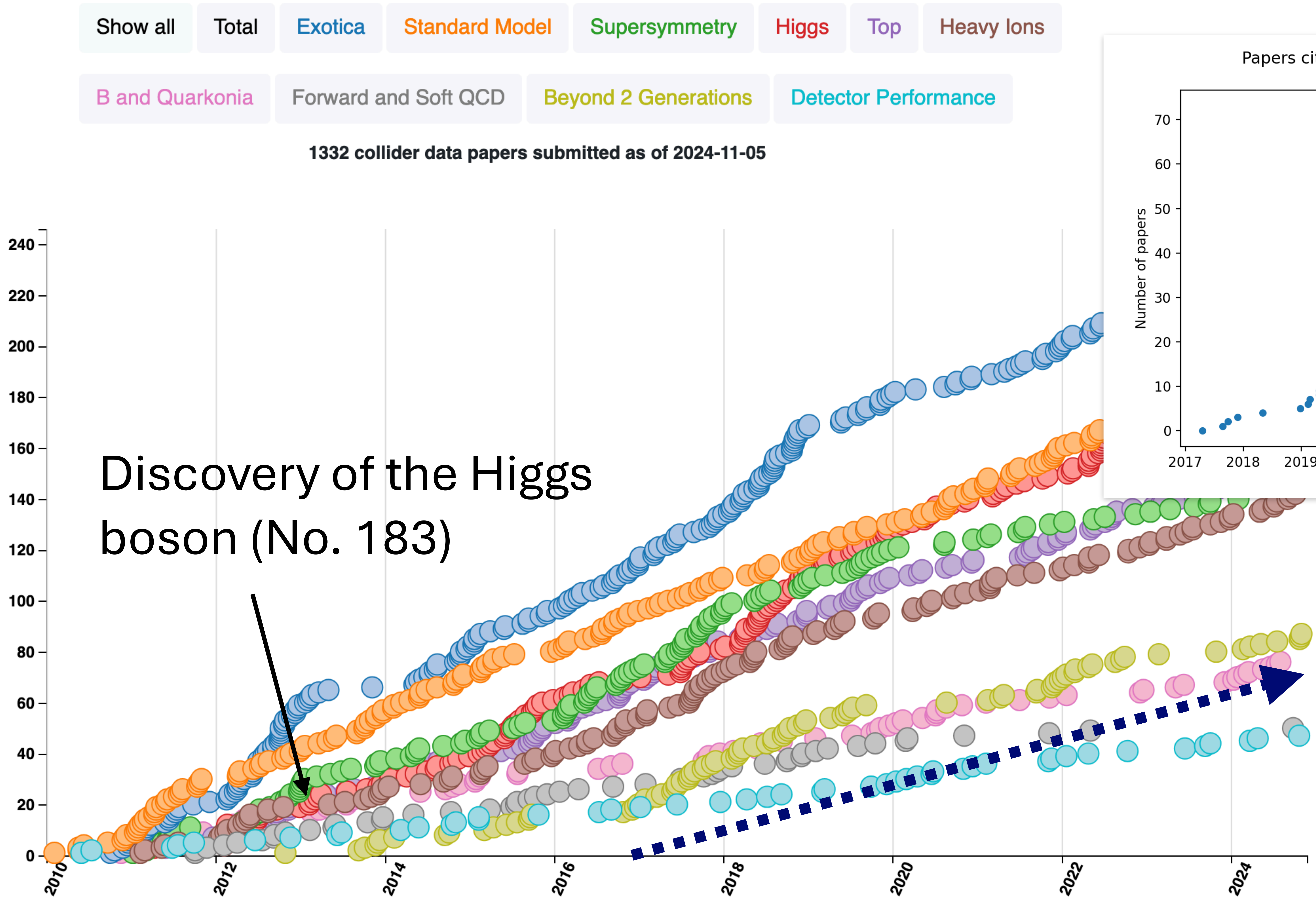
A separate CMS Open Data Guide

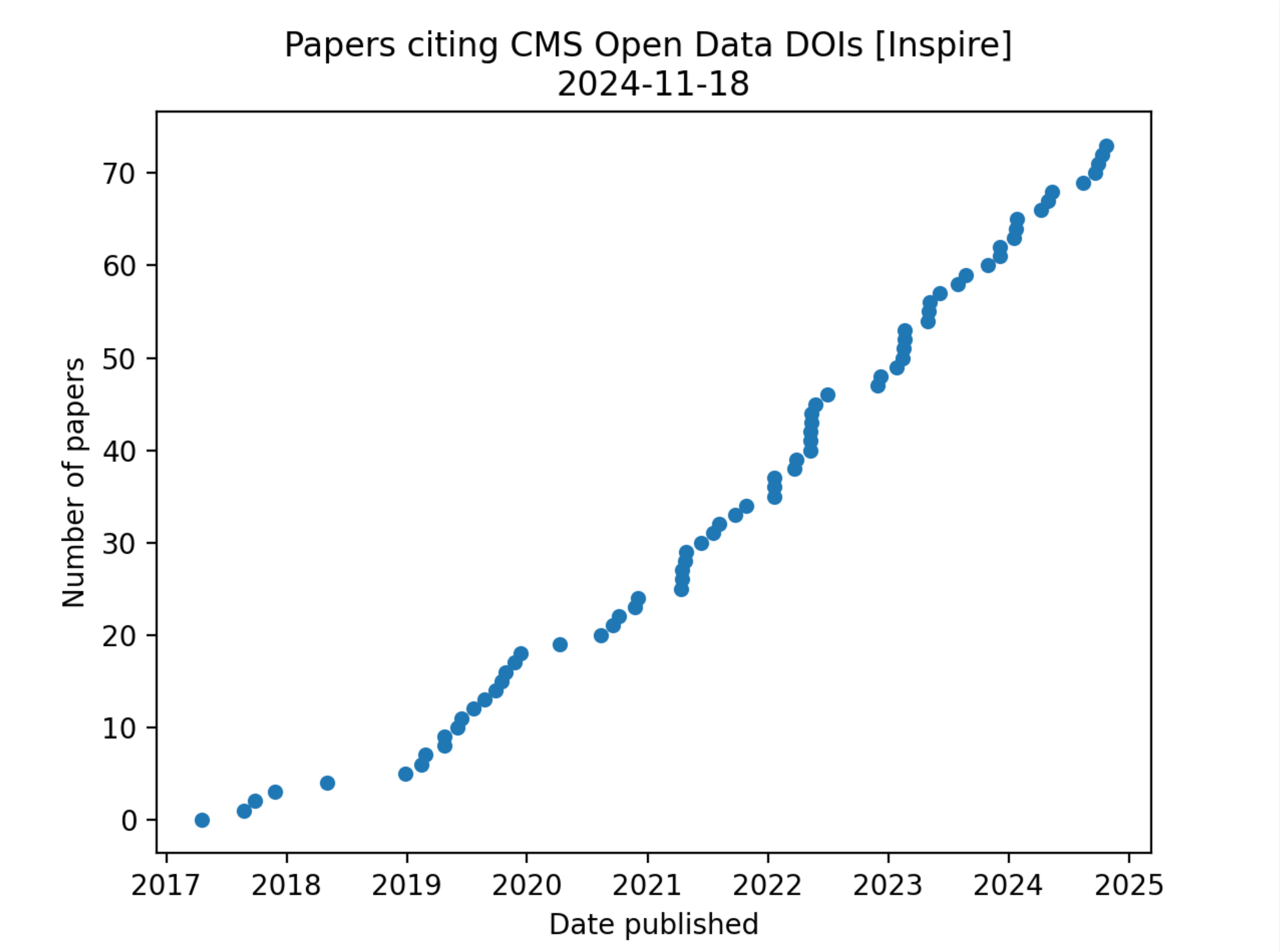
- In particular, trying to explain **how to use** the data and **what to do** with them in addition to **what is** in the data

Workshops with Software Carpentry style tutorials:

- 2020 CMS Open Data Workshop for Theorists
- 2021 CMS Open Data Workshop
- 2022 CMS Open Data Workshop at CERN
- 2023 CMS Open Data Workshop at Fermilab LPC
- 2024 CMS Open Data Workshop & Hackathon at CERN

CMS Publications





OD results: Equivalent of a new working group (but we are not authors)

Interactive version at http://cms-results.web.cern.ch/cms-results/public-results/publications-vs-time/

Beyond Open Data

Current State of Analysis Preservation in CMS (2)

Event generation+simulation as well as reconstruction (both data and MC) centralised

> Software and database tags preserved and archived

Internal documentation (analysis notes) preserved

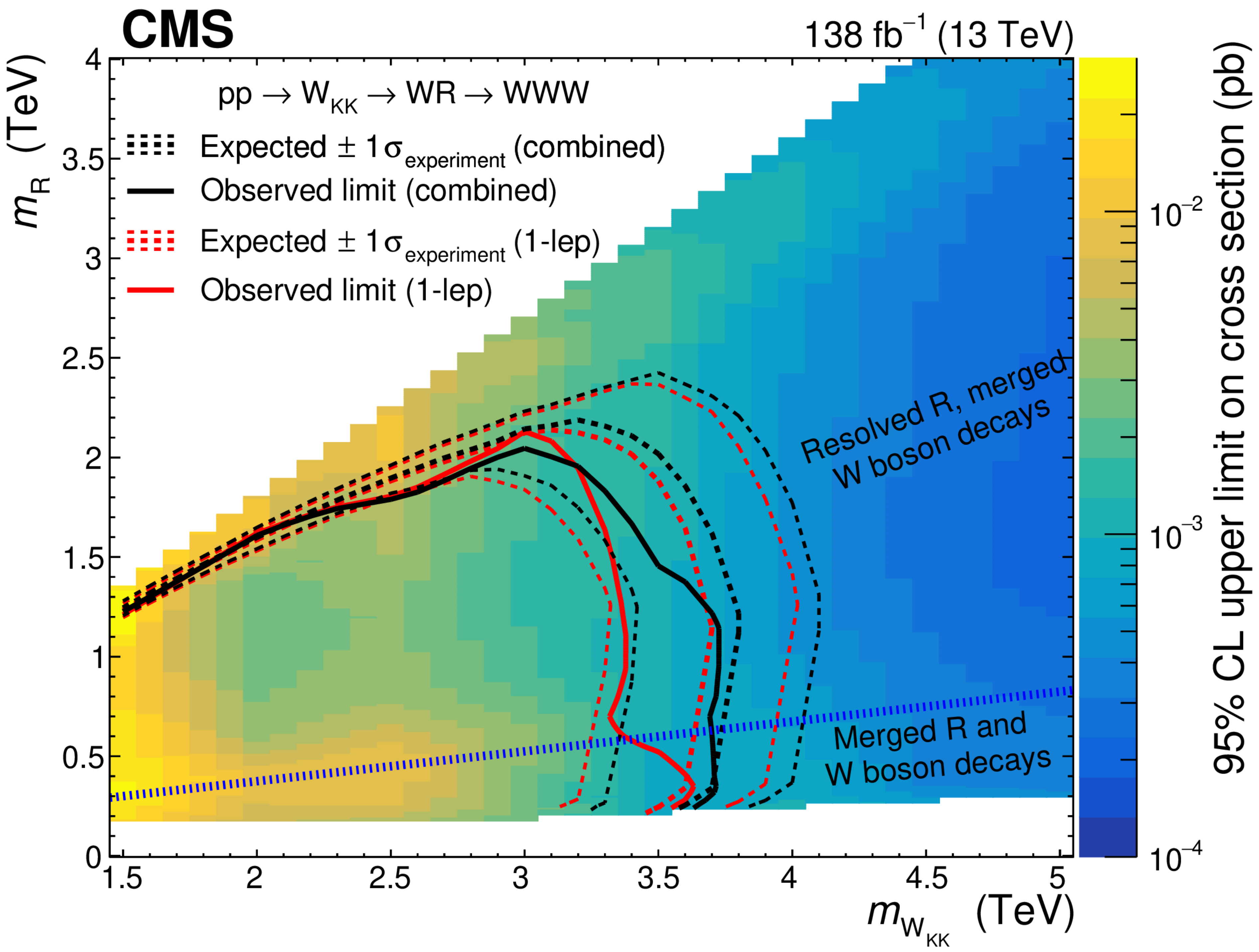
Paper publications **open access** (since 2014 under SCOAP³), Preprints available on arXiv

HEPData

Paper publications typically in PDF format — nice to read and print, but challenging to extract actual data

Goal: provide figures and tables in tabulated, machine-readable format

Can be used for comparisons, reinterpretations etc.



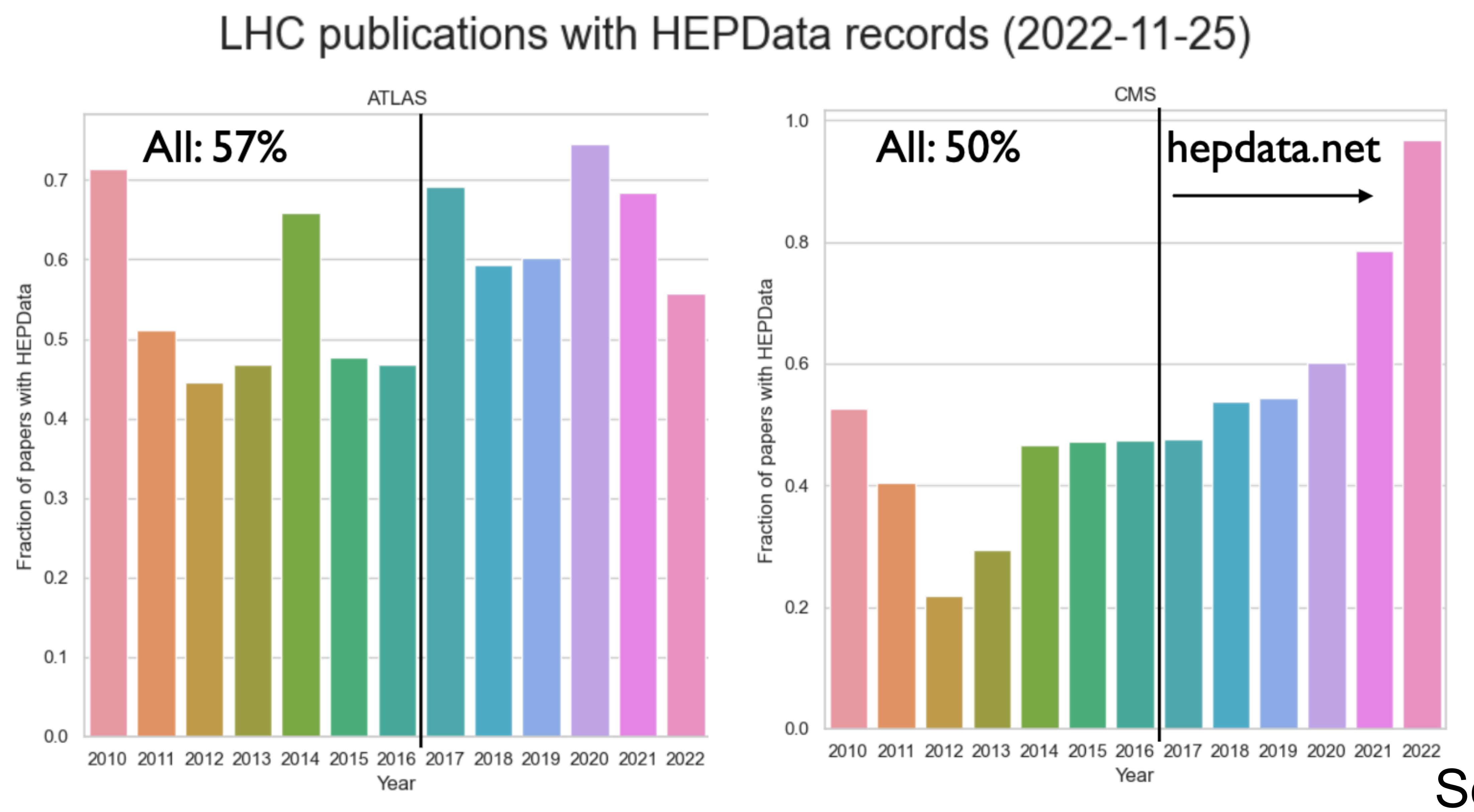
Showing 50 of 18000 values — Show All 18000 values

W_KK mass [TeV]	Radion mass [TeV]	upper limit on signal cross sections × branching fraction to 3 W bosons [pb]
1.5	0.0	0.0
1.5	0.01	0.0
1.5	0.02	0.0
1.5	0.03	0.0
1.5	0.04	0.0
1.5	0.05	0.0
1.5	0.06	0.0
1.5	0.07	0.0
1.5	0.08	0.0

HEPData

doi.org/10.17182/hepdata.102646

Increasing HEPData Adoption



Source

CMS has HEPData records for almost all analyses since 2022

How?

In 2018, hepdata_lib was born: “[Python] Library for getting your data into HEPData” → provide users with **good tooling**, show that it is **easy**, **make it mandatory** in the collaboration (2021)

Open Software and Hardware

Side Note: Open-Sourcing Software and Hardware

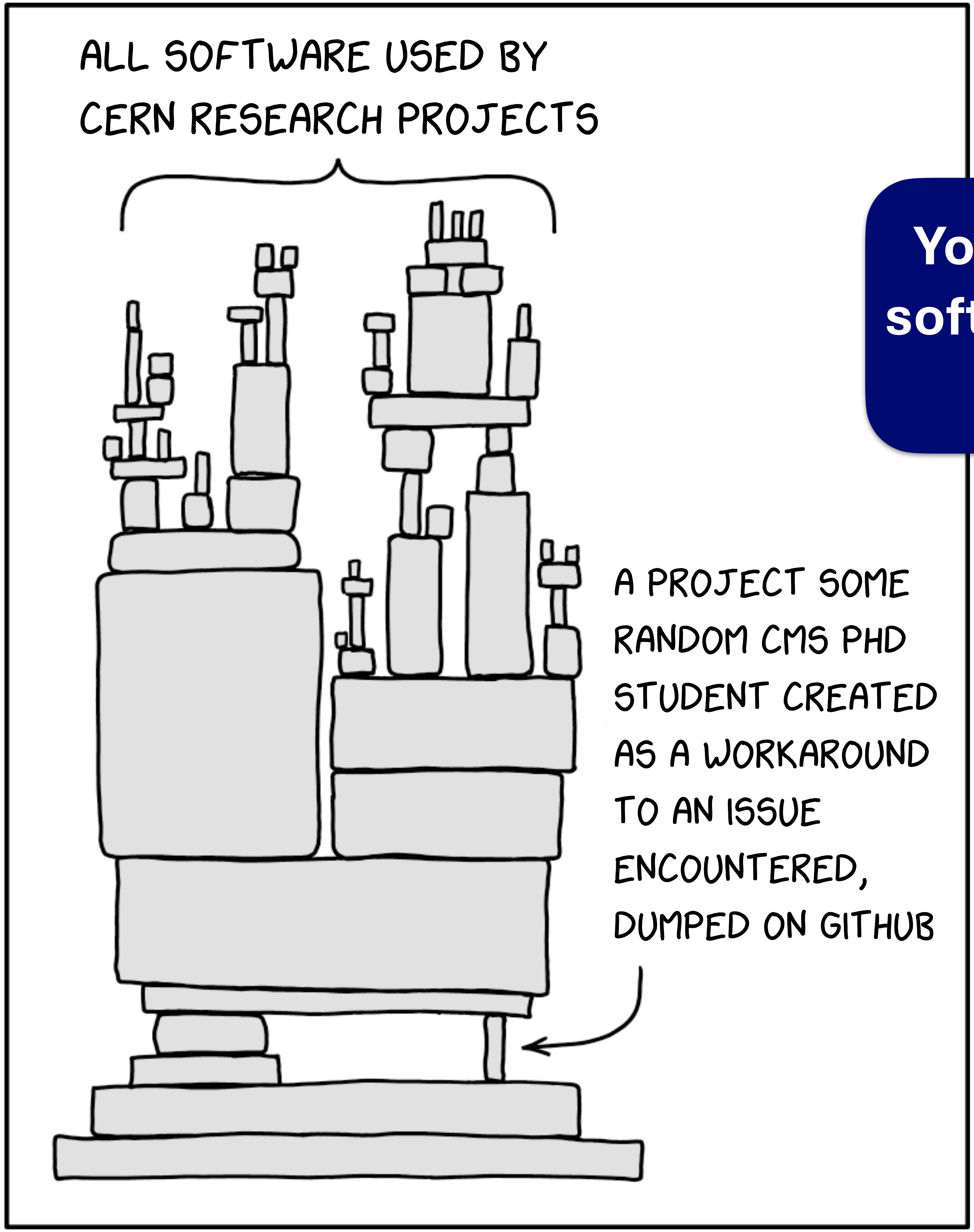
CERN has recently established an Open Source Programme Office (OSPO)

Goal is to support the CERN community in the process of **making internal projects public**, e.g.:

- Identify and apply suitable license
- Guidelines for project maintenance and support

Also, provide a public catalogue of CERN's open source projects

More information: opensource.cern



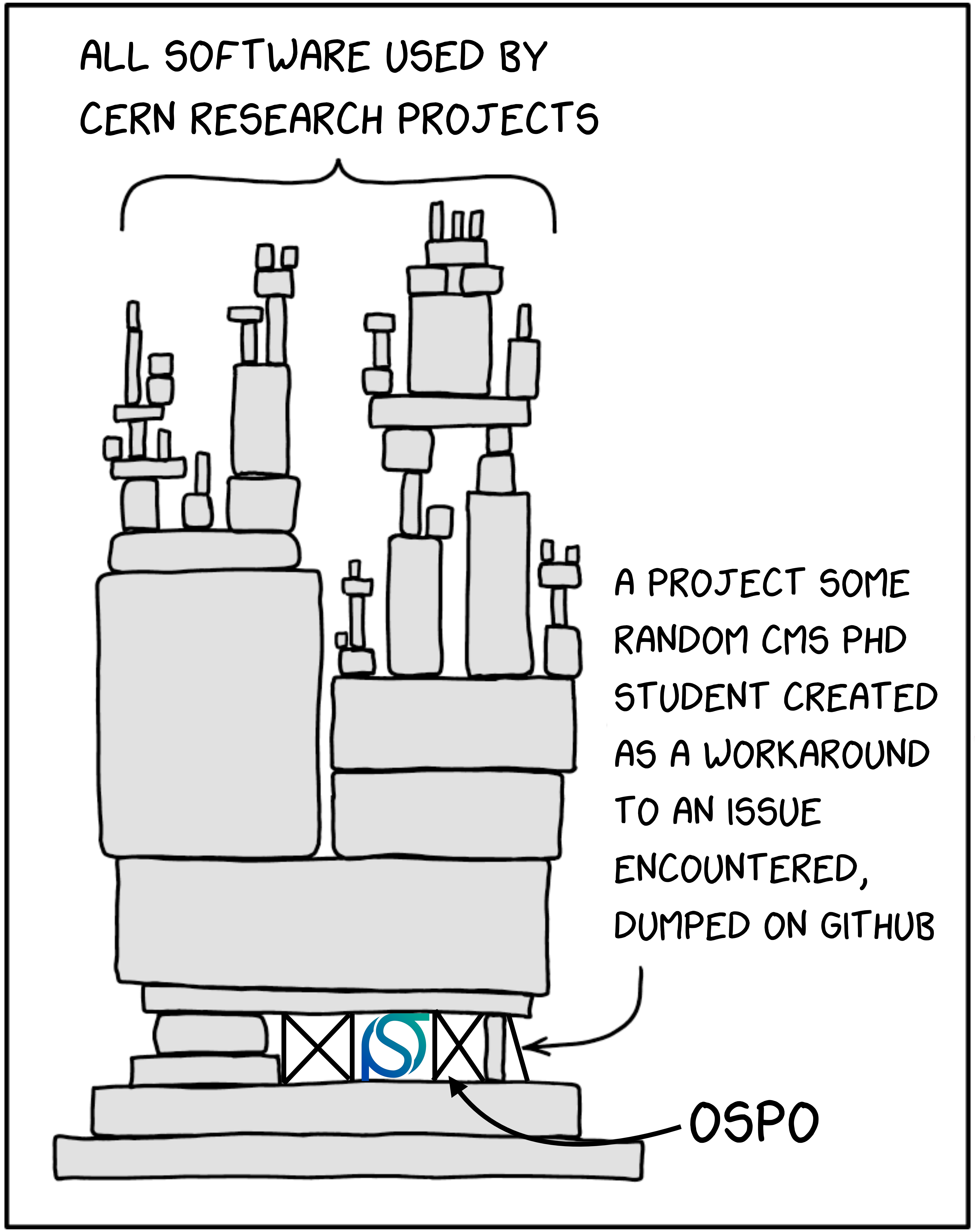
You probably also write software that is useful for others

Adapted from https://xkcd.com/2347/

Sharing?

Maintenance?

Contributing?



License?

Findable?

...

Adapted from https://xkcd.com/2347/

Preserving Physics Analyses

Current State of Analysis Preservation in CMS (2)

Event generation+simulation as well as reconstruction (both data and MC) centralised

> Software and database tags preserved and archived

Internal documentation (analysis notes) preserved

Paper publications **open access** (since 2014 under SCOAP3), Preprints available on arXiv

Helping Your Future You

"Your closest collaborator is you six months ago… and your younger self doesn't reply to emails" → ***preserving your analysis pipelines will help you in your immediate future***

Steps towards Reusable Analyses

1.Capture software	2. Capture commands	3. Capture workflow
Individual analysis stages in an executable way (including all dependencies)	How to run the captured software?	How to connect the individual analysis steps?

Capturing analysis code almost trivial today

Requires e.g. two additional files in a GitLab repository → something you will learn this week

CMS Open Data simplified analysis example

https://opendata.cern.ch/record/5500

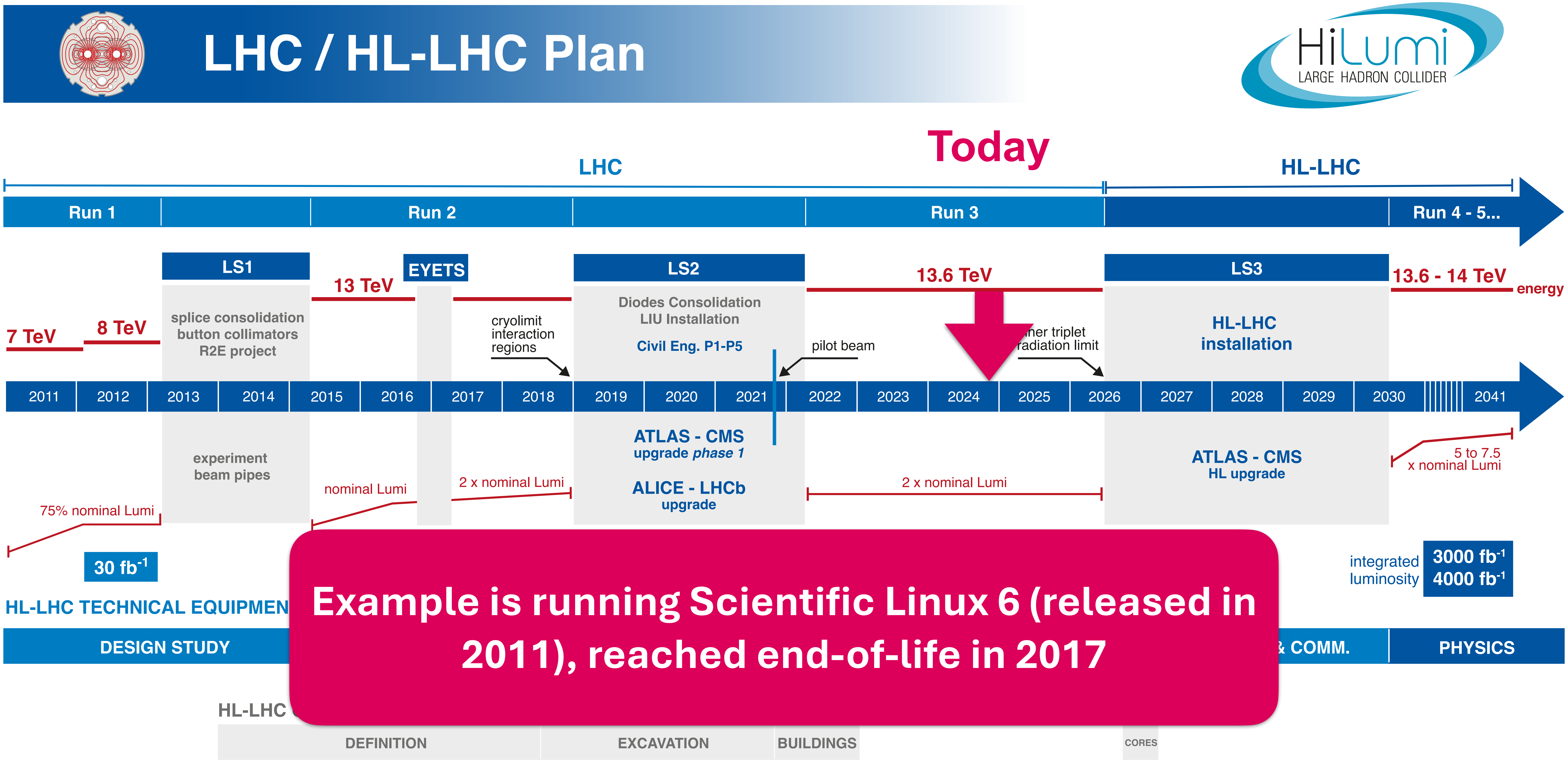
Steps towards Reusable Analyses

1. Capture software	2. Capture commands	3. Capture workflow
Individual analysis stages in an executable way (including all dependencies)	How to run the captured software?	How to connect the individual analysis steps?

Capturing analysis code almost trivial today

Requires e.g. two additional files in a GitLab repository → something you will learn this week

Why Did the Demo not Work Everywhere?



Example is running Scientific Linux 6 (released in 2011), reached end-of-life in 2017

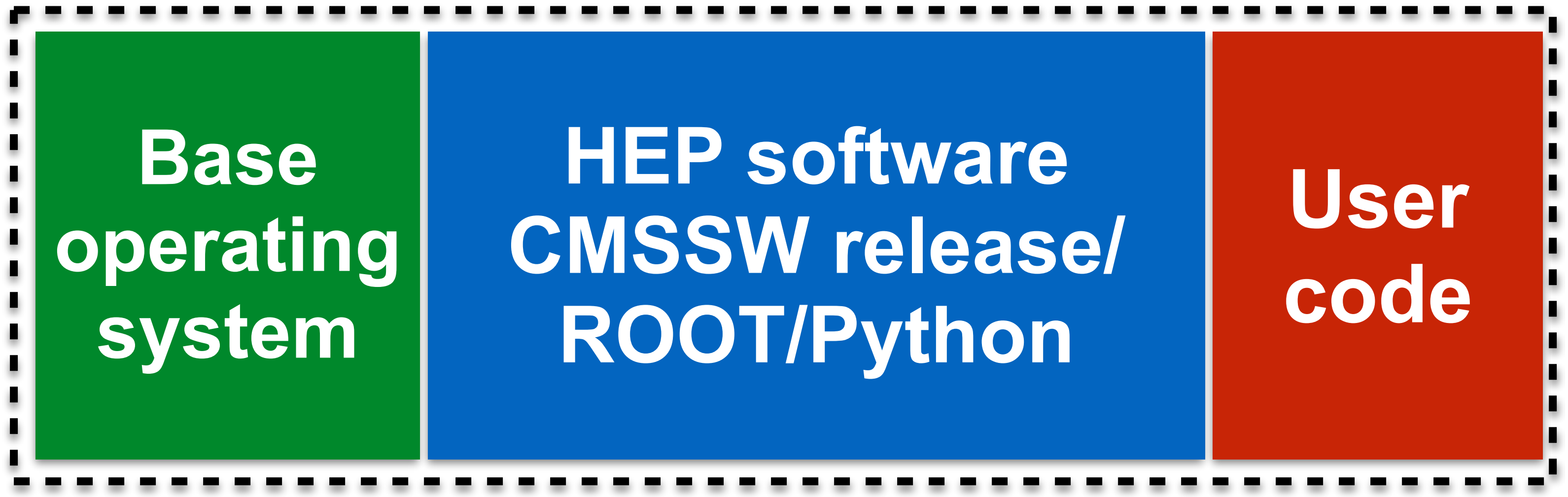
Source: https://project-hl-lhc-industry.web.cern.ch/content/project-schedule

Tooling: Software Containers

Software containers enable portability of (compiled) code

They allow e.g. to compile and run old and recent CMSSW versions on today's operating systems and processor architectures (but also your analysis code from last year)

> "Works on my *and your* machines" — from laptop to batch/grid/cloud



Advantage: **You know exactly which version of your code is running**

> Ideally built automatically using continuous integration (e.g. GitHub/GitLab)

Also useful for analysis development in general (or e.g. DAQ software, machine learning, …)

Steps towards Reusable Analyses

1.Capture software	2. Capture commands	3. Capture workflow
Individual analysis stages in an executable way (including all dependencies)	How to run the captured software?	How to connect the individual analysis steps?

Capturing analysis code almost trivial today

Requires e.g. two additional files in a GitLab repository → something you will learn this week

Once commands have been captured, can run individual analysis steps

Steps towards Reusable Analyses

1.Capture software	2. Capture commands	3. Capture workflow
Individual analysis stages in an executable way (including all dependencies)	How to run the captured software?	How to connect the individual analysis steps?

Capturing analysis code almost trivial today

Requires e.g. two additional files in a GitLab repository → something you will learn this week

Once commands have been captured, can run individual analysis steps

Capturing the workflow can be achieved in various ways

Several tools exist, e.g. SnakeMake (available in REANA)

Towards Common Analysis Tools in CMS

CMS established a new Common Analysis Tools (CAT) group at the end of 2022

This group is now working with various groups in CMS towards improved data processing tools, analysis workflows and their preservation as well as statistical inference tools (and much more)

Open Science in Action

The CMS Collaboration recently released the full statistical model ("set of likelihood functions") of the measurements that contributed to establishing the **discovery of the Higgs boson** in 2012 including the required software

doi.org/10.17181/c2948-e8875

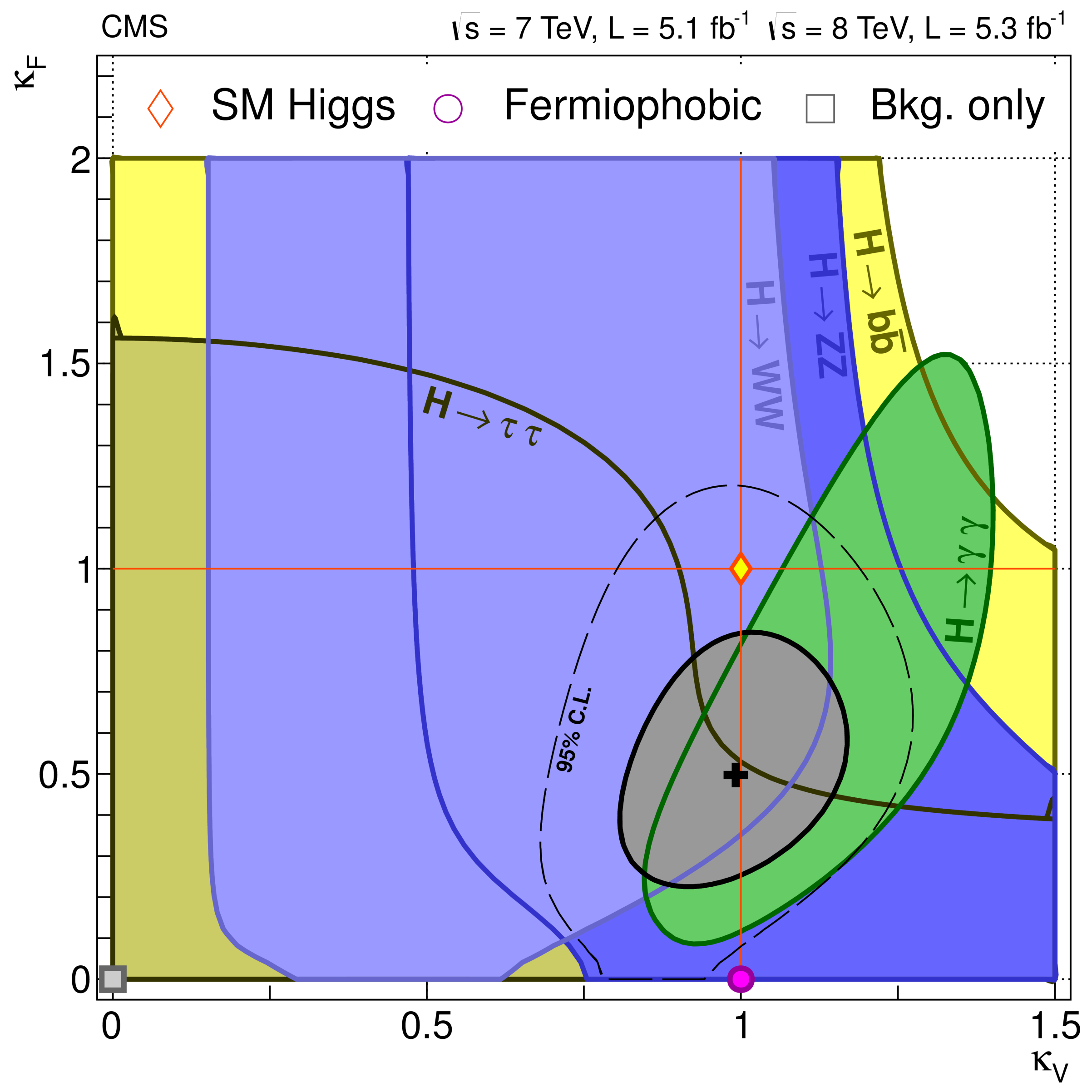
opendata CERN

CDS

HEPData

CMS Combine

FOSS software



SCOAP³

Open-Access Publications

reana

Reusable physics analyses

Already being used outside CMS!

More information at https://openscience.cern/index.php/node/435 and https://cms.cern/news/cms-commitment-open-science-takes-next-step

Towards Analysis Preservation and Reusability

The new CMS CAT group works towards closing the gap in analysis preservation and reusability

However, **analysts need to be part of this change**

Lessons Learned

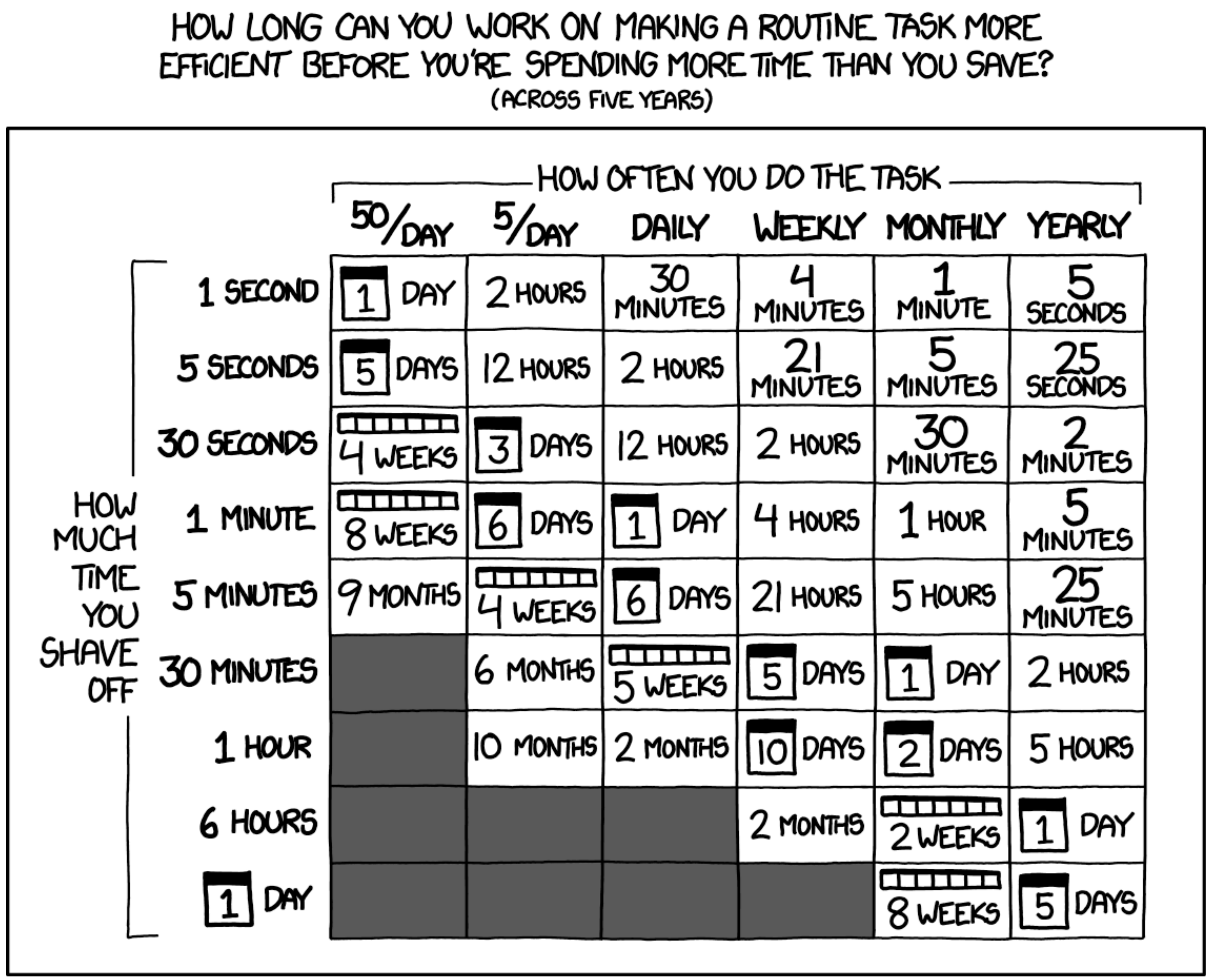
Large collaborations move slowly, **policies help enforce standards — grassroots initiatives can make a difference**

Try to use **clear and understandable naming**, e.g. for data sets

Make **recurring/useful workflows** a **routine**/ automate them

Use **version control** — for collaboration with others ideally also have **tests**

Your computing environment might change within a few months time — **software containers provide portability**



	50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
1 HOUR		10 MONTHS	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
6 HOURS				2 MONTHS	2 WEEKS	1 DAY
1 DAY					8 WEEKS	5 DAYS

Source: https://xkcd.com/1205/

On to You!

CERN has its first Open Science Policy plus an Implementation Plan

- Openly available and meant as "inspiration" for other institutions

The LHC experiments are making an effort to preserve larger parts of the physics analysis chain

- Whether this is successful will depend a lot on the analysts themselves
- There are several examples how Open Science made even internal collaboration better

I hope you will see the advantages of a more structured/systematic approach — this week's training will provide with the skills required

- Your future self will probably thank you

PSI

New: CERN Open Science Policy

Captures current practice and states vision across multiple Open Science domains:

Open Access to Publications
Open Research Data
Open Software
Open Hardware
Citizen Science

Research Integrity, Reuse & Reproducibility
Infrastructure for Open Science
Research Assessment & Evaluation
Education, Training & Outreach

v1.0 released Oct 2022: https://cds.cern.ch/record/2835057

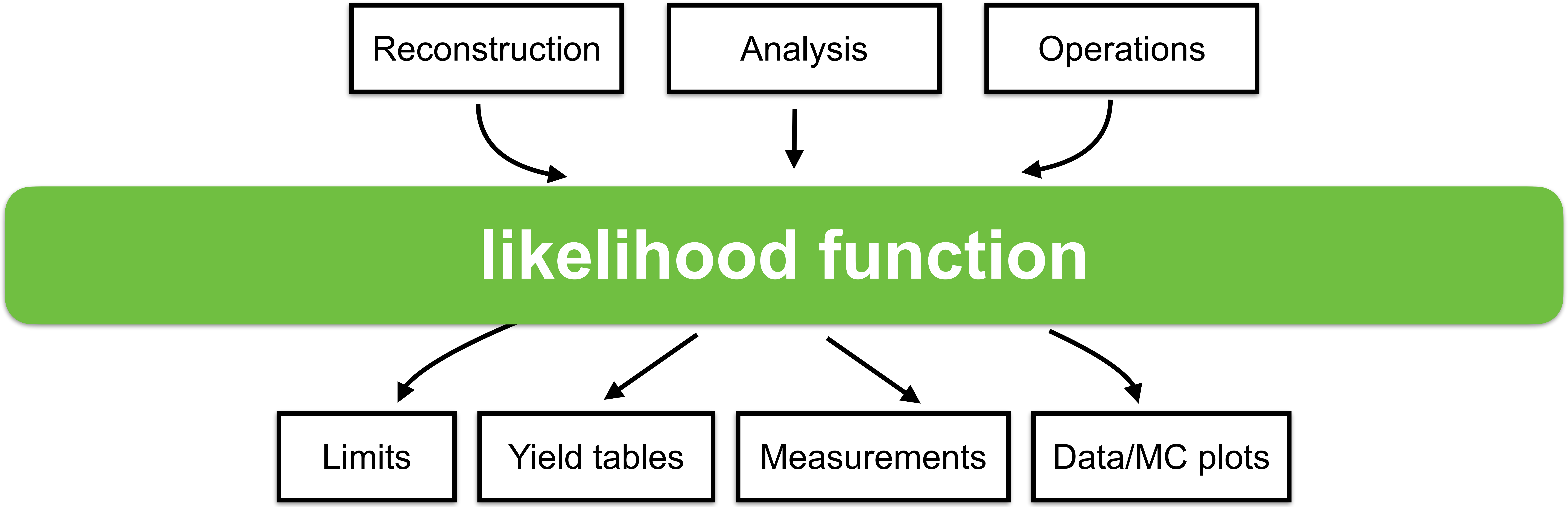
For more information, see https://openscience.cern/

> Have a look at the implementation plan!
> Data Management Plan template: https://openscience.cern/index.php/DMP

Publishing Full Likelihoods

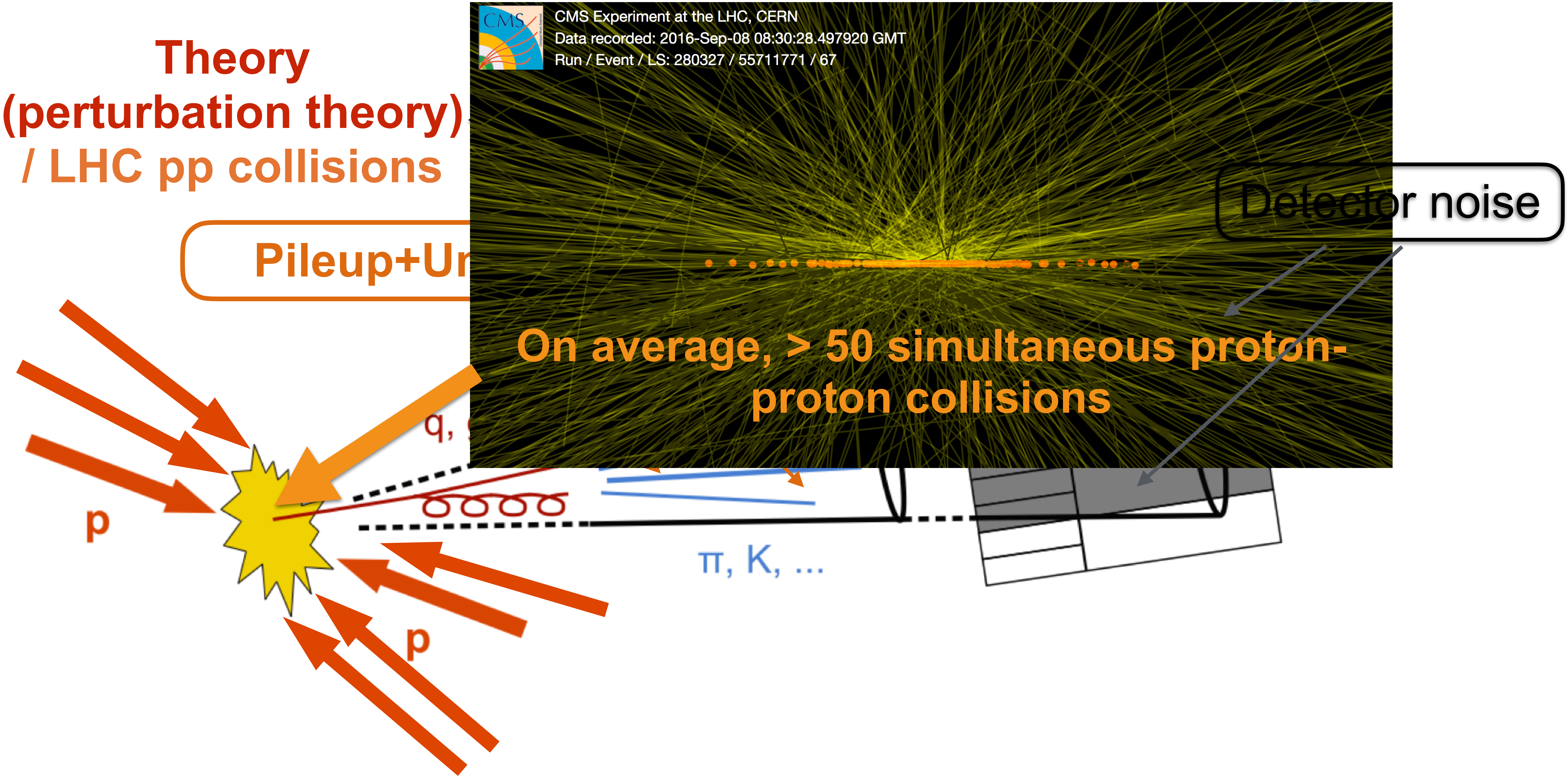
The **likelihood function** is a particularly special data product

- Small, information-dense, overall summary of the analysis
- Almost every analysis decision is reflected in the likelihood



Expect to see more full likelihoods from CMS in the next few months

Collider data is complex



Further challenges in high-energy physics data analysis

Collaborations make huge internal review effort (months to years) to **ensure accurate interpretation of the data**

- False claims (also from OD users) could risk erosion of public trust

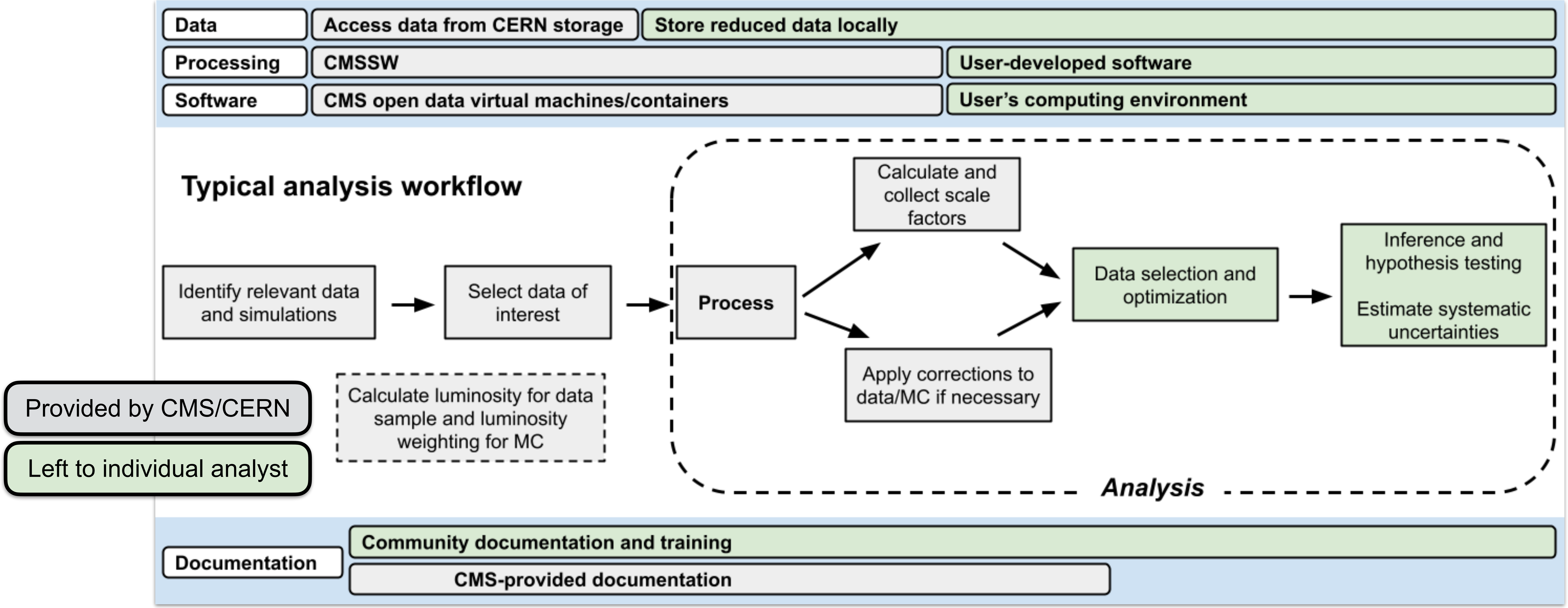
Small deviations can make a big difference

- A few events could mean a discovery

Physics objects definitions are analysis-dependent

- An electron in one analysis might not be one in another due to different reconstruction algorithms used

Analysis flowchart



The analysis part usually takes a lot of iterations

Analysing Collider Data is very Challenging

We can **only store 0.05‰ of the collisions** (1 in 20,000 events or 2,000 events per second)

- A multi-stage trigger system selects events of interest — this bias needs to be taken into account when performing an analysis

A raw event has the size of about 2 megabytes

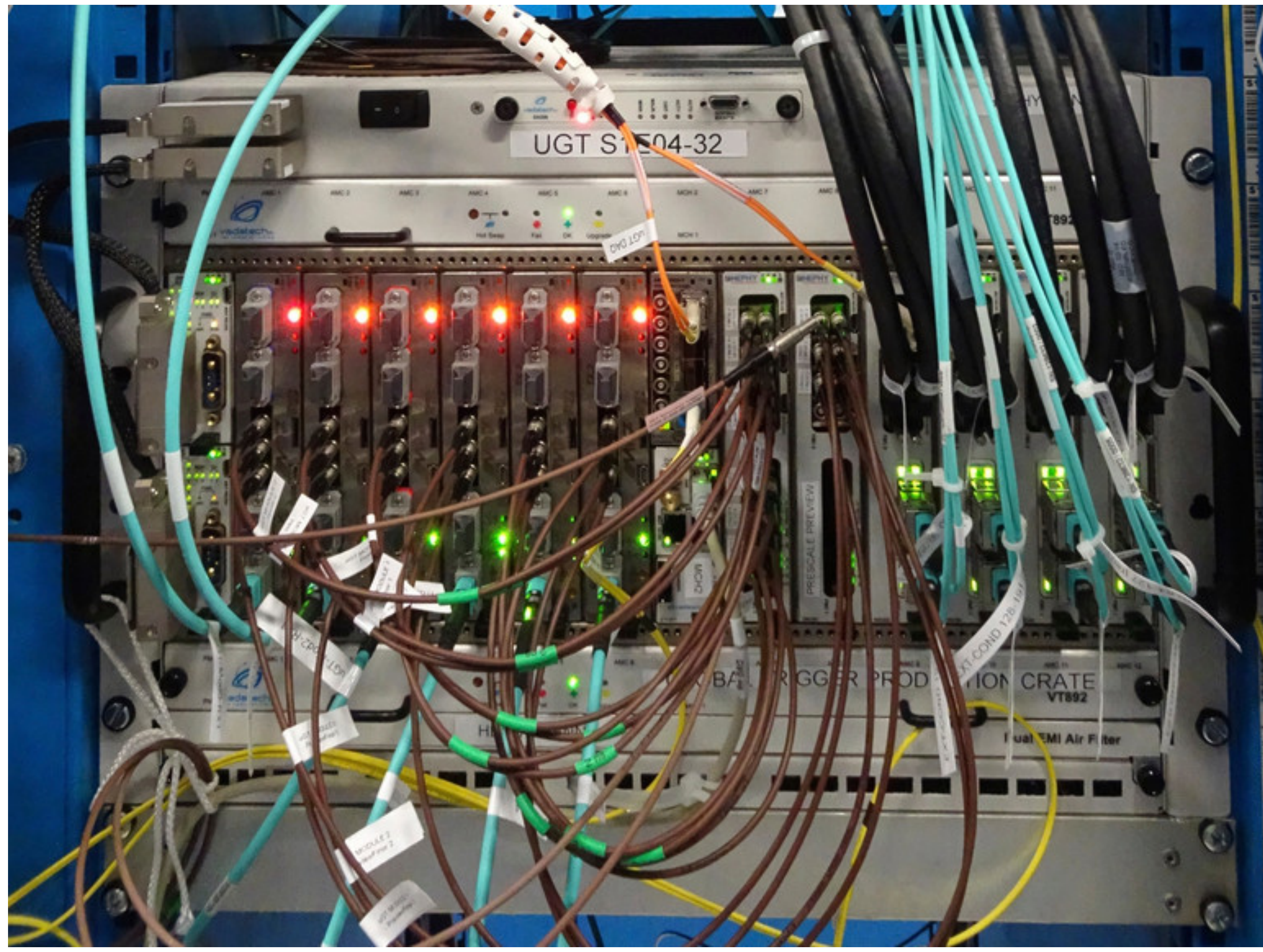
- We have recorded tens of billions of events, and simulated even more
- **Size can be reduced at the cost of information loss** — expertise required
- For Run 1, we have largely released “Analysis object data” (500 kB/event)
- For Run 2 (2015+), we release MiniAODs and NanoAODs (2 kB/event)

Billions of events need **significant computing power** for processing

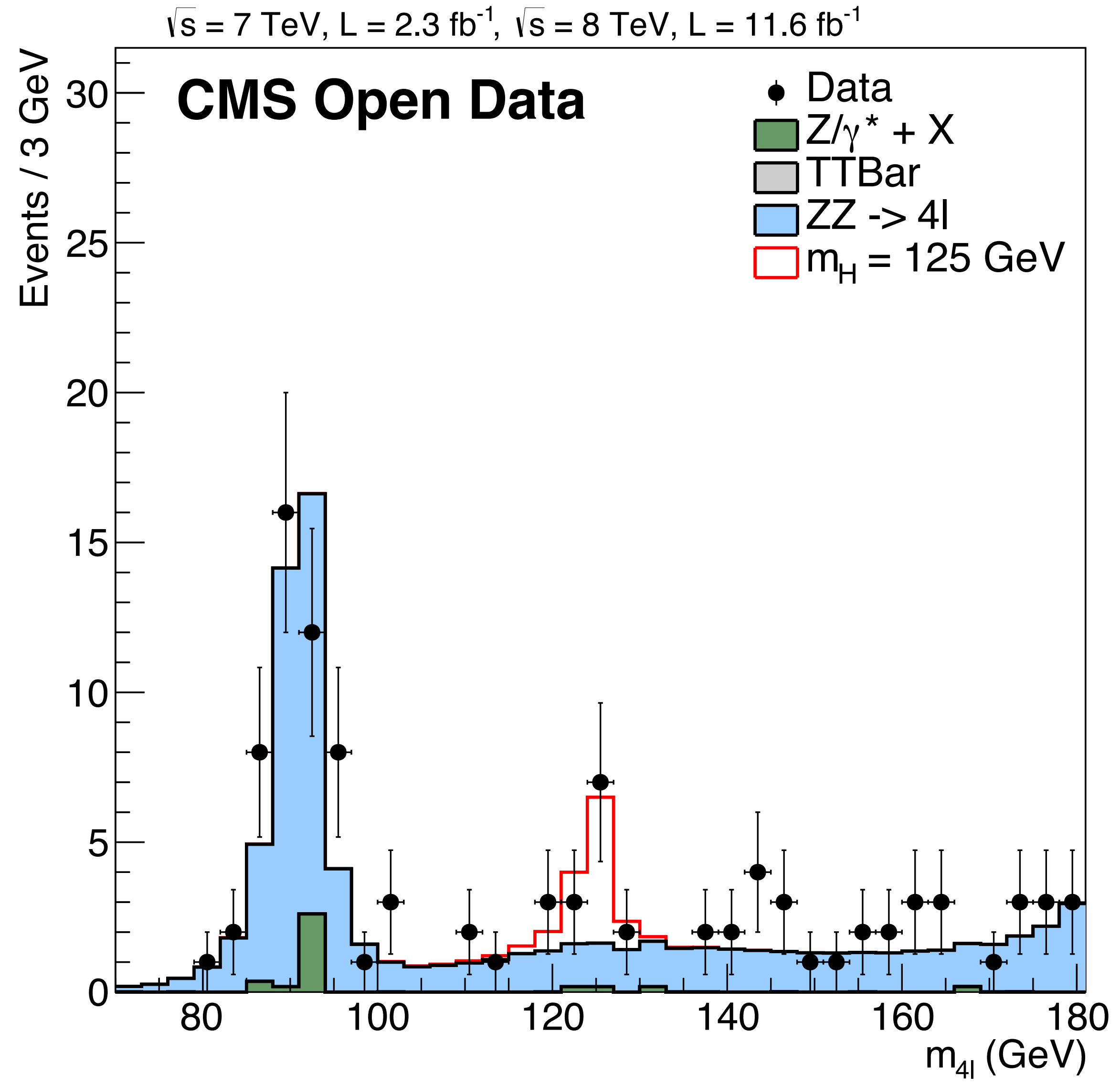
A complete physics analysis needs to take **dozens of systematic uncertainties** into account

- Understanding the relevance of individual uncertainties needs expertise

Statistical interpretation needs particular care



Computational Challenges



DOI:10.7483/OPENDATA.CMS.JKB8.RR42

We provide simplified analysis examples to lower the threshold to get started

- Pro: users can obtain a result/plot rather quickly
- Contra: these are usually far from realistic

At least the first step of the analysis chain requires substantial computing resources, ideally high-throughput batch processing systems

- Data sets can be processed in an “embarrassingly parallel” way
- We provide examples/tutorials on using public cloud resources

kubernetes

Simulation of new processes needs CMSSW

- Parts of the software are more than a decade old ➜ interfacing can be difficult

Keeping up

When developing examples, we now aim to use open tools combined with container technologies for automatic and regular validation

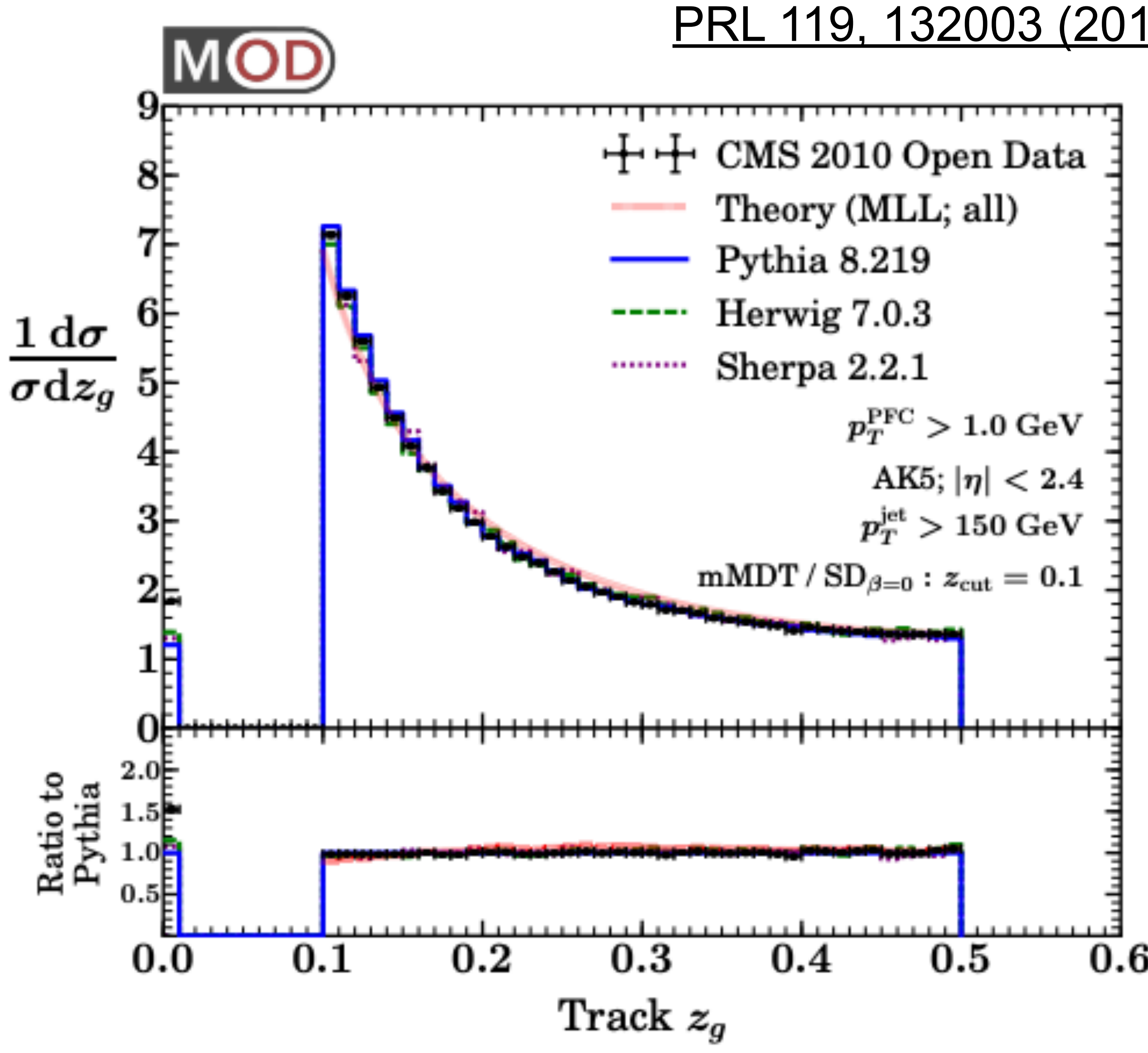
- Continuous integration using CERN's GitLab installation
- Simpler examples also run as GitHub actions

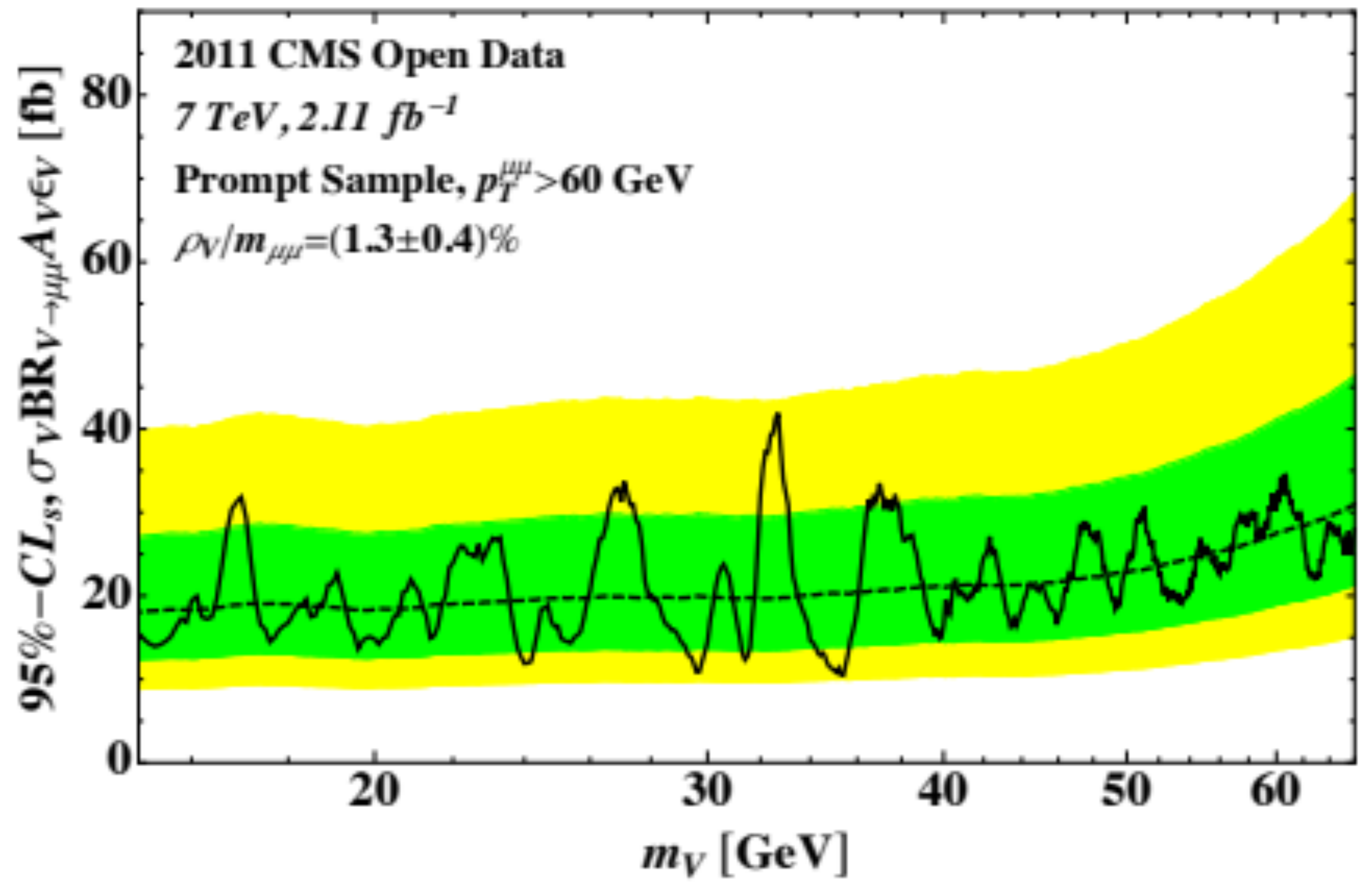
For easier usability, we provide examples on how get out of the HEP-specific software tool chain to industry standard tools

By now, CMS Open Data have been used for both actual physics results and also several computing-related projects

Eventually, the data might be used to unveil hidden physics!

PRL 119, 132003 (2017)





Phys. Rev. D 100, 015021 (2019)

CERN Code of Conduct: Creativity & Professionalism

As CERN contributors, we

- >Follow developments within our domain.
- >Use our professional experience in a constructive manner.
- >Contribute to the evolution of CERN by committing to **sharing our knowledge**.
- >**Share** with internal parties **any information that could benefit them in their work**.
- >Are open to new ideas and approaches.
- >Adopt alternative outlooks in order to generate new thoughts and concepts.
- >Conduct our work in a **structured way** to **enhance knowledge transfer and continuity**.

https://cds.cern.ch/record/2240689/