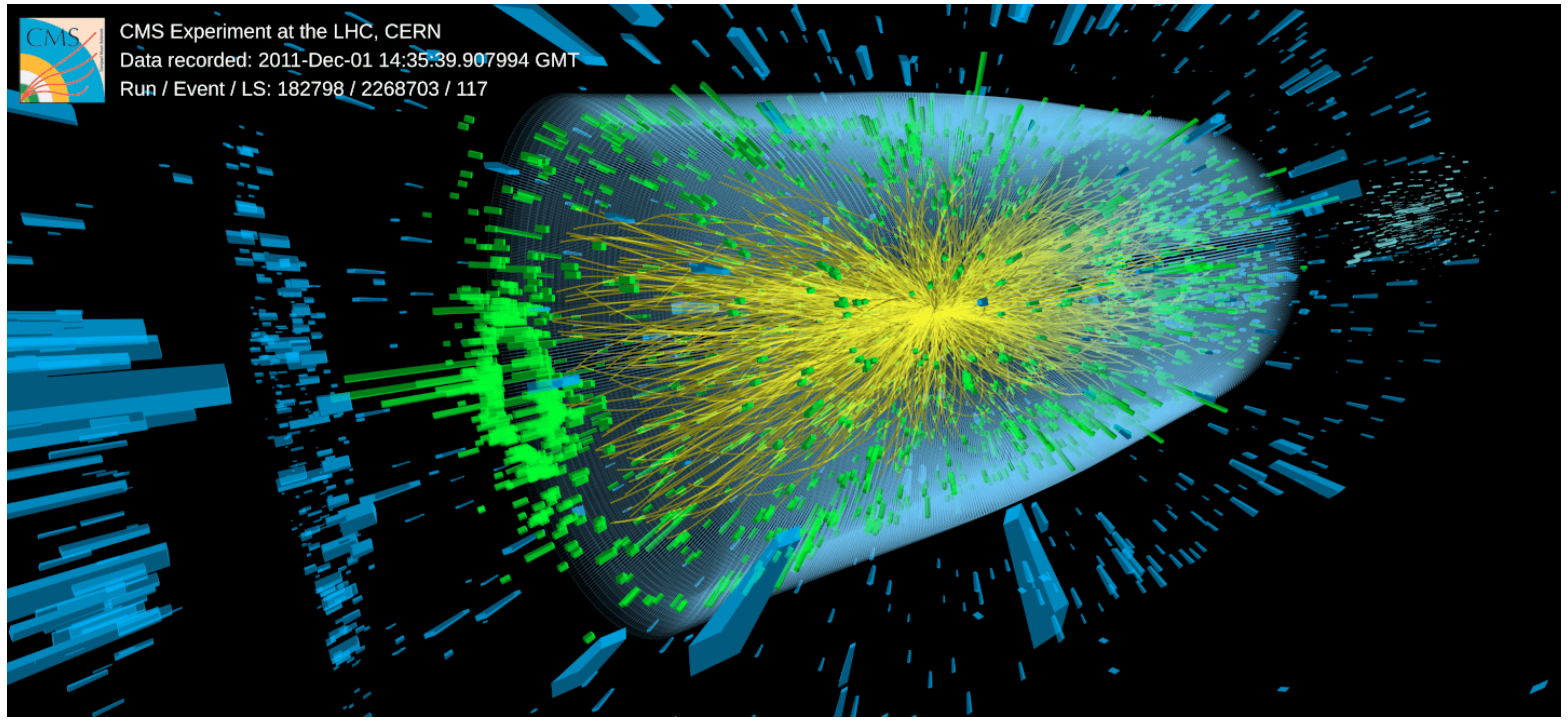
Analysis preservation at CMS

Current status and plans

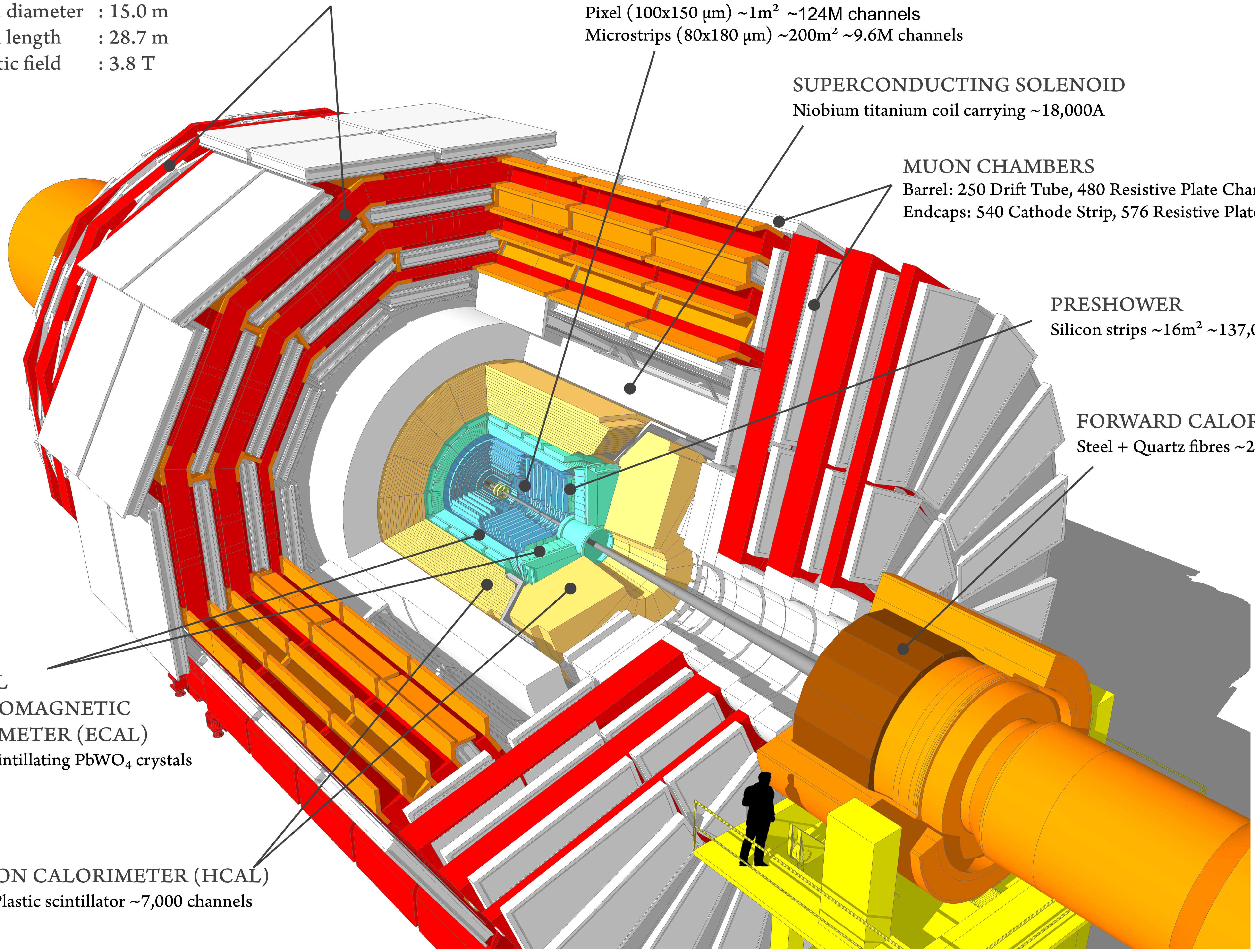

CMS Experiment at the LHC, CERN
Data recorded: 2011-Dec-01 14:35:39.907994 GMT
Run / Event / LS: 182798 / 2268703 / 117

Clemens Lange (Paul Scherrer Institute PSI)

HSF Analysis Preservation Training

16th January 2023

The CMS experiment



CMS DETECTOR

Total weight	: 14,000 tonnes
Overall diameter	: 15.0 m
Overall length	: 28.7 m
Magnetic field	: 3.8 T

STEEL RETURN YOKE
12,500 tonnes

SILICON TRACKERS
Pixel (100x150 μm) ~1m^2 ~124M channels
Microstrips (80x180 μm) ~200m^2 ~9.6M channels

SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A

MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers

PRESHOWER
Silicon strips ~16m^2 ~137,000 channels

FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels

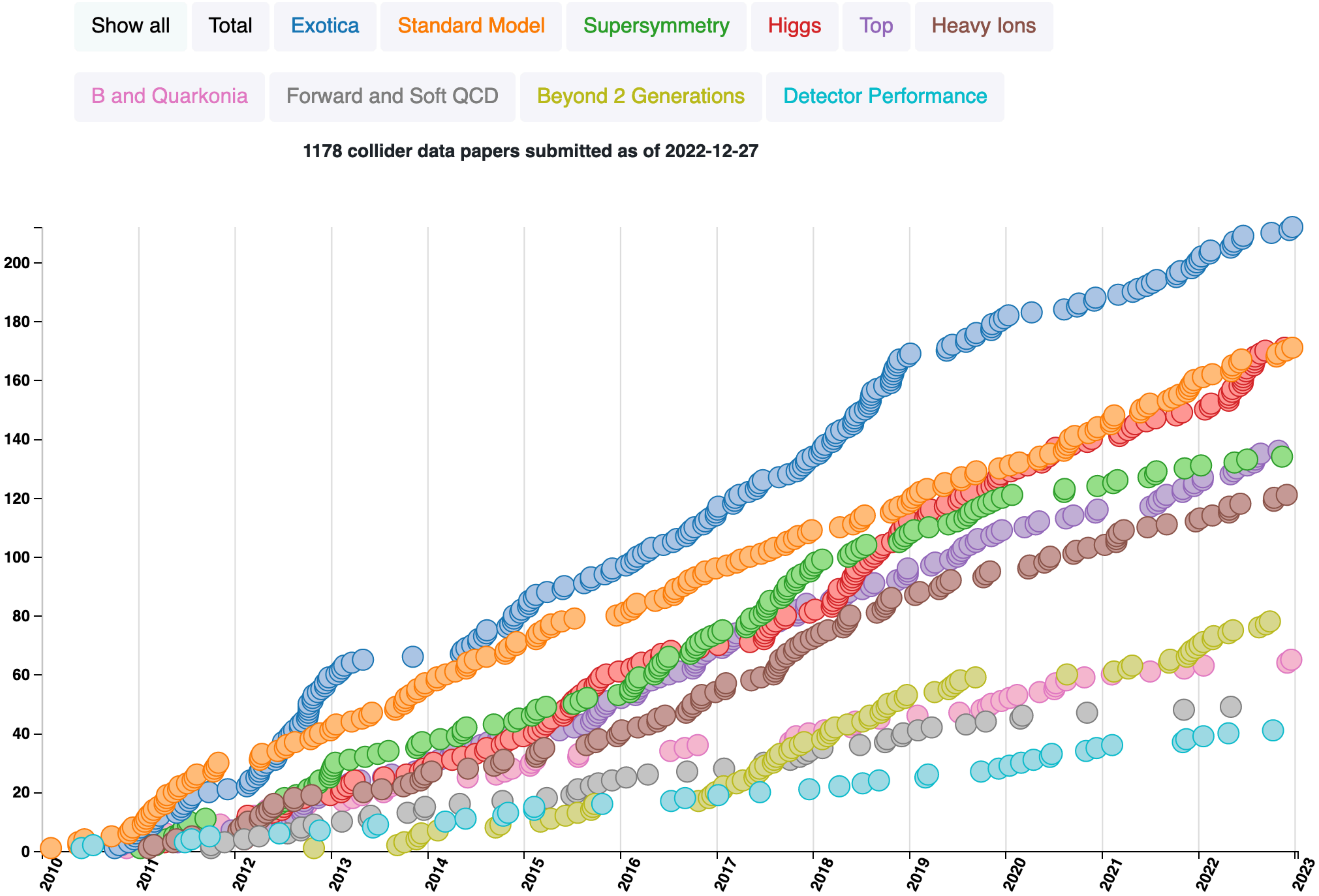
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating $PbWO_4$ crystals

HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

- 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

Producing huge amounts of data over decades!

CMS publications

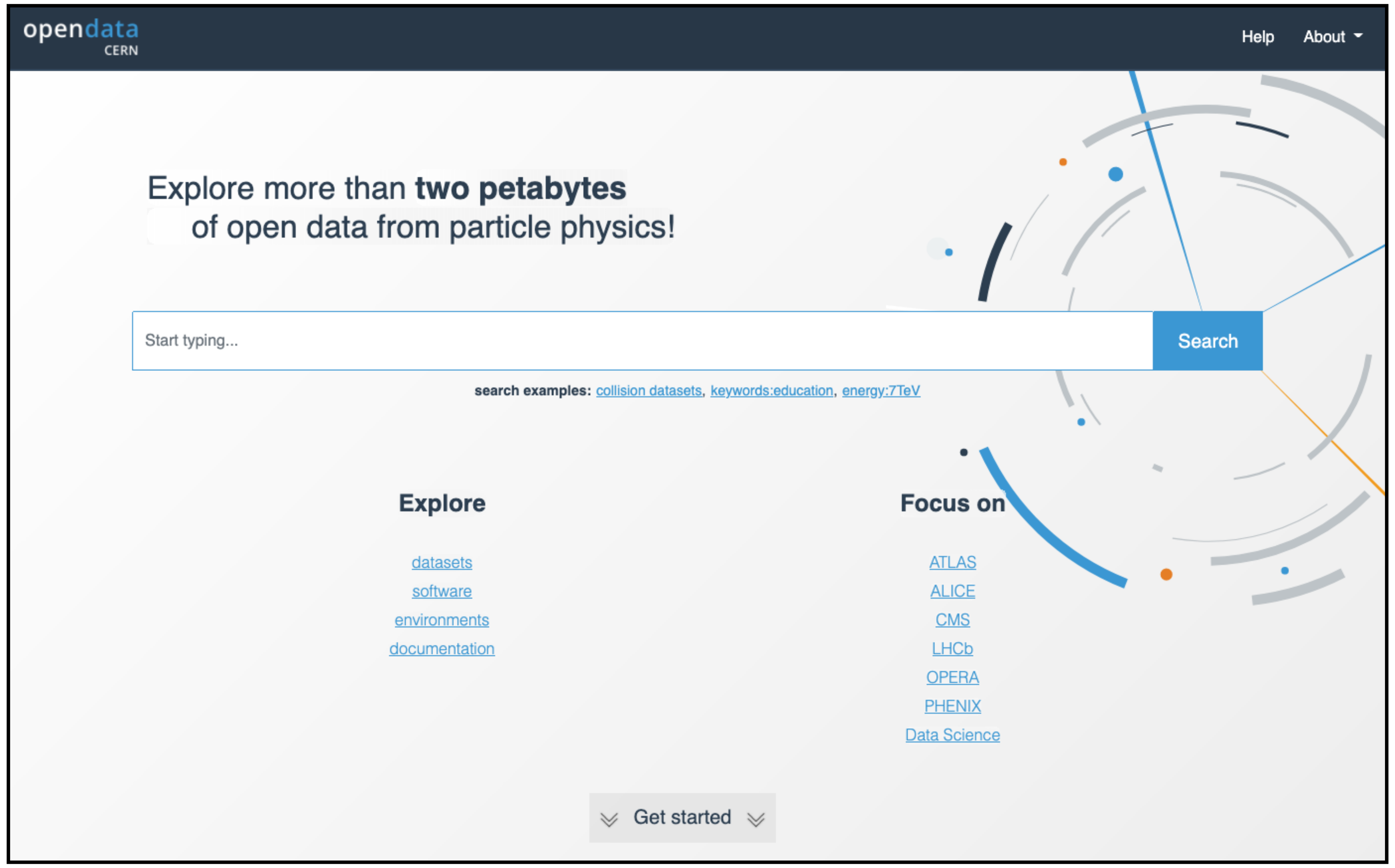


1178 collider data papers submitted as of 2022-12-27

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

Physics results

- Since 2008, >1000 peer-reviewed papers published
  - Among them the discovery of the Higgs boson (No. 183)
- All published under open access (since 2014 under SCOAP^3)
  - Preprints available on arXiv
  - Tabulated results largely available on HEPData portal
- Since 2014, have released > 3 petabytes of open data available on the CERN Open Data Portal
  - Entire Run-1 + 2015 data sets



Explore more than **two petabytes** of open data from particle physics!

Start typing...

search examples: collision datasets, keywords:education, energy:7TeV

Explore

datasets
software
environments
documentation

Focus on

ATLAS
ALICE
CMS
LHCb
OPERA
PHENIX
Data Science

CMS 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 make scientific studies**
- Data and simulation will start to be released approximately five years after collection (50%)
  - Released under the Creative Commons CC0 waiver
  - Full dataset by the close of the experiment

higher computational effort ↓

- Level 1: Open access publication and additional numerical data
- Level 2: Simplified data for Outreach and Education
- **Level 3**: Reconstructed data and the software to analyse them
- Level 4: Raw data, and the software to reconstruct and analyse them

Challenges of providing open data

Data: available ≠ usable

Open Data needs to be FAIR:

- **F**indable ➜ CERN Open Data Portal records
- **A**ccessible ➜ reliable storage and access technology
- **I**nteroperable ➜ provide good documentation, avoid jargon
- **R**eusable ➜ preserve software (and hardware to run it if needed), data provenance, workflows

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

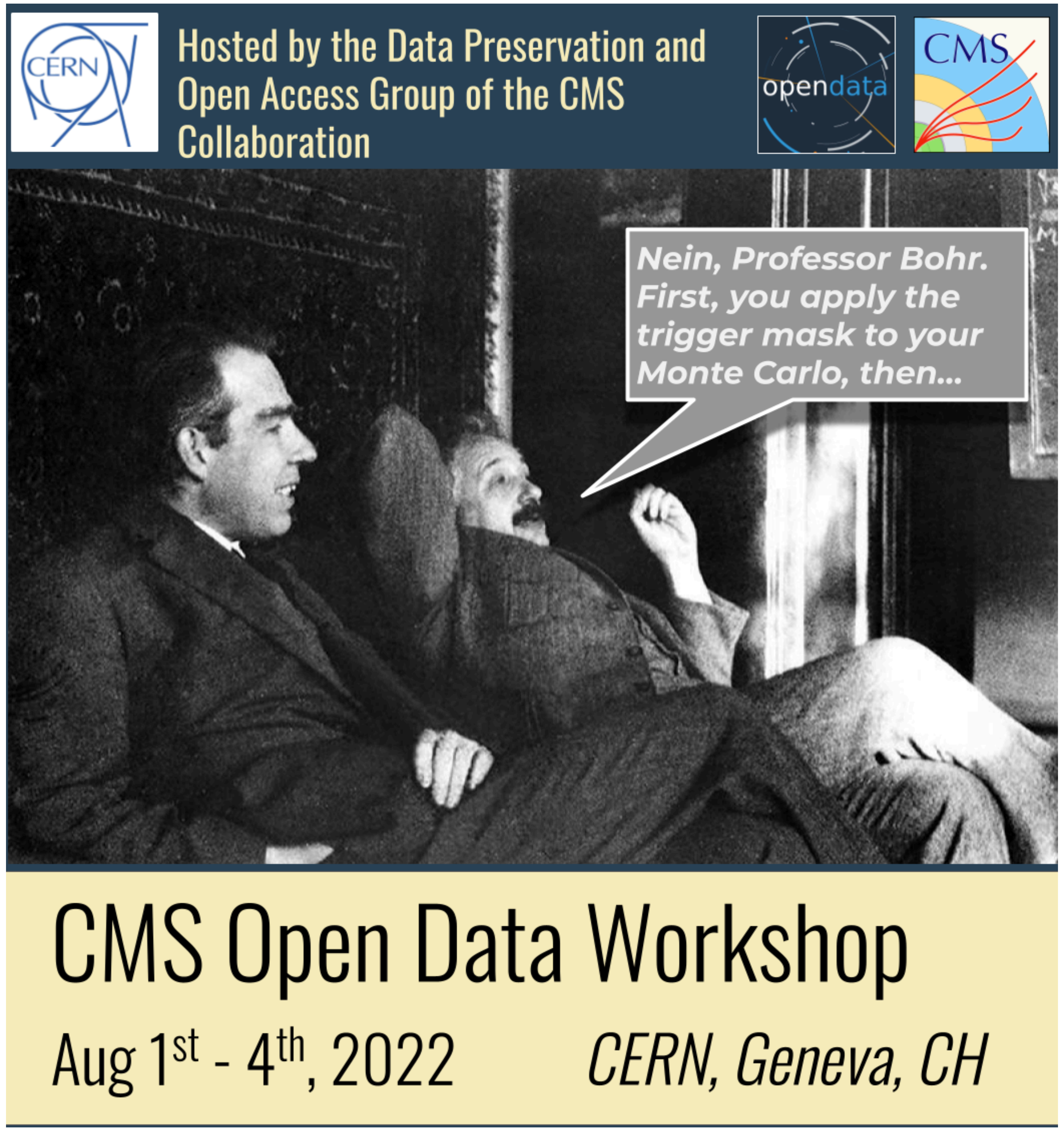
v1.0 released Oct 2022: https://cds.cern.ch/record/2835057

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

Addressing the challenges

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


CERN
opendata
CMS
Nein, Professor Bohr. First, you apply the trigger mask to your Monte Carlo, then...

Hosted by the Data Preservation and Open Access Group of the CMS Collaboration

CMS Open Data Workshop

Aug 1st - 4th, 2022 *CERN, Geneva, CH*

Demo time

Let’s rediscover the Higgs boson in 5 minutes

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

The analysis disconnect

- Open Science is also a sociological challenge
- The main reason we provide only analysis examples is that we do not (yet) systematically preserve the actual analyses

Theory model/ LHE file

CMS GEN-SIM step

CMS DIGI step

CMS RECO step

geometry/ conditions database

CMS specific

Experiment data

Reproducible

ntuplisation/selection

Statistical analysis

Analysis specific

Currently not preserved

Internal documentation

HEPData

Preserved

N
fit
signal
data
m_{jj}
PUBLISHED

Ongoing work

- Preserve code in CMS-provided repository
- Build analysis software containers automatically
- Connect analysis steps using workflow languages/engines
- Use e.g. also for reinterpretations

GitLab

workflow

reana
Reproducible research data analysis platform

Theory model/ LHE file

geometry/ conditions database

CMS GEN-SIM step

CMS DIGI step

CMS RECO step

CMS ntuple step

CMS specific

Experiment data

Reproducible

selection

Statistical analysis

Analysis specific

Internal documentation

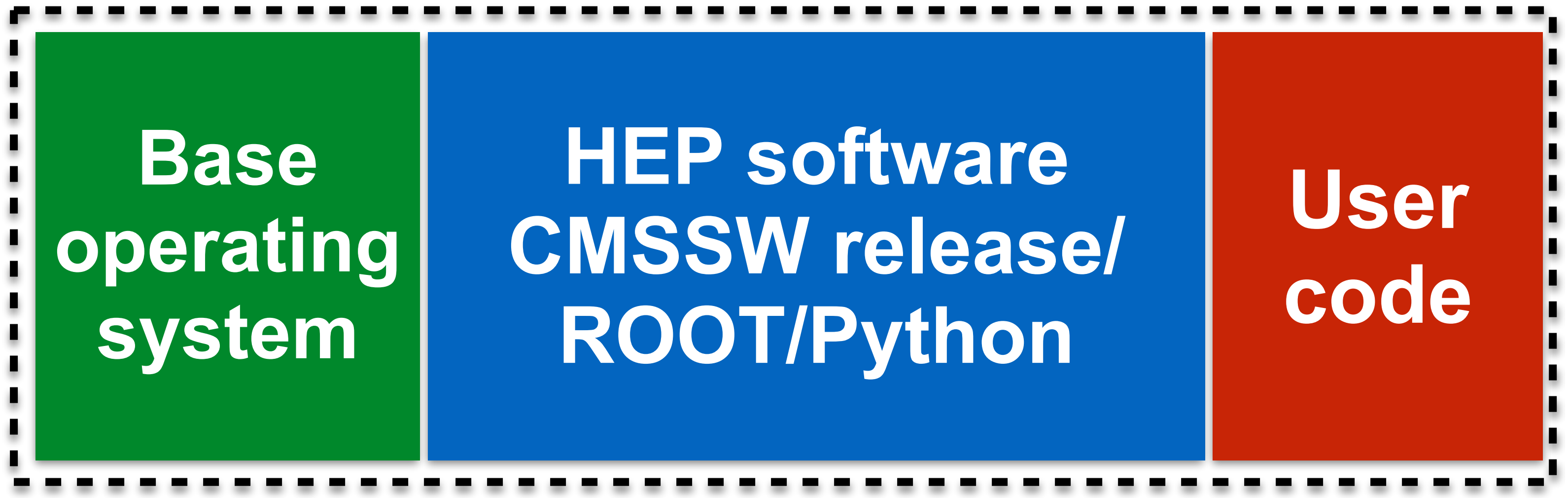
HEPData

Preserved

N
fit
signal
data
m_{jj}
PUBLISHED

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
  - “Works on my *and your* machines” — from laptop to batch/grid/cloud



Base operating system

HEP software CMSSW release/ ROOT/Python

User code

- 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, …)

Keeping things running with continuous improvements

- 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

GitLab

GitHub

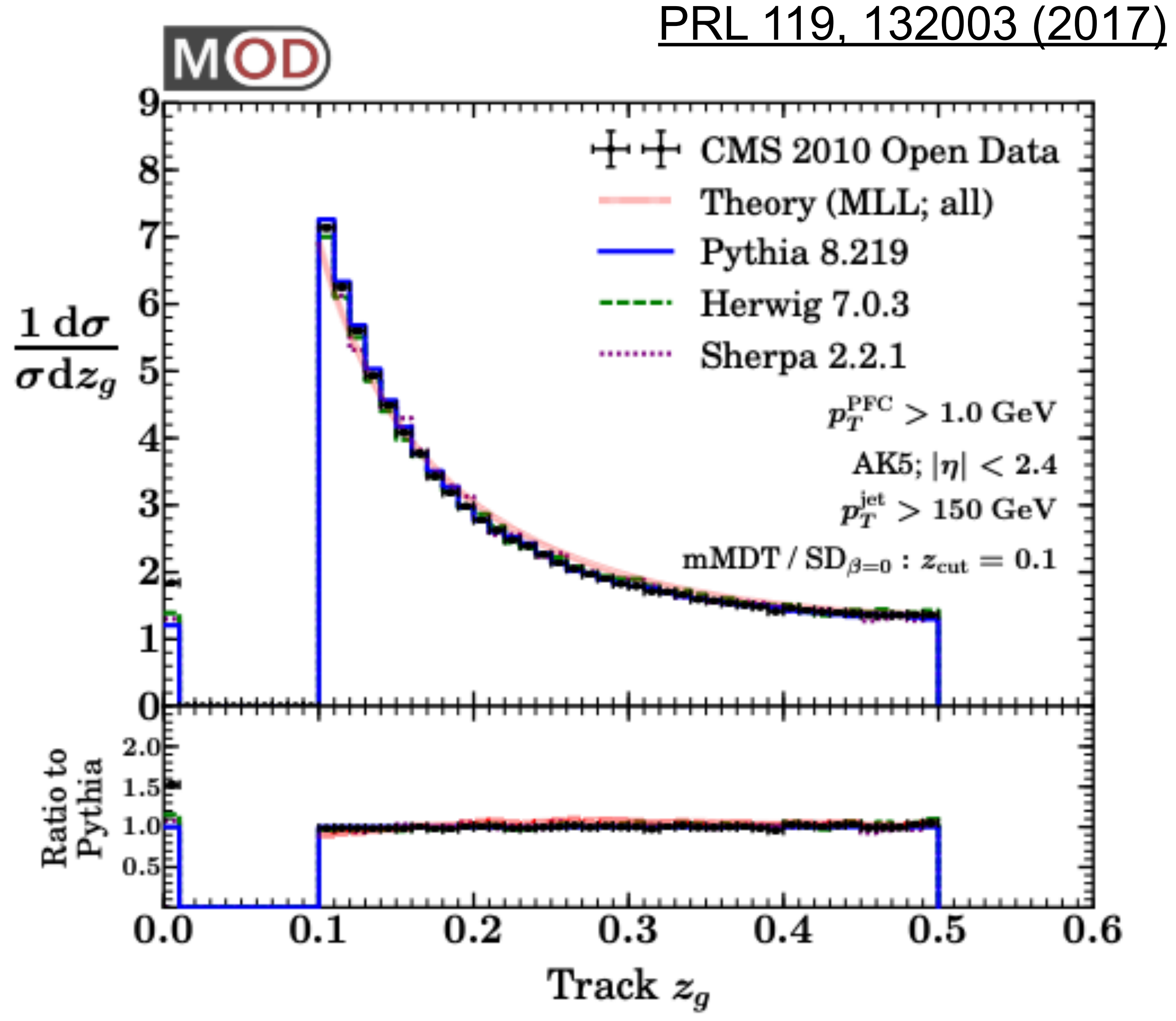
- For easier usability, we provide examples on how get out of the HEP-specific software tool chain to industry standard tools

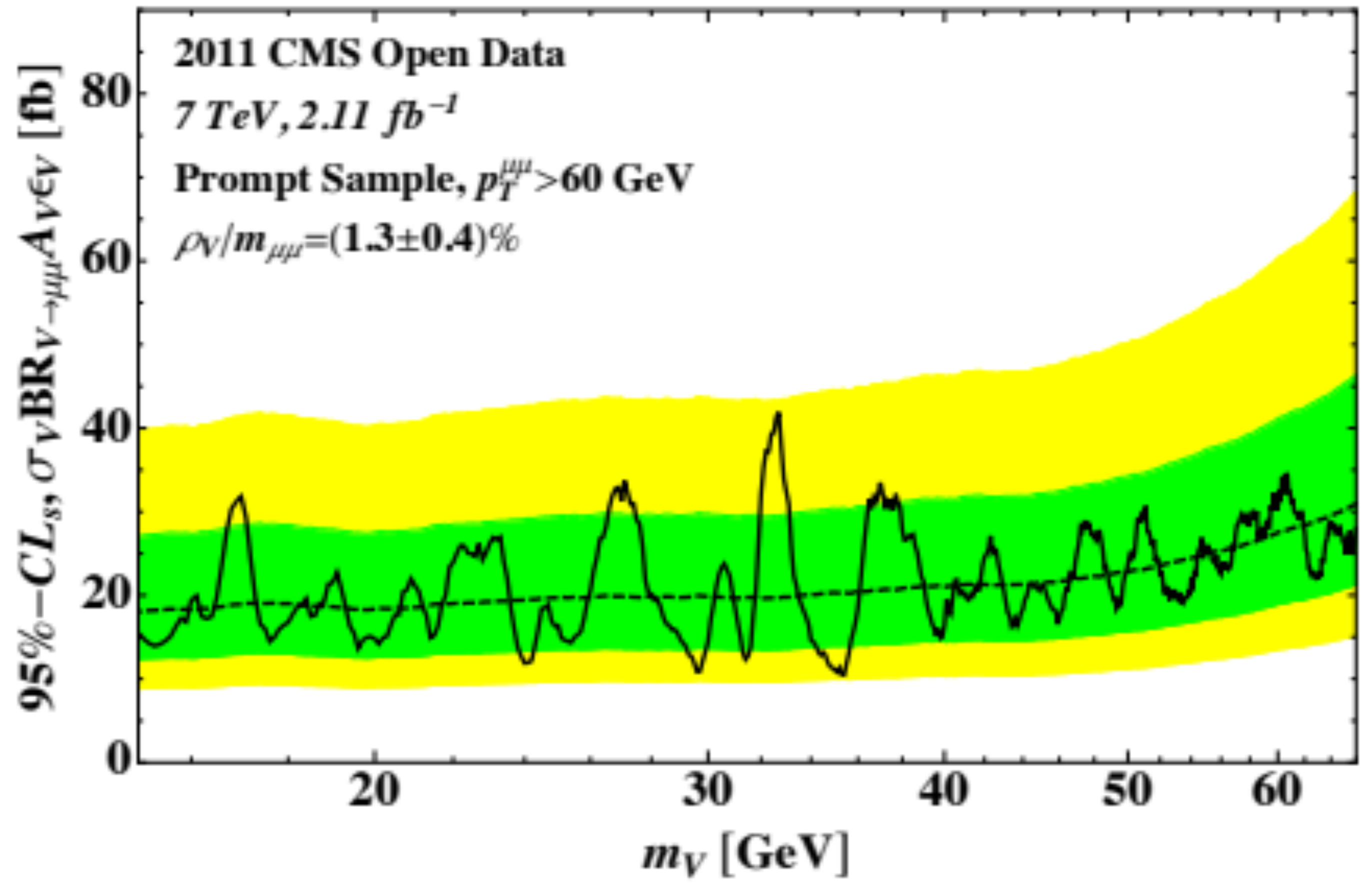
ROOT Data Analysis Framework → NumPy

- 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 for measurements we have not thought of today!

PRL 119, 132003 (2017)

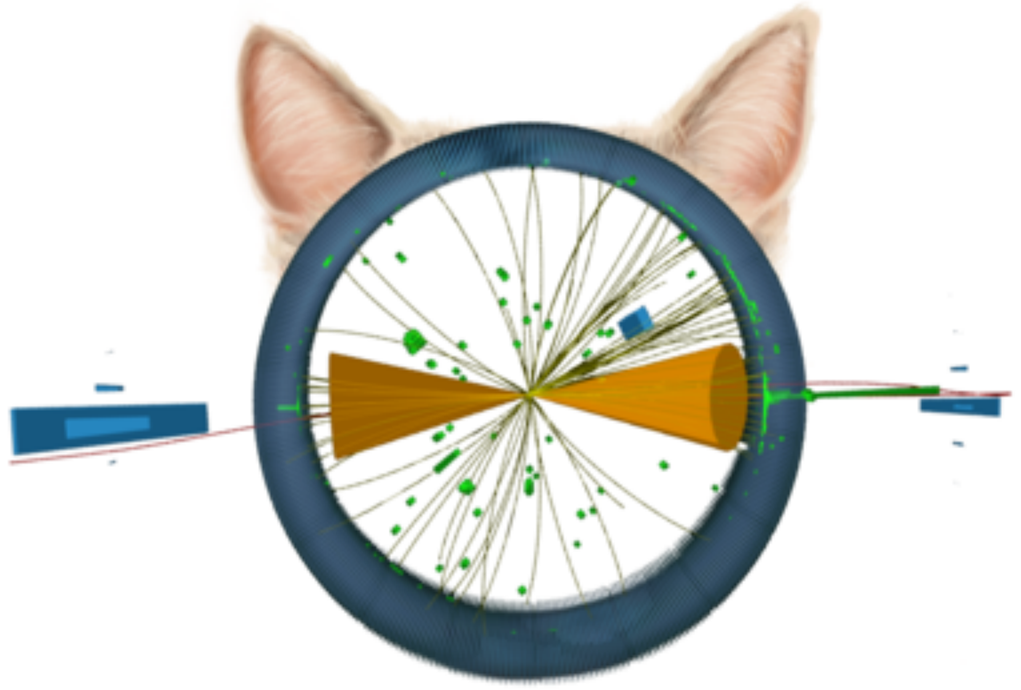




Phys. Rev. D 100, 015021 (2019)

A new group in CMS

- Recently, a new group has been formed in the CMS Physics Coordination domain: **Common Analysis Tools**
- Work is ramping up this month
- Subgroups:
  - Data processing tools → get to analysis-level quantities including corrections
  - Workflow orchestration and analysis preservation
  - Statistical interpretation tools



On to you!

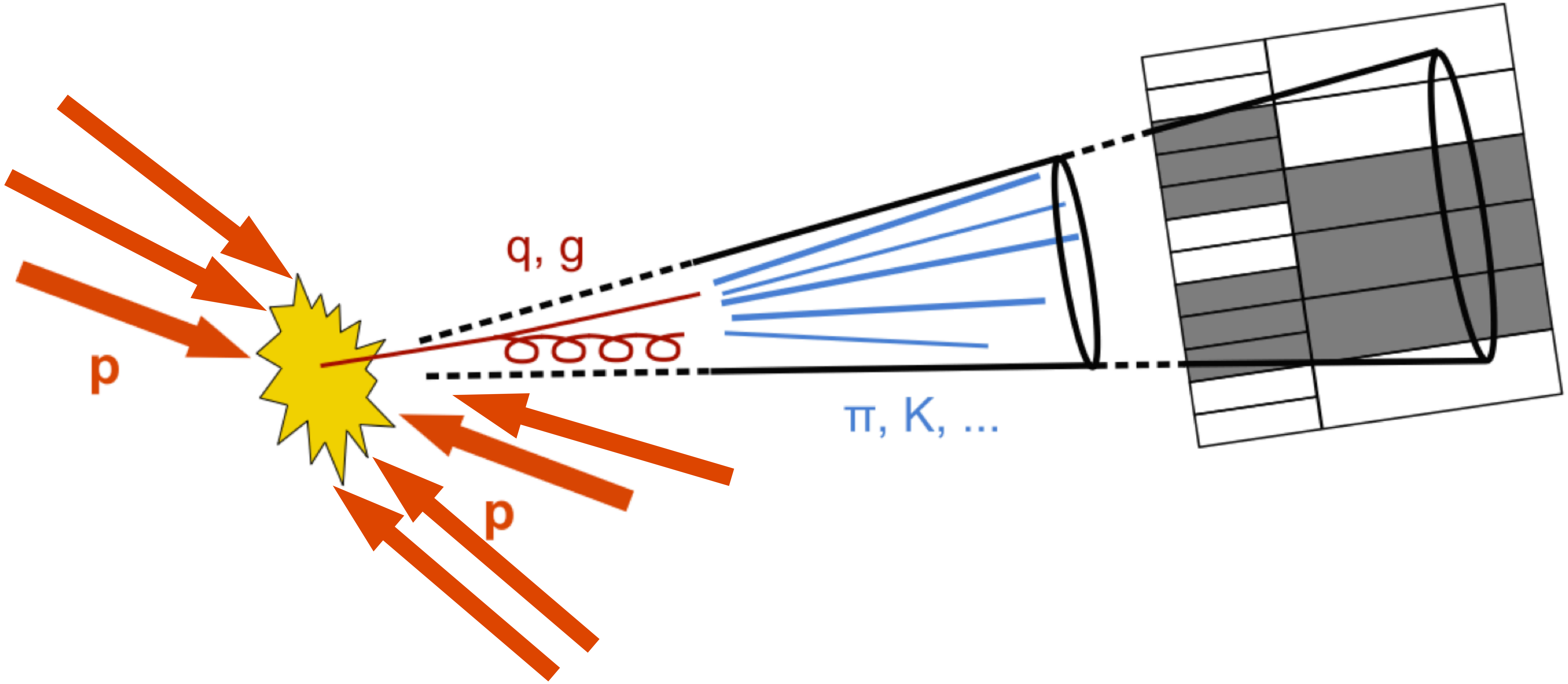
- CMS is making an effort to preserve larger parts of the physics analysis chain
  - Whether this is successful will depend a lot on the analysts themselves
- This week’s training will provide with the knowledge to perform better science
- I hope you will see the advantages of a more structured/systematic approach
  - Your future self will probably thank you

PAUL SCHERRER INSTITUT

PSI

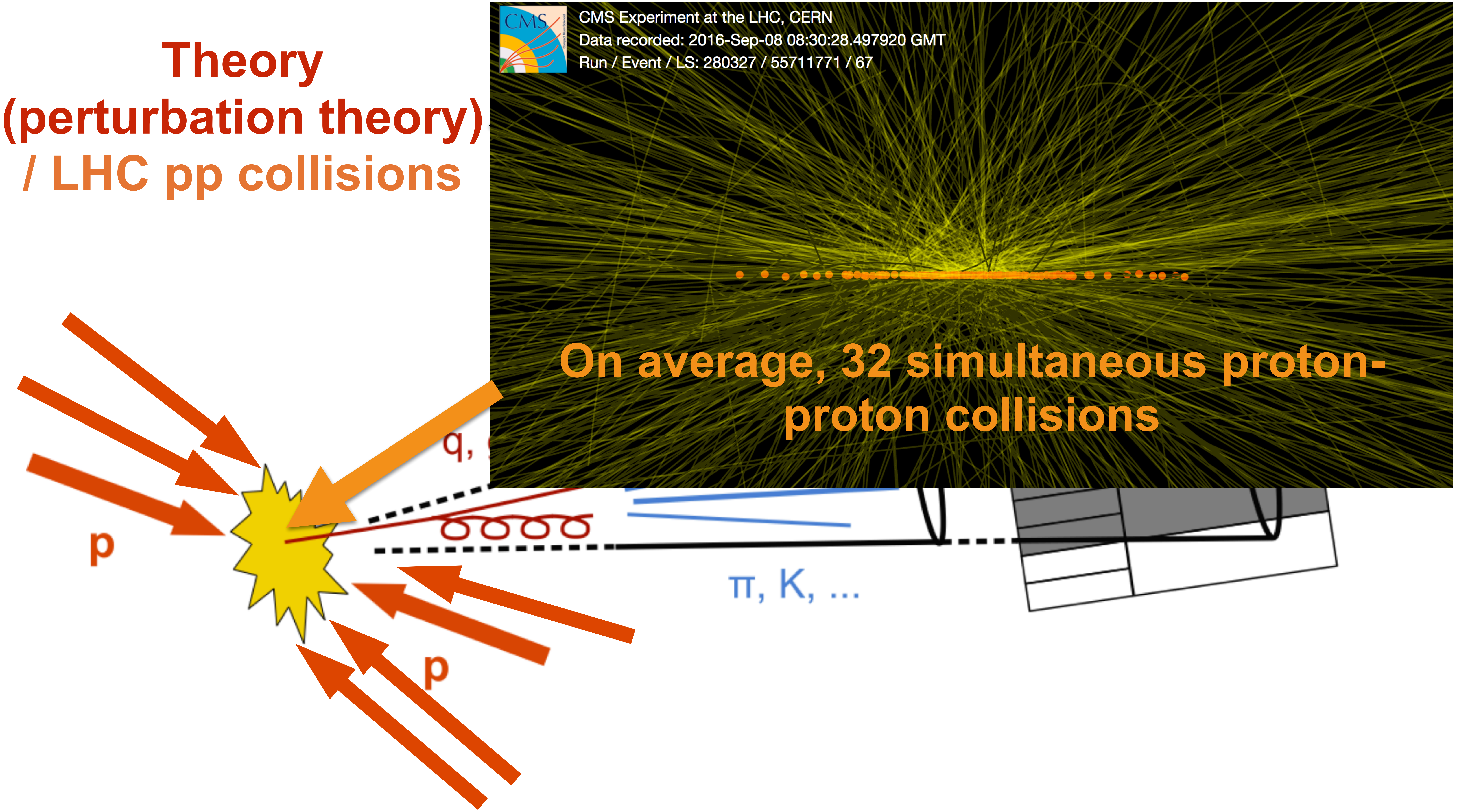
Collider data is complex

Theory (perturbation theory) / LHC pp collisions ↔ **Parton Shower + Hadronisation (non-perturbative)** ↔ **Experiment**



Collider data is complex

Theory (perturbation theory) / LHC pp collisions

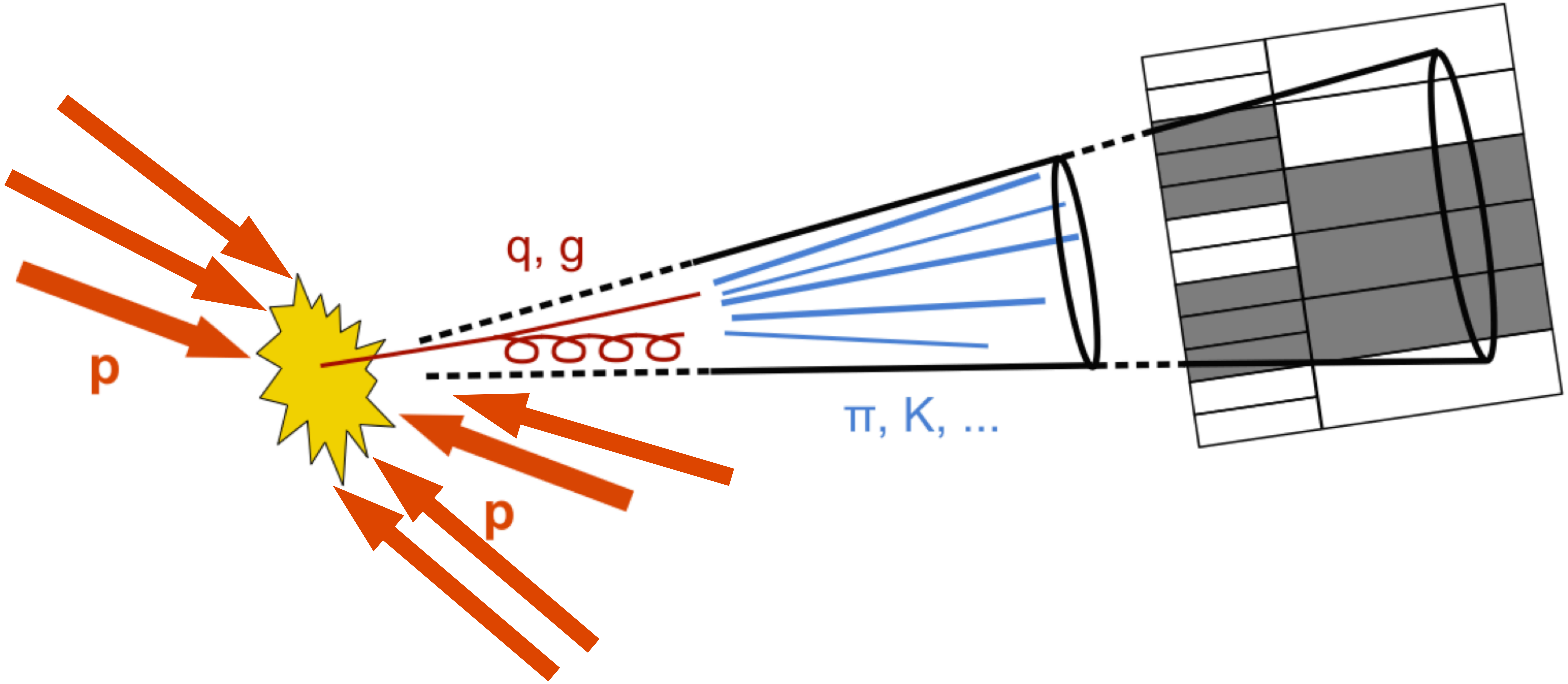


CMS Experiment at the LHC, CERN
Data recorded: 2016-Sep-08 08:30:28.497920 GMT
Run / Event / LS: 280327 / 55711771 / 67
q,
p
p
π, K, ...

On average, 32 simultaneous proton-proton collisions

Collider data is complex

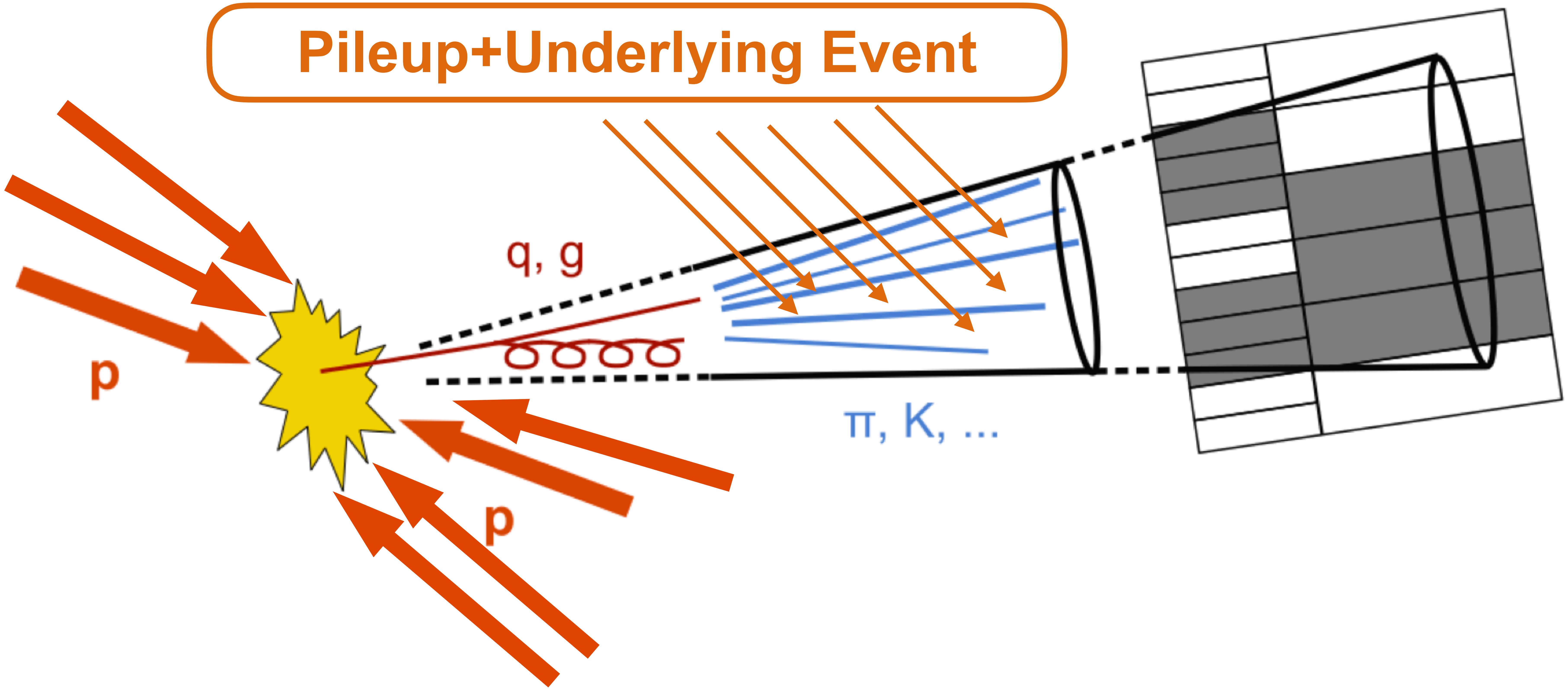
Theory (perturbation theory) / LHC pp collisions ↔ Parton Shower + Hadronisation (non-perturbative) ↔ Experiment



Collider data is complex

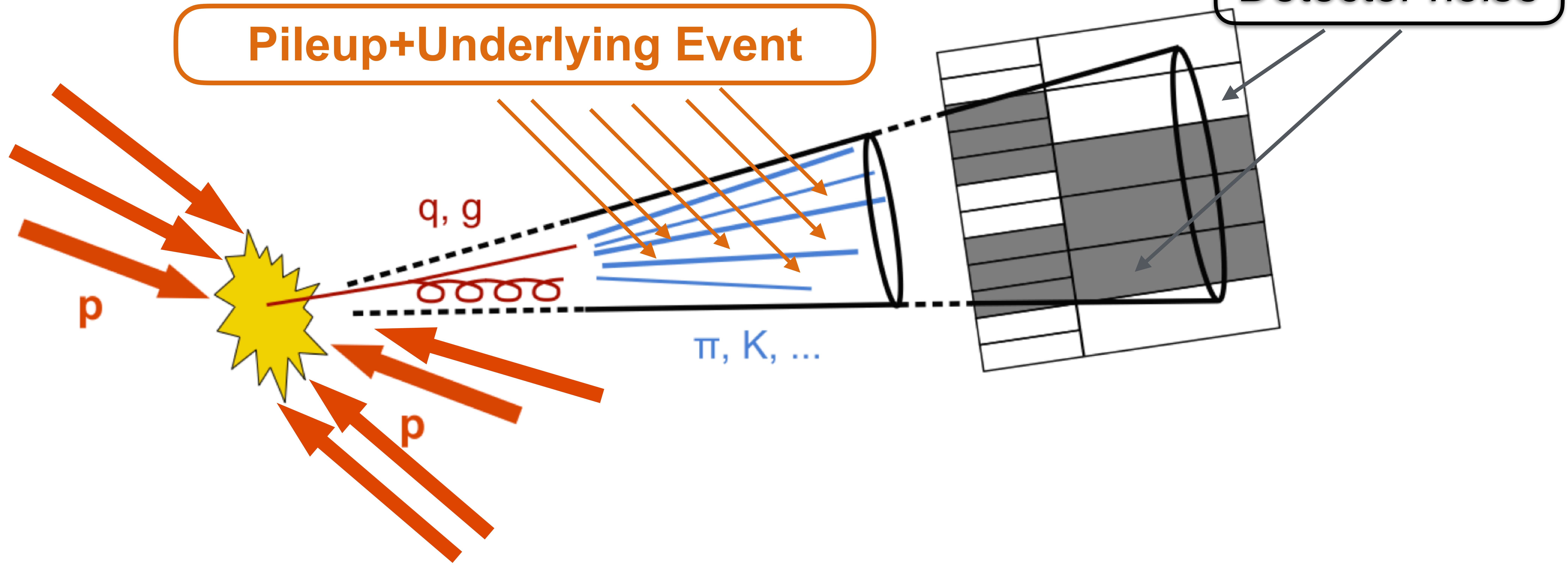
Theory (perturbation theory) / LHC pp collisions ↔ Parton Shower + Hadronisation (non-perturbative) ↔ Experiment

Pileup+Underlying Event



Collider data is complex

Theory (perturbation theory) / LHC pp collisions ↔ Parton Shower + Hadronisation (non-perturbative) ↔ Experiment

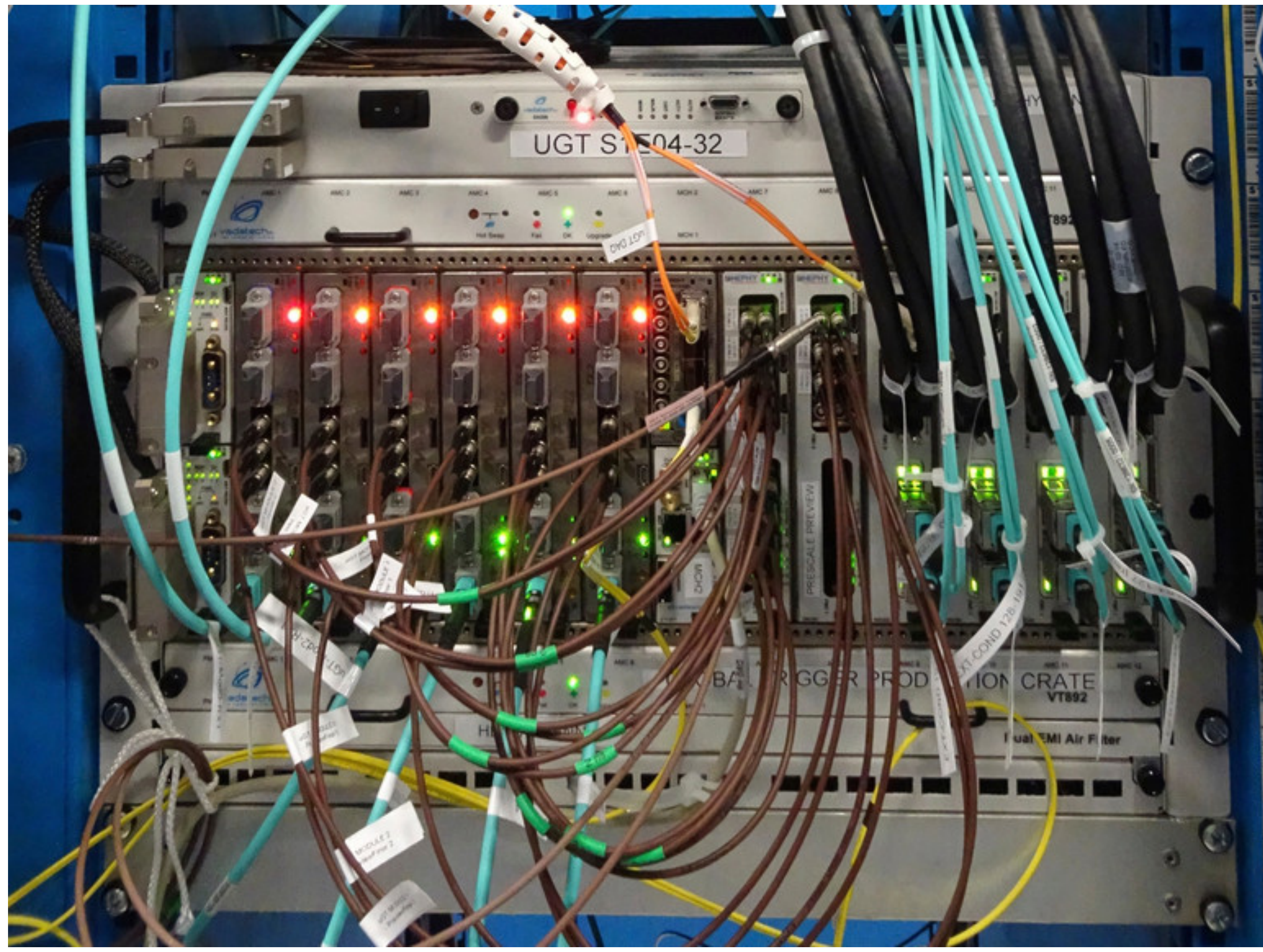


Detector noise

Pileup+Underlying Event

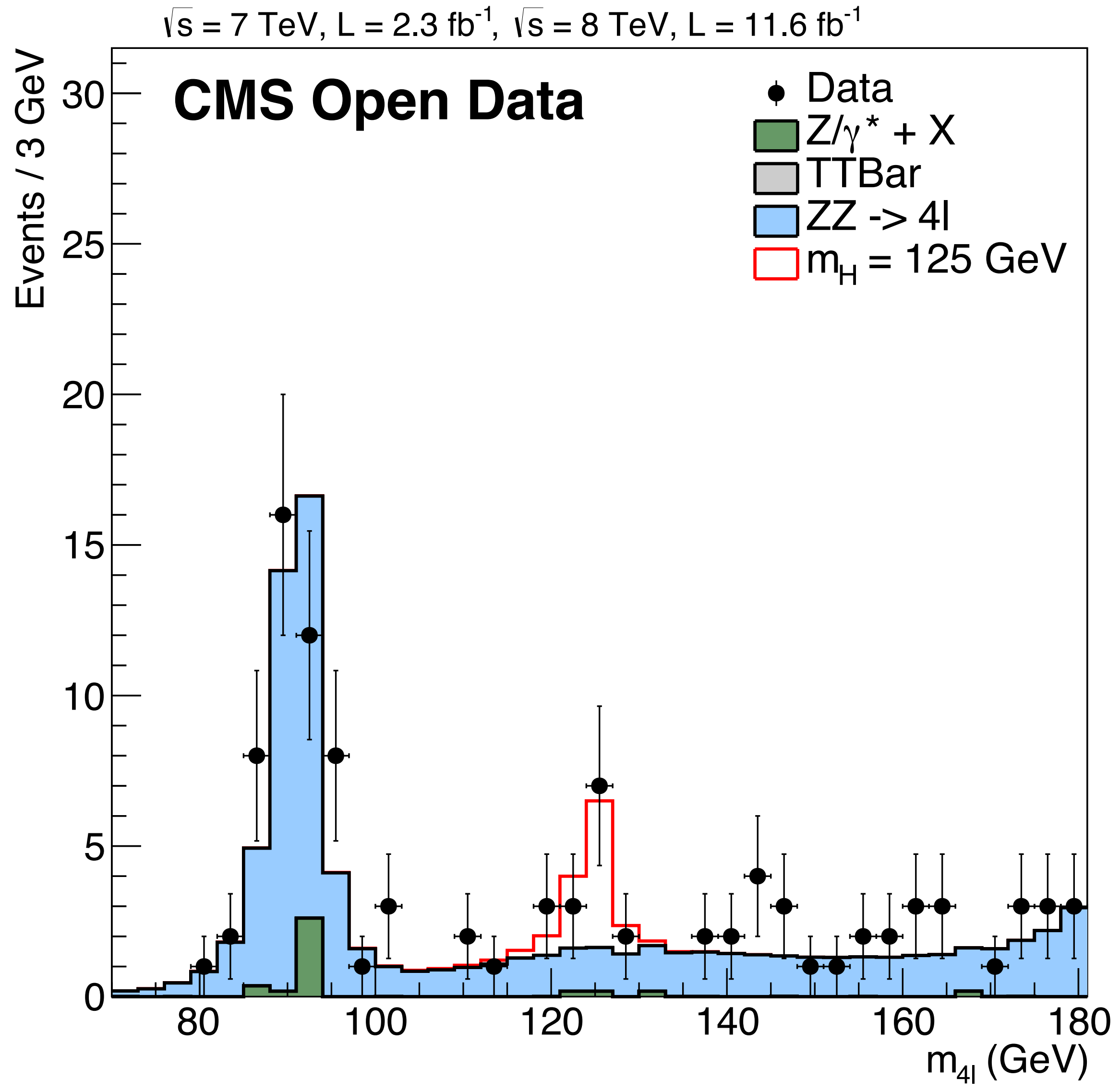
Analysing collider data is very challenging

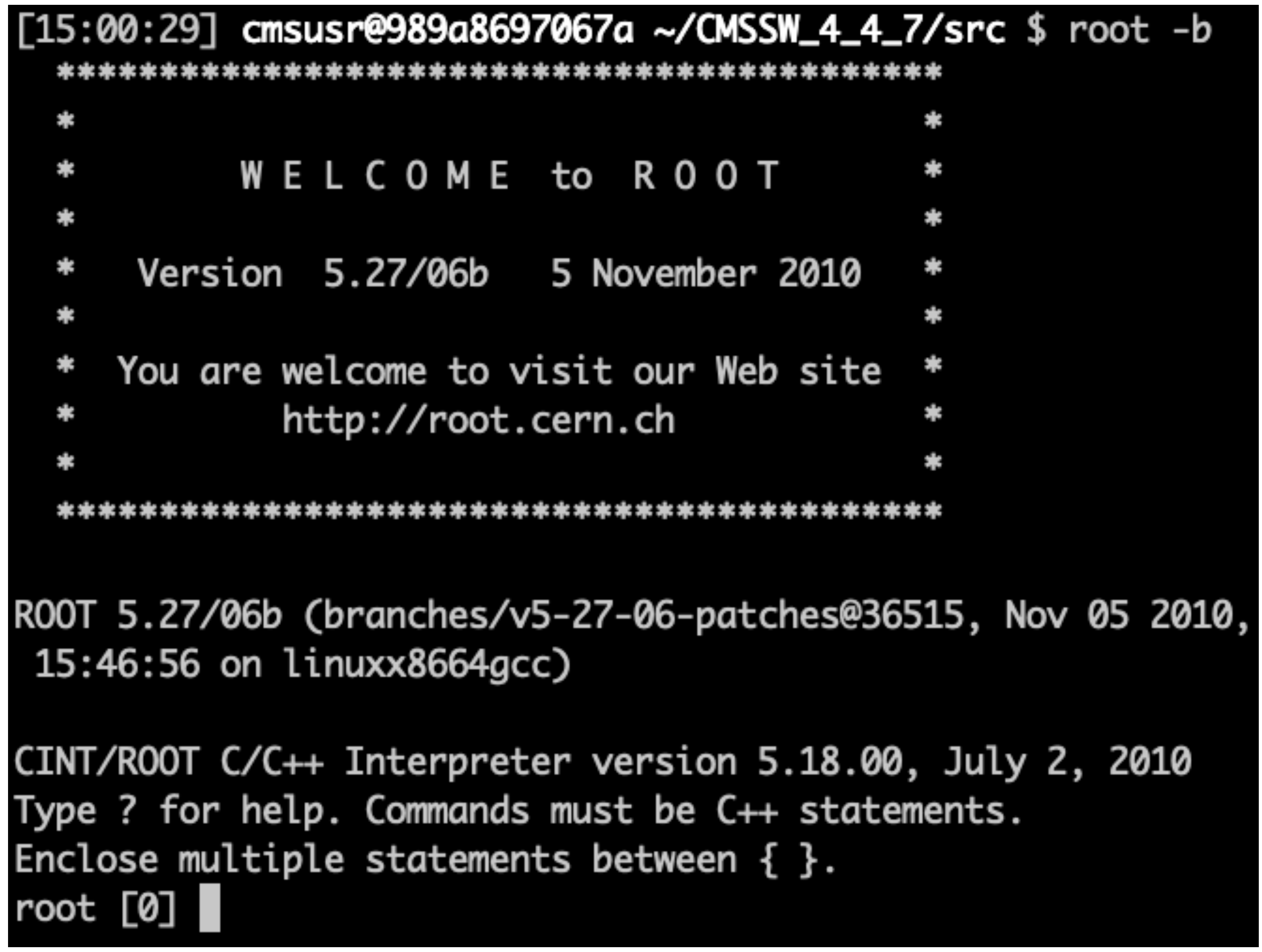
- We can **only store 0.025‰ of the collisions** (1 in 40,000 events or 1,000 events per second)
  - A multi-stage trigger system selects events of interest — this bias needs to be taken into account when performing an analysis
- A raw event has the size of about 2 megabytes
  - We have recorded tens of billions of events, and simulated even more
  - **Size can be reduced at the cost of information loss** — expertise required
  - We currently release largely “Analysis object data” (500 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

- 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
- Simulation of new processes needs CMSSW
  - Parts of the software are more than a decade old ➜ interfacing can be difficult





```
[15:00:29] cmsusr@989a8697067a ~/CMSSW_4_4_7/src $ root -b
  *******************************************
  *                                         *
  *        W E L C O M E  to  R O O T       *
  *                                         *
  *   Version   5.27/06b    5 November 2010 *
  *                                         *
  *  You are welcome to visit our Web site  *
  *          http://root.cern.ch            *
  *                                         *
  *******************************************

ROOT 5.27/06b (branches/v5-27-06-patches@36515, Nov 05 2010,
 15:46:56 on linuxx8664gcc)

CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
root [0] █
```