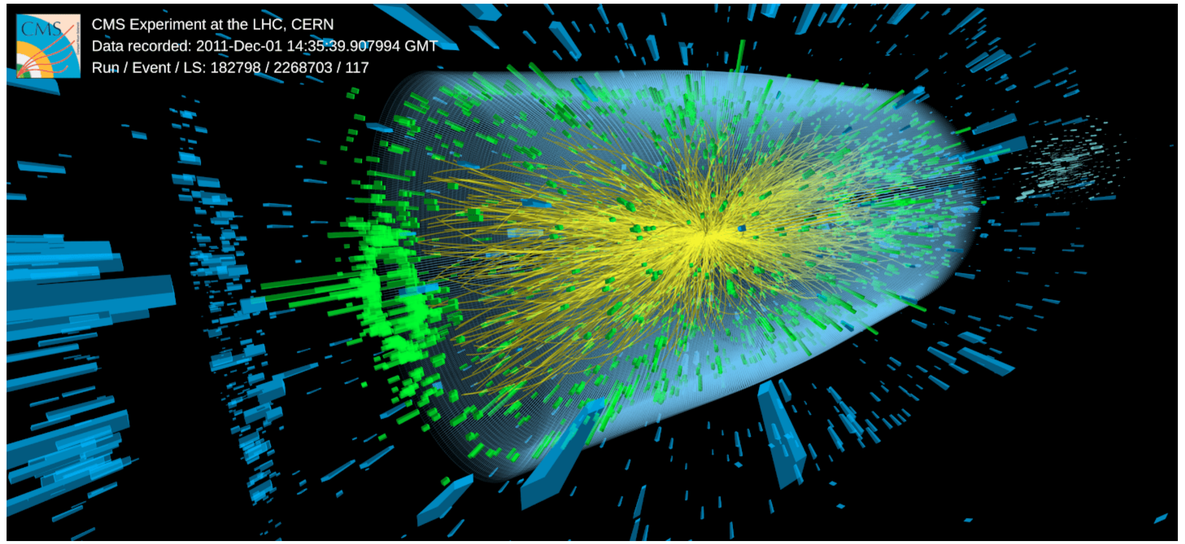
PAUL SCHERRER INSTITUT

Analysis preservation at CMS

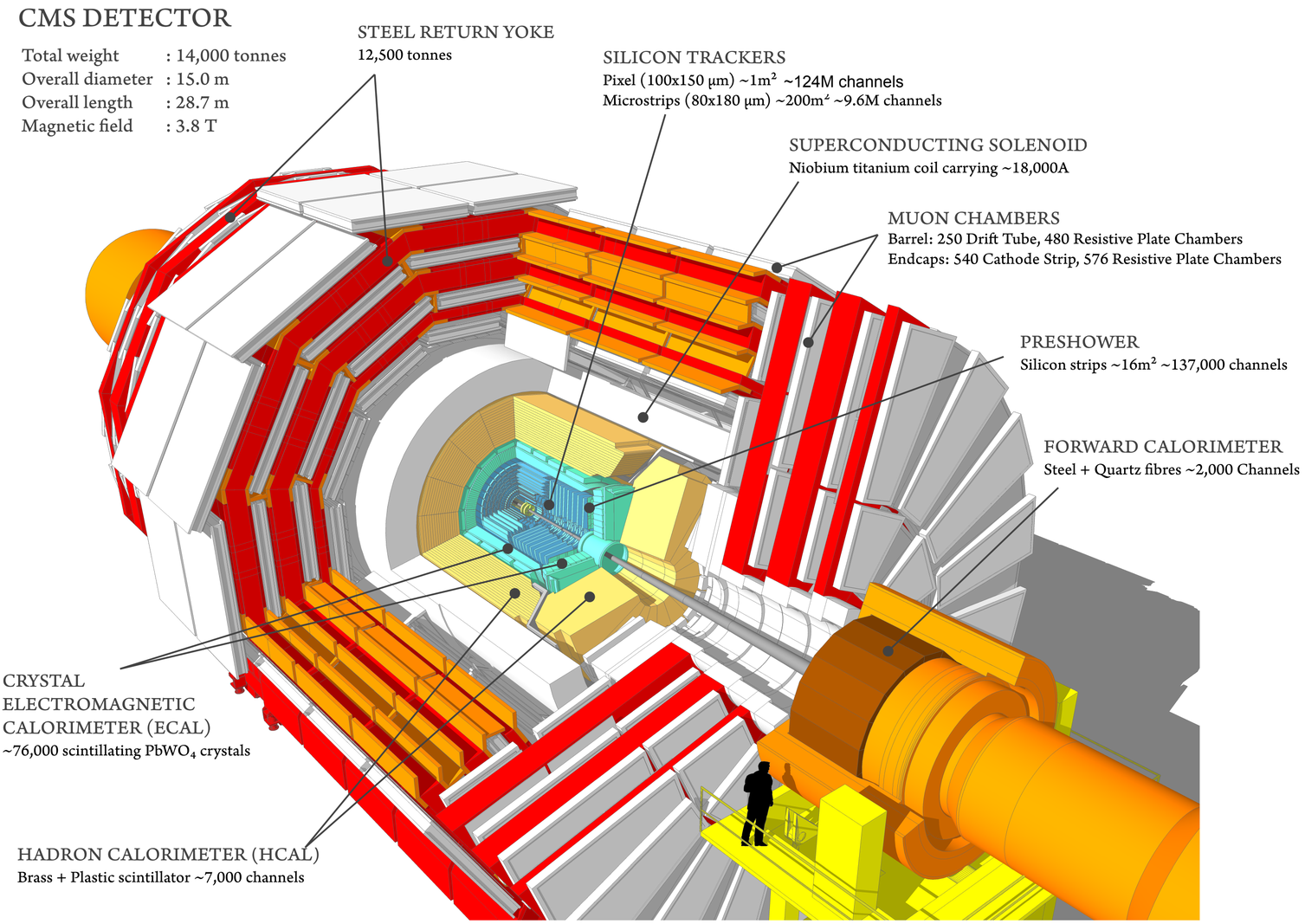
The why and how



Clemens Lange (Paul Scherrer Institute PSI)
HSF Training on Analysis Pipelines

26th February 2024

The CMS experiment



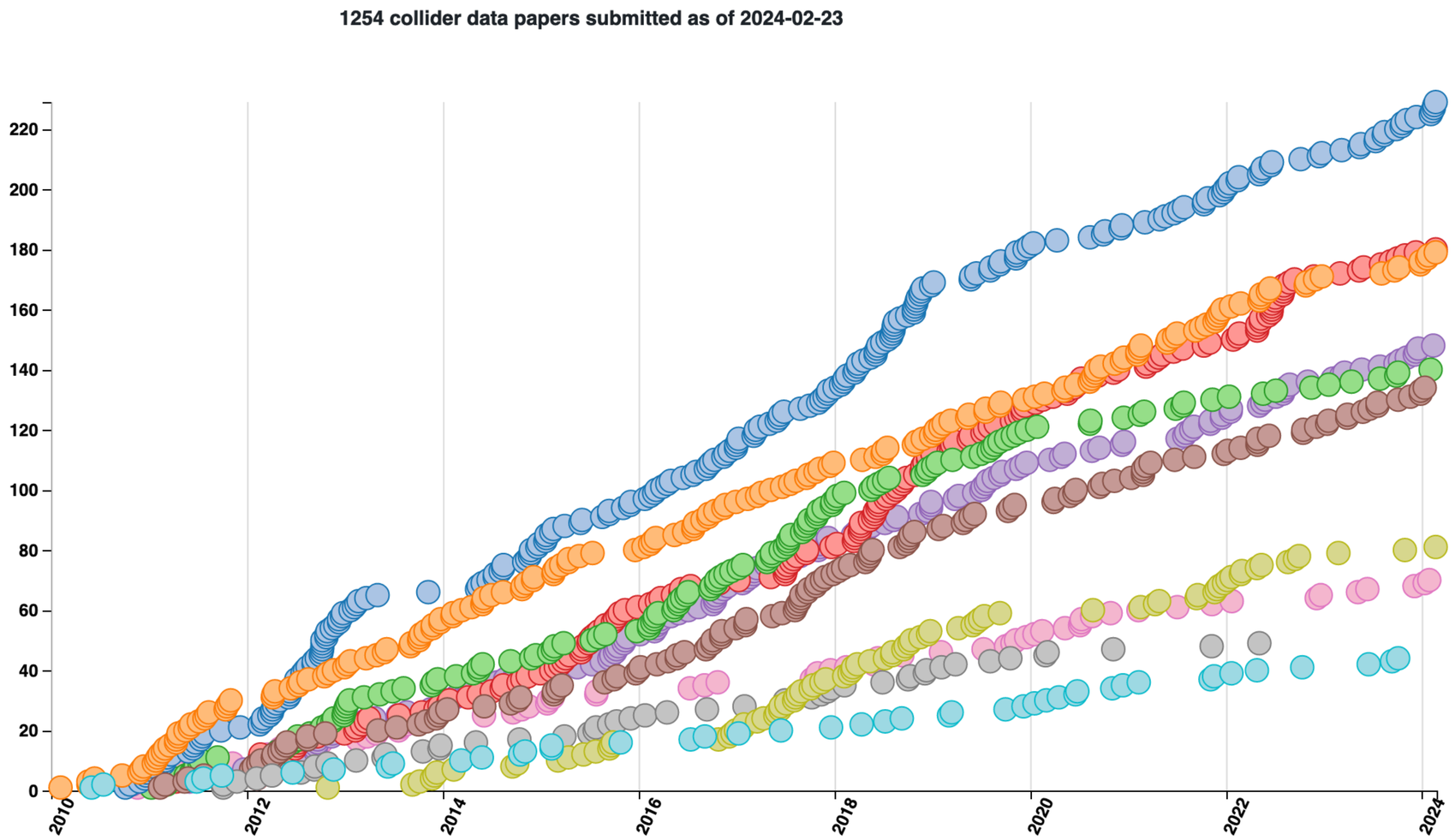
- Record up to **40,000,000 events** of the LHC collisions **per second**, 24/7 (almost) all year long
- Goal: understand the smallest building blocks of matter
- **~134 million readout channels** — extraordinary levels of technical sophistication

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

CMS publications

Show all | Total | Exotica | Standard Model | Supersymmetry | Higgs | Top | Heavy Ions

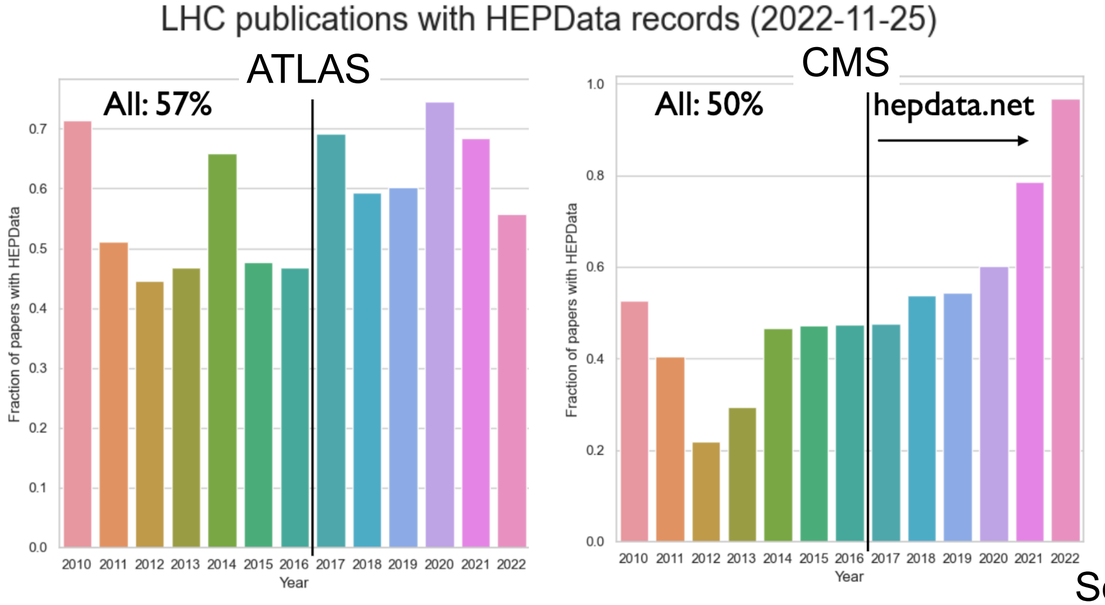
B and Quarkonia | Forward and Soft QCD | Beyond 2 Generations | Detector Performance



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

Physics results

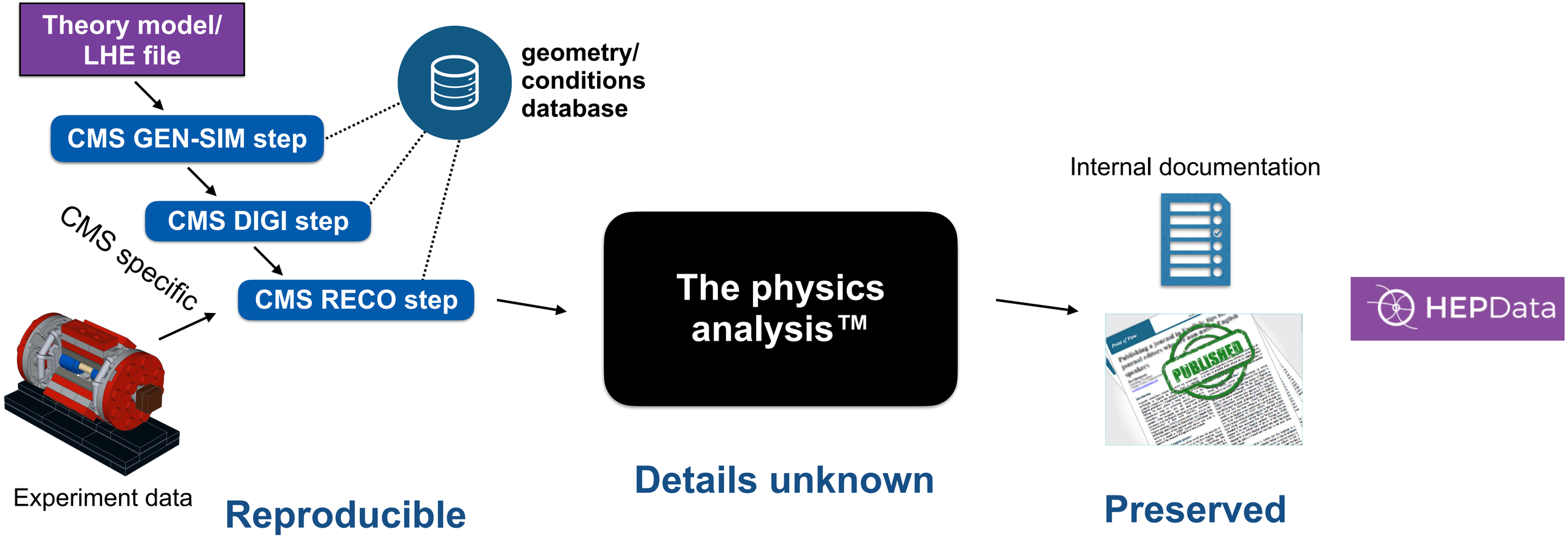
- Since 2008, >1200 peer-reviewed papers published
  - Among them the discovery of the Higgs boson (No. 183)
- All published under **open access** (since 2014 under SCOAP³)
  - Preprints available on arXiv
  - Tabulated results largely available on HEPData portal
- Since 2014, have released > 3 petabytes of open data available on the CERN Open Data Portal
  - Both collision and simulation data sets
  - Entire Run-1 + 2015 data sets



Source

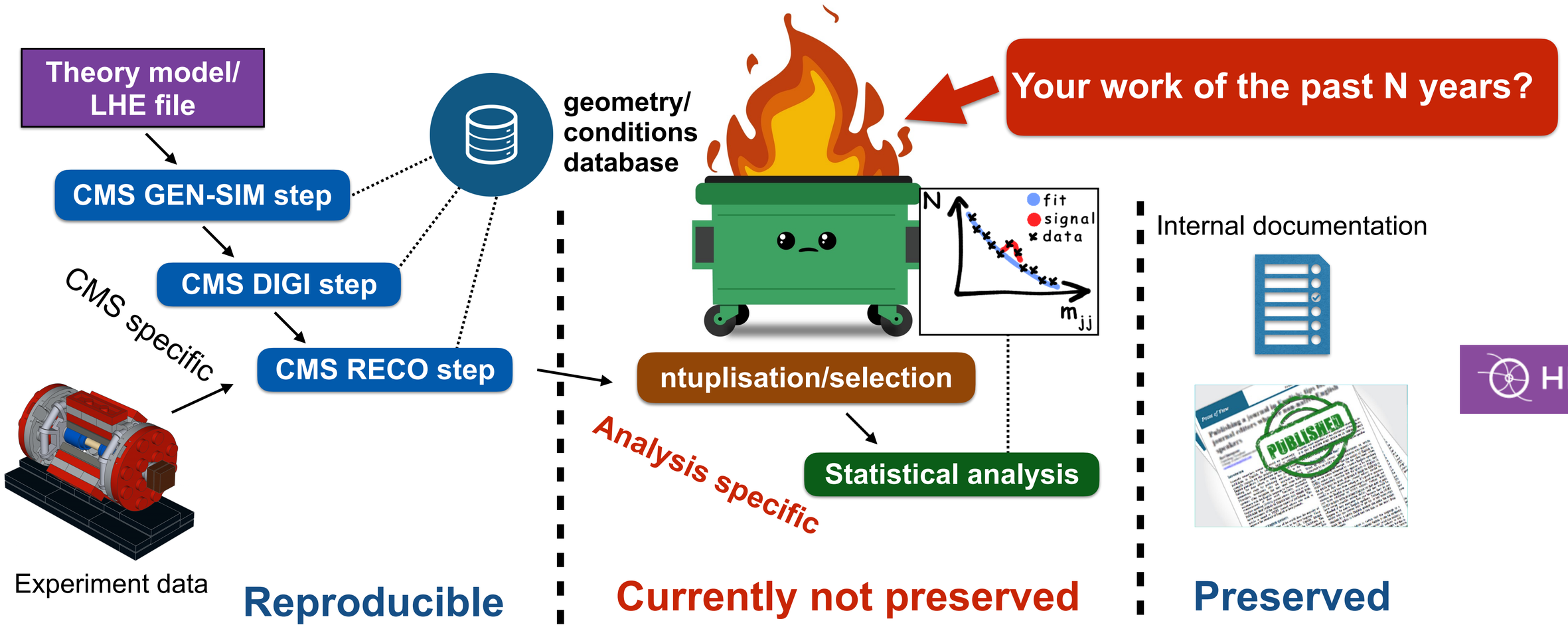
Current state of analysis preservation in CMS

- Event generation+simulation as well as reconstruction (both data and MC) centralised
  - Software and database tags preserved and archived
- Internal documentation (analysis notes) preserved
- We have not (yet) agreed in CMS to systematically preserve actual analyses and prepare them for reuse



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

Individual analysis stages in an executable way (including all dependencies)

2. Capture commands

How to run the captured software?

3. Capture workflow

How to connect the individual analysis steps?

PAUL SCHERRER INSTITUT

Demo time

CMS Open Data simplified analysis example

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

Steps towards reusable analyses

1.Capture software

Individual analysis stages in an executable way (including all dependencies)

2. Capture commands

How to run the captured software?

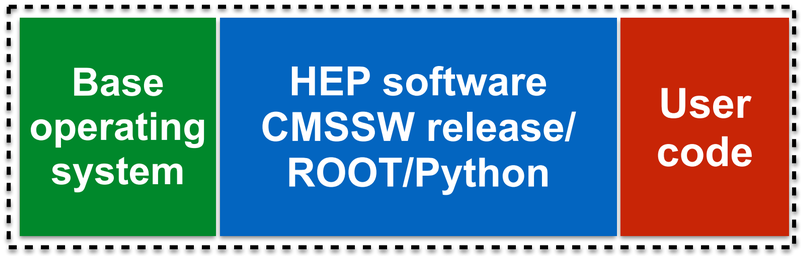
3. Capture workflow

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

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

Individual analysis stages in an executable way (including all dependencies)

2. Capture commands

How to run the captured software?

3. Capture workflow

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

Individual analysis stages in an executable way (including all dependencies)

2. Capture commands

How to run the captured software?

3. Capture workflow

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, selected examples:
  - SnakeMake (available in REANA)
  - LAW (Luigi Analysis Workflow, used by several analyses in CMS)

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)

Experiment data

Theory model/ LHE file

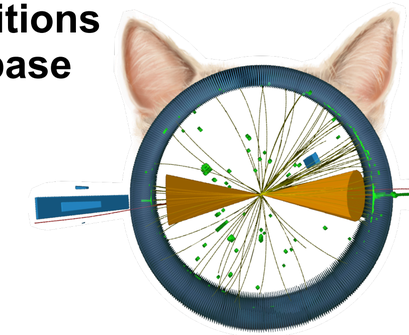
CMS GEN-SIM step

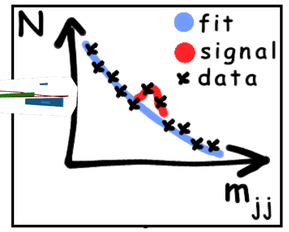
CMS DIGI step

CMS RECO step

CMS specific

geometry/ conditions database





ntuplisation/selection

Statistical analysis

Analysis specific

Reproducible

Shared tools with expert-approved recipes

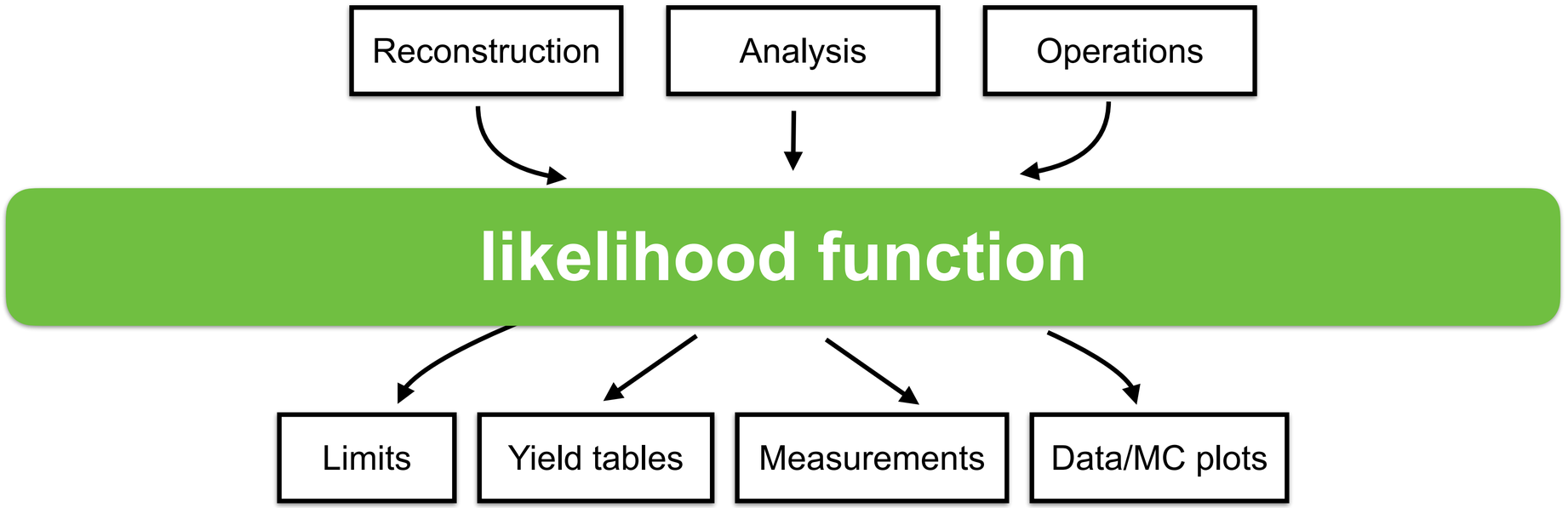
Internal documentation

HEPData

Preserved

Publishing full likelihoods

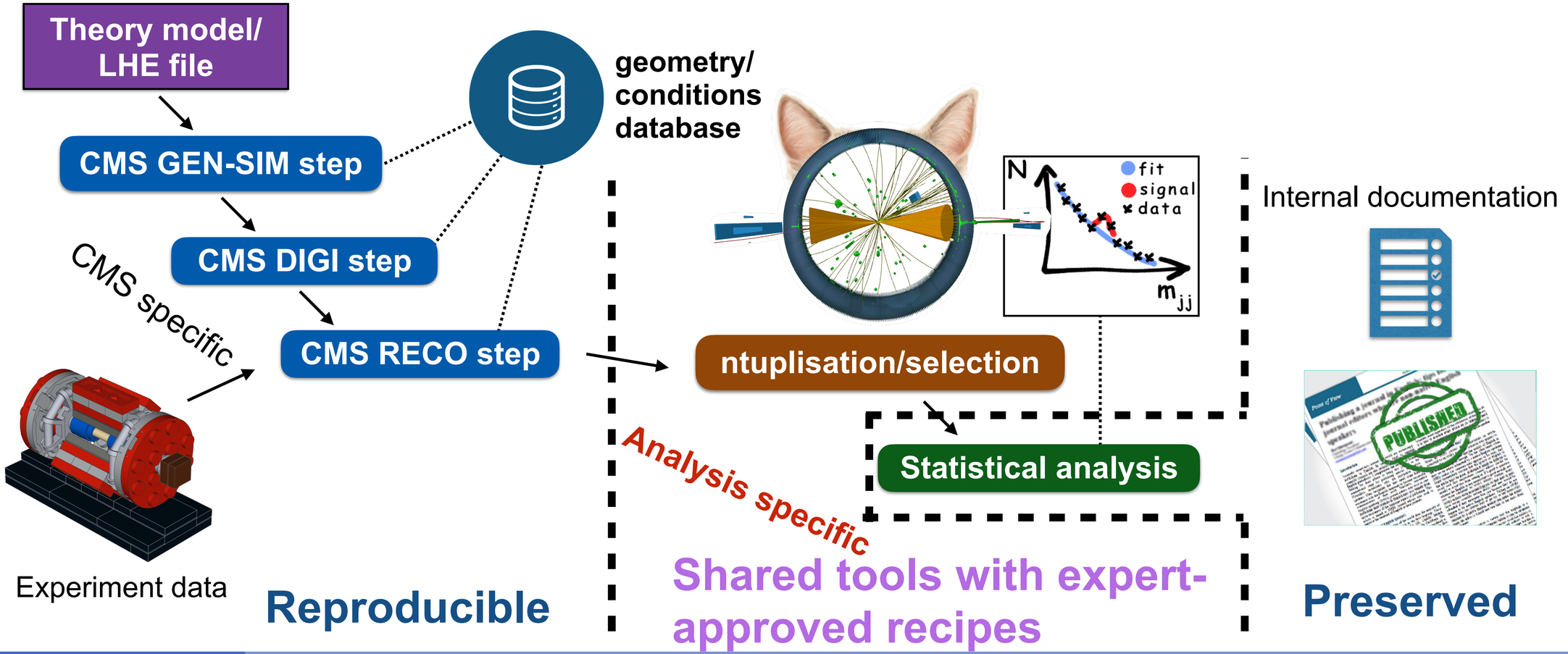
- The **likelihood function** is a particularly special data product
  - Small, information-dense, overall summary of the analysis
  - Almost every analysis decision is reflected in the likelihood



- Expect to see first full likelihoods from CMS in the next few months
  - These will be accompanied by the release of the statistical inference tool “Combine”

Towards analysis preservation and reusability

- The new CMS CAT group will work to close the gap in analysis preservation and reusability
- Will need analysts to be part of this change



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

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!

CERN Open 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

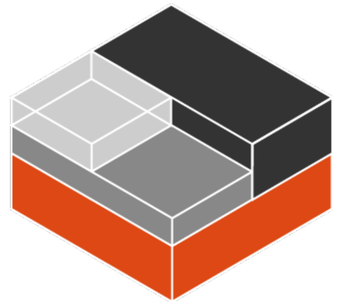
- Level 1: Open access publication and additional numerical data
- Level 2: Simplified data for Outreach and Education
- **Level 3**: Reconstructed data and the software to analyse them
- Level 4: Raw data, and the software to reconstruct and analyse them

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



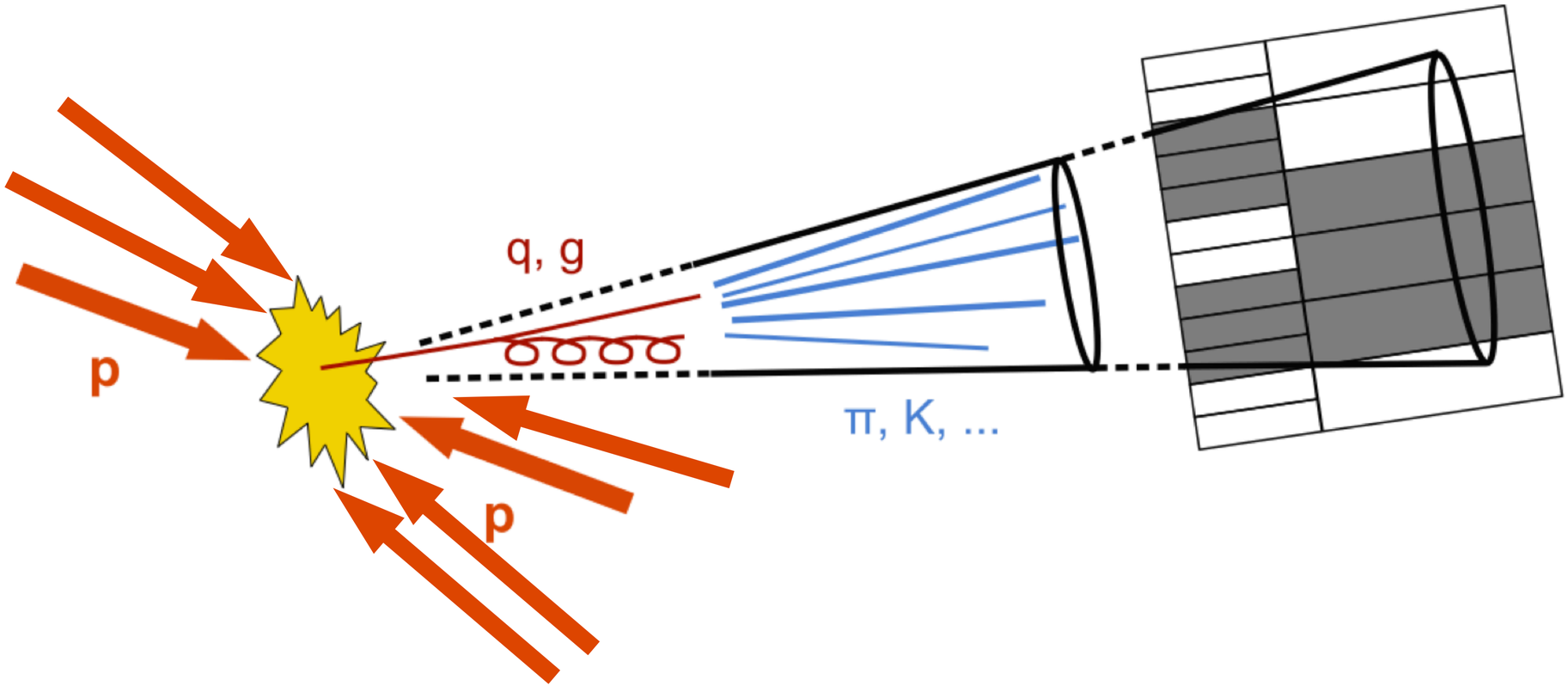
Addressing the challenges

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
  - 2023 CMS Open Data Workshop at Fermilab LPC

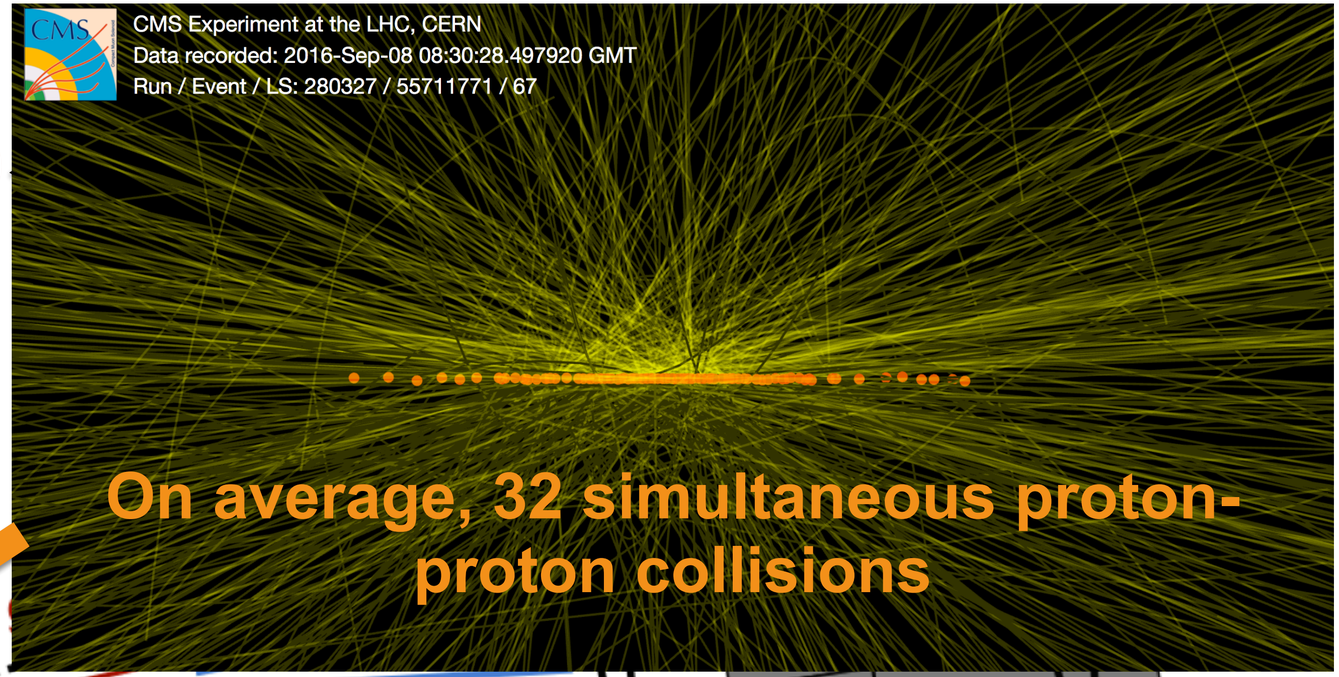
Collider data is complex

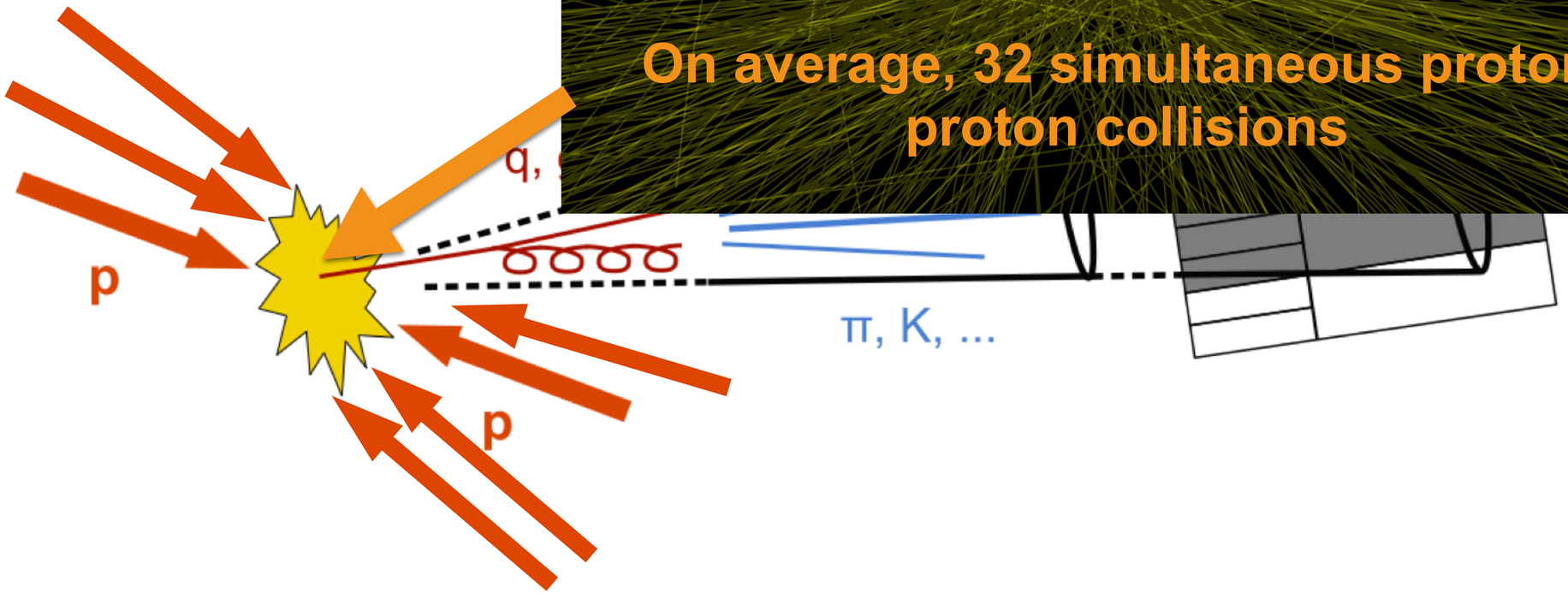
Theory (perturbation theory) / LHC pp collisions ↔ **Parton Shower + Hadronisation (non-perturbative)** ↔ **Experiment**



Collider data is complex

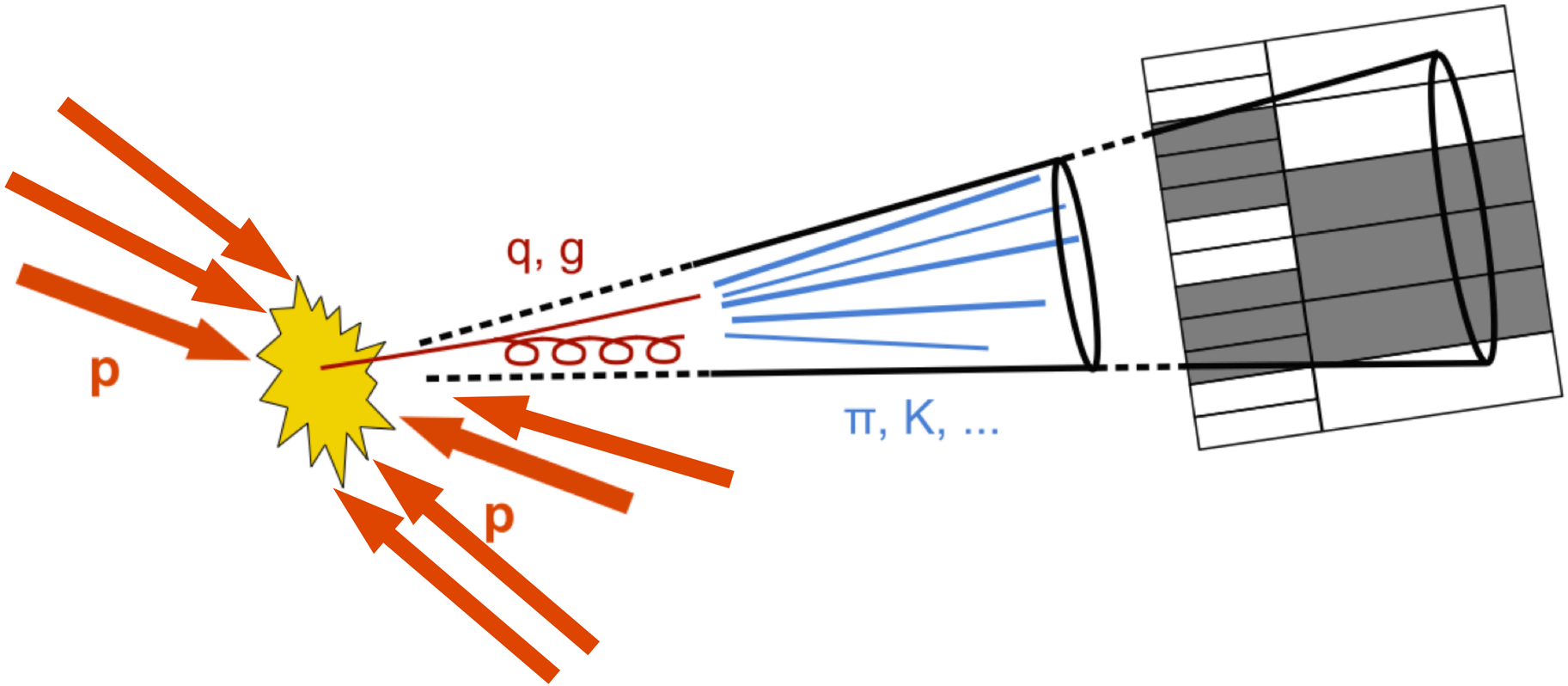
Theory (perturbation theory) / LHC pp collisions





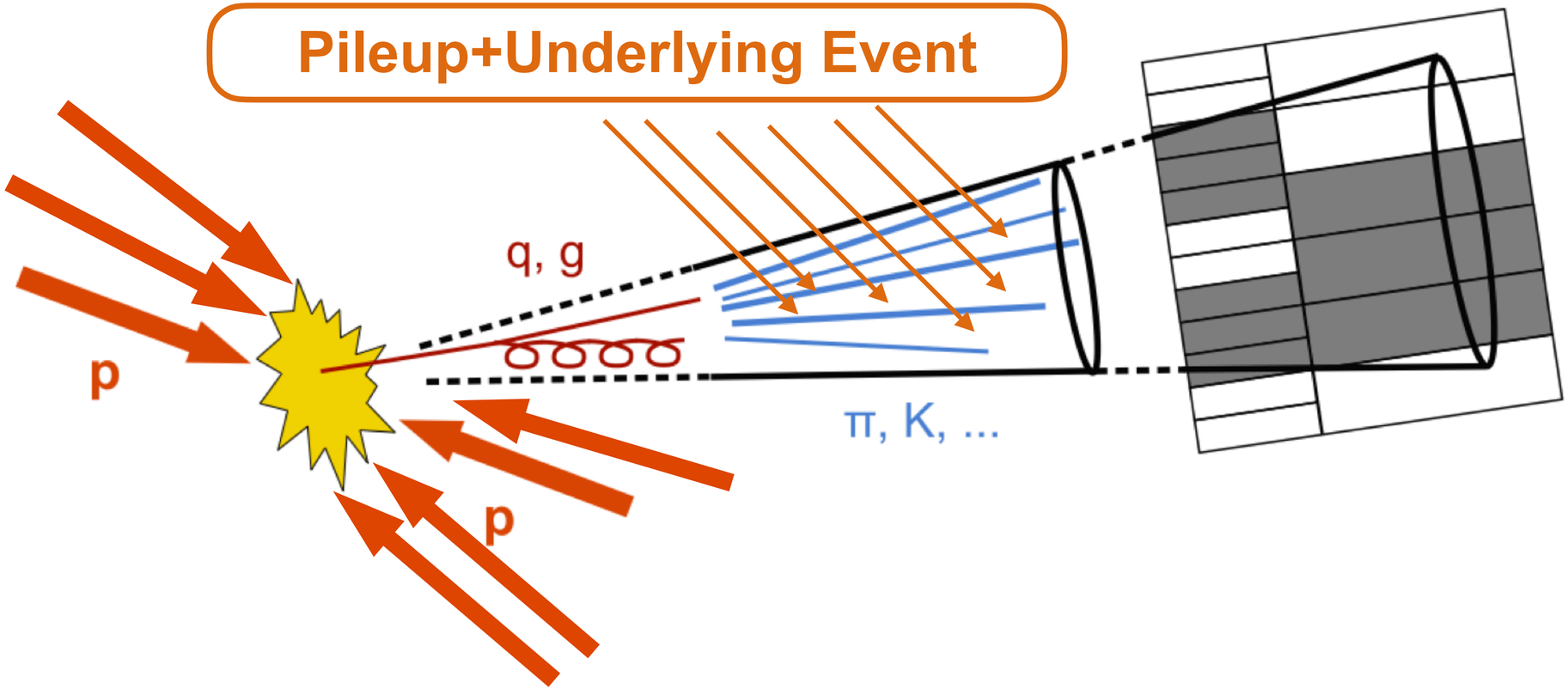
Collider data is complex

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



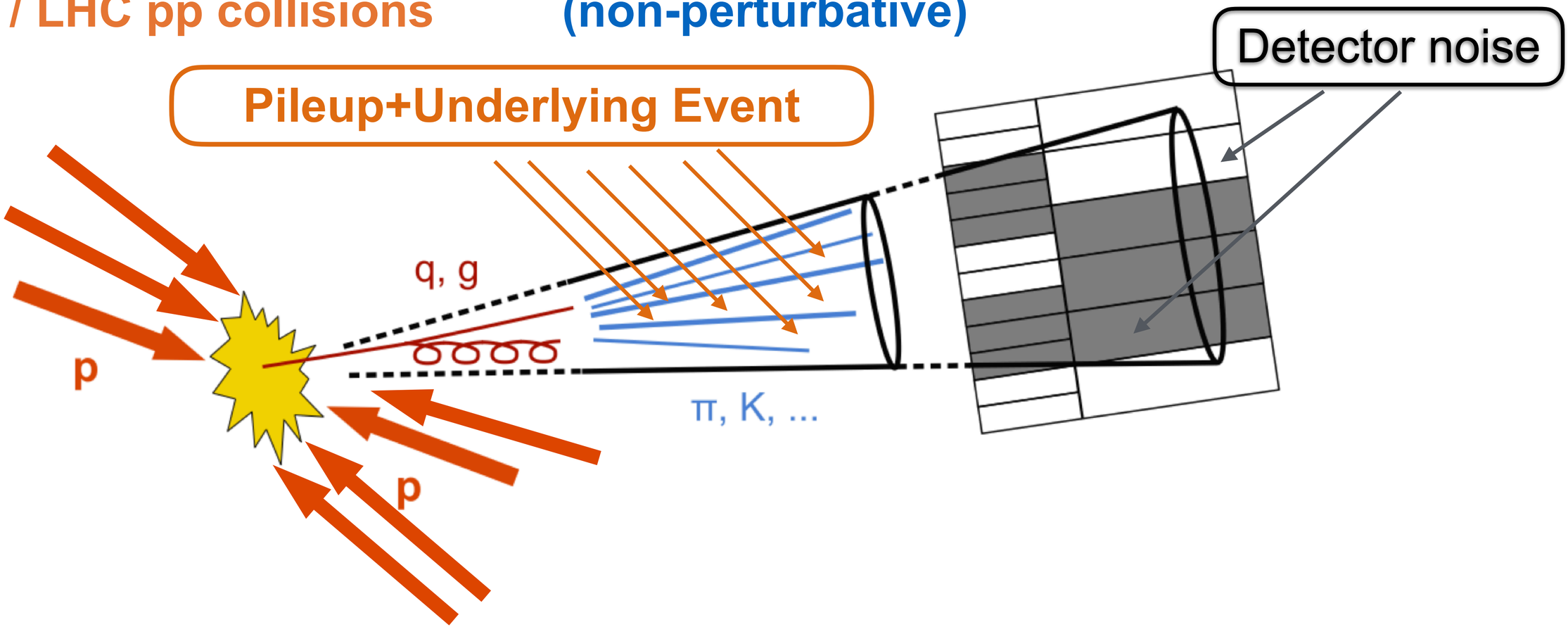
Collider data is complex

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



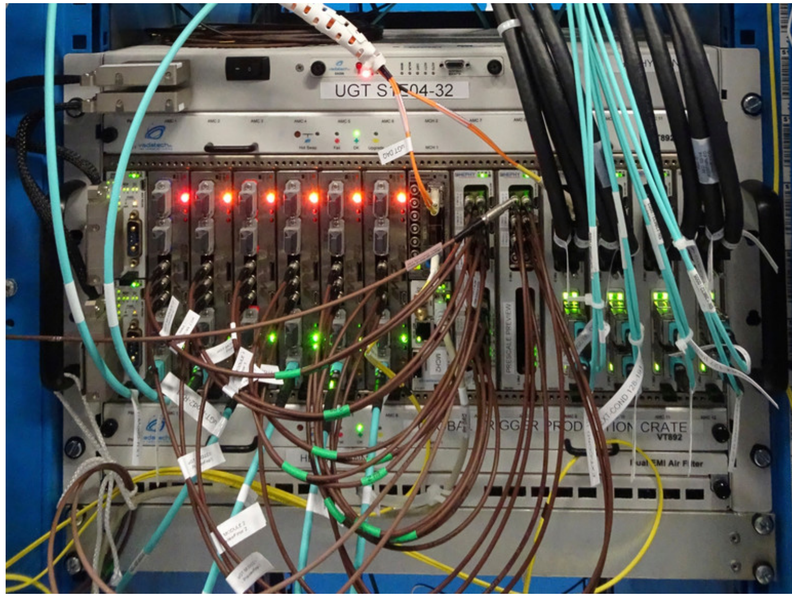
Collider data is complex

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



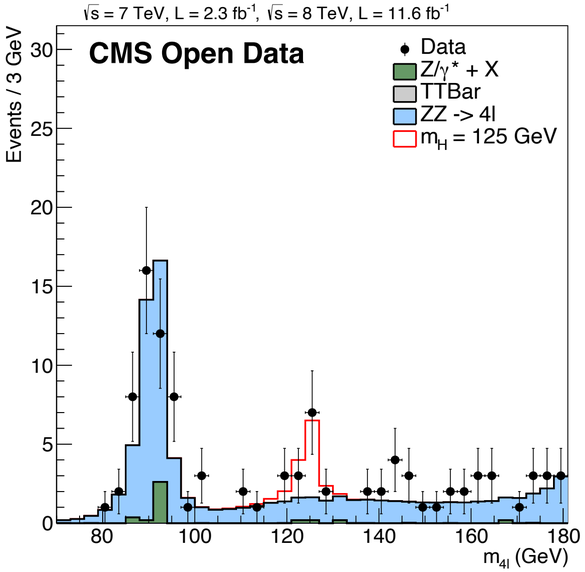
Analysing collider data is very challenging

- 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]
```