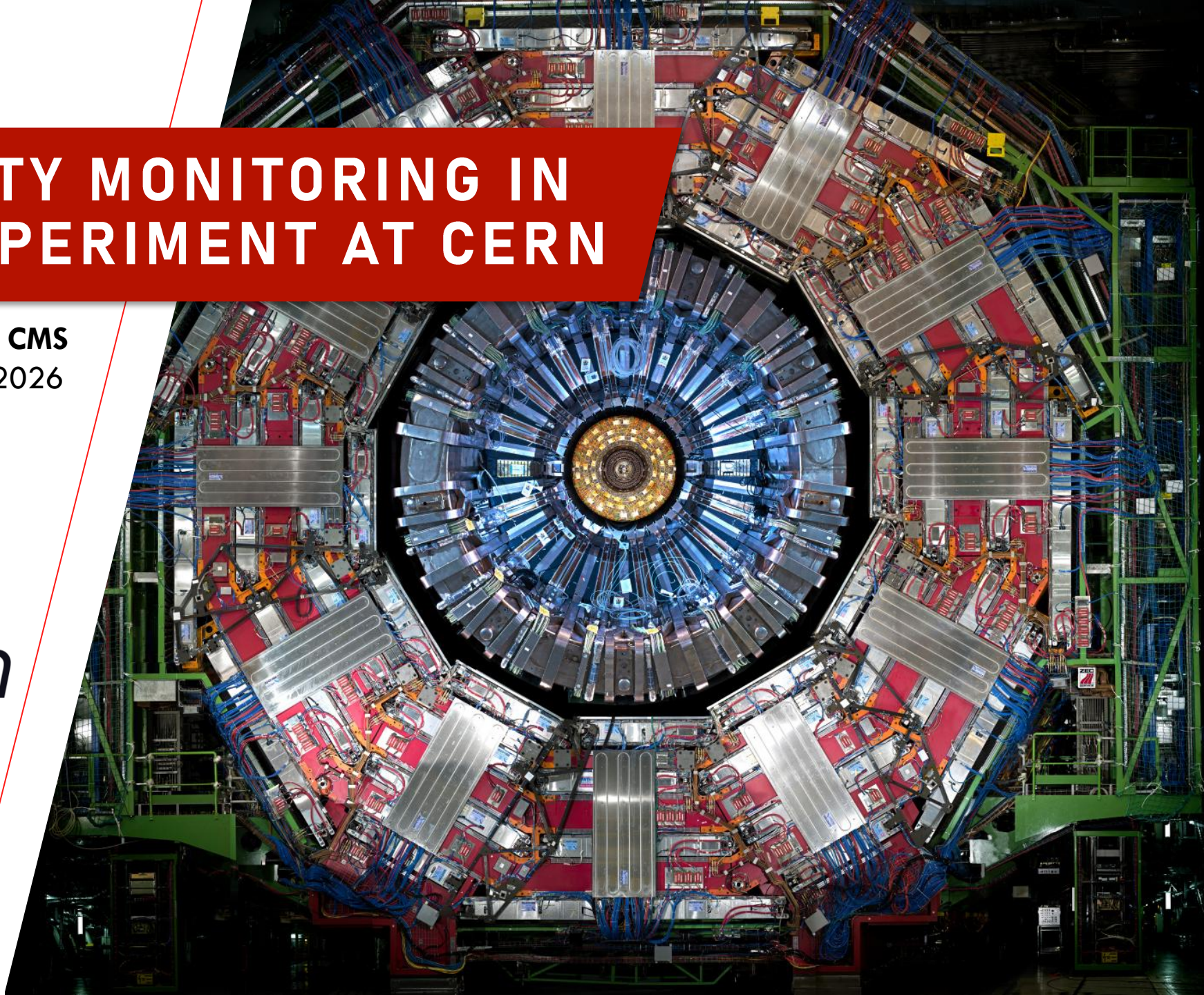
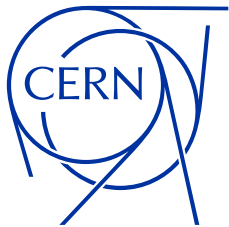


DATA QUALITY MONITORING IN THE CMS EXPERIMENT AT CERN

Roberto Seidita – on behalf of CMS
EUROFusion seminar – 07/10/2026

ETH zürich



The setting

A pairing of remarkable machines

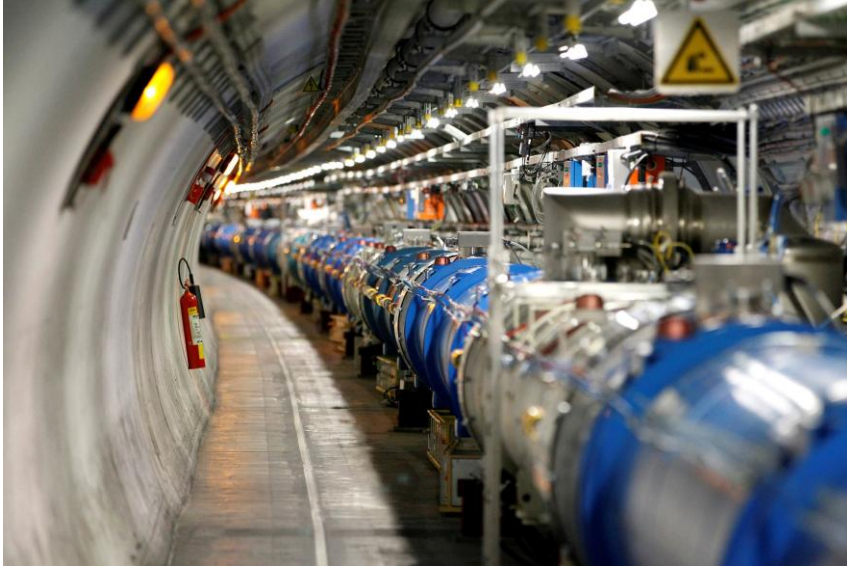
OUR SCIENTIFIC GOAL

An accelerator is a microscope for fundamental matter and interactions

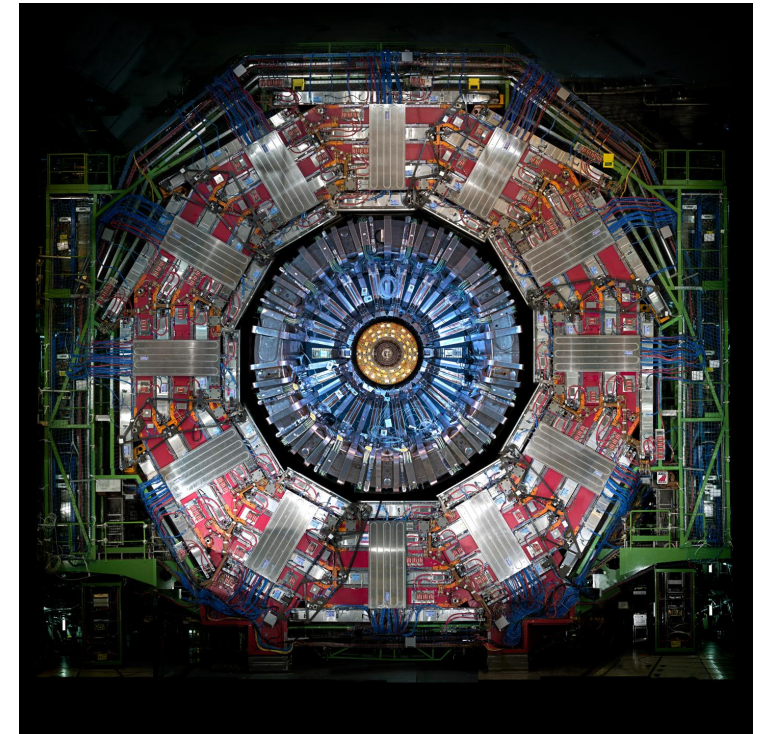
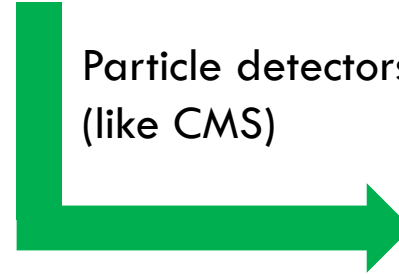
⇒ Collide particles at the highest achievable energies to resolve the smallest possible structures



The large hadron collider



Particle detectors
(like CMS)





Mont Blanc

Geneva

CERN

LHCb

ATLAS

CERN Prévessin

SPS 7 km

ALICE

CMS

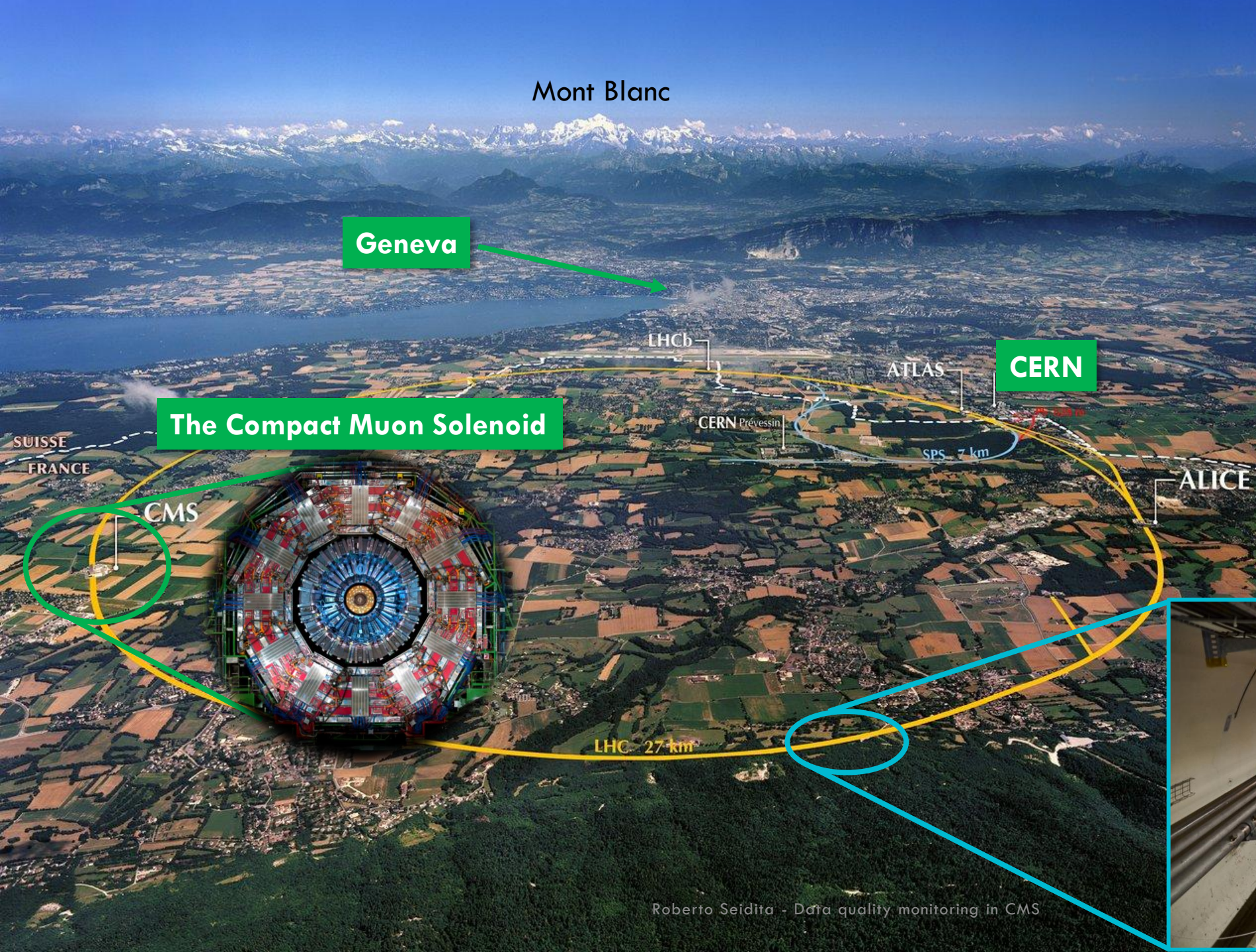
LHC 27 km

A 27 km tunnel, up to 100 m underground right next to lake Geneva

Two beams of protons are accelerated by radiofrequency cavities to up to 6.8 TeV of energy

Superconducting 8 T magnets cooled to 1.9 K keep the beams in orbit





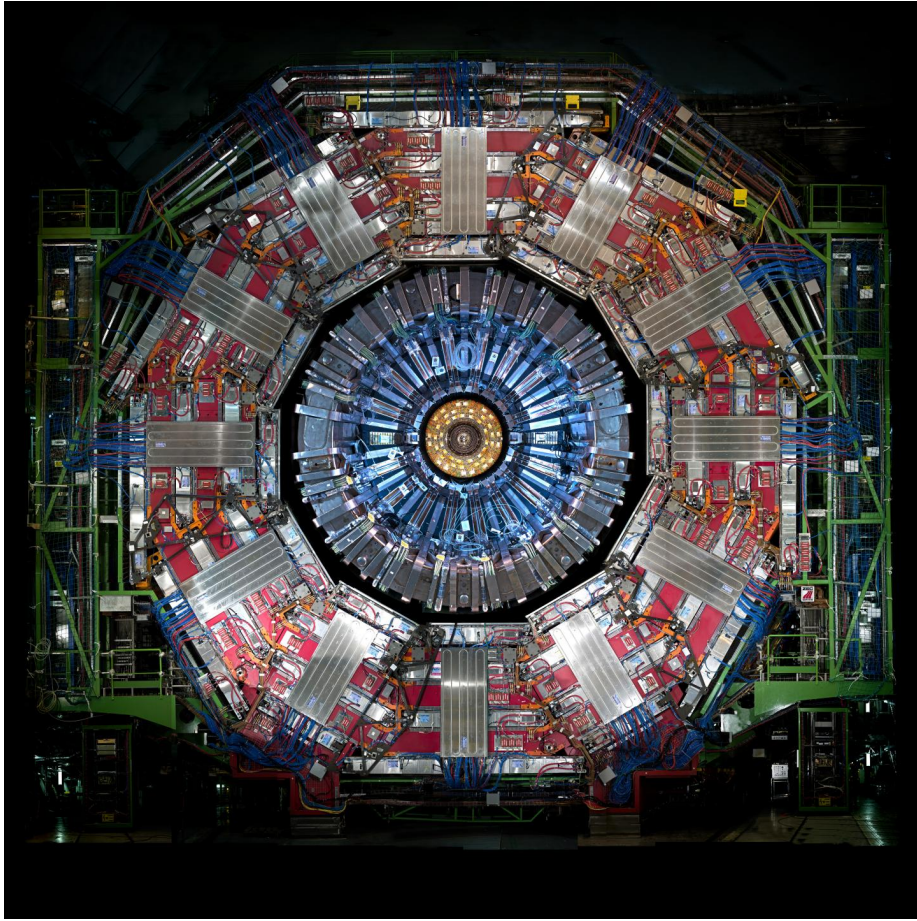
A 27 km tunnel, up to 100 m underground right next to lake Geneva

Two beams of protons are accelerated by radiofrequency cavities to up to 6.8 TeV of energy

Superconducting 8 T magnets cooled to 1.9 K keep the beams in orbit



CMS IN NUMBERS



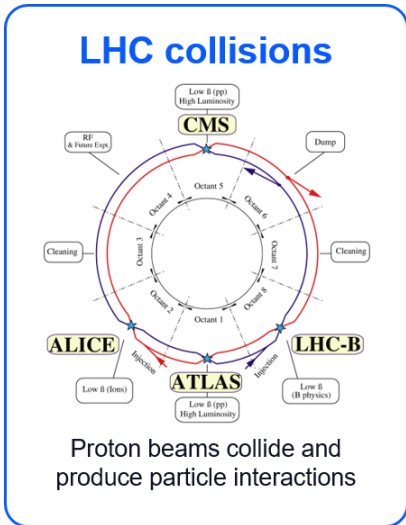
15 m wide, 21 m long

Around 6000 collaborators

Spread over 259 institutes

In 60 countries around the globe

CMS DATA FLOW IN A NUTSHELL

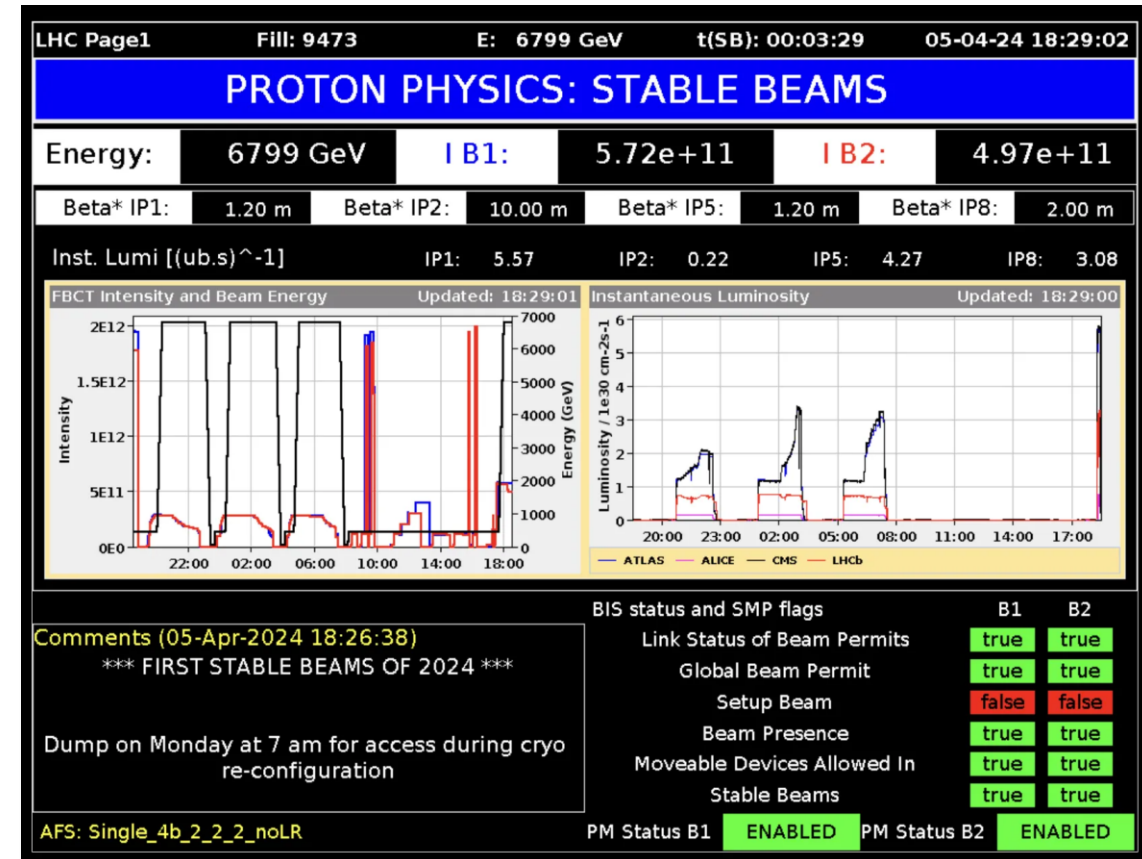


Two beams of protons are accelerated to up to 7.8 TeV each

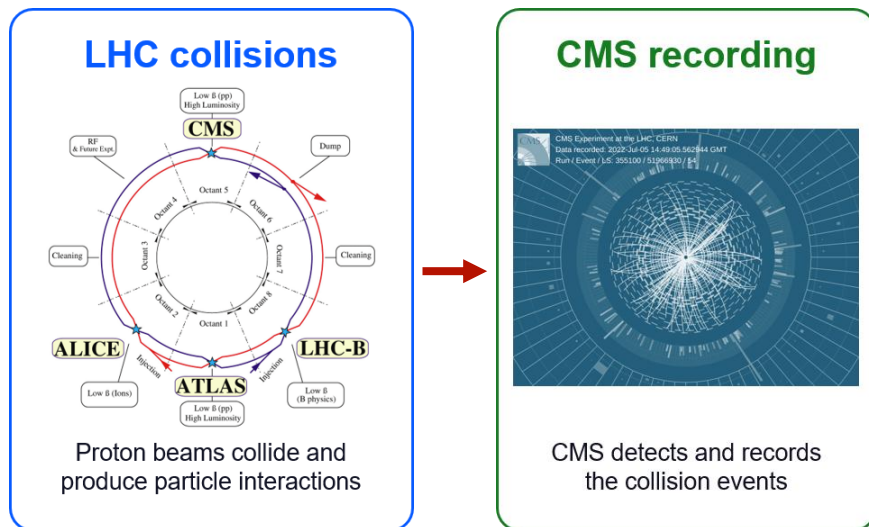
They collide in 4 interaction points, one being CMS, every 25 ns or at a rate of 40 MHz

The largest dataset at the collider energy frontier with over 500 /fb of data

→ Higgs boson discovery with ~ 10 /fb



CMS DATA FLOW IN A NUTSHELL

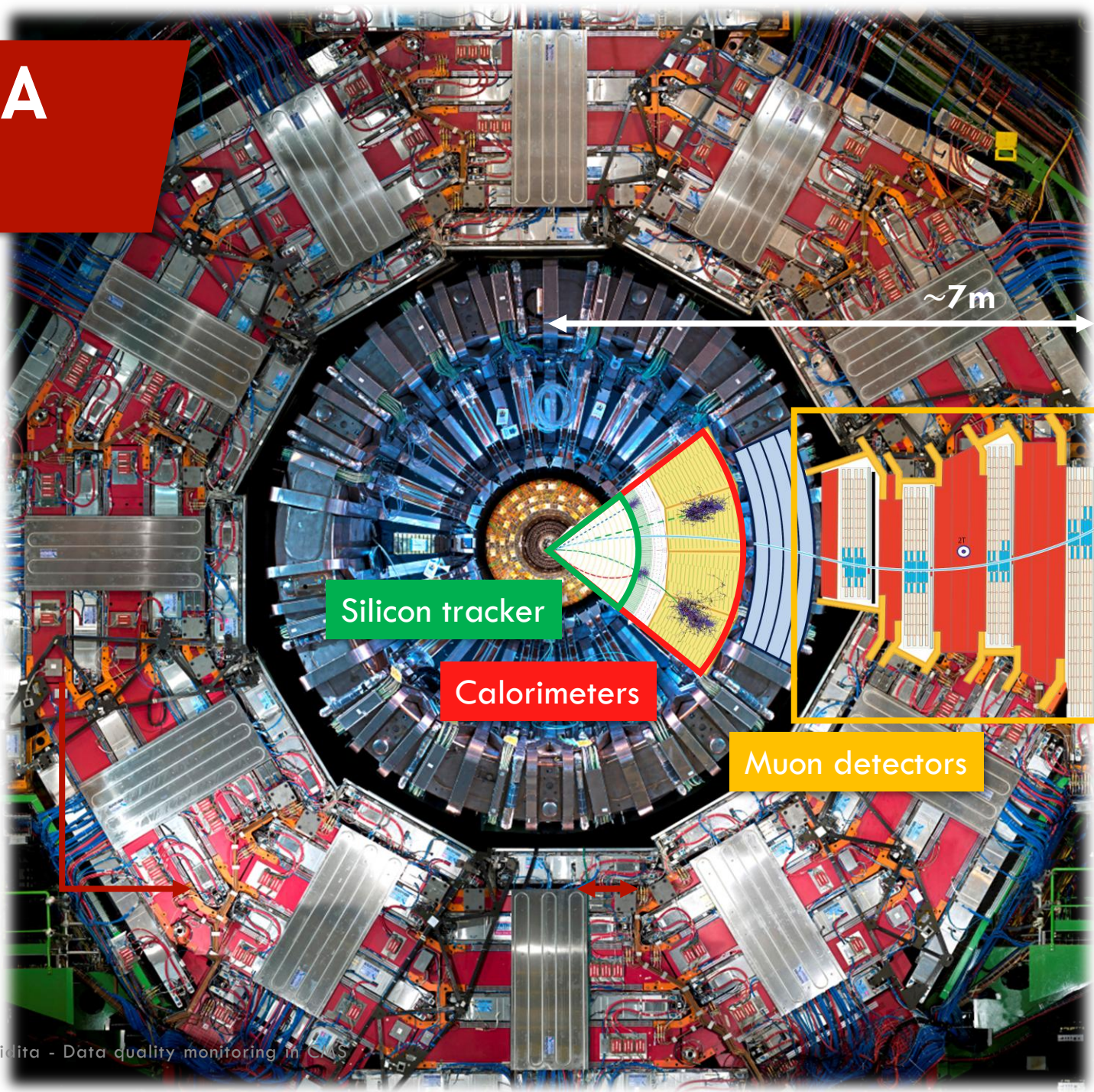


Collisions are recorded by the CMS detector

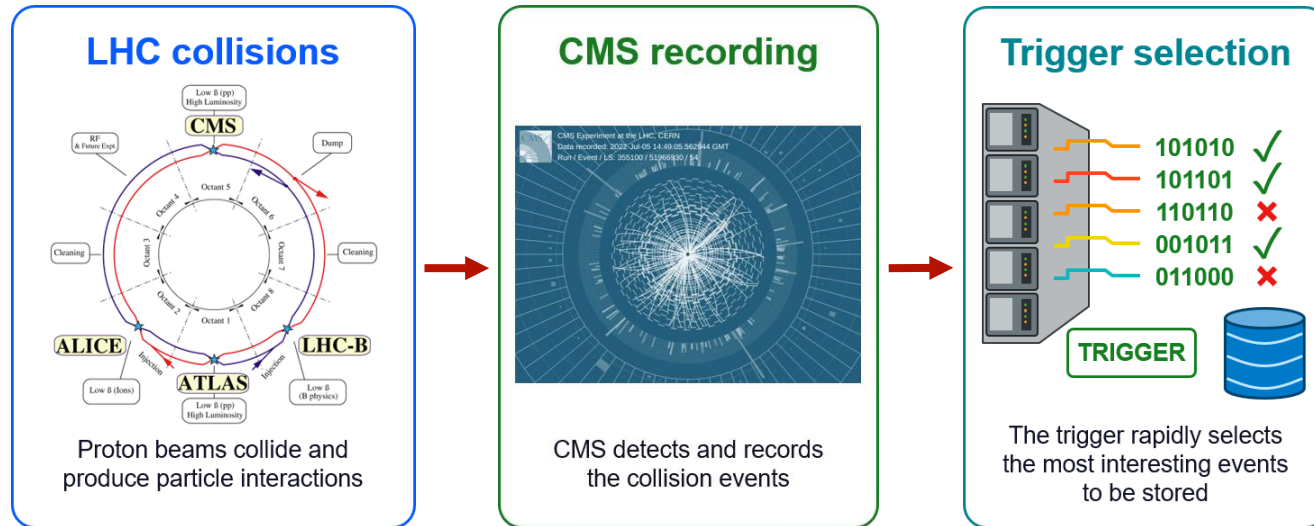
~135M readout channels across detectors

~ 1 – 3 MB per event → 40 TB/s of output

Real- and near-real-time monitoring crucial



CMS DATA FLOW IN A NUTSHELL



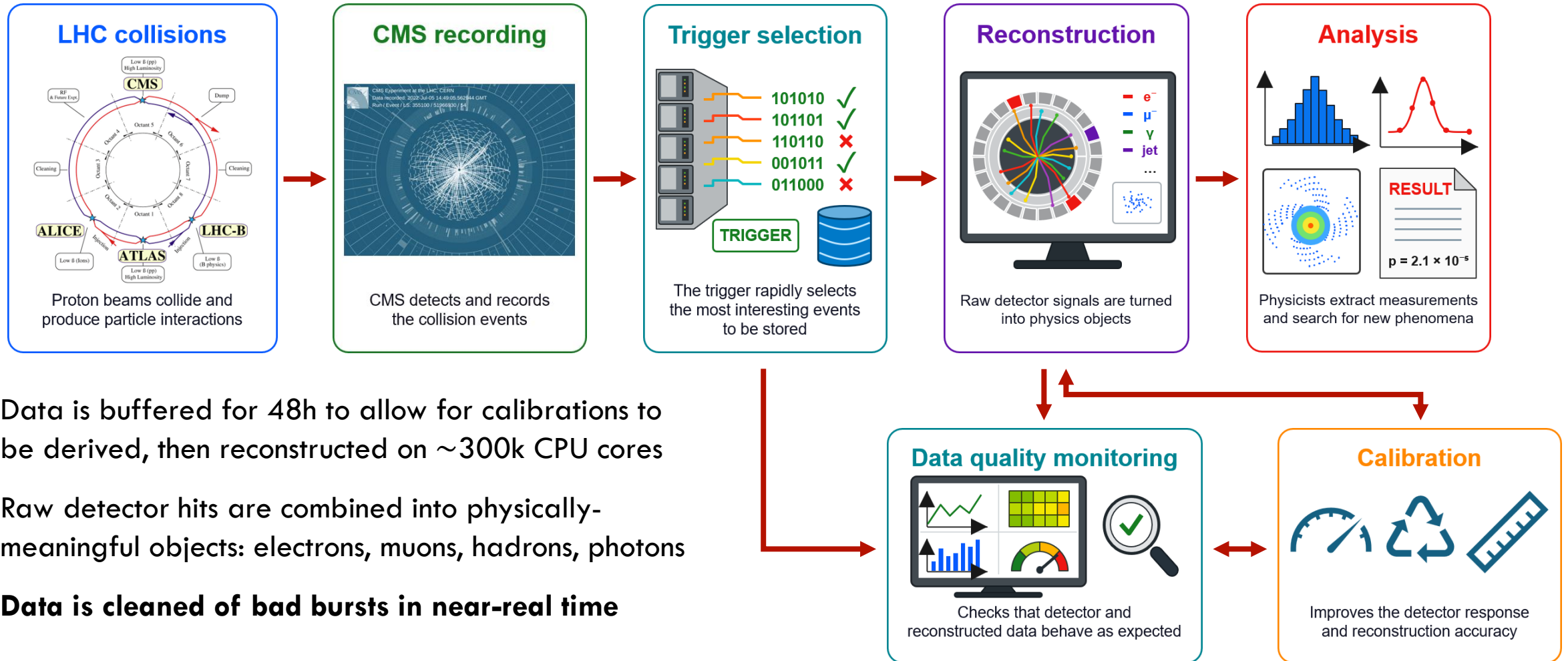
A two-tiered trigger system reduced the rate for storage

Level 1 trigger (FPGAs) reduces rate from ~ 40 MHz to ~ 120 kHz

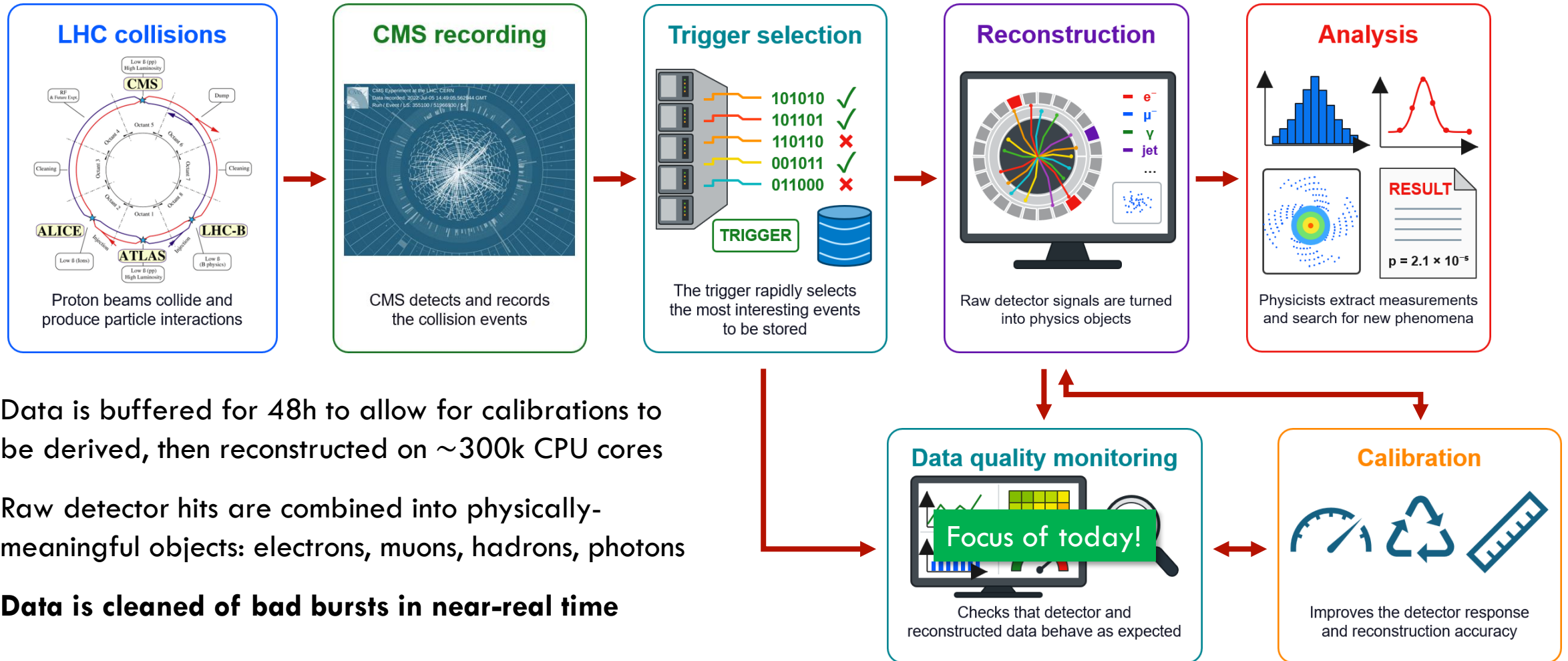
High level trigger (CPUs + GPUs) further cuts from ~ 120 kHz to ~ 3 kHz

→ ~ 10 GB/s written to disk/tape for permanent storage

CMS DATA FLOW IN A NUTSHELL



CMS DATA FLOW IN A NUTSHELL



Data is buffered for 48h to allow for calibrations to be derived, then reconstructed on $\sim 300k$ CPU cores

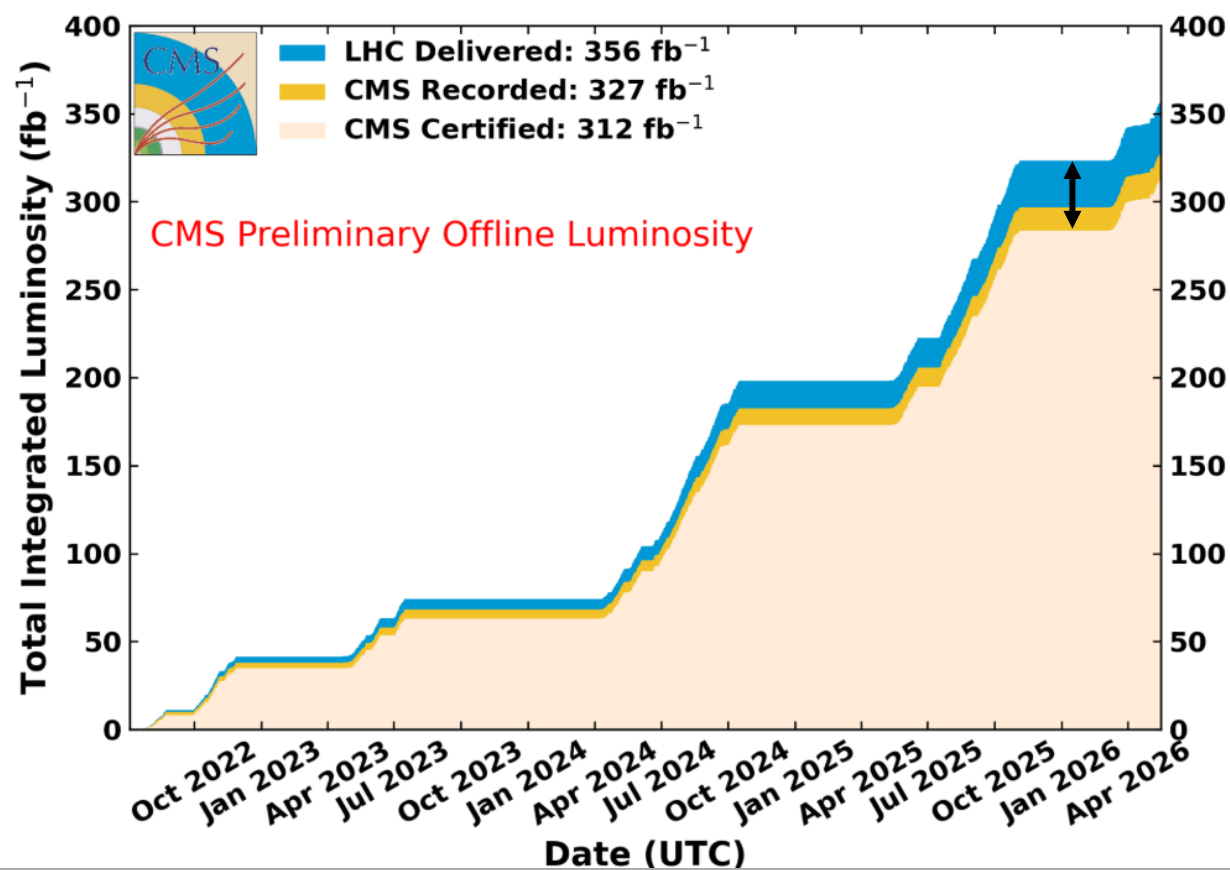
Raw detector hits are combined into physically-meaningful objects: electrons, muons, hadrons, photons

Data is cleaned of bad bursts in near-real time

THE ENDGAME

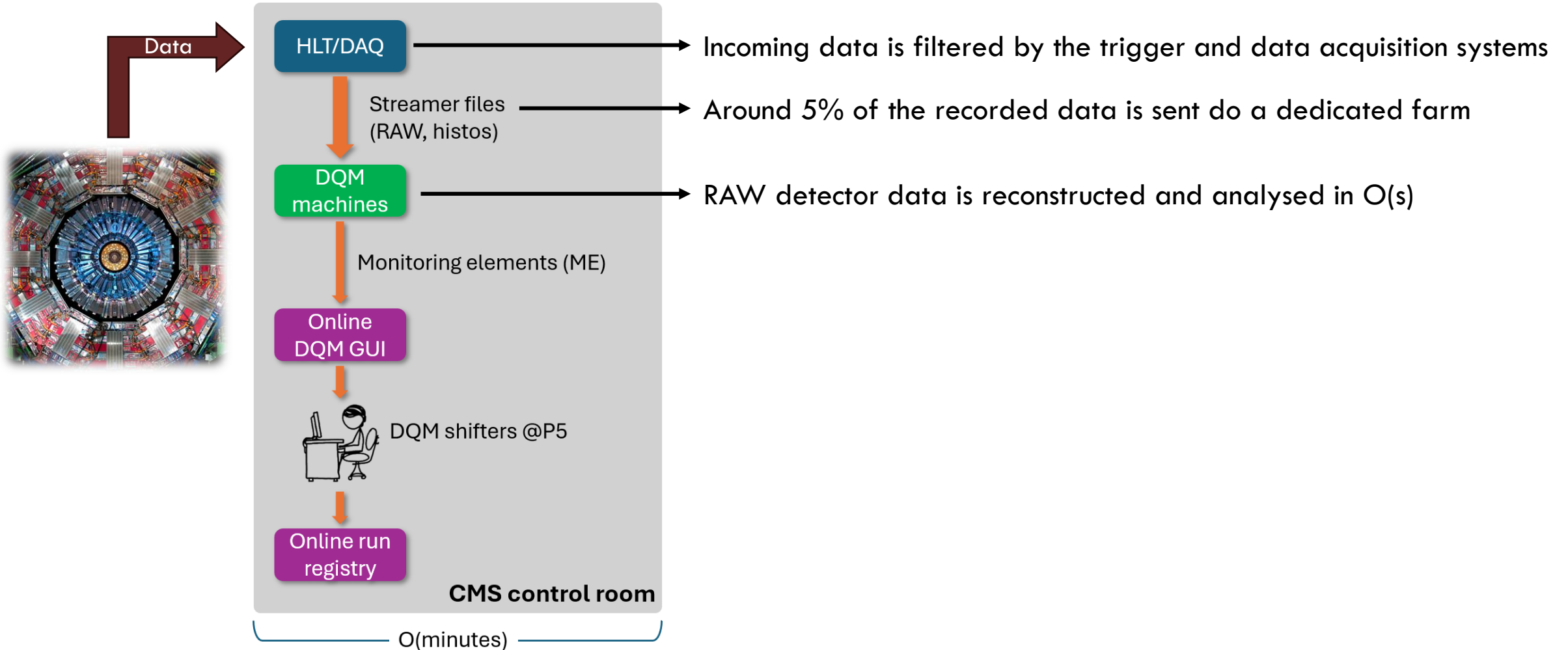
CMS Integrated Luminosity, pp, 2022-26, $\sqrt{s} = 13.6$ TeV

Date included from 2022-07-05 14:47:25 to 2026-05-16 22:27:37 UTC

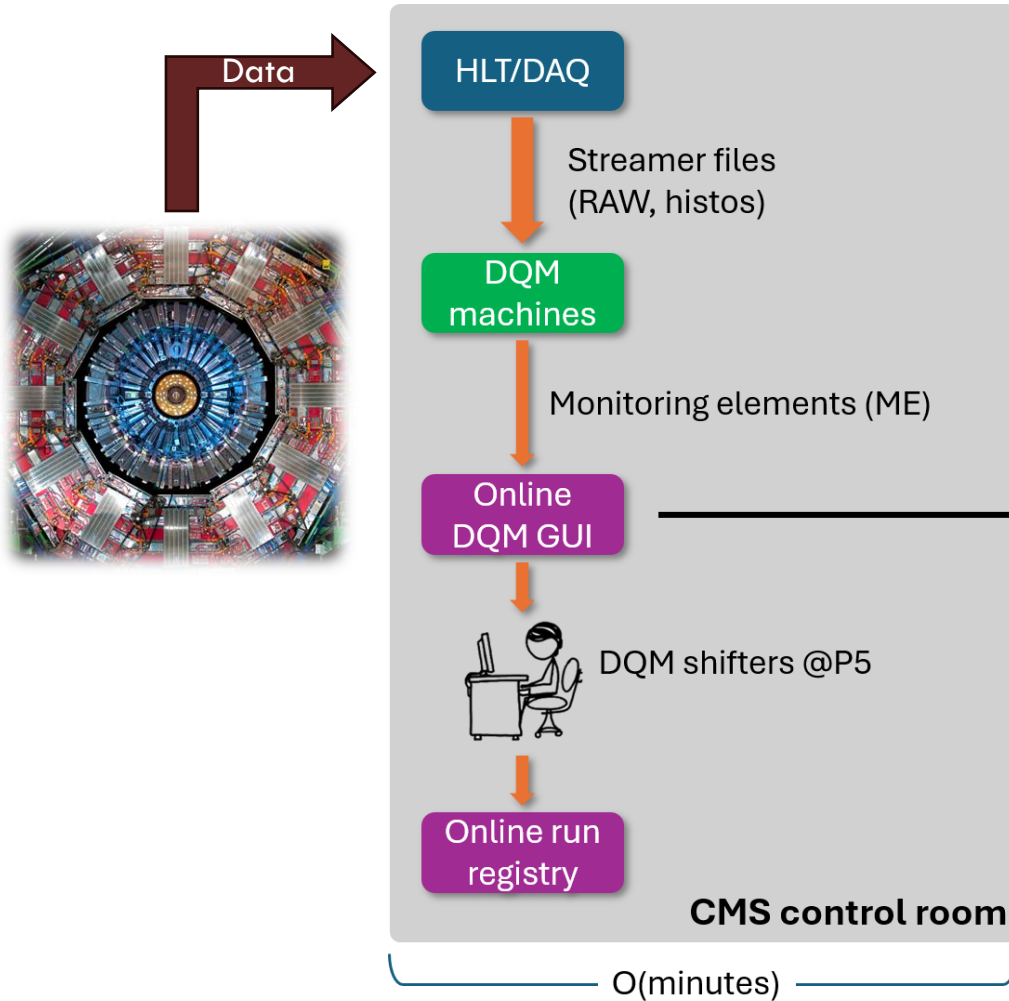


Minimize this gap!

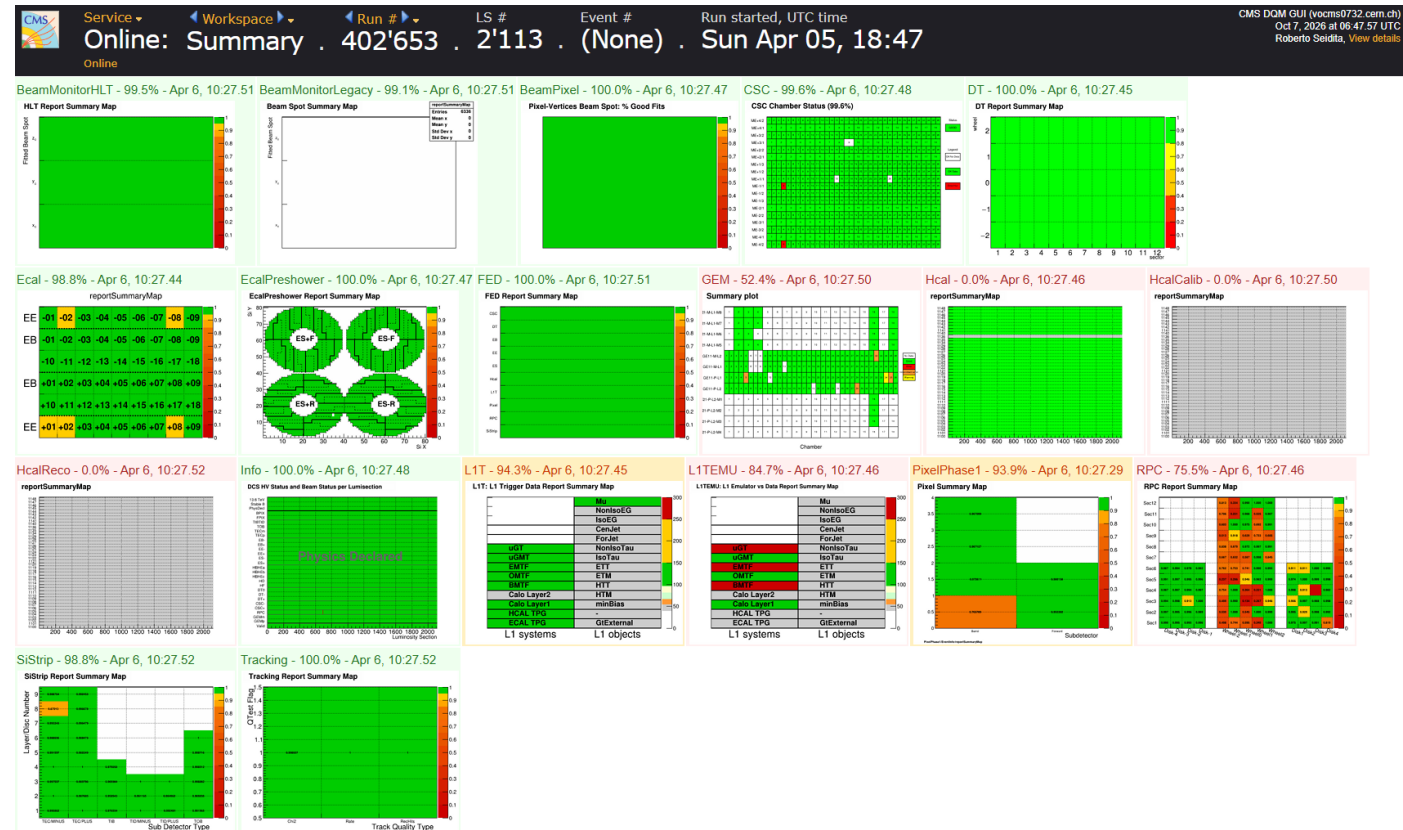
CMS DATA QUALITY MONITORING



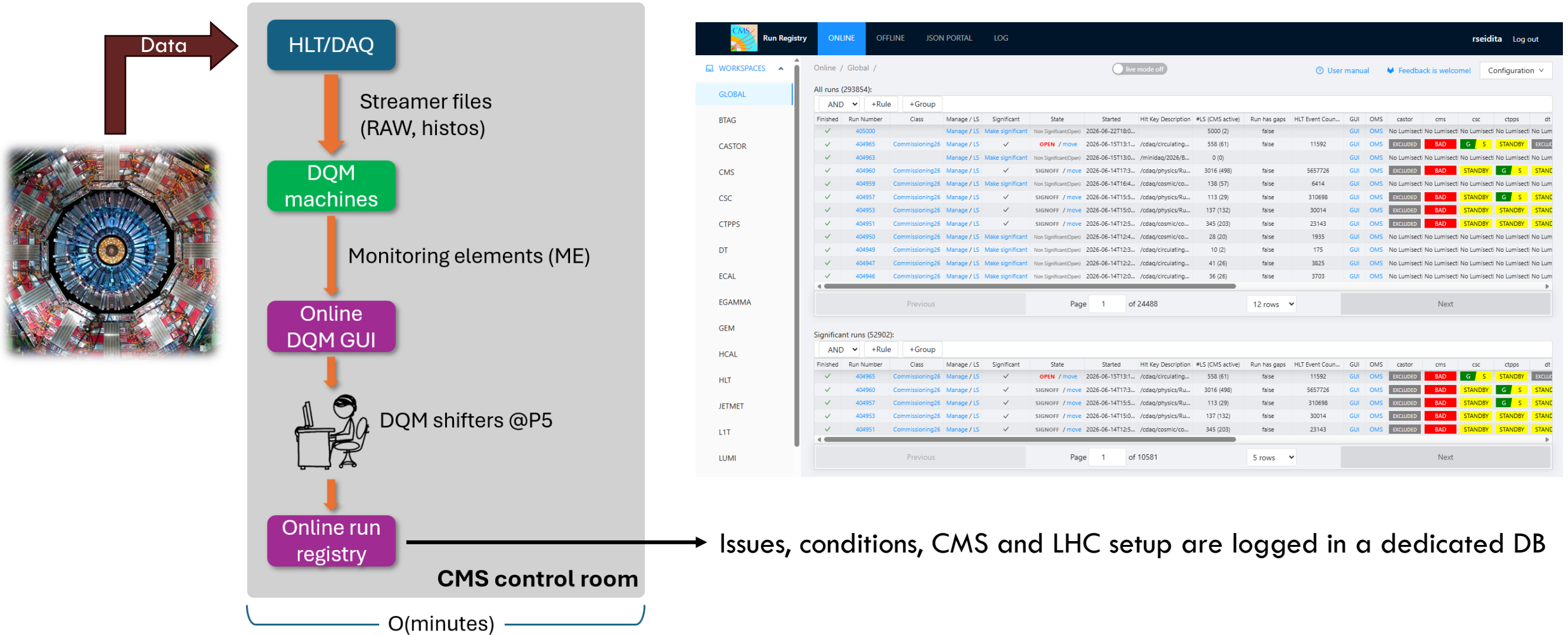
CMS DATA QUALITY MONITORING



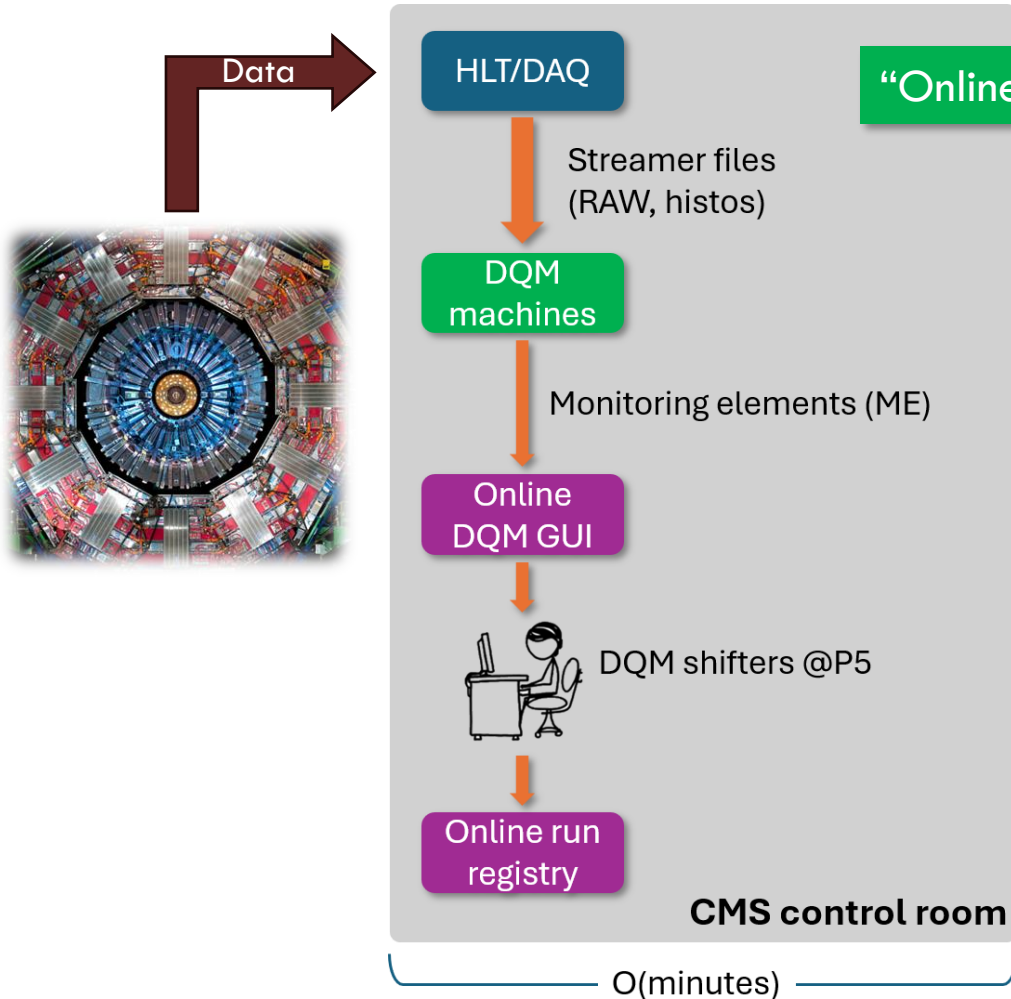
Observables are visualized for expert inspection. Automatic alarms (based on predefined conditions) are raised in case of problems



CMS DATA QUALITY MONITORING



CMS DATA QUALITY MONITORING

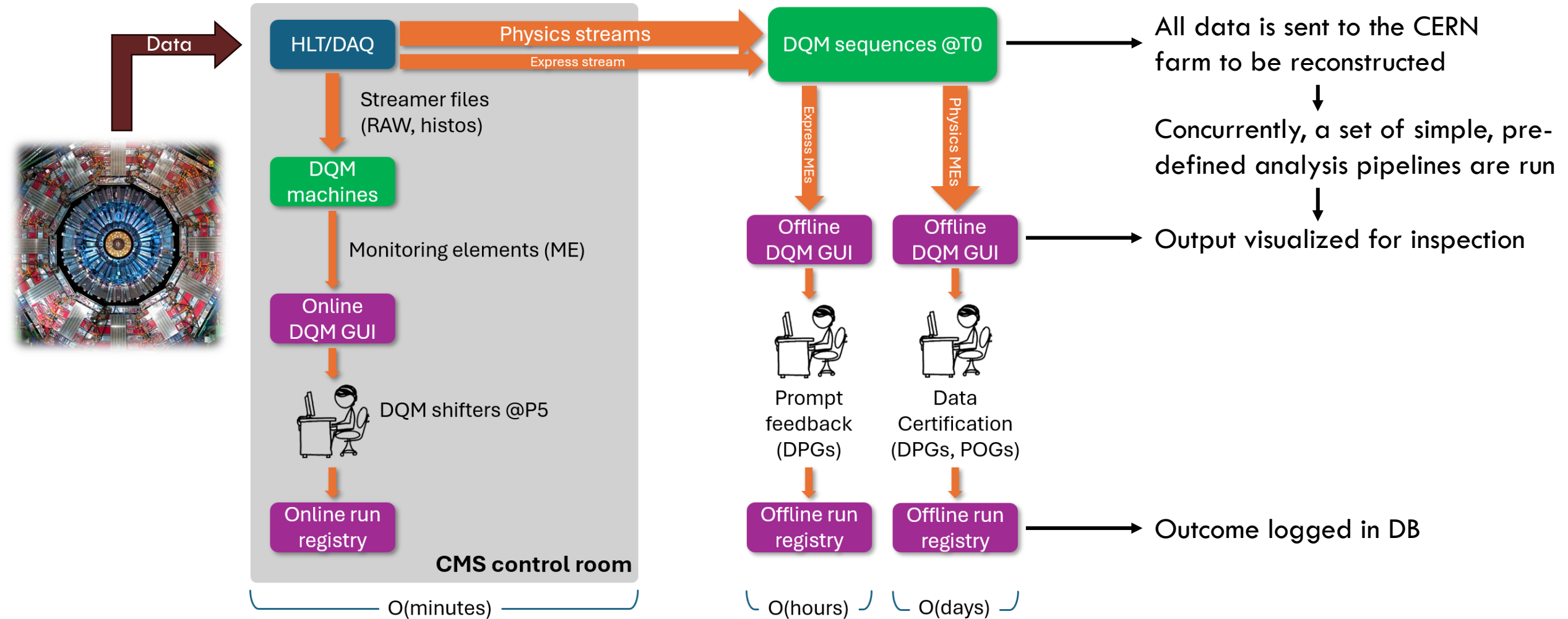


Main goal: promptly identify issues for experts to act as quickly as possible → minimize data loss

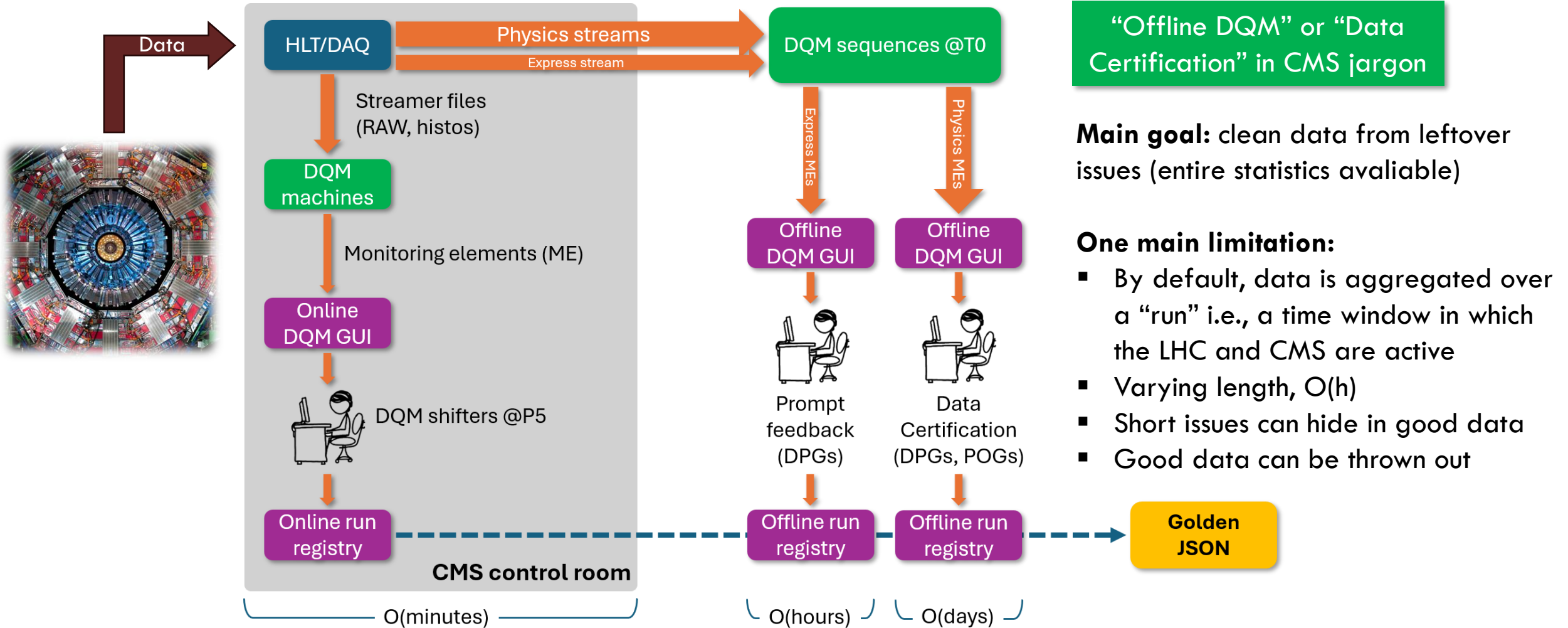
Two main limitations:

- One shifter oversees everything for 8h → fatigue and attention span play a role (automatic alarms help to some extent)
- Limited statistics at around 5% of all recorded data (limited by compute)

CMS DATA QUALITY MONITORING



CMS DATA QUALITY MONITORING



A PROBLEM OF SCALE

The problem

CMS generates an enormous quantity of rich data

For every run (i.e., $O(h)$ of data), $\sim 10k$ observables are plotted for inspection by experts

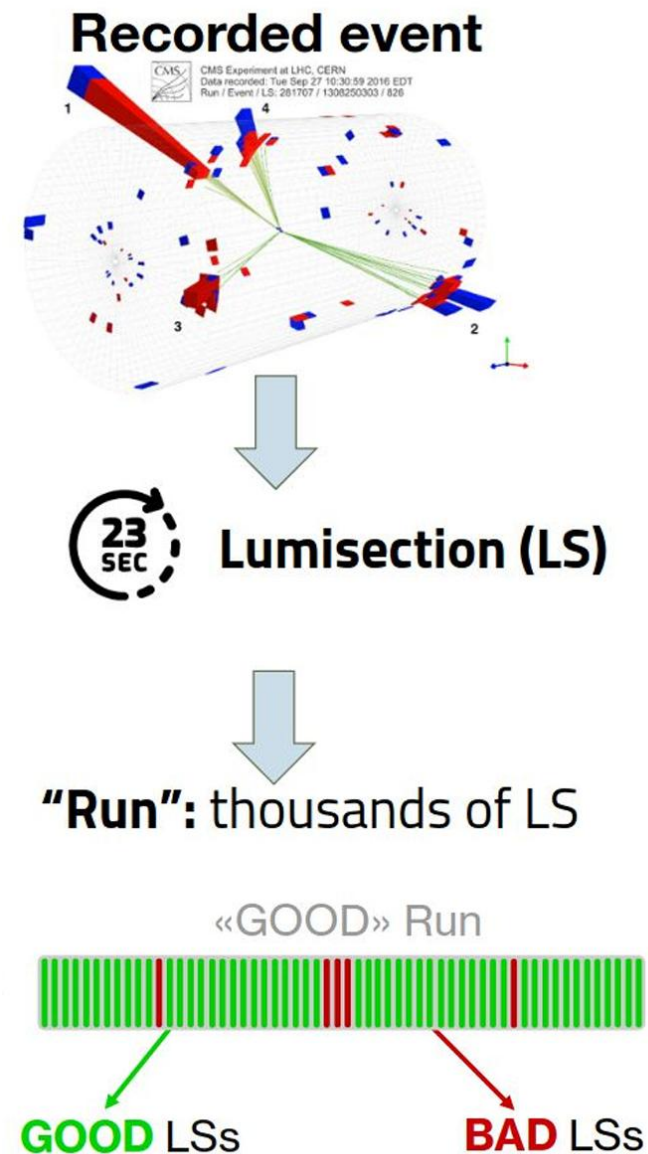
→ Ideally, go finer in time → limited by sheer volume of data

AD is the perfect tool for the job

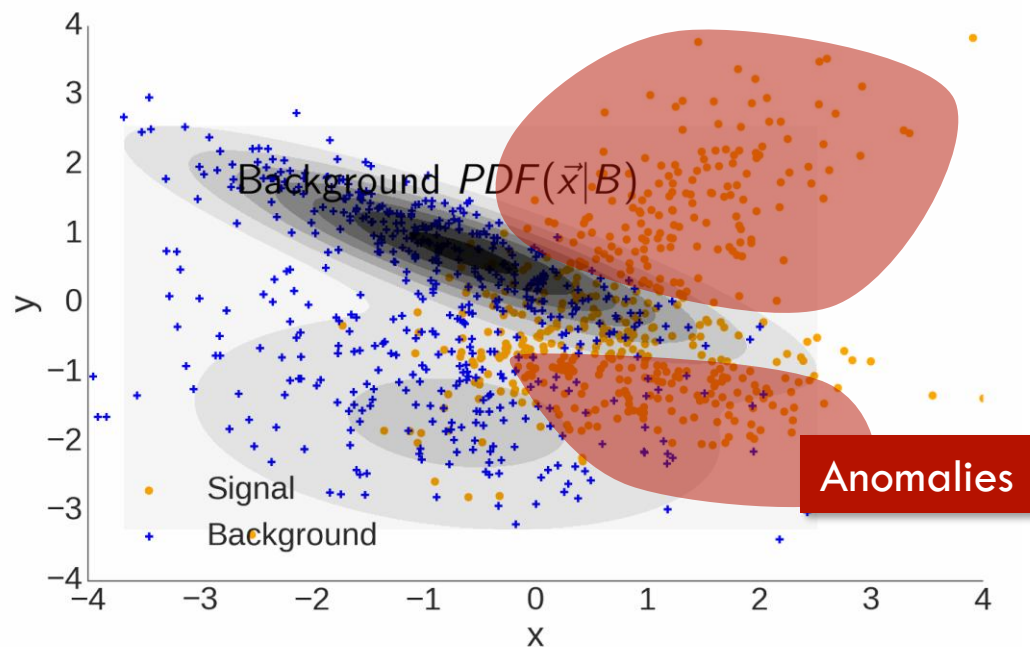
Learn what normal data looks like

Flag issues at arbitrary granularity in near-real time

First steps in this direction already around 2017 [[10.1007/s41781-018-0020-1](https://doi.org/10.1007/s41781-018-0020-1)]



ANOMALY DETECTION IN A NUTSHELL



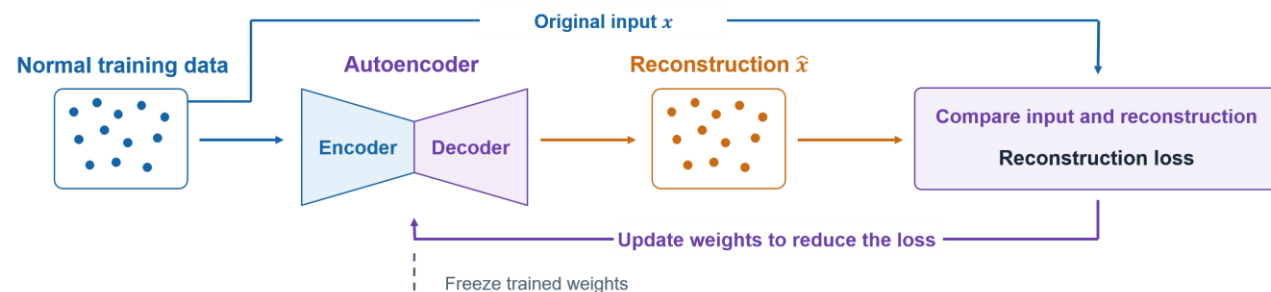
- Model the distribution of “normal” data
- Flag examples that are “different”

→ Falls under the umbrella of unsupervised or semi-supervised ML

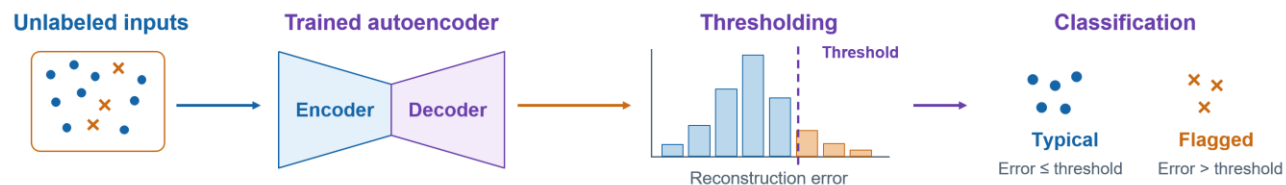
THE ANOMALY DETECTION PARADIGM

Modern unsupervised machine learning gives us a powerful new tool!

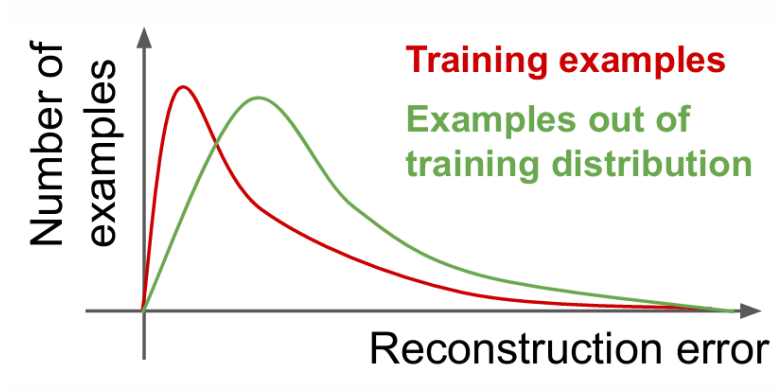
Training



Inference



Underlying assumption: the model learns a “meaningful” representation of the “normal” data, and will fail to encode-decode data with a different structure

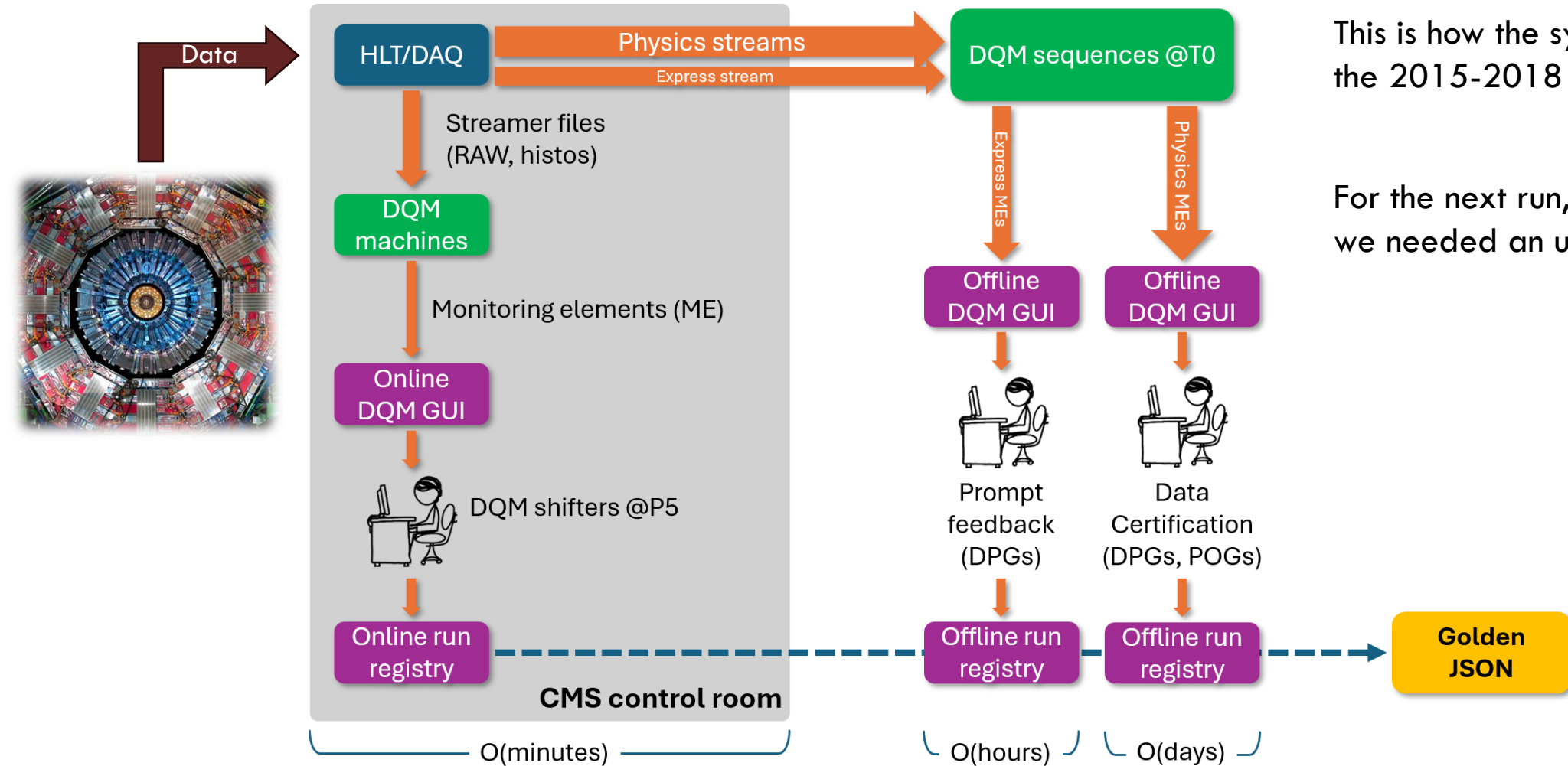


Why? Train only on normal data, no need to specify what anomalies look like a priori

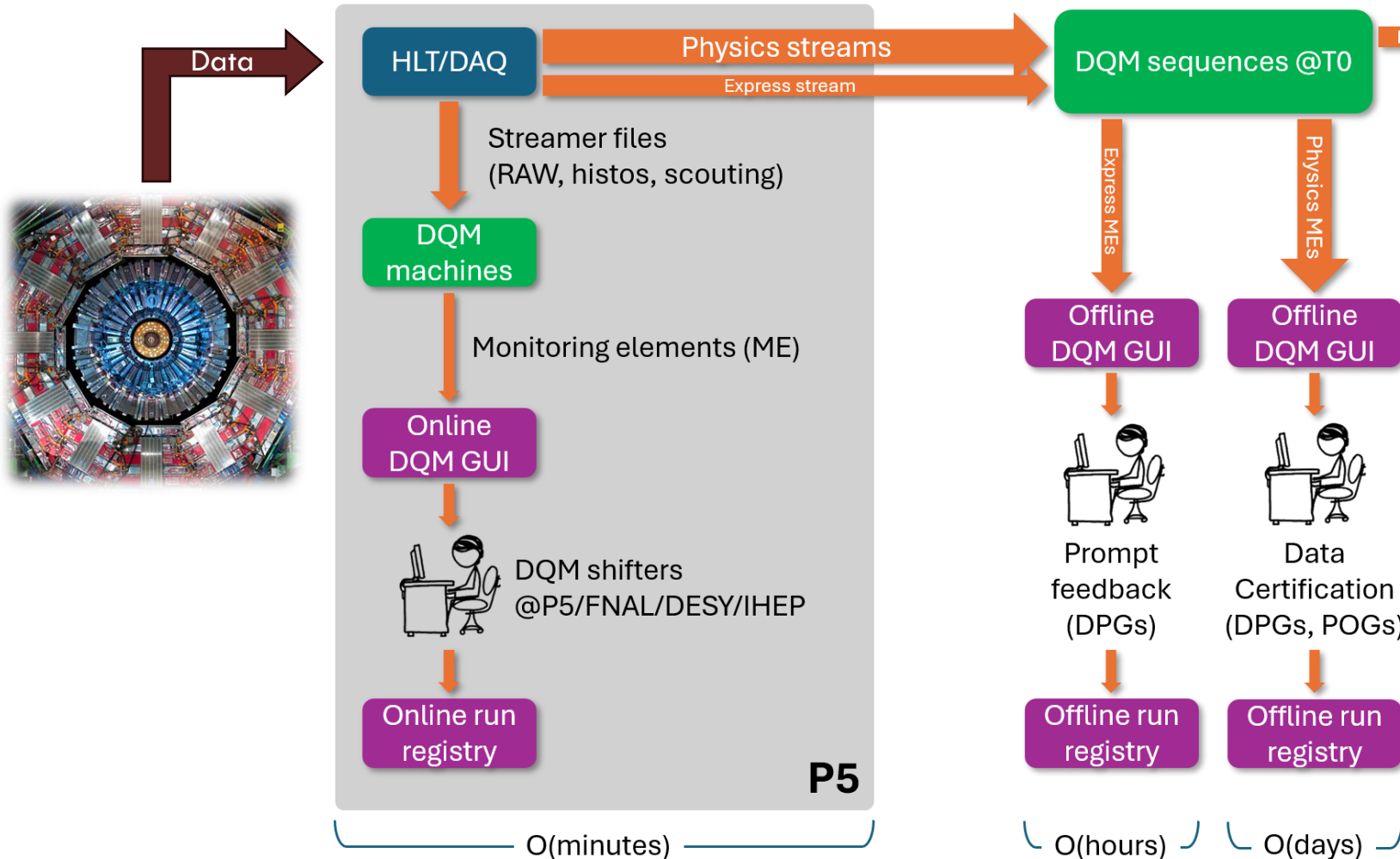
UPGRADING THE PIPELINE

This is how the system operated during the 2015-2018 data taking period

For the next run, from 2022 to 2026, we needed an upgrade



UPGRADING THE PIPELINE



A selection of metrics are aggregated and store every 23s instead of an entire run

Twofold aim:

- Database with high-granularity data for AD model training
- Host trained models to inference on new data in near-real time

Two "modes":

- **Purity:** clean up transient issues
- **Efficiency:** minimize loss of good data

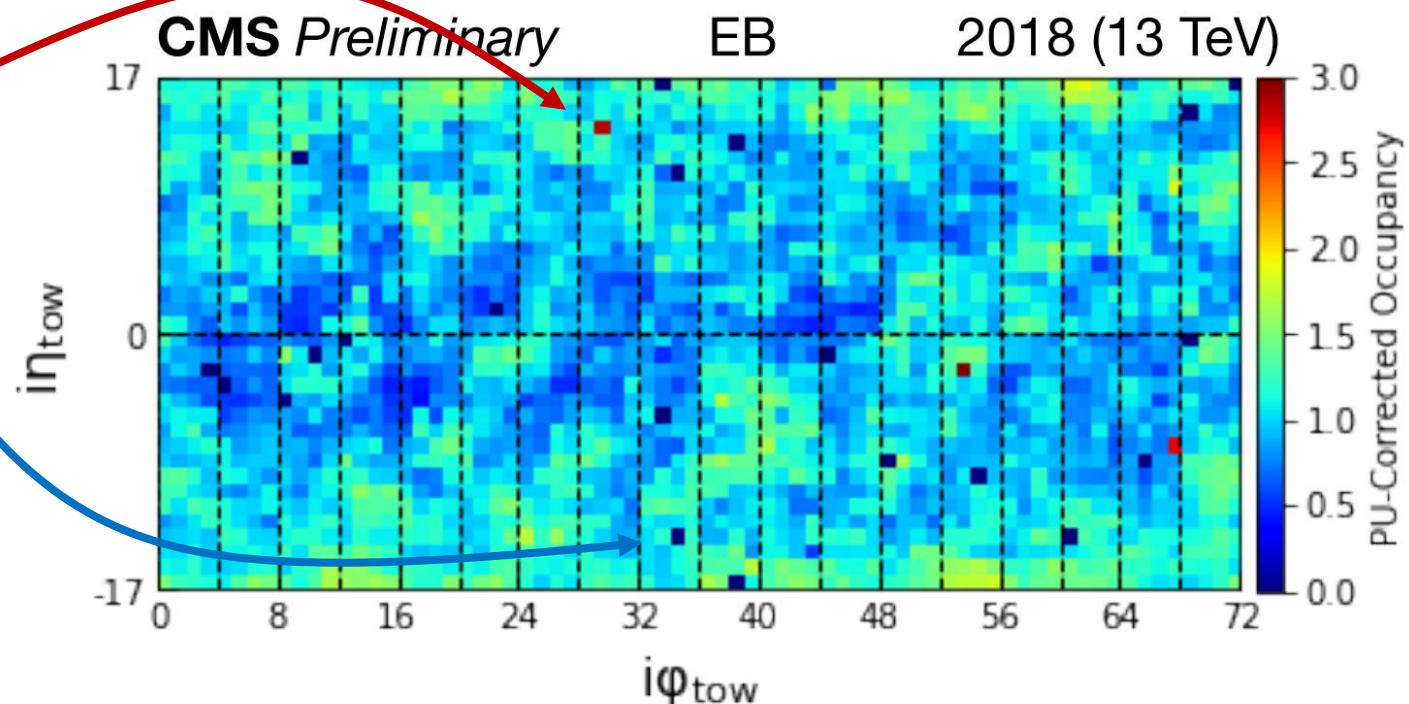
An online example

The electromagnetic calorimeter

THE CMS ECAL

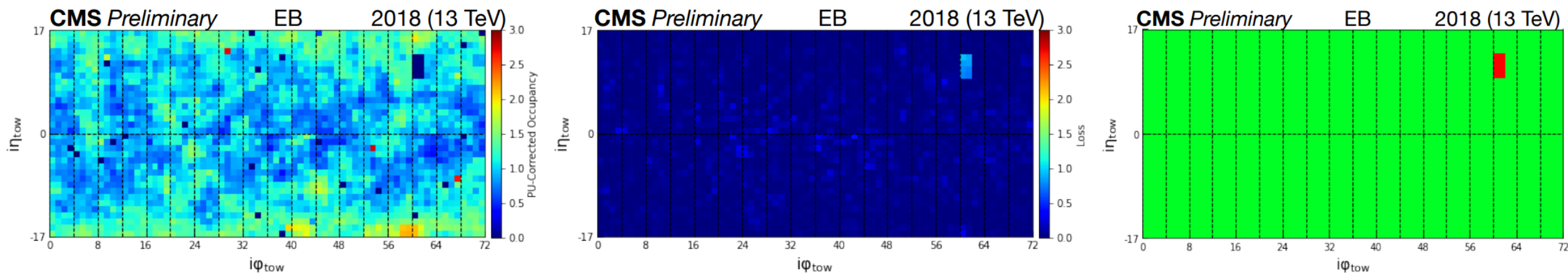
75'848 PbWO_4 scintillating crystals, each read out individually

- Two typical problems:
 - **Noisy cells** manifest as higher than average occupancy
 - **Low-response cells** manifest as lower than average (or zero) occupancy
- This data is gathered with LS granularity i.e., every 23 seconds, in the online regime
- **Anomalies can be missed by human operator over hours of shift**



THE CMS ECAL

- Train an AE to reconstruct the occupancy map as an image
- ResNET-based architecture with conv layers for both the encoder and decoder
- The pixel-by-pixel reconstruction error (loss map) makes anomalies evident
- Lower false positive rate w.r.t. simple threshold on cell occupancy



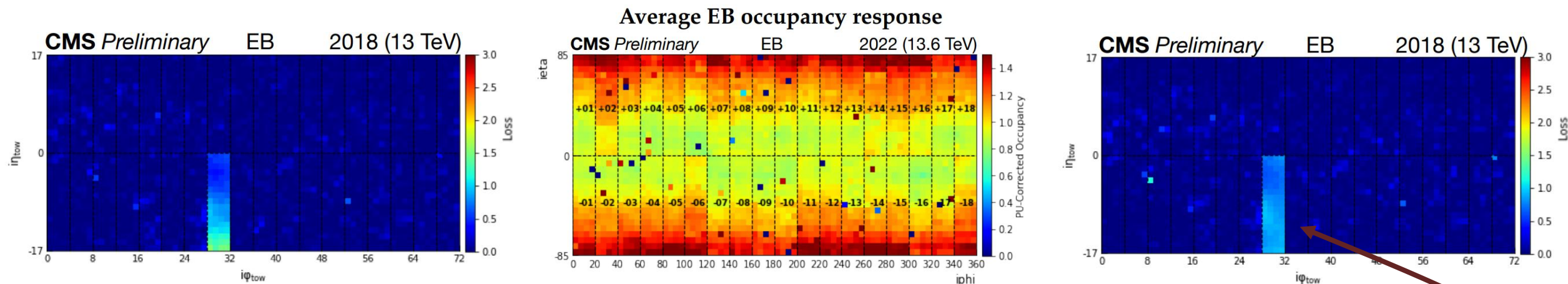
Original occupancy map

Loss map of the AE

Loss map after thresholding

THE CMS ECAL

- Correct processing of the data is crucial (*garbage in, garbage out*), for example:
- Because of collider physics, high pseudorapidity (η) regions have higher occupancy
- But we still want the same relative discrepancies to be flagged → normalize by average



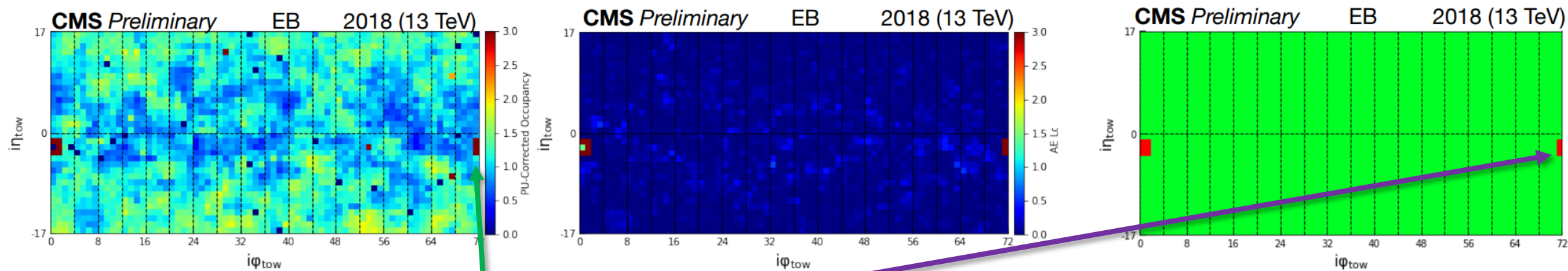
- Whole block off, but
- AE learned to expect more hits at high η , so
- Higher loss at higher η

Normalize loss to average occupancy

Loss is flat over η for the same anomaly

THE CMS ECAL

- This technology was deployed in CMS for 2022 data taking
- Strong performance on new data

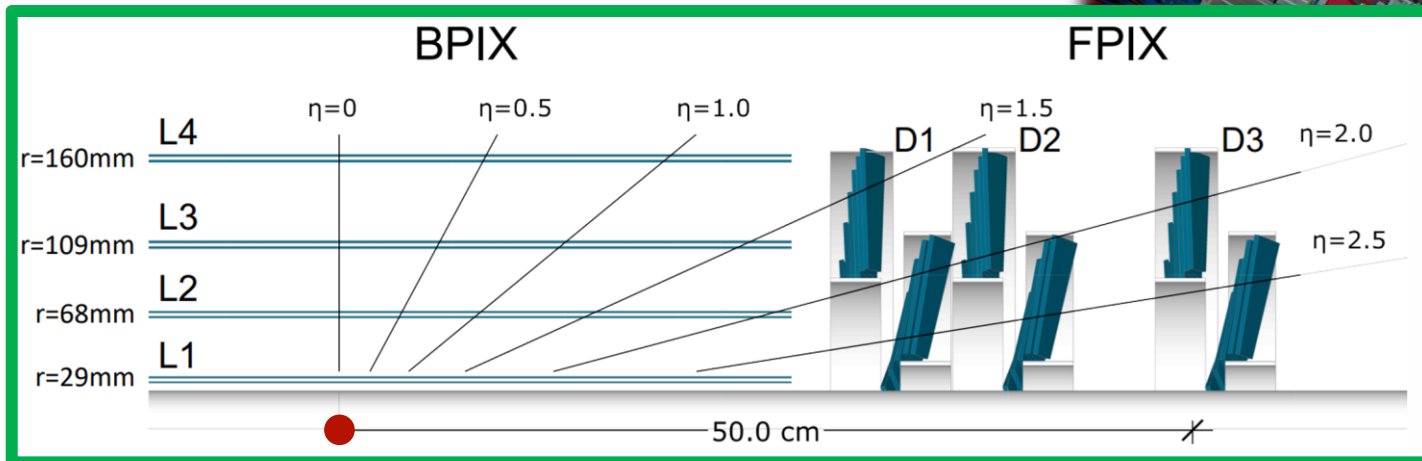


Noticing **this** instead of **this** in real time is much, much easier, especially after 8h of shift :)

A “purity mode” example

The silicon tracker

EXAMPLE: THE CMS SILICON PIXEL TRACKER

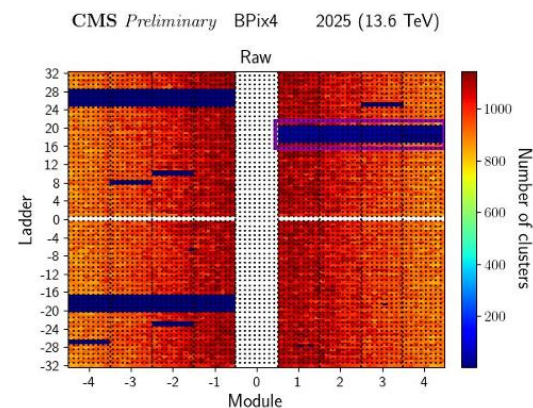
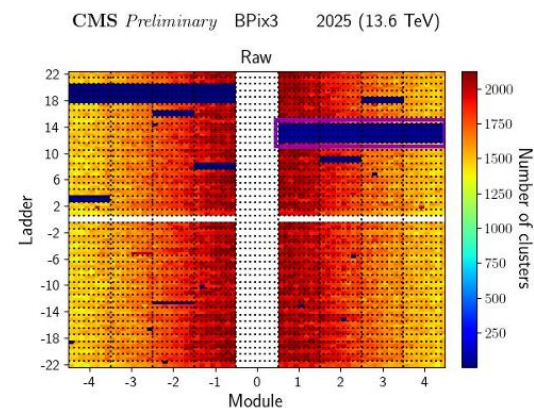
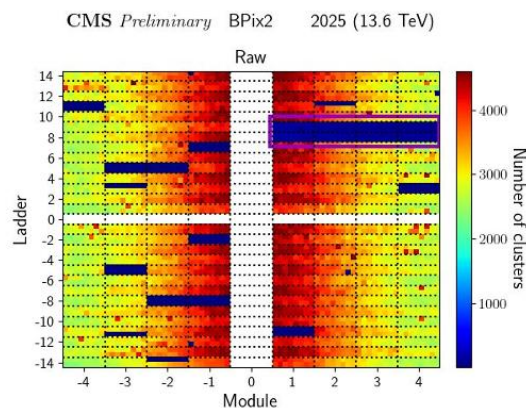
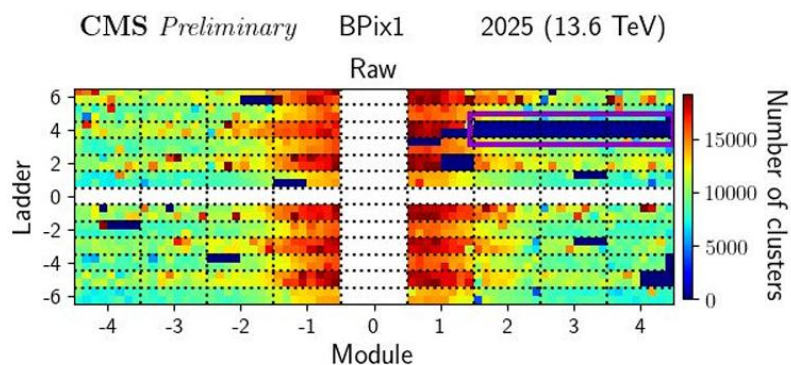
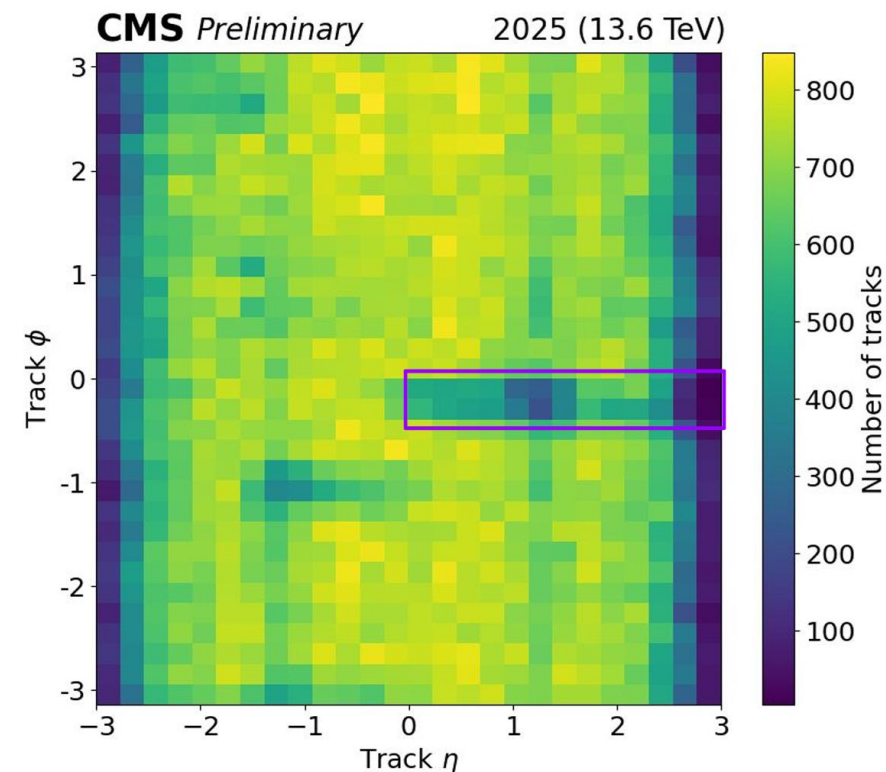
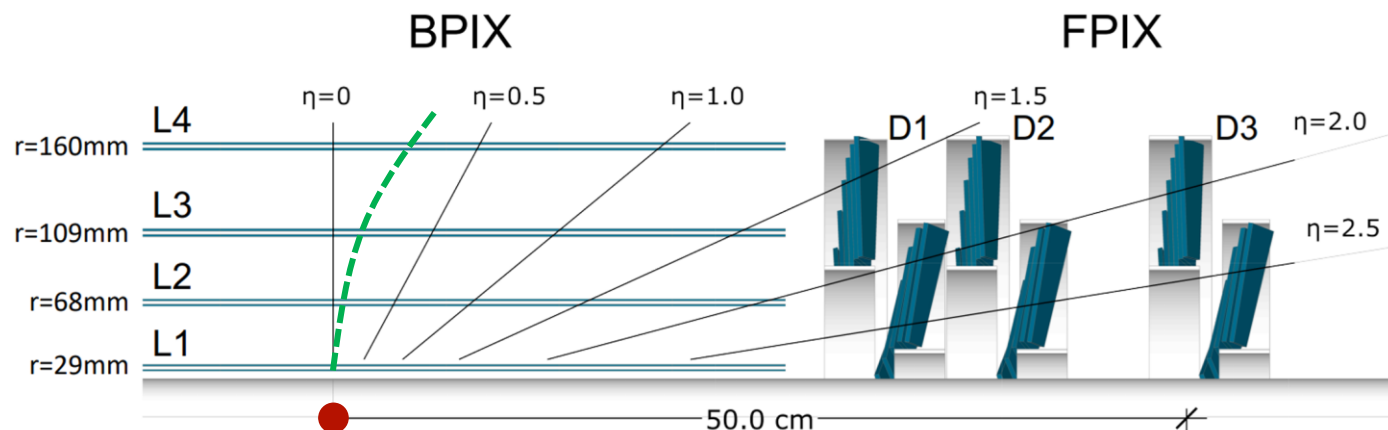


1856 silicon sensors with 160×416 pixels each

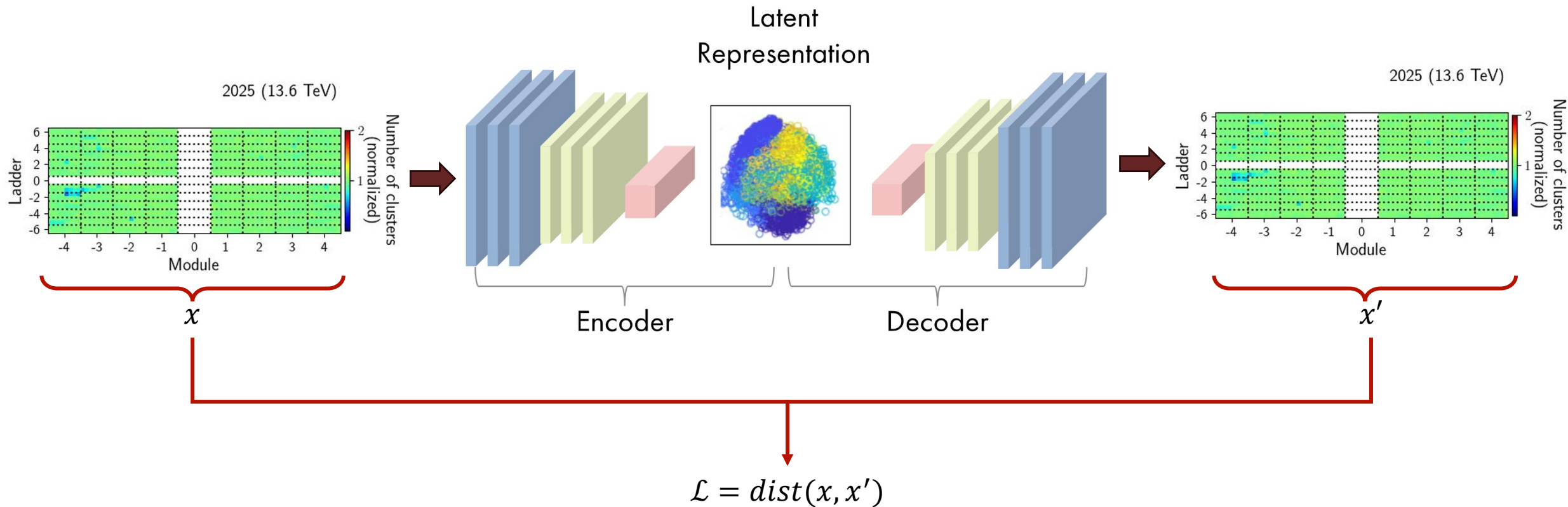
Assembled in cylinders in the central part and disks in the forward part

Sometimes, some modules fail temporarily in unlucky ways

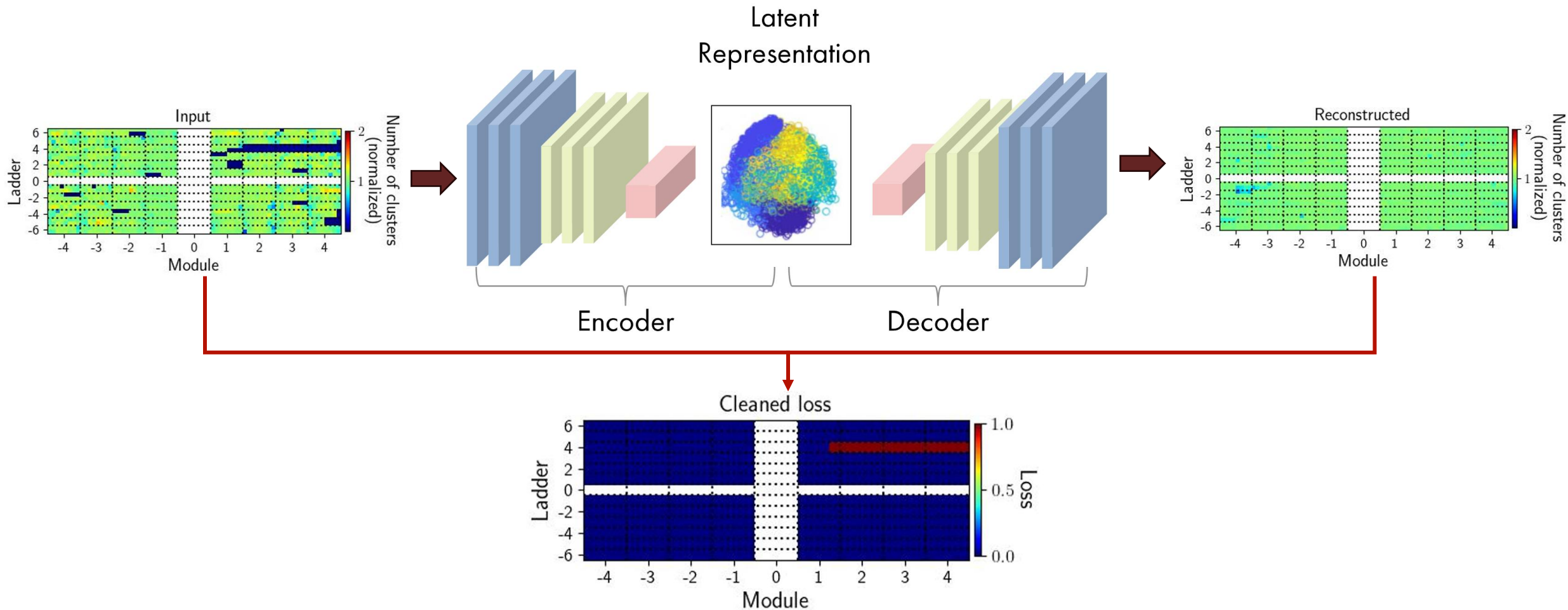
EXAMPLE: THE CMS SILICON PIXEL TRACKER



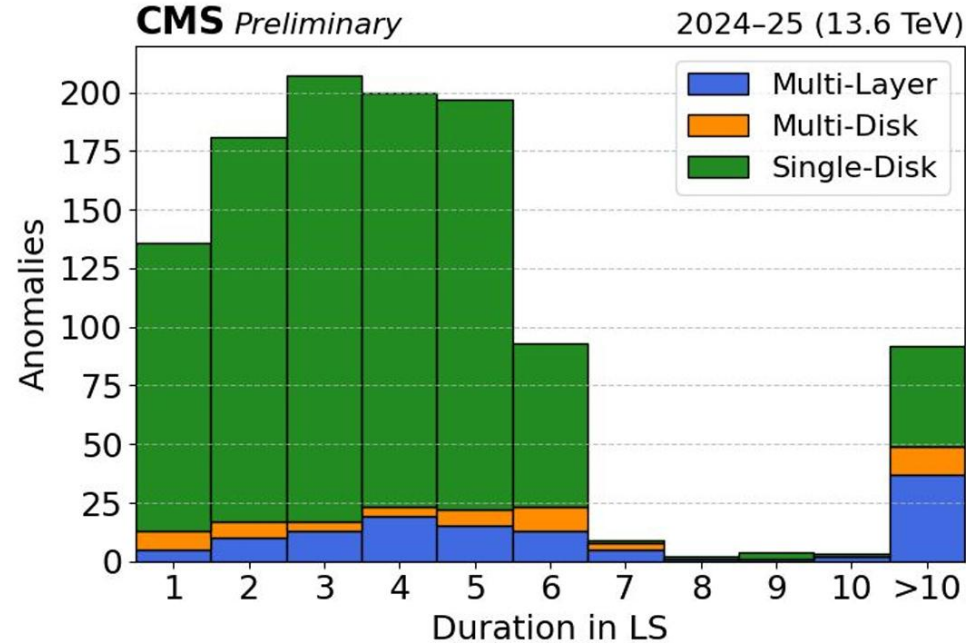
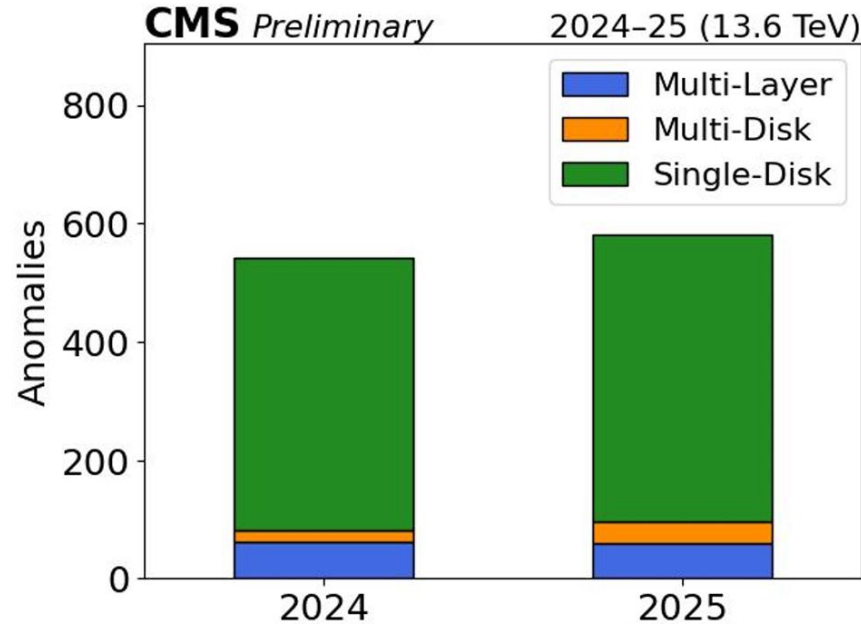
EXAMPLE: THE CMS SILICON PIXEL TRACKER



EXAMPLE: THE CMS SILICON PIXEL TRACKER



EXAMPLE: THE CMS SILICON PIXEL TRACKER



We deployed the system online for 2024 data taking and it has served us very well

→ Remove faulty data with 24" granularity, **only when it affects physics**

→ All automatically!

**Retroactively
applied to all data
collected since 2022**

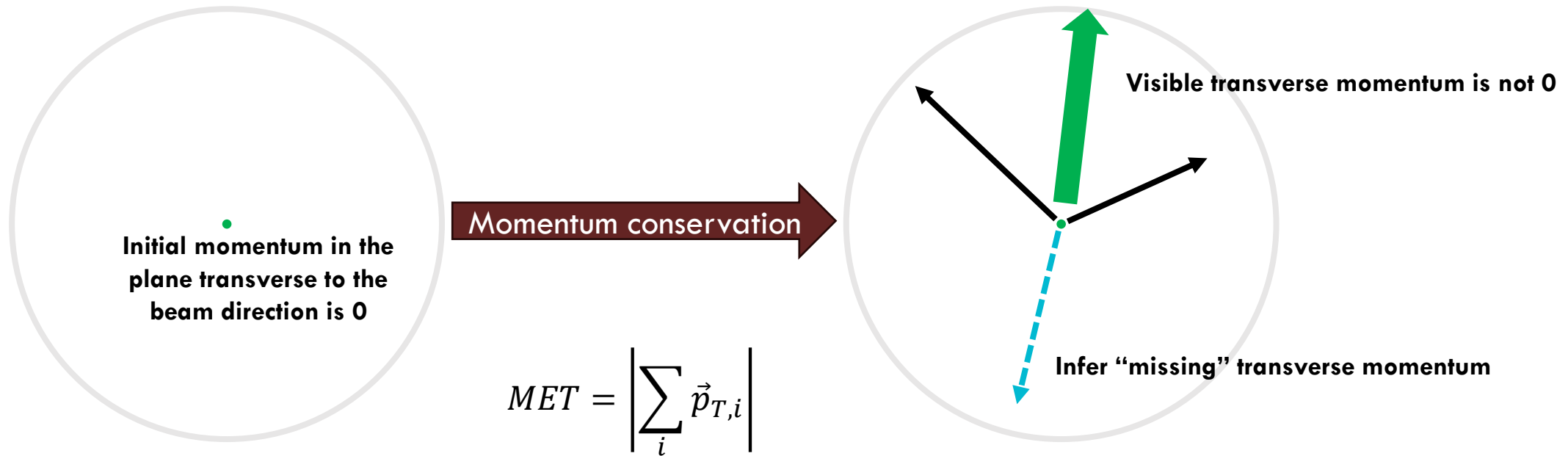
An “efficiency mode” example

Hadronic jets and missing
momentum

MISSING TRANSVERSE MOMENTUM

Some particles (neutrinos, maybe dark matter) cannot be detected in CMS

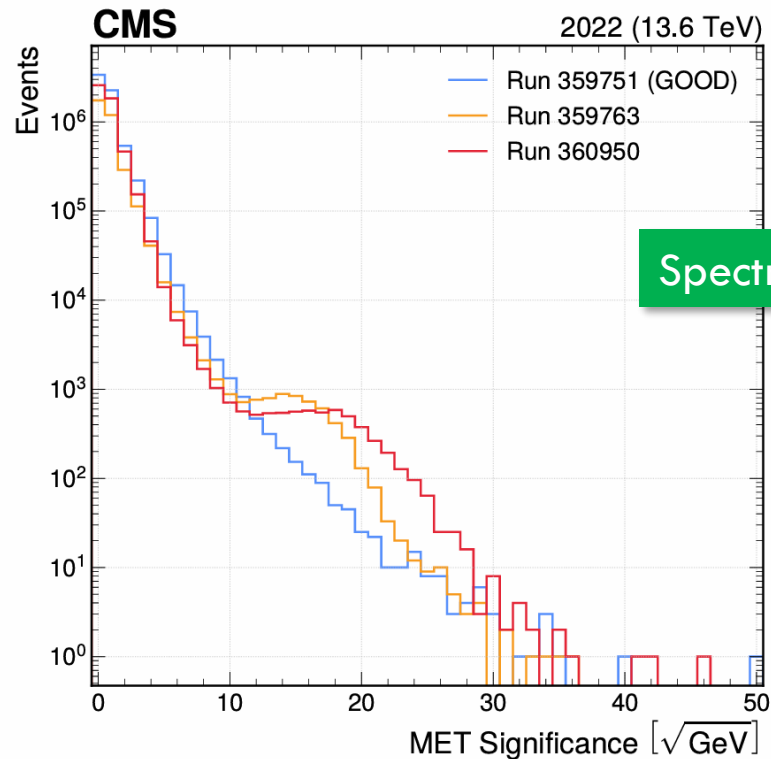
→ We can still indirectly infer something about their presence by exploiting momentum balance



MISSING TRANSVERSE MOMENTUM

MET is a very high-level quantity → literally a product of every single signal in the detector

→ Affected by any issue in principle → powerful diagnostic tool, but traceback can be a challenge



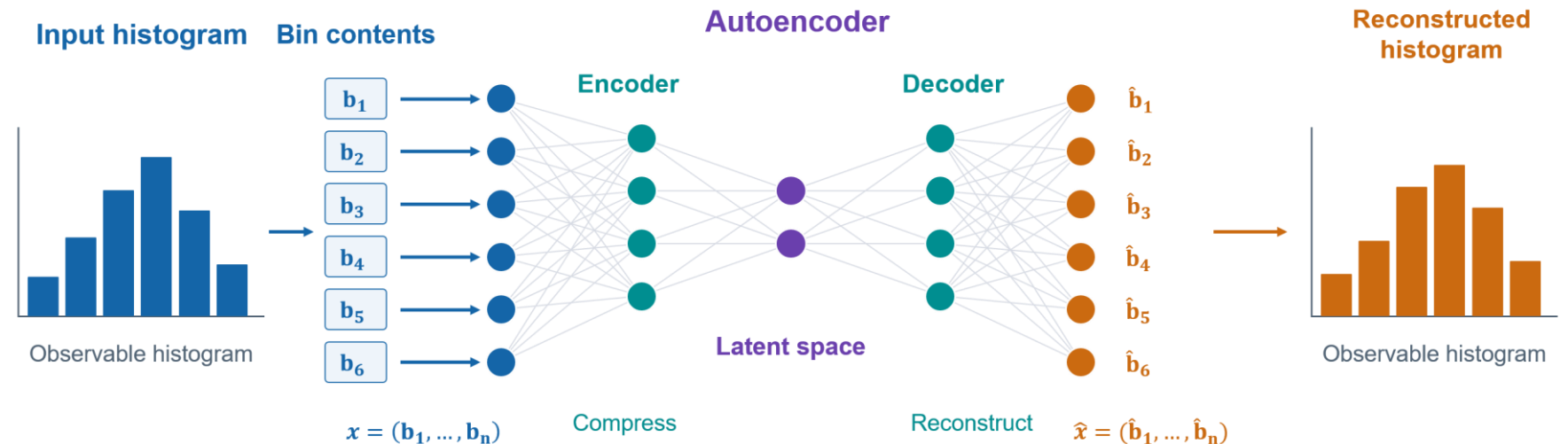
Spectrum obviously unphysical, origin unclear

MISSING TRANSVERSE MOMENTUM

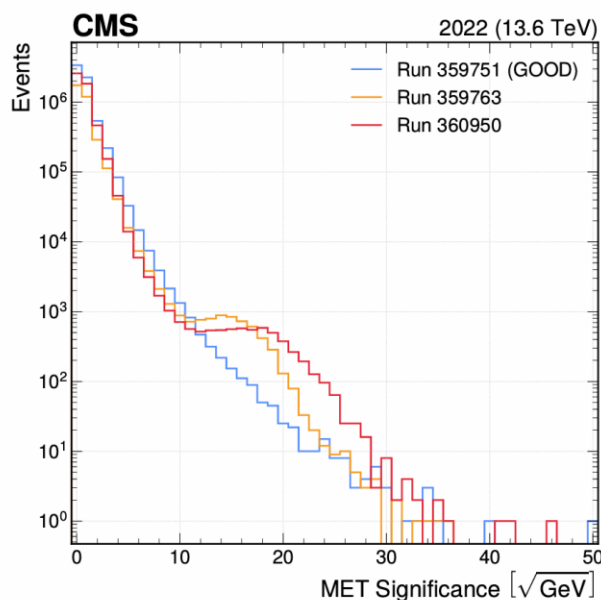
MET is a very high-level quantity → literally a product of every single signal in the detector

→ Affected by any issue in principle → powerful diagnostic tool, but traceback can be a challenge

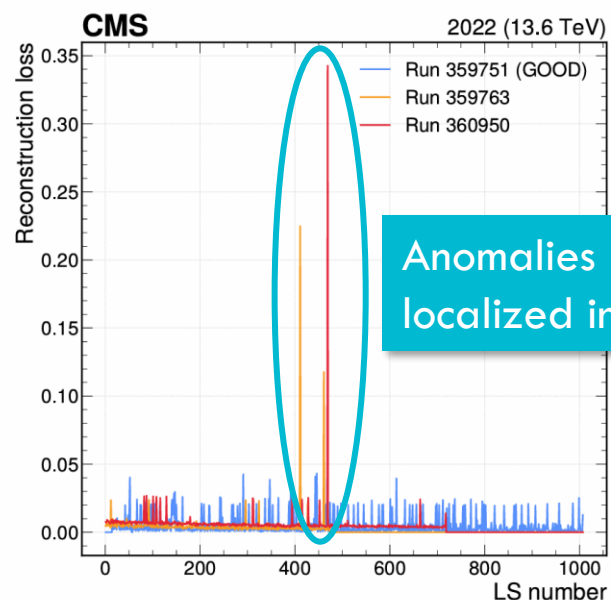
- Long-short-term-memory autoencoder trained on sequences of good data slices
- Each example is a histogram filled over 23s of data recording



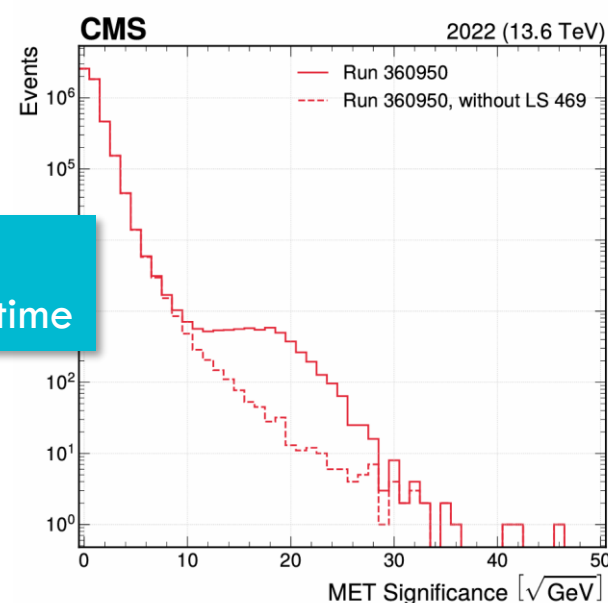
MISSING TRANSVERSE MOMENTUM



Some runs show clear issues

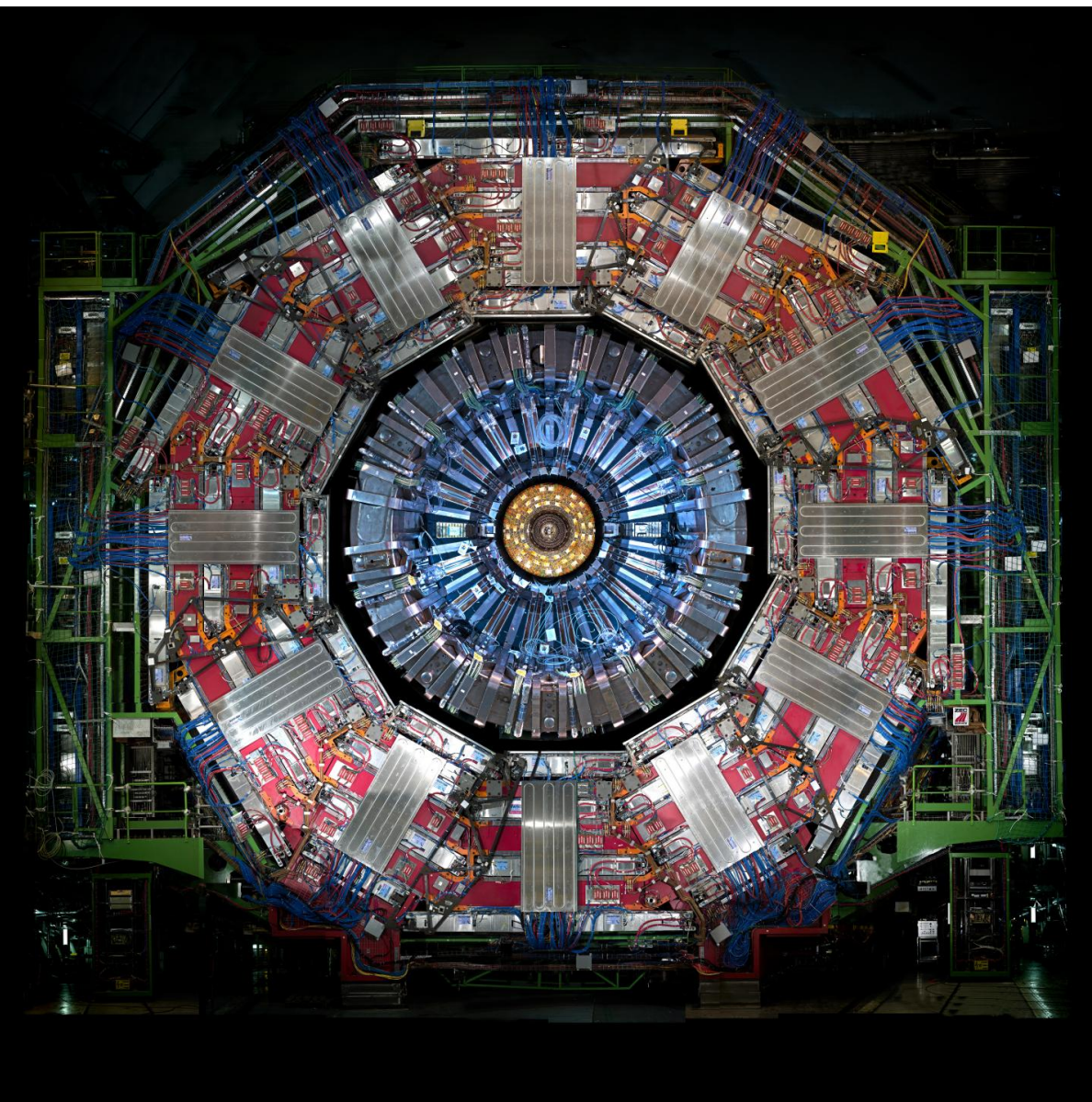


AE can quickly scan
the run in 23s slices



Removing 1 out of
730 LSs fixes the issue

Systematically pinpoint localized issues in time, surgically remove them



SUMMARY

- Anomaly detection is proving to be a powerful tool in data monitoring and curation in HEP
- Some references: [ECAL](#), [Tracker](#), [AutoDQM](#), [JME](#)
- Currently in shutdown, gearing up for 2030
- Plan to redesign the system in an “ML first, expert later” scheme

THANK YOU

*« Ce qui est admirable, ce n'est pas que
le champ des étoiles soit si vaste,
c'est que l'homme l'ait mesuré. »*

Jacques Anatole François Thibault

