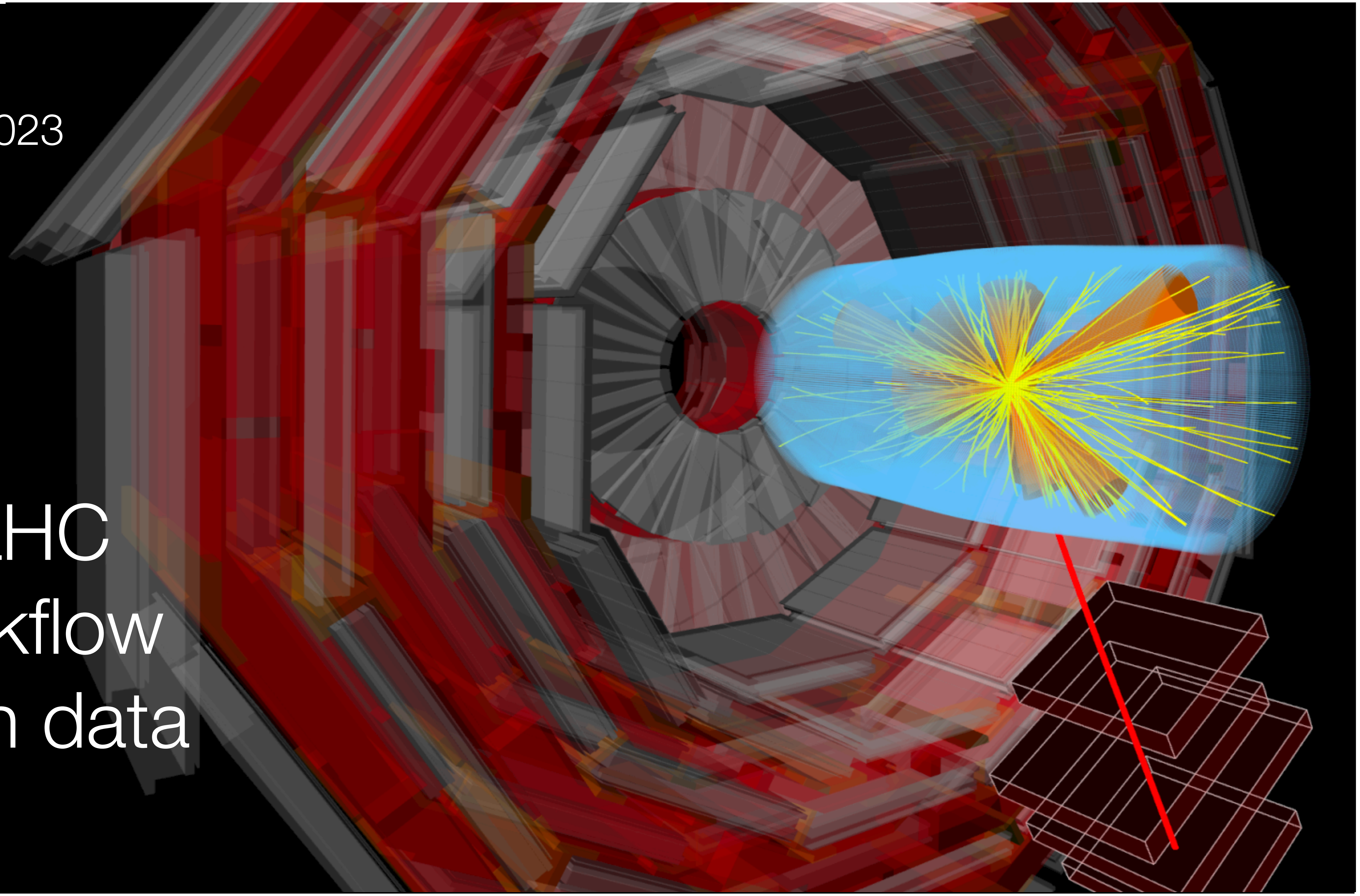
ETH zürich

IPA

Annapaola de Cosa
IPA ML Workshop 22 March 2023

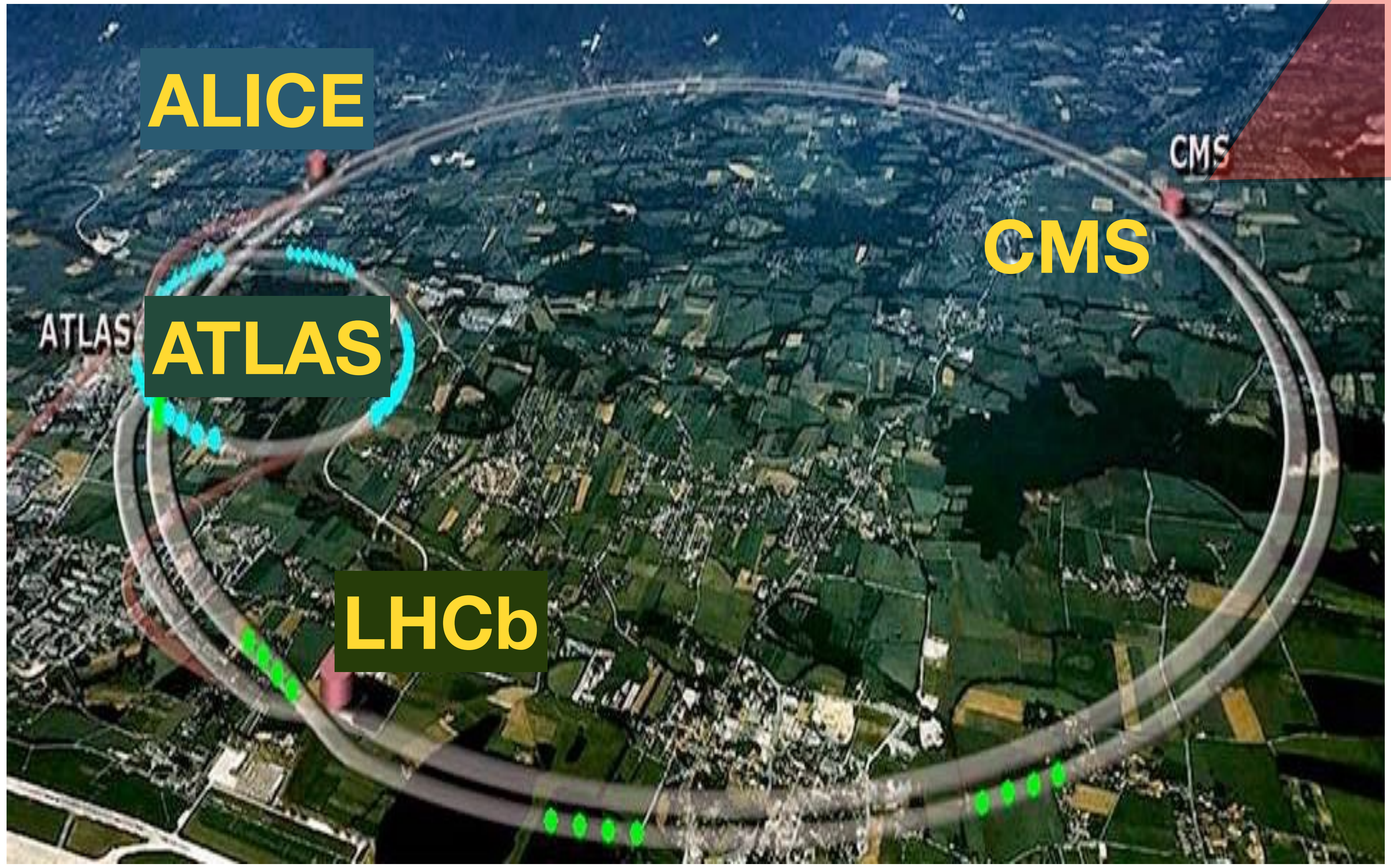
Introduction on LHC experiments workflow with emphasis on data challenges

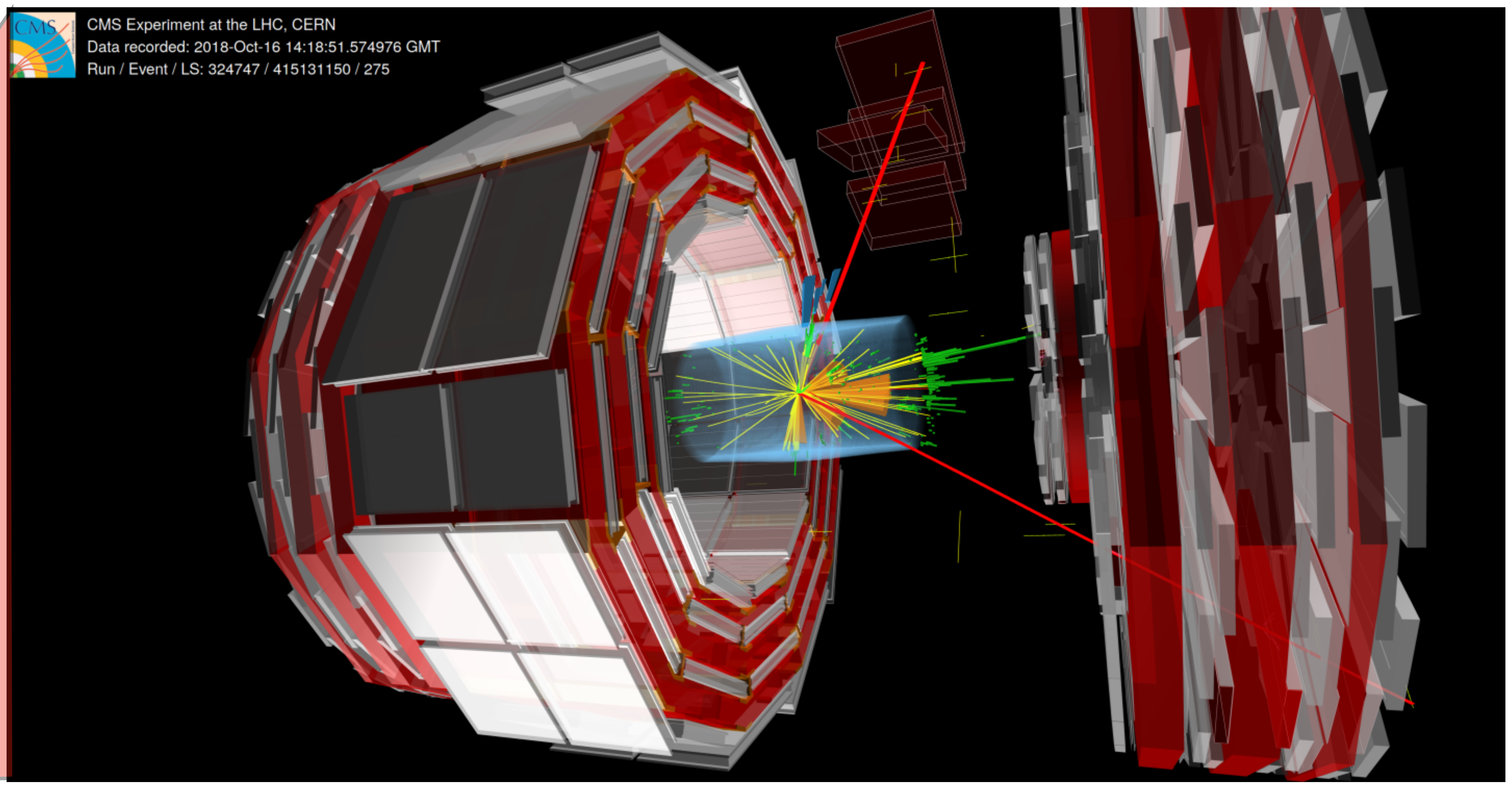


LHC and its experiments

The Large Hadron Collider is the largest and most powerful particle accelerator

- Collides beams of protons up to **13.8 TeV**
- Enable investigation at the TeV scale
- Proton bunches collided **every 25 ns**





Broad physics programme

- Measurements of SM processes
  - Higgs, Flavour, EWK physics, …
- Search for new physics
  - SUSY, Hidden Valleys, Dark Matter, …

The high intensity challenge

Physics of interest usually quite rare

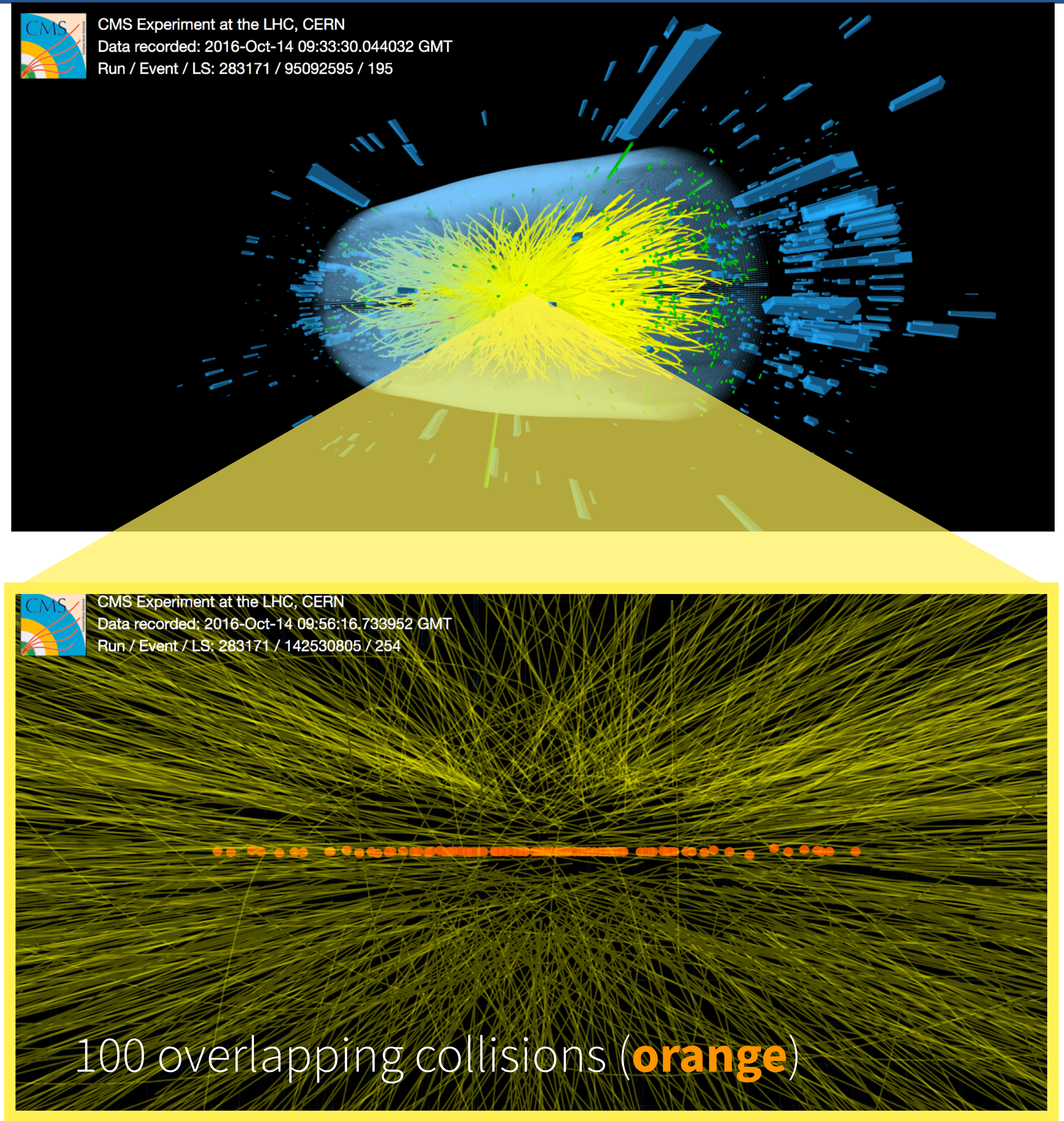
- Rates ~10-15 orders of magnitude lower than most common background processes

High collision rate increases probability to observe physics of interest

- **600 million collisions/s**
- **Tens overlapping collisions (pileup)**
- Just few containing interesting particles
- Interesting physics look very similar to background

Challenges:

- **need to handle large amount of particles,**
- **disentangle collision products,**
- **identify interesting physics in a see of particles**



The magnitude of the data problem

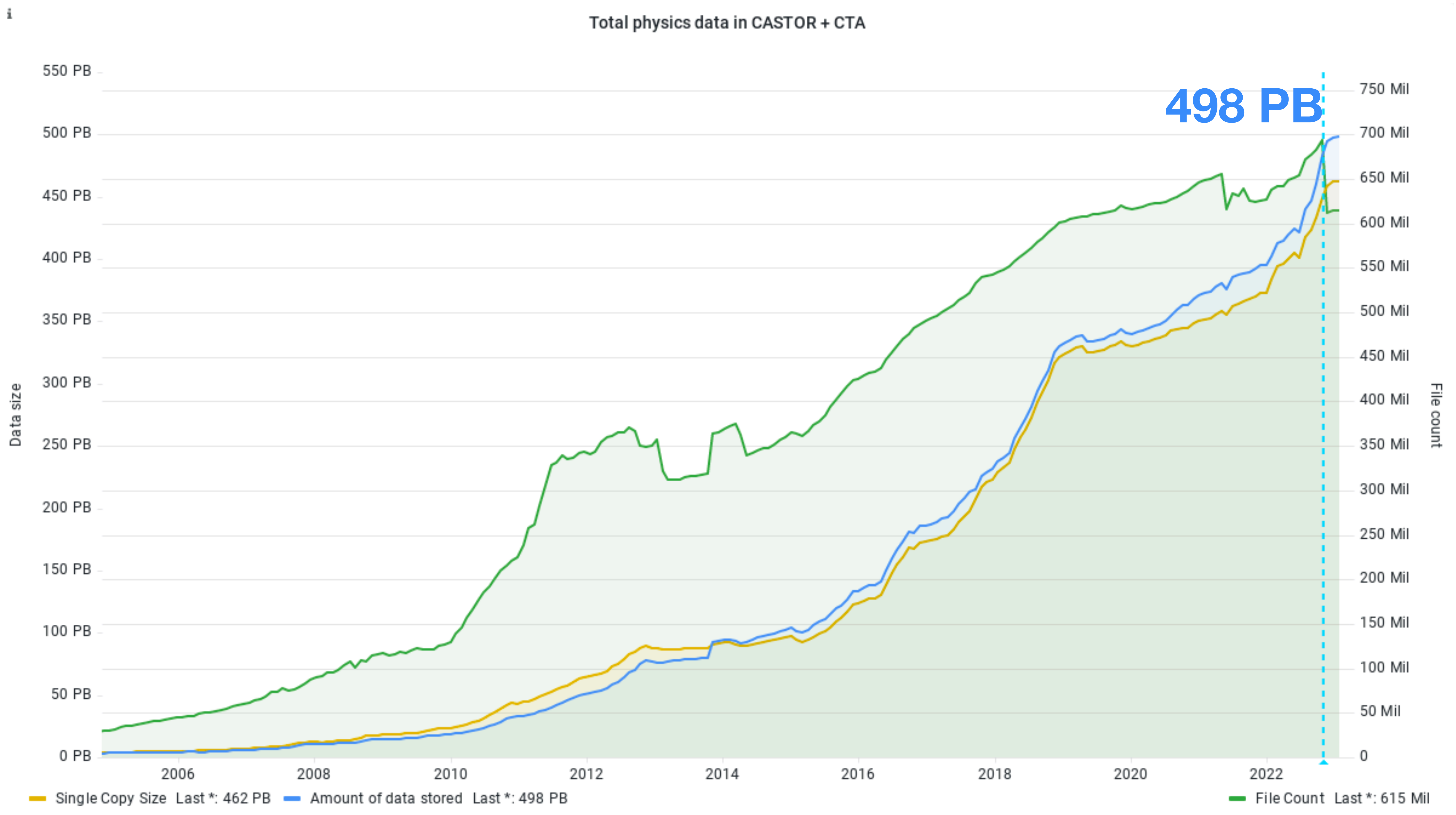
High collision rate increases probability to observe physics of interest

- ~600 million collisions/s
  - ➡ **~ 1PB/s**

As of today we have 498 PB of data stored on tapes

Challenges:

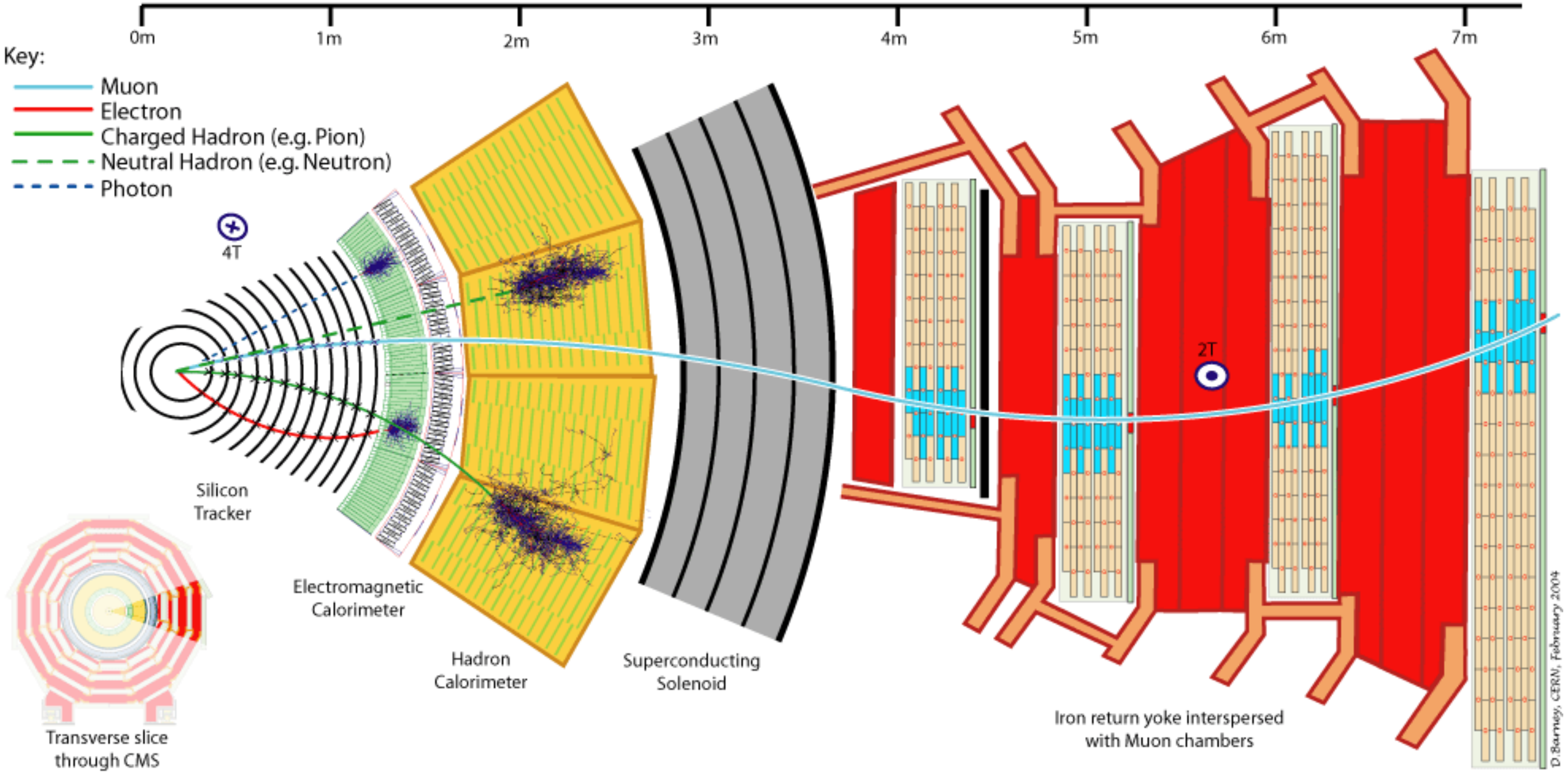
- **cope with large amounts of data**



From interaction to data

Large amounts of heterogeneous and complex data from multiple sub-detectors

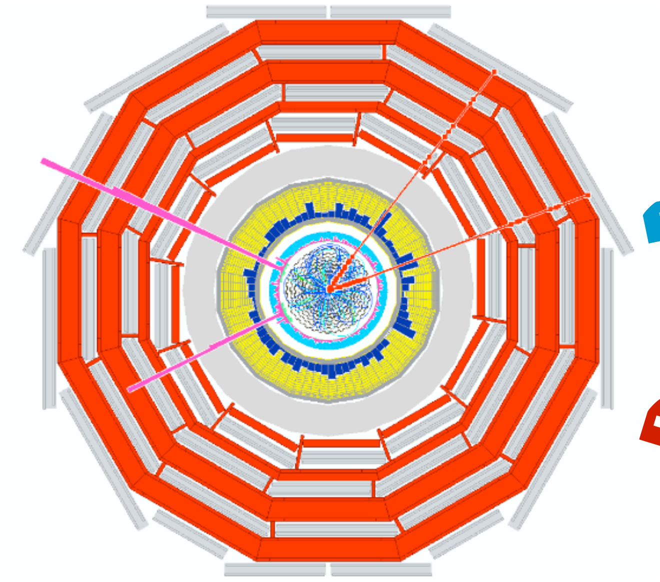
$\mathcal{O}(10^8)$ **sensors used to record particles from p-p collisions**



From the collisions to the physics result

Challenges:

- **Large amounts of high-dimensional data**
- **Large collisions production rate**
- **Limited frontend output bandwidth**
- **Limited storage space**



Collision Data

40 MHz

Trigger

~100 kHz

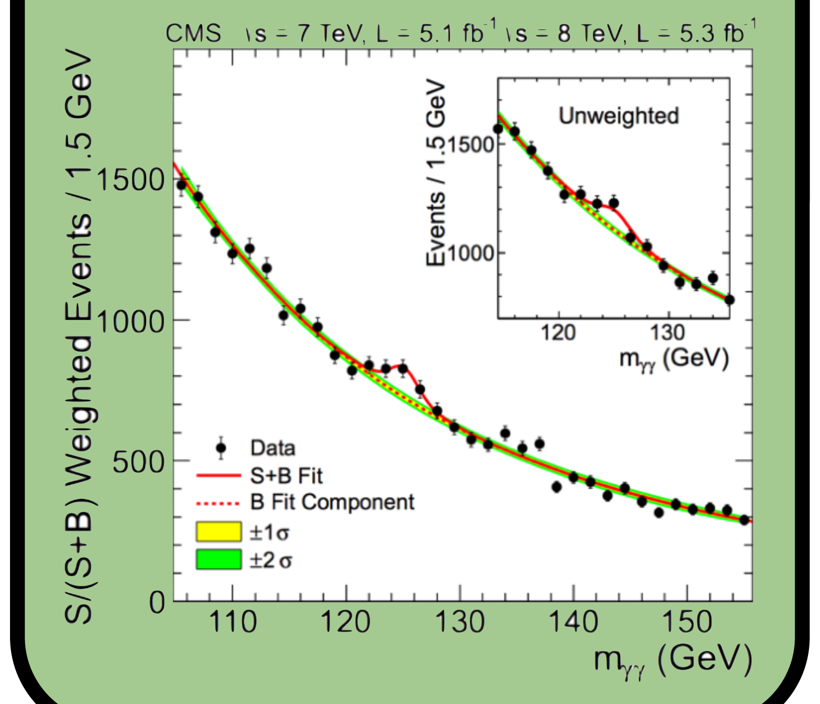
DataAcQuistion and Event Filter (HLT)

~1 kHz

Storage

Event Reconstruction

Offline analysis



Simulated Data

Simulation

Online data selection

Impossible to handle subdetector outputs at LHC rates

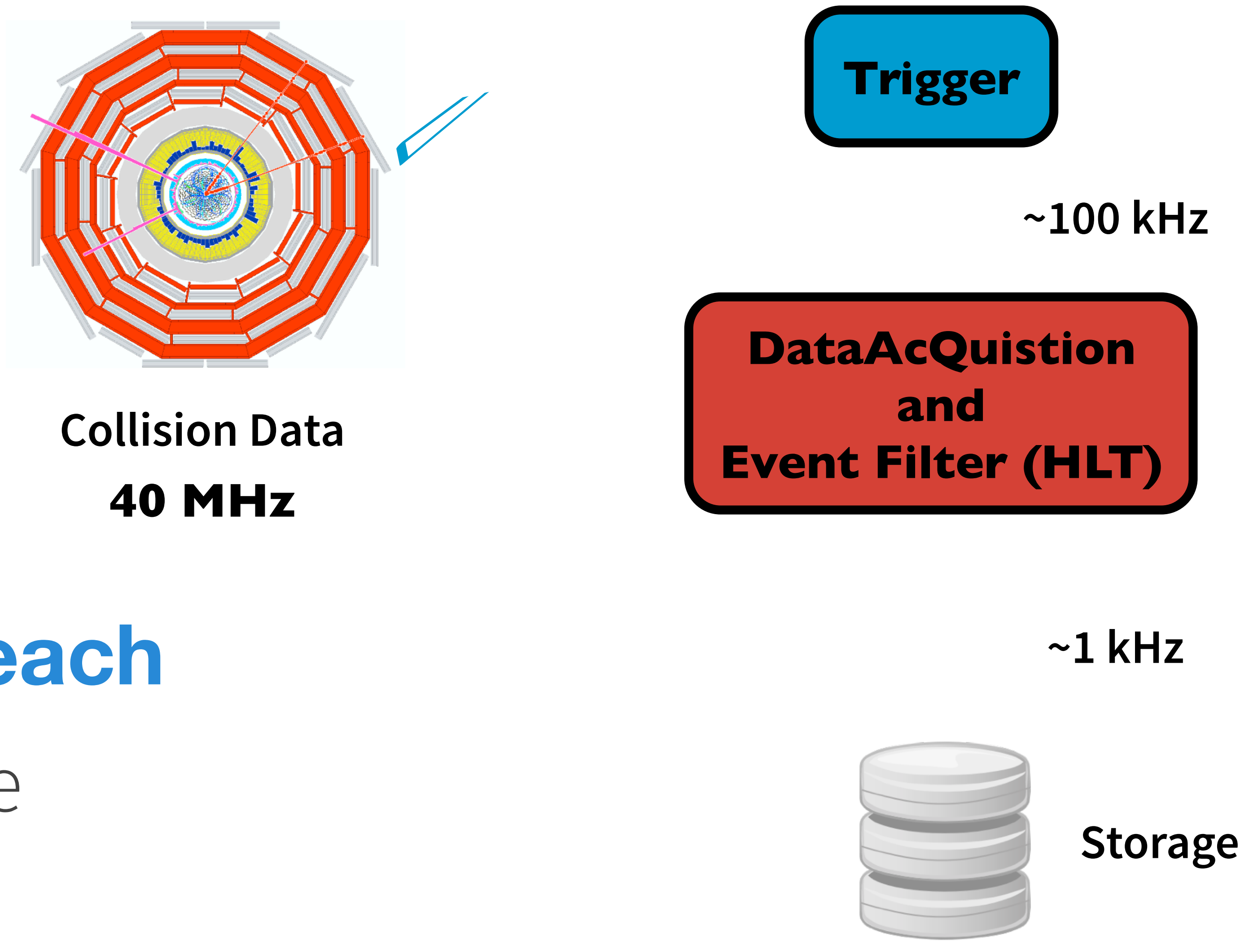
- Need **to decide what to discard and what to keep** for further analysis
- Online event selection performed by **L1 (hardware)** and **HLT (software)** triggers

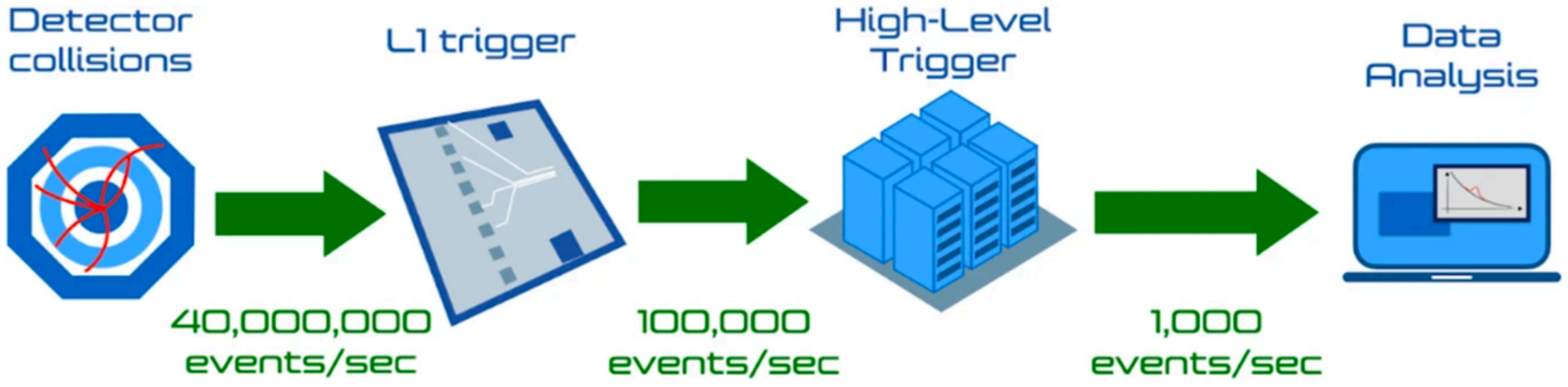
But trigger selection is currently a huge limitation to experiments sensitivity reach

- High threshold cuts (e.g. particle momentum) applied on trigger objects to limit data rate

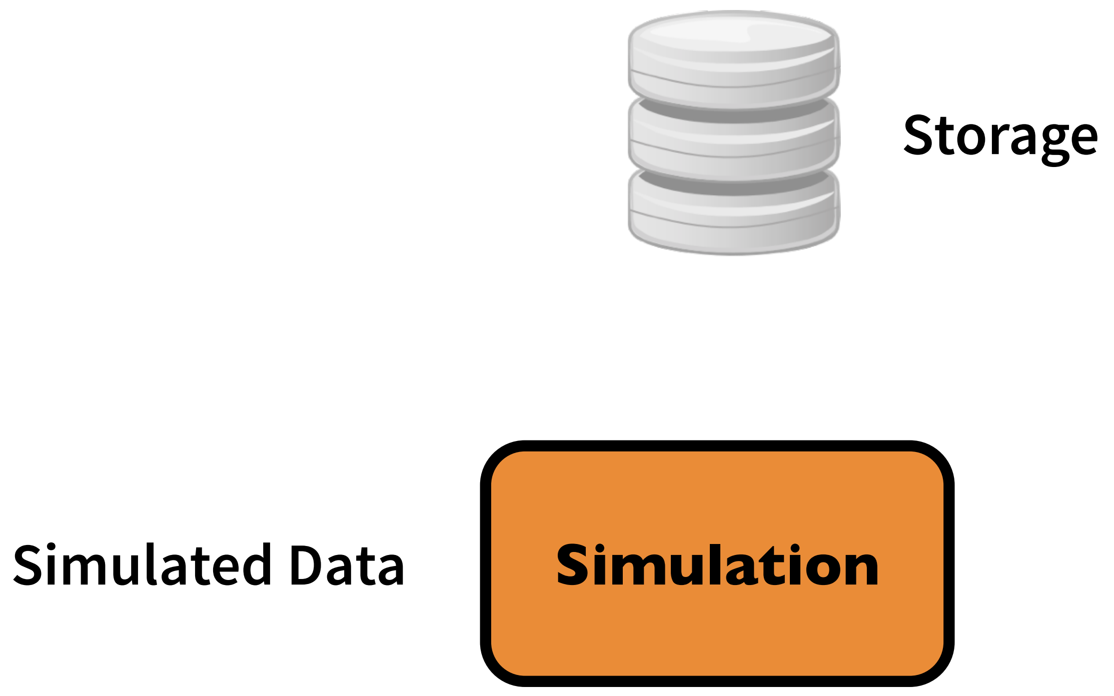
Great benefits from the full exploitation of data high-dimensionality

- ML embedding in on-detector FPGA systems





Data Simulation



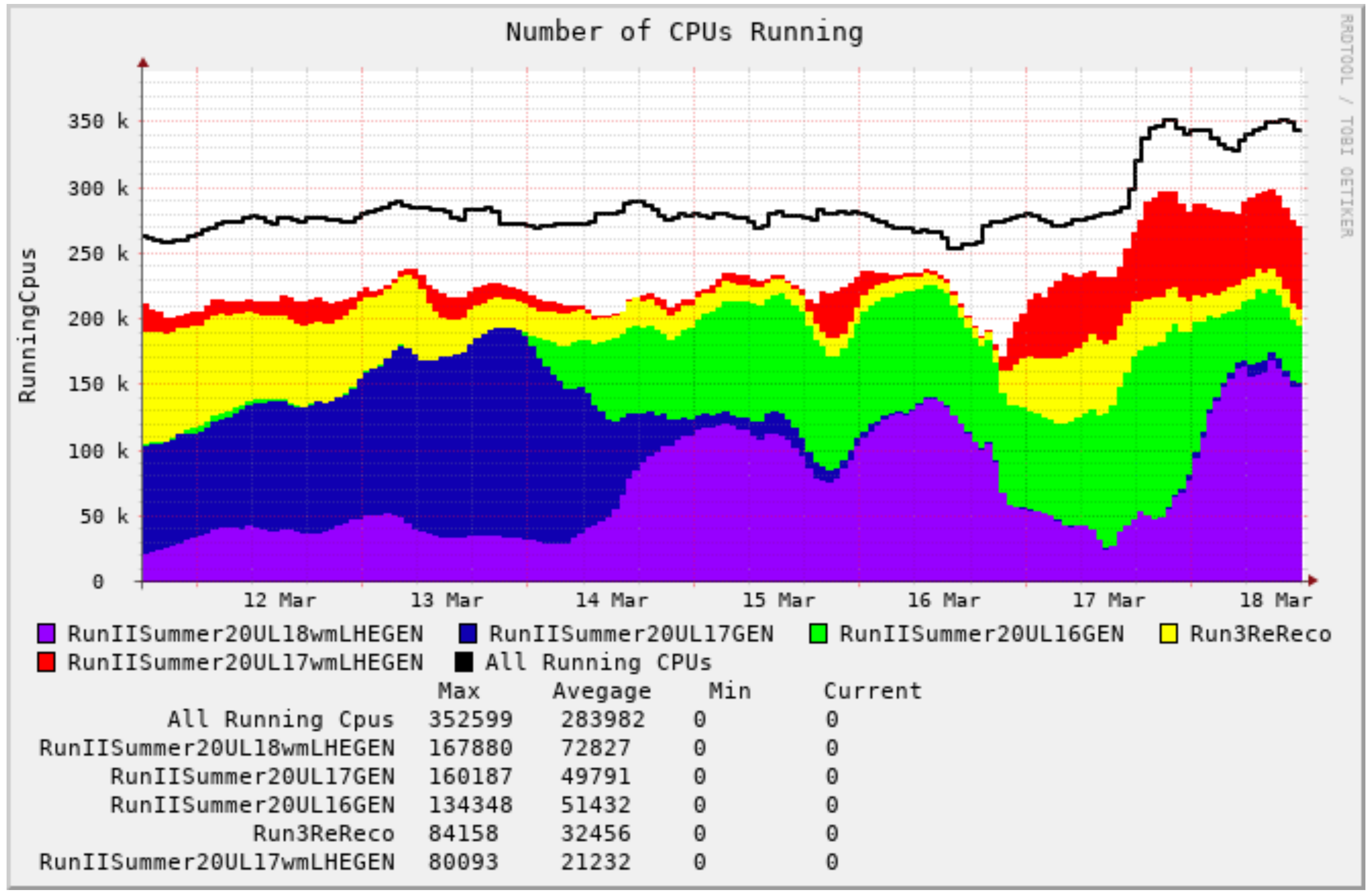
Simulated data is crucial for LHC experiments:

- to design optimal analyses for SM measurements and new physics searches
- to develop new detector technologies

Large data samples are needed to reduce systematical uncertainties

Production of MC simulated data is computationally intensive

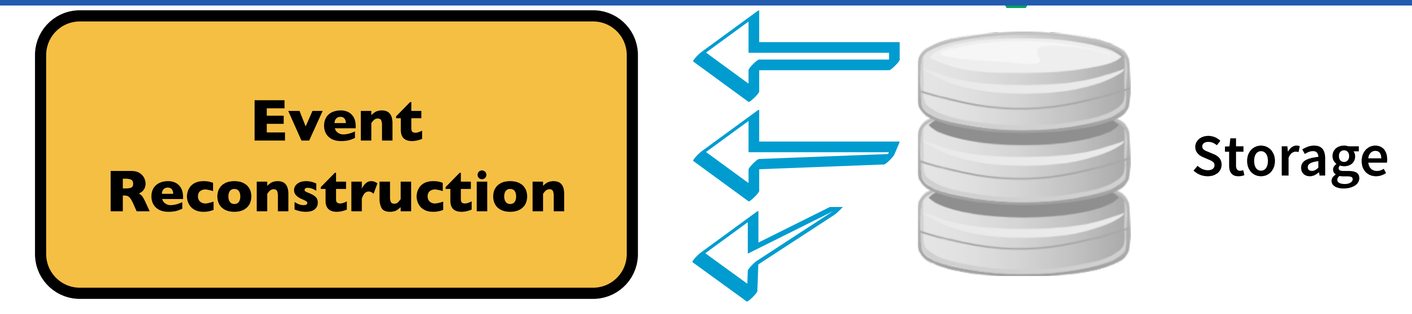
- Requires a lot of resources (CPU and data storage)



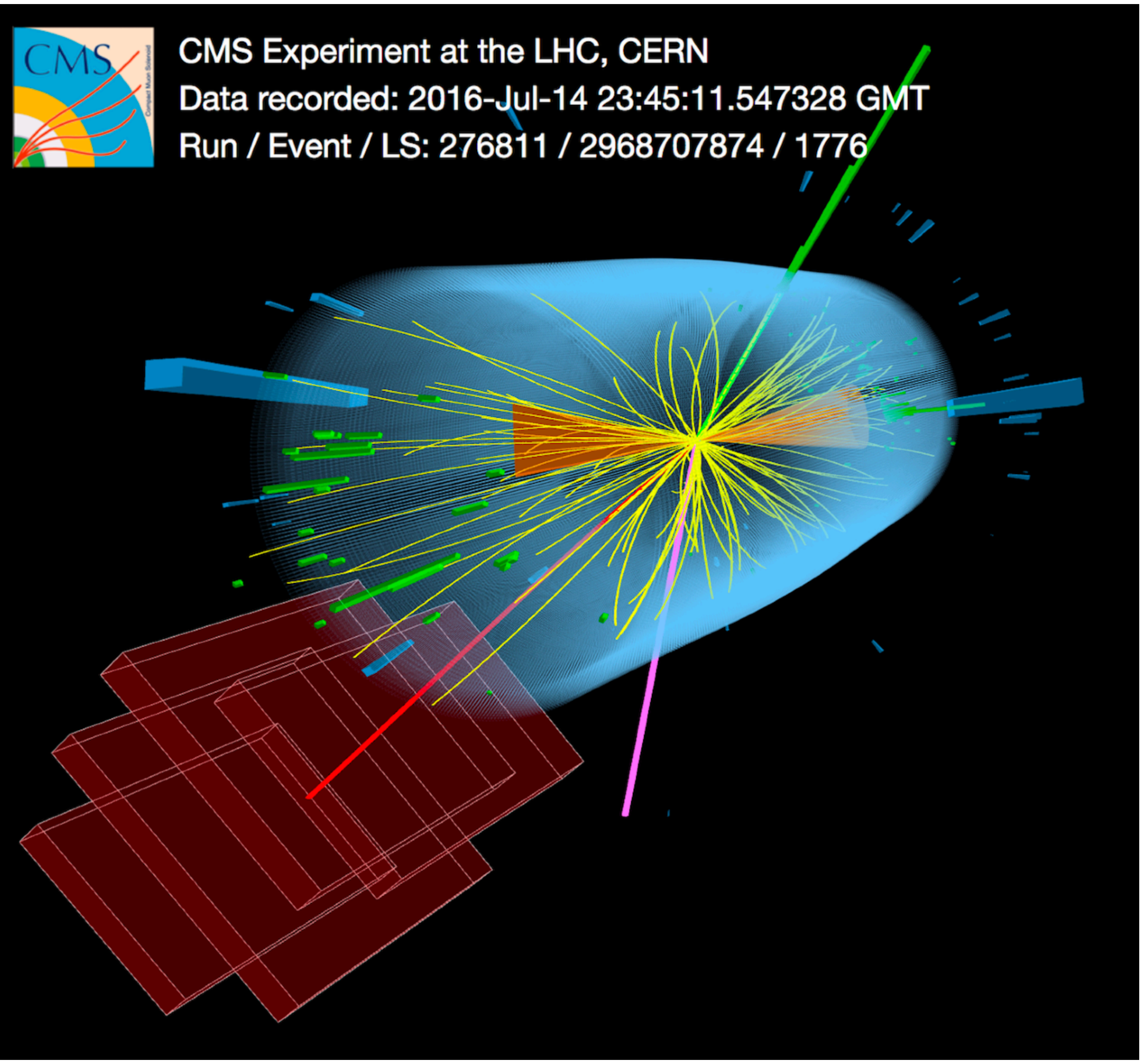
	Max	Avegage	Min	Current
All Running Cpus	352599	283982	0	0
RunIISummer20UL18wmLHEGEN	167880	72827	0	0
RunIISummer20UL17GEN	160187	49791	0	0
RunIISummer20UL16GEN	134348	51432	0	0
Run3ReReco	84158	32456	0	0
RunIISummer20UL17wmLHEGEN	80093	21232	0	0

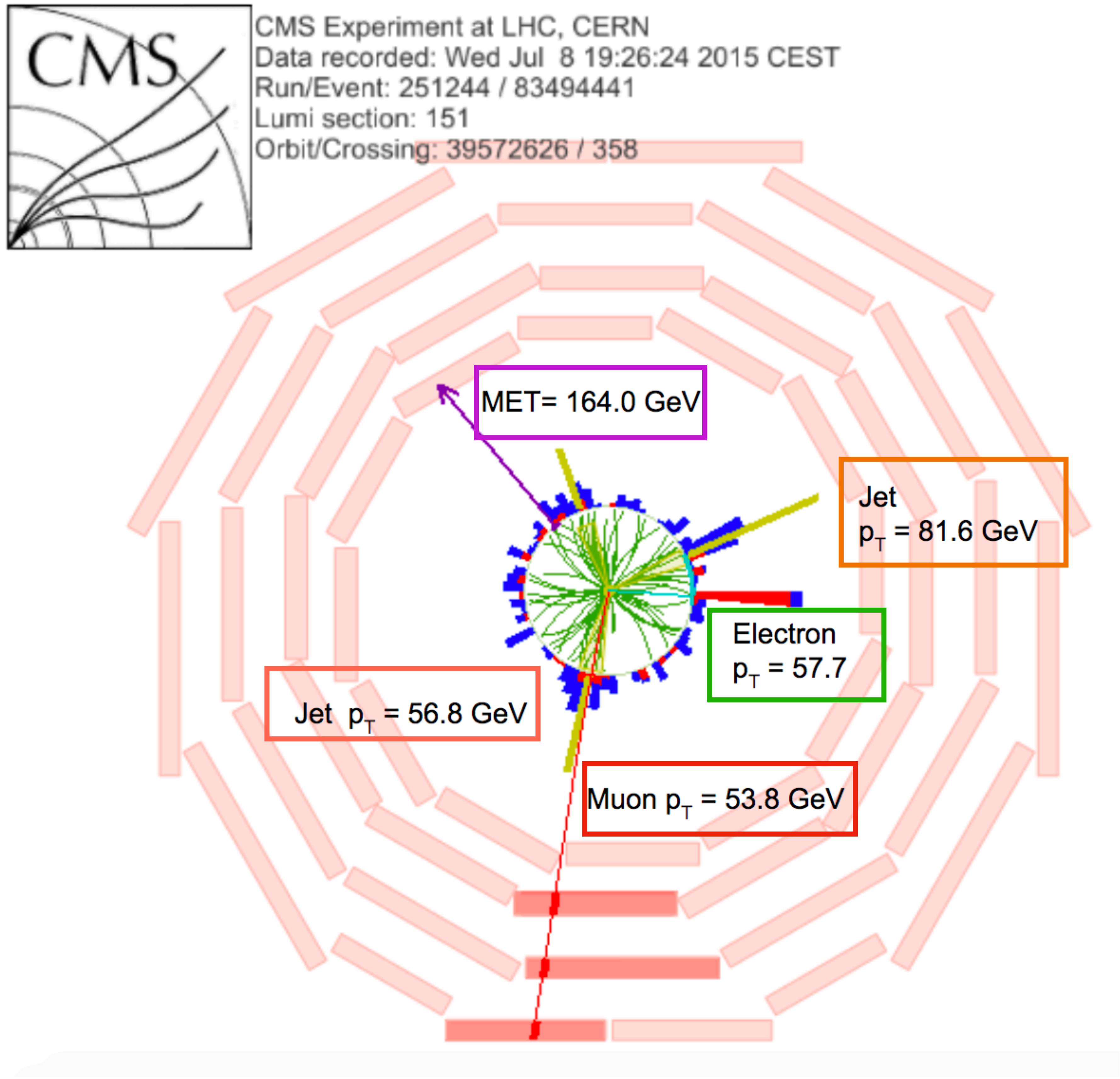
ML techniques (e.g. Generative Adversarial Networks) offer a promising approach to this issue

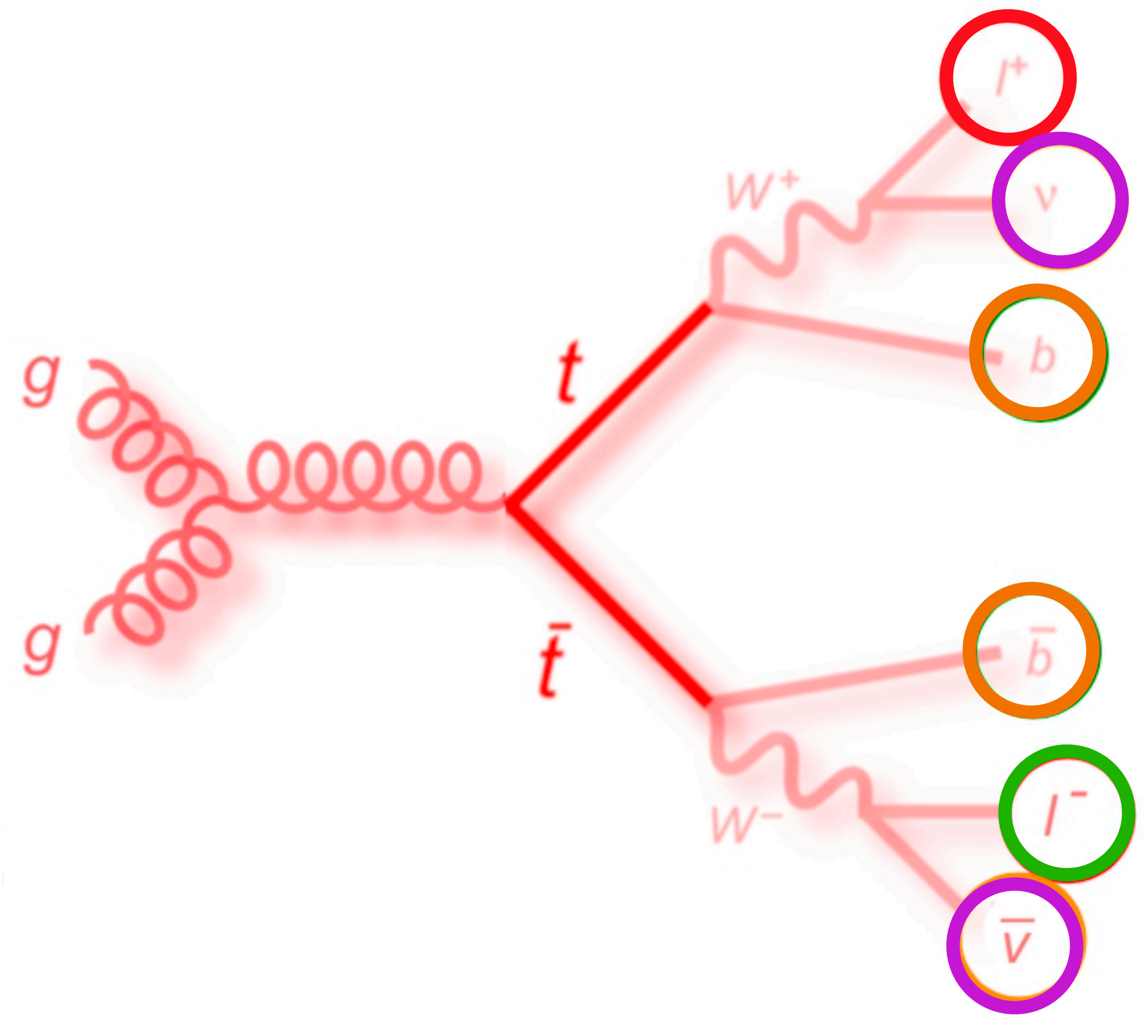
From data to physics



Raw data are processed to reconstruct tracks and energy clusters, which are combined to identify and measure physics objects (leptons, hadrons) produced in the collisions







Reconstruction

Event Reconstruction

Storage

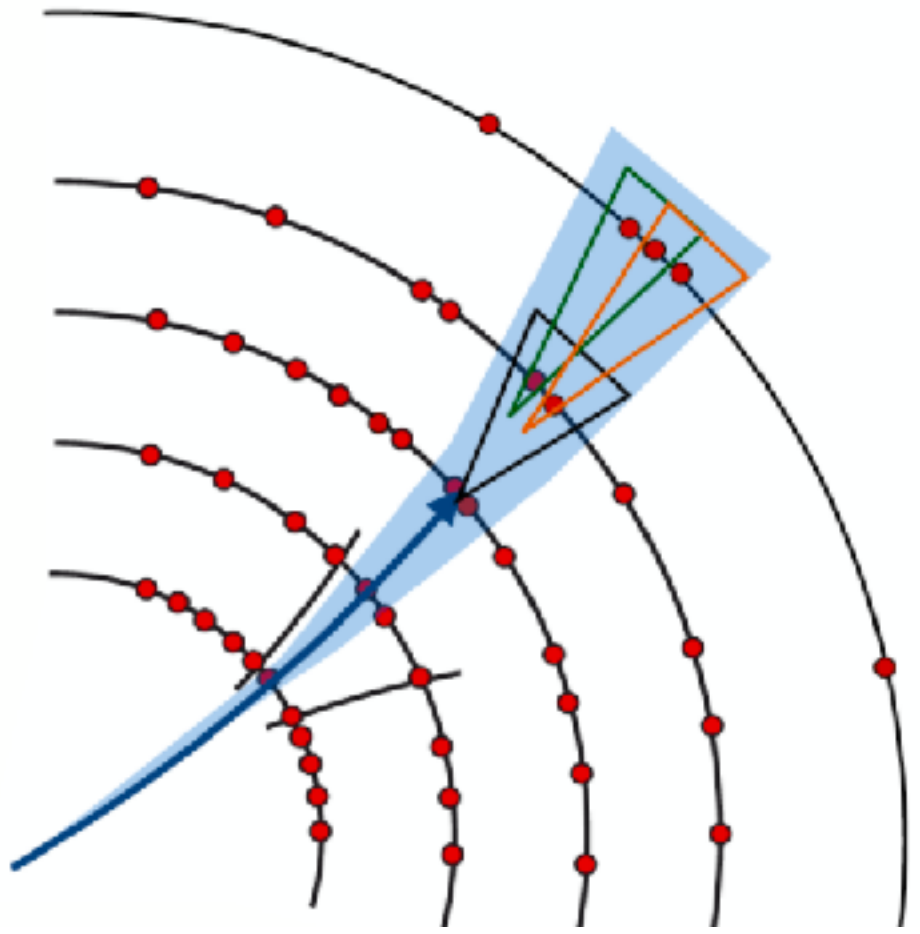
Low-level reconstruction algorithms (tracking and calorimeter clustering) require huge resources in terms of CPU

Example: Tracking algorithm (~10^8) channel:

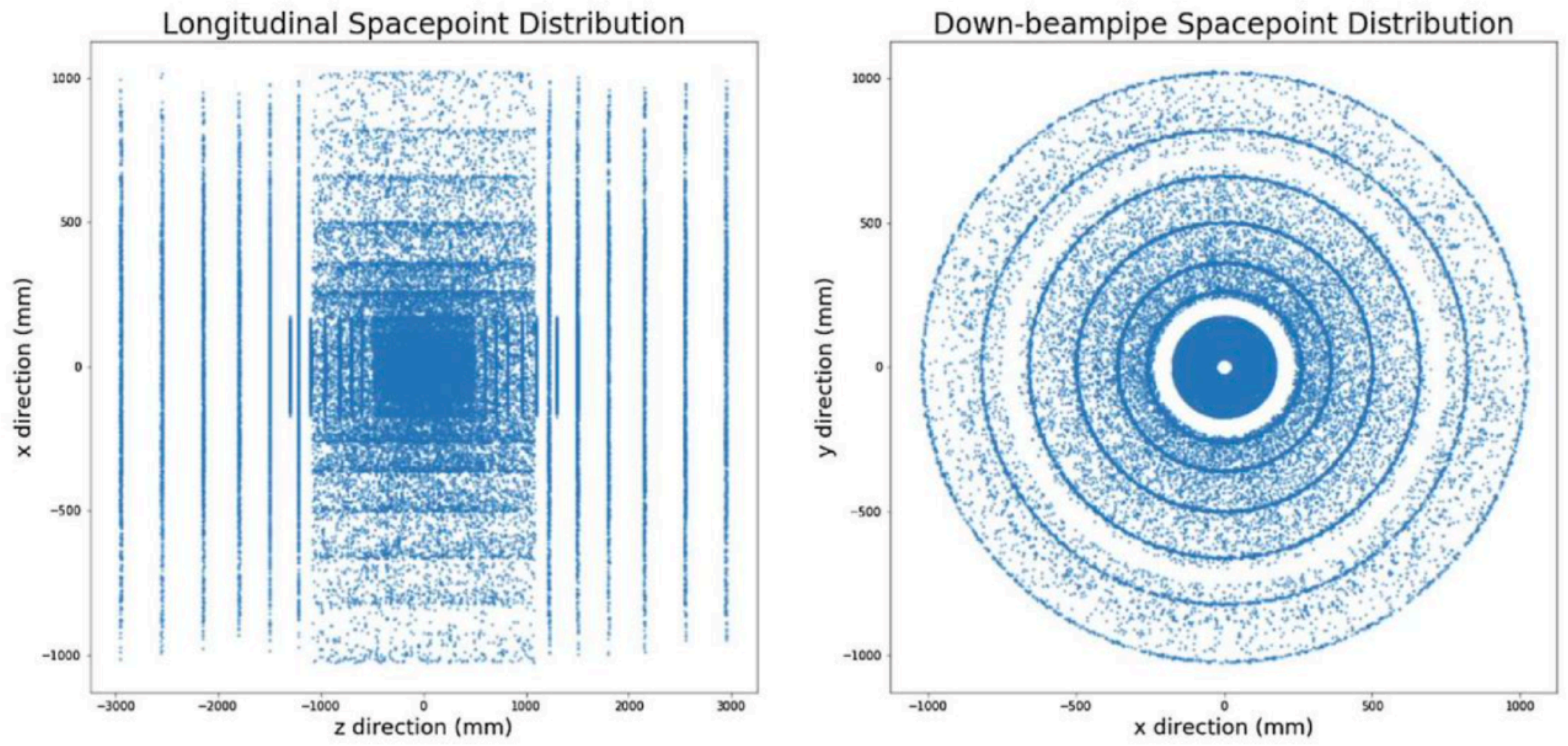
- Track seeds from hits
- Seeds are extended to full tracks
- Tracking software must be fast for reconstruction at HLT (100 KHz)

Graph neural network techniques are being explored for tracking and calorimeter clustering

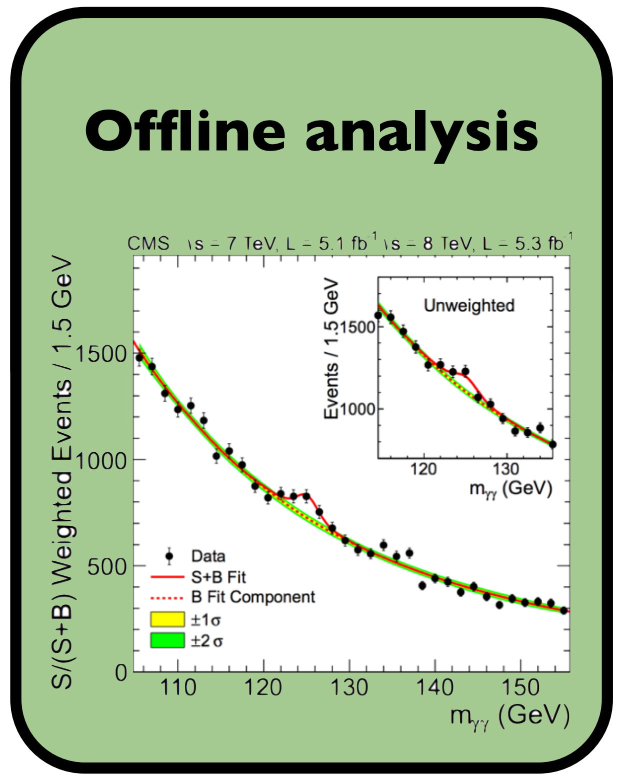
- **Inter-experiments efforts** on going **towards HL-LHC** (including tracking at L1 trigger)



Mean number of hits 110K

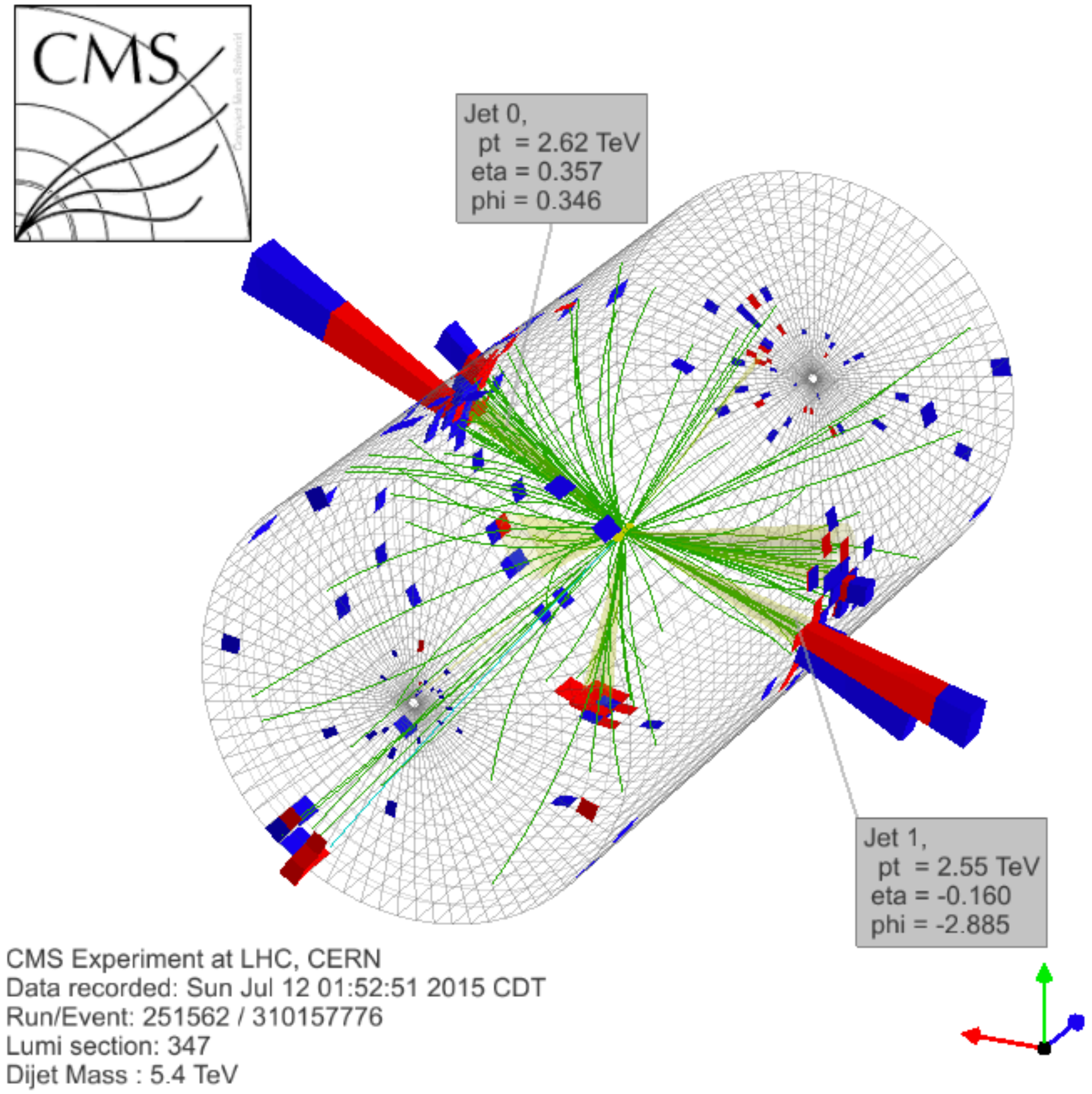


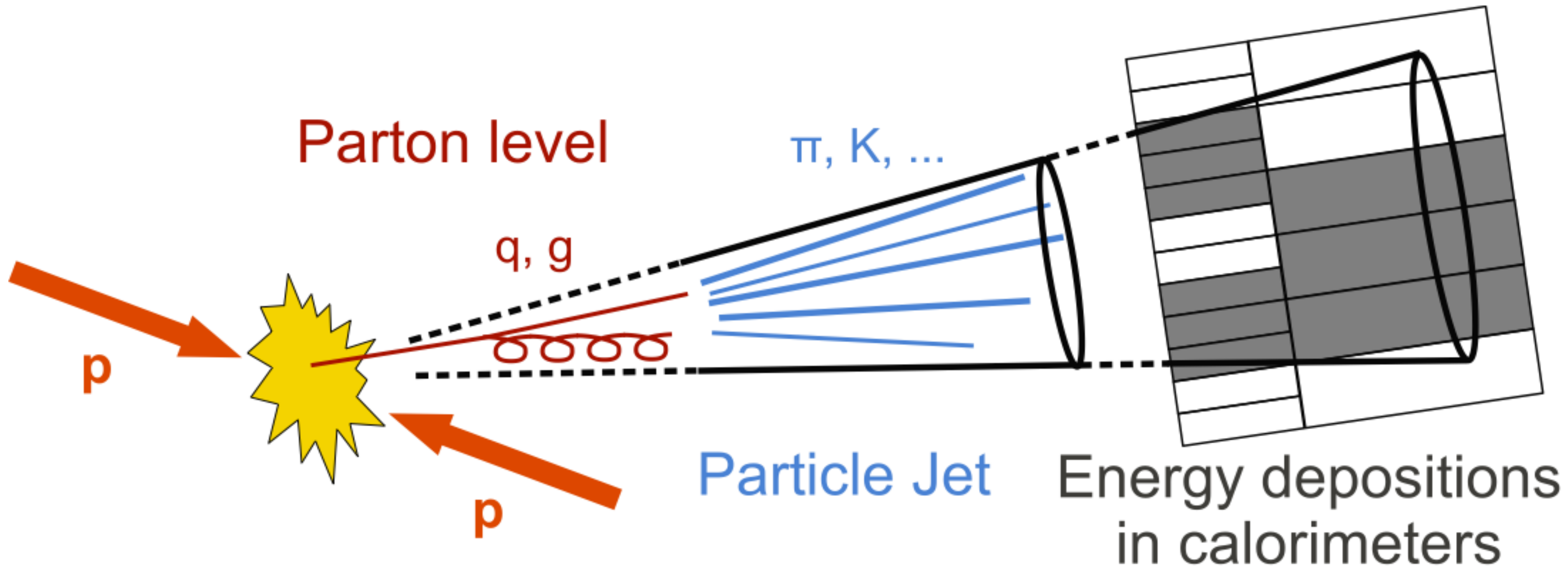
Jet classification



Quarks and gluons produced in p-p collisions shower and hadronize appearing as jets of particles

- Jets are complex objects to reconstruct and identify
- Jets features can give an important insight on their origin

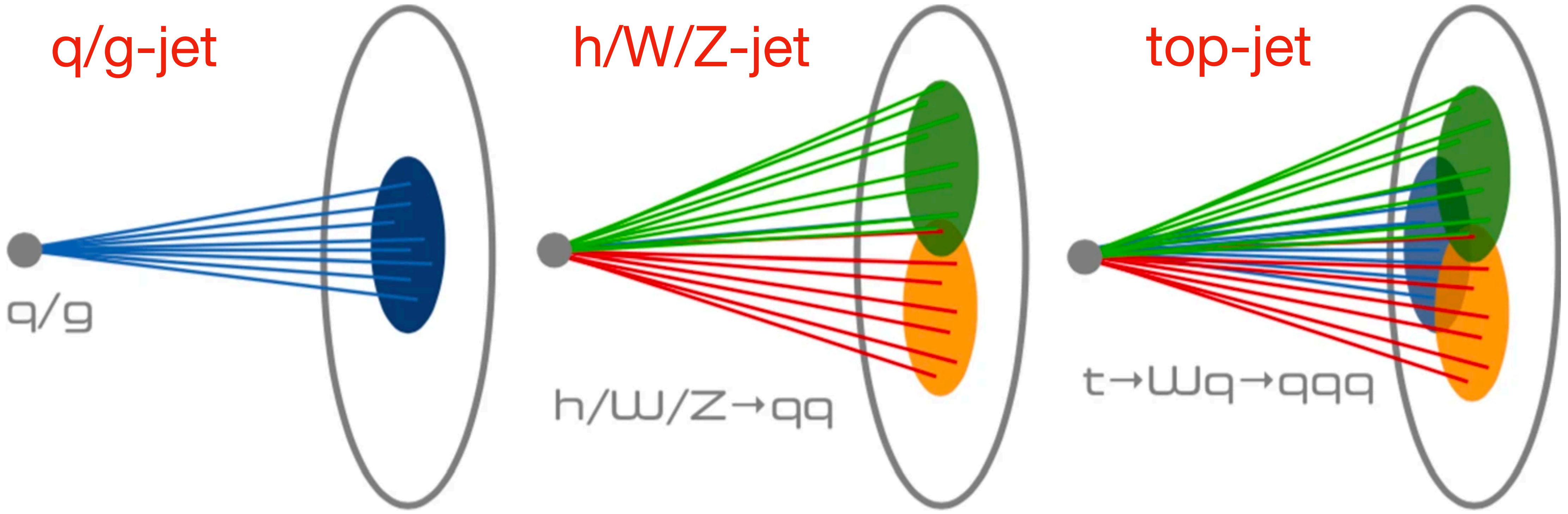


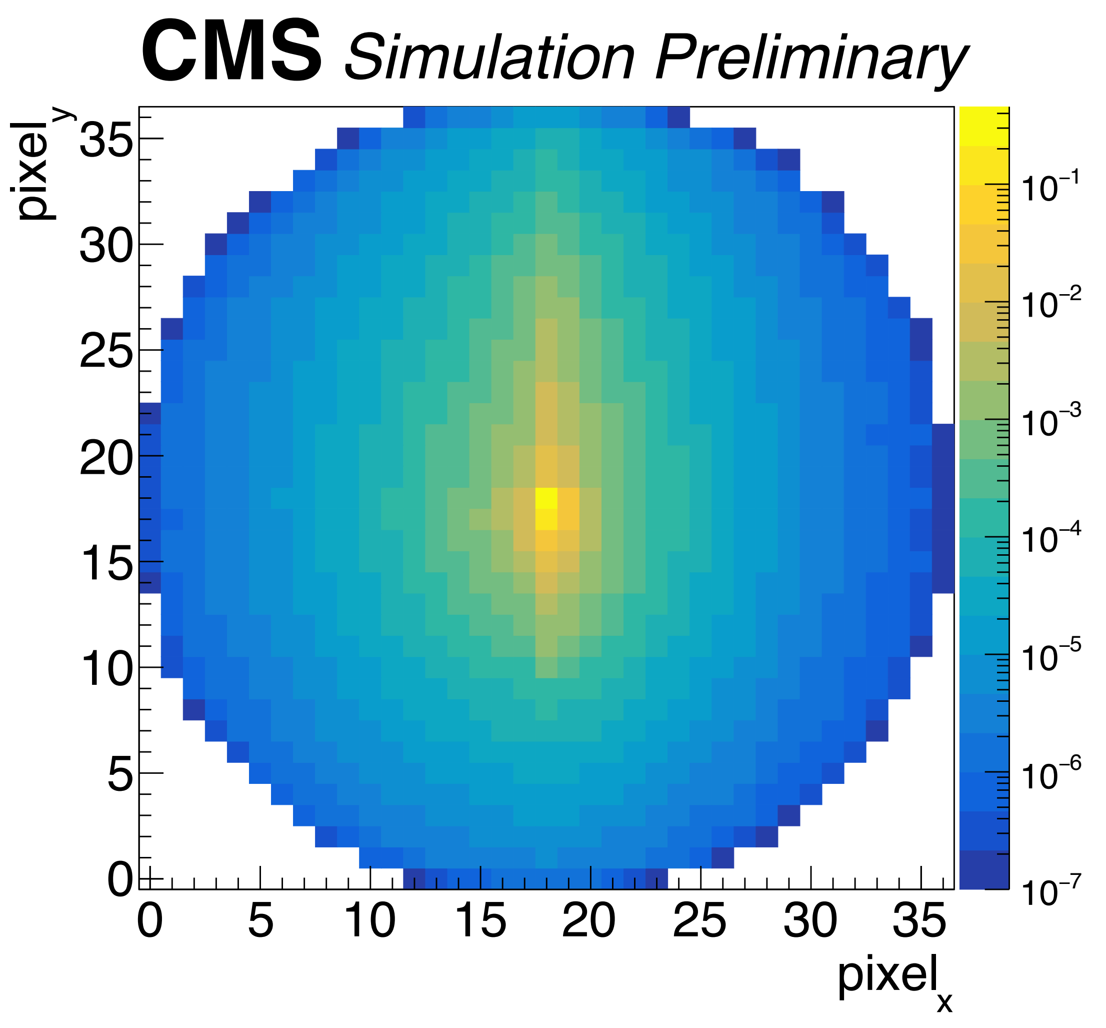


Jets classification:

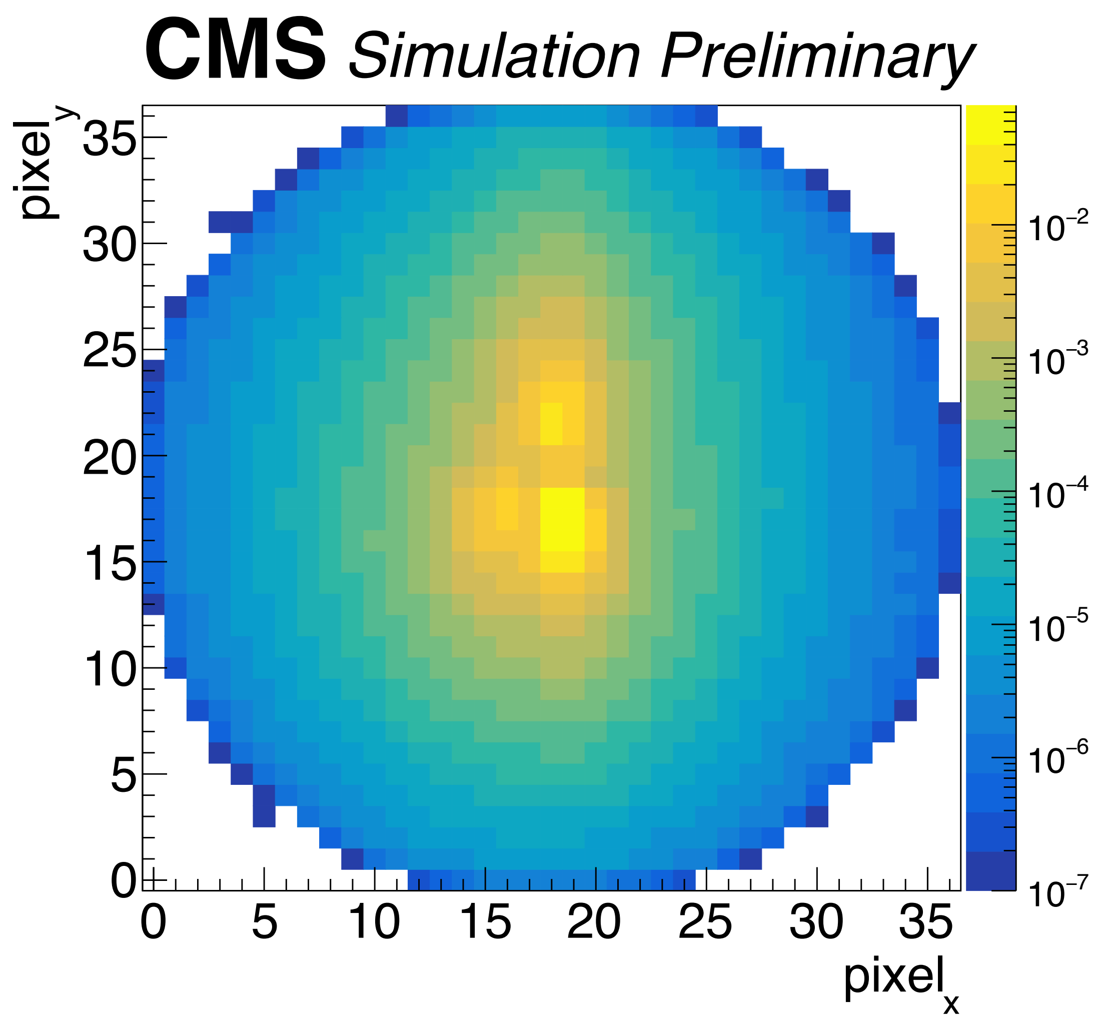
- Quark vs gluon jets
- Heavy vs light flavour jets
- Boosted objects (top, Higgs, W/Z)
- New physics jets

Jets substructure

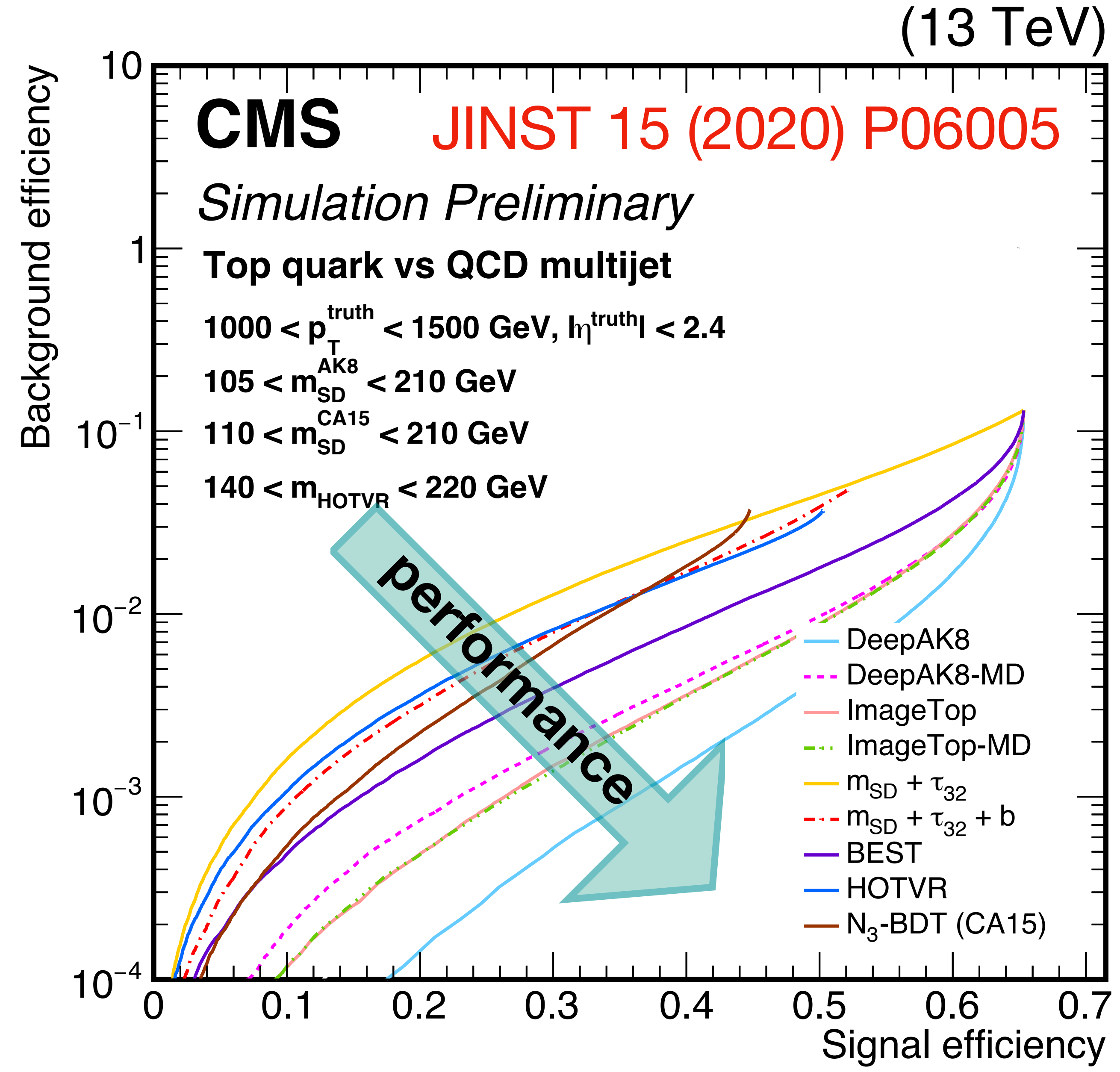




q/g-jet



top-jet



Highly energetic jets from heavy objects (t, H, W, Z) overlap

- Decay products appear as collimated
- **High level jet substructure variables** widely used for jet classification
- Modern approach: **make use of low-level features** (particle momenta)
  - feeding minimally processed data into a DDN

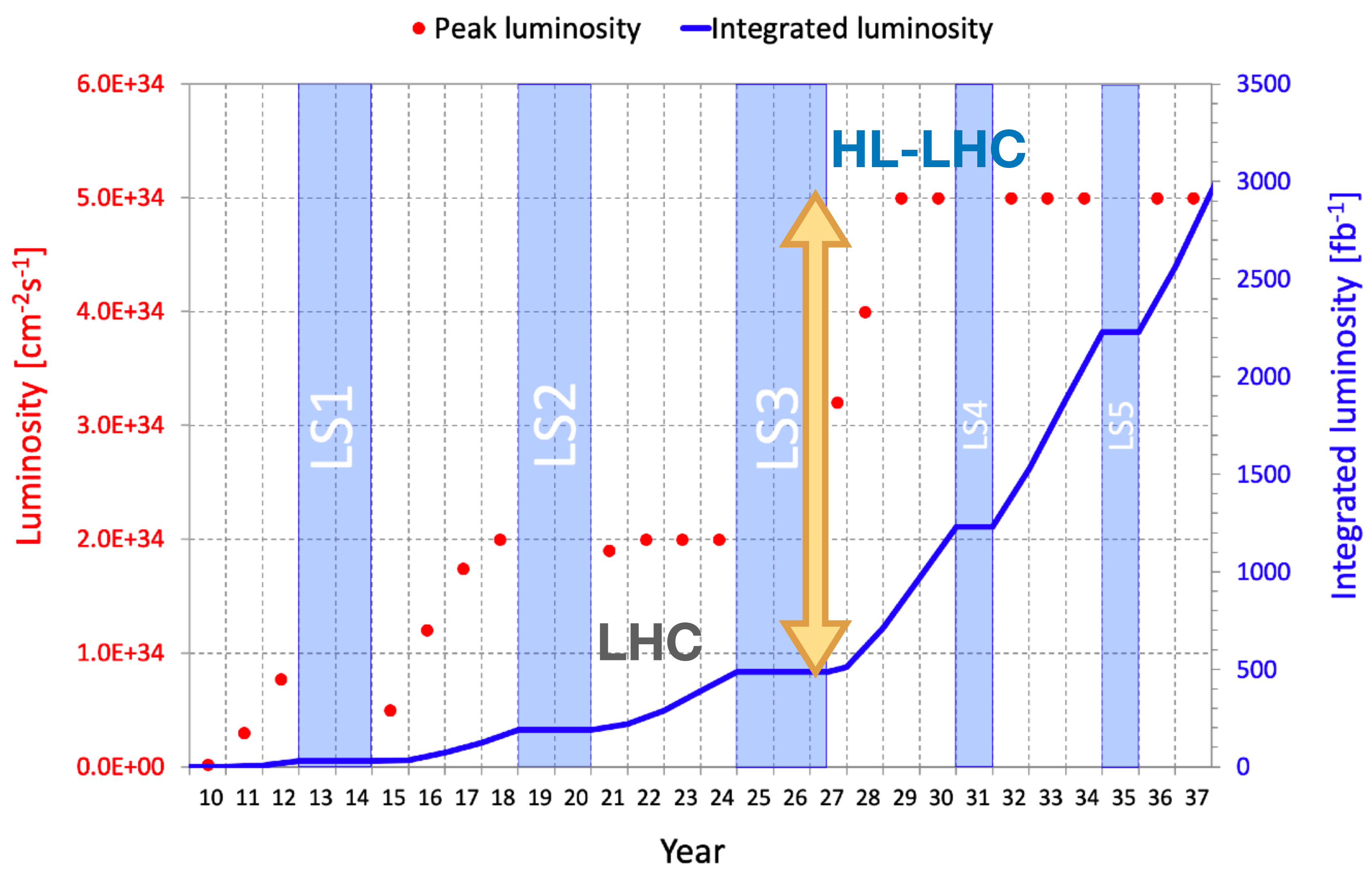
Similar considerations apply to q-, g- and nonSM-jets identification

HL-LHC: the high luminosity challenge

High Lumi LHC: Particle rate expected to increase by a factor 10

- Luminosity will go up to $5 \times 10^{32} cm^{-1} s^{-1}$
- Average pileup: $< \mu > = 140$
- $3000\ fb^{-1}$ of data (10 x LHC)

HL-LHC data will allow precision Higgs boson measurements and enlarge the explorations of new physics possibilities



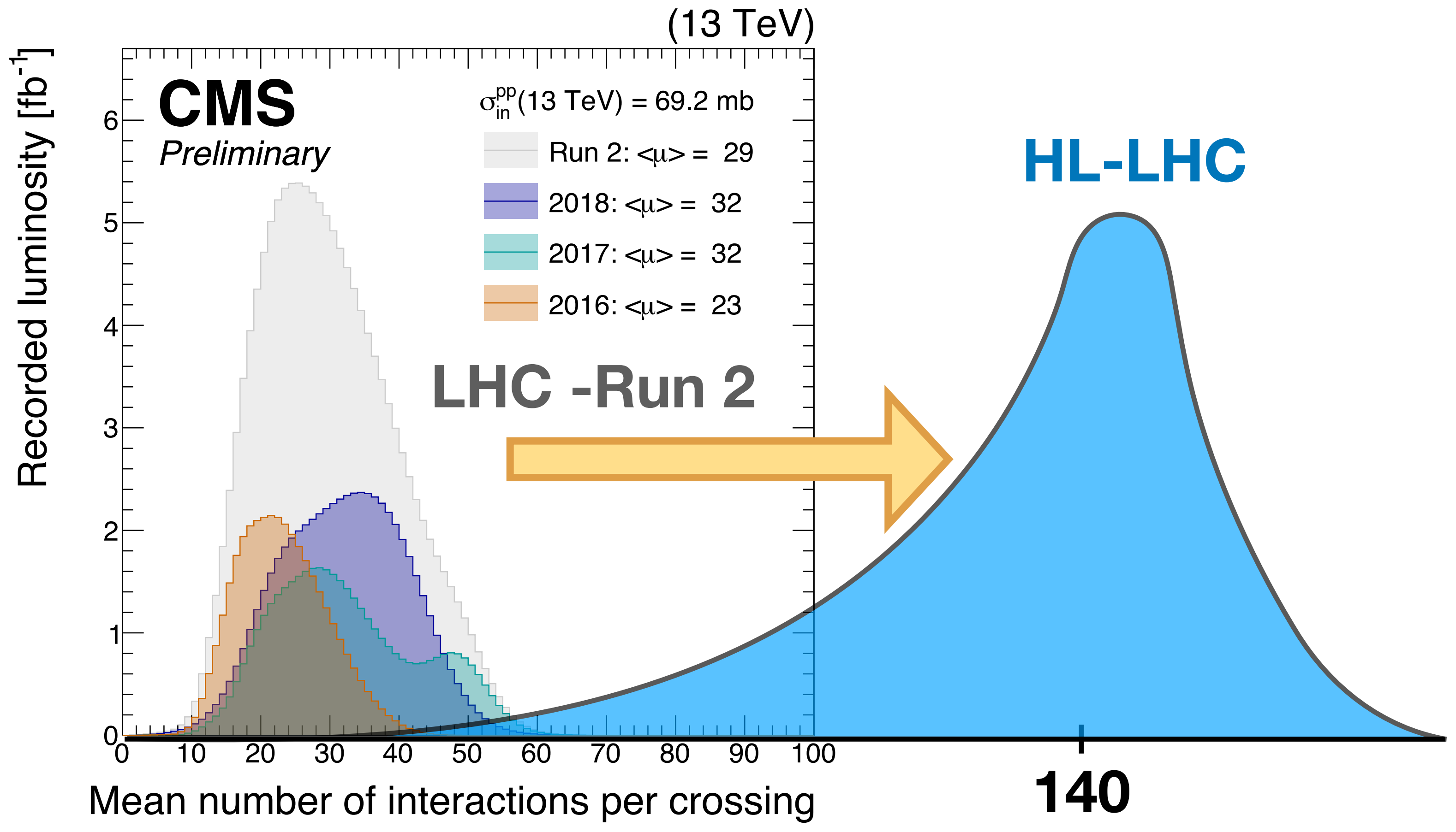
Increase in particle rate comes with

- increase in complexity of data-handling
- increase in storage and CPU requirements

Requiring:

- new tools for data processing

Overlapping collisions



The LHC Session

Today

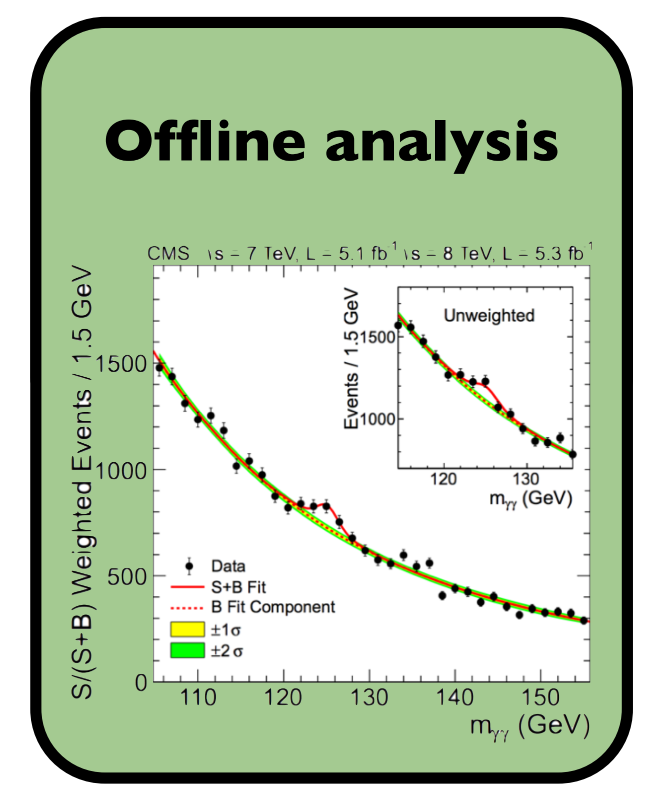
15:00	**Introduction on LHC experiments workflow with emphasis on challenges**	*Prof. Annapaola de Cosa*	*HIT E 51, ETH Zurich*	15:10 - 15:30
	Object identification and reconstruction	*Dr Alessandro Calandri*	*HIT E 51, ETH Zurich*	15:30 - 15:50
16:00	**Ultrafast ML inference for triggering**	*Dr Thea Aarrestad*	*HIT E 51, ETH Zurich*	15:50 - 16:10
	Machine learning for data quality monitoring	*Roberto Seidita*	*HIT E 51, ETH Zurich*	16:10 - 16:30
	Probabilistic models in ML for HEP	*Davide Valsecchi*	*HIT E 51, ETH Zurich*	16:30 - 16:50

Tomorrow

09:00	**Data-MC matching**	*Massimiliano Galli*	*HIT E 51, ETH Zurich*	09:00 - 09:20
	Analysis techniques	*Florian Eble*	*HIT E 51, ETH Zurich*	09:20 - 10:00
10:00	**Quantum machine learning**	*Vasilis Belis*	*HIT E 51, ETH Zurich*	10:00 - 10:20

Additional material

Heavy Flavour jet tagging



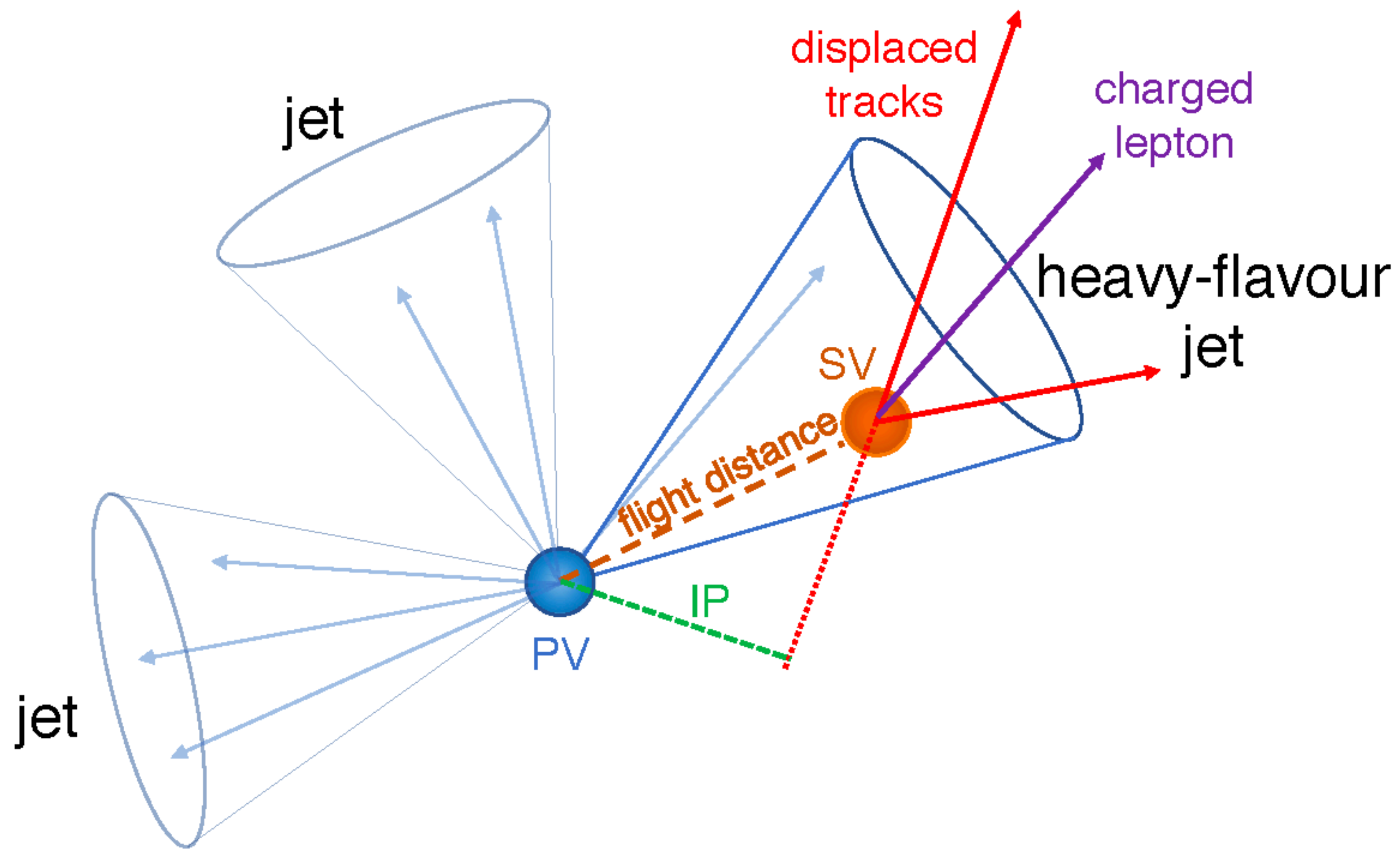
Jets originated from b-quark fragmentation appears in final states of several interesting processes:

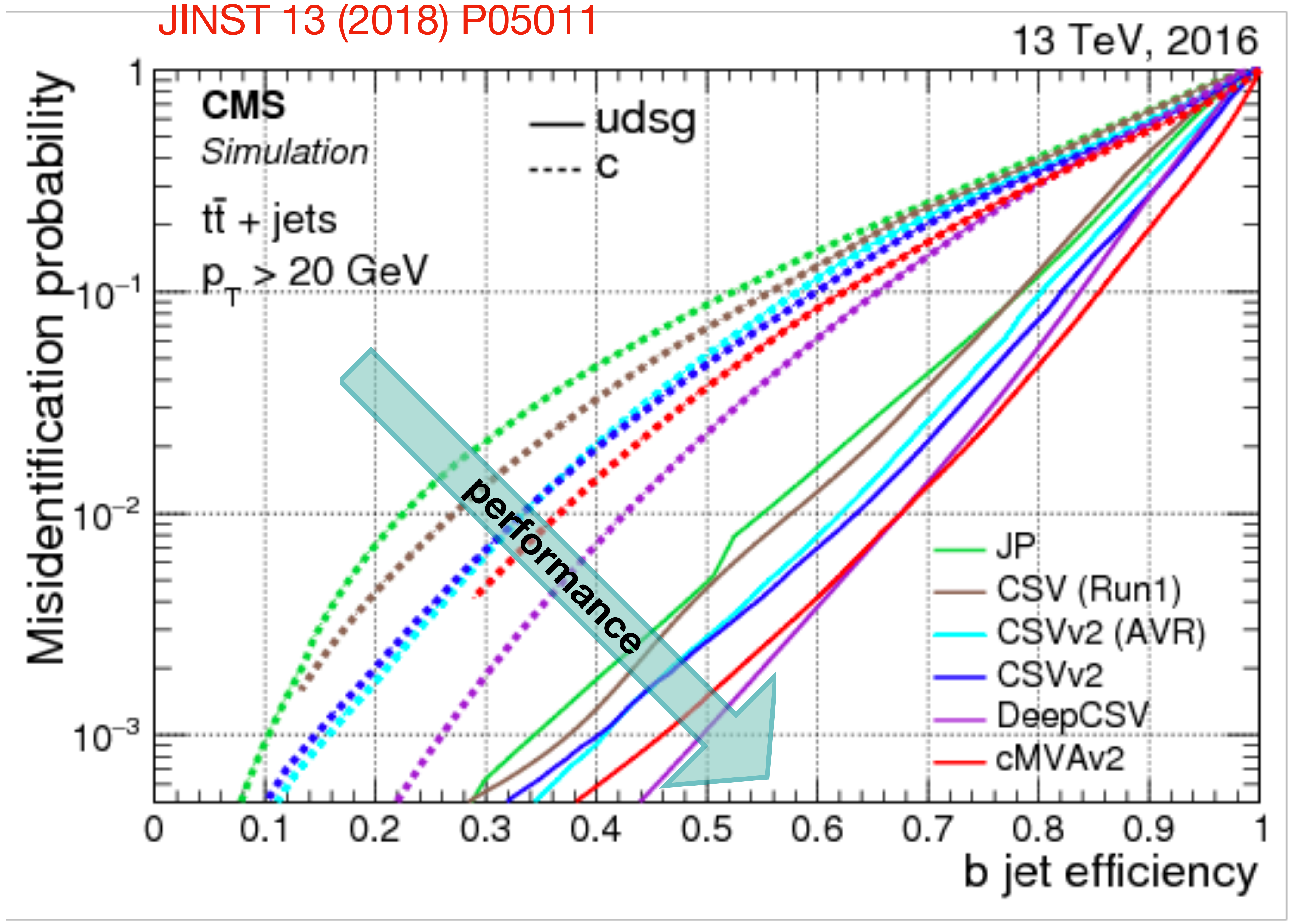
- e.g.: $t\bar{t} \to WbWb, VH \to Vbb$

b-jets tagging exploits B mesons decay features

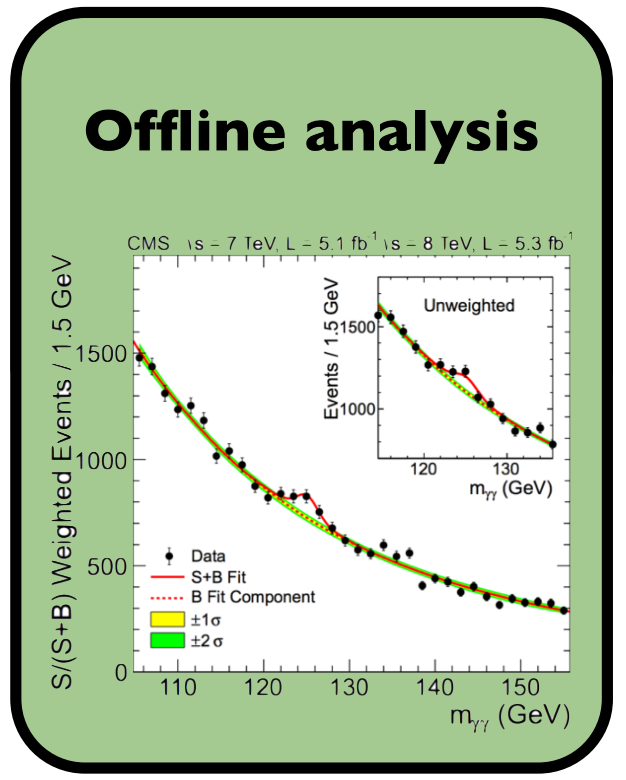
- B mesons long lifetime leads to displaced decay
- Presence of leptons from semileptonic decays

Modern b-tagging methods use deep learning techniques



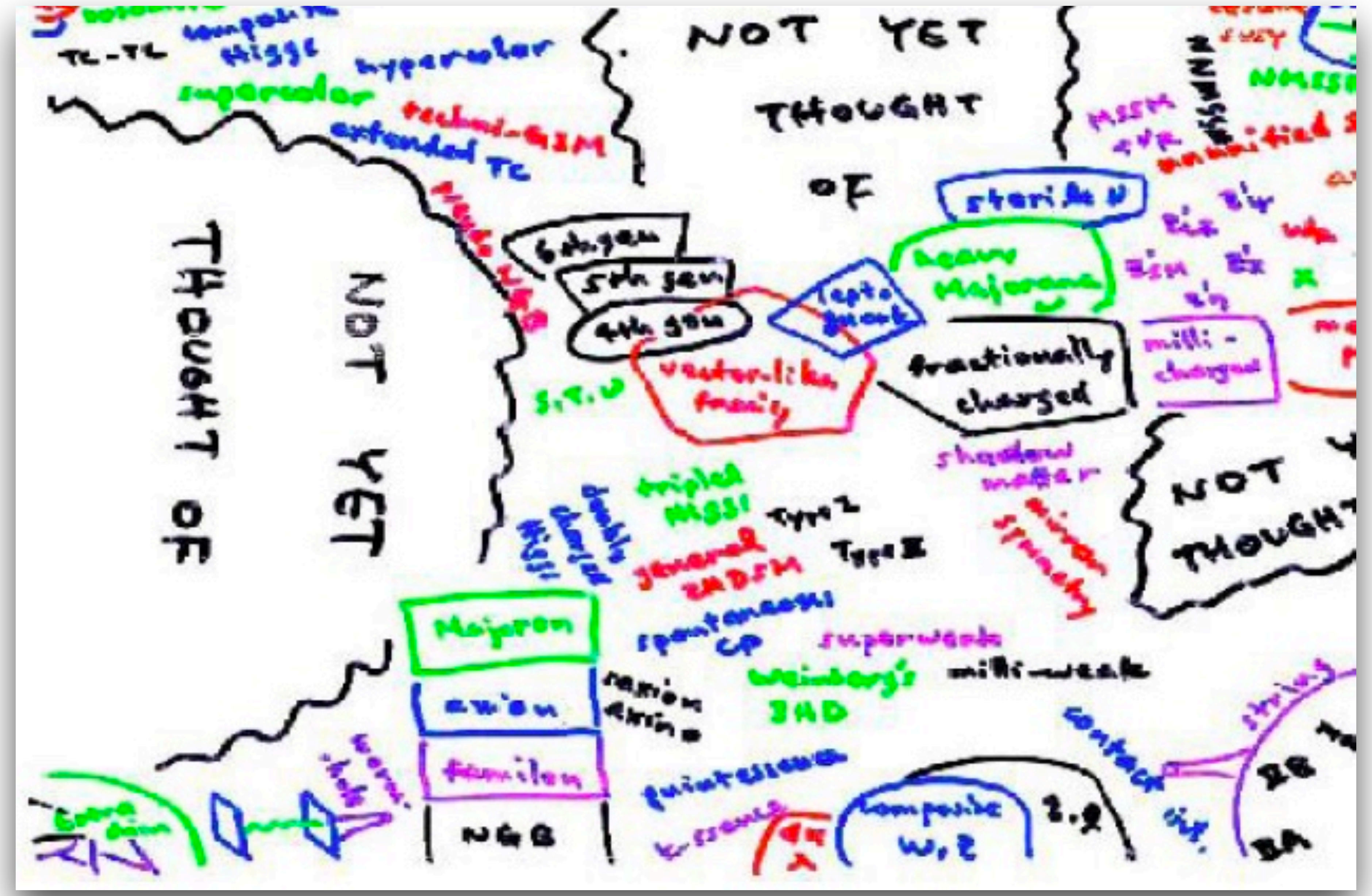


Anomaly detection



Lack of a priori knowledge on what new physics might look like

- Vast landscape of new physics models being probed at the LHC experiments up to the TeV scale
- Still lots of possibilities remain unexplored



Model agnostic strategies may open the doors to new physics

- Anomaly detection techniques based on DNN (e.g Autoencoders) are gaining traction