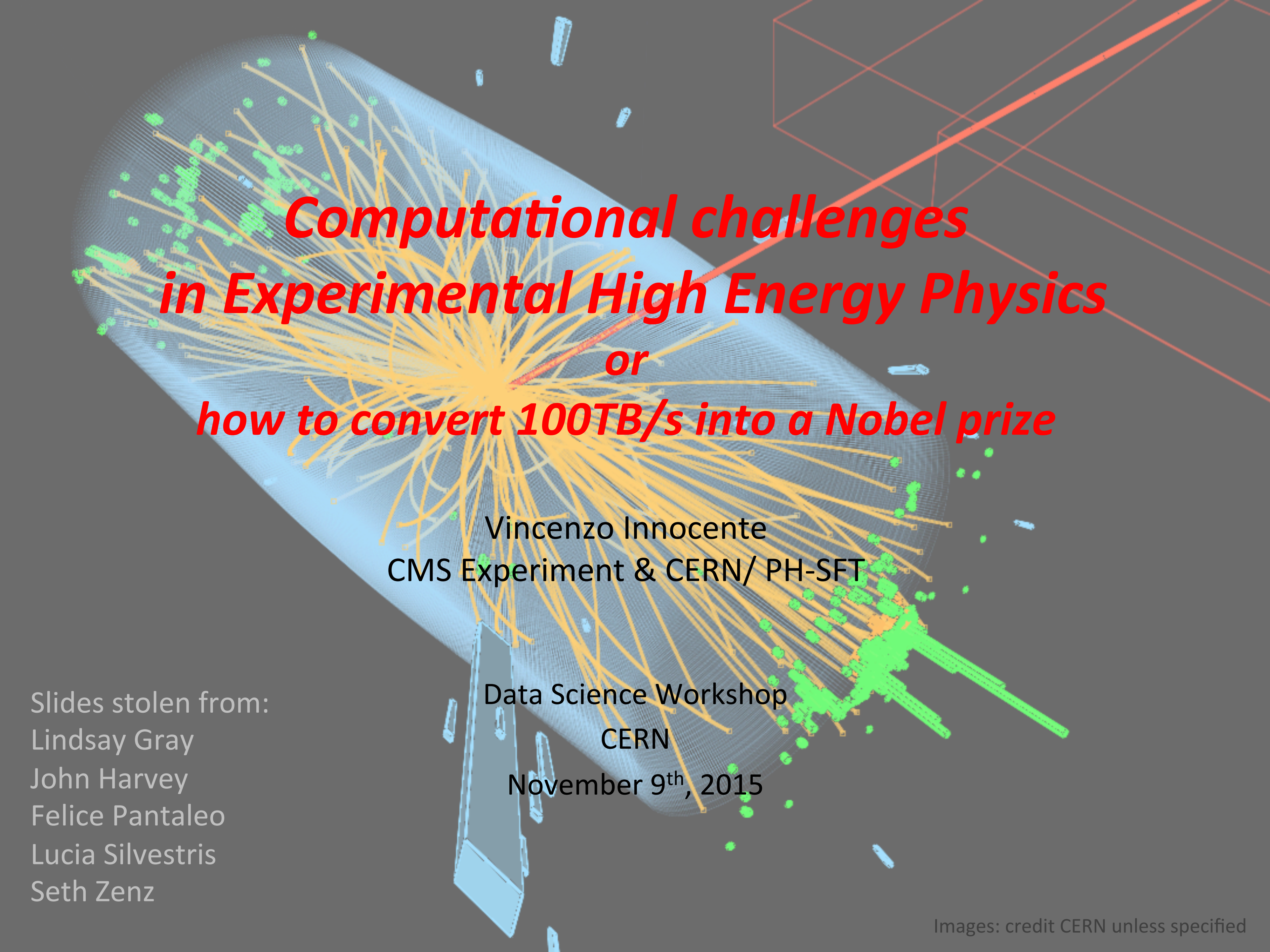


Computational challenges in Experimental High Energy Physics

or

how to convert 100TB/s into a Nobel prize

Vincenzo Innocente
CMS Experiment & CERN/ PH-SFT

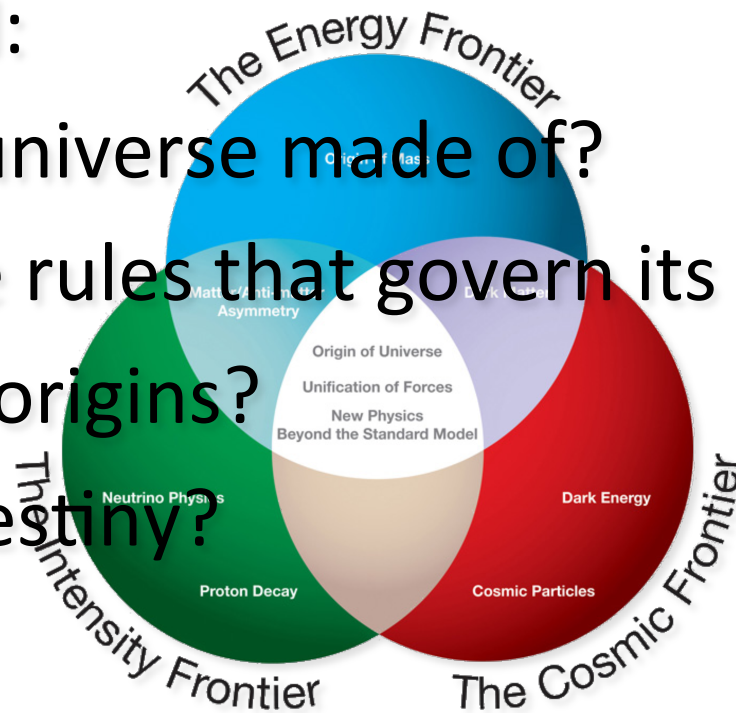
Data Science Workshop
CERN
November 9th, 2015

Slides stolen from:
Lindsay Gray
John Harvey
Felice Pantaleo
Lucia Silvestris
Seth Zenz

Particle Physics Frontiers

Only by exploring these 3 frontiers we can find the answers to the most fundamental questions of the mankind:

- What is the universe made of?
- What are the rules that govern its evolution?
- What are its origins?
- What is its destiny?



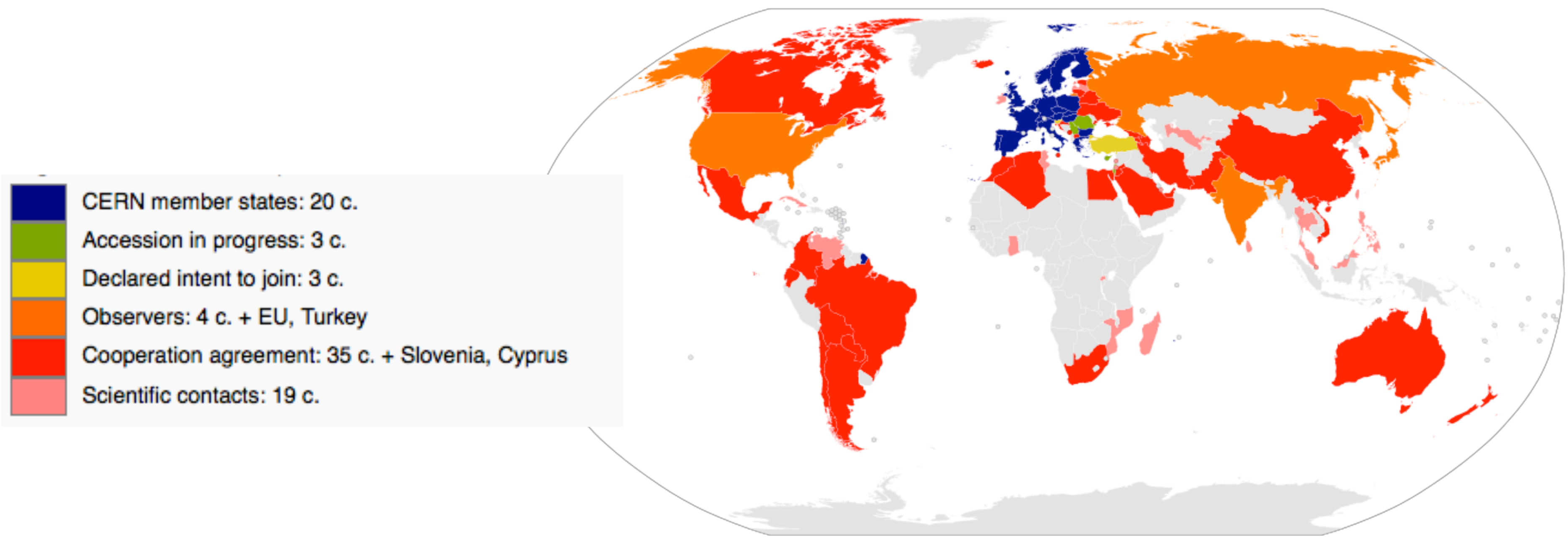
CERN

European Organization for Nuclear Research

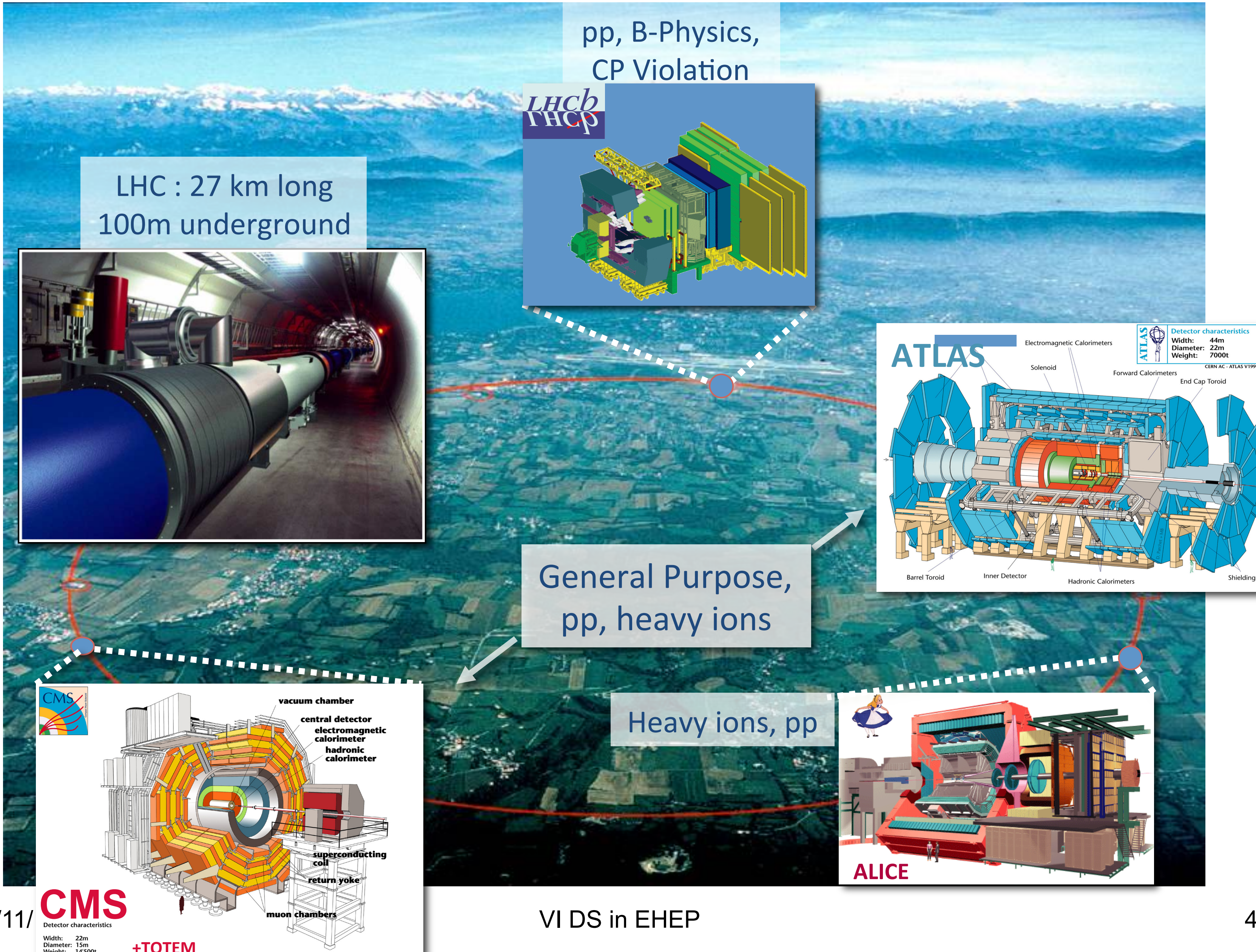
Founded in 1954 by 12 countries.

2012: 20 member states

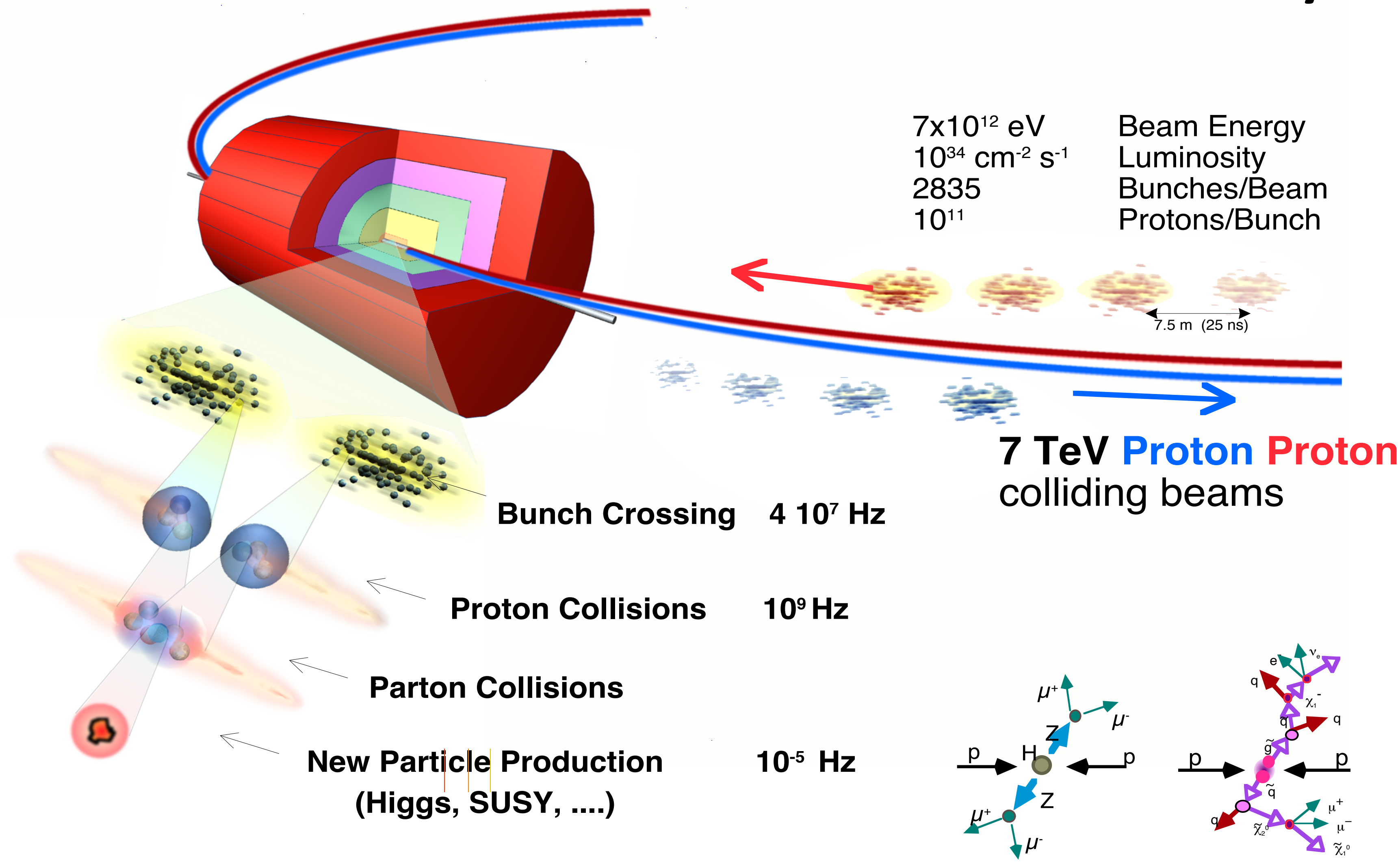
More than 10,000 users all around the world



The Large Hardon Collider at CERN



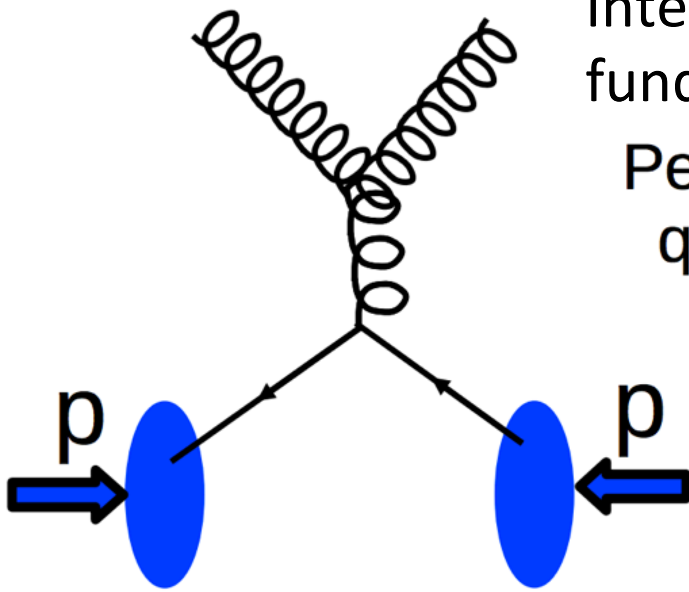
Collisions at the LHC: summary



Selection of 1 event in 10,000,000,000,000

When protons collide...

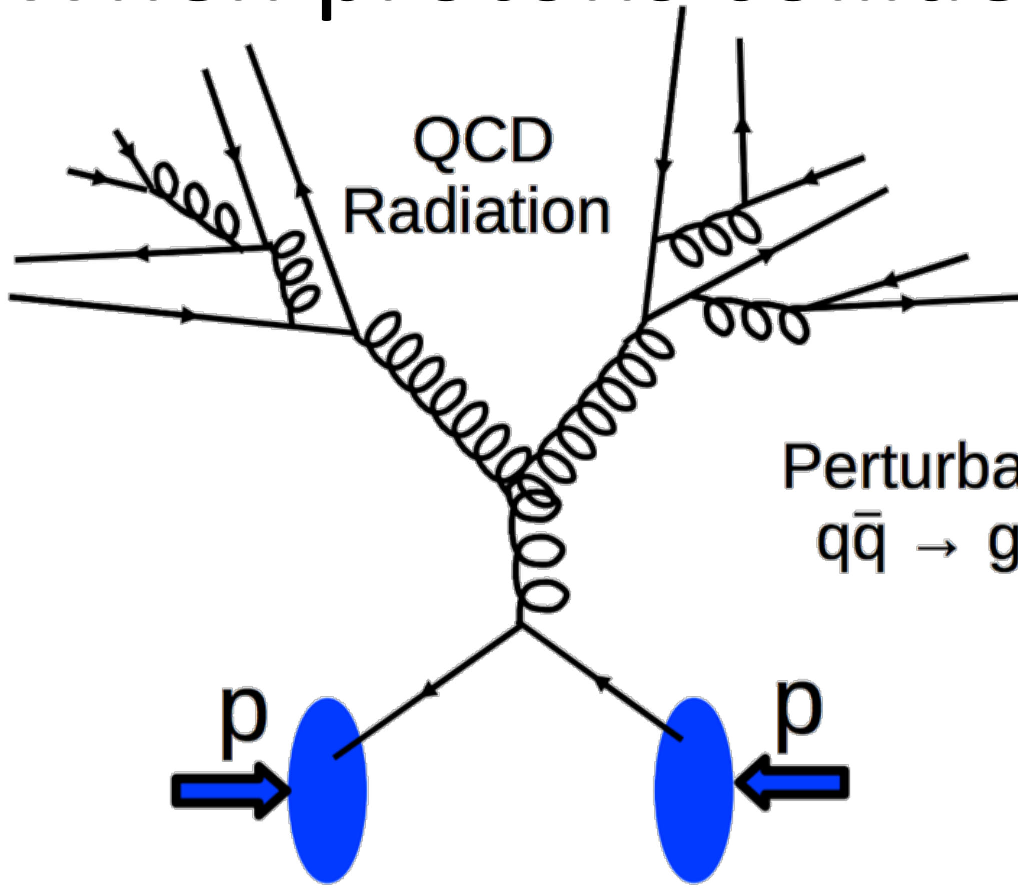
When protons collide...



Interaction among fundamental constituent

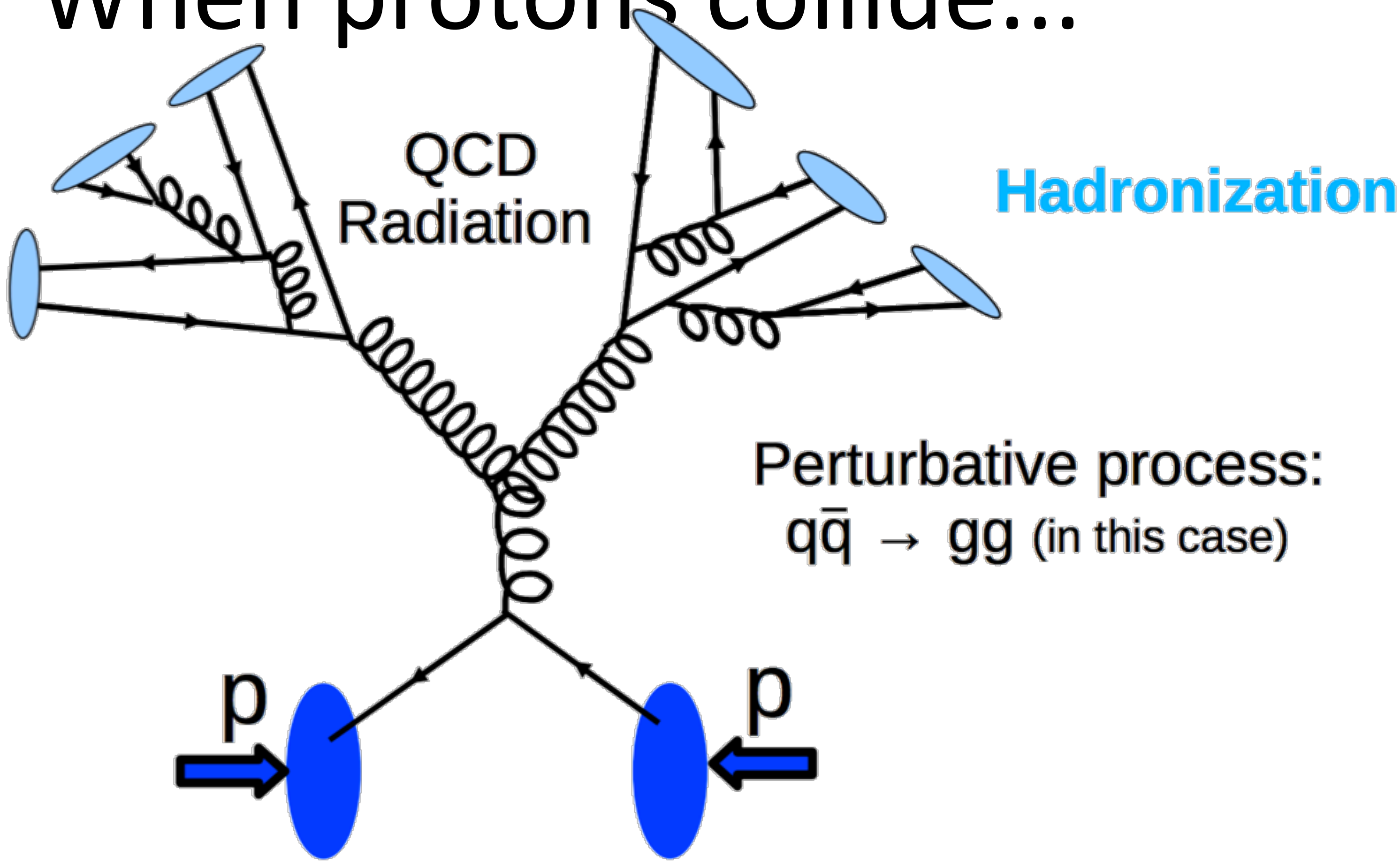
Perturbative process:
$q\bar{q} \rightarrow gg$ (in this case)

When protons collide...



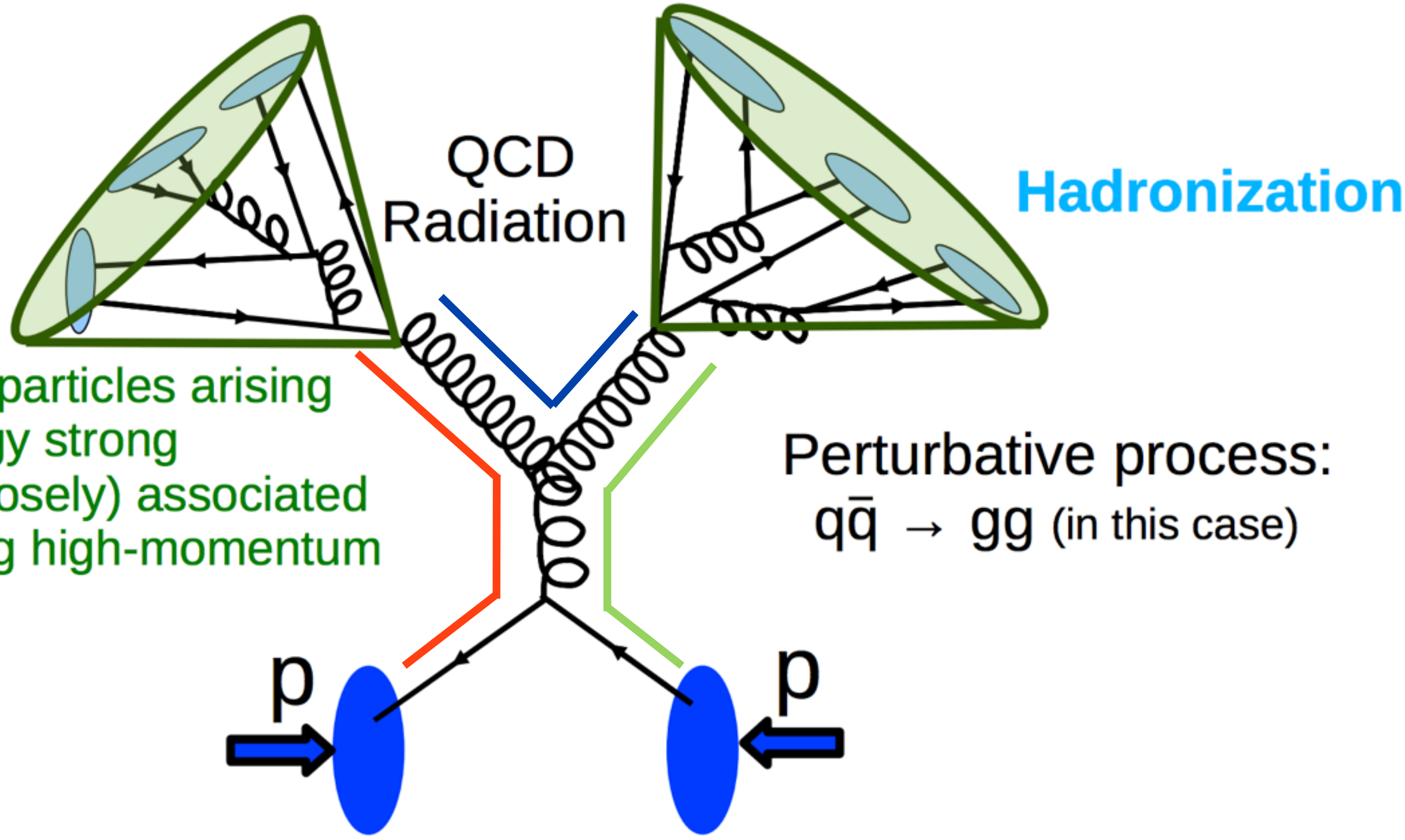
Perturbative process:
$q\bar{q} \rightarrow gg$ (in this case)

When protons collide...

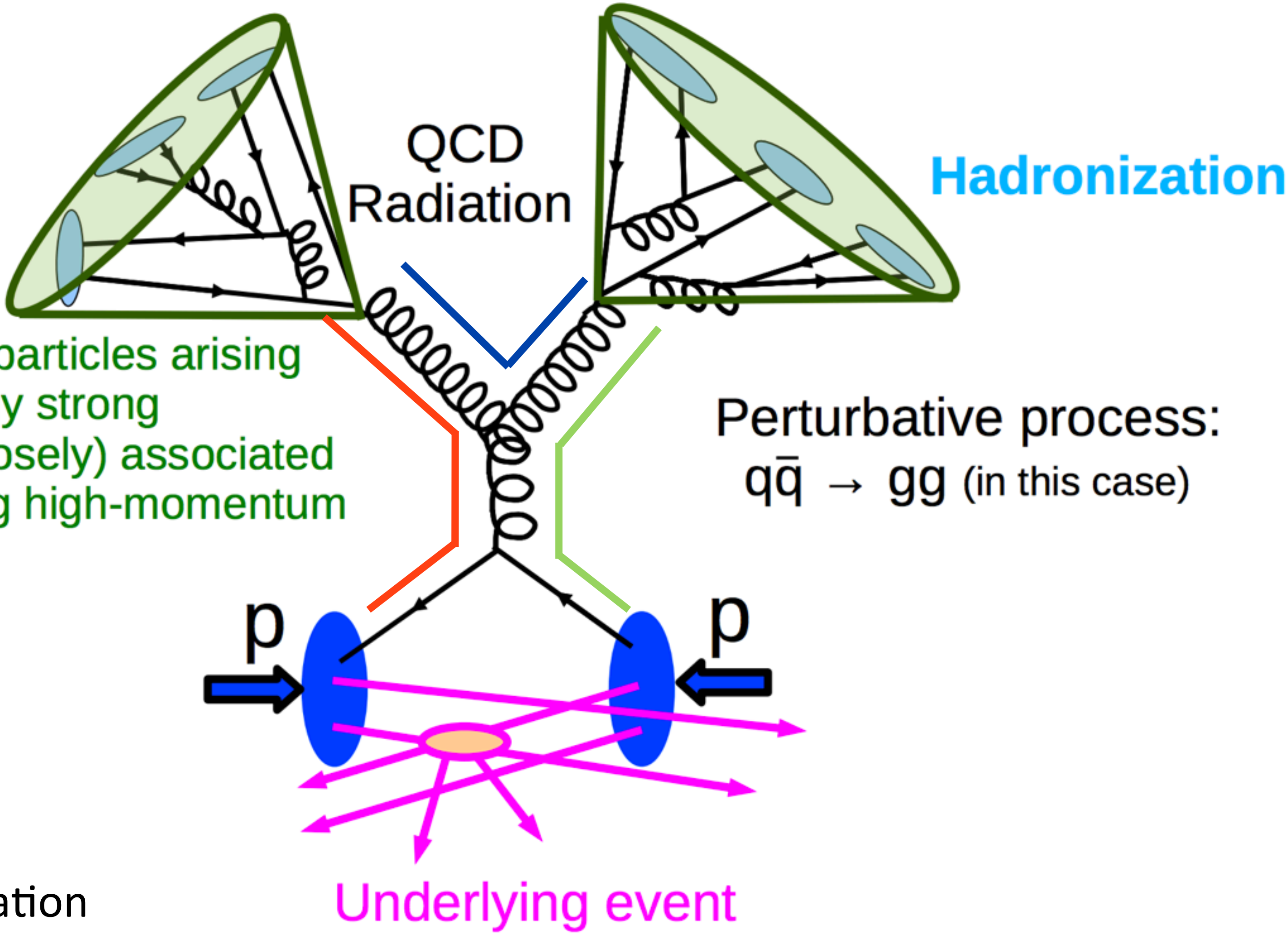


When protons collide…

Jet: collimated particles arising from high-energy strong interactions, (loosely) associated with an outgoing high-momentum parton



When protons collide...

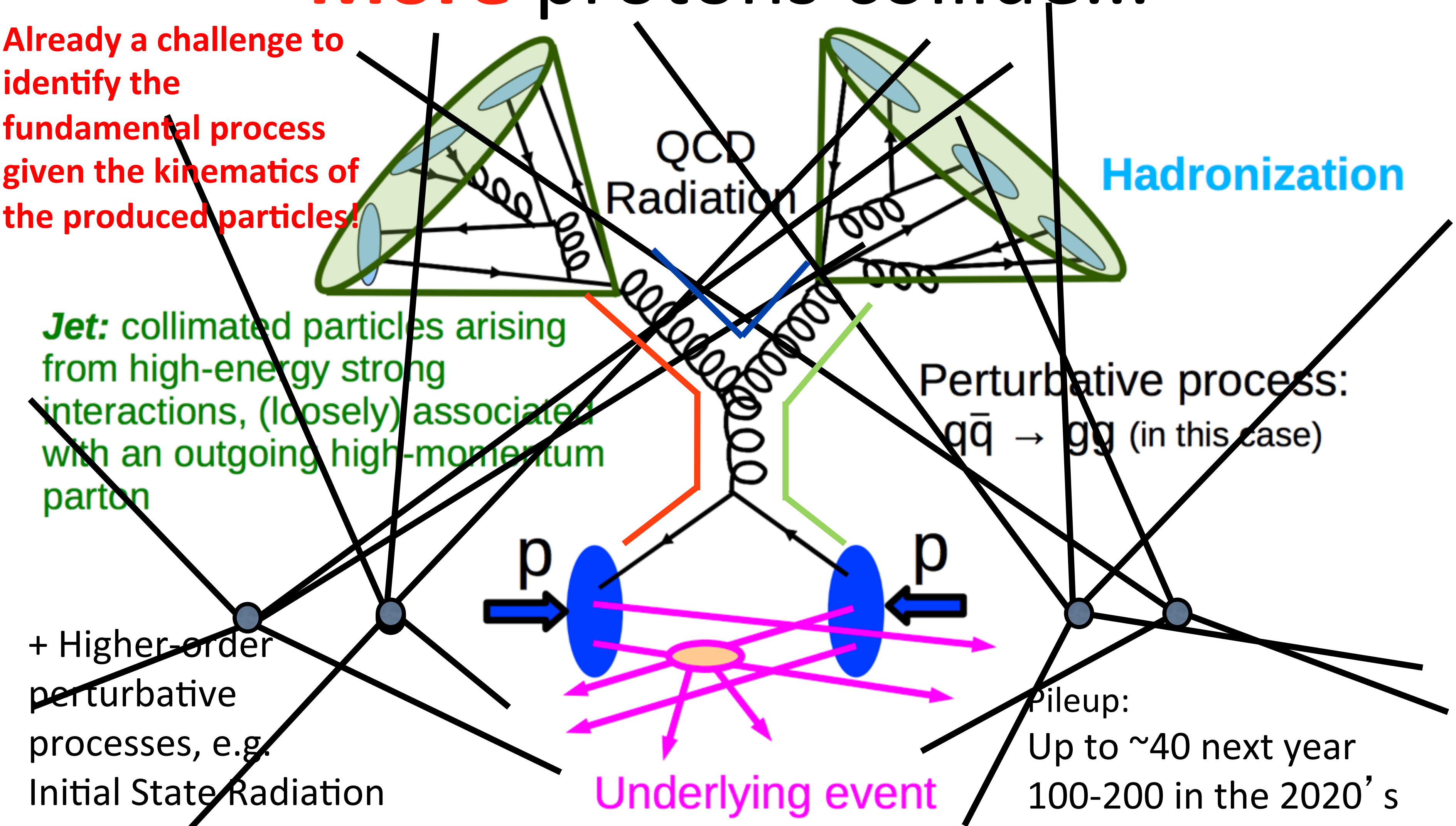


Jet: collimated particles arising from high-energy strong interactions, (loosely) associated with an outgoing high-momentum parton

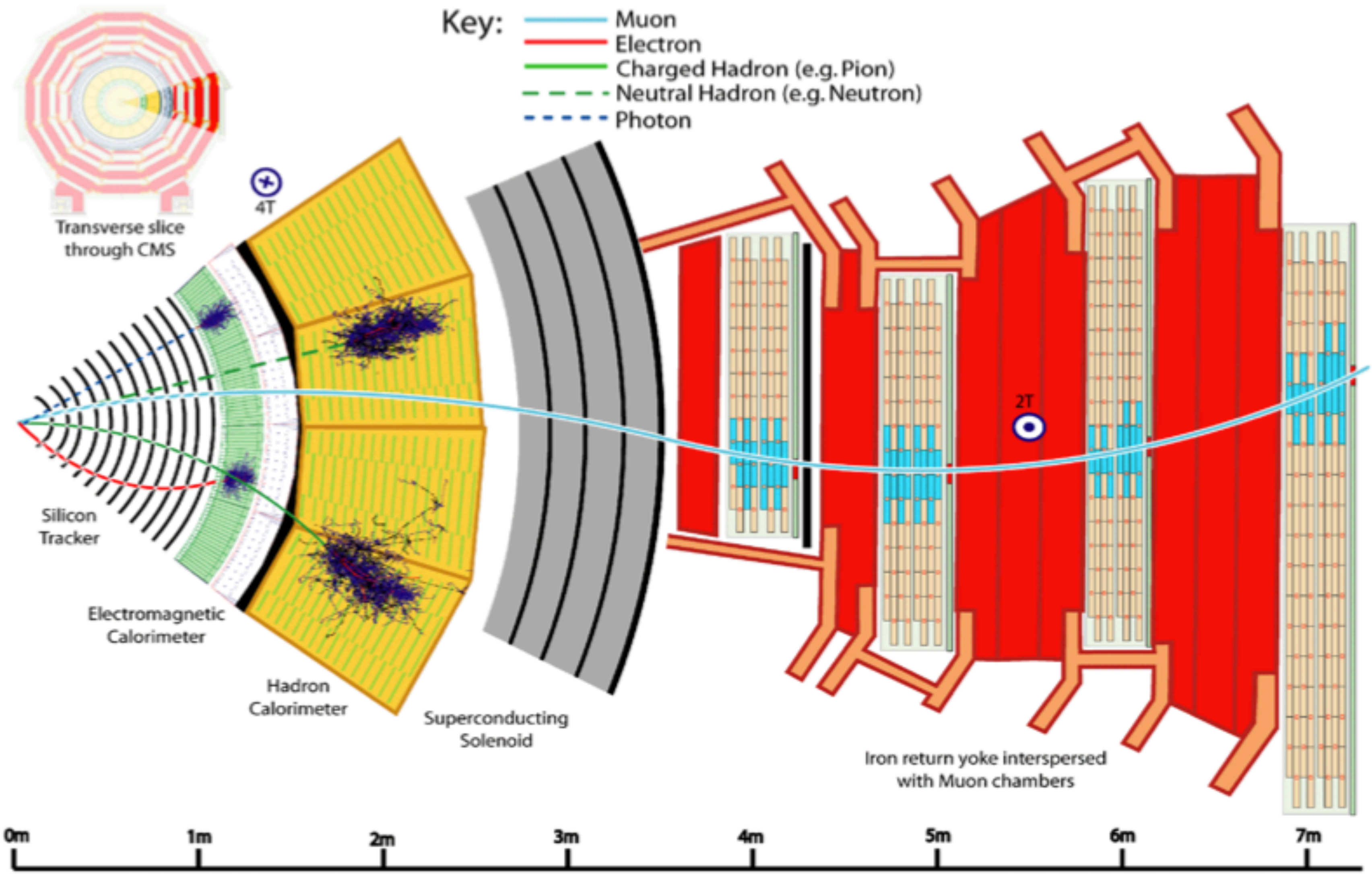
Perturbative process:
$q\bar{q} \rightarrow gg$ (in this case)

+ Higher-order perturbative processes, e.g. Initial State Radiation

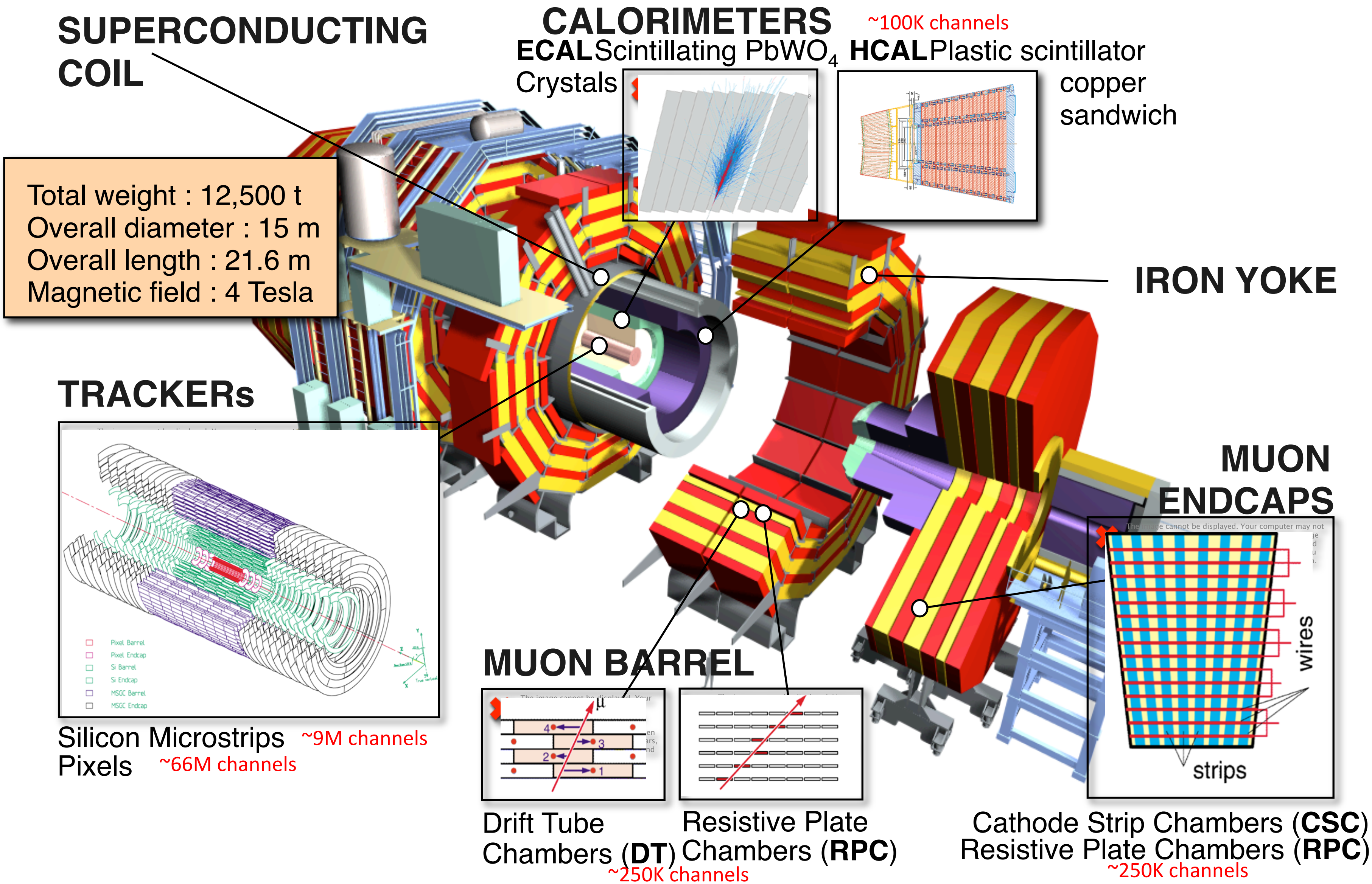
More protons collide...



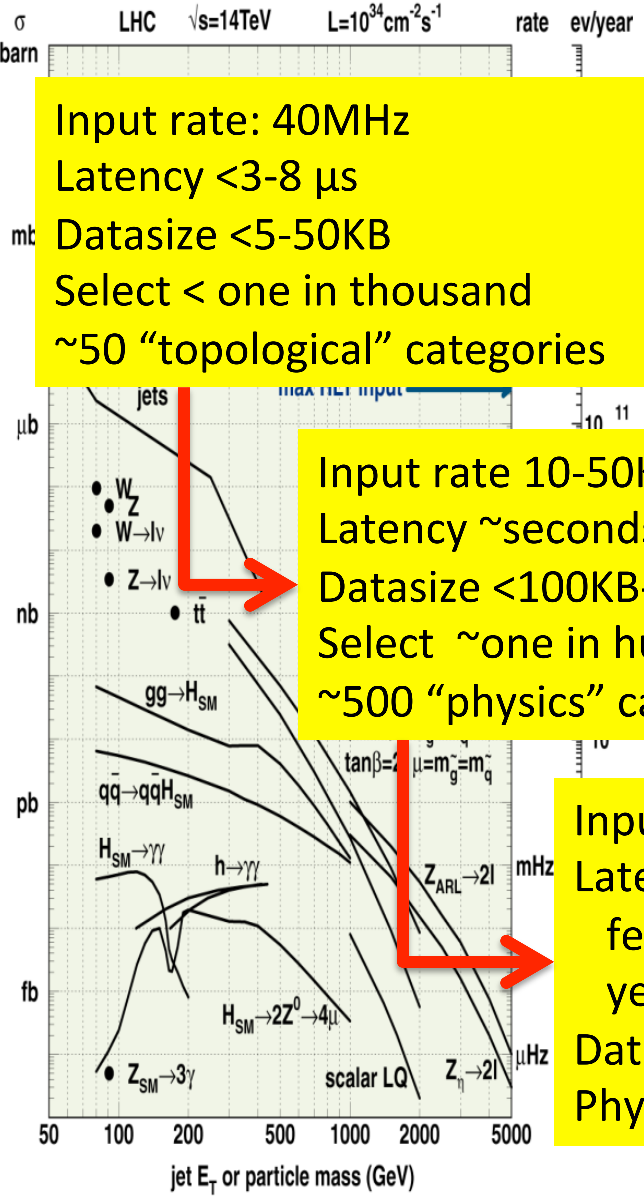
Detector “onion” structure



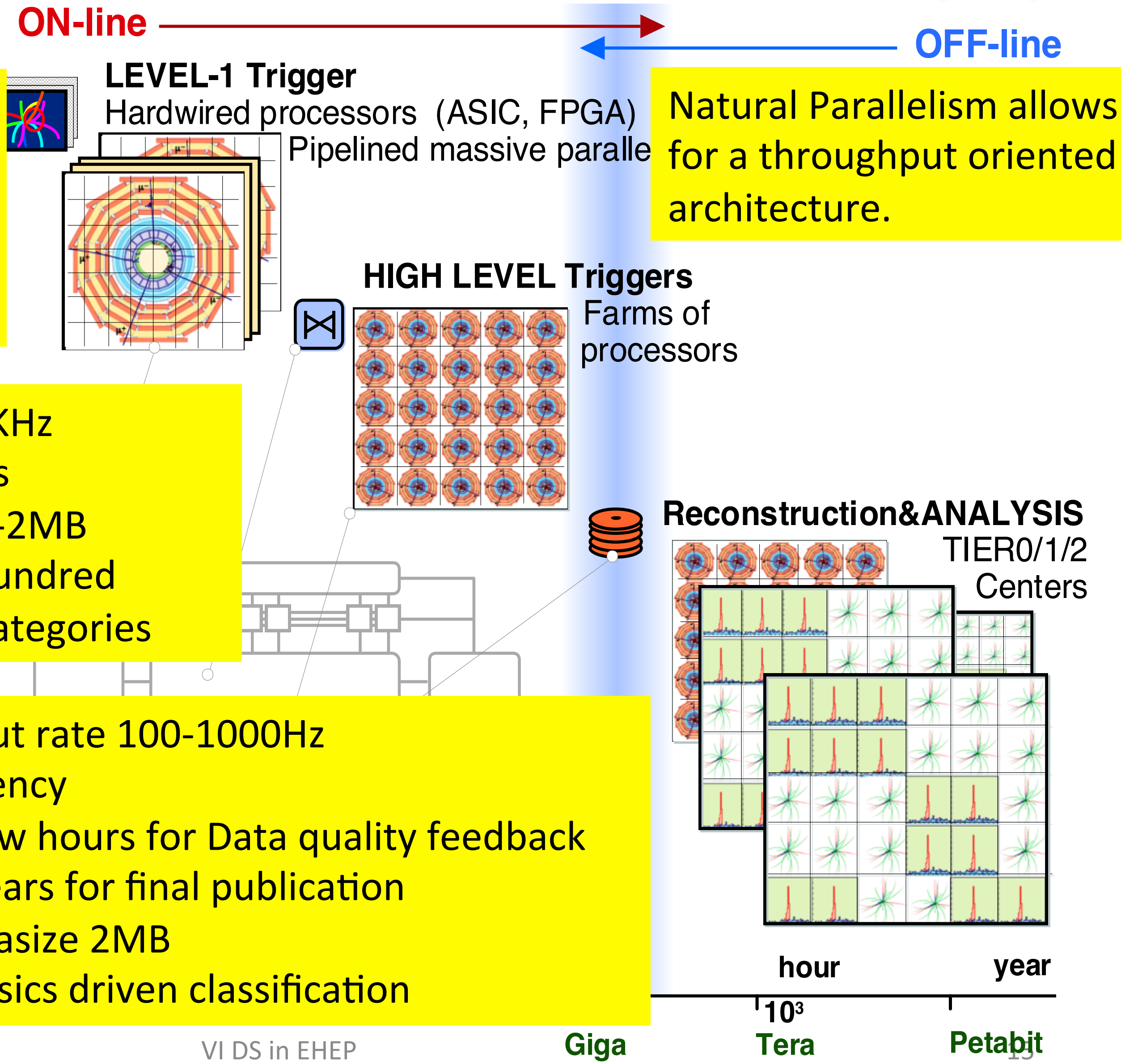
An experiment: CMS



Data Flow



Input rate: 40MHz
Latency <3-8 μs
Datasize <5-50KB
Select < one in thousand
~50 "topological" categories



Natural Parallelism allows for a throughput oriented architecture.

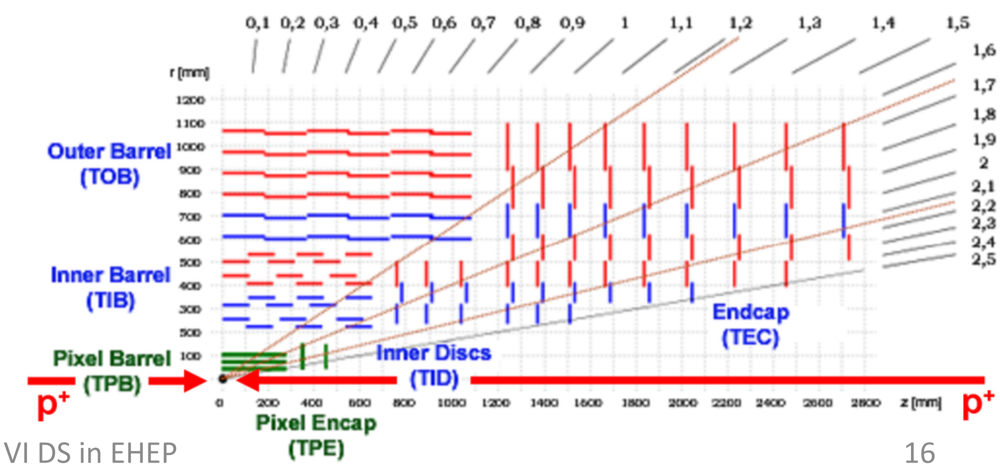
Input rate 10-50KHz
Latency ~seconds
Datasize <100KB-2MB
Select ~one in hundred
~500 "physics" categories

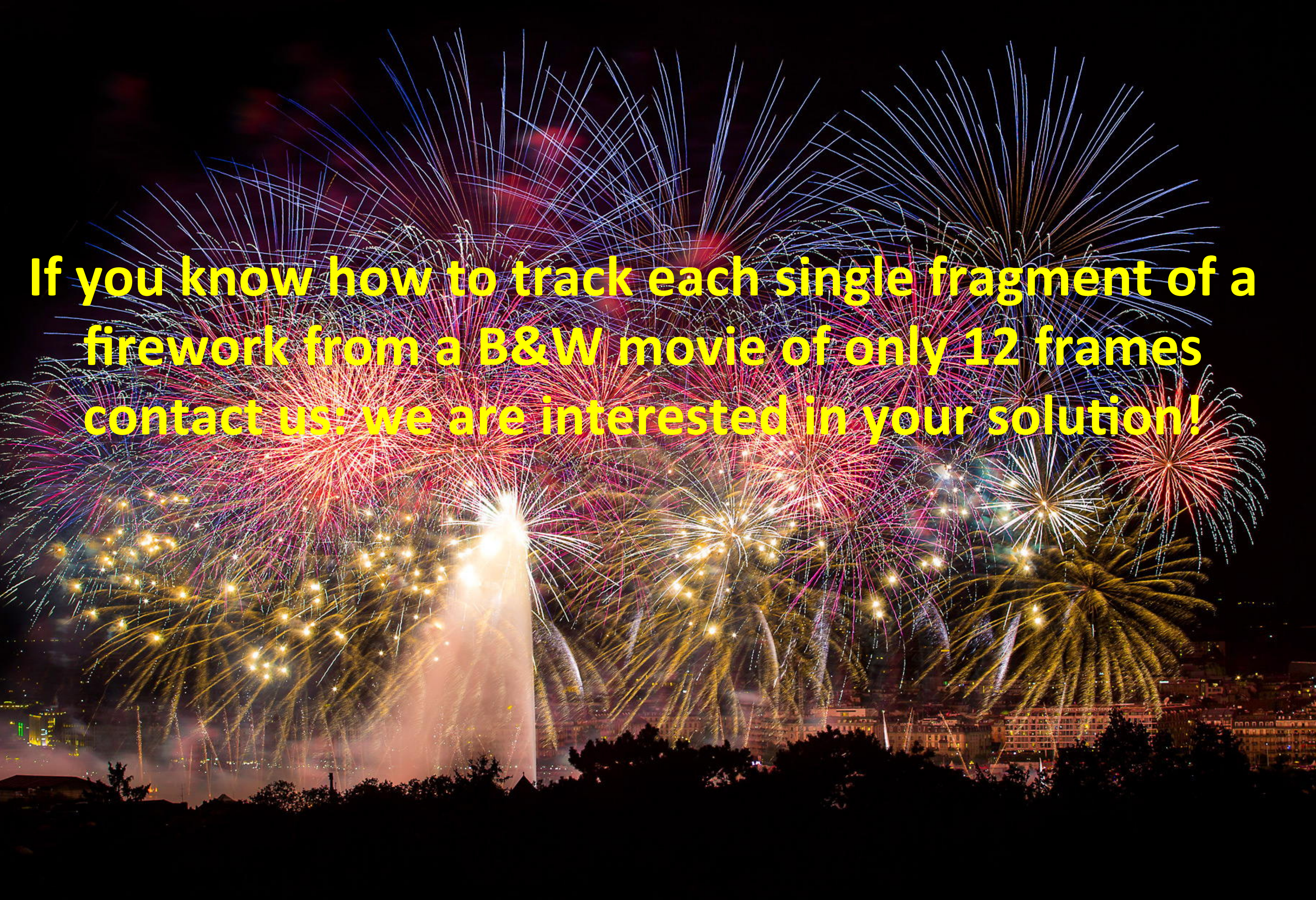
Input rate 100-1000Hz
Latency
- few hours for Data quality feedback
- years for final publication

Datasize 2MB
Physics driven classification

Tracking at CMS

- Particles produced in the collisions leave traces (hits) as they fly through the detector
- The innermost detector of CMS is called **Tracker**
- **Tracking**: the art of associate each hit to the particle that left it
- The collection of all the hits left by the same particle in the tracker along with some additional information (e.g. momentum, charge) defines a **track**
- **Pile-up**: # of p-p collisions per bunch crossing





If you know how to track each single fragment of a firework from a B&W movie of only 12 frames contact us: we are interested in your solution!

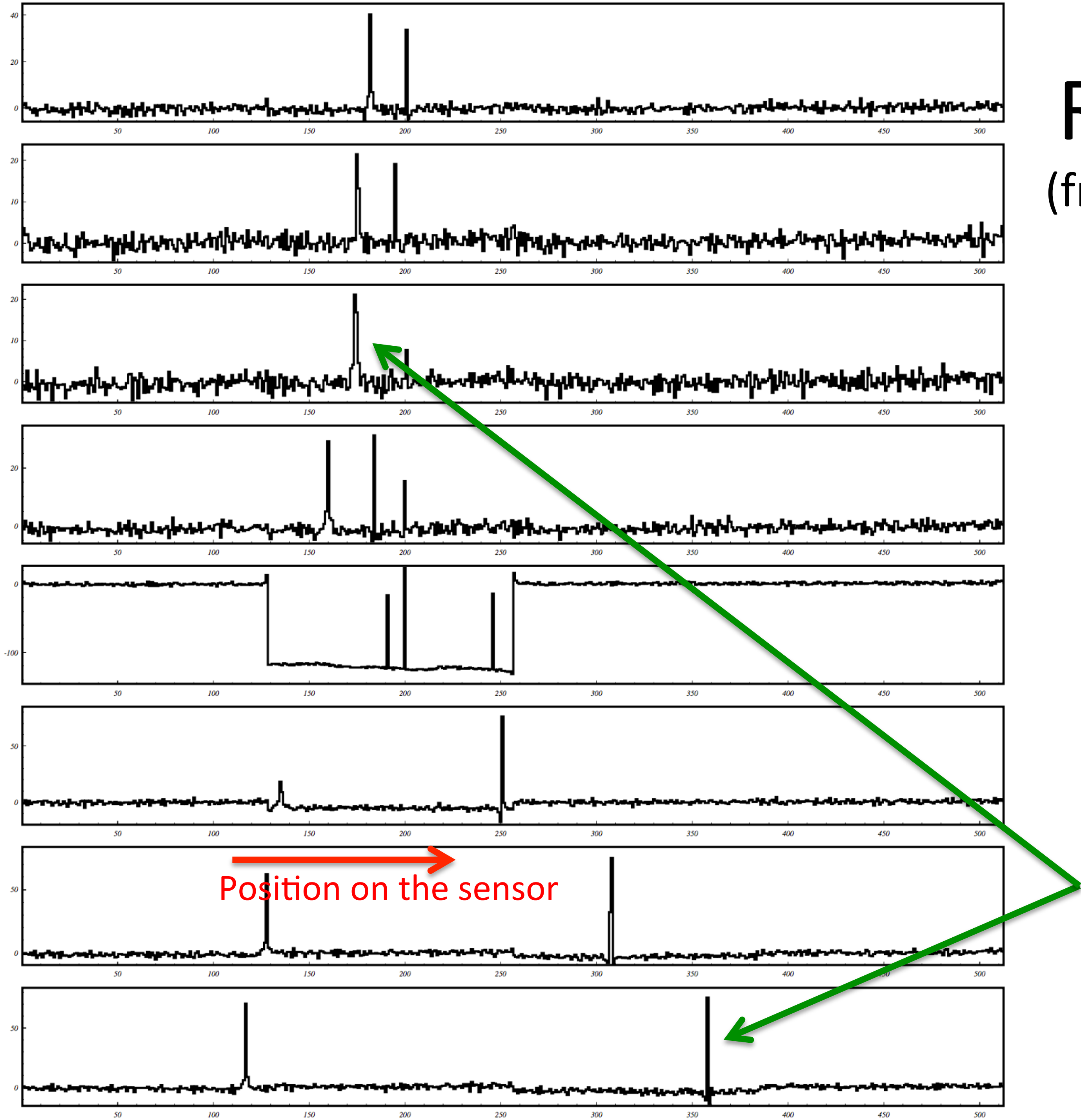
credit © Genève Tourisme

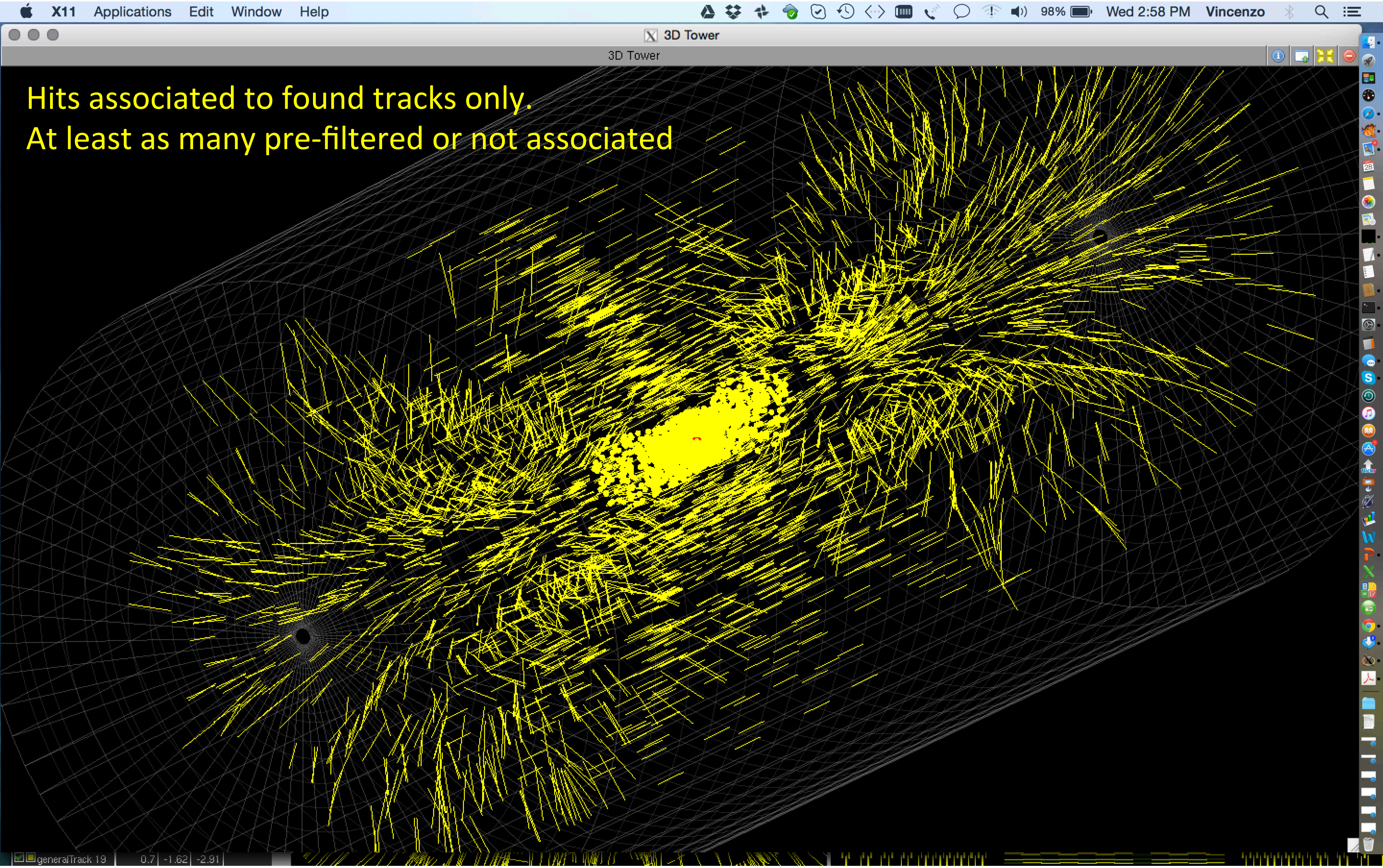
Raw Data

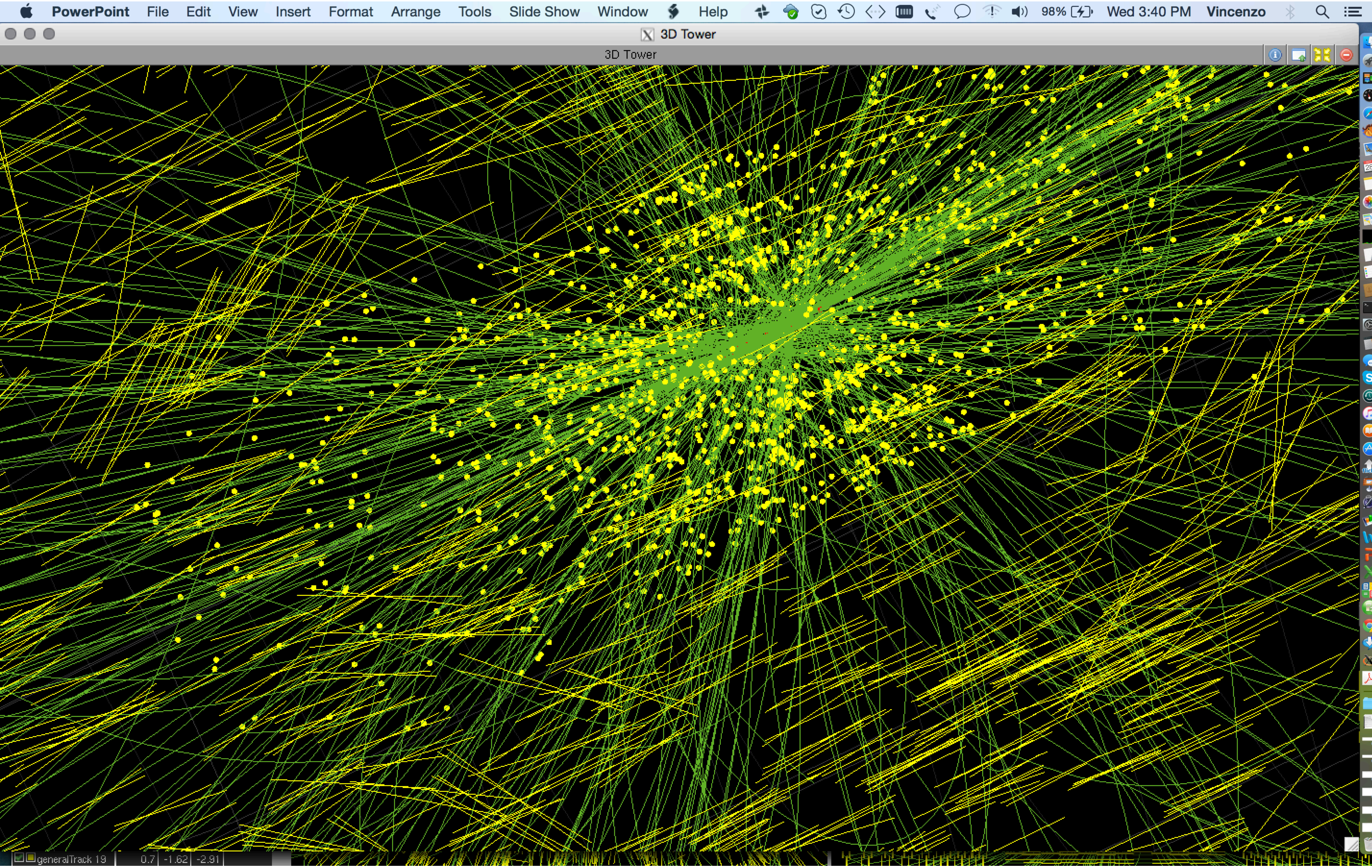
(from a test setup)

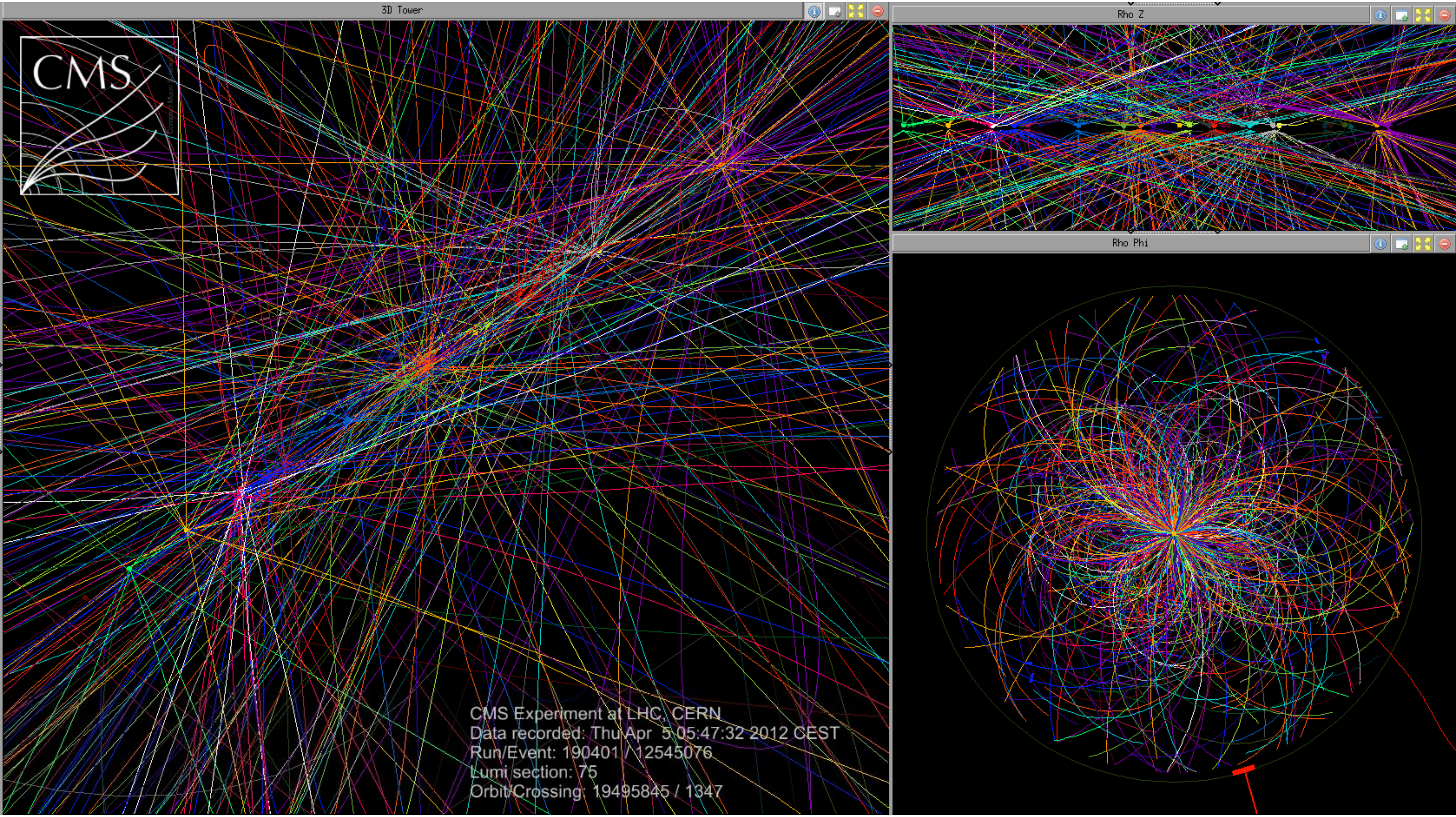
~25K sensors

Zero suppressed in readout hardware (still mostly noise)



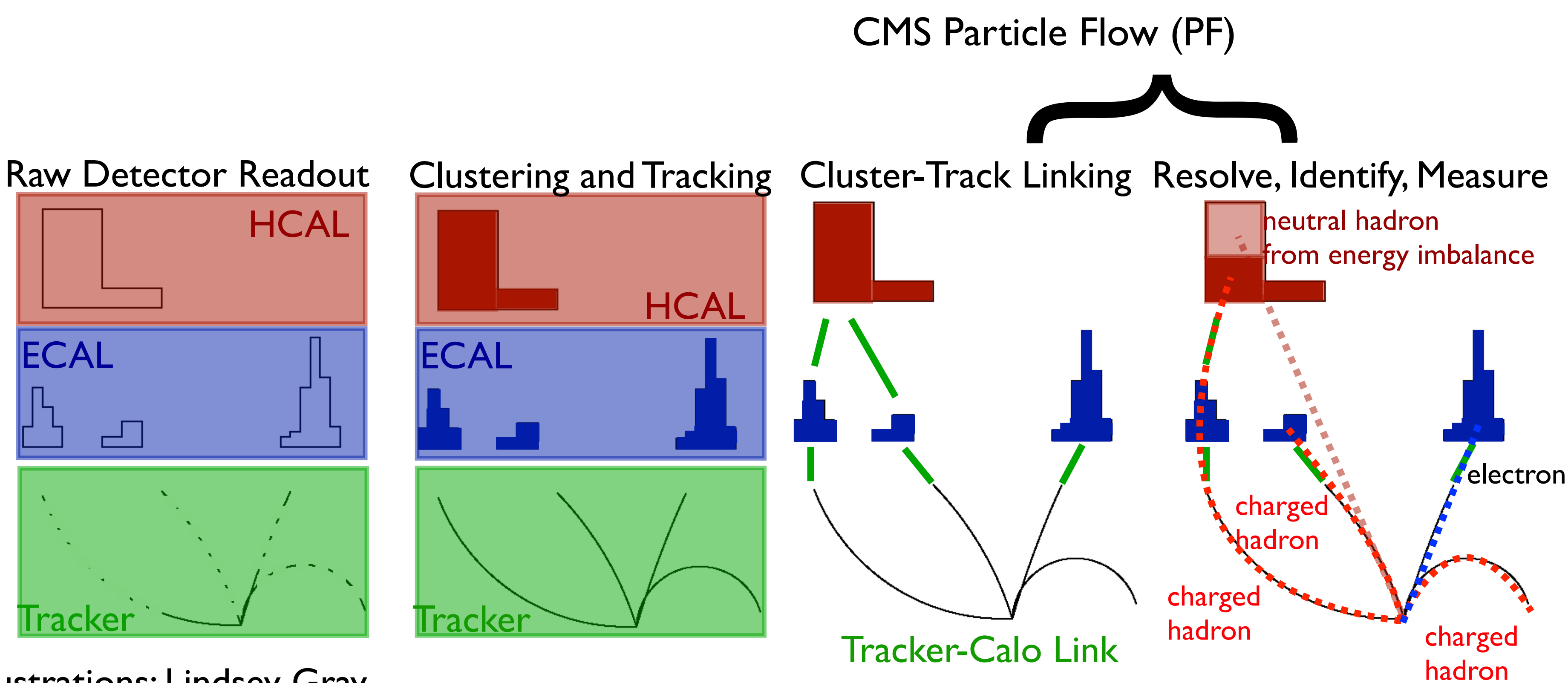

X11
Applications
Edit
Window
Help
98%
Wed 2:58 PM
Vincenzo
3D Tower
3D Tower
Hits associated to found tracks only.
At least as many pre-filtered or not associated
generalTrack 19
0.7
-1.62
-2.91


PowerPoint File Edit View Insert Format Arrange Tools Slide Show Window Help
98% Wed 3:40 PM Vincenzo
3D Tower
3D Tower
generalTrack 19
0.7 -1.62 -2.91



Not just identification and classification, also precise parameter determination

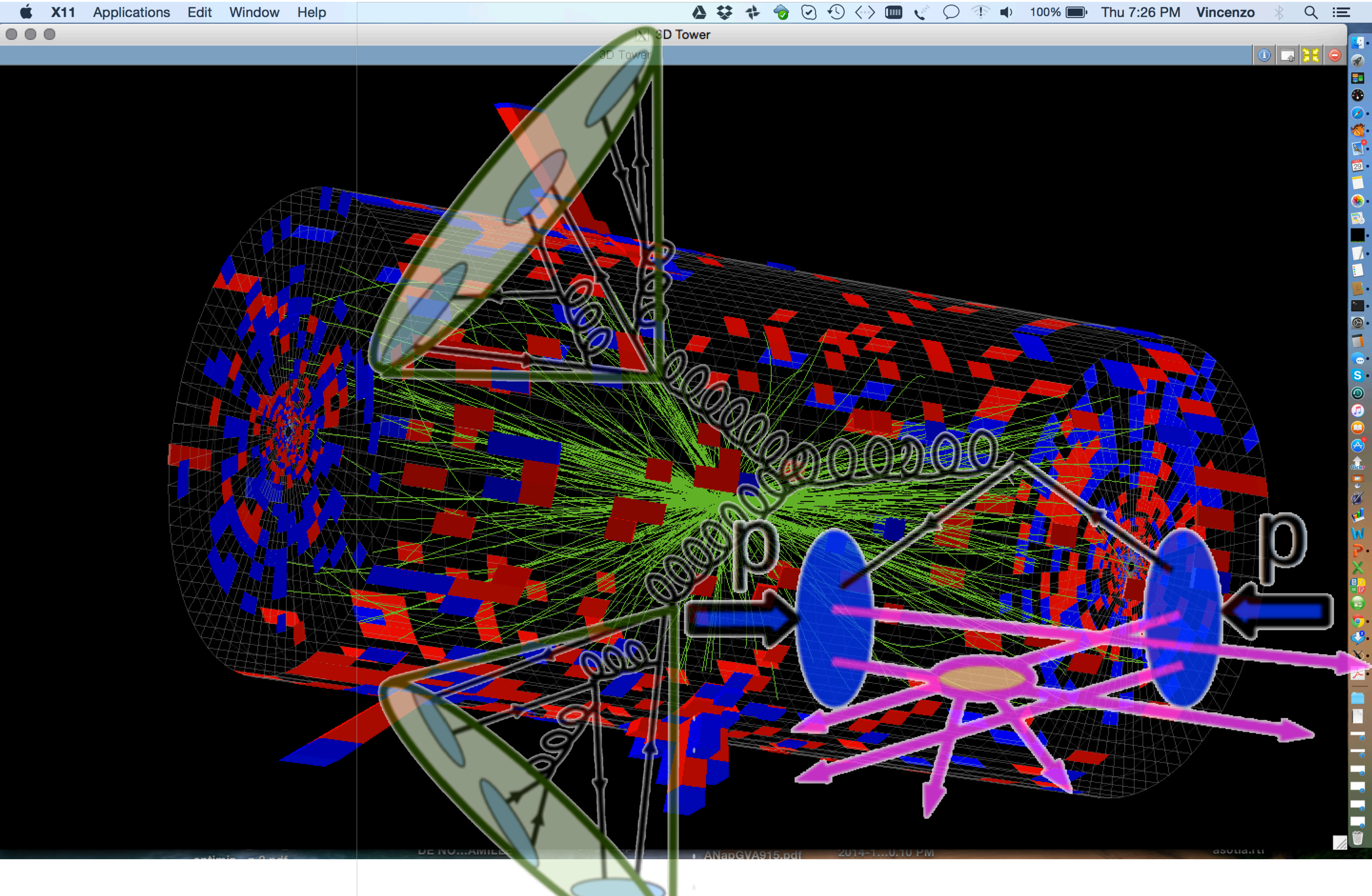
Reconstructing Jet Constituents



Illustrations: Lindsey Gray

Non trivial regression to compute best estimation of particle energy combining all available information taking into account non-uniformity in detector response

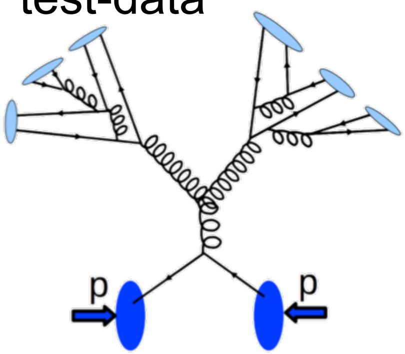
Based on intensive, iterative statistical analysis of data themselves to extract alignment and calibration constants

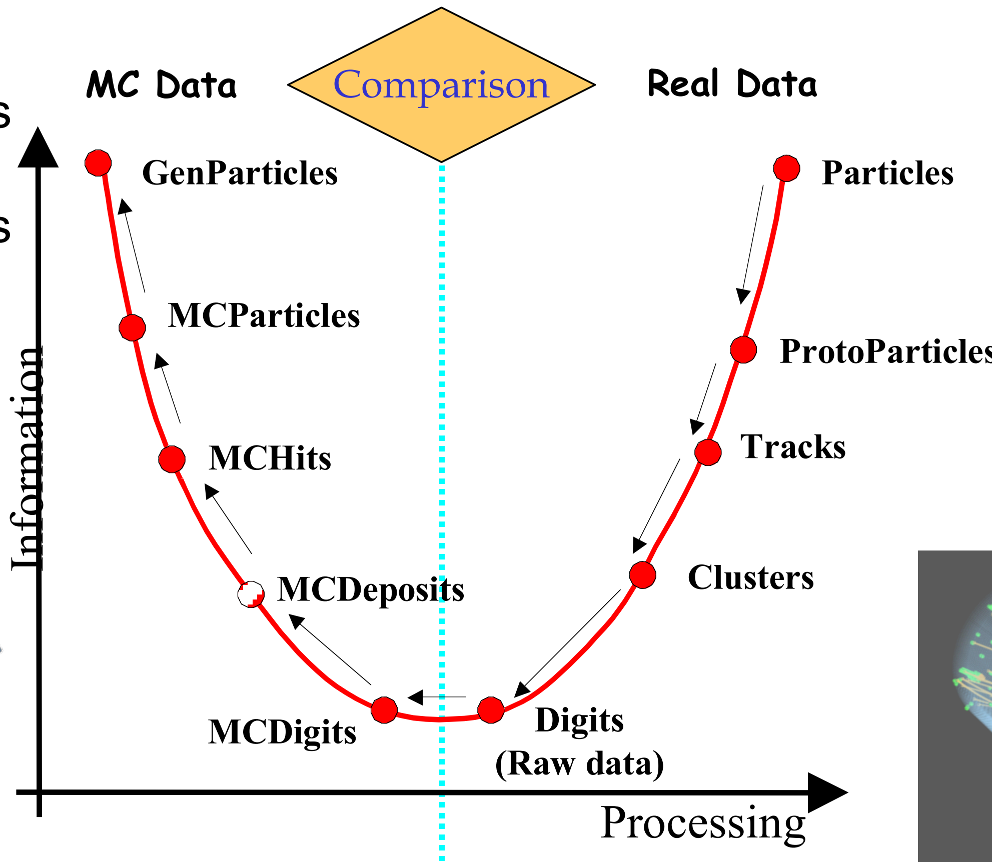


Actual granularity of red towers is ~100 times finer

High Energy Analysis Model

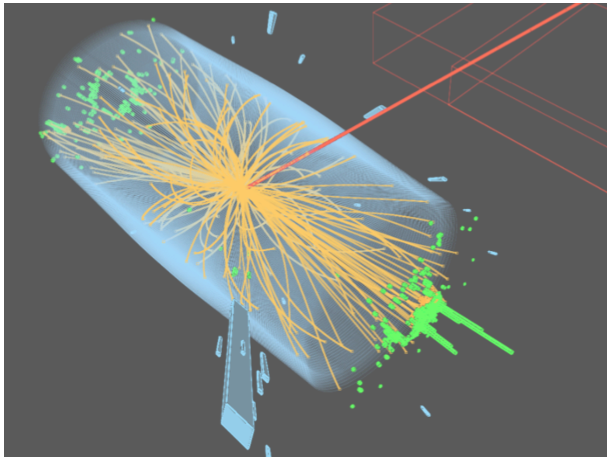
MonteCarlo Simulation models the evolution of physics processes from collision to digital signals using knowledge from theory and test-data





Reconstruction “goes back in time”, step-by-step, from digital signals to the original particles produced in the collision.

Each step uses an algorithm based on a model of the physics behind it

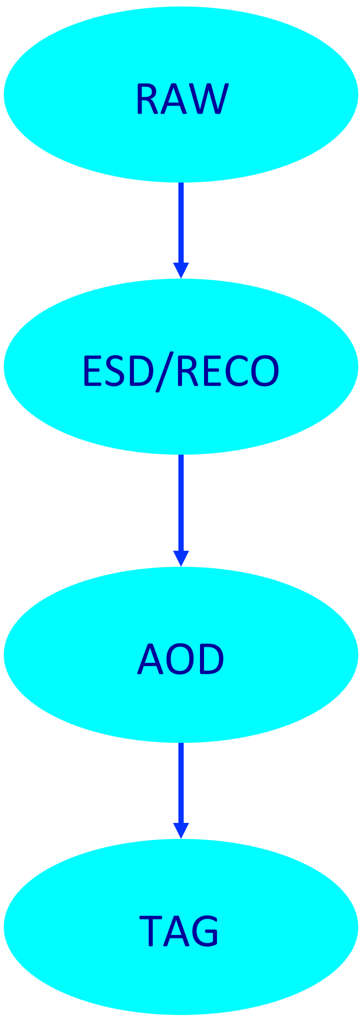


Analysis compares (at statistical level) reconstructed events from real data with those predicted by simulation

[now also for the general public: http://cms.web.cern.ch/content/cms-physics-masterclass](http://cms.web.cern.ch/content/cms-physics-masterclass)

Data Hierarchy: Our solution to BigData

"RAW, ESD, AOD, TAG"



~2 MB/event — RAW — Triggered events recorded by DAQ — Detector digitisation

~100 kB/event — ESD/RECO — Reconstructed information — Pseudo-physical information: Clusters, track candidates

~10 kB/event — AOD — Analysis information — Physical information: Transverse momentum, Association of particles, jets, (best) id of particles,

~1 kB/event — TAG — Classification information — Relevant information for "fast" event selection

Analogies with Industry

- Signal/image processing
  - Digital-Analog Conversions (including calibrations)
  - Pattern recognition, “clustering”
- Topological problems
  - Closest neighbor, minimum path, space partitioning
- Gaming *(our main source of inspiration!)*
  - “walk-through” complex 3D geometries
  - Detection of “collisions” (particles with surfaces!)
- Navigation/Avionics (Kalman filtering)
  - Tracking in a force field in presence of “noise”
  - Trajectory identification and prediction
- Regression, classification, statistical analysis
  - Determination of physical parameters
  - Assign probabilities at various level of the data hierarchy
  - Statistic analysis with full data sample

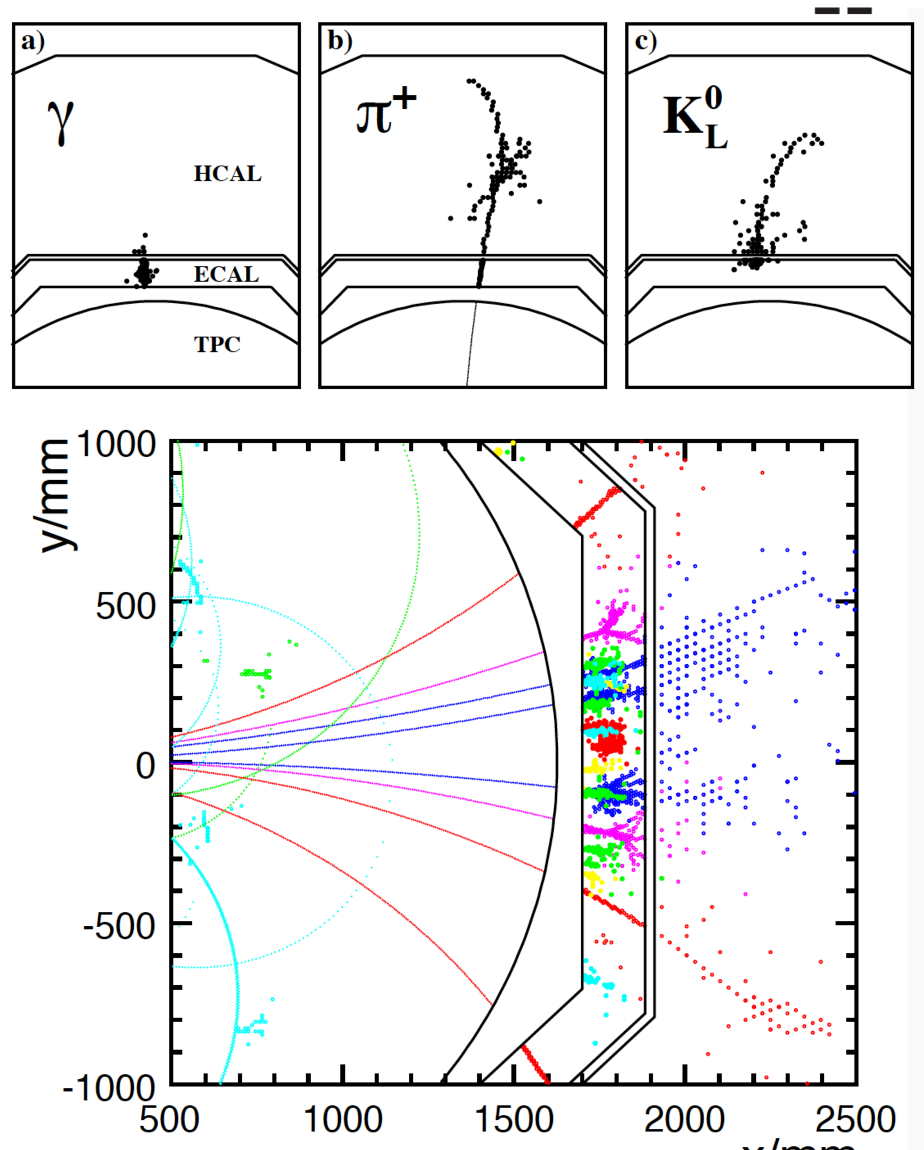
The dream of every experimental HEP Physicist:

Identify and measure each single particle produced in a collision

This may need high resolution calorimetry that will compete with trackers in complexity and data volume

Still, using current data-processing approach, most of this information will reach the physicists only in a very condensed form

Difficult to estimate the real impact of such a detector on physics analysis w/o a new data-processing paradigm

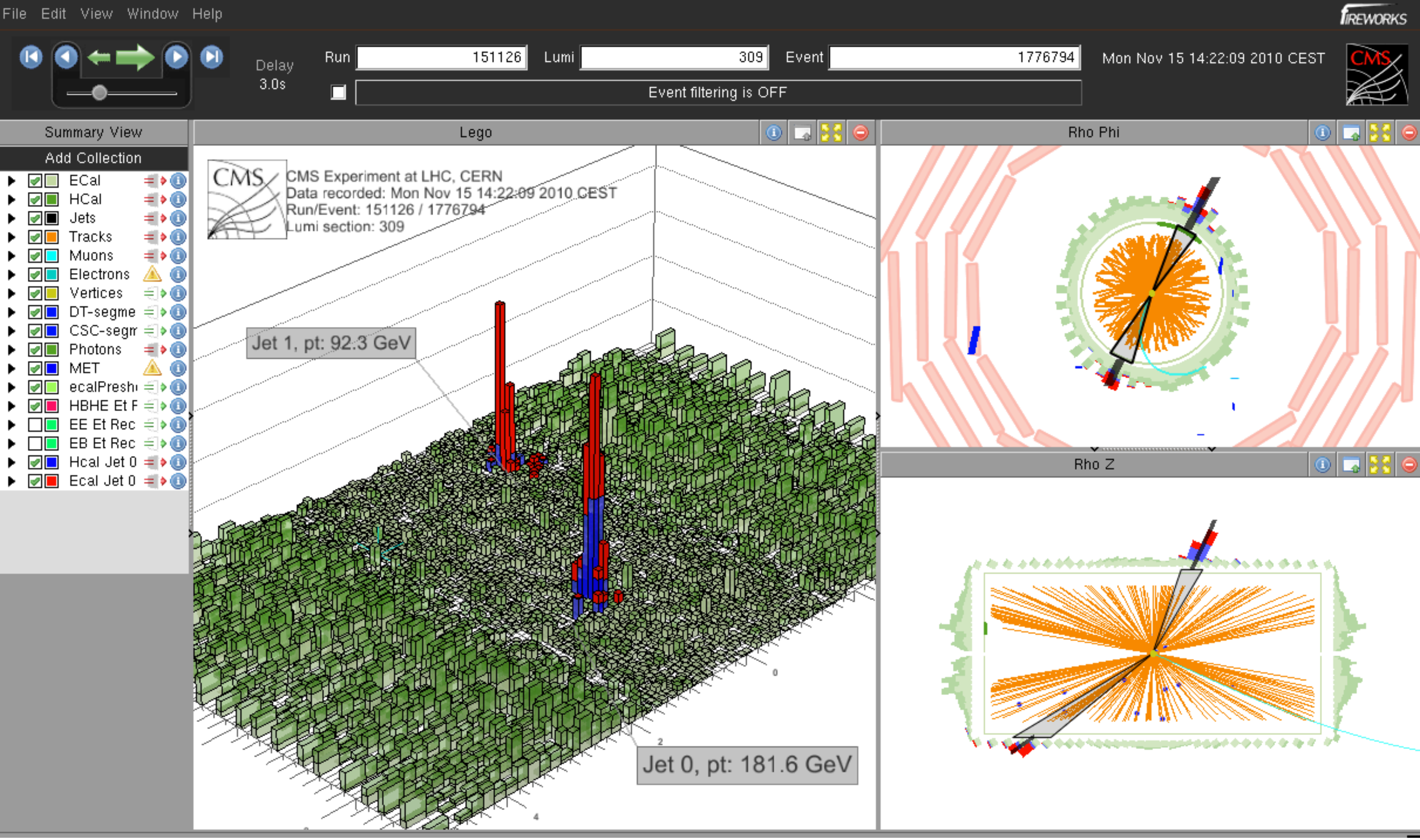


http://arxiv.org/abs/0907.3577

Summary

- HEP Data processing encompasses many solutions and technological challenges that share commonalities with industry and other natural sciences
- We build knowledge by verifying quantitative predictions based on (mathematical) modeling of the causality connection among phenomena step-by-step
- Traditional data processing pipeline is a very lossy process that precludes the access to detailed information by the physicists at the end of the chain
- New Data science technologies that can bring the whole information collected by our detectors at the fingertips of the physicists may revolutionize not just physics analysis, also the design of the experiments themselves

BACKUP



Tracker is surrounded by two layers of "calorimeters"