From Raw Data to Physics Results

Paul Laycock

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Credits for material

- **This lecture is adapted from a series of 3 lectures I gave at the CERN summer student lecture program this year**

- **All ATLAS, CMS and NA62 material is copyright CERN**
  - I invite you to look for more educational material on the CERN site, it's a very useful resource

- I have relied heavily on material from
  - **Anna Sfyrla** and **Jamie Boyd** from previous CERN Summer Student Lecture Program courses
  - Thanks also to **Dave Barney**

- I do not always credit original sources when the provenance was lost
  - apologies where that is the case

Aims, assumptions and disclaimers

- **Aim**
  - Learn about the journey of data, from the ***raw data*** we read out from the detectors that make up our experiments, to the ***highly refined data*** we ***publish*** in scientific journals

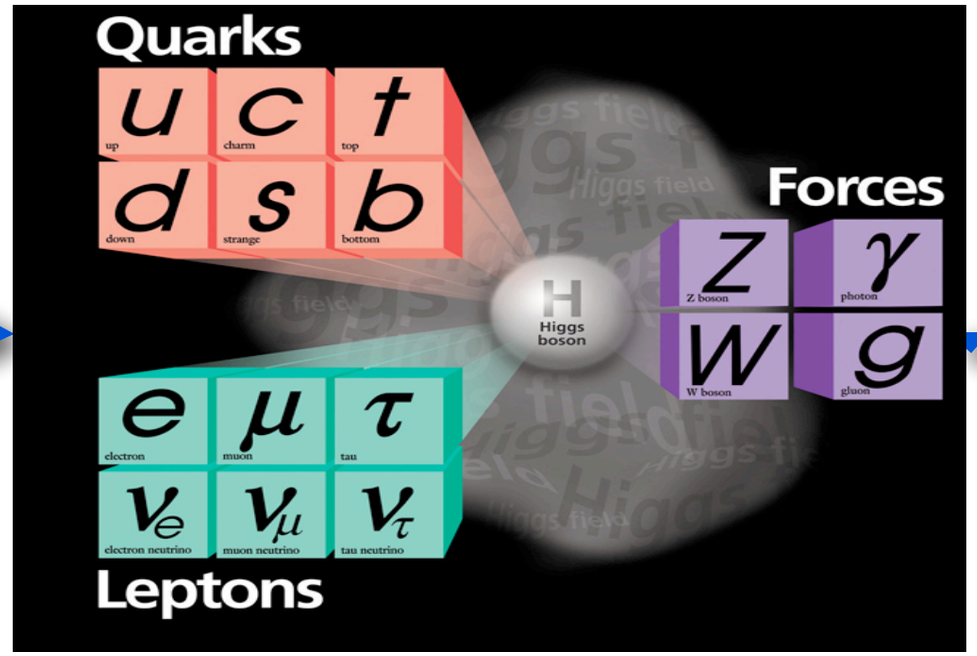
- **Assumptions**
  - You have some (first) ideas about particle physics and the questions that we're trying to answer in physics experiments worldwide
  - You know something about high energy physics detectors and their data
    - i.e. I will assume you have attended some of the other lectures in this series

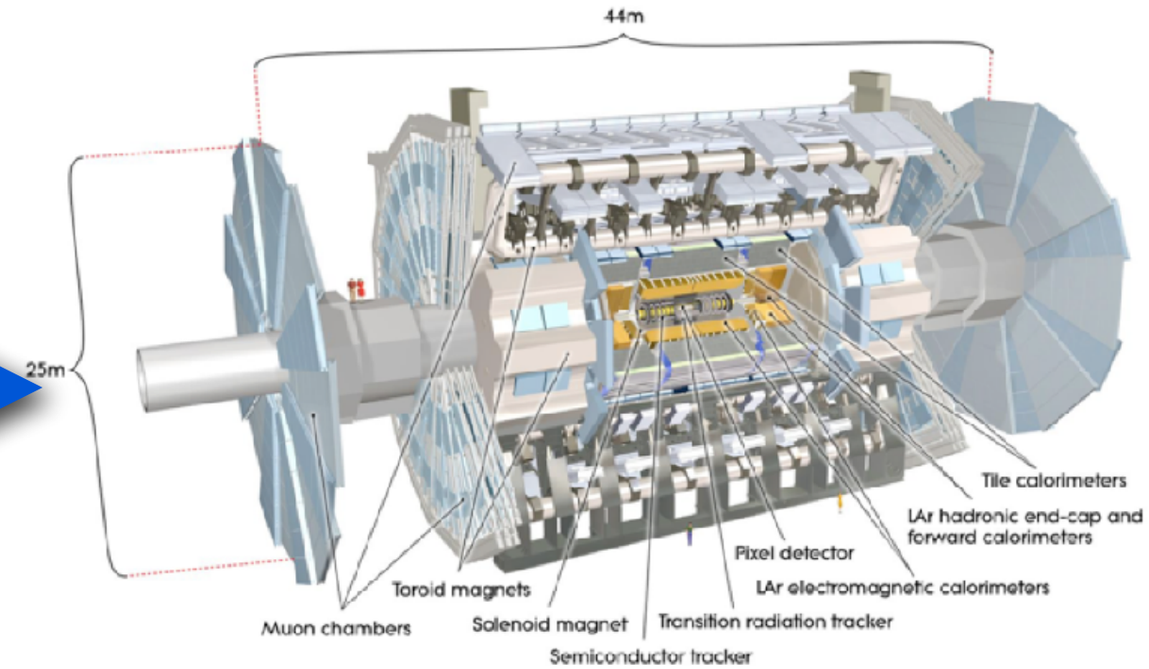
- **Disclaimer**
  - Choice of examples is often based on those experiments I've personally worked on, i.e. **H1, ATLAS, NA62** and **Belle II**

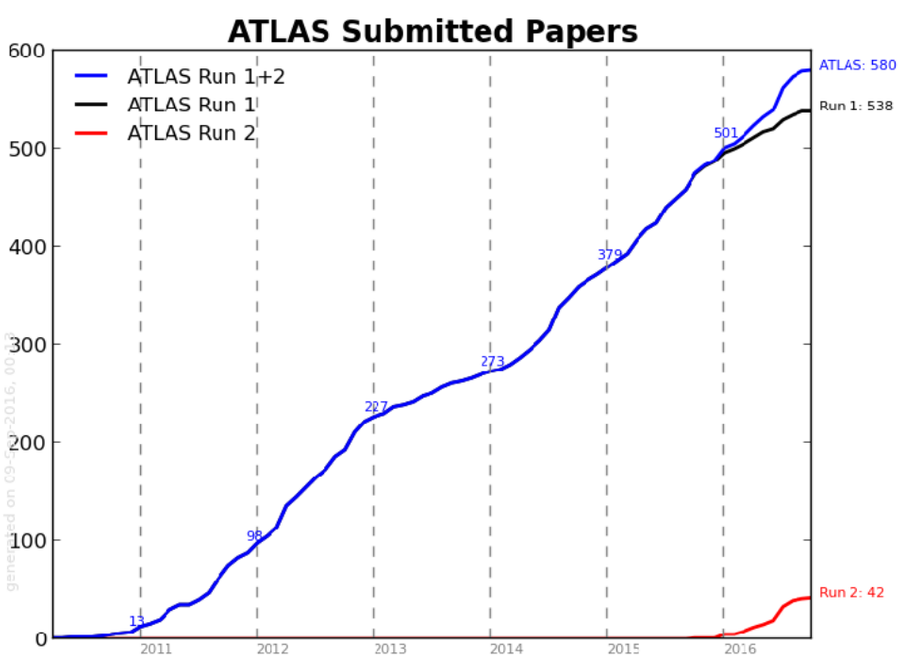
- Feedback and questions are welcome: laycock@bnl.gov
  - *email is a good way to contact me*

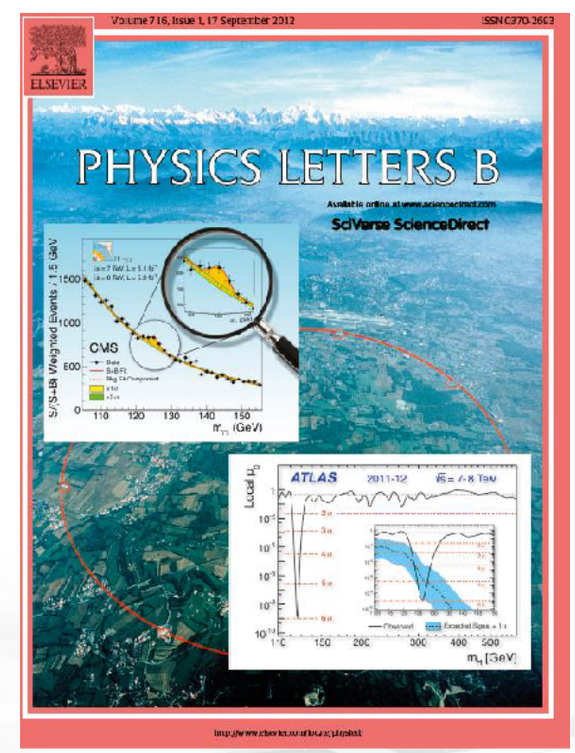

Brookhaven National Laboratory
RHIC
New York Blue Supercomputer
Interdisciplinary Energy Science Building
NSLS
CFN
NSLS-II
Long Island Solar Farm
Taken from Ketevi Assamagan

The particle physics cycle





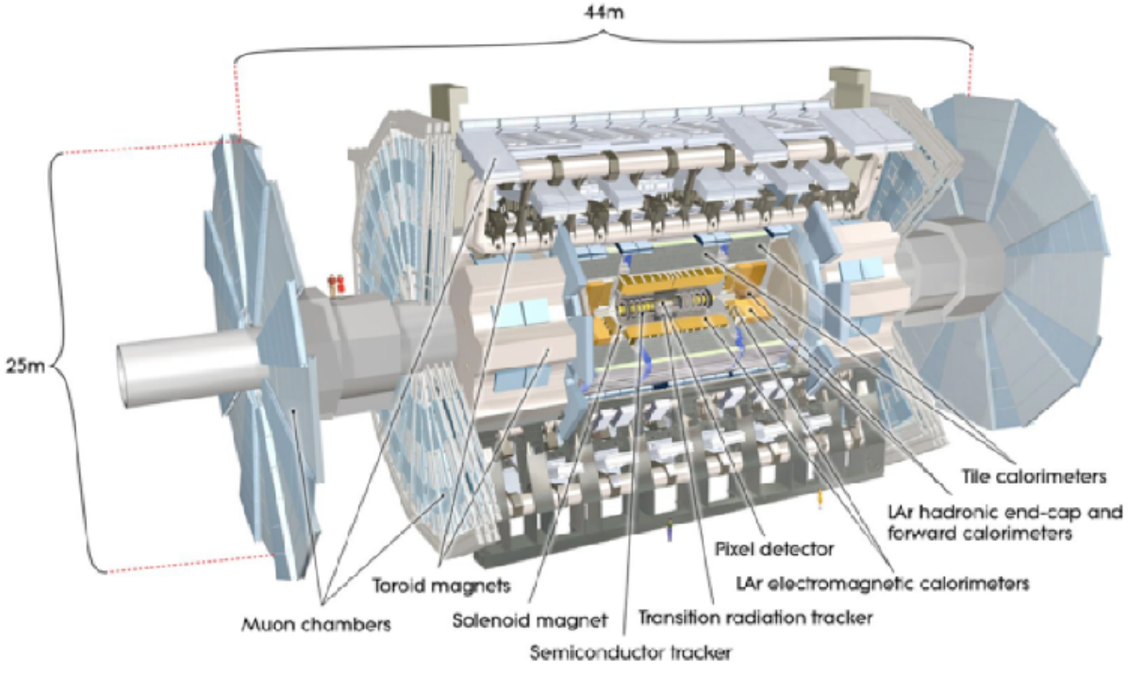




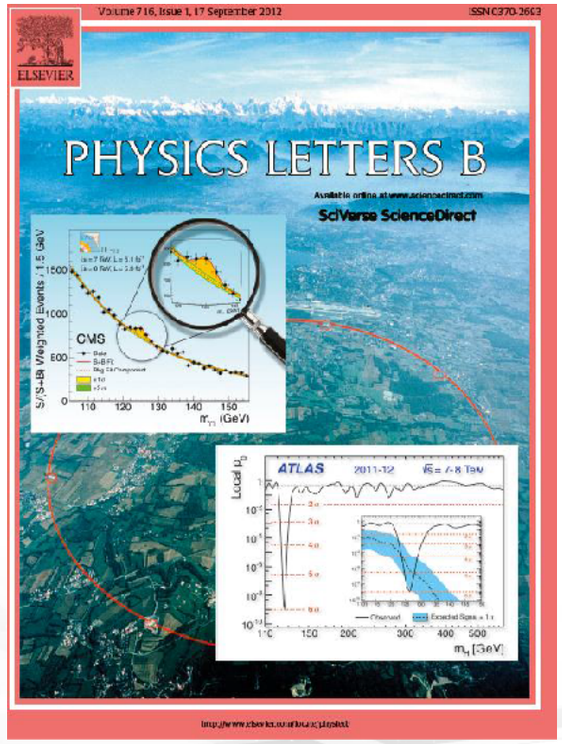
Experimental physics

- Much of the work of the experimental physicist is running experiments and extracting measurements from them

- ***Note*** *- Experimental physicists also need to propose, design and build new experiments (see previous slide)*

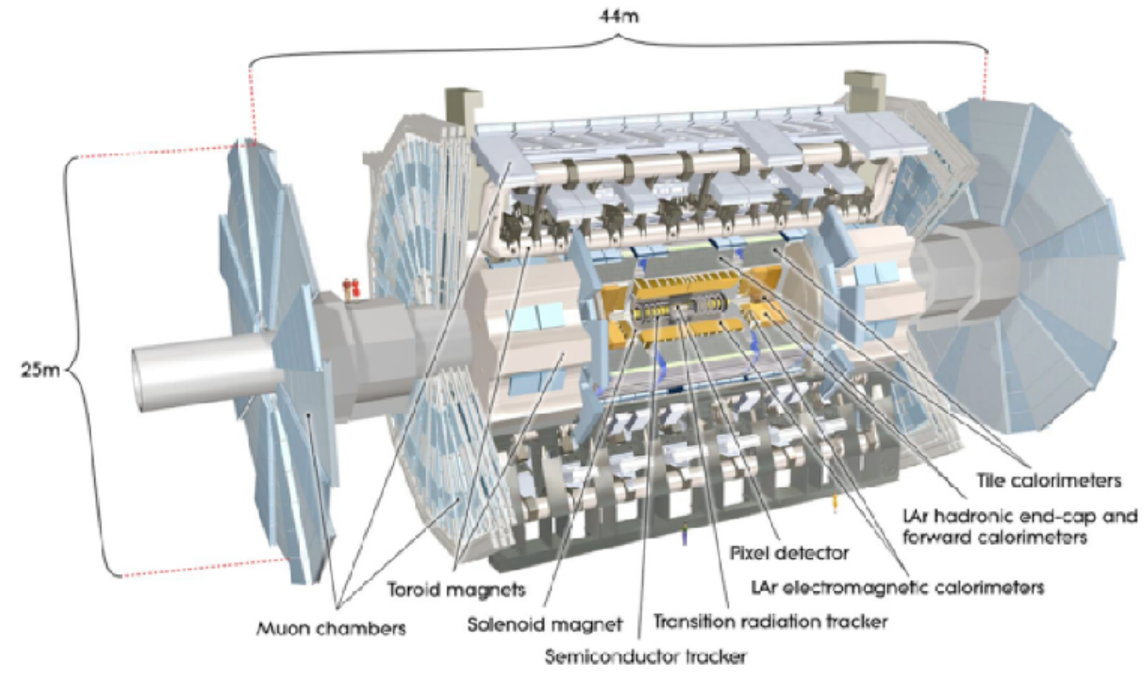


- These lectures are focused on understanding how we turn raw experimental detector data into physics results that we can publish
  - Results must be ***accurate***
  - with well understood ***precision***
  - It's important to understand the difference between these two words, we often confuse them

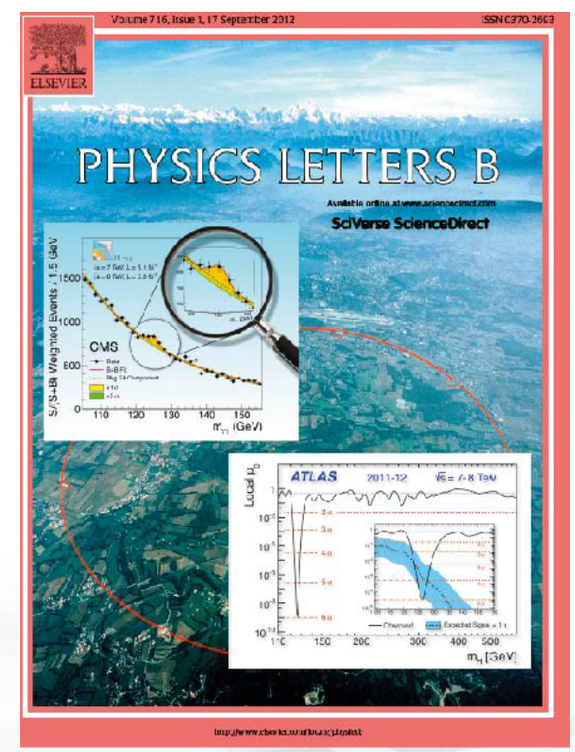


This lecture

- **Focus**
  - The journey of raw data from the detector to a publication



- **Requirements**
  - How we reconstruct fundamental physics processes from raw detector data



  - How we extract our signals from the mountain of data, finding needles in the haystack

Testing theoretical predictions

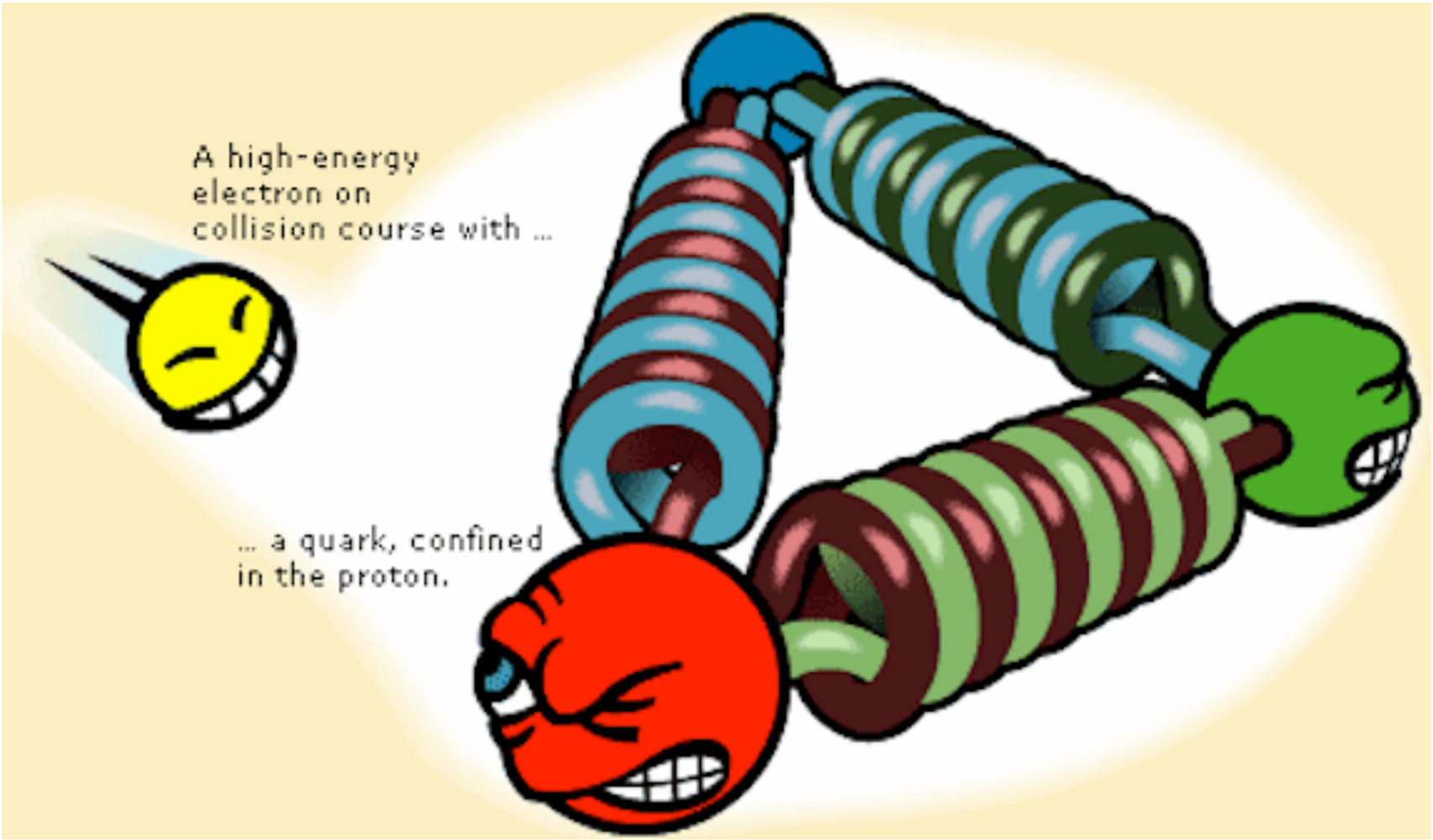
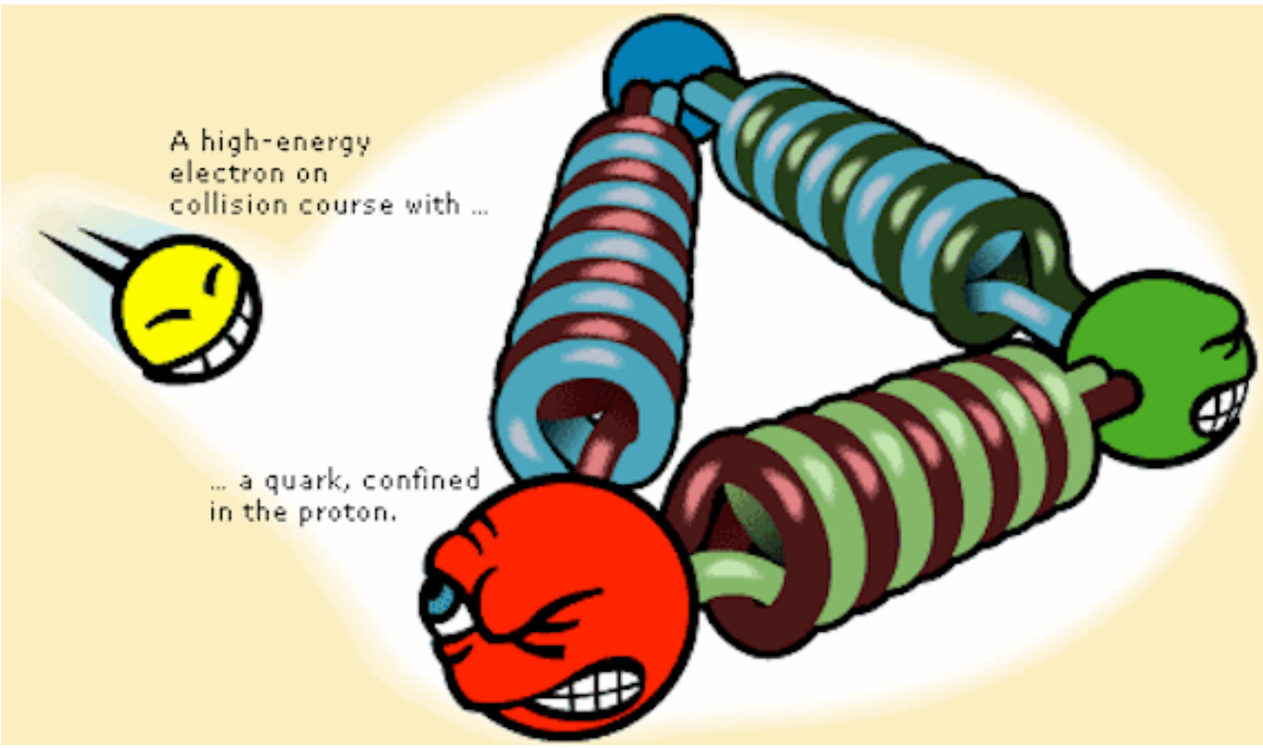


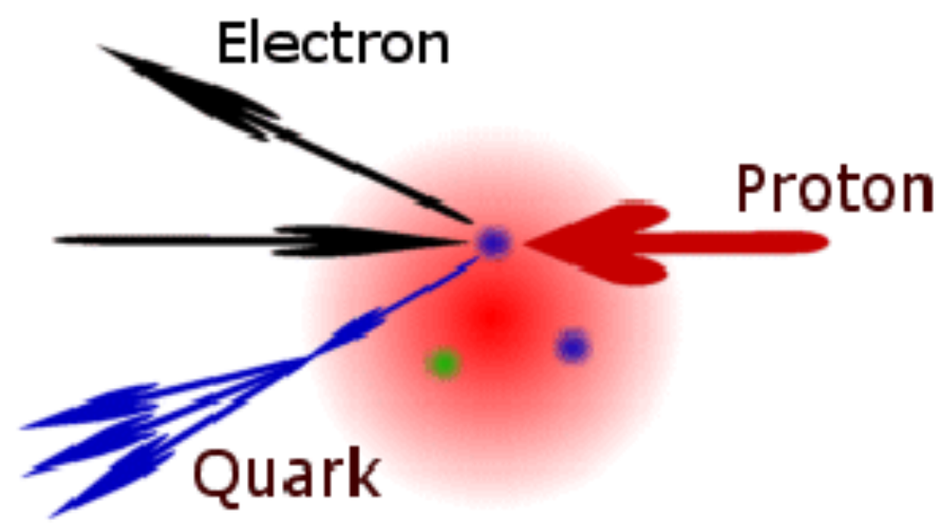
Image taken from the announcement of the winners of the 2004 Nobel Prize in Physics for

"the discovery of asymptotic freedom in the theory of the strong interaction"

- In **Quantum Chromodynamics**, which describes the strong nuclear force, the mathematics says that the closer quarks are to one another, the weaker the force becomes. Conversely the further the quarks move apart, the greater the force is.
- Critical to understanding hadronic matter (e.g. the protons at the **LHC**)
- ***How can this be tested ?***

Theory meets phenomenology

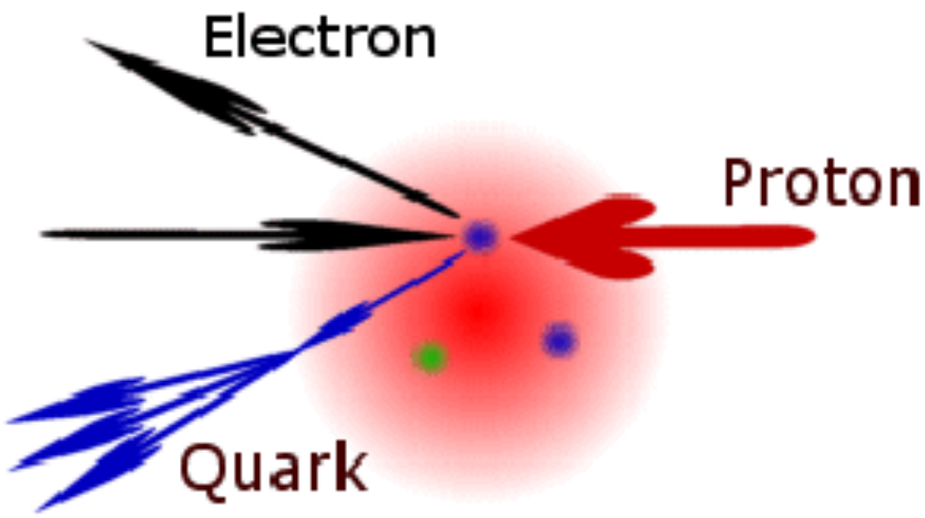


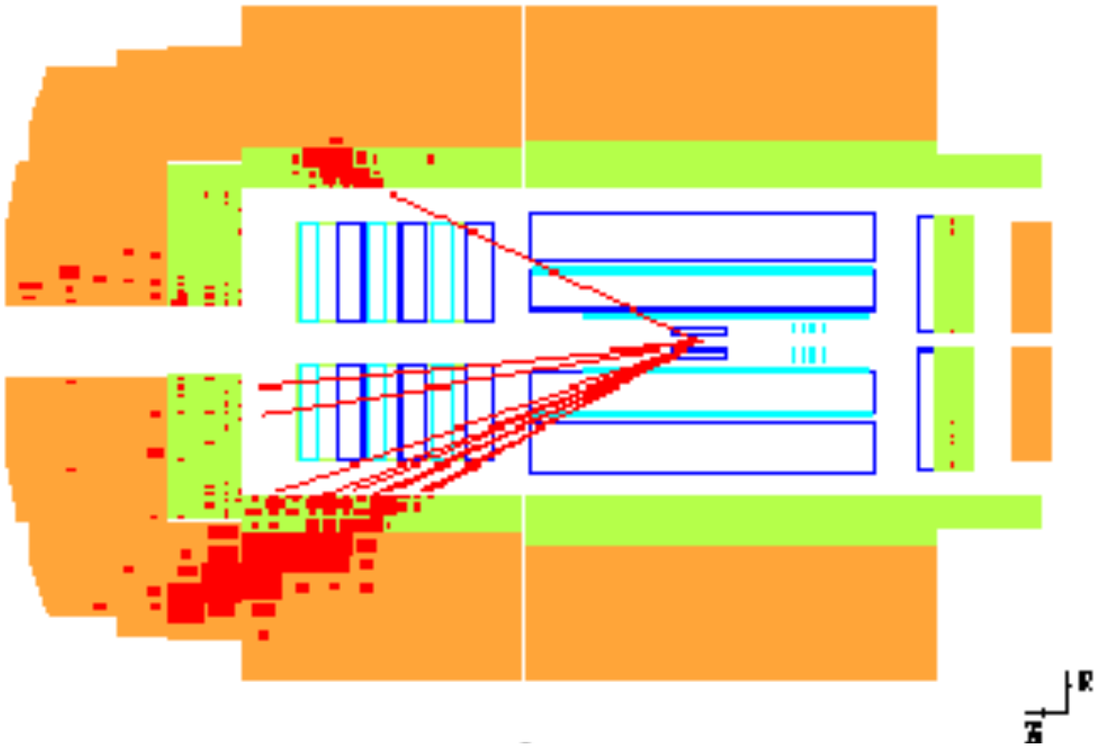


- Test in the old-fashioned particle physicist way
- ***Smash particles together!***

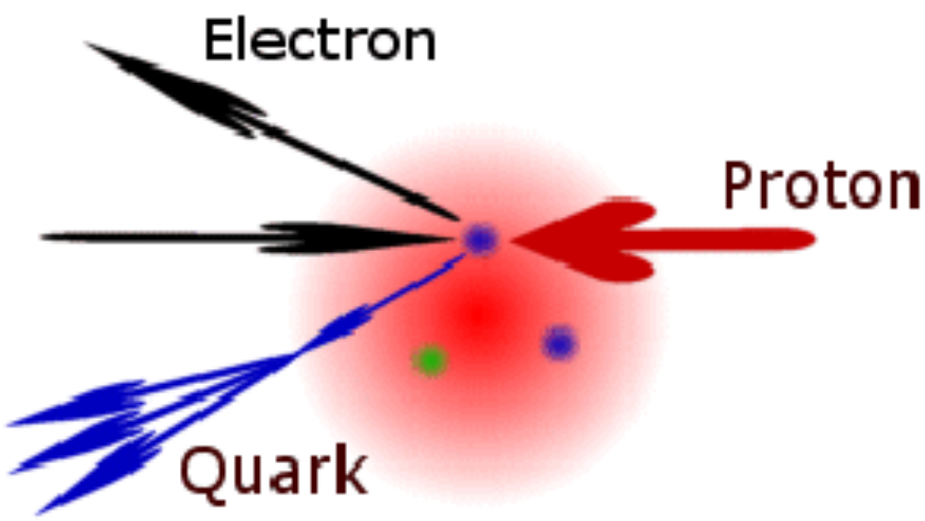
- Model the **proton** as a fuzzy bag of **quarks** and **gluons**, hit it with a high energy (*point-like*) **electron** - what will happen?

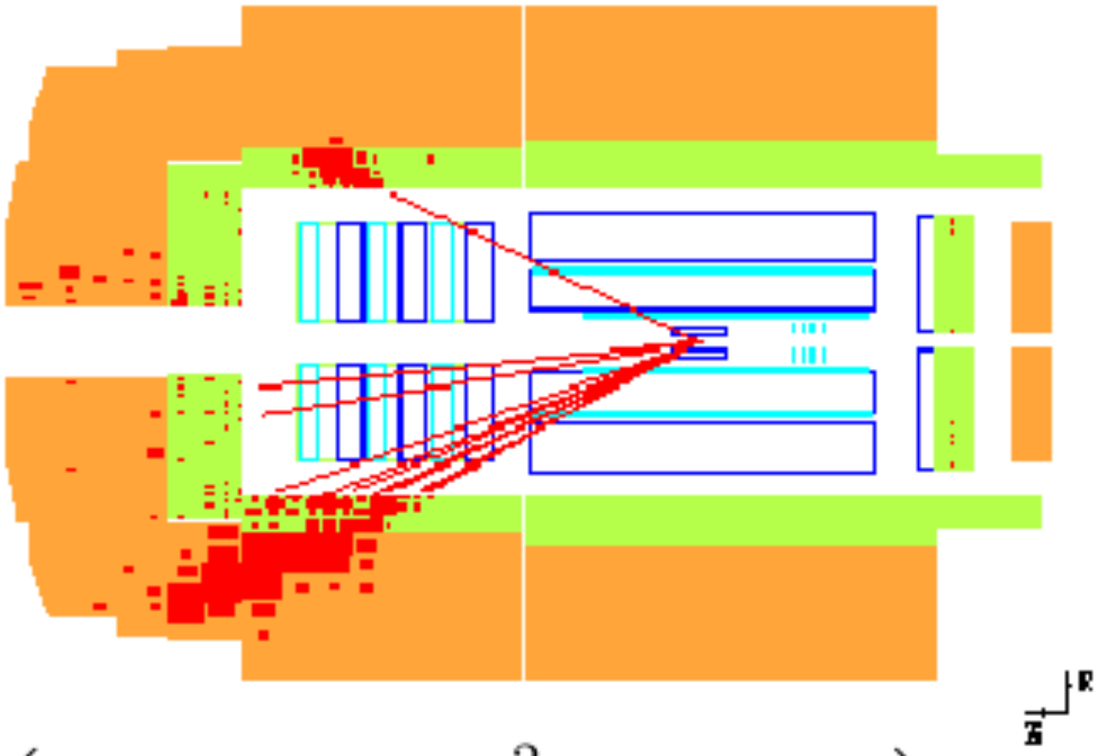
Phenomenology to experiment





Phenomenology to experiment

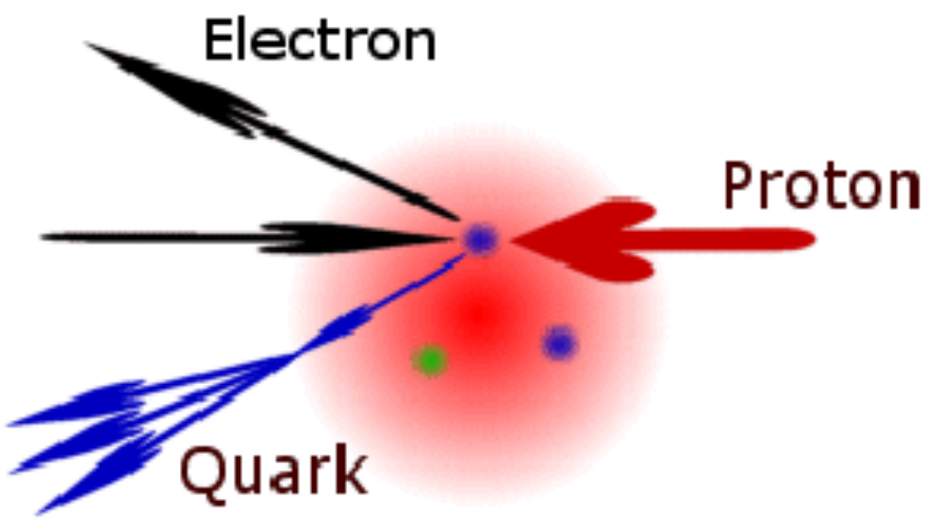


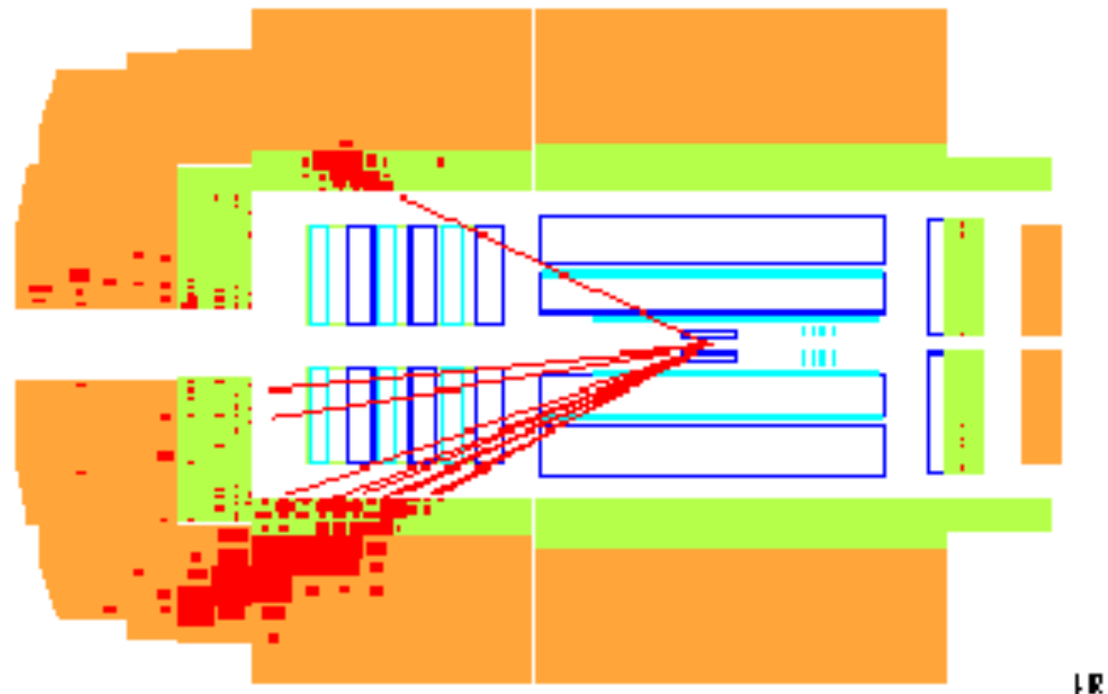


Measure:

$$\frac{\mathrm{d}^2\sigma_{NC}^{ep}}{\mathrm{d}x\mathrm{d}Q^2} = \frac{2\pi\alpha^2 Y_+}{xQ^4}\left(F_2(x,Q^2) - \frac{y^2}{Y_+}F_L(x,Q^2)\right)$$

Phenomenology to experiment





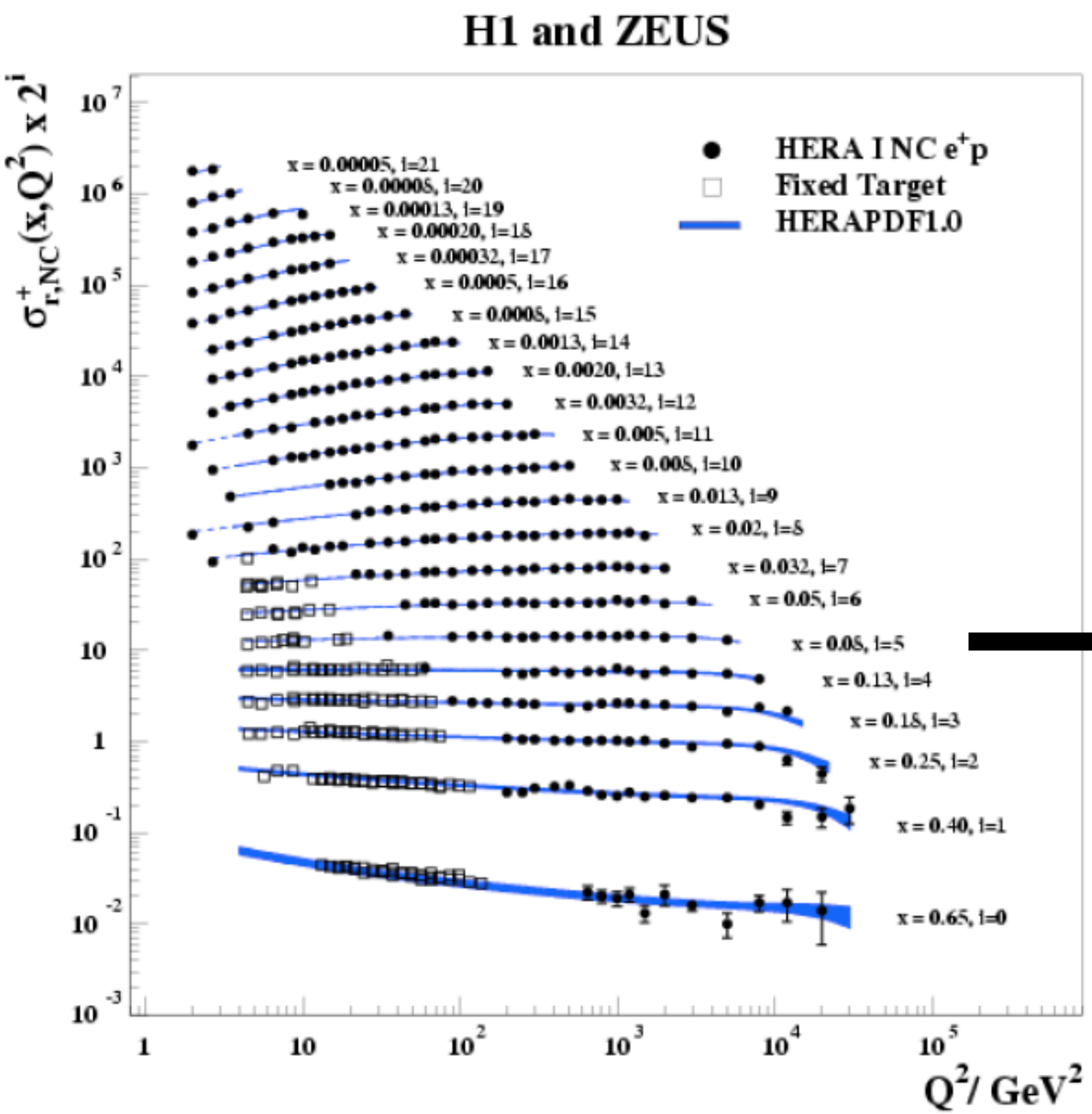
Measure:

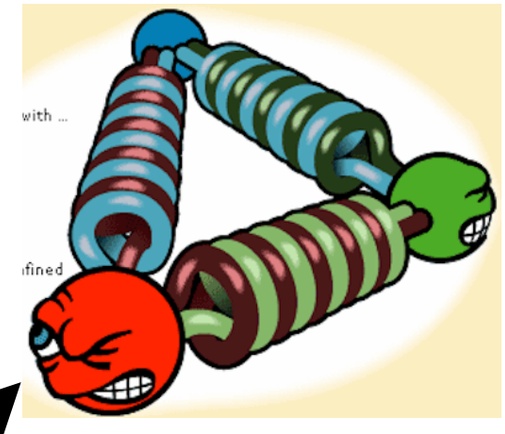
$$\frac{d^2\sigma^{ep}_{NC}}{dx\,dQ^2} = \frac{2\pi\alpha^2 Y_+}{xQ^4}\left(F_2(x,Q^2) - \frac{y^2}{Y_+}F_L(x,Q^2)\right)$$

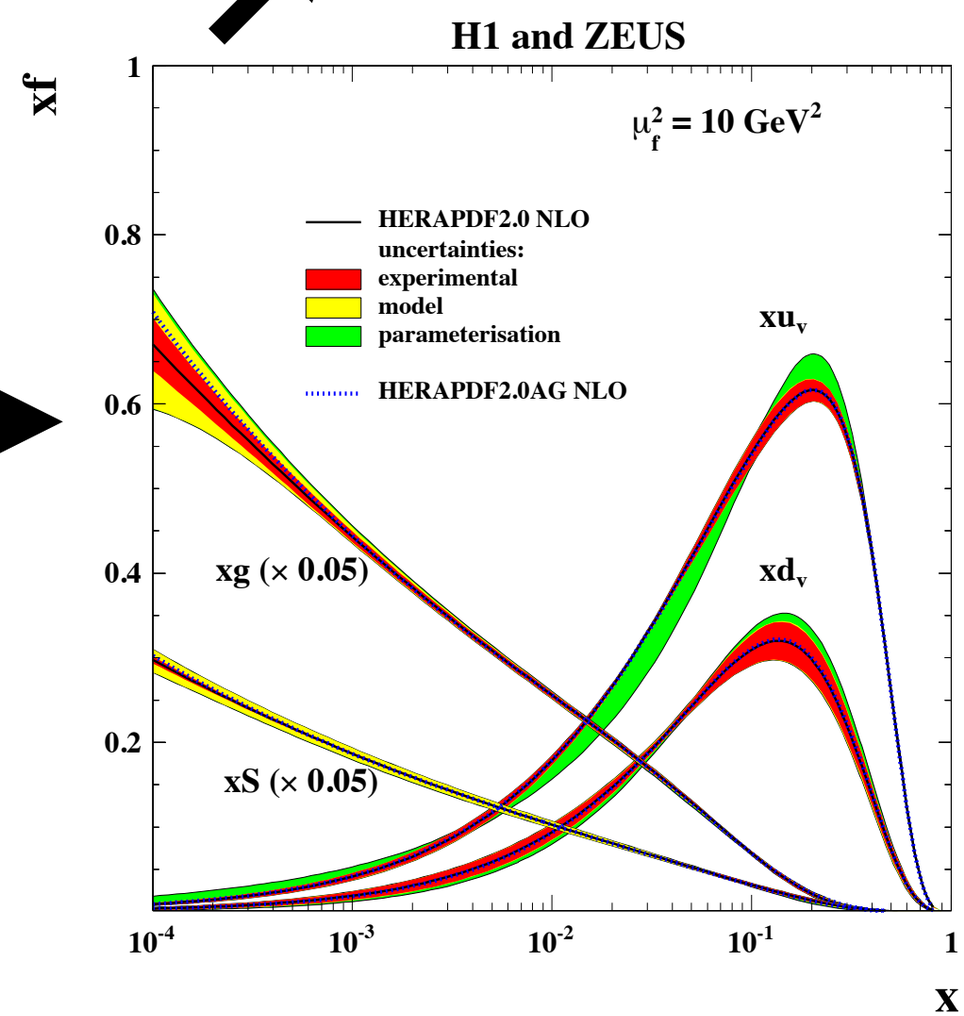
Extract:

- F_2 directly related to (PDFs) quark content: $F_2 \sim x\Sigma e^2(q+q)$
- $dF_2/d\ln Q^2$ (scaling violations) sensitive to gluon content
- F_L only non-zero in higher order QCD – independent access to gluon density and QCD dynamics

Extracting observables





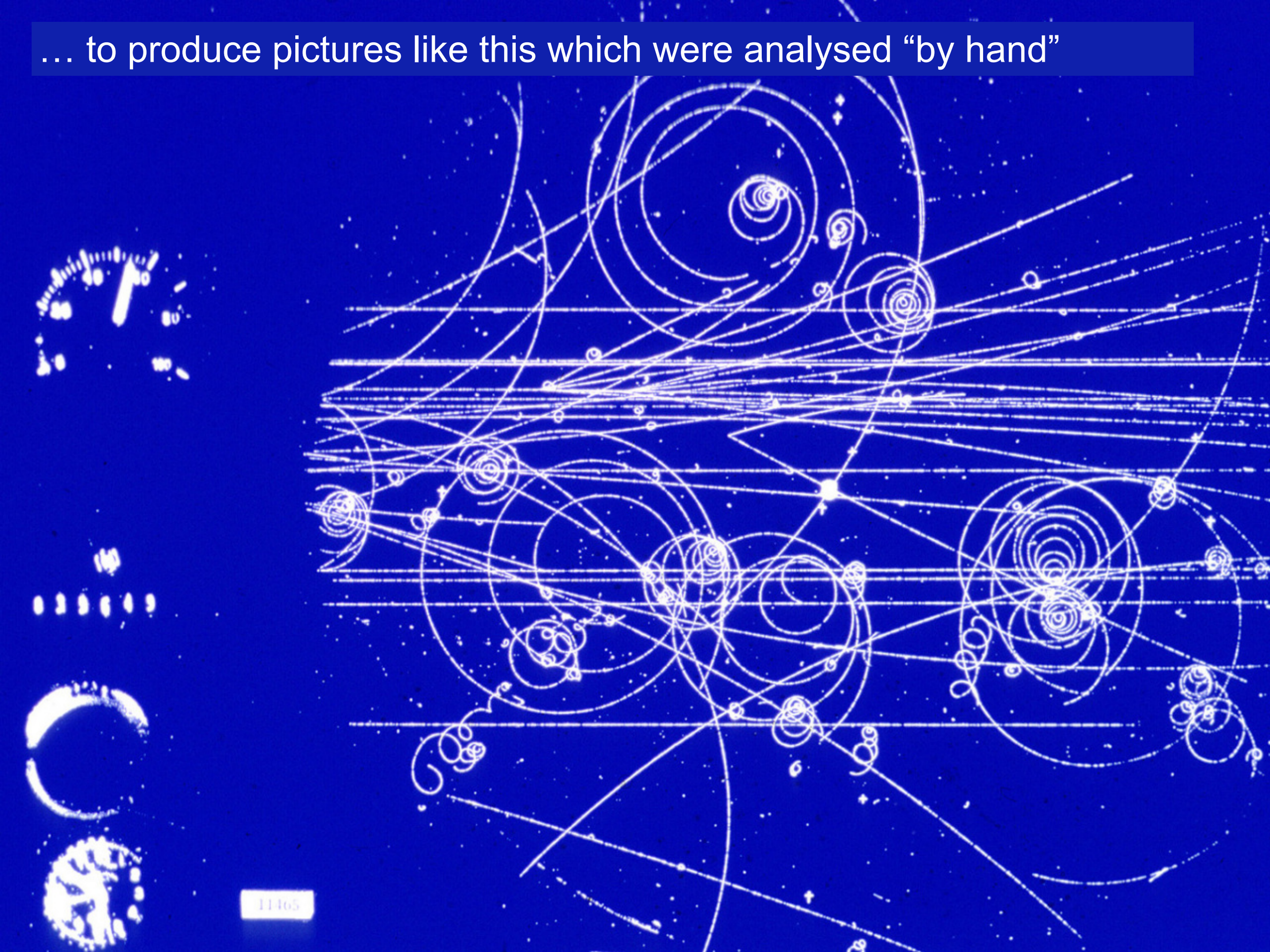


- Experimental confirmation of **QCD** allowing us to infer the **quark** and **gluon** structure of the **proton**

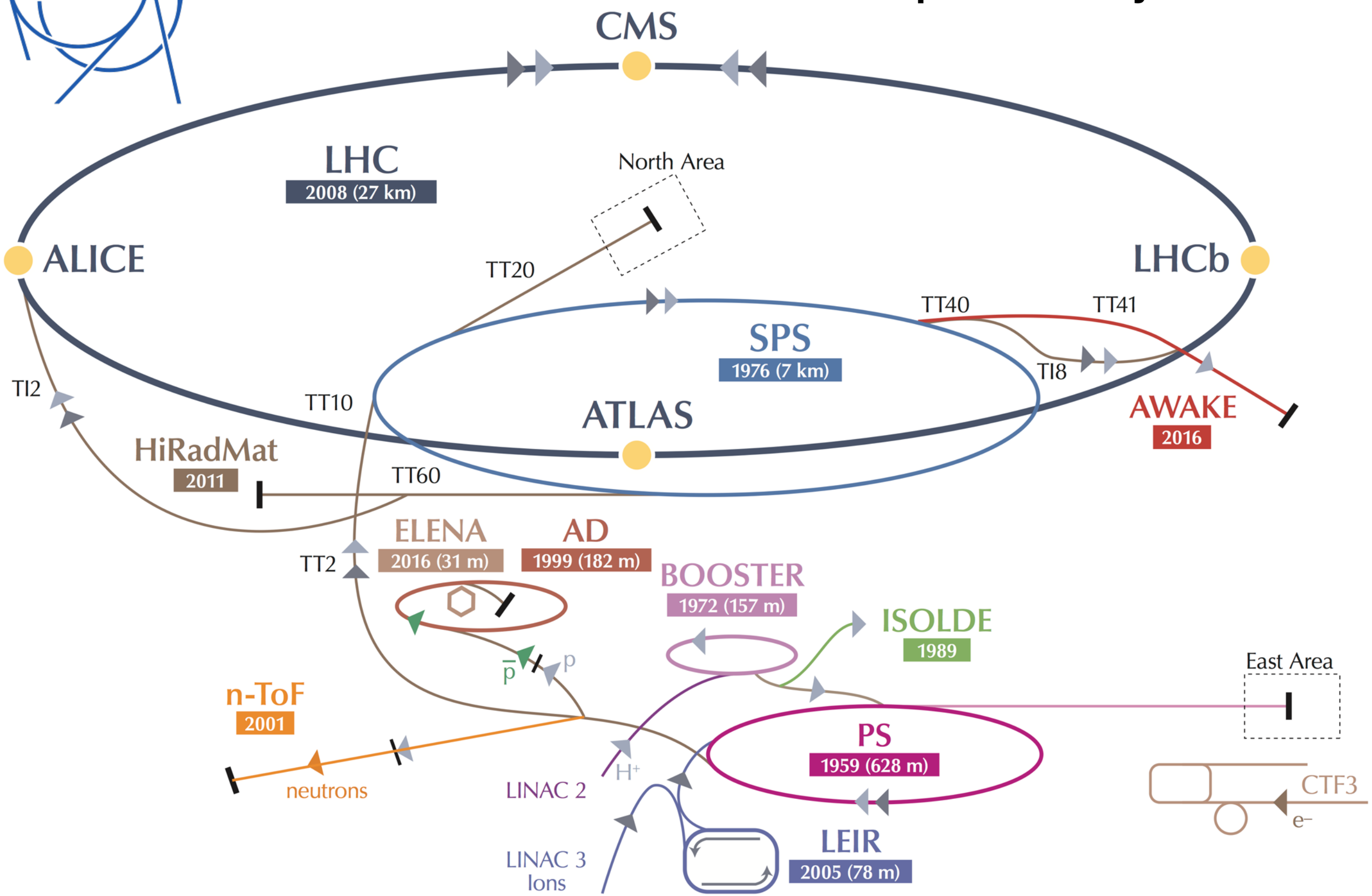
Particle physics detectors

- In the 1960s we used Bubble chambers, like the one pictured here on public display at CERN…

… to produce pictures like this which were analysed “by hand”



Today @ CERN we have huge rates of collisions so that we can produce very rare events



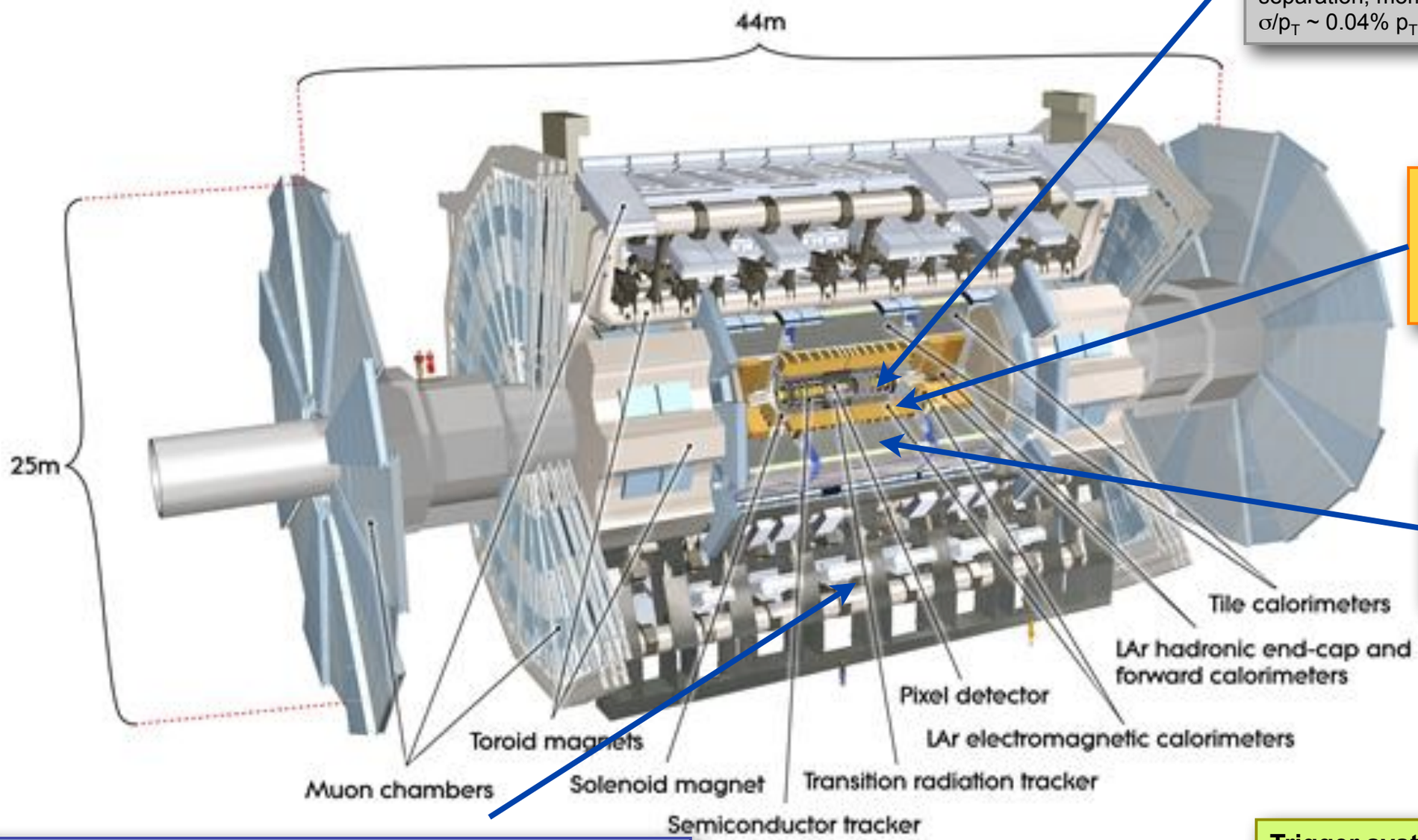
The ATLAS Detector @ LHC

L ~ 46 m, ∅ ~ 22 m, 7000 tons
~10^8 electronic channels

Inner Tracker (|η|<2.5, B=2T): Si Pixels, Si strips, Trans. Rad. Det. Precise tracking and vertexing, e/π separation, momentum resolution: $\sigma/p_T \sim 0.04\%\ p_T\,(\mathrm{GeV}) \oplus 1.5\%$

EM calorimeter: Pb-LAr Accordion, e/γ trigger, id. and meas., energy res.: $\sigma/E \sim 10\%/\sqrt{E} \oplus 0.7\%$

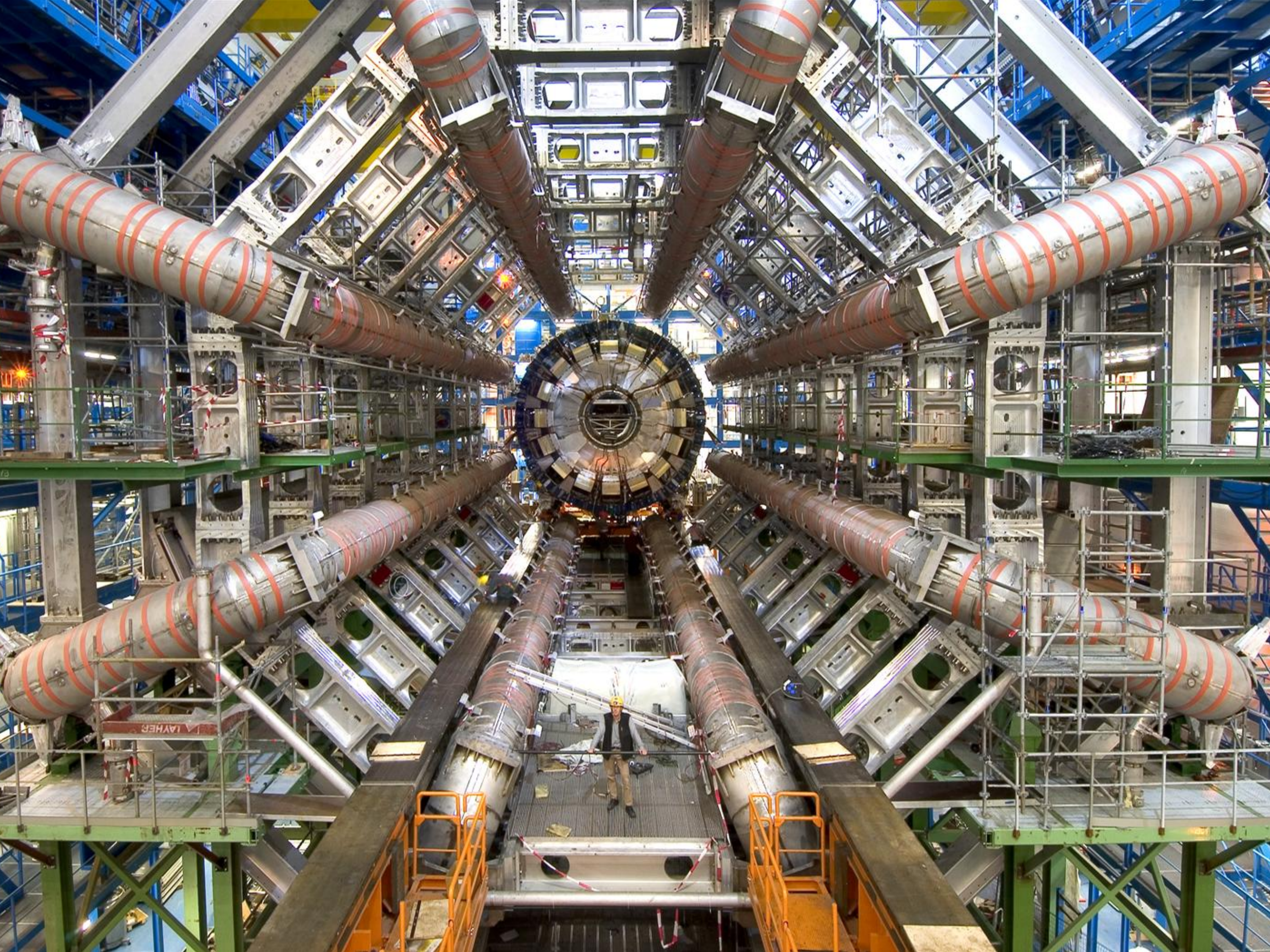
HAD calorimetry (|η|<5): Fe/scintillator Tiles (cen), Cu/W-LAr (fwd). trigger and meas. of jets and $E_{T,miss}$, energy res.: $\sigma/E \sim 50\%/\sqrt{E} \oplus 3\%$

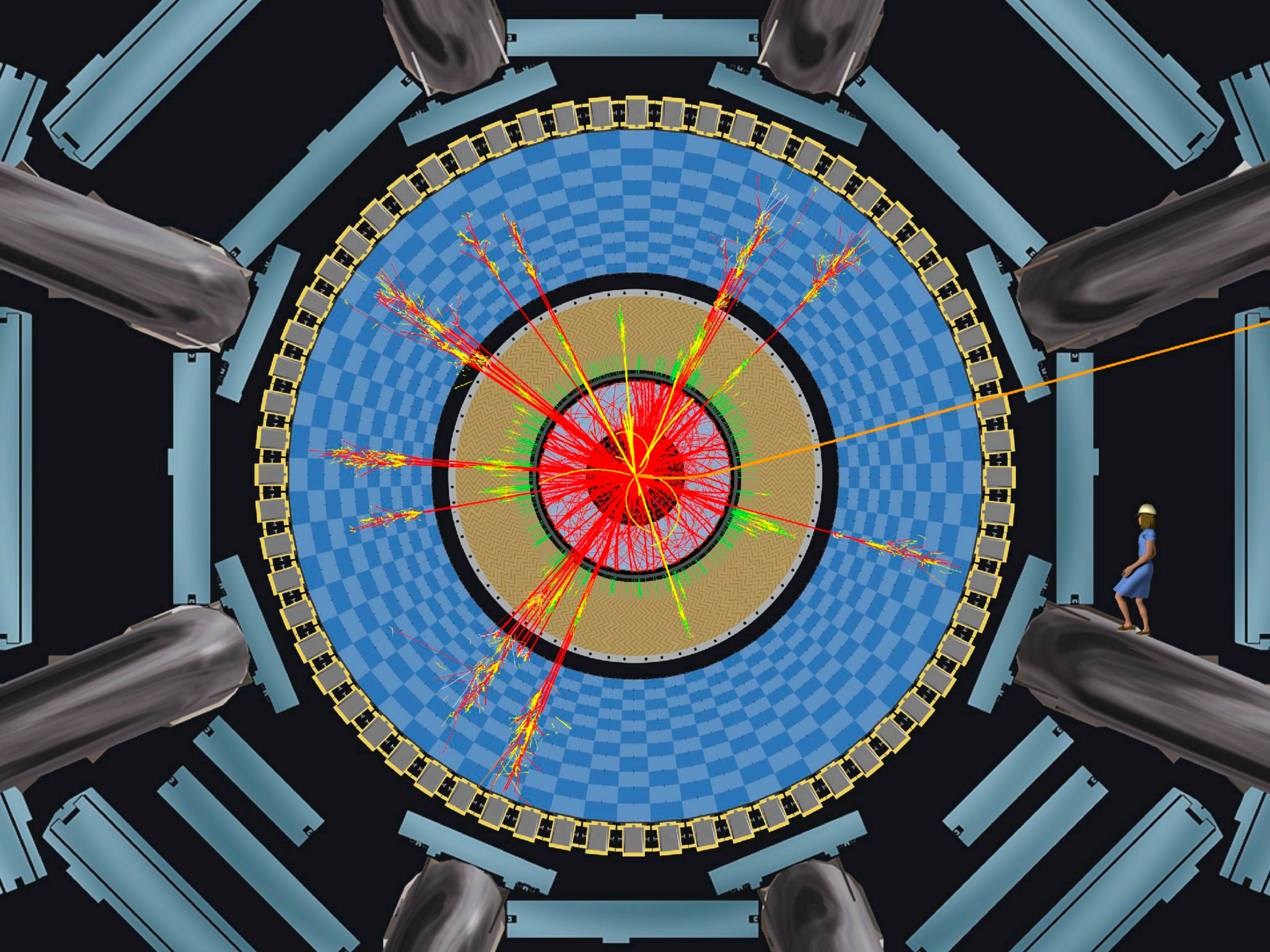


Muon Spectrometer: air-core toroids with gas-based muon chambers. trigger and meas. with momentum resolution < 10% up to $E_\mu \sim 1$ TeV

Trigger system: 3-levels reducing the IA rate from 40 MHz to ~200 Hz

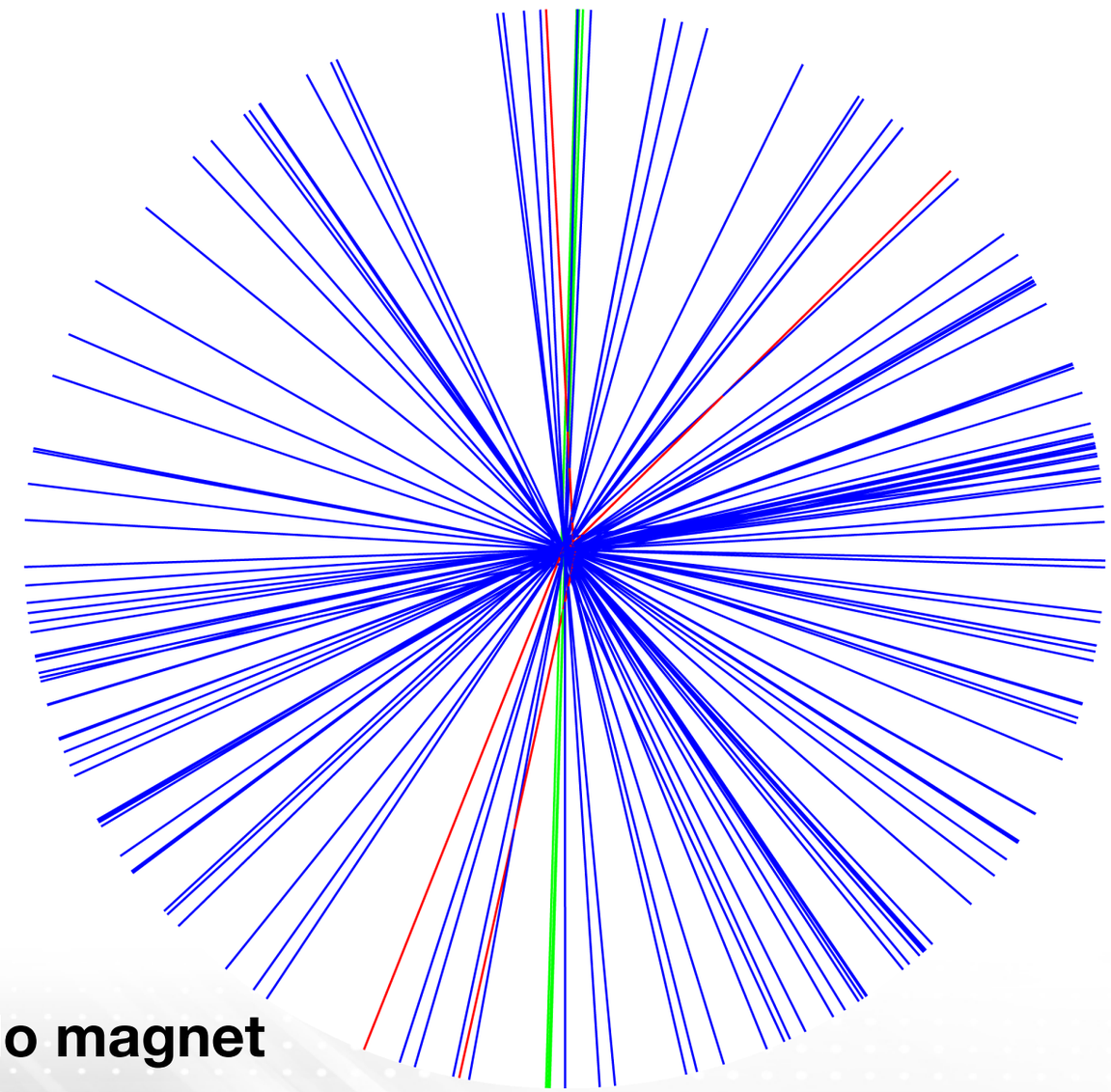
Millions of detector readout channels read out to reconstruct one "event"





Before the detector, came the simulation

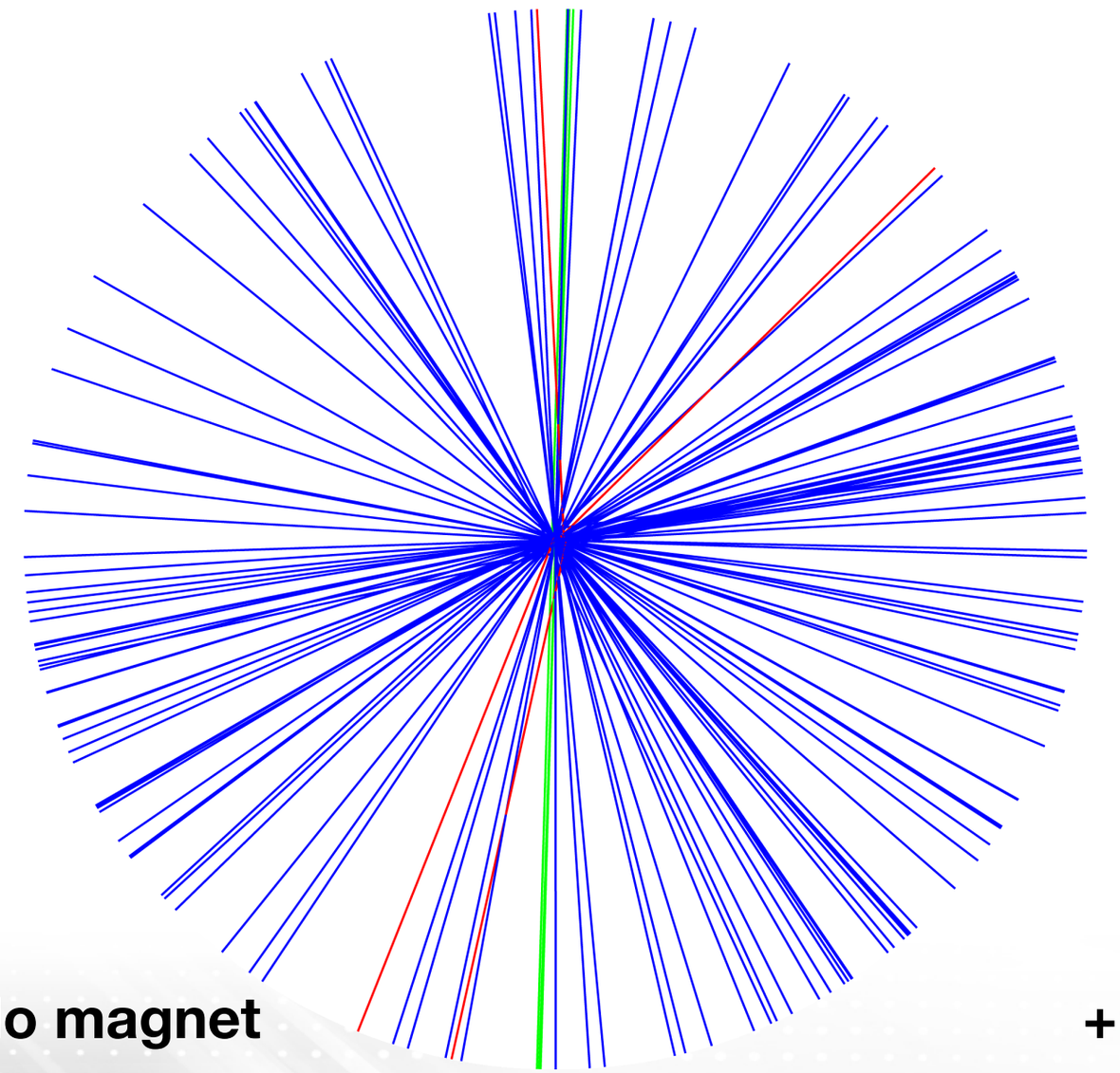
- When designing detectors, we ***simulate detector response*** to physics of interest
- Adding a ***solenoid magnet*** makes it possible to measure momentum (and charge) in our tracker by measuring curvature in the transverse plane
- Interesting physics is often at ***high momentum***, e.g. four high momentum muon tracks here



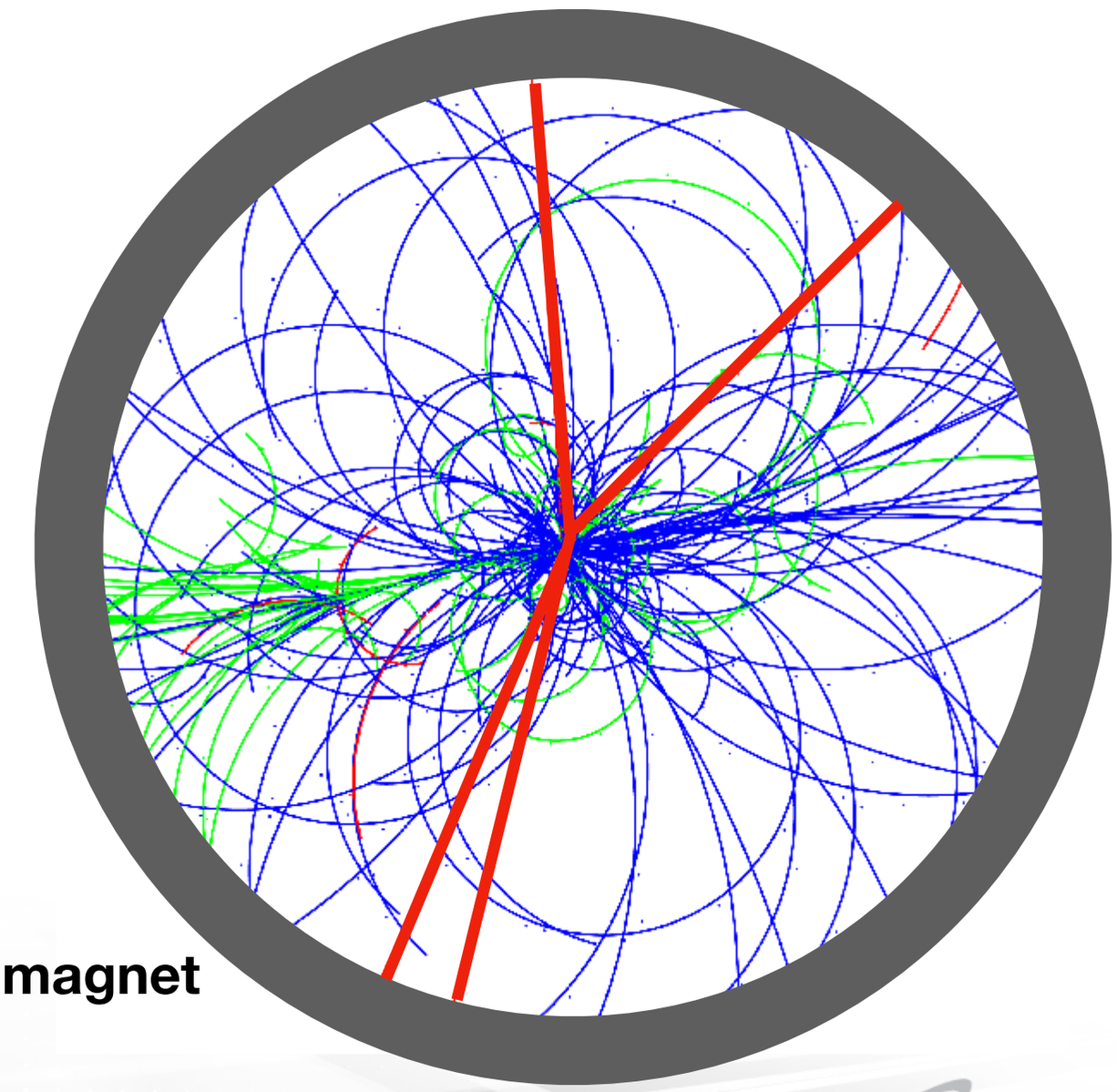
No magnet

Before the detector, came the simulation

- When designing detectors, we ***simulate detector response*** to physics of interest
- Adding a ***solenoid magnet*** makes it possible to measure momentum (and charge) in our tracker by measuring curvature in the transverse plane
- Interesting physics is often at ***high momentum***, e.g. four high momentum muon tracks here

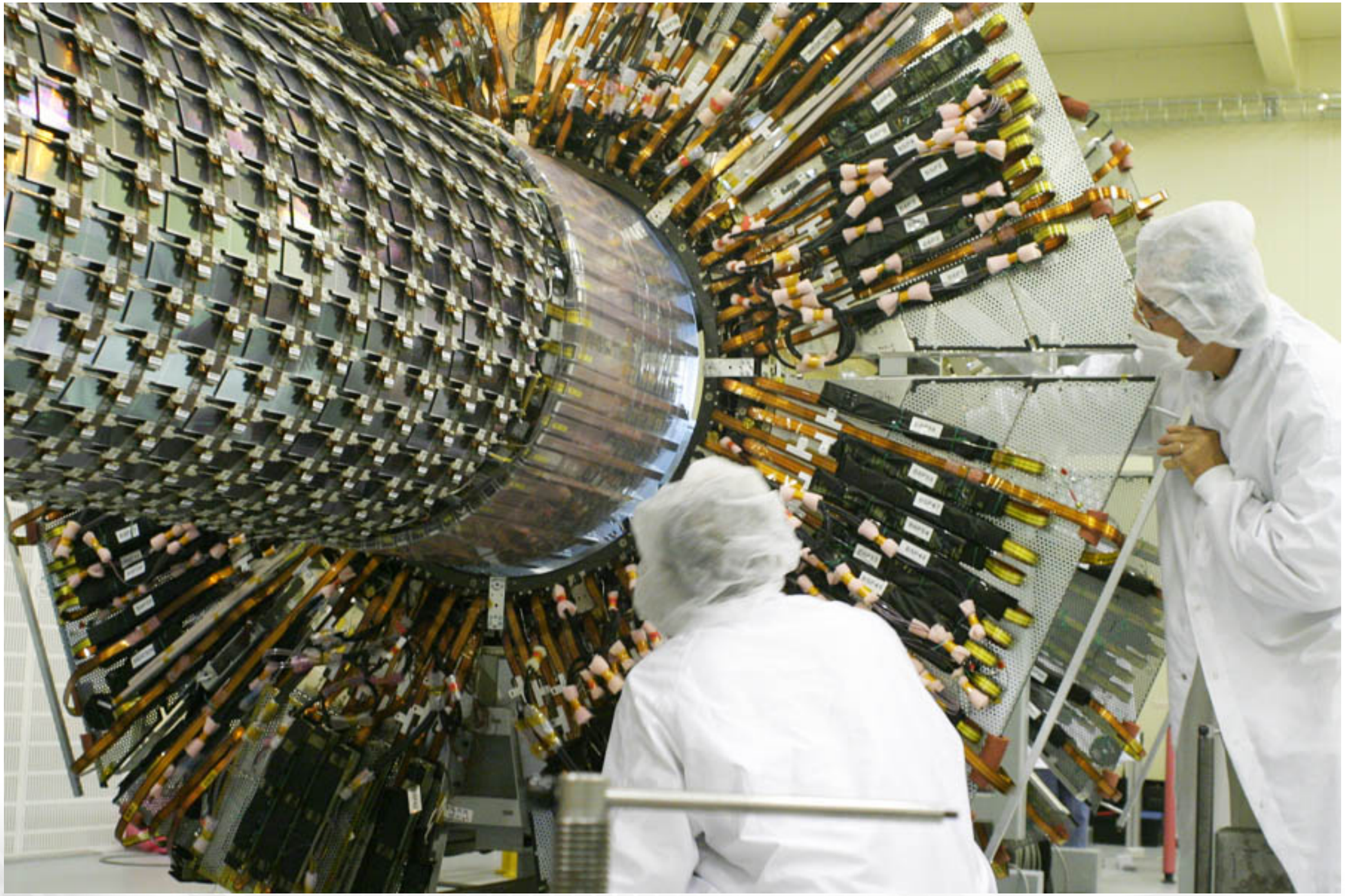


No magnet



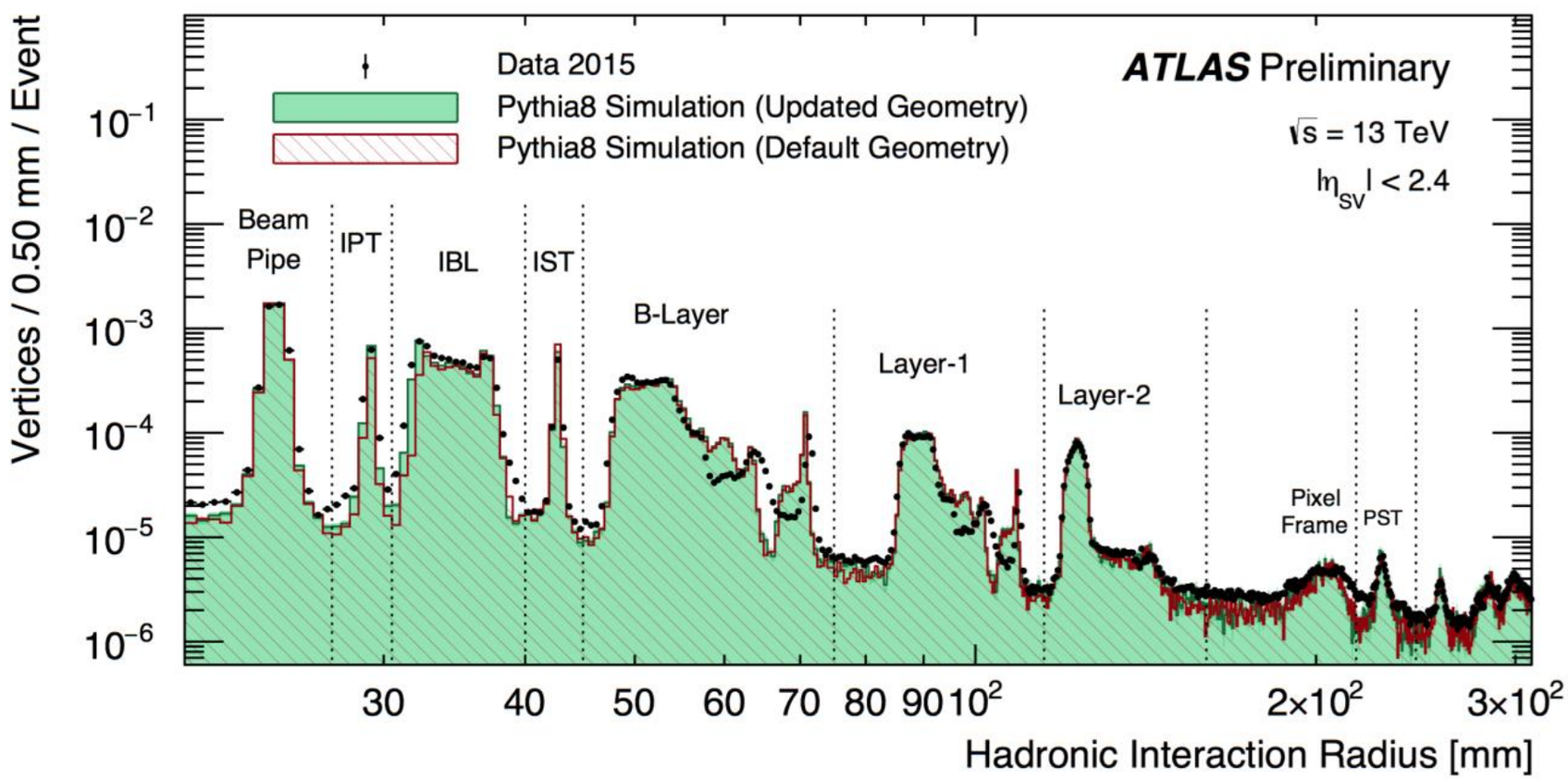
+ magnet

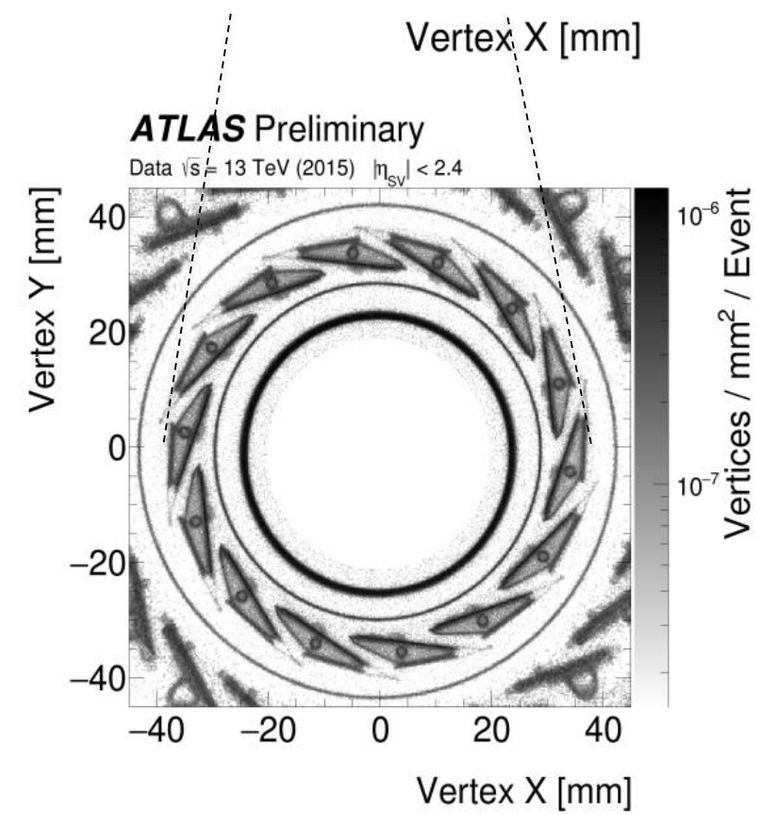
Building detectors



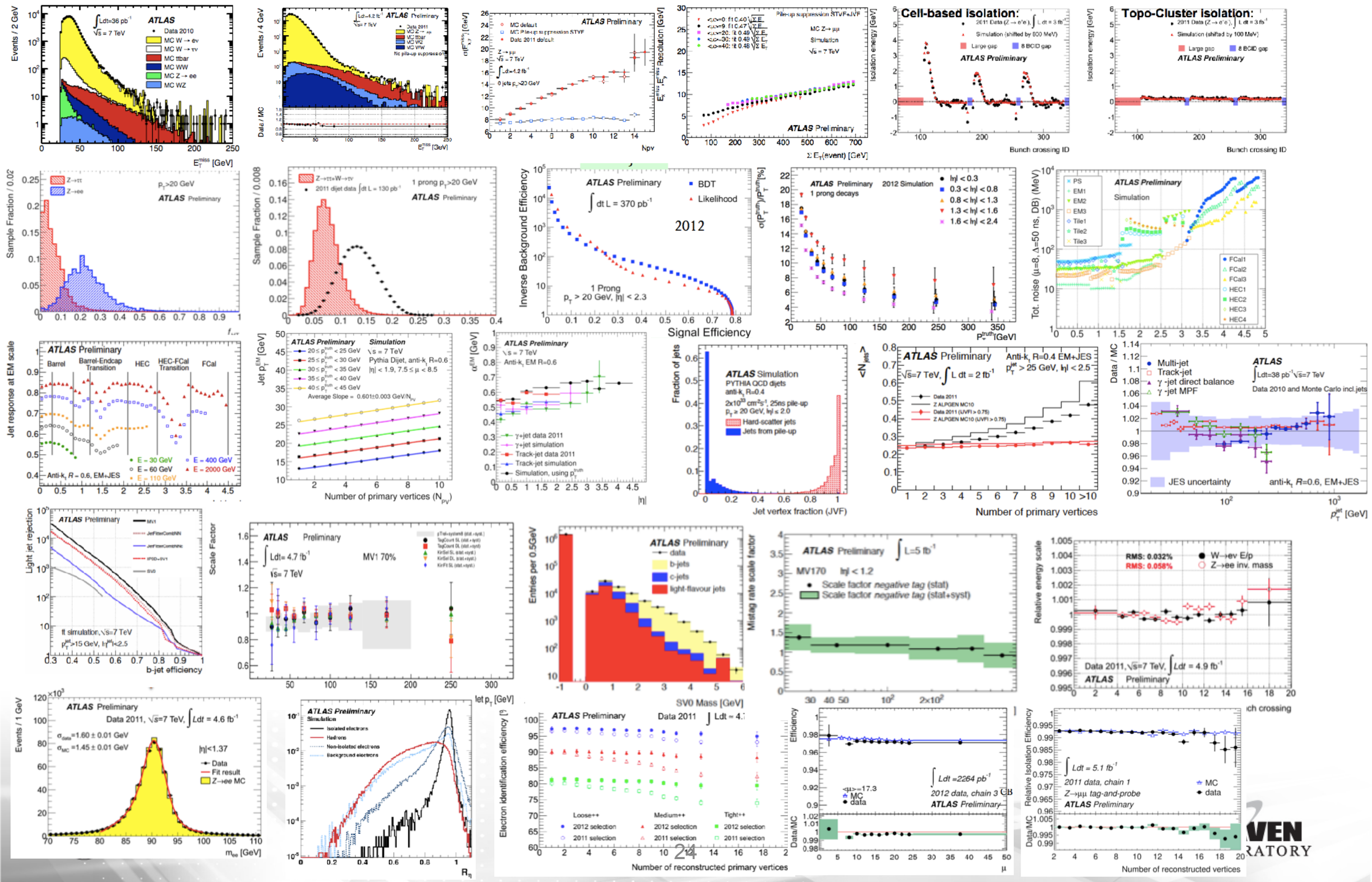
Simulation and understanding detectors

- We use **simulations** to model the detector as **accurately** and **precisely** as possible
- We then ***test*** that our simulations are accurate ***using real data***
- We make corrections to our simulations if necessary
  - *A common problem - missing some material in the simulation*
- Once we can rely on an ***accurate model*** of our detector, we can rely on the simulation modelling the real detector's response to particles
  - This allows us to ***correct the data for detector response***

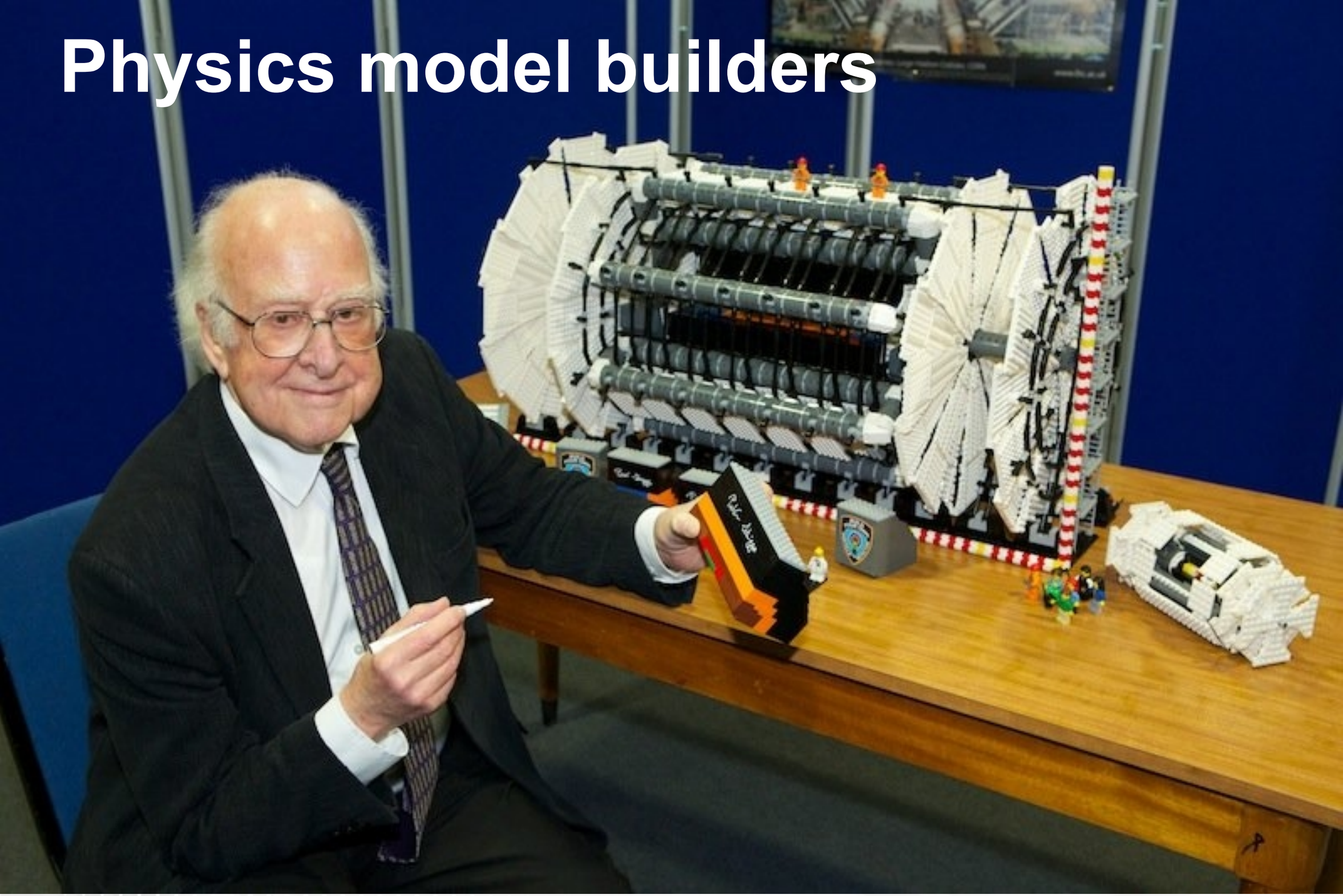




Ingredients to the ATLAS physics program



Physics model builders



Physics event generators

Courtesy: Z. Marshall

Simulation chain

How much processing time needed for each step?

Event Generation
simulate the physics process.

From < 1s to a few hours / event.

Detector Simulation
simulate the interaction of the particles with the detector material.

From 1 to 10min / event

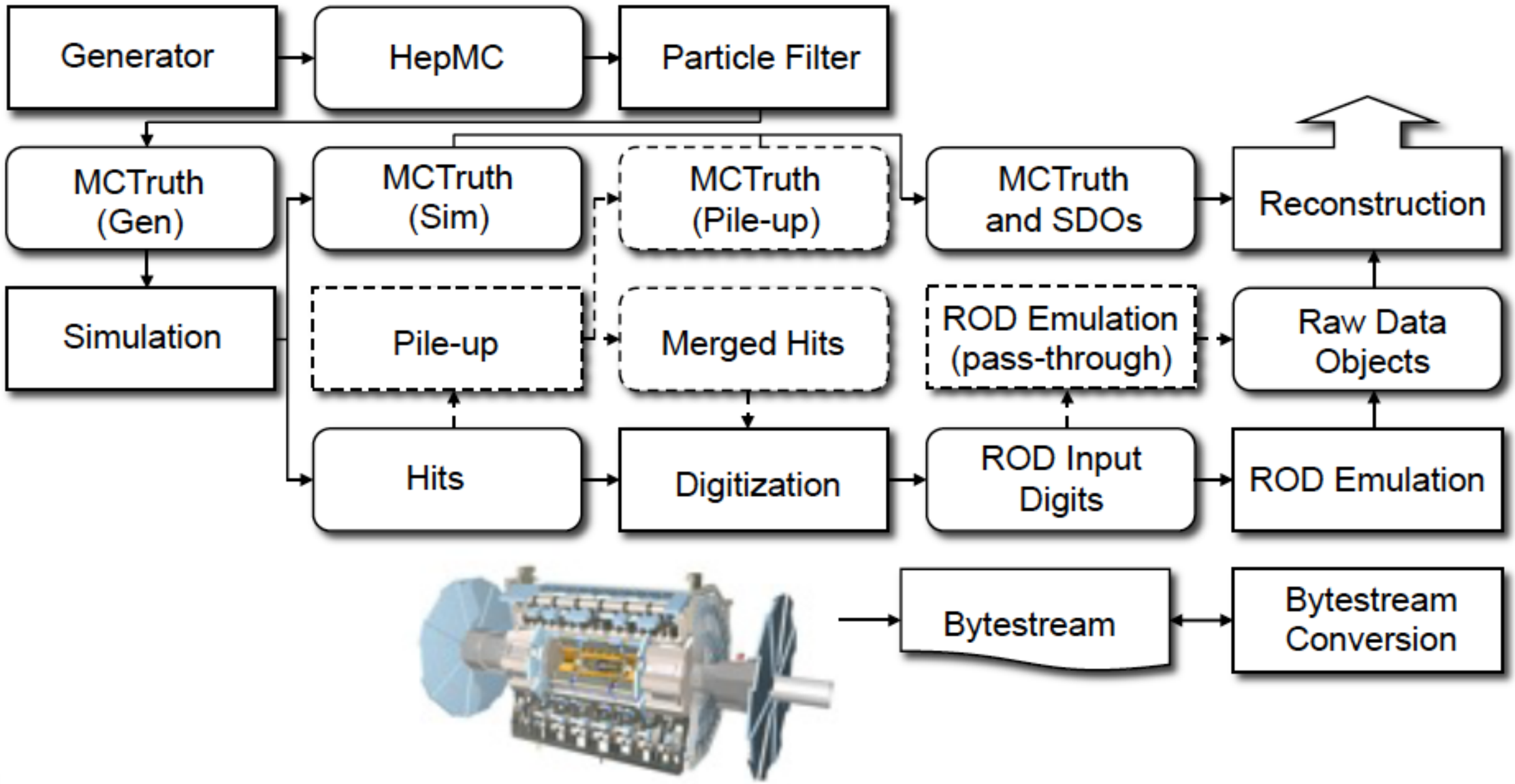
Digitization
Translate interactions with detector into realistic signals.

From 5 to 60s / event

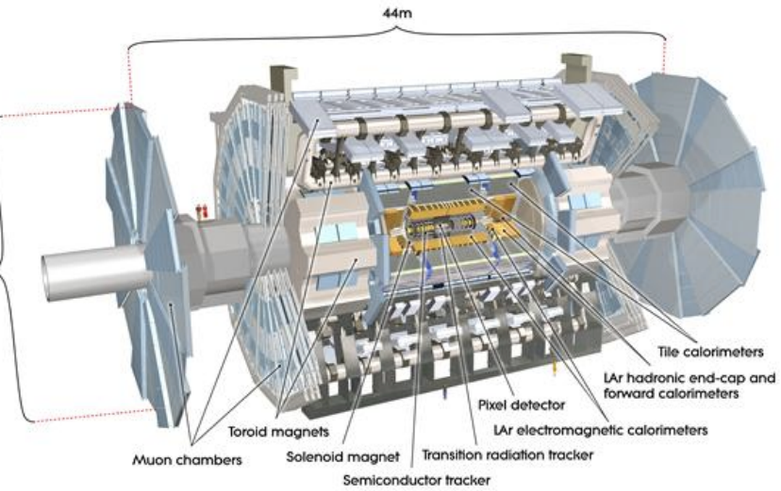
Reconstruction
Go from signals back to particles, as for real data.

Require simulations for signal and background processes with better statistics than data

Complete ATLAS simulation workflow



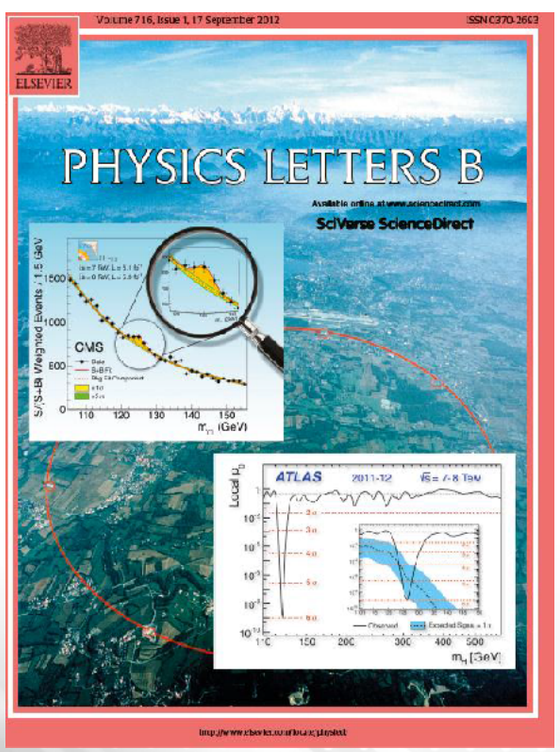
Data's journey



Trigger/DAQ

Data Preparation

Distributed computing

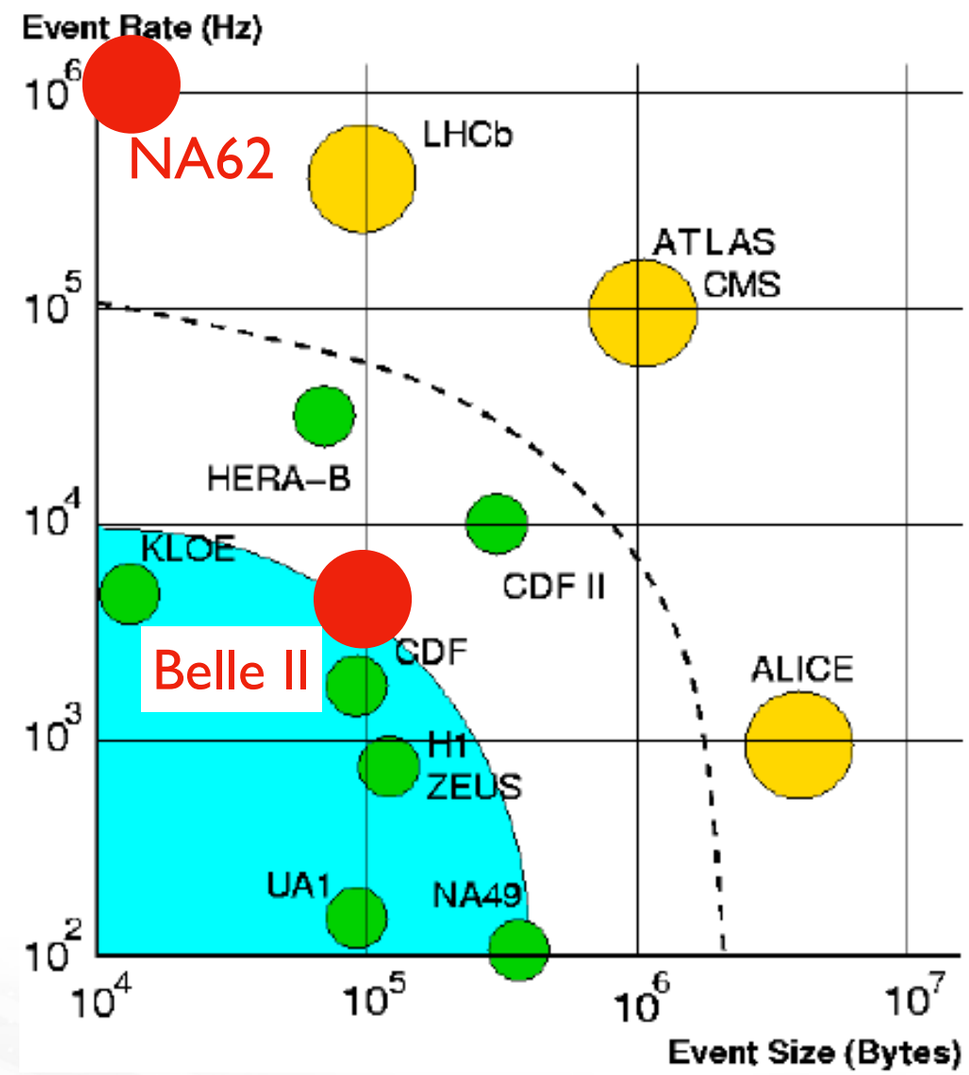


Raw data throughput

Plot modified from:
"GridPP: development of the UK computing Grid for particle physics."

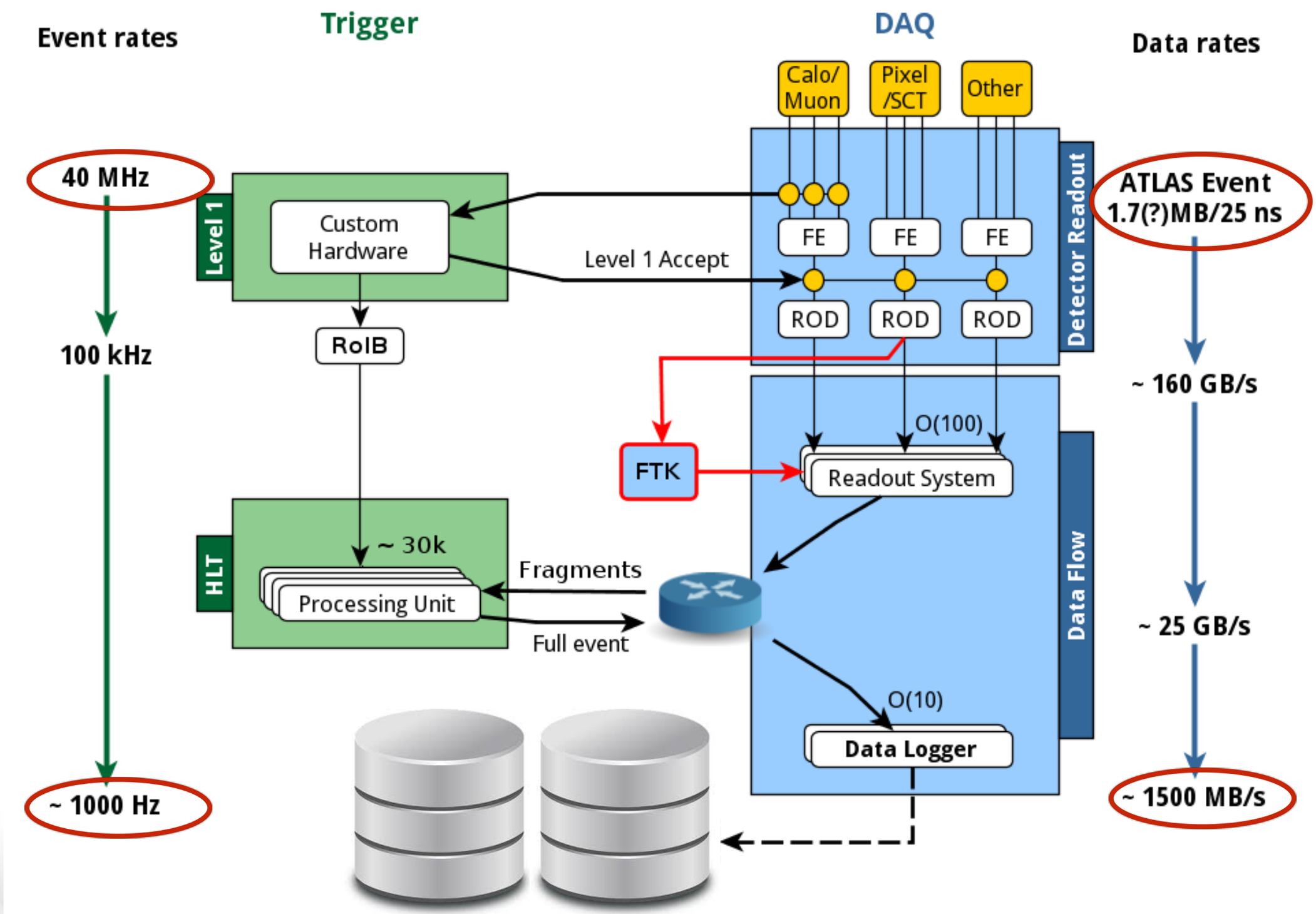
- H1
  - Proton structure and QCD
  - small event sizes and rates
- ATLAS
  - Higgs, searches for new physics
  - big event sizes and rates
- NA62
  - Ultra-rare kaon decays
  - huge rates of small size events
- Belle II
  - Ultra-rare B decays
  - modest event sizes and rates

DAQ throughput = event rate * event size



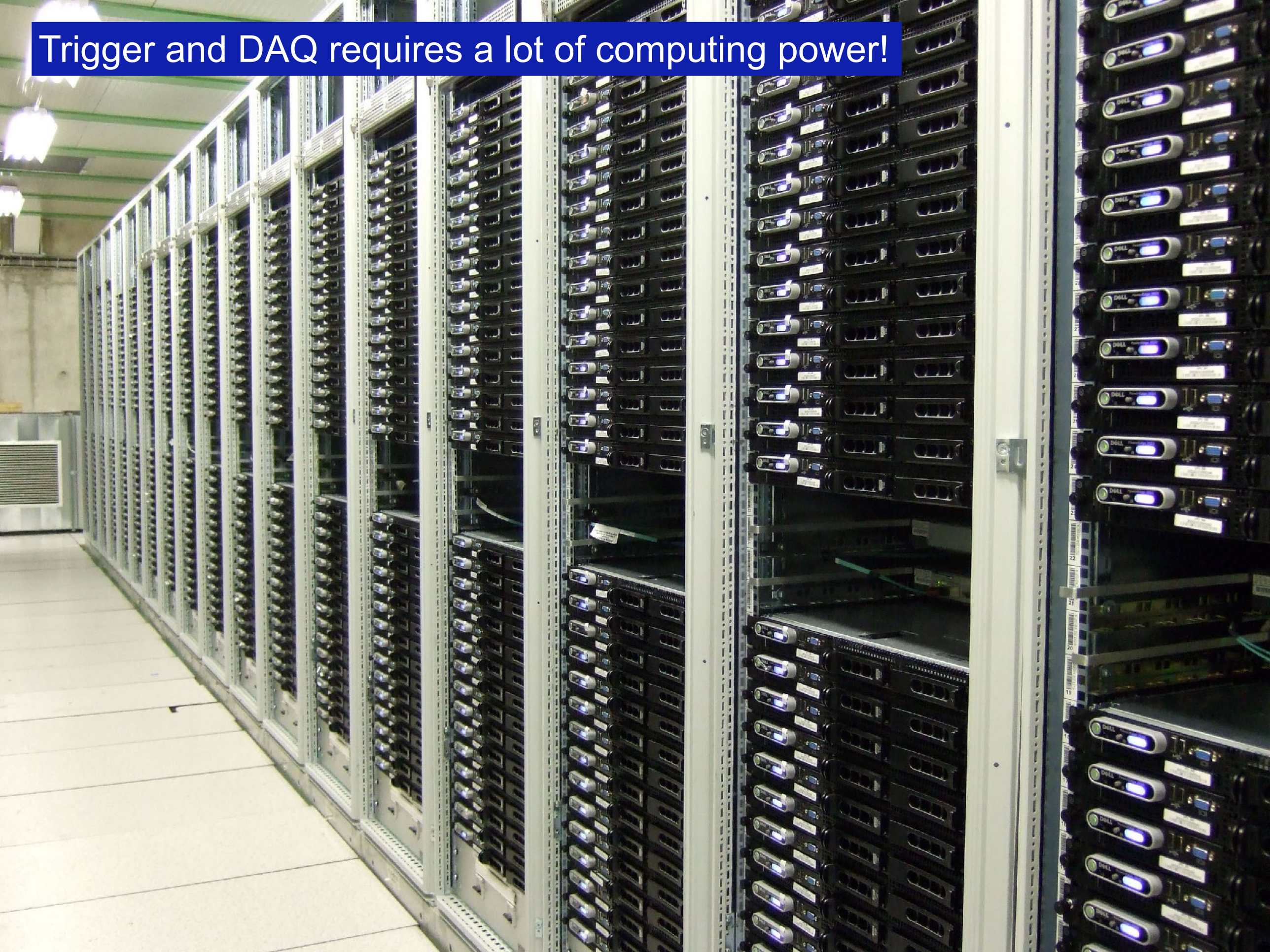
- Triggers are critical to reduce the amount of data we record and analyse later

The Atlas Trigger and DAQ

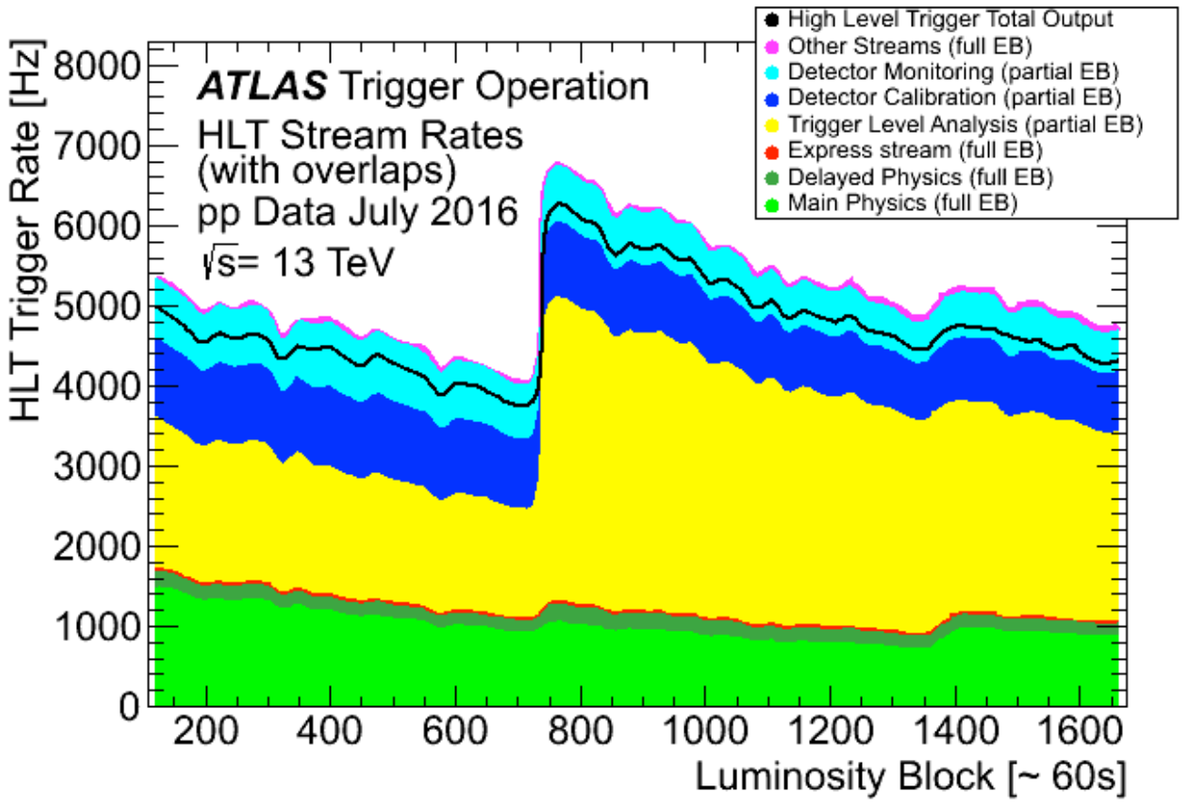


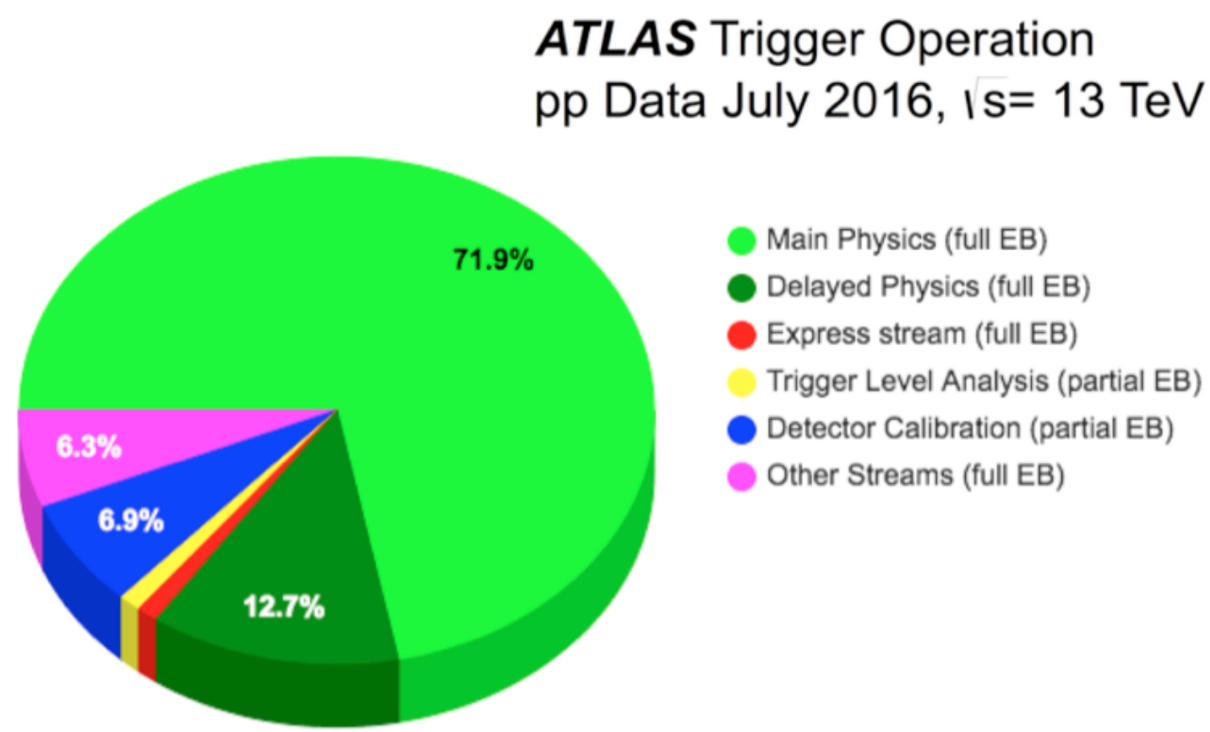
Permanent Storage

Trigger and DAQ requires a lot of computing power!



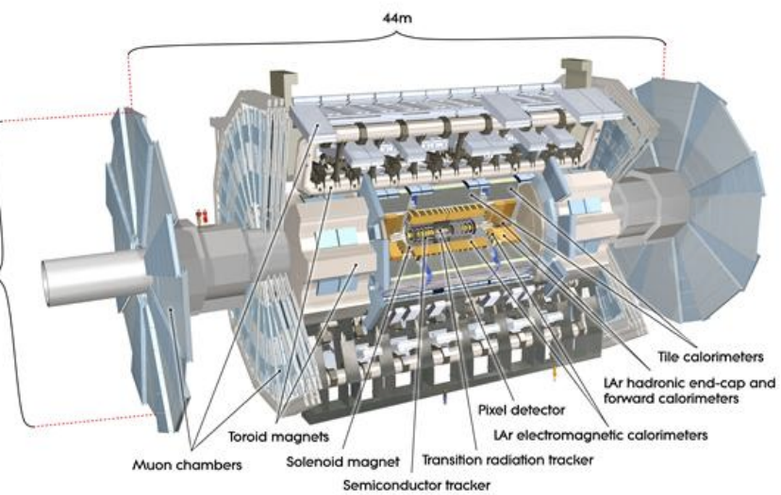
Trigger streams

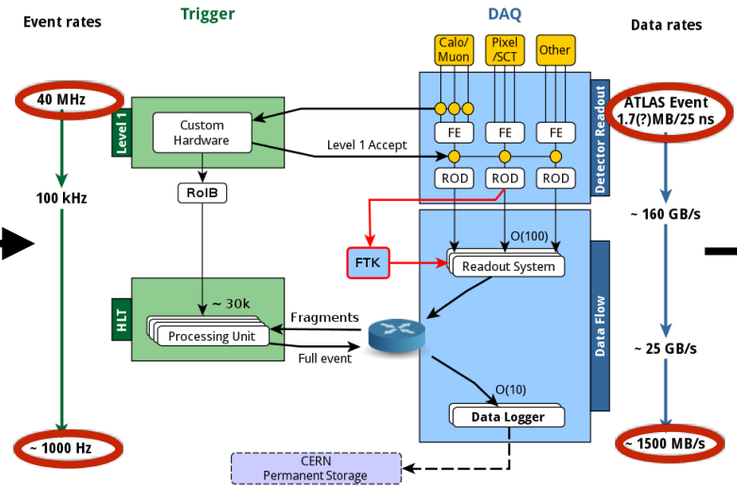




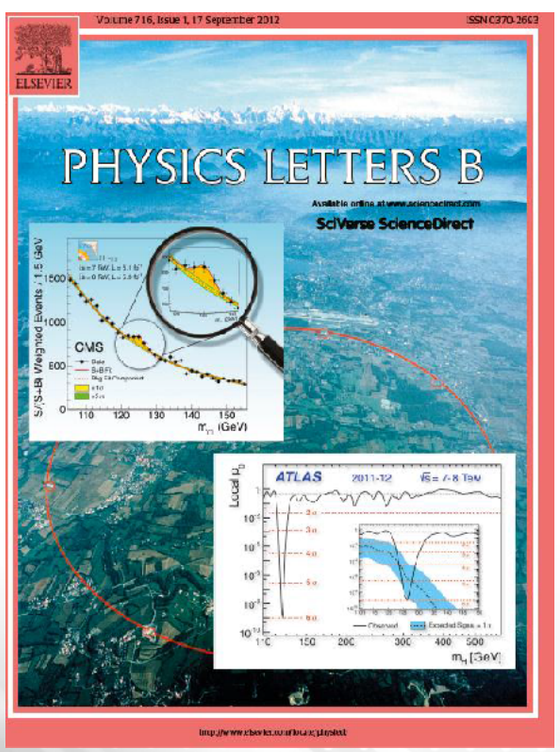
- We know in advance that in addition to the main physics data, we will need some **dedicated data** for:
  - *performing calibrations*
  - *assessing data quality*
- Writing **dedicated output streams** (written to different physical files) provides people with just the data they need

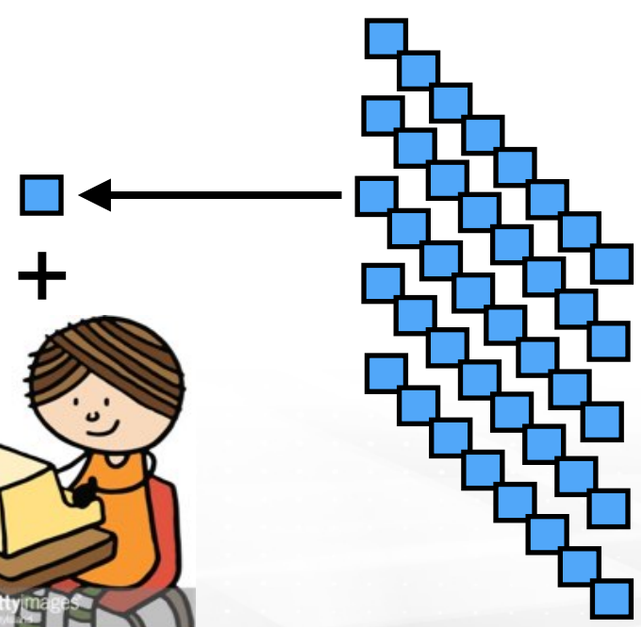
Data's journey





Data Preparation





Distributed computing

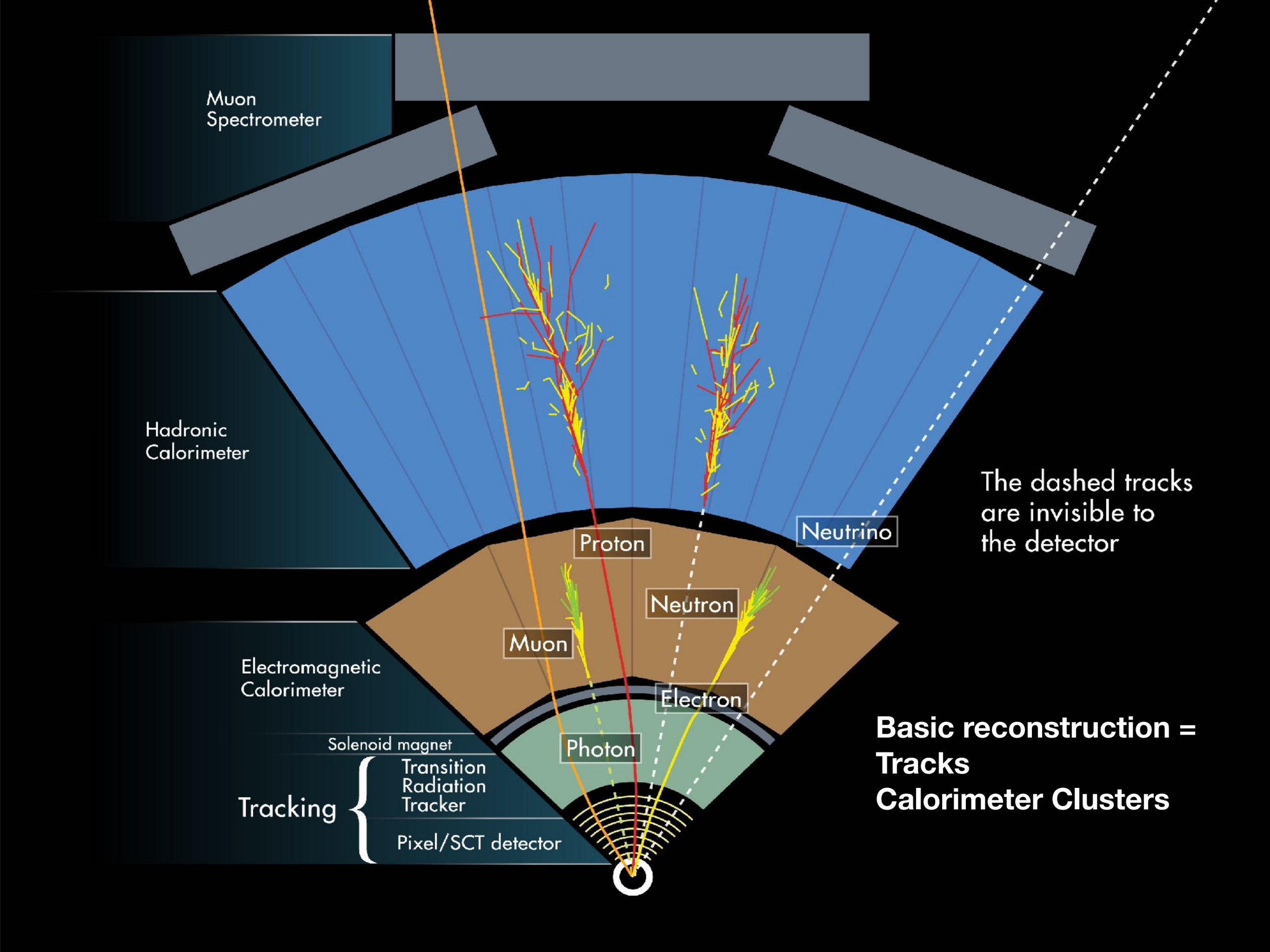
Data Preparation

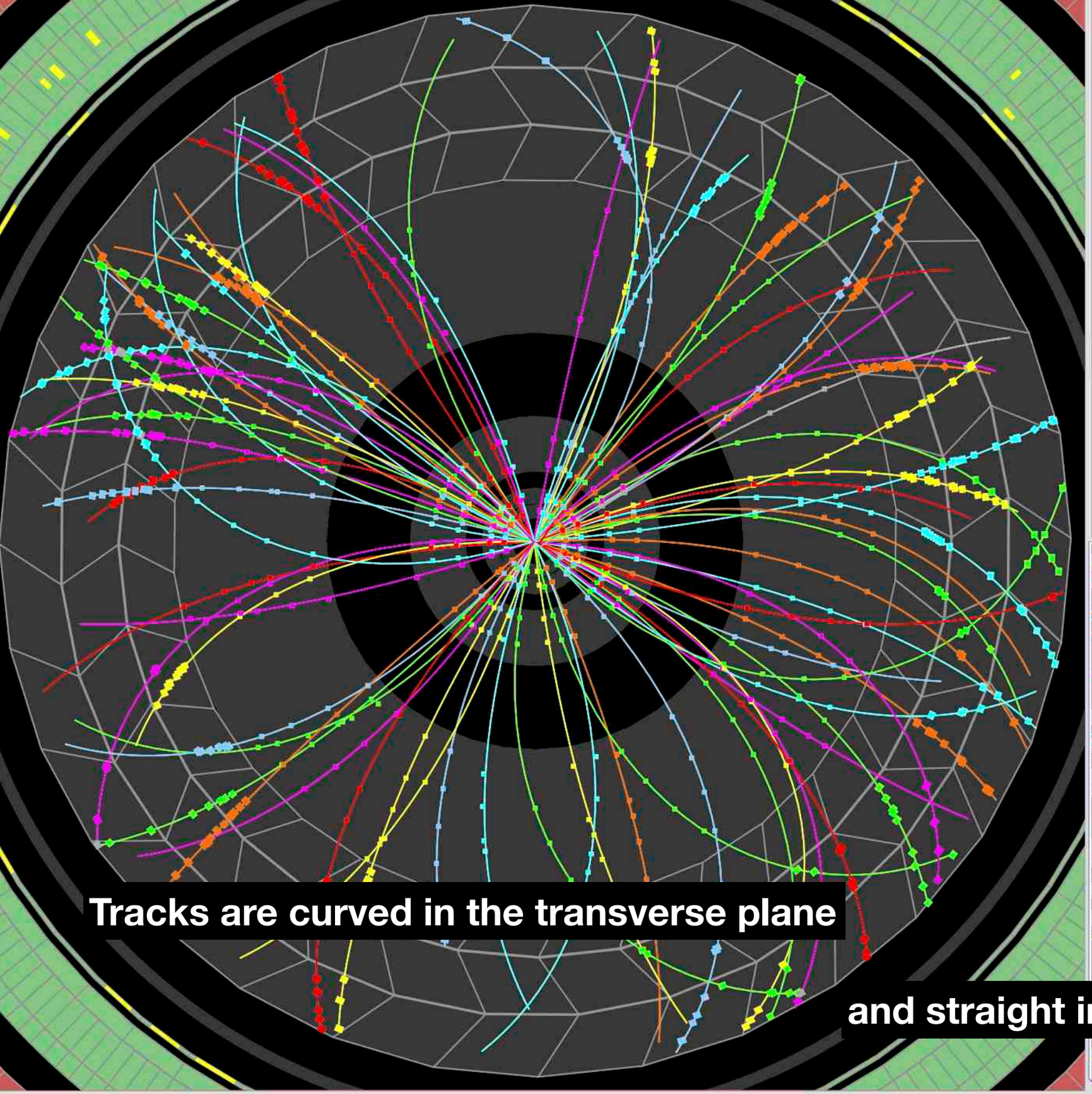
- Three major steps to **prepare data for physics analysis** and achieve
  - reliable, high quality data (yes, we ***reject*** low quality data)
  - the ***best performance*** from our detectors
  - readiness for ***physics analysis***

1. Make sure that the **data quality** is excellent, also in real time
   - Maximise the amount of data that is useful

2. **Calibrate** the detectors
   - Correct for imperfections in the detectors, account for changes over time, etc.

3. **Reconstruct physics signals** from the data
   - Produce analysis object data which contains physics analysis level information like how many muons does the event have

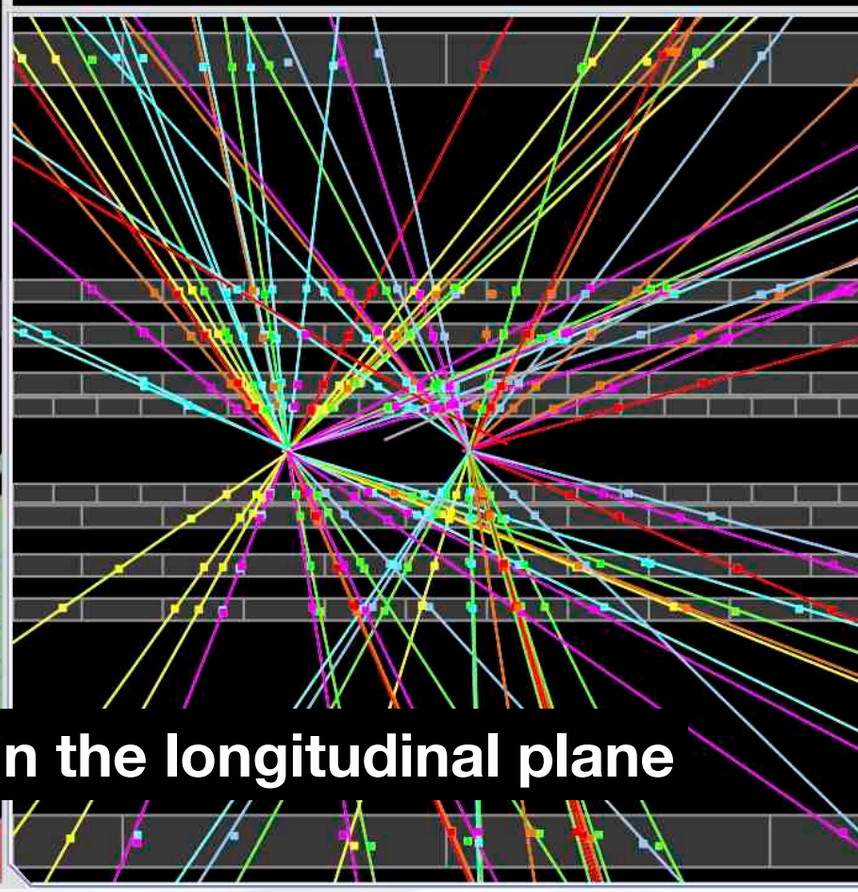

Muon
Spectrometer
Hadronic
Calorimeter
Electromagnetic
Calorimeter
Solenoid magnet
Tracking
Transition
Radiation
Tracker
Pixel/SCT detector
Proton
Neutrino
Neutron
Muon
Electron
Photon
The dashed tracks
are invisible to
the detector
Basic reconstruction =
Tracks
Calorimeter Clusters



ATLAS EXPERIMENT

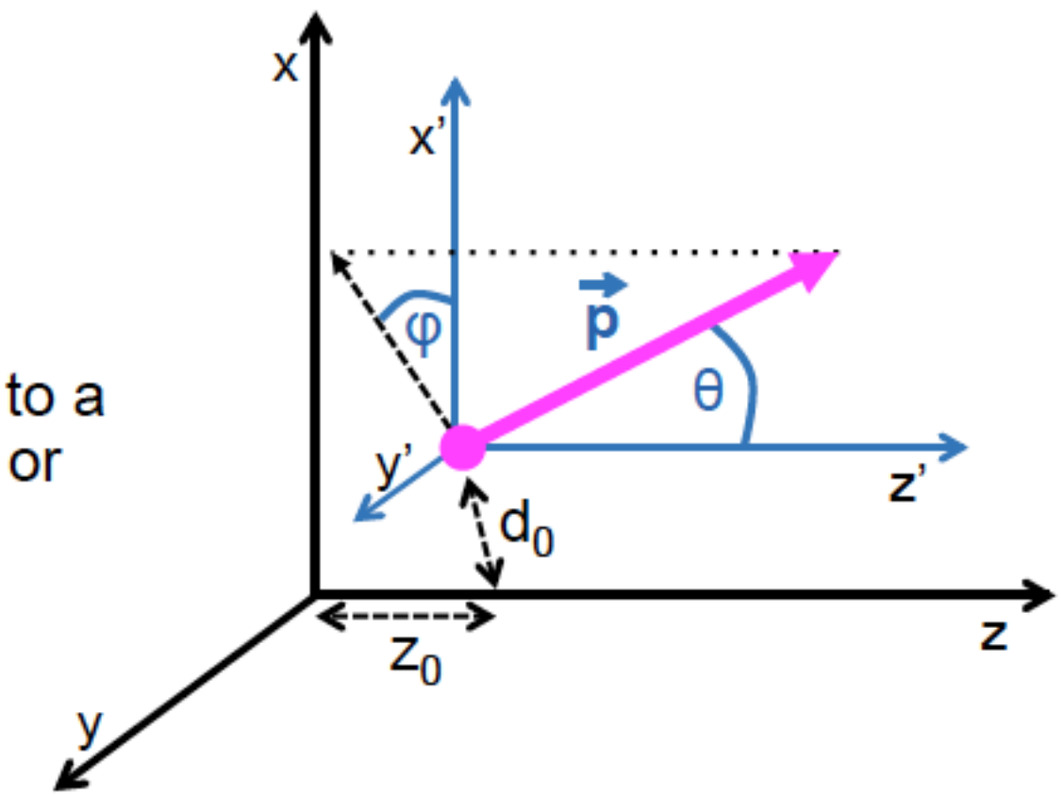
Run Number: 265545, Event Number: 5720351

Date: 2015-05-21 10:39:54 CEST



Reconstructing particle tracks

- **For a track we measure:**
  - Its momentum;
  - It's direction;
  - Its charge;
  - Its "perigee": the closest point to a reference line, transverse (d_0) or longitudinal (z_0).

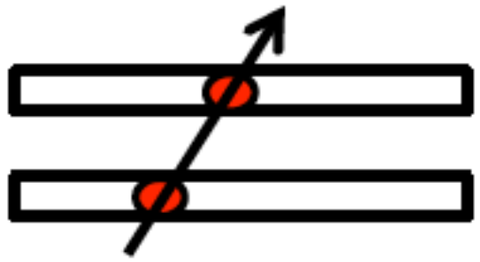


- The track ***curves in the transverse plane*** because of the magnet
- It travels in a ***straight line in the longitudinal plane***

- ***Question: What kind of trajectory describes a particle traveling through our tracking detectors ?***

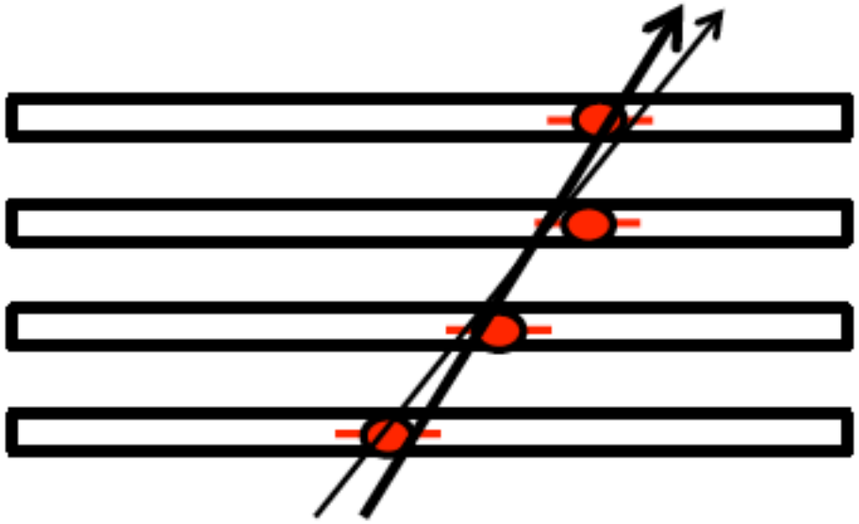
Track fitting

- Perfect measurement – ideal

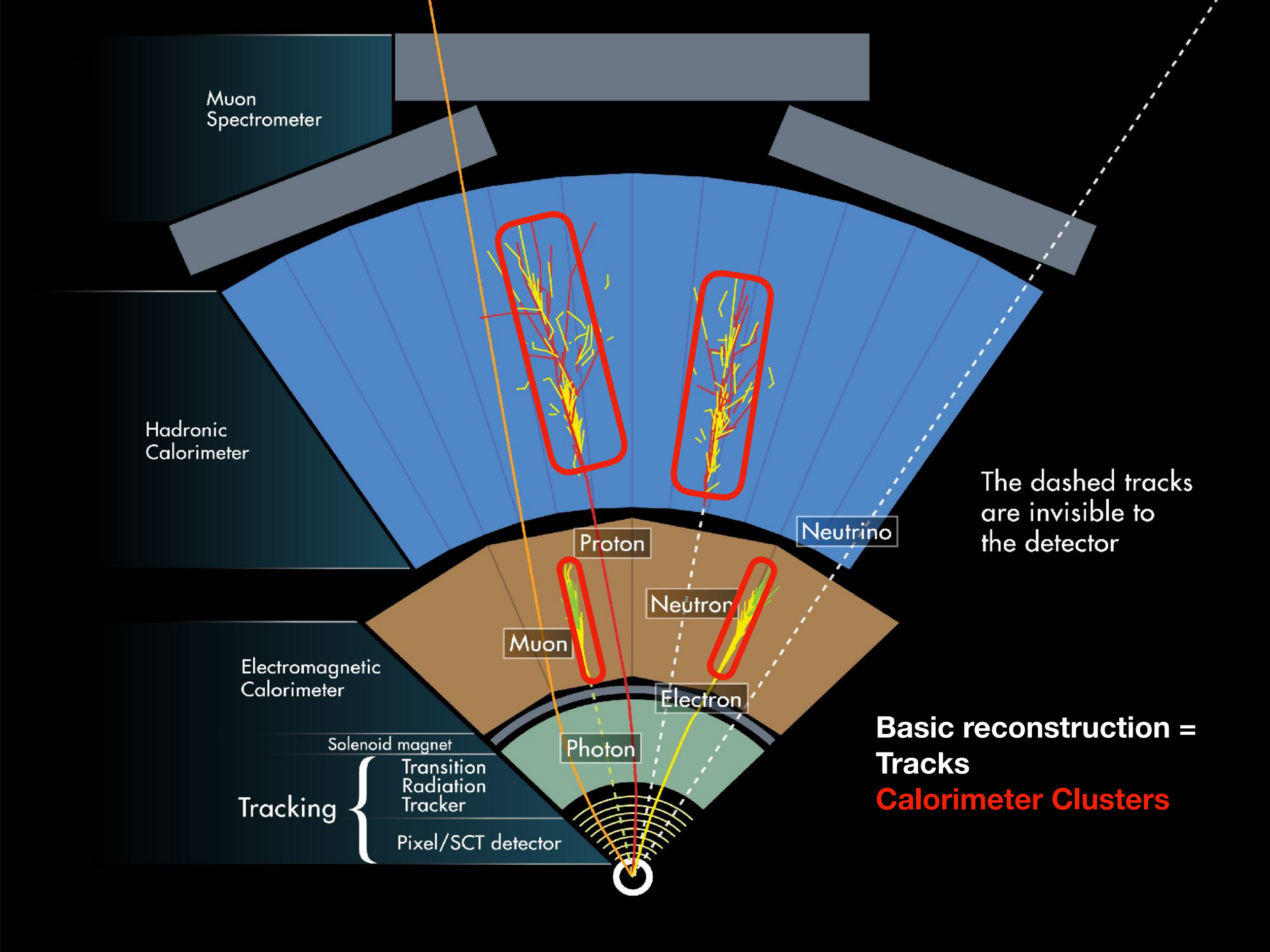


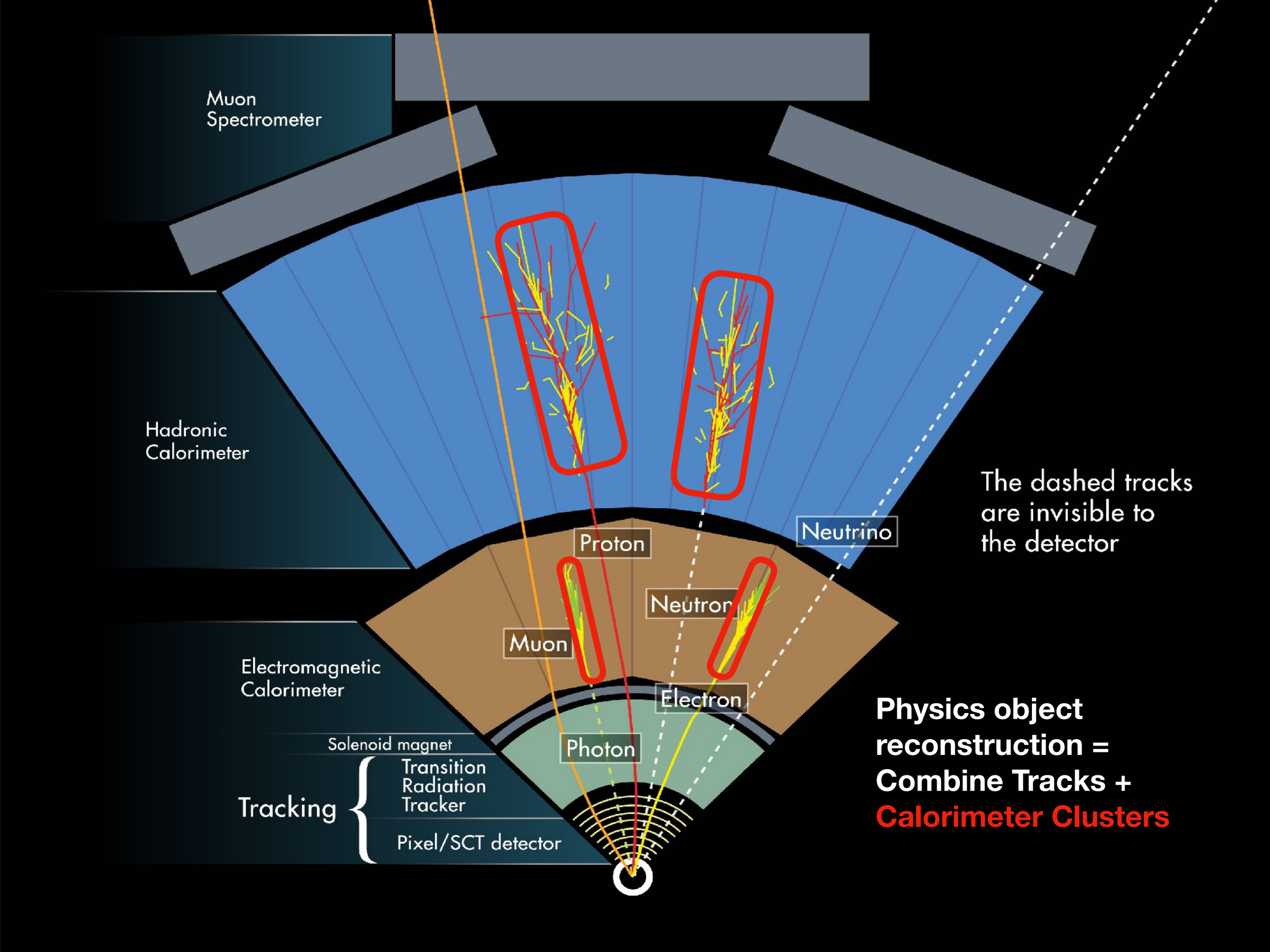
- Imperfect measurement – reality

- Small errors and more points help to constrain the possibilities

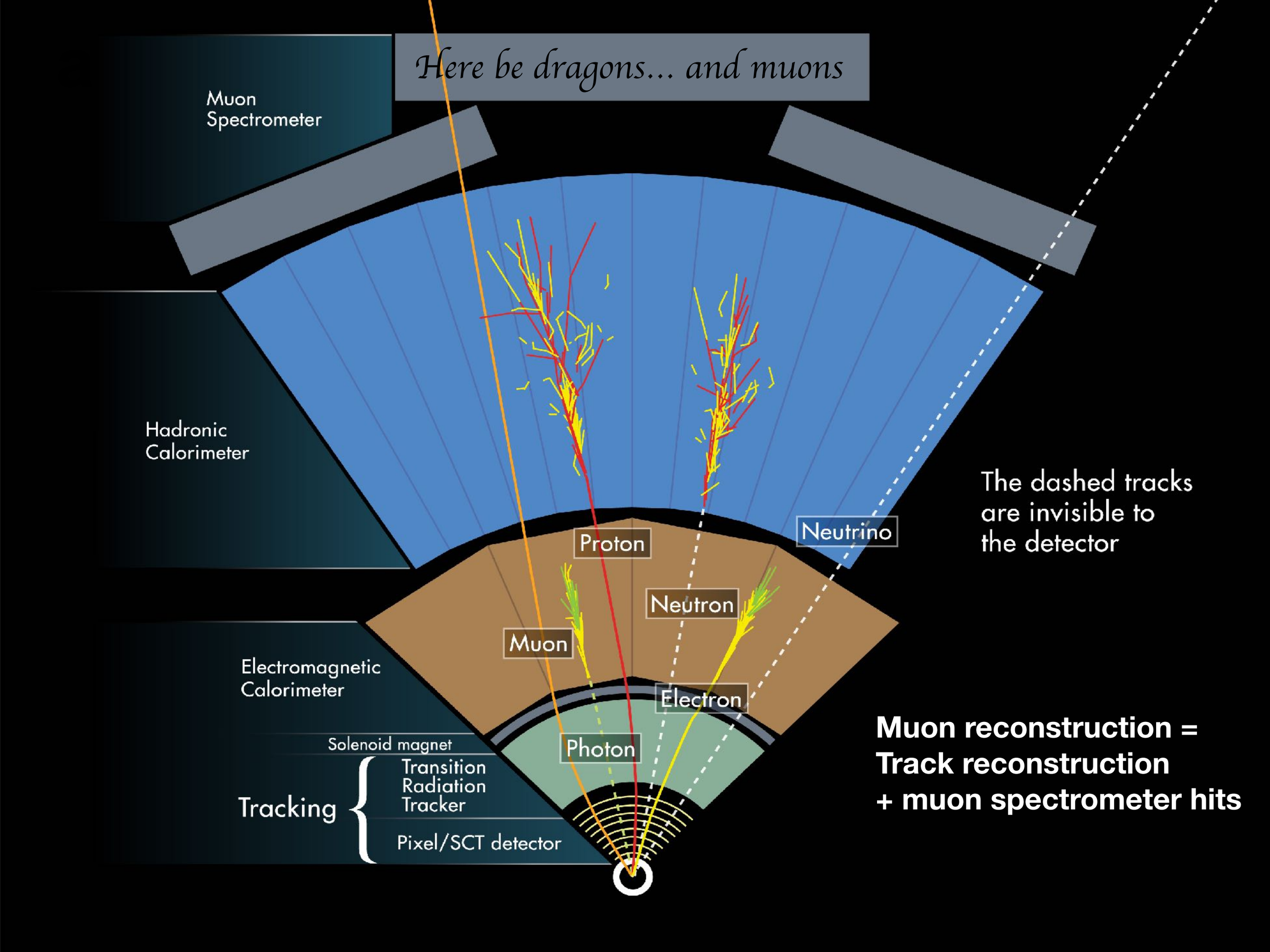


- **Quantitatively:**
  - Parameterize the track;
  - Find parameters by Least-Squares-Minimization;
  - Obtain also uncertainties on the track parameters.


Muon
Spectrometer
Hadronic
Calorimeter
Electromagnetic
Calorimeter
Solenoid magnet
Tracking
Transition
Radiation
Tracker
Pixel/SCT detector
Proton
Neutrino
Neutron
Muon
Electron
Photon
The dashed tracks
are invisible to
the detector
Basic reconstruction =
Tracks
Calorimeter Clusters

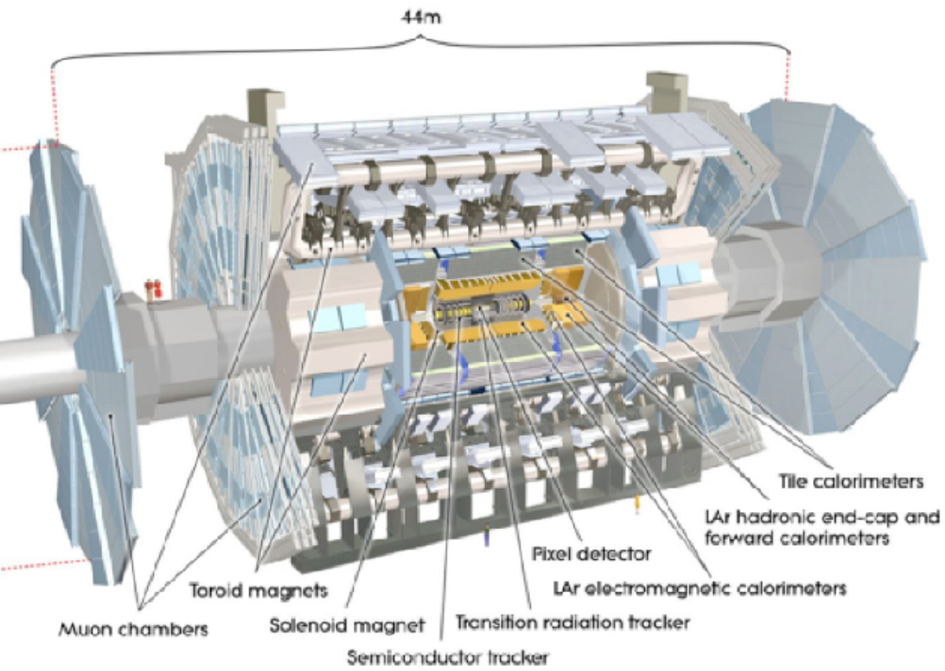

Muon Spectrometer
Hadronic Calorimeter
Electromagnetic Calorimeter
Solenoid magnet
Tracking
Transition Radiation Tracker
Pixel/SCT detector
Proton
Neutrino
Neutron
Muon
Electron
Photon
The dashed tracks are invisible to the detector
Physics object reconstruction = Combine Tracks + Calorimeter Clusters

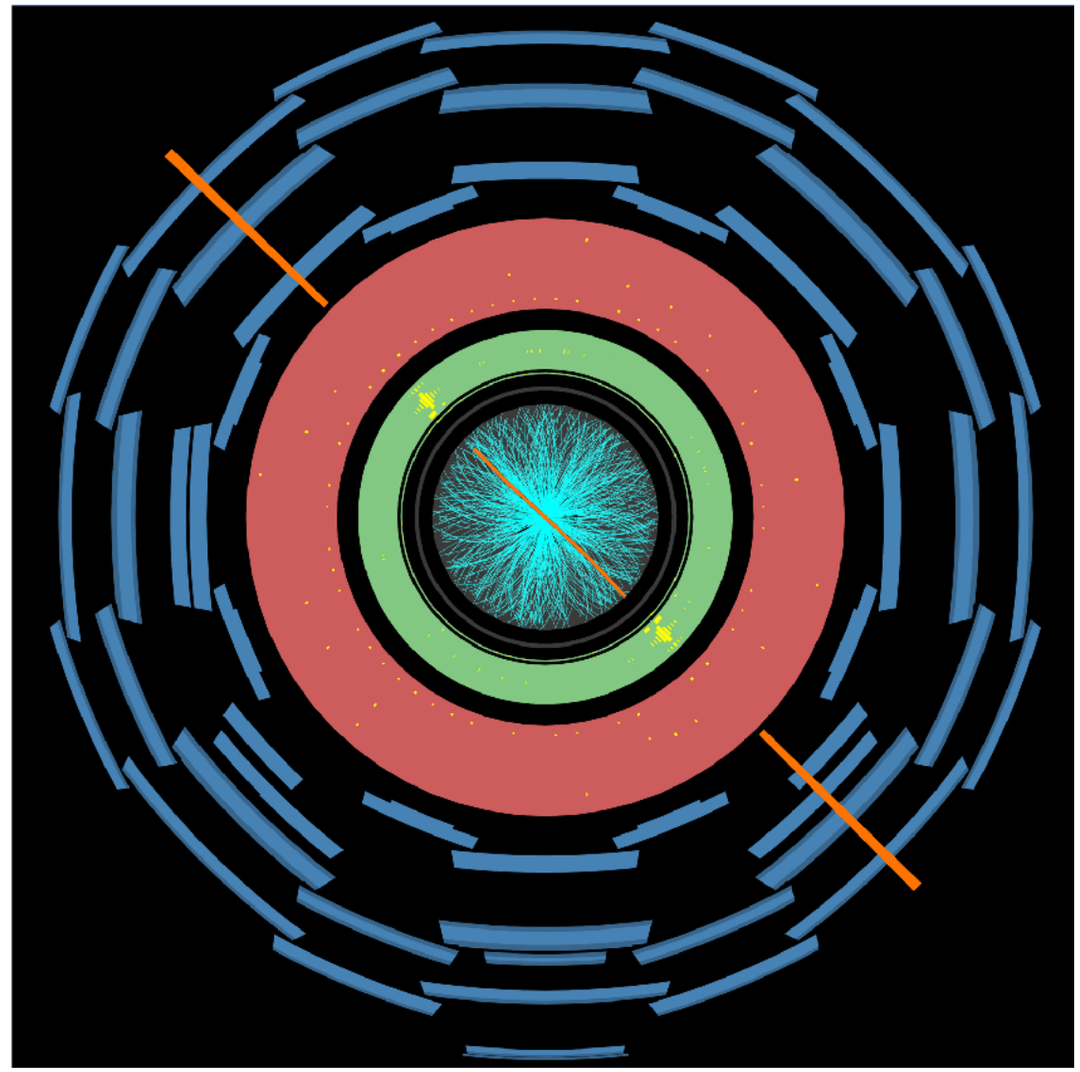
Here be dragons… and muons



The dashed tracks are invisible to the detector

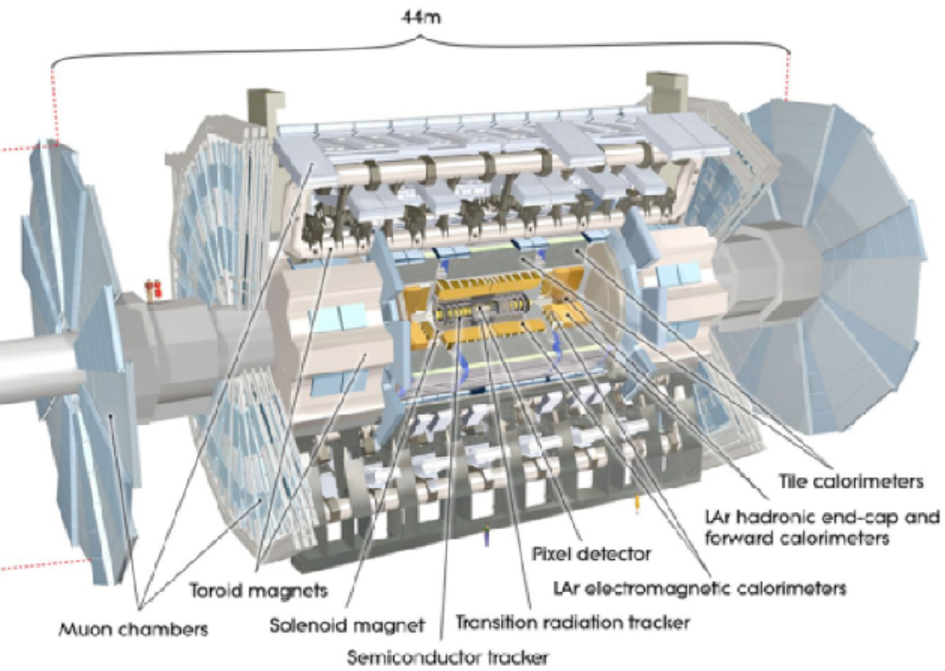
Muon reconstruction = Track reconstruction + muon spectrometer hits

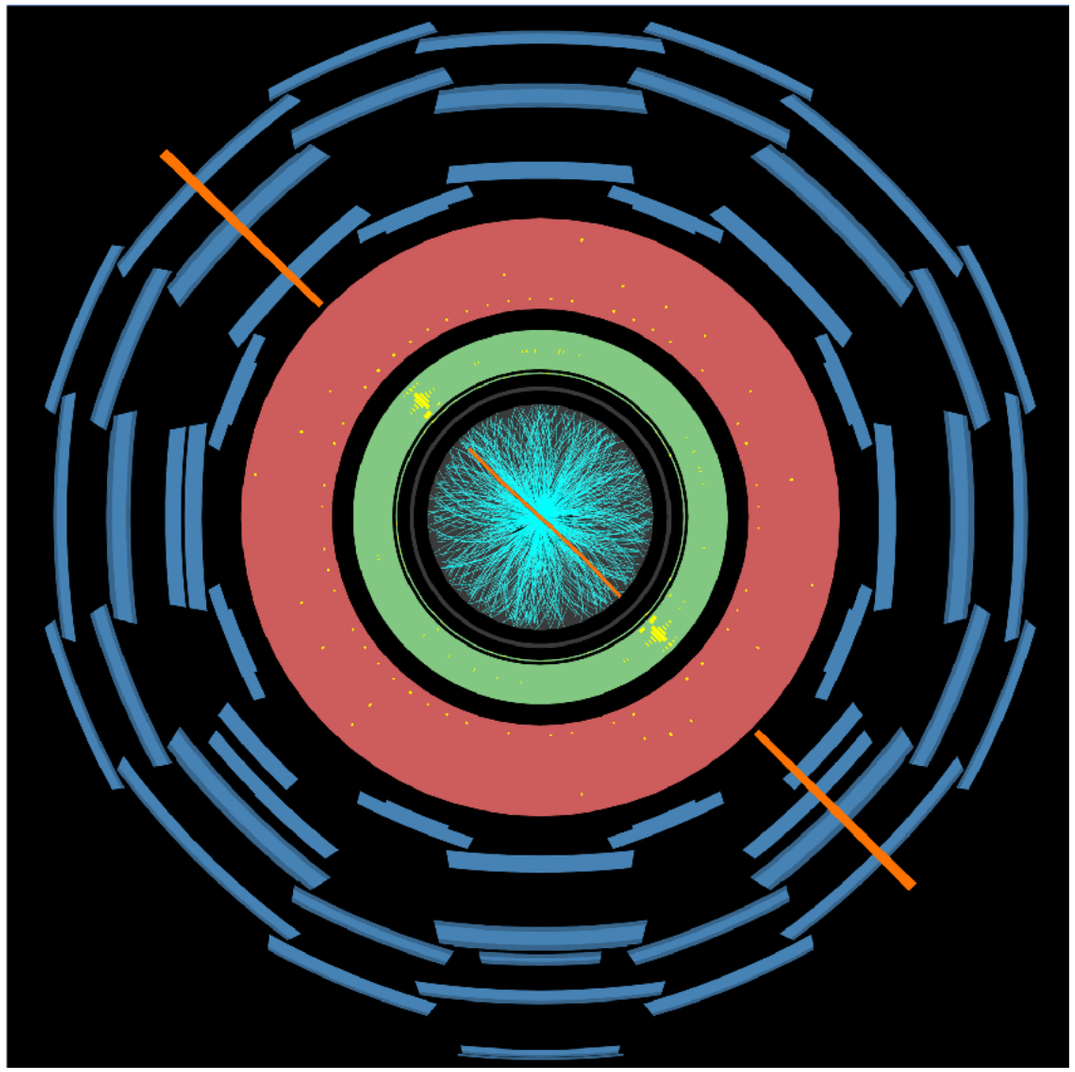




- **Question**: what physics process is observed in this event?

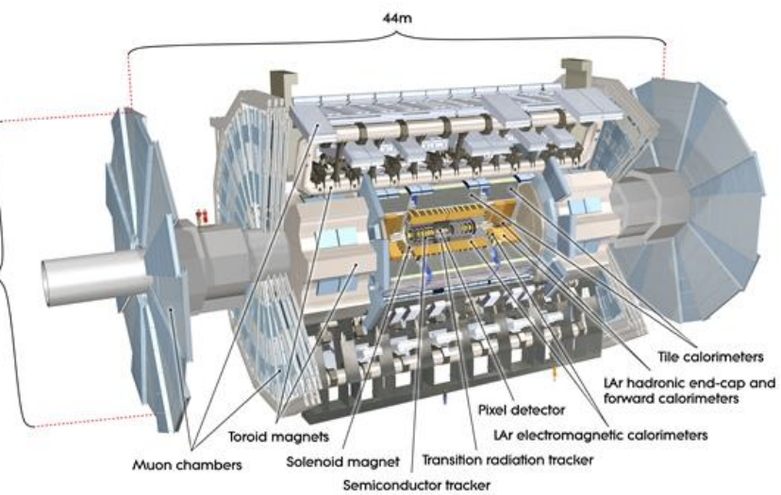
Neutrinos

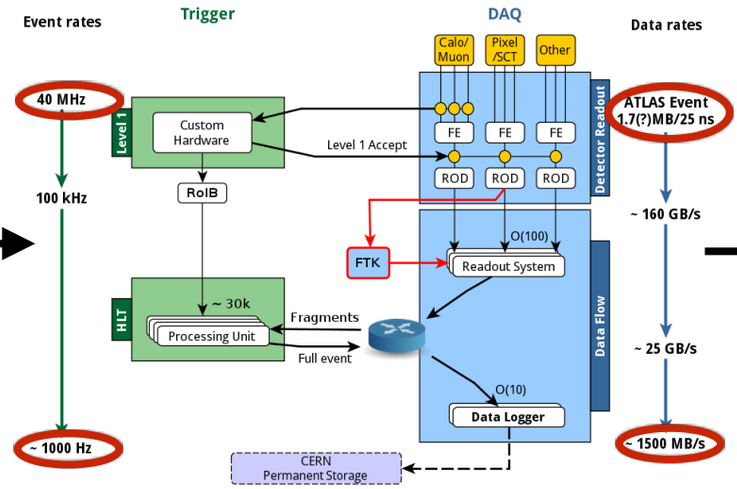




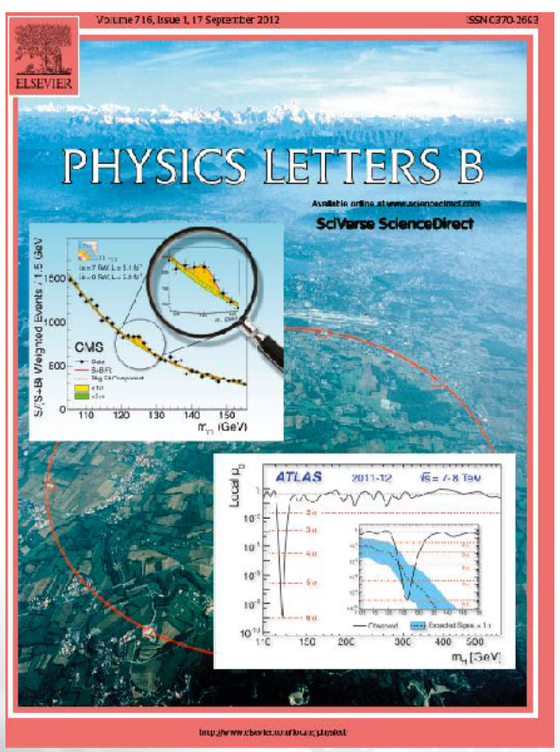
- Let's look at the simplest case for reconstructing neutrinos
- Remember, we are looking down the beam pipe, so the plane of the display is transverse to the proton beam direction
- ***Question**: Can you quantify the momentum in this plane **before** the proton collision*
  - *What does that tell you about the distribution of momentum **after** the collision?*
  - *How would this look if we had a **W boson** instead of a **Z boson** ?*

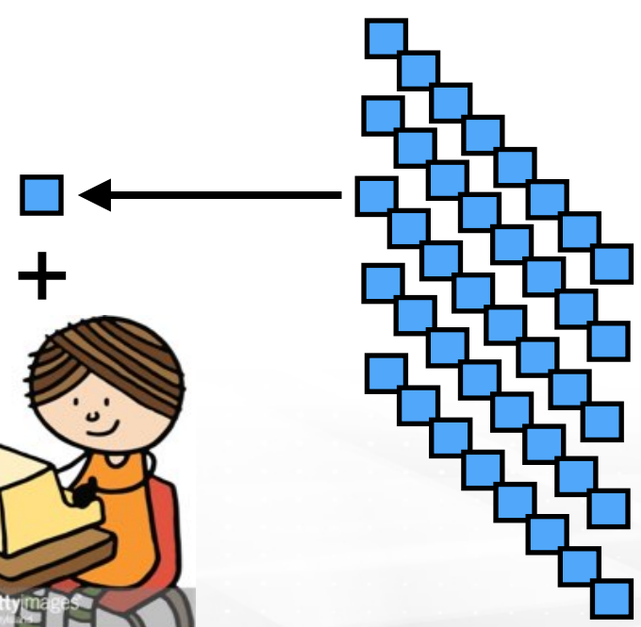
Data's journey





Data Preparation





Distributed computing

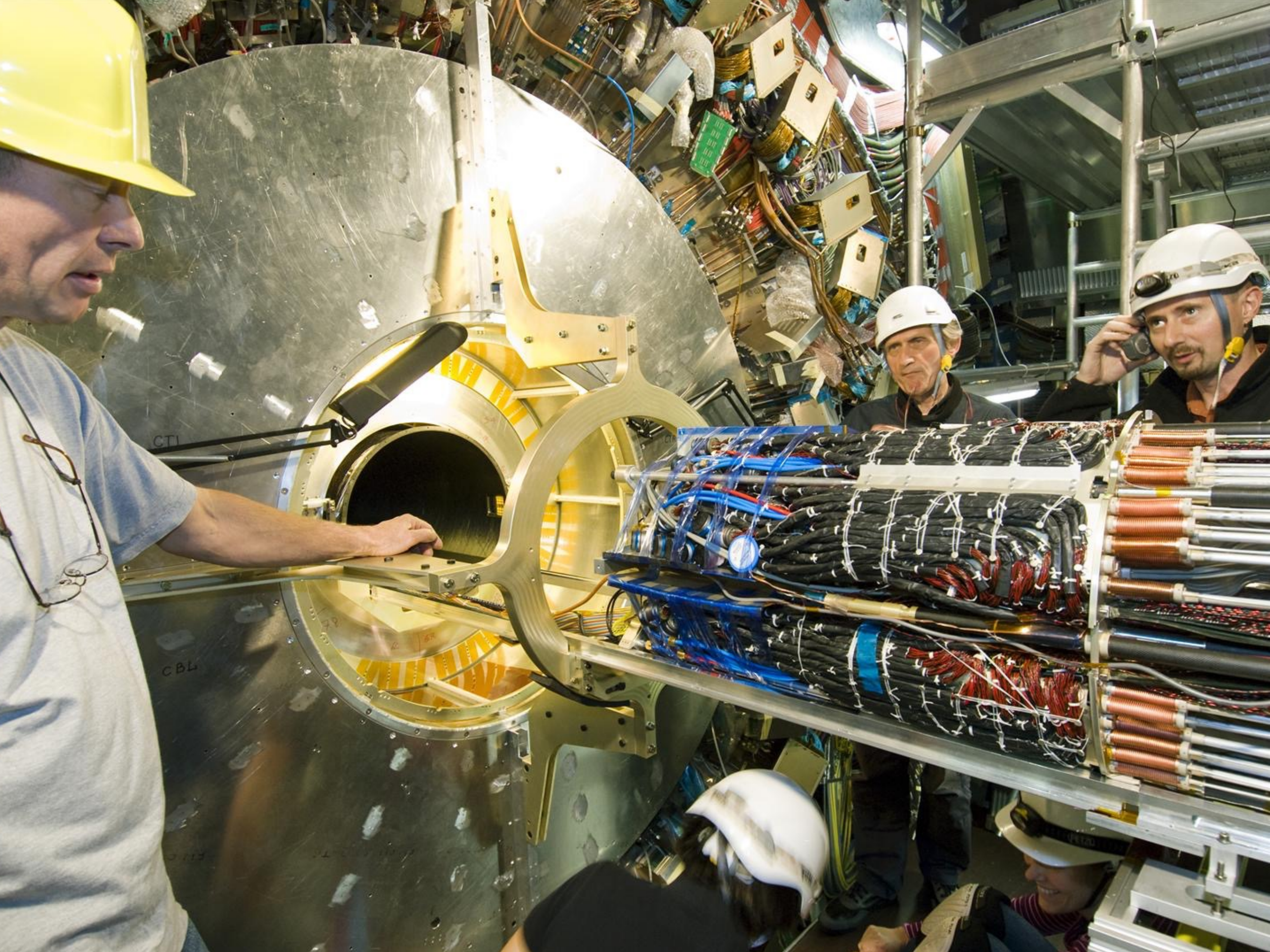
Data Preparation

- Three major steps to **prepare data for physics analysis** and achieve
  - reliable, high quality data (yes, we ***reject*** low quality data)
  - the ***best performance*** from our detectors
  - readiness for ***physics analysis***

1. Make sure that the **data quality** is excellent, also in real time
   - Maximise the amount of data that is useful

2. **Calibrate** the detectors
   - Correct for imperfections in the detectors, account for changes over time, etc.

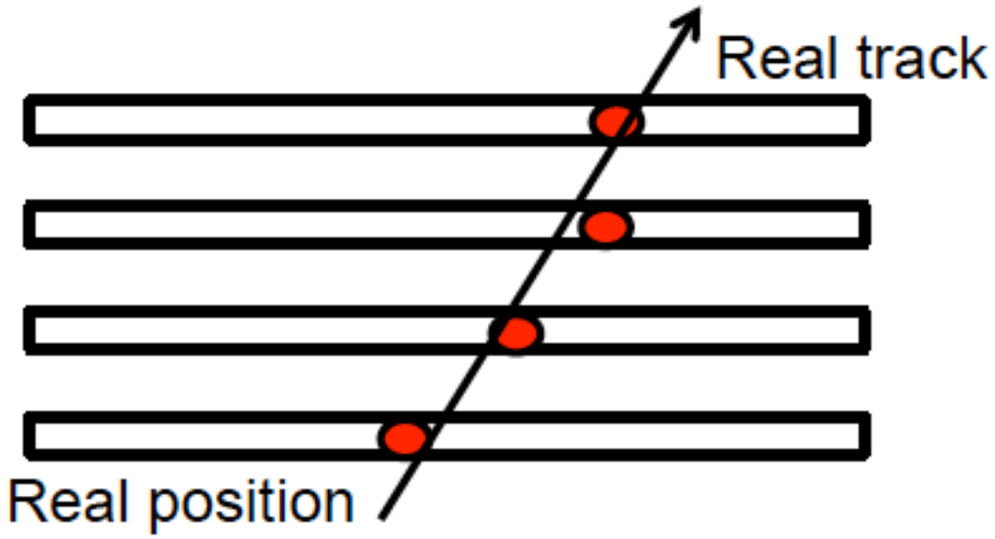
3. **Reconstruct physics signals** from the data
   - Produce analysis object data which contains physics analysis level information like how many muons does the event have



Misalignment and detector effects

- **Presence of Material**
  - Coulomb scattering off the core of atoms
  - Energy loss due to ionization
  - Bremsstrahlung
  - Hadronic interaction

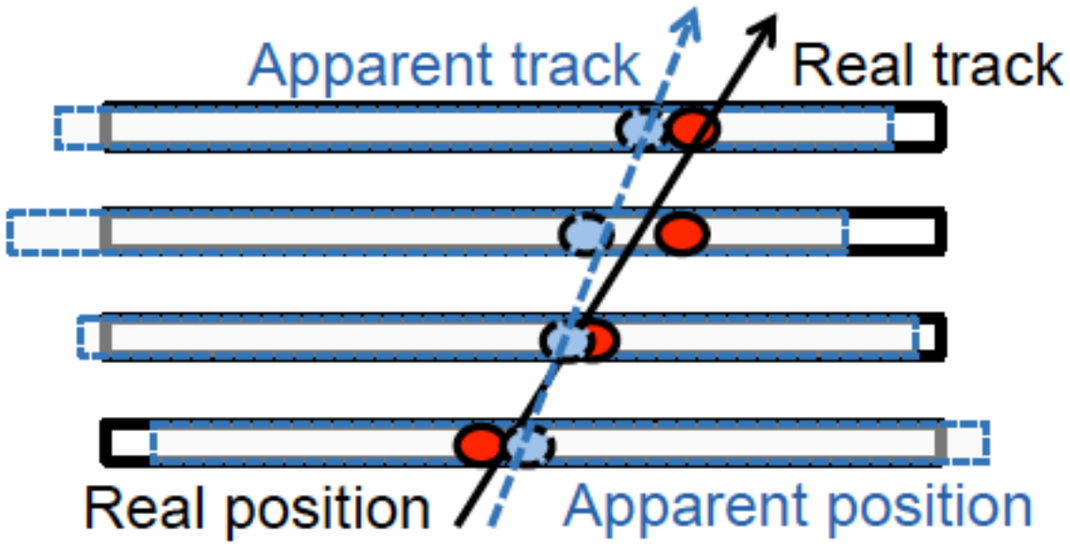
- **Misalignment**
  - Detector elements not positioned in space with perfect accuracy.
  - Alignment corrections derived from data and applied in track reconstruction.



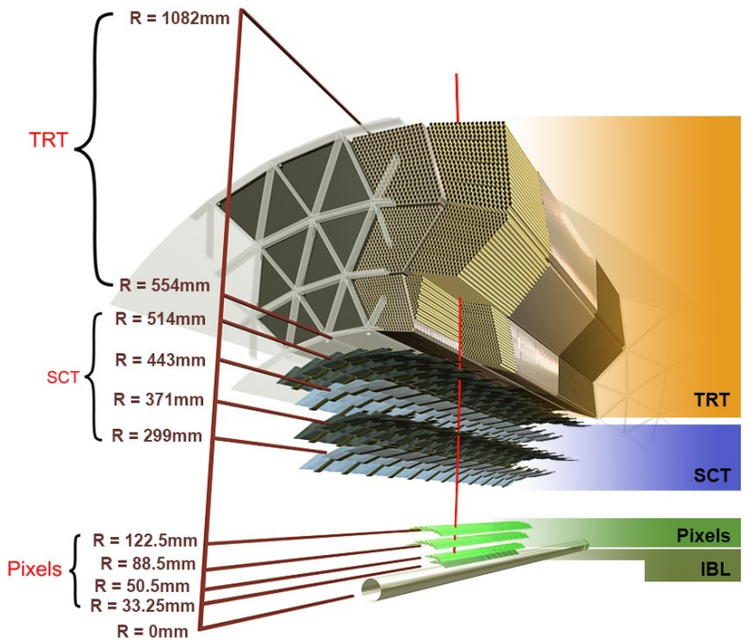
Misalignment and detector effects

- **Presence of Material**
  - Coulomb scattering off the core of atoms
  - Energy loss due to ionization
  - Bremsstrahlung
  - Hadronic interaction

- **Misalignment**
  - Detector elements not positioned in space with perfect accuracy.
  - Alignment corrections derived from data and applied in track reconstruction.

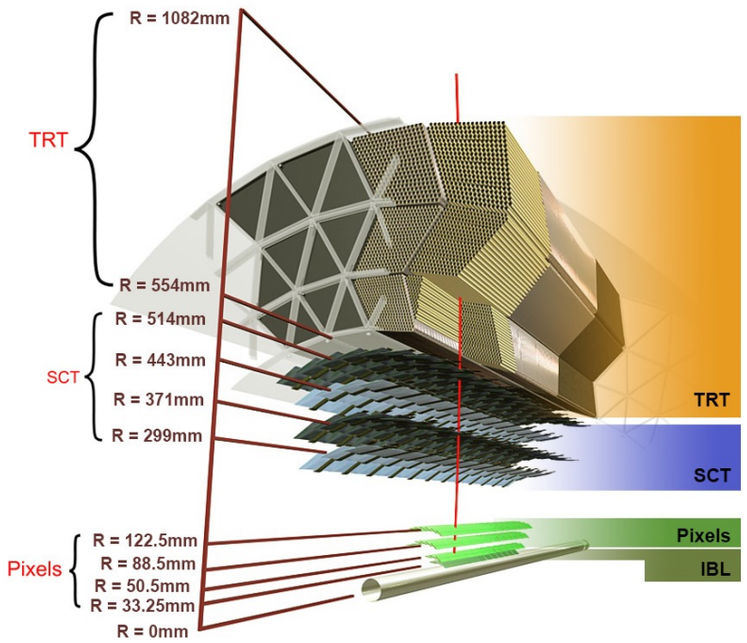


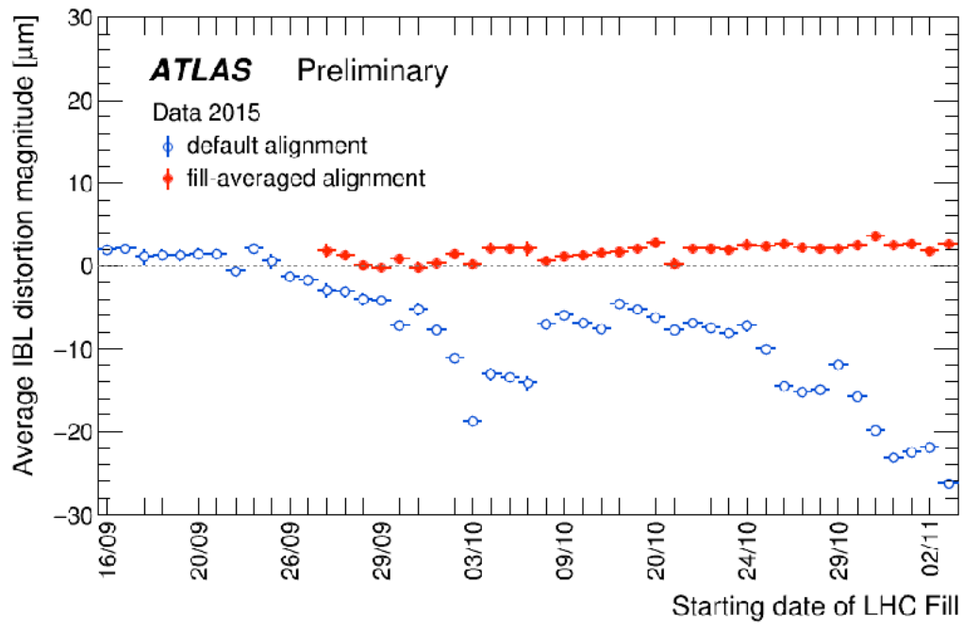
Calibration



- During the break between Run 1 and Run 2, ATLAS inserted the IBL, an extra layer of silicon tracker close to the beam pipe

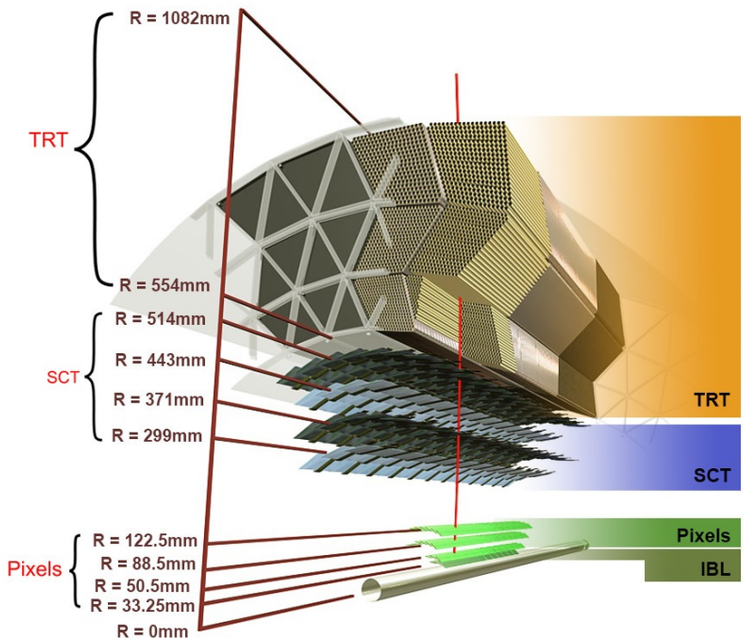
Calibration

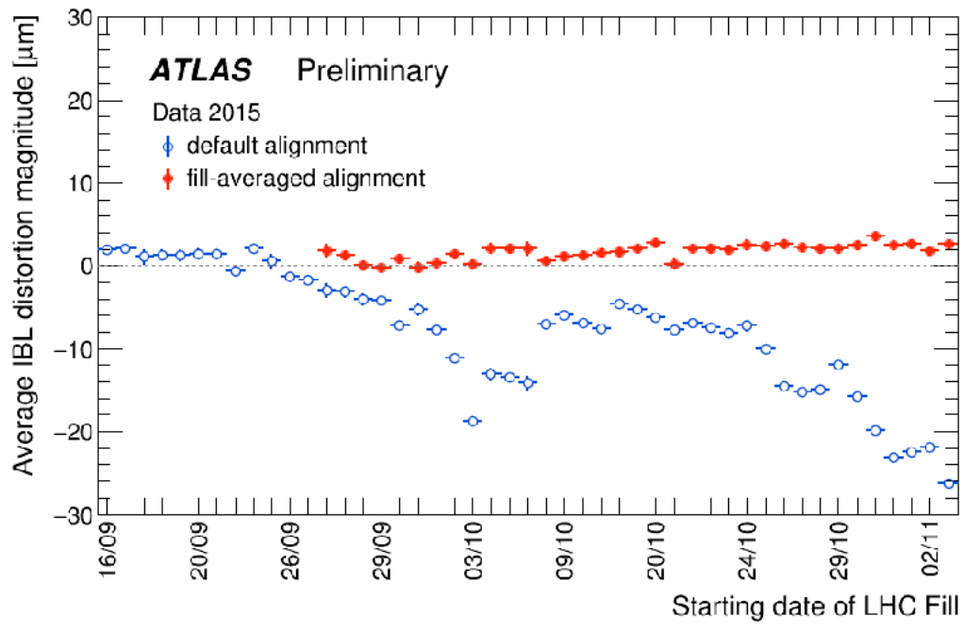


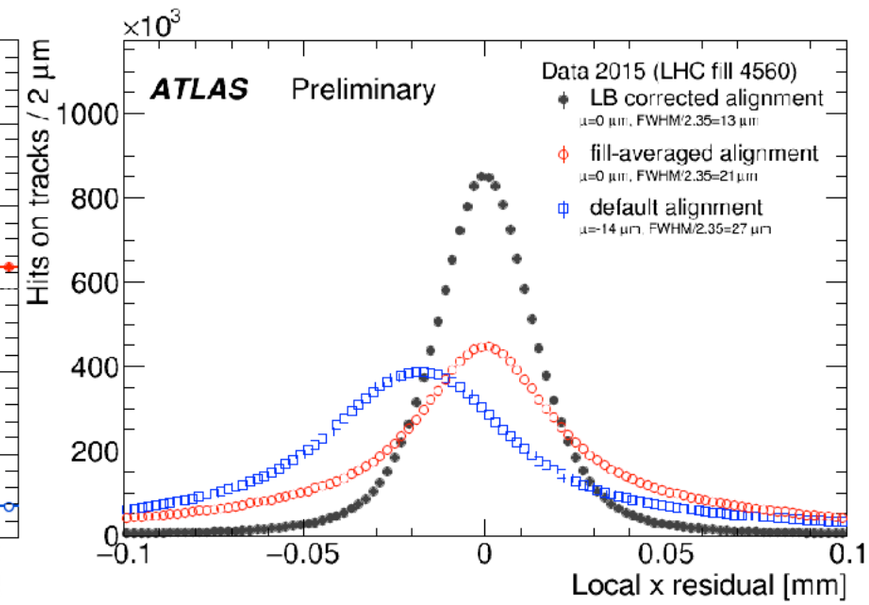


- During the break between Run 1 and Run 2, ATLAS inserted the IBL, an extra layer of silicon tracker close to the beam pipe
- At the start of data taking in Run 2, it started to move
- As time went on, the movement was very significant, much more than the detector precision so the movement could really be seen in physics distributions and data quality

Calibration







- During the break between Run 1 and Run 2, ATLAS inserted the IBL, an extra layer of silicon tracker close to the beam pipe
- At the start of data taking in Run 2, it started to move
- As time went on, the movement was very significant, much more than the detector precision so the movement could really be seen in physics distributions and data quality
- ATLAS quickly implemented and commissioned a correction procedure as part of its calibration process
- Following the correction the performance of the detector was back to nominal

Data Preparation

- Three major steps to **prepare data for physics analysis** and achieve
  - reliable, high quality data (yes, we ***reject*** low quality data)
  - the ***best performance*** from our detectors
  - readiness for ***physics analysis***

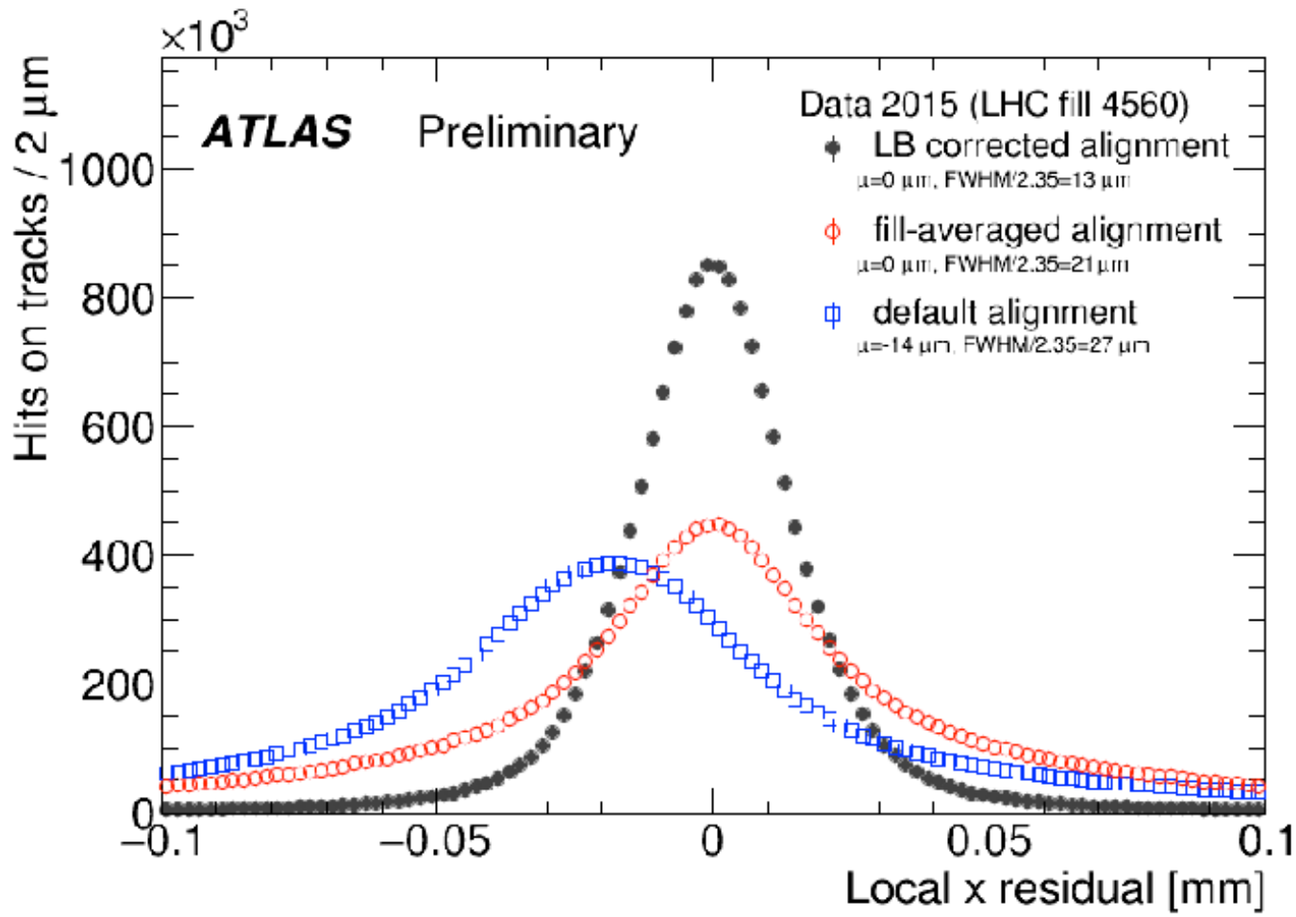
1. Make sure that the **data quality** is excellent, also in real time
   - Maximise the amount of data that is useful

2. **Calibrate** the detectors
   - Correct for imperfections in the detectors, account for changes over time, etc.

3. **Reconstruct physics signals** from the data
   - Produce analysis object data which contains physics analysis level information like how many muons does the event have

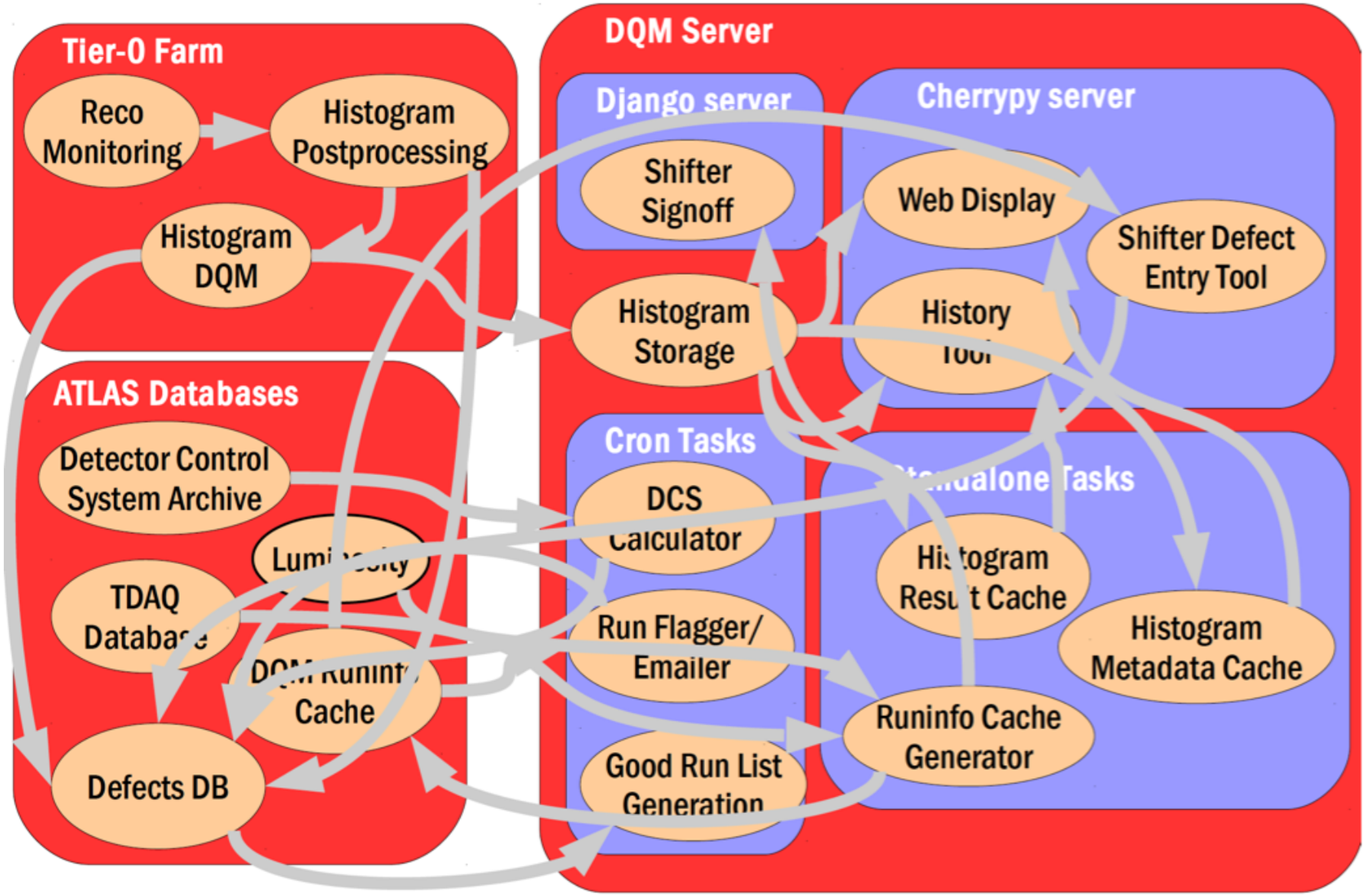
What makes good data quality?

What makes good data quality?



- The plot that shows a misalignment of the **ATLAS IBL** is a good example of a data quality plot
- The reference would be the **black** histogram, this is what data should look like
  - ***If the shifter sees the blue or red histogram, she will raise the alarm !***
- ***Before the calibration***, data quality assessment might ***reject*** this data for physics analysis
- ***After the calibration***, the data quality is ***good*** and the data can be used for physics analysis

Data describing data - metadata and databases



- In addition to the **calibration databases**, we need **databases** for several other important metadata tasks, **data quality** is a particularly important example
- Understanding our data requires us to keep precise track of our **metadata** too

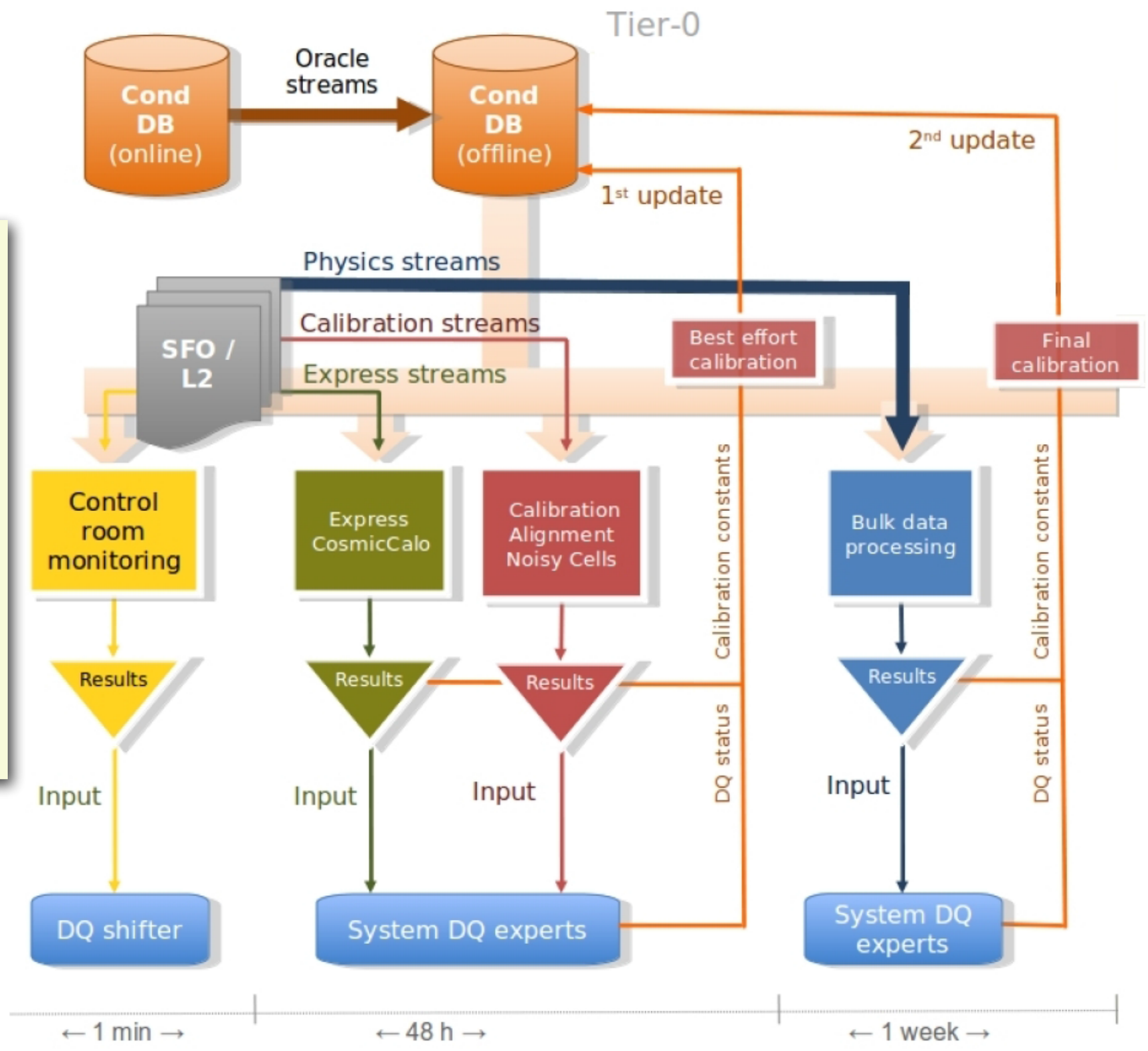
Data Preparation workflows

Runs on dedicated computing resources at CERN

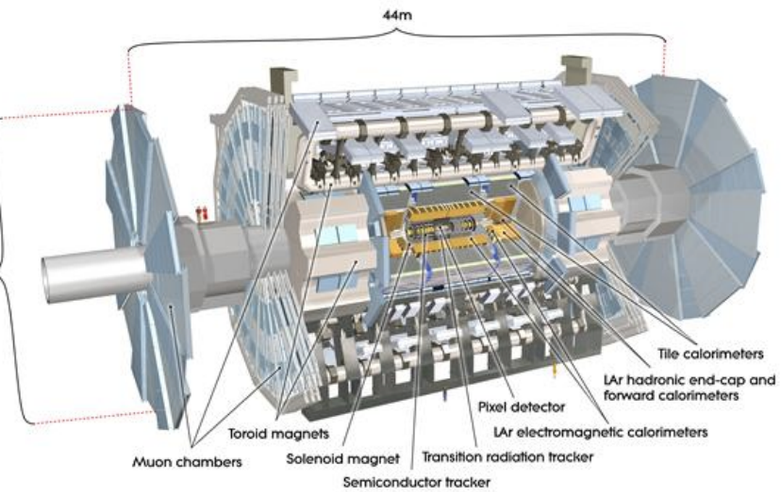
Providing data quality feedback at three levels

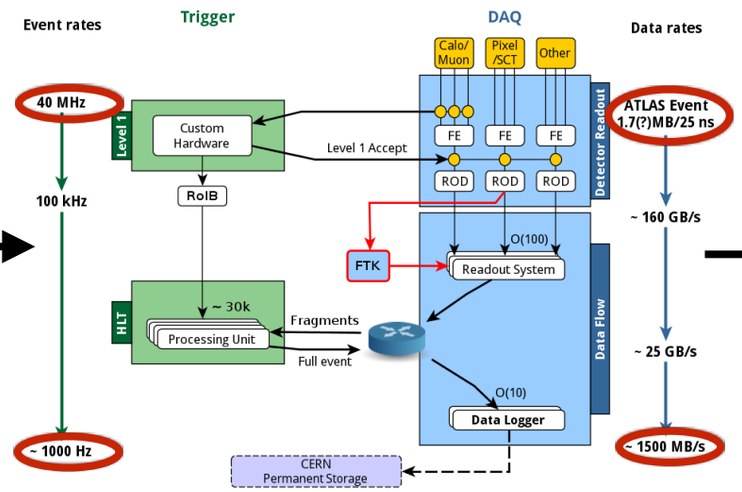
Once calibrations are ready, run reconstruction on all data

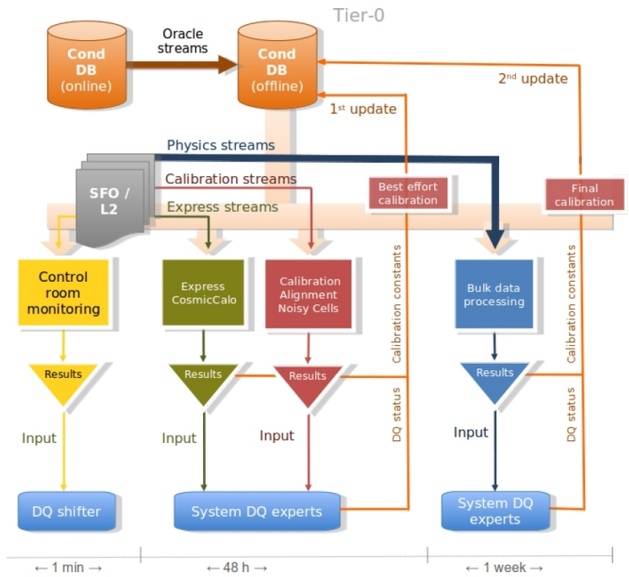
ATLAS typically processes
~60M events per day
~60 TB per day

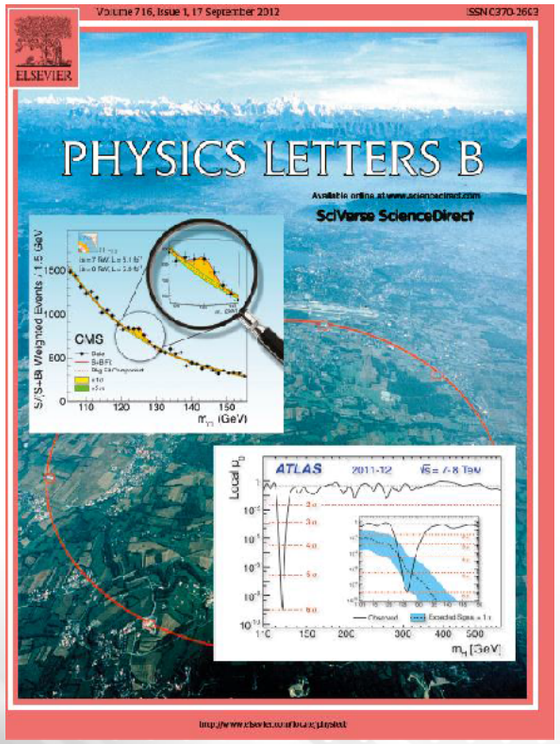


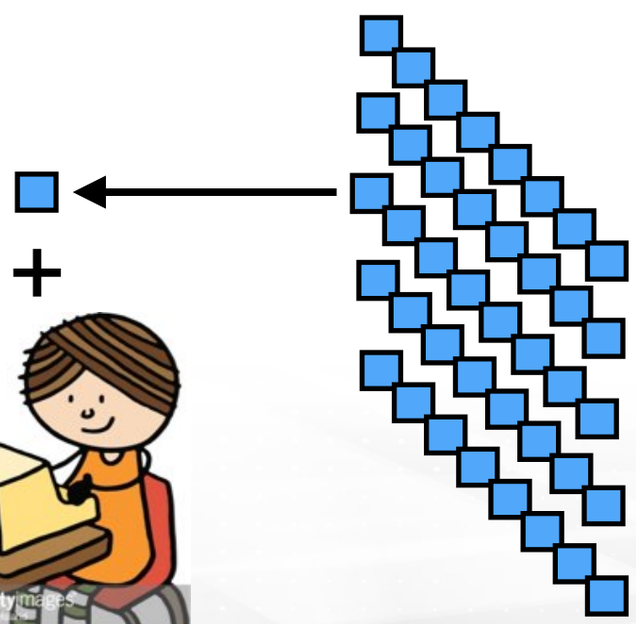
Data's journey









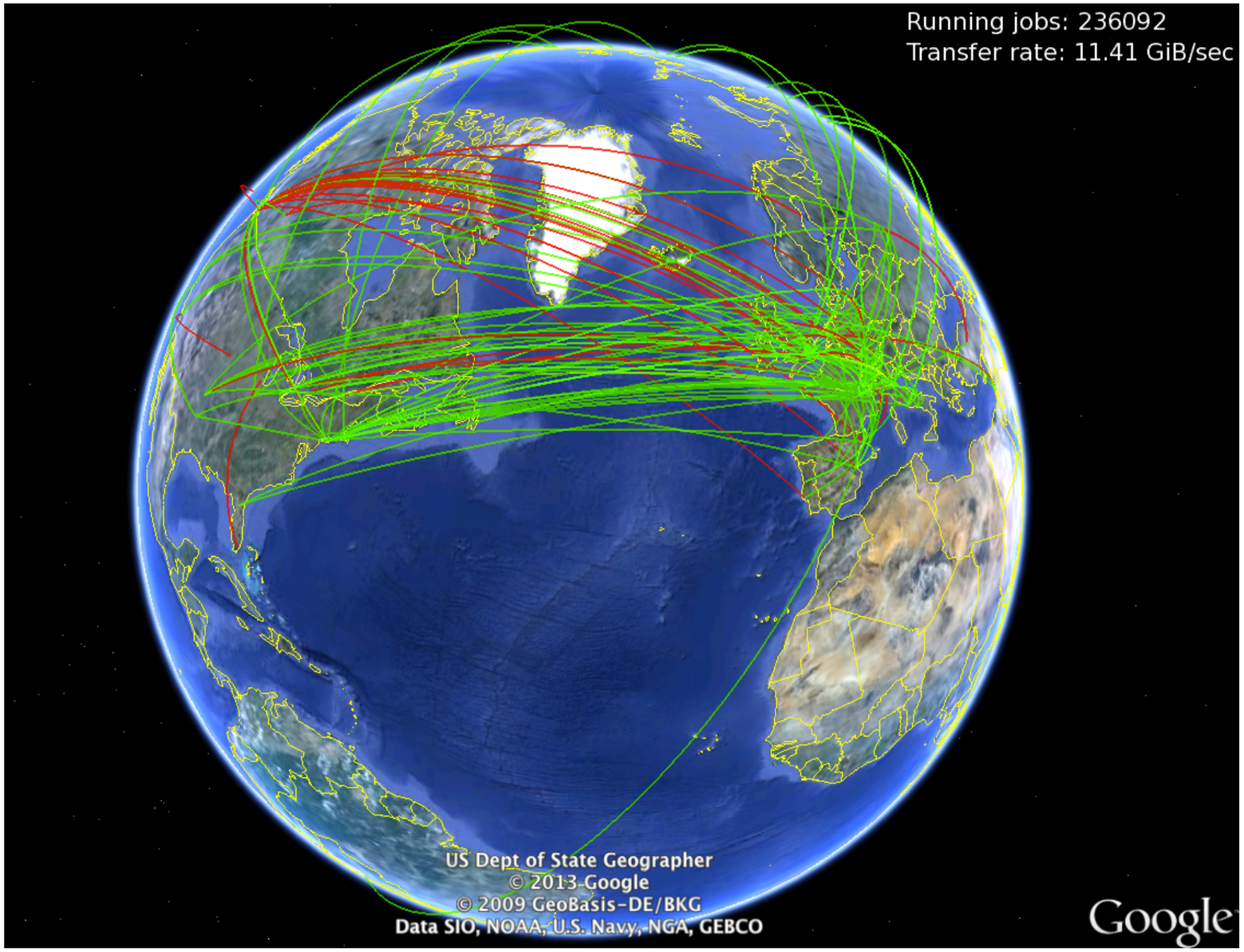


Distributed computing

The Worldwide LHC Computing Grid

- Now the data has been ***prepared for physics analysis***, it's time to extract our favourite physics signal!

- Many experiments, particularly those at the **LHC**, use computing sites all over the world via **the grid** to
  - harness all of that ***computing power***
  - enable collaborators ***worldwide*** to access the data



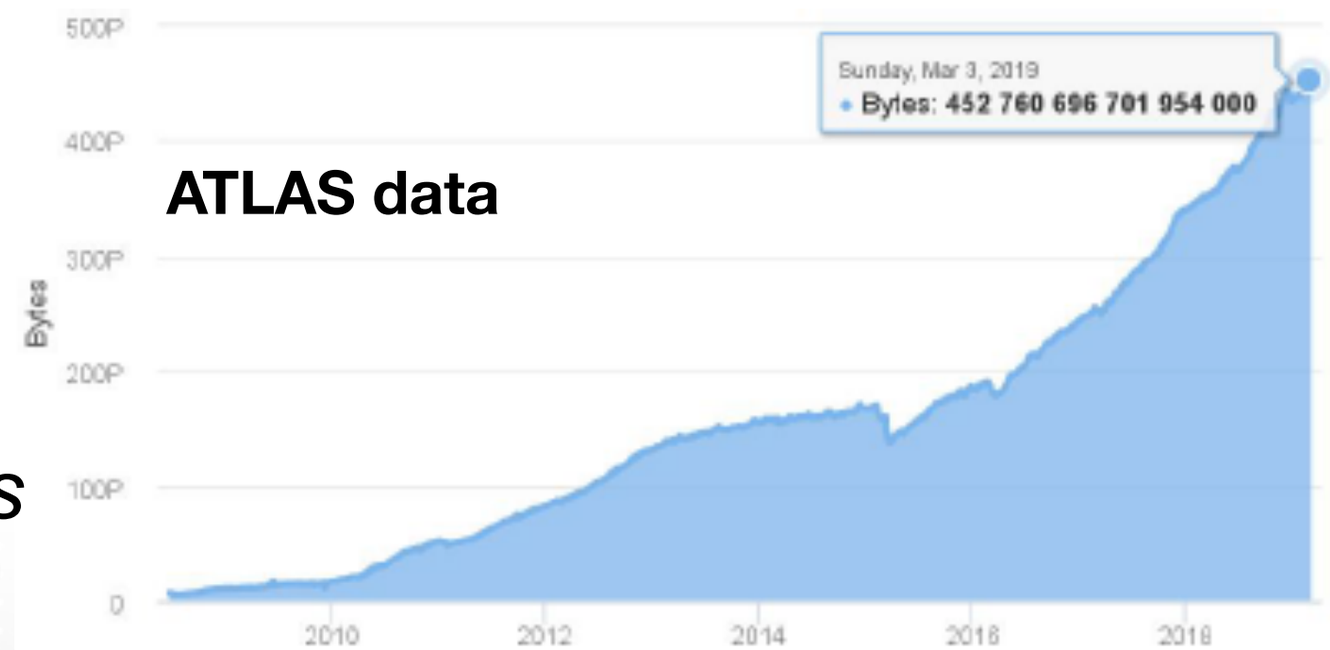
- Image taken from the **WLCG** GoogleEarth Dashboard
  - http://wlcg.web.cern.ch/wlcg-google-earth-dashboard

Analysing a lot of data

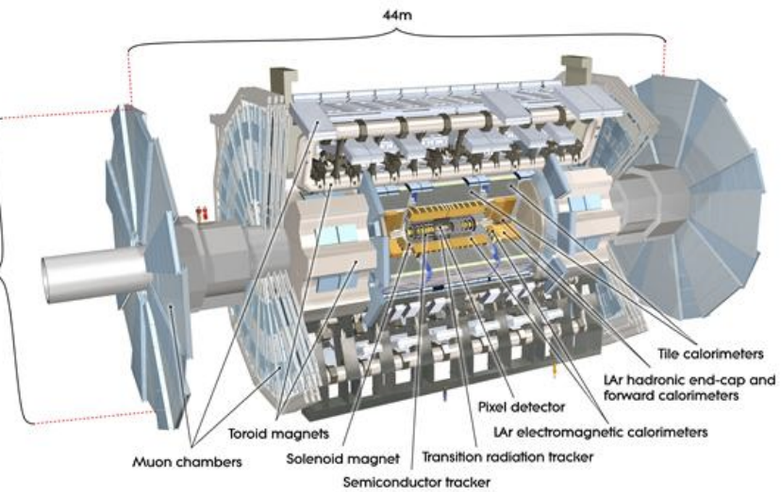
- Our data is calibrated and with good data quality
- and we've reconstructed the physics objects in the data
  - ***it is reliable, accurate and ready for physics analysis***

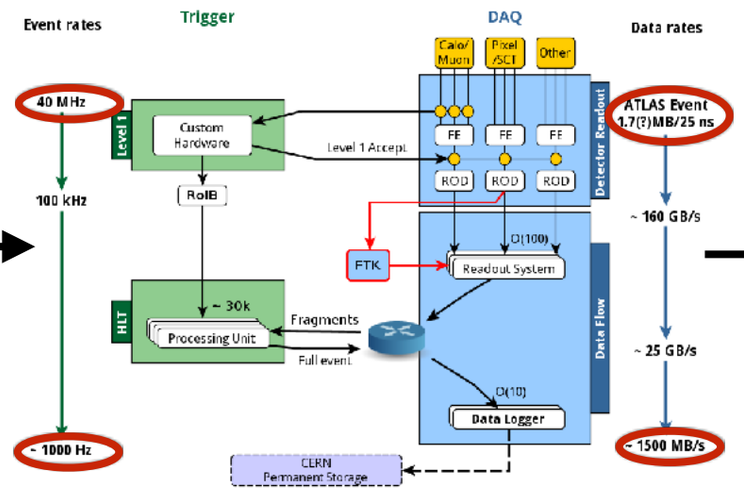
- ***Now we can extract our measurements***

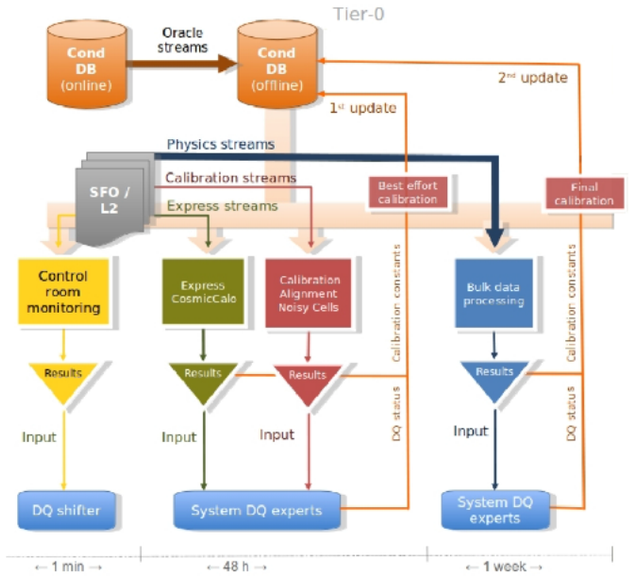
- ***Question****: How long would it take to read all of the ATLAS data? (Assume for simplicity you have off-the-shelf SSDs with read speed ~500MB/s)*

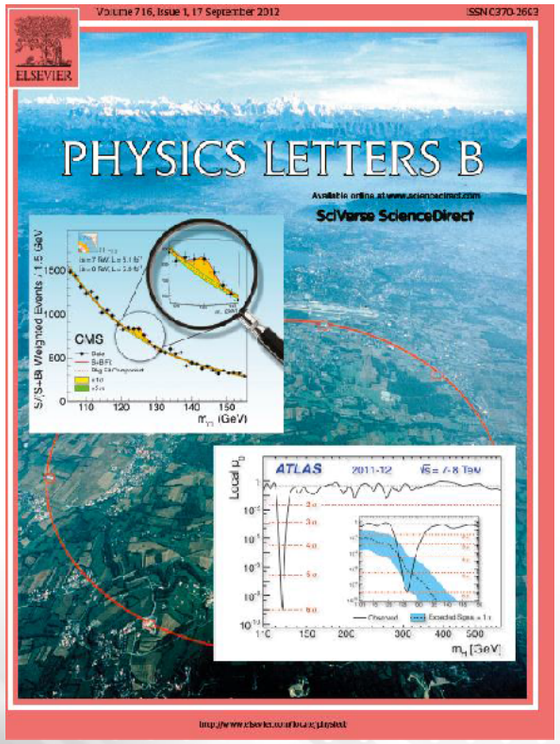


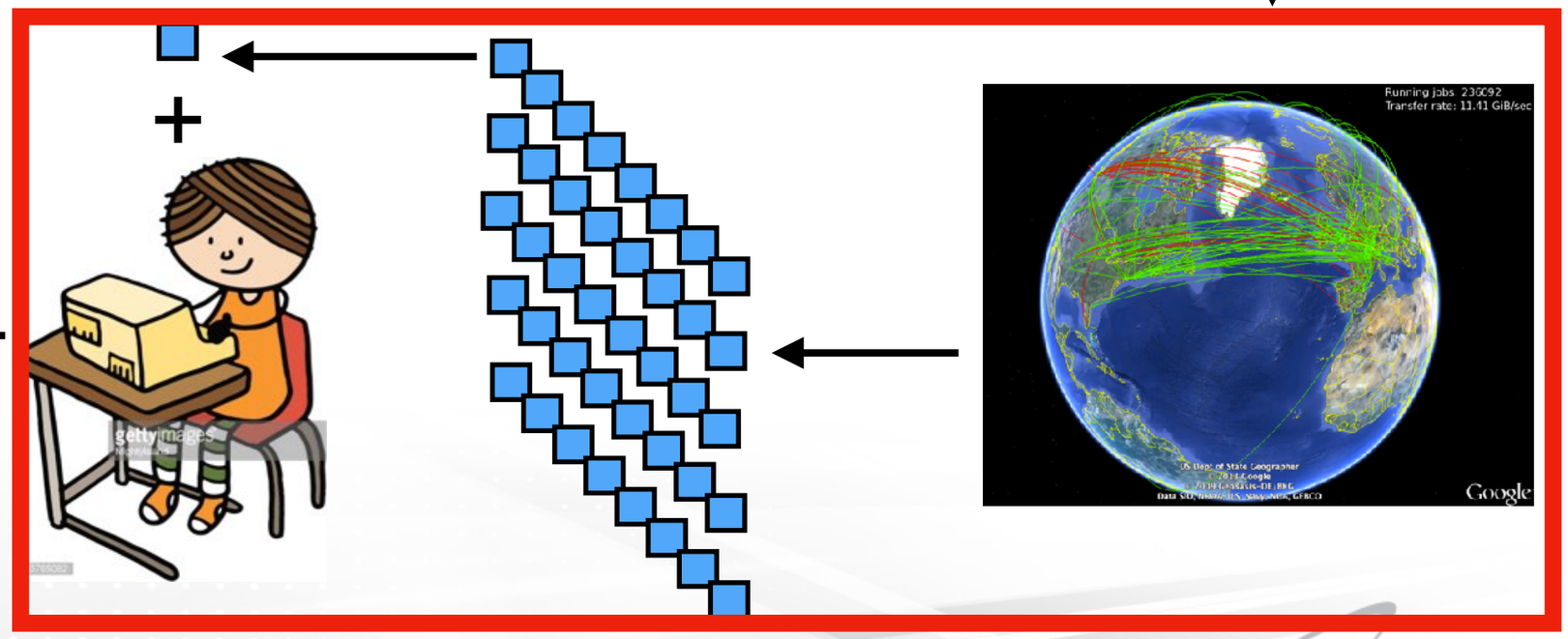
Data's journey



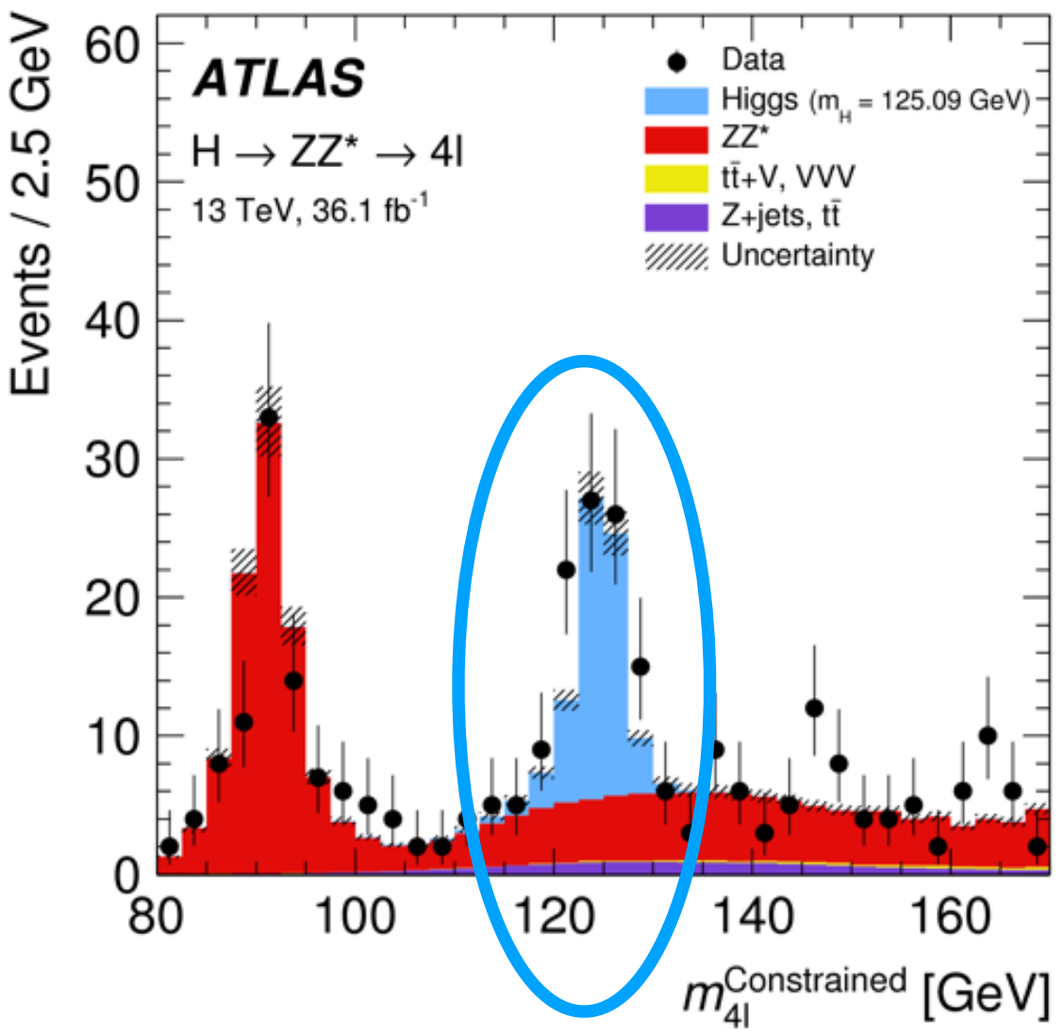


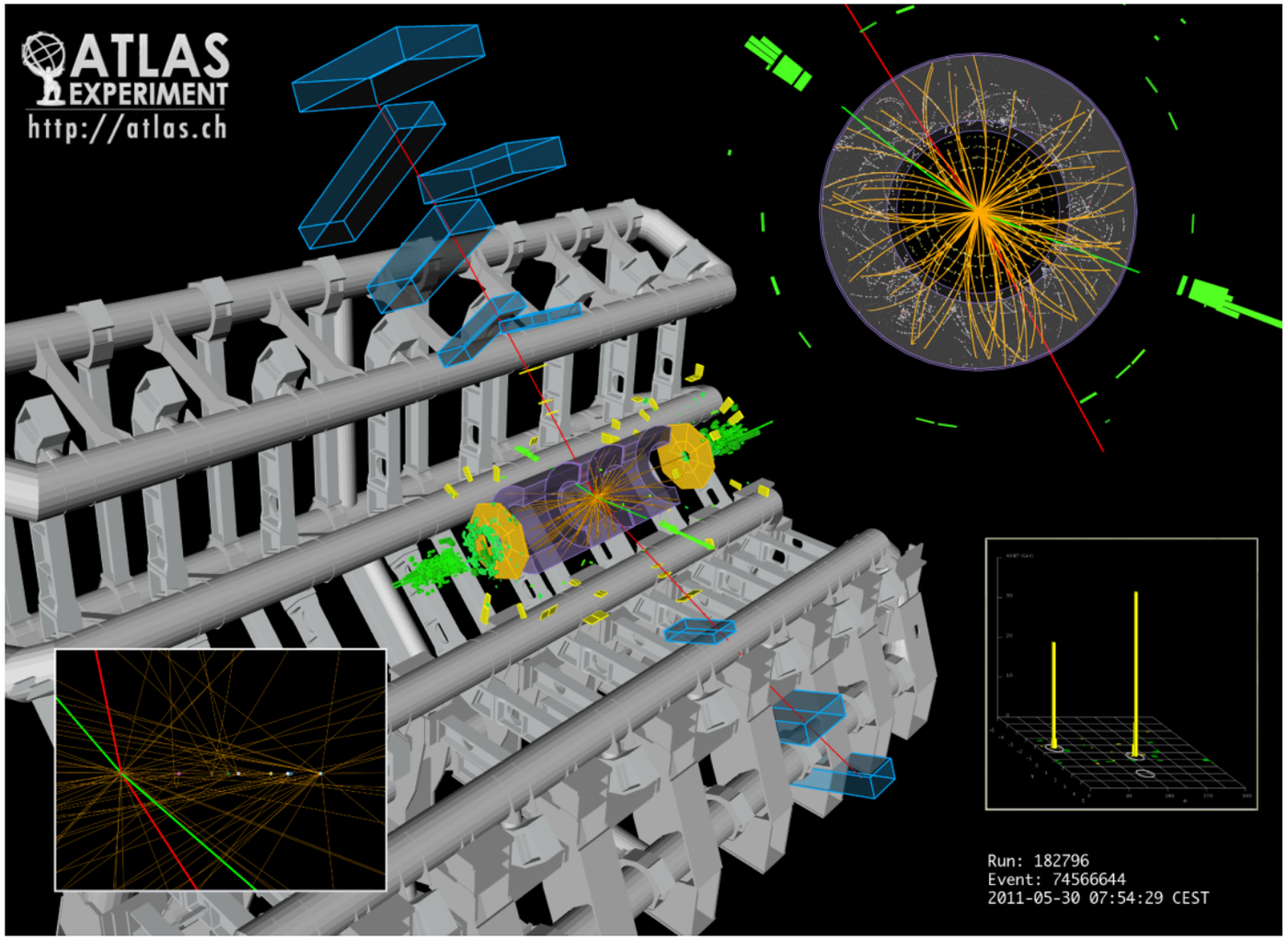






Discovering the Higgs Boson





- Strategy for this channel (there are several) - we look for events with **two Z bosons** that have decayed to **four leptons**, e.g. two electrons and two muons in the event on the right
- If the **two Z bosons** were produced by the **decay of a Higgs boson**, when we reconstruct the invariant mass of the system we should see a **peak** at the **Higgs boson mass**

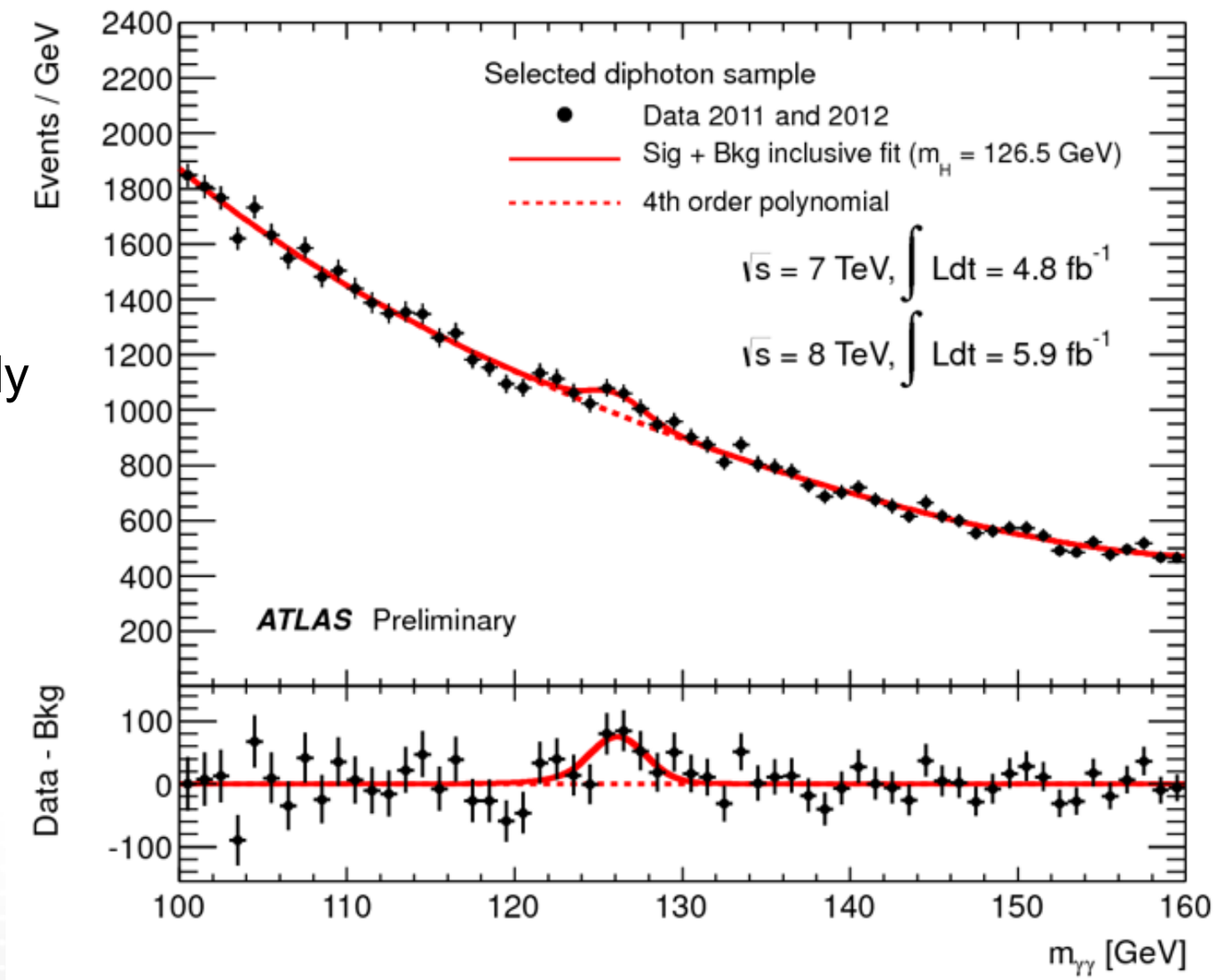
Needles in haystacks

- There are billions of events and the ones we are really interested in are ***very rare***
- Often the interesting events are also ***very difficult to distinguish*** from background
  - Requires ***high precision detectors***, which means ***lots of data*** for each event
- The data are structured but each event is different - ***unique data science challenge***

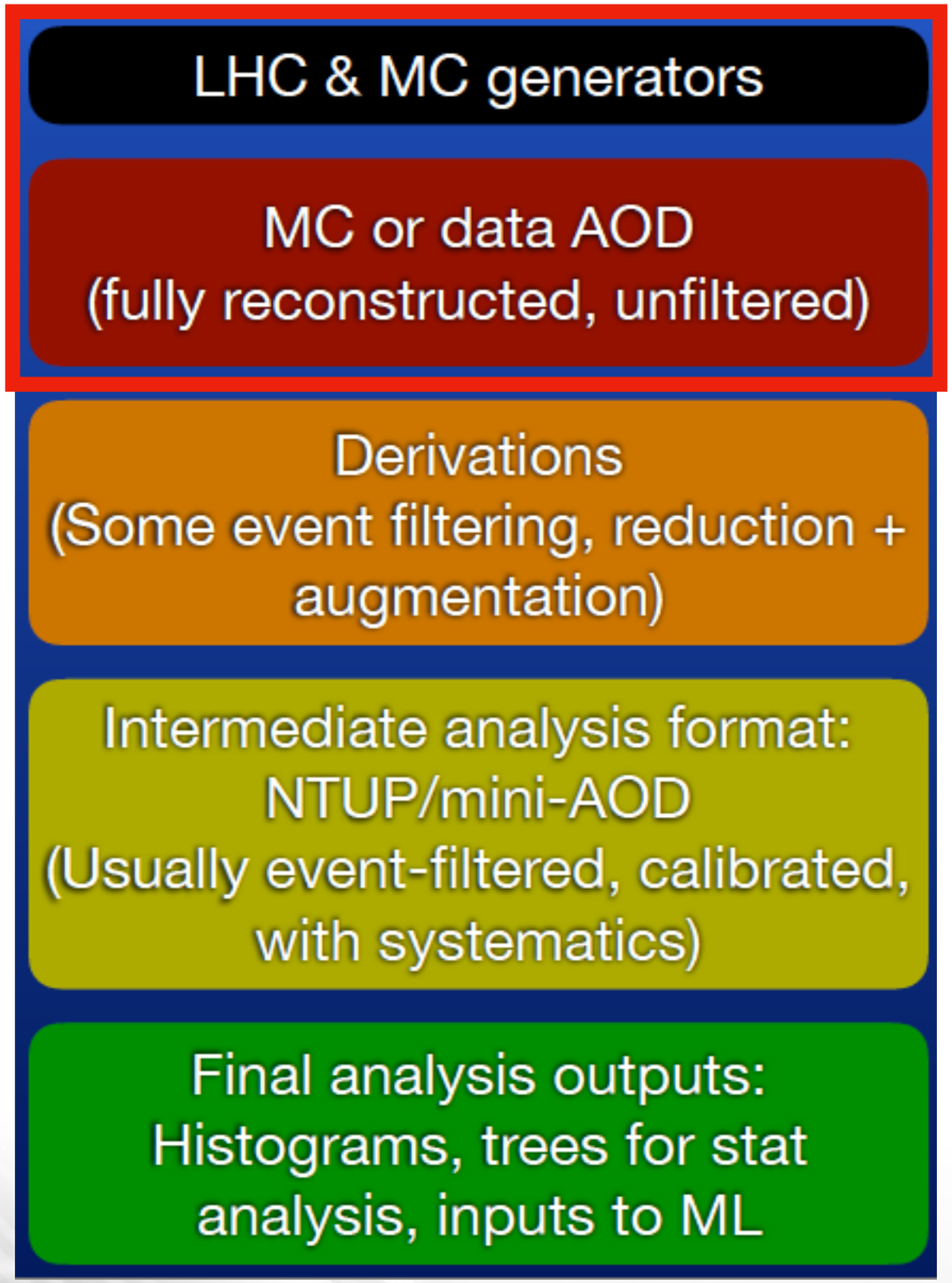
- Data reduction proceeds via a two-pronged approach:

- Select only the events that you are interested in
  - For example, this analysis may only consider events with two photons
- Retain only the information you need
  - In this example the muon information and other physics objects could be thrown away

- Final statistical inference is only performed on the reduced data

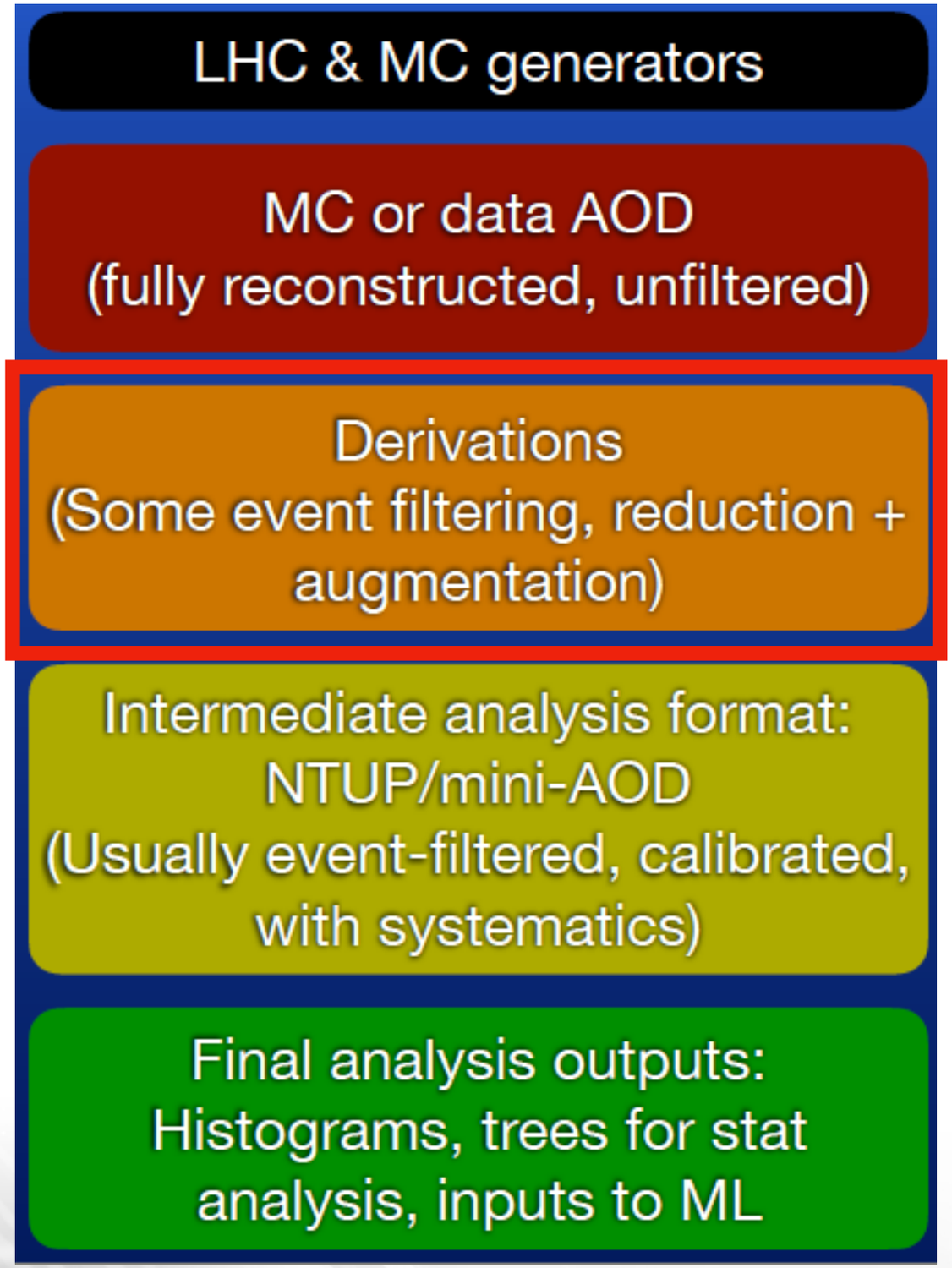


Data analysis workflows



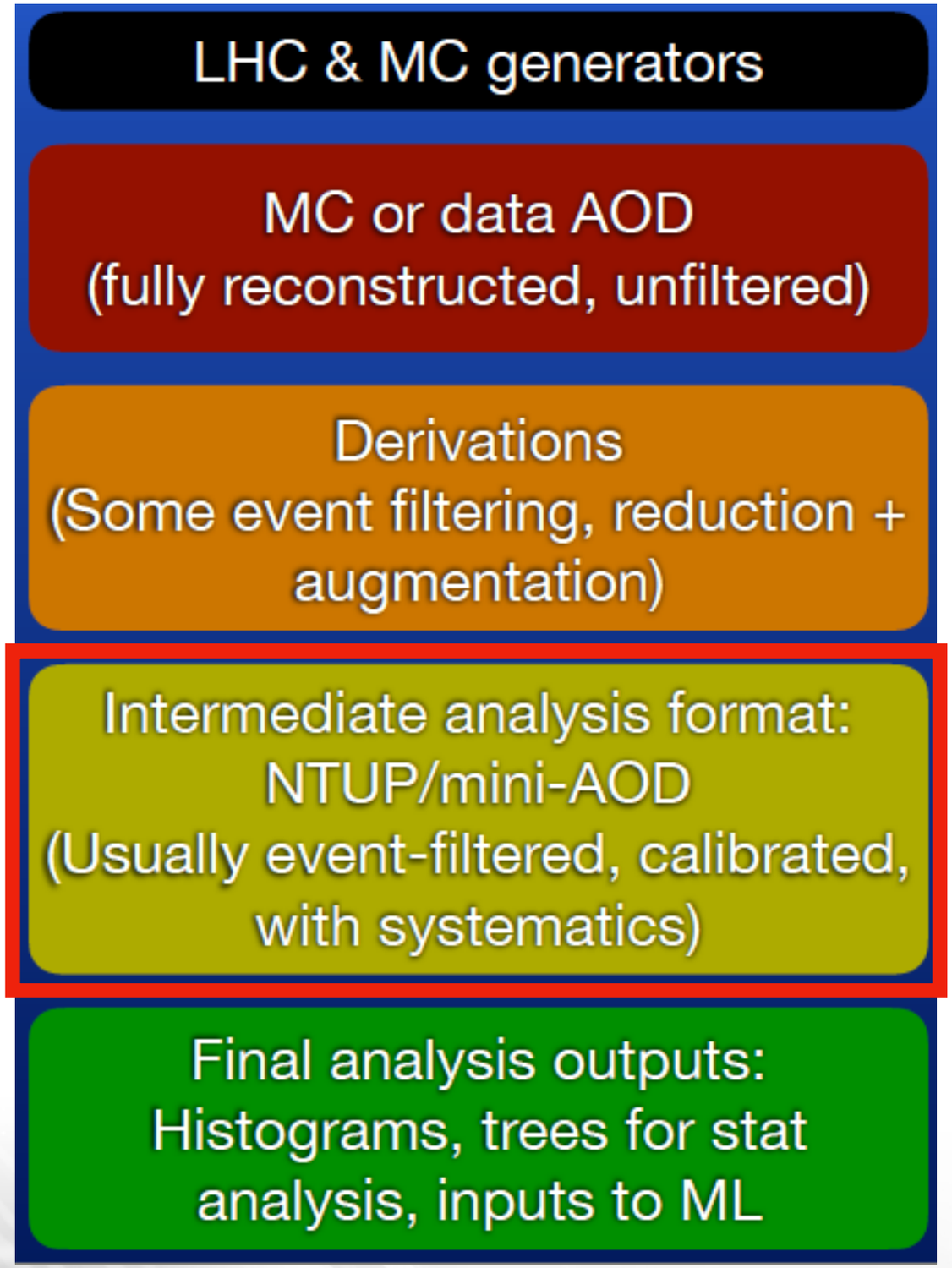
- Typically, event generation, simulation and reconstruction are performed centrally, Data Preparation got us to here.

Data analysis workflows



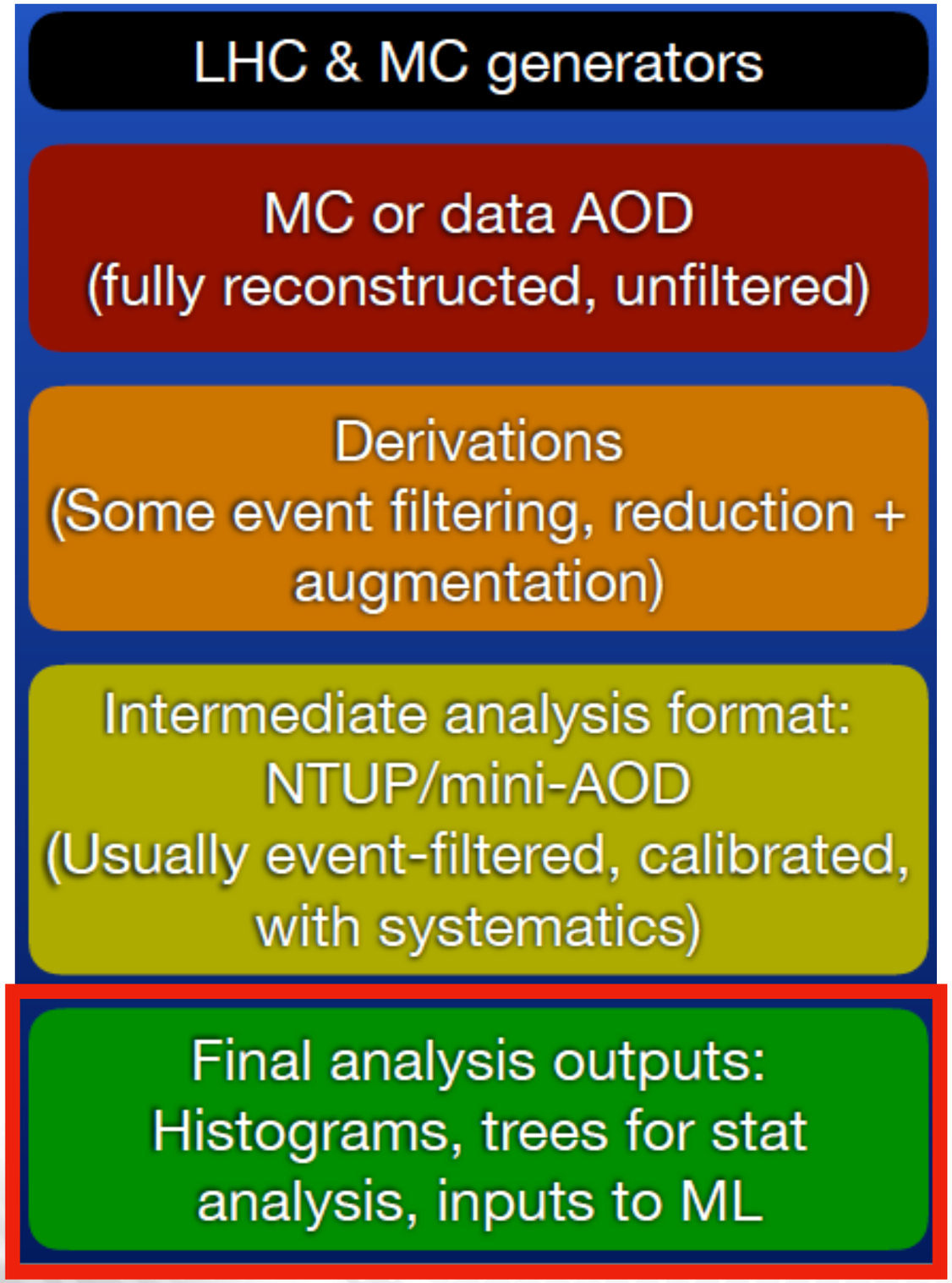
- We then like to make event selections, e.g. select events with two muons. Often we'll calculate some extra variables here, like in the invariant mass of those two muons. We're adding more data per event, but output fewer events. In principle we could throw away some information here, but people are often reluctant to do that

Data analysis workflows



- Next we take our selected events, correct efficiencies and calculate systematic uncertainties

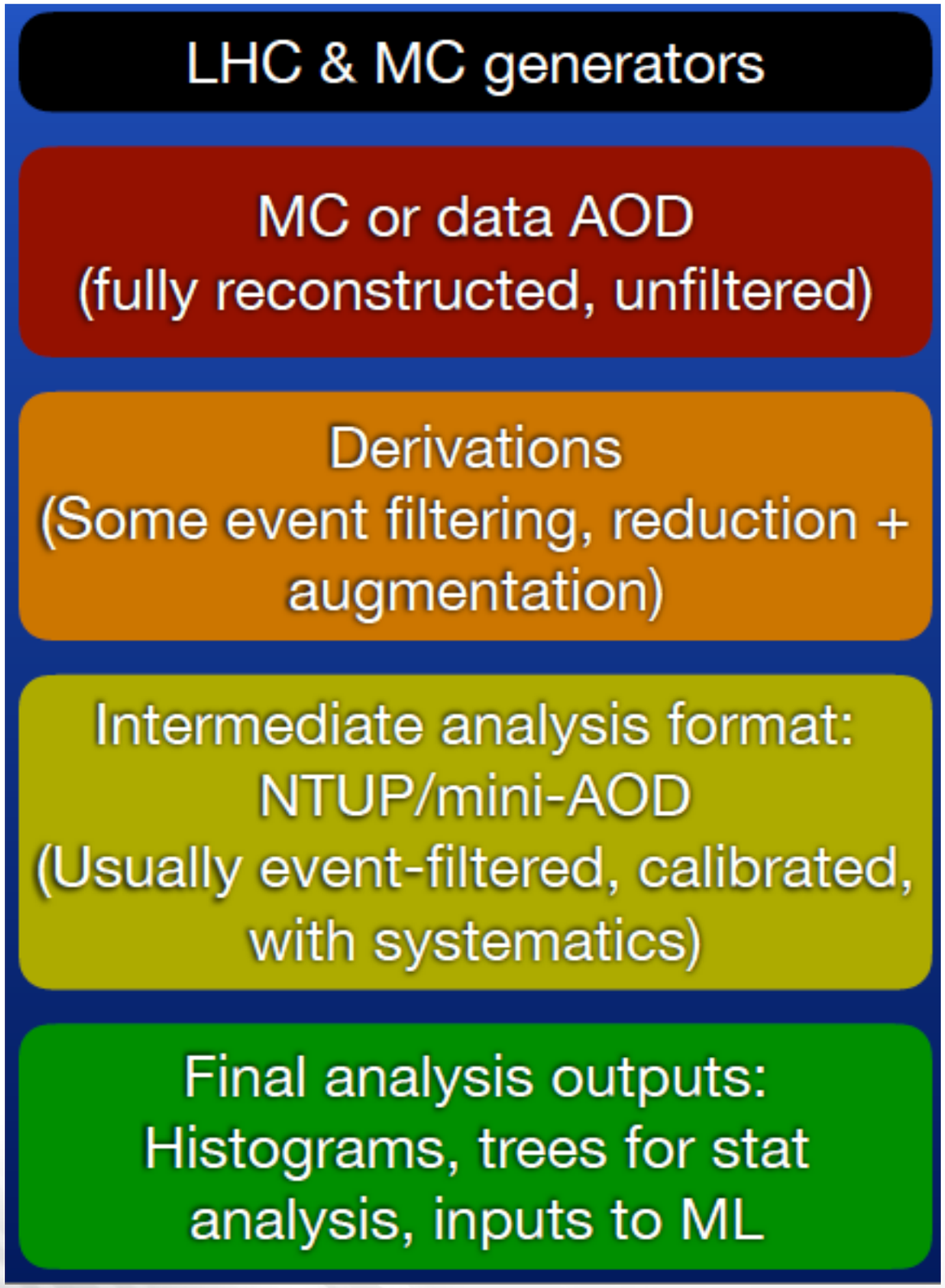
Data analysis workflows



- Finally we extract cross sections (or limits) and produce the final statistical analysis and plots to publish

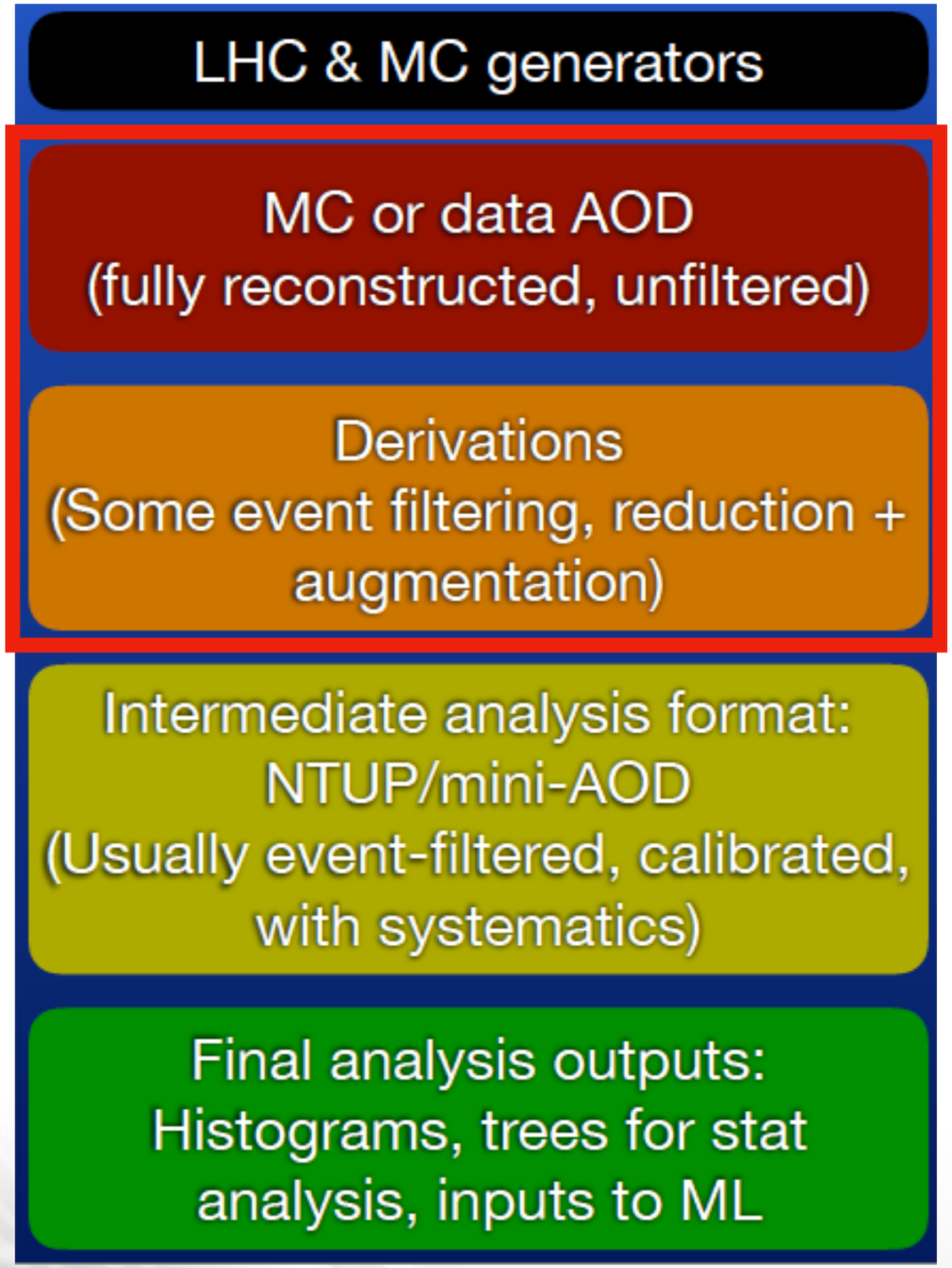

SYAHRUL RAMADAN photography

Data analysis workflows



- This process is a chain
- If we have to redo something in the chain, everything below needs to be redone
- At least once
  - Usually more than that
- The reality is that this happens (too) often

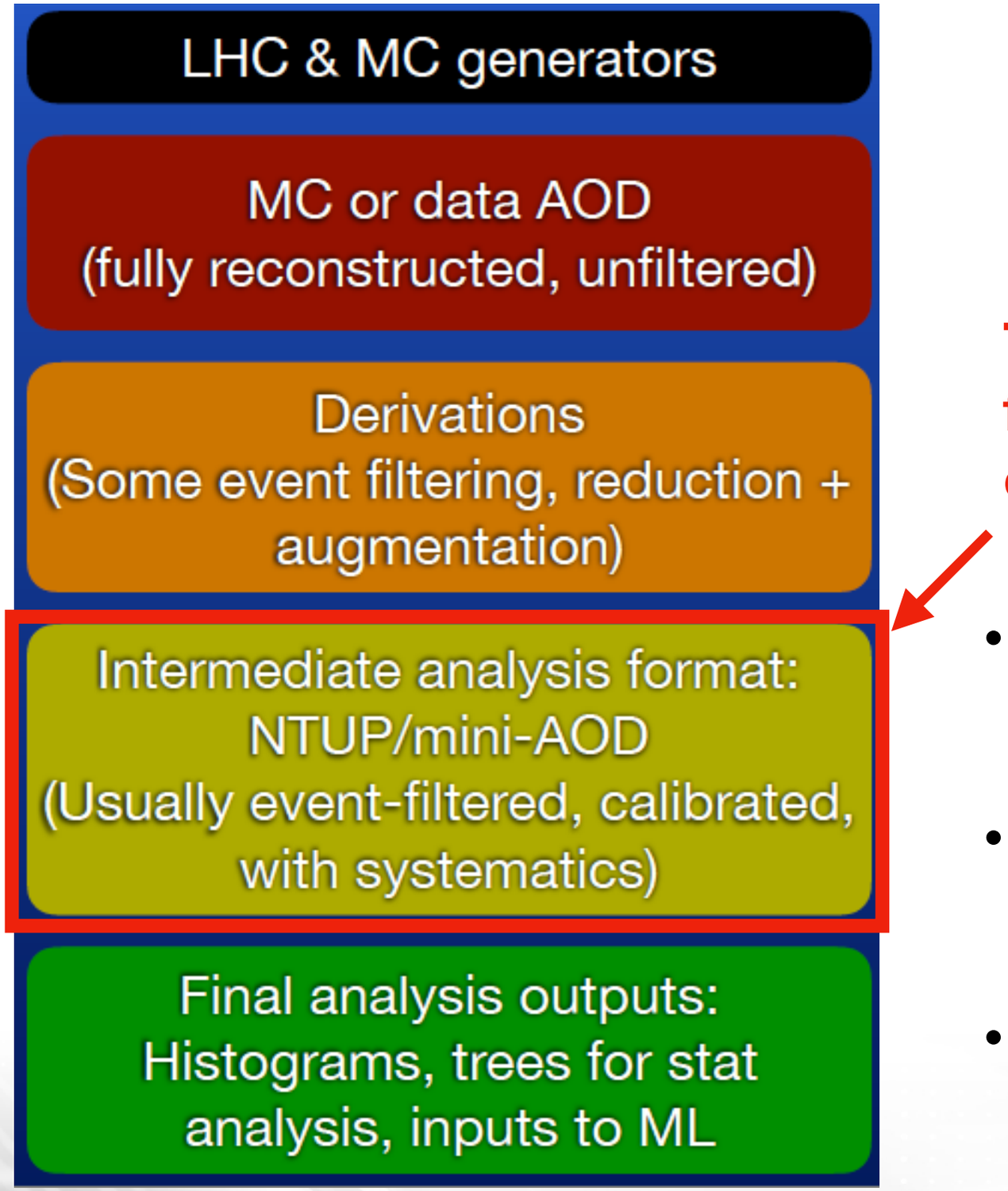
Data analysis workflows



- Reiterating steps at this level means improving calibrations and reconstruction algorithms
- This reduces systematic uncertainties and improves the final measurement

This data is petabyte-scale for one analysis

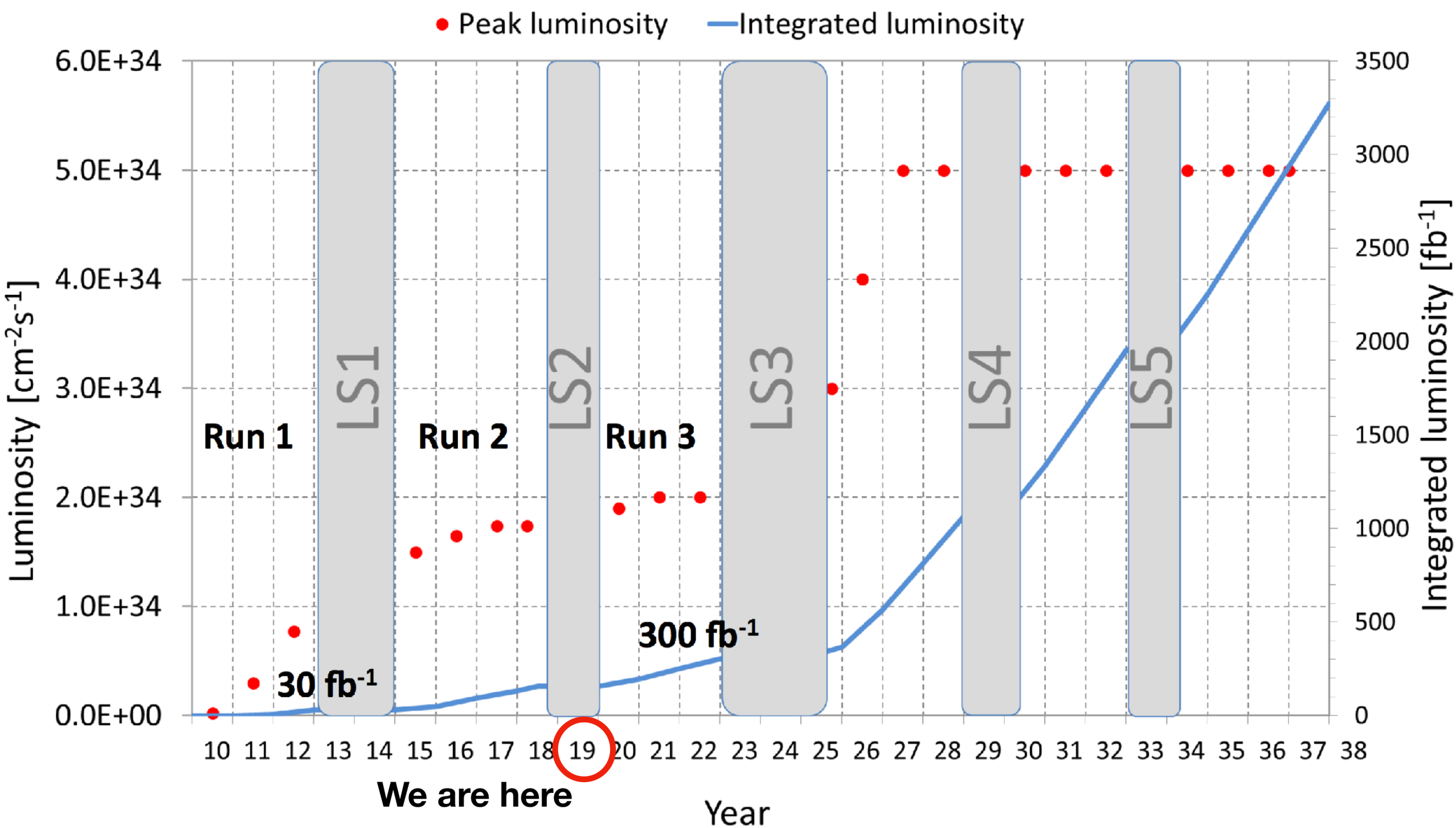
Data analysis workflows

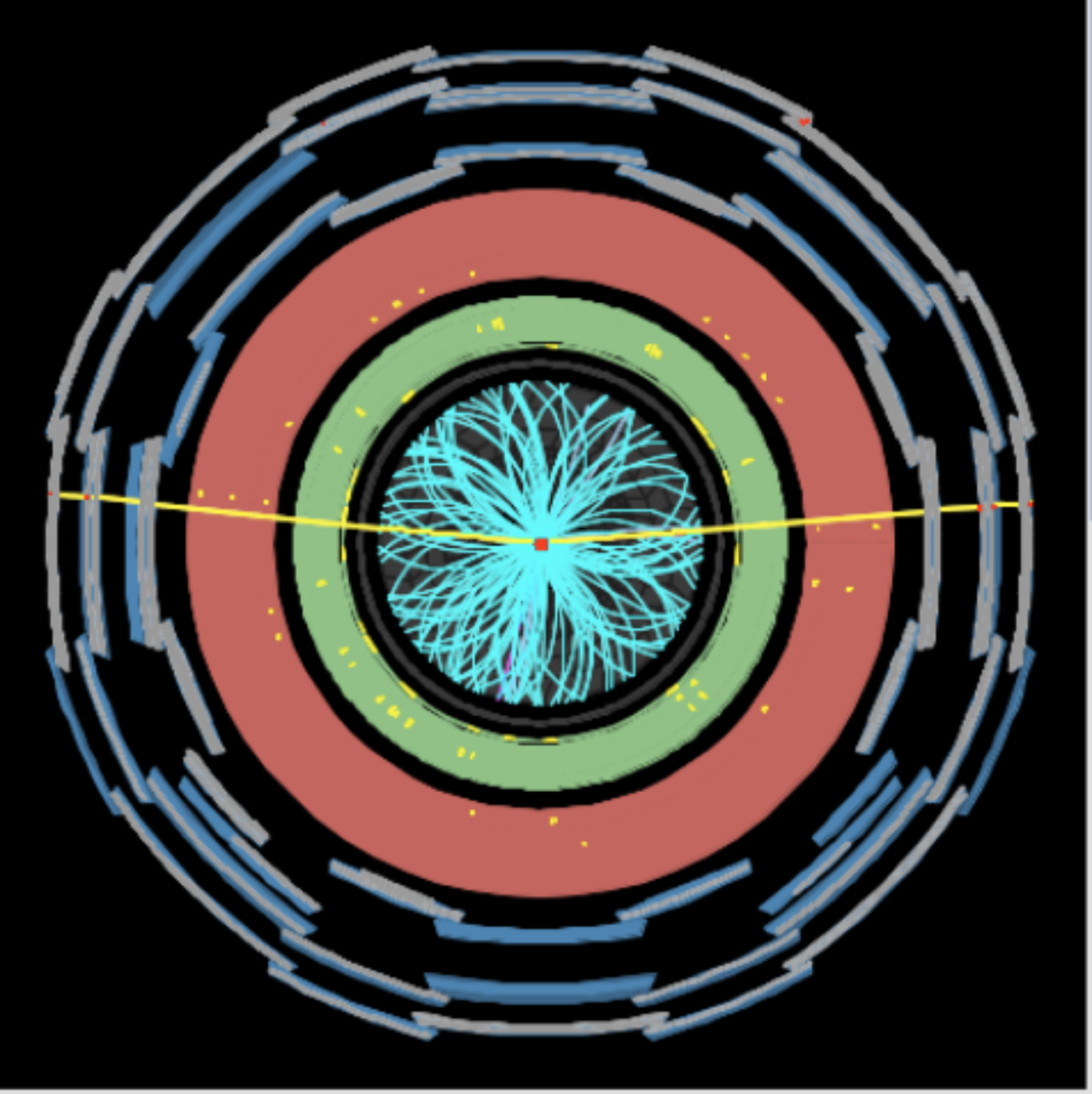


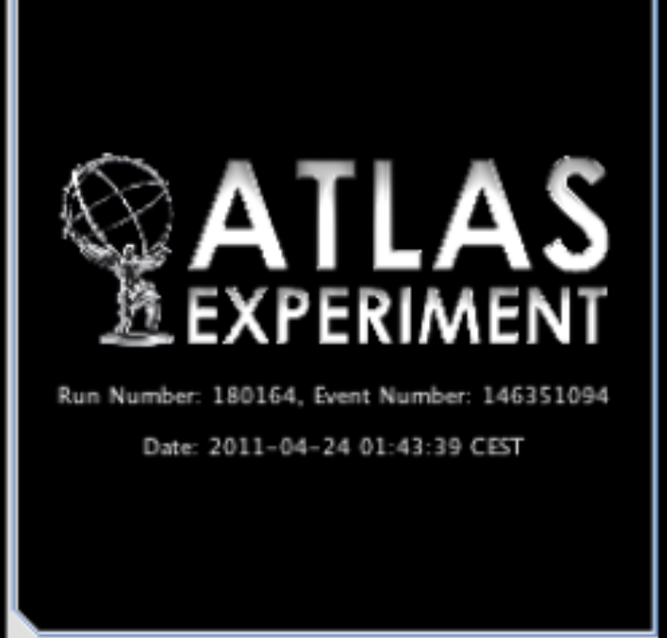
This data is terabyte-scale for one analysis, final analysis outputs are gigabytes

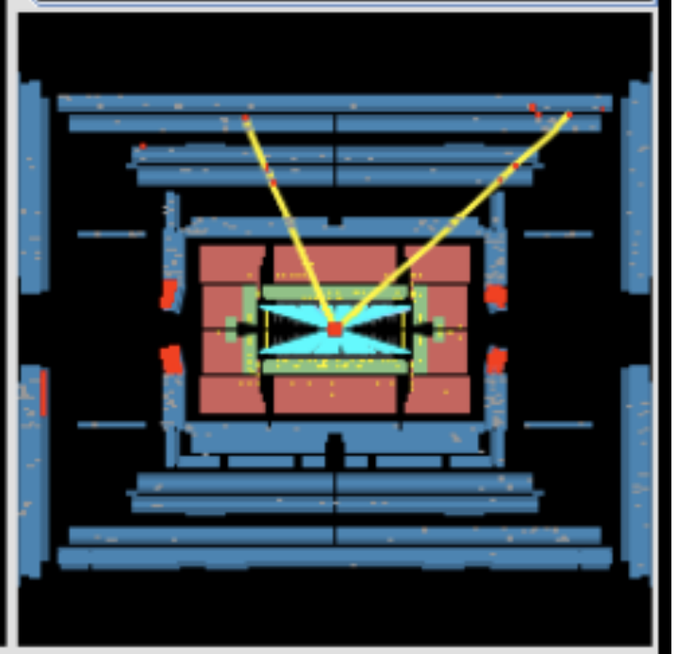
- Reiterating steps here is also physics analysis, optimising the final algorithms
- Limited by the information available, hence the reluctance to throw information away
- Sometimes the conclusion is we need to rerun a step higher up the chain, e.g. Derivations

The long game







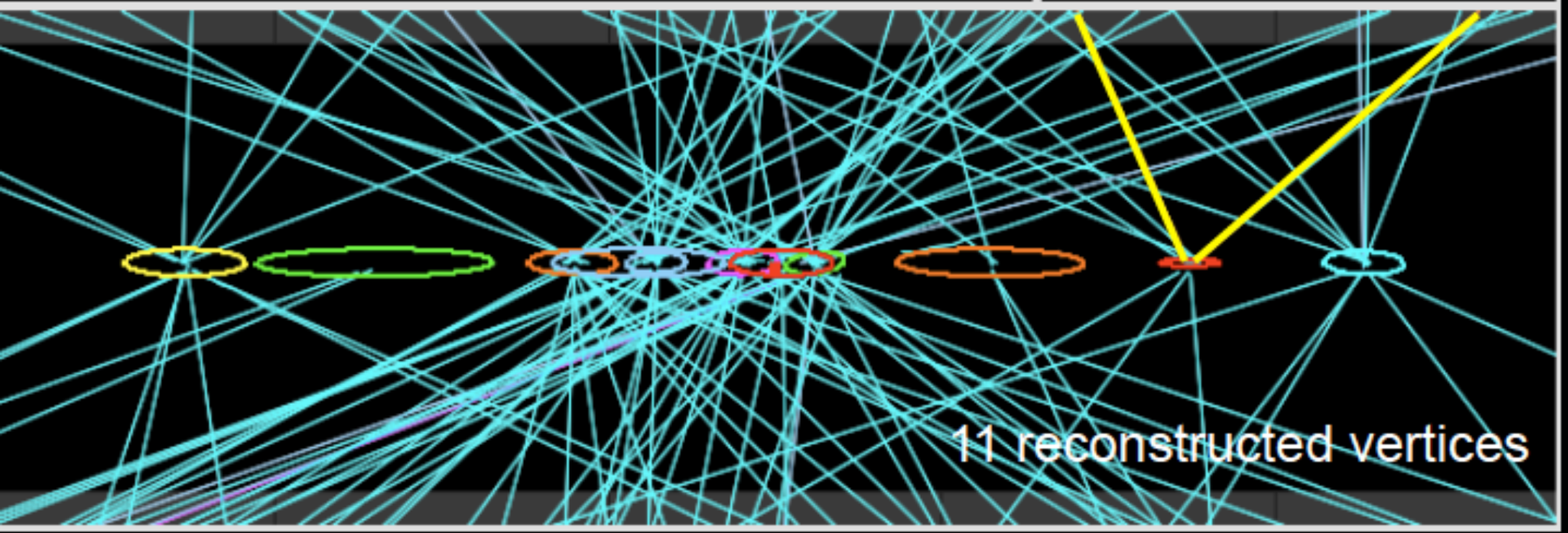


This event is taken from 2011 data

In 2017 the average number of collisions was ~40

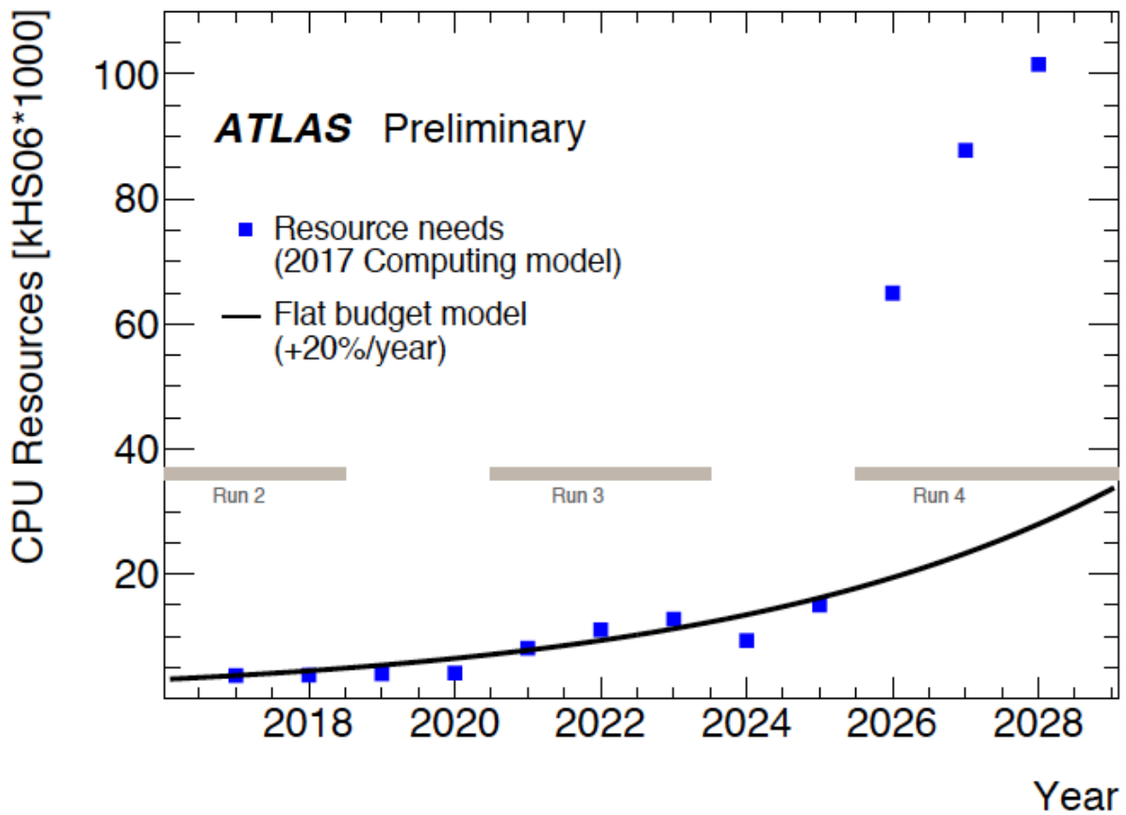
For HL-LHC this could be as high as 200 collisions

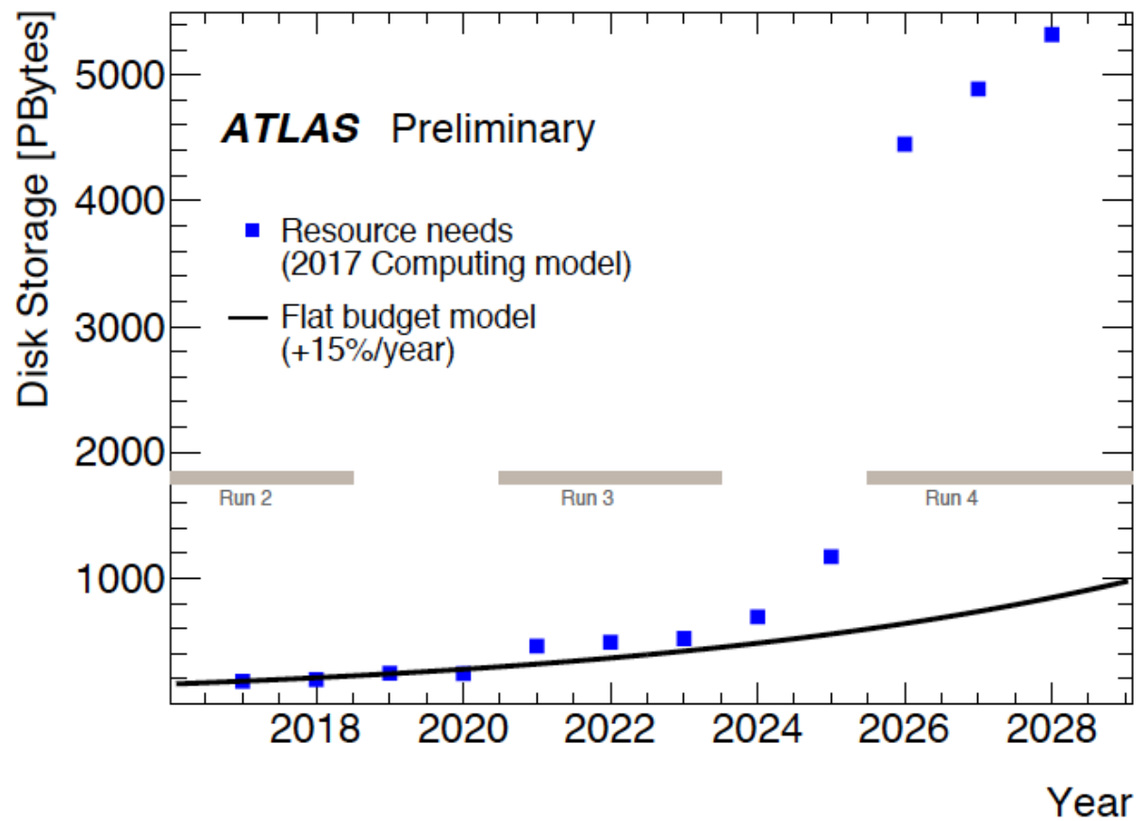
That costs a lot more CPU and improved algorithms



Track pT > 0.5 GeV

LHC future computing

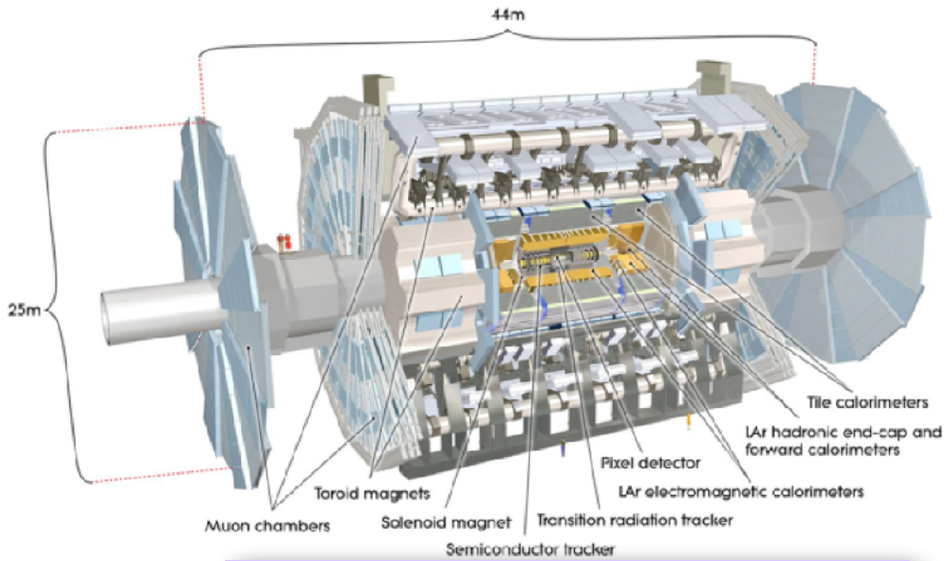


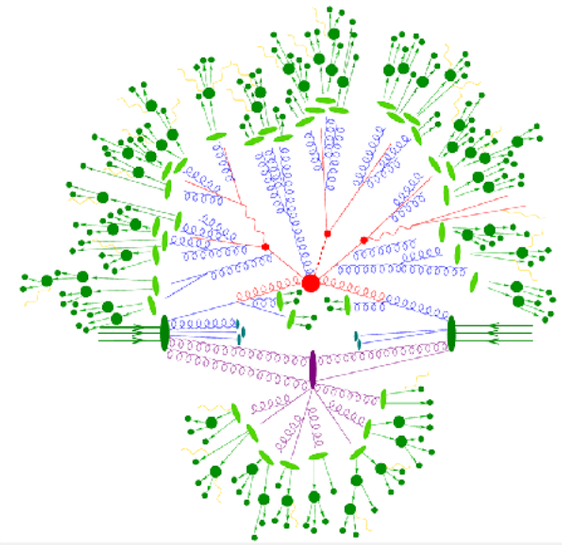


The data challenge is going to get a lot harder than standard progress can handle, c.f. HEP Software Foundation paper on software and computing R&D for the 2020s:

http://arxiv.org/pdf/1712.06982.pdf

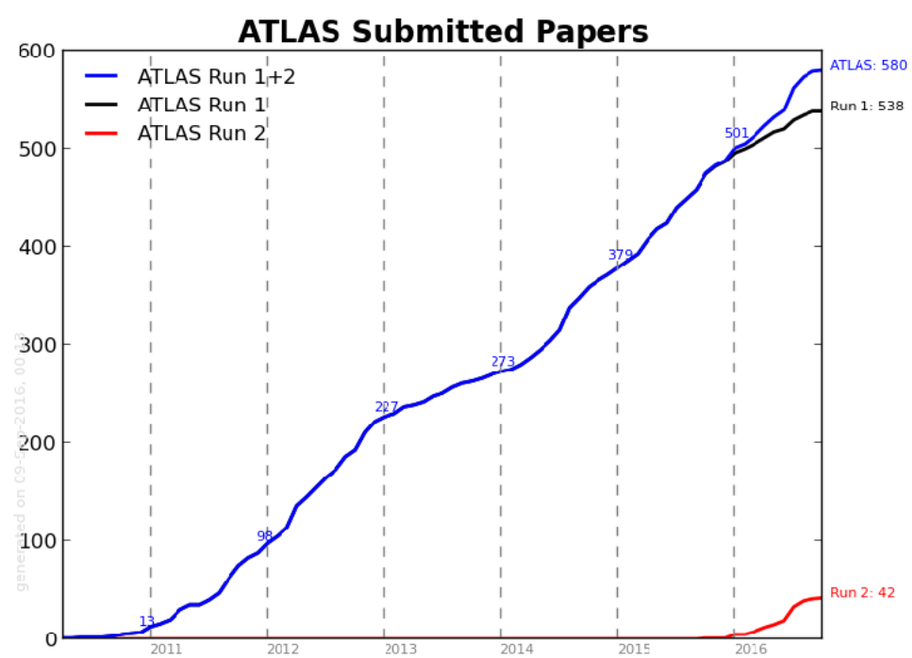
Exabyte-scale physics analysis





Exabytes of Data

Exabytes of Simulation



Publish!

Physics


"WANTS
YOU"
JOIN YOUR COUNTRY'S ARMY!
GOD SAVE THE KING
Reproduced by permission of LONDON OPINION
Printed by the Victoria House Printing Co., Ltd., Tudor Street, London, E.C.