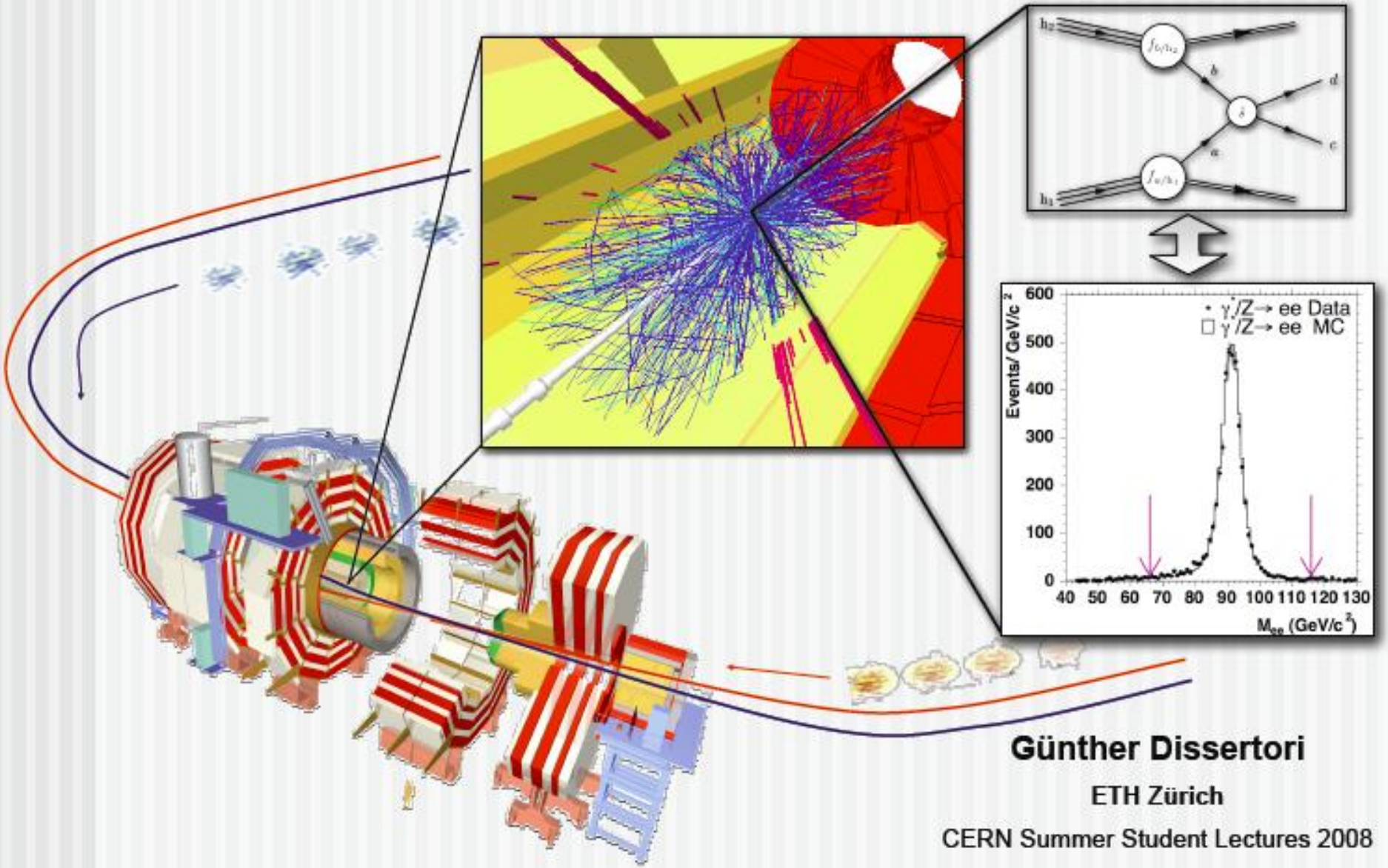
Lecture 5

Computing for HEP experiments

From RAW data
to Physics Analysis

Based on lectures given by G. Dissertori
CERN Summer Students Lectures
July 2008

From Raw Data to Physics Results

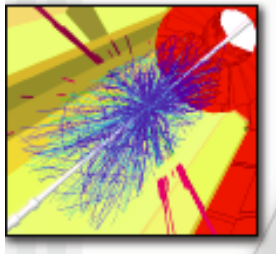


Günther Dissertori

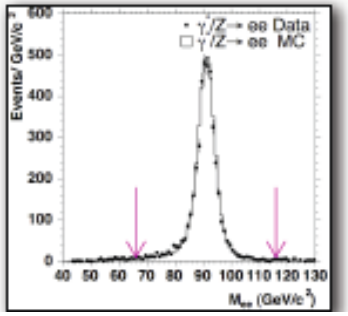
ETH Zürich

CERN Summer Student Lectures 2008

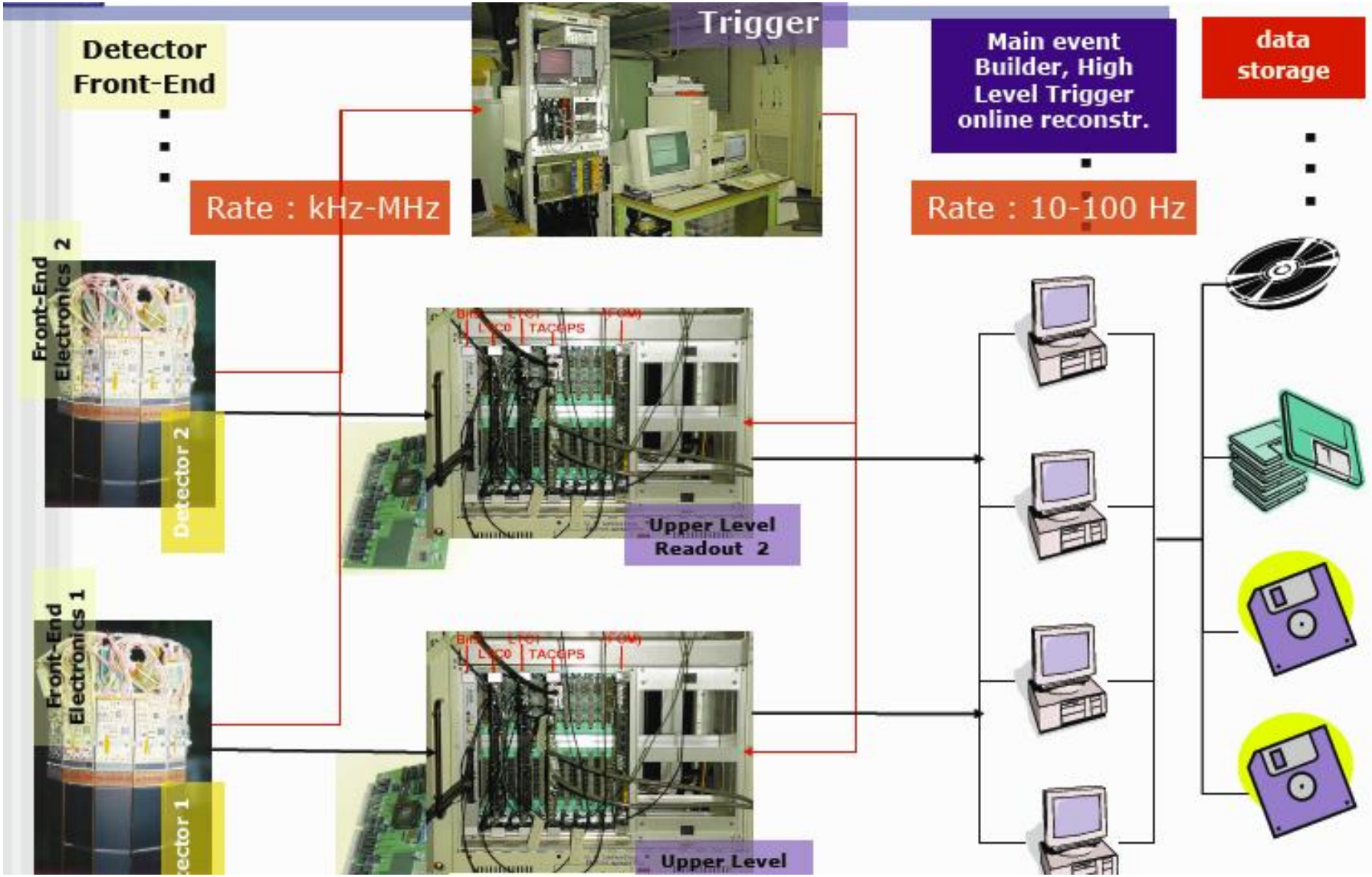
Data analysis chain



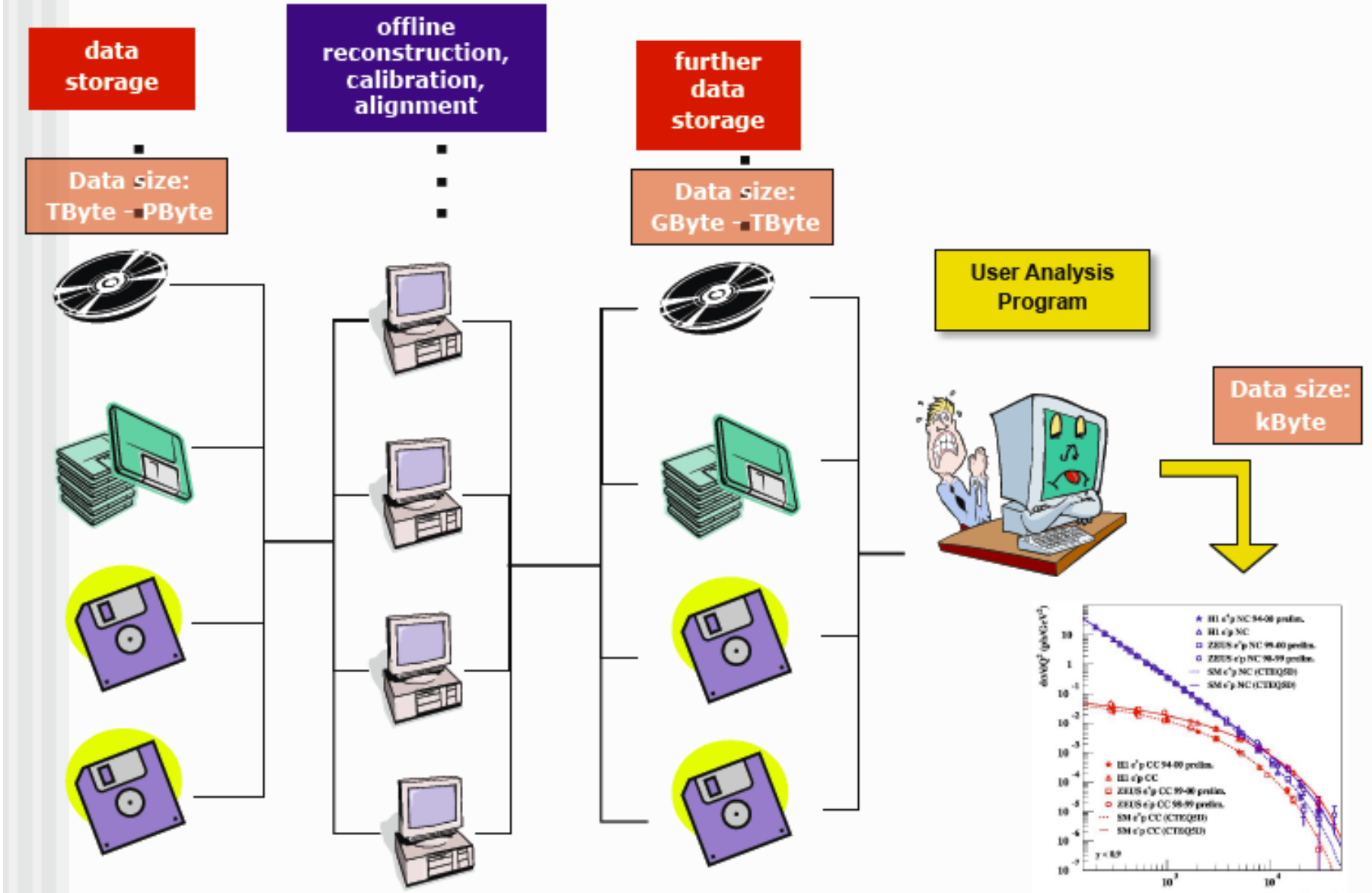
- Have to collect data from many channels on many sub-detectors (millions)
- Decide to read out everything or throw event away (Trigger)
- Build the event (put info together)
- Store the data
- Analyze them
  - reconstruction, user analysis algorithms, data volume reduction
- do the same with a simulation
  - correct data for detector effects
- Compare data and theory



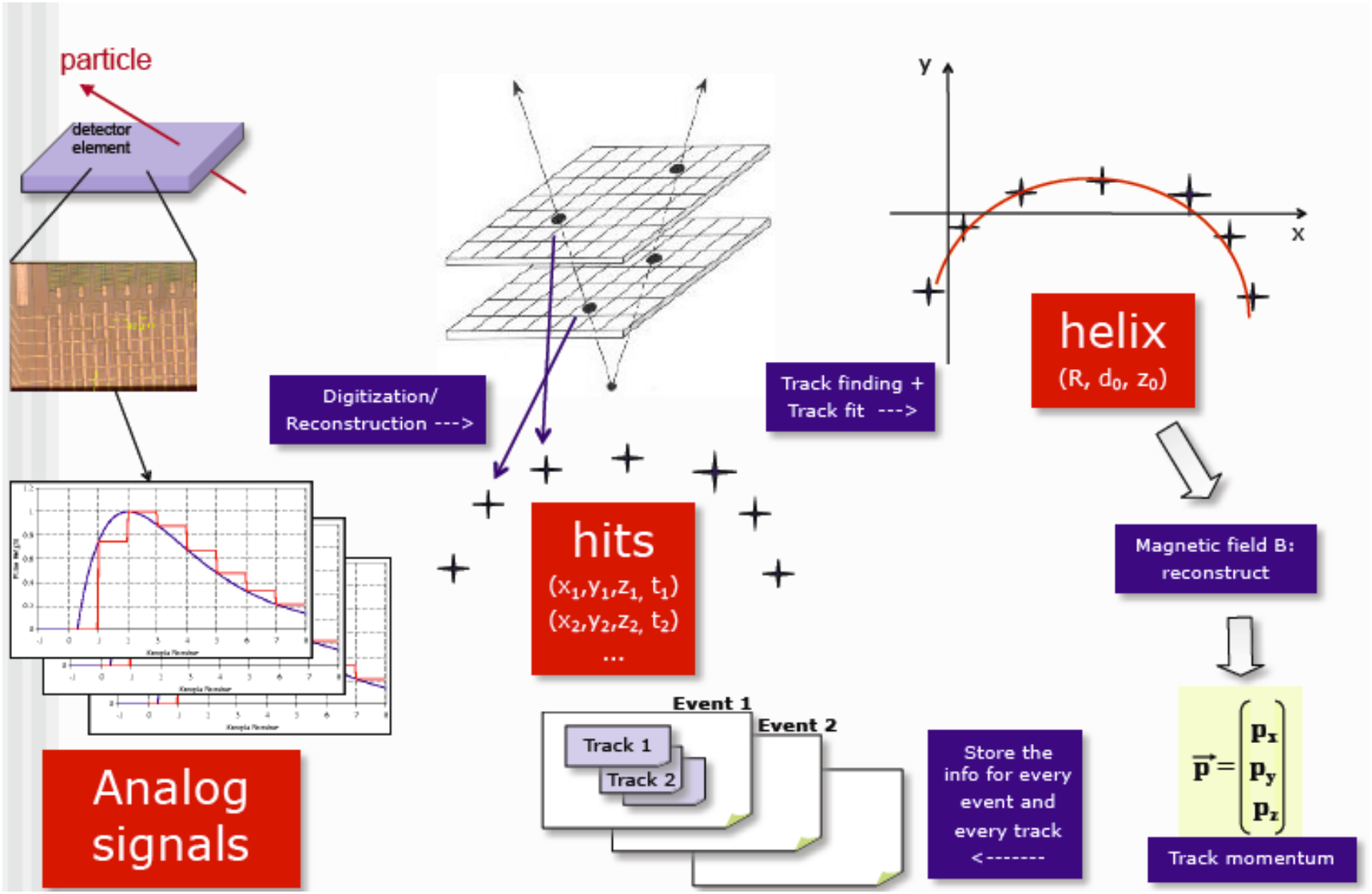
DAQ chain


Detector
Front-End
Trigger
Main event
Builder, High
Level Trigger
online reconstr.
data
storage
Rate : kHz-MHz
Rate : 10-100 Hz
Front-End
Electronics 2
Detector 2
Upper Level
Readout 2
Front-End
Electronics 1
Upper Level

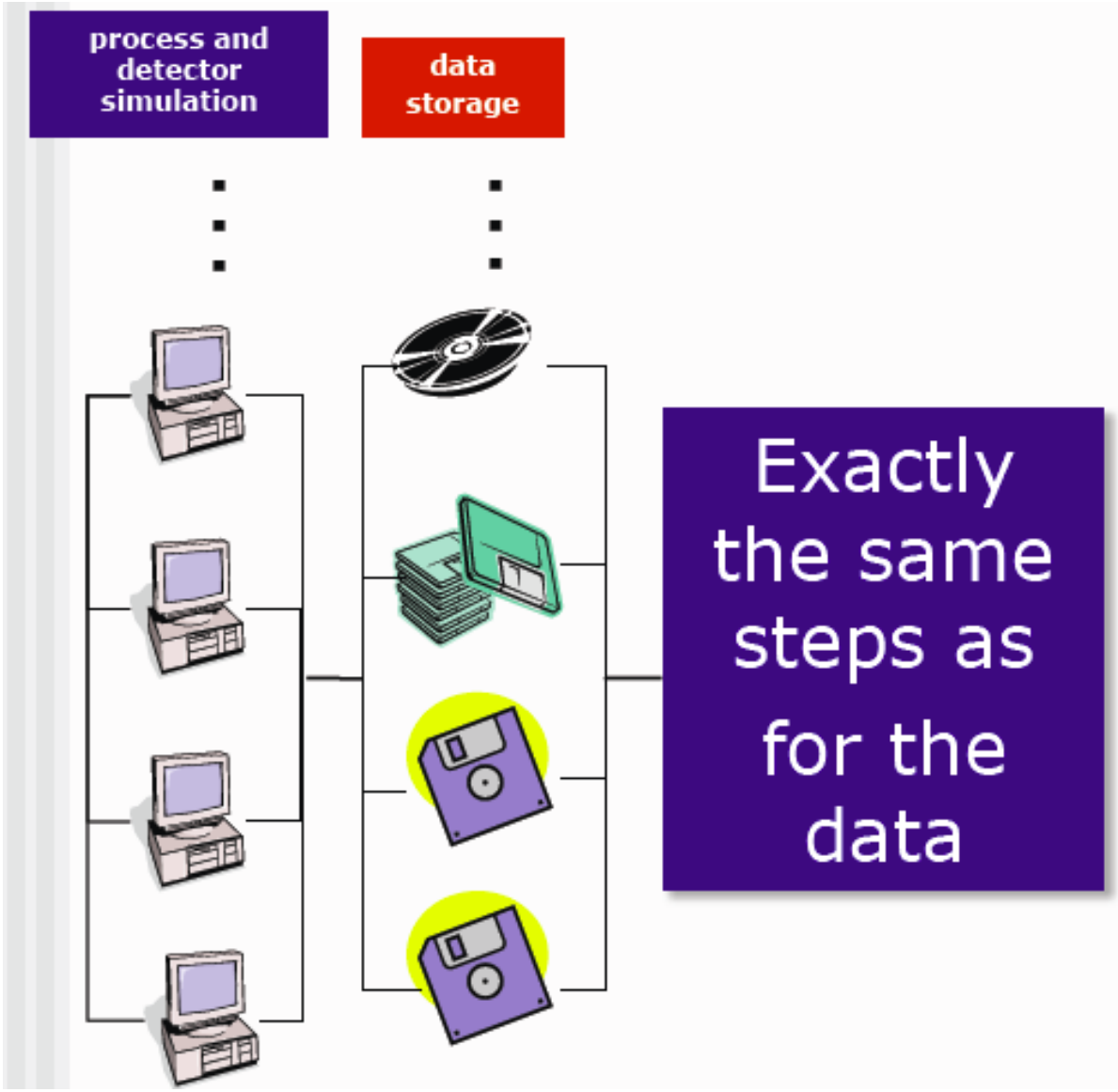
Offline analysis chain



Data reduction/abstraction



Simulation



Simulation of many (millions) of events

- simulate physics process
  e.g. $e^+e^- \rightarrow$ hadrons
  or p p $\rightarrow$ jets
- plus the detector response to the produced particles
- understand detector response and analysis parameters (lost particles, resolution, efficiencies, backgrounds)
- and compare to real data
- Note : simulations present from beginning to end of experiment, needed to make design choices

Our task

We use experiments to inquire about what "reality" (nature) does

We intend to fill this gap

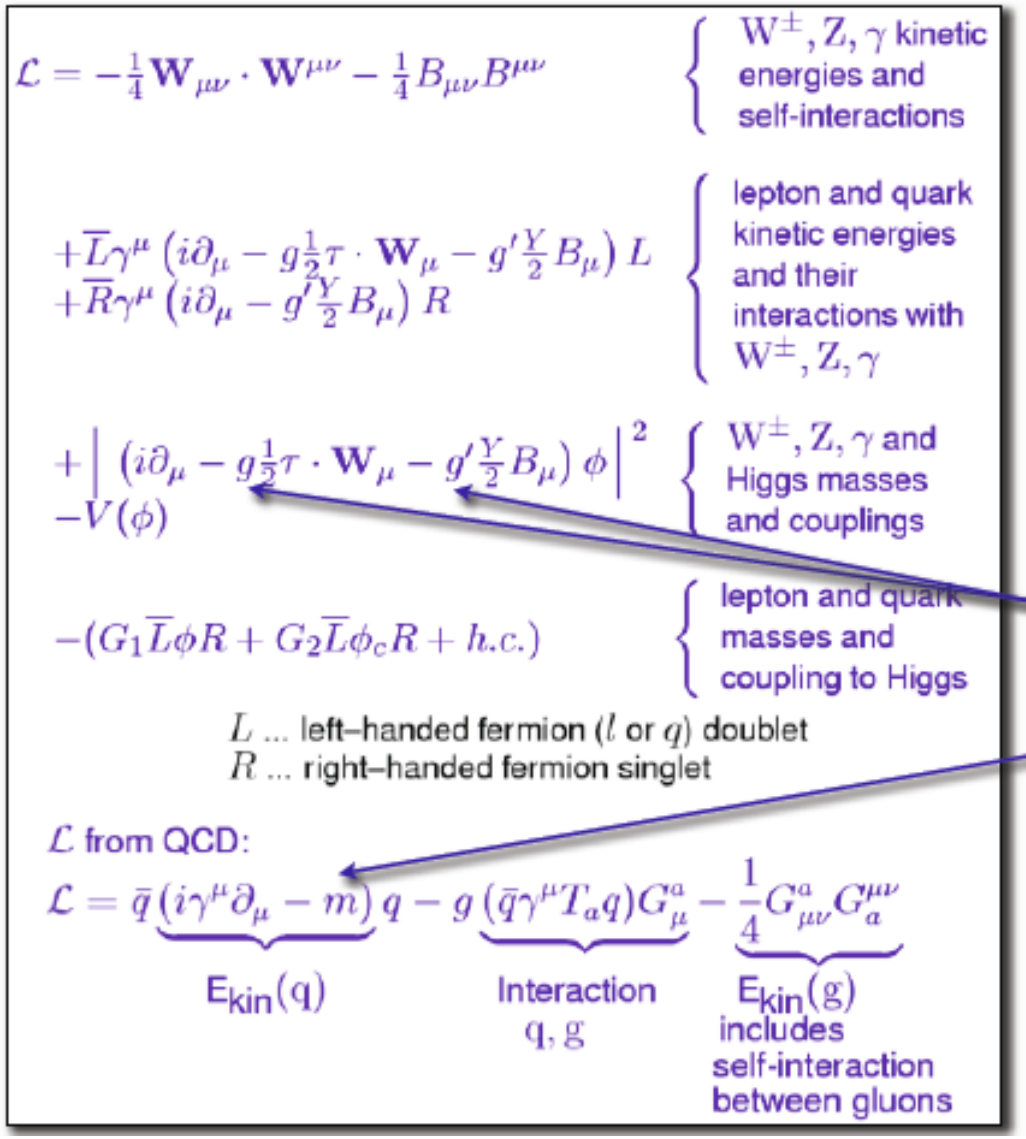
Theory

$$S = \mathrm{i} \int \mathrm{d}^4 x \, \mathcal{L}(x)$$

The goal is to understand in the most general; that's usually also the simplest.

- A. Eddington

Theory



eg.
the Standard Model

has parameters

coupling constants

masses

predicts:
cross sections,
branching ratios,
lifetimes, ...

Experiment

eg.

1/30th of an event in the BaBar detector

- get about 100 evts/sec

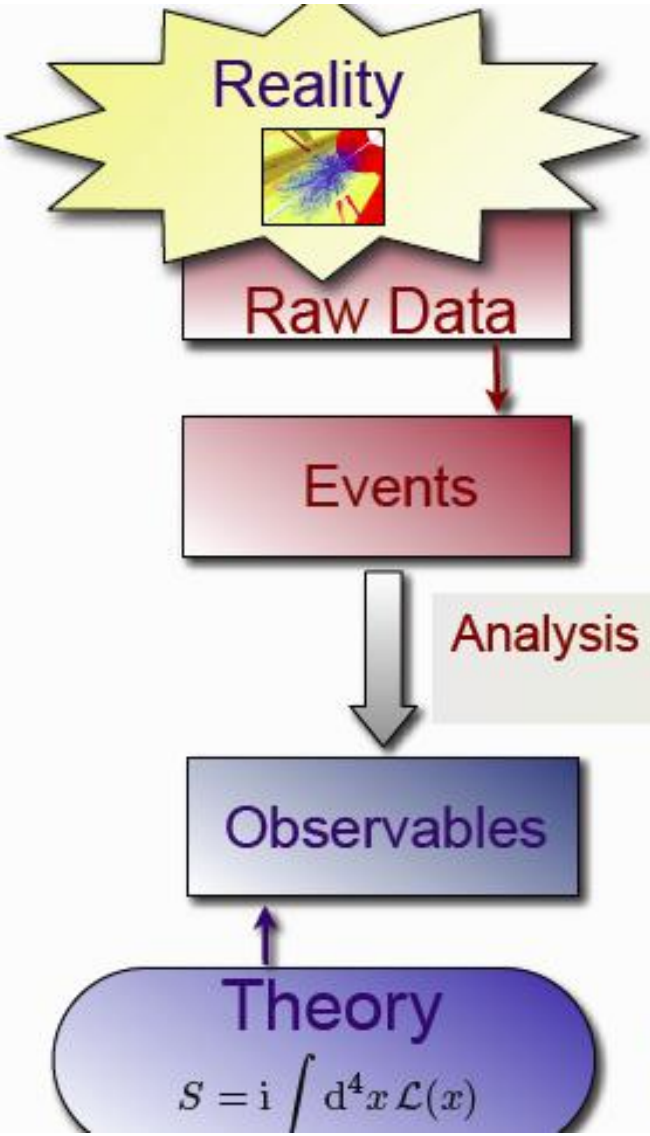
"Address" :

- which detector element took the reading

"Value(s)" :

- what the electronics wrote out

Making the connection



The imperfect measurement of a (set of) interactions in the detector

A unique happening:
eg. Run 23458, event 1345
which contains a $Z \rightarrow \mu^+\mu^-$ decay

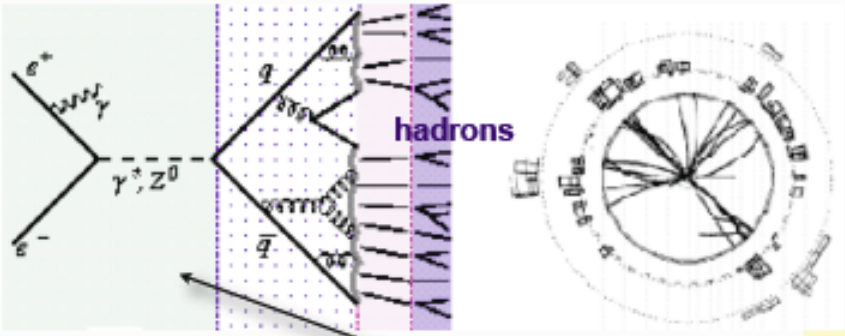
Analysis : We "confront theory with experiment" by comparing the measured quantity (observable) with the prediction.

cross sections (probabilities for interactions), branching ratios (BR), ratios of BRs, specific lifetimes, ...

A small number of general equations, with some parameters (poorly or not known at all)

A simple example

- Measurement of e^+e^- annihilation into hadrons and muons:



Hadronic final state

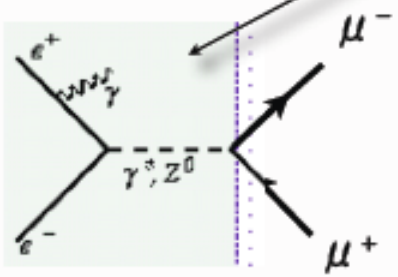
- many charged tracks (>~ 10)
- sum of energy deposits in calorimeters not too far from centre-of-mass energy

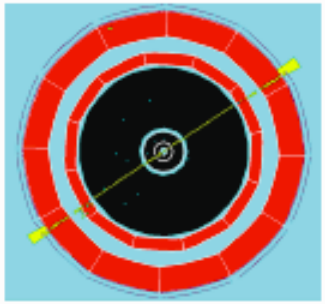
sum over all quark flavours, which can be produced at a certain e^+e^- centre-of-mass energy E_{CM}, ,eg. d, u, s, c, b, t

$$R := \frac{\sigma(e^+e^- \rightarrow q_f \bar{q}_f)}{\sigma(e^+e^- \rightarrow \mu^+\mu^-)} = N_c \sum_f z_f^2$$

Number of colours

electric charges of quarks, in units of electron charge



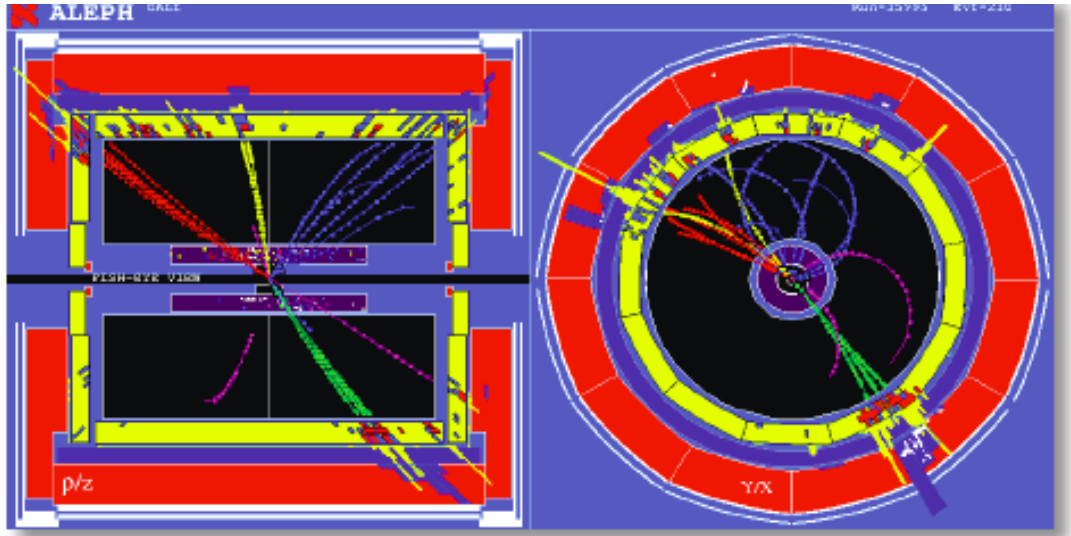


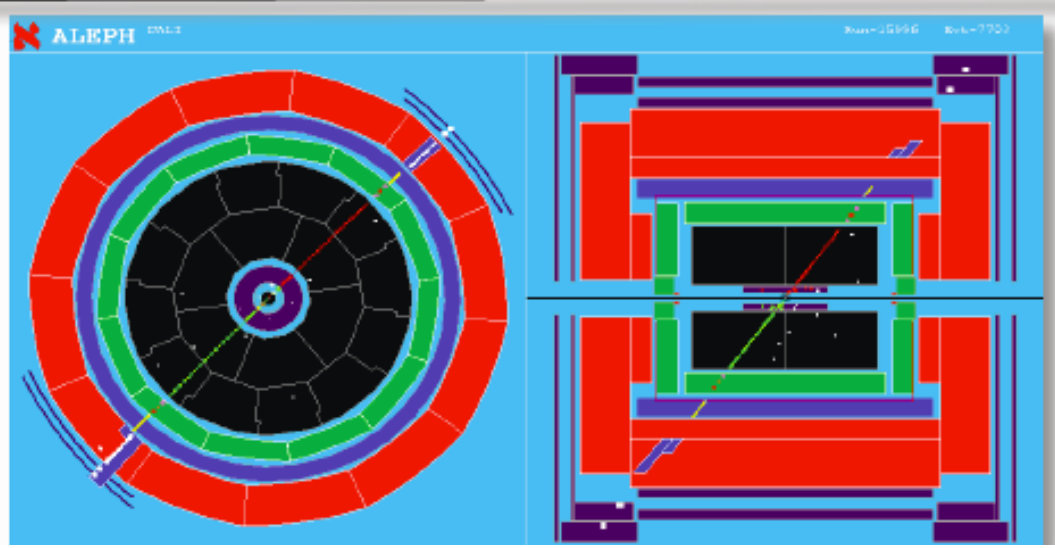
Muonic final state

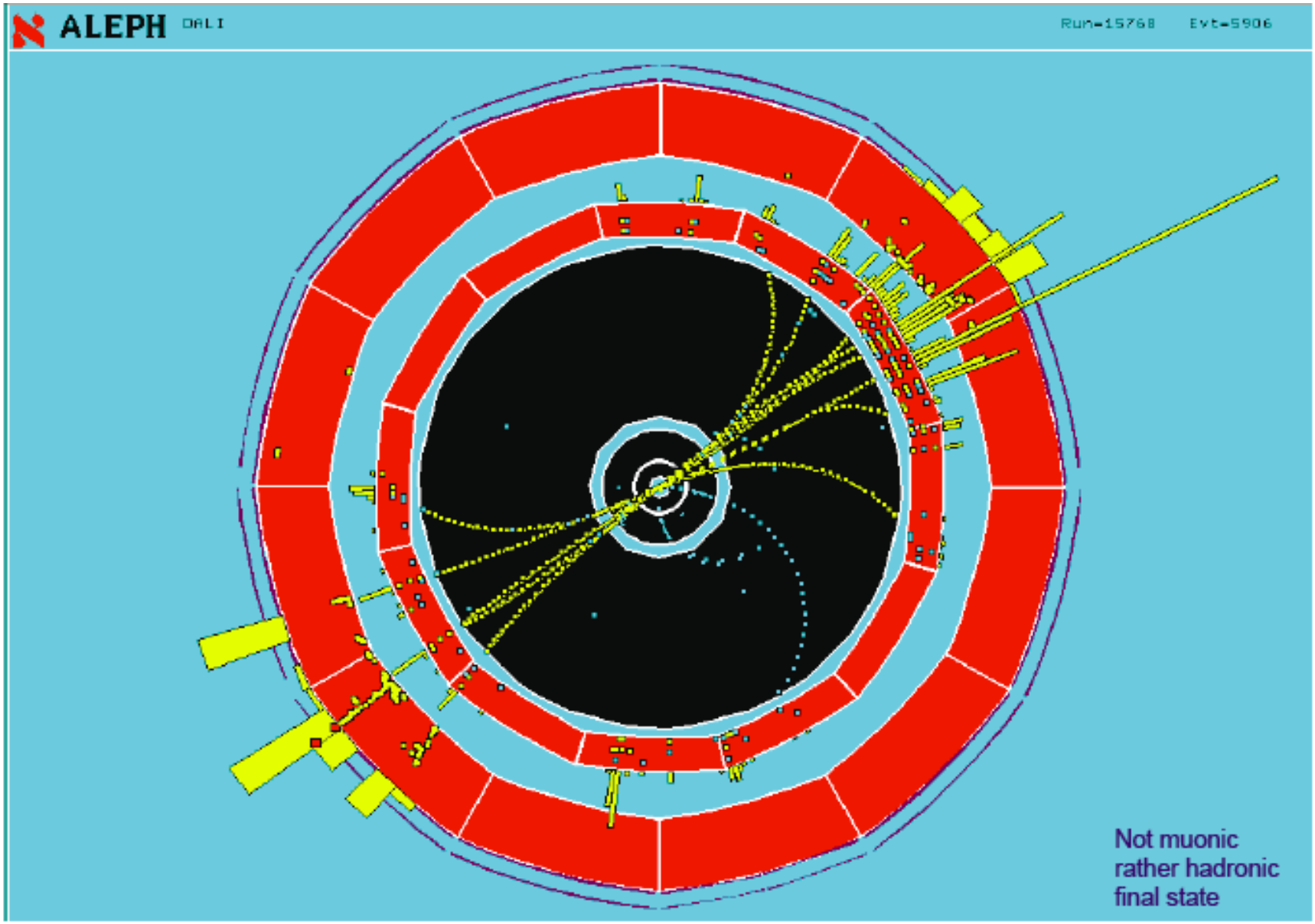
- two charged tracks, approx. back-to-back, with expected momentum (~ 1/2 E_{CM})
- right number of muon hits in outer layers (muons very penetrating, traverse whole detector)
- expected energy in calorimeter (electrons deposit all their energy, muons leave little)

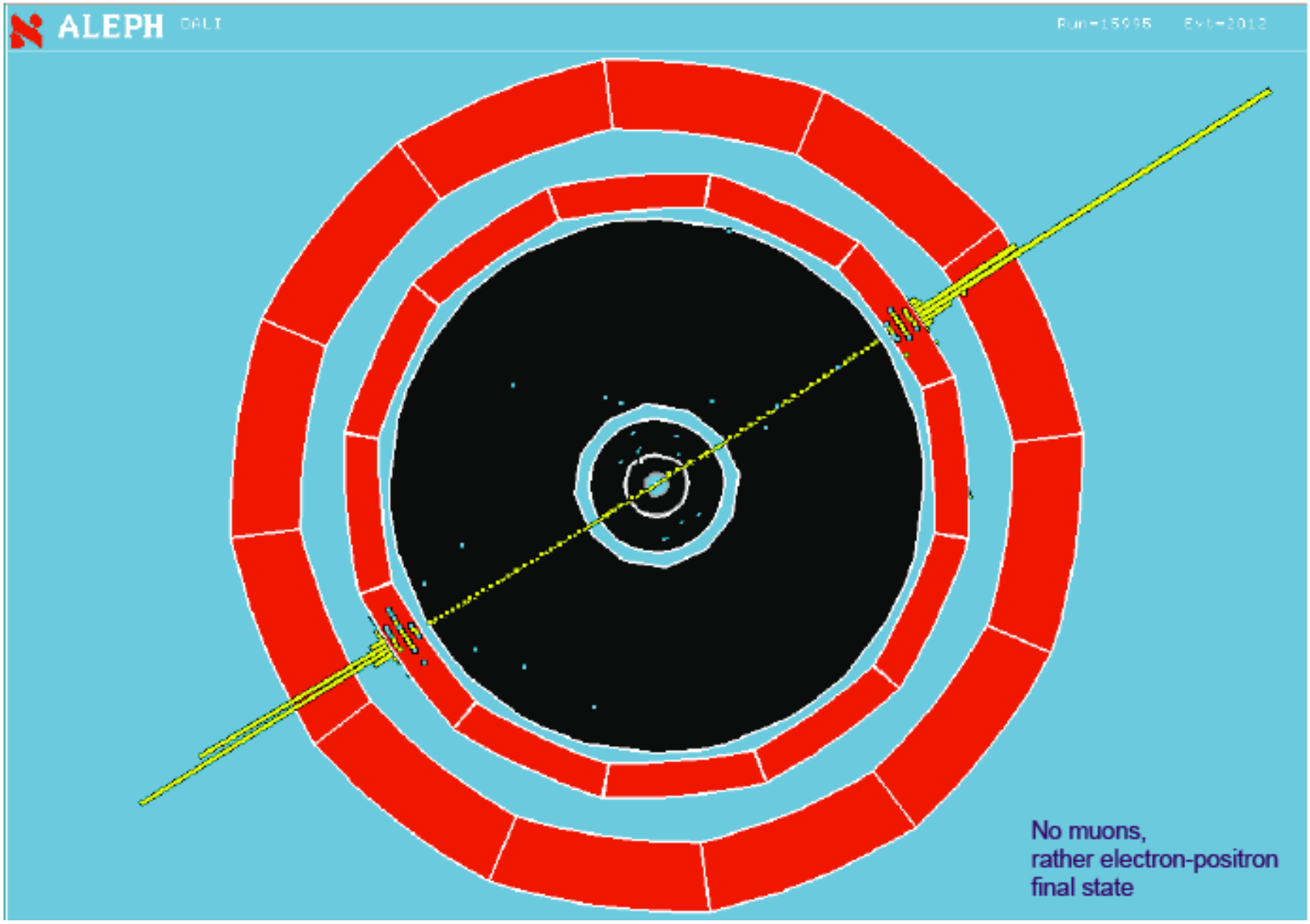
A simple „counting experiment”

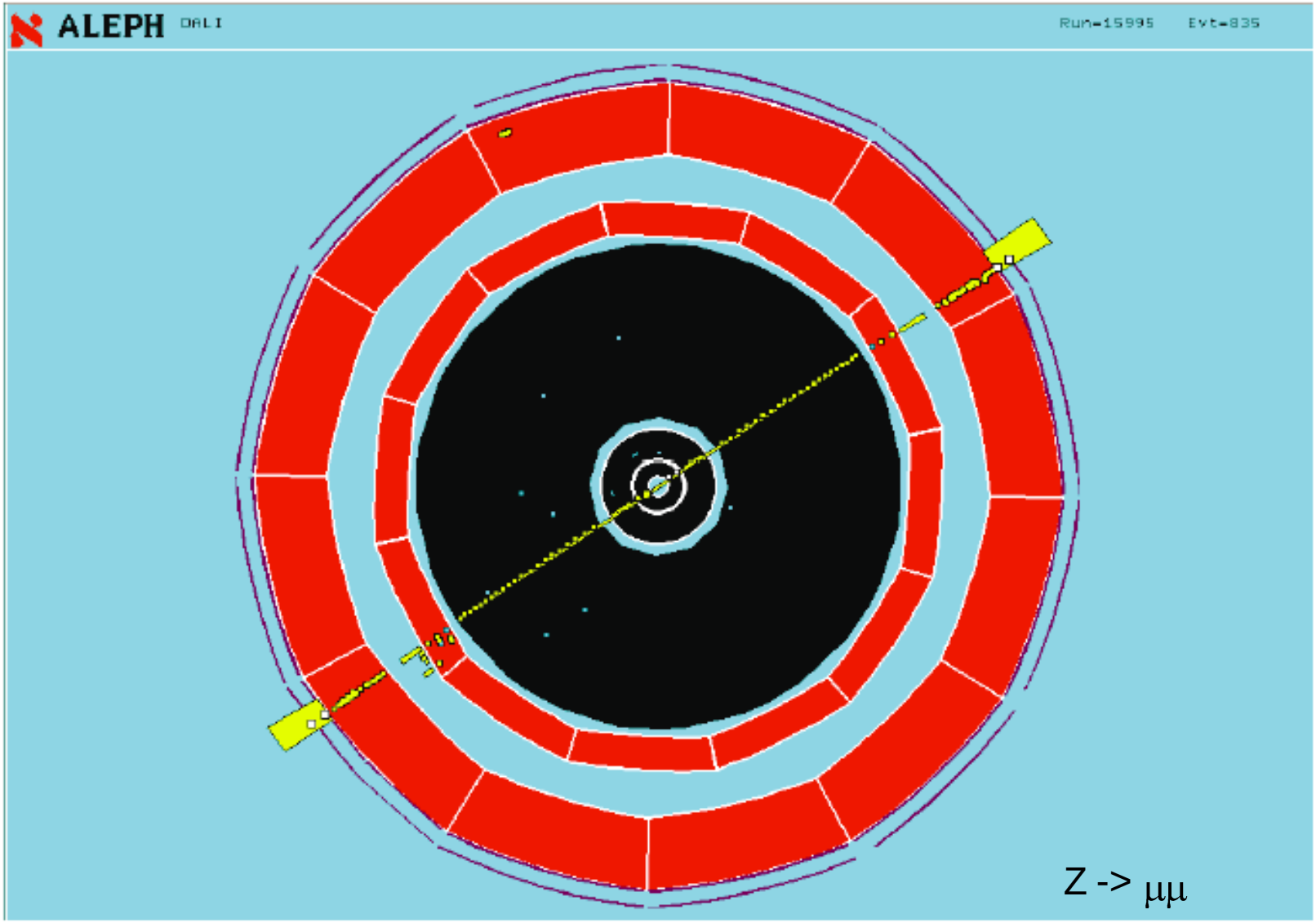
$$R = \frac{\#\text{ of Hadron final states}}{\#\text{ of Muon final states}}$$

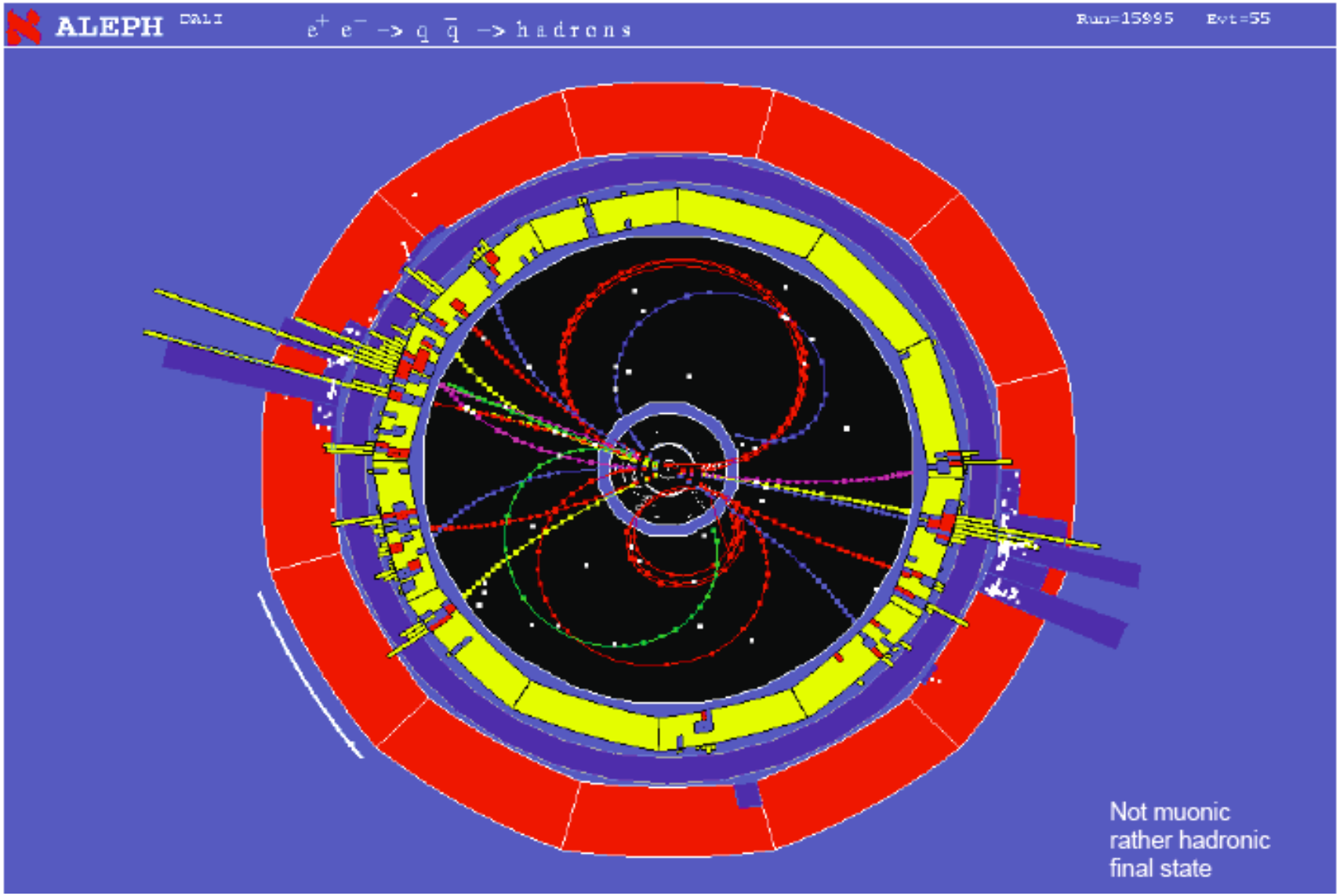





ALEPH
DALI
Run=15768
Evt=5906
Not muonic
rather hadronic
final state


ALEPH
DALI
Run=15995 Evt=2012
No muons,
rather electron-positron
final state


ALEPH
DALI
Run=15995
Evt=835
Z -> μμ


ALEPH DALI
e+ e− -> q q̄ -> hadrons
Run=15995 Evt=55
Not muonic
rather hadronic
final state

Uncertainties

- Just having a "counting result" is not all, there's lot more to do!
- Statistical error
  - We saw 2 muon events, could easily have been 1 or 3
  - Those fluctuations go like the square-root of the number of events

$$BR\left(Z^0 \to \mu^+\mu^-\right) = \frac{N_{\mu\mu}}{N_{total}} \pm \frac{\sqrt{N_{\mu\mu}}}{N_{total}}$$

  - To reduce this uncertainty, you need to record lots (millions) of events in the detector, and process them
- Systematic error
  - What if you only see 50% of the $\mu^+\mu^-$ events?

"efficiency"

$$N_{\mu\mu\,\text{seen}} = \varepsilon N_{\mu\mu}$$

    - because of event selection (cut), detector imperfections, poor understanding, etc.

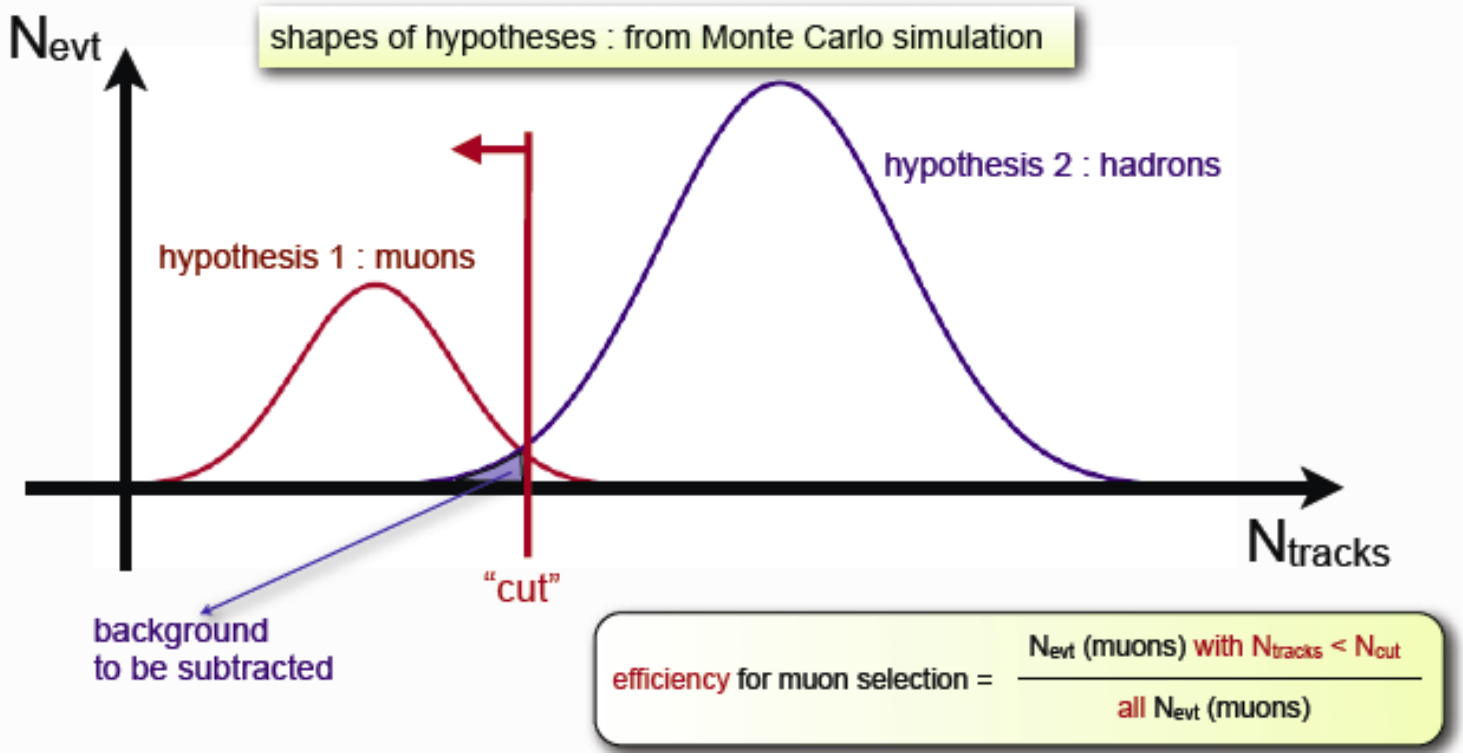
$$BR\left(Z^0 \to \mu^+\mu^-\right) = \frac{N_{\text{seen}}/\varepsilon}{N_{total}}$$

$$\varepsilon = 0.50 \pm 0.05$$

- from statistical error of detector simulation
- imperfect modeling of geometry in simulation
- model of muon interactions in simulation, etc

Event selection

- Event per event have to decide how to categorize it
  - eg. do we call it a muon event, or a hadronic event?
  - how do we estimate the efficiency?
  - Define an event selection, eg. "cut-based"
  - see statistics lectures, *hypothesis testing* etc...

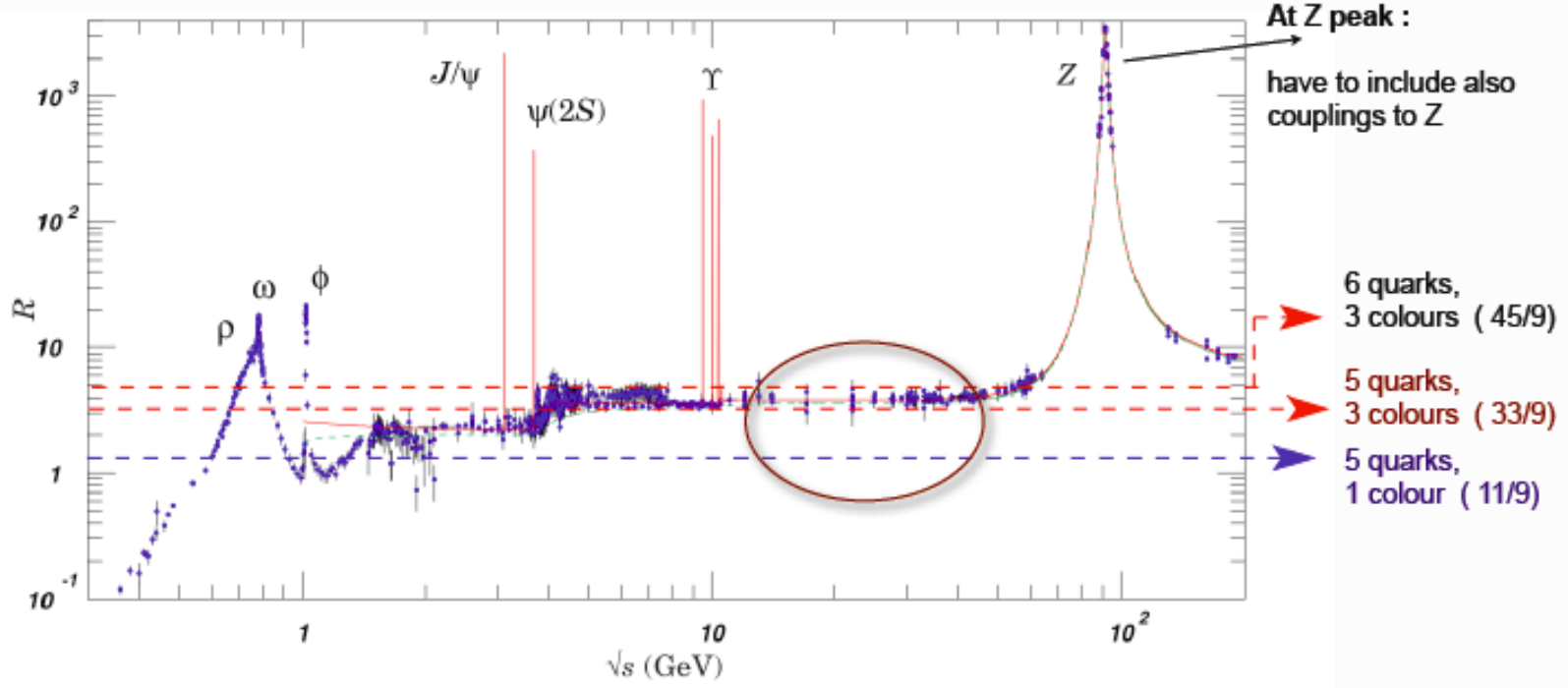


Result

For E_{CM} below the Z peak and above the Υ resonance we expect:

$$R = N_c \sum_f z_f^2 = N_c \cdot [(\tfrac{2}{3})^2 + (-\tfrac{1}{3})^2 + (-\tfrac{1}{3})^2 + (\tfrac{2}{3})^2 + (-\tfrac{1}{3})^2] = N_c \cdot \tfrac{11}{9}$$

u d s c b

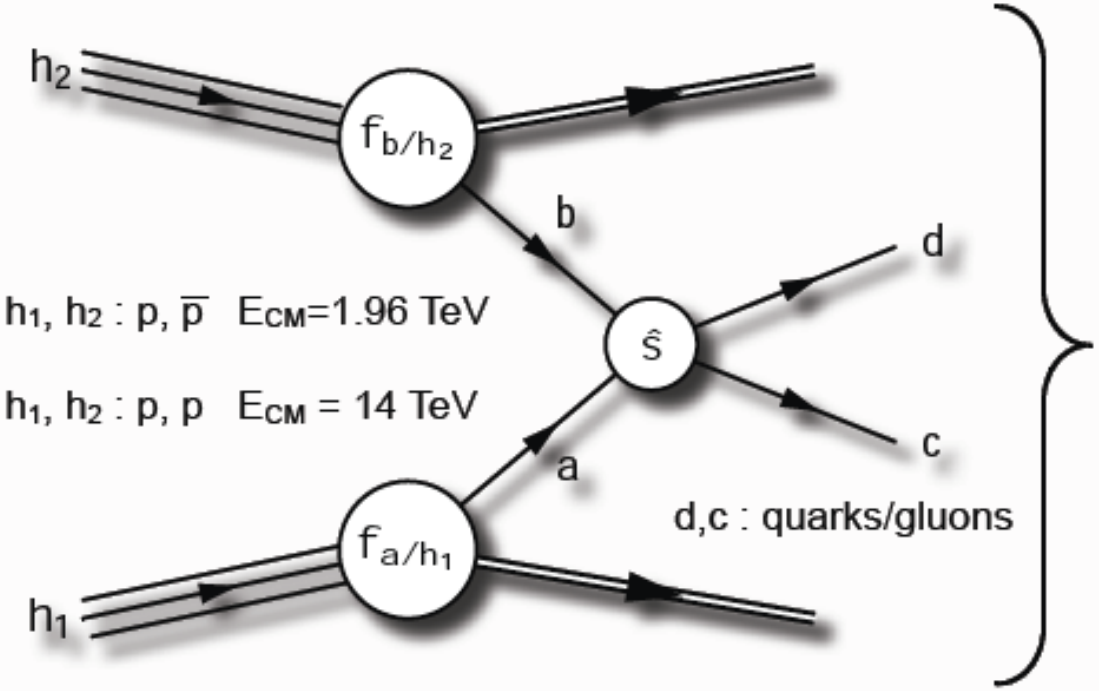


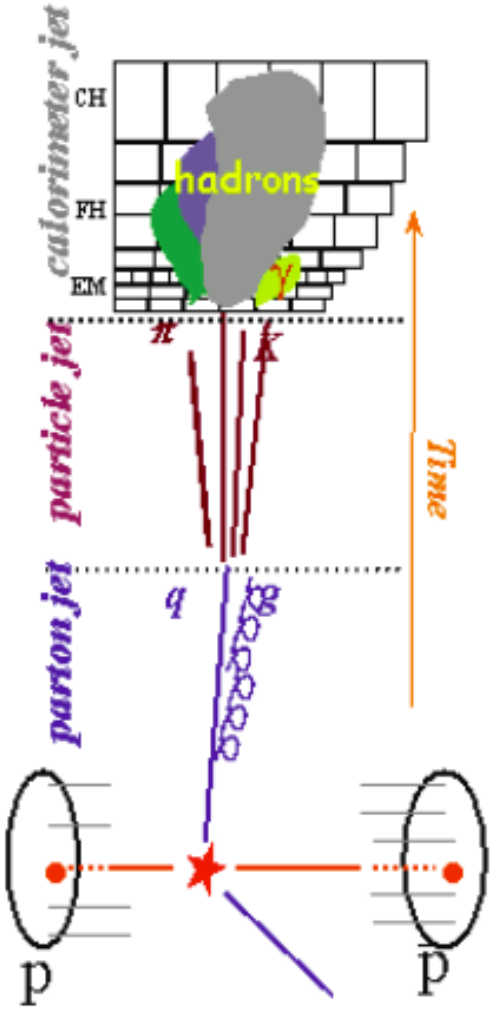
- Confirmation of : Number of colours = 3 !

Note : small remaining difference : because of QCD correction (gluon radiation) = 1 + α_s / π

A more complicated example

- at the Tevatron, or in the future at the LHC





Goal

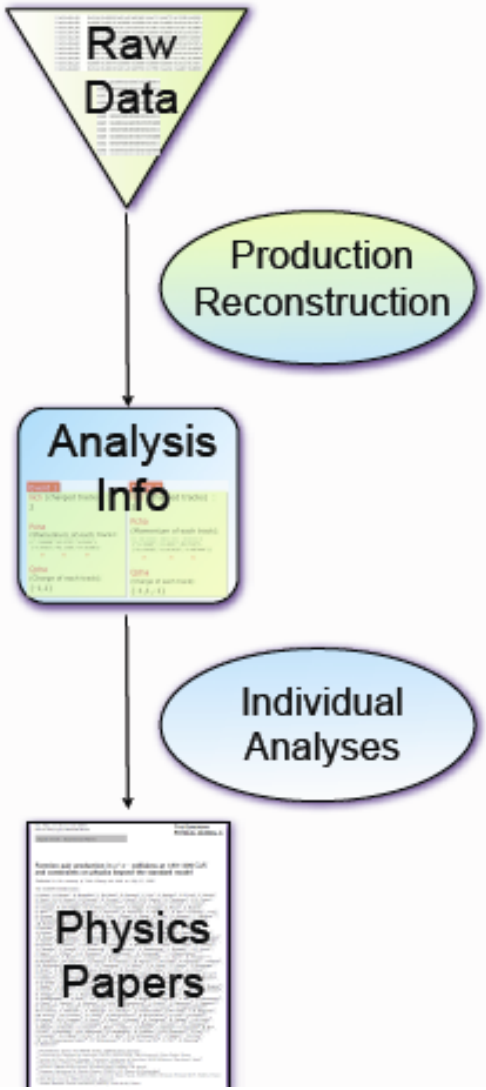
- measure probability that quarks/gluons are produced with a certain energy, at a certain angle
- Problem : do not observe quarks and gluons directly, only hadrons, which appear collimated into **jets**
- Reconstruct tracks and/or energy clusters in the calorimeter

Where do we stand now?

- After data flow from DAQ: data reduction and abstraction
  - reconstruct tracks, energy deposits (clusters) in calorimeters
  - calculate "high-level" physics quantities
    - eg. momentum of charged particles, energy of neutral particles
  - apply even higher-level algorithms, eg. jet finding
  - store all these quantities/objects event per event
- The data analysis
  - define the theoretically computed observable(s) to be measured
  - apply event selection (cuts)
  - estimate efficiencies and backgrounds, eg. from MC simulation
  - if distributions are measured : take care of absolute calibrations and effects because of detector resolution/smearing
    - correct for these effects
  - determine statistical and systematic uncertainties
  - compare with theory, found a deviation, something new?
    - if yes, book the ticket to Stockholm
  - determine parameters, eg. by fitting the prediction to the data

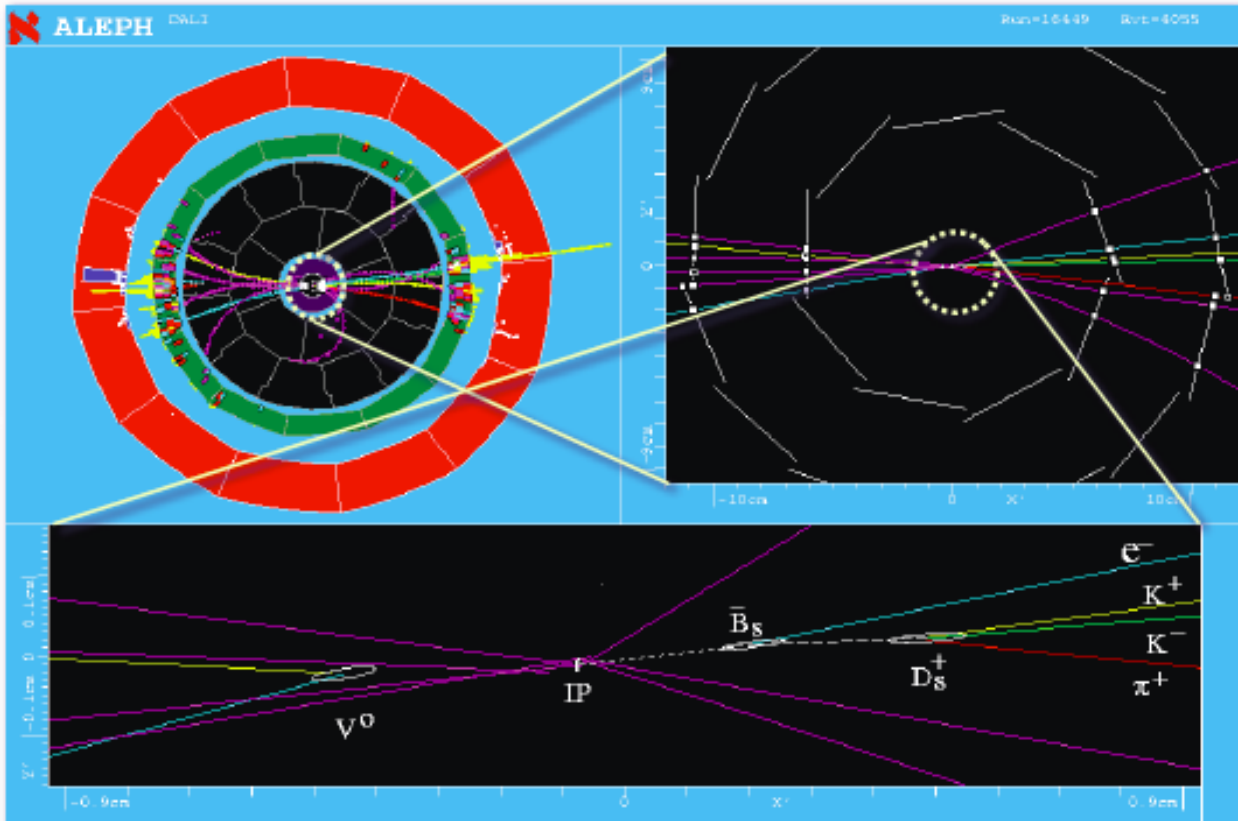
The process in practice

- The reconstruction step is usually done in common
  - "Tracks", "particle ID", "calorimeter towers" etc are general concepts, not analysis-specific. Common algorithms make it easier to understand how well they work
  - "very coordinated" data access
- Analysis is a very individual thing
  - Many different measurements being done at once
  - Small groups working on topics they are interested in
  - Many different time scales for these efforts
  - "chaotic" data access
- Collaborations build offline computing systems to handle all this



Why tracking needs to be done well

- Determine how many charged particles were created in an event
- Measure their momentum
  - direction, magnitude
  - combine these to look for decays of particles with known masses
  - only final stable particles are visible
- Measure spatial trajectories
  - combine to look for separated vertices, indicating particles with long lifetimes



Tracks fitting

- 1D straight line fit as simple case
- Two perfect measurements
  - away from interaction point
  - no measurement uncertainty
  - just draw a straight line through them and extrapolate

- Imperfect measurements give less precise results
  - the farther you extrapolate, the less you know

- Smaller errors and more points help to constrain the possibilities.
  But how to find the best point from a large set of points?

- Quantitatively
  - parameterize a track:
  - In case of straight line $y(x) = \theta\, x + d$ or, eg., helix in case of magnetic field present

predicted track position at i^{th} hit

position of i^{th} hit

$$\chi^2 = \sum_{i=1}^{n_{\text{hits}}} \frac{\left(y_i - y(x_i)\right)^2}{\sigma_i^2}$$

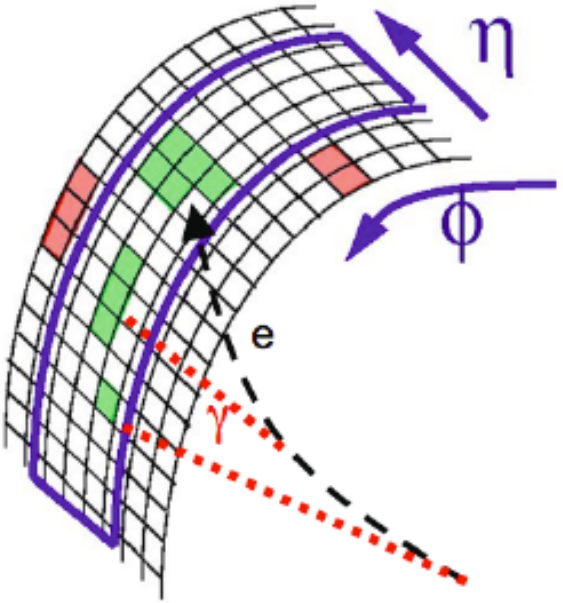
uncertainty of i^{th} measurement

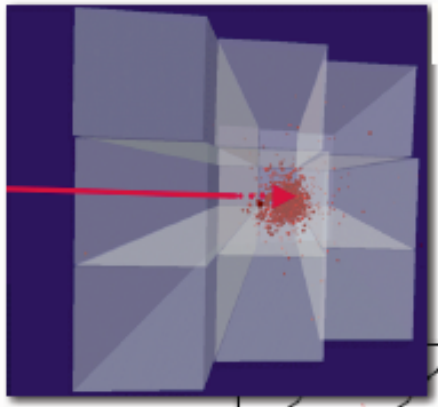
  - Find track parameters by Least-Squares-Minimization
  - Obtain also uncertainties on track parameters

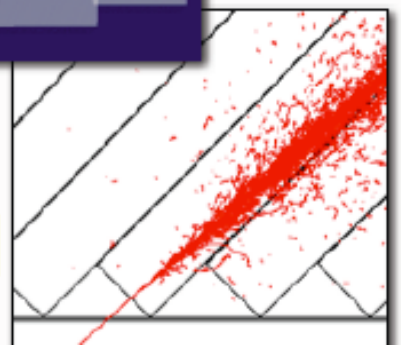
$\delta\theta$ δd

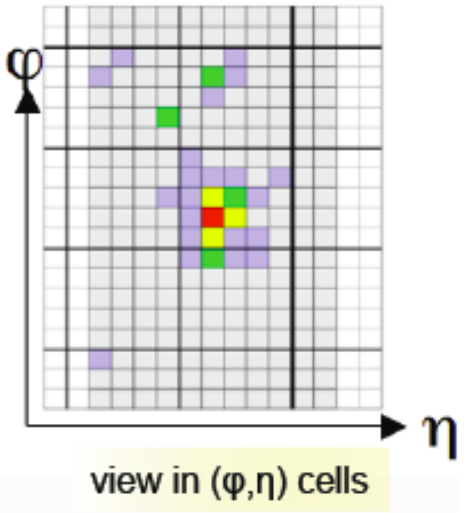
Cluster of energy

- Calorimeters are segmented in cells
- Typically a shower extends over several cells
  - Useful to reconstruct precisely the impact point from the "center-of-gravity" of the deposits in the various cells
- Example CMS Crystal Calorimeter:
  - electron energy in central crystal ~ 80 %, in 5x5 matrix around it ~ 96 %
- So task is : identify these clusters and reconstruct the energy they contain



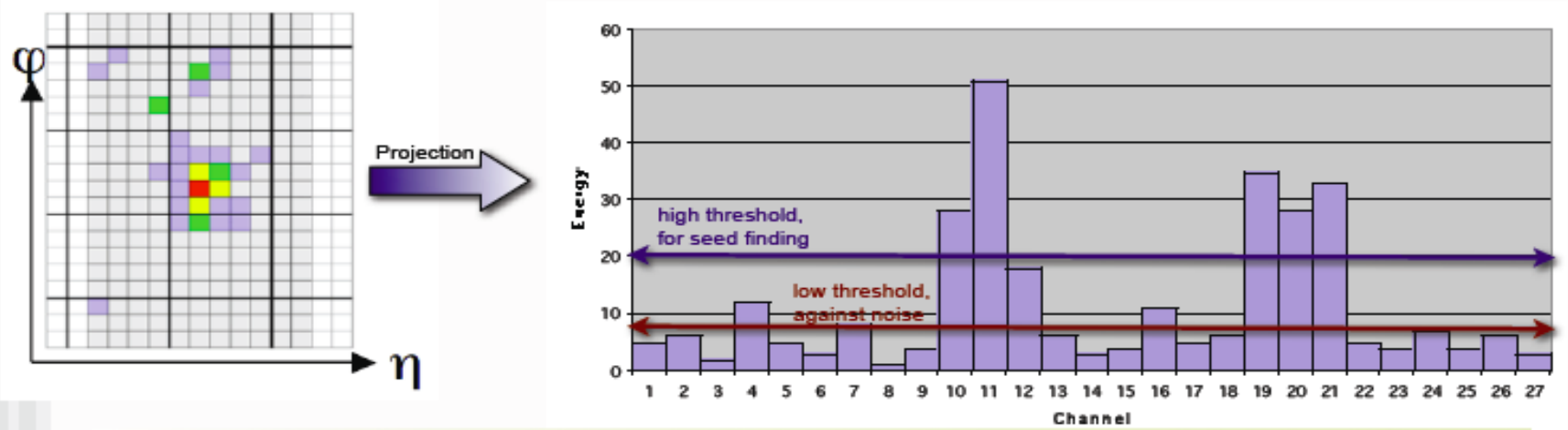






Cluster finding

- Clusters of energy in a calorimeter are due to the original particles
  - Clustering algorithm groups individual channel energies
  - Don't want to miss any; don't want to pick up fakes

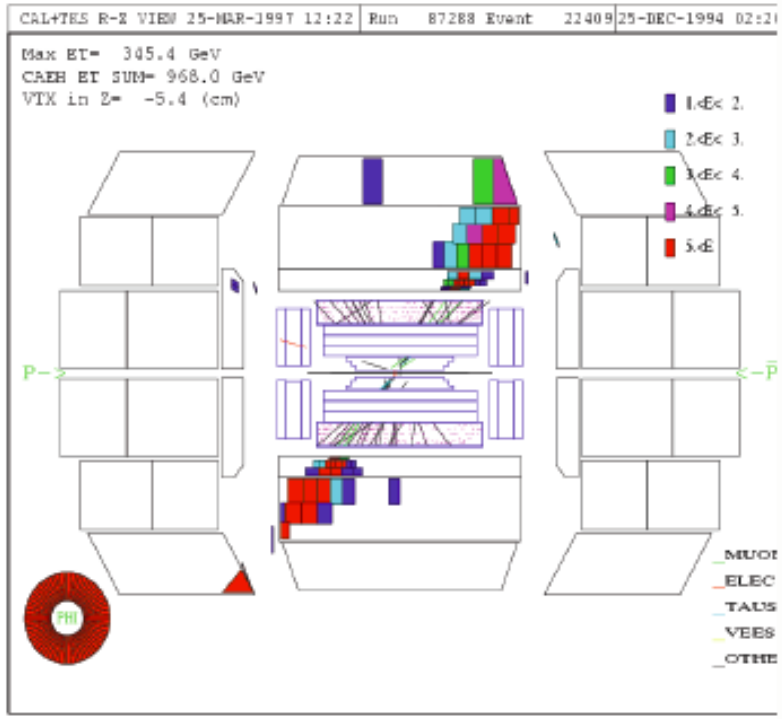


Simple example of an algorithm

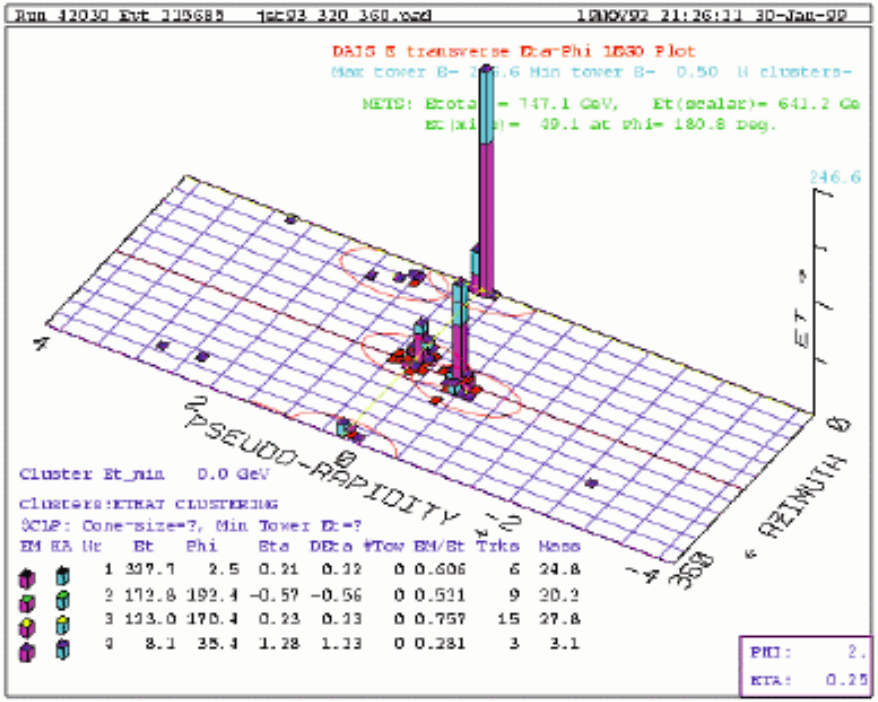
- Scan for seed crystals = local energy maximum above a defined seed threshold
- Starting from the seed position, adjacent crystals are examined, scanning first in φ and then in η
- Along each scan line, crystals are **added to the cluster if**
  1. The crystal's energy is above the noise level (lower threshold)
  2. The crystal has not been assigned to another cluster already
  3. The previous crystal added (in the same direction) has higher energy

Jets in hadron colliders

Jets in DØ



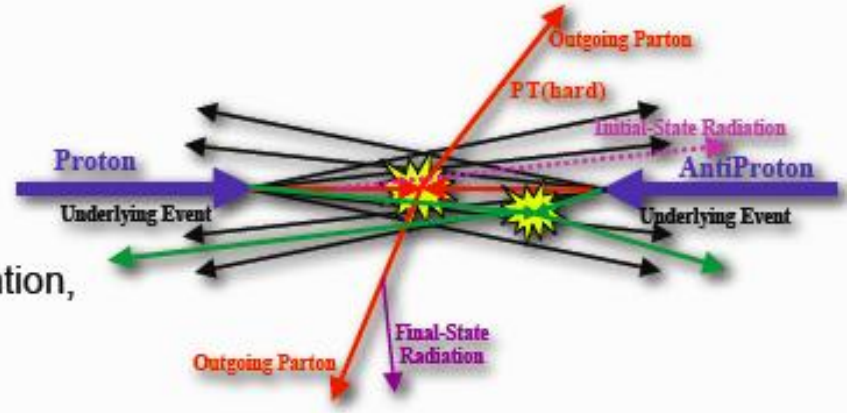
CDF



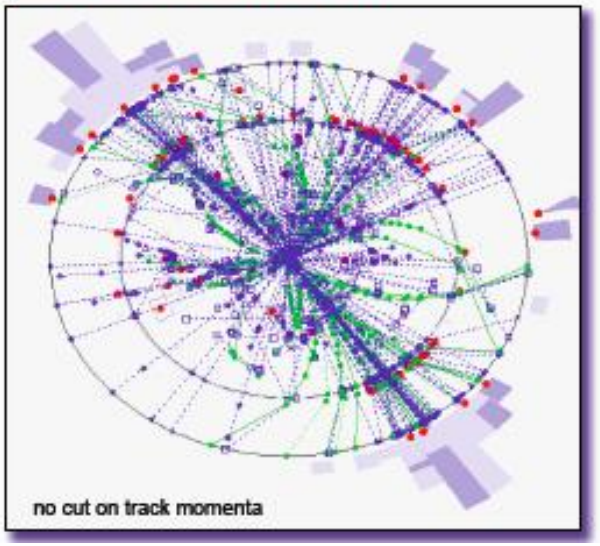
- Introducing a cone prescription seems “natural”...
- But how to make it more quantitative?
  - don’t want people “guessing” at whether there are 2,3, ... jets

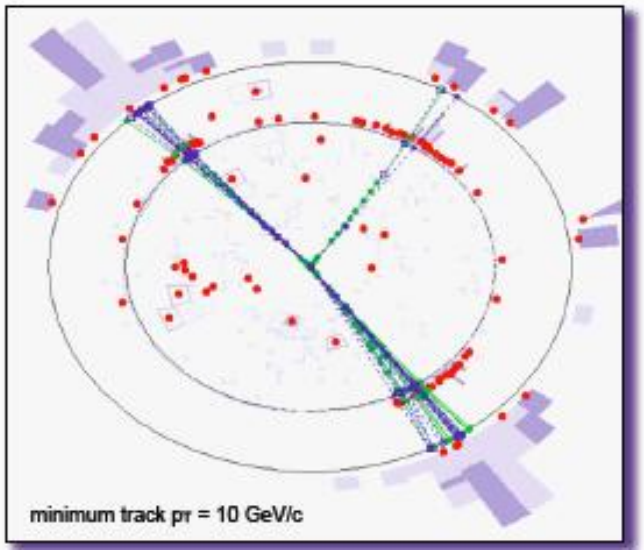
Further difficulties

- Pile Up : many additional soft proton-proton interactions
  - up to 20 at highest LHC luminosity
- Underlying event
  - beam-beam remnants, initial state radiation, multiple parton interactions
  - gives additional energy in the event



- All this additional energy has nothing to do with jet energies
  - **have to subtract it**





Some numbers

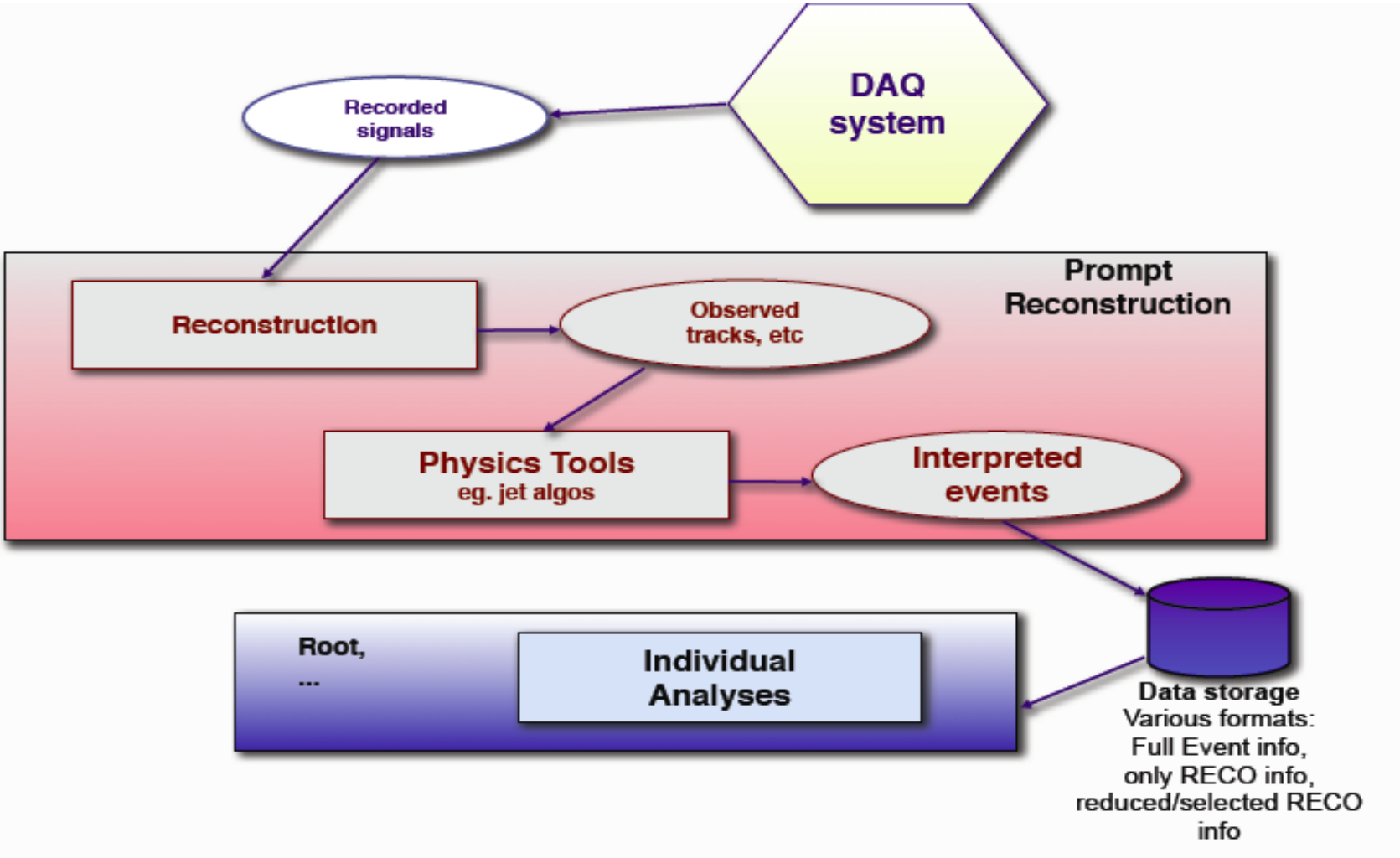
- Examples from CMS, estimates
  - Rate of events streaming out from High-Level Trigger farm ~150 Hz
  - each event has a size of the order of 1 MByte

- CMS will record ~100k top-quark events per day
  - among about 10^7 events in total per day
  - will have roughly 150 "physics" days per year
  - thus about 10^9 evts/year, a few Pbyte

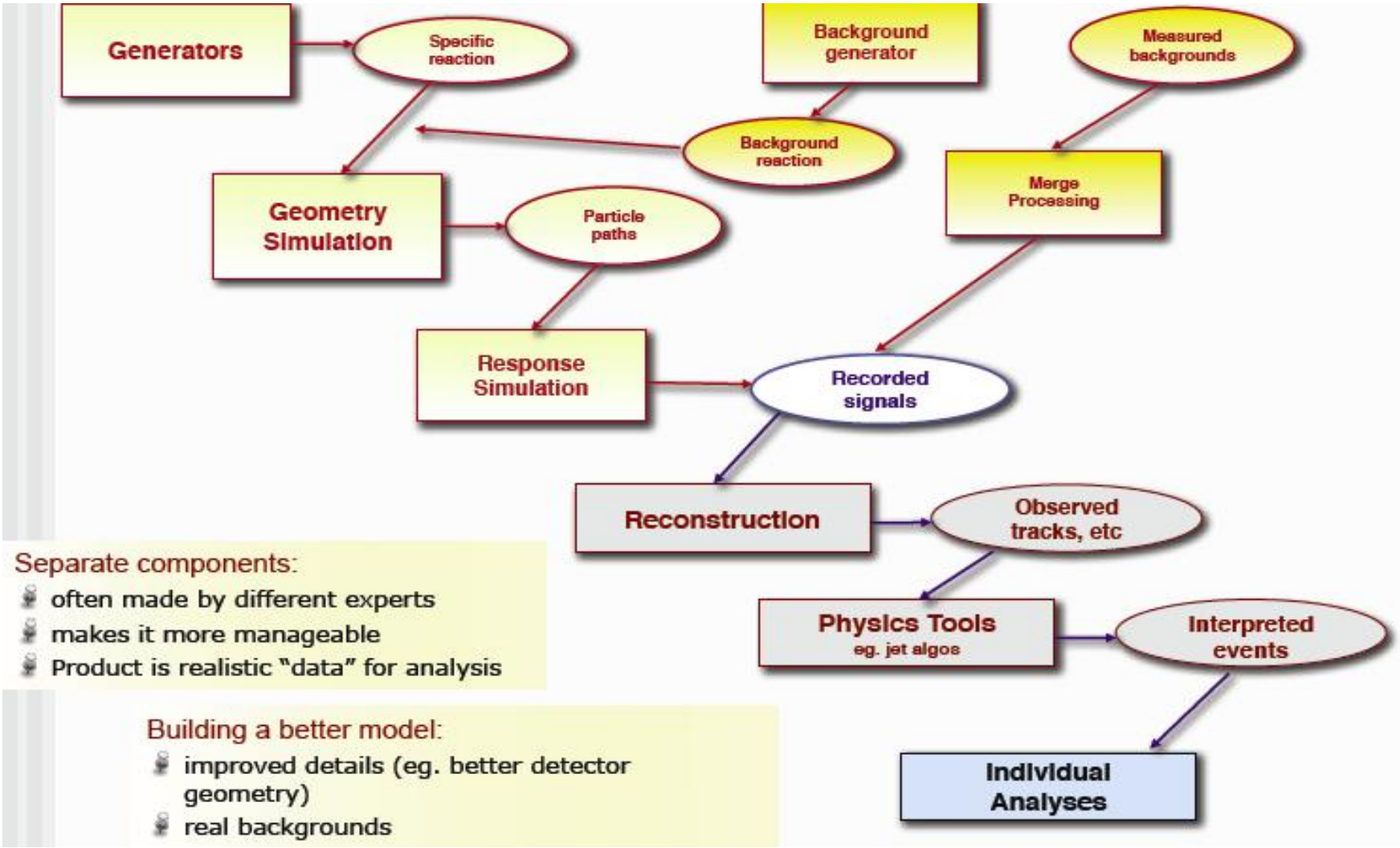
- "prompt" processing
  - Expect to do first reprocessing step within one day
  - Reco time per event on std. CPU: < 5 sec (on lxplus)
  - Note : will have to reprocess several times
    - new/better algorithms, updated calibrations, etc.
- Expect to simulate several 100s to 1000s of millions of events
  - will be mostly done at computing centres outside CERN
  - Simulation time per event now ~ 100 secs (eg. for QCD or top evts)

- Now : ~2 million lines of code (reconstruction and simulation)

Reconstruction flow



Flow of simulated data



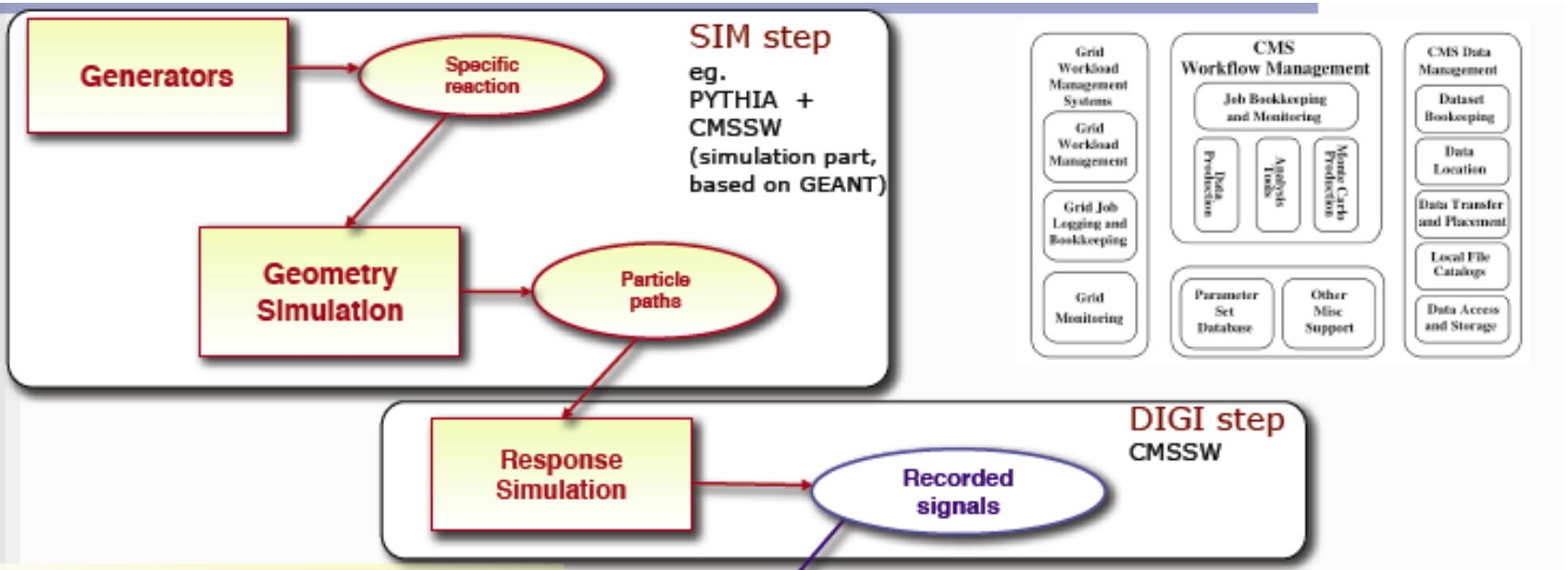
Separate components:

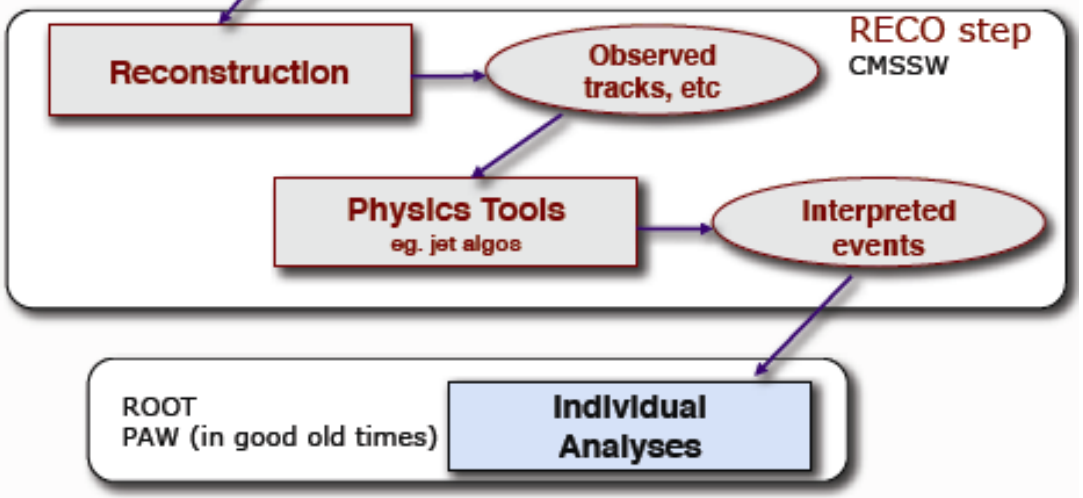
- often made by different experts
- makes it more manageable
- Product is realistic "data" for analysis

Building a better model:

- improved details (eg. better detector geometry)
- real backgrounds

Partitioning production system





Partitioning

- there can be event stores between individual components

Why this structure

- flexibility,
  have different versions of pieces
- efficient for repeated studies
  don't have to start all over again if some improvement in later stage
- Manageability
  large programs, hard to build, understand, debug, maintain, ...

- Reconstruction and Analysis
  is how we get from raw data to physics papers
- On your way
  - first you have too much information → reduce
  - sometimes too little information or little prior knowledge
    - make hypotheses
- What makes it hard, but also exciting
  - many many cross checks
  - more cross checks
  - sometimes some “art” involved
  - tuning, evolutionary improvement

- Even to me it is often a miracle that we can generate wonderful results from these complicated instruments!

Athena components (i)

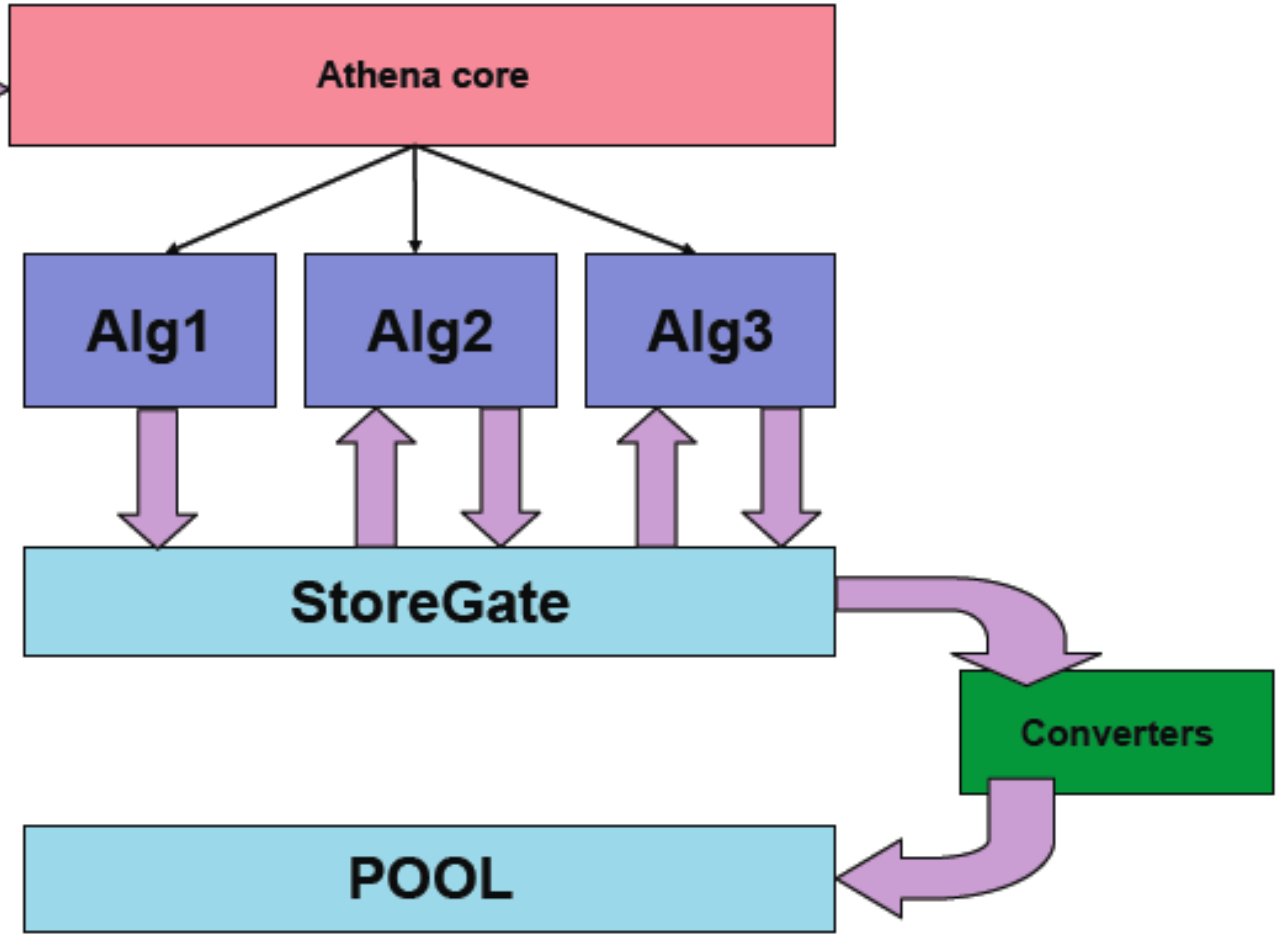
- **Algorithm**: an application - a piece of code that "does something"
  - All algorithms inherit from the Algorithm class, which contains three methods:
    - *Initialize*() - run once at the start
    - *Execute*() - run n times
    - *Finalize*() - run once at the end
  - Algorithms are invoked centrally by the framework
  - Many algorithms can be run in a single job - one after the other
- **Data object**: result of an algorithm, or the input to it
  - E.g. Track, Cluster, Muon, Electron, McEvent
- **Service**: globally available software entity which performs some common task
  - Message printing
  - Histogram drawing
- **Event**: a single pass of the execute() method, roughly corresponding to a physics event
- **JobOptions**: Python script which passes user instructions to Athena
  - Which algorithms to run, what order, configuration
  - Control of number of cycles, input/output files, runtime variables etc

Athena components (ii)

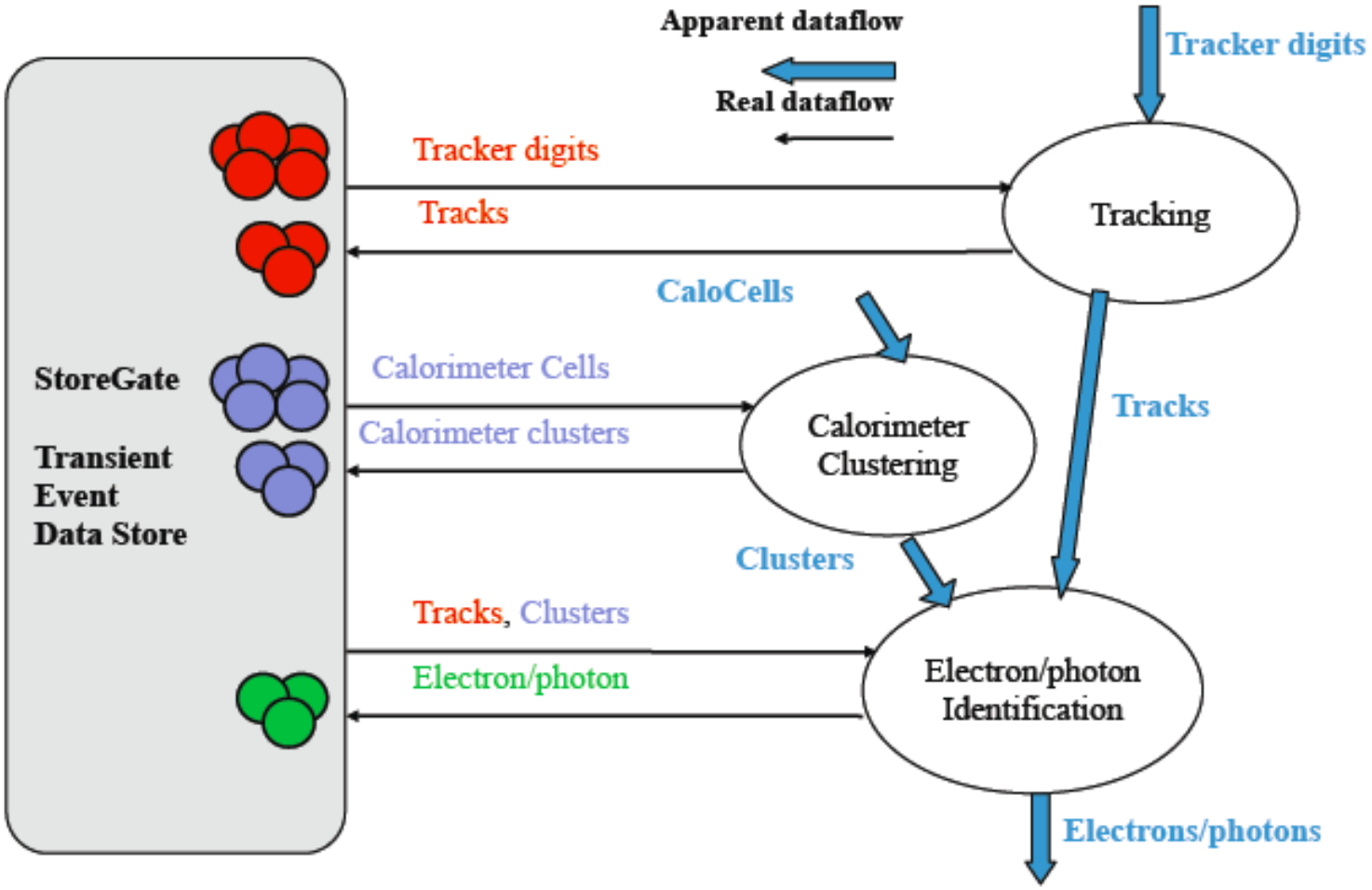
- **Tool**: piece of code that is shared between algorithms - it can be executed as many times as you need in the execute() method of your algorithms
- **Auditors**: software which monitors the other components of the framework
- **Sequence**: execution order of the algorithms
- **Filters**: software which allows or forbids an event from passing to the next algorithm in the sequence or being written to disk
- **Transient Store (StoreGate)**: service which stores results of algorithms (data objects) and passes them to the next algorithm.
  - The data is held in the computer memory
- **Persistent Store (POOL)**: format in which the data objects are written to disk
- **Converter**: software which enables the data objects used in the code to be written to and read from POOL without the details of the persistency being included in the objects themselves

Athena scheme (simplified)

JobOptions

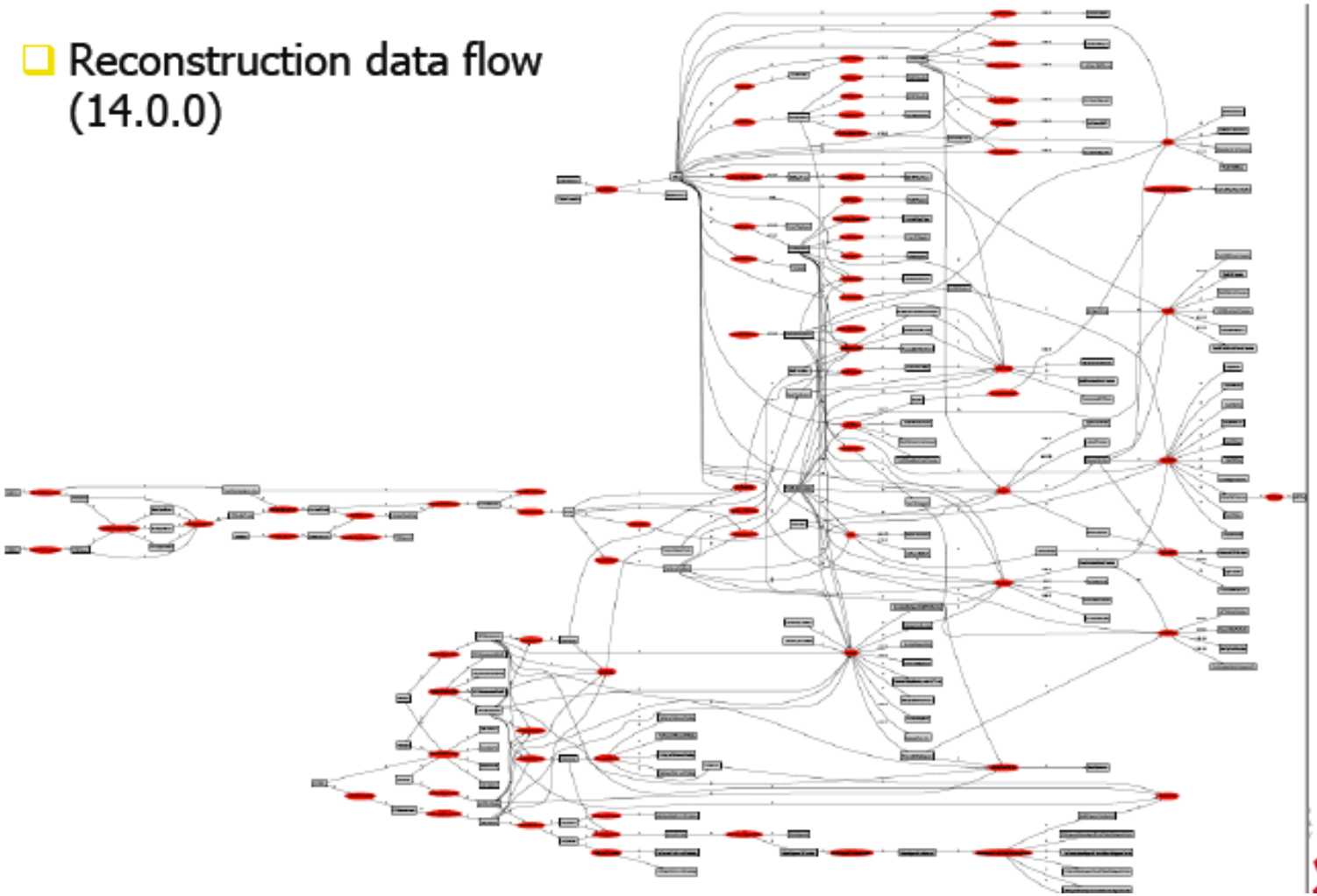


Athena scheme (a bit less simplified)



Athena scheme (even less simplified)

- Reconstruction data flow (14.0.0)



Don't panic!

- You don't have to worry about most of the complications
- Physics analysis is the simplest part of the framework
- If you're going to be working on a certain area of the software you'll just concentrate on a few pieces of code, not the whole framework!
- It is still useful to hear about the full picture, so you have some idea of what the software is doing "under the bonnet"