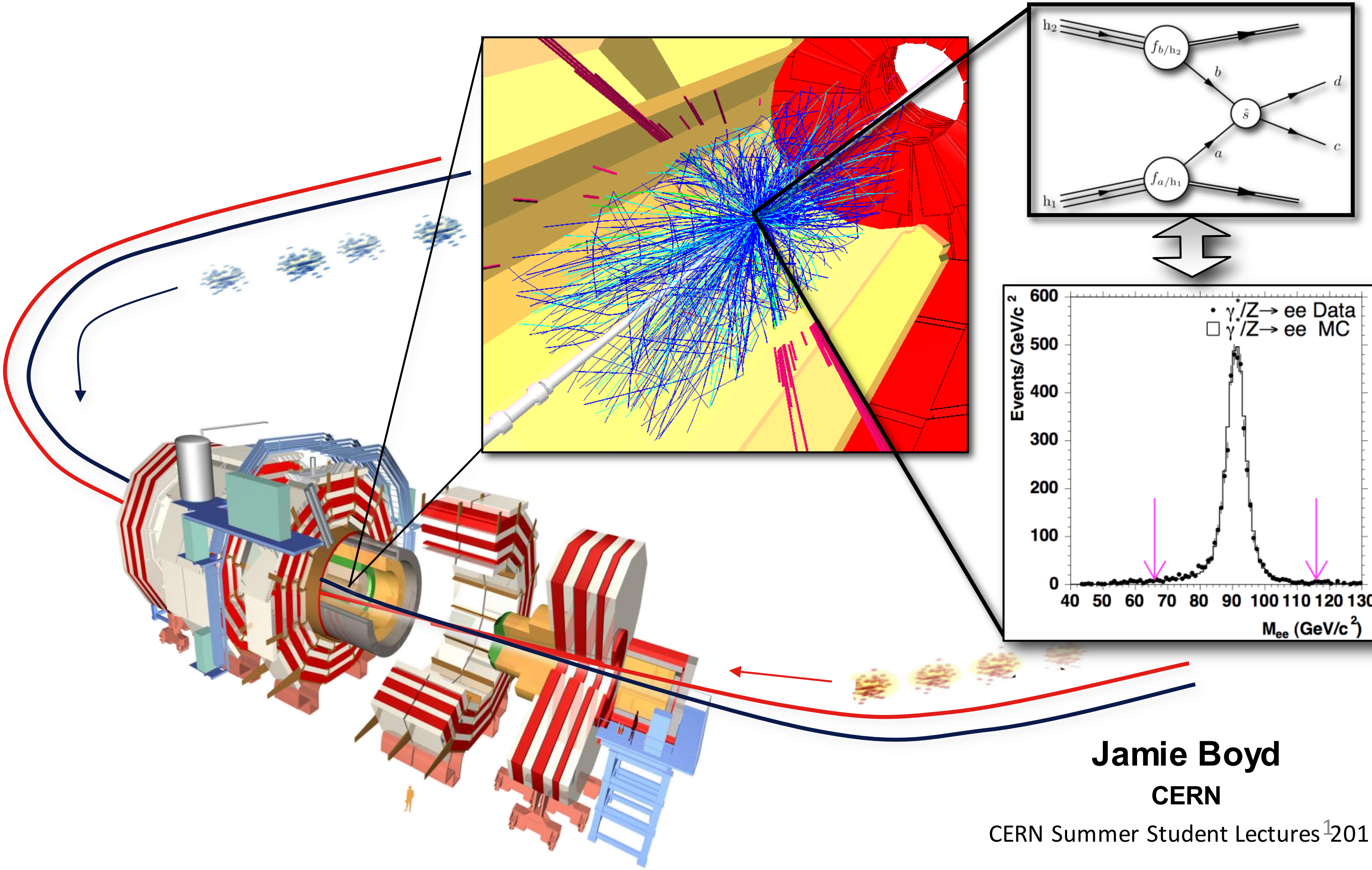
From Raw Data to Physics Results



Jamie Boyd

CERN

CERN Summer Student Lectures 2011

Outline

- Summary
    - Brief overview of the full lecture course
- A simple example
    - Measuring the Z^0 cross-section
- Reconstruction & Simulation
    - Track reconstruction
    - Calorimeter recostruction
    - Physics object reconstruction
    - Simulation
- Physics Analysis
    - Data Quality
    - Jet cross-section
    - Z'->ll
    - H->γγ
    - H->ZZ->4l
- Computing infrastructure

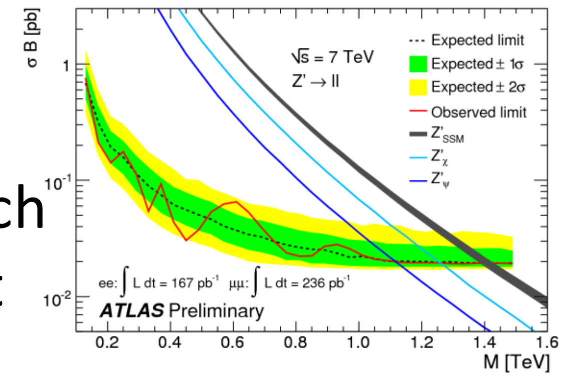
Todays Lecture

Disclaimer : Much of the content based on previous years lectures Thanks to G. Dissertori

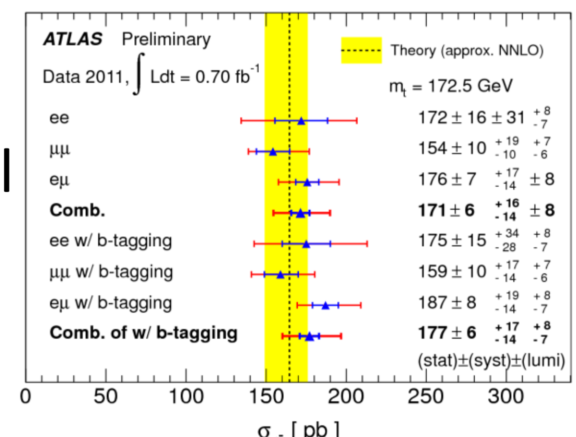
Section 4: Physics Analysis

Physics Analysis

- Two main types of physics analysis at LHC
  - Searching for new particles
  - Making precision measurements
- Searches statistically limited
  - More data is the way of improving the search
  - If don't see anything new set limits on what you have excluded

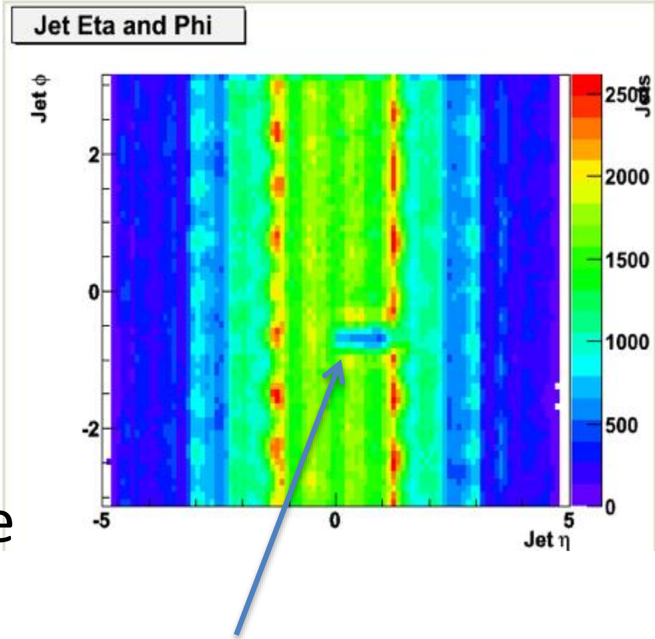


- Precision measurements
  - Precision often limited by the systematic uncertainties
  - Precision measurements of Standard Model parameters allows important tests of the consistency of the theory



Data Quality

- Particle physics detectors are very complex instruments – often there can be small problems
  - Noise in the detectors
  - Regions of the detector that are not working
- These effect the physics object reconstruction and the physics analysis
- Can exclude events with problems from the analysis
- Or include them, but take into account the problem
- Need a thorough system to check that the data is of good quality for physics analysis
  - Lots of people, checking lots of histograms…
- Data Quality very important to get correct physics results



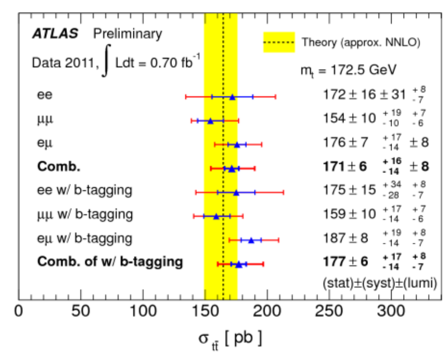
A dead region in the calorimeter in ATLAS

Physics Analysis Steps

~TB

- Start with the output of reconstruction
- Apply an event selection based on the reconstructed object quantities
  - Often calculate new information e.g masses of combinations of particles
  - Event selection designed to improve the 'signal' to 'background' in your event sample
- Estimate
  - Efficiency of selection
  - Background after selection
  - Can use simulation for these – but have to use data-driven techniques to understand the uncertainties
- Make final plot
  - Comparing data to theory
  - Correcting for efficiency and background in data
  - Include the statistical and systematic uncertainties

~kB



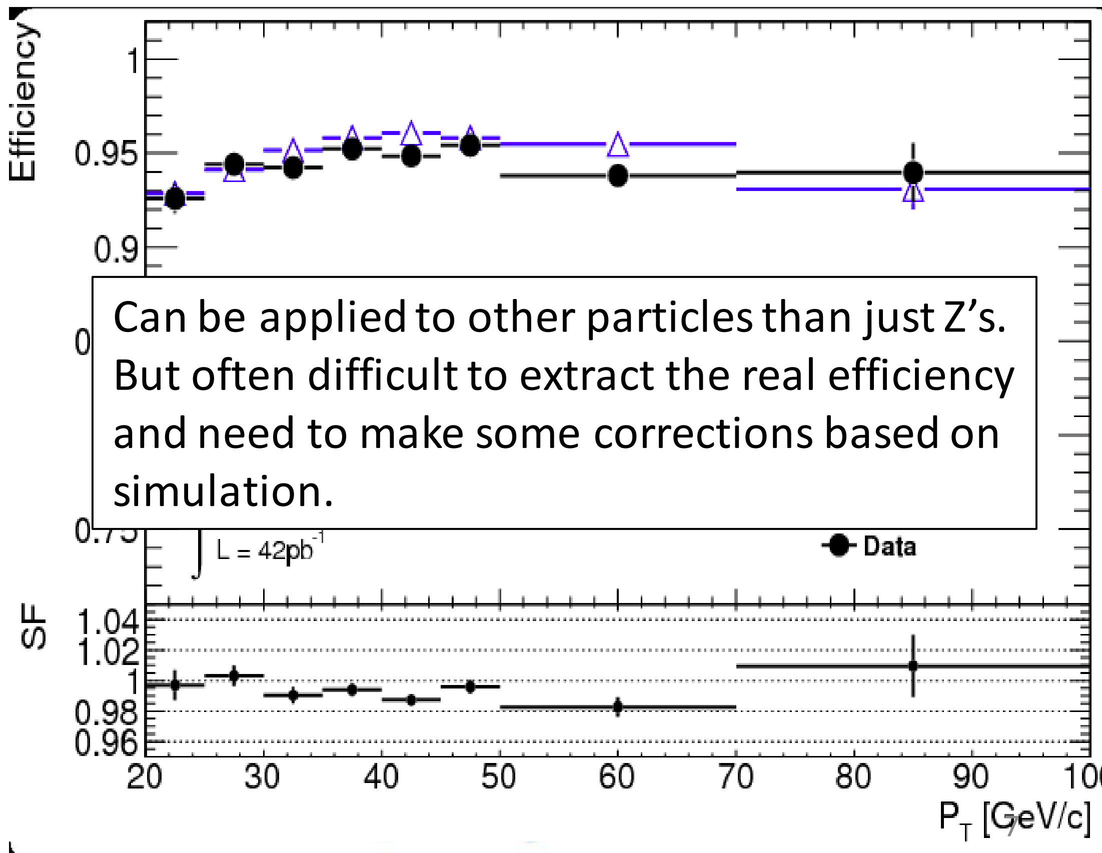
Measuring efficiencies from the data

- Can use simulation to get reconstruction efficiency
- But difficult to know how well the simulation describes the detector
- Need to try to use real data to estimate the efficiency too
- Example: Tag & Probe

idea: use Z->μμ decays to give us a pure sample of muon to use to measure the efficiency.

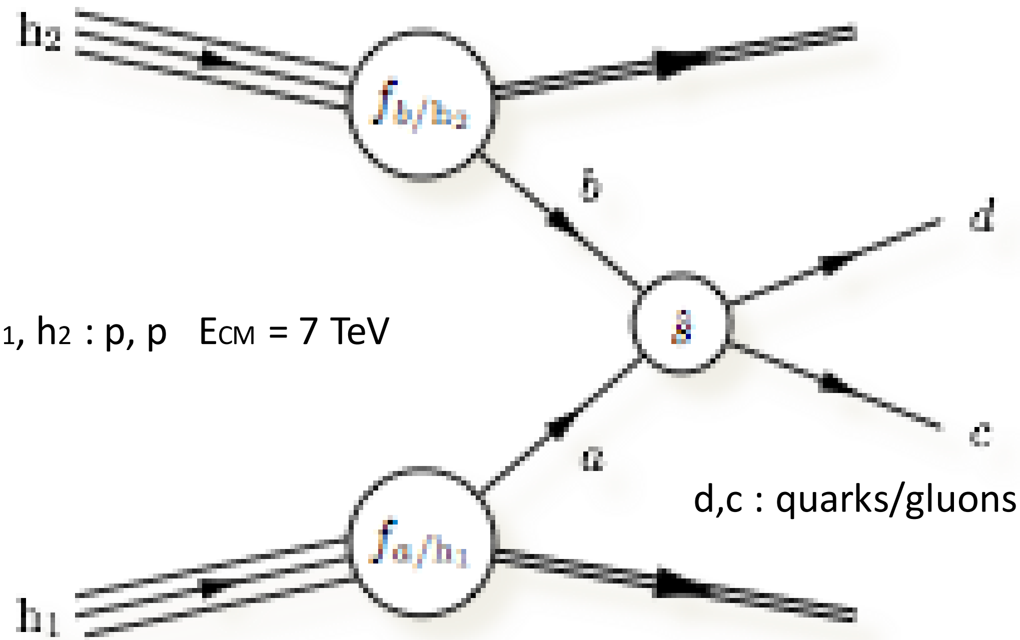
recipe:

1. select events with 1 reconstructed muon and 1 high momentum track
2. require that the mass of the track and the muon is consistent with Z mass
3. test if track is also reconstructed as a muon



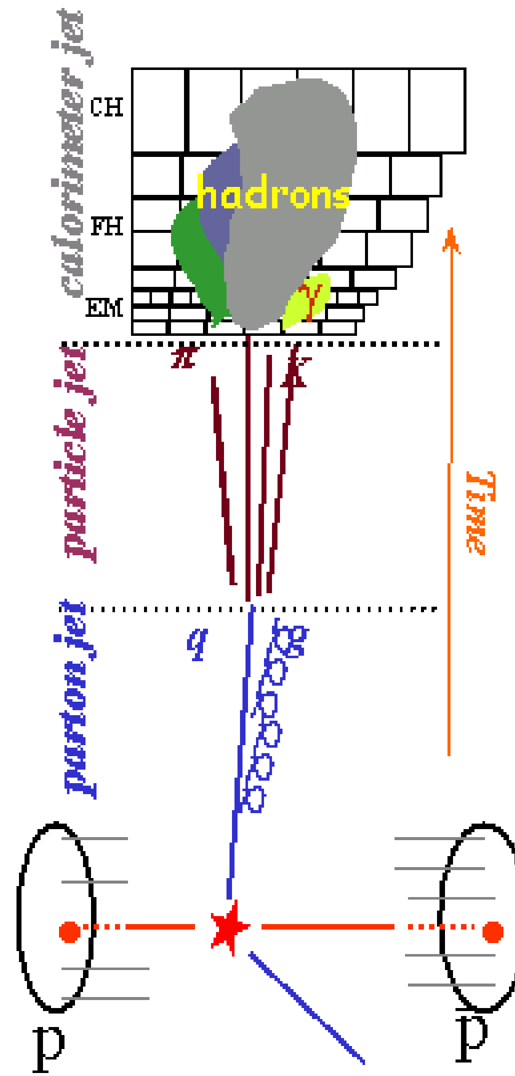
Can be applied to other particles than just Z's. But often difficult to extract the real efficiency and need to make some corrections based on simulation.

JET production at hadron colliders



h_1, h_2 : p, p E_{CM} = 7 TeV

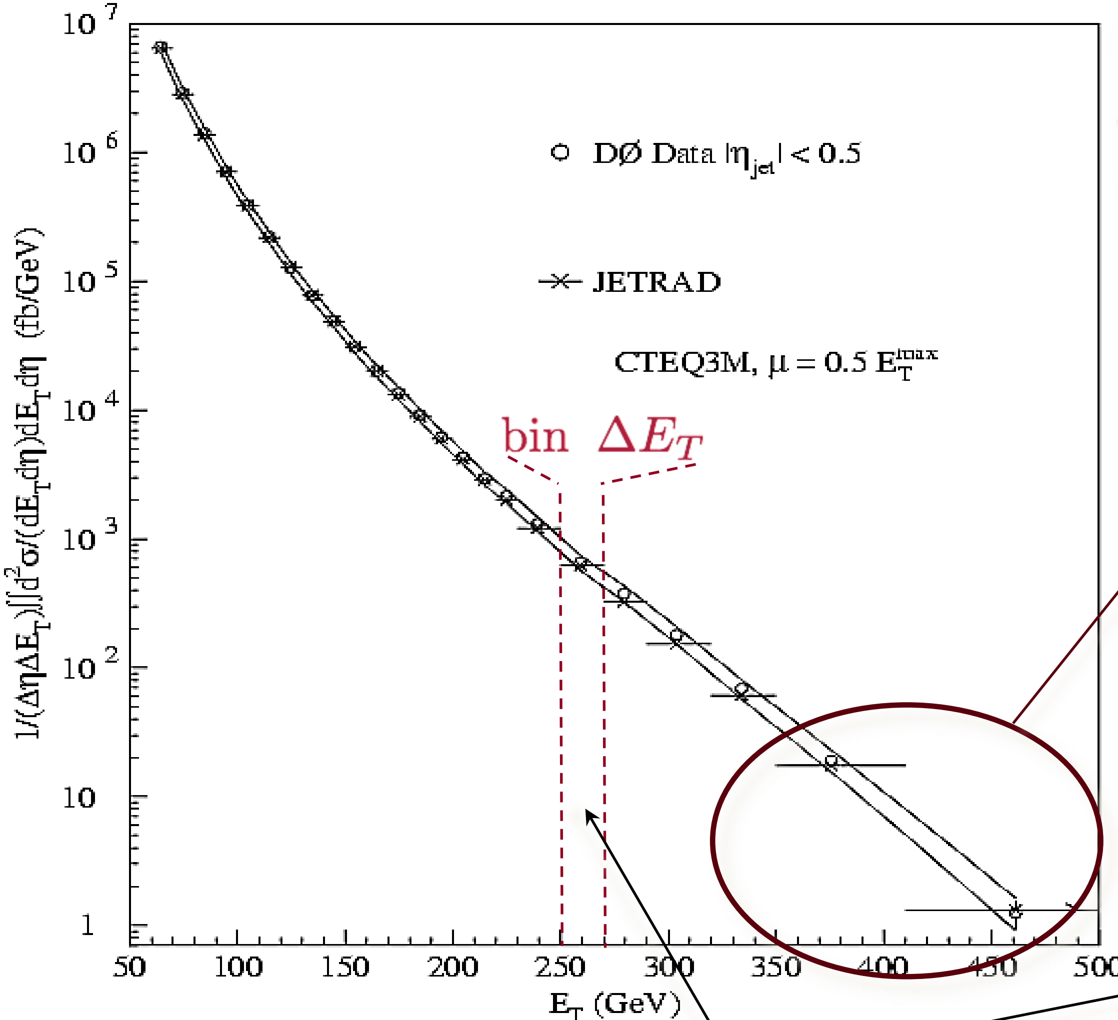
d,c : quarks/gluons



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 from energy clusters in the calorimeter
  - Unfortunately don't have time to go into details of jet reconstruction

What do we have to measure?



Goal

- measure cross section (probability) that jets are produced with a certain transverse energy E_T, within a certain rapidity range
- Test of perturbative QCD, over many orders of magnitude!
- Look at very high energy tail, new physics could show up there in form of excess (eg. sub-structure of quarks?)

can be calculated in pert. QCD

$$\left\langle \frac{d^2\sigma}{dE_T d\eta} \right\rangle = \frac{N}{\Delta E_T \Delta\eta\, \epsilon\, \mathcal{L}}$$

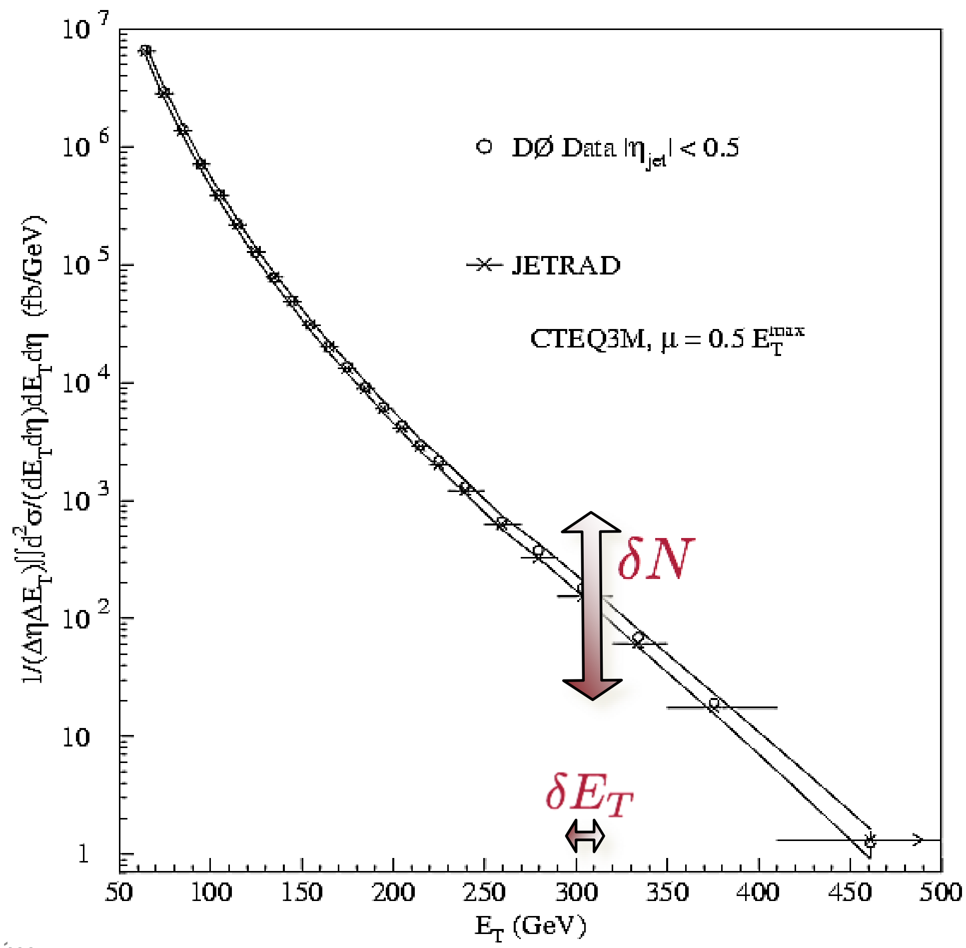
efficiency to reconstruct jets

integrated accelerator luminosity

- count number of events, N, in this bin
- for a certain range in rapidity (angle) $\Delta\eta$

Problem : Energy scale

- Question : how well do we know the energy calibration?
- Critical because of very steeply falling spectrum!



$$\frac{d^2\sigma}{dE_T d\eta} \approx \text{const} \cdot E_T^{\,-6}$$

relative uncertainties

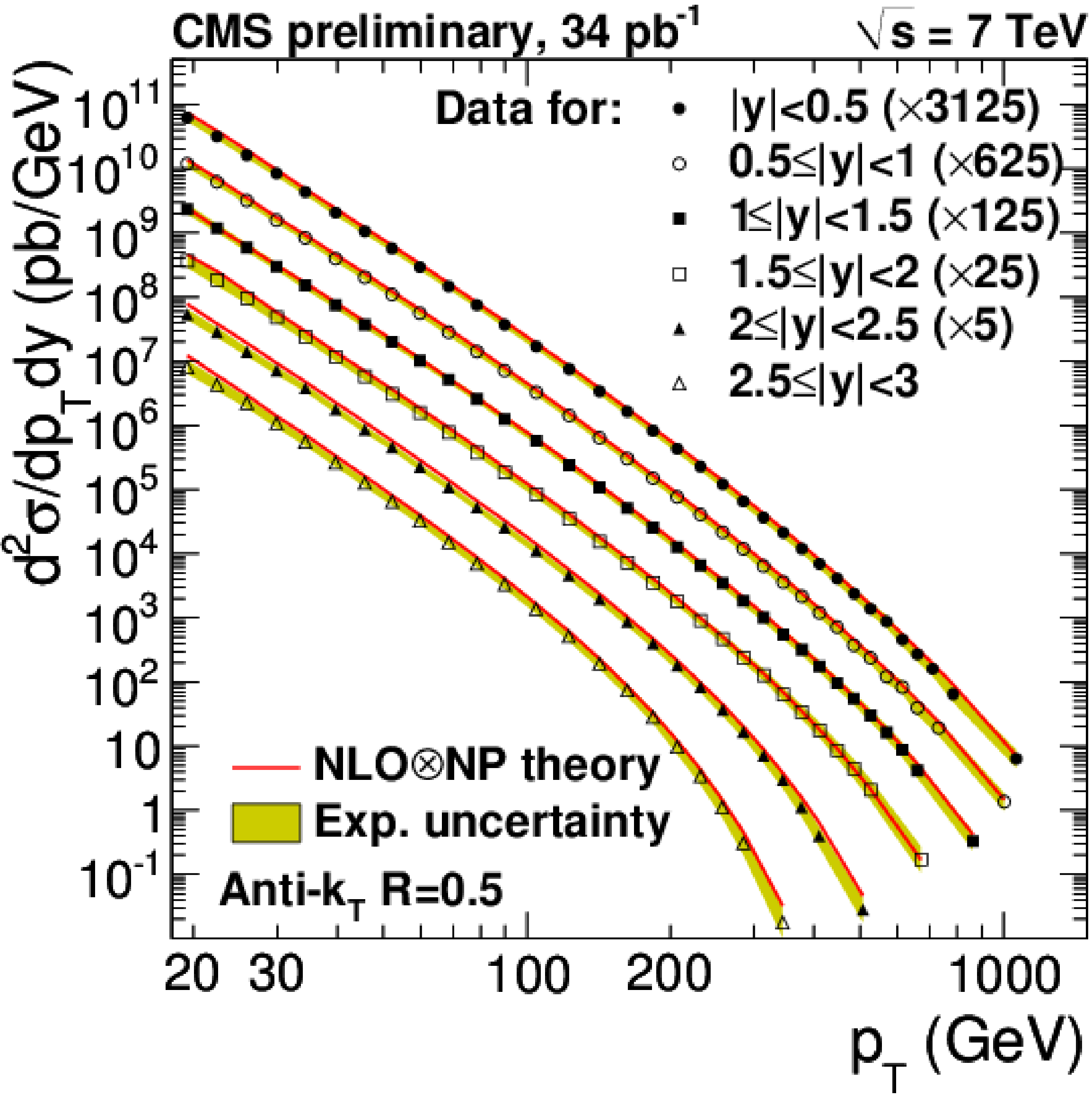
$$\frac{\delta N}{N} \approx 6 \cdot \frac{\delta E_T}{E_T}$$

so beware:

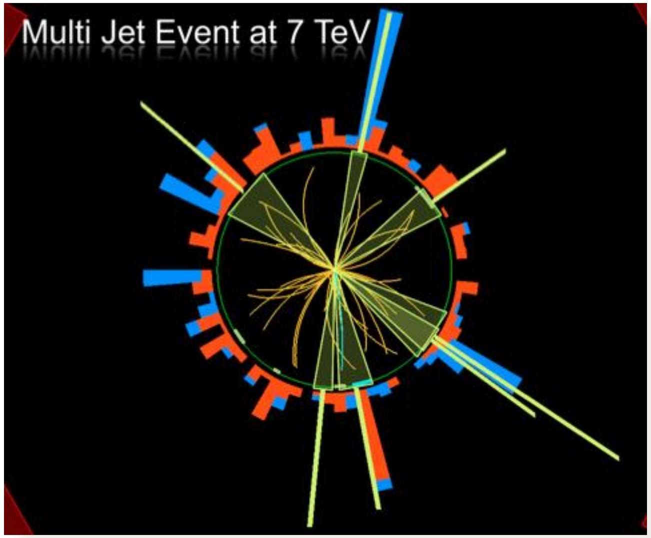
eg. an uncertainty of **5%** on absolute energy scale (calibration)

➒ an uncertainty of **30%** (!) on the measured cross section

Jet cross-section at the LHC



Data agrees with theoretical prediction over many orders of magnitude!

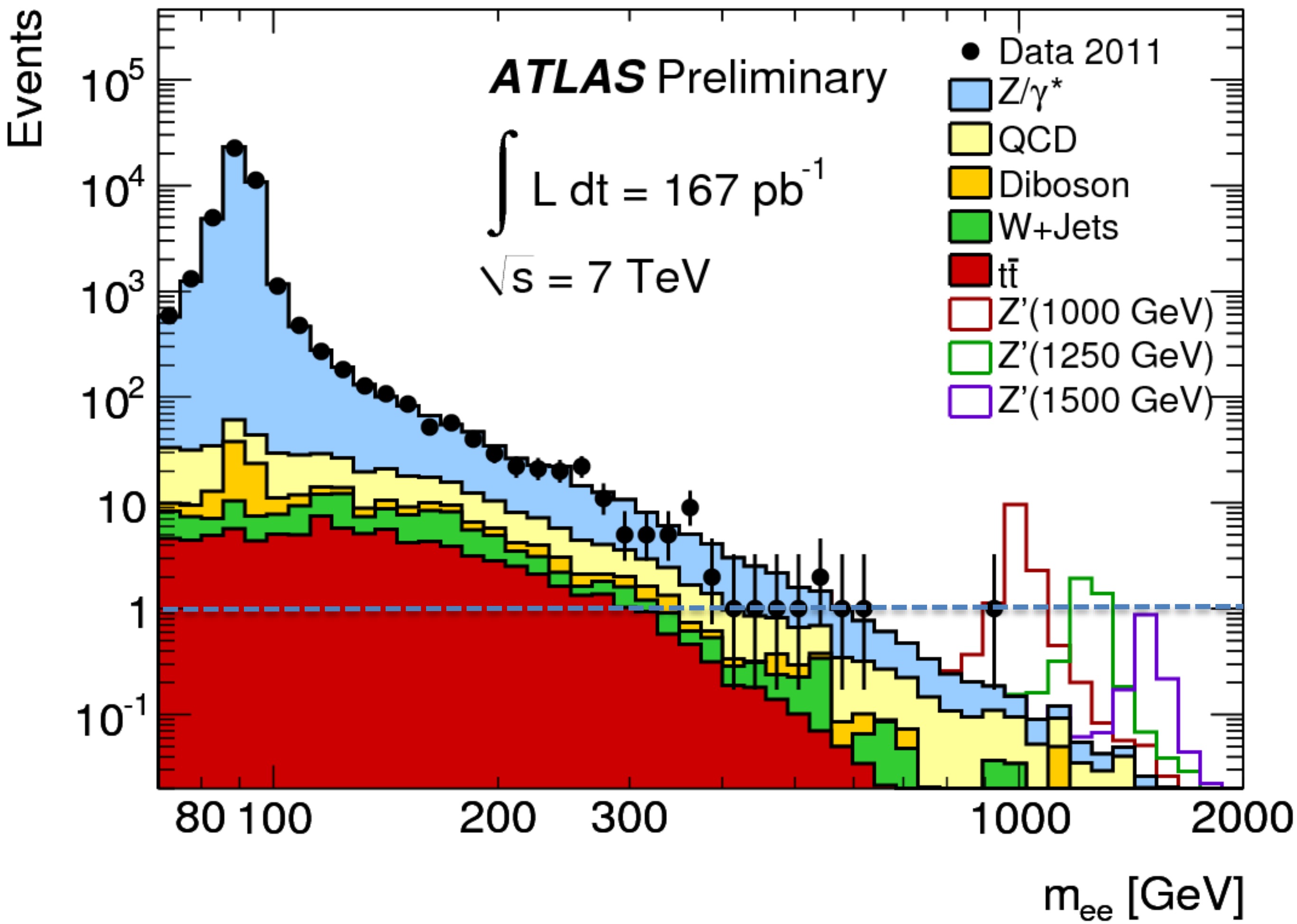


Example – search for a new heavy Z'

Many new physics models have a new heavy gauge boson which can decay to leptons.
Like a Z but heavier - called Z'.
Important to search for such new particles at the LHC.

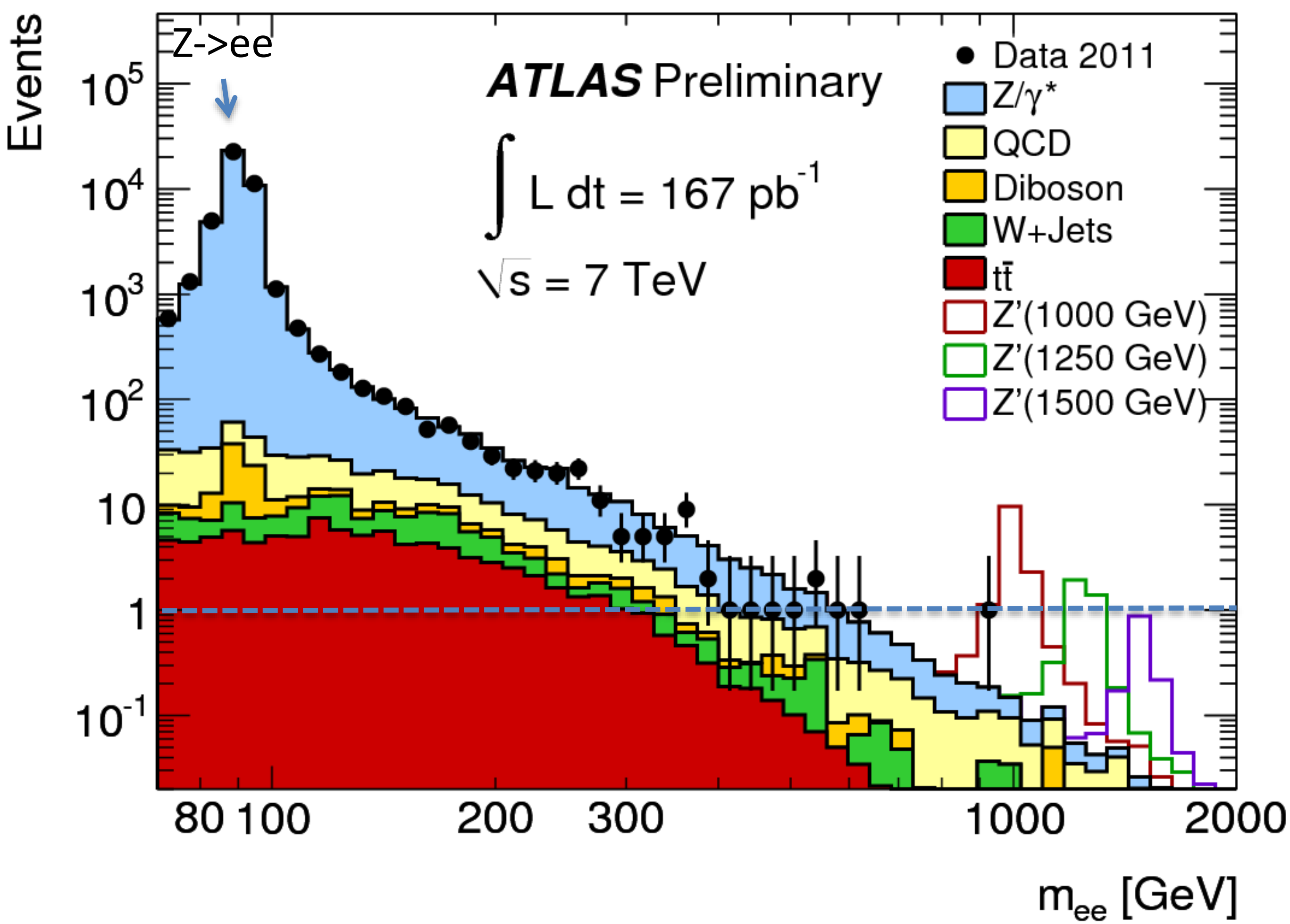
Example – search for a new heavy Z'

Like Z->ee but at higher mass.



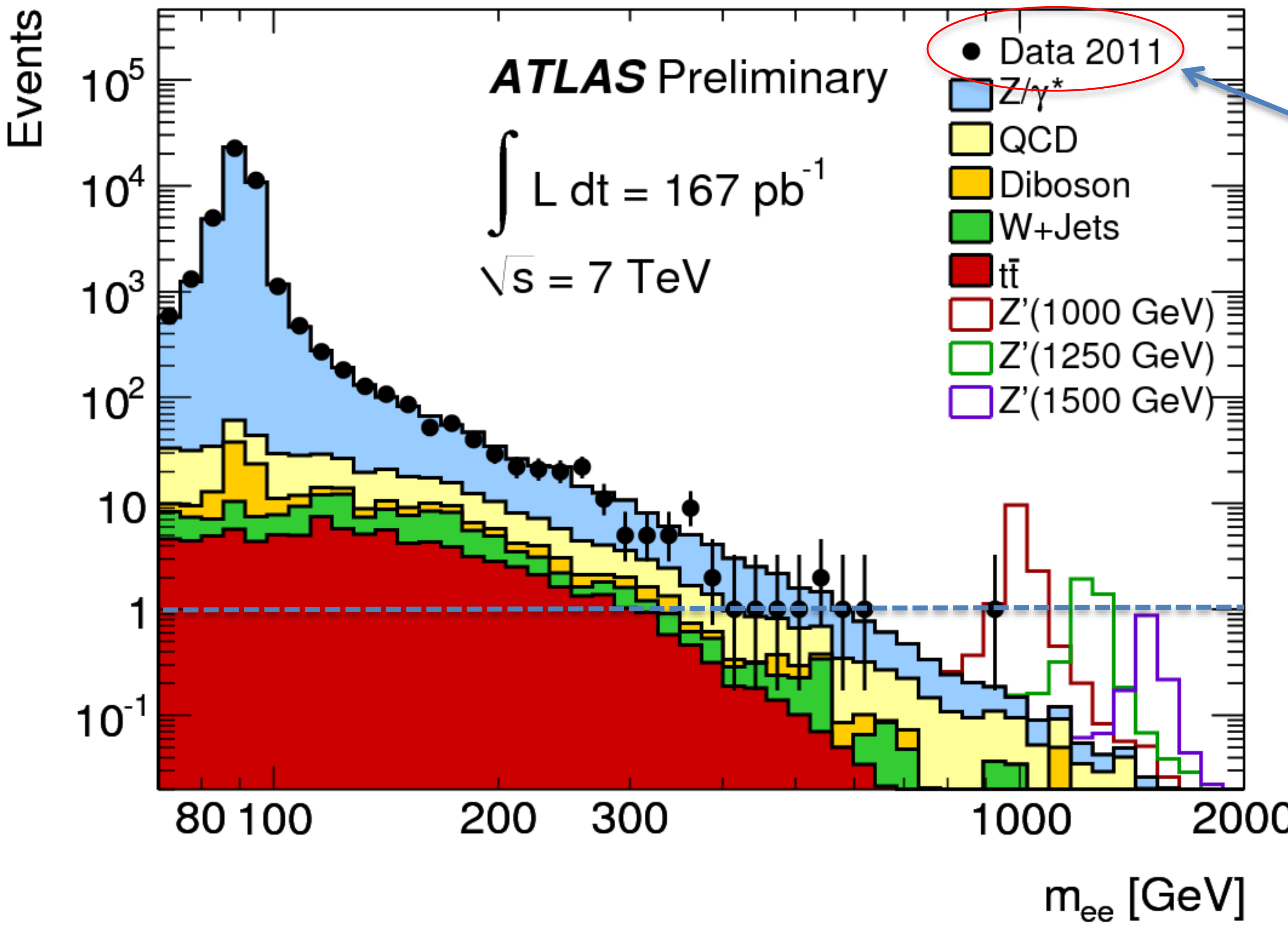
Example – search for a new heavy Z'

Like Z->ee but at higher mass.



Example – search for a new heavy Z'

Like Z->ee but at higher mass.

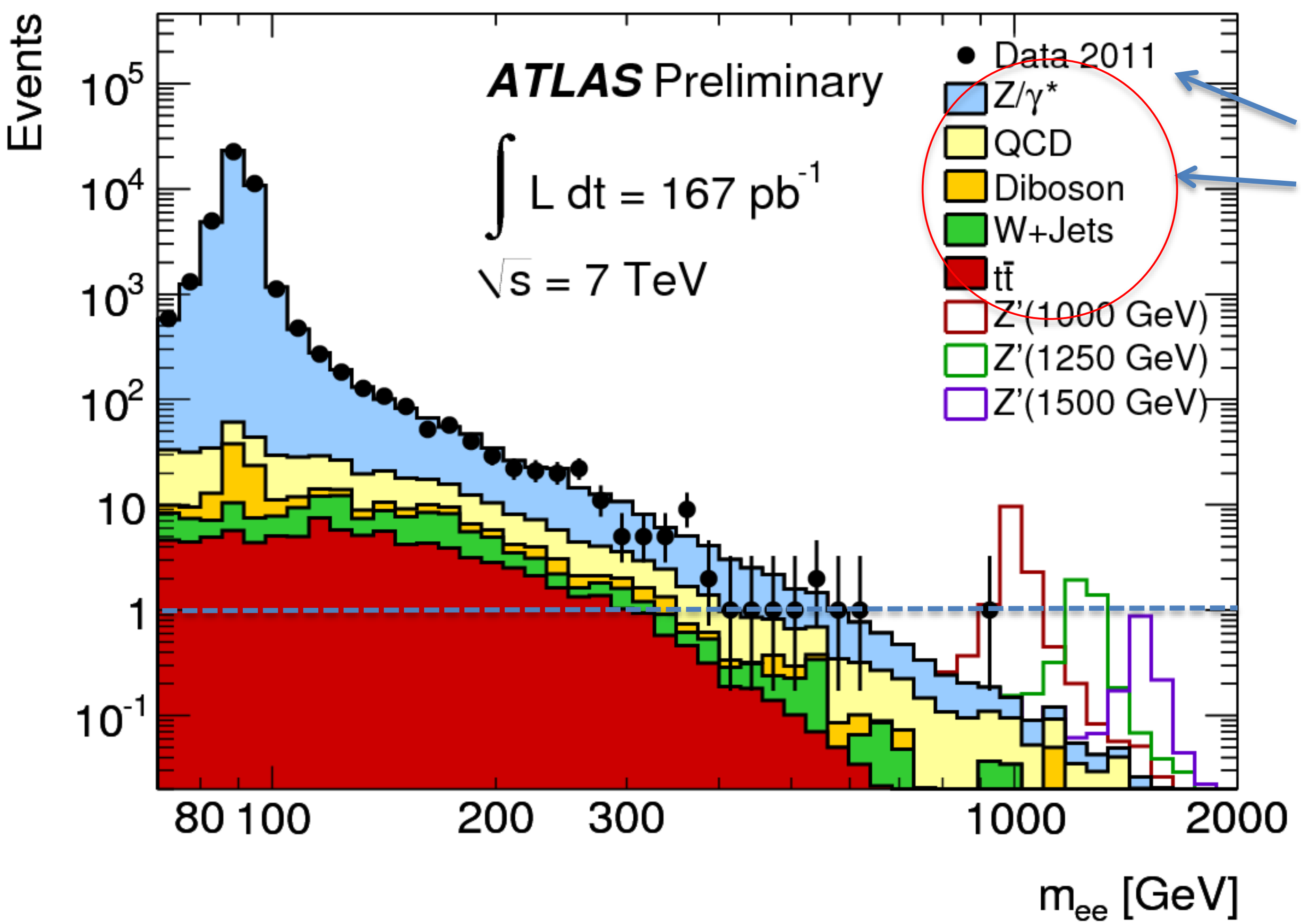


Select 2 electron candidates and plot their invariant mass for

1. **Data**

Example – search for a new heavy Z'

Like Z->ee but at higher mass.

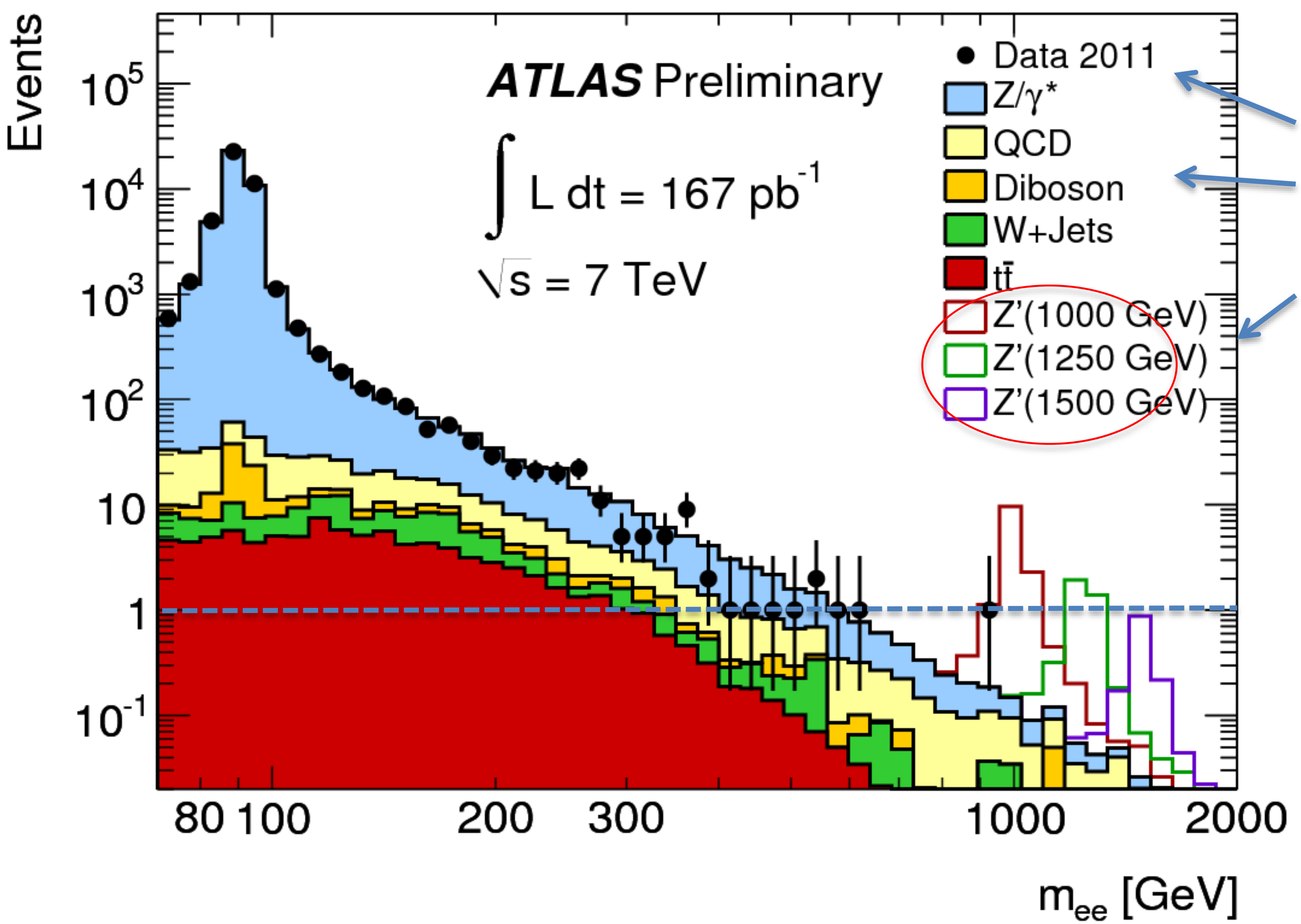


Select 2 electron candidates and plot their invariant mass for

1. Data
2. **Simulated backgrounds events**

Example – search for a new heavy Z'

Like Z->ee but at higher mass.

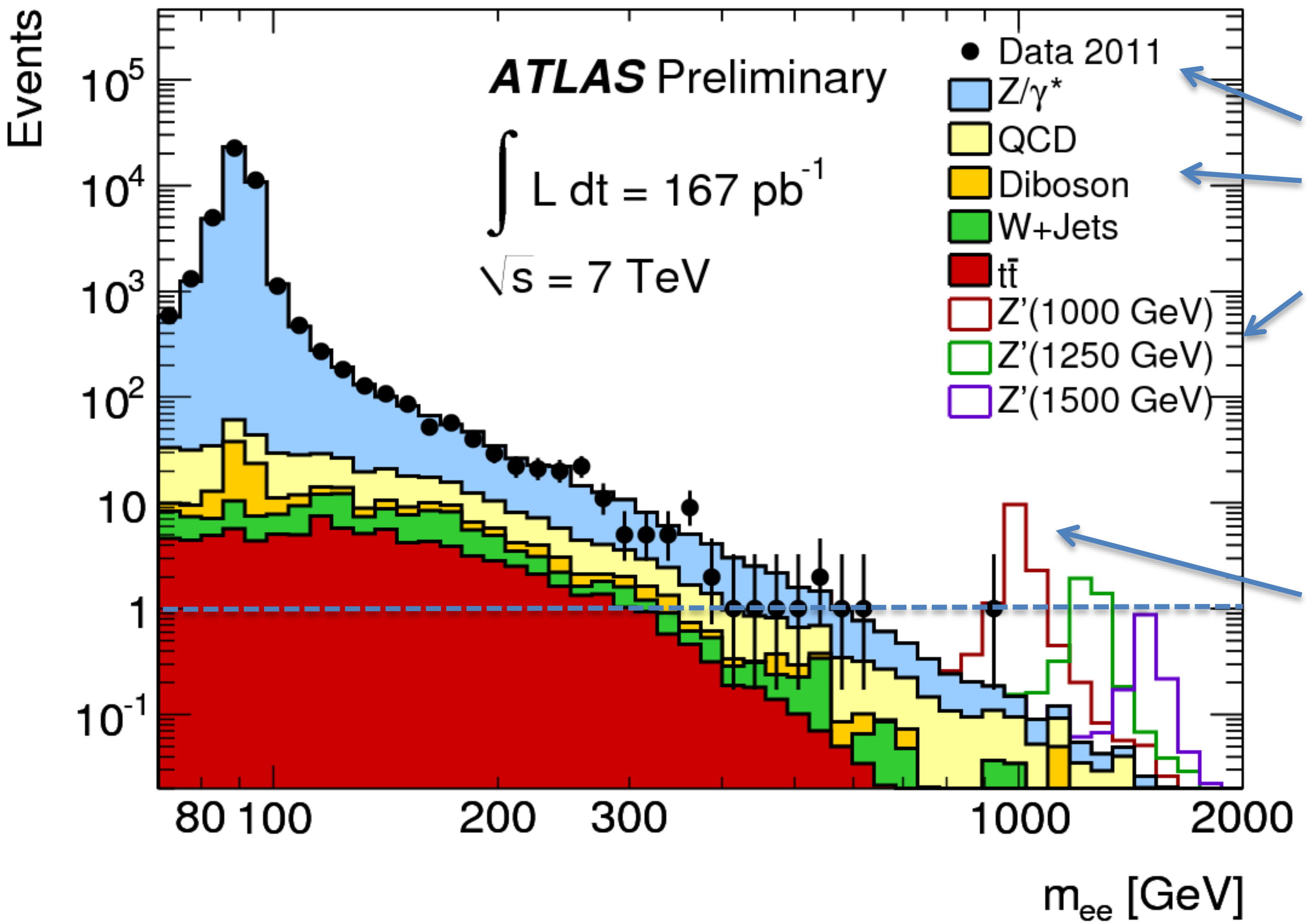


Select 2 electron candidates and plot their invariant mass for

1. Data
2. Simulated backgrounds events
3. **Simulated signal (Z') with different masses**

Example – search for a new heavy Z'

Like Z->ee but at higher mass.



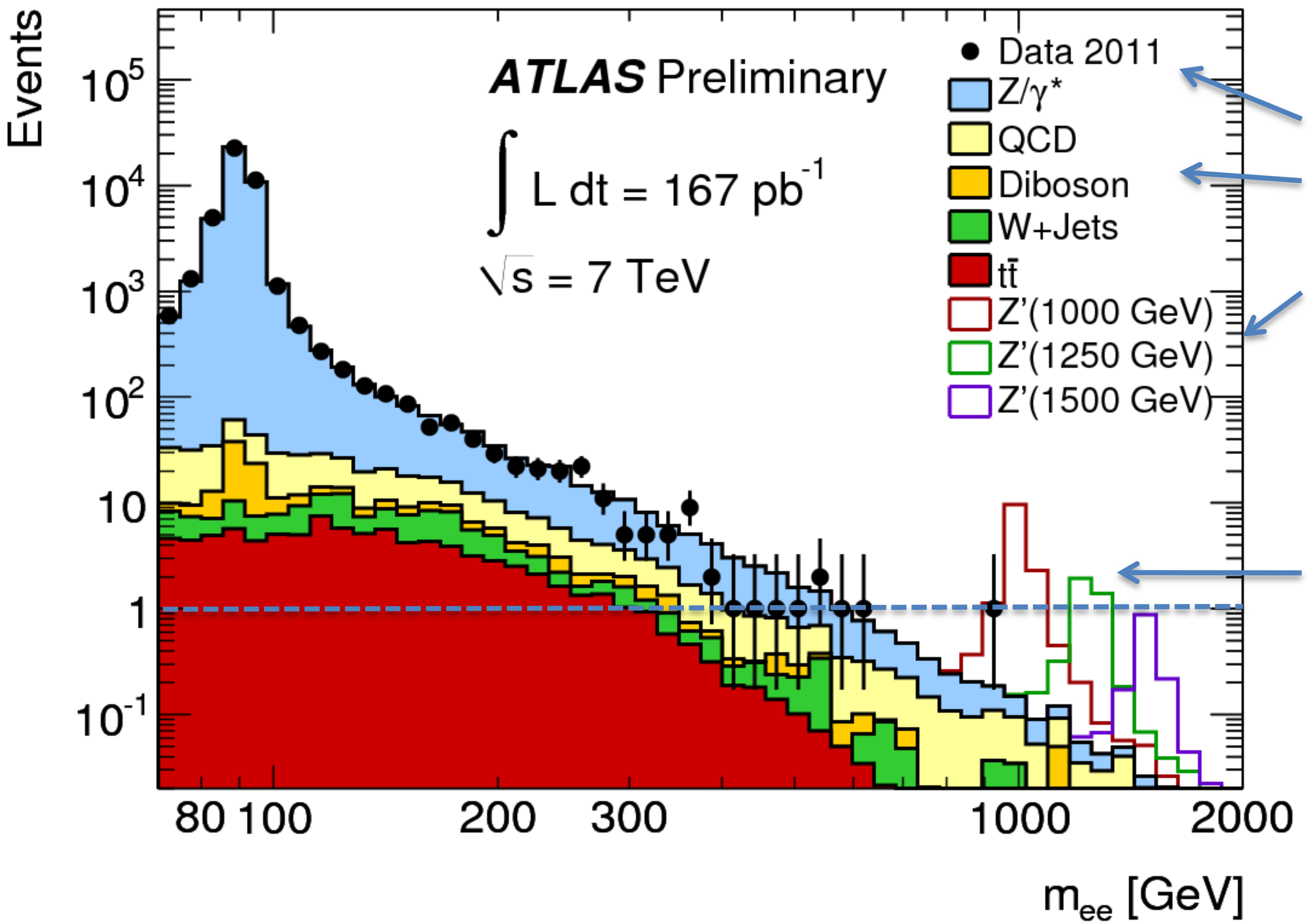
Select 2 electron candidates and plot their invariant mass for

1. Data
2. Simulated backgrounds events
3. Simulated signal (Z') with different masses

Data inconsistent with a 1TeV Z'

Example – search for a new heavy Z'

Like Z->ee but at higher mass.



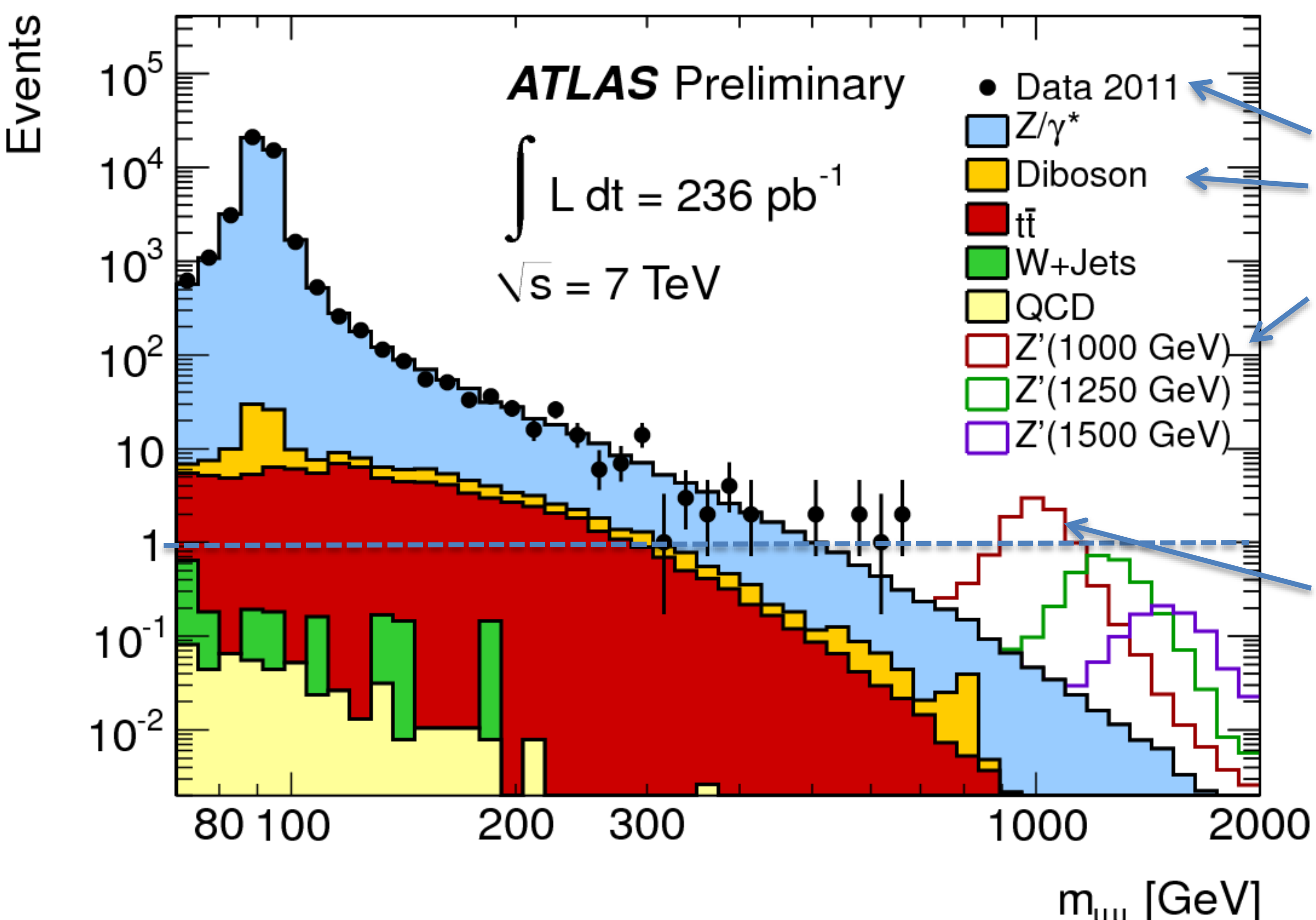
Select 2 electron candidates and plot their invariant mass for

1. Data
2. Simulated backgrounds events
3. Simulated signal (Z') with different masses

Cross-section decreases with mass (higher the mass of the Z', the more data needed to discover it)

Example – search for a new heavy Z’

Now for **muons**!



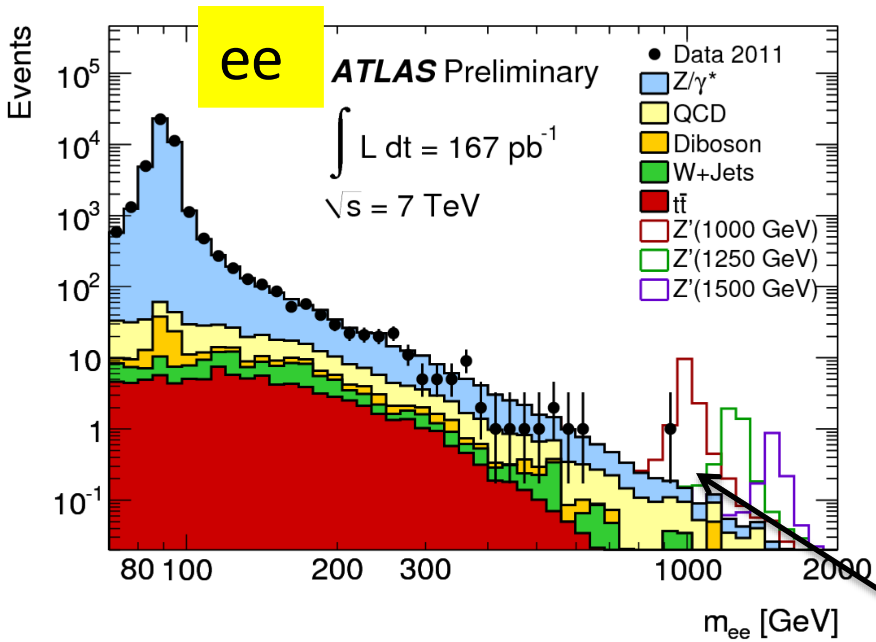
Select 2 muon candidates and plot their invariant mass for

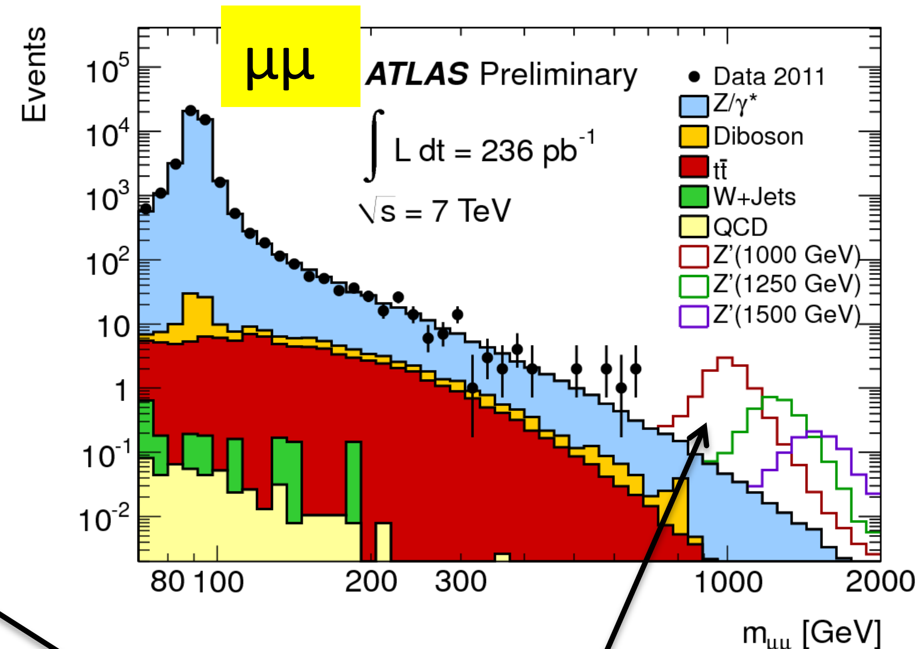
1. Data
2. Simulated backgrounds events
3. Simulated signal (Z’) with different masses

Data inconsistent with a 1TeV Z’

Z' analysis

Combining the electron and muon channel the data exclude Z' upto mass of 1.4 TeV
Need to take into account the statistical and systematic uncertainties!

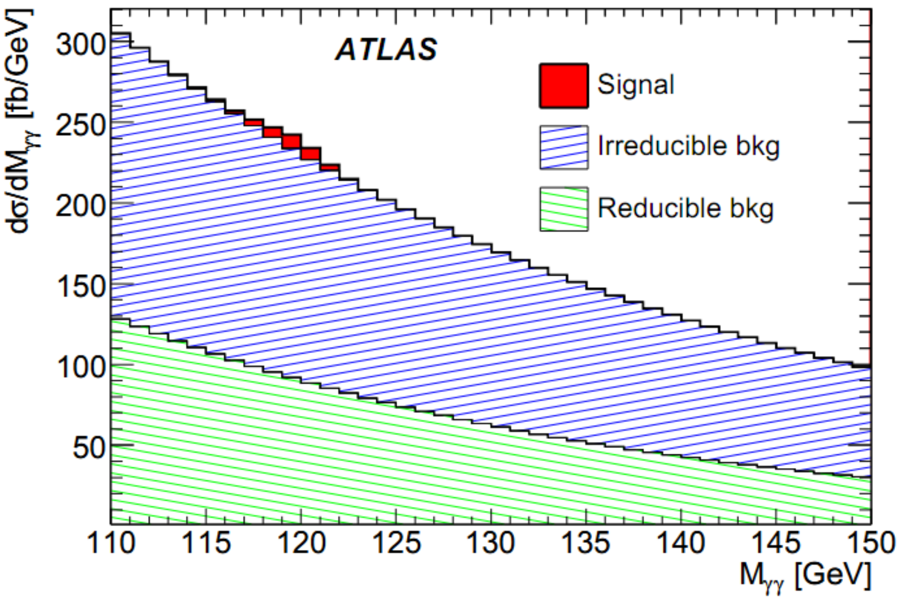


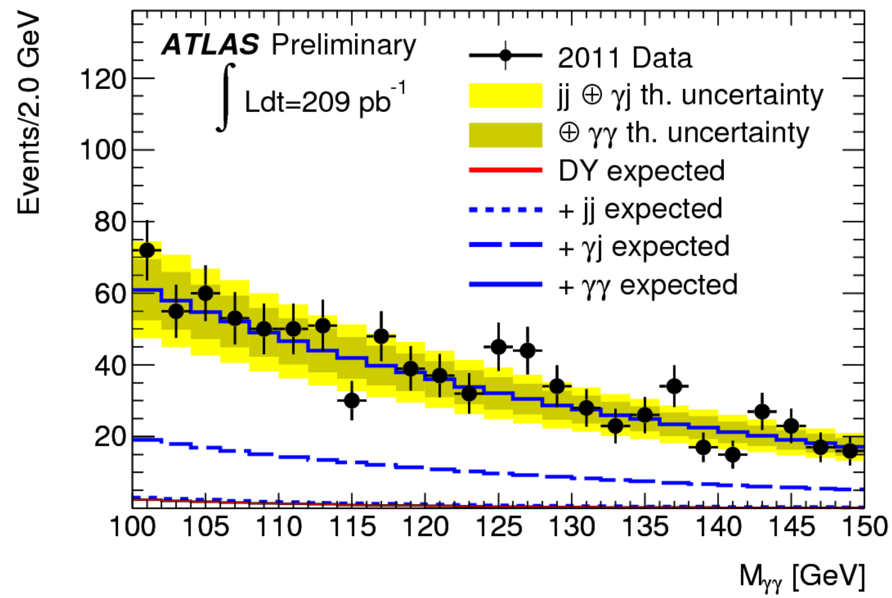


Simulation tells us that the 1TeV Z' is narrower in electron decay mode than muon decay mode
(Electron momentum resolution better at high energy)
Background composition different in the electron and muon channels.

H->γγ

H->γγ is the best way of discovering a low mass Higgs at the LHC.





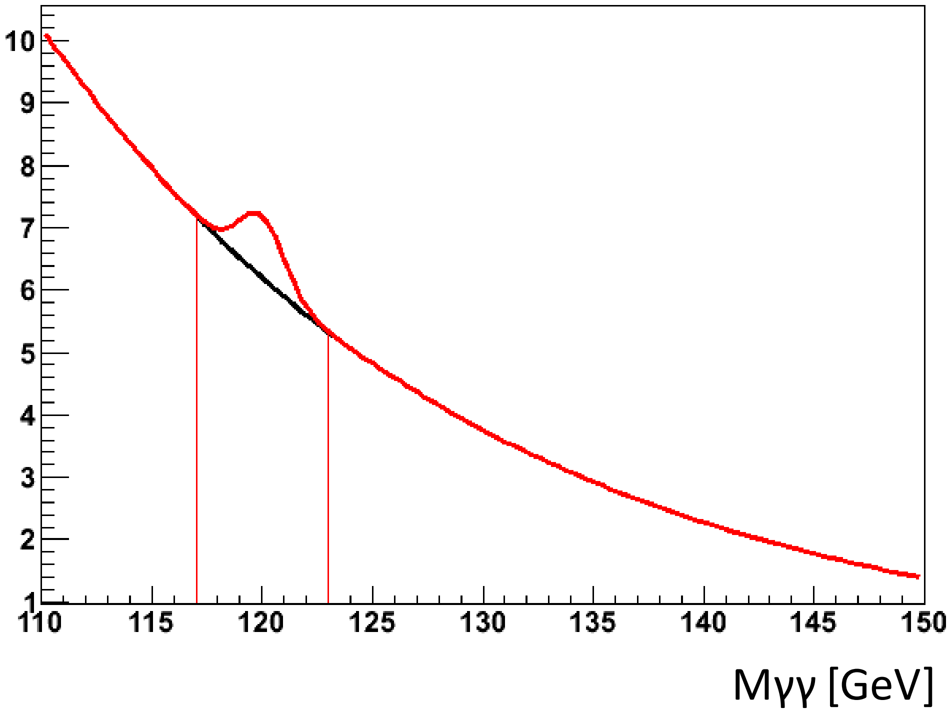
Tiny signal on top of large background. Good resolution is essential to being able to observe this.

No evidence of a bump in the ATLAS data (now we have 5x more data than this to look at!)
Result with ~1fb^{-1} will be presented later today at the EPS conference!
http://eps-hep2011.eu/

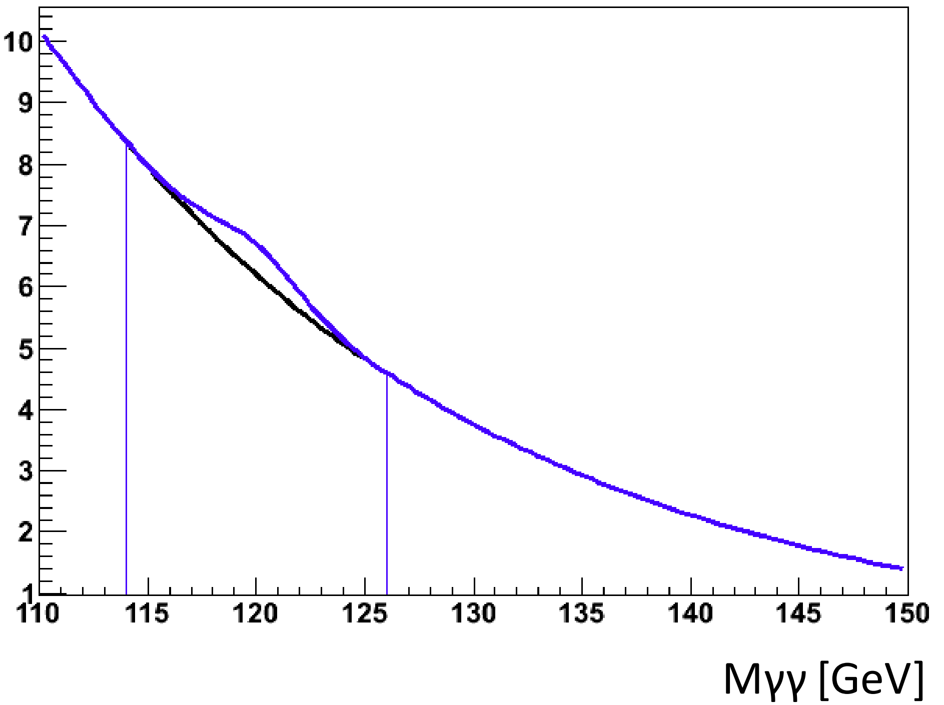
Importance of Resolution

Toy example: Signal peak on exponential background.
2 different signal resolutions. Same number of signal events in each peak
Would discover the left hand signal much quicker!

Mass resolution 1 GeV
Signal over background in cut range ~10%



Mass resolution 2 GeV
Signal over background in cut range ~5%



Very important to build the detector to give you the best resolution.
But also to optimize the reconstruction algorithms and calibrations to give the best resolution possible for that detector.

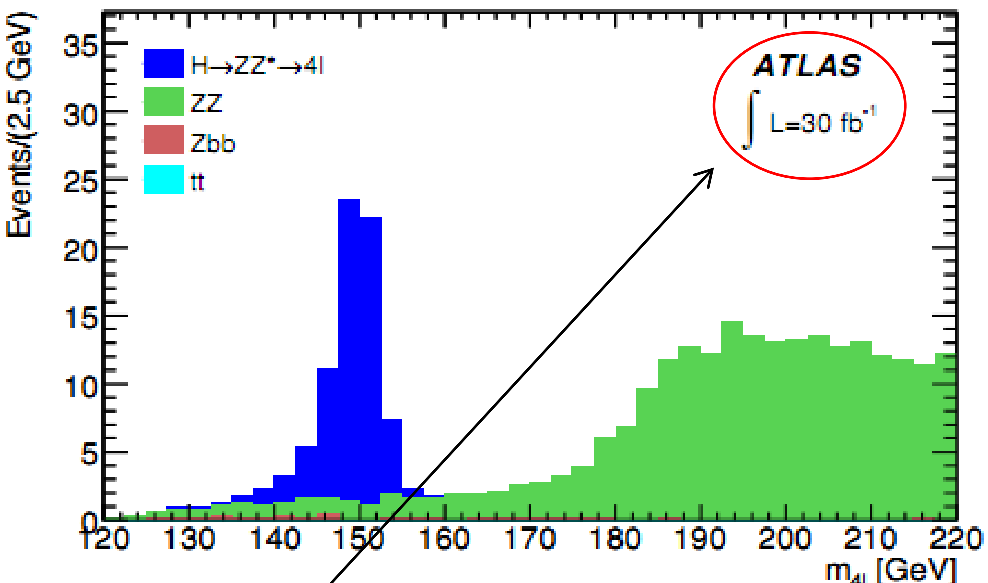
Another Example: H->ZZ

- Searching for the Higgs boson in the decay H->ZZ
- The Z can decay like
  - Z->qq (quarks seen as jets in the detector)
  - Z->ll (electrons, muons or taus)
  - Z->vv (neutrino's do not interact with the detector and so are only 'seen' as missing energy)
- H->ZZ->$l^+l^-l^+l^-$ is by far the easiest to detect experimentally (we only look for l = electron or muon, as the tau decays quickly and so is not directly seen in the detector)
- H->4l is called the 'golden mode' as experimentally it is by far the easiest
  - Leptons have low background
  - Leptons reconstruction has high efficiency
  - Leptons have good momentum resolution

Simple H->ZZ analysis steps

- 1. Select events which contain reconstructed: $e^+ e^- \mu^+ \mu^-$
- 2. make sure mass of the lepton pairs is consistent with the Z mass
- 3. Histogram the mass of the 4 leptons
- 4. See a peak corresponding to the Higgs (hopefully!)

Simulated data!



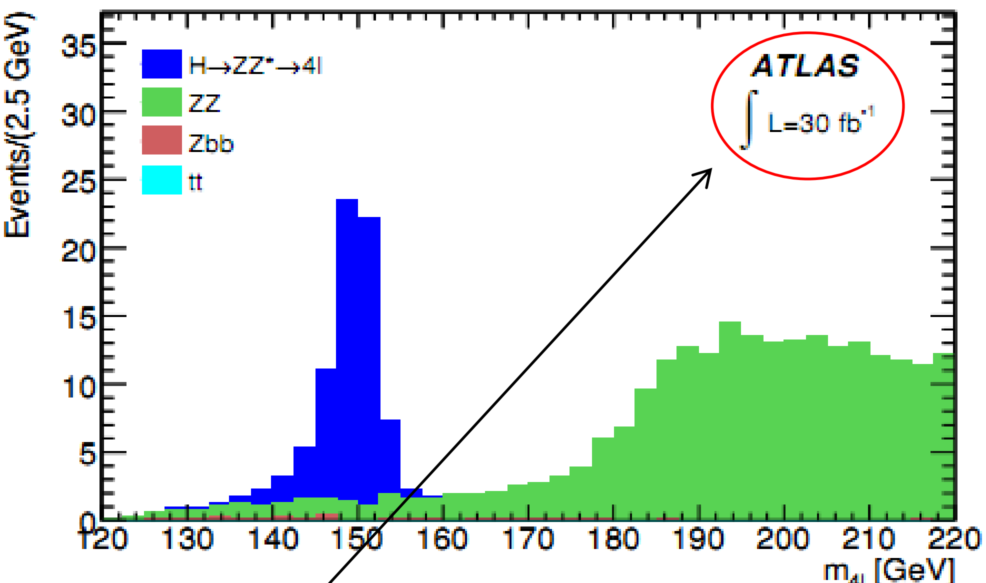
30 times more data than now!

Lepton efficiency very important for this mode.
If we have a 90% efficiency per lepton that gives 0.90^4=66% efficiency for the Higgs!

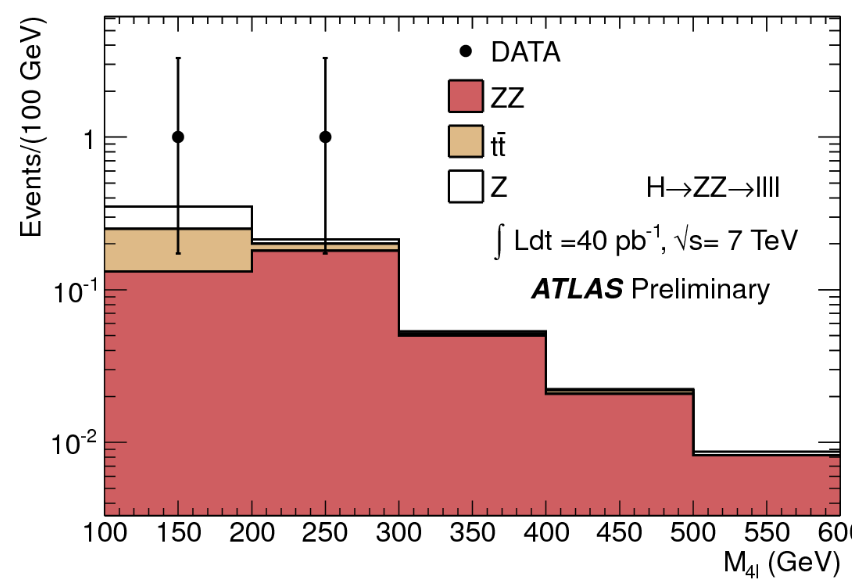
Simple H->ZZ analysis steps

- Where we are now:
  - Two ZZ events seen in $0.04 fb^{-1}$ of data in ATLAS first analysis
  - New analysis with $1 fb^{-1}$ to be shown at EPS conference later today (http://eps-hep2011.eu/)
  - Beginning to be very interesting! (depending on if the Higgs exists and with what mass)

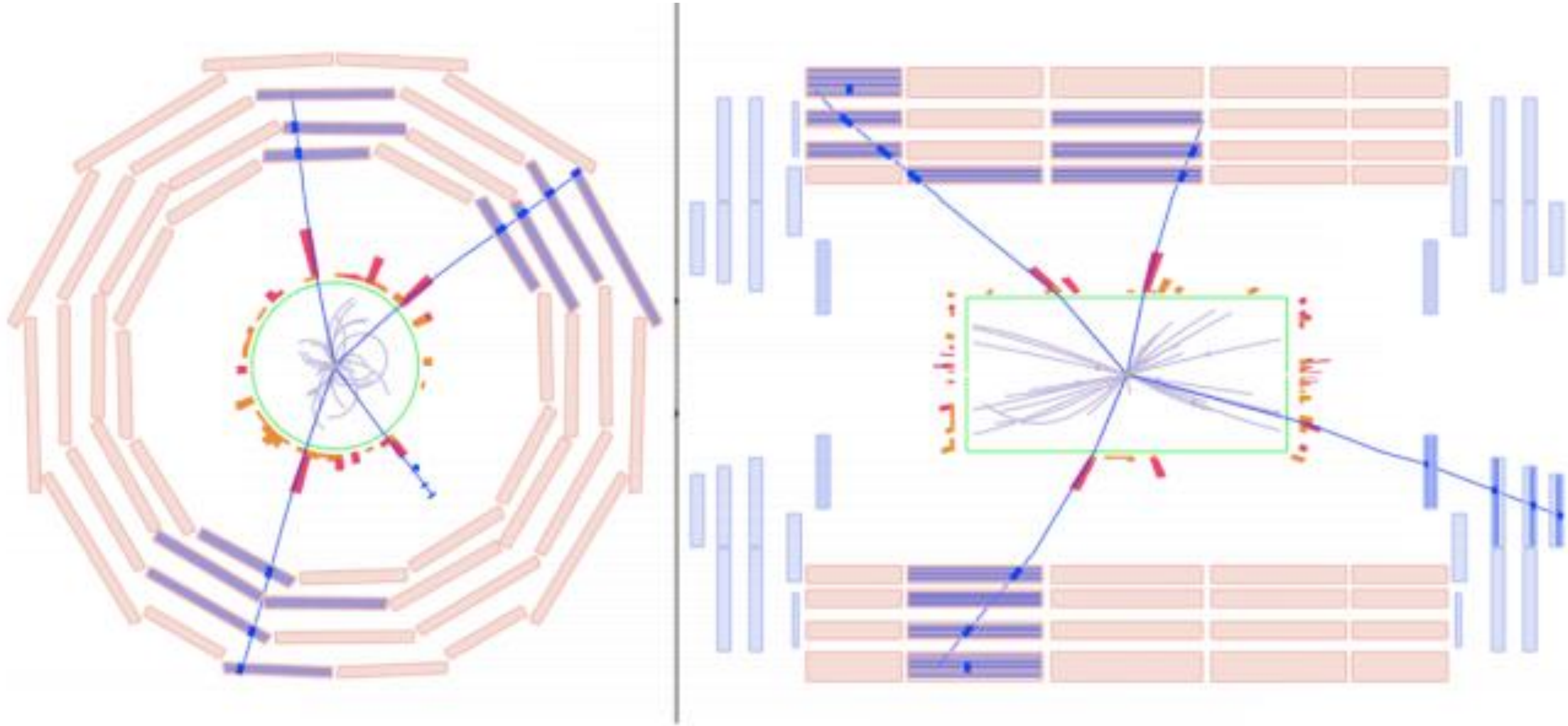
Simulated data!



30 times more data than now!



H->4μ event in CMS



Invariant Masses

$\mu_0 + \mu_1$: 92.15 GeV (total(Z) p_T 26.5 GeV, ϕ -3.03),
$\mu_2 + \mu_3$: 92.24 GeV (total(Z) p_T 29.4 GeV, ϕ +.06),
$\mu_0 + \mu_2$: 70.12 GeV (total p_T 27 GeV),
$\mu_3 + \mu_1$: 83.1 GeV (total p_T 26.1 GeV).

Invariant Mass of 4μ: 201 GeV

Summary from Physics examples

- Data Quality very important to not include junk data in the analysis
- Jet cross-section
  - Energy calibration uncertainty leads to a big uncertainty due to the sharply falling spectrum
- Z':
  - use of simulation to see what a new physics signal would look like
- H->γγ:
  - importance of resolution
  - Importance of background rejection
- H->ZZ->4l:
  - importance of high efficiency
- Many other types of physics analyses (measuring cross-sections, masses, lifetimes)
  - Require also accurate knowledge of the efficiency, and the resolutions and the background rates

Se… Computing behind this all…

Somewhere, something went terribly wrong

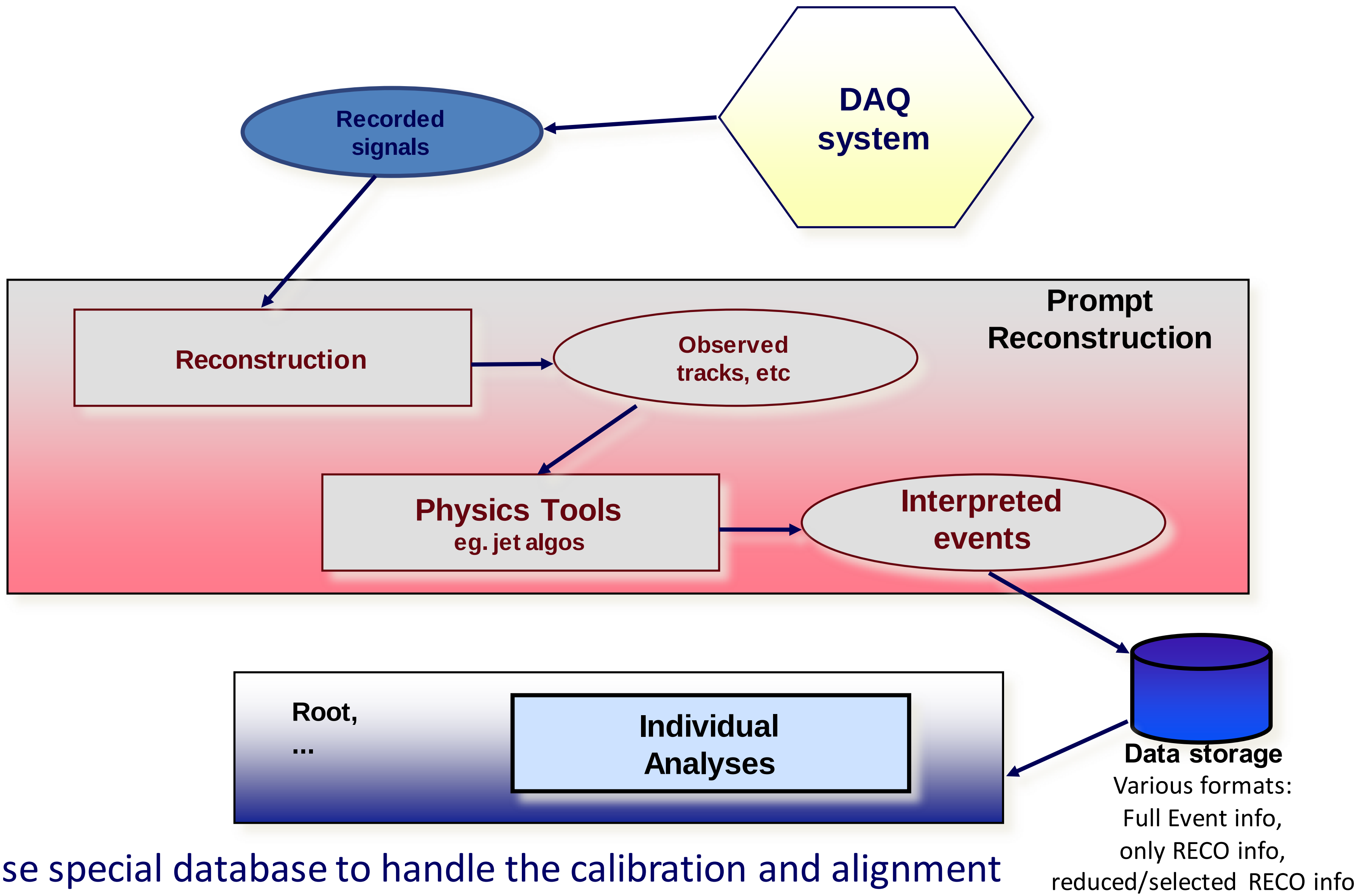
Some numbers

- Examples from ATLAS
  - Rate of events streaming out from High-Level Trigger farm ~300 Hz
  - each event has a size of the order of 1.5MB
  - 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
  - Reco time per event on std. CPU: < 15 sec (on lxplus)
  - increases with pileup (more combinatorics in the tracking)

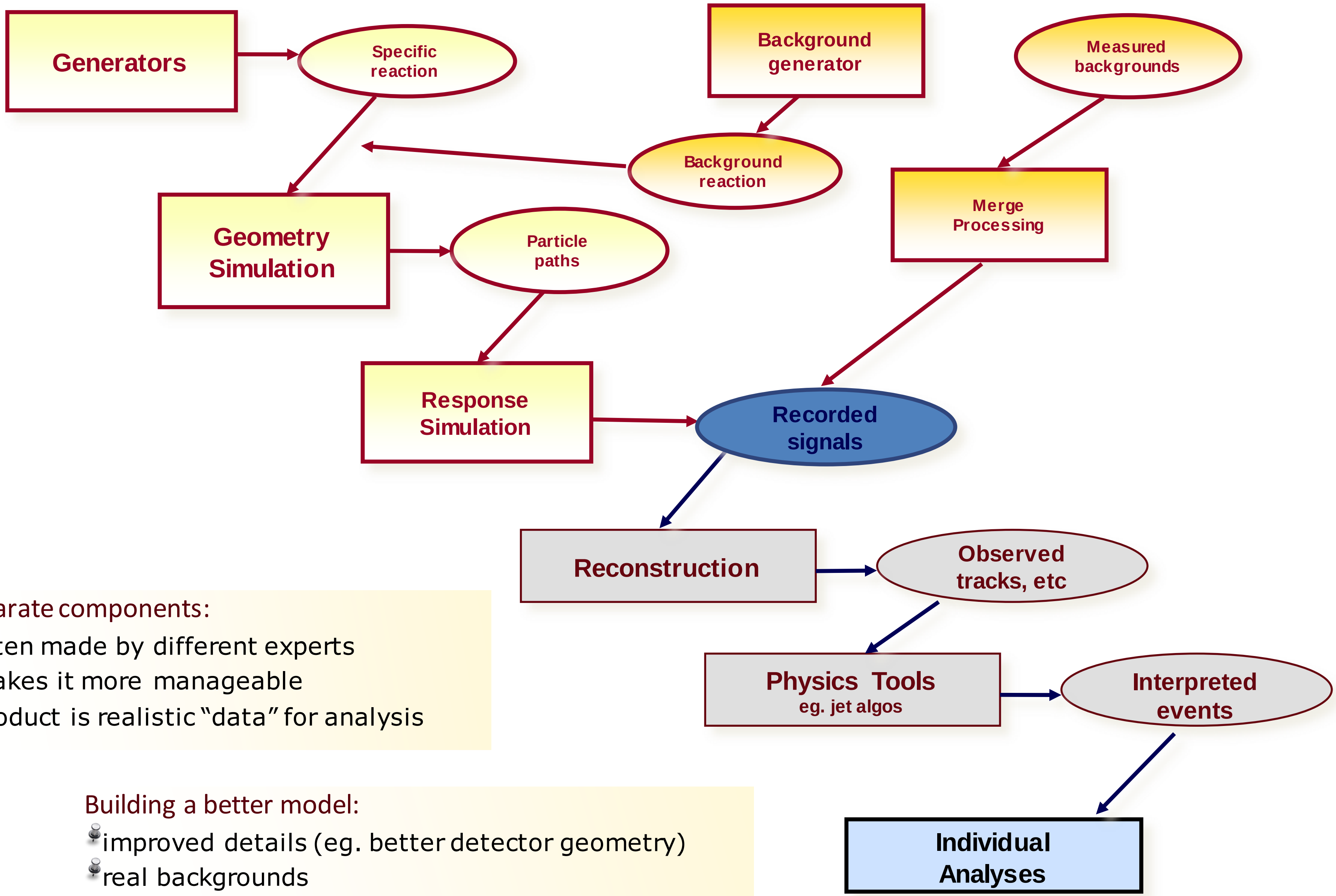
- simulating a few billions of events
  - are mostly done at computing centres outside CERN
  - Simulation very CPU intensive
- ~4 million lines of code (reconstruction and simulation)
  - ~1000 software developers on ATLAS

RECO flow



- Use special database to handle the calibration and alignment data needed in reconstruction

Flow of simulated data



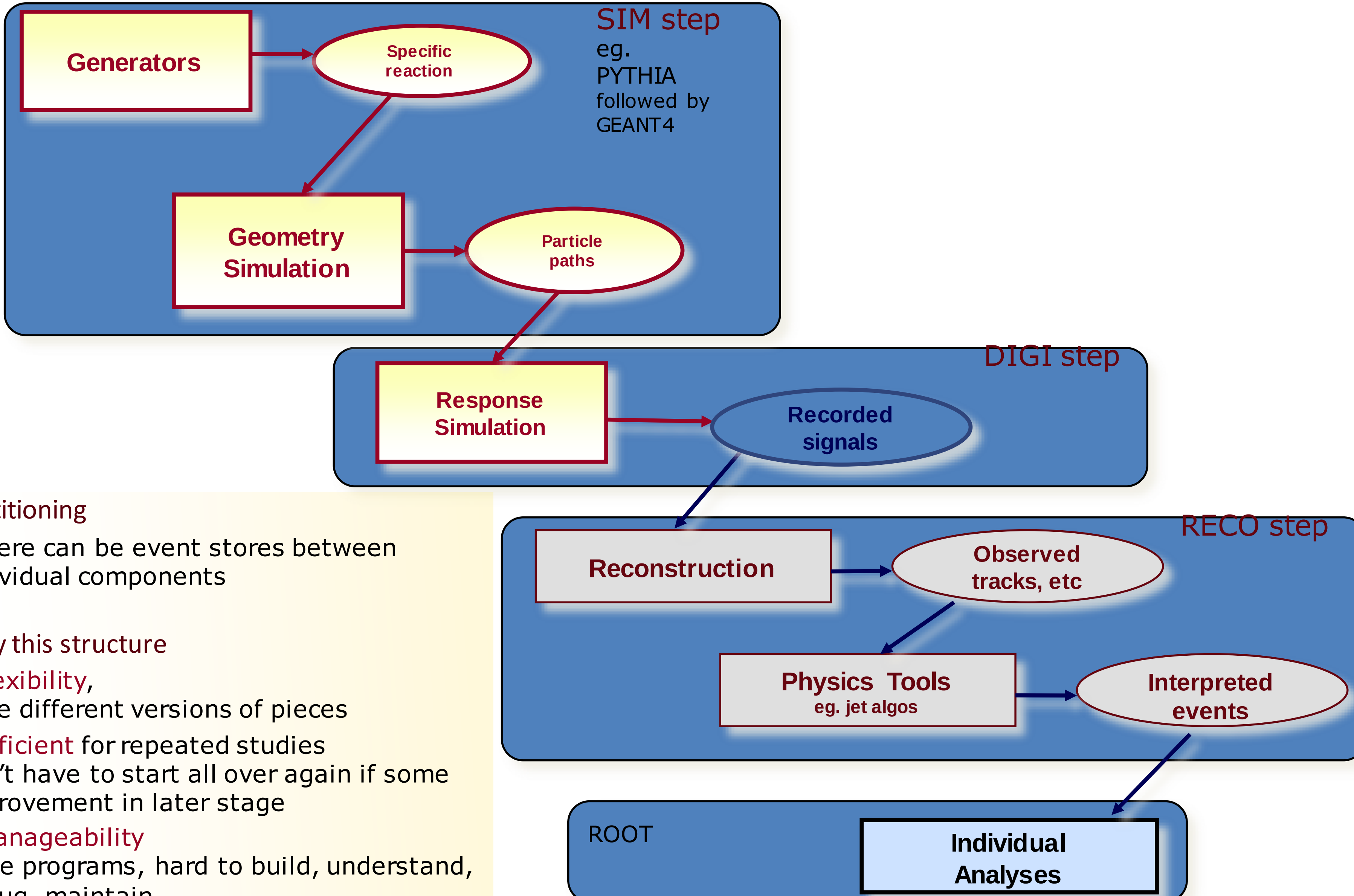
Separate components:

- 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 systems



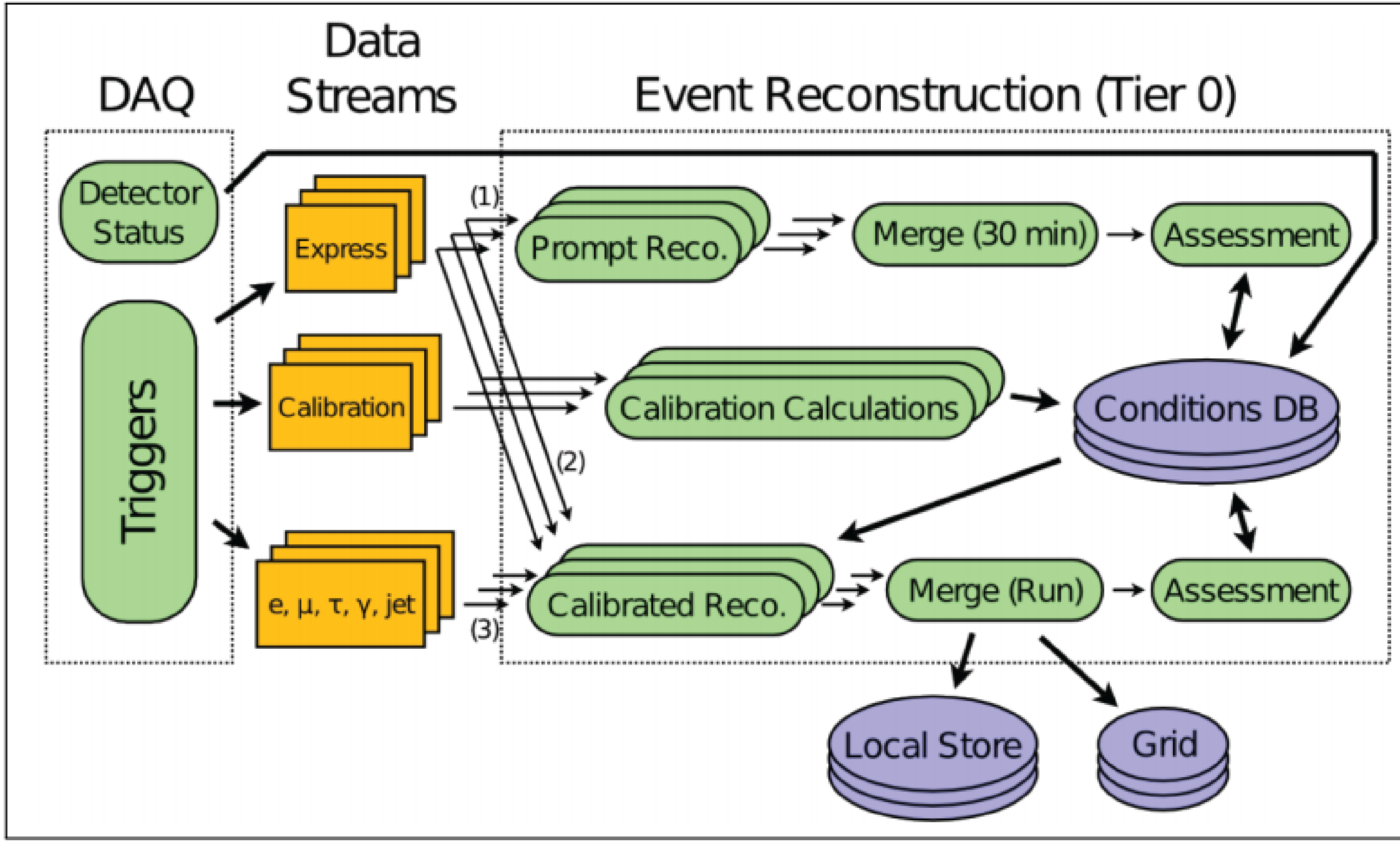
Partitioning

- 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, ...

Prompt reconstruction



In ATLAS we reconstruct the data ~36hours after it is recorded.

This time is used to derive updated calibrations from the data, that are needed in the reconstruction.

Once a year we reprocess all the data with updated software and calibrations.

Conclusions

- Reconstruction and Analysis
  is how we get from raw data to physics papers

- Sophisticated reconstruction algorithms + calibrations and alignments needed
  - High efficiencies, good resolutions and low fake rates
  - Important to get the best physics out of the experiment
- Detailed simulation also plays a key role
- Complex software infrastructure needed to be able to obtain the final physics results
- Any discoveries at the LHC will rely on the data-quality, simulation, reconstruction and analysis chain working well

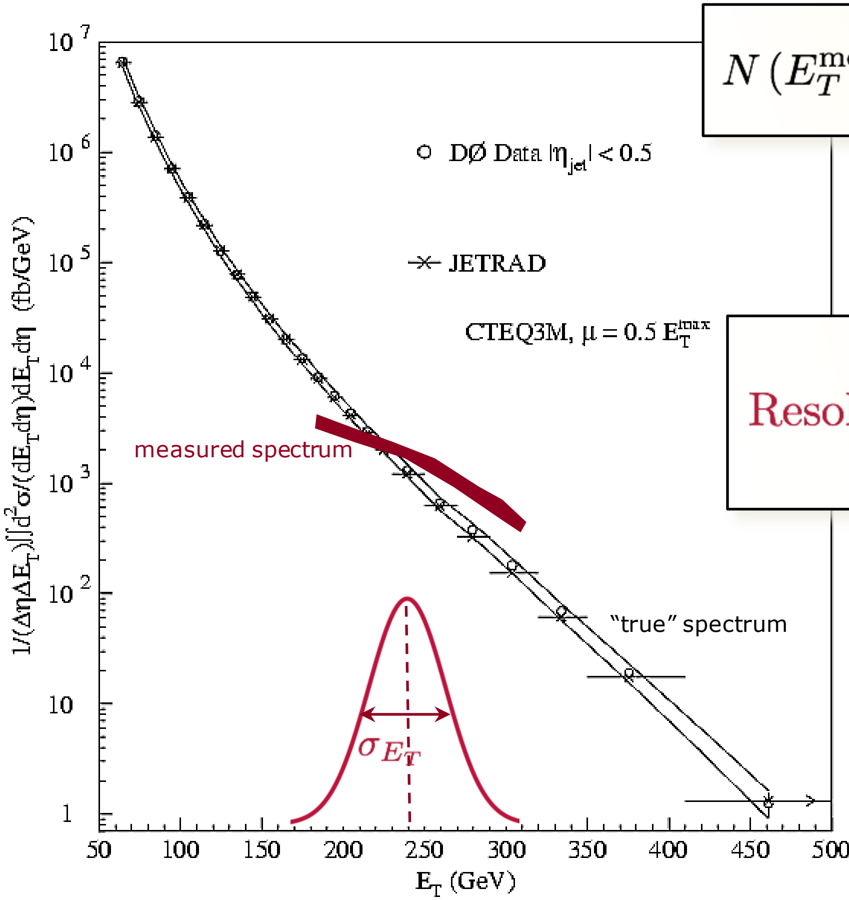
- Even to me it is often a miracle that we can generate wonderful results from these complicated instruments!

Final remarks

- This is **GREAT** time to be a CERN summer student
- The experiments have a lot of high quality data
- The LHC is working great
- The experiments all are working very well
- There could be an exciting physics discovery at any time
- Work hard
    - Learn what you can!
- And most of all - Enjoy yourself!!!

Problem 2 : Energy resolution

- The energy resolution can distorts the spectrum
- Again : Critical because of very steeply falling spectrum!



$$N\left(E_T^{\mathrm{meas}}\right) = \int_0^{\infty} N\left(E_T^{\mathrm{true}}\right) \cdot \mathrm{Resol}\left(E_T^{\mathrm{meas}}, E_T^{\mathrm{true}}\right) \mathrm{d}E_T^{\mathrm{true}}$$

eg. Gaussian resolution function

$$\mathrm{Resol}\left(E_T^{\mathrm{meas}}, E_T^{\mathrm{true}}\right) \propto \exp\left[-\frac{\left(E_T^{\mathrm{meas}} - E_T^{\mathrm{true}}\right)^2}{\sigma_{E_T}^2}\right]$$

so beware:

A bad energy resolution can distort the true spectrum

➒ have to determine the energy resolution

➒ have to "unfold" the measured spectrum

➒ problem is minimized if bin width ~ σ_{E_T}