Physics Analysis at LHC-II

Shamona Fawad Qazi
National Centre for Physics

Where do we stand now?

- After data flow from DAQ: data reduction and abstraction
  - reconstruct tracks, energy deposits in calorimeters
  - calculate high level physics quantities e.g. momentum
  - apply even higher-level algorithms e.g. 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 e.g. from MC simulation
  - if distributions are measured : take care of calibrations and effects due to detector resolution → correct for these effects
  - determine statistical and systematic uncertainties
  - Compare with theory, found a deviation, something new?

A Real Analysis

- Select decay channel
- Define the variable to be measured
- Determine the backgrounds
- Obtain signal variable distribution for required process as well as background processes using Monte carlo
- Make a best fit to find mean, RMS and other statistical parameters
- Repeat the same procedure for real data
- Include the detector parameters and make a measurement
- Study systematic errors

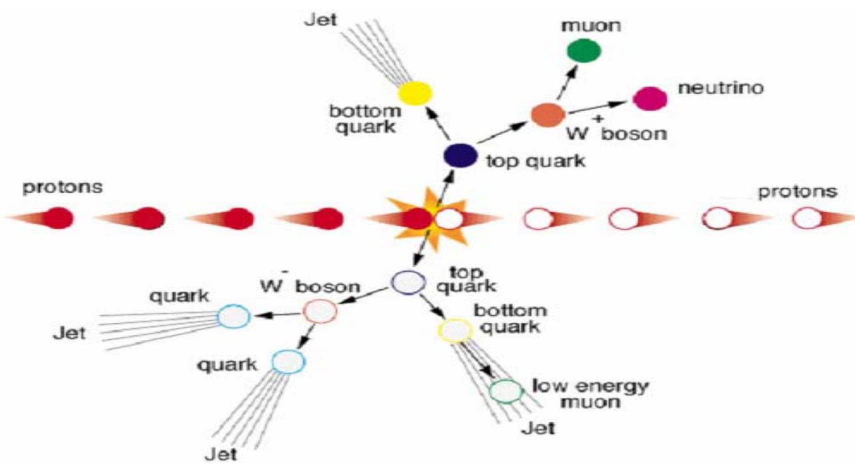
Selection of Decay Channel

- Have all the necessary physics objects (lepton, photons, jets) in our hand to start doing some real measurement
- First of all select a decay channel for analysis; this will define our end products
- For example

$$\bar{t}t \to \bar{b}bW^{+}W^{-} \to \bar{b}b\mu\nu_{\mu}q_1q_2$$

– where q_1 gives one jet and q_2 gives the other jet

– b and $\bar{b}$ will also give jets

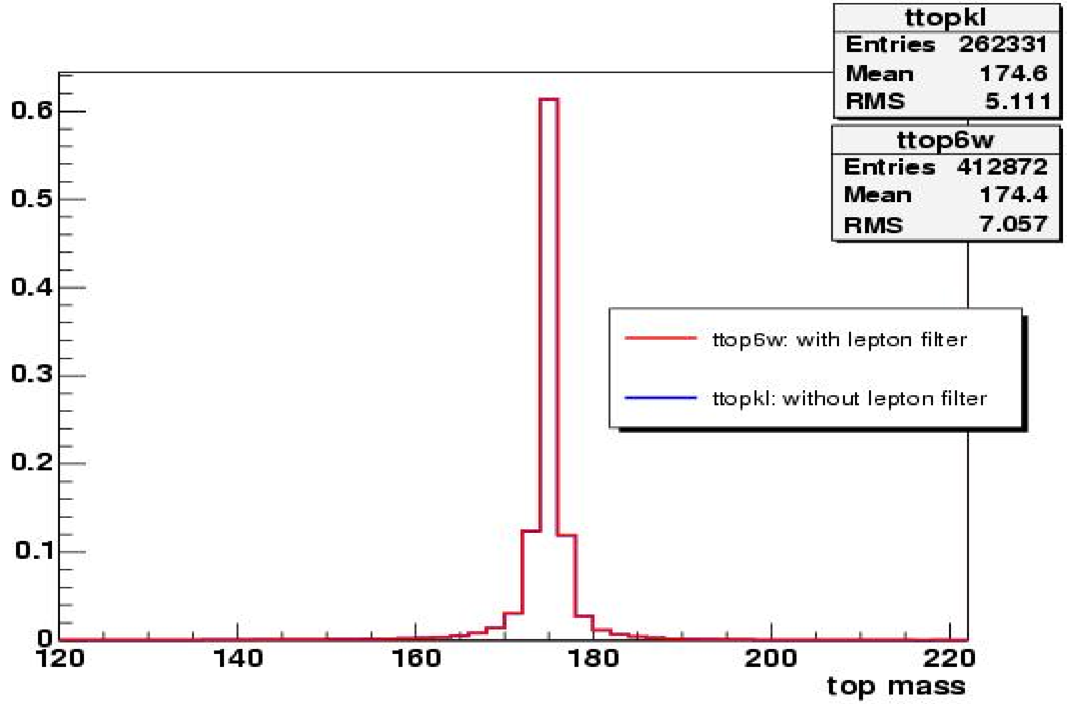


- Determine the possible backgrounds for the channel
  – Background processes are those which have similar end products

Selection of Variable

- Choose variable easy to work with e.g. mass of t-tbar (combined from all final state particle)
- A plot of invariant mass of final state particle gives a distribution for
  - Signal
    - ★ Should be narrow and high
  - Background
    - ★ Could be wider or narrower than signal
  - If signal and background are of similar width then selection criteria for events (usually know as cuts) are chosen such so that the signal distribution becomes narrow and can be differentiated from background → allows the measurement of the variable

Sample Variable : Top Mass



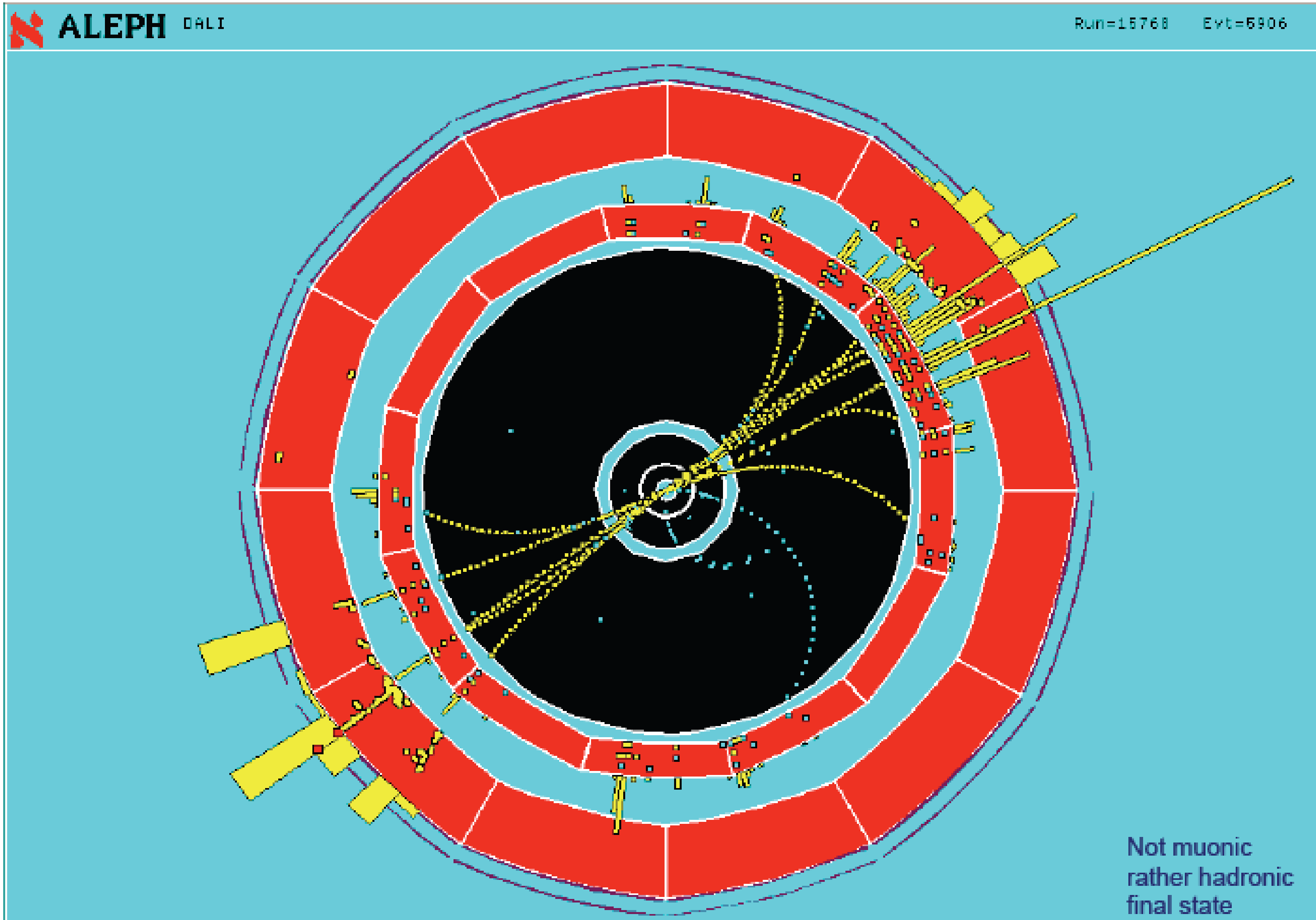
Top Mass plotted using two different cuts

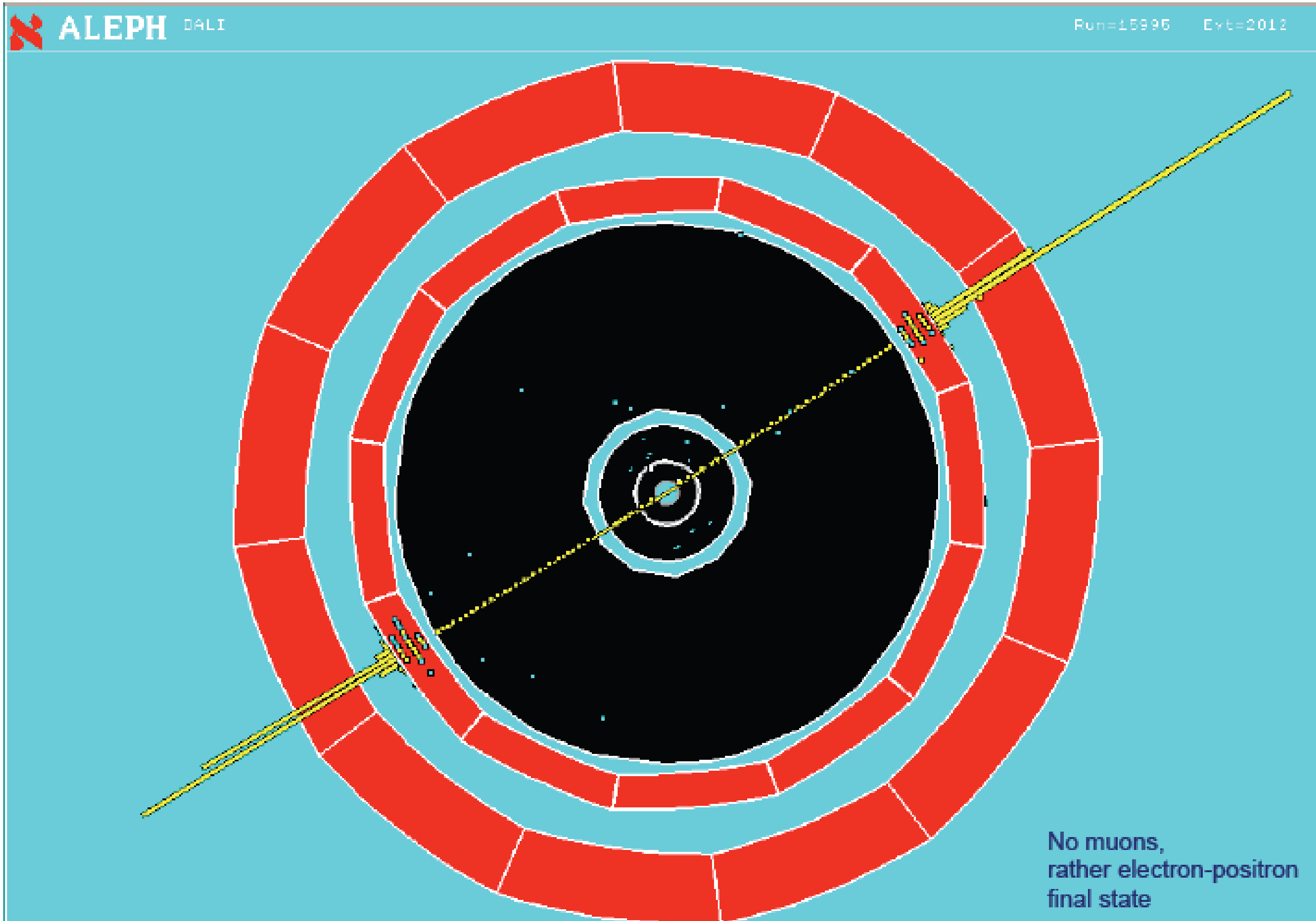
These overlaid top mass distributions show that the effect of lepton filter is very small.

Some Other Variables

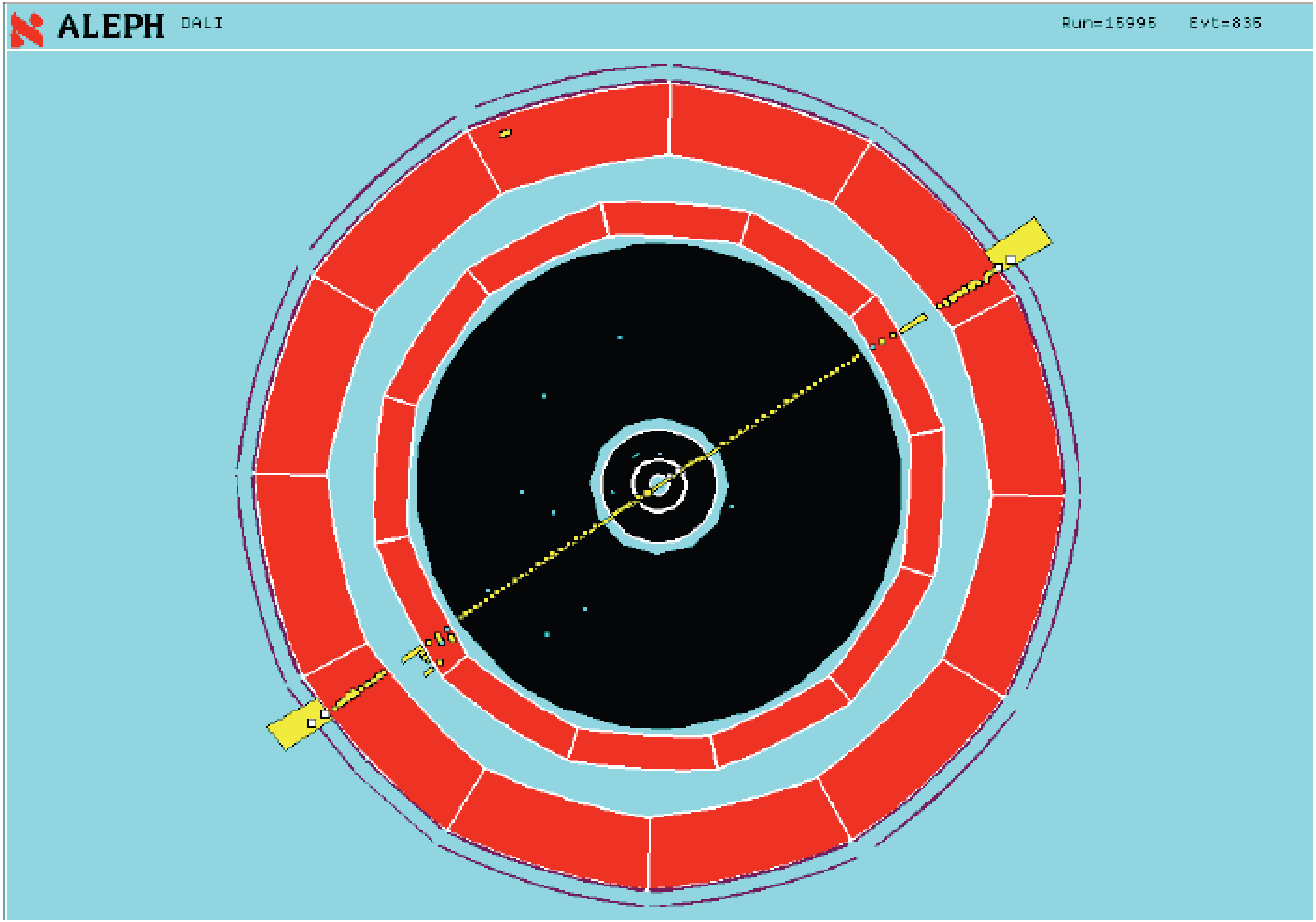
- Branching ratios
- Cross-sections
- Lifetimes

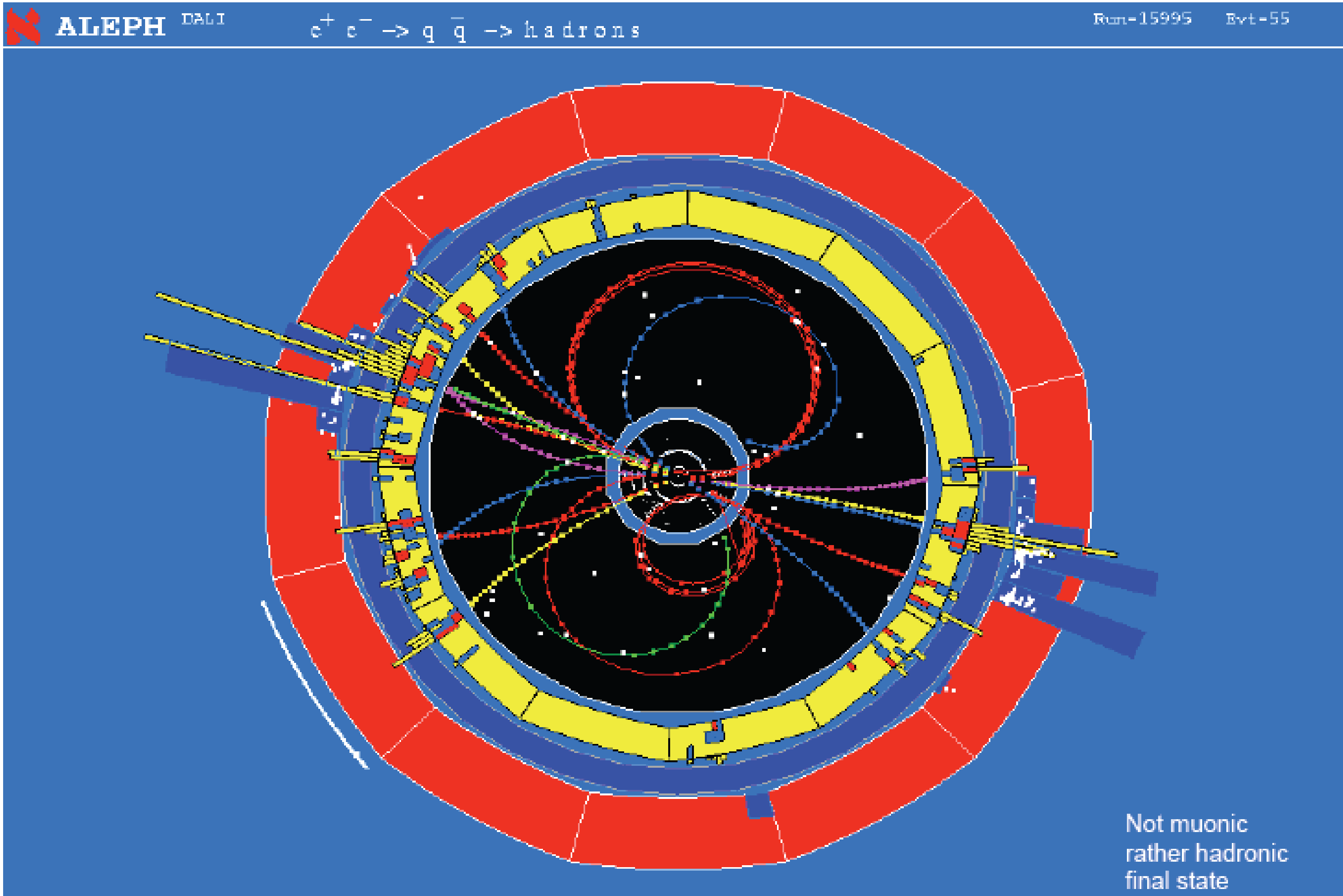
Some Example Events as seen in the Detector


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

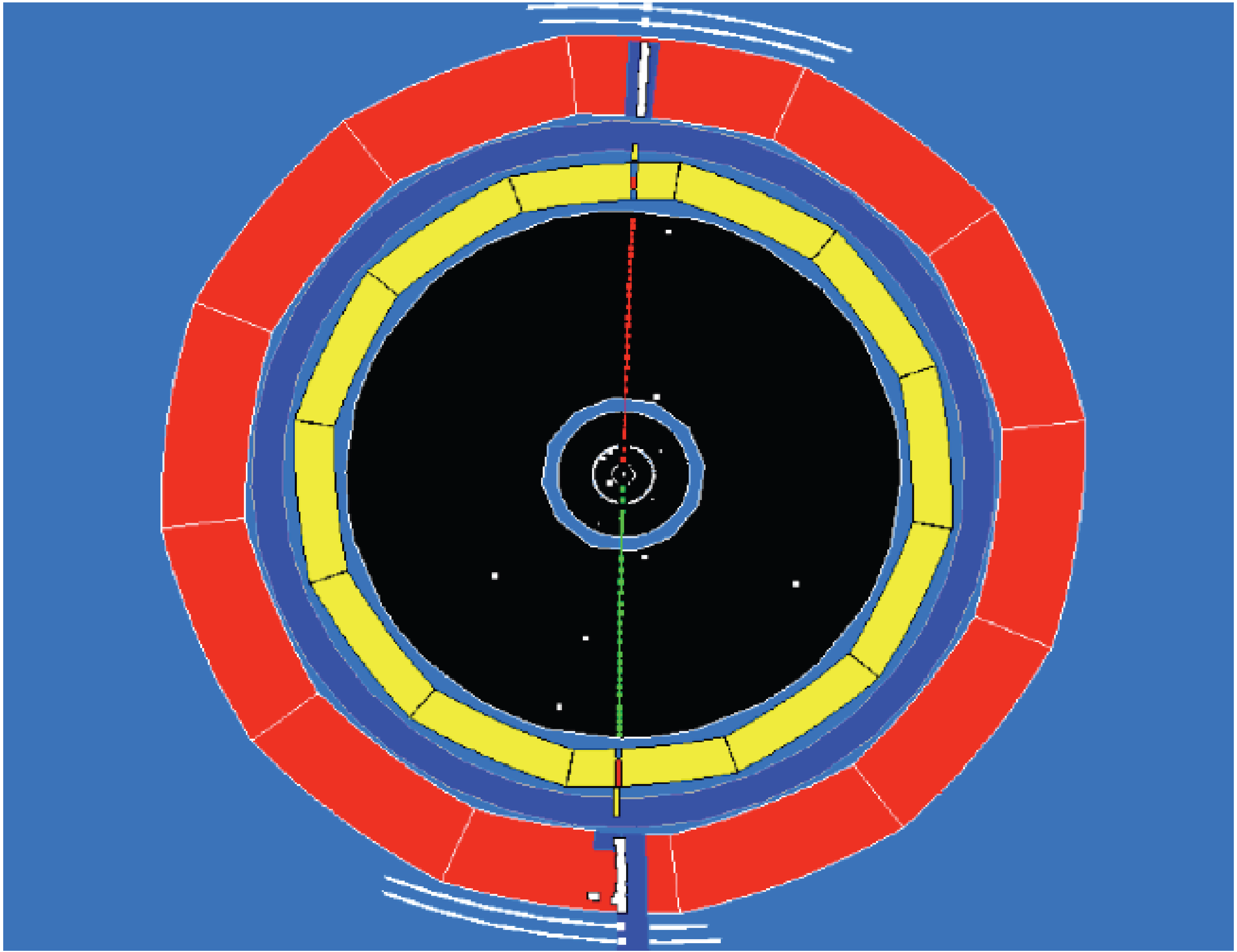
$$Z \rightarrow \mu^{+}\mu^{-}$$





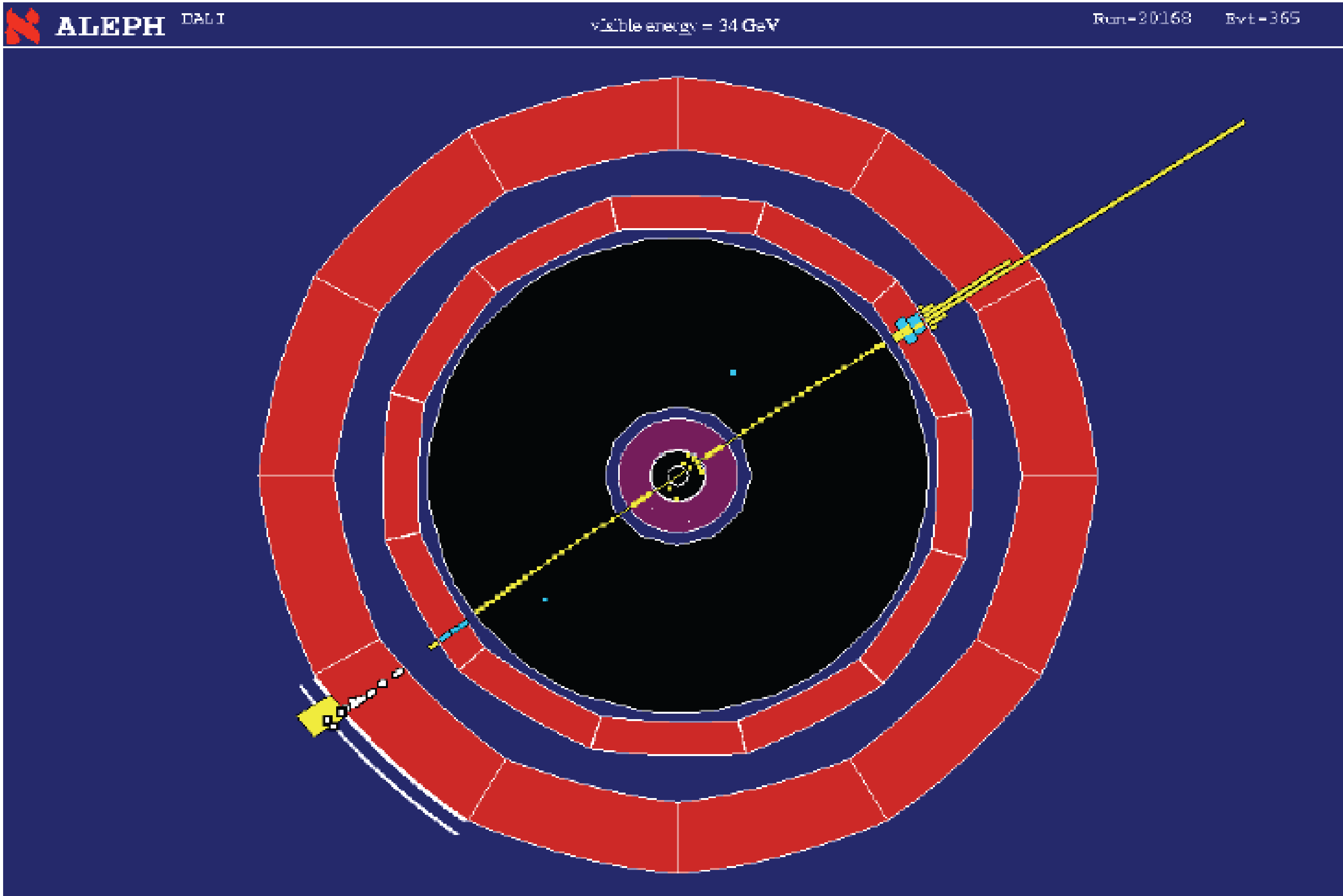
Not muonic
rather hadronic
final state

$$Z \to \mu^{+}\mu^{-}$$

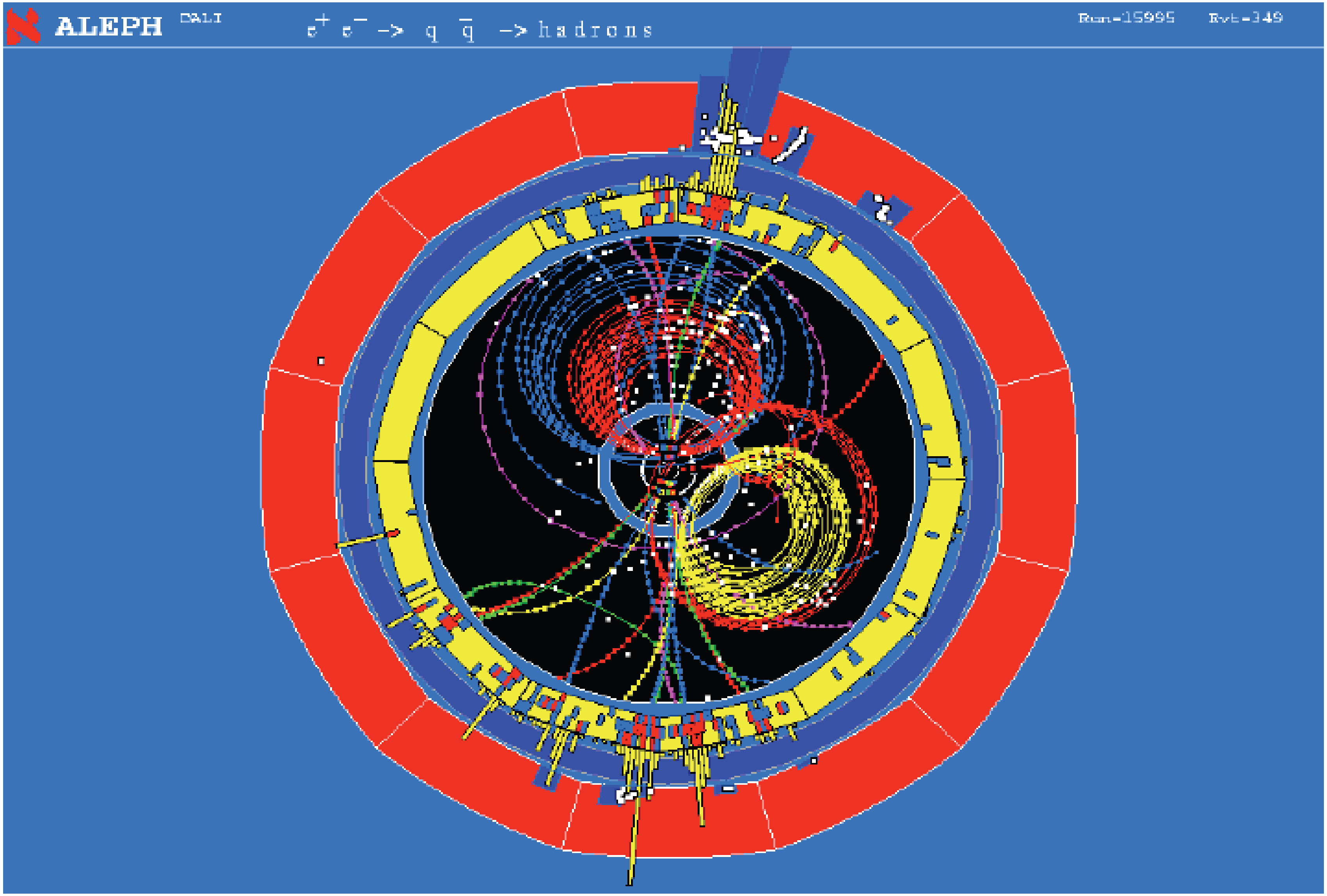


not $Z \to \mu^+\mu^-$

rather Z decay to $\tau^+\tau^-$,
one tau decayed to electron + 2 neutrinos
the other tau decayed to muon + 2 neutrinos



Not muonic rather hadronic final state



And so on....

Why do we need Statistics?

General picture

1 **Physical phenomena**

Described by a theory

2 Sampling a reality

Experiment

3 **Data sample**

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$$

4 **Data analysis**

5 **Results**

- parameter estimates
- confidence limits
- hypothesis tests

Described by PDFs, depending on p **uknown** parameters with **true values**

$$\boldsymbol{\theta}^{true} = (\theta_1^{true}, \theta_2^{true}, \ldots, \theta_p^{true})$$

For example:

$$\boldsymbol{\theta}^{true} = (m_H^{true}, \Delta m_s^{true}, \ldots, \sigma_{tot}^{true})$$

For example:

$$\boldsymbol{x} = (event_1, \ldots, event_N)$$

In statistics x is a multivariate **random variable** *(each event has many properties, all potential variables)*

Event Selection

- Iterate over the event list and select possible candidates
- Take all possible combinations of the daughter particles and perform a constarined vertex fit
- Iterate over the list and calculate the invariant mass of the parent particle using the four momentum of daughters

$$m^2 = E^2 - p^2 = \sum_i \varepsilon_i^2 - \sum_i p_i^2$$

  – for i number of decay products, each having momentum p_i and energy ε_i

  – if calculated mass is within the defined mass limits → save the event

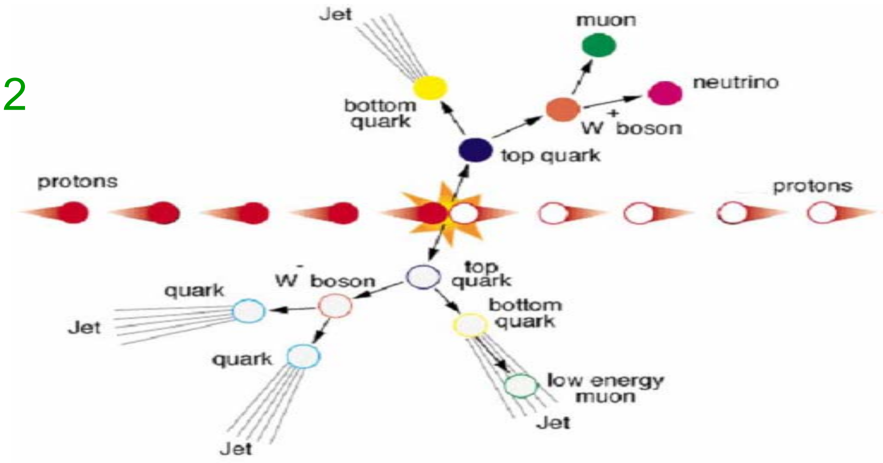
  – if calculated mass is beyond the limits → do not save

- At the end of event list you will have a list of events having possible combinations or candidates for the selected decay channel

  – All the candidates may (or should) not be the correct ones

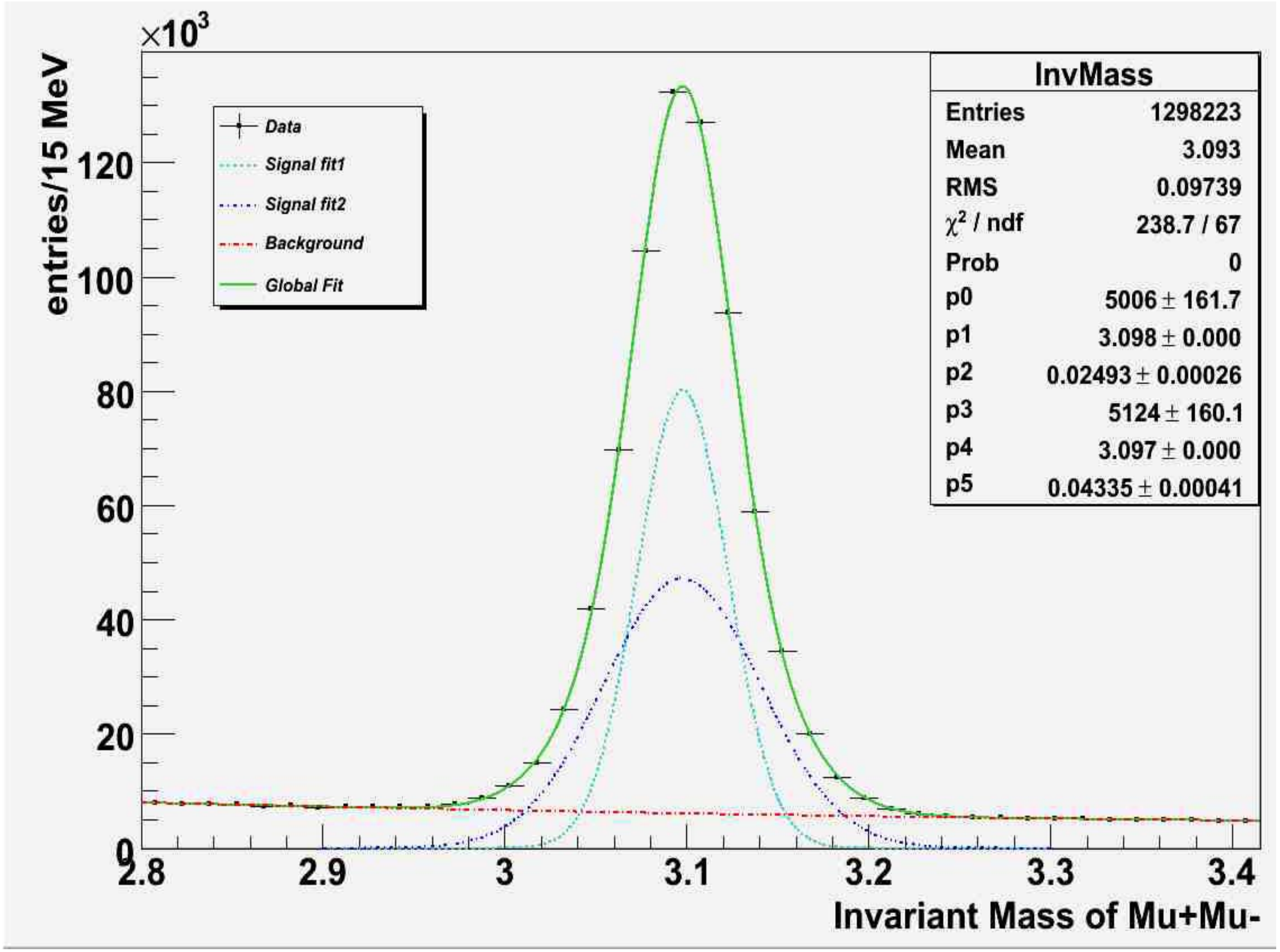
Event Selection : $\bar{t}t$ decay

$$\bar{t}t \rightarrow \bar{b}bW^{+}W^{-} \rightarrow \bar{b}bl\nu_{l}q_{1}q_{2}$$



- Look for events with at least 4 jets
- Look for at least one lepton in the event
- Run b-tagging algorithms to make sure there are at least 2 b-jets
- Form possible combinations; 2 possibilities

Signal with Fit



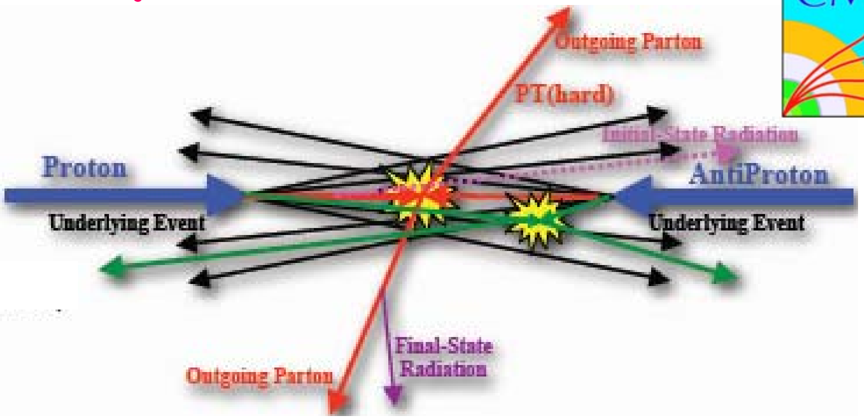
Invariant mass of $\mu^-\mu^+$

Background

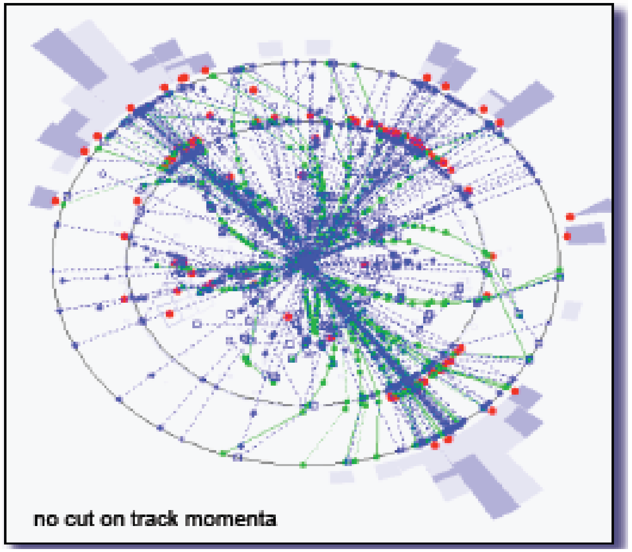
- Any other process giving SAME FINALSTATE as our required channel
- For our case where final state contains 4 jets and one muon; the neutrino is not detected
  - so any channel which gives 4 jets and sometimes a fake lepton can be our background
- Such backgrounds can be generated from direct WW, WZ, ZZ production
- Other sources
  - combinatorics
  - Pile up events
  - Underlying events
- To reduce or get rid of background we need to do a detailed study of various cuts

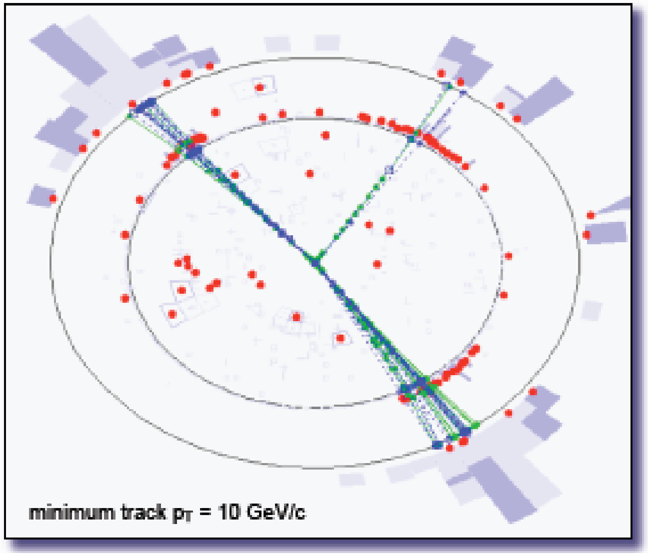
Background for Jets

- Pile up : many additional soft proton-proton collisions
  - **up to 20 at highest LHC luminosity**



- Underlying event:
  - **another event which happens at the same time and appears as a part of first event (require info. from timing detectors to separate the two)**
  - **beam-beam remnants, initial state radiation, multiple parton interactions**
- All this additional energy has nothing to do with jet energies
  - **needs to be subtracted**



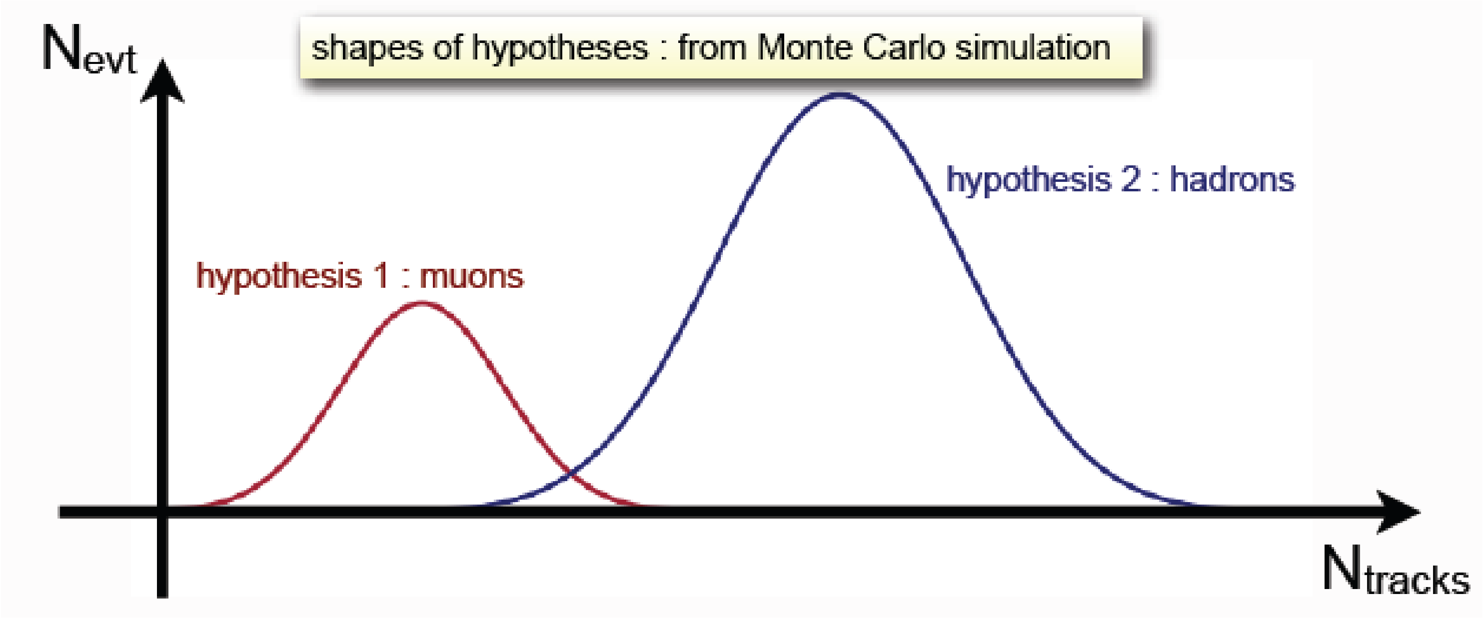


Monte Carlo Studies

- Select cuts for optimization
- Incorporate backgrounds
  - To reduce or get rid of background we need to do a detailed study of various cuts
- Determine efficiency, acceptance etc.
- Study detector resolution
- In some cases define some fit parameters or their limits

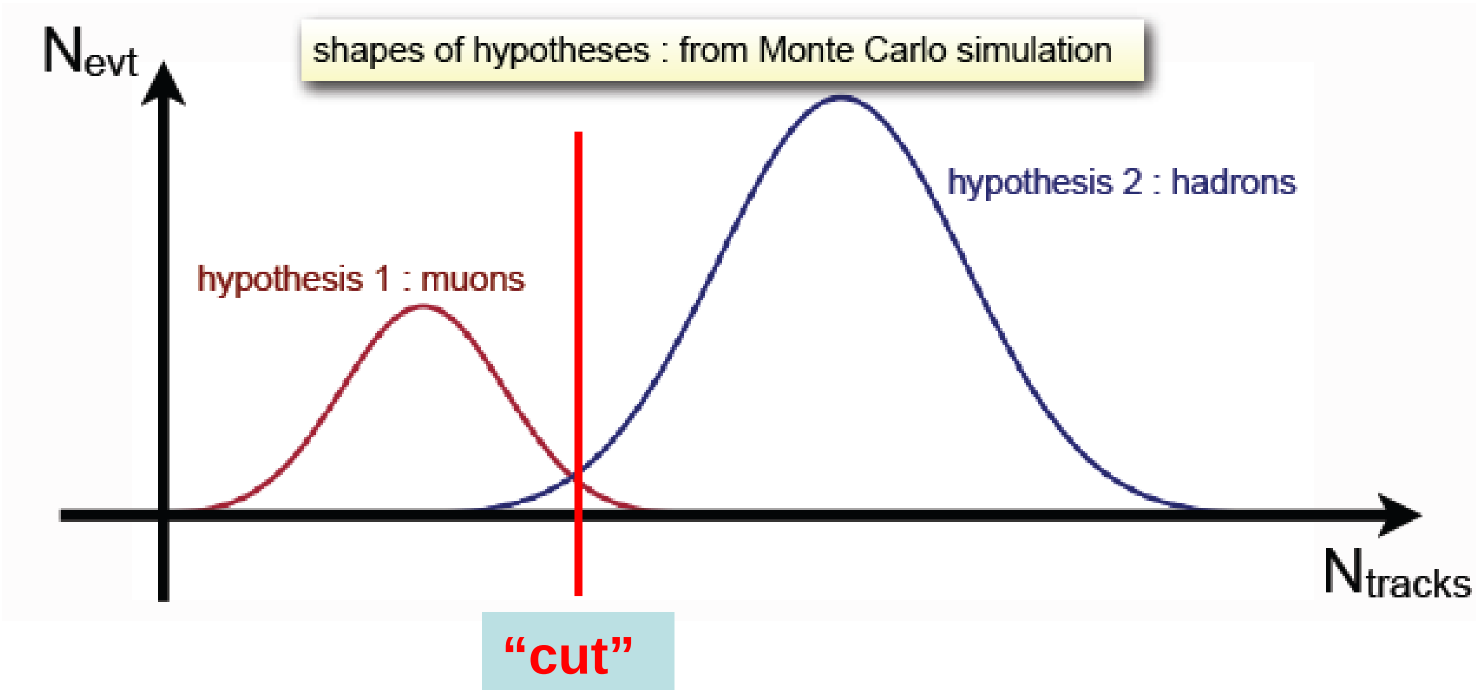
Selection Cuts

For Example: $Z^0 \rightarrow \mu^+\mu^-$



Selection Cuts

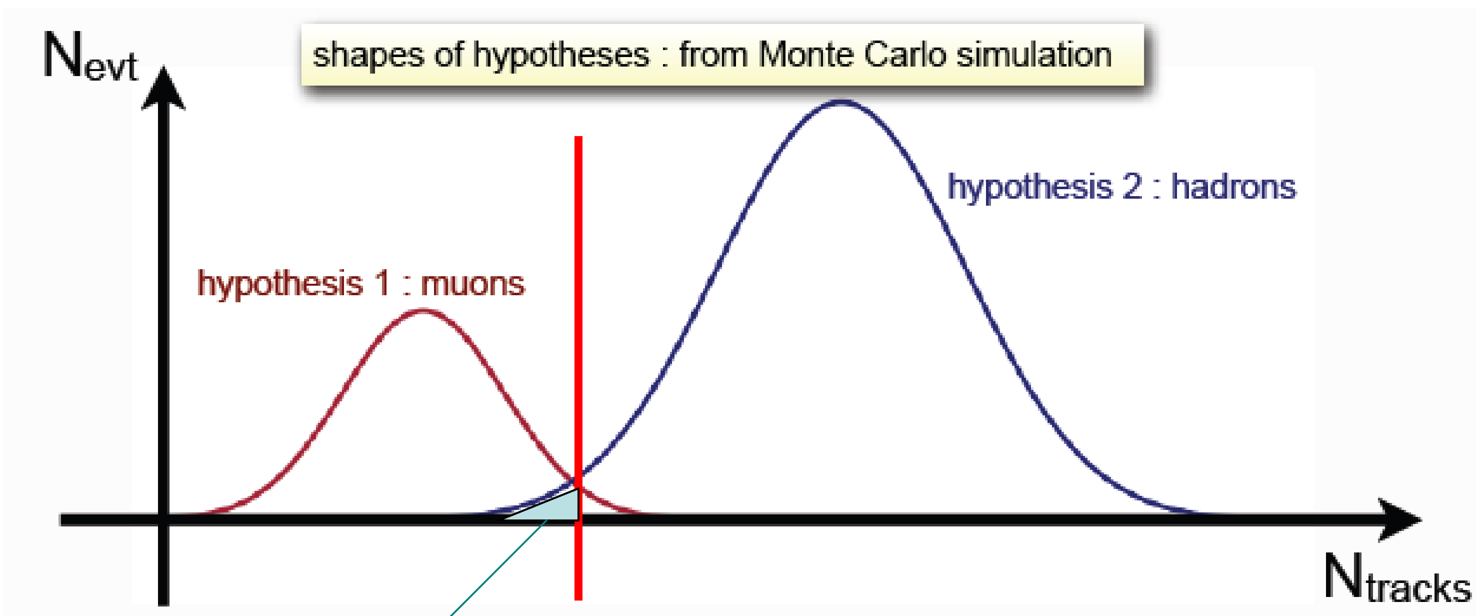
For Example: $Z^0 \rightarrow \mu^+ \mu^-$



$$\text{efficiency for muon selection} = \frac{N_{evt}\,(\text{muons})\ \text{with}\ N_{tracks} < N_{cut}}{\text{all}\ N_{evt}\,(\text{muons})}$$

Selection Cuts

For Example: $Z^0 \rightarrow \mu^+ \mu^-$



"cut"

background to be subtracted

$$\text{efficiency for muon selection} = \frac{N_{evt}\text{ (muons) with } N_{tracks} < N_{cut}}{\text{all } N_{evt}\text{ (muons)}}$$

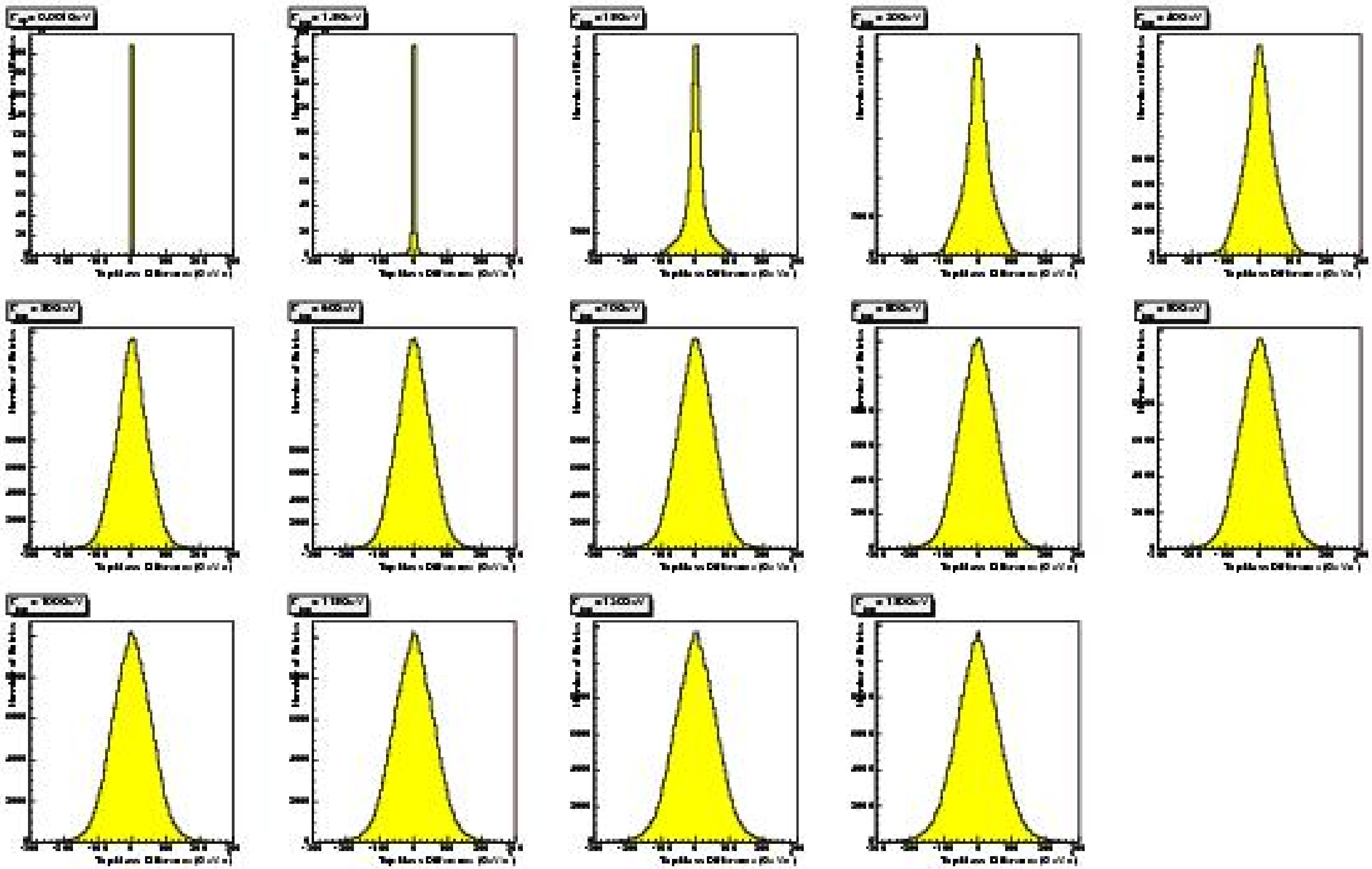
Selection Cuts : Optimization

- The event is selected with no or minimum cuts e.g. may be a vey few track quality cuts or PID cuts
- Check the effect of different variables on signal and background distributions e.g. momentum, event shape variables etc.
- Start with the lowest value of a variable and gradually increase the limit to see the effect of the cut
- Study all possible cuts one by one and select the ones which maximize the signal to background ratio i.e.

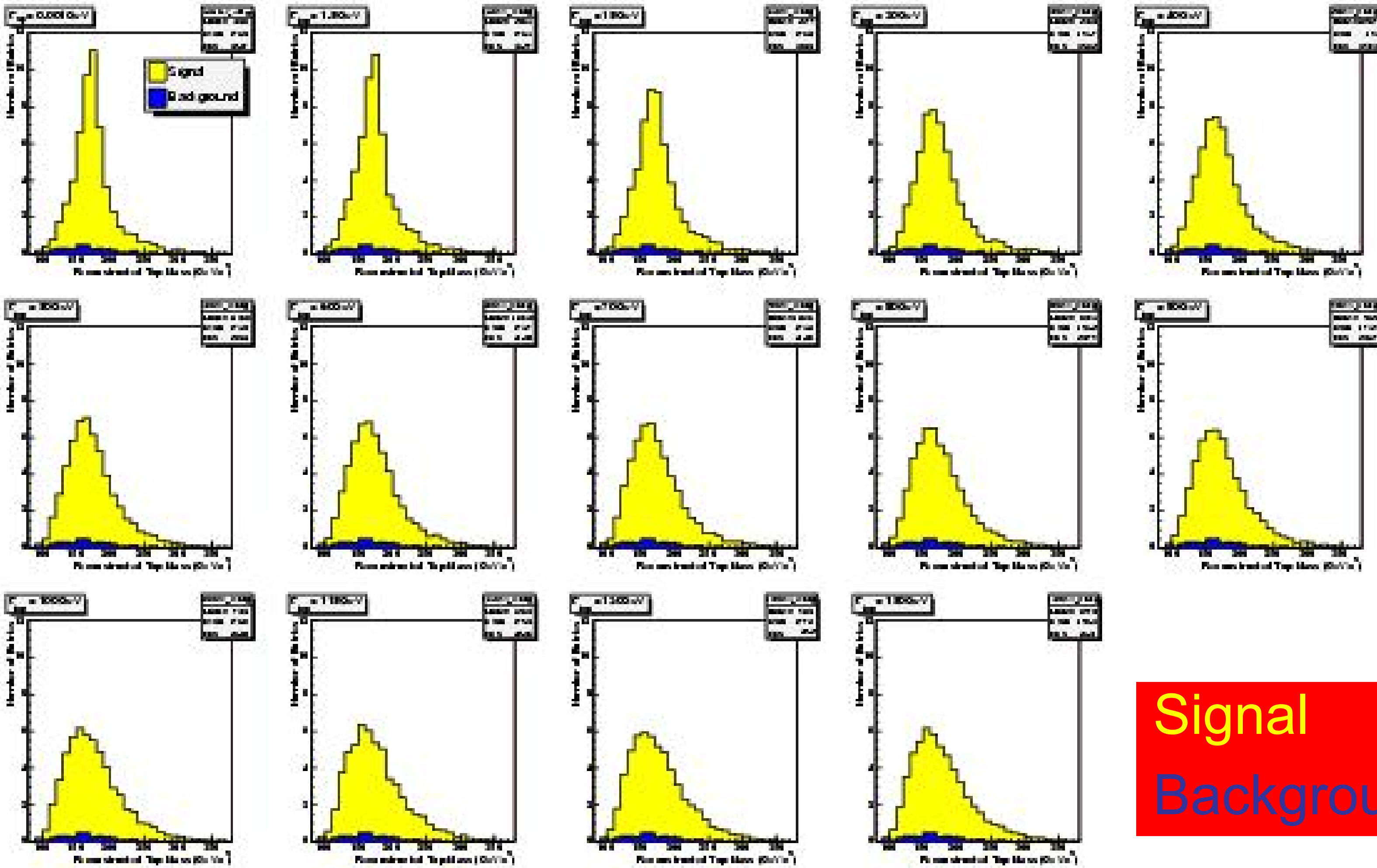
$$\frac{s^2}{s+B}$$

– where s is the signal yield of selected candidates and B is the background count

Distribution of a Signal Variable



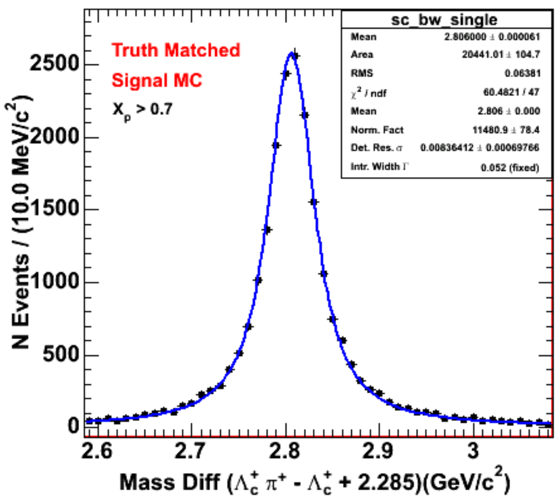
Effect of cuts on Signal and Background



Signal

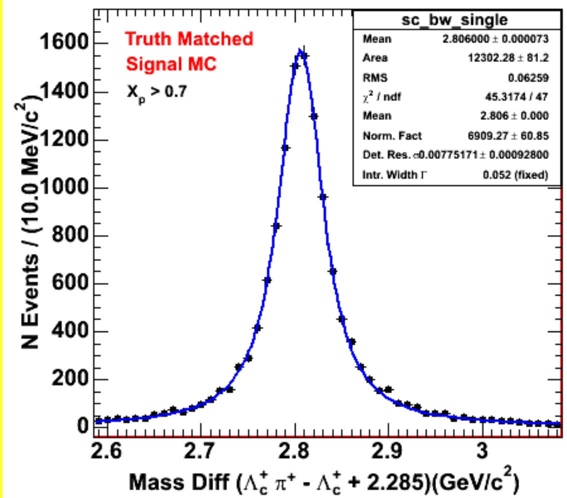
Background

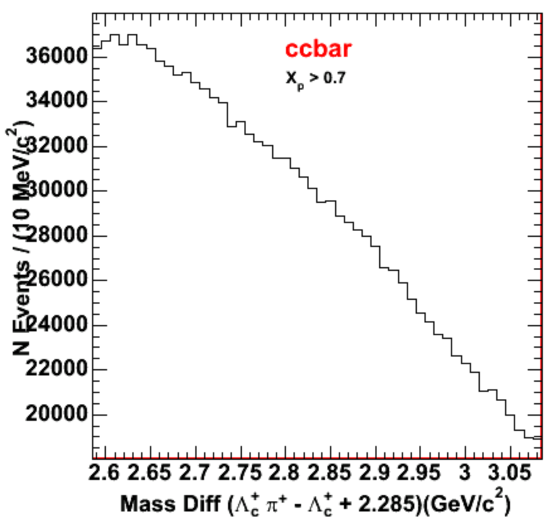
Effect of cuts on Signal and Background-II



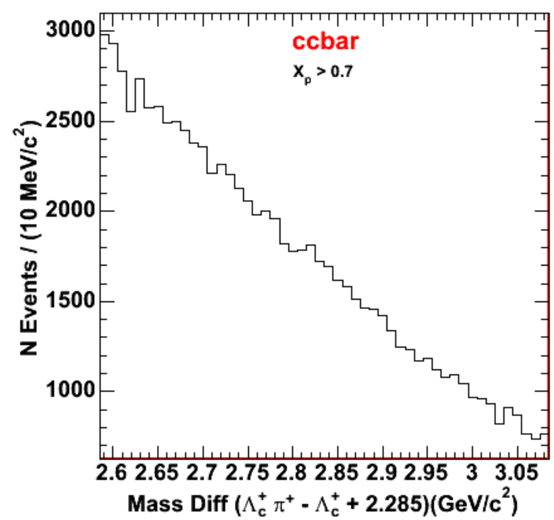
Signal and Background

BEFORE





AFTER



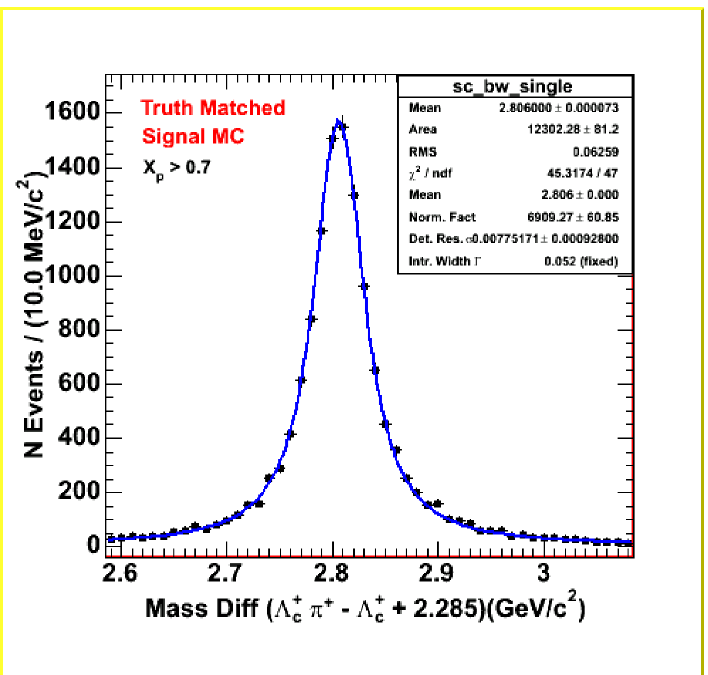
Other Backgrounds

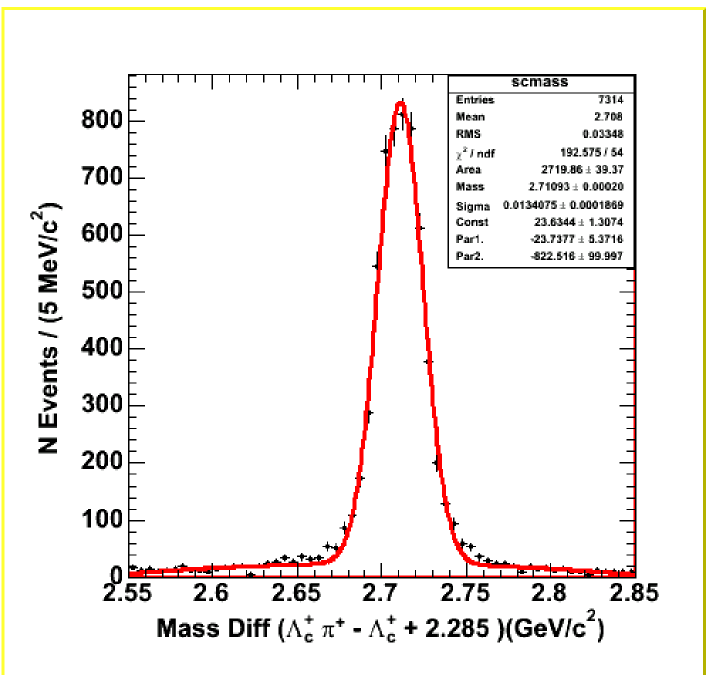
- Backgrounds due to other decays with same FINAL STATE products need to be incorporated
- Study their contribution to your signal and effect on your measurement through Monte Carlo

$$\Sigma_c^{++}(2800) \rightarrow \Lambda_c^{+}(2285)\ \pi^{+}$$

$$\Lambda_c^{+}(2880) \rightarrow \Sigma_c^{0}(2455)\ \pi^{+}$$

$$\Sigma_c^{0}(2455) \rightarrow \Lambda_c^{+}(2285)\ \pi^{-}$$



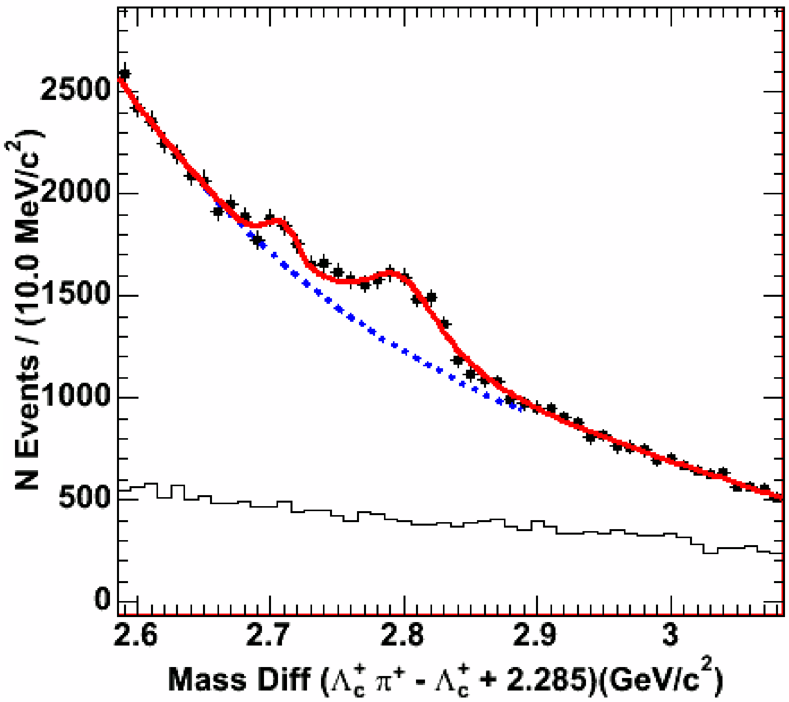


Background (Reflection) in Real Data

$\Sigma_c^{++}(2800) \rightarrow \Lambda_c^{+}(2285)\ \pi^{+}$

$\Lambda_c^{+}(2880) \rightarrow \Sigma_c^{0}(2455)\ \pi^{+}$

$\Sigma_c^{0}(2455) \rightarrow \Lambda_c^{+}(2285)\ \pi^{-}$



Detector Parameters

- Acceptance : The number of reconstructed particles (signal area) divided by the total number of generated particles

$$\text{accepatance} = \frac{\text{reconstructed}}{\text{generated}}$$

- Efficiency : The number of particles which passed through the selection algorithms and cuts divided by the total number of generated particles

$$\text{detector efficiency} = \frac{\text{selected}}{\text{generated}}$$

- Check the loss of efficiency due to each selection cut
- Can estimate the number of expected signal area for a given luminosity
  – allows comparison with the theory and real data

Detector Resolution

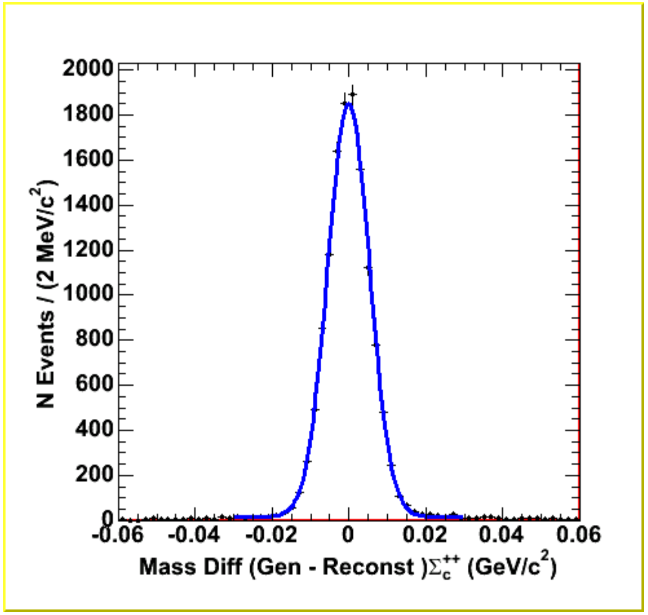
- Detector is not perfect
- Can have error in measurements
  - – loss of energy or charge can reduce the actual signal
  - – contribution from nearby elements can increase the signal
- Monte Carlo : Plot the Generated – Reconstructed mass difference
  - – Fitting a single gaussian gives the resolution
- The resolution should be less than the natural decay width

Gaussian Fit parameters:

$$\text{Area} = 12478 \pm 116$$

$$\text{Mean} = -0.00011 \pm 0.00005 \ \text{GeV/c}^2$$

$$\sigma = 0.00543 \pm 0.00004 \ \text{GeV}$$



Real Data

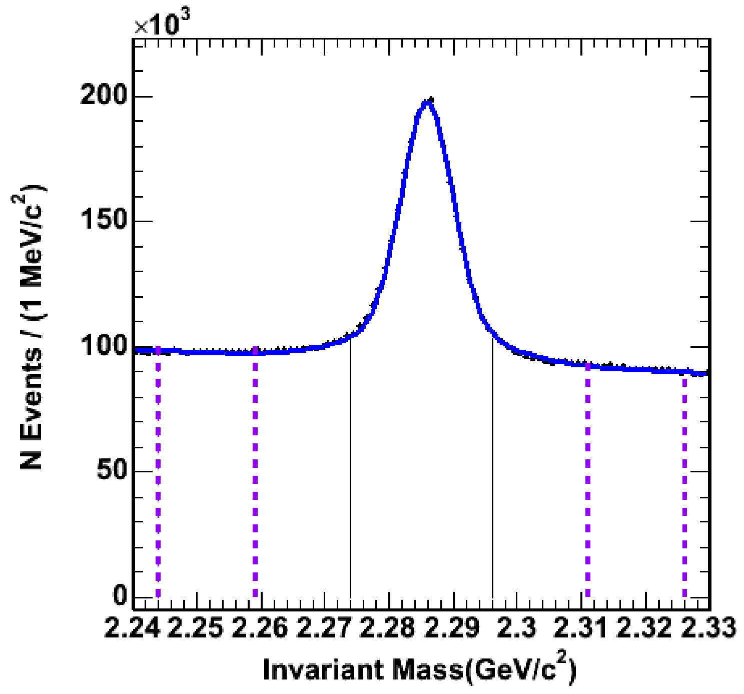
- Select the data sample to be used
- Run the same .cc code on the data files and save the output root files having all the possible candidates and their event info.
- The size (luminosity fb^{-1}) of MC should be at least twice of the real data
- Fit the histogram
- Compare your fitted values to the theoretical / expected ones
- Measure your selected variable using appropriate method

Sample Plot from Data

- For example

$$\Lambda_c^+ \rightarrow p k^- \pi^+$$

– 231fb^{-1} from BABAR



Double-Gaussian fit parameters :

$$\text{Signal Yield} = 1270100 \pm 5122$$

$$\text{Signal Mean} = 2.28588 \pm 0.00001\ \text{GeV/c}^2$$

$$\sigma_1 = 0.00397 \pm 0.00002\ \text{GeV}$$

$$\sigma_2 = 0.01052 \pm 0.00026\ \text{GeV}$$

$$\text{fraction of } \sigma_1 = 0.7$$

Goodness of Fit

- We are now interested in this kind of questions
  - Is the fit good or not?
  - How **significant** is discrepancy between data and obtained functional form?
  - How well does the vector of measurements in the histogram $\boldsymbol{n} = (n_1, \ldots, n_k)$ compare with predicted values $\boldsymbol{\nu} = E[\boldsymbol{n}] = (\nu_1, \ldots, \nu_k)$?
- These questions can be answered with a **goodnes-of-fit test**
  - Which is itself a part of a so called HYPOTHESIS TESTING

- So called **NULL hypothesis** H_0 is:

  The functional form (or predicted values) describes well our data!

- The form (i.e. the parameters that form depends on) is found by one of the methods for parameter estimation (moments, ML, chi-square)
- We are now looking for a **statistic** t (usually a single number) whose value reflects an agreement between the data and the hypothesis
  - The most commonly used statistic is the $\chi^2_{\min}$

χ^2 Distribution

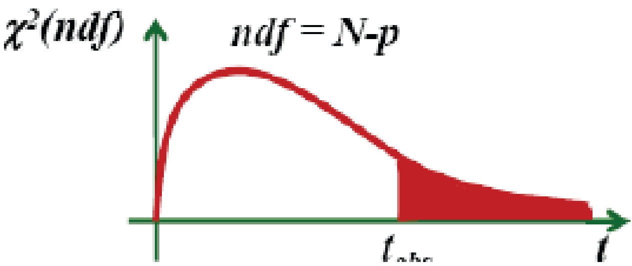
- The most commonly used statistics for Goodness of Fit test is distributed as χ^2 distribution

  – when fitting histograms with N bins, with a function depending on p parameters, then the obtained in the fit is distributed according to the (N-p) functions,

  ★ (N-p) is called the number of degrees of freedom (ndf)

- Suppose we have
  - A set of precisely known values $\boldsymbol{x} = (x_1,\ldots,x_N)$
    - *For example histograms bins*
  - At each x_i
    - a measured value y_i
      - *For example number of events in the given histogram bin*
    - corresponding error on measured value σ_i
    - predicted value of measurement that depends on parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$ we want to estimate: $F(x_i;\boldsymbol{\theta})$
  - Suppose that measurements are independent
- To find best estimate of θ we minimize the suitably weighted summ of squared differences between measured and predicted values → so called "**least squares**" or "**chi-square**"

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - F(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2}$$



Uncertainties

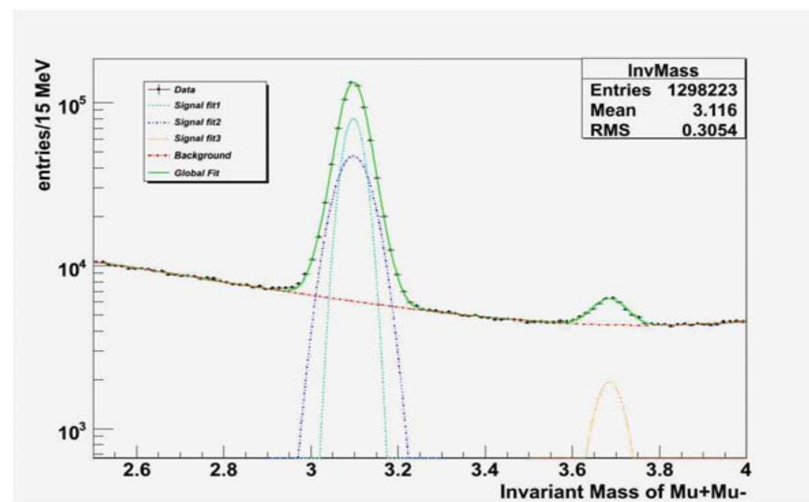
- Statistical error
  - these go like the square-root of number of events

$$N \pm \sqrt{N}$$

  - to reduce these you need to record lots (millions) of events in the detector and process them
- Systematic error
  - if you see less number of events due to
    - ★ event selection (cuts)
    - ★ detector imperfections
    - ★ poor understanding etc.

Statistical Errors

- Statistical error is deviation of a measurement from the mean of the sample
- Should not be confused with the background
- Generally an error of less that 10% on physics measurement is an acceptable value
- e.g. in the plot
  - Signal to background ratio
    - ★ Large for left signal
    - ★ Small for signal on right
  - But statistical error
    - ★ Less than 10% for both



Systematic Errors

- Reproducible inaccuracy introduced by faulty equipment, calibration, or technique
- Experiment related : due to hardware / detector used
  – e.g. detector resolution, luminosity
- Analysis related : software or measurement technique
  – e.g. fitting technique, generated mass, error in some theoretical parameter
- All independent systematic errors are added in quadrature to give one final value

$$\sigma^2 = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}$$

Comparison with Theory

- An experimental measurement is given by

$$x \pm x_{stat} \pm x_{syst}$$

- The theoretical models may predict the value to be x_1, x_2 or x_3 depending on the model used
- If the predicted values lie within the measured error limits then the theory and experiment are comparable

Physics Groups

Physics Coordinator
P.Sphicas
Deputy: J.Incandela

- **e/γ** — D. Futyan, P.Vanlaer
- **Muons** — N.Amapane, N.Neumeister
- **Jets/MissET** — G.Dissertori, N.Varelas
- **b-tagging** — T.Speer, I.Tomalin
- **ParticleFlow/τ** — R.Cavanaugh, P.Janot
- **Onl Selection** — S.Dasu, C.Leonidopoulos
- **Diffraction** — M.Groethe
- **Heavy Ions** — D.D'Enterria, B.Wyslouch
- **Super LHC** — D.Denegri
- **QCD** — K.Rabbertz
- **EWK** — R.Tenchini
- **Top** — J.D'Hondt, J.Mnich
- **B physics** — U.Langenegger
- **Higgs** — A.Nikitenko, Y.Sirois
- **SUSY-BSM** — S.Eno, M.Spiropulu
- **MC generators** — P.Bartalini, F.Moortgat

Publishing Results

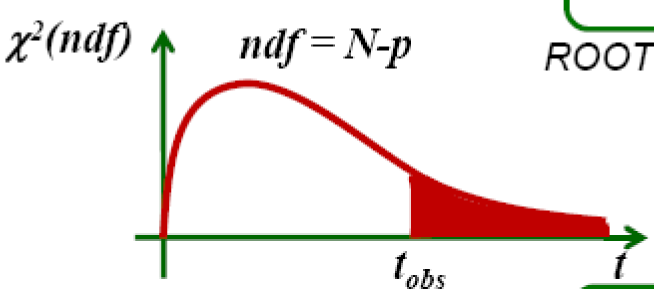
- Present your work in front of the conveners and other members of your physics group
  - should be done at every important step
    - ★discuss the problems and issues and make changes / corrections as required
- Write a CMS Internal Note
  - ask for a paper committee
    - ★ any members of collaboration from any institution can be the members
- Committee will review the whole analysis
  - may suggest changes or some new things to be included
    - ★ prepare the paper draft
- Collaboration wide review
  - send for publishing

THAND YOU!!!!

Questions?
Comments…

Fitting Overview

1	**Data sample**	$(y_1, y_2, \ldots, y_N)$		$\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$	*Event counts in bins*
2	**Parameter estim.** (ML, chi-square)	$(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)$		$\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_N)$	*Predicted counts in bins*
3	**test statistic** t	t_{obs}	3.1	$\chi^2_{\min} = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i^2}$	*t in case θ estimated by χ^2 method*
				or	
4	p **- value**	$p = \int_{t_{obs}}^{\infty} g(t \mid H_0)\, dt$	3.2	$-2\ln\lambda \approx -2\ln\frac{L(\boldsymbol{n};\boldsymbol{\nu})}{L(\boldsymbol{n};\boldsymbol{n})}$	*t in case θ estimated by ML method (Pearson's χ^2 can also be used as test statistic)*
		ROOT `TMath::Prob`			
5	**Decision about** H_0	If ($p < 1-\alpha$) reject H_0 at CL α		If p << try with new fit	*Very often use just common sense*

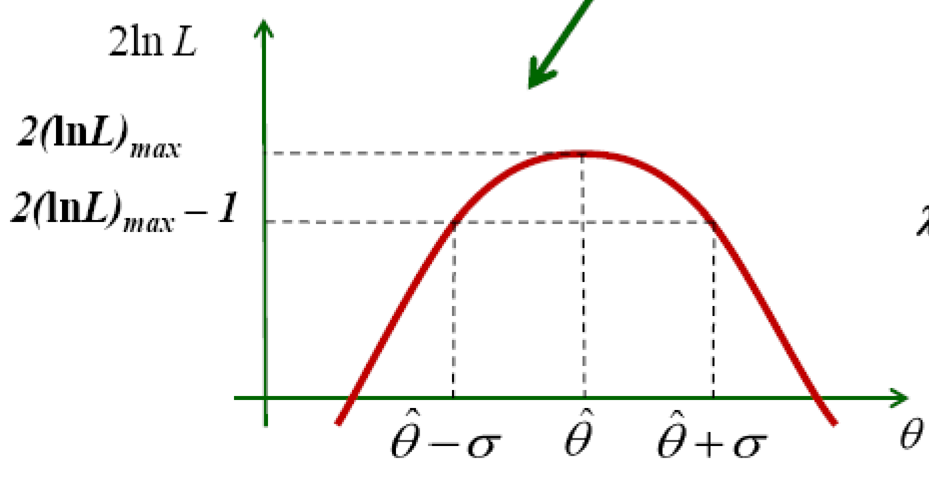


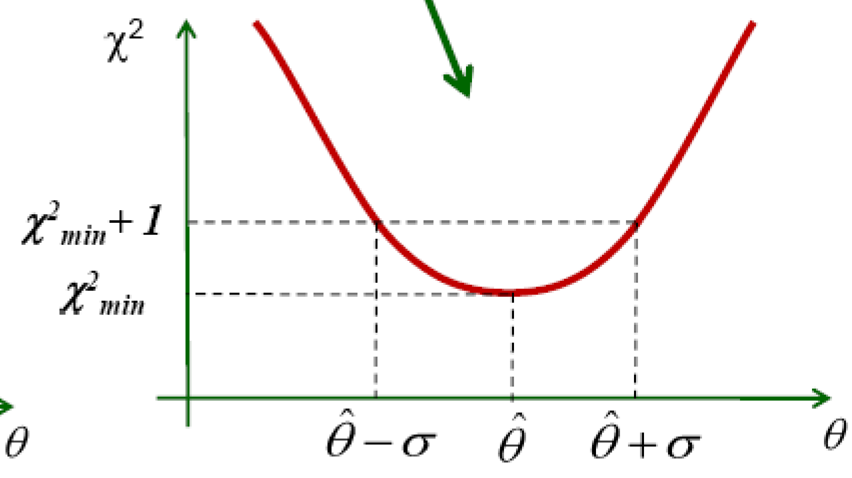
In theory α is predefined (f.g. 95%); in practice p-value is converted to z-value (f.g. significance = 5), see lecture 3

Parameters and Errors

- **The best values** of parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)$ are found by solving p equations
$$\frac{\partial\chi^2(\boldsymbol{\theta})}{\partial\theta_i} = 0, \quad i = 1,\ldots,p$$
- **Errors** (or limits) on parameters are found in the equivalent was as for the ML method
  - Matrix inversion
  - Shape of χ^2 arround it's minimum value

$$\mathrm{Prob}(2\ln L) \geq 2\ln L_{\max} - \Delta_L \quad \Leftrightarrow \quad \mathrm{Prob}(\chi^2) \leq \chi^2_{\min} + \Delta_{\chi^2}$$





Title

- First line
  - sub
    - ★ subsub
- Second line
  - sub
    - ★ subsub
- Fourth line
  - sub
    - ★ subsub
- Fifth line
  - sub
    - ★ subsub

Title

- First line
  - sub
    - ★ subsub
- Second line
  - sub
    - ★ subsub
- Fourth line
  - sub
    - ★ subsub
- Fifth line
  - sub
    - ★ subsub