

Basic Statistical Methods in HEP – Part 1

Andrés Flórez
Universidad de Los Andes

Disclaimer...


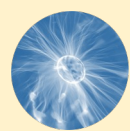
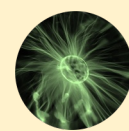
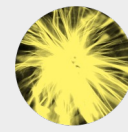

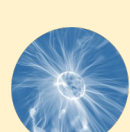
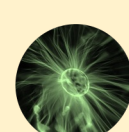
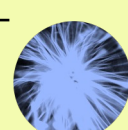


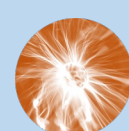






- This lecture is meant to be an introductory class with some fundamental concepts.
- **This subject is very broad and there are several topics I will not cover because of time...**
- **Hopefully, this is a seed for future lectures on these topics that could help many students...**



IT'S BECAUSE HOT AIR RISES. THE SUN'S HOT IN THE MIDDLE OF THE DAY, SO IT RISES HIGH IN THE SKY.

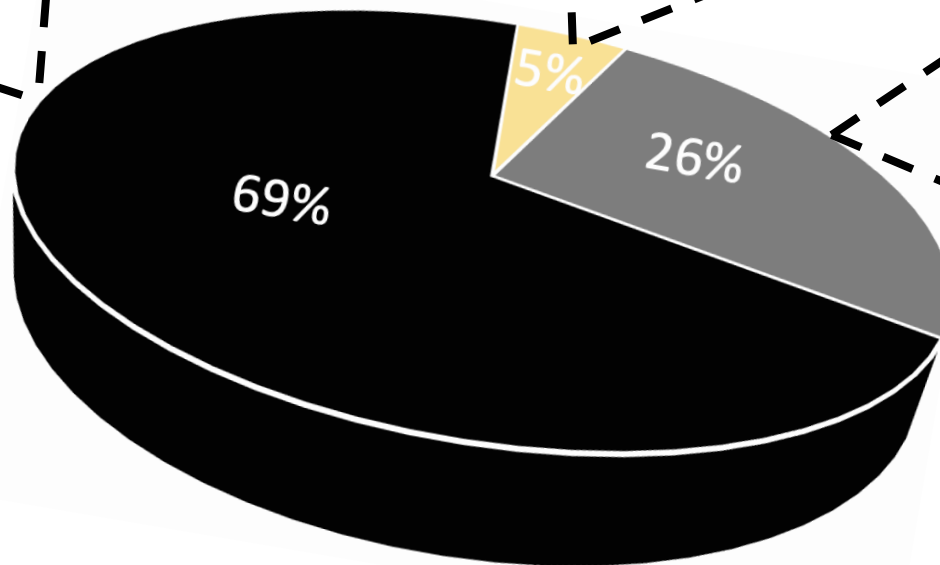
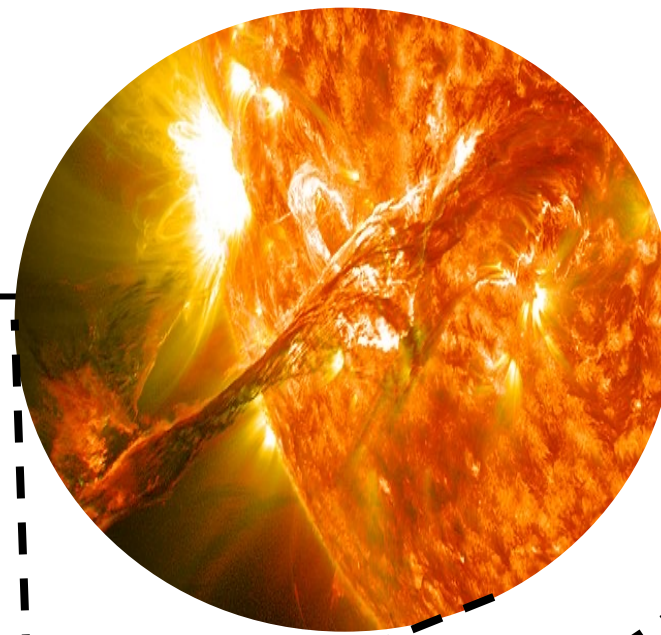
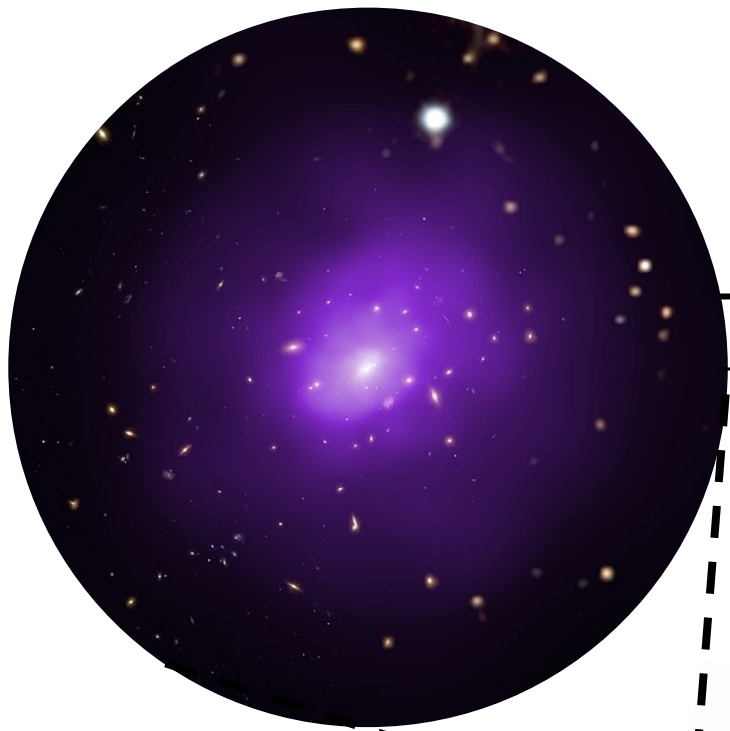


Standard Model

FERMIONS						BOSONS	
Quarks			Leptons				
1 ^{ra}		2 ^{da}		3 ^{ra}			
u $+\frac{2}{3}$ $\frac{1}{2}$		c $+\frac{2}{3}$ $\frac{1}{2}$		t $+\frac{2}{3}$ $\frac{1}{2}$		γ 0	 1
d $-\frac{1}{3}$ $\frac{1}{2}$		s $-\frac{1}{3}$ $\frac{1}{2}$		b $-\frac{1}{3}$ $\frac{1}{2}$		W^- -1	 1
e -1 $\frac{1}{2}$		μ -1 $\frac{1}{2}$		τ -1 $\frac{1}{2}$		W^+ -1	 1
ν_e 0 $\frac{1}{2}$		ν_μ 0 $\frac{1}{2}$		ν_τ 0 $\frac{1}{2}$		Z^0 0	 1
						g 0	 1

HIGGS	
H^0	0
	

- ❖ Fermions compose matter while bosons mediate fundamental interactions.
- ❖ Fermions are divided in different groups (generations), that differ only on the mass of the different types of particles.



12/5/22

Andrés Flórez

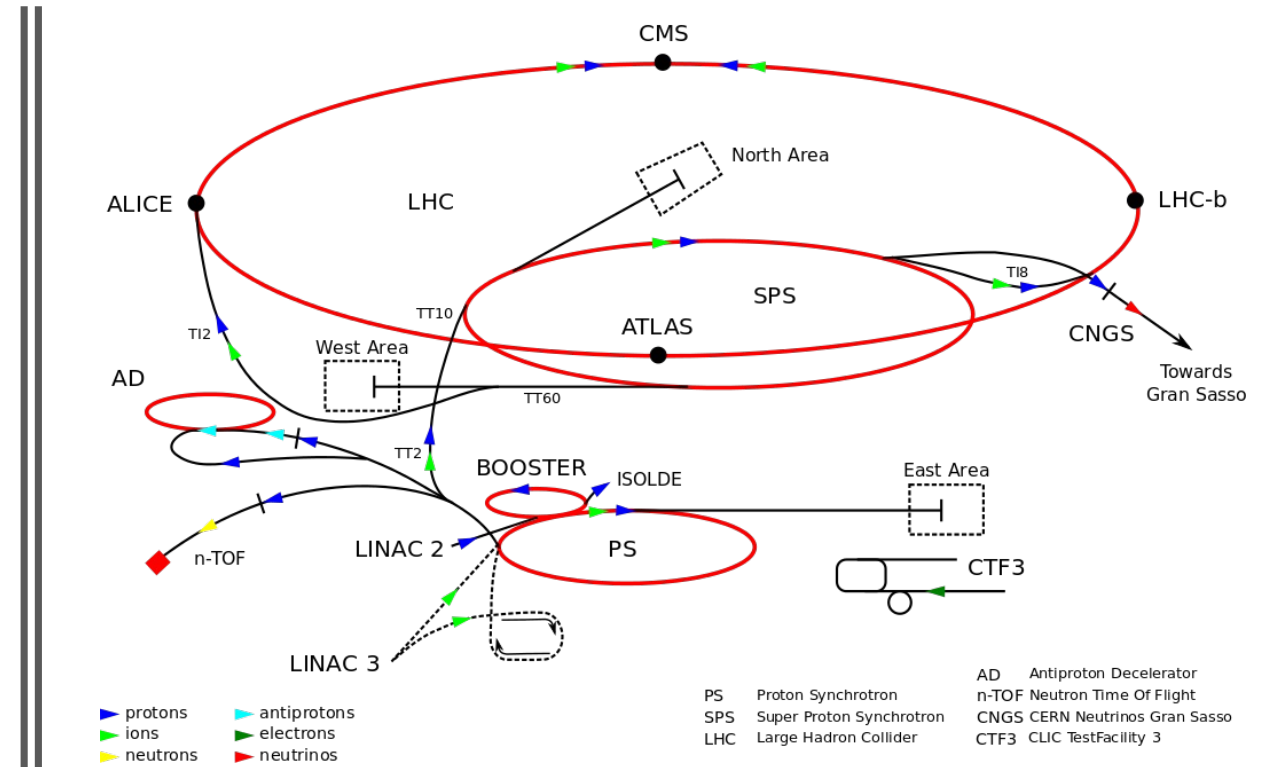
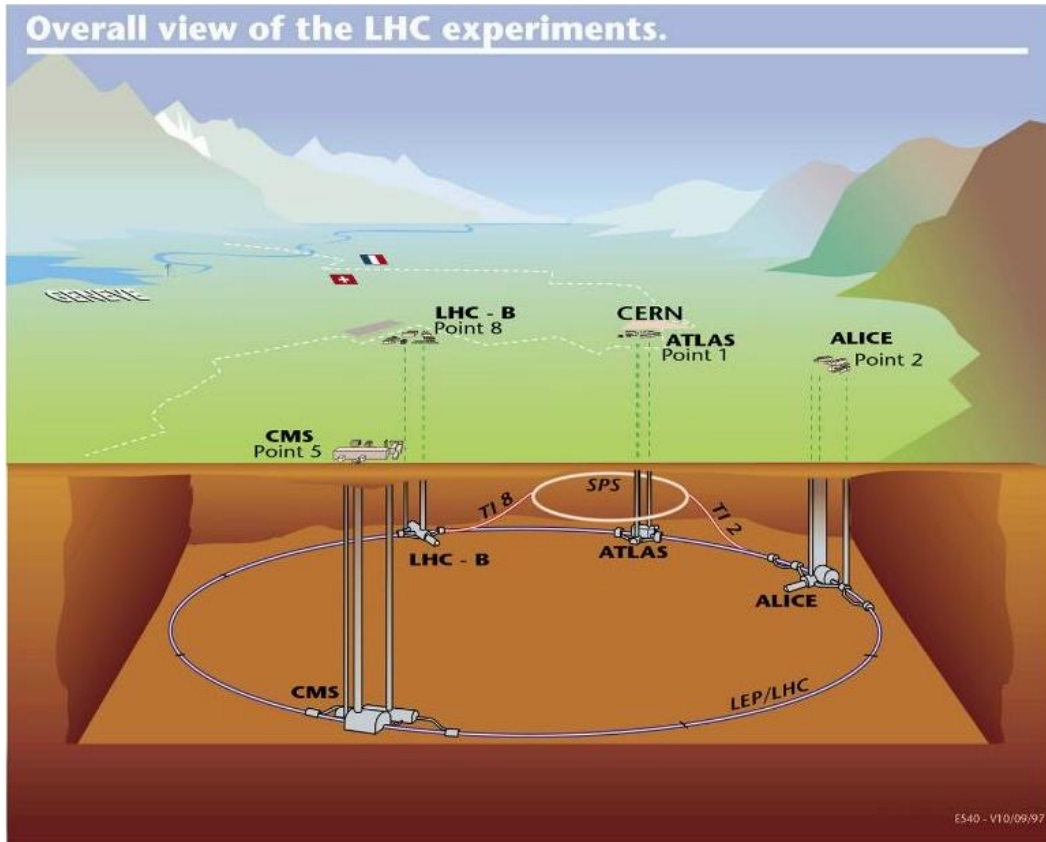


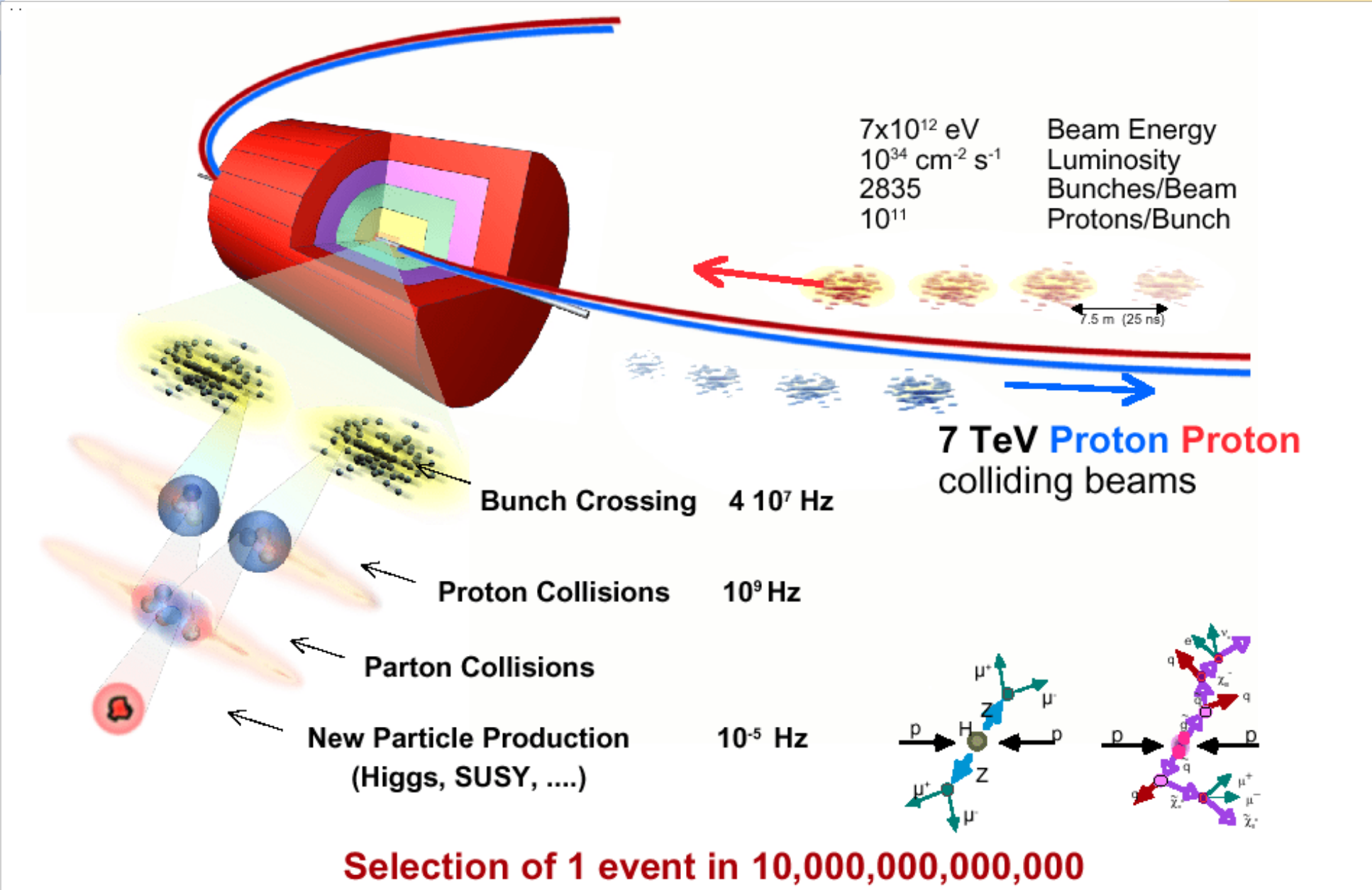
The Large Hadron Collider (LHC)

12/5/22

Andrés Flórez

The LHC





CMS DETECTOR

Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T

STEEL RETURN YOKE
12,500 tonnes

SILICON TRACKERS

Pixel ($100 \times 150 \mu\text{m}$) $\sim 16\text{m}^2 \sim 66\text{M}$ channels
Microstrips ($80 \times 180 \mu\text{m}$) $\sim 200\text{m}^2 \sim 9.6\text{M}$ channels

SUPERCONDUCTING SOLENOID

Niobium titanium coil carrying $\sim 18,000\text{A}$

MUON CHAMBERS

Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers

PRESHOWER

Silicon strips $\sim 16\text{m}^2 \sim 137,000$ channels

FORWARD CALORIMETER

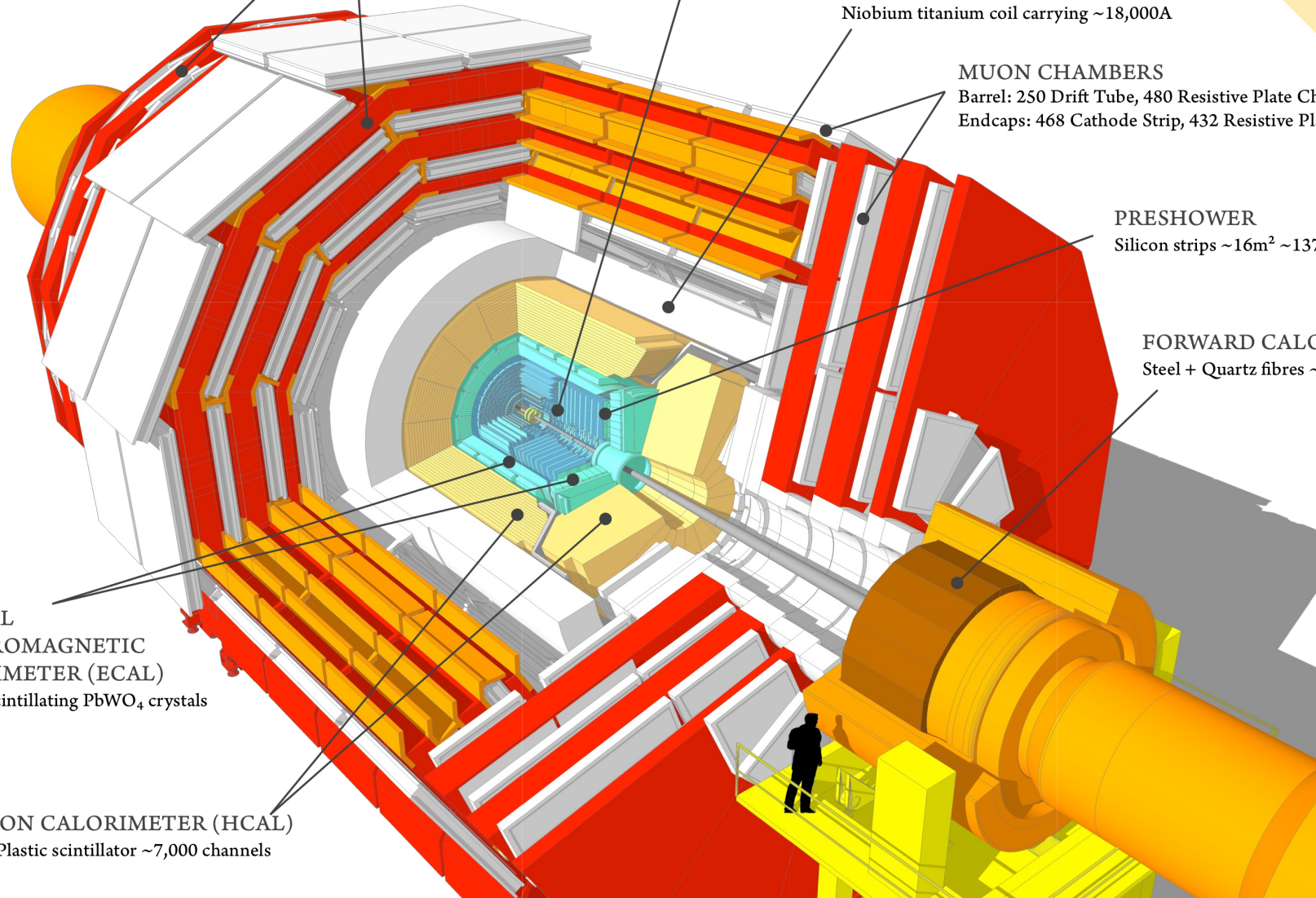
Steel + Quartz fibres $\sim 2,000$ Channels

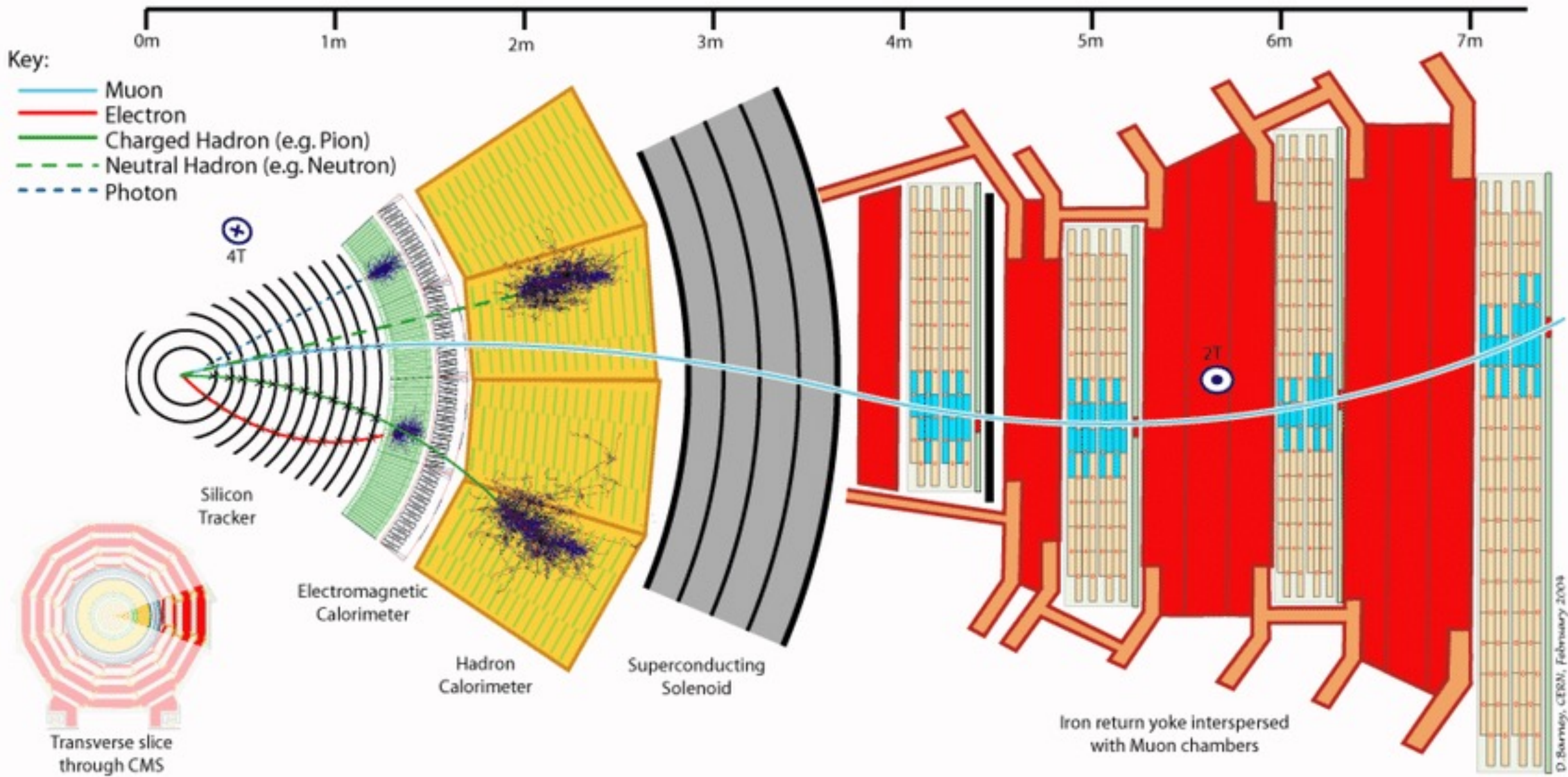
CRYSTAL
ELECTROMAGNETIC
CALORIMETER (ECAL)

$\sim 76,000$ scintillating PbWO_4 crystals

HADRON CALORIMETER (HCAL)

Brass + Plastic scintillator $\sim 7,000$ channels





Observables

X number of leptons and or photons

e

μ

τ

γ

Veto jets (j) and/or b-jets (b) and/or select nj or nb jets

0j

nj

0b

nb

Use some observables based on the topology of the event

E_T^{miss}

H_T

α_T

m_T

.....

Observables

e	μ	τ	ν
g	nj	Ob	ν
E_{T}^{miss}	H_T	\cancel{E}_T	\cancel{m}_T

e	μ	τ	ν
g	nj	Ob	ν
E_{T}^{miss}	H_T	\cancel{E}_T	\cancel{m}_T

Some Fundamental Concepts

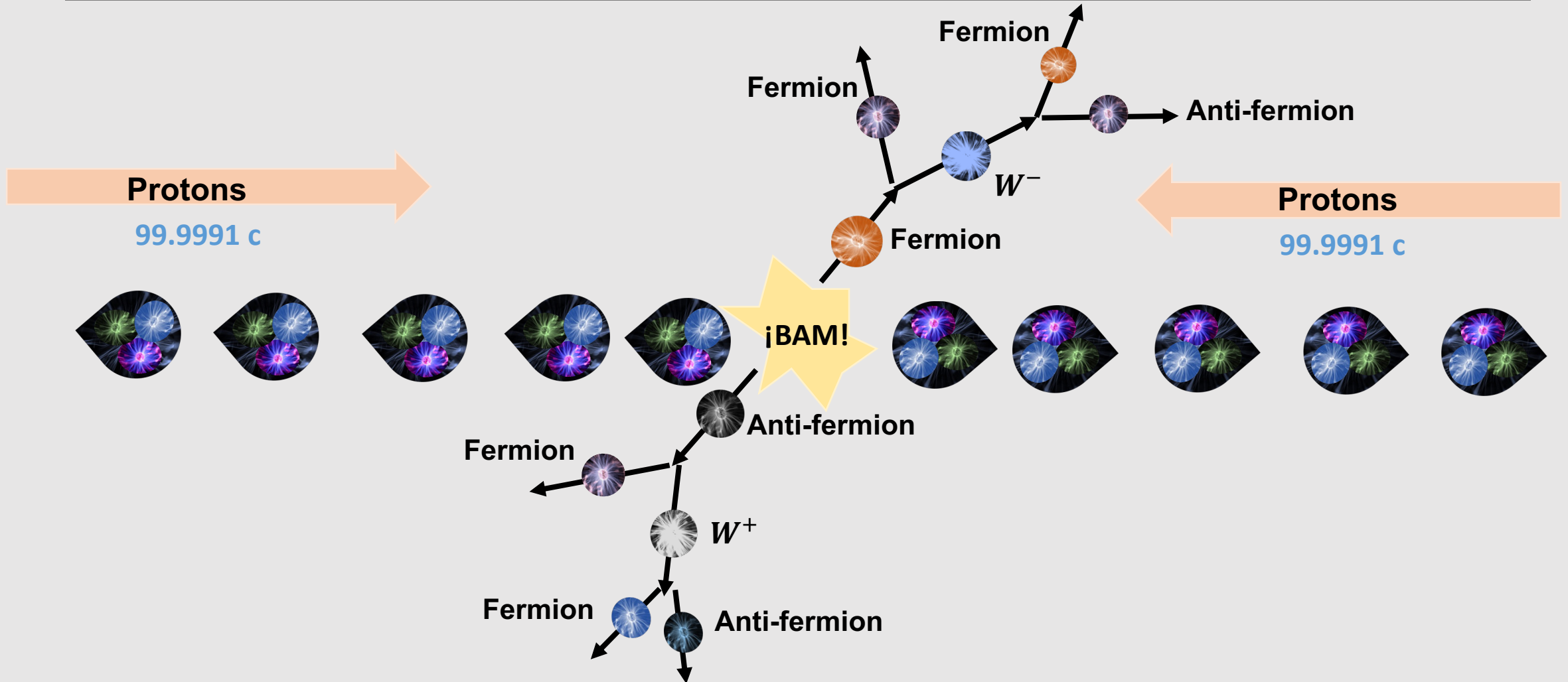
- **Event:**

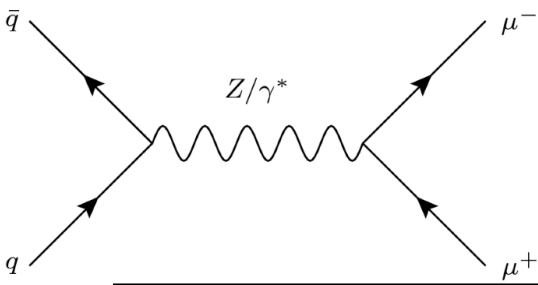
- It is understood as the result or outcome after a fundamental interaction among particles takes place.
- For example, we consider an event as the result after the interaction between two proton beams at the LHC or after the collision of heavy ion beams, or between a beam of particles and a fix target, etc.

- **Cross section:**

- Is the probability of production of a specific process. This quantity is related with the level of the interaction between the beam and the target, or between two beams, and it depends on the energy of collisions.

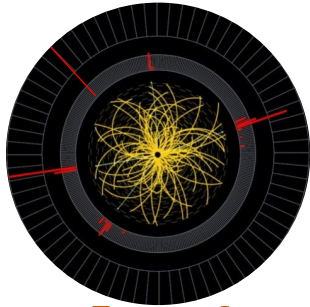
Proton-Proton Collisions



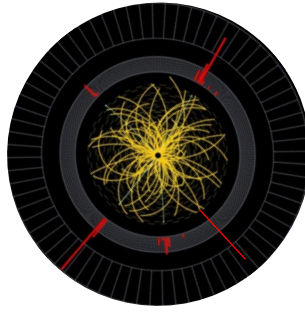


Toy Example

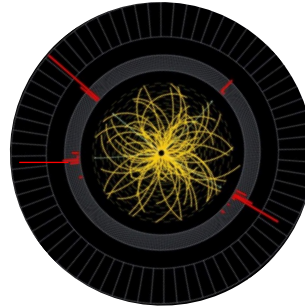
Event 1



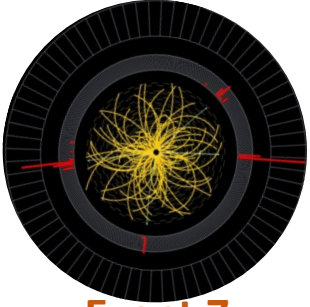
Event 2



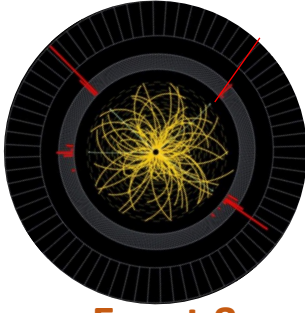
Event 3



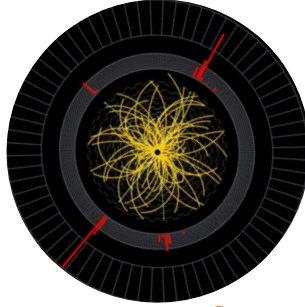
Event 4



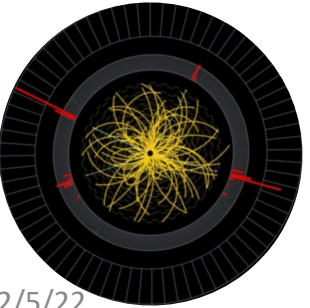
Event 5



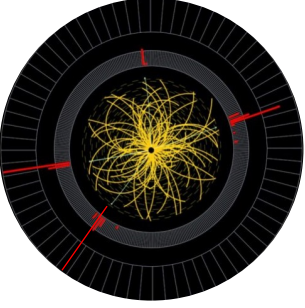
Event 6



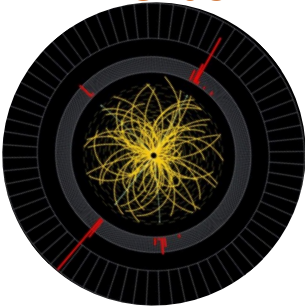
Event 7



Event 8



Event 9

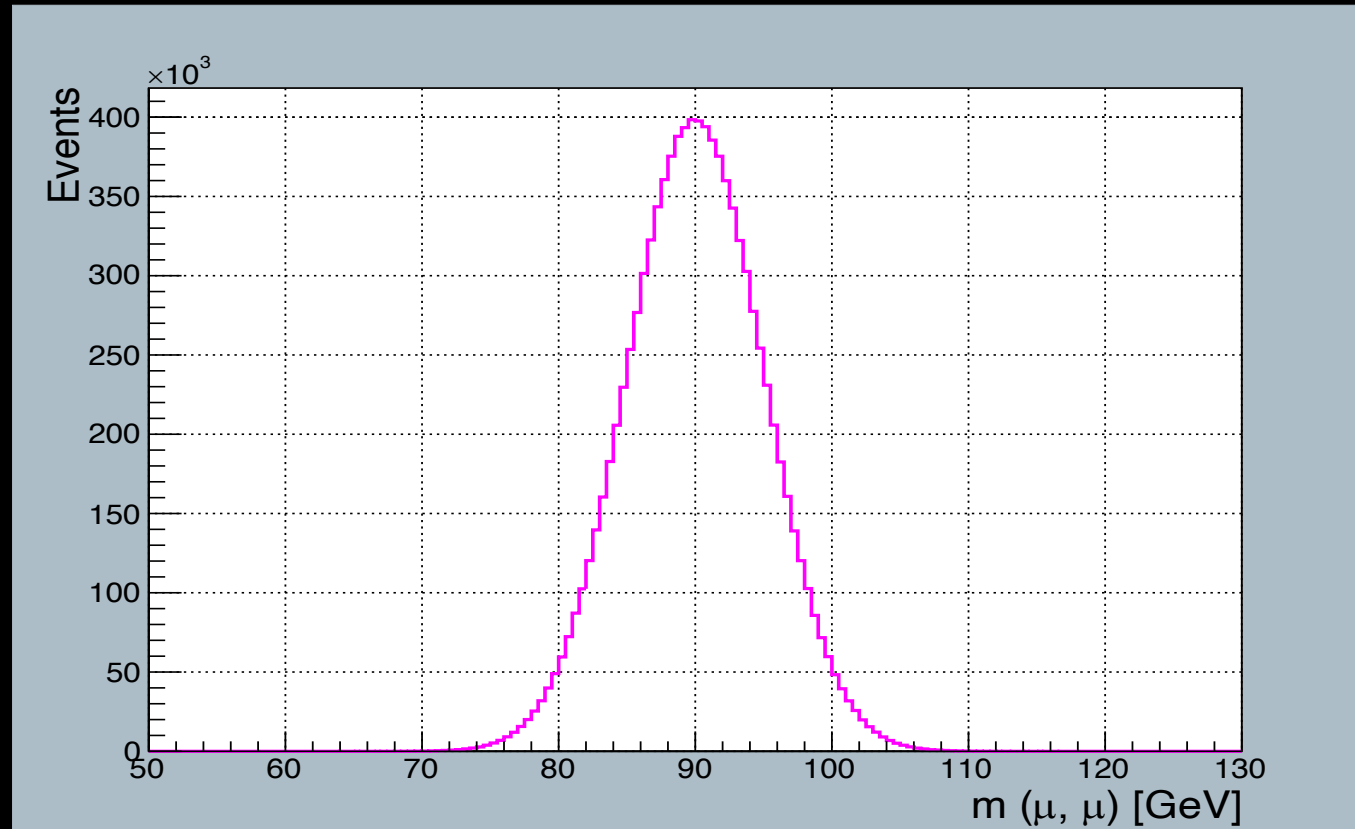


Select events with at least two muons.....

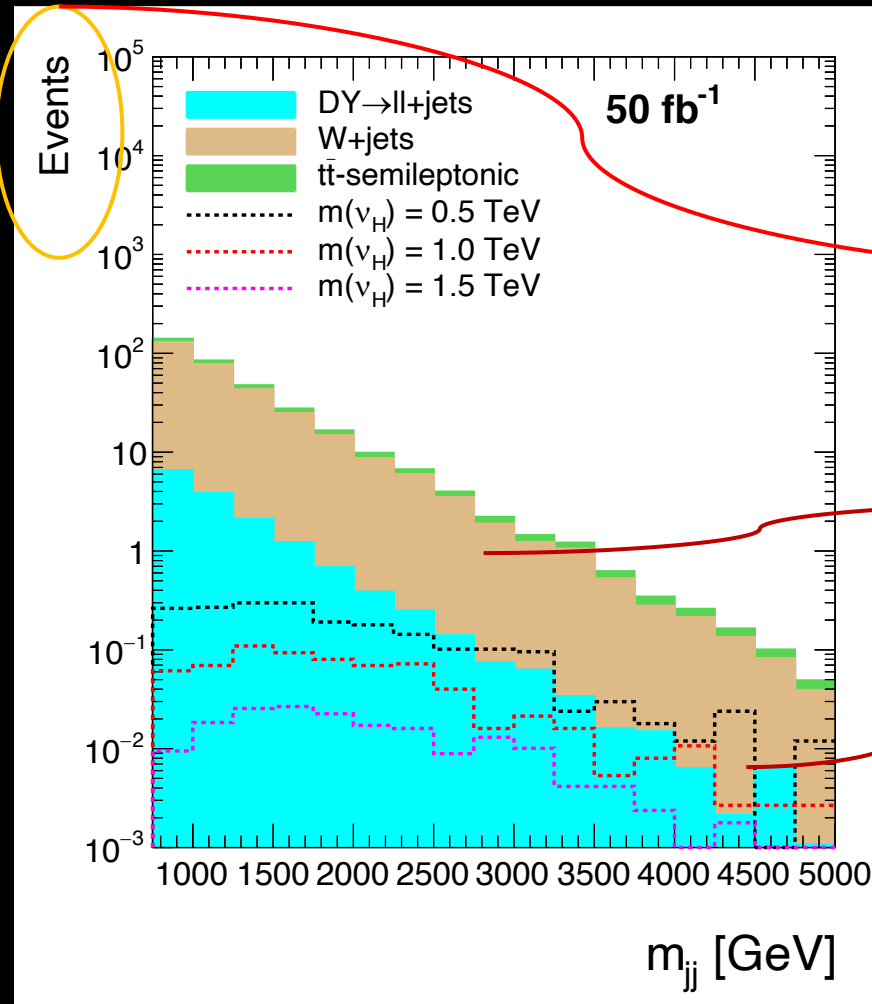
Event	$p_T(\mu)_1$	$\eta(\mu)_1$	$\phi(\mu)_1$	$p_T(\mu)_2$	$\eta(\mu)_2$	$\phi(\mu)_2$
1	20.5	-2.1	3.1	37.3	2.3	-3.2
2	25.2	1.8	2.1	32.6	-1.3	-2.2
3	10.3	0.1	3.3	44.6	2.3	-3.0
4	30.5	0.3	-1.1	39.3	-2.2	1.5
5	40.8	-1.5	2.5	38.0	1.9	-2.0
6	35.0	-1.4	-1.4	45.1	0.0	1.2
7	42.1	2.2	2.8	22.5	-1.3	-2.8
8	44.2	0.7	2.2	35.8	1.9	-1.2
9	27.4	0.8	-1.4	33.3	-2.3	1.2

$Z \rightarrow \mu^+ \mu^-$ Reconstructed mass

$$m(\mu, \mu) = \sqrt{(E_1^\mu + E_2^\mu)^2 - (\vec{p}_1^\mu + \vec{p}_2^\mu)^2}$$



Example: MC



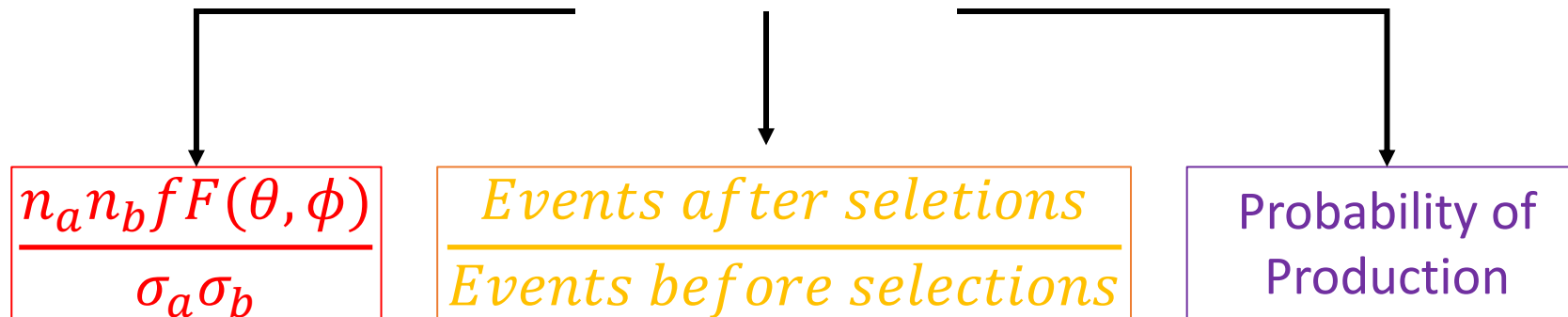
$$N_i = L \times \epsilon_i \times \sigma_i$$

Expected background processes from the SM

Expected number of signal events

Expected number of events in MC

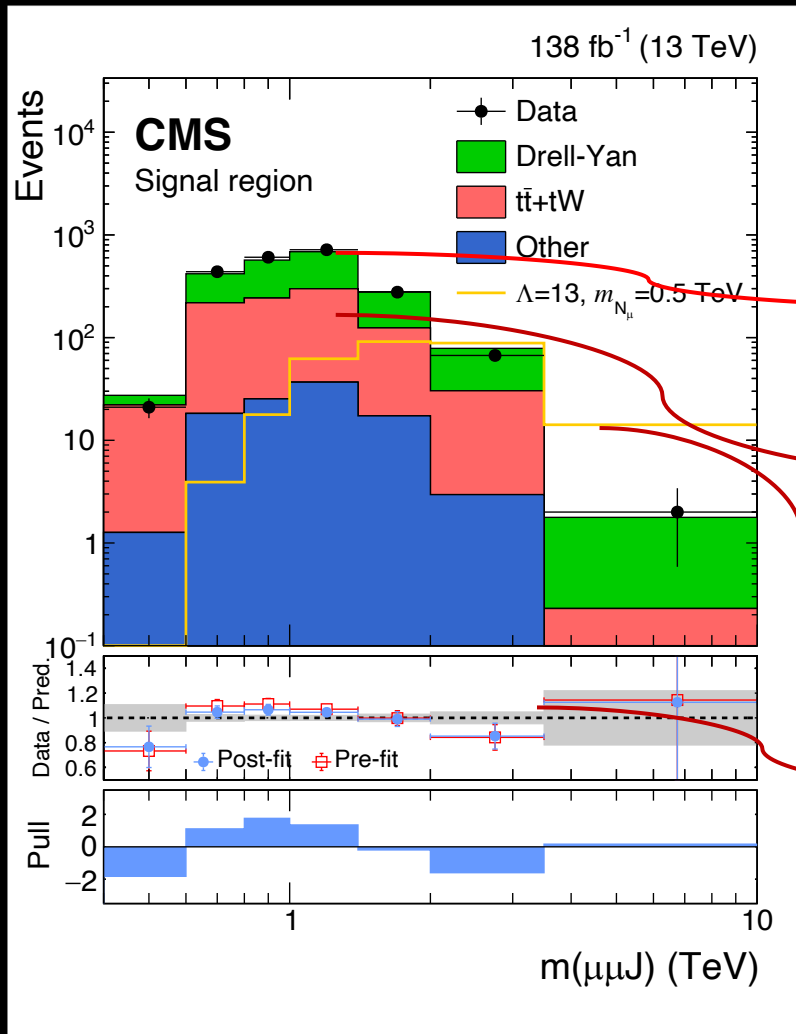
$$N_i = L \times \epsilon_i \times \sigma_i$$



$$\epsilon_i = \frac{N_{pass}}{N_{tot}} \Rightarrow \delta\epsilon_i = \sqrt{\frac{\epsilon_i}{N_{tot}} (1 - \epsilon_i)} \rightarrow \text{Binomial error...}$$

Statistical errors are quite important, and they can be quantified using different approaches...we will get back to this later...

More Realistic Example: Data



Observed data

Expected background processes from the SM

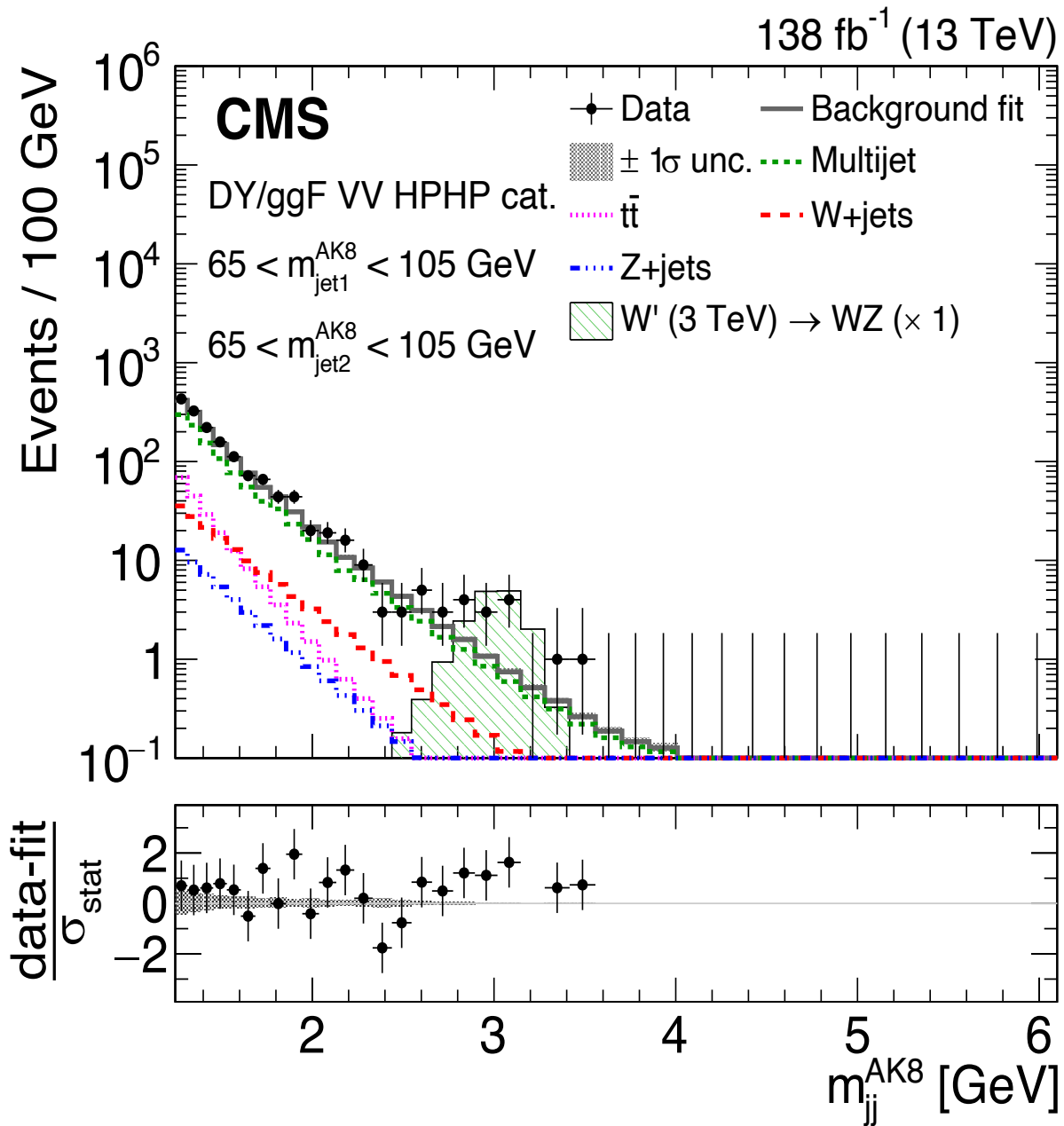
Expected number of signal events

Data over BKG prediction

Understanding the data...

- Note that in general we have a signal model that we want to test experimentally.
- **We apply some event selection criteria (filters or cuts), based on the characteristics of the hypothetical signal, in order maximize the probability to observe the process of interest above the background (BKG) (noise).**
- Note that after we apply our filters to the actual data, we will have some events passing our selections.
- The main questions are:
 - **What is the composition of these events?**
 - **Is the BKG prediction consistent with the observed data?**
 - **Do we have an excess of events above the BKG prediction, consistent with the signal hypothesis? If so, how significant, based upon the statistical and systematic errors?**

Understanding the data...



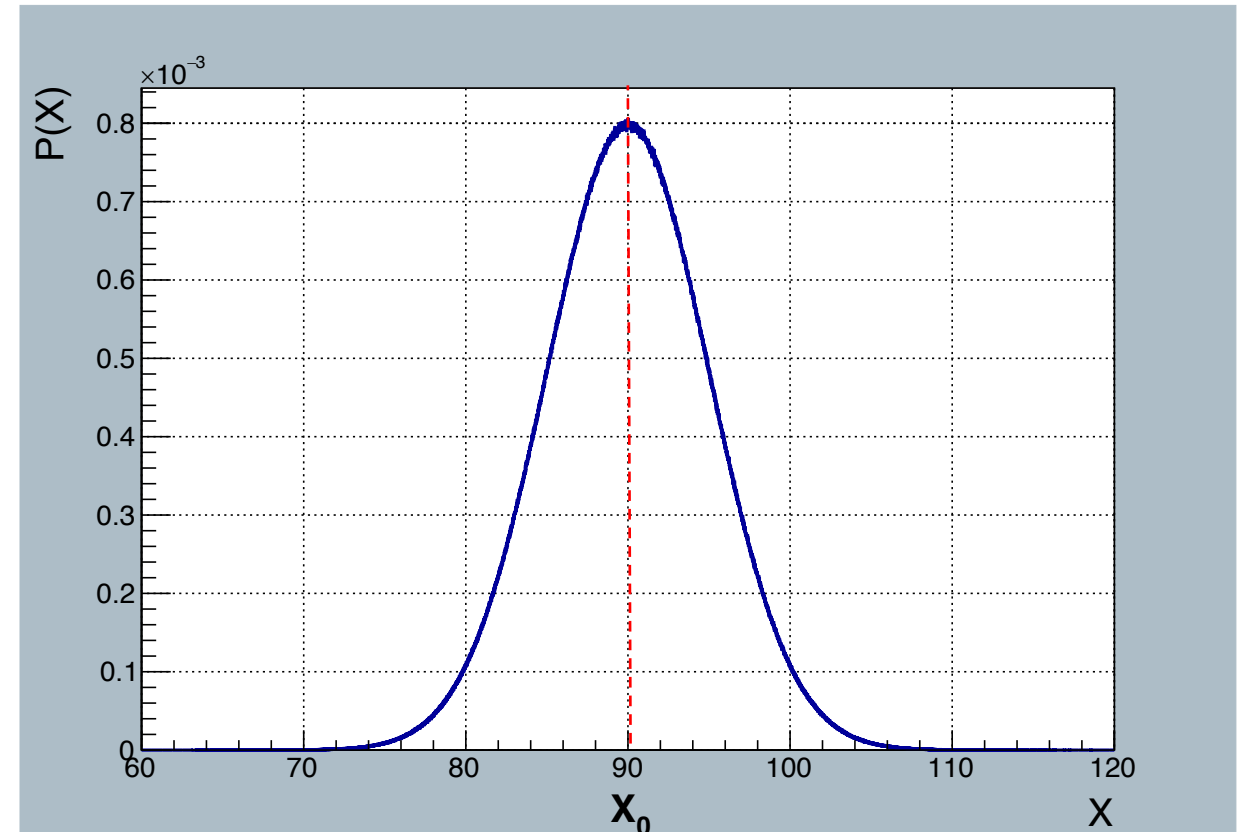
Probability Distributions

- We are concerned about estimating the probability to detect signal events inside data, **knowing that the data is composed of events from different background processes and possibly signal.**
- Suppose we are performing measurements of a random observable X .

- **Normalization:** $\int_{-\infty}^{+\infty} P(x) dx = 1$

- **Moments:** $\alpha_n = \int x^n P(x) dx$

Probability Density Function (PDF)



Probability Distributions

- The mean is the average result of many measurements

$$\mu \equiv \int x P(x) dx = \langle x \rangle$$

- The variance is the width of the PDF about the mean

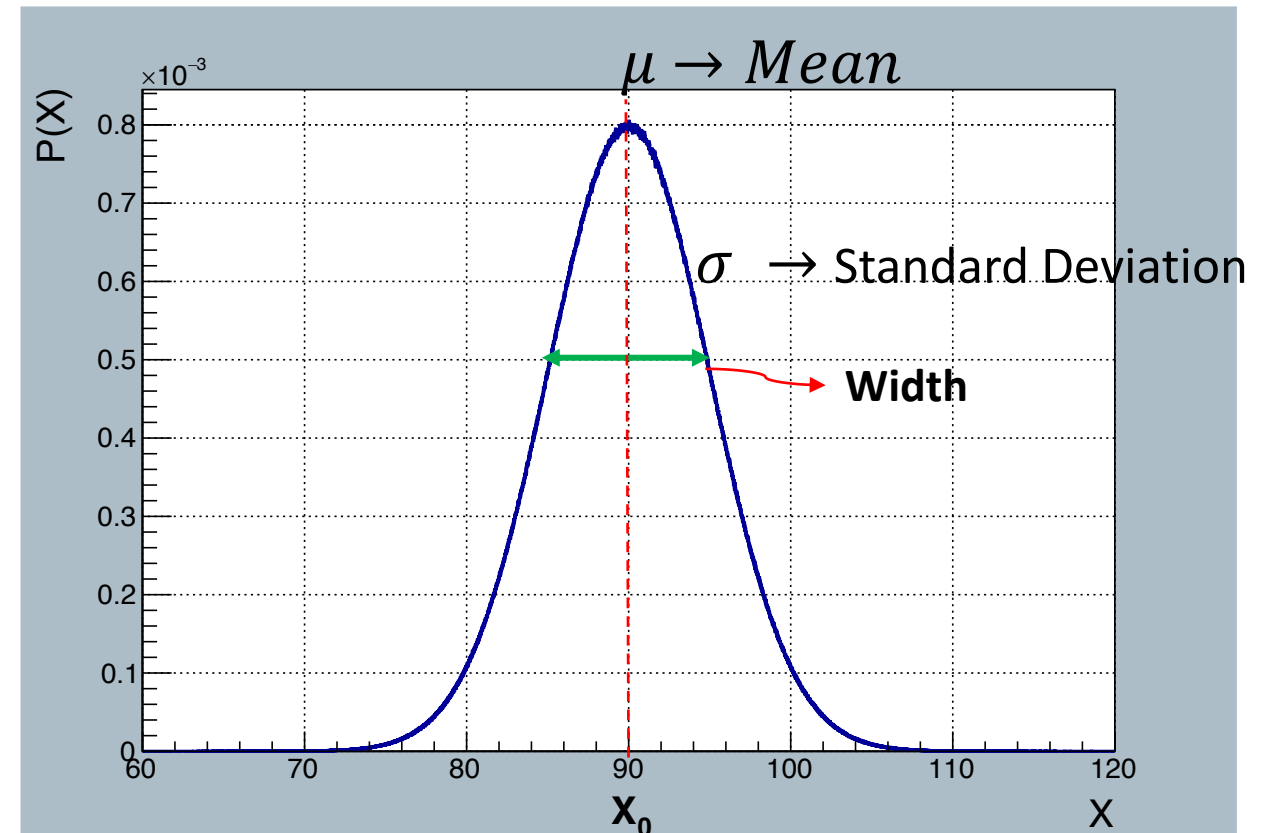
$$\text{Var}(x) \equiv \int \langle (x - \mu)^2 \rangle P(x) dx = \sigma^2 = \langle x^2 \rangle - \mu^2$$

Mean of squares

$$\int x^2 P(x) dx$$

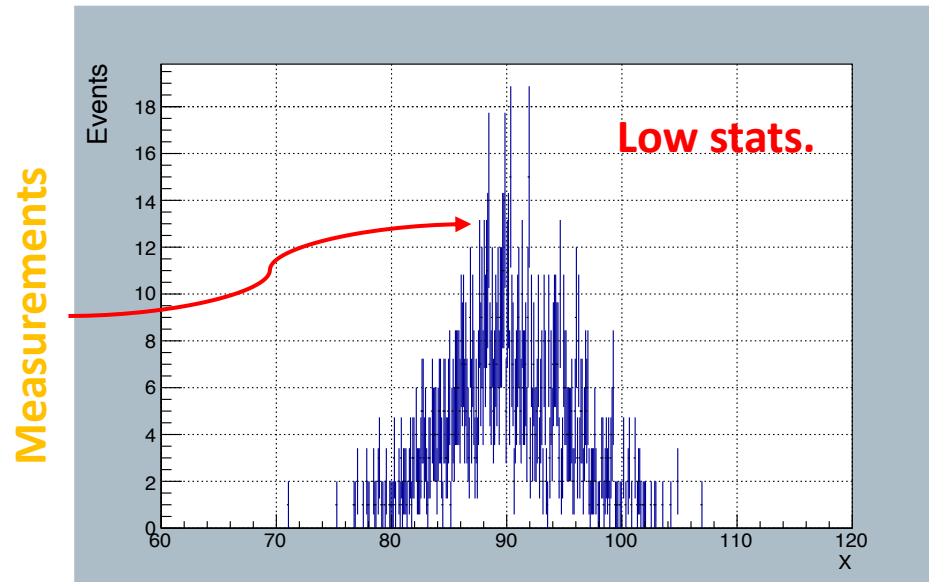
We do not know the PDF, we just have several measurements distributed according to a PDF

Probability Density Function (PDF)

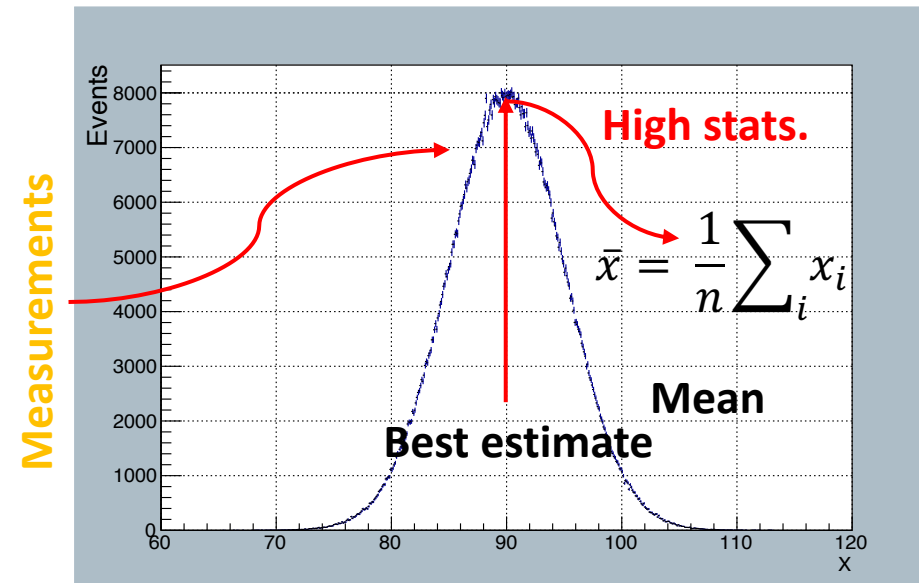


Probability Distributions

- In general, we do not know the exact shape of the PDF.
- What we have experimentally, is a set of measurements that we use to build the PDF.
- Note that the accuracy of the shape of the PDF will be determined by the number and quality of the measurements.



12/5/22

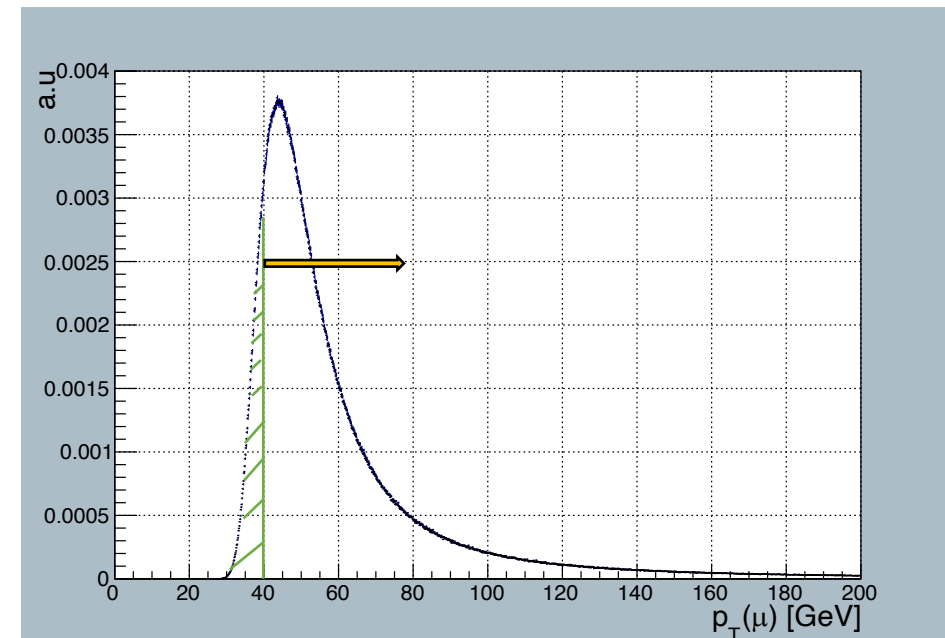
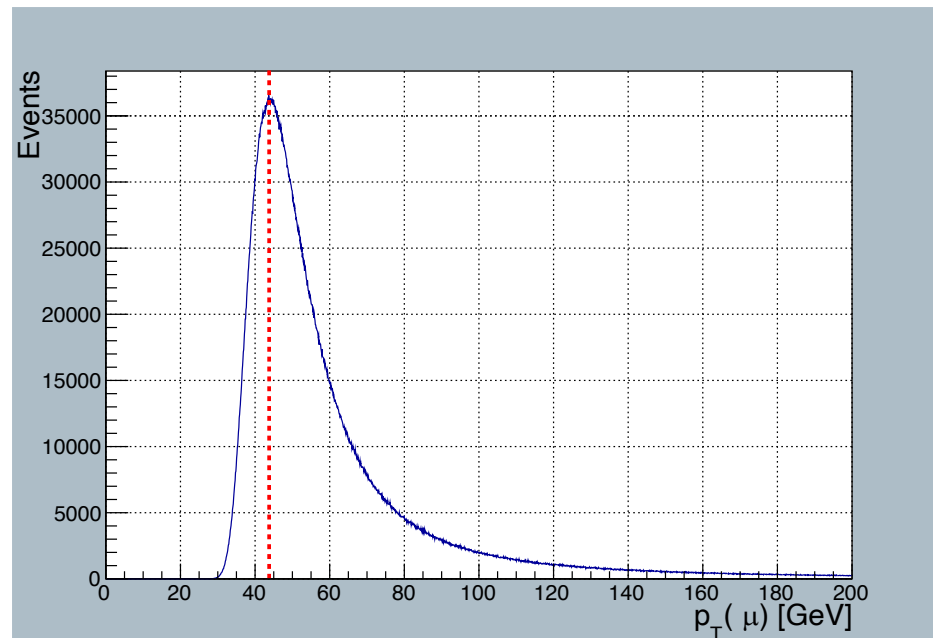


Andrés Flórez

Population vs Sample

- **Two important concepts:**

- **Population:** The entire set of data from which we want to perform a statistical study.
- **Sample:** Is a subset of the population that contains some specific characteristics.



Variance..

- Variance of the sample:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Single measurement → x_i Mean → \bar{x}

- In a perfect world with infinite statistics, where we know the true PDF, how are related the sample variance, S^2 , and the true variance σ^2 ?

$$\langle S^2 \rangle = \langle (x_i - \bar{x})^2 \rangle = \langle x_i^2 \rangle - 2\langle x_i \bar{x} \rangle + \langle \bar{x}^2 \rangle = \langle x_i^2 \rangle - 2 \left\langle x_i \frac{1}{n} \sum_j x_j \right\rangle + \left\langle \frac{1}{n^2} \left(\sum_j x_j \right)^2 \right\rangle$$

$$\langle S^2 \rangle = \langle x_i^2 \rangle - \frac{2}{n} \langle x_i^2 + \sum_{i \neq j} x_i x_j \rangle + \frac{1}{n^2} \left(n \langle x_i^2 \rangle + n(n-1) \langle x_i x_j \rangle_{i \neq j} \right) = \langle x^2 \rangle - \frac{1}{n} \langle x^2 \rangle + \frac{n-1}{n} \langle x_i x_j \rangle_{i \neq j}$$

$$\langle S^2 \rangle = \frac{n-1}{n} \left(\langle x^2 \rangle - \langle x_i x_j \rangle_{i \neq j} \right) \Rightarrow \frac{n-1}{n} (\langle x^2 \rangle - \mu^2) = \frac{n-1}{n} \sigma^2$$

Variance..

- Note that the sample variance is $\frac{n-1}{n}$ smaller than the true variance.
- Also, note that if “n” is very large $S^2 \rightarrow \sigma^2$.
- It is easy to see that the sample mean is the best unbiased estimate of the true mean:

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

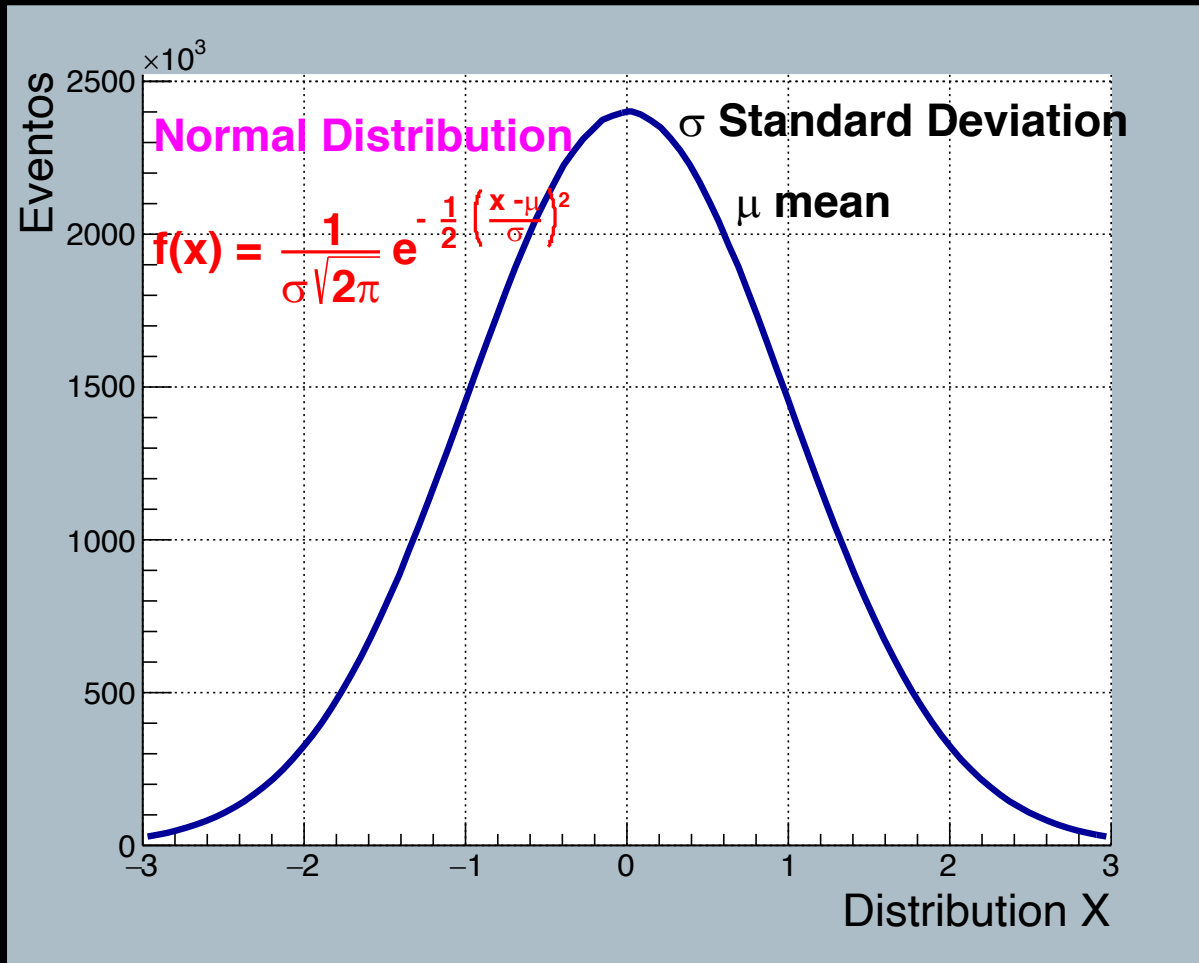
- Naturally, we must ask, what is the error on the sample mean?

$$\sigma_{\bar{x}}^2 = \langle (\bar{x} - \mu)^2 \rangle = \left\langle \left(\frac{1}{n} \sum_i x_i - \mu \right)^2 \right\rangle \Rightarrow \frac{\langle x^2 - \mu^2 \rangle}{n} = \frac{\sigma^2}{n}$$

Since σ^2 is the uncertainty on a single measurement, it means that the uncertainty on the mean is a factor \sqrt{n} smaller! Therefore, if our number of independent measurements increase, we reduce our uncertainty accordingly. Finally, note that this is a general result, independent of the distribution (PDF).

Now, lets see some popular distributions

Normal Distribution (Gaussian)



```

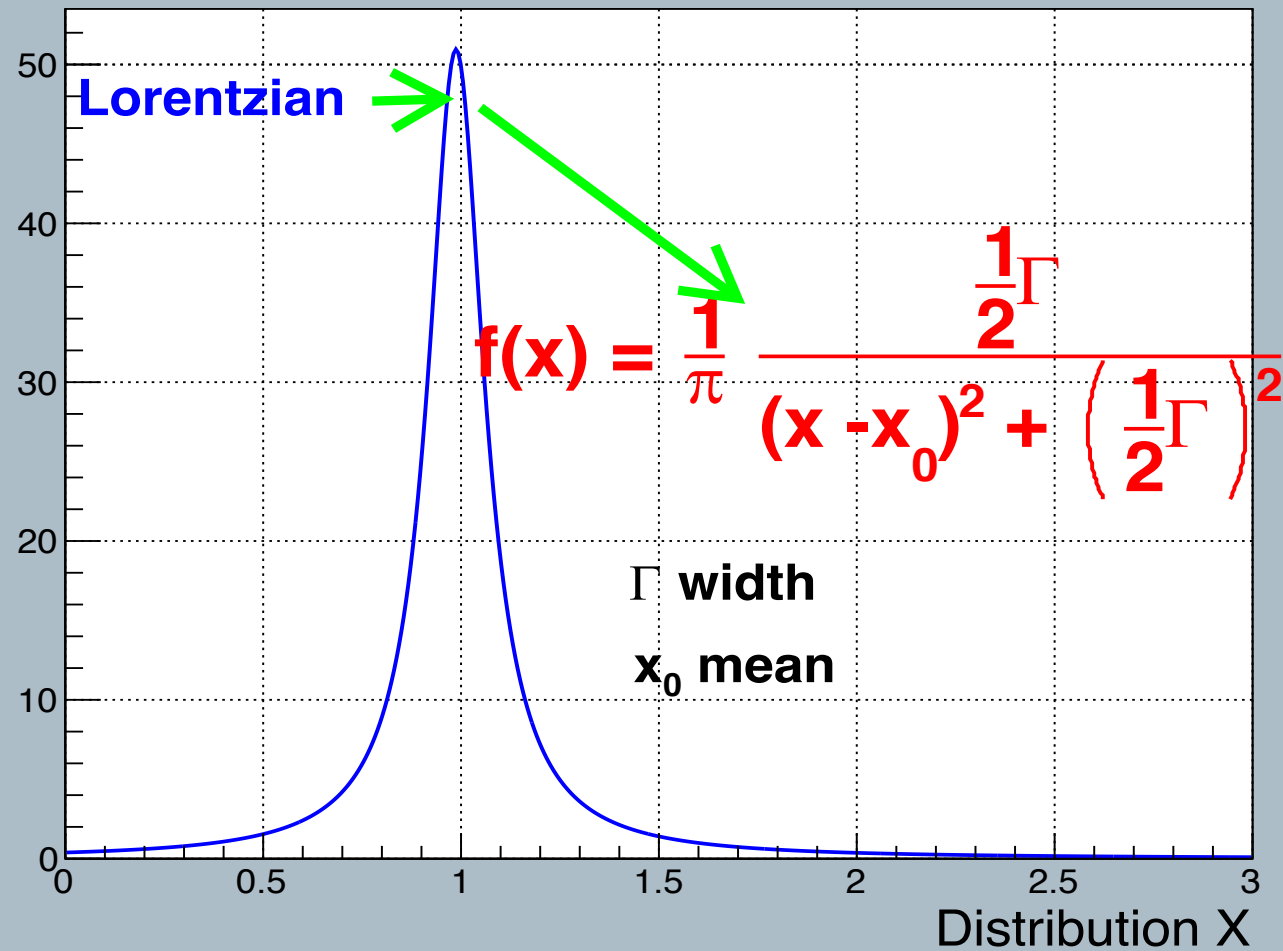
#include<iostream>
#include<TH1D.h>
#include<TCanvas.h>
#include<TRandom.h>
#include<TStyle.h>
#include<TROOT.h>
#include<TPaveText.h>
void Gaussian_0(){
  TCanvas *c1 = new TCanvas("c1","Fitting Demo",10,10,700,500);
  c1->SetFillColor(33);
  c1->SetFrameFillColor(10);
  c1->SetGrid();

  TH1D* h1 = new TH1D("h1", " ", 600,-3.0,3.0);

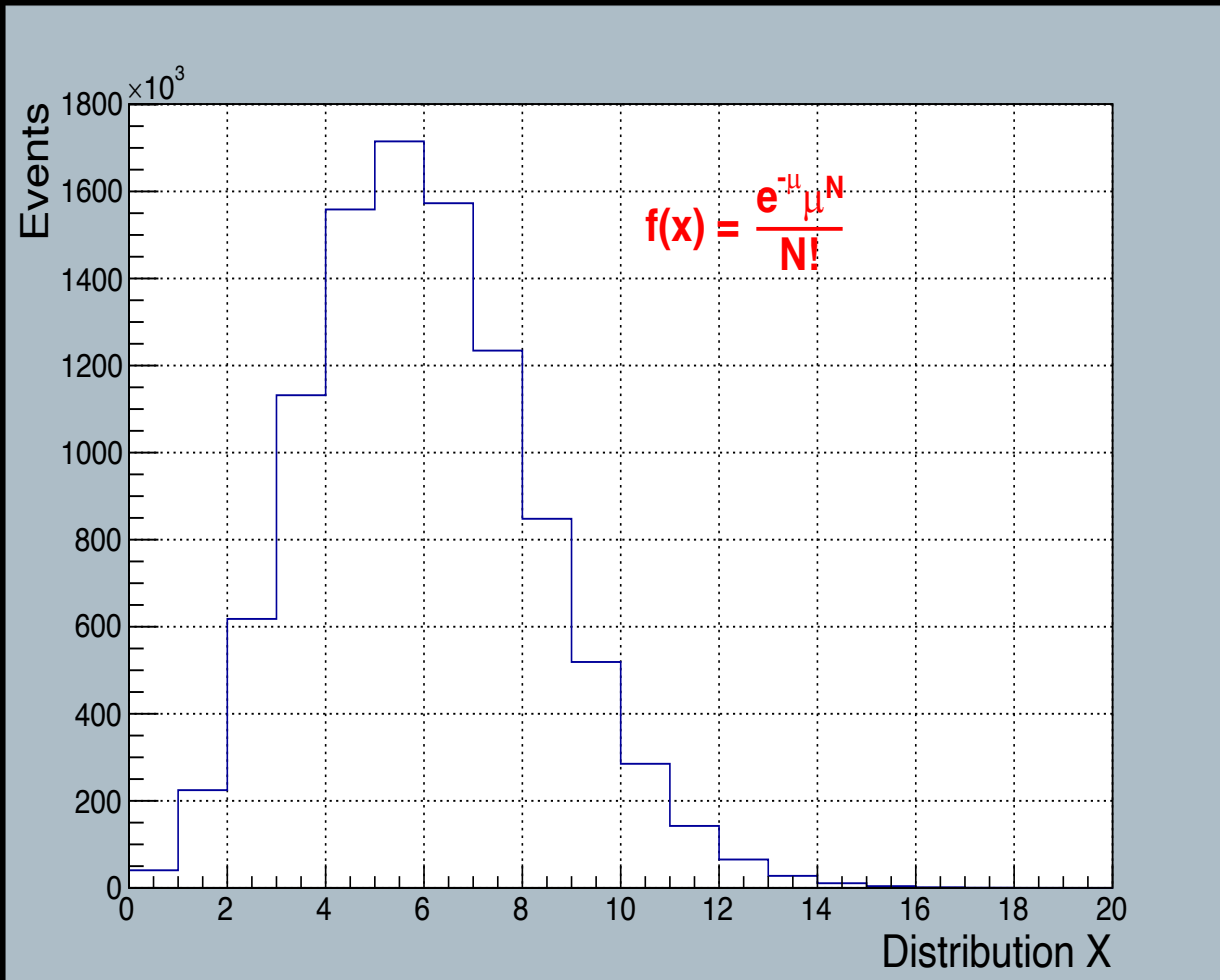
  TRandom3 rndgen;
  for(double i = 0; i < 10000000; i++) {
    double rnd = rndgen.Gaus(0., 1.0);
    h1->Fill(rnd);
  }

  h1->Draw("HIST1");
  c1->cd();
}
  
```

Lorentzian



Poissonian



```

#include<iostream>
#include<TH1D.h>
#include<TF1.h>
#include<TCanvas.h>
#include<TRandom.h>
#include<TStyle.h>
#include<TLegend.h>
#include<TROOT.h>
#include<TPaveText.h>
void Pois(){
    TCanvas *c1 = new TCanvas("c1","Fitting Demo",10,10,700,500);
    c1->SetFillColor(33);
    c1->SetFrameFillColor(41);
    c1->SetGrid();

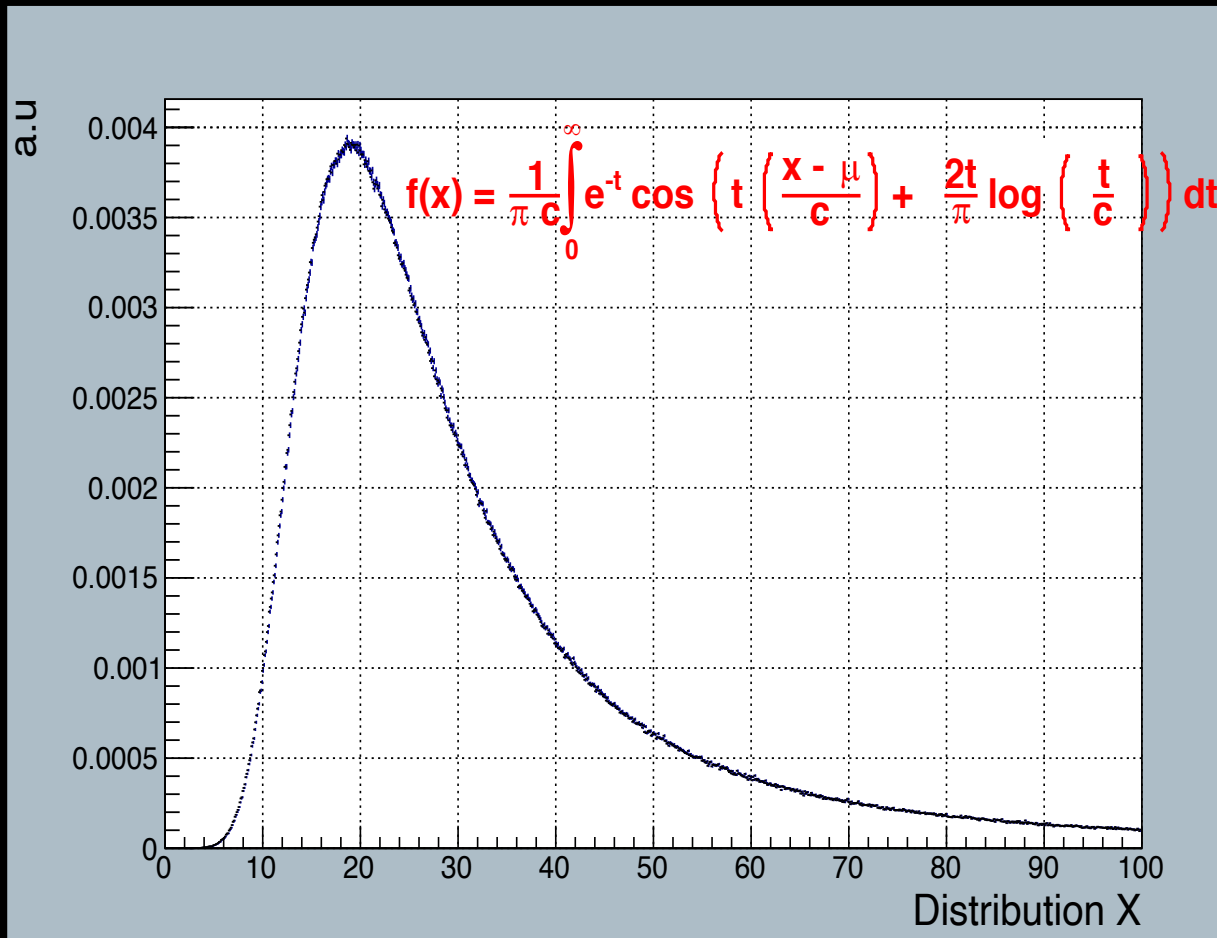
    TH1D* h1 = new TH1D("h1", " ", 20, 0.,20.);

    TRandom3 rndgen;
    for(double i = 0; i < 10000000; i++) {
        double rnd = rndgen.Poisson(5.5);

        h1->Fill(rnd);
    }

    h1->Draw();
    c1->cd();
}
  
```

Landau



```

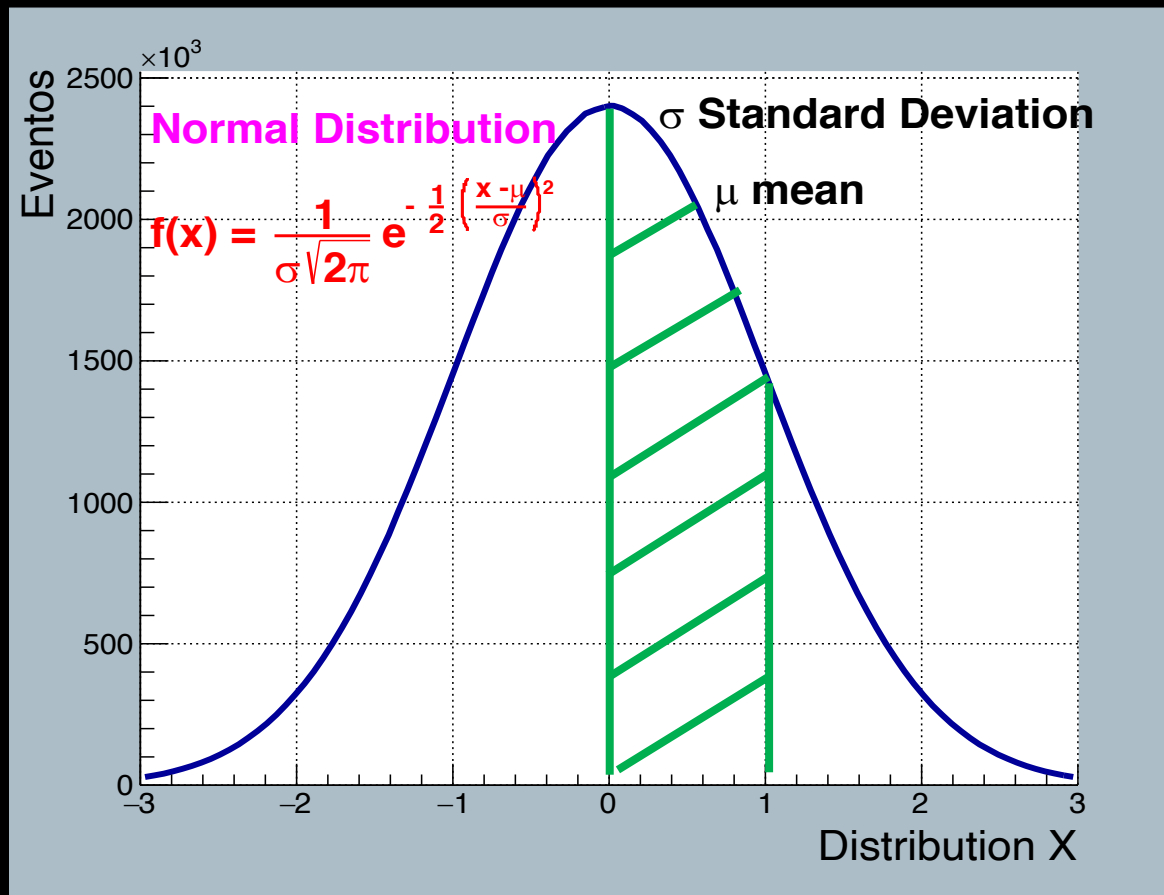
#include<iostream>
#include<TH1D.h>
#include<TCanvas.h>
#include<TRandom.h>
#include<TStyle.h>
#include<TR00T.h>
#include<TPaveText.h>
void Landau(){
    TCanvas *c1 = new TCanvas("c1","Fitting Demo",10,10,700,500);
    c1->SetFillColor(33);
    c1->SetFrameFillColor(10);
    c1->SetGrid();

    TH1D* h1 = new TH1D("h1", " ", 1000,0.,100);

    TRandom3 rndgen;
    for(double i = 0; i < 10000000; i++) {
        double rnd = rndgen.Landau(20,5.0);
        h1->Fill(rnd);
    }
    h1->Scale(1./h1->Integral());

    h1->Draw("HIST1");
    c1->cd();
}
  
```

Probability







- We understand probability as the area under the curve in a distribution, with respect to the total area.
- We interpret it as the feasibility of a process to occur, given a known PDF.

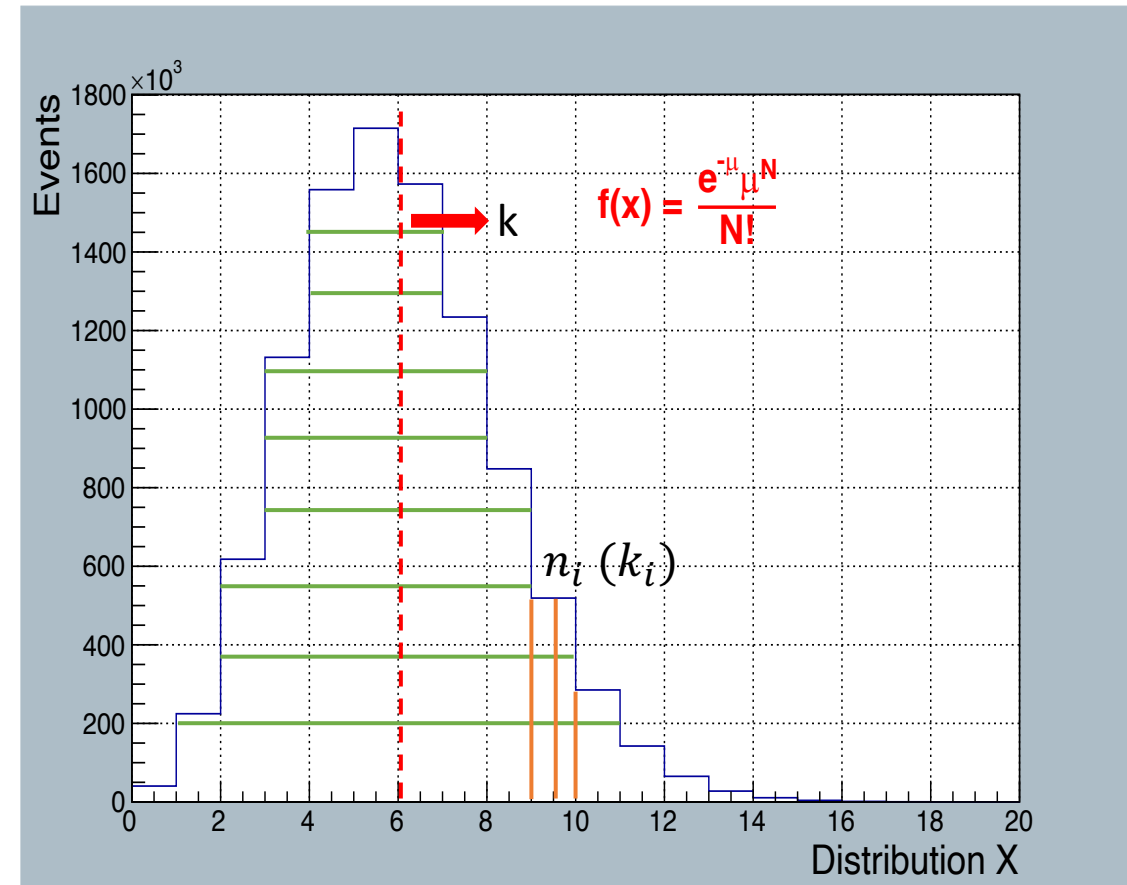
Common Probability Distributions

- **Binomial**: Describes a random process with two possible outcomes.
 - p → probability of one of the outcomes and $(1-p)$ → probability of the second outcome.
 - If we repeat the process a given number of times, then we obtain a **distribution of outcomes**.
- **Poissonian**: Discrete and random process with a fixed mean.
- **Gaussian**: continuous limit, obtained with high statistics (**see next**)

Let's define some important concepts in order to understand the use of probabilities in particle physics

Important concepts

- n : Total number of events collected. 
- n_i : Total number of events in bin "i". 
- k : Total number of events passing some selection criteria. 
- k_i : Total number of events passing some selection criteria in bin "i". 
- $\epsilon = \frac{k}{n} \rightarrow$ efficiency



Binomial Statistics

- We have “ k ” positive outcomes (success) out of “ n ” total independent measurements.
- The probability of success depends on the efficiency, ϵ

$$P(k; n, \epsilon) = \frac{n!}{k! (n - k)!} \epsilon^k (1 - \epsilon)^{n-k}$$

- For a binomially distributed case, the estimator is $k: n\epsilon$
- Variance:

$$\sigma_k^2 = \langle (k - \mu)^2 \rangle = \langle k^2 \rangle - \mu^2$$
$$\langle k^2 \rangle = \sum k^2 P(k; n, \epsilon) = \sum_{k=0}^n k^2 \frac{n!}{k!(n-k)!} \epsilon^k (1 - \epsilon)^{n-k}$$

Binomial Statistics

- Variance:

$$\langle k^2 \rangle = n\epsilon \sum_{k=1}^n k \epsilon^{k-1} (1-\epsilon)^{n-k} \frac{(n-1)!}{(k-1)!(n-k)!}$$

$$\langle k^2 \rangle = n\epsilon \sum_{k'=0}^{n-1} (k'+1) \epsilon^{k'} (1-\epsilon)^{n-1-k'} \frac{(n-1)!}{(n-1-k')!}$$

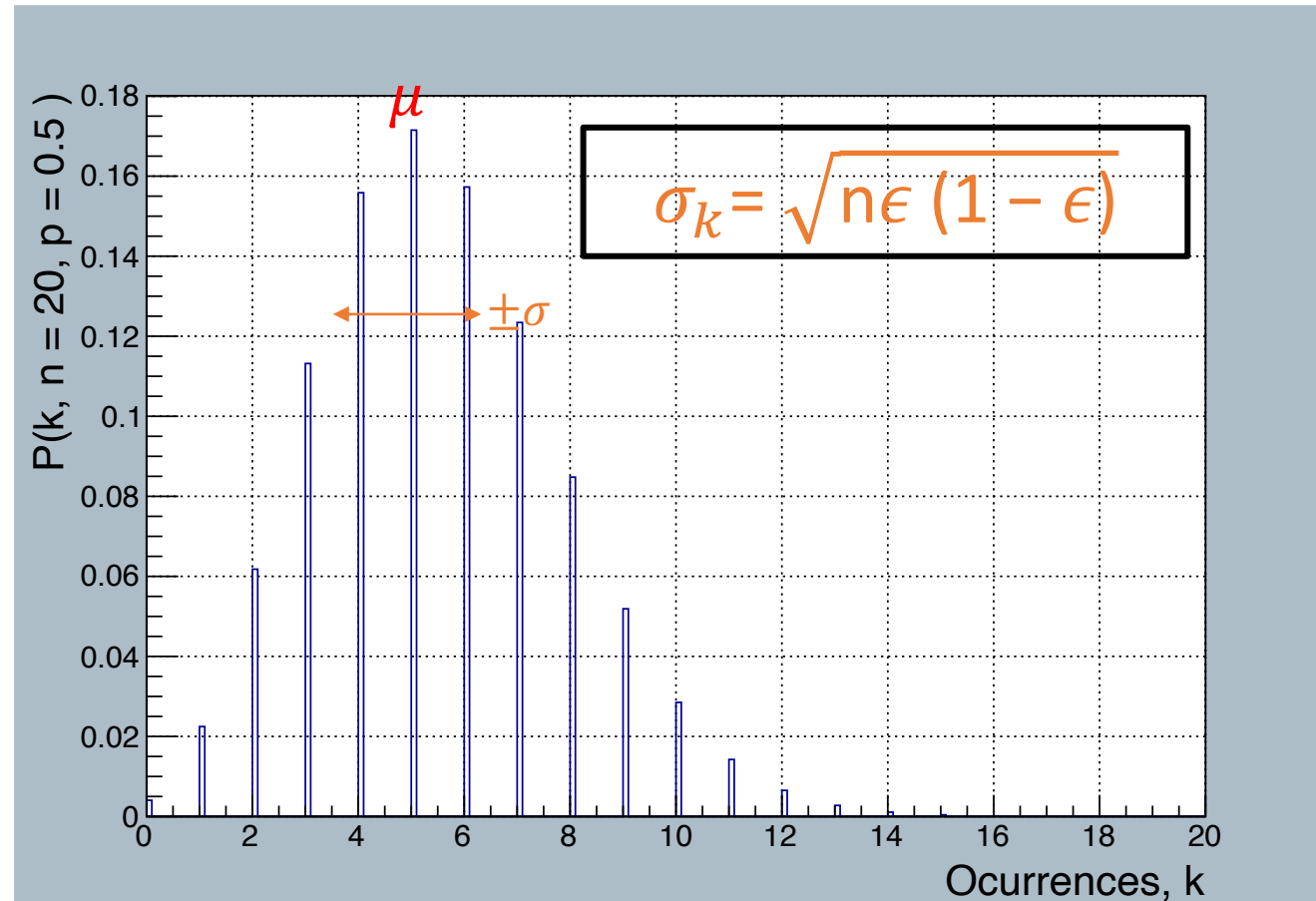
$$\langle k^2 \rangle = n\epsilon \sum_{k'=0}^{n-1} P(k; n-1, \epsilon) + n\epsilon \sum_{k'=0}^{n-1} k' P(k; n-1, \epsilon)$$

$$\langle k^2 \rangle = n\epsilon + n\epsilon \times \epsilon(n-1) = n\epsilon(n\epsilon - \epsilon + 1)$$

$$\sigma_k^2 = n\epsilon(n\epsilon - \epsilon + 1) - (n\epsilon)^2 = n\epsilon(1 - \epsilon)$$

$$\sigma_k^2 = n\epsilon(1 - \epsilon)$$

Binomial Statistics



Binomial Statistics: Efficiency

- Now, imagine we have a MC signal sample for a given theoretical model and we want to determine the efficiency of applying to a criterion in order to reduce the BKG rate.
- Then, we have N_{pass} events out of N_{tot} total, passing the criterion.
- Therefore, as shown before, the efficiency is $\epsilon_c = \frac{N_{pass}}{N_{tot}}$
- Note that now, our estimator is ϵ_c .
- Therefore

$$P(N_{pass}; N_{tot}) = \frac{N_{tot}!}{N_{pass}! (N_{tot} - N_{pass})!} \epsilon_c^{N_{pass}} (1 - \epsilon_c)^{N_{tot} - N_{pass}}$$

Binomial Statistics: Efficiency

- Note that for this case, the best estimate of the efficiency is clearly the efficiency itself!
- So, $\sigma_c^2 = \langle \epsilon_c^2 \rangle = N_{tot} \epsilon_c (1 - \epsilon_c) \times \frac{1}{N_{tot}^2} = \frac{\epsilon_c (1 - \epsilon_c)}{N_{tot}}$
- For example, if we have 80 events passing out of 100 events:
$$\epsilon_c = 0.80 \pm 0.04$$
- Note that we performed the calculation considering a single measurement for ϵ_c . Imagine we run the simulation of the signal process many times, **varying randomly the initial seed**, in order to obtain a more accurate measurement of the efficiency.

Binomial Statistics: Efficiency

- With this in mind:

$$\langle \epsilon_c \rangle = \frac{\langle N_{pass} \rangle}{N_{tot}} = \frac{N_{tot} \epsilon}{N_{tot}} = \epsilon$$

- And the variance? Note that for a single measurement:

$$\sigma_c^2 = \frac{\epsilon_c (1 - \epsilon_c)}{N_{tot}} = \frac{N_{pass} (N_{tot} - N_{pass})}{N_{tot}^3}$$

- Therefore

$$\langle \sigma_c^2 \rangle = \frac{N_{tot} \langle N_{pass} \rangle}{N_{tot}^3} - \frac{\langle N_{pass}^2 \rangle}{N_{tot}^3} = \frac{N_{tot}^2 \epsilon}{N_{tot}^3} - \frac{N_{tot}^2 \epsilon^2 - N_{tot} \epsilon^2 + N_{tot} \epsilon}{N_{tot}^3} = \frac{(N_{tot} + 1)}{N_{tot}} \sigma^2$$

See slide 33

Summary

- Understanding some fundamental statistical concepts is fundamental in particle physics: mean, variance, PDF....
- Some PDF are widely used in our field, and many others, and it is important to understand their similarities and differences: Gaussian, Poissonian, Lorentzian, etc.

12/5/22



Thank you!

Andrés Flórez