XXIII DAE-BRNS HIGH ENERGY PHYSICS SYMPOSIUM 2018

# CUT-BASED PHOTON ID TUNING OF CMS USING GENETIC ALGORITHM

Debabrata Bhowmik

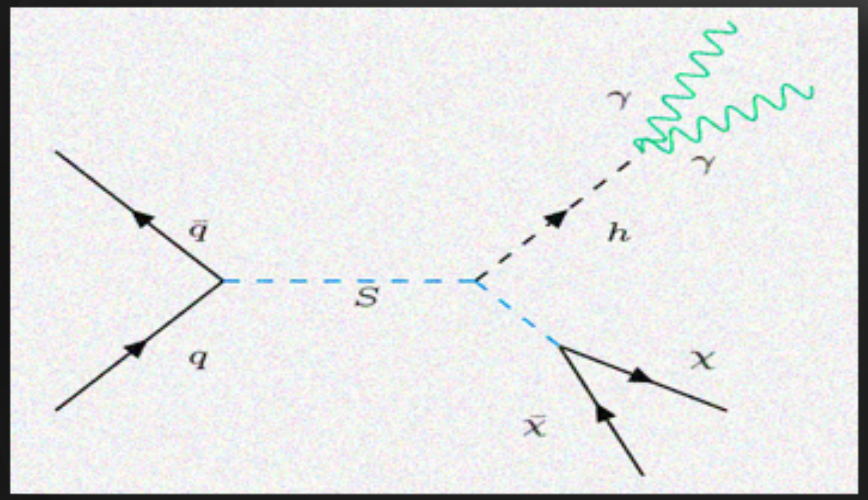
*Saha Institute of Nuclear Physics*

# Outline

- Motivation
- Aim
- Pile up correction and how Effective Area comes into play
- How Isolation depends on Photon Pt
- Genetic Algorithm
- Extraction of Cut values

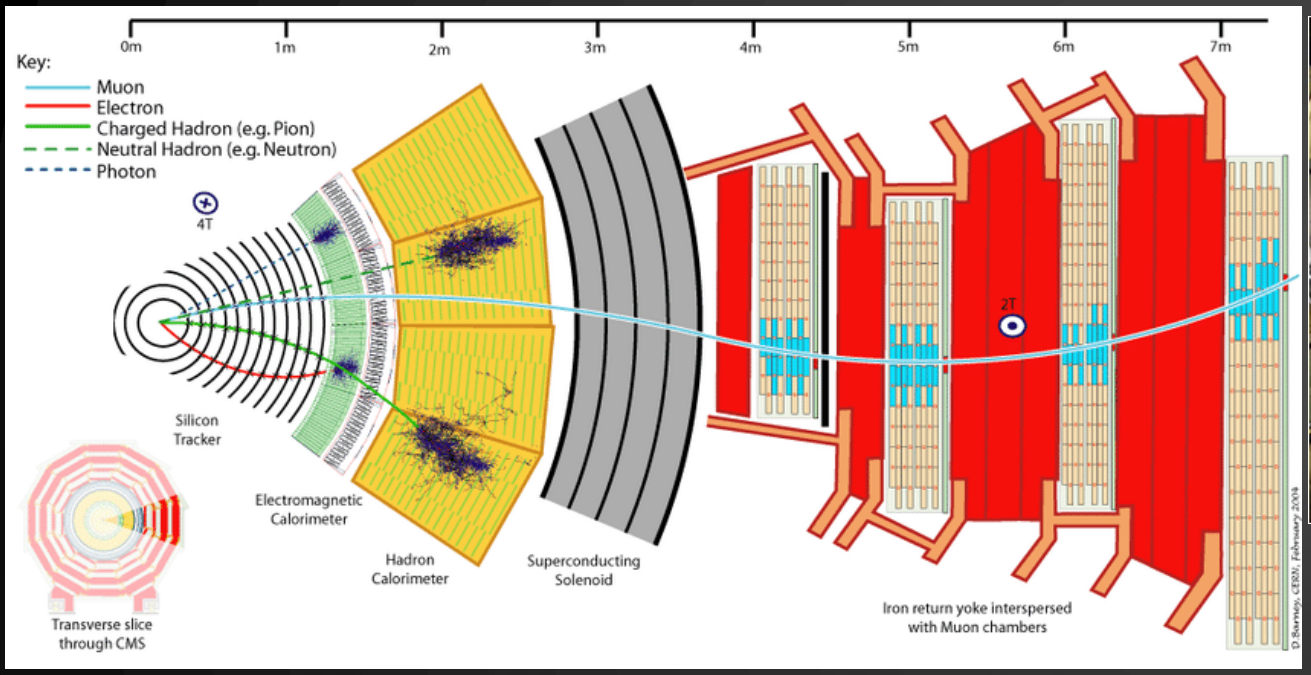
# Motivation

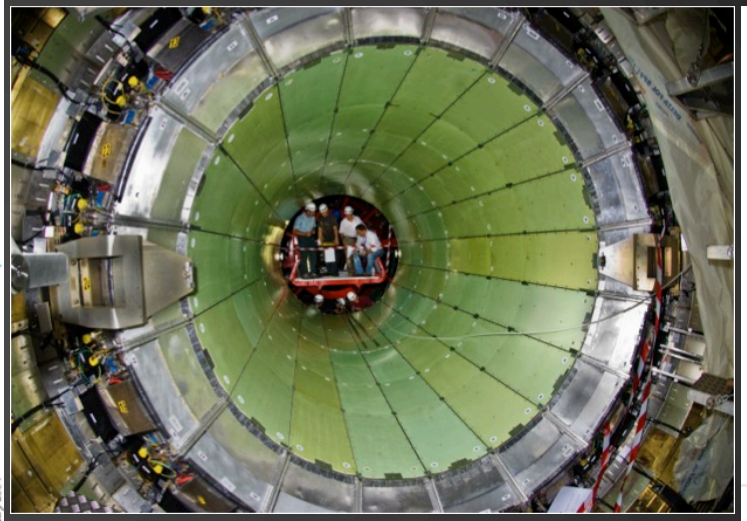
- Let's say we are interested in following kind of events or any other event which has photon(s) in its final state.

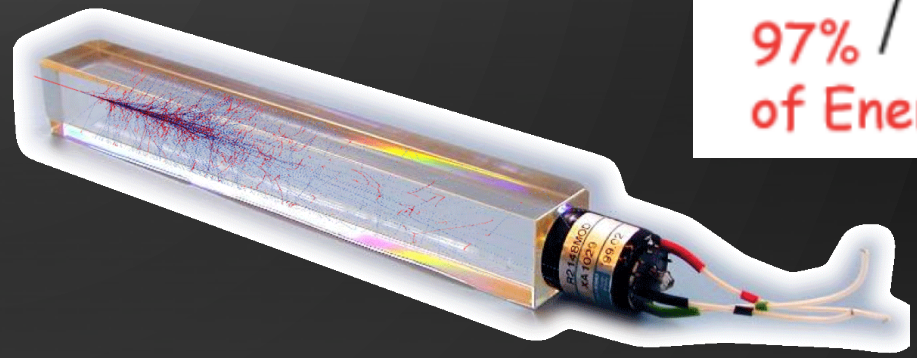


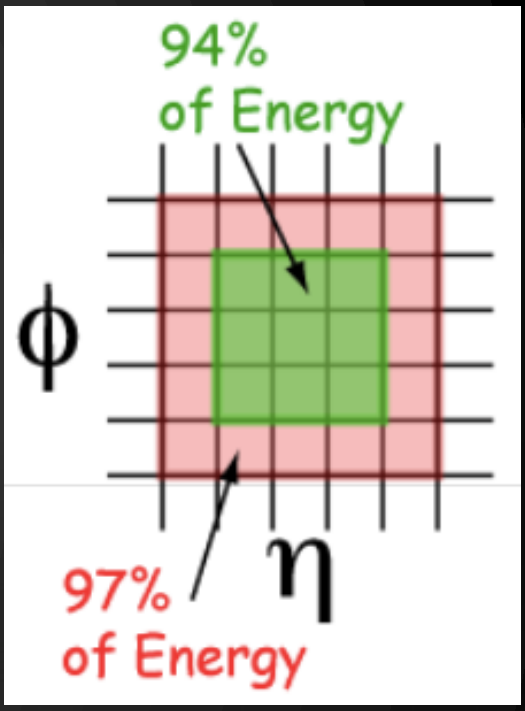

- So, in final state we expect only some missing energy balanced by a Higgs decaying to two photons
- The question is how do you know which photon is actually the one you are looking for, as you will see lots of them detected by your detector.
- In other word, which will you call a "photon" object to select a event of your interest

# CMS ELECTROMAGNETIC CALORIMETER

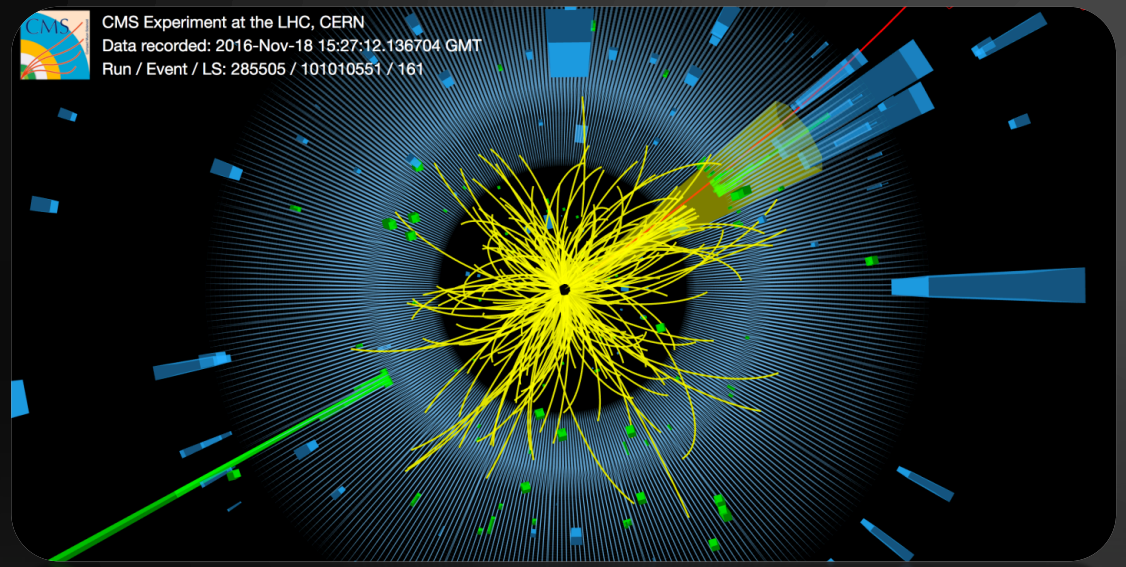




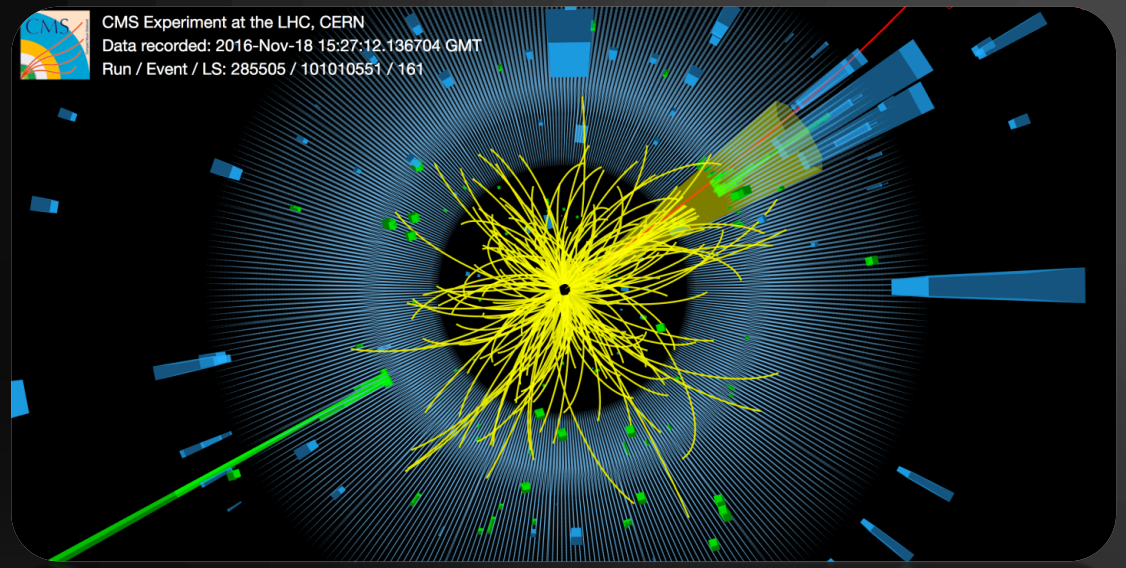


# AIM

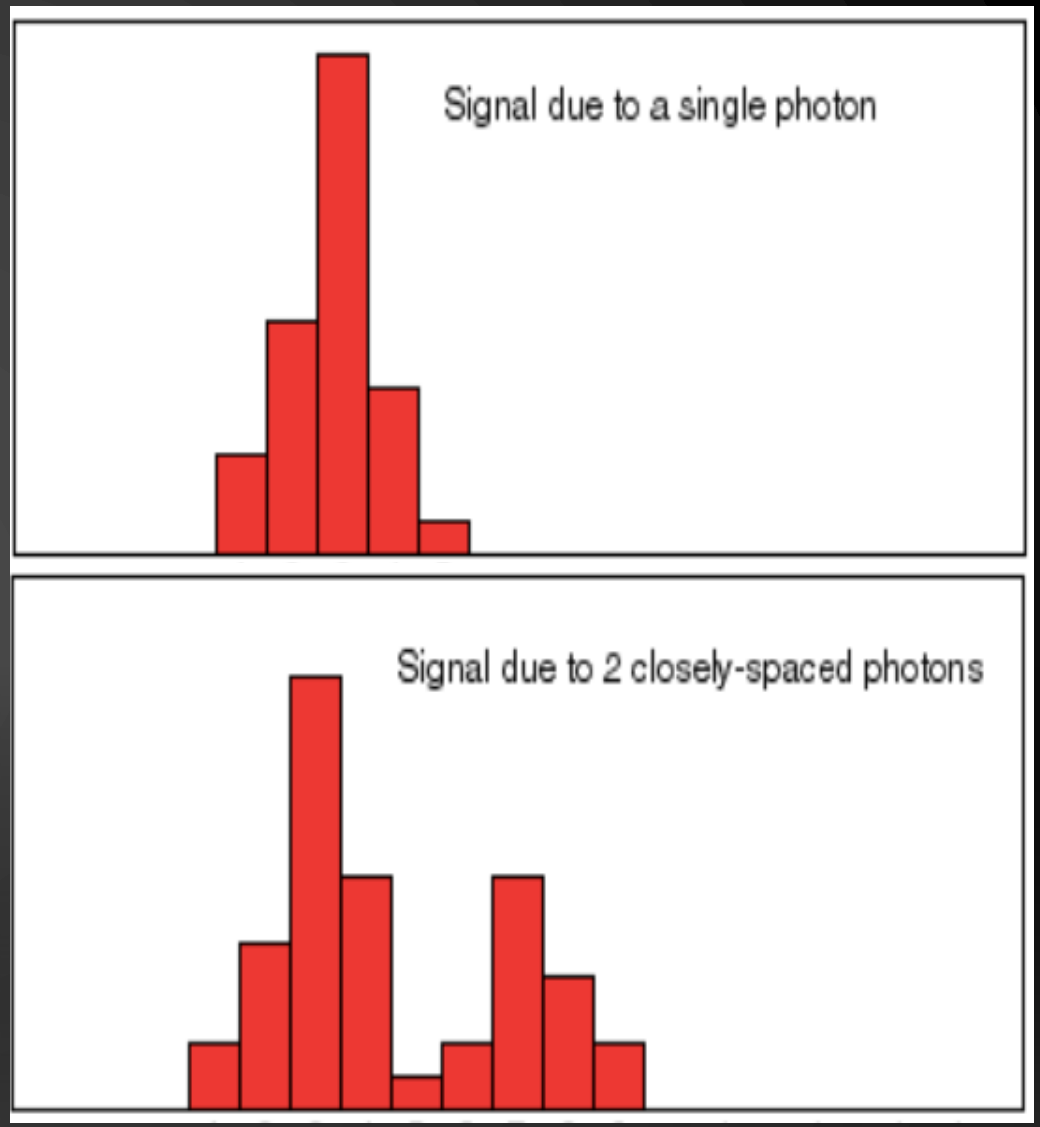



- Identify ptompt photons from the background ones.
- Five variables is chosen : shower shape variable $\sigma_{i\eta i\eta}$, H/E and three isolations (Photon, charged and neutral hadron)
- Optimize the cut values of these variables

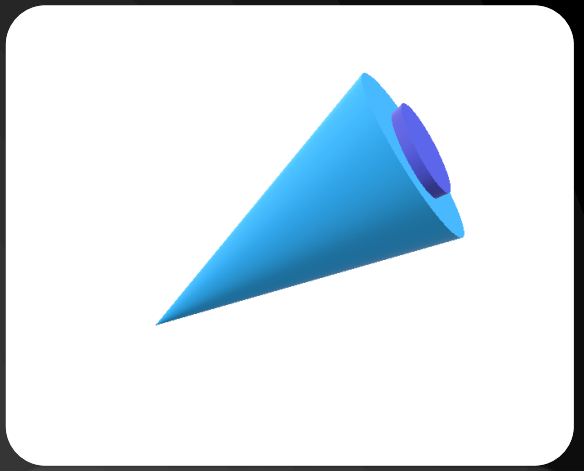
# AIM

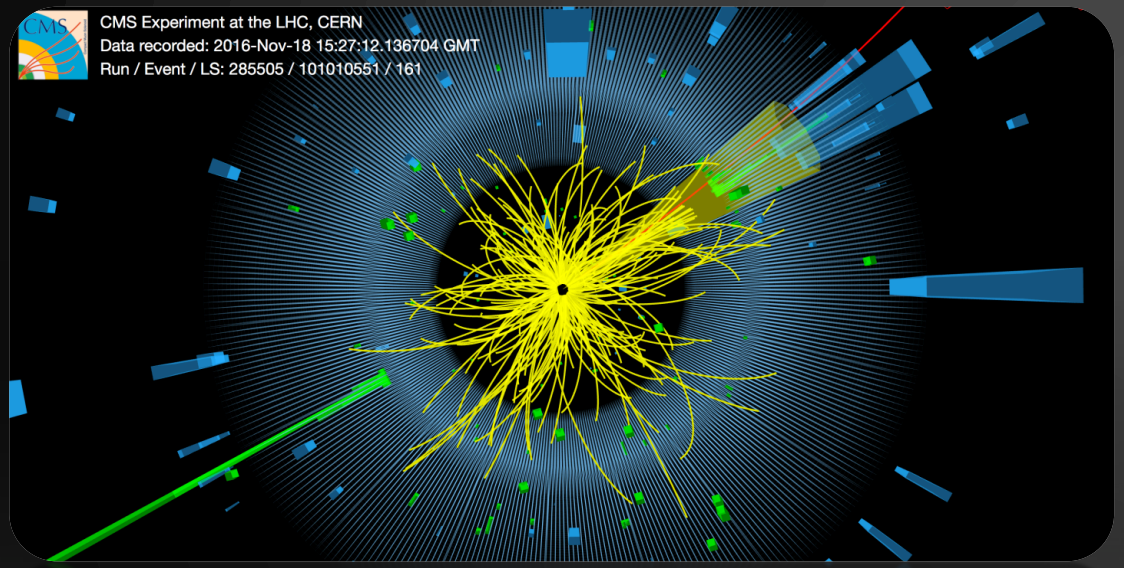



$$\sigma_{i\eta i\eta} = \left( \frac{\sum (\eta_i - \bar{\eta})^2 \omega_i}{\sum \omega_i} \right)^{1/2} ; \quad \bar{\eta} = \frac{\sum \eta_i \omega_i}{\sum \omega_i} ; \quad \omega_i = \max \left( 0, 4.7 + \log \frac{E_i}{E_{5\times5}} \right)$$



# AIM

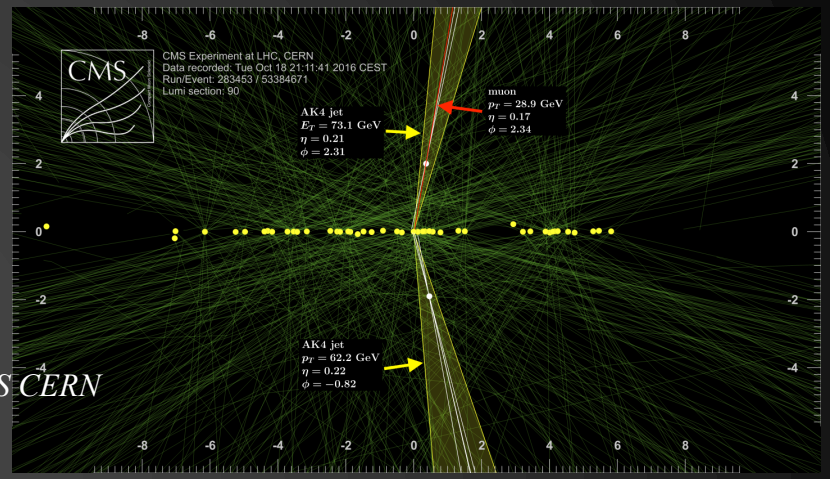




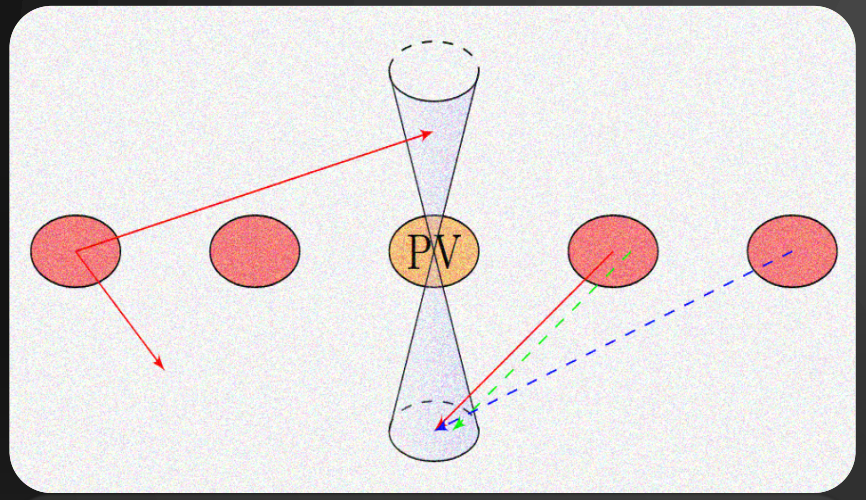

- Identify ptompt photons from the background ones.
- Five variables is chosen : shower shape variable $\sigma_{i\eta i\eta}$, H/E and three isolations (Photon, charged and neutral hadron)
- Optimize the cut values of these variables

# Why pile up correction is necessary

- In collision we will detect the particles not only coming from primary vertices but from the soft collisions too.




*Ref : CDS CERN*

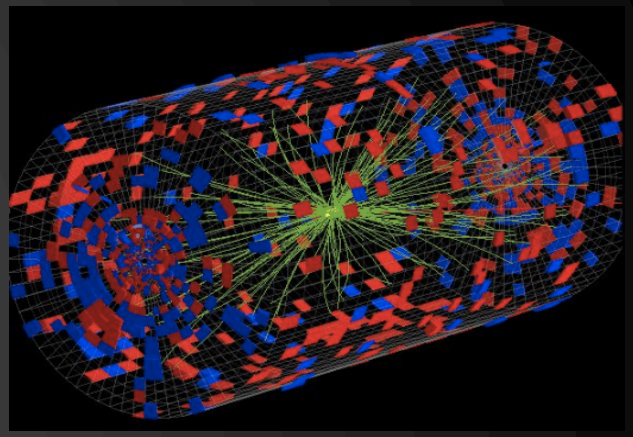



*Ref : Pileup measurement and mitigation techniques in CMS - A Perloff, on behalf of the CMS collaboration*

- If we are looking for isolated photon then the momentum of other particles coming from pile up may be counted.
- For charged particles we have tracks in the tracker
- But for neutral hadrons and photons it's obvious to remove the effect of pileup.
- But the question is how

# PU correction using ρ

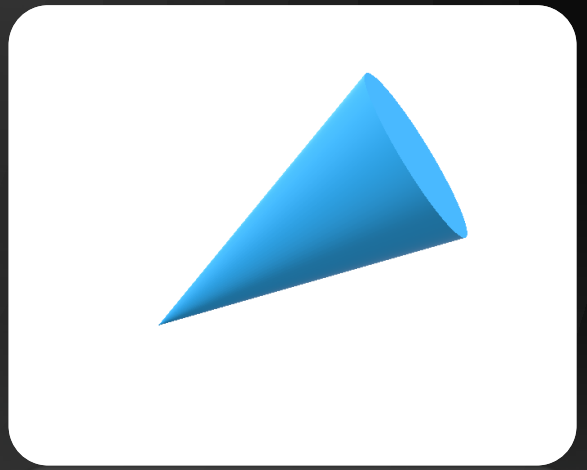
- If whole Ecal η-φ plane in N patches and area of the i-th patch is $A_i$ .
- Then ρ is defined as $\rho = \underset{i \in \text{patches}}{\text{median}} \left\{ \frac{p_{ti}}{A_i} \right\}$



- ρ is effective contribution in $P_t$ sum from pile up per unit area.
- If isolation sum is defined as the sum of pt(excluding seed particle) inside a cone, it should vary linearly with ρ.

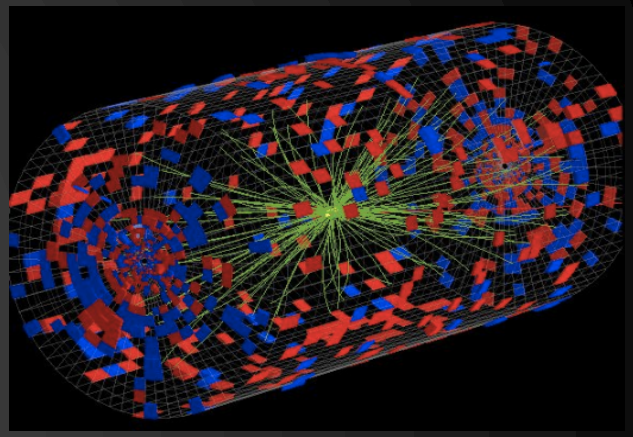
$$\text{Isolation} = \text{EA} \times \rho + \text{Isolation}_{\text{Corr}}$$

$$( \ Y \quad = \quad m \ . \ X \ + \quad C \ )$$



# PU correction using ρ

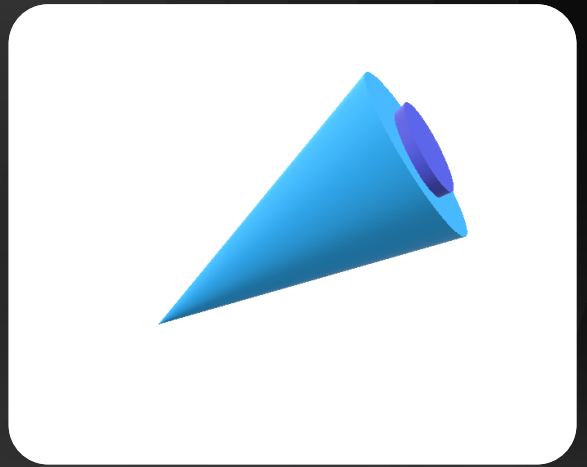
- If whole Ecal η-φ plane in N patches and area of the i-th patch is $A_i$ .
- Then ρ is defined as $\rho = \underset{i \in \text{patches}}{\text{median}} \left\{ \frac{p_{ti}}{A_i} \right\}$



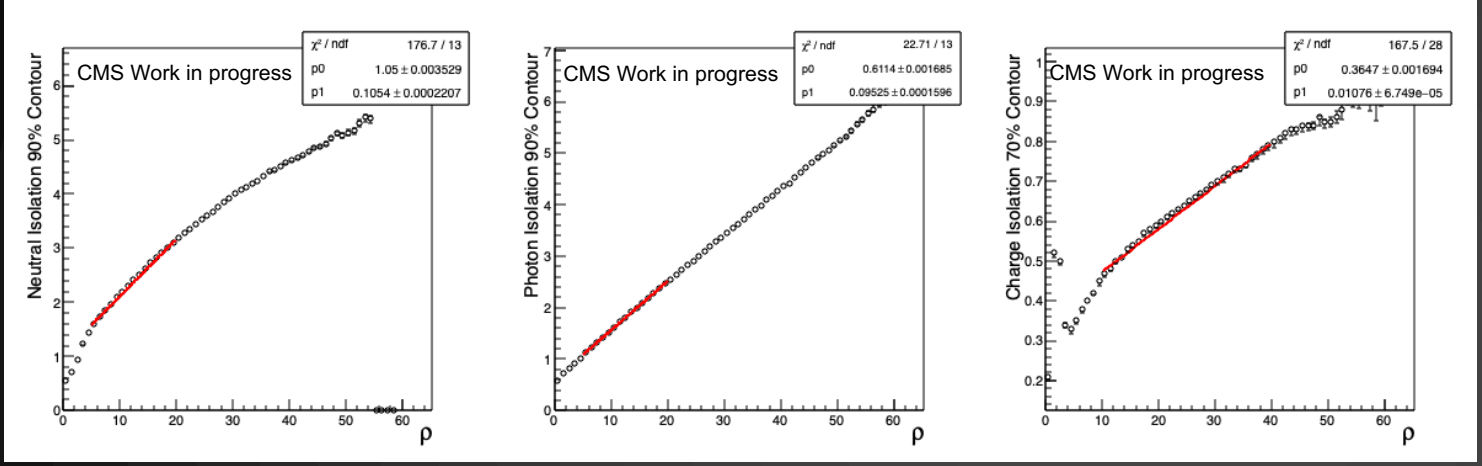
- ρ is effective contribution in $P_t$ sum from pile up per unit area.
- If isolation sum is defined as the sum of pt(excluding seed particle) inside a cone, it should vary linearly with ρ.

$$\text{Isolation} = \text{EA} \times \rho + \text{Isolation}_{\text{Corr}}$$

$$( \ Y \quad = \quad m \ . \ X \ + \quad C \ )$$



# Isolation vs ρ plots and effective area




| bin | EA charged hadrons(70% cont) | EA neutral hadrons(90% cont) | EA photons(90% cont) |
|---|---|---|---|
| abs(η)<1.0 | 0.0112 | 0.0668 | 0.1113 |
| 1.0<abs(η)<1.479 | 0.0108 | 0.1054 | 0.0953 |
| 1.479<abs(η)<2.0 | 0.0106 | 0.0786 | 0.0619 |
| 2.0 <abs(η) < 2.2 | 0.01002 | 0.0233 | 0.0837 |
| 2.2 <abs(η) < 2.3 | 0.0098 | 0.0078 | 0.1070 |
| 2.3 <abs(η) < 2.4 | 0.0089 | 0.0028 | 0.1212 |
| abs(η) > 2.4 | 0.0087 | 0.0137 | 0.1466 |

$1.0<\eta<1.479$

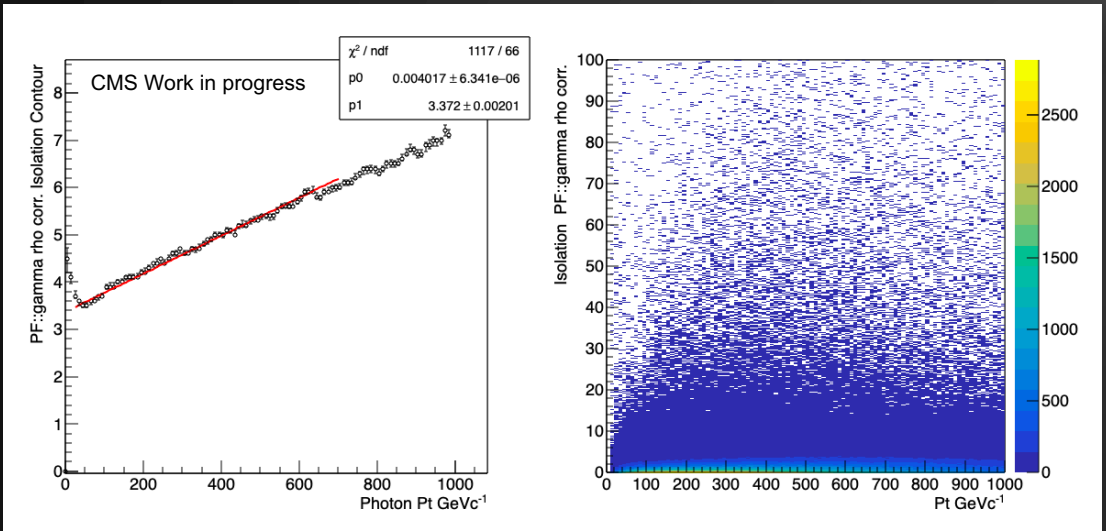
Slope is the Effective Area(EA)

$\text{Isolation} = \text{EA} \times \rho + \text{Isolation}_{\text{Corr}}$

$\text{Isolation}_{\text{Corr}} = \text{Isolation} - \text{EA} \times \rho$

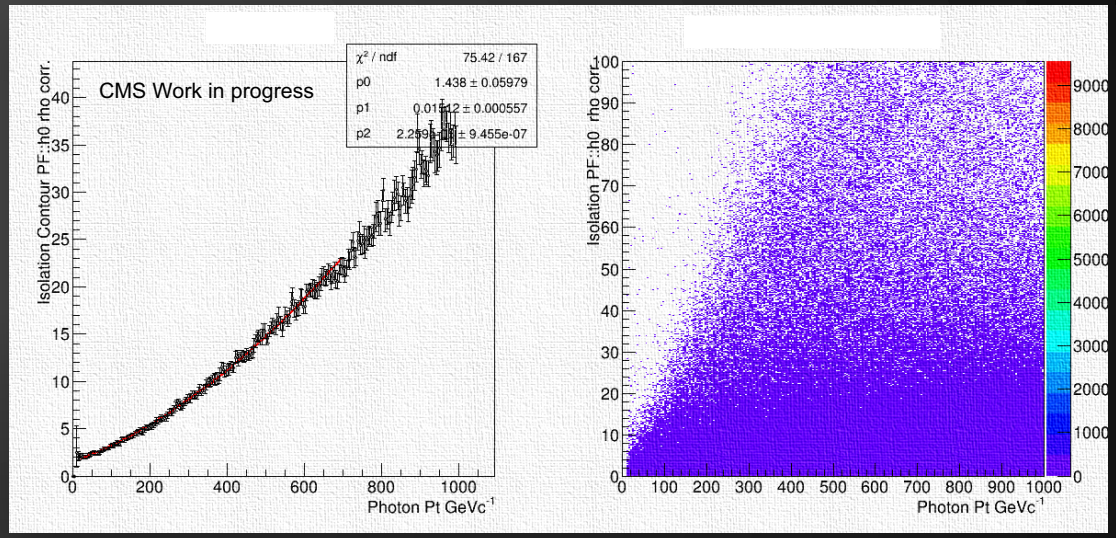
# Pt scaling of Isolation

After pile up correction we check if the pile up corrected isolations have any Pt dependence, and we notice they do have Pt dependency




**Barrel Photonl Isolation Pt scaling**

0.004017*Pt



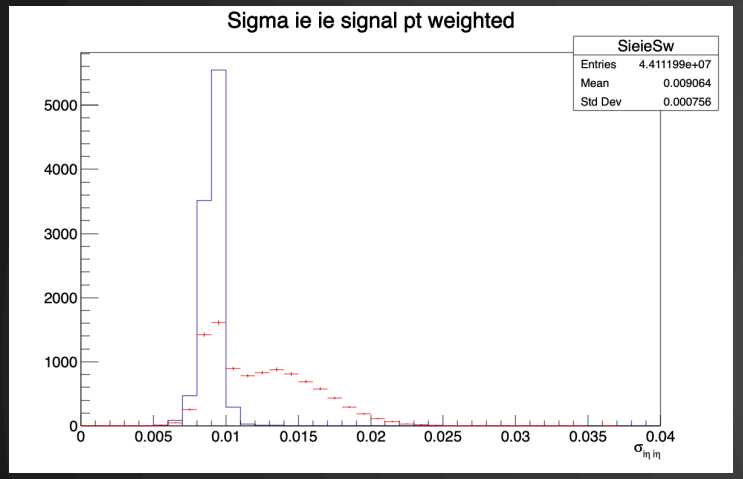

**Barrel Neutral Isolation Pt scaling**

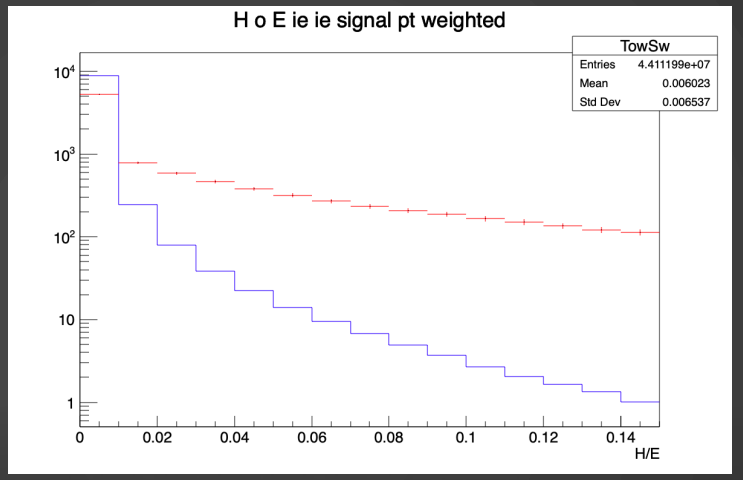
$0.01512*Pt+0.000023*Pt^2$

# Extraction of cut values

- In MC samples we do have generator level information of signal and background
- With the help of this information we feed the network corrected values of the variables for both signal and background events

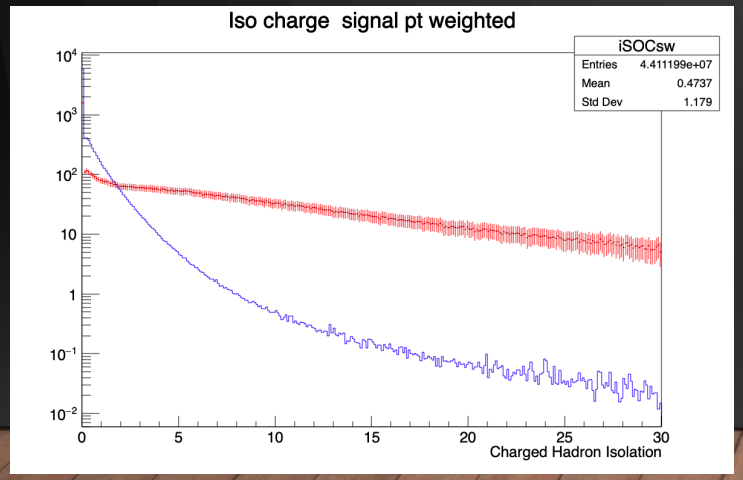
# Extraction of cut values

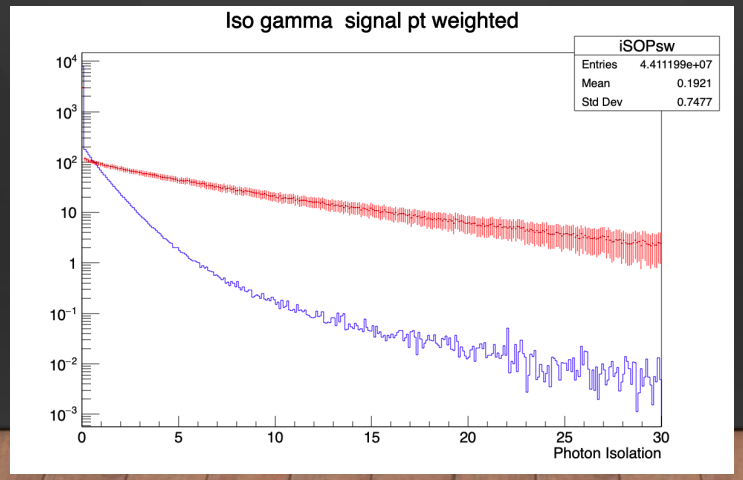


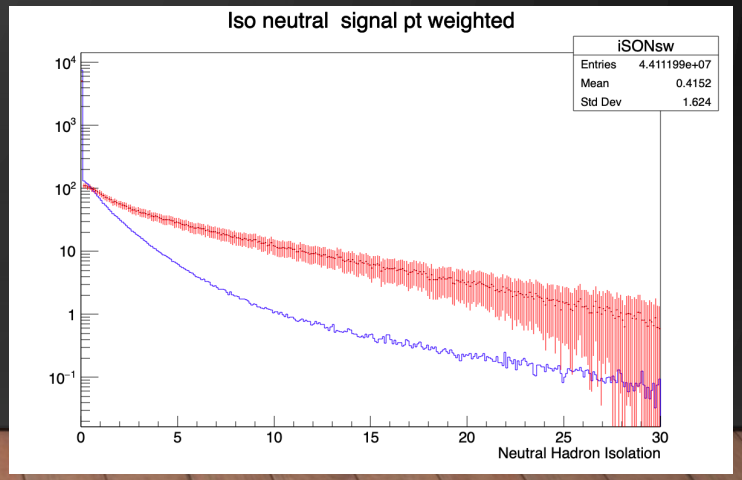




Signal

Background



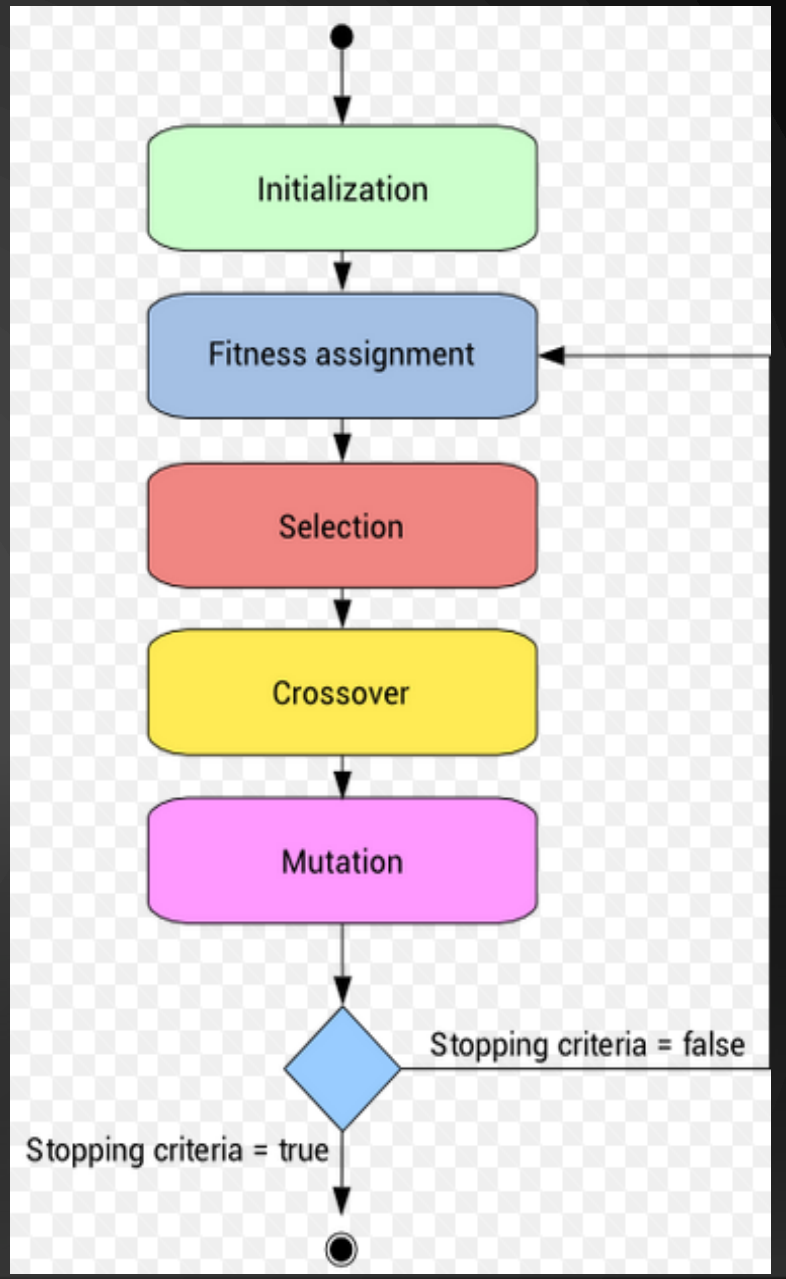



# Extraction of cut values

- In MC samples we do have generator level information of signal and background
- With the help of this information we feed the network corrected values of the variables for both signal and background events
- To train and test, genetic algorithm is used

# GENETIC ALGORITHM

> *It is not the strongest of the species that survives, nor the most intelligent , but the one most responsive to change.*



# INITIALISATION

- If there are different survival items, each having its own "Survival Points"
- Carry items(n) in a backpack of maximum weight of 30 kg
- Objective is to maximize the survival points.

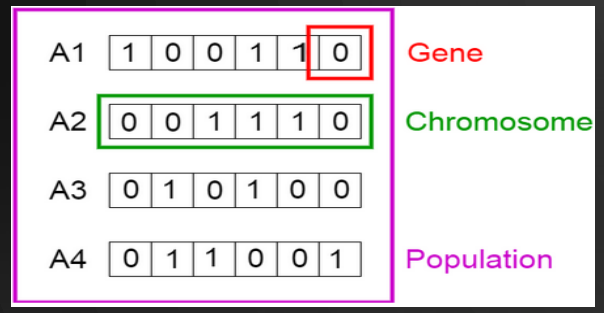
| ITEM | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| SLEEPING BAG | 15 | 15 |
| ROPE | 3 | 7 |
| POCKET KNIFE | 2 | 10 |
| TORCH | 5 | 5 |
| BOTTLE | 9 | 8 |
| GLUCOSE | 20 | 17 |

# INITIALISATION

- If there are different survival items, each having its own "Survival Points"
- Carry items(n) in a backpack of maximum weight of 30 kg
- Objective is to maximize the survival points.

| ITEM | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| SLEEPING BAG | 15 | 15 |
| ROPE | 3 | 7 |
| POCKET KNIFE | 2 | 10 |
| TORCH | 5 | 5 |
| BOTTLE | 9 | 8 |
| GLUCOSE | 20 | 17 |

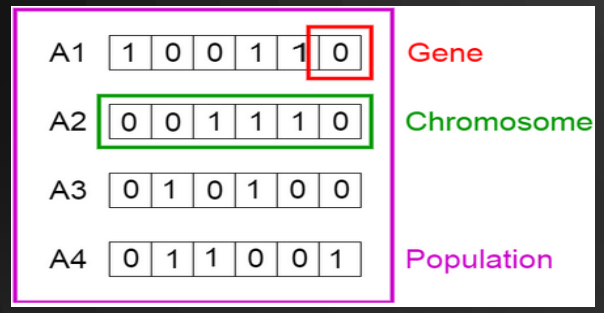
Check every possible solution to find out the right solution? Choice is $2^n$



# INITIALISATION

- If there are different survival items, each having its own "Survival Points"
- Carry items(n) in a backpack of maximum weight of 30 kg
- Objective is to maximize the survival points.

| ITEM | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| SLEEPING BAG | 15 | 15 |
| ROPE | 3 | 7 |
| POCKET KNIFE | 2 | 10 |
| TORCH | 5 | 5 |
| BOTTLE | 9 | 8 |
| GLUCOSE | 20 | 17 |




Check every possible solution to find out the right solution? Choice is $2^n$

Items ~ 5 variables, weight~signal efficiency, points~background rejection.

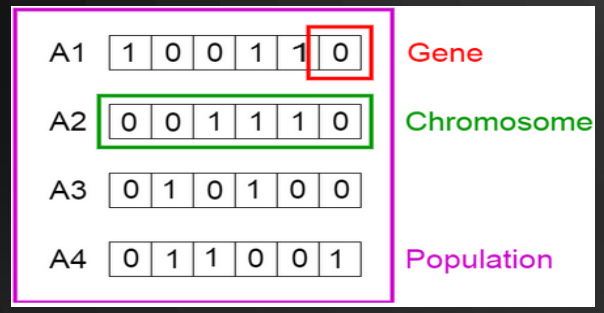
every variable can take at least 1000 values. Choice is at least $10^{15}$

Need to check over a large test sample to estimate background rejection

# INITIALISATION

- If there are different survival items, each having its own "Survival Points"
- Carry items(n) in a backpack of maximum weight of 30 kg
- Objective is to maximize the survival points.

| ITEM | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| SLEEPING BAG | 15 | 15 |
| ROPE | 3 | 7 |
| POCKET KNIFE | 2 | 10 |
| TORCH | 5 | 5 |
| BOTTLE | 9 | 8 |
| GLUCOSE | 20 | 17 |




Check every possible solution to find out the right solution? Choice is $2^n$

Items ~ 5 variables, weight~signal efficiency, points~background rejection.

every variable can take at least 1000 values. Choice is at least $10^{15}$

Need to check over a large test sample to estimate background rejection

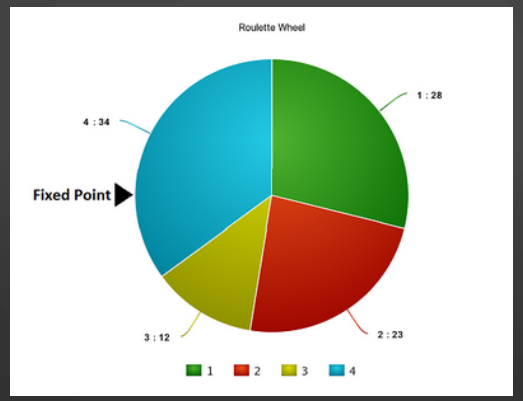
To initialize, forget right solution, make a population of solutions
For us it's values of training samples

This set of chromosome is considered as our initial population.

# SELECTION

| | Survival Points | Percentage |
|---|---|---|
| Chromosome 1 | 28 | 28.9% |
| Chromosome 2 | 23 | 23.7% |
| Chromosome 3 | 12 | 12.4% |
| Chromosome 4 | 34 | 35.1% |

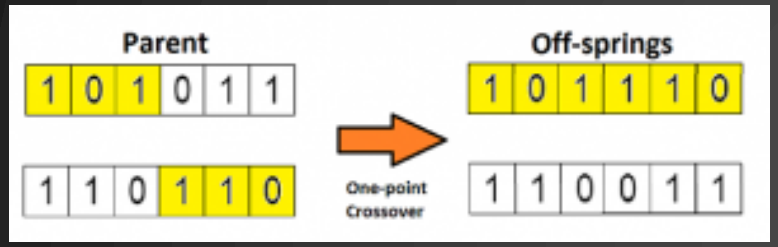
Based on these values, make a wheel to choose parents




Advantage is, the solutions which have less background rejection ability, will contribute less in the next generation. Thus, a faster way towards right solution.

# CROSSOVER & MUTATION

- In this previous step, parent chromosomes has been selected that will produce off-springs.




- One point crossover. A random crossover point is selected and the tails of both the chromosomes are swapped to produce a new off-springs.

- There is some change in the genes of children which makes them different from its parents.
- A random tweak in the chromosome, which also promotes the idea of diversity in the population.
- A simple method of mutation is

# STOPPING CONDITIONS

- There are different termination conditions, which are listed below:

1. There is no improvement in the population for over x iterations.
2. We have already predefined an absolute number of generation for our algorithm.
3. When our fitness function has reached a predefined value.

Main advantage is it estimates, given a signal efficiency, maximum background rejection.

# EXTRACTION OF CUT VALUES

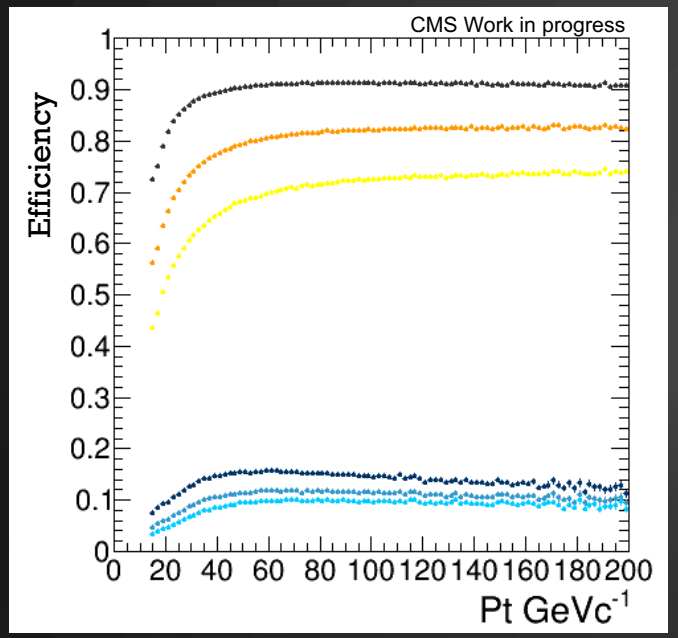
- After this training, the network gets ready to optimize the values of all the 5 variables for particular efficiencies of course
- That means we actually get the cut values from which given a unknown mixed sample of Signal and background, eventually which will be the case for data, we will be able to say for which (loose/medium/tight) cut we will have what amount of signal efficiency with how much background rejection

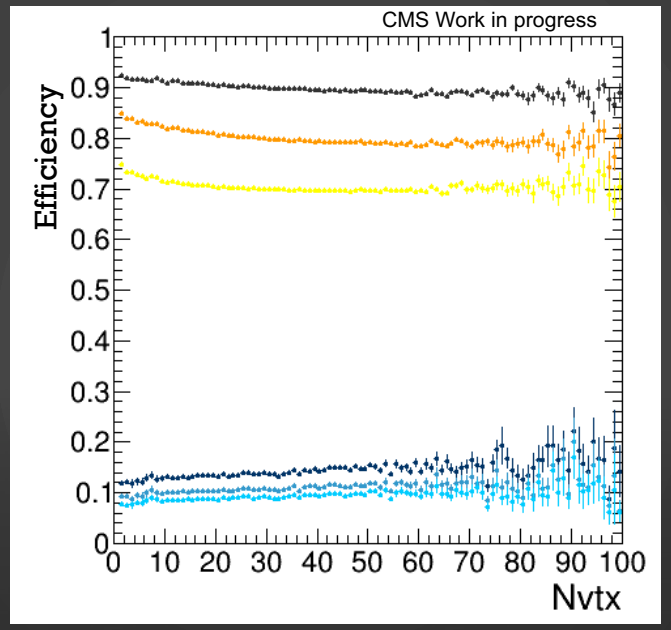
# Optimized ID

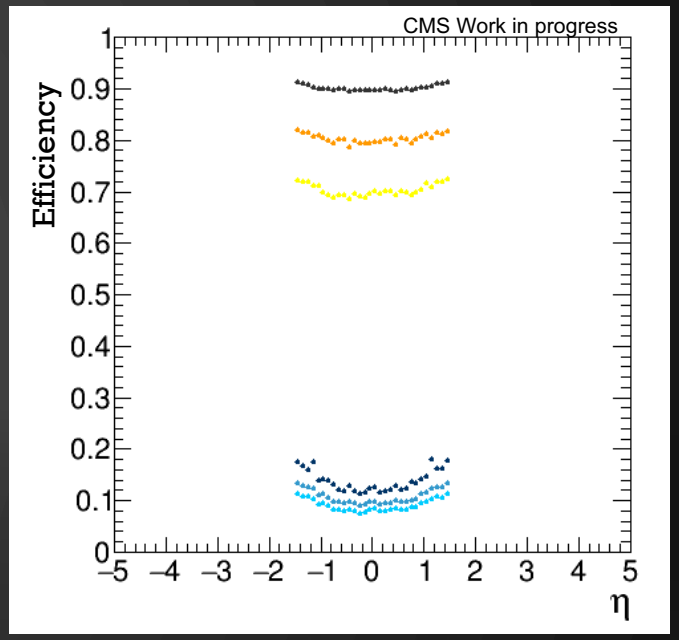
| BARREL | Loose (90.08%) | Medium (80.29%) | Tight (70.24%) |
|---|---|---|---|
| Background Rejection | Loose (86.25%) | Medium (89.36%) | Tight (90.97%) |
| HoverE | 0.04596 | 0.02197 | 0.02148 |
| $sigma_{ietaieta}$ | 0.0106 | 0.01015 | 0.00996 |
| Rho corrected PF charged hadron isolation | 1.694 | 1.141 | 0.65 |
| Rho corrected PF neutral hadron isolation | 24.032 + 0.01512*pho_pt + 2.259e-05*pho_pt^2 | 1.189 + 0.01512*pho_pt + 2.259e-05*pho_pt^2 | 0.317 + 0.01512*pho_pt + 2.259e-05*pho_pt^2 |
| Rho corrected PF photon isolation | 2.876 + 0.004017*pho_pt | 2.08 + 0.004017*pho_pt | 2.044 + 0.004017*pho_pt |

From loose to medium, if we sacrifice ~10% signal efficiency, background efficiency get reduced to ~21%

# Pt, Nvtx and η dependency of signal and background efficiencies




Signal efficiency — Background efficiency for 90% signal efficiency
Signal efficiency — Background efficiency for 80% signal efficiency
Signal efficiency — Background efficiency for 70% signal efficiency

# SUMMARY

- Pile up correction of isolations are studied
- Pt dependency of the isolations is studied
- Using genetic algorithm CMS official cut based photon ID for 2017 analyses is optimized
- Checked the pt, NVtx and eta dependency of the ID

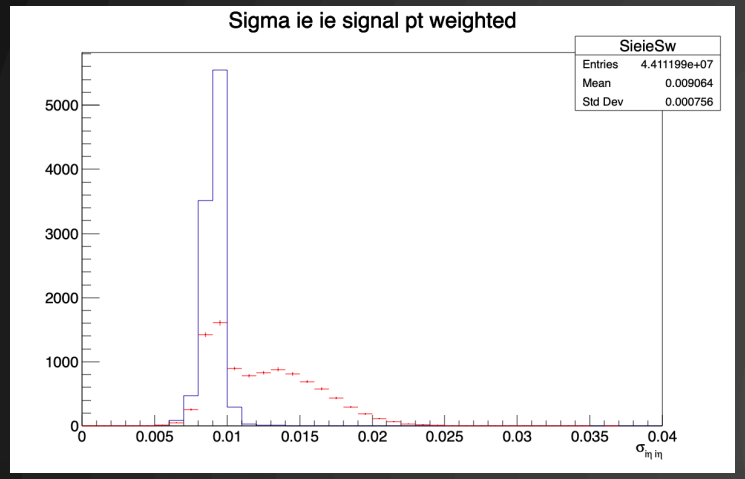
# SUMMARY

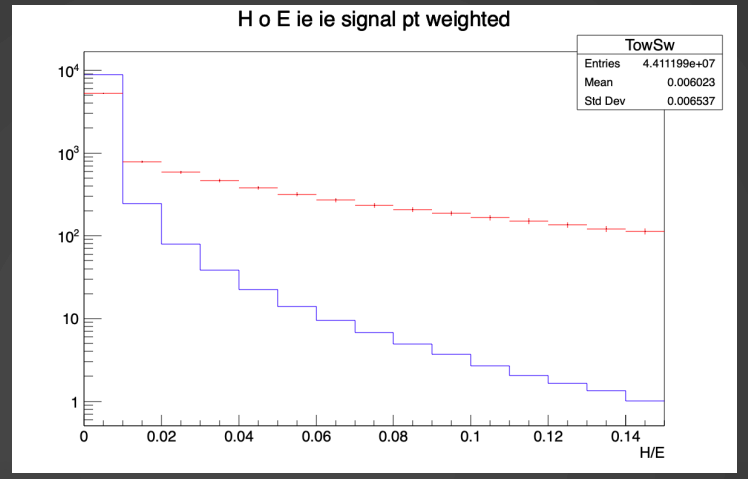
- Pile up correction of isolations are studied
- Pt dependency of the isolations is studied
- Using genetic algorithm Official cut based photon ID for 2017 analyses are optimized
- Checked the pt, NVtx and eta dependency of the ID

THANK YOU

# BACKUP

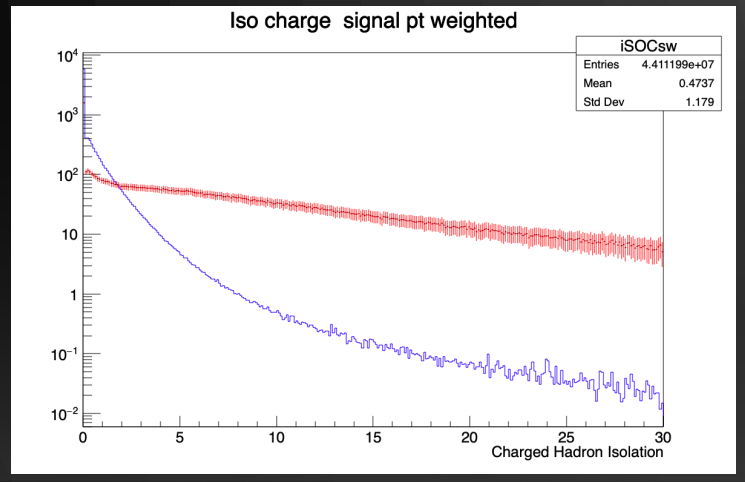
# 30

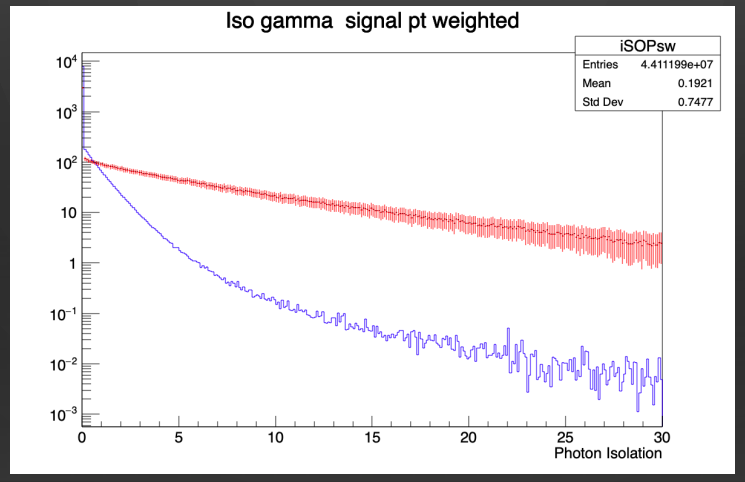


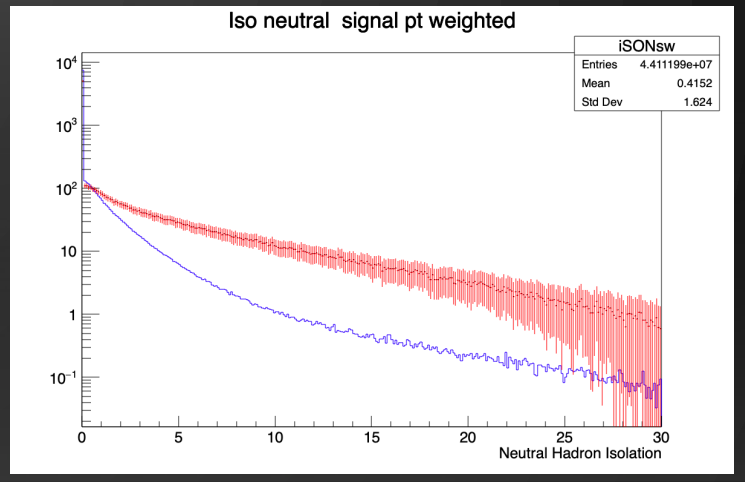




Signal

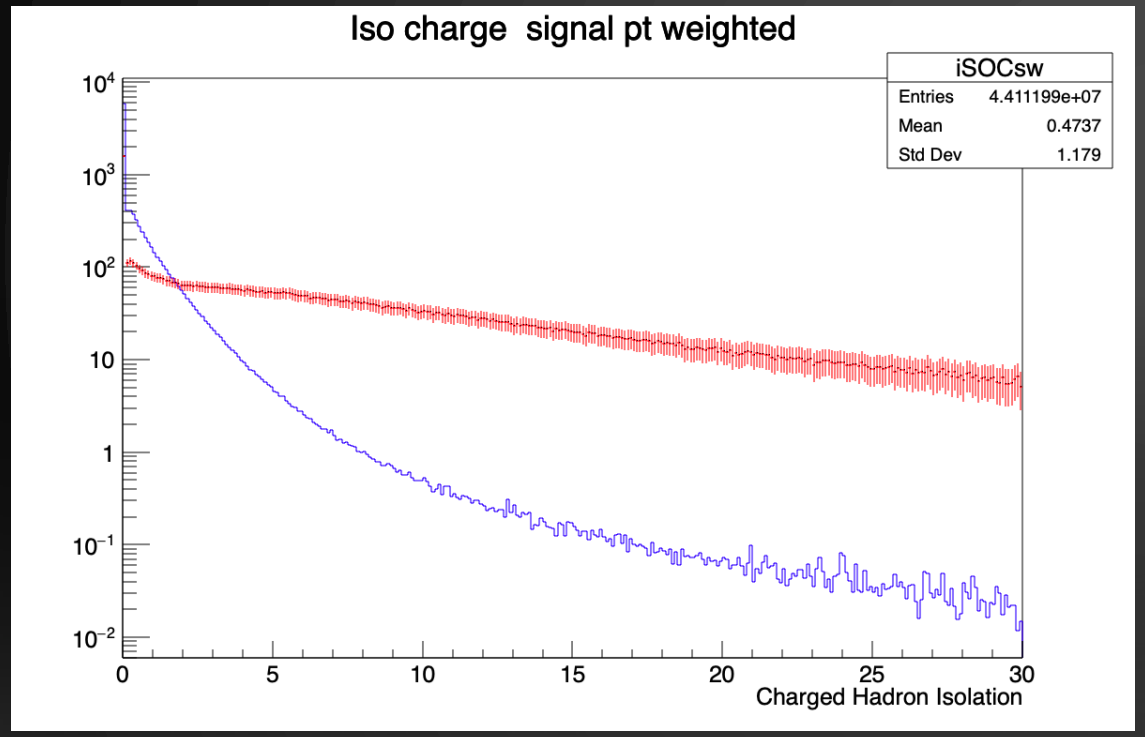
Background

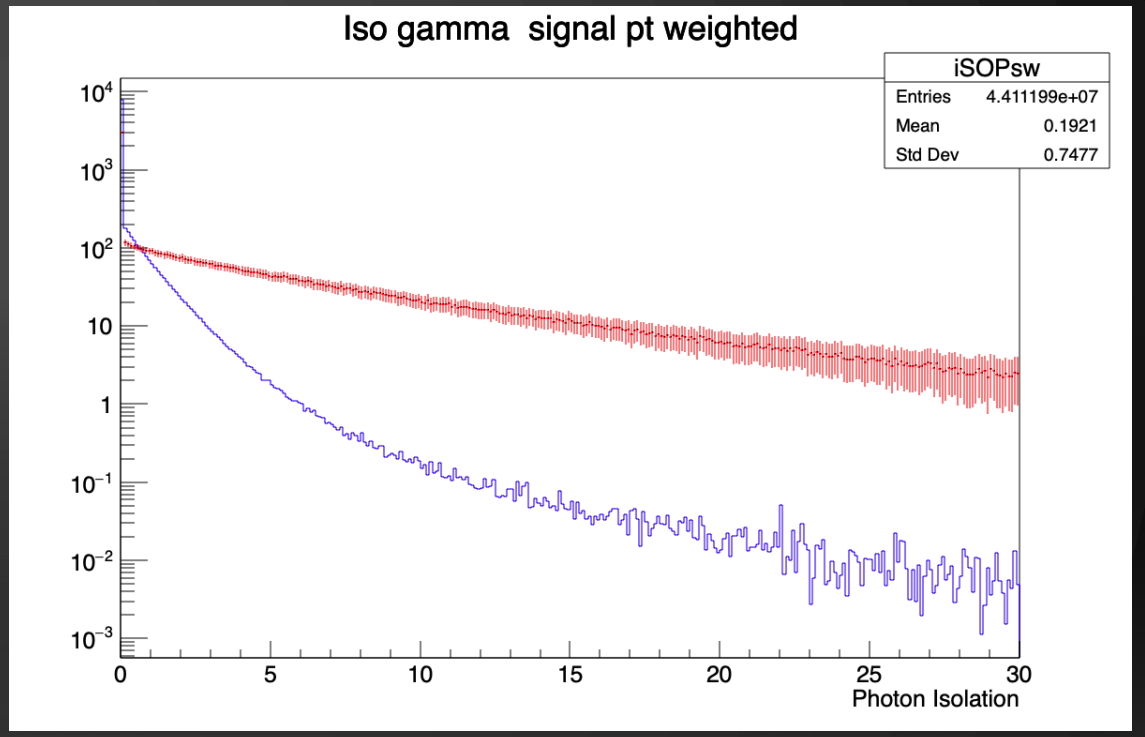


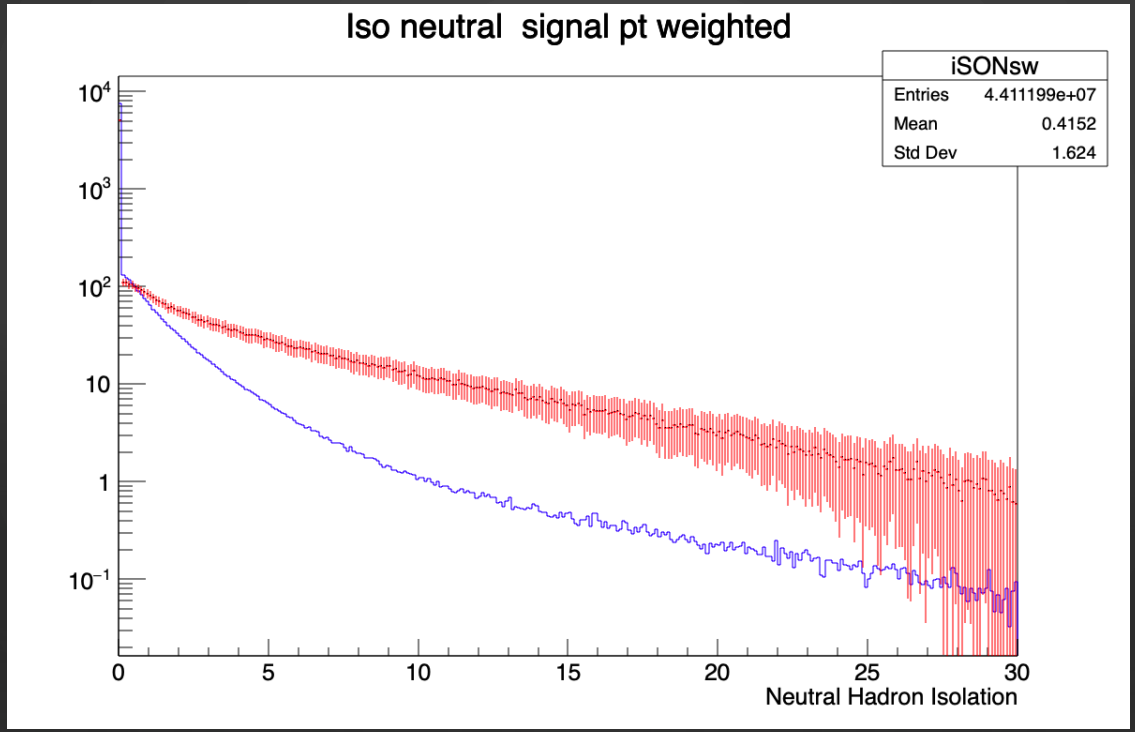



# 31


Iso charge signal pt weighted
iSOCsw
Entries 4.411199e+07
Mean 0.4737
Std Dev 1.179
Charged Hadron Isolation

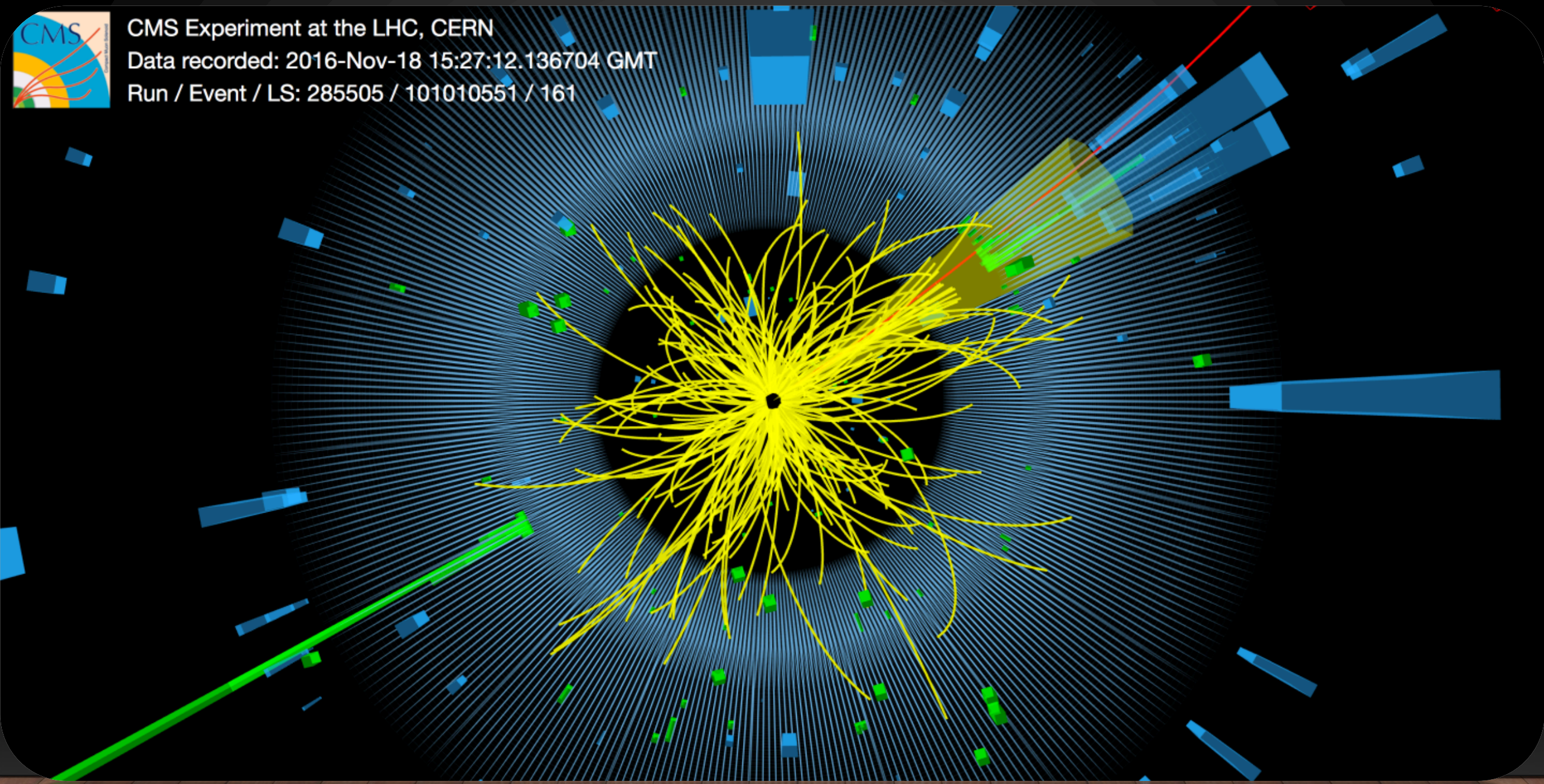

Iso gamma signal pt weighted
iSOPsw
Entries 4.411199e+07
Mean 0.1921
Std Dev 0.7477
Photon Isolation

# 32

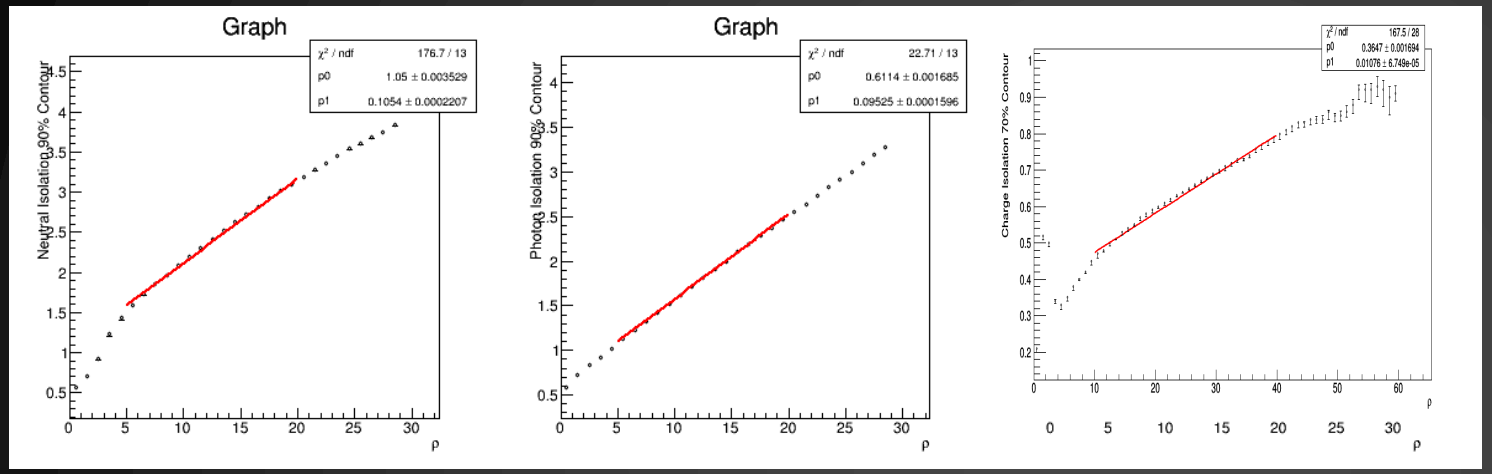

Iso neutral signal pt weighted
iSONsw
Entries 4.411199e+07
Mean 0.4152
Std Dev 1.624
$10^4$
$10^3$
$10^2$
10
1
$10^{-1}$
0
5
10
15
20
25
30
Neutral Hadron Isolation

# 33

$$\sigma_{i\eta i\eta} = \left( \frac{\sum (\eta_i - \bar{\eta})^2 \omega_i}{\sum \omega_i} \right)^{1/2}; \quad \bar{\eta} = \frac{\sum \eta_i \omega_i}{\sum \omega_i}; \quad \omega_i = \max\left(0, 4.7 + \log \frac{E_i}{E_{5\times5}}\right)$$


CMS Experiment at the LHC, CERN
Data recorded: 2016-Nov-18 15:27:12.136704 GMT
Run / Event / LS: 285505 / 101010551 / 161

# Isolation vs ρ plots and effective area




| bin | EA charged hadrons(70% cont) | EA neutral hadrons(90% cont) | EA photons(90% cont) |
|---|---|---|---|
| abs(η)<1.0 | 0.0112 | 0.0668 | 0.1113 |
| 1.0<abs(η)<1.479 | 0.0108 | 0.1054 | 0.0953 |
| 1.479<abs(η)<2.0 | 0.0106 | 0.0786 | 0.0619 |
| 2.0 <abs(η) < 2.2 | 0.01002 | 0.0233 | 0.0837 |
| 2.2 <abs(η) < 2.3 | 0.0098 | 0.0078 | 0.1070 |
| 2.3 <abs(η) < 2.4 | 0.0089 | 0.0028 | 0.1212 |
| abs(η) > 2.4 | 0.0087 | 0.0137 | 0.1466 |

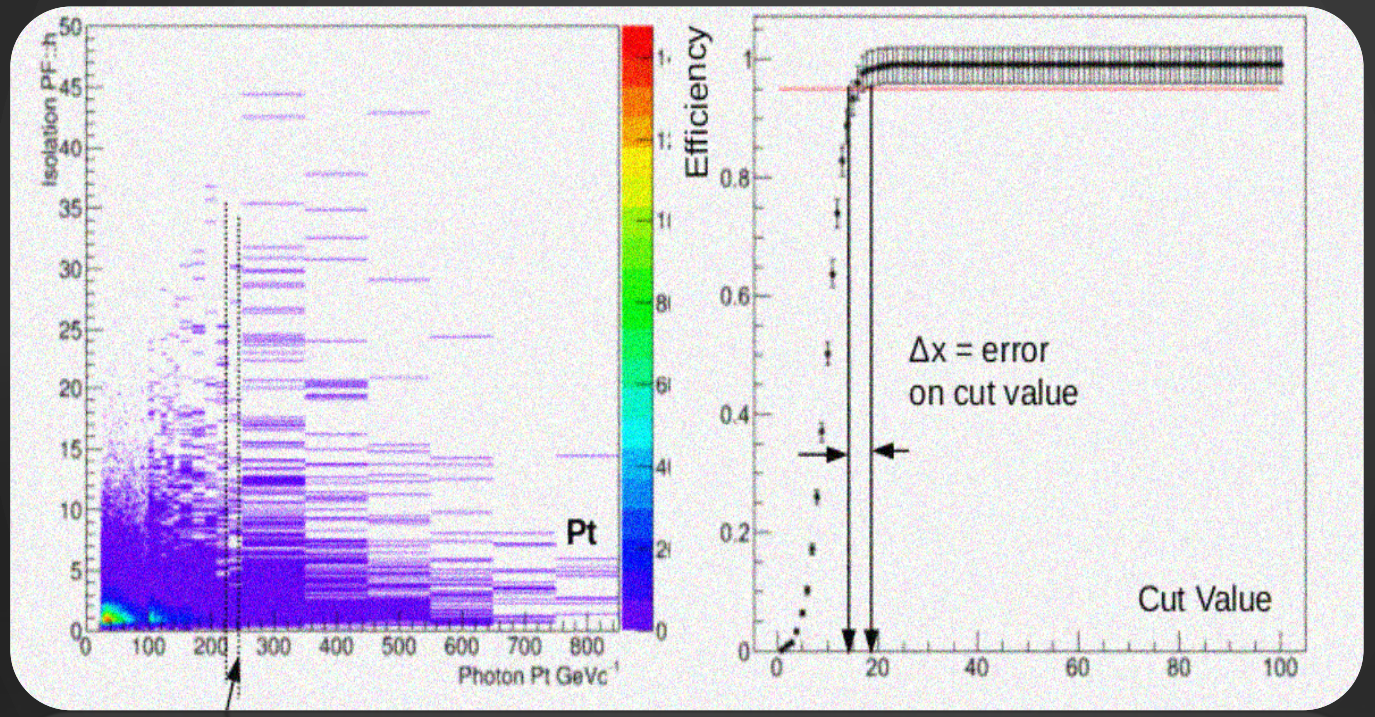
η<1.0

Slope is the Effective Area(EA)

$\text{Isolation} = \text{EA} \times \rho + \text{Isolation}_{\text{Corr}}$

$\text{Isolation}_{\text{Corr}} = \text{Isolation} - \text{EA} \times \rho$

# 36


Isolation PF
Photon Pt GeVc
Pt
Efficiency
Δx = error
on cut value
Cut Value

What has this quote got to do with genetic algorithm ?

Actually, the entire concept of a genetic algorithm is based on the above line.

A basic example:

Let's take a hypothetical situation where, in order to keep city safe from bad things, head of a country implements a policy like this.

- Select all the good people, and ask them to extend their generation by having their children.
- This repeats for a few generations.
- As a result there will be an entire population of good people.

The example above is to illustrate the concept and not completely realistic

The basic idea was that we changed the input (i.e. population) such that we get better output (i.e. better country).

# WHAT IS A GENETIC ALGORITHM ?

In the example discussed above :

1. Firstly, initial population, as countrymen, was defined.
2. A function to classify whether a person is good or bad needed to be defined.
3. Good people were selected to produce their off-springs.
4. And finally, these off-springs replace the bad people from the population and this process repeats.

This is how genetic algorithm actually works, which tries to mimic the human evolution to some extent.

To formalize a definition of a genetic algorithm :

**It is an optimization technique, which tries to find out such values of input so that we get the best output values or results.**

# STEPS INVOLVED IN GENETIC ALGORITHM

To make things easier, let us understand it by the famous Knapsack problem

- Let's say, you are going to spend a month in the wilderness.
- Only thing you are carrying is the backpack of maximum weight of **30 kg**.
- Now you have different survival items, each having its own "Survival Points"
- So, your objective is maximize the survival points.

| ITEM | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| SLEEPING BAG | 15 | 15 |
| ROPE | 3 | 7 |
| POCKET KNIFE | 2 | 10 |
| TORCH | 5 | 5 |
| BOTTLE | 9 | 8 |
| GLUCOSE | 20 | 17 |

# FITNESS FUNCTION

- Let us calculate fitness points for our first two chromosomes.

For A1 chromosome [100110]

| ITEMS | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| Sleeping bag | 15 | 15 |
| Torch | 5 | 5 |
| Bottle | 9 | 8 |
| TOTAL | 29 | 28 |

Similarly for A2 chromosome [001110]

| ITEMS | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| Pocket Knife | 2 | 10 |
| Torch | 5 | 5 |
| Bottle | 9 | 8 |
| TOTAL | 16 | 23 |

For this problem, our chromosome will be considered as more fit when it contains more survival points.

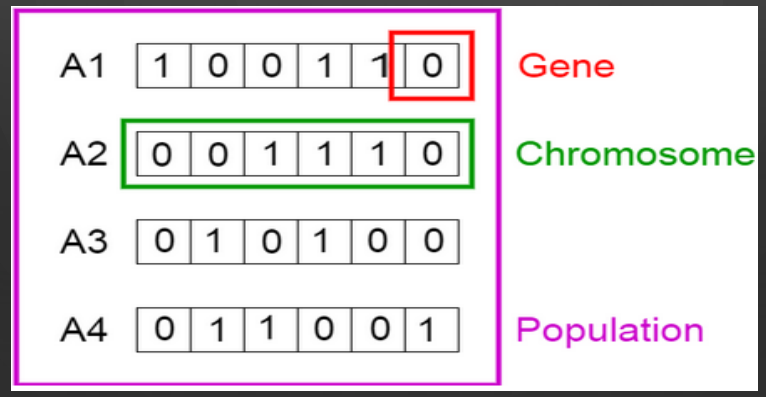
➢ Therefore chromosome 1 is more fit than chromosome 2.

# INITIALISATION



- To solve this problem using genetic algorithm, our first step is defining population.
- Population will contain individuals, each having their own set of chromosomes.

| ITEM | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| SLEEPING BAG | 15 | 15 |
| ROPE | 3 | 7 |
| POCKET KNIFE | 2 | 10 |
| TORCH | 5 | 5 |
| BOTTLE | 9 | 8 |
| GLUCOSE | 20 | 17 |




This set of chromosome is considered as our initial population.

# FITNESS FUNCTION



- Let us calculate fitness points for our first two chromosomes.

For A1 chromosome [100110]

| ITEMS | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| Sleeping bag | 15 | 15 |
| Torch | 5 | 5 |
| Bottle | 9 | 8 |
| TOTAL | 29 | 28 |

Similarly for A2 chromosome [001110]

| ITEMS | WEIGHT | SURVIVAL POINTS |
|---|---|---|
| Pocket Knife | 2 | 10 |
| Torch | 5 | 5 |
| Bottle | 9 | 8 |
| TOTAL | 16 | 23 |

For this problem, our chromosome will be considered as more fit when it contains more survival points.

➢ Therefore chromosome 1 is more fit than chromosome 2.

# SELECTION

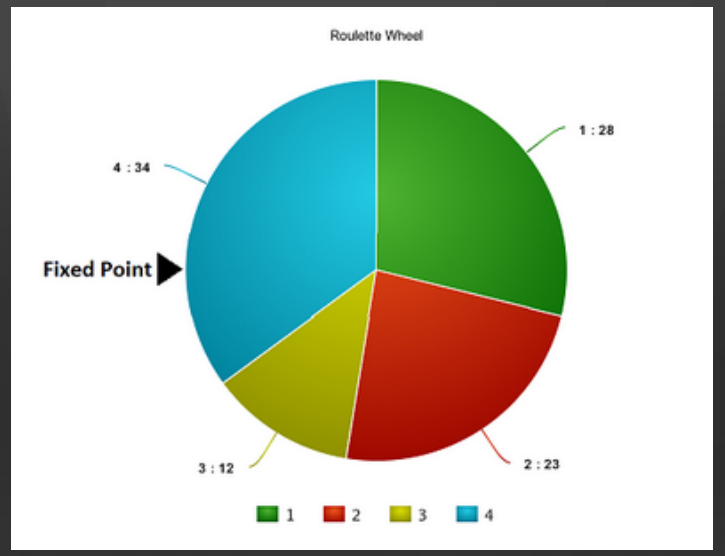


- General thought is that we should select the fit chromosomes and allow them to produce off-springs
- But that would lead to chromosomes that are more close to one another in a few next generation, and therefore less diversity
- For that reason we generally use Roulette Wheel Selection method

# 45

| | Survival Points | Percentage |
|---|---|---|
| Chromosome 1 | 28 | 28.9% |
| Chromosome 2 | 23 | 23.7% |
| Chromosome 3 | 12 | 12.4% |
| Chromosome 4 | 34 | 35.1% |

Based on these values, let us create our roulette wheel



# 46

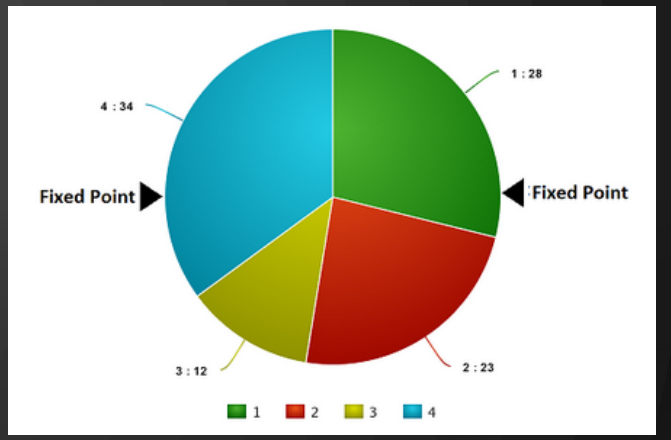
Now this wheel is rotated and the region of wheel which comes in front of the fixed point is chosen as the parent

For the second parent, the same process is repeated

Sometimes we mark two fixed points.

In this method we can get both our parents in one go

This method is known as Stochastic Universal Selection method

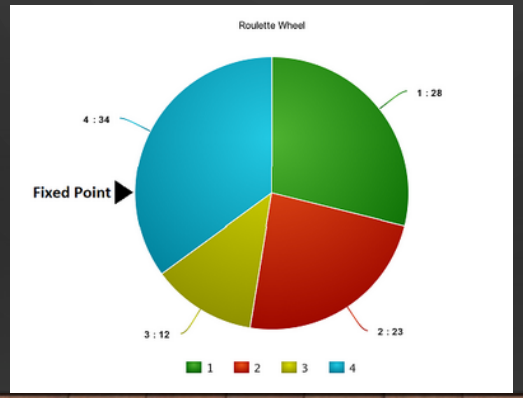


# SELECTION

- General thought is that we should select the fit chromosomes and allow them to produce off-springs
- But that would lead to chromosomes that are more close to one another in a few next generation, and therefore less diversity

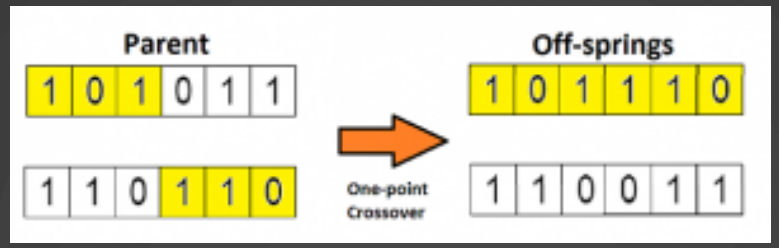
| | Survival Points | Percentage |
|---|---|---|
| Chromosome 1 | 28 | 28.9% |
| Chromosome 2 | 23 | 23.7% |
| Chromosome 3 | 12 | 12.4% |
| Chromosome 4 | 34 | 35.1% |

Based on these values, make a wheel to choose parents

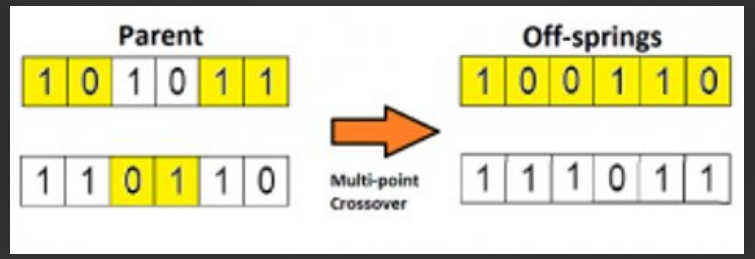


# CROSSOVER

- In this previous step, parent chromosomes has been selected that will produce off-springs.




- This is the most basic form of crossover, known as one point crossover. Here a random crossover point is selected and the tails of both the chromosomes are swapped to produce a new off-springs.
- If you take two crossover point, then it will called as multi point crossover which is as shown below.



# MUTATION

- There is some change in the genes of children which makes them different from its parents.
- This may be defined as a random tweak in the chromosome, which also promotes the idea of diversity in the population.
- A simple method of mutation is

# Effective Area in η bins

| bin | EA charged hadrons(70% cont) | EA neutral hadrons(90% cont) | EA photons(90% cont) |
|---|---|---|---|
| abs(η)<1.0 | 0.0112 | 0.0668 | 0.1113 |
| 1.0<abs(η)<1.479 | 0.0108 | 0.1054 | 0.0953 |
| 1.479<abs(η)<2.0 | 0.0106 | 0.0786 | 0.0619 |
| 2.0 <abs(η) < 2.2 | 0.01002 | 0.0233 | 0.0837 |
| 2.2 <abs(η) < 2.3 | 0.0098 | 0.0078 | 0.1070 |
| 2.3 <abs(η) < 2.4 | 0.0089 | 0.0028 | 0.1212 |
| abs(η) > 2.4 | 0.0087 | 0.0137 | 0.1466 |

$$\text{Isolation} = \text{EA} \times \rho + \text{Isolation}_{\text{Corr}}$$

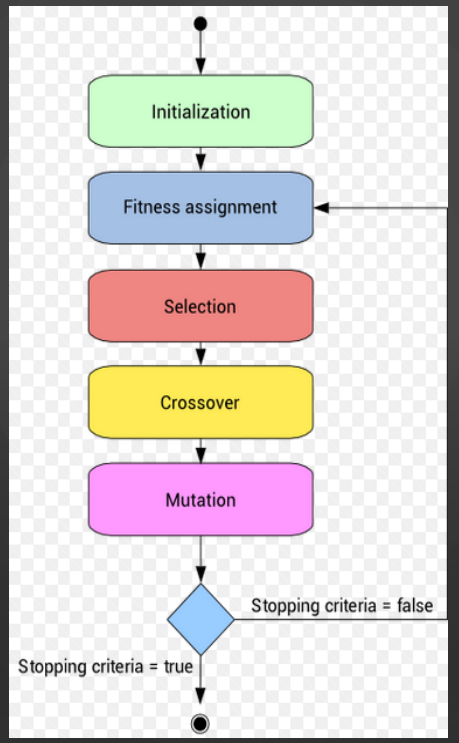
$$\text{Isolation}_{\text{Corr}} = \text{Isolation} - \text{EA} \times \rho$$

# AIM

- Main goal is to set some criteria on basis of which we can identify ptompt photons from the background ones.
- Five variables is chosen : shower shape variable $\sigma_{i\eta i\eta}$, H/E and three isolations (Photon, charged and neutral hadron)
- Optimize the cut values of these variables to use as the discriminator of prompt and background photons so that given a desired signal efficiency we get maximum background rejection
- For isolations we need corrections before using

# STEPS INVOLVED IN GENETIC ALGORITHM

The working of a genetic algorithm is also derived from biology, which is as shown in the image below.




Main advantage is it estimates, given a signal efficiency, maximum background rejection.