Studio di stati esotici charmonium-like e di algoritmi di tracking in CMS con il supporto di GPU

Attività di ricerca del I anno di Dottorato in Fisica XXXI ciclo

Dottorando

Adriano Di Florio

Tutore

Dott. Alexis Pompili

Outline

- **GPU computing applications within studies of charmonium-like exotic states**
  - **Statistical significance estimation by MC toys with *GooFit***
  - **Amplitude analysis fit of $B^0 \to J/\psi K^+ \pi^-$ with *GooFit***
  - **Inclusive search of the production of the Y(4140) state**
- **New tracking algorithm on heterogeneous computing systems**
- **Schools, courses, conferences and workshops**

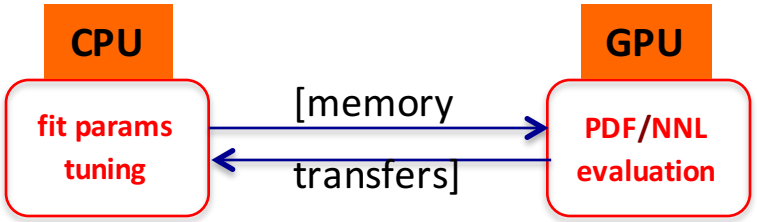
Statistical significance estimation by MC toys with *GooFit*

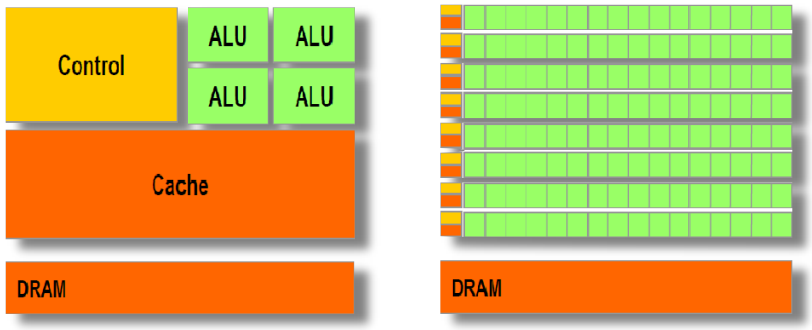
GooFit & GPUs

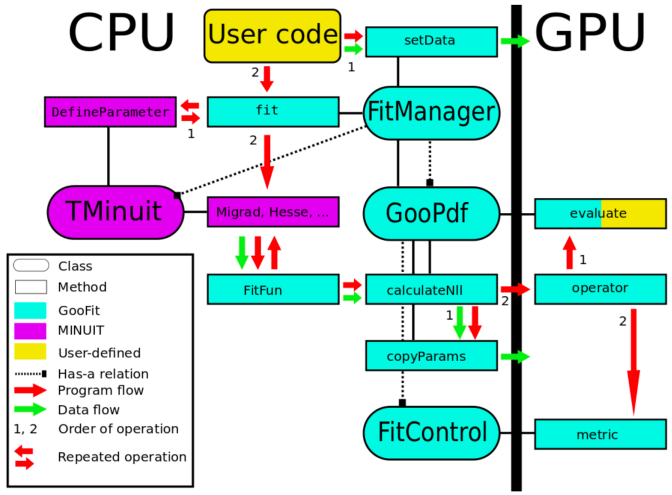
Hetherogeneous GPU-acccelerated computing is the use of a **Graphics Processing Unit** to accelerate scientific applications (among other apps). Features of a GPU architecture:

- **Thousands** of cores
- **Big loads of data**
- **Low frequency** clock (~1GHz)
- **Arithmetical operations in a single clock cycle** (*sin,cos,sqrt,1/x, …*)

ROOT and **RooFit** are two of the most used analysis tools in HEP. **GooFit** is a tool that allows to analyse massive ammounts of data: it acts as interface bewteen MINUIT on CPU and a GPU. It allows a p.d.f. to be evaluated in parallel exploiting thousands of GPU cores.





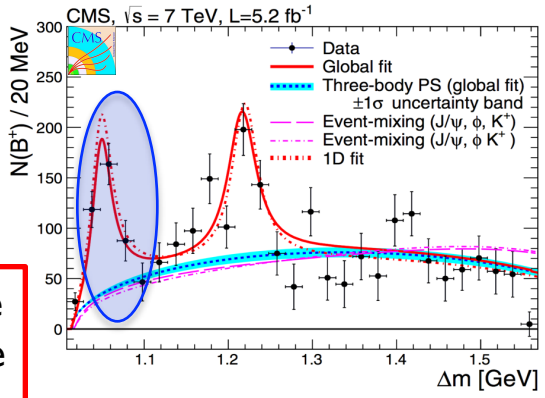


GooFit: a library for massively parallelising maximum-likelihood fits
R.Andreassen et al., *J.Phys.:Conf.Ser.* 513 (2014) 052003

GooFit : LEE & scanning technique

The *previous* test case

A **high-statistics toy Monte Carlo technique** has been implemented both in ***ROOT/RooFit*** and ***GooFit*** frameworks to estimate the (local) statistical significance of an "expected" signal [CMS, PLB 734 (2014) 261]. Results presented at ***ACAT 2016*** and ***CHEP 2016.***



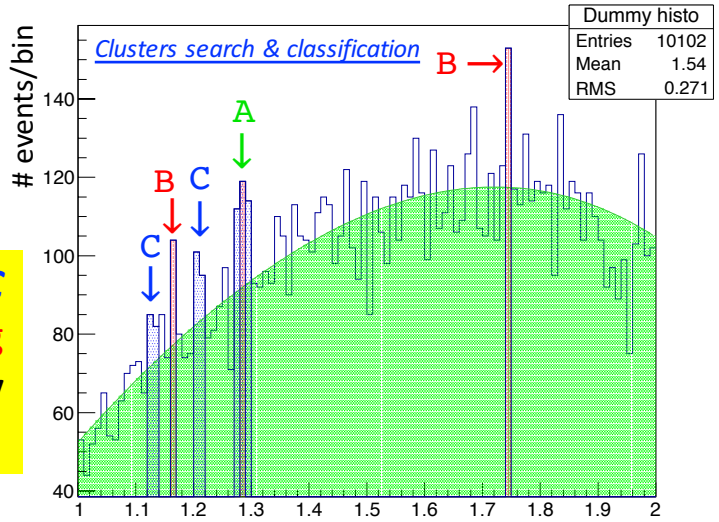
Ongoing step

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking fluctuations with respect to the background model everywhere in the overall mass spectrum.

The scanning technique is configured on the basis of a clustering approach and has been designed with the aim to satisfy two concurrent requirements:

1) Do **not** miss any interesting fluctuation

2) Do **not** select too many marginal fluctuations



Recently we have configured the procedure on the new ***Recas HPC Cluster***. We are running the application with different clustering configurations to be able to estimate the **systematic uncertainty** on the p-value estimation related to the method itself.

Documentation: CMS Analysis Note AN-2015-334

Amplitude analysis fit of $B^0 \to J/\psi K^+ \pi^-$ with *GooFit*

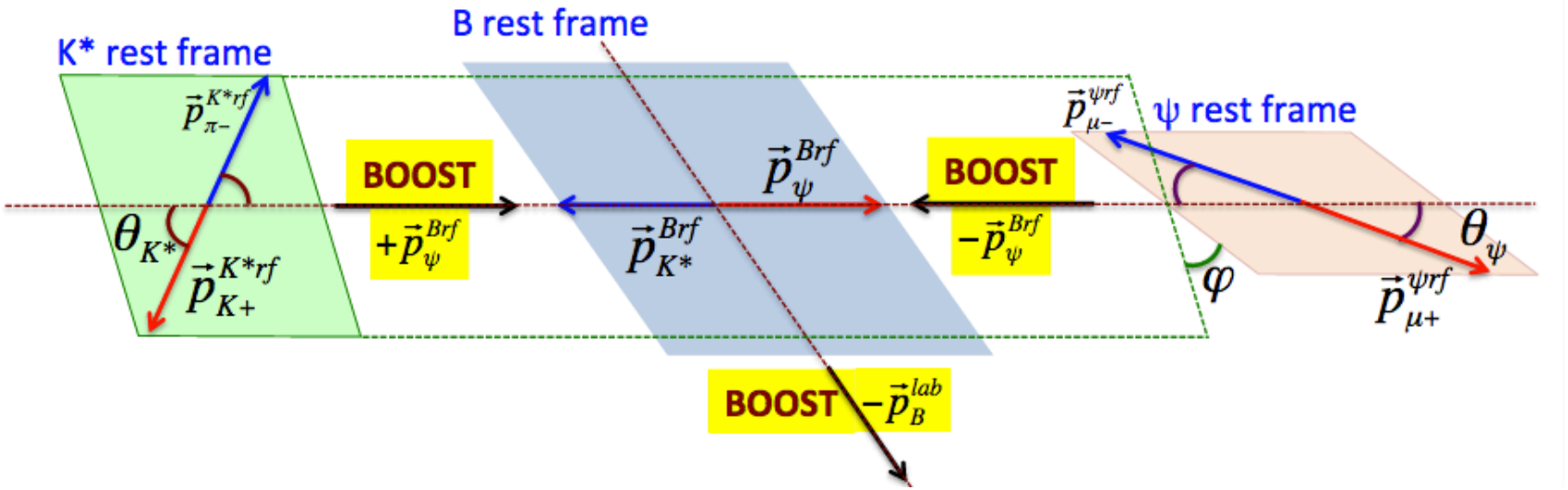
GooFit : amplitude analysis

Traditional Dalitz Plot analyses deal with 3-body decays into pseudoscalars. In that case the decay amplitudes are calculated on <u>2D</u> parameter space, namely the *Dalitz Plot* space itself. In 3-body decays with vectors in the final state the decay amplitude is calculated on an ***n-dimensional*** parameter space within the helicity formalism.

Decay modes needing an amplitude analysis under study in CMS (tetraquark search)

- $B^0 \to J/\psi K^+ \pi^-$ to search for Z(4430), $Z_c(4240)$ and $Z_c(3900)$
- $B^+ \to J/\psi \phi K^+$ to search for Y(4140) and other structures in $J/\psi\phi$ systems

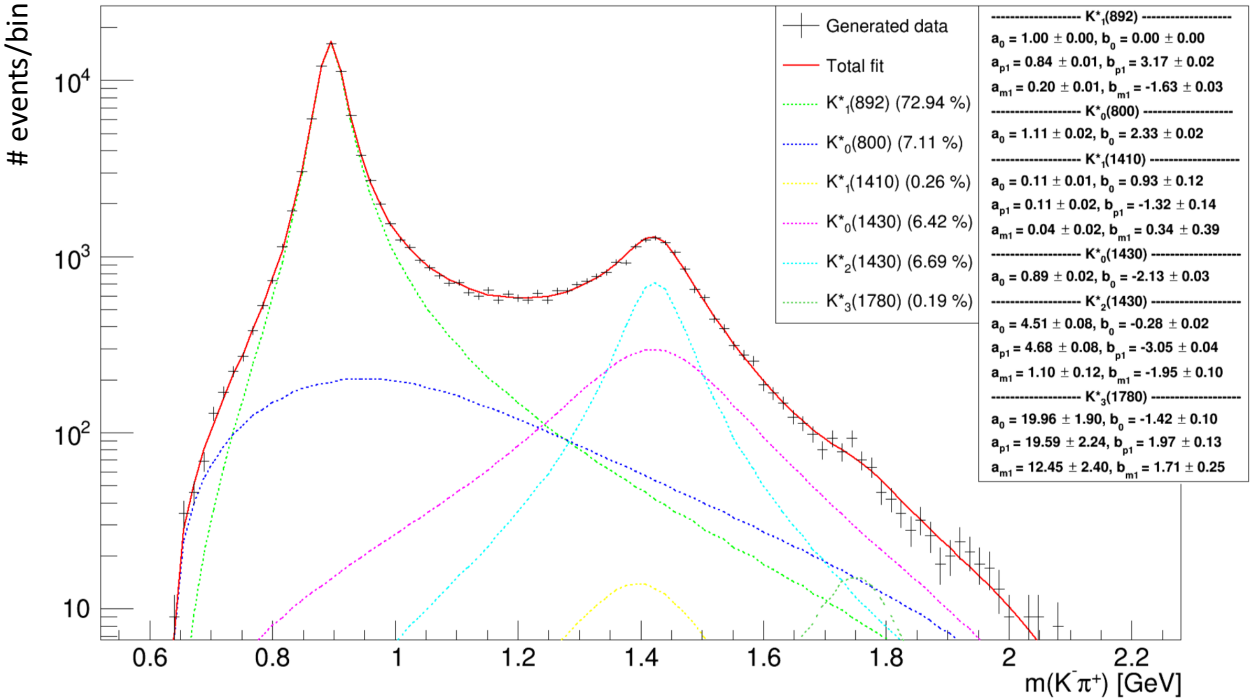
The parameter space is 4D: $\Phi = (m^2_{K\pi}, m^2_{\psi\pi}, \vartheta_\psi, \varphi_{\psi K^*})$



GooFit : amplitude analysis

Porting of A.A fit for $B^0 \to J/\psi K^+\pi^-$ from *RooFit* to *GooFit*

Considering only the most relevant intermediate K* resonances [$B^0 \to J/\psi\, K^*$] contributing with **28 fit parameters** (1 absolute value & 1 phase for each helicity amplitude; one amplitude for each spin-0 K*, three for each spin>0 K*). Since RooFit requires **very long fitting times** (many hours), the whole fitting code is being ported on **GooFit** to run it **on GPUs**.



First result: fit timing very promising, the GooFit fit takes 10' performing over 1k MIGRAD calls (RooFit needs 1h to fit 4 parameters)!

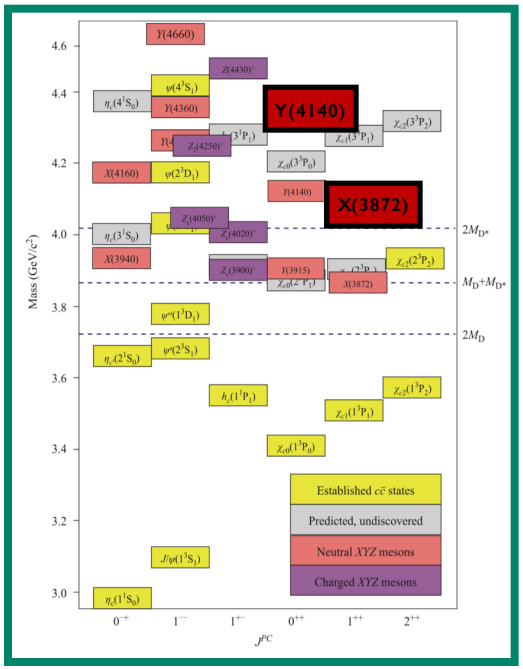
Inclusive search of the production of the Y(4140) state

Inclusive search for Y(4140)

In the last 13 years about 30 states **(known as *X, Y, Z* states)** observed while decaying to charmonium inspite of being **above the open-charm thresholds**. They show "exotic" characteristics.

Two main hypotheses:

- hadronic molecule
- tetraquark



Confirmed (2014) two structures seen by CDF (2011) in the $J/\psi\phi$ mass system, with a 1D analysis of $B^+ \to J/\psi\phi K^+$.

Prompt and non prompt production of Y(4140) state in $\bar{p}p$ collisions [$\bar{p}p \to Y(4140) + X$ with $Y \longrightarrow J/\psi\phi$] by studying inclusively the $J/\psi\phi$ mass spectrum (2015).

First amplitude analysis (2016) of the $B^+ \to J/\psi\phi K^+$ observing 4 structures in the $J/\psi\phi$ mass system; among them the Y(4140) structure, the closest to the kinematical threshold, which is slightly *better described as a* $D_sD_s^*$ *cusp* (resonant interpretation also possible)

It becomes crucial to confirm the D0 result for the inclusive search. In case of a positive confirmation **this would rule out a cusp interpretation of the Y(4140).**

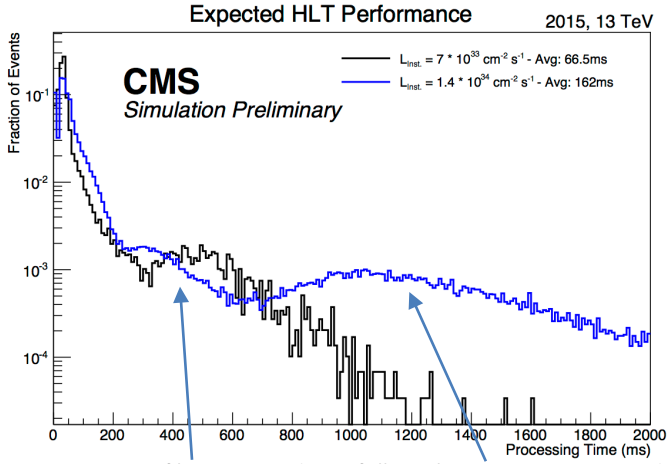
By now 8 TeV Run I data being explored.
Plan to use 13 TeV Run II data.

New tracking algorithm on heterogeneous computing systems

Future Tracking on GPU

Tracking at HLT. **Status**, **issues** and **perspectives**:

- The online farm consists of ~20k CPU Xeon cores
- A single event per logical core
- At the moment tracks are not reconstructed for all the events at the HLT
- This will be even more difficult at higher pile-up
- GPUs are becoming wider, cheaper and better supported.



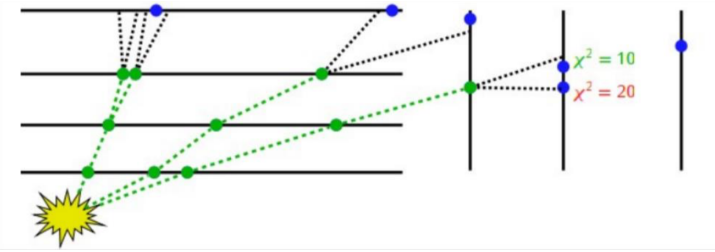
reconstruction of leptons and calorimetric object e.g. **muons**, photons

full track reconstruction and particle flow e.g. **jets**, tau

Future-proof solution to this issues: **rethinking** of tracking algorithms **in parallel** (not to be run necessarily on a GPU!).

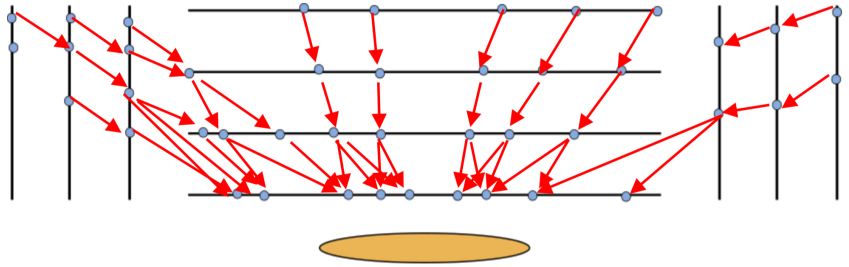
Track creation as it is : triplets propagation

Propagate 1-2-3 triplet to 4th layer and search for compatible hits



Cellular Automaton

Create hit pairs from pairs of adjacent layers. Join compatible pairs that share hits checking their compatibility. Calculations are simple, and localized in memory, straightforward to parallelize efficiently



Physics & timing performance

Events with PU50 do not exploit the full computational capabilities of the GPU

- Only **2-5%** of the GPU busy
- **~100MB** GPU DRAM used per event (compared to 10s of GB available)

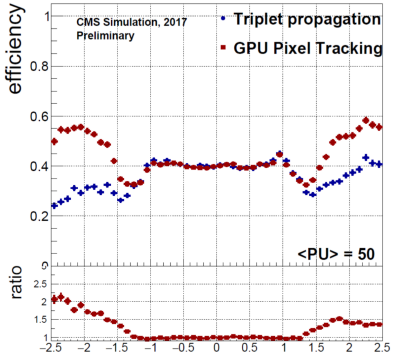
Hardware setup

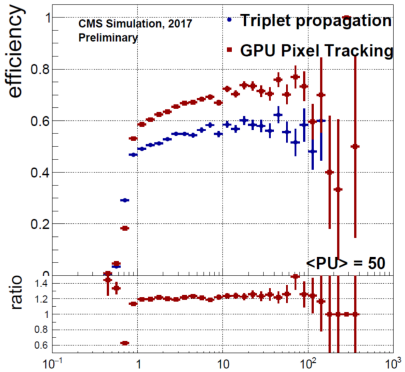
CPU Intel 4771K

GPU NVIDIA K40

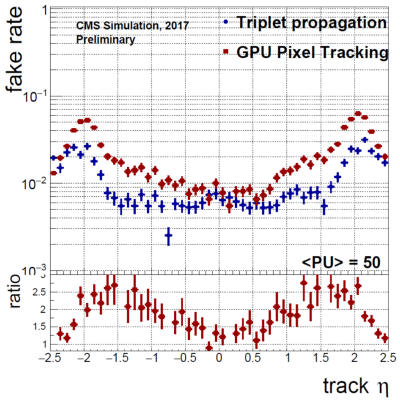
Physiscs performance

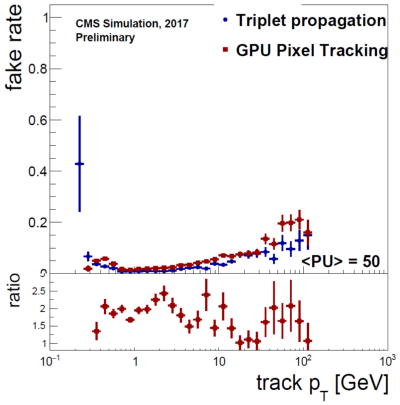
Tracking Efficiency





Tracks fake rate



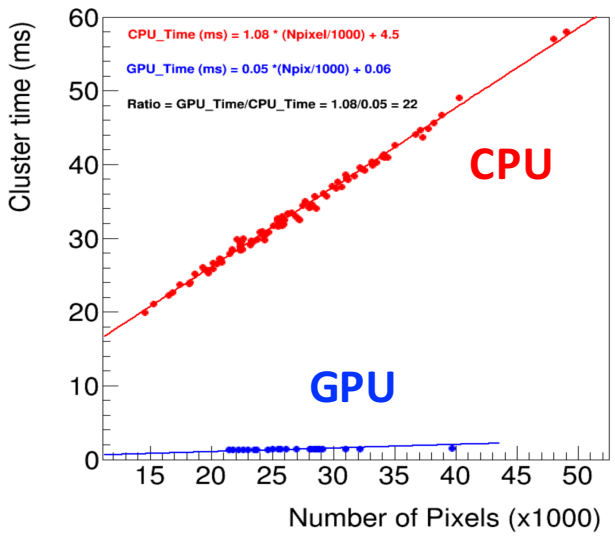


Timing performance

	time per event CPU (ms)	time per event GPU (ms)
Triplet propagation	66.3	N/A
CA	22	1.6 (15.2)

CPU triplet propagation

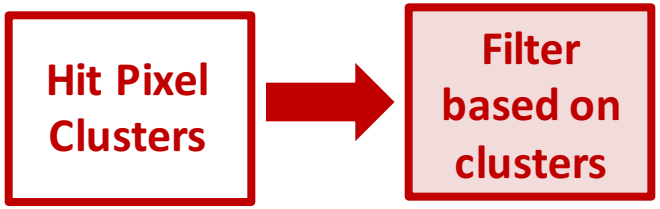
GPU Cellular automaton

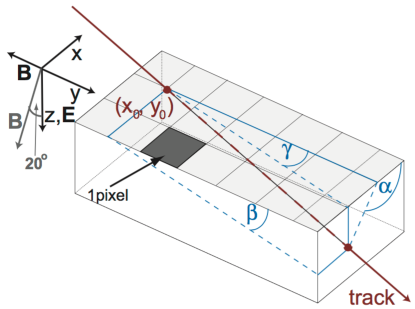


Clustering, filtering & ML

Doublets construction based mainly on geometry

***BUT** we can get some further information form the RECO Hits*





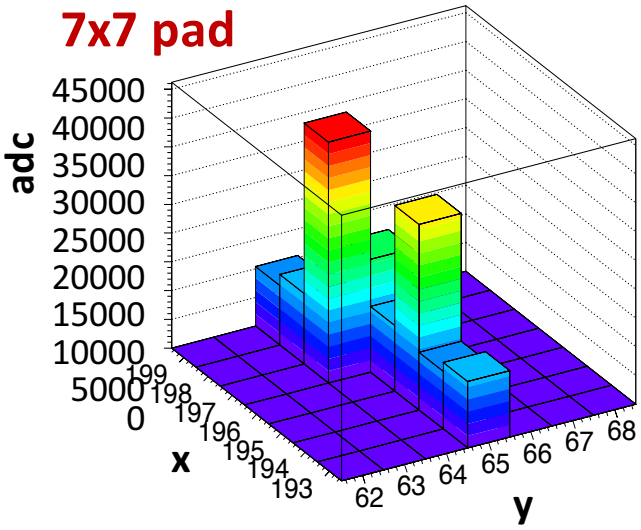
How is a Pixel Cluster represented in the CMSSW?

`class SiPixelCluster` "collection" of `Pixel`

A matrix whose indices correspond to *position* and elements to *adc* values (**only** for pixels turned on).

Work in progress . . .

Use Machine Learning & Image Recognition techniques to add an additional filtering for doublets based on clusters shapes.



Matching Tracks

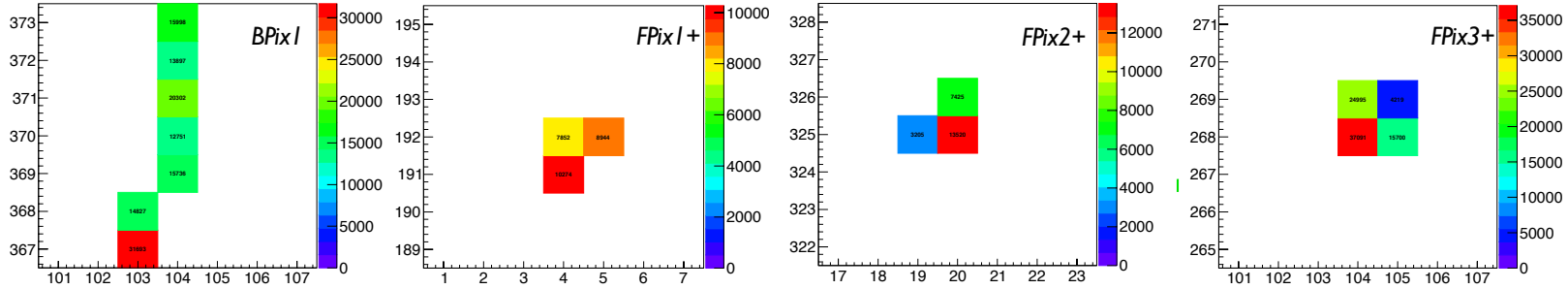
Example – RECO Tracks

RECO – SIM matching

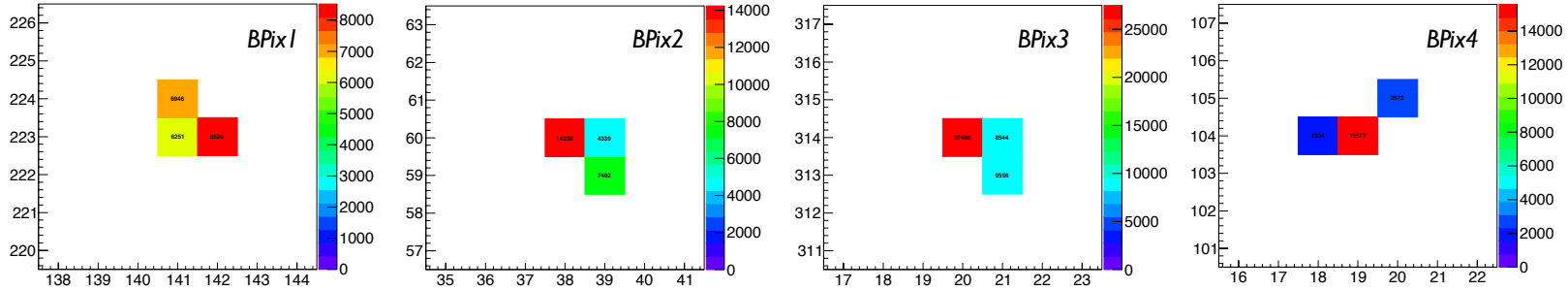
(MultiTrackValidator.cc)

proton (id: 2212)

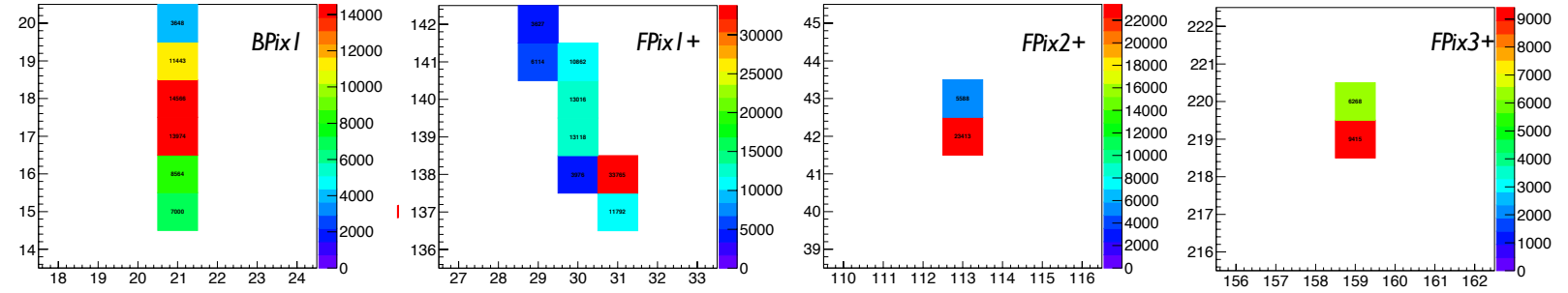
TTbar_13+TTbar_13TeV



electron (id: 11)



fake track



III – Schools, courses, conferences and workshops

Schools, courses, conferences and workshops

Workshops & Conferences

- 22nd International Conference on **C**omputing in **H**igh **E**nergy and **N**uclear Physics, CHEP 2016 (San Francisco, 10-14 October 2016). Poster presentation title: *"Performance studies of GooFit on GPUs versus RooFit on CPUs while estimating the statistical significance of a new physical signal"*,
- CCR (Commisione Calcolo e Reti) meeting (Rome, 5 July 2016) : "GPUs for Statistical Data Analysis in HEP: a performance study of GooFit on GPUs vs RooFit on CPUs "
- CMS Physics Week (CERN, 8-12 February 2016)
- WLCG Workshop (San Francisco, 8-9 October 2016)

National and International Schools

- "Programming graphic boards with CUDA, an intensive course"(Bari, 11-13 May 2016)
- CERN School of Computing 2016 (Mol, Belgio, SKC-CEN, 28 Agosto - 9 Settembre 2016)
  Final exam passed - recognized 6 ETFS CREDITS
- XXVIII Seminario Nazionale di Fisica Nucleare e Subnucleare "Francesco Romano"
  (Otranto, 3-10 October 2016)

Conference Proceedings

- GPUs for statistical data analysis in HEP: a performance study of GooFit on GPUs vs. RooFit on CPUs", *17th International workshop on Advanced Computing and Analysis Techniques* (ACAT 2016)
- Performance studies of GooFit on GPUs versus RooFit on CPUs while estimating the statistical significance of a new physical signal", *22nd International Conference on Computing in High Energy and Nuclear Physics* (CHEP 2016) [proceeding in preparation]

Schools, courses, conferences and workshops

Attended PHD School courses

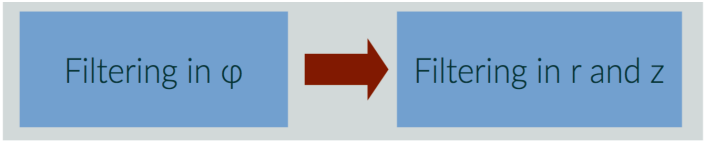
Exam date planned – Exam submitted – Exam passed

"Programming with Python"	2 CFU
"Gaseous Detectors"	2 CFU
"Management and knowledge of European research model and promotion of research results"	2 CFU
"How to prepare a technical speech in English"	2 CFU
"Statistical and computational model of data analysis"	2 CFU
"Standard model and beyond"	2 CFU
"Introduction and advanced C++ programming"	2 CFU
"Complex Systems"	2 CFU

Backup

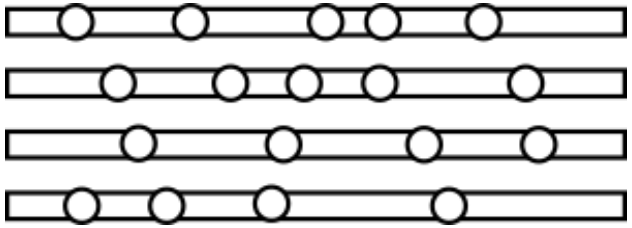
Tracking workflow

Cell/Doublet Generation



1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

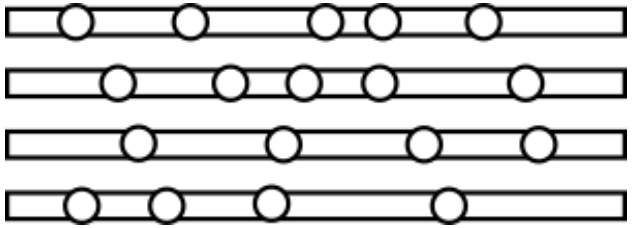
Cellular Automaton



Tracking workflow

Cell/Doublet Generation

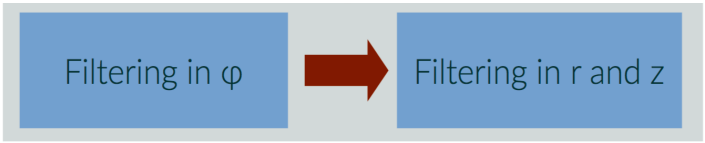
Cellular Automaton



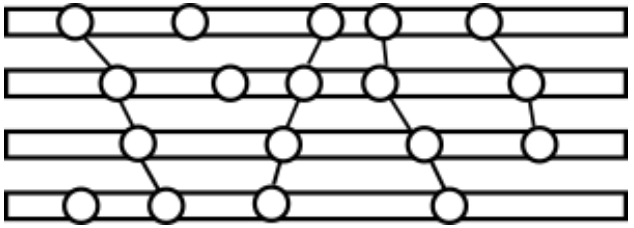
Each Hit makes connection with a compatible Hit on the next layer: Cell

Tracking workflow

Cell/Doublet Generation



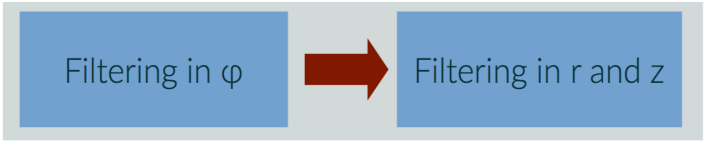
Cellular Automaton



Each Hit makes connection with a compatible Hit on the next layer: Cell

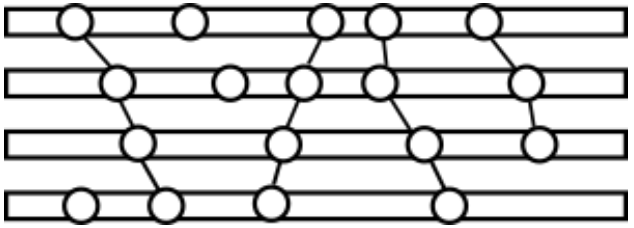
Tracking workflow

Cell/Doublet Generation



1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

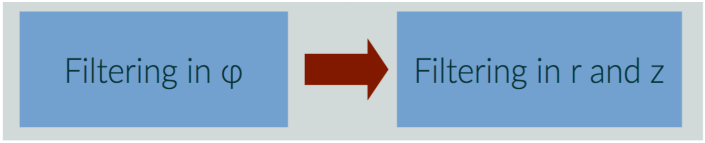
Cellular Automaton



Each Cell color is set to black or level to 0

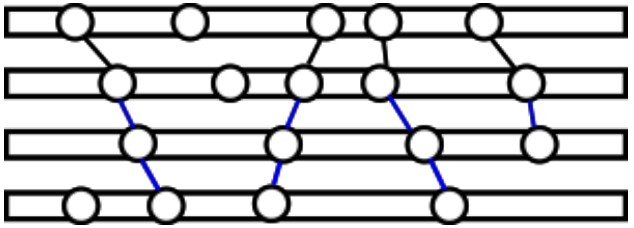
Tracking workflow

Cell/Doublet Generation



1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

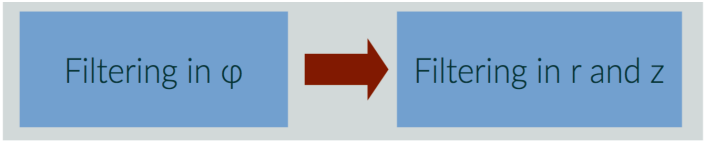
Cellular Automaton



If a Cell is connected to an upper cell with the same color, its color becomes lighter, or level higher.

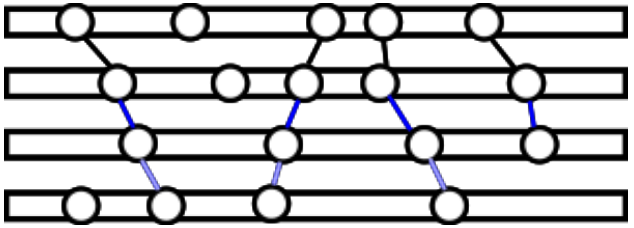
Tracking workflow

Cell/Doublet Generation



1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

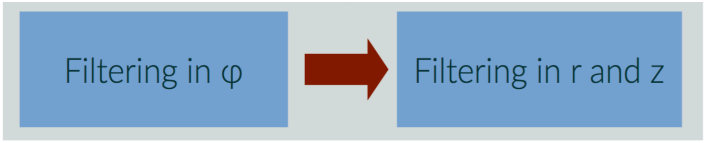
Cellular Automaton



A Track will be constructed traversing the cells from lighter to darker colors

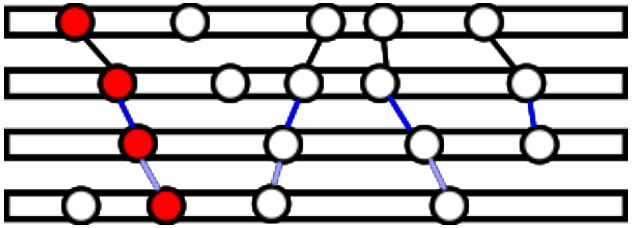
Tracking workflow

Cell/Doublet Generation



1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

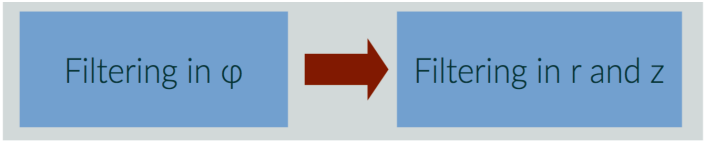
Cellular Automaton



A Track will be constructed traversing the cells from lighter to darker colors

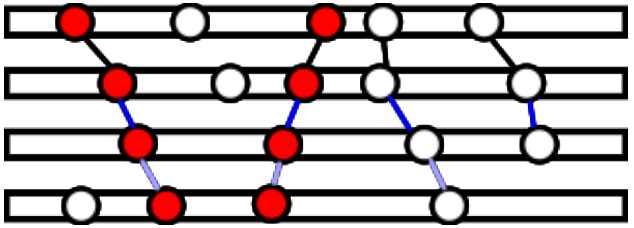
Tracking workflow

Cell/Doublet Generation



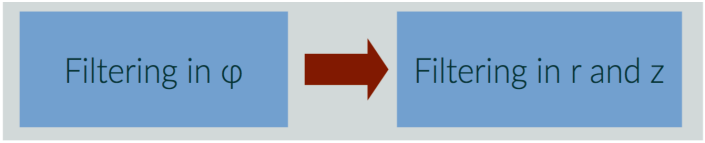
1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

Cellular Automaton



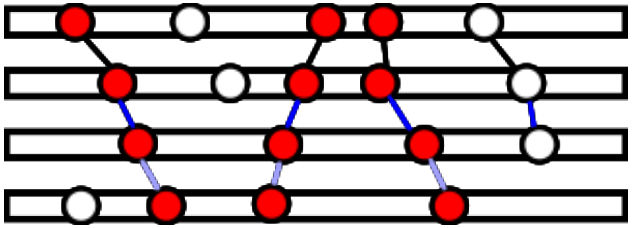
Tracking workflow

Cell/Doublet Generation



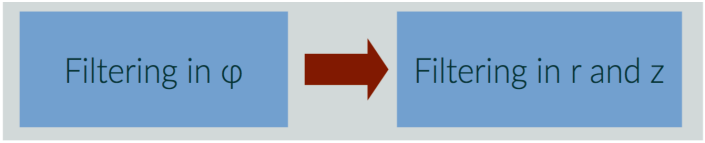
1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

Cellular Automaton



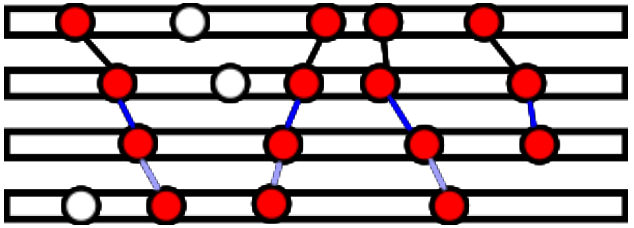
Tracking workflow

Cell/Doublet Generation



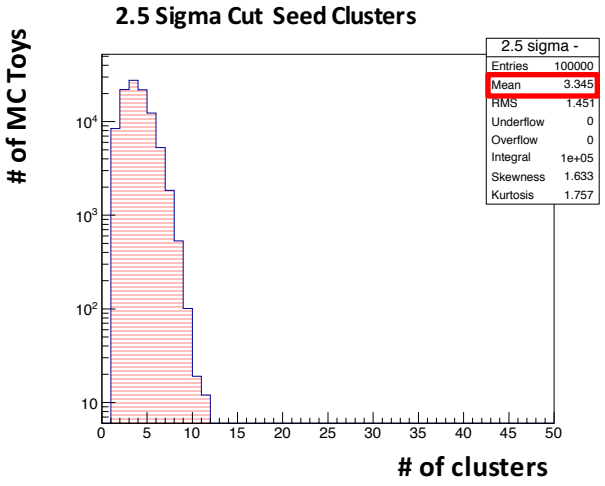
1. A **pair** of consecutive pixel layers is considered
2. The hits on the inner layer are sorted in phi
3. A hit in the outer layer is considered and a ϕ, r and z compa bility range is determined
4. Exploiting the sort, hits on the inner layer are filtered in ϕ
5. **r** and **z** selections are applied to each hit passing the ϕ lter
6. Remaining hits are used to form **doublets\cells** with the hit on the outer layer

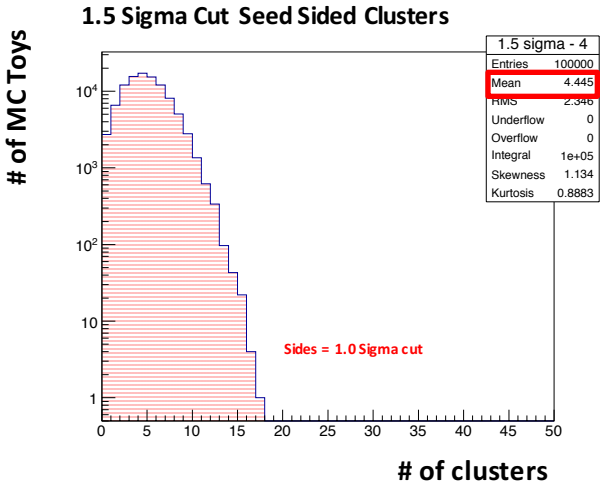
Cellular Automaton

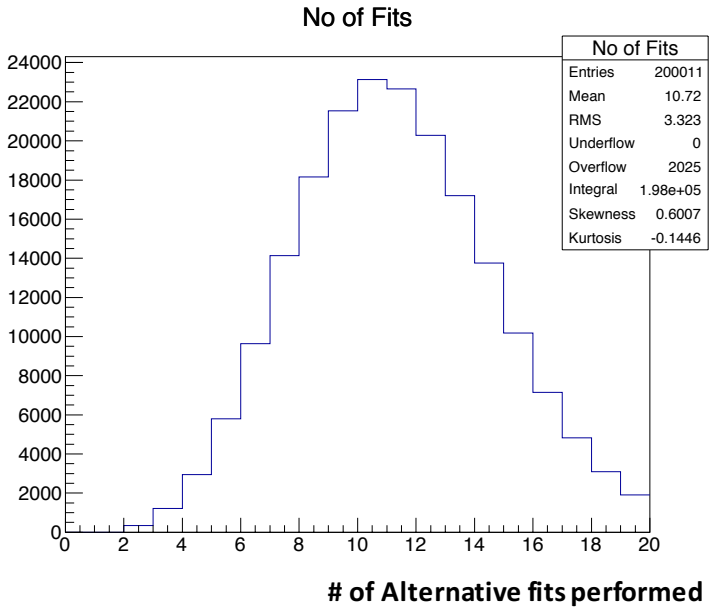


A Track will be constructed traversing the cells from lighter to darker colors

GooFit & clustering







Clustering, filtering & ML

Clustering, filtering & ML

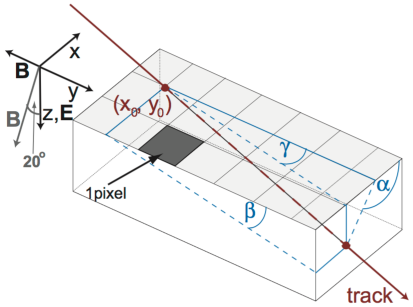
Doublets construction based mainly on geometry

Clustering, filtering & ML

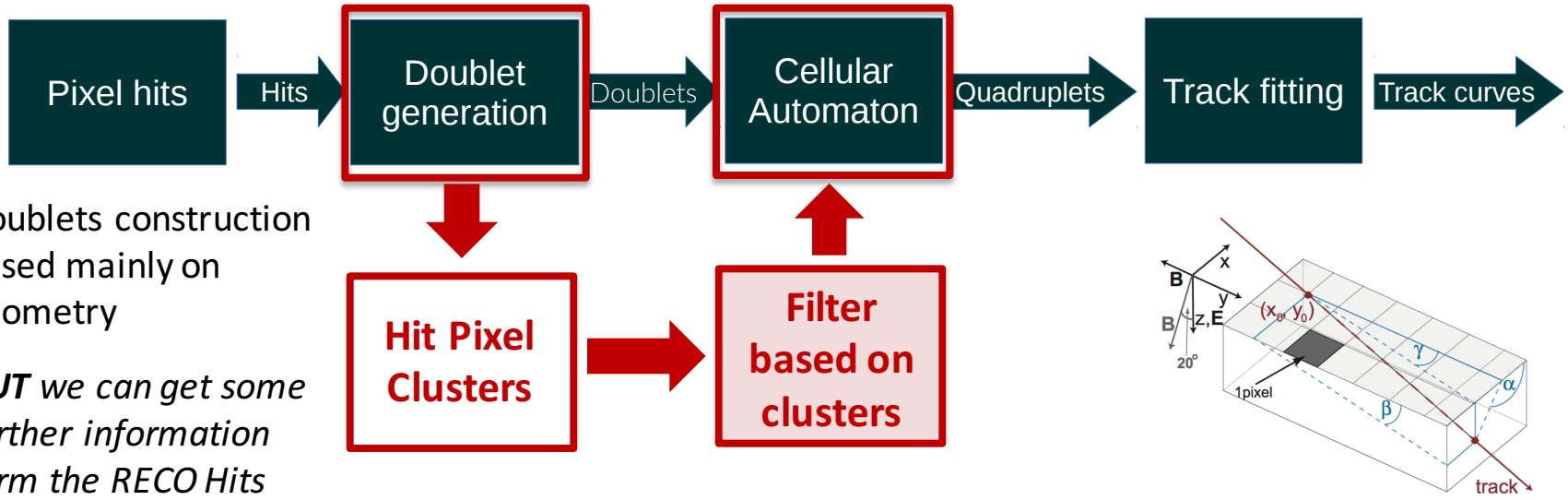
Doublets construction based mainly on geometry

BUT *we can get some further information form the RECO Hits*

Hit Pixel Clusters



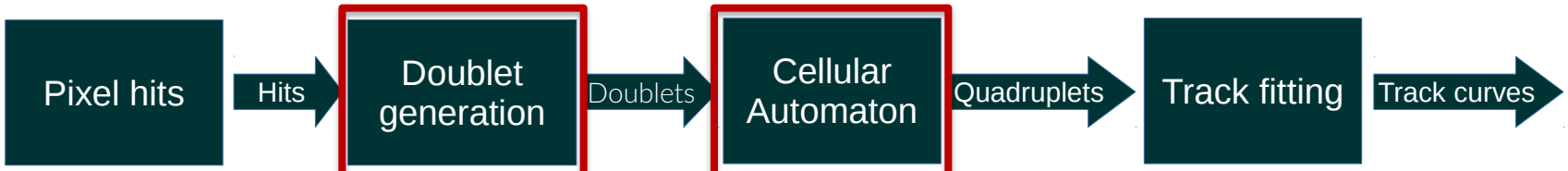
Clustering, filtering & ML



Doublets construction based mainly on geometry

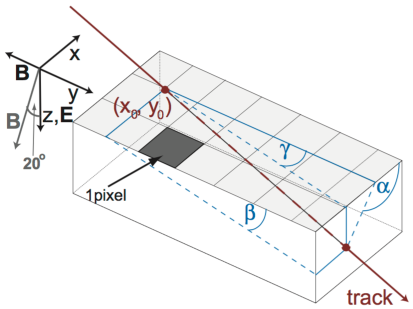
BUT *we can get some further information form the RECO Hits*

Clustering, filtering & ML



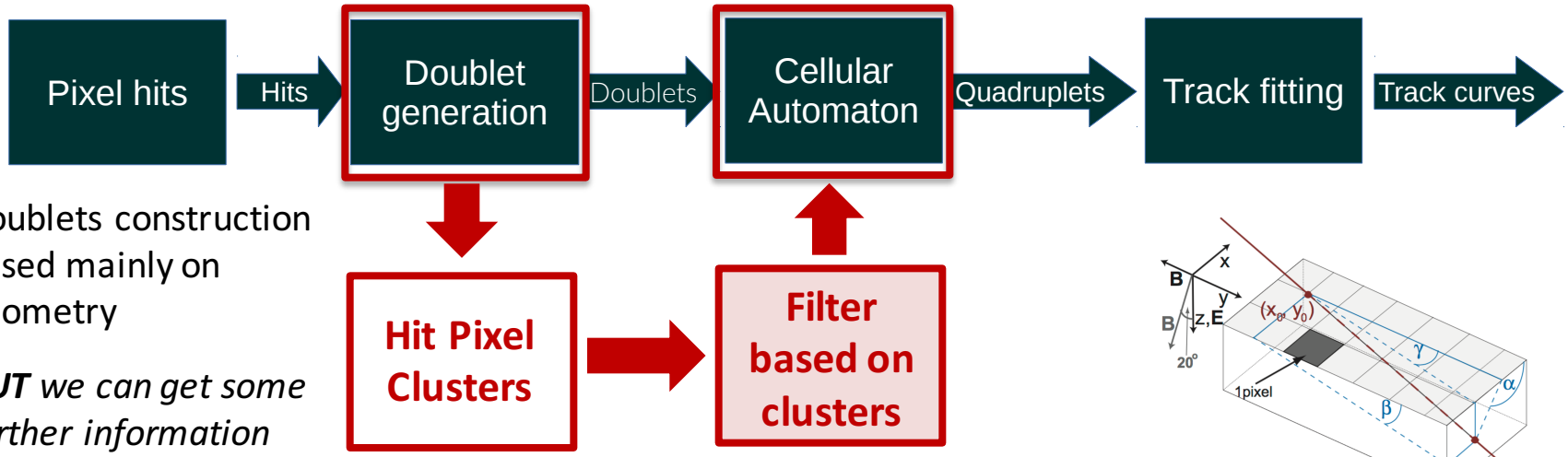
Doublets construction based mainly on geometry

BUT *we can get some further information form the RECO Hits*



How is a Pixel Cluster represented in the CMSSW?

Clustering, filtering & ML



Doublets construction based mainly on geometry

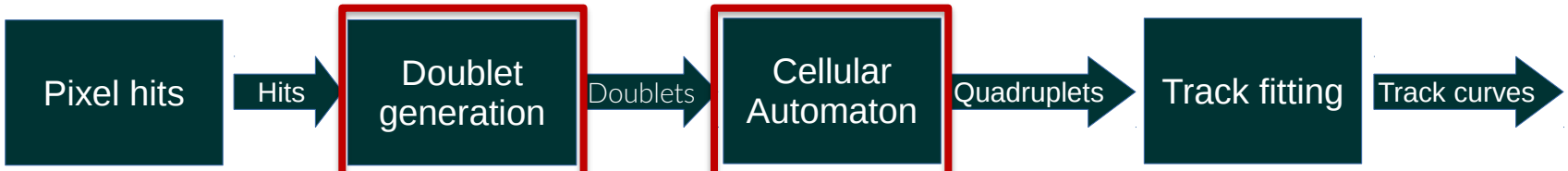
***BUT** we can get some further information form the RECO Hits*

How is a Pixel Cluster represented in the CMSSW?

`class SiPixelCluster` "collection" of `Pixel`

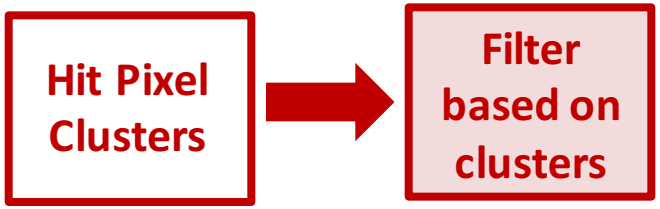
A matrix whose indices correspond to *position* and elements to *adc* values (**only** for pixels turned on).

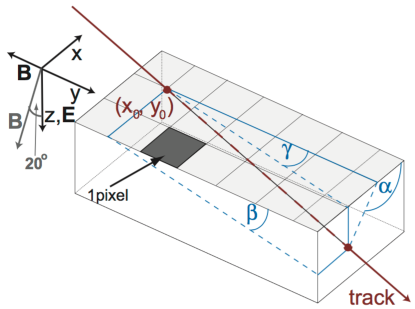
Clustering, filtering & ML



Doublets construction based mainly on geometry

***BUT** we can get some further information form the RECO Hits*

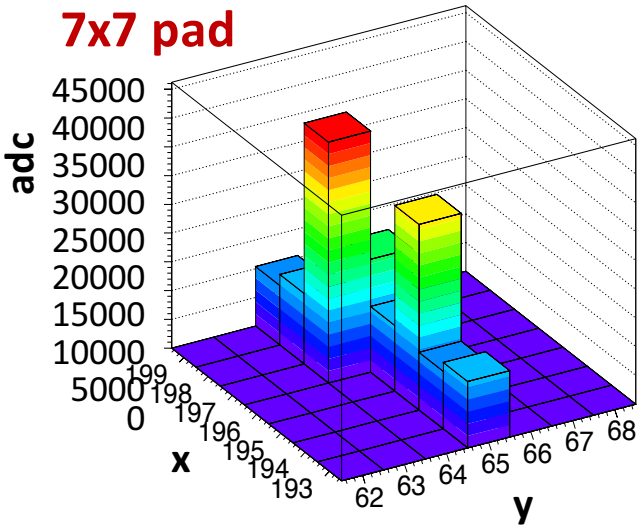




How is a Pixel Cluster represented in the CMSSW?

`class SiPixelCluster` "collection" of `Pixel`

A matrix whose indices correspond to *position* and elements to *adc* values (**only** for pixels turned on).

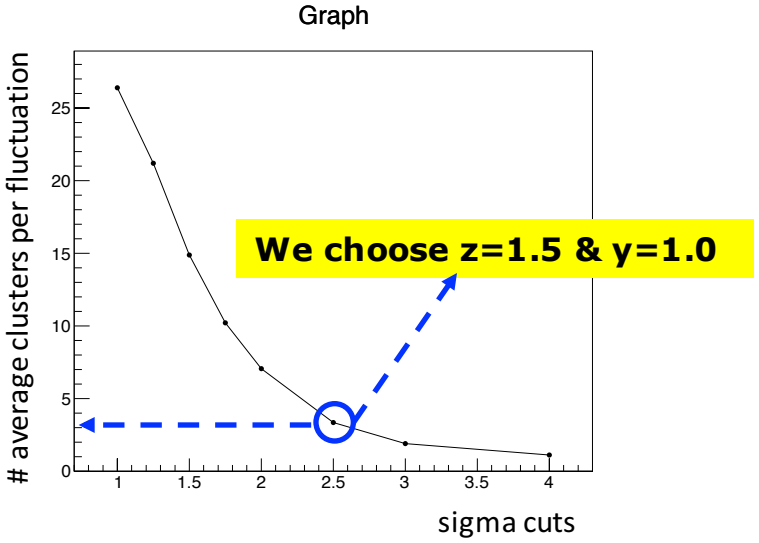


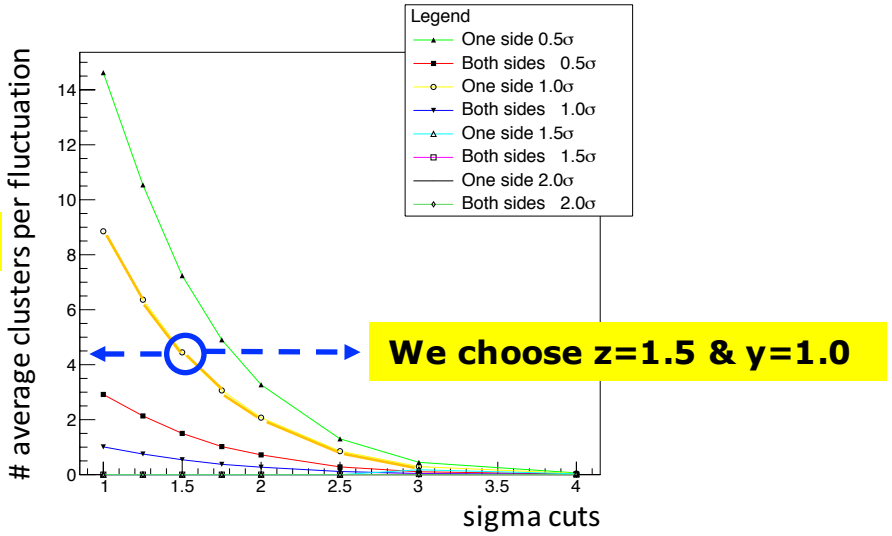
GooFit : LEE & scanning technique

For the physics case of study, as first step, we studied how to choose:

- **x (single seed threshold);**
- **y (side bin threshold);**
- **z (additional sided seed threshold)**

... by counting the mean value of the distribution of the number of seeds/clusters per single fluctuation.





These cuts assure us to build ~9 clusters in average for each Toy MC distribution and that at least 1 cluster is always found in order to perform at least 1 *Alternative Hypothesis* fit. As first test We have produced (in ~10 days, by using only 1 server equipped with 2 nVidia Tesla K20 GPUs) ~9M of MC Toys.

GooFit : LEE & scanning technique

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking behaviours with respect to the background model everywhere in the mass spectrum.

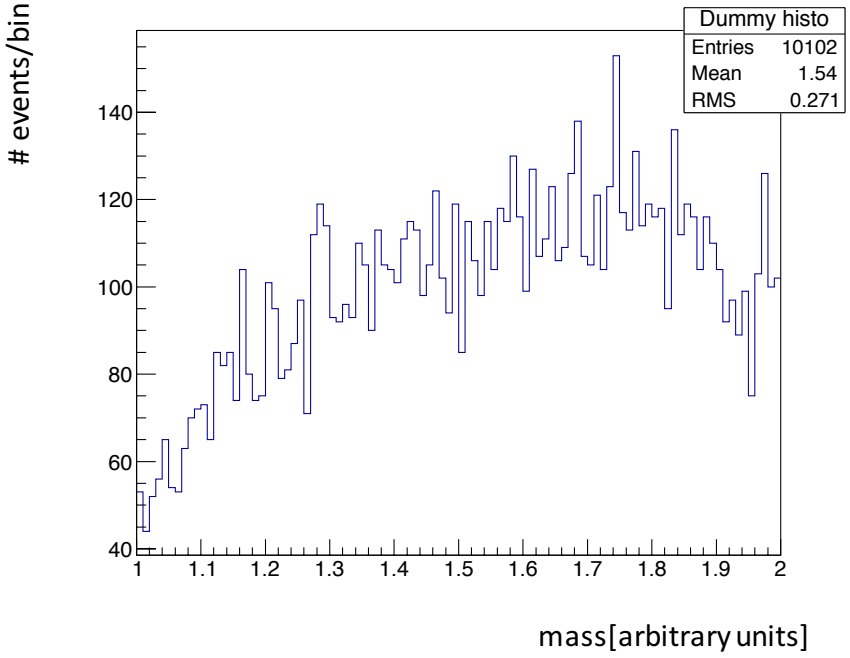
The scanning step has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do **not** miss any interesting fluctuation

B) Do **not** select too many small fluctuations

The procedure:

N.B. Here we show a *dummy* distribution for the sake of a clear visualisation of the procedure, since the actual data have not been published yet.



GooFit : LEE & scanning technique

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking behaviours with respect to the background model everywhere in the mass spectrum.

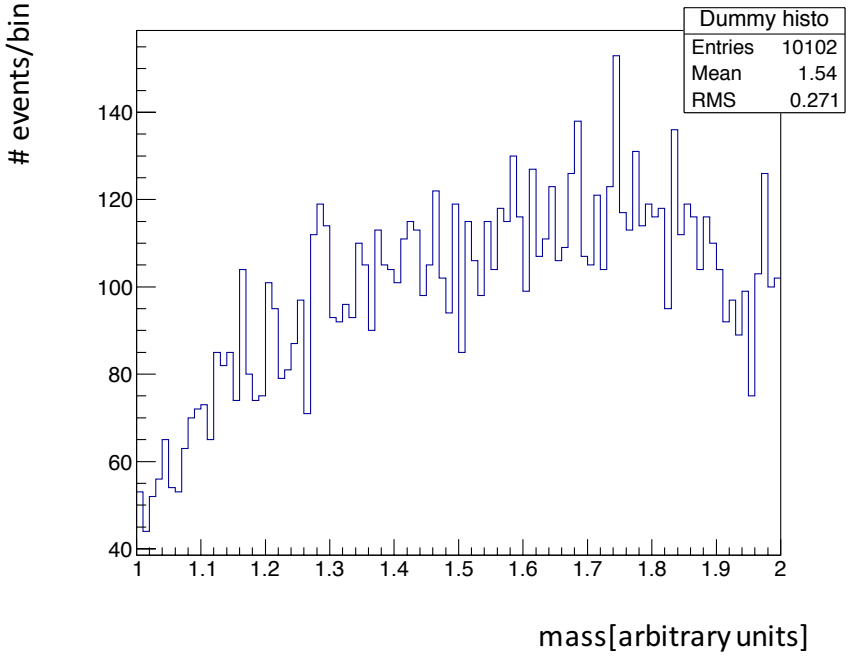
The scanning step has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do **not** miss any interesting fluctuation

B) Do **not** select too many small fluctuations

The procedure:

N.B. Here we show a *dummy* distribution for the sake of a clear visualisation of the procedure, since the actual data have not been published yet.



GooFit : LEE & scanning technique

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking behaviours with respect to the background model everywhere in the mass spectrum.

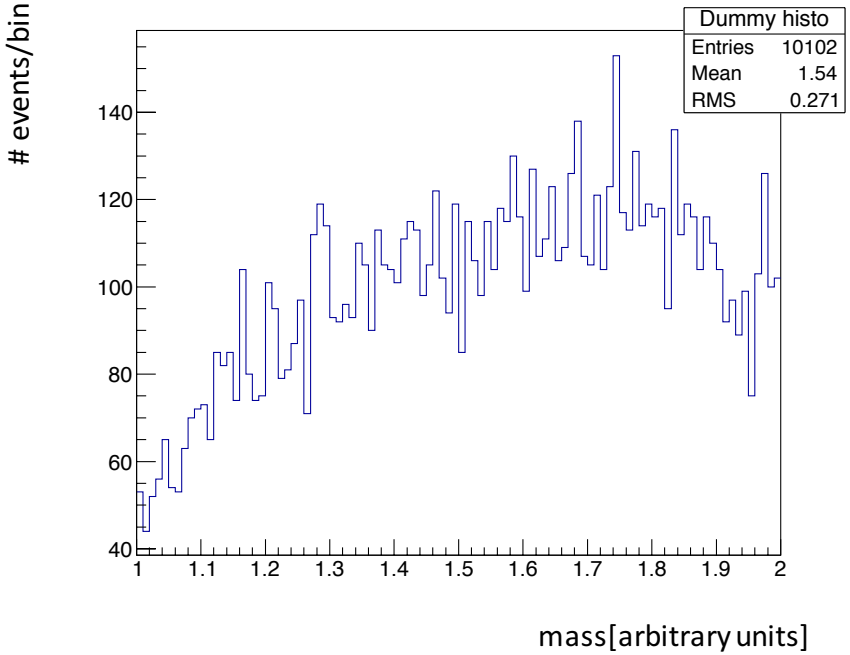
The scanning step has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do **not** miss any interesting fluctuation

B) Do **not** select too many small fluctuations

The procedure:

1. For **each MC Toy iteration** a distribution based on the **background p.d.f**. model is generated in the range whole mass spectrum via *Hit or Miss procedure*. The # of events is fixed by the # of events found in the data.



GooFit : LEE & scanning technique

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking behaviours with respect to the background model everywhere in the mass spectrum.

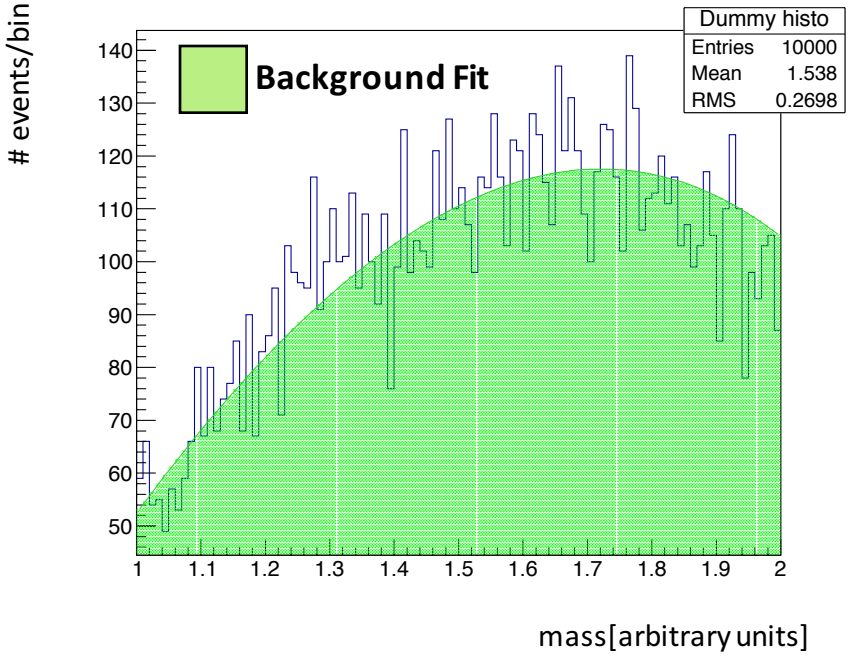
The scanning step has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do **not** miss any interesting fluctuation

B) Do **not** select too many small fluctuations

The procedure:

2. The *Null Hypothesis* fit is performed with the background function only (the same used to generate the data) in order to set up the clustering procedure.



GooFit : LEE & scanning technique

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking behaviours with respect to the background model everywhere in the mass spectrum.

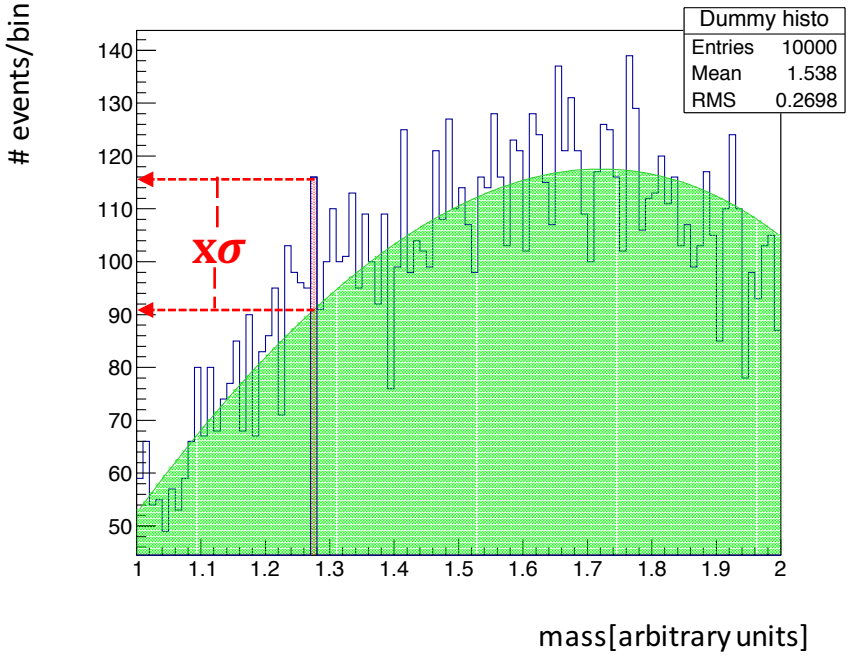
The scanning step has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do **not** miss any interesting fluctuation

B) Do **not** select too many small fluctuations

The procedure:

3. Search for a **seed** defined as a bin whose content fluctuates more than $x\sigma$ strictly above the value of the background function in the center of that bin.



GooFit : LEE & scanning technique

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking behaviours with respect to the background model everywhere in the mass spectrum.

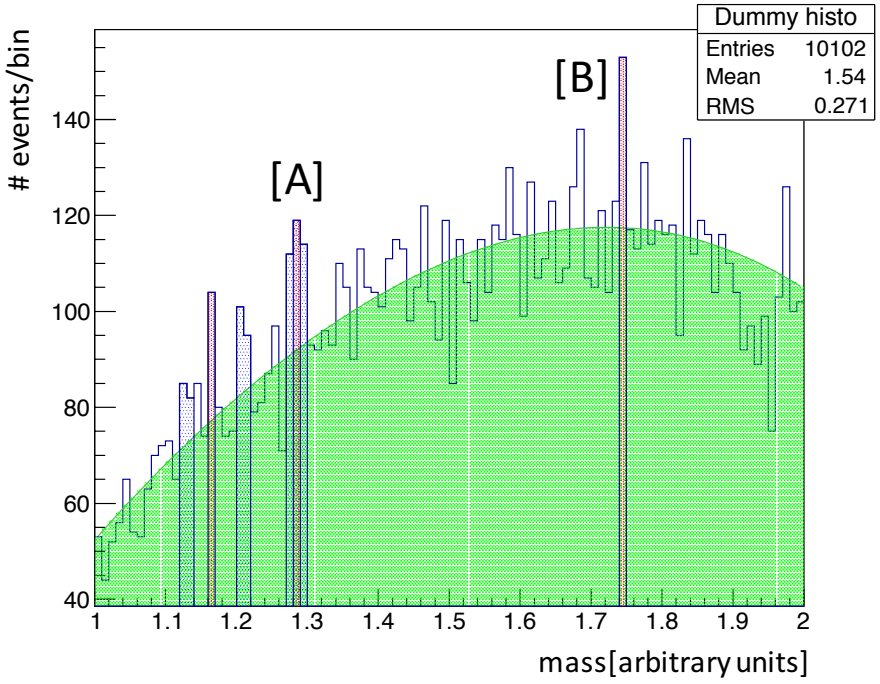
The scanning step has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do **not** miss any interesting fluctuation

B) Do **not** select too many small fluctuations

The procedure:

4. Check if the **seed's side bins** show a content that fluctuates more than ***yσ strictly*** above the value of the background function in the center of that bin. In case of positive result **the side bin(s) is(are) attached to the seed thus forming a cluster [A]**. In case of negative result the **seed bin is taken alone [B]** .



GooFit : LEE & scanning technique

When an unexpected signal is found a ***global significance*** must be estimated. Thus the LEE must be considered and a *scanning technique* must be implemented in order to consider all the relevant peaking behaviours with respect to the background model everywhere in the mass spectrum.

The scanning step has been configured on the basis of a clustering approach and has been designed in advance with the aim to satisfy two concurrent requirements:

A) Do **not** miss any interesting fluctuation

B) Do **not** select too many small fluctuations

The procedure:

4. Check also for **"light" seeds:** bins that fluctuates more than $z\sigma$ with $z < x$ and with at least a side bin fluctuating more than $y\sigma$. In case of positive result a cluster is formed [C].