Neutrino Group Meeting

DBSCAN selection: parameters, cuts, dEdx, point resolution

C. Jesús-Valls
cjesus@ifae.es

TestBeam Analysis

DBSCAN algorithm

DBSCAN algorithm:

Given a set of N entries of 3D coordinates (i,j,t), where i,j are pad coordinates with maxADC at time t.
Given two initial parameters DIST, and minHITS.

For all non selected entries:

Start a NEW cluster:

- Take one entry that has not been selected and compute the distance to all non selected entries in the sample.

- If the number of entries closer than DIST is bigger than minHITS add the analyzed entry to cluster the entries closer than DIST to a list of potentially selectable clusters.

- For all potentially selectable clusters, repeat 2.
- If 2. has been tested on all potential entries: End the cluster and start a new One.

NOTATION

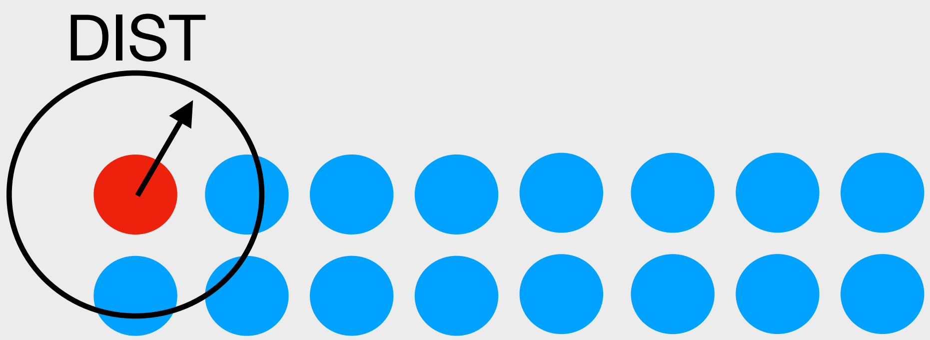
I will use many different sketches to explain how DBSCAN works and to illustrate the procedure I followed to optimize parameters. Then it is useful to have in mind the following legend:

- (blue dot) A entry
- (orange dot) A potentially selectable entry
- (red dot) The entry being analyzed
- (green dot) Any other color form a cluster
- ✗ Entries potentially selectable are not clustered
- ✓ Entries potentially selectable are clustered

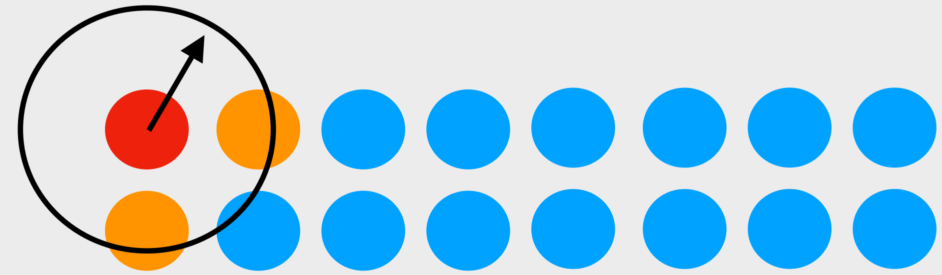
The example works

The example fails

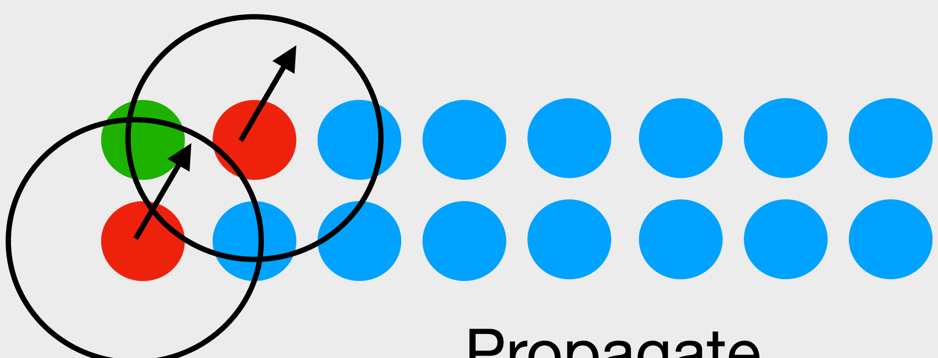
DBSCAN algorithm (2D)



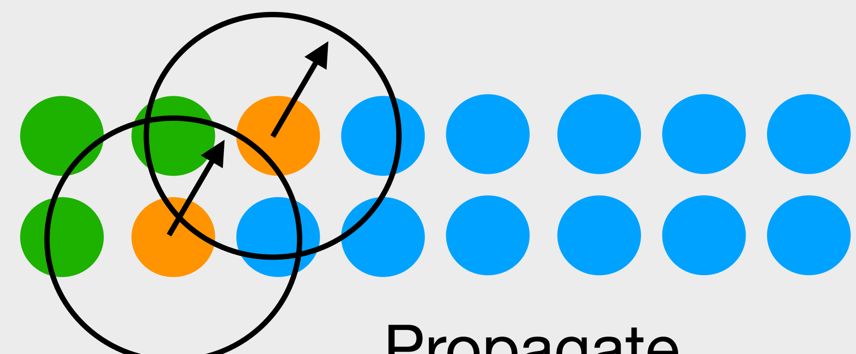
1 We take one entry



2 Compute distance to other tracks. If more than minHITS are inside DIST, cluster them.

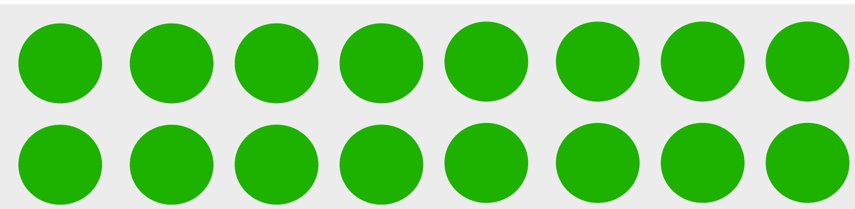


3 Propagate…



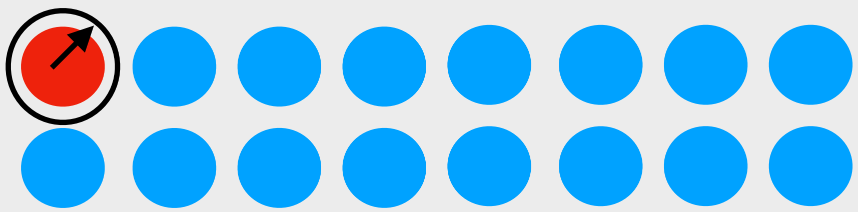
5 Propagate…

…iterations… →

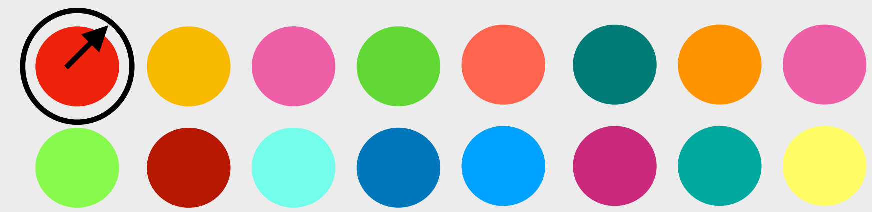


6 End

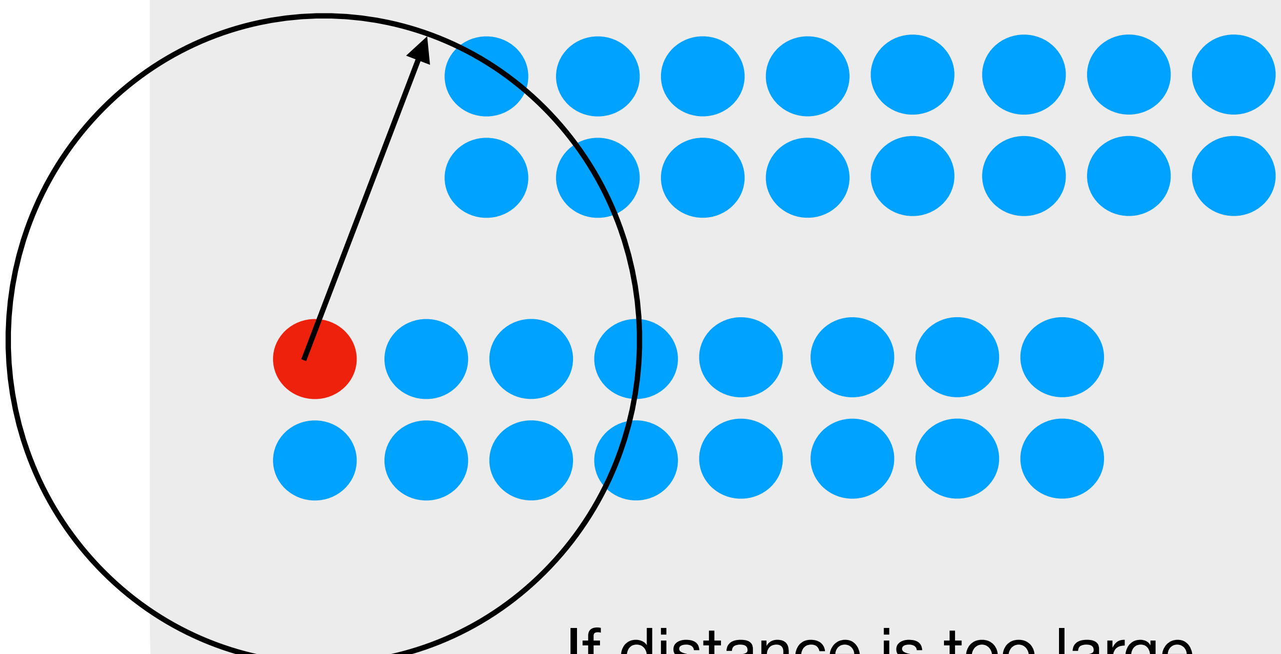
DBSCAN algorithm (2D)



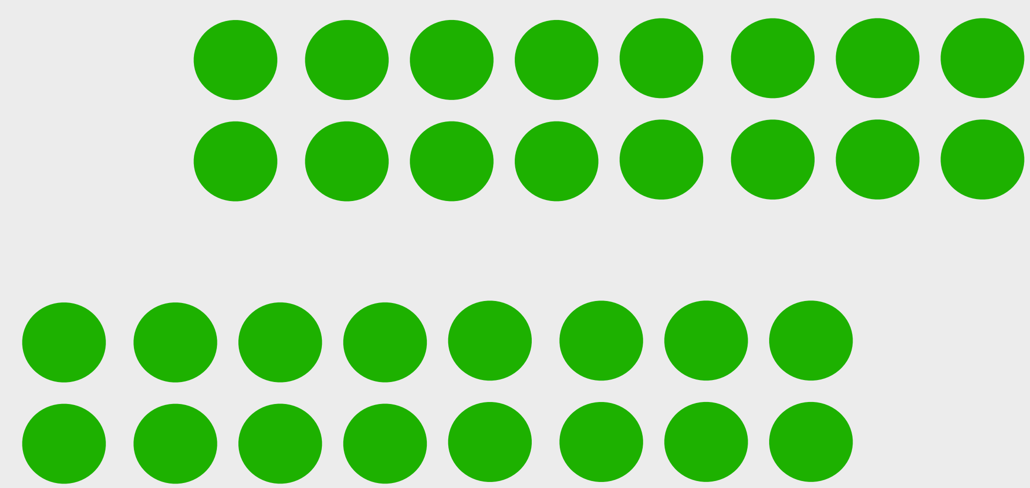
If distance is too short…



Real cluster is split in subclusters



If distance is too large…

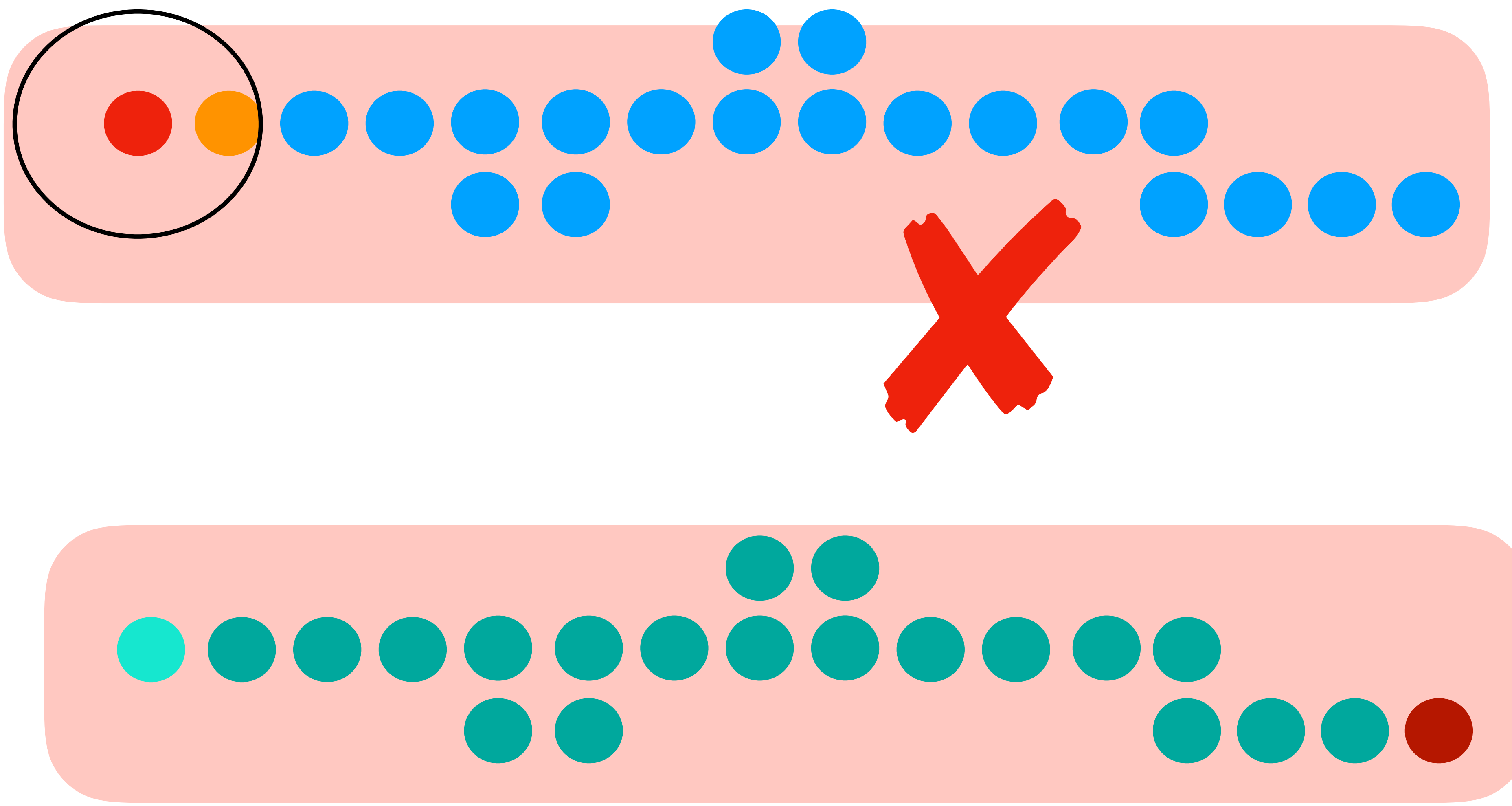


We merge different tracks

Origin
Selected
Others

DIST ~SQRT(2) & minHITS 1

Same parameters:

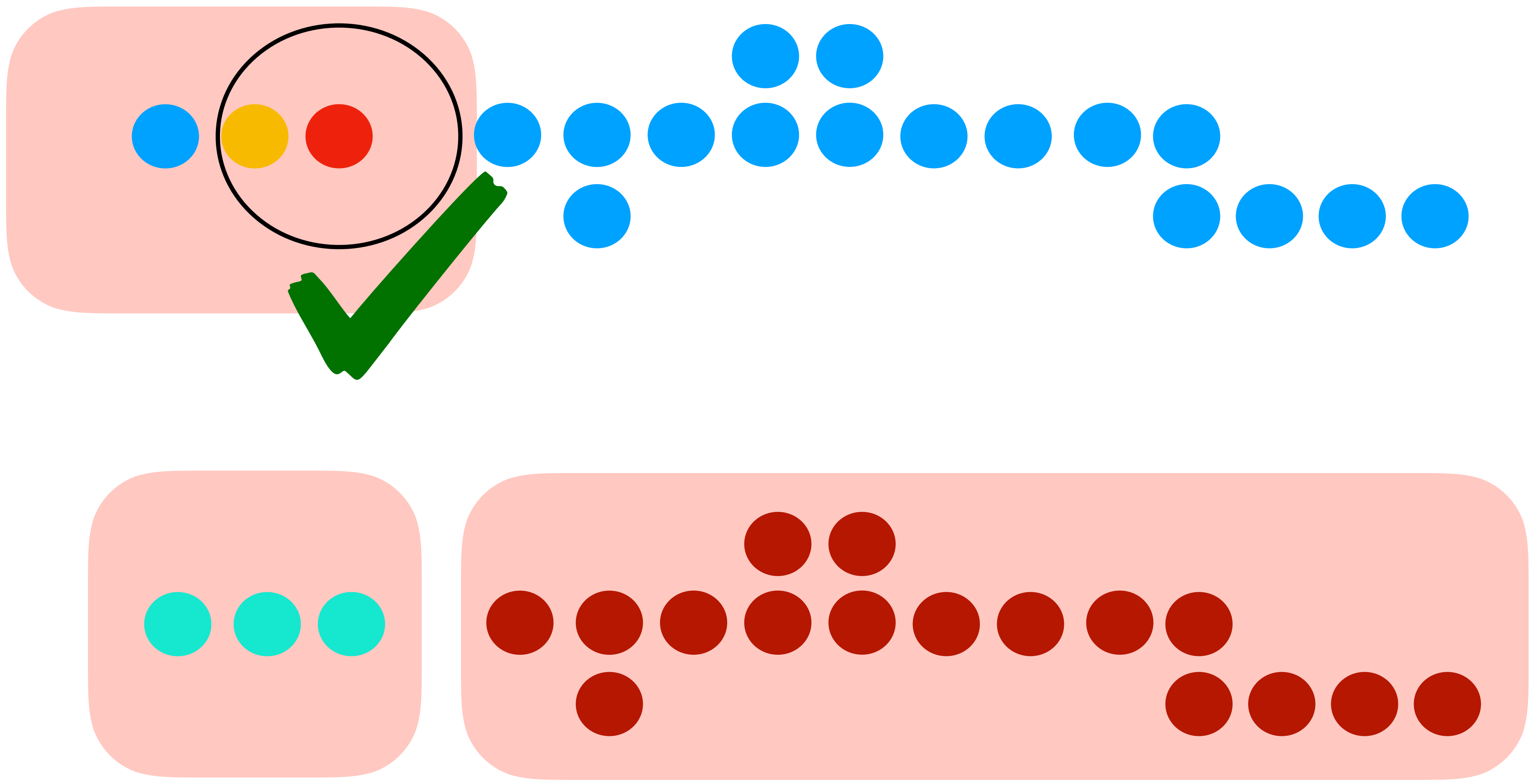


Potentially fail on the edges if minHITS is 2 for thin tracks.

Origin
Selected
Others

DIST ~SQRT(2) & minHITS 1

Same parameters:

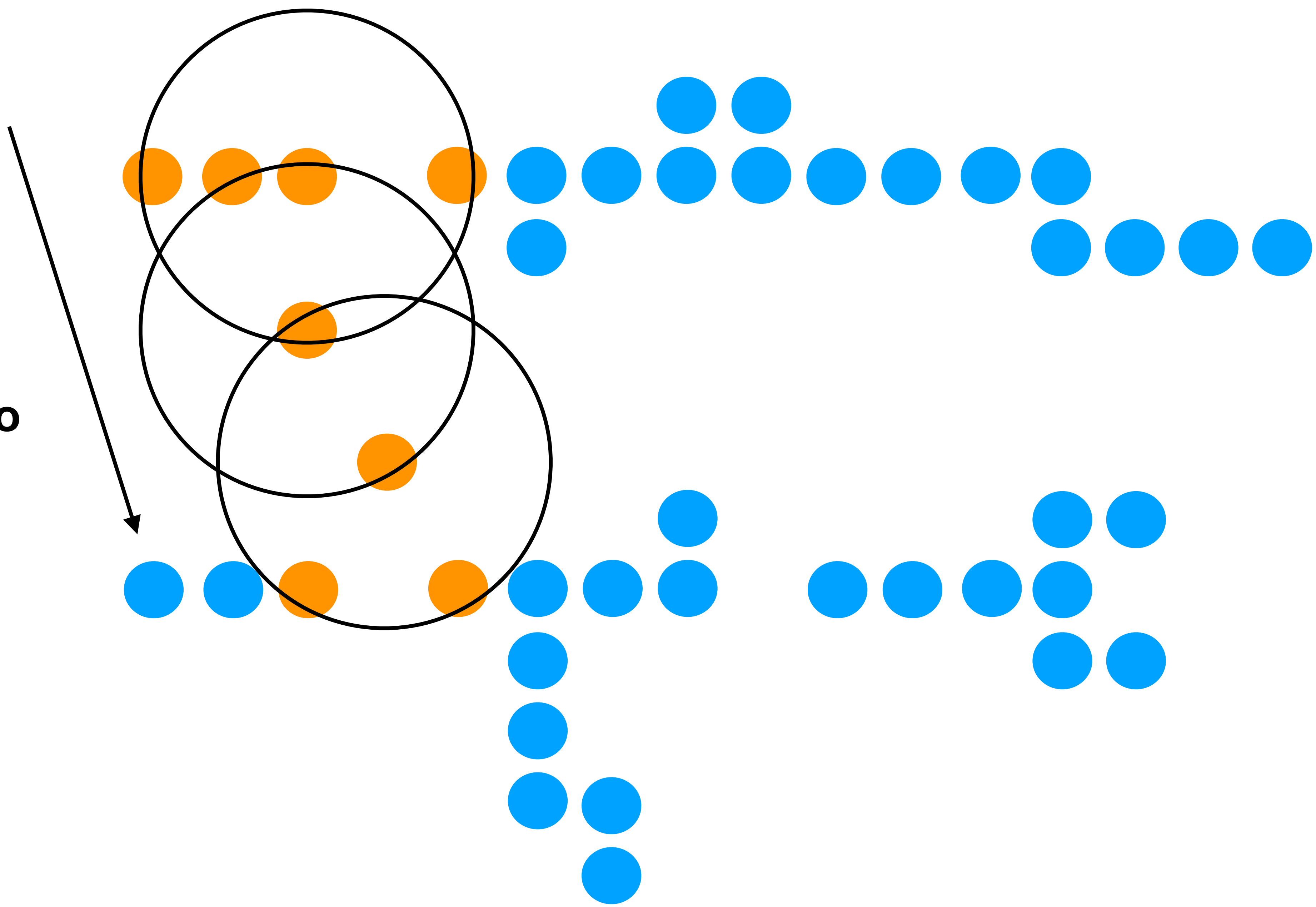


Potentially can cut tracks!

DBSCAN Parameters selection (2D)

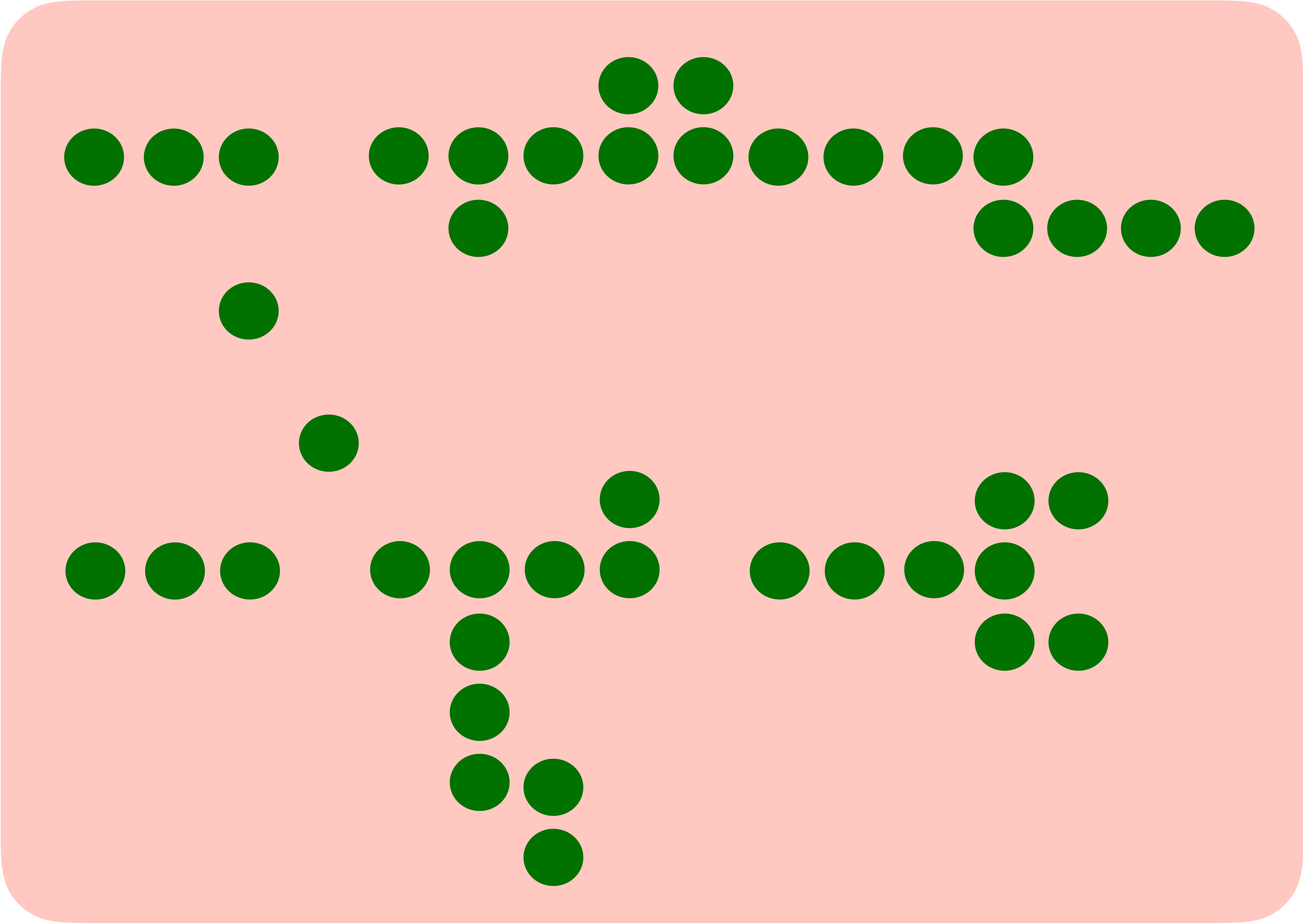
The only possibility is to increase distance… at the risk of including noise, and merging tracks.

DBSCAN allows the clustering to jump!!!



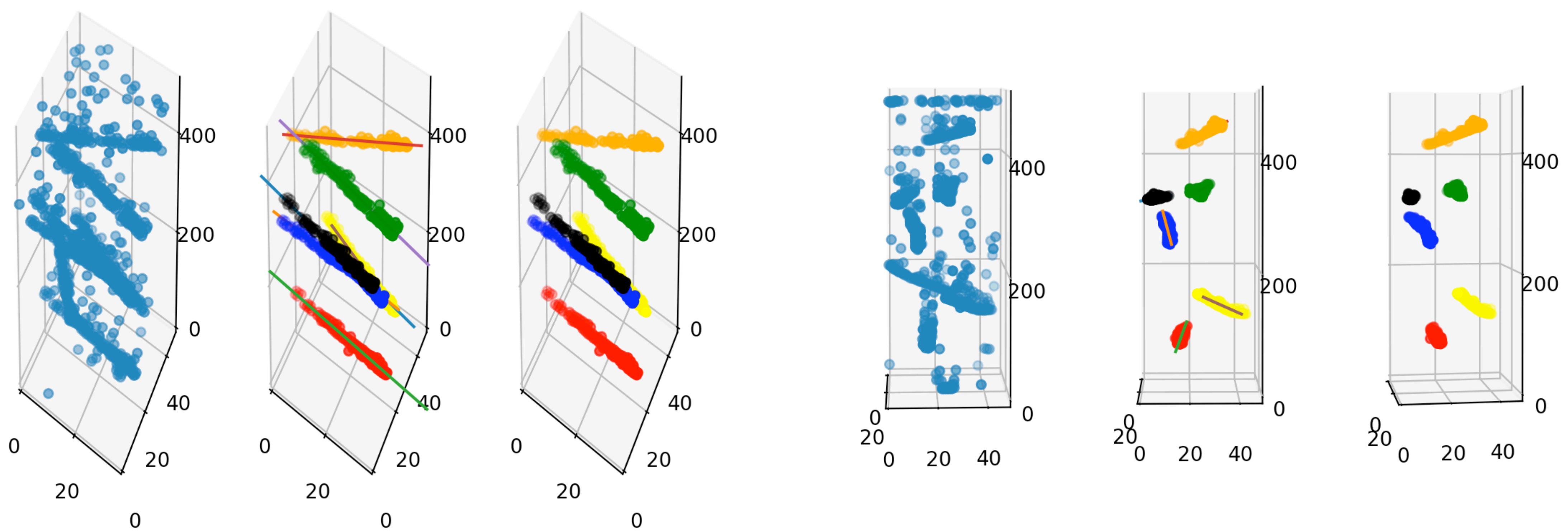
DBSCAN Parameters selection (2D)

With a more or less reasonable selection of parameters this event will be selected as a single track



DBSCAN Parameters selection (2D)

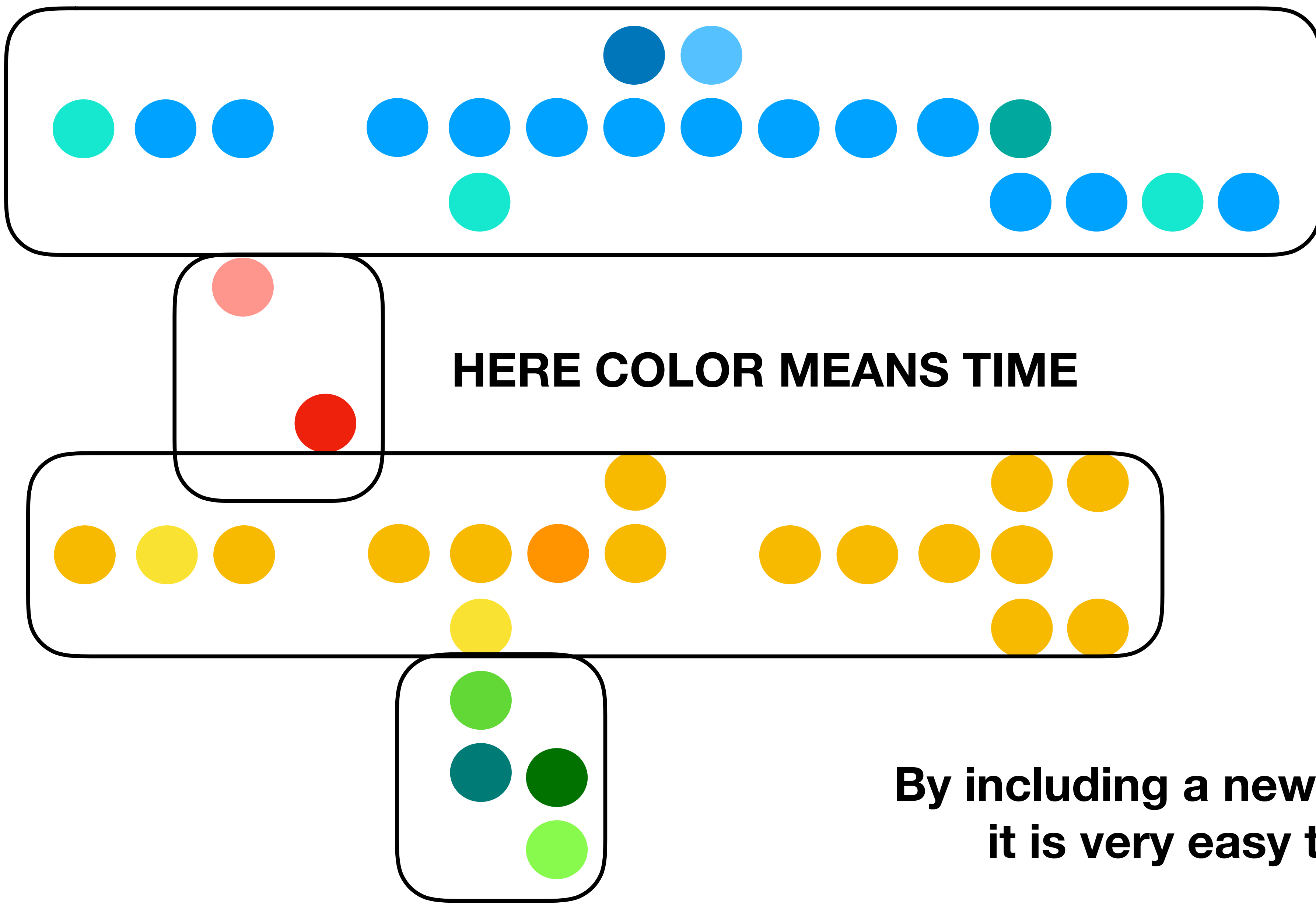
Being honest, that is terrible, and frightening… it can ruin up the analysis if is not controlled.
However, how is it possible to go from that to this?



Adding time is a critical factor.

Consider the same situation but including time in the analysis. time, goes ~ from 0 to 500. For a full window of ~32μs. If we compute the distance as: $d = (\Delta x^2 + \Delta y^2 + \Delta t^2 / f)^{0.5}$

We can fiscally tune the window of selection by setting the f factor!



f has to be big enough to collect all charge spreading, but no too large to merge tracks or include noise!

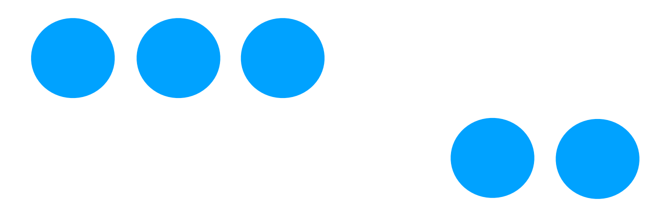
By including a new variable (time == color) it is very easy to distinguish tracks

DBSCAN Parameters selection

According to Sergey is 80. Have to check

Let's do some math:

Time units (from 0 to 500) are 32μs, then 1 a.u is in reality 64ns. f has to be related to the typical time of spreading between pads. We want the same level of discrimination in time that we discussed for space, allowing missing pads, and keeping all spreading.



If $f = \infty$; we have pure spatial separation. $d = (3^2+1^2+0)^{0.5} = 3.16$

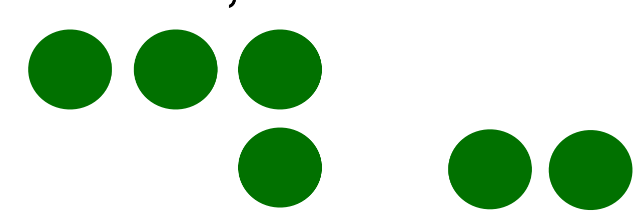
If $f = 12.5$; we have distance would be: $d = (3^2+1^2+(40/12.5)^2)^{0.5} = 4.5$

40 a.u = 2.5μs

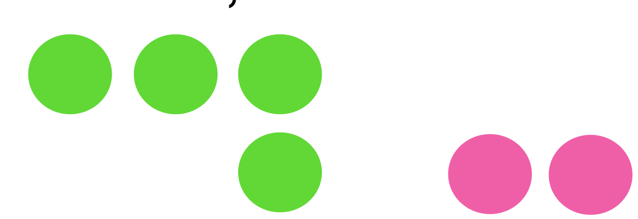
$\boldsymbol{f = 12.5}$

40/12.5 = 3.2, we penalize in the same way being at 2.5μs than being at one pad away.

If DIST = 4, and minHITS = 1



If DIST = 5, and minHITS = 1



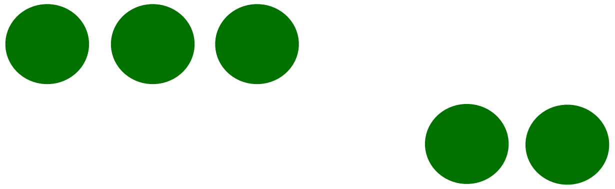
DBSCAN Parameters selection

- As discussed before minHITS must be 1 to avoid non selecting the edges of thin tracks.
- Time condition is imposed by electronics.

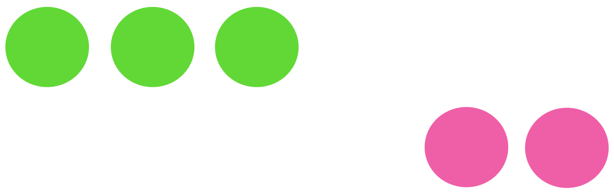
Following very simple arguments we have proven that all the selection discussion can be made by analyzing the impact of a **single** parameter: **the distance.** The other two have been chosen by means of understandable and **physical** requirements.

Distance is of course a physical quantity. However, it is not trivial cut to apply.

If DIST = 5, and minHITS = 1



If DIST = 4, and minHITS = 1



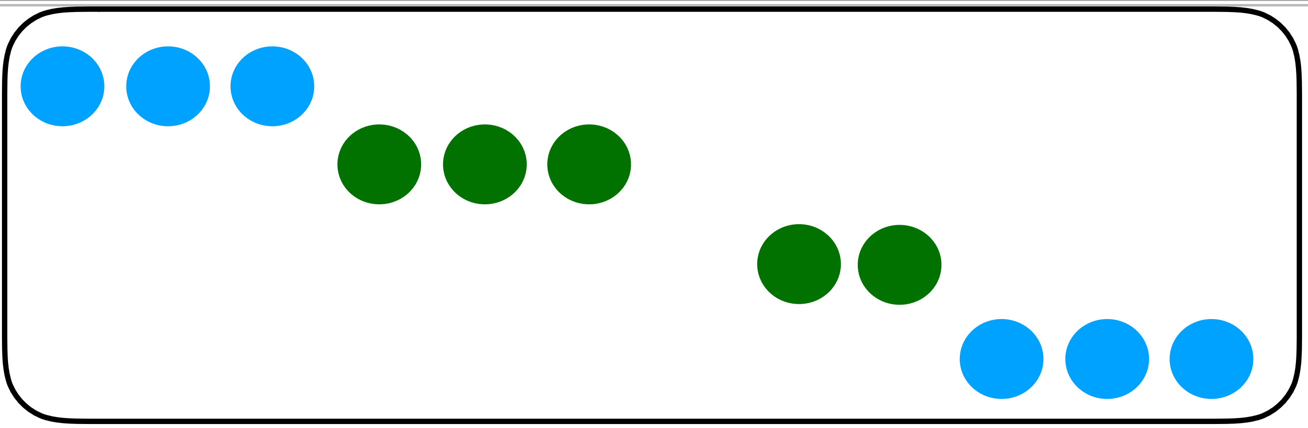
In this situation even given all the information of the system is not very clear if all entries must be clustered.

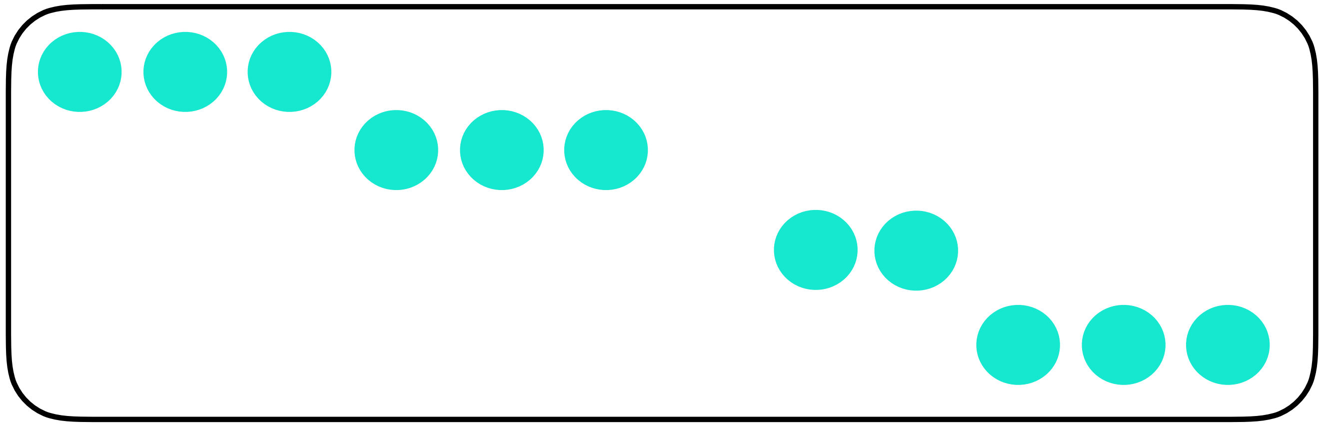
Why?

We need some context

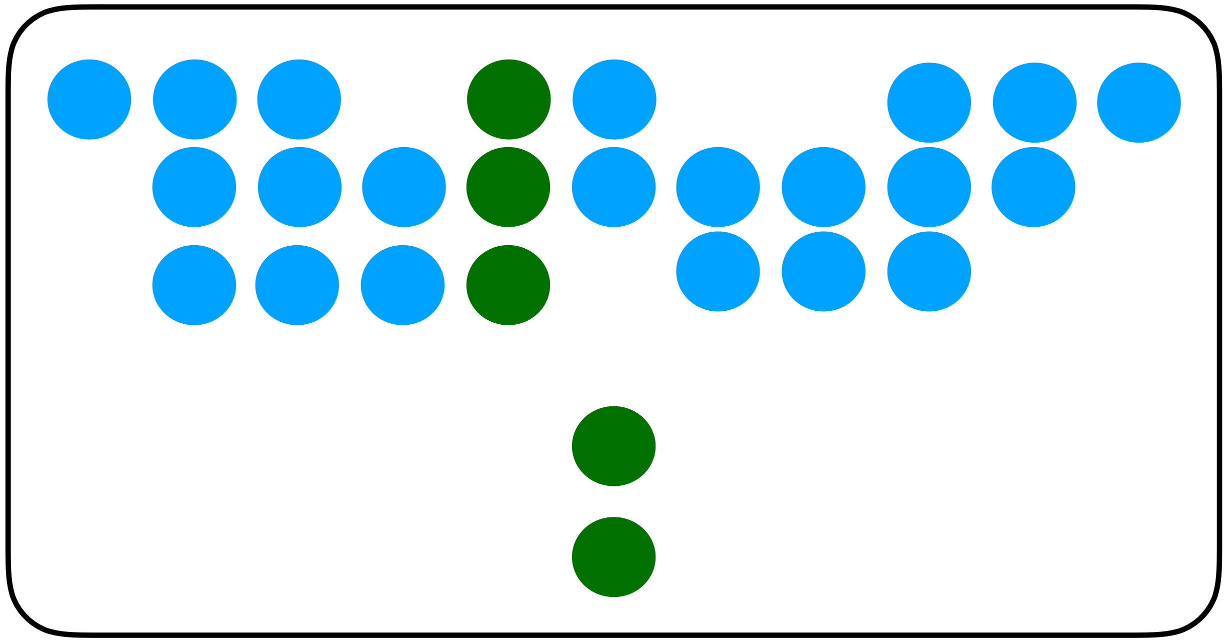
DBSCAN cuts

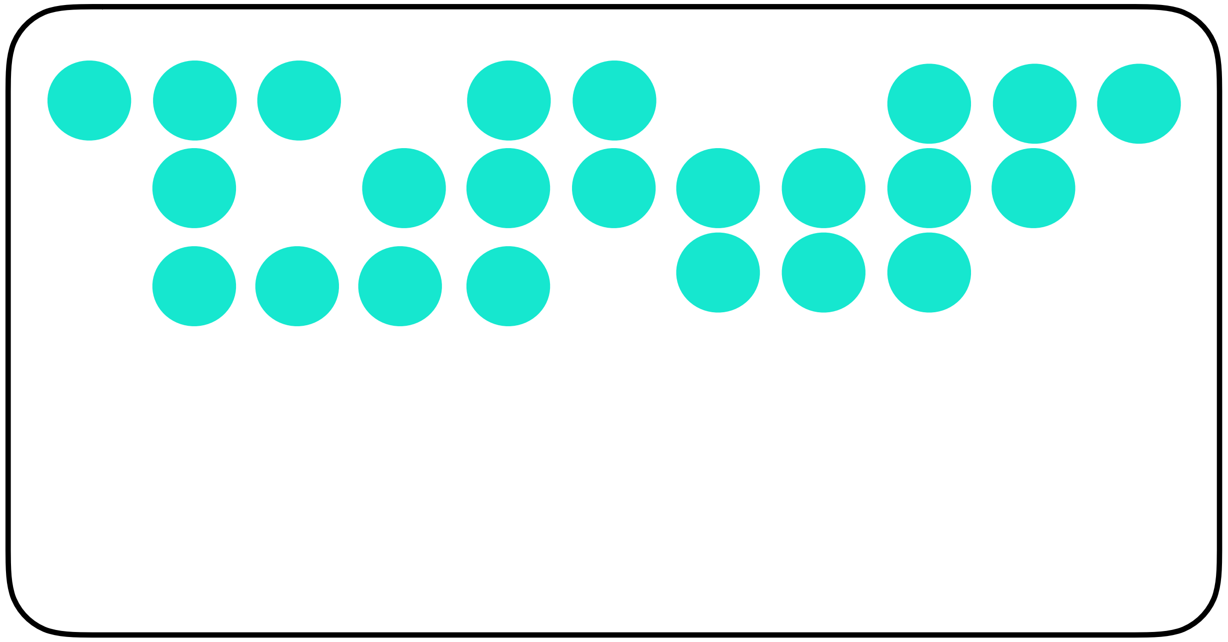
Cluster them!





Dont cluster them!



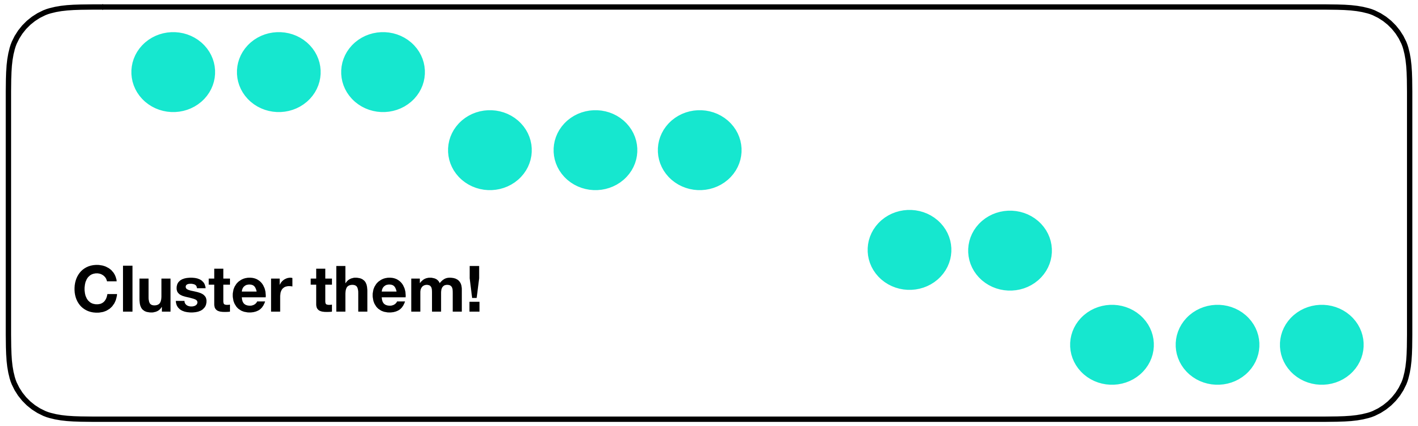


But distance is the same how do you distinguish in DBSCAN? **Is not possible**

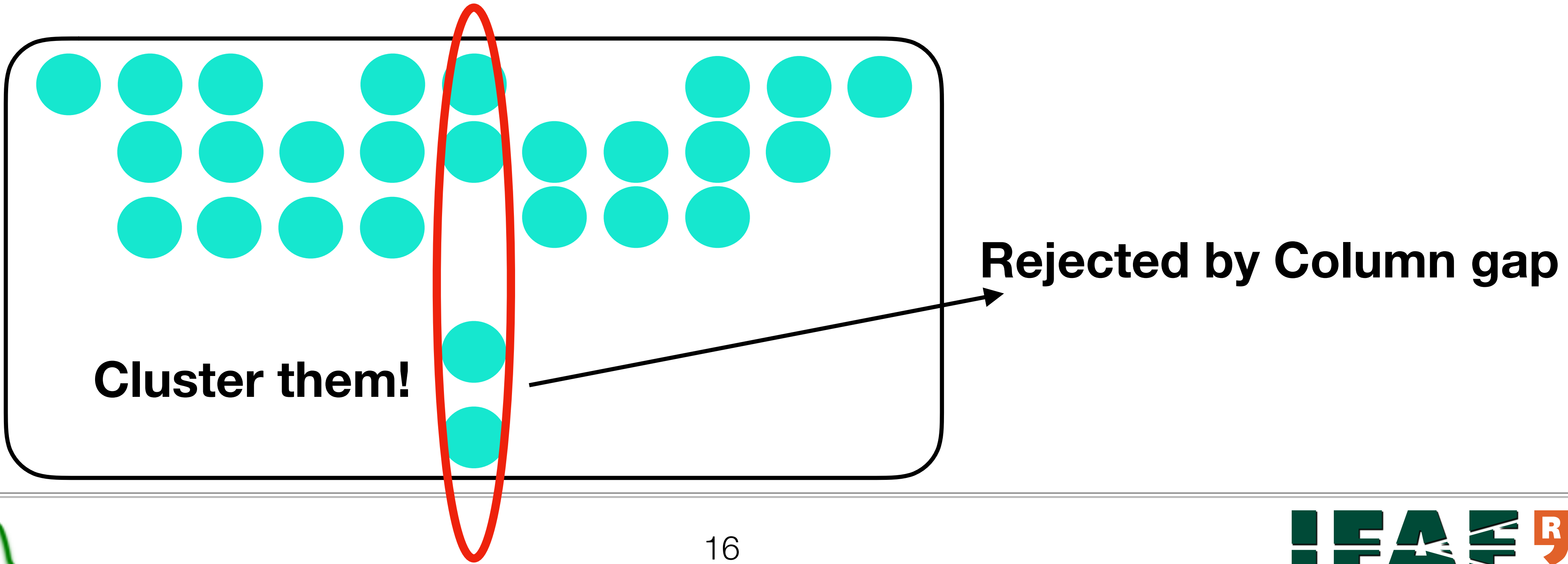
Most conservative: DIST = 5, and minHITS = 1

Does it mean we have to reject DBSCAN because is does not care about the context? No.

Step 1: Take a relatively permissive distance that allow to select physical tracks without missing pads.



Step 2:Impose some extra conditions to forbid some selected tracks:
e.g. **if (there is a gap in a column, reject the event):**

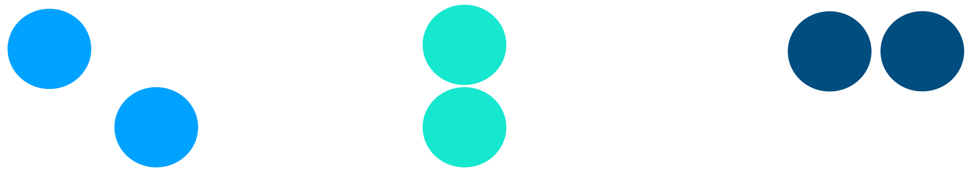


DBSCAN 3D and multitrack

As it was commented the other day, having control of the waveform it is crucial in order to ensure high data-quality. If waveform is not taken into account, the contamination from overlapping tracks highly bias the selected data.

3D analysis has also been implemented in DBSCAN. It works in 2 steps:

1. DBSCAN clustering is applied and clusters passing the cuts are stored.
2. A time filter is applied. This time filter iterates over the selected sample looking to the waveform information of the neighboring previously non selected pads. If there is waveform information in a 2.5μs window around the time of the neighbor pad, the collected charged of that pad (looking to max in window) is considered, adding it to the cluster.
3. The DIST for adding this clusters is consider to be only spatial: We are looking for pads at a maximum distance of sqrt(2)

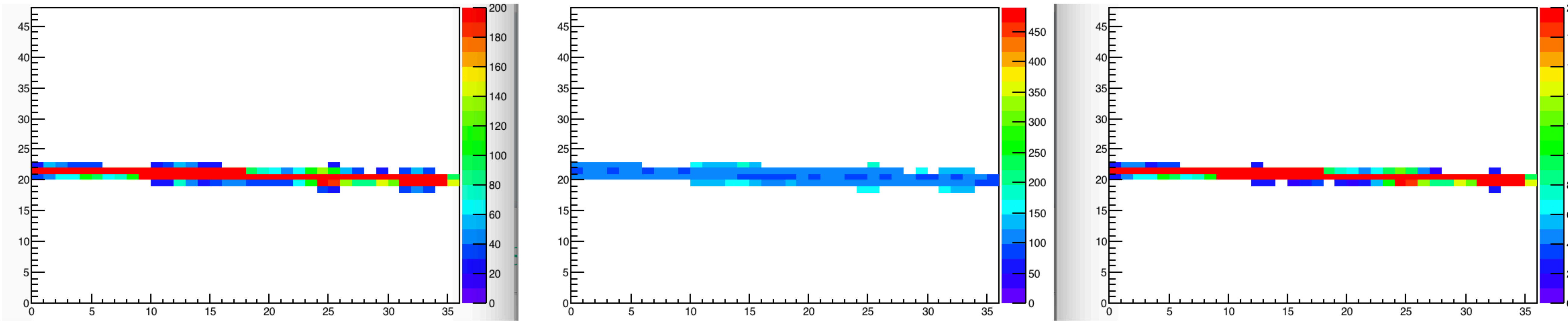


DBSCAN Track examples

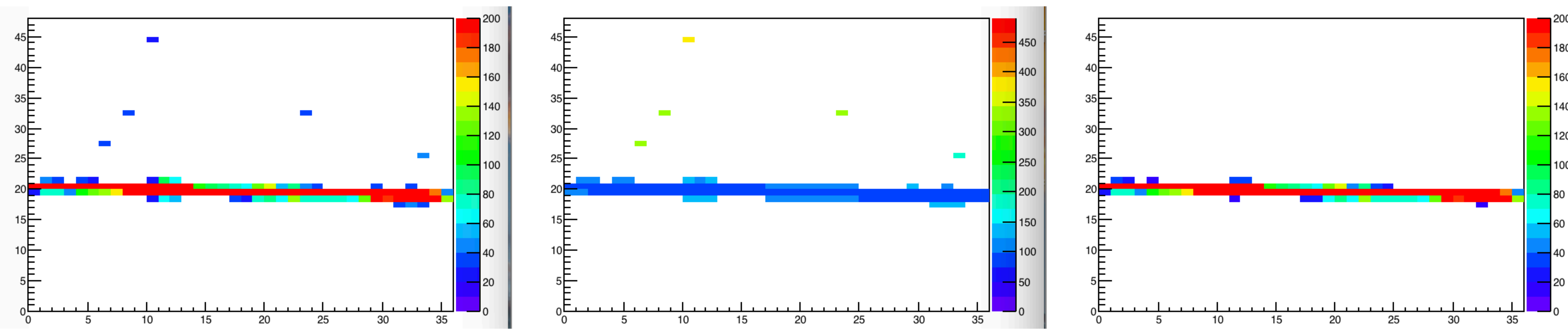
Sergey's selection (TestSelection3D)

Missing Pads at long time!

1st sel events in run 386



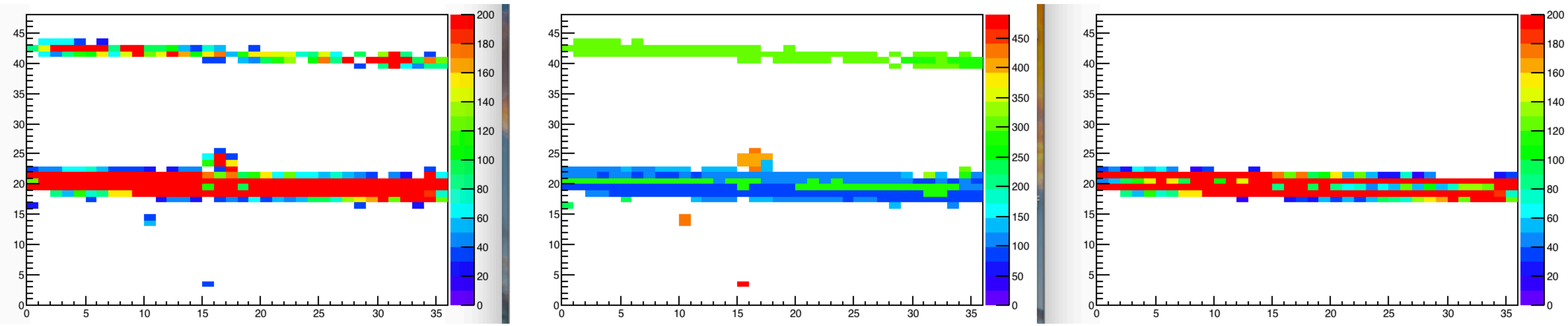
3rd sel events in run 386



DBSCAN Track examples

Sergey's selection (TestSelection3D)

35th sel events in run 386

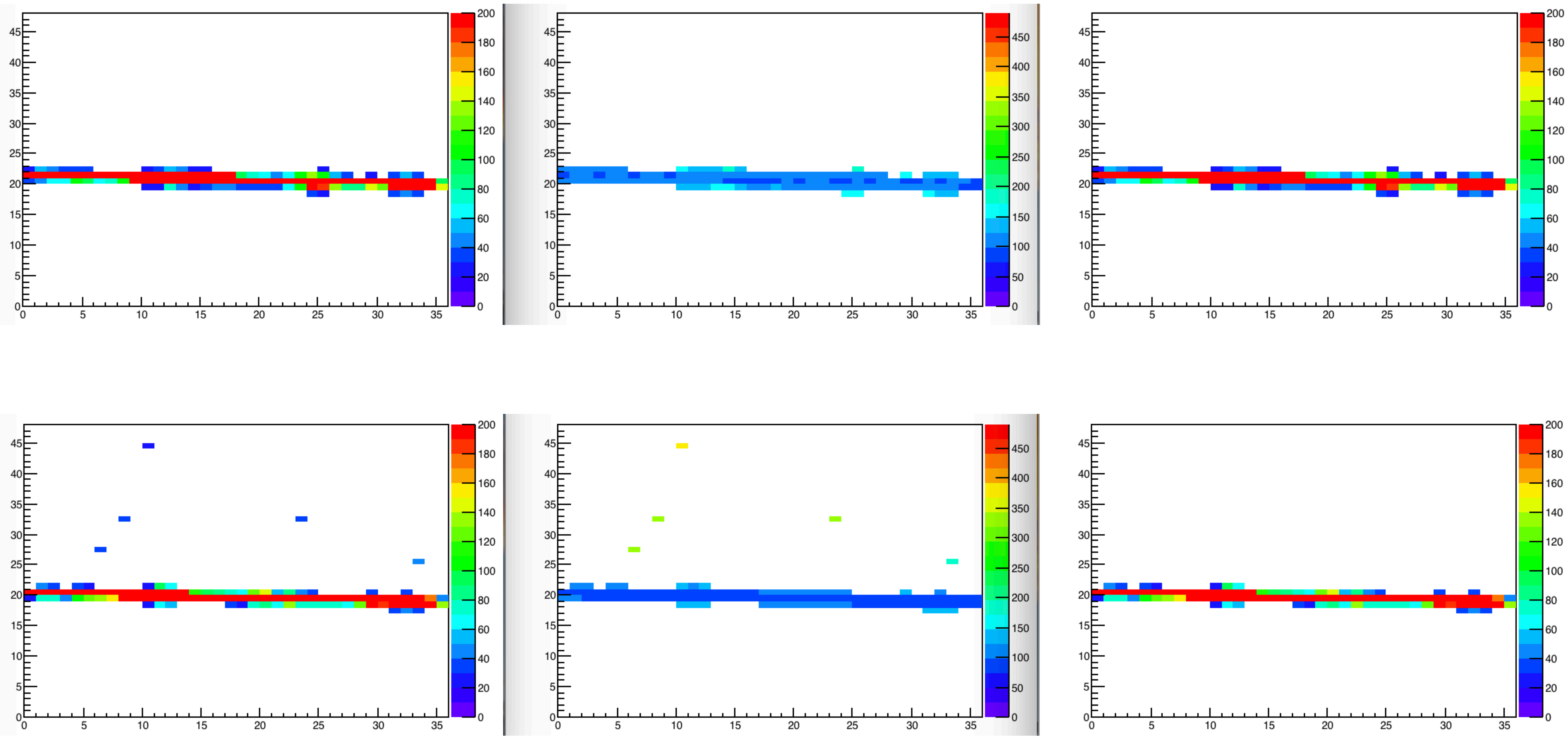


The charge pattern has been changed since it is 3D (It is able to take information from the waveform and therefore change vale of maxADC)

However: Here it is merging 3 tracks!

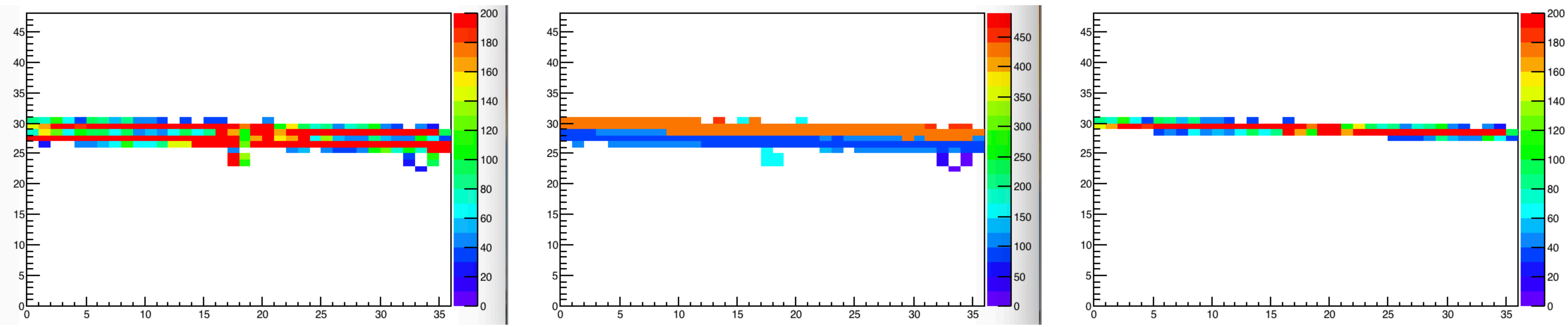
DBSCAN Track examples

All the spreading is collected



DBSCANSelection

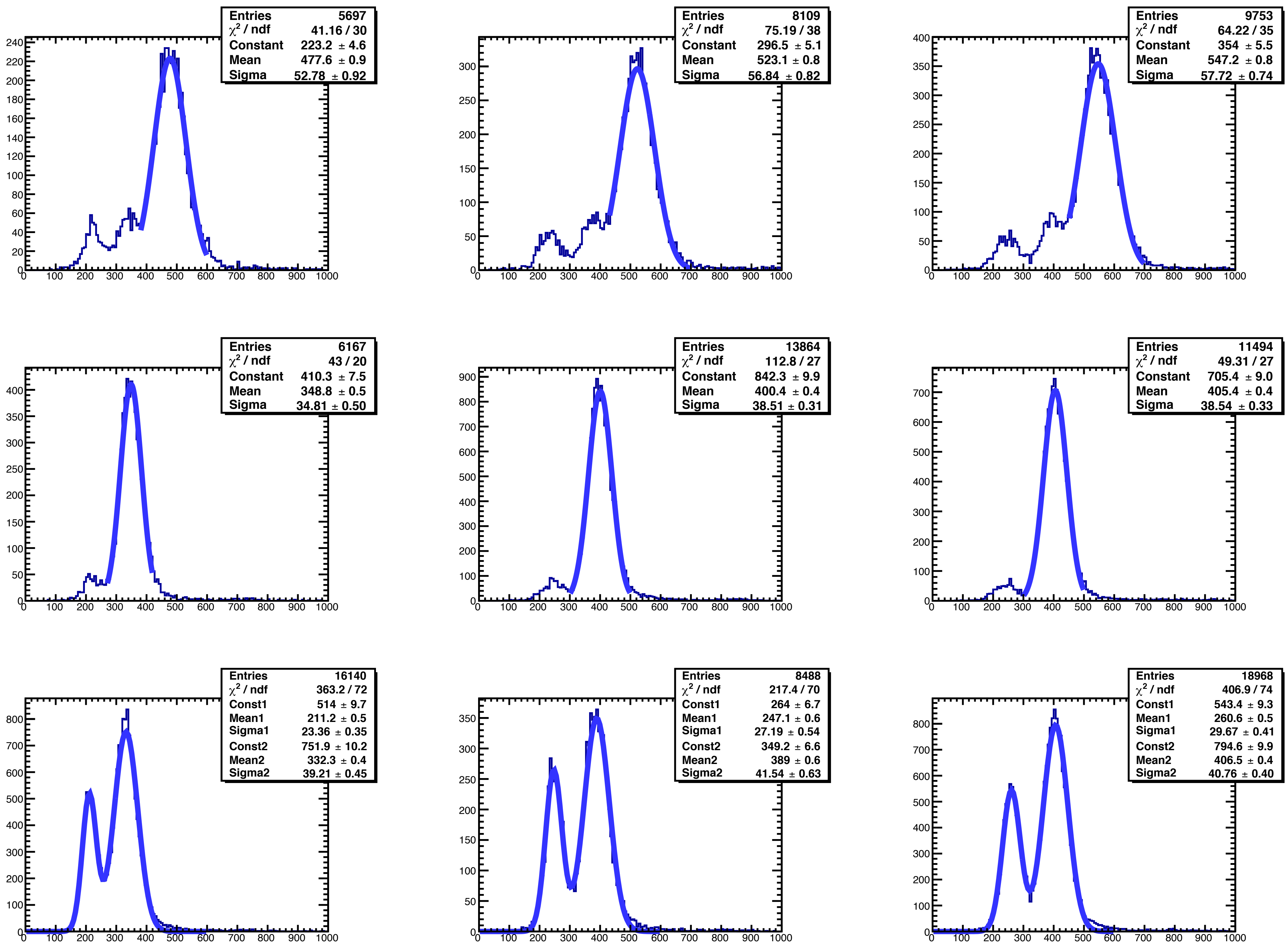
All the spreading is collected



It correctly works like a 3D selection, no merging problems found.

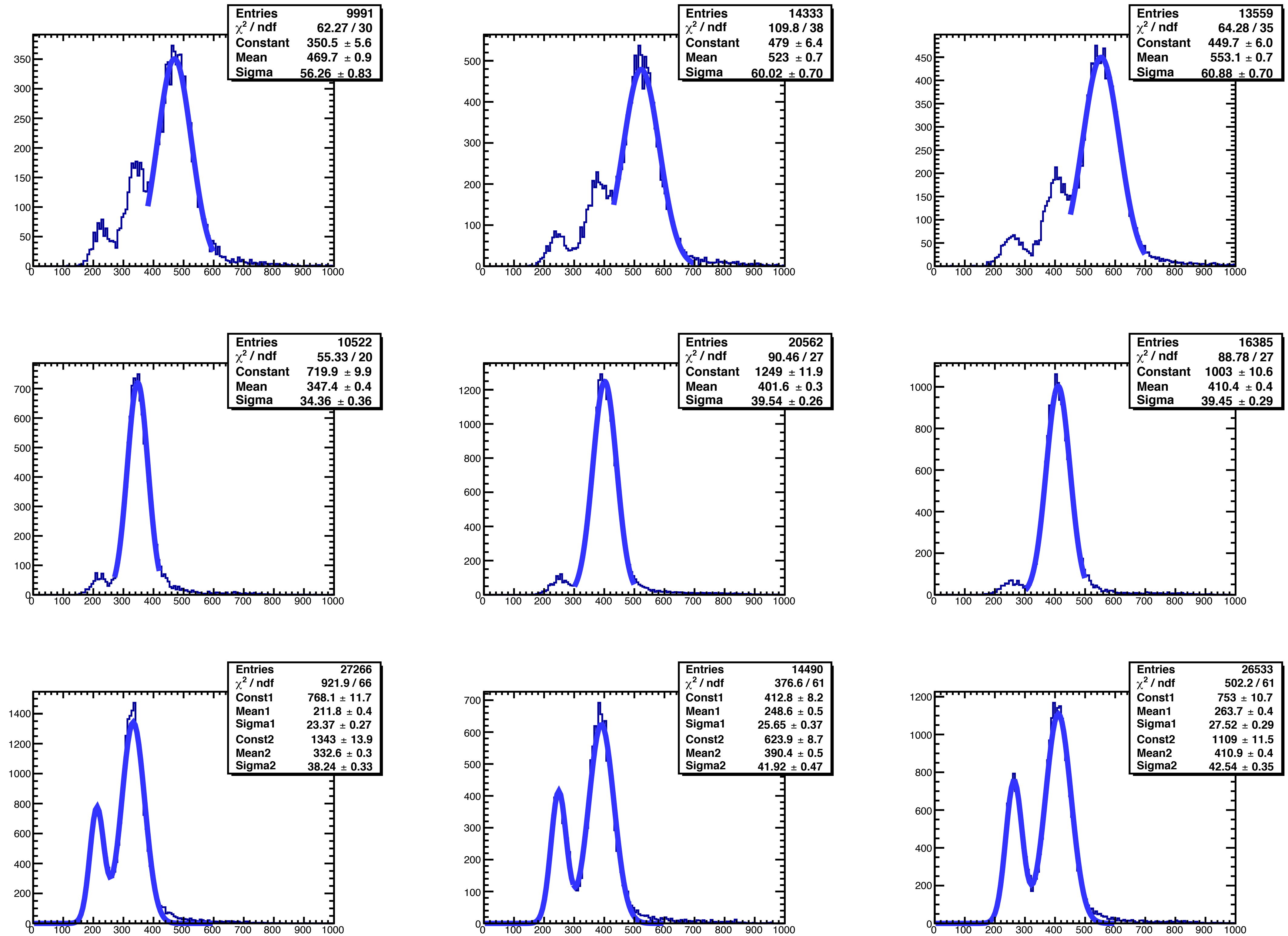
DBSCAN dE/dx comparison

98.8K

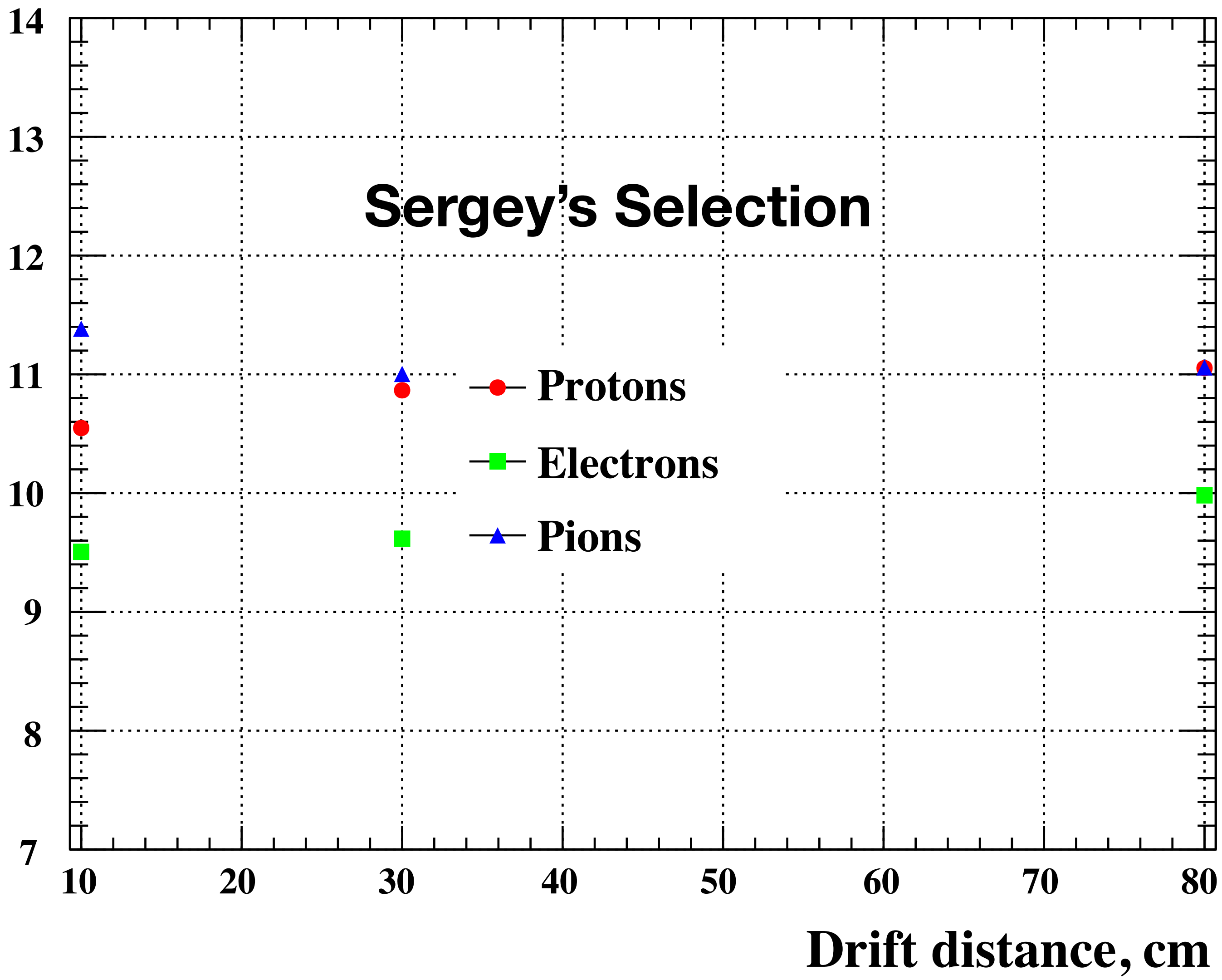


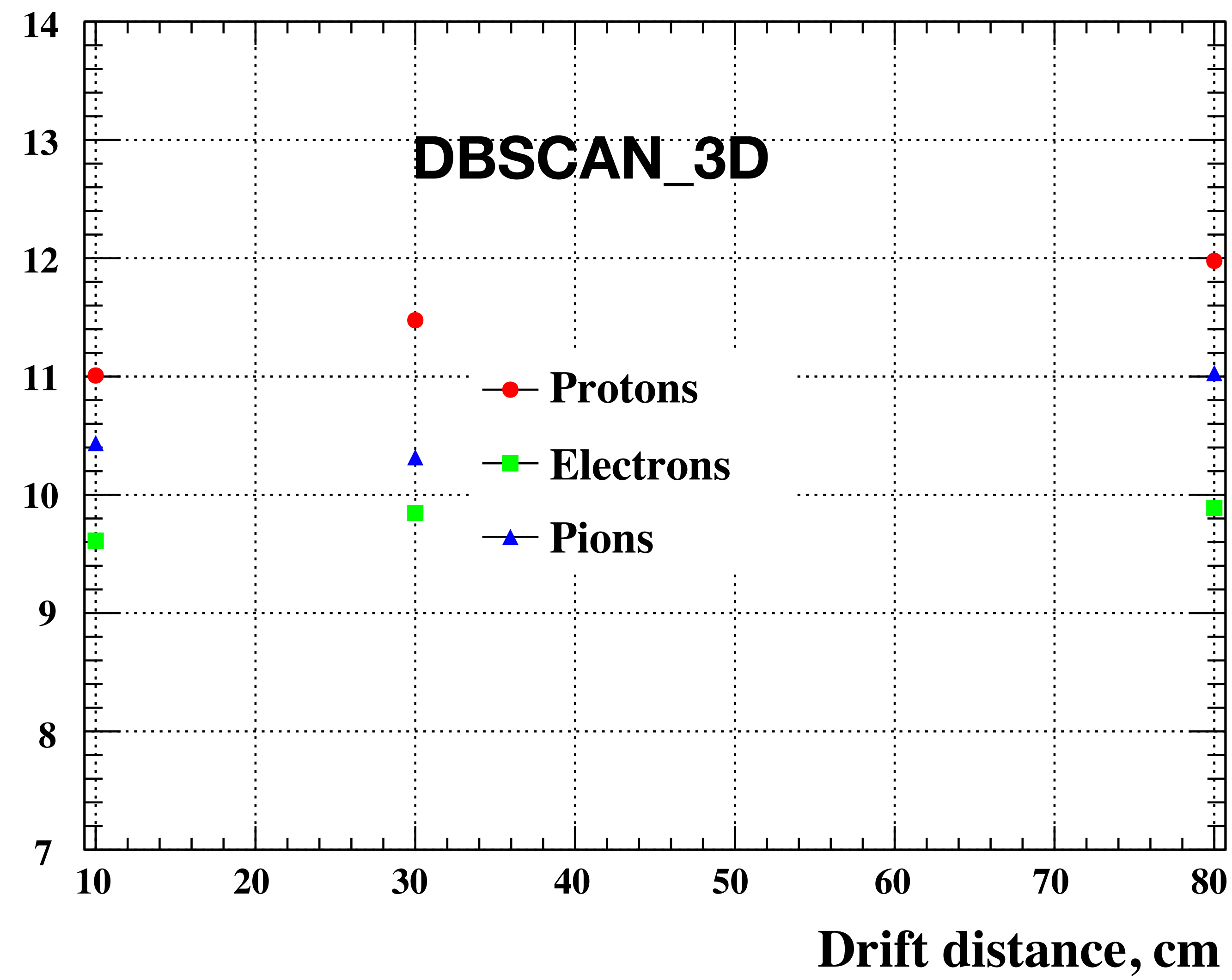
DBSCAN dE/dx comparison

55.5% More statistics



DBSCAN dE/dx comparison



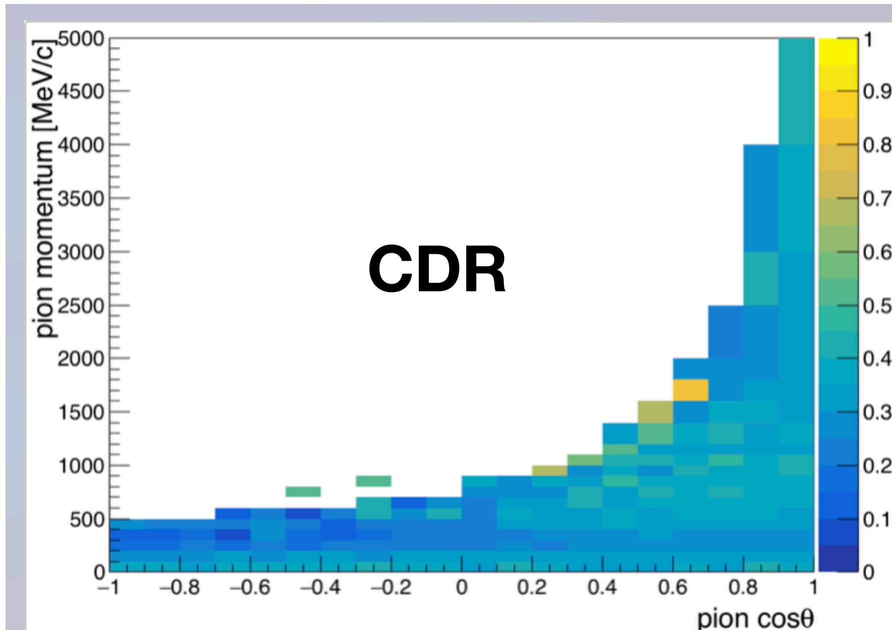


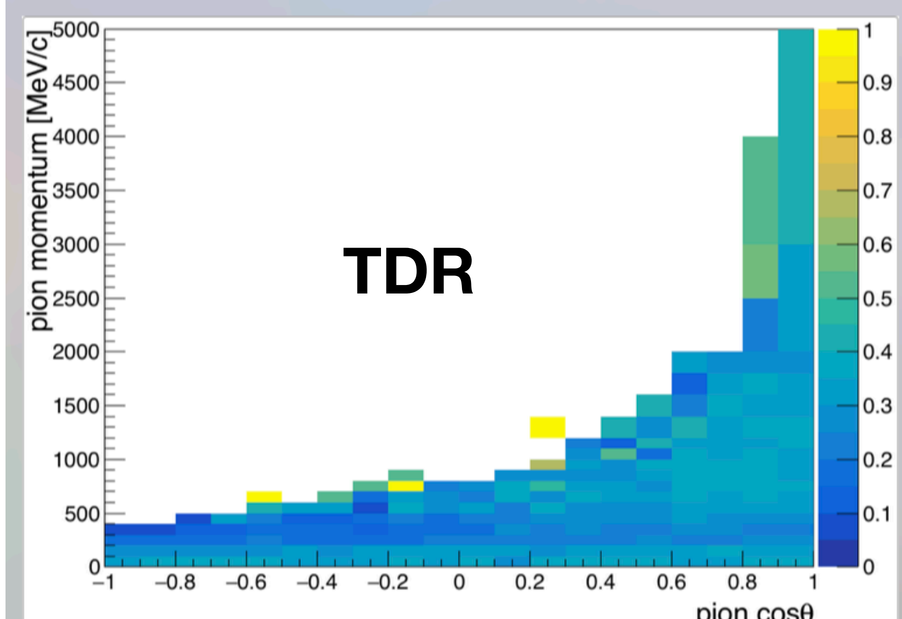
TDR / Sim&Opt

Pion Efficiencies, input for BANFF

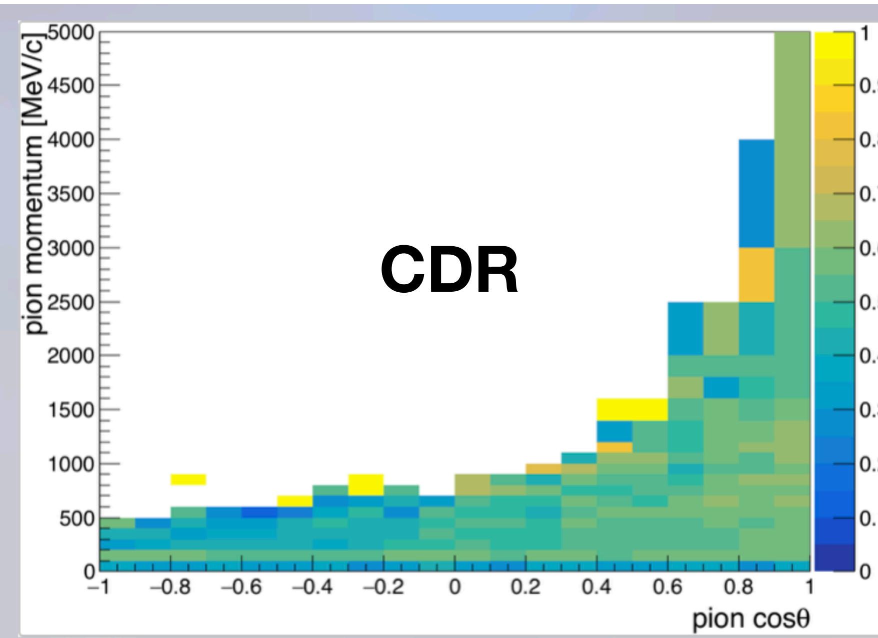
REMINDER FROM LAST PRESENTATION

Cesar





Mathieu/Simon



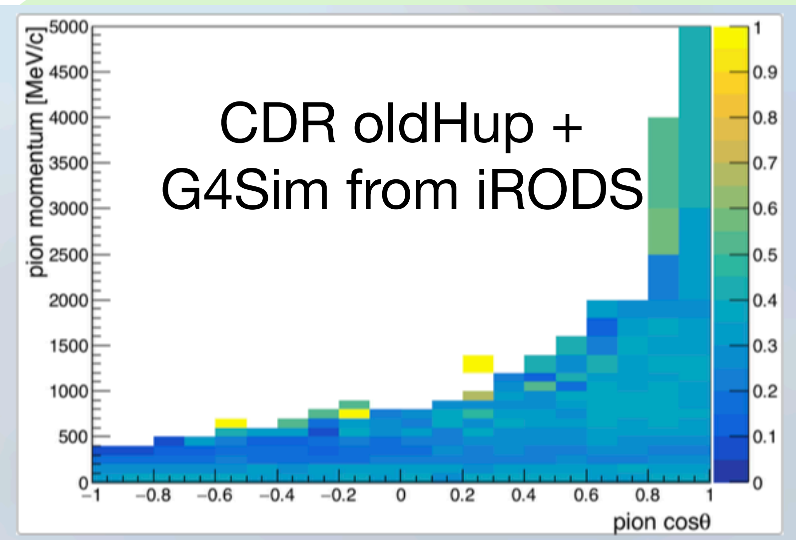
$\pi+ \rightarrow \pi+$

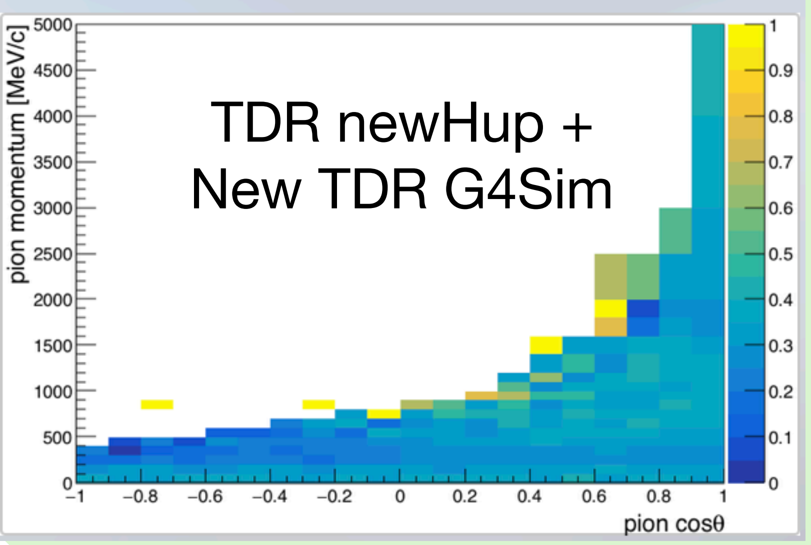
????

Simon / Mathieu
Help is necessary

Pion Efficiencies, input for BANFF

REMINDER FROM LAST PRESENTATION



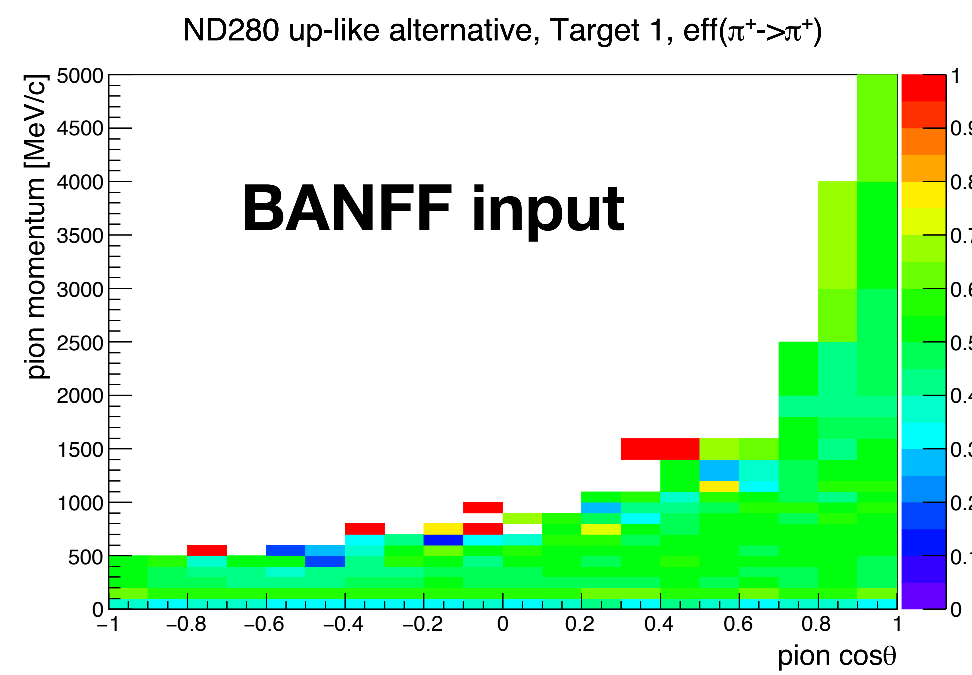


Pion eff do match

Therefore Highland seems to be doing the job as before.

I checked commits:
No suspicious modification in highland Concerning pions for more than 1 year.

Mathieu provided some input:



It seems to be in favor of what Simon had.
I will redo the colormaps for clarity soon.

Purity and selected events Tables

Purities and selected events tables

- As it must, current matches.
- For upgrade FGD1 match, SFGD is a bit different (we made some modifications, so its reasonable).
- FGD2, is not that clear. There is an obvious increase of the number of selected events. The increase in efficiency is 1% (uncertainty computed to be 0.2%).

Did we modified something that can explain this improvement?

NEW:

		# of events ($/10^{21}$ POT)	Purity (%) CC0π	CC1π	CC Other
current	FGD 1	50507	72.5%	64.0%	68.2%
	FGD 2	50125	71.5%	62.3%	63.8%
upgrade	FGD 1	52655	72.9%	64.1%	64.7%
	FGD 2	51460	71.6%	62.9%	63.3%
	SuperFGD	95490	72.5%	70.3%	72.7%
	SuperFGD*	135561	62.0%	66.7%	58.8%

Mathieu's talk:

		# evts ($/10^{21}$ POT)	purity (in %) CC0π	CC1π	CCother
current	FGD 1	50500	72.6%	64.2%	67.9%
	FGD 2	50161	71.4%	62.9%	63.5%
upgrade	FGD 1	52518	72.3%	63.1%	63.7%
	FGD 2	50677	70.6%	62.5%	61.3%
	Horiz.Target	101858	74.1%	73.4%	70.6%
	Horiz.Target*	141080	64.0%	69.7%	57.9%

Purities and selected events tables

I have nothing to compare with this 2 samples

			# of events (/10^{21} POT)	Purity (%) CC0π	CC1π	CC Other
RHC $\bar{\nu}_\mu$	current	FGD 1	16504	77.8%	76.3%	53%
		FGD 2	16167	78.2%	76.7%	54.5%
	upgrade	FGD 1	16487	77.9%	76.9%	54%
		FGD 2	16181	77.8%	77.4%	54.3%
		SuperFGD	28095	75%	78.7%	61.1%
		SuperFGD*	135561	62.0%	66.7%	58.8%
RHC ν_μ	current	FGD 1	8379	56.3%	58.1%	70.8%
		FGD 2	8158	54.9%	58.2%	67.6%
	upgrade	FGD 1	8373	56%	57.4%	70.1%
		FGD 2	8111	55.6%	57.4%	67.8%
		Horiz.Target	13109	55.4%	62.1%	76.9%

Purities and selected events tables

Ignore red, is just a warning.
Do not compare results with CDR

		# of events ($/10^{21}$ p.o.t.)	Purity CC0π	CC1π	CC Other
Current	FGD1	47337	75.9%	64.4%	61.8%
	FGD2	45939	75.7%	65.1%	64.4%
Upgrade	FGD1	48374	74.7%	64.5%	70.2%
	FGD2	45719	73.4%	63.8%	70.1%
	SuperFGD	100295	74.1%	72.9%	70.1%

CDR:

Selection	Current-like	Upgrade-like
ν_μ (ν beam)	93,401	194,654
$\bar{\nu}_\mu$ ($\bar{\nu}$ beam)	33,437	63,687
ν_μ ($\bar{\nu}$ beam)	17,998	33,773

		# evts ($/10^{21}$ POT)	purity (in %) CC0π	CC1π	CCother
current	FGD 1	50500	72.6%	64.2%	67.9%
	FGD 2	50161	71.4%	62.9%	63.5%
upgrade	FGD 1	52518	72.3%	63.1%	63.7%
	FGD 2	50677	70.6%	62.5%	61.3%
	Horiz.Target	101858	74.1%	73.4%	70.6%
	Horiz.Target*	141080	64.0%	69.7%	57.9%

TDR:

Selection	Current-like	Upgrade-like
ν_μ (ν beam)	100632	199605
$\bar{\nu}_\mu$ ($\bar{\nu}$ beam)	32671	60763
ν_μ ($\bar{\nu}$ beam)	16537	29593

We should compare with this more recent numbers!
Mathieu talk March 14th, 2018
Nu beam (FHC)
Current like: 100661
Upgrade like (tpc-ecal): 205 053

Purities and selected events tables

Ignore red, is just a warning. Do not compare results with CDR

- Why the numbers for RHC are so much better???

Mathieu's talk:

		CC-inclusive	$CC-0\pi$	$CC-1\pi$	$CC-$other
ν_μ (FHC)	TPC+ECal	97.0%	74.1%	73.4%	70.6%
	Target	67.1%	35.1%	62.4%	31.1%
$\bar{\nu}_\mu$ (RHC)	TPC+ECal	85.8%	71.1%	31.4%	24.5%
	Target	25.3%	13.9%	2.9%	4.4%
ν_μ (RHC)	TPC+ECal	80.3%	43.4%	39.3%	61.0%
	Target	17.0%	7.4%	5.9%	11.3%

Mathieu's opinion:

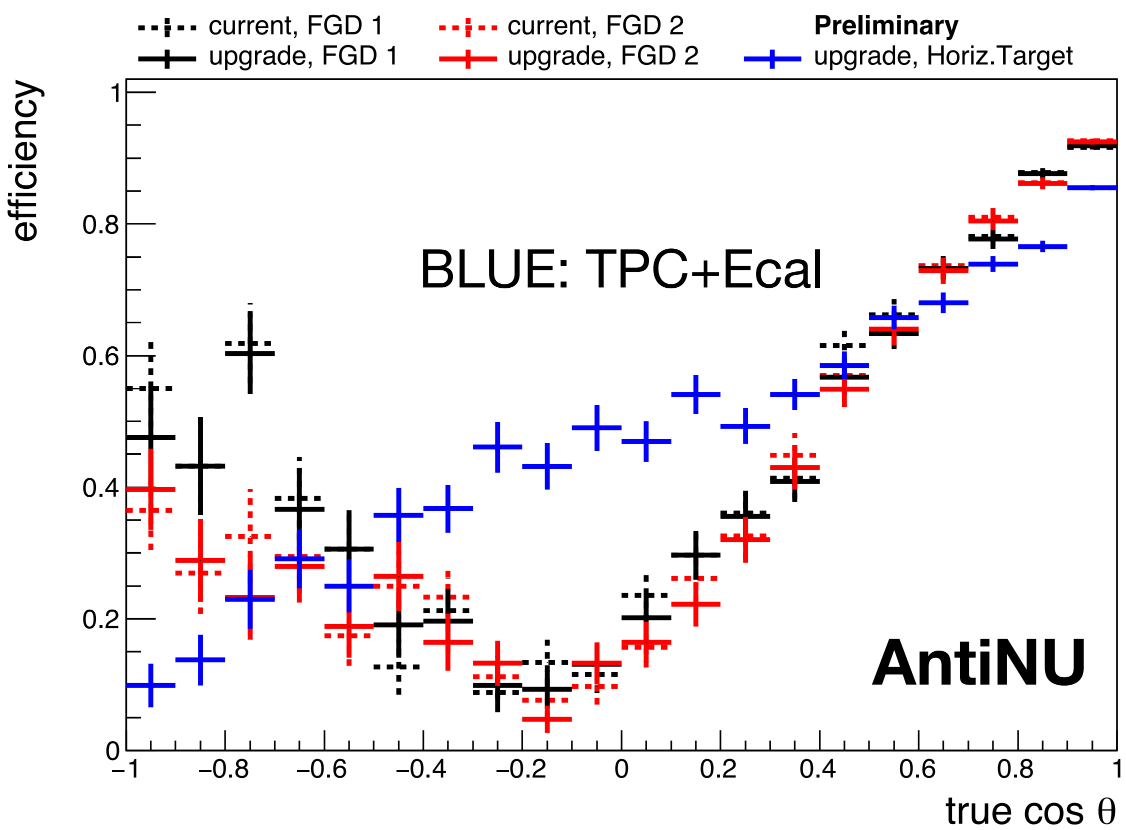
"improvement in TPC+ECal for RHC may be related to the fact that we were using potentially bugged production for RHC at that time (you can ask to Davide but I think the story was that FHC was reprocessed after bug discovery but not RHC)"

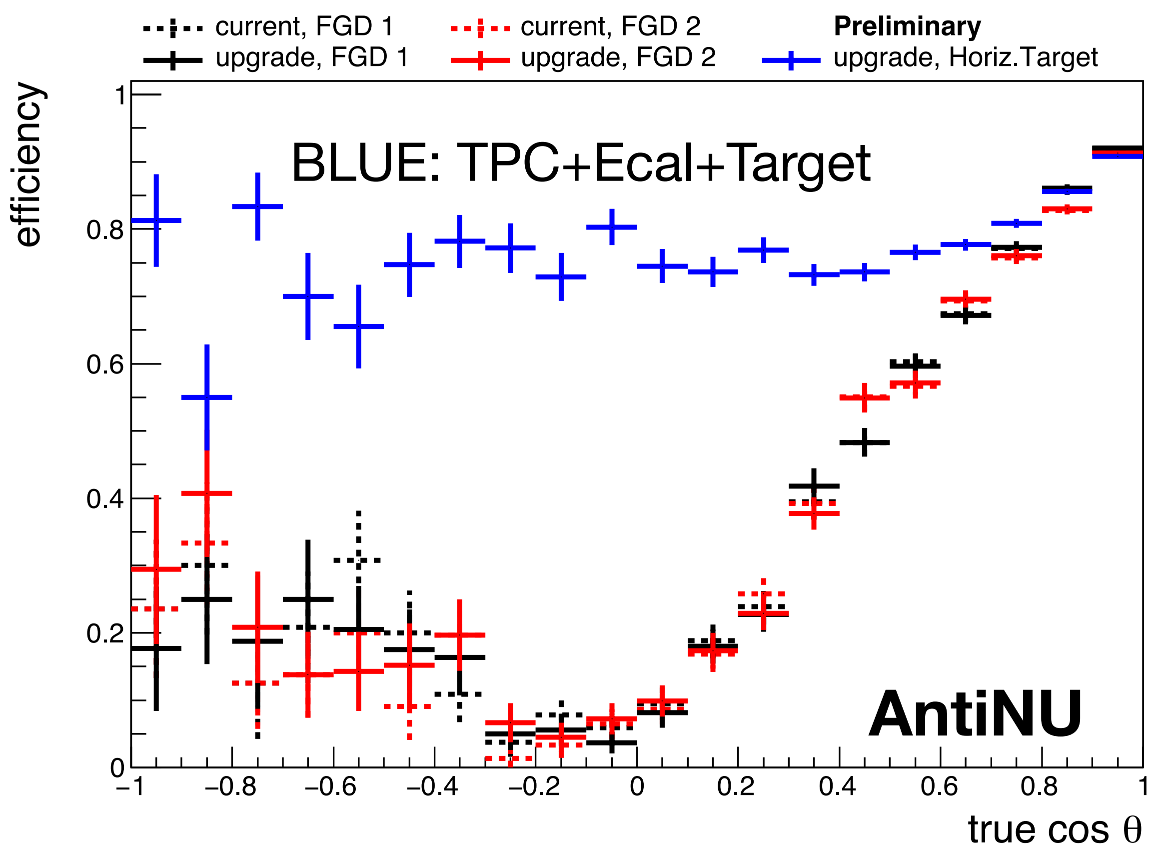
TDR:

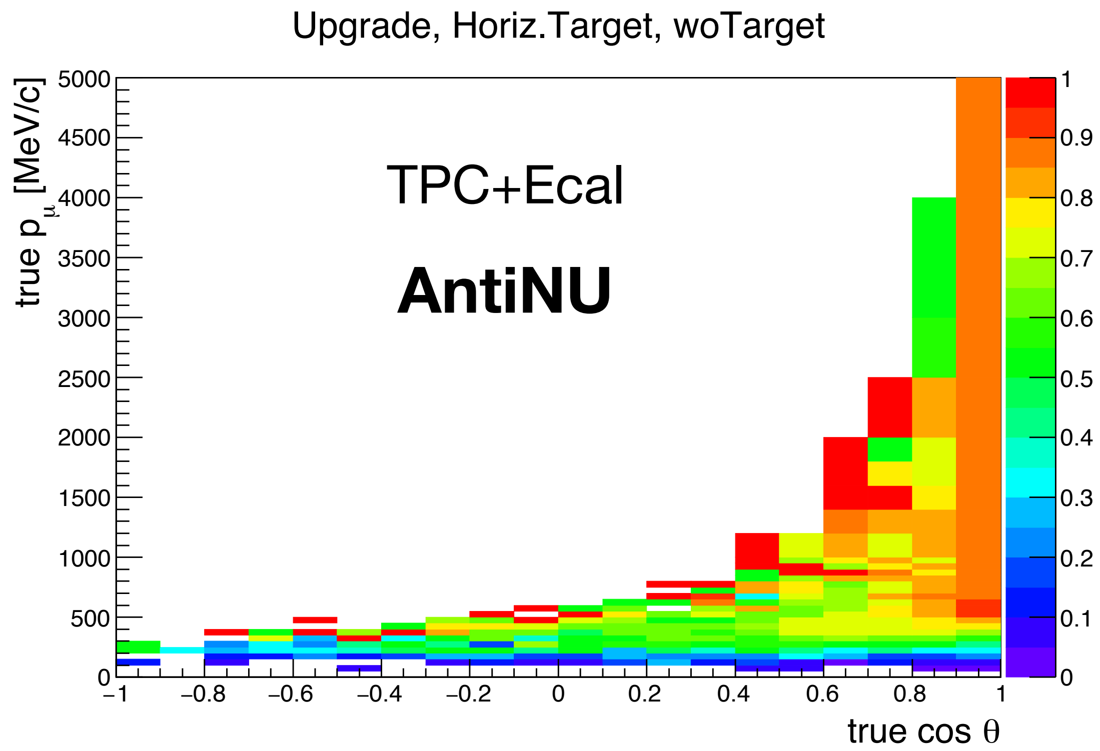
		CC-inclusive	$CC0\pi$	$CC1\pi$	CC Other
ν_μ (FHC)	TPC+ECal	96.8%	72.5%	70.3%	72.7%
	Target	65.9%	34.7%	60.0%	31.0%
$\bar{\nu}_\mu$ (RHC)	TPC+ECal	97.6%	75.0%	78.7%	61.1%
	Target	52.7%	19.1%	45.5%	12.9%
ν_μ (RHC)	TPC+ECal	94.4%	55.4%	62.1%	76.9%
	Target	50.0%	19.5%	36.8%	32.7%

Efficiency plots for RHC

Efficiency plots for RHC



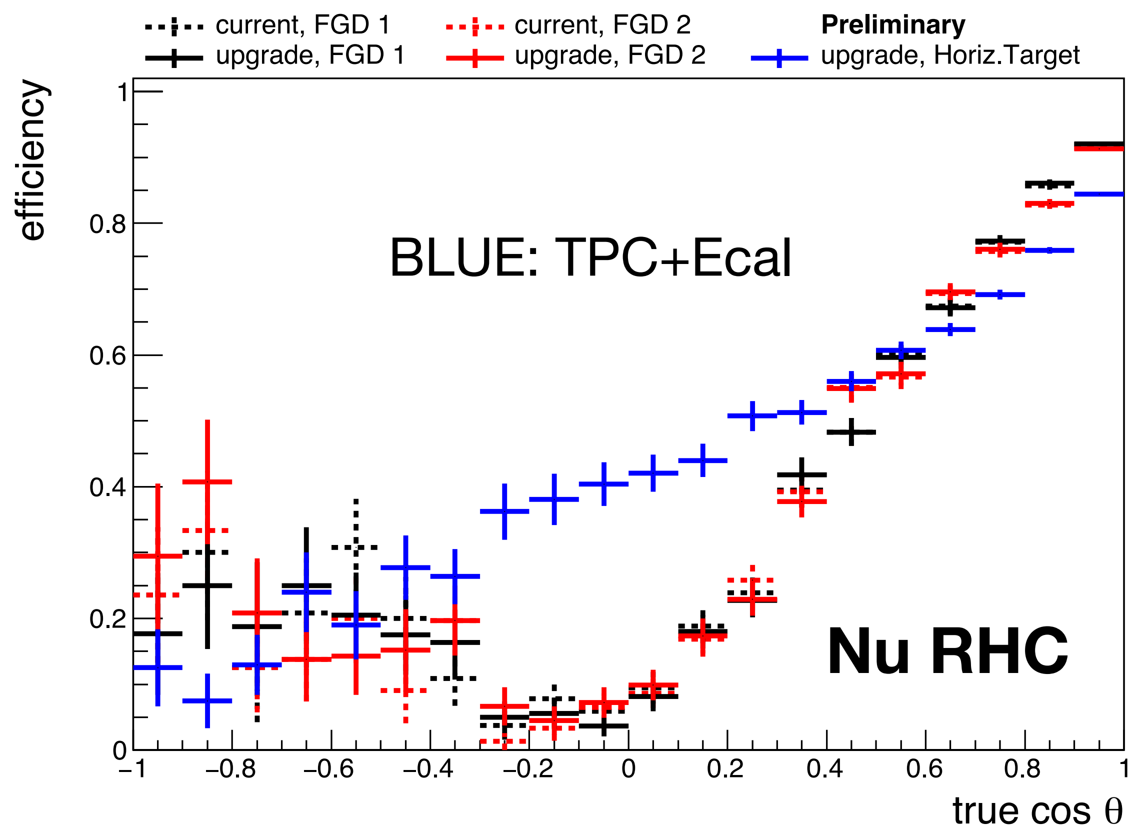


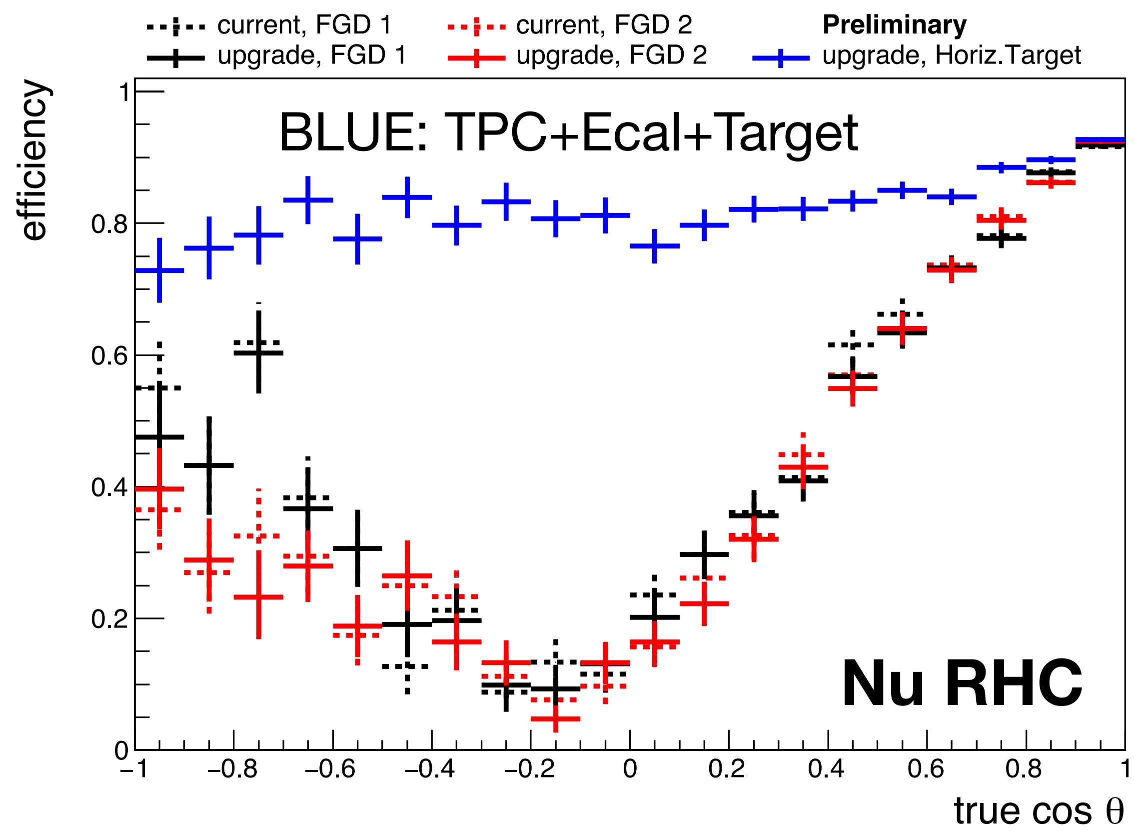


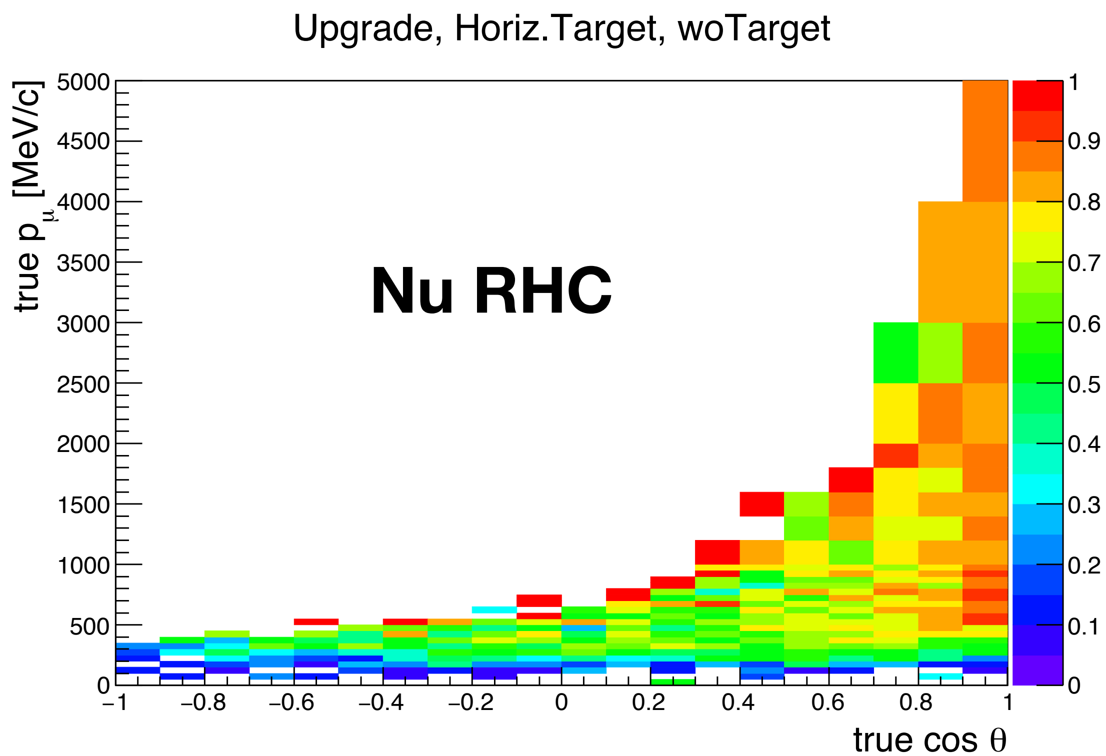
2D TPC+Ecal+Target

Missing! Do you want it?

Efficiency plots for RHC







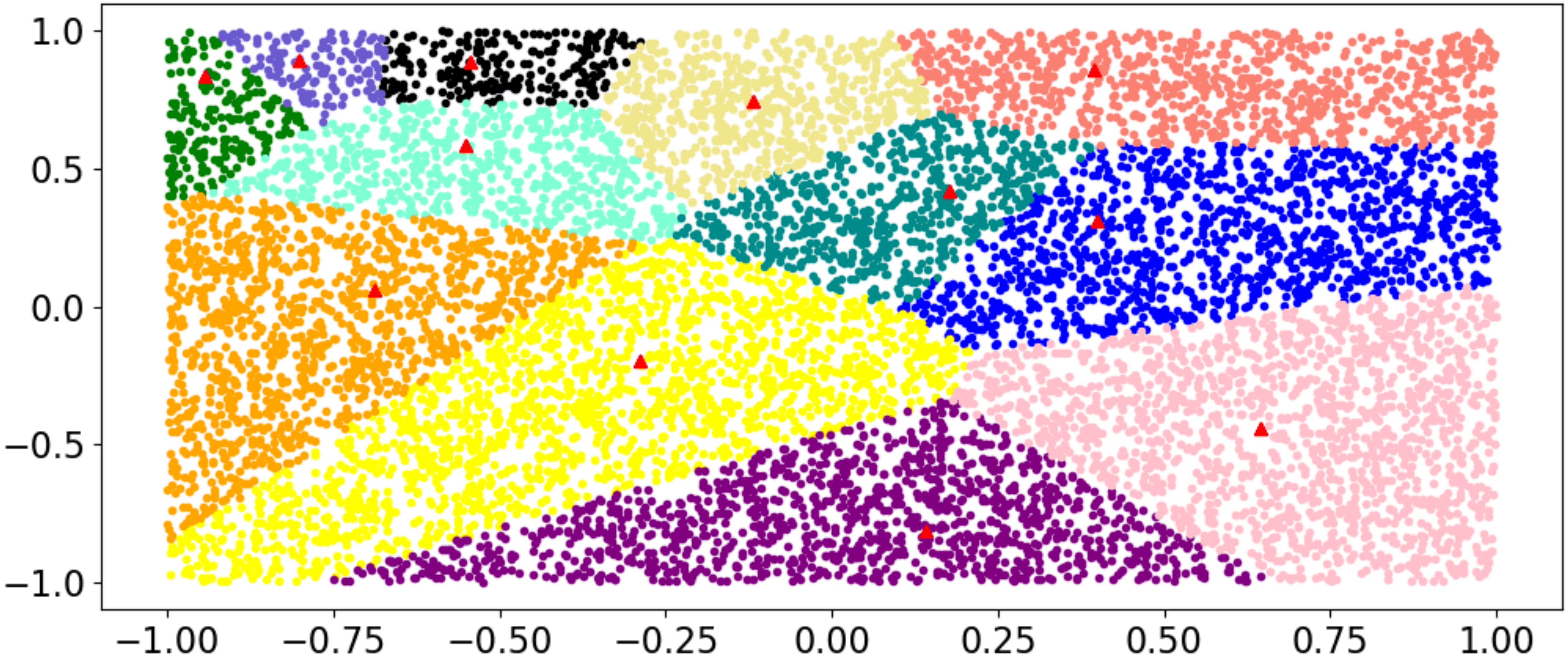
2D TPC+Ecal+Target

Missing! Do you want it?

Additional activities

Additional activities:

Learning python & ML basics.
Learning some notions in chaos theory.
Mounting IKEA furniture: home miniLab is closer :)

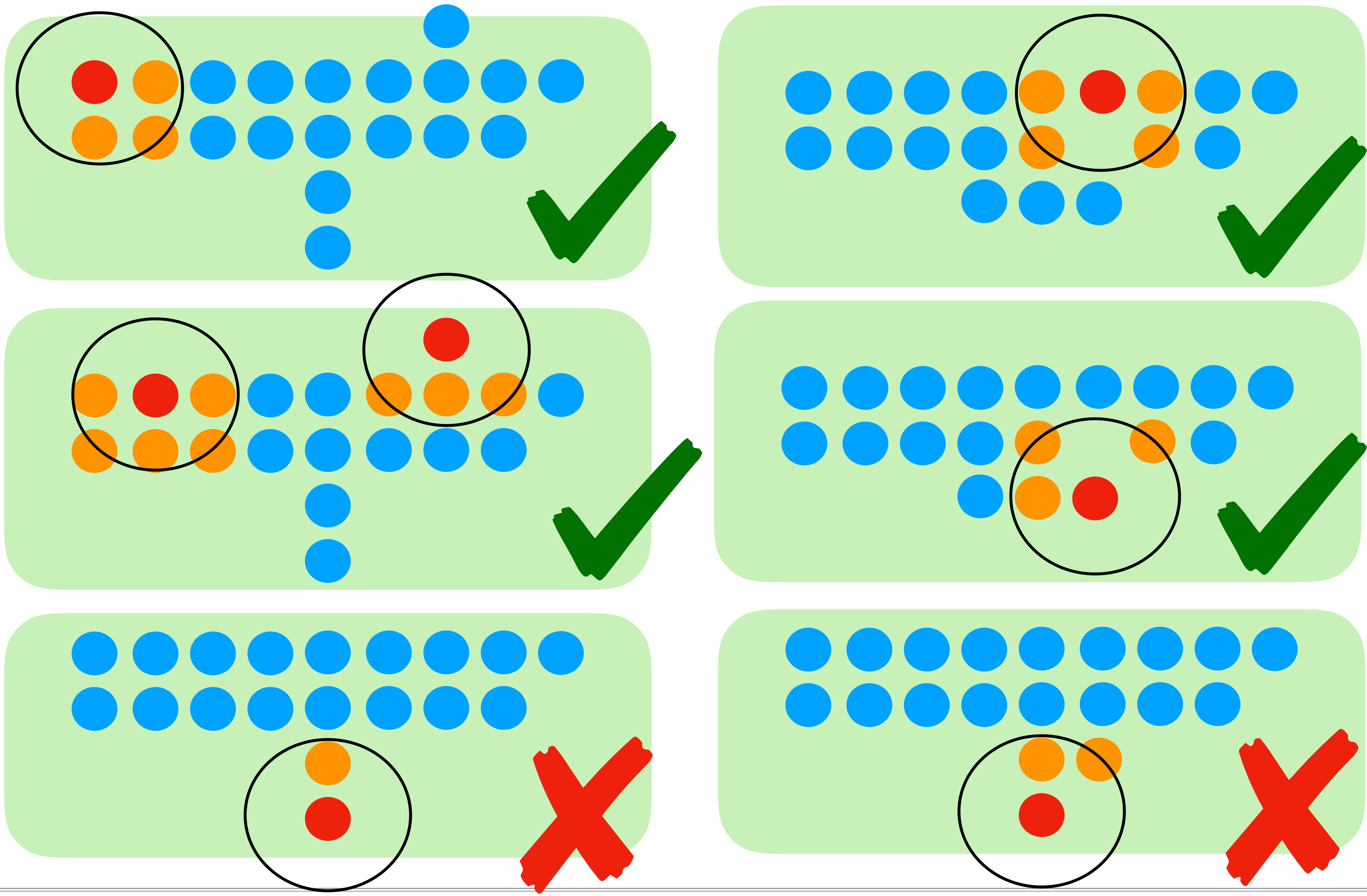


BACK UP slides

DBSCAN Parameters selection (2D)

Origin
Selected
Others

DIST ~SQRT(2) & minHITS 3



DBSCAN Parameters selection (2D)

This looks very nice … and naive!
Looking only to one situation is potentially DANGEROUS.

Parameters should work universally for all tracks and data samples.
Let's take a closer look…

DBSCAN Parameters selection (2D)

Reduce minHITS from 2 to 1?

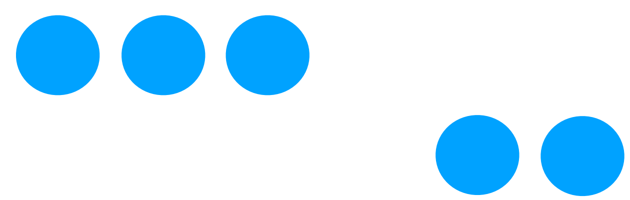
Summary

Along this presentation we have seen that:

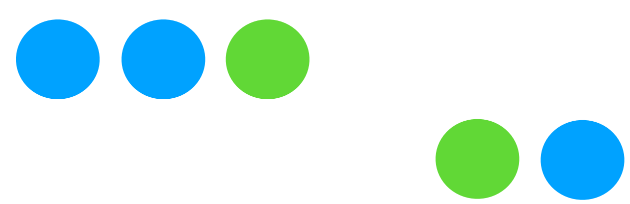
DBSCAN is physical, namely all its parameters minHITS, DIST, and f have a very intuitive physical interpretation.

The intuition about the parameters allow us to discuss the different geometries of hit tracks, and therefore impose:

minHITS = 1 to not cut first and last point of thin tracks
f = 12.5 to penalize as much being on the edge of electronics window (2.5μs) than being in non-likely spatial configuration:



DIST (including time coordinate) = 5 to allow this configuration with green dots at 2.5μs time difference to be selected



Summary

The results supersede the TestSelection3D in data quality (does not miss charge spreading, and is fully 3D in the sense of able to modify time structure, without allowing track merging).

Now multitrack selection is supported. The results from DBSCAN supersede statistics from TestSelection3D. ~55.5% more statistics.

I will develop this week a visualization macro to check in a evt by evt basis the original information, the information of Sergey's selection and DBSCAN selection. Thus you can see by eye what is doing and what is not doing each selection.

At this point, I deeply rely on DBSCAN selection, even if it could be improved a bit if necessary.

PRF Info is missing sorry.

Unrelated comment

We are considering maxADC for the analysis. However, when we move further in the chamber the signal broadens and the maxADC is lower !

This explains the movement in dEdx resolution.