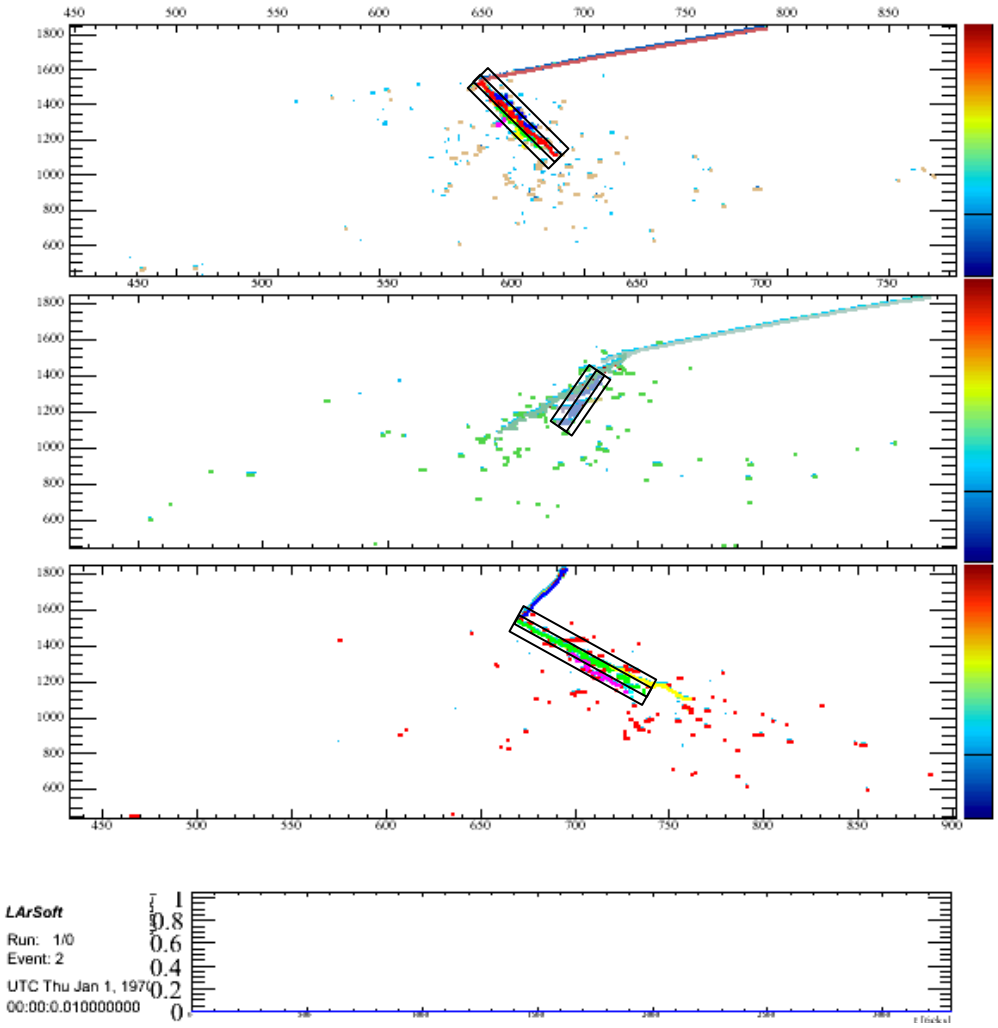
Status of clustering

Ben Carls

Status as of last update

- Employing fuzzy clustering, followed by merging of nearby clusters
- Using Hough line finder to identify tracks, spin off into new clusters
- Vetoed Hough lines on isolation criteria, this resulted in too many good lines getting vetoed

The older method of dealing with Hough lines in electron showers

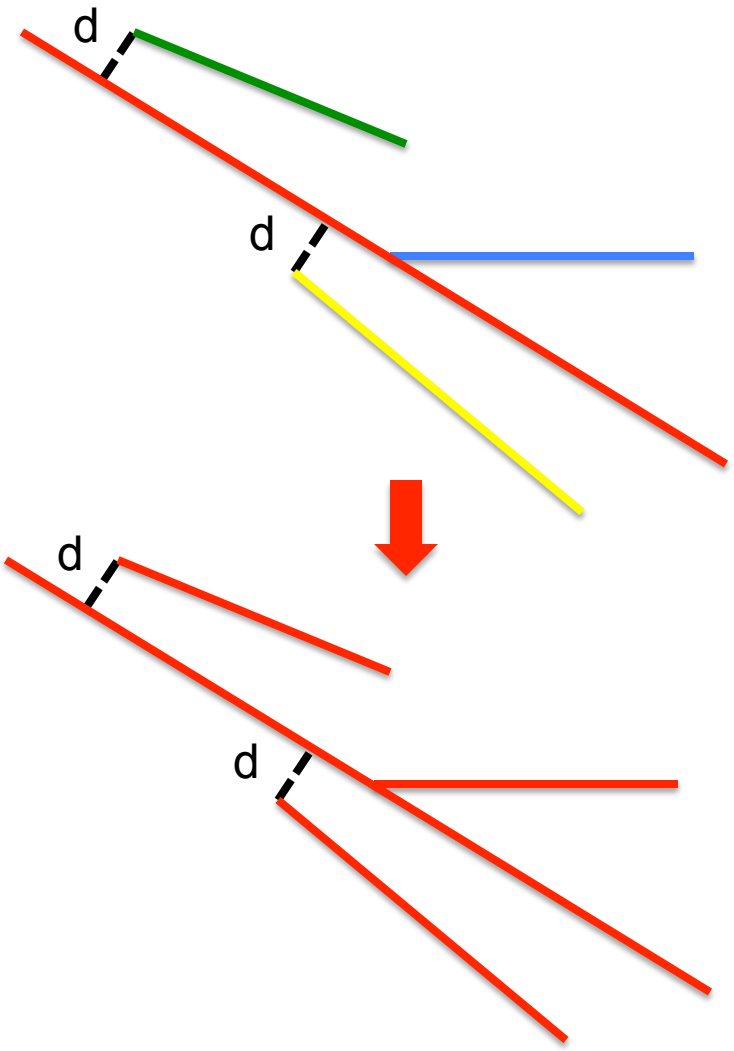


Hough line finder identifies lines in electron showers, need to veto

Examining area around lines and checking isolation, applying a cut

This cut is applied after the lines have been merged

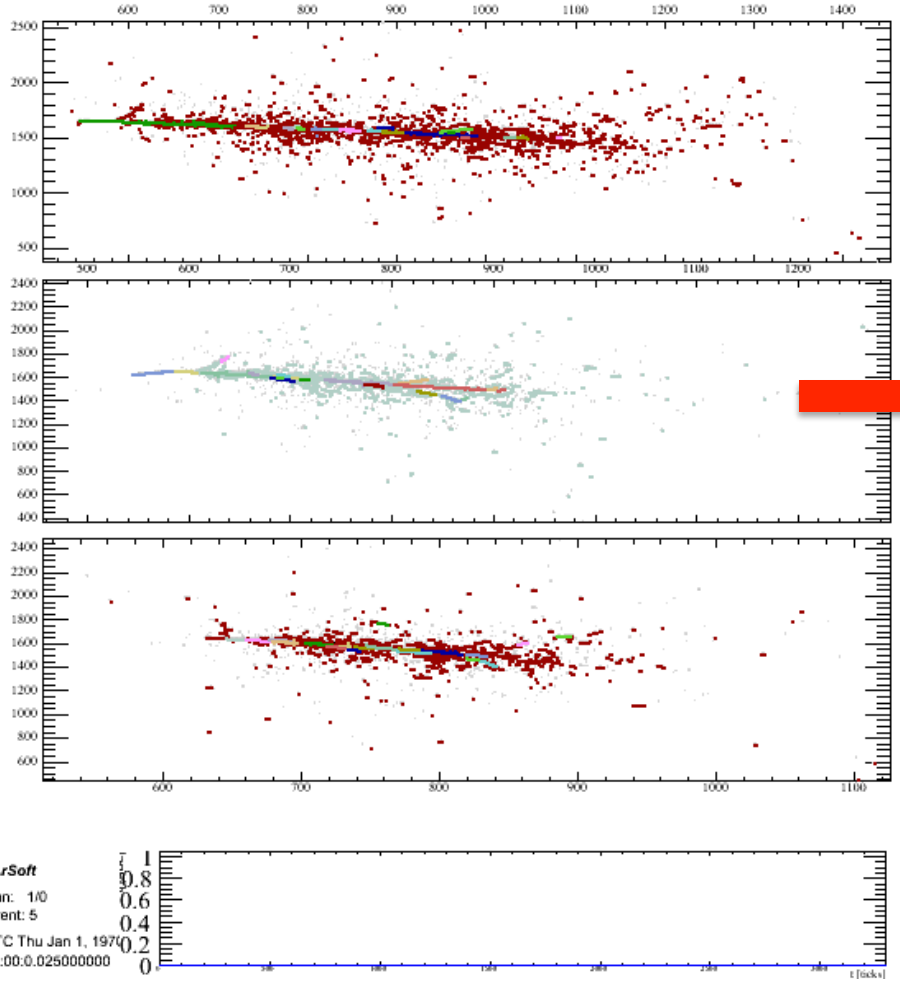
Showers with Hough lines

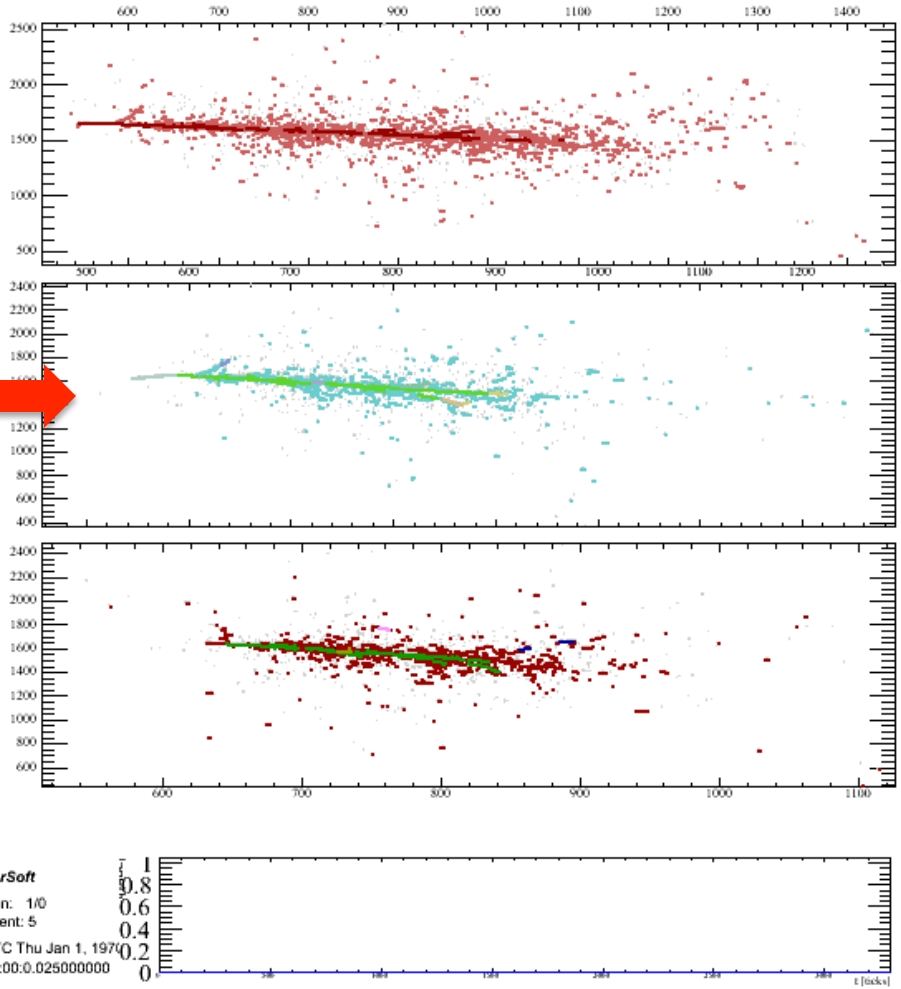


Hough line finder very effective at finding lines in showers

Now aiming to capitalize on this, merging Hough lines together to construct showers

Merge Hough lines using the distance between the line segments and angle between slopes $< 20°$

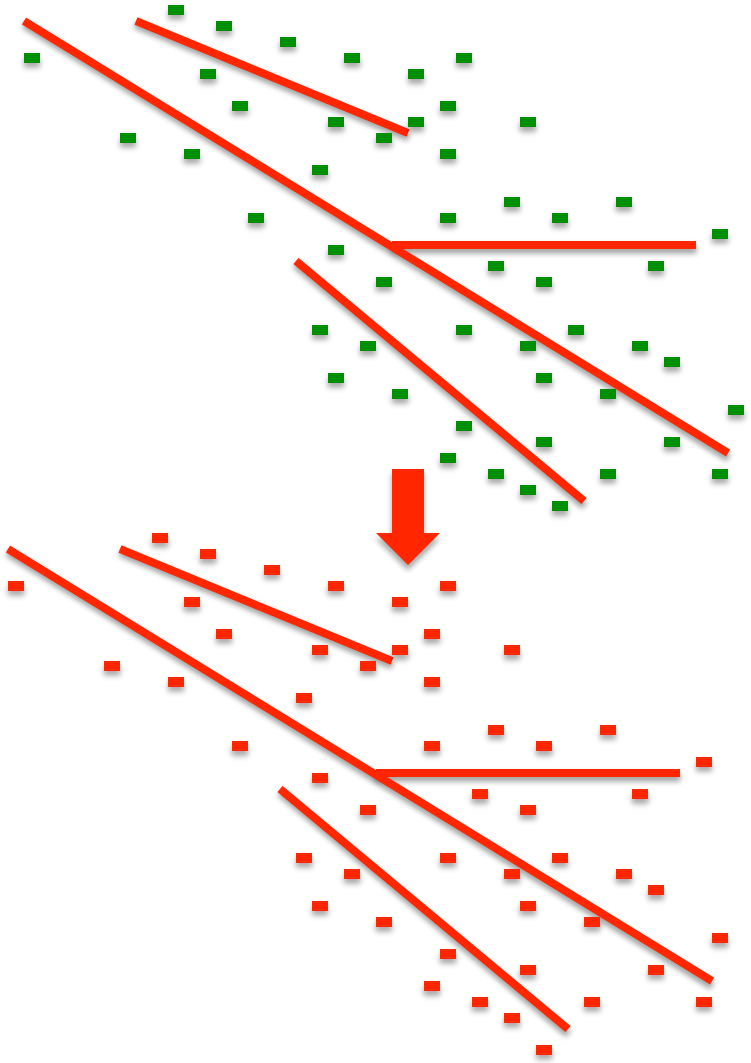




LArSoft
Run: 1/0
Event: 5
UTC Thu Jan 1, 1970
00:00:0.025000000

LArSoft
Run: 1/0
Event: 5
UTC Thu Jan 1, 1970
00:00:0.025000000

The fuzzy cluster remnants

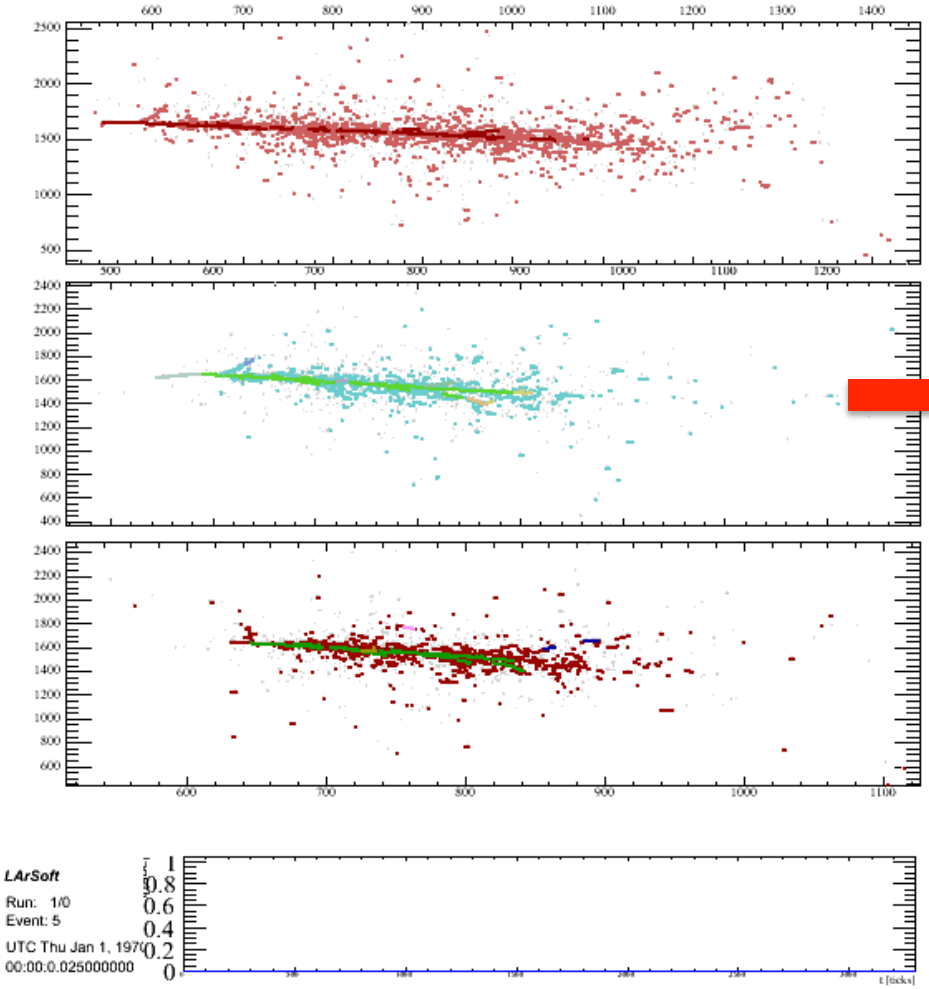


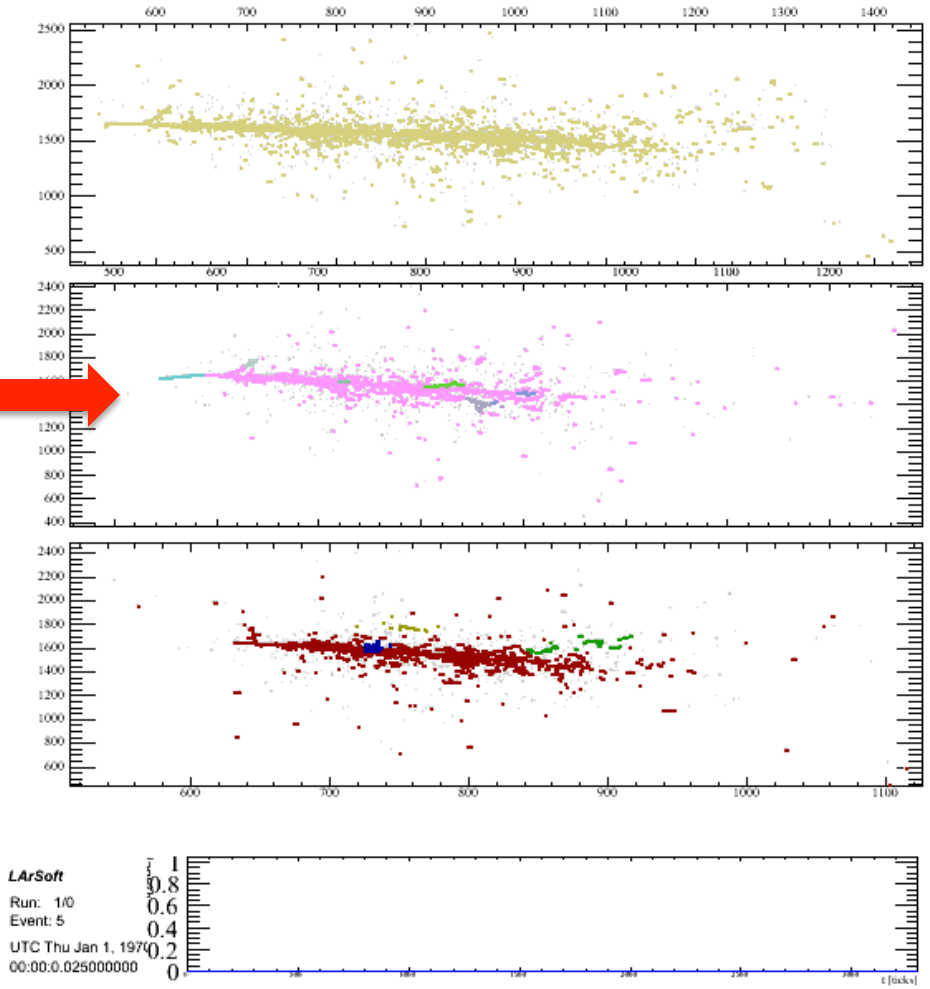
The issue remains with what to do with the left over fuzzy cluster points

Currently merging these points into the nearest Hough line segment using:

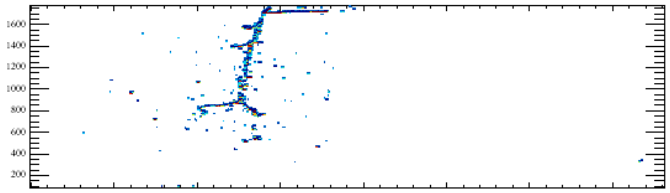
$$d_{weighted} = \frac{distance\ point\ to\ line\ segment}{length\ of\ line\ segment}$$

Goal is to give longer line segments more weight


LArSoft
Run: 1/0
Event: 5
UTC Thu Jan 1, 1970
00:00:0.025000000


LArSoft
Run: 1/0
Event: 5
UTC Thu Jan 1, 1970
00:00:0.025000000

Overview of the algorithm



Run fuzzy clustering

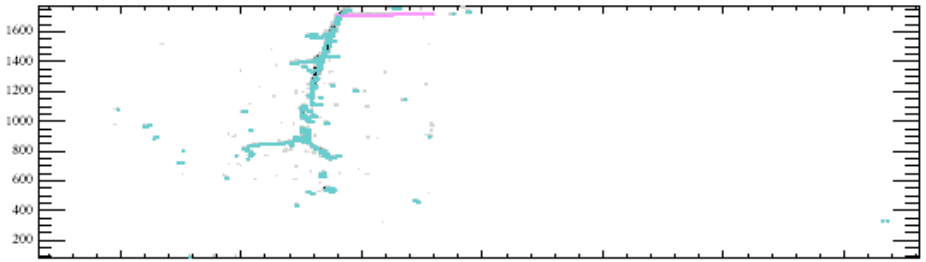
Check Xie-Beni index, determine number of clusters

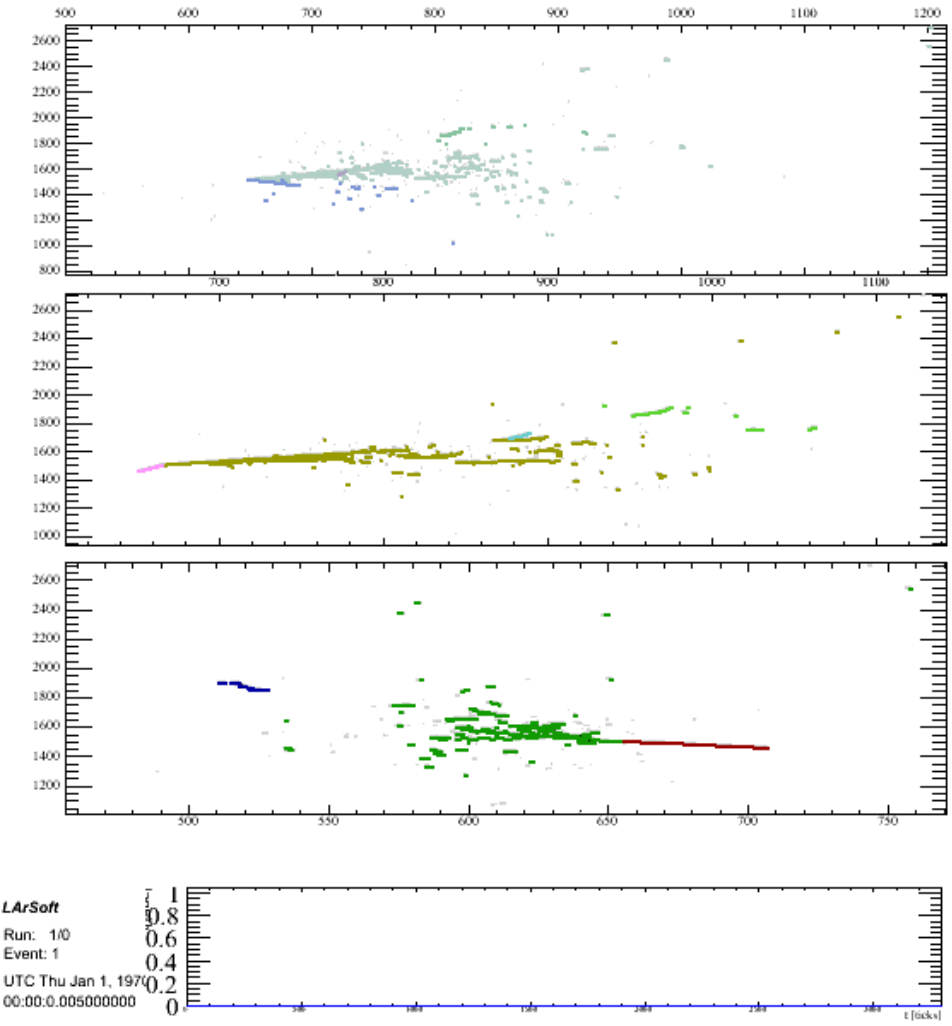
Merge nearby fuzzy clusters

Run Hough line finder

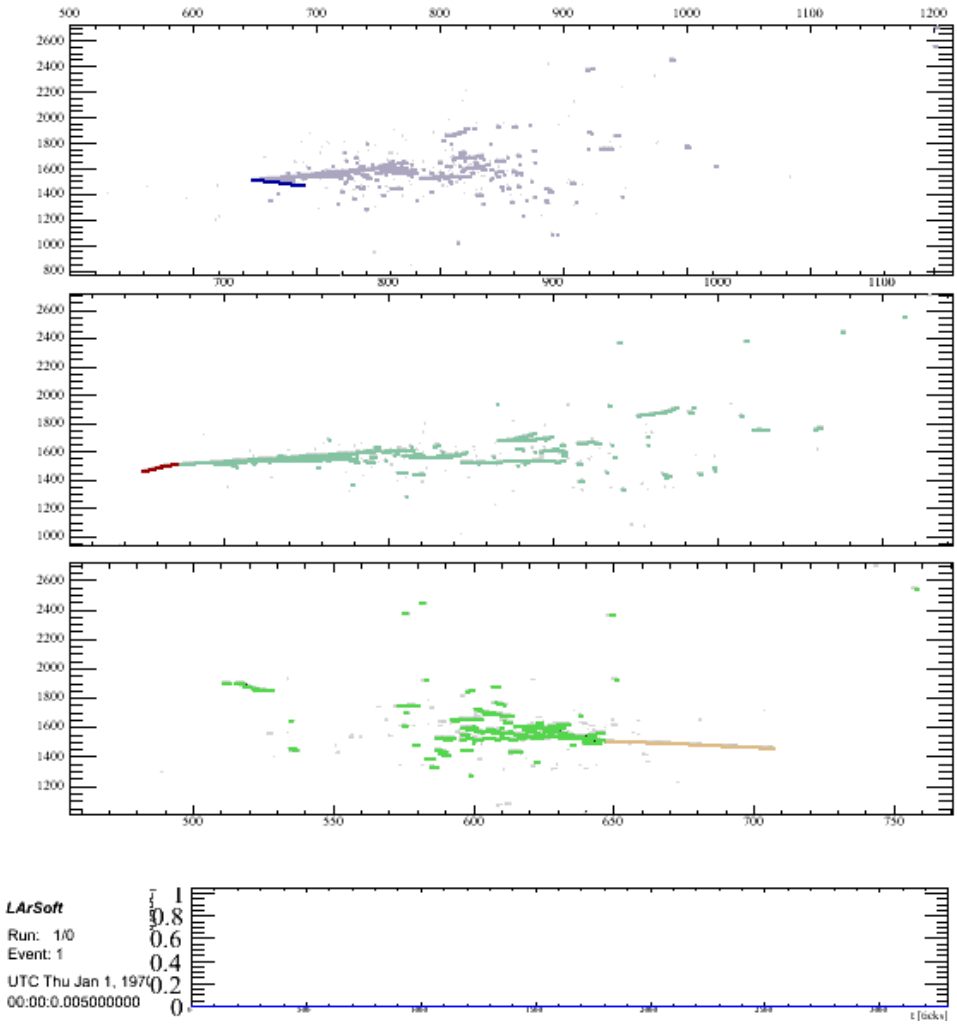
Merge Hough lines

Merge remaining hits with Hough lines

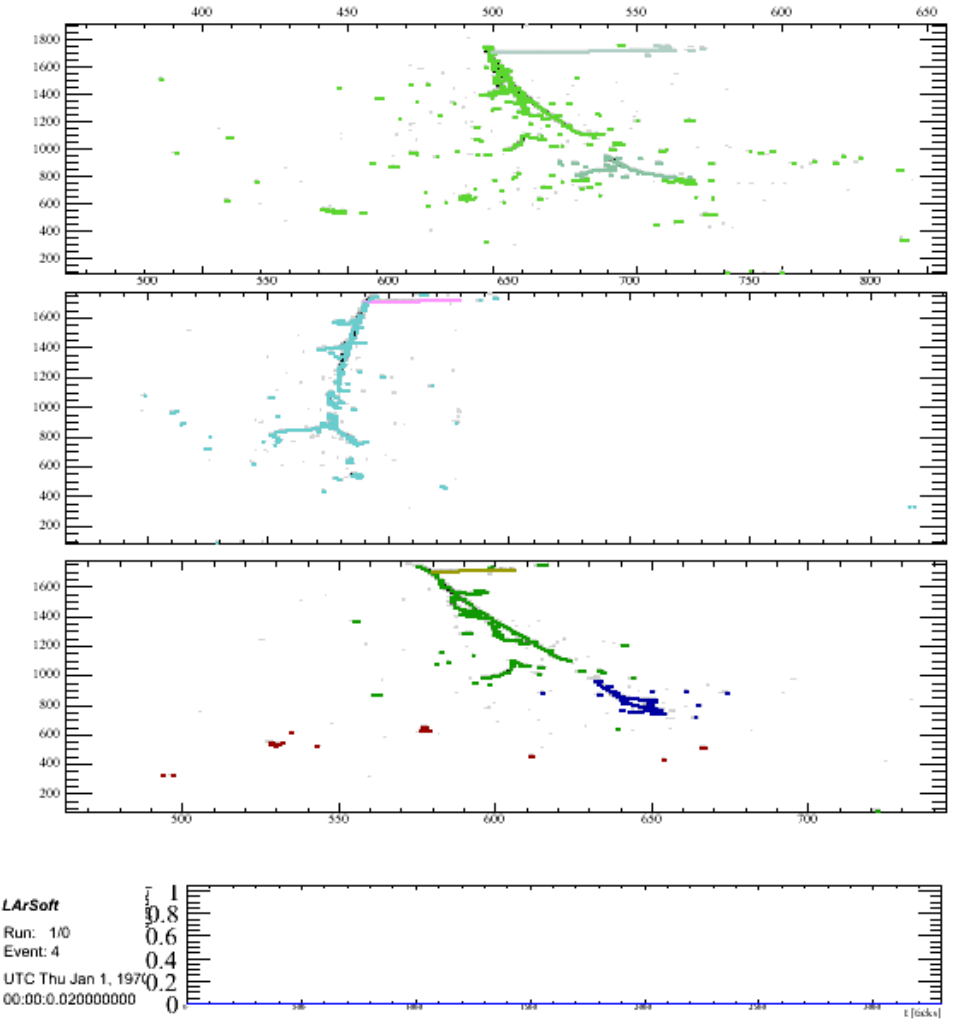




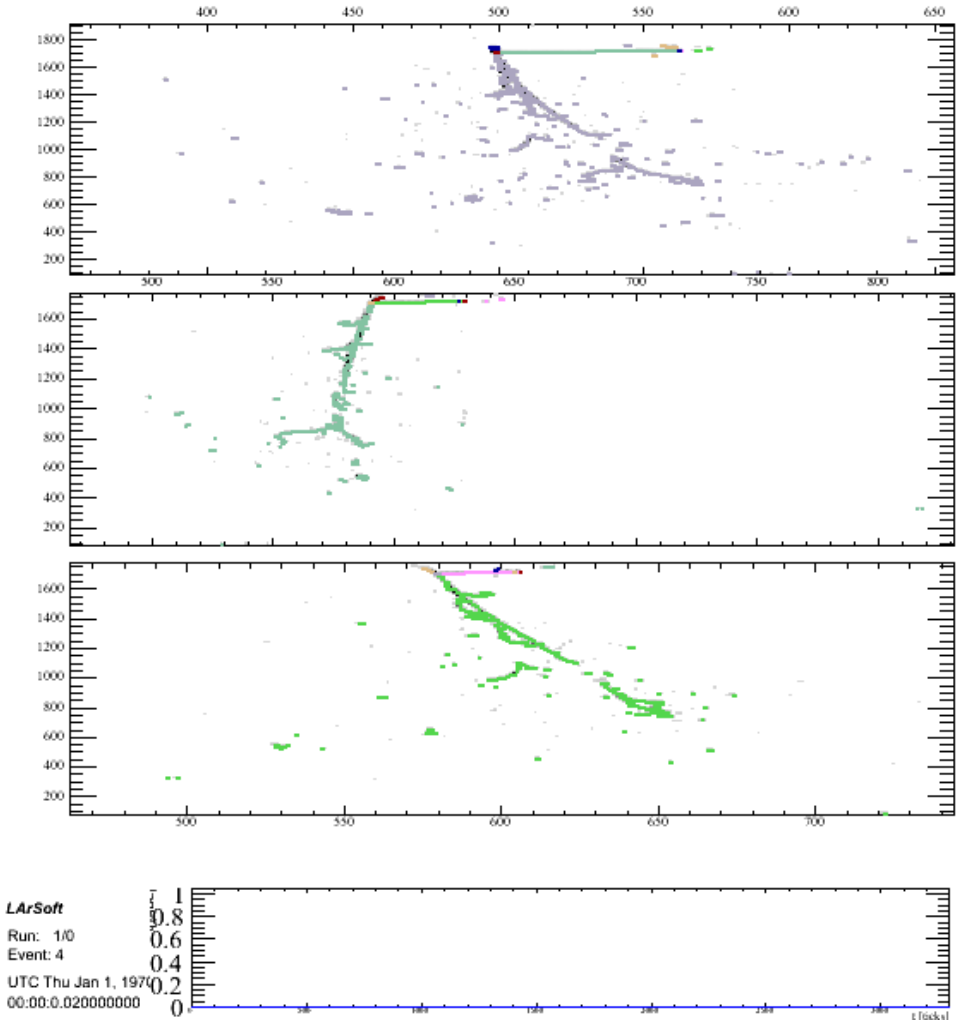
Fuzzy and Hough line clustering



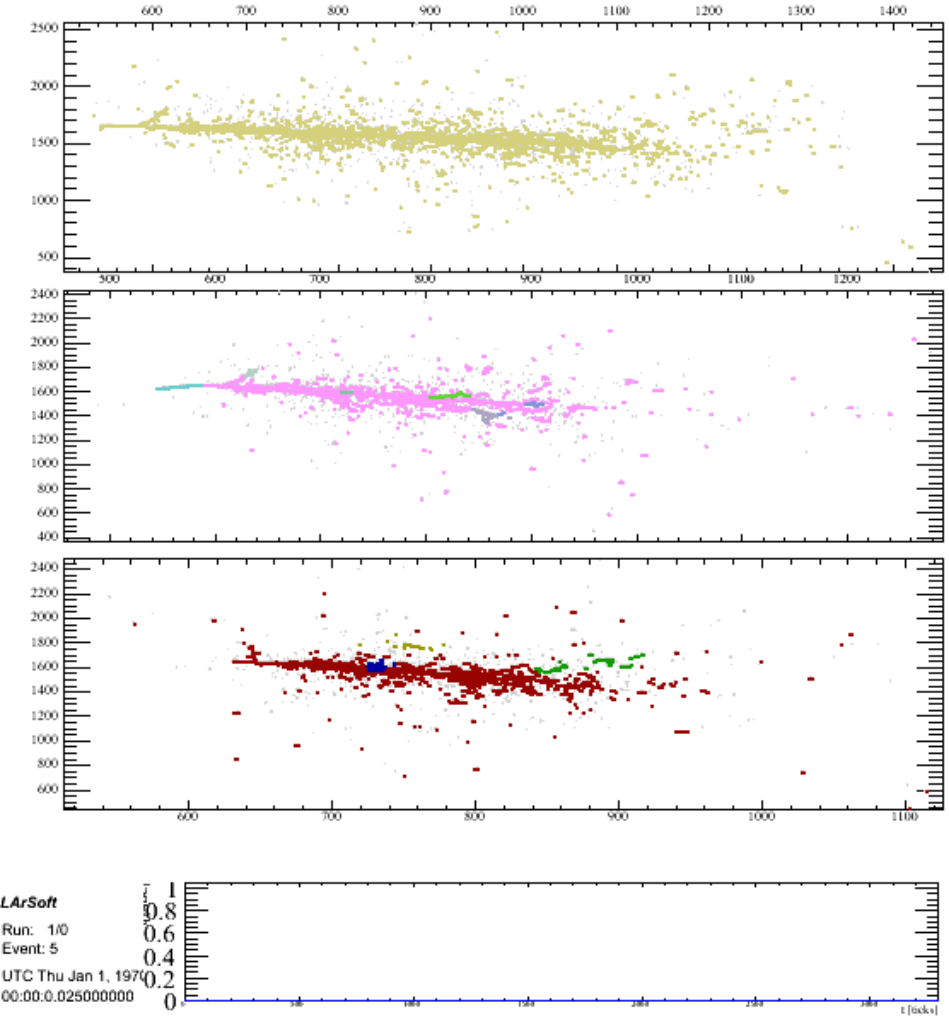
Cluster cheater



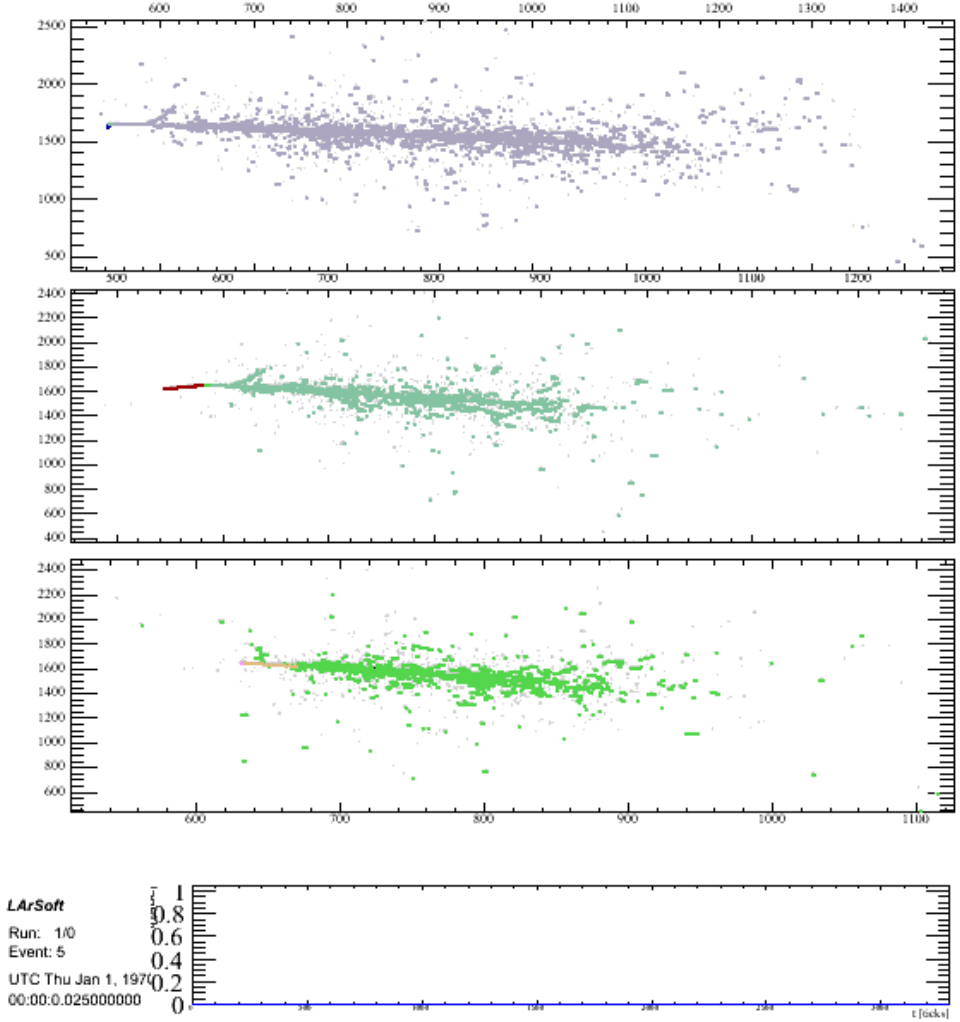
Fuzzy and Hough line clustering



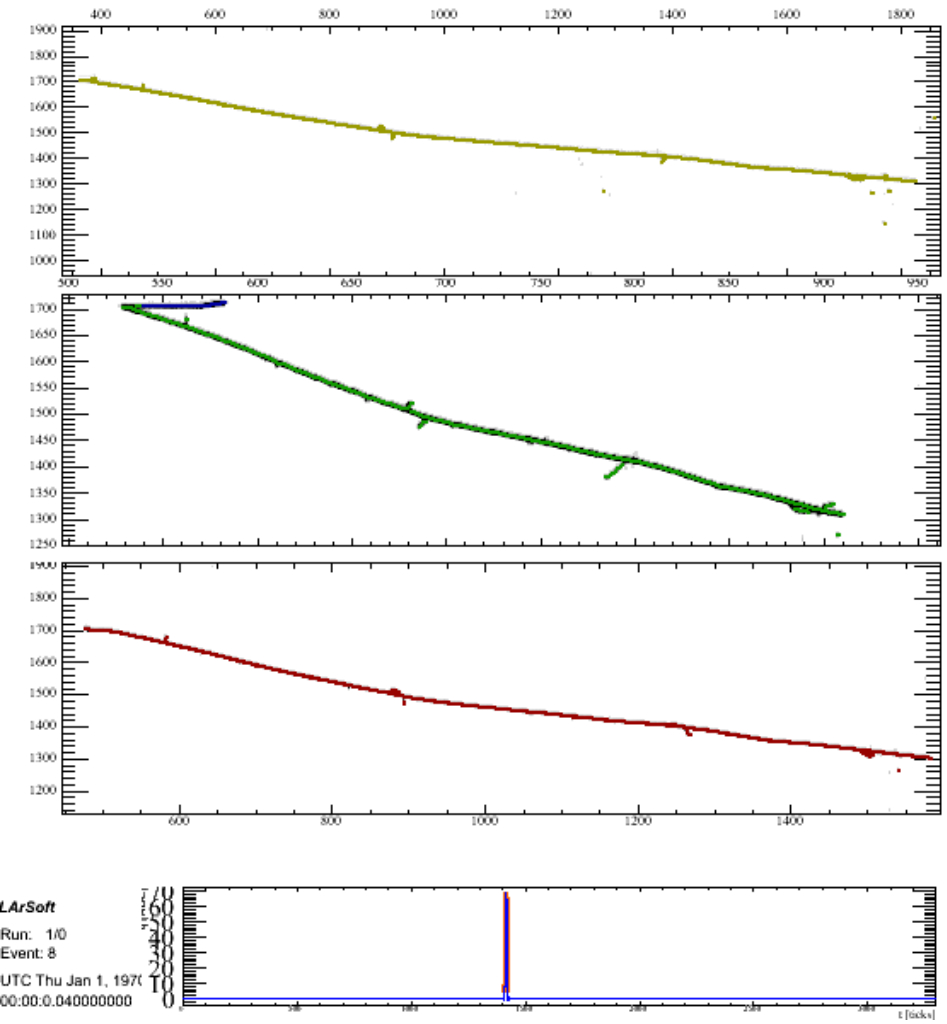
Cluster cheater



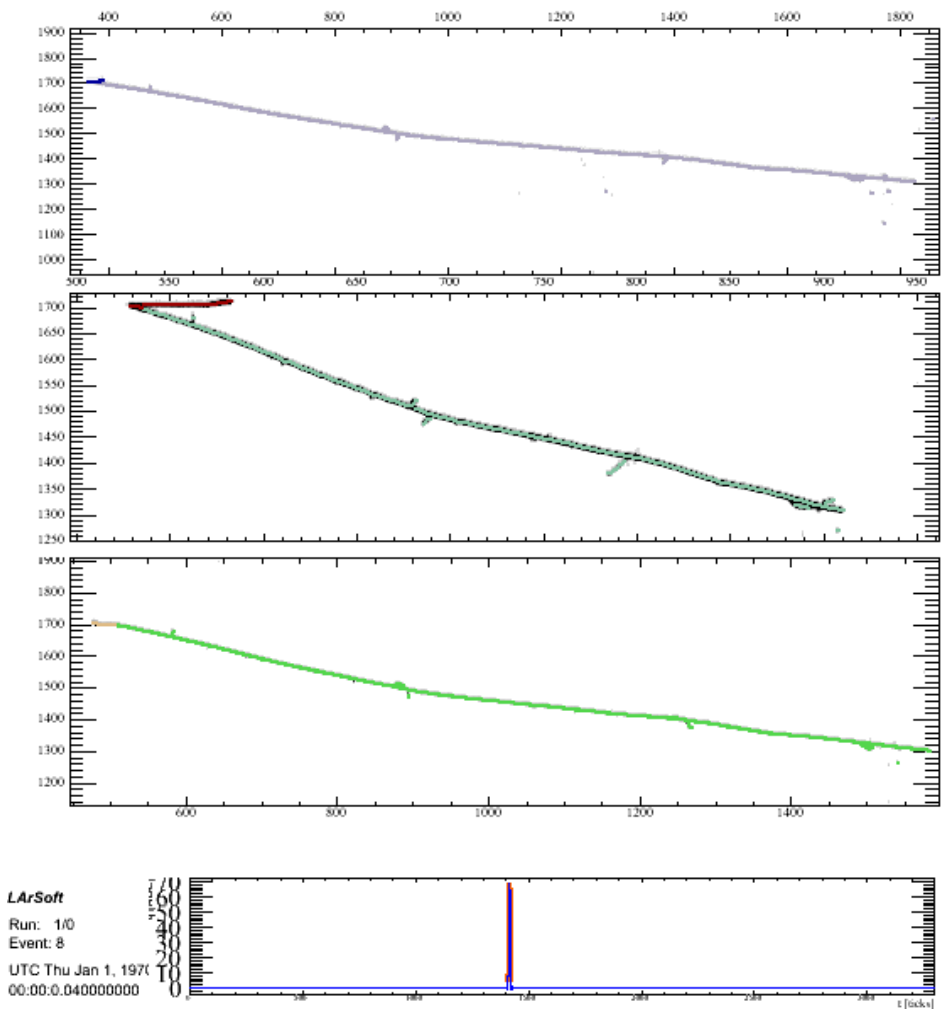
Fuzzy and Hough line clustering



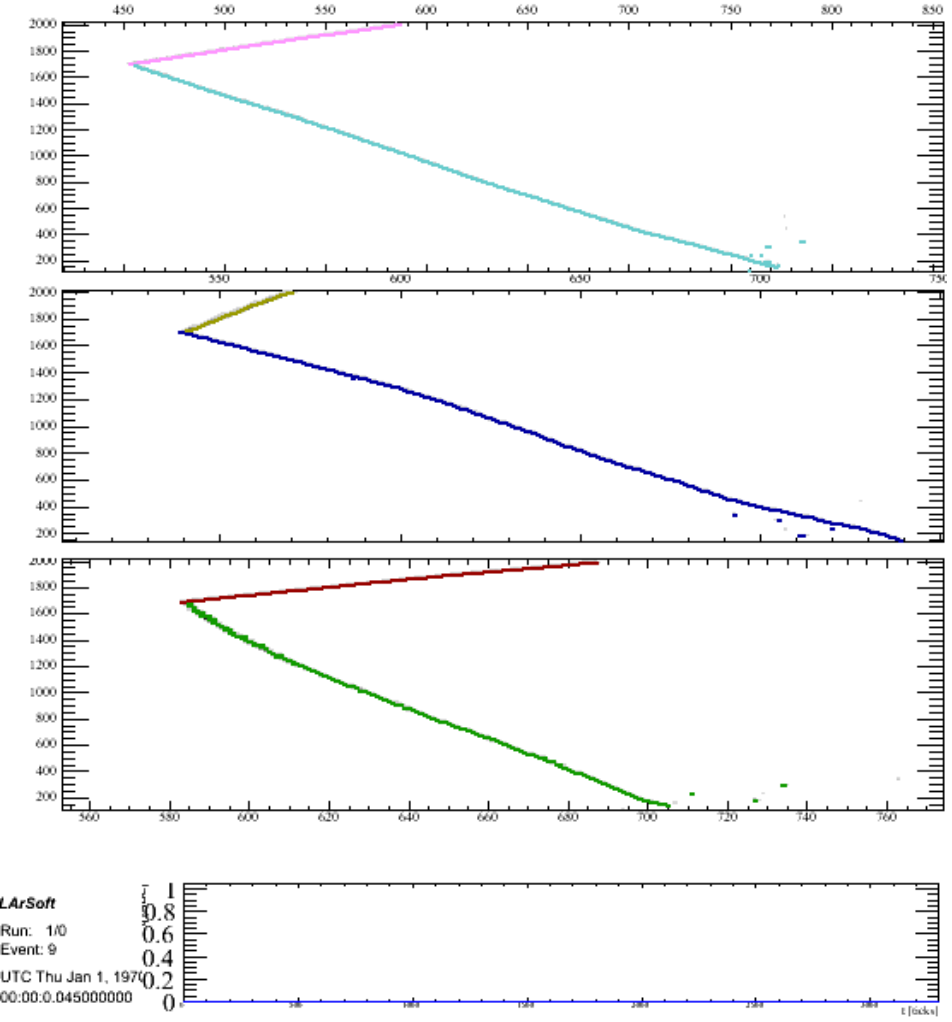
Cluster cheater



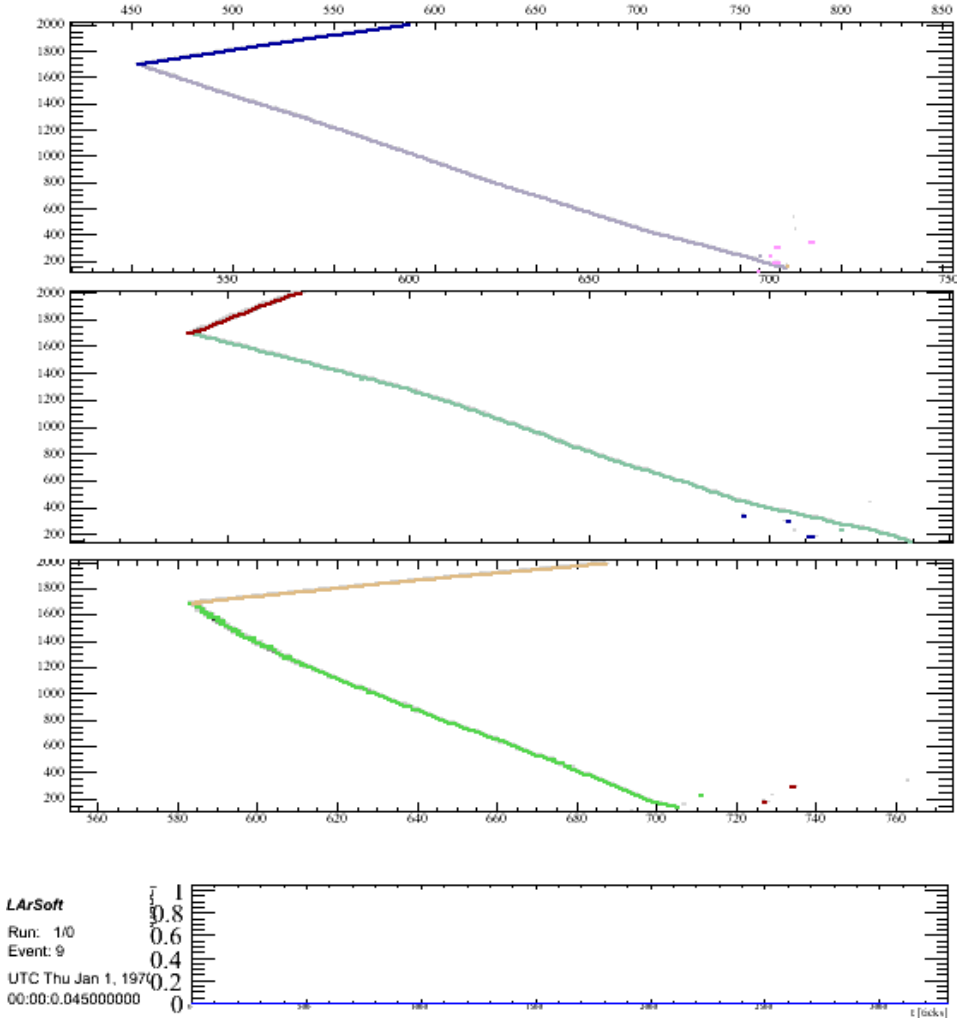
Fuzzy and Hough line clustering



Cluster cheater



Fuzzy and Hough line clustering

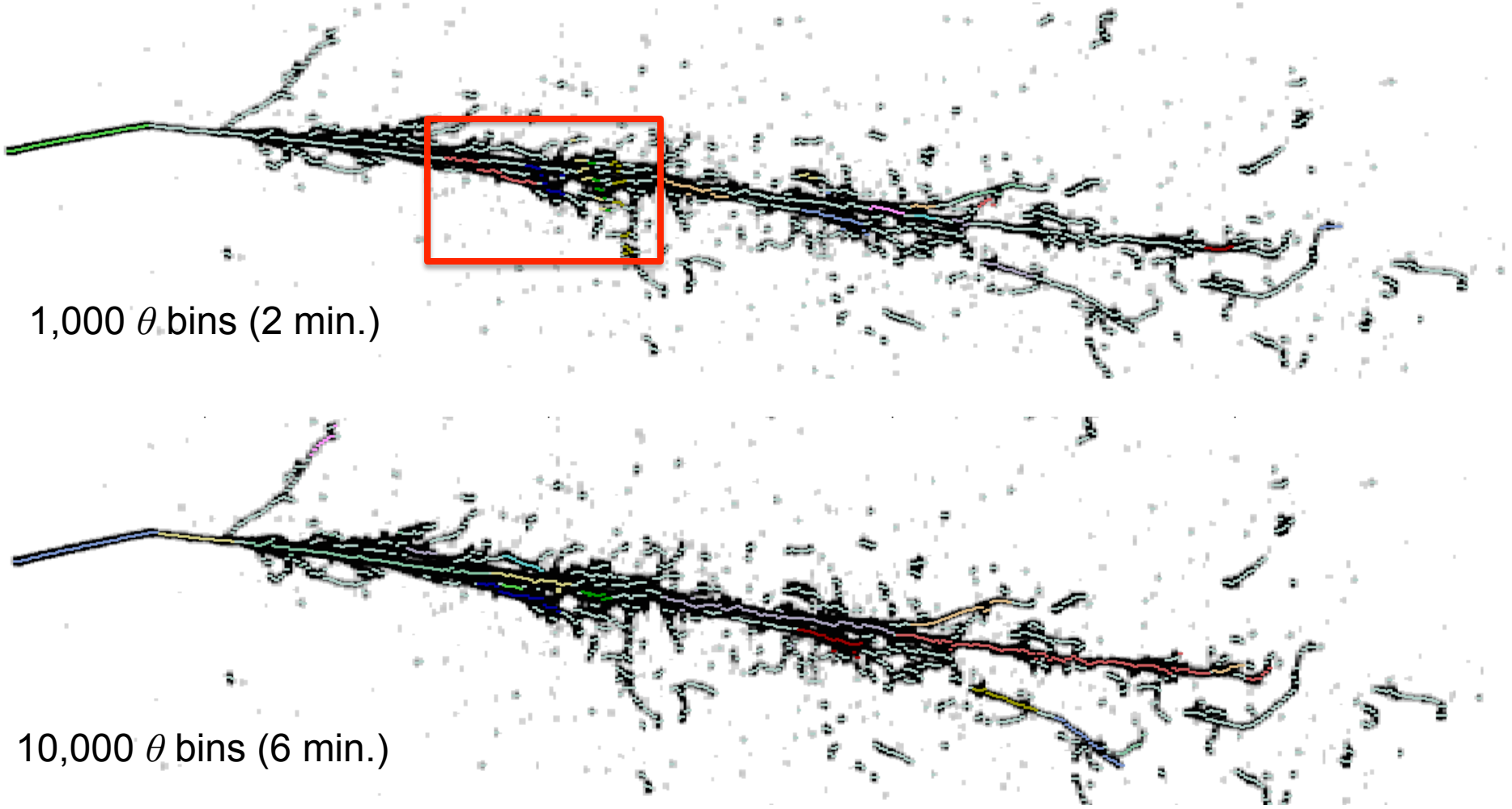


Cluster cheater

Hough transform performance

- Quality Hough lines are very dependent on resolution in the accumulator
- Low resolution (fast) has a higher fake rate than a higher resolution (slow)
- Ideally would run at low resolution, with fake veto

Hough transform performance



1,000 θ bins (2 min.)

10,000 θ bins (6 min.)

Compare to 37 minutes (5,000 θ bins) for original Hough transform code

Next steps

- Finish optimization
- Streamline code, might be too many unnecessary classes and functions
- Examine more events (only CCQE so far) to improve robustness
- Finalize the needed modules

Hough transform

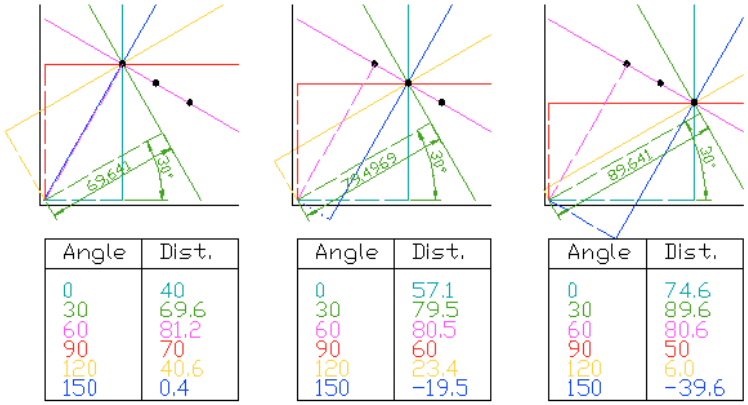
Use the parameterization:

$$y = \left(-\frac{\cos\theta}{\sin\theta}\right)x + \left(\frac{r}{\sin\theta}\right)$$

Fill a matrix (accumulator) in (r,θ) space with:

$$r(\theta) = x_0 \cdot \cos\theta + y_0 \cdot \sin\theta$$

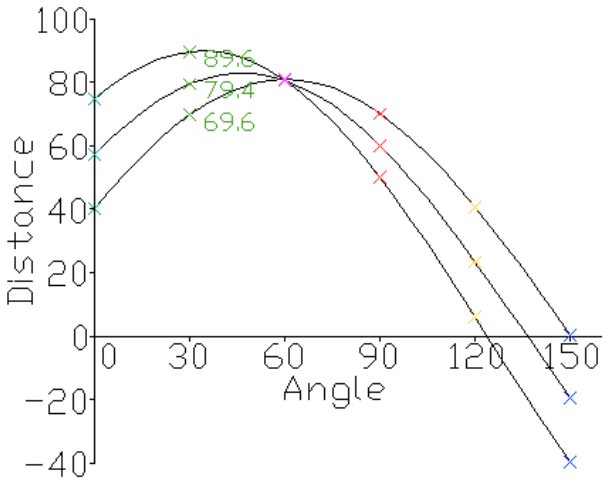
We then search the accumulator for the most populated bin, rather computationally demanding



Angle	Dist.
0	40
30	69.6
60	81.2
90	70
120	40.6
150	0.4

Angle	Dist.
0	57.1
30	79.5
60	80.5
90	60
120	23.4
150	-19.5

Angle	Dist.
0	74.6
30	89.6
60	80.6
90	50
120	6.0
150	-39.6



The accumulator, binned

Back up

Where clustering was

- Want to finalize clustering occurring right after hit reconstruction
- Lots of tools already available (e.g. DBSCAN, Hough line finder, and End Point Finder)
- Goal is to bring them all together

The basic algorithm

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ then STOP; otherwise return to step 2.*

The objective function we are trying to minimize:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2$$

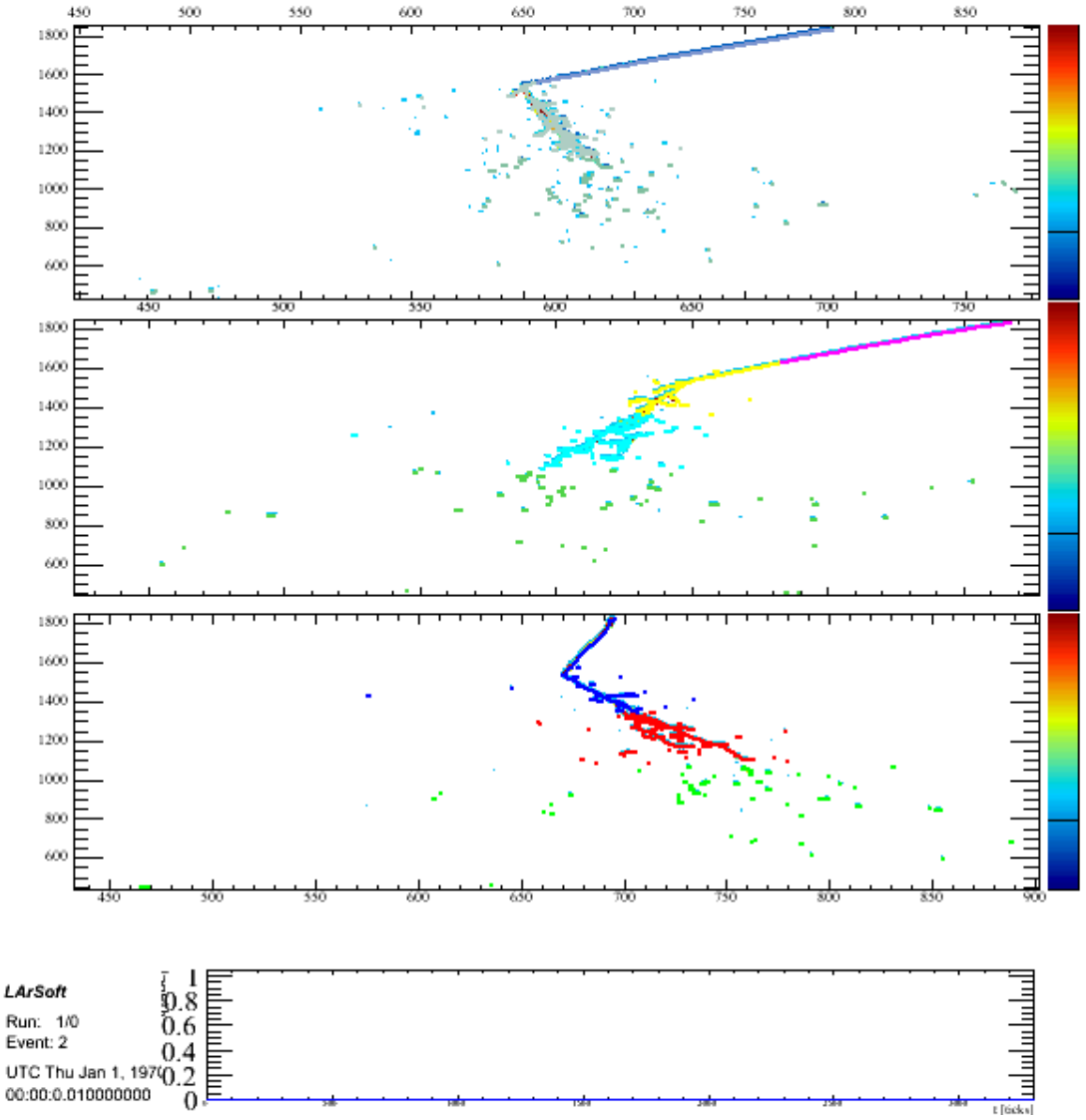
Number of clusters

- The number of clusters needs to be given as an input
- The Xie-Beni index gives us a way to evaluate how well a certain cluster number works

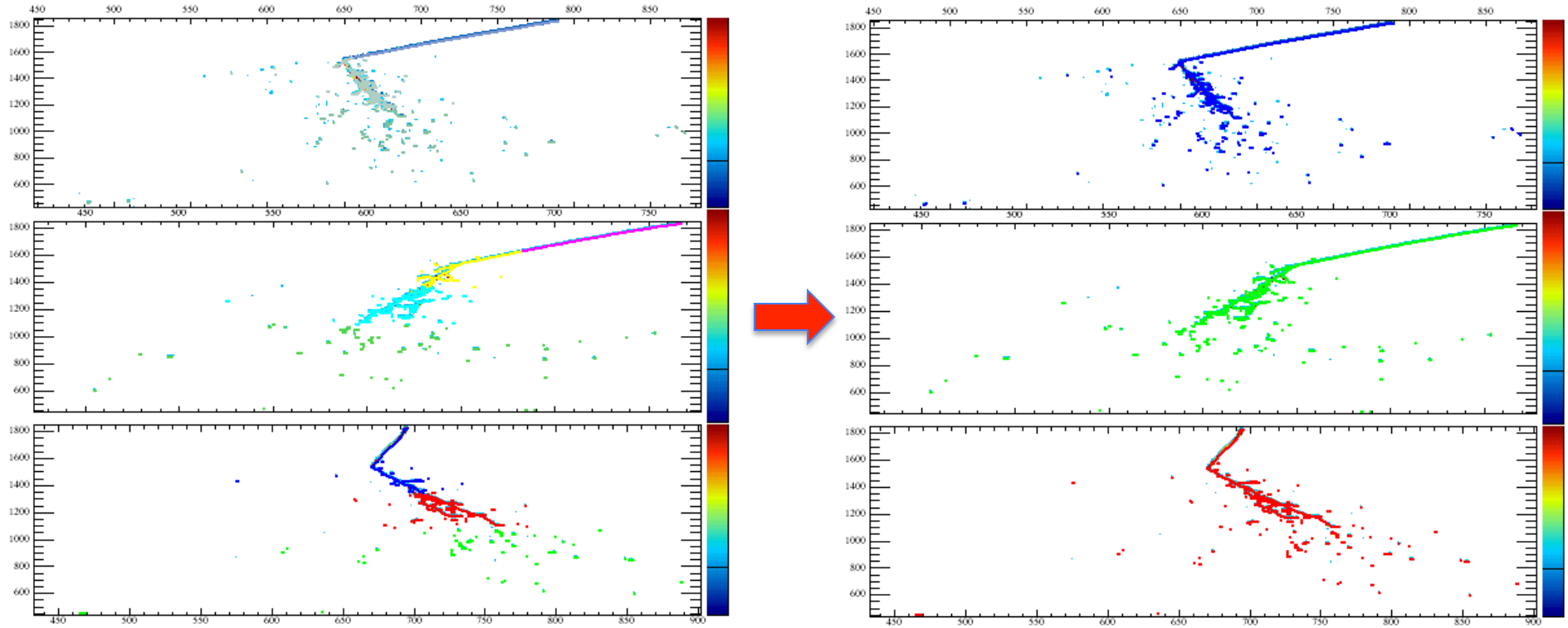
$$\frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{m} \left\| x_i - c_k \right\|^2}{\min_{k,\ell} \left\| c_k - c_\ell \right\|^2}$$

Does not work with single cluster events by definition, we'll get to that later

There is a problem with the Xie-Beni Index though

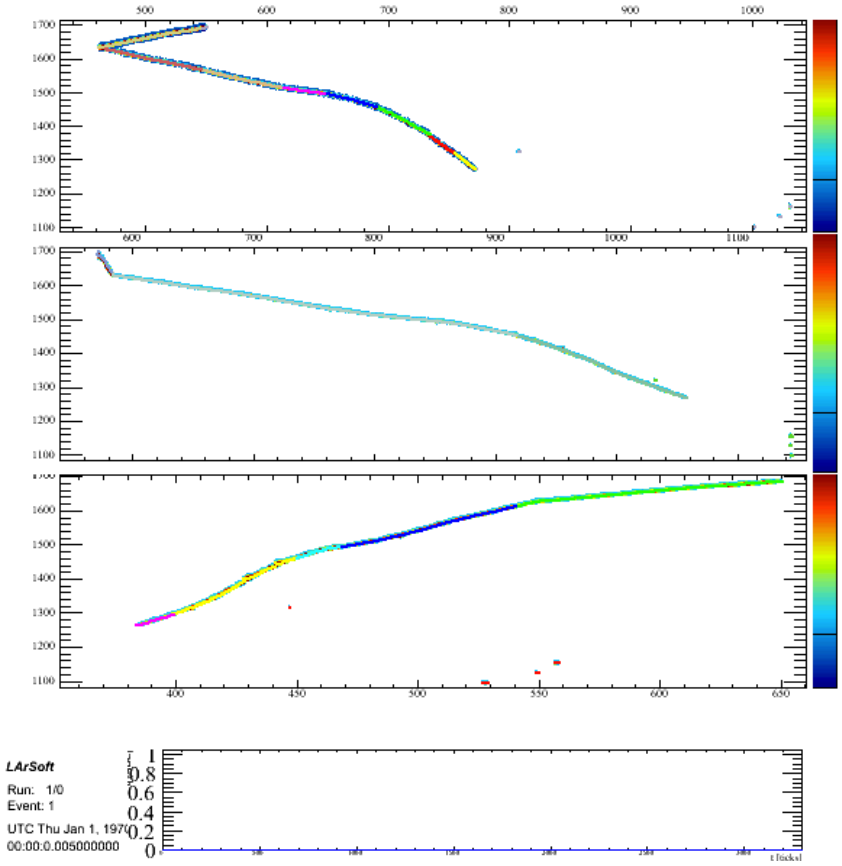


Look to merge clusters sing Euclidean distance



1. Look for point in cluster i closest to centroid in cluster j
2. Compare that point to the point in the cluster j closest to centroid of cluster i
3. Merge the clusters

Now look for Hough lines and expand them into clusters

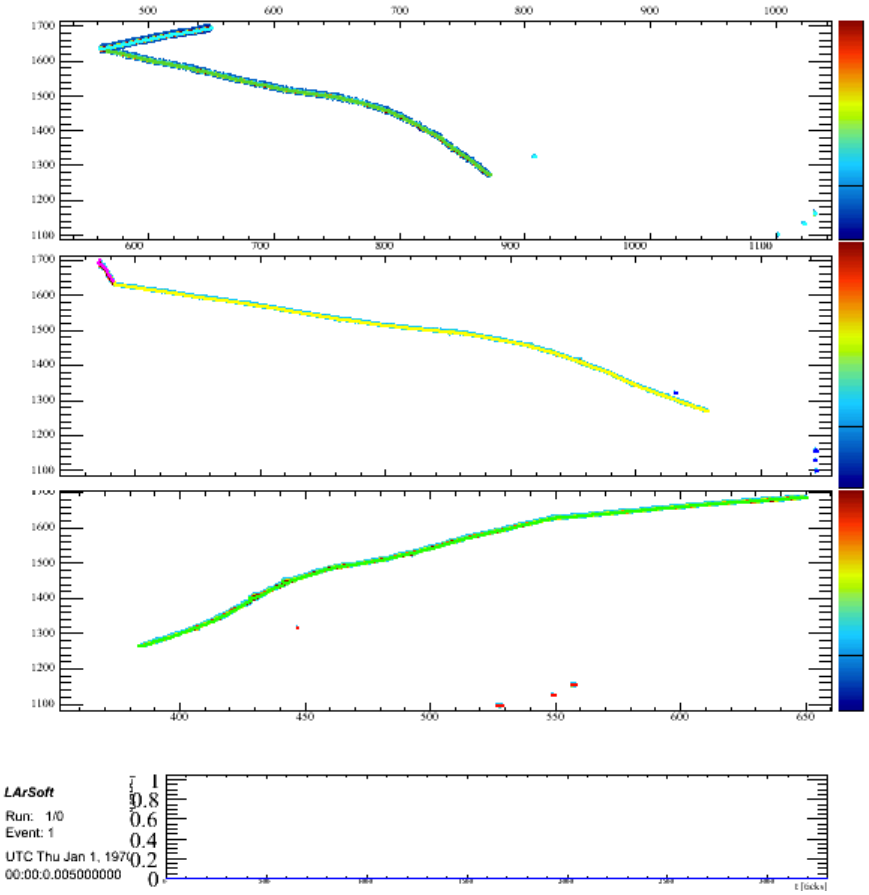


Search for Hough lines in the already identified fuzzy clusters

Hough line finder was sped up to make this practical

An issue remains with needing to merge the Hough lines though

Merging the Hough lines



Pick a line and check at its end points

If a nearby line is found, check the angle between the slopes

If the angle between the slopes is < 30°, merge the lines