A Novel Two Stage Approach for Single Pulse Post-processing

-Shubham Singh and the SKA-PSS team:

(A. Karastergiou, A. Naidu, B. Stappers, B. Shaw, B. Posselt, D. Lumba, G. Berriman, J. Taylor, K. Rajwade, L. Levin, M. Droog, M. Mickaliger, P. Thiagaraja, R. Hombal)

13 Oct 2025: FTSky, Bengaluru

SKAO

PSS

MANCHESTER 1824

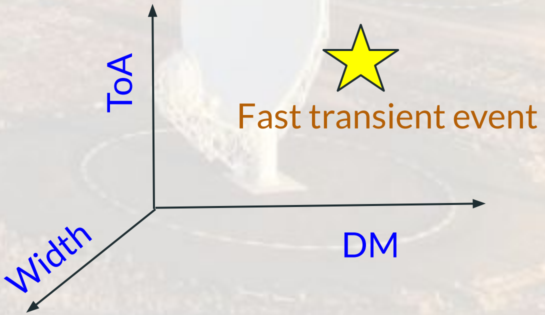
The University of Manchester

The Post-processing Parameter Space

Fast transient search

Fast transients can be described by three parameters in addition to their sky coordinates,

1. Distance from observer (Indicated by electron column density, DM)
2. When did the event happen (Time of occurrence/arrival, ToA)
3. For how long the transient event last (Duration of the event, Width)

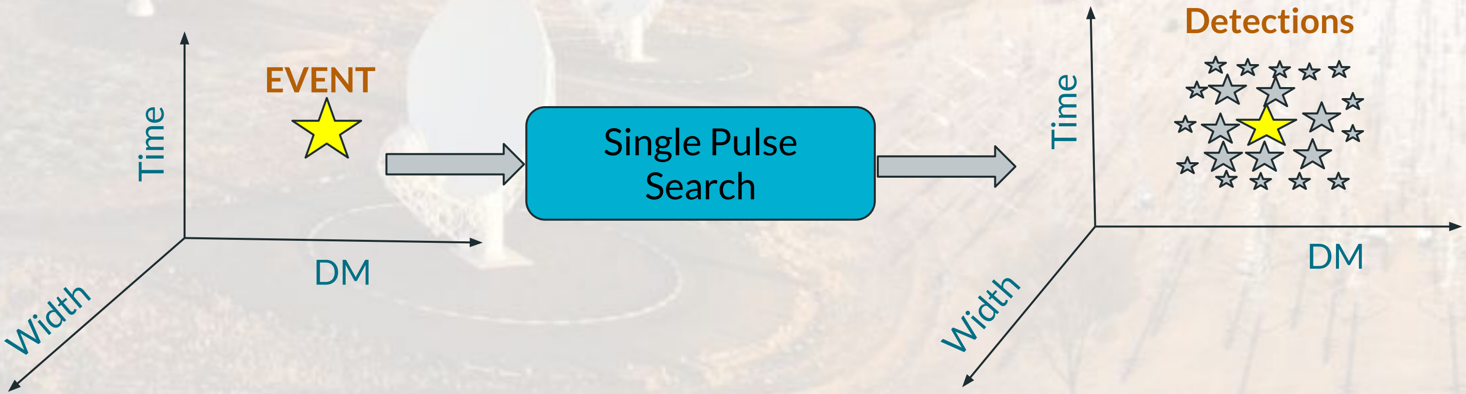


Final result of single pulse search

- The telescope search mode data is searched over,

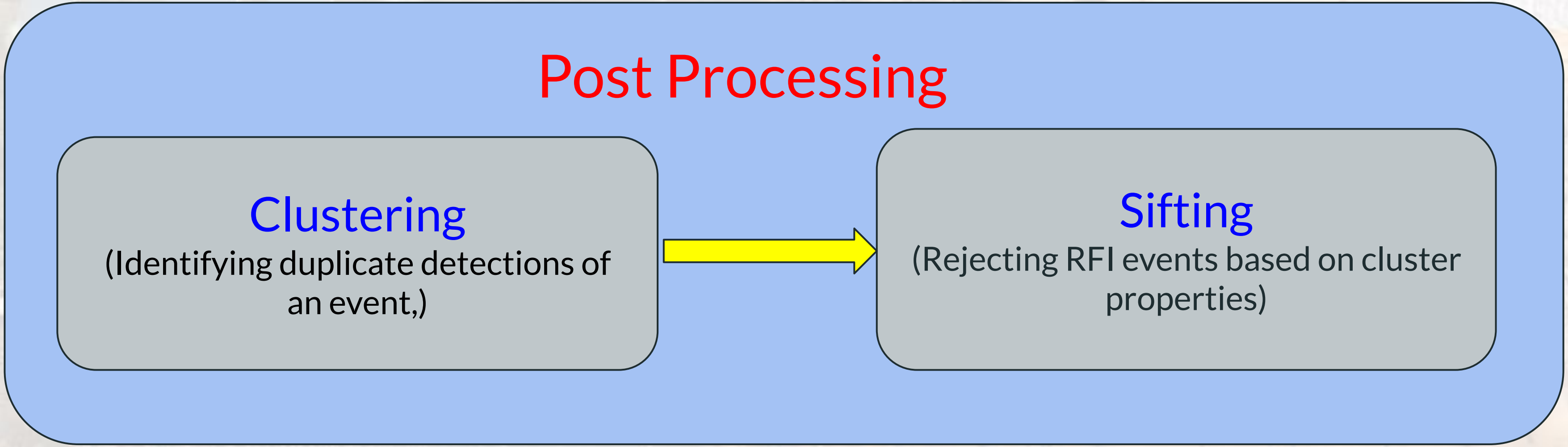
1. All time samples 2. Set of Trial DMs 3. Set of trial Width

- The result of the search is a set of detections in the Time-DM-Width space with associated S/N.



Need for post processing

- A strong and wide event can sometimes produce millions of duplicate detections.
- A big fraction of candidates are RFI generated.



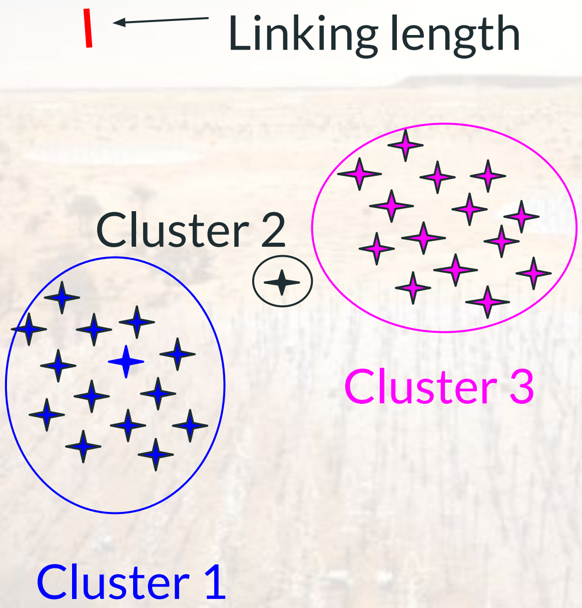
Clustering methods

Friends of friends clustering

- Fast
- Can handle large number of detections
- User defined linking lengths
- The linking lengths need to be tuned according to the target pulse
- Can not handle variable density (caused by the DM plans)

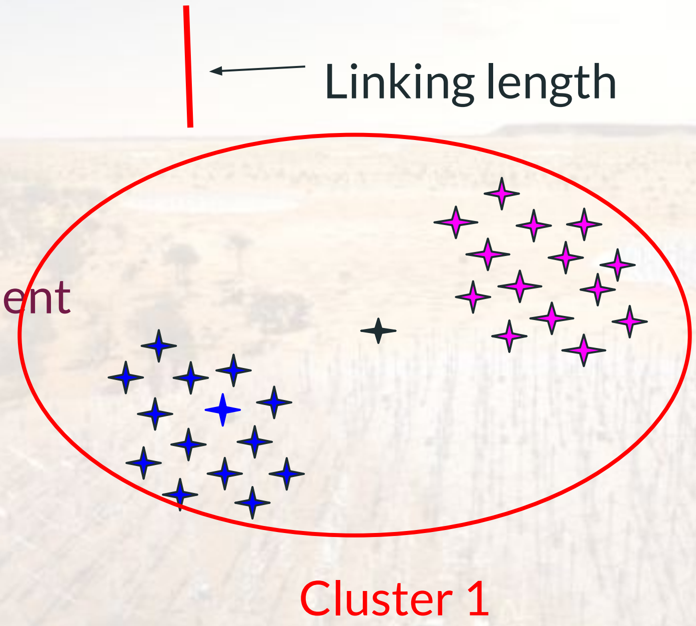
Friends of friends clustering

Smaller Linking length can make too many cluster. It can also split a single cluster into multiple clusters.



Friends of friends clustering

Large linking length might merge multiple independent clusters along with noise.



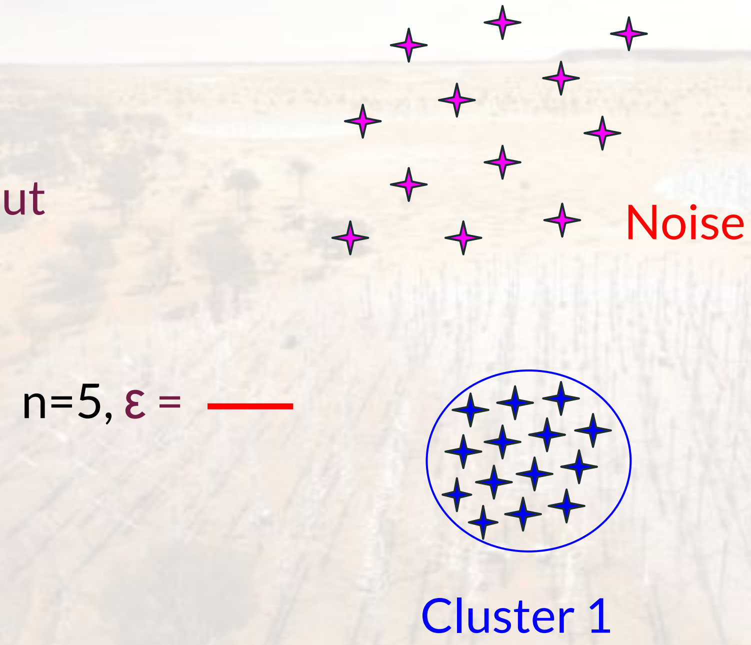
DBSCAN Clustering

- Fast
- Can handle large number of detections
- Uses density instead of linking lengths
- Can handle noise
- Still needs a length scale (ε) and a number (n) to decide density cut-off.
- Can not handle variable density

DBSCAN Clustering

A predefined n and ε may work for one density but not for other.

The DM plan used in search creates different detection density for different DM range.

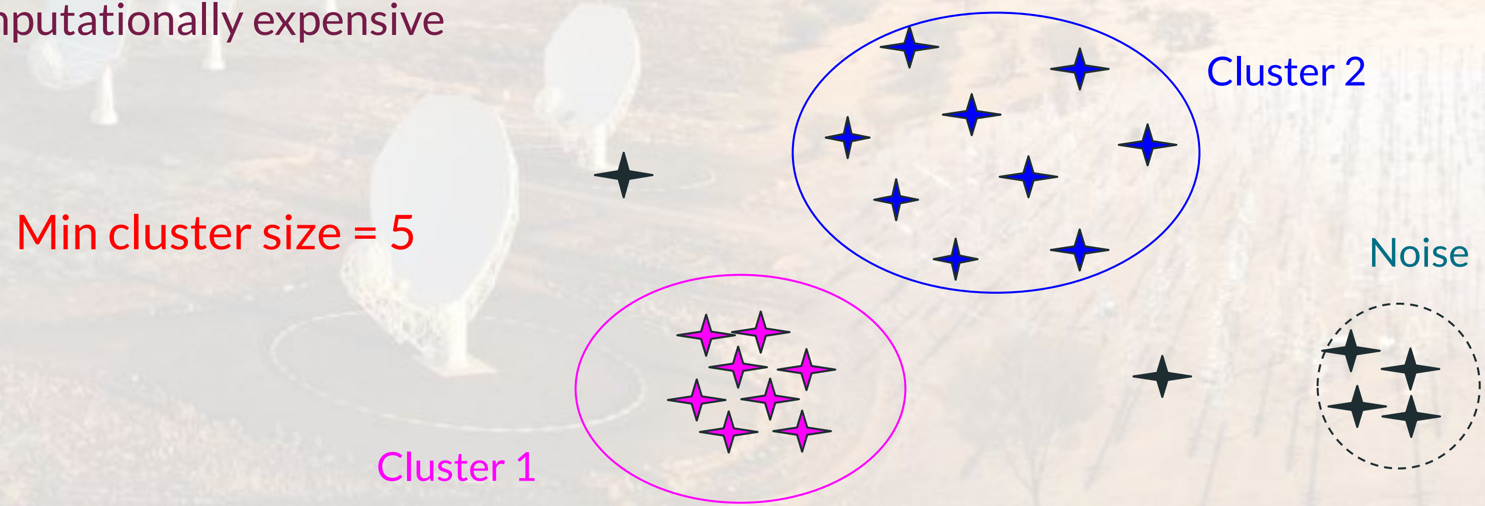


HDBSCAN Clustering

- Needs only one user defined parameter (minimum size of the cluster), can create bias against faint pulsars
- Tolerant to density variations, hence can handle density variations due to DM-plan
- Robust against noise, can identifies noise points
- Computationally expensive

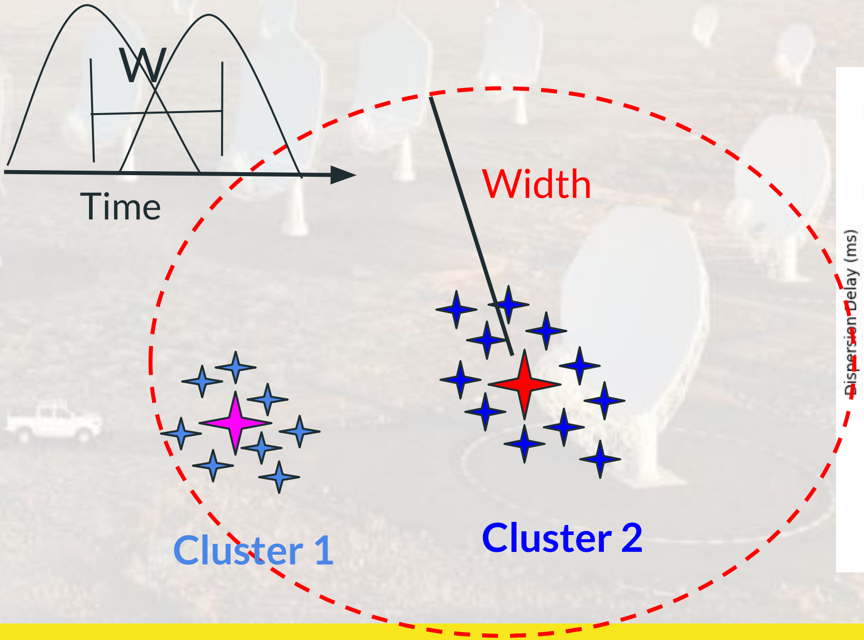
HDBSCAN Clustering

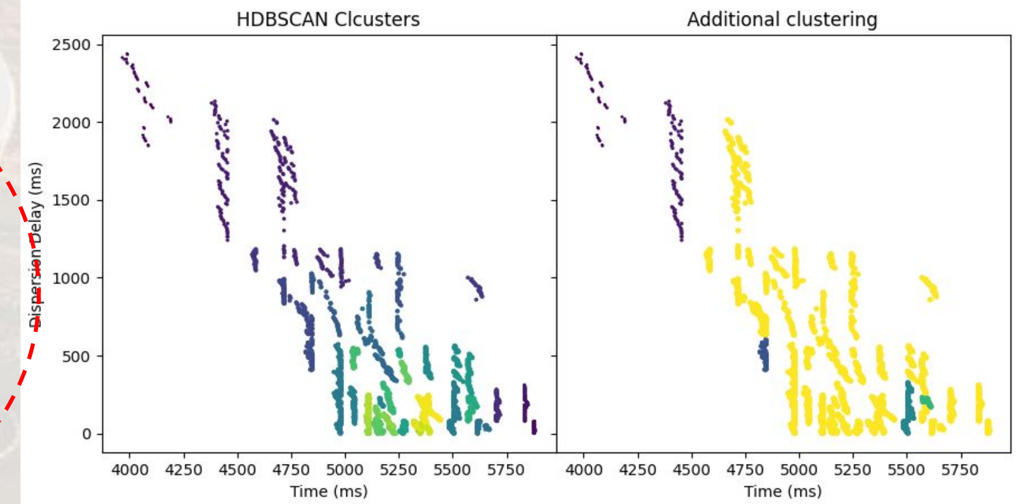
- Uses only one user defined parameter (minimum size of the cluster), can create bias against faint pulsars
- Tolerant to density variations, hence can handle density variations due to DM-plan
- Robust against noise, can identifies noise points
- Computationally expensive



Need for additional clustering step

- In case of wide pulses, multiple clusters are formed for a single event.
- We propose and additional step of clustering that unifies the clusters within the detection width.





Optimal approach for clustering

HDBSCAN Clustering
(With one clustering parameter: Minimum cluster size)

\+

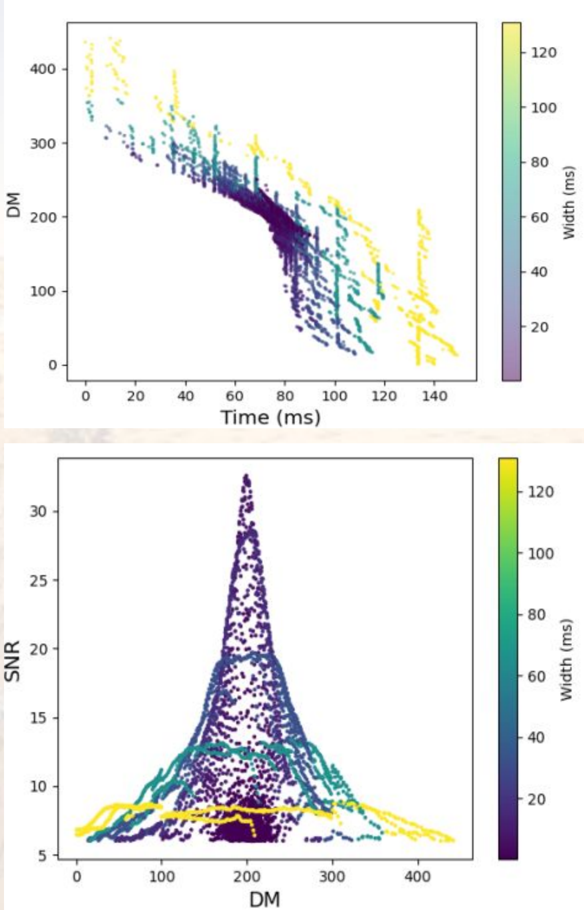
Additional Width Based Clustering
(Uses detection width of brightest detection as clustering radius)

Sifting methods

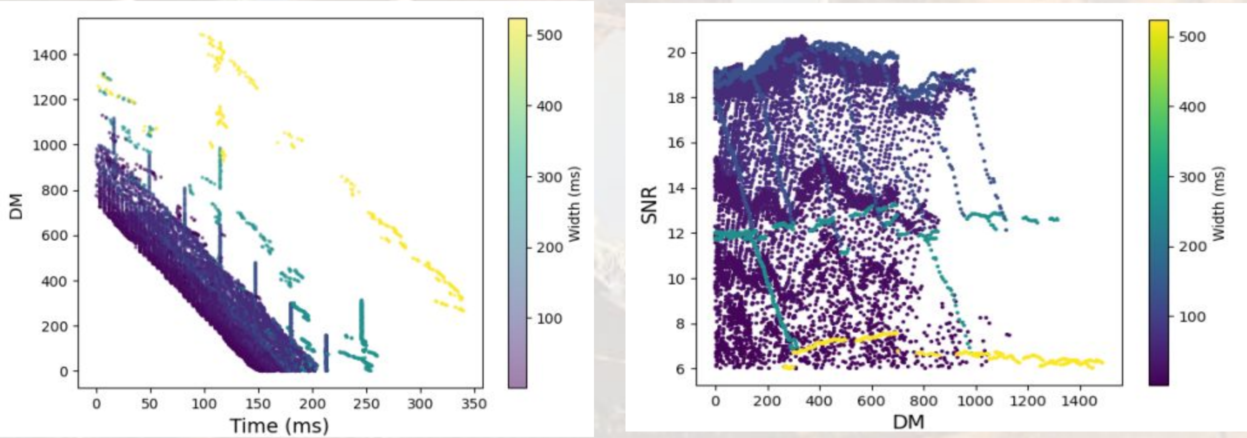
Selecting cluster features for sifting

- Limited number of intuitive features
- No complex fitting on the data: expensive for large clusters and noisy for small clusters
- Select carefully: feature can introduce bias

Astrophysical Event



RFI Event



Manual thresholding

- A lower cut-off on the size of cluster can get rid of noise detections.
- A simple lower cut-off on the DM of best detection can remove most of the wideband RFI.
- But excluding narrowband RFI clusters is non-trivial.

Machine learning sifting methods

We tried three supervised machine learning methods:

1. **Decision tree:** Simple and interpretable, but prone to overfitting and imprecise (accuracy: 70-80%)
2. **Random Forest:** Collection of decision trees, interpretable and accurate (accuracy: 90-100%)
3. **Neural Network:** Performance similar to Random Forest, but hard to interpret results

We decided to use Random Forest for its accuracy, flexibility, and interpretability.
We easily achieve **96%** accuracy with our selected features and Random Forest sifting.

Reducing false negative rate in random Forest

We can use the flexibility of the Random forest classifier to reduce the loss of astrophysical candidates.

Regular Majority Voting:
An object belongs to class A, if more than 50% trees say so.

Majority Voting to reduce the loss of astrophysical class:
A cluster is astrophysical if 40% or more trees say so.

Optimal approach for sifting

Simple set of cluster features

(Small number of intuitive, easy to compute features)

+

Random Forest Classifier

(Classifier with higher weightage to astrophysical class to reduce the loss of astrophysical signal)

No Optimal Approach, but this does the job !!

Single pulse search post-processing in PSS-Cheetah

Available methods for Clustering

1. FoF (fast approach)
2. HDBSCAN + Width based clustering (robust approach)

Available methods for Sifting

1. Simple thresholding on features
2. Random Forest classification (RfSift)

Recommended Combination: HDBSCAN clustering followed by Random Forest Sifting

Summary

- The single pulse post-processing has two parts: clustering and sifting
- The clustering step aims to identify clusters belonging to a specific event.
- HDBSCAN followed by a width based clustering is the optimal way to cluster single pulse detection.
- The sifting step aims to identify RFI clusters and remove them.
- The Random Forest classification of clusters with simple set of features is an effective and flexible method for sifting.
- The Random Forest Classifier can be tuned to reduce false negative rates, reducing the loss of astrophysical signal.

Thanks !

Difference between RFI and Astrophysical clusters

Astrophysical Signal vs Narrowband RFI

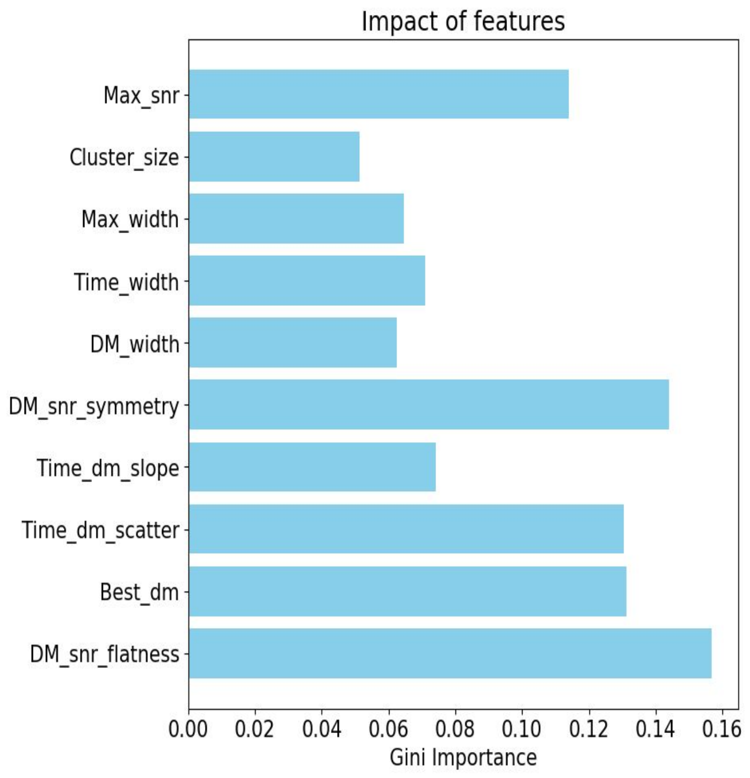
A few fundamental differences how narrowband and wideband signals appear in DM-Time plane and DM-S/N plot

Astrophysical Signal vs wideband RFI

Very similar, only difference is the DM corresponding to best S/N detection, which is very small in case of wideband RFI

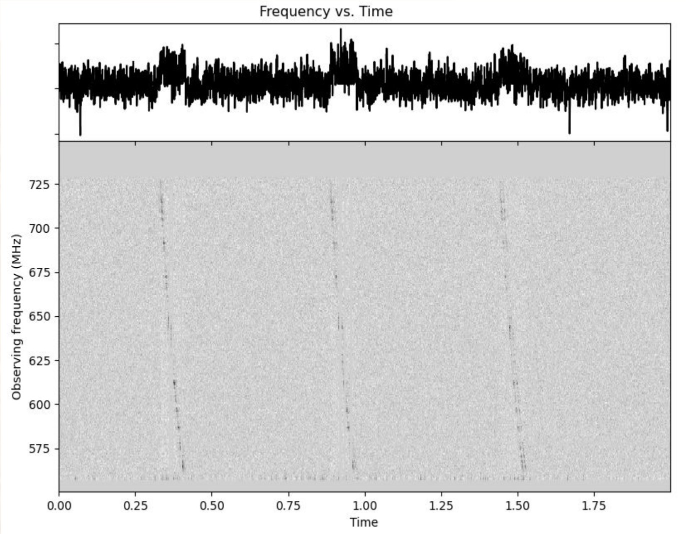
Performance of cluster features with random forest

1. **Max_snr:** Maximum detection SNR in the cluster
2. **Cluster_size:** Number of detections in cluster
3. **Max_width:** Detection width of brightest member
4. **Time_width:** Extent in time axis
5. **DM_width:** extent in DM axis
6. **DM_snr_symmetry:** measure of how symmetric DM-SNR plot is
7. **Time_dm_slope:** Slope of the Time DM plot
8. **Time_dm_scatter:** Scatter in Time-DM plane
9. **Best_dm:** DM of best detection
10. **DM_snr_flatness:** The flatness measure of DM-SNR plot



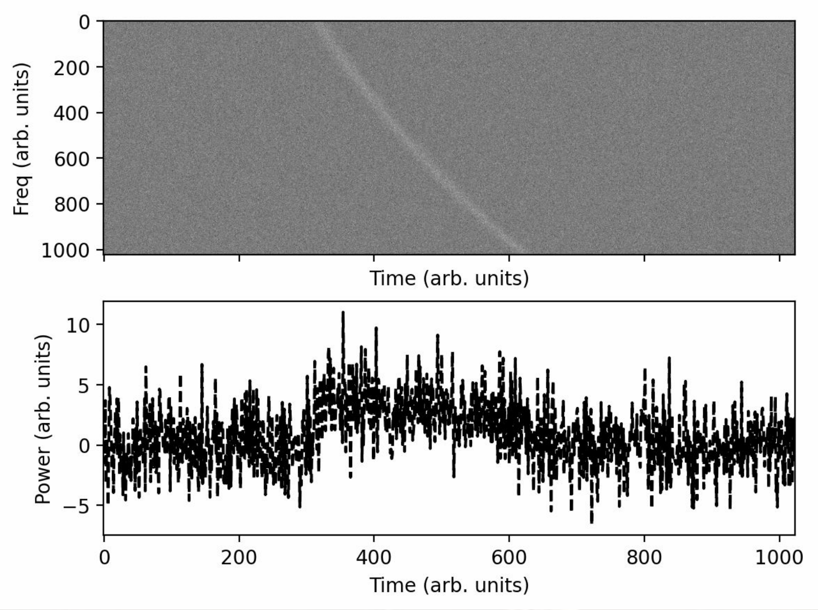
Radio data from the telescope

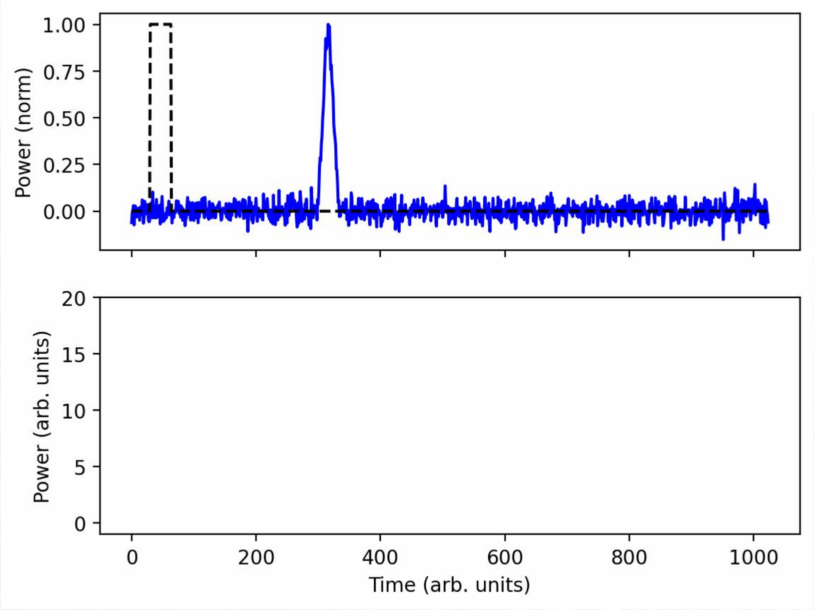
- The radio telescope is pointed towards the interesting sky coordinates (RA, Dec)
- High time resolution data with enough frequency channels is recorded
- The final product is Time-frequency data corresponding to a sky coordinates.



What is PSS (Pulsar Search Subsystem)

The subsystem aims to find pulsars and fast transients in real time

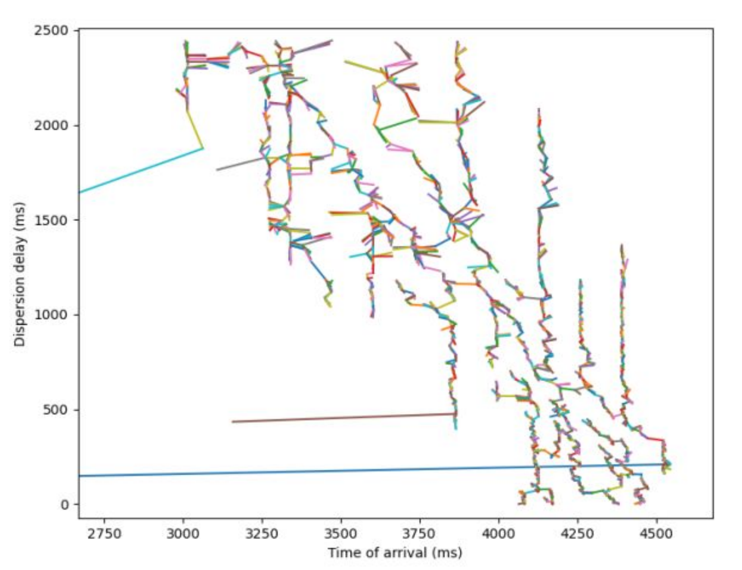




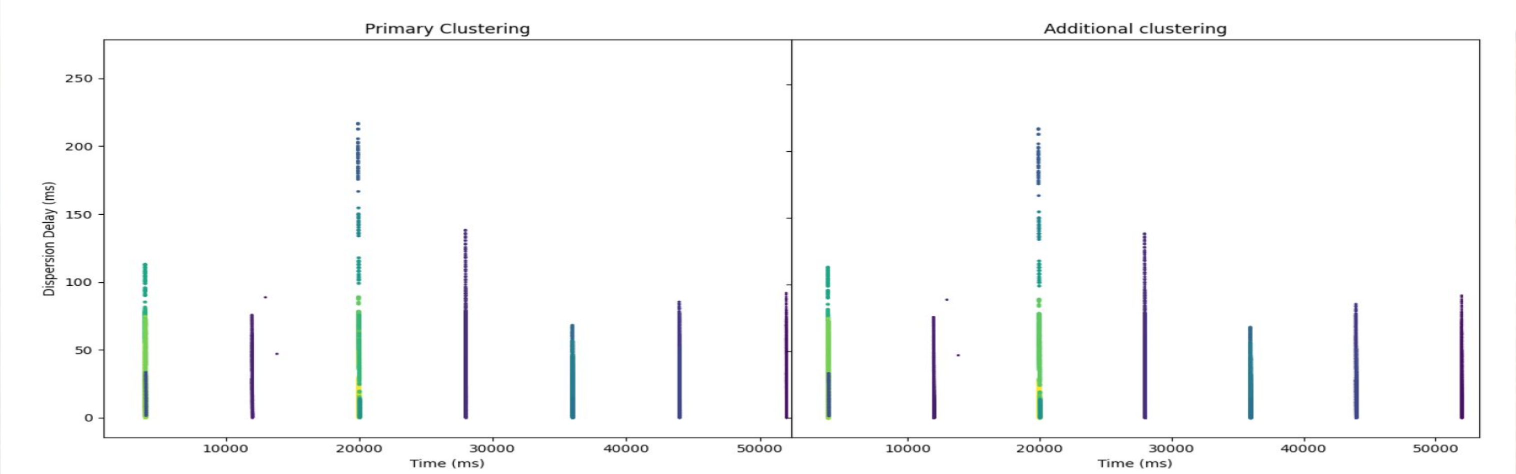
Credits: Kaustubh Rajwade

Constructing MST based on MRD

- ‘Spanning_tree_MRD’ class provides tools to compute MRD and construct MST.
- Method ‘Compute_coredist()’ is first used to compute the core distance for all data points.
- Then method ‘construct_tree()’ is used to construct the MST based on MRD.



Examples of additional clustering step based on width



Case of narrow pulses (10 ms wide, 24 clusters are reduced to 20)