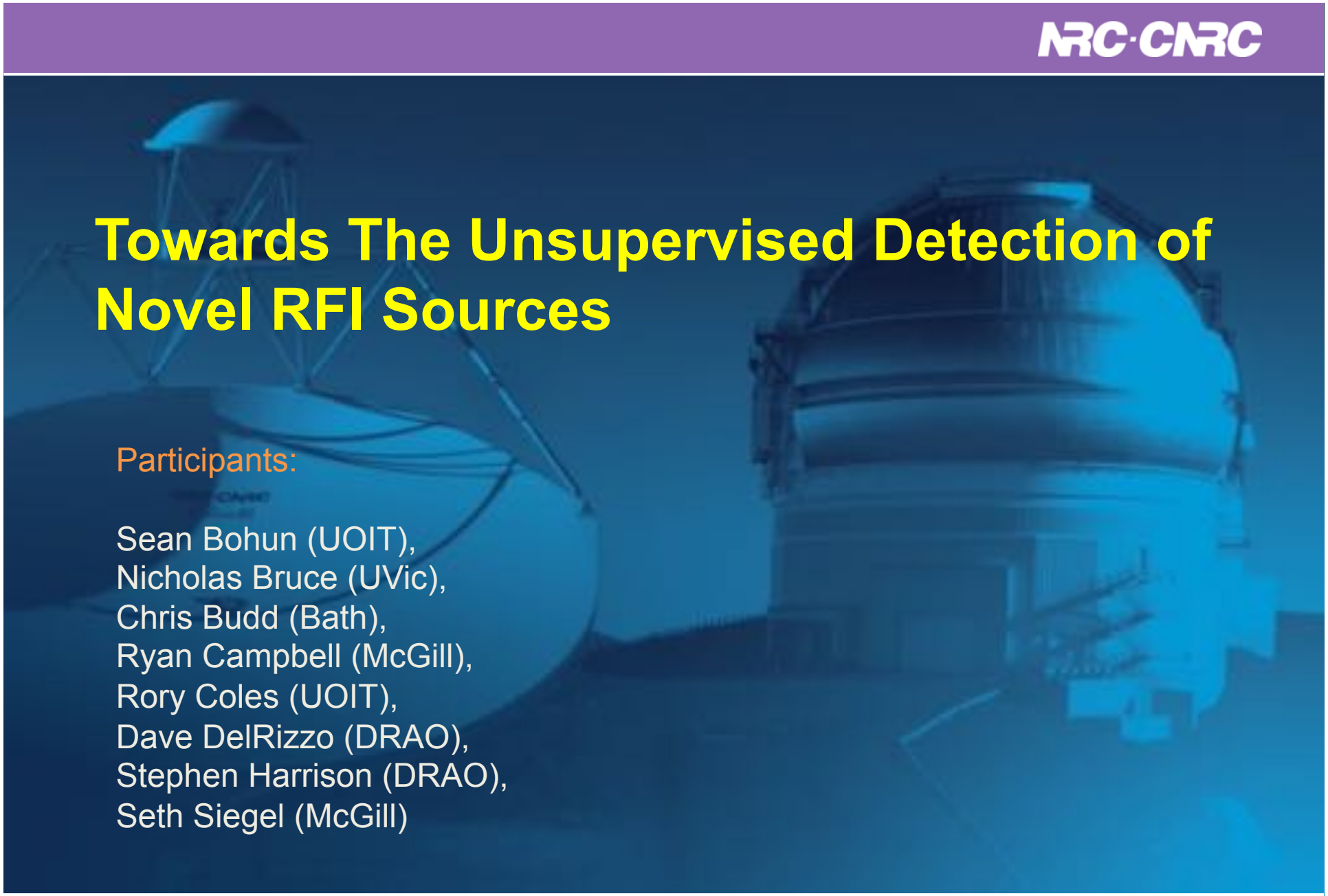


# Towards The Unsupervised Detection of Novel RFI Sources


Participants:

Sean Bohun (UOIT),
Nicholas Bruce (UVic),
Chris Budd (Bath),
Ryan Campbell (McGill),
Rory Coles (UOIT),
Dave DelRizzo (DRAO),
Stephen Harrison (DRAO),
Seth Siegel (McGill)

# Introduction

The Dominion Radio Astrophysical Observatory (DRAO) is located in a geographically isolated region near Penticton B.C.

This is the site of a multitude of sensitive RF detectors that require a quiet RF spectrum to effectively operate.

For any given day, the site can experience any number of transient RF signals due to a variety of sources.

# IPSW Questions

DRAO would like to develop the capability to:

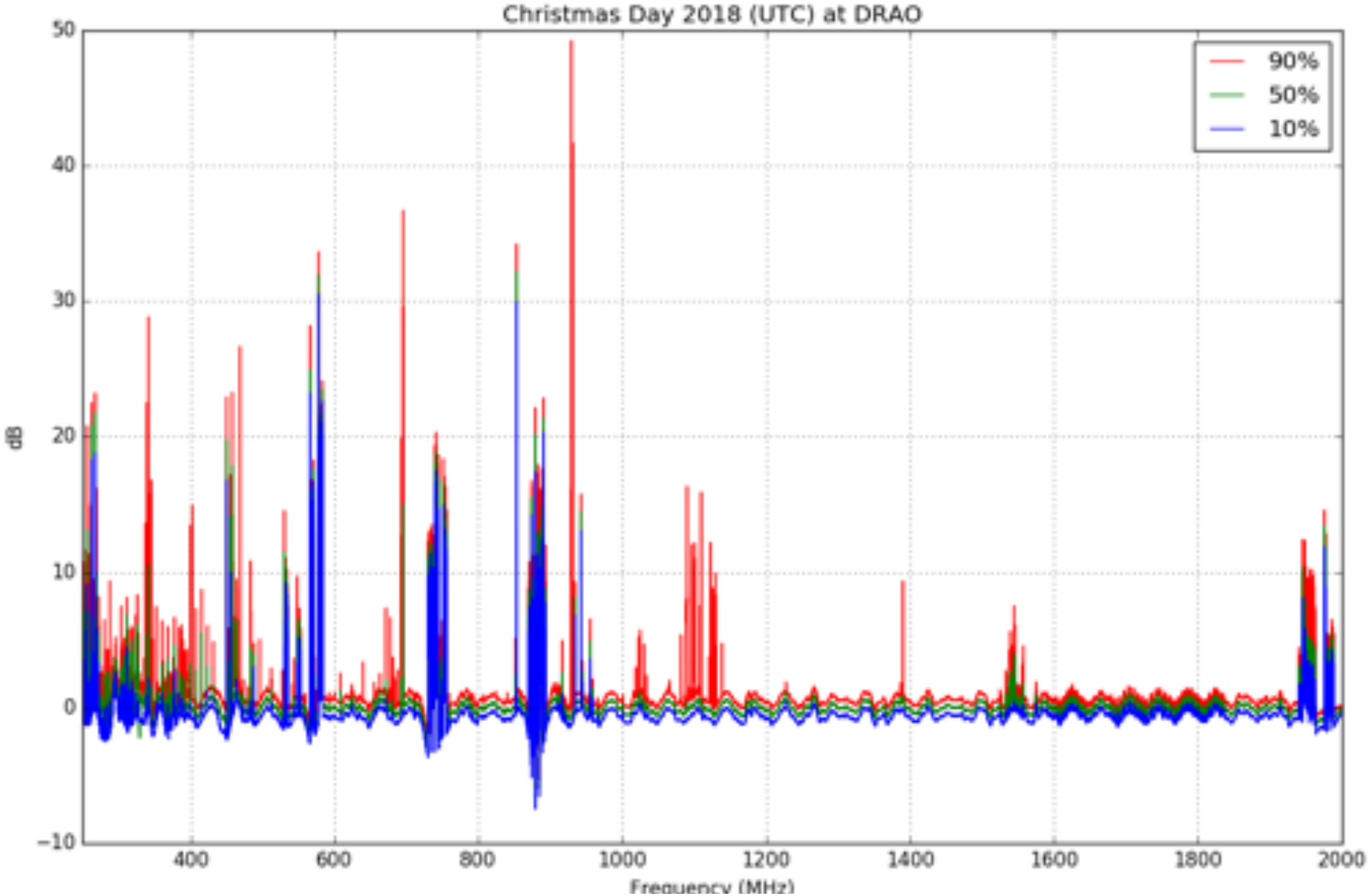
- Classify and Cluster the set of known RF sources as they are determined
- Identify any novel RF sources that have not been previously classified
- Provide a set of descriptors for each novel source
- Update the clusters dynamically as novel sources are identified

Once identified, the hope is that any novel sources can be eliminated with this technique.

# Site map

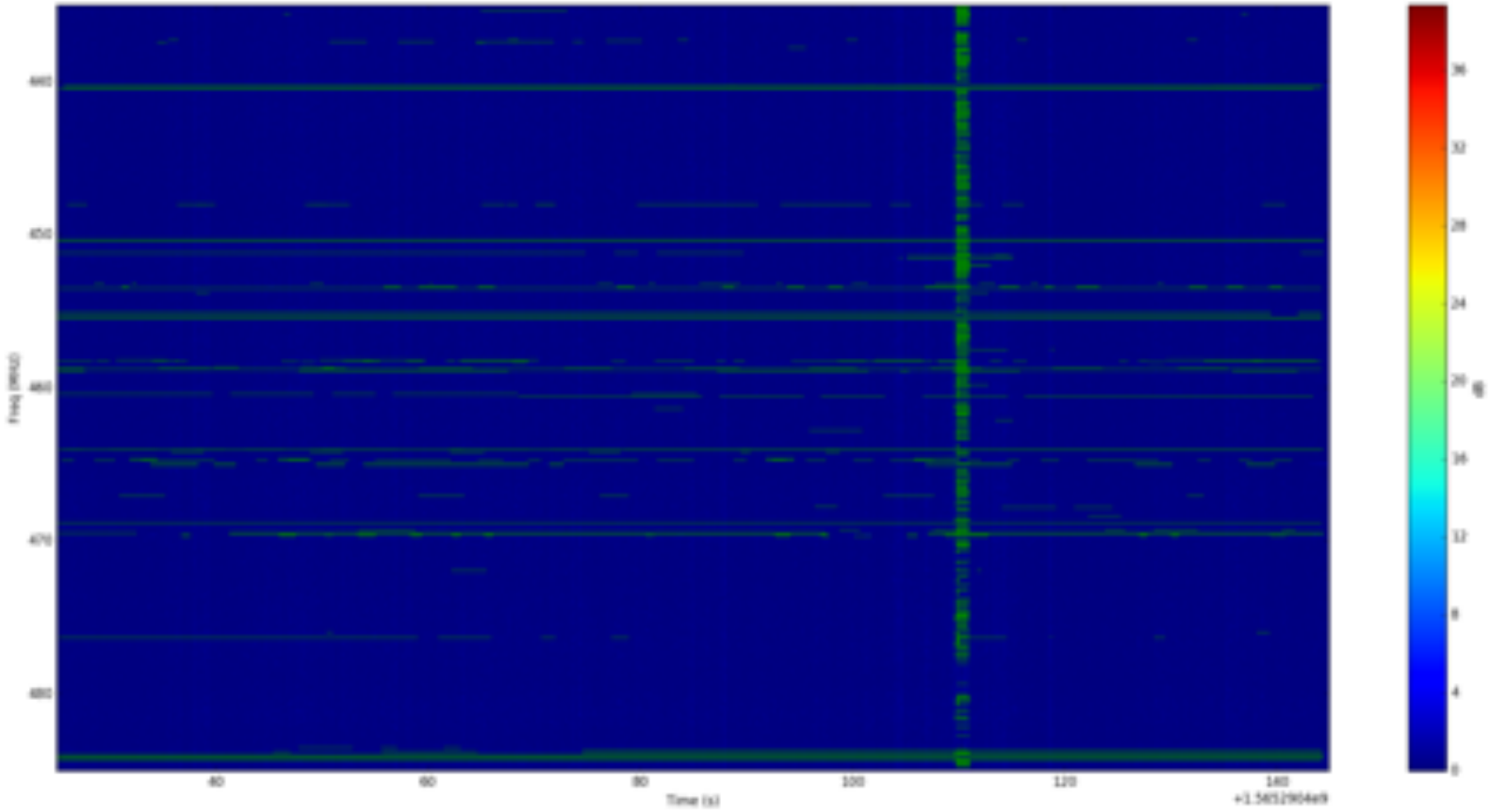

St Andrews Golf Course (2km)
Penticton (20km)
Us!
Twin Lakes Golf Course (8km)
OK Falls, Cell Towers (5km)
Oliver (16km)

# RFI: Static

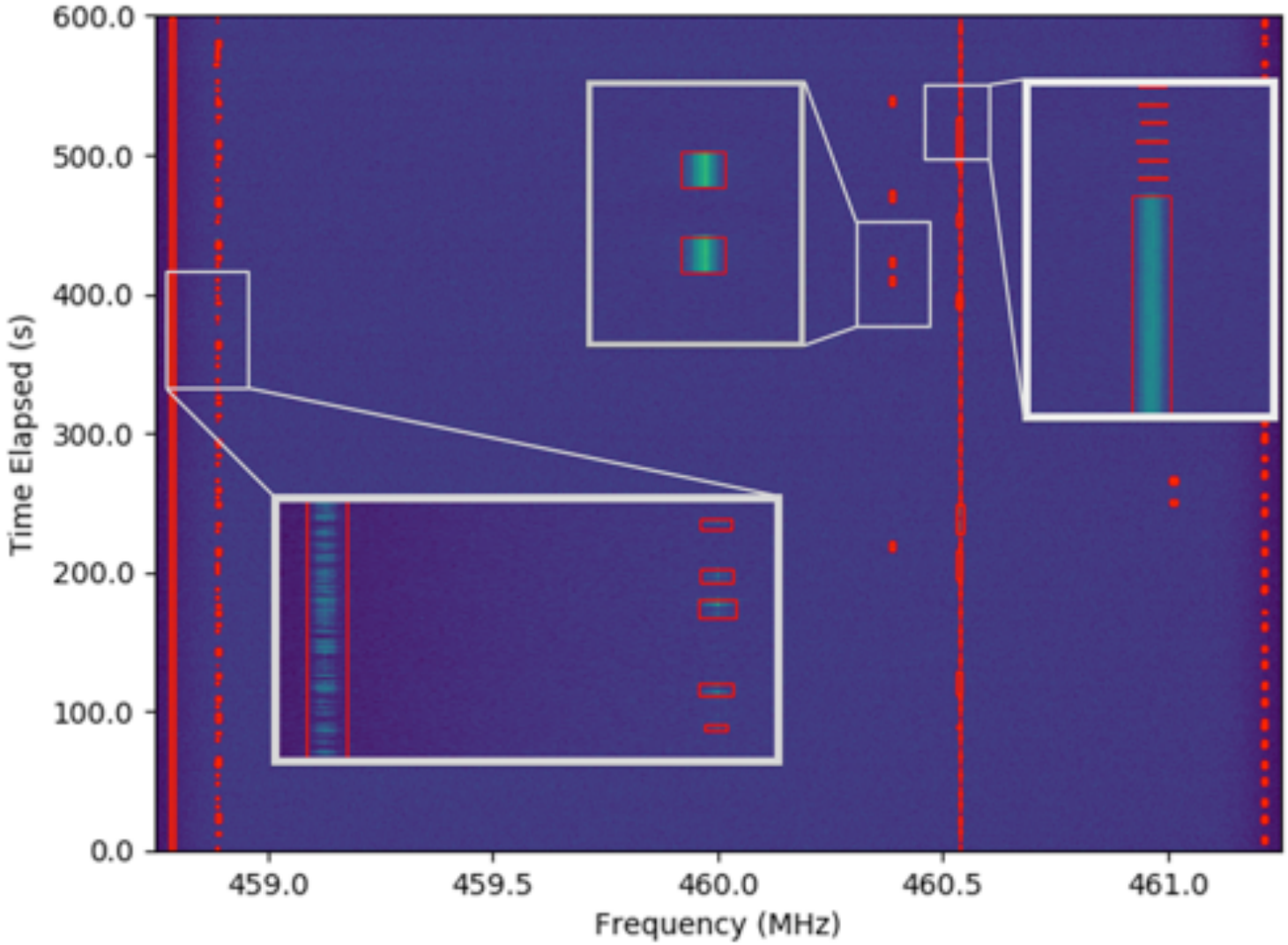

Christmas Day 2018 (UTC) at DRAO
90%
50%
10%
dB
50
40
30
20
10
0
−10
400
600
800
1000
1200
1400
1600
1800
2000
Frequency (MHz)

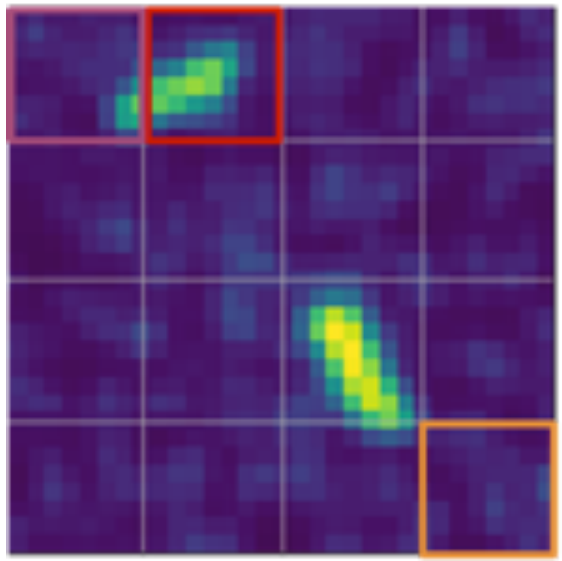
# The Dynamic RF Scene

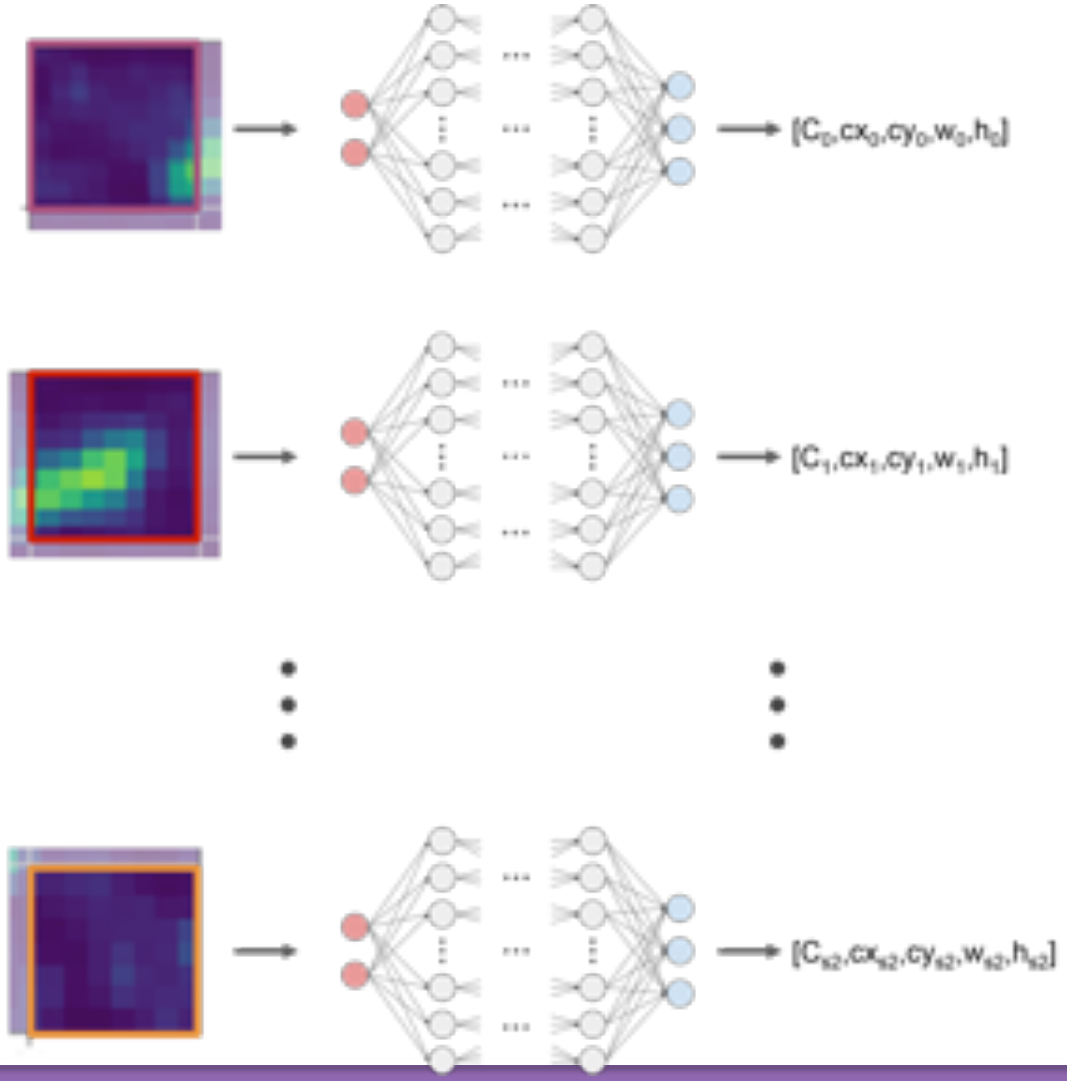
- Detection of RFI in time-frequency space.
  ML bounding box approach.



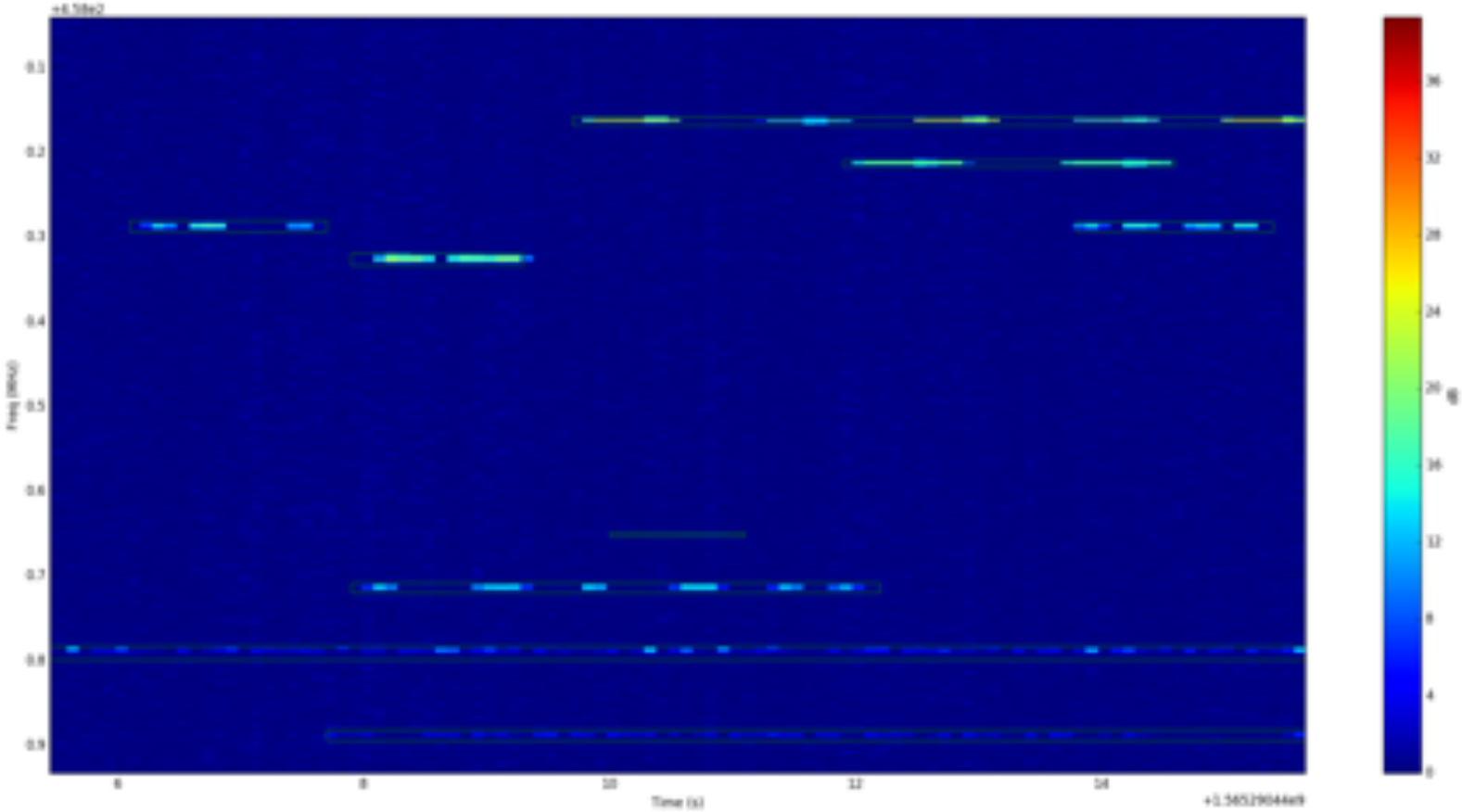
# Separate the signals by using a bounding box





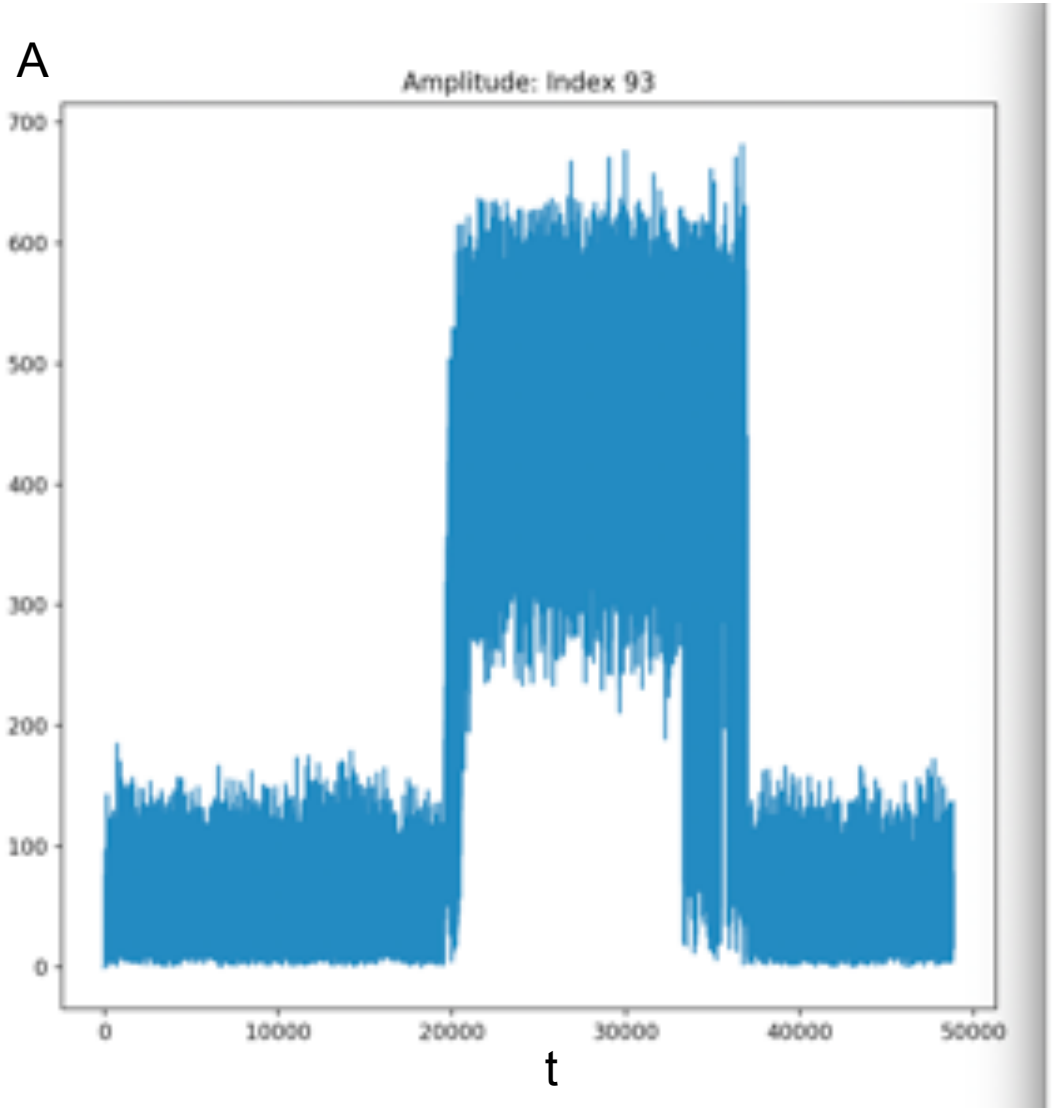

$[C_0, cx_0, cy_0, w_0, h_0]$
$[C_1, cx_1, cy_1, w_1, h_1]$
$[C_{s2}, cx_{s2}, cy_{s2}, w_{s2}, h_{s2}]$

# Results: Separated signals

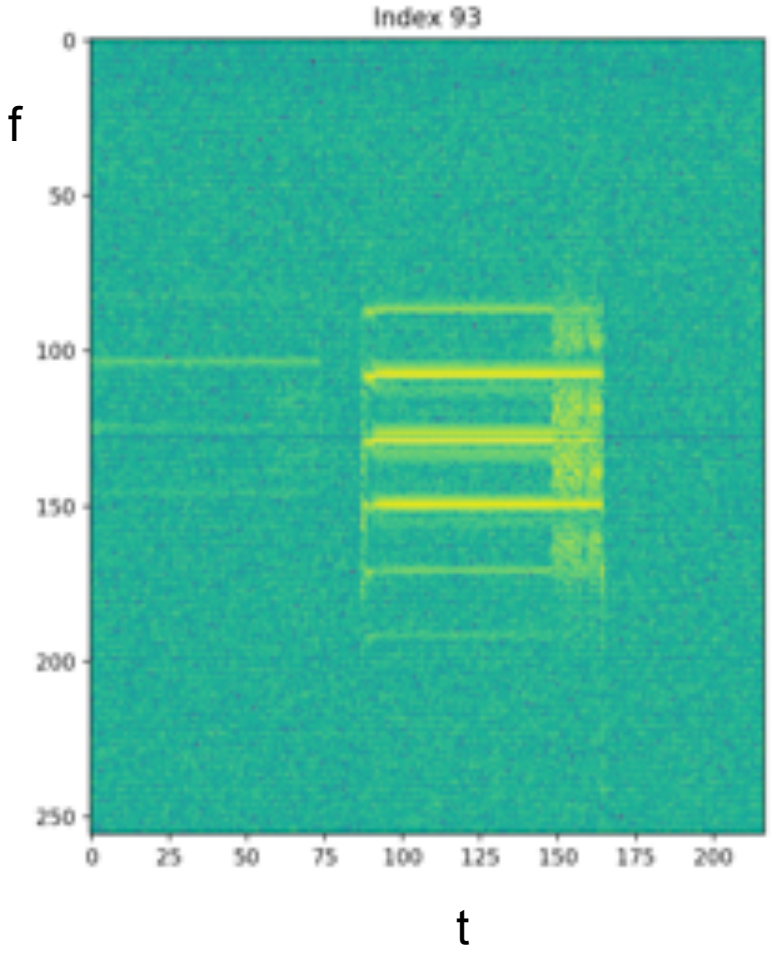


# Sample signal types

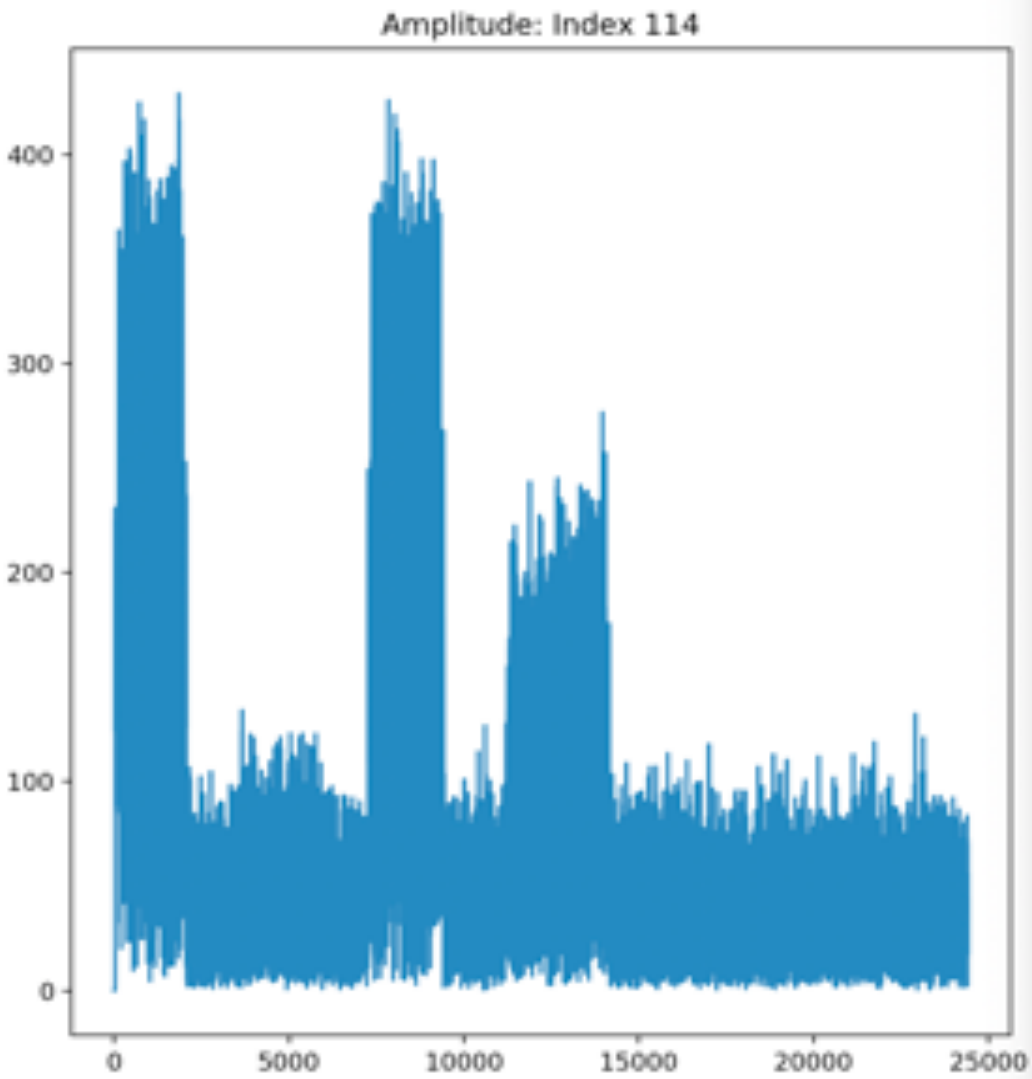
## Short tone burst

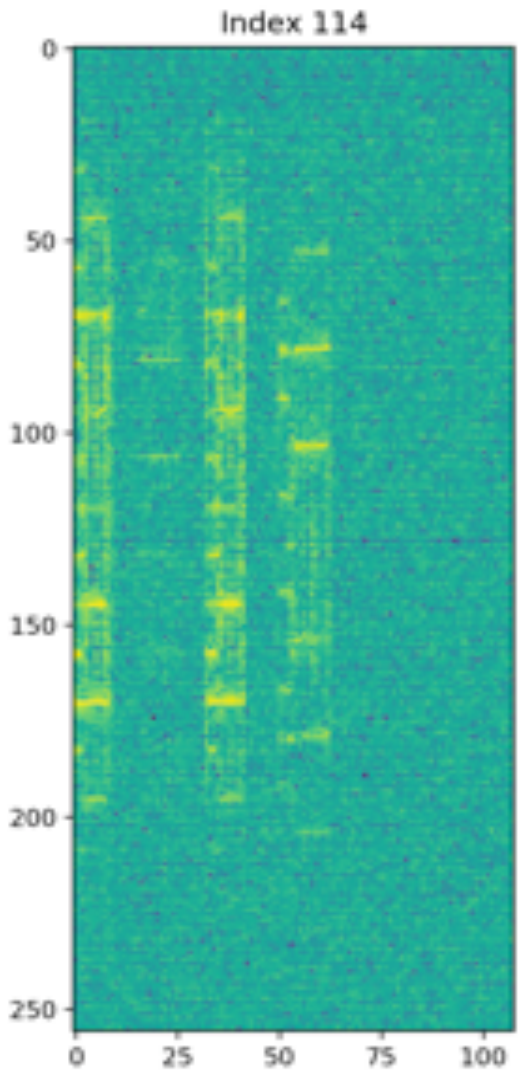



Spectrograph

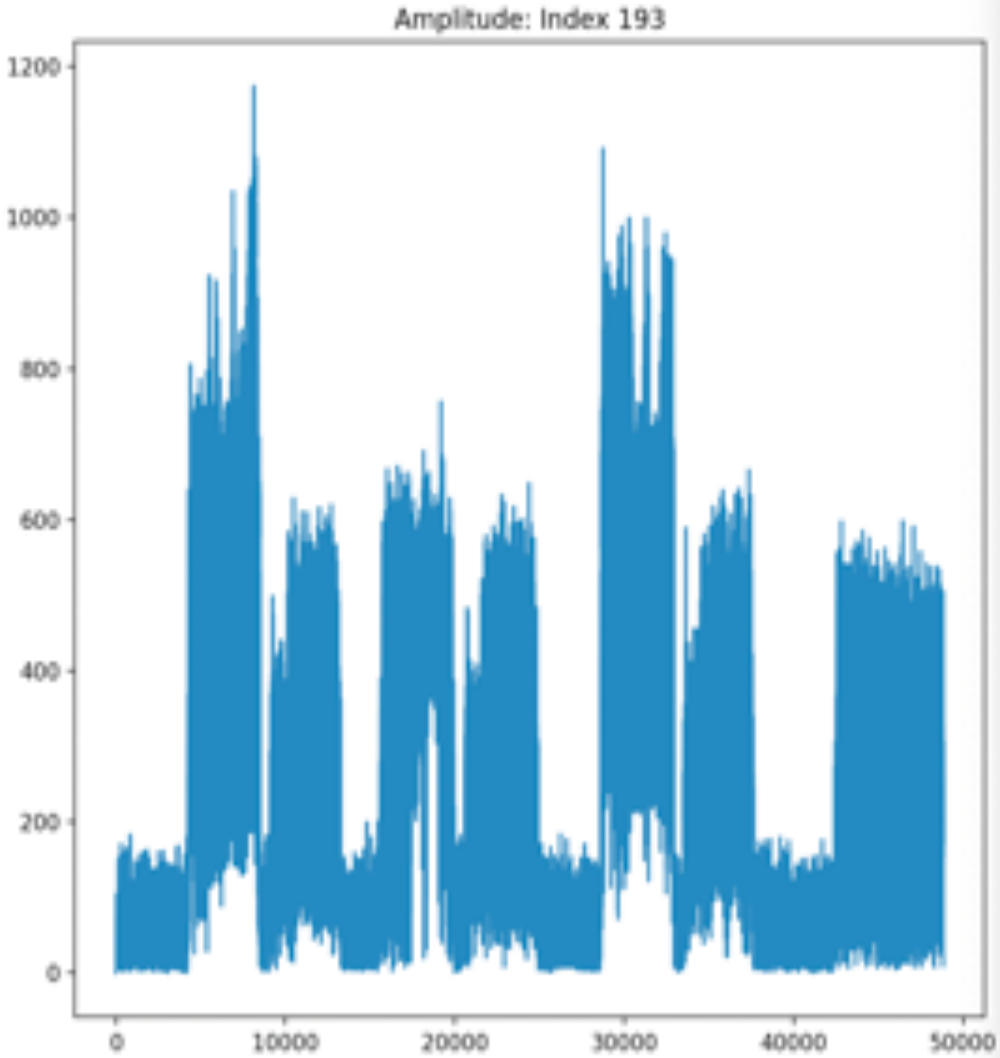


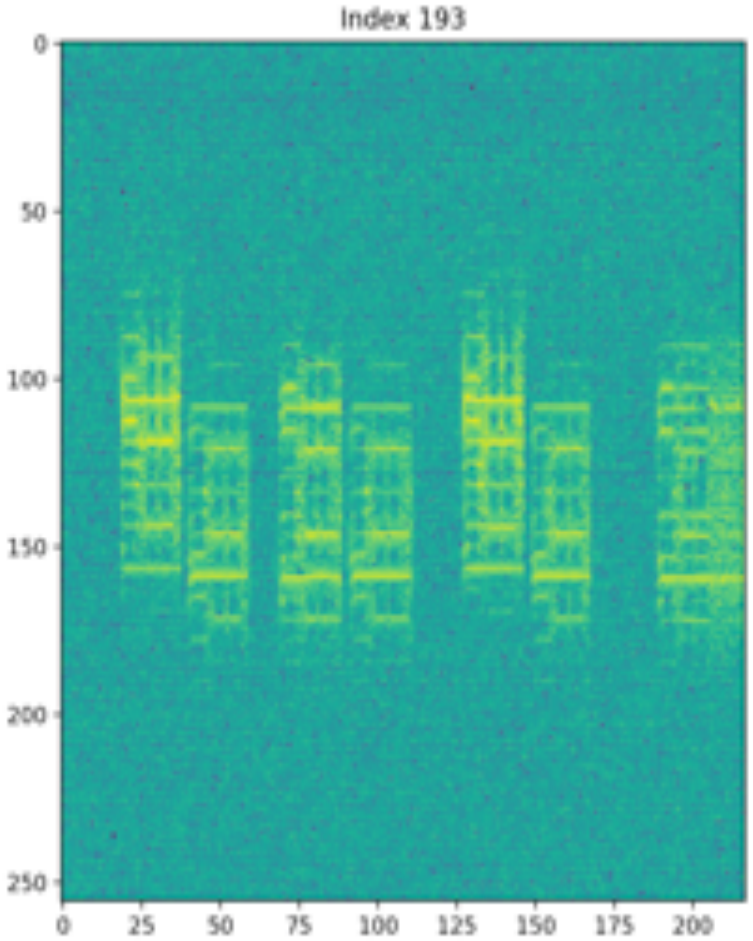
# Frequency Shift Key  (FSK)



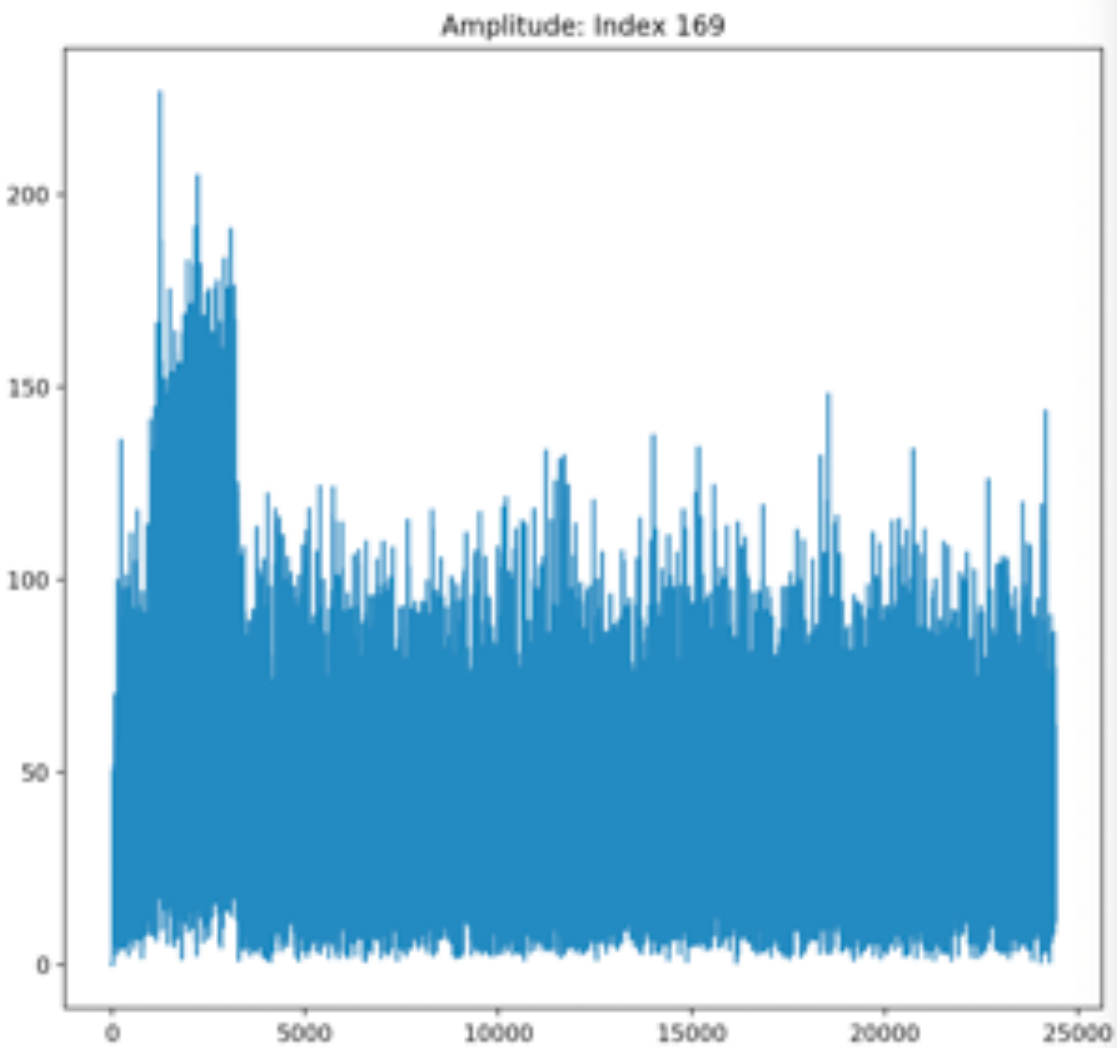



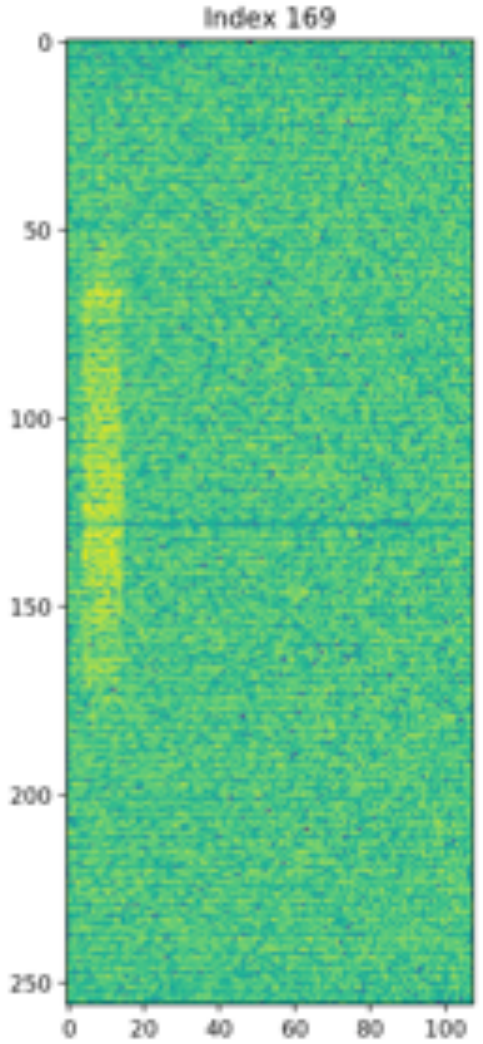
# Amplitude Shift Key  (ASK)



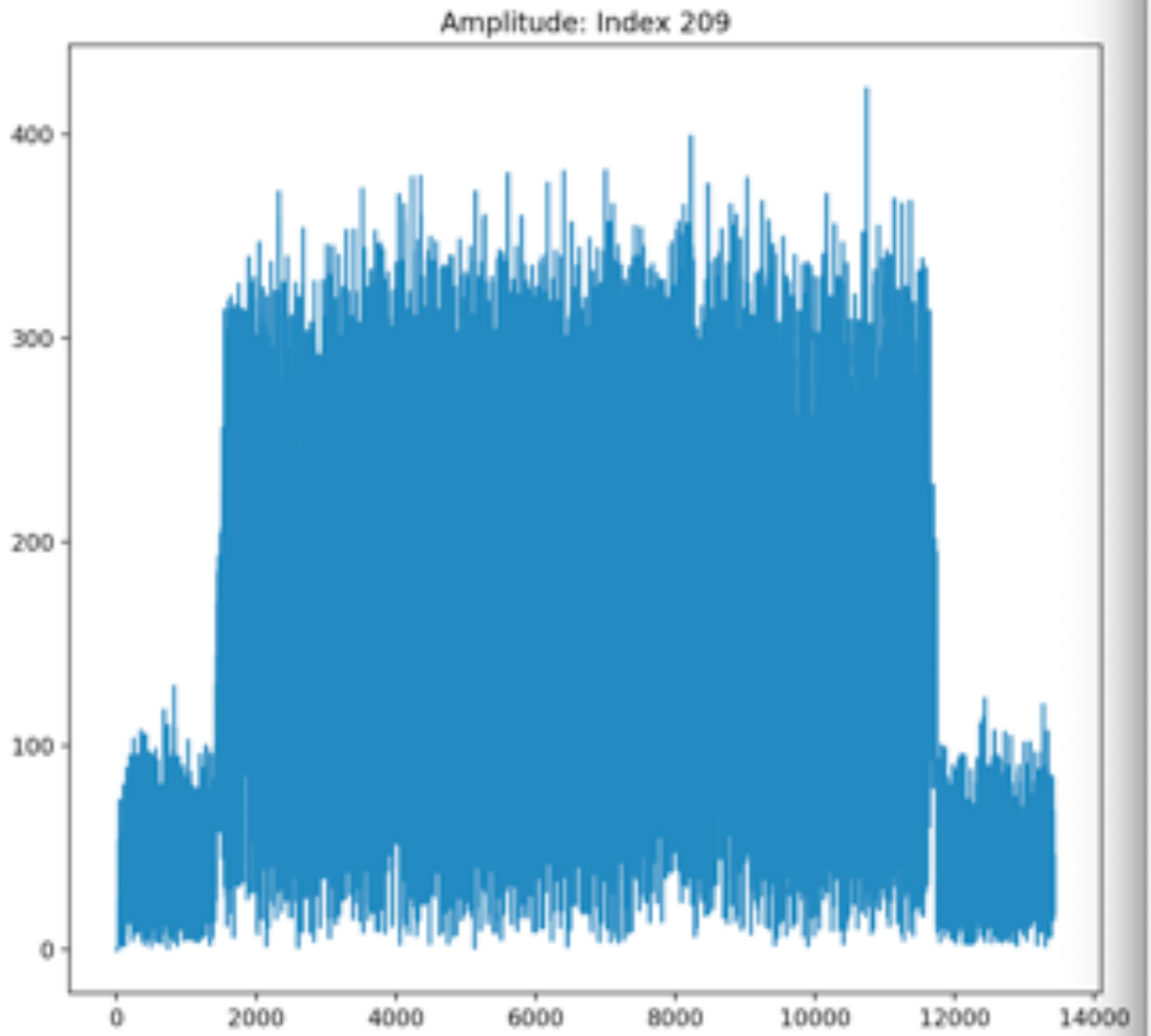



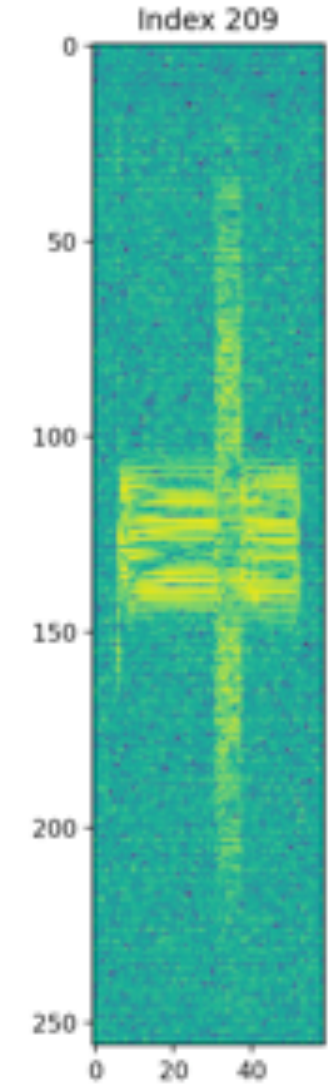
# Noisy burst


Amplitude: Index 169
200
150
100
50
0
0
5000
10000
15000
20000
25000


Index 169
0
50
100
150
200
250
0
20
40
60
80
100

# Unknown (Box)


Amplitude: Index 209
400
300
200
100
0
0
2000
4000
6000
8000
10000
12000
14000


Index 209
0
50
100
150
200
250
0
20
40

# Want to differentiate between different signals

Signals have different modulation types

eg:

Analogue: AM, SSB,DSB, FM, PM,

Digital: ASK, FSK, PSK, QAM, BPSK

Can separate signals by looking at their modulation types

# Cumulants

Higher order cumulants are claimed to be able to do this

Higher order moment

$$M_{pq} = E\left[\mathbf{u}^{p-q}(\mathbf{u}^*)^q\right]$$

Higher order cumulant

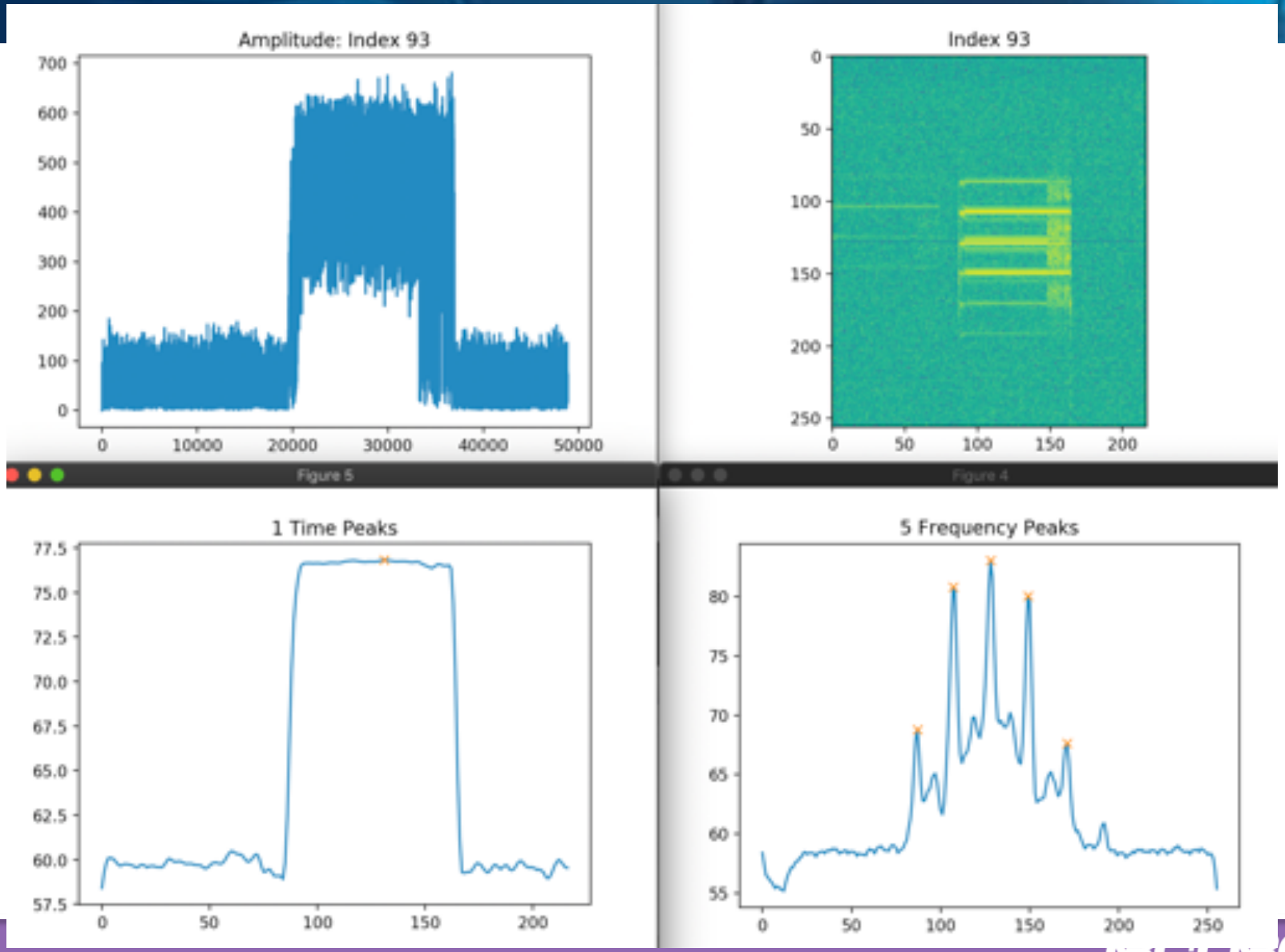
$$C_{42} = M_{42} - |M_{20}|^2 - 2M_{21}^2$$

$$C_{63} = M_{63} - 9M_{21}M_{42} + 12M_{21}^3 - 3M_{20}M_{43} - 3M_{22}M_{41} + 18M_{20}M_{21}M_{22}$$

Papers claiming this base their claims on using polynomial supervised learning from synthetic data!

We are using unsupervised learning from real data

# Also: Count the number of time and frequency peaks



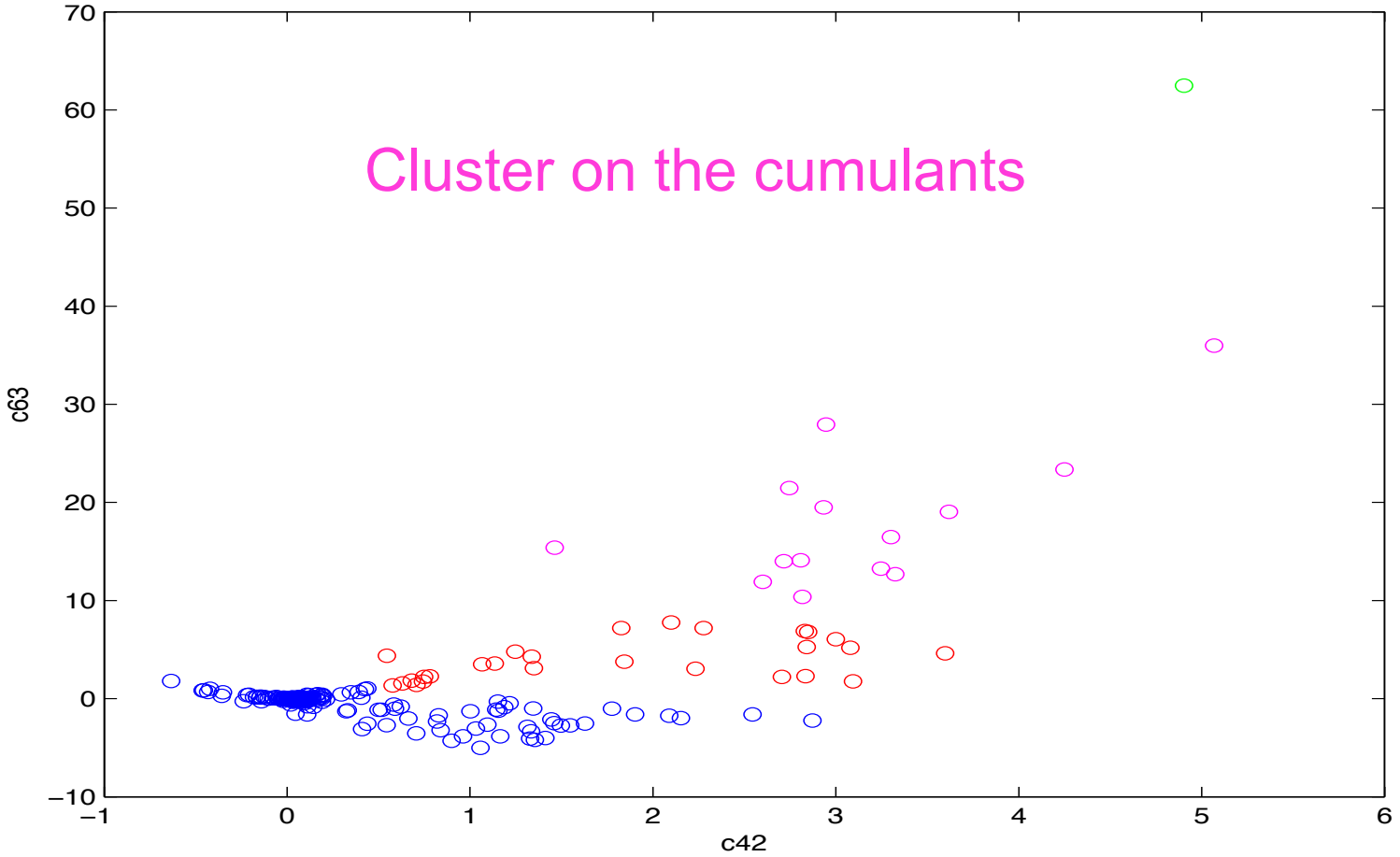
# Features vector

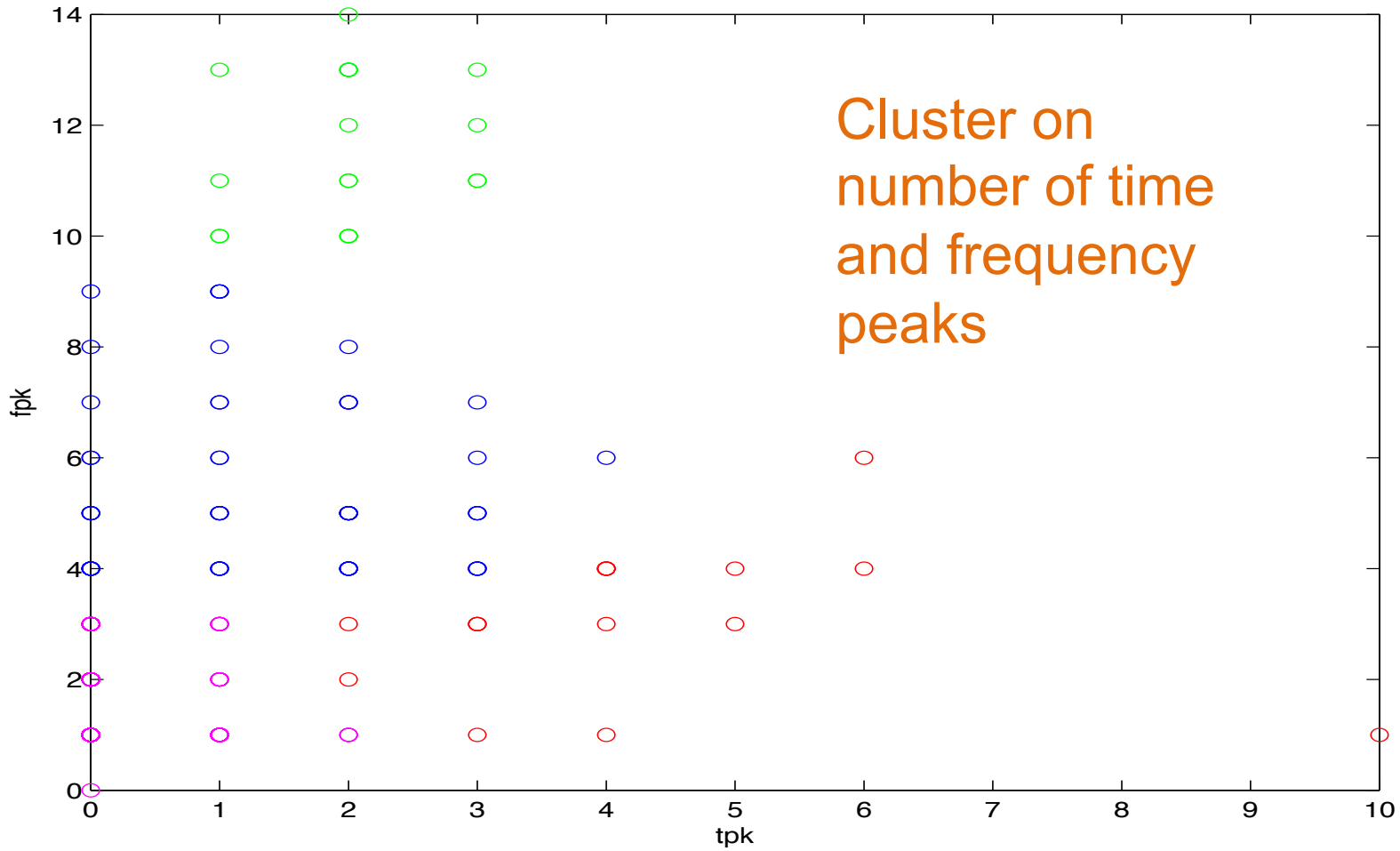
**From each detection, the following 13 features were extracted for use in clustering:**

- Center Frequency (Hz)
- Bandwidth (Hz)
- $C_{42}$ ($4^{th}$ order power cumulant)
- $C_{63}$ ($6^{th}$ order power cumulant)
- Transmission length (seconds)
- $P_{db}$ ($2^{nd}$ order power cumulant)
- Number of peaks in the frequency direction
- Number of peaks in the time direction
- Prominence of the major frequency peak
- Avg. spacing of the frequency peaks
- Avg. spacing of the time peaks
- Normalized power centroid in the time direction
- Width in the frequency direction (channels)

# Clustering the feature vectors in the data

Various algorithms exist for clustering: DBSCAN, WARD (Matlab)



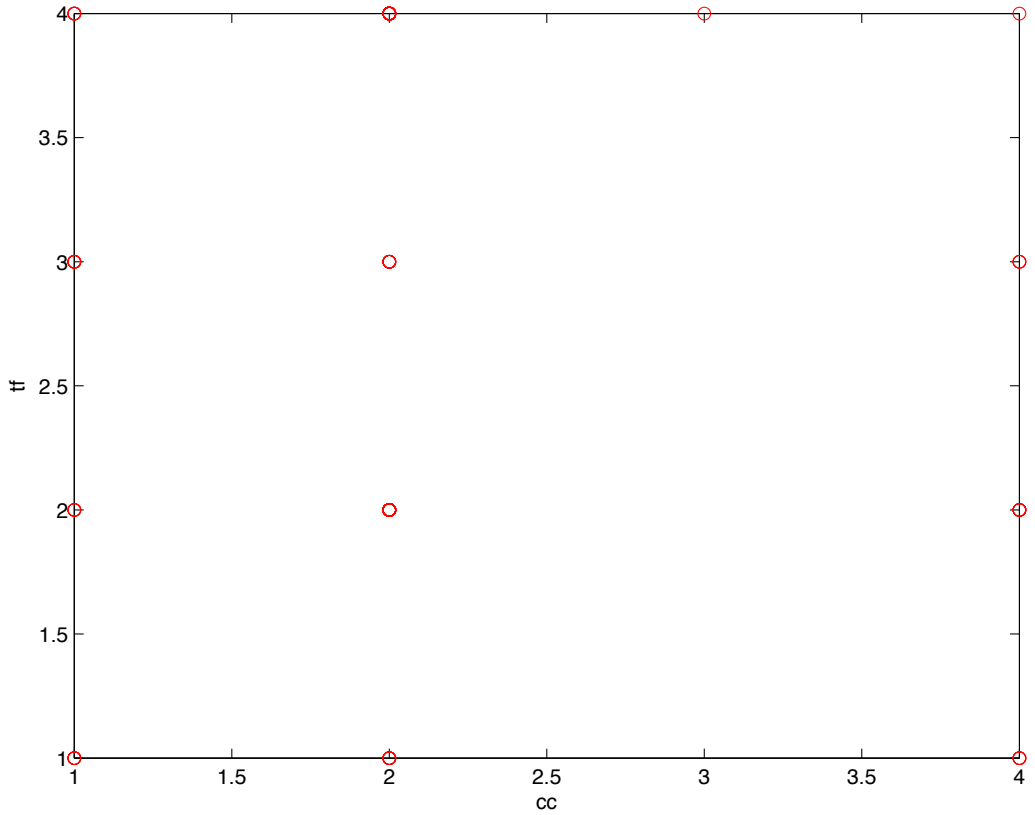



Cluster on number of time and frequency peaks

# Hierarchical clustering

The two cluster indices appear to be independent

Time-frequency cluster index: tf




Cumulant index: cc

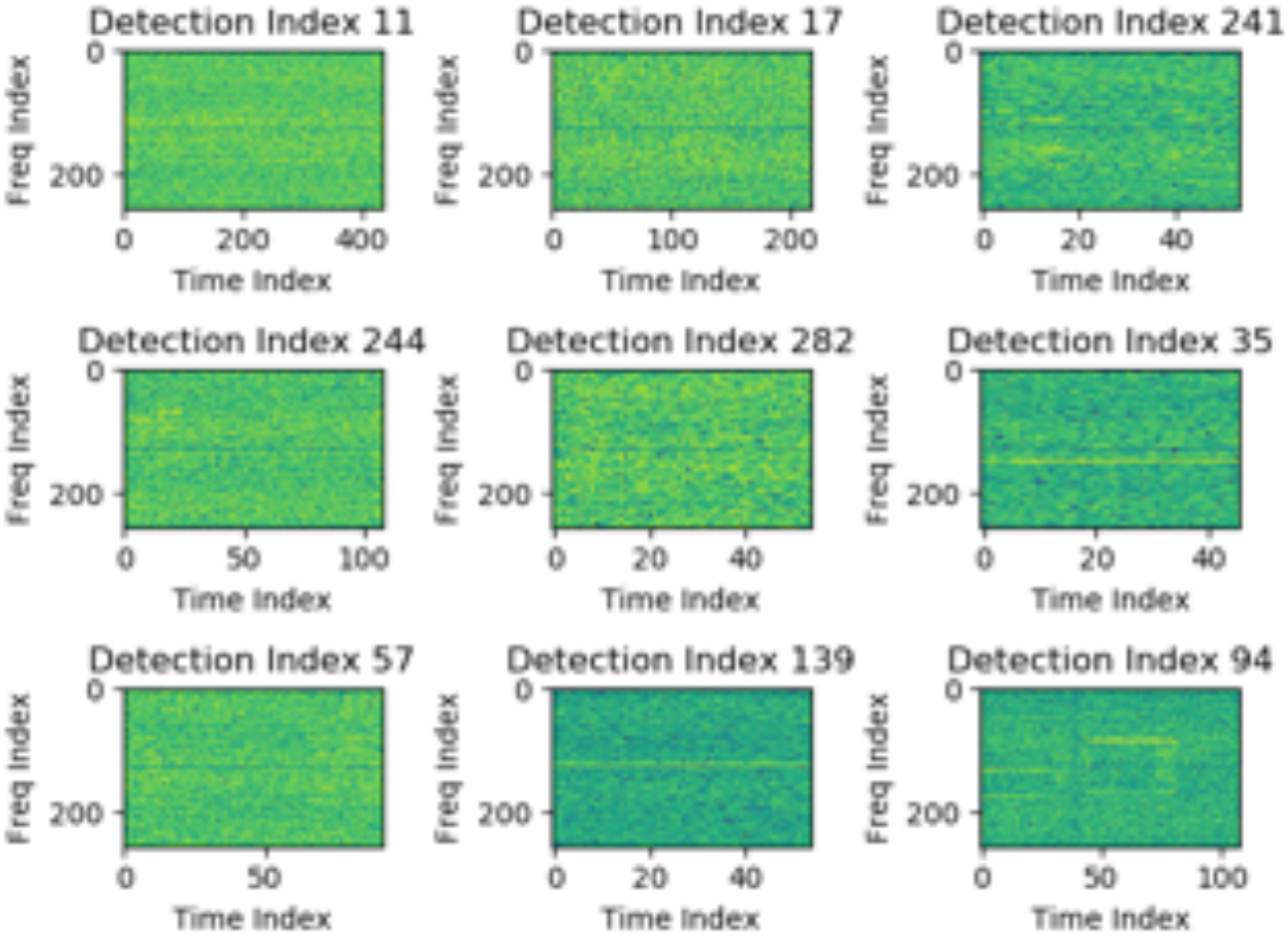
# Combine the cluster indices

Give a combined cluster index vector

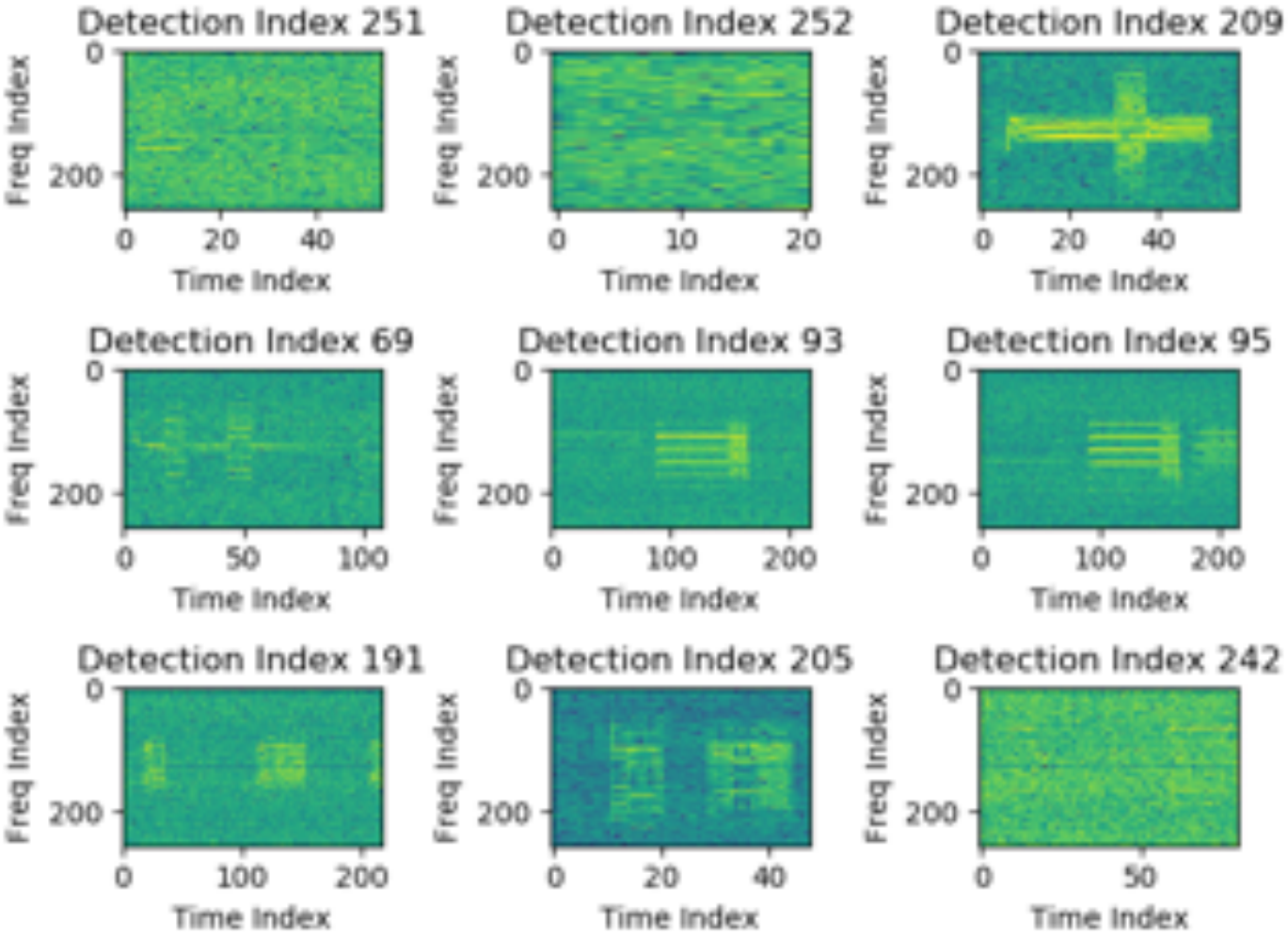
$$I = (cc, tf)$$

This seems to be effective in classifying the signals and identifying new signals

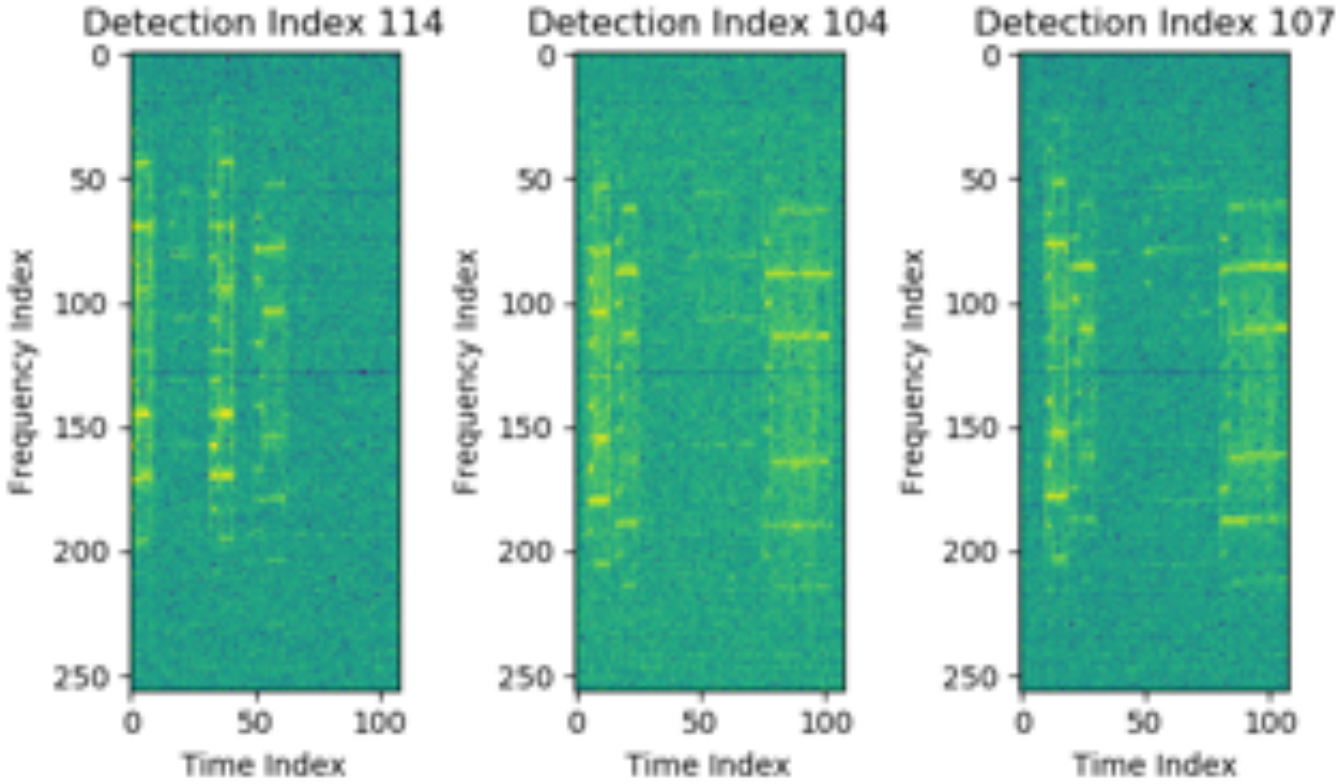
# Cluster 1: I = (2,4)



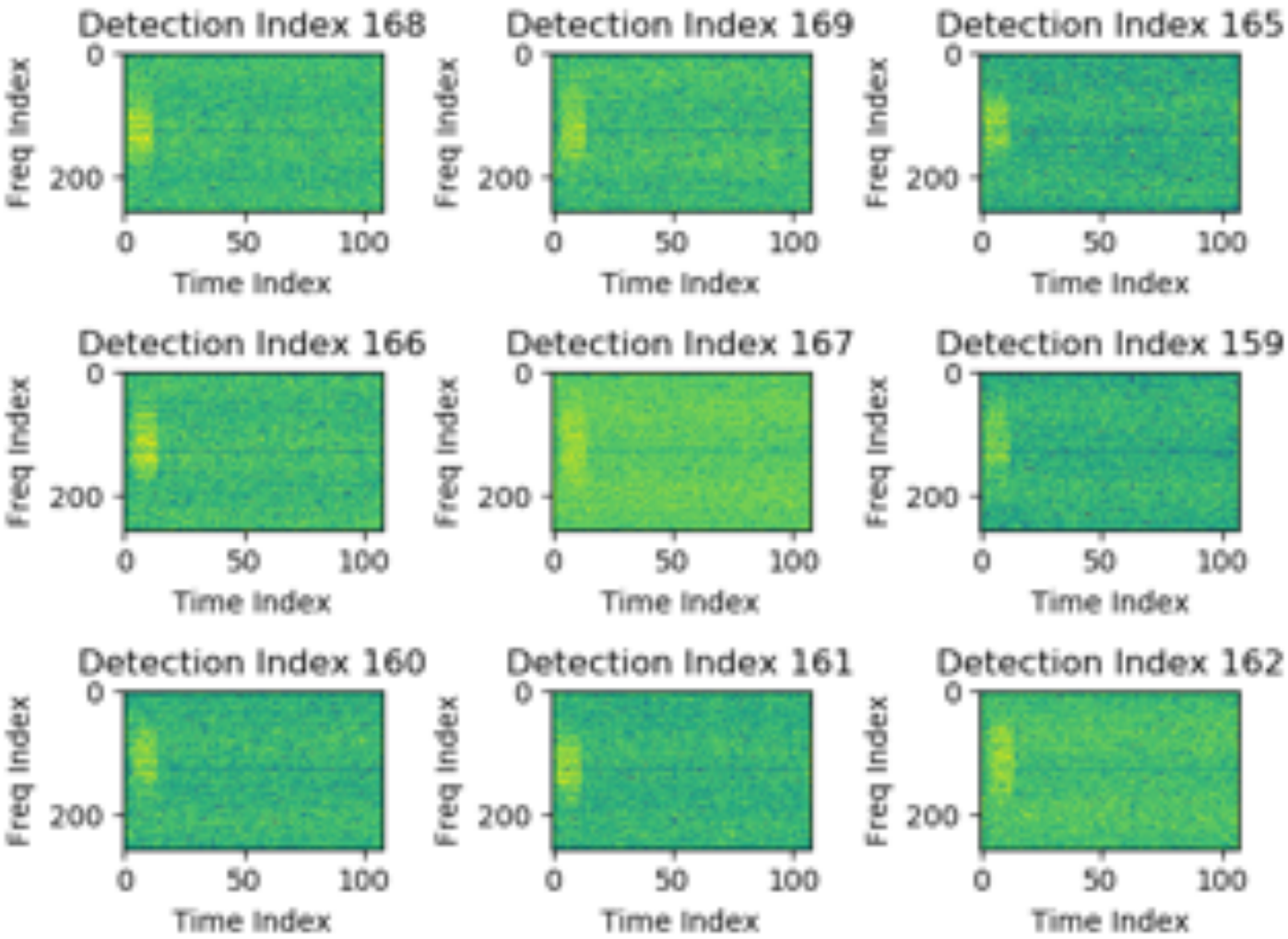
# Cluster 2: I = (2,2)



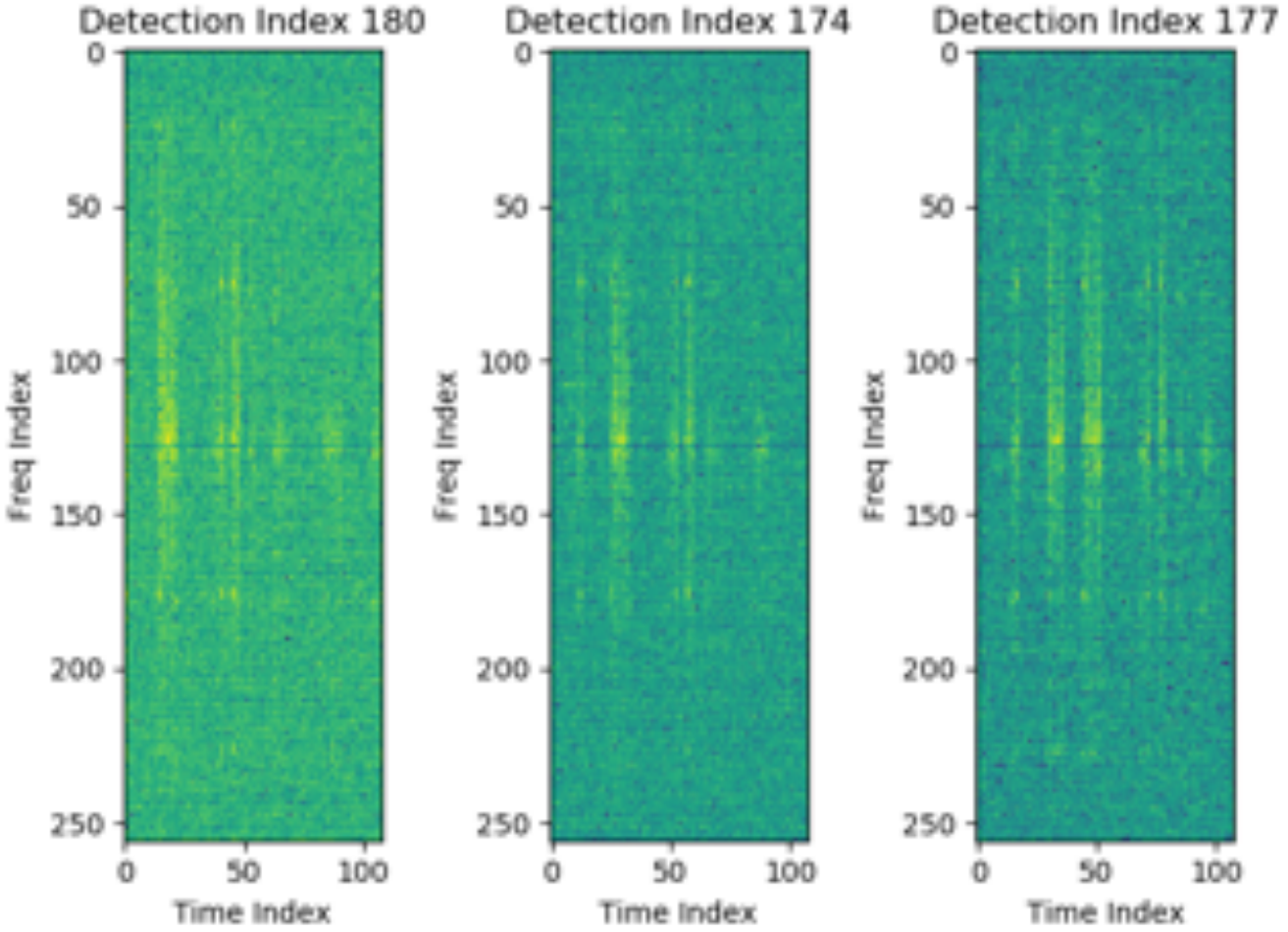
# Cluster 3: I = (1,3)



# Cluster 7: I = (1,4)



# Cluster 8: I = (1,1)

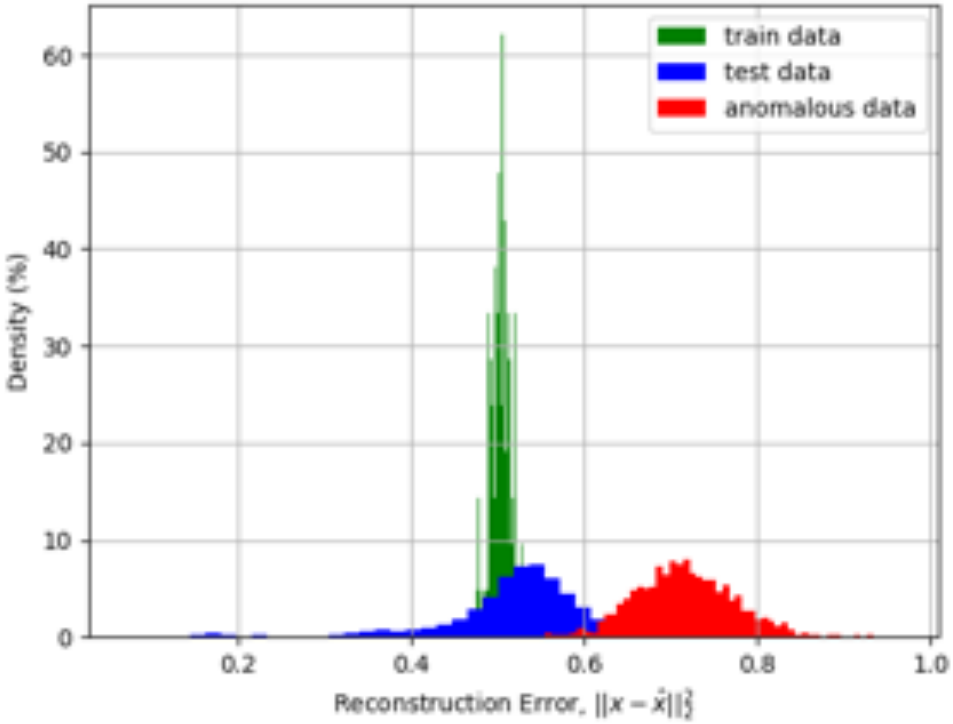


# Autoencoder approach

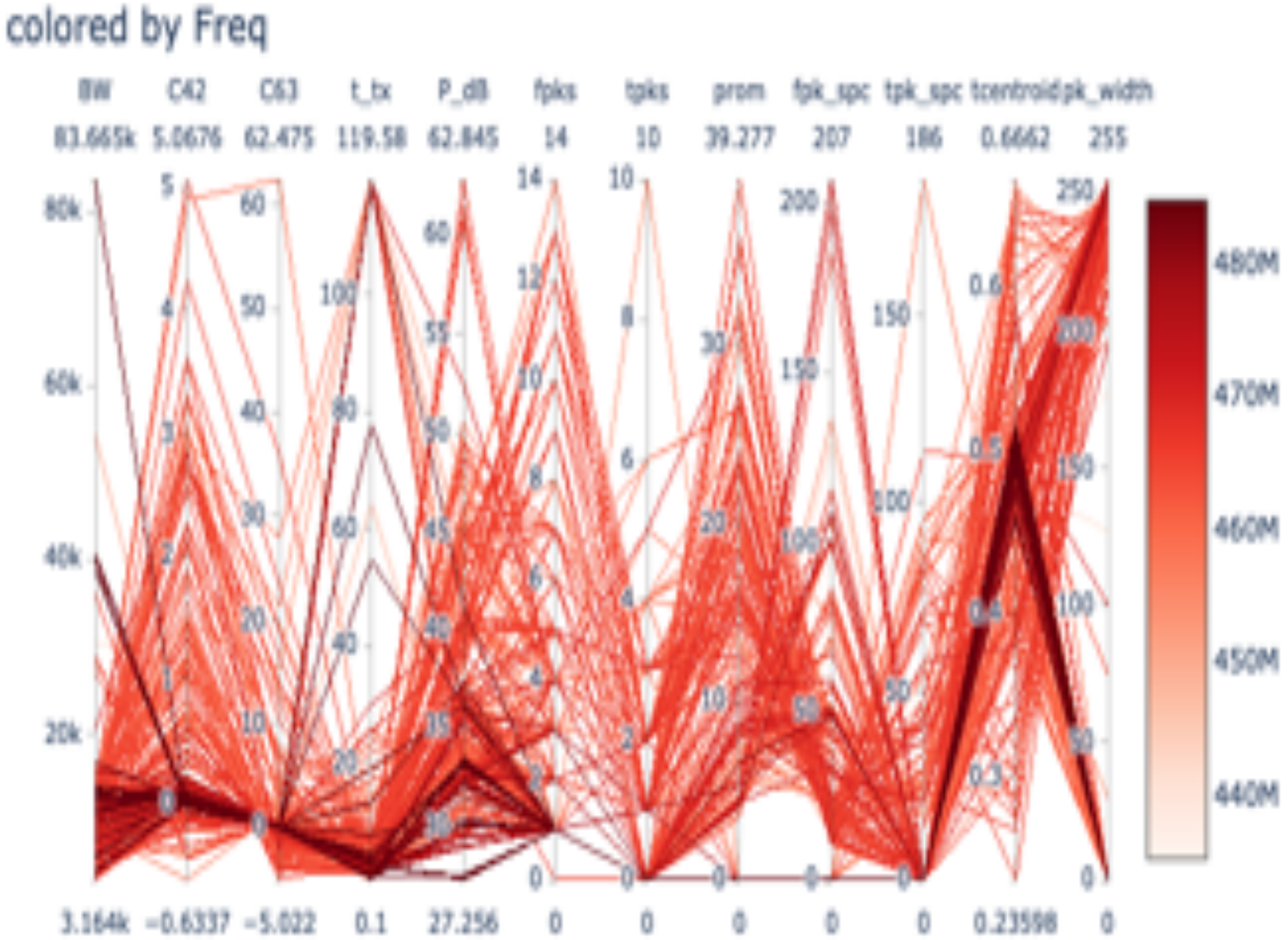
- Separately, two methods were tried using a convolutional autoencoder on the baseband time-series data:
  - The first idea is to shove all the signals into the encoder, and use the latent feature space for clustering (in progress)
  - The second idea is to not cluster, instead training the autoencoder on the full dataset, then compute the difference between the input and the reconstruction. If the error is large enough, it's "anomalous".

# Autoencoder approach

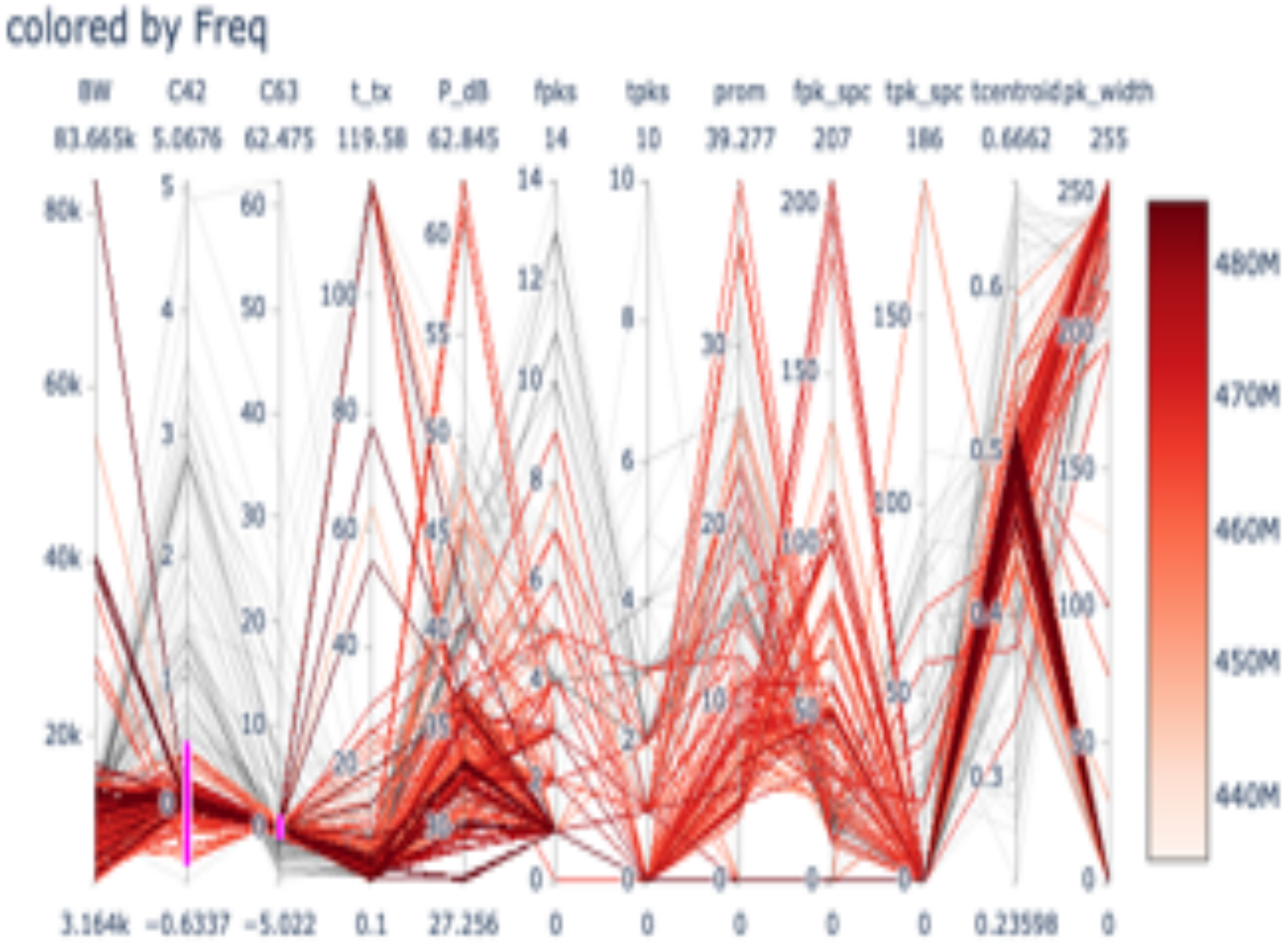
- First it was trained on all of the dataset with a $P_{db}$ greater than 40 dB, then compared with simulated pure white noise (on the right)
- Next a collection of real signals known to be noisy was used for training, and then the rest were for testing (in progress)
- The following slides show the process of separating those signals using a parallel coordinates plot



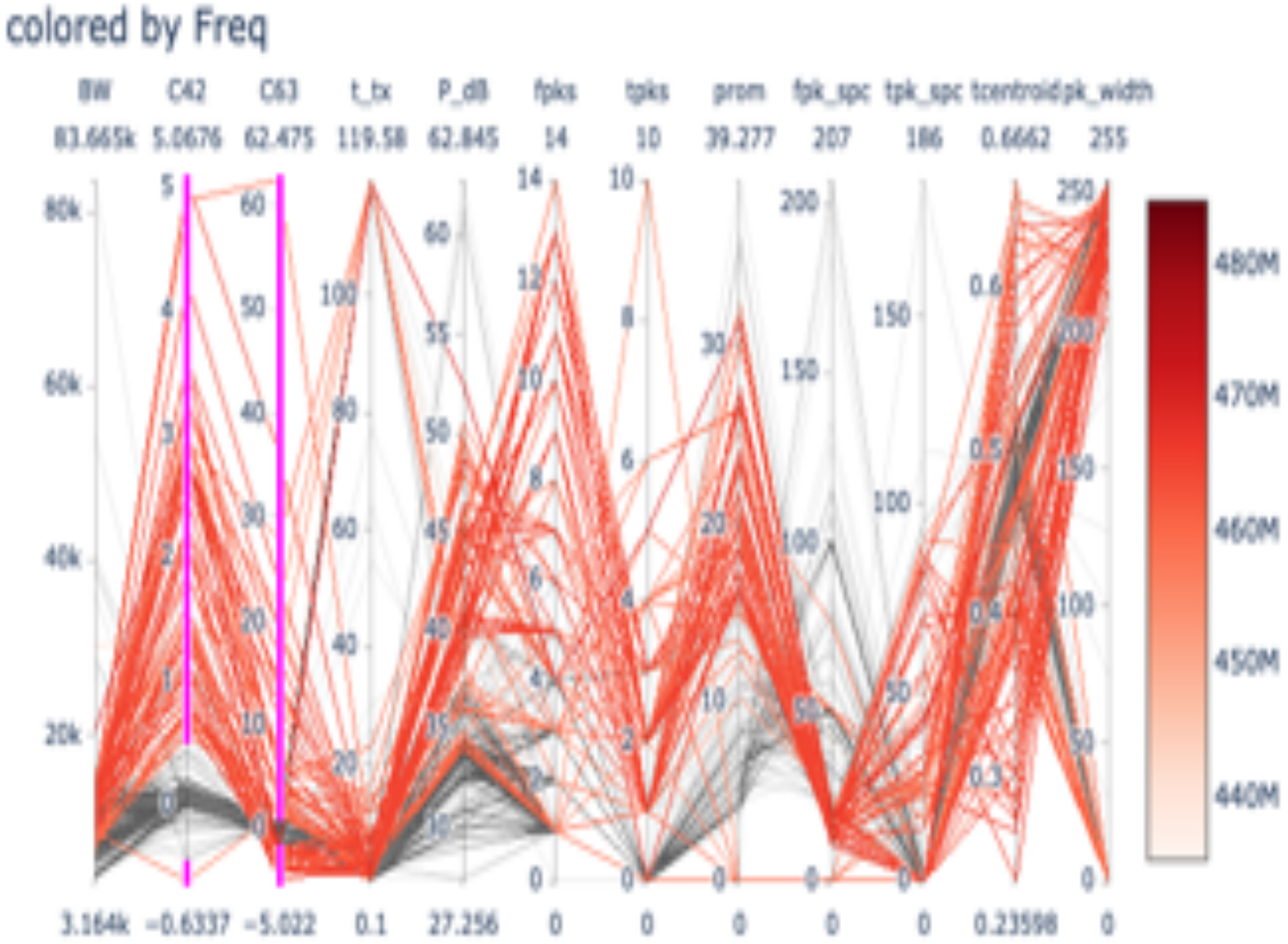
# Full dataset



# Dataset for training (see pink)



# Data for testing (see pink)



# Conclusions

Building on the unsupervised detection, it was found that cumulants C42 and C63 alone were insufficient to cluster incoming RFI signals.

This is in contrast to other researcher claims that cumulants are sufficient. However this claim was based on supervised learning algorithms using synthetic data and not unsupervised leaning on real data.

Cumulants combined with extra information of the number of peaks in the time domain and the frequency domain, were sufficient to classify the incoming signals.

When tested on a subsequent dataset, this scheme detected a new cluster providing some confidence as to its future capability.

# Future work

Autoencoder implementations are currently in progress.

1) One uses all the data and is attempting to cluster based on the latent feature space.

2) A second also uses all the data but tried to reconstruct the signal.

An inability to be able to do this task would identify a signal as novel.

A system that combines the bounding box technique to separate signals with the clustering is in the process of being implemented.