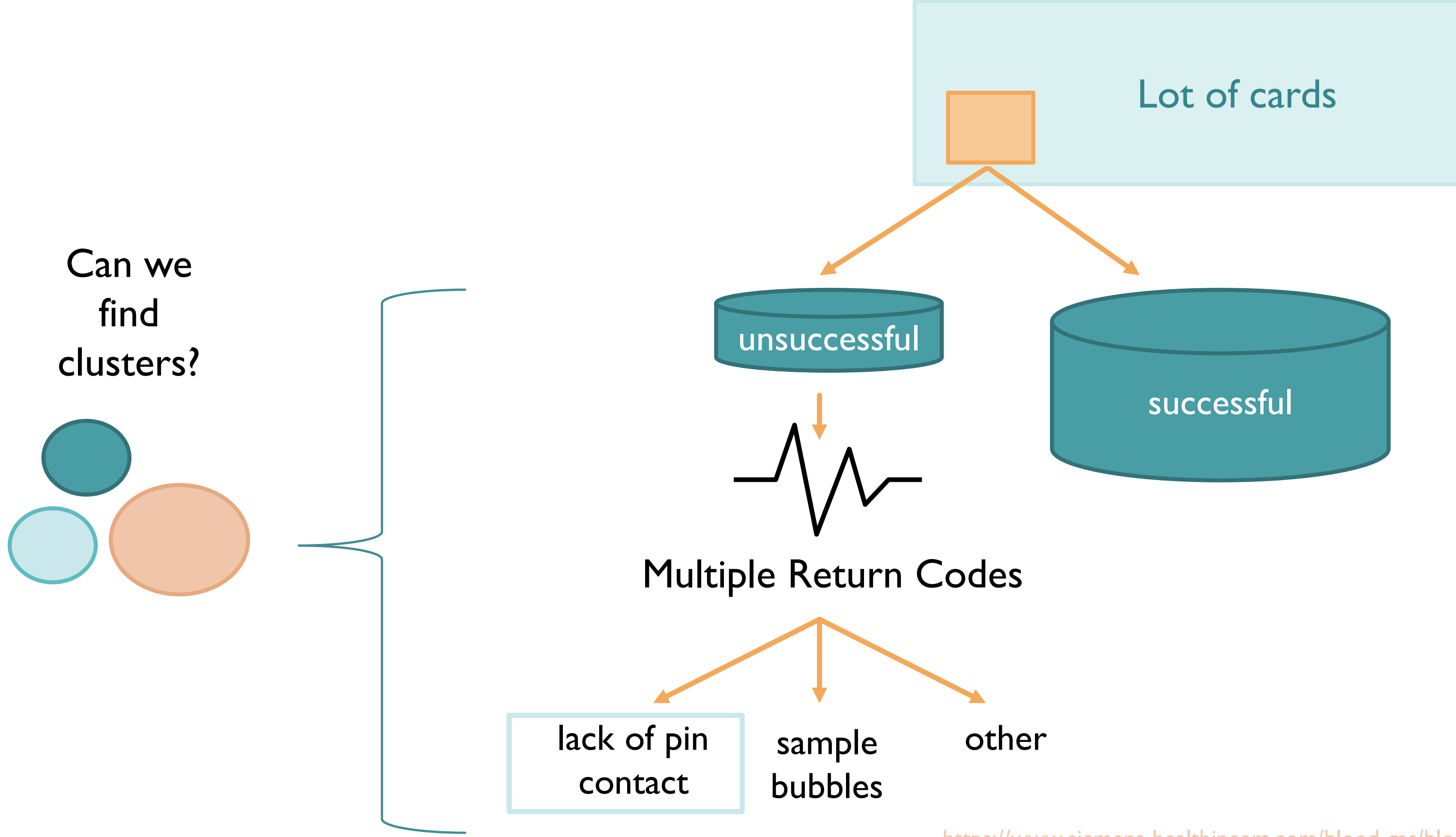
MAY 24 UPDATE

JUSTINE FILION, NEETHU GOPALAKRISHNA, SAISREE GR, SARA HALL

RECAP: ANOMALY DETECTION IN BIOSENSOR WAVEFORMS



Can we find clusters?

Lot of cards

unsuccessful

successful

Multiple Return Codes

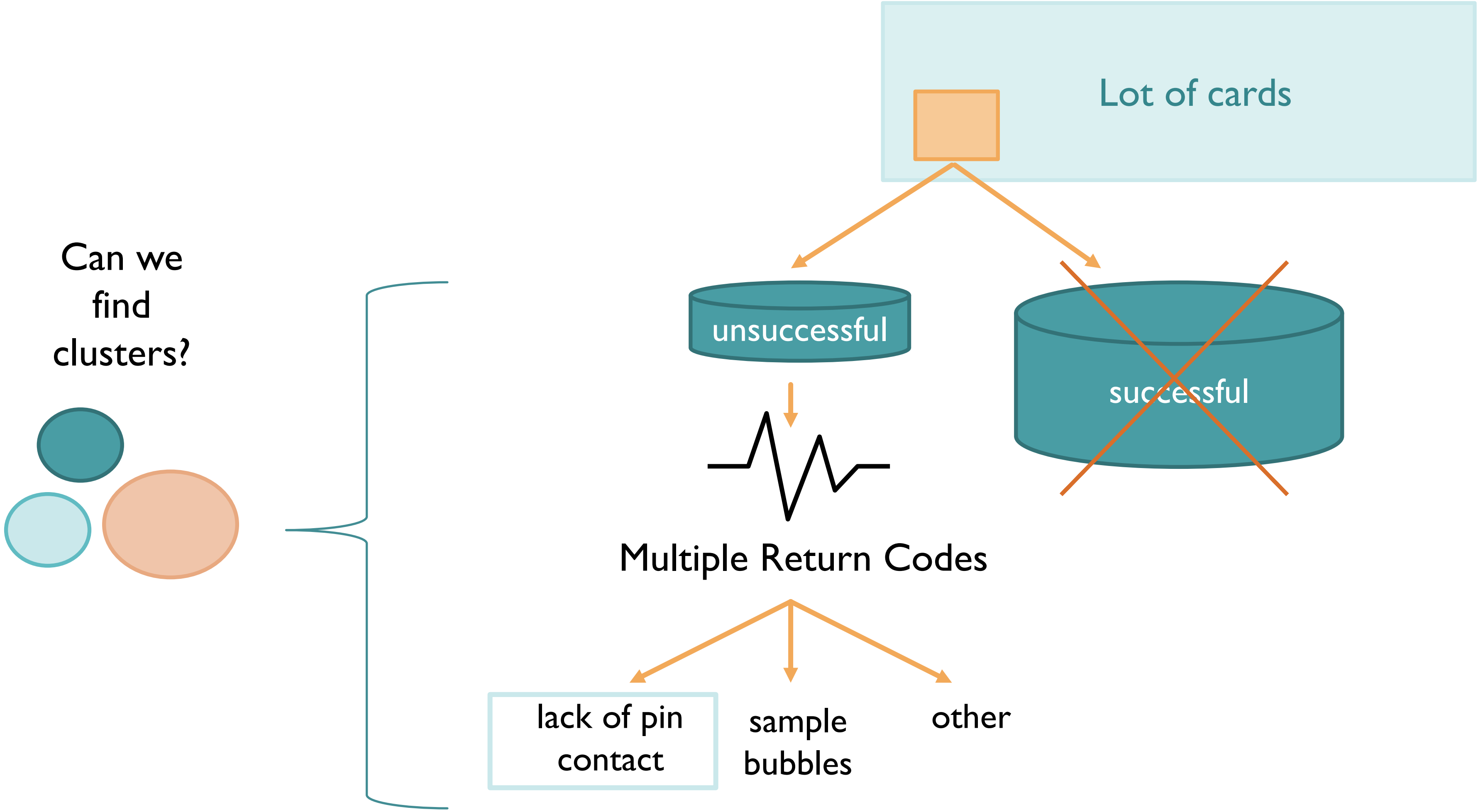
lack of pin contact

sample bubbles

other

https://www.siemens-healthineers.com/blood-gas/blood-gas-systems/epoc-blood-analysis-system

RECAP: ANOMALY DETECTION IN BIOSENSOR WAVEFORMS



RECAP: ANOMALY DETECTION IN BIOSENSOR WAVEFORMS

Research Questions:

Can we develop machine learning pipelines to cluster readings and isolate pin contact errors?

Determine which methods are effective and which are not for identifying anomalies in biosensor readings?

Deliverables:

- Well commented Python code for everything we have tried
- A final report detailing our attempts

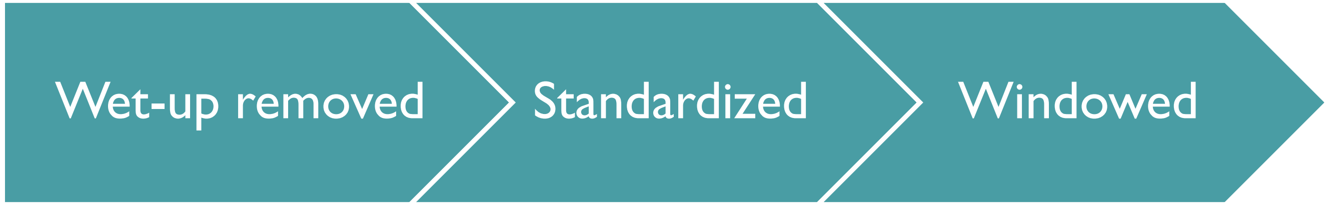
PROGRESS DURING THIS WEEK'S CYCLE

OVERALL PROGRESS

- Preprocessing
  - Time series standardization
  - Time series windowing
- Modelling
  - Visualization/clustering attempts on whole windows.
  - Clustering attempts on predictors from the tsfresh package[1]
  - Feature generation with an autoencoder
  - Clustering attempts with a self-organizing map

[1] https://tsfresh.readthedocs.io/en/latest/

PREPROCESSING - WINDOWING



Wet-up removed

Standardized

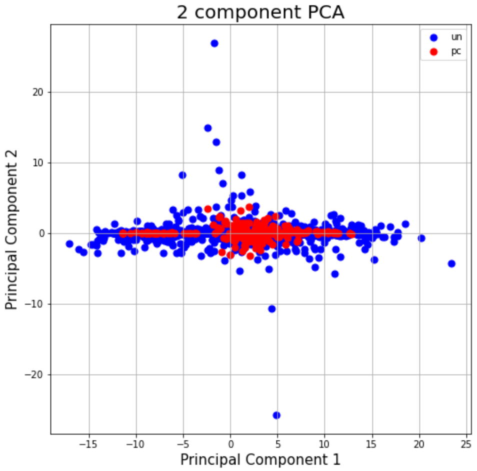
Windowed

- Got rid of unsuccessful readings with sample detect time 0 (all had return code = cannot calculate)
- Windowed w.r.t sample detect time:
  - Calibration: -15 to -3 seconds
  - Post: 12 -16 seconds
  - Sample: 32 – 35 seconds

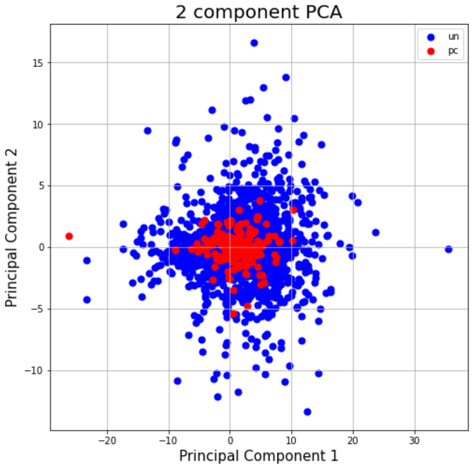
Excluded readings with empty windows (all unsuccessful)

MODELLING – WHOLE WINDOWS

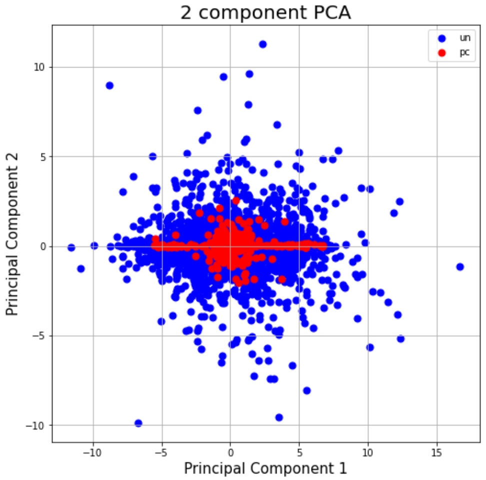
- PCA to get a 2D Representation of the readings:



Calibration Window



Post Window

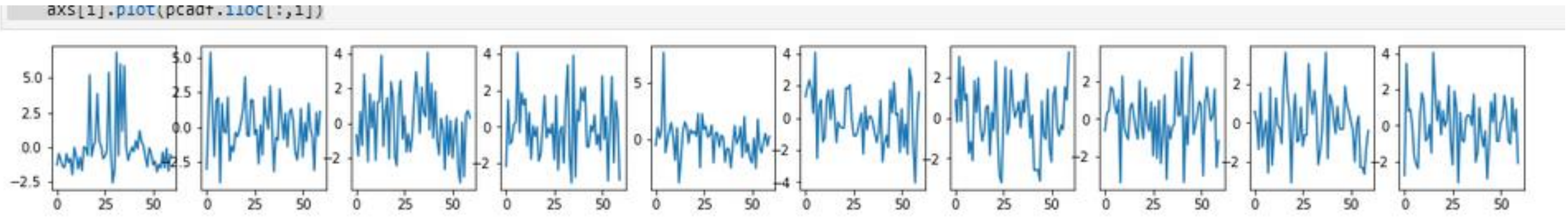


Sample Window

MODELLING – WHOLE WINDOWS

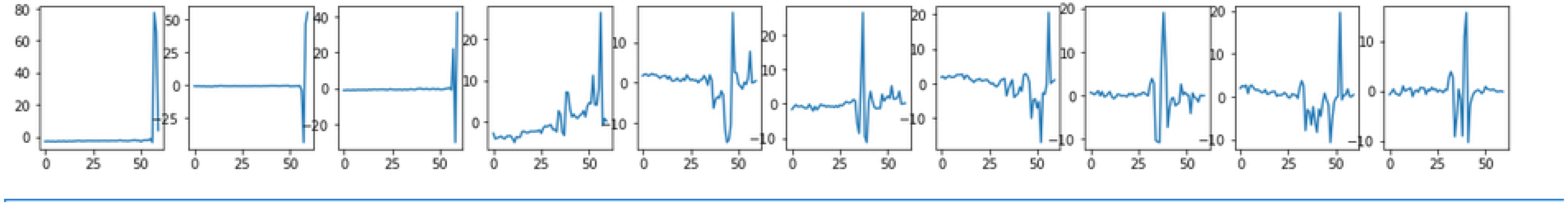
- PCA to get 'representative' readings – calibration window

Pin



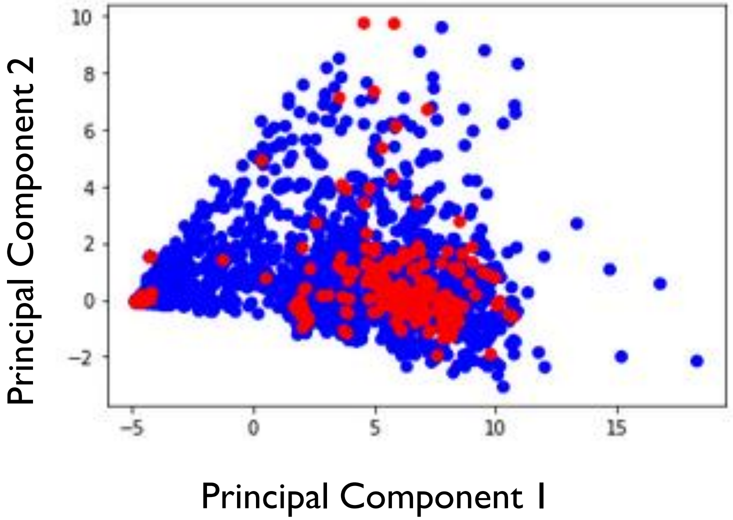
```
axs[1].plot(pcadf.iloc[:,1])
```

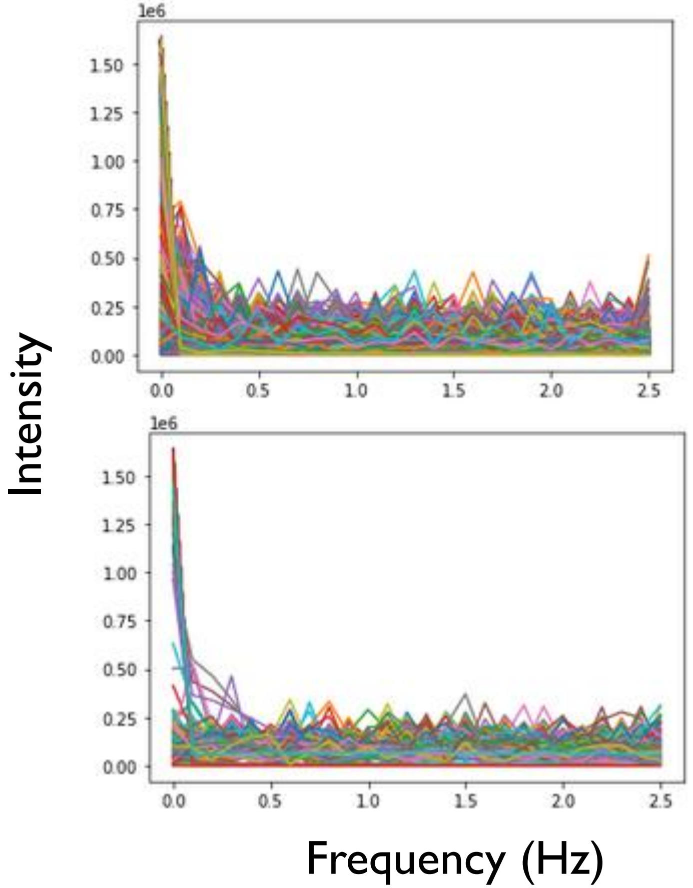
Un



MODELLING – WHOLE WINDOWS

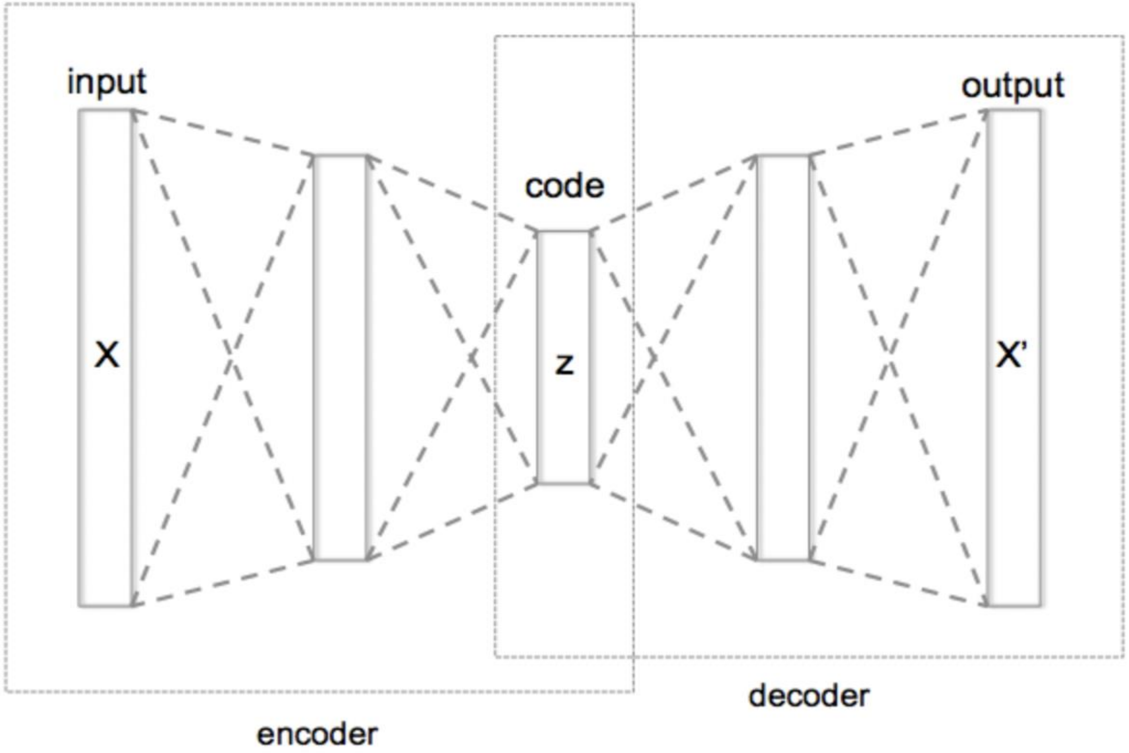
- Frequency Representation – calibration



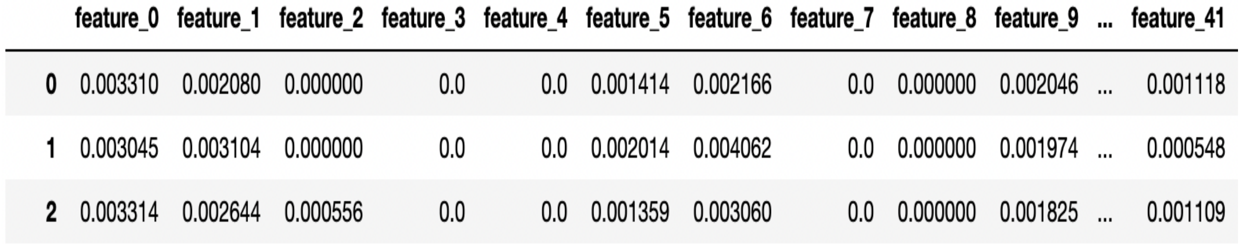


MODELLING - AUTOENCODER

- Dimensionality reduction
- Extract features



MODELLING - AUTOENCODER

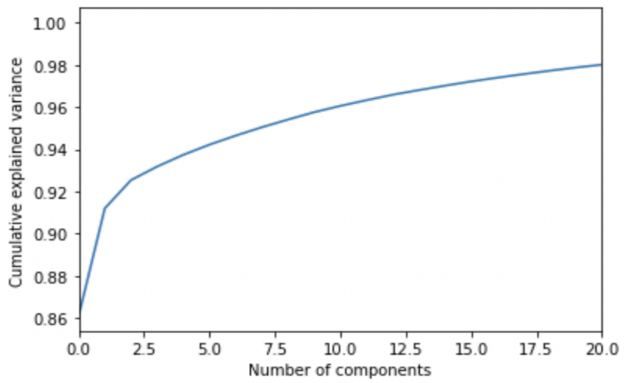


	feature_0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	...	feature_41
0	0.003310	0.002080	0.000000	0.0	0.0	0.001414	0.002166	0.0	0.000000	0.002046	...	0.001118
1	0.003045	0.003104	0.000000	0.0	0.0	0.002014	0.004062	0.0	0.000000	0.001974	...	0.000548
2	0.003314	0.002644	0.000556	0.0	0.0	0.001359	0.003060	0.0	0.000000	0.001825	...	0.001109

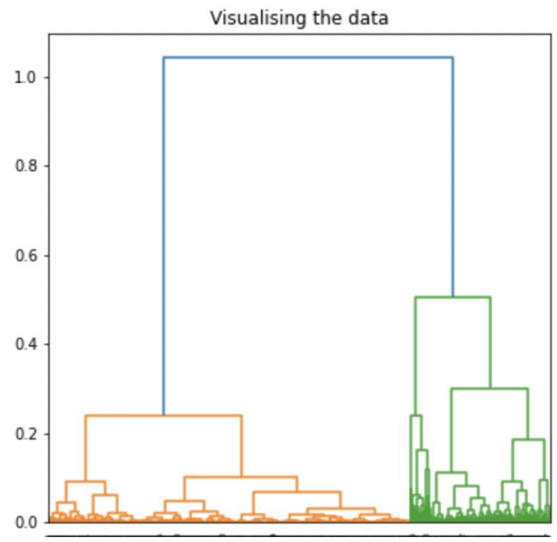
- Features obtained by using autoencoder.(~100)
  - Dropped features with zero value.

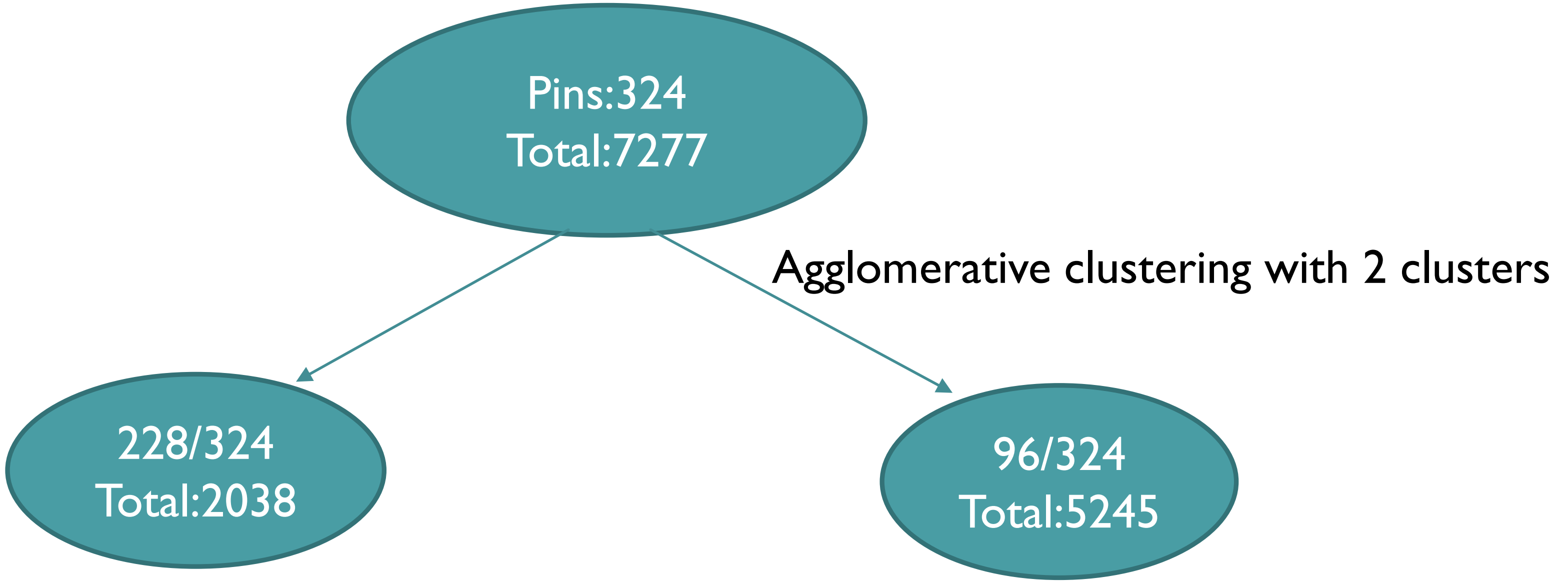
- PCA on the features obtained, variation explained = 0.95.
  - 8 Principal components selected.

```
Text(0, 0.5, 'Cumulative explained variance')
```



MODELLING - AUTOENCODER





Pins:324
Total:7277

Agglomerative clustering with 2 clusters

228/324
Total:2038

96/324
Total:5245

MODELLING – SELF ORGANIZING MAP

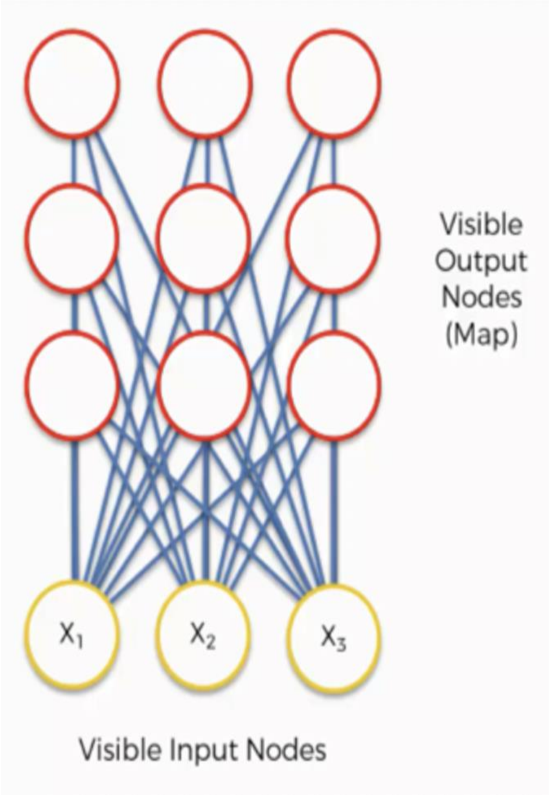
TRAINING:

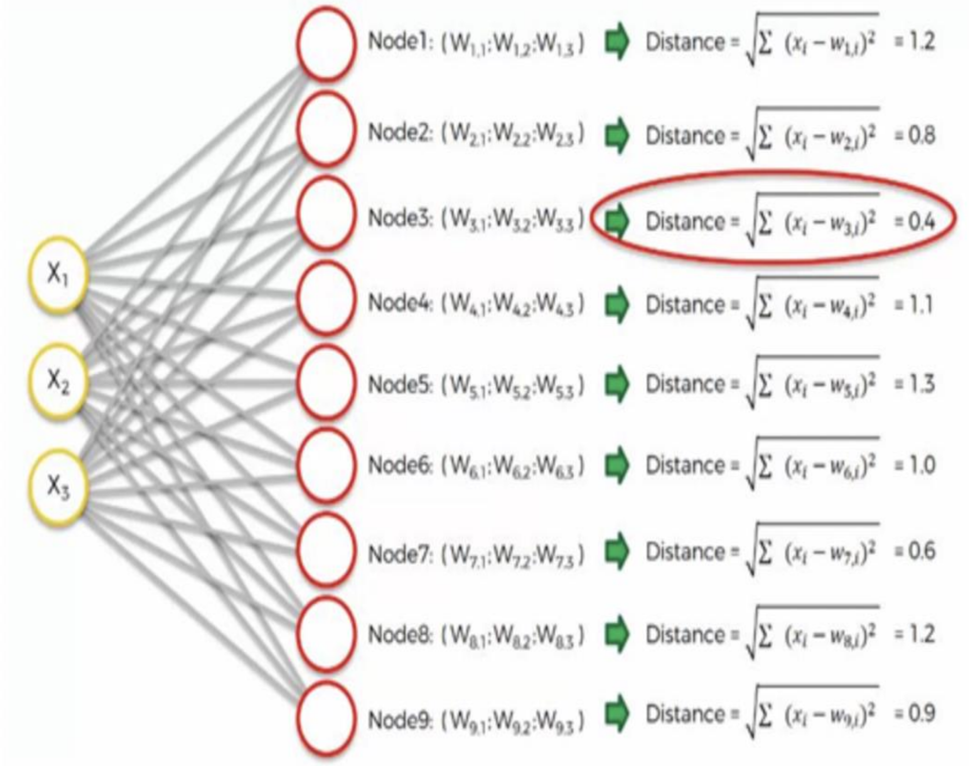
1. Weight initialisation

2. Choosing vector input randomly

3. Choosing Best Matching Unit

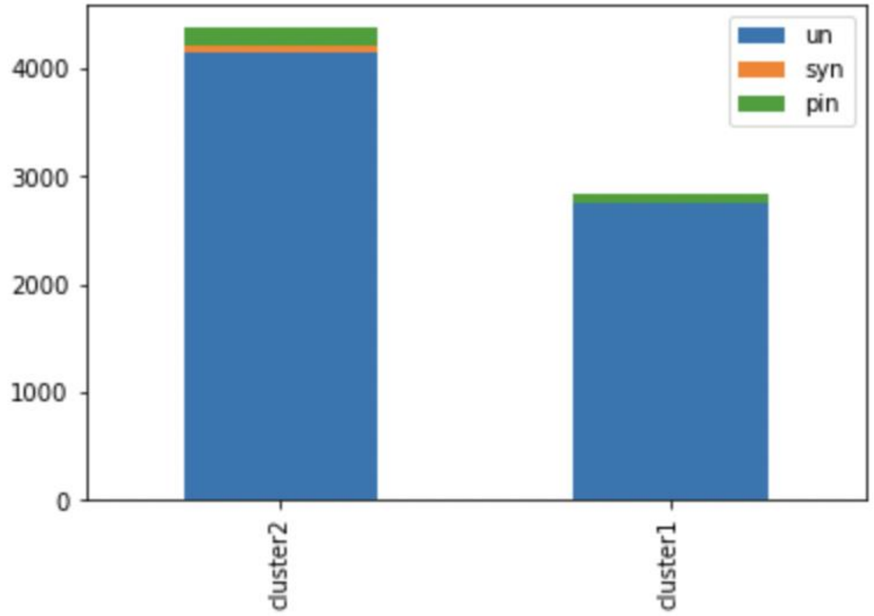
4. Repeat 2 and 3 for all data points



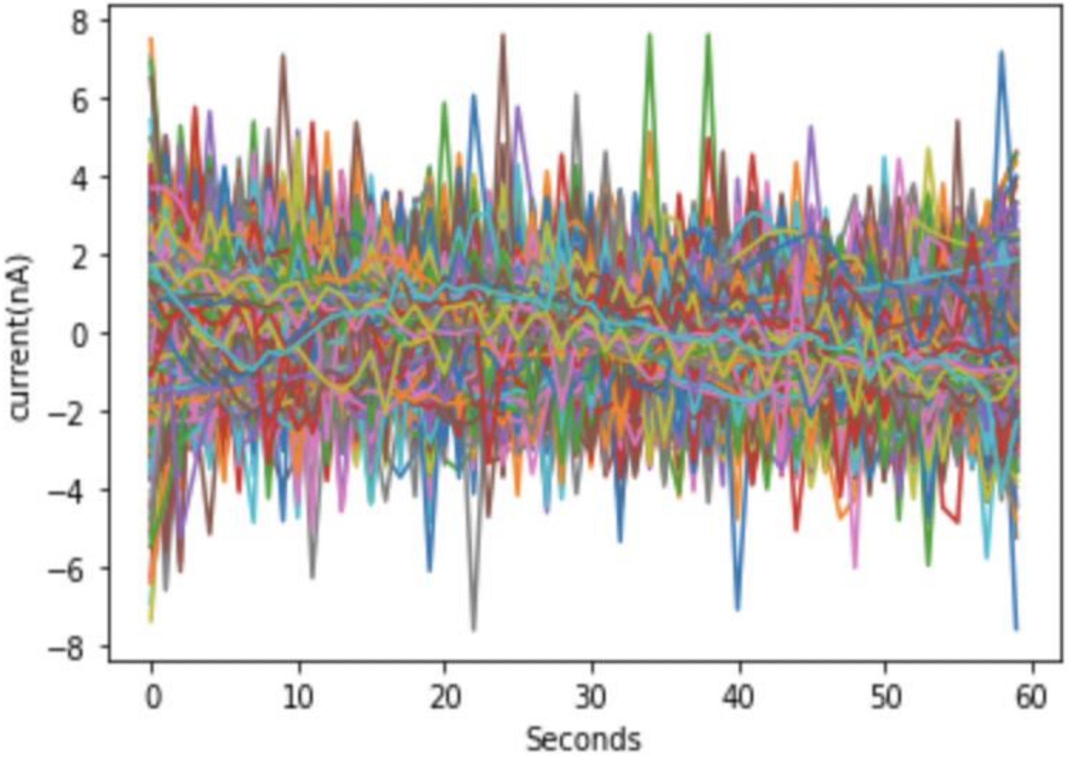

X_1
X_2
X_3
Node1: $(W_{1,1}:W_{1,2}:W_{1,3})$ Distance = $\sqrt{\sum (x_i - w_{1,i})^2}$ = 1.2
Node2: $(W_{2,1}:W_{2,2}:W_{2,3})$ Distance = $\sqrt{\sum (x_i - w_{2,i})^2}$ = 0.8
Node3: $(W_{3,1}:W_{3,2}:W_{3,3})$ Distance = $\sqrt{\sum (x_i - w_{3,i})^2}$ = 0.4
Node4: $(W_{4,1}:W_{4,2}:W_{4,3})$ Distance = $\sqrt{\sum (x_i - w_{4,i})^2}$ = 1.1
Node5: $(W_{5,1}:W_{5,2}:W_{5,3})$ Distance = $\sqrt{\sum (x_i - w_{5,i})^2}$ = 1.3
Node6: $(W_{6,1}:W_{6,2}:W_{6,3})$ Distance = $\sqrt{\sum (x_i - w_{6,i})^2}$ = 1.0
Node7: $(W_{7,1}:W_{7,2}:W_{7,3})$ Distance = $\sqrt{\sum (x_i - w_{7,i})^2}$ = 0.6
Node8: $(W_{8,1}:W_{8,2}:W_{8,3})$ Distance = $\sqrt{\sum (x_i - w_{8,i})^2}$ = 1.2
Node9: $(W_{9,1}:W_{9,2}:W_{9,3})$ Distance = $\sqrt{\sum (x_i - w_{9,i})^2}$ = 0.9

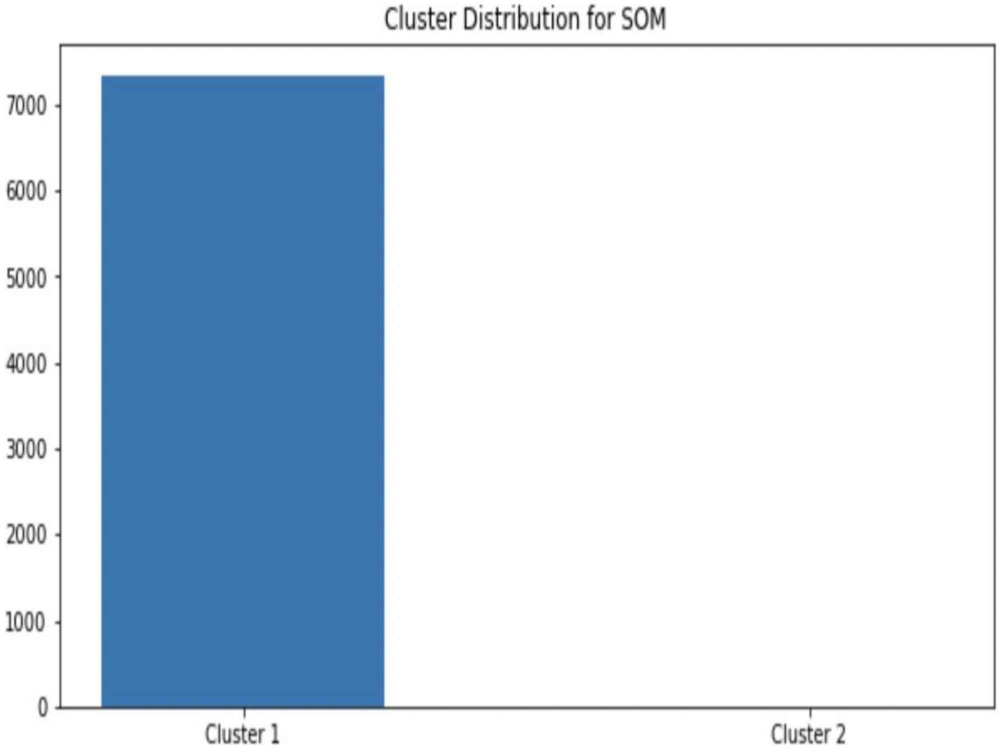
CLUSTERS IN SOM

Count in each cluster



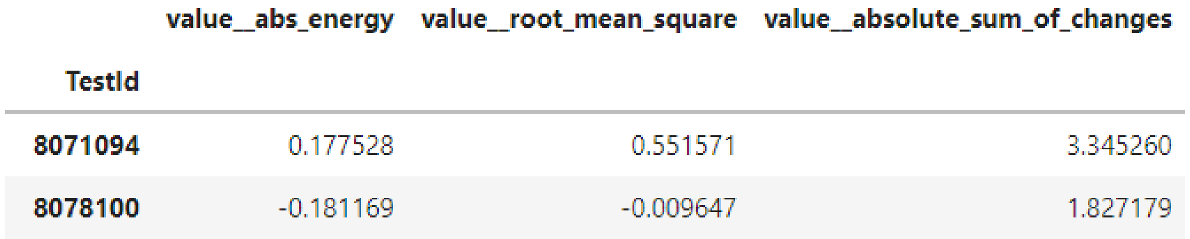
RESULTS OF STANDARDISING WINDOWS - SOM





MODELLING – TSFRESH PREDICTORS

- Phase 1: Feature Extraction (~ 450 features)
  - skewness(x)
  - sample_entropy(x)
  - autocorrelation(x, lag)

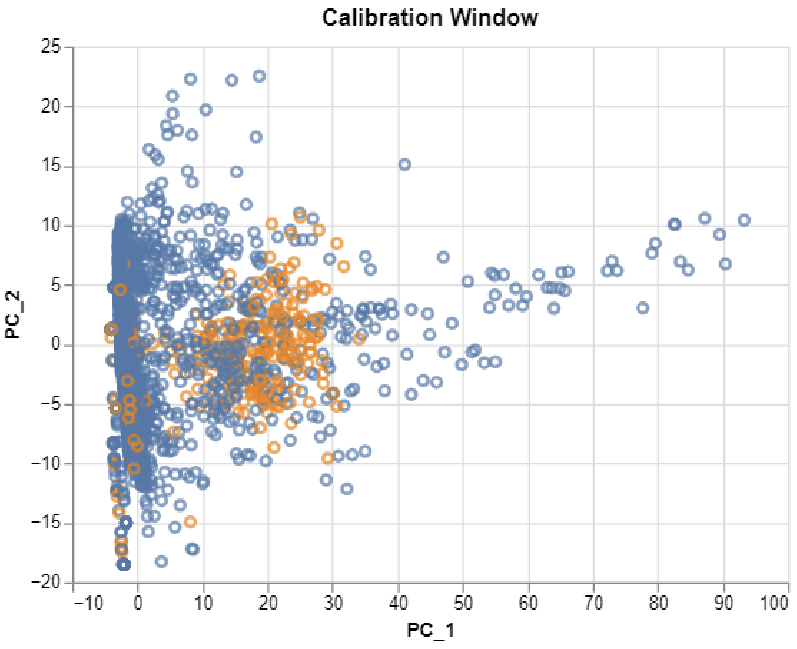


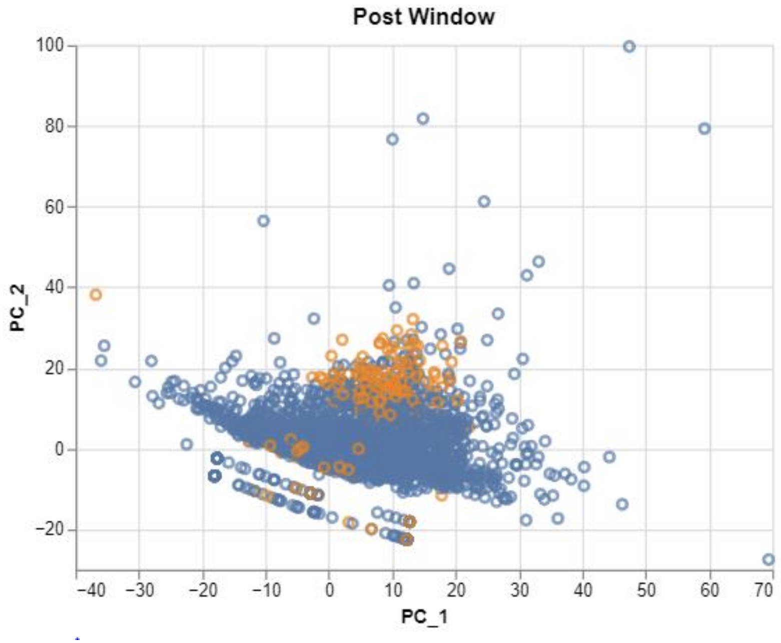
	value__abs_energy	value__root_mean_square	value__absolute_sum_of_changes
TestId			
8071094	0.177528	0.551571	3.345260
8078100	-0.181169	-0.009647	1.827179

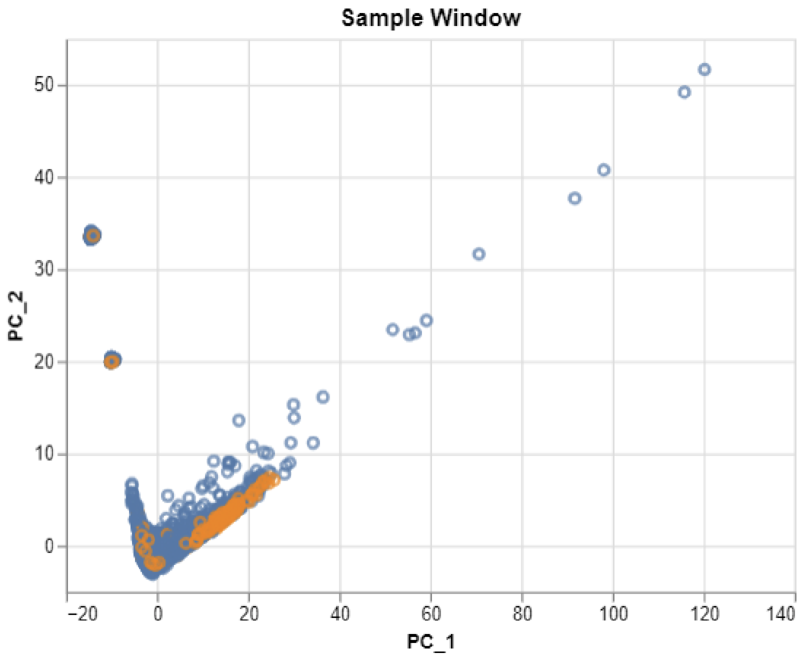
- Phase 2: Feature Significance Testing (~250 features)
  - Only the features that are significant with respect to classifying the readings are kept.

MODELLING – DIMENSION REDUCTION

- Phase 3: PCA for dimension reduction (~ 30 components)
  - 95 % accumulated amount of variance explained


Calibration Window

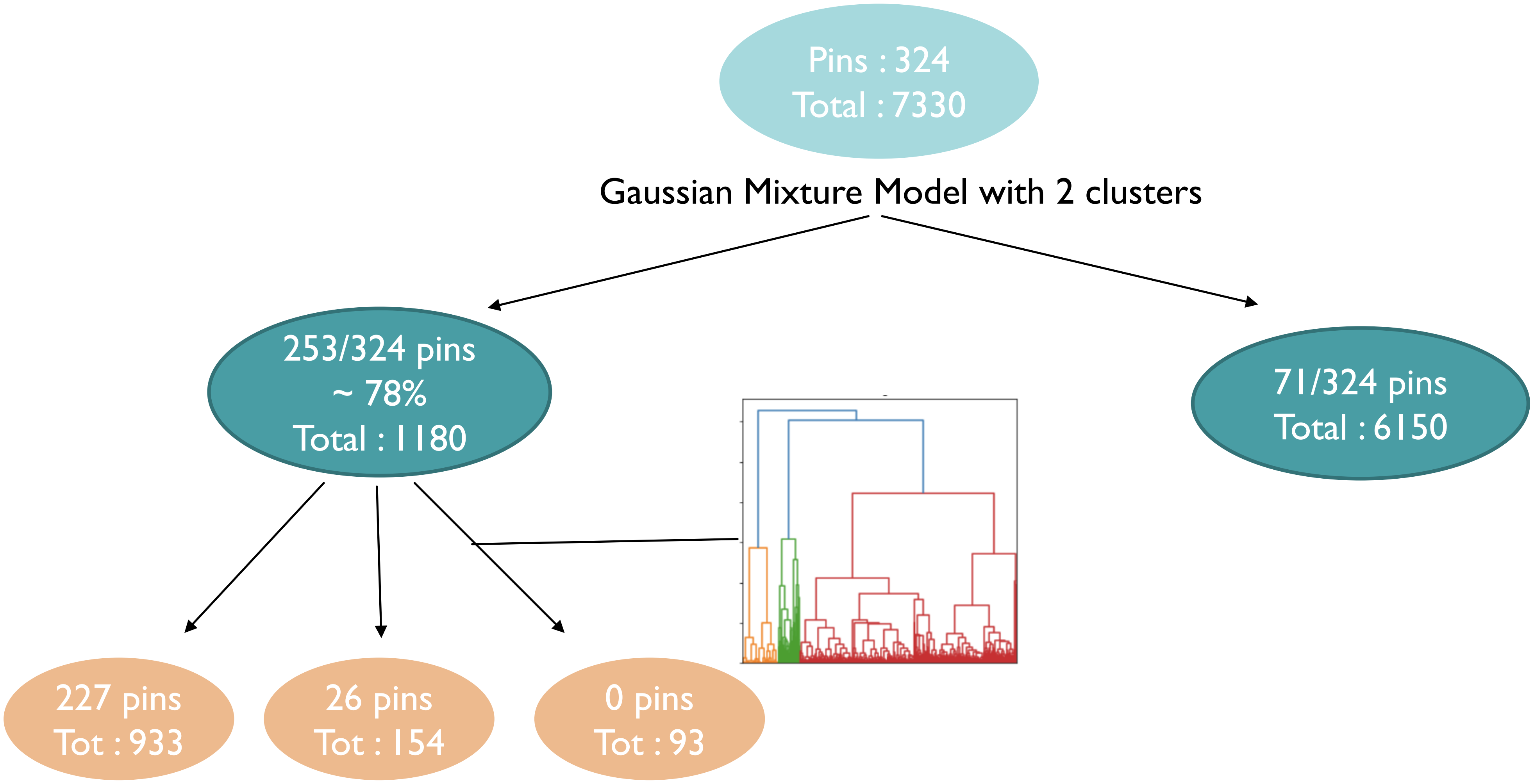

Post Window


Sample Window

MODELLING – CLUSTERING

- Phase 4: Clustering the components
- Apply to all windows separately
  - Create clusters using various algorithms (Gaussian Mixture Model/Agglomerative Clustering)
    - Try to get a cluster with most of the pins and a small amount of total readings
  - Cluster the subcluster that contains most of the pins

MODELLING – CLUSTERING



Pins : 324
Total : 7330

Gaussian Mixture Model with 2 clusters

253/324 pins
~ 78%
Total : 1180

71/324 pins
Total : 6150

227 pins
Tot : 933

26 pins
Tot : 154

0 pins
Tot : 93

PRELIMINARY/INCREMENTAL RESULTS

- Completed the first round of preprocessing (time series standardization and windowing)
- Tried a few clustering methods on both the whole signal in the windows and derived predictors.
  - Deliverables: notebooks that we are cleaning up/documenting as we go to give the client.

ROADBLOCKS

- Need to find a standard way to describe the readings contained in clusters that we are getting.
- So far not getting a cluster with pure pin contacts
  - There are different ways things go wrong – need to look at more clusters.
- Standardizing might be leading to a loss of information
  - Pin contacts usually go to 0 apparently

PLANNING AND ACTIONS FOR THE NEXT CYCLE

Sara

- Look into different ways of windowing and standardizing.
- Attempt whole time-series clustering/visualization with new windows.
- Dig into Fourier and other transforms more.
- Record meeting minutes for Siemens check-in on Tuesday.

Saisree

- Try other clustering algorithms on shape-based TS(DTW-SOM applied to different windows)
- Attempt feature extraction using 1-D CNN if the shape-based approach is not promising
- Record meeting minutes for Siemens check-in on Friday.

Neethu

- Using auto-encoder to extract features.
- Use the extracted features for clustering.
- Use LSTM for anomaly detection
- Record meeting minutes for Advisory committee on Tuesday.

Justine

- Do more research on anomaly detection in timeseries
- Record meeting minutes for check-in meeting with Capstone advisors.

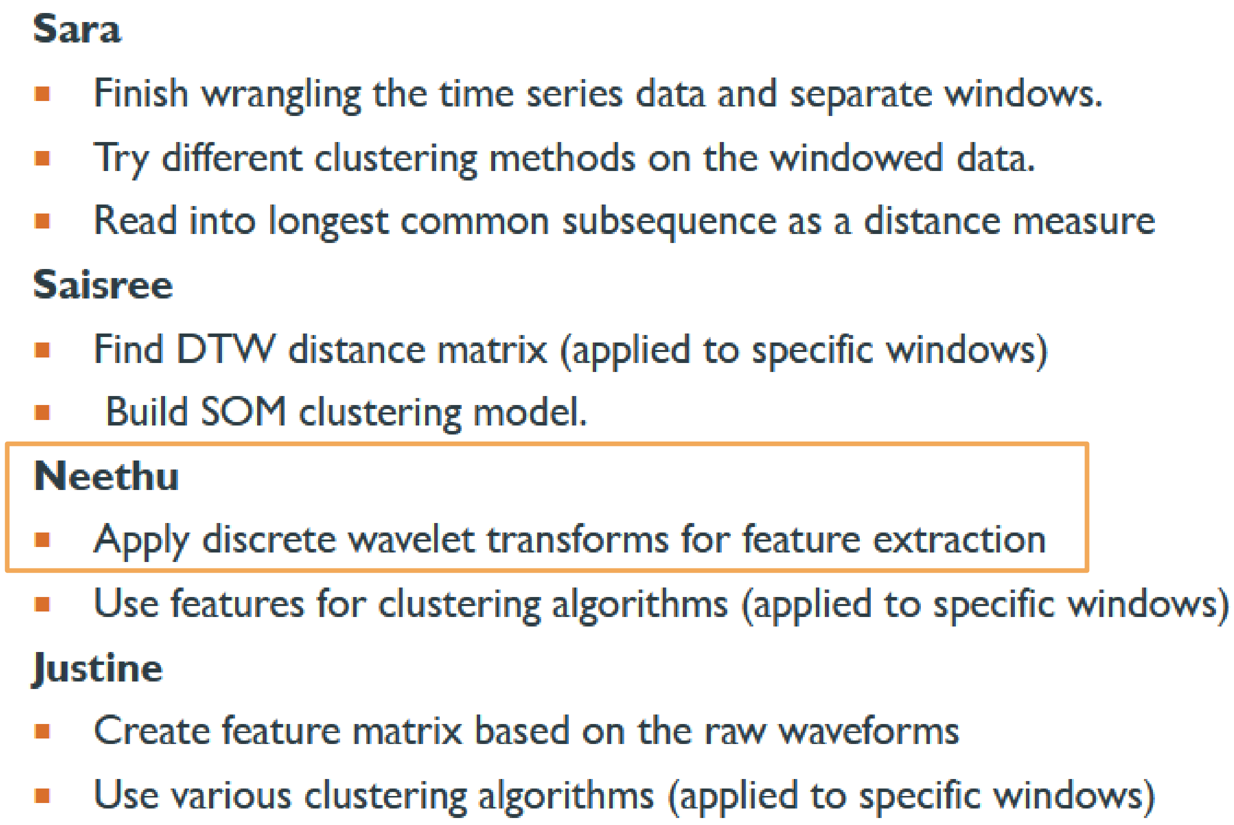
Team

- Create slides for and present work in meetings (split equally).
- More discussion on tasks today, points here are flexible after morning's meeting with our client.

DEVIATION FROM THE ORIGINAL PLAN/SCHEDULE

ACCOMPLISHED ALL LAST WEEK'S TASKS?

- Last week's tasks:



Sara

- Finish wrangling the time series data and separate windows.
- Try different clustering methods on the windowed data.
- Read into longest common subsequence as a distance measure

Saisree

- Find DTW distance matrix (applied to specific windows)
- Build SOM clustering model.

Neethu

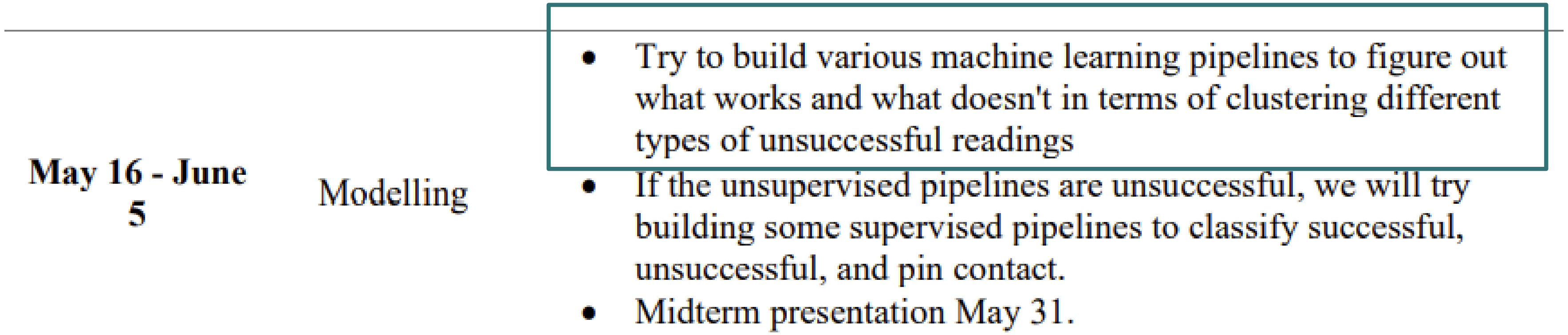
- Apply discrete wavelet transforms for feature extraction
- Use features for clustering algorithms (applied to specific windows)

Justine

- Create feature matrix based on the raw waveforms
- Use various clustering algorithms (applied to specific windows)

DEVIATION FROM THE ORIGINAL PLAN/SCHEDULE

AHEAD/BEHIND/ON TRACK?



May 16 - June 5	Modelling	• Try to build various machine learning pipelines to figure out what works and what doesn't in terms of clustering different types of unsuccessful readings • If the unsupervised pipelines are unsuccessful, we will try building some supervised pipelines to classify successful, unsuccessful, and pin contact. • Midterm presentation May 31.

SUMMARY OF INTERACTIONS WITH THE CLIENT

- Exchanged a few emails throughout the week
  - Our data was updated again.
- Meeting on Monday, May 16th
  - Discussed our plan for the week.
- Meeting on Friday, May 20th
  - Presented them with the progress we made over the week and got feedback
  - Learned more about what is looked for when diagnosing a pin contact

SUMMARY OF INDIVIDUAL AND TEAM EFFORTS MAY 16 - 22

- Sara :
  - Data wrangling: 10 hours
  - Clustering Attempts: 22 hours
  - Administrative work: 5 hours
  - **Total : ind. + team = 41.5**

- Neethu :
  - Feature extraction and modelling : 24 hrs
  - Researching : 7.5 hrs
  - Others : (slides, windowing, standardize) : 4.5
  - **Total : ind. + team = 40.5**

Saisree :

- Researching : 10 hrs
- Clustering : 17 hrs
- Outlier detection/wrangling : 5 hrs
- Administrative work: 5 hrs
- **Total : ind. + team : 37 hrs**

- Justine:
  - Feature extraction and clustering : 27 hours
  - Data processing (scaling/windowing) : 3.5 hours
  - Others (writing minutes, slides for presentations, meetings etc.) : 8 hours
  - **Total : ind. + team = 38.5**
- Team:
  - **Time spent in meetings : 4.5 hours**

FEEDBACK?