PROCESS CONTROL AND ANOMALY DETECTION WITH CLASSIFICATION OF OPTICAL EMISSION SPECTRA

AUSTIN BOCK

SPRINGBOARD CAPSTONE 2 – MILESTONE

OCT, 2017

PROJECT OVERVIEW & GOALS

- **Overview:**
  - Customer has come to me with a dataset that includes optical emission spectra from monitors on a plasma etch tool and the measured Error data from subsequent production runs. (detailed descriptions provided in "Data Dictionary.txt" on Github)
  - Customer requests exploratory analysis for outlier classification which could be used for process control. 2 methods of using this data will be explored as Key Goals.
  - The process uses statistical process control in a manufacturing environment to control 'Error' on production
  - The key starting assumption from the customer is that other noise/error sources are not important and this assumption may not be true in that other sources could be present.

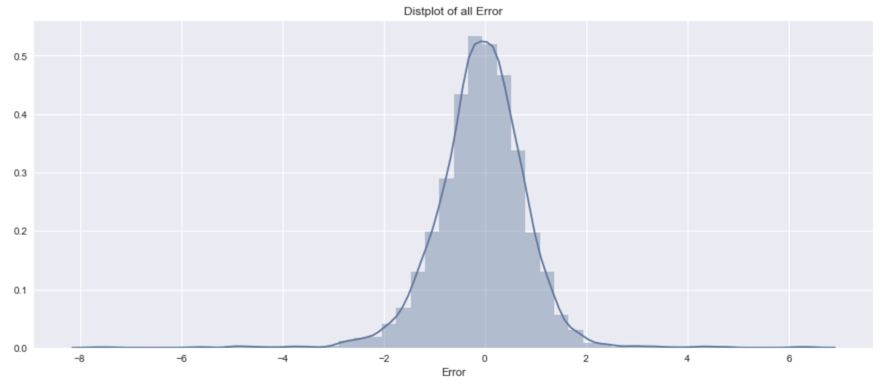
- **Key Goals:**
  - Can classifier be used as UP/DOWN monitor that captures the likelihood of failing with Error outside statistical limits of the population on production
  - Can classifier be used as monitor the process condition for any shifts that will require investigation.
  - Advise on next steps to improve modeling capability with other data sources

DATA CLEANING

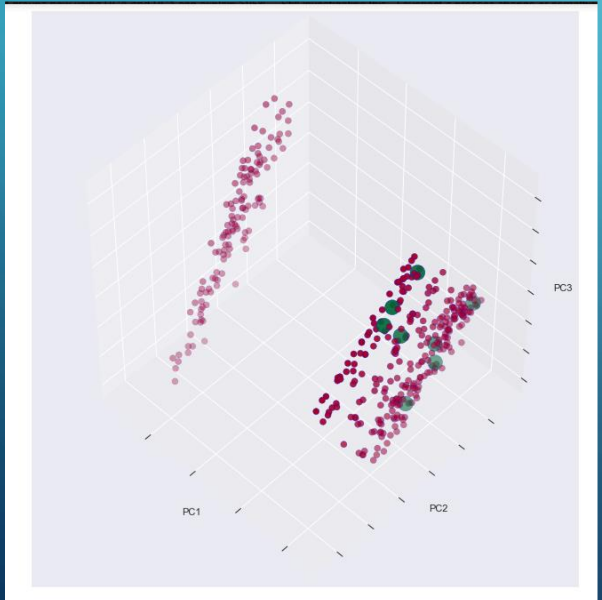
- **Cleaning:**
  - Customer supplied dataset contains a large anonymized dataset of time series sequence through identical toolsets.
  - Timestamps are removed and converted to sequential runs, thus protecting IP metrics.
  - Error columns were normalized to zero
  - Spectra are also included as separate files. These were married to the subsequent ‘Error’ datapt by the following:
    - Renaming all files with windows creation stamp to allow perfect sort by run order and converting from binary to csv.
    - Algorithm to search all IDs with ‘Error’ point and then appending the run ID identifier to the end of the monitor that has optical emission data. Ex. 2017-01-06-08-32-45_monitorID_ErrorID.csv
    - The monitorID spectra file can now be matched to ‘Error’ data pt for ErrorID
    - The labels are then created for each run based on ‘Error’ according to population limits of

  - See “Data Dictionary” on Github for more explicit summary

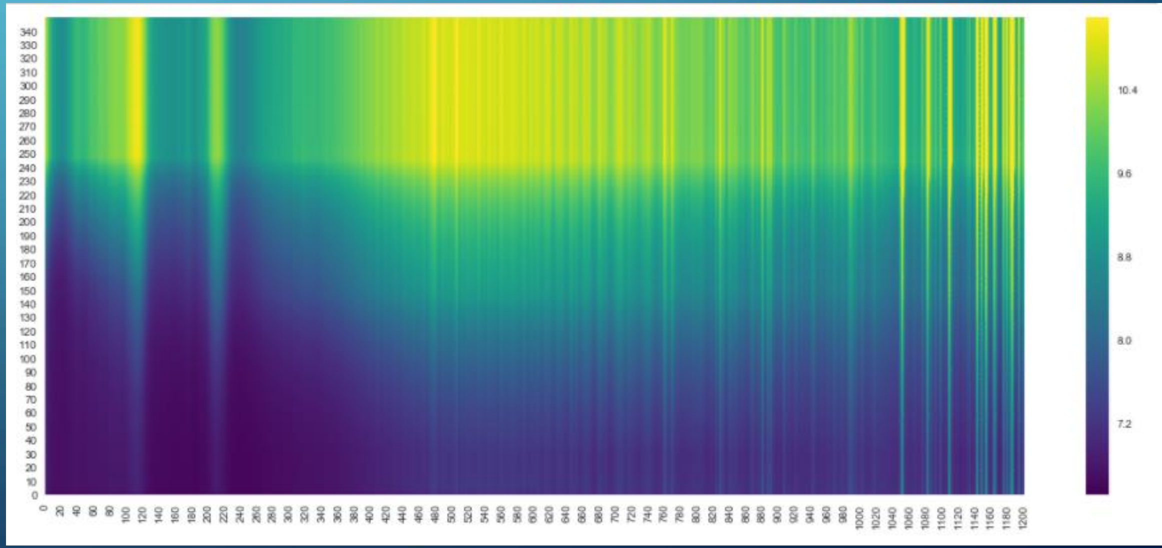
EXPLORATION



PCA successfully found a tool shift – this was confirmed with process expert who found hardware fix was root cuase for process shift below

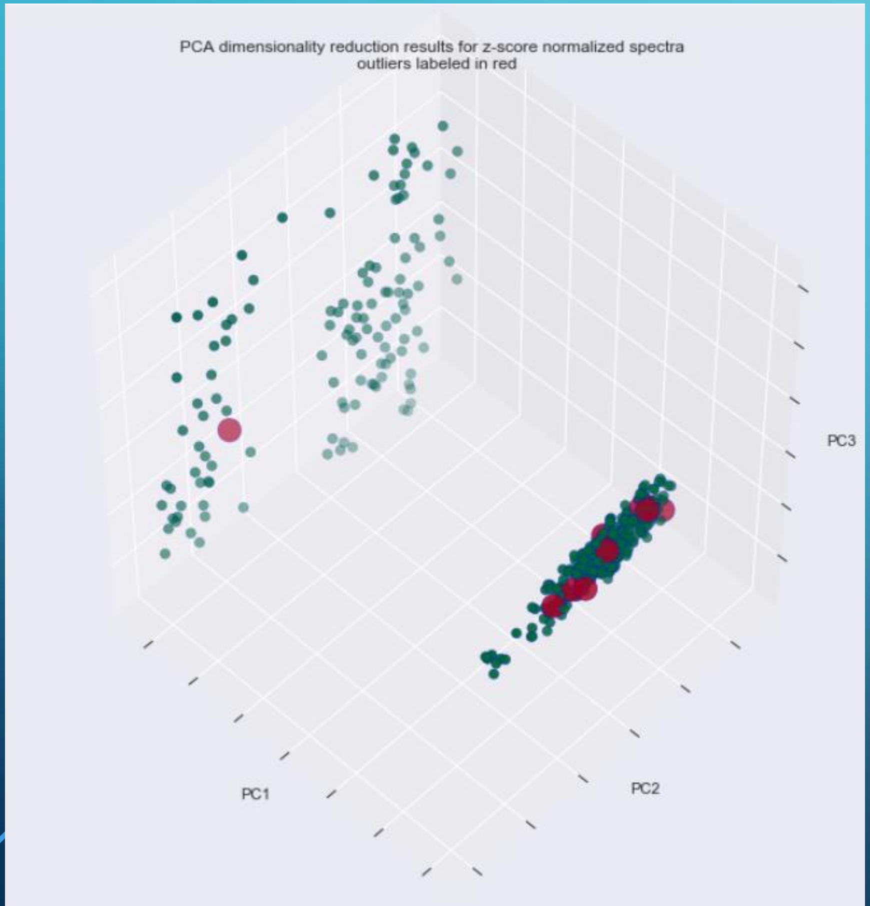


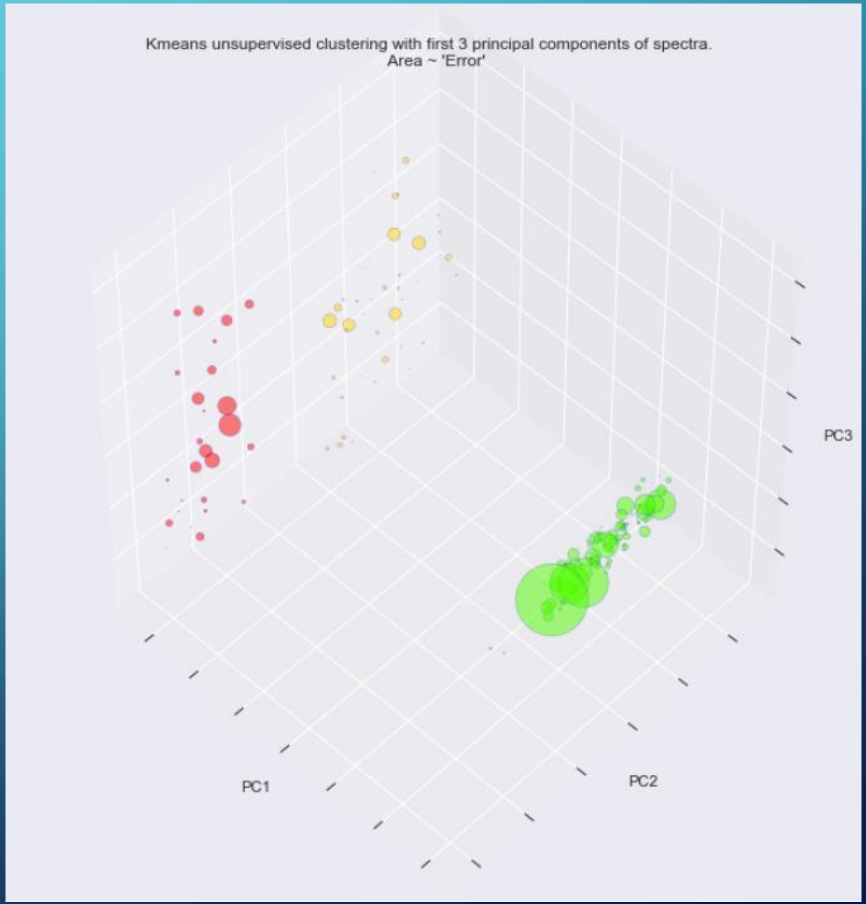
Log scale emission



PCA

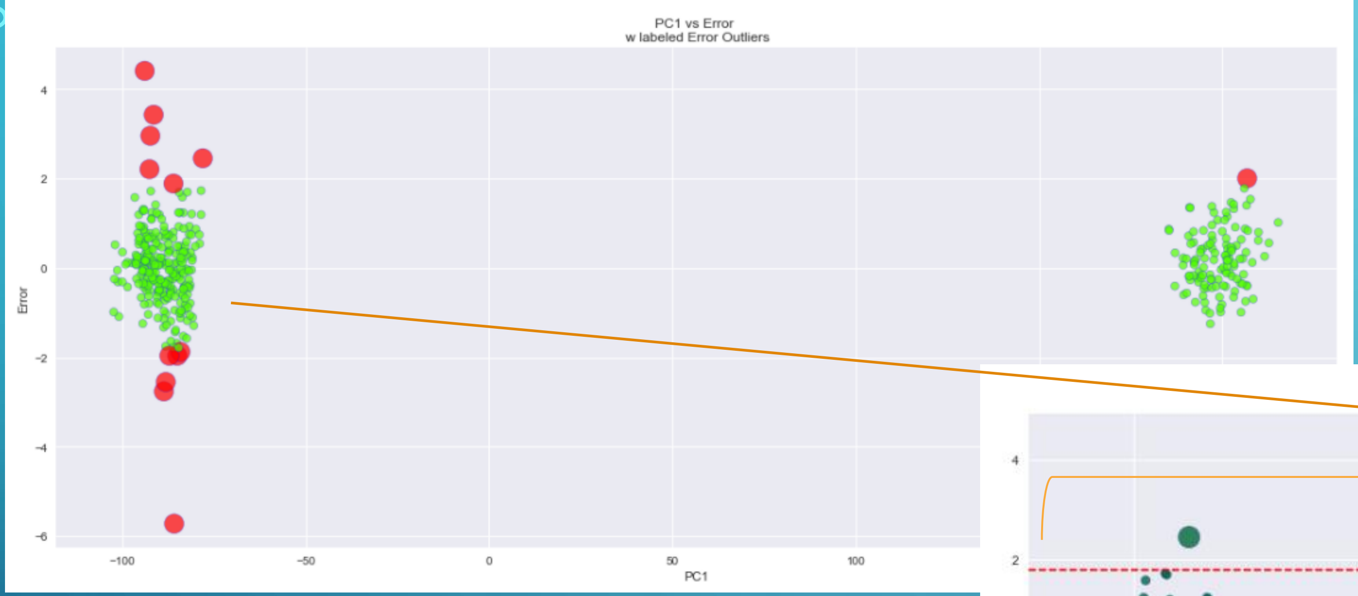
- **Kmeans clusters (n=3) for reduced spectra:**
  - Clear delineation of cluster groups is possible for large sample set
    - What drives the clusters from a hardware level?

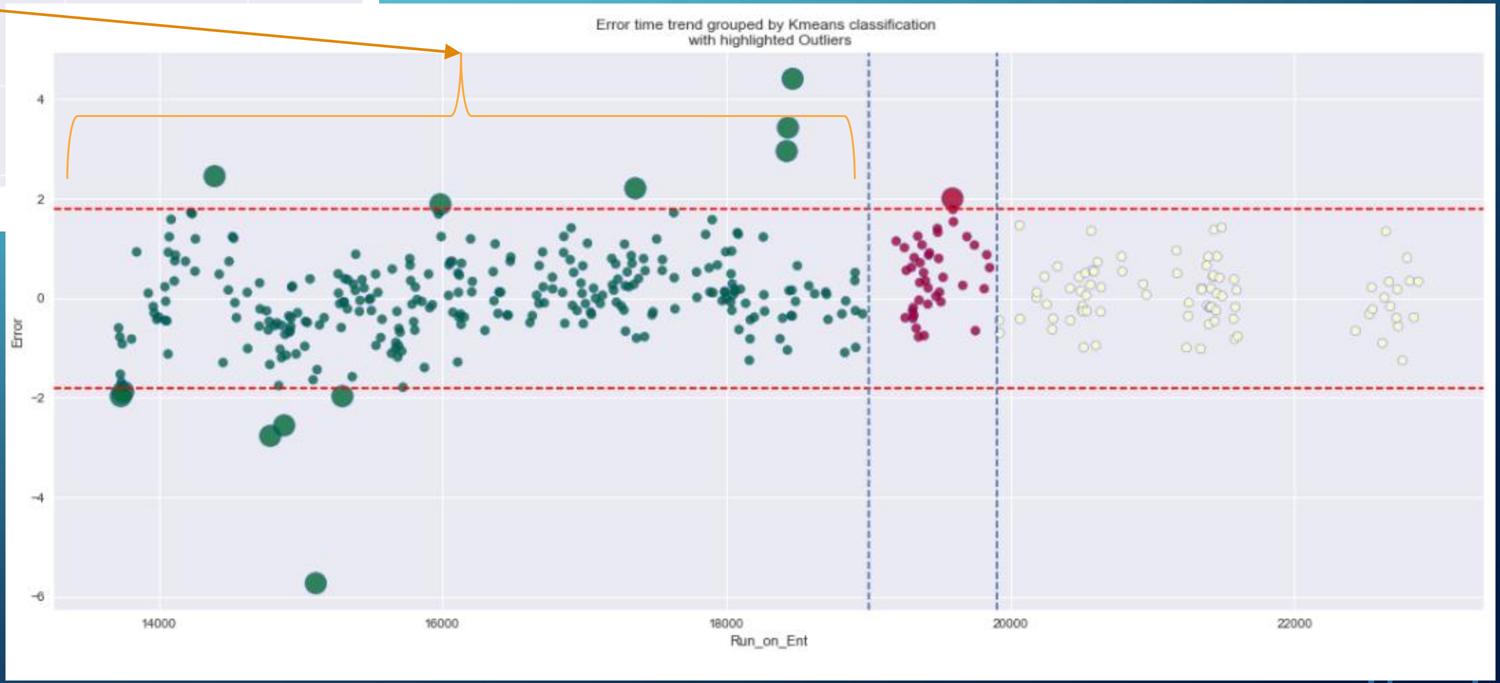




INVESTIGATION

- Clusters are clearly correlated to confirmed hardware work (per process expert, marked as vertical blue lines).
- Focus to within hardware cycle analysis (dark green for anomaly detection)



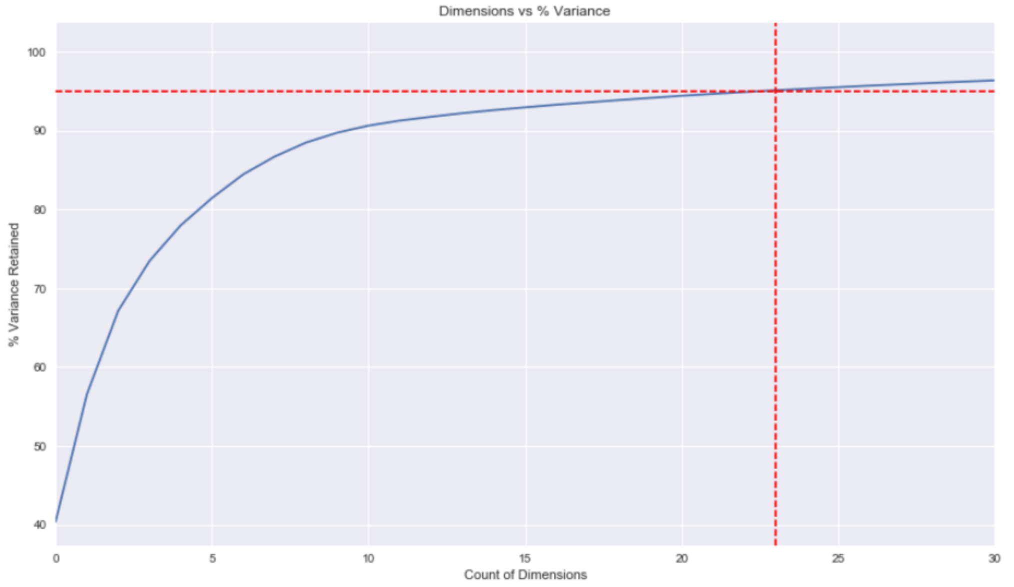


ANOMALY DETECTION MODELING

- Within single cluster, Train/Test split on labeled benign samples
- PCA fit/transform dimension reduction (n = 24 dimensions) benign train
- PCA transform benign test, malign test
- Final malign outlier Test using benign model fit
- 2 outlier detection models (Isolation Forest vs. 1-class SVM)

```
The minimum number of dimensions to retain 95% variance =  24
[ 40.38  56.53  67.13  73.45  77.98  81.45  84.45  86.7   88.47  89.73
  90.62  91.26  91.74  92.2   92.59  92.94  93.27  93.57  93.86  94.14
  94.41  94.65  94.88  95.1   95.3   95.5   95.69  95.87  96.04  96.2 ]

<matplotlib.text.Text at 0x22e4d961b38>
```



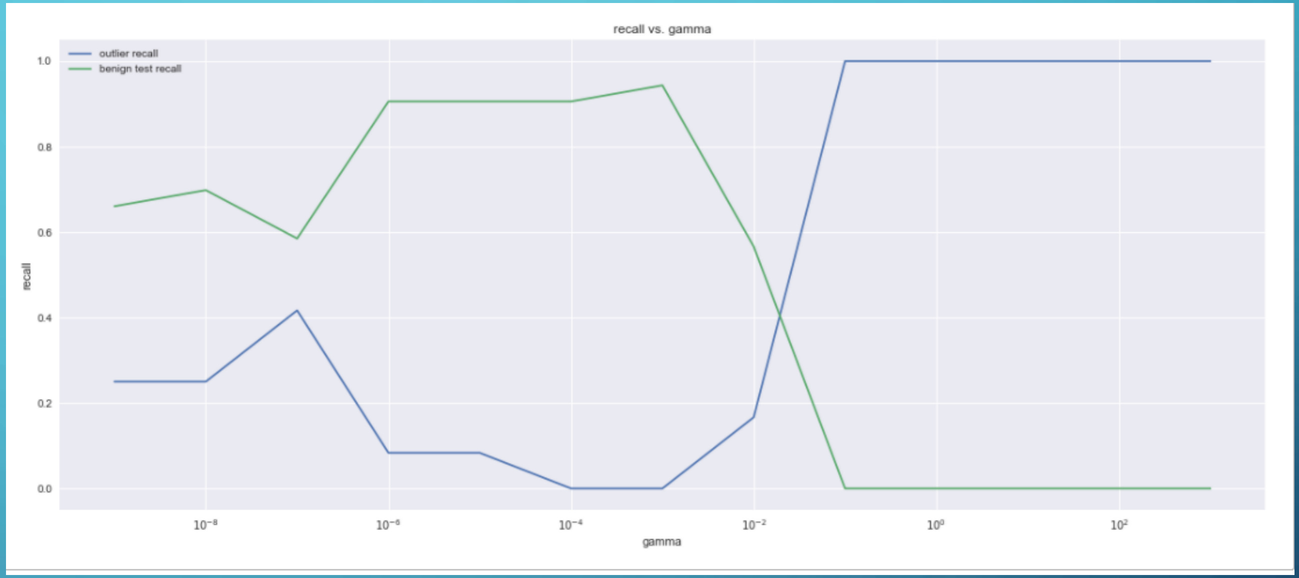
RESULTS

- Isolation Forest best result:

```
y_pred_train recall: 0.8962
y_pred_test recall: 0.9245
y_pred_outliers recall 0.1667
```

  - Poor outlier recall
  - Cost of FN (TypeII error) of outliers outweighs some small allowable false positive rate (TypeI error)

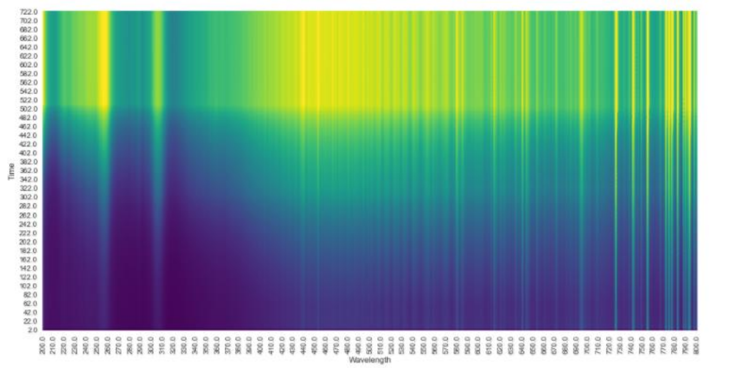
- 1 class SVM:
  - Poor recall ratio for benign/malign



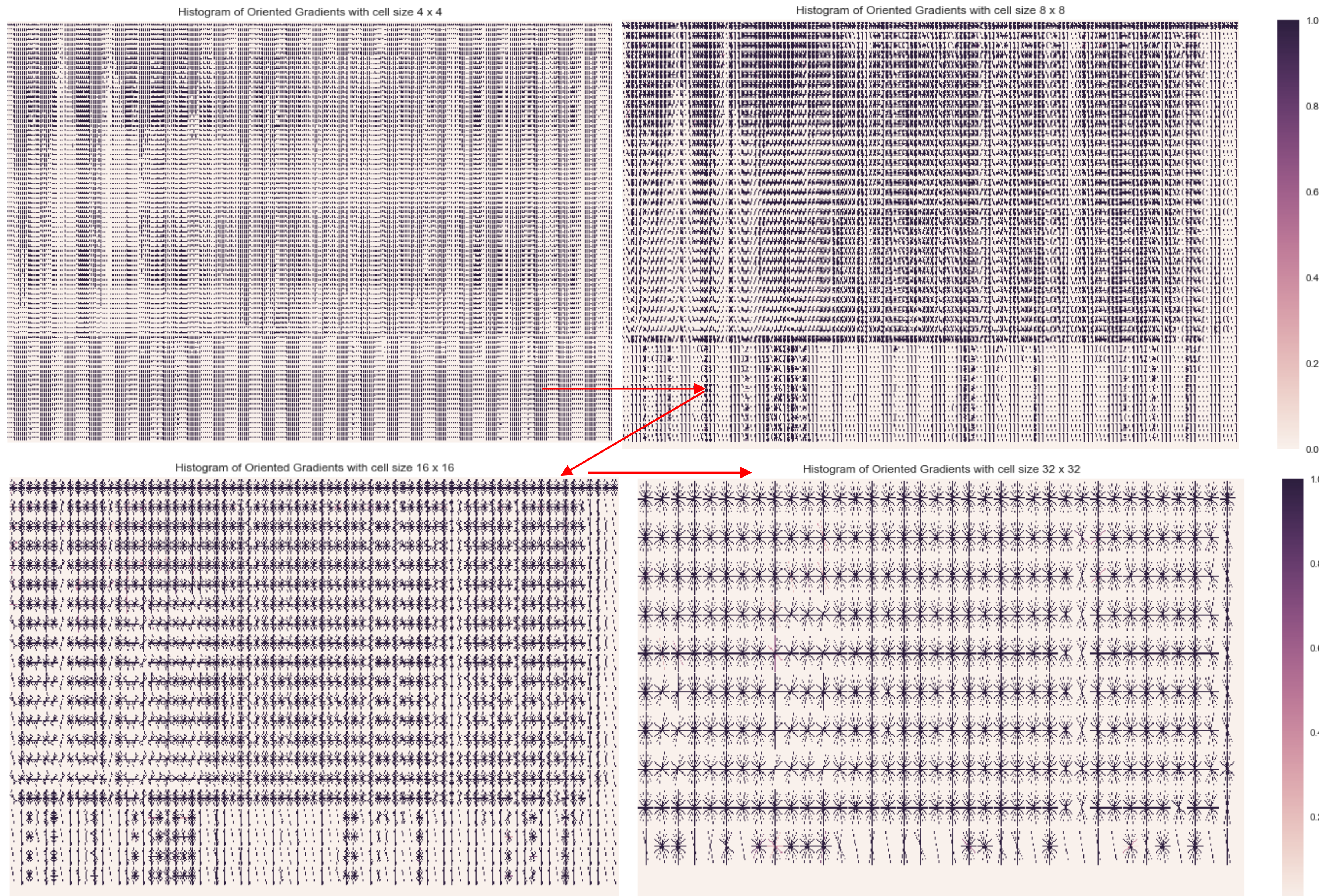
- Neither anomaly classifier is predicting outliers well

TREAT RAW SPECTRA AS IMAGES

- Histogram of oriented gradients: (optimization of cells)

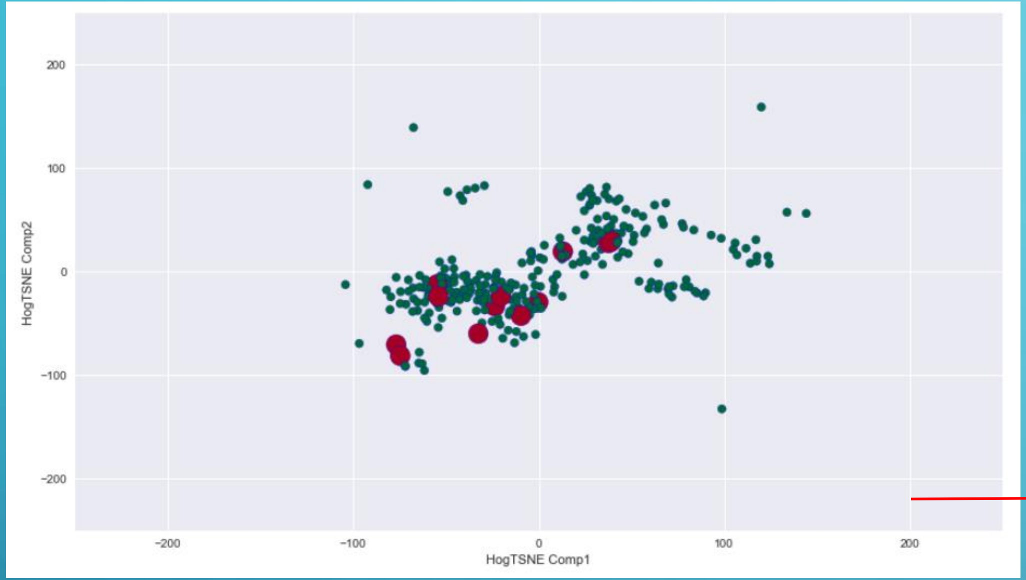


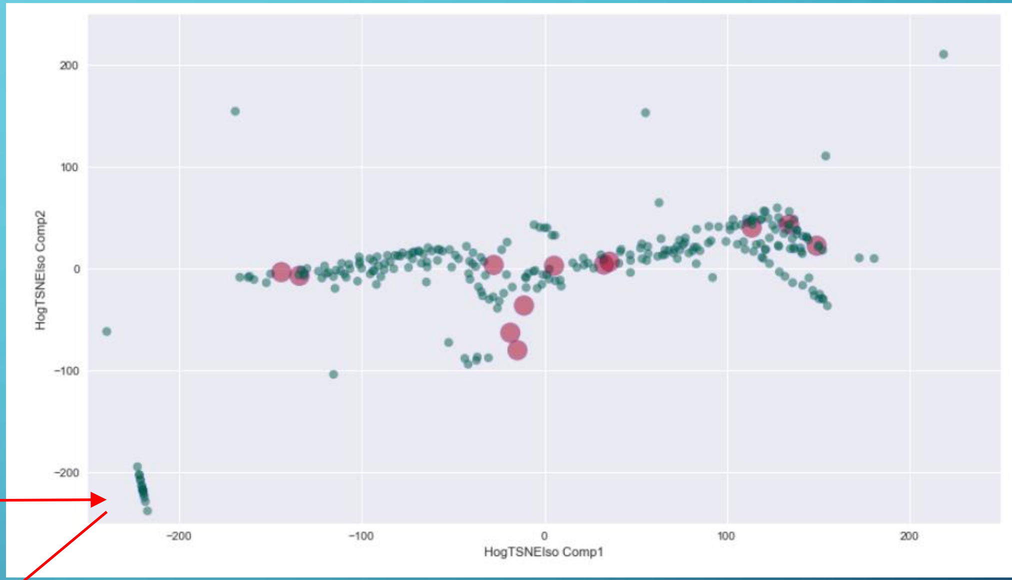
Counts occurences of gradient orientation in localized cells of image.

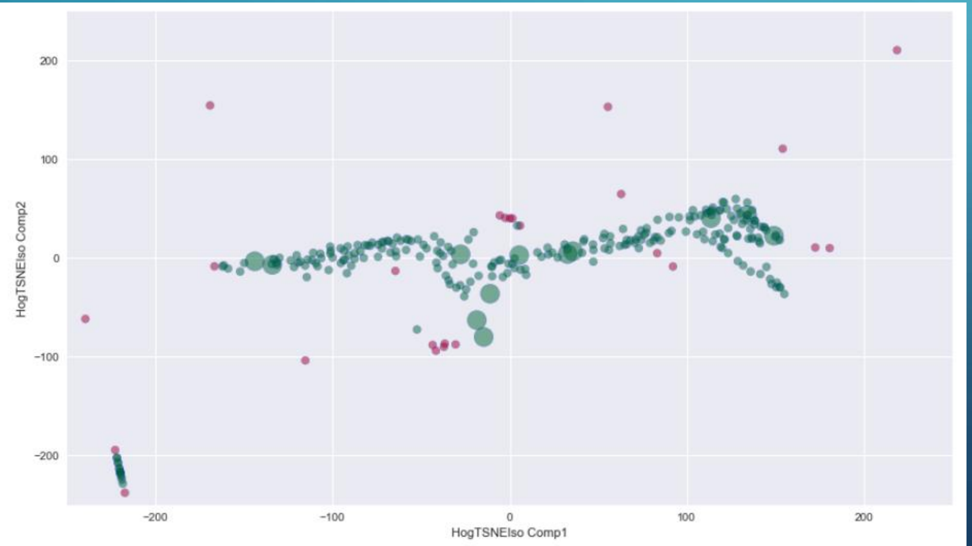


HOG FEATURE VECTOR->TSNE->ISO

- t-distributed Stochastic Neighbor Embedding
- **Isomap manifold** Non-linear dimensionality reduction through Isometric Mapping



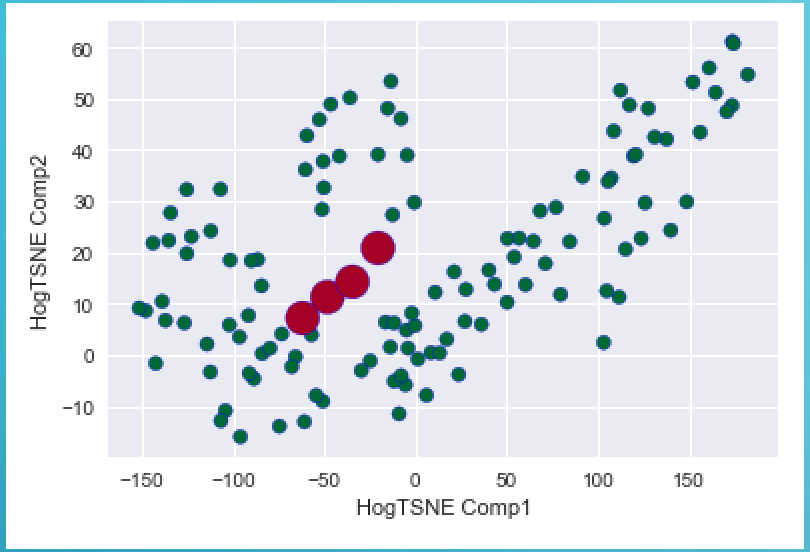


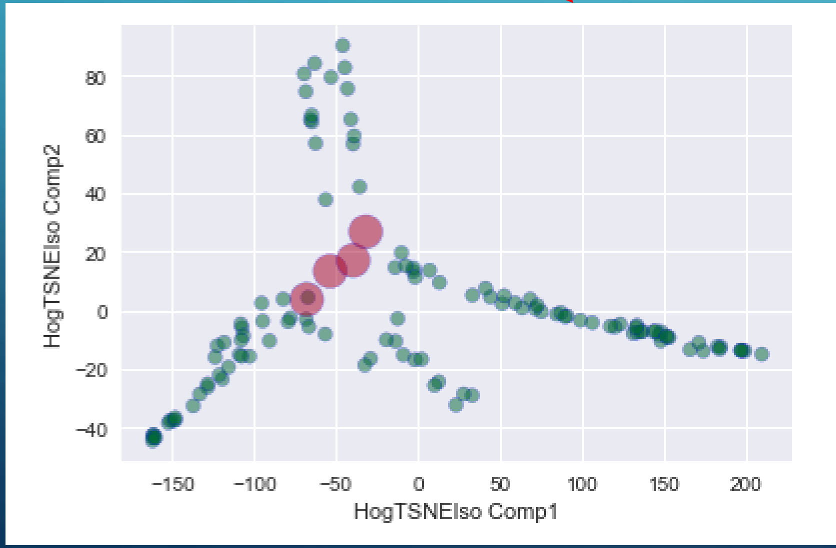


Final clustering doesn't follow prescribed outliers per expert model. Clustering does signify clear outliers for investigation:
Local Outlier Factor classification for outliers

HOG FEATURE VECTOR->TSNE->ISO

- Repeat outlier detection pipeline on ‘clean’ dataset: no hardware work recorded, shorter timeframe





Final clustering seems to enable model prediction of outliers

SUMMARY OF FINDINGS

- Goals and assumption:
    - Customer requests exploratory analysis for outlier classification which could be used for process control.
    - Key hypothesis that monitor spectra can predict critical error
    - The key starting assumption from the customer is that other noise/error sources are not important and this assumption may not be true in that other sources could be present.
- Findings and recommendations
    - Verified hardware state changes from maintenance are dominating classification schemes.
        - Recommend investigation to use as quality control methodology for hardware maintenance.
    - Spectral images can detect outliers not correlated to customer defined labels.
        - Recommend investigation into new outliers is warranted for root cause analysis.
    - Controlled hardware timeframe is possibly capable to predict Error outliers, but not repeatable on each cycle.
        - Recommendation to extend dataset to include hardware RF data with spectral data
        - Hypothesis that hardware RF system is driving some portion of Error outliers