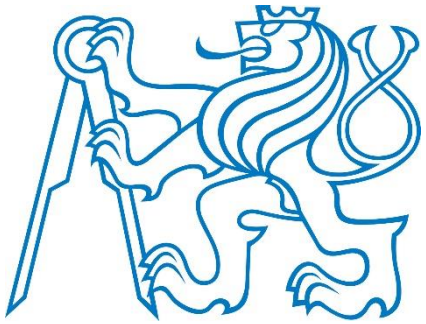
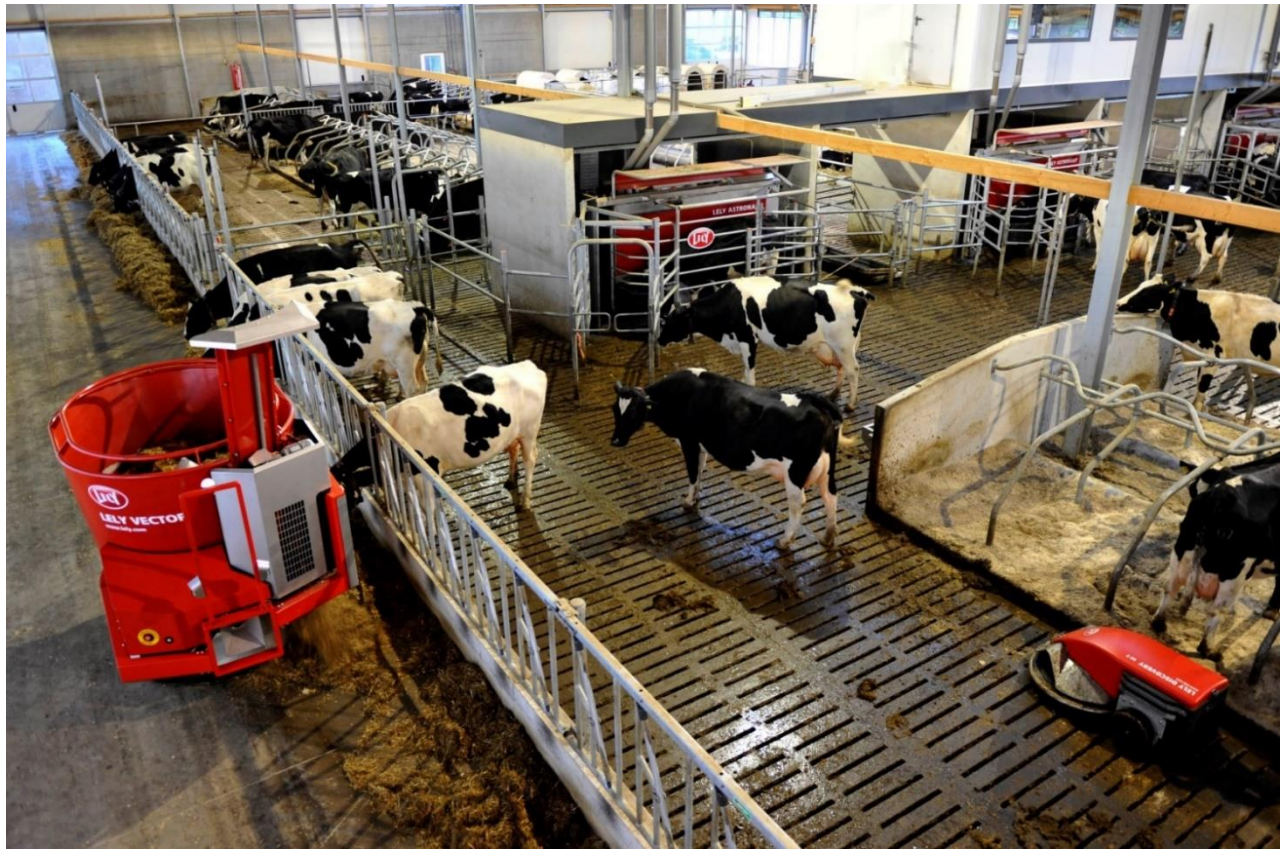
Machine learning on automated farms

Tomáš Borovička, David Veselý



Barn

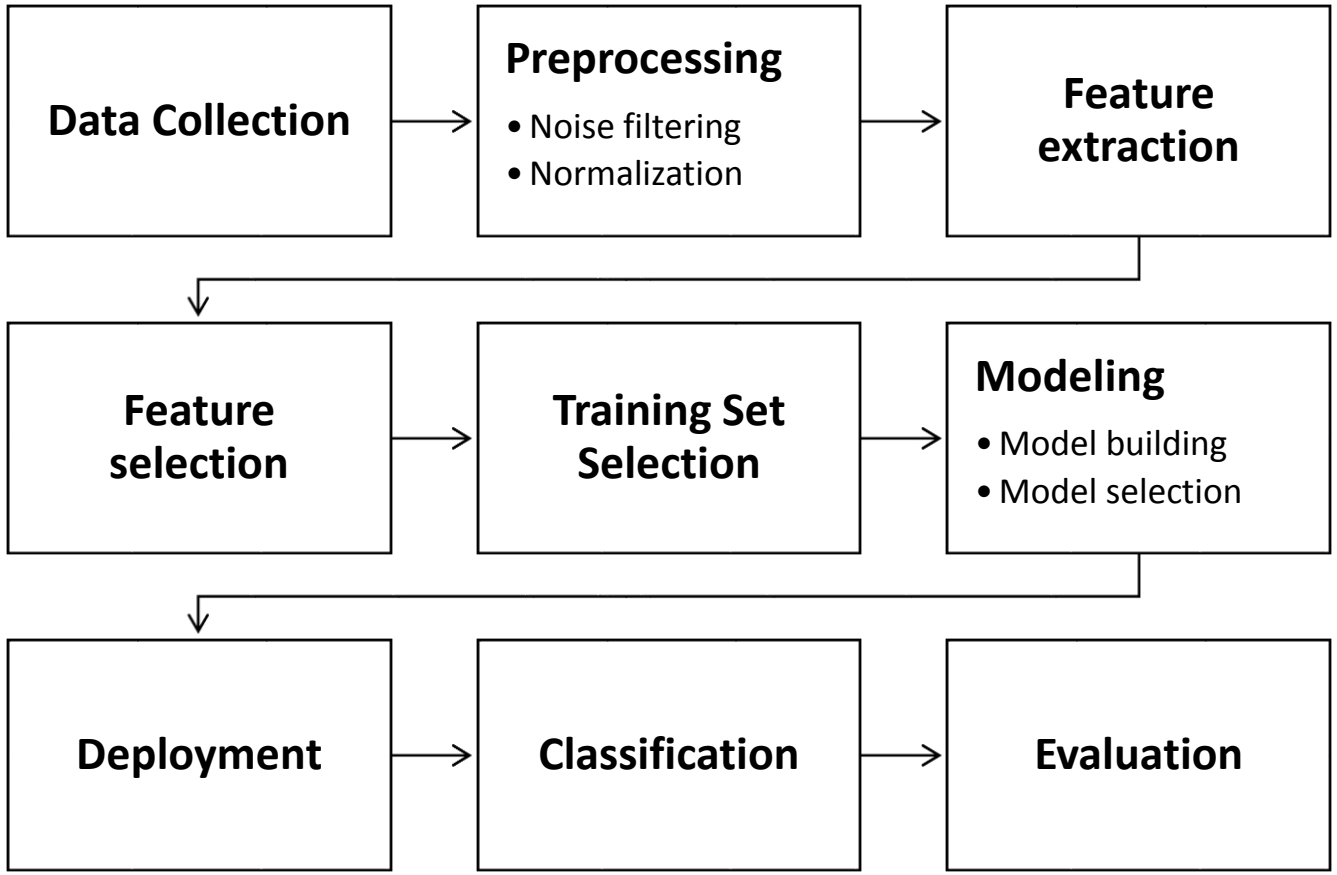


Disease Detection

Project Goal

- The goal was to detect sick cows and recognize particular diseases (mastitis, ketosis, lameness, milk fever…) using data available from robots.

Data Mining Process



Data Set

- 8 farms, ~2000 cows, ~2.5 million records.

Preprocessing

- Out of range values filtering (correct when possible).
- Replace missing values.
- Normalization
  - Z-score

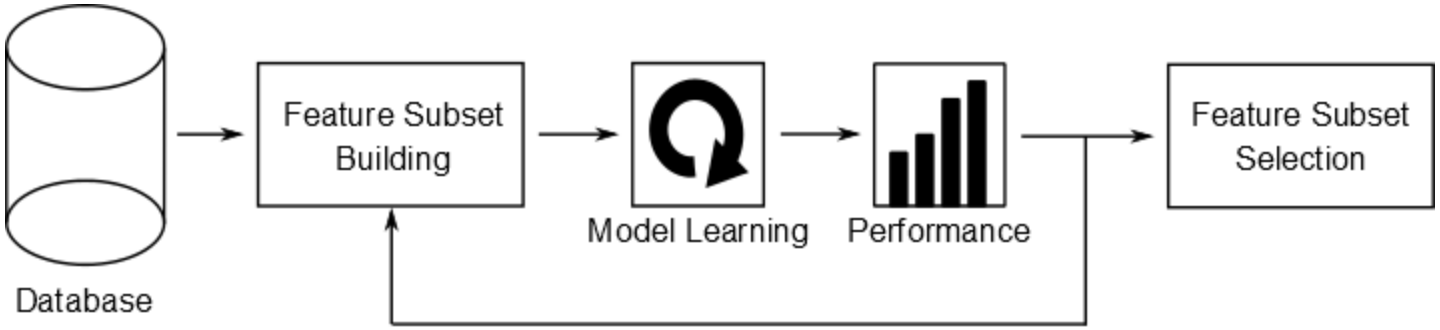
$$z_\tau^c = \frac{\mu_\tau^c - x^c}{\sigma_\tau^c}$$

where x denotes row value, μ and σ denotes average respectively standard deviation computed for a particular cow c in the time window of size τ.

Feature Extraction and Selection

- 46 row attributes.
  - Domain knowledge aggregation / feature extraction.
  - Time windows.
- **Total ~1176 features.**

- Bidirectional search (forward / backward selection).
- Genetic algorithm feature set selection.



Problem of Labeled Data

- Labeling is done by farmers.
- Farmers register diseases in farm management system.
- A lot of diseases are not registered.
- Many cases when disease is registered late or just wrongly.
- From four cows with exactly same behavior is just one registered as sick.

- **Data labeled by farmers are not reliable!**

Supervised or Unsupervised?

- Supervised
  - Not enough good quality data to build really good and reliable model.
- Unsupervised
  - That is more or less just the anomaly detection.

- We don’t have representative set of sick cows.
- Rules have not support.

- **Semi-supervised**

Semi-supervised Learning

Let X_l be a set of examples for which we know the label y_i and let X_u be far bigger set of examples without known label. Semi-supervised learning attempts to utilize unlabeled data in order to yield greater performance than standard supervised method if only labeled data are used.

Shahshahani, Behzad M., and David A. Landgrebe. "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon." Geoscience and Remote Sensing, IEEE Transactions on 32.5 (1994): 1087-1095.

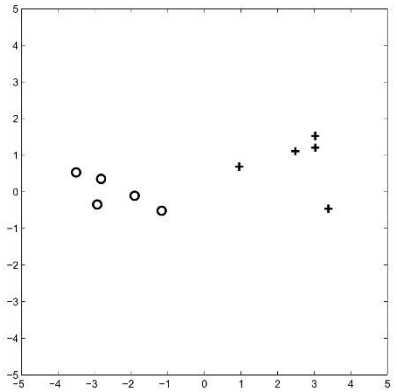
Semi-supervised Learning

If a parametric model can be decomposed as

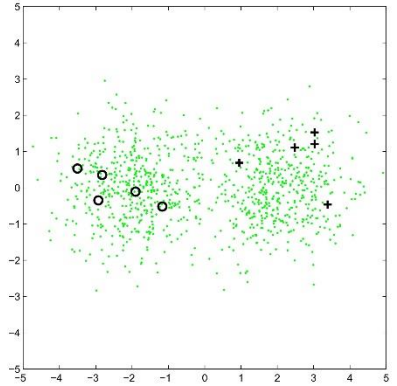
$P(x, y|\theta) = P(y|x, \theta)P(x|\theta)$,

the use of unlabeled examples can help to reach a better estimate of the model parameters.

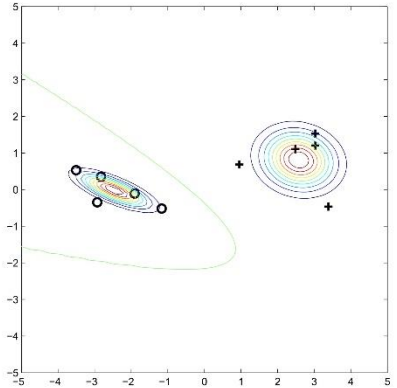
(Zhang and Oles, 2000)



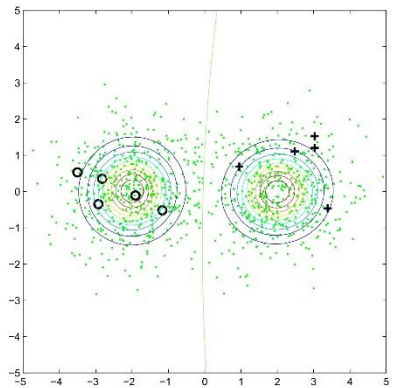
(a) labeled data



(b) labeled and unlabeled data (small dots)



(c) model learned from labeled data



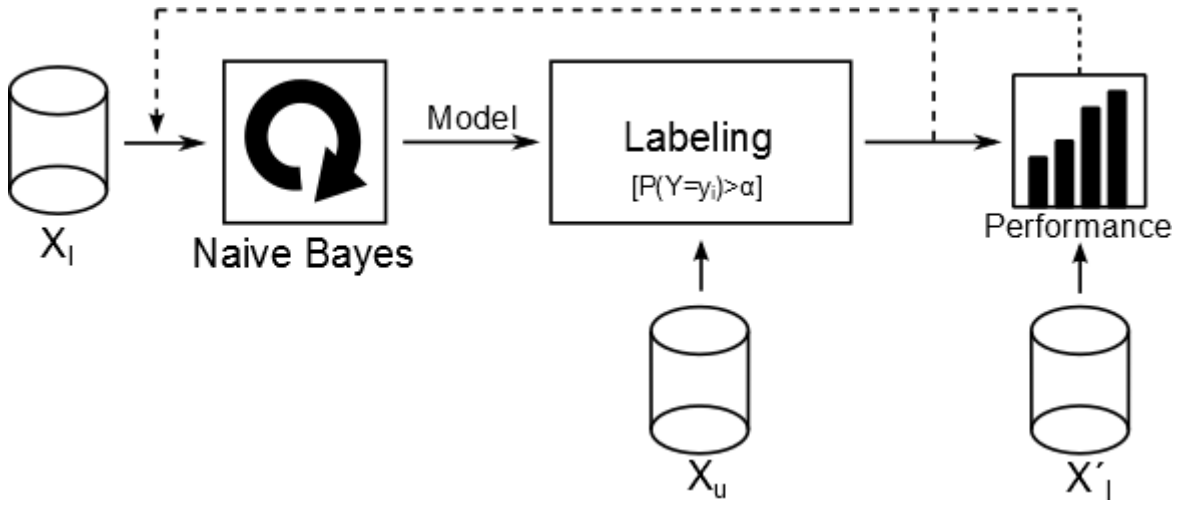
(d) model learned from labeled and unlabeled data

Semi-supervised Learning Assumptions

- **Smoothness assumption**
  - If two examples x1 and x2 in a high density region are close to each other then also corresponding outputs y1 and y2 should be close.
- **The Cluster Assumption**
  - If the examples are in the same cluster, they are likely to be of the same class. The decision boundary should lie in a low density region.
- **The Manifold Assumption**
  - The (high-dimensional) data lie (roughly) on a low dimensional manifold.

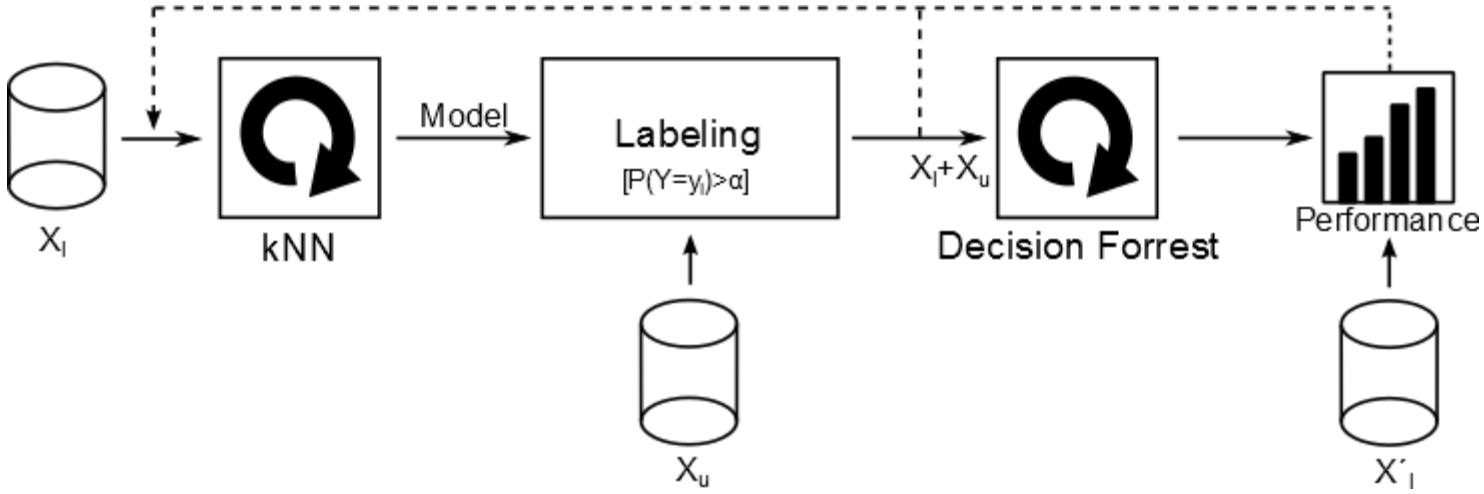
Self training

- The assumption of self-training is that its own predictions, at least the high confidence ones, tend to be correct.
- Experiments with several learning algorithms.
- Stopping criteria and certainty threshold appeared as really non trivial task.



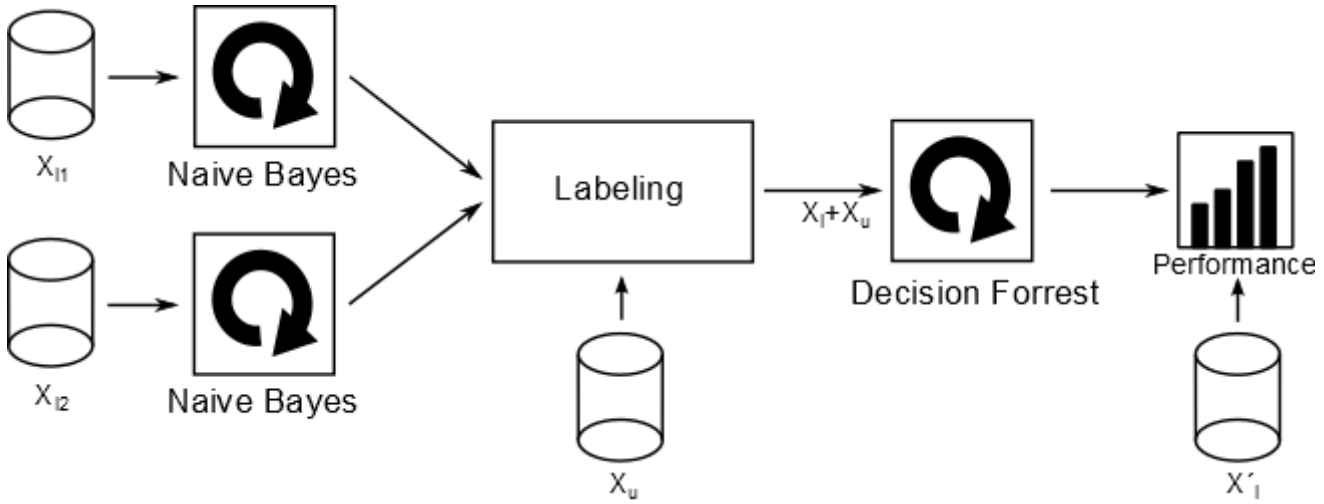
Co-training

- Incorporating new final classifier to prevent model bias.
- Minimizing the error on second classifier.



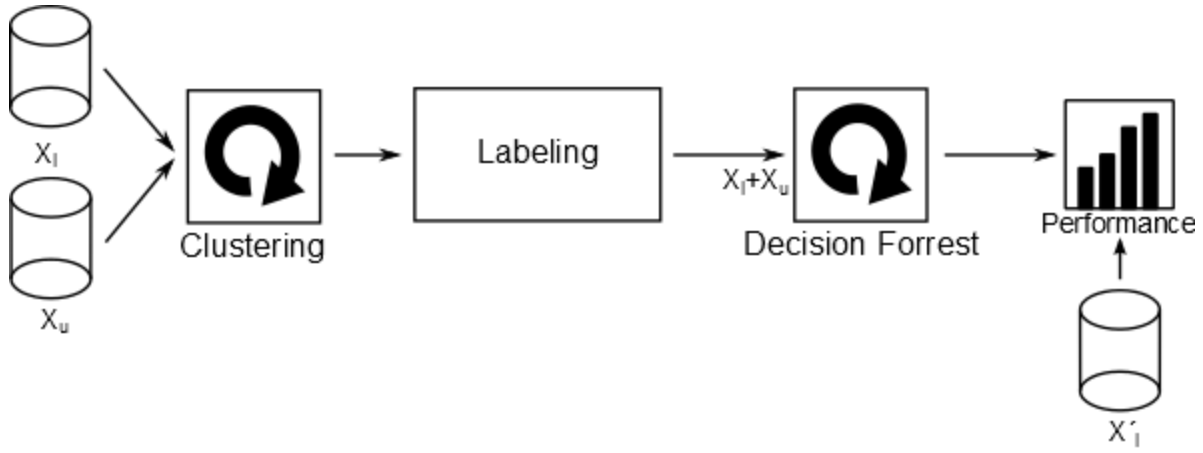
Co-training

- Two classifiers with different feature subsets.
- When they agree on classification use the sample to train third classifier.

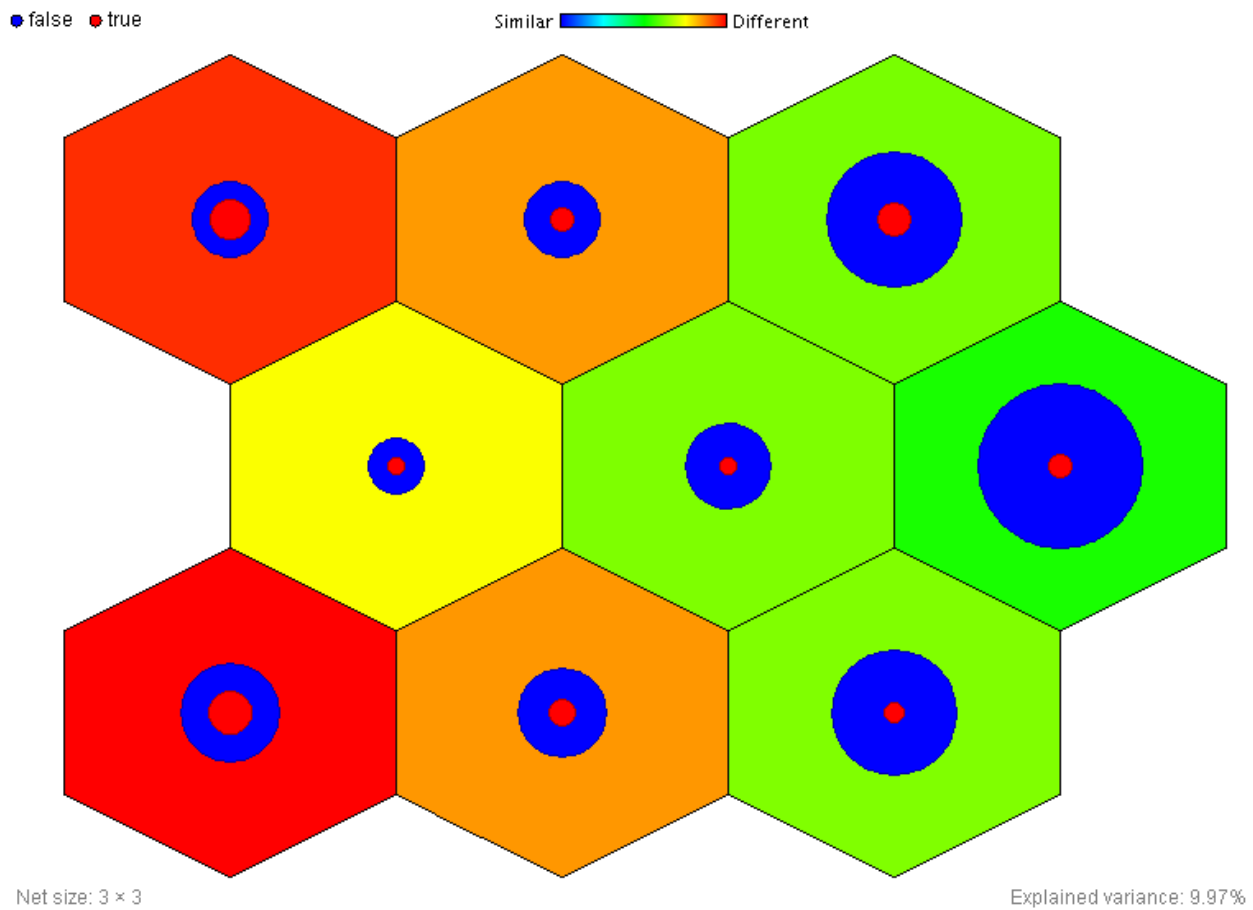


Cluster And Label

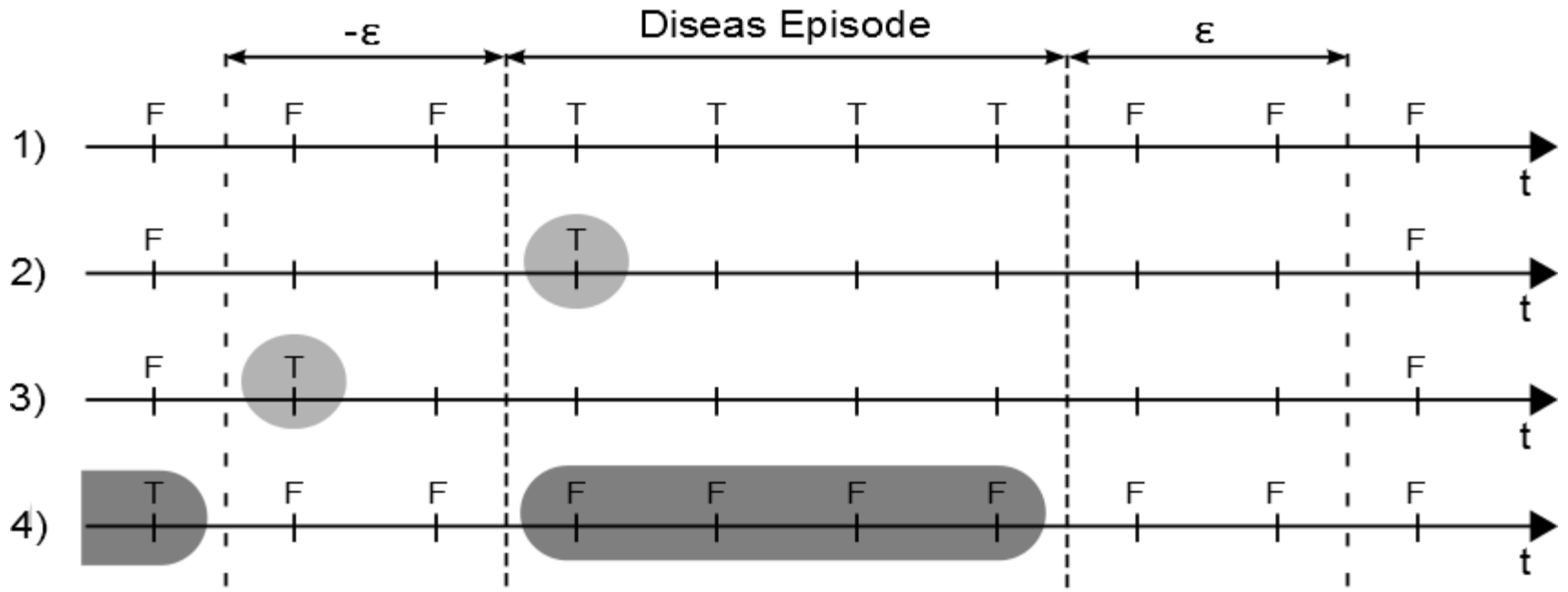
- Cluster whole data set using some clustering technique.
- Label each cluster with labeled examples within the cluster.



Cluster and Label



Evaluation



- Problem of Early Detection
  - Difficult to find cut off between early detection and high precision.

How to evaluate the final model?

- Reliable evaluation data set is missing.
- We have self created training set.
- Can we self create the evaluation set?

- Take only true positives – registered and treated diseases.

Heat Detection

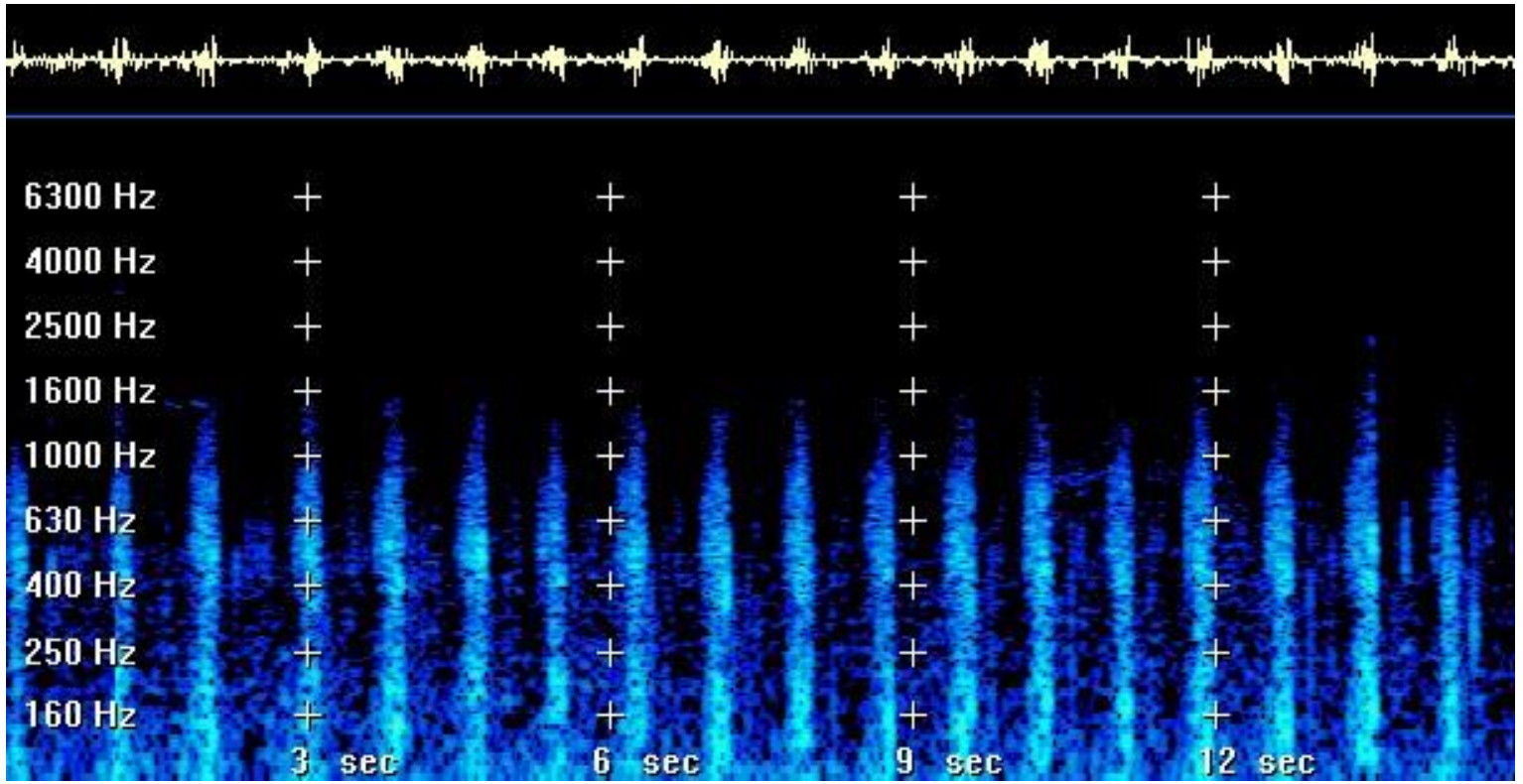
Project Goal

- The goal was to detect heat of cows as early as possible from activity and rumination data.

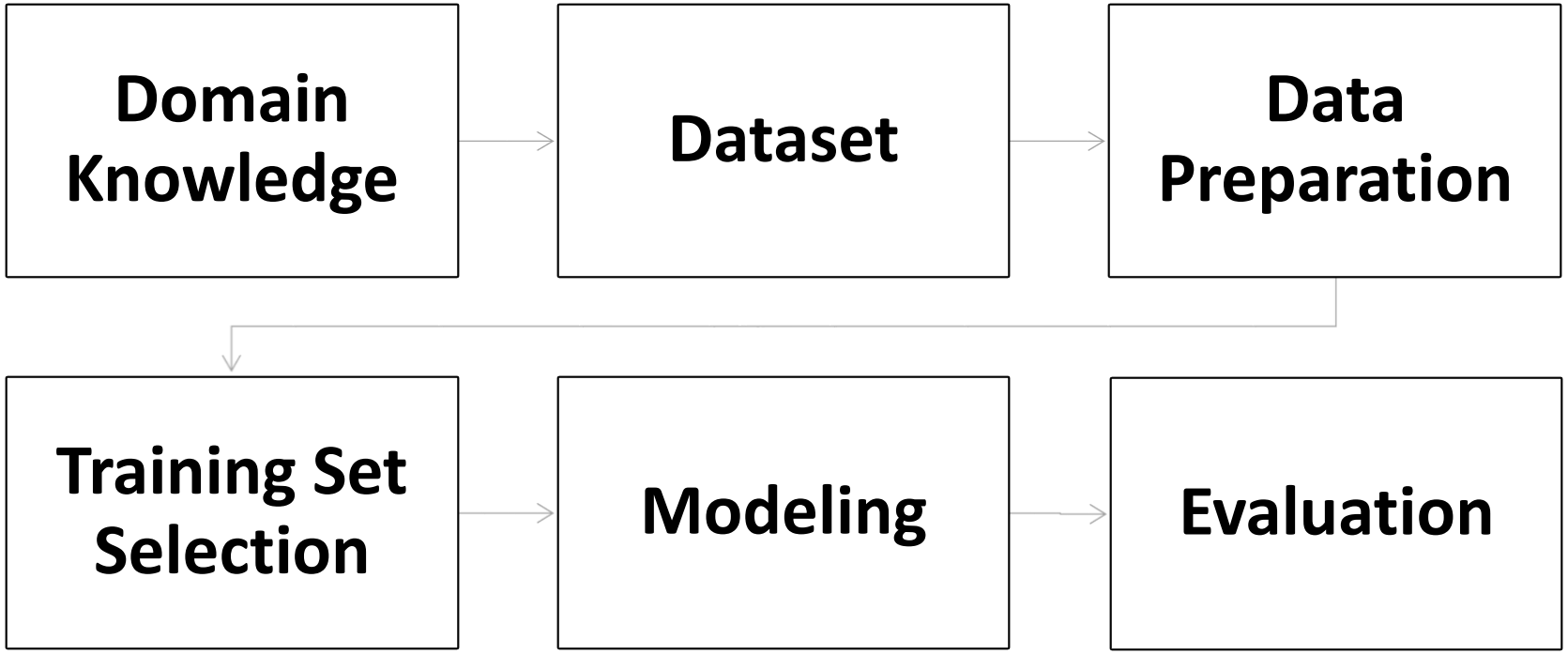
Heat detection

- Activity sensors
- Wireless communication
- Rumination

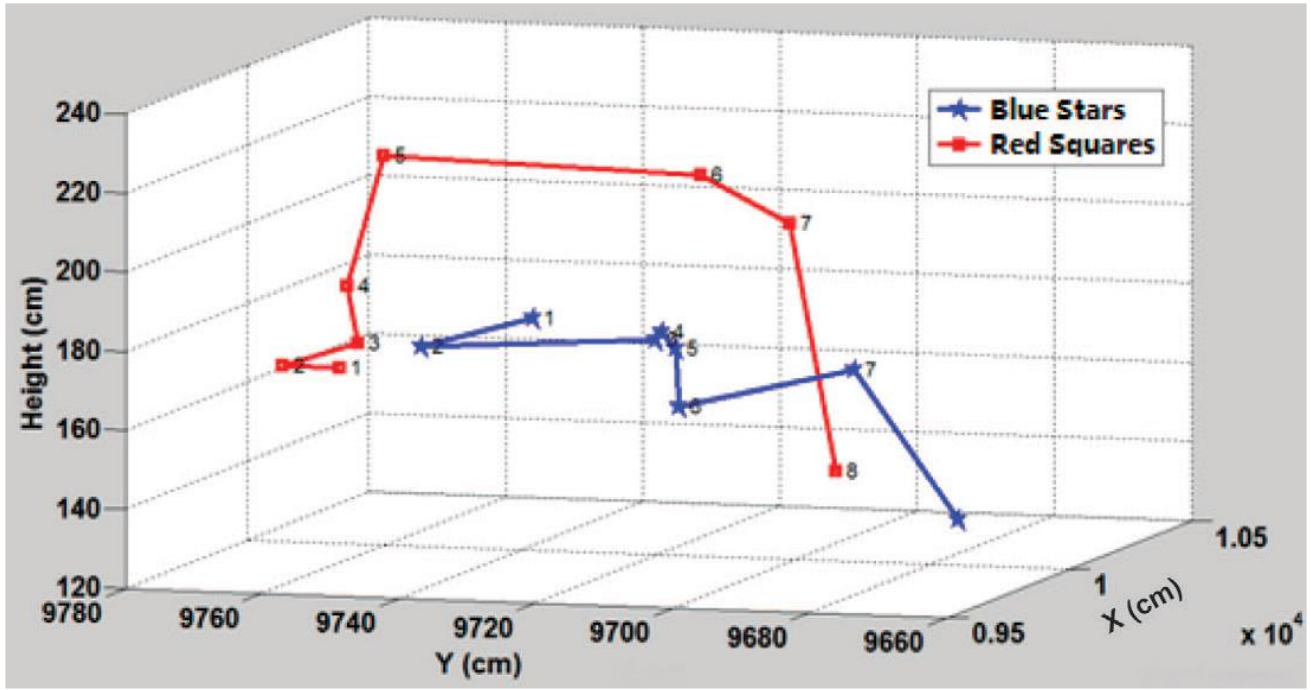
Typical Chewing Audio Signal



Data Mining Process



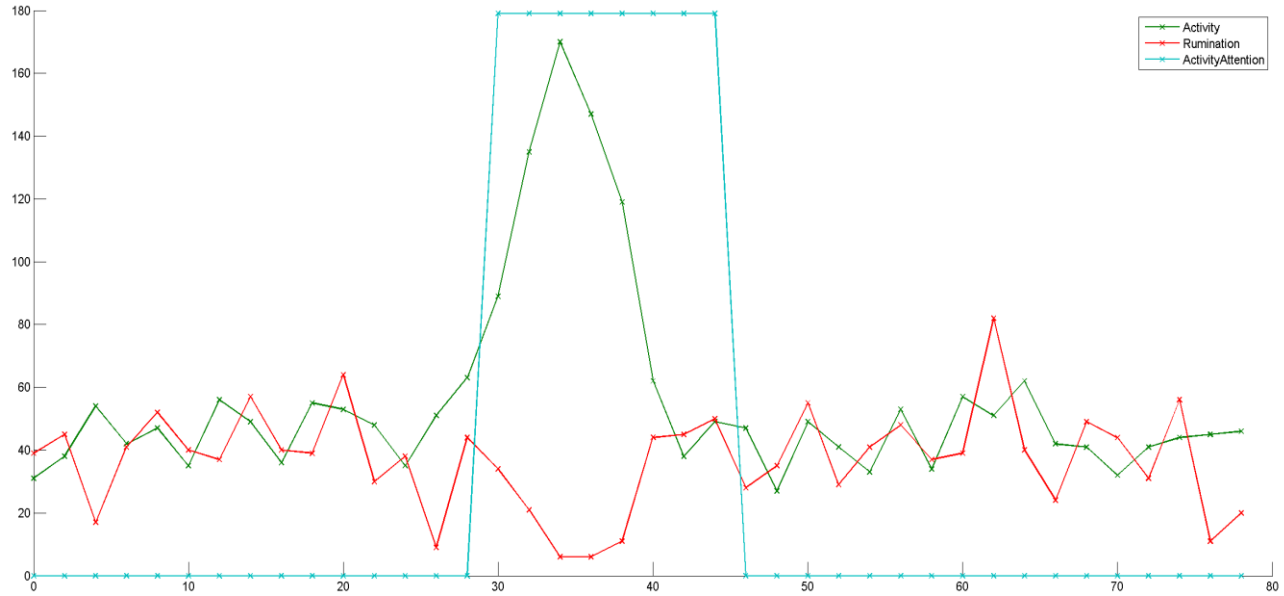
Domain Knowledge



Homer, E.; Gao, Y.; Meng, X.; et al. Technical note: A novel approach to the detection of estrus in dairy cows using ultra-wideband technology. Journal of dairy science, volume 96, no. 10, 2013: pp. 6529-6534.

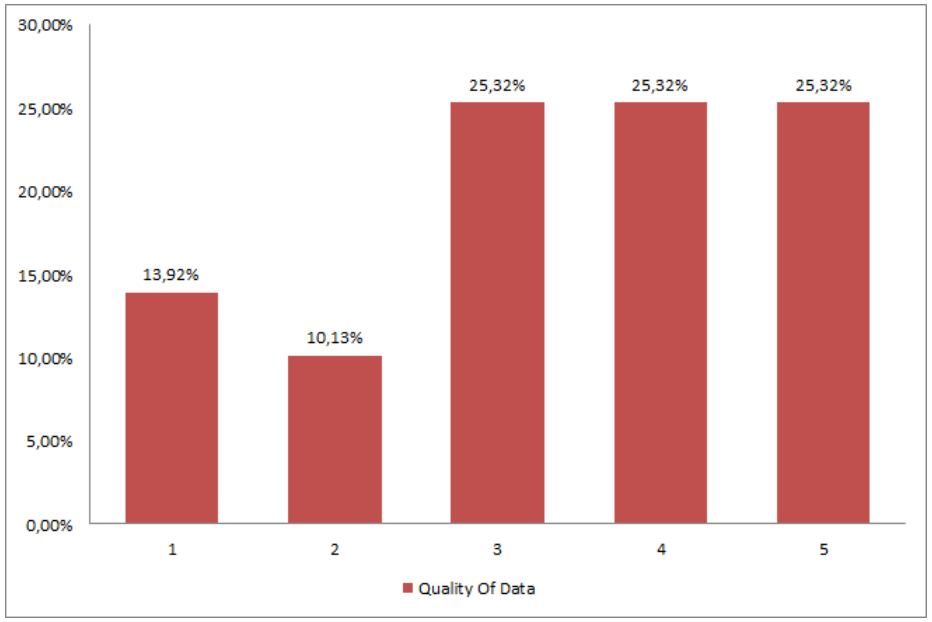
Domain Knowledge

- During heat
  - activity is increasing.
  - rumination is decreasing.
- Heat is every 3 weeks.



Data Quality

- Every cow marked by value from scale bad ★★★★★ good.

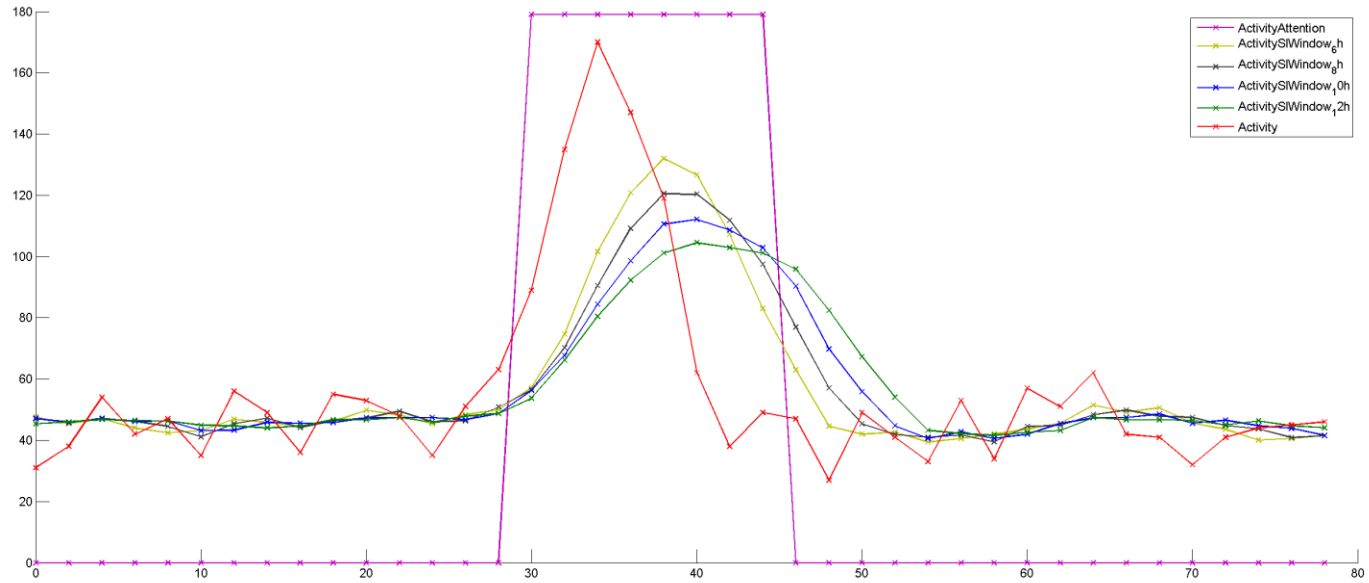


Data Preparation

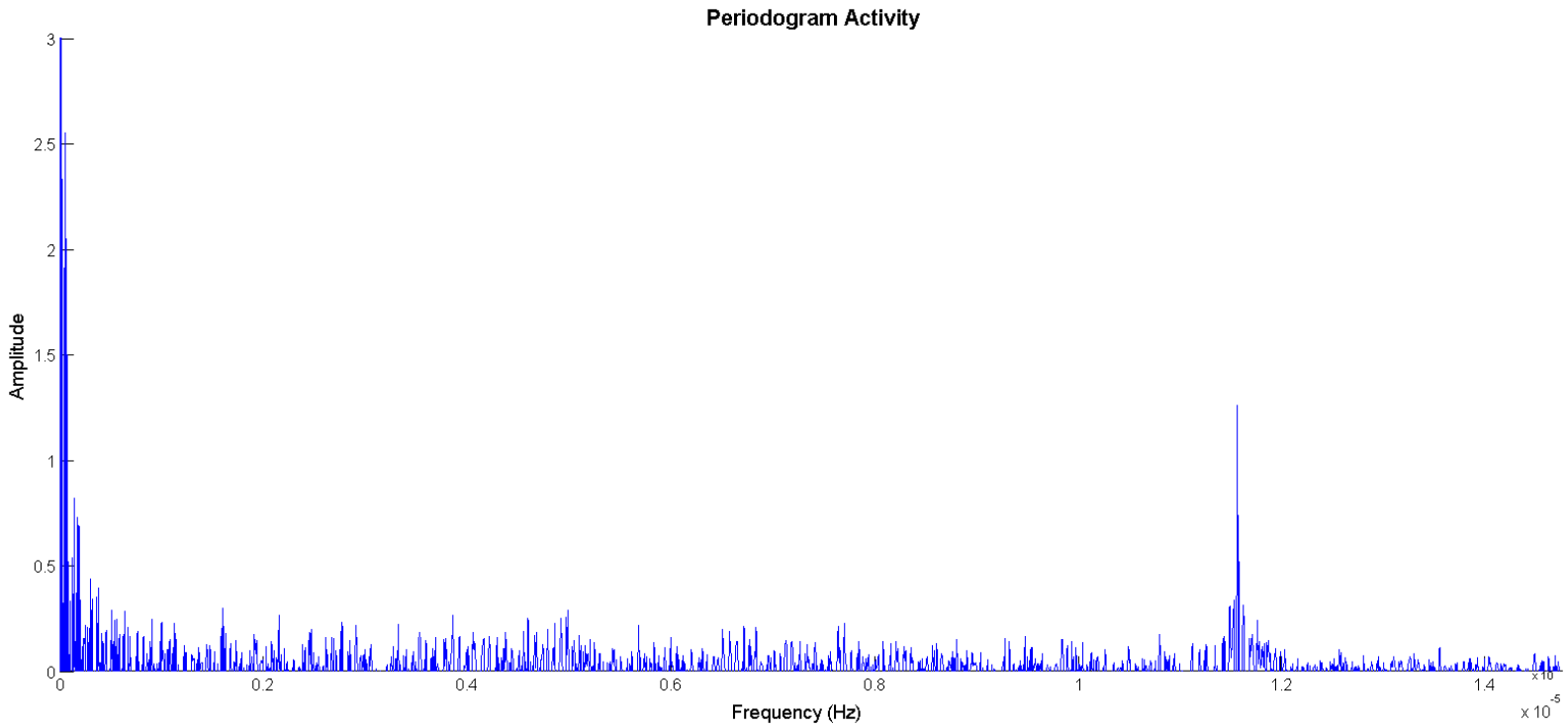
- Data Smoothing
  - Moving average
  - Butterworth filter
- Feature Extraction
- Data Normalization

Moving Average

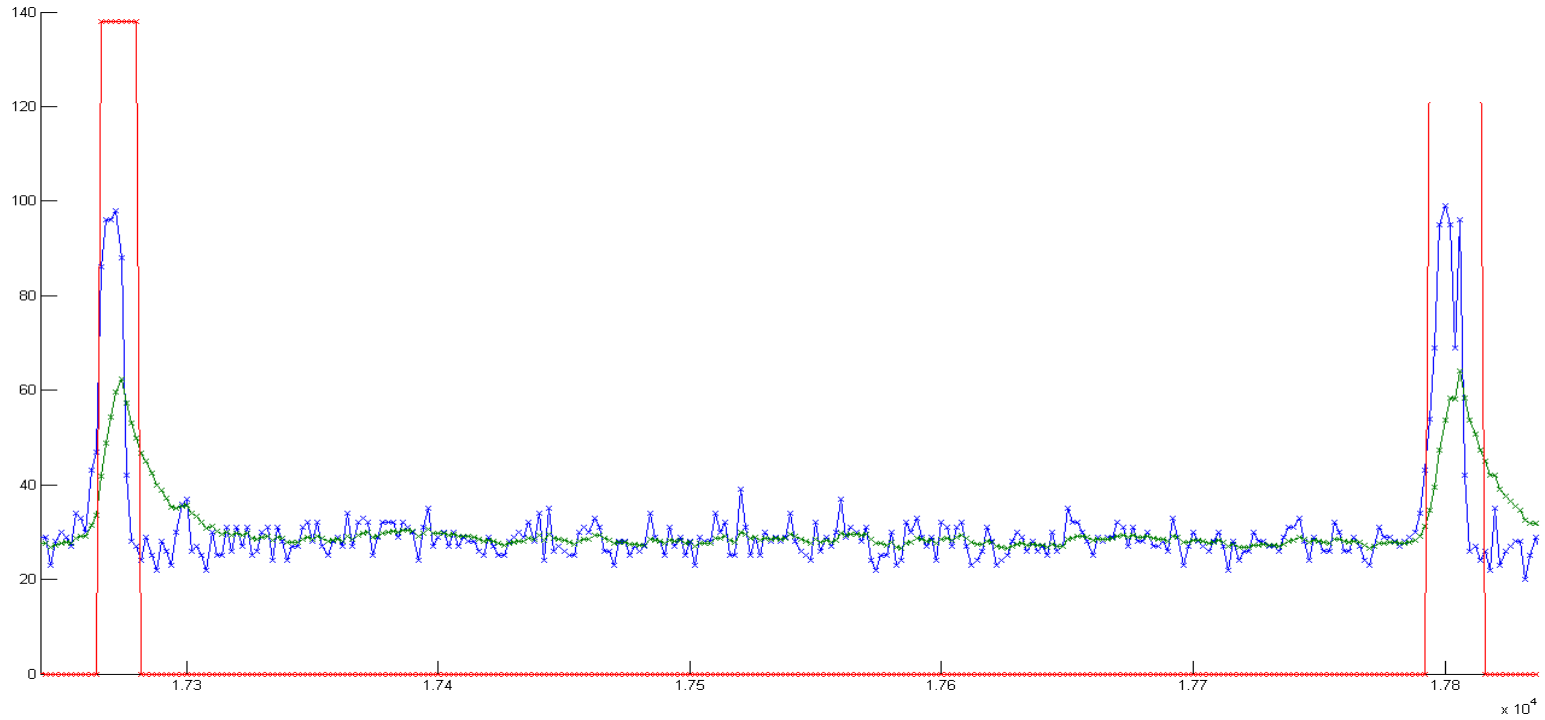
- Event is delaying => worse early detection.



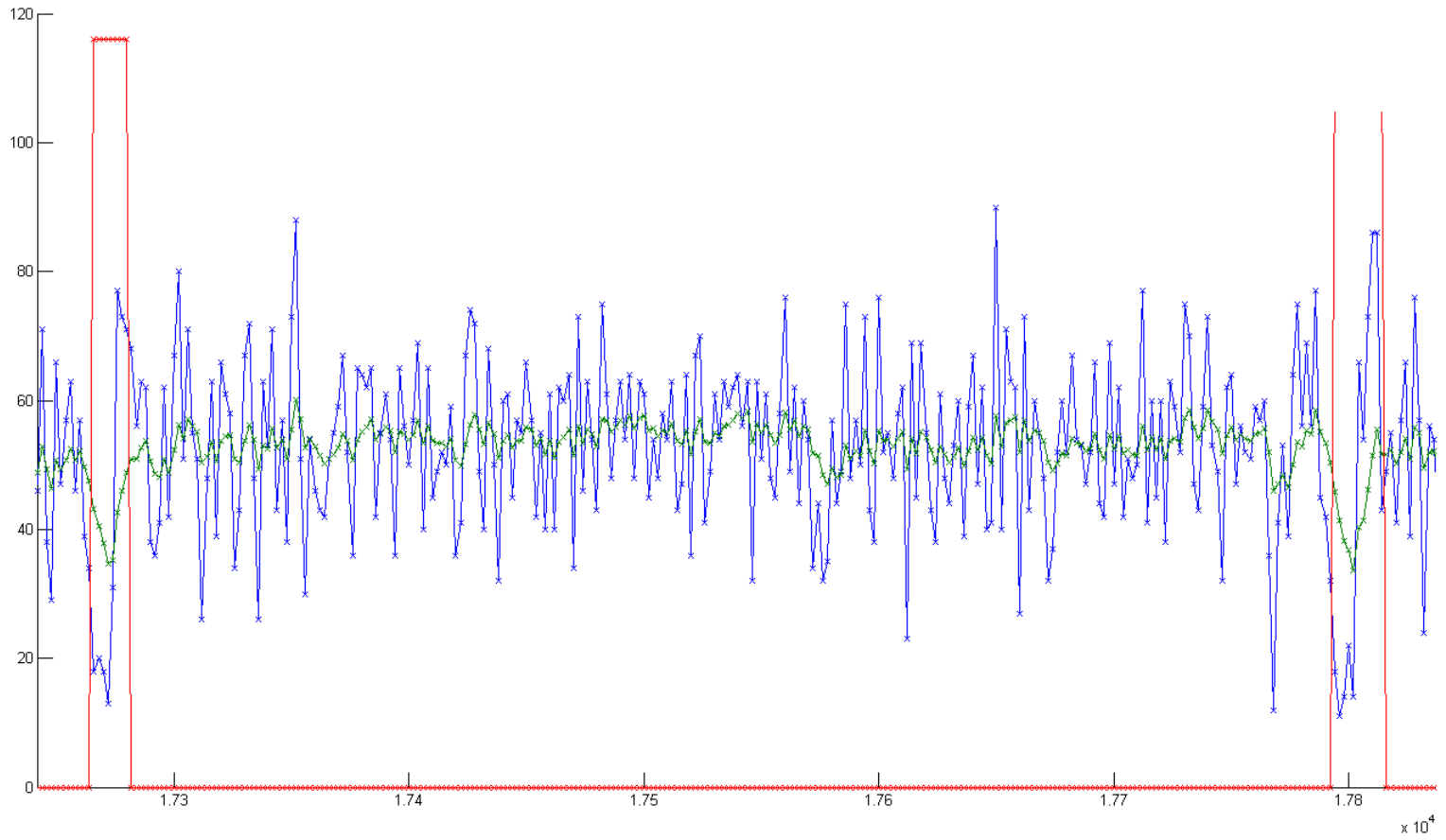
Fourier Analysis



Filter Activity



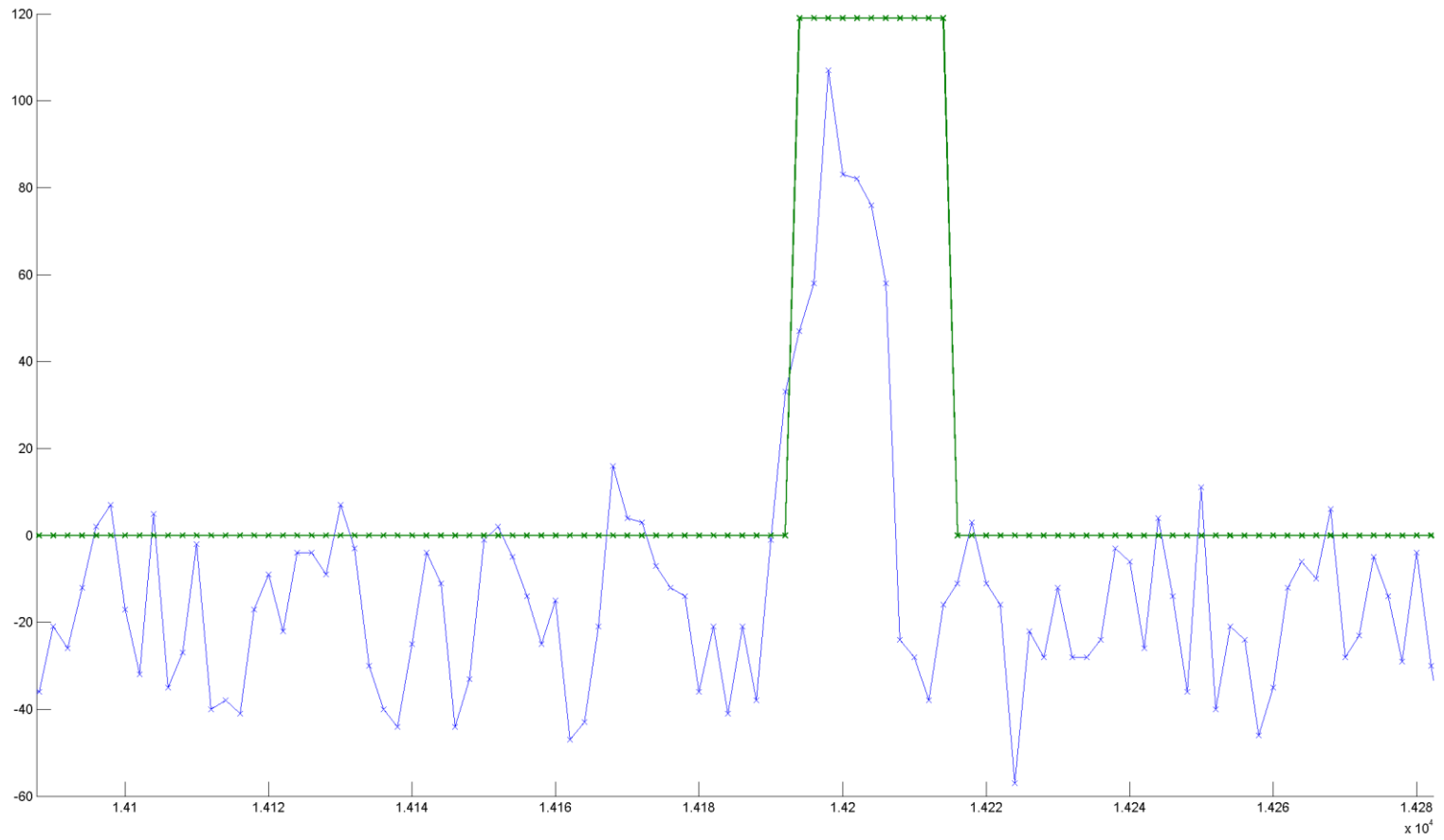
Filter Rumination



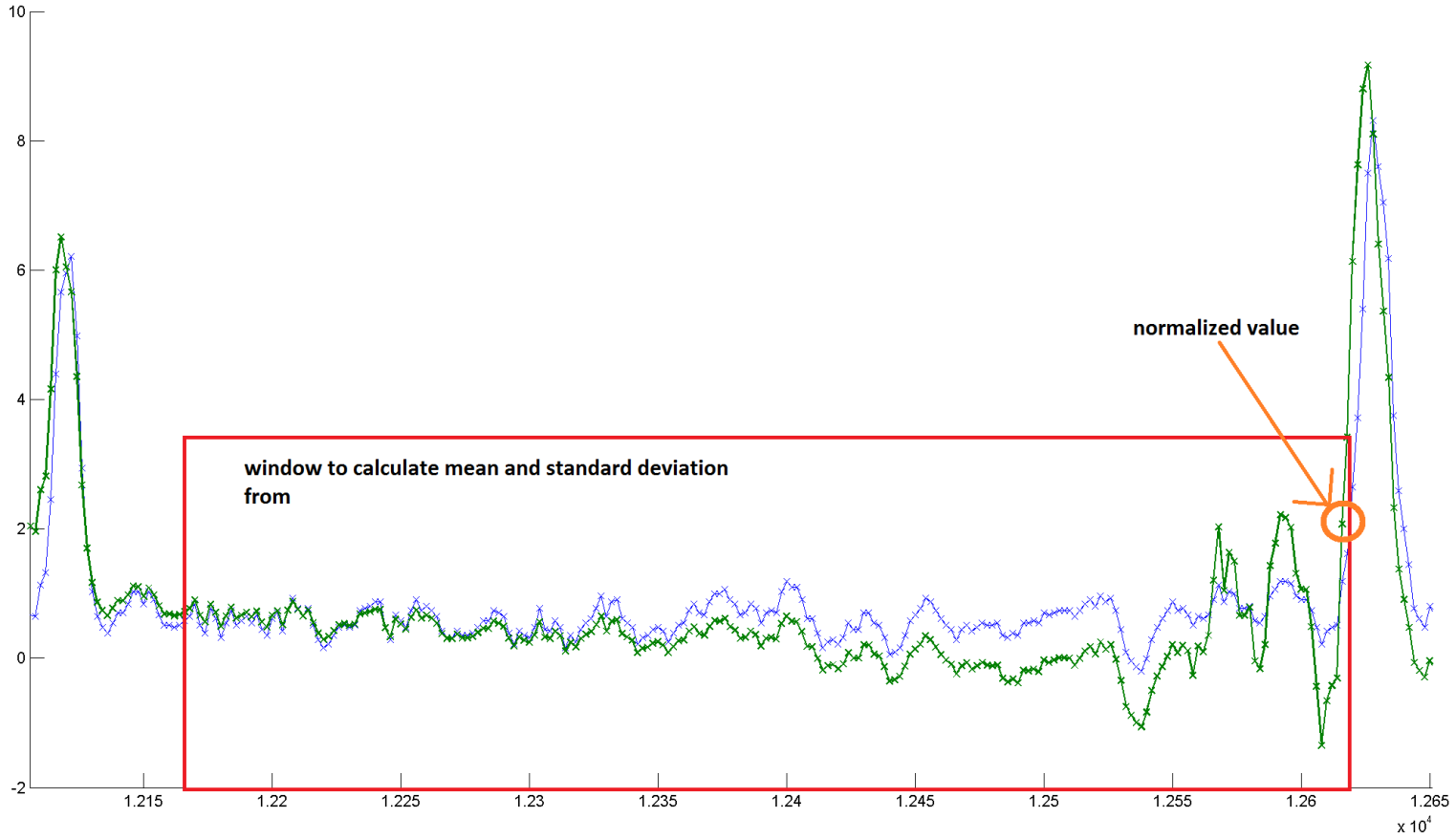
Feature Extraction

- Some features have parameters (like size of a window).
- More than **600** features in total.

Feature Extraction



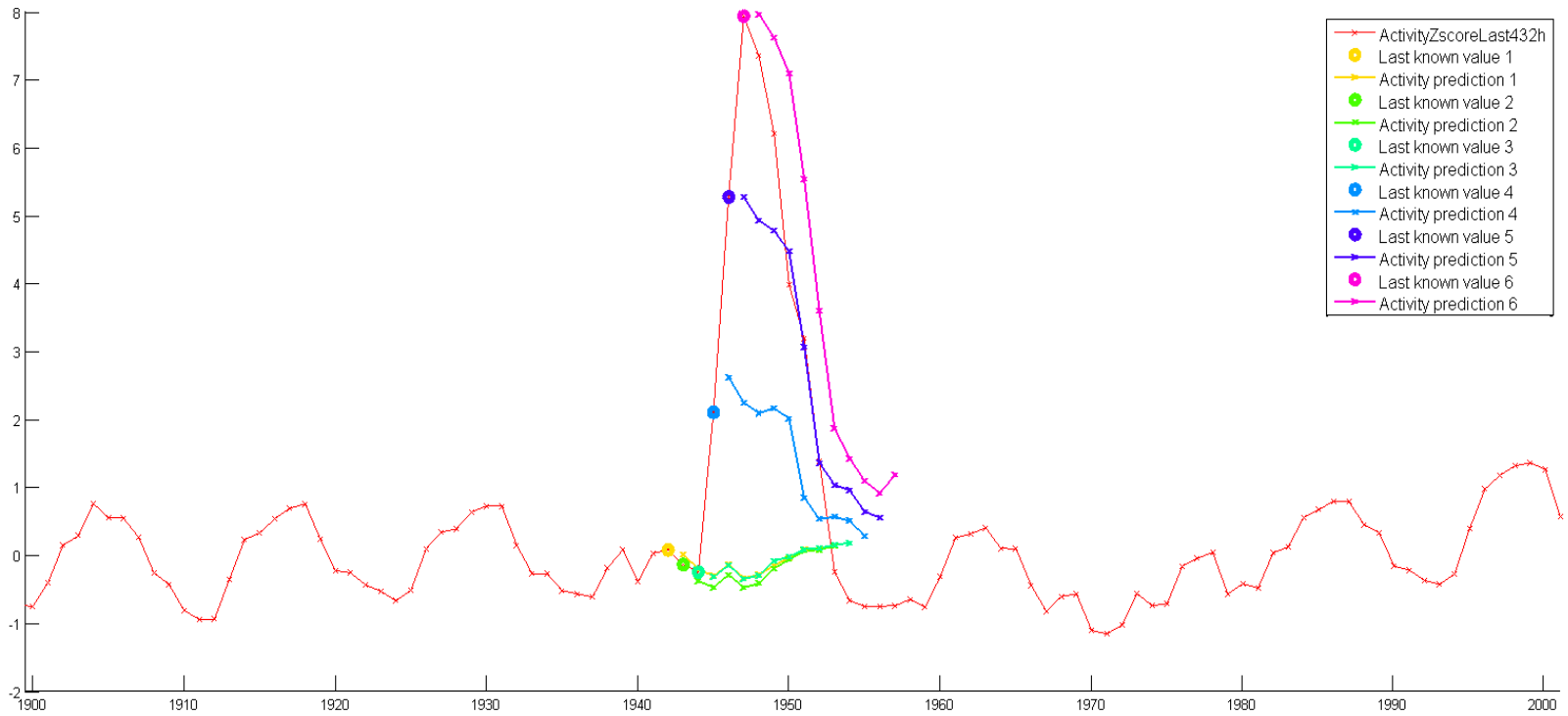
Data Normalization



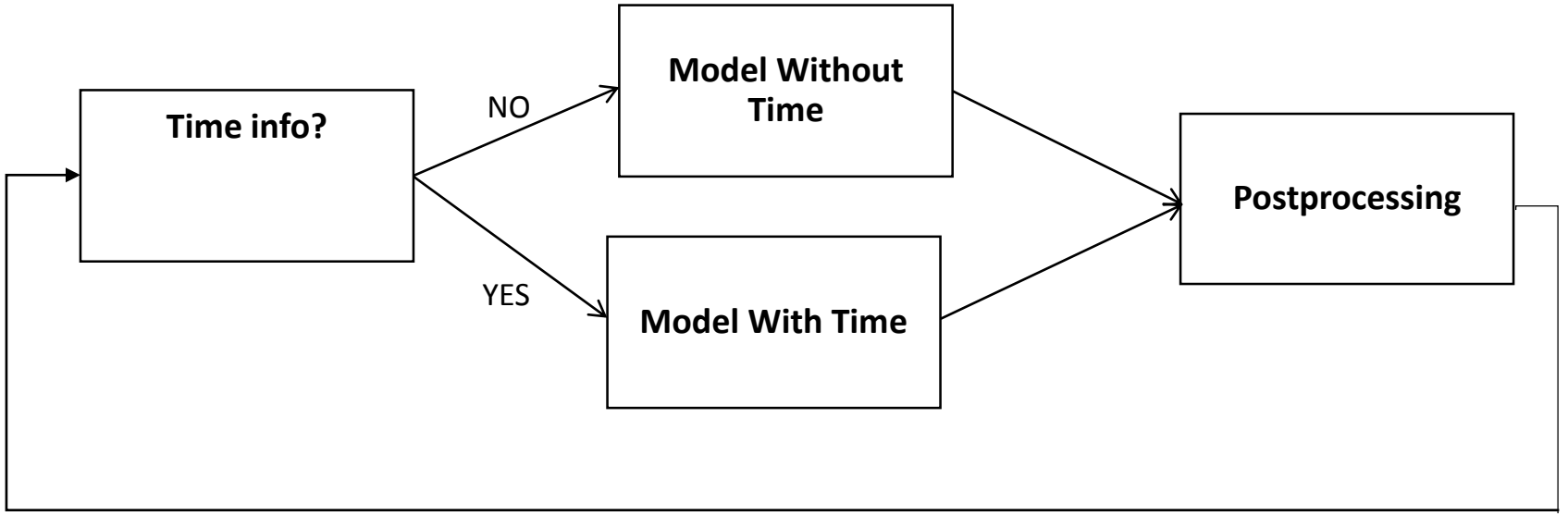
Modeling

- Classification
  - Naive Bayes, Decision tree, Random forest
- Event detection
  - Moving average detection
- Time series analysis
  - ARIMA

Evaluation - ARIMA



Model With Postprocessing



Questions?