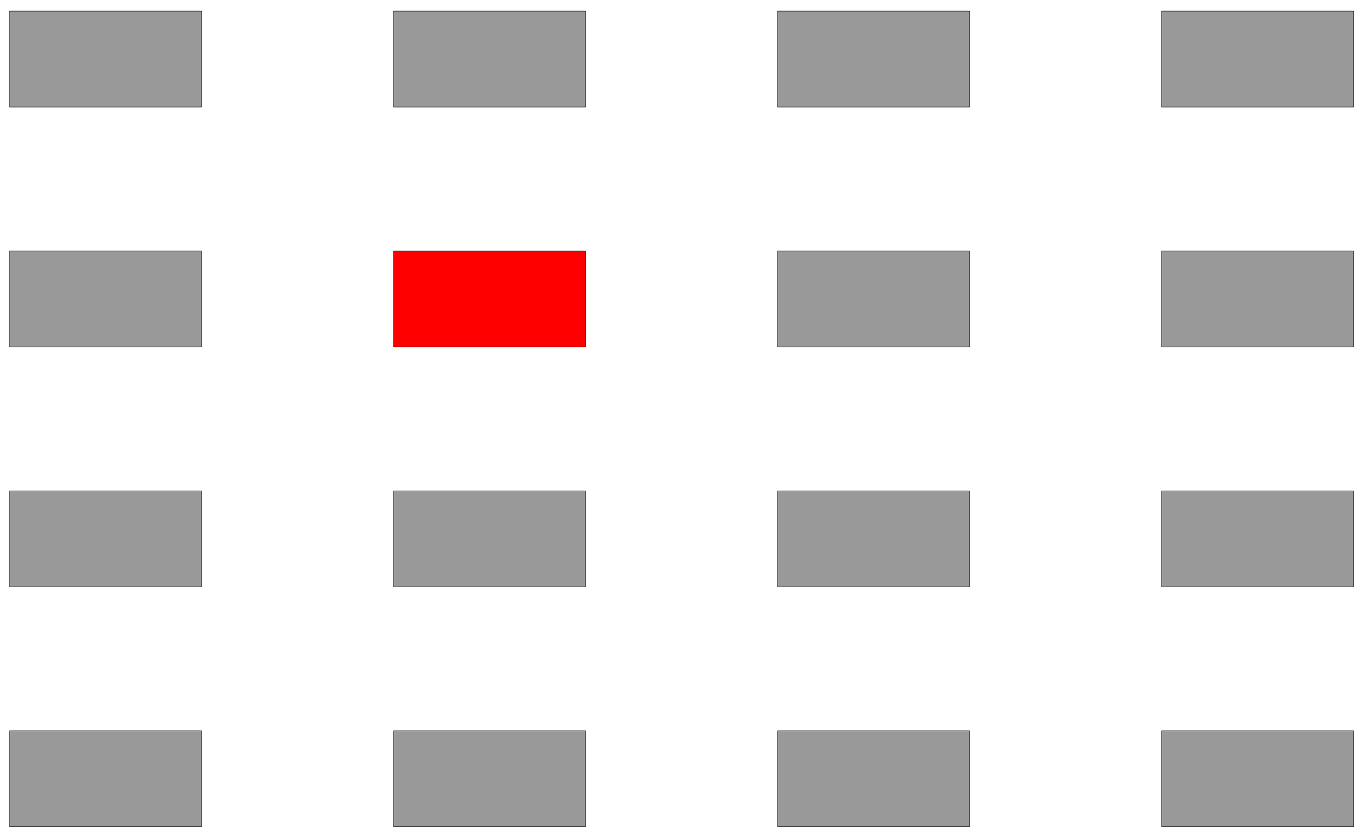
Anomaly Detection

Nathan Dautenhahn

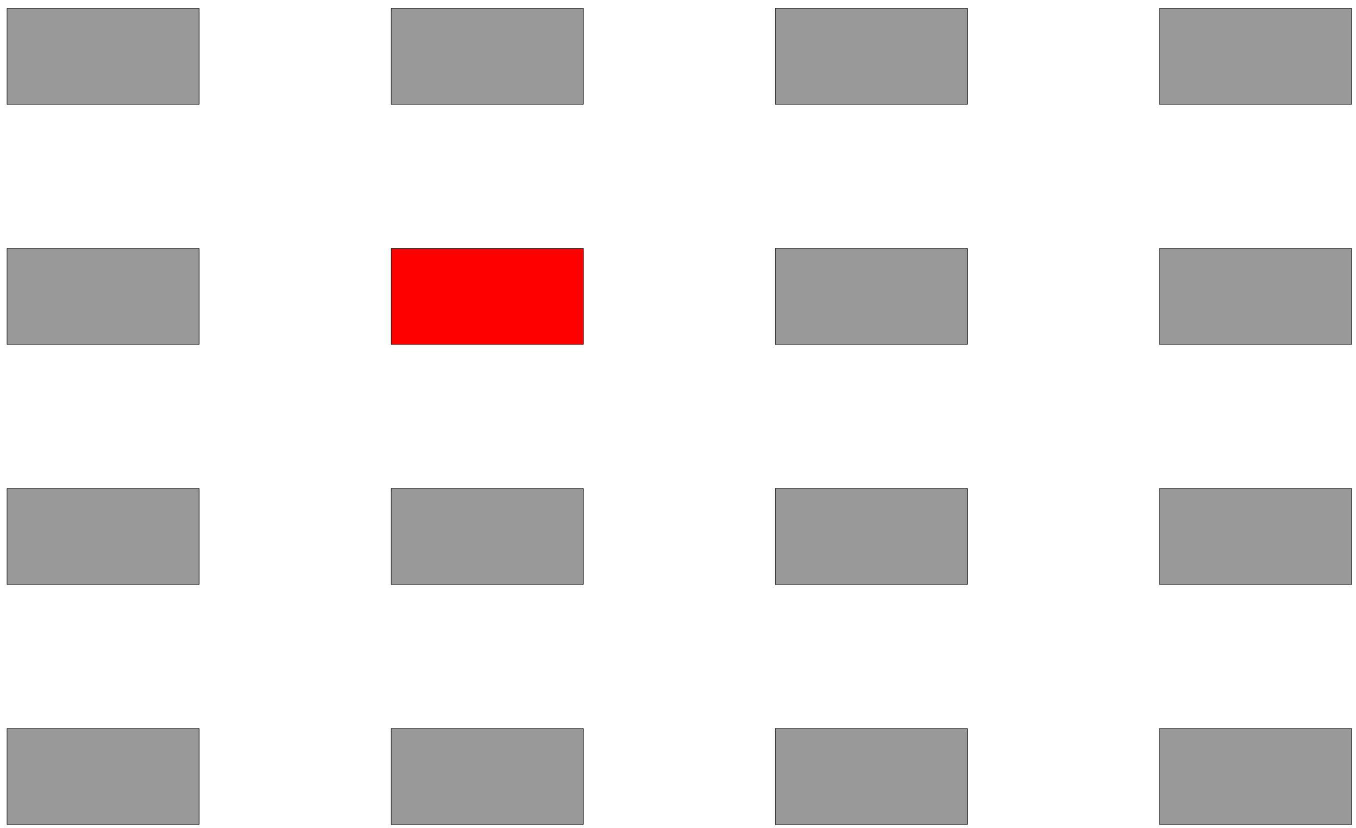
CS 598 Class Lecture

March 3, 2011

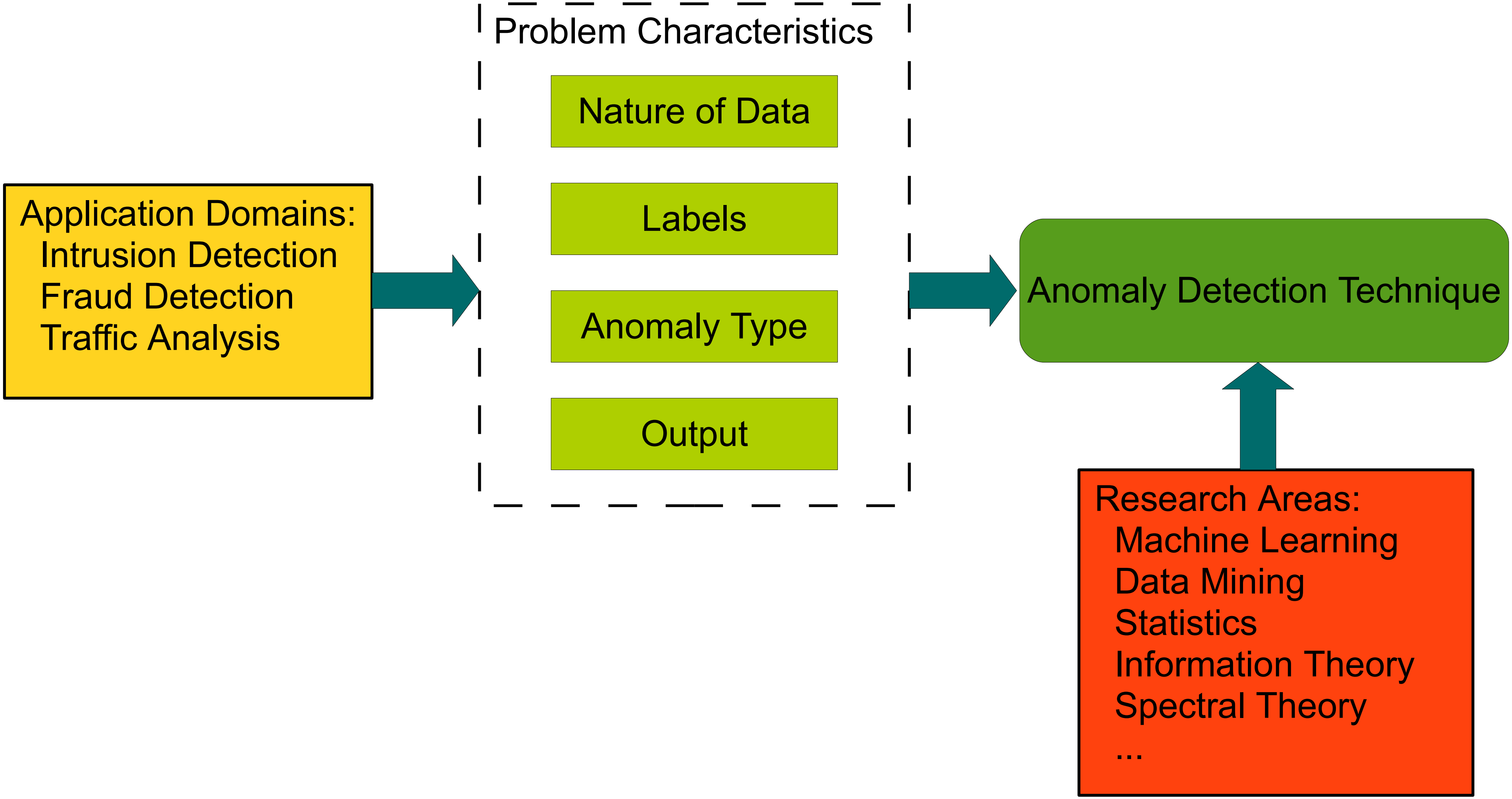
An anomaly is a deviation from the normal or expected behavior



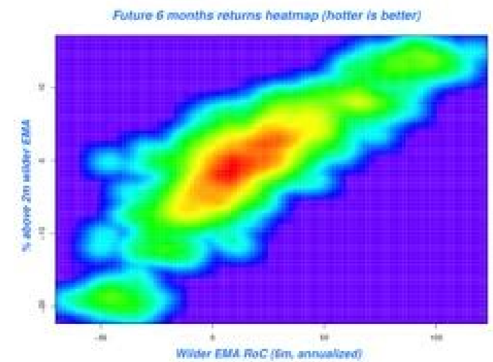
This seems easy, why even worry about it?



Anomaly detection solves these problems in several diverse ways...



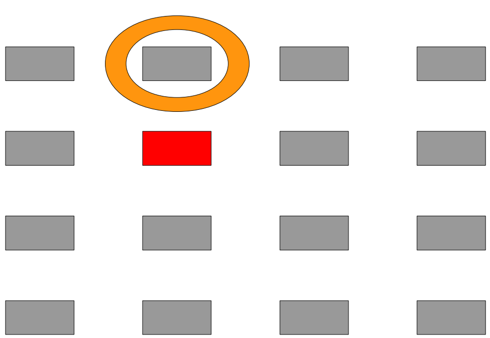
What types of data do we have?



Data Instance ↔ Point | Relational

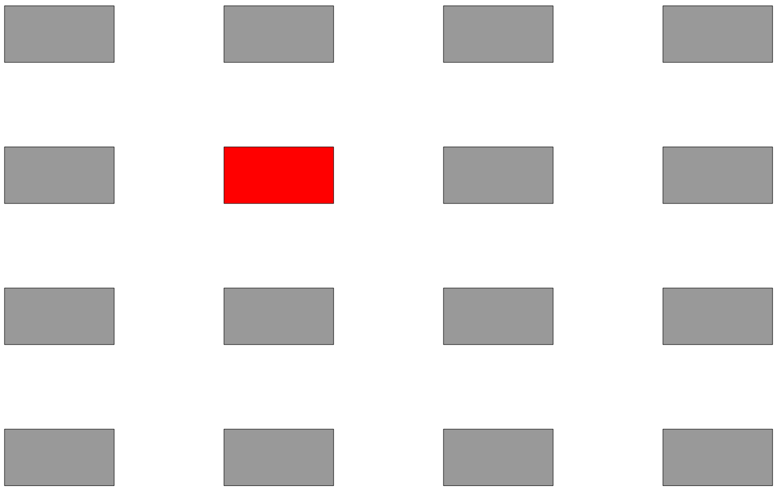
Attribute Set ↔ Univariate | Multivariate → {color,area,location}

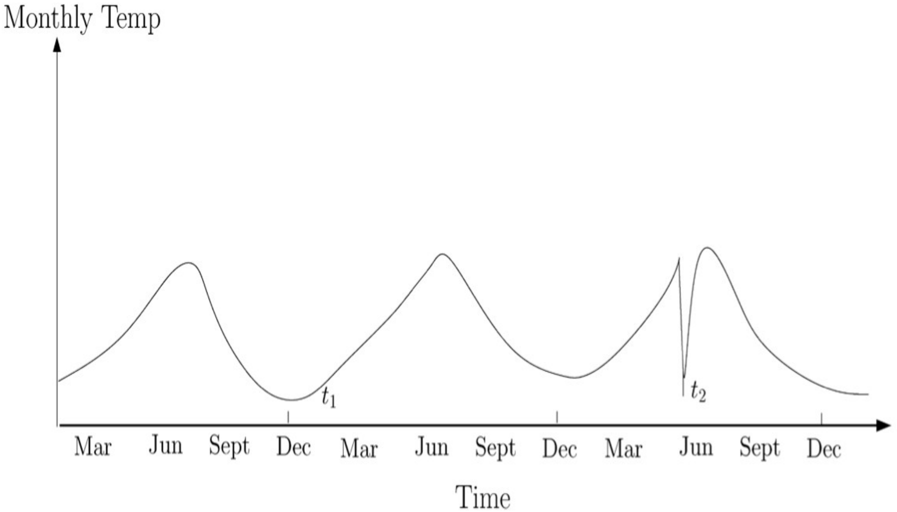
Type ↔ Binary | Categorical | Continuous



Categorical	Continuous	Binary
Color	Area	In Set X

Anomalies can be classified as point, contextual, or collective





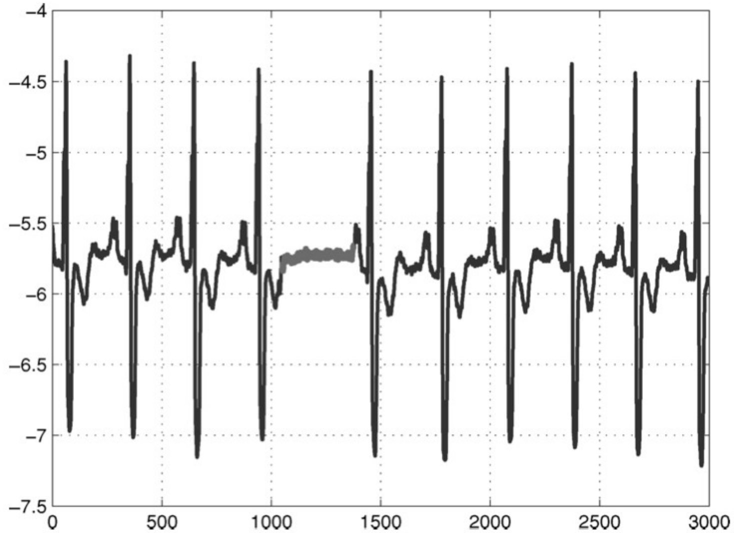
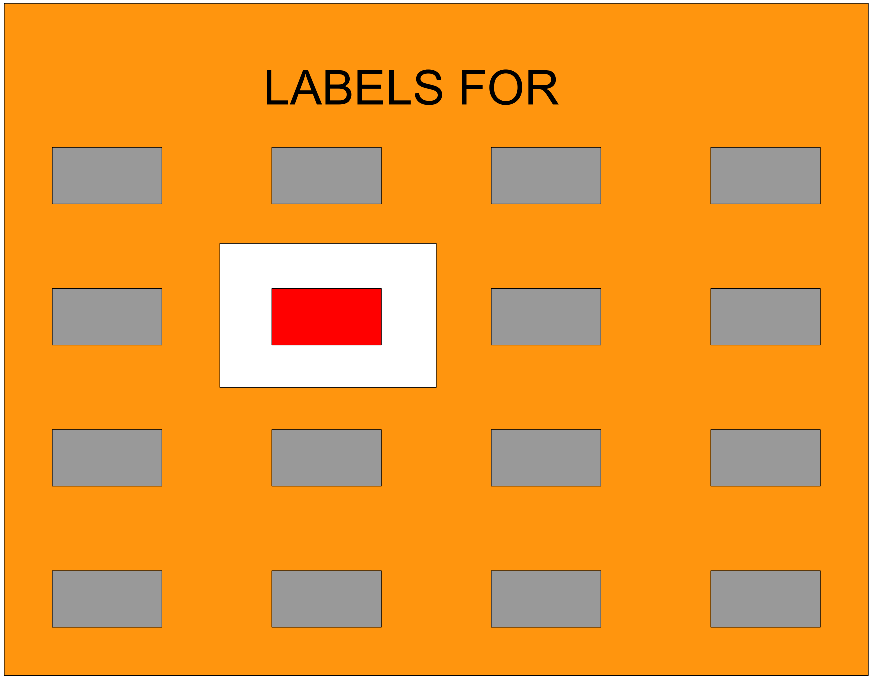


Fig. 4. Collective anomaly corresponding to an *Atrial Premature Contraction* in an human electrocardiogram output.

[Images from Chandola 09]

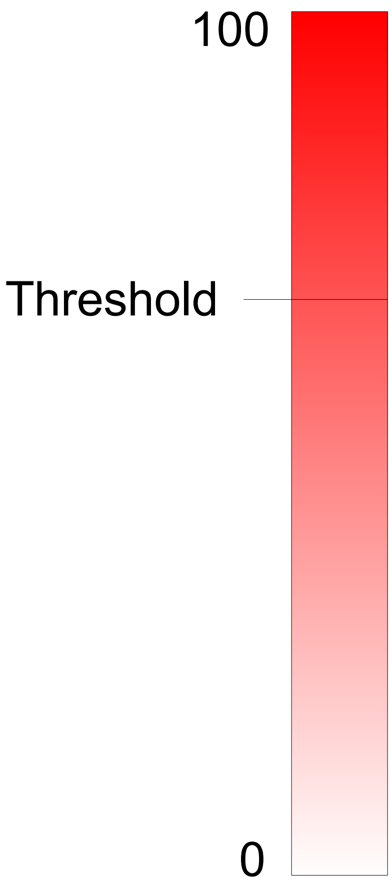
Data must be labeled as anomalous or normal in a training phase

Semi-supervised Unsupervised



Anomaly output in the form of either scores or labels

Anomalous Level Score



Labeled

{ Normal, Anomalous }

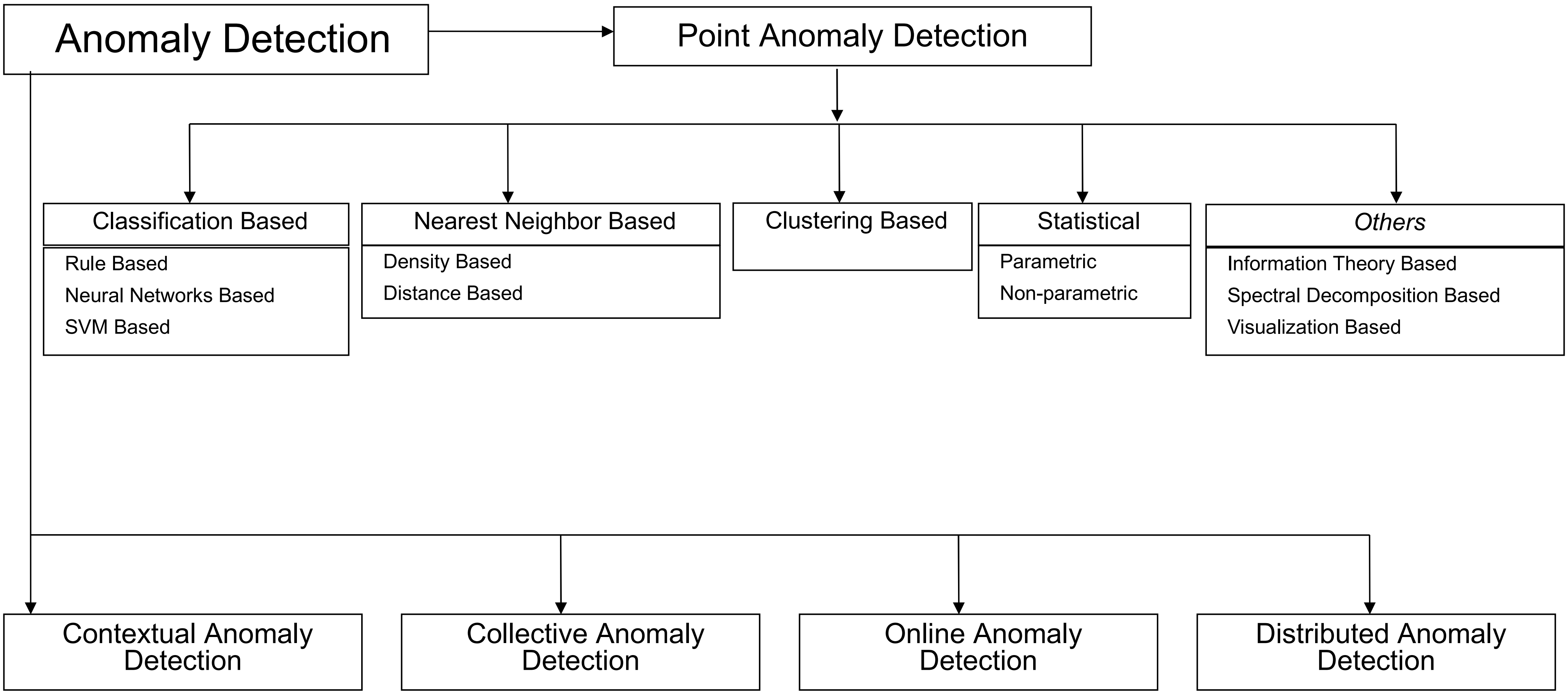
Classifying techniques is hard but their exists a set of high level areas

Statistical | Machine Learning | Data Mining

Categorical	Nearest Neighbor	Clustering
Spectral	Information Theory	Statistical

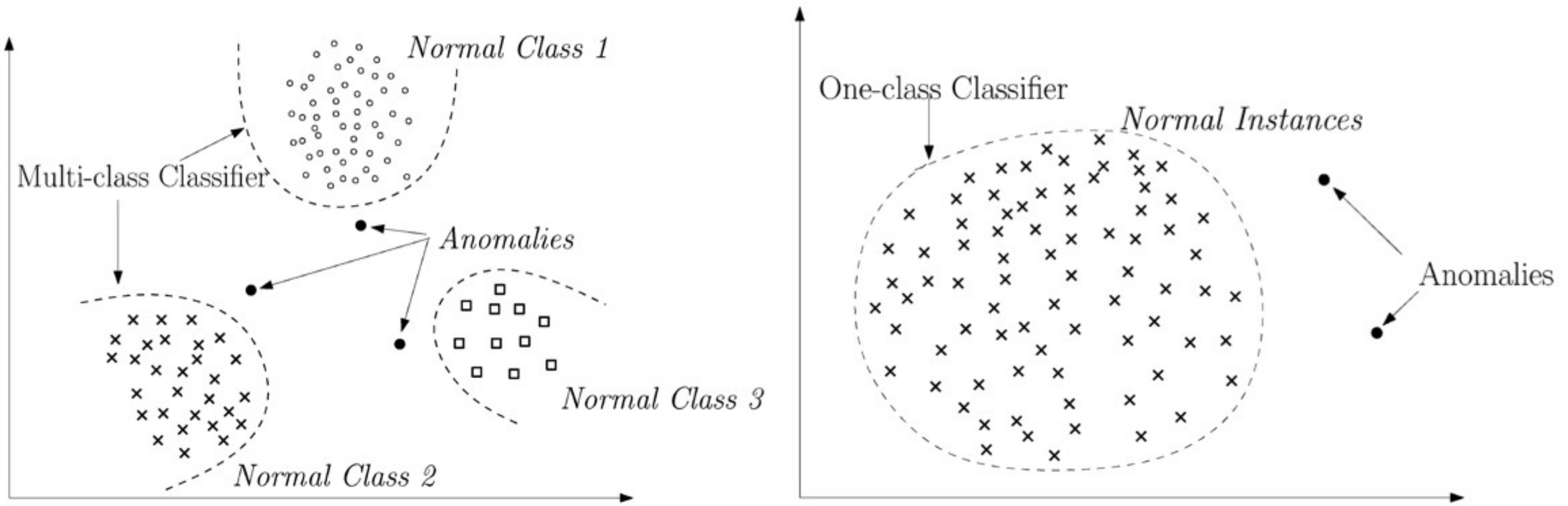
Irrelevant for this discussion

Taxonomy*



* Outlier Detection – A Survey, Varun Chandola, Arindam Banerjee, and Vipin Kumar, Technical Report TR07-17, University of Minnesota

Classification based techniques learn a model on training data and classify test instances

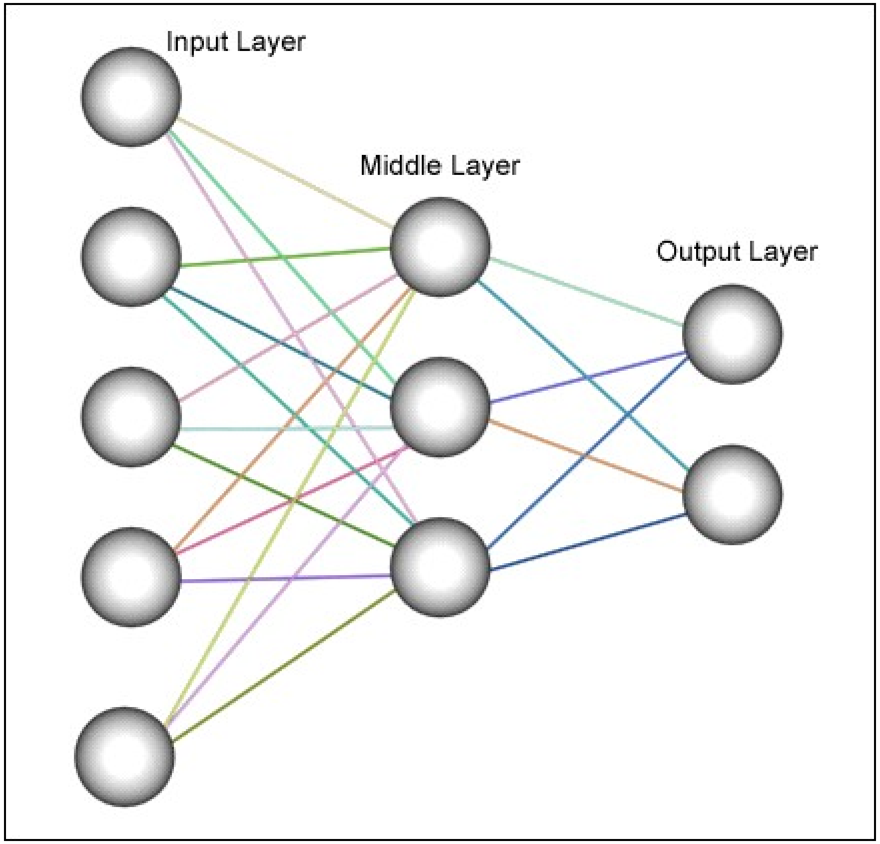


(a) Multi-class Anomaly Detection (b) One-class Anomaly Detection

- Neural Networks
- Rule Based
- SVM
- Bayesian
- Fuzzy Logic
- Genetic Algorithms

[Image from Chandola 09]

Neural Networks



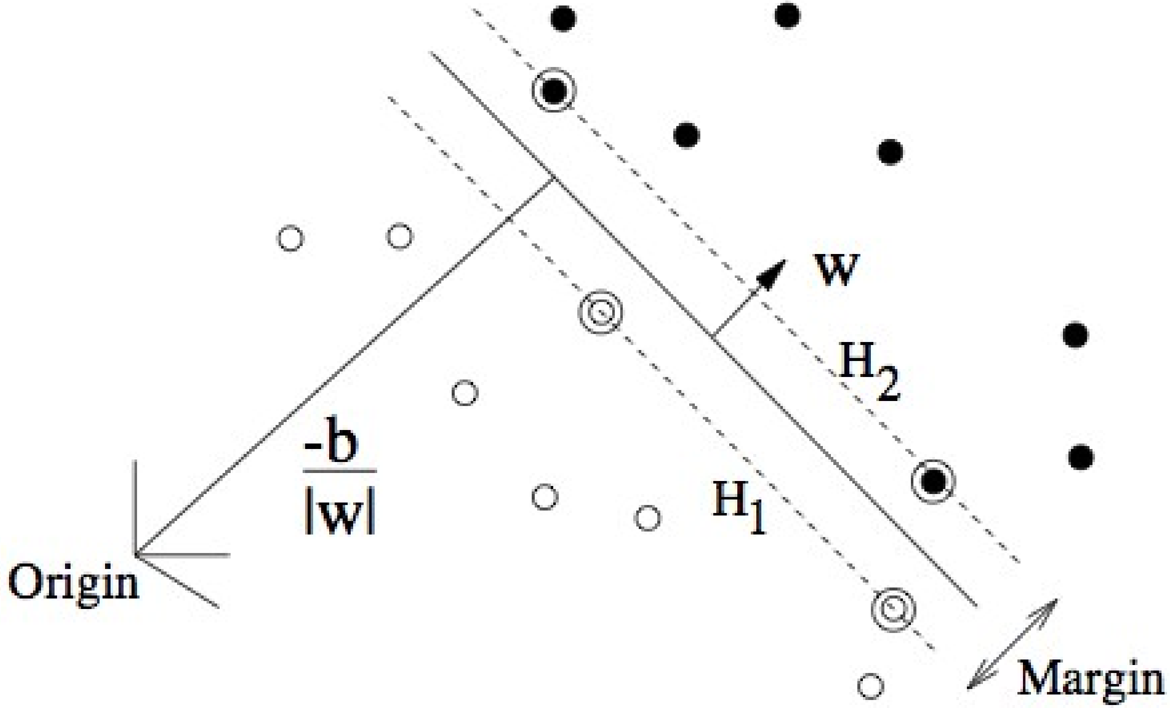
- Good when dealing with huge data sets and handles noisy data well
- Bad because learning takes a long time

Rule based – Misuse Detection

- Rule: Set of permissible actions (if classifying normal data) – categorical
- Approach
  - Learn rule from training data using algorithm: RIPPER, Decision trees
  - Rule has confidence value proportional to the number of training samples matched by the rule

Support Vector Machines

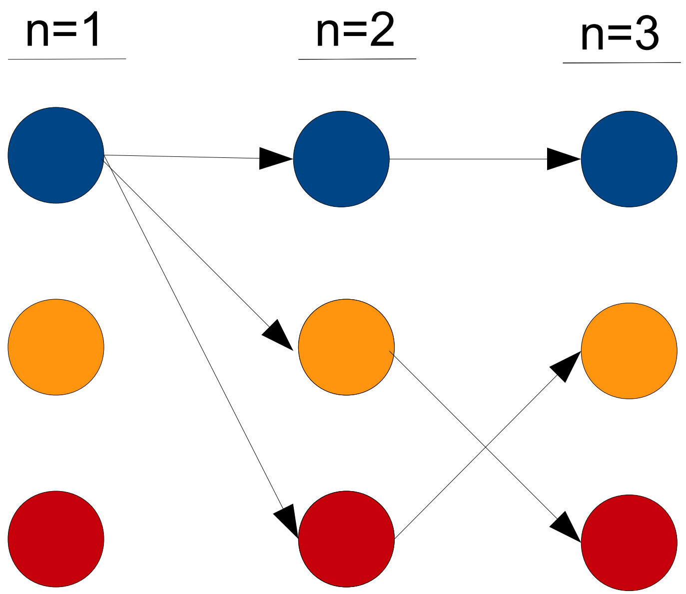
- Problem: Find hyperplane separating two classes of data instances



Markov Chain

- Problem: Determine whether color of an object is anomalous

	Blue	Yellow	Red
Blue	.01	.75	.249
Yellow	.249	.01	.75
Red	.75	.249	.01



Bayesian Networks Example

- Assume independence
- Infected milk example
- Hypothesis: infected
- Information Variable: positive
- Positive conditionally depends on infected
- Given output of information variable, calculate the a-posteriori probability of the hypothesis

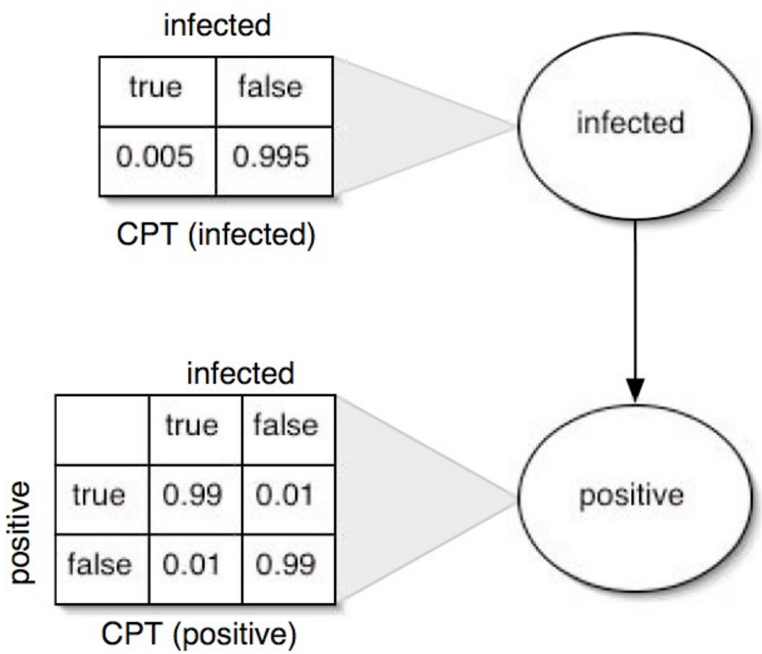


Figure 1. Bayesian Network and CPTs

Given [positive=1] A-posteriori Probability of Infection: 0.33

Bayesian is great, but what if we do not know the conditional probabilities?

- Bayesian Decision Theory shows us how to design an optimal classifier if we know the prior probabilities $P(\Omega_i)$ and the class-conditional densities $P(\mathbf{X}|\Omega_i)$
- Unfortunately: we rarely have complete knowledge of these class-conditional probabilities
- However: we can often find training data that include particular representatives of the patterns we want to classify

[www.csc.kth.se/utbildning/]

There are two general approaches to solving the problems with Bayesian decision theory

- **Parametric**: Assume some parametric form for the conditional densities and estimate its parameters using training data. Then use Bayesian decision rule to classify data instances
- **Non-Parametric**: Make no assumption of the underlying class-conditionals and estimate them completely from the training data.

[www.csc.kth.se/utbildning/]

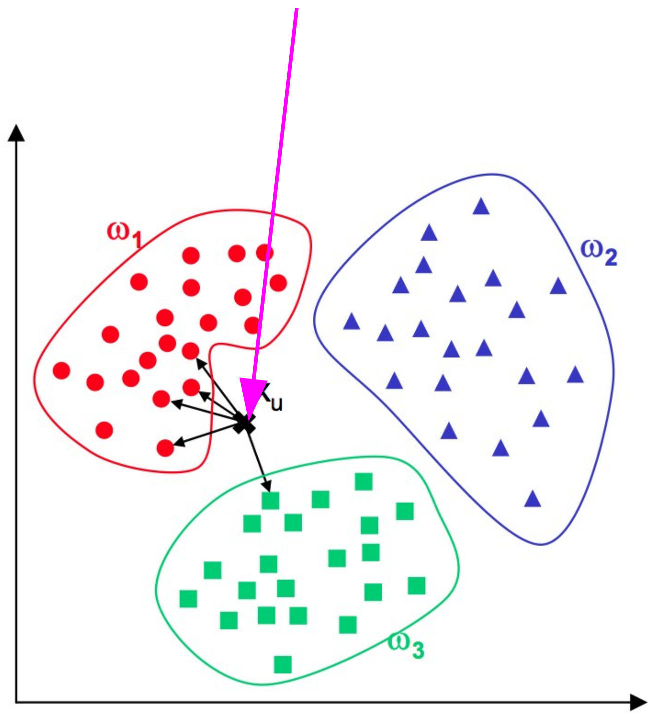
Parametric: Statistics Based Techniques

- Advantage
  - Utilize existing statistical modeling techniques to model various type of distributions
- Challenges
  - With high dimensions, difficult to estimate distributions
  - Parametric assumptions often do not hold for real data sets

[Chandola 09]

Kth Nearest Neighbors Distance – Uses distance metric to classify

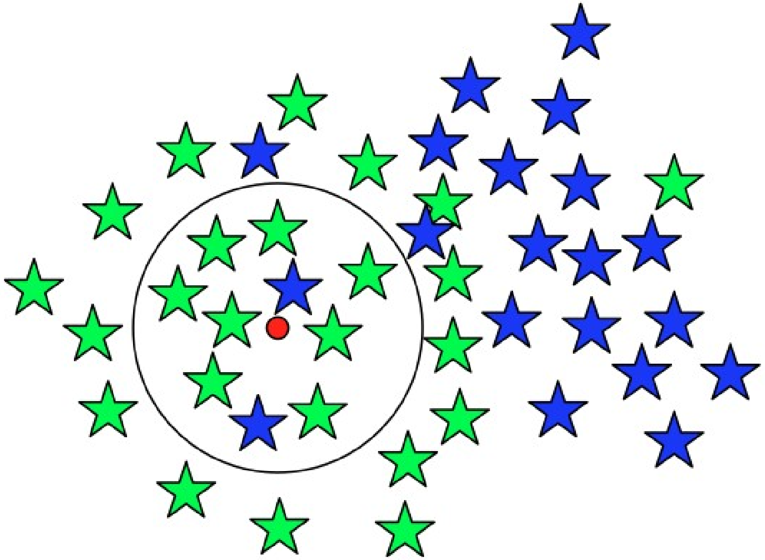
Classified as RED



Distance: Generally Euclidean Distance

Nearest Neighbor Density

- Estimate pdf of target function
- Frequentist notion



E.G. Local Outlier Factor

Characteristics of the kth NN classifier

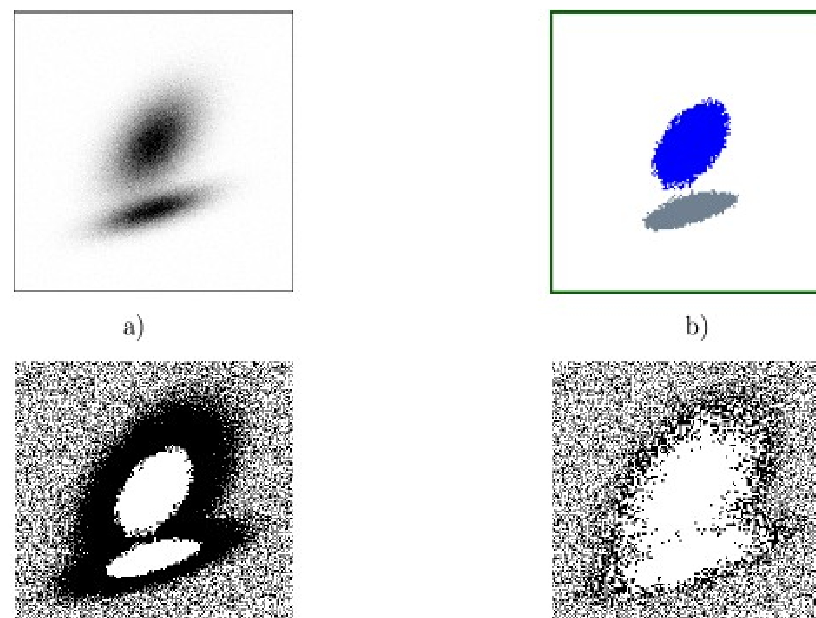
P(error)_Bayes < P(error)_1NN < 2p(error)_Bayes

- Advantages
  - Analytically tractable
  - Simple implementation
  - Nearly optimal in large sample limit
  - Uses local information → highly adaptive
  - Lends itself to parallel implementation
- Disadvantages
  - Large storage requirements
  - Computationally intensive recall
  - Highly susceptible to curse of dimensionality

[www.csc.kth.se/utbildning/]

Cluster Based: FindOut

- FindOut algorithm* by-product of *WaveCluster*
- Main idea: Remove the clusters from original data and then identify the outliers
- Transform data into multidimensional signals using wavelet transformation
  - High frequency of the signals correspond to regions where is the rapid change of distribution – boundaries of the clusters
  - Low frequency parts correspond to the regions where the data is concentrated
- Remove these high and low frequency parts and all remaining points will be outliers



* D. Yu, G. Sheikholeslami, A. Zhang, FindOut: Finding Outliers in Very Large Datasets, 1999.

* Outlier Detection – A Survey, Varun Chandola, Arindam Banerjee, and Vipin Kumar, Technical Report TR07-17, University of Minnesota

Clustering Based Techniques

- Advantages:
  - No need to be supervised
  - Easily adaptable to on-line / incremental mode suitable for anomaly detection from temporal data
- Drawbacks
  - Computationally expensive
    - Using indexing structures (k-d tree, R* tree) may alleviate this problem
  - If normal points do not create any clusters the techniques may fail
  - In high dimensional spaces, data is sparse and distances between any two data records may become quite similar.
    - Clustering algorithms may not give any meaningful clusters

Many Many Many More...

- Fuzzy Logic
- Genetic Algorithms
- Principle Component Analysis
- ARIMA
- EWMA
- HOLT-Winters
- FFT