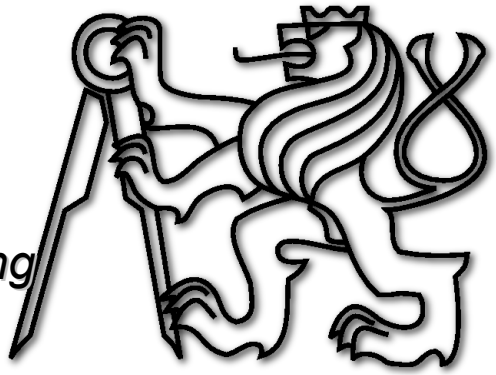
Anomaly detection

http://cig.felk.cvut.cz

Computational Intelligence Group
Department of Computer Science and Engineering
Faculty of Electrical Engineering
Czech Technical University in Prague



Anomaly detection

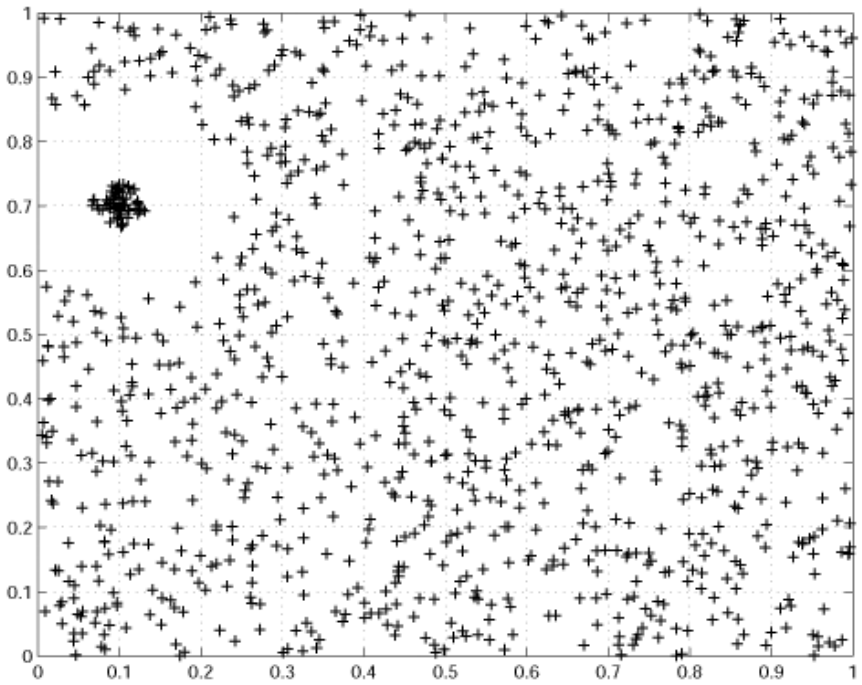
- Refers to a finding patterns in data which do not conform to expected behavior.
  - Generally referred as anomalies or outliers.
- Anomalous data usually means something unexpected is happening. And probably is going wrong :).

Usages of anomality detection

- Anomaly detection is commonly used in
  - Fraud detection (credit cards, healt care, insurance, …)
  - Computer system intrusion detection
  - Near fault conditions
  - Fault detection

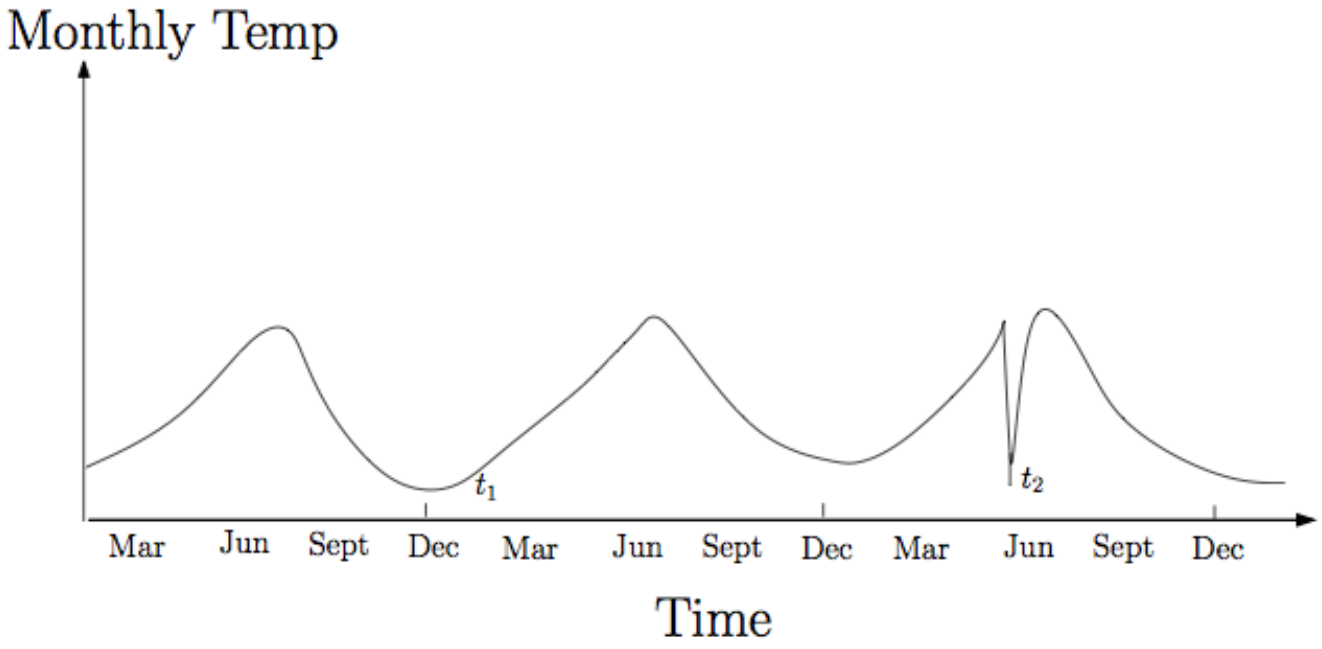
Types of anomaly

- Point anomalies
  - Individual instances, that are “far” from their normal positions.
  - Eg. Amount of money spend in one transaction, temperature of a device.



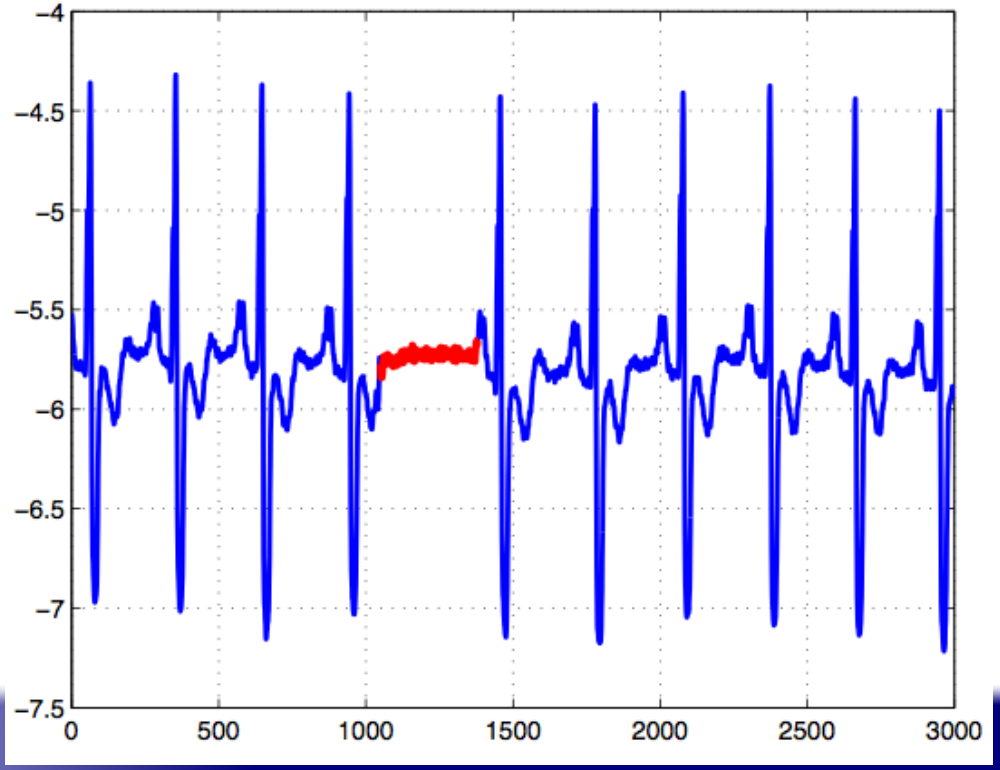
Types of anomaly

- Contextual anomalies
  - Data instances anomalous in given context. (But in other context they may be perfectly OK).
  - There have to be a context (ordering) defined in the data.



Types of anomaly

- Collective anomalies
  - A collection (subset) of related data is anomalous with respect to entire dataset.



Detection of anomalies

- Supervised anomaly detection
  - Classification problem => each instance have to be labeled as normal or anomalous.
  - **A classifier can distinguish between anomalous or normal data in given feature space.**
  - **+** Many possible modeling methods.
  - **-** Anomalous classes are very rare (otherwise they are not anomalies :)).
  - **-** Have to obtain correct labels for training data.
  - Generally not used.

- Anomalies classification uses (to us) well known methods.
  - Neural networks, SVM, Bayes Net, Decision trees,
- One-class classification
  - Special approach when classifier accepts or rejects the instance.
  - Methods define a boundary within which all instances should lie.
  - SVM, K-Means, Kernel Fisher Discriminants

Nearest neighbor based anomaly detection

- Normal data occur in dense neighborhood and anomalies are very sparse.
- There are two basic approaches
  - Compute distances of first k-nearest neighbors
  - Compute relative density of each data instance
    - LOF, COF algorithms
- Improvements of these approaches lies in speed-ups of computation time and better computation of relative density.

Clustering based AD

- Normal data forms clusters while anomalies do not belong into any cluster.
  - Clustering algorithms that do not force a cluster for any instance.
  - DBSCAN, ROCK, SNN
- Normal data instances lie near centroids (cluster representatives) but anomalies lie far from them.
  - Clustering algorithms utilizing reprezentatives.
  - KNN, SOM

Clustering based AD (2)

- Normal data belongs to large and dense clusters, while anomalies belongs to small and/or sparse clusters.
  - Some combination of LOF algorithm (Density of data instances) and clustering.
  - Eg. FindCBLOF algorithm.

Statistical approaches

- Compute probability of occurring of given instance. Normal instances are highly probable, anomalies are not.
  - Classical statistical approaches, like MIN-MAX range, variance of variable, inter-quantile range.
  - Statistical tests, regression based models.
  - And also non-parametric techniques like histograms.