Anomaly detection

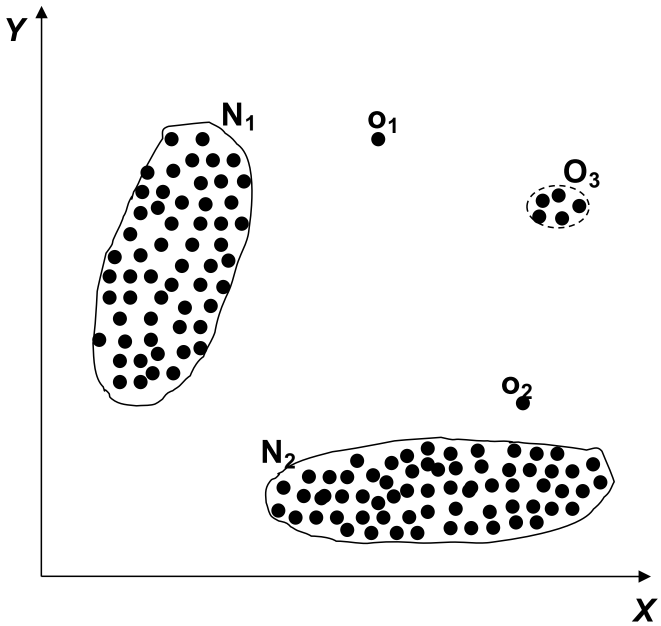
Today

- Types of anomalies and detection methods
- Detecting anomalies in:
  - sequences
  - multivariate data sets
  - multivariate sequences
- Evaluating anomaly detection
- Deep learning for anomaly detection

TU Delft

Point Anomalies

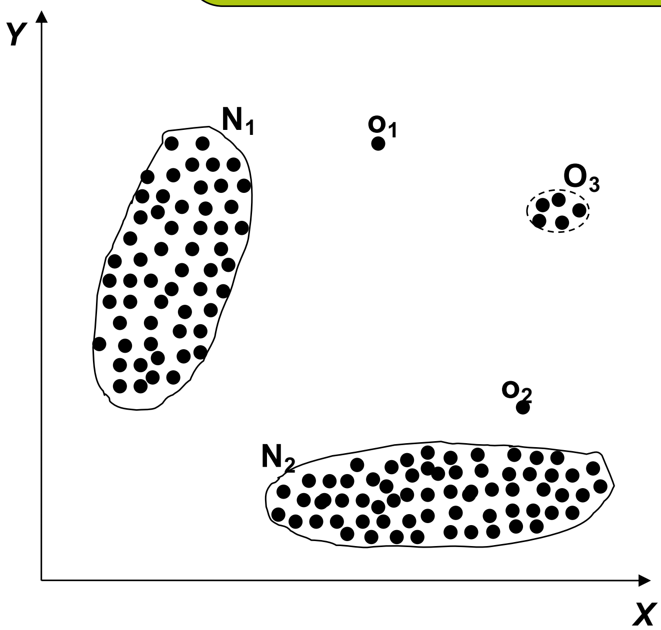
- An individual data instance is anomalous w.r.t. the data



Point Anomalies

- An individual data

Q: Which outliers are anomalies?



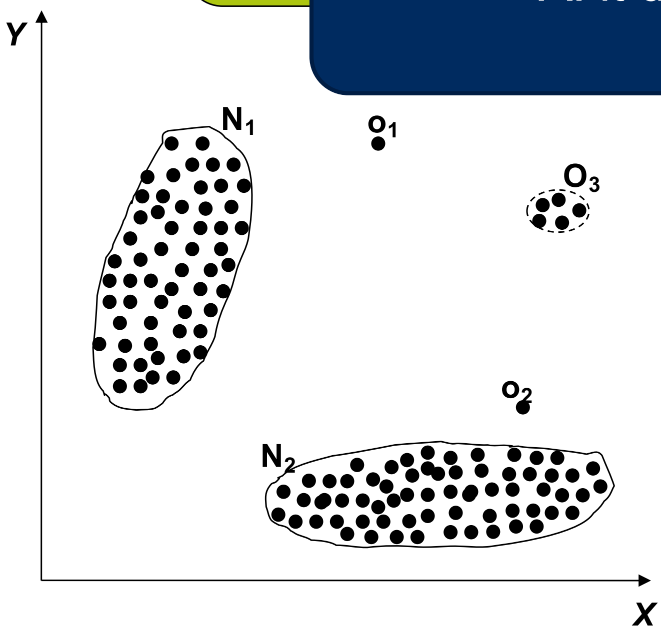
TU Delft

Point Anomalies

- An individual data

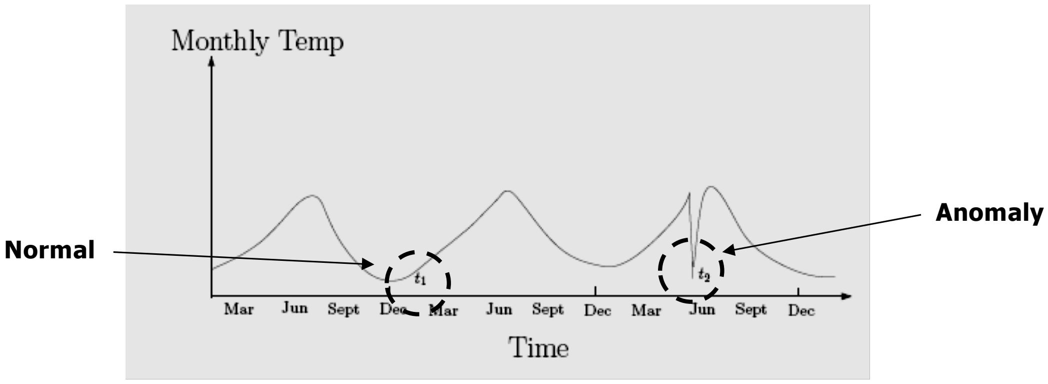
Q: V

A: it depends



Contextual Anomalies

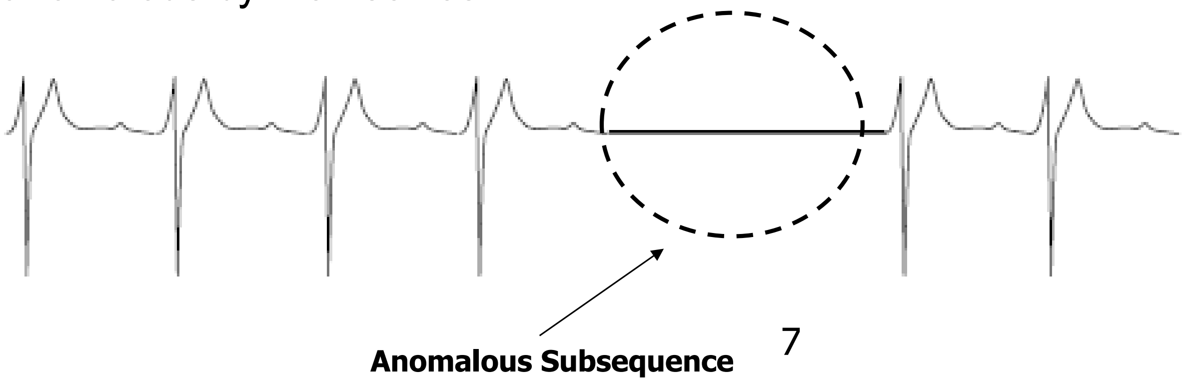
- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies*



* Xiuyao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka, Conditional Anomaly Detection, IEEE Transactions on Data and Knowledge Engineering, 2006.

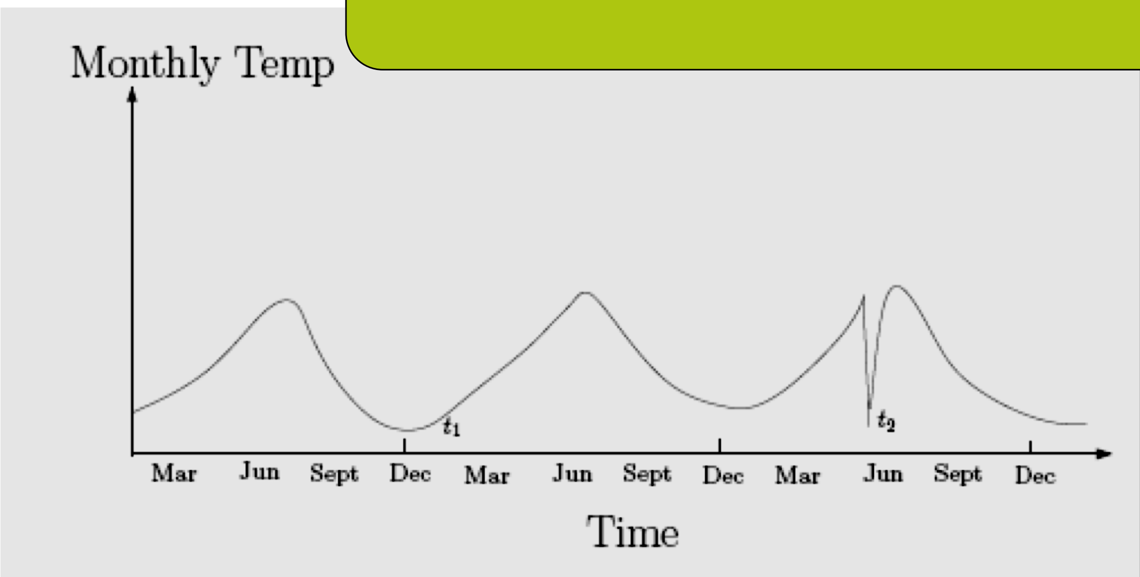
Collective Anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a collective anomaly are not anomalous by themselves



Anomalies in time series

Q: How to detect?



TU Delft

Compute the residual

- Use training data to learn a model:
- Use past test data to make a one-step prediction:

$$y_k = f(y_{k-1}) + \epsilon$$

- Compute the residual:

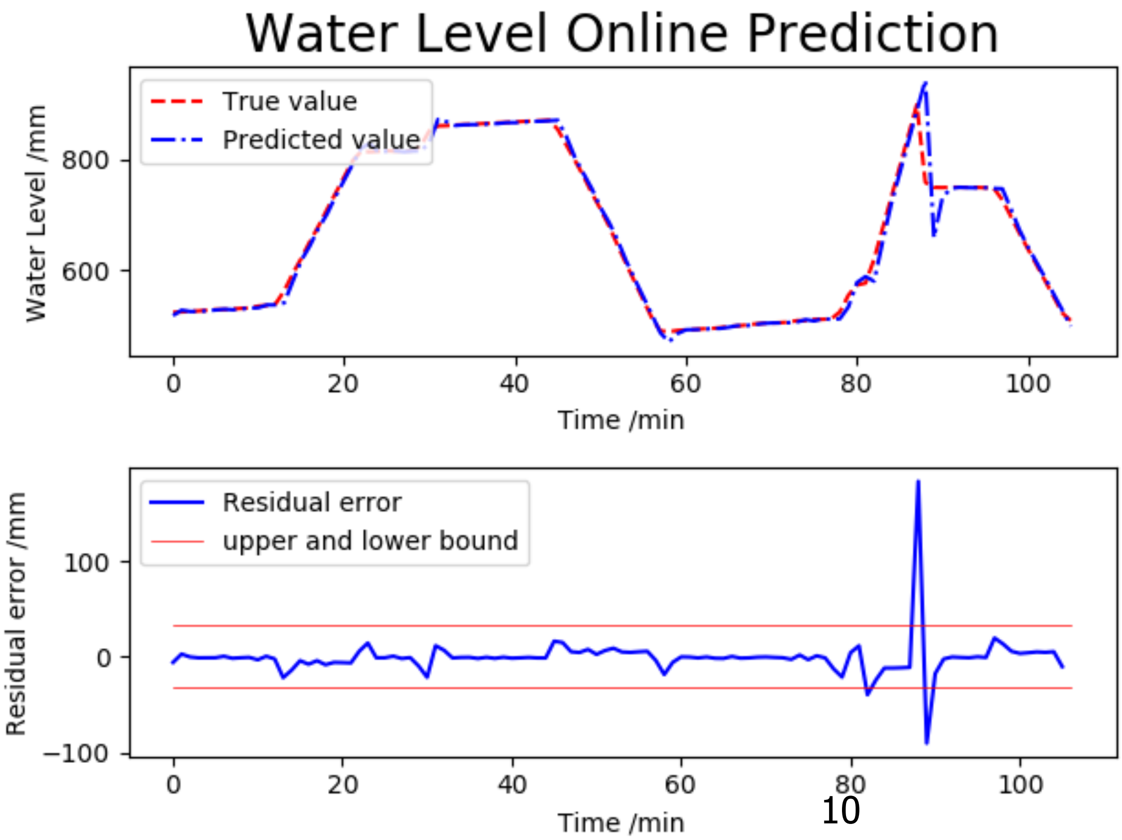
$$\hat{y}_{k|k-1} = f(y_{k-1})$$

- Evaluate the residual error through statistical test (depends on noise assumptions)

$$r_k = y_k - \hat{y}_{k|k-1}$$

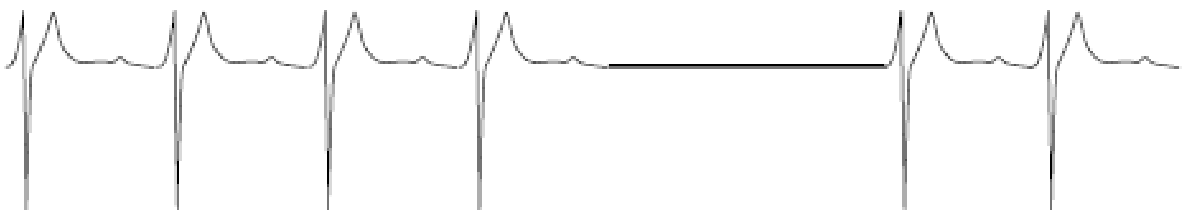
- Or simply using a decision threshold, typical:
  - 2 or 3 times the standard error
  - or simply sort on residual error and return largest ones
- Works with almost every time series model

An example (in SCADA)



Anomalies in time series

Q: How to detect?



Possibilities

1. Use sliding windows:
   - translate to standard point anomaly detection, or
   - compute distances using sequence alignment

1. Learn a sequential model (e.g. n-gram), and
   - compute the probability of observing a sequence
   - if below a threshold, raise an alarm

1. …

TU Delft

Today

- Types of anomalies and detection methods
- Detecting anomalies in:
  - sequences
  - **multivariate data sets**
  - multivariate sequences
- Evaluating anomaly detection
- Deep learning for anomaly detection

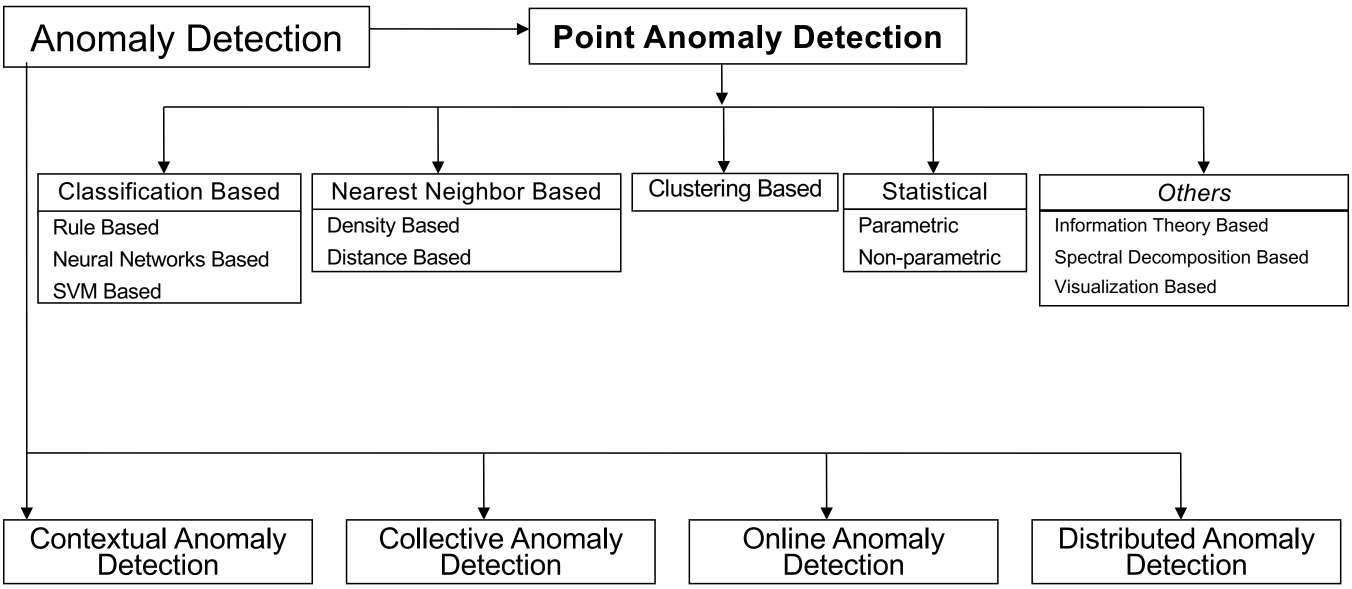
TU Delft

Anomaly detection, non-sequential

- A mess of possible methods:
  - Clustering
  - (One-class) Classification
  - Nearest Neighbors
  - Statistical
  - Spectral
  - ...
  - ***Key ingredient: assumption of what is an anomaly***

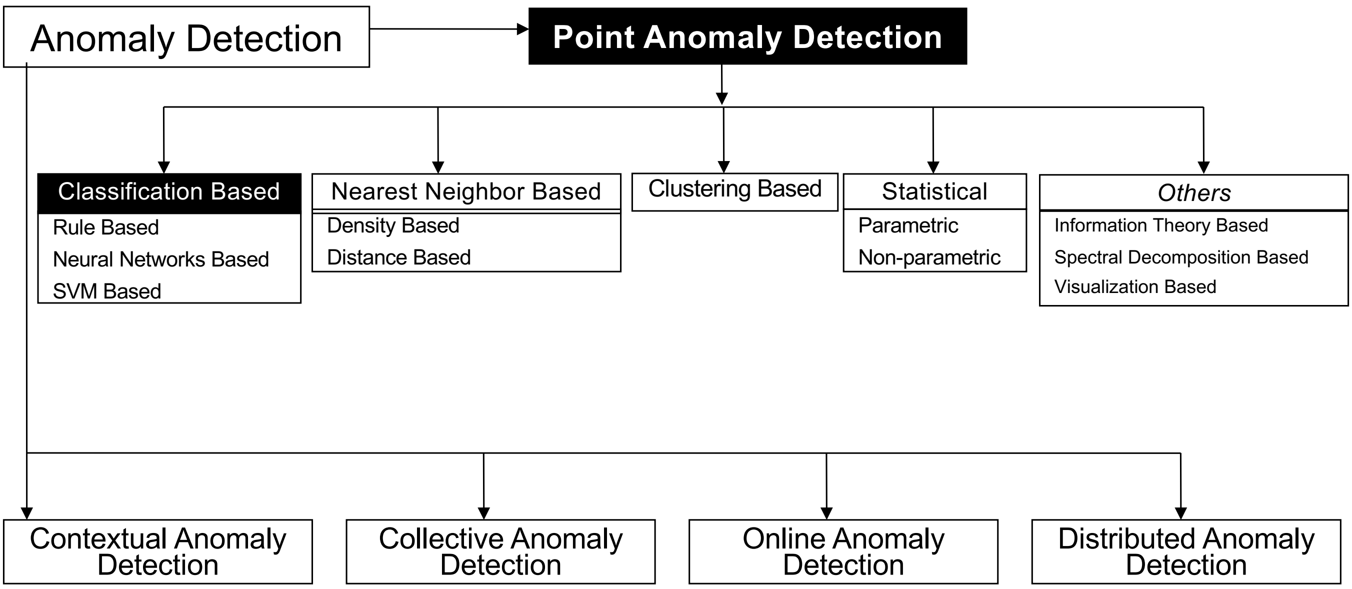
TU Delft

Taxonomy*



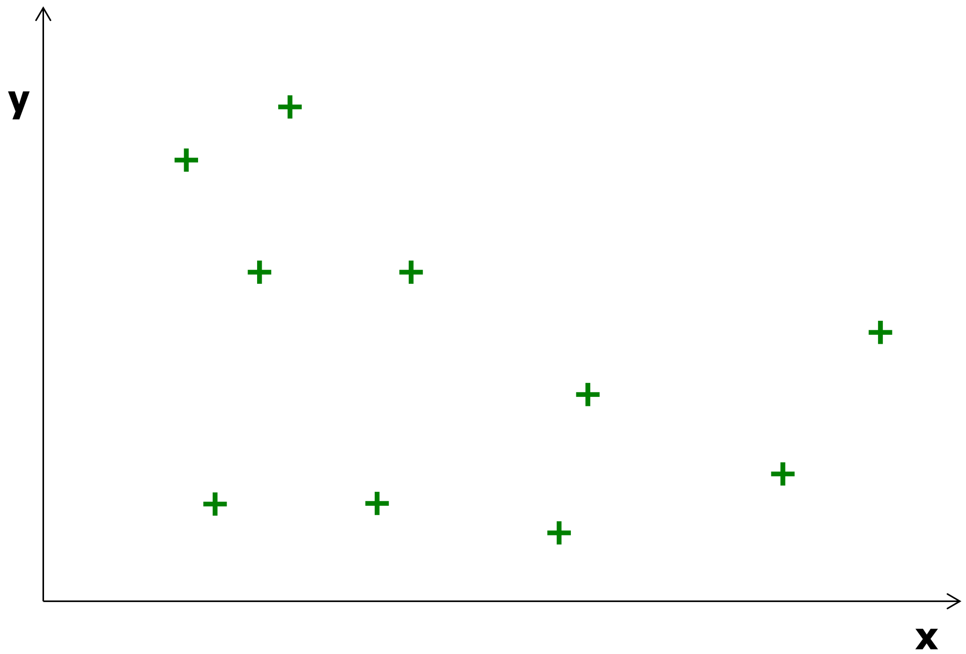
* Outlier Detection – A Survey, Varun Chandola, Arindam Banerjee, and Vipin Kumar

Taxonomy*

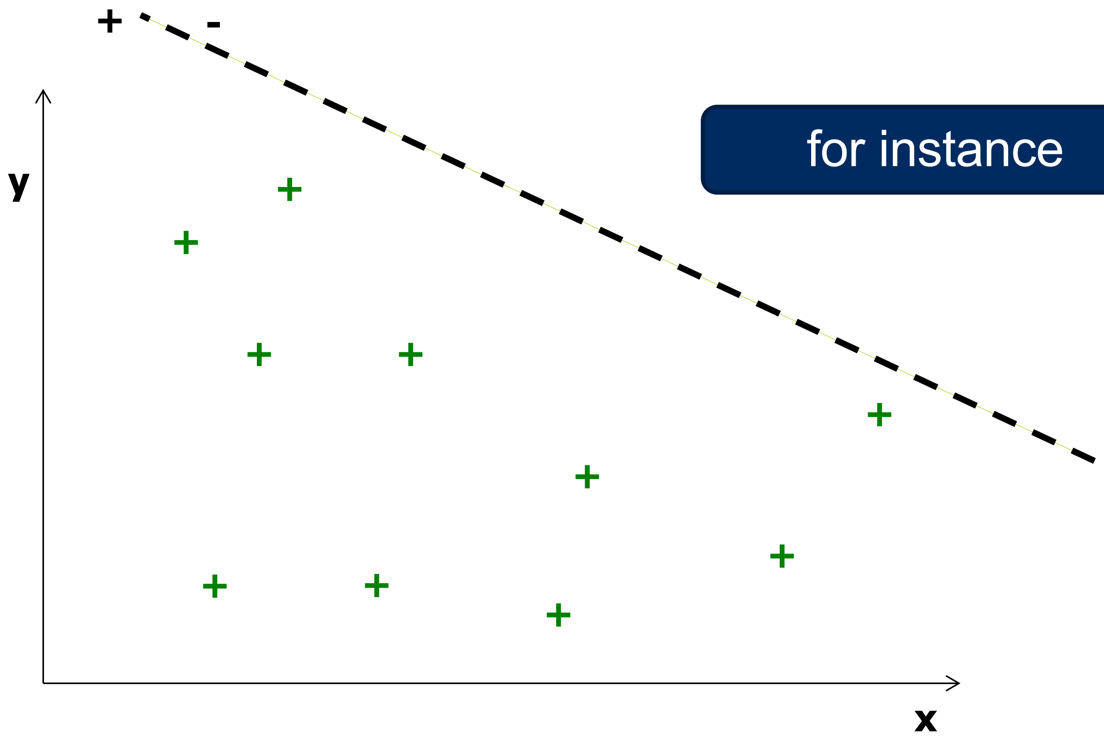


* Outlier Detection – A Survey, Varun Chandola, Arindam Banerjee, and Vipin Kumar

Where to put the decision boundary?

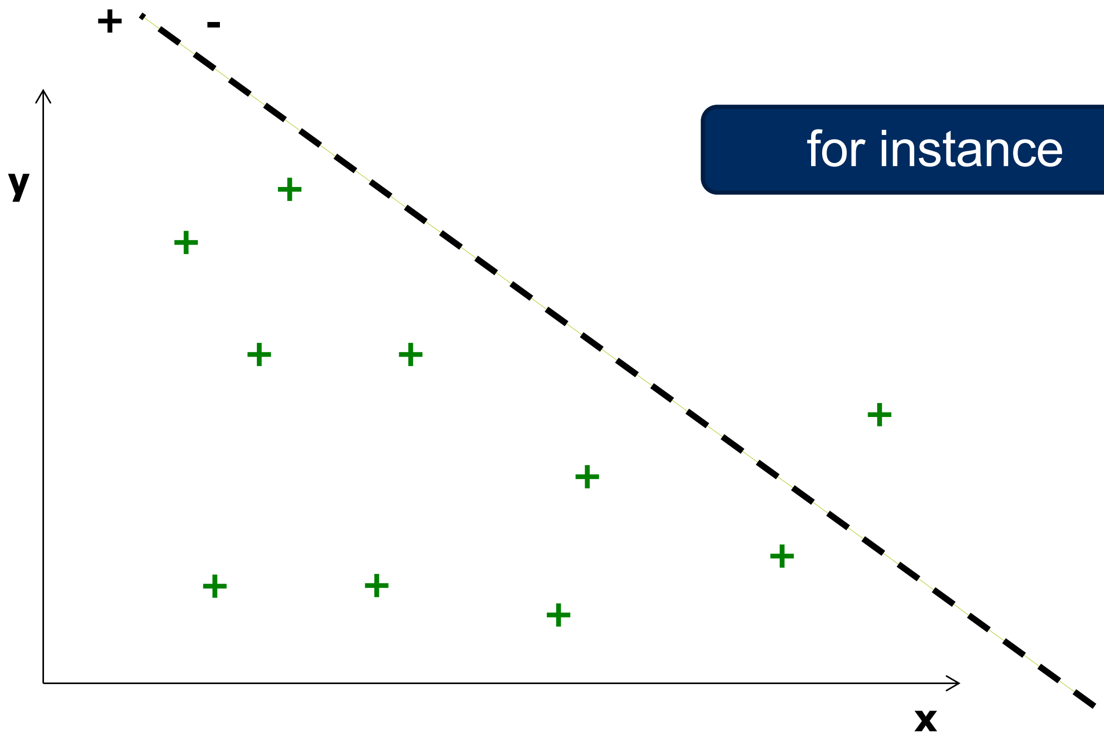


Where to put the decision boundary?



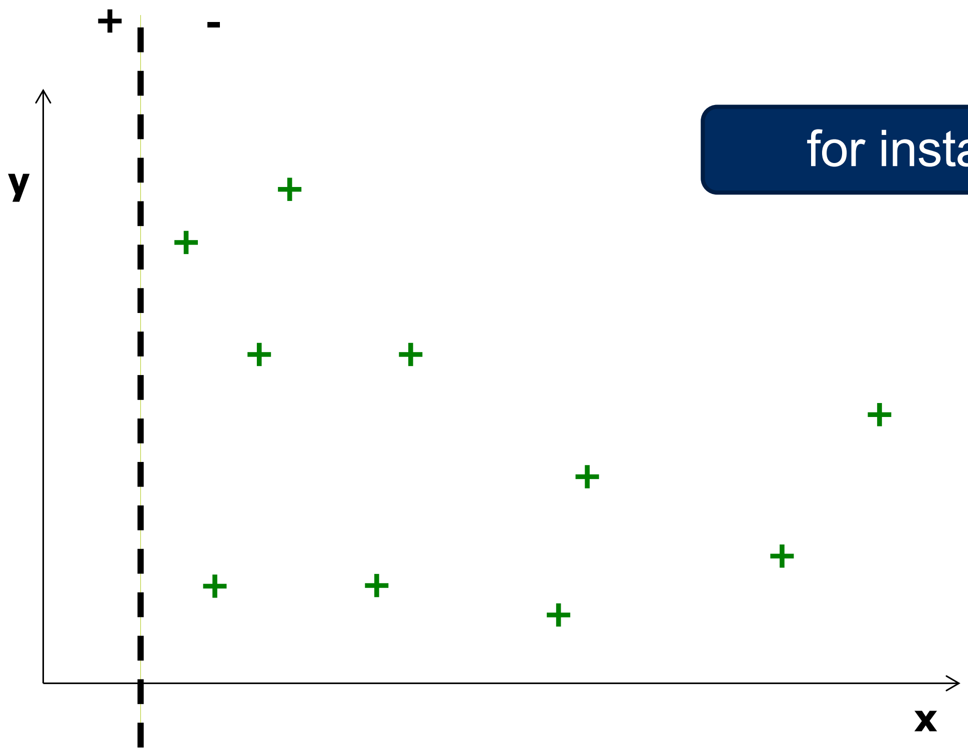
for instance

Where to put the decision boundary?



for instance

Where to put the decision boundary?



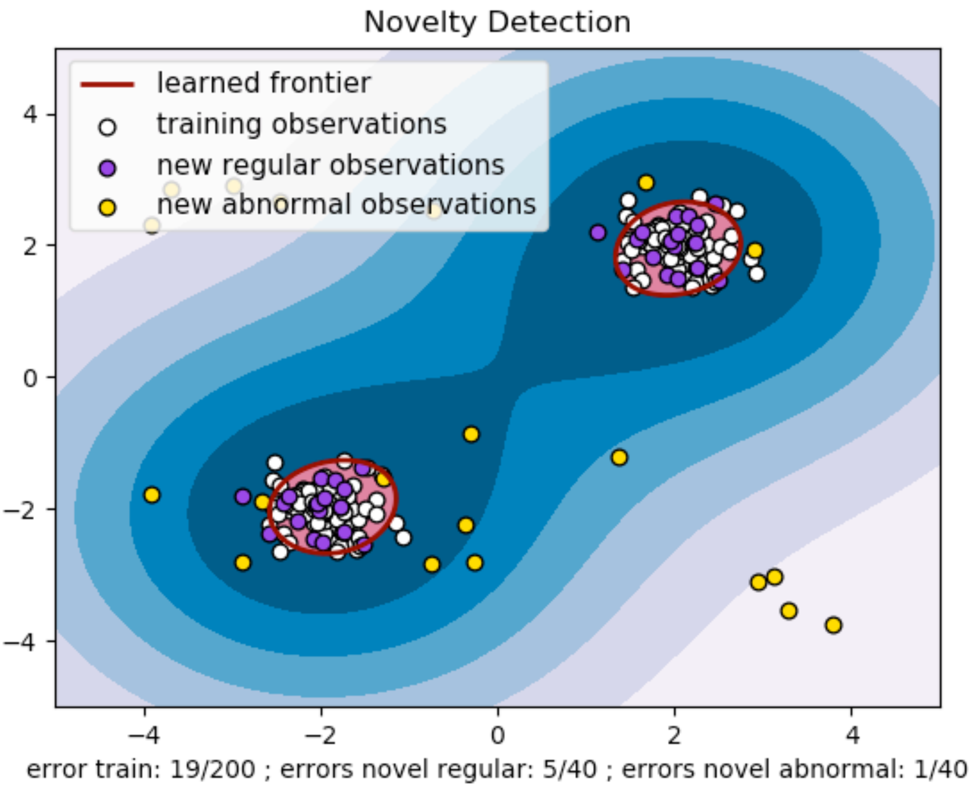
for instance

Classification based methods: OSVM

- Supervised → fraud detection exercise
- Only positive data:
  separate positive data from remaining input space
- Assumptions:
  - Further away from origin is anomalous (one-class SVM)
  - Close to the origin is anomalous (different one-class SVM)
  - Close to origin is also anomalous (improved one-class SVM)
  - Further from centroid is anomalous (non-linear one-class SVM)
  - …
- Key ingredient:
  - ***Maximize negative/outlier space***
  - ***Minimize positive/normal space***

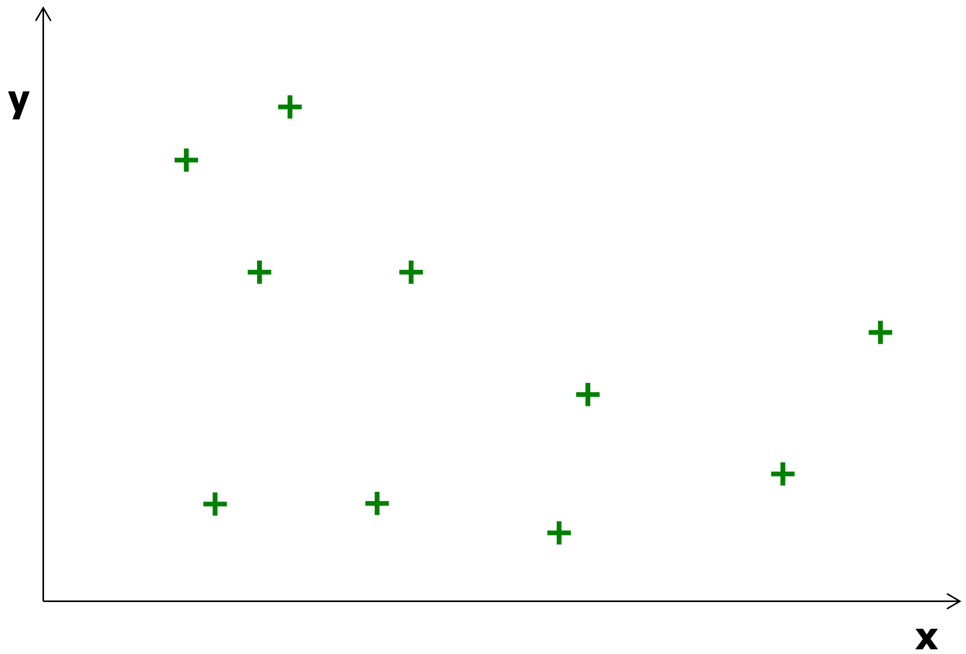
TU Delft

RBF one-class SVM



TUDelft

Different ways to use classification?



Other methods

- Add synthetic anomalous records
  - i.e. uniformly over a hypersphere or hypercube
- Learn a classifier from the data

- Predict X using Y, predict Y using X
  - (or more general predict X using everything but X, …)
- Label as anomalous when sufficient predictions are off

- Of course many studies use ensembles….

- Use models in innovative ways
  - e.g. Isolation Forests

TU Delft

Isolation Forest

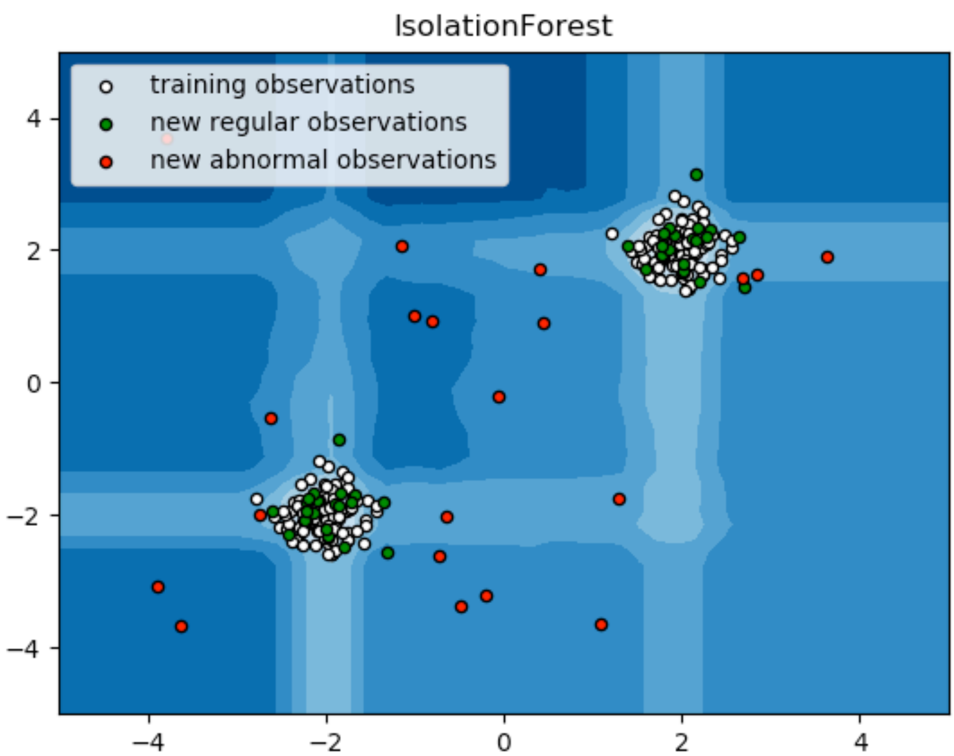
1. Repeat N times:
    1. Randomly pick a feature f
    2. Split the f uniformly at randomly between [min,max]
    3. Continue until all leafs contain singletons

- The path length to reach a leaf is the isolation score
- Average this length over all trees to get the anomaly score
- Intuition:

  isolating anomalies is easier because only a few conditions are needed to separate those cases from the normal observations

TU Delft

Isolation Forest



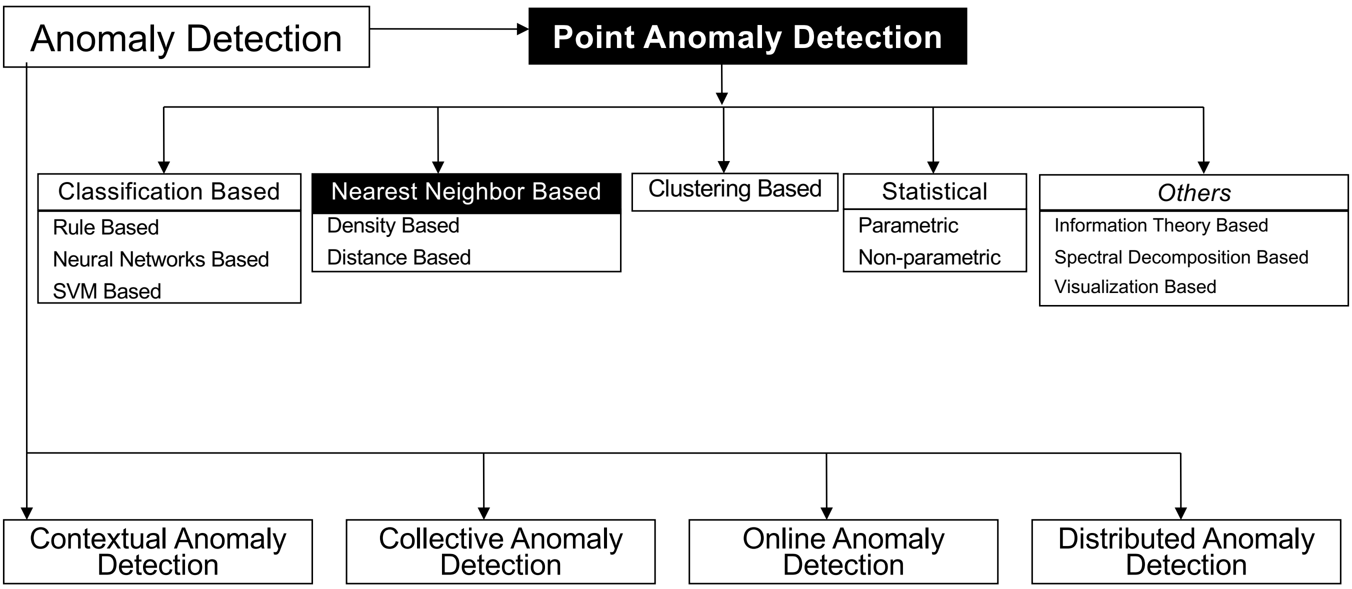
TUDelft

Classification Based Techniques

- Advantage
  - Can be used in unsupervised setting
  - Models can be (easily) understood
  - Computationally inexpensive when testing
- Drawback
  - Make assumptions about data distribution
    - Where is the origin? Is it normal or anomalous?
    - Intuitively less appealing

TU Delft

Taxonomy



Anomaly Detection
Point Anomaly Detection
Classification Based
Rule Based
Neural Networks Based
SVM Based
Nearest Neighbor Based
Density Based
Distance Based
Clustering Based
Statistical
Parametric
Non-parametric
Others
Information Theory Based
Spectral Decomposition Based
Visualization Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Nearest Neighbor Based Techniques

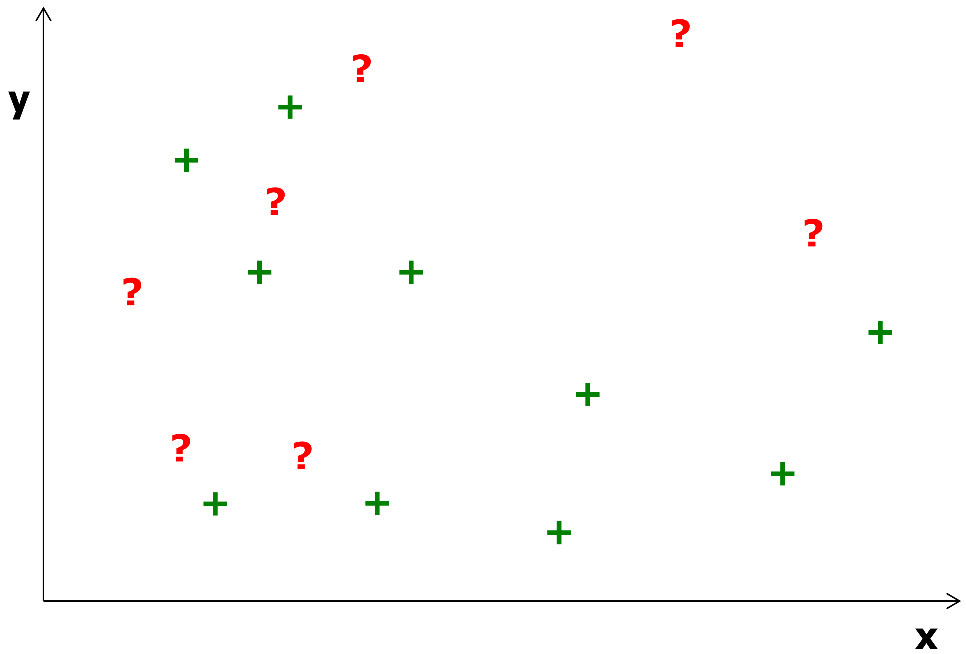
- Key assumption:
  normal points have close neighbors while anomalies are located far from other points

- Two-step approach

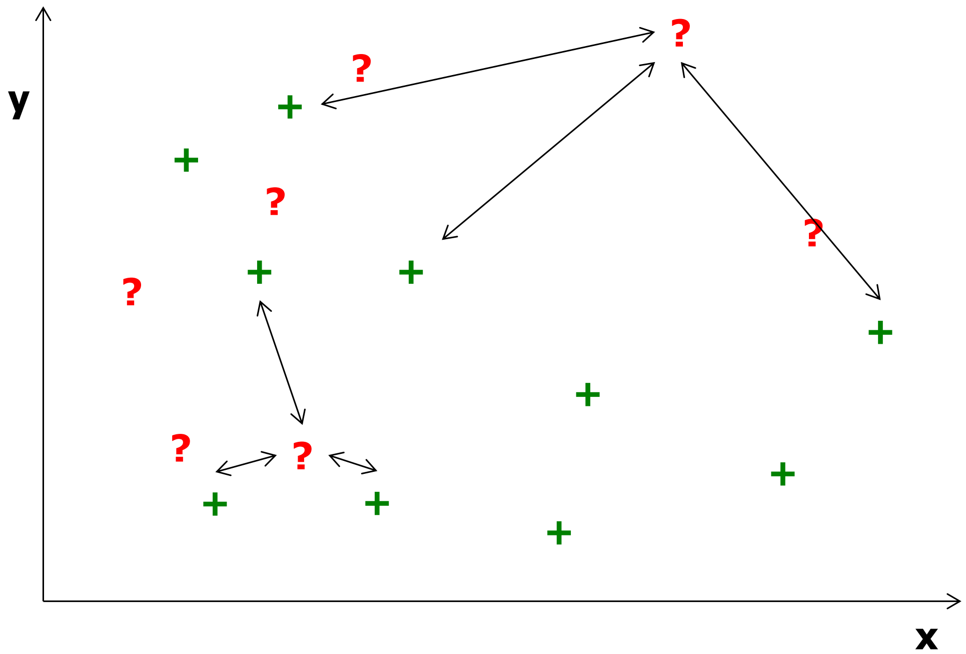
  1. Compute neighborhood for each data record
  2. Analyze the neighborhood to determine whether data record is anomaly or not

TU Delft

How to use neighbors?

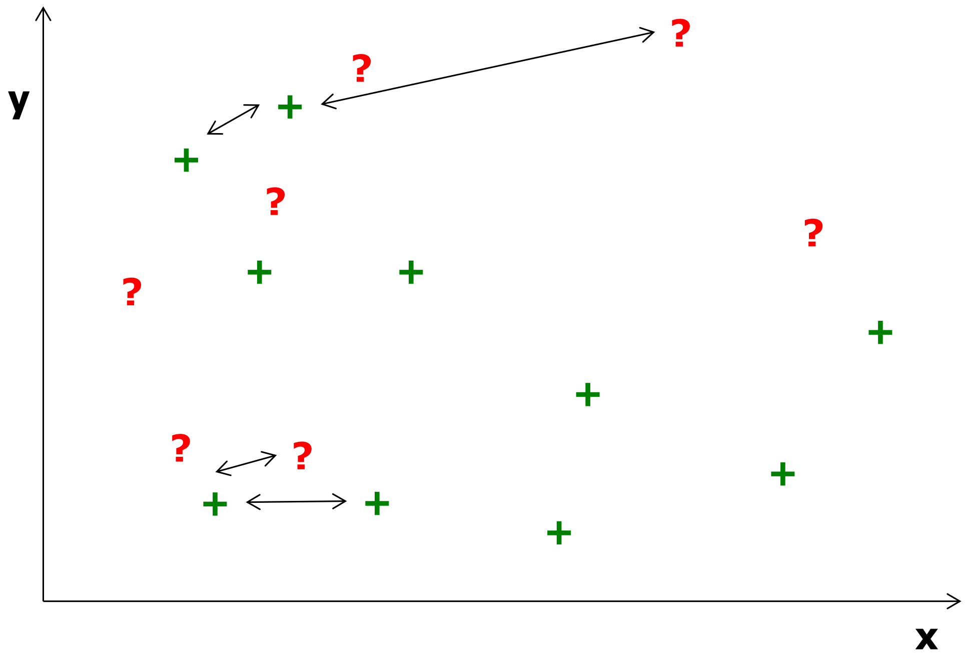


How to use neighbors? Distance



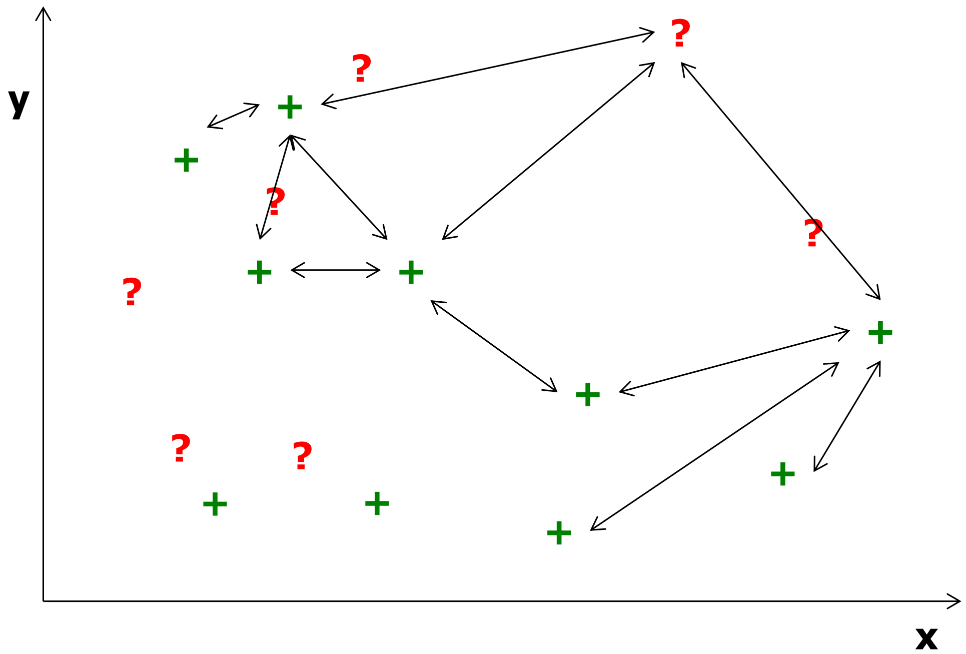
Anomaly if distance from other points is above threshold

How to use neighbors? Compare Distances



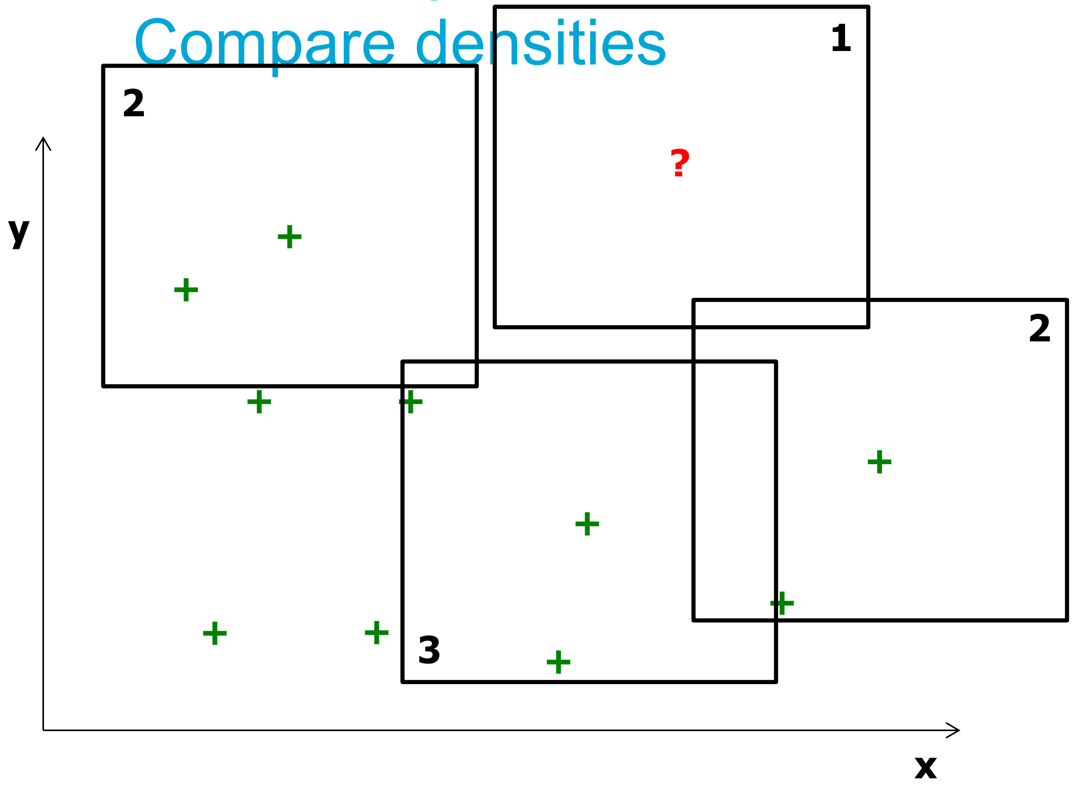
Anomaly if distance to nearest neighbor n compared to distance from n to nearest neighbor is above threshold

How to use neighbors?
Compare densities



Anomaly if density is substantially lower then neighbor's density, or average density

How to use neighbors?
Compare densities



Similar to density estimation using e.g., Parzen windows

Two key approaches

- Distance-based:
  - A point is anomalous when it is far from other points
- Density-based:
  - A point is anomalous when it is in a low density region

TU Delft

Distance based Outlier Detection

- Nearest Neighbor (NN) approach*,**
  - For each data point d compute the distance to the k-th nearest neighbor dk
  - Sort all data points according to the distance to dk
  - Outliers are points that have the largest distance dk and therefore are located in the more sparse neighborhoods
  - Usually data points that have top n% distance dk are identified as outliers
    - n – user parameter
  - Not suitable for datasets that have modes with varying density

TU Delft

- Knorr, Ng,Algorithms for Mining Distance-Based Outliers in Large Datasets, VLDB98

** S. Ramaswamy, R. Rastogi, S. Kyuseok: Efficient Algorithms for Mining Outliers from Large Data Sets, ACM SIGMOD Conf. On Management of Data, 2000.

Density based Outlier Detection
Local Outlier Factor (LOF)*

- The local outlier factor is based on a concept of a local density
  - locality is given by k nearest neighbors
  - distance to k neighbors is used to estimate the density
- Points that have a substantially lower density than their neighbors are considered to be anomalies
- The local density is estimated by the typical distance at which a point can be "reached" from its neighbors.
  - The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters
  - See next slides for further details

TU Delft

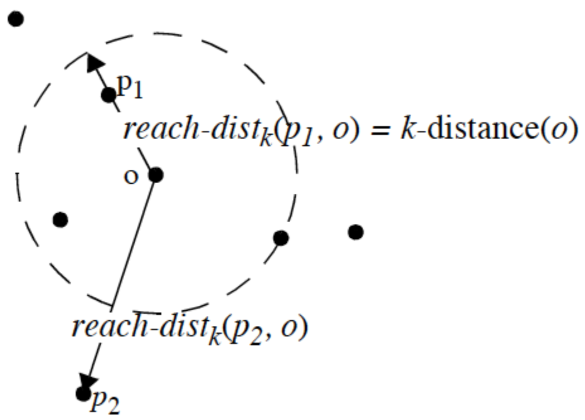
* - Breunig, et al, LOF: Identifying Density-Based Local Outliers, KDD 2000.

Density based Outlier Detection
Local Outlier Factor (LOF)*

- For each data point q compute the distance to the k-th nearest neighbor (k-distance(q))

- Compute reachability distance (reach-dist) for each data example q with respect to data example p as:

reach-dist(q, p) = max{k-distance(p), d(q,p)}



TU Delft

* - Breunig, et al, LOF: Identifying Density-Based Local Outliers, KDD 2000.

Density based Outlier Detection
Local Outlier Factor (LOF)*

- Compute local reachability density (lrd) of data example q as inverse of the average reachability distance based on the MinPts (k) nearest neighbors of data example q

$$\text{lrd}(q) = \frac{MinPts}{\sum_{p} reach\_dist_{MinPts}(q, p)}$$

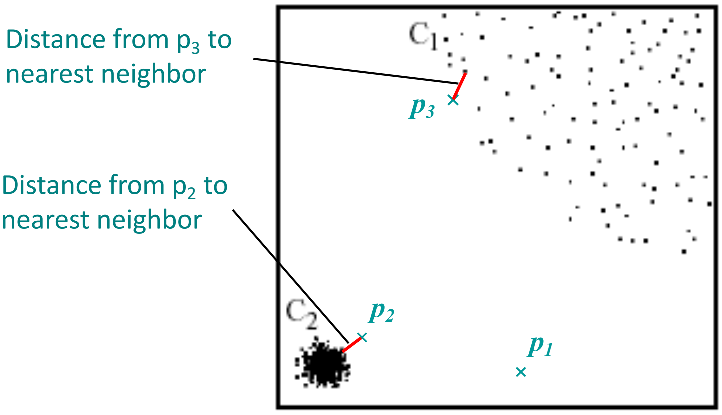
- Compute LOF(q) as ratio of average local reachability density of q's k-nearest neighbors and local reachability density of the data record q

$$\text{LOF}(q) = \frac{1}{MinPts} \cdot \sum_{p} \frac{lrd(p)}{lrd(q)}$$

* - Breunig, et al, LOF: Identifying Density-Based Local Outliers, KDD 2000.

Advantages of Density based Techniques

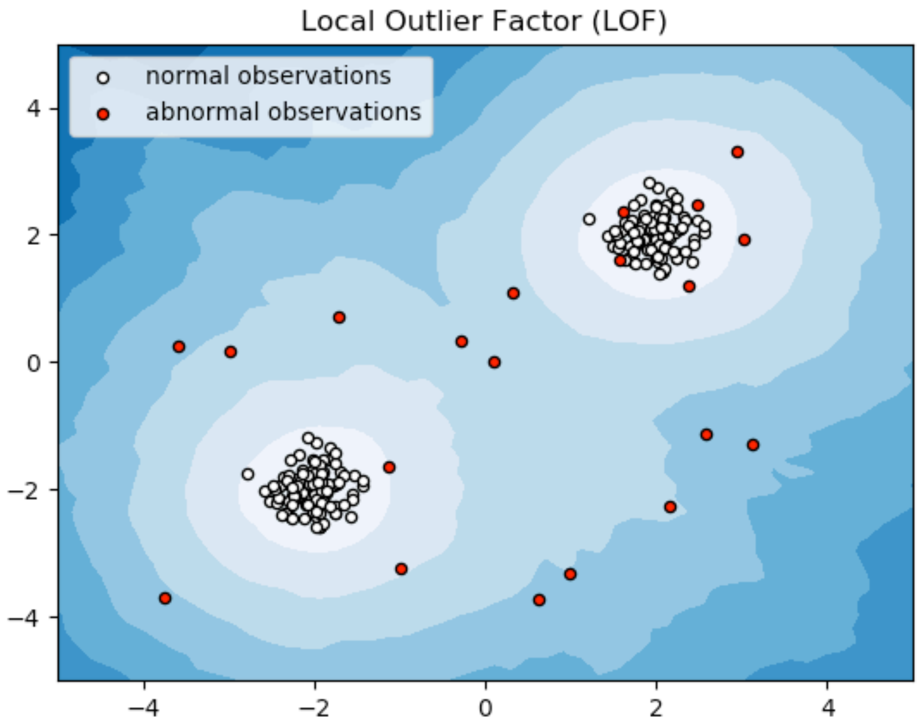
- Local Outlier Factor (LOF) approach
  - Example:



In the *NN* approach, p_2 is not considered as outlier, while the *LOF* approach find both p_1 and p_2 as outliers

NN approach may consider p_3 as outlier, but LOF approach does not

Local Outlier Factor



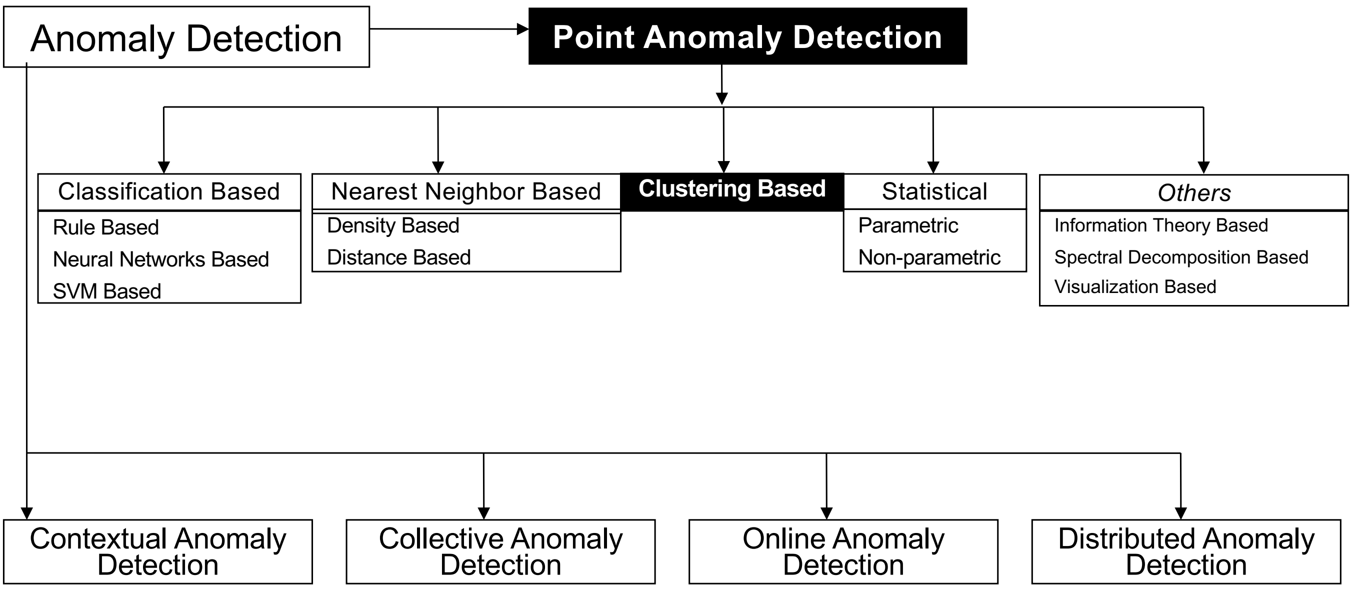
Nearest Neighbor Based Techniques

- Advantage
  - Can be used in unsupervised setting
  - Do not make any assumptions about data distribution
  - Intuitively appealing, uses distances
- Drawbacks
  - Computationally expensive (when testing)
  - Requires distances, may be unintuitive
  - In high dimensional spaces, data is sparse and the concept of similarity may not be meaningful anymore:

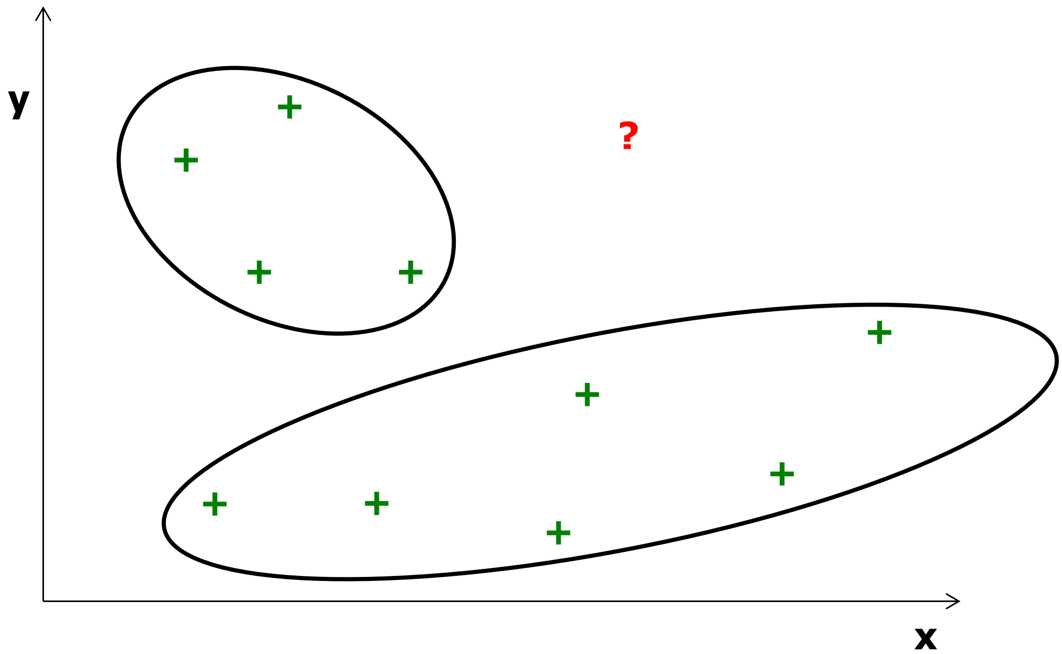
    Due to the sparseness, distances between any two data records may become quite similar
    => Each data record may be considered as potential outlier!

TU Delft

Taxonomy


Anomaly Detection
Point Anomaly Detection
Classification Based
Rule Based
Neural Networks Based
SVM Based
Nearest Neighbor Based
Density Based
Distance Based
Clustering Based
Statistical
Parametric
Non-parametric
Others
Information Theory Based
Spectral Decomposition Based
Visualization Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

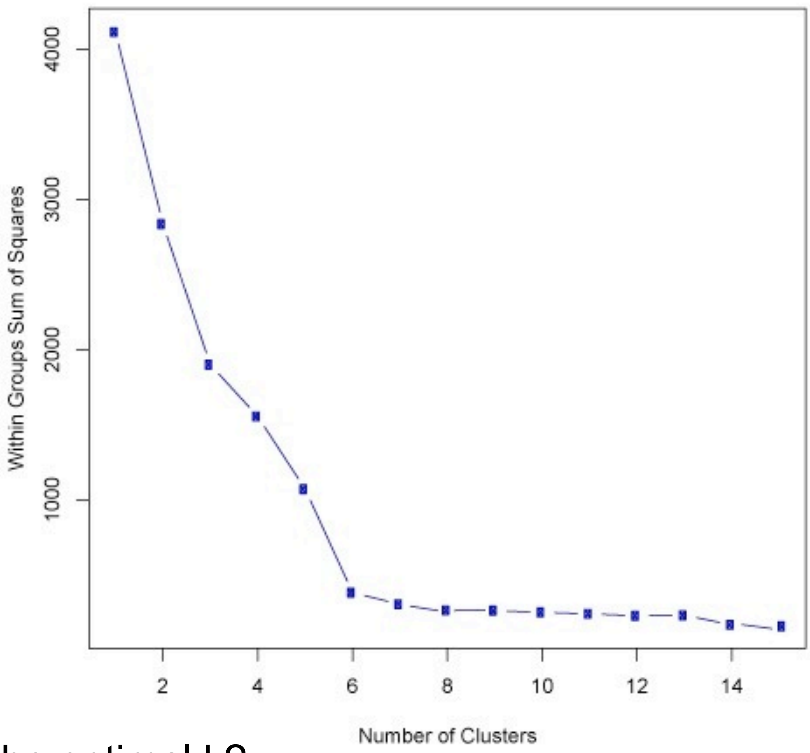
How to use clusters?



Clustering Based Techniques

- Key assumption:
  - *normal data records belong to large and dense clusters*
  - *anomalies do not belong to any cluster or form very small clusters*
- Local density using clustering:
  - Local anomalies are distant from other points within the same cluster

Deciding the number of clusters: ELBOW



- what is the optimal k?

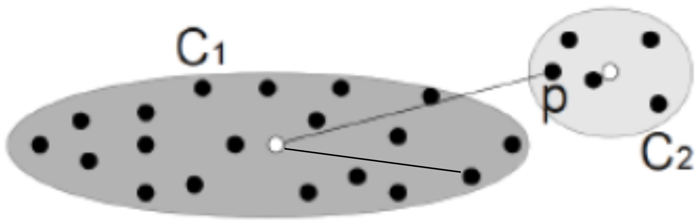
Clustering Based Techniques

- 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

TU Delft

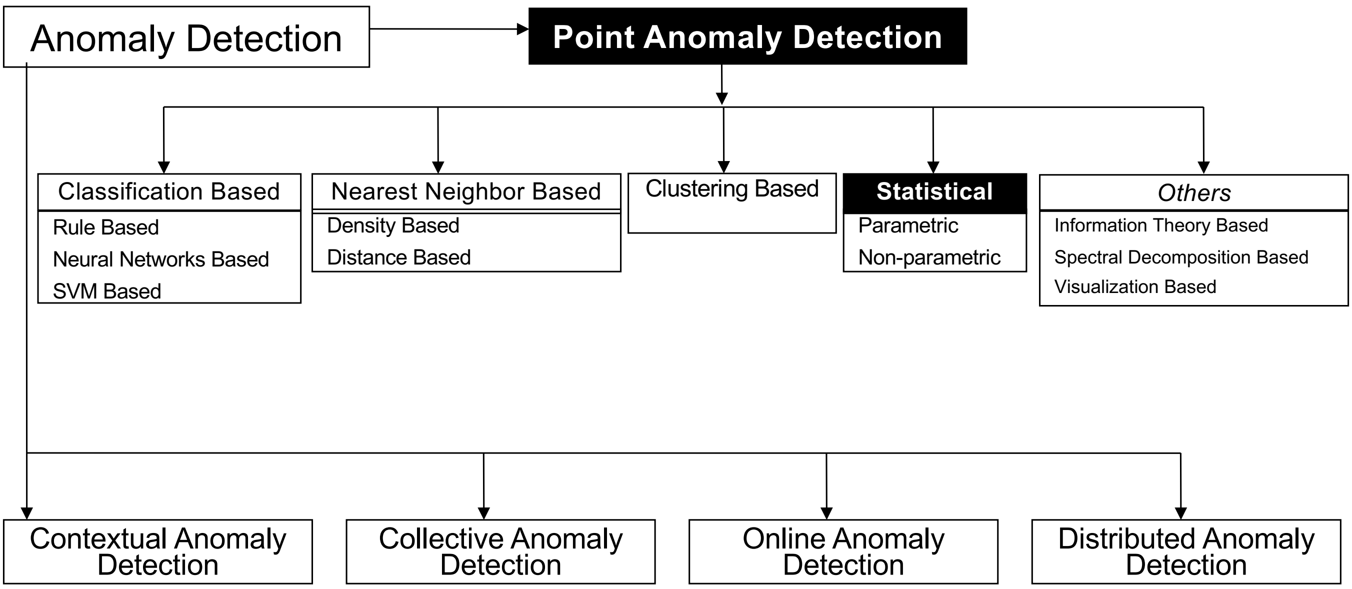
Cluster-based Local Outlier Factor (CBLOF)

- Determine CBLOF for each point using the cluster size and cluster distance:
    - if point is in a small cluster, CBLOF is the product of the cluster size and its distance to the closest larger cluster
    - if point is in a large cluster CBLOF is the product of the cluster size and the distance between the point and its own cluster



TU Delft

Taxonomy


Anomaly Detection
Point Anomaly Detection
Classification Based
Rule Based
Neural Networks Based
SVM Based
Nearest Neighbor Based
Density Based
Distance Based
Clustering Based
Statistical
Parametric
Non-parametric
Others
Information Theory Based
Spectral Decomposition Based
Visualization Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Today

- Types of anomalies and detection methods
- Detecting anomalies in:
  - sequences
  - multivariate data sets
  - **multivariate sequences**
- Evaluating anomaly detection
- Deep learning for anomaly detection

TU Delft

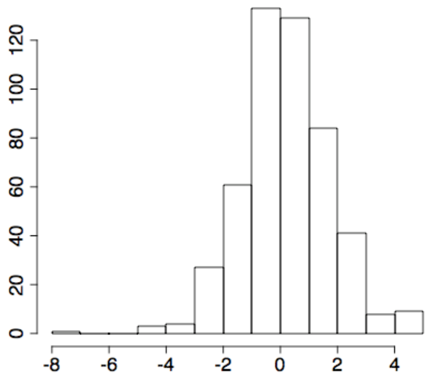
Statistics Based Techniques

- Data points are modeled using stochastic distribution
  - points are determined to be outliers depending on their relationship with this model
- 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

TU Delft

Hypothesis testing

$H_0 : \mu = 0$



$H_0 : \mu = 0$ null hypothesis

$H_1 : \mu > 0$ alternative hypothesis

Test statistic (t-student): $t = \frac{\overline{X}}{s}$

Reject H_0 if $t > c_\alpha$

for desired false negative rate α

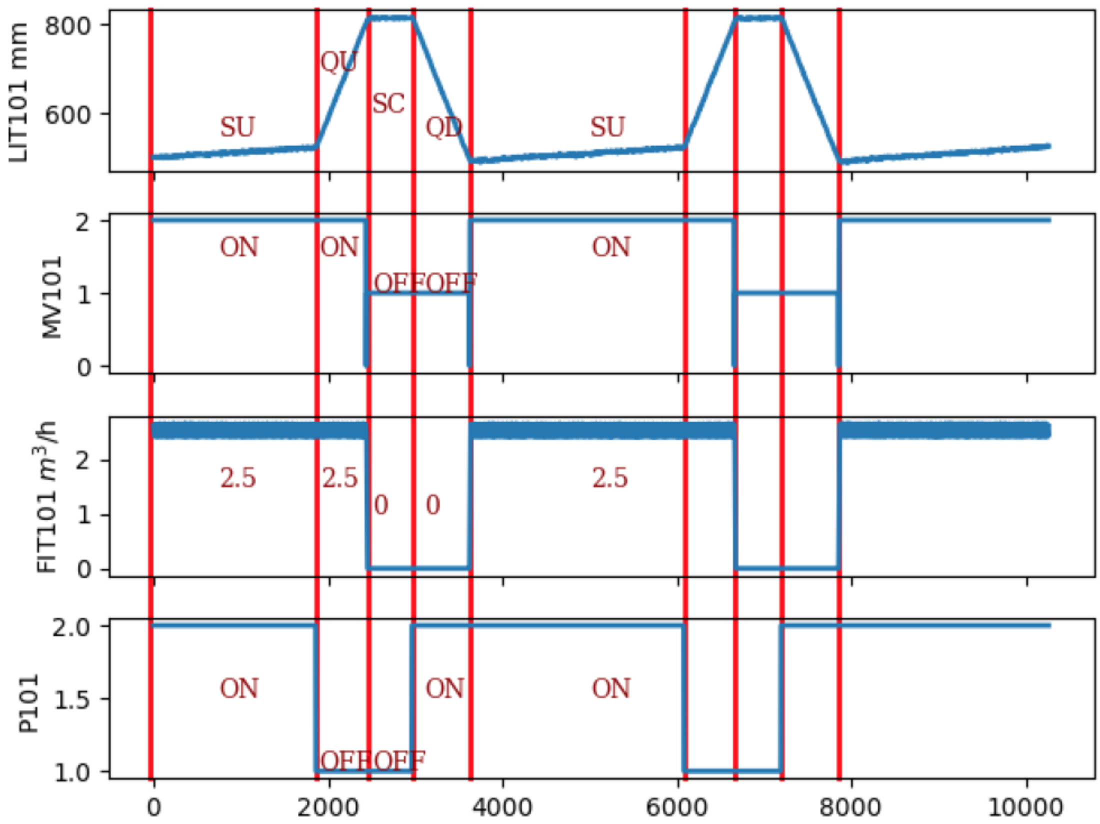
TU Delft

Types of Statistical Techniques

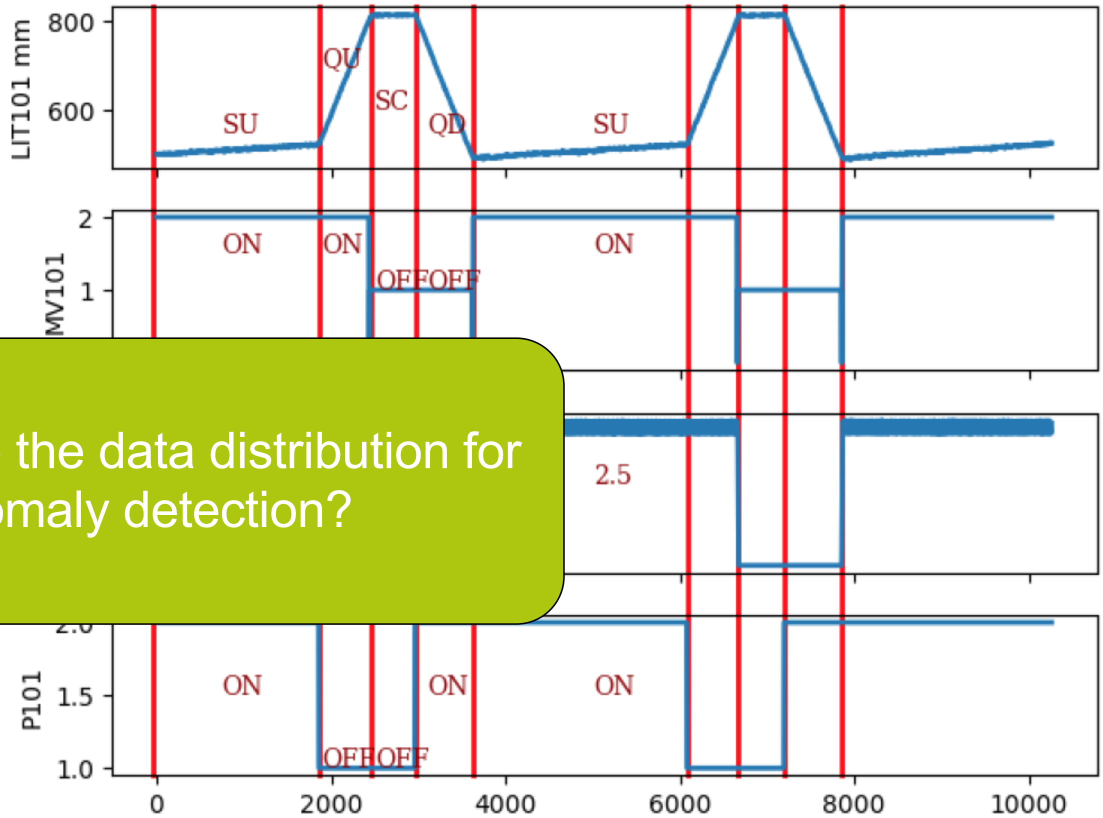
- Parametric Techniques
  - Assume that the normal (and possibly anomalous) data is generated from an underlying parametric distribution
  - Learn the parameters from the normal sample
  - Determine the likelihood of a test instance to be generated from this distribution to detect anomalies
- Non-parametric Techniques
  - Do not assume any knowledge of parameters
  - Use non-parametric techniques to learn a distribution
    - e.g. parzen window estimation

TU Delft

An example on SCADA signals



An example on SCADA signals



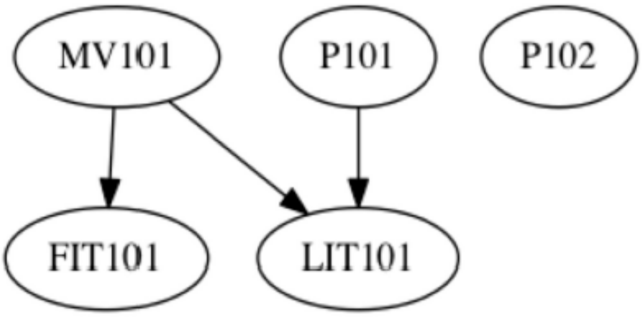
How to use the data distribution for anomaly detection?

Learn a Bayesian Network distribution

P(v=1)	P(v=2)
0.4890	0.5110

P(v=1)	P(v=2)
0.4890	0.5110

P(v=1)
1

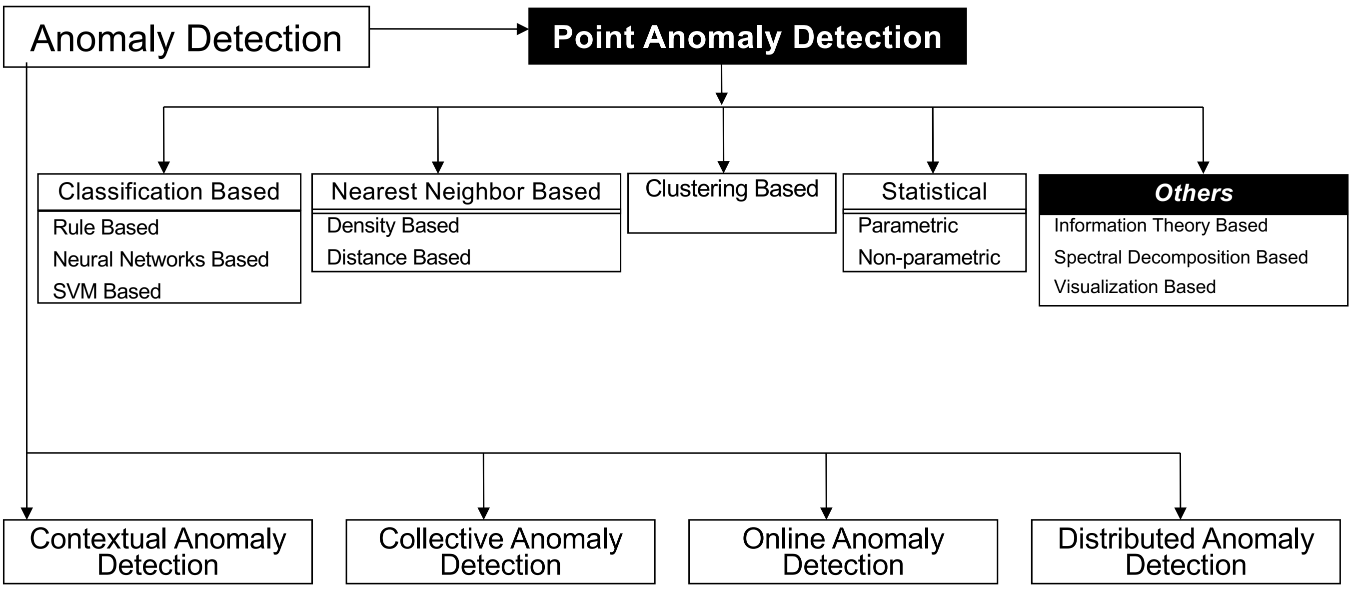


MV101	P(v=1)	P(v=2)
1	1	0
2	0	1

MV101 P01	P(v=1)	P(v=2)	P(v=3)	P(v=4)
1 1	0	1	0	0
2 1	0	0	0.0521	0.9479
1 2	0.9479	0.0521	0	0
2 2	0	0.0444	0.9556	0

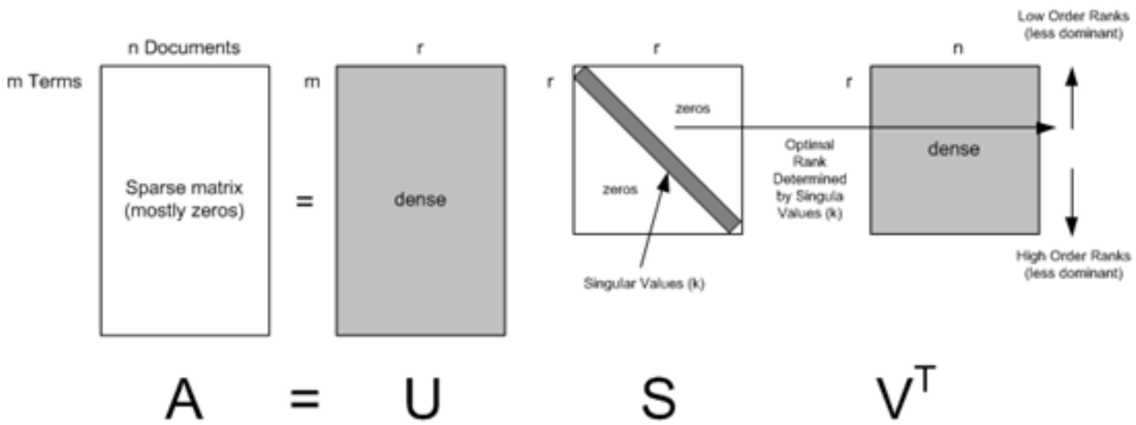
- Model conditional (in)dependencies between attributes

Taxonomy


Anomaly Detection
Point Anomaly Detection
Classification Based
Rule Based
Neural Networks Based
SVM Based
Nearest Neighbor Based
Density Based
Distance Based
Clustering Based
Statistical
Parametric
Non-parametric
Others
Information Theory Based
Spectral Decomposition Based
Visualization Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Spectral Techniques

- Analysis based on Eigen decomposition of data
- PCA (Principal Component Analysis)
  - Orthogonal transformation to reduce dimension
  - Most data patterns are captured by the several principal vectors

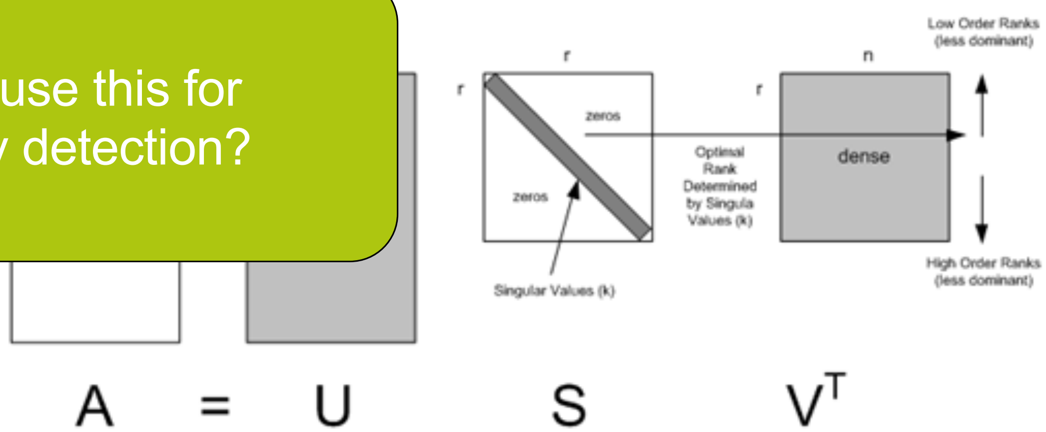


TU Delft

Spectral Techniques

- Analysis based on Eigen decomposition of data
- PCA (Principal Component Analysis)
  - Orthogonal transformation to reduce dimension
  - Most (linear) data patterns are captured by several principal vectors

How to use this for anomaly detection?



Spectral Techniques

- Key Idea
  - Find combination of attributes that capture bulk of variability
  - Reduced set of attributes can explain normal data well
  - But do not necessarily explain the outliers
- Several methods use Principal Component Analysis
  - Top few principal components capture variability in normal data
  - Smallest principal component should have constant values
  - Outliers have variability in the smallest component

TU Delft

PCA (Principal Component Analysis)

- Deriving principal vectors
  - Deriving the principal vector which captures the maximum variance

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \operatorname{Var}\{\mathbf{w}^T\mathbf{X}\} = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\mathbf{X}\right)^2\right\}$$

  - Find next component

$$\hat{\mathbf{X}}_{k-1} = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{X}$$

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\hat{\mathbf{X}}_{k-1}\right)^2\right\}.$$

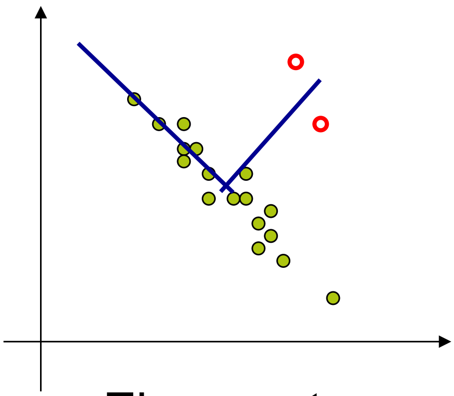
TU Delft

Example: Network traffic

Data matrix

$$\mathbf{Y} = \begin{bmatrix} \cdots \\ 100 \;\; 30 \;\; 42 \;\; 212 \;\; 1729 \;\; 13 \\ \cdots \end{bmatrix}$$

Perform PCA on matrix **Y**



Low-dimensional data

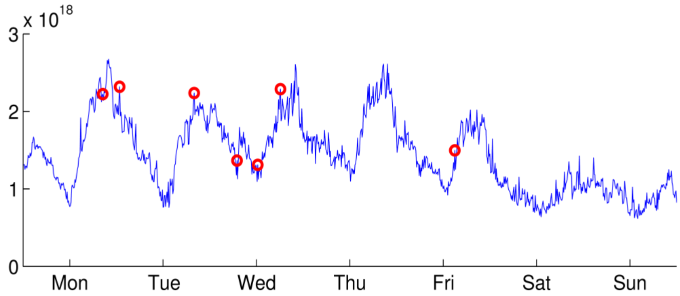
$$\mathbf{Yv} = \begin{bmatrix} \cdots \\ \mathbf{y}_t^T\mathbf{v}_1 \;\; \mathbf{y}_t^T\mathbf{v}_2 \\ \cdots \end{bmatrix}$$

Eigenvectors (components)

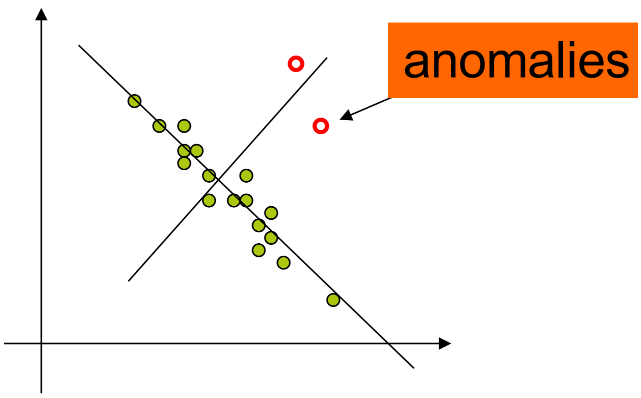
$$\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \ldots \end{bmatrix}$$

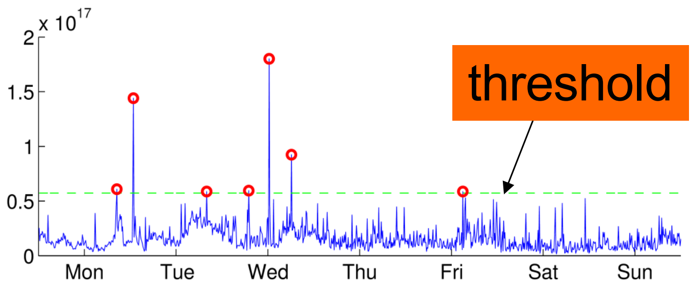
Example: Network traffic

Abilene backbone network traffic volume over 41 links collected over 4 weeks



Perform PCA on 41-dim data
Select top 5 components





Projection to residual subspac

$$\mathbf{y} = \hat{\mathbf{y}} + \tilde{\mathbf{y}} \qquad \hat{\mathbf{y}} = \mathbf{P}\mathbf{P}^T\mathbf{y} = \mathbf{C}\mathbf{y} \qquad \tilde{\mathbf{y}} = \left(\mathbf{I} - \mathbf{P}\mathbf{P}^T\right)\mathbf{y} = \tilde{\mathbf{C}}\mathbf{y}$$

$$P = [v_1 \ v_2 \ \ldots v_r]$$

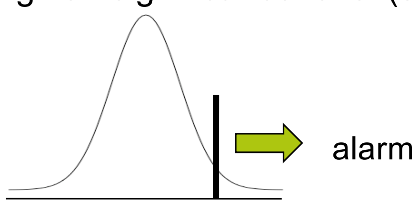
Using Robust PCA*

- Variability analysis based on robust PCA
  - Compute the principal components of the dataset
  - For each test point, compute its projection on these components
  - If yi denotes the ith component, then the following has a chi-squared distribution

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \ldots + \frac{y_q^2}{\lambda_q}, q \leq p$$

  - An observation is outlier if for a given significance level (statistical test)

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > \chi_q^2(\alpha)$$



- Have been applied to intrusion detection, outliers in space-craft components, etc.

Spectral Techniques

- Remember to first normalize your data, if your PCA method does not do it for you

- Advantage
  - Useful for multi-variate signals
  - Computationally efficient (use graphic cards!)

- Disadvantage
  - Based on the assumption that anomalies and normal instances are distinguishable in the reduced space
  - Does not take context into account
  - PCA is sensitive to outliers…

TU Delft

Today

- Types of anomalies and detection methods
- Detecting anomalies in:
  - sequences
  - multivariate data sets
  - multivariate sequences
- **Evaluating anomaly detection**
- Deep learning for anomaly detection

TU Delft

Anomaly detection is hard to evaluate

- Often little/no information on positives
  - Rely on quality of clustering, no clear quality measure exists
  - Good distances are often hard to find
- Anomalies are usually time periods instead of points
  - An attack starts and stops
  - Is every detection within that period a true positive?
- Unclear how to count positives
  - Many alarms are raised in a few seconds, is this a single positive?
  - Should we group them over time or per host/group?

TU Delft

Evaluation example

- Results from network anomaly detection paper (using LOF)

KDDcup99 dataset			
Normal	0.8902	0.9096	0.8998
R2L	0.7164	0.7283	0.7223
DoS	0.8733	0.8929	0.8829
Probe	0.8399	0.8550	0.8474
U2R	0.6092	0.6140	0.6115
Average	0.7858	0.7999	0.7928
NSL-KDD dataset			
Normal	0.9186	0.9314	0.9249
R2L	0.6897	0.7043	0.6969
DoS	0.8611	0.8752	0.8681
Probe	0.8549	0.8612	0.8580
U2R	0.6107	0.6231	0.6168
Average	0.7870	0.7990	0.7929

Normal	55119	5294	163	15	2	60593
R2L	4165	11791	14	216	3	16189
DoS	22412	17	205258	2164	2	229853
Probe	493	20	91	3562	0	4166
U2R	75	10	2	1	140	228
Total	82264	17132	205528	5958	147	311029
Normal	9045	480	124	56	6	9711
R2L	770	1939	0	43	1	2753
DoS	792	37	6529	94	8	7460
Probe	39	7	287	2085	3	2421
U2R	64	8	3	0	124	199
Total	10710	2471	6943	2278	142	22544

- Is this good or bad? It is useful?

TU Delft

Evaluation example

- Results from network anomaly detection paper (using LOF)

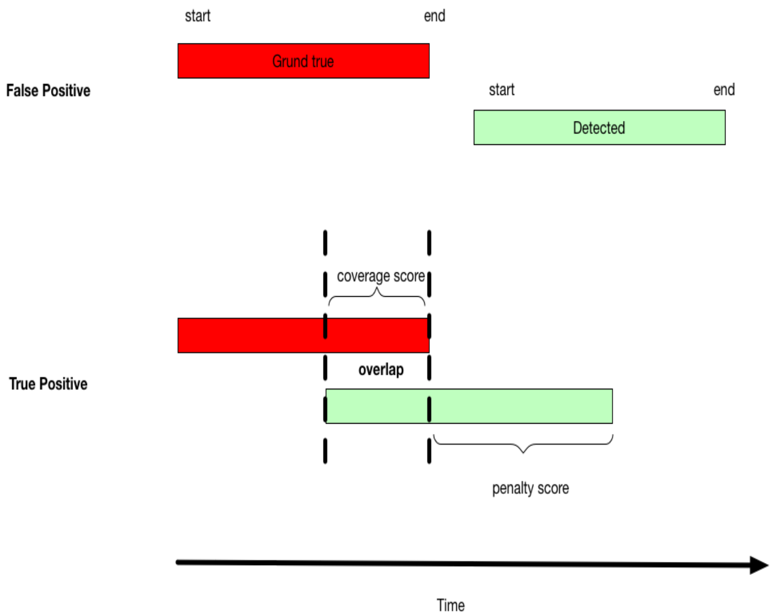
KDDcup99 dataset										
Normal	0.8902	0.9096	0.8998	Normal	55119	5294	163	15	2	60593
R2L	0.7164	0.7283	0.7223	R2L	4165	11791	14	216	3	16189
DoS	0.8733	0.8929	0.8829	DoS	22412	17	205258	2164	2	229853
Probe	0.8399	0.8550	0.8474	Probe	493	20	91	3562	0	4166
U2R	0.6092	0.6140	0.6115	U2R	75	10	2	1	140	228
Average	0.7858	0.7999	0.7928	Total	82264	17132	205528	5958	147	311029
NSL-KDD dataset										
Normal	0.9186	0.9314	0.9249	Normal	9045	480	124	56	6	9711
R2L	0.6897	0.7043	0.6969	R2L	770	1939	0	43	1	2753
DoS	0.8611	0.8752	0.8681	DoS	792	37	6529	94	8	7460
Probe	0.8549	0.8612	0.8580	Probe	39	7	287	2085	3	2421
U2R	0.6107	0.6231	0.6168	U2R	64	8	3	0	124	199
Average	0.7870	0.7990	0.7929	Total	10710	2471	6943	2278	142	22544

- Is this

How many hosts should be investigated?
Are they easy to investigate?
How quick is the detection?

Evaluation in SCADA systems

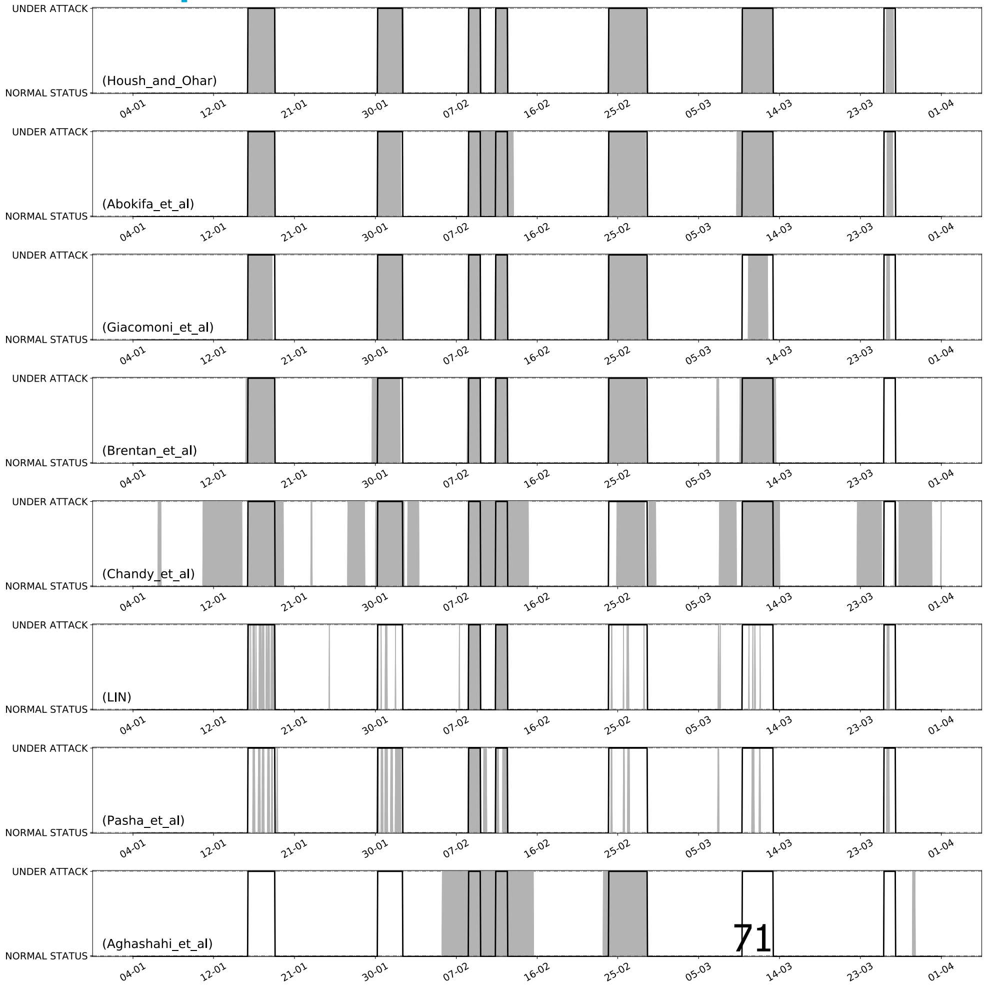
- Time series discord (unusual motif) [1]
- More meaningful than isolated points



[1] Senin P, Lin J, Wang X, Oates T, Gandhi S, Boedihardjo AP, Chen C, Frankenstein S. Time series anomaly discovery with grammar-based compression. In EDBT 2015 Mar 23 (pp. 481-492).

TU Delft

An attempt in BATADAL



TU Delft

An attempt in BATADAL

$$S_{TTD} = 1 - \frac{1}{n_a} \sum_{i}^{n_a} \frac{TTD_i}{\Delta t_i}$$

- TTD = time till detection
- Δti = duration of attack

$$S_{CM} = \frac{TPR + TNR}{2}$$

- point-based values

$$S = \gamma \cdot S_{TTD} + (1 - \gamma) \cdot S_{CM}$$

- gamma is parameter set by organizers

Take-away message

- Evaluating anomaly detection is hard
- You should evaluate it in a way that makes sense
    - Is host-detection important, or packet-based?
    - Is time untill detection relevant?
    - Is detecting a single attack a single positive?
    - …

TU Delft