CS659: Anomaly detection

Daniel Barbará

(Some slides adopted from a tutorial by Arindam Banerjee, Varun Chandola, Vipin Kumar, Jaideep Srivastava, and Aleksandar Lazarevic)

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Variants of Anomaly/Outlier Detection Problems
  - Given a database D, find all the data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
  - Given a database D, find all the data points $\mathbf{x} \in D$ having the top-n largest anomaly scores f($\mathbf{x}$)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D
- Applications:
  - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection

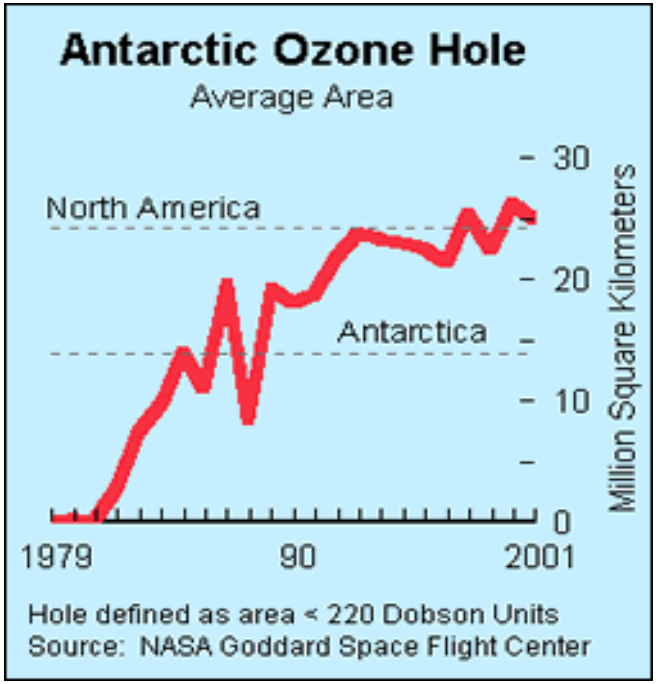
Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels

- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Sources:

http://exploringdata.cqu.edu.au/ozone.html

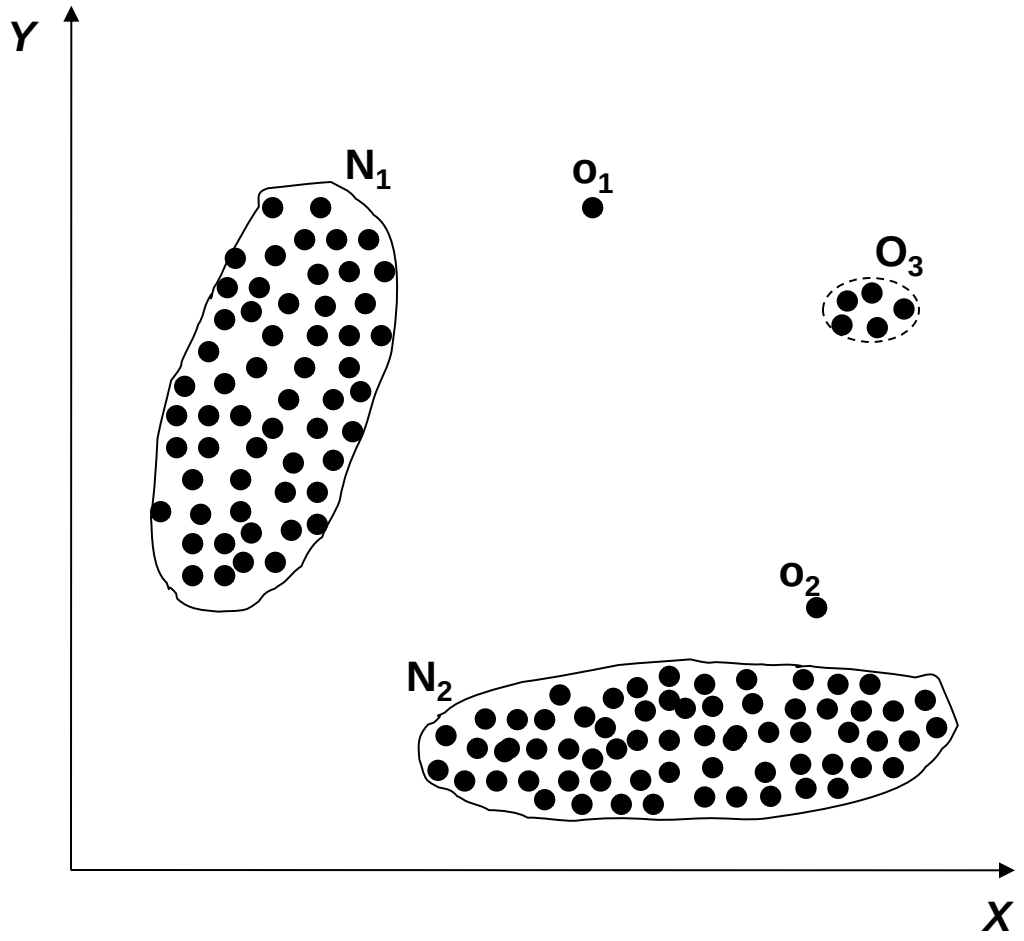
http://www.epa.gov/ozone/science/hole/size.html

Applications of Anomaly Detection

- Credit card fraud
  - Millions of transactions stored, only a small fraction are fraudulent
- Intrusion detection
  - Intrusions are a very small fraction of sessions
- Medical diagnosis
  - E.g., when analyzing an image a small fraction of pixels can be cancerous

Simple Example

- N_1 and N_2 are regions of normal behavior
- Points o_1 and o_2 are anomalies
- Points in region O_3 are anomalies



Related problems

- Rare Class Mining
- Chance discovery
- Novelty Detection
- Exception Mining
- Noise Removal
- Black Swan*

* N. Talleb, The Black Swan: The Impact of the Highly Probable?, 2007

Key Challenges

- Defining a representative normal region is challenging
- The boundary between normal and outlying behavior is often not precise
- The exact notion of an outlier is different for different application domains
- Availability of labeled data for training/validation
- Malicious adversaries
- Data might contain noise
- Normal behavior keeps evolving

Anomaly Detection

- Working assumption:
  - There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data

Data Labels

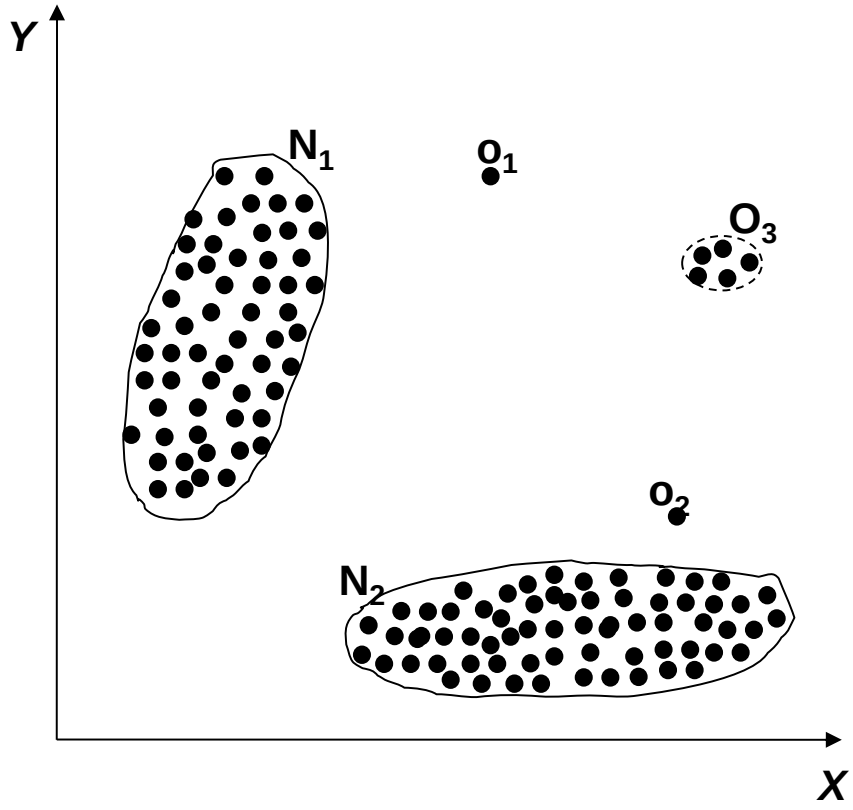
- Supervised Anomaly Detection (a bad idea…)
  - Labels available for both normal data and anomalies
  - Similar to rare class mining
- Semi-supervised Anomaly Detection
  - Labels available only for normal data
- Unsupervised Anomaly Detection
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data

Type of Anomaly

- Point Anomalies
- Contextual Anomalies
- Collective Anomalies

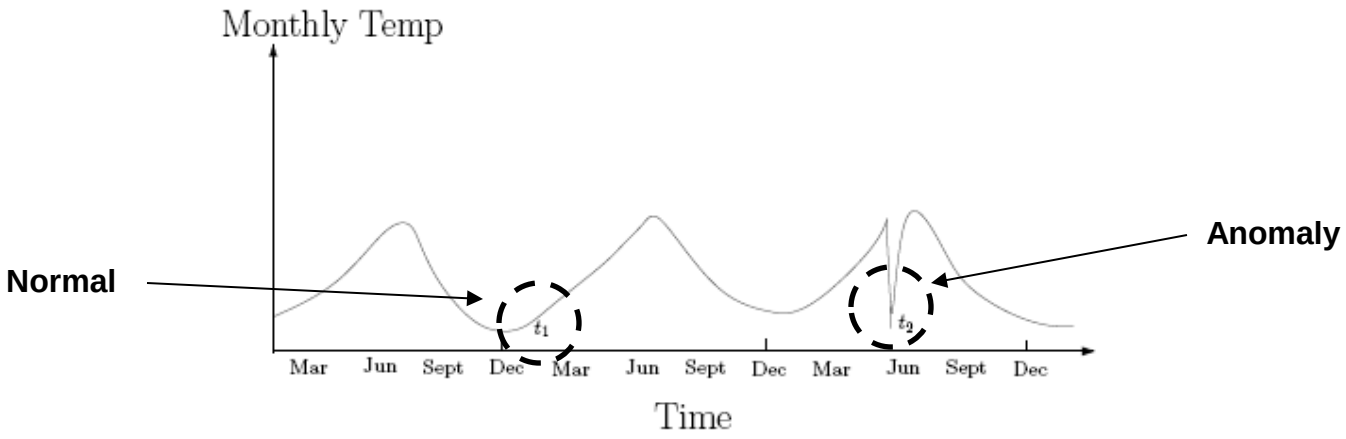
Point Anomalies

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



Contextual Anomalies

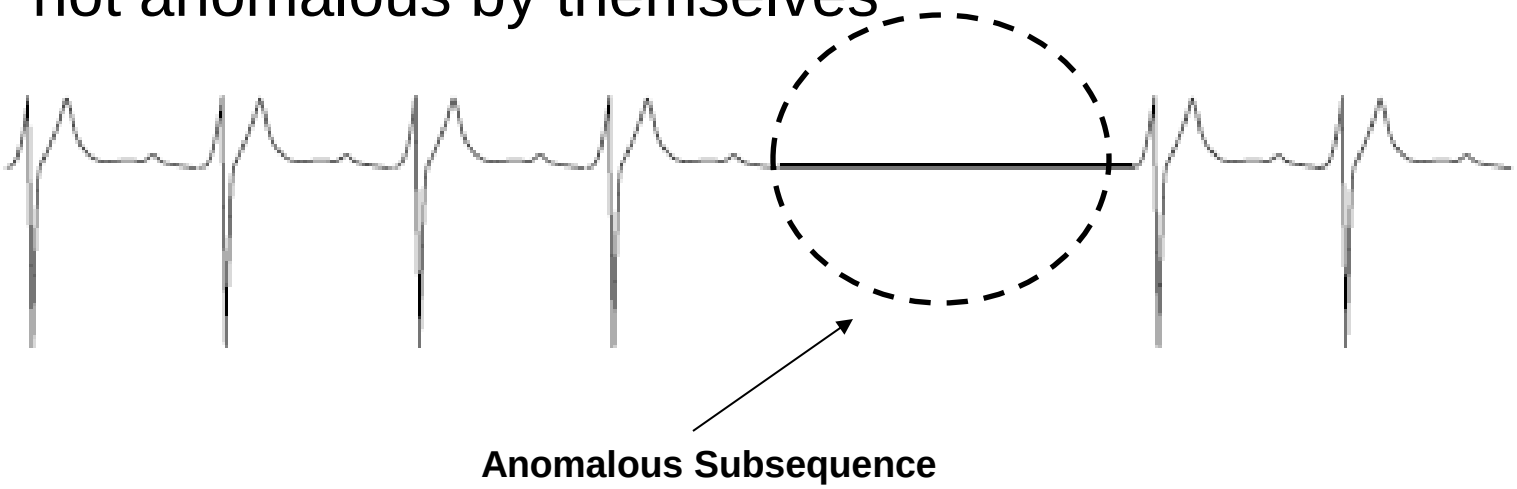
- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies*



* 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



Output of Anomaly Detection

- Label
  - Each test instance is given a *normal* or *anomaly* label
- Score
  - Each test instance is assigned an anomaly score
    - Allows the output to be ranked
    - Requires an additional threshold parameter

Evaluation of Anomaly Detection – F-value

- Accuracy is not sufficient metric for evaluation
  - Example: network traffic data set with 99.9% of normal data and 0.1% of intrusions
  - Trivial classifier that labels everything with the normal class can achieve 99.9% accuracy !!!!!

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	TN	FP
	C	FN	TP

anomaly class – C

normal class – NC

- **Focus on both recall and precision**
  - **Recall (R) = TP/(TP + FN)**
  - **Precision (P) = TP/(TP + FP)**
- **F – measure = 2*R*P/(R+P)**

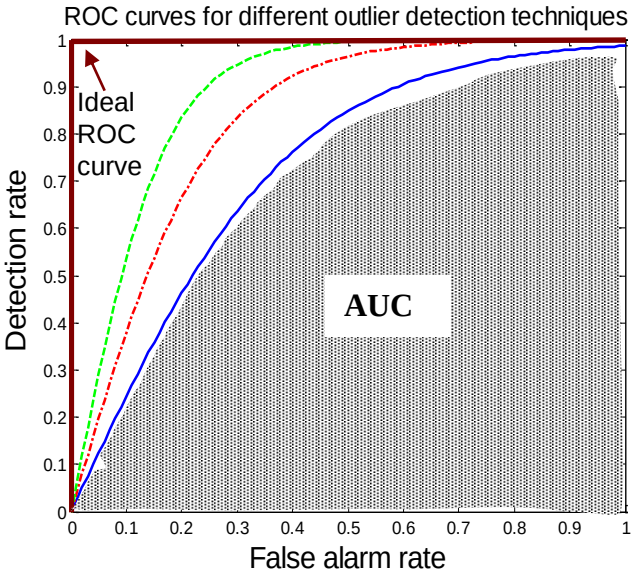
Evaluation of Outlier Detection – ROC & AUC

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	TN	FP
	C	FN	TP

anomaly class – C

normal class – NC

- Standard measures for evaluating anomaly detection problems:
  - *Recall (Detection rate)* - ratio between the number of correctly detected anomalies and the total number of anomalies
  - *False alarm (false positive) rate* – ratio between the number of data records from normal class that are misclassified as anomalies and the total number of data records from normal class
  - *ROC Curve* is a trade-off between detection rate and false alarm rate
  - *Area under the ROC curve (AUC)* is computed using a trapezoid rule



Base Rate Fallacy

- Bayes theorem:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

- More generally:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)}$$

Base Rate Fallacy (Axelsson, 1999)

The base-rate fallacy is best described through example.[2] Suppose that your doctor performs a test that is 99% accurate, i.e. when the test was administered to a test population all of whom had the disease, 99% of the tests indicated disease, and likewise, when the test population was known to be 100% free of the disease, 99% of the test results were negative. Upon visiting your doctor to learn the results he tells you he has good news and bad news. The bad news is that indeed you tested positive for the disease. The good news however, is that out of the entire population the rate of incidence is only 1/10000, i.e. only 1 in 10000 people have this ailment. What, given this information, is the probability of you having the disease? The reader is encouraged to make a quick "guesstimate" of the answer at this point.

Base Rate Fallacv

$$P(S|P) = \frac{P(S) \cdot P(P|S)}{P(S) \cdot P(P|S) + P(\neg S) \cdot P(P|\neg S)}$$

$$P(S|P) = \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} = \\ = 0.00980\ldots \approx 1\%$$

- Even though the test is 99% certain, your chance of having the disease is 1/100, because the population of healthy people is much larger than sick people

Base Rate Fallacy in Intrusion Detection (An example of AD)

- I: intrusive behavior,
  $\neg$I: non-intrusive behavior
  A: alarm
  $\neg$A: no alarm

- Detection rate (true positive rate): P(A|I)
- False alarm rate: P(A|$\neg$I)

- Goal is to maximize both
  - Bayesian detection rate, P(I|A)
  - P($\neg$I|$\neg$A)

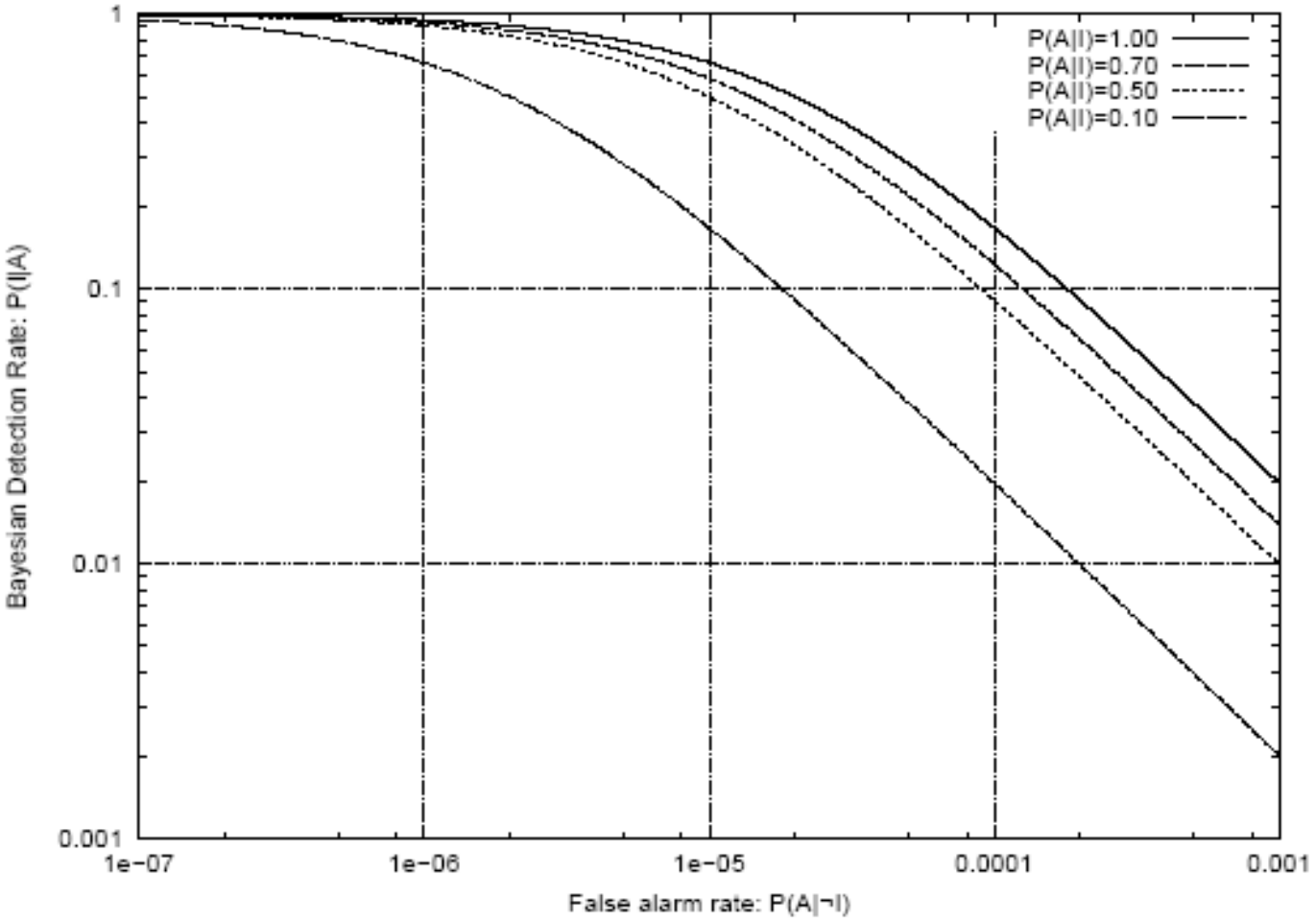
Detection Rate vs False Alarm Rate

$$P(I|A) = \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)}$$

- Suppose: $P(I) = 1 \Big/ \frac{1 \cdot 10^6}{2 \cdot 10} = 2 \cdot 10^{-5};$ $P(\neg I) = 1 - P(I) = 0.99998$
- Then: $P(I|A) = \frac{2 \cdot 10^{-5} \cdot P(A|I)}{2 \cdot 10^{-5} \cdot P(A|I) + 0.99998 \cdot P(A|\neg I)}$

- False alarm rate becomes more dominant if P(I) is very low

Detection Rate vs False Alarm



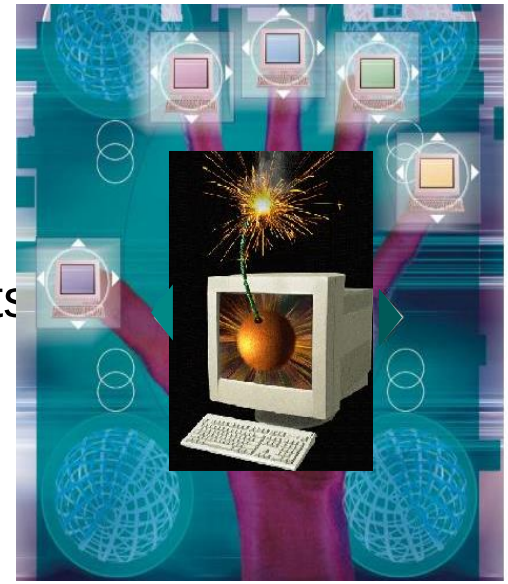
- Axelsson: We need a very low false alarm rate to achieve a reasonable Bayesian detection rate

Applications of Anomaly Detection

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining
- …

Intrusion Detection

- Intrusion Detection:
  - Process of monitoring the events occurring in a computer system or network and analyzing them for intrusions
  - Intrusions are defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional signature-based intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations (But the FPR is key!)



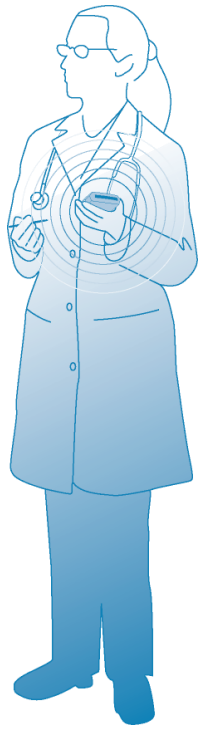
Fraud Detection

- Fraud detection refers to detection of criminal activities occurring in commercial organizations
  - Malicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft).
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading

- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high
    - FP means bothering the client and wasting resources
    - FN means losing the money

Healthcare Informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key Challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal



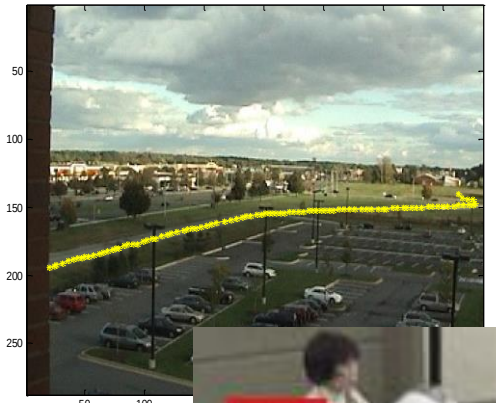
Industrial Damage Detection

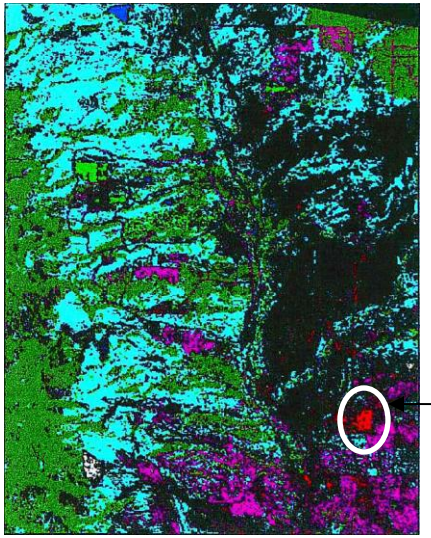
- Industrial damage detection refers to detection of different faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.
  - Example: Aircraft Safety
    - Anomalous Aircraft (Engine) / Fleet Usage
    - Anomalies in engine combustion data
    - Total aircraft health and usage management

- Key Challenges
  - Data is extremely huge, noisy and unlabelled
  - Most of applications exhibit temporal behavior
  - Detecting anomalous events typically require immediate intervention

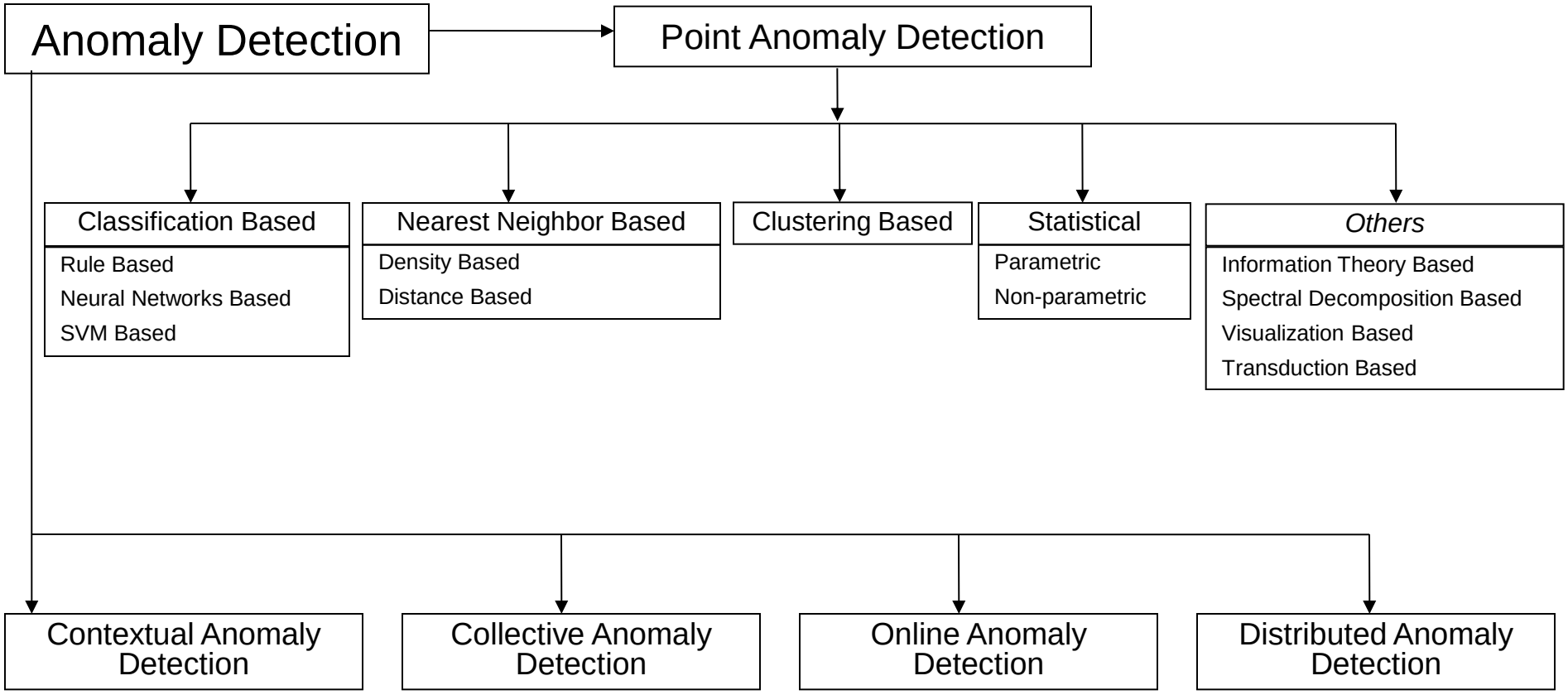
Image Pro

- Detecting outliers in a image monitored over time
- Detecting anomalous regions within an image
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Detecting collective anomalies
  - Data sets are very large





Taxonomy*



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

Classification Based Techniques

- Main idea: build a classification model for normal (and anomalous (rare)) events based on labeled training data, and use it to classify each new unseen event
- Classification models must be able to handle skewed (imbalanced) class distributions
- Categories:
  - *Supervised learning techniques*
    - Require knowledge of both ***normal*** and ***anomaly*** class (this is never possible!)
    - Build classifier to distinguish between normal and known anomalies
  - *Semi-supervised classification techniques*
    - Require knowledge of ***normal*** class only!
    - Use modified classification model to learn the normal behavior and then detect any deviations from normal behavior as anomalous

Classification Based Techniques

- Advantages:
  - *Supervised learning techniques*
    - Models that can be easily understood
    - High accuracy in detecting many kinds of known anomalies
  - *Semi-supervised classification techniques*
    - Models that can be easily understood
    - Normal behavior can be accurately learned
- Drawbacks:
  - *Supervised learning techniques*
    - Require both labels from both normal and anomaly class (a huge mistake)
    - Cannot detect unknown and emerging anomalies (e.g., no zero-day attacks captured)
  - *Semi-supervised classification techniques*
    - Require labels from normal class
    - Possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies

MY VIEW: Anomaly detection & Supervised Learning

- Argument: AD should not be framed as a SL problem:
  - The anomalous ‘class’ is actually an unbounded number of classes
    - E.g.: Normal traffic vs. attacks – there can be an unbounded number of types of attacks
  - As such, modeling anomalies as a single class is a form of overfitting

Supervised Classification Techniques

- Manipulating data records (oversampling / undersampling / generating artificial examples)
- Rule based techniques
- Model based techniques
  - Neural network based approaches
  - Support Vector machines (SVM) based approaches
  - Bayesian networks based approaches
- Cost-sensitive classification techniques
- Ensemble based algorithms (SMOTEBoost, RareBoost

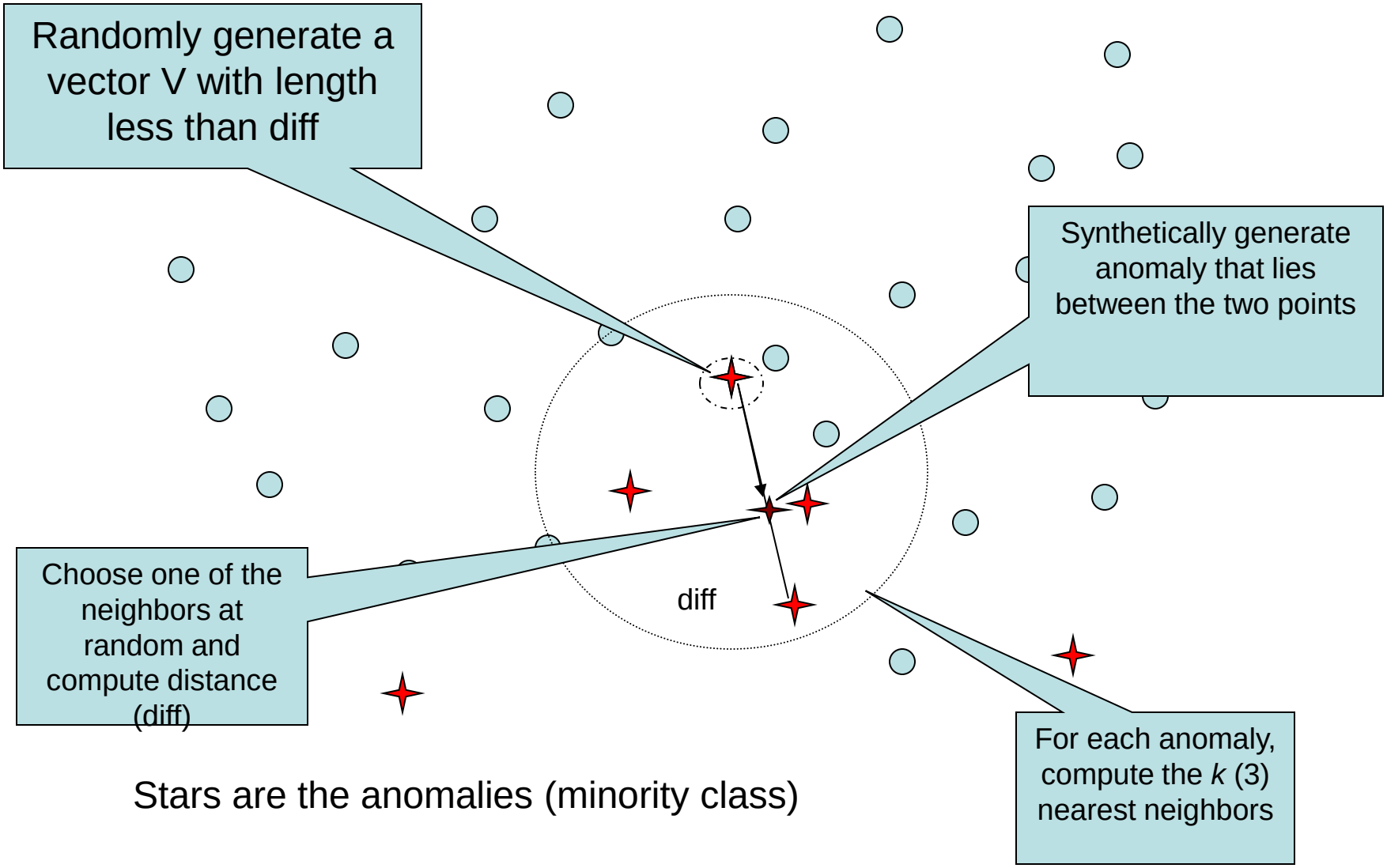
Manipulating Data Records

- **Over-sampling the rare class** [Ling98]
  - Make the duplicates of the rare events until the data set contains as many examples as the majority class => balance the classes
  - Does not increase information but increase misclassification cost
- **Down-sizing (undersampling) the majority class** [Kubat97]
  - Sample the data records from majority class (Randomly, Near miss examples, Examples far from minority class examples (far from decision boundaries)
  - Introduce sampled data records into the original data set instead of original data records from the majority class
  - Usually results in a general loss of information and overly general rules
- **Generating artificial anomalies**
  - SMOTE (Synthetic Minority Over-sampling TEchnique) [Chawla02] - new rare class examples are generated inside the regions of existing rare class examples
  - Artificial anomalies are generated around the edges of the sparsely populated data regions [Fan01]
  - Classify synthetic outliers vs. real normal data using active learning [Abe06]

Example: SMOTE

- SMOTE (Synthetic Minority Oversampling Technique)
  - The idea is not to repeat examples (like in sampling with replacement)
  - Rather we are to generate synthetic examples based on the limited available minority cases, i.e., generate artificial anomalies systematically

SMOTE for continuous features



Stars are the anomalies (minority class)

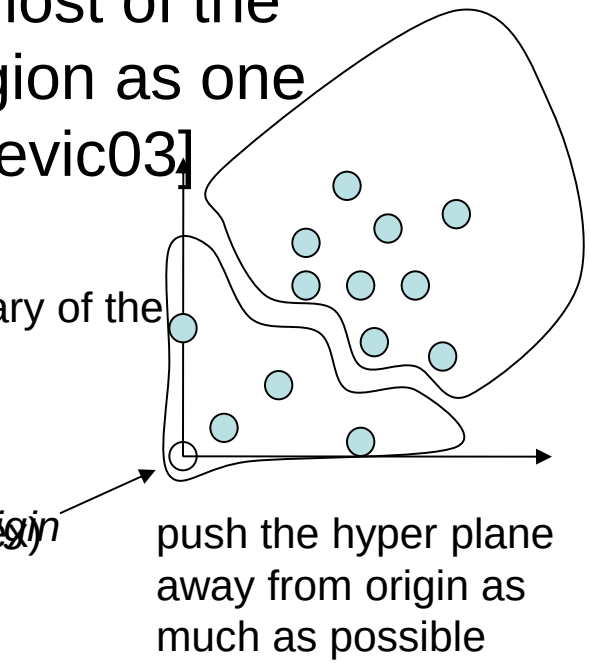
Semi-supervised Classification Techniques

- Use modified classification model to learn the normal behavior and then detect any deviations from normal behavior as anomalous

- Recent approaches:
  - Neural network based approaches
  - Support Vector machines (SVM) based approaches
  - Markov model based approaches
  - Rule-based approaches
  - Bayesian approaches

Using Support Vector Machines

- Converting into one class classification problem (One-class SVM)
  - Separate the entire set of training data from the origin, i.e. to find a small region where most of the data lies and label data points in this region as one class [Ratsch02, Tax01, Eskin02, Lazarevic03]
    - Parameters
      - Expected number of outliers (alternatively, the boundary of the Normal region)
      - Variance of rbf kernel (As the variance of the rbf kernel gets smaller, the number of support vectors is larger and the separating surface gets more complex)

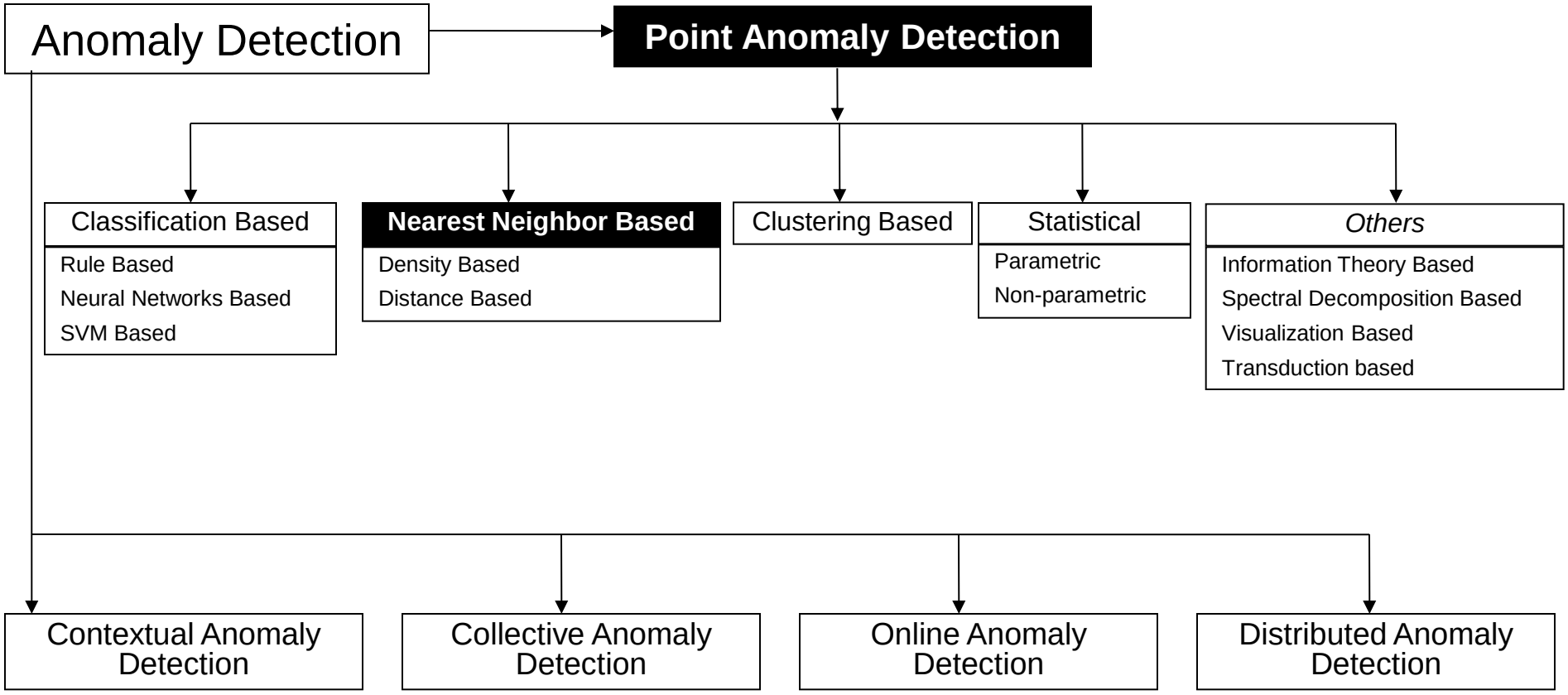


  - Separate regions containing data from the regions containing no data [Scholkopf99]
  - Challenge: Adjust the parameters to define the extent of the hypersphere covering the normal data

Bayesian methods

- Pseudo-Bayes estimators [Barbara01]
  - I stage: learn prior and posterior of unseen anomalies from the training data
  - II stage: use Naive Bayes classifier to classify the instances into normal instances, known anomalies and new anomalies (discovering new classes)

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
Transduction 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
- General two-step approach
  1. Compute neighborhood for each data record
  2. Analyze the neighborhood to determine whether data record is anomaly or not
- Categories:
  - Distance based methods
    - Anomalies are data points most distant from other points
  - Density based methods
    - Anomalies are data points in low density regions

Nearest Neighbor Based Techniques

- Advantage
  - Can be used in unsupervised or semi-supervised setting (do not make any assumptions about data distribution)
- Drawbacks
  - If normal points do not have sufficient number of neighbors the techniques may fail
  - Computationally expensive
  - 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!

Nearest Neighbor Based Techniques

- Distance based approaches
  - A point *O* in a dataset is an *DB*(*p*, *d*) outlier if at least fraction *p* of the points in the data set lies greater than distance *d* from the point *O**
- Density based approaches
  - Compute local densities of particular regions and declare instances in low density regions as potential anomalies
  - Approaches
    - Local Outlier Factor (LOF)
    - Connectivity Outlier Factor (COF
    - Multi-Granularity Deviation Factor (MDEF)

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

Distance based Outlier Detection

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

* 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.

Local Outlier Factor (LOF)*

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

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

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

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

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

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

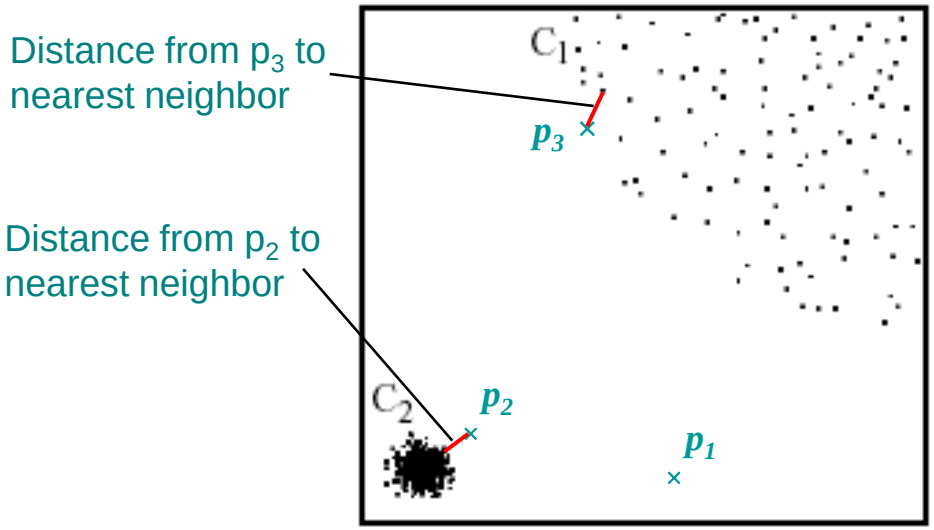
$$LOF(q) = \frac{1}{MinPts} \cdot \sum_{p} \frac{lrd(p)}{lrd(q)}$$

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

Advantages of Density based Techniques

- *Local Outlier Factor (LOF) approach*

  – Example:



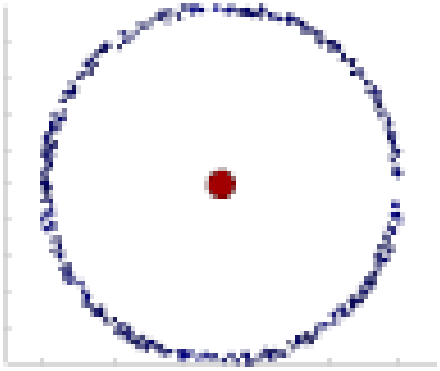
In the *NN* approach, p_2 is not considered as outlier, while the *LOF* approach find both p_1 and p_2 as outliers

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

Connectivity Outlier Factor (COF)*

- Outliers are points *p* where average chaining distance $ac\text{-}dist_{kNN(p)}(p)$ is larger than the average chaining distance (*ac-dist*) of their k-nearest neighborhood kNN(p)
- Let p_2 is a point from G-{p_1} closest to p_1, let p_3 be a point from G-{p_1, p_2} closest to the set {p_1,p_2},…, let p_i be a point from G-{p_1,… p_{i-1}} closest to the set {p_1,…,p_{i-1}}, etc. o_i *is* a point from {p_1,… p_i} closest to p_{i+1}.

$$ac-dist_G(p_1)=\sum_{i=1}^{r-1}\frac{2(r-i)}{r(r-1)}dist(o_i,p_{i+1})$$

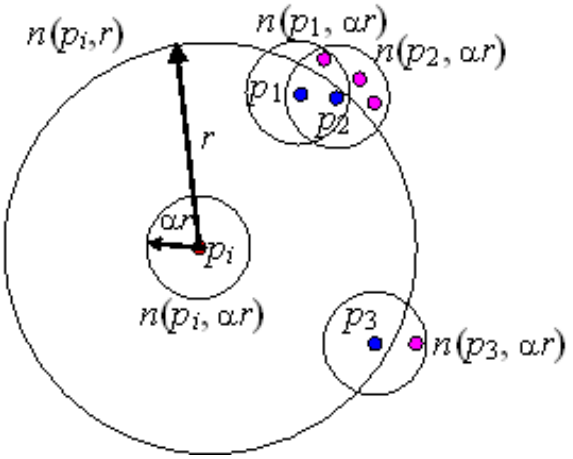


- *dist*(o_i, p_i+*1*) - the single linkage distance between sets {p_1,… p_i} and G-{p_1,… p_i}
- $r = |G|$ is the size of the set G
- $ac\text{-}dist_G(p_1)$ is larger if $dist(o_i, p_{i+1})$ is large for small values of the index *i*, which corresponds to the sparser neighborhood of the point p_1.
- COF identifies outliers as points whose neighborhoods is sparser than the neighborhoods of their neighbors

* J. Tang, Z. Chen, A. W. Fu, D. Cheung, “A robust outlier detection scheme for large data sets,” Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, Taïpeh, Taiwan, 2002.

Multi-Granularity Deviation Factor - LOCI*

- LOCI computes the neighborhood size (the number of neighbors) for each point and identifies as outliers points whose neighborhood size significantly vary with respect to the neighborhood size of their neighbors
- This approach not only finds outlying points but also outlying micro-clusters.
- LOCI algorithm provides LOCI plot which contains information such as inter cluster distance and cluster diameter



Outlier are samples p_i where for any $r \in [r_{min}, r_{max}]$, $n(p_i, \alpha \cdot r)$ significantly deviates from the distribution of values $n(p_j, \alpha \cdot r)$ associated with samples p_j from the r-neighborhood of p_i. Sample is outlier if:

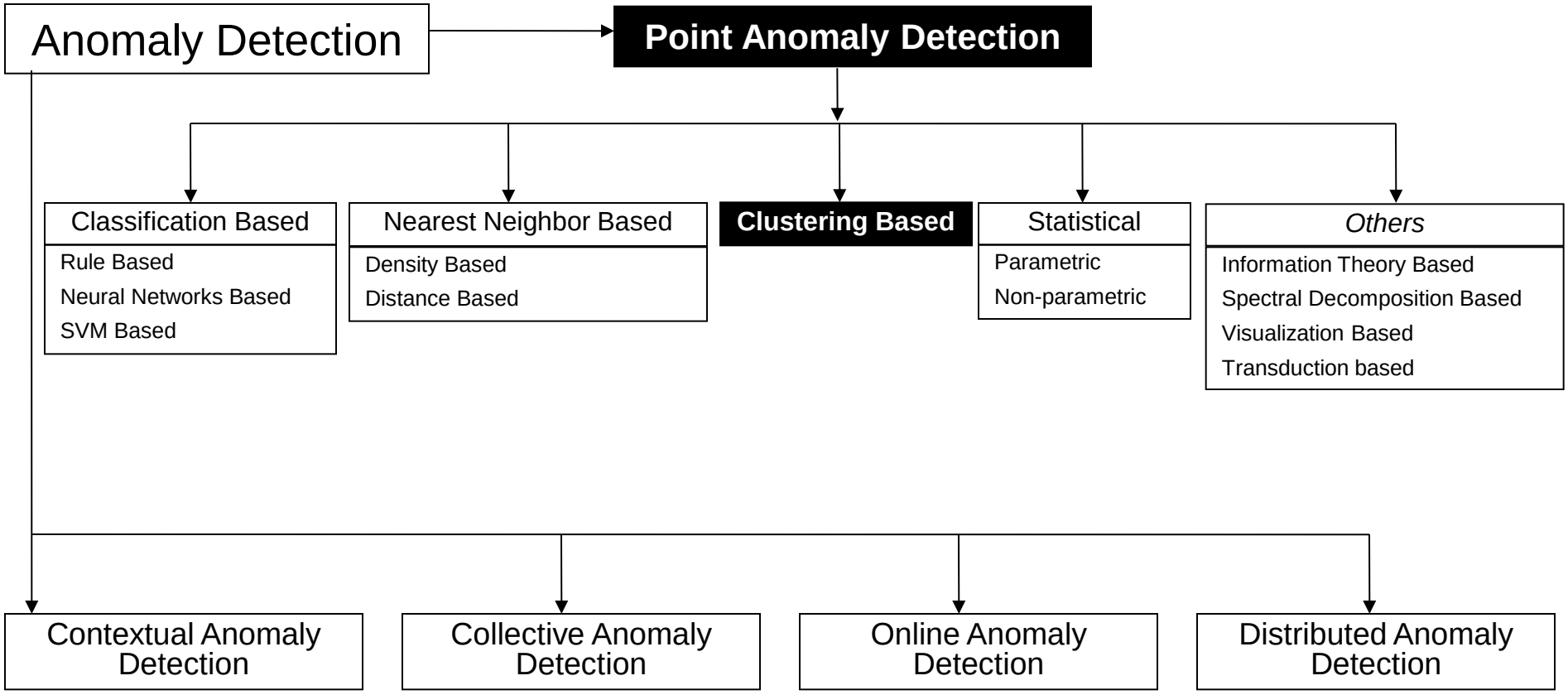
$$n(p_i, \alpha r) < \hat{n}(p_i, r, \alpha) - k_\sigma \sigma_{\hat{n}}(p_i, r, \alpha)$$

Example:

$n(p_i,r)=4$, $n(p_i,\alpha \cdot r)=1$, $n(p_1,\alpha \cdot r)=3$, $n(p_2,\alpha \cdot r)=5$, $n(p_3,\alpha \cdot r)=2$, $\hat{n}(p_i, r, \alpha) = (1+3+5+2) / 4 = 2.75$, $\sigma_{\hat{n}}(p_i, r, \alpha) \approx 1.479$; $\alpha = 1/4$.

*- S. Papadimitriou, et al, “LOCI: Fast outlier detection using the local correlation integral,” *Proc. 19th Int’l Conf. Data Engineering* (*ICDE'03*), Bangalore, India, March 2003.

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
Transduction based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Clustering Based Techniques

- *Key assumption*: normal data records belong to large and dense clusters, while anomalies belong do not belong to any of the clusters or form very small clusters
- Categorization according to labels
  - Semi-supervised – cluster normal data to create modes of normal behavior. If a new instance does not belong to any of the clusters or it is not close to any cluster, is anomaly
  - Unsupervised – post-processing is needed after a clustering step to determine the size of the clusters and the distance from the clusters is required fro the point to be anomaly
- Anomalies detected using clustering based methods can be:
  - Data records that do not fit into any cluster (residuals from clustering)
  - Small clusters
  - Low density clusters or local anomalies (far from other points within the same cluster)

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

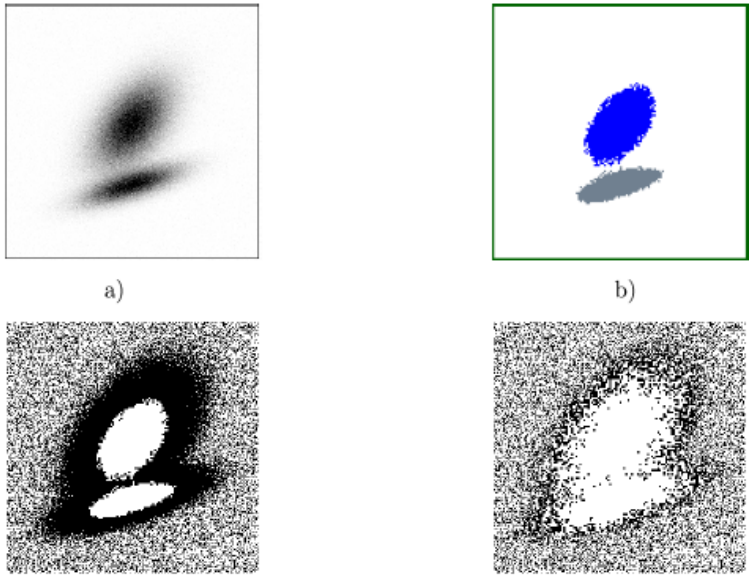
Simple Application of Clustering

- Radius ω of proximity is specified
- Two points x_1 and x_2 are "near" if $d(x_1, x_2) \leq \omega$
- Define $N(x)$ – number of points that are within ω of x
- Time Complexity $O(n^2) \Rightarrow$ approximation of the algorithm
- Fixed-width clustering is first applied
  - The first point is a center of a cluster
  - If every subsequent point is "near" add to a cluster
    - Otherwise create a new cluster
  - Approximate N(x) with N(c)
  - Time Complexity – $O(cn)$, c - # of clusters
- Points in small clusters - anomalies

* E. Eskin et al., A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data, 2002

FindOut

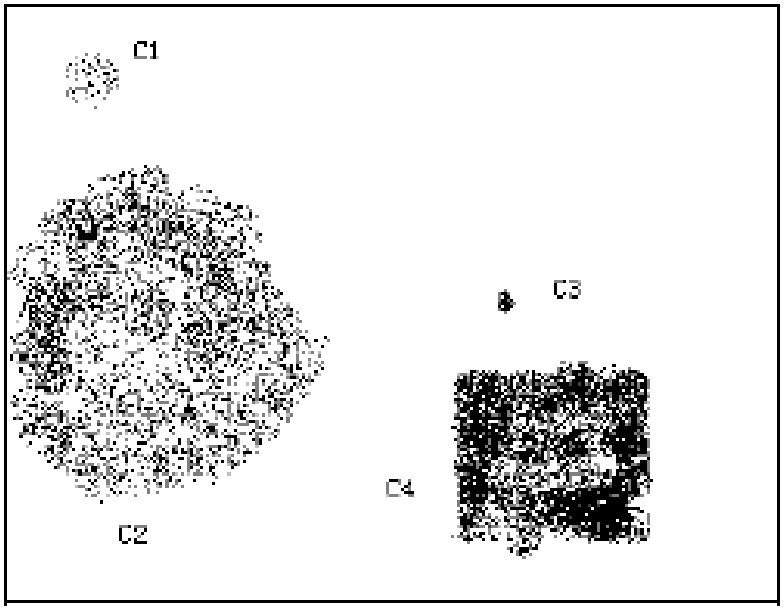
- FindOut algorithm* by-product of *WaveCluster*
- Main idea: Remove the clusters from original data and then identify the outliers
- Transform data into multidimensional signals using wavelet transformation
  - High frequency of the signals correspond to regions where is the rapid change of distribution – 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



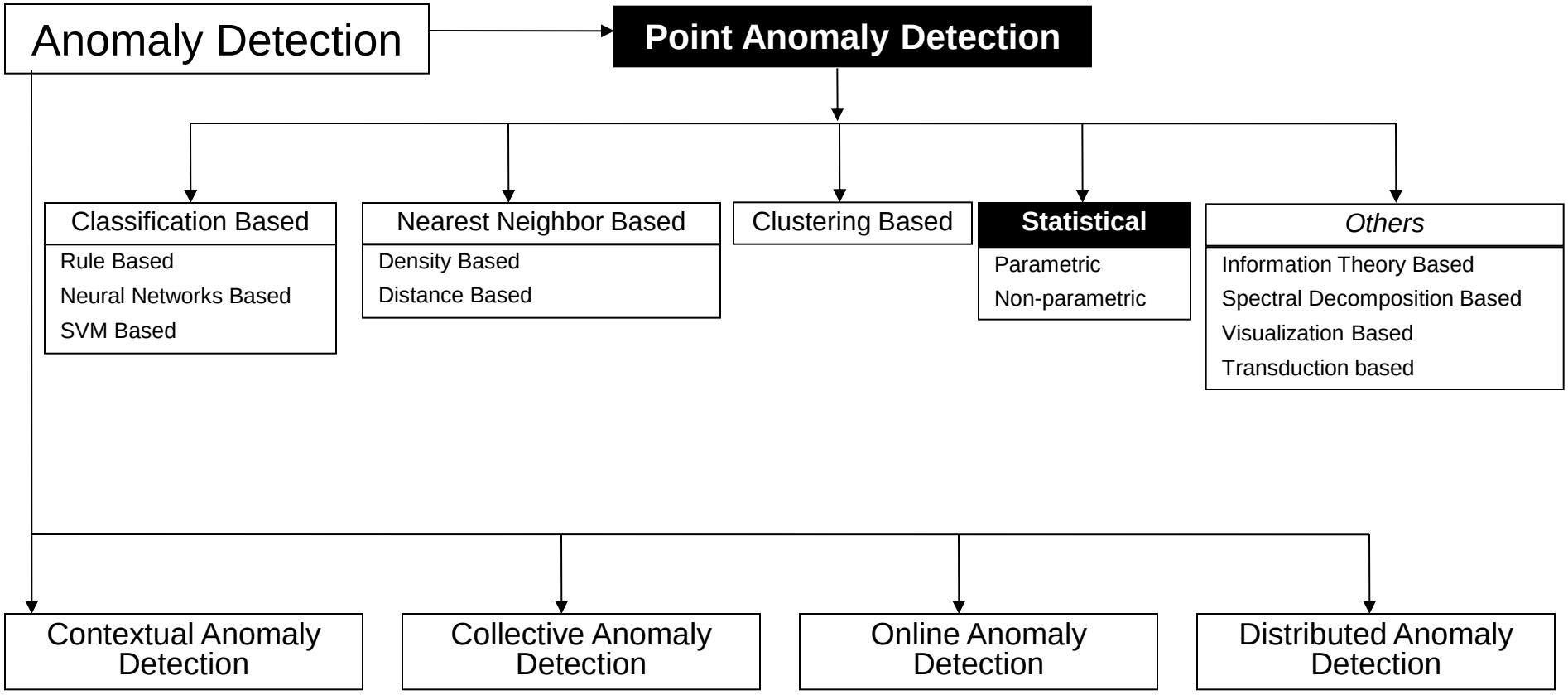
* D. Yu, G. Sheikholeslami, A. Zhang, FindOut: Finding Outliers in Very Large Datasets, 1999.

Cluster based Local Outlier Factor (CBLOF)

- Use squeezer clustering algorithm to perform clustering
- Determine CBLOF for each data record measured by both the size of the cluster and the distance to the cluster
  - if the data record lies in a ***small*** cluster CBLOF is measured as a product of the size of the cluster the data record belongs to and the distance to the closest larger cluster
  - if the object belongs to a ***large*** cluster CBLOF is measured as a product of the size of the cluster that the data record belongs to and the distance between the data record and the cluster it belongs to (this provides importance of the local data behavior)



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
Transduction based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Statistics Based Techniques

- Data points are modeled using stochastic distribution $\Rightarrow$ 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

Types of Statistical Techniques

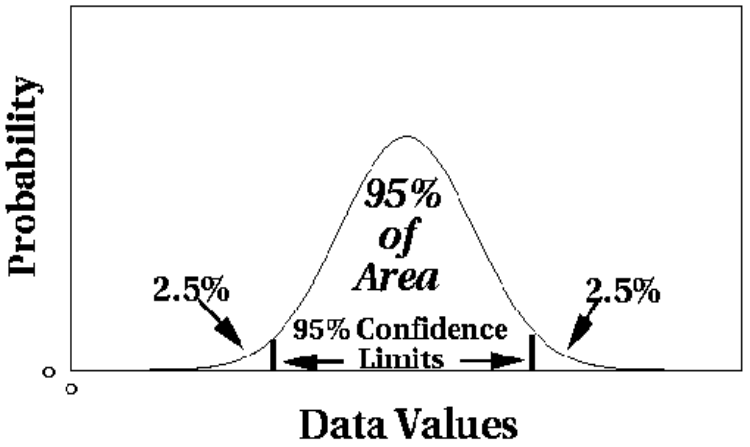
- Parametric Techniques
  - Assume that the normal (and possibly anomalous) data is generated from an underlying parametric distribution
  - Learn the parameters from the normal sample
  - Determine the likelihood of a test instance to be generated from this distribution to detect anomalies
- Non-parametric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

Statistical Approaches (Parametric)

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)

- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)

$$\alpha = prob(|x| \geq c)$$



Specify α

Find c

Any point x such that |x| >= c is an outlier

Grubbs' Test

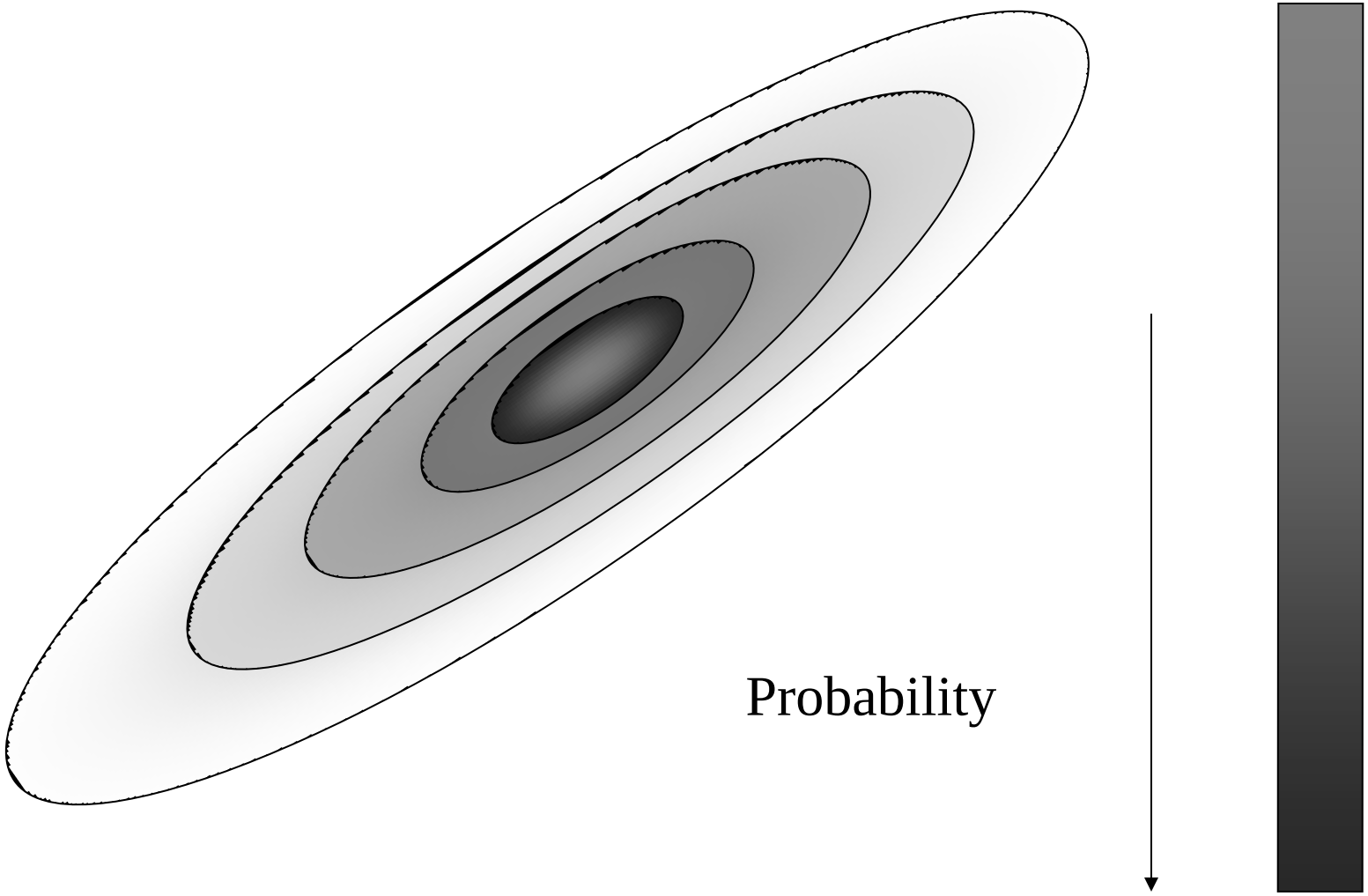
- Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - H_0: There is no outlier in data
  - H_A: There is at least one outlier
- Grubbs' test statistic:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

- Reject H_0 if:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

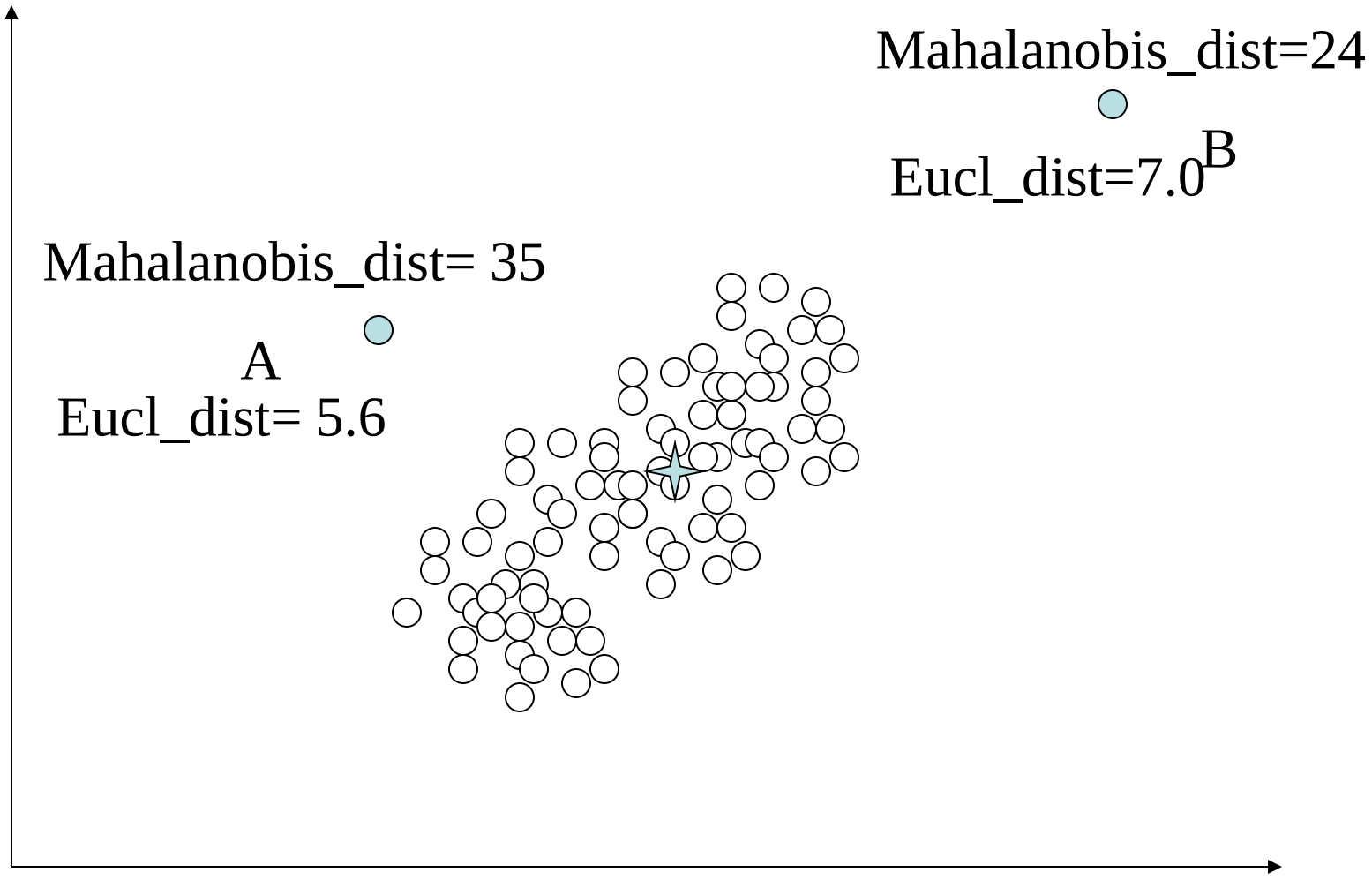
Multivariate data



Mahalanobis distance

$$Mahalanobis(x, \bar{x}) = (x - \bar{x})S^{-1}(x - \bar{x})^T$$

S is the covariance matrix of the data



Even though A is closer to the centroid, its M distance is smaller

Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach

- Data distribution, D = (1 – λ) M + λ A
- M is a probability distribution estimated from data
  - Can be based on any modeling method (naïve Bayes, maximum entropy, etc)
- A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

SmartSifter (SS)*

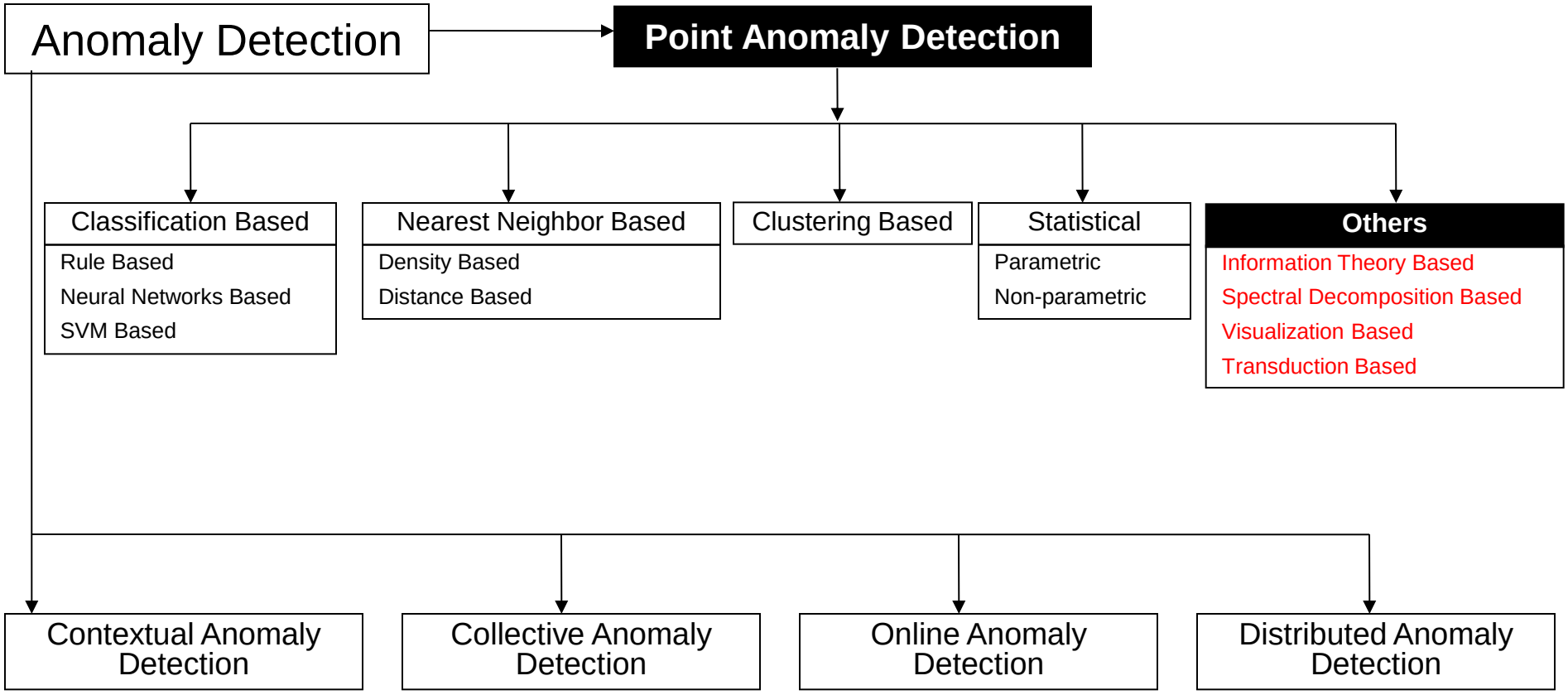
- Uses Finite Mixtures
- SS uses a probabilistic model as a representation of underlying mechanism of data generation.
  - Histogram density used to represent a probability density for categorical attributes
    - SDLE (Sequentially Discounting Laplace Estimation) for learning histogram density for categorical domain
  - Finite mixture model used to represent a probability density for continuous attributes
    - SDEM (Sequentially Discounting Expectation and Maximizing) for learning finite mixture for continuous domain
- SS gives a score to each example x_i on the basis of the learned model, measuring how large the model has changed after the learning

* K. Yamanishi, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, KDD 2000

Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

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
Transduction Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Information Theory Based Techniques

- Compute information content in data using information theoretic measures, e.g., entropy, relative entropy, etc.
- Key idea: Outliers significantly alter the information content in a dataset
- Approach: Detect data instances that significantly alter the information content
  - Require an information theoretic measure
- Advantage
  - Operate in an unsupervised mode
- Challenges
  - Require an information theoretic measure sensitive enough to detect irregularity induced by very few outliers

Information Theory Based Techniques

- Using a variety of information theoretic measures [Lee01]
- Kolmogorov complexity based approaches [Arning96]
  - Detect smallest data subset whose removal leads to maximal reduction in Kolmogorov complexity
- Entropy based approaches [He05]
  - Find a k-sized subset whose removal leads to the maximal decrease in entropy

Spectral Techniques

- Analysis based on eigen decomposition of data
- Key Idea
  - Find combination of attributes that capture bulk of variability
  - Reduced set of attributes can explain normal data well, but not necessarily the outliers
- Advantage
  - Can operate in an unsupervised mode
- Disadvantage
  - Based on the assumption that anomalies and normal instances are distinguishable in the reduced space
- 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

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 y_i denotes the *i*th 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

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

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

* Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. 2003. A novel anomaly detection scheme based on principal component classifier, In Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop.

Temporal analysis of dynamic graphs

- Based on principal component analysis [Ide04]
  - Applied to network traffic data
  - For each time *t*, compute the principal component
  - Stack all principal components over time to form a matrix
  - Left singular vector of the matrix captures normal behavior
  - For any *t*, angle between principal component and the singular vector gives degree of anomaly
- Matrix approximation based methods [Sun07]
  - Form approximation based on CUR decomposition
  - Track approximation error over time
  - High approximation error implies outlying network traffic

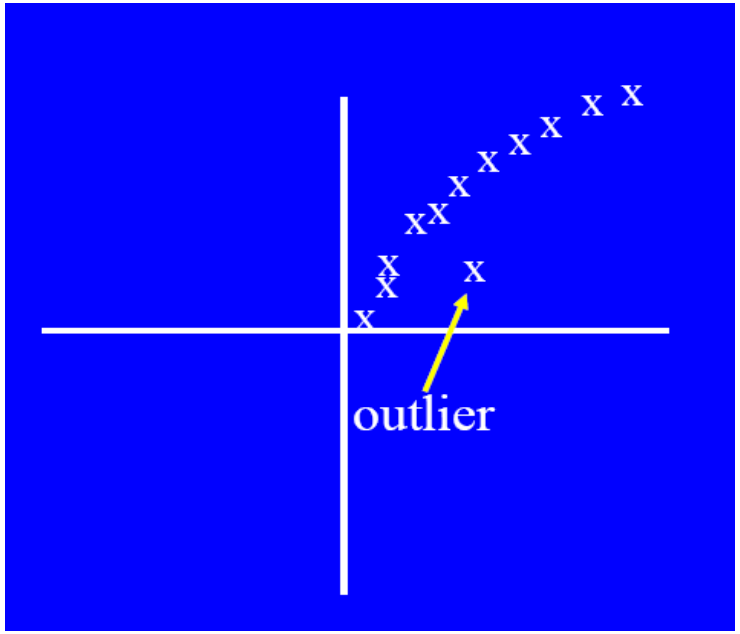
Visualization Based Techniques

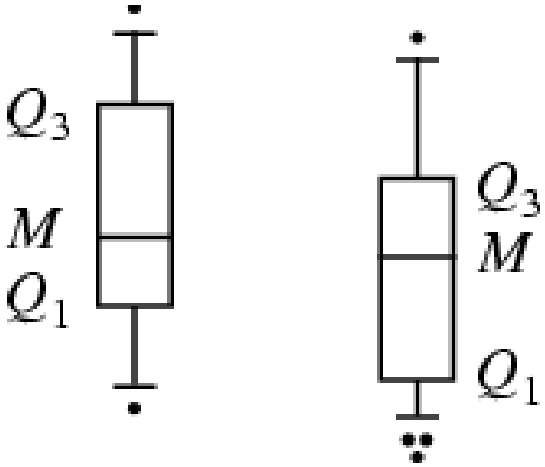
- Use visualization tools to observe the data
- Provide alternate views of data for manual inspection
- Anomalies are detected visually
- Advantages
  - Keeps a human in the loop
- Disadvantages
  - Works well for low dimensional data
  - Can provide only aggregated or partial views for high dimension data

Graphical Approaches

- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)

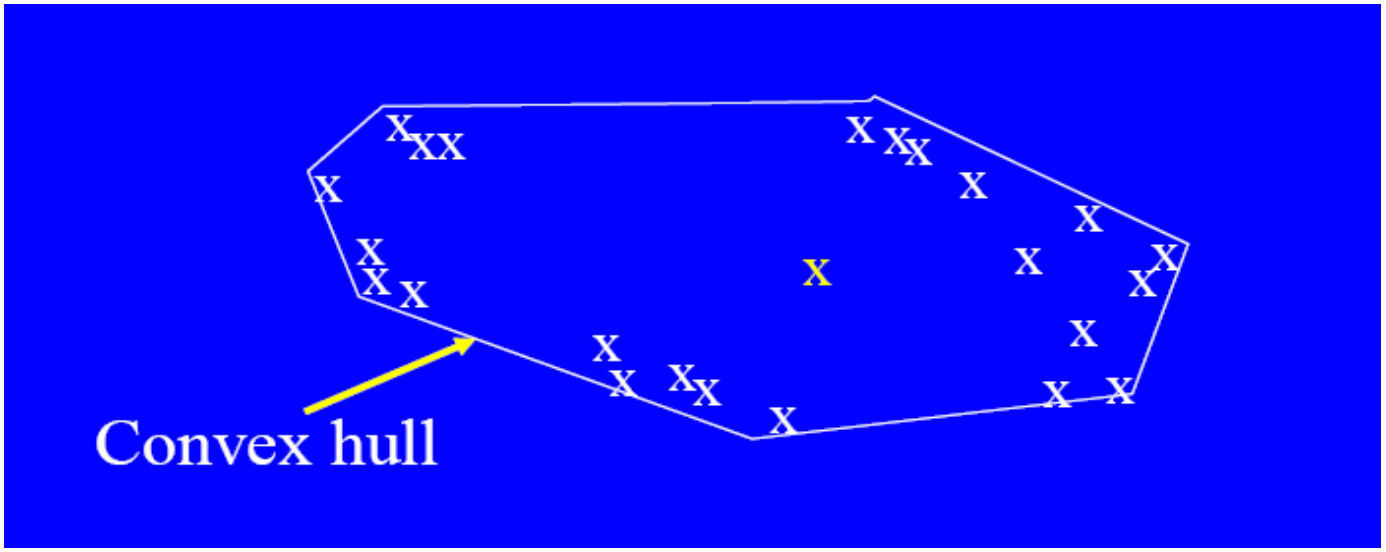
- Limitations
  - Time consuming
  - Subjective





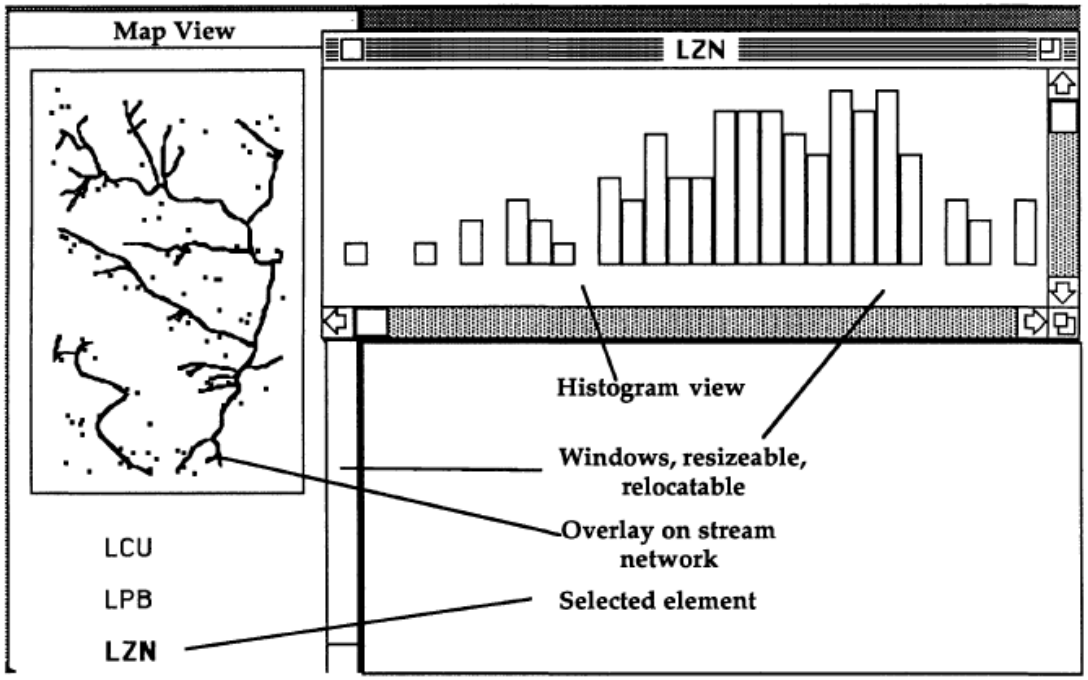
Convex Hull Method

- Extreme points are assumed to be outliers
- Use convex hull method to detect extreme values



Application of Dynamic Graphics*

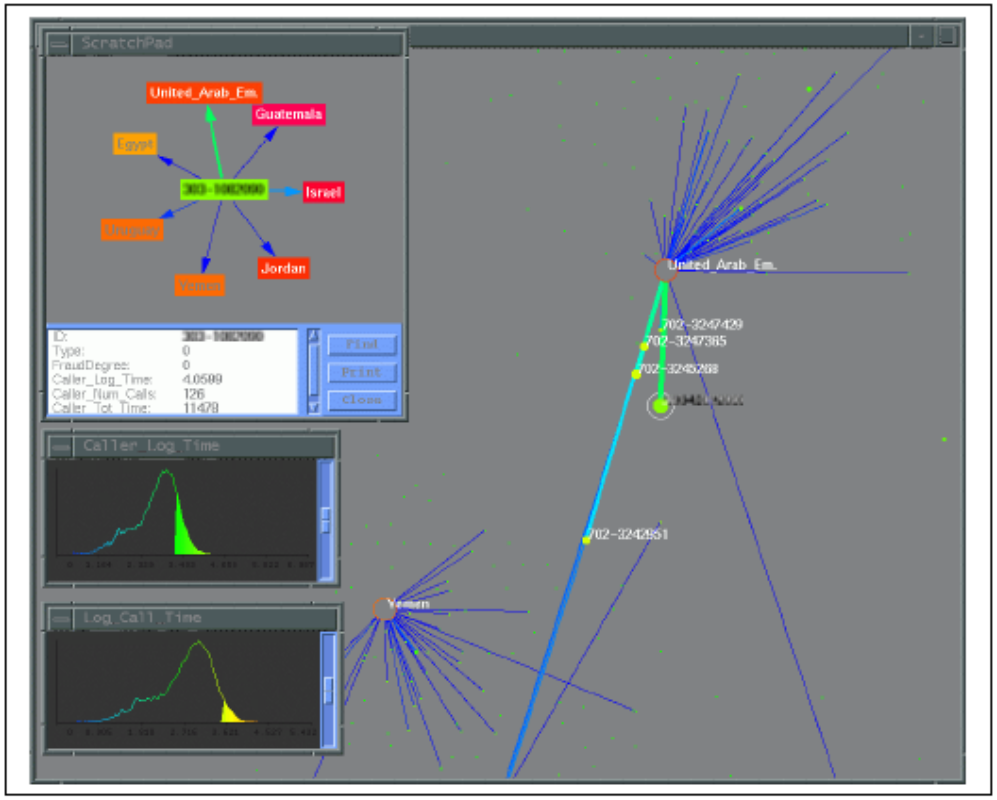
- Apply dynamic graphics to the exploratory analysis of spatial data.
- Visualization tools are used to examine local variability to detect anomalies
- Manual inspection of plots of the data that display its marginal and multivariate distributions



* Haslett, J. et al. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*

Visual Data Mining*

- Detecting Tele-communication fraud
- Display telephone call patterns as a graph
- Use colors to identify fraudulent telephone calls (anomalies)



* Cox et al 1997. Visual data mining: Recognizing telephone calling fraud. *Journal of Data Mining and Knowledge Discovery*

Transduction Based

- Transduction [Vapnik 1998] (the opposite of induction) is the procedure that reasons from specific cases (training) to specific cases (test)
- It avoids creating a model by making only decisions about individual points at a time

Transduction

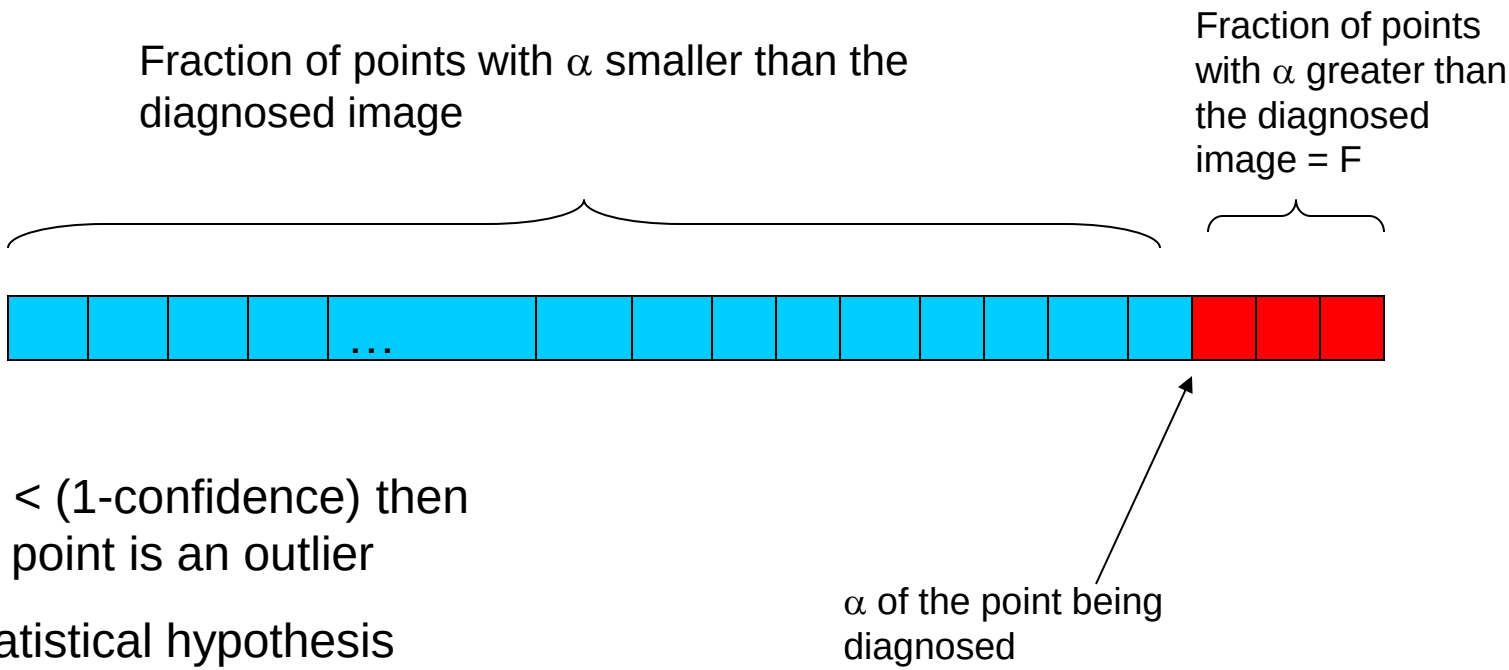
- Vapnik: "When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one."
- Russell: "we shall reach the conclusion that Socrates is mortal with a greater approach to certainty if we make our argument purely inductive than if we go by way of 'all men are mortal' and then use deduction"

Transduction

- In practice:
  - E.g.: in supervised learning, instead of building a model, try to judge the fitness of the point you want to classify by incorporating it into each of the classes
  - To do so, you need a test of fitness, based on a function (strangeness) – how strange is this point in this class?
  - In unsupervised learning you can use the same principle to check if a point belongs in a group of other points (e.g., the normal points)

Transduction Based: StrOUD [Barbará 2006]

- The strangeness function (α) measures how strange an item is. E.g.: sum of the distances to nearest neighbors
- Given a distribution of alphas for the general population, compute the likelihood of being an outlier for a given point



If F < (1-confidence) then the point is an outlier

(Statistical hypothesis testing)

Strangeness

- Examples:
  - Sum of the distances to the *k* nearest neighbors (the larger this is, the farther away the neighbors, and the point is more strange)
  - Size of the ‘codeword’ to describe an entity
    - Components of the entity have each a likelihood *p*
    - Surprise of the components is *–log p*
    - Shannon’s theory says that the optimal coding can be implemented using codewords of size *–log p*
    - Hence the total size of the description is

$$\sum -\log p$$

    - This has been used to find outliers in images (components are objects and the entity is the image)

StrOUD: kNN

Using k-Nearest Neighbors as the strangeness.

$$\alpha_i = \sum_{j \in kNN(i)} d(i,j)$$

(Sum of the distances to the k-nearest neighbors of *i)*

Procedure (part I)

- Sample the dataset
- For every point in the sample:
  - Compute the distances to its k-Nearest neighbors in the sample
  - Add them to compute its strangeness
- Sort the strangeness within the sample (baseline)

Procedure (Part II)

- For every point *i* to be tested:
  - Compute the distances of the point to the points in the sample before and find its k-Nearest Neighbors
  - Sum the distances to k-Nearest Neighbors to compute its strangeness
  - Using the baseline of strangeness, figure out how many strangeness values are bigger than or equal to the one you computed. Call that *b*
  - Compute $pvalue_i = \frac{b+1}{N+1}$ (where *N* is the size of the sample used before)

Procedure (final)

- Compare $pvalue_i$ with (1-confidence) (E.g., 1-0.95 =0.05)
  - If $pvalue_i < (1 - confidence)$ declare *i* an outlier
  - Otherwise *i* is normal

Using LOF as strangeness

Procedure (part I)

- Sample the dataset
- For every point in the sample:
  - Compute its LOF
  - Use LOF as its strangeness
- Sort the strangeness within the sample (baseline)

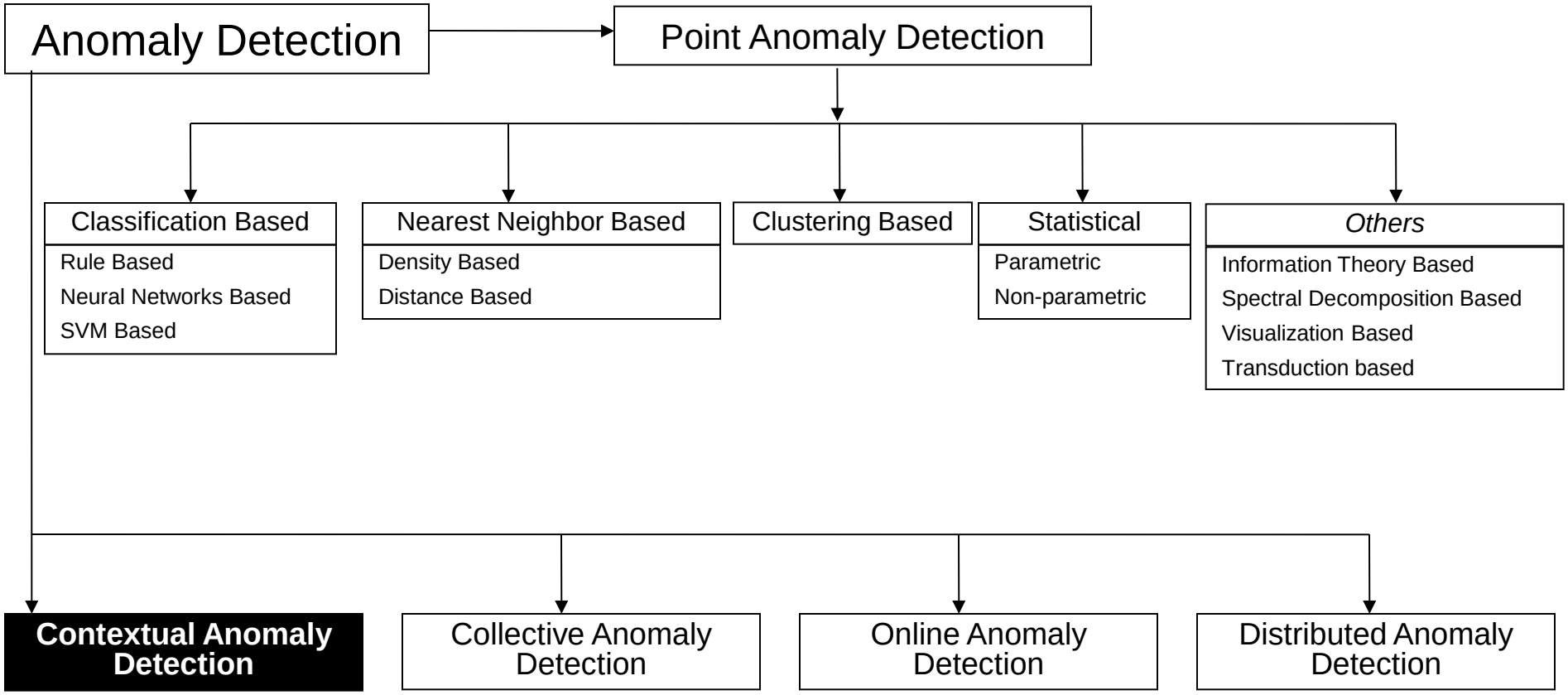
Procedure (Part II)

- For every point in the test set
  - Compute its LOF (with respect to the baseline)
  - Using the baseline of strangeness, figure out how many strangeness values are bigger than or equal to the one you computed. Call that *b.*
  - Compute $p - value_i = \frac{b+1}{N+1}$

Procedure (final)

- Compare $pvalue_i$ with (1-confidence) (E.g., 1-0.95 =0.05)
  - If $pvalue_i < (1 - confidence)$ declare *i* an outlier
  - Otherwise *i* is normal

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
Transduction based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Contextual Anomaly Detection

- Detect context anomalies
- General Approach
  - Identify a context around a data instance (using a set of *contextual attributes*)
  - Determine if the data instance is anomalous w.r.t. the context (using a set of *behavioral attributes*)
- Assumption
  - All normal instances within a context will be similar (in terms of behavioral attributes), while the anomalies will be different

Contextual Anomaly Detection

- Advantages
  - Detect anomalies that are hard to detect when analyzed in the global perspective
- Challenges
  - Identifying a set of good contextual attributes
  - Determining a context using the contextual attributes

Contextual Attributes

- Contextual attributes define a neighborhood (context) for each instance
- For example:
  - Spatial Context
    - *Latitude, Longitude*
  - Graph Context
    - *Edges, Weights*
  - Sequential Context
    - *Position, Time*
  - Profile Context
    - *User demographics*

Contextual Anomaly Detection Techniques

- Techniques
  - Reduction to point outlier detection
    - Segment data using contextual attributes
    - Apply a traditional point outlier within each context using behavioral attributes
  - Utilizing structure in data
    - Build models from the data using contextual attributes
      - E.g. – Time series models (ARIMA, etc.)

Conditional Anomaly Detection*

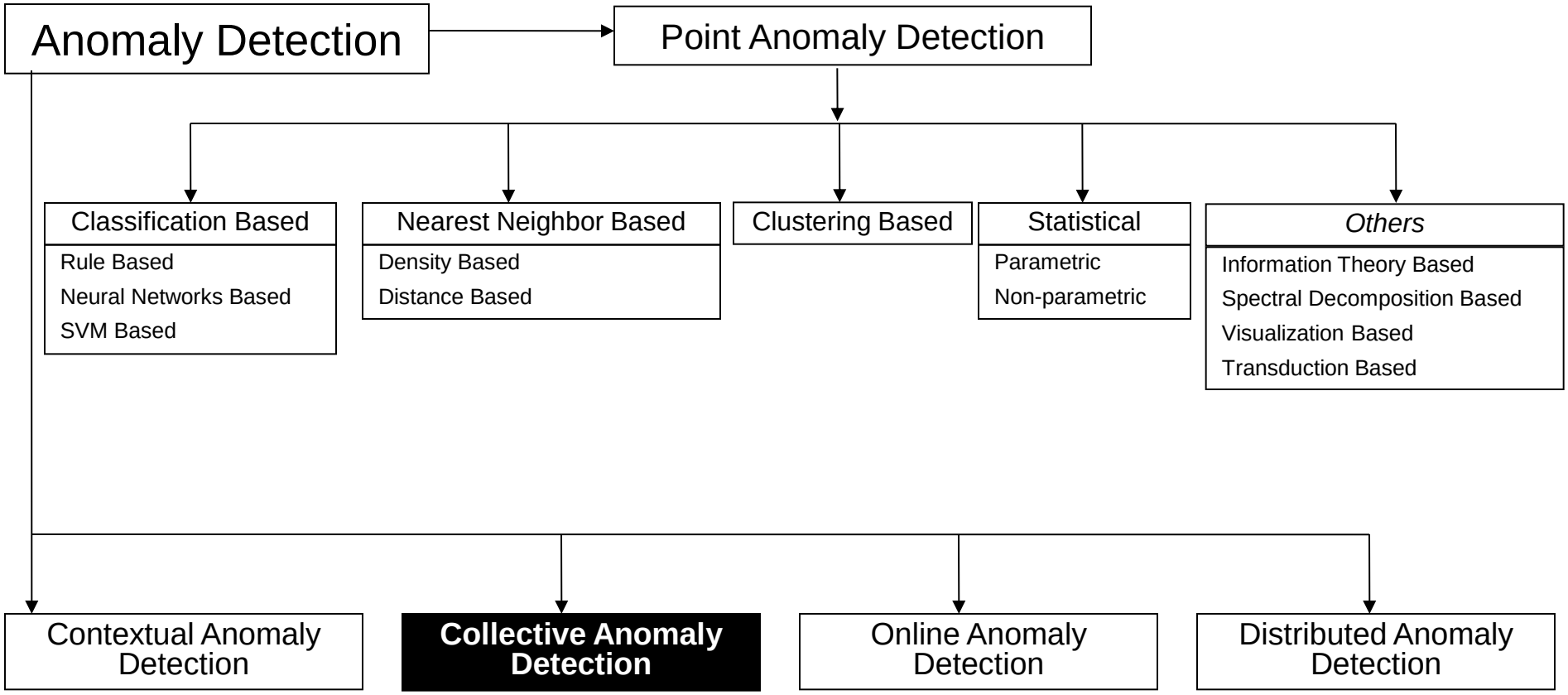
- Each data point is represented as [x,y], where x denotes the *environmental (contextual) attributes* and y denotes the *indicator (behavioral) attributes*
- A mixture of N_U Gaussian models, U is learnt from the contextual data
- A mixture of N_V Gaussian models, V is learn from the behavioral data
- A mapping $p(V_j|U_i)$ is learnt that indicates the probability of the behavioral part to be generated by component V_j when the contextual part is generated by component U_i

- Outlier Score of a data instance ([x,y]):

$$Outlier\ Score = \sum_{i=1}^{n_U} p(x \in U_i) \sum_{j=1}^{n_V} p(y \in V_j) p(V_j|U_i)$$

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

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
Transduction Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Collective Anomaly Detection

- Detect collective anomalies
- Exploit the relationship among data instances
- Sequential anomaly detection
  - Detect anomalous sequences
- Spatial anomaly detection
  - Detect anomalous sub-regions within a spatial data set
- Graph anomaly detection
  - Detect anomalous sub-graphs in graph data

Sequential Anomaly Detection

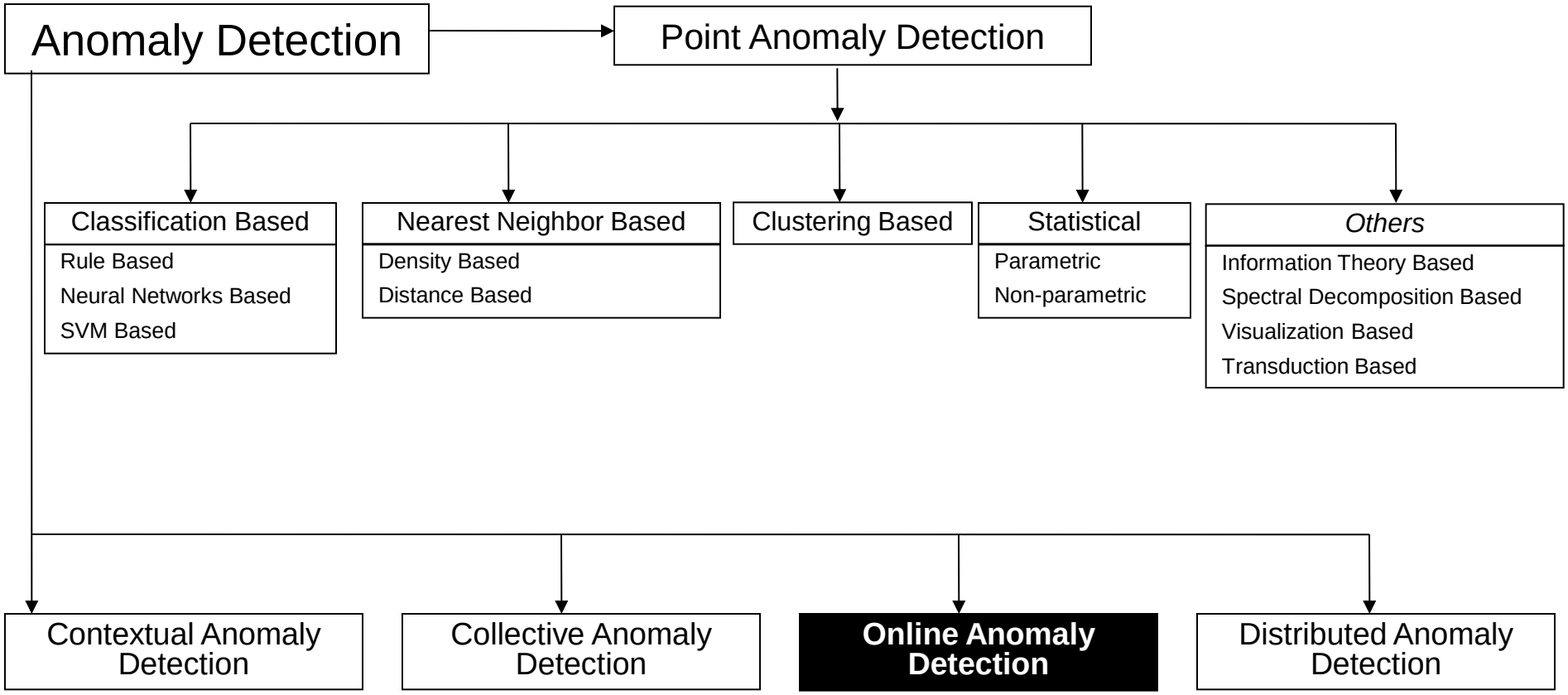
- Detect anomalous sequences in a database of sequences, or
- Detect anomalous subsequence within a sequence
- Data is presented as a set of symbolic sequences
  - System call intrusion detection
  - Proteomics
  - Climate data

Sequence Time Delay Embedding (STIDE)*

- Assumes a training data containing normal sequences
- Training
  - Extracts fixed length (k) subsequences by sliding a window over the training data
  - Maintain counts for all subsequences observed in the training data
- Testing
  - Extract fixed length subsequences from the test sequence
  - Find empirical probability of each test subsequence from the above counts
  - If probability for a subsequence is below a threshold, the subsequence is declared as anomalous
  - Number of anomalous subsequences in a test sequence is its anomaly score
- Applied for system call intrusion detection

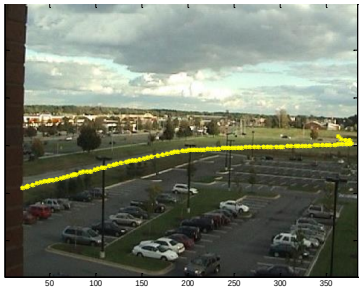
* Warrender, Christina, Stephanie Forrest, and Barak Pearlmutter. Detecting Intrusions Using System Calls: Alternative Data Models. To appear, 1999 IEEE Symposium on Security and Privacy. 1999.

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
Transduction Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Motivation for On-line Anomaly Detection

- Data in many rare events applications arrives continuously at an enormous pace
- There is a significant challenge to analyze such data
- Examples of such rare events applications:
  - Video analysis

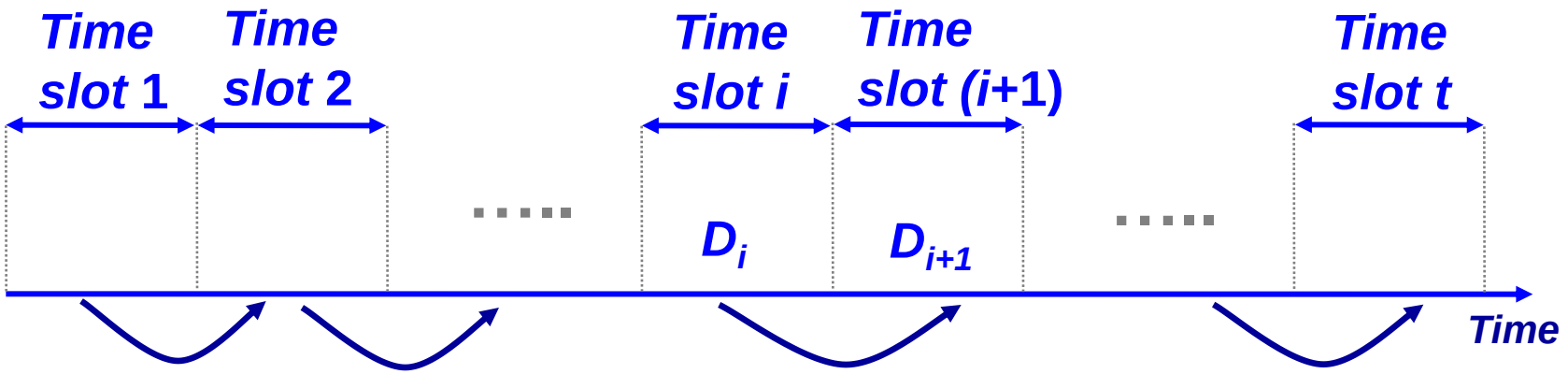


  - Network traffic monitoring
  - Aircraft safety

  - Credit card fraudulent transactions

On-line Anomaly Detection – Simple Idea

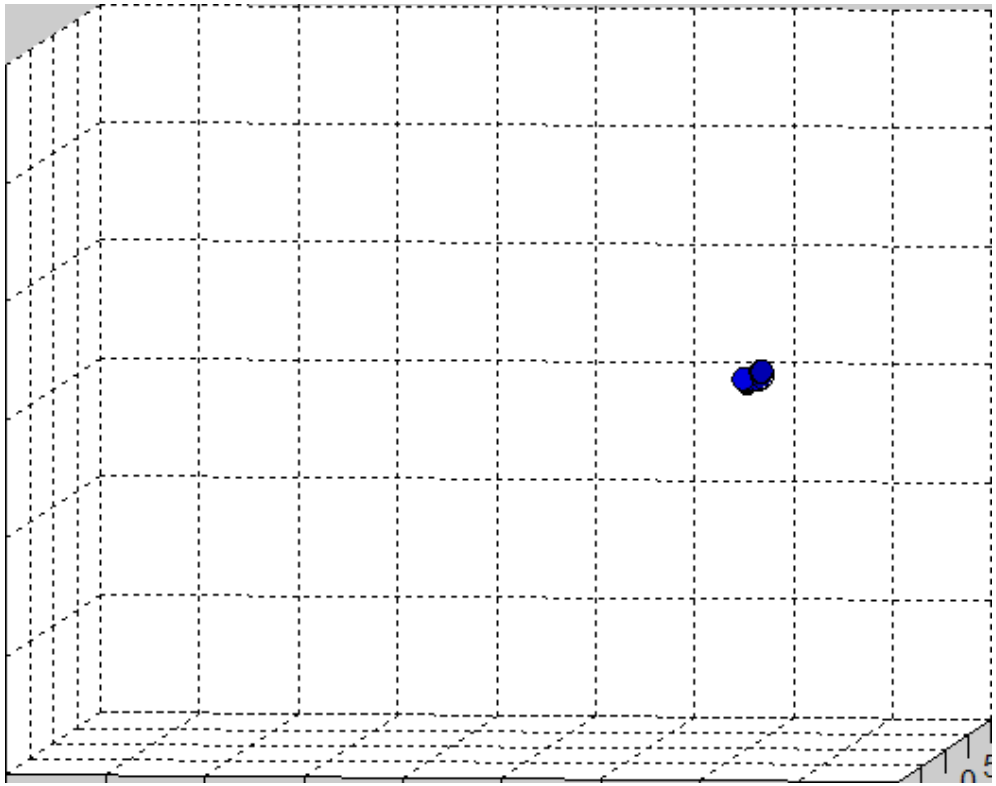
- The normal behavior is changing through time
- Need to update the "normal behavior" profile dynamically
  - Key idea: Update the normal profile with the data records that are "probably" normal, i.e. have very low anomaly score



  - Time slot i – Data block D_i – model of normal behavior M_i
  - Anomaly detection algorithm in time slot ($i+1$) is based on the profile computed in time slot i

Drawbacks of simple on-line anomaly detection algorithm

- If arriving data points start to create a new data cluster, this method will not be able to detect these points as outliers neither the time when the change occurred



United Technologies

Incremental LOF algorithm

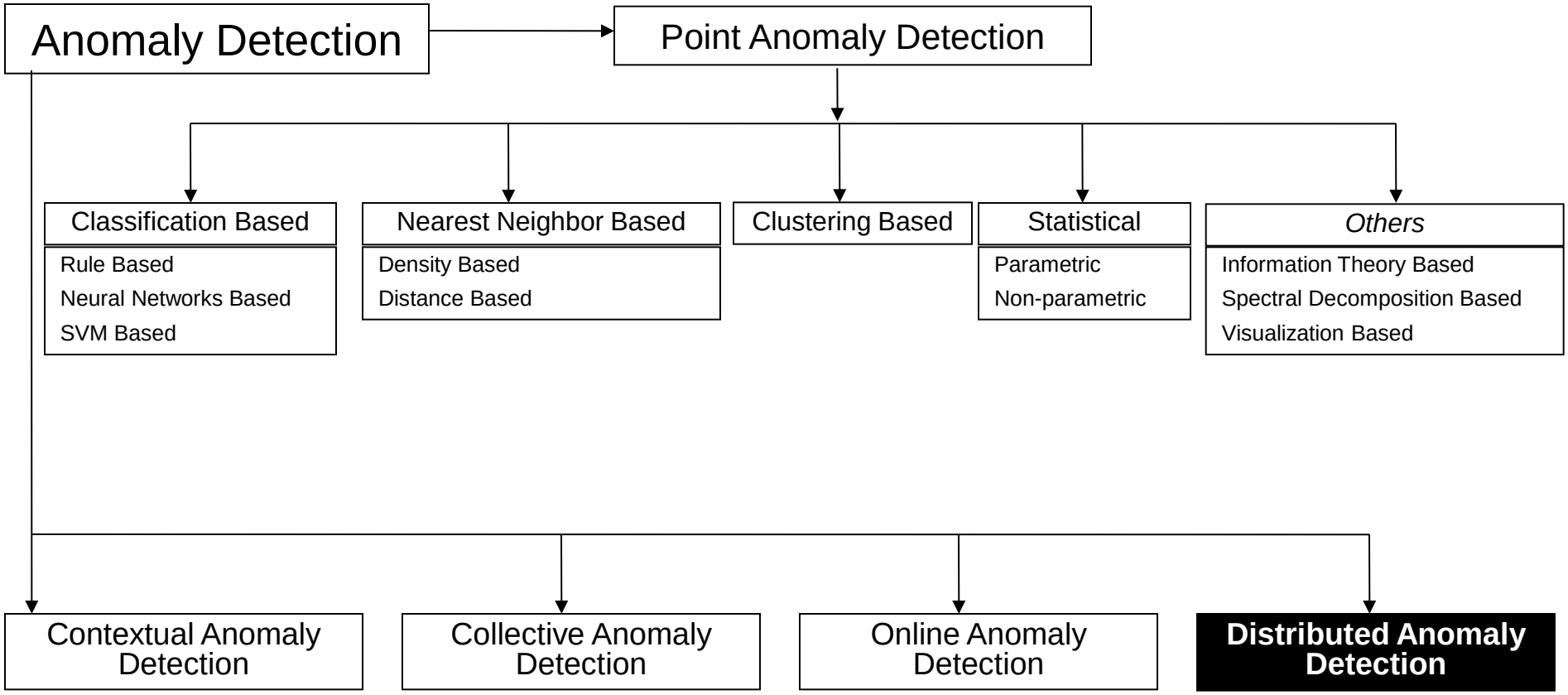
- Incremental *LOF* algorithm computes *LOF* value for each inserted data record and instantly determines whether that data record is outlier
- *LOF* values for existing data records are updated if necessary

Incremental LOF_insertion(*Dataset S*)

- Given: Set S $\{p_1, \ldots, p_N\}$ $p_i \in \mathbf{R}^D$, where ***D*** corresponds to the dimensionality of data records.
- For each data point p_c coming into *data set S*
  - insert(p_c)
  - *Compute* kNN(p_c)
  - ($\forall p_j \in k$NN(p_c))
    compute *reach-dist*$_k(p_c, p_j)$ using Eq. (1);

  //Update_neighbors of p_c
  - $S_{update\_k\_distance}$ =kRNN(p_c);
  - ($\forall p_j \in S_{update\_k\_distance}$)
    update k-distance(p_j) using Eq.(5);
  - $S_{update\_lrd} = S_{update\_k\_distance}$;
  - ($\forall p_j \in S_{update\_k\_distance}$), ($\forall p_i \in k$NN($p_j$)\{$p_c$})
    reach-dist$_k(p_i, p_j)$ =k-distance(p_j);
    if $p_j \in k$NN(p_i)
    $S_{update\_lrd} = S_{update\_lrd} \cup \{p_i\}$;
  - $S_{update\_LOF} = S_{update\_lrd}$;
  - ($\forall p_m \in S_{update\_lrd}$)
    update lrd(p_m) using Eq. (2);
    $S_{update_LOF} = S_{update_LOF} \cup k$RNN($p_m$);
  - ($\forall p_l \in S_{update\_LOF}$)
    update LOF(p_l) using Eq.(3);
  - compute lrd(p_c) using Eq.(2);
  - compute LOF(p_c) using Eq.(3);
- End //for

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

Need for Distributed Anomaly Detection

- Data in many anomaly detection applications may come from many different sources
  - Network intrusion detection
  - Credit card fraud
  - Aviation safety
- Failures that occur at multiple locations simultaneously may be undetected by analyzing only data from a single location
  - Detecting anomalies in such complex systems may require integration of information about detected anomalies from single locations in order to detect anomalies at the global level of a complex system
- There is a need for the high performance and distributed algorithms for correlation and integration of anomalies

Distributed Anomaly Detection Techniques

- Simple data exchange approaches
  - Merging data at a single location
  - Exchanging data between distributed locations
- Distributed nearest neighboring approaches
  - Exchanging one data record per distance computation – computationally inefficient
  - privacy preserving anomaly detection algorithms based on computing distances across the sites
- Methods based on exchange of models
  - explore exchange of appropriate statistical / data mining models that characterize normal / anomalous behavior
    - identifying modes of normal behavior;
    - describing these modes with statistical / data mining learning models; and
    - exchanging models across multiple locations and combing them at each location in order to detect global anomalies

Conclusions

- Many more methods
- Very important area of research
- Lots of interest
  - Commercial: e.g., banking, telco
  - Intelligence analysis