Project 2: 46 days left

Detection-Based Cybersecurity: Anomaly Detection

CS 459/559: Science of Cyber Security
15th Lecture

Instructor:

Guanhua Yan

Agenda

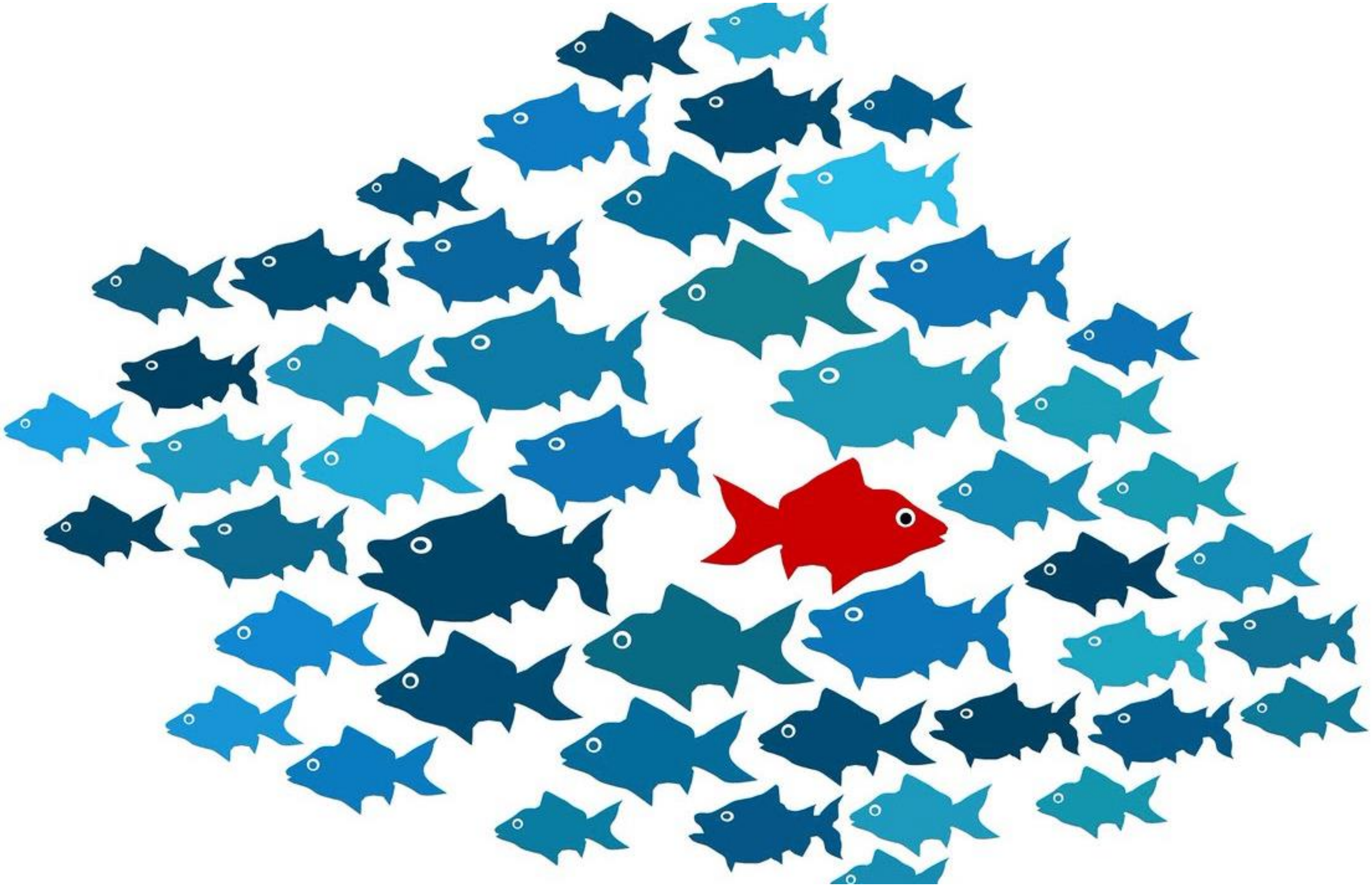
- ~~Quiz 1: September 29 (closed book)~~
- ~~Project 1 (offense): October 10~~
- Quiz 2: November 12
- Presentations: 11/17, 11/19, 11/24, 12/1, 12/3
- CTF competition: November 26
- Project 2 (defense): December 5
- Final report: December 15

Outline

- What is anomaly detection?
- Anomaly detection techniques
- Anomaly detection applications

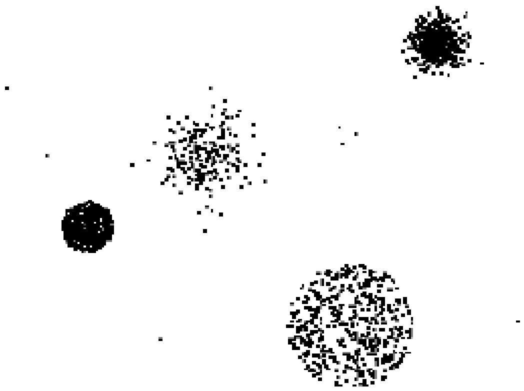
What is anomaly?

What is an anomaly?



Anomaly/Outlier Detection

- **What are anomalies/outliers?**
  - The set of data points that are considerably **different** than the remainder of the data



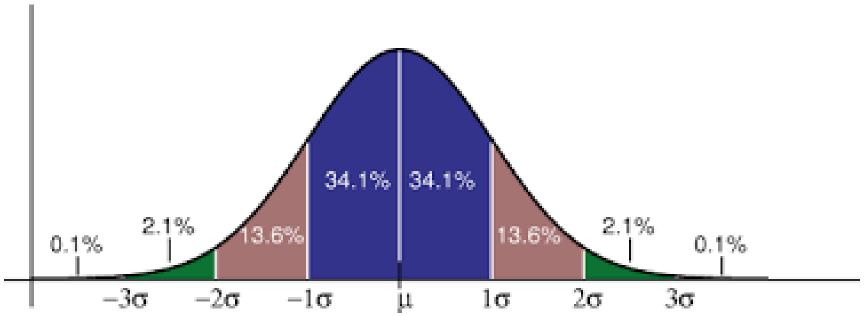
- **Natural implication is that anomalies are relatively rare**
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July

- **Can be important or a nuisance**
  - Unusually high blood pressure
  - 200 pound, 2 year old

Question 1: What causes anomalies?

Causes of anomalies

- **Data from a different class of object or underlying mechanism**
  - Disease vs. non-disease
  - Fraud vs. non-fraud
- **Natural variation**
  - Tails on a Gaussian distribution



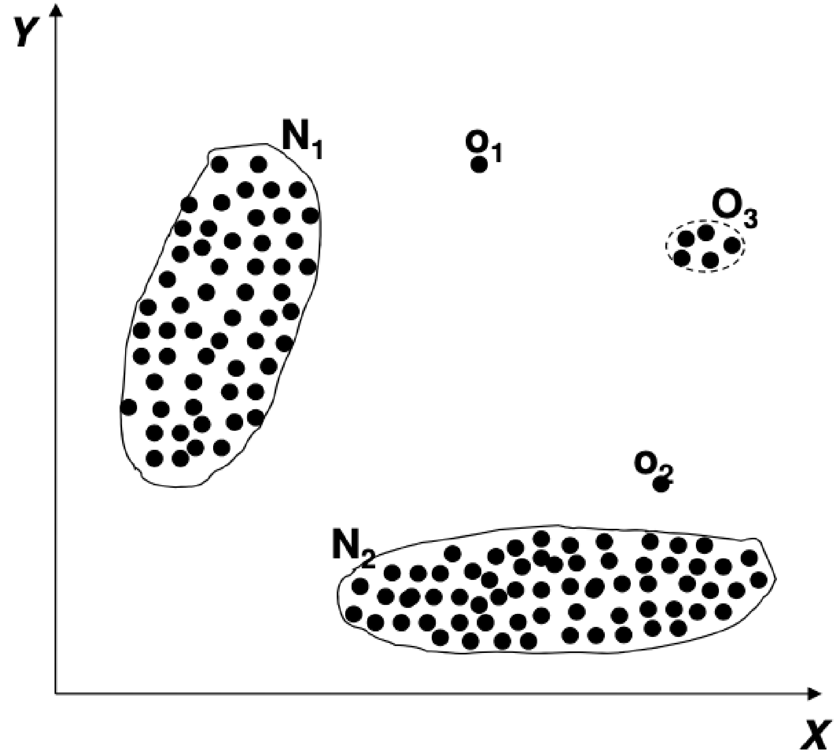
- **Data measurement and collection errors**

Structure of anomalies

- Point anomalies
- Contextual anomalies
- Collective anomalies

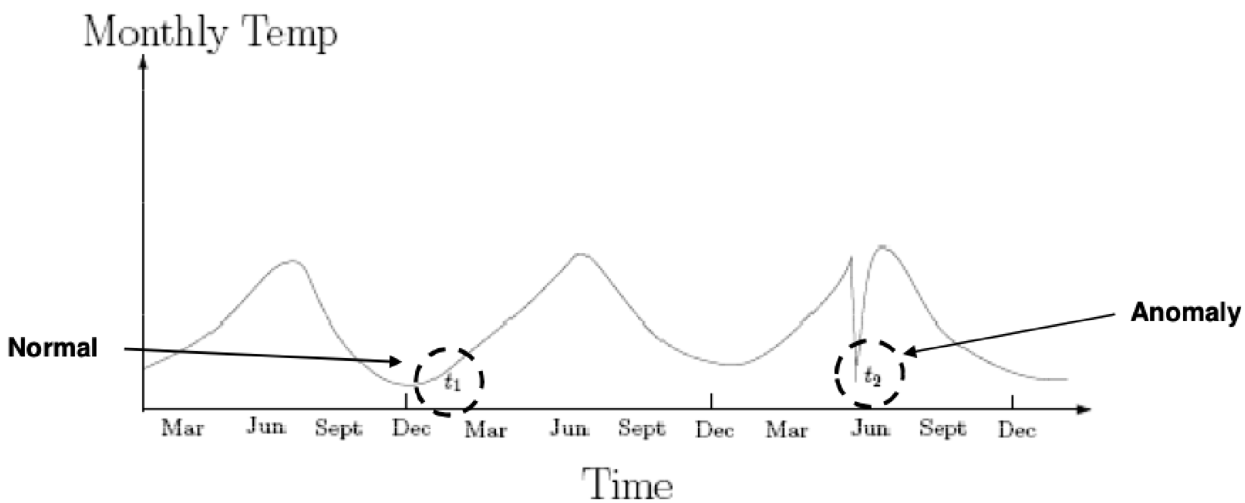
Point Anomalies

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



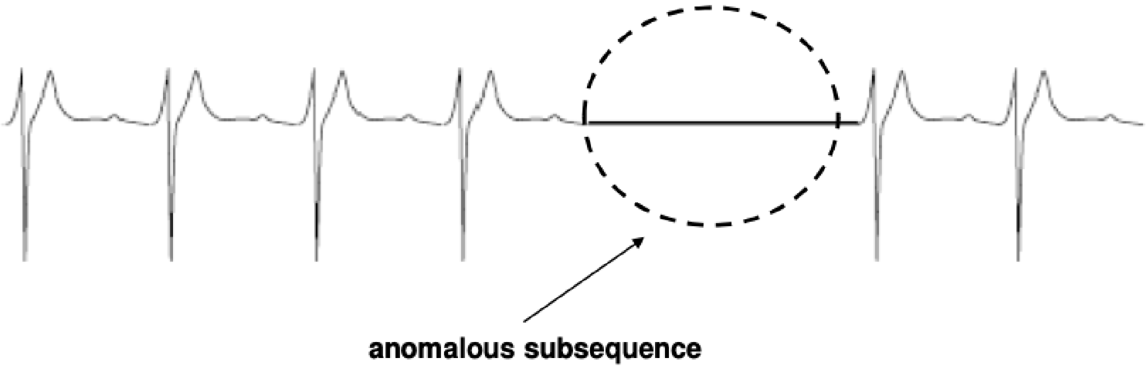
Contextual anomalies

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



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



Anomaly detection techniques

Anomaly detection

- Anomaly detection is the process of identifying rare, unusual, or suspicious data points or events that deviate significantly from the norm.

Question 2: How to detect anomalies?

How it works?

Establish a baseline

Based on historical data, we need to understand what "normal" looks like.

Identify deviations

New data is continuously compared against this established baseline.

Flag anomalies

Data points that fall outside the expected range are identified as anomalies.

Model-based vs Model-free

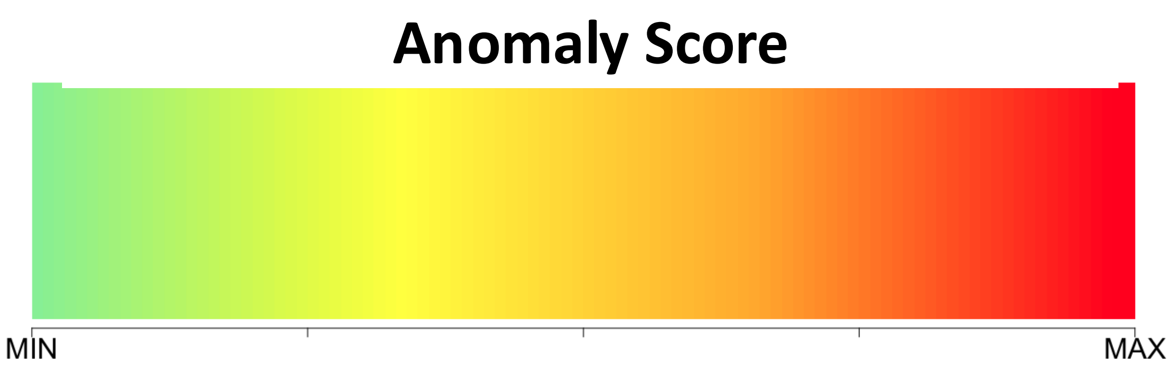
- **Model-based Approaches**
  - Model can be parametric or non-parametric
  - Anomalies are those points that don’t fit well
  - Anomalies are those points that distort the model

- **Model-free Approaches**
  - Anomalies are identified directly from the data without building a model
  - Often the underlying assumption is that most of the points in the data are normal

General Issues: Label vs Score

- **Some anomaly detection techniques provide only a binary categorization**
  - Anomaly vs. Normal

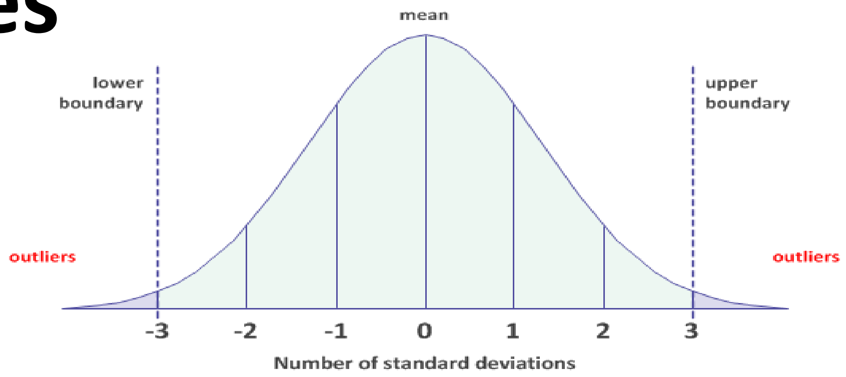
- **Other approaches measure the degree to which an object is an anomaly**
  - This allows objects to be ranked
  - Scores can also have associated meaning (e.g., statistical significance)



Anomaly Detection Techniques

- Statistical approaches
- Proximity-based
- Density-based
- Clustering-based
- Reconstruction Based
- One-class SVM
- Information theory-based

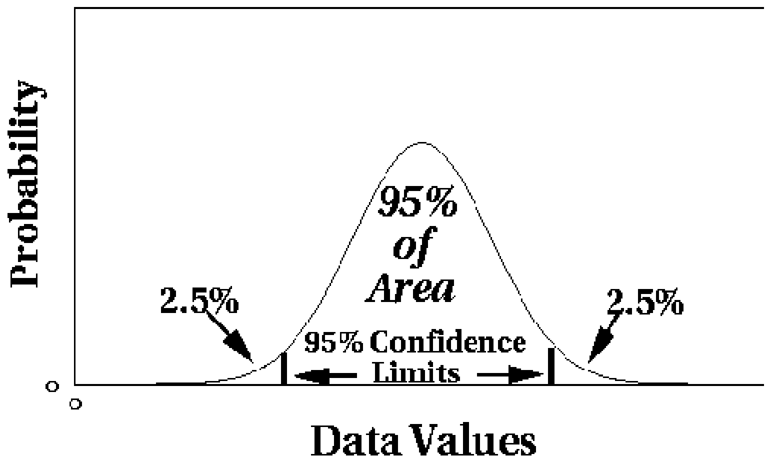
1. Statistical Approaches



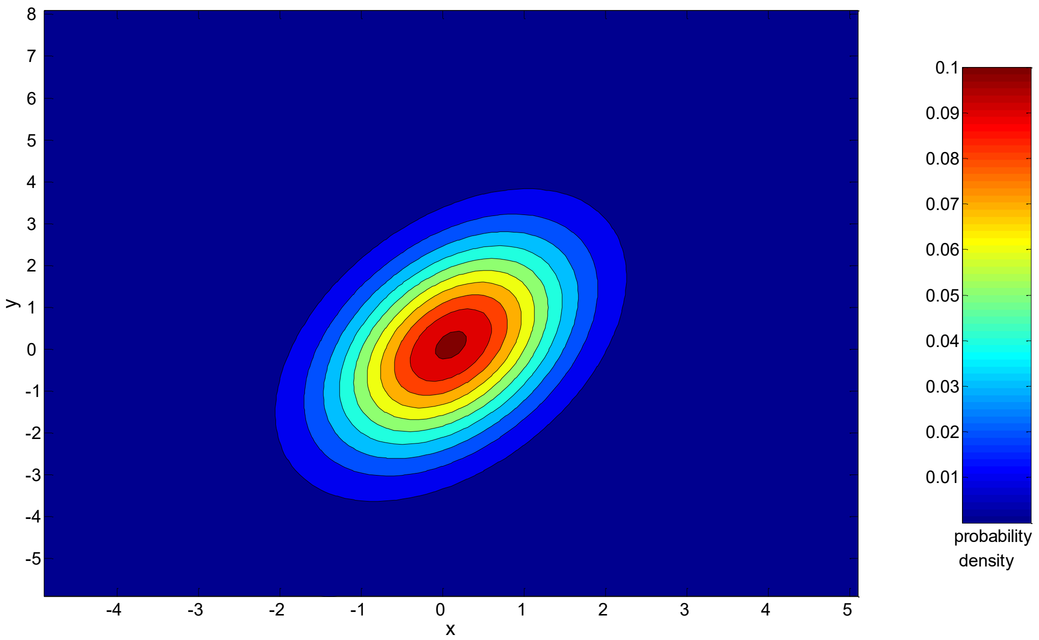
Probabilistic definition of an outlier: An outlier is an object that has a low probability with respect to a probability distribution model of the data.

- **Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)**
- **Apply a statistical test that depends on**
  - Data distribution
  - Parameters of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)

Normal Distributions



One-dimensional Gaussian



Two-dimensional Gaussian

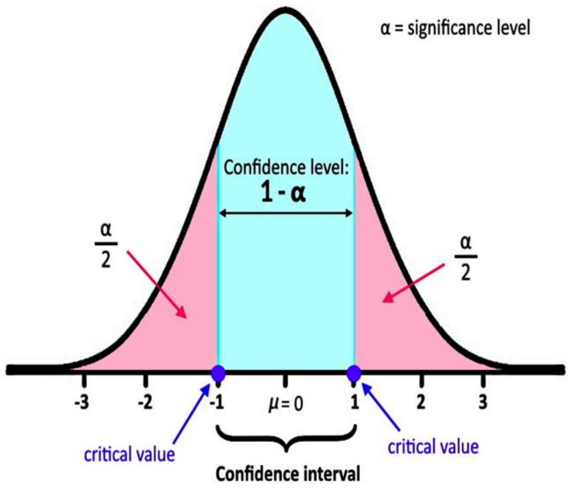
Grubbs' Test

- **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:**

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



with $t_{\alpha/(2N),N-2}$ denoting the upper critical value of the t-distribution with $N-2$ degrees of freedom and a significance level of $\alpha/(2N)$.

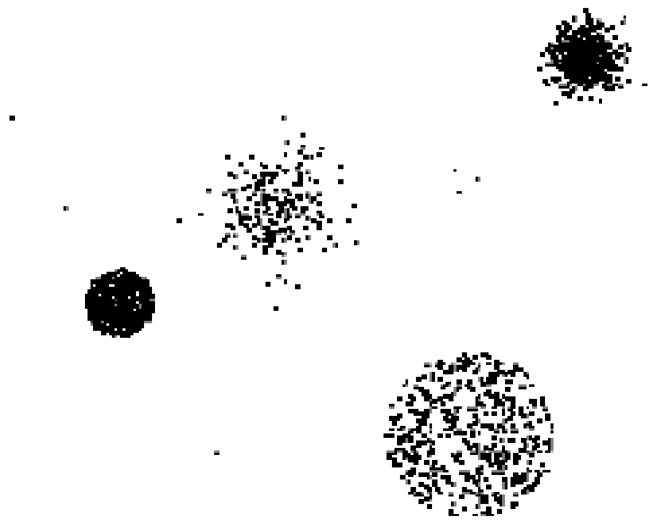
Strengths/Weaknesses

- **Strengths**
  - Firm mathematical foundation
  - Can be very efficient
  - Good results if distribution is known

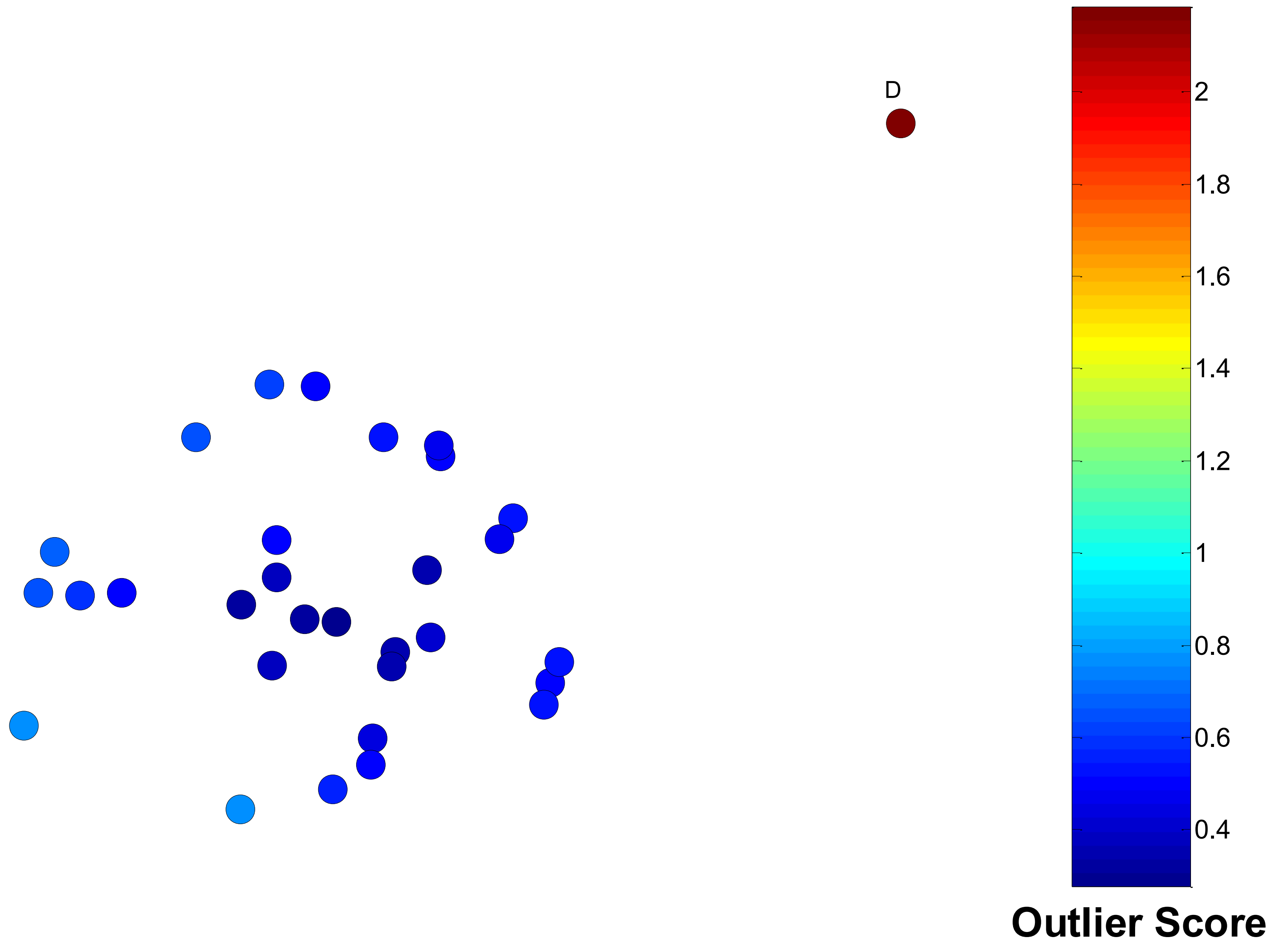
- **Weaknesses**
  - In many cases, data distribution may not be known
  - For high dimensional data, it may be difficult to estimate the true distribution
  - Anomalies can distort the parameters of the distribution

2. Distance-Based Approaches

- **The outlier score of an object is the distance to its k-th nearest neighbor**

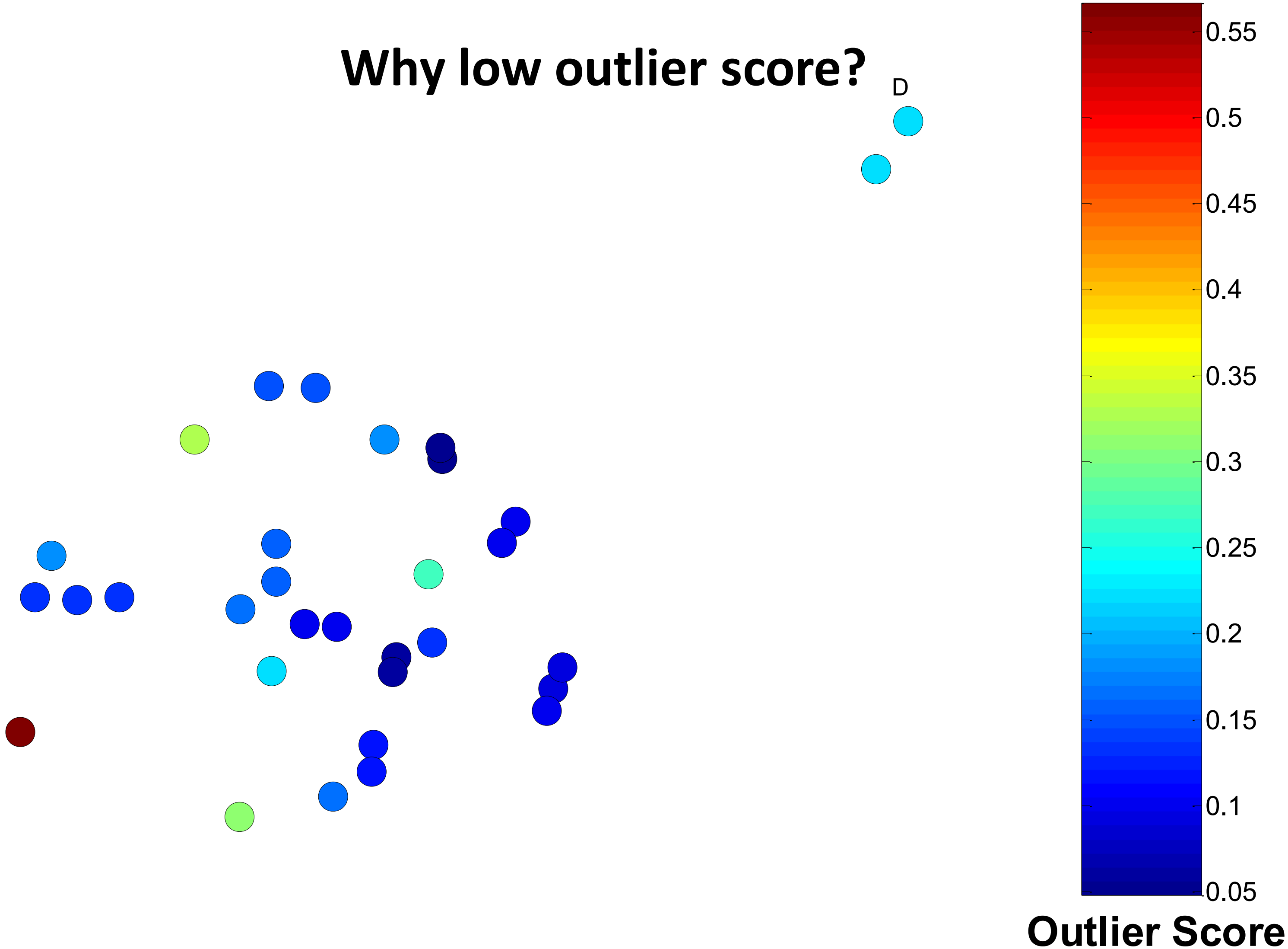


One Nearest Neighbor - One Outlier



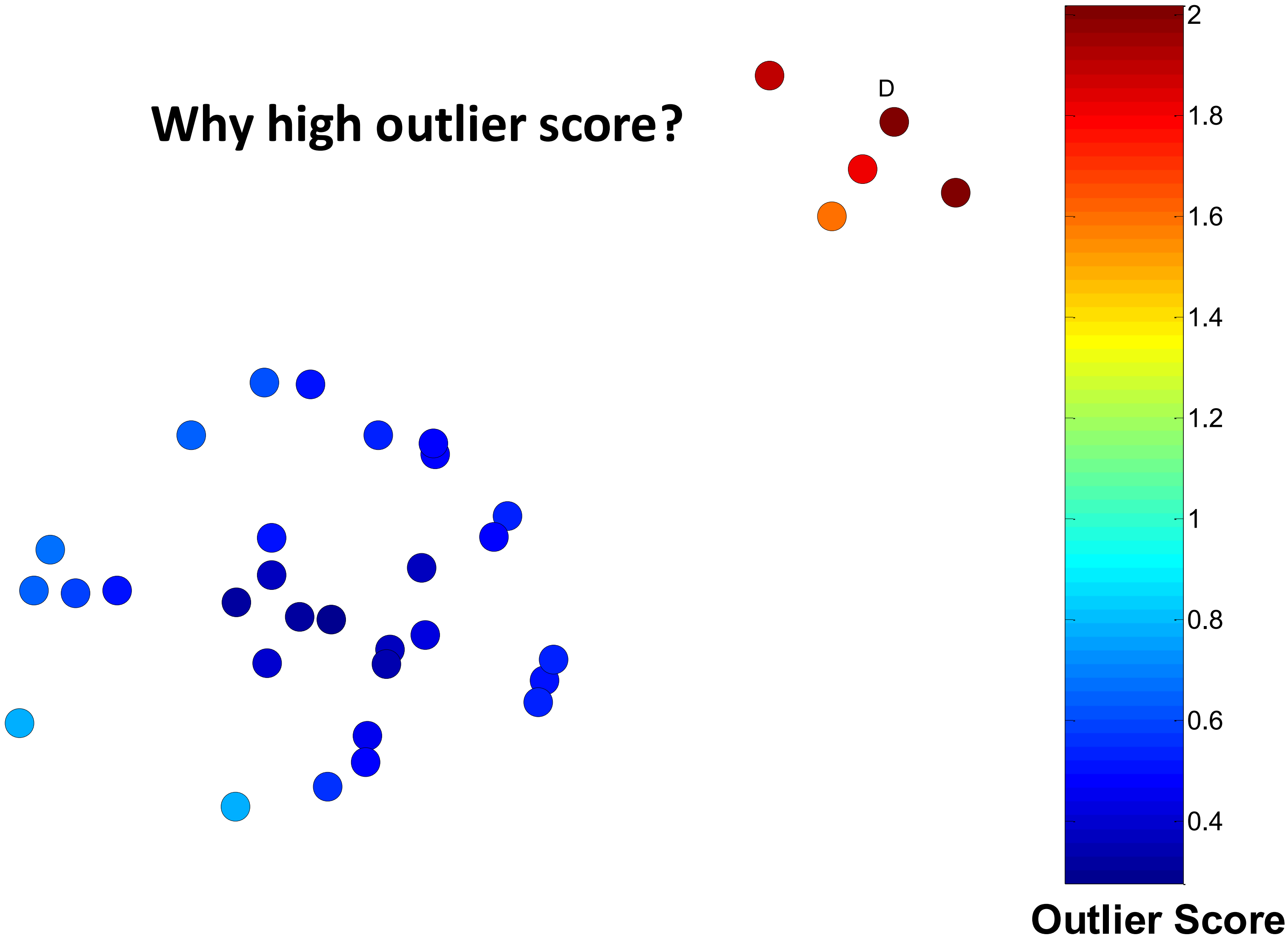
One Nearest Neighbor - Two Outliers

Why low outlier score?

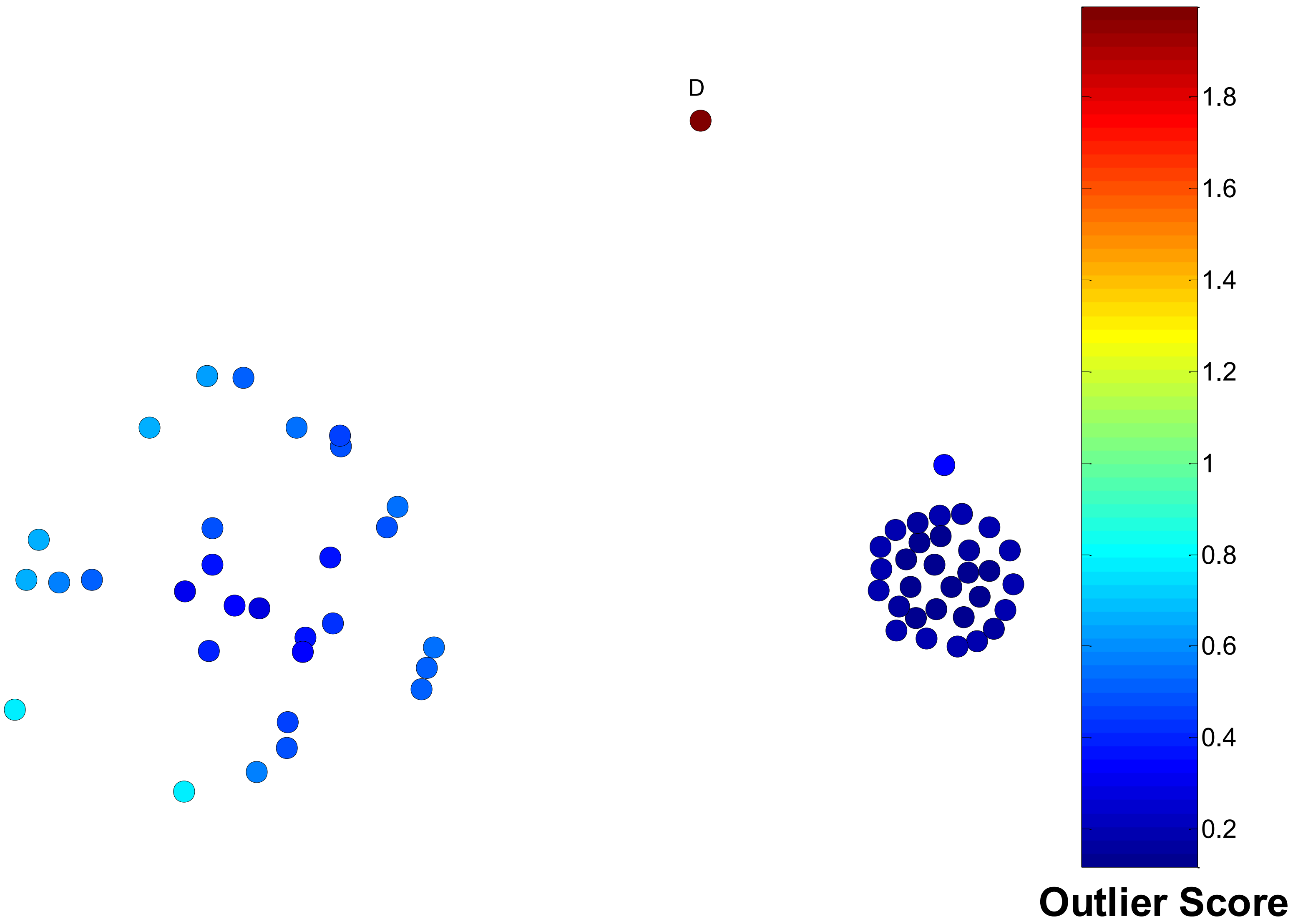


Five Nearest Neighbors - Small Cluster

Why high outlier score?



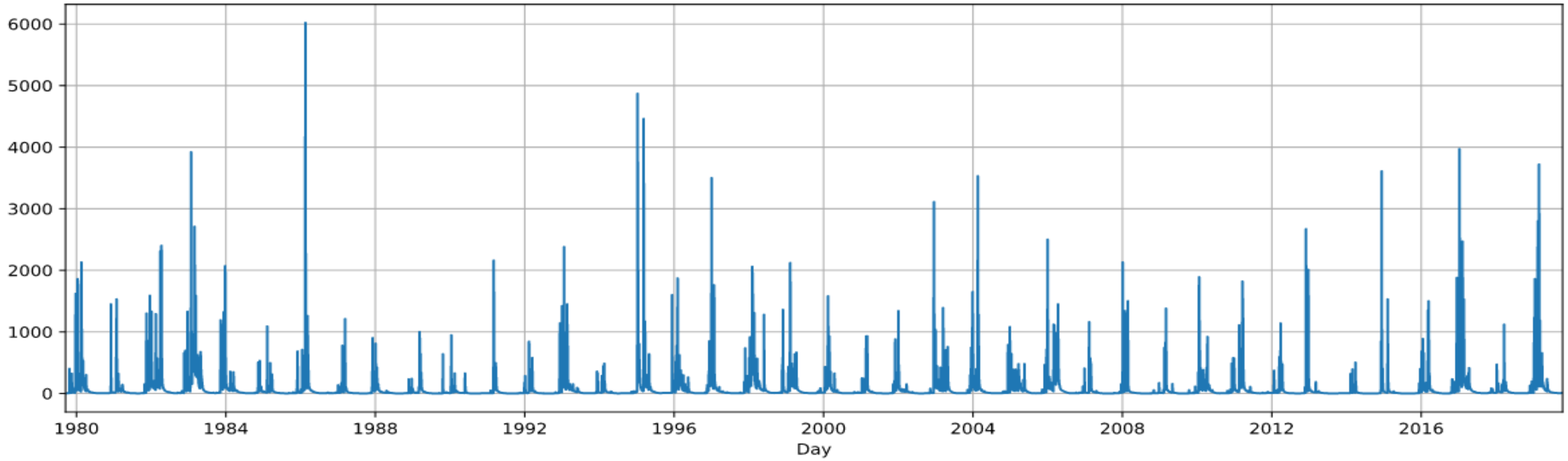
Five Nearest Neighbors - Differing Density



Strengths/Weaknesses

- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Sensitive to variations in density
- Distance becomes less meaningful in high-dimensional space

Time series data

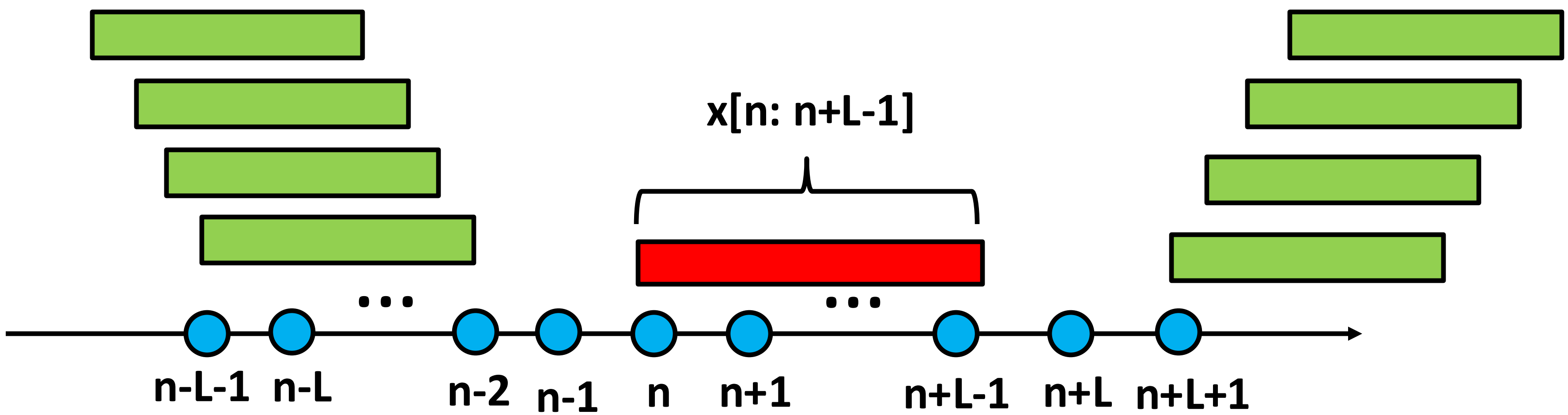


Matrix profile

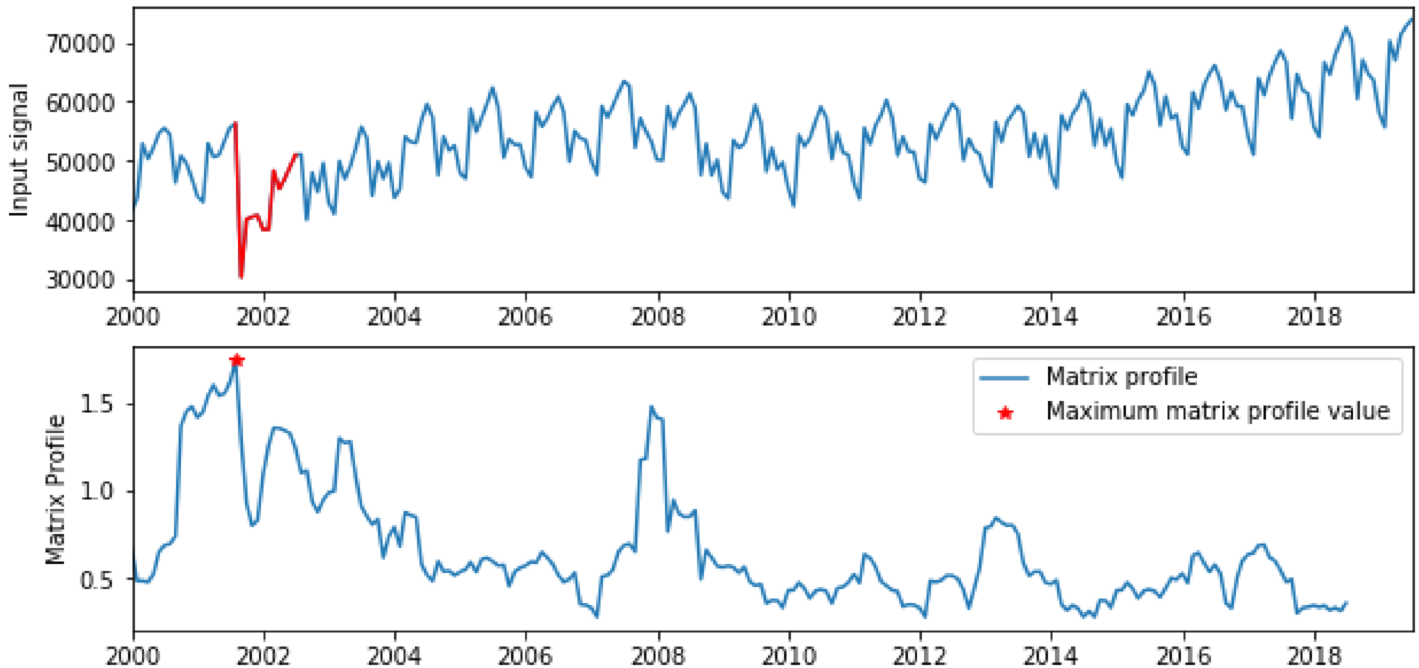
- Reminder : **Matrix profile** [Yeh et al., 2016] : given a pattern length L, compute

$$m[n] = \min_{i>n+L \text{ or } i<n-L} d(x[n:n+L-1], x[i:i+L-1])$$

- Small matrix profiles values indicate that the subsequence has been found elsewhere in the time series, suggesting that it could be a pattern



Anomaly detection based on matrix profile



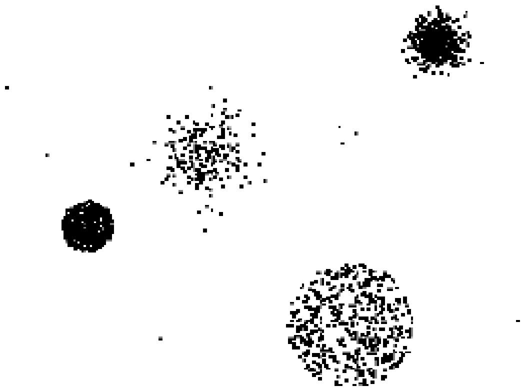
Matrix profile with window of length L = 12 months

Comments on Matrix Profile

- By examining large values on the matrix profile, anomalies can be detected
- Subsequences that are *far* from all subsequences in the signal: likely to correspond to new behaviors
- Advantages: no need for a parametric model
- Necessitates to have a rough idea of the scale of the anomaly (parameter L)

3. Density-Based Approaches

- **Density-based Outlier: The outlier score of an object is the inverse of the density around the object.**
  - The sparser, the more anomalous!
- **Density can be defined in terms of the k nearest neighbors**
  - One definition: Inverse of distance to k-th neighbor
  - Another definition: Inverse of the average distance to k neighbors
  - DBSCAN definition



Relative Density

- Consider the density of a point relative to that of its k nearest neighbors
- Let $y_1, \ldots, y_k$ be the k nearest neighbors of $\boldsymbol{x}$

$$density(\boldsymbol{x}, k) = \frac{1}{dist(\boldsymbol{x}, k)} = \frac{1}{dist(\boldsymbol{x}, \boldsymbol{y}_k)}$$

Inverse of distance from its k-th neighbor

$$relative\ density(\boldsymbol{x}, k) = \frac{\sum_{i=1}^{k} density(\boldsymbol{y}_i, k)/k}{density(\boldsymbol{x}, k)}$$

Average density of its k neighbors

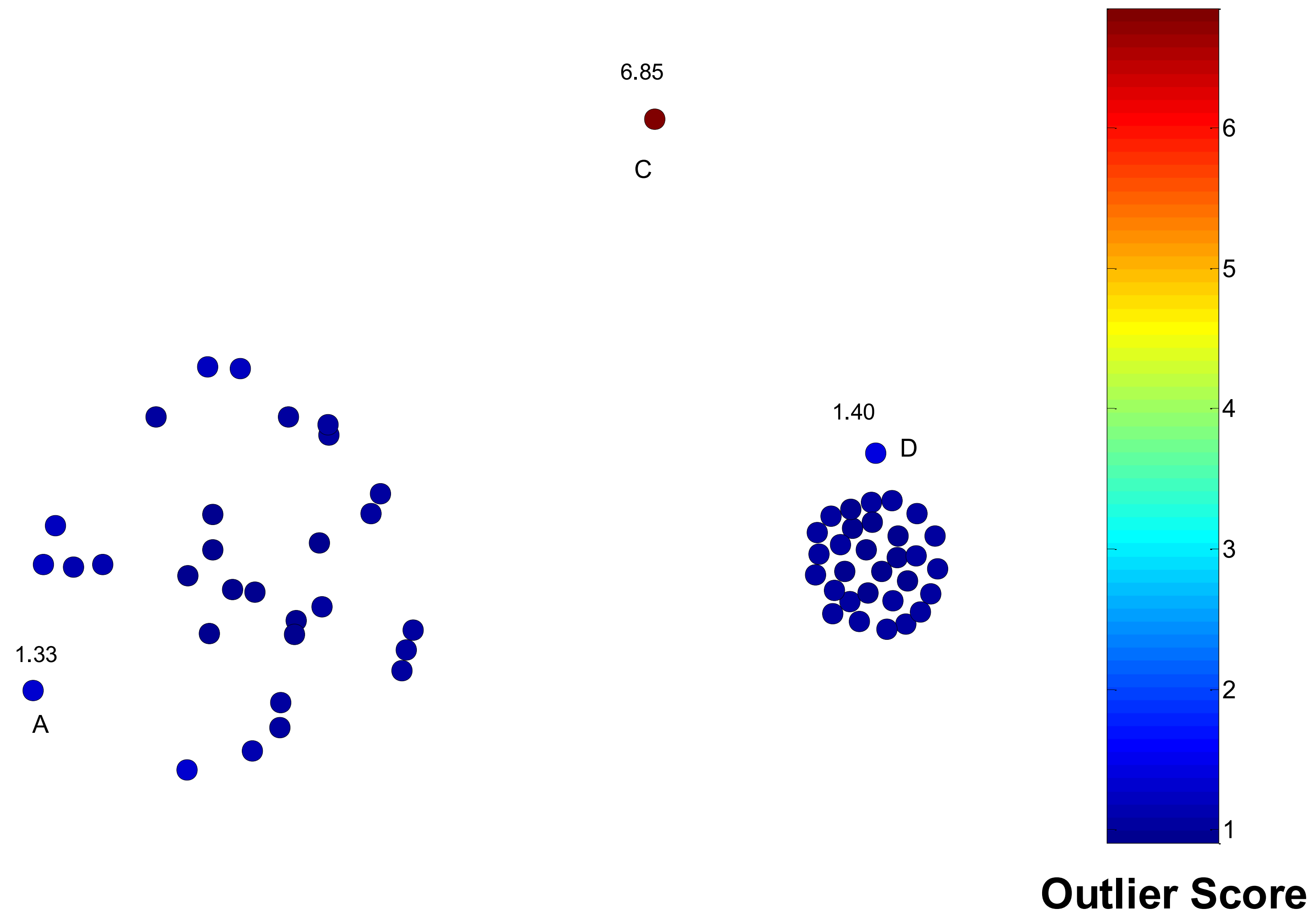
Its own density

$$= \frac{dist(\boldsymbol{x}, k)}{\sum_{i=1}^{k} dist(\boldsymbol{y}_i, k)/k}$$

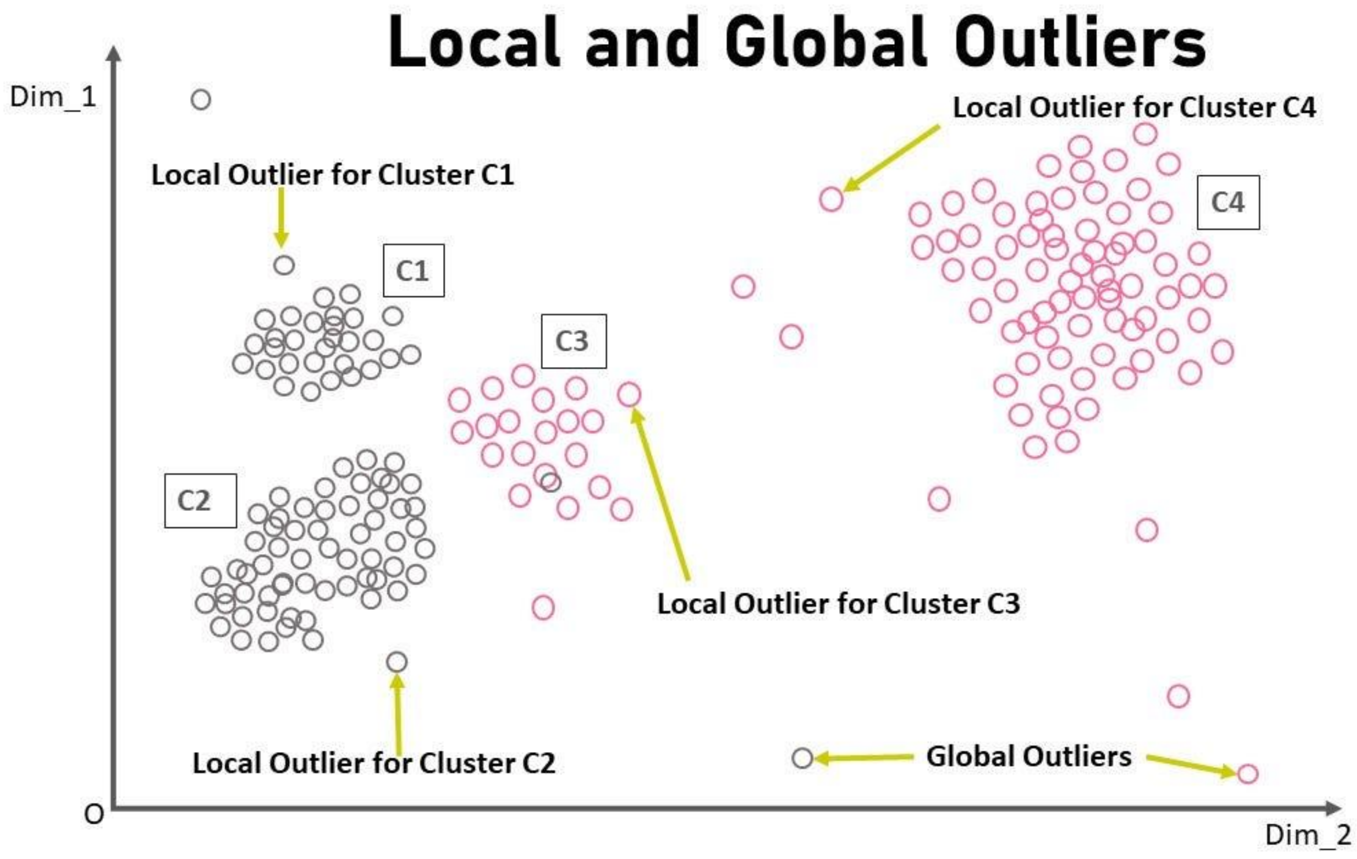
Distance from its k-th neighbor

Its neighbors' average distance from their k-th neighbors

Relative Density Outlier Scores



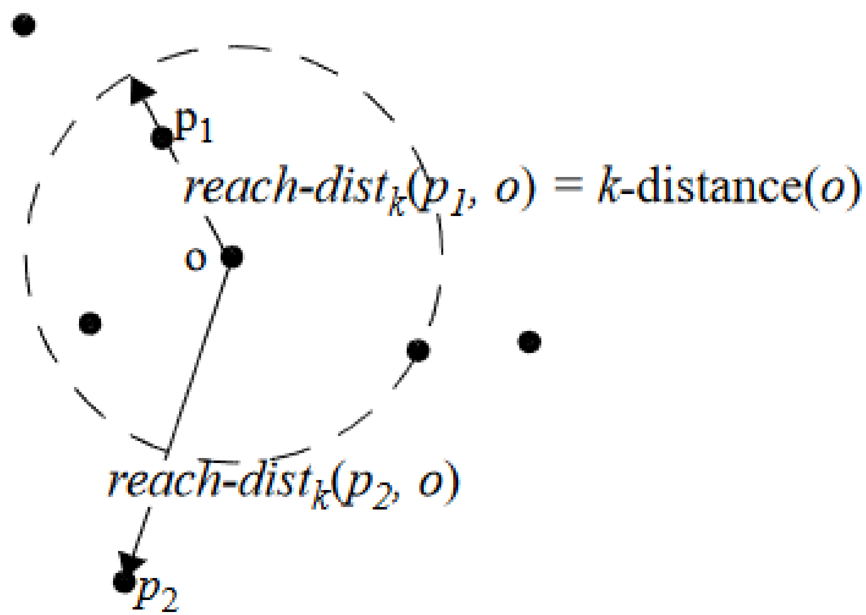
Global vs. local outliers



Density-based outlier detection: LOF

- **LOF: Local Outlier Factor**

- **K-distance**, it is the distance of a point to its *k-th* neighbor
- **Reachability distance (RD): max(k-distance(B), distance(A, B))**



LOF continued

- **Local reachability density (LRD)**: the inverse of the average RD of its neighbors.

$$LRD_k(x) = 1 \Big/ \left(\frac{\sum_{o \in N_k(x)} d_k(x, o)}{|N_k(x)|} \right)$$

- **The LOF score is computed by comparing the LRD of a record with the LRD's of its k-neighbors.**

$$LOF(x) = \frac{\sum_{o \in N_k(x)} \frac{LRD_k(o)}{LRD_k(x)}}{|N_k(x)|}$$

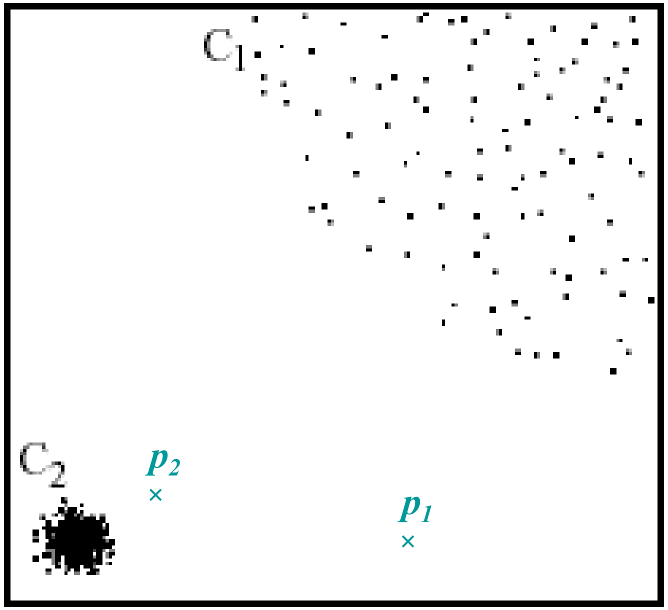
Neighbors' LRDs

Its own LRD

LOF(k) ~ 1 means **Similar density as neighbors,**
LOF(k) < 1 means **Higher density than neighbors (Inlier),**
LOF(k) > 1 means **Lower density than neighbors (Outlier)**

Anomaly detection based on LOF approach

- **Outliers are points with largest LOF value**



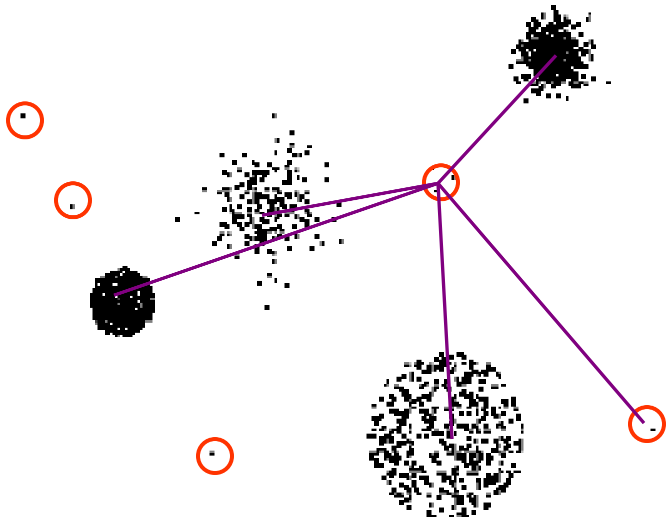
In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Strengths/Weaknesses

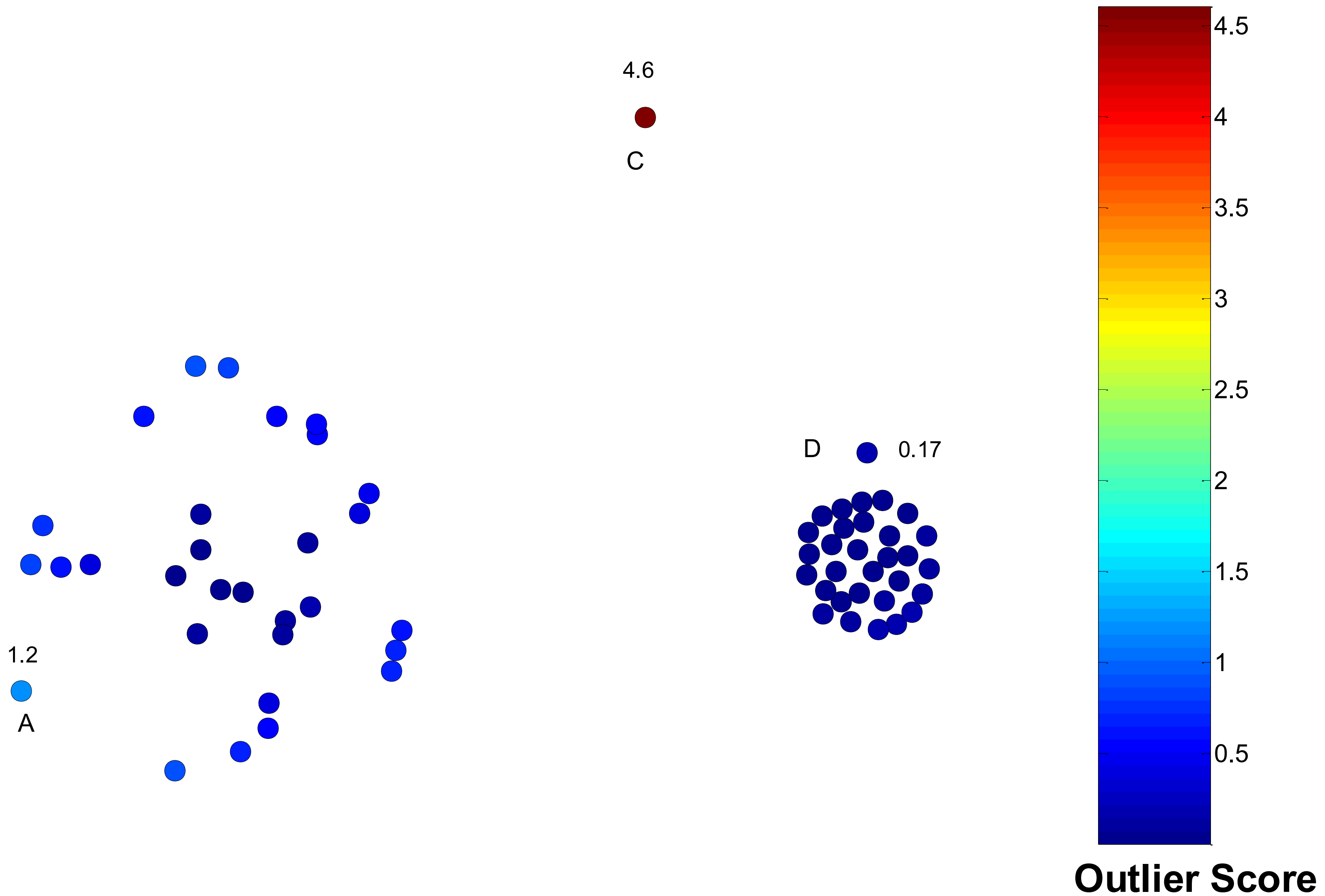
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

4. Clustering-Based Approaches

- **An object is a cluster-based outlier if it does not strongly belong to any cluster**



Distance of Points from Closest Centroids

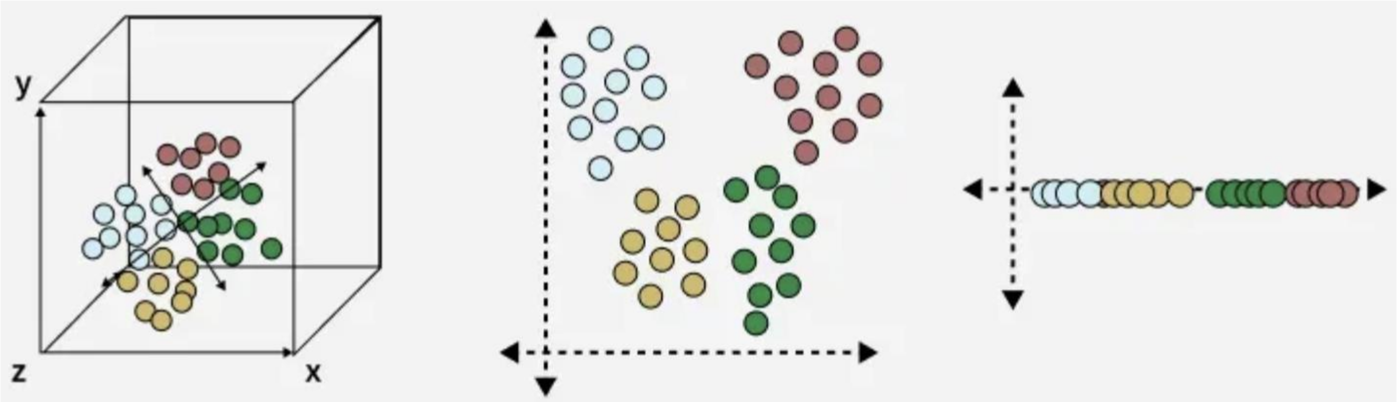


Strengths/Weaknesses

- **Strengths**
  - Simple
  - Many clustering techniques can be used
- **Weaknesses**
  - Can be difficult to decide on a clustering technique
  - Can be difficult to decide on number of clusters
  - Outliers can distort the clusters

5. Reconstruction-Based Approaches

- **Based on assumptions there are patterns in the distribution of the normal class that can be captured using lower-dimensional representations**
- **Reduce data to lower dimensional data, e.g., using Principal Components Analysis (PCA) or Auto-encoders**
- **Measure the reconstruction error for each object**
  - The difference between original and reduced dimensionality version



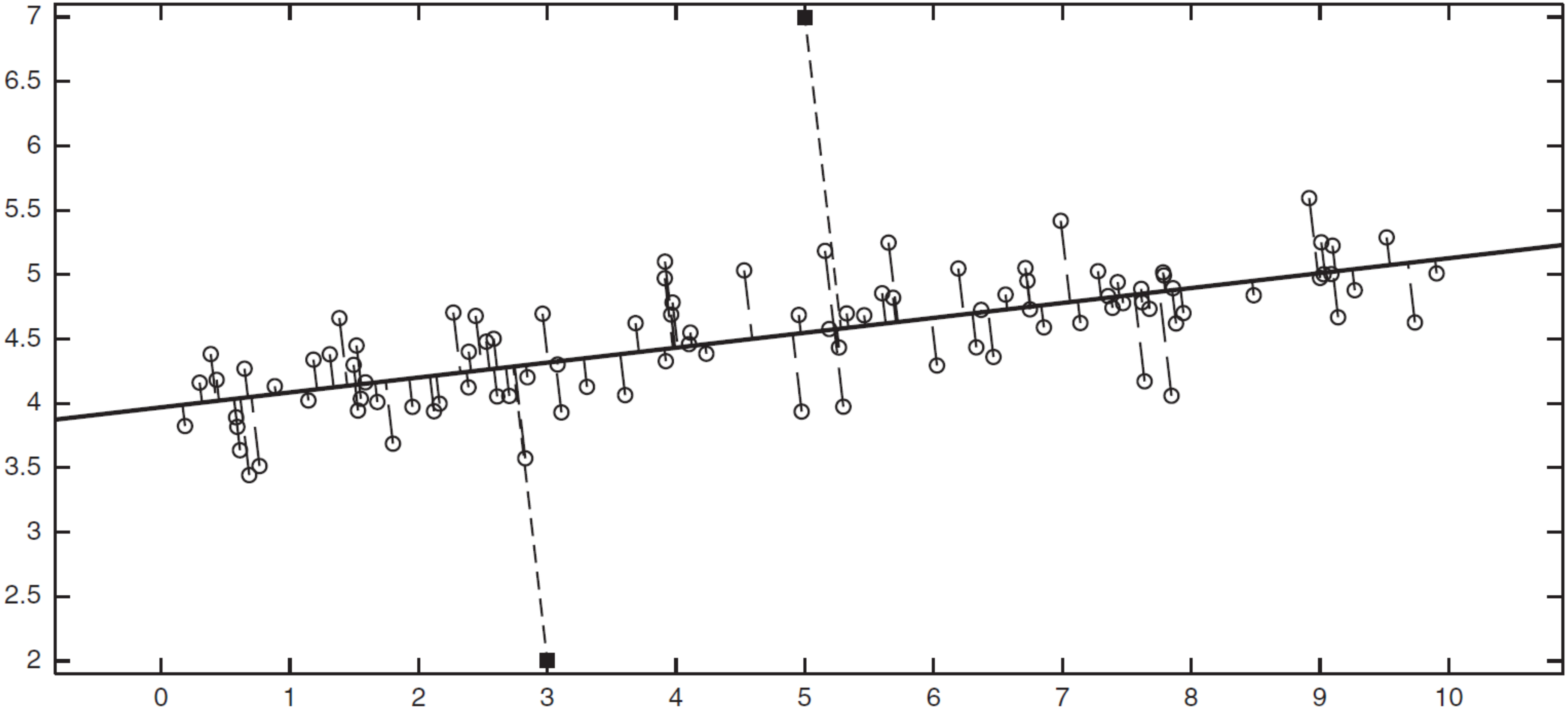
Reconstruction Error

- Let $\mathbf{x}$ be the original data object
- Find the representation of the object in a lower dimensional space
- Project the object back to the original space
- Call this object $\hat{\mathbf{x}}$

$$\text{Reconstruction Error}(\mathbf{x}) = ||\mathbf{x} - \hat{\mathbf{x}}||$$

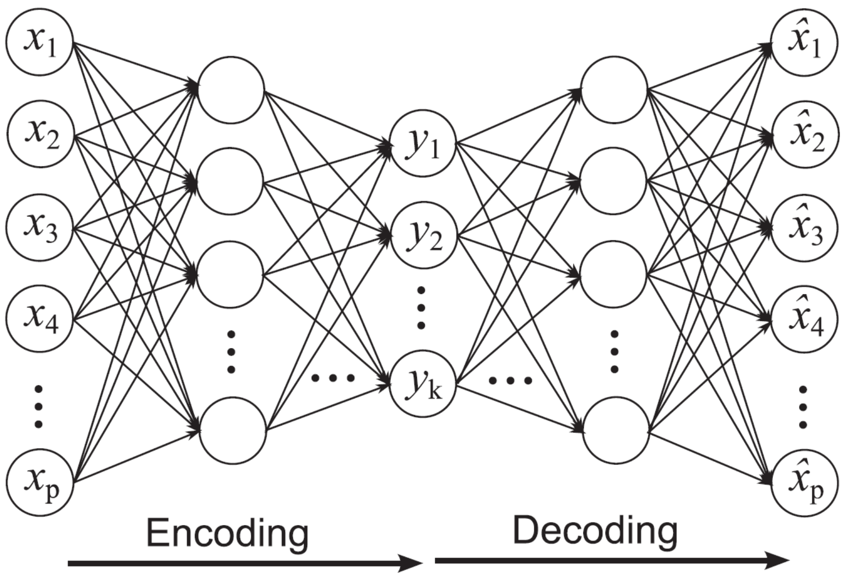
- Objects with large reconstruction errors are anomalies

Reconstruction of two-dimensional data



Basic Architecture of an Autoencoder

- **An autoencoder is a multi-layer neural network**
- **The number of input and output neurons is equal to the number of original attributes.**

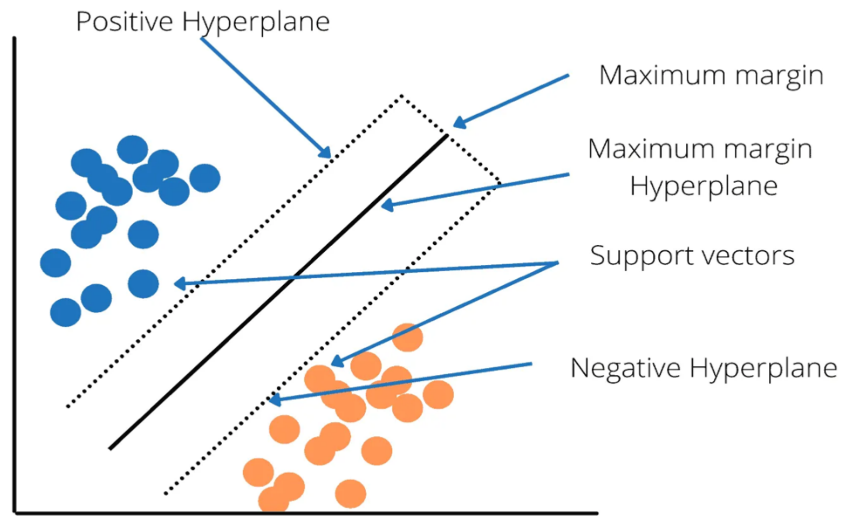


Strengths and Weaknesses

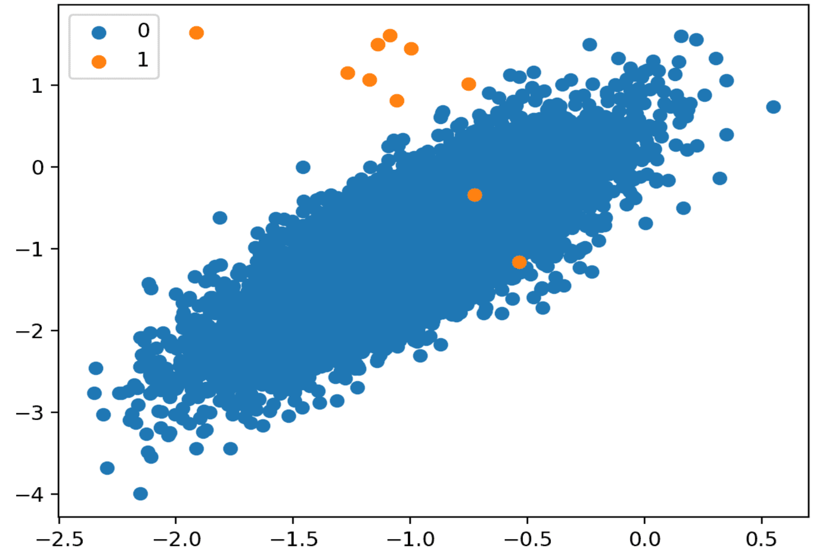
- **Does not require assumptions about distribution of normal class**
- **Can use many dimensionality reduction approaches**
- **The reconstruction error is computed in the original space**
  - This can be a problem if dimensionality is high

6. One-Class SVM (Support Vector Machine)

- Uses an SVM approach to classify normal objects
- Uses the given data to construct such a model
- This data may contain outliers
- But the data does not contain class labels
- How to build a classifier given one class?

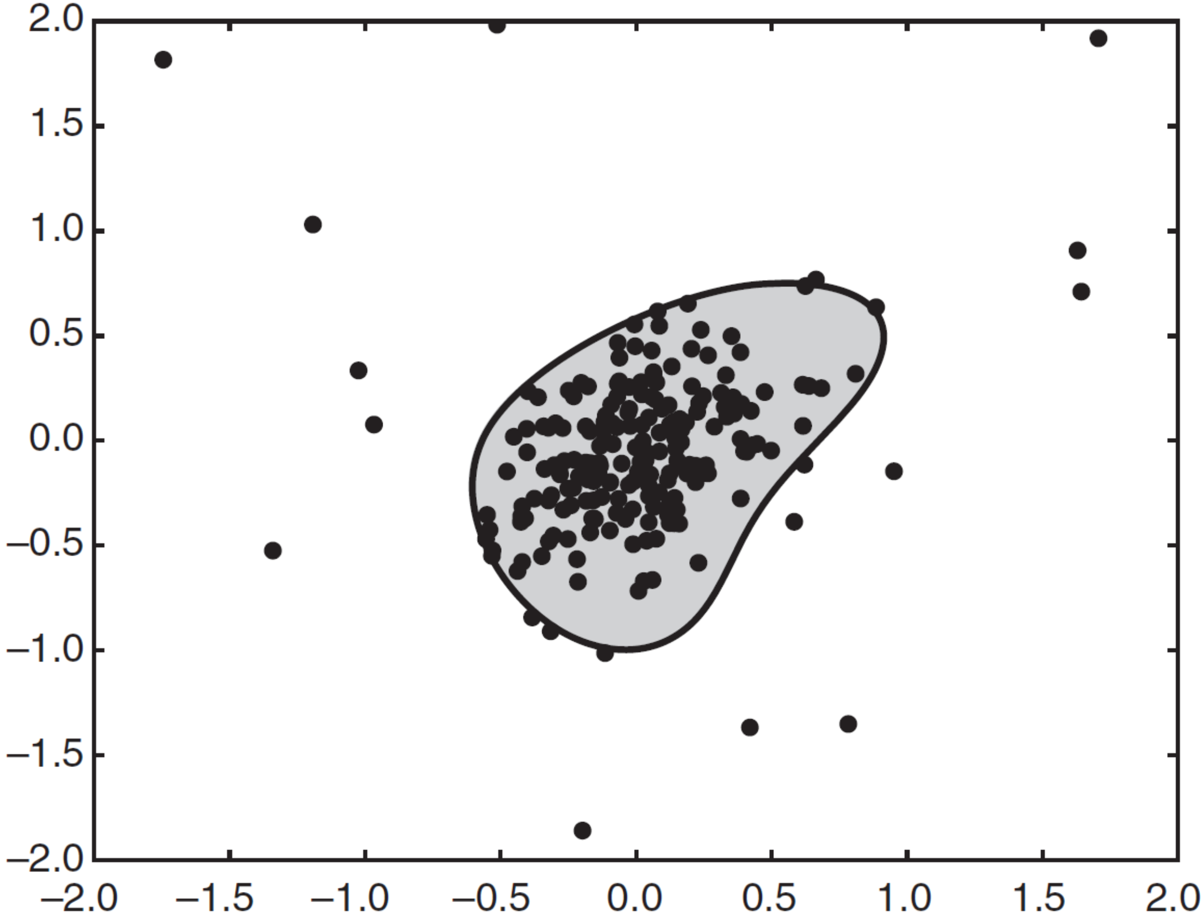


Two-Class SVM



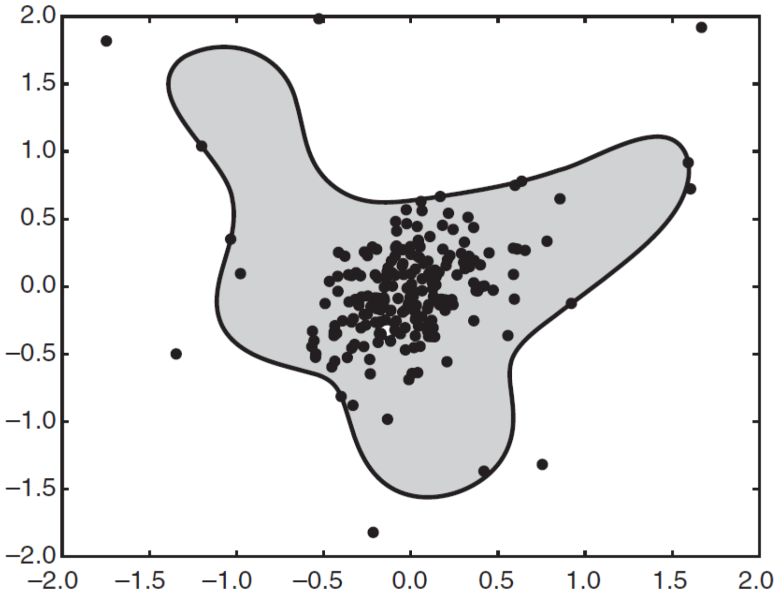
Finding Outliers with a One-Class SVM

- **Decision boundary with $\nu = 0.1$ (ν is fraction of outliers)**

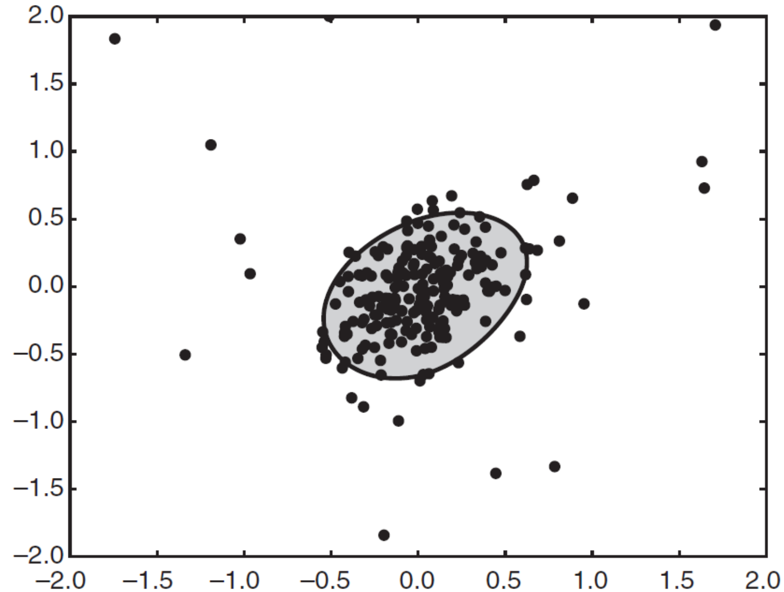


Finding Outliers with a One-Class SVM

- **Decision boundary with** $\nu = 0.05$ and $\nu = 0.2$



(a) $\nu = 0.05$.



(b) $\nu = 0.2$.

Strengths and Weaknesses

- Strong theoretical foundation
- Choice of ν is difficult
- Computationally expensive

7. Information Theoretic Approaches

- **Key idea is to measure how much information decreases when you delete an observation**

$$Gain(x) = Info(D) - Info(D \setminus x)$$

- **Anomalies should show higher gain**
- **Normal points should have less gain**

Entropy

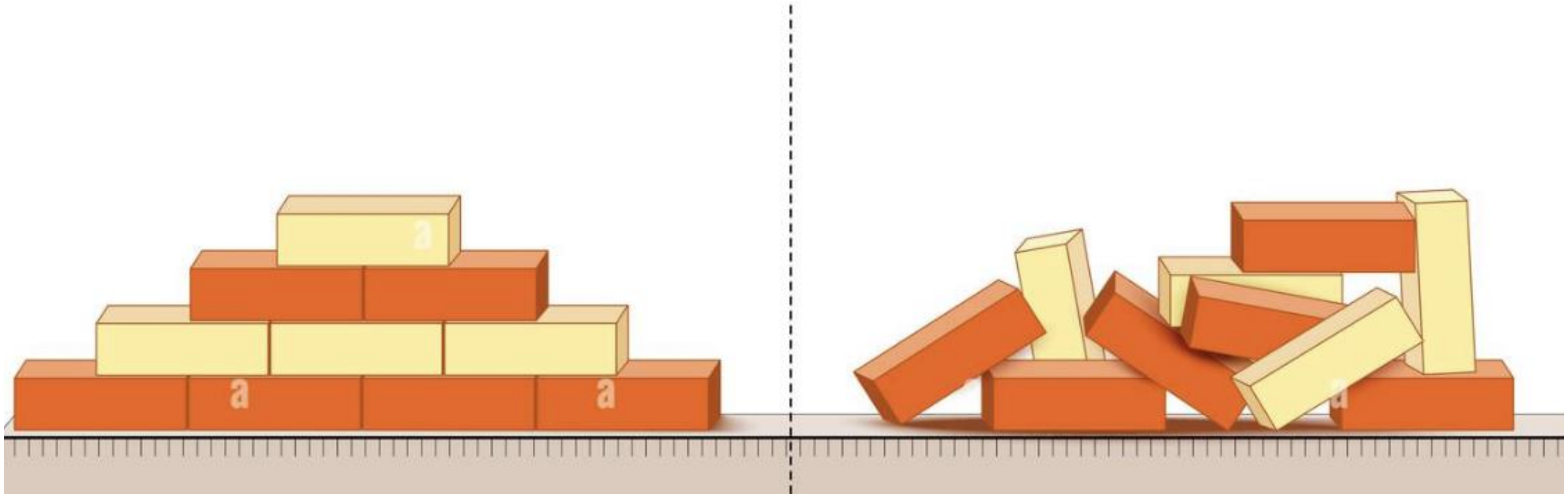
"surprise" of event

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

chance of it happening

add them all up

Low vs. high entropy



Question: which side has higher entropy?

Information Theoretic Example

- **Survey of height and weight for 100 participants**

weight	height	Frequency
low	low	20
low	medium	15
medium	medium	40
high	high	20
high	low	5

- **Eliminating last group give a gain of**
2.08 – 1.89 = 0.19

Strengths and Weaknesses

- Solid theoretical foundation
- Theoretically applicable to all kinds of data
- Difficult and computationally expensive to implement in practice

Anomaly detection applications

Applications

- 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
- Lots more!

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**

Anomaly detection on real network data

- Three groups of features
  - Basic features of individual TCP connections
    - source & destination IP *Features 1 & 2*
    - source & destination port *Features 3 & 4*
    - Protocol *Feature 5*
    - Duration *Feature 6*
    - Bytes per packets *Feature 7*
    - number of bytes *Feature 8*

dst …	service …	flag
h1	http	S0
h1	http	S0
h1	http	S0
h2	http	S0
h4	http	S0
h2	ftp	S0

existing features useless

syn flood

normal

dst …	service …	flag	%S0
h1	http	S0	70
h1	http	S0	72
h1	http	S0	75
h2	http	S0	0
h4	http	S0	0
h2	ftp	S0	0

construct features with high information gain

  - **Time based features**
    - For the same source (destination) IP address, number of unique destination (source) IP addresses inside the network *in last T seconds – Features 9 (13)*
    - Number of connections from source (destination) IP to the same destination (source) port *in last T seconds – Features 11 (15)*
  - Connection based features
    - For the same source (destination) IP address, number of unique destination (source) IP addresses inside the network *in last N connections - Features 10 (14)*
    - Number of connections from source (destination) IP to the same destination (source) port *in last N connections - Features 12 (16)*

Typical anomaly detection output

score	srcIP	sPort	dstIP	dPort	protocc	flags	packets	bytes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
37674.69	63.150.X.253	1161	128.101.X.29	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.59	0	0	0	0	0
26676.62	63.150.X.253	1161	160.94.X.134	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.59	0	0	0	0	0
24323.55	63.150.X.253	1161	128.101.X.185	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
21169.49	63.150.X.253	1161	160.94.X.71	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
19525.31	63.150.X.253	1161	160.94.X.19	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
19235.39	63.150.X.253	1161	160.94.X.80	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
17679.1	63.150.X.253	1161	160.94.X.220	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.81	0	0.58	0	0	0	0	0
8183.58	63.150.X.253	1161	128.101.X.108	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.58	0	0	0	0	0
7142.98	63.150.X.253	1161	128.101.X.223	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
5139.01	63.150.X.253	1161	128.101.X.142	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
4048.49	142.150.Y.101	0	128.101.X.127	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
4008.35	200.250.Z.20	27016	128.101.X.116	4629	17	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3657.23	202.175.Z.237	27016	128.101.X.116	4148	17	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3450.9	63.150.X.253	1161	128.101.X.62	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
3327.98	63.150.X.253	1161	160.94.X.223	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2796.13	63.150.X.253	1161	128.101.X.241	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2693.88	142.150.Y.101	0	128.101.X.168	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
2683.05	63.150.X.253	1161	160.94.X.43	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2444.16	142.150.Y.236	0	128.101.X.240	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
2385.42	142.150.Y.101	0	128.101.X.45	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
2114.41	63.150.X.253	1161	160.94.X.183	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
2057.15	142.150.Y.101	0	128.101.X.161	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1919.54	142.150.Y.101	0	128.101.X.99	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1634.38	142.150.Y.101	0	128.101.X.219	2048	1	16	[2,4)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1596.26	63.150.X.253	1161	128.101.X.160	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
1513.96	142.150.Y.107	0	128.101.X.2	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1389.09	63.150.X.253	1161	128.101.X.30	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
1315.88	63.150.X.253	1161	128.101.X.40	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.82	0	0.57	0	0	0	0	0
1279.75	142.150.Y.103	0	128.101.X.202	2048	1	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1237.97	63.150.X.253	1161	160.94.X.32	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0
1180.82	63.150.X.253	1161	128.101.X.61	1434	17	16	[0,2)	[0,1829)	0	0	0	0	0	0	0	0	0.83	0	0.56	0	0	0	0	0

- Anomalous connections that correspond to the “slammer” worm
- Anomalous connections that correspond to the ping scan
- Connections corresponding to Univ. Minnesota machines connecting to “half-life” game servers

End of Lecture 15