AIN2002
Introduction to Data Science

Outlier/Anomaly Detection

Dr. Fatih KAHRAMAN
fatih.kahraman@bau.edu.tr

Anomaly Detection

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

- **Applications**
  - Credit card fraud detection
  - telecommunication fraud detection
  - network intrusion detection
  - fault detection
  - many more

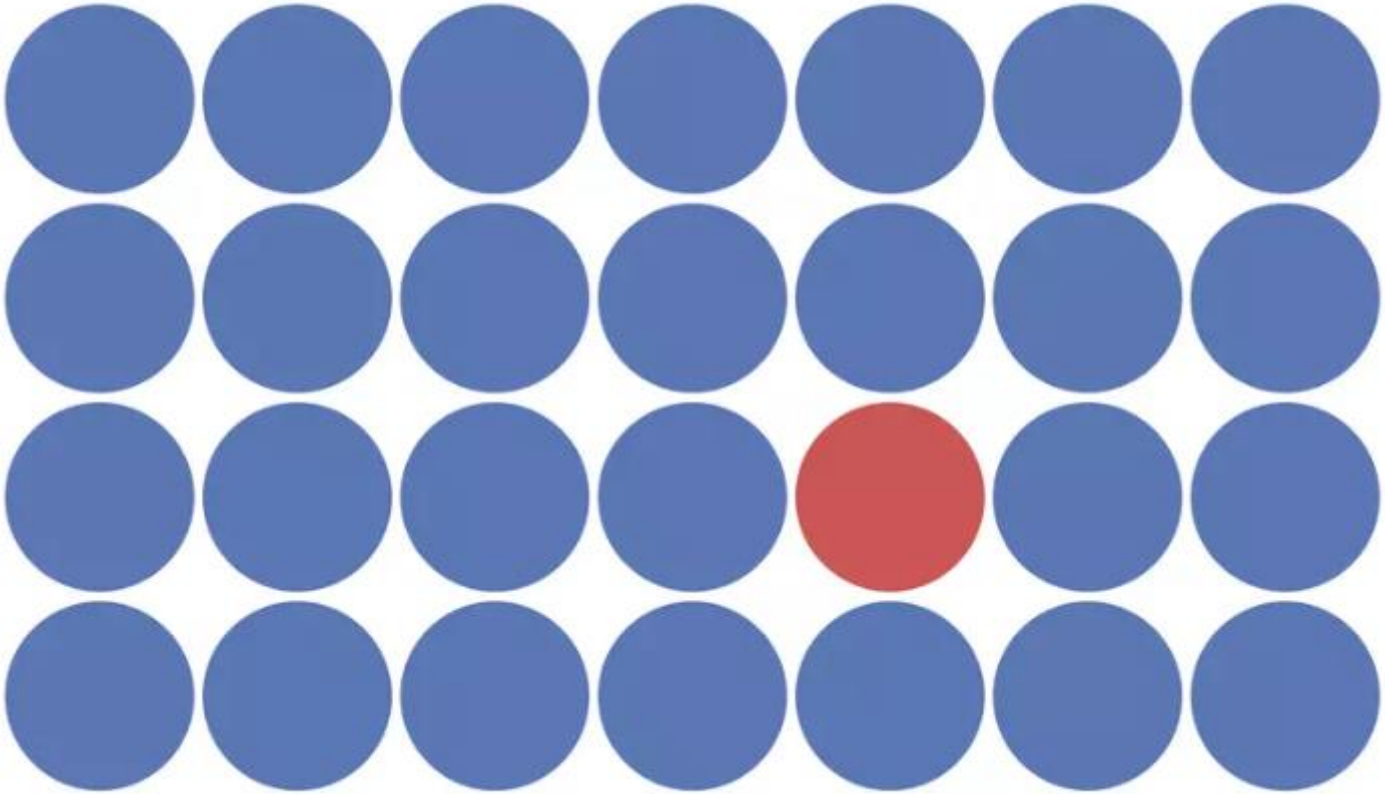
Anomaly Detection

- **Challenges**
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging (just like for clustering)

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

Anomaly Detection

> **What is Anomaly?**

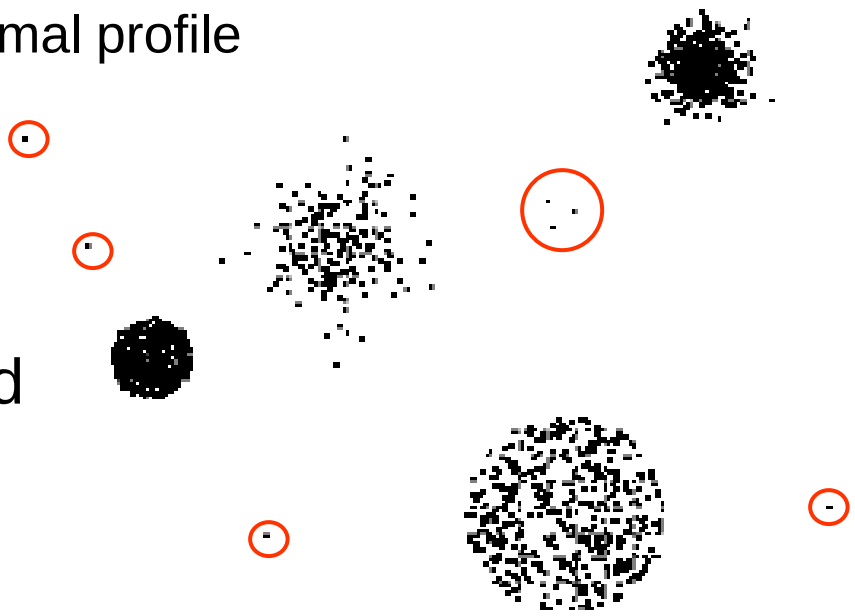


Anomaly: Deviation in an event from its expected value within a group

Anomaly Detection Steps

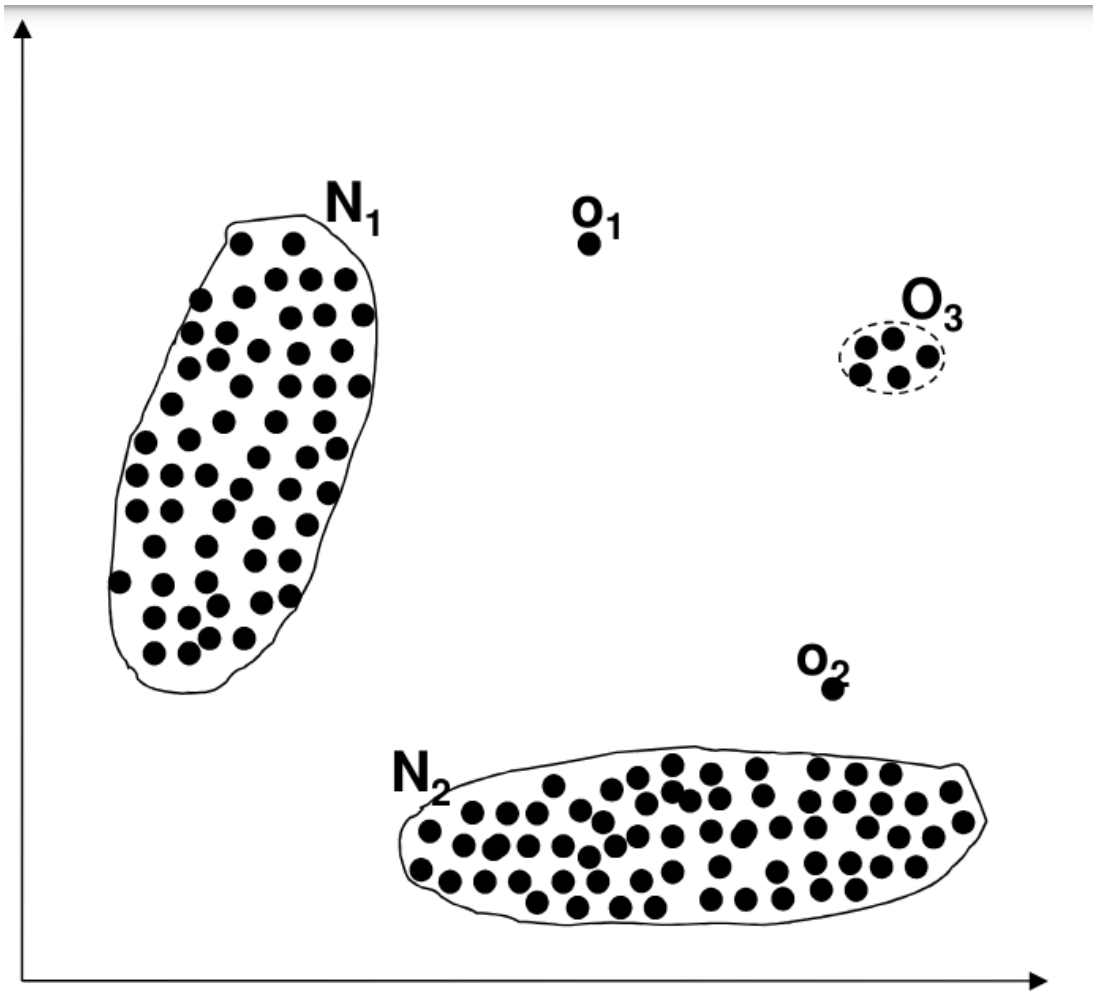
- **General Steps**
  - Build a profile of the "**normal**" behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the "**normal**" profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile

- **Types of anomaly detection methods**
  - Graphical & Statistical-based
  - Distance-based
  - Model-based



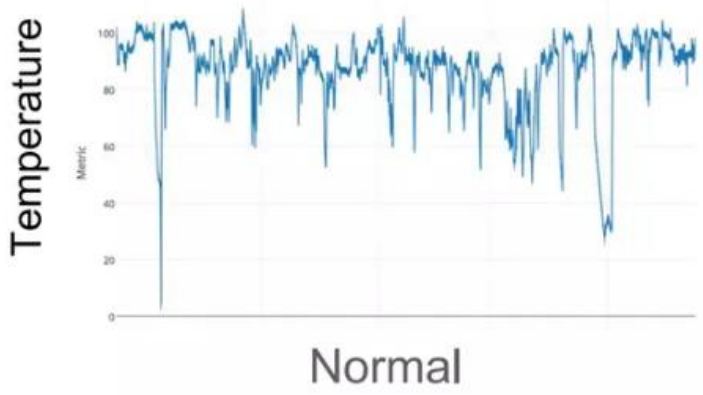
Anomaly Detection Steps

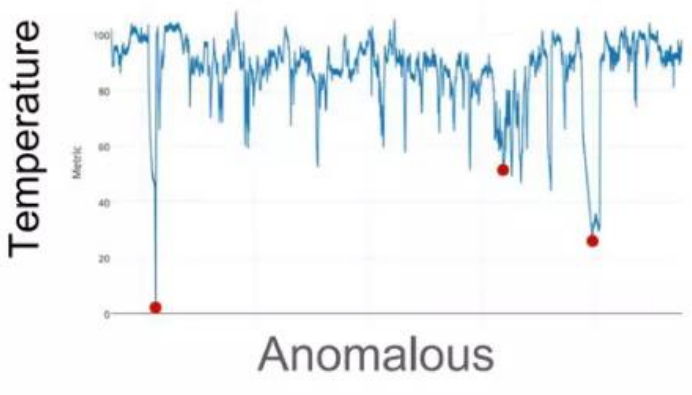
- **An individual data instance is anomalous w.r.t. the Data**

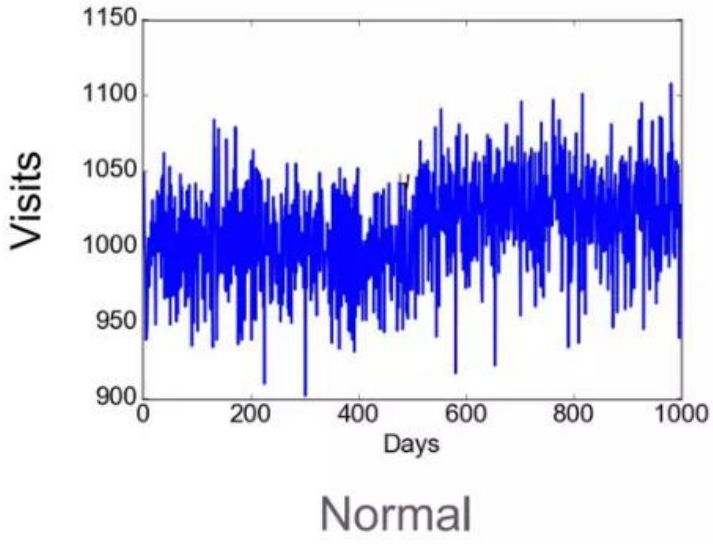


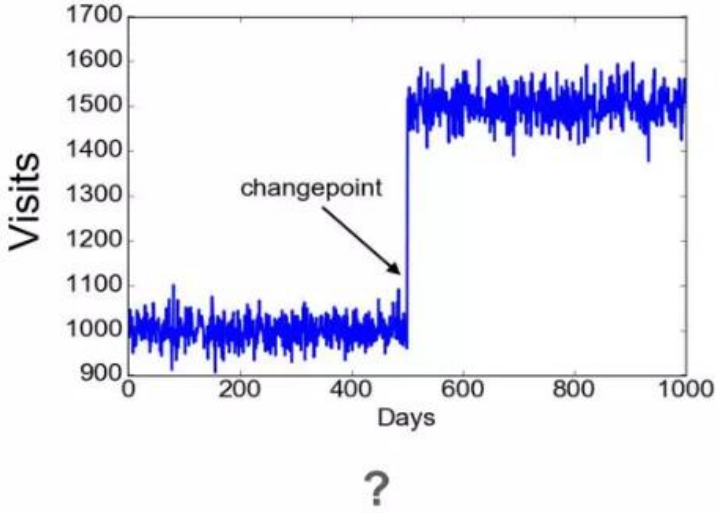
Identification of Anomalies

> **Notice that…**



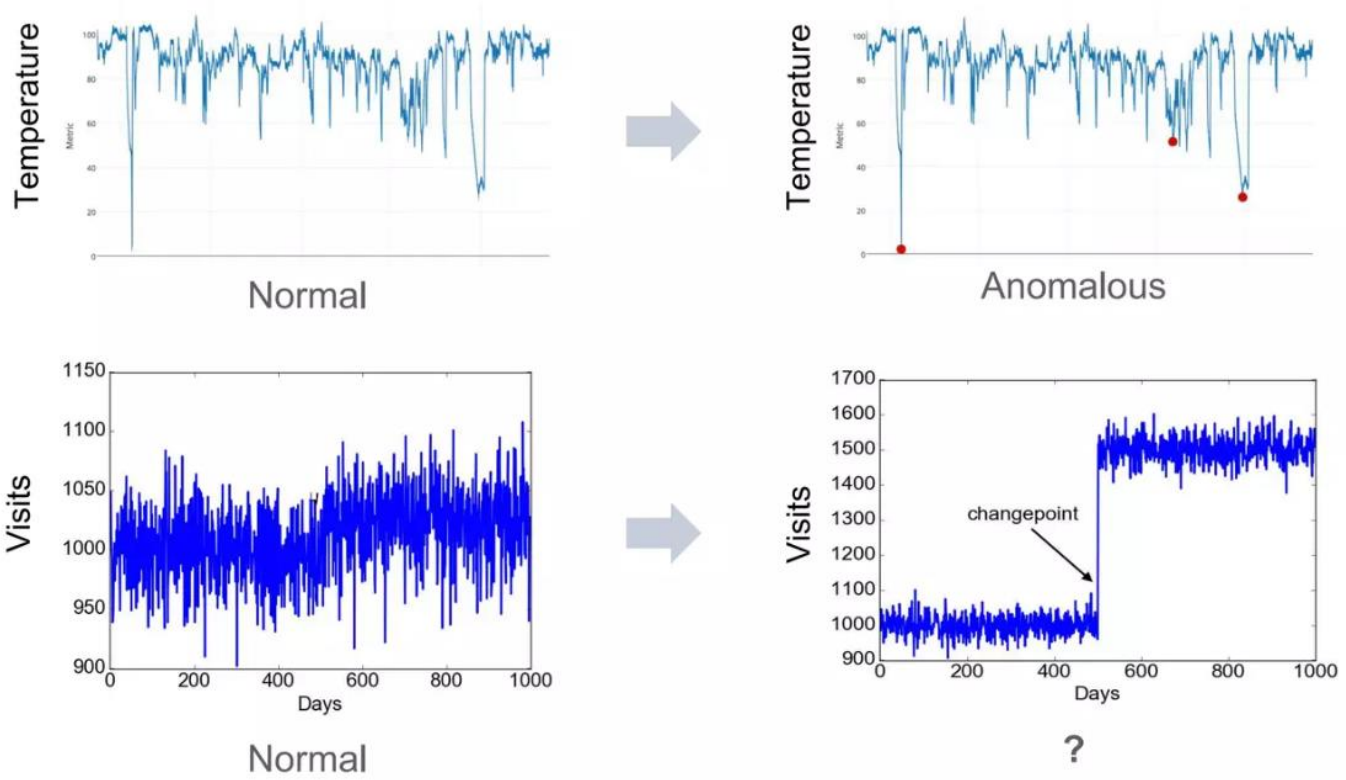






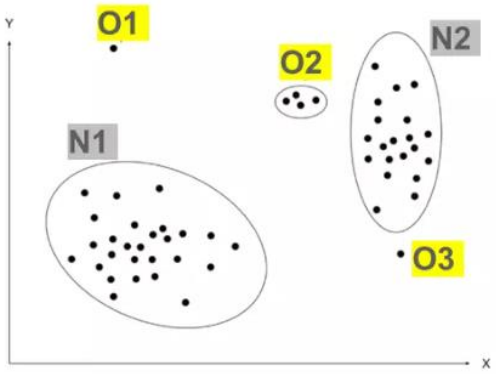
Identification of Anomalies

- **Why is it difficult?**
  - Normal behavior changes over time
  - Anomalies differ with domains
  - Noise tends to be similar to anomalies

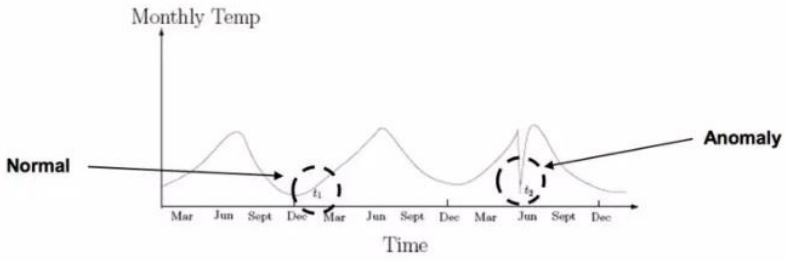


Types of Anomalies

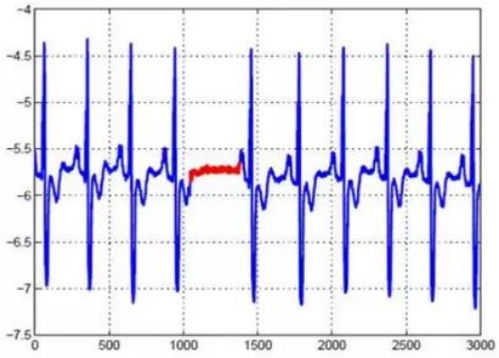
Point anomaly



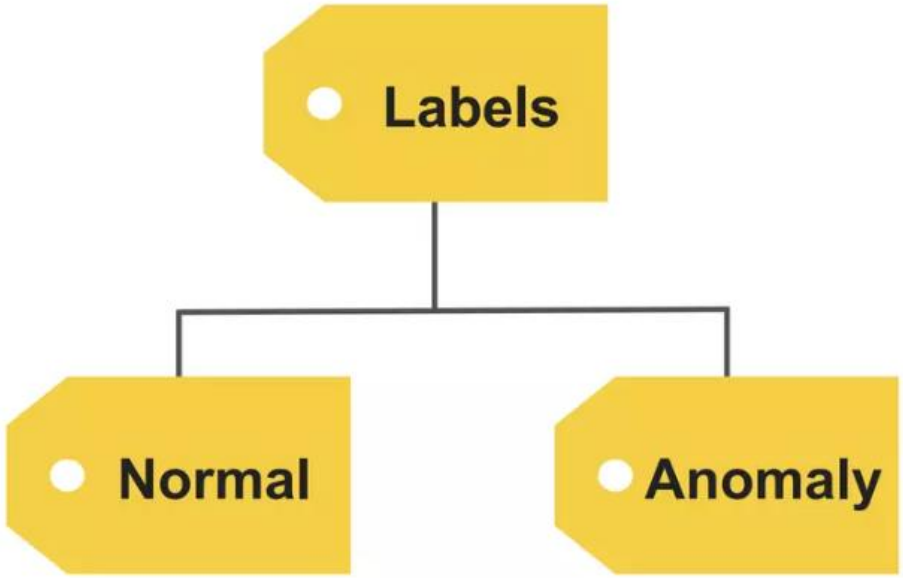
Contextual anomaly



Collective anomaly



Anomaly Score!



Target: Class

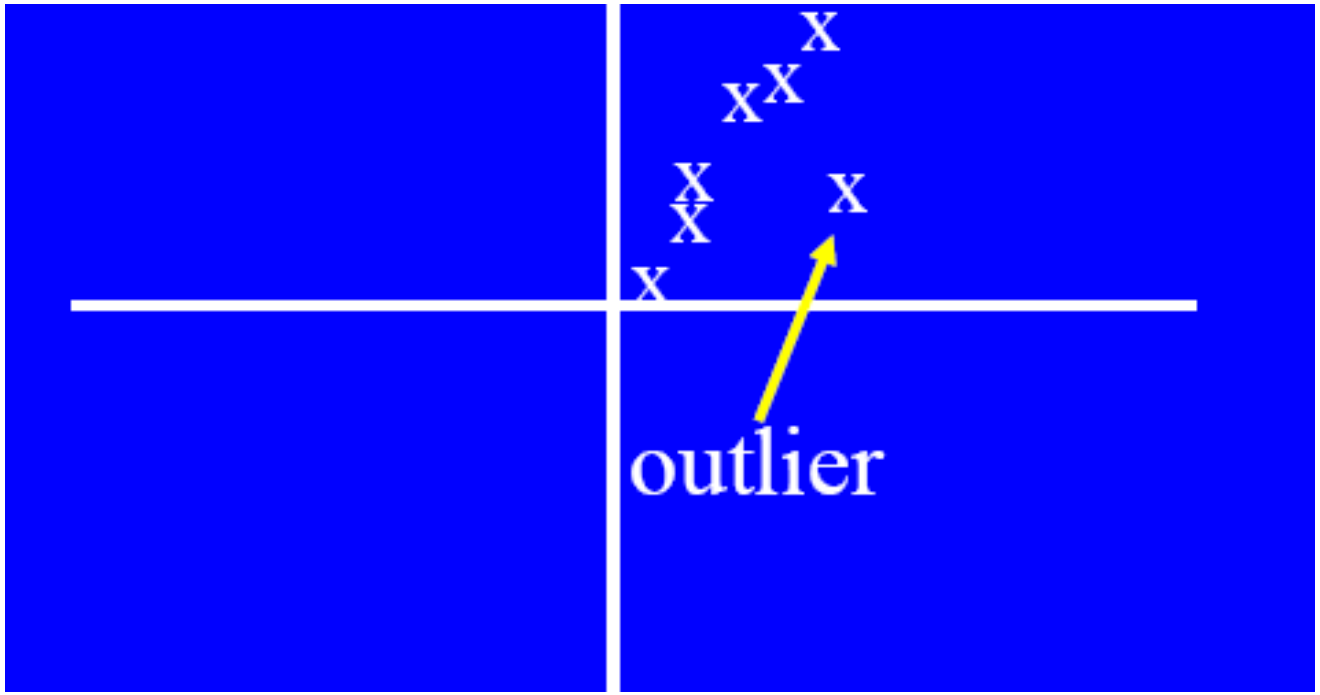
Target: Score

Graphical Approach

- Use visualization tools to observe the data
- Provide alternate views of data for manual inspection
- Anomalies are detected visually
- **Advantage**
  - 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 Approach

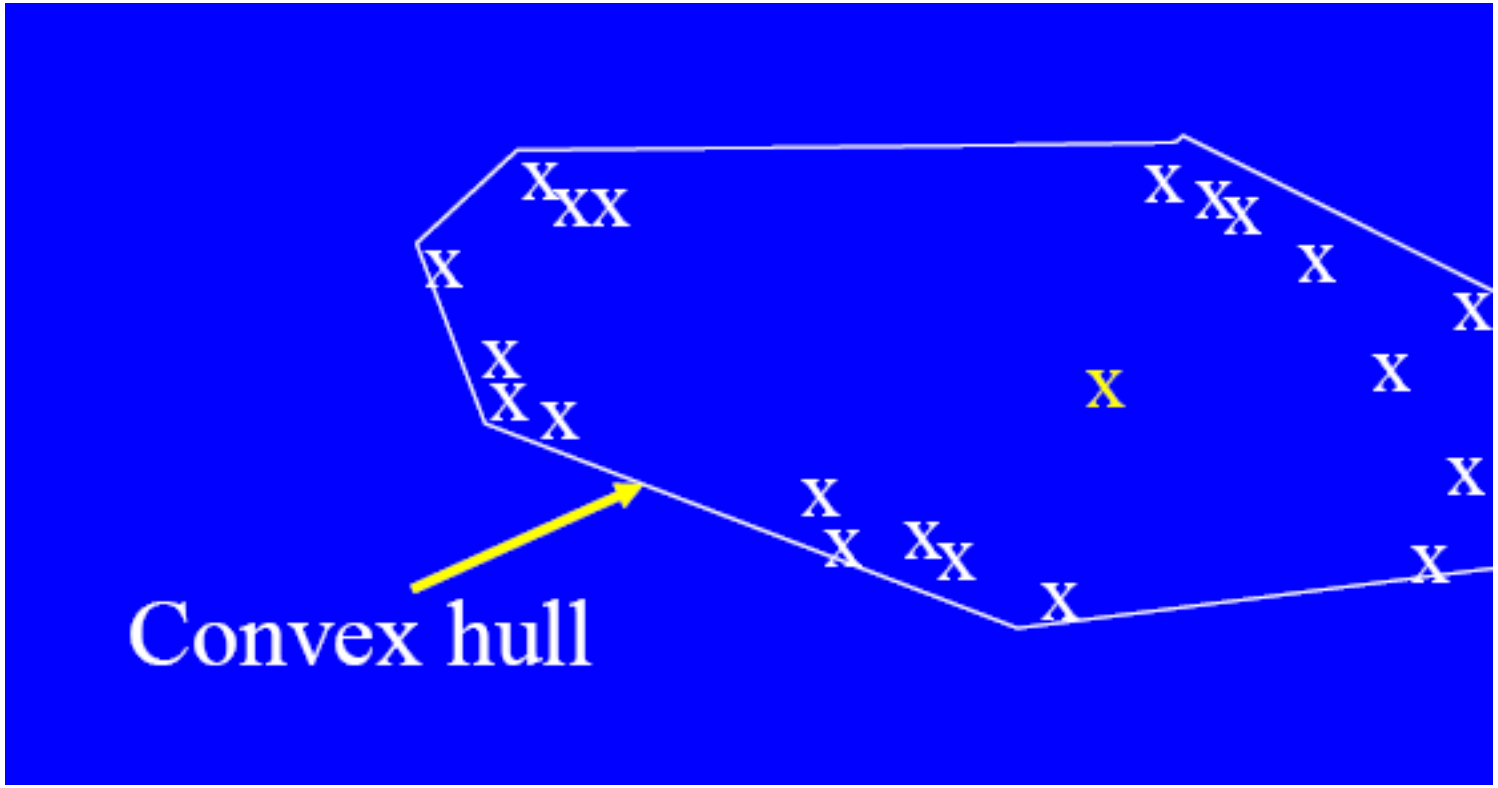


Limitations

- Time consuming
- Subjective

Graphical Approach

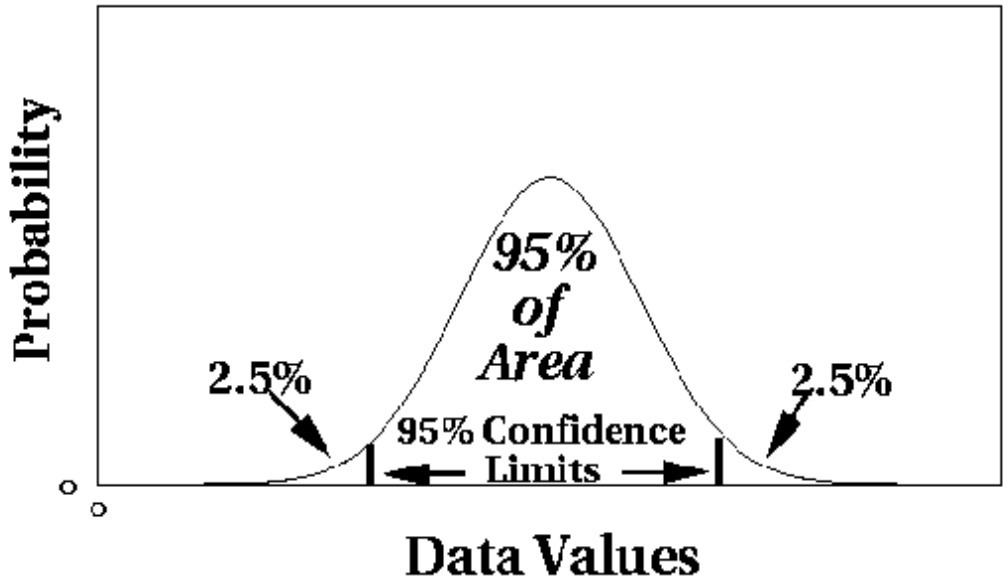
- Extreme points are assumed to be outliers
- Use convex hull method to detect extreme values
- What if the outlier occurs in the middle of the data?



Statistical Approach

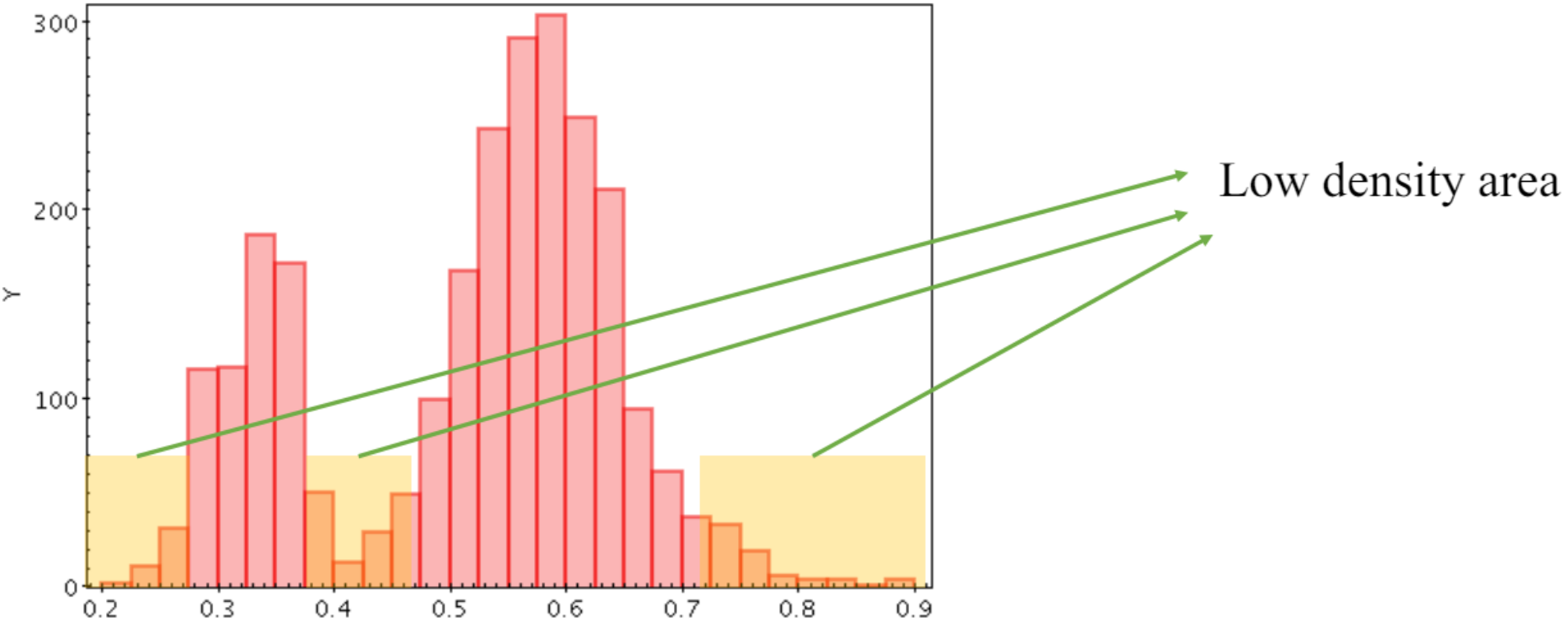
- Assume a parametric model describing the distribution of the data (e.g., normal distribution)

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



Statistical Approach

> Histogram-based Outlier Score (HBOS)



Distance-based Approaches

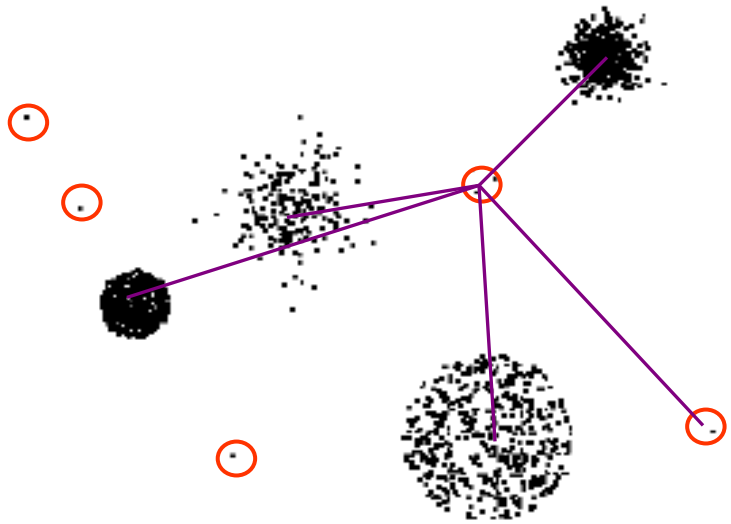
- **Data is represented as a vector of features**

- **Three major approaches**
  - Nearest-neighbor based
  - Density based
  - Clustering based

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

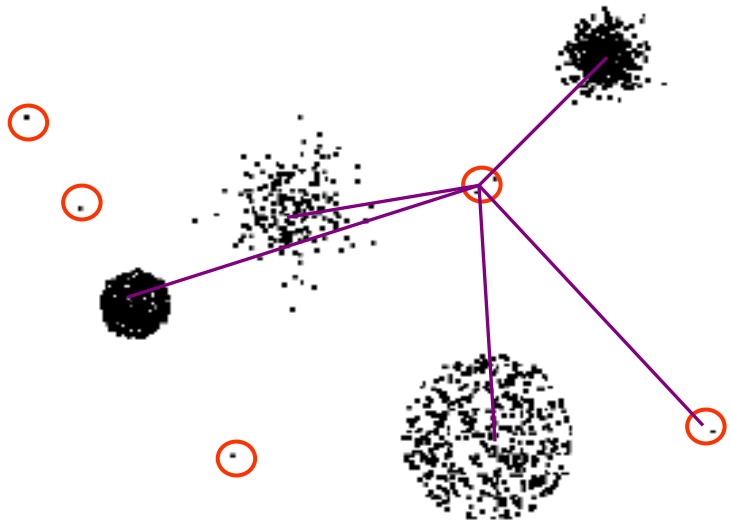
Clustering Based Approaches

- **Basic idea:**
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, **they are outliers**



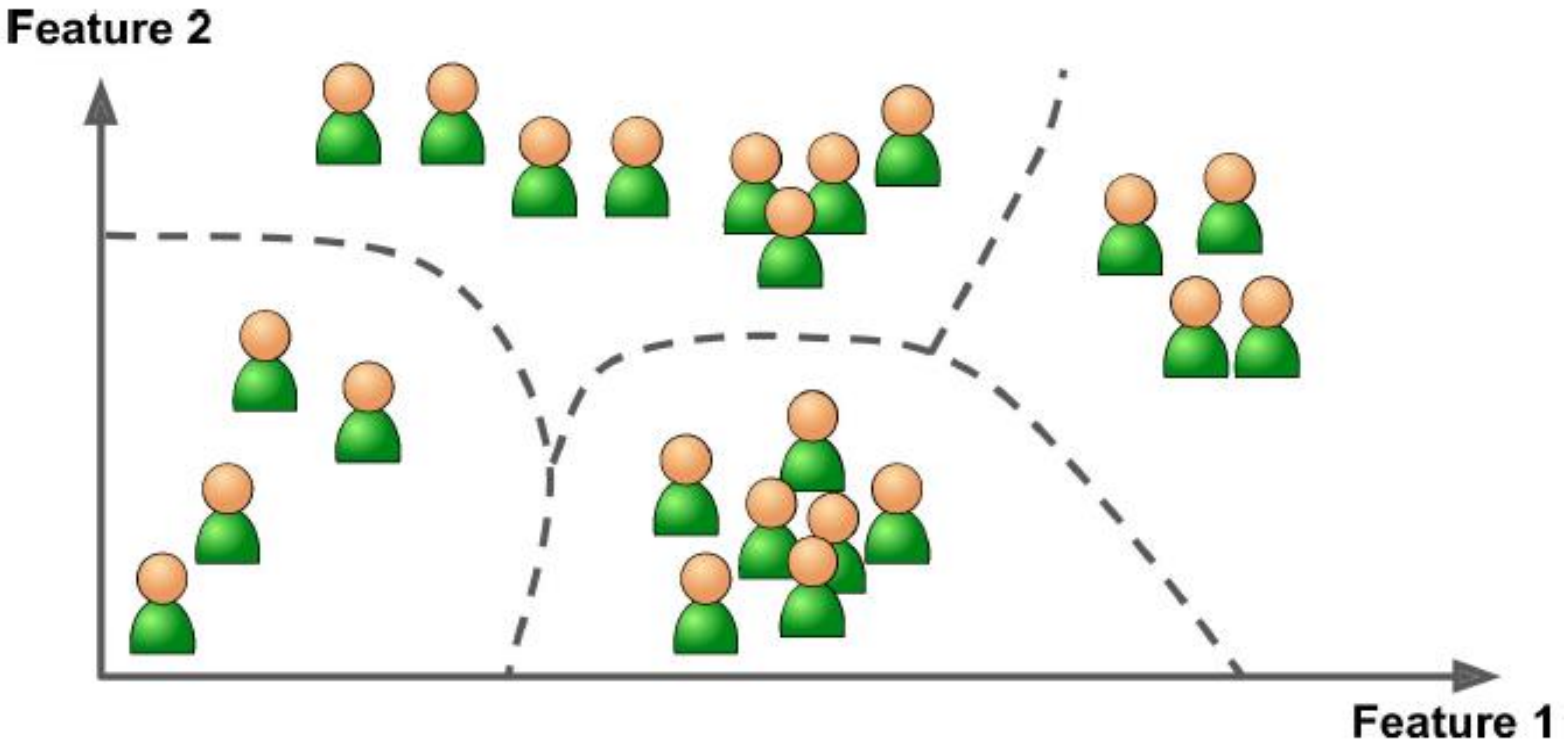
Clustering Based Approaches

- **Basic idea:**
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, **they are outliers**



Clustering Based Approaches

- **Example**
  - You have a lot of data about your blog's visitors
  - A clustering algorithm tries to detect groups of similar visitors

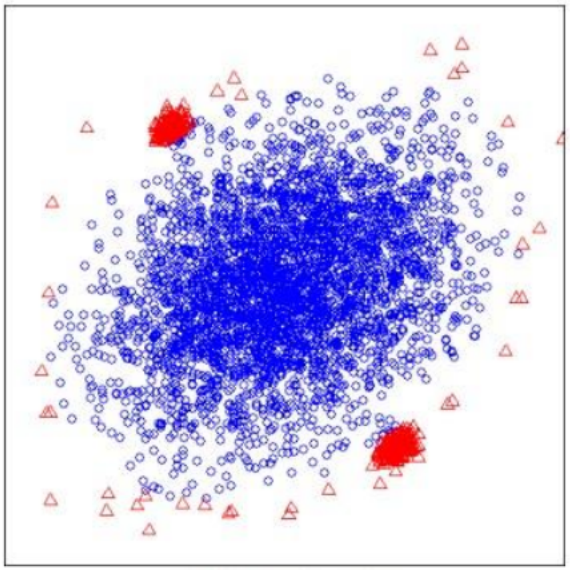


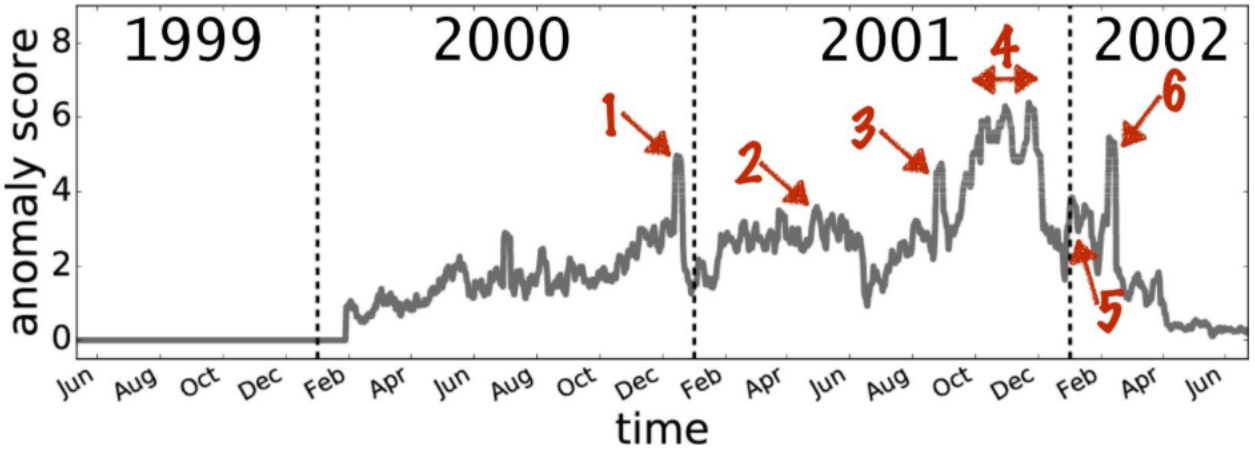
Clustering Based Approaches

> **Anomaly Detection**

Anomaly

> **What is that?**





Anomaly Detection

> Applications

Credit card fraud detection	Health monitoring	Fault detection in mechanical units
Clustering	Nearest neighbor technique	Random forest
Decision trees	Naïve bayes	Gradient boosted trees
SVM	Parametric statistical modeling	Spectral methods
Neural networks	Neural networks	Neural networks

Thanks

Dr. Fatih KAHRAMAN
fatih.kahraman@bau.edu.tr