# Data Mining: Anomaly Detection

Tom Heskes

Radboud University Nijmegen

# Anomaly/Outlier Detection

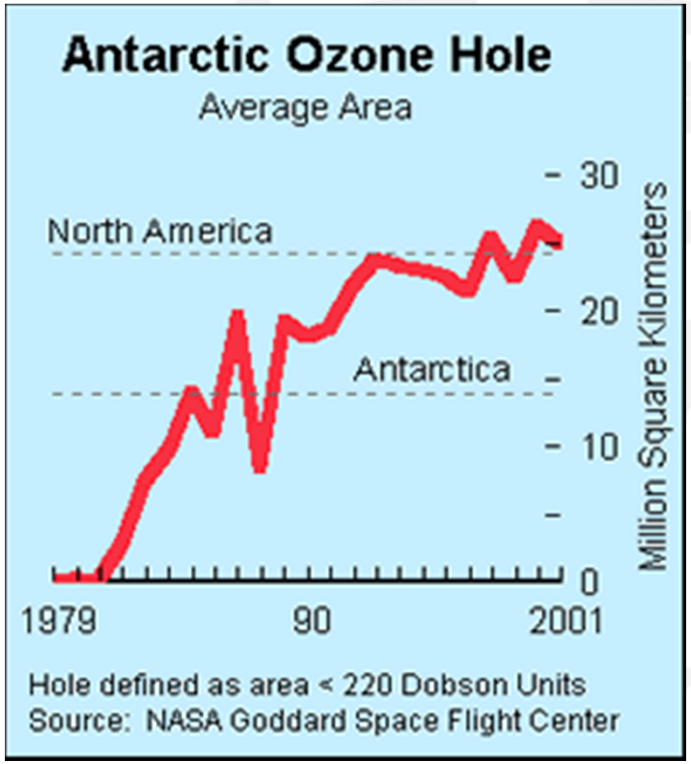
- What are anomalies/outliers?
  - The set of data points that are considerably different from the remainder of the data
- Variants of Anomaly/Outlier Detection Problems
  - Given a database *D*, find all the data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold *t*
  - Given 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

- 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!



# Anomaly Detection

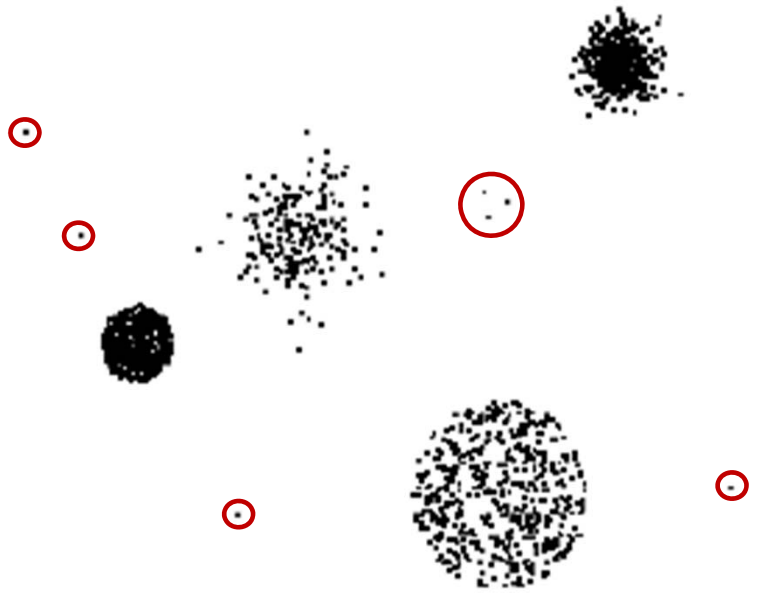
- Challenges
  - How many outliers are there in the data?
  - Method is unsupervised, so validation can be quite challenging (just like for clustering)
  - Finding needle in a haystack

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

# Anomaly Detection Schemes

- General Steps
  - Build a profile of the “normal” behavior, e.g., 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

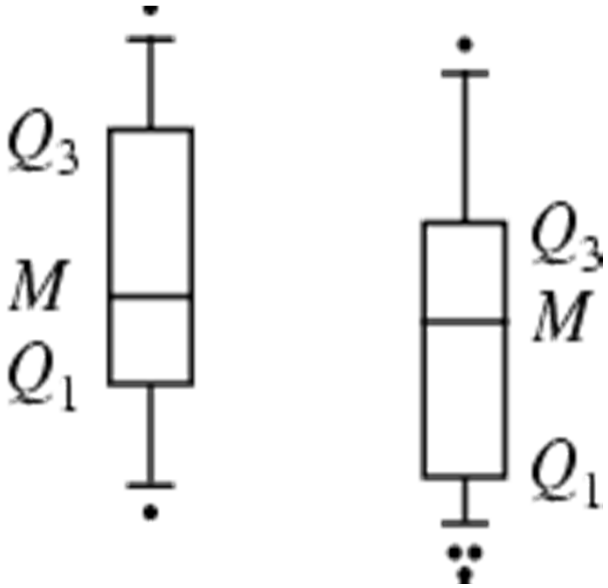
- Types of anomaly detection schemes
  - Graphical & statistical-based
  - Distance-based
  - Model-based

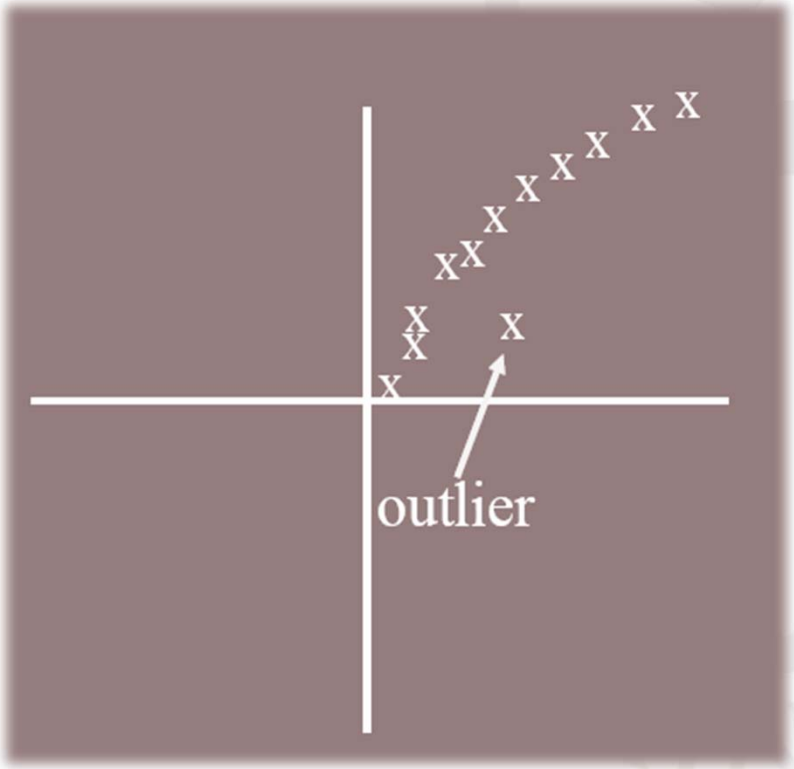


# Graphical Approaches

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

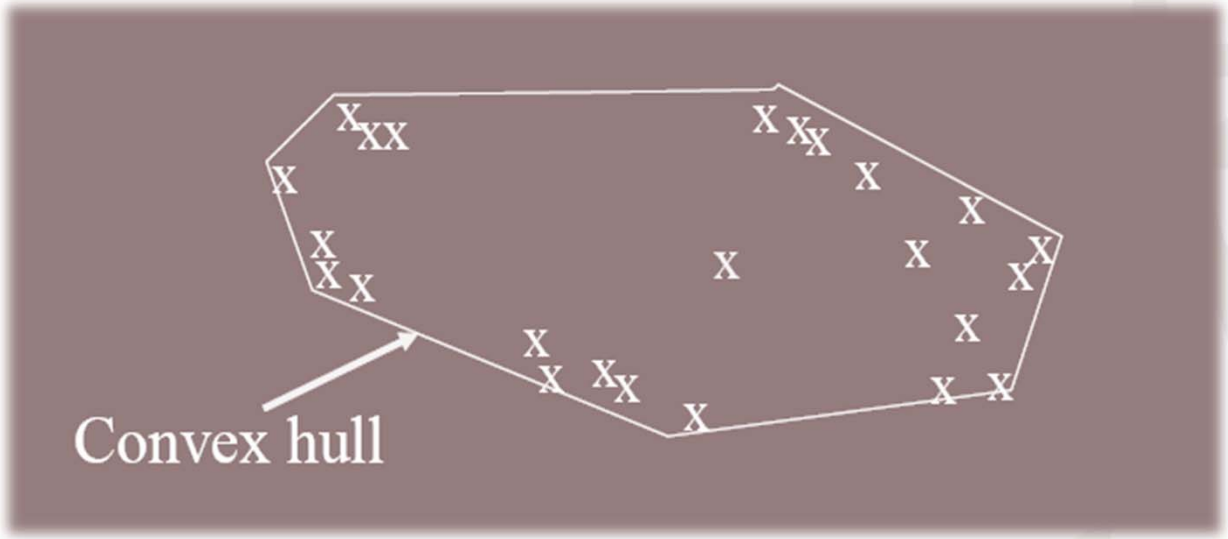
- Limitations
  - Time consuming
  - Subjective



# Convex Hull Method

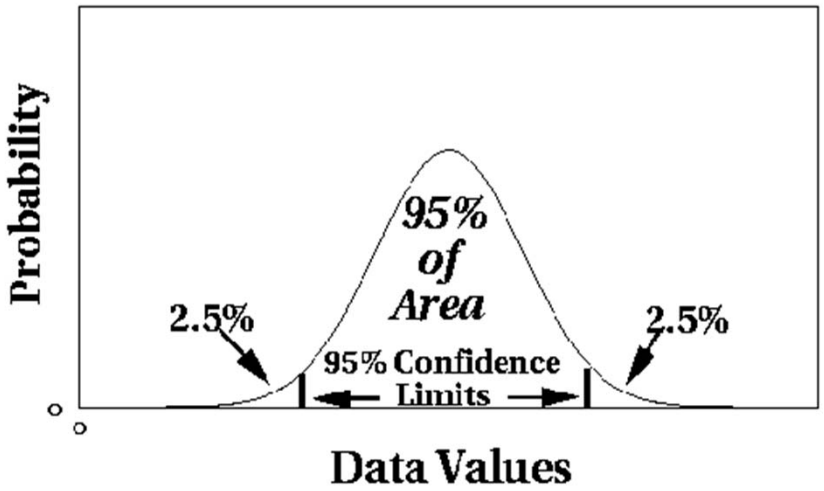
- Extreme points are assumed to be outliers
- Use convex hull method to detect extreme values




- What if the outlier occurs in the middle of the data?

# Statistical Approaches

- 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)



# Grubb's 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: $$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

# 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

# Distance-Based Approaches

- Data is represented as a vector of features
- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based

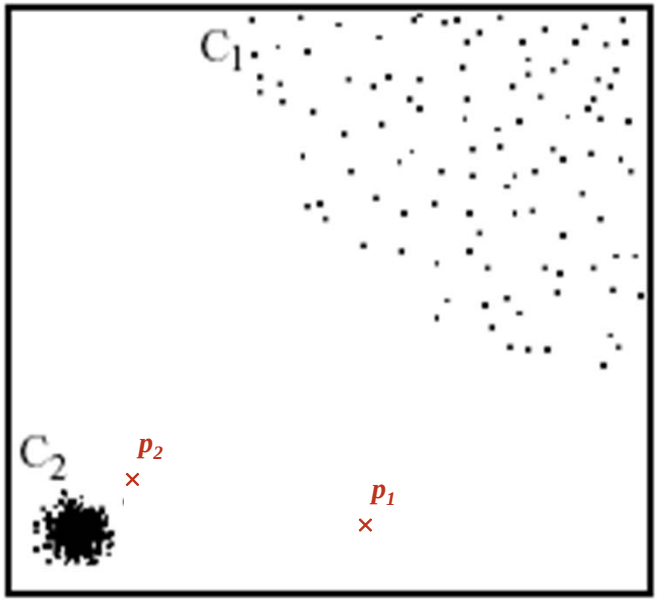
# Nearest-Neighbor Based Approach

- Compute the distance between every pair of data points
- There are various ways to define outliers:
  - Data points for which there are fewer than *p* neighboring points within a distance *D*
  - The top *n* data points whose distance to the *k*th nearest neighbor is greatest
  - The top *n* data points whose average distance to the *k* nearest neighbors is greatest

# Density-Based: Local Outlier Factor

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value

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



## Clustering-Based

- 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 these candidate points are far from all other non-candidate points, they are outliers