Outlier Detection: Extreme Values

Mining Massive Datasets

Prof. Carlos Castillo

Topic 19

Universitat Pompeu Fabra Barcelona

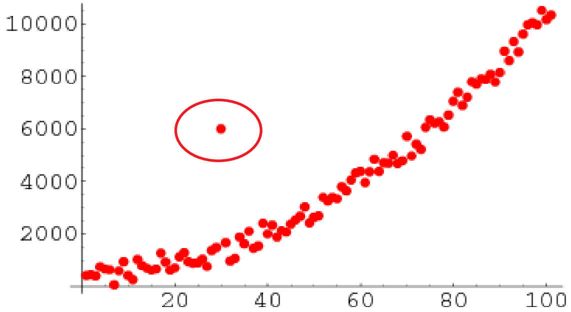
Sources

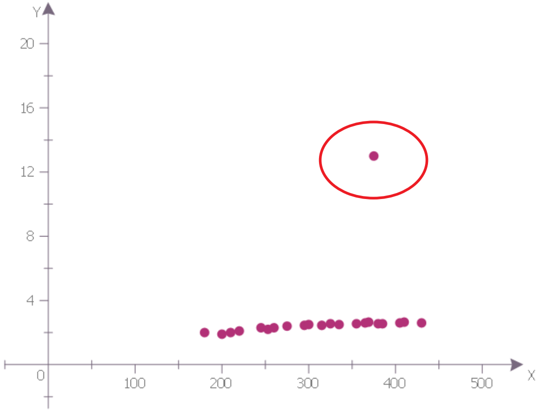
- Data Mining, The Textbook (2015) by Charu Aggarwal (chapter 8) – slides by Lijun Zhang

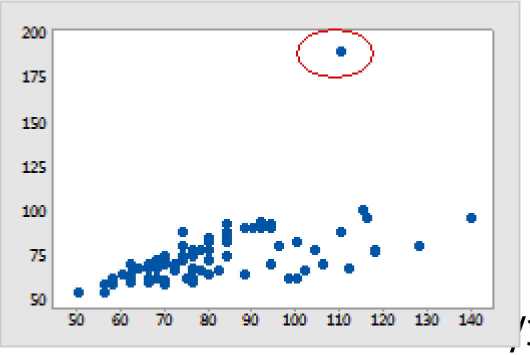
Outliers


Serena Williams

Sultan Kösen





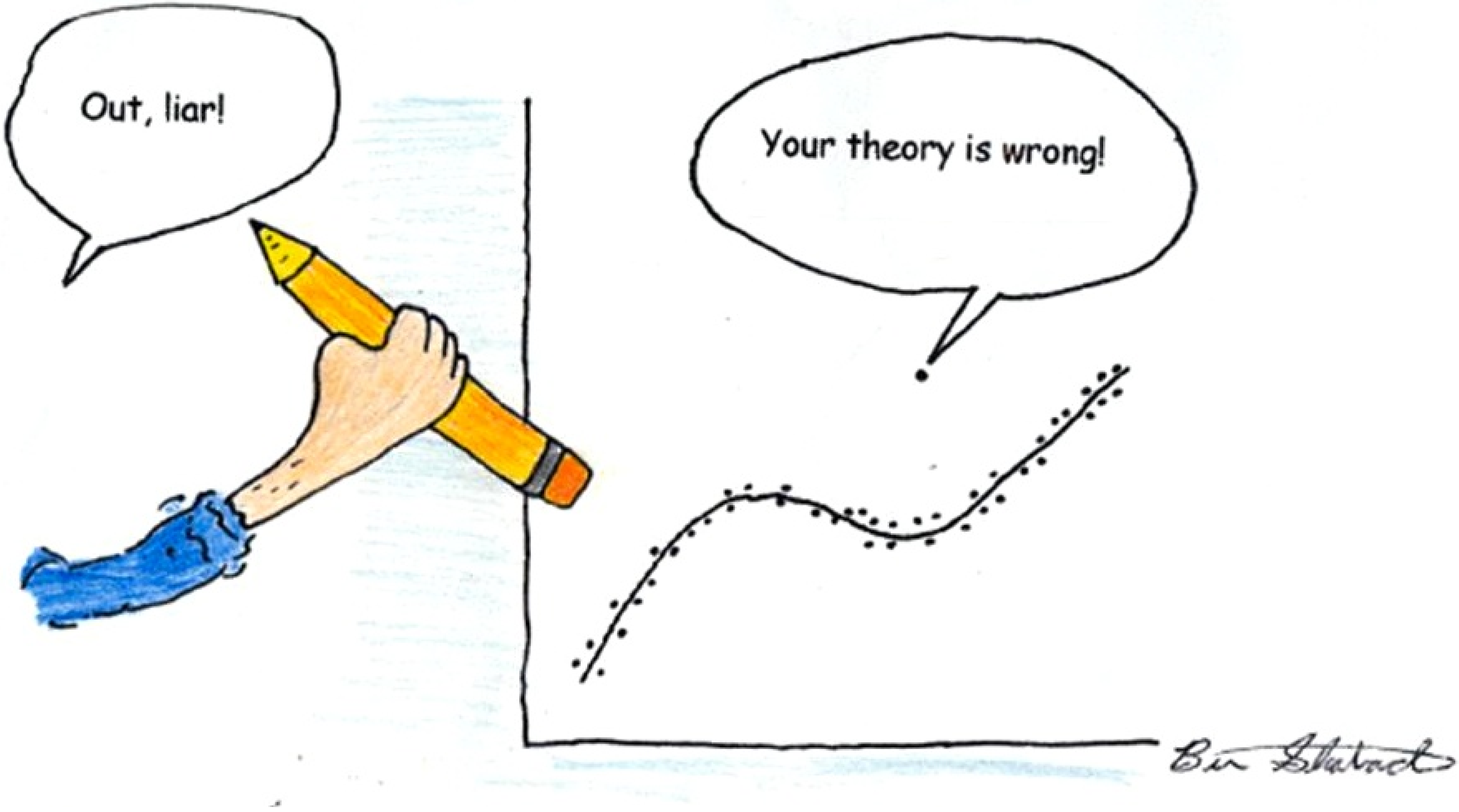


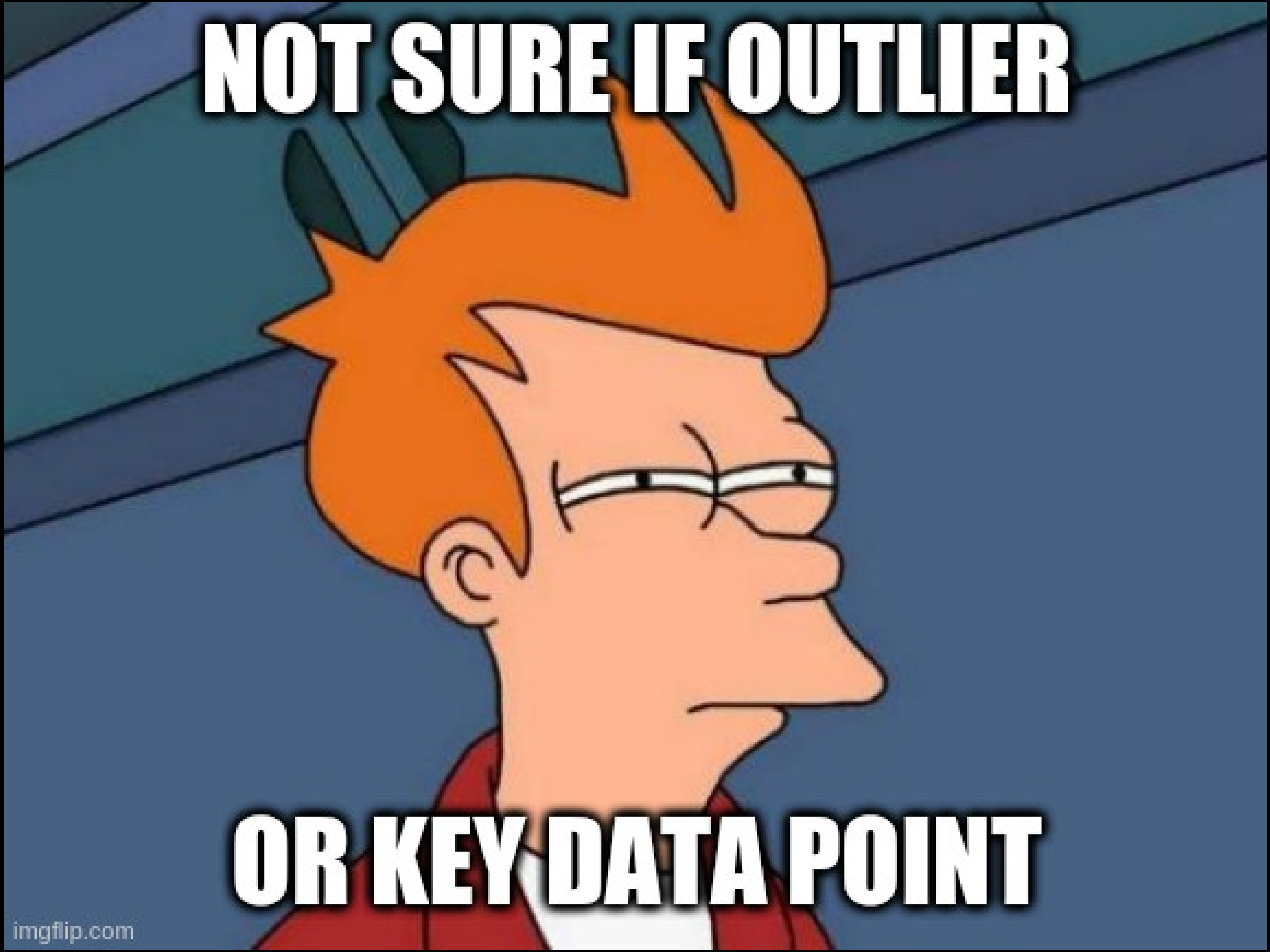
What is an outlier?

- Informally, "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."
  - **Clustering** seeks to group points that are **similar**
  - **Outlier detection** seeks points that are **different** from the remaining data

Outliers happen all the time

- Sensor malfunction
- Error in data transmission
- Transcription error
- Fraudulent behaviour
- Sample contamination
- Interesting natural occurrence 


Out, liar!
Your theory is wrong!


NOT SURE IF OUTLIER
OR KEY DATA POINT
imgflip.com

Some applications

- Data cleaning
  - Remove noise in data
- Credit card fraud
  - Unusual patterns of credit card activity
- Network intrusion detection
  - Unusual records/changes in network traffic

Outlier detection methods

- Key idea
  - Create a model of normal patterns
  - Outliers are data points that do not naturally fit within this normal model
  - The "*outlierness*" of a data point is quantified by a outlier score
- Outputs of Outlier Detection Algorithms
  - Real-valued outlier score
  - Binary label (outlier / not outlier)

Evaluation
(outlier validity)

Internal (unsupervised) criteria

- Rarely used in outlier analysis
- For any method, a measure can be created that will favor that method (~*overfit*)
- Solution space is small
  - Maybe there is just one outlier, finding it or missing it makes all the difference between perfect and useless performance

External (supervised) criteria

- **Known outliers** from a synthetic dataset or **rare items** (e.g., belonging to smallest class)
- Suppose D is the data, G are the real outliers, and S(t) are found when threshold t is used

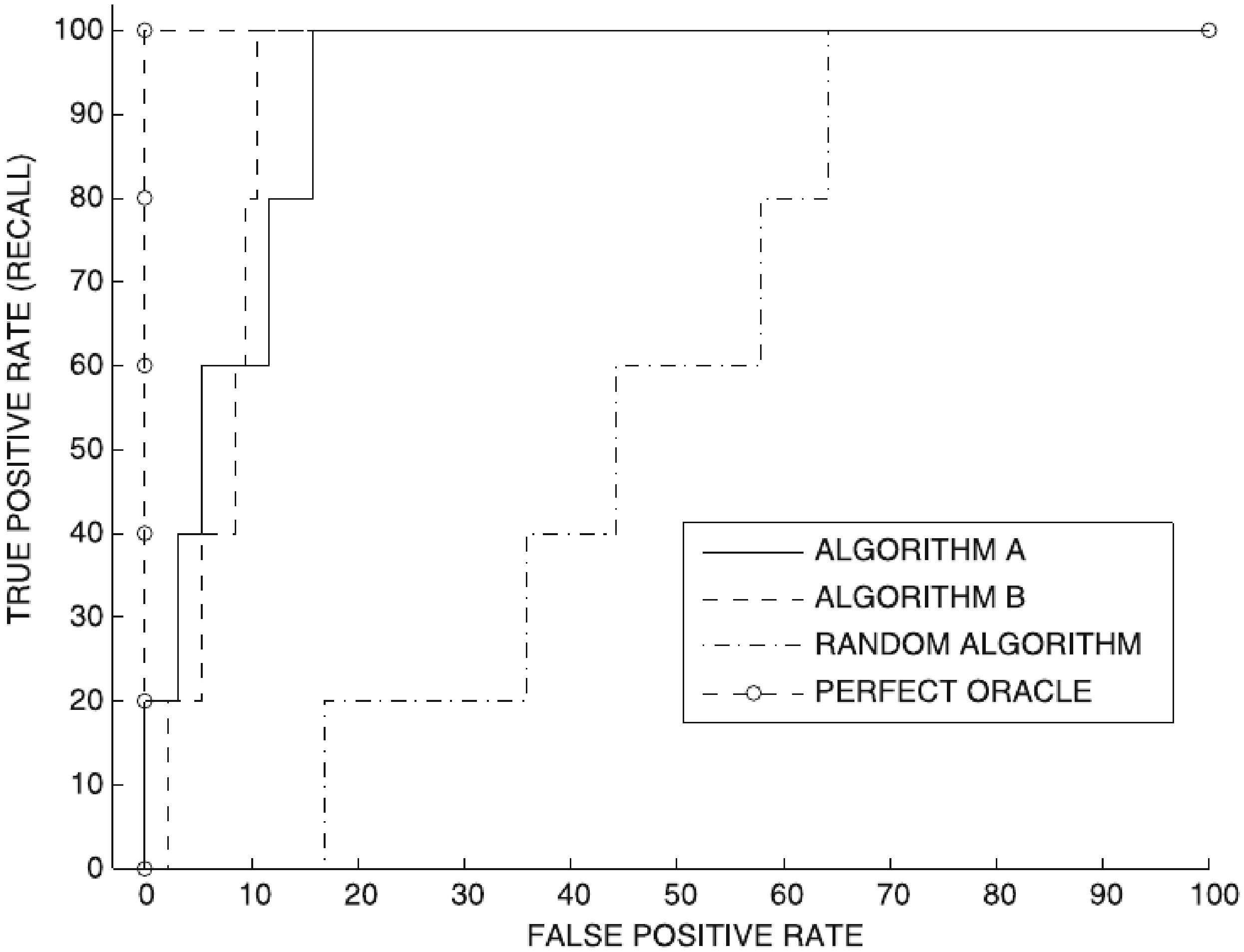
$$TPR(t) = \mathrm{Recall}(t) = 100 \cdot \frac{|S(t) \cap G|}{|G|}$$

$$FPR(t) = 100 \cdot \frac{|S(t) - G|}{|D - G|}$$

ROC curve = (FPR(t), TPR(t))$_t$

Rank of ground-truth outliers:

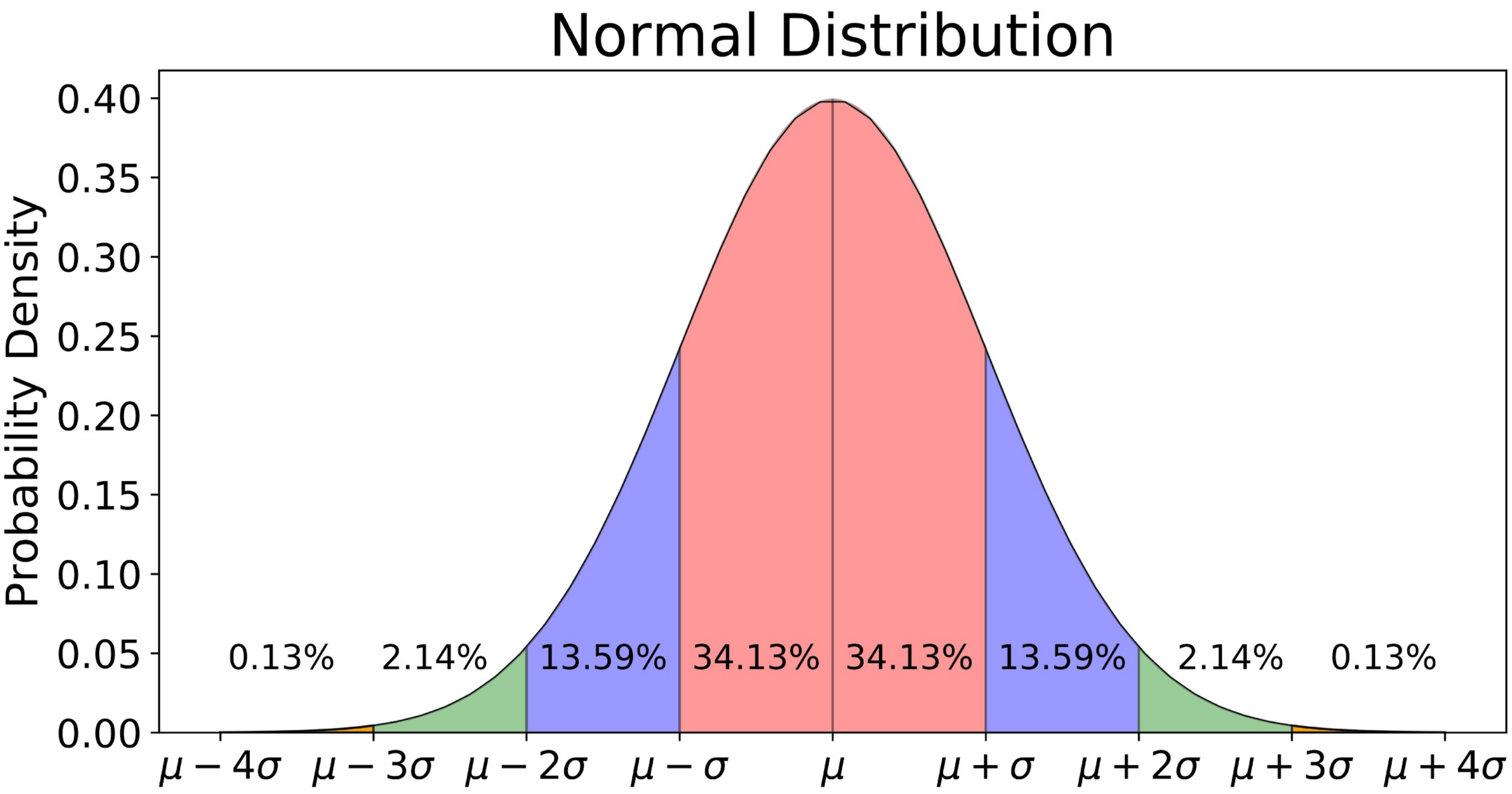
- Algorithm A
  1, 5, 8, 15, 20
- Algorithm B
  3, 7, 11, 13, 15
- Random
  17, 36, 45, 59, 66
- Perfect oracle
  1, 2, 3, 4, 5



Remainder of TT19-TT21:
Outlier detection methods

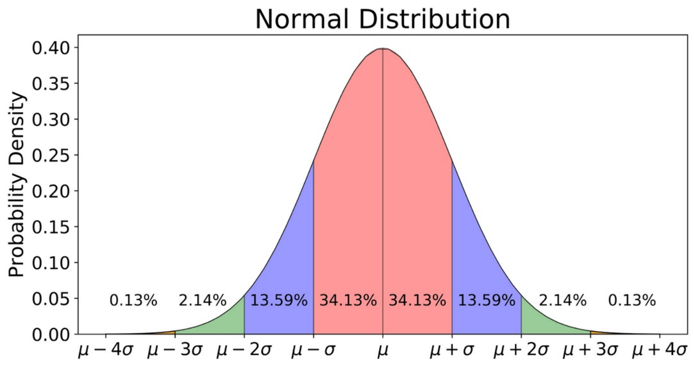
Extreme values analysis

Extreme value analysis: Statistical Tails



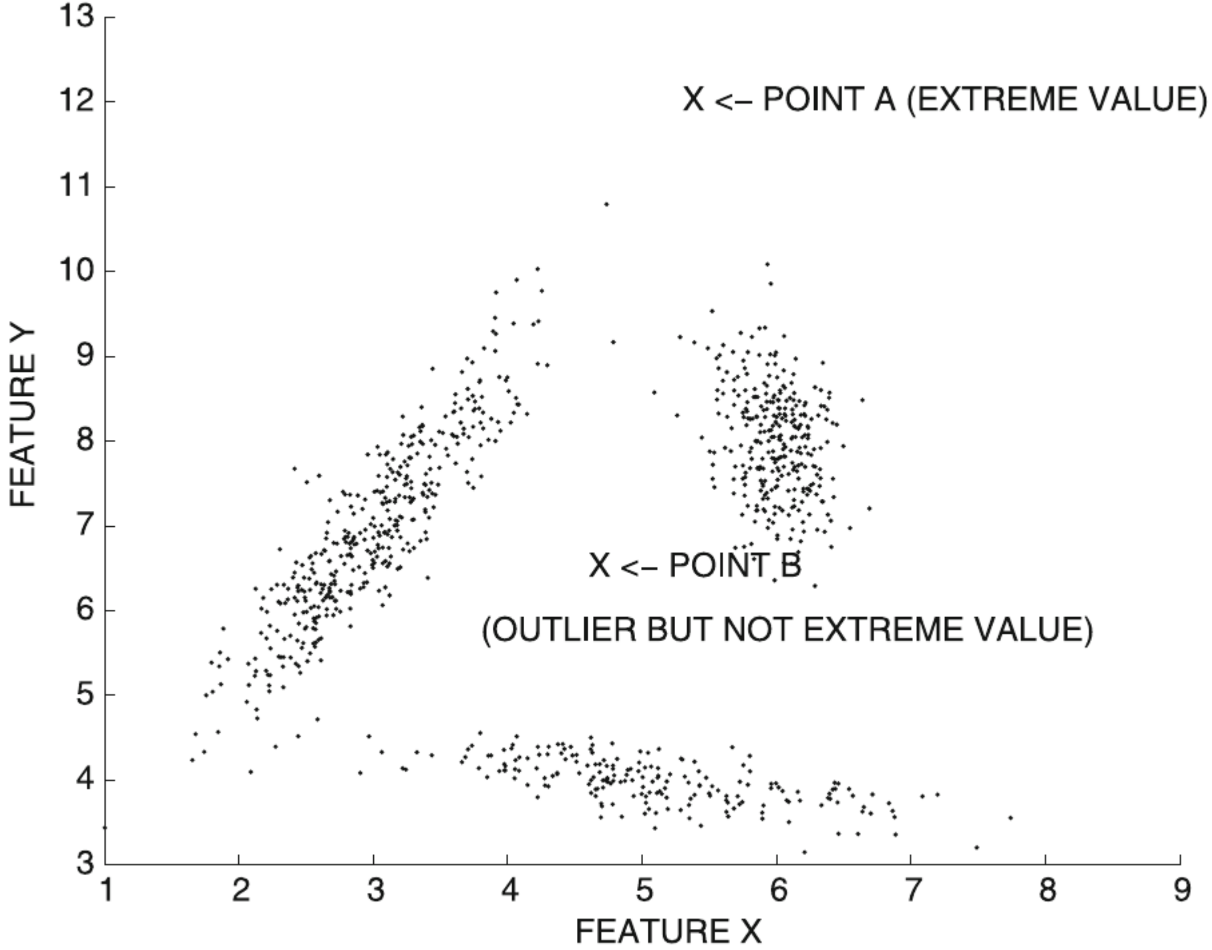
Extreme value analysis (cont.)

- Hypothesis: all extreme values are outliers
- However, outliers may not be extreme values:
  - {1,3,3,3,50,97,97,100}
  - 1 and 100 are extreme values
  - 50 is an outlier but no an extreme value



Extreme value analysis (cont.)

- Point A is an extreme value (hence, an outlier)
- Point B is an outlier but not an extreme value



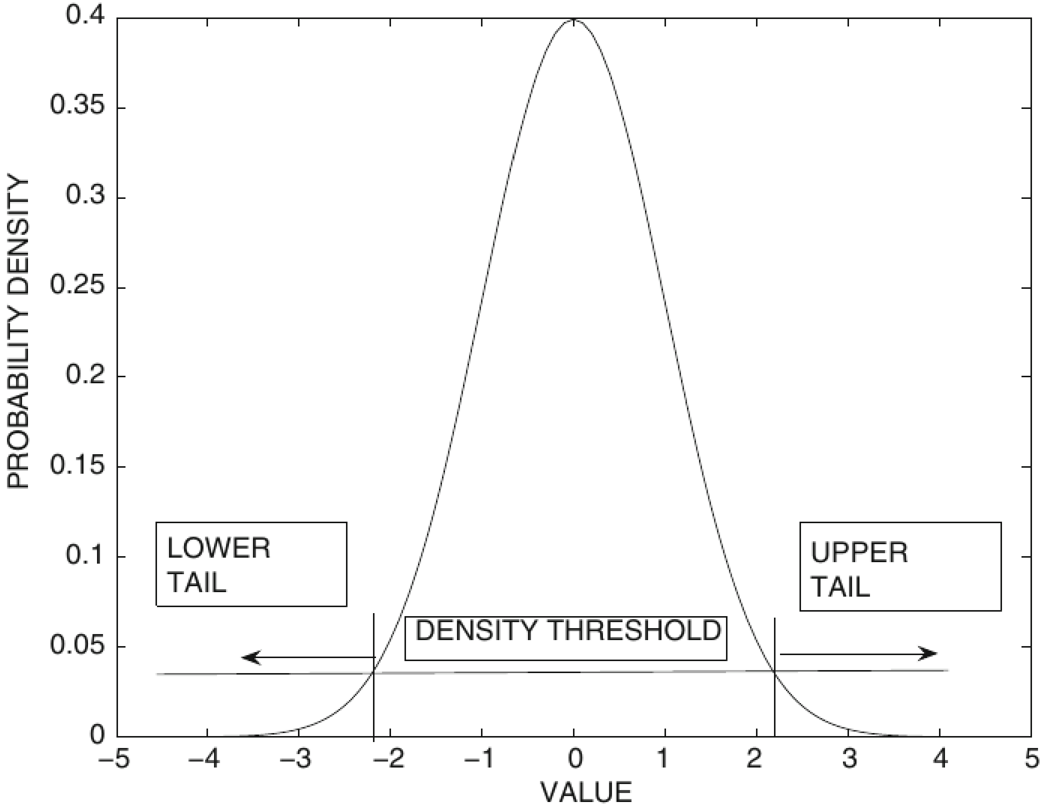
Univariate extreme value analysis

- Statistical tail confidence test
  - Let density be $f_X(x)$
  - Tails are **extreme** regions s.t. $f_X(x) \leq \theta$

Univariate extreme value analysis (cont.)

Let density be $f_X(x)$; tails are extreme regions s.t. $f_X(x) \leq \theta$

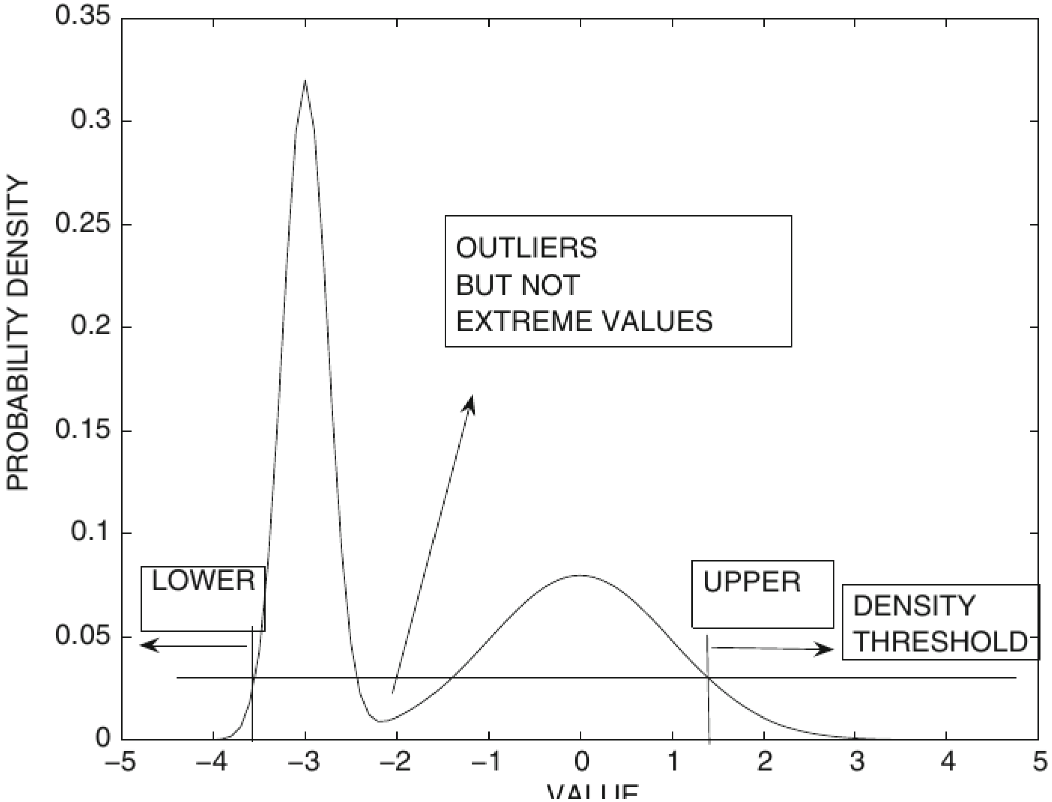
- Symmetric distribution
  - Two symmetric tails
  - Areas inside tails represent cumulative distribution



Univariate extreme value analysis (cont.)

Let density be $f_X(x)$; tails are extreme regions s.t. $f_X(x) \leq \theta$

- Asymmetric distribution
  - Areas in two tails are different
  - Regions in the interior are not tails



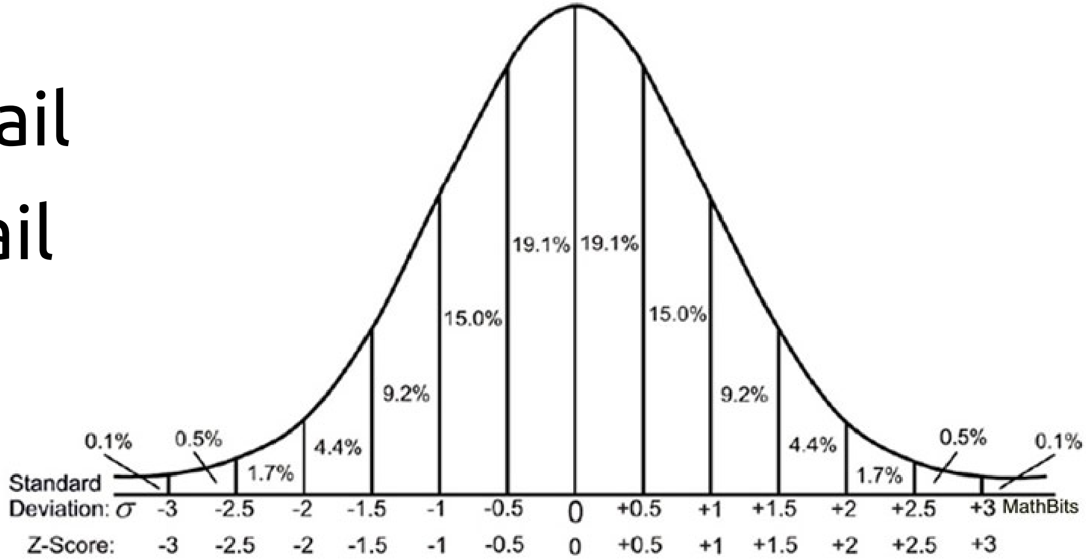
Standardization for univariate outlier analysis

- Normal distribution assumed $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
- Parameters
  - From prior knowledge
  - Estimated from data

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Standardization for univariate outlier analysis (cont.)

- z-value $$z_i = \frac{x_i - \mu}{\sigma}$$
- Follows normal distribution with mean 0 and standard deviation 1
  - Large z-value: upper tail
  - Small z-value: lower tail



Exercise

Answer in
Google Spreadsheet

- Find the absolute z-score of each feature in the electrical scooters dataset
- Compute max z-score across features as an outlier metric
- Indicate which are the #1 and #2 most unusual electric scooters in this list and why

Multivariate extreme values

- Probability density function of a Multivariate Gaussian distribution in d dimensions

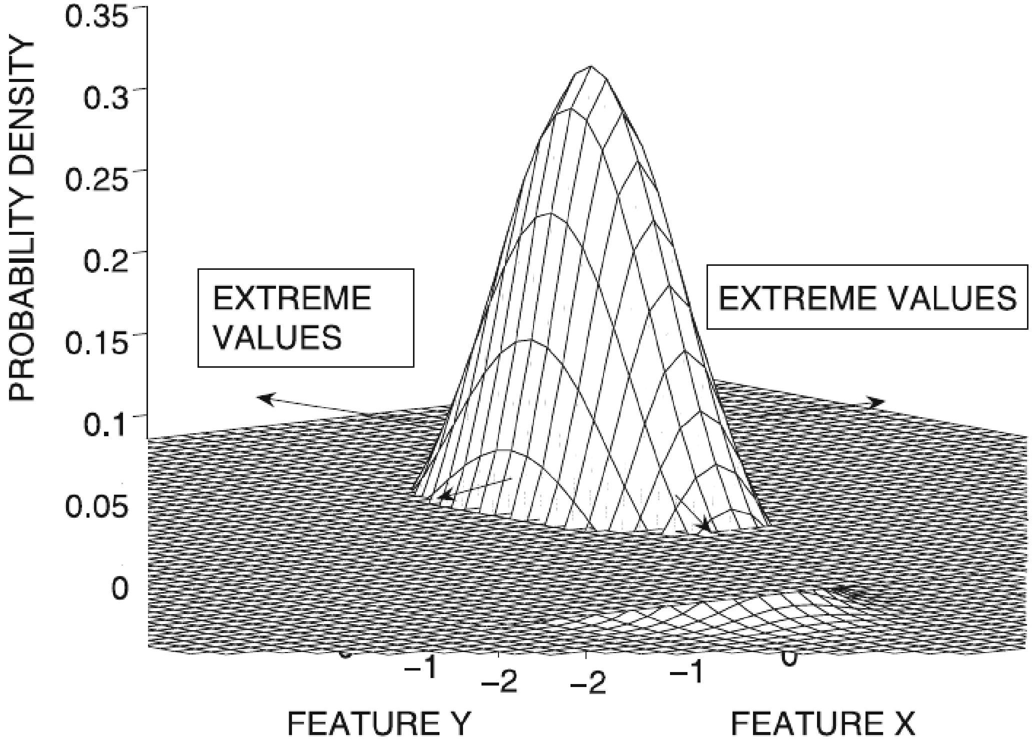
$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot (\overline{X} - \overline{\mu}) \Sigma^{-1} (\overline{X} - \overline{\mu})^T}$$

$$= \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot Maha(\overline{X}, \overline{\mu}, \Sigma)^2}$$

- $Maha(\overline{X}, \overline{\mu}, \Sigma)$ is the Mahalanobis distance
- $|\Sigma|$ is the determinant of the covariances matrix

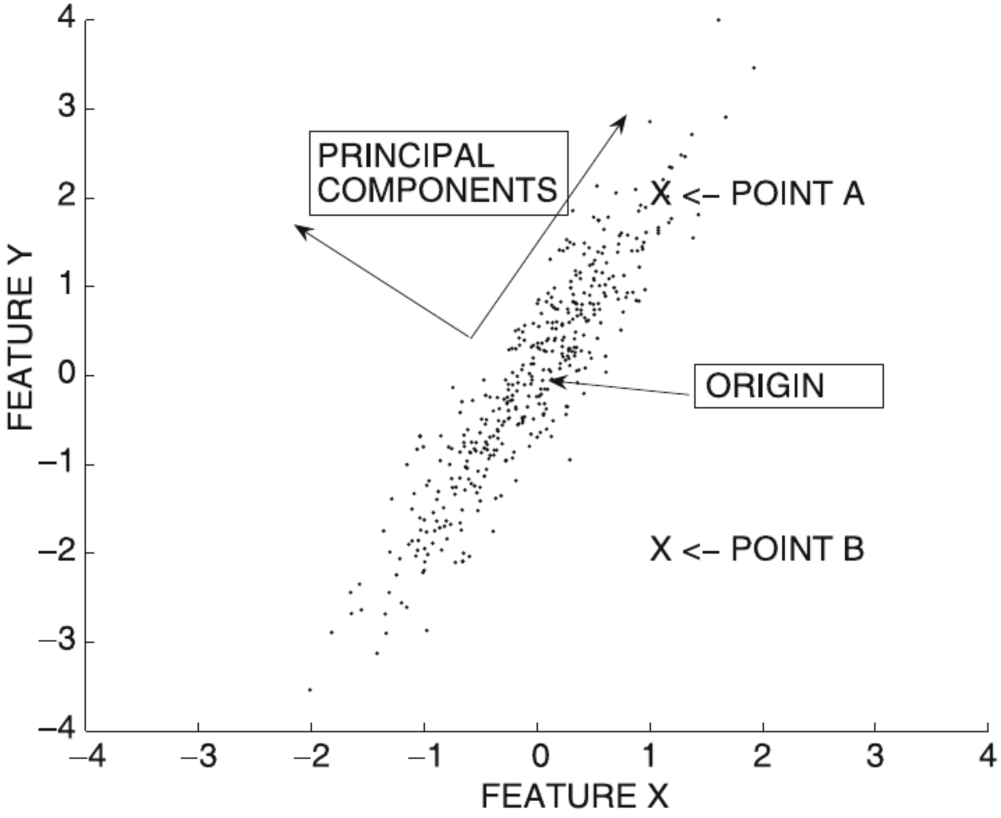
Multivariate extreme values (cont.)

- Extreme value score of $\overline{X}$
  - $Maha(\overline{X}, \overline{\mu}, \Sigma)$
  - Mahalanobis distance to the mean of the data
  - Larger values imply more extreme behavior



Multivariate extreme values (cont.)

- Extreme value score of $\overline{X}$
  - $Maha(\overline{X}, \overline{\mu}, \Sigma)$
  - Mahalanobis distance to the mean of the data
  - Larger values imply more extreme behavior
  - The Mahalanobis distance is the Euclidean distance in a transformed (axes-rotated) data set after dividing each of the transformed coordinate values by the standard deviation along its direction

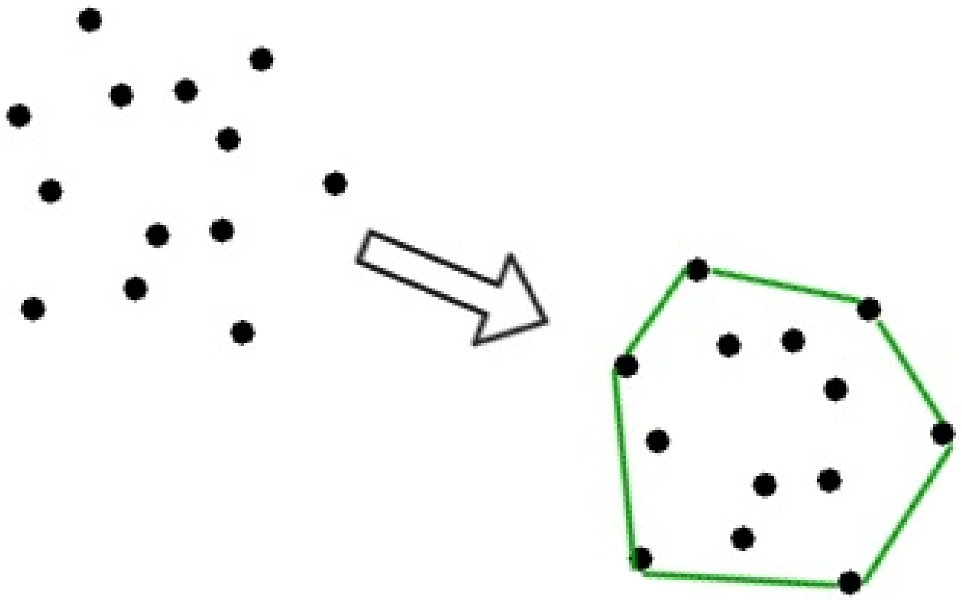


Depth-based methods

Key concept: convex hull

The convex hull of a set C is the set of all convex combinations of points in C

$$\begin{aligned}\text{conv } C = \{\theta_1 x_i + \cdots + \theta_k x_k | \\ x_i \in C, \\ \theta_i \geq 0, \\ \theta_1 + \cdots + \theta_k = 1\}\end{aligned}$$

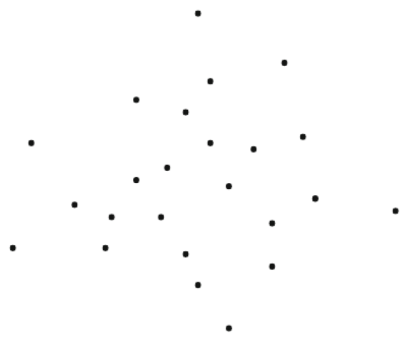


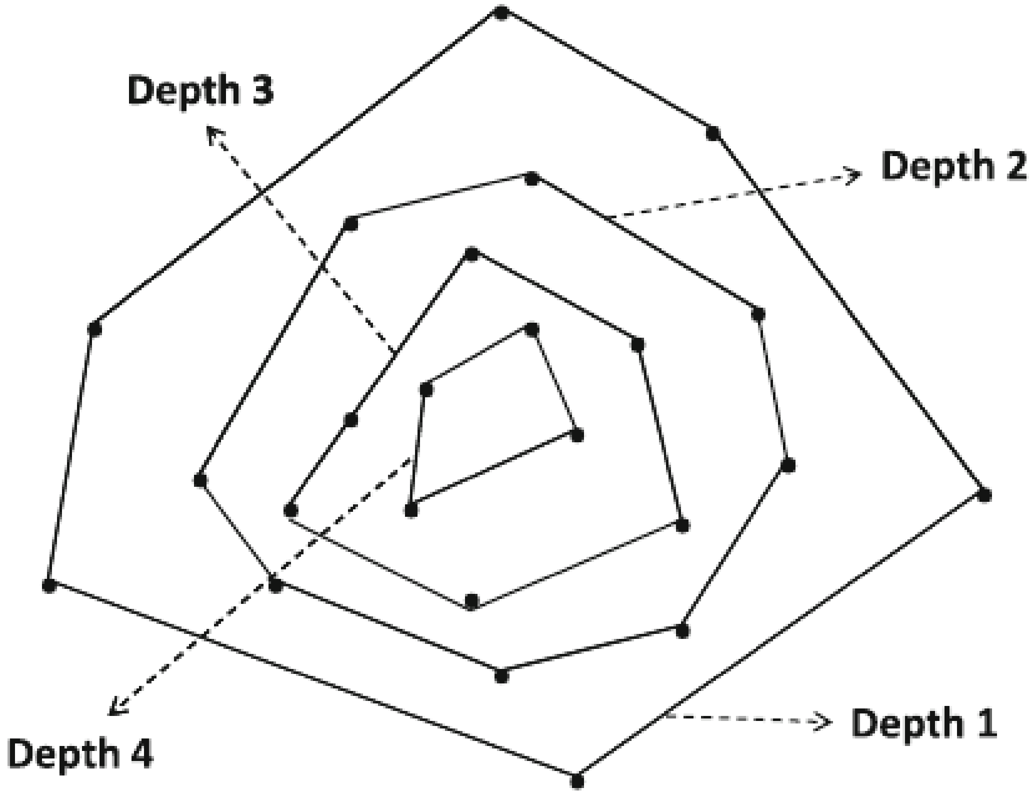
Algorithm

Algorithm *FindDepthOutliers*(Data Set: $\mathcal{D}$, Score Threshold: r)
begin
  $k = 1$;
  repeat
    Find set S of corners of convex hull of $\mathcal{D}$;
    Assign depth k to points in S;
    $\mathcal{D} = \mathcal{D} - S$;
    $k = k + 1$;
  until(D is empty);
  Report points with depth at most r as outliers;
end

Explanation: peeling layers

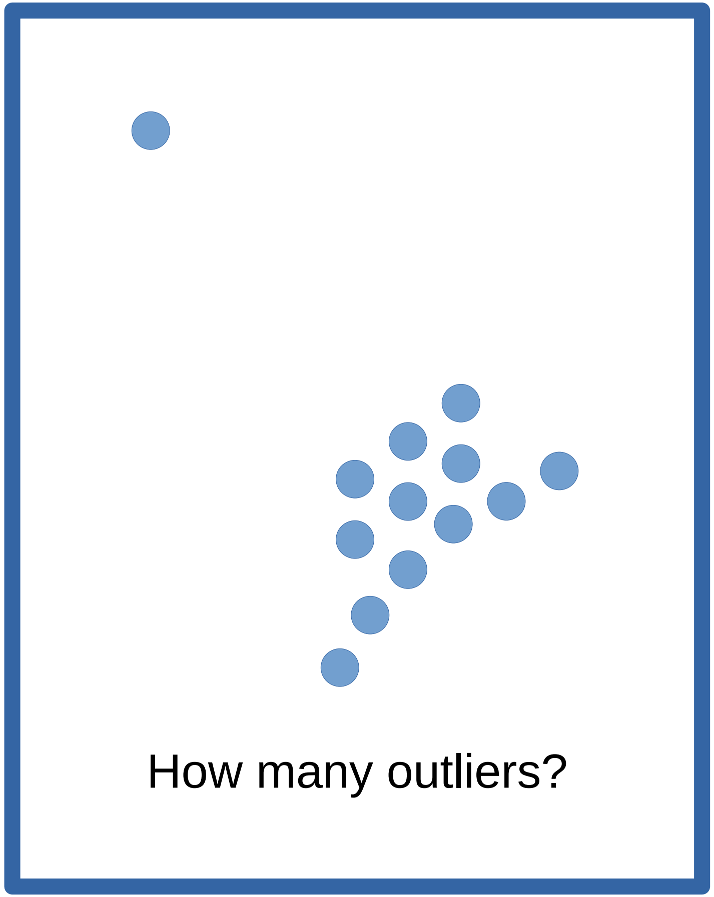
Algorithm *FindDepthOutliers*(Data Set: $\mathcal{D}$, Score Threshold: r)
begin
  $k = 1$;
  repeat
    Find set S of corners of convex hull of $\mathcal{D}$;
    Assign depth k to points in S;
    $\mathcal{D} = \mathcal{D} - S$;
    $k = k + 1$;
  until(D is empty);
  Report points with depth at most r as outliers;
end





Limitations of this method

- No normalization
- Computational complexity increases significantly with dimensionality
- Many data points are indistinguishable



Summary

Things to remember

- Extreme value analysis
  - Univariate, multivariate, depth based
- Outlier evaluation/validity

Exercises for TT19-TT21

- Data Mining, The Textbook (2015) by Charu Aggarwal
  - Exercises 8.11 → all except 10, 15, 16, 17