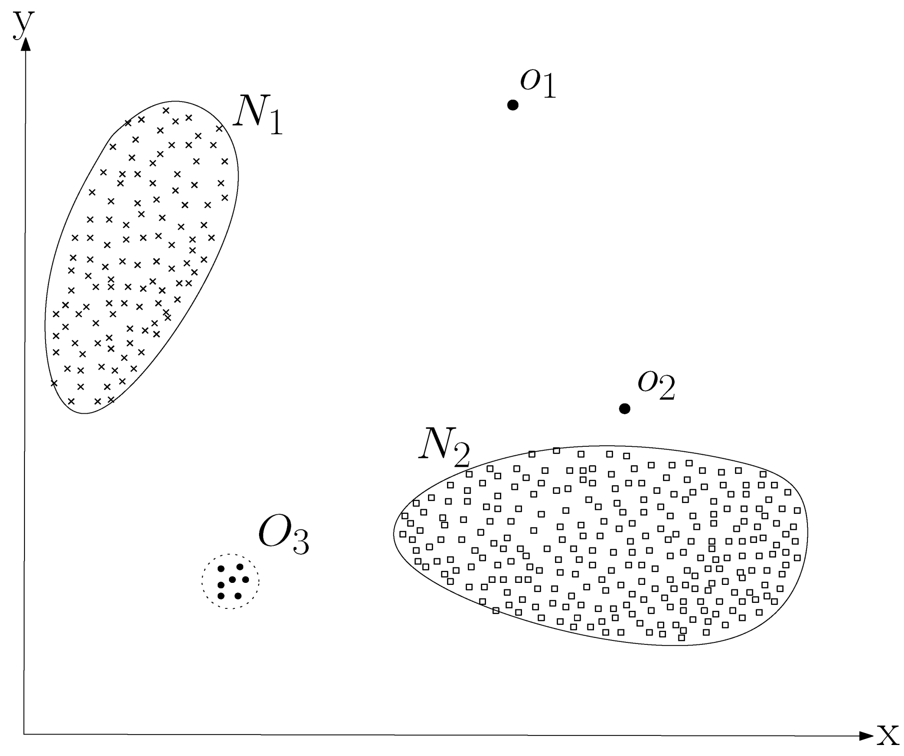
Anomaly detection

Luboš Popelínský
Luboš Popelínský
KDLab, Faculty of Informatics, MU

Thanks to Luis Torgo, Karel Vaculík and other members of the KDLab

What is an Outlier ?

- Definition of Hawkins [Hawkins 1980]:
- *"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"*



Applications of Outlier Detection

- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
- Medicine
  - Unusual symptoms or test results may indicate potential health problems of a patient
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
- Detecting measurement errors
  - Data derived from sensors may contain measurement errors
  - Removing such errors can be important in other data mining and data analysis tasks
- Intrusion detection
- Language learning “irregularities”
  - *Jedu do Porta. Jedu do hor.* vs. *Jedu na hory.*

Types of Outliers

Point outliers

Cases that either individually or in small groups are very different from the others.

Contextual outliers

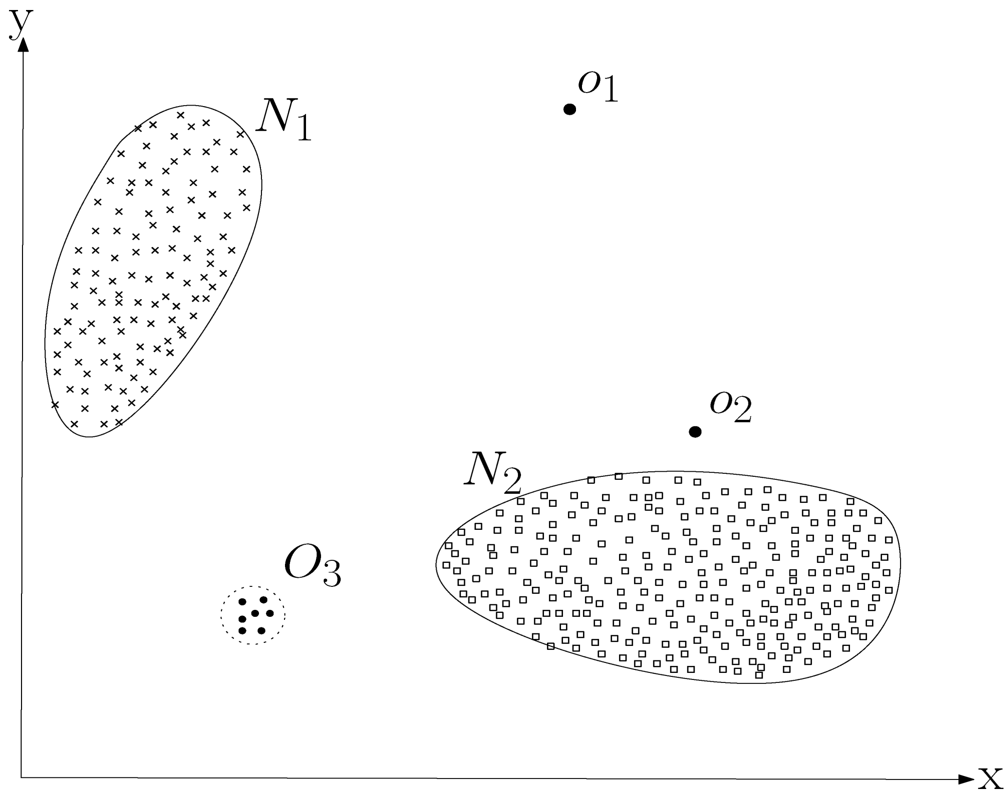
Cases that can only be regarded as outliers when taking the context where they occur into account.

Collective outliers

Cases that individually cannot be considered strange, but together with other associated cases are clearly outliers.

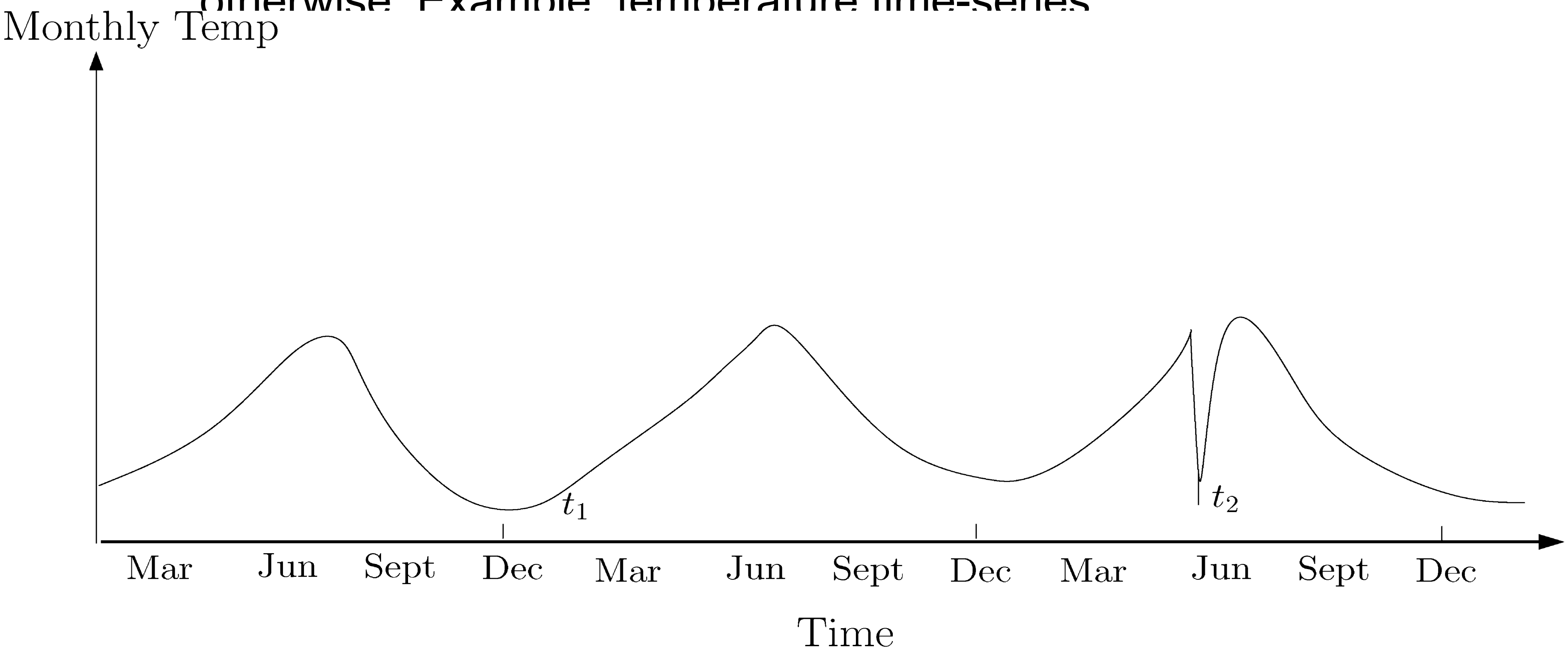
Types of Outliers

- Point Anomalies
  - An individual data instance can be considered as anomalous with respect to the rest of data
  - Example:
    credit card fraud detection



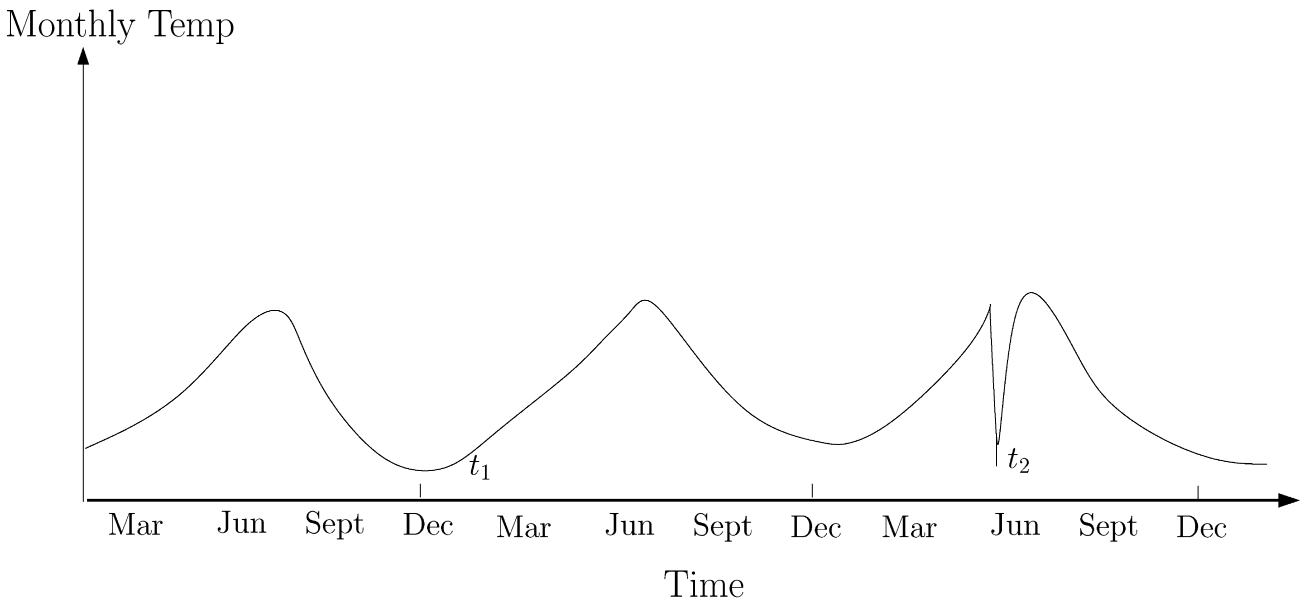
Types of Outliers

- Contextual Anomalies
  - If a data instance is anomalous in a specific context, but not otherwise. Example: temperature time-series



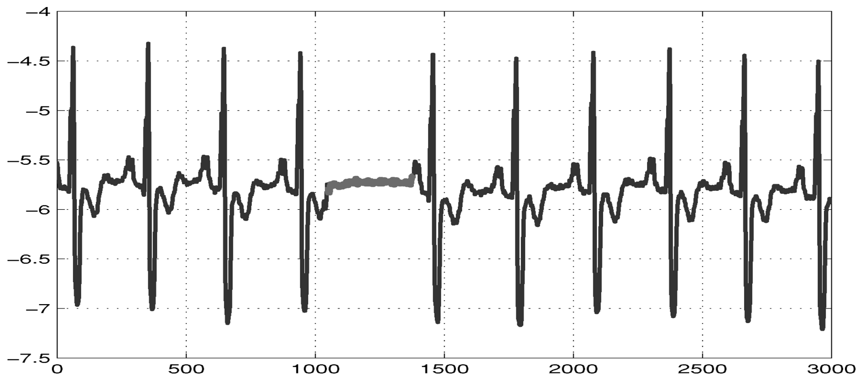
Context-based Approach

- Is the temperature 28°C outlier?
- If we are in Brno in summer NO
- If we are in Brno in winter YES
- → it dependes on the location and time – CONTEXT



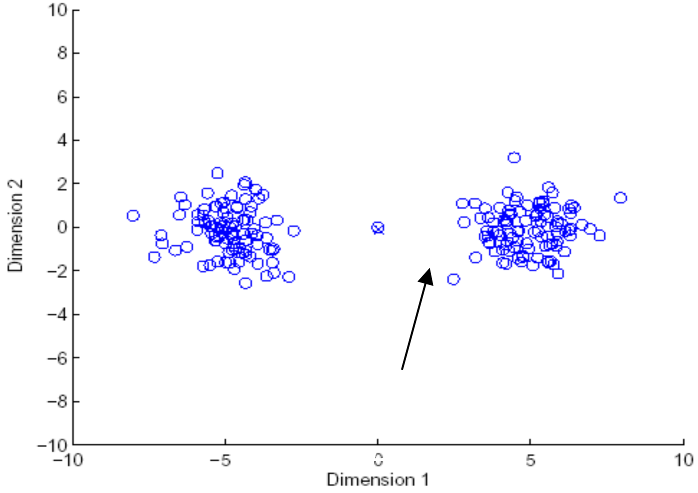
Types of Outliers

- Collective Anomalies
  - A collection of related data instances is anomalous with respect to the entire data set.
  - The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous. *Example: human cardiogram*



Outlier Detection Methods

- Statistical Methods
  - normal data objects are generated by a statistical (stochastic) model, and data not following the model are outliers
  - Example: statistical distribution: Gaussina
    - Outliers are points that have a low probability to be generated by Gaussian distribution
  - Problems: Mean and standard deviation are very sensitive to outliers
    - These values are computed for the complete data set (including potential outliers)
  - Advantage: existence of statistical proof why the object is an outlier



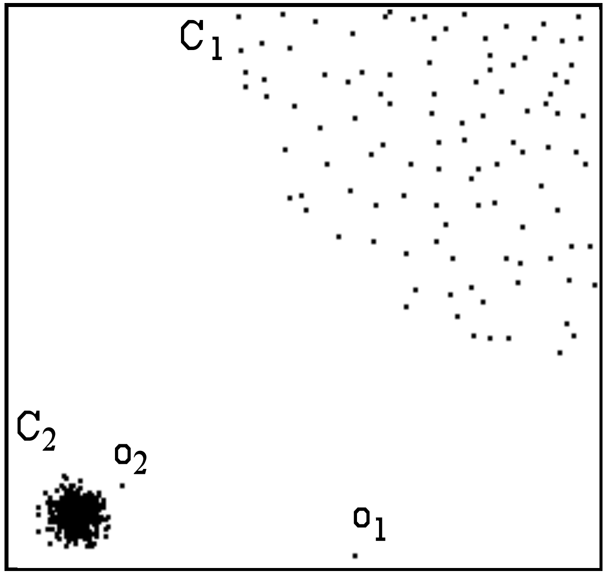
Outlier Detection Methods

- **Proximity-Based Methods**
- An object is an outlier if the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set.

- Distance-based Detection
  - Radius *r*, *k* nearest neighbors

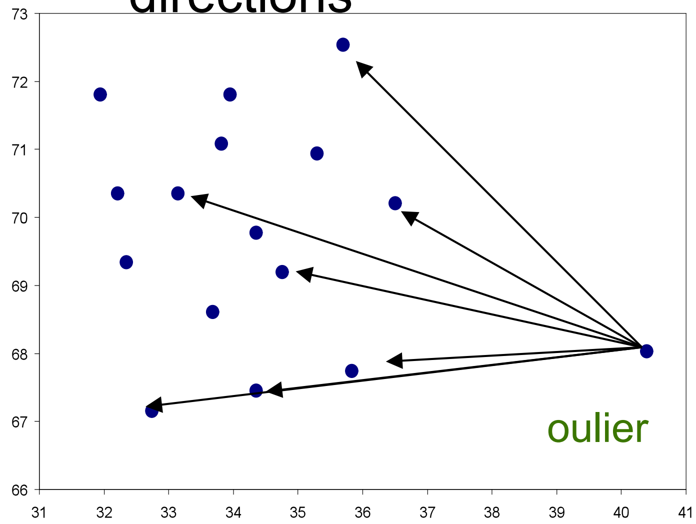
- Density-based Detection
  - Relative density of object counted
  - from density of its neighbors

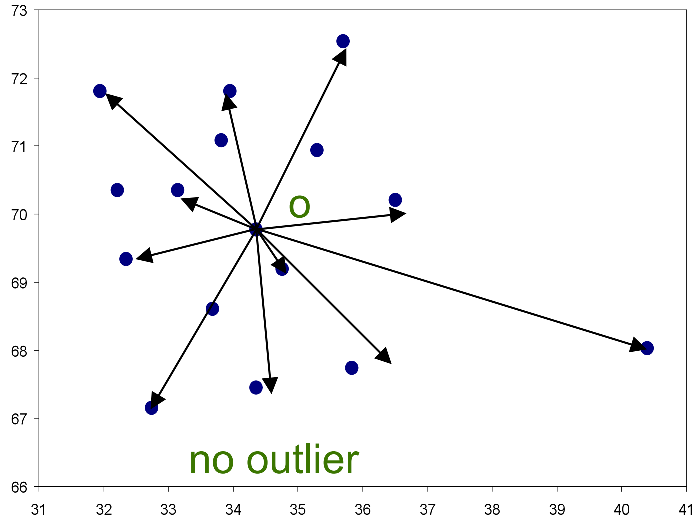
- Clustering-Based Methods
- Normal data objects belong to large
- and dense clusters, whereas outliers
- belong to small or sparse clusters,
- or do not belong to any clusters.



High-dimensional Outlier Detection Methods

- ABOD – angle-based outlier degree
  - Object o is an outlier if most other objects are located in similar directions
  - Object o is no outlier if many other objects are located in varying directions





Outlier Detection Methods Types

- Supervised Methods
  - building a predictive model for normal vs. anomaly classes
- Semi-supervised Methods
  - training data has labeled instances only for the normal class
- Unsupervised Methods
  - no labels, most widely used

Supervised Methods

- building a predictive model for normal vs. anomaly classes
  - problem is transformated to **classification problem**

- Any supervised learning algorithm
  - E.g. a decision tree
  - how to detect outliers

Supervised Methods (cont.)

- Problems:
  - anomalous instances are far fewer than normal instances
  - obtaining acurate labels for the anomaly class is challenging

Semi-supervised Methods

- training data has labeled instances only for the normal class
- one-class learning
- e.g. One-class SVM
- Clustering (e.g. EM algorithm)
- Normal data instances lie close to their closest cluster centroid,
- while anomalies are far away from their closest cluster centroid.
-

Unsupervised Methods

- no labels, most widely used
- assumption: normal instances are far more frequent than anomalies in the data and they make clusters
- Proximity-based methods, clustering
- **Global methods**:

  kNN – outlier factor == sum of distances to k nearest neighbors
- **Local methods**:
- **LOF**, Local Outlier Factor

Local Outlier Factor (LOF)

$dist_k(o)$ . . . k-distance of an object o . . . **distance from o to its kth nearest neighbor**

$N_k(o)$ k-distance neighborhood of o . . . set of k nearest neighbors of o

reach.distk(o, p) = max{$dist_k(p)$, dist(o,p)} . . .
reachability-distance of an object o with respect to another object p

The local reachability-distance is the inverse of the average reachability-distance of its k-neighborhood.

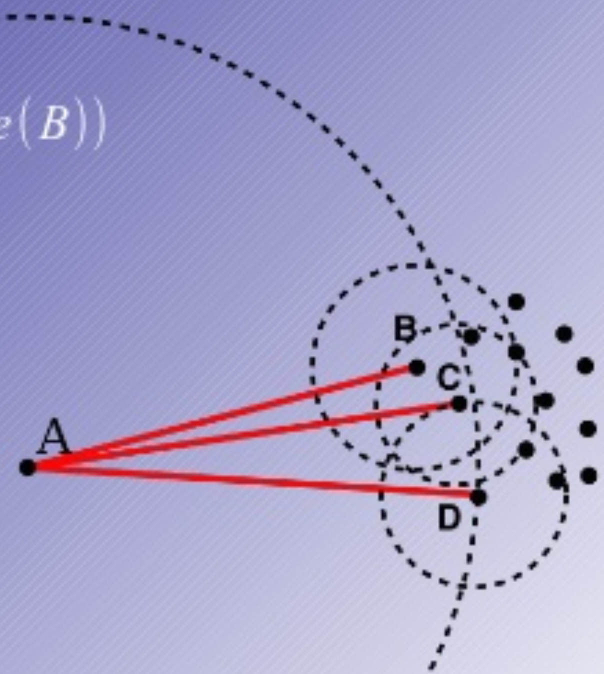
LOF is the average of the ratio between the local reachability-distance of o and those of its k-nearest neighbors.

Local Outlier Factor

$$reach-dist_k(A,B)=max(d(B,A),k-distance(B))$$

$$lrd(A)=\frac{1}{\sum_{B\in KNN(A)} reach-dist_k(A,B)/k}$$

$$LOF(A)=\frac{\frac{1}{k}\sum_{B\in KNN(A)} lrd(B)}{lrd(A)}$$



https://en.wikipedia.org/wiki/Local_outlier_factor

Evaluation of anomaly detection methods

Supervised settings – easy, precision/recall

Semi-supervised, unsupervised methods:

Need for classified data

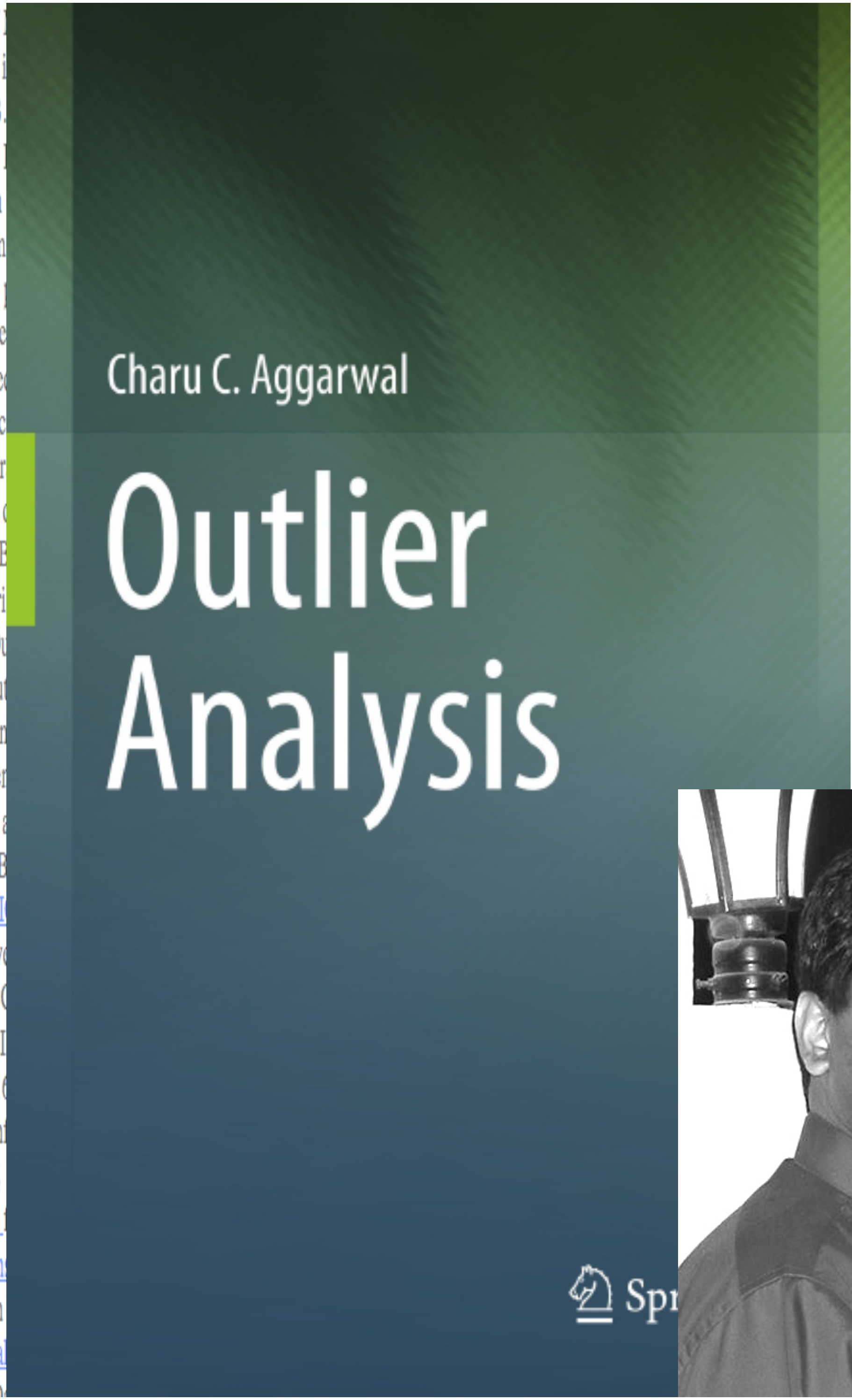
1. Two class data, e.g. from UCI,
   1st class aka normal, the 2nd is a source of anomalies

2. Artificial data generator
   more flexible

Implementations

- R : e.g. mvoutliers, DMwR and many others
- scikit-learn: Robust covariance, One Class SVM, Isolation Forest, Local Outlier Factor
- ELKI https://elki-project.github.io/
- OutRules: A Framework for Outlier Descriptions in Multiple Context Spaces, Univ. Saarbruecken
  http://www.ipd.kit.edu/~muellere/OutRules/
  based on WEKA http://www.cs.waikato.ac.nz/ml/weka/

- Charu Aggarwal is a Distinguished Research Staff
(DRSM) at the IBM T. J. Watson Research Center i
Yorktown Heights, New York. He completed his B.
Kanpur in 1993 and his Ph.D. from Massachusetts
Technology in 1996. He has worked extensively in
data mining, with particular interests in data stream
uncertain data and social network analysis. He has
14 (3 authored and 11 edited) books, over 250 pape
refereed venues, and has applied for or been grante
patents. His h-index is 80. Because of the commerc
the above-mentioned patents, he has received sever
invention achievement awards and has thrice been
a Master Inventor at IBM. He is a recipient of an IB
Corporate Award (2003) for his work on bio-terrori
detection in data streams, a recipient of the IBM Ou
Innovation Award (2008) for his scientific contribut
privacy technology, a recipient of two IBM Outstan
Technical Achievement Awards (2008) for his scien
contributions to high-dimensional and data stream
He has received two best paper awards and an EDB
Time Award (2014). He is a recipient of the IEEE I
Research Contributions Award (2015). He has serv
general or program co-chair of the IEEE Big Data
(2014), the ICDM Conference (2015), the ACM CI
Conference (2015), and the KDD Conference (201
co-chaired the data mining track at the WWW Con
2009. He served as an associate editor of the IEEE
Transactions on Knowledge and Data Engineering
to 2008. He is an associate editor of the ACM Tran
Knowledge Discovery and Data Mining , an action
the Data Mining and Knowledge Discovery Journa
associate editor of the IEEE Transactions on Big D




Outlier Detection: Beauty and the Beast in Data Analytics
Because of ….