Machine Learning Rhein-Neckar Meetup

An Overview of Outlier detection techniques and applications

Ying Gu

28.02.2016

Anomaly/Outlier Detection

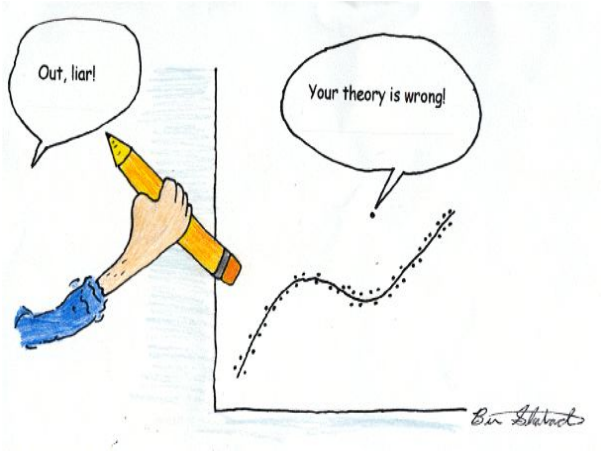
What are anomalies/outliers?

- The set of data points that are considerably different than the remainder of the data anomaly
- On a scatter plot of the data, they lie far away from other data

Goal of anomaly detection

- Find objects that are different from most other objects

Anomaly detection, deviation detection, outlier detection, novelty detection...



Applications

- **Fraud Detection:** The purchasing behavior of someone who steals credit card is probably different from that of the original owner. We can detect a theft by
  - looking for buying patterns that characterize theft
  - noticing a change from typical behavior

- **Intrusion Detection:** Attacks on computer systems and computer networks
  - Attacks designed to disable or overwhelm computers and networks are obvious
  - Attacks designed to secretly gather information are difficult to detect
    - Can only be detected by monitoring systems and networks for unusual behavior

Applications

- **Ecosystem Disturbances:** typical events in nature
  - Hurricanes, floods, droughts, heat waves and fires
    - Predict the likelihood of these events
    - Predict the causes of them

- **Medicine:** unusual symptoms or test results may indicate potential health problems for a particular patient
  - Whether a particular test result is anomalous may depend on other characteristics of the patient, such as age and sex
  - The categorization of a result as anomalous or not incurs a cost!

Definition

Definition of Hawkins [Hawkins 1980]

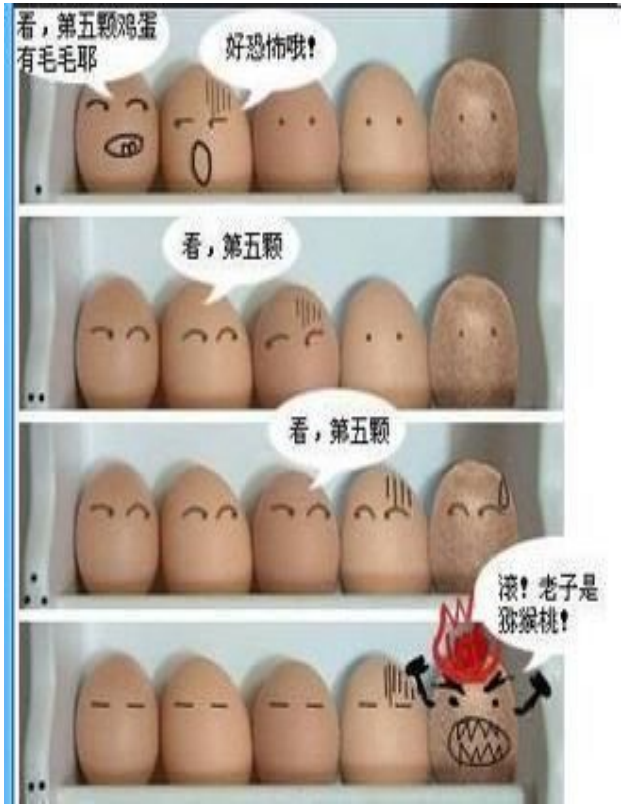
- “An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism”

Statistics-based intuition

- Normal data objects follow a “generating mechanism”, e.g. some given statistical process
- Abnormal objects deviate from this generating mechanism

Cause of Anomalies

- **Data from different classes**
  Anomalous, becauses it is of a different type or classes
  - Some committing credit card fraud belongs to a different class of credit card users than those people who use credit cards legitimately
  - Fraud, intrusion, outbreaks of disease, abnormal test results



Causes of Anomalies

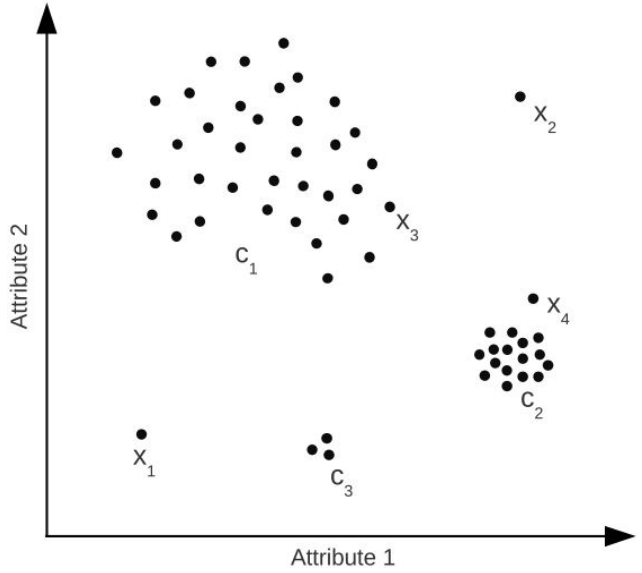
- **Natural Variation**
  Most of the objects are near a center (average) and the likelihood that an object differs significantly from this average object is small.

- **Data Measurement and Collection Errors**
  Errors in the data collection or measurement process
  - Because of human error, a problem with measuring device or the presence of noise.
  - Goal: Eliminate such anomalies, since they provide no interesting information but only reduce the quality of the data. (data preprocessing, specifically data cleaning)

Type of Anomalies

- **Point Anomalies**

  A single instance, which differs from the norm with respect to its attribute values
  - Easily identifiable by humans with an appropriate visualization
  - Anomaly detection algorithms, which are able to detect such anomalies, are sometimes called **point anomaly detection algorithms**



The two anomalies x_1 and x_2 are easily identifiable, x_4 could also be an anomaly with respect to its direct neighborhood whereas x_3 seems to be one of the normal instances.

Type of Anomalies

- **Contextual Anomalies**

  Anomalies, which can only be identified with respect to a specific context

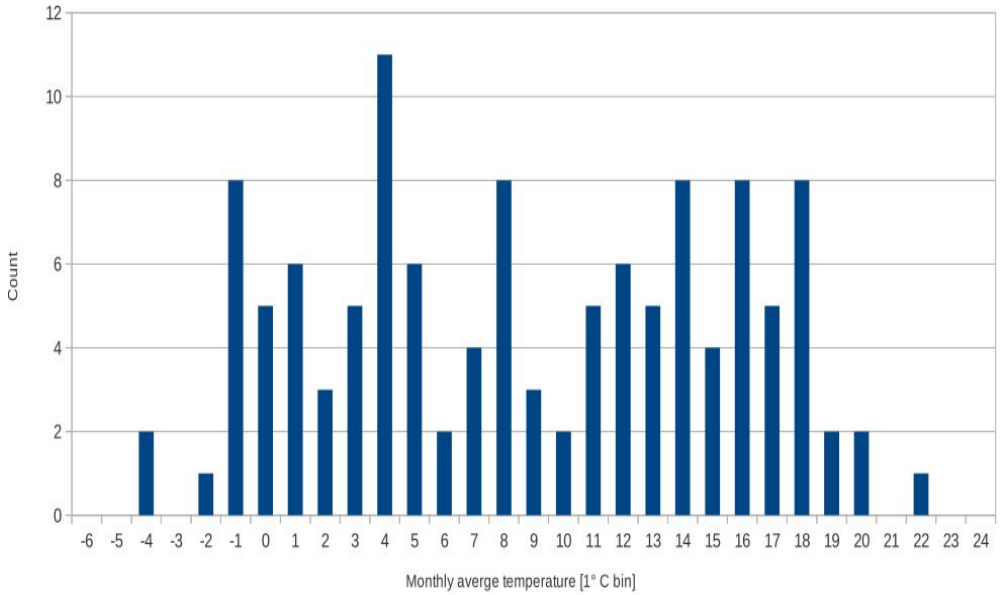
  - Need to integrate the context in the data, so that point anomaly detection algorithms can be applied.

    There are many ways to do this, but it usually needs domain knowledge. For example:

    - Add additional attributes
    - Aggregation and binning with respect to multiple dimensions of data
  - The procedure of integrating the context is also known as the generation of a **data view**.

Example of Contextual Anomalies

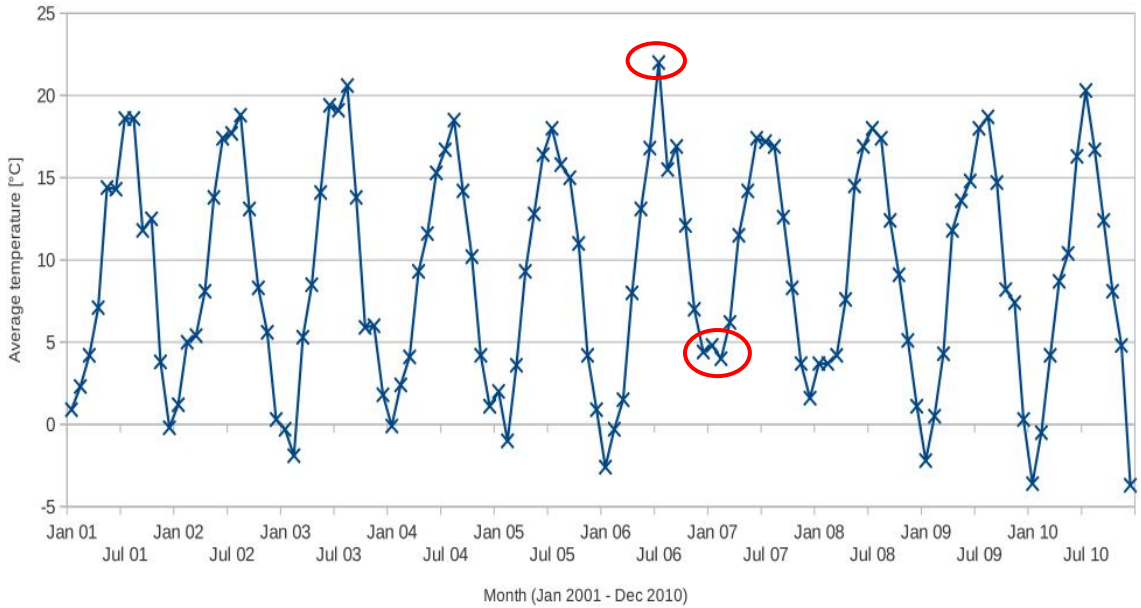
Measurements of the average monthly temperatures in Germany over ten years from 2001 to 2010. (N=120 measurements).



A histogram plot of the average monthly temperatures in Germany from 2001 to 2010. One obvious anomaly is one very hot month in the bin 22-23°C

Example of Contextual Anomalies

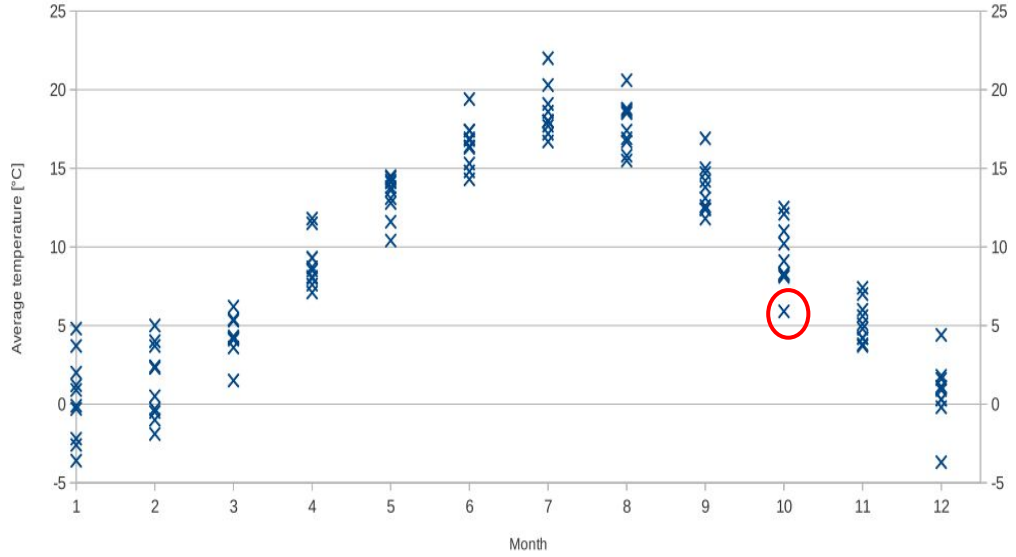
Take the context into account.
The following figure plotted the measurements with respect to the context time



We can not only see the previously mentioned extreme outliers, but also two very mild winters in 2006 and 2007, which we could not observe previously.

Example of Contextual Anomalies

We could add the month as a numerical attribute. A point anomaly detection algorithm would now be able to detect a few more outliers besides the extreme temperature values



The month of October in 2003 turns out to be an anomaly with only 5.9°C on average, which is only half of the average temperature in that month.

Type of Anomalies

- **Collective Anomalies**

  A combination of multiple instances forms an anomaly, whereas each single one of these contributing instances is not necessarily an anomaly in itself. They are the most complex type of anomalies.

  - Generate data view to transfer collective anomalies into point anomalies
  - Domain knowledge are needed
  - Example: host-based intrusion detection system
    - Buffer overflow happens alone can not be considered as anomalous
    - Buffer overflow followed by creating a new user account, then it will be very suspicious

Approaches to Anomaly Detection

- Model-based Techniques
  - Build a model for the data, the anomalies do not fit the model well.
  - statistic methods, classifications
- Proximity-based techniques
  - Define a proximity measure between objects, anomalous objects are those that are distant from most of the other objects. Also called distance-based outlier detection techniques.
  - k-NN based methods
- Clustering-based techniques
  - Outlier detection can be regarded as complementary of clusterings.
  - k-Means

The Use of Class Labels

- **Supervised anomaly detection**
  Uses normal instances and anomalies for training and testing. The model typically classifies the instances into two classes

- **Semi-supervised anomaly detection**
  Uses only normal data for training. The model should be able to detect deviations in the test data from that norm.

The Use of Class Labels

- **Unsupervised anomaly detection**
  Uses no labeling information at all. The algorithm only takes intrinsic information of the data into account in order to detect anomalous instances being different from the majority.

Output

- Binary class label, normal or anomalous (Early years)
- Numerical output (anomaly score), can be converted to binary output
  - Used by semi-supervised and unsupervised algorithms
- The interpretation of outlier score depends on the algorithm.
  - Some algorithms: only an ordering according to the scores make sense(ordinal scale)
  - Other algorithms: reference point, 1.0 is normal behavior
  - Probability (not possible with most algorithms), so the scores can be compared by different data set.

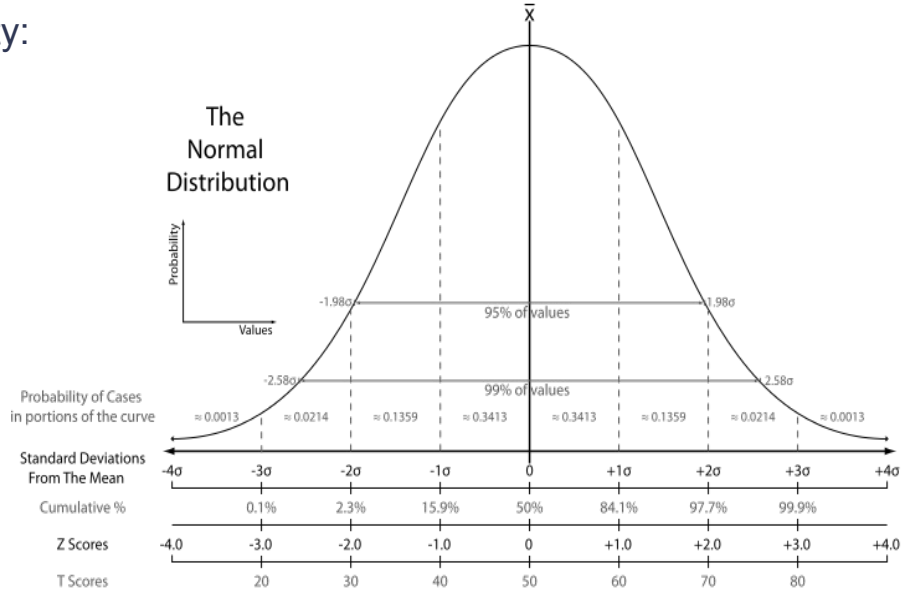
Approach

– Detecting Outliers in Gaussian Distribution

- Gaussian Distribution has a nice property:
  - $p(x>\mu+\sigma)$ and $p(x<\mu-\sigma)=31.73\%$
  - $p(x>\mu+2\sigma)$ and $p(x<\mu-2\sigma)=4.5\%$
  - $p(x>\mu+3\sigma)$ and $p(x<\mu-3\sigma)=0.27\%$

- In practice, define:
  $|x|>\mu+2\sigma$, the x is an anomaly.

  This is the reason why outliers should not more than 5%



Approach
– k-NN Distance-based

- Based on the distances to the neighbors
  - The distance to the k-th Nearest-Neighbor is computed
  - The average distance to all the k-Nearest-Neighbors is used as a score

$$score_{knn}(x) = \frac{\sum_{o \in N_k(x)} d(x,o)}{|N_k(x))|}$$

$N_k(x)$: the k-nearest neighbor set

- d(x,o): the distance between x and o
- This approach is preferable, because it results in a much more robust local density estimation

Evaluation
ROC Curve (Receiver Operating Characteristic)

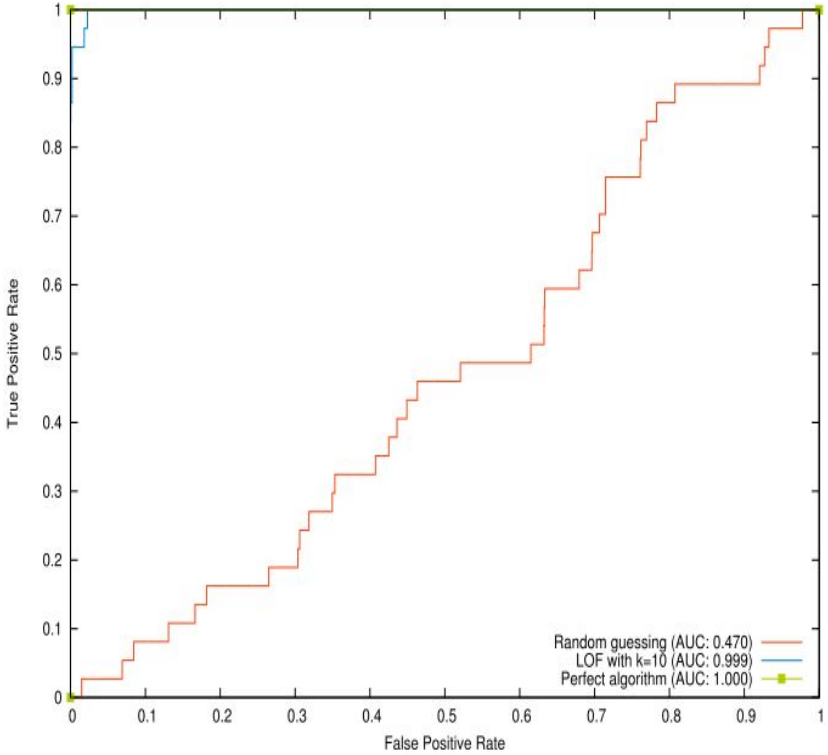
ROC curve can also be used in the evaluation of the algorithms
Computation is the same as the typical ROC computation :

- TPR: True Positive Ratio
- FPR: False Positive Ratio

Notice: The different rates are computed by varying the threshold of the outlier score.

AUC is the integral of the ROC. It is the probability that an anomaly detection algorithm will assign a randomly chosen normal instance a lower score than an randomly chosen outlying instance. Hence, the AUC is a good good quality measure, the higher the value, the more likely it is that anomalies are detected.

Example of ROC Curves



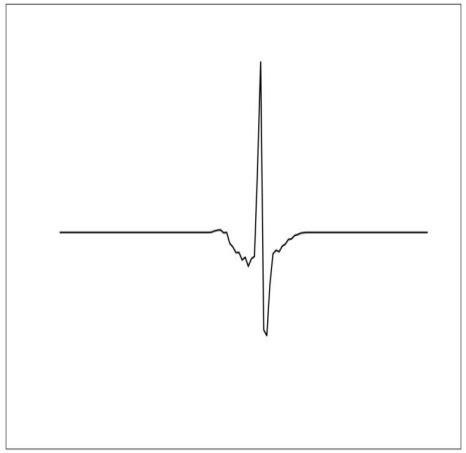
The blue ROC originates from the LOF algorithm applied on the 2D artificial dataset

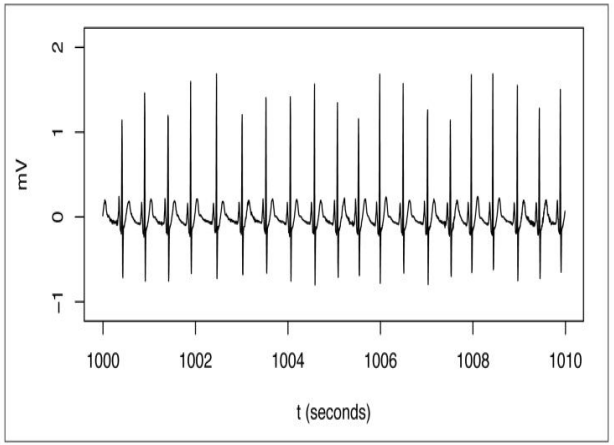
The red curve represents a random guessing approach and the green curve is the result of a perfect algorithm.

Time Series Anomaly Detection

Anomalies are defined not by their own characteristics, but in contrast to what is normal.

-- Ted Dunning



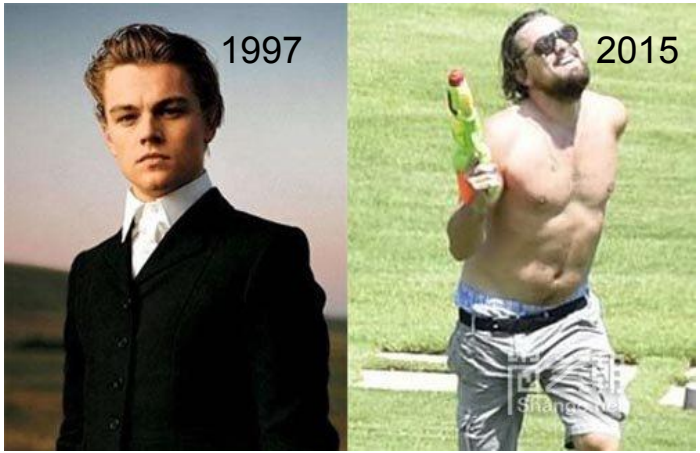


Methods:

- statistic methods
- Turn the time series into points

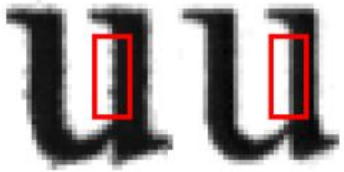
Research

- Speed up the Algorithms
- Big Data (Cloudera Hadoop, Apache Flink)
- Evolving Algorithms
  - Time can changing everything



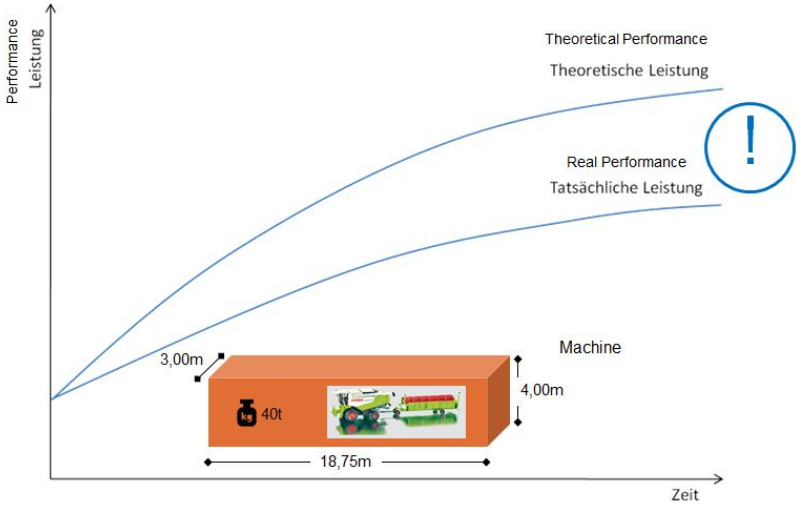
Projects (DFKI)

1. Find the faking documents



2. Intrusion Detector (with Telecom)
3. Indoor location tracking (with VW)

4. Machine Learning for Self-propelled Harvesters (with CLAAS)
   a. Fault Detection and Maintenance Studies
   b. Productivity Studies



Outlier Detection with RapidMiner

Dataset 1

DFKI-artificial-3000:

download: http://www.madm.eu/_media/downloads/dfki-artificial-3000-unsupervised-ad.zip

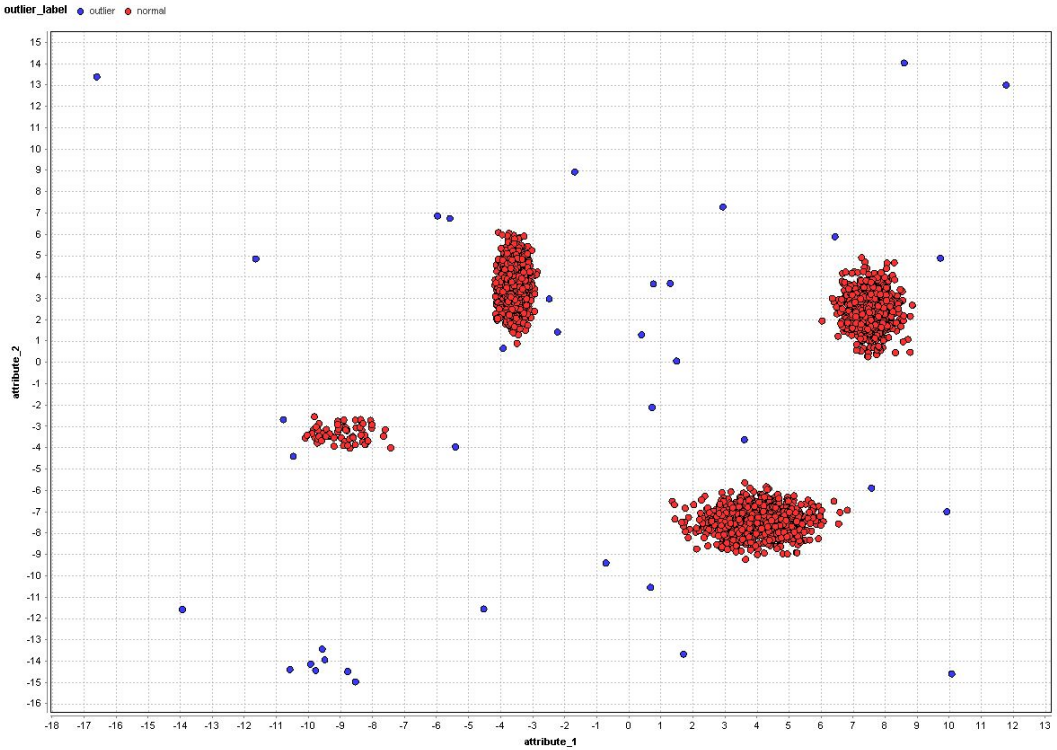
3000 Records, 2 attributes

outlier_label:

- outlier (37)
- normal (2963)

Row No.	outlier_label	attribute_1	attribute_2
1	outlier	-9.799	-14.403
2	outlier	-10.606	-14.356
3	outlier	-9.959	-14.101
4	outlier	-9.516	-13.903
5	outlier	-8.798	-14.451
6	outlier	-8.556	-14.934
7	outlier	-9.596	-13.403
8	outlier	-3.946	0.692
9	outlier	-5.997	6.899
10	outlier	-11.670	4.890
11	outlier	-2.256	1.458

Dataset 1



Dataset 2

Diagnose breast cancer

Download: https://dataverse.harvard.edu/api/access/datafile/2711924?format=original

30 Attributes, 367 Records

Label:

- o = Outlier (Malignant, 10)
- n = Normal (Benign, 357)

Row No.	att31	att1	att2	att3	att4	att5
1	o	17.990	10.380	122.800	1001	0.118
2	o	20.570	17.770	132.900	1326	0.085
3	o	19.690	21.250	130	1203	0.110
4	o	11.420	20.380	77.580	386.100	0.142
5	o	20.290	14.340	135.100	1297	0.100
6	o	12.450	15.700	82.570	477.100	0.128
7	o	18.250	19.980	119.600	1040	0.095
8	o	13.710	20.830	90.200	577.900	0.119
9	o	13	21.820	87.500	519.800	0.127
10	o	12.460	24.040	83.970	475.900	0.119
11	n	13.540	14.360	87.460	566.300	0.098
12	n	13.080	15.710	85.630	520	0.107
13	n	9.504	12.440	60.340	273.900	0.102
14	n	13.030	18.420	82.610	523.800	0.090
15	n	8.196	16.840	51.710	201.900	0.086
16	n	12.050	14.630	78.040	449.300	0.103