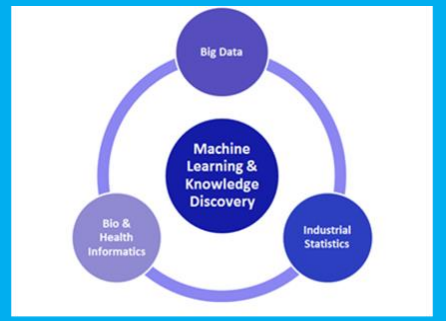
521156S
TOWARDS DATA MINING

MATKALLA TIEDONLOUHINTAAN

UNIVERSITY OF OULU

BISG Oulu



Data Analysis and Inference Group

How to **prepare** your data to make sure that your data mining process will be successful…

Tuomo Alasalmi
tuomo.alasalmi@oulu.fi

TOPICS OF THE LECTURES

1. Introduction to data preprocessing
2. Data ethics, data security, privacy and open data
3. Data management and databases
4. Data gathering
5. Missing data
6. **Noise, outliers, signal saturation**
7. Normalization, transformation, dependence, distributions
8. Feature selection and ensuring the generality of the results

Data on uusi vesi

Datan mahdollistamat uudet palvelut ja liiketoiminnot ovat innoittaneet analyytikoita ja it-taloja vertaamaan dataa öljyyn. Vertauksissa data mahdollistaa samanlaisen murroksen kun öljyn hyödyntäminen 1900-luvun taitteessa.

Näissä analogioissa on aina ongelmansa. Nykyisin on mahdollista tehdä esimerkiksi analytiikkaa, tiedolla johtamista ja tekoälysovelluksia ennennäkemättömällä tavalla. Dataa on saatavilla ylen määrin joka puolella. Tässä kohtaa vertaus öljyyn ontuu.

ÖLJY ON rajallinen resurssi, jonka viimeisten pisaroiden hyödyntäminen on koko ajan vaikeampaa. Dataa puolestaan on yllin kyllin. Hyödyntämisessä ongelmana on enemmänkin se, että läheskään kaikki saatavilla oleva data ei ole yrityksille hyödyllistä. Virheellinen tieto on jopa haitallista.

Yhtä pätevä vertaus voisikin olla vesi. Vettä on maapallolla todella paljon, mutta vain pieni osa siitä on sellaisenaan ihmiselle juomakelpoista. Datastakaan ei voi hyödyntää kuin pienen osan ilman puhdistamista.

JOS PUHDISTETTUUN dataan sekoittaa likaista dataa, menee lopputulos helposti käyttökelvottomaksi. Jos pohjalla olevat bitit eivät ole kohdallaan, on asiakkaan kokema palvelu vähintäänkin kulmakarvoja kohottavaa.

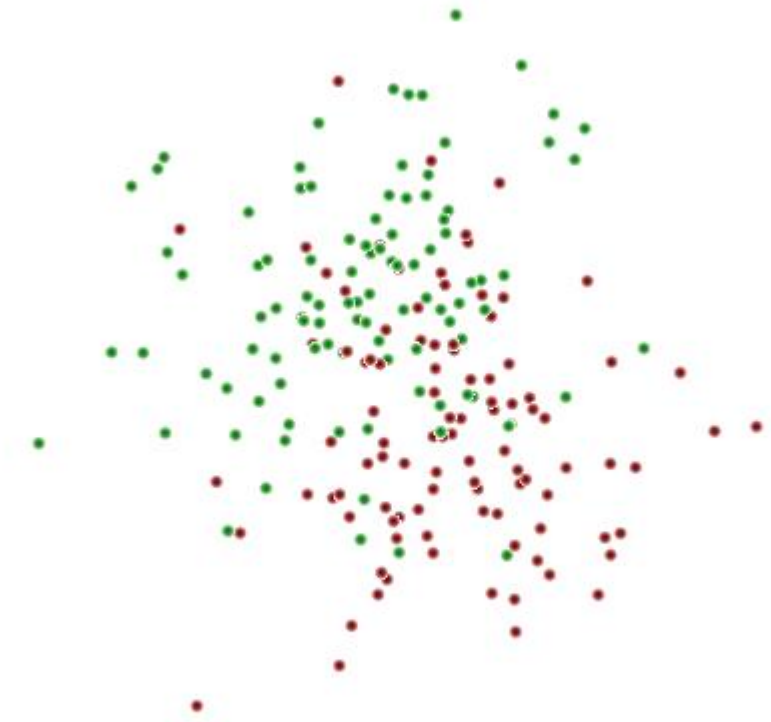
Näin äskettäin esimerkin rengashotellista. Asiakkaan mobiilisovellukseen tuli raportti säilytyksessä olevien renkaiden kulumisesta. Raportissa oli yksi selvästi kulunut rengas, johon piti kiinnittää huomiota. Palvelu on sinällään upea, mutta valitettavasti asiakkaan hotelliin toimittamat renkaat olivat iskemättömät.

Erityisen vaativaa on luoda luotettavaa pohjaa erilaisista big data -lähteistä. Tiedon rakenne on aina puutteellista ja sen ajantasaisuutta ei pysty varmistamaan.

MYÖNNETÄÄN, ei data ole myöskään uusi vesi. Vertailukohtia on mukava hakea kaikille tutuista elementeistä. Data on raaka-aineena kiitollinen, sen määrä kasvaa eksponentiaalisesti. Hyvääkin dataa riittää aika monelle, kunhan vain pidämme huolen sen puhdistamisesta.

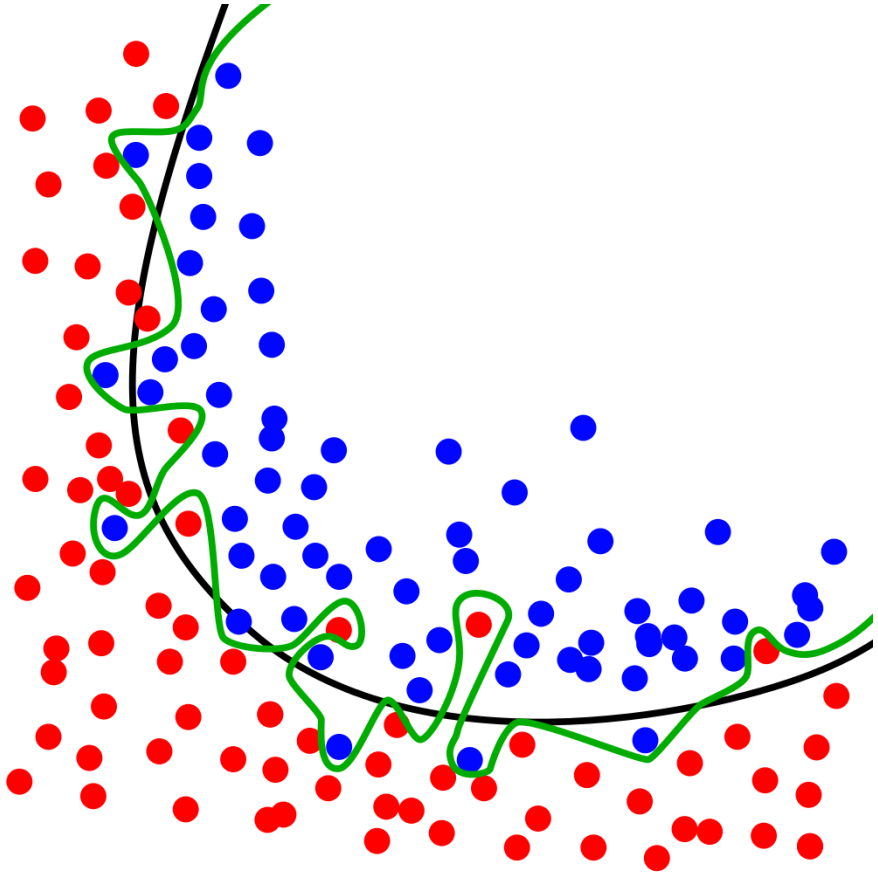
MOTIVATION

- **Data is everywhere and everyone wants to make use of it**
- **Raw, uncleaned data is not always very usefull**
- **Even the mainstream is starting to wake up to this (it's about time!)**
  - On the left is an editorial of an IT magazine TIVI from wk 40/2017 making the point that only some of data is usefull without cleaning
- **Even if we (well I don't) like to talk about artificial intelligence nowadays, there really isn't any real intelligence in these systems (at least yet)**
- **Without intelligence what we'll get instead is garbage in, garbage out**
- **We need to be the real intelligence in the process to make sure garbage is not what our models output**



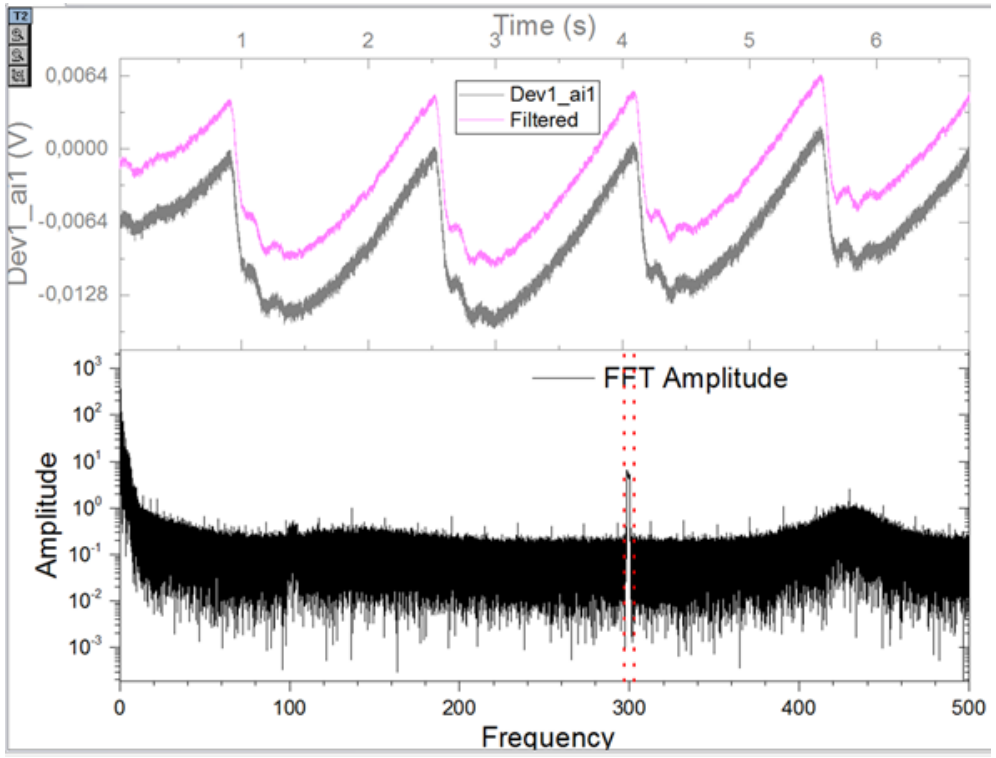
NOISY DATA

- **In addition to normal variation, data can contain noise from various sources**
  - Label noise: incorrect label (mistake, subjective error, etc.)
  - Attribute noise: errors, missing or unknown values
- **In classification, noise can negatively affect the system performance**
- **Noise may create small clusters of instances of a particular class in parts of the instance space corresponding to another class, remove instances located in key areas within a concrete class or disrupt the boundaries of the classes and increase overlapping among them**
- **Especially relevant in supervised problems**



NOISY DATA

- **Dealing with noisy data**
- **Robust learners**
  - The capability of an algorithm to build models that are insensitive to data corruptions and suffer less from the impact of noise
  - Models are similar with clean and noisy data
  - E.g. C4.5 classification tree uses pruning strategies to reduce the possibility that the tree overfits to noise
- **Data polishing**
  - Aim to correct noisy instances prior to modelling
  - Only feasible with small data sets



NOISY DATA

- **Dealing with noisy data**
- **Noise filters**

- Identify and remove noisy (and possibly borderline) instances
- From continuous signals (e.g. IR absorbance in the figure), filter out the clearly identifiable noise frequencies
  - Equipment errors
  - 50 Hz from the power lines
  - Can also lose valuable information!

NOISY DATA

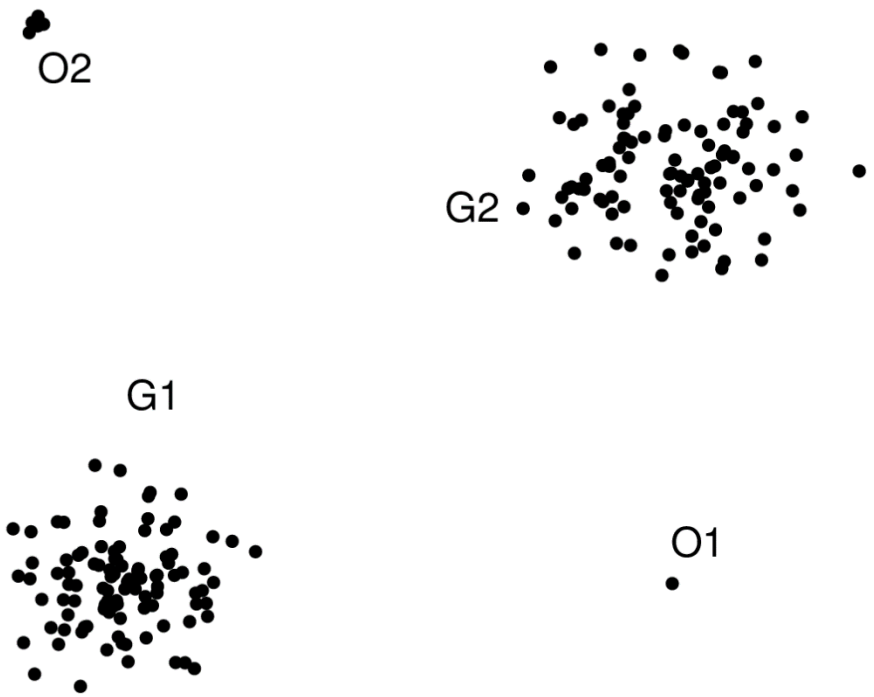
– Data pollution

– Data sets can contain values that are not intended to be in it

- Gender field might contain values male, female and business because the system was originally designed to handle personal data but is used with business customers, too

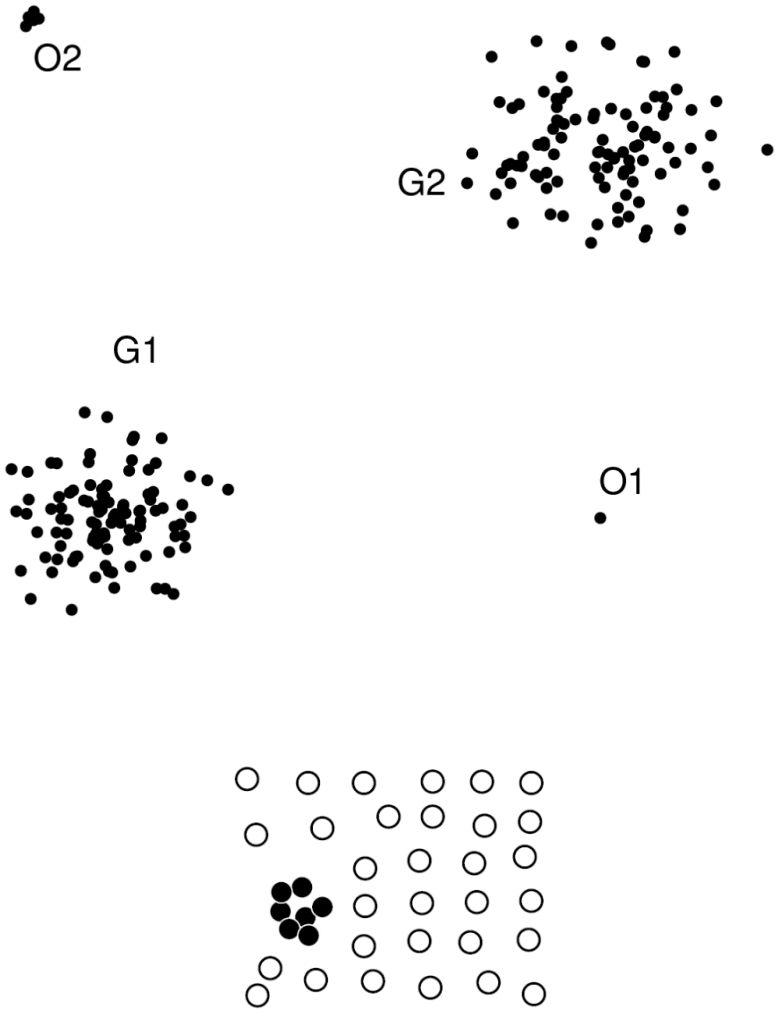
– Sometimes fields contain unidentifiable garbage

- Manual copy and paste errors
- In a csv file, the field values contain the delimiter character
  - Free text fields
  - Decimal point
  - Thousand separator
- Can be hard to find the source of the problem



OUTLIERS

- **Data examples with behaviours that are very different from the expectation**
  - Individual cases (O1) or clumps (O2)
- **It is important to investigate why there are outliers**
  - What sort of process could account for them?
  - Are they just extremes of the range? Different equipment with different biases or calibration? Etc.
  - Can they be translated back into normal range if they are indeed errors?
- **If no rationale can account for the outliers, you can treat them as you would MVs**
  - Same caveats apply here as with MV handling
- **Some modelling methods suffer more (e.g. NN) than others (e.g. SVM)**



OUTLIER TYPES

– Global outliers (O1 and O2 in the figure)

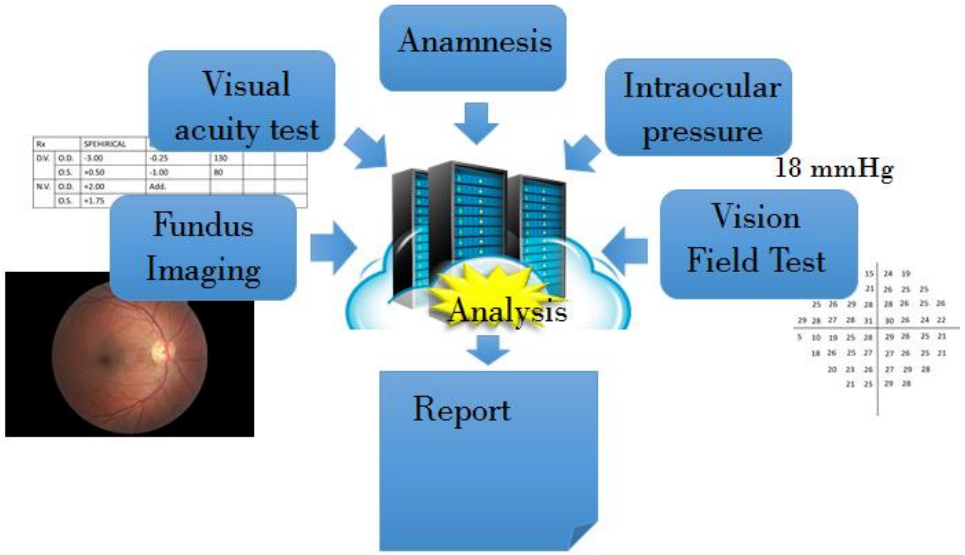
- Deviates significantly from the rest of the data set
- Most outlier detection algorithms are aimed at finding global outliers

– Contextual outliers

- Whether a value is considered outlier or not depends on the context
- E.g. +25°C outside temperature in Oulu in January is an outlier but in July it is not

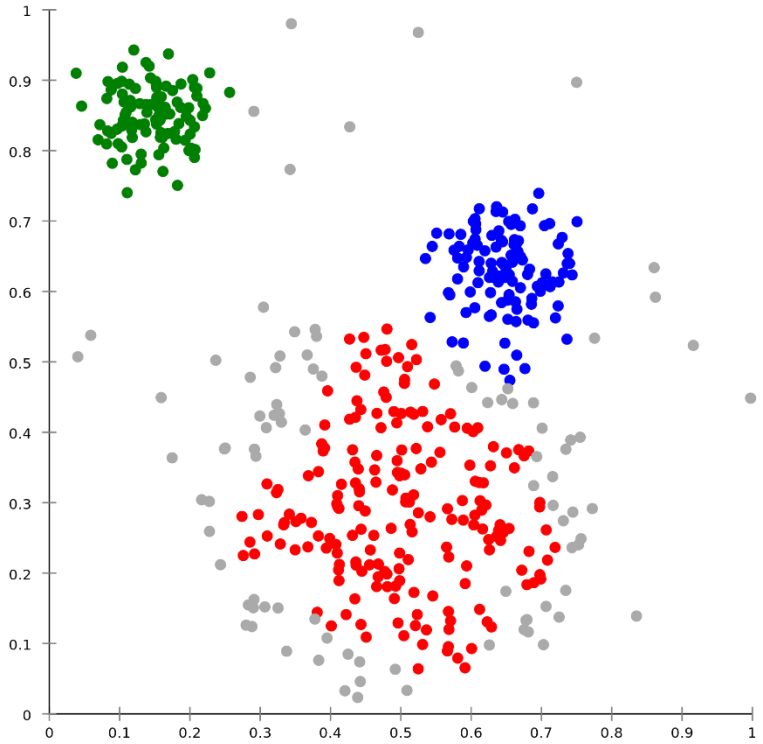
– Collective outliers (lower figure)

- A group of objects as a whole deviate significantly from the entire data set
- In the group, individual instances would not be considered outliers
- A stock transaction between two parties: normal
- Large set of transactions of the same stock in a short period of time between the same parties: outlier? (market manipulation?)



OUTLIER DETECTION

- **Building a comprehensive model for data normality is challenging**
  - Hard to enumerate all possible normal behaviours
  - The border between normal data and outliers is not clear cut, more like a gray area
  - The cutoff point is application specific: no universal rules
  - Noise in the data makes outlier detection harder
- **In some application scenarios, it is important to not just detect outliers but to also provide justification for the detection**
  - CRYSTAL-project example: detecting early signs of eye disease, the model cannot be a black box or it won't be taken to use at all
- **Also related to novelty detection**
  - E.g. Identify novel topics and trends on social media



OUTLIER DETECTION

– Supervised

- If labels are available, a classifier can be trained to detect outliers
  1. Model the normal, detect instances not matching the model
  2. Model the outliers
- The number of instances can be very imbalanced between normal and outlier classes
- The challenge is to balance outlier detection sensitivity and to not mislabel too many normal instances

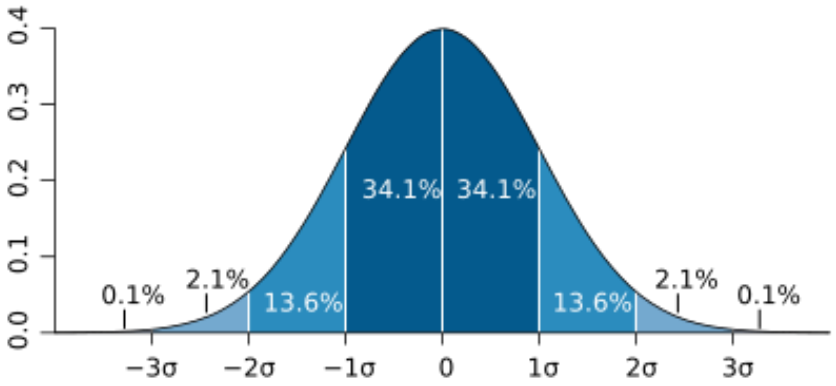
– Unsupervised

- Assumption: normal instances are somewhat clustered (can be several clusters)
- Detect instances far away in feature space from the normal clusters

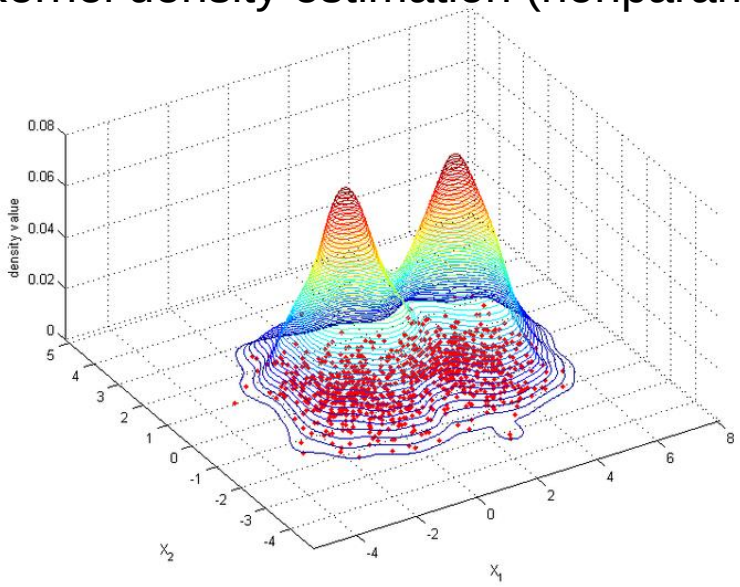
– Semi-Supervised

- Use labeled and close by unlabeled instances together
  - Better for modelling normality than abnormality

Normal distribution (paramteric):



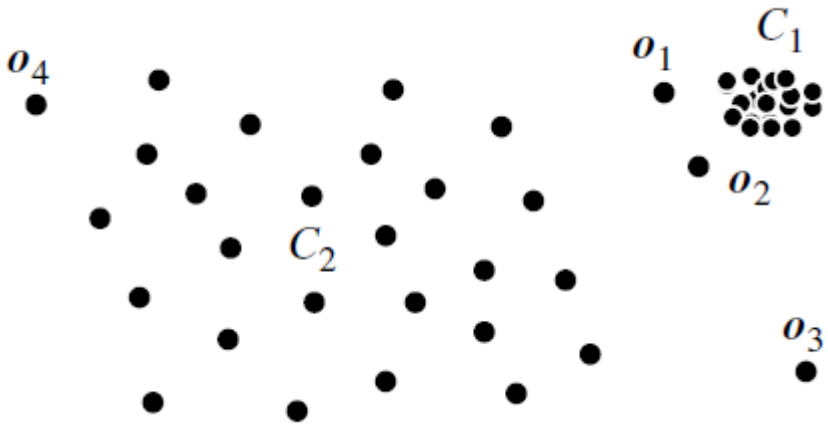
Kernel density estimation (nonparamteric):



OUTLIER DETECTION

– Statistical (model based) methods

- Assume that normal data objects are generated by a statistical model
- Data not following the model are considered outliers
- Effectiveness depends on how well the assumptions for the model hold true
- Parametric models
  - Normal data objects are generated from a parametric probability distribution
  - Probability that an instance is generated from the distribution can be calculated
- Nonparamteric models
  - No a priori statistical model is assumed
  - The model is determined from the input data



OUTLIER DETECTION

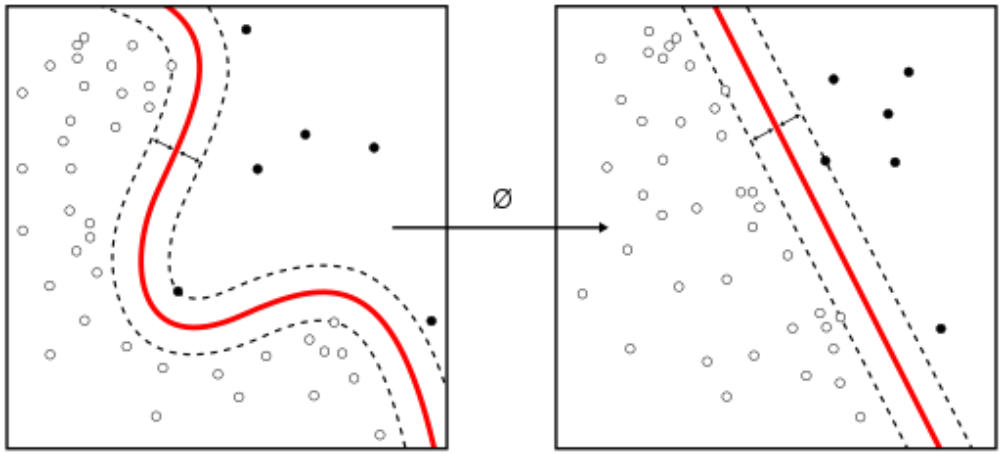
– Proximity based methods

- Assume that an object is an outlier if the nearest neighbours are far away in the feature space compared to other objects in the same data set
- Effectiveness relies on the distance measure used
- Hard to detect groups of outliers if they are close to each other
- Distance based
  - Determine how many objects reside within a predefined distance (r) from the object
  - If this falls below a threshold (t), the object is considered an outlier
  - Global parameters (r,t)
  - Good for global outliers
- Density based
  - Compare density of objects around the object and compare with the density of its local neighbours

OUTLIER DETECTION

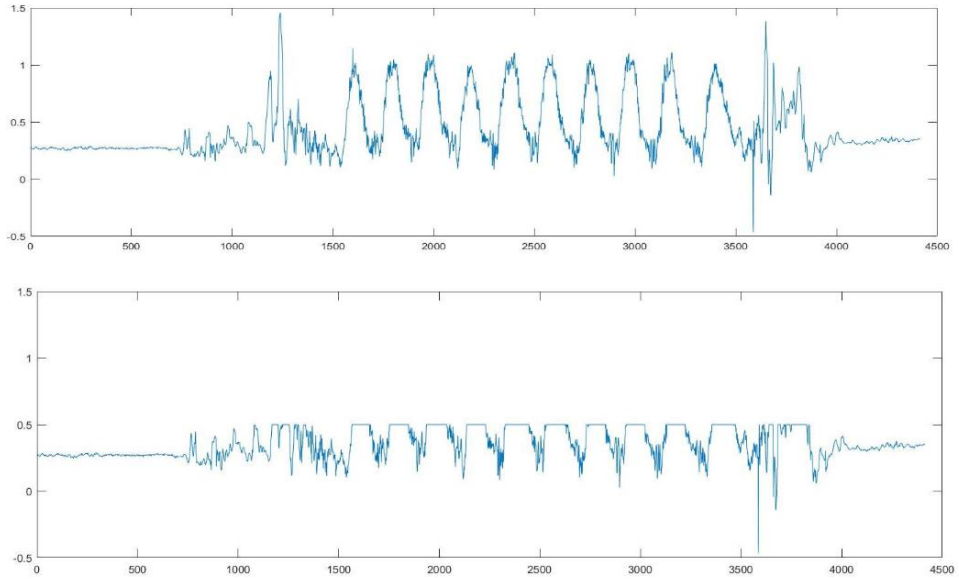
– Clustering based methods

- Assume that normal data belong to large and dense clusters whereas outliers belong to small or sparse clusters or do not belong to any cluster
- Clustering is an expensive method and does not scale very well to large data sets



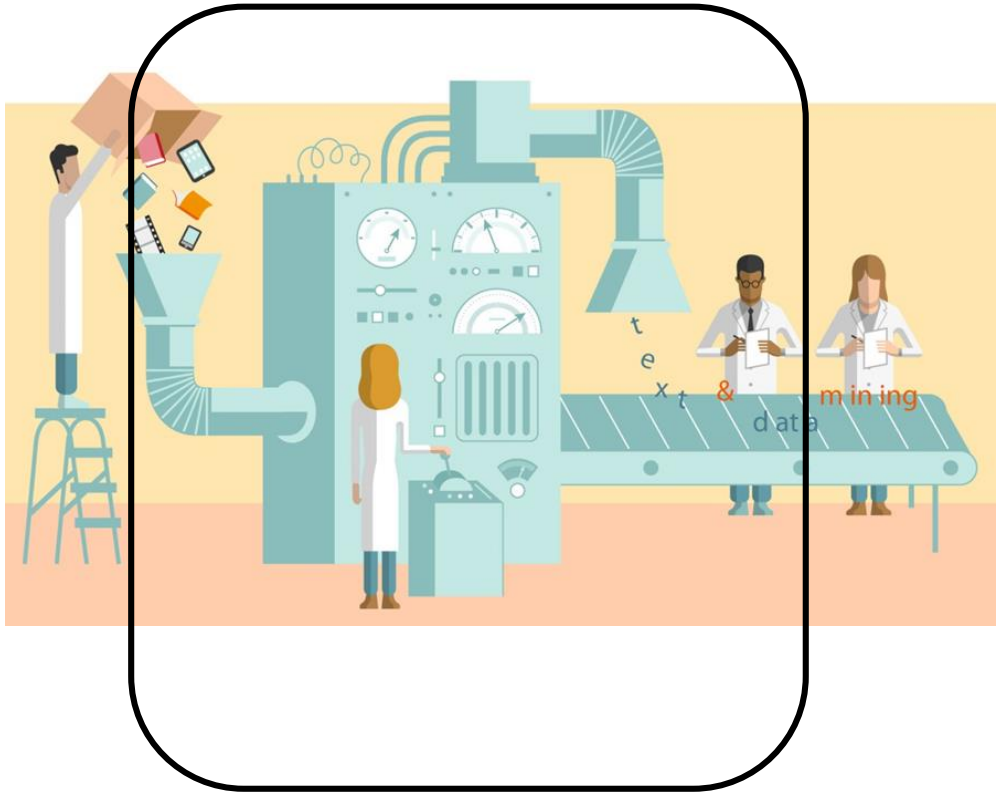
– Classification based methods

- Any classifier can be used
- Problem is having enough and a representative sample of outliers in the training data
  - One solution is to use one-class models
  - A classifier is build to describe the normal class
  - Any samples that don't belong to the normal class are considered outliers



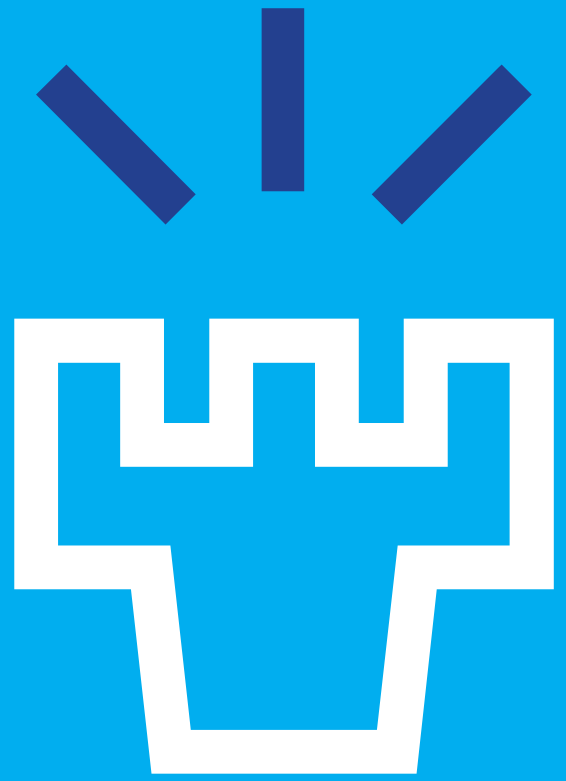
SIGNAL SATURATION

- Not much can be done after data collection but it is important to realize there is a problem
- Minimum or maximum values (or both) are overrepresented in the data
- Several minimum or maximum values occur as consecutive samples



CONCLUSIONS

- Real world data sets contain problems like outliers and noise
- Be aware of these
- With small data sets, data polishing can be used to handle noise
- Using robust learning algorithms is suggested for larger data sets
- There are no hard rules on how to handle or detect outliers, it always depends
- Get to know your data and find out why there are outliers (and if they really are outliers), this will lead you into right direction on what to do



UNIVERSITY
OF OULU