Data Quality & Preprocessing

Lecture 18

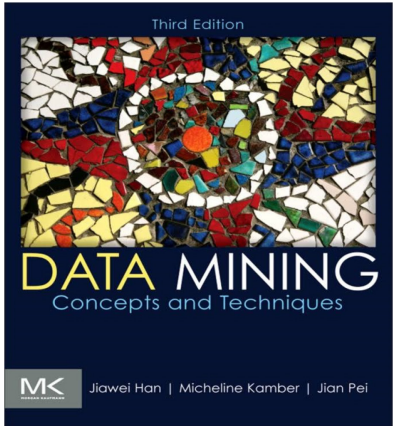
IDS-L18

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Outline of Today's Lecture

- Data quality aspects
- Data cleaning
- Data integration
- Data reduction
- Data transformation

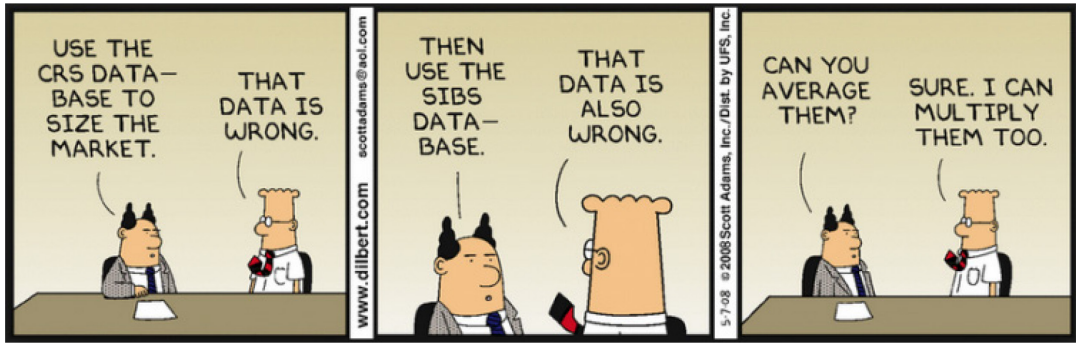


Based on chapter 3 of "data mining concepts and techniques" by Jiawei Han

Introduction


input
output
Avoiding GIGO

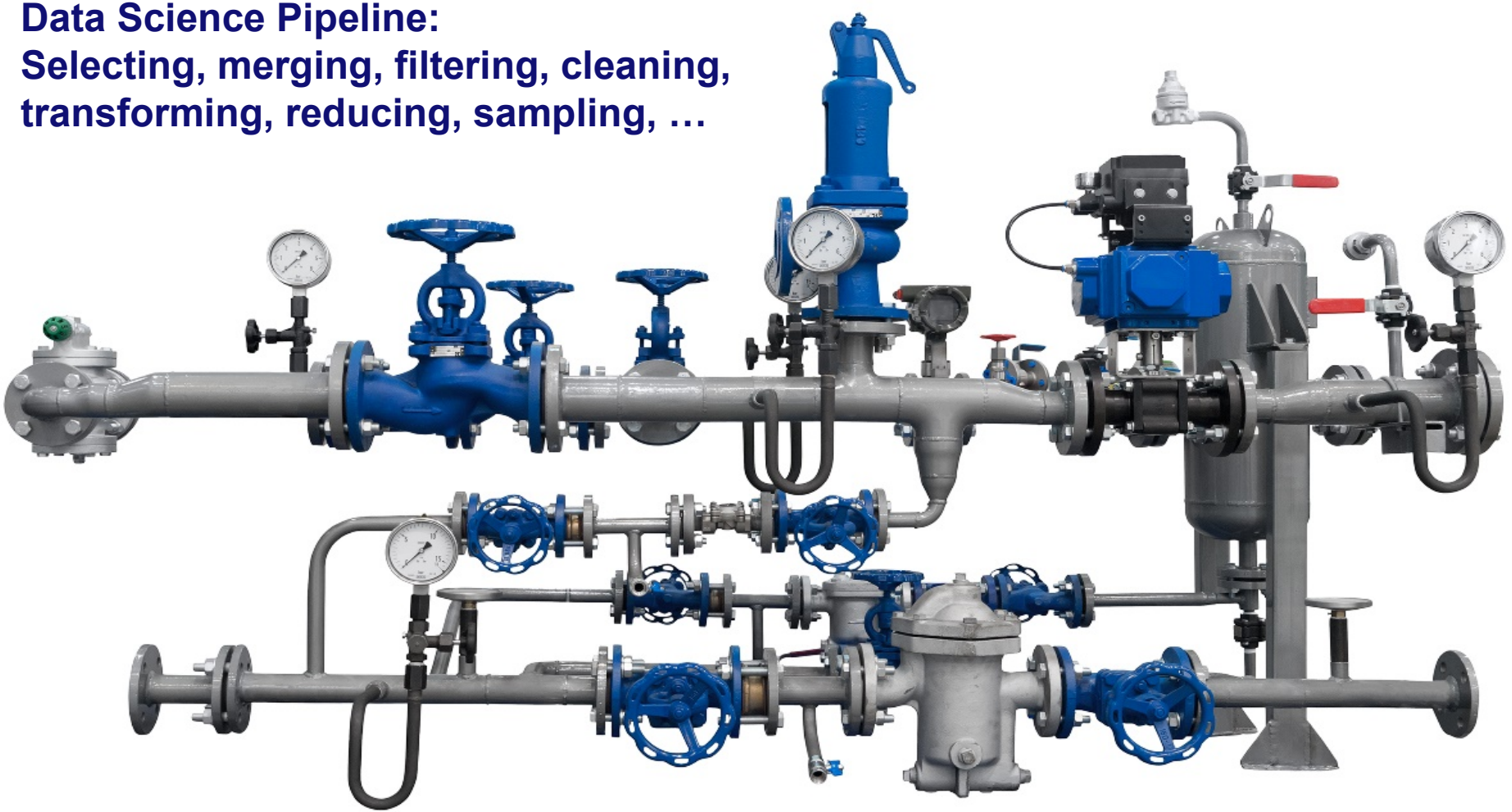
Introduction



- Real life data is rarely provided in a format ready for the analysis
- Data types, format and quality depend on the data sources
  - Not all formats are suitable for each analysis technique
  - Quality issues can alter the obtained results
- **Overall goal:** Increase data quality and modify the data to suit the analysis question and applied techniques

Chair of Process and Data Science

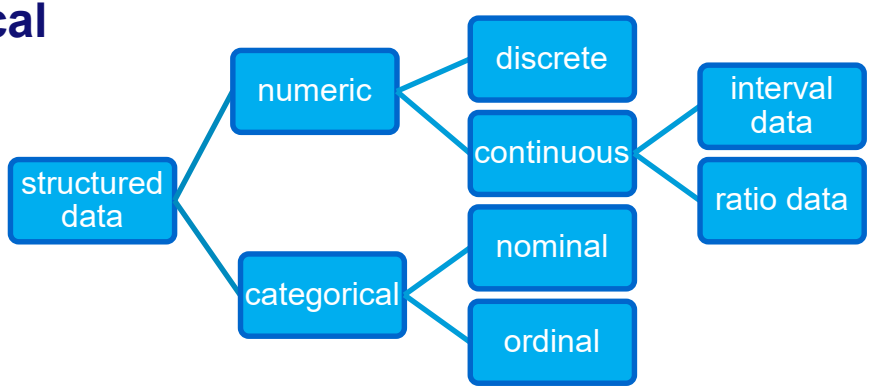
Data Science Pipeline:
Selecting, merging, filtering, cleaning, transforming, reducing, sampling, …



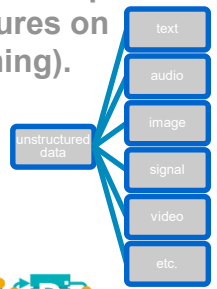
Data types - Overview

The attributes of our data can have various types – we need to know them to be able to analyze the data correctly

- **Nominal/Categorical**
- **Binary**
- **Ordinal**
- **Numeric**
  - **Interval-scaled**
  - **Ratio-scaled**



Unstructured data is turned into structured data (see for example the lectures on text mining).



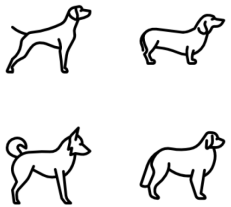
Chair of Process and Data Science

Data types – Nominal/Categorical

The attributes of our data can have various types – we need to know them to be able to analyze the data correctly

- **Values of nominal attributes represent some kind of category, code or state**
- **Ordering the values has no meaning (e.g. black hair is not better than brown hair)**
- **<u>Example</u>: Attribute “hair color” with values “blonde”, “brown”, “black” etc.**

Gender



Dog breed

Hair color

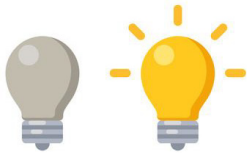
Chair of Process and Data Science

Data types - Binary

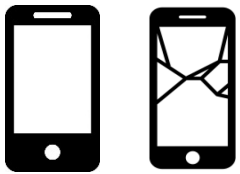
The attributes of our data can have various types – we need to know them to be able to analyze the data correctly

- **Nominal attribute with only two categories (often 1 and 0)**
- **Attribute is *symmetric* if both values are "equal" (subjective or freq. based)**
- **Attribute is *asymmetric* if one value represents the normal/default case and the other value an exceptional one**
- **<u>Example</u>: Outcome of a blood test which can either be "positive" or "negative"**

Gender



Status of the light



Working or broken screen

Chair of Process and Data Science

Data types - Ordinal

The attributes of our data can have various types – we need to know them to be able to analyze the data correctly

- **The possible values of the attribute have a meaningful order**
- **The difference between successive values can't be quantified**
- **<u>Example:</u> Customer satisfaction measured "very satisfied", "satisfied", "neutral" etc.**

How satisfied were you with our service?

Please rate your visit at our restaurant.

Chair of Process and Data Science

Data types - Numeric

- Attributes are measurable quantities
- Mean, median, mode values of these attributes exist
- Interval-scaled
  - Measured on a scale of equal-sized units that allows to measure the difference between values
  - The scale includes zero (A temperature of 0°C or 0°F is realistic)
  - Example: Temperature (in °C or °F), Coordinates, etc.
- Ratio-scaled
  - It is possible to identify multiples/ratios between values
  - The scale ends at zero (0kg, 0km/h or 0°K)
  - Example: Weight, speed, distance or monetary values

Temperature

Weight

Chair of Process and Data Science

Data Quality Aspects

Data Quality Aspects

- There is a set of commonly agreed quality aspects
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Interpretability
- Should be considered when setting up databases etc.
- Can be used to gain overview of quality of provided data

Example:
We want to analyze a company's sales database which has the attributes "item", "price" and "units sold".

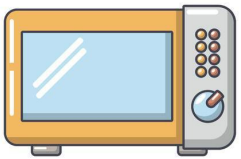
Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer	289,99	22
Microwave	12.095	56
Water cooker	30,50	45
…	…	…

Data Quality Aspects

Example: We want to analyze a company's sales data base which has the attributes "item", "price" and "units sold". Which quality aspects do we have to consider?

- **Accuracy**
  - **Are the values of our attributes correct?**
  - **Is it possible to identify errors in the data?**

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer	289,99	22
Microwave	**12,095**	**56**
Water cooker	30,50	45
…	…	…



The price of the microwave seems very low. Was it wrongly entered as 12,095€ instead of 120,95€?

Chair of Process and Data Science

Data Quality Aspects

Example: We want to analyze a company's sales data base which has the attributes "item", "price" and "units sold". Which quality aspects do we have to consider?

- **Completeness**
  - **Do we have missing values?**
  - **"Disguised missing data" occurs when users enter the default value presented to them**
    - **E.g. pre-selected dates for entering your birthday in an online form**
    - **E.g. default names: John Doe (US) and Max Mustermann (D)**

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer	289,99	22
Microwave	12.095	56
Water cooker	30,50	45
Smart fridge		**0**

The smart fridge is not in stock yet and its price is still unknown.

Data Quality Aspects

Example: We want to analyze a company's sales data base which has the attributes "item", "price" and "units sold". Which quality aspects do we have to consider?

- **Consistency**
  - **Does the data follow naming conventions, common formats etc.**
  - **Are these naming conventions and formats used in the same way throughout the entire data?**
  - **E.g. different date formats**

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer	**289,99**	22
Microwave	12.095	56
Water cooker	**30,50**	45
…	…	…

Inconsistent usage of "." and "," as decimal separator.

13/12/2018
Thursday, 13 December 2018
13 Dec 2018
13 December 2018
12/13/18
13-Dec-18

Data Quality Aspects

Example: We want to analyze a company's sales data base which has the attributes "item", "price" and "units sold". Which quality aspects do we have to consider?

- **Timeliness**
  - **For a period of time during the collection process the data might be incomplete (aging, lost updates, etc.)**
  - **Once it is all received, it is correct and complete**

Item	Price	Units sold
Vacuum cleaner	52.85	**34**
Dish washer	289,99	**22**
Microwave	12.095	**56**
Water cooker	30,50	**45**
…	…	…

Sales employees enter their sales into the system in the last 5 days of the month. During these days the data might be incomplete.

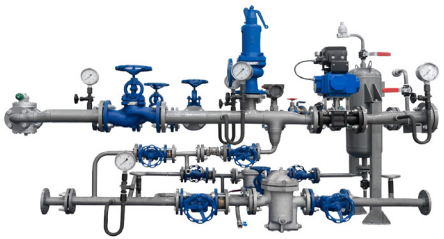
Data Quality Aspects

Example: We want to analyze a company's sales data base which has the attributes "item", "price" and "units sold". Which quality aspects do we have to consider?

- **Believability**
  - **How much does the user trust the data?**
  - **User should believe the data is true, real and credible**
  - **Depends on the source of the data and its processing history**

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer	289,99	22
Microwave	12.095	56
Water cooker	30,50	45
…	…	…

Previous errors in the sales numbers have decreased the trust in the system



Data Quality Aspects

Example: We want to analyze a company's sales data base which has the attributes "item", "price" and "units sold". Which quality aspects do we have to consider?

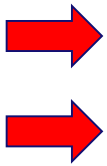
- **Interpretability**
  - **Is the data understandable without much explanation?**
  - **Does it leave for ambiguousness?**

Item	Price	Units sold
300F2	52.85	34
601X2	289,99	22
505D8	12.095	56
210R0	30,50	45
…	…	…

Items are not identified by their name but some code "Item 300F2". Without having the explanations of the codes it is difficult to understand the sales records.

Data Cleaning

Data Cleaning - Revisited

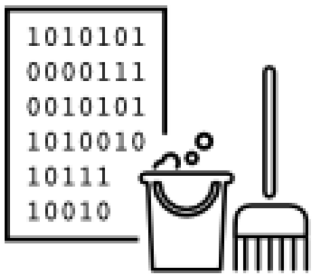


Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2	*Decision trees and data visualization/explora*	*19/10/2018*	*Friday*

- **In Lecture 3 and Instruction 2 we already learned about data cleaning techniques**
  - **Handling missing values**
  - **Handling impossible values**
  - **Outlier detection**
- **Today, we will go into some more detail**

Data Cleaning

- Data cleaning aims to improve these quality aspects by
  - Filling in missing values
  - Smoothing noise
  - Identifying outliers
  - Correcting inconsistencies in the data etc.
- We focus in this lecture on:
  - Missing values
  - Noise



How do we handle missing values

- There are multiple ways to treat missing values in the data set
  - Ignore the tuple
  - Fill in the value manually
  - Fill with overall mean/median
  - Fill with mean/median of all samples belonging to the same class
  - Fill with the most probable value
- The appropriate method has to be chosen based on the data set

Missing values – Ignore the Tuple

- The entire data entry is discarded
- Usually done when data entry becomes unusable (e.g. labeling attribute for decision tree classification is missing)
- If the data set is missing a lot of values, this technique might make your data set unusable

Back to our previous example:
We want to clean our data regarding the missing values

Item	Price	Units sold
Vacuum cleaner	52.85	34
~~Dish washer~~		~~22~~
Microwave	120.95	56
~~Water cooker~~	~~30.50~~	
…	…	…

Missing values – Fill in the value manually

- **Assumes that due to domain knowledge we know the missing values**
- **Time consuming if there are a lot of missing values**

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer		**22**
Microwave	120.95	56
Water cooker	30.50	
…	…	…

Missing values – Fill with overall mean/median

- Compute mean, median etc. and fill gaps with it
- E.g. missing yearly income in customer data base and mean income is 56,000€
- If the values are spread over a wide range, the method may introduce values very far away from the real value.

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer		22
Microwave	120.95	56
Water cooker	30.50	37.33
…	…	…

Mean of the sold units:
(34 + 22 + 56) / 3 = 37.33

Missing values – Fill with mean/median of all samples belonging to the same class

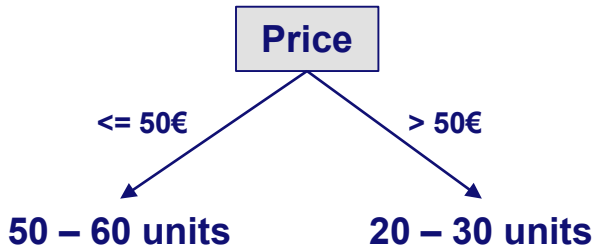
- Higher chances of being accurate than overall mean/median
- E.g. we know the customer is a 20 year old student living in Aachen West and compute the mean yearly income for this demographic group
- Only valuable if we can form meaningful groups in our data

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer	289.99	22
Microwave	120.95	56
Water cooker	30.50	
Fridge	449.95	65
Oven	529.00	**60.5**
…	…	…

To estimate the sold units of the oven we calculate the mean units sold of the dish washer and fridge (category: big kitchen devices):
(56+65) / 2 = 60.5

Missing values – Fill with most probable value

- **Based on a more complex prediction technique (Decision tree induction, Regression etc.)**



Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer		22
Microwave	120.95	56
Water cooker	30.50	20 - 30
…	…	…

Simple decision tree predicting the number of sold units based on the item price.

Regression is suited for this:

$$\text{RENTAL PRICE} = -0.1513 + 0.6270 \times \text{SIZE} - 0.1781 \times \text{FLOOR} + 0.0714 \times \text{BROADBAND RATE}$$

How do we handle missing values

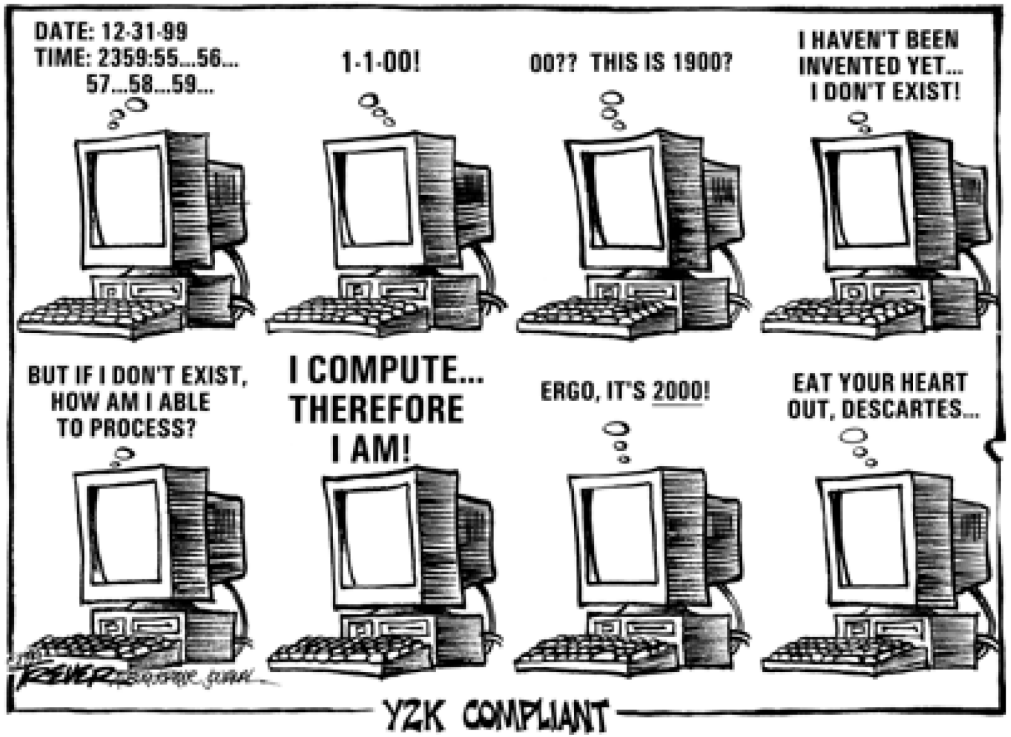
- **There are multiple ways to treat missing values in th**
  - **Ignore the tuple**
  - **Fill in the value manually**
  - **Fill with overall m**
  - **Fill wi**

Are all missing values "bad"? What about intentionally left blanks?
E.g. Nowadays customers might not have landline phone numbers anymore and leave it blank

→ Good systems should handle these values as "NA" etc. but we are working with real-life data!

the data set

IT is used in ways not envisioned



6 digit date format to save storage space

Chair of Process and Data Science

IT is used in ways not envisioned

trimester	2

trimester

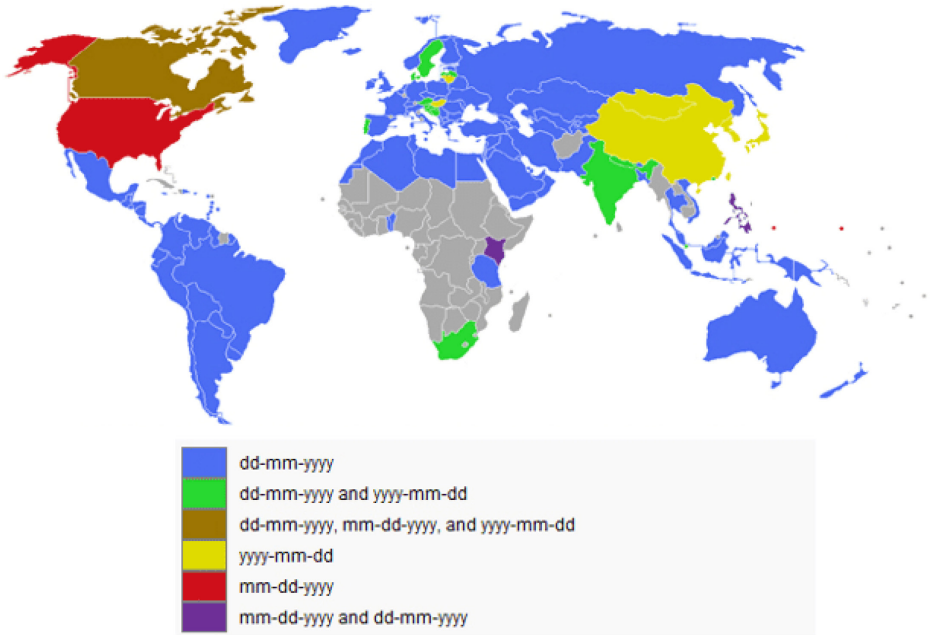
semester

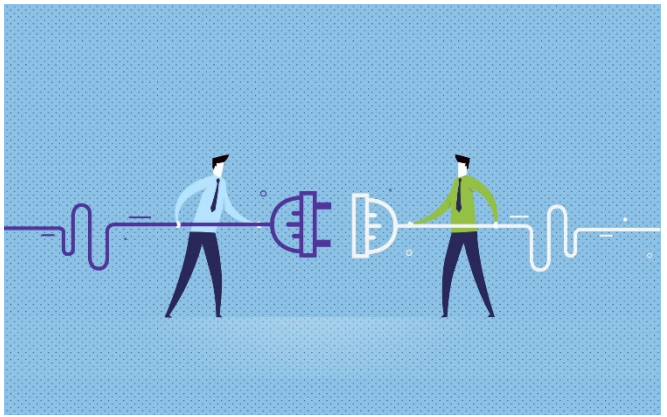
quarters

TU/e

Chair of Process and Data Science

IT is used in ways not envisioned



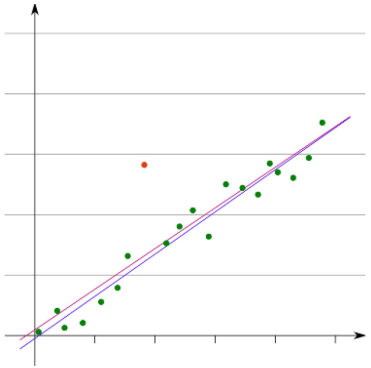


IT systems after a merger

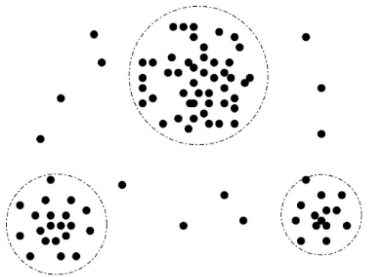
Chair of Process and Data Science

How do we handle noise? (1/2) – Detect and Remove

- Noise: Random error or variance in a variable
- One possibility is to detect and remove outliers
- Possible outlier detection techniques:
  - Techniques already discussed in previous lectures
  - Using boxplots, histograms (Lecture 3)
  - Linear regression (Lecture 5): Classify data points over a certain error threshold as outliers
  - Clustering (Lecture 8): Values that fall outside the clusters could be considered outliers
- After identification, simply remove the detected outliers



Outlier detection using linear regression



Outlier detection using clustering

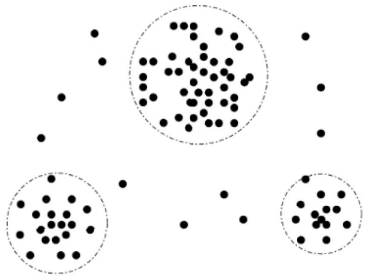
Chair of Process and Data Science

How do we handle noise? (1/2) – Detect and Remove

- One possibility is to detect and remove outliers
- Possible outlier detection techniques:
  - Techniques already discussed in previous lectures
  - Using boxplots, histograms (Lecture 3)
  - Linear regression (Lecture 5): C…
    over a certain error t…
  - Clusteri… …ll outside the
    …ed outliers
- …simply remove the detected outliers

What if we don't want to remove data?
What if we are not confident about our outlier detection?

Outlier detection using linear regression



Outlier detection using clustering

Chair of Process and Data Science

What is noise? Not so easy!

	correctly recorded	incorrectly recorded
frequent / many neighbors		
infrequent / few neighbors		

How do we handle noise? (2/2) - Binning

- **There might be scenarios when we don’t want to actually remove data**
- **We can use binning to smooth the data values**
- **Binning smooths data values by “consulting its neighborhood” (the values around it)**

Example data set:
4, 28, 8, 34, 15,
21, 24, 21, 25

How do we handle noise? (2/2) - Binning

- **Step 1: Sort data**

Example data set:
4, 28, 8, 34, 15,
21, 24, 21, 25

Sorted data set:
4, 8, 15, 21, 21, 24,
25, 28, 34

How do we handle noise? - Binning

- **Step 1: Sort data**
- **Step 2: Put data into a number of bins**
  - **Equal-width: Each bin has the same width**
  - **Equal-frequency: Each bin contains the same number of values**

Example data set:
4, 28, 8, 34, 15,
21, 24, 21, 25

Sorted data set:
4, 8, 15, 21, 21, 24, 25,
28, 34

3 Equal-width bins (interval width of 10):
Bin 1: 4, 8
Bin 2: 15, 21, 21, 24
Bin 3: 25, 28, 34

3 Equal-frequency bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

How do we handle noise? - Binning

- **Step 1: Sort data**
- **Step 2: Put data into a number of bins**
  - **Equal-width: Each bin has the same size**
  - **Equal-frequency: Each bin contains the same number of values**
- **Step 3: Smooth**
  - **By bin mean: Each value is replaced by the bin mean**
  - **By bin median: Each value is replaced by the bin median**
  - **By bin boundaries: Minimum and maximum value in each bin are identified and each value is replaced by closest boundary**

Example data set:
4, 28, 8, 34, 15,
21, 24, 21, 25

Sorted data set:
4, 8, 15, 21, 21, 24, 25,
28, 34

3 Equal-frequency bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smooth by bin mean:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smooth by boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

How do we handle noise? - Binning

- **The effectiveness of the method is questionable**
  - **Is it necessary to repair minor noise inside a bin?**
  - **Major noise will not be in the correct bin**
- **There are multiple other (much better) smoothing techniques:**
  - **Smoothing by cluster: use properties of the cluster to smooth the data**
  - **Smoothing by regression: fit the data to a regression function**
  - **Etc.**

Example data set:
4, 28, 8, 34, 15,
21, 24, 21, 25

Sorted data set:
4, 8, 15, 21, 21, 24, 25,
28, 34

3 Equal-frequency bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smooth by bin mean:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smooth by boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data Integration

Data Integration

- **Our problem: The data we need is collected by several systems and stored in separately**
- **Solution sounds simple: Obtain the data from each system and merge it into one data set**
- **This simple task can introduce a lot of issues:**
  - **Redundancies**
  - **Inconsistencies**
  - **Heterogenic semantics**
  - **Structures of data**

Data Integration Challenges – Tuple Duplication

- There are multiple data entries referring to the same case
- Often caused by inconsistency in the data maintenance
- E.g. the client's address was not correctly updated and there are now two entries for the same purchase but with different addresses

Item	Price	Units sold
Vacuum cleaner	52.85	34
Dish washer	289.99	22
Microwave	**120.95**	**56**
Microwave	**100.95**	**56**
…	…	…

Do the entries refer to two different Microwaves?
Did the price change for one product and the system duplicated the entry instead of updating it?

Data Integration Challenges – Value Conflicts

- **Data Value Conflicts**
  - Differences because of *representation, scaling or encoding*
    - E.g. Product prices depend on currency and might be represented with or without VAT
  - Differences because of *abstraction level*
    - E.g. In one database total sales might refer to company wide sales while another database may refer to one specific region

Item	Price	Units sold
…	…	…
Microwave	120.95	56
…	…	…

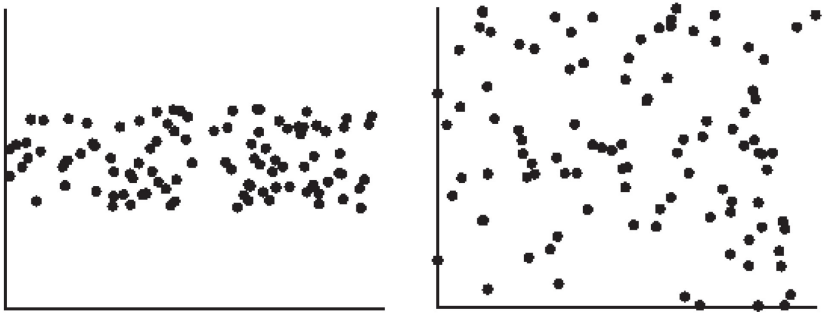
Item	Price	Units sold
…	…	…
Microwave	120.95	130
…	…	…

One database collects sales for a special region, while the other collects them world wide.
Be careful to merge the relevant one to your data set!

Data Integration – Challenges

- **Redundancy**
  - Attributes might be redundant if they can be derived from other attributes
  - This takes up storage space but can also affect the results of our analysis
  - Some redundancies can be detected by correlation analysis
    - Quantifies the correlation of attributes
    - Beware that correlation does not automatically imply causality
    - E.g. chi-square test for nominal data or correlation coefficient for numeric data

Scatter plots showing a positive and a negative correlation between attributes



Scatter plots showing no correlation between the two attributes

Chair of Process and Data Science

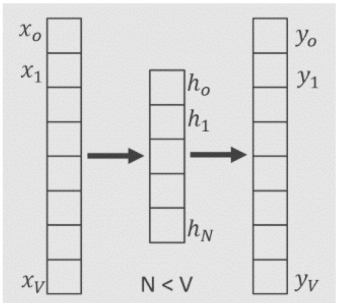
Data Reduction

Data Reduction

- Depending on our data sources the data set might grow big with a lot of attributes
- Problem: Running our analysis might not be efficient more
- Goal: Reduce the data set to be smaller but produce the same analysis results
- Two types of reduction:
  - Dimensionality reduction
  - Numerosity reduction
- Time spend on data reduction should not grow longer than the potential time saved during the analysis because of it

Data Reduction – Dimensionality Reduction

- Reduce the number of attributes used for the analysis
- Achieved by applying a certain encoding mechanism to reduce the data set size
- The techniques can be split in two main classes
  - Reduce dimensionality by *projecting data on fewer dimensions*
    - Autoencoders (see text mining)
    - Principal components analysis
  - Reduce dimensionality by *detecting and removing irrelevant/redundant attributes*
    - Attribute subset selection

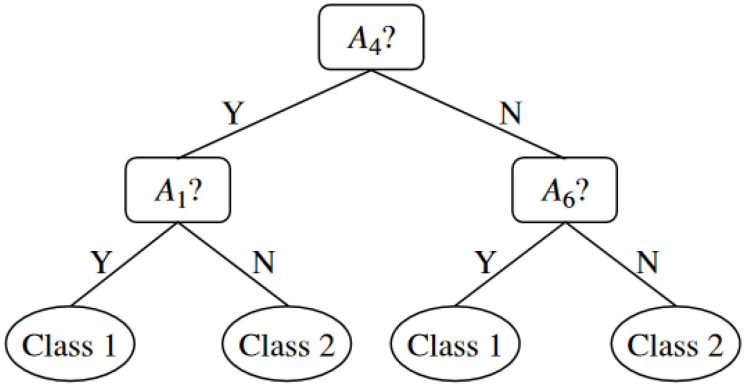


Chair of Process and Data Science

Data Dimensionality Reduction – Attribute Subset selection

- Removes irrelevant or redundant attributes
- Makes analysis more effective and results might be easier to understand
- E.g. decision tree induction which selects best attributes to describe the different classes of the dataset

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



The decision tree shows that attributes A1, A4 and A6 are relevant to describe the label of the data set.
Note: Each of these attributes might have a set of correlated attributes which are not displayed by the tree since they classify the data with the same quality as A1/A4/A6.

Data Reduction – Numerosity Reduction

- Reduce the data by replacing or estimating the original data with alternative, smaller representations
- There are two main classes of techniques
  - Parametric methods: a *model* estimates the data and *only the data parameters are stored*
    - Regression (Lecture 5)
  - Non-parametric methods: store reduced representations of the data
    - Histogram (Lecture 3)
    - Sampling data (Lecture 3)
    - Clusters (Lecture 10)
    - Data Cubes

Data Numerosity Reduction – Data Cube Aggregation

- **Main principle: Detailed data can be aggregated to a less detailed level**
- **The level of detail should fit the analysis**
- **Possible aggregation functions:**
  - **Summation**
  - **Counting elements**
  - **Obtaining the minimum/ maximum**
  - **Obtaining the total**

The company saves the total sales numbers per product per year, but we are only interested in the yearly total.

2016
2017
2018

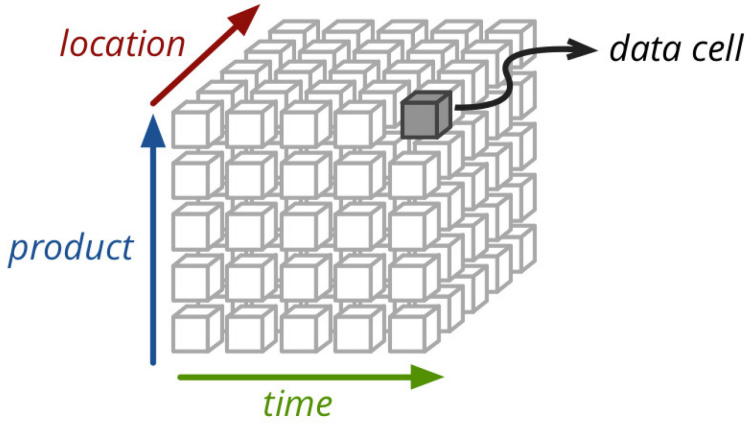
Item	Price	Units sold
Vacuum cleaner	52.85	3857
Dish washer	289.99	2568
Microwave	120.95	5698
…	…	…

Aggregation

Year	Units sold
2016	11557
2017	13584
2018	12587

Data Numerosity Reduction – Data Cube Aggregation

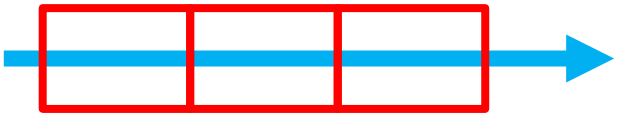
- The data cube stores multidimensional aggregated data
- In the cube each cell holds an aggregated data value
- Provides fast access to precomputed data
- Data can be stored hierarchically, which allows for different levels of detail
  - E.g. a hierarchy for location allows to group data into regions



Example: Predict weather

rain
temperature
humidity
wind
…

1km², 10km², 100km²



minute, hour, day, week

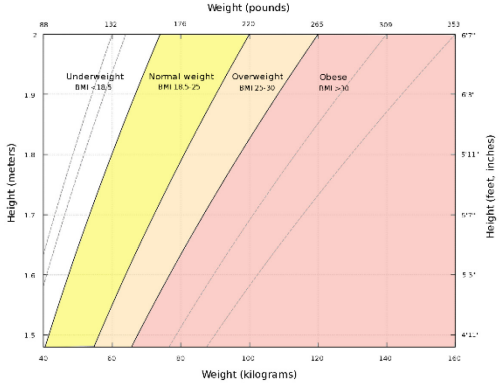
Data Transformation

Data Transformation

- Goal: Manipulate data such that
  - The noise is removed
  - The analysis is more efficient
  - The results are easier to understand
- Multiple transformation possibilities
  - Smoothing (see noise removal)
  - Attribute construction
  - Aggregation (see data reduction)
  - Normalization
  - Discretization (see binning in Lecture 3)
  - Concept hierarchy generation (see data reduction)
- Data transformation is a broad topic and therefore has a big overlap with topics discussed earlier during this lecture

Data Transformation – Normalization

- ***Problem:*** **Analysis results might depend on the chosen measurement unit**
  - **E.g. Meters vs. inches, kilogram vs. pounds**
- **Attributes with smaller measurement units might have a bigger range of values**
- **Algorithms tend to give more weight to attributes with a big range**
- **Could result in an unfair distribution of weights between the attributes**



dB	Power ratio	Amplitude ratio
100	10 000 000 000	100 000
90	1 000 000 000	31 623
80	100 000 000	10 000
70	10 000 000	3 162
60	1 000 000	1 000
50	100 000	316.2
40	10 000	100
30	1 000	31.62
20	100	10
10	10	3.162
6	3.981 ≈ 4	1.995 ≈ 2
3	1.995 ≈ 2	1.413 ≈ $\sqrt{2}$
1	1.259	1.122
0	1	1
−1	0.794	0.891
−3	0.501 ≈ ½	0.708 ≈ $\sqrt{½}$
−6	0.251 ≈ ¼	0.501 ≈ ½
−10	0.1	0.316 2
−20	0.01	0.1
−30	0.001	0.031 62
−40	0.000 1	0.01
−50	0.000 01	0.003 162
−60	0.000 001	0.001
−70	0.000 000 1	0.000 316 2
−80	0.000 000 01	0.000 1
−90	0.000 000 001	0.000 031 62
−100	0.000 000 000 1	0.000 01

Data Transformation – Normalization

- ***Solution:*** **Scale all data to fit within a smaller interval**
  - **E.g.** -1.0 to 1.0 **or** 0.0 to 1.0
- **Will assign equal weight to all attributes**
- **Also useful to speed up algorithms such as training a neural network**
- **Many normalization methods exist such as**
  - **Min - max normalization**
  - **Standard score normalization (z-score/ zero-mean normalization)**
  - **Decimal scaling**

Data Transformation – Min-max Normalization

Recall from Lecture 3

- **Min-max normalization maps values of an attribute onto the range** $[low, high]$

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low$$

- **Preserves relationship among the original data values**

Data Transformation – Z-score Normalization

- **Standard score uses the standard deviation to normalize.**

$$a_i' = \frac{a_i - \bar{a}}{sd(a)} \qquad [-\infty, \infty]$$

Recall from Lecture 3

- **Useful when the actual minimum or maximum of attribute A are unknown**
- **Useful when outliers might influence the min-max normalization**

Data Transformation – Decimal scaling

- Normalizes by moving the decimal point of the values of attribute A
- The number of decimal points moved depends on the maximum absolute of A

$$a'_i = \frac{a_i}{10^{\mathrm{j}}}$$

With j as smallest integer such that $\max(|a'_i|)<1$

Conclusion

Short summary of lecture

- Already before the analysis we face many challenges
- Careful data preprocessing can avoid errors and increase quality of our results
- Main challenges:
  - Merge our data from multiple sources
  - Detect and remove outliers and noise
  - Deliver data in a format that fits analysis question and technique

Relevant Literature

- Chapter 3 of Data Mining: Concepts and Techniques by Han J. et al (2012), 3rd Edition
- Chapter 3 of Fundamentals of Machine Learning for Predictive Data Analytics by J. Kelleher, B. Mac Namee and A. D'Arcy.
- Chapter 2 of Analytics in a Big Data World: The Essential Guide to Data Science and its Applications by B. Baesens

#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2			
Instruction 3			
Instruction 4			
Instruction 5			
Instruction 6			
Lecture 14	Process mining (unsupervised)	28/11/2018	Wednesday
Lecture 15	Process mining (supervised)	29/11/2018	Thursday
Instruction 7	*Process mining and sequence mining*	*30/11/2018*	*Friday*
Lecture 16	Text mining (1/2)	05/12/2018	Wednesday
Instruction 8	*Text mining and process mining*	*06/12/2018*	*Thursday !!*
Lecture 17	Text mining (2/2)	12/12/2018	Wednesday
Lecture 18	Data preprocessing, data quality, binning, etc.	13/12/2018	Thursday
Lecture 19	Visual analytics & information visualization	19/12/2018	Wednesday
backup		20/12/2018	Thursday
Instruction 9	*Text mining, preprocessing and visualization*	*21/12/2018*	*Friday*
Lecture 20	Responsible data science (1/2)	09/01/2019	Wednesday
Lecture 21	Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	*Responsible data science*	*11/01/2019*	*Friday*
Lecture 22	Big data (1/2)	16/01/2019	Wednesday
Lecture 23	Big data (2/2)	17/01/2019	Thursday
Instruction 11	*Big data*	*18/01/2019*	*Friday*
Lecture 24	Closing	23/01/2019	Wednesday
backup		24/01/2019	Thursday
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*
backup		30/01/2019	Wednesday
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*

- **Register for the exam!**
- **Exchange students** (see post).
- **Assignments.**