4.4 How and why to harmonize, rescale and cleansing data

Why is data cleaning needed?

1) Pre-existing databases (secondary data)

- Data is collected for other purposes
- Quality of data is driven by what was originally important
- So far no need to satisfy the quality requirements for the new analysis

2) Data collected via surveys (primary data)

- Individuals asked to fill out an extensive survey will get fatigued and simply put default values, may check the neutral response or indicate that they are satisfied with everything or satisfied with nothing
- Individuals my be offended by questions about their age, income,… and will often deliberately fill in a false answer
- Most people, when asked to fill in a survey, simply refuse.

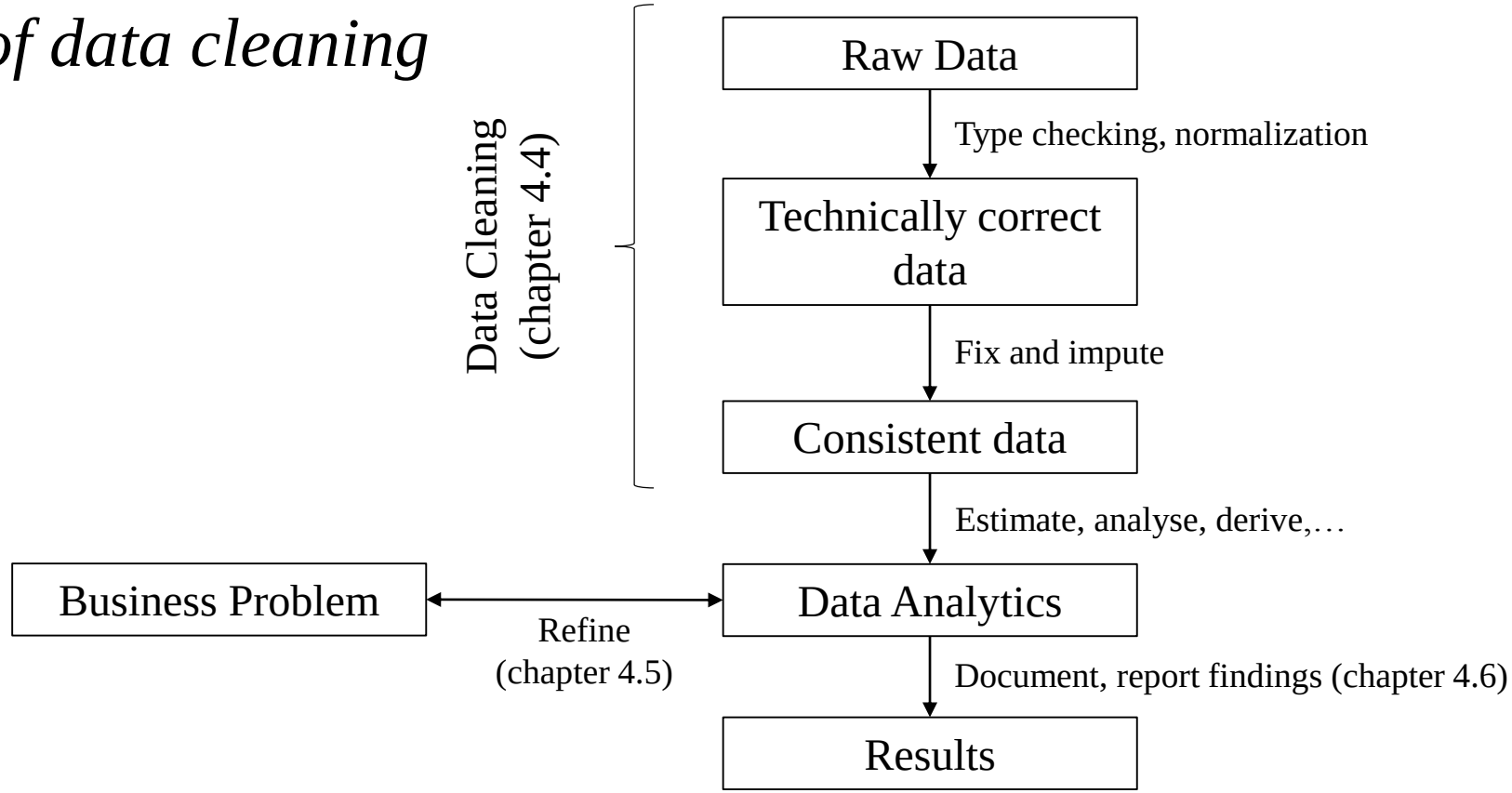
4.4 How and why to harmonize, rescale and cleansing data

Example

- Vendors have to fill in various forms to get reimbursed for their services. As a result, the decision-relevant fields get scrutinized carefully and the rest do not.

4.4 How and why to harmonize, rescale and cleansing data

Process of data cleaning



4.4 How and why to harmonize, rescale and cleansing data

The difference of *technically correct data* to *consistent data*

1) Technically correct data

- Each value can be directly recognized as belonging to a certain variable
- Each value is stored in a data type that represents the value domain of the real-world variable (e.g. a text variable should be stored as text, a numeric value as number)

2) Consistent data

- Missing values, special values, (obvious) errors and outliers are either removed, corrected or imputed.
- Data is consistent with constraints based on real-world knowledge about the subject that the data describes

⇨ Data that is “fit” for analysis

4.4 How and why to harmonize, rescale and cleansing data

Data cleaning includes identifying

Assess technical correctness

- the range of valid responses
- invalid data responses (e.g. letters in number fields)

Assess consistency

- inconsistent data encodings (e.g. different abbreviations for countries)
- suspicious data responses (e.g. questionable responses)
  ⇨ ask: do the outliers make sense?
- suspicious distribution of values (e.g. 99% of respondents live in poor neighborhoods but have income higher than 1 mCHF)
  ⇨ use descriptive statistics to identify (e.g. histograms, Box plot, stem and leaf plots, scatter plots,…)
- suspicious interrelationships between fields
  ⇨ identify correlations (e.g. factor analysis, principle component analysis)

Key question: DOES THE DATA MAKE SENSE?

4.4 How and why to harmonize, rescale and cleansing data

Handling null or missing values

1) **Deletion**: Drop observation containing missing value
2) **Deletion when necessary**: Only drop observation, when missing value is needed for analysis (e.g. for one kind of analysis 1’000 people, for the other 950 people)
3) **Imputing a value**: use of regression to predict an answer
4) **Randomly imputing a value**: Imputing a value as in 3) might understate the uncertainty about the value. Therefore rerun analysis for all possible outcomes and weight by regression-based probability.

4.4 How and why to harmonize, rescale and cleansing data

Problems faced when combining data from different sources

1) **Level of granularity** is different (e.g. in one data source location is given in another data source not)
   ⇨ *Solution*: if location is not needed for analysis, skip this variable. If location is needed, aggregate data and treat records without location as if they have missing values (see slide before)
2) Different **data architecture** (e.g. missing value can be represented by spaces, NA, na, Not/Available)
   ⇨ *Solution*: harmonize before aggregating the data

4.4 How and why to harmonize, rescale and cleansing data

Problems faced when combining data from different sources

3) Data is **stored in different ways** (e.g. some information on vehicles is stored in one database being called “two door Chevrolet”, other information is stored in database called “Chevy Cruzes”. Problem: no single categorization for both databases.)
   ⇨ *Solution*: define a record which has enough fields to contain the information of each of the databases.
4) **Different weights** of observations (e.g. one observation might reflect the responses of 10’000 people, the other of 100 people)
   ⇨ *Solution*: introduce a weighting field

4.4 How and why to harmonize, rescale and cleansing data

Helpful tips and tricks for data cleaning

1) Create a field with the date of each observation (“date stamp”)
2) Create a field identifying the data source from which the data is collected
3) Model might require information which is not in database but can be computed.
   ⇨ *Solution*: create a new field with computed variable
4) Certain fields have the same value across all datasets
   ⇨ *Solution*: it might be worth deleting those fields
5) Store input data for each stage (raw, technically correct, consistent, aggregated and formatted)

4.4 How and why to harmonize, rescale and cleansing data

Frameworks for data and information quality:

- the 10 C’s (data science)
- ACCURATE (accounting)

4.4 How and why to harmonize, rescale and cleansing data

Final check on data quality: ***the 10 C's***

Completeness: are all the fields of the data complete?
Correctness: is the data accurate?
Consistency: is the data consistent with the definition of that field and concept?
Currency: is the data obsolete?
Collaborative: is the data based on one opinion or on a consensus of experts?
Confidential: is the data secure from unauthorized use?
Clarity: is the data legible and comprehensible?
Common Format: is the data in a format easily used?
Convenient: can the data be conveniently and quickly accessed?
Cost-effective: is the cost of collecting and using the data commensurate with its value?

4.4 How and why to harmonize, rescale and cleansing data

Final check on data quality: ***ACCURATE***

Accurate: is the information fair and free from bias? Are there any arithmetical or grammatical errors?

Complete: are there any facts or figures missing or concealed?

Cost-beneficial: can the money spent on information be recovered?

User-targeted: does the style, format, detail and complexity of the information address the needs of the users of the information?

Relevant: is the information communicated to the right person?

Authoritative: does the information come from a reliable source?

Timely: does the receiver of the information have enough time to decide appropriate actions based on the information received?

Easy to use: is the information understandable to the users?

4.4 How and why to harmonize, rescale and cleansing data

Case Study: Claim data of health insurance company

Starting position:

- client provided us with dataset ① and dataset ② of claims data
- “;”-separated csv files
- 13m rows for both datasets; 250 columns in dataset ①, 150 columns in dataset ②
- Analysis planned to perform on dataset ①

4.4 How and why to harmonize, rescale and cleansing data

Case Study: Claim data of health insurance company

B1019004_TARIFFTEXT	B1019004_REQTARIF	B1019004_	B1019004_E
Übergangszuschlag, pro Analyse	317	4708.00	
TEMESTA 1.0 EXPIDET Tabl 1 mg 50 Stk	400	1233539	
TRAMAL RETARD Tabl 50 mg 50 Stk	400	3166333	
Zuschlag für jede Analyse, die kein Suffix C aufweist	nur anwendbar in V	317	4707.20
UTROGESTAN Kaps 100 mg 3 x 30 Stk	400	5497382	
Instruktion von Selbstmessungen, Selbstbehandlung	1	00.0610	
Conseiis en nutrition 2éme ä 6éme séance	999	7812	
Röntgen: Thorax und/oder Rippen, inkl. Sternum, ers	1	39.0190	
Rundungsdifferenzen	405	5999.99	

DR
B1019004_QUANTITY
-1
-1
-1
-1
-1
-1
1
10

4.4 How and why to harmonize, rescale and cleansing data

Case Study: Claim data of health insurance company

We started with some data checks and noticed the following (extract):

- **Deletion of fields:** Many columns with all empty entries ⇨ delete columns as this leads to increased performance
- **Currency**: one field was indicating the treatment reason (accident, motherhood,..). In 13m rows only 8 treatments for motherhood were listed ⇨ question: are only 8 entries reasonable? ⇨ client conversation showed, that the entries of this field are not reliable ⇨ no use of this field for analysis
- **Suspicious data responses**: Quantity of claim was negative for some rows ⇨ question: does this make sense? ⇨ after conversation with client the answer was yes (discounts and rounding was booked as negative quantity)

4.4 How and why to harmonize, rescale and cleansing data

Case Study: Claim data of health insurance company

- **Suspicious data responses**: scatter plot showed, that several of the patient had age above 100 ⇨ analysis and client conversations showed, that these values were true
- **Consistency**: Some data rows were shifted by one or more columns ⇨ analysis showed, that a text field could contained “;” which was also used as separator ⇨ correction of shift
- **Data is stored in different ways**: Birth date, address needed for the analysis was only contained in dataset ② ⇨ define new records which contain birth date and address and join dataset ① and ②
- …. and more and more and more questions were asked to understand the data (different fields for date, for amount, …)

4.5 Discovery of relationship in data

To understand the data and its relationships, 9 steps can be performed

Source: Booz-Allen-Hamilton: The Field Guide to Data Science (including detailed table of methods described on the following slides)

1) Filtering
2) Imputation
3) Dimensionality Reduction
4) Normalization & Transformation
5) Feature Extraction
6) Aggregation
7) Enrichment
8) Clustering
9) Regression

Processing
How do I clean and separate my data?

Describe
How do I develop an understanding of the content of my data?

Discover
What are the key relationships in the data?

4.5 Discovery of relationship in data

1) *Filtering: How do I identify data based on its absolute or relative values?*

- Involves using **relational algebra projection and selection** to add or remove data based on its value
- Involves **outlier removal**, **exponential smoothing** and the use of either **Gaussian or median filters**

4.5 Discovery of relationship in data

2) Imputation: How do I fill in missing values in my data?

- Values for missing data can be generated using **random sampling** or **Monte Carlo Markov Chain methods** (if other observations in the dataset can be used)
- Using **mean**, **regression models** or **statistical distributions** based on existing observations

4.5 Discovery of relationship in data

3) Dimensionality Reduction: How do I reduce the number of dimensions in my data

- **Principle component analysis** or **factor analysis** can help determine whether there is correlation across different dimensions in the data
- For unstructured text data: **term frequency / inverse document frequency** identifies the importance of a word in some document
- **Feature hashing** is an efficient method for creating a fixed number of features which form the indices of an array
- **Sensitivity Analysis** and **wrapper methods** when important features are not known.
- **Self-organizing maps** and **Bayes nets** are helpful in understanding the probability distribution of the data.

4.5 Discovery of relationship in data

4) Normalization & Transformation: How do I reconcile duplication representations in the data?

- Correct duplicate data elements with **de-duplication**
- **Normalization** ensures your data stays within a common range
- **Format conversion** is typically used when data is in binary format
- **Fast Fourier Transforms** and **Discrete wavelet transforms** are used for frequency data
- **Coordinate transformations** are used for geometric data defined over Euclidean

4.5 Discovery of relationship in data

5) Feature extraction

Extracting features out of data is the process of determining the set of features with the highest information value to the model.

Features can be filtered out and tested with statistics. Filtering out features is independent of the finally applied model.

With so-called wrapper methods (e.g. regression, k-nearest neighbour) on the other side, one evaluates feature sets by constructing models and measuring performance.

4.5 Discovery of relationship in data

6) Aggregation: How do I collect an summarize my data

- Basic statistics (**raw counts**, **means**, **medians**, **standard deviations**, **ranges**) are helpful in summarizing data.
- **Box plots** and **scatter plots** provide compact representation
- **“Baseball card” aggregation** is a way of summarizing all the information available on an entity

4.5 Discovery of relationship in data

7) Enrichment: How do I add new information to my data?

- **Annotation** for tracking source information and other user-defined parameters
- **Relational algebra rename and feature addition** (e.g. geography, weather) can be helpful in processing certain data fields together or use one field to compute the value of another

4.5 Discovery of relationship in data

8) Clustering: How do I segment the data to find natural groupings?

- **Connectivity-based** methods: hierarchical clustering
- **Centroid-based** methods: when the number of clusters is known: **k-means**. When the number is unknown: **x-means** or **Canopy clustering**
- **Distribution-Based** methods: **Gaussian mixture models**
- **Density-based** methods: **Fractal and DB scan** are useful for non-elliptical clusters
- **Graph-based** methods: useful when you only have knowledge of how one item is connected to another
- **Topic modeling:** for segmentation of text data

4.5 Discovery of relationship in data

9) Regression: How do I determine which variables are important?

- **Tree-based methods**, when structure of data is unknown
- **Generalized linear models**, when statistical measure of importance are needed.
- **Regression with shrinkage** (e.g. LASSO, elastic net) and **stepwise regression**, when statistical measure of importance is not needed.

4.6 Documentation and reporting of findings

Raw data and relationships will not hold the attention of your stakeholders

⇨ Tie your finding to the analytics problem and from there to the business problem.

⇨ State what the impact of your findings to the business are. This might be complemented with recommendations.

⇨ Document relationships which are higher than “normal”.

Remark: The data warehouses and data marts should also be documented in a way that makes them usable by external parties. There might be requests to revisit the data months or years after the analysis is done.

4.7 Re-definition of the business and analytics problem statement by use of the data analytics result

Solid data and relationships will allow you to find that:

- the true constraints of the system isn't what you thought it was and therefore the analytics problem needs to be **reframed**.
- the business problem itself missed a key facet (e.g. unexpected relationship, time-series effect,…) that needs to be **included**.

Remark: only data of good quality enables you to do a first true refinement of your analytics and business problem.

Excurse: European Statistics Code of Practice

***Vision** of the European Statistical System:*

“ The European Statistical System will be a world leader in statistical information services and the most important provider for the European Union and its Member States. Based on scientific principles and methods, the European Statistical System will offer and continuously improve a programme of harmonized European statistics that constitutes an essential basis for democratic processes and progress in society.”

Excurse: European Statistics Code of Practice

***Mission** of the European Statistical System:*

“We provide the European Union, the world and the public with independent high quality information on the economy and society on European, national and regional levels and make the information available to everyone for decision-making purposes, research and debate.”

⇨ To realize mission and vision: Code of Practice based on 15 Principles

Excurse: European Statistics Code of Practice

1) Professional Independence
2) Mandate for data collection
3) Adequacy of resources
4) Commitment to quality
5) Statistical confidentiality
6) Impartiality and objectivity
7) Sound methodology
8) Appropriate statistical procedures
9) Non-excessive burden on respondents
10) Cost effectiveness
11) Relevance
12) Accuracy and reliability
13) Timeliness and punctuality
14) Coherence and comparability
15) Accessibility and clarity

Institutional Environment
Institutional and organisational factors have a significant influence on the effectiveness and creditability of a statistical authority developing, producing and disseminating European Statistics.

Statistical Processes
European and other international standards, guidelines and good practices are fully observed in the processes used by the statistical authorities to organise, collect, process and disseminate European Statistics. The credibility of the statistics is enhanced by a reputation for good managment and efficiency.

Statistical Output
Available statistics meet users' needs. Statistics comply with the European quality standards and serve the needs of European institutions, governments, research institutions, business concerns and the public generally.