# Chapter 4
# Data

# 4.1 Content of this Chapter

All the aspects of data:

1. Identification and prioritization of data
2. How to collect and get data
3. How and why to harmonize, rescale and cleansing data
4. Discovery of relationship in data
5. Documentation and reporting of findings
6. Re-definition of the business and analytics problem statement by use of the data analytics result

# 4.2 Identification and prioritization of data

**What we have already done:**

1) We have a business problem statement: first assessment of available data
2) We have translate the what into the why, and thus, had to think about different options / alternatives to perform the analytics project
3) We proposed a set of drivers: input variable and outcomes
4) We proposed a set of (not necessarily causal) relationship between the variables

# 4.2 Identification and prioritization of data

**Identification and prioritization**

1) Decide the characteristic of the input variable (mean, distribution, text, etc)
2) Determine which data can cover this characteristic (e.g. personal data for determination of the age / distributions of ages)
3) Determine which data types are most preferable and prioritize them
4) List all data one already has or one knows they are available
5) If missing data, either one has to go for doing an inventory of additional data with your customer or if not available one has to collect them.
6) If data are neither available nor collectable in due time, one has to refine or redefine the analytical or the business problem eventually.

# 4.2 Identification and prioritization of data

**But, you must become clear about the type of data you need and you finally have / can receive.**

A thorough understanding of the data is required.

# 4.2 Identification and prioritization of data

**Types of data:**

There are many different types of classification of data:

- Hard data vs. soft data
- Numerical vs. categorical data
- Ordinal vs. nominal data
- Discrete vs. continuous data
- Cross-sectional data vs. time series
- Structured vs. unstructured data

# 4.2 Identification and prioritization of data

**Types of data:**

There are many different types of classification of data:

- Primary vs. secondary data
- Meta data
- Dummy variable
- Binned (or discretized) data
- Binary data

# 4.2 Identification and prioritization of data

**Some definitions (repetition):**

A *population* includes all of the members / items of interest in a study

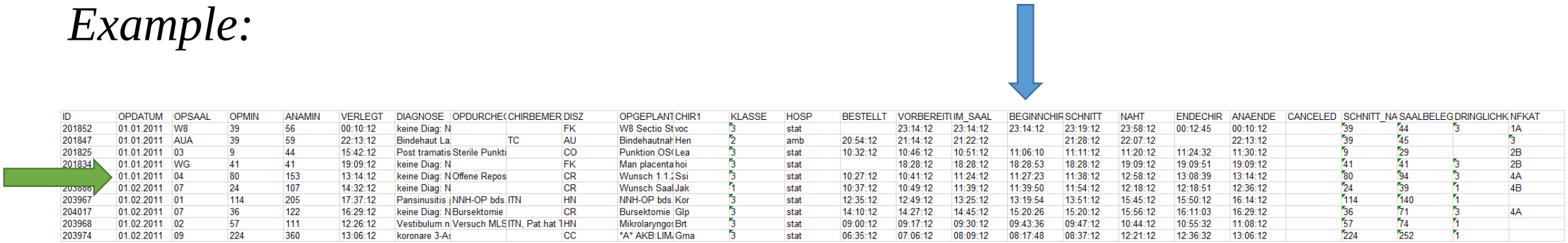
A *sample* is a subset of the population. A sample is of determined randomly and preferably a representative of the whole population.

# 4.2 Identification and prioritization of data

**Some definitions (repetition):**

A *data set* (of structured data) is usually an array of data with variables in columns and observations in rows.

*Example:*



| ID | OPDATUM | OPSAAL | OPMIN | ANAMIN | VERLEGT | DIAGNOSE | OPDURCHE | CHIRBEMER | DISZ | OPGEPLANT | CHIR1 | KLASSE | HOSP | BESTELLT | VORBEREIT | IM_SAAL | BEGINNCHIR | SCHNITT | NAHT | ENDECHIR | ANAENDE | CANCELED | SCHNITT_NA | SAALBELEG | DRINGLICHK | NFKAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 201852 | 01.01.2011 | W8 | 39 | 56 | 00:10:12 | keine Diag: N | | | FK | W8 Sectio St | voc | 3 | stat | | 23:14:12 | 23:14:12 | 23:14:12 | 23:19:12 | 23:58:12 | 00:12:45 | 00:10:12 | | 39 | 44 | 3 | 1A |
| 201847 | 01.01.2011 | AUA | 39 | 59 | 22:13:12 | Bindehaut La | | TC | AU | Bindehautnal | Hen | 2 | amb | 20:54:12 | 21:14:12 | 21:22:12 | | 21:28:12 | 22:07:12 | | 22:13:12 | | 39 | 45 | | 3 |
| 201825 | 01.01.2011 | 03 | 9 | 44 | 15:42:12 | Post tramatis | Sterile Punkti | | CO | Punktion OS( | Lea | 3 | stat | 10:32:12 | 10:46:12 | 10:51:12 | 11:06:10 | 11:11:12 | 11:20:12 | 11:24:32 | 11:30:12 | | 9 | 29 | | 2B |
| 201834 | 01.01.2011 | WG | 41 | 41 | 19:09:12 | keine Diag: N | | | FK | Man placenta | hoi | 3 | stat | | 18:28:12 | 18:28:12 | 18:28:53 | 18:28:12 | 19:09:12 | 19:09:51 | 19:09:12 | | 41 | 41 | 3 | 2B |
| [illegible] | 01.01.2011 | 04 | 80 | 153 | 13:14:12 | keine Diag: N | Offene Repos | | CR | Wunsch 1.1.2 | Ssi | 3 | stat | 10:27:12 | 10:41:12 | 11:24:12 | 11:27:23 | 11:38:12 | 12:58:12 | 13:08:39 | 13:14:12 | | 80 | 94 | 3 | 4A |
| [illegible] | 01.02.2011 | 07 | 24 | 107 | 14:32:12 | keine Diag: N | | | CR | Wunsch Saal | Jak | 1 | stat | 10:37:12 | 10:49:12 | 11:39:12 | 11:39:50 | 11:54:12 | 12:18:12 | 12:18:51 | 12:36:12 | | 24 | 39 | 1 | 4B |
| 203967 | 01.02.2011 | 01 | 114 | 205 | 17:37:12 | Pansinusitis | NNH-OP bds | ITN | HN | NNH-OP bds | Kor | 3 | stat | 12:35:12 | 12:49:12 | 13:25:12 | 13:19:54 | 13:51:12 | 15:45:12 | 15:50:12 | 16:14:12 | | 114 | 140 | 1 | |
| 204017 | 01.02.2011 | 07 | 36 | 122 | 16:29:12 | keine Diag: N | Bursektomie | | CR | Bursektomie | Glp | 3 | stat | 14:10:12 | 14:27:12 | 14:45:12 | 15:20:26 | 15:20:12 | 15:56:12 | 16:11:03 | 16:29:12 | | 36 | 71 | 3 | 4A |
| 203968 | 01.02.2011 | 02 | 57 | 111 | 12:26:12 | Vestibulum n | Versuch MLS | ITN, Pat.hat T | HN | Mikrolaryngo | Brt | 3 | stat | 09:00:12 | 09:17:12 | 09:30:12 | 09:43:36 | 09:47:12 | 10:44:12 | 10:55:32 | 11:08:12 | | 57 | 74 | 1 | |
| 203974 | 01.02.2011 | 09 | 224 | 360 | 13:06:12 | koronare 3-As | | | CC | *A* AKB:LIM | Gma | 3 | stat | 06:35:12 | 07:06:12 | 08:09:12 | 08:17:48 | 08:37:12 | 12:21:12 | 12:36:32 | 13:06:12 | | 224 | 252 | 1 | |

# 4.2 Identification and prioritization of data

**Some definitions (repetition):**

A *variable* (or *field* or *attribute*) is a characteristic of the items of the population.

An *observation* (or *case* or *record*) is a list of all variable values for a single member / item of a population.

# 4.2 Identification and prioritization of data

**Hard data vs. soft data**

- *Hard data* are data that is collected by scientific observation and measurement (e.g. experimentation)
- *Soft data* are data that is explored from interviews and reflective opinions and preferences.

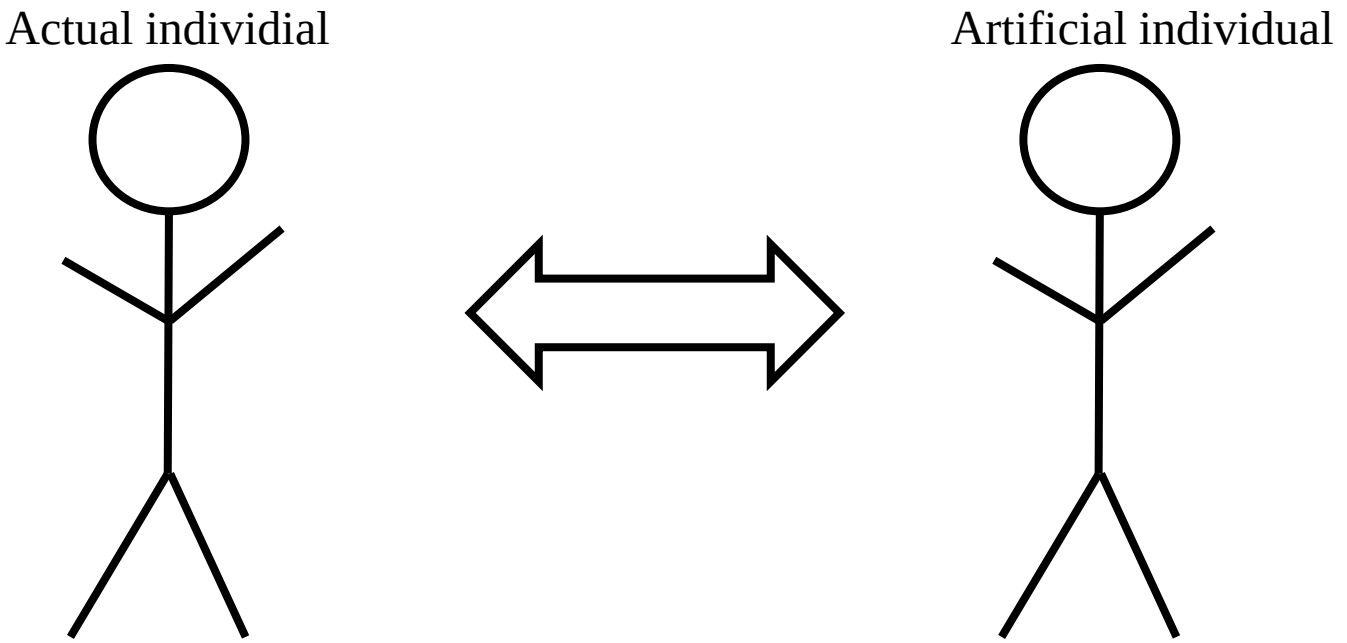
*Example:*

Testing two new products based on sold items and revenues vs. explore in interviews with consumers opinions and preferences of these products.

# 4.2 Identification and prioritization of data

Unfortunately, in most cases we have soft data.

Thus, how to translate soft data into hard data?




Hypothesize an artificial individual whose preferences and beliefs can be completely described with hard data

# 4.2 Identification and prioritization of data

*Example (economy):* The rationale investors that have all the same and full market information.

# 4.2 Identification and prioritization of data

*Example (operating room optimization):*

Soft data: interviews with surgeons and management how decisions are made when and who can operate an emergency.

Translation into hard data: Development a set of rules to achieve the same behavior:

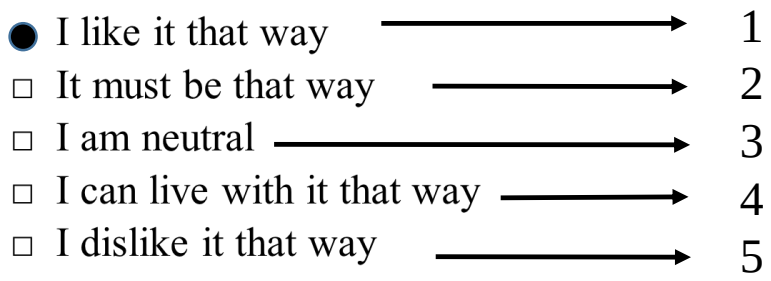
- If there is a sufficient long free slot in the discipline of the emergency, then allocate it first there
- if the time until the operating rooms are closing is less than 6 hours for an emergency category 2 (maximum waiting time for an operation is 6 hours), then allocate it to emergency shift (after the closing of the standard operating rooms)
- And so on….

# 4.2 Identification and prioritization of data

**Numerical vs. categorical data**

A data or variable is called numerical if meaningful arithmetical operations can be performed. Otherwise, it is called catecorical.

*Example categorical:*



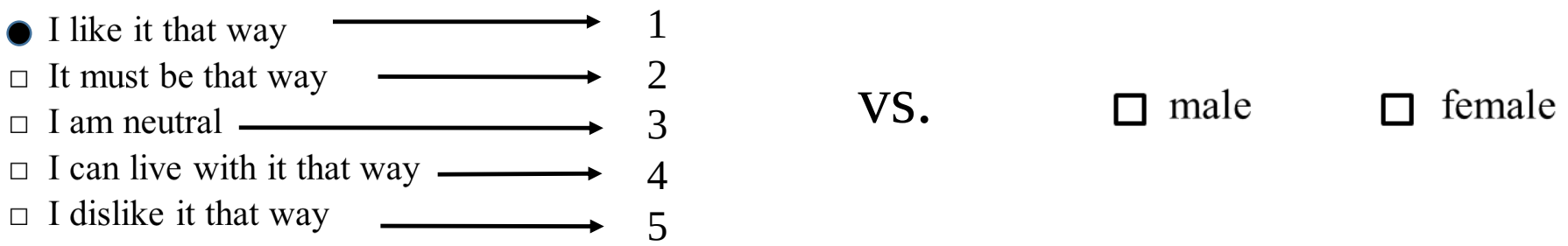
# 4.2 Identification and prioritization of data

**Ordinal vs. nominal data**

A categorical data is *ordinal* if there is a natural ordering of its possible categories.

If there is no such natural ordering, it is called *nominal*.

*Example:*



# 4.2 Identification and prioritization of data

**Discrete vs. continuous data**

A numerical variable is *discrete* if it results from a count, such a number of customers who buying a certain product.

A *continuous* variable is the result of an essentially continuous measurement, such as the waiting time of patient.

# 4.2 Identification and prioritization of data

**Cross-sectional data vs. time series**

*Cross-sectional* data are data on a cross section of a population at a distinct point in time.

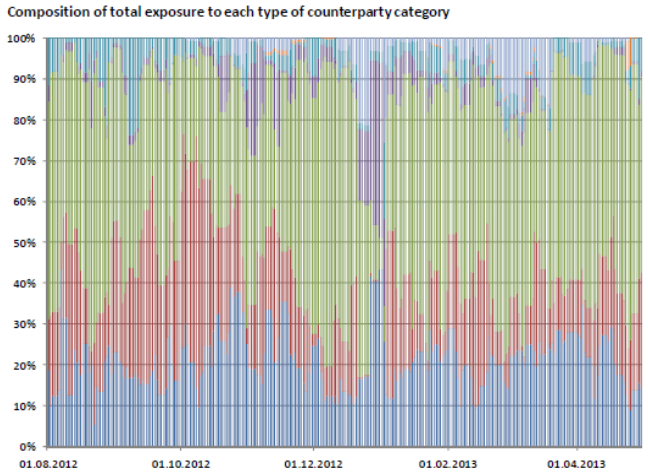
*Time series data* are data collected over time.

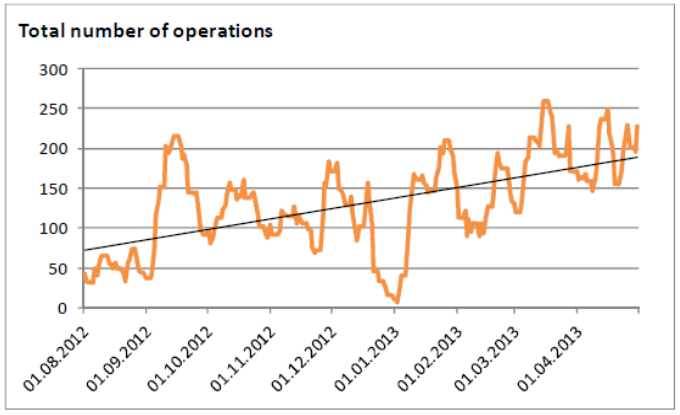
# 4.2 Identification and prioritization of data

**Cross-sectional data vs. time series**

*Example: Trading activities of a bank over a certain time horizon.*

1) *Proportional composition of the traded instruments per day*
2) *Development of the number of daily trades*



# 4.2 Identification and prioritization of data

**Structured vs. unstructured data**

*Structured data* can be put into rows and columns

*Unstructured data* cannot be put anymore in rows and columns e.g. a text or a movie

# 4.2 Identification and prioritization of data

*Examples:*

The Connected Business

Demand for analytics skills outstrips supply in all sectors

By Paul Solman

The rapid growth of big data has been such that few sectors can ignore it. A survey last year by SAP, the business software group, found that 92 per cent of respondents had seen the volume of data in their organisation increase during the past 12 months, while three-quarters of respondents believed their organisation needed new data science skills.

The problem for many is finding the analytics expertise to understand what this information means as the demand for data scientists outstrips supply.

"We have seen a significant year-on-year increase in the demand for data scientists in both the UK and overseas," says Jon Palling, global sector head for technology at STHree, the recruitment group. "The laws of supply and demand have encouraged many permanent employees to switch to contracting for greater financial reward."

Yet the signs are that analytics software companies are moving to help fill the skills gap by offering training, not only on their own products but in data science. Cloudera, which provides software based on the Hadoop data storage and processing framework, reports strong demand for its courses.

The group claims to have trained 20,000 people in the classroom, and another 100,000 through online courses and partnerships.

"We focus on the entire breadth of audiences," says Ryan Goldman, senior product marketing manager at Cloudera. "The primary focus is on current data professionals who want to first take on big data and move into building data science tools.

"But we also understand there is a larger audience trying to learn more about big data," he says. "So we also have online courses for those people who are just coming out of college or thinking about moving into big data."

Hands-on experience and a programme of certification means Cloudera's training can produce professionals who are "ready to be hired", says Mr Goldman. "We're taking what we know from the market . . . to understand what their end-user needs to know how to do. So we build our training around those real-world skill sets," he says.

"Part of the value of the certification process is bringing some kind of standard definition to what it means to be a data scientist," adds Sean Owen, director of data science at the company. "Not everyone is going to agree, but at least we're out there with a working definition as defined by the training course and the certification."

Alwin Magimay, head of digital and analytics at KPMG UK, the consultancy, says: "The analytics market is a bit like computer programming was in the 1990s – we're in the phase of building up competence and it's the beginning of a long journey."

To help meet demand in the finance, retail, technology and media sectors, KPMG is sponsoring the first UK Data Science Summer School, which aims to recruit 100 PhD students in Europe and turn them into data scientists during a five-week intensive course.

Again, the emphasis is on practical skills. "The summer school will give clients a platform to see how students can work on the data," says Mr Magimay. "Sixty per cent of the summer school will be theoretical but 40 per cent will be live client issues."

He adds: "The mobile app industry didn't exist five years ago. Some companies are blazing a trail and some are just trying to figure out what they can do. The title 'data scientist' implies an academic slant, but it's increasingly important to business."

"Businesses increasingly need insight to drive decision making," says Lance Fisher, chief information officer at STHree. "There is undoubtedly an increase in demand for good data analysts who understand the business and can look for trends in masses of data."

Nevertheless, some believe demand for data skills may only be satisfied when analytics becomes accessible to a wider cross-section of employees.

Twenty-eight per cent of workers use predictive tools regularly, according to the SAP survey, and that figure is expected to rise to 42 per cent over the next five years.

Further, 84 per cent of respondents said they wanted training to integrate analytics into their daily work.

US company QlikTech has developed QlikView, analytics software it says can be used without any data science expertise. The group's clients range from retailers, healthcare trusts and local governments to the Swedish police and even a Peruvian fishing co-operative.

"People are natural data scientists and have the capacity to carry out quite complex evaluations of data," says Donald Farmer, a vice-president at QlikTech. "The problem has been that traditional analytics systems have often become highly technical.

"QlikView is constructed in a natural way, so people can understand it easily. Only very rudimentary training is needed to find your way around the screen, and certainly no training needed in statistical analysis."

Mr Farmer adds: "We're not negating the role of data scientists – they will always be needed at some level. But it sometimes seems that the only way to do business is for a data scientist to learn the business or the business person to learn data science. We think there is another way."

RELATED TOPICS United Kingdom

Technology tackles health hazards in sport

Can the 'connected' car save carriers?

Mobile World Congress: Top three innovations


# 4.2 Identification and prioritization of data

**Structured vs. unstructured data**

But: data analytics is typically only applied /can only be applied to structured data.

Thus, how to analyze unstructured data?

Typically, data analytics algorithms cannot access the unstructured data directly.

The unstructured data are first “structured” to be an input for a data analytics process.

# 4.2 Identification and prioritization of data

*Example:* Matching of fingerprints

- Identify important points
- Set up a map / polygon
- The structured maps are the analysed

Picture source: wikipedia «fingerprint»

# 4.2 Identification and prioritization of data

*Example:* Text Mining

One have to extract content and not single words. Thus, the relevant words have to be extracted in a structured way.

*Example:* Big Data

Typically, first application of so-called Hadoop and/or Map-Reduce, and then the data are structured for data analytics and statistical tools.

# 4.2 Identification and prioritization of data

**Primary vs. secondary data**

*Primary data* are data which are not yet available and has to be measured and collected first.

*Example:*

- Company starts producing, recording and storing certain data e.g. new reporting data or new regulatory data
- IFRS 9, Impairment: there will be new requirements that companies must quantify expected credit losses on financial instruments.
- This has not been required and done in the past thus, they have to start collect the required risk data, calculate the expected credit losses, have to store them and have to report them in the financial statements

# 4.2 Identification and prioritization of data

**Primary vs. secondary data**
*Secondary data* are data already collected by someone else.

*Example:*

- Internet
- Accounting / reporting data in companies
- All your student data at ETH / UZH
- Statistical data e.g. census (Swiss Federal Statistical Office)
- Log data in IT systems
- And so on

# 4.2 Identification and prioritization of data

**Primary vs. secondary data**

*Advantage of secondary data:*

- It safes time as the data are already available
- Some data are already structured and cleaned

*Disadvantage of secondary data:*

- Data may be outdated
- Data may already be processed (or manipulated)

# 4.2 Identification and prioritization of data

**Meta data**

*Meta data* are data about the data i.e. describes the data.

*Examples:*

- Date when data has been collected
- Purpose for the collection
- How the data was collected
- Size / volume of the data
- Image resolution
- Dates of changes in the data / dates of access to the data
- Tags in social media

# 4.2 Identification and prioritization of data

**Dummy variable (binary data, indicator)**

A *dummy variable* is a 0 – 1 coded variable for a specific category. Typically, 1 labels the observations in this category and 0 for all other observations not in that category.

**Binned (or discretized) data**

*Binned data* correspond to a numerical variable that has been categorized into discrete categories.

These categories are called *bins*.

# 4.2 Identification and prioritization of data

*Use Case: Web Page Analytics*

Business issue: A car brand wanted to optimize the success rate of customer contacts and thus, to increase the sales of cars.

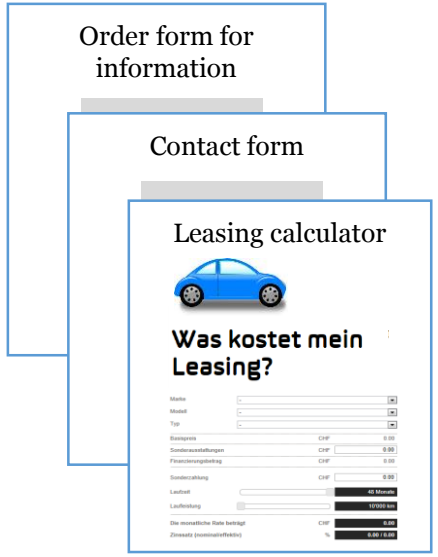
Analytics problem:

- Based on the access date of the web page of the car brand, classification of the different types of web page visitors e.g. with which probability such a visitor will buy a car
- Based on these classes, optimization of the optimal contact procedures

# 4.2 Identification and prioritization of data

*Use Case: Web Page Analytics (cont'd)*

Data available:




**Log-File:**

Recording and storage of the web page behavior of the visitors e.g.

- Which page have been visited
- How they navigate through the pages
- Which information has been ordered
- What calculation in the leasing calculator has been performed
- What time they have visited the web page
- Data entered into the forms (name, address, age, age of the car, etc)
- Which cells of a form has been filled out
- Recurring visits of the page or one-off visit
- By which device the page was visited

**Sales data:**

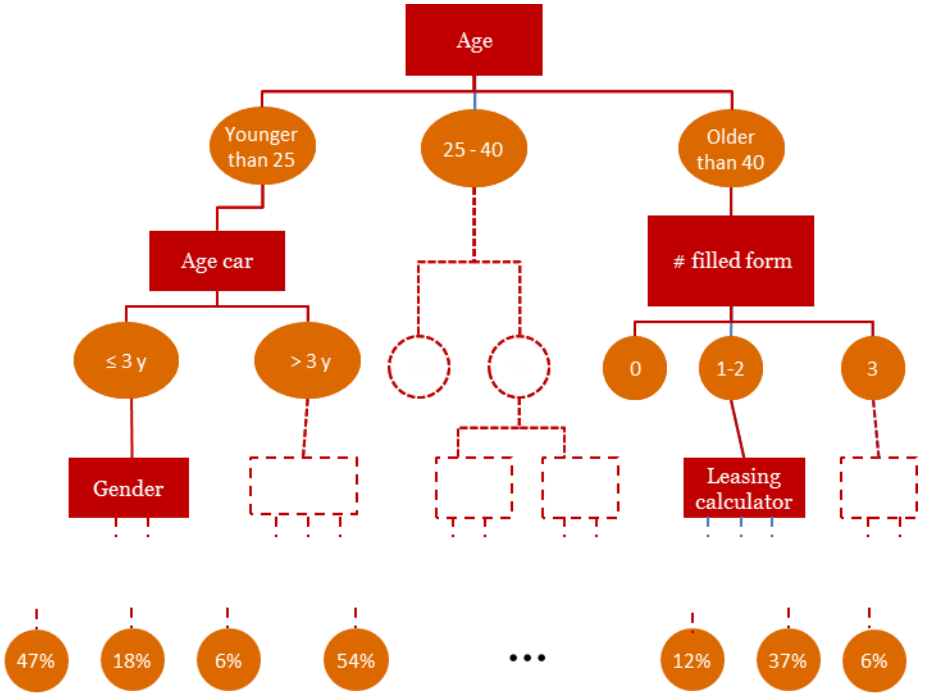
Customers who have bought a car

- when,
- price segment,
- where,
- payment option
- personal information

# 4.2 Identification and prioritization of data

*Use Case: Web Page Analytics (cont'd)*

Classification:

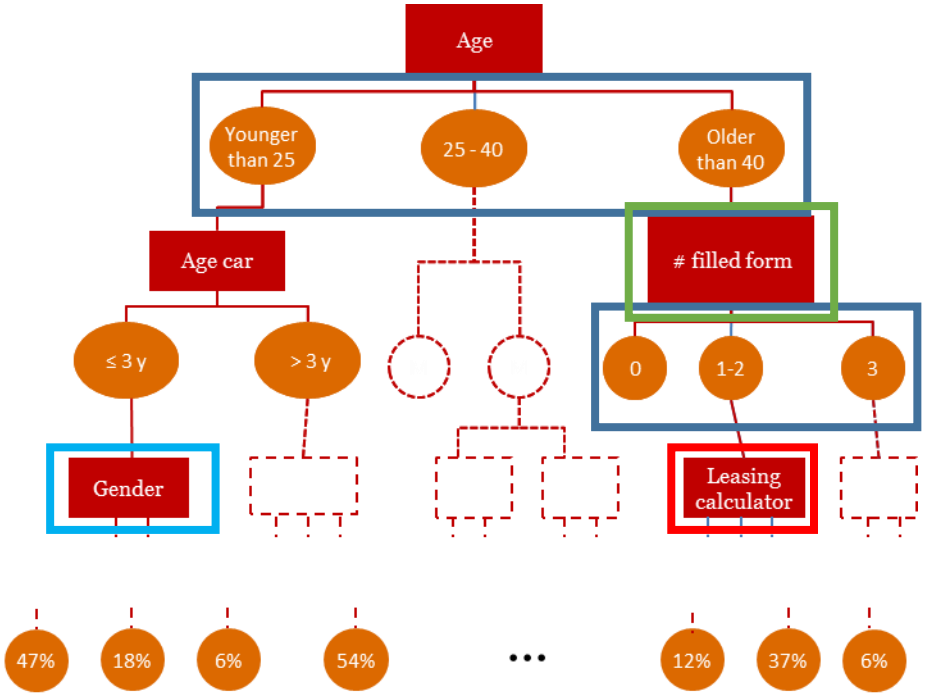



Probability of a car sale per customer class

# 4.2 Identification and prioritization of data

*Use Case: Web Page Analytics (cont'd)*

Data used:




Probability of a car sale per customer class

Binned data

Meta data

Categorial data (nominal)

Continuous data

Numerical
Cross-sectional

# 4.2 Identification and prioritization of data

*Use Case: Web Page Analytics (cont'd)*

Finally, based on the probability of buying a new car, the response medium has been optimized (cost – return optimization) and some response time analysis:

- > 43%: call and offer them a test drive
- < 8%: send them twice a year the usual advertising brochure
- 8% - 27%: send them some general car brochures within the next 2 months
- 27% - 43%: send them tailored brochures within the next 3 weeks; follow up brochures every 2 months