Data collection and analysis

Requirements for the Project work and Thesis courses

Aims and scope

- To base the business problem/theoretical research under investigation in the Project Work (and in the subsequent Thesis) to draw valid and reliable conclusions.
  - These conclusions lay the foundation for managerial decisions that have an impact on resource allocation.
  - Therefore, it is of high importance to have relevant and reliable data and to identify the right tools and methods that allow us to process these data, since it influences the success of the problem solving.
- Together with data collection methodological issues should also be considered which are discussed in another presentation in more detail.

Basic concepts

- **Primary and secondary data**
  - During a research, both types may be needed, as it is not certain that secondary data alone provide enough information about the problem itself or serve adequately the problem solving.
  - It is also possible that secondary data is not available in a given business case.

- **Primary and secondary research methods**
  - Secondary data is collected and processed using secondary methods, while primary data is collected and analysed using primary methods.

- First, we usually examine the availability of secondary data using secondary methods.

Methodological issues

Secondary methods	Primary methods

• How to find the right secondary methods to collect and analyse secondary data?
    • Desk research: library, offline and online research methods
	• How to find the right primary methods to collect and analyse primary data?
	    • Survey
	    • Focus group
	    • Interview
	    • Observation
	    • Experiment

1. Introduction
   - The importance and role of data collection in the project work
   - Purpose of data collection and expectations regarding data collection
2. Classification of data
   - Primary vs secondary data: When? What is it good for? How to collect?
   - Quantitative vs qualitative data: When? What is it good for? How to analyse?
     - Connection with measurement levels: which data collection solution fits which measurement level?
3. Quantitative data
   - Sampling, graphical display of collected data
4. Qualitative data
   - Sampling

Introduction

The importance and role of data collection in the research process of the project work

Aims and requirements of data collection

Data collection in the project work

- A clearly defined business problem or research question should lie at the heart of the project task, the solution of which is based on the utilization of knowledge acquired during the previous studies in our various business programs.

- The multifaceted purpose of data collection and processing is
    - to support the actuality of the business problem, research question, and to highlight the need for solving the problem or answering the theoretical research question.
    - to explore and analyse the system of conditions and circumstances that determine the success of the solution
    - to identify alternatives for the problem solution
    - to develop viable solution proposals,
    - to evaluate the success of the proposed solution, and by that way
    - to support the soundness of the business problem solution.

Data collection in the project work

- ## What is the purpose of data collection?
  - To give an adequate answer to the business problem / research question defined clearly in the project work.
  - To evaluate the data collected and thus try to get as much information as possible regarding the investigated business problem. This serves as the basis for choosing the right method for problem solving.

- ## What is NOT the purpose of data collection?
  - To collect too much data,
  - To collect irrelevant data that are not directly related to the business problem or research question
  - To choose data processing methods that are not relevant to the properties of the collected data
    - E.g. the normality of data is not tested, therefore, not relevant statistical methods are used when analysing data
    - E.g. we calculate the average and the standard deviation of data on a Likert scale.

Data collection in the project work

- The chosen business problem, research question therefore determines
  - what the population will be to be investigated, what conditions surround it,
  - what available data sources can be relied upon to examine the population,
  - whether further primary data collection is necessary for problem solving.

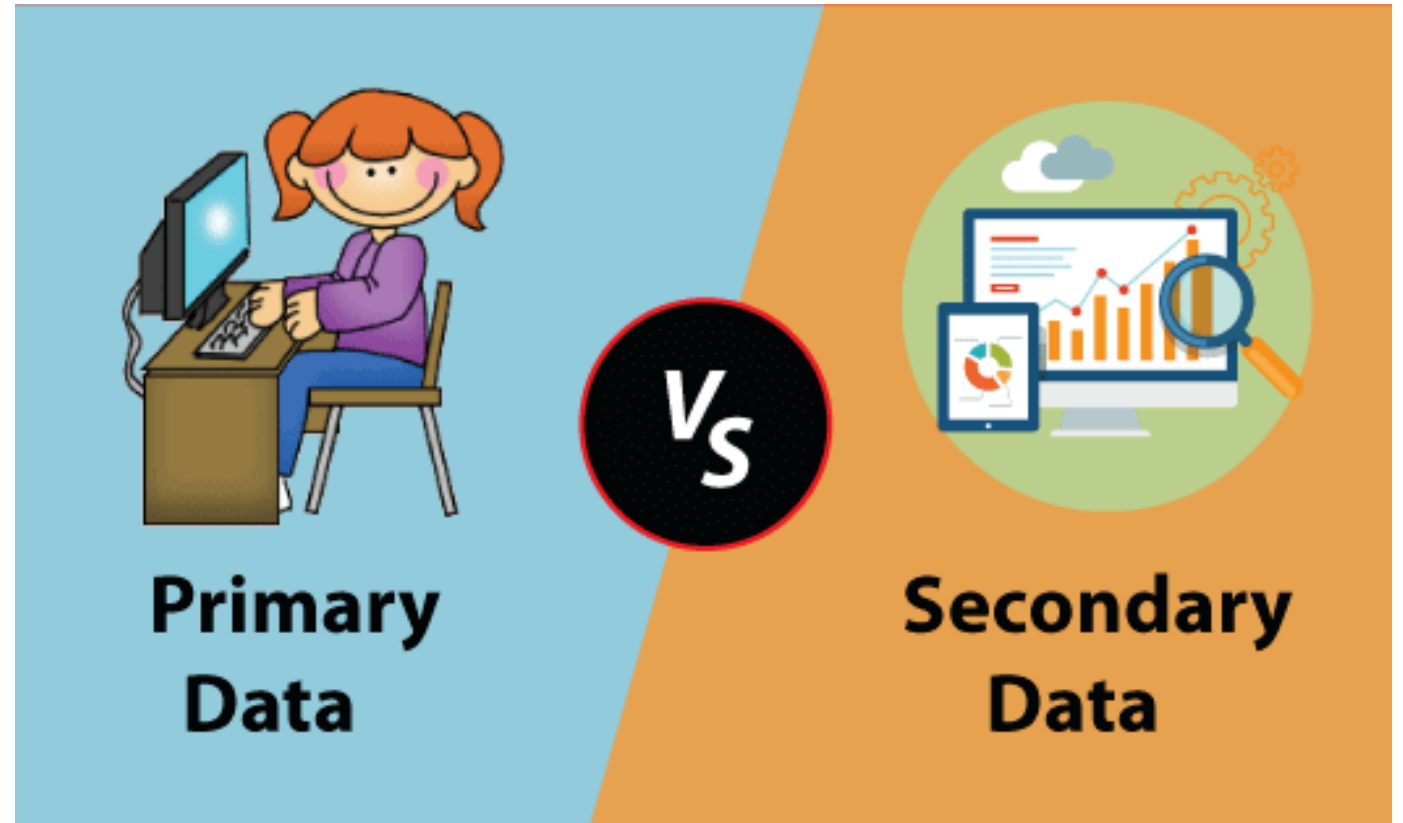- The success of problem solving and thus the answering of the research question largely depends on the secondary and/or primary data used during the research.

Danger ahead

- If we list personal opinions, intuitions, judgments, and statements.
- The purpose of data collection and processing is to increase relevance, validity and objectivity!

Classification of data

Secondary Data

Secondary data

Secondary data is data collected by someone else, not for the purpose of investigating the given problem. The essence of secondary data collection is the utilization of materials that have already been processed according to one's own perspective.

Secondary data could help in

- ✓ identifying and understanding the problem

- ✓ defining the problem more precisely

- ✓ the development of the research plan (for example, in the identification of key variables)

- ✓ answering certain research questions and testing hypotheses,

- ✓ the interpretation of primary data.

Sources of secondary data

External sources:

- Online statistical databases, databases available in libraries, scientific books and articles in the given topic

- KSH, World Bank, EU and OECD statistics, statistical yearbooks

- Monthly and quarterly reported data

- Regional, national, European, global data

- Specialized publications, industry analyses, studies, conference publications, forums

Internal sources:

- Various reports (e.g. quality records, market research reports, customer databases, accounting data, registers, financial reports)

- Internal company and business data (e.g. sales data per product group, customer / employee satisfaction measurements, financial and accounting statements, data)

Primary data

During **primary data collection**, first-hand information is collected and analysed for specific research purposes. Primary data collection arises when the necessary data does not exist (outdated, imprecise, unreliable), or the topic has not yet been worked on by others, and secondary data proves to be insufficient. Through surveys, focus groups, interviews, observations, measurements or experimentations the additional information could be collected.

Primary data collection is worthwhile when

✓ no suitable database is available,

✓ the available data is no longer valid,

✓ the information is not target group specific,

✓ there is an information limit for secondary data,

✓ any new information is needed.
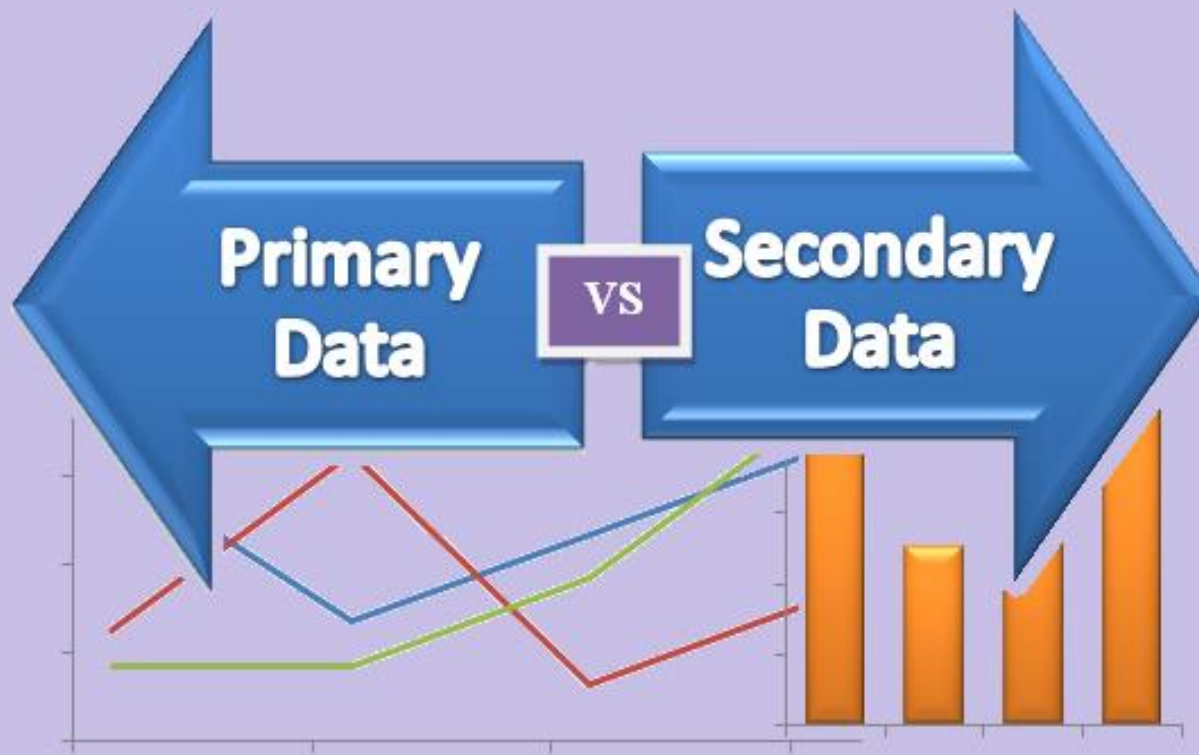
Secondary vs primary data

- **Secondary data is data collected for purposes other than the specific problem.**
  - They are usually available quickly, relatively cheap, could be collected in a shorter time.
  - They contribute to understanding and solving a research problem, but depending on the complexity of the research, primary research may also be necessary.
- **Primary data is information primarily collected to solve a specific research problem. The process of primary data collection is presented in the research process.**
  - The primary data collection process is more complex, usually longer, requires more resources and more expensive
- Both secondary and primary data can be qualitative or quantitative.

SECONDARY VS PRIMARY DATA

Secondary data – background information:

- The overview of the results originating from primary research

- Information related to trends, specific topics and background

- Less details about many things

- Well established facts

Primary data – collected for the specific problem:

- Specific data from a single source

- Observations, results in their original context

- More details, information about the specific problem

- New findings

Secondary data problems

- **Validity and reliability**
  - These are important concerns in research and they cannot be taken for granted when using secondary data.
  - While using secondary sources of information it should be well-researched that the content is genuine and authentic.
- **Personal bias**
  - In secondary sources the chances of bias are higher as compared to that in primary sources.
  - Newspapers, magazines and websites do not use rigorous and well-controlled methods in documentation. Most of the time such writings are opinion-based and they are far from facts. In these publications writers can distort the facts to make the situation look better or worse.
- **Availability of data**
  - To obtain highly rigorous, scientific and valid data from secondary sources you need to work hard as such data is hardly available in secondary sources.

Secondary data problems

- **Format of data**
  - The format of the data can be totally different and the researcher cannot use it in his/her research. Using another format in data collection that is not related to your research format can give biased and invalid results.
- **Quality of data**
  - Quality of the data is related to its accuracy and accuracy comes with rigorousness in collecting the data. It depends on the source that you are using in your research; scientific books and journals can provide you quality data.
  - Again newspapers and magazines cannot provide good data for research, they should be avoided.
- **Obsolete data**
  - Sometimes secondary sources are available to be used in the research but they are very old. Old data is of no use to be used in the research.

- For primary data collection you should consider
  - which population variables should the data collection cover
    - What else should I know in addition to what has been available so far to investigate the problem, i.e. what circumstances justify the further collection of data to increase the validity of problem solving and drawing conclusions?

- Primary data collection can also encourage the choice between the methods that can be used
  - What additional data and information are needed defines how to collect the data, what type of data is generated, and also basically defines the methods used to draw conclusions.

Primary data

Types of empirical research data

Source: Malhotra: Marketing research
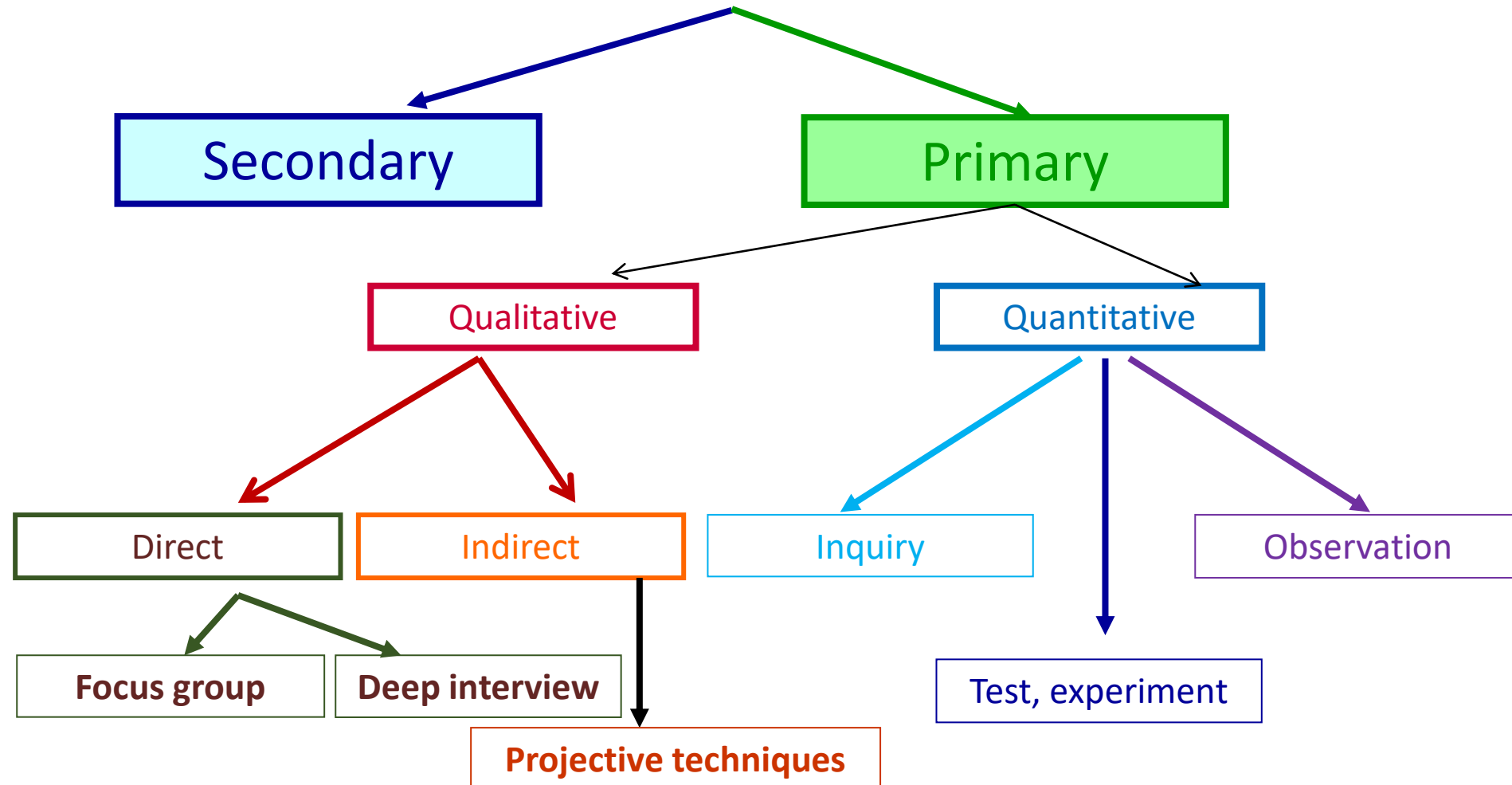
Classification of data

Qualitative vs quantitative
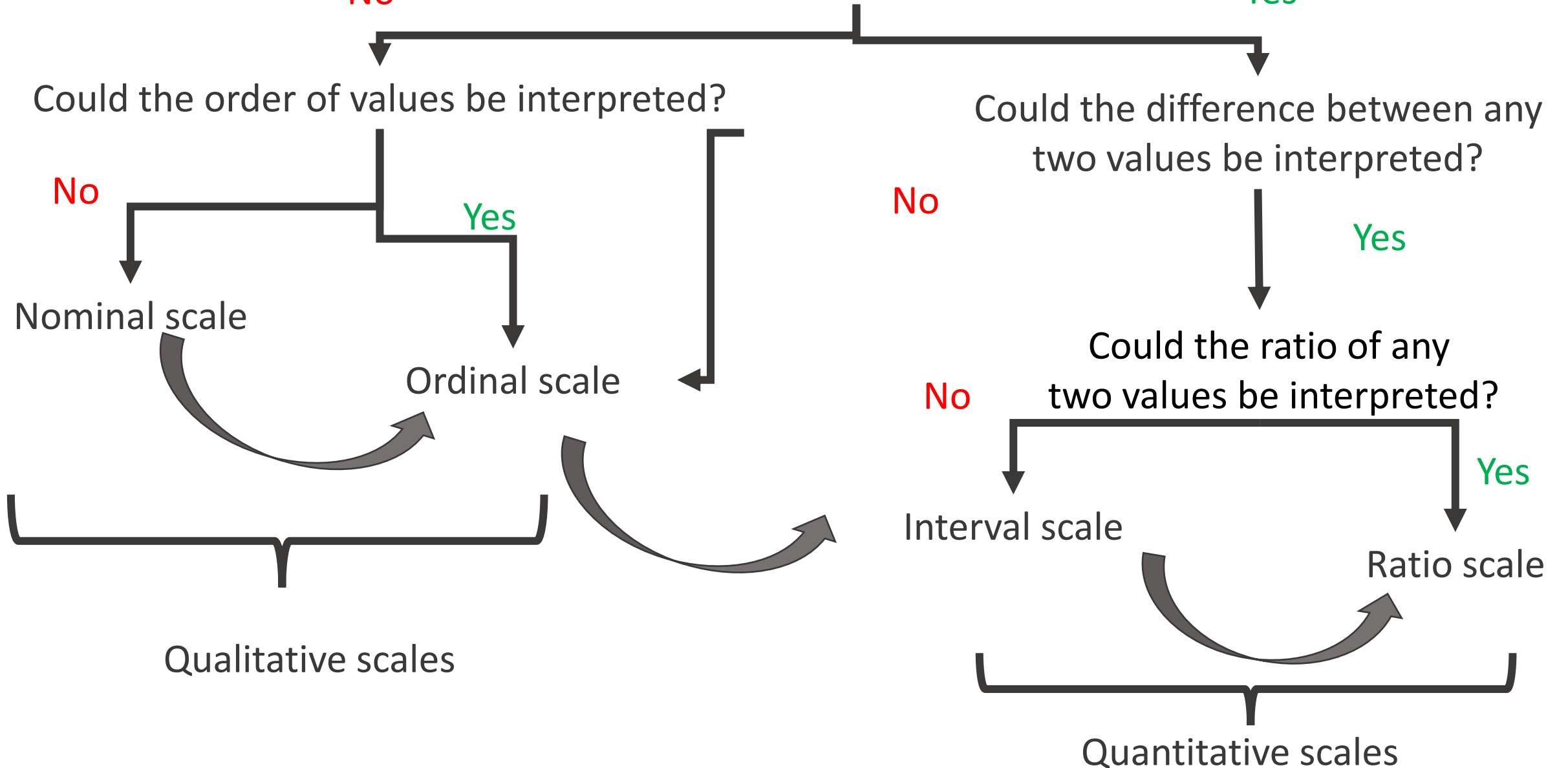
Quantitative data

We are looking for answers to quantitative questions (e.g. yield, age, sales)

- **When to collect?**
  - if we need exact numbers or statistical data.
- **What are they good for?**
  - to reach conclusions that can be expressed in numbers.
- **How to collect?**
  - by questioning (questionnaire, survey), observation, measurement, experiment → See methodological issues
- **What methods could be used to process?**
  - Generally, well-standardized methods are available for summarizing and organizing quantitative data. It is possible to analyse them e.g. with mathematical, statistical, accounting, etc. methods. (For methodological issues see also Business Statistics, Marketing research and Econometrics courses)
- **In principle, it can be reproduced and repeated!**

Qualitative data

We are looking for answers to qualitative questions (e.g. attitude, perception, motivation, understanding of thinking)

- **When to collect?**
  - If we want to better understand the basic question, or wish to discover new information (e.g. motivation, aspect, decision alternative).
- **What is it good for?**
  - It can help to bring thinking trends, perception, inner values to the surface and to interpret and understand the examined business problem in its context. The results are rather subjective, and we often do not receive tangible, quantifiable data.
- **How to collect?**
  - Most often, questioning (focus group, interview), observation
- **What methods could be used to process it?**
  - E.g. content analysis, specifically with qualitative analysis tools → see methodological issues
- **No two qualitative studies are the same!**

Measurement scales

- We use four measurement levels to describe the data, depending on the content of the values obtained through measurement and observation.
  - Qualitative data:
    - Nominal – survey, interview
    - Ordinal – ranking, Likert scale evaluations, the difference between the ranked factors cannot be interpreted
  - Quantitative data:
    - Interval - between any two adjacent values, there is the same distance both in terms of content and number
    - Ratio - the ratio of differences can also be measured
    - Point and interval estimation, hypothesis testing, relationship testing methods, time series analysis

Research methods

Quantitative

- Its purpose is to collect and analyze objective, quantifiable, statistically evaluable data

- SAMPLING is an essential issue

- We collect sample data so that we can confirm or reject our assumptions.

- Tools: mathematical-statistical methods, the possibility of generalization

- E.g. questionnaire surveys, satisfaction surveys

Qualitative

- Its purpose is a deeper understanding of problems, processes, human behaviour and trends (collection and analysis of opinions, attitudes and beliefs)

- It does not provide representative or statistically evaluable results

- It cannot be used to verify hypotheses, at most to formulate them. E.g. observation, experiment, in-depth interview, expert interview, focus group

- Sometimes it is more difficult to do it „right". It is more difficult to generalize, validity and reliability are questionable.

Measurement level of data

- It is necessary to choose a data processing methodology that matches the measurement level of the data!

- IMPORTANT: these measurement levels fundamentally influence the design of the data collection and the course of the data processing.

- STATISTICAL METHODS ARE SCALE DEPENDENT!

- However, the chosen methodology must reflect how I wish to collect data! – the relationship is mutual!

Collecting quantitative data

Sampling issues

Sampling issues

- In most cases, it is not possible to examine the entire population, so we have to select a random sample by examining which we can draw general conclusions about the entire population (See Business Statistics: Population vs sample)

- Basic issues when deciding on sampling
  - Identification of the sampling frame
  - Representativeness: the sample represents the population, i.e. the overall characteristics, properties, and parameters of the sample closely approximate the same overall characteristics of the population.
  - It is not an expectation to ensure representativeness if it is not possible or too expensive (must be stated within limits of the research)
  - Inclusion and exclusion criteria
  - Sample size
    - How important the results are, the heterogeneity of the population, the method used, the time and cost of the study

For further details see Hunyadi-Vita (2006): Statisztika közgazdászoknak, Malhotra (2009): Marketing research

Random sampling methods

- Random sampling is a part of the sampling technique in which **each sample has an equal probability of being chosen**. A sample chosen randomly is meant to be an unbiased representation of the total population.

- Types:
  - Simple non-random sampling
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling

For further details see: Hunyadi-Vita (2006): Statisztika közgazdászoknak, Malhotra (2009): Marketing research

Non random sampling methods

- This type of sample does not ensure representativeness, the conclusions that can be drawn are limited. Each element has a greater or lesser chance of being included in the sample.

- Sampling error cannot be quantified. It cannot be used for inferential statistical purposes.

- Non random sampling methods:
  - Arbitrary sampling
  - Expert selection
  - Snowball method
  - Quota sampling

For further details: Hunyadi-Vita (2006): Statisztika közgazdászoknak, Malhotra (2009): Marketing research
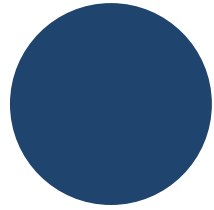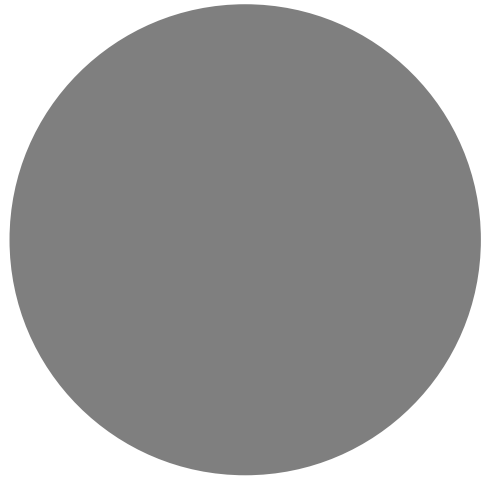
Collecting quantitative data

Graphical representation

Basic principles of graphic representation

- See more in Business Statistics course materials
- The chosen visualization depends on the measurement level of the data to be represented!
- Principles of graphical representation :
  - Clarity: shape, colors, backgrounds
  - Goal orientation: preferably a representation of a phenomenon
  - Simplicity: to help you get the right overview, less is more
  - Ability to reconstruct: indication of data source and methodological elements
  - Optical neutrality, area principle: sizing and scaling in accordance with what is being said
  - Homogeneity: use one type of diagram for one type of phenomenon

Aim of graphical presentation

What do I want to show with the visualization?

- Comparison
- Presentation of composition
- Overview of the distribution
- Expressing a relationship

How to choose?

- How many variables do we want to show? (one, two, three or more?)
- How much data do we want to display for the single variables? (A few or many?)
- Do we wish to present the values of the variable(s) as a function of time, or e.g. do we wish to display the differences between individual elements and groups?
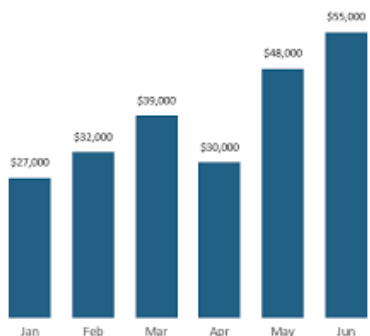
One variable
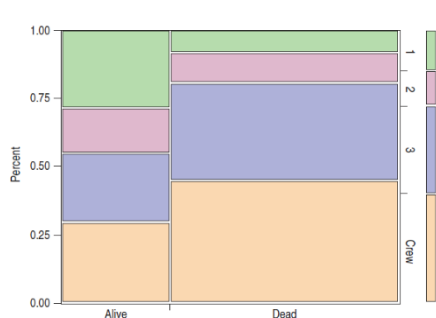
Two variables
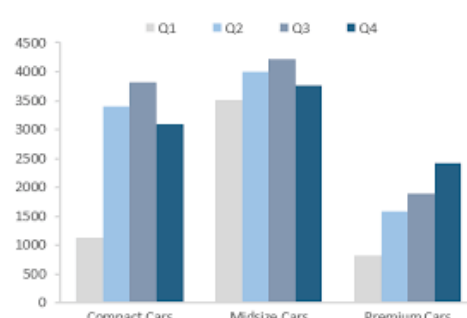
Three or more variables

Cyclic

Not cyclical

Many periods

Comparing variables

Comparing in time

1. Comparison

What is the purpose of visualization?

Few periods
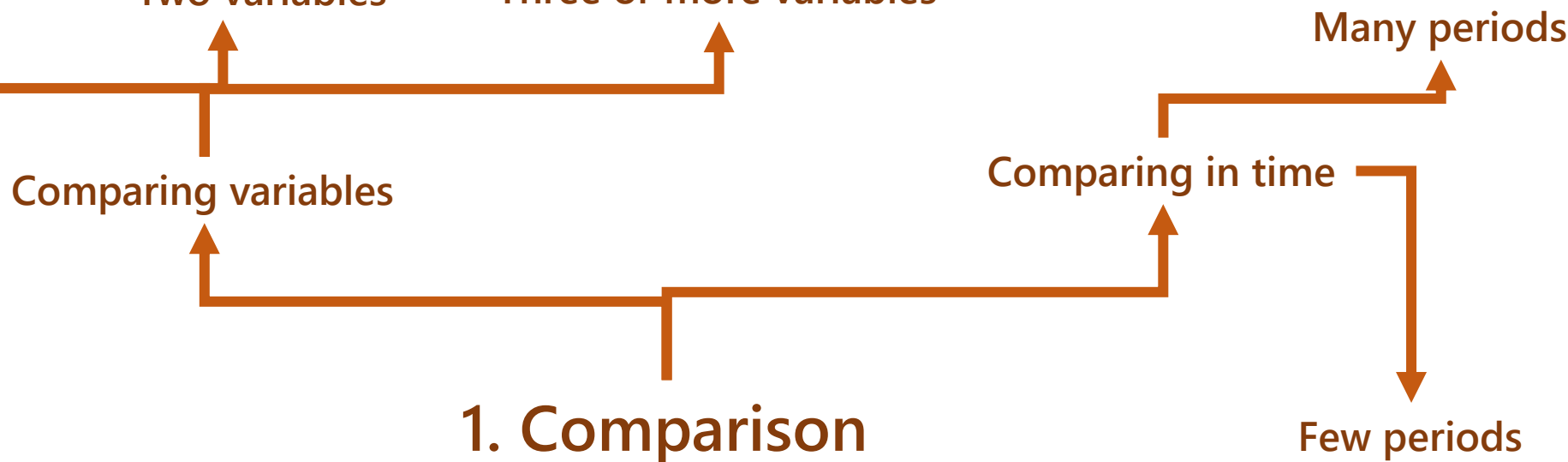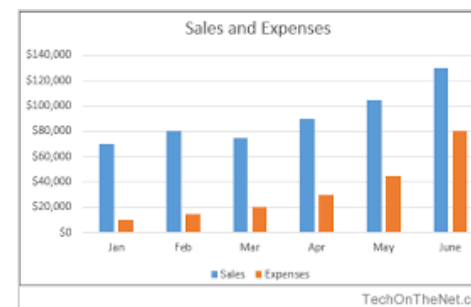
Few variables

Many variables
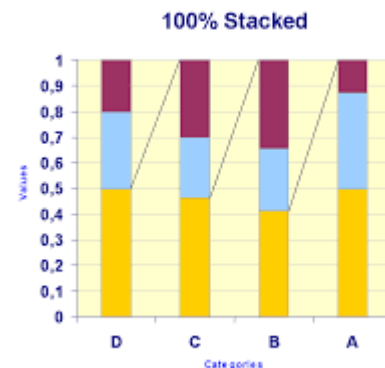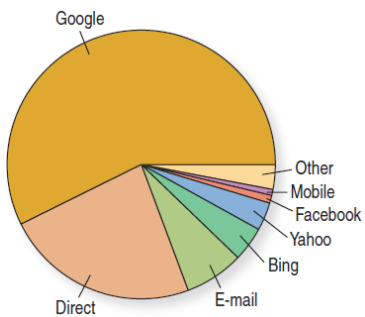
2. Illustrating composition

What is the purpose of visualization?

It does not change over time

It changes over time

Several periods

Few periods

Only the relative difference matters
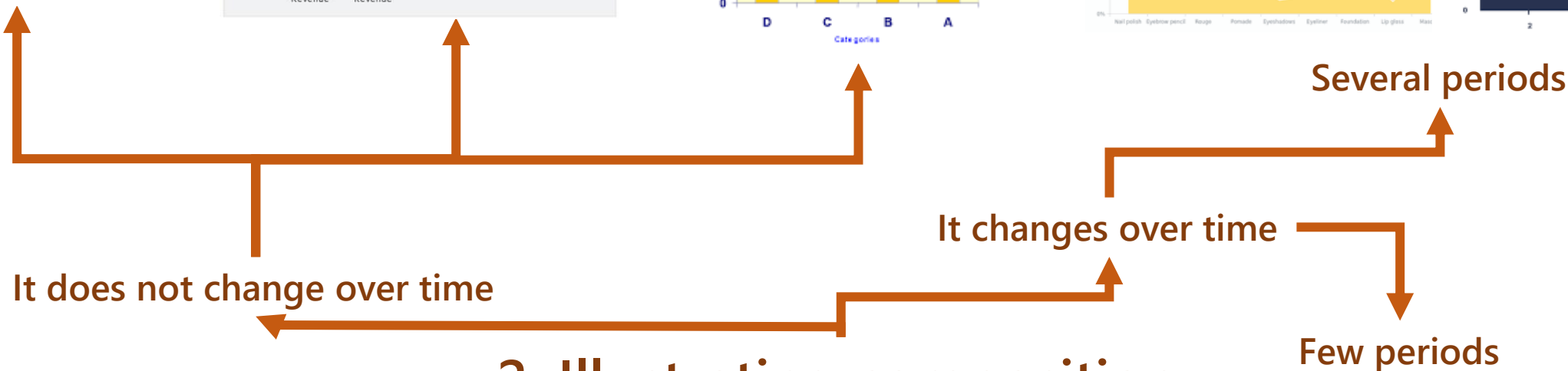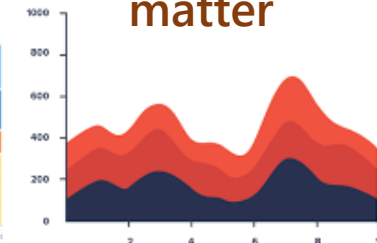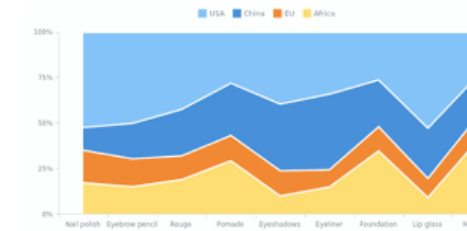
Both absolute and relative differences matter
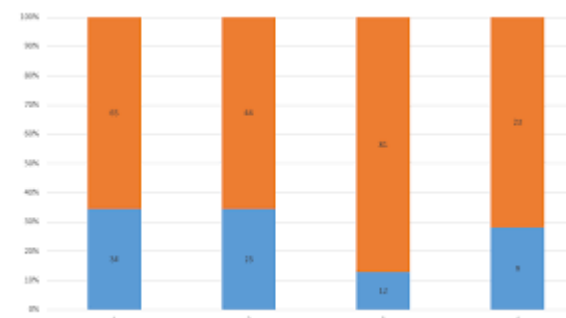
Only the relative difference matters
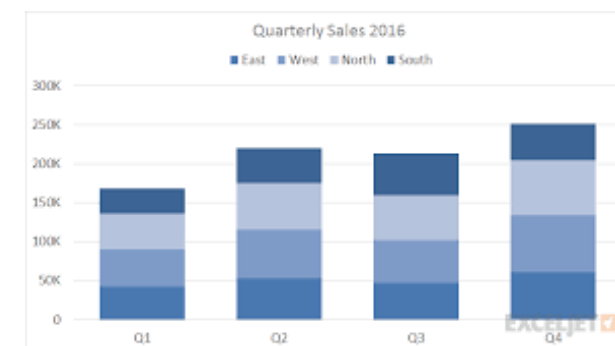
Both absolute and relative differences matter
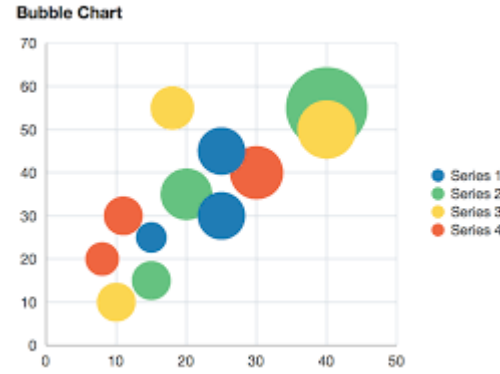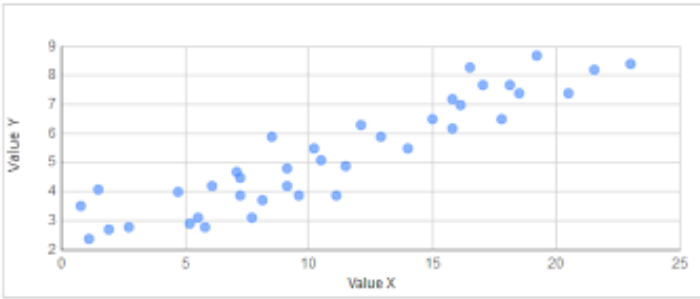
Two variables

Three variables

3. Illustrating relationship
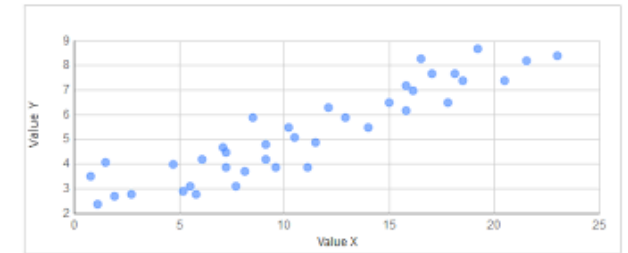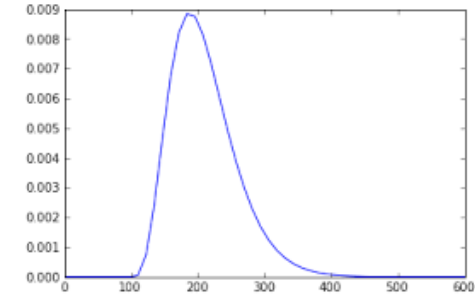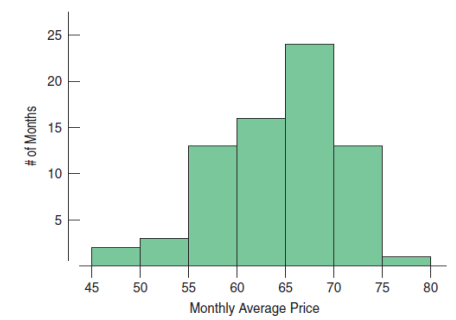
What is the purpose of visualization?

Few data

One variable

Many data

Two variables

4. Overview of the distribution

Three variables

In the business field, a significant part of the results are measurable (quantitative) data, which can be summarized and analysed using mathematical, statistical, accounting, financial analysis, etc. methods. However, they often record only the information that appears on the surface and do not provide an opportunity for a deeper understanding of problems, processes, and human behaviour.

Qualitative data
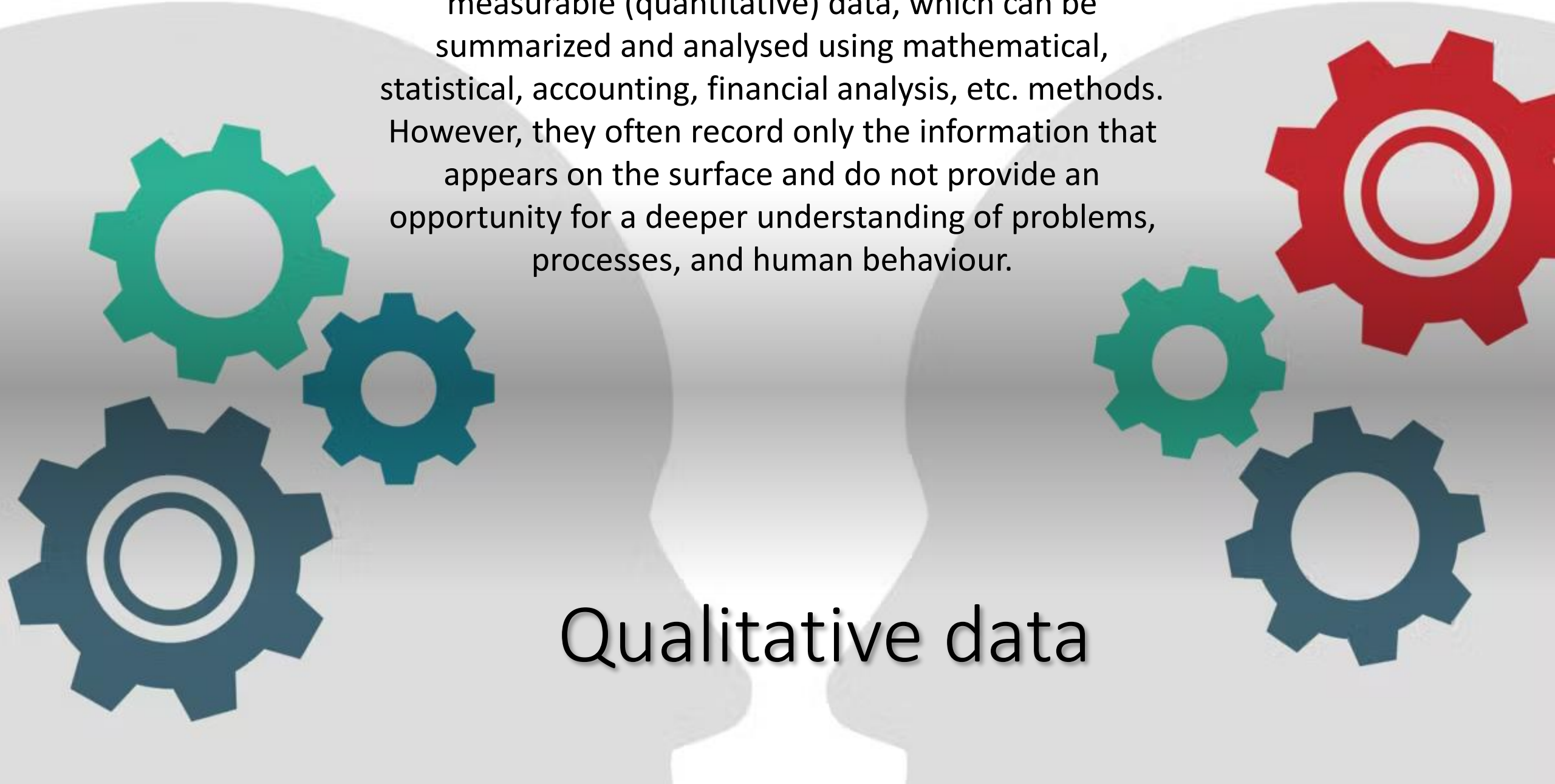
Sampling issues

Representativeness is not the goal!

Non-random sampling: typically, arbitrary or quota sampling

Typically, smaller sample size (5-20 people)

Inclusion and exclusion criteria (screening survey)

Typically, textual processing of results

Qualitative methods see in Methodological issues.

THE END.