October 17, 2025

Training Programme

09:15	Introduction
09:30	Data Mining
10:30	Coffee break
11:00	Planning a data analysis in the audit
12:00	Data Management
13:00	Lunch
14:30	Example: Detection of clusters and outliers
15:30	Data visualisation
16:15	Conclusion
16:30	End of training

Making a Difference Through Data

Day 3

UNIVERSITÄT DES SAARLANDES

Appropriate assessing, analysing and visualising your results Part 1: Data Mining

Wolfgang Meyer, Saarland University

Introduction to the Course

UNIVERSITÄT DES SAARLANDES

Data Mining

- Identifying and extracting relevant data from sources

Data Processing

- Sizeable effort in preparation for analysis

Data Analysis

- Using methods to interpret data and develop insights

Data Visualization

- Communicating results including visualization and data storytelling

Chapter 1

1.1 Data Sources
1.2 Systematic Research
1.3 Data Mining

Chapter 2

1.1 Evaluating Catalogue
1.2 Assessing Data Quality
1.3 Errors and Missings

Chapter 3

1.1 Data Management
1.2 Machine Learning

Chapter 4

1.1 Descriptive
1.2 Clustering
1.3 Combining
1.4 Longitudinal

Chapter 5

1.1 Basic Graphs
1.2 Advanced Visualization

Source: Hyman et al (2024): Data Analytics and Visualization. Hoboken: John Wiley & Sons, p. 13 (modified)

Content Chapter 1.1

- Types of Data Sources
  - Public Data (Statistics)
  - Databases
  - Process produced Data
  - Internet Data (“Big Data”)
- Access Options
  - Complete Data Sets
  - Limited Data Use
  - Aggregated Data
- Systematic Research
  - Principles
  - Process of Data Mining

Chapter 1

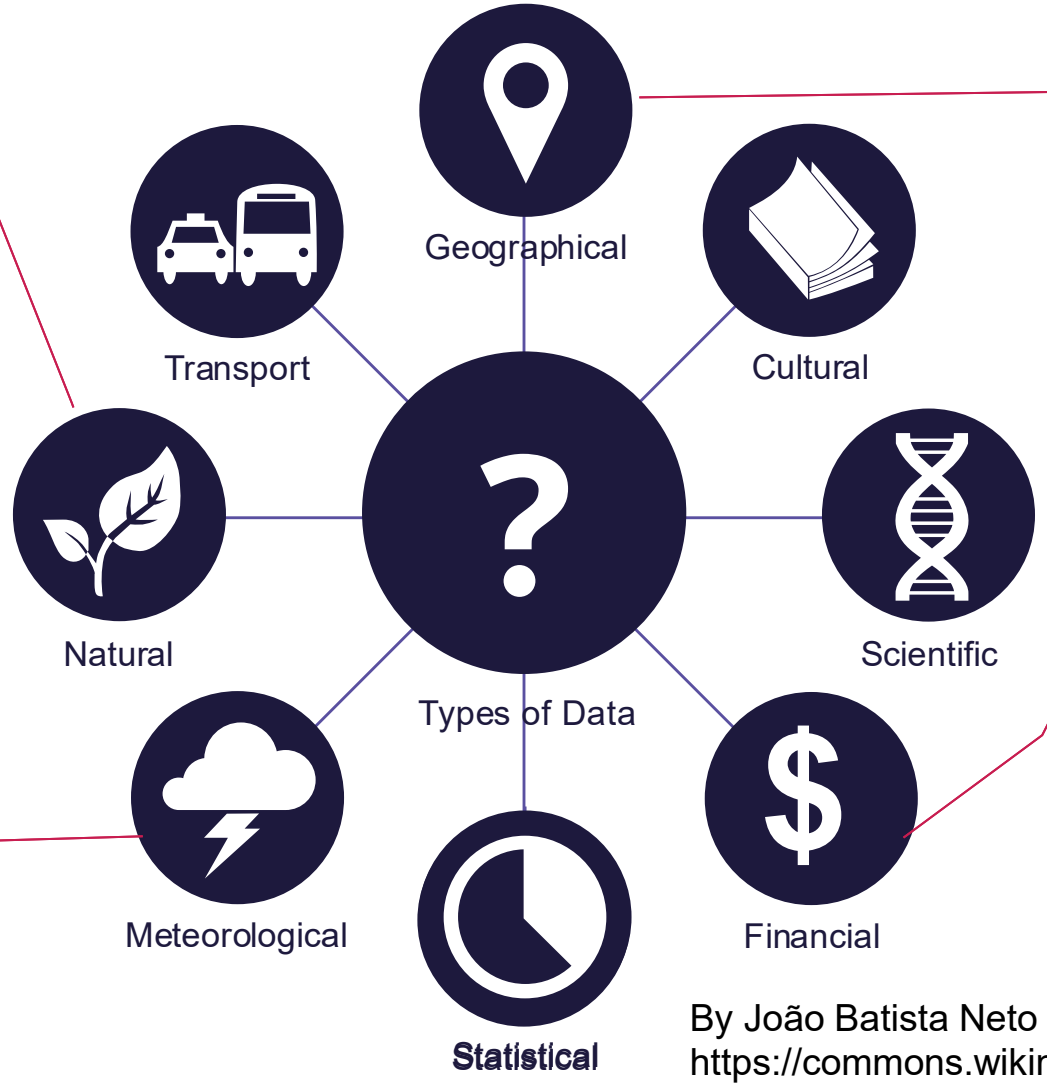
What is data?

"Data is a **collection of facts**, numbers, words, observations or other useful **information**. Through **data processing** and **data analysis**, organizations **transform raw data points into valuable insights** that **improve decision-making** and drive better business outcomes."

A definition by IBM

Types of Data Sources: some examples

Development of bio-indicators since 1980, use for polluter screening, growing mapping systems merely local in use.



Earliest forms of mapping known from (6.000 BC) Systematic mapping since 18th century, today GPS satelite data

First accounting system 3.000 BC (Mesopotania) Today International Financial Reporting Standards (IFRS) and National Rules (e.g. SWISS GAAP FER)

Since 1781 (Societas Metrologica Palatina); today global network AWEKAS (real time satelite data)

By João Batista Neto - Data types - pt br.svg, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=43063497

Public Data (Statistics)

UNIVERSITÄT DES SAARLANDES

Characteristics

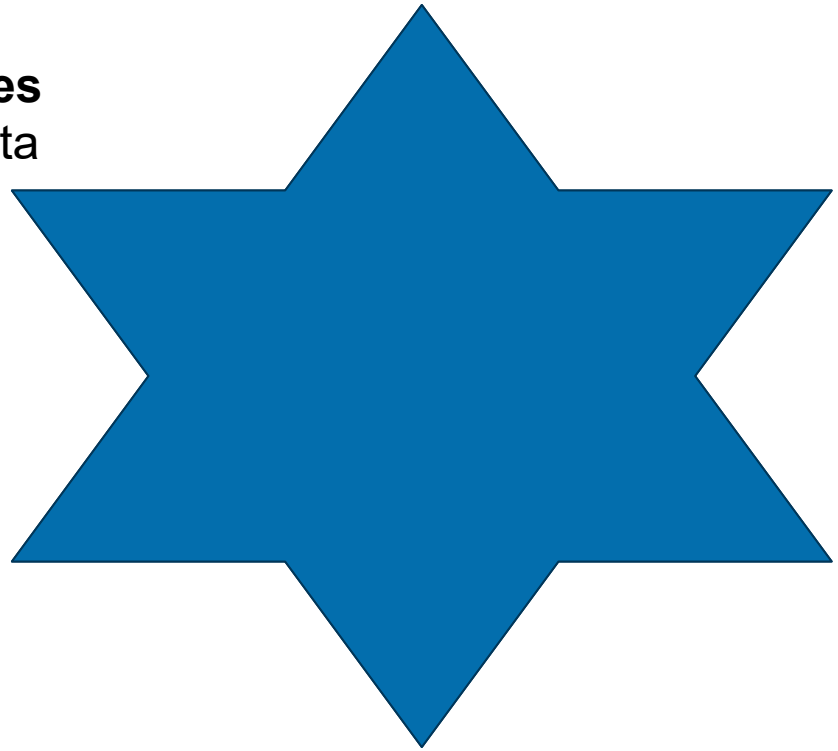
- Data provided by **public agencies** (statistical offices)
- Data collected on base of **laws** and **precise definitions**
- **Quality control** and **standardized data processing**
- Long time-series and **comparability**
- Regional and social **differentiation** possible
- **Free access** to data (in some cases even to raw data)

Databases

Access to **Raw Data**

Access to **NoSQL-Databases** (processing unstructured data – “Big Data”) Access to single data structures

Retrieval in **Relational Databases (SQL)** Access to tables



Time-series Databases (Apache) Access to pairs of time and values

Object-oriented Databases (ODMS) Access to **objects**

Hierarchical Database (DBMS, XML) Access to a tree-like model

Process-produced data

"information generated automatically as a **byproduct of ongoing, real-world activities**, such as machine operations or administrative processes, **rather than being collected** through direct surveys or observations".

A definition provided by KI

Internet Data („Big Data“)

“Big data primarily refers to **data sets that are too large or complex** to be dealt with by traditional data-processing software. Data with many entries (rows) offer **greater statistical power**, while data with higher complexity (more attributes or columns) may lead to a **higher false discovery rate**”.

Wikipedia

Access Options

Complete Data Sets = The original (raw) data is available and can be used without any restrictions

Limited Data Use = Only parts of the whole data set is accessible (e.g. because of data privacy)

Aggregated Data = No individual data sets are available, only aggregated/grouped data

Principles of Systematic Research

What should be searched for?

- Define the research principle
- Define the research components

Where should be searched?

- Define the databases
- Define keywords and search strings

How should be searched?

- Define the retrieval process

How can be ensured that nothing is overlooked?

- Define complementing research strategies

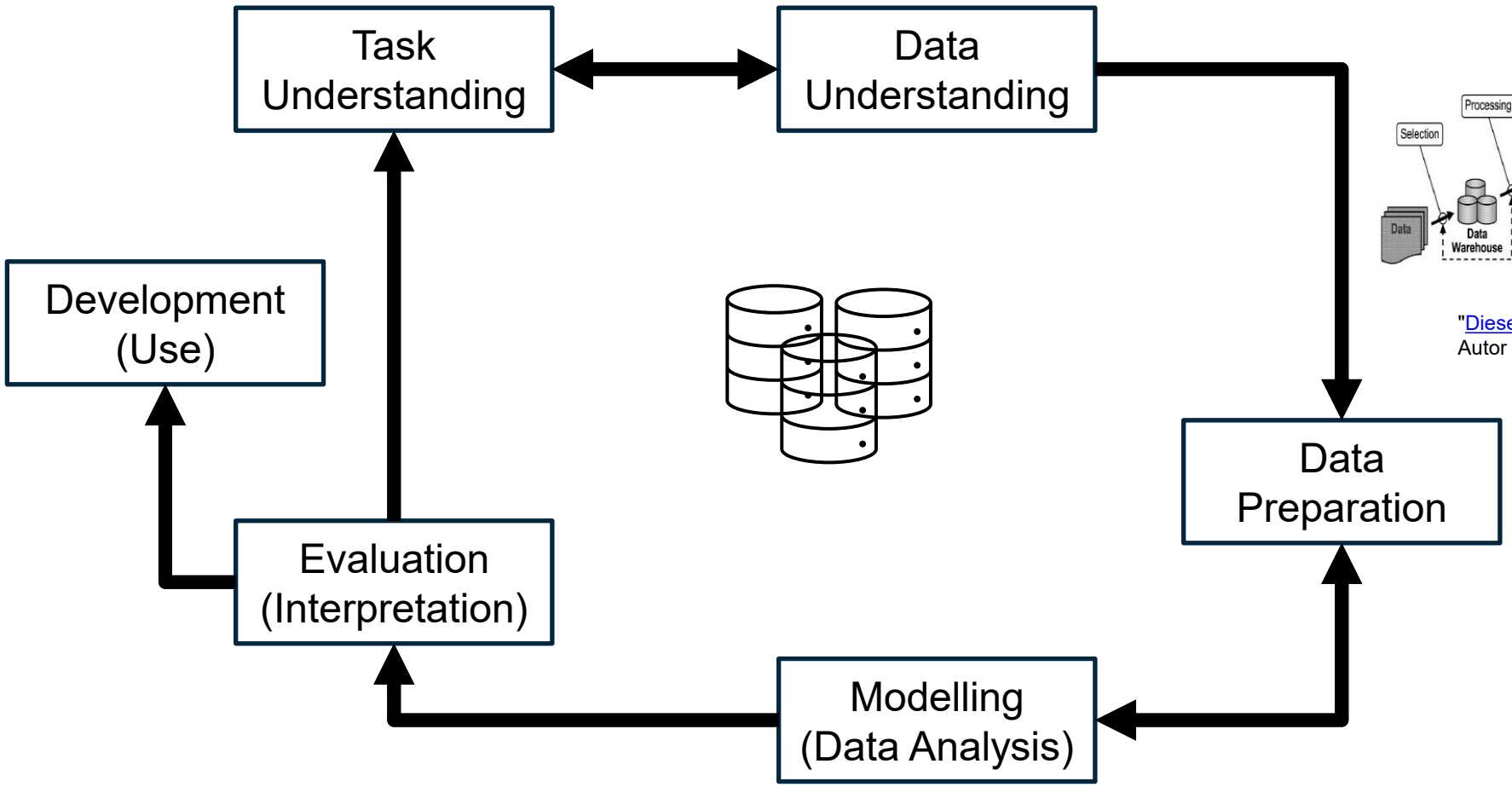
Data-mining (Definition)

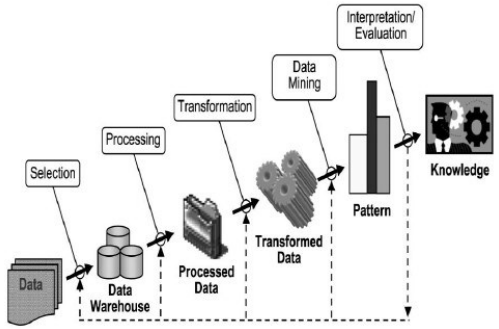
"Data mining is the **process of extracting meaningful patterns** … from large volumes of data using techniques like statistical analysis and machine learning.

It involves **analyzing large datasets to discover hidden patterns** and trends, which can then be **used to improve decision-making** and predict future outcomes.

Wikipedia

Process of Data-mining (CRISP-Model)





Cleve, J. & Lämmel, U. (2016). Data Mining. Berlin/Boston: de Gruyter, S. 7

EIDGENÖSSISCHE FINANZKONTROLLE
CONTRÔLE FÉDÉRAL DES FINANCES
CONTROLLO FEDERALE DELLE FINANZE
CONTROLLA FEDERALA DA FINANZAS
SWISS FEDERAL AUDIT OFFICE

Making a Difference Through Data

WGEPPP Forum 2025

Dr. Roger Pfiffner
Bern, 17.10.2025

Planning a Data Analysis in Auditing

Agenda

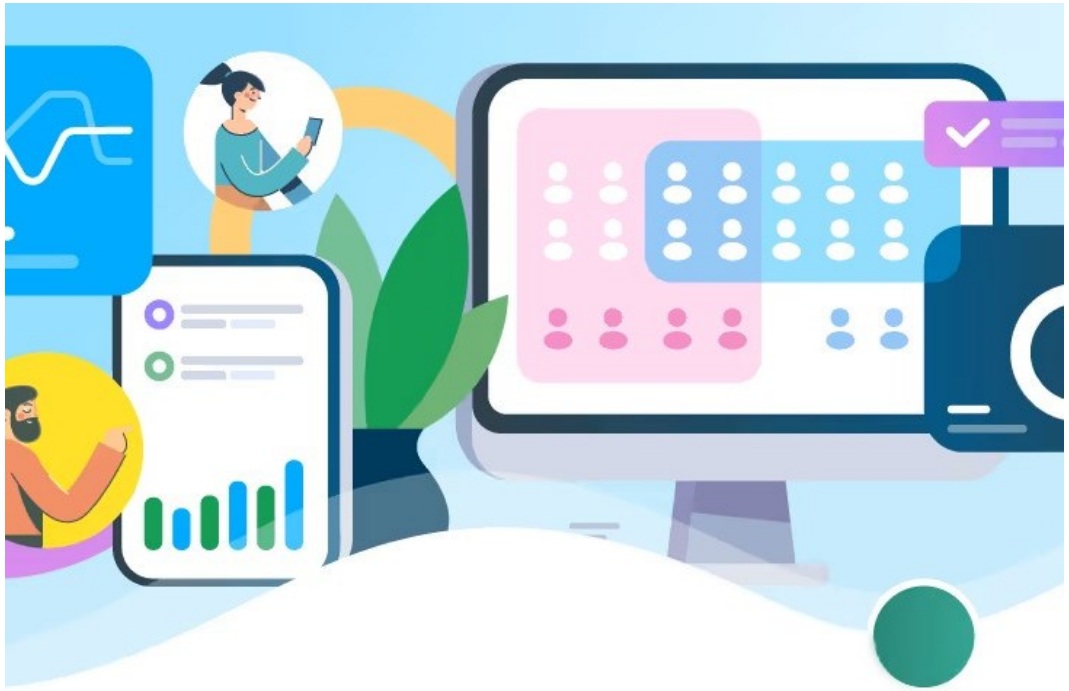
1. The Benefits of Data Analysis for SAI
2. Planning Data Analyses in Your Audit

Roger Pfiffner, PhD
Performance Auditor @ Swiss Federal Audit Office

Please connect with me on LinkedIn 😊

Why SAI are increasingly performing data analyses

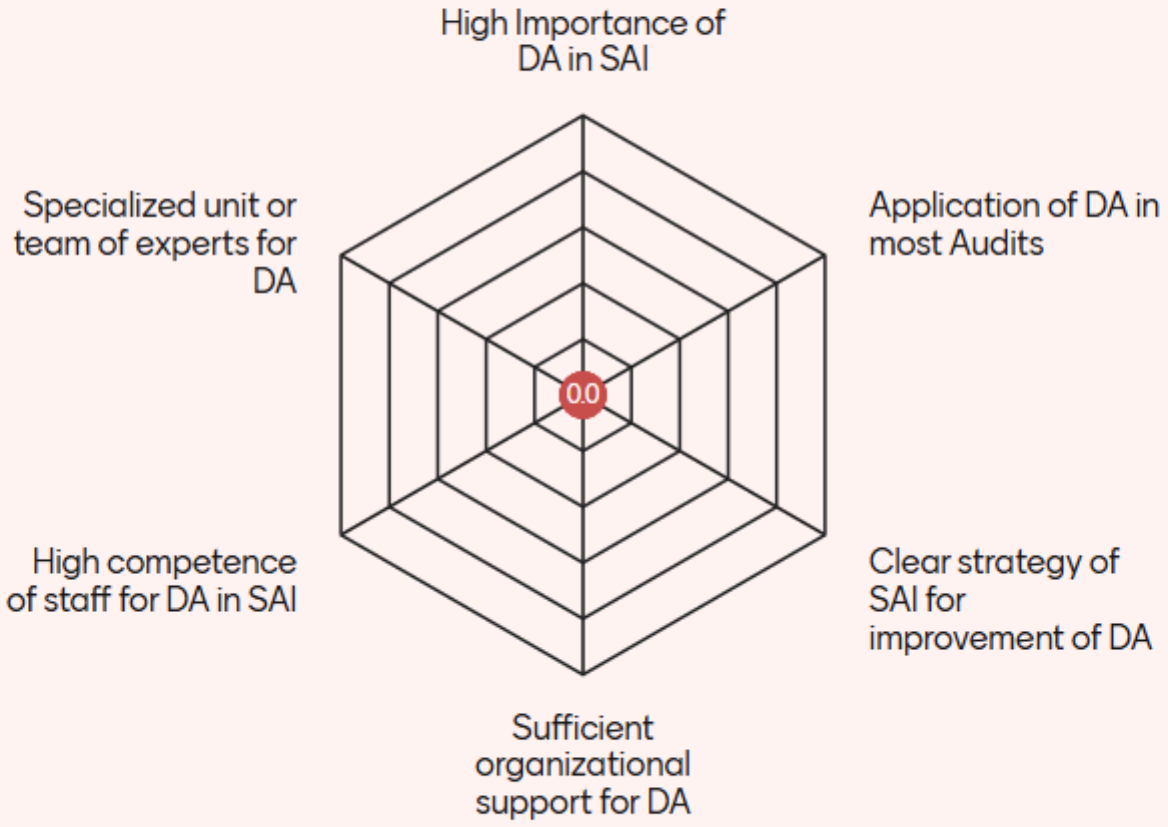
- **Data-based and quantified results** are more objective, better traceable and provable
- More **comprehensive audit basis** (Examination of entire populations instead of random samples)
- Better **risk identification** (Detection of anomalies, irregularities and patterns that indicate risks, errors or misuse)
- **Digitalization** and Innovation (e.g. through AI technologies)



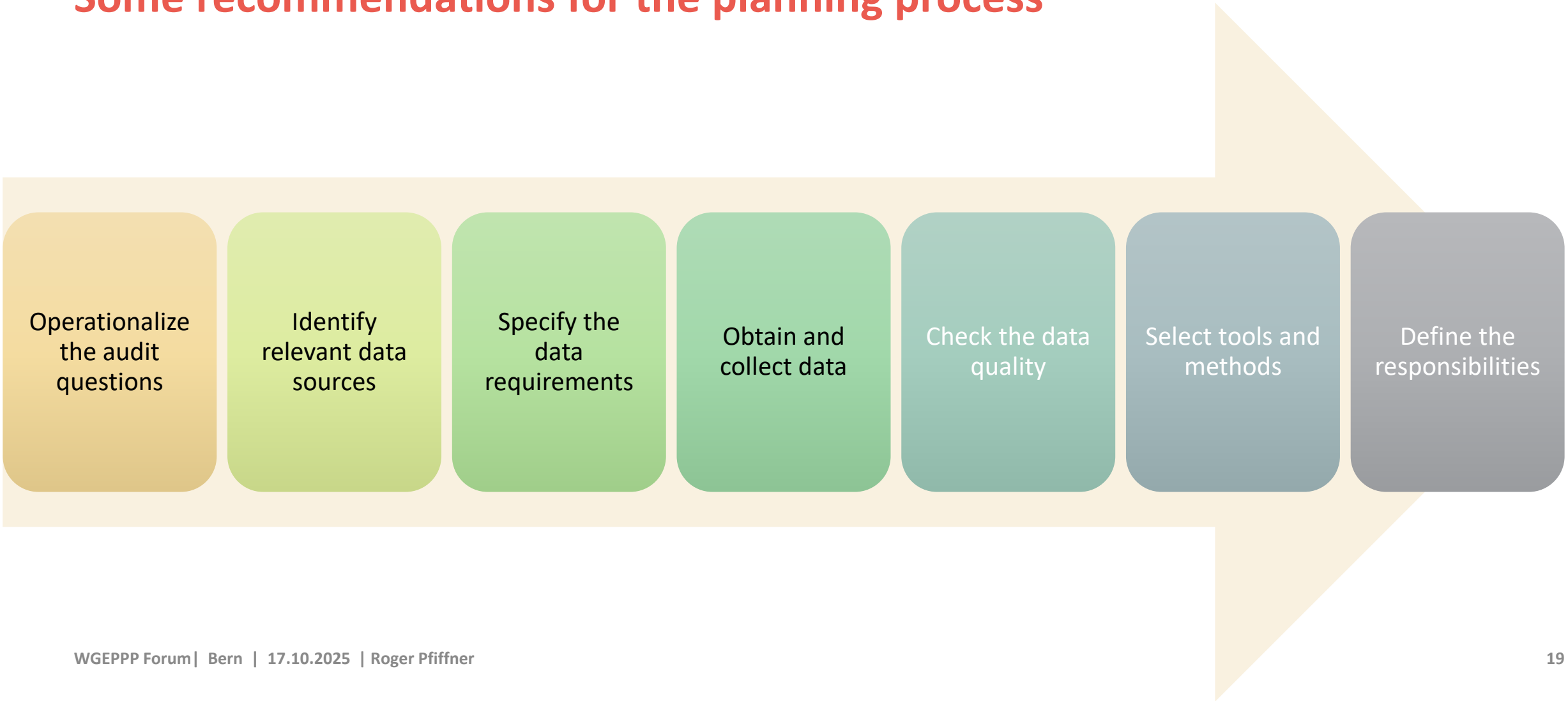
Join at menti.com | use code 4200 6951

Mentimeter

Please rate the following aspects concerning data analysis (DA) in your SAI.



Some recommendations for the planning process



Operationalize the audit question

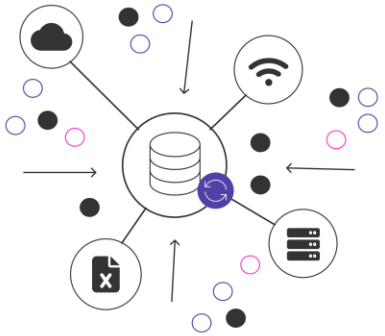
Broad questions (such as *Is the subsidy program for renewable energy projects being implemented appropriately and is it effective?*) lack clarity and measurability.

That's why we must …

1. **Clarify the audit objectives** (improve productivity, reduce costs, improve impact?)
2. **Narrow the Focus of the audit question** by focusing on specific projects, processes, metrics or goals to enable targeted and feasible analysis
    - Example: To what extent do the funded renewable energy projects achieve their planned CO2 reduction targets within the first three years of implementation?
3. Critical review of validity, interpretability, relevance (risk orientation) and stakeholder expectations

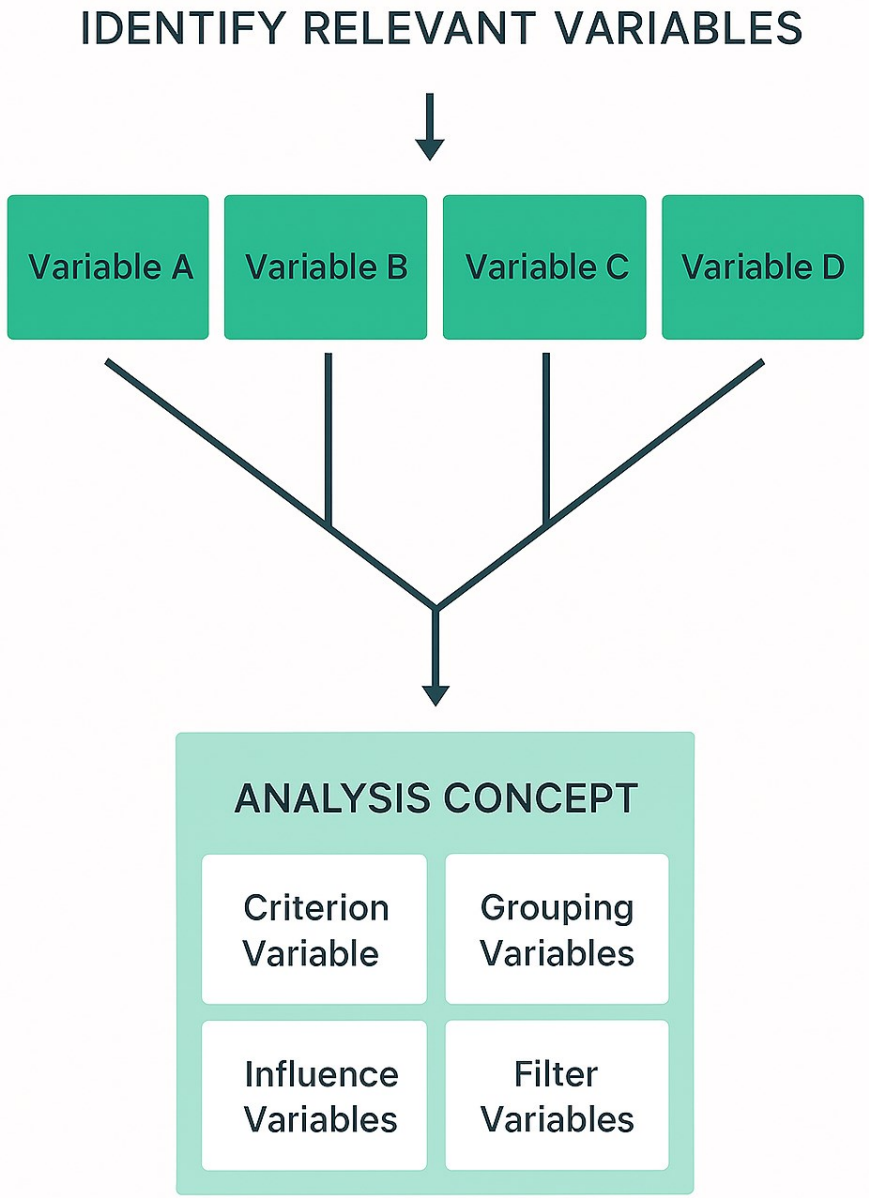
Identify relevant data sources

- Identification of existing data sources, in particular register data, administrative data, **geospatial information** or external data such as market and industry information.
  - Inclusion of **IT system log files**?
  - Inclusion of **unstructured data** (text documents, social media posts, emails, photos, videos)?
- Request data documentation if available
- Clarification of accessibility: How easy is it to obtain data, what rights and technical requirements are there?
- Conduct your own survey if necessary



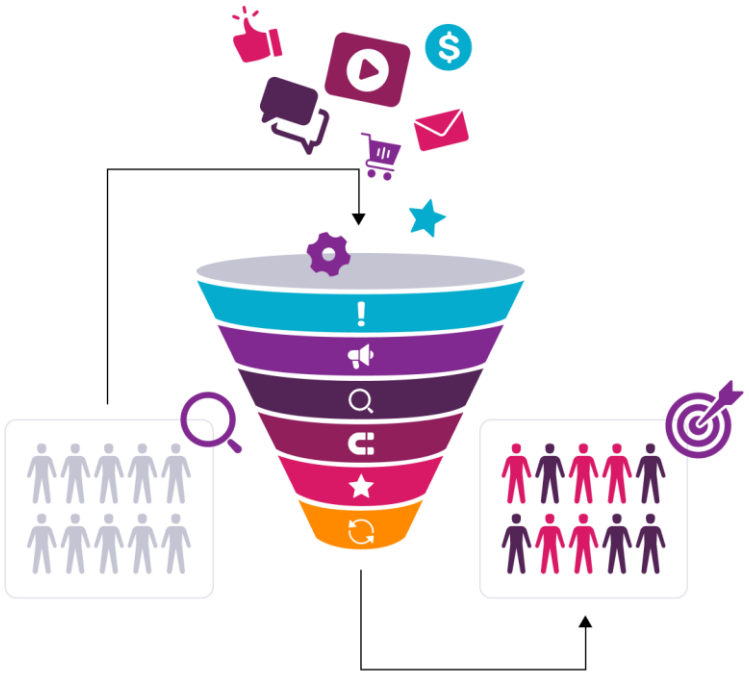
Specify data requirements

- → Identify relevant variables
- → Break down the variables into roles (analysis concept)
- → Define the observation period
- → Decide on the appropriate data granularity: individual raw data vs. aggregated data/key figures
- → Identifier for merging (and pseudonymisation)?
- → Determine available data formats (e.g. Excel, CSV, JSON, TXT, SQL)?



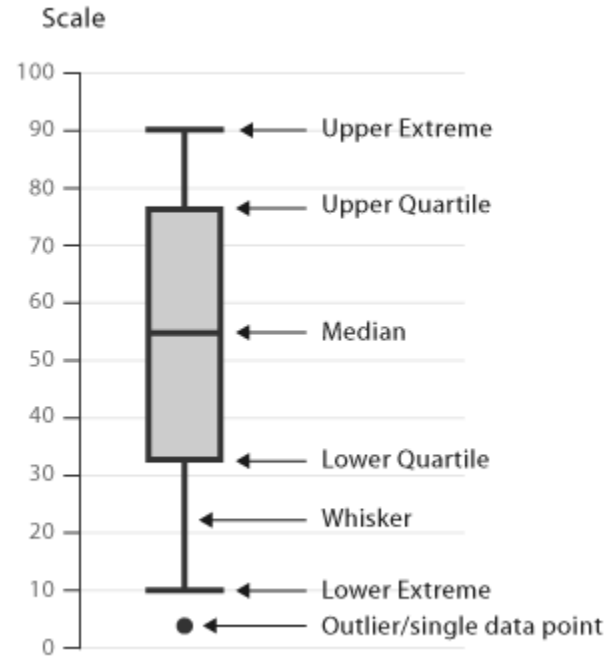
Obtain and collect data

- Discuss your analysis concept with the data owner or the auditee
- Ensure the data owner's willingness to collaborate (adequate time and scope)
- Agree on a schedule for delivery (and plan the following steps such as preparation and analysis)
- Ensure data protection and data security issues (through data minimization, anonymization, access rights in audit team, data storage, data protection agreement)



Check data quality

- Data exploration: Check the number of observations, distribution, missing values, outliers and inconsistencies (e.g. with summary statistics, box plots, etc.)
- Check for duplicates
- Plausibility checks: Do the case numbers, sums and averages match existing or published analyses? Are the classification codes complete?
- Selection of test objects and testing/comparison of the selected cases.




```
> describe(data)
                          vars   n   mean     sd median trimmed    mad  min    max  range  skew kurtosis   se
Survived                     1 887   0.39   0.49   0.00    0.36   0.00 0.00   1.00   1.00  0.47    -1.78 0.02
Pclass                       2 887   2.31   0.84   3.00    2.38   0.00 1.00   3.00   2.00 -0.62    -1.29 0.03
Name*                        3 887 444.00 256.20 444.00  444.00 329.14 1.00 887.00 886.00  0.00    -1.20 8.60
Sex*                         4 887   1.65   0.48   2.00    1.68   0.00 1.00   2.00   1.00 -0.61    -1.63 0.02
Age                          5 887  29.47  14.12  28.00   28.96  11.86 0.42  80.00  79.58  0.45     0.28 0.47
Siblings.Spouses.Aboard      6 887   0.53   1.10   0.00    0.27   0.00 0.00   8.00   8.00  3.67    17.64 0.04
Parents.Children.Aboard      7 887   0.38   0.81   0.00    0.19   0.00 0.00   6.00   6.00  2.73     9.63 0.03
Fare                         8 887  32.31  49.78  14.45   21.50  10.32 0.00 512.33 512.33  4.76    32.99 1.67
```

Select tools and methods

- Define the analysis tools (e.g. Excel, R, Python, MAXQDA)
- Consider their advantages and disadvantages, e.g. in terms of
  - Verifiability
  - Reproducibility & Versioning
  - Teamwork
  - Flexibility
- Determine which methods and analyses are planned (transformation, linking, descriptive statistics, data mining, cluster analysis, correlation analyses, test for significance, etc.)

Define the responsibilities

Important points to be clarified:

- → Who is responsible for the analysis? Is a data specialist available?
- → When will the data analyst be involved? Only during the analysis?
- → How do audit manager and data specialist work together?
- → Who is responsible for the quality assurance?

Questions?

UNIVERSITÄT DES SAARLANDES

Appropriate assessing, analysing and visualising your results Part 3: Data Management

Wolfgang Meyer, Saarland University

Introduction to the Course

Data Mining
- Identifying and extracting relevant data from sources

Data Processing
- Sizeable effort in preparation for analysis

Data Analysis
- Using methods to interpret data and develop insights

Data Visualization
- Communicating results including visualization and data storytelling

Chapter 1

1.1 Data Sources
1.2 Systematic Research
1.3 Data Mining

Chapter 2

1.1 Evaluating Catalogue
1.2 Assessing Data Quality
1.3 Errors and Missings

Chapter 3

1.1 Data Management
1.2 Machine Learning

Chapter 4

1.1 Descriptive
1.2 Clustering
1.3 Combining
1.4 Longitudinal

Chapter 5

1.1 Basic Graphs
1.2 Advanced Visualization

Source: Hyman et al (2024): Data Analytics and Visualization. Hoboken: John Wiley & Sons, p. 13 (modified)

Content Chapter 3

- Comparison
  - Principles
  - Preparing Data for Comparison
  - Variables and Scales
  - Forms of Comparison
- Clustering
  - Principles
  - Forms of Clustering
- Qualitative Cluster Analysis
  - Principles
  - Developing Types and Classification Theory

Chapter 3

1.1 Data Management
1.2 Machine Learning

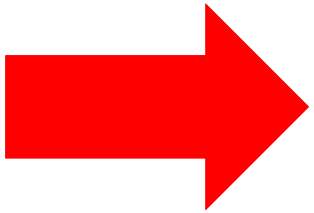
Why comparing?

UNIVERSITÄT DES SAARLANDES

An Example

"One has to be careful not to compare apples and oranges"

A British idiom, referring to the differences between items which are popularly thought to be incomparable or incommensurable (Wikepedia)



In fact, comparing apples and oranges is easy – everything can be compared, it is a question of criteria

Why comparing?

Making things comparable means:

Making things measurable !

Measuring = comparing objects with standardized scales

How to compare incomparable things (an example)

Task: comparing Apple and Oranges

Requirements:

assumptions on unique characteristics of apples

and oranges

First step: theory on variations

How to compare incomparable things (an example)

Task: Defining a variable with at least two categories

Assumption:

The color of oranges is orange, apples are green,

red, yellow – but not orange

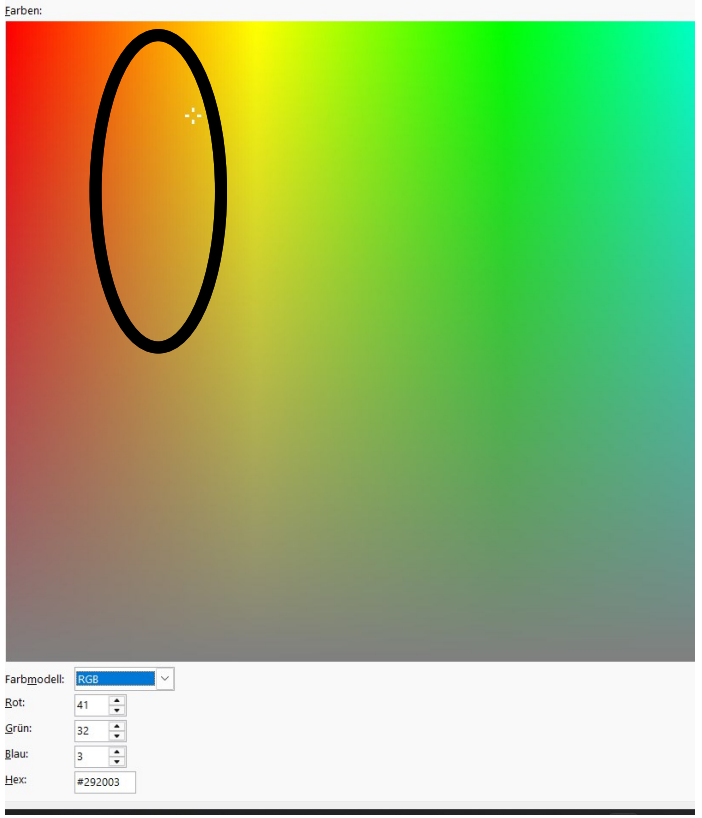
Second step: defining a variable and a scale

How to compare incomparable things (an example)

Task: Operationalization

Requirement:

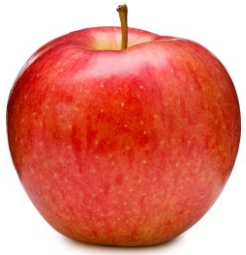
Operational definition of “color”



Third step: operationalization of the variable

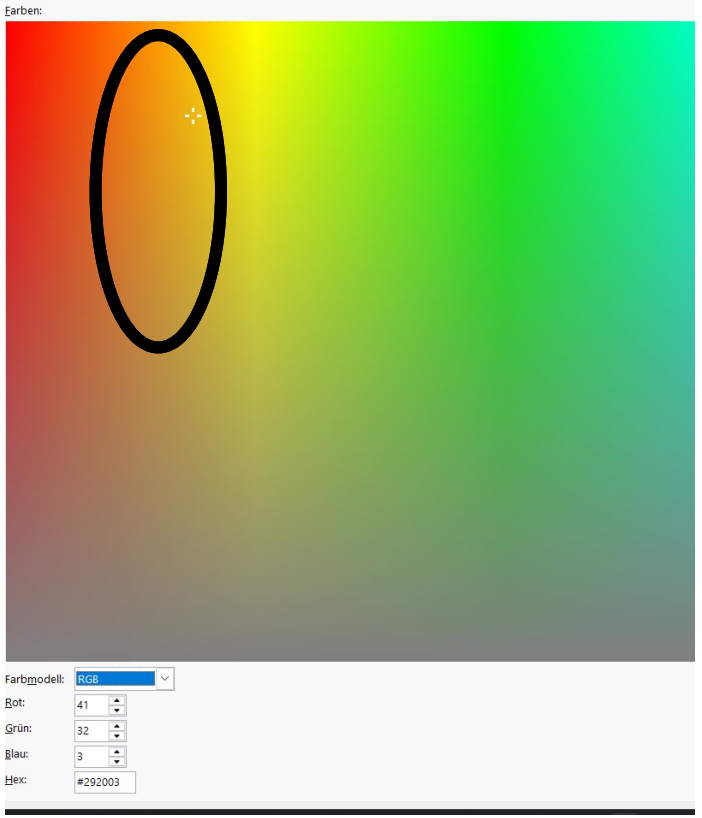
How to compare incomparable things (an example)

Task: Measurement



Procedure:

Comparing the object values with the scale



Fourth step: comparing elements with scale

How to compare incomparable things (Summary)

1. Theory on variations of objects (what are the differences)
2. Defining variables and scales (based on this assumption)
3. Operationalization of the variables (how to compare)
4. Comparing elements with the scale (sorting)

What we need:
A Theory
A Method
A Target

What can we do if we do not know about categories?

The **task** is again to sort elements,

but we do not know about the number of **categories**

The **theory** is still the same: the elements in one group should share similar characteristics (e.g the color orange) and the elements in the other groups should be different to this.

The **method** is called “clustering”: sorting elements AND defining groups (clusters) is now part of the analysis

How to cluster elements (an example)

Task: comparing Fruits

Requirements:

assumptions on unique characteristics of fruits

First step: theory on variations (as before)

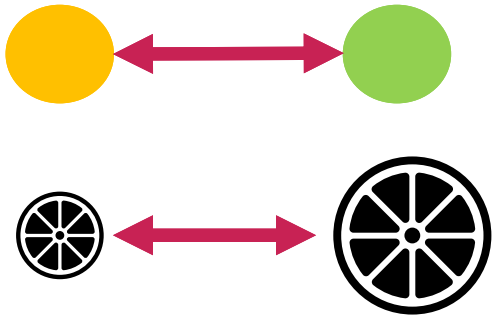
How to cluster elements (an example)

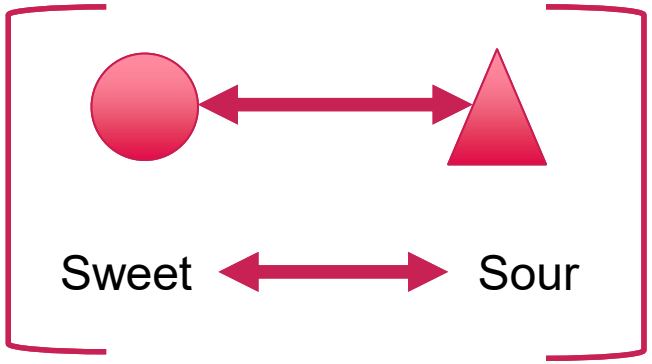
Task: Defining variables

Assumption:

Each variable differentiates elements

on one dimension





Second step: defining dimensions, variables and scales

How to cluster elements (an example)

Task: Operationalization and Standardization of scales

Requirement:

Z-Transformation $$Z = \frac{X - \mu}{\sigma}.$$

Third step: standardization of scales

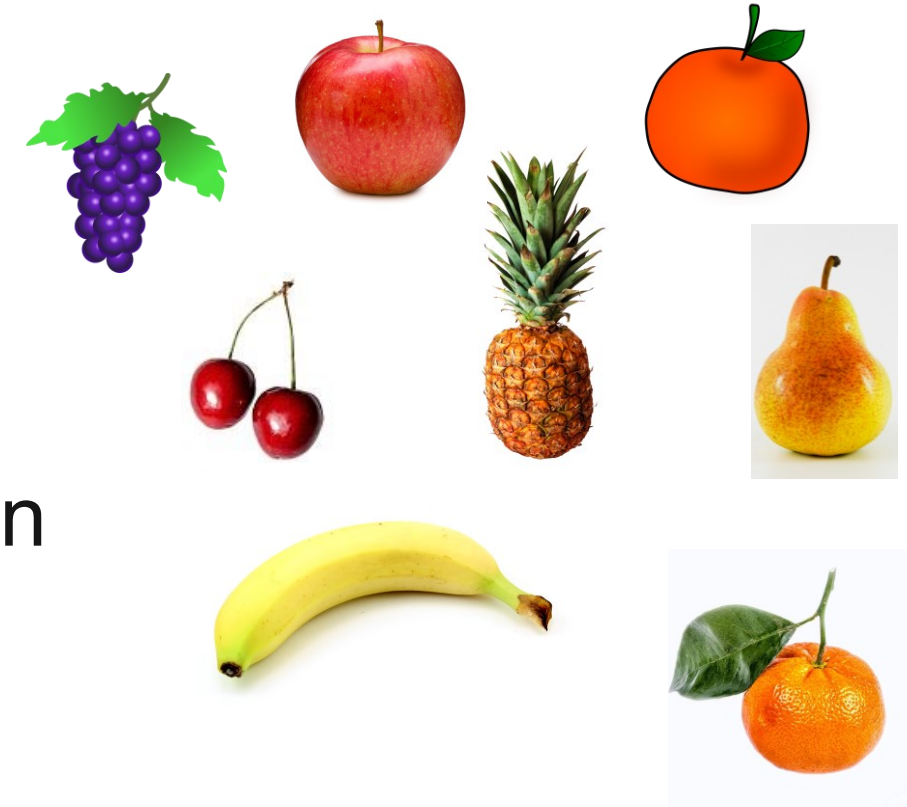
How to cluster elements (an example)

Task: Clustering

Procedure:

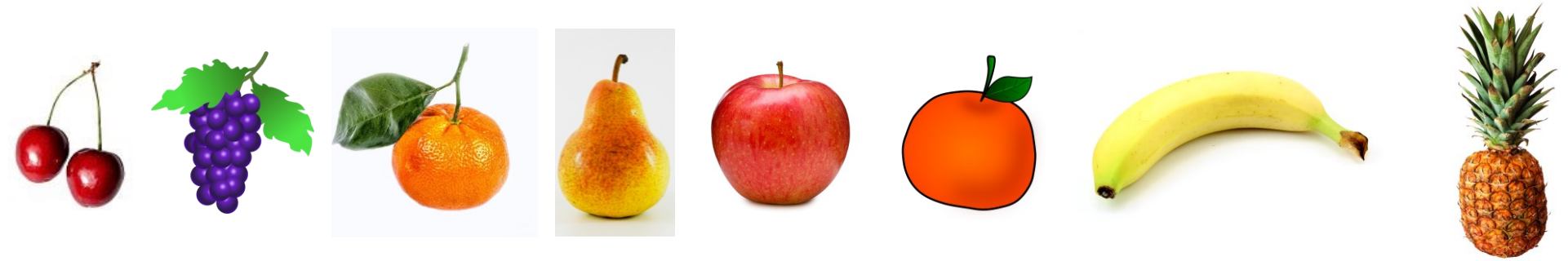
Comparing the distances between

elements on various scales



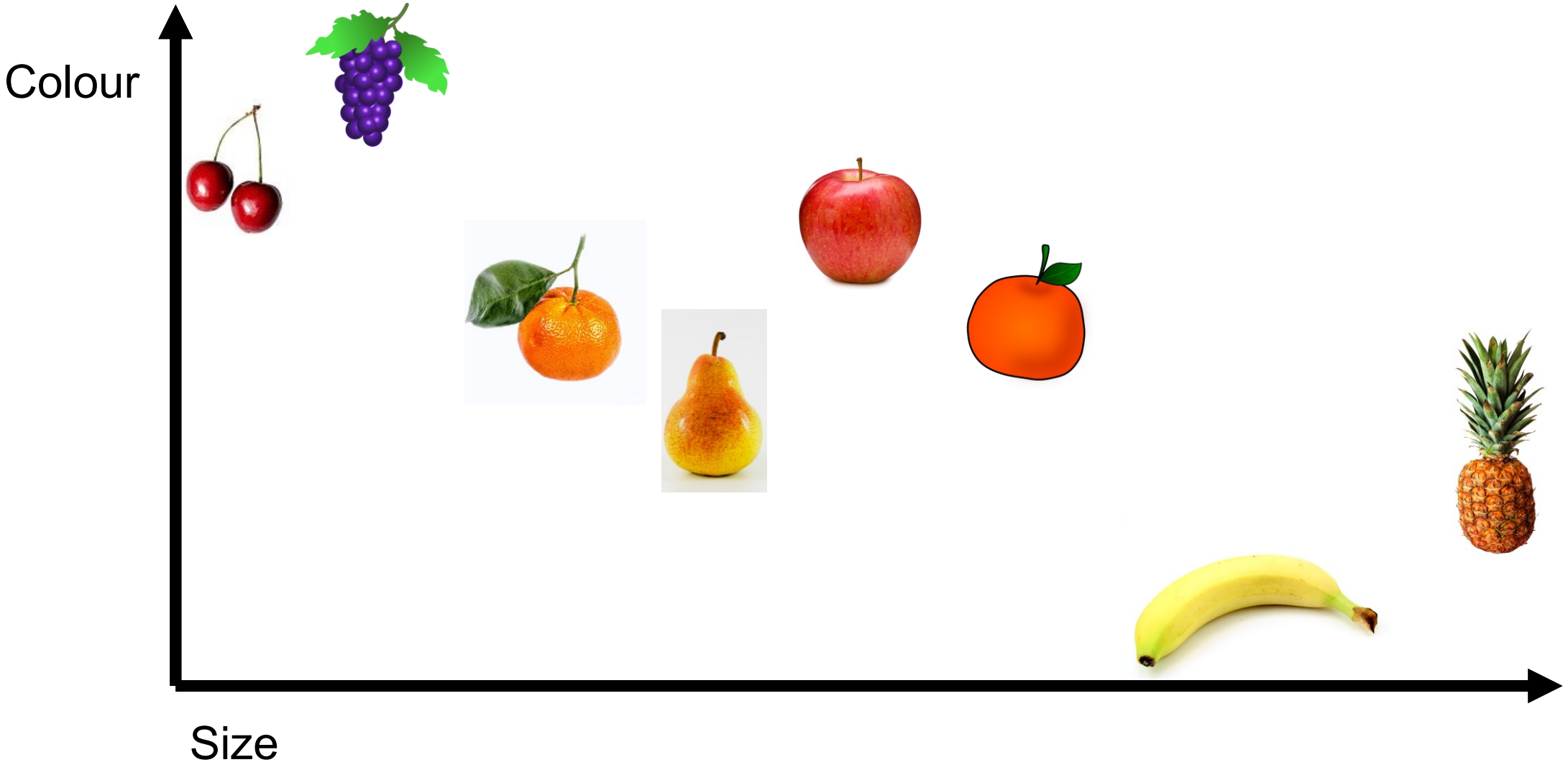
Fourth step: multidimension comparison of elements

How to cluster elements (an example)

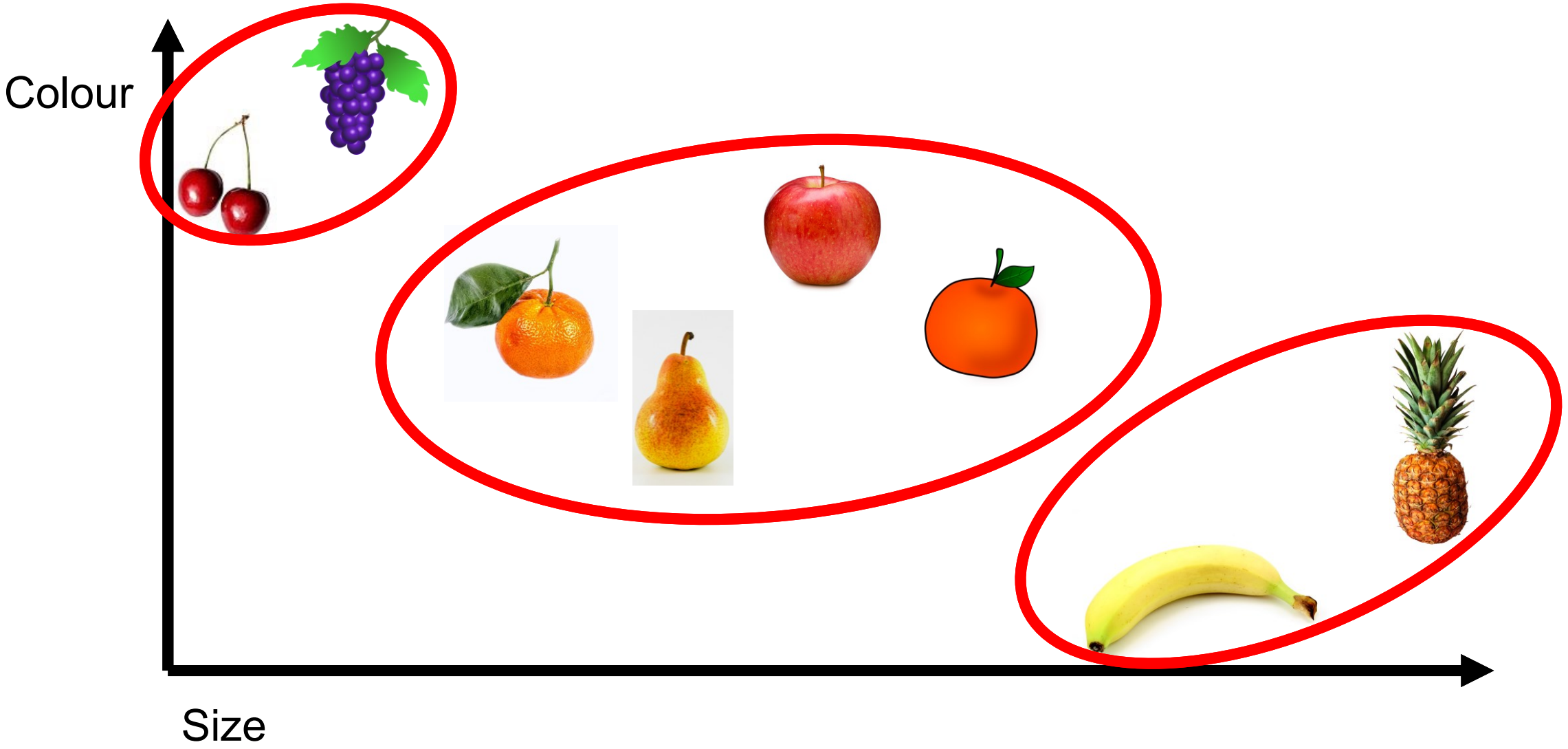


Size

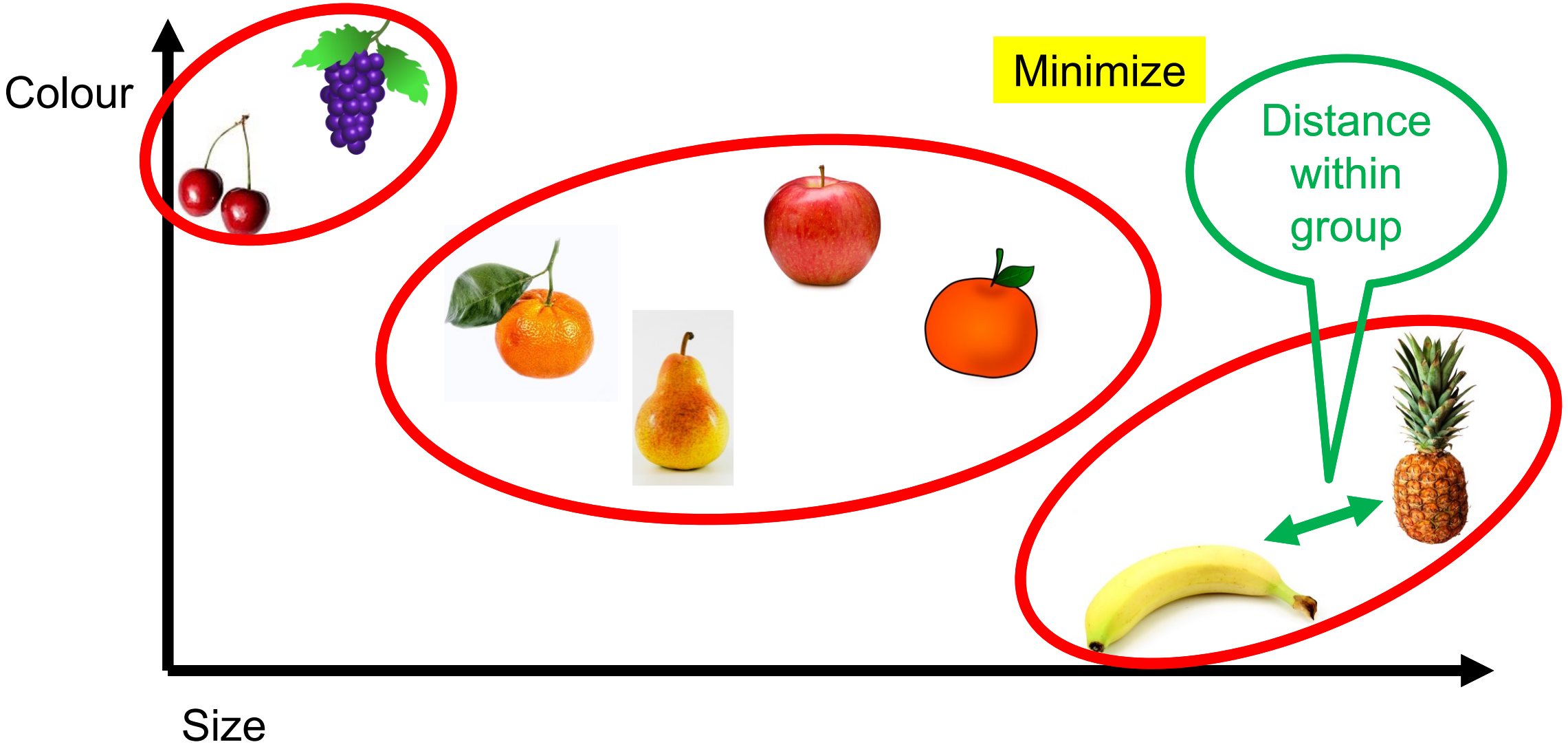
How to cluster elements (an example)



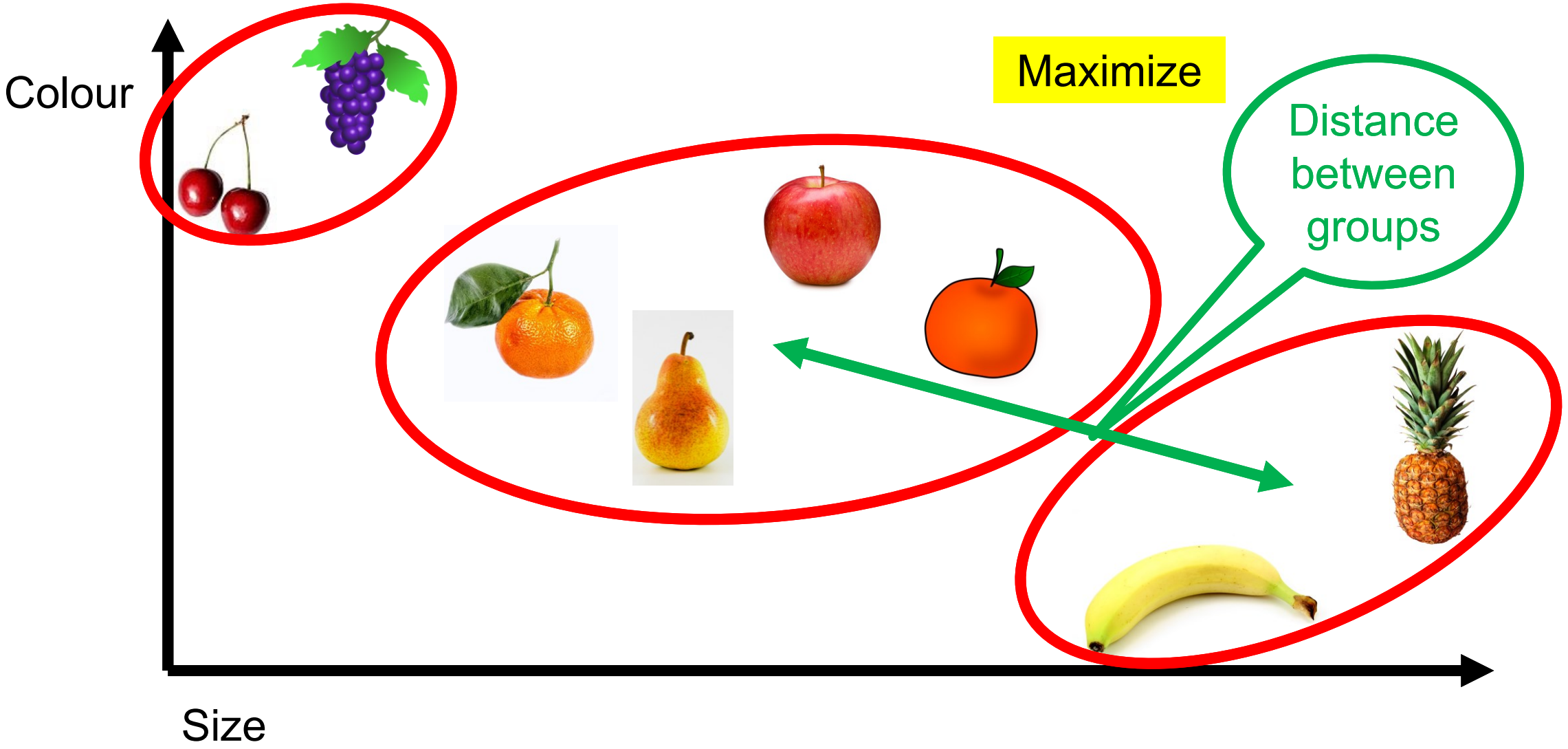
How to cluster elements (an example)



How to cluster elements (an example)



How to cluster elements (an example)



How to cluster elements (Summary)

1. Theory on variations of objects (what are the differences)
2. Defining dimensions, variables and scales
3. Standardization of Scales (z-score)
4. Multidimensional comparison of elements (sorting)
5. Grouping (building clusters)

What we need:
A Theory
A Method

What is the difference between Cluster and Factor Analysis?

UNIVERSITÄT DES SAARLANDES

Cluster Analysis

- Grouping of Cases
- Finding Similarities and Differences of Cases
- Building Types
- Correlation of Cases

Factor Analysis

- Grouping of Variables
- Finding Latent Structures of Variables
- Building Factors
- Correlation of Variables

What can we do if we do not have a theory?

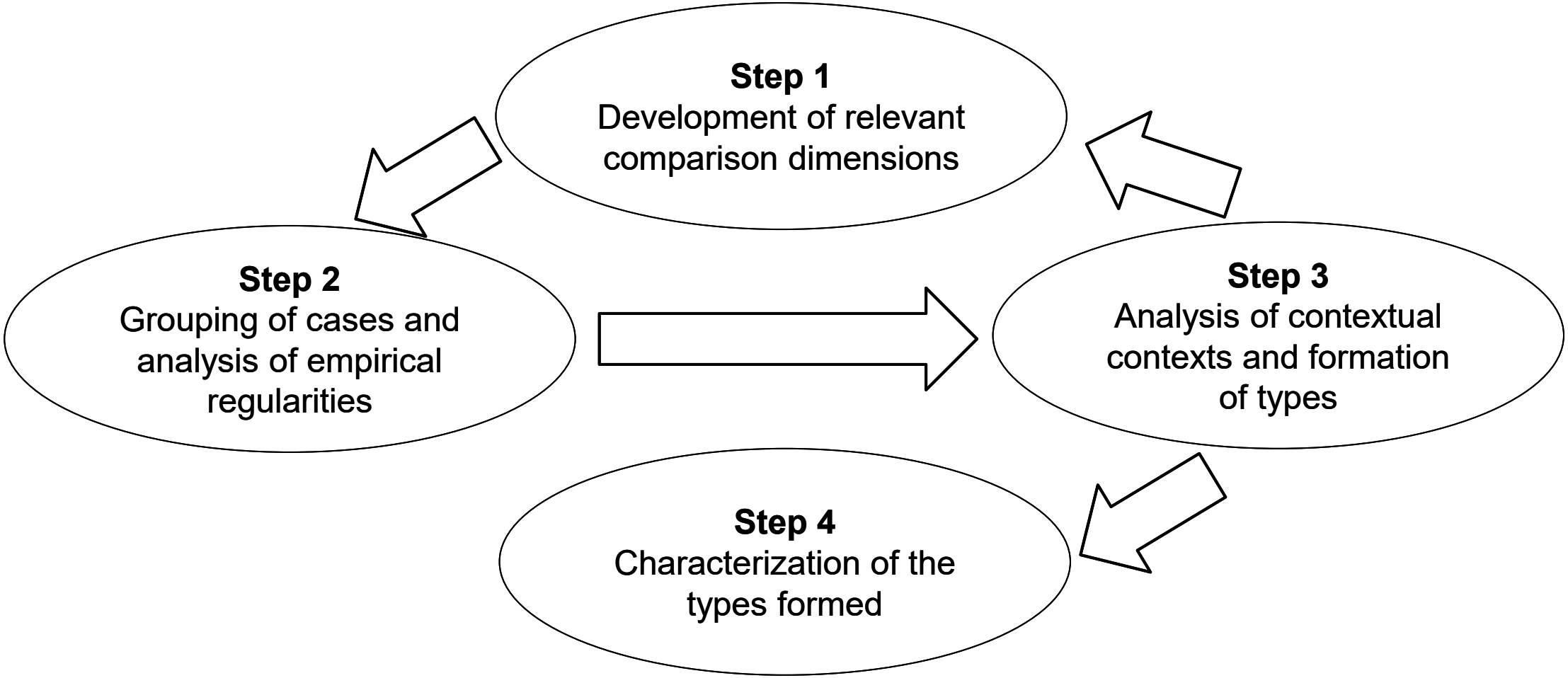
The **task** is again to sort elements,

but we do not know the **variables**

We have to develop a **theory** while doing the analysis: this **inductive approach** is an **iterative process**.

The **method** is called “qualitative clustering”: sorting elements defining groups (clusters) AND developing a theory about variables are now parts of the analysis

How to do qualitative clustering



Kluge, S. (1999): Empirisch begründete Typenbildung. Zur Konstruktion von Typen und Typologien in der qualitativen Sozialforschung. Opladen: Leske+Budrich, p. 261.

What can we do with „clusters“?

Using clusters for typologies (descriptive)

- Characterizing Types (specifics)
- Identifying specific Opportunities and Risks*

Using clusters for comparisons (comparing clusters)

- Comparing extreme groups
- Comparing with average
- Comparing with best practice (benchmarking)

Using clusters for comparisons (analysing development)

- Before – After comparison
- Changing of Types
- Changing of Typologies

Pros and Cons of Qualitative Clustering

UNIVERSITÄT DES SAARLANDES

Pros

- Useful for small numbers of cases
- Identifies the Meaning of Types
- Understanding the variables
- Can be linked to quantitative analysis

Cons

- Not Usefull for big number of cases
- No precise Measurement
- No standardization of variables
- Can be biased by personal views

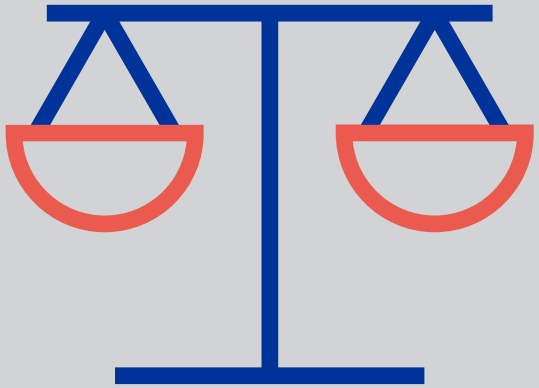
EIDGENÖSSISCHE FINANZKONTROLLE
CONTRÔLE FÉDÉRAL DES FINANCES
CONTROLLO FEDERALE DELLE FINANZE
CONTROLLA FEDERALA DA FINANZAS
SWISS FEDERAL AUDIT OFFICE

Making a Difference Through Data

WGEPPP Forum 2025

Dr. Roger Pfiffner

Bern, 17.10.2025



Detection of Clusters and Outliers Using Multivariate Methods

Agenda

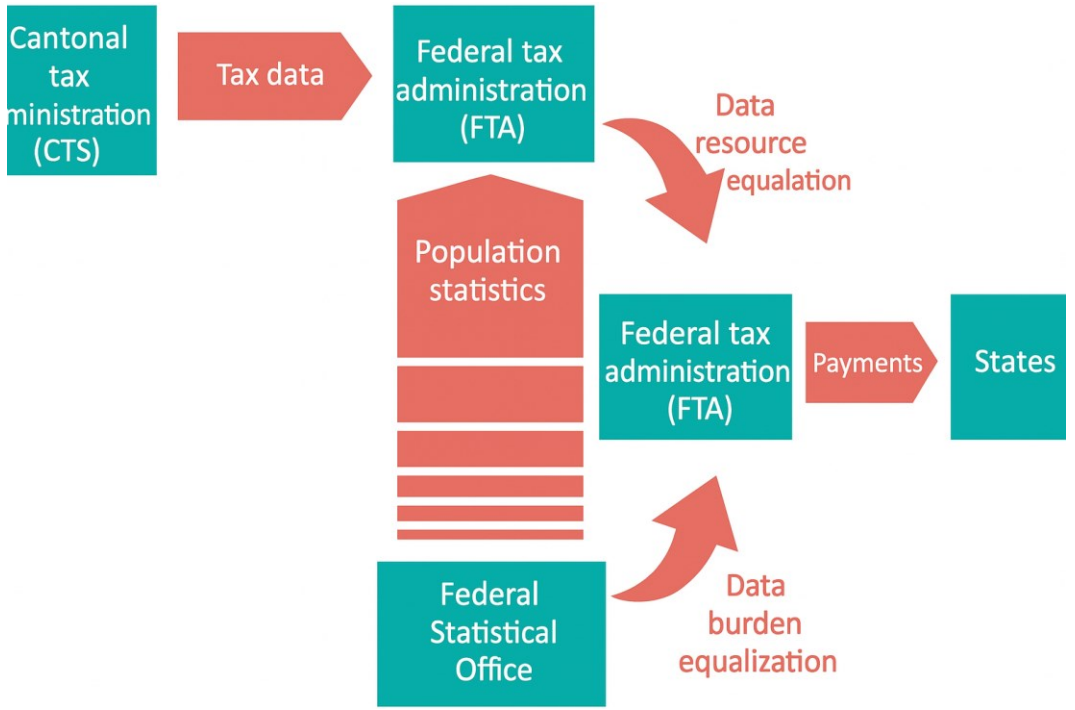
1. Starting Point and Audit
2. Anomaly Detection Using a Machine Learning-based Method
3. Data Compression with PCA, Clustering and Visualization of Groups
4. Conclusion

Roger Pfiffner, PhD
Performance Auditor @ Swiss Federal Audit Office

Please connect with me

Starting Point

- **The National Fiscal Equalization (NFA):** System in Switzerland for redistributing financial resources between the cantons. Economically stronger cantons (contributors) transfer funds to financially weaker cantons (recipients).
- **Objectives**: Reduction of cantonal disparities in financial capacity and ensuring the uniform delivery of public services across the country.
- **Calculations**: based on tax data (taxable resources per capita) from the cantons.

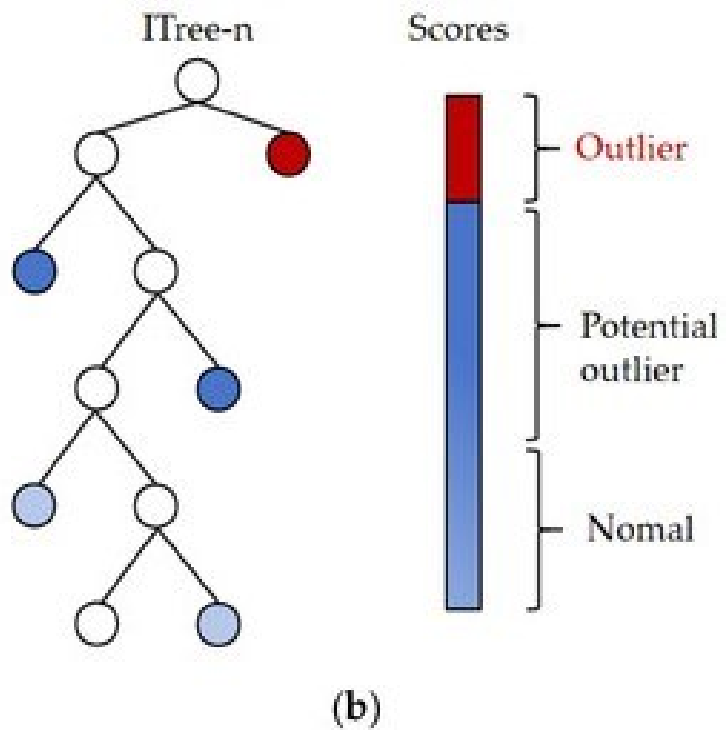


The Audit

- **Task** of the Swiss Federal Audit Office:
  - Assessment of the redistribution calculations
  - Verification of the data provided by the cantons. Sample-based assessment of data delivery and processing using a risk-based approach
  - All 26 cantons will be inspected on site at least once in four years
- **Challenges** for data verification:
  - Large amount of data (many observations and variables)
  - Diversity of the population, ensuring the representativeness of tests
  - Definition of correct risk criteria (if no hypotheses about complex relationships are available)
- **Objective** of the analysis: selection of cases for detailed examination

Detection of Anomalies Using the Isolation Forest Algorithm

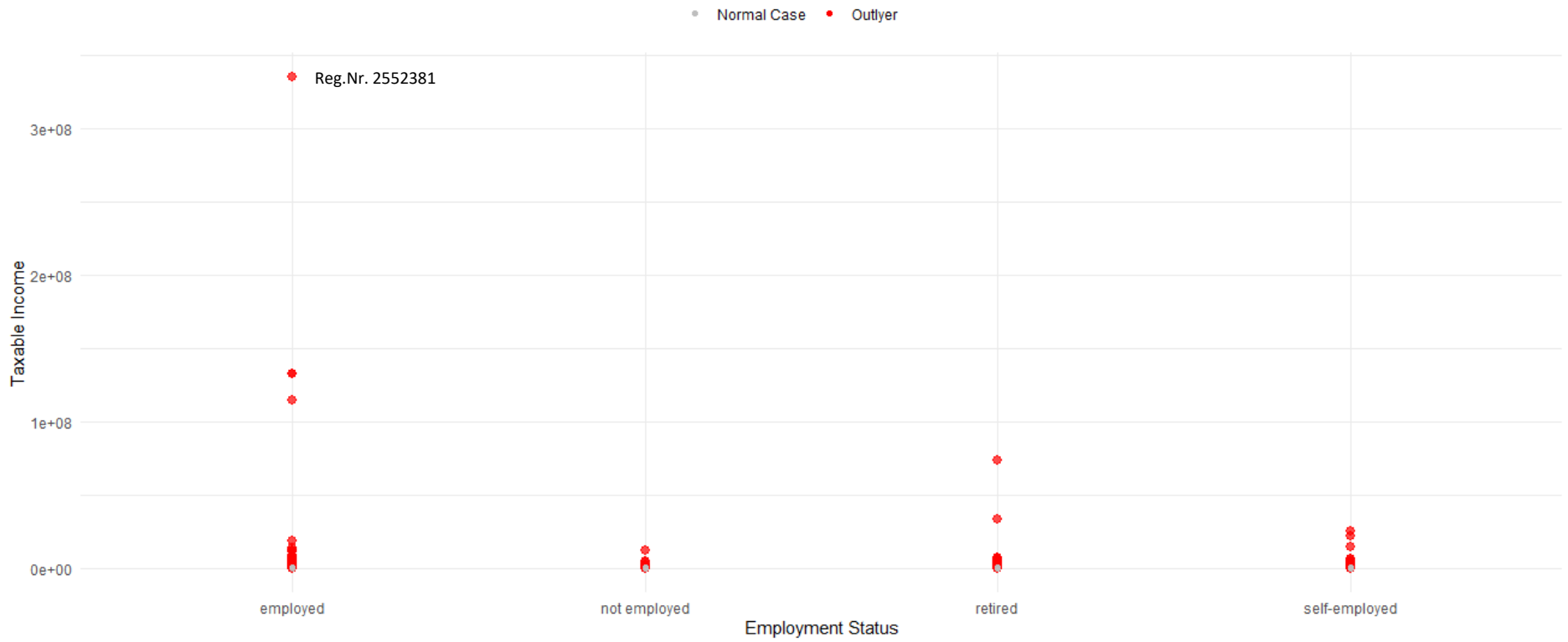
- Analyses multiple variables *simultaneously* instead of considering each one in isolation
- Unsupervised learning method: patterns, groups or relationships are not specified by the auditor
- Independent recognition of latent (hidden) structures
- Easily interpretable anomaly score
- Alternatives: K-means, one-class SVM



binary decision tree to separate data points

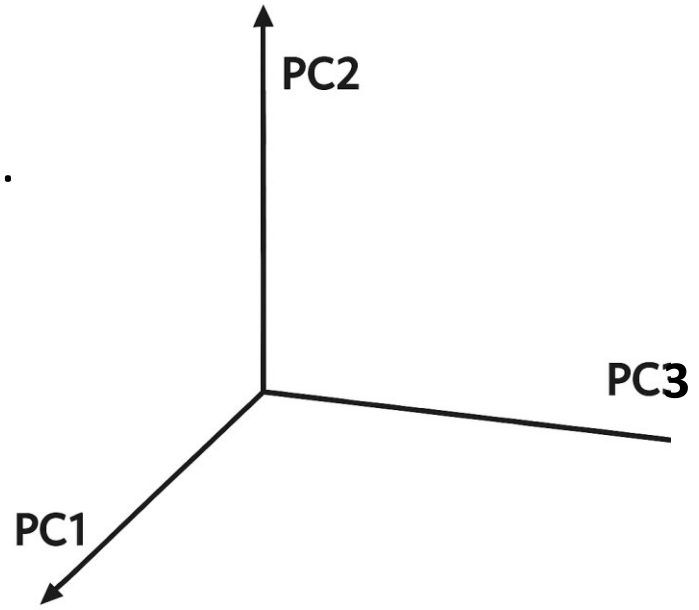
Detection of Anomalies Using the Isolation Forest Algorithm

Anomalies (top 5%) presented by taxpayer and income groups



Reduce Complexity With Principal Component Analysis (PCA)

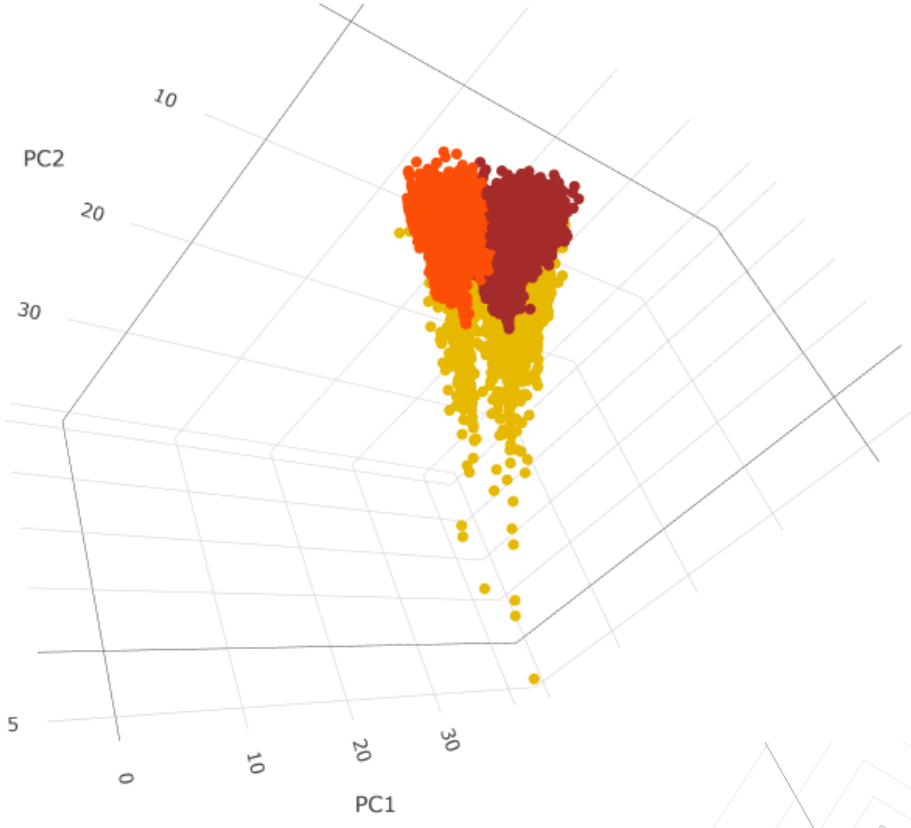
- Many variables often mean redundancy, as they measure similar things. Therefore, they can be reduced to fewer dimensions without losing much information.
- PCA summarizes many variables into a few principal components that explain as much variance as possible (ideally 80 to 95%).
- Benefits:
  - The principal components are uncorrelated (orthogonal), which is ideal for many clustering algorithms such as K-means.
  - Clustering of observations that are similar in the new space (e.g. 2D or 3D) facilitates the interpretability and visualization of groups.

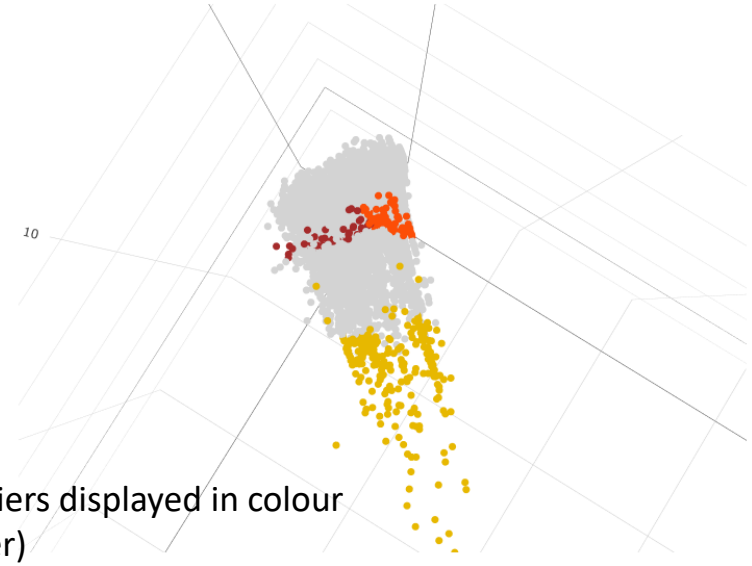


Three Groups of Tax Payers

Approached used (in R):

- → Removal of outliers
- → Standardization of Variables
- → 14 variables reduced to 3 PCs that explain 54% of variance
- → Clustering of data points as a vector in the 3-dimensional space with k-means
- → Visualization as a scatter plot (space balls)
- → Highlighting anomalies in each cluster





Only outliers displayed in colour (by cluster)

Conclusion

- Good Applicability
  - Efficient methods with low computational effort for large data sets
  - Flexible and applicable methods for data with different scale levels
  - Detects even subtle anomalies (and therefore relatively many)
- The results depend heavily on the underlying data. Little variance in the data prevents clear separation of clusters (“grape-shaped” patterns).
- Next steps:
  - Validation of outliers
  - Technical interpretation of outlier characteristics
  - Sample selection for individual case testing, testing and documentation
  - Evaluation of methodology

Questions?

UNIVERSITÄT DES SAARLANDES

Appropriate assessing, analysing and visualising your results Part 5: Data Visualization

Wolfgang Meyer, Saarland University

Introduction to the Course

Data Mining
- Identifying and extracting relevant data from sources

Data Processing
- Sizeable effort in preparation for analysis

Data Analysis
- Using methods to interpret data and develop insights

Data Visualization
- Communicating results including visualization and data storytelling

Chapter 1

1.1 Data Sources
1.2 Systematic Research
1.3 Data Mining

Chapter 2

1.1 Evaluating Catalogue
1.2 Assessing Data Quality
1.3 Errors and Missings

Chapter 3

1.1 Data Management
1.2 Machine Learning

Chapter 4

1.1 Descriptive
1.2 Clustering
1.3 Combining
1.4 Longitudinal

Chapter 5

1.1 Basic Graphs
1.2 Advanced Visualization

Source: Hyman et al (2024): Data Analytics and Visualization. Hoboken: John Wiley & Sons, p. 13 (modified)

Content Chapter 5

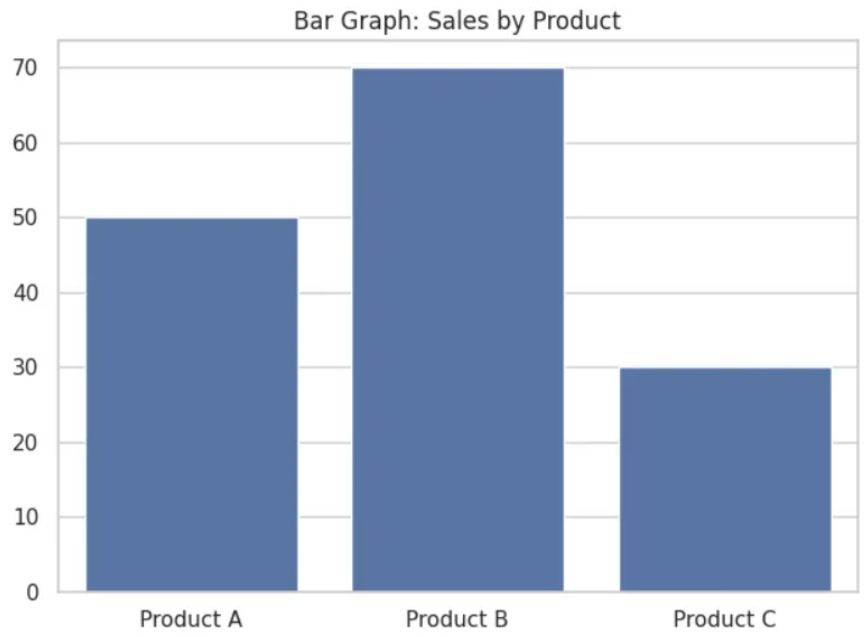
- Basic Graphis
  - Bar Graph
  - Line Graph
  - Pie Graph
  - Histogram
  - Scatter Plot
  - Area Chart
  - Bubble Chart
  - Spider Chart
  - Time Series Chart
  - Box Plot Chart
- Advanced Visualization (How to lie with statistics)
  - The Gee-Whiz Graph
  - The Dimensional Picture
  - Round Numbers are always false

Chapter 5

1.1 Basic Graphs
1.2 Advanced Visualization

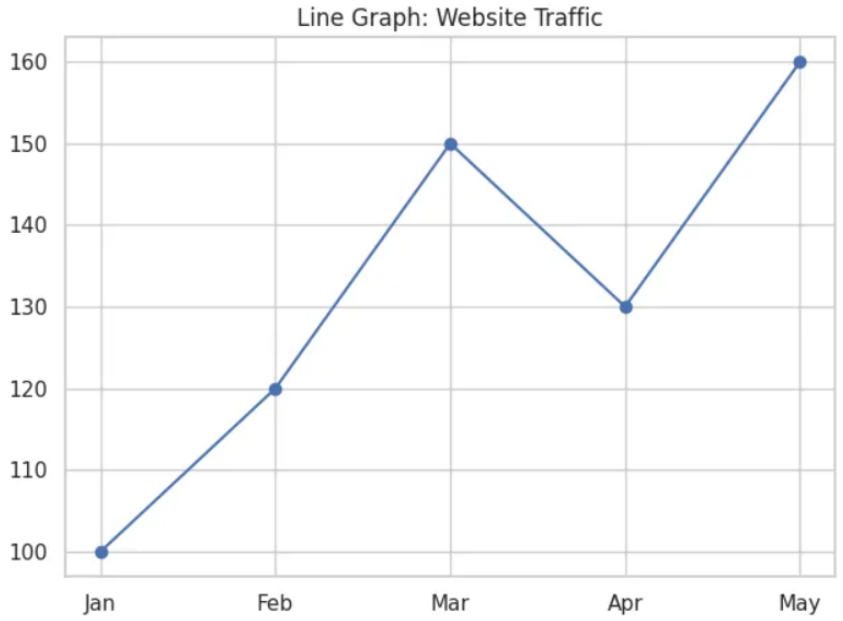
Basic Graphs

Bar Graphs



Example of a Bar Graph

Line Graphs

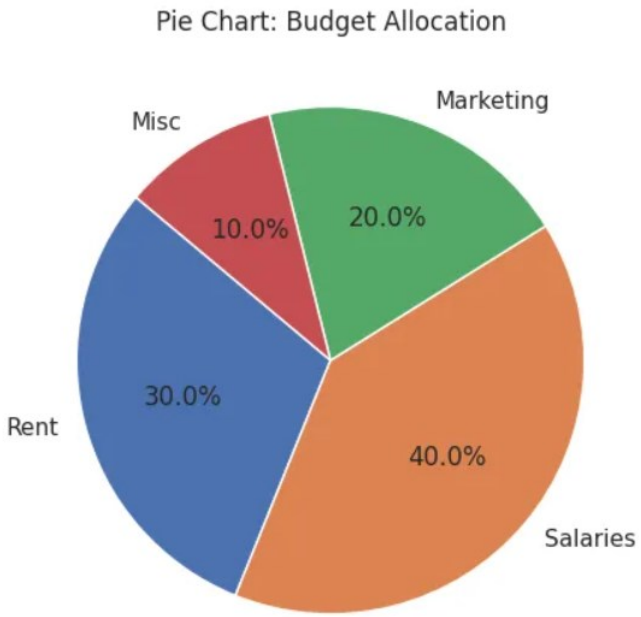


Line Graph Example: These Types of Graphs are Best for Showing Trends Over Time

https://graphtutorials.com/types-of-graphs/

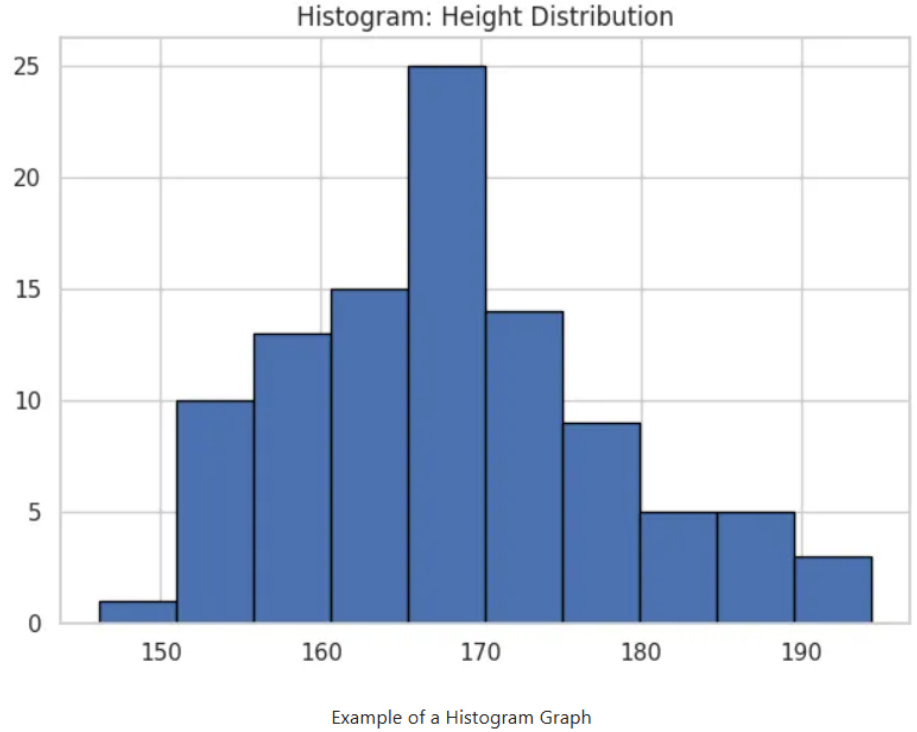
Basic Graphs

Pie Graphs



Pie Chart Example: These Types of Graphs are Best for Showing Proportions or Percentages of a Whole

Histogram

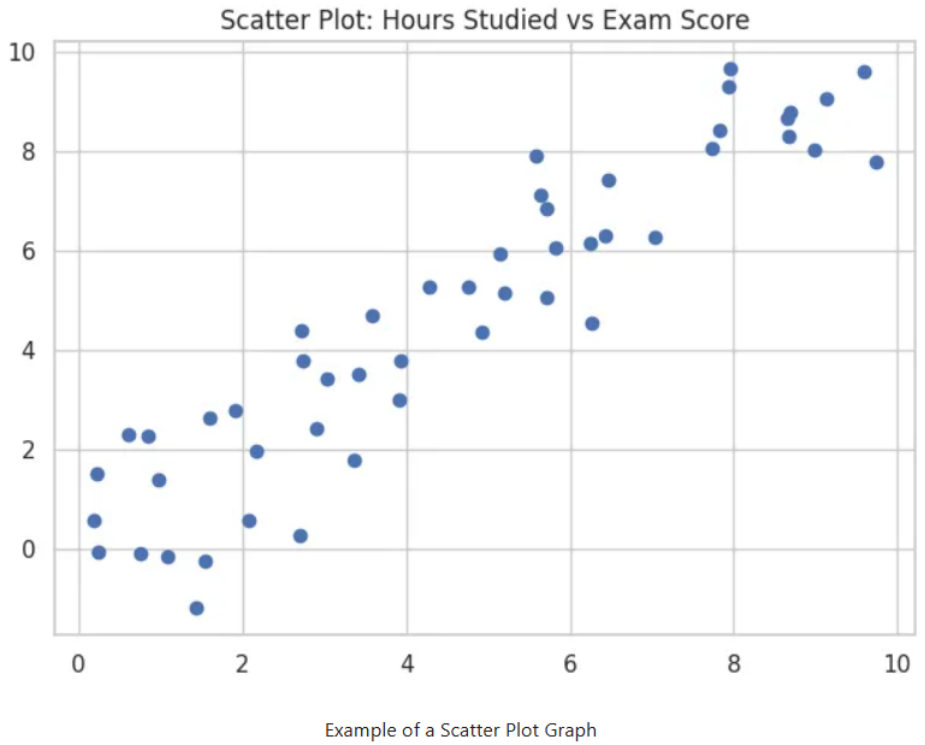


Example of a Histogram Graph

https://graphtutorials.com/types-of-graphs/

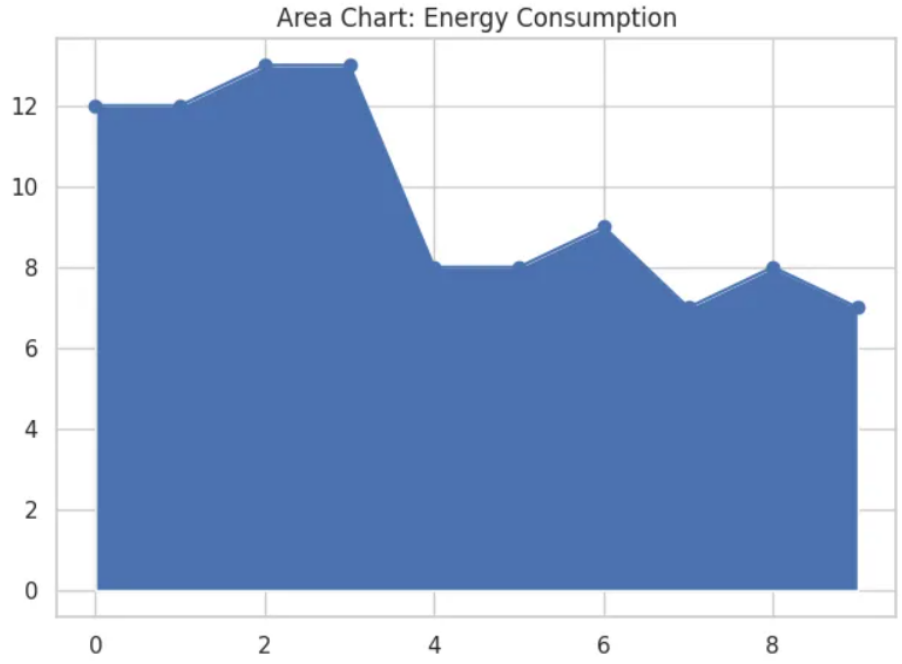
Basic Graphs

Scatter Plot



Example of a Scatter Plot Graph

Area Chart

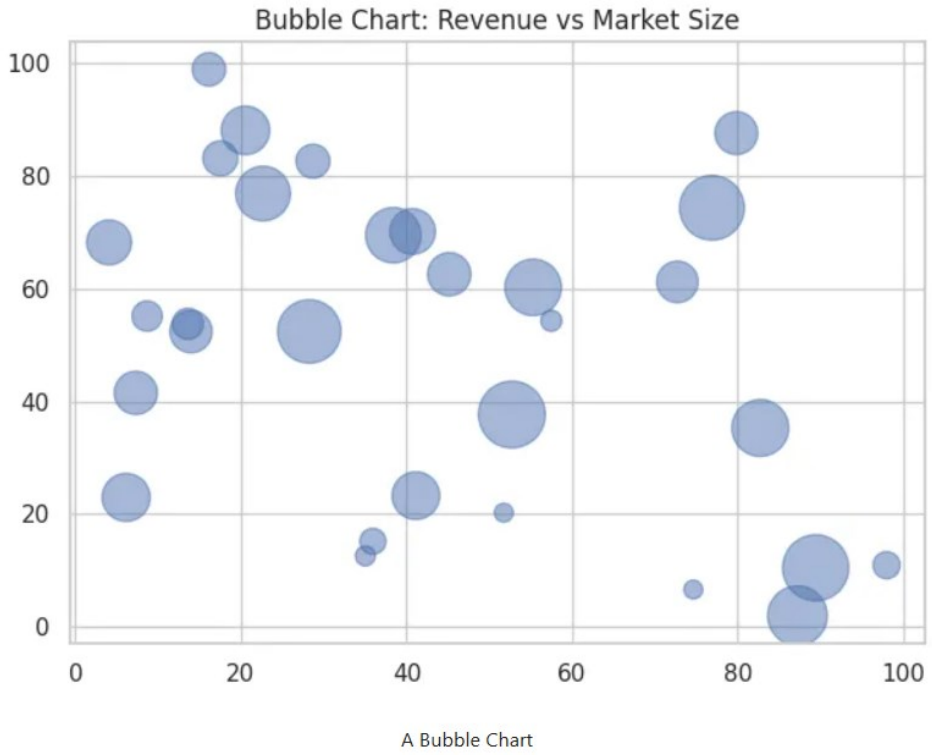


Area Chart

https://graphtutorials.com/types-of-graphs/

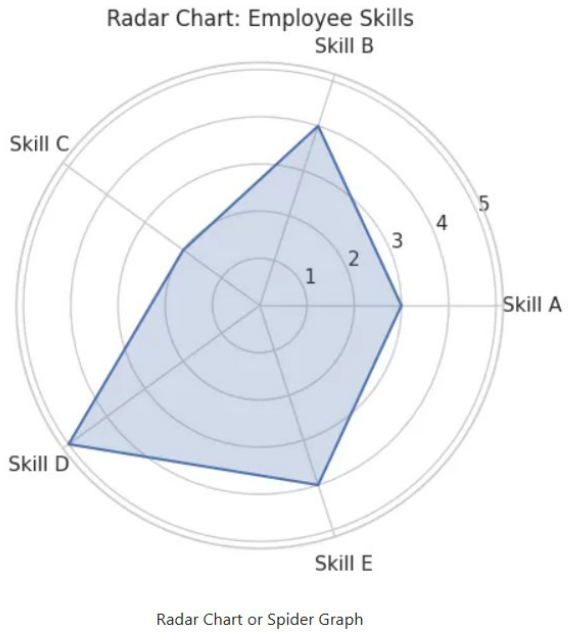
Basic Graphs

Bubble Chart



A Bubble Chart

Spider Chart

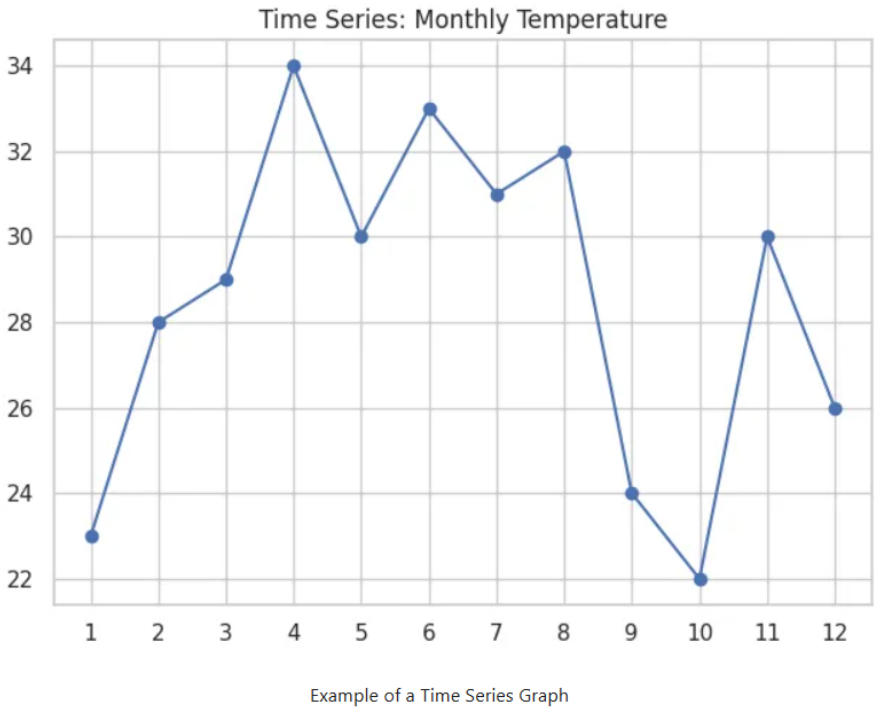


Radar Chart or Spider Graph

https://graphtutorials.com/types-of-graphs/

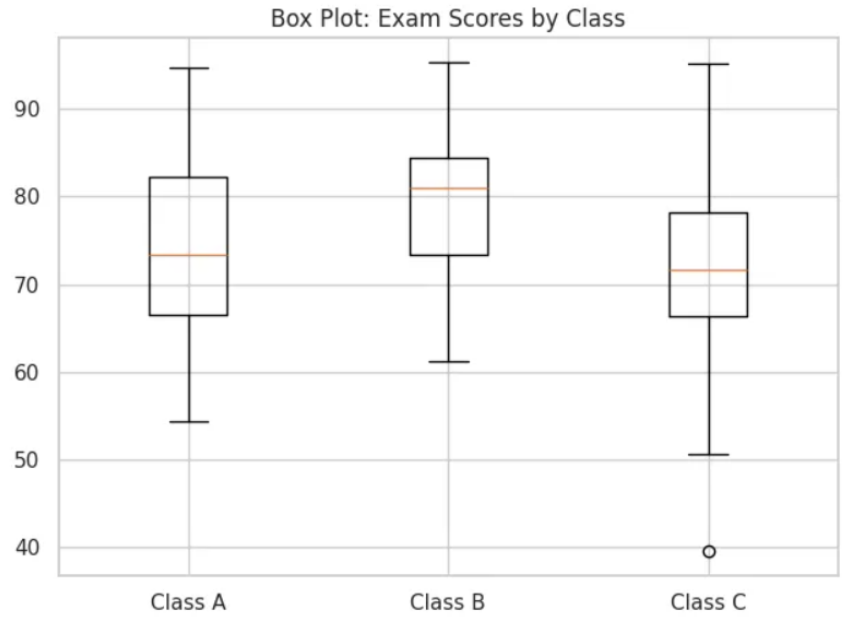
Basic Graphs

Time Series Chart



Example of a Time Series Graph

Box Plot Chart



Box Plot / Box-and-Whisker Plot Graph

https://graphtutorials.com/types-of-graphs/

Basic Graphs – How to Use them

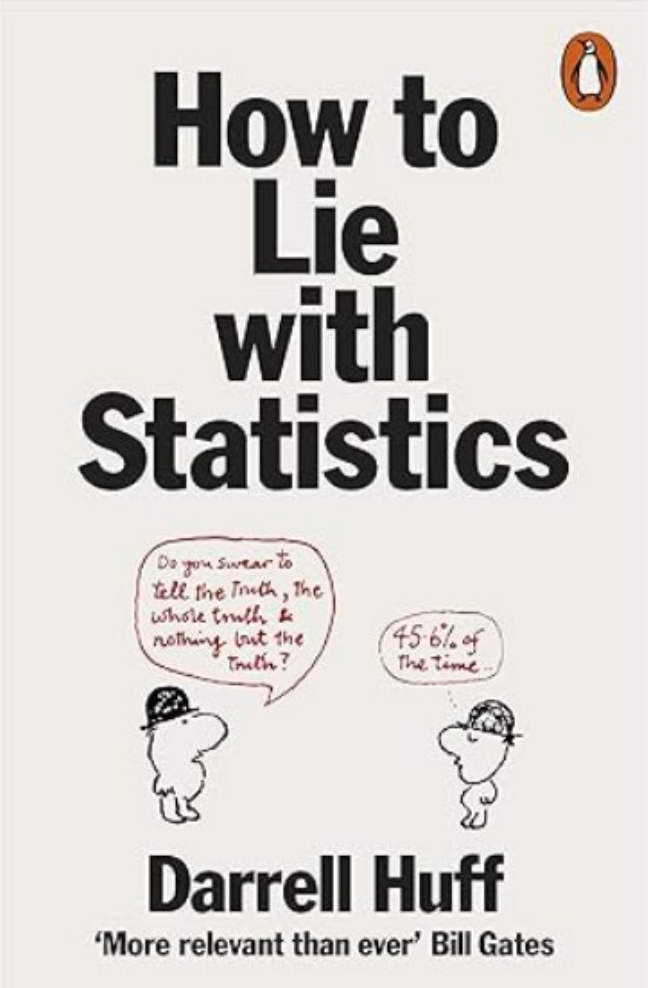
Type	Best for
Bar Graph	Comparing quantities across different categories
Line Graph	Showing trends over time
Pie Graph	Showing proportions or percentages of a whole
Histogram	Showing frequency distributions of continuous data
Scatter Plot	Exploring relationships or correlations between two variables
Area Chart	Visualizing part-to-whole relationships over time
Bubble Chart	Showing relationships with three variables (X, Y, and size)
Spider Chart	Comparing multiple variables across categories
Time Series Chart	Tracking changes in data over time
Box Plot Chart	Displaying distribution, median, and outliers

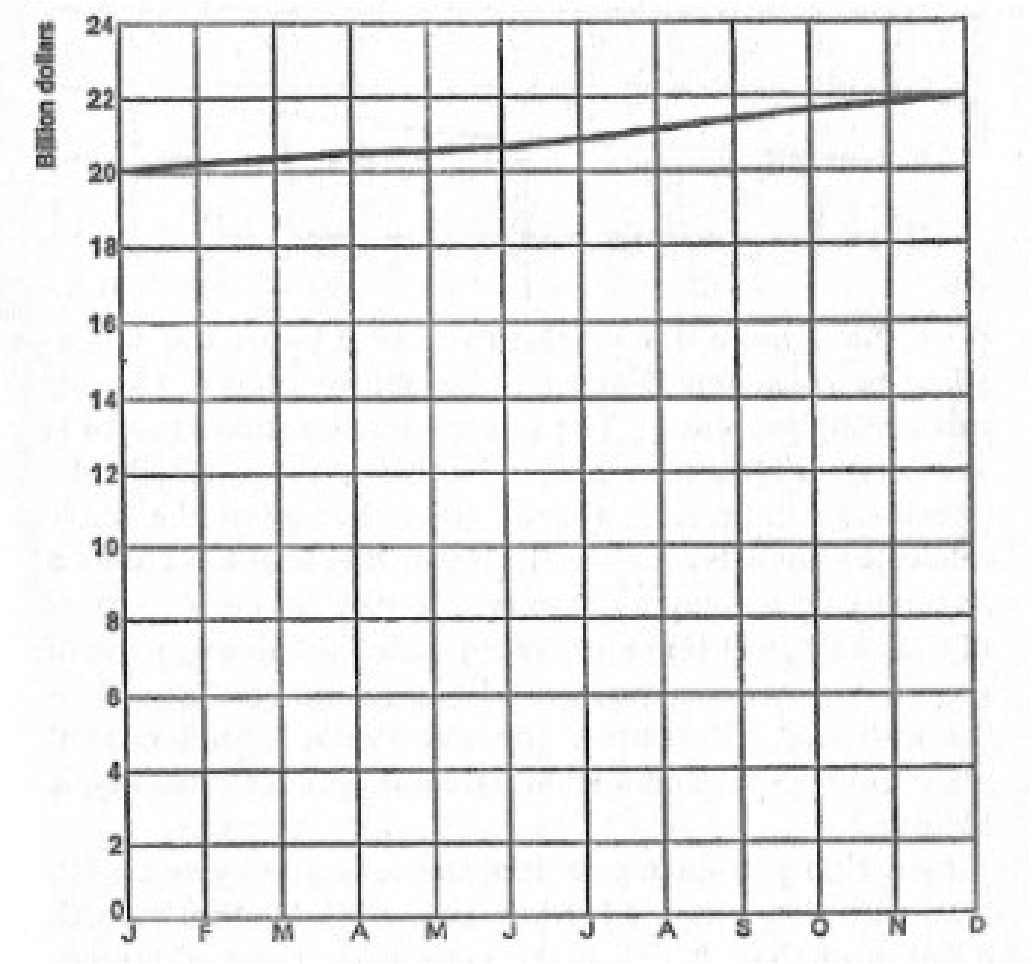
https://graphtutorials.com/types-of-graphs/

Basic Graphs – Making comparisons visual

Comparison	Useful Graphs
Groups	Bar, Box Plot
Correlation	Line, Scatter Plot, Bubble Chart
Development	Time Series, Area Charts
Variables	Spider Chart
Distribution	Pie, Histogram, Spider Chart, Box Plot

How to lie with Statistics – The Ghee-Whiz-Graph

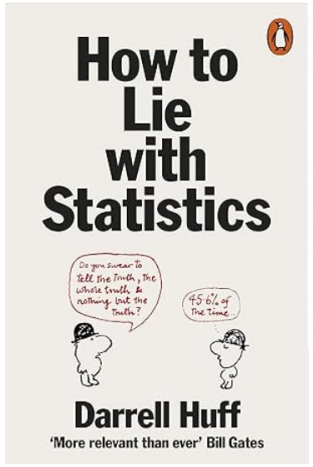


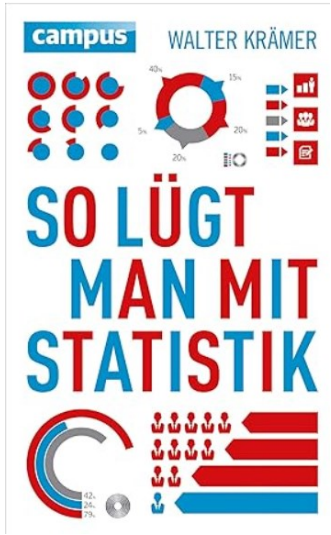


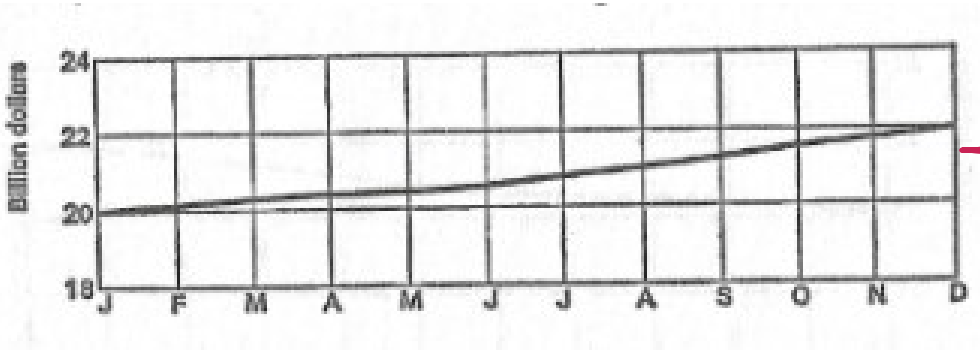
The real thing

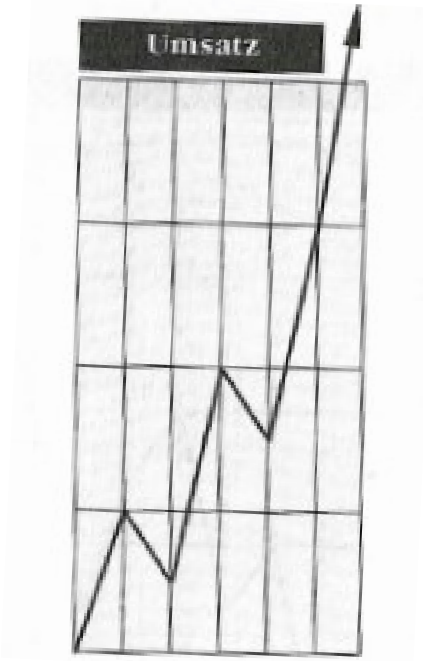
Huff, D. (1954, reissued 2023). How to lie with Statistics. Penguin, pp.50ff.

How to lie with Statistics – The Ghee-Whiz-Graph

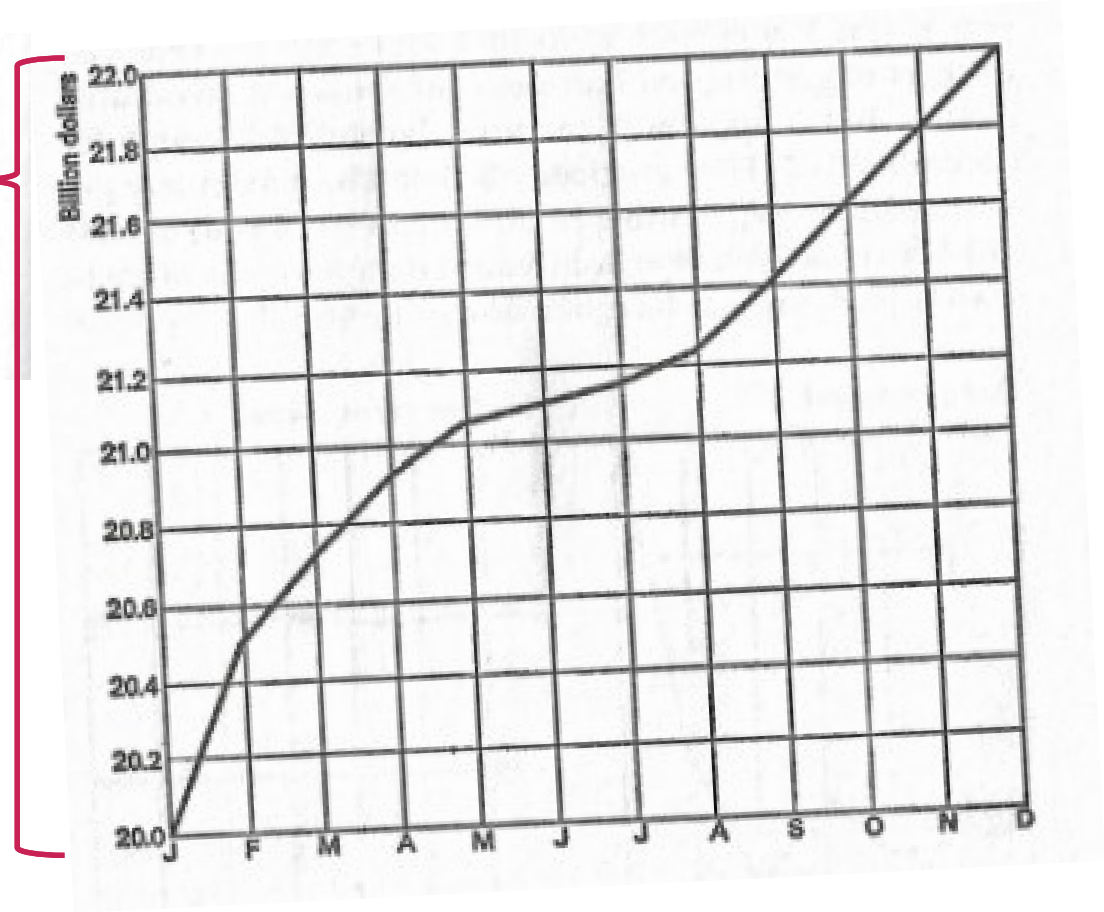




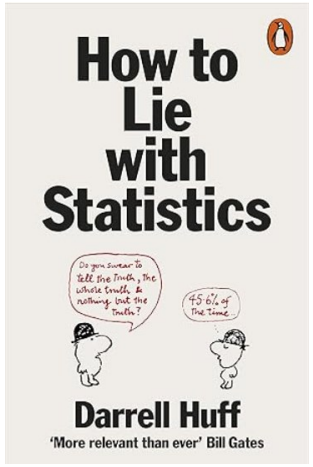


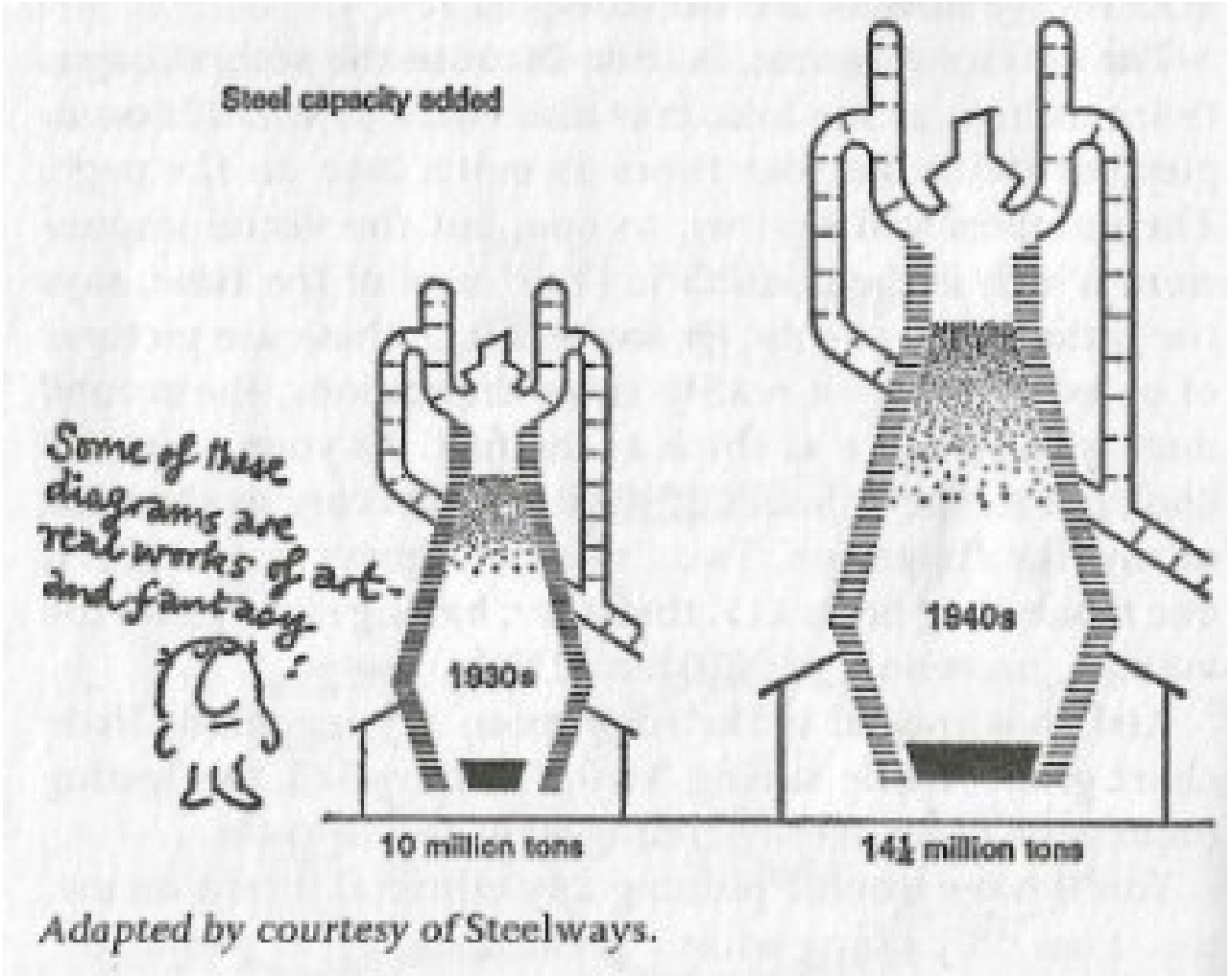


To make a mountain out of a molehill



How to lie with Statistics – the Dimensional Picture

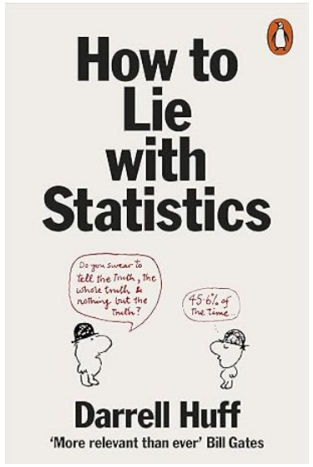


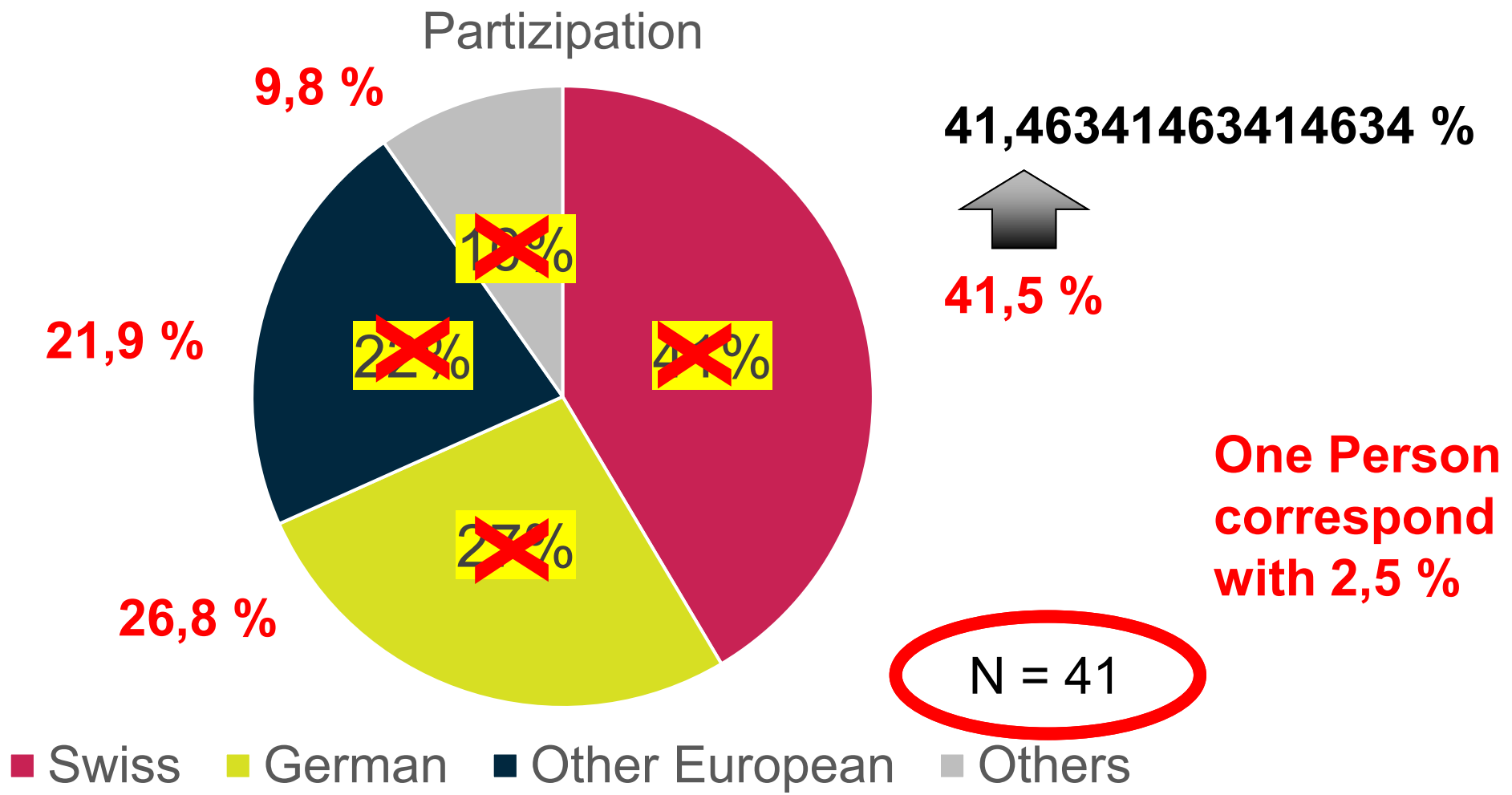


Growth Rate 42% (Data)

Growth Rate 115% (Graph)

How to lie with Statistics – Round Numbers are always false





Finally – One do not need graphs to lie with statistics

The Result

„96,5% of all Dentists recommend XXX Tooth Paste“

The Question

„What do you recommend for dental hygiene?

a) XXX Tooth Paste

b) Strychnine“

Source: MAD Magazine

Even the best statistical analysis is not able to correct wrong data