FRANKFURT UNIVERSITY OF APPLIED SCIENCES

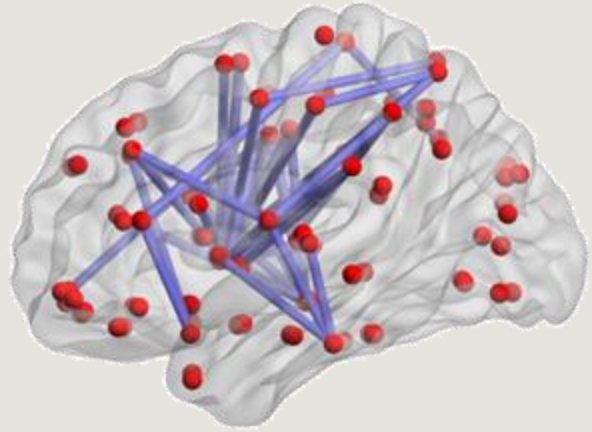
Introductory Data Analysis

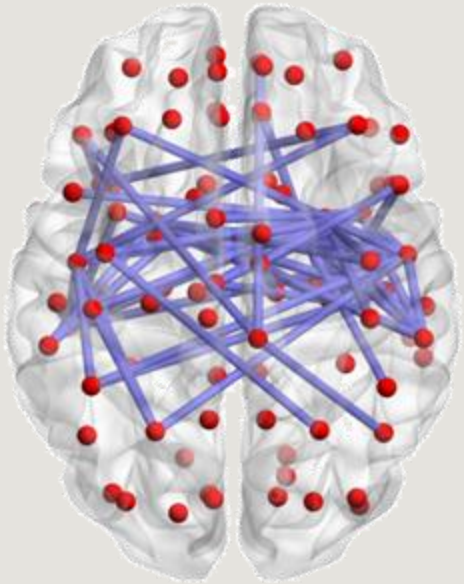
Prof. Dr. Matteo Marouf

14.10.2025 & 21.10.2025

Agenda

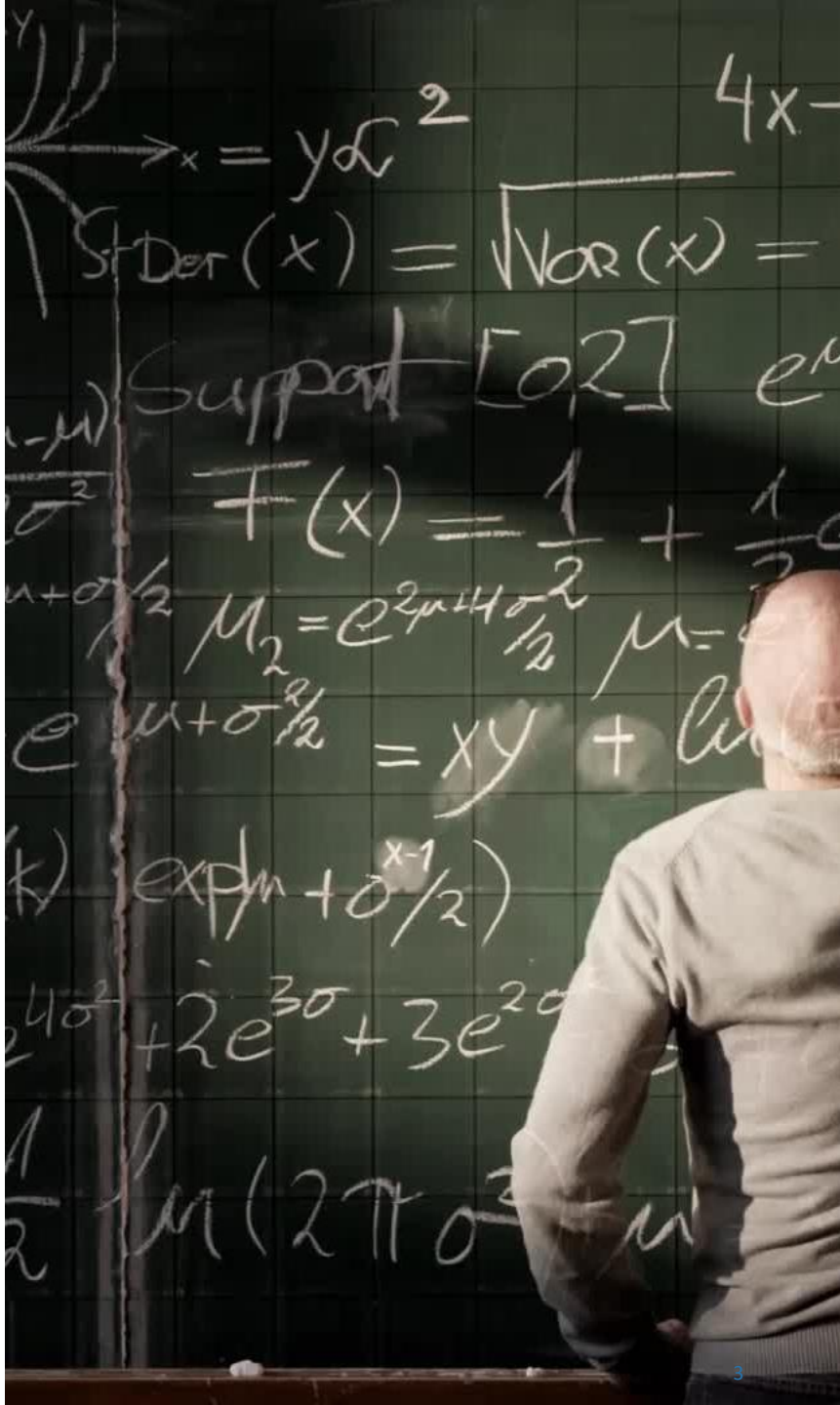
- Who is your professor?
- establishing classroom agreements
- Logistics
- what we will study and which tool we will use
- Introduction to Data analysis
- Data types
- Statistical Inference cycle





Who is your professor?

- Who am I?
- What should you expect from me?
- What do I appreciate?



slido

Please download and install the Slido app on all computers you use

How would your rate your knowledge in topics like probability and statistics?

ⓘ Start presenting to display the poll results on this slide.

Classroom agreements: How can we behave to have a productive and enjoyable learning experience?

Classroom Agreements

- Interactive Teaching- you need to solve small tasks or answer short questions in the classroom
- Stay silent unless you have a question and mute your smartphone.
- Blocks start at 11:45, please arrive on time.
- Once you are assigned to a group, switching is not allowed!
- We are a large group, so please avoid emails with excuses and personal requests.

Logistics about this model

- Read through the instructions provided on CampUAS page **carefully.**

- **Important: select a group by 17.10.2025 based on your program:**
  - HIS M.Sc. students starting in Summer Semester 2025:
    - Tuesday (14:15–15:45)
    - Tuesday (16:00–17:30)
    - Thursday (11:45–13:15)
  - HIS M.Sc. students who started in Winter Semester 2025/2026:
    - Tuesday (14:15–15:45)
    - Tuesday (17:45–19:15)
    - Thursday (14:15–15:45)
  - Wirtschaftsinformatik – Master (WI-M):
    - Tuesday (14:15–15:45)
  - Allgemeine Informatik – Master (AI-M):
    - Thursday (14:15–15:45)

For Ineligible students

Students who cleared the prerequisite should practice at home. I will test a group assignment option on CampUAS.

Logistics about this model

Groups:

- Unfortunately, your group selection is non-binding (you may land in another group).
- Only eligible students are allowed to join exercise groups.
- Priority is given to those who have not cleared the prerequisite yet.

Requirement for the final exam eligibility:

- **80% active participation in the exercises is mandatory for exam eligibility.**
- **You'll also have to take a rehearsal exam with a small task to achieve exam eligibility.**

About the module:

- We will program with Python and use common statistical Python frameworks.
- The module is conducted in English.
- Attending all Blocks is strongly recommended for exam preparation.

What will we use to analyze data?

Python: Development Frameworks for bold and muscular data analysts

Category	Python-Based
Statistical analysis	Statistics, Statsmodels
Essential packages	Numpy, Pandas, Matplotloib, Seaborn
Interactive Data Visualization	Streamlit, plotly
Machine Learning	Scikit-learn, XGBoost, RAPIDS
Big Data Analytics	Apache Spark (PySpark)
Large Language Models	Hugging Face Transformers, OpenAI GPT
Deep Learning	TensorFlow, PyTorch

- Non-Python alternatives: Matlab, RapidMiner, R, Excel, SAS..etc.

So, what will we address throughout this course?

We are drowning in information but starved for knowledge.

John Naisbitt

In nutshell, we will learn statistical analysis

Statistical analysis is all about converting data into useful information (insights).
Statistics is a process in which we:

Collect Data

Summarize Data,

Interpret Data and Analyze

How do you pass this module successfully?

1. Attend the classroom

2. Attend and try to solve the exercises in the session

3. Practice the exercises at home without AI assistance

4. Try to reach out to your colleagues first and try to open a discussion about the lectured topics before you resort to AI assistance.

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Introduction to Data Analysis

Prof. Dr. Matteo Marouf

14.10.2025

Turn to your neighbor and discuss a potential use-case of Data Analysis use-case.

- It should be from the HIS context

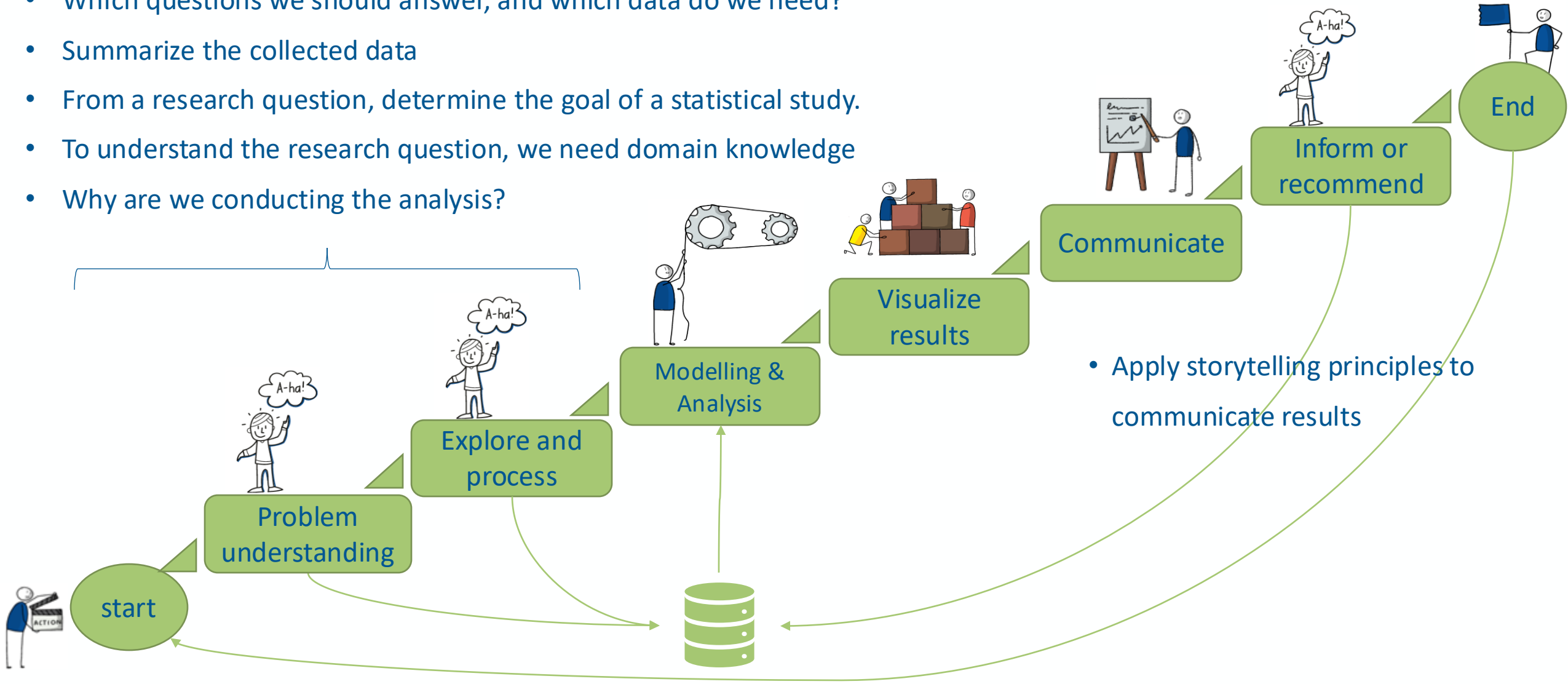
- Requires Computer Science skills (e.g., programming, databases).

- Impactful and relevant (leading to better decisions, efficiency, or outcomes).

Describe input data, required modeling/Analysis technique, outcome, how to use-results

Statistical investigation

- Which questions we should answer, and which data do we need?
- Summarize the collected data
- From a research question, determine the goal of a statistical study.
- To understand the research question, we need domain knowledge
- Why are we conducting the analysis?



- Apply storytelling principles to communicate results

Start: Informative Data representation

Data exist in different formats

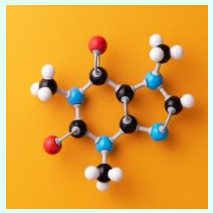
Social media

Imaging

EHR

Genomics

graphs



Bioactive molecules

Pharmaceutical packaging machines mainly focus on large-volume products: Quantities of more than 100,000 units are standard. They are not working well for medicines that are produced and packaged only in small quantities, so-called microbatches. This is where conventional packaging machin are mostly inefficient due to the long setup and changeover times. A problem that production experts from Boehringer Ingelheim have been working on in recent years.

Textual data

Vital signals

Data Representation

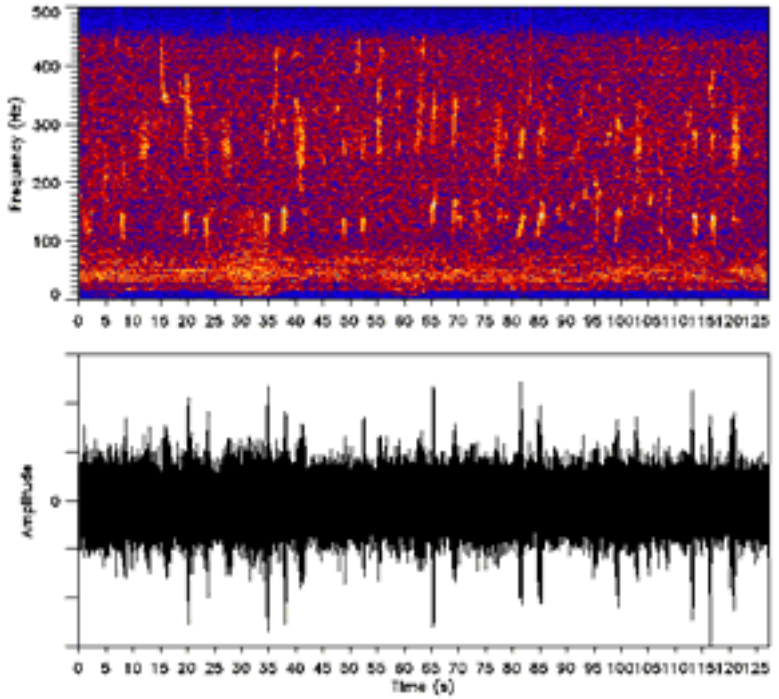
To analyze data, we must create an **informative and numerical representation of the data** that can be manipulated by electronic devices.

Examples

- **Textual data:** Word Embedding, or tokenization is a term used for the representation of words for text analysis, typically in real valued vectors or tokens.
- **Images**: Image Pixel intensity values are used.

Examples on finding a proper Data representation

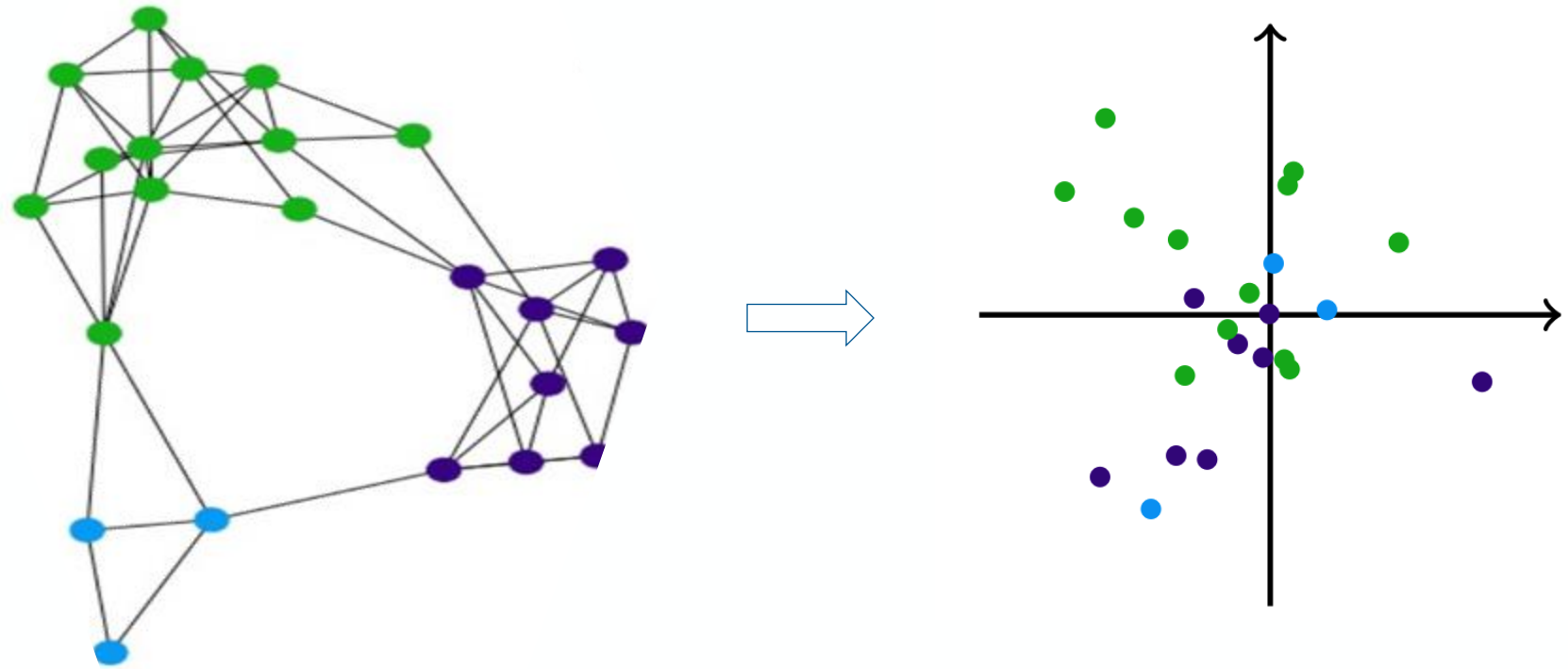
Spectrograms: An audio signal can be transformed into a spectrogram, representing the signal's frequency spectrum over time. This visual representation is useful for speech and music analysis.



https://commons.wikimedia.org/w/index.php?curid=34672291

Examples on finding a proper Data representation

- **Node Embeddings (e.g., Node2Vec) for graph data :** Learn continuous feature representations for nodes in a graph, capturing the network's structural information.



Examples on finding a proper Data representation

- **Term Frequency-Inverse Document Frequency (TF-IDF)** converts textual information into numerical vectors, reflecting the importance of words in documents.

TF

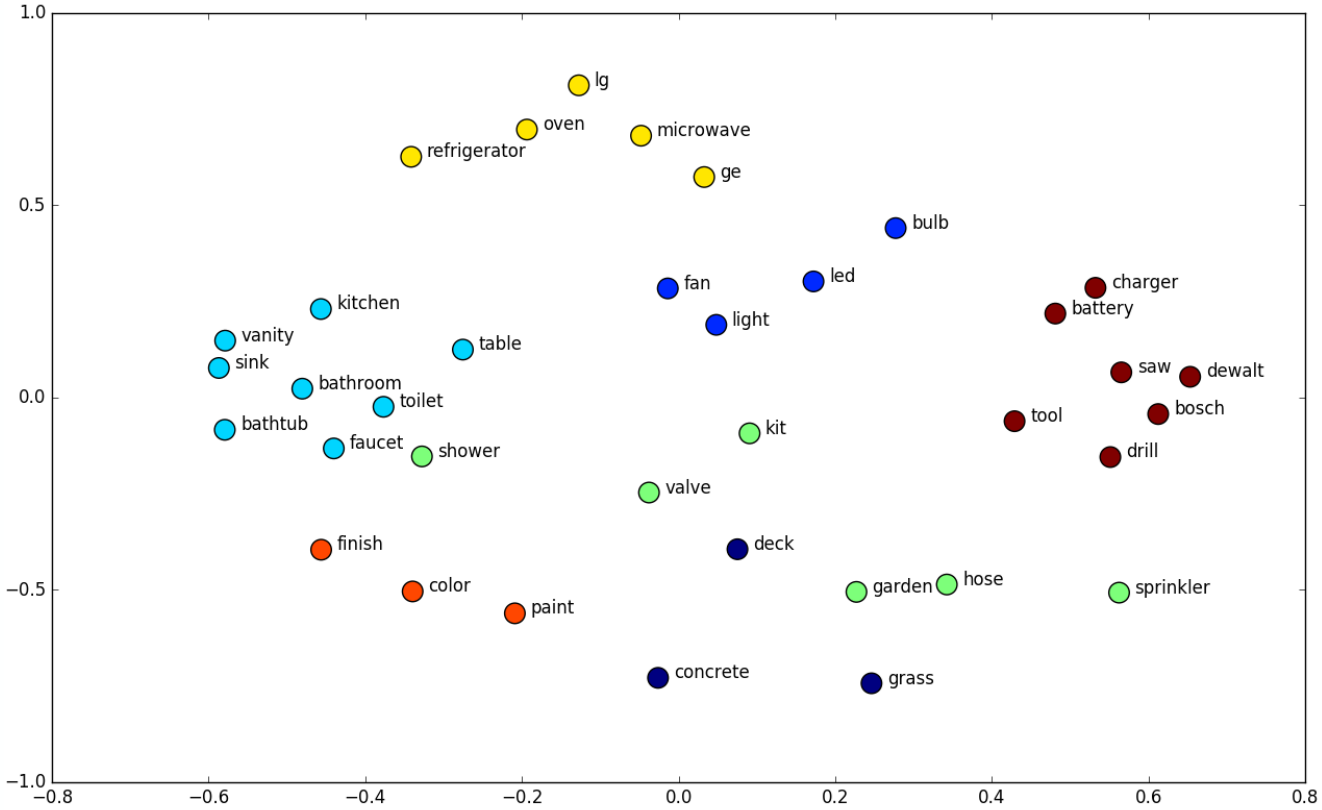
Frequency of a word withing a document

IDF

Frequency of a word across the corpus

Examples on finding a proper Data representation

- **Word Embeddings (e.g., Word2Vec, GloVe):** Word embeddings map words into continuous vector spaces where semantically similar words are positioned closely. These vectors capture contextual relationships and can be pre-trained on large corpora.



Outcome: Storytelling is a crucial skill when presenting and communicating data mining results

Key components:

- **Data** – the evidence source
- **Visualization** – charts, graphs, dashboards, etc.
- **Narrative** – the storyline that guides interpretation (what happened, why, what now?)

Common types of visualizations used in business presentation

Story telling with data, Cole Nussbaumer Knaflic

Pitfalls of Statistics: number can mislead if you don't watch out

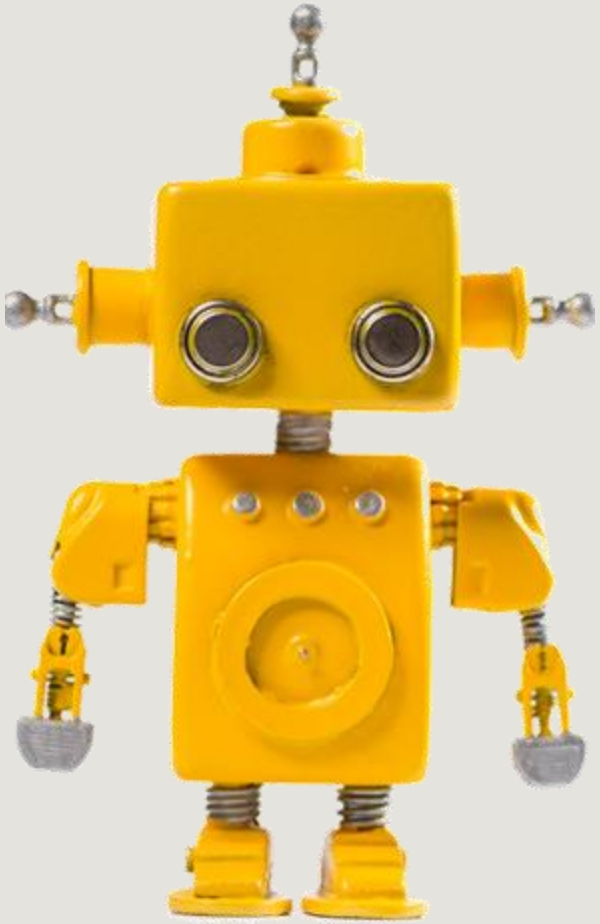
Example: Which treatment is better to better cure kidney stones?

Stone size	Treatment A (success / total)	A %	Treatment B (success / total)	B %
mixed	**273 / 350**	**78%**	**289 / 350**	**83%**

Look at the subgroup rows

Stone size	Treatment A (success / total)	A %	Treatment B (success / total)	B %
Small stones	**81 / 87**	**93%**	234 / 270	87%
Large stones	192 / 263	**73.0%**	**55 / 80**	69%
Combined	**273 / 350**	**78%**	**289 / 350**	**83%**

Simpson Paradox — the aggregated result gives the opposite conclusion from the subgroup results.
Never trust a single summary number — always ask: What's the denominator? Are there hidden subgroups?



Quiz with AI

- **Give me another real example of a politician from any country who used a misleading statistic. The example should involve one of the following**:
  - Ignoring the denominator
  - Leaving out broader context
  - Falsely amplifying a pattern
  - Focusing on total numbers and not mentioning the subgroups

Briefly explain what the statistic was and why it was misleading

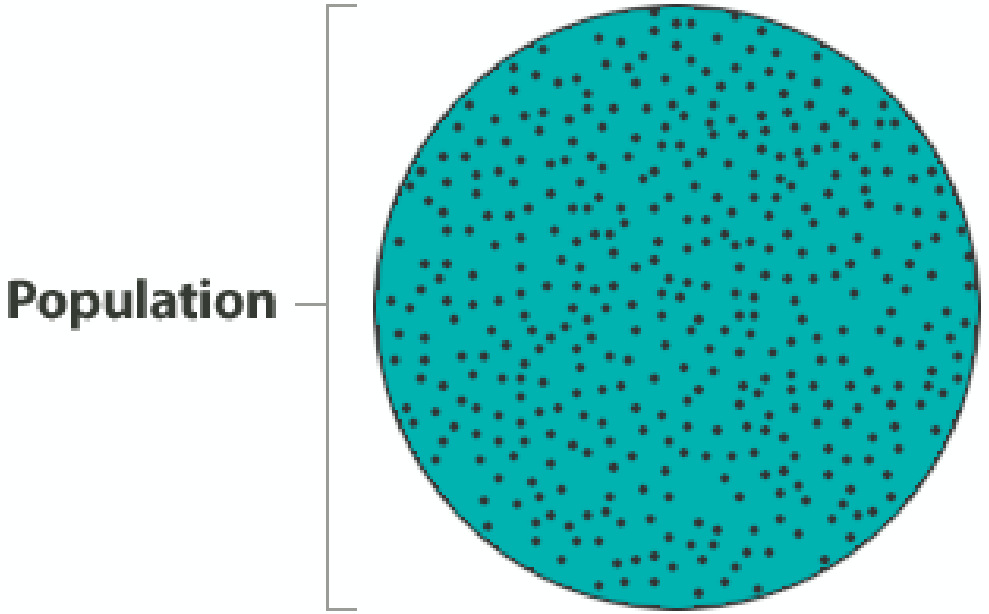
FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Statistical Inference cycle

- Producing data
- Exploratory analysis
- Probability
- Inference

Producing data from the population

- The process of statistical analysis starts with identifying the group we would like to describe and analyze.
- Population, then, is the entire group that is the target of our interest.
- In most cases, the **population is so large** that, as much as we want to, there is absolutely no way we can study all of it.



Quiz

A researcher wants to understand the eating habits of all university students in the United States but only collects data from 500 students at five universities. Which of the following best describes the *population* in this study?

a. The 500 students who were surveyed

b. All students at the five universities involved

c. All university students in the United States

d. Only the students who responded to the survey questions

We study population but we analyze a sample (data)

A more practical approach would be to examine and collect data only from a subgroup of the population, which we call a *sample*. We call this first step, which involves choosing a sample and collecting data from it, **producing data**.

What should we consider when we sample?

✅ Population Clear Definition

- The population must be clearly defined (who exactly are we talking about?). Without this, you risk sampling from the wrong group.

✅ Representativeness

- The sample should reflect the characteristics of the population (e.g., age, gender, location, behavior). If the sample is skewed, conclusions won't generalize.

✅ Random Selection

- Random sampling helps reduce bias and gives every member of the population a fair chance of being chosen. Non-random methods often introduce systematic errors.

✅ Sample Size

- Too small → unreliable results and high variability.
  Too large → costly and often unnecessary.
  Choose a size that provides enough statistical power to detect effects.

✅ Avoiding Bias

- Be mindful of: **Selection bias** (certain groups over/underrepresented)
- **Nonresponse bias** (those who don't participate differing from those who do)
- Other biases

✅ Sampling Method

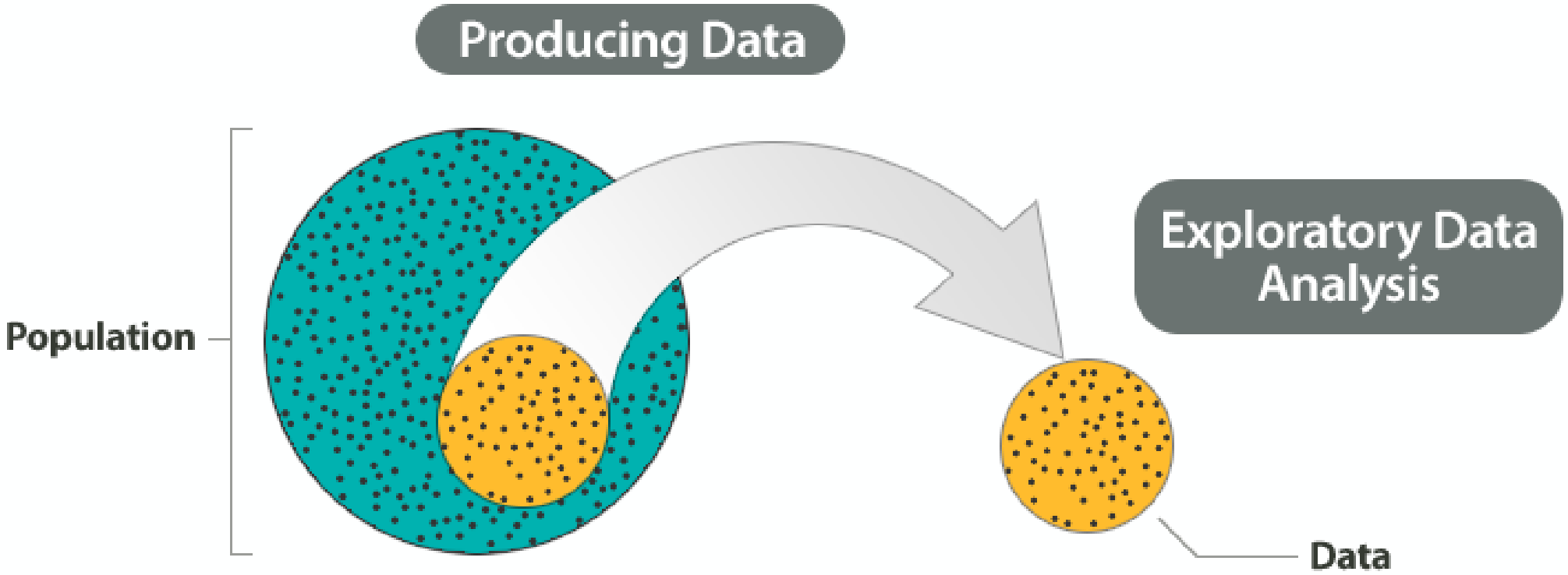
- Different approaches suit different studies:
  -Simple random sample
  -Stratified sample
  -Cluster sample
  -Systematic sample

✅ Practical Constraints

- Resources like time, cost, and accessibility can limit how the sample is chosen. These should be balanced with the need for validity

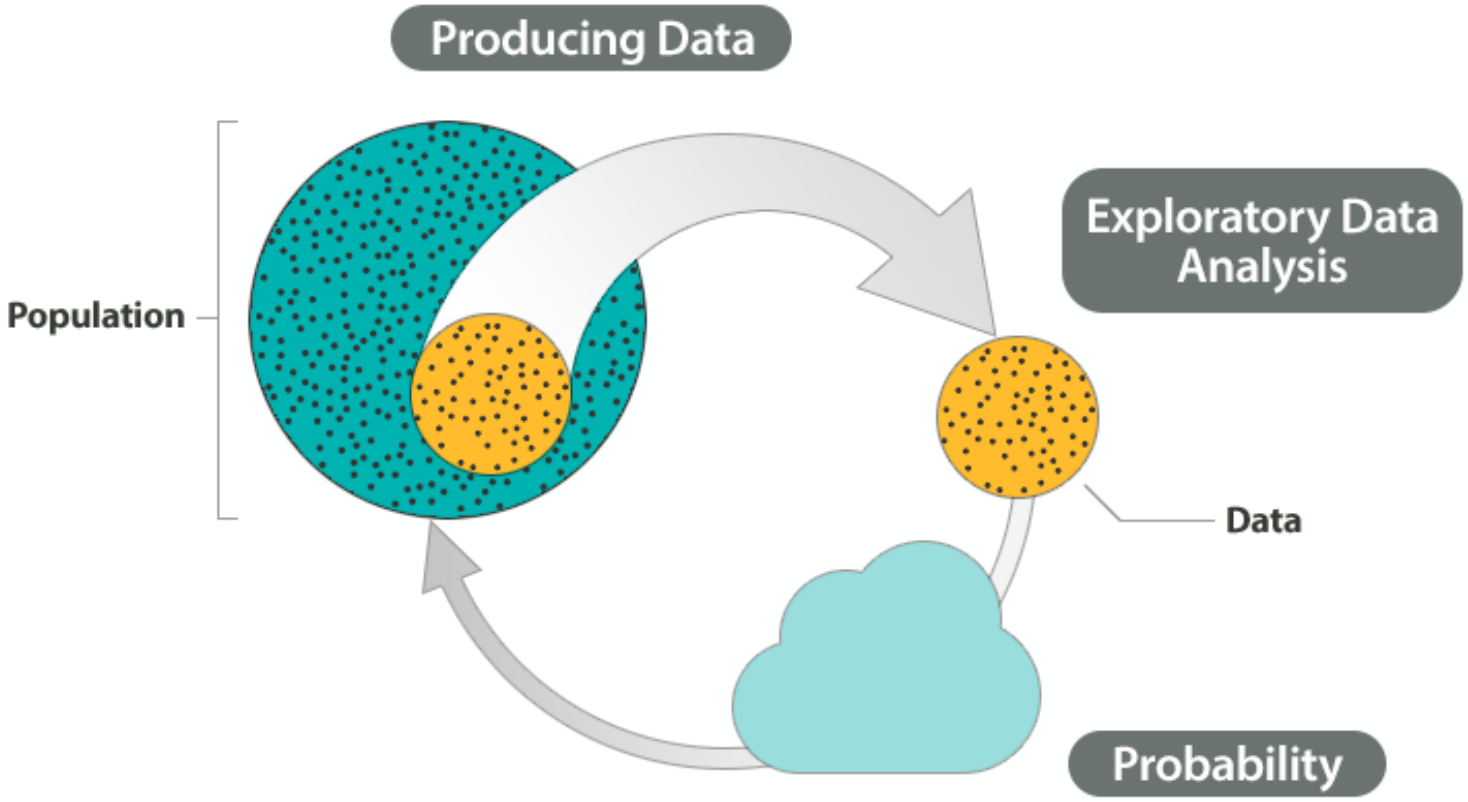
Exploratory Data Analysis is usually the first step if you have sampled data

- After collecting data, we have a long list of answers or numbers. To explore and understand the data, we need to summarize it meaningfully. This step, called **exploratory data analysis**, involves summarizing the collected data.



Probability to evaluate the sample fitness to draw conclusions about population

- Remember what we want is to be able **to draw conclusions about the population** based on the sample results. However, we must find a way to understand **how much the sample might differ** from the population. That's where **probability** comes in.



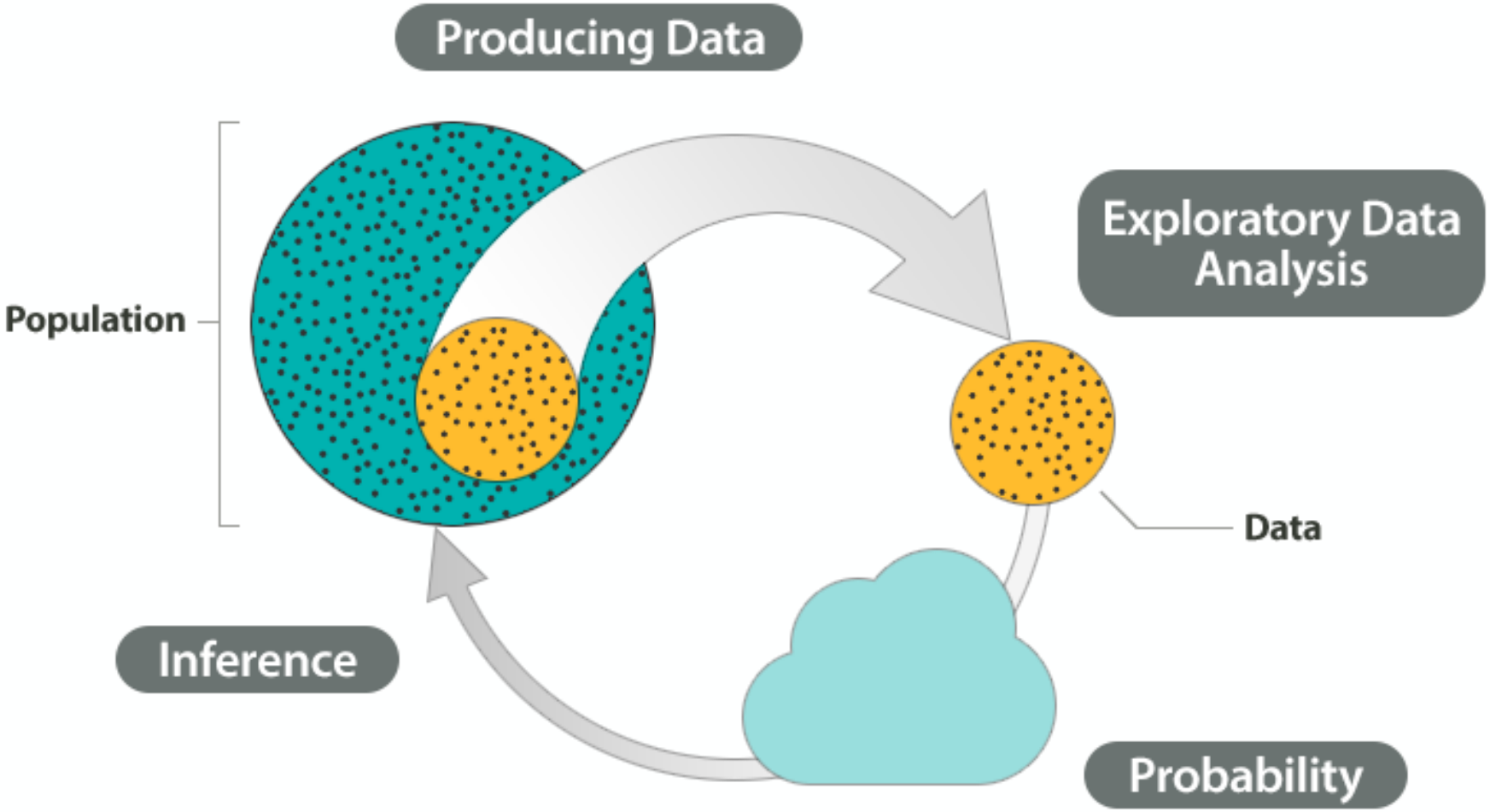
Real-World Example: Estimating Customer Satisfaction

- A store wants to know what **percentage of *all* customers** are satisfied with their service. That's the **population**, but they can't ask every customer.
- ✅ **Step 1: Take a Sample**
  - They randomly survey **100 customers** (the **sample**).
    Out of those, **82 say they're satisfied** → sample proportion = **0.82** (82%).
  - If we stop here, all we know is what this *sample* said — not the whole population.
- ✅ **Step 2: Use Probability to see if and how can we Infer Back to the Population**
  - Probability helps us account for the fact that the sample is only one possible group out of many and might not reflect the population perfectly.
  - So we can answer : **"What can we say about the true satisfaction rate in the *entire* customer base based on just this sample?"**

We'll delve into this in more detail later.

Inference

- Finally, we can draw conclusions about our population based on what we've discovered about our sample. This final step in the process is called **inference**.



Types of Statistical analysis

Main types of Statistical Analysis

Observational Studies

- describe a group of individuals or to investigate an association between two variables.
- Researcher observes without interfering
- **Example: Comparing smoking habits and lung health**

Experimental Studies

- Researcher manipulates variables
- Participants are assigned to groups
- Used to provide evidence for a cause-and-effect relationship between variables
- **Example: Testing effectiveness of a new drug**

Other types of Statistical Analysis

Descriptive Studies

- Summarize or Describe the data
- Answers “what is happening?”
- ***Example:* Average age of patients at a clinic**

Surveys

- Data collected through questionnaires or interviews
- Useful for opinions, behaviors, demographics
- **Example: Student satisfaction survey**

Case Studies

- In-depth analysis of a single individual or group
- **Example: Detailed review of a rare medical condition**

Quiz: Identify the most appropriate description of this study

Educational psychologists conduct research to understand the effects of various instructional methods on learning. In a study, researchers taught a math lesson to 9th graders using the "Inventing to Prepare for Learning (IPL)" instructional cycle. A control group received traditional "tell and practice" instruction. After the lessons, both groups independently studied a worked example of a math problem. Subsequently, they took a test that included problems similar to the worked example. The findings of this study were published in the journal Cognition and Instruction in 2004.

a. An observational study designed to estimate or make a claim about a population.

b. An observational study aimed at establishing an association between two variables.

c. An experiment that seeks to establish a cause-and-effect relationship between two variables.

Quiz: Identify the most appropriate description of this study

Environmental scientists studied whether office workers' stress levels could be reduced by access to green spaces. They recruited employees from a large company, measured their proximity to parks and natural areas using GPS, and recorded their stress levels through surveys. The researchers analyzed whether employees living closer to green spaces reported lower stress. Their findings were published in the Journal of Environmental Psychology.

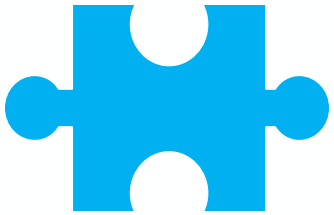
a. An observational study designed to estimate or make a claim about a population.

b. An observational study aimed at establishing an association between two variables.

c. An experiment that seeks to establish a cause-and-effect relationship between two variables.

Preliminary plan of topics you will learn with varying levels of depths

- Block 1: Introduction to the course, Data Types, and Python Basics
- Block 2: Descriptive Statistics & Exploratory Data Analysis (EDA)
- Block 3: Probability Theory and Conditional Probability
- Block 4: Random Variables and Statistical Distributions (Binomial, Normal)
- Block 5: Sampling Distributions & Central Limit Theorem
- Block 6: Confidence Intervals
- Block 7: Hypothesis Testing
- Block 8: Applied Data Cleaning & Preparation
- Block 9: Data Visualization and Storytelling with Data
- Block 10: Causality vs. Correlation
- Block 11: Other selected topics of Data Analysis
- Block 12: Wrap-up, Integration, and Discussion



Questions

Answers