FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Introduction to Data Mining

Prof. Dr. Matteo Marouf

08.04.2025

slido

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

How would your rate your knowledge in topics like AI, Data Science and Data mining?

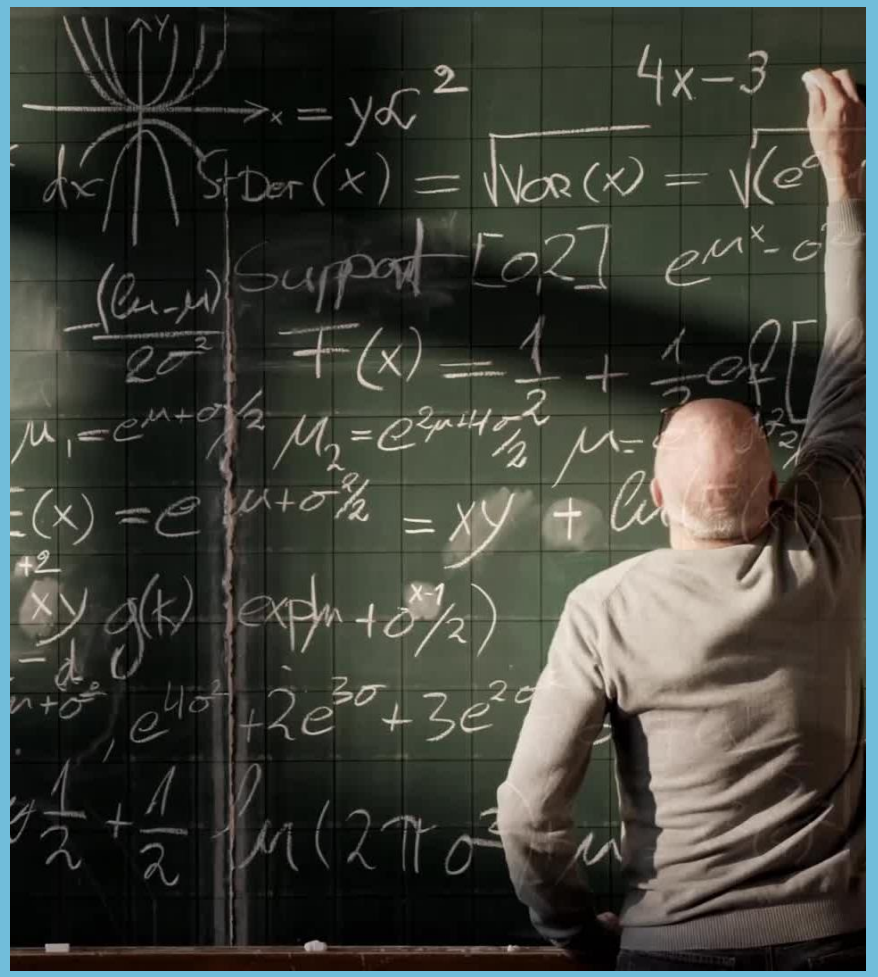
ⓘ Start presenting to display the poll results on this slide.

If you have the right data, what would you like to discover?

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Who is your professor?

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



About this module

Logistics about this model

- Read through the instructions provided on CampUAS page carefully.
- **Important: Join an exercise Group by 15.4.2025 based on your program:**
  - **Monday** (11:45-13:15 & 14:15-15:45): For HIS M.Sc. students who started in Winter Semester 2024/2025.
  - **Tuesday** (10:00-11:30& 11:45-13:15): For HIS M.Sc. students starting in Summer Semester 2025.
  - **Monday** (14:15-15:45): For Allgemeine Informatik – Master students.
  - ~~Once you choose a group, switching is not allowed, so pick carefully!~~
  - I've opened the groups choice, to allow switching groups. Only eligible students are allowed to join exercise groups.
- **80% active participation** in exercises (**Übungen**) is mandatory for exam eligibility.
- Attending **all lectures** is strongly recommended for exam preparation.
- We will program with Python and use common ML, AI and Data Mining Python frameworks.
- We are a large group, so please avoid emails with execuses and personal requests.
- Stay silent unless you have a question and mute your smartphone.
- Lectures start at 8.15, please arrive on time
- The module is conducted in English.

About this module

Pre-read

- **Mathematical Notations:**
  - Vectors & Matrices: Matrix multiplication, vector norms, eigenvalues.
  - Derivatives, Gradients: Single-variable and multivariable derivatives, Gradient vectors and Jacobian matrix.
  - Exponential & Logarithmic Functions
  - Basic Statistics: Bayes' Theorem, Mean, variance, probability distribution
  - Covariance Matrix
  - Dimensionality Reduction with PCA

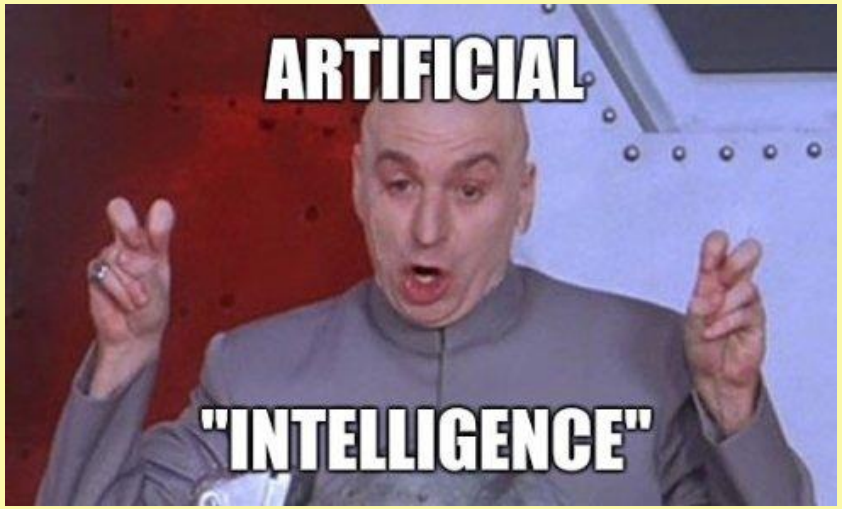
- **Key Machine Learning Terminology:** Model, Backpropagation, preprocessing, Feature, Target, training, performance evaluation, Overfitting/Underfitting.
- Introduction to Python programming if you are not familiar with Python

- **Further Reads:**

References will be provided after each lecture. Students are encouraged to gain hands-on experience by tackling AI and Data Science challenges on platforms like Kaggle.

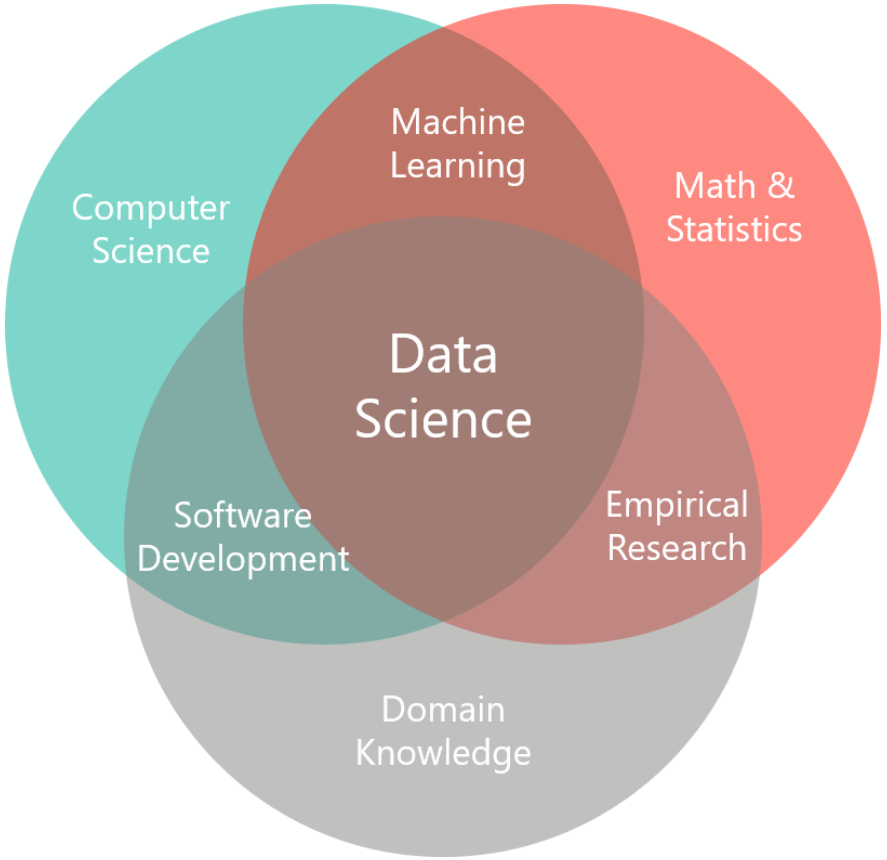
How do you pass this module successfully?

What is data mining and how did it all start?

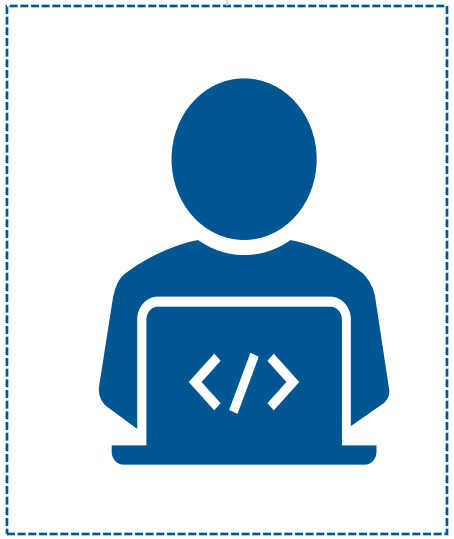


Data Science, data mining, and Big data analytics

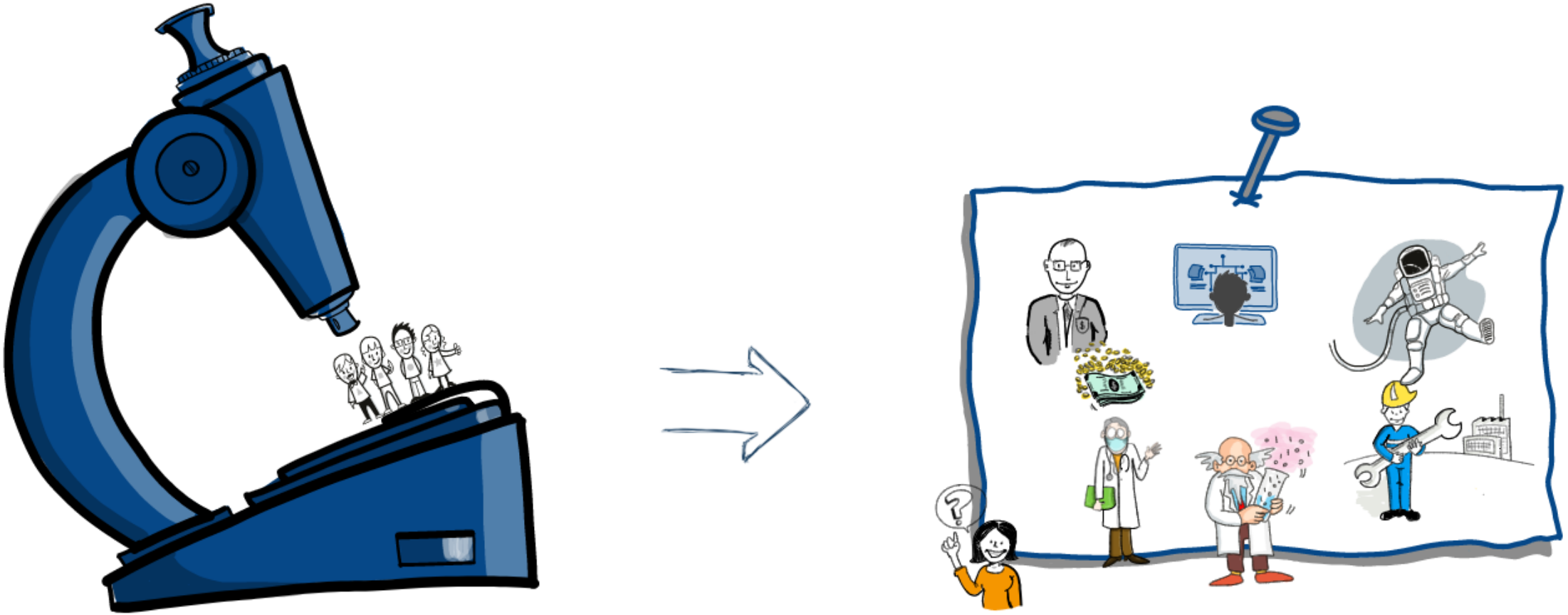
Data mining is an ***interdisciplinary*** field that uses **scientific methods, processes, algorithms** and **systems** to extract ***knowledge and insights*** from structured and unstructured data across a broad range of **application domains**.



Example: Data Science and Big data analytics



Data Mining is interdisciplinary



Wide range of applications rely on data mining (data science)

Manufacturing & Industry
- Predictive maintenance
- Quality control
- Supply chain optimization

E-commerce & Retail
- Recommendation systems
- Dynamic pricing
- Inventory management

Social Media & Internet
- Spam filtering
- Fake news detection
- Influencer identification

Cybersecurity
- Intrusion detection
- Malware detection
- Identity theft prevention

Science & Research
- Genomics & bioinformatics
- Climate change analysis
- Astronomy

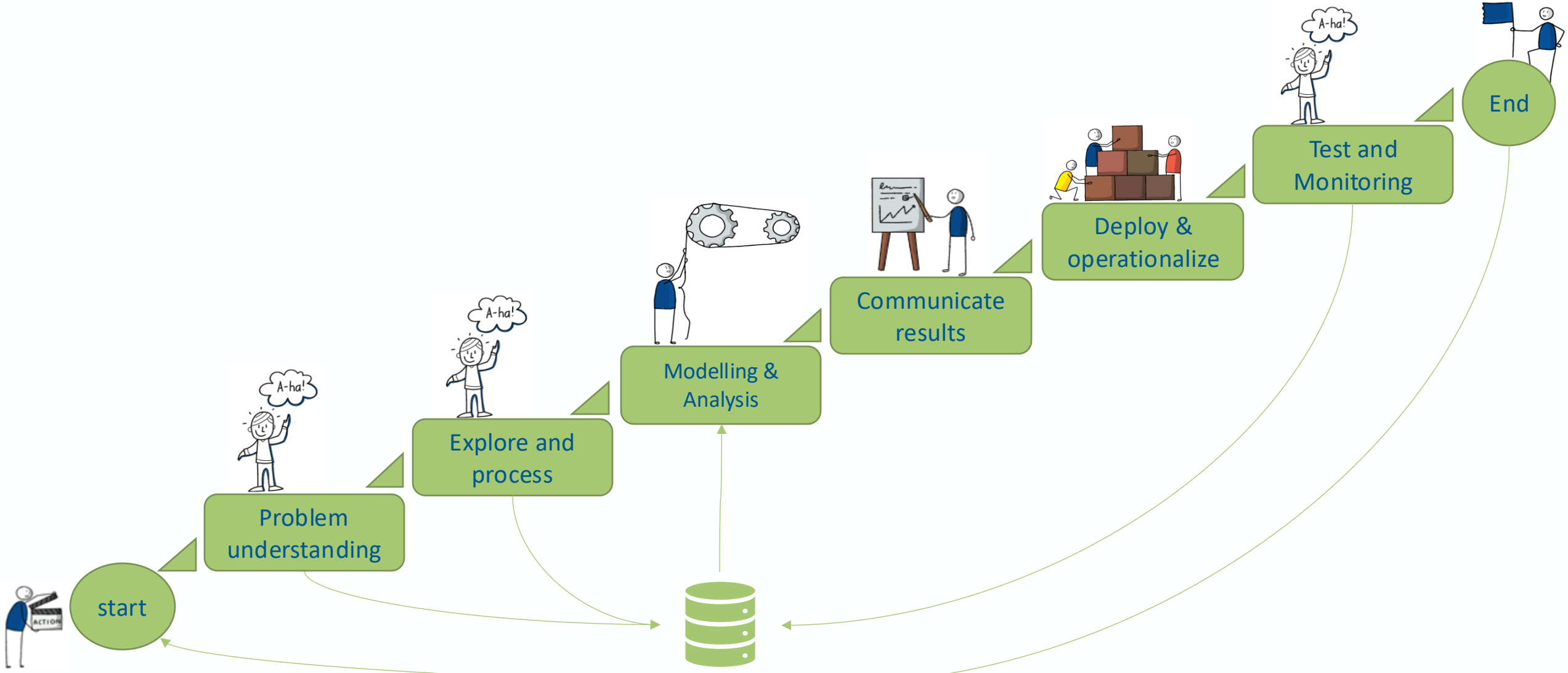
Business & Marketing
- Process optimization
- Customer segmentation
- Churn prediction
- Sentiment analysis

Finance & Banking
- Fraud detection
- Risk management
- Algorithmic trading

Healthcare & Medicine
- Disease prediction & diagnosis
- Drug discovery
- Patient monitoring

Smart Cities & Transportation
- Traffic prediction
- Route optimization
- Accident analysis

Data Mining project life cycle may involve all or some of these stages



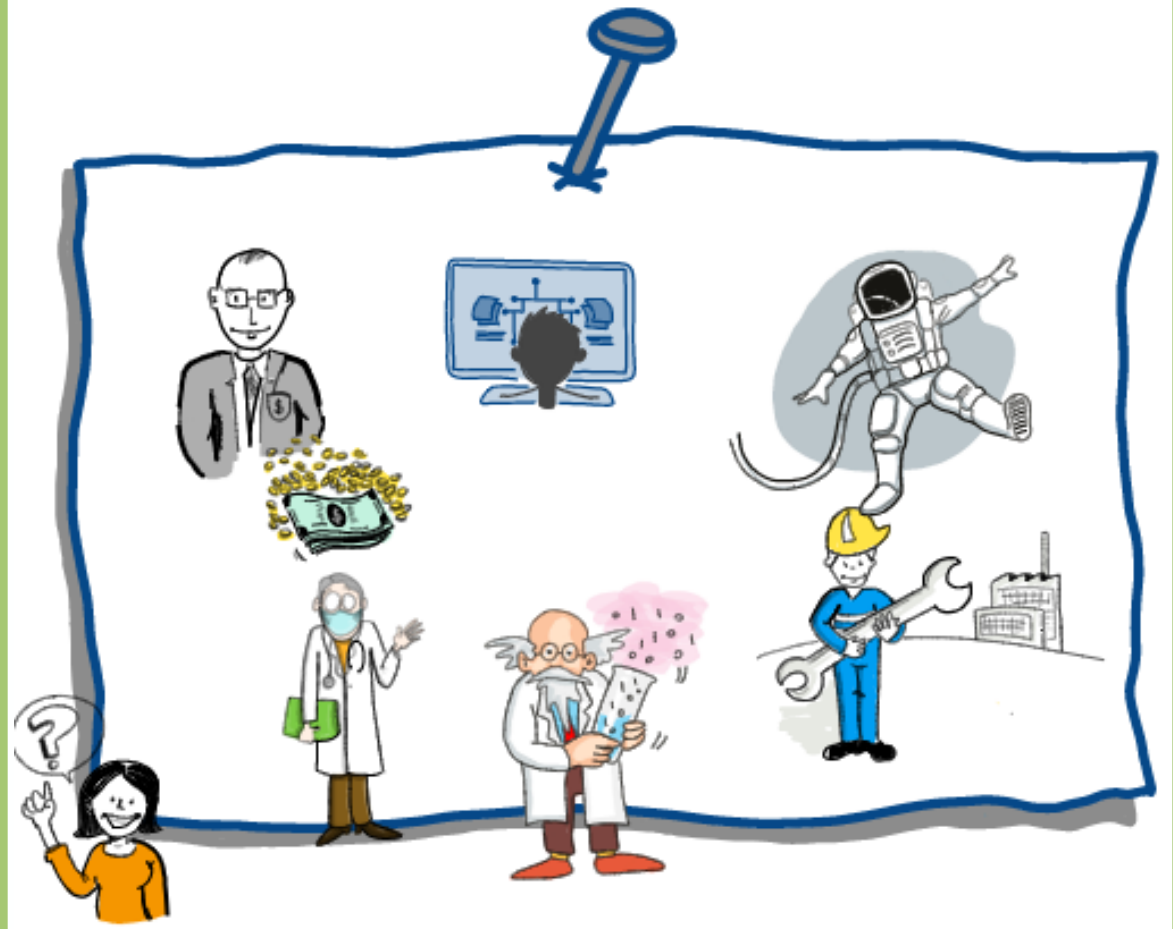
Domain knowledge plays a crucial role in successful data mining project

Domain Knowledge:

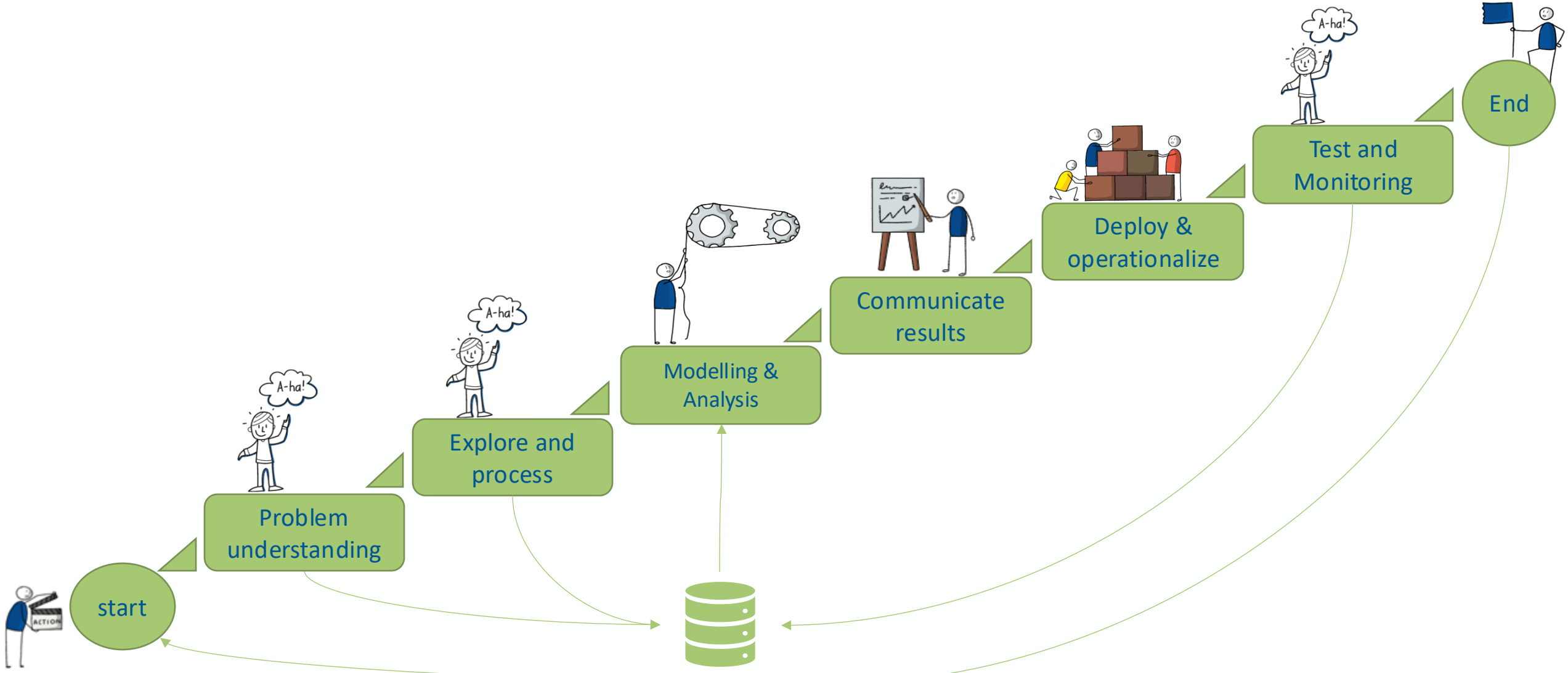
- guides the successful preprocessing
- Influnces the modeling
- Provides the the context to explain the results

The key takeaway

Always dive into the domain behind the data — it's the key to engineering meaningful features, focusing your analysis, and telling a story your audience understands.

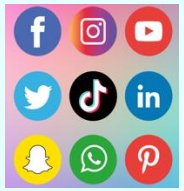


Data Mining project life cycle may involve all or some of these stages



Informative Data representation is a fundamental step in Data Mining

Data exist in different formats



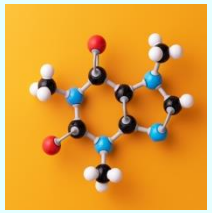
Social media

Imaging

EHR

Genomics

Knowledge 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

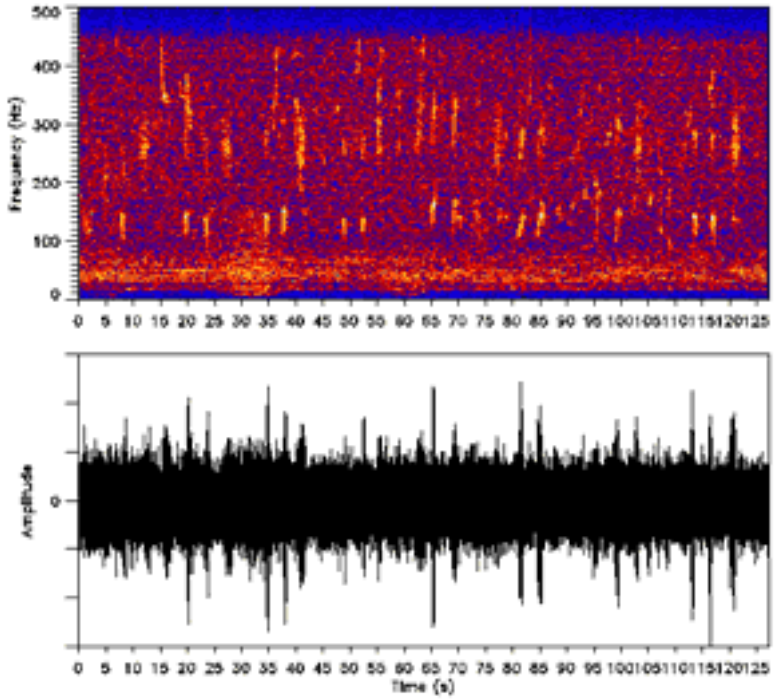
To train a model, we must create an **informative and numerical representation of the data** that can be manipulated by electronic devices.

Examples

- **Textual data:** Word Embedding is a term used for the representation of words for text analysis, typically in real valued vectors.
- **Images**: Pixel values are used.
- **Molecular structure:** Could be represented as mathematical graph where each atom is a node, and each bond is an edge.

Examples on finding a proper Data representation

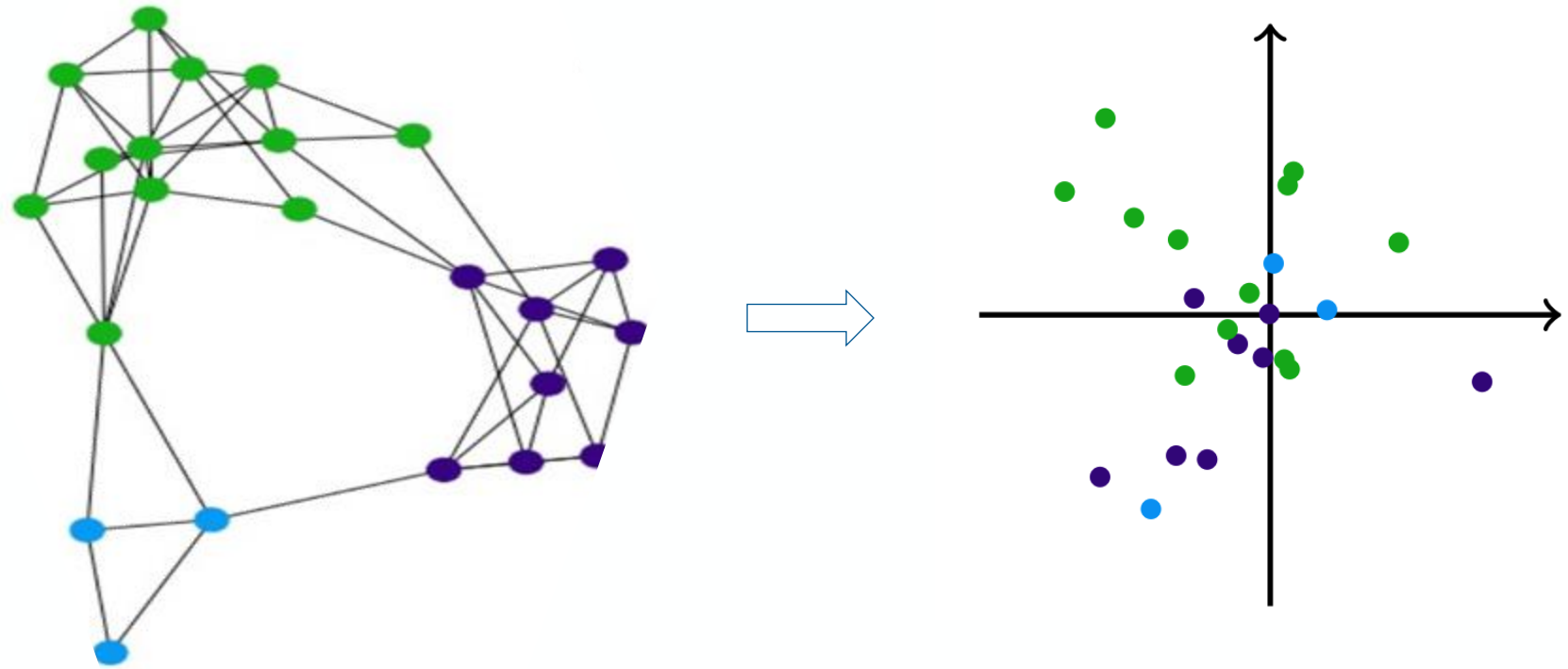
Spectrograms: An audio signal can be transformed into a spectrogram, representing the signal's frequency spectrum over time. This visual representation is useful for speech and music analysis.



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

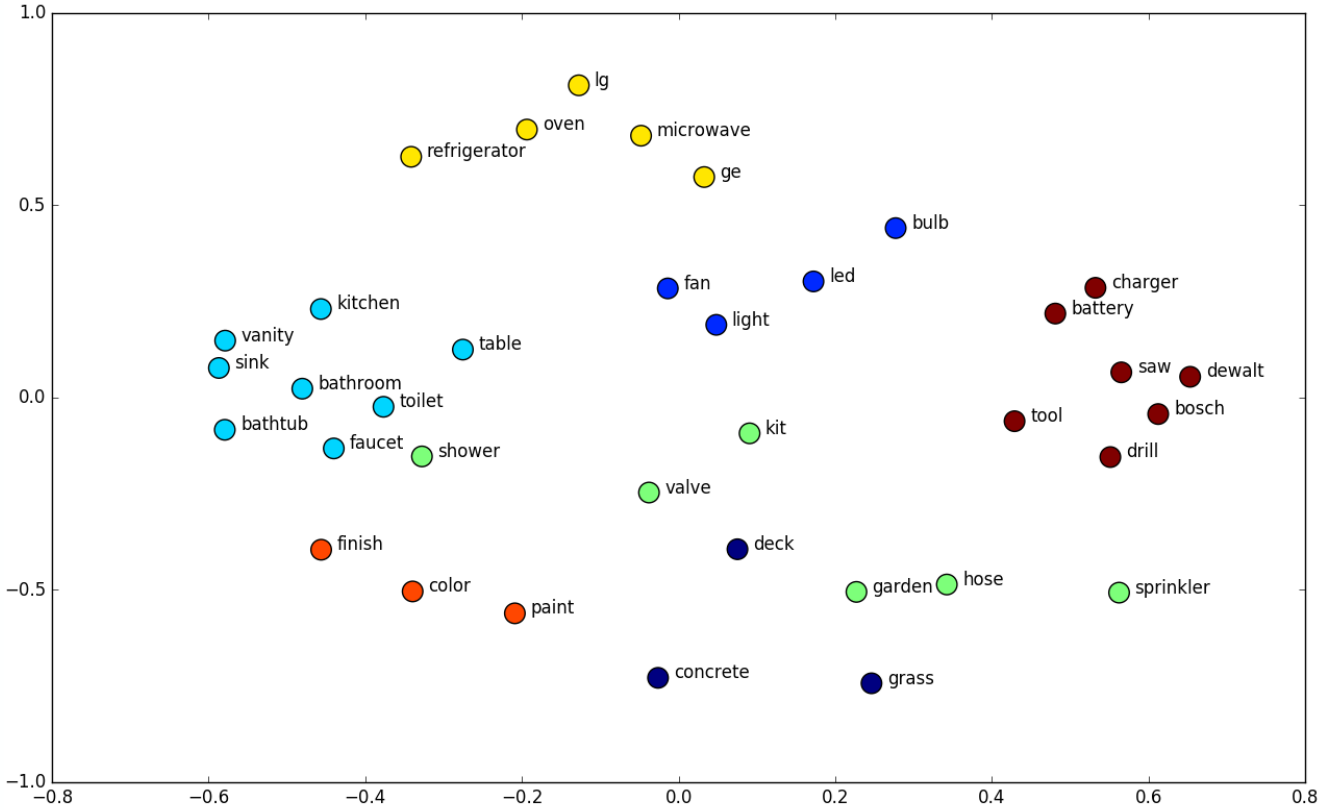
Frequency of a word withing a document

IDF

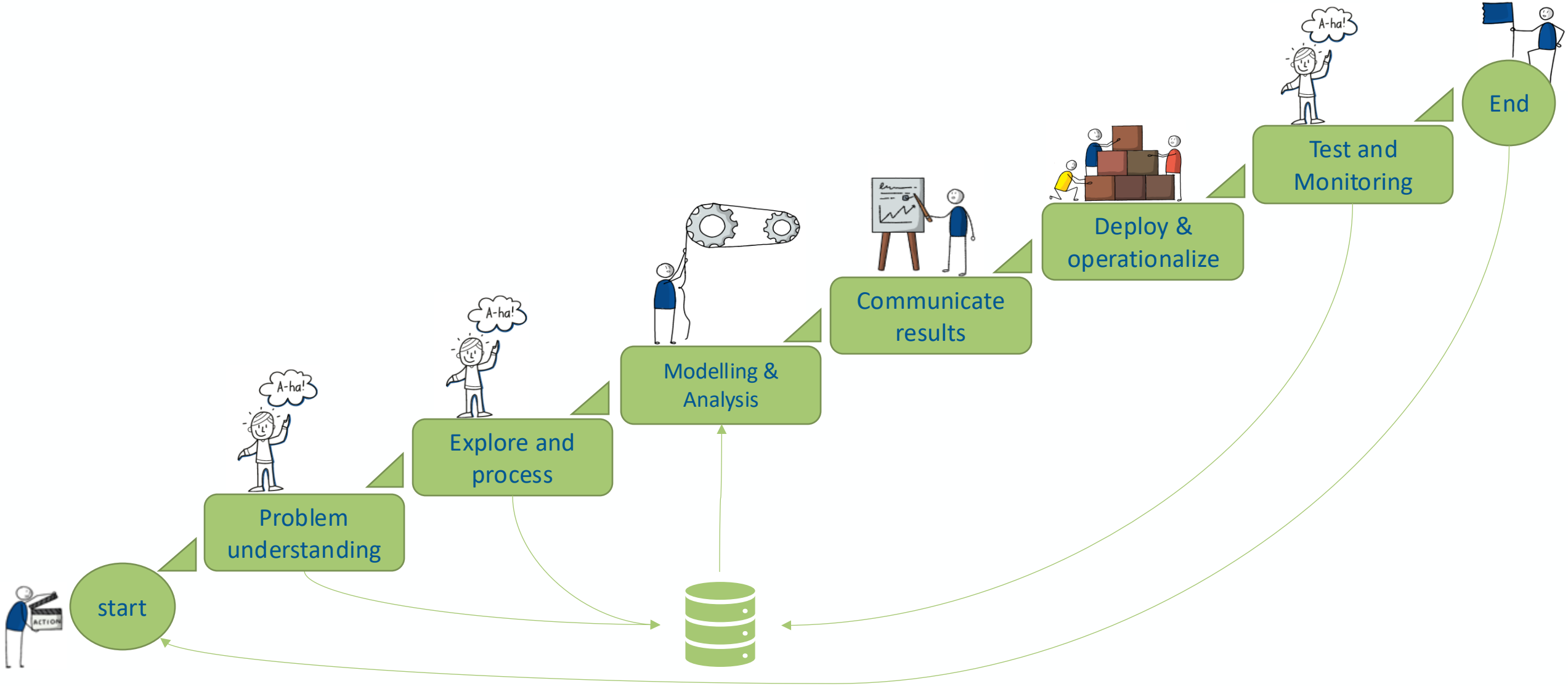
Frequency of a word across the corpus

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.



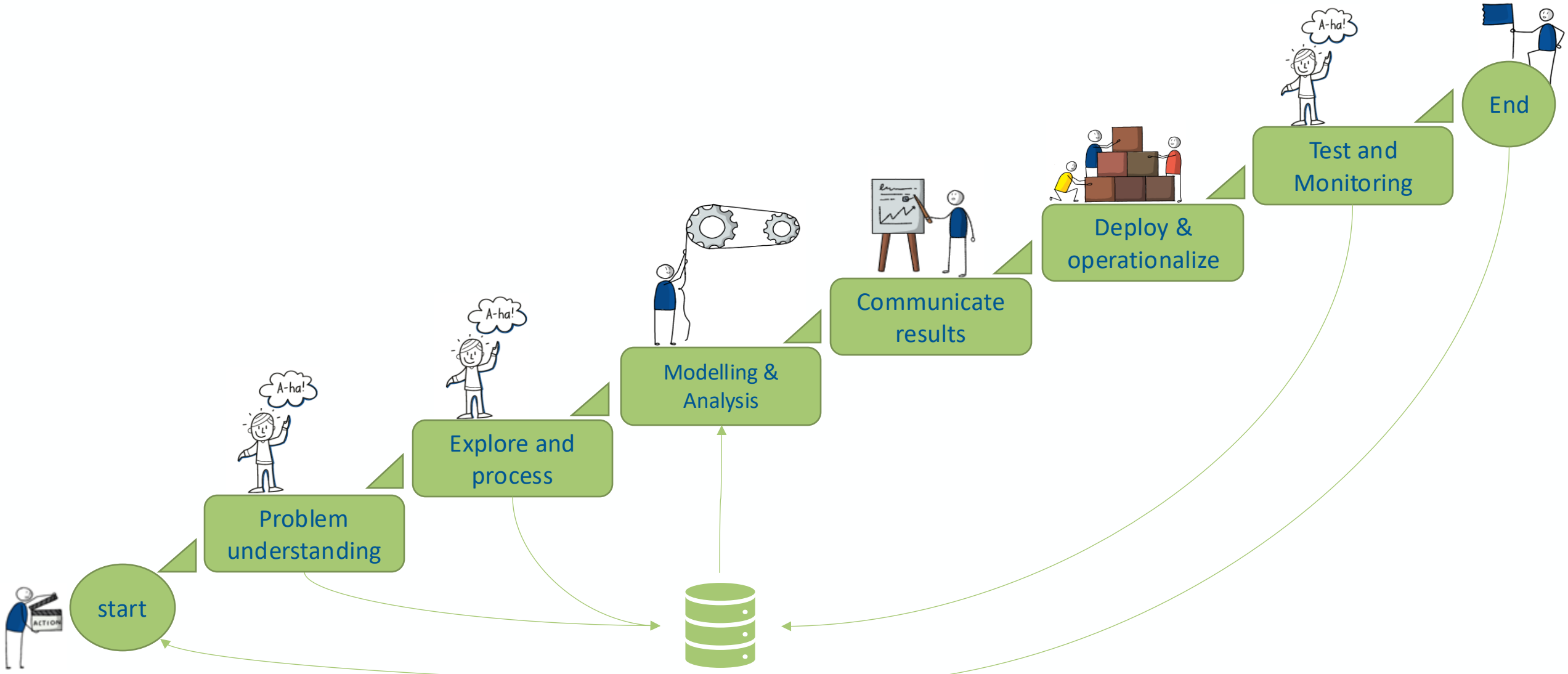
Data Science project life cycle



Main topics you may stumble across while working in this field

Classification:	Sorting data into predefined categories (e.g., spam detection).
Association Analysis:	identifying relationships between items in large datasets
Regression:	Predicting continuous values (e.g., stock prices, recovery time for hospitalized patients).
Clustering:	Grouping data based on similarity without predefined labels (e.g., customer segmentation).
Optimization:	AI often finds the best solutions to problems within constraints (e.g., route planning, resource allocation).
Recommendation:	Suggesting items or content based on user preferences (e.g., Netflix or Amazon recommendation engines).
Anomaly Detection:	Identifying unusual patterns or outliers in data (e.g., fraud detection, fault prediction).
Personalization:	Tailoring user experiences or interactions (e.g., personalized marketing, adaptive learning systems).
Decision Support:	Inform decision-making by analyzing large datasets and providing insights (e.g., medical diagnostics, business intelligence).
Simulation:	model complex systems or environments to predict behavior and outcomes (e.g., weather modeling, financial market).
Text mining	Understanding, mining, and generating human language (e.g., sentiment analysis, language translation).

Data Mining project life cycle may involve all or some of these stages



Story telling is a crucial skill when presenting and communicating data mining results

Key components:

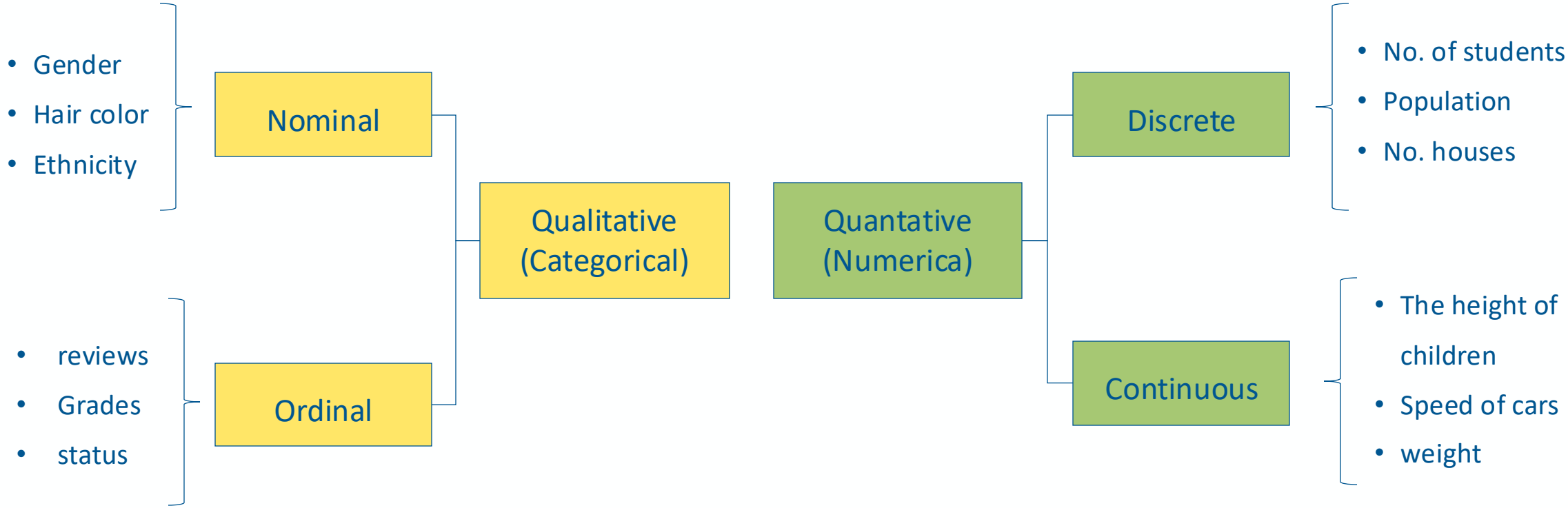
- **Data** – the evidence or facts
- **Visualization** – charts, graphs, dashboards, etc.
- **Narrative** – the storyline that guides interpretation (what happened, why, what now?)

Common types of visualizations used in business presentation

Story telling with data, Cole Nussbaumer Knaflic

Which data types can we analyze with data mining?

Types of Data



Types of Data

Structured Data

- Organized data that adheres to a predefined schema, typically stored in relational databases.
- **Examples:** Customer information tables, transaction records.

Unstructured Data

- Data without a predefined format, making it more challenging to collect, process, and analyze.
- **Examples:** Emails, videos, social media posts.

Semi-Structured Data

- Data that does not reside in a relational database but still has some organizational properties, such as tags or markers to separate data elements.
- **Examples:** JSON files, XML documents, graph data

What will we use to analyze and mine data?

Development Frameworks for bold and muscular data miners

Category	Python-Based
Other essential packages	Numpy, Pandas, Matplotloib, Seaborn
Machine Learning	Scikit-learn, XGBoost, RAPIDS
Big Data Analytics	Apache Spark (PySpark)
Large Language Models	Hugging Face Transformers, OpenAI GPT
Deep Learning	TensorFlow, PyTorch
Hyperparameter Optimization	Optuna, Hyperopt, GridSearchCV

Non-Python alternatives: Matlab, RapidMiner, R,

So, what will we address throughout this course?

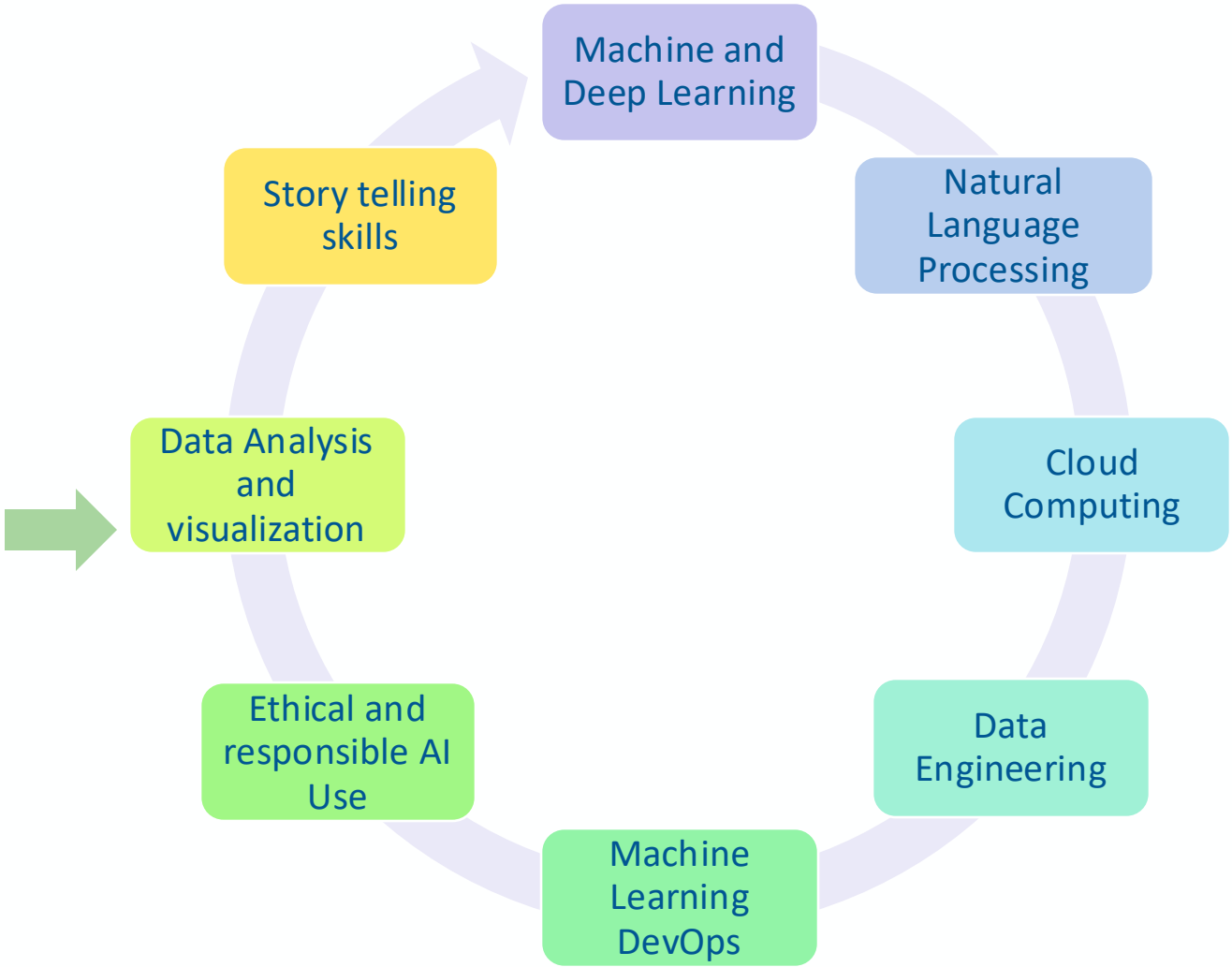
What is needed for successful AI related career as a computer scientist?

1

Fundamentals computer science, Math, and programming subjects

2

Building a solid foundation in one of more of the below listed fields.



3

Deepen your knowledge in an area you are passionate about, follow emerging technologies and keep an eye on job market trends

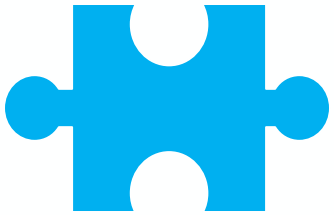
What will you learn?

- Sprint 1+2: Basic Data Mining Practice and data science pipeline
- Sprint 4: Clustering Analysis
- Sprint 5: Classification Analysis
- Sprint 6: Storytelling, Features Importance and Explainability
- Sprint 7: Regression Analysis
- Sprint 8: From Regression to a Neural Network
- Sprint 9: other forms of Data Mining
- Sprint 10: Mining Sequence Data
- Sprint 11: GenAI for Data Mining: Retrieval Augmented Generation
- Sprint 12: Mining Big Data
- Sprint 13: Ethical and Responsible Data Mining
- Sprint 14: Wrap-up & discussion

I want to commence a career in this field what are my options afterwards?

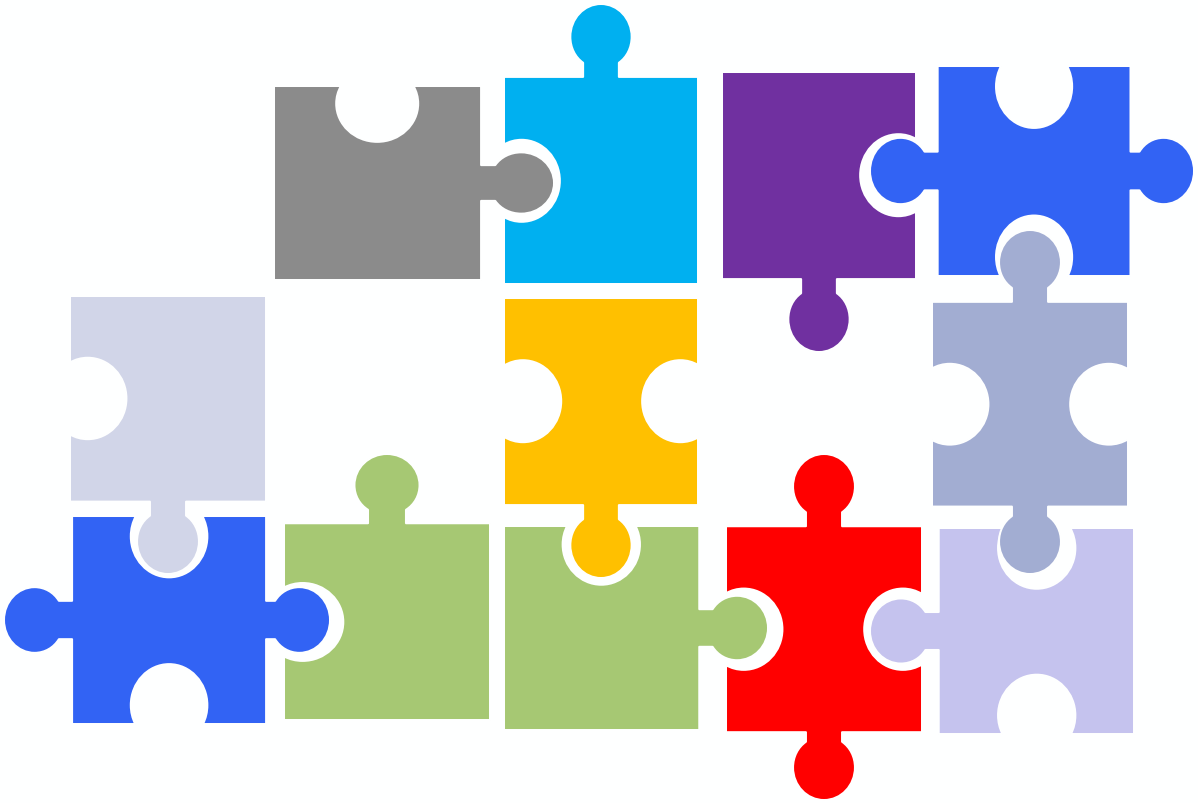
- AI Engineer
- Data Analyst
- Data scientist
- Robotics and Automation expert
- Machine Learning DevOps (ML Engineer)
- Data Engineer
- Software Engineer
- Whatever!

FRANKFURT UNIVERSITY OF APPLIED SCIENCES



Questions

Answers



FRANKFURT UNIVERSITY OF APPLIED SCIENCES