FH KREMS
UNIVERSITY OF APPLIED
SCIENCES / AUSTRIA

Machine Learning, Artificial Intelligence, and Big Data Analytics (IL, 4th Semester)

Lecture 1

Agenda (Lecture 1)

- About your teacher
- Introduction round
- Program of the Course
- Organizational Information

- Intro to Data Science (AI, ML, and Big Data)
- Supervised vs. Unsupervised Learning
- The Data Science process
- Recap of R

About me (Danilo)

- Born in Messina, Italy
- Studies
  - BSc Engineering at TU Bari
  - MSc Engineering at FH Technikum-Wien
  - PhD Computer Science at Uni Wien
- Work
  - Researcher @ Telecommunication Research Center Vienna (FTW)
  - Lecturer “Automotive Telecommunications” @ FH Technikum-Wien
  - Senior Research Scientist at Siemens AG Austria
- Interests
  - Technology, Innovation, and Trends; Data Science; AI ethics; Human Rights
- Private
  - 2 kids, play guitar, like biking/swimming, play videogames

Typical projects

Customer and Sales Analytics

Cross/up-selling, lead generation, pricing, customer behavior, …

Urban and Building analytics

Smart sensing, energy optimization, renewables, sustainability…

DSS for Product Life Cycle

Supply-chain diagnostic, product configuration, …

Industrial AI

Shopfloor monitoring, predictive maintenance, production optimization, …

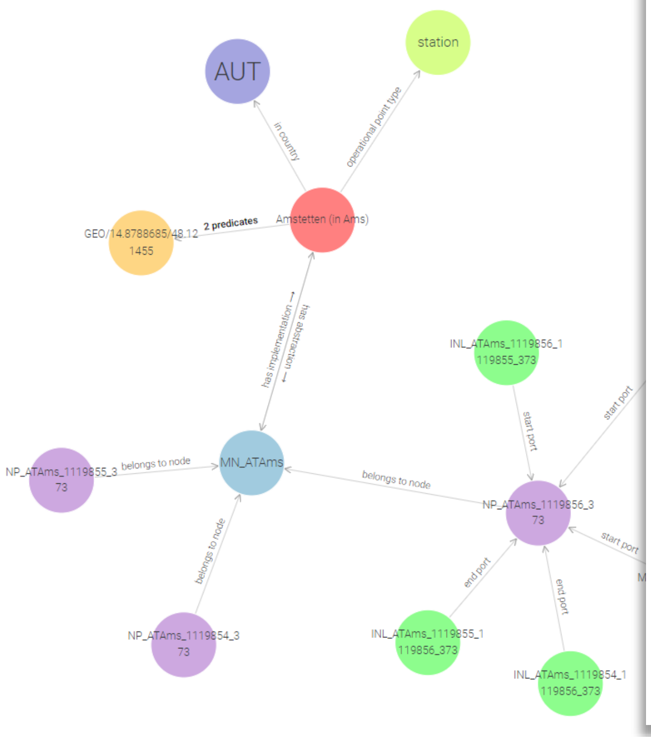
Research topics of interest

- Predictive Machine Learning in all its flavors (statistical learning, sub-symbolic learning, …)
- Unsupervised learning, pattern recognition, anomaly detection, clustering, …
- Recommender systems, collaborative filtering, content-based filtering, …
- Visual analytics and UX, Explainable AI (XAI), Interpretable Machine Learning, …
- (new) Neural-Symbolic AI.

About me (Stefan)

- Born in Vorarlberg
- Studies
  - BSc Software & Information Engineering @ TU Vienna
  - MSc Information & Knowledge Management @ TU Vienna
  - PhD Computer Science @ TU Vienna
- Work
  - Researcher @ DERI Galway, Ireland
  - Project Assistant @ TU Vienna and WU Vienna
  - Research Scientist @ Siemens AG Austria
- Interests
  - Knowledge Graphs, AI
- Private
  - 2 kids

Typical projects



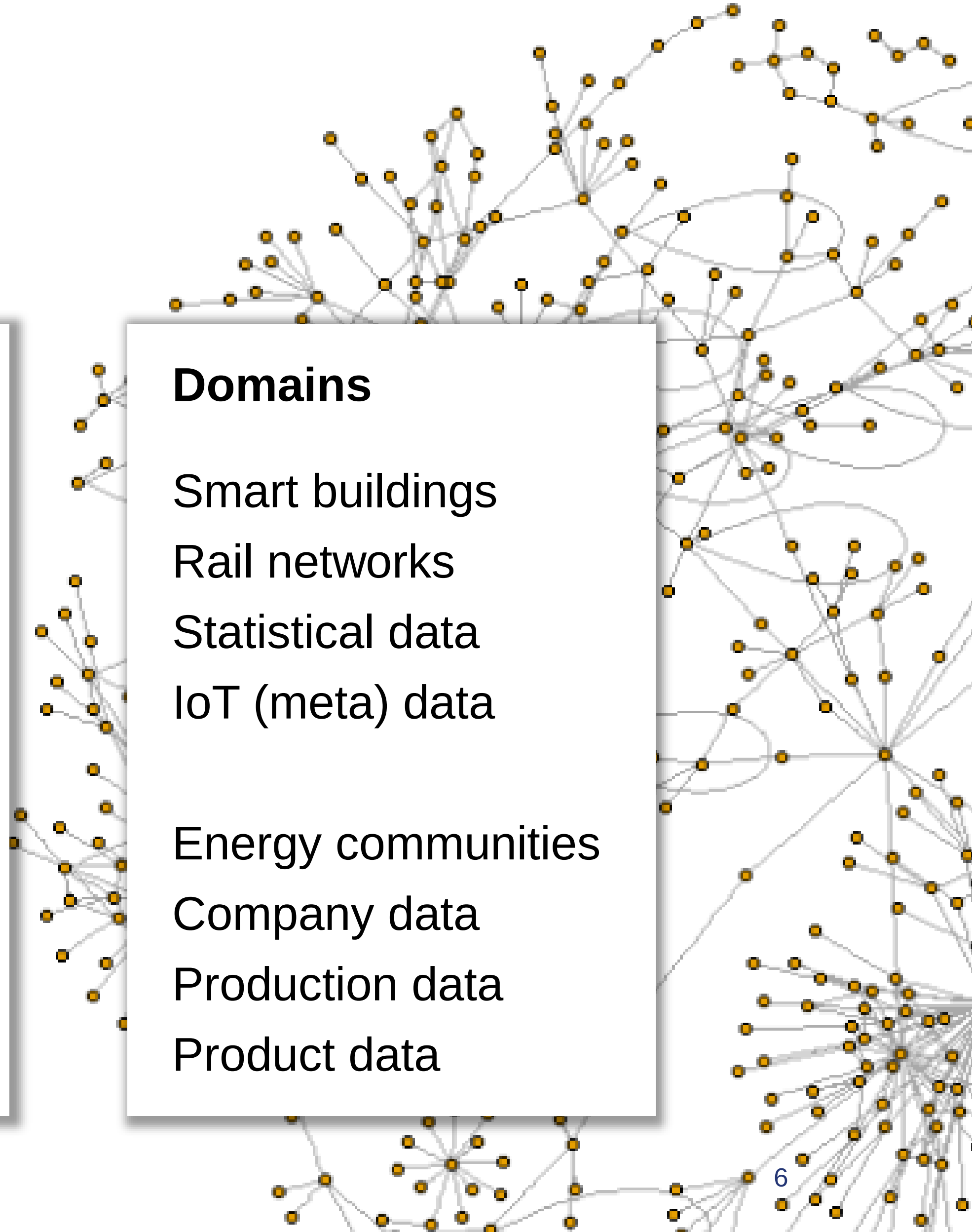
Semantic Technologies and KGs

- Data integration and enrichment
- Reasoning (existential rules)
- Data modelling and data mgmt.
- SPARQL query rewriting
- Declarative calculation
- Combination with config. tech.
- Graph machine learning
- Neuro-symbolic AI
- Rapid prototyping

Domains

- Smart buildings
- Rail networks
- Statistical data
- IoT (meta) data

- Energy communities
- Company data
- Production data
- Product data



About you

- Anything you want to share about you, e.g.
  - Where are you from?
  - Why did you join this BSc program?
  - What are your interests?
  - Do you work in parallel to your studies?
  - …

Setting the baseline

- Scan the barcode from your mobile phone

OR

- go to http://sli.do and insert this code: 30191

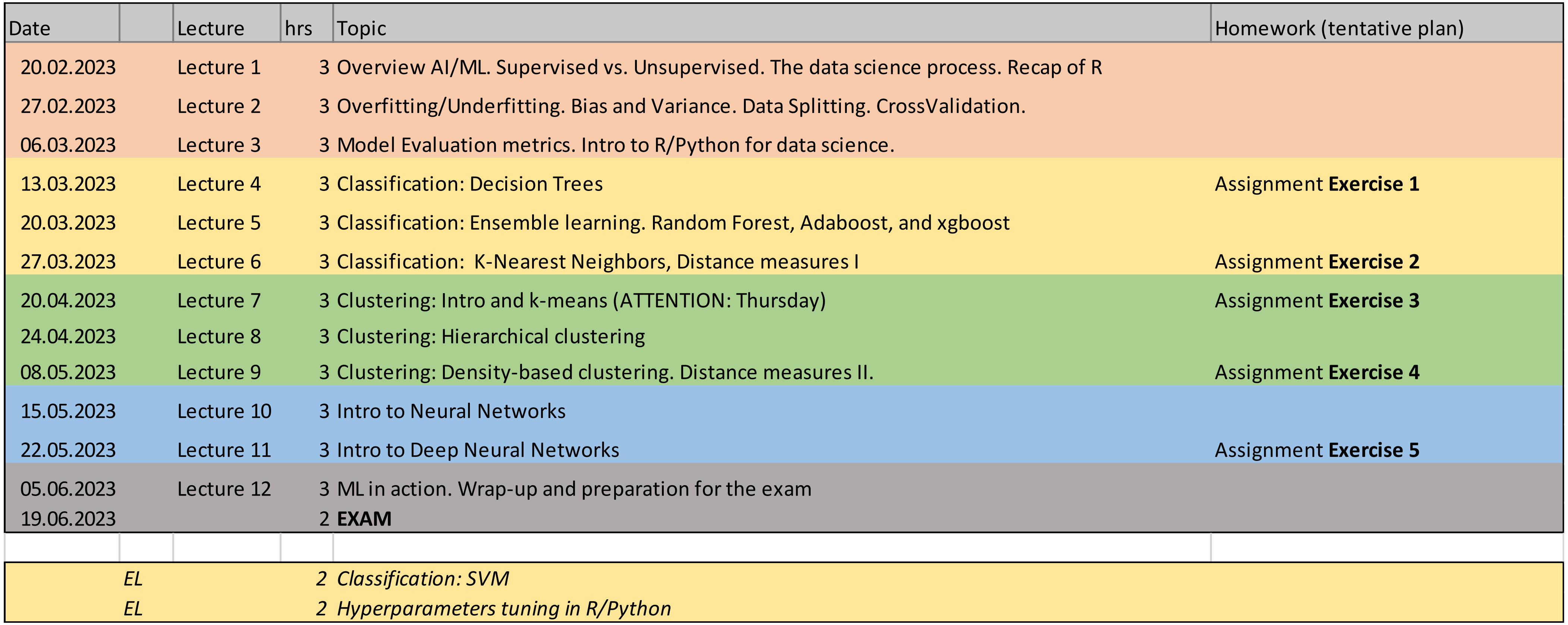
and follow my instructions.

Program of the Course

- Recap and Introduction
- Module 1: Supervised Learning
- Module 2: Unsupervised Learning
- Module 3: Neural networks
- Practical hints, examples, and how to manage a ML project

- EXAM

Detailed Plan



Date		Lecture	hrs	Topic	Homework (tentative plan)
20.02.2023		Lecture 1	3	Overview AI/ML. Supervised vs. Unsupervised. The data science process. Recap of R	
27.02.2023		Lecture 2	3	Overfitting/Underfitting. Bias and Variance. Data Splitting. CrossValidation.	
06.03.2023		Lecture 3	3	Model Evaluation metrics. Intro to R/Python for data science.	
13.03.2023		Lecture 4	3	Classification: Decision Trees	Assignment **Exercise 1**
20.03.2023		Lecture 5	3	Classification: Ensemble learning. Random Forest, Adaboost, and xgboost	
27.03.2023		Lecture 6	3	Classification: K-Nearest Neighbors, Distance measures I	Assignment **Exercise 2**
20.04.2023		Lecture 7	3	Clustering: Intro and k-means (ATTENTION: Thursday)	Assignment **Exercise 3**
24.04.2023		Lecture 8	3	Clustering: Hierarchical clustering	
08.05.2023		Lecture 9	3	Clustering: Density-based clustering. Distance measures II.	Assignment **Exercise 4**
15.05.2023		Lecture 10	3	Intro to Neural Networks	
22.05.2023		Lecture 11	3	Intro to Deep Neural Networks	Assignment **Exercise 5**
05.06.2023		Lecture 12	3	ML in action. Wrap-up and preparation for the exam	
19.06.2023			2	**EXAM**	
	EL		*2*	*Classification: SVM*	
	EL		*2*	*Hyperparameters tuning in R/Python*	

Assumption (pre-requisite)

- Basic knowledge of statistics
  - Probability, conditional probability, Bayes, …
  - Main distributions
  - Descriptive statistics (mean, median, mode, variance, …)

- Basic programming skills

- Curiosity to try out new things and learn from data

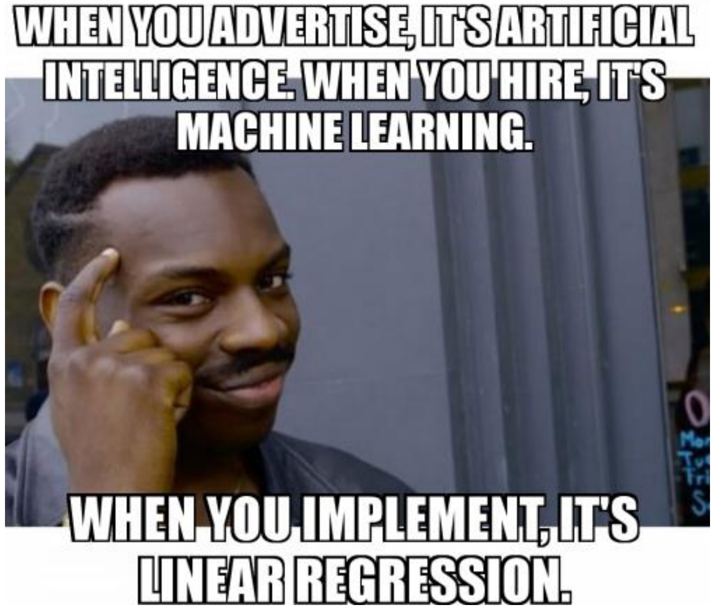
Evaluation

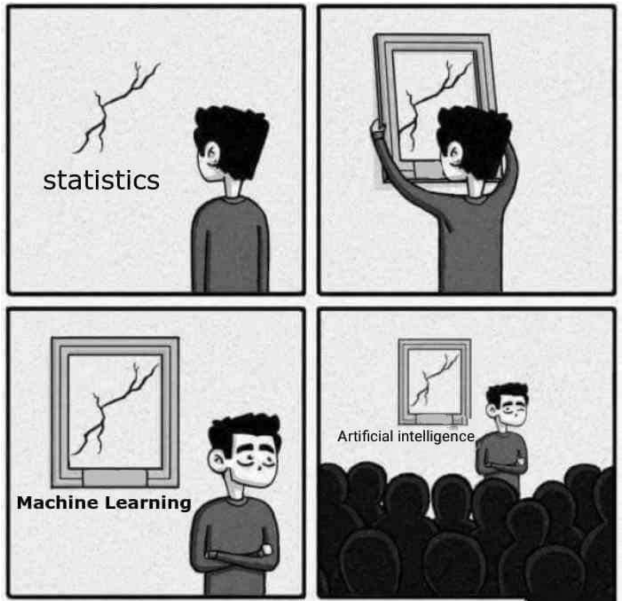
- 40% Exercise
  - Homework
    - 5 exercises
  - Class work
    - After important topics (mostly small coding exercise)
  - Bonus given to those who present their solution in class

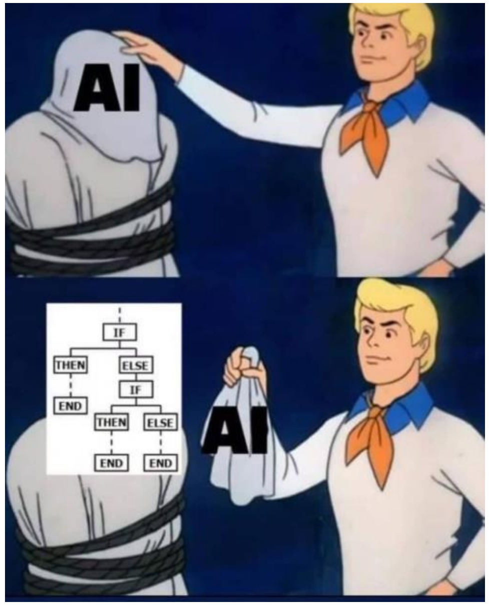
- 60% Written test

Introduction to data science

What is Artificial Intelligence

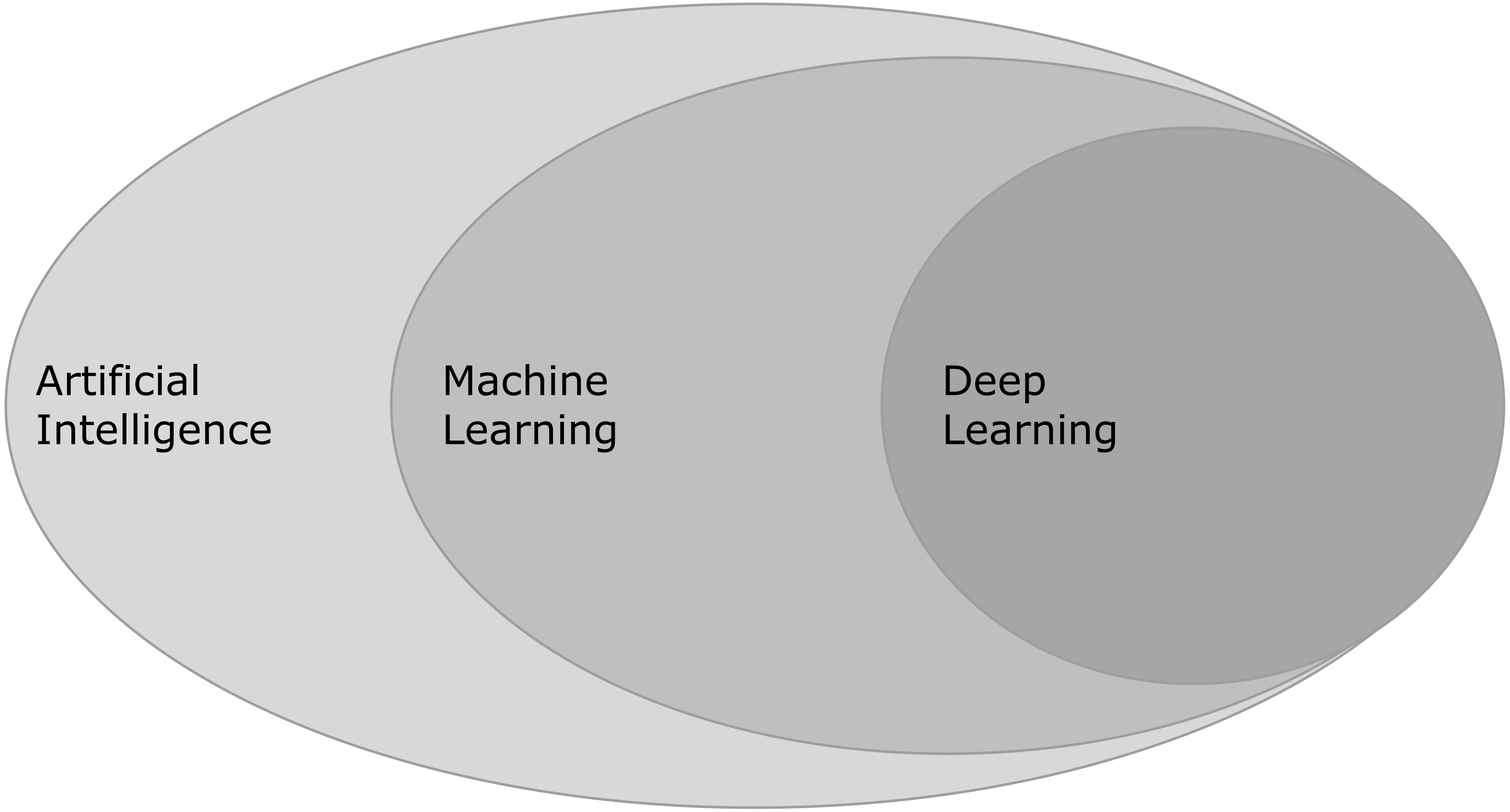

WHEN YOU ADVERTISE, IT'S ARTIFICIAL INTELLIGENCE. WHEN YOU HIRE, IT'S MACHINE LEARNING.
WHEN YOU IMPLEMENT, IT'S LINEAR REGRESSION.





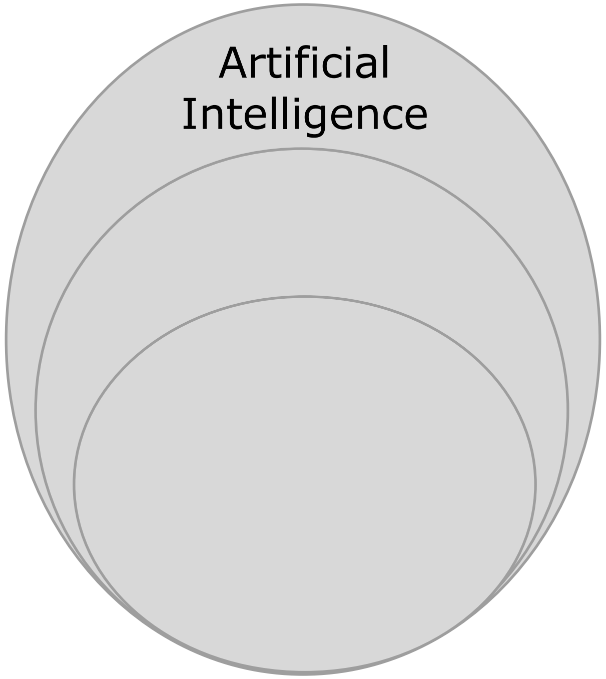
A term that has been misused too often…

What is Artificial Intelligence



Artificial Intelligence

Definition



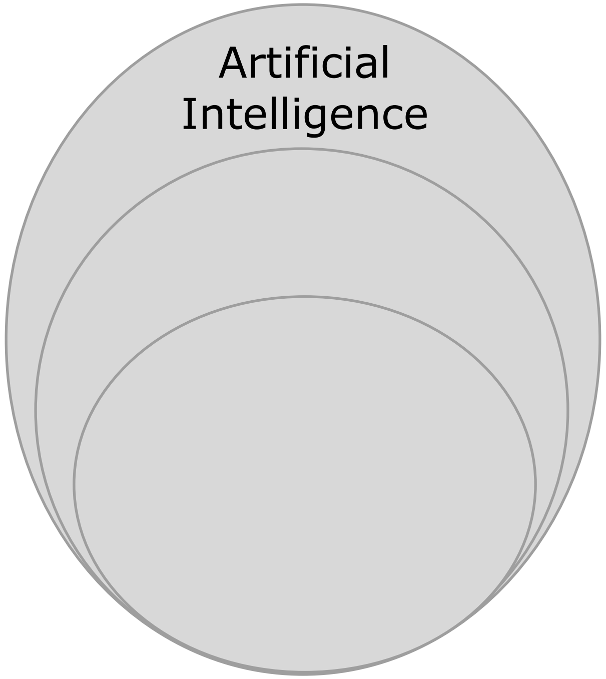
- *"The theory and development of computer systems **able to perform tasks normally requiring human intelligence**, such as visual perception, speech recognition, decision-making, and translation between languages."* (Oxford Living Dictionary)

- *"The ability of a digital computer or computer-controlled robot to **perform tasks commonly associated with intelligent beings**."* (Encyclopedia Britannica)

- *"The field of computer science dedicated to **solving cognitive problems commonly associated with human intelligence**, such as learning, problem solving, and pattern recognition."* (Amazon)

Artificial Intelligence

Types of AI



- **Symbolic AI** (aka classical AI or rule-based AI)

Relies on an explicit representation of a domain, hard-coded by humans as a set of symbols and rules. Uses deductive reasoning, logical inference, and other deterministic approaches to solve problems. Very useful in high-risk domains.

- **Non-Symbolic AI**

Tries to approach intelligence without specific representations of knowledge/domain. It learns autonomously by being *trained* with enough raw information to construct its own implicit knowledge. It can be further divided into sub-symbolic AI and statistical AI.

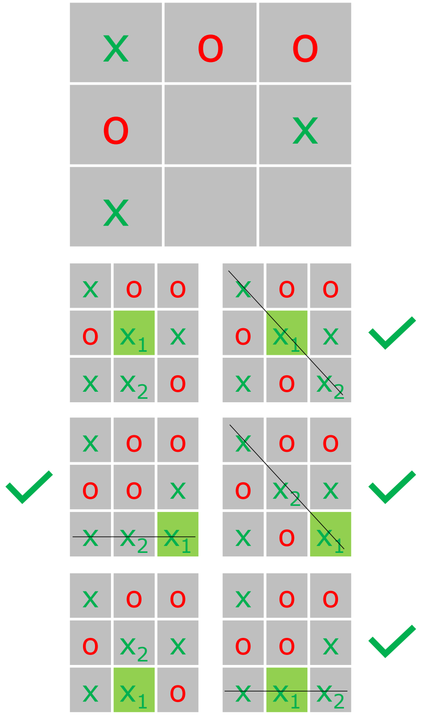
Artificial Intelligence

Types of AI (Example – TicTacToe)

X	O	O
O		X
X		

Artificial Intelligence

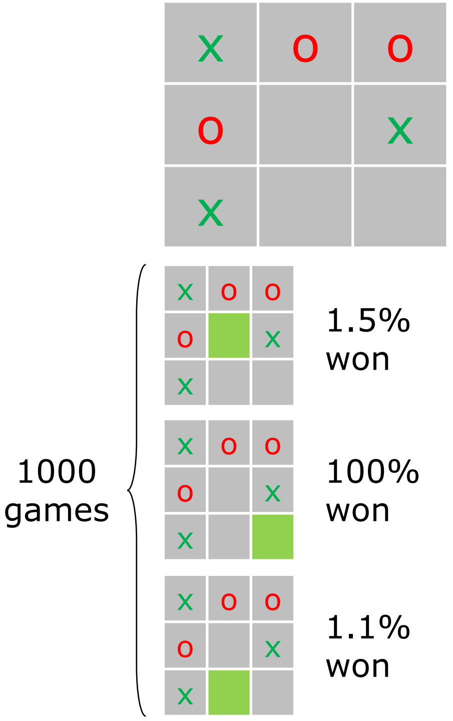
Types of AI (Example – TicTacToe)



- **Symbolic AI** (aka classical AI or rule-based AI)
  - Define and model the problem
  - Traverse possible solutions (e.g., search algorithm)
  - Deduct the best move
  - Challenge: Branching can make the possible solution space quickly increase requiring more sophisticated searches or modeling additional knowledge. (e.g. solution space for chess → 10^{120})

Artificial Intelligence

Types of AI (Example – TicTacToe)

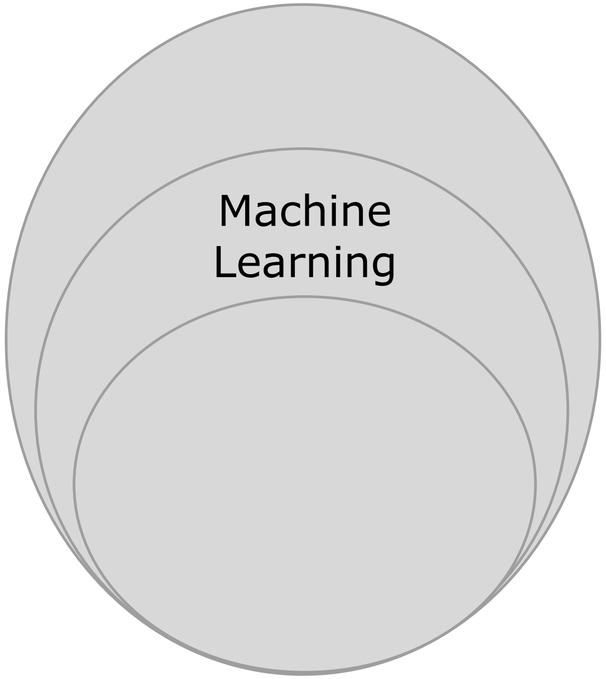


- **Symbolic AI** (aka classical AI or rule-based AI)
  - Define and model the problem
  - Traverse possible solutions (search algorithm)
  - Deduct the best move
  - Challenge: Branching can make the possible solution space quickly increase requiring more sophisticated searches or modeling additional knowledge. (e.g. solution space for chess → 10^{120})

- **Non-Symbolic AI**
  - Observe previously played games
  - Infer the move that leads to higher chance of success.
  - Challenge: The data about previously played games could not be sufficient to make an inference

Machine learning

Definition



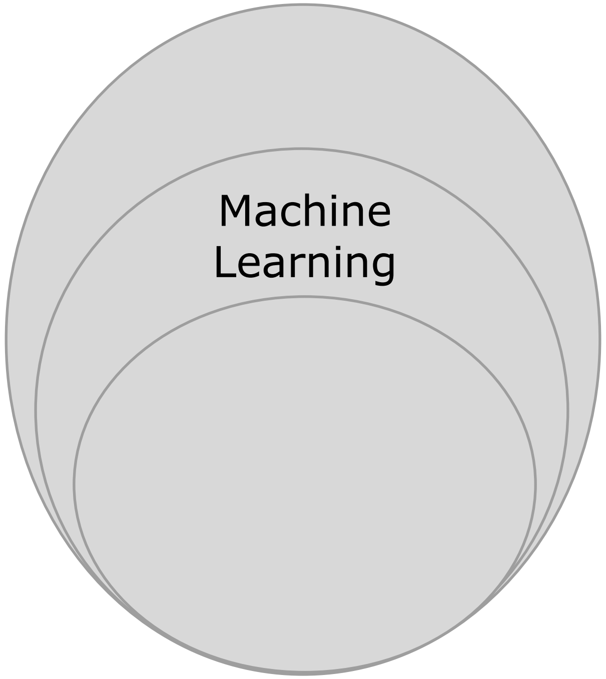
- *“the use and development of computer systems that are able* ***to learn and adapt without following explicit instructions****, by using algorithms and statistical models to analyze and draw inferences from patterns in data”* (Oxford living Dictionary)

- *“Machine learning (in AI) is a discipline concerned with the implementation of computer software that* ***can learn autonomously****”* (Encyclopedia Britannica)

- *“A subset of artificial intelligence (AI) that provides systems the ability to* ***automatically learn and improve from experience*** *without being explicitly programmed.”* (IBM)

Machine learning

What is “learning”



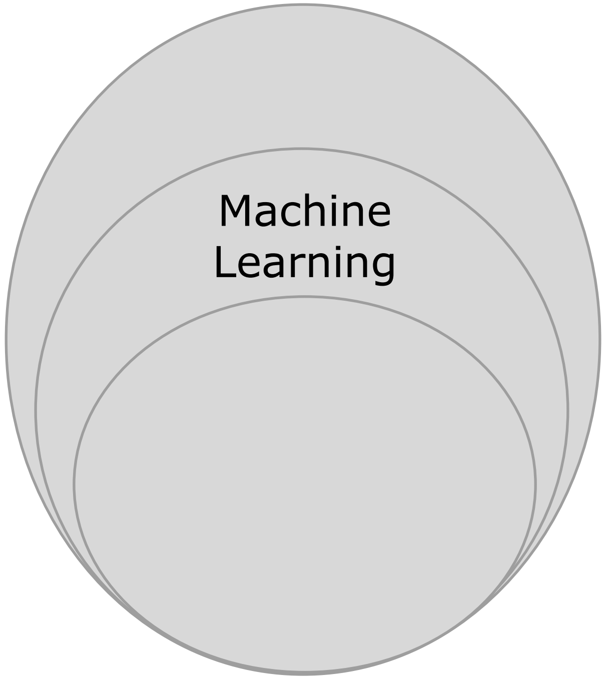
”A computer program is said to learn from experience ***E*** with respect to some class of tasks ***T*** and performance measure ***P*** if its performance at tasks in ***T***, as measured by ***P***, improves with experience ***E.****”*

Tom Mitchell (*Machine Learning*. McGraw Hill).

Learning on experience E over tasks T measured by performance P

Machine learning

Types of ML



- **Supervised Learning**: The machine is presented with a set of **input** data and **output** data and learns the relation between input and output such that it can predict the output for new input data.

- **Unsupervised Learning**: The machine is presented with **just input** data and autonomously search for structure and patterns within the input data.

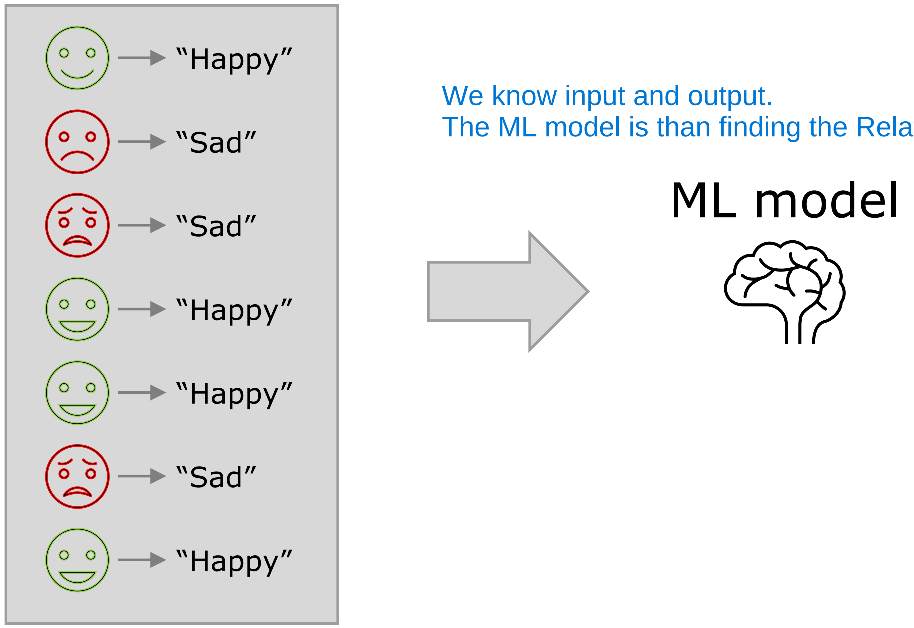
- **Semi-supervised Learning**: A mix of the above.

- **Reinforcement Learning**: The machine does not only learn from a static dataset but continuously learns through trial after receiving positive or negative feedback as reinforcement.

Machine learning

Supervised learning

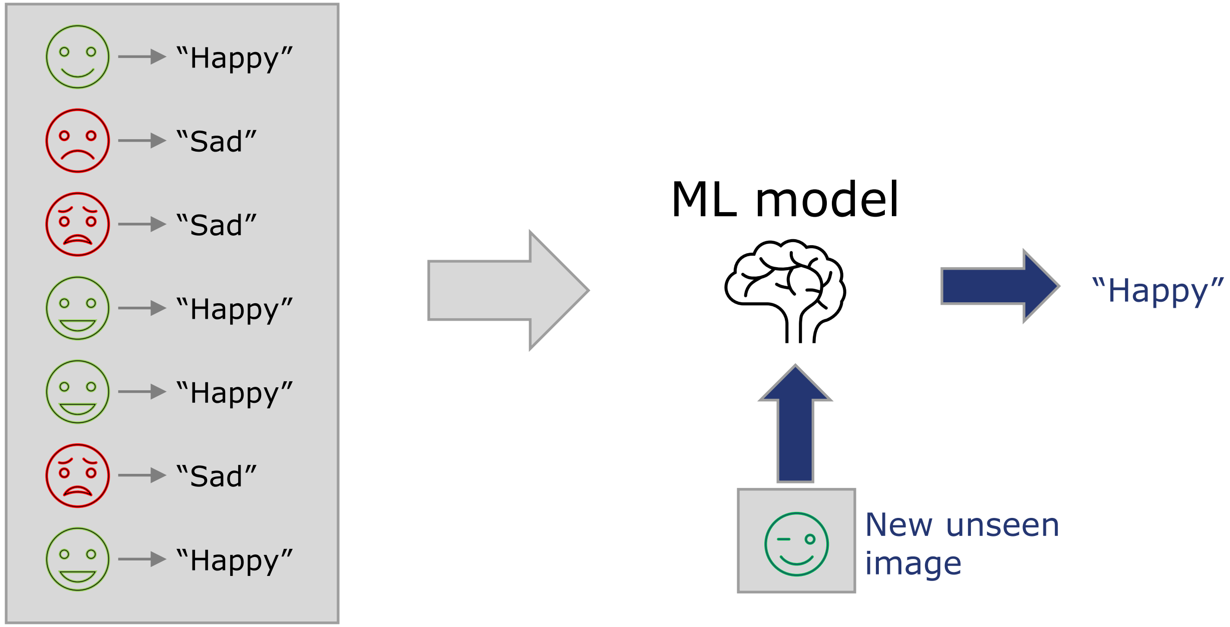
We know input and output.
The ML model is than finding the Relation between input and output



Training data (contains both input and output)

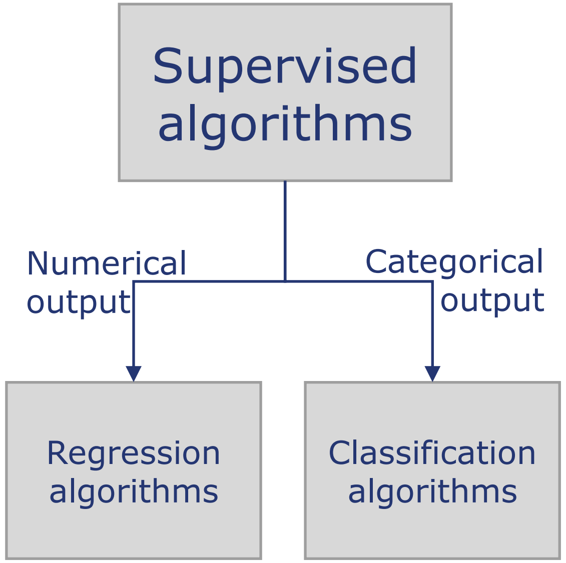
Machine learning

Supervised learning



Training data (contains both input and output)

Machine learning
Supervised learning



Two main types of supervised learning algorithms

- **Regression**: The variable to predict is numerical
  - Examples: Predict the price of an object. Predict the age of a subject. Predict the income of a person. Predict the cost of a project. Predict the lifetime of a device. Etc.
- **Classification**: The variable to predict is categorical (a label)
  - Examples: Spam detection (“spam” vs. “not spam”), image recognition (“table” vs. “chair” vs. “TV” vs. …), malfunctioning component (“proper” vs. “malfunctioning”), credit reliability of a person (“reliable” vs “unreliable”), etc.

Machine learning

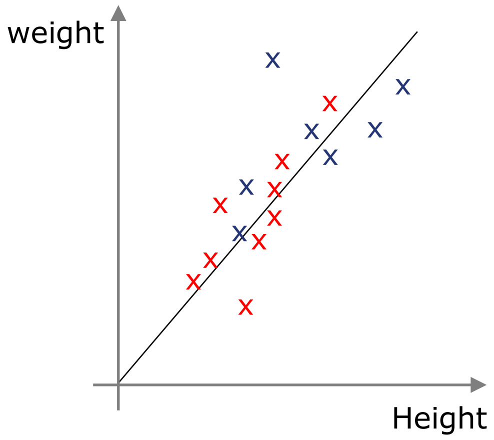
Supervised learning examples

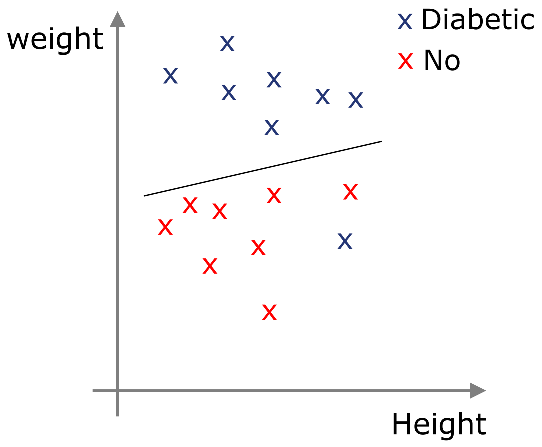
Linear regression

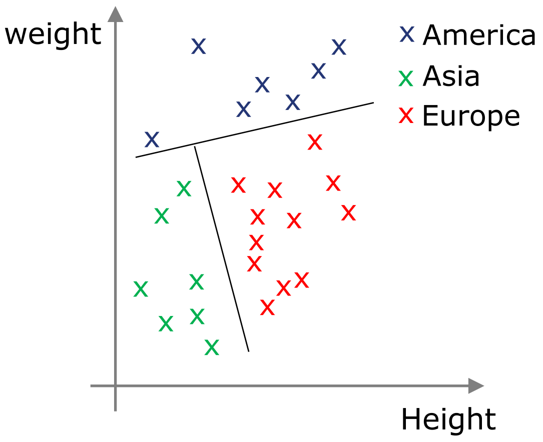
Support vector Machine

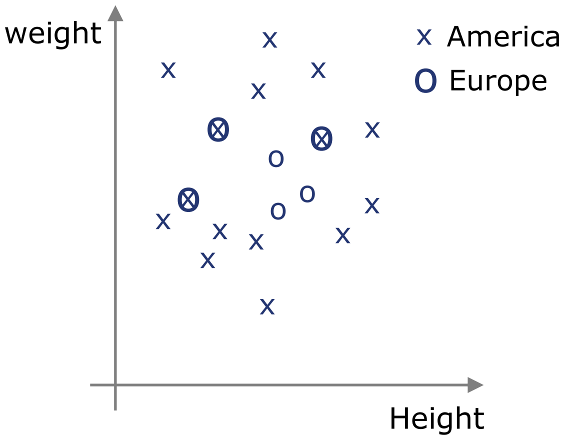
KNN

Regression	Classification		
	Binary	Multi-class	Multi-label
Predict weight based on height	Predict if subject has diabetes or not	Predict in which continent the subject is living	Predict in which continent subject has lived



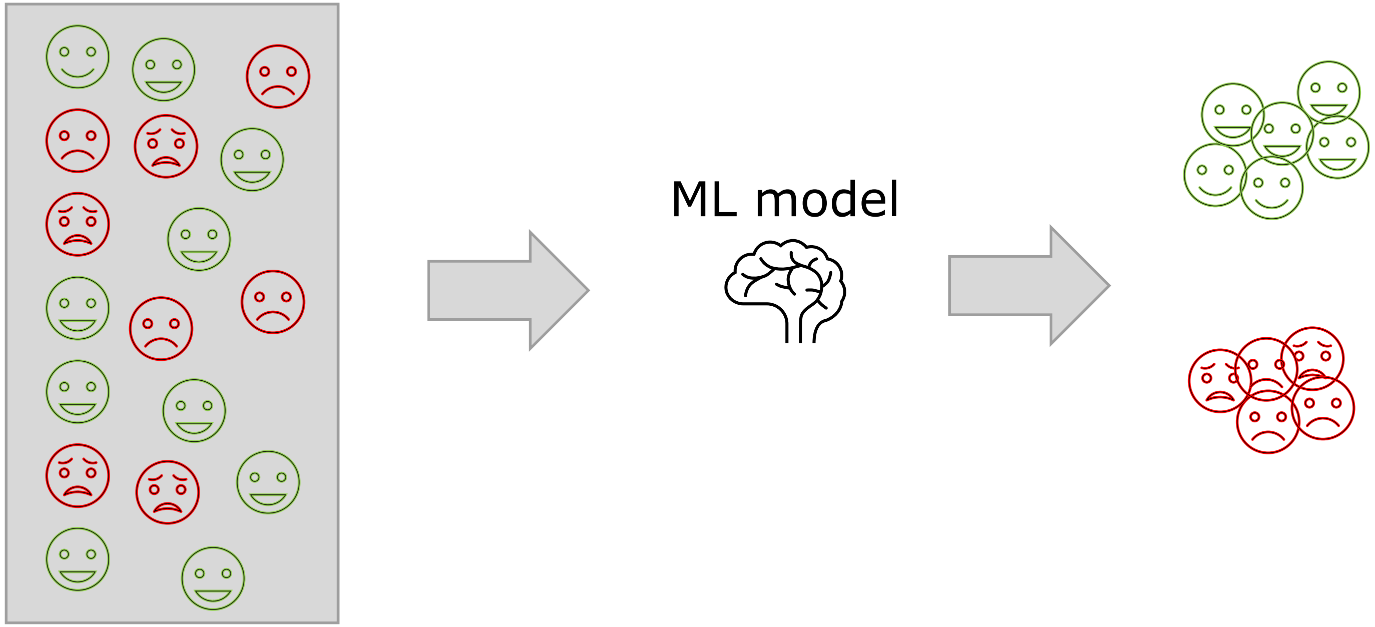






Machine learning

Unsupervised learning

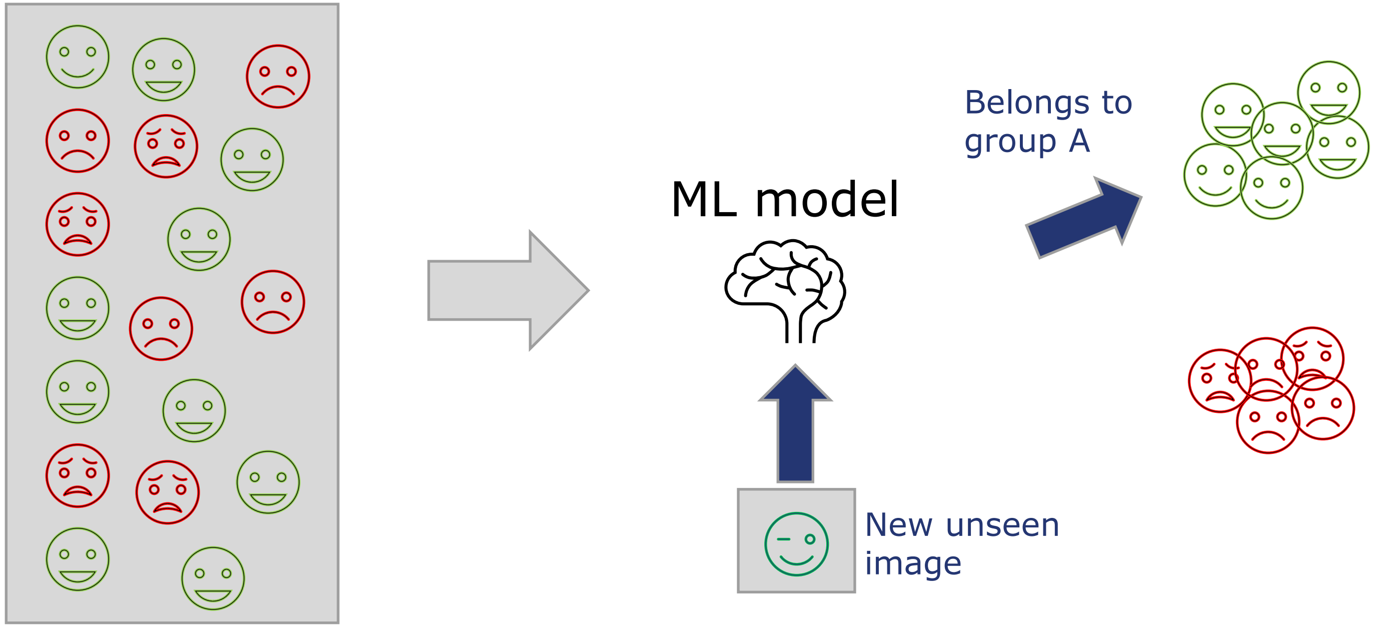


ML model

Training data (contains only input)

Machine learning

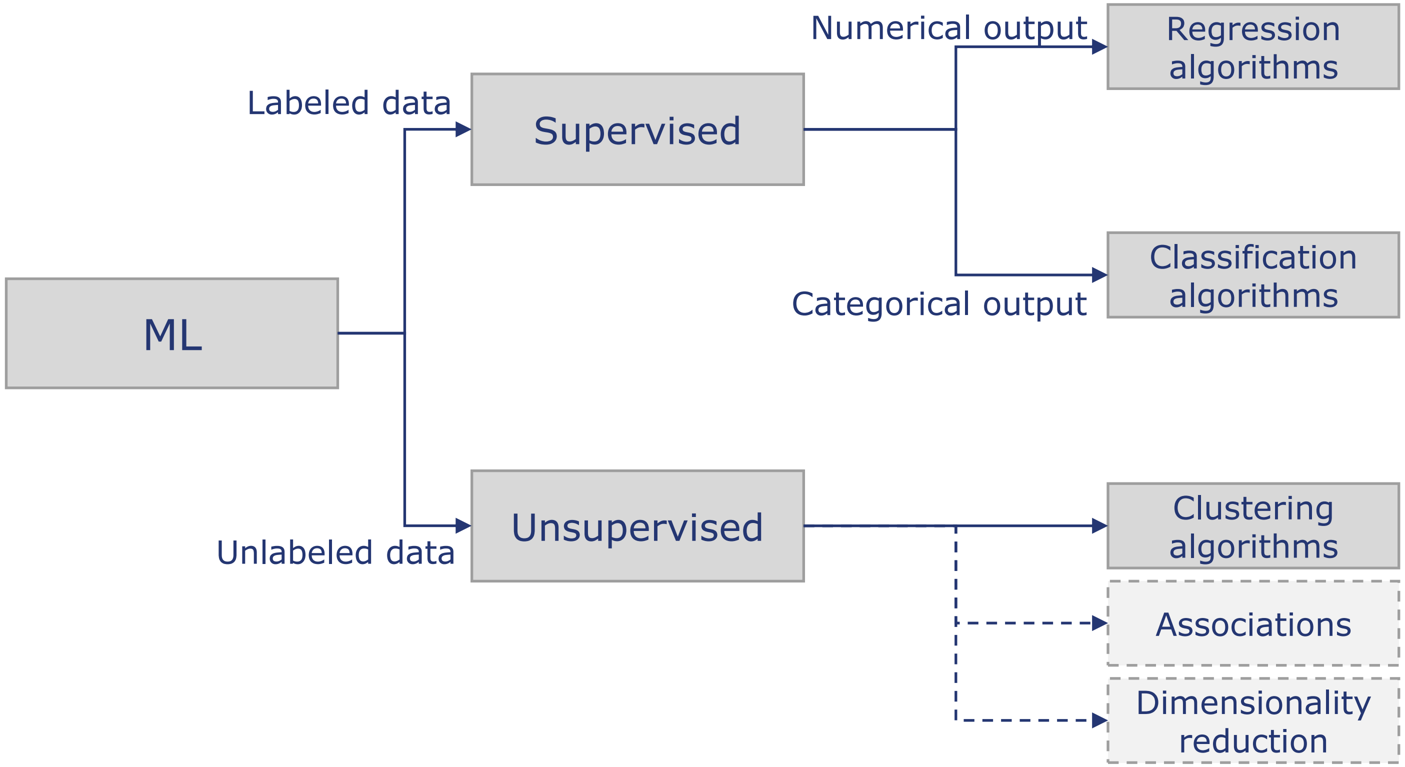
Unsupervised learning



Training data (contains only input)

Machine Learning

Supervised vs. Unsupervised Summary

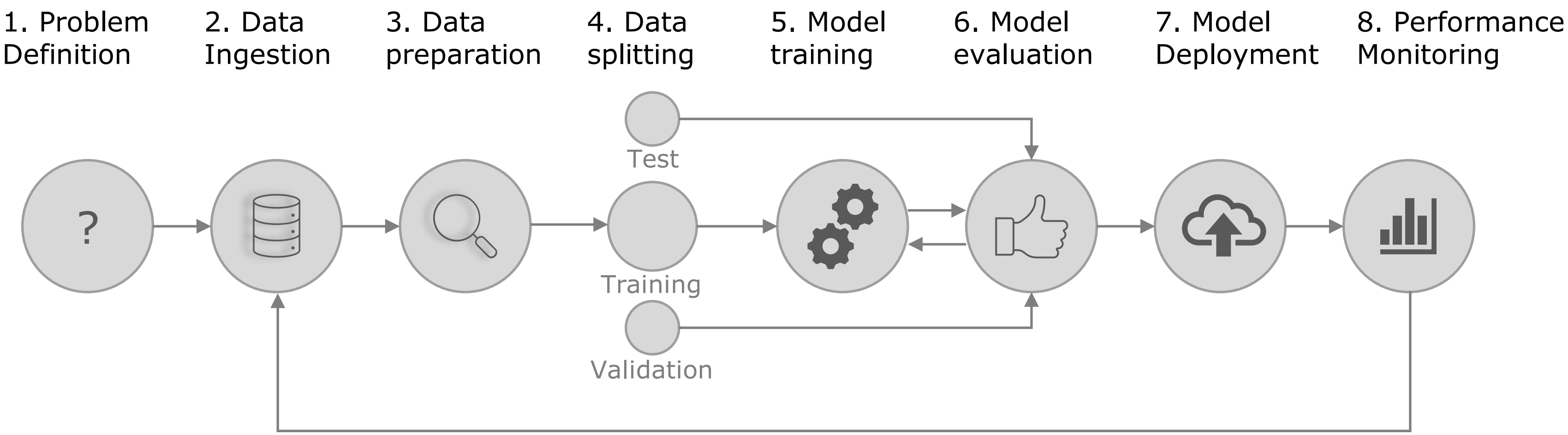


- Linear regression
  - Simple
  - Multiple
  - Multivariate
- Non-linear regression
- Decision Tree
- Random Forest

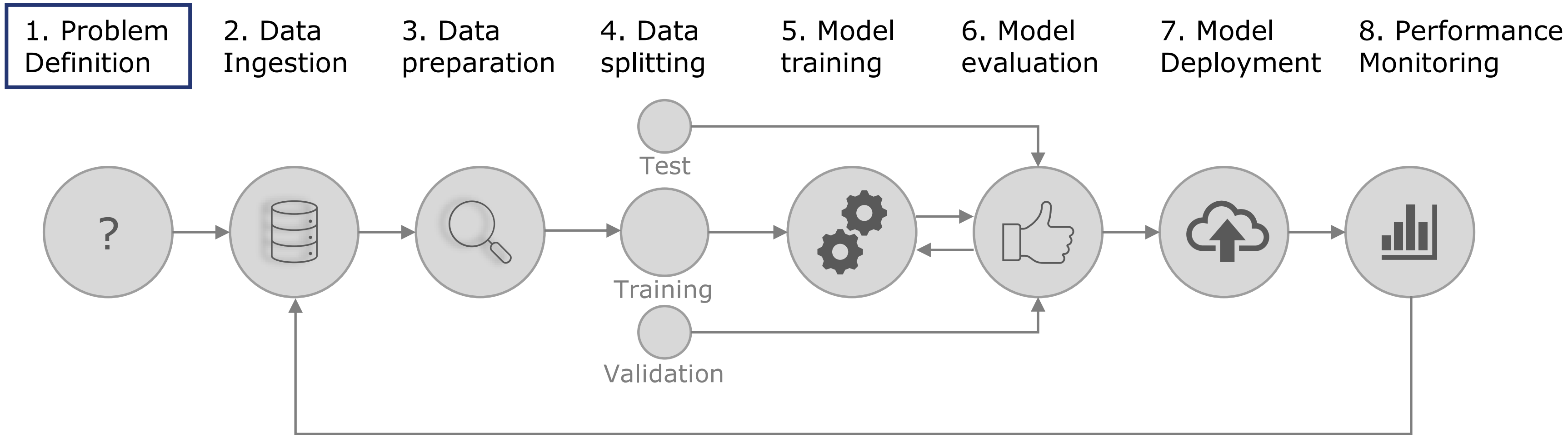
- Naïve Bayes
- Logistic regression
- Support Vector Machines
- K-nearest neighbors
- Decision Tree
- Random Forest
- …

- K-means
- Hierarchical clustering
- Density-Based clustering
- Model-based clustering
- …

The Machine Learning pipeline



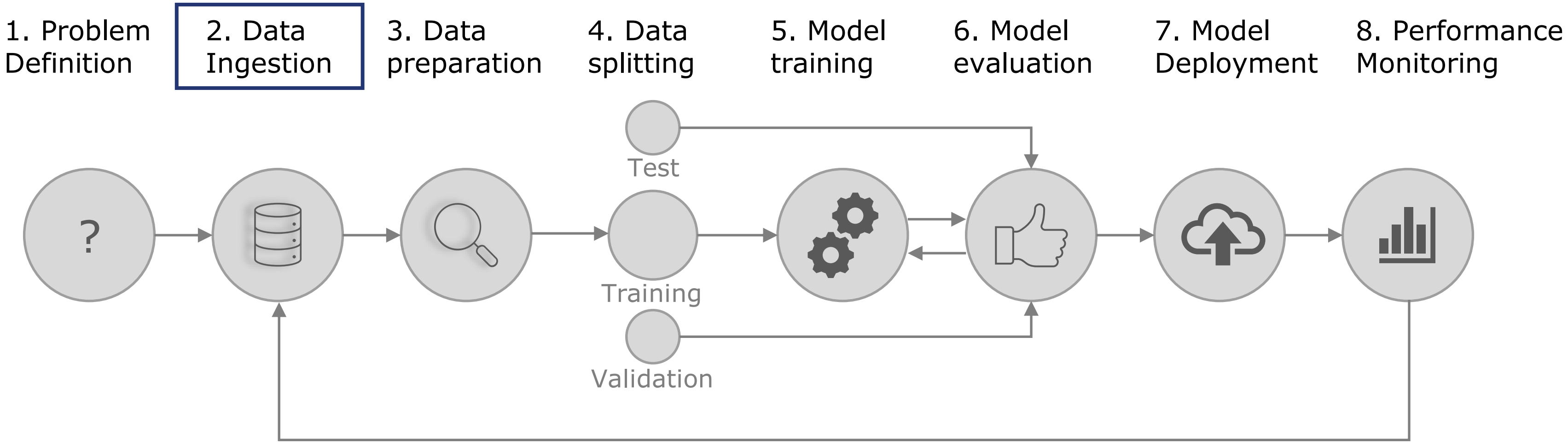
The Machine Learning pipeline



1. Problem Definition
2. Data Ingestion
3. Data preparation
4. Data splitting
5. Model training
6. Model evaluation
7. Model Deployment
8. Performance Monitoring

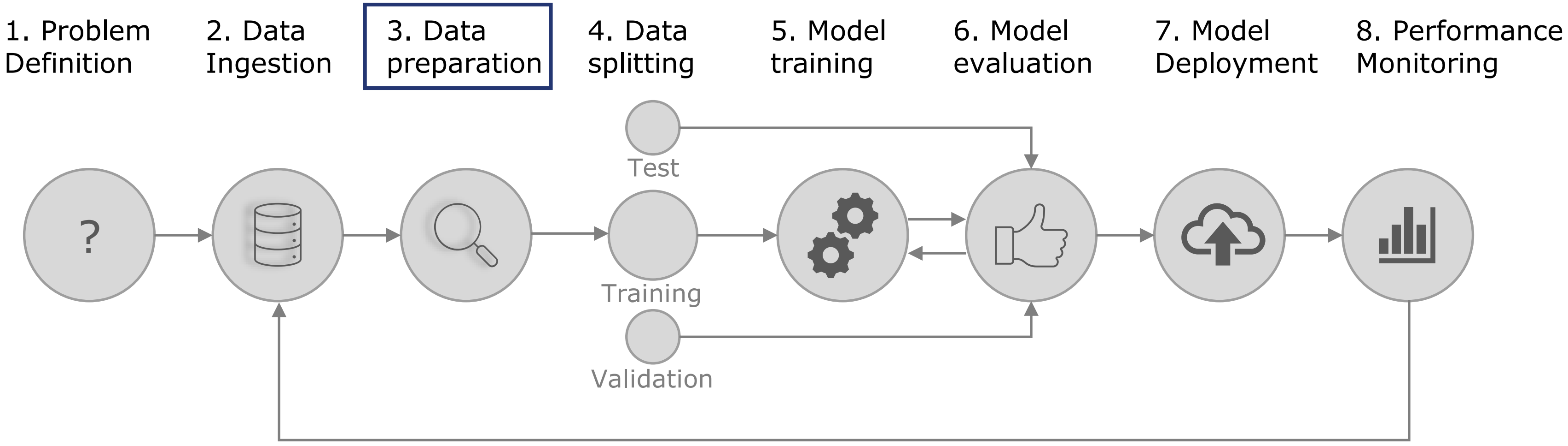
- Define the Business Question.
- Learn the domain.
- Translate the business question into a data science question.

The Machine Learning pipeline



- Identify which data is necessary and how to collect it
- Offline vs. Online
- Database vs. Files

The Machine Learning pipeline



1. Problem Definition

2. Data Ingestion

3. Data preparation

4. Data splitting

5. Model training

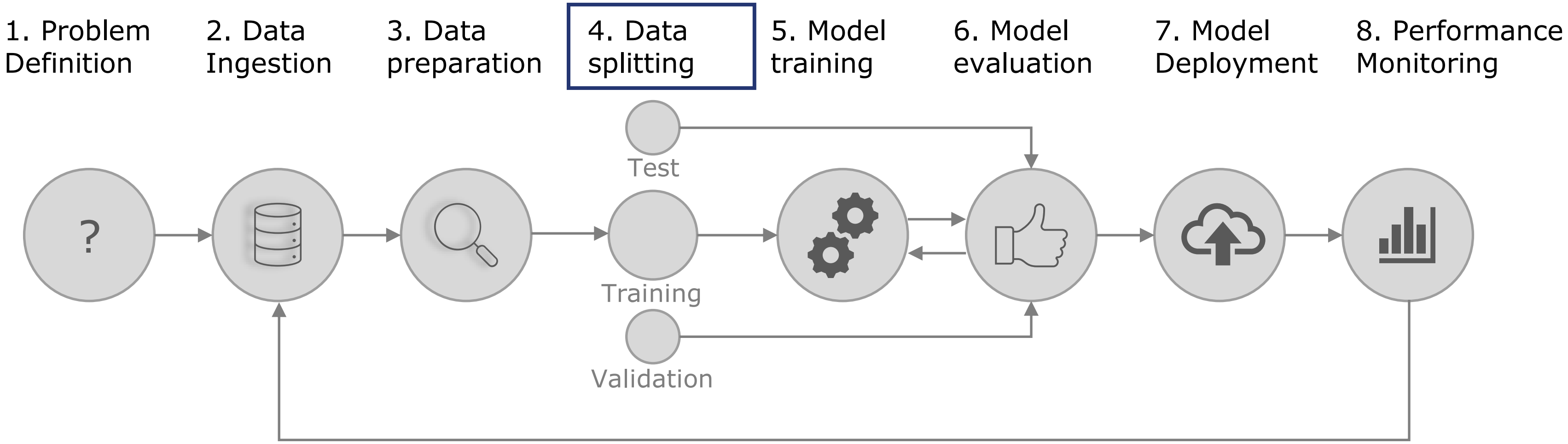
6. Model evaluation

7. Model Deployment

8. Performance Monitoring

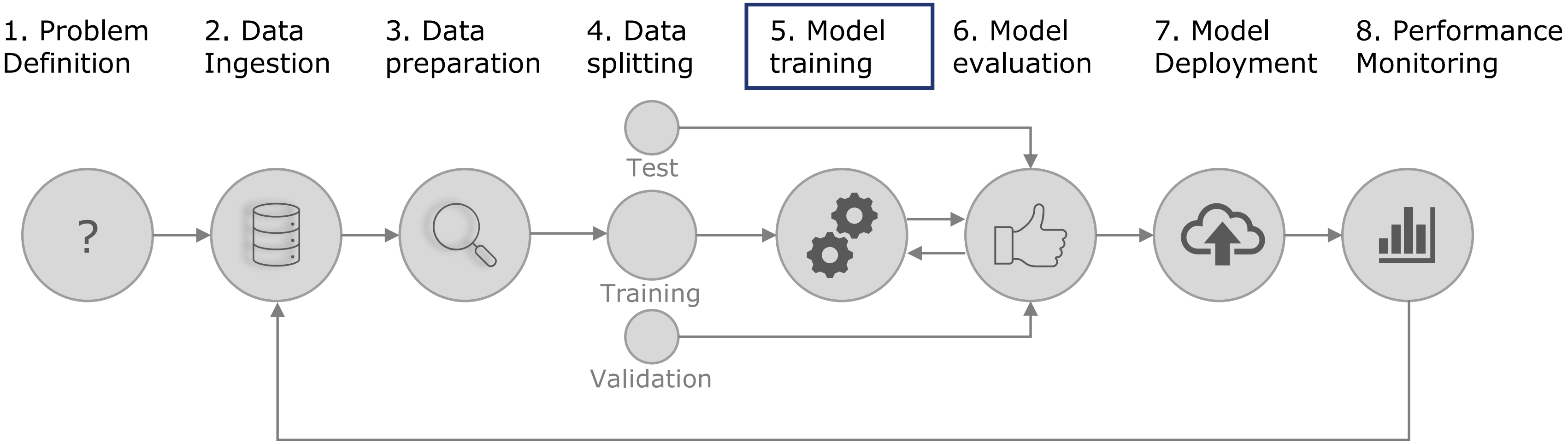
- Cleansing (remove erroneous entries, fill missing values, remove duplicates, etc.)
- Feature selection
- Feature engineering

The Machine Learning pipeline



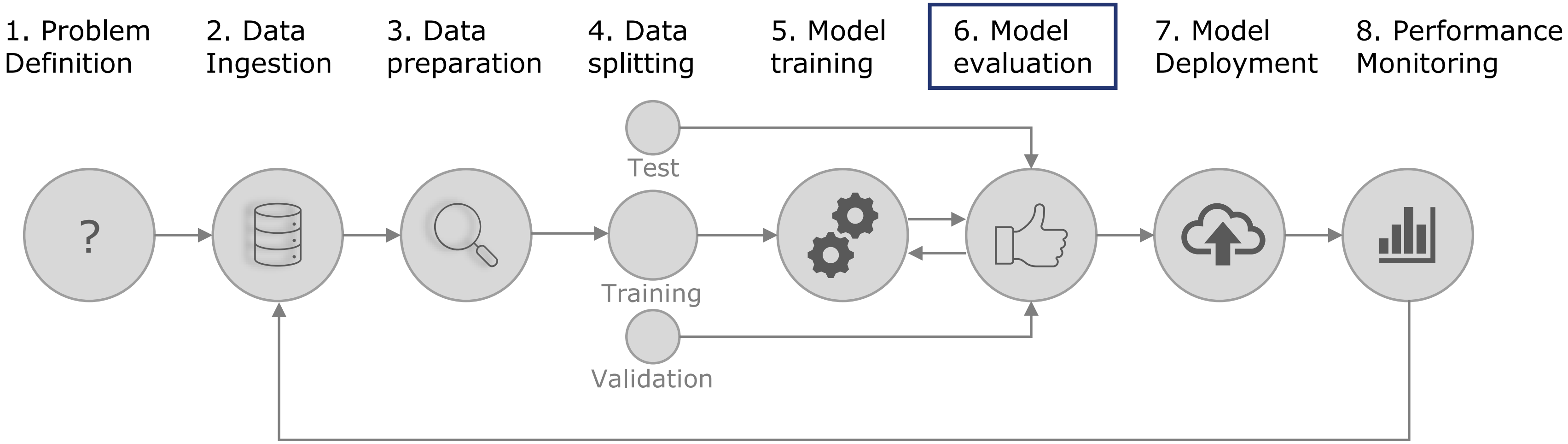
- Split into two or three subsets
- *Training* set → Used to train your model
- (*Validation* set) → Used for validate the trained model (e.g., tuning the parameters)
- *Test* set → Used to the evaluate the final tuned model to other final models

The Machine Learning pipeline



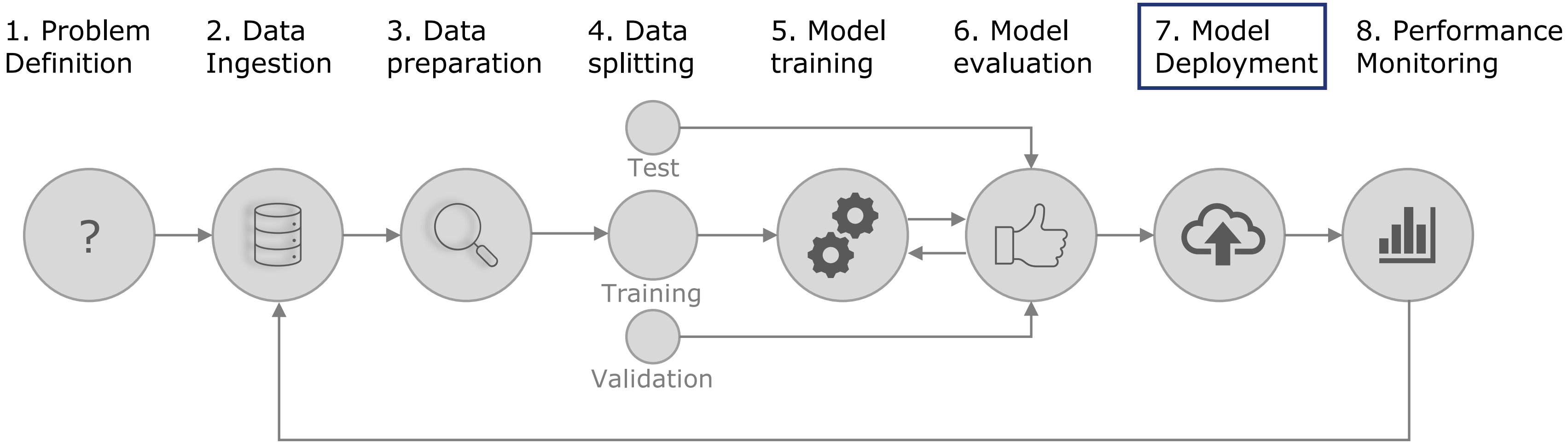
- Identify which type of algorithm is appropriate for training your model
    - Desired Accuracy,
    - Desired Interpretability,
    - Desired Scalability,
    - Constraints on processing power

The Machine Learning pipeline



- Use the validation and test data sets to assess the model accuracy
- Iterate 5. and 6. until appropriate

The Machine Learning pipeline



1. Problem Definition

2. Data Ingestion

3. Data preparation

4. Data splitting

5. Model training

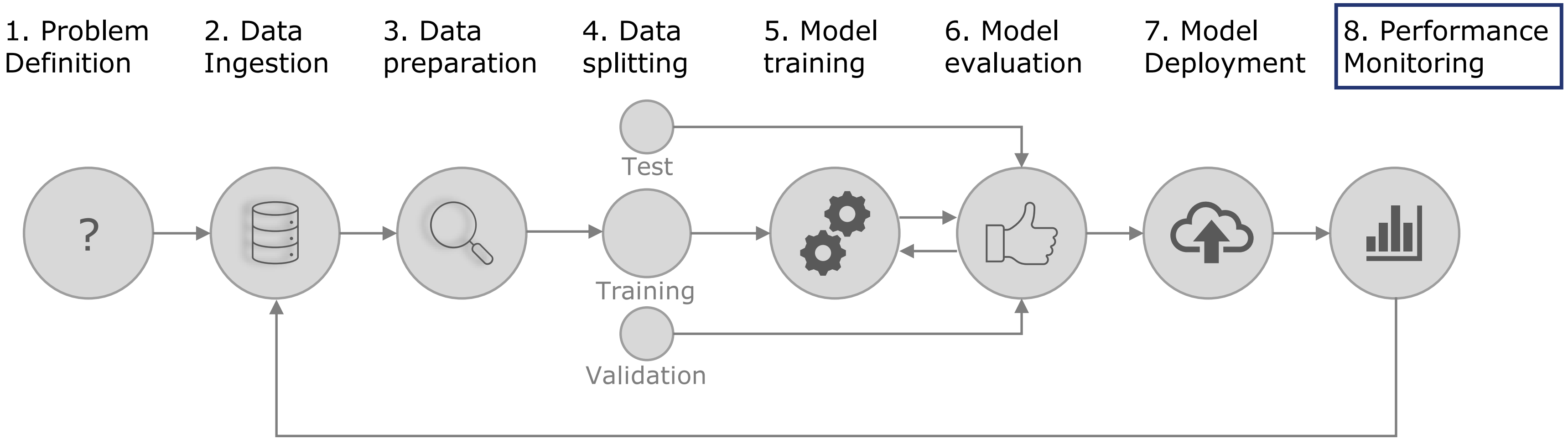
6. Model evaluation

7. Model Deployment

8. Performance Monitoring

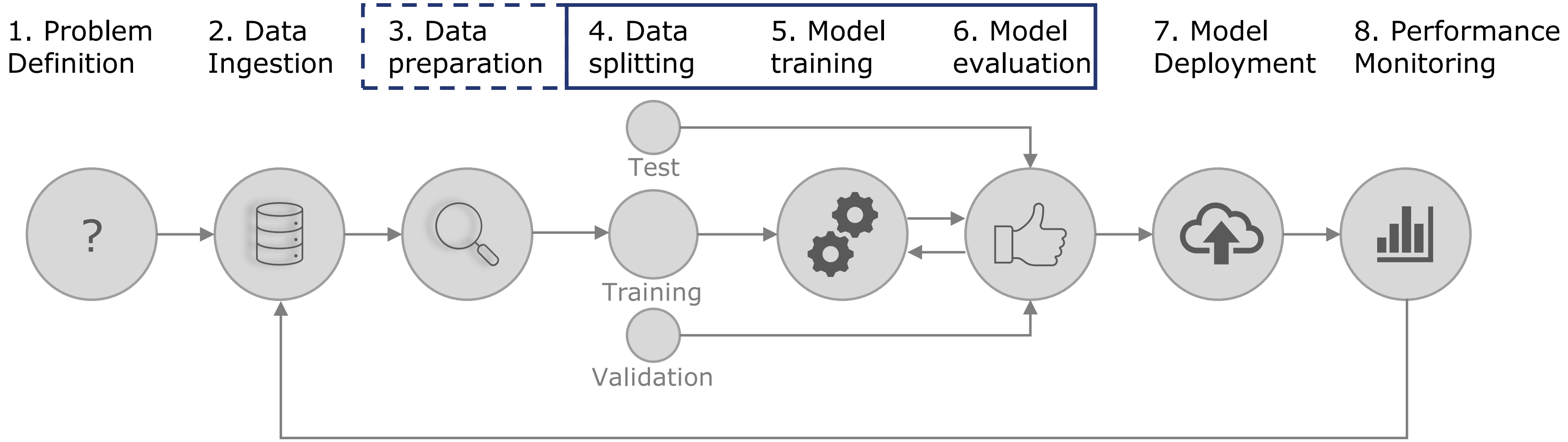
- Define how to deploy the model. API? Dashboard? Integrated Pipeline? Etc.

The Machine Learning pipeline



- Continuously monitor the performance of the model to identify deviations and deterioration of accuracy.

Focus of this course



Recap of R

- Setup for the course
- The R language
  - *R (>=4.0.0, suggested >=4.2.x)* – https://cran.r-project.org/
- Common IDEs (pick one)
  - *RStudio* (Simply the best. I strongly suggest to update to v2022.xx.x)
  - *Visual Studio Code* (the emerging one. Requires this)
  - *RGui* (the lightest. Comes with R. Lacks several functionalities)
  - *Jupyter Notebook* or *Jupyter Lab*
- During the course we’ll also use ***Python***… and if we have time we’ll introduce ***Julia*** as well.

Recap of R

Lecture continues on Rstudio