A Practical Introduction to Data Science

Part 1

Gergely Zsombor Haász

haasz.zsombi@gmail.com

About me

- Financial Mathematics MSc
- Over five years of experience in Data Science
  - Credit Risk Management
  - Big Data Consulting
  - Telecommunications
- Lifelong learner

Gergely Zsombor Haász

haasz.zsombi@gmail.com

Course Agenda

I. Introduction to Data Science

II. Business and Data Understanding

III. Introduction to Supervised Learning

IV. Advanced Supervised Learning

V. Unsupervised Learning

VI. Time Series Analysis

VII. Deep Learning

VIII. Machine Learning Operations

Introduction to Data Science

Data Science

Machine Learning

The ML Lifecycle

ML Algorithms

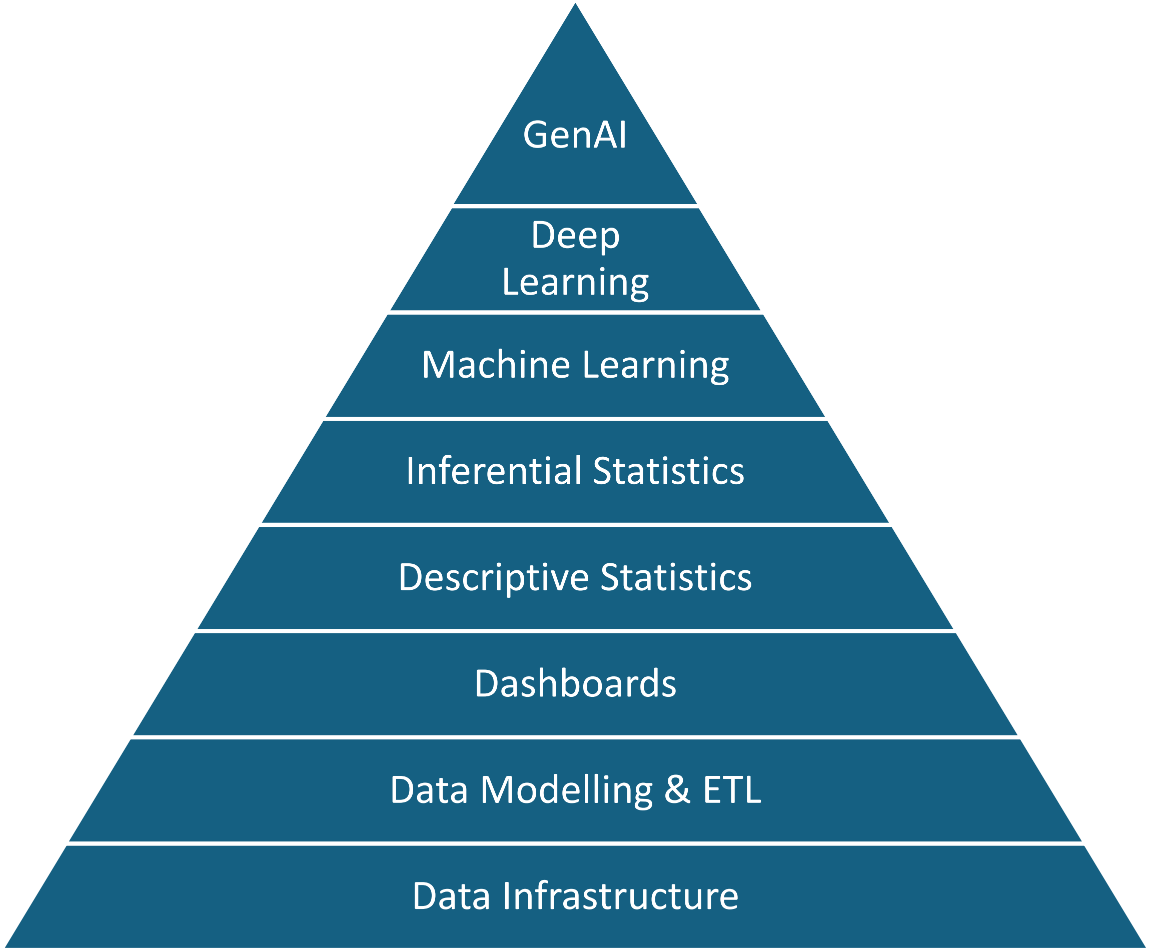
The Goal

Deep Learning

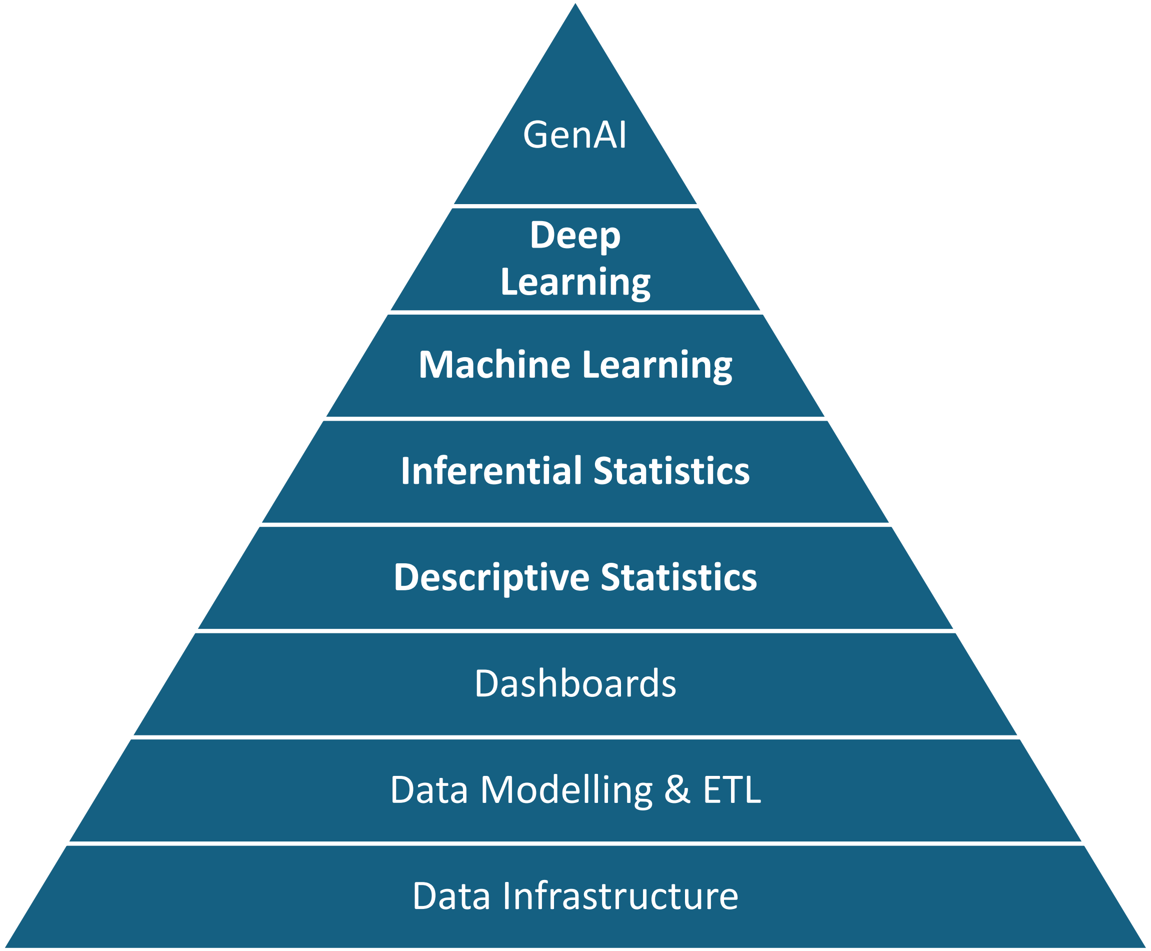
Tools

Data Science

The Data Analytics Iceberg



The Data Analytics Iceberg



GenAI

Deep Learning

Machine Learning

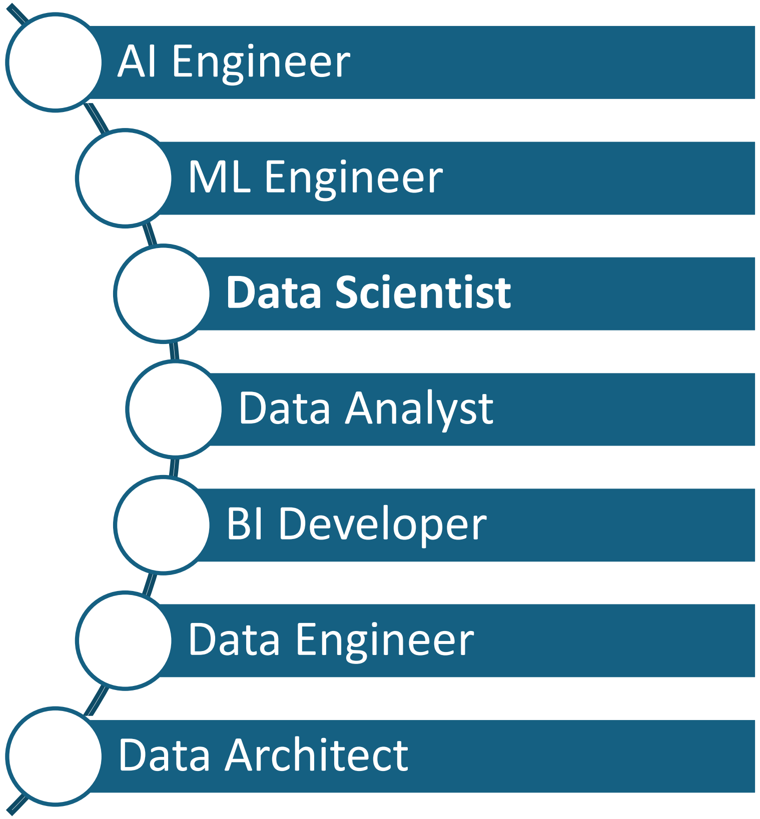
Inferential Statistics

Descriptive Statistics

Dashboards

Data Modelling & ETL

Data Infrastructure



AI Engineer

ML Engineer

Data Scientist

Data Analyst

BI Developer

Data Engineer

Data Architect

What is Data Science?

***Data Science** is about **extracting information** from data by **finding patterns**, to make better data-driven decisions.*

Data types

- Tabular
- Image & video
- Text & speech
- Other unstructured

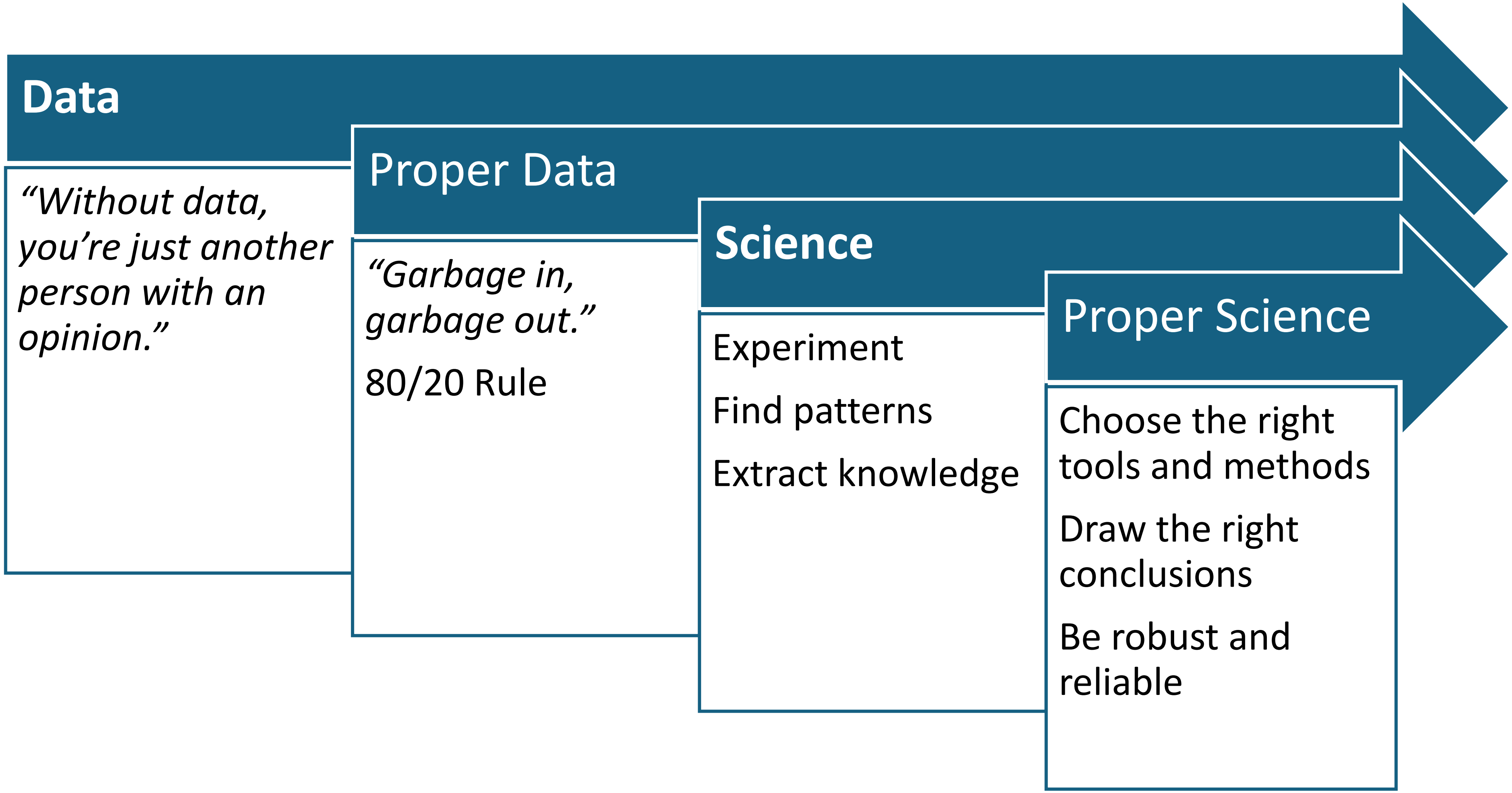
Methods

- Descriptive statistics
- Dashboards
- Inferential statistics
- Machine Learning
- Deep Learning
- Optimization

Applications

healthcare, psychology, finance, marketing, entertainment, sport, transportation, logistic, manufacture, energy, telecommunications, customer relationships, etc.

Data & Science



Data

"Without data, you're just another person with an opinion."

Proper Data

"Garbage in, garbage out."

80/20 Rule

Science

Experiment

Find patterns

Extract knowledge

Proper Science

Choose the right tools and methods

Draw the right conclusions

Be robust and reliable

Statistical Analysis

Descriptive

- Goals:
  - Understand as it is
  - Gather insights

- Tools:
  - Measures of Central Tendency
  - Measures of Dispersion
  - Distribution Properties
  - Correlation

Inferential

- Goals:
  - Generalization
  - Measure uncertainty
  - Causality

- Tools:
  - Estimation
  - Hypothesis Testing
  - Regression Analysis

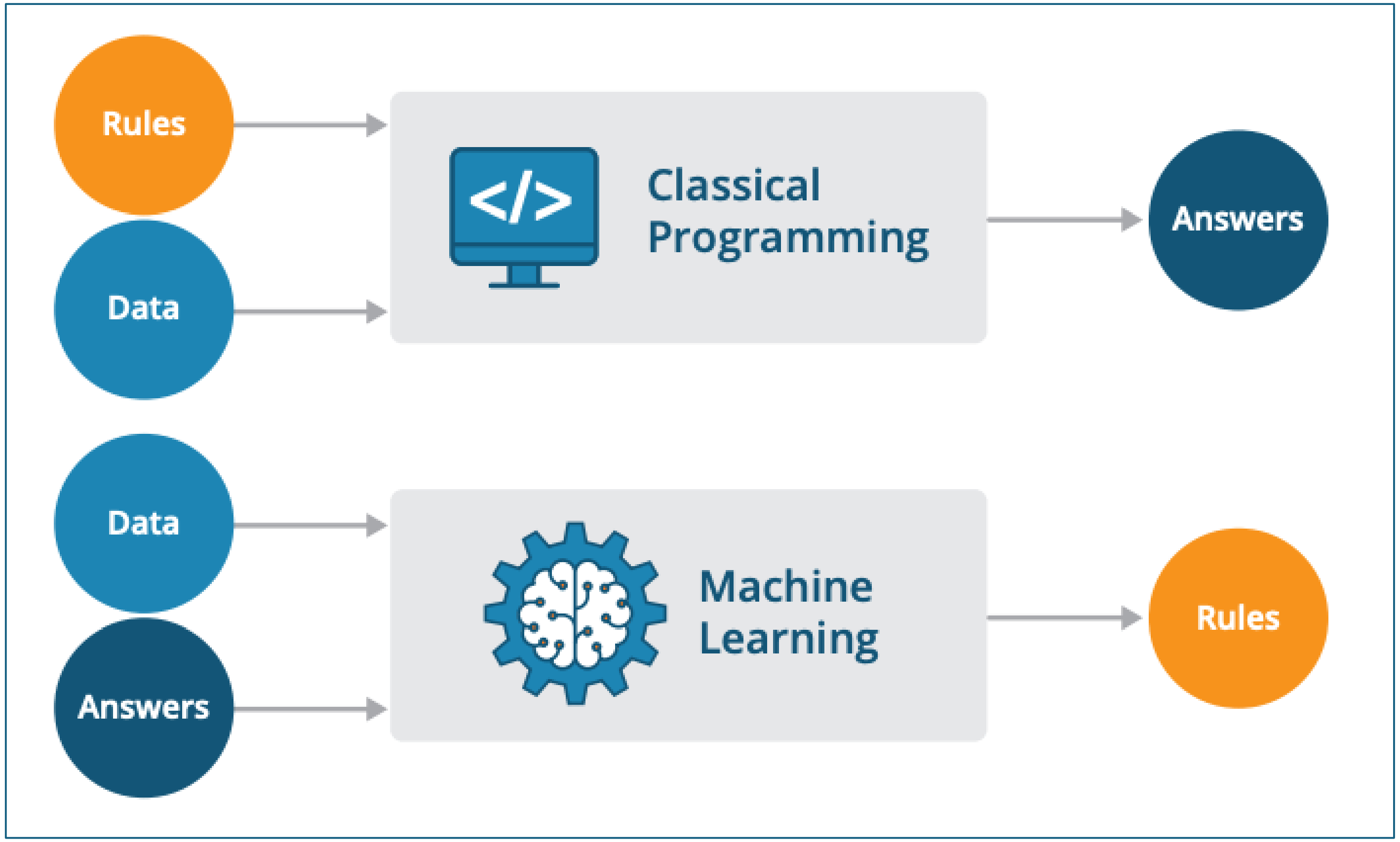
Machine Learning

What is Machine Learning?

”the ability to learn without being explicitly programmed” (Arthur Samuel, 1959)

An **ML model** is simply a mathematical function or algorithm that maps **input data** to an output (**prediction**).

Instead of explicit programming, an **optimization algorithm** is used to determine the best **parameters** of the model that minimize the prediction **error**.



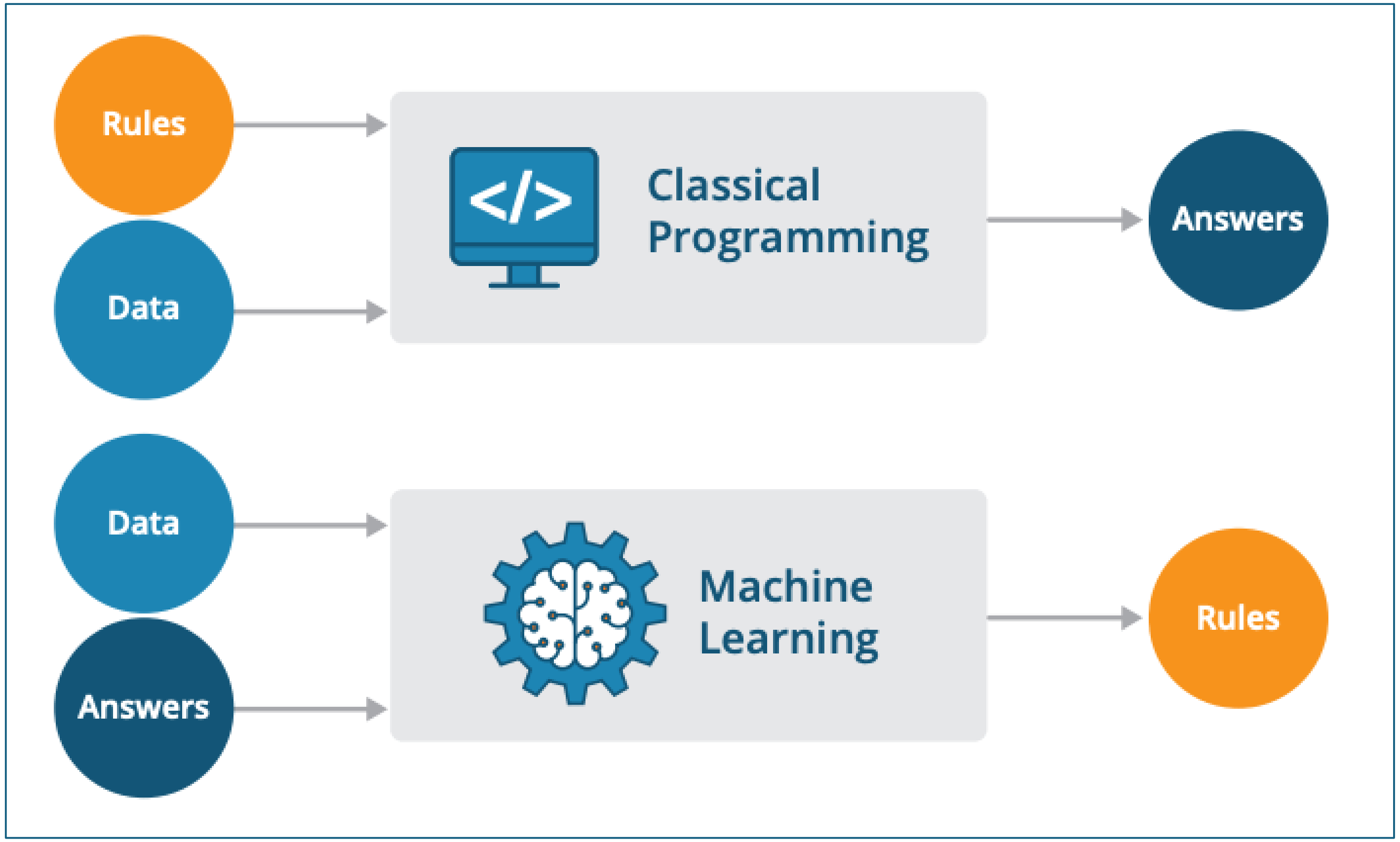
Source: Machine Learning is Forcing Software Development to Evolve - ML in Production

What is Machine Learning?

”the ability to learn without being explicitly programmed” (Arthur Samuel, 1959)

When to use Machine Learning?

1. Large data
2. Automation
3. Complex or changing patterns



Source: Machine Learning is Forcing Software Development to Evolve - ML in Production

What is Machine Learning?

Supervised Learning

predict, forecast, explain

- □ Classification
- □ Regression
- □ Survival Analysis

Unsupervised Learning

recognize patterns & structure

- □ Clustering
- □ Dimensionality Reduction
- □ Anomaly Detection
- □ Recommendation Systems

Reinforcement Learning

optimize actions

- □ Model-based
- □ Model-free

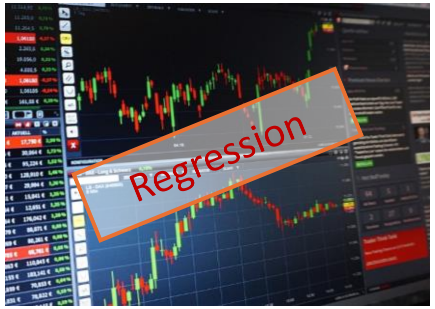
Supervised Learning

Supervised Learning

Classification

Regression

RESIGNATION
Classification



SPAM
Classification

Unsupervised Learning

Clustering

- Customer segmentation

Dimensionality Reduction

- Noise reduction, Visualization, Latent Variables

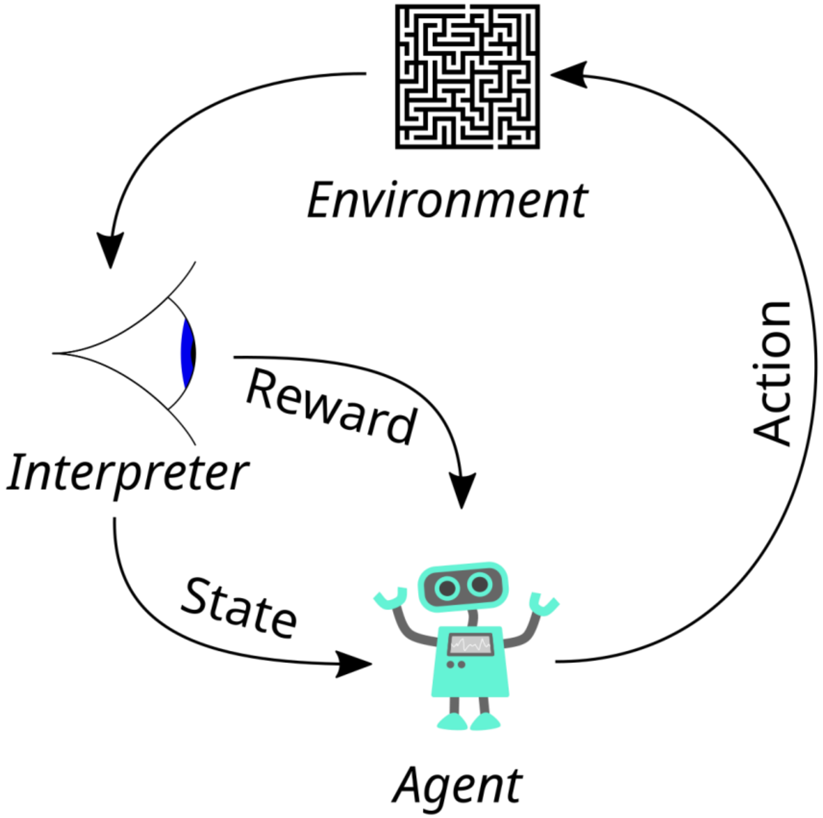
Anomaly Detection

- Fraud detection, fault detection

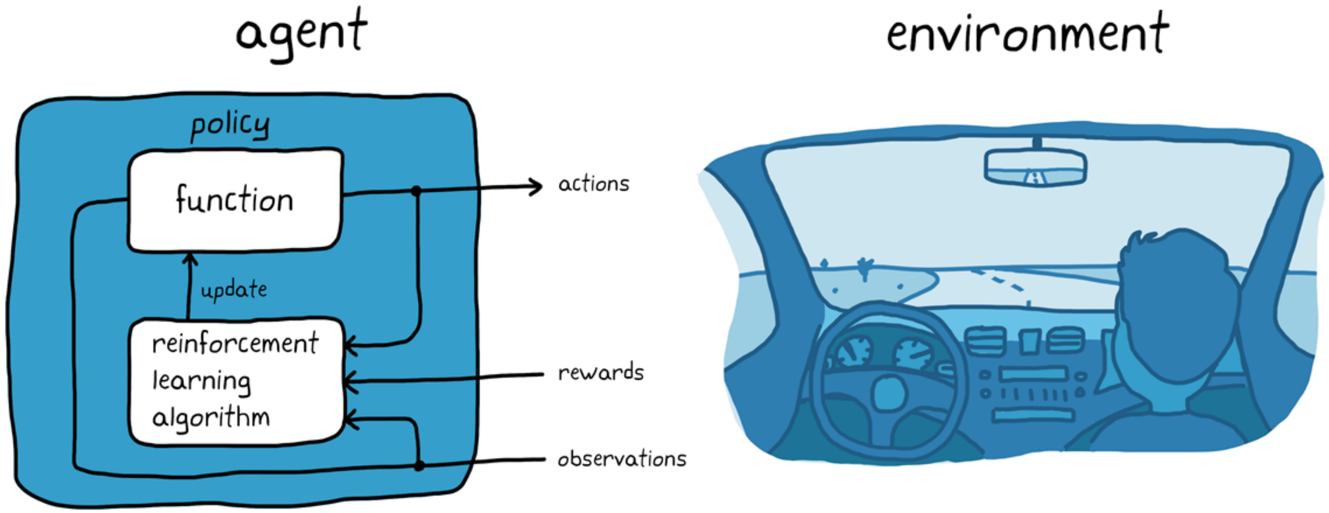
Recommendation Systems

- Personalized product/movie/news recommendations

Reinforcement Learning



Source: Reinforcement learning – Wikipedia

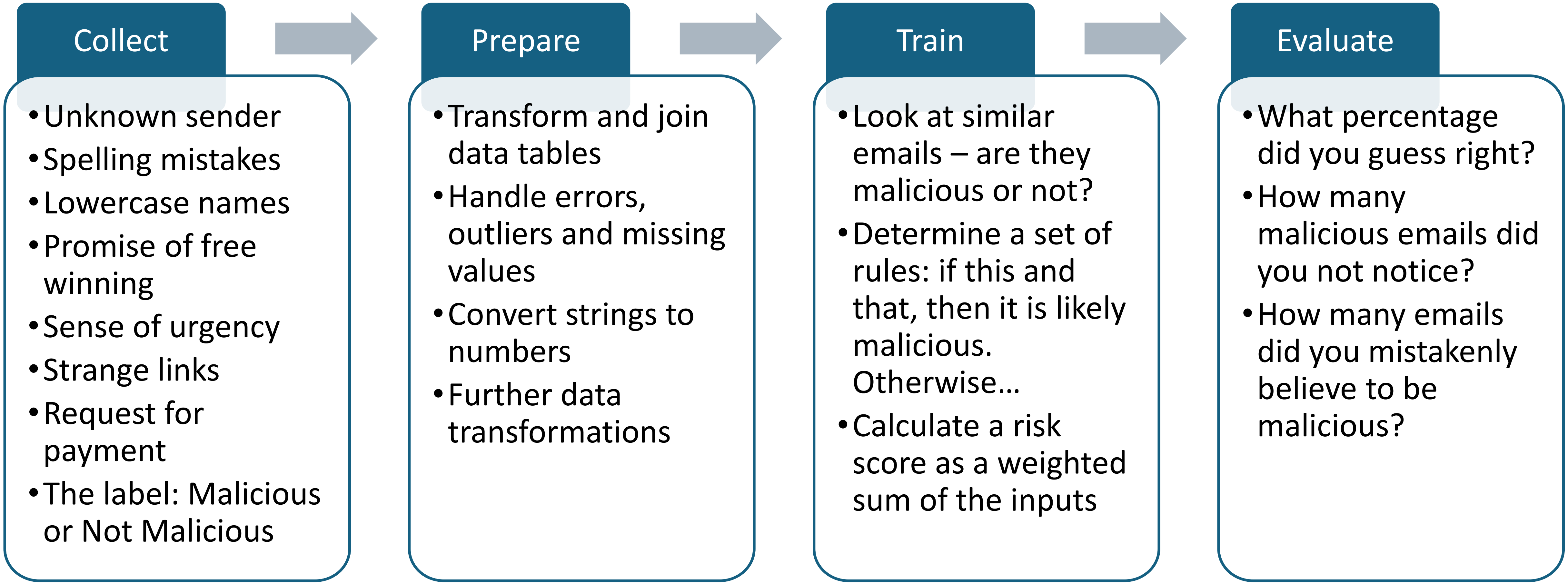


Source: What Is Reinforcement Learning? - MATLAB & Simulink

The ML Lifecycle

How does it work?

How does it work?



Collect

- Unknown sender
- Spelling mistakes
- Lowercase names
- Promise of free winning
- Sense of urgency
- Strange links
- Request for payment
- The label: Malicious or Not Malicious

Prepare

- Transform and join data tables
- Handle errors, outliers and missing values
- Convert strings to numbers
- Further data transformations

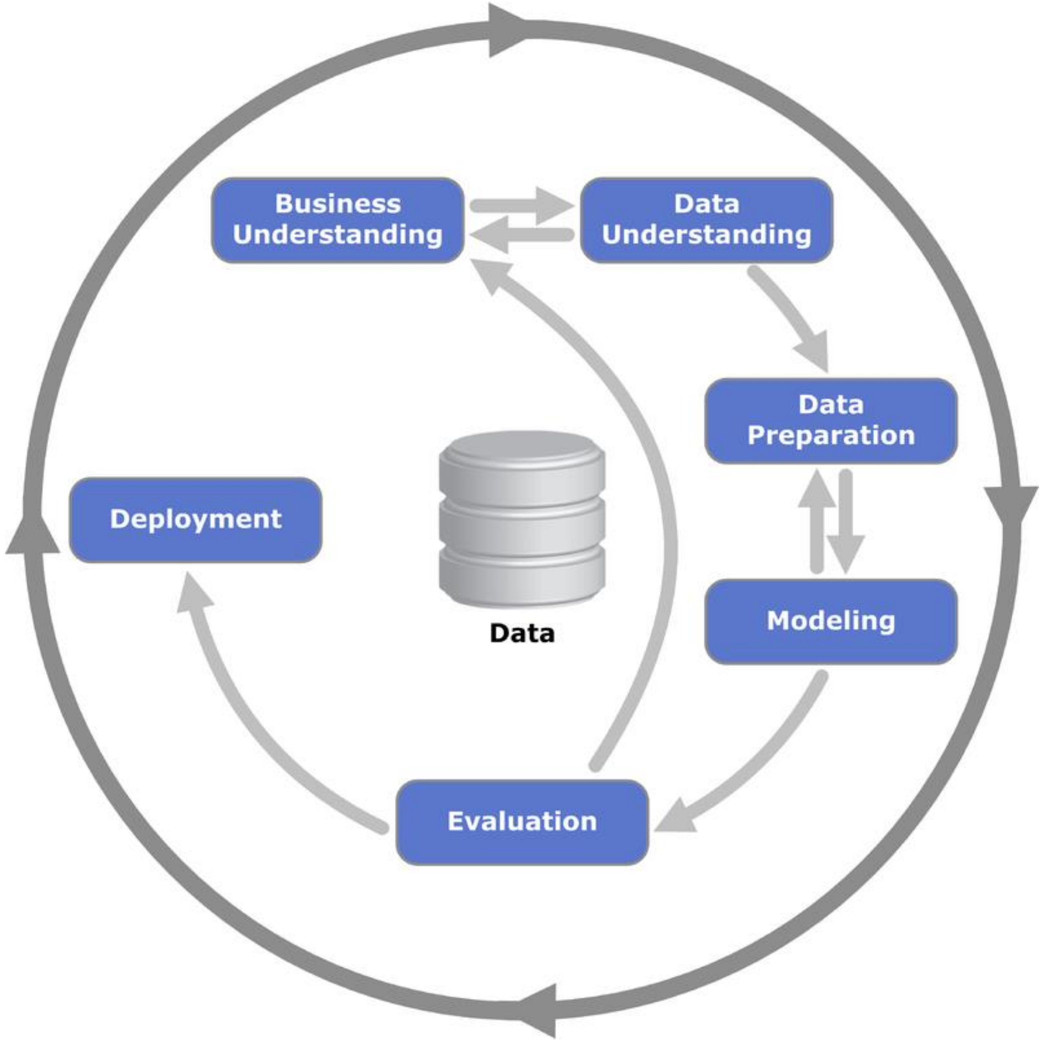
Train

- Look at similar emails – are they malicious or not?
- Determine a set of rules: if this and that, then it is likely malicious. Otherwise…
- Calculate a risk score as a weighted sum of the inputs

Evaluate

- What percentage did you guess right?
- How many malicious emails did you not notice?
- How many emails did you mistakenly believe to be malicious?

The Machine Learning Lifecycle



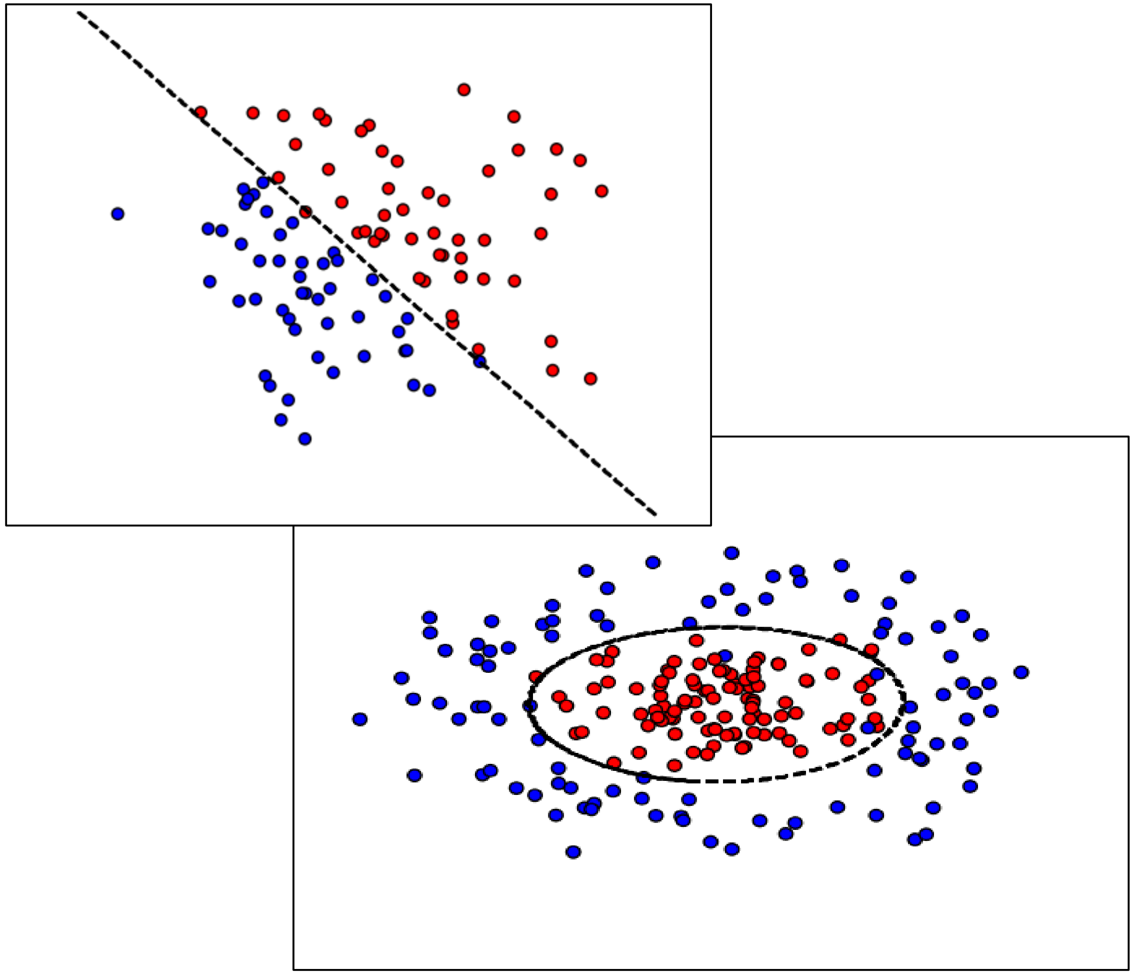
Source: Cross-industry standard process for data mining - Wikipedia

ML Algorithms

Machine Learning Algorithms

Supervised learning / Classification

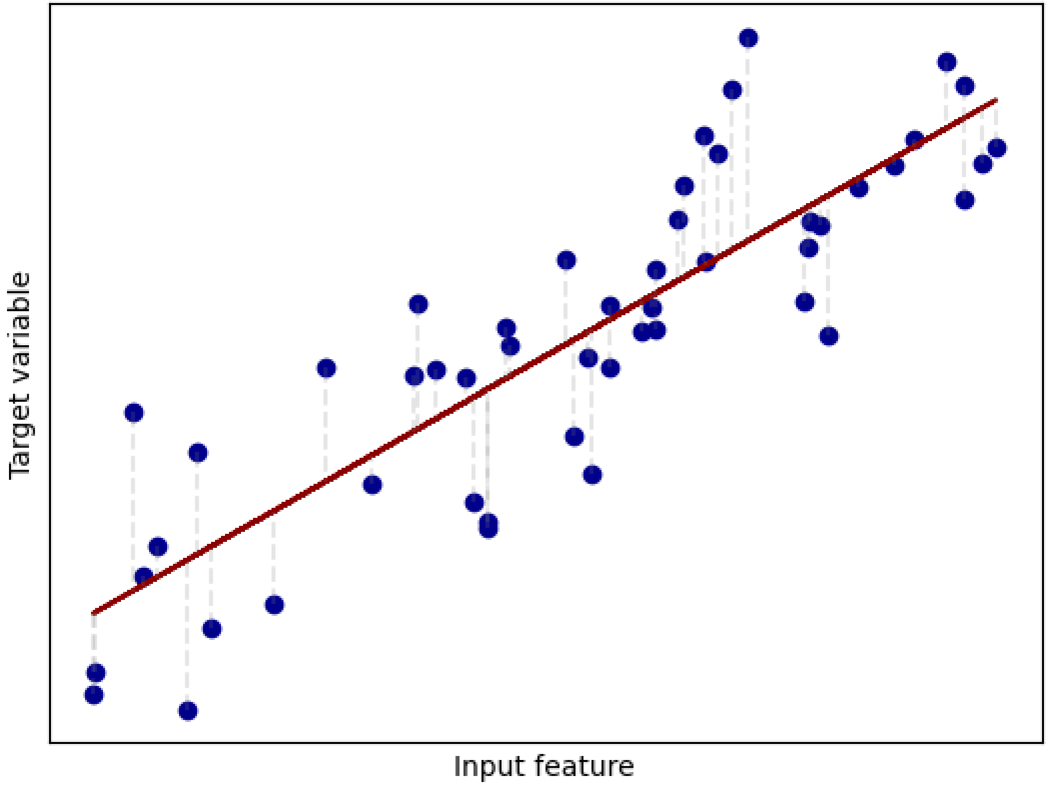
- Logistic Regression
- Decision Tree
- k-nearest Neighbors (kNN)
- Support Vector Machines (SVM)
- Ensembles (Random Forest, Gradient Boosting)



Machine Learning Algorithms

Supervised learning / Regression

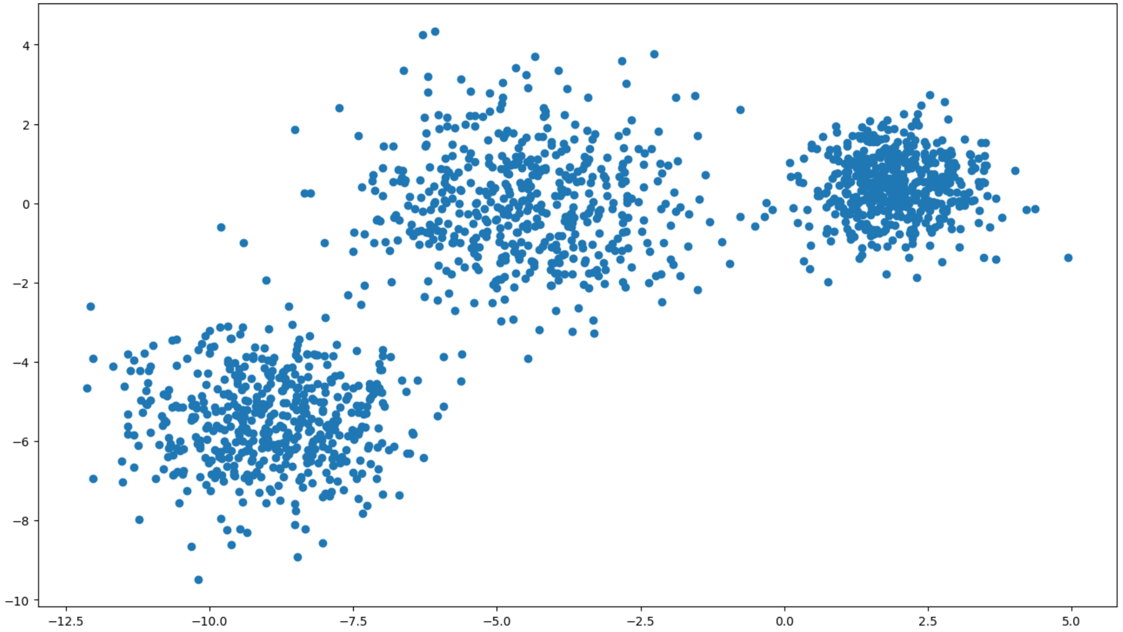
- Linear Regression
- Decision Tree
- k-nearest Neighbors (kNN)
- Support Vector Machines (SVM)
- Ensembles (Random Forest, Gradient Boosting)



Machine Learning Algorithms

Unsupervised learning / Clustering

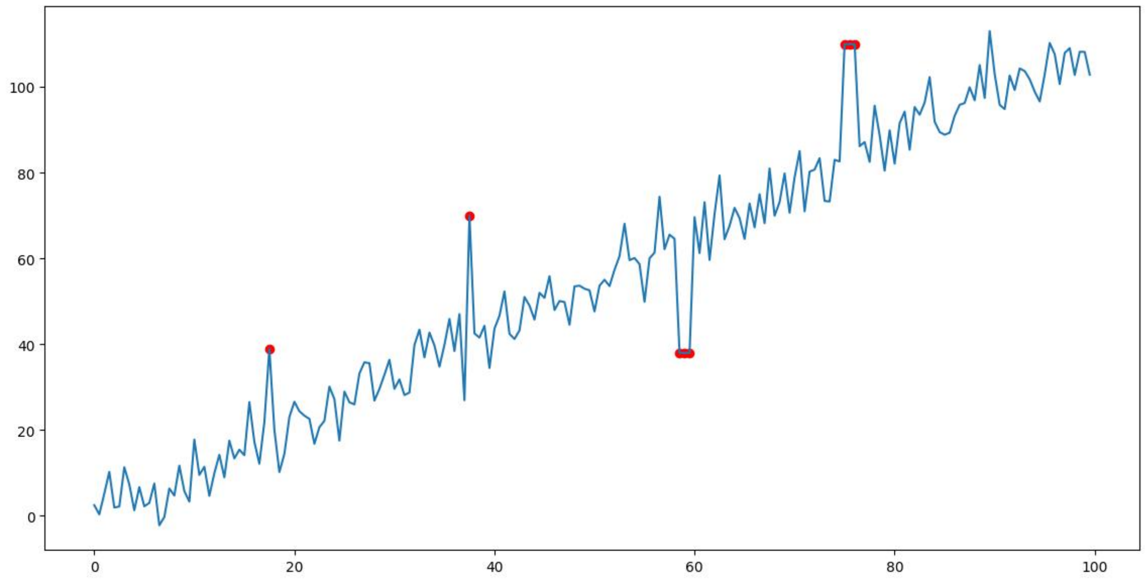
- K-means
- Hierarchical
- Density based (e.g. DBSCAN)



Machine Learning Algorithms

Unsupervised learning / Anomaly Detection

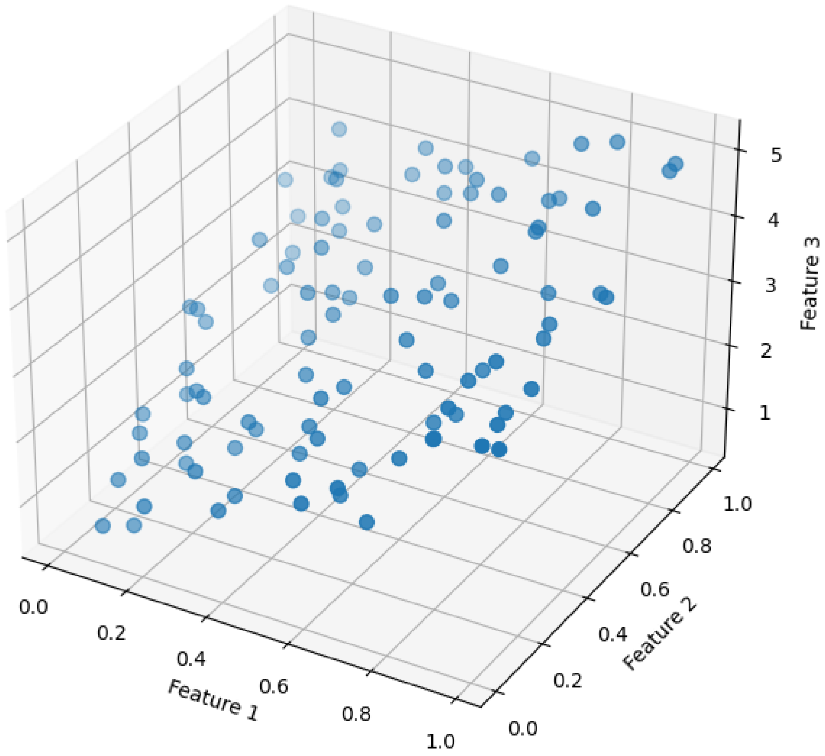
- Statistical Outlier Detection
- Isolation Forest
- One-class SVM
- Autoencoder

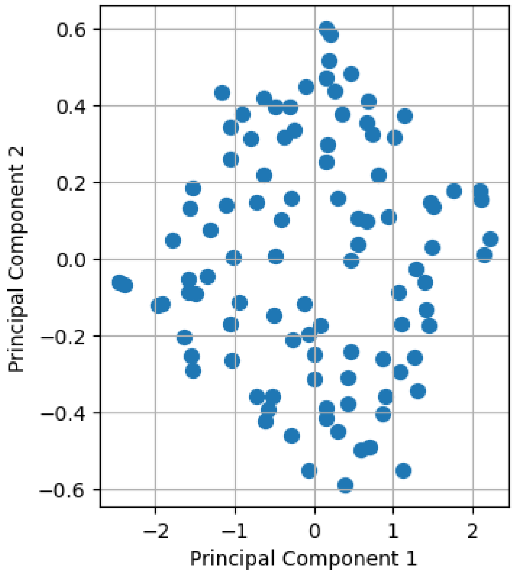


Machine Learning Algorithms

Unsupervised learning / Dimensionality Reduction

- PCA
- Factor Analysis
- Autoencoder



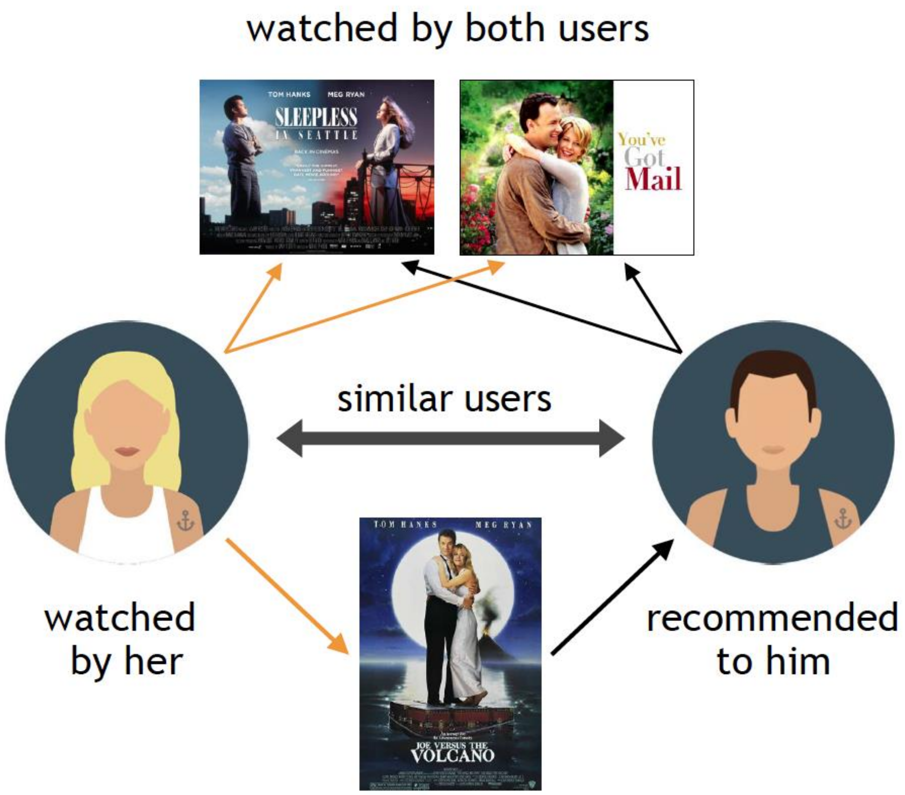


Machine Learning Algorithms

Unsupervised learning / Recommender Systems

- Collaborative filtering
- Content-based filtering
- Matrix factorization

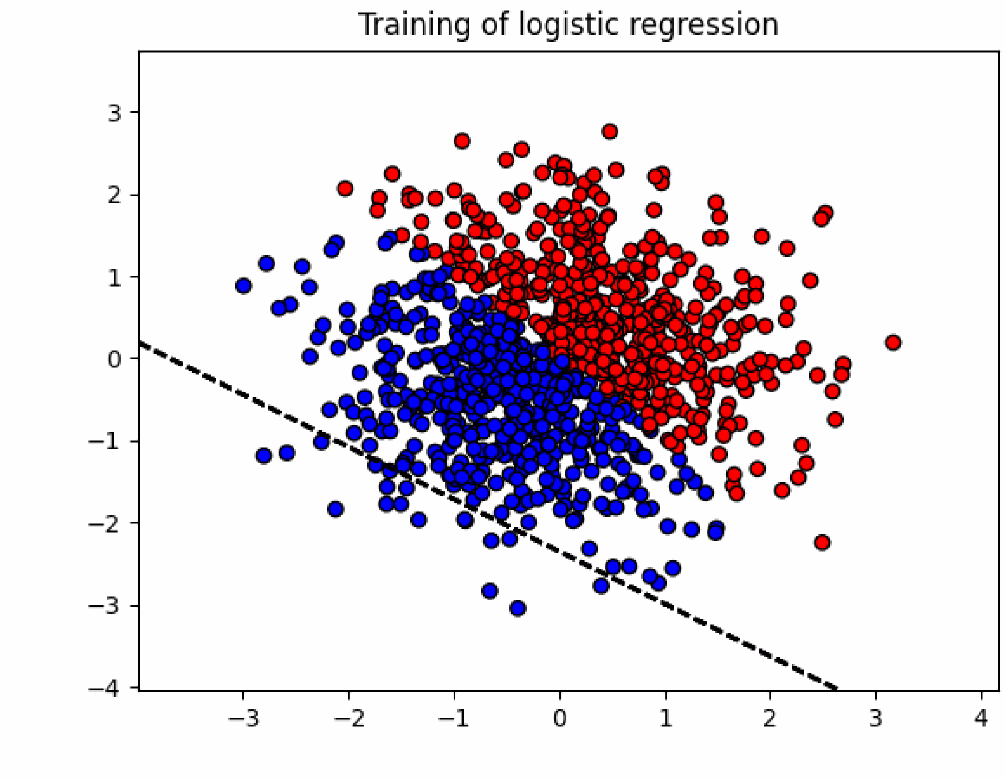
Collaborative Filtering



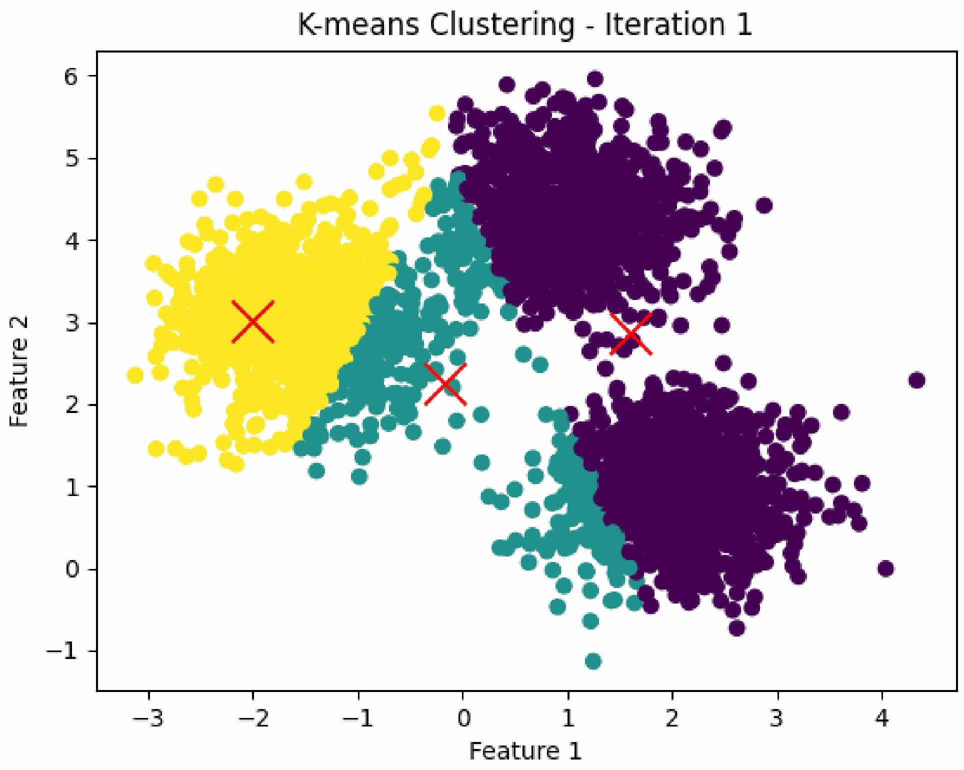
Source: What is a Recommendation System? | Data Science | NVIDIA Glossary

Model Training

Logistic Regression



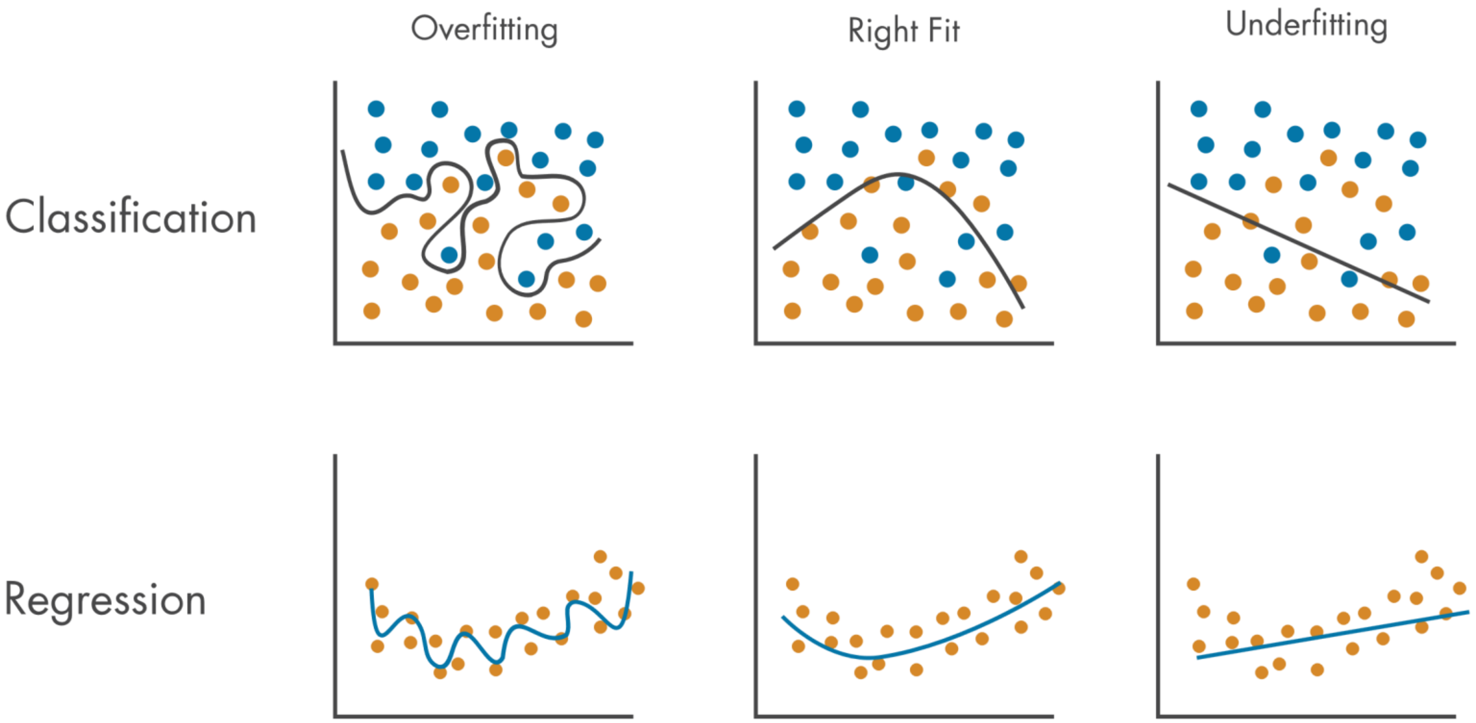
K-means Clustering

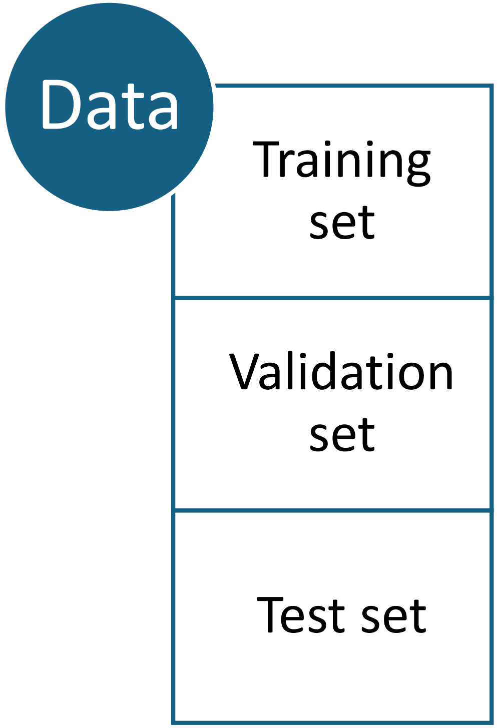


The Goal

Generalization

The goal of ML is to develop models that perform well on unseen data.





Data

Training set

Validation set

Test set

Source: [Overfitting - MATLAB & Simulink]

Deployment

More than 80% of data science projects never make it to production

Pitfalls preventing deployment

- Lack of engagement
- No real business value
- Data issues
  - Availability
  - Quality
  - Regulation
- Implementation issues
  - Technical Integration
  - Lack of expertise
  - Budget

Pitfalls after deployment

- Bad model
  - Bias
  - Poor performance
- Different data/environment
- Bad model-based decisions
- External factors

To do

- Consider deployment from the beginning of a project
- Talk to the business
- Model monitoring
- Be ready to intervene

Deep Learning

What is Deep Learning?

A subset of machine learning inspired by the human brain, utilizing deep neural networks to learn from data. A neural network is basically a sequence of nonlinear mathematical functions.

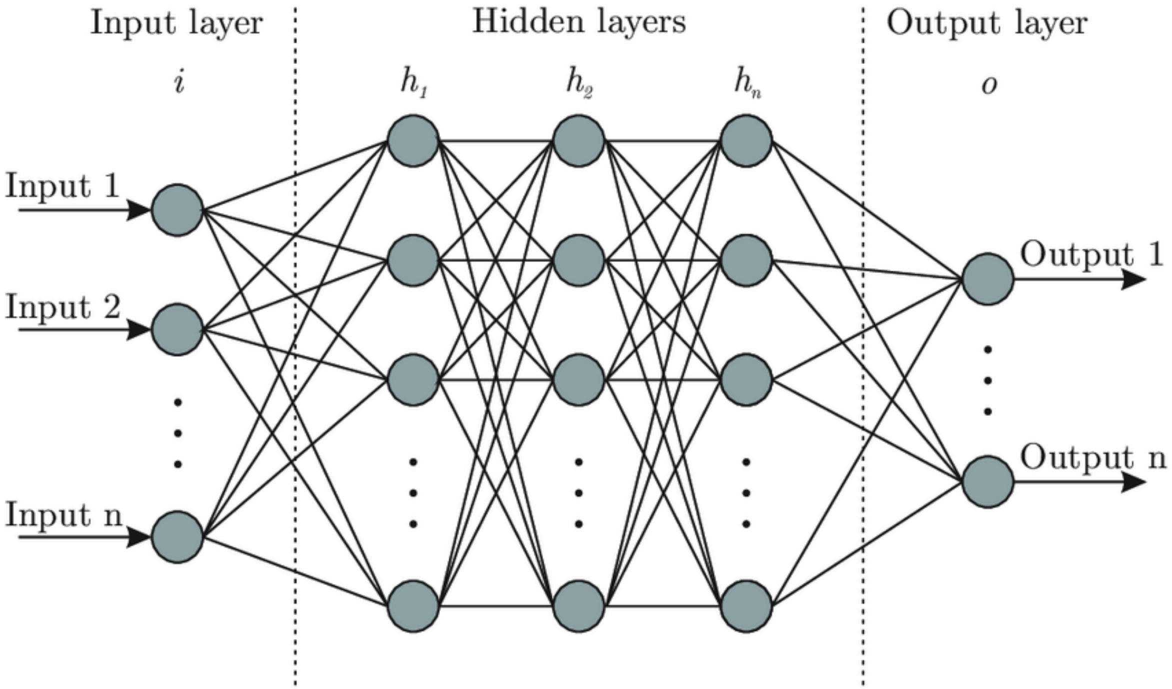
Model architecture outlines the sequence and connectivity of layers and neurons, along with the functions they perform.

Elements of a Neural Network:

- Layers
- Neurons (Nodes)
- Weights and biases
- Activation functions

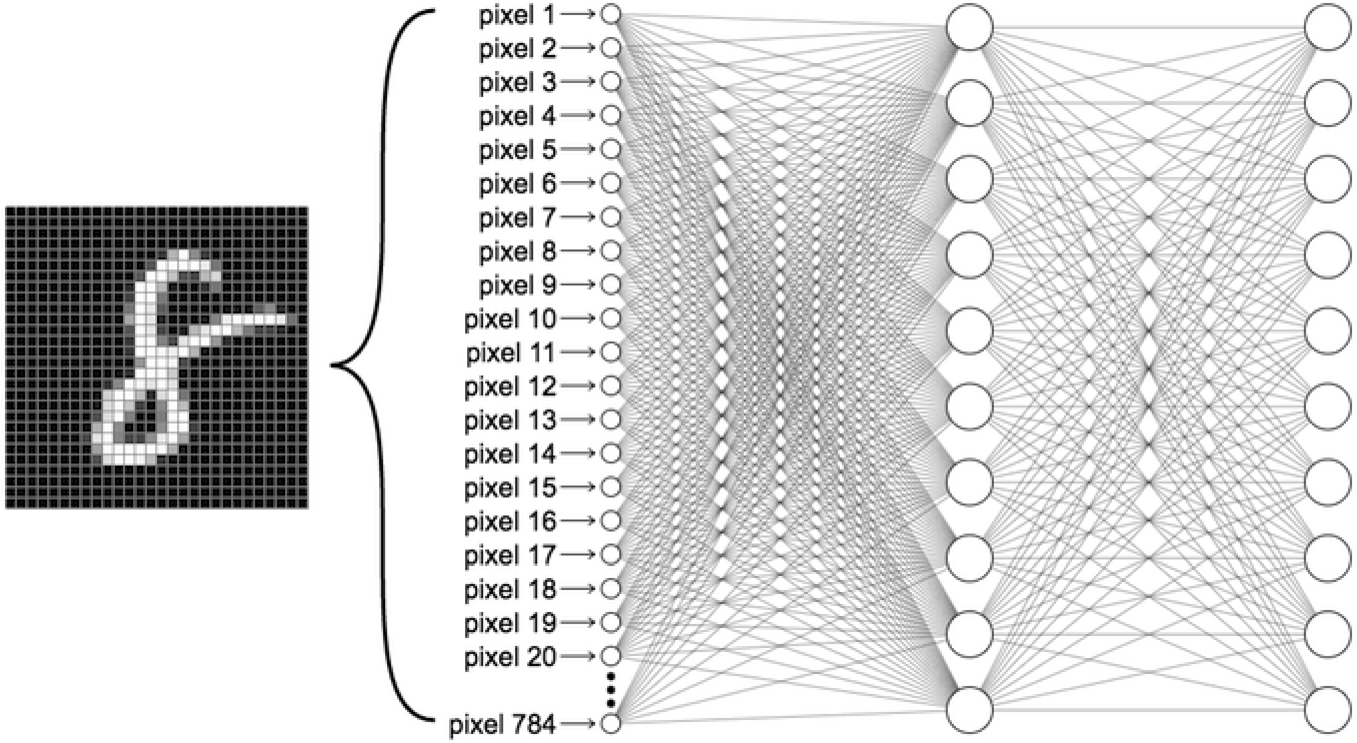
Model Training:

- Gradient Descent
- Backpropagation



Source: Designing Your Neural Networks. A Step by Step Walkthrough | by Lavanya Shukla | Towards Data Science

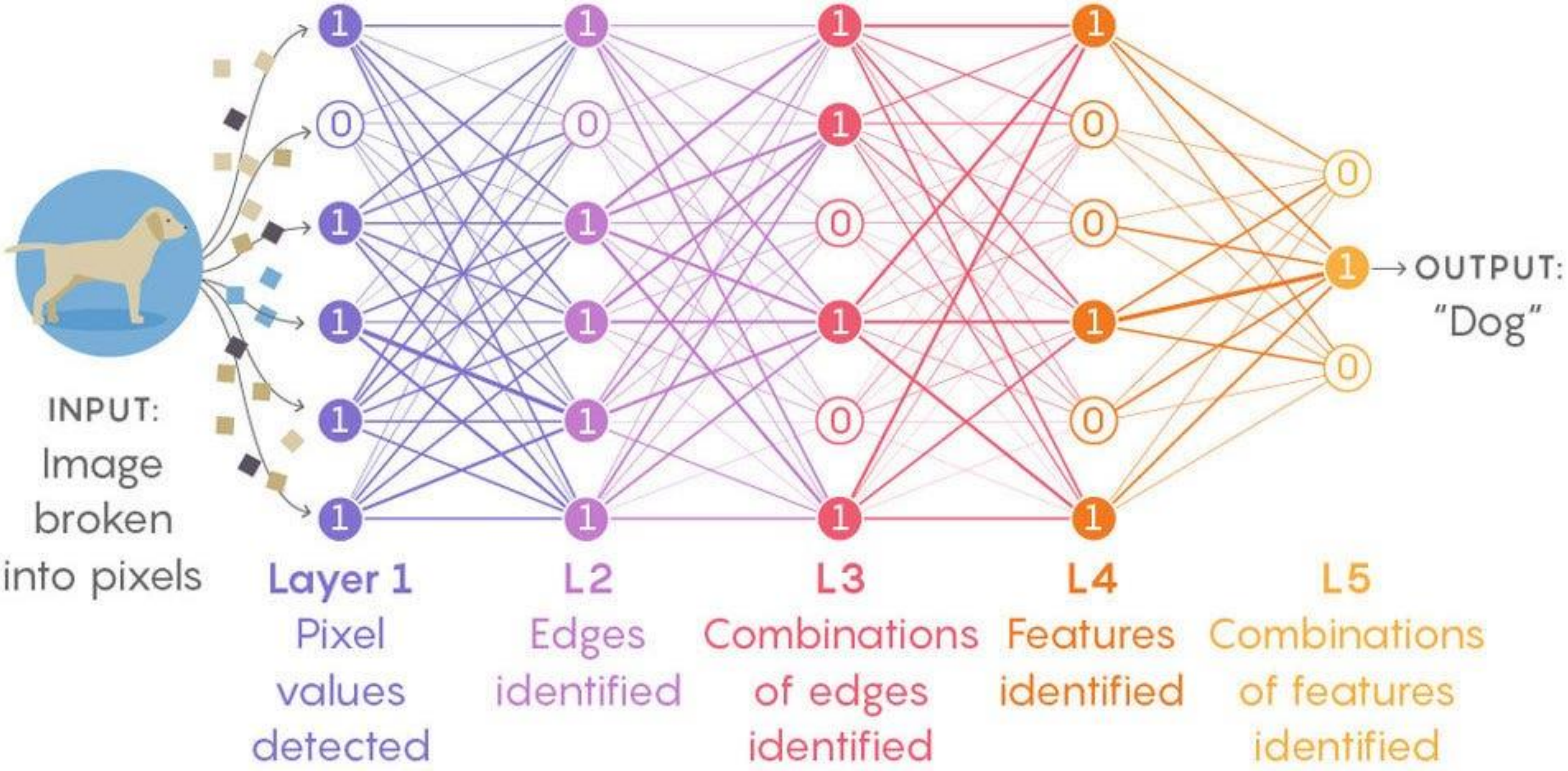
What is Deep Learning?



Handwritten Digit Recognition with a Feedforward Neural Network

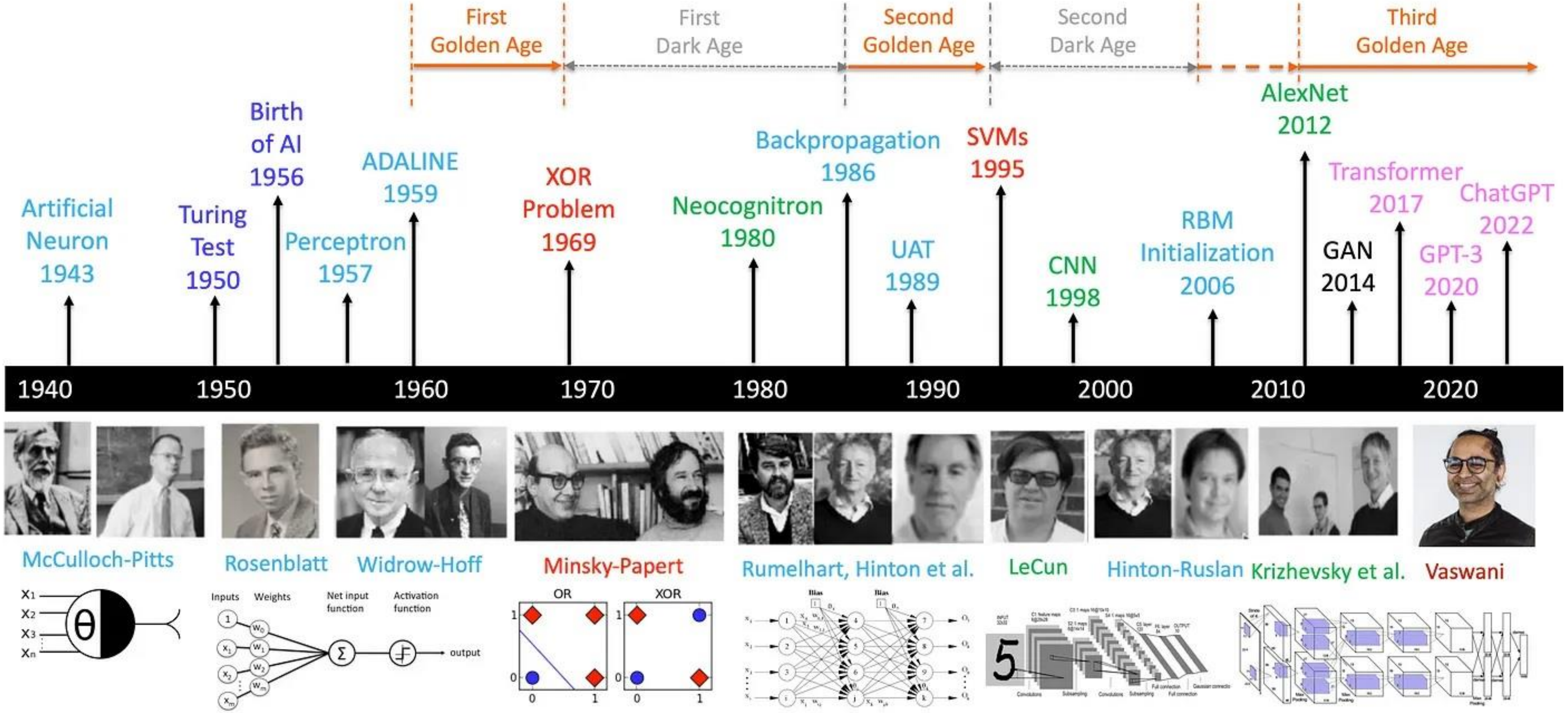
Source: Designing Your Neural Networks. A Step by Step Walkthrough | by Lavanya Shukla | Towards Data Science

What is Deep Learning?



Source: Designing Your Neural Networks. A Step by Step Walkthrough | by Lavanya Shukla | Towards Data Science

A Brief History of AI with Deep Learning



Golden Age from 2012 – Data, Compute, Algorithms

Source: A Brief History of AI with Deep Learning | by LM Po | Medium

Sequence Models

Time Series

- Forecasting
- Anomaly Detection
- Classification
- Imputation

Text

- Text classification
- Named Entity Recognition
- Sentiment Analysis
- Text Summarization
- Machine Translation
- Text Generation

Audio

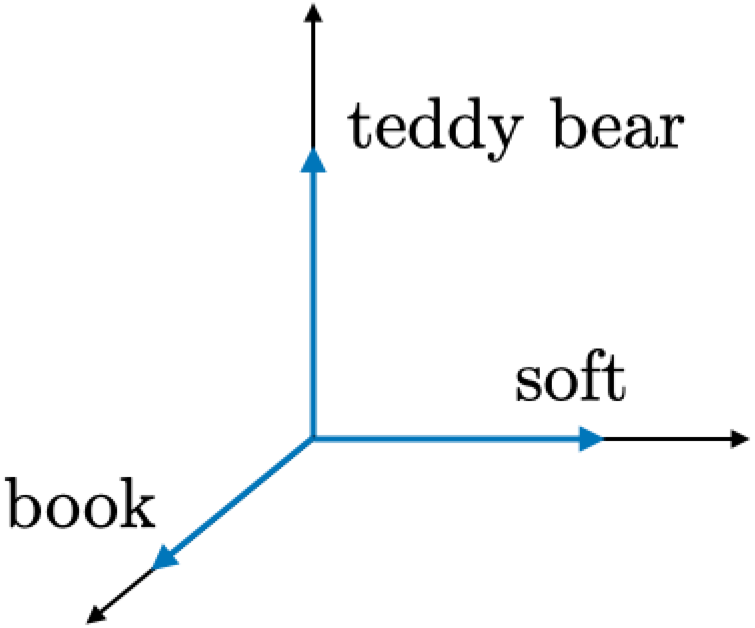
- Text-to-speech
- Speech-to-text
- Music Generation

Architectures

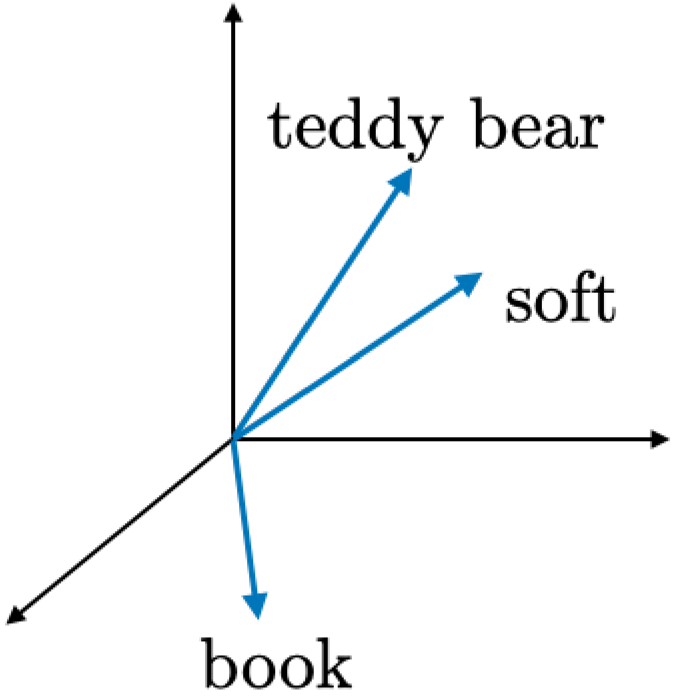
- RNN
- LSTM
- GRU
- Encoder-decoder
- Transformer

Deep Learning for Text

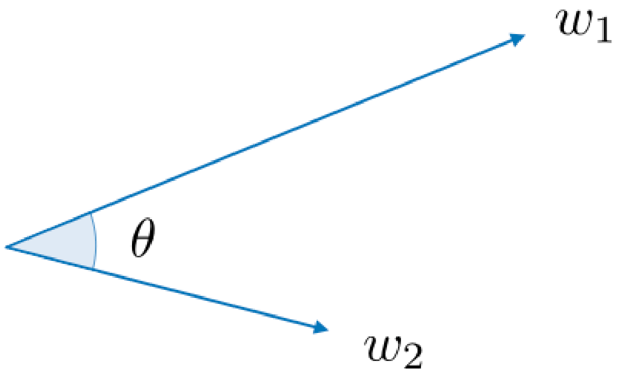
One-hot representation



Vector Embedding



Cosine Similarity

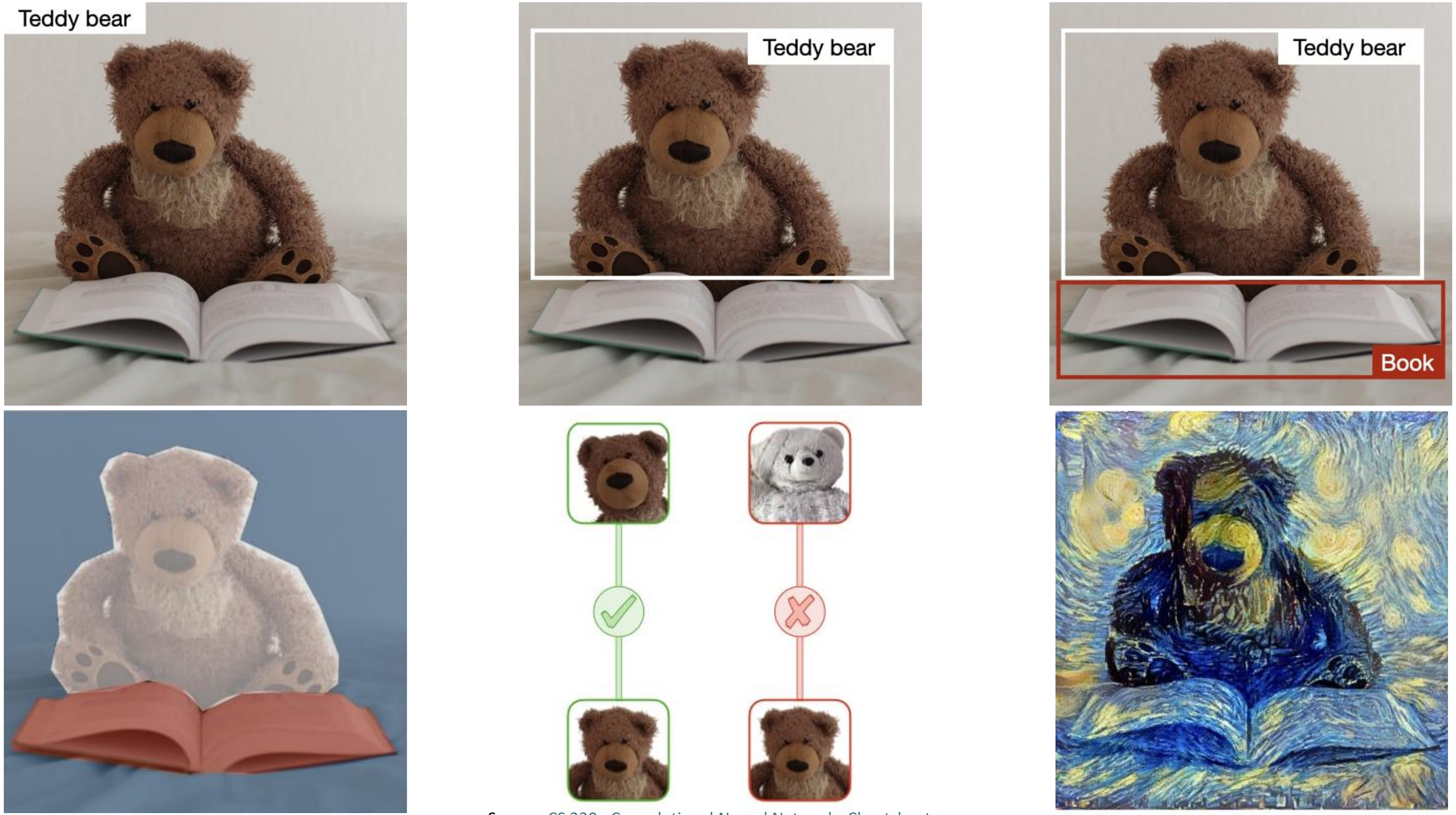


$$similarity = \frac{\omega_1 \omega_2}{\|w_1\|\|w_2\|} = \cos(\theta)$$

Embedding: words (tokens) are represented as multidimensional vectors. Directions and distances describe the relationships between words. Embeddings are optimized to best model these relationships.

Source: CS 230 - Recurrent Neural Networks Cheatsheet

Deep Learning for Vision



Source: CS 230 - Convolutional Neural Networks Cheatsheet

Computer Vision

Image

- Image recognition
- Object detection
- Semantic /Instance Segmentation
- Optical Character Recognition
- Facial Recognition

Generative

- Image Enhancement
- Style Transfer
- Image Captioning
- Video Manipulation
- Image/Video Generation

Video / 3D

- Object Tracking
- Pose estimation
- Depth Estimation
- 3D Reconstruction
- SLAM

Architectures

- Traditional CV
- Deep Learning:
  - CNN
  - Autoencoder
  - GAN
  - Diffusion Model
  - Vision Transformer
  - Multimodality

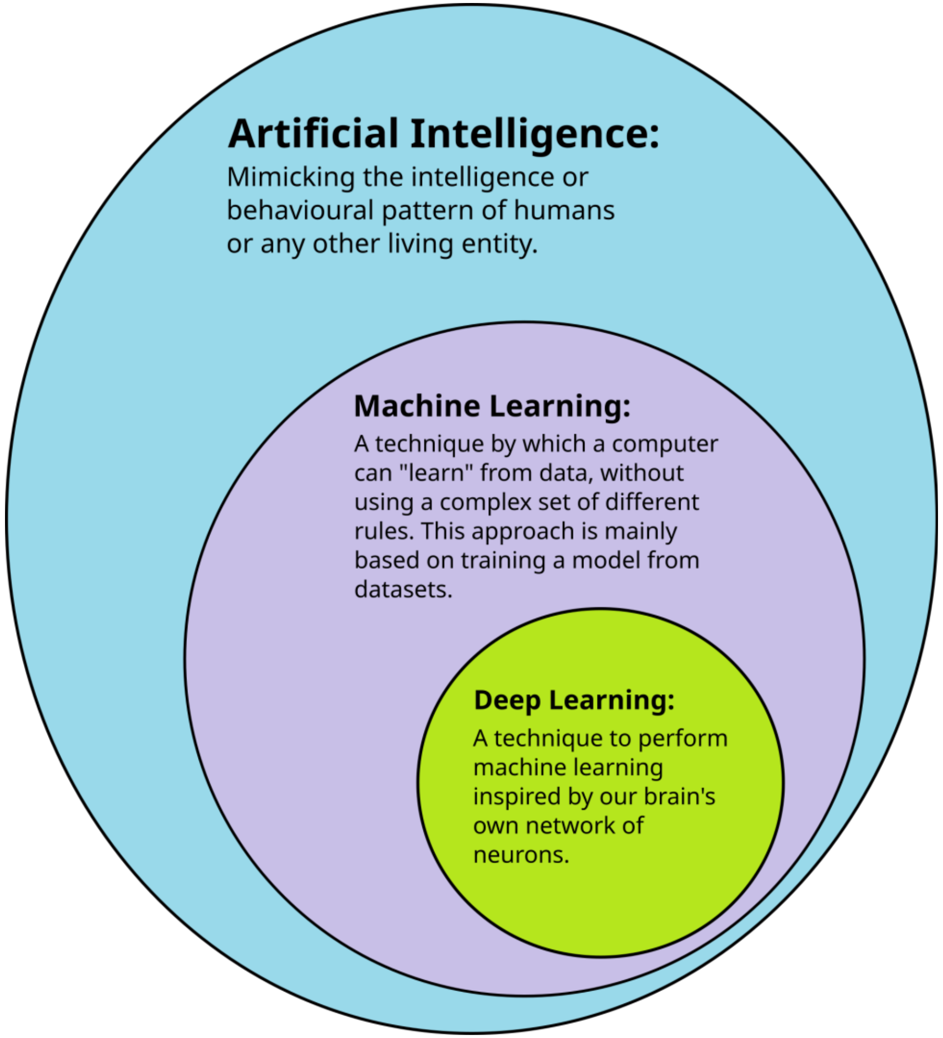
What are LLMs?

- **Deep learning models** designed and optimized specifically for conversations.
- **Transformer:** groundbreaking model architecture based on the **attention** mechanism – the word representation is influenced by the context
- **Pre-training** on large data, **post-training** on high quality data with methods like RLHF
- **System prompt:** instructions to determine how the AI model responds to the **user prompt**
- The trained weights determine the model's behaviour by storing the model's understanding of the world (memory). But this is not human intelligence – more like "autocorrect on steroids"
- Learn more:
  - 3Blue1Brown – Neural networks
  - Andrej Karpathy – Deep Dive into LLMs like ChatGPT
  - Stanford CS229 I Machine Learning I Building Large Language Models (LLMs)

Deep Learning versus Machine Learning

Deep Learning:

- Built-in feature extraction
- Better architectures for text and vision:
  - Word embedding for text
  - Convolution for images
  - Sequence modelling for time series and text
- Performs well on large amounts of data and complex problems
- Computationally expensive, but great parallelism (GPUs)



Artificial Intelligence: Mimicking the intelligence or behavioural pattern of humans or any other living entity.

Machine Learning: A technique by which a computer can "learn" from data, without using a complex set of different rules. This approach is mainly based on training a model from datasets.

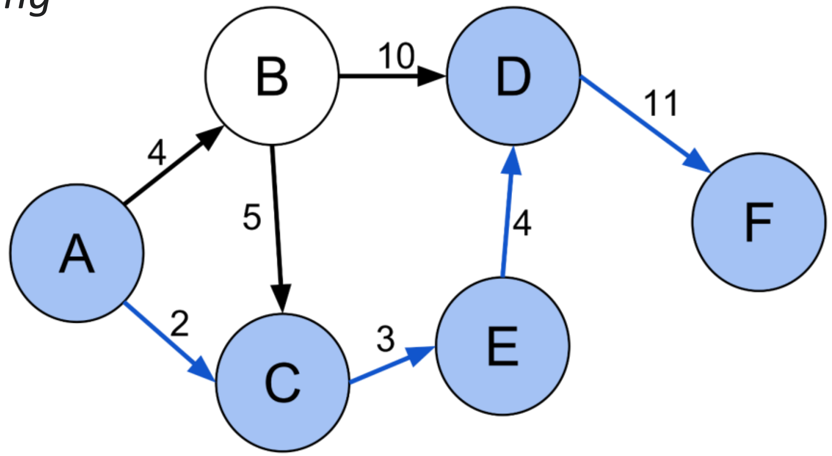
Deep Learning: A technique to perform machine learning inspired by our brain's own network of neurons.

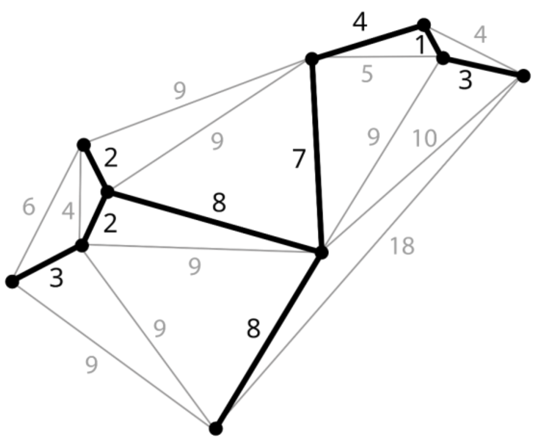
Source: Deep learning - Wikipedia

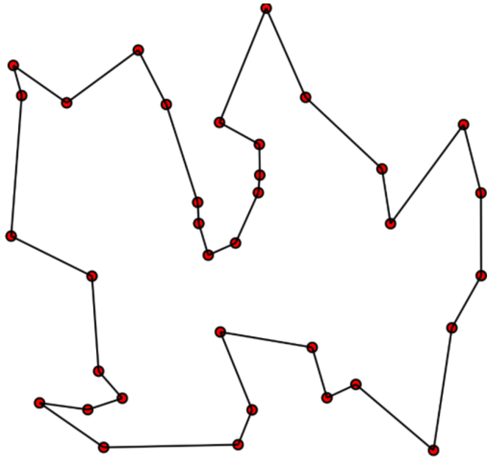
Mathematical Optimization

Maximize/minimize a target function (cost, time), or find a solution if it exists, with certain constraints, limited choices or limited computing resources.

- Search Algorithms
  - State space search
    - Shortest path
    - Minimum Spanning Tree
  - Local search
    - Hill climbing
    - Simulated Annealing
    - Travelling Salesman





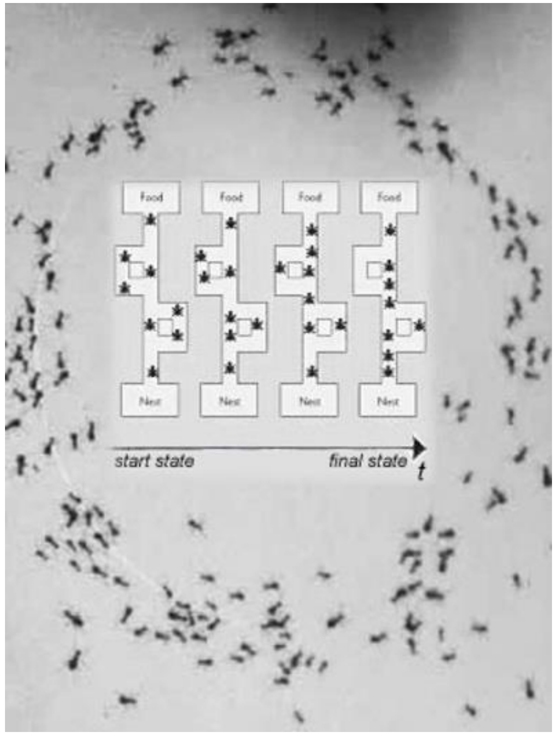


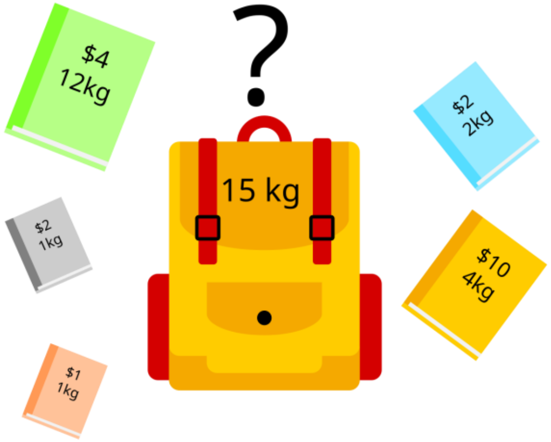
Source: Wikipedia (Shortest path problem, Minimum spanning tree, Travelling salesman problem)

Mathematical Optimization

- Evolutionary Algorithms
- Swarm Intelligence
- Logic (Knowledge Representation)
- Constrained Optimization
  - LP, NLP, IP, QP
  - Constraint Satisfaction
  - Knapsack Problem
- Stochastic Optimization
  - Monte Carlo Simulation

Learn more: CS50's Introduction to Artificial Intelligence with Python

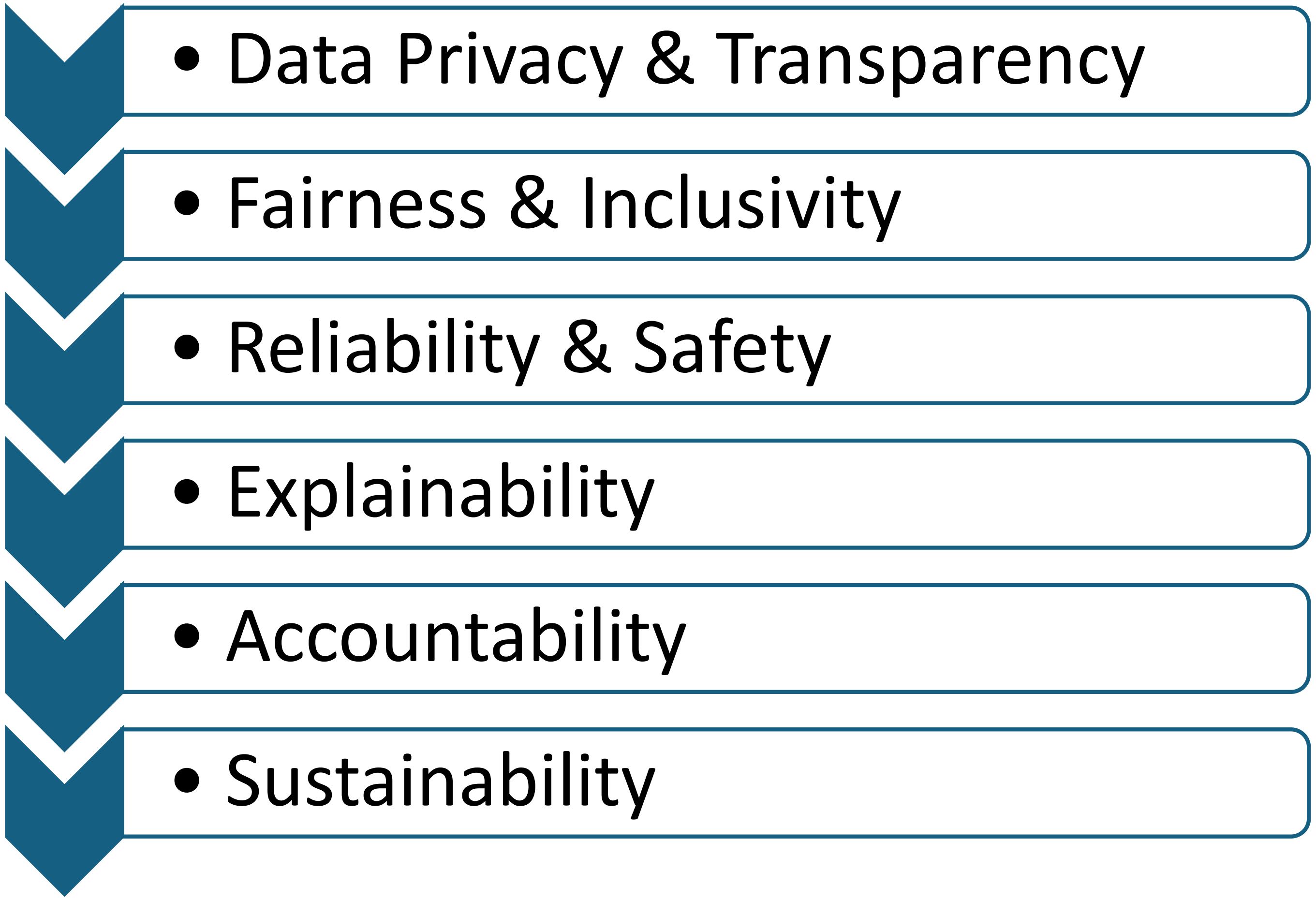




$$\max c^T x$$
$$\textit{subject to } Ax \leq b$$
$$\textit{and } x \geq 0$$

Source: Wikipedia (Ant colony optimization algorithms, Knapsack problem)

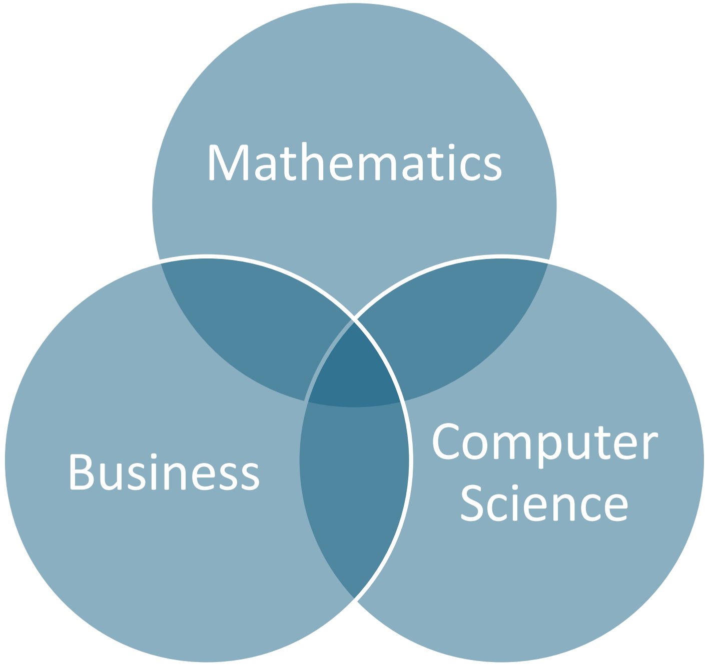
Responsible AI

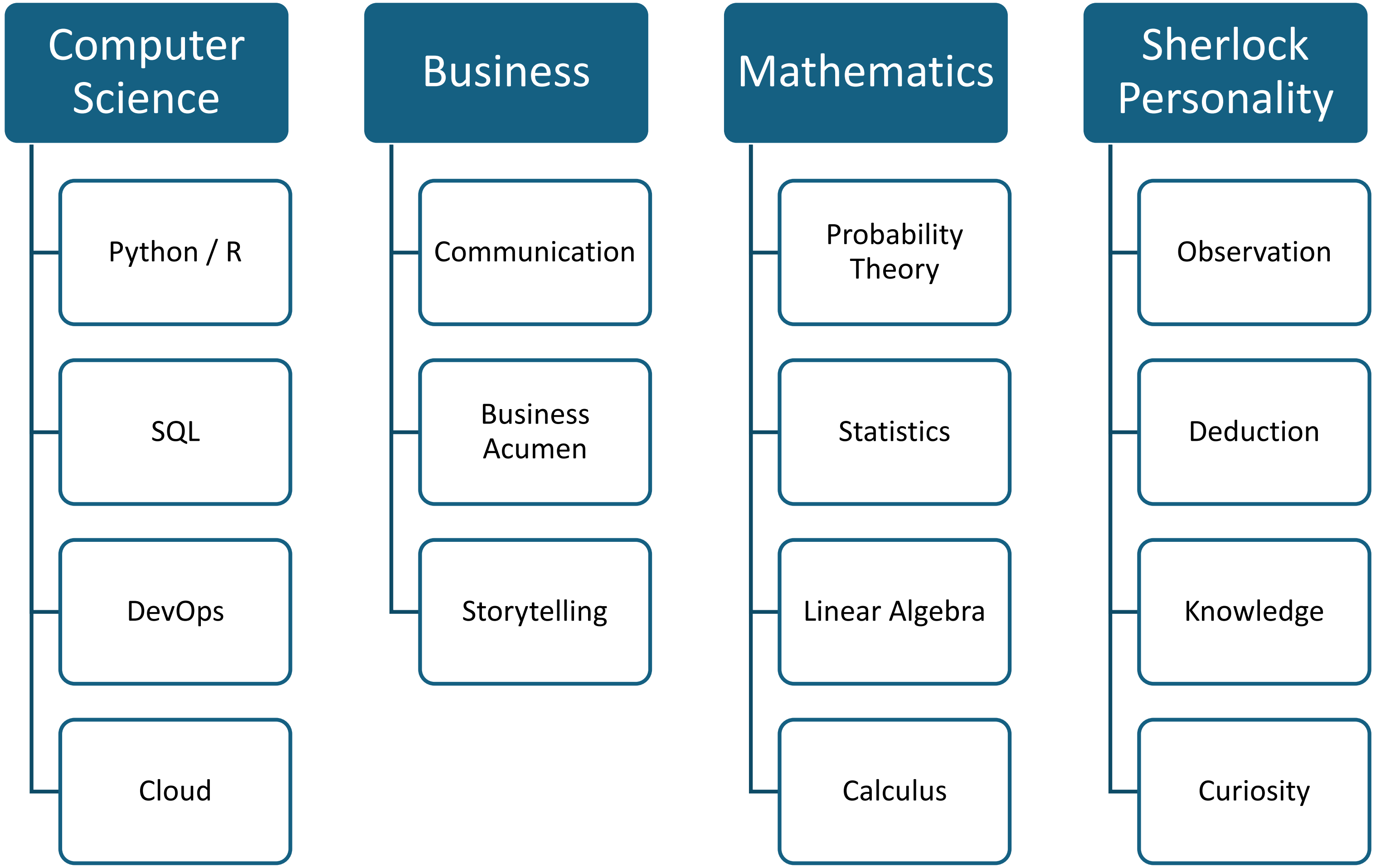


- Data Privacy & Transparency
- Fairness & Inclusivity
- Reliability & Safety
- Explainability
- Accountability
- Sustainability

Tools

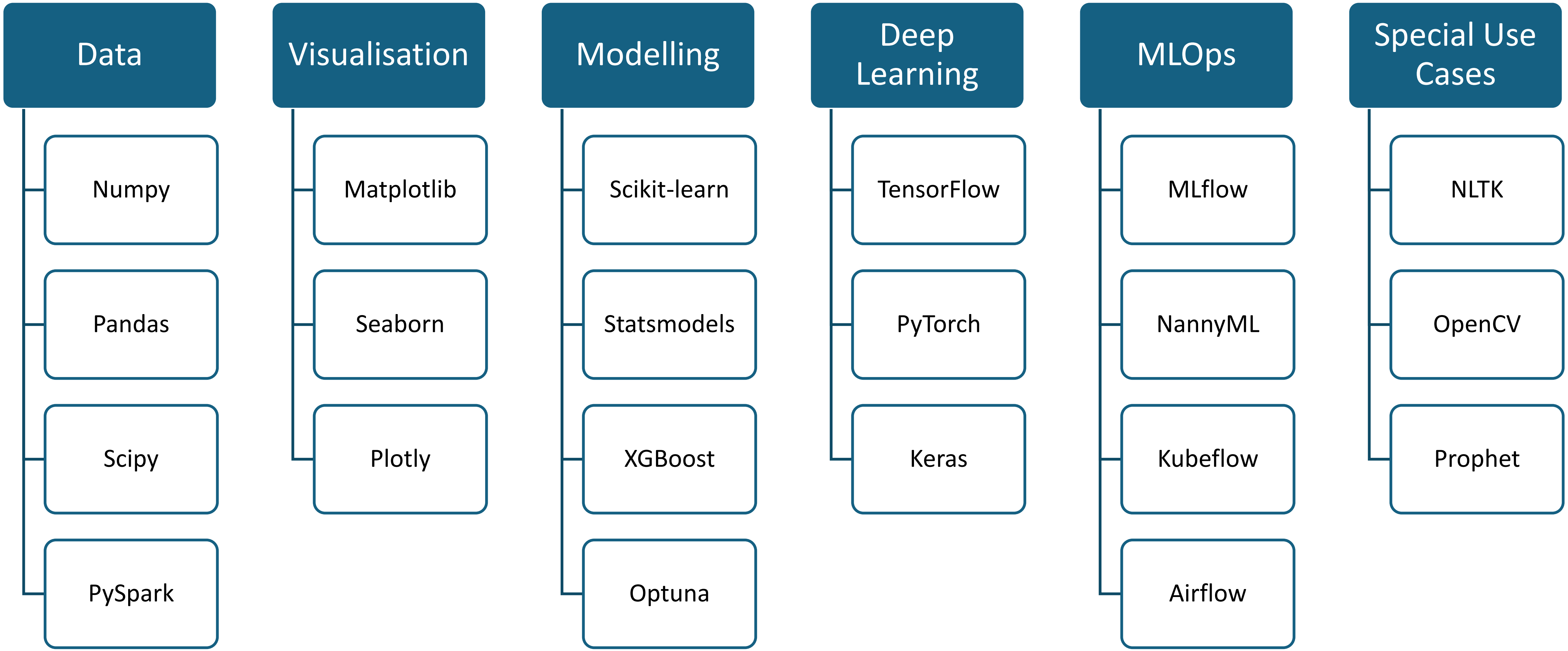
Skills for Data Science





- Computer Science
  - Python / R
  - SQL
  - DevOps
  - Cloud
- Business
  - Communication
  - Business Acumen
  - Storytelling
- Mathematics
  - Probability Theory
  - Statistics
  - Linear Algebra
  - Calculus
- Sherlock Personality
  - Observation
  - Deduction
  - Knowledge
  - Curiosity

Python for Data Science



Data
- Numpy
- Pandas
- Scipy
- PySpark

Visualisation
- Matplotlib
- Seaborn
- Plotly

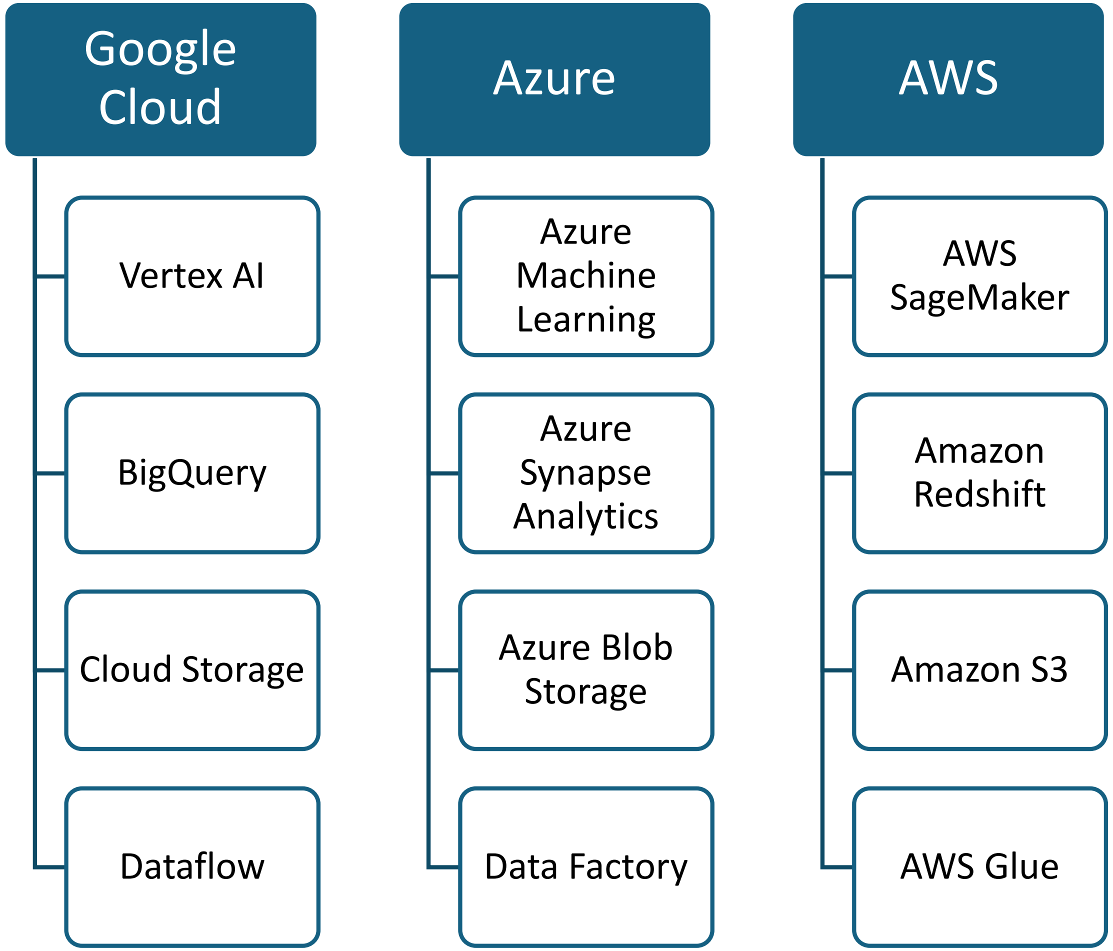
Modelling
- Scikit-learn
- Statsmodels
- XGBoost
- Optuna

Deep Learning
- TensorFlow
- PyTorch
- Keras

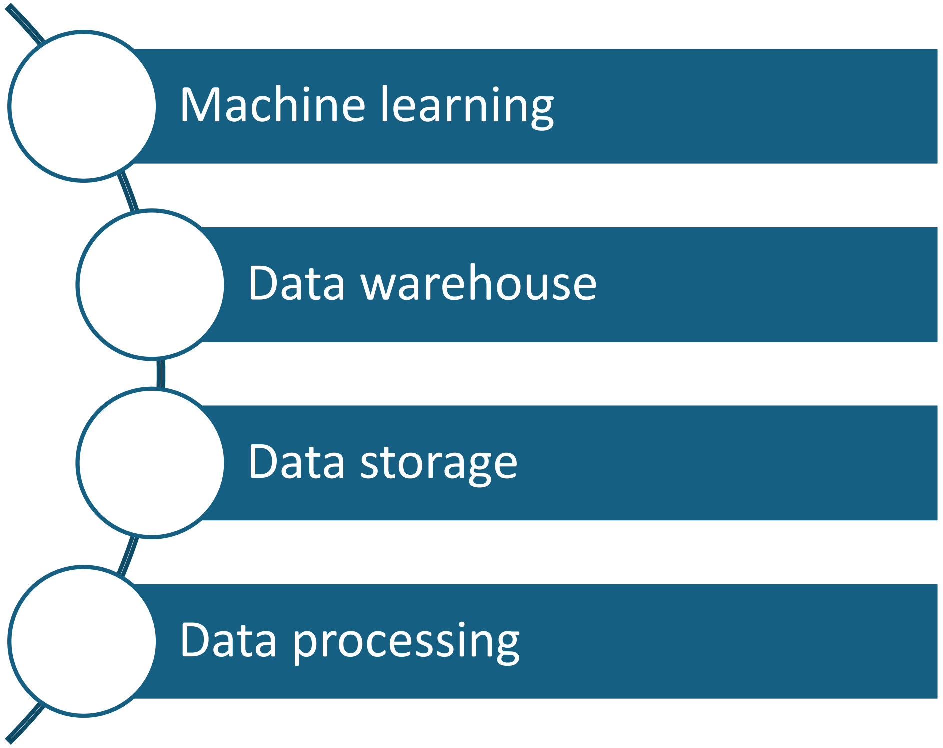
MLOps
- MLflow
- NannyML
- Kubeflow
- Airflow

Special Use Cases
- NLTK
- OpenCV
- Prophet

Cloud Platforms for Data Science



Google Cloud	Azure	AWS
Vertex AI	Azure Machine Learning	AWS SageMaker
BigQuery	Azure Synapse Analytics	Amazon Redshift
Cloud Storage	Azure Blob Storage	Amazon S3
Dataflow	Data Factory	AWS Glue



- Machine learning
- Data warehouse
- Data storage
- Data processing

Introduction to Data Science

I. Introduction to Data Science

II. Business and Data Understanding

III. Introduction to Supervised Learning

IV. Advanced Supervised Learning

V. Unsupervised Learning

VI. Time Series Analysis

VII. Deep Learning

VIII. Machine Learning Operations

- Data Science
- Machine Learning
- The ML Lifecycle
- ML Algorithms
- The Goal
- Deep Learning
- Tools

Thank you for your attention!

Your feedback would be much appreciated:

Any Questions?

Gergely Zsombor Haász

haasz.zsombi@gmail.com