Artificial Intelligence

Corso di Laurea in Computational Finance
2nd semester - 9 CFU

Luca Pajola, Luca Pasa, Elisa Tosetti

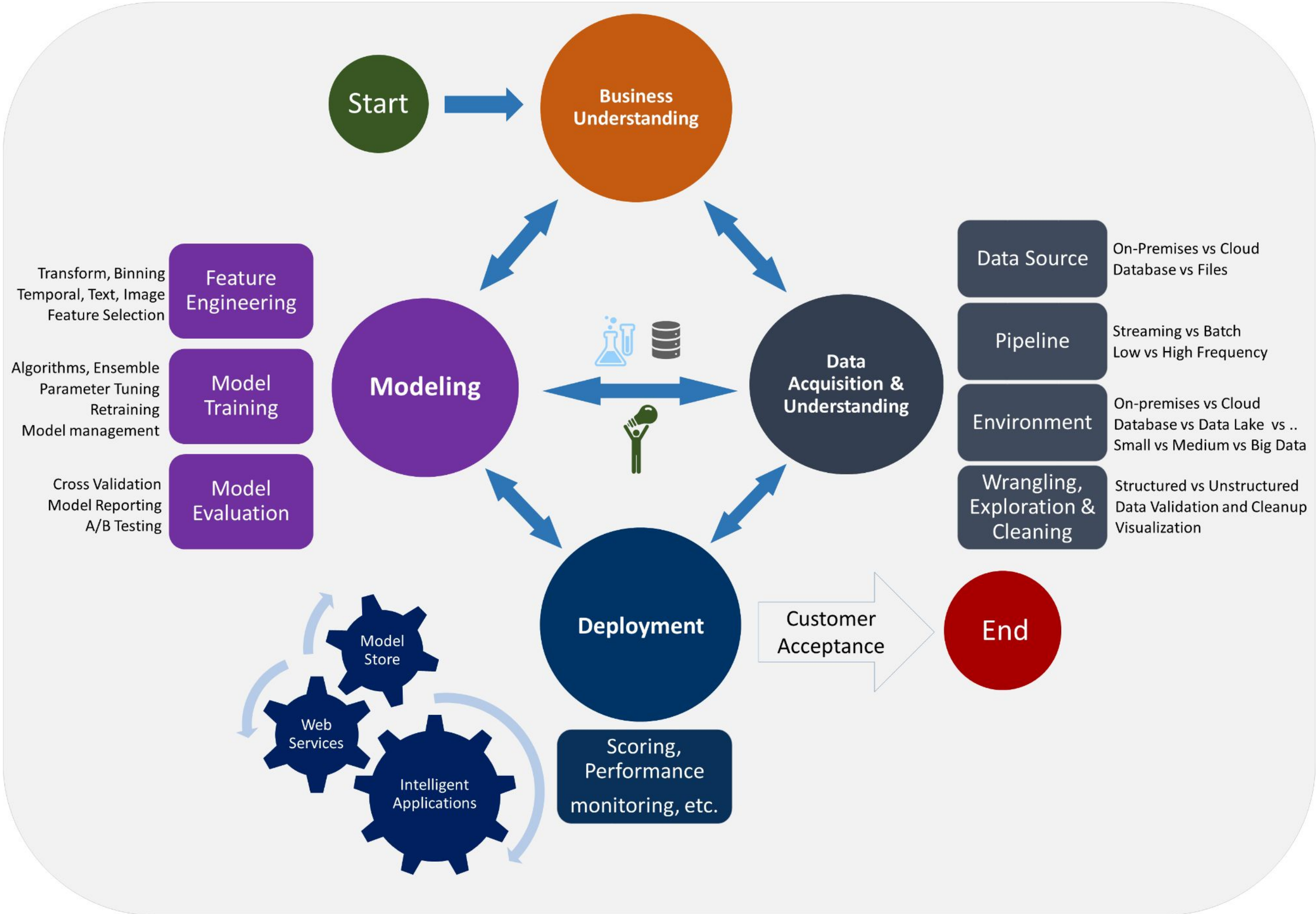
AI Pipeline

Let’s start …

- What is a “daily” job of an AI Engineer?

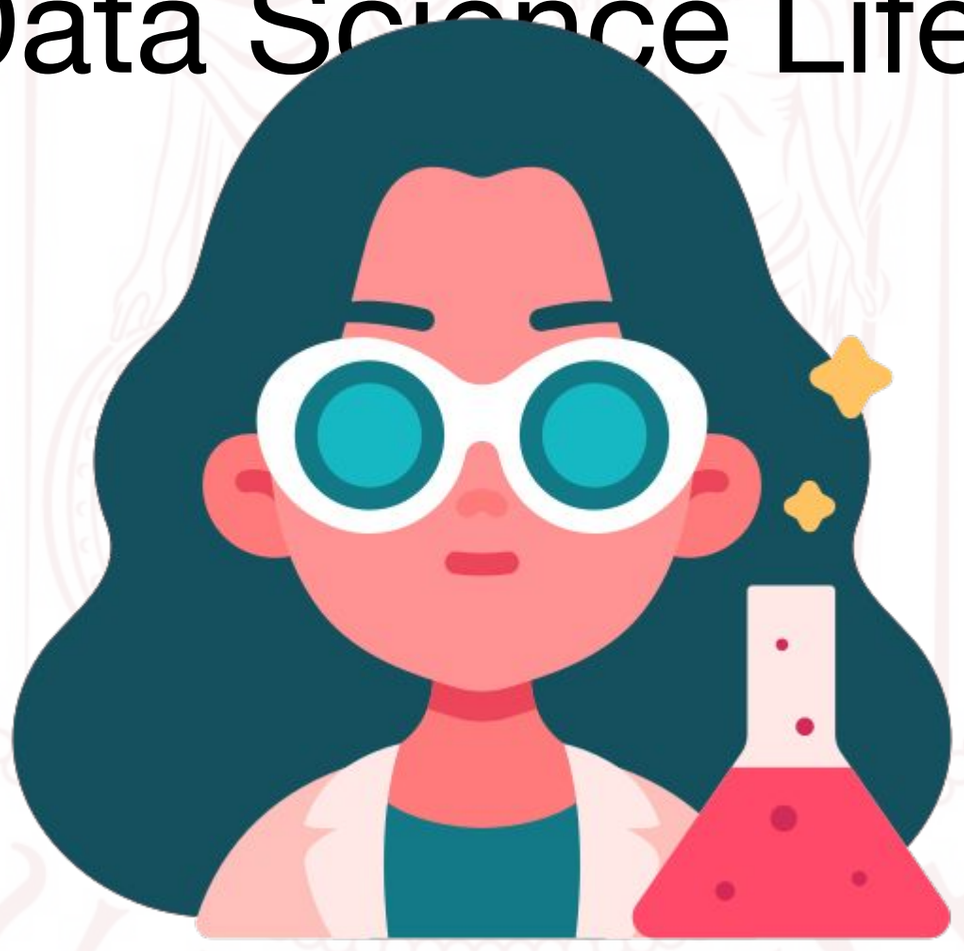
Let’s start …

- The role of a data scientist is **manifold**
- Often confused with “training a model”
- … the development is much more complex!



Source: Microsoft

Pt.1: Data Science Life-cycle



5 Steps

1. Business Understanding
2. Data Acquisition and Understanding
3. Modeling
4. Deployment
5. Customer Acceptance

Business Understanding

- Two tasks addressed in this stage
  - Define the objectives
  - Identify data sources
- Objectives
  - Formalize the **model targets**
    - E.g., sales forecast, order being fraudulent
  - And the appropriate **metrics**
    - E.g., model accuracy, efficiency

Business Understanding

- Examples of questions
  a. How much or how many? (regression)
  b. Which category? (classification)
  c. Which group? (clustering)
  d. Is this weird? (anomaly detection)
  e. Which option should be taken? (recommendation)

Business Understanding

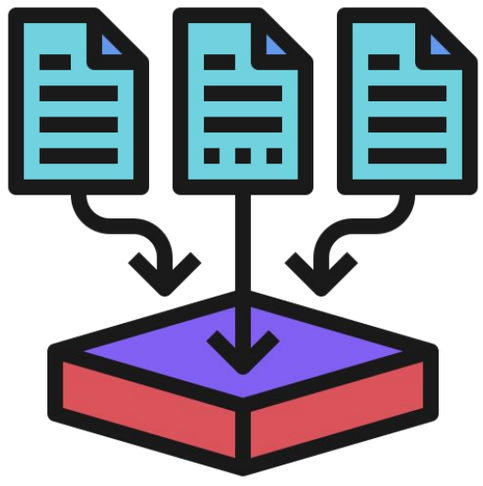
- The SMART metric to evaluate a project
  a. Specific - define a clear goal
  b. Measurable - that can be measured
  c. Achievable - and achievable with the resources
  d. Relevant - what you achieve must be relevant
  e. Time-bound - a clear timeline to respect

Business Understanding

- Identify the data sources
  a. Data that's relevant to address the task
- Understand if something else is needed
  a. Additional data?
  b. Validity check?
  c. Labelling operations?
  d. Internal vs external data sources

Data Acquisition & Understanding

- Three tasks addressed in this stage
  a. Ingest the data
  b. Explore the Data
  c. Set up a **data pipeline**
- Different type of data pipeline
  a. E.g., Batch-based, real-time

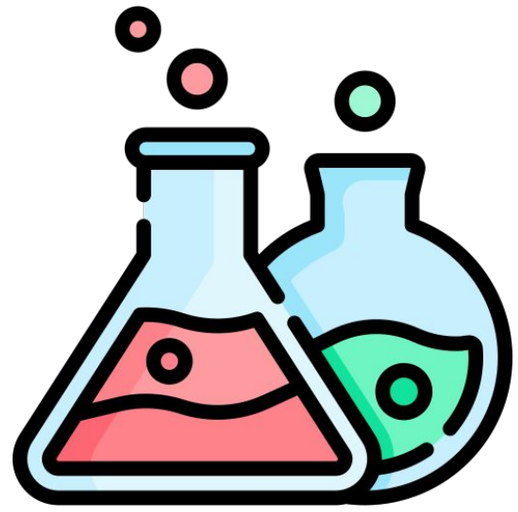


Modeling

- Steps addressed
  - Feature engineering
  - Model training
  - Validation

- Feature engineering
  - Transform the raw data in features utilized in the analysis
  - Essential to understand the problem to produce a valid feature set
  - Requires **domain-experts**

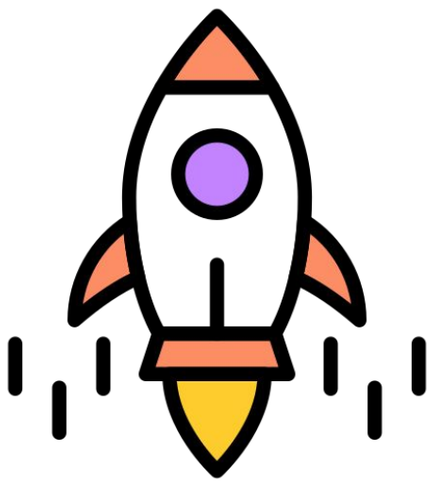
Modeling



- Model training
  - Different models (techniques)
    i. Based on what we are trying to answer
  - Steps:
    i. Split the data for training and evaluation purposes
    ii. Train models
    iii. Evaluate models
    iv. Determine which model is more suitable
- Model Evaluation
  - Does the model satisfy the production requirement?
  - Should we try alternative approaches?
  - Debug models
  - Evaluate different aspects of the model

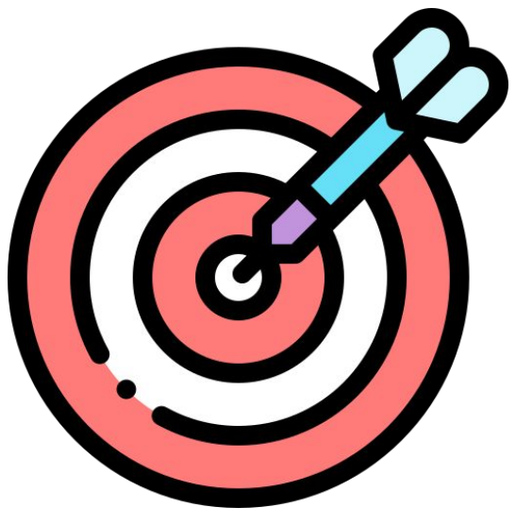
Deployment

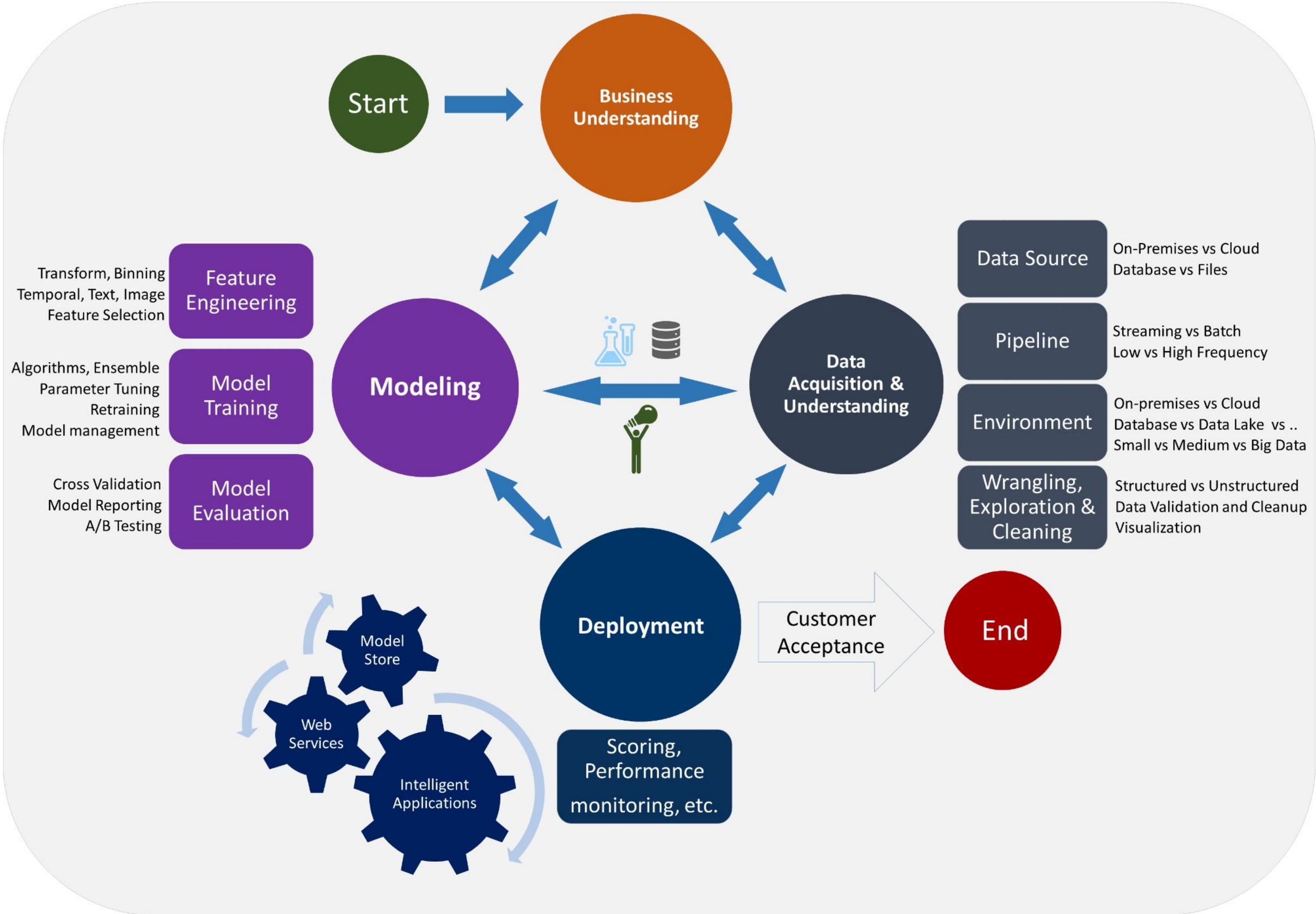
- Goal
  - Operationalize a model
- What to do
  - Move the model into a realistic environment
    - E.g., server
  - Develop an open API interface



Customer Acceptance

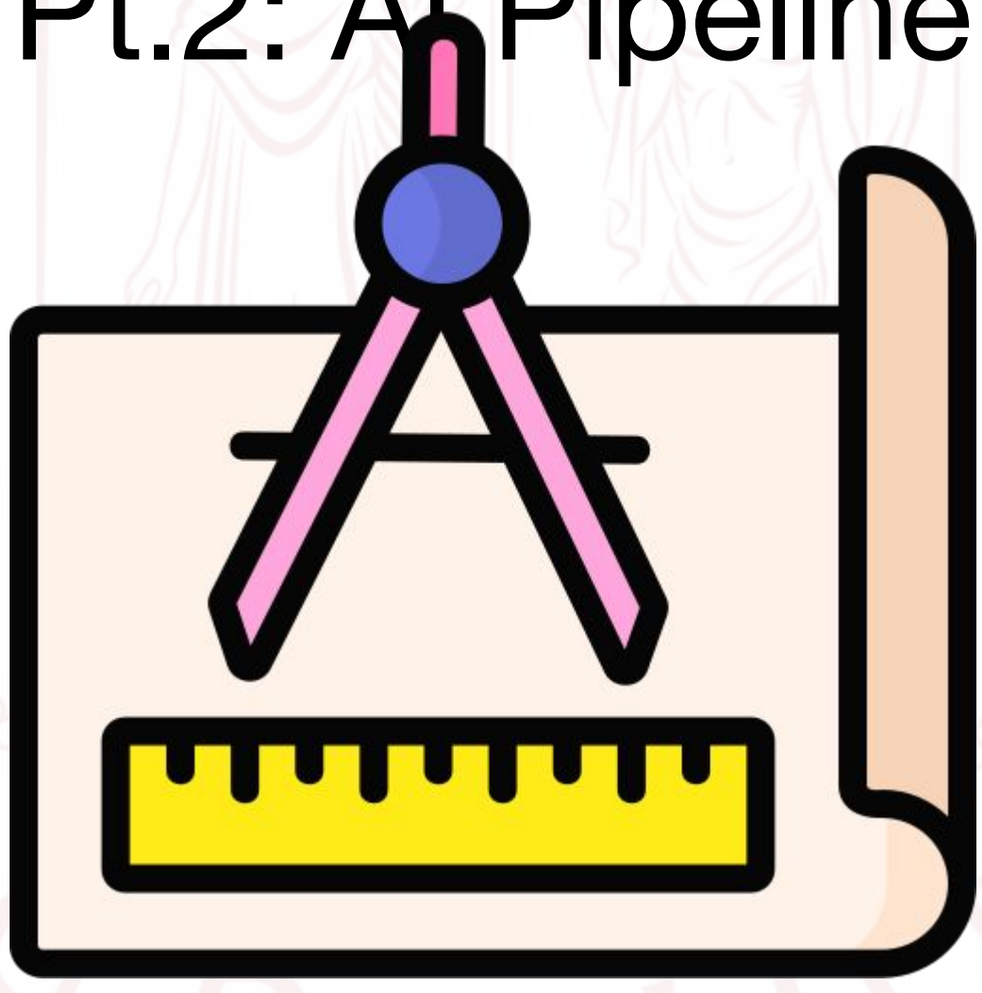
- Goal
  - Finalize project deliverables
- What to produce
  - System Validation with customer's need
  - Project hand-off





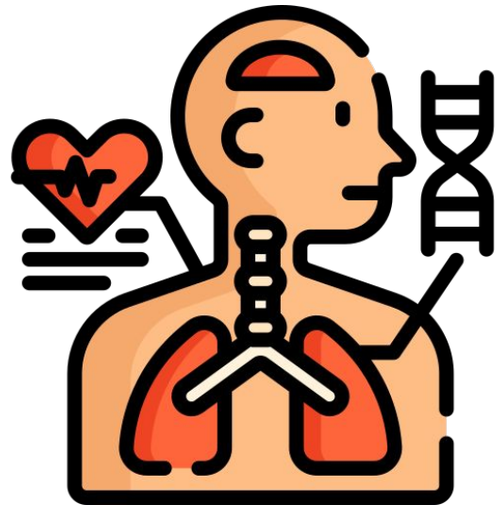
Source: Microsoft

Pt.2: AI Pipeline



The Machine Learning Anatomy

- When we design an AI application we talk about a complex system
  - Many operations are executed behind the scenes
- Many steps
  - Preprocessing
  - Feature Engineering
  - AI-Algorithm
  - Post Processing
  - Visualization and Dashboard
- Each essential for the overall AI execution



Preprocessing

- Goal of preprocessing
  - From raw data to data utilizable by an automatized process
- Many techniques can be applied
  - Based on the nature of the data
- **Data Cleaning**
  - process of fixing or removing
    - Incomplete data
    - Incorrectly formatted data

Preprocessing

Incorrect or inconsistent data leads to false conclusions

Preprocessing

- Data quality
  - Validity - data conform to rules (temperature of room cannot be 1000 celsius)
  - Accuracy - a zip code that does not exist
  - Completeness - any missing data?
  - Consistency - multiple features are consistent? (someone 10 years old cannot be as marital status “divorced”)
  - Uniformity - is data written with the same unit measure (e.g., everything expressed in Kg)
- How to perform data quality assurance?
  - Inspect
  - Correct
  - Verify
  - Report

Preprocessing

- Inspection
  - Data profiling or summary statistics
  - Visualization
  - Software Packages
- Cleaning
  - Irrelevant data
  - Duplicates
  - Type Conversion
  - Syntax Errors
  - Standardize
  - Scaling
  - Missing Values
    - Drop, Impute
  - Outlier

Preprocessing

- Verification
  - After fixing data, you re-verify
  - After filling data, your data might violate some rules
- Reporting
  - Having a document that log errors that occurred and the type of actions you took to clean the data
  - Essential to better understand

Feature Engineering

Feature engineering in refers to the process of selecting, transforming, and manipulating raw data into features that can be used in machine learning models

Feature Engineering

- Feature: any measurable input that can be used in a predictive model
- In FE we transform raw data into expresfull representations
  - Often requires domain experts
- Different type of operations
  - Feature creation: age -> is legal age
  - Transformation: weight of a person expressed in grams -> kg
  - Feature Extraction
  - Exploratory Data Analysis
- Some overlap with the data cleaning

Machine Learning Models

- A machine model is a tool
  - Learns to solve a task based on some data
- There are many families of ML algorithms
  - Each has some properties
  - And can solve a determinate set of tasks
- Tasks is the objective
  - How much or how many? (regression)
  - Which category? (classification)
  - Which group? (clustering)
  - Is this weird? (anomaly detection)
  - Which option should be taken? (recommendation)

Machine Learning Models

- Terminology
  - Hyper-parameters: values that defines the ML algorithm structure
  - Parameters: values that are learned from the data to solve a specific task
- Two major families of tasks
  - Supervised
    - X is the dataset
    - y is the ground-truth
  - Unsupervised
    - Only X

Machine Learning Models

- Evaluation: the process of evaluating the quality of a ML model
- Different aspects can be measured
  - E.g., performance, efficiency
- For instance, in a classification task
  - Accuracy = 100 * (#corr / #errors)
- Evaluation metrics differ based on the type of task

Post Processing

- Once the model is trained we debug it
  - Errors analysis
  - Explainable AI
  - Model robustness
  - Bias analyses
- Definition of interfaces
  - REST API
  - Dashboards

Visualization and Dashboard

- Data visualization is the graphical representation of some data
  - E.g., charts, maps, graphs
- Multipurpose
  - Understand the data and how it is distributed
  - Analyze the data to spot insights
  - Communicate insights
- A dashboard is a visualization that:
  - Is real-time
  - Is interactive
- Data visualization is dangerous
  - And often deceiving

If you are curious about

- https://www.youtube.com/watch?v=E91bGT9BjYk
- https://www.youtube.com/watch?v=bVG2OQp6jEQ
- https://www.youtube.com/watch?v=DcYLT37ImBY

Lecture Sources

- https://learn.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle
- https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4
- https://builtin.com/articles/feature-engineering