Introduction

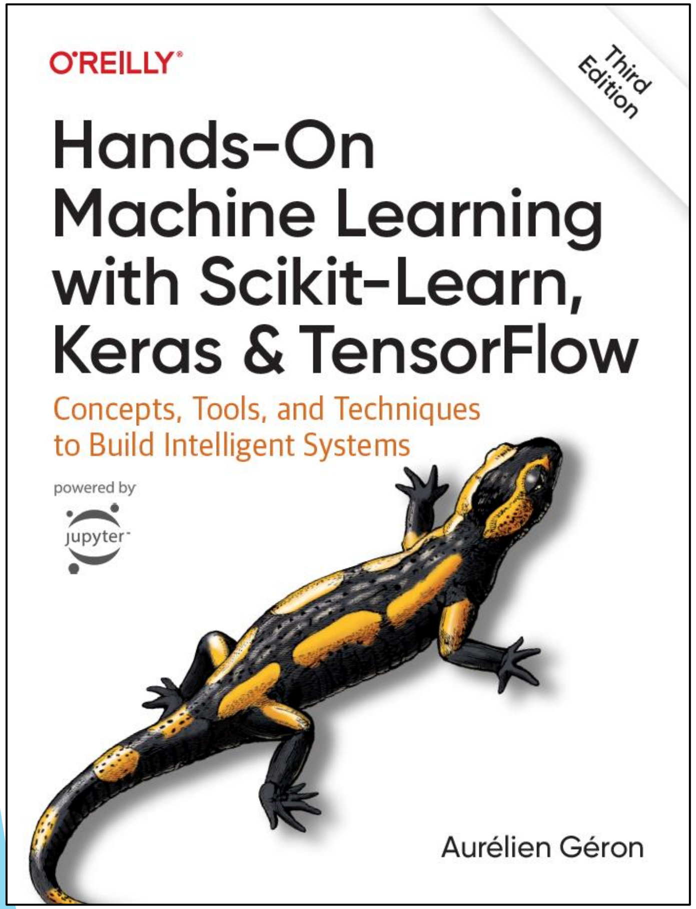
Fall 2024

Maryam Abdolali

Syllabus – Topics we are going to cover!

- **Feature Engineering**
  - Cleaning & Transforming data
- **Association Rule Mining**
  - Apriori
  - Eclat
- **Mining patterns using machine learning**
  - Supervised
    - k-Nearest Neighbors
    - Linear Regression
    - Logistic Regression
    - Support Vector Machines (SVMs)
    - Decision Trees and Random Forests
    - Neural networks
  - Unsupervised
    - Clustering
      - K-Means
      - DBSCAN
      - Hierarchical Cluster Analysis (HCA)
    - Visualization and dimensionality reduction
      - Principal Component Analysis (PCA)
      - t-Distributed Stochastic Neighbor Embedding (t-SNE)
- **Anomaly detection and novelty detection**
  - One-class SVM
  - Isolation Forest

Main Textbooks & Grading


O'REILLY®

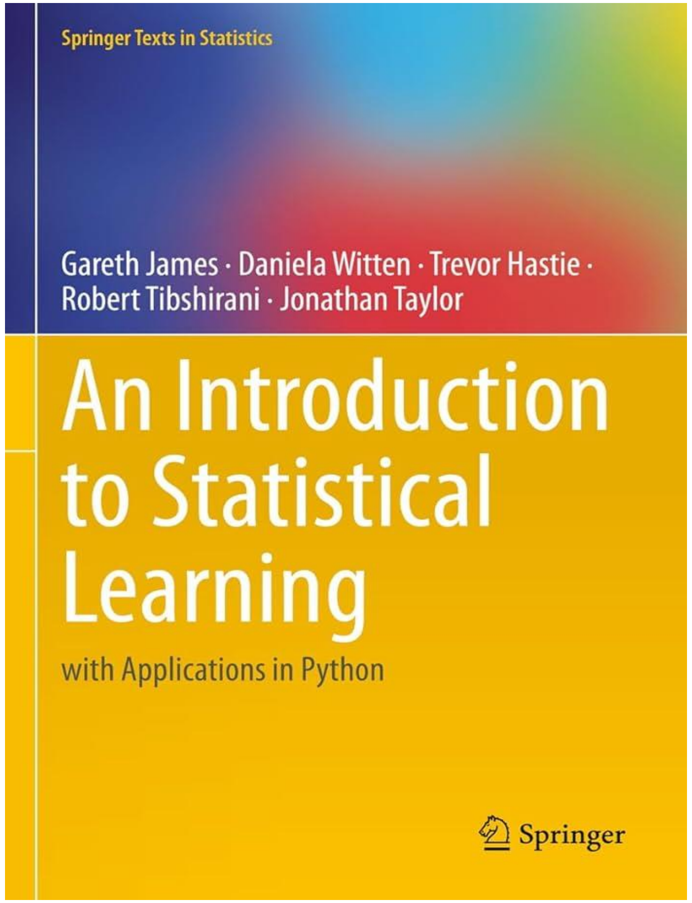
Third Edition

Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques to Build Intelligent Systems

powered by jupyter

Aurélien Géron


Springer Texts in Statistics

Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor

An Introduction to Statistical Learning

with Applications in Python

Springer

Grading:

- Final Exam 40%
- HWs 30%
- Final Project 30%
- BONUS: Surprise practical problem-solving!

This semester, we will dive deeply into practical programming, emphasizing the hands-on use of concepts through Scikit-Learn (and PyTorch)

What is even Data Mining?

Data mining is the process of **discovering/mining patterns** in **large data sets** involving methods at the intersection of machine learning, statistics and database systems

data mining can transform raw data into valuable insights

Pattern Discovery
A retail store analyzes customer purchase data and discovers that people who buy bread often also buy butter.
Apriori Algorithm

Predictive Analysis
An e-commerce website uses historical data on customer behavior to predict which products are likely to be popular in the upcoming holiday season, allowing them to stock accordingly.
Linear Regression, ARIMA

Decision Making
A bank uses data mining to analyze loan applicants' to decide on whom to approve for loans.
ML: Decision Trees

Anomaly Detection
A credit card company uses data mining to detect fraudulent transactions.
One-Class SVM

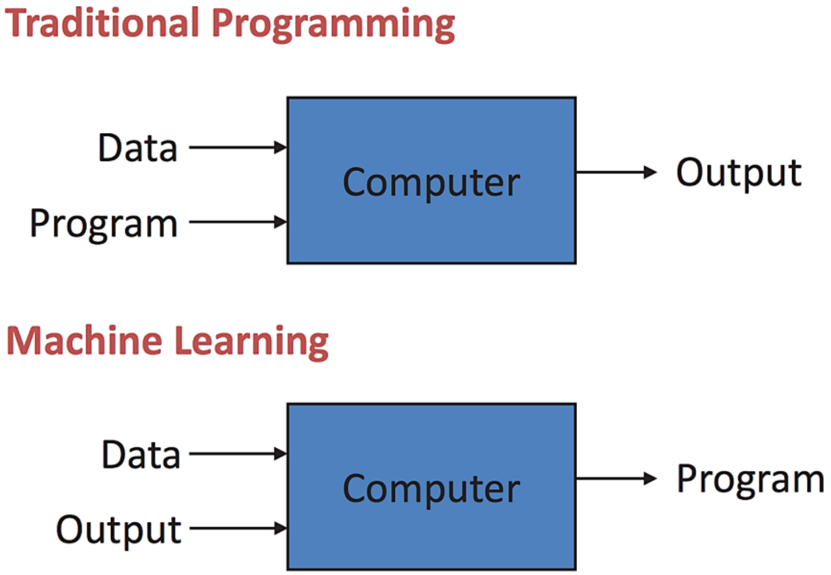
Knowledge Discovery
In healthcare, researchers analyze patient data to discover that certain lifestyle factors significantly increase the risk of developing diabetes.
Statistical Analysis (e.g., t-test)

Why do we need to “mine”?

- **The Explosive Growth of Data:**
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube, social media, mobile devices, …
- ***We are drowning in data, but starving for knowledge!***

What is Machine Learning?

Machine learning is the science (and art) of programming computers so they can ***learn from data***.



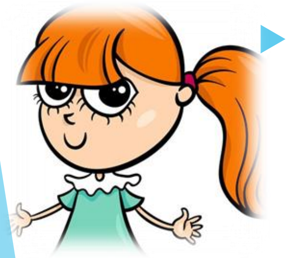
- What exactly does it mean for a machine to ***learn*** something?
  - I downloaded a copy of Wikipedia, has my computer really learned something?

The goal of ML: Generalization

- Real world Example

- Consider two college students diligently preparing for their final exam.

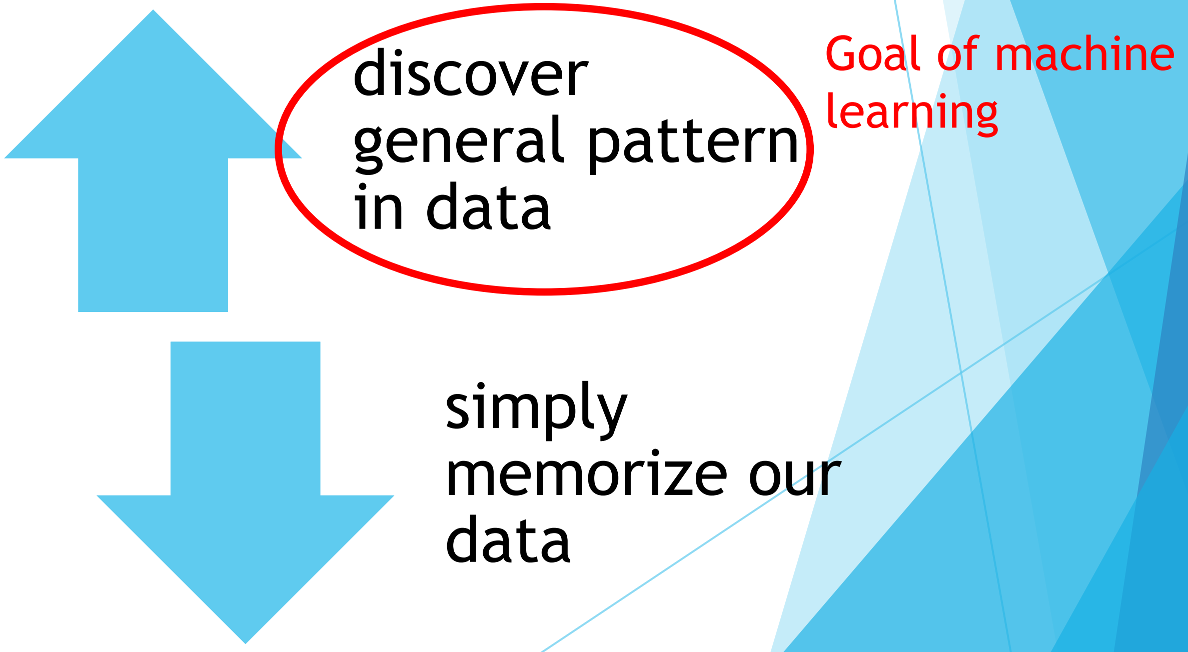
  - Extraordinary Ellie:
    - whose preparation consisted entirely of memorizing the answers to previous years' exam questions.
    - Ellie has an extraordinary memory, and thus could perfectly recall the answer to any *previously seen* question, she might nevertheless freeze when faced with a new (*previously unseen*) question.



  - Inductive Irene:
    - with comparably poor memorization skills, but a knack for picking up patterns.

- If the exam truly consisted of recycled questions from a previous year, Ellie would handily outperform Irene.

- However, even if the exam consisted entirely of fresh questions, Irene might maintain her 90% average.



discover general pattern in data

Goal of machine learning

simply memorize our data

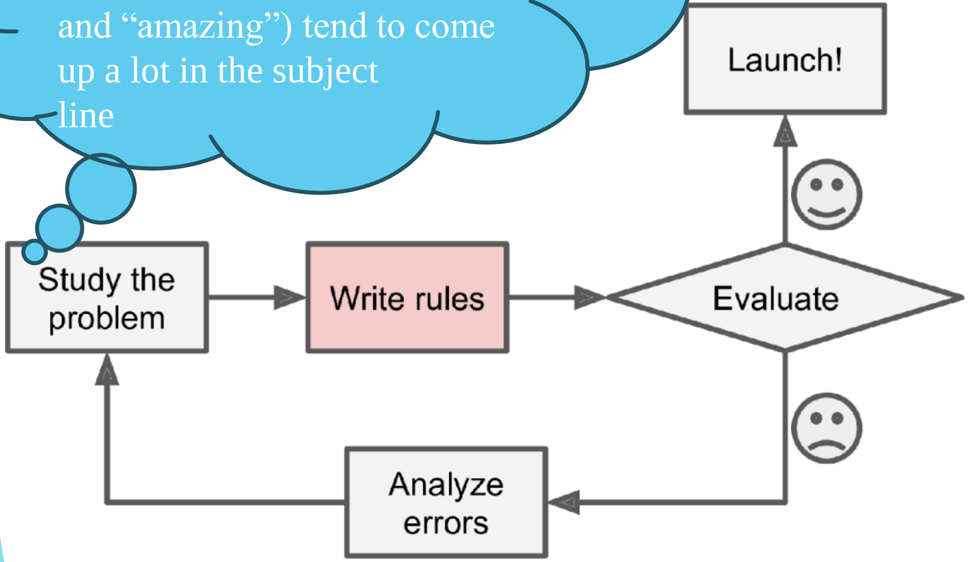
Why ML?

Example: Spam Filtering

Traditional

what spam typically looks like?

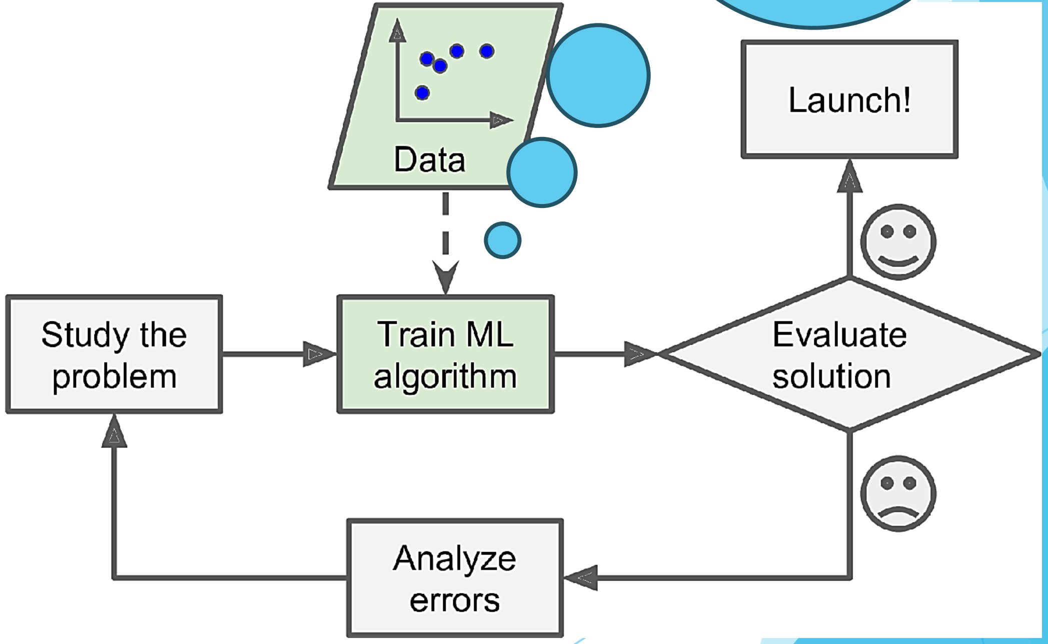
some words or phrases (such as “4U,” “credit card,” “free,” and “amazing”) tend to come up a lot in the subject line



your program will likely become a long list of complex rules—pretty hard to maintain

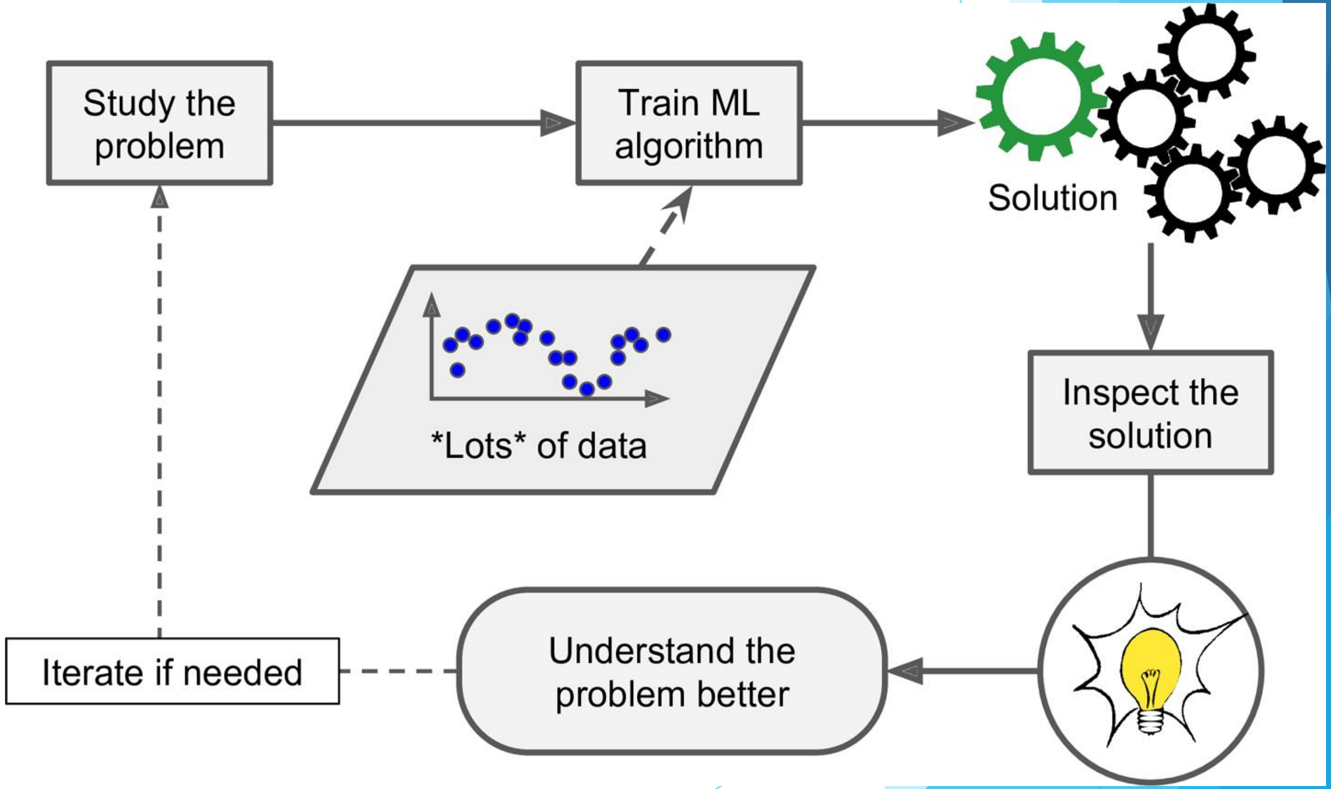
ML

Automatically learns which words and phrases are good predictors of spam by detecting unusually frequent patterns of words in the spam examples compared to the ham examples.



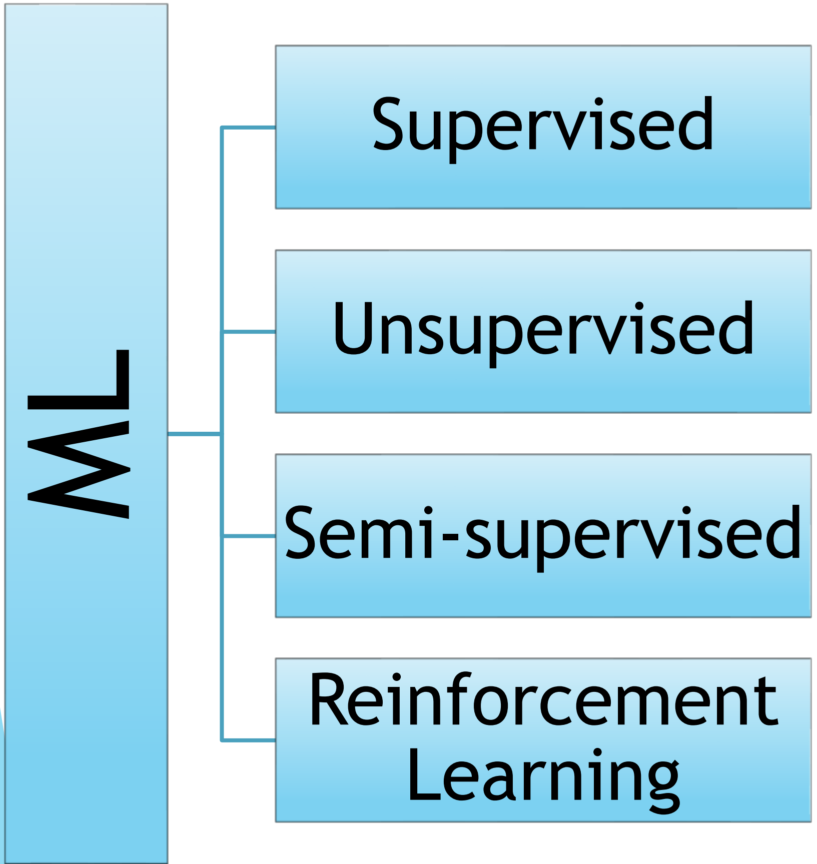
Machine Learning is great for:

- Problems for which existing solutions require a lot of fine-tuning or long lists of rules
- Complex problems for which using a traditional approach yields no good solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data. **(data mining)**



Types of Machine Learning Systems

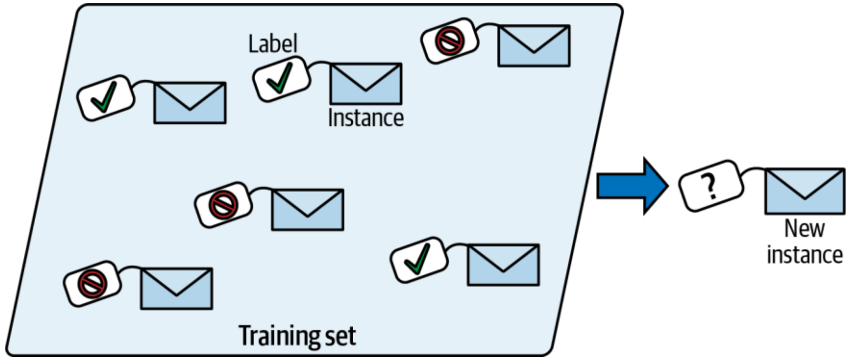
Whether or not they are trained with human supervision



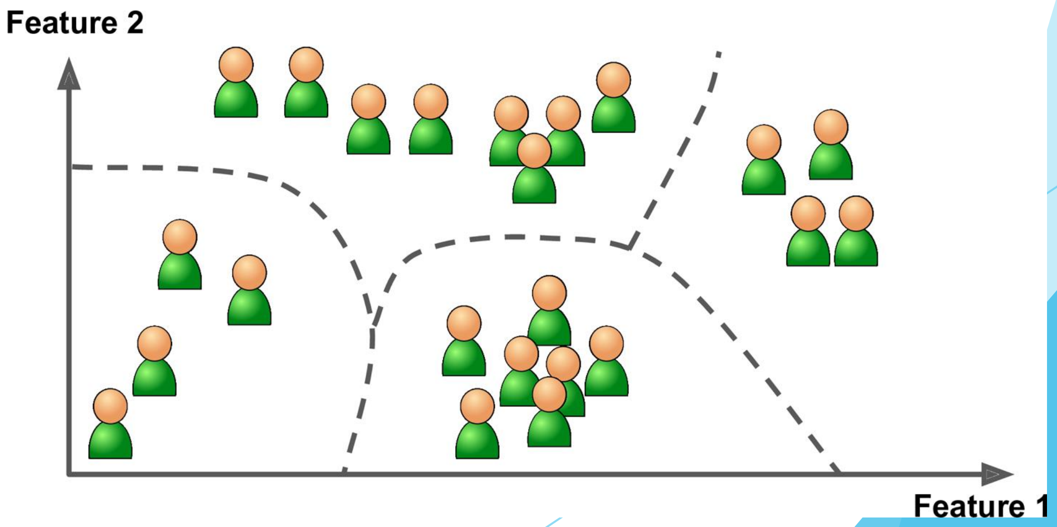
ML

- Supervised
- Unsupervised
- Semi-supervised
- Reinforcement Learning

- In *supervised learning*, the training set you feed to the algorithm includes the desired solutions, called *labels*

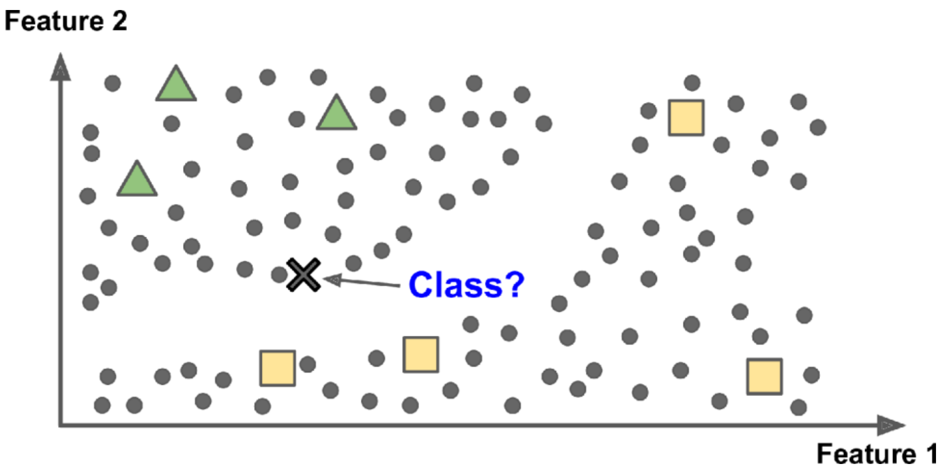


- In *unsupervised learning*, as you might guess, the training data is unlabeled



-cont-

- Labeling data is **time-consuming** and **costly**, you will often have plenty of unlabeled instances, and few labeled instances. Some algorithms can deal with data that's *partially labeled*. This is called *semi-supervised learning*



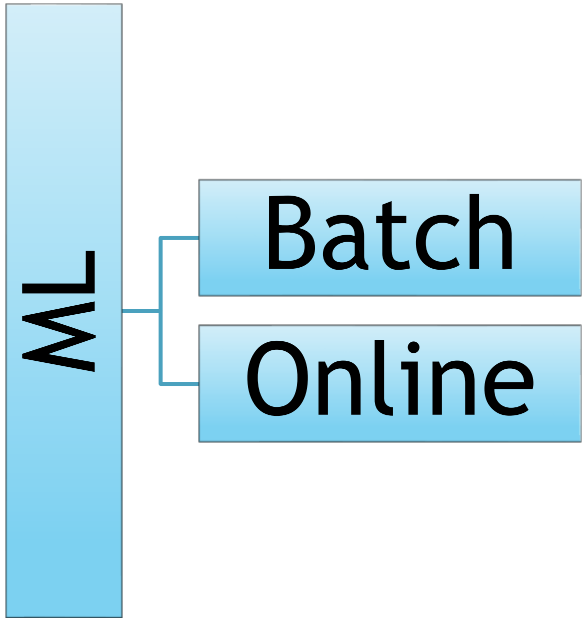
Reinforcement Learning

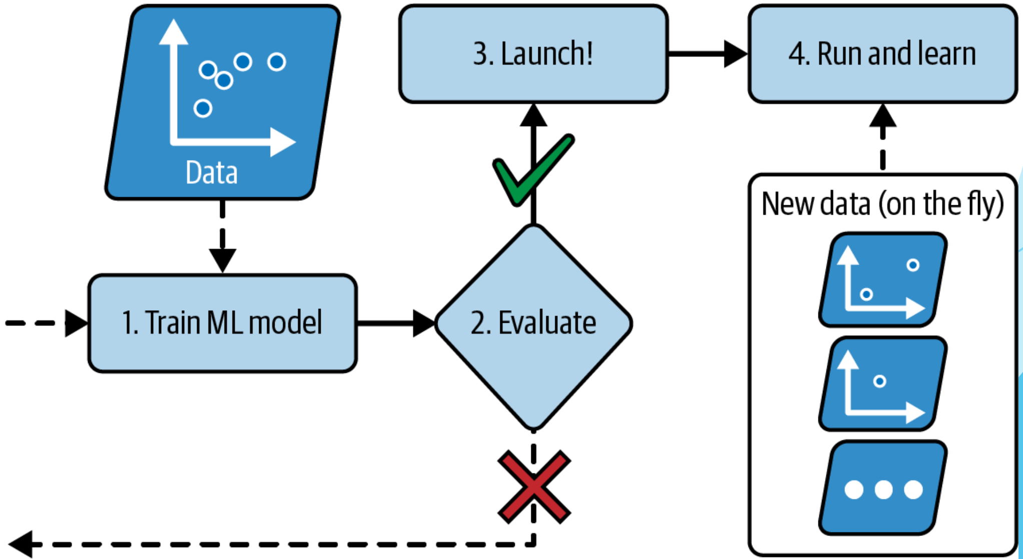
Reinforcement Learning is a very different beast. The *agent* can observe the environment, select and perform actions, and get *rewards* in return. It learns by itself what is the best strategy, called a *policy*, to get the most reward over time.

We cover this in AI course!

Types of machine learning systems

- Batch Learning:
  - It must be trained using all the available data.
  - If new data arrives, you need to train a new version of the system from scratch on the full dataset.
    - But computationally inefficient
- In online learning, you train the system incrementally by feeding it data instances sequentially

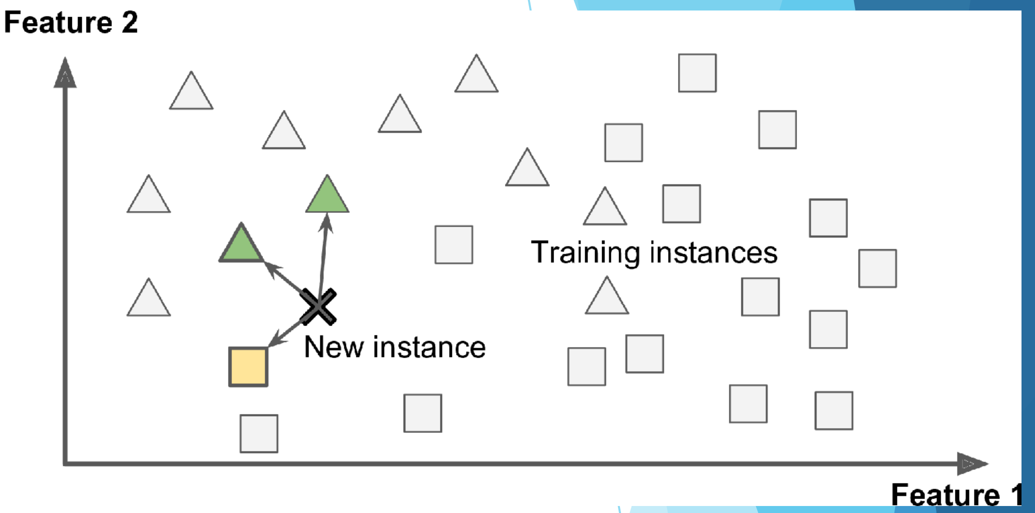


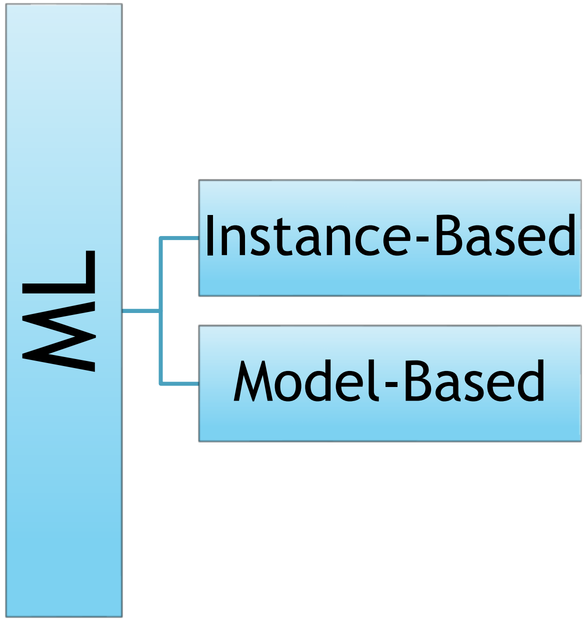


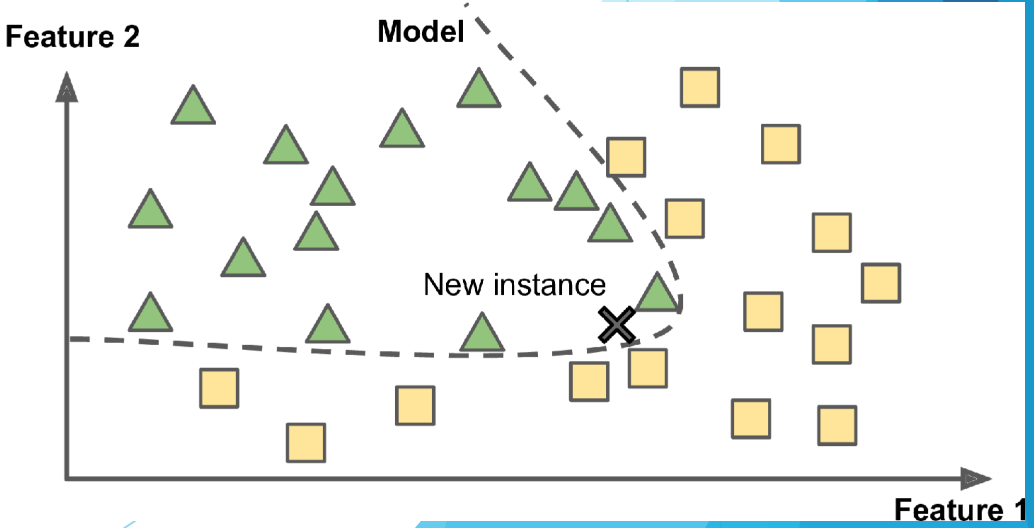
Types of machine learning systems

- *Instance-based learning*: the system learns the examples by heart, then generalizes to new cases by using a similarity measure to compare them to the learned examples

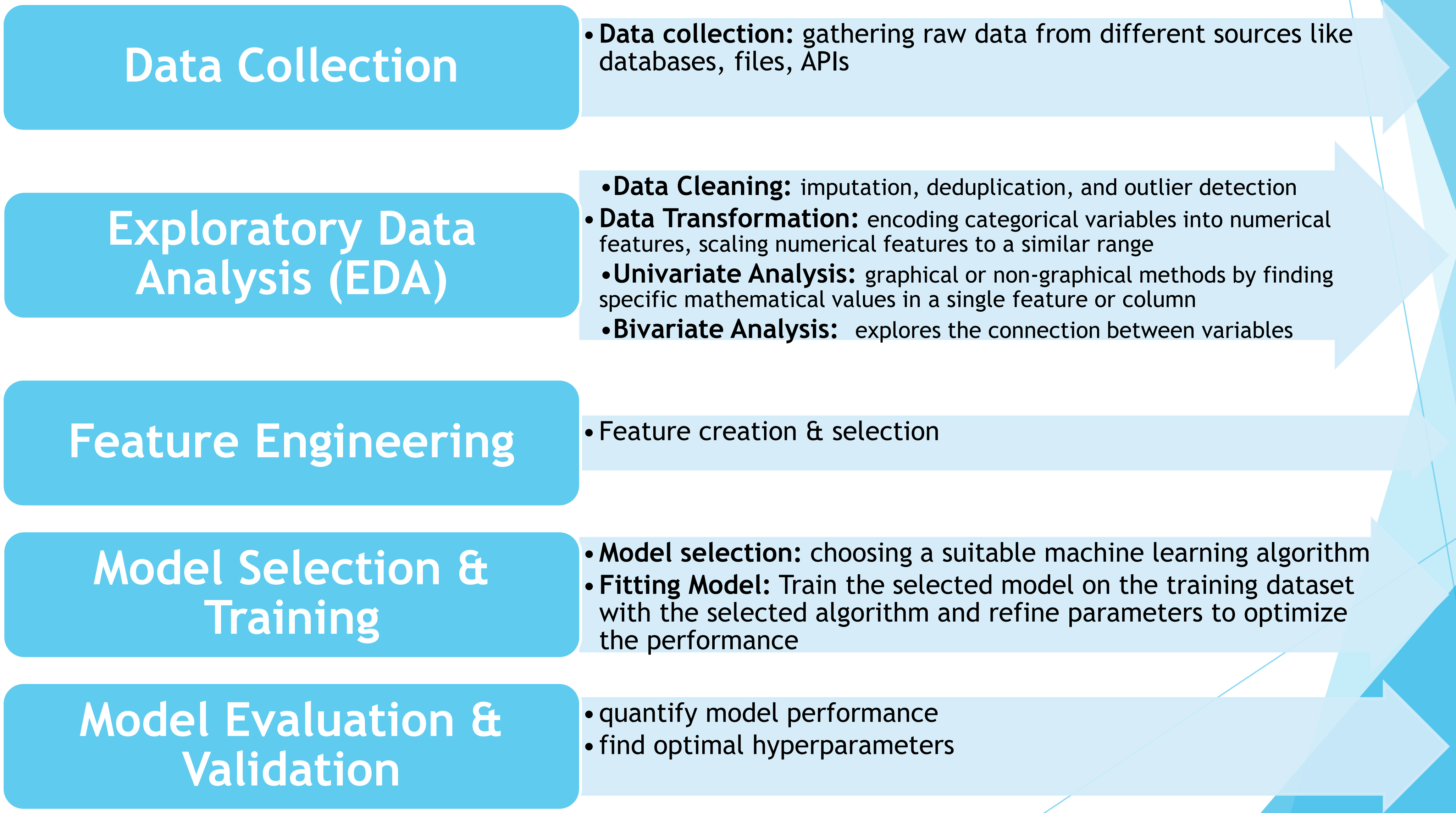
- Another way to generalize from a set of examples is to build a model of these examples and then use that model to make *predictions*







Typical Machine Learning pipeline



Data Collection

- **Data collection:** gathering raw data from different sources like databases, files, APIs

Exploratory Data Analysis (EDA)

- **Data Cleaning:** imputation, deduplication, and outlier detection
- **Data Transformation:** encoding categorical variables into numerical features, scaling numerical features to a similar range
- **Univariate Analysis:** graphical or non-graphical methods by finding specific mathematical values in a single feature or column
- **Bivariate Analysis:** explores the connection between variables

Feature Engineering

- Feature creation & selection

Model Selection & Training

- **Model selection:** choosing a suitable machine learning algorithm
- **Fitting Model:** Train the selected model on the training dataset with the selected algorithm and refine parameters to optimize the performance

Model Evaluation & Validation

- quantify model performance
- find optimal hyperparameters

Main Challenges of Machine Learning

- **“bad data”**
  - Insufficient Quantity of Training Data
  - Nonrepresentative Training Data
    - “Your training data be representative of the new cases you want to generalize to”
    - Famous Example: US presidential election in 1936
  - Poor-Quality Data
    - errors, outliers, and noise
  - Irrelevant Features
    - Garbage in, garbage out
    - Your system will only be capable of learning if the training data contains enough relevant features and not too many irrelevant ones.
- **“bad model”**
  - Overfitting/underfitting (we will dive into this later)