Program Content

Subjects	Tools
Supervised Learning • **Linear Regression** • **Classification (logistic regression, KNN …)** **Data wrangling:** 1. Data Exploration 2. Data Cleaning (selecting, filtering, grouping data …) 3. Dealing with missing values (different techniques of imputation) 4. Selecting the best model **Capstone Project:** • **Analysing real projects**	• **R programming Language** ➢ **R and Rstudio** ➢ **Caret package** ➢ **Tidyverse package (dplyr, tidyr, readr, stringr, purrr, ggplot2, broom)** ➢ **Writing reports with Rmarkdown.**

Goals:

What are my goals?

- **First:** Setting a goal to achieve.
- **Second:** you have to work for it no matter what.
- **Third:** Always be consistent.
- **Fourth:** Put a deadline in your mind.

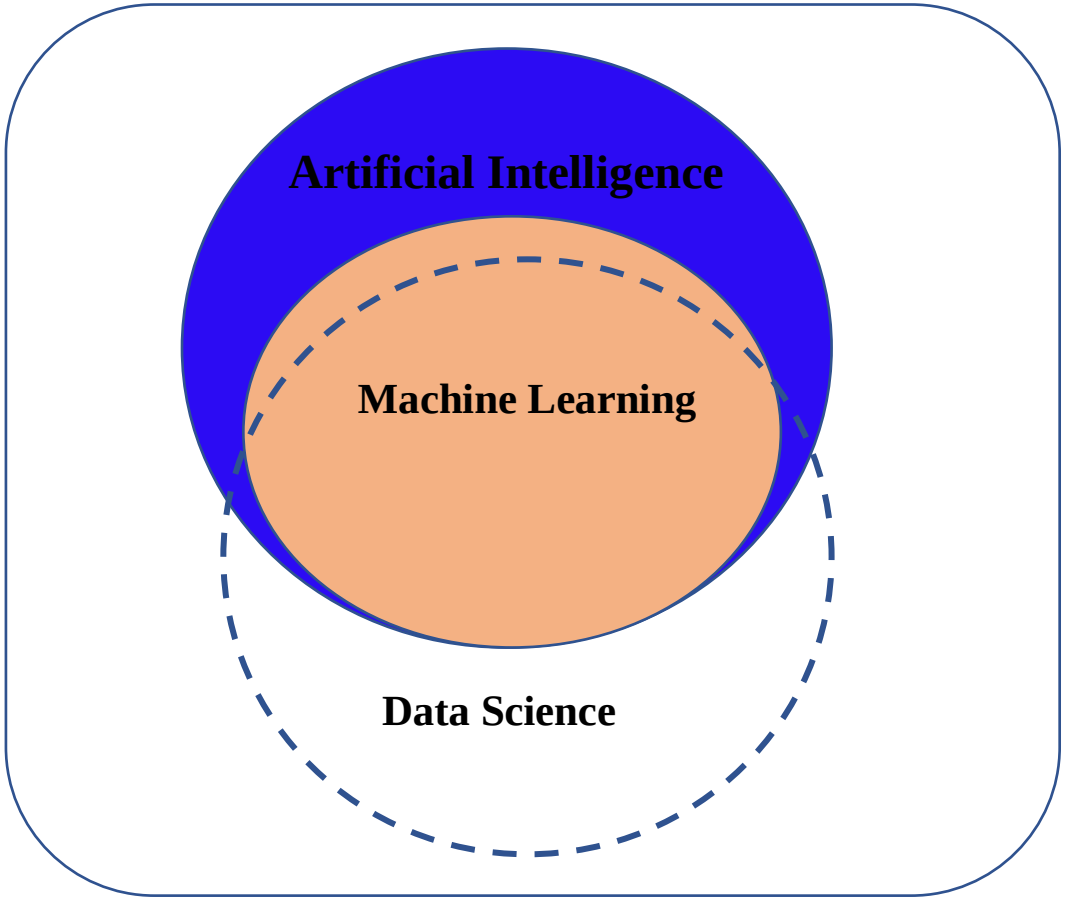
Introduction to Machine Learning

Let us First Consider Some Questions:

1. Have you ever thought about while searching on the internet how the webpages are ranked?
   "**The search engine will return them in the order of relevance. The same happens on YouTube or any other search engines**".
2. If you search on **Facebook** or any other social media, such Instagram or Twitter for something, like videos, articles or images, shortly after the same type will start popping up? How did that happen?
3. Receiving Emails: certain emails are considered spam and some not.
4. Self-Driving Cars, how is that possible?

What is Machine Learning (ML)? & What is Artificial Intelligence (AI)?

A Diagram Showing the Relationship between AI, ML and Data Science



Machine Learning: machine learning is a branch of Artificial Intelligence that automates the building of systems that learn iteratively from data, identify patterns, and predict future results – with minimum human intervention.

Artificial Intelligence: AI is about making computers able to perform the thinking tasks like humans and animal are capable of.

Data Science: is about discovering and communicating insights form data. Machine learning is considered an important tool for data science. This is very obvious in recent years. Especially in making predictions.

Similarities and differences between ML and AI

Artificial Intelligence	Machine Learning ML	Data Science
• AI: is a huge set of tools for making computers behave intelligently. • It comprises of several sub-fields Such as Robotics and Machine Learning. • Making decisions is based on programmable rules derived from theory. • A computer is programmed by step-by-step instructions.	• A set of tools for programming Computers to behave intelligently. • Taking decisions is based on data. • Models Learn from historical data without being programmed to do certain tasks. • Machine Learning is an interdisciplinary mix between statistics and computer science.	Data science: is about discovering and communicating insights form data. Machine learning is considered an important tool for data science.

Requirements to Learn Machine Learning

You can ask yourself, what should I have to learn ML?

1. Statistical Background.
2. Mathematical Background (preferred but not required). (Statistics, Economics, Mathematics, physics …)
3. Domain Expertise.
4. Computer Skills: programming Languages or a Modern software. (**R, Python, SAS …)**

Types of Models

Terminology	Explanatory Models	Predictive Models
Dependent	Dependent variable	The Target The Outcome The Predicted The Response The Class (Categorical)
Independent	Independent Variables Explanatory Variables Regressors	Inputs Features Attributes Predictors Descriptors (less common)

In Explanatory models, interpretability is of importance.

Predictive Models: Accuracy is of importance.

Explanatory Model:

$$\boldsymbol{Dep = Intercept + slope01 * Indep01}$$

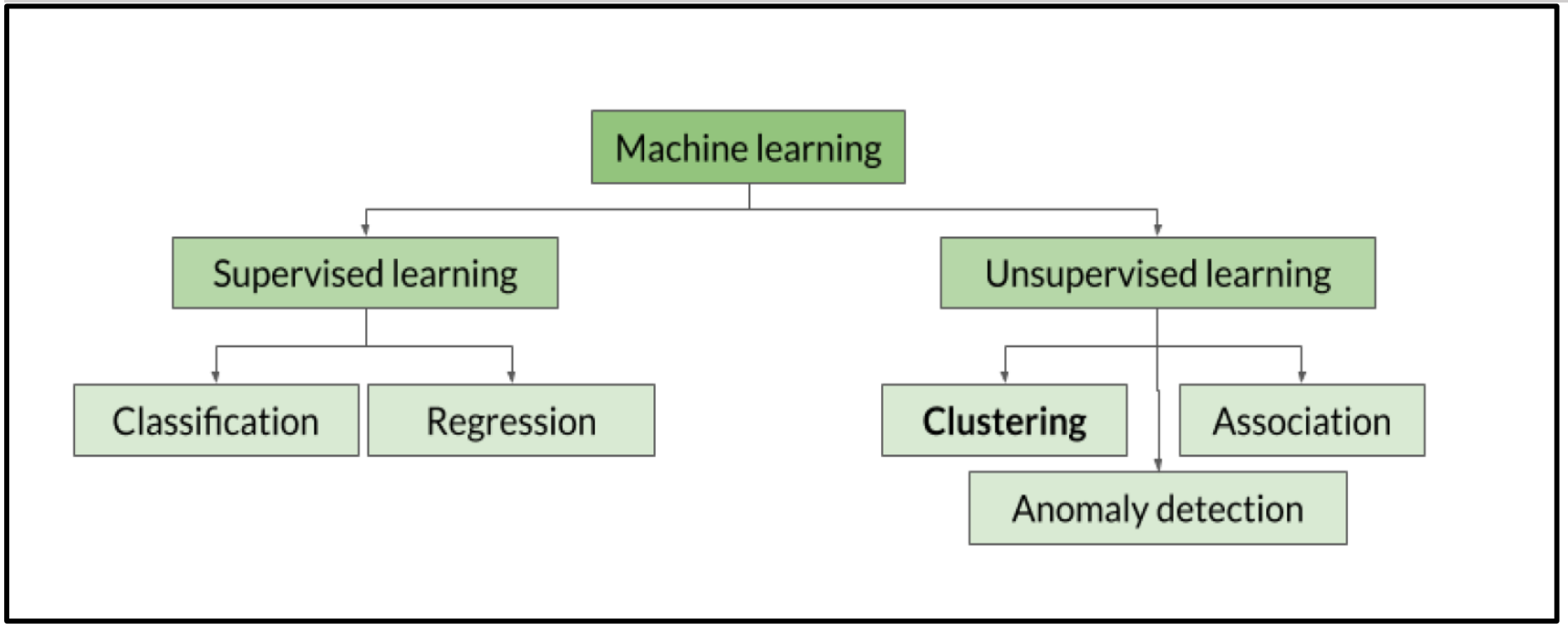
We interpret the slope as the average effect on the Dep variable of a one unit change in The Indep variable. (holding all other variables unchanged)

Example:

$$\boldsymbol{home_{price} = Intercept + slope * square_{feet}}$$

if the surface of home increases by one square feet, the price will increase on average with the value of slope (unit of price 100 dollars as example)

Machine Learning Hierarchy



Types of Machine Learning

- **Supervised Learning:** Means a model with a dependent variable such as Linear regression, Logistic Regression … (this is our only focus in this course)
- **Unsupervised Learning:** a model without a dependent variable such as clustering, k-means …
- **Reinforcement Learning:** This subfield is mostly related to robotics, and it uses complex mathematics like game theory. Example: how a robot moves of plays chess. **(not considered in this course)**

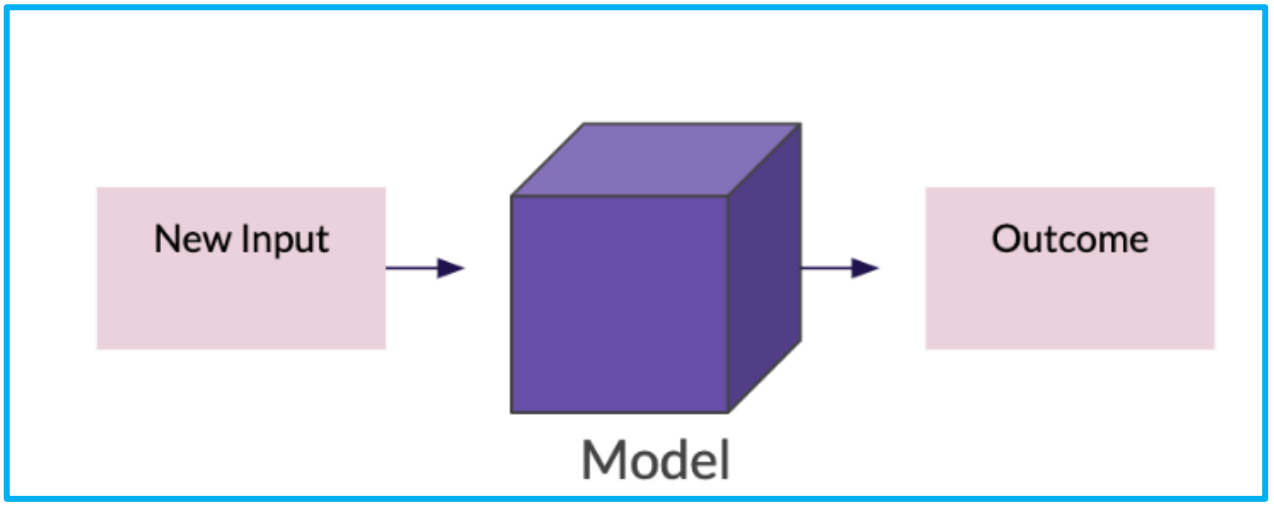
What is the difference between supervise learning and unsupervised Learning?

Supervised Learning	Unsupervised Learning
Target variable: labelled and known **Features:** we have features or inputs **Examples:** Linear Regression Logistic regression, Regression Trees, KNN	**Target:** Unknown (no guidance) **Features:** we have features or inputs **Examples:** Anomaly Detection such as Fraud detection or hacking. Clustering, market segmentation, finding associations

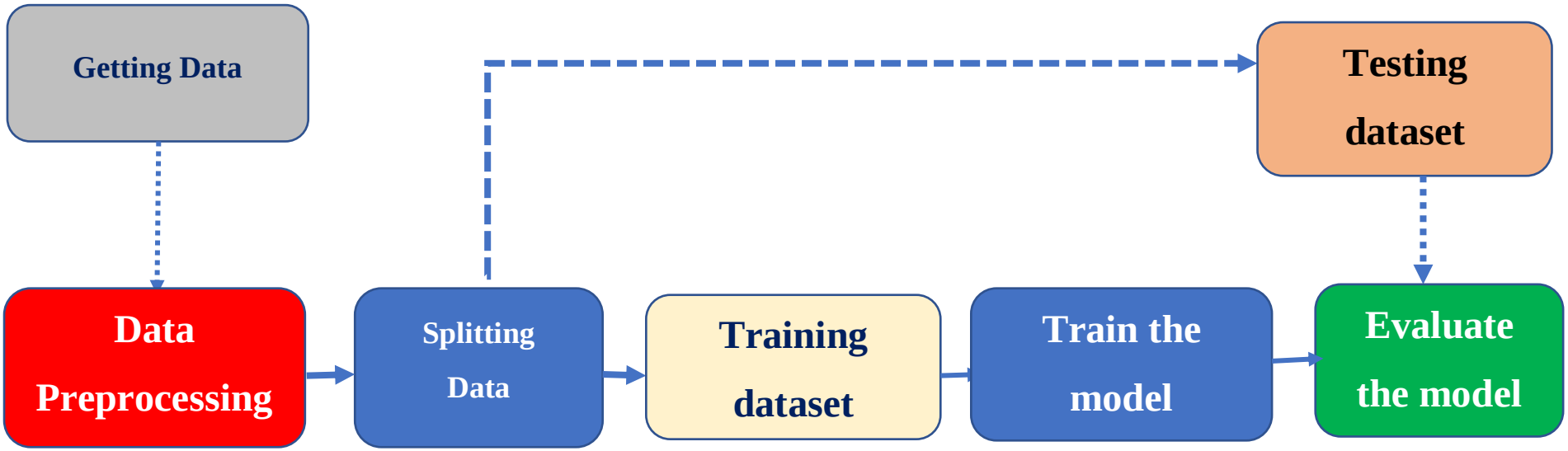
Machine Learning Model

Building a ML model requires data (an existing data) this data is called **Training Data.** This is the data where a model can learn patterns.

When building a model on the existing data, we call this process: **Training a model.**



Machine Learning Workflow



Pre-Processing Data has many Steps: which is called data wrangling or data cleaning: 80% of work will be done here.

- **Renaming Variables**
- **Dealing with Missing Data (imputation)**
- **Transforming Variables (Normalize variables, scaling, centring, log transformation, reciprocal transformation …)**
- **Creating new variables**
- **Collapsing the levels of categorical variables**

- **Feature Selection (Step-wise, forward, backward selection or best method)**

Supervised Learning

Regression	Classification
Target: Continuous **Examples:** • Predicting a stock price in the stock market • Predicting a house price. • Predicting Child's height **Note: A regression problem can be turned into a classification problem. We can do that by cutting (or categorise the target variable)** **Example:** • In the stock market, we can divide the price into three levels: low, medium or high. • In a problem when trying to predict age: the levels can be child, teenager, adult or elder.	**Target:** Categorical **Target:** two categories (binary) **Target:** more than two categories. In classification, a target variable can have more than two categories, example, handwritten digit reader, the target has 10 categories (0, 1, …, 9) **Classification:** means assigning a category to an observation. The purpose is predict a category **Examples: Binary variables** • Churning (will a customer stop buying a company's product or stay loyal) • Will a customer stop subscription • Is a cell cancerous (in medicine) or not • Is this email a spam or not • Is a transaction Fraud or not

	• Will a customer default or no **Multi-categorical variable:** • Classifying animals (cat, dog, fish …) • Classifying flowers

Exercises: Is it Classification or Regression?

1- You have the features (Age, education, job title, area ….), predict Income: ……..

2- Based on the features (Income, Age, credit record, education …), will a customer default or not: ……..

3- Based on (area, house age, surface, ….) predict the price:

4- Based on (airplane condition, pilot experience, weather condition) will the plane fall: …….

An example of Supervised Learning Model: A training data for building a model for predicting whether a patient has a heart disease.

Target Variable

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
50	F	196	0	False	non-anginal pain	98	False
53	F	215	0	True	asymptomatic	110	True
62	M	245	3	False	typical angina	126	True
48	M	190	0	True	non-anginal pain	99	False
70	M	201	0	True	typical angina	105	False

Labels

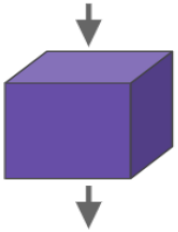
The variables shaded in green as the inputs or features, while **heart disease** is the outcome or the target variable (shaded in orange).

Note **(we may here in such kind of models; especially, in classification labelled models to mean supervised models)**

After training the model. We can give the model a new input to predict to new outcome.

After training (supervised learning)

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
65	F	208	2	False	typical angina	105	???



Heart disease
False

Unsupervised Learning

Unsupervised Learning: Learns from the existing datasets, and tries to detect patterns. This technique is so powerful. You can find insights without knowing much about the datasets.

Application of unsupervised Learning:

- **Clustering:** this consists of identifying groups in the dataset. Grouping the observations is based on the strong similarities among them. Grouping is not certain prior running the algorithm. An algorithm may come with different groups based on the data. Example: if you have eight objects (dogs and cats), after running the Algorithm it may come with two groups based the kind, or with three groups based on the colour, or more groups based on other features. The model will never tell why or how the groups are grouped by. you have to figure that out yourself. Some useful algorithms are **K-means, DBSCAN (Density-Based Spatial Clustering of applications with Noise).**
- **Anomaly Detection (Detecting Outliers):** In data analysis, anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Example Intrusion of Fraud detection.

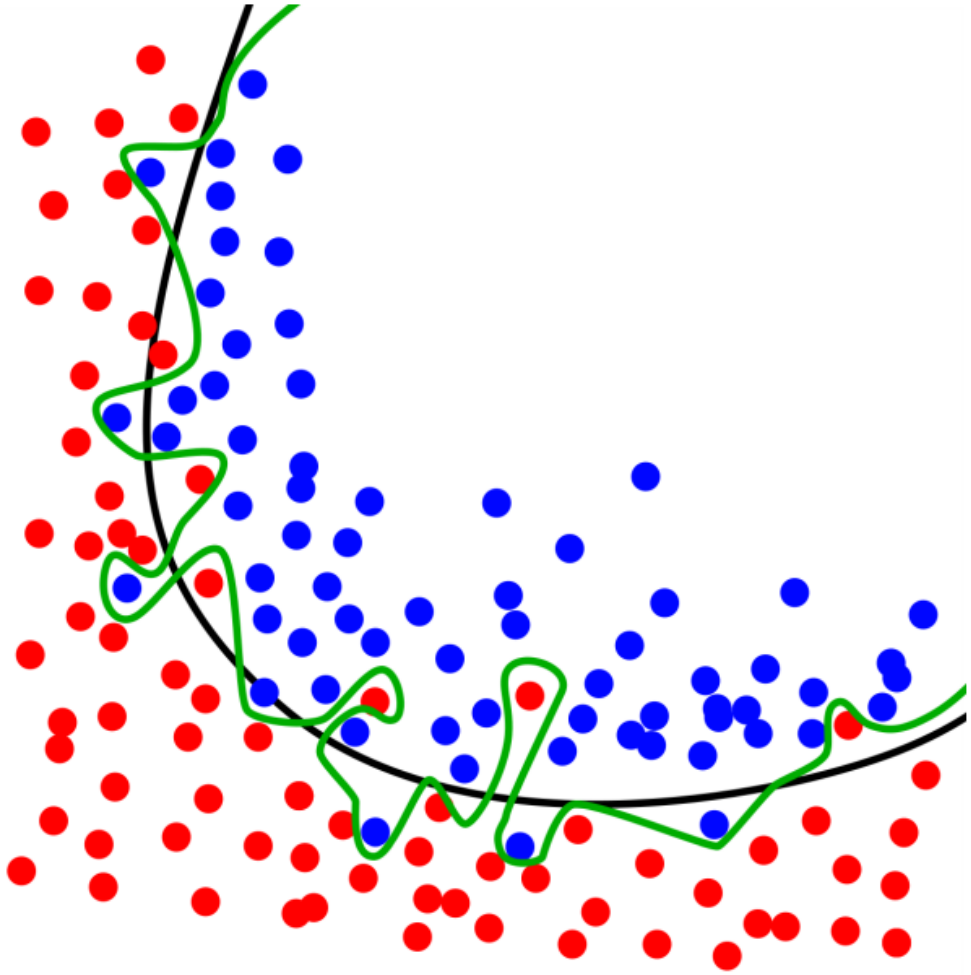
- **Associations:** Finding relationship between observations, in other words finding events that happened together. Example **Market basket analysis** which means which items are bought together**. For example,** people who buy diaper are likely to buy milk, or people who buy cheese is likely to buy bread and so on.

Model Evaluation

The story: Man with a fancy suit, you train a man inside the room who will detect everything correctly, but when you go out and start asking for directions the man will guide you the wrong direction. Or you train him in local city, but you go to a different city the man will know nothing about the directions. It is the same for the model, you want the model to predict as accurately as possible, but of course you don't want to wait for some time to know whether the model predicts the outcome correctly or not. There the term **honest assessment** comes in.

Overfitting: The model performs well on the training data, but poorly on the testing data, we say the model does not generalize well. This is known as **overfitting.**

Illustrating overfitting



Under-fitting: this means the model is very simple and not flexible enough to fit the data, for example if the relationship is quadratic where we fit a simple model.

Computational Tools for learning Machine Learning:

1. **Programming Languages:**
    a. **R programming Language:**
        i. **Caret package: https://topepo.github.io/caret/index.html**
        ii. **Tidymodels package https://github.com/tidymodels**
        iii. **rtemis https://rtemis.netlify.com**
        iv. **mlr3 https://mlr3.mlr-org.com/index.html**
        v. **h2o https://github.com/h2oai/h2o-3/tree/master/h2o-r**
        vi. **Open ML : https://github.com/openml/openml-r**
    b. **Python:**
        1. **Anaconda (https://www.anaconda.com/products/individual)**
        or Enthought (https://assets.enthought.com/downloads/)
            a. **Pandas**
            b. **Numpy**
            c. **Scikit-learn**
            d. **Statsmodels**
            e. **Scipy and many other packages**

f. **Seaborn and matplotlib**

c. **SAS:**

i. **SAS Viya (has many products)**

ii. **SAS Enterprise Miner**

d. **Other programming Languages including Java, C++.**

References:

Getting free datasets: https://data.ca.gov/dataset