### **Machine Learning**

Machine learning can be defined as the process of solving a practical problem by collecting a dataset, and algorithmically training a **statistical model** based on that dataset.

### **Supervised Learning**

The goal of a supervised learning algorithm is to use a dataset to produce a model that takes a feature vector **x** as input and outputs information that allows deducing a label for this feature vector. For instance, a model created using a dataset of patients could take as input a feature vector describing a patient and output a probability that the patient has cancer.

### **Unsupervised Learning**

The goal of an unsupervised learning algorithm is to create a model that takes a feature vector **x** as input and either transforms it into another vector or into a value that can be used to solve a practice problem. **Clustering** is useful for finding groups of similar objects in a large collection of objects, such as images or text documents.

### **Reinforcement Learning**

Reinforcement learning is a subfield of machine learning where the machine also called an agent lives in an environment and is capable of perceiving the state of that environment as a vector of features. A common goal of reinforcement learning is to learn a function that takes the feature vector of a state as input and outputs an optimal action to execute in that state. The action is optimal if it maximizes the expected average long-term reward.

### **Raw and Tidy Data**

Raw data is a collection of entities in their natural form: they cannot always be directly fed to machine learning algorithms. Tidy data can be seen as a spreadsheet, in which each row represents one example, and columns represent various attributes of an example.

### **Training and Holdout Sets**

The first step in a machine learning project is to shuffle the examples and split them into three distinct sets: training, validation, and test. The learning algorithm uses the training set to produce the model. The validation set is used to choose the learning algorithm, and find the best configuration values for that learning algorithm, also known as hyperparameters. The test set is used to assess the model performance before delivering it to the client or putting the model into production.

### **Parameters vs. Hyperparameters**

Hyperparameters are inputs of machine learning algorithms or pipelines that influence the performance of the model. They don’t belong to the training data and cannot be learned from it.

Parameters are variables that define the model trained by the learning algorithm. Parameters are directly modified by the learning algorithm based on the training data to find the optimal values of the parameters.

### **When to use Machine Learning**

1. When the problem is too complex for coding.
2. When the problem is constantly changing.
3. When it is a perceptive problem such as speech, image, and video recognition.
4. When it is an unstudied phenomenon.
5. When the problem has a simple objective.

### **Machine Learning Engineering**

Machine learning engineering is the use of scientific principles, tools and techniques of machine learning and traditional software engineering to design and build complex computing systems. MLE ecompasses all stages from data collection, to model training, to making the model available for use by the customers. MLE includes any activity that lets machine learning algorithms be implemented as a part of an effective production system.

A machine learning project life cycle consists of the following steps:

1. Goal definition.
2. Data collection and preparation.
3. Feature engineering.
4. Model training.
5. Model evaluation.
6. Model deployment.
7. Model serving.
8. Model monitoring.
9. Model maintenance.

### **Bias**

Bias in data is an inconsistency with the phenomenon that data represents. **Selection bias** is the tendency to skew your choice of data sources to those that are easily available, convenient, and cost-effective. **Omitted variable bias** happens when your featurized data doesn’t have a feature necessary for accurate prediction. **Sampling bias** also known as distribution shift occurs when the distribution of examples used for training doesn’t reflect the distribution of the inputs the model will receive in production.

It is usually impossible to know exactly what biases are present in a dataset. Biases can be avoided by questioning everything: who created the data, what were their motivations and quality criteria, and more importantly, how and why the data was created.