# Supervised and unsupervised learning

One of the main ideas behind that is to find patterns in data and make predictions on that data without the need to develop each and every use-case from scratch. Therefore, a certain number of algorithms are available. These algorithms can be “classified” by how they work. the main two principles (which then can also be spilt) are:

* Supervised Learning
* Unsupervised Learning
* Semi-supervised Learning

## Supervised learning

With supervised learning, the algorithm learns basically by existing data and learning “from the past”. This means that there is basically a lot of learning data that allows the algorithm to find the patterns by learning from this data. We can also call this “a teacher”. It works closely to how we as humans learn: we get information from our parents, teachers and friends and combine this to make future predictions. Examples are:

* Manufacturing: if several properties of a material were of specific properties, the quality was either good or bad (or maybe scaled from several numbers). Now, if we produce a new material and we look at the properties of the material, based on the existing data we have from former productions, we can say how the quality will be. Properties of a material might be: hardness, color, …
* Banking: based on several properties of a potential borrower, we can predict if the person is capable of paying back the loan. This can be based on existing data of former customers and what the bank “learned” from them. The algorithm takes a lot of different variables into consideration. This variables can be income, montly liability to pay, education, job, etc.

## Unsupervised learning

With unsupervised learning we have no “teacher” available. The algorithms get data, and the algorithms try to find patterns in that. This can either be by clustering data (e.g. customer with high income, customer with low income, …) and make predictions based on that. An unsupervised learning algorithm can be useful for several use-cases. Below are some samples:

* Manufacturing: find anomalies in the production lines (e.g. the average output of units per hour was between 200 and 250, but on day D at time T, the output was only 20 units. The algorithm can cluster data into normal output and an anomaly that was detected.
* Banking: normally, the customer would only spend money in his home country. The algorithm detects an abnormal behavior, like money transferred in a country that he normally isn’t in. This can be an indicator for fraud.

## Semi supervised learning

Last, but not least, there is **Semi supervised learning**, which is a combination of both. In many machine learning projects, not all training data that is used for supervised learning is available, so values might need to get predicted. This can be done by combining supervised and unsupervised learning algorithms and then work with the “curated” data on it.

# Features and Label in Machine Learning

**Features** are known values, which are often used to calculate results. This are the variables that have an impact on a prediction. If we talk about manufacturing, we might want to reduce junk in our production line. Known features from a machine could then be: Temperature, Humidity, Operator, Time since last service. Based on these Features, we can later calculate the quality of the machine output

**Labels** are the values we want to build the prediction on. In training data, labels are mostly known, but for the prediction they are not known. When we focus on the machine data example from above, a label would be the quality. So all of the features together make up for a good or bad quality and algorithms can now calculate the quality based on that.

Let’s now go on another “classification” of machine learning techniques. We “cluster” them by supervised/unsupervised.

# Clustering, Classification, Regression

## Clustering

The first one is **clustering**. Clustering is an unsupervised technique. With clustering, the algorithm tries to find a pattern in data sets without labels associated with it. This could be a clustering of buying behavior of customers. Features for this would be the household income, age, … and clusters of different consumers could then be built.

## Classification

The next one is **classification**. In contrast to clustering, classification is a supervised technique. Classification algorithms look at existing data and predicts what a new data belongs to. Classification is used for spam for years now and these algorithms are more or less mature in classifying something as spam or not. With machine data, it could be used to predict a material quality by several known parameters (e.g. humidity, strength, color, …). The output of the material prediction would then be the quality type (either “good” or “bad” or a number in a defined space like 1-10). Another well-known sample is if someone would survive the titanic – classification is done by “true” or “false” and input parameters are “age”, “sex”, “class”. If you would be 55, male and in 3rd class, chances are low, but if you are 12, female and in first class, chances are rather high.

## Regression

The last technique for this post is **regression.**Regression is often confused with clustering, but it is still different from it. With a regression, no classified labels (such as good or bad, spam or not spam, …) are predicted. Instead, regression outputs continuous, often unbound, numbers. This makes it useful for financial predictions and alike. A common known sample is the prediction of housing prices, where several values (FEATURES!) are known, such as distance to specific landmarks, plot size, … The algorithms could then predict a price for your house and the amount you can sell it for.

# K-means clustering

