# Supervised - Introduction

## Definitions

**dataset**: data is provided as pairs of inputs and desired outputs.
**model**: the algorithm finds a way to produce the desired output given an input.
**generalization**: the algorithm generate autonomously an output for an input it has never seen before.

## Main steps in a ML analysis

- Understand the problem we are trying to solve and if the data can solve the problem
- Formalize the problem
- Collect enough data to solve the problem
- Identify features and algorithms which allow right predictions
- Define metrics for the performances measurement
- Generate the predictive model and integrate the ML solution within a business product

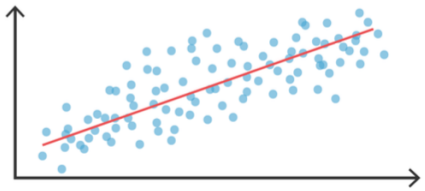
## Regression

If feature has continuity in the output (numerical), it's regression.

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + ... + w[i] * x[i] + b$$

One-feature regression example:

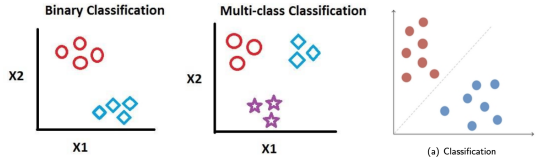
$$\hat{y} = w * x + b$$



(b) Regression

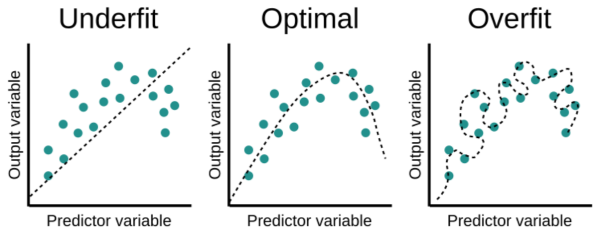
## Classification

Classification predicts a category/label (Non numerical encoded).
**Classification algorithms**: Linear Classifiers, Support Vector Machines, Decision Trees, K-Nearest Neighbor, Random Forest, Neural Networks, Naive Bayes, ...




## Overfitting and Underfitting

We expect simple models to generalize better to new data. Therefore, we always want to find the simplest model. Building a model that is too complex for the amount of information we have, is called overfitting.




Usually, collecting **more data points** will yield **more variety**, so larger datasets allow building **more complex** models.