Introduction to Machine Learning

Course Overview

Mohamad GHASSANY

École supérieure d'ingénieurs Léonard-de-Vinci

09/01/2018

Course Introduction

Machine Learning

You probably use it dozens of times a day without even knowing it.

Application examples :

- Effective web search.
- Social networks recognize friends from photos or suggest friends.
- Email spam detection.
- Handwriting recognition.
- Understanding the human genome.
- Medical diagnostics.
- Predict possibility for a certain disease on basis of clinical measures.
- Fraud detection.
- Drive vehicles.
- Recommendations (eg, Amazon, Netflix).
- Natural language processing.

The aim of ML is to build computer systems that can adapt to their environments and learn form experience.

Machine Learning

What is Machine Learning ?

- A science of getting computers to learn without being explicitly programmed [1].
- Study of algorithms that improve their performance P at some task T with experience E [2].

T : recognition of a handwritten letter “a” from its image.

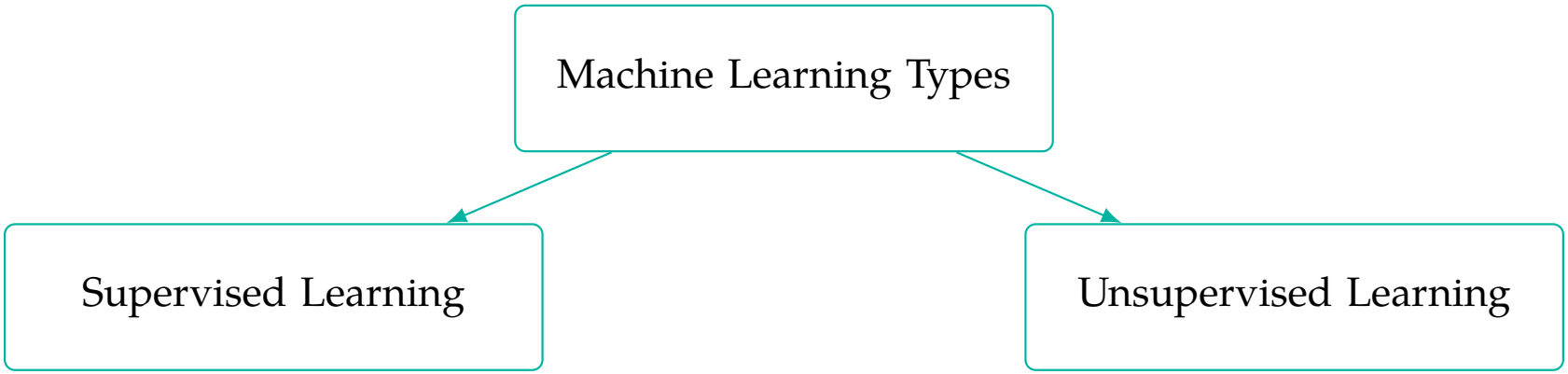
E : images of a handwritten “a”.

P : recognition rate.

1. Arthur Samuel.
2. Tom Mitchell.

Types of Machine Learning Problems

In general, any machine learning problem can be assigned to one of two broad types :

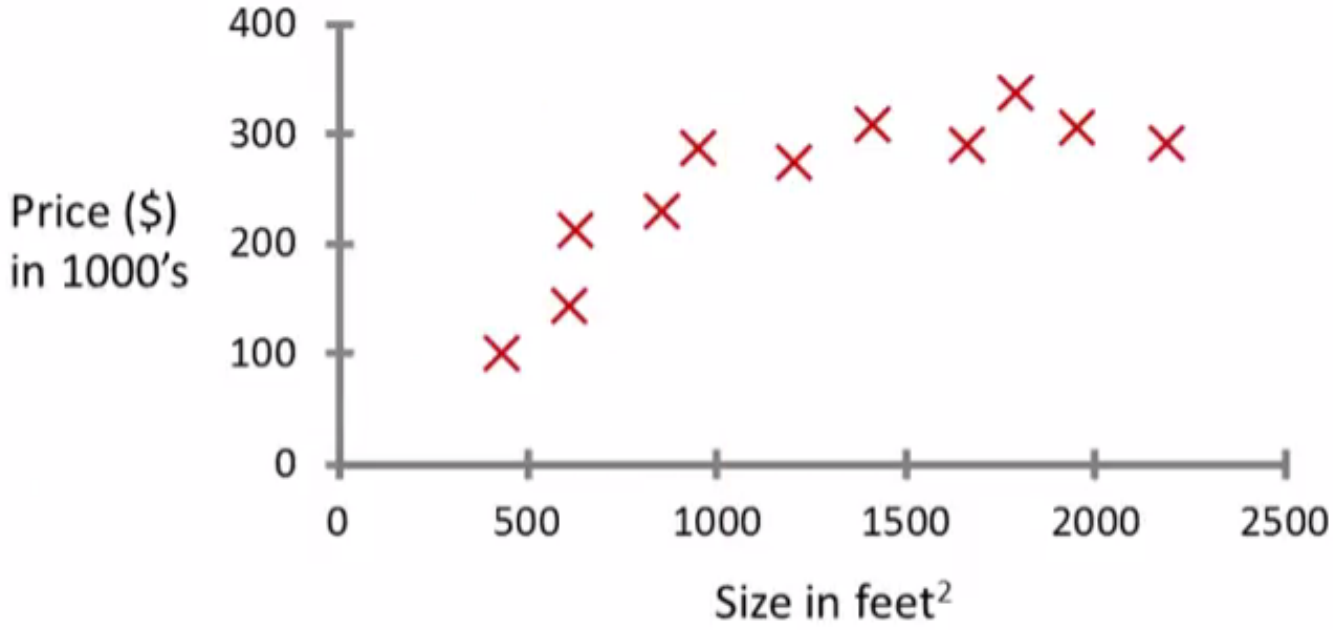


Other : Semi-supervised Learning, Reinforcement learning, Recommender system, etc...

Supervised Learning

Example : House price prediction [3]

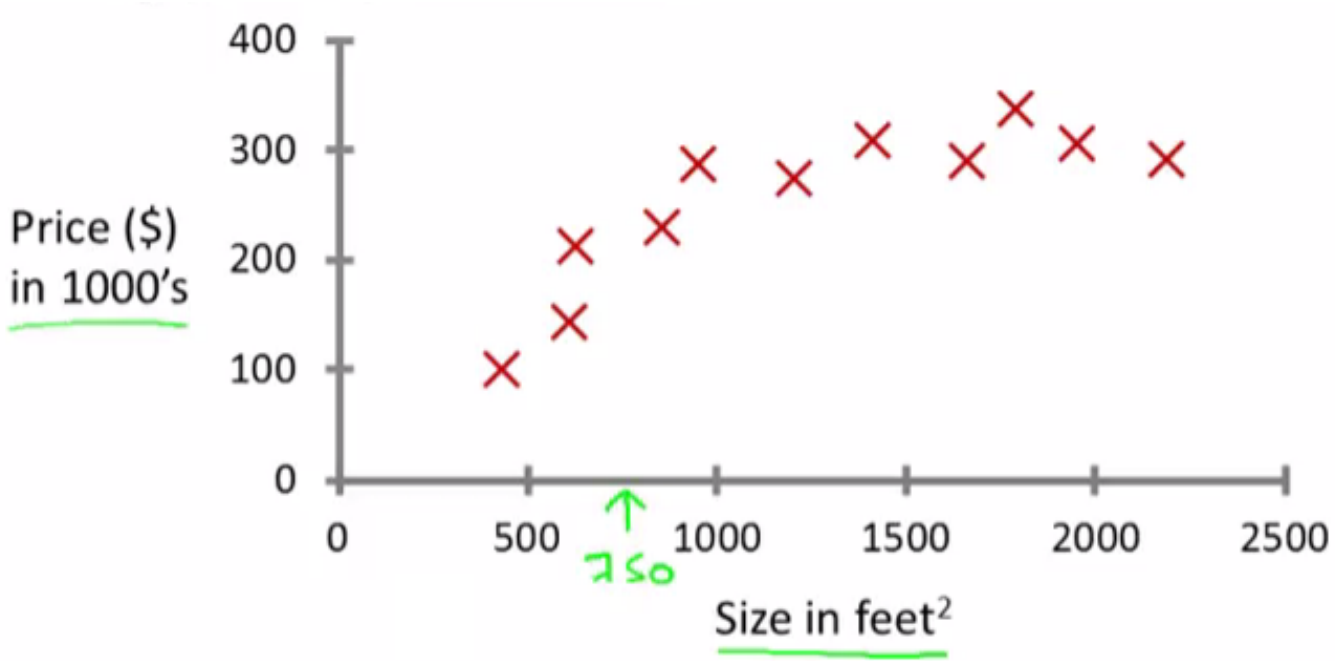
Let's say we want to predict housing prices. We plot a data set and it looks like this :



3. Example from Andrew Ng's MOOC.

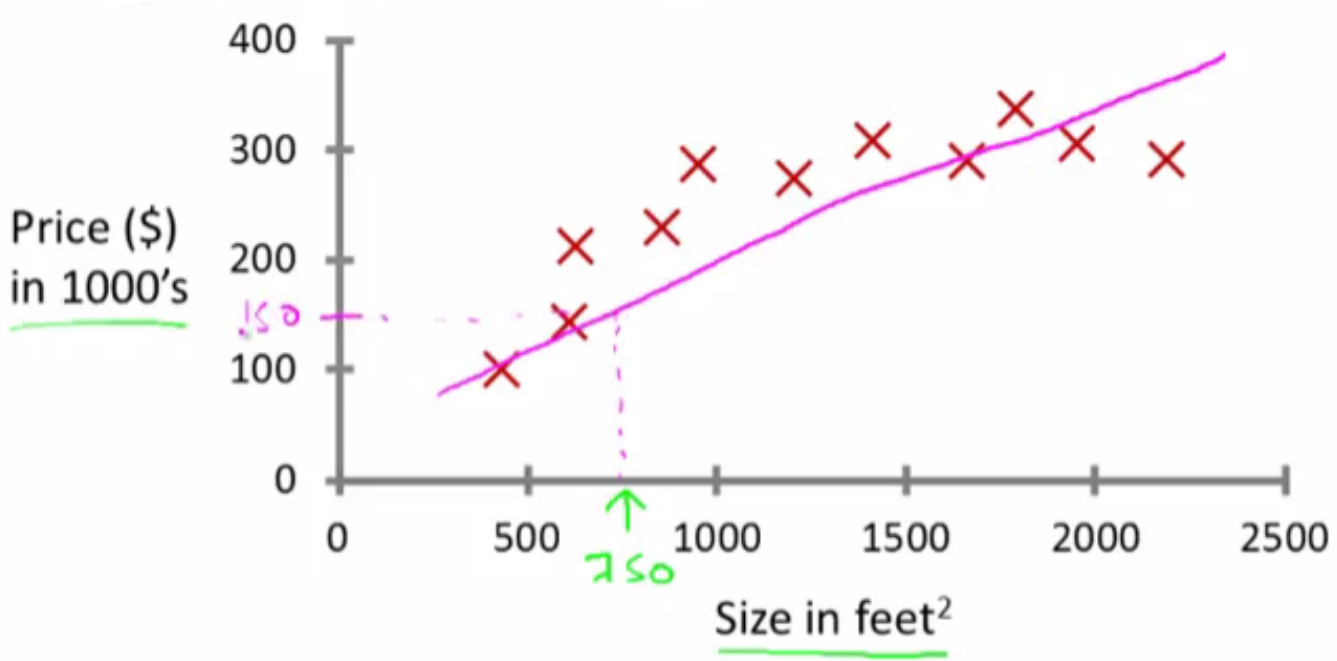
Example : House price prediction

Let's say we own a house that is, say 750 square feet and hoping to sell the house and we want to know how much we can get for the house.



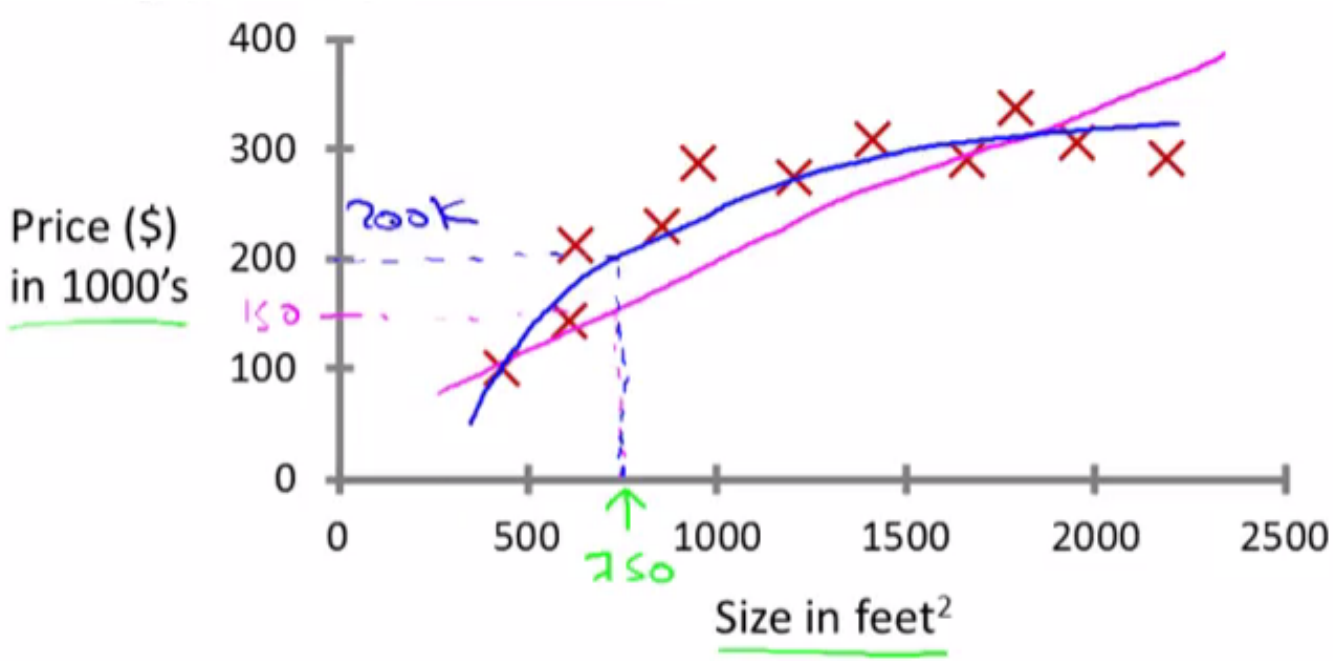
Example : House price prediction

A learning algorithm can for example **"fit"** a straight line to the data and, based on that, it looks like maybe the house can be sold for maybe about 150 000$.



Example : House price prediction

There might be a better learning algorithm ! Maybe a ***quadratic function*** to this data.

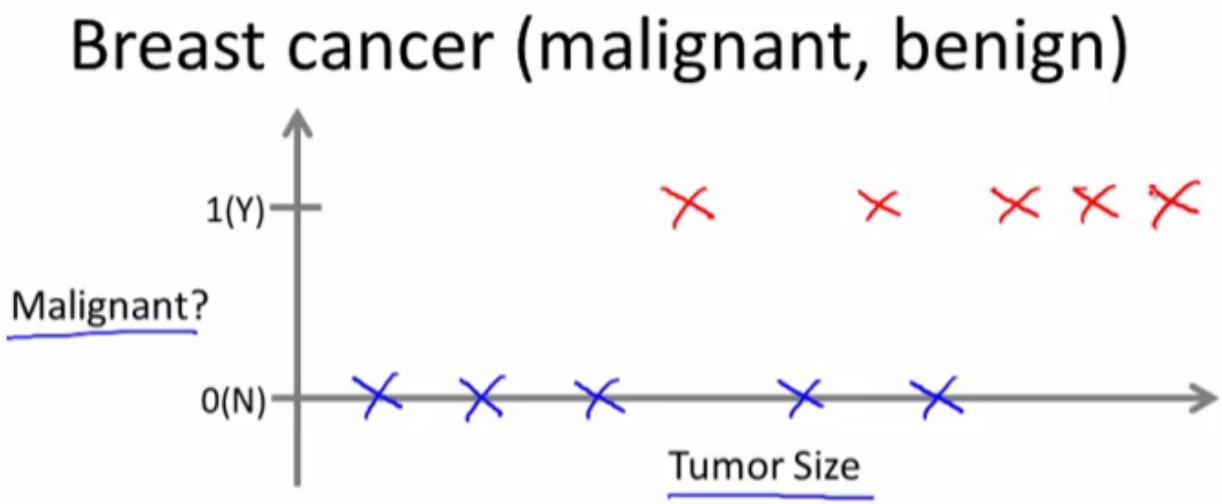


If we do that, and make a prediction here, then it looks like maybe we can sell the house for closer to 200 000$.

In this example, there is target variable "Price". It is a continuous variable.

Example : Medical diagnosis

Let's say we want to look at medical records and try to predict of a breast cancer as malignant or benign [4].
A collected data set gave the following :

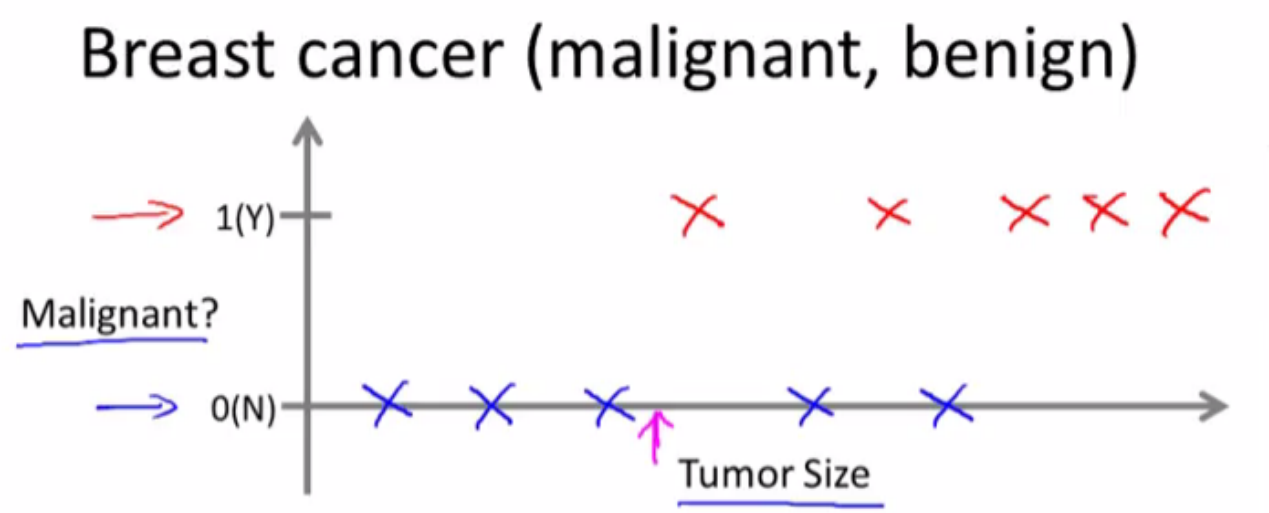


- In this example, there is target variable "malignant or benign". It is a discrete variable.

4. A malignant tumor is a tumor that is harmful and dangerous and a benign tumor is a tumor that is harmless.

Example : Medical diagnosis

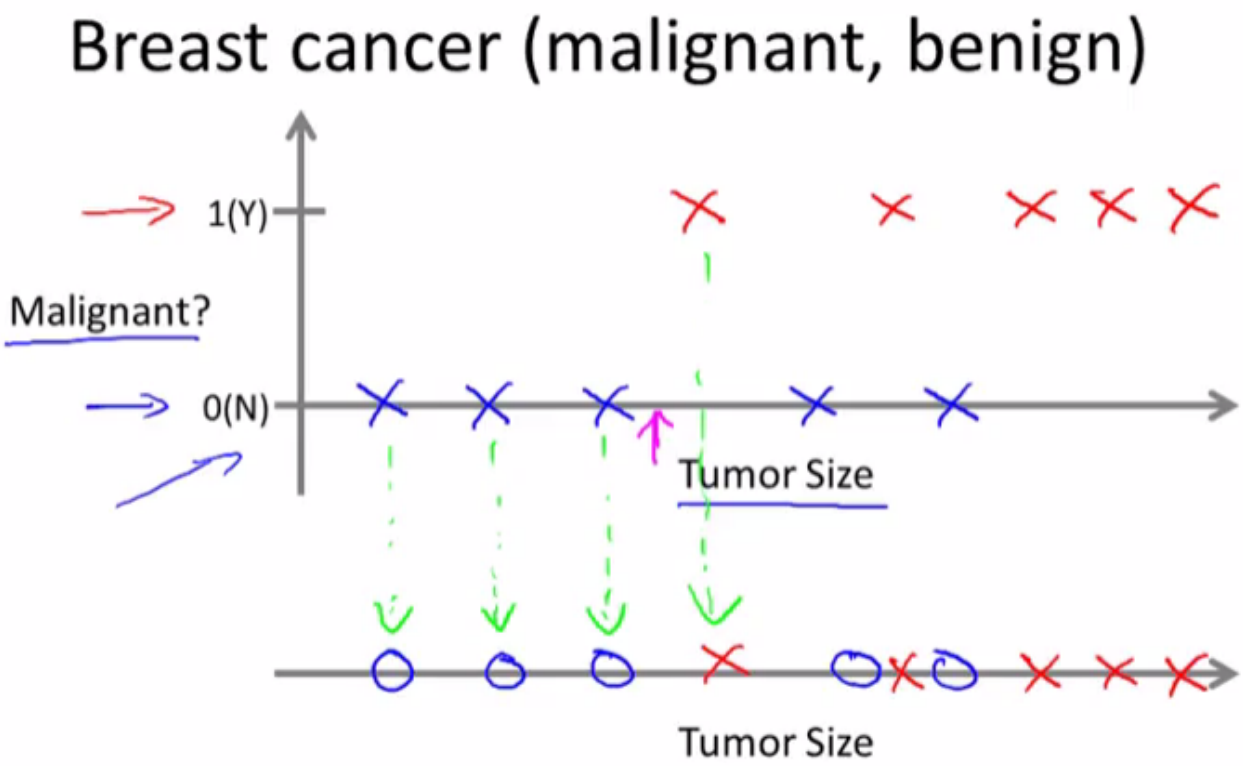
Let's say a person has a breast tumor, and her breast tumor size is known.



- The machine learning question here is, can you estimate what is the **probability** that a tumor is malignant versus benign ?

Example : Medical diagnosis

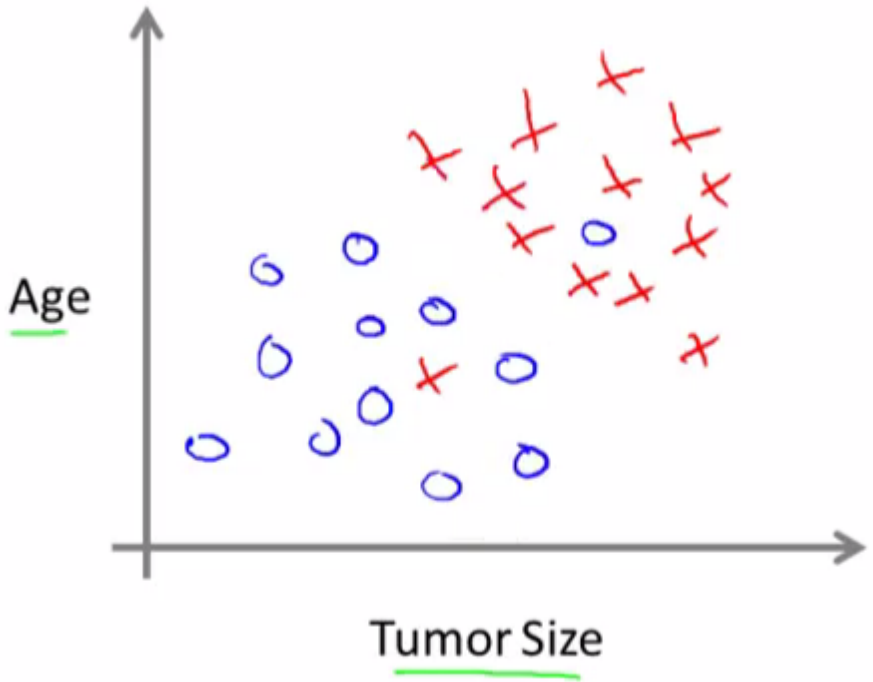
We can take the data set on top and map it down using different symbols.



In this example, there is only one input.

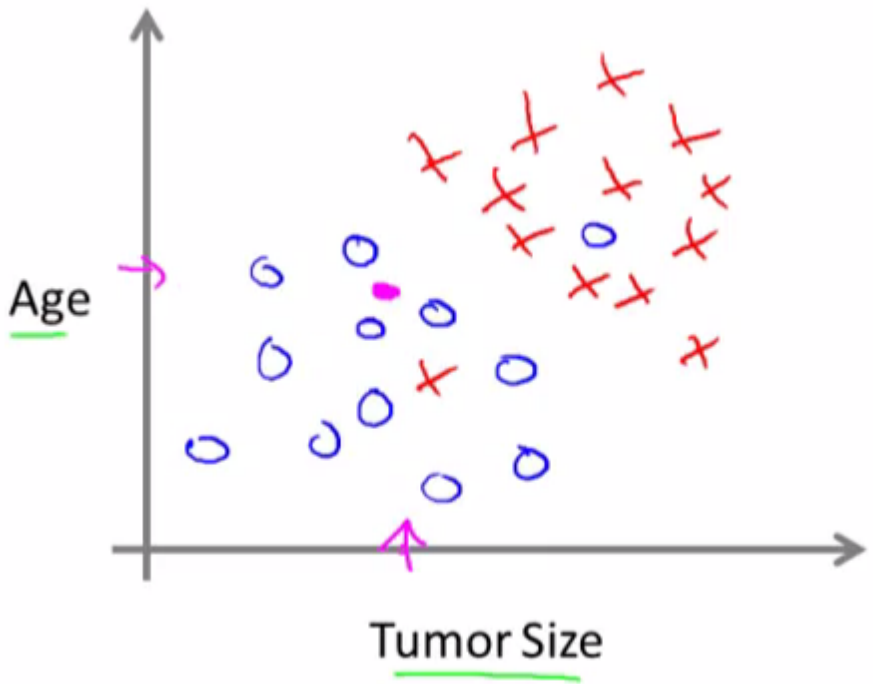
Example : Medical diagnosis

Let's say that we know both the age of the patients and the tumor size. In that case maybe the data set will look like this.



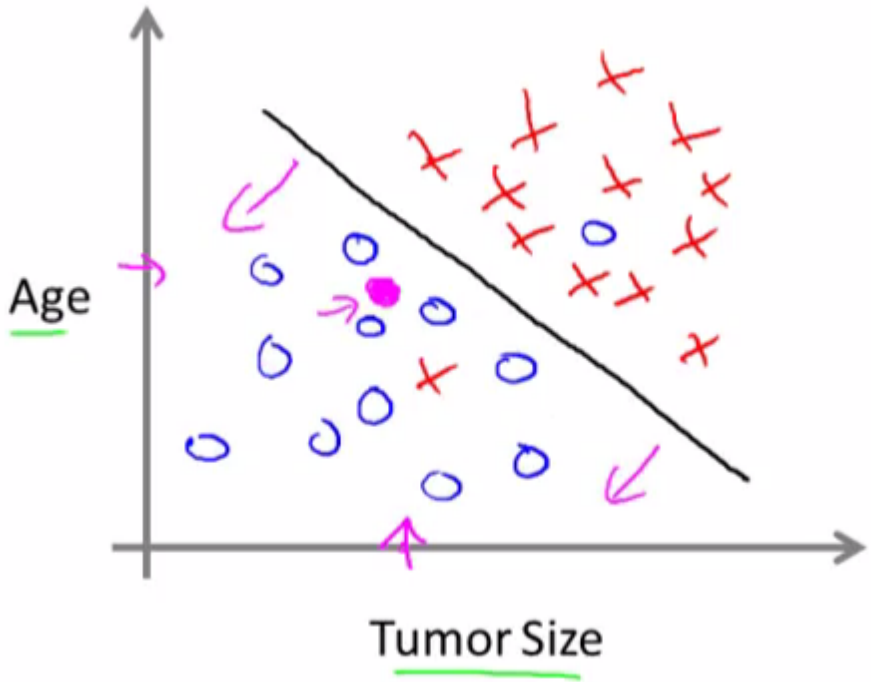
Example : Medical diagnosis

So, let's say a person who tragically has a tumor. And maybe, their tumor size and age falls around there (rose point) :



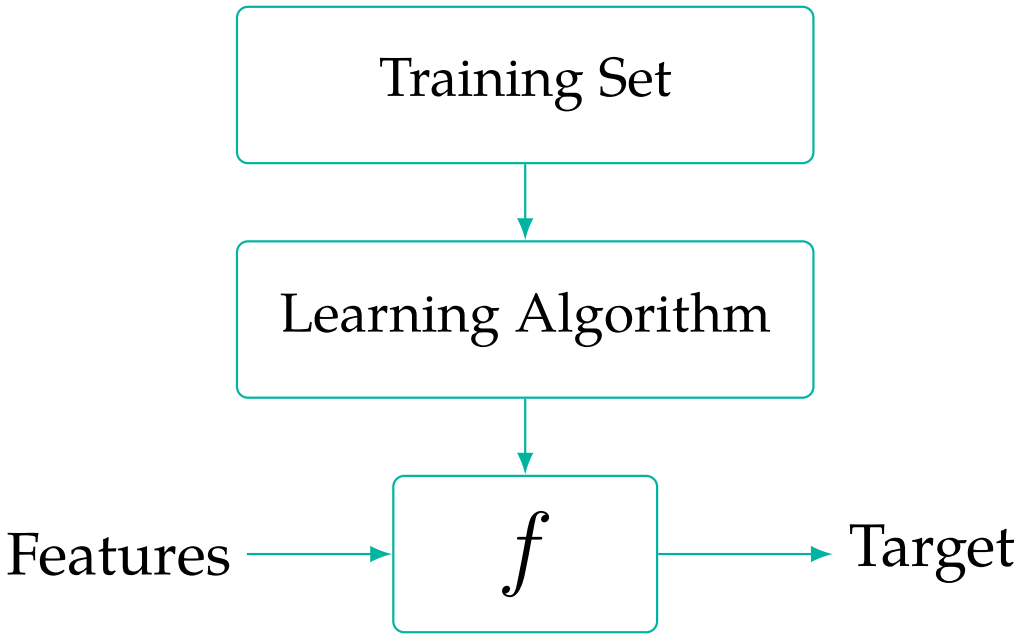
Example : Medical diagnosis

Given a data set like this, a learning algorithm may throw a straight line through the data to try to separate out the malignant tumors from the benign ones.



Supervised Learning : Definition & Model

The term supervised learning refers to the fact that we gave the algorithm a data set in which the **"right answers"** (known as **labels**) were given.



- Supervised Learning refers to a set of approaches for estimating f.
- f is also called ***hypothesis*** in Machine Learning.

Regression and Classification

Regression

- The example of the house price prediction is also called a **regression** problem.
- A regression problem is when we try to predict a **quantitative (continuous)** value output. Namely the price in the example.

Classification

- The process for predicting **qualitative (categorical, discrete)** responses is known as classification.
- Methods : Logistic regression, Support Vector Machines, etc..

Supervised Learning : Notations

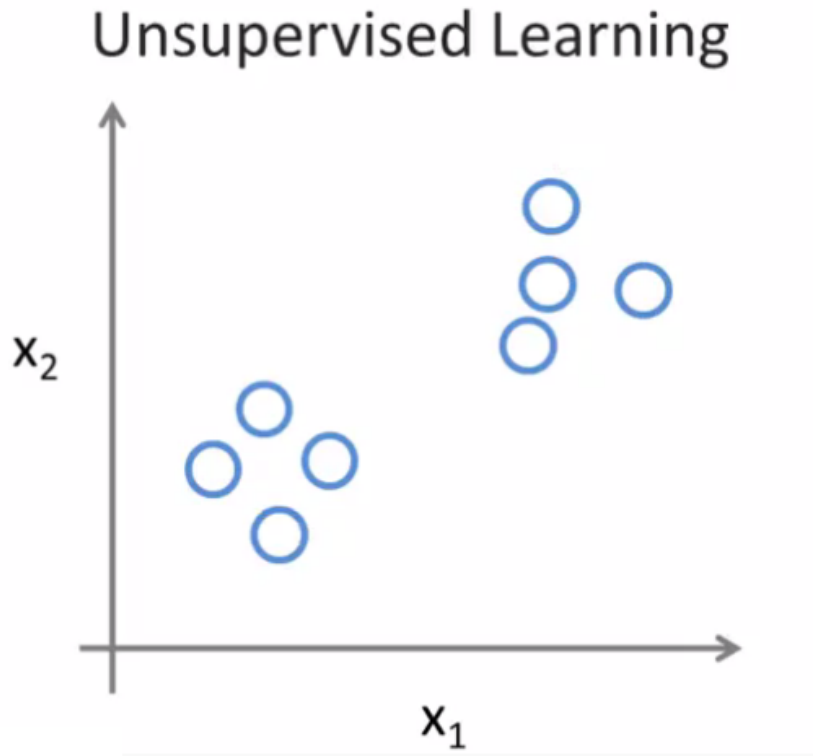
Notations :

- The size of the house in the first example, tumor size and age in the second example, are the **input** variables. Typically denoted by X.
- The inputs go by different names, such as *predictors*, *independent variables*, *features*, *predictor* or sometimes just *variables*.
- The house price in the first example and the diagnosis in the second example are the **output** variables, and are typically denoted using the symbol Y.
- The output variable is often called the *response*, *dependent variable* or *target*.

Unsupervised Learning

Unsupervised Learning : "No labels"

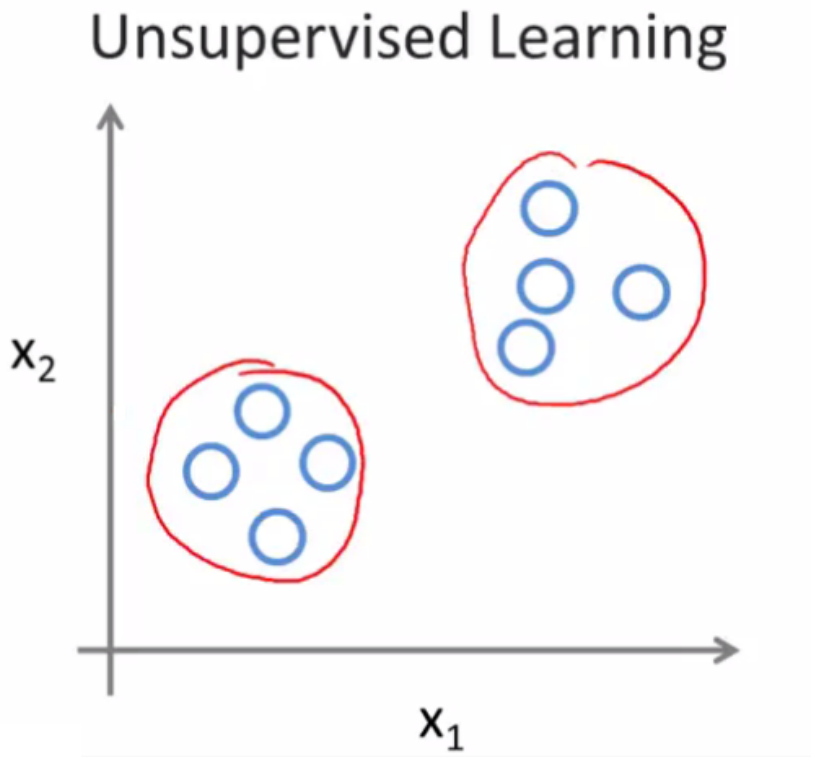
In Unsupervised Learning, we're given data that doesn't have any **labels**.
For example :



Question : Can you find some structure in the data ?

Unsupervised Learning : Structure

Given this data set, an Unsupervised Learning algorithm might decide that the data lives in two different clusters.

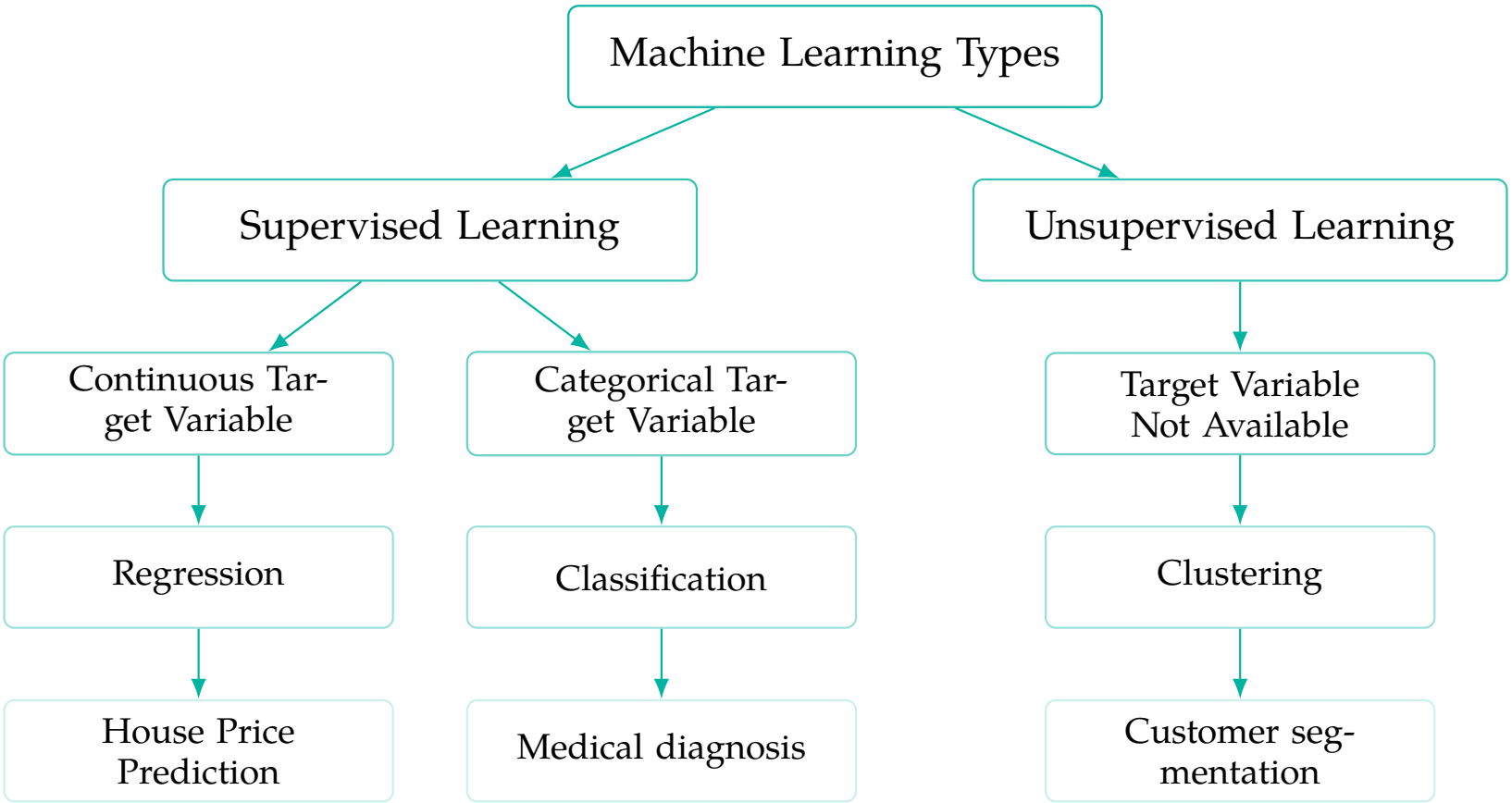


This is called a **clustering** algorithm. Example : Market segmentation.

Unsupervised Learning : Example

One example where clustering is used is in Google News (news.google.com)

Types of Machine Learning Problems



Course overview

Aims & Prerequisites

Aims

After the course you should be able to :

- Recognize which learning method is suitable for a given task.
- Describe the theory behind the methods.
- Apply the method to example problems with few data.
- Undertake an experimental assessment of learning methods and report the results.

Prerequisites

- Math.
- Linear algebra.
- Statistics & Probabilities.
- Programming in R.

Course Information

Format :

- 3h per week $\times$ 9 weeks = 27h (3h projects presentations)

Language :

- English

Material :

- Main course webpage : `http://mghassany.com/MLcourse`

Additional readings :

- The Elements of Statistical Learning (by Friedman, Tibshirani and Hastie).
- Pattern Recognition and Machine Learning (by Bishop).
- Andrew Ng.'s Machine Learning course on Coursera.

Course Information

Evaluation :

- 15 mins Quiz every week (Quiz i is about week $i-1$ and week i).
- Reports at the end of every session.
- Small project to be presented in the last session (in groups of 3).

Communication :

- Announcements on Yammer (Group : ESILV - Promo 2019 - Initiale- A4).
- Announcements via Moodle.
- Email, Desk L521.

Course Contents

Supervised Learning :

- Week 1 : Simple Linear Regression.
- Week 2 : Multiple Linear Regression.
- Week 3 : Logistic Regression.
- Week 4 : Linear Discriminant Analysis.
- Week 5 : Support Vector Machines.

Unsupervised Learning :

- Week 6 : Dimensionality Reduction.
- Week 7 : Clustering k-means.
- Week 8 : Hierarchical Clustering.
- Week 9 : Projects presentations.

About the instructors

Mohamad GHASSANY

- Associate Professor at ESILV.
- PhD in Computer Science from Université Paris 13.
- Master 2 in Applied Mathematics & Statistics from Université Grenoble Alpes.
- Works on Machine Learning research subjects.
- Teaches Machine Learning, Probability and Statistics.
- Worked as :
  - Research engineer at Telecom Business School.
  - Post-doctoral researcher at ENS Cachan.
- Personal website : mghassany.com

About the instructors

Aline ELLUL

- Teacher at ESILV since 2010.
- Worked in AI for 10 years.
- Software product manager for 11 years.
- BSc in history of art (Louvre School).
- Contact : `aline.ellul@devinci.fr`

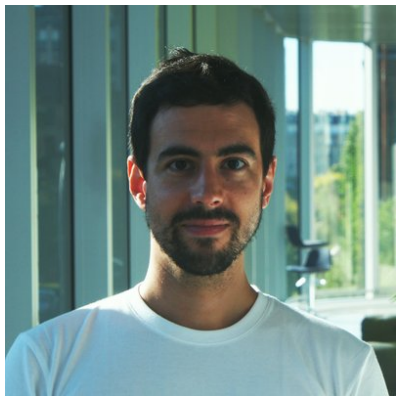
Denis OBLIN

- Marketing director in Bank & Insurance (14 years).
- Specialized Master in Big Data from Télécom ParisTech in 2014.
- Mémorandum Consulting Firm : Partner since 2014.
- Contact : `denis.oblin@gmail.com`

Fatma CHAMEKH

- Post-doc researcher at ESILV.
- PhD in CS from Université Lyon.
- Expert in semantic web.
- Contact : `fatma.chamekh@devinci.fr`

About the instructors

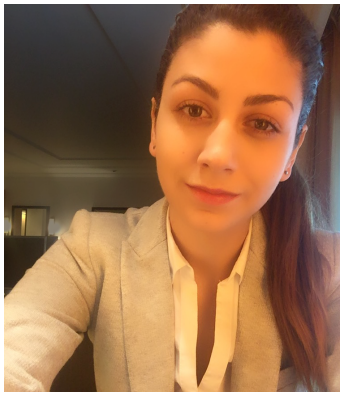


Mathieu CATTENOZ

- Expert in aerospace industry and start-up.
- PhD in Signal Processing from Paris-Saclay.
- Engineer from Centrale-Supélec & Tech. Univ. of Munich.
- Contact : `mathieu.cattenoz@gmail.com`

Mahmoud ZEIN

- Senior Manager - Analytics & Information Management at Deloitte.
- Head of BA, HSBC International Trade Finance (12 years).
- PhD in Quantitative Finance, Aix-Marseille.
- Contact : `neworizon@gmail.com`



Maggie MHANNA

- Data Scientist at Renault Digital.
- PhD in ML & Signal Processing from Paris-Saclay.
- M.Sc. in Renewable Energy from École Polytechnique.
- Contact : `maggiemhanna@gmail.com`

Have a nice course!