

# Introduction to Machine Learning

## Course Overview

**Mohamad GHASSANY**

École supérieure d'ingénieurs Léonard-de-Vinci

09/01/2018



ÉCOLE  
**D'INGÉNIEURS**  
PARIS-LA DÉFENSE

## 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 from experience.

## MACHINE LEARNING

## What is Machine Learning?

- ▶ A science of getting computers to learn without being explicitly programmed<sup>1</sup>.
- ▶ Study of algorithms that **improve** their **performance P** at **some task T** with **experience E**<sup>2</sup>.



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

**E**: images of a handwritten “a”.

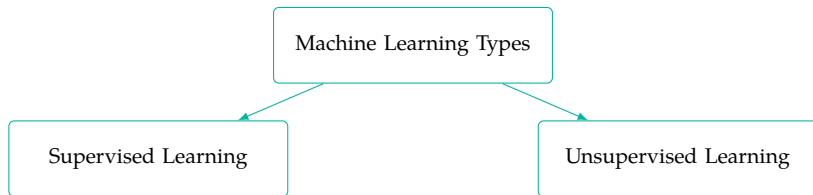
**P**: recognition rate.

---

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 :

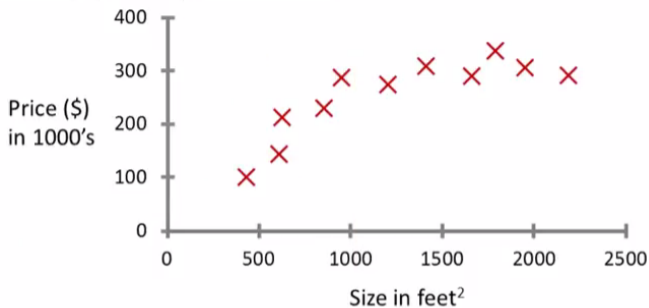


**Other** : Semi-supervised Learning, Reinforcement learning, Recommender system, etc...

# Supervised Learning

EXAMPLE : HOUSE PRICE PREDICTION<sup>3</sup>

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

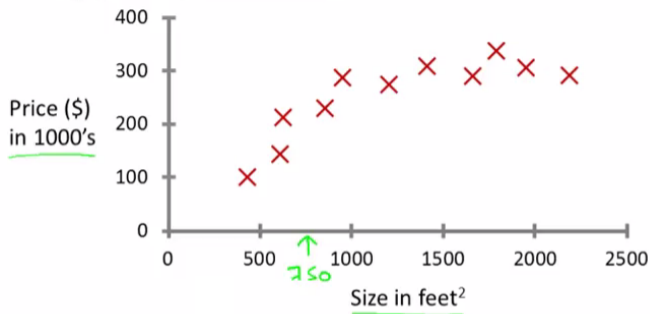


---

3. Example from Andrew Ng's MOOC.

## EXAMPLE : HOUSE PRICE PREDICTION

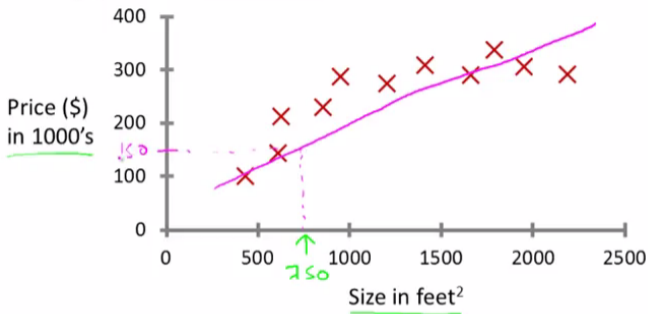
Let's say we own a house that is, say 750 square feet and hoping to sell the house and we want to know how much we can get for the house.





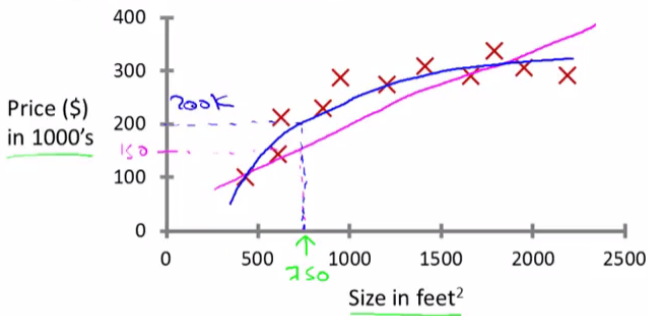
## EXAMPLE : HOUSE PRICE PREDICTION

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.



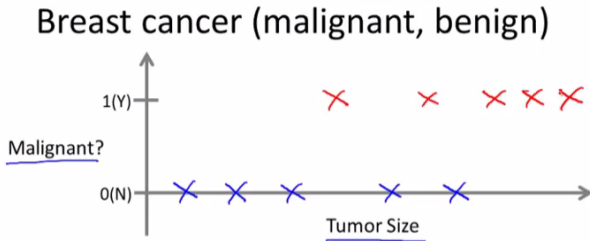
If we do that, and make a prediction here, then it looks like maybe we can sell the house for closer to 200 000\$.

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 if a breast cancer is malignant or benign<sup>4</sup>.

A collected data set gave the following :



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

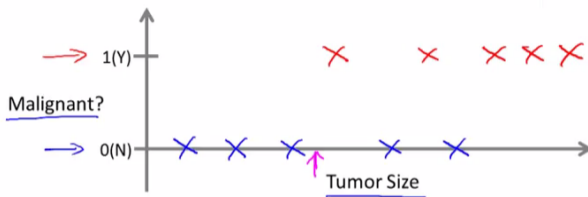
---

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.

## Breast cancer (malignant, benign)

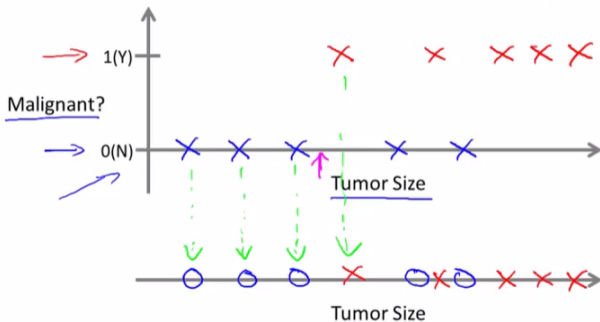


- The machine learning question here is, can you estimate what is the **probability** that a tumor is malignant versus benign?

## EXAMPLE : MEDICAL DIAGNOSIS

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

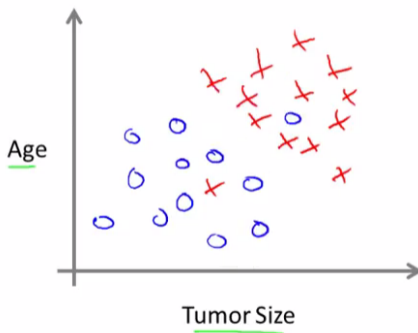
## Breast cancer (malignant, benign)



In this example, there is only one **input**.

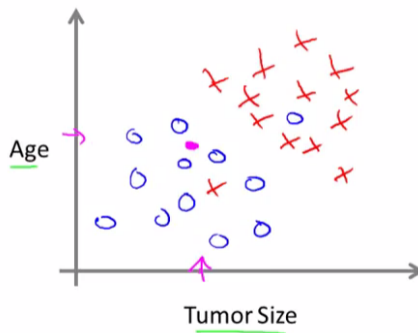
## EXAMPLE : MEDICAL DIAGNOSIS

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.



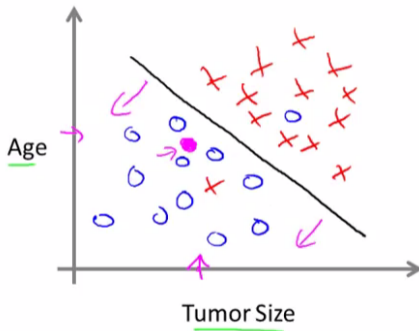
## EXAMPLE : MEDICAL DIAGNOSIS

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

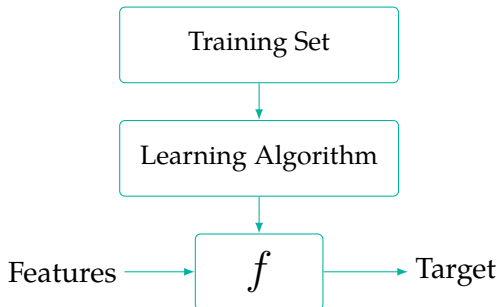
Given a data set like this, a learning algorithm may throw a straight line through the data to try to separate out the malignant tumors from the benign ones.





## SUPERVISED LEARNING : DEFINITION &amp; 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

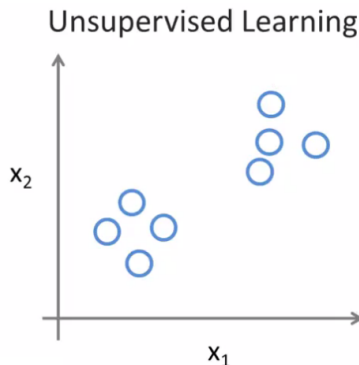
Notations :

- ▶ 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"

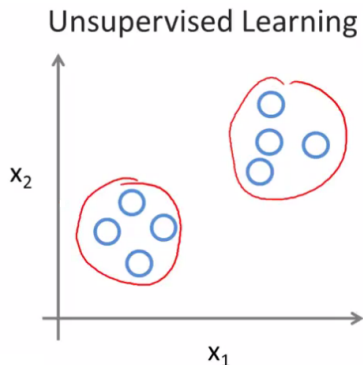
In Unsupervised Learning, we're given data that doesn't have any **labels**.  
For example :



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 : EXEMPLE

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

The screenshot shows the Google News homepage with a search bar at the top. Below the search bar, there are several news articles organized into sections. The first section is titled 'Actualités' and features an article about 'Peillon et Valls présentent leurs programmes'. The second section is titled 'A la une' and features an article about 'Un nourrisson est mort après une prise de vitamine D'. The third section is titled 'Foot - Transfert' and features an article about 'Le PSG officialise l'arrivée de Julian Draxler'. The fourth section is titled 'Météo à Arcueil, Île-de-France' and shows weather information. The fifth section is titled 'Le choix des rédactions' and features the EUROSPORT logo. The sixth section is titled 'La L1 cherche des renforts à prix raisonnable' and features an article about 'Draxler débarque officiellement au PSG'. The seventh section is titled 'Le PSG va prolonger Cavani mais peine à trouver un nouveau point de...' and features an article about 'Griezmann: "Je me posez plus de questions sur mon avenir"'. The eighth section is titled 'Le mercato EN DIRECT : Arsenal pense à Planić' and features an article about 'Fabian Borne'.

Google

Actualités - Edition France

A la une

**Peillon et Valls présentent leurs programmes**

Le Monde - il y a 2 heures

Les deux candidats à la primaire à gauche ont exposé les grandes lignes de leurs projets, mardi 3 janvier. Le Monde | 03.01.2017 à 11h54 • Mis à jour le 03.01.2017 à 12h35 | Par Bastien Bonnefous. Manuel Valls, sur le marché de Noël de Strasbourg, le ...

Autres: Manuel Valls »

**Un nourrisson est mort après une prise de vitamine D**

Le Monde - il y a 4 heures

Des malaises ont déjà été signalés après l'administration d'Uvestérol. Des investigations doivent déterminer si le décès survenu en décembre est imputable à ce médicament. Le Monde | 02.01.2017 à 21h02 • Mis à jour le 03.01.2017 à 12h06 | Par Paul ...

Autres: Ergocalcérol »

**Foot - Transfert - Le PSG officialise l'arrivée de Julian Draxler (Wolfsburg)**

L'Espresso.fr - il y a 2 heures

Le PSG a annoncé la signature de Julian Draxler, ce mardi, en provenance de Wolfsburg. Le milieu offensif allemand s'est engagé jusqu'en 2021. Partager sur Facebook Twitter Google+ 0 partages. Football - Transferts Football - Julian Draxler a signé un ...

Autres: Paris Saint-Germain Football Club » Julian Draxler »

**Météo à Arcueil, Île-de-France**

Aujourd'hui mer. 3° - 1°

Dem. 7° - 0°

Mer. 6° - 3°

Ven. 2° - 3°

**Arcueil, Île-de-France »**

Paris: 14 000 habitants ont quitté la capitale depuis 2009

MCE Ma Chaine Etudiante - il y a 2 heures

Théâtre du Châtelet : la Britannique Ruth Mackenzie prend la direction

Le Point - il y a 35 minutes

Où vivre en sécurité en région parisienne ?

Une étude inédite au cas par cas

20minutes.fr - il y a 1 heure

**Le choix des rédactions**

**EUROSPORT**

La L1 cherche des renforts à prix raisonnable ? Le onze des bonnes pioches

Ludovic Alard

Draxler débarque officiellement au PSG

Cyril Morin

Le PSG va prolonger Cavani mais peine à trouver un nouveau point de...

Pierre-Alexandre Comte

Griezmann: "Je me posez plus de questions sur mon avenir"

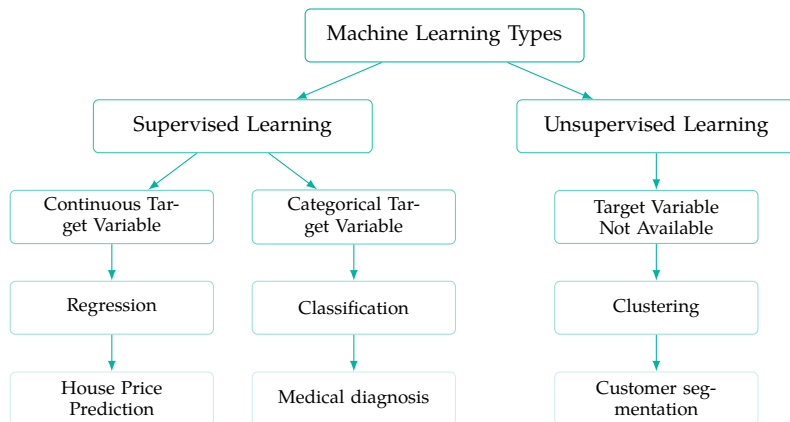
Europe1

Le mercato EN DIRECT : Arsenal pense à Planić

Fabian Borne

Téléchargez l'application Eurosport.fr

# 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](http://mghassany.com)



ÉCOLE  
D'INGÉNIEURS  
PARIS-LA DÉFENSE



Institut  
Mines-Télécom



## 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](mailto:aline.ellul@devinci.fr)



**DENIS OBLIN**

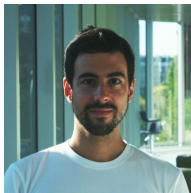
- ▶ Marketing director in Bank & Insurance (14 years).
- ▶ Specialized Master in Big Data from Télécom ParisTech in 2014.
- ▶ Mémorandum Consulting Firm : Partner since 2014.
- ▶ Contact : [denis.oblin@gmail.com](mailto:denis.oblin@gmail.com)



**FATMA CHAMEKH**

- ▶ Post-doc researcher at **ESILV**.
- ▶ PhD in CS from Université Lyon.
- ▶ Expert in semantic web.
- ▶ Contact : [fatma.chamekh@devinci.fr](mailto:fatma.chamekh@devinci.fr)

## ABOUT THE INSTRUCTORS



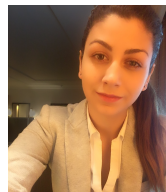
**MATHIEU CATTENOZ**

- ▶ 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](mailto: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](mailto: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](mailto:maggiemhanna@gmail.com)



Have a nice course!