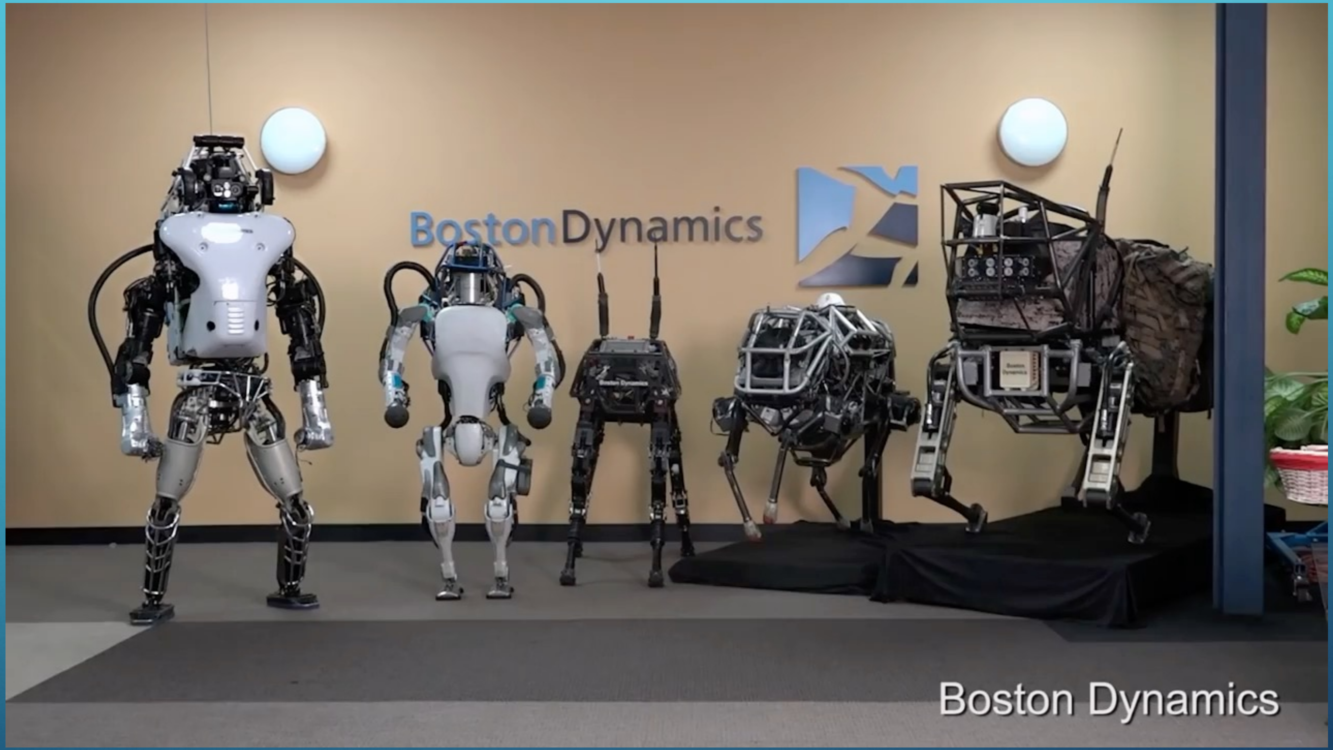
MÉTHODE SUPERVISÉE

OnBoardingSchool 2023

VALÉRIE GAUTARD

FASCINANTE IA



OUTLINE

- ML basis
- Decision trees
- Ensemble learning
- SVM

IA, A DEFINITION

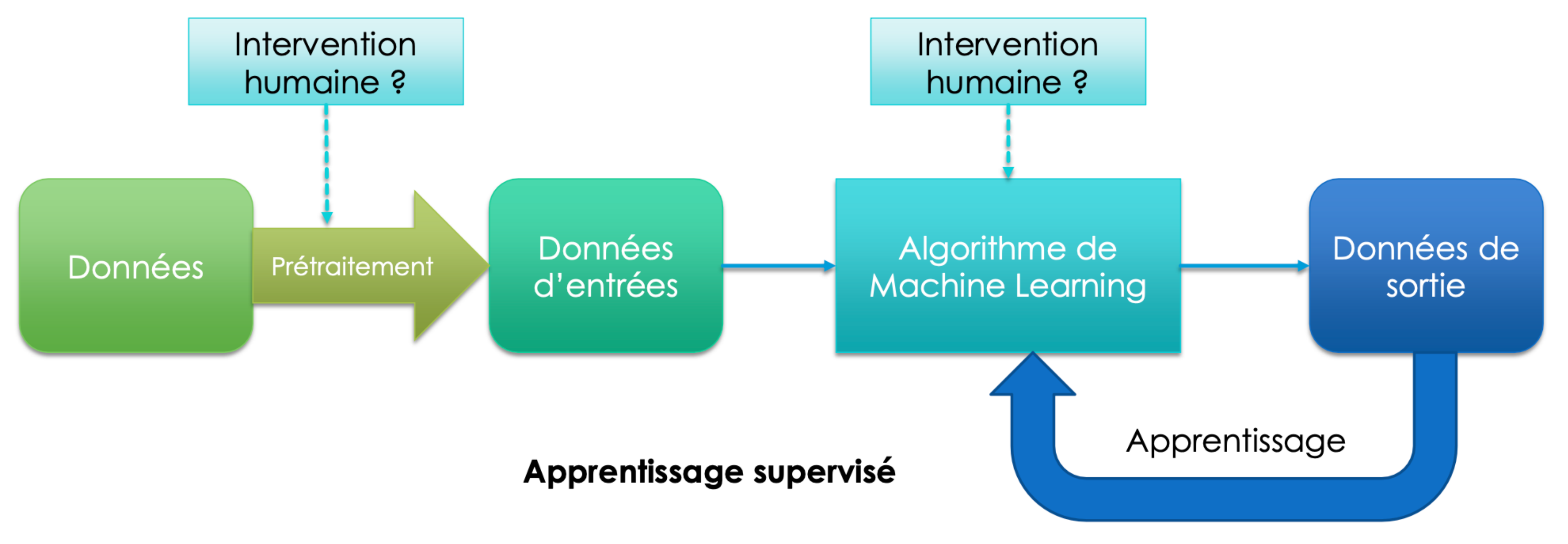
- "A computer program is said to learn a class of tasks T from experience E with a performance measure P if its performance on tasks T, as measured by P, improves with experience E ."

  Tom Mitchell (1997)

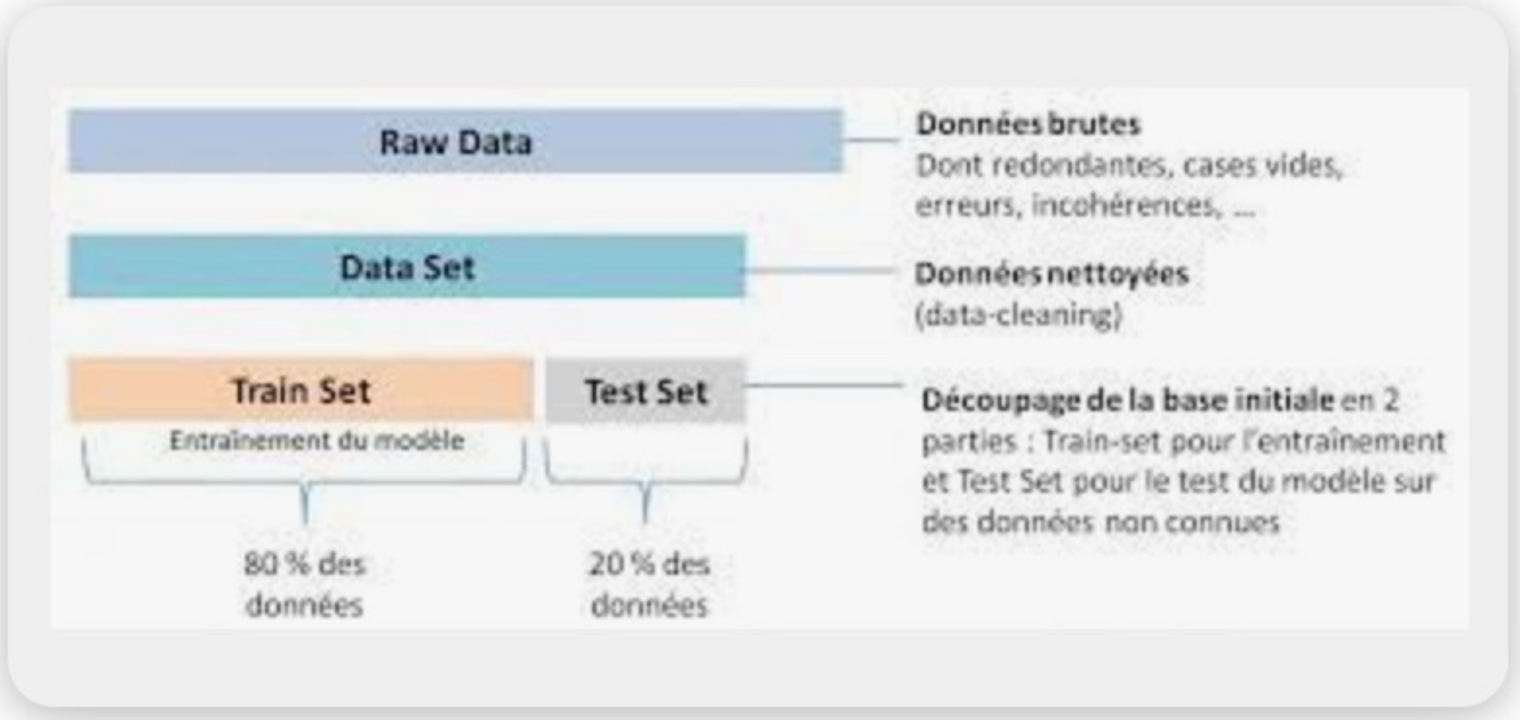
- Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed

  attributed to Arthur Samuel (1959)

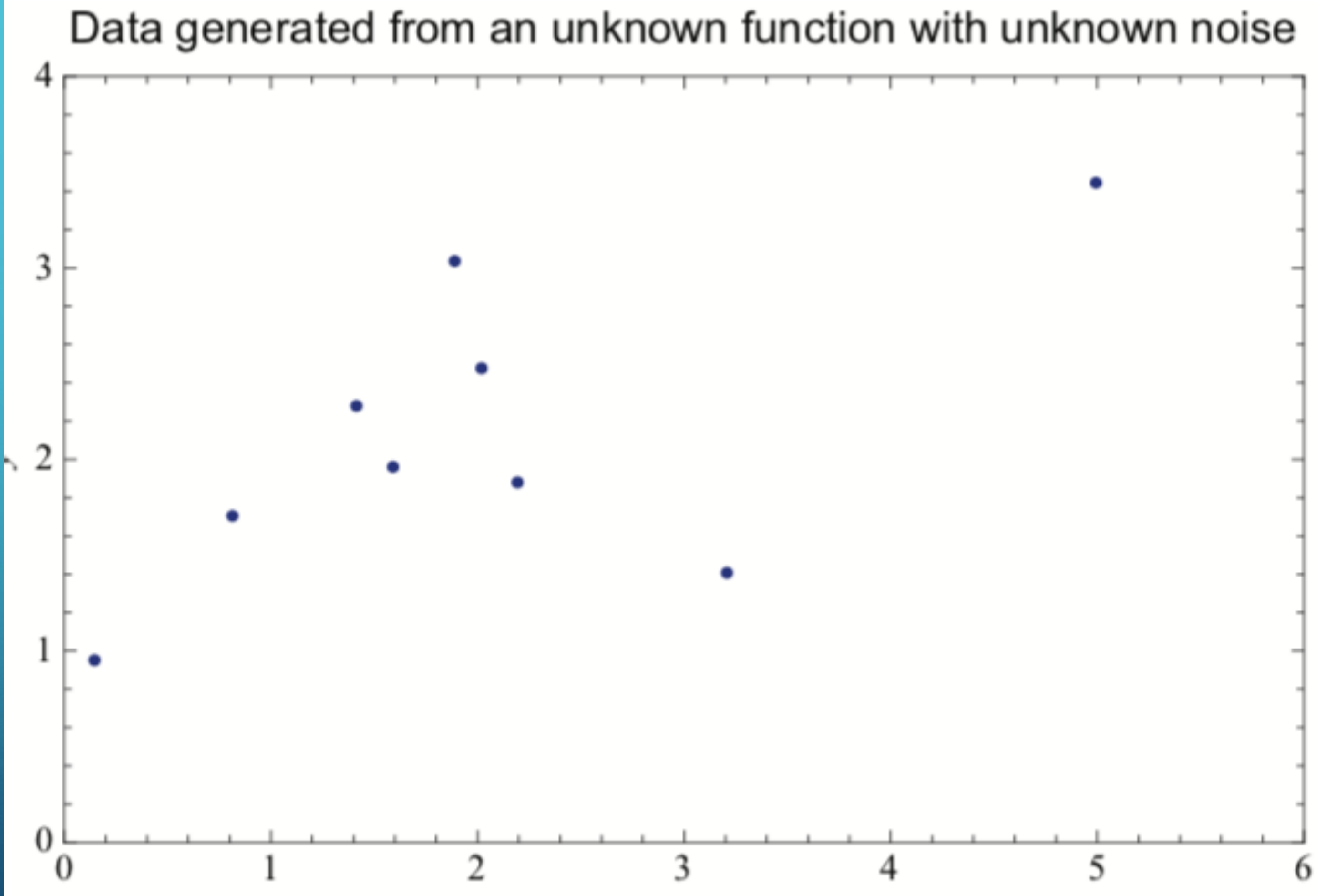
THE ML PROCESS (1/2)



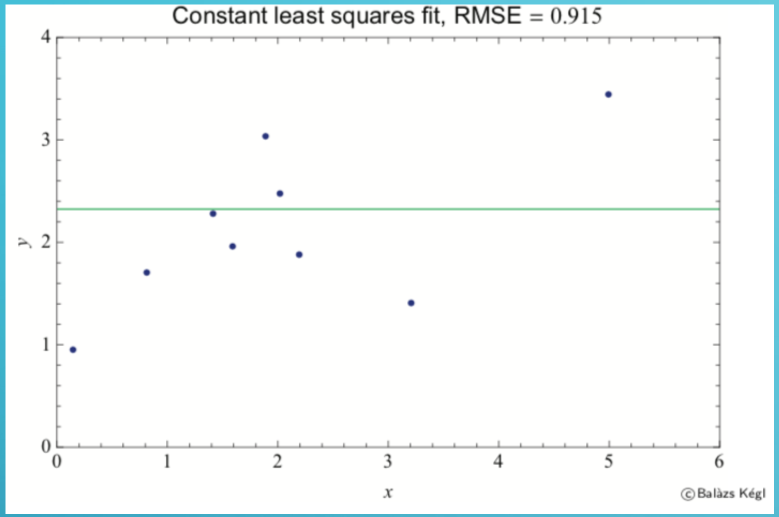
THE ML PROCESS (2/2)

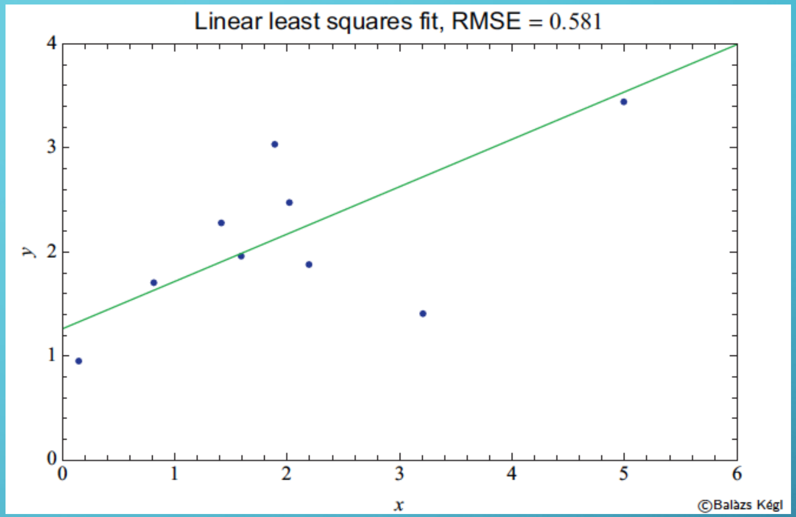


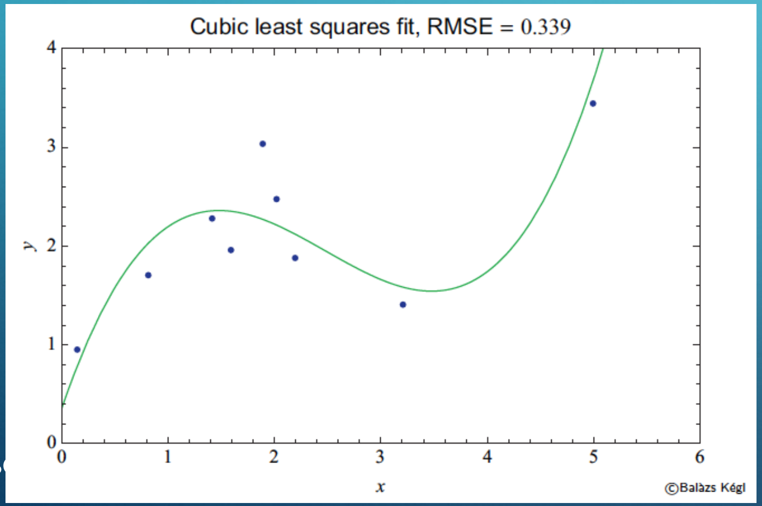
CHOICE OF FUNCTION CLASS

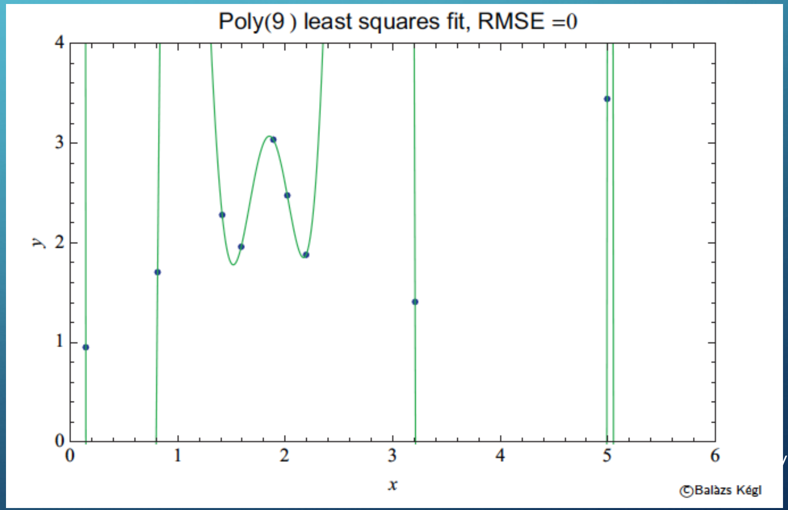


CHOICE OF FUNCTION CLASS

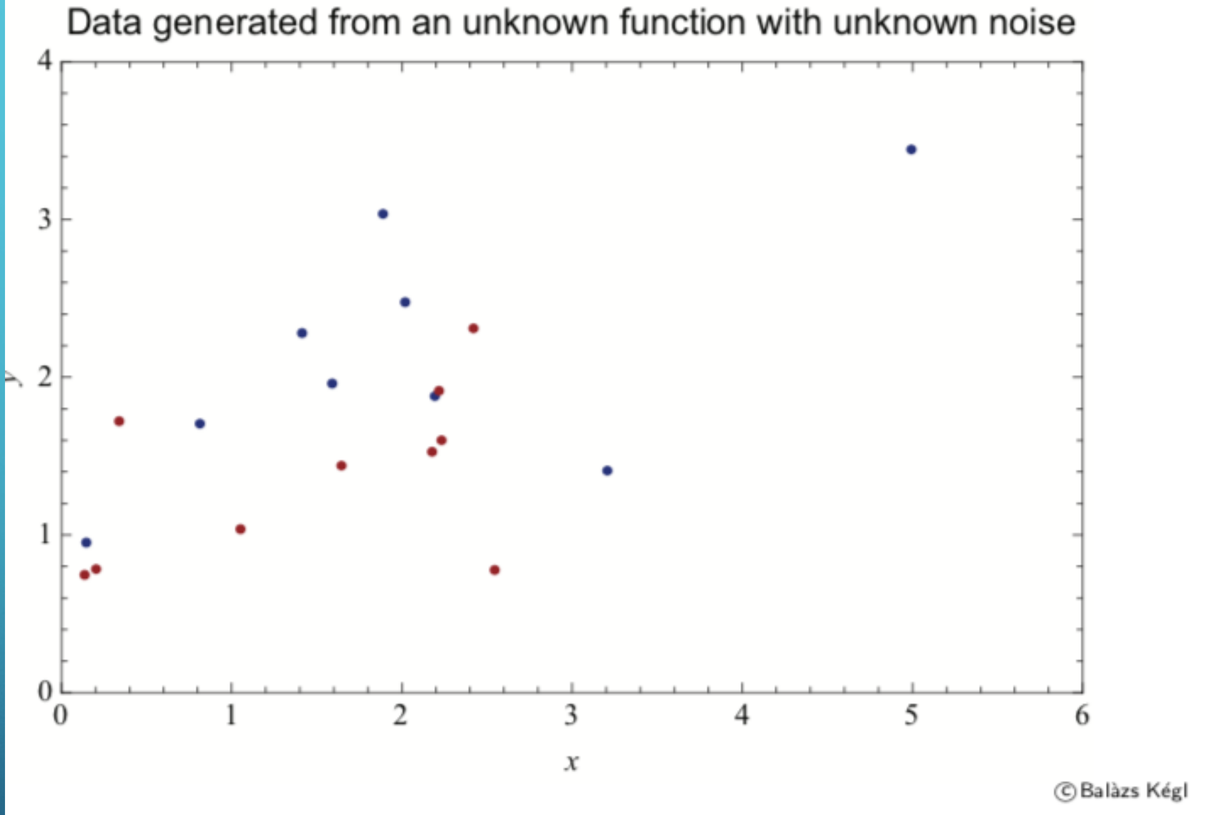




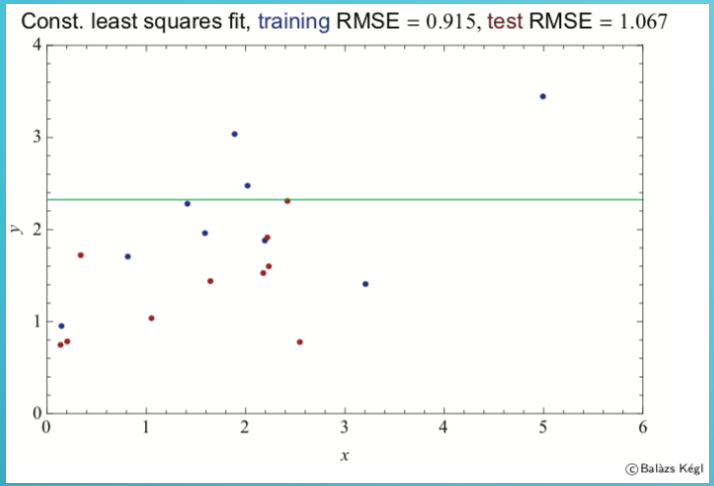


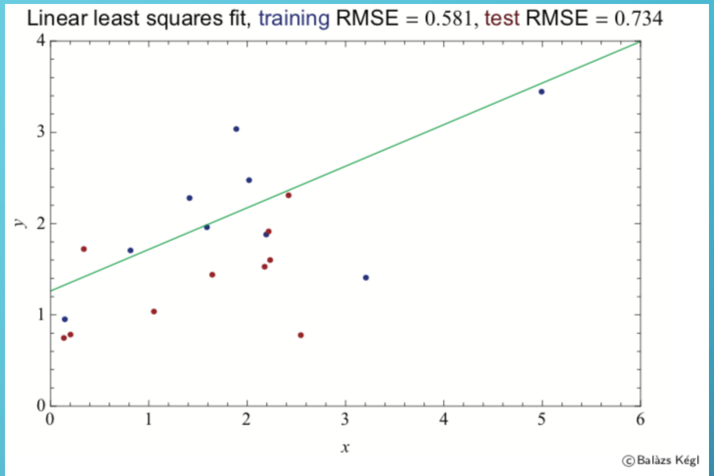


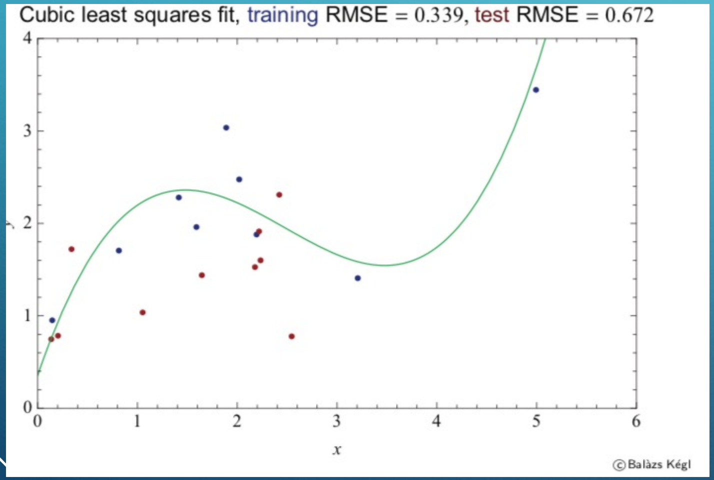
CHOICE OF FUNCTION CLASS

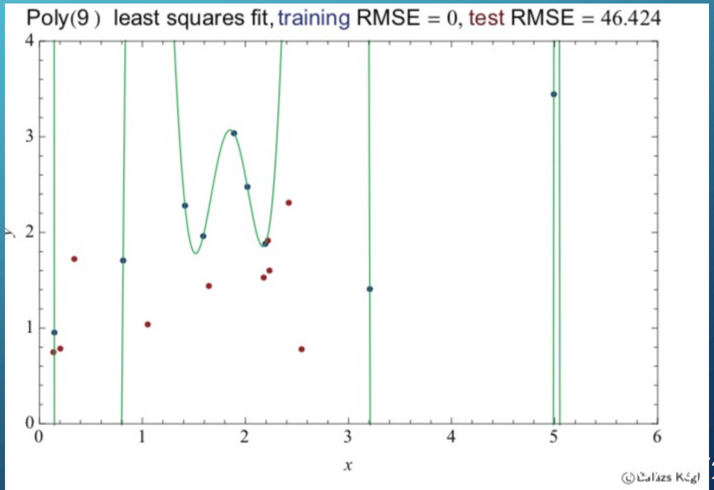


CHOICE OF FUNCTION CLASS

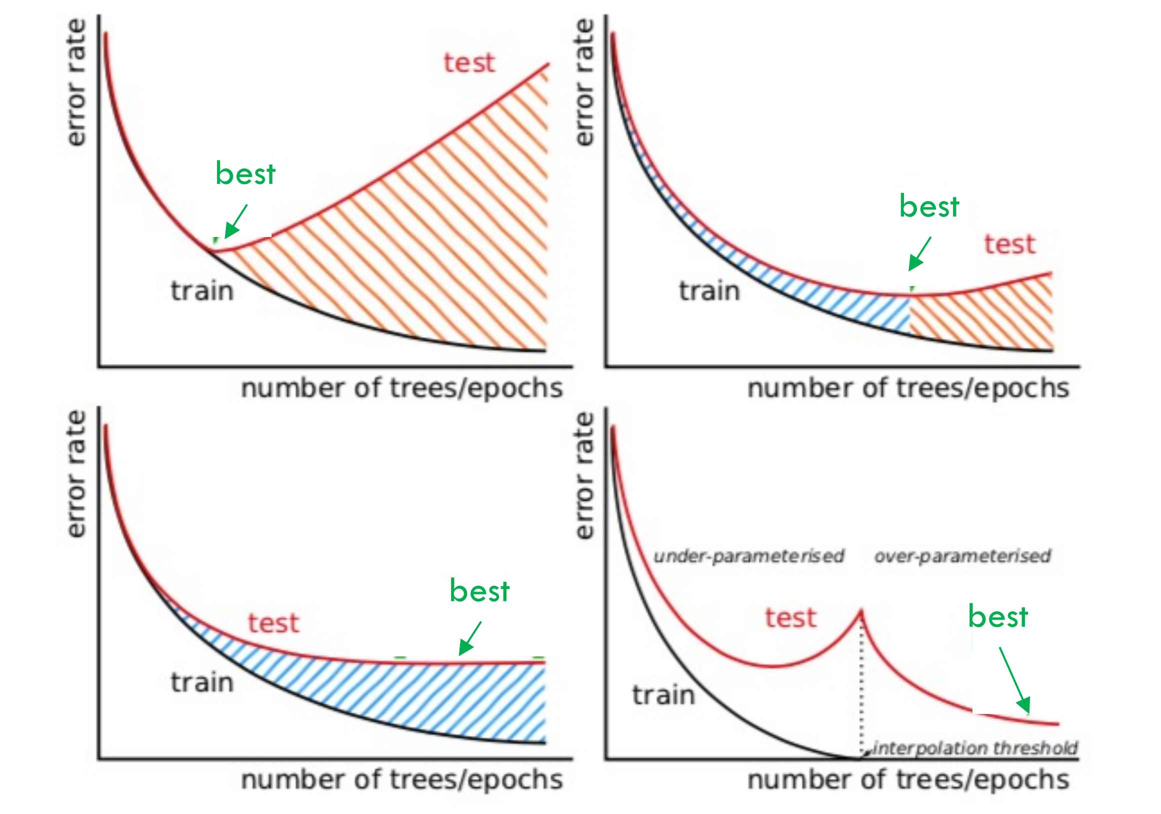








SUR-ENTRAINEMENT/SOUS-ENTRAINEMENT



Trop de sur-entrainement (overfitting) / bon sur-entrainement (encore sous entrainement)

SUPERVISED LEARNING METHODS

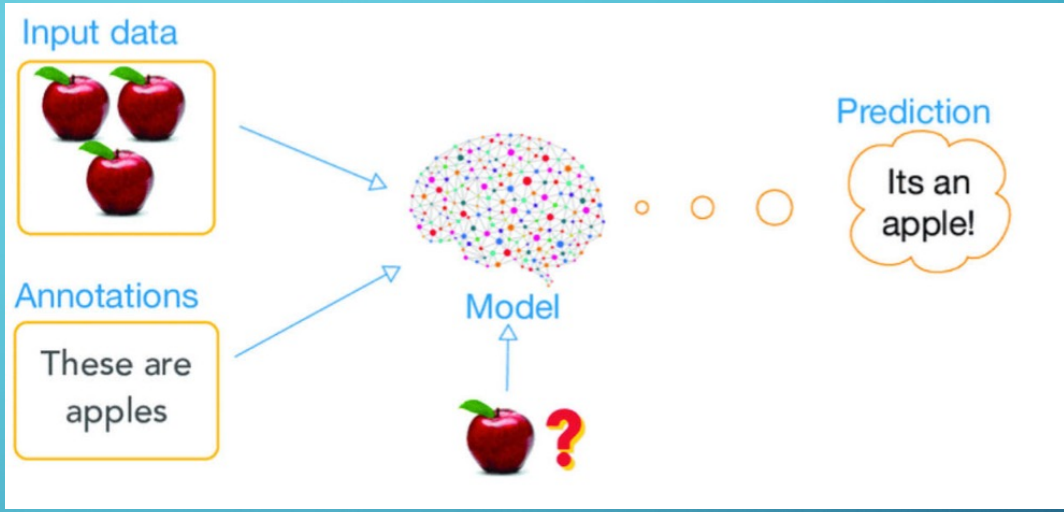
- Labelled data
- Let's take data:

    N examples

    $(x,y)_1, (x,y)_2, \ldots, (x,y)_N$

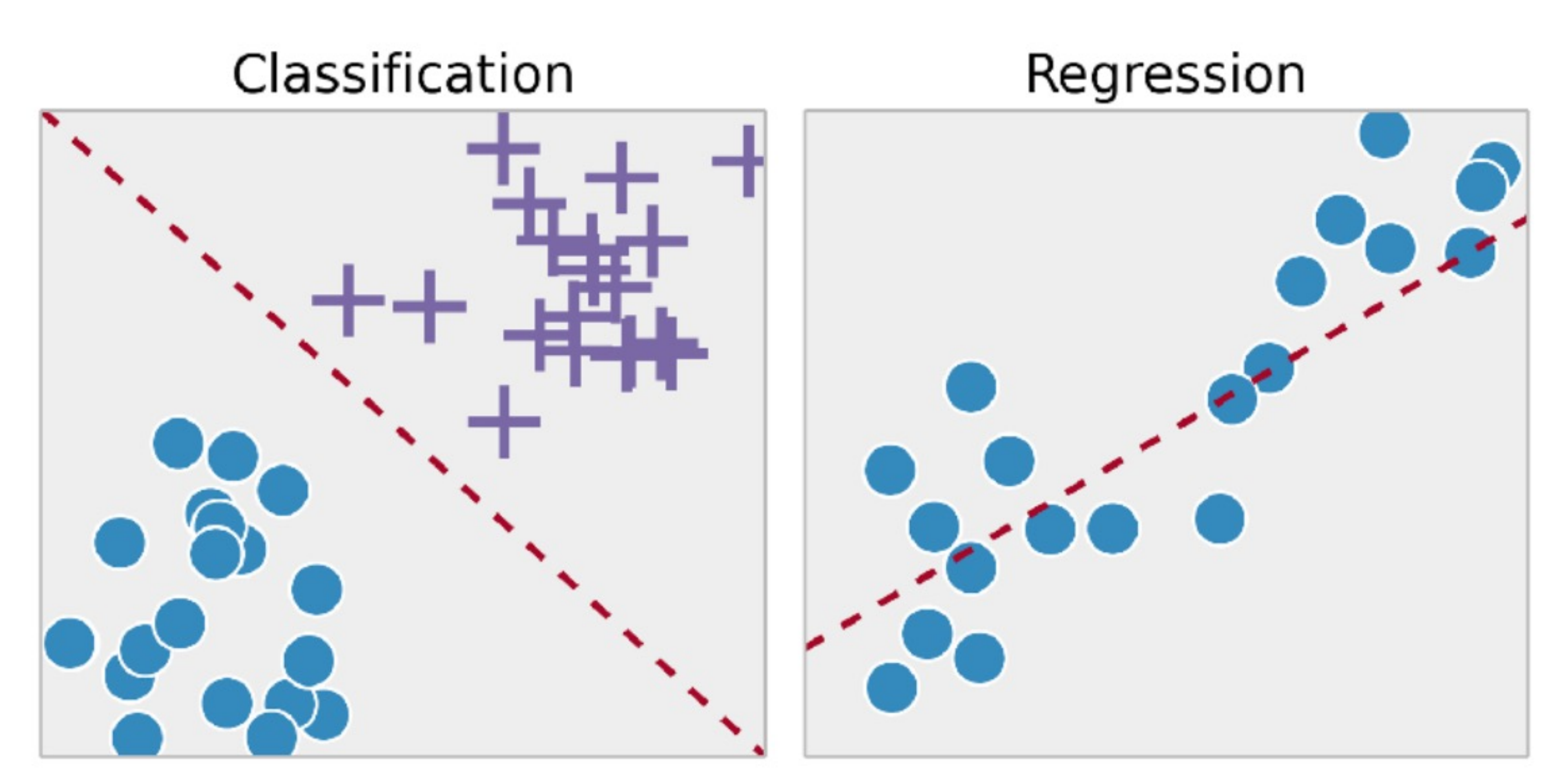
    x : feature

    y : label

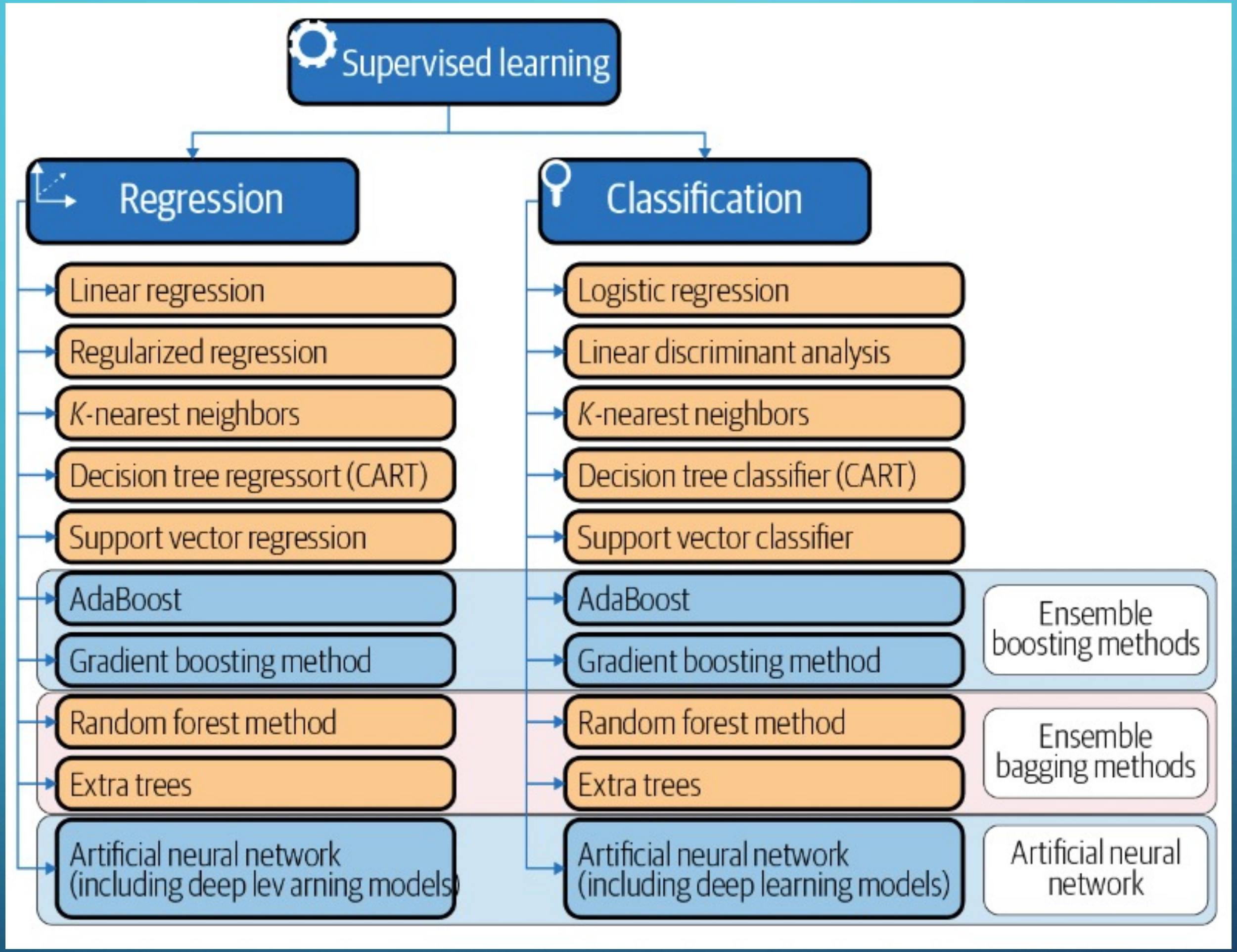


- Use labelled example to know how good algorithm is doing
- Example :
    - Evaluate the evolution of the price of a house
    - Grant or not of a loan

SUPERVISED LEARNING METHODS



SUPERVISED LEARNING METHODS

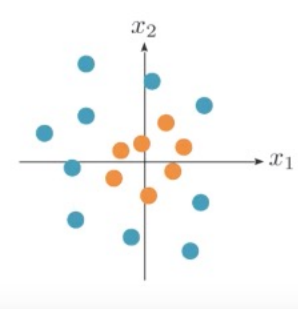


OUTLINE

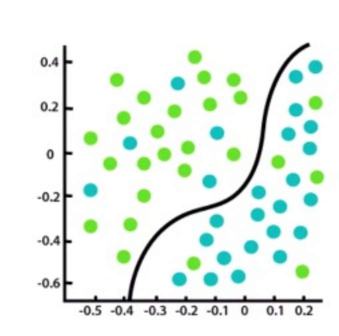
- ML basis
- Decision trees
- Ensemble learning
- SVM

DECISION TREE

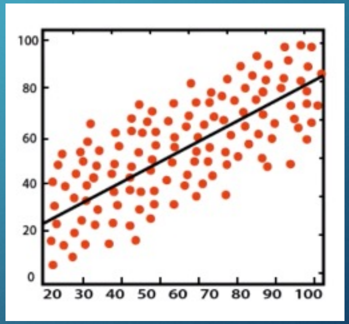
- Supervised method
- Classification or regression
- Transparent algorithm



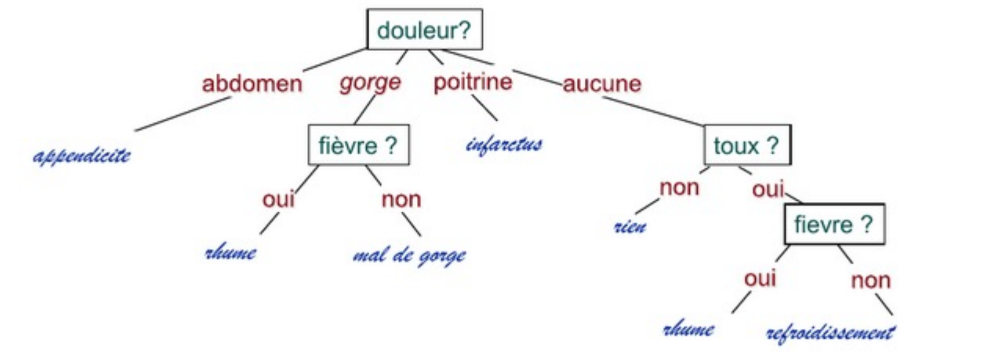
Classification



Regression



	Toux	Fièvre	Poids	Douleur
Marie	non	oui	normal	gorge
Fred	non	oui	normal	abdomen
Julie	oui	oui	maigre	aucune
Elvis	oui	non	obese	poitrine



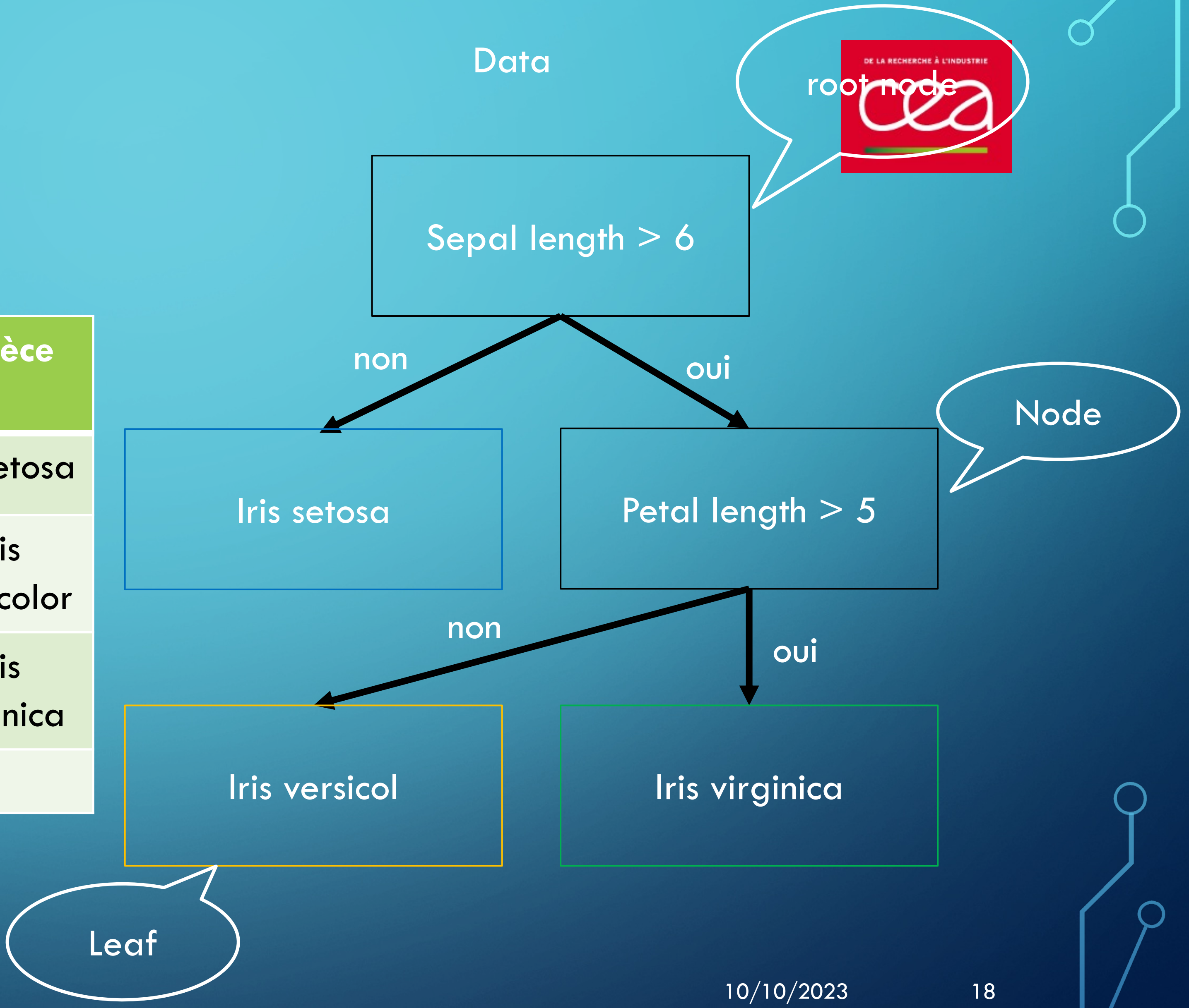
DECISION TREE

- Example with the Iris database

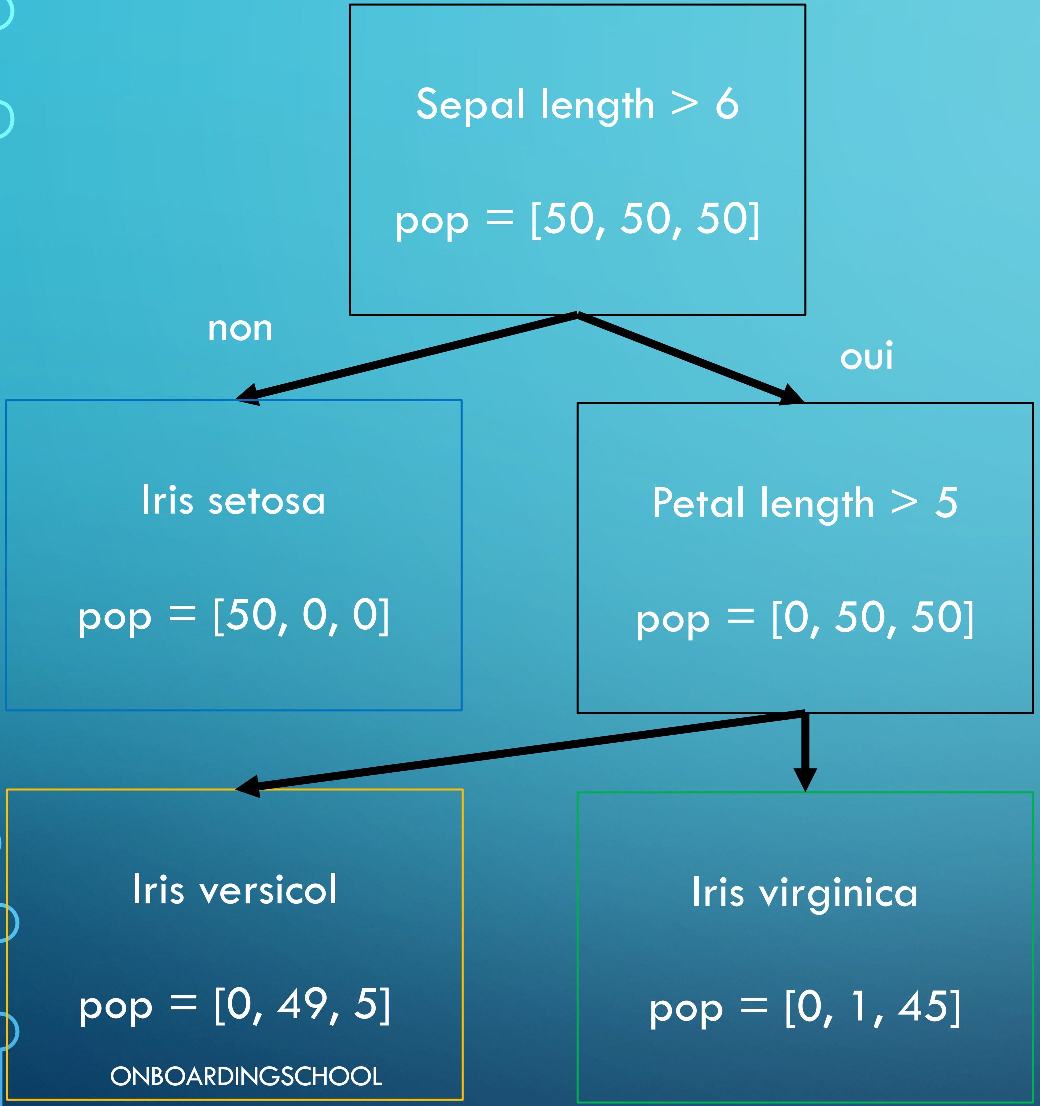
Sepal Length	Sepal width	Petal Length	Petal width	Espèce
5.1	3.5	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica

DECISION TREE

Sepal Length	Sepal width	Petal Length	Petal width	Espèce
5.1	3.5	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica



DECISION TREE



probabilité d'appartenance à une classe

pop = [nb setosa, nb versicolor, nb virginica]

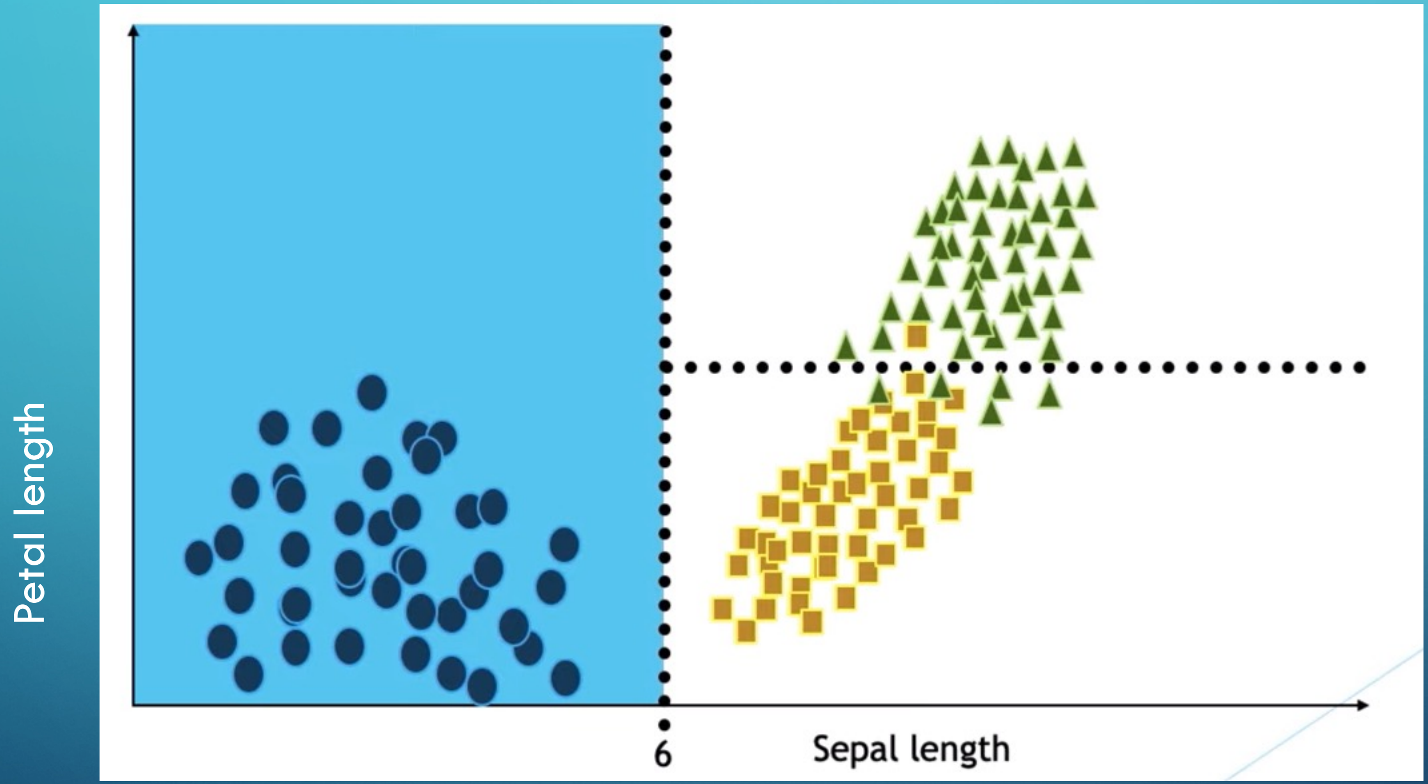
Par exemple :

Identifions une fleur avec un sépale de 6.5 cm et un pétale de 4 cm

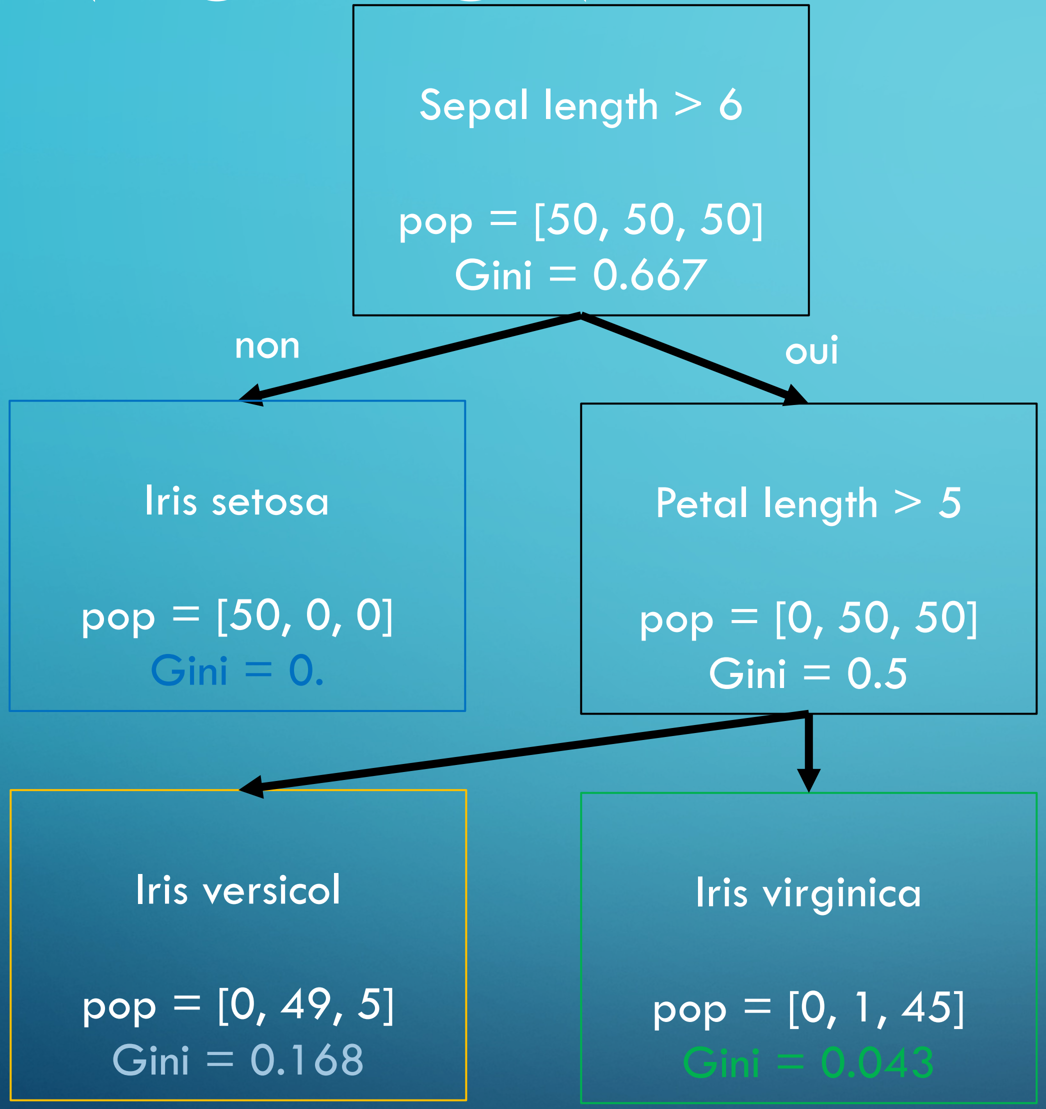
Setosa :	$0/54 = 0$
Versicolor :	$49/54 = 0.91$
virginica :	$5/54 = 0.09$

DECISION TREE

Frontière de décision



INDICE DE GINI



$$G_i = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$

$p_{i,k}$ est le ratio du nombre d'individus de la classe k parmi la population du $i^{ème}$ noeud.

$$\text{gini} = 1 - \left(\frac{50}{150}\right)^2 - \left(\frac{50}{150}\right)^2 - \left(\frac{50}{150}\right)^2 =$$

$$\text{gini} = 1 - \left(\frac{50}{50}\right)^2 - \frac{0}{50} - \frac{0}{50} = 0$$

$$\text{gini} = 1 - \frac{0}{50} - \left(\frac{50}{100}\right)^2 - \left(\frac{50}{100}\right)^2 = 0.5$$

$$\text{gini} = 1 - \frac{0}{54} - \left(\frac{49}{54}\right)^2 - \left(\frac{5}{54}\right)^2 = 0.168$$

$$\text{gini} = 1 - \frac{0}{46} - \left(\frac{1}{46}\right)^2 - \left(\frac{45}{46}\right)^2 = 0.043$$

DECISION TREE, PURITY

- Purity represents the cost of the node

$$J(k) = \frac{m_{gauche}}{m} G_{gauche} + \frac{m_{droite}}{m} G_{droite}$$

$$Ou \begin{cases} G_{gauche/droite} \text{ mesure l'impureté du sous ensemble droite/gauche} \\ m_{gauche/droite} \text{ est la proportion de notre population du sous ensemble droite/gauche} \end{cases}$$

Sepal length > 6
pop = [50, 50, 50]
Gini = 0.667

non

Iris setosa
pop = [50, 0, 0]
Gini = 0.

Petal length > 5
pop = [0, 50, 50]
Gini = 0.5

$G_{gauche} = 0$

$\frac{m_{gauche}}{m} = 50/150$

$G_{droite} = 0.5$

$\frac{m_{droite}}{m} = 100/150$

$$J(0) = \frac{50}{150}\, 0 + \frac{100}{150}\, 0.5 = 0.33$$

DECISION TREE, CHOICE OF THE NODE

Sepal length > 6
pop = [50, 50, 50]
Gini = 0.667

non

Iris setosa
pop = [50, 0, 0]
Gini = 0.

oui

Petal length > 5
pop = [0, 50, 50]
Gini = 0.5

Sepal length > 6
pop = [50, 50, 50]
Gini = 0.667

non

Iris setosa
pop = [50, 20, 0]
Gini = 0.41

oui

Petal length > 5
pop = [0, 30, 50]
Gini = 0.47

$$J(0) = \frac{50}{150}\,0 + \frac{100}{150}\,0.5 = 0.33$$

$$G_i = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$

$$J(0) = \frac{70}{150}\,0.41 + \frac{80}{150}\,0.47 =$$

DECISION TREE

- Advantage :
  - easy to train
  - easy to use
  - interpretable (transparent algorithm)

- But…
  - imprecise
  - unreliable generalization

OUTLINE

- ML basis
- Decision trees
- Ensemble learning
- SVM

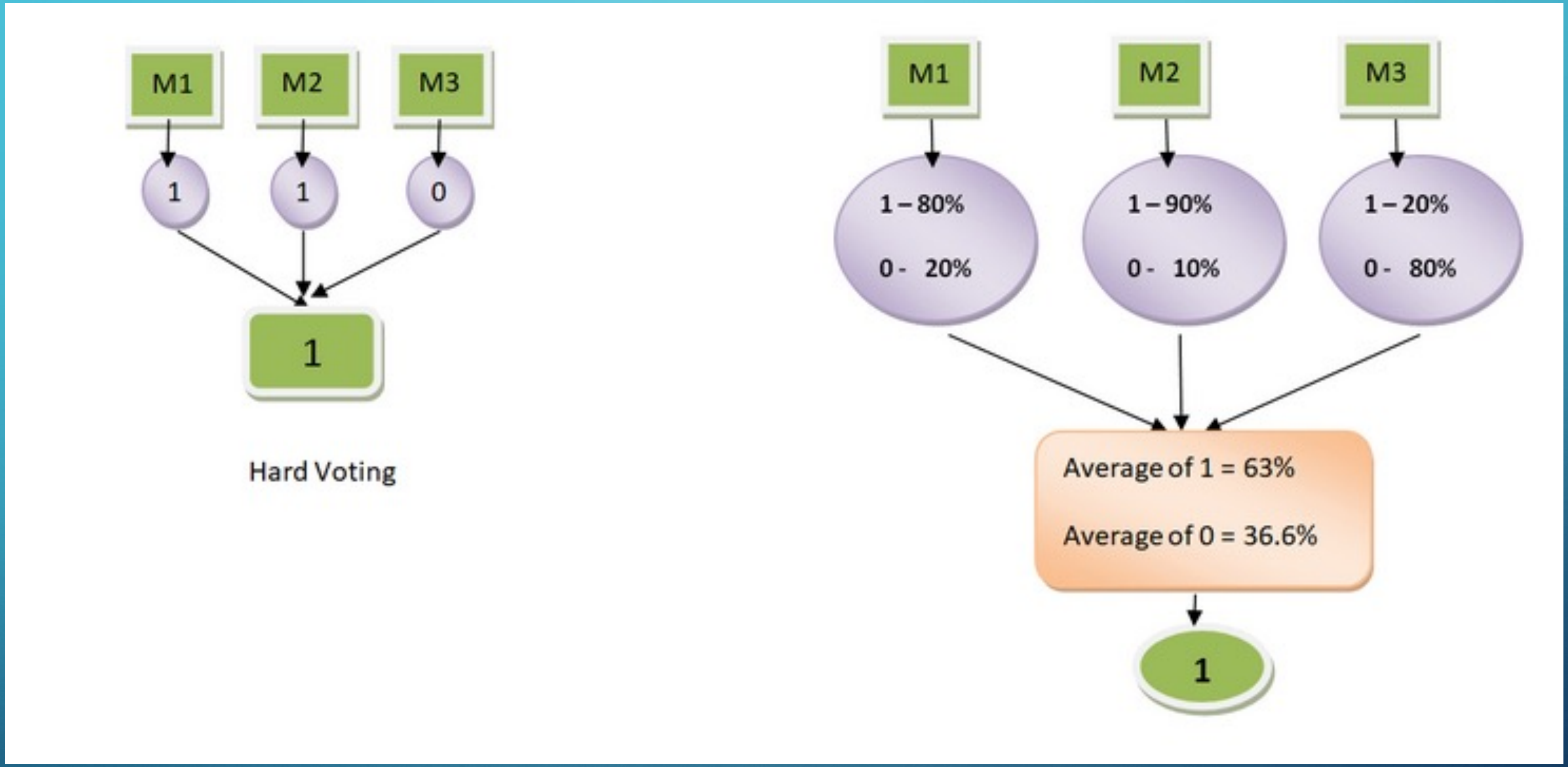
ENSEMBLE LEARNING

- Training several small ML models to make a better performing model
- Inter- and intra-operator variability
  - intra-operator: variability for the same operator
  - Inter-operator: variability for different operators
- example
  - tumor segmentation
  - estimate the price of a house
- Assumes that solving a problem is more effective by a crowd than by an expert alone
- Crowd hypothesis:
  - diversity
  - independence
  - decentralization: judgments add up, no higher authority to decide
- Ensemble learning methods use several learning algorithms and take into account the results of these models in order to obtain better predictive performance than the models taken separately.

ENSEMBLE LEARNING, EXAMPLE

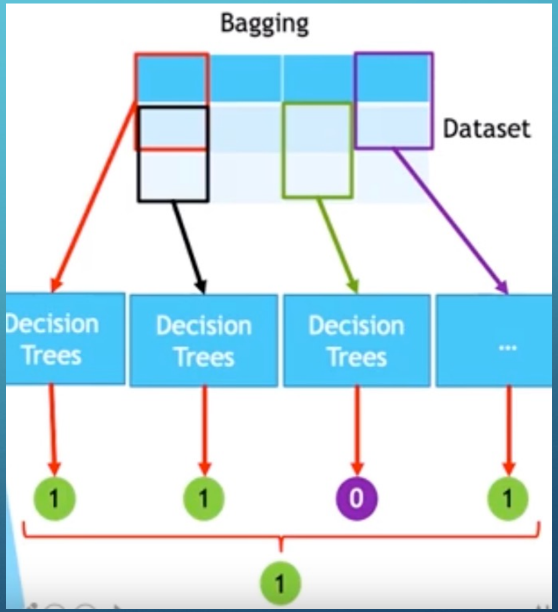
Index	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Mélange
1	1	1	0	0	1	1
2	1	1	1	0	0	1
3	0	0	1	1	1	1
4	0	1	1	1	0	1
5	1	0	1	0	1	1
	60%	60%	60%	60%	60%	100%

ENSEMBLE LEARNING



RANDOM FOREST

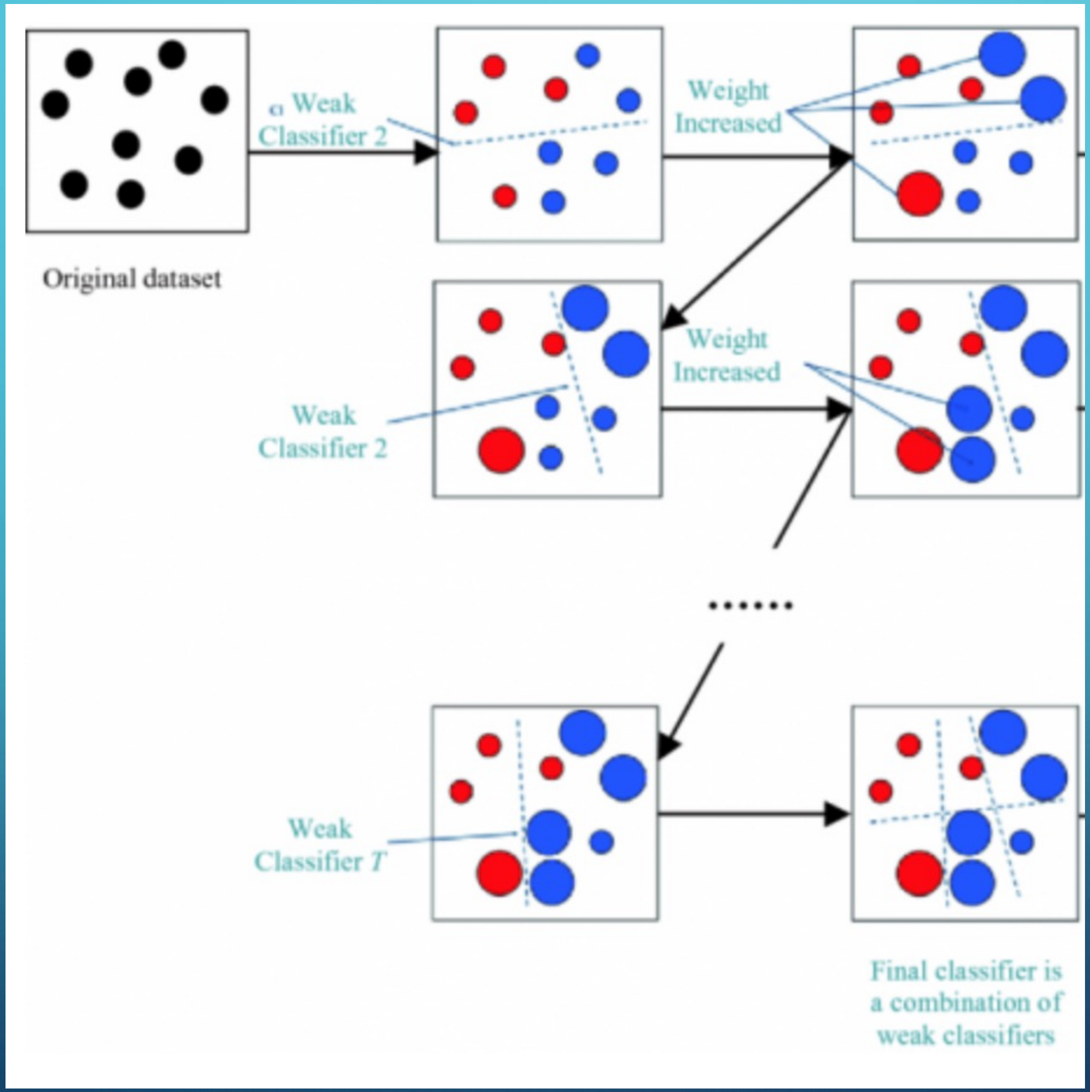
- To overcome the problem of generalizing decision trees, several decision trees are used: Random Forest
- Bagging : We train the decision tree only on
  - part of the data
  - part of the variables



RANDOM FOREST

- Steps
  - Repeat the steps until you have the desired number of trees
    - Create a dataset (with bagging you can take the same data several times). Random selection enhances tree variability
    - Train trees
  - Obtaining a forest of trees.
  - We average the results of the forest of trees

BOOSTED TREE

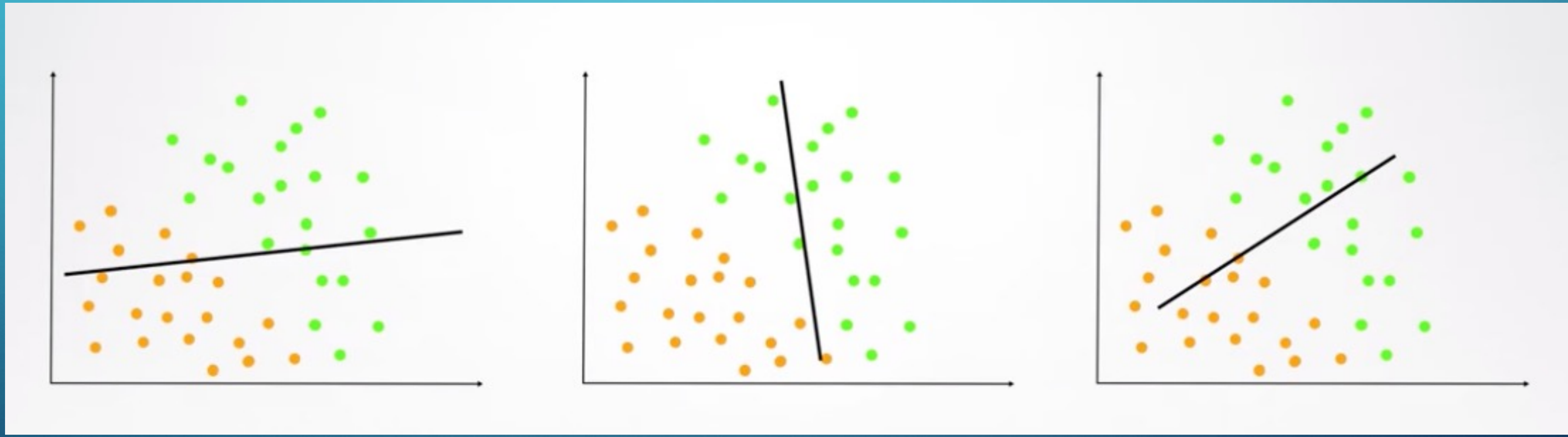


OUTLINE

- ML basis
- Decision trees
- Ensemble learning
- SVM

SVM

- Méthode supervisée
- Clustering
- 2 idées principales :
  - hyperplan à marge maximale
  - noyau



SVM

- La marge est la distance entre la frontière de séparation et les points les plus proches de cette frontière, appelés vecteurs supports
- Droite qui maximise la marge

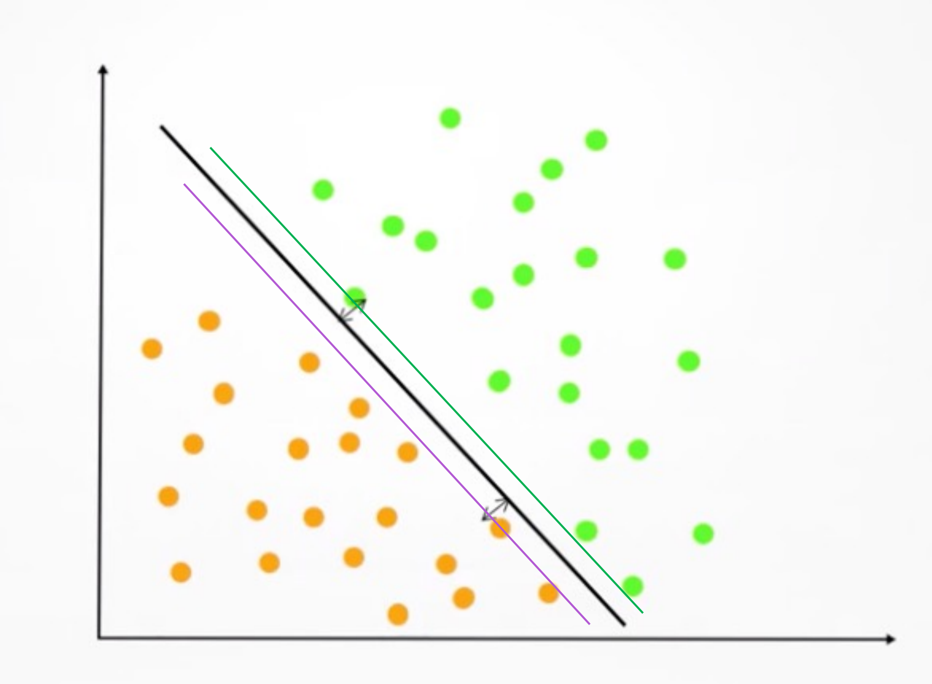
Données : $x = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T,$

Poids : $w = (\mathbf{w}_1, \ldots, \mathbf{w}_N)^T$

Fonction discriminante linéaire :

$$h(x) = w^T x + w_0$$

$$\max_{w,w_0} \min_k \{||x - x_k|| : x \in \mathbb{R}^N, w^T x + w_0 = 0\}$$

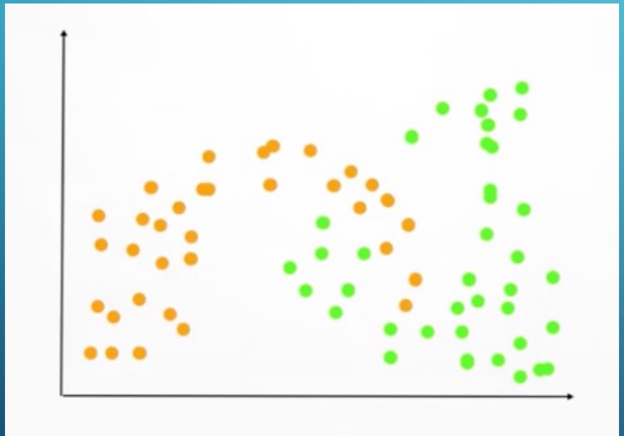


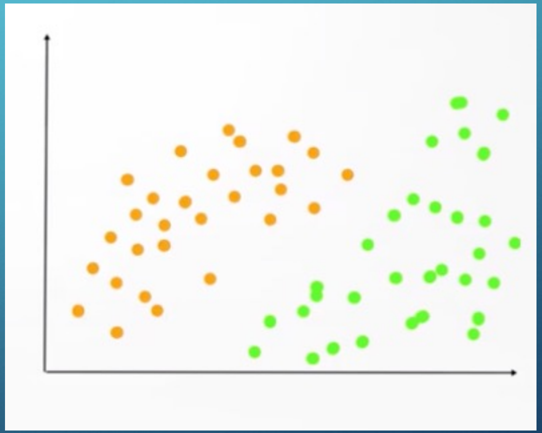
En utilisant les multiplicateurs de lagrange :

$$\text{Maximiser } \tilde{L}(\alpha) = \sum_{k=1}^{p} \alpha_k - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j l_i l_j x_i^T x_j \qquad (2) \text{ sous les contraintes } \alpha_k \geq 0 \text{ , et } \sum_{k=1}^{p} \alpha_k l_k = 0$$

SVM

- Astuce du noyau : l’idée est d’appliquer une transformation sur nos données pour pouvoir les séparer linéairement.
- On transforme l’espace de représentation des données d’entrées en un espace de plus grande dimension (potentiellement de dimension infinie) dans lequel il existe une séparation linéaire

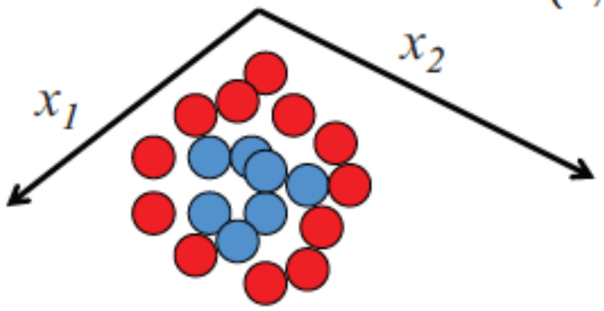


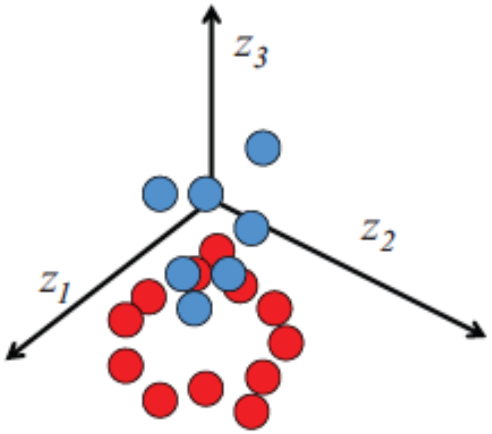


SVM, EXEMPLE

$$h : (x_1, x_2) \rightarrow (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\begin{aligned} h(x) \cdot h(y) &= (x_1^2, \sqrt{2}x_1x_2, x_2^2) \cdot (y_1^2, \sqrt{2}y_1y_2, y_2^2) \\ &= (x \cdot y)^2 \\ &= K(x, y) \end{aligned}$$





- En réalité, on ne connaît pas à priori le bon noyau. Il faut tester différents noyaux et utiliser le meilleur

REMERCIEMENTS

- Yann Coadou – CPPM
- Geoffrey Daniel – CEA-Saclay