

# Chapitre2: plan

1. Définition
2. Concepts de base du ML
3. Problèmes résolus par le ML  
(Classification, Régression et clustering)
4. Types d'algorithmes de ML ( Supervisés, non Supervisés, semi-supervisés et Renforcement)
5. Etapes ( data cleansing, Feature selection and extraction, Model training, model deployment)

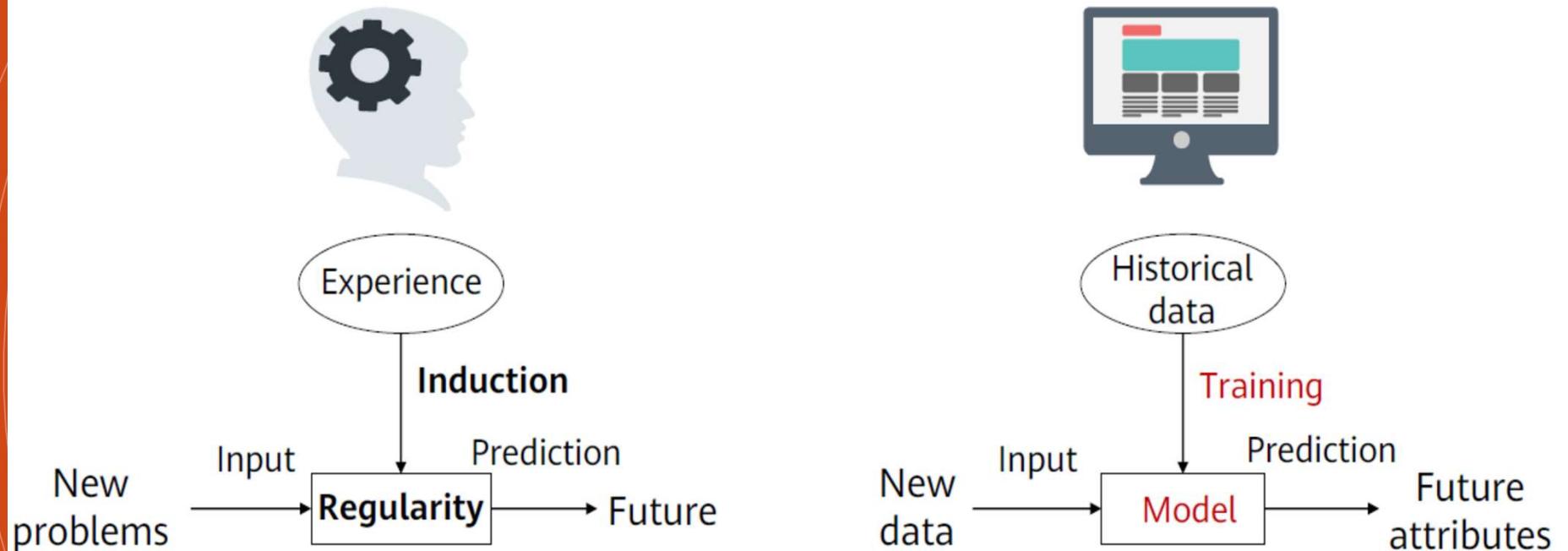
## Définition



*"Machine Learning is the science of getting computers to learn without being explicitly programmed."*  
Arthur Samuel, 1959.

« Donner à la machine la capacité d'apprendre sans la programmer d'une façon explicite ». Arthur Samuel, 1959.

## Définition



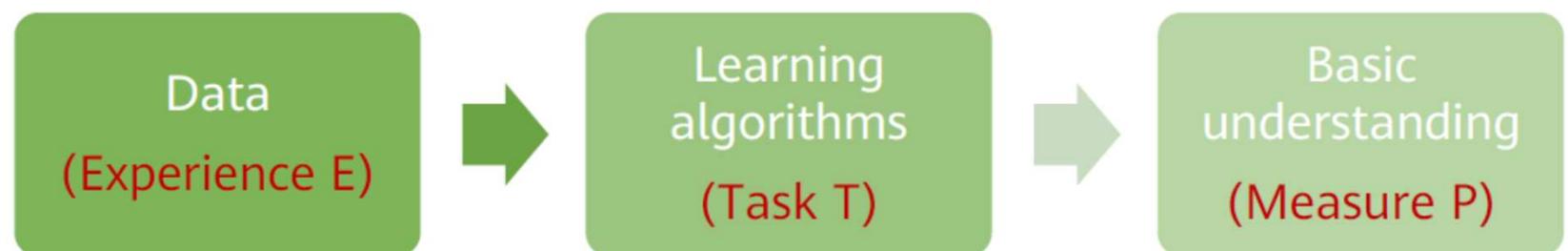
## Définition

« On dit qu'un programme apprend de l'expérience E en ce qui concerne une tâche T et une mesure de performance P, si sa performance sur T, mesurée par P, s'améliore avec l'expérience E ».

**Tom Mitchell (1997)**

### Trois caractéristiques:

- ✓ tâche T
- ✓ mesure de performance à améliorer P
- ✓ source d'expérience E



## Différence entre un algorithme classique et le machine learning (1)

### Algorithme classique

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ x & \longrightarrow & f(x) \end{array}$$

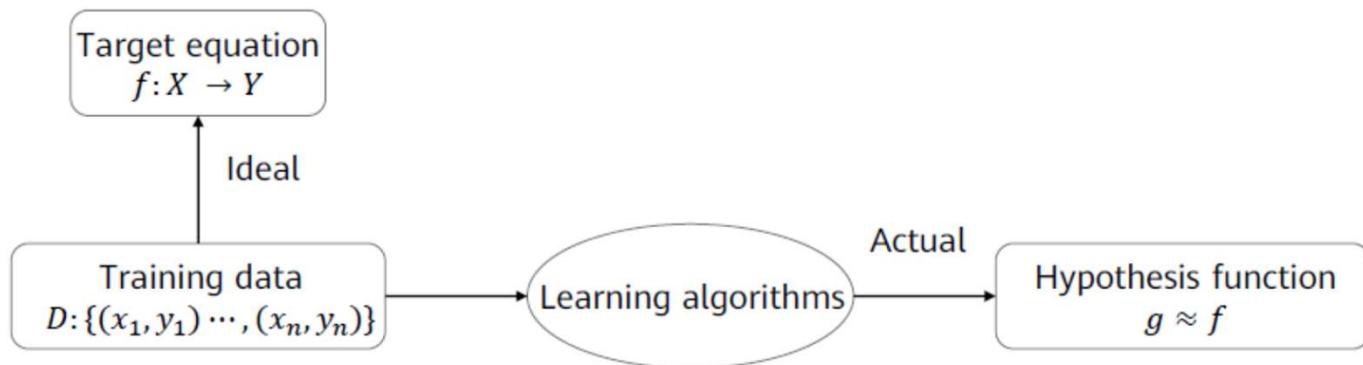
- ✓ On fournit à la machine **les règles de calcul** ( la fonction  $f$  ) pour trouver  $y = f(x)$

### Machine Learning

$$X \xrightarrow{?} Y$$

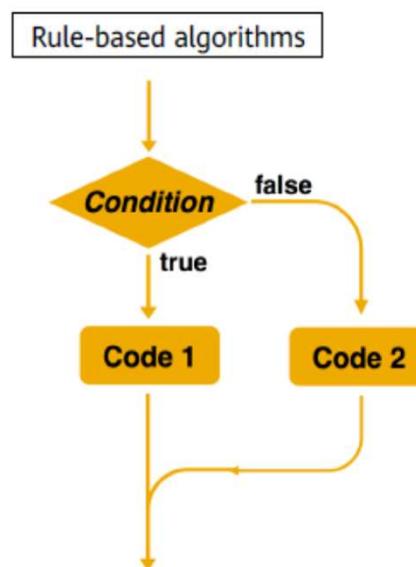
- ✓ On fournit à la machine **les données**  $X, Y$
- ✓ C'est à la machine de trouver **l'approximation de la fonction  $f$**
- ✓ La machine trouve le **modèle  $g$**  qui relie  $X$  à  $Y$

## Différence entre un algorithme classique et le machine learning (2)

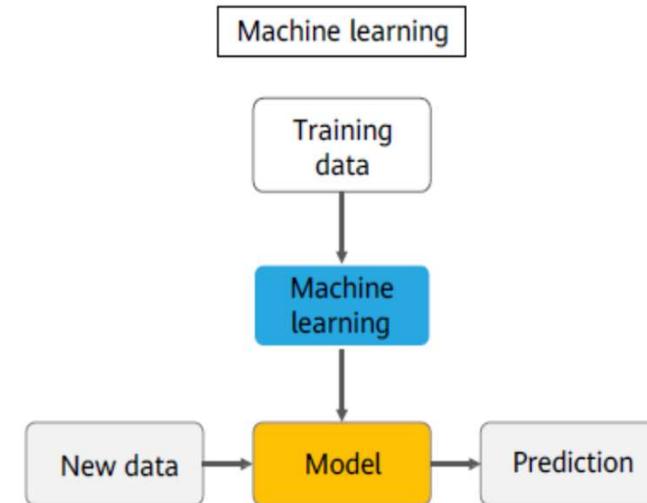


- Target function  $f$  is unknown. Learning algorithms cannot obtain a perfect function  $f$ .
- Assume that hypothesis function  $g$  **approximates** function  $f$ , but may be different from function  $f$ .

## Différence entre un algorithme classique et la machine learning (3)

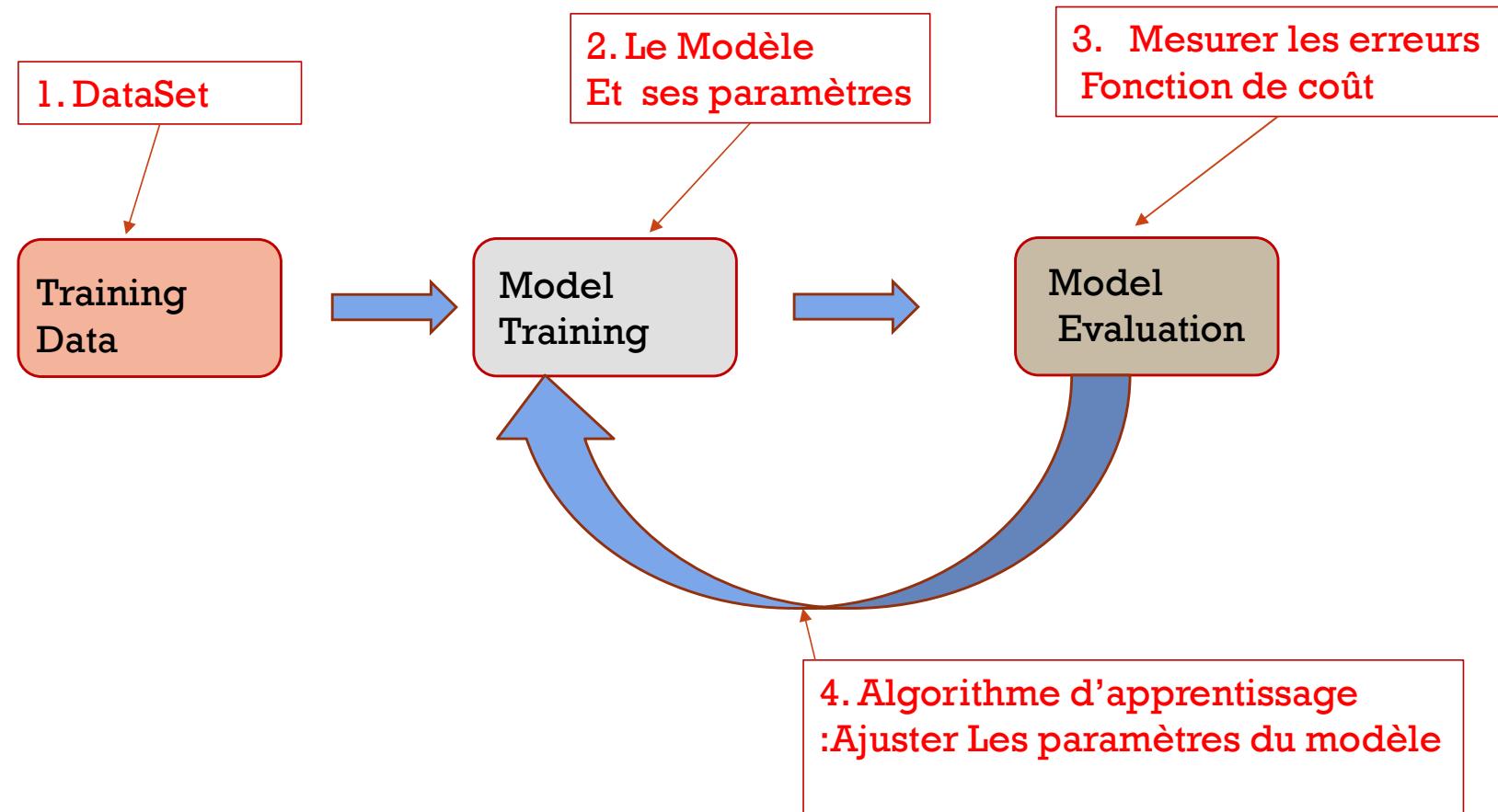


- Explicit programming is used to solve problems.
- Rules can be manually specified.



- Samples are used for training.
- The decision-making rules are complex or difficult to describe.
- Rules are automatically learned by machines.

## Machine Learning: 4 concepts de base



## ML Concept de base - DataSet

- ✓ **DataSet:** Une collection de données enregistrées dans un tableau,  
Chaque ligne d'enregistrement s'appelle a **Sample**, les attributs  
s'appellent **Features**,
- ✓ Un **DataSet** est subdivisé en deux parties:
  - ✓ **Training Set:** les données qui sont utilisées dans la phase d'apprentissage  
(Training Process),
  - ✓ **Test Set:** Les données qui sont utilisées dans la phase de l'évaluation du  
modèle (Evaluation Process)

## Basic ML Concepts- DataSet Example

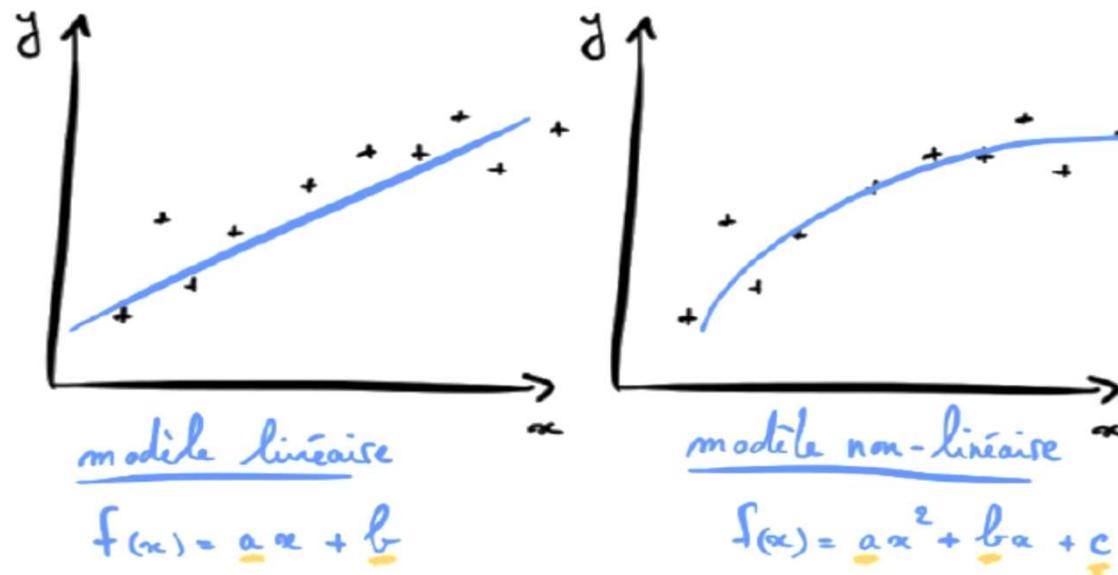
Samples	Target <i>y</i>	<i>x</i> <sub>1</sub>	Features <i>x</i> <sub>2</sub> <i>x</i> <sub>3</sub>	
	Prix	Surface m <sup>2</sup>	N chambres	Qualité
n1	€ 313,000.00	124	3	1.5
n2	€ 2,384,000.00	339	5	2.5
.	€ 342,000.00	179	3	2
	€ 420,000.00	186	3	2.25
	€ 550,000.00	180	4	2.5
N	€ 490,000.00	82	2	1
	€ 335,000.00	125	2	2

Training Set

Test Set

## ML Concept de base -Modèle

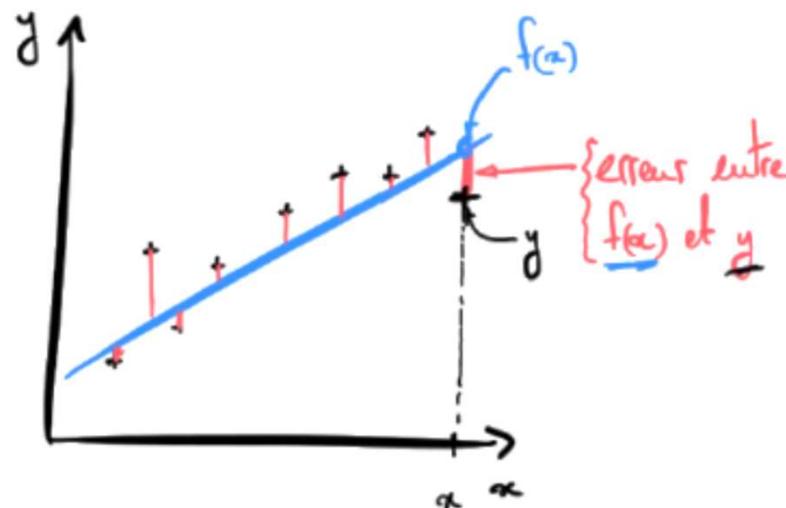
- A partir d'un Dataset, on crée un modèle, qui n'est autre qu'une fonction mathématique. Les coefficients de cette fonction sont les paramètres du modèle.



$a, b, c$  dans l'exemple représentent les paramètres du modèle à ajuster itérativement par l'algorithme d'apprentissage.

## ML Concept de base -Fonction coût

Un modèle appliqué sur un dataset retourne des **erreurs**. L'ensemble de ces erreurs est appelé la **Fonction Coût (cost function)** (le plus souvent Il s'agit de la moyenne quadratique des erreurs).



## ML Concept de base -Training Algorithm

L'algorithme d'apprentissage consiste à chercher **les paramètres du modèle** qui **minimisent** la Fonction Coût.

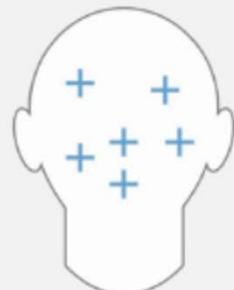
L'algorithme d'apprentissage le plus courant est l'algorithme de **Gradient Descent**

## Quand utiliser une ML? (1)

La solution d'un problème est complexe, pas de règles claires,

Le problème intègre un grand volume de données sans une distribution claire,

*Rules are complex or cannot be described, such as facial recognition and voice recognition.*



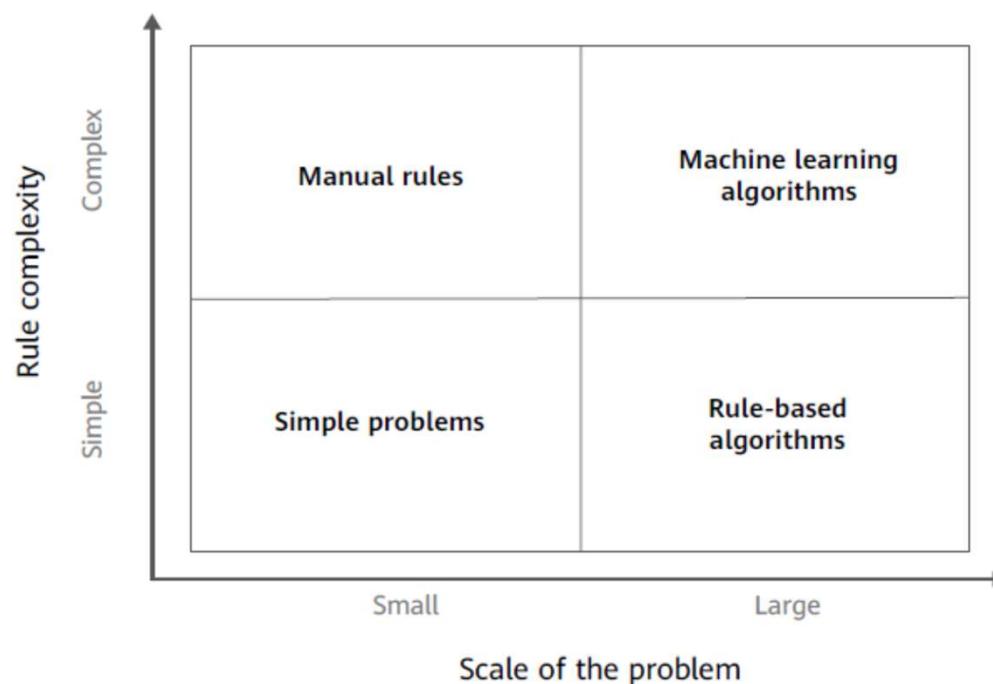
*Task rules change over time. For example, in the part-of-speech tagging task, new words or meanings are generated at any time.*



*Data distribution changes over time, requiring constant readaptation of programs, such as predicting the trend of commodity sales.*



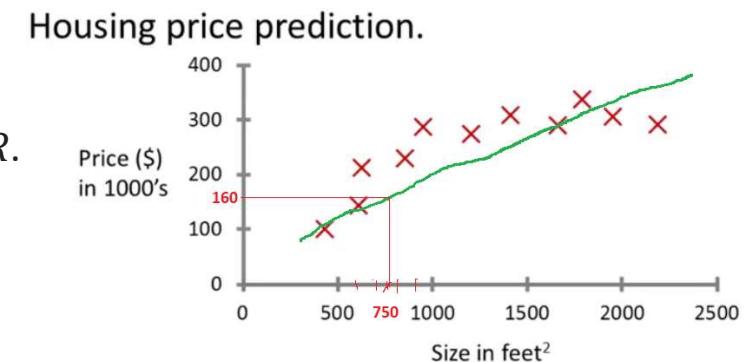
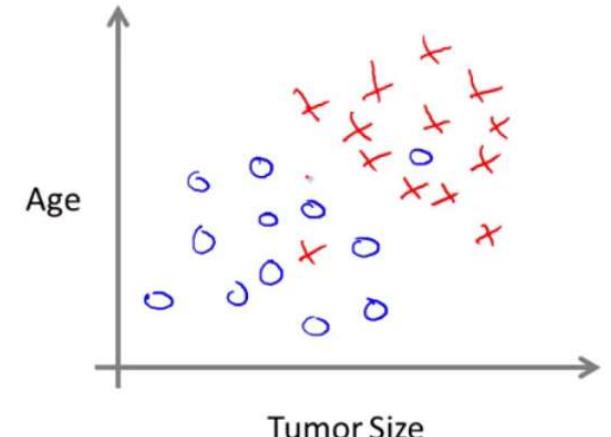
## Quand utiliser une ML? (2)



# Problèmes résolus par le ML

37

- ✓ **Classification:** Un programme informatique doit spécifier à laquelle des  $k$  catégories une entrée appartient.  
Pour accomplir cette tâche, les algorithmes d'apprentissage génèrent généralement une fonction  $f:R^n \rightarrow \{1, 2, \dots, k\}$ .
- ✓ **Regression:** un programme informatique prédit la sortie pour l'entrée donnée.  
les algorithmes génèrent généralement une fonction  $f:R^n \rightarrow R$ .
- ✓ **Clustering:** une grande quantité de données d'un ensemble de données non étiqueté est divisée en plusieurs catégories selon à la similitude interne des données. Les données d'une même catégorie sont plus similaires que celles de différentes catégories.



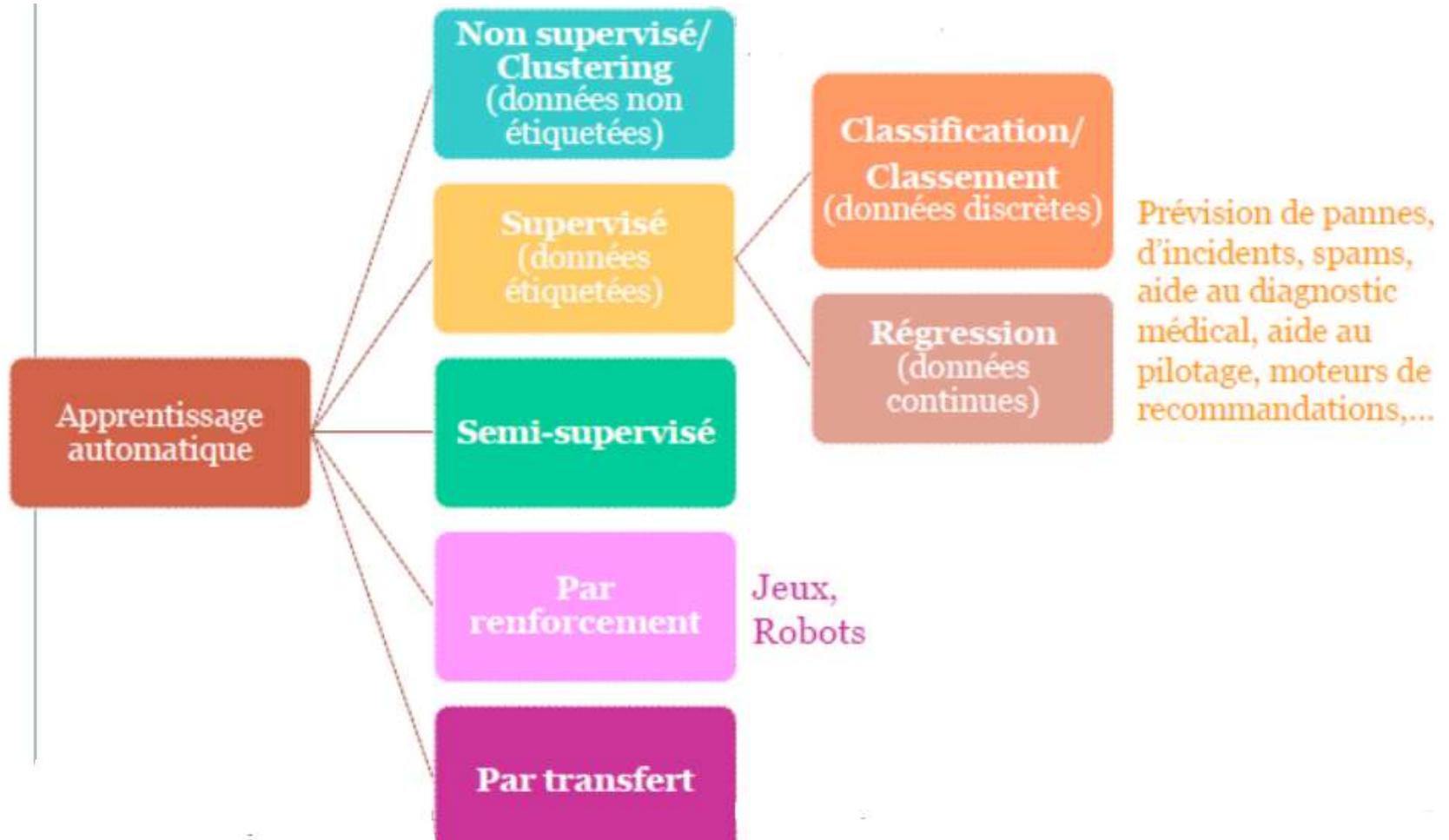
# Les types de la ML

38

- ✓ **Apprentissage supervisé** : obtenir un modèle optimal avec les performances requises grâce à l'apprentissage basé sur des échantillons labellisés de catégories connues.
- ✓ **Apprentissage non supervisé** : pour les échantillons non labellisés, les algorithmes d'apprentissage modélisent directement les ensembles de données d'entrée. Le regroupement en classe se fait selon le degré de similarité entre les données,
- ✓ **Apprentissage semi-supervisé** : un modèle d'apprentissage automatique qui utilise automatiquement une grande quantité de données non étiquetées pour aider à l'apprentissage directement d'une petite quantité de données étiquetées.
- ✓ **Apprentissage par renforcement** : C'est un domaine d'apprentissage machine concerné comme les agents cherchaient à agir dans un environnement pour maximiser une récompense cumulative. La différence entre apprentissage par renforcement et apprentissage supervisé est le signal enseignant.
- ✓ **Apprentissage par transfert** : consiste à compléter l'apprentissage d'un modèle de machine learning, préalablement entraîné à résoudre une tâche donnée, en vue de lui permettre de résoudre une tâche similaire.

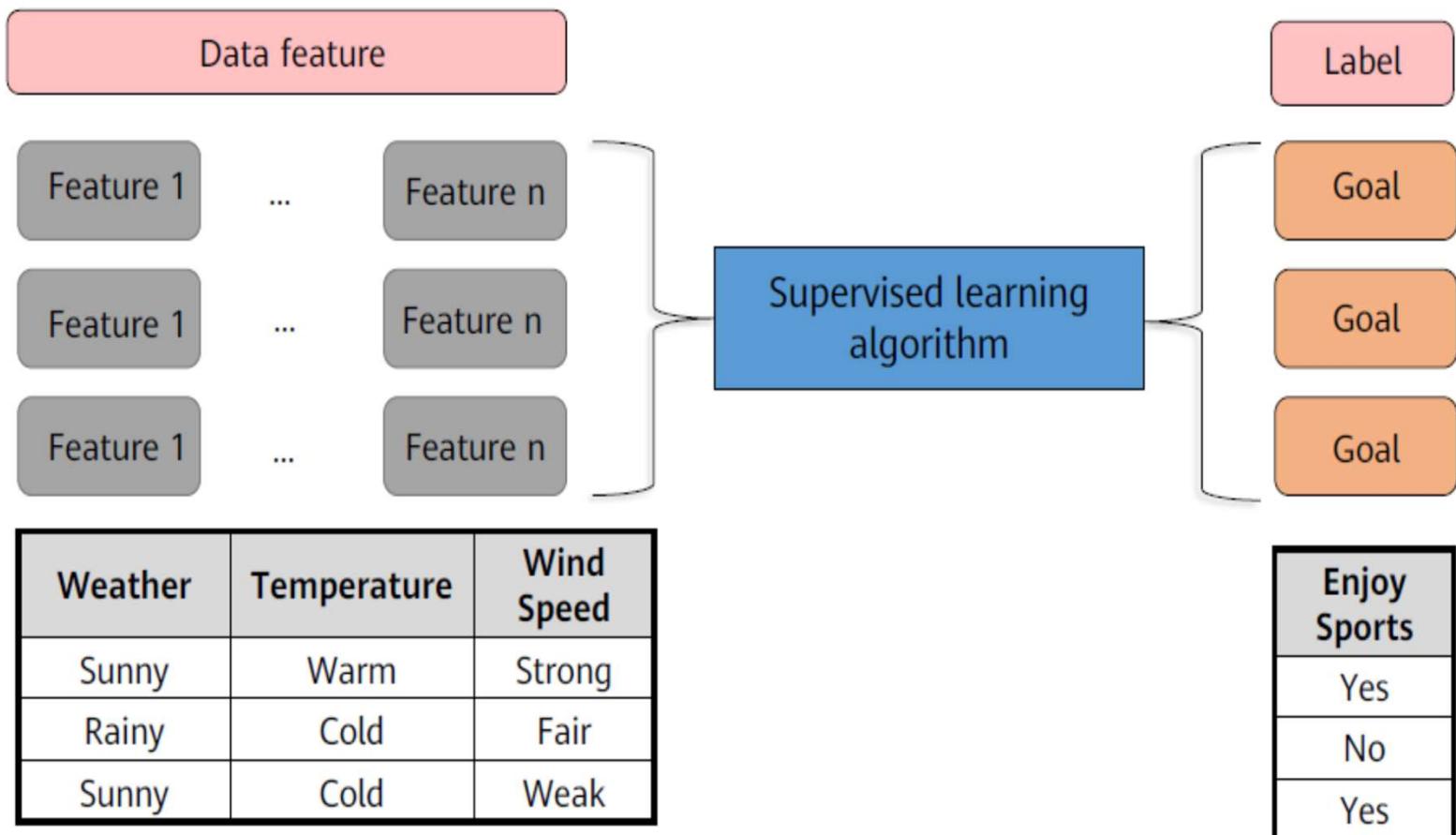
# Les types de la ML

39



# Supervised Learning- Classification

40



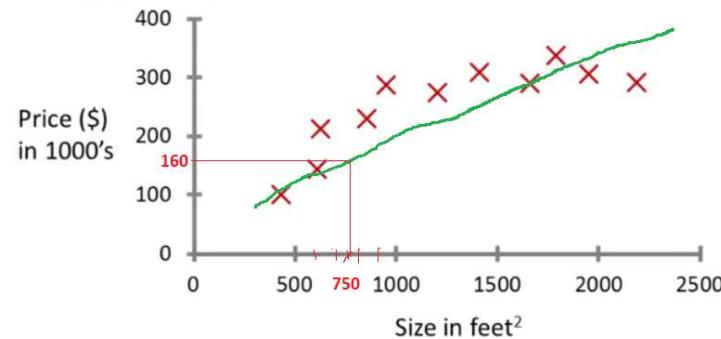
# Supervised Learning- Regression

41

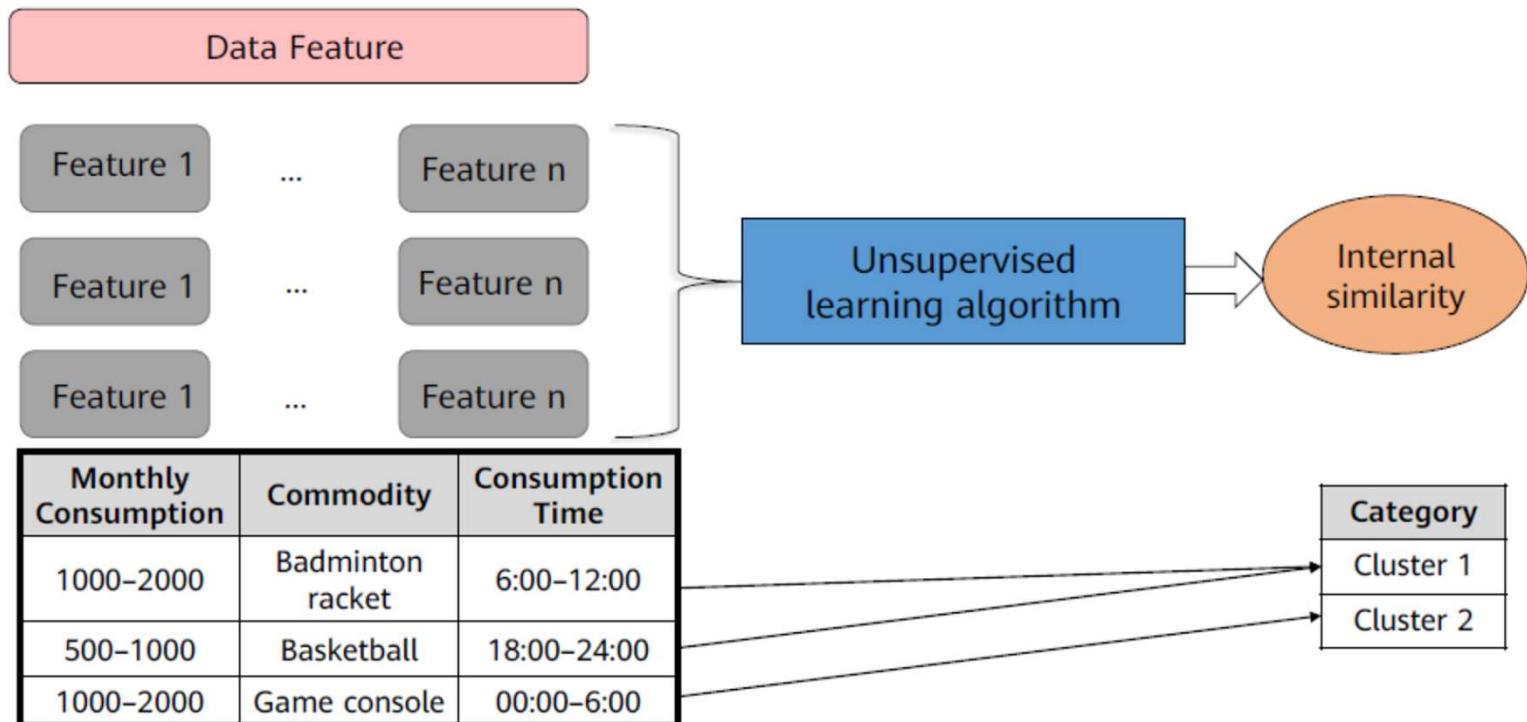
Regression: reflects the features of attribute values of samples in a sample dataset. The dependency between attribute values is discovered by expressing the relationship of sample mapping through functions.

- How much will I benefit from the stock next week?
- What's the temperature on Tuesday?

Housing price prediction.

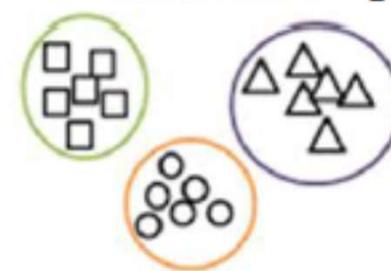


# UnSupervised Learning



# UnSupervised Learning

- Les données d'apprentissage n'incluent pas les résultats souhaités.
- Le système ne dispose que d'exemples, mais non d'étiquettes, et le nombre de classes et leur nature n'ont pas été prédéterminées.
- Aucun expert n'est requis.
- Le système doit découvrir par lui-même la **structure** plus ou moins cachée des données.



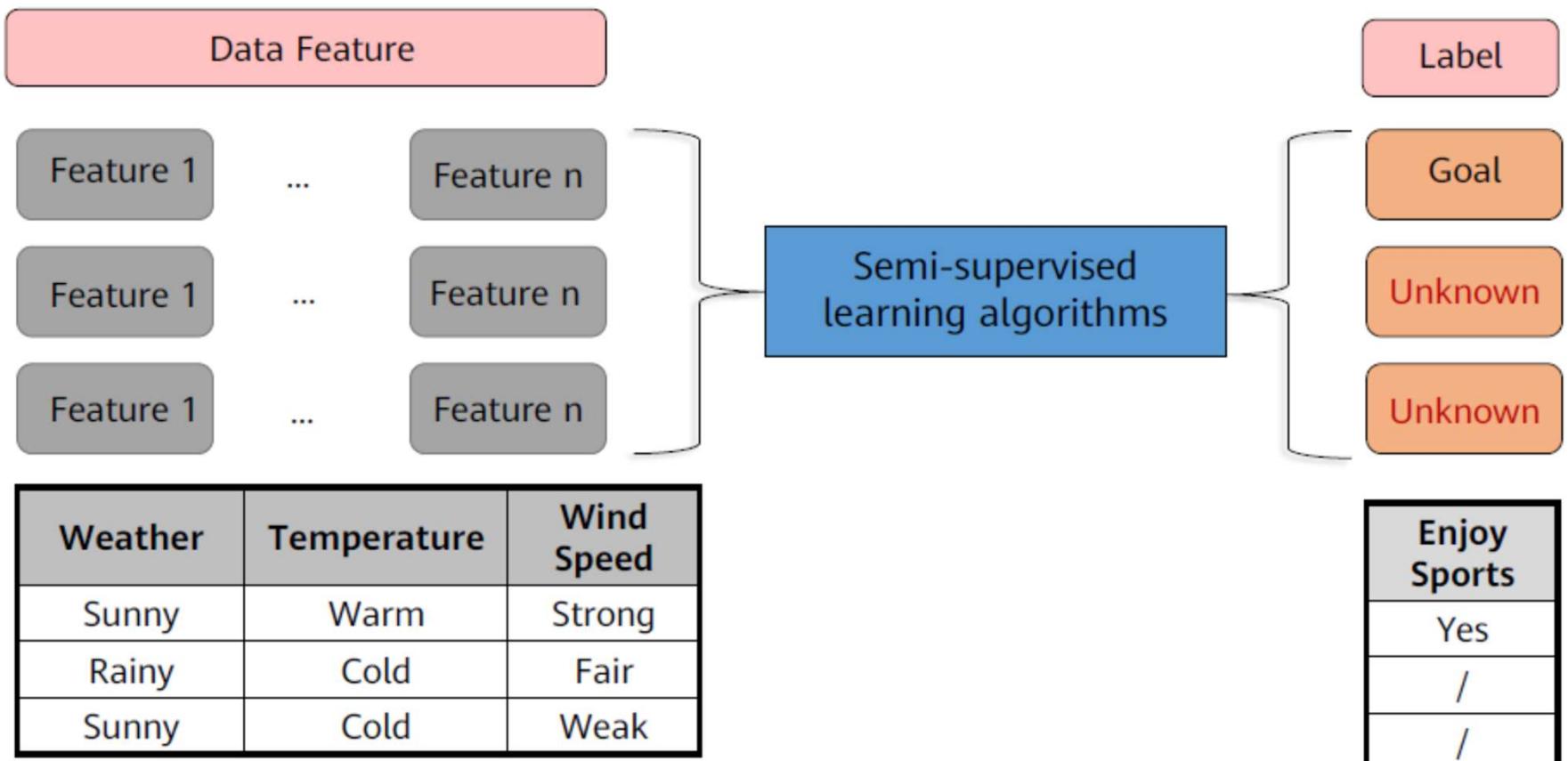
- Il doit cibler les données selon leurs attributs disponibles, pour les classer en **groupes homogènes** d'exemples.
- La **similarité** est généralement calculée selon une fonction de distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer ou déduire du **sens** pour chaque groupe.

# Semi Supervised Learning

Les données d'apprentissage incluent quelques résultats souhaités.

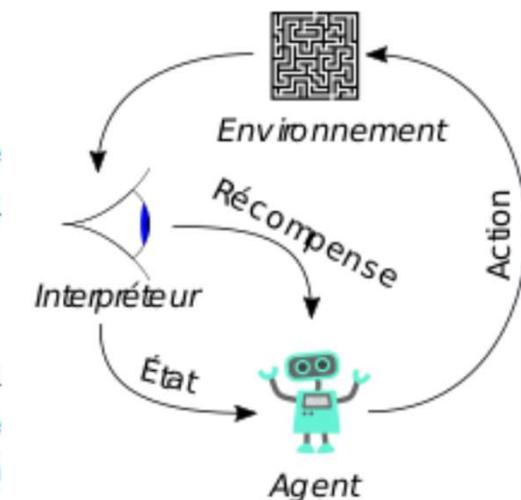
- Il est mis en œuvre quand des données (ou «étiquettes») manquent.
- Le modèle doit utiliser des exemples étiquetés pouvant néanmoins renseigner.
- Ex.: En médecine, il peut constituer une aide au diagnostic.

# Semi Supervised Learning

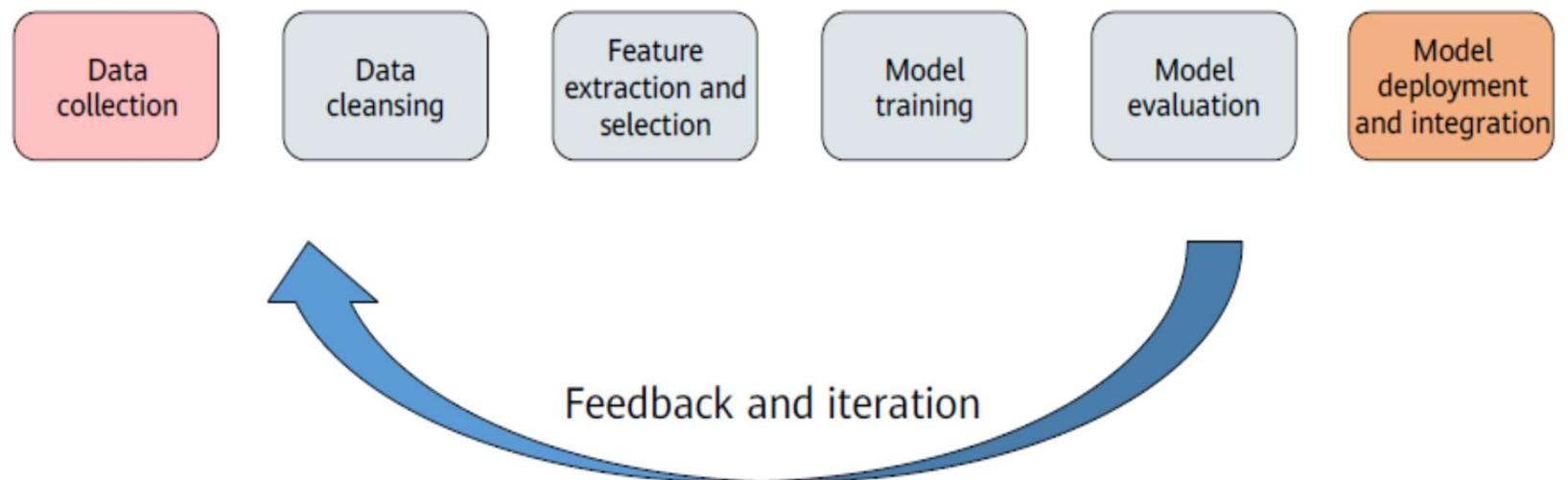


## Reinforcement Learning (Apprentissage par renforcement)

- Il consiste, pour un agent autonome (robot), à apprendre les **actions** à prendre, à partir **d'expériences**, de façon à optimiser une récompense quantitative au cours du temps.
- L'agent est plongé au sein d'un environnement, et prend ses décisions en fonction de son état courant.
- En retour, l'environnement procure à l'agent une récompense (signal), qui peut être positive ou négative.
- L'agent cherche, au travers d'expériences itérées, un comportement décisionnel (*stratégie* ou *politique*: une fonction associant à l'état courant l'action à exécuter) optimal, en ce sens qu'il maximise la somme des récompenses au cours du temps.

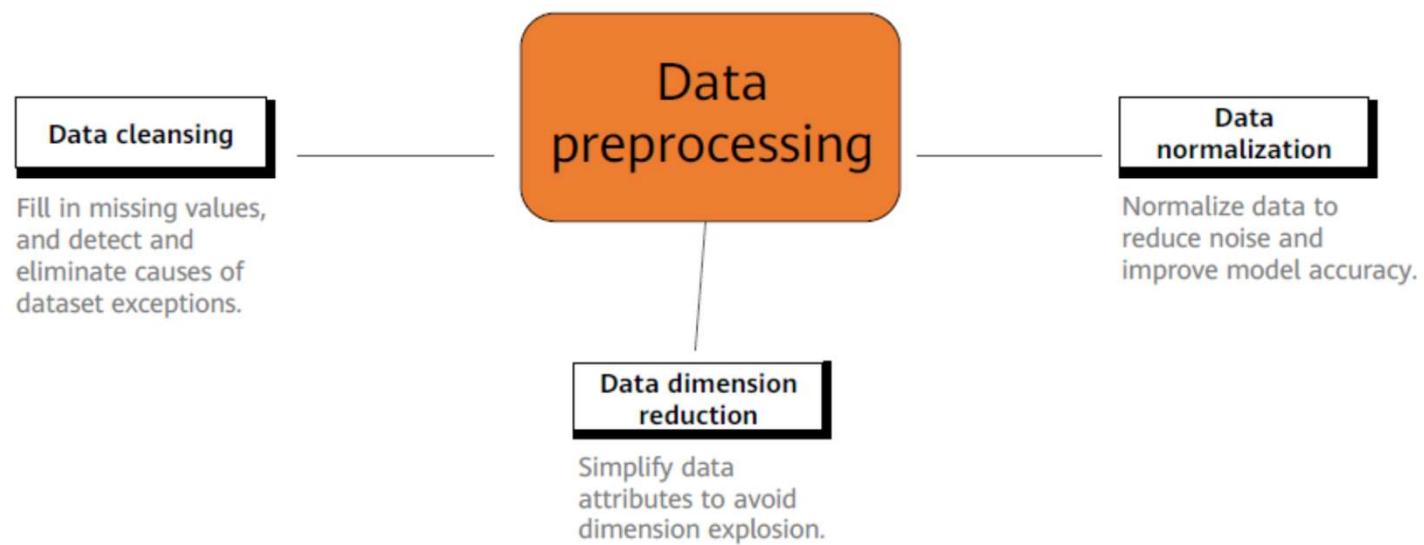


## Etapes de ML



## Etape1: Prétraitement des données

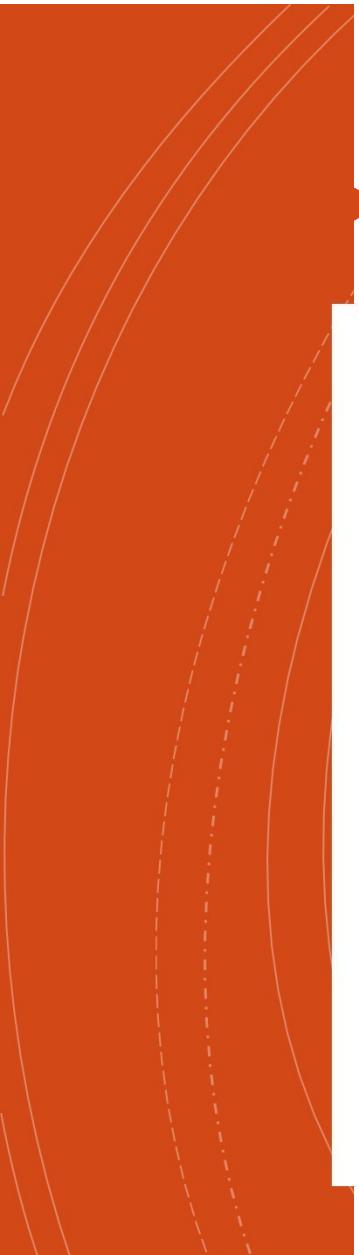
- Une étape cruciale dans le processus: **Sans de bonnes données, on ne peut pas avoir de bons modèles,**



## Etape1- Qu'est ce que la Dirty Data?

- En général, les données réelles peuvent présenter des problèmes de qualité.
  - Incomplétude : contient des valeurs manquantes ou des données sans attributs
  - Bruit : contient des enregistrements ou des exceptions incorrects.
  - Incohérence : contient des enregistrements incohérents.

## Exemple: Dirty Data



#	Id	Name	Birthday	Gender	IsTe ach er	#Stu dent s	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Annotations pointing to specific dirty data points:

- Invalid duplicate item: Points to the row where Id 555 appears twice.
- Missing value: Points to the empty cell in the City column for Iceland.
- Invalid value: Points to the invalid gender value 'A' in row 5.
- Value that should be in another column: Points to the country 'Italy' listed under the City column for Italy.
- Incorrect format: Points to the date '1983-12-01' in row 6.
- Attribute dependency: Points to the row where the student count (23) is higher than the number of students (0).
- Misspelling: Points to the misspelled country name 'Ytali' in row 10.

# Conversion des données

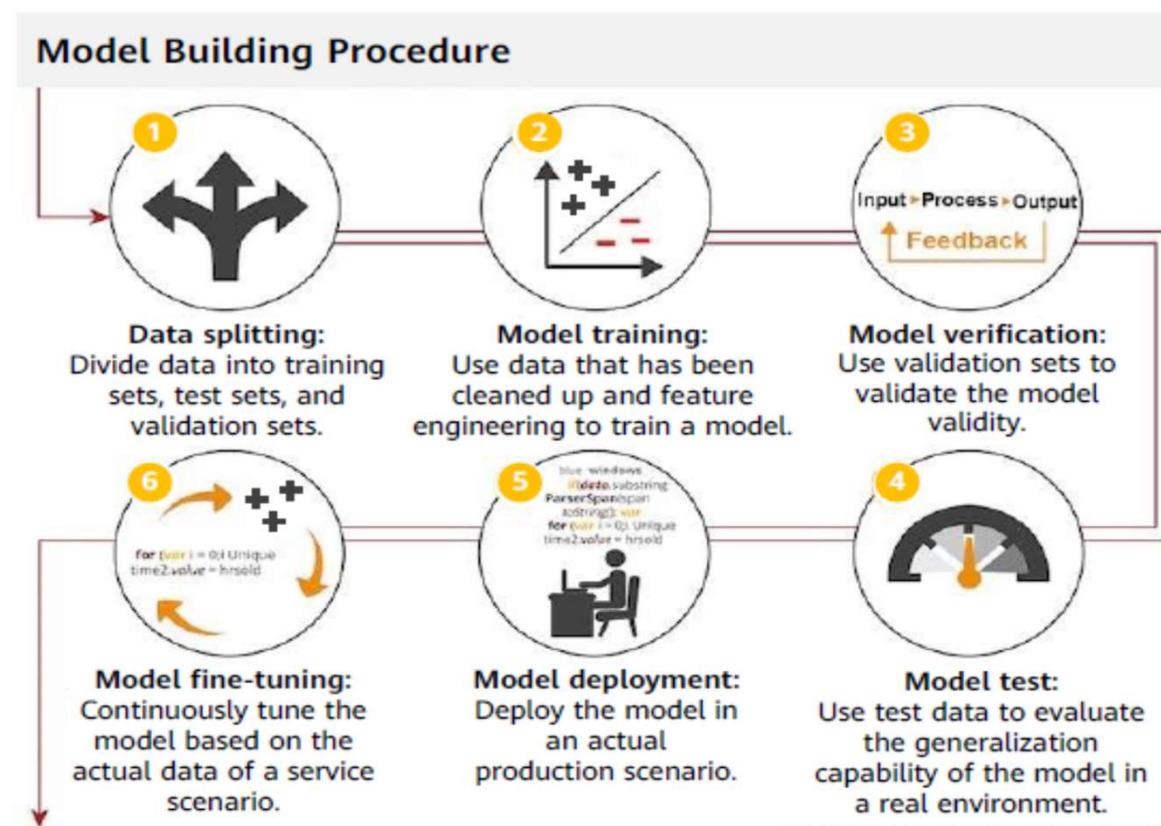
- Après avoir été prétraitées, les données doivent être converties en un **format de représentation** adapté au modèle d'apprentissage automatique. Les formats de conversion de données communs incluent les éléments suivants :
  - En ce qui concerne la classification, les données de catégorie sont codées dans une représentation numérique correspondante.
  - Une valeur est convertie en une catégorie pour réduire la valeur des variables (segmentation par âge)
  - Dans le texte, **le mot est converti en un vecteur** (généralement en utilisant le modèle word2vec, le modèle BERT, etc.).
  - **Traiter les données d'image** (espace colorimétrique, niveaux de gris, changement géométrique, Haarfeature et amélioration de l'image)
- **Ingénierie des features:**
  - Normaliser les features pour garantir les mêmes plages de valeurs pour les variables d'entrée du même modèle.
  - Extension de features : combinez ou convertissez des variables existantes pour générer de nouvelles fonctionnalités, telles que la moyenne.

## Sélection des features

- Généralement, un ensemble de données comporte de nombreuses caractéristiques, dont certaines peuvent être redondantes ou sans rapport avec la valeur à estimer.



## Procédure totale



## Validation de modèle: Capacité de généralisation (Generalization Capability)

L'objectif de l'apprentissage machine est que le modèle obtenu après l'apprentissage soit performant sur de nouveaux échantillons de données, et pas seulement sur des échantillons utilisés pour l'apprentissage.

**Généralisation:** La capacité d'appliquer un modèle à de nouveaux exemples.

**Erreur d'Apprentissage (Training Error)** : erreur qu'on obtient lorsqu'on exécute le modèle sur les données d'apprentissage.

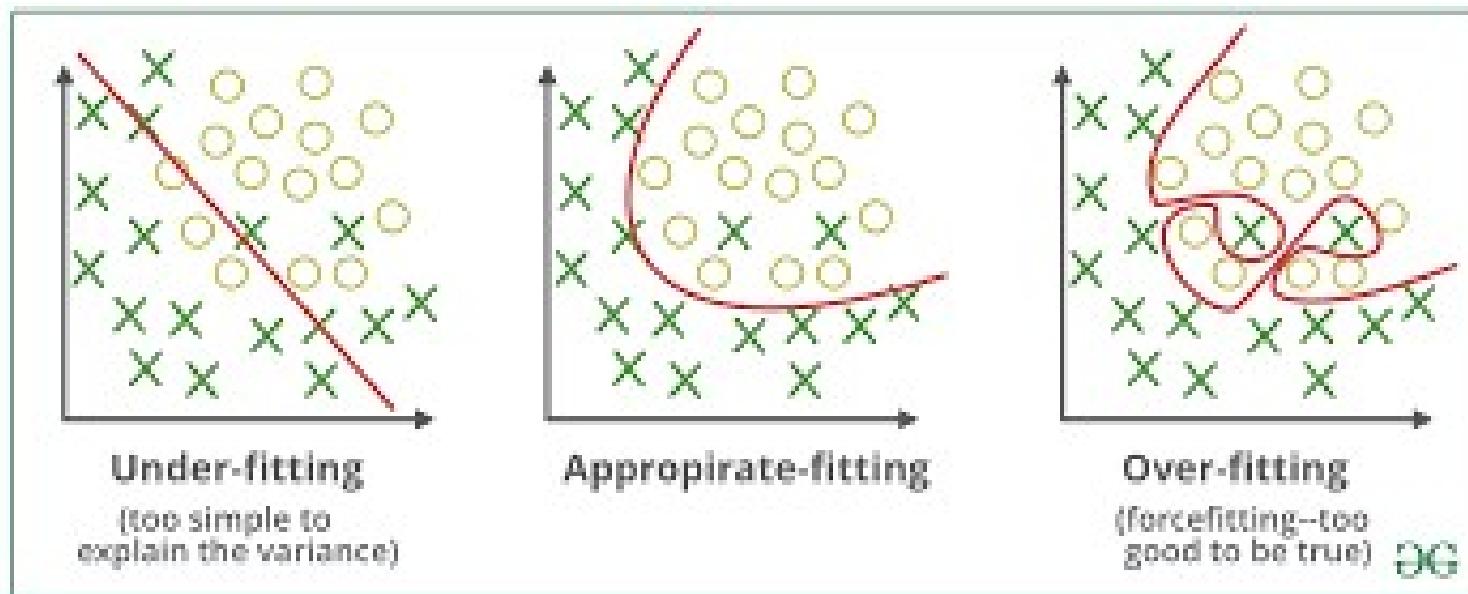
**Erreur de généralisation (Generalization Error)** : erreur qu'on obtient lorsqu'on exécute le modèle sur de nouveaux échantillons. Évidemment, nous préférerons un modèle avec une erreur de généralisation plus faible.

## Validation de modèle: **Capacité du modèle (Model complexity)**

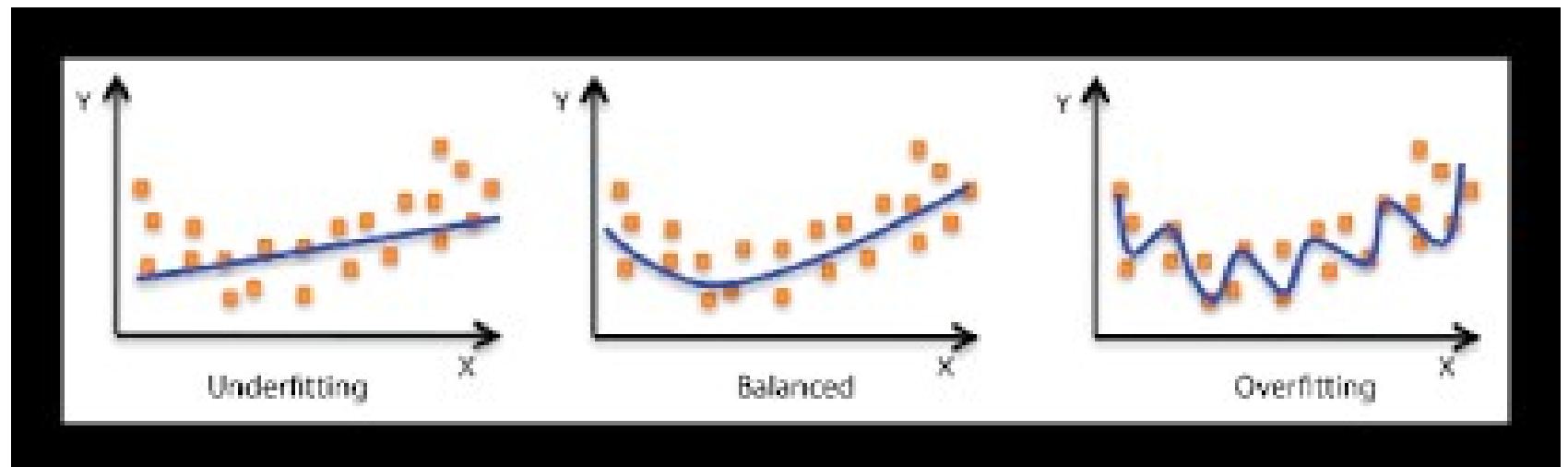
**Model complexity: La capacité du modèle à s'ajuster aux fonctions réelles,**

- ✓ Lorsque la capacité correspond à la complexité de la fonction réelle et à la quantité de données d'apprentissage fournies, l'algorithme est généralement **optimal**.
- ✓ Les modèles avec une capacité insuffisante ne peuvent pas résoudre des fonctions complexes et un **underfitting** peut se produire.
- ✓ Un modèle à haute capacité peut résoudre des tâches complexes, mais un surapprentissage **Overfitting** peut se produire si la capacité est supérieure à celle requise par une tâche.

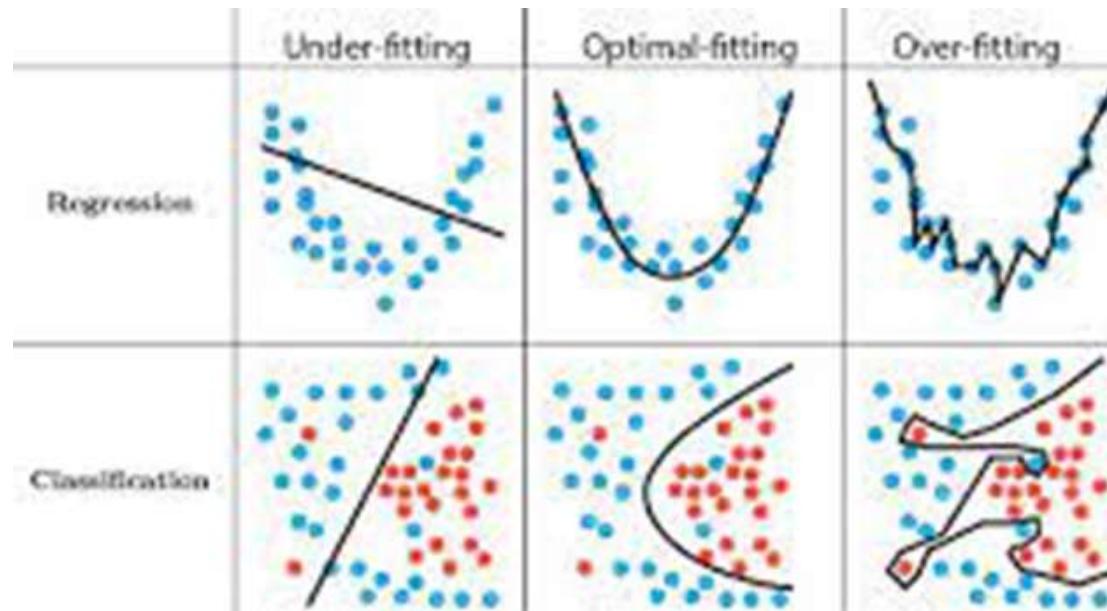
## Validation de modèle: Overfitting et Underfitting



## Validation de modèle: Overfitting et Underfitting

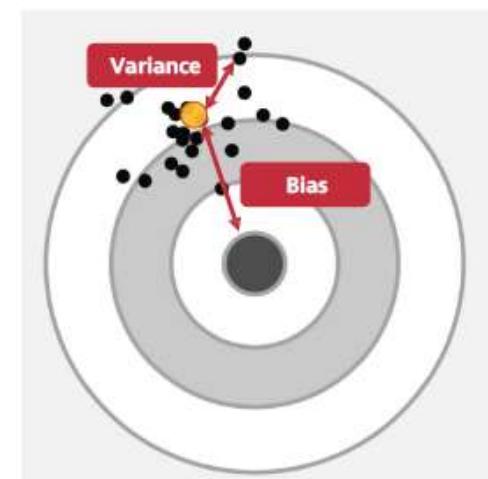


## Validation de modèle: Overfitting et Underfitting



## Validation de modèle: **Overfitting -Cause**

- Généralement, l'erreur de prédiction peut être divisée en deux types :
  - ✓ **Erreur causée par un "biais"**
  - ✓ **Erreur causée par « variance »**
- **Variance :**
  - ✓ Décalage du résultat de prédiction par rapport à la valeur moyenne
  - ✓ Erreur causée par la sensibilité du modèle aux petites fluctuations de l'ensemble d'apprentissage
- **Biais :**
  - ✓ Différence entre la valeur de prédiction attendue (ou moyenne) et la valeur correcte que nous essayons de prédire.



## Validation de modèle: **Variance & Biais TradeOff**

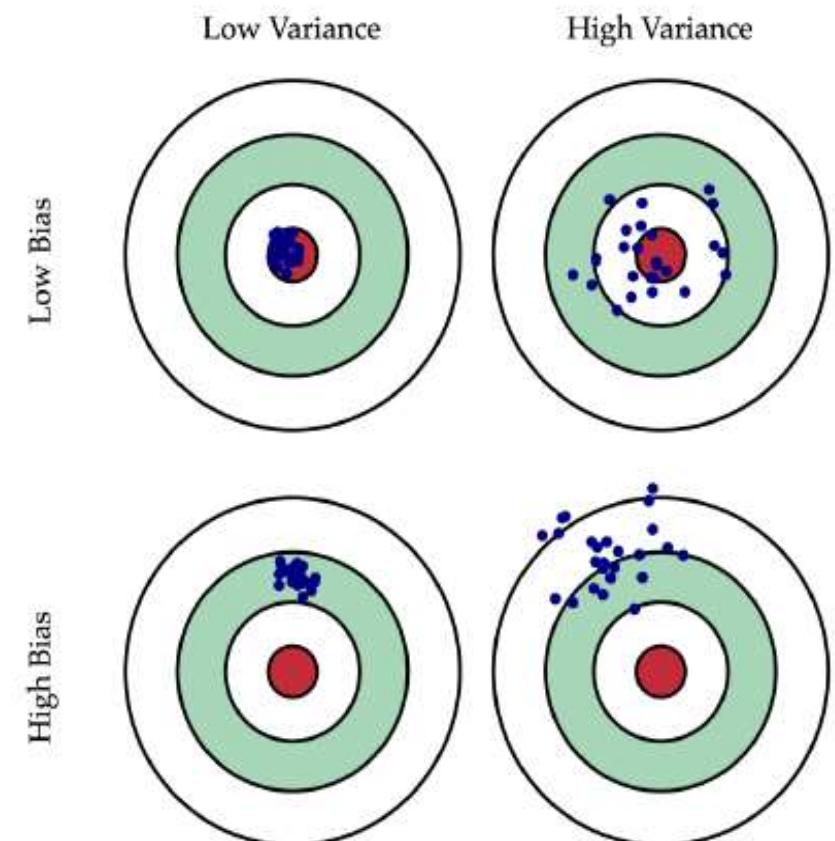
✓ Combinations entre variance et biais

Low bias & low variance → Good model

Low bias & high variance → Overfitting

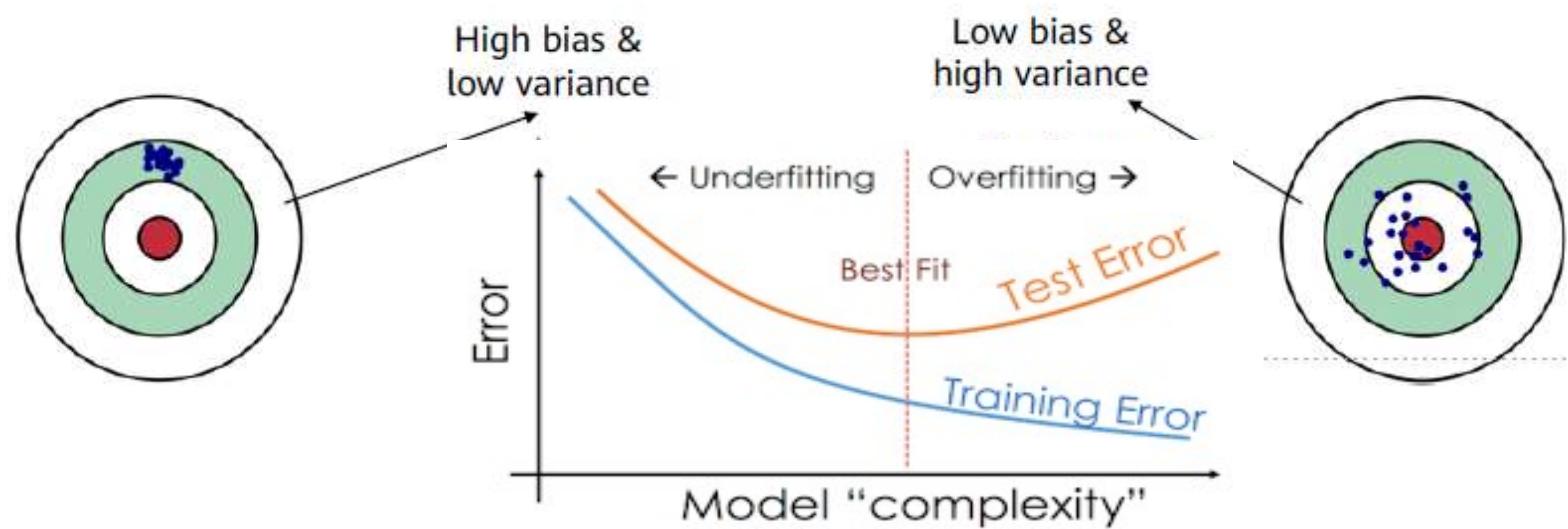
High bias & low variance → Underfitting

High bias & high variance → Poor model



## Validation de modèle: Variance & Biais TradeOff

- ✓ À mesure que la complexité du modèle augmente, l'erreur d'apprentissage (Training Error) diminue.
- ✓ Au fur et à mesure que la complexité du modèle augmente, l'erreur de test (Testing Error) diminue jusqu'à un certain point puis augmente dans le sens inverse, formant une courbe convexe.



## Regression Evaluation: **MSE & MAE**

Plus l'erreur absolue moyenne (MAE) est proche de 0, mieux le modèle est adapté aux données d'entraînement

**Mean Absolute Error (MAE)**  
Erreur Absolue Moyenne

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i|$$

Plus l'erreur quadratique moyenne (MSE) est proche de 0, mieux le modèle est adapté aux données d'entraînement

**Mean Square Error(MSE)**  
Erreur quadratique Moyenne

$$MSE = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

## Classification Evaluation: **Confusion Matrix**

### Termes et définitions:

- ✓ **P**: positifs, indiquant le nombre de cas réels positifs dans les données.
- ✓ **N** : négatifs, indiquant le nombre de cas réels négatifs dans les données.
- ✓ **TP** : True positives (vrai positif), indiquant le nombre de cas positifs correctement classés par le classifieur.
- ✓ **TN** : True Negatives (vrai négatif), indiquant le nombre de cas négatifs correctement classés par le classifieur.
- ✓ **FP** : faux positif, indiquant le nombre de cas positifs mal classés par le classificateur.
- ✓ **FN**: faux négatif, indiquant le nombre de cas négatifs mal classés par le classificateur.

		Estimated amount	yes	no	Total
Actual amount	yes	TP	FN	P	
	no	FP	TN	N	
Total		P'	N'	P + N	

Confusion matrix

## Classification Evaluation: Mesures d'évaluation

Measurement	Ratio
Accuracy and recognition rate	$\frac{TP + TN}{P + N}$
Error rate and misclassification rate	$\frac{FP + FN}{P + N}$
Sensitivity, true positive rate, and recall	$\frac{TP}{P}$
Specificity and true negative rate	$\frac{TN}{N}$
Precision	$\frac{TP}{TP + FP}$
$F_1$ , harmonic mean of the recall rate and precision	$\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
$F_\beta$ , where $\beta$ is a non-negative real number	$\frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$

## Classification Evaluation: Exemple

Nous avons entraîné un modèle d'apprentissage automatique pour identifier si l'objet d'une image est un chat.

Désormais, nous utilisons 200 images pour vérifier les performances du modèle.

Parmi les 200 images, les objets sur 170 images sont des chats, tandis que d'autres ne le sont pas.

$$\text{Precision: } P = \frac{TP}{TP+FP} = \frac{140}{140+20} = 87.5\%$$

$$\text{Recall: } R = \frac{TP}{P} = \frac{140}{170} = 82.4\%$$

$$\text{Accuracy: } ACC = \frac{TP+TN}{P+N} = \frac{140+10}{170+30} = 75\%$$

Actual amount \ Estimated amount	yes	no	Total
yes	140	30	170
no	20	10	30
Total	160	40	200

## Exercice pratique

Visualiser et traiter la base de données home prices in Melbourne, Australia.

Accéder à <https://www.kaggle.com/dansbecker/basic-data-exploration>