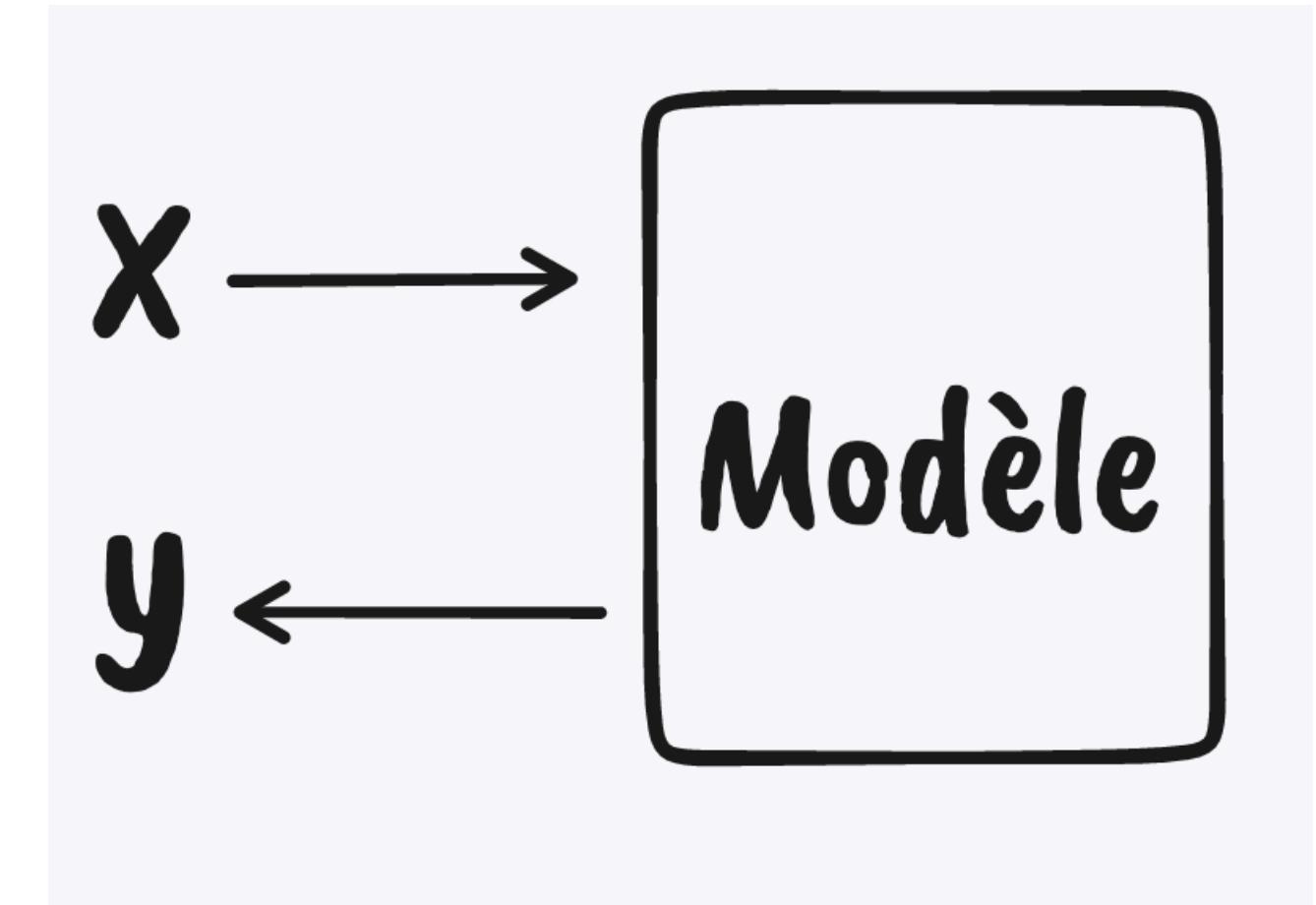
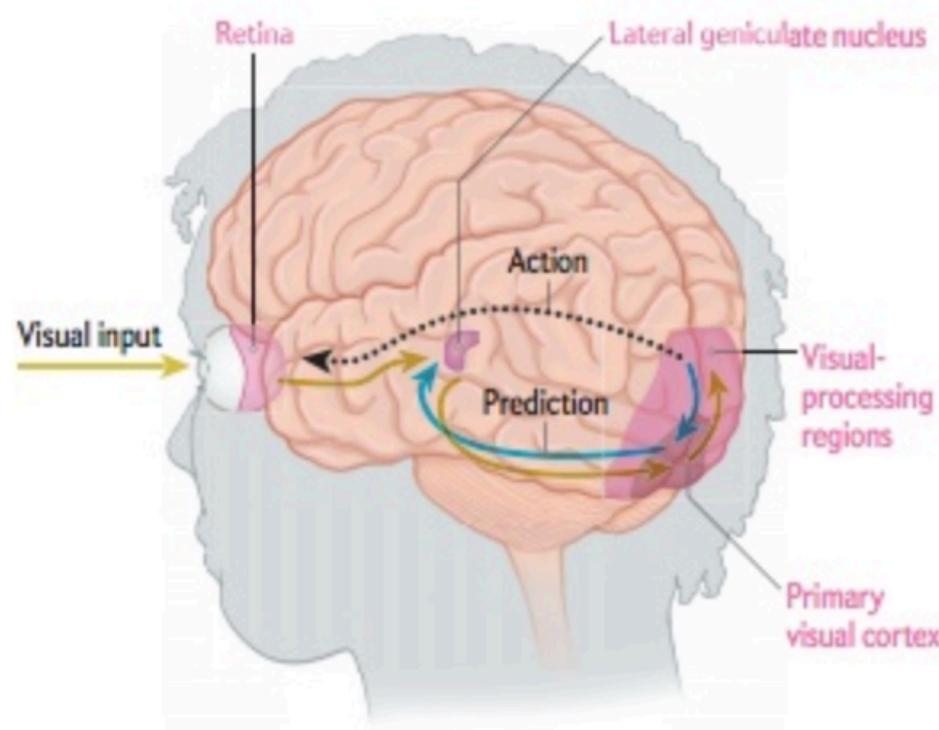


# **Comment les machines "apprennent" ?**

# Le ML c'est trouver un modèle



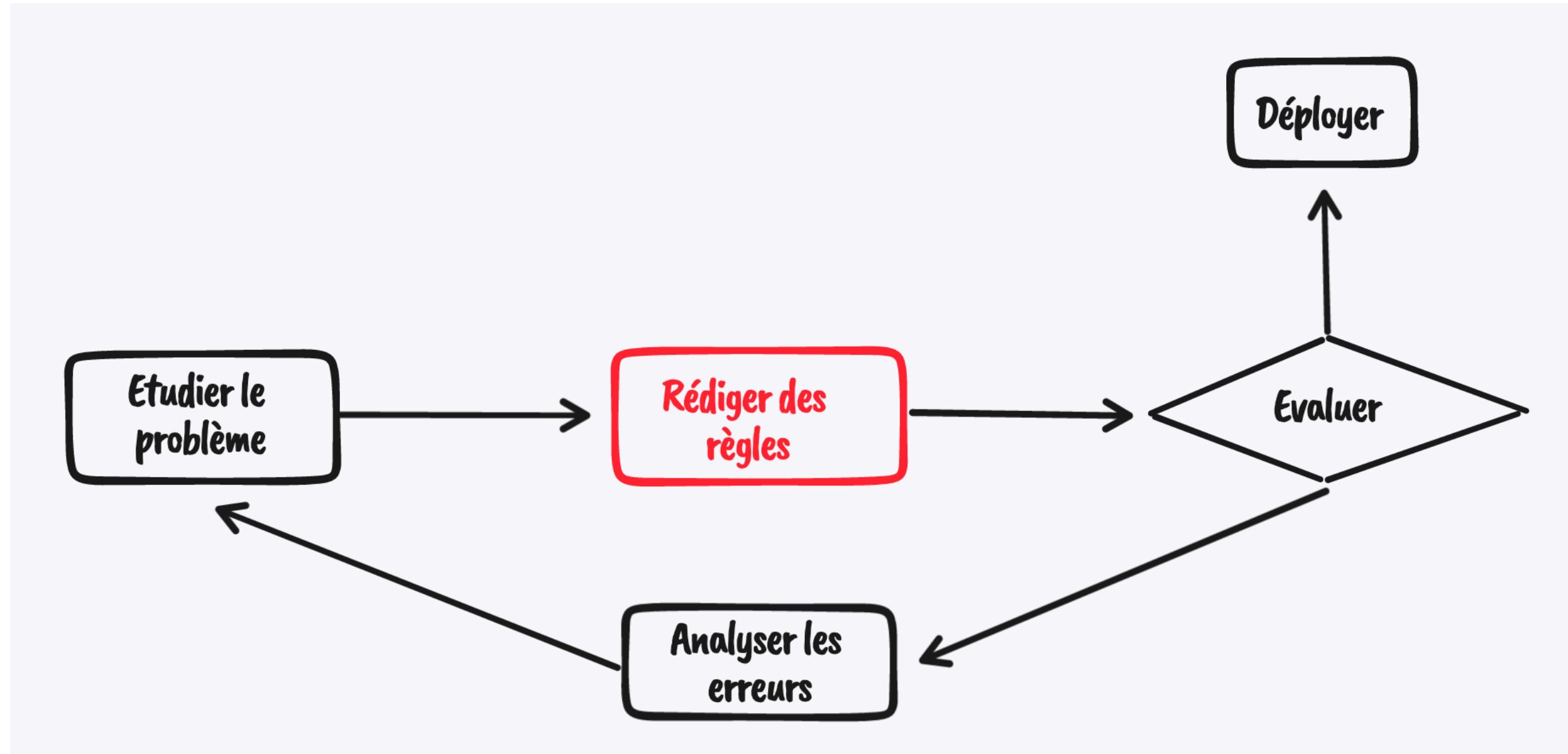
# C'est quoi un modèle ?



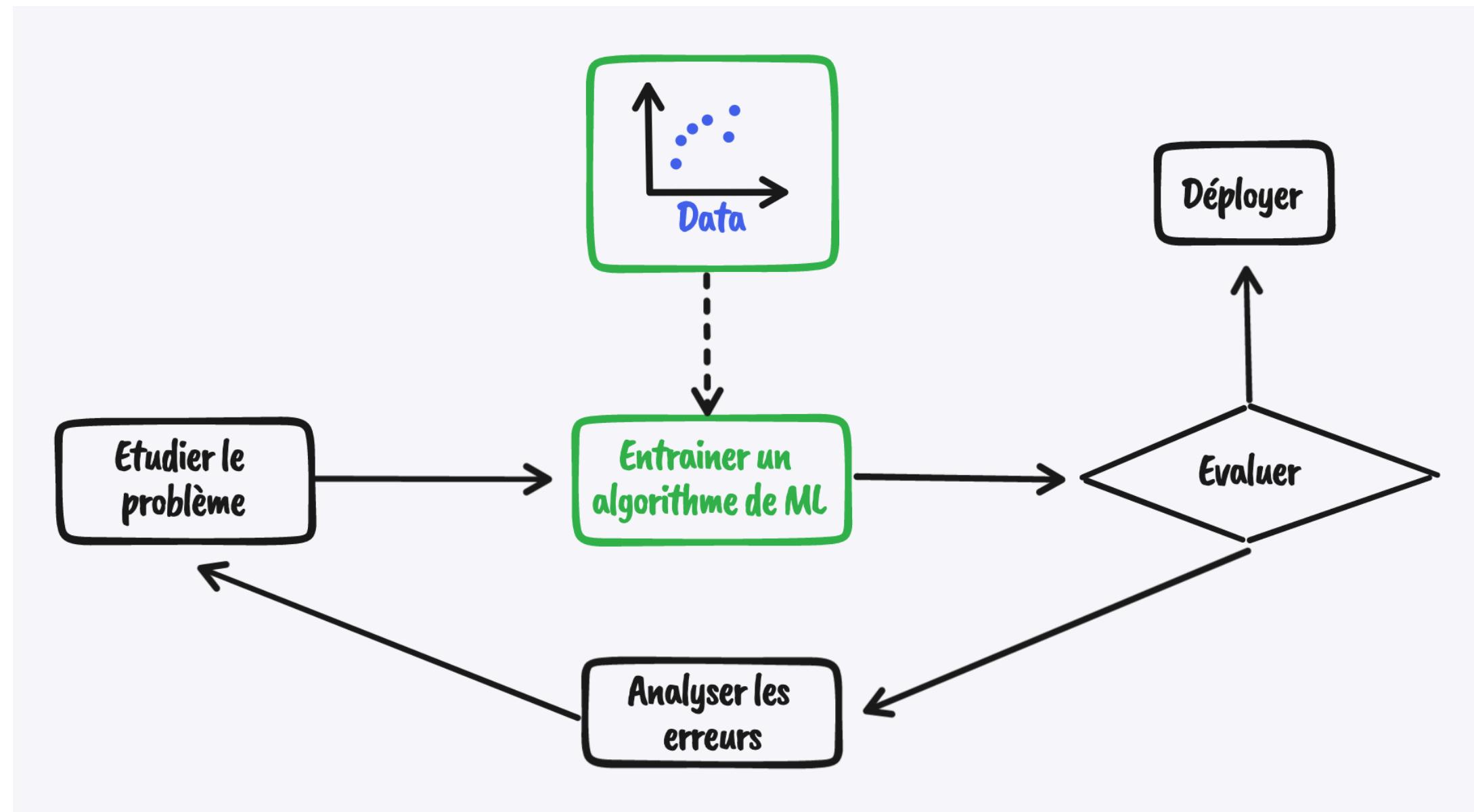
# Comment construire un modèle



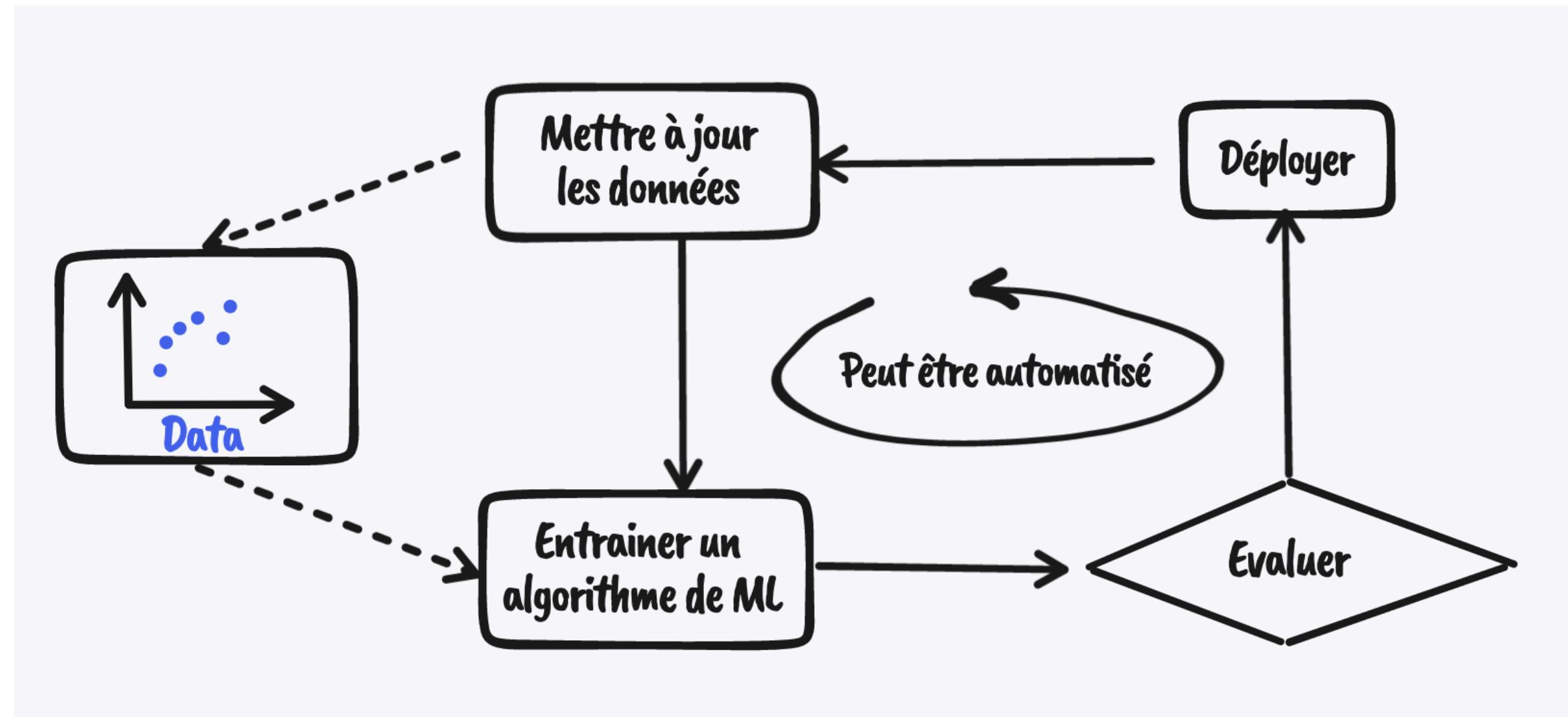
# Approche traditionnelle



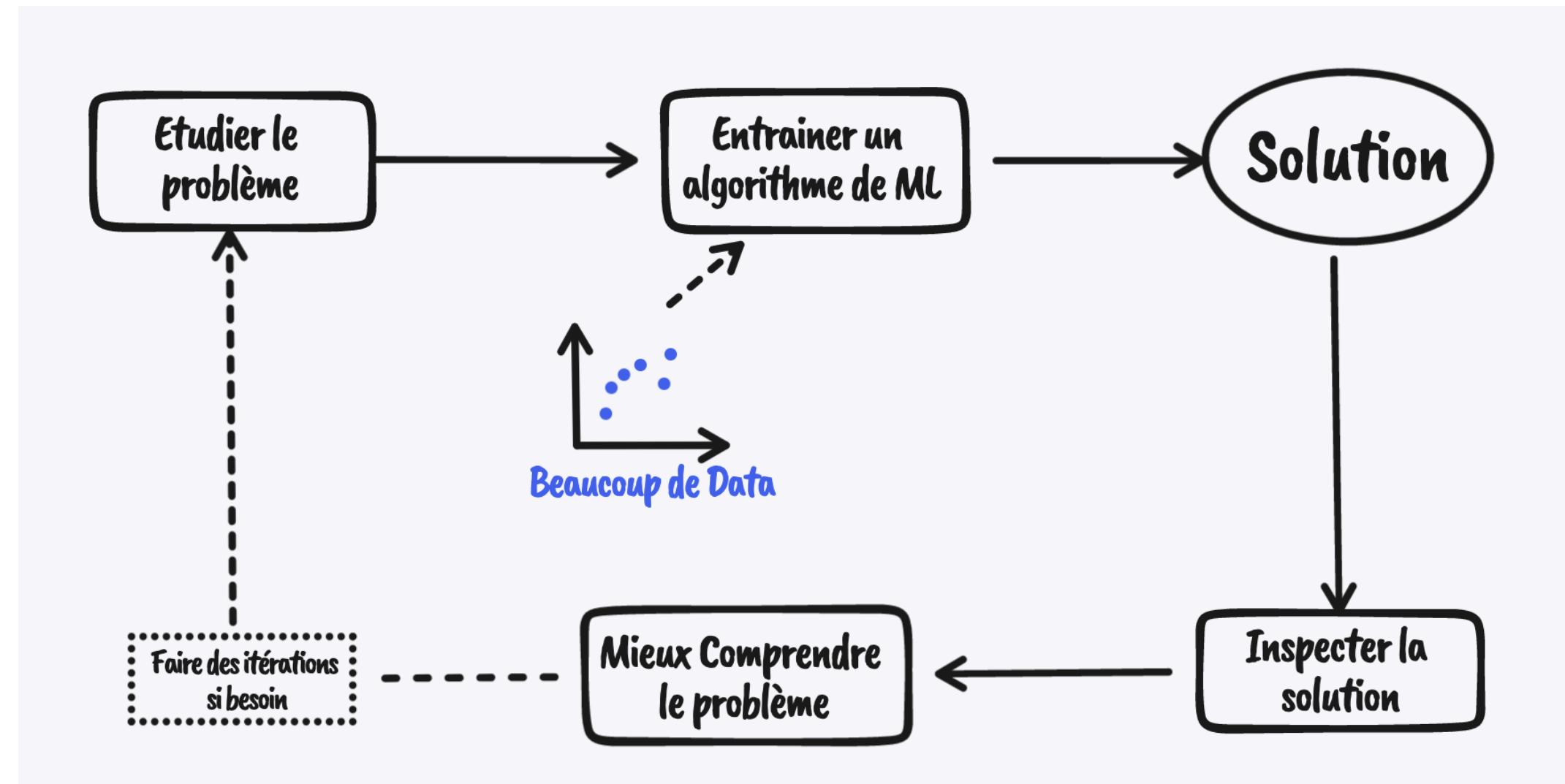
# Approche Machine Learning



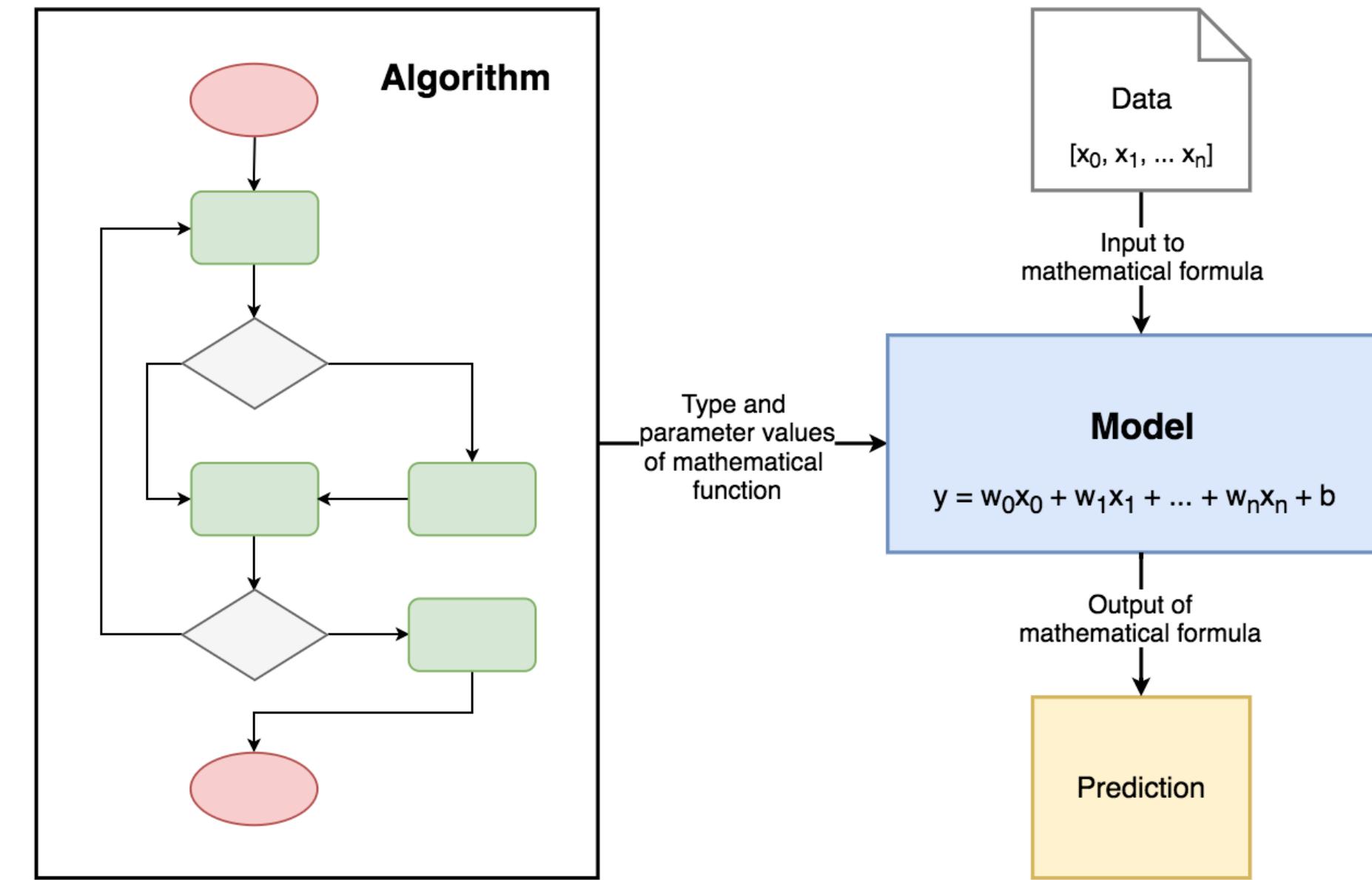
# Le Machine Learning permet de s'adapter au changements



# Le Machine Learning peut aider l'humain à apprendre

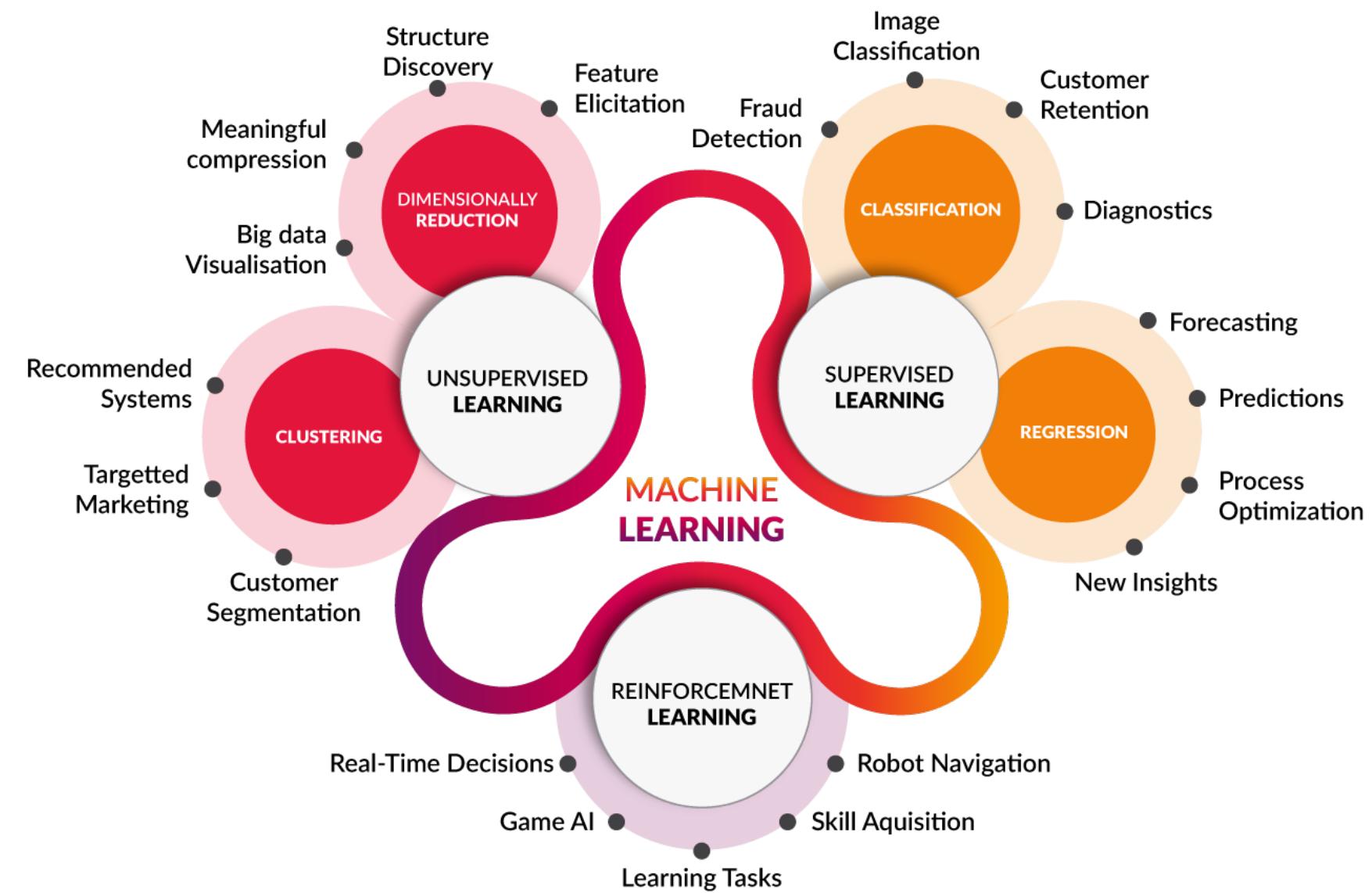


# Attention ! Algorithme ML $\neq$ Modèle ML



# Type d'apprentissage en ML

# Type d'apprentissage en ML



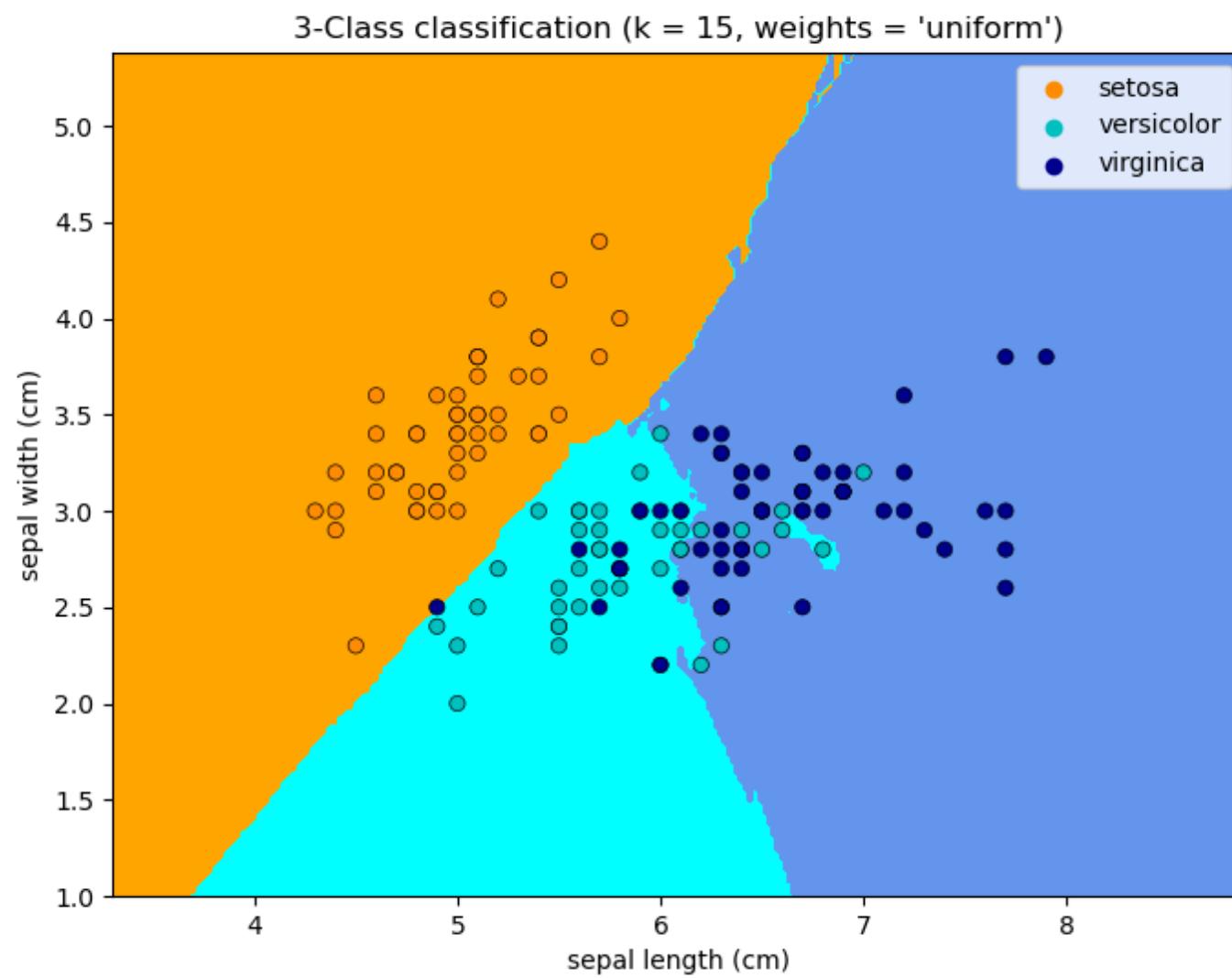
# Apprentissage supervisé

- Étant donné X, peut-on prédire Y ?
  - **Classification** – processus d'attribution d'une catégorie à l'échantillon de données d'entrée. Exemples d'usages : prédire si une personne est malade ou non, détecter les transactions frauduleuses, classifier les visages.
  - **Régression** - processus de prédiction d'une valeur numérique continue pour un échantillon de données d'entrée. Exemples d'utilisations : évaluation du prix d'une maison, prévision de la demande alimentaire des épiceries, prévision de la température.

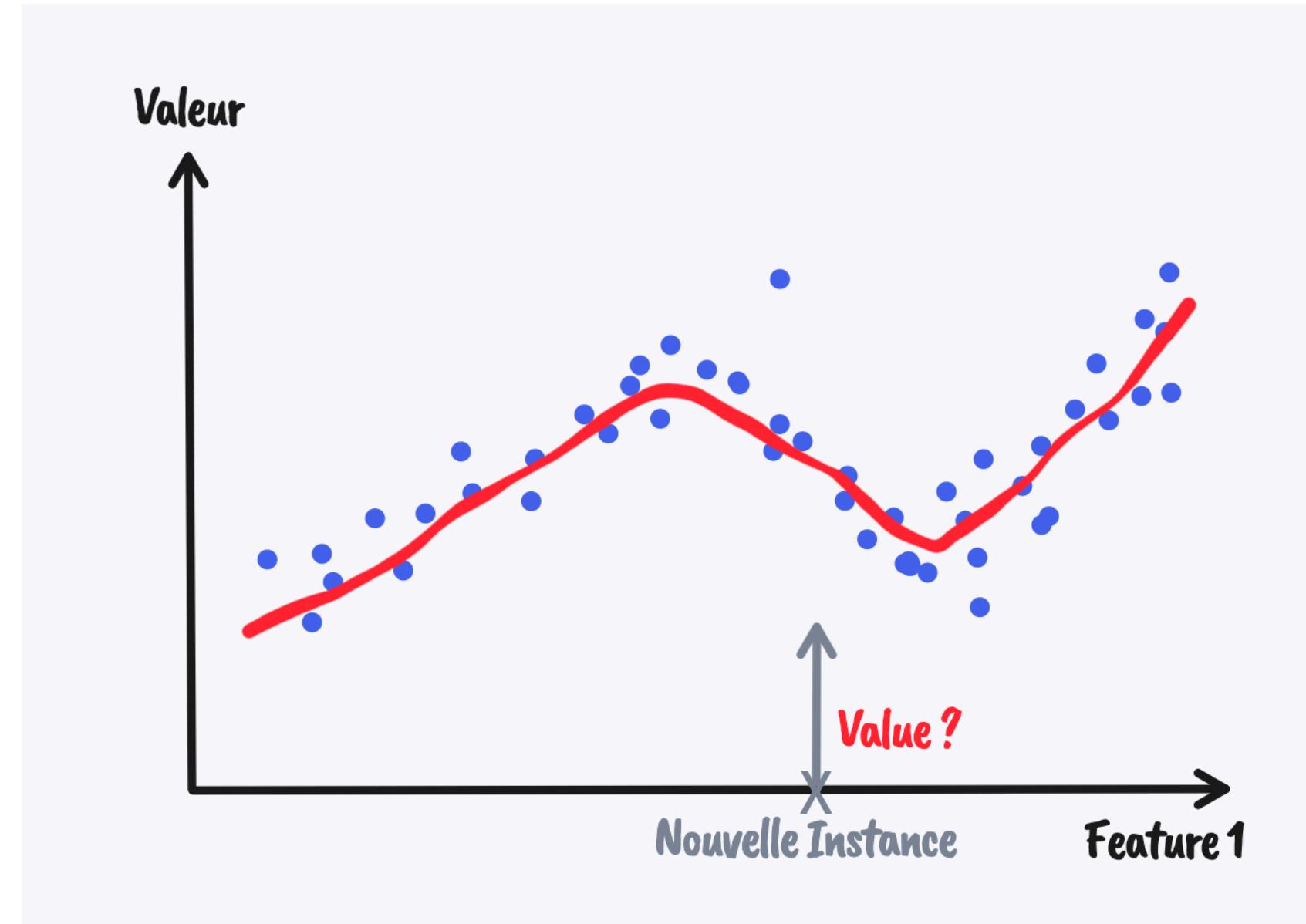
# Classification

Leaf Width	Leaf Length	Species
2.7	4.9	stetosa
3.2	5.5	versicolor
2.9	5.1	versicolor
3.4	6.8	virginica

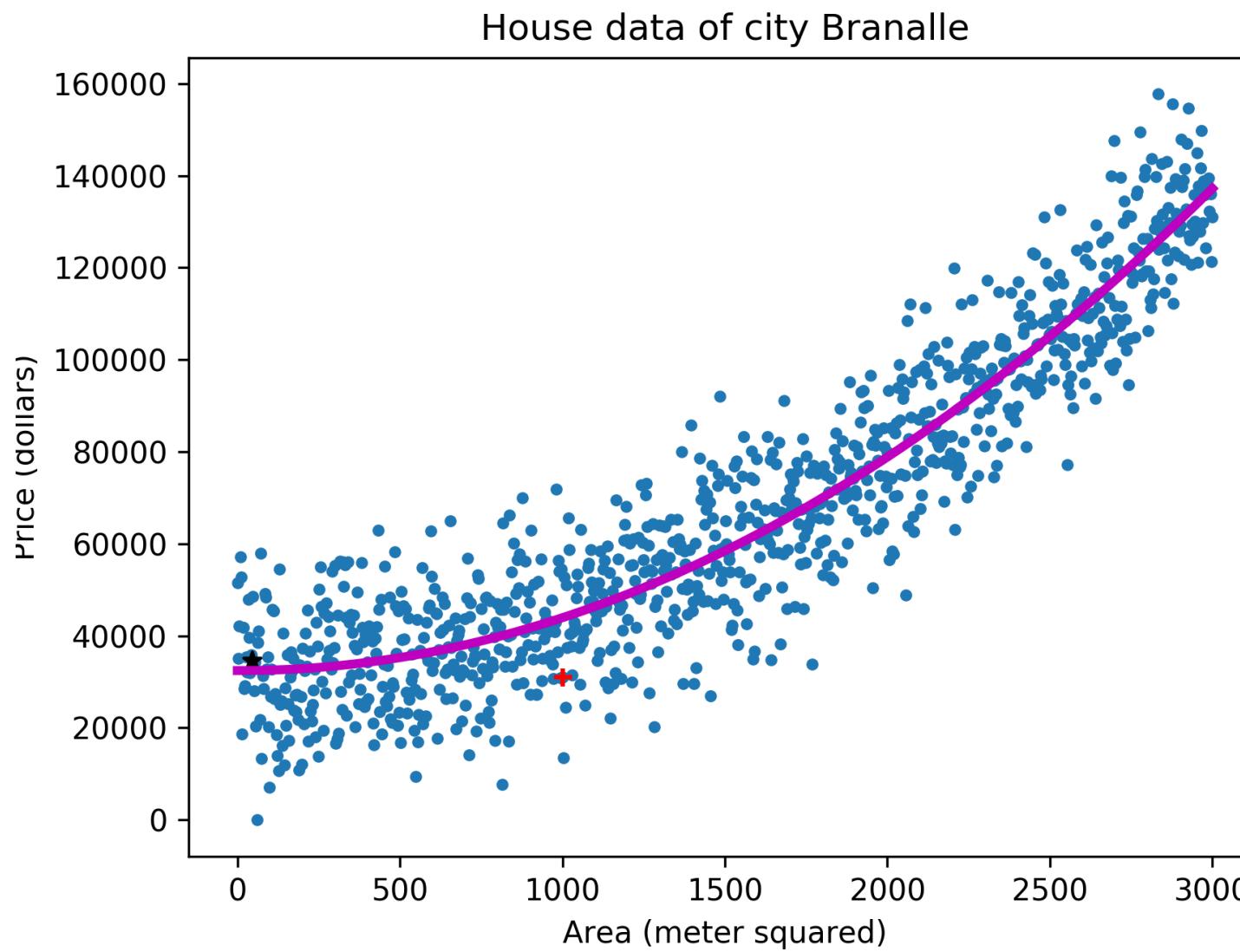
# Classification



# Régression



# Régression



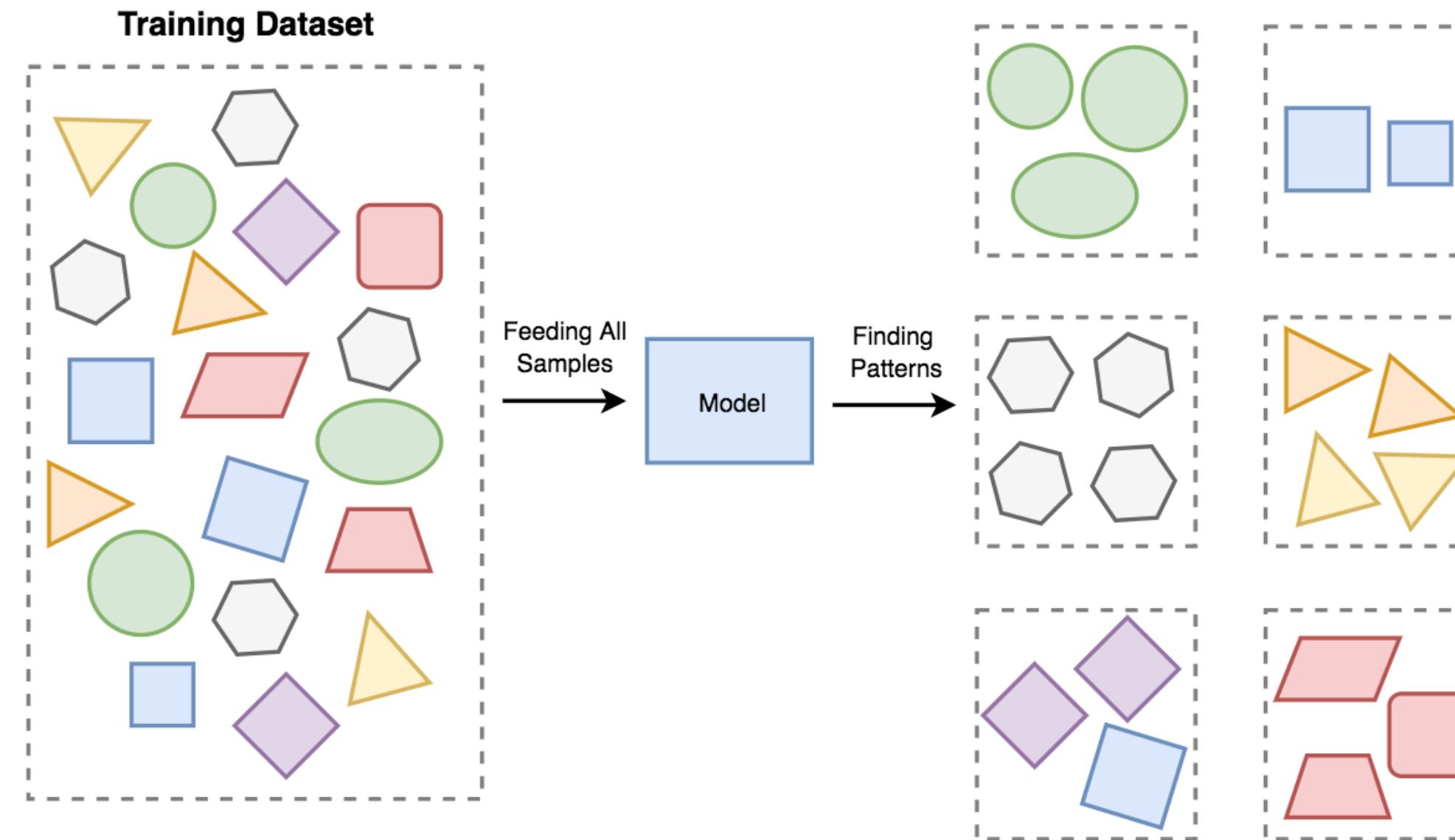
# Apprentissage non-supervisé

- Étant donné  $X$ , peut-on inférer  $Y$ ?<sup>\*</sup>
  - Clustering
  - Détection d'anomalies
  - règles d'association
  - réduction de dimensionnalité

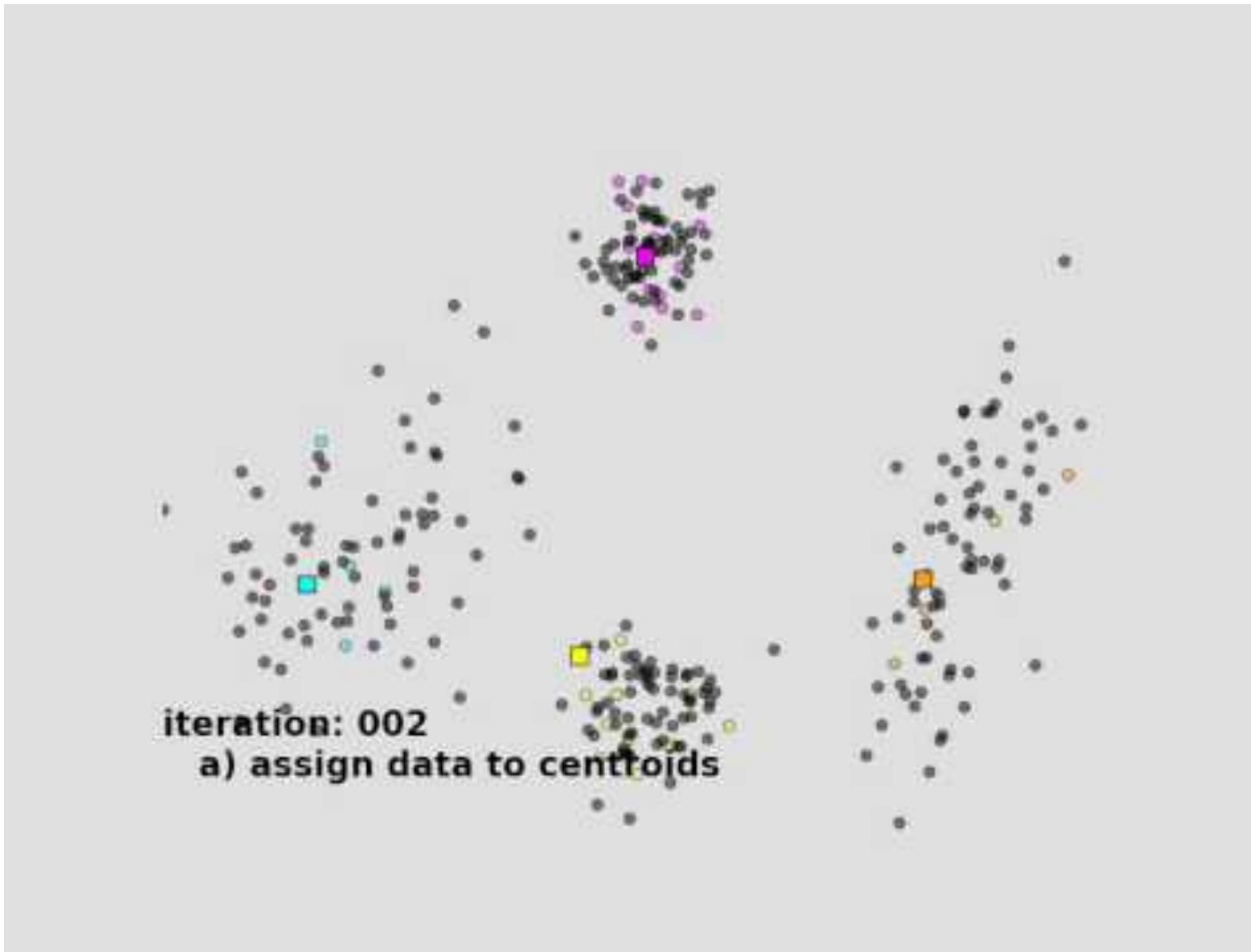
---

<sup>\*</sup>Groupe d'algorithmes qui tentent de tirer des inférences à partir de données non étiquetées (sans référence à des résultats connus ou étiquetés). Dans l'apprentissage non supervisé, il n'y a pas de bonnes réponses.

# Clustering



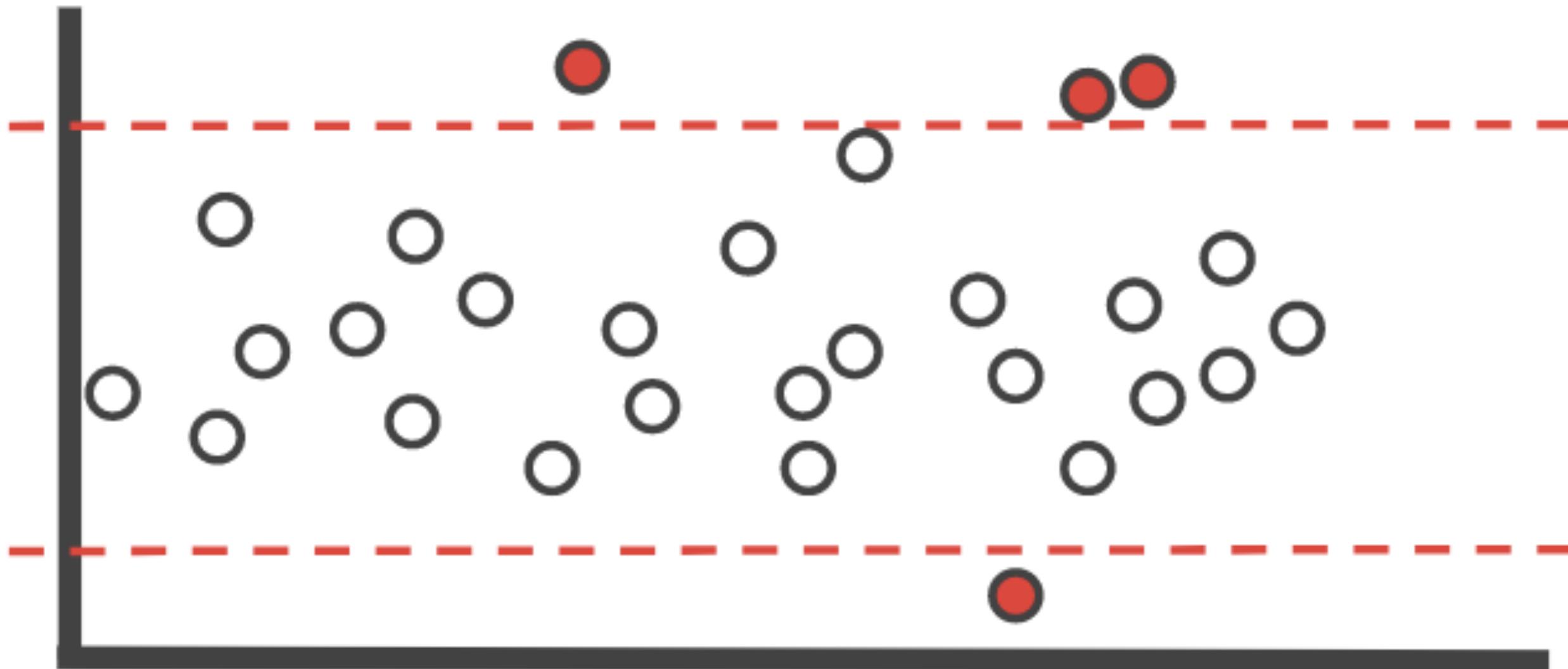
# Clustering



# Règles d'association

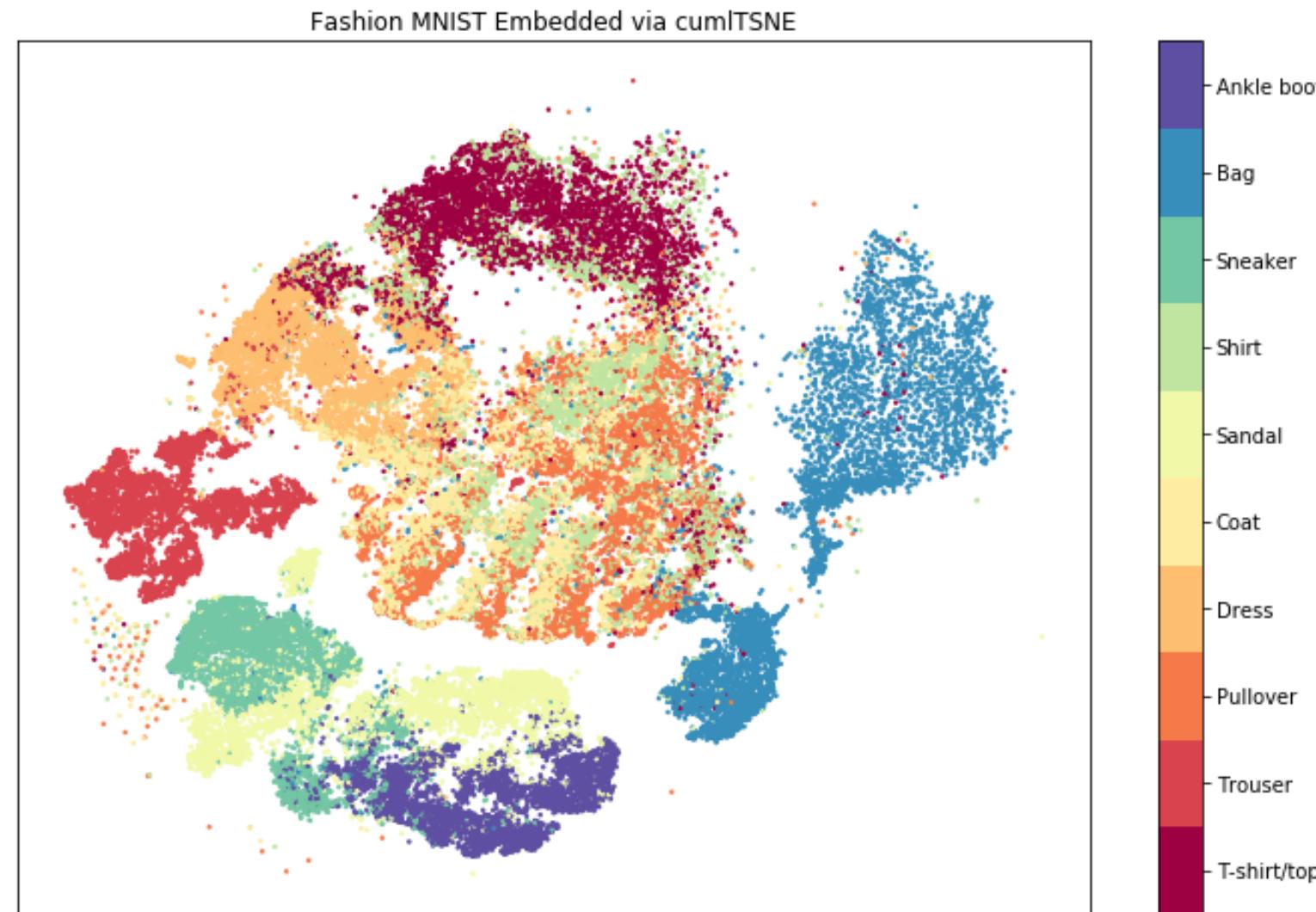
Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

# Détection d'anomalies



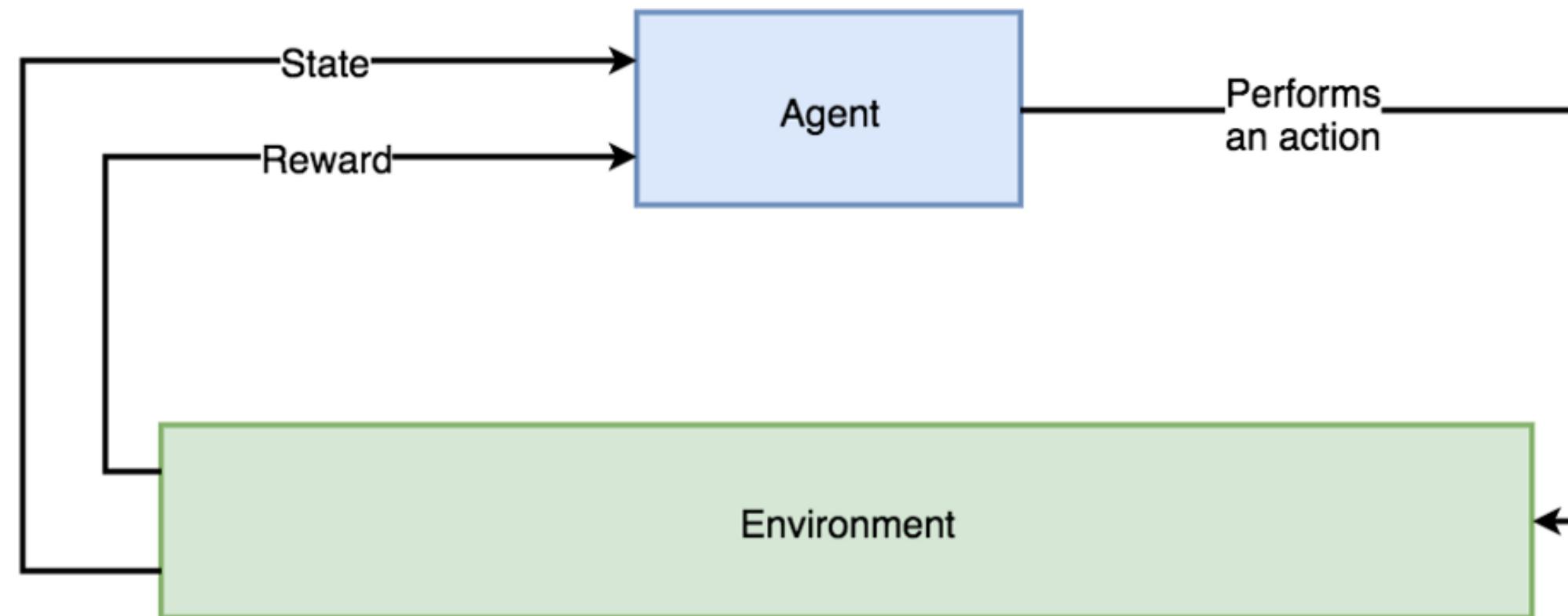
# Réduction de la dimensionnalité

```
CPU times: user 2.02 s, sys: 896 ms, total: 2.91 s
Wall time: 2.9 s
```



# L'apprentissage par renforcement

- L'**apprentissage par renforcement** consiste, pour un agent autonome (robot, etc.), à apprendre les actions à prendre, à partir d'expériences, de façon à optimiser une récompense quantitative au cours du temps.



# Type de problèmes

Type de problème	Description	Example
Classification	Choisissez l'une des N étiquettes	chat, chien, cheval, etc.
Regression		
Clustering		
Règles d'association		
Prédiction de données structurée		
Hiérarchisation (Recommendation)		
Apprendre des actions dans un environnement		

# Question

Quel problème de ML est un exemple d'apprentissage non supervisé ?

- [ ] Clustering
- [ ] Régression
- [ ] Prédictiton de données structurée
- [ ] Classification



# Question

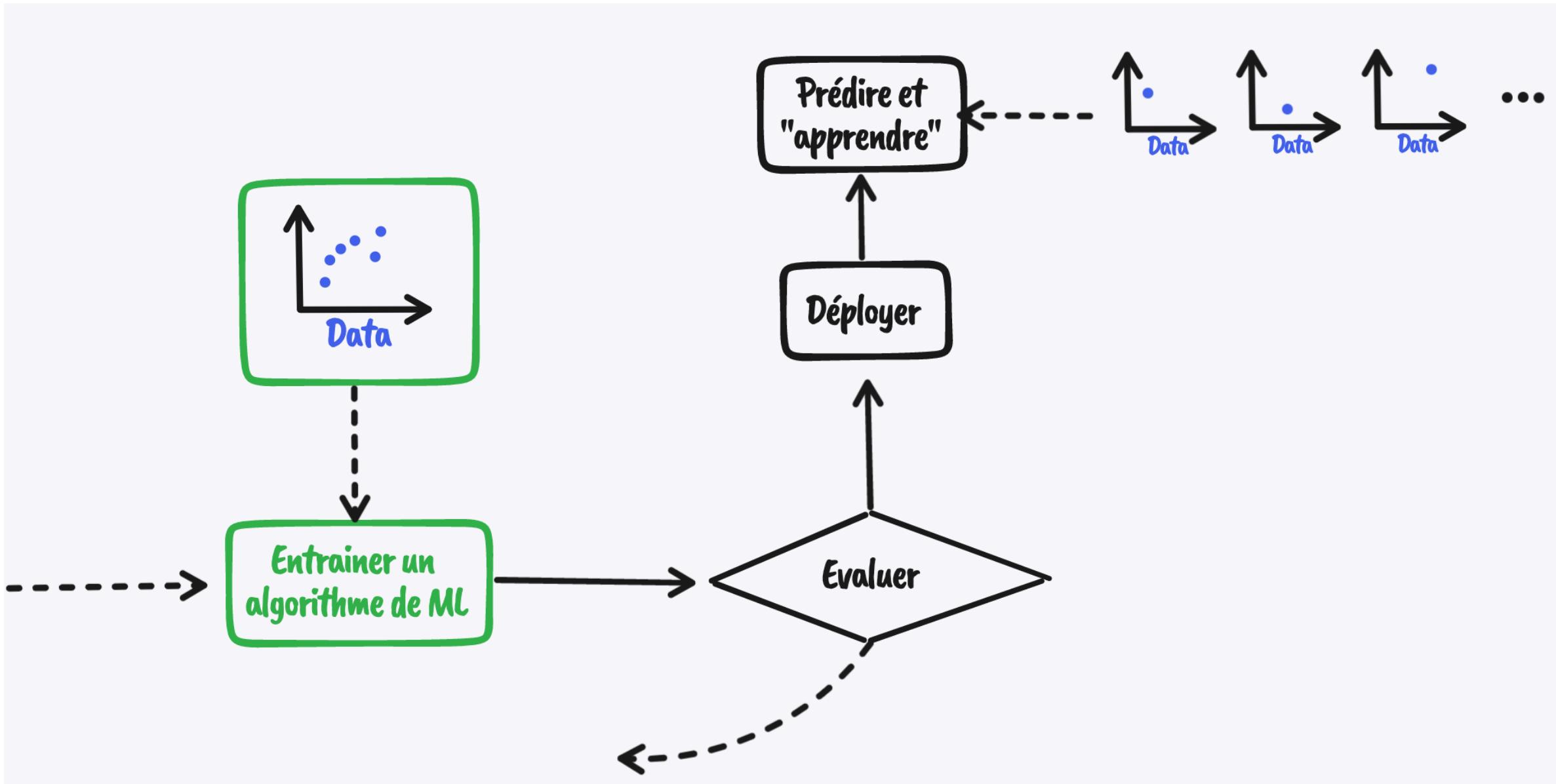
Quel problème de ML est un exemple d'apprentissage non supervisé ?

- [X] Clustering
- [ ] Régression
- [ ] Prédictiton de données structurée
- [ ] Classification

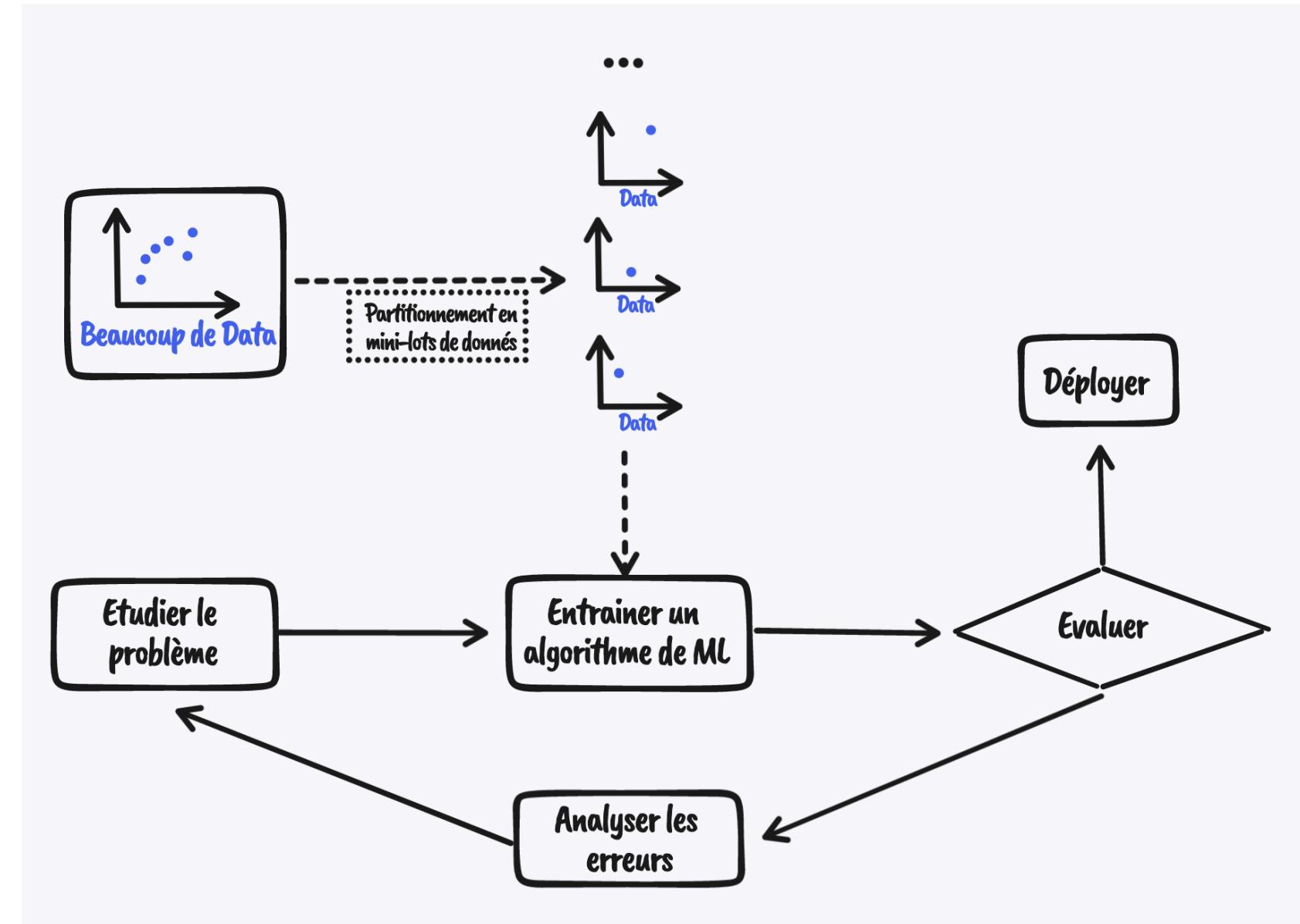


# **Apprentissage en ligne ou par batch**

# Apprentissage en ligne



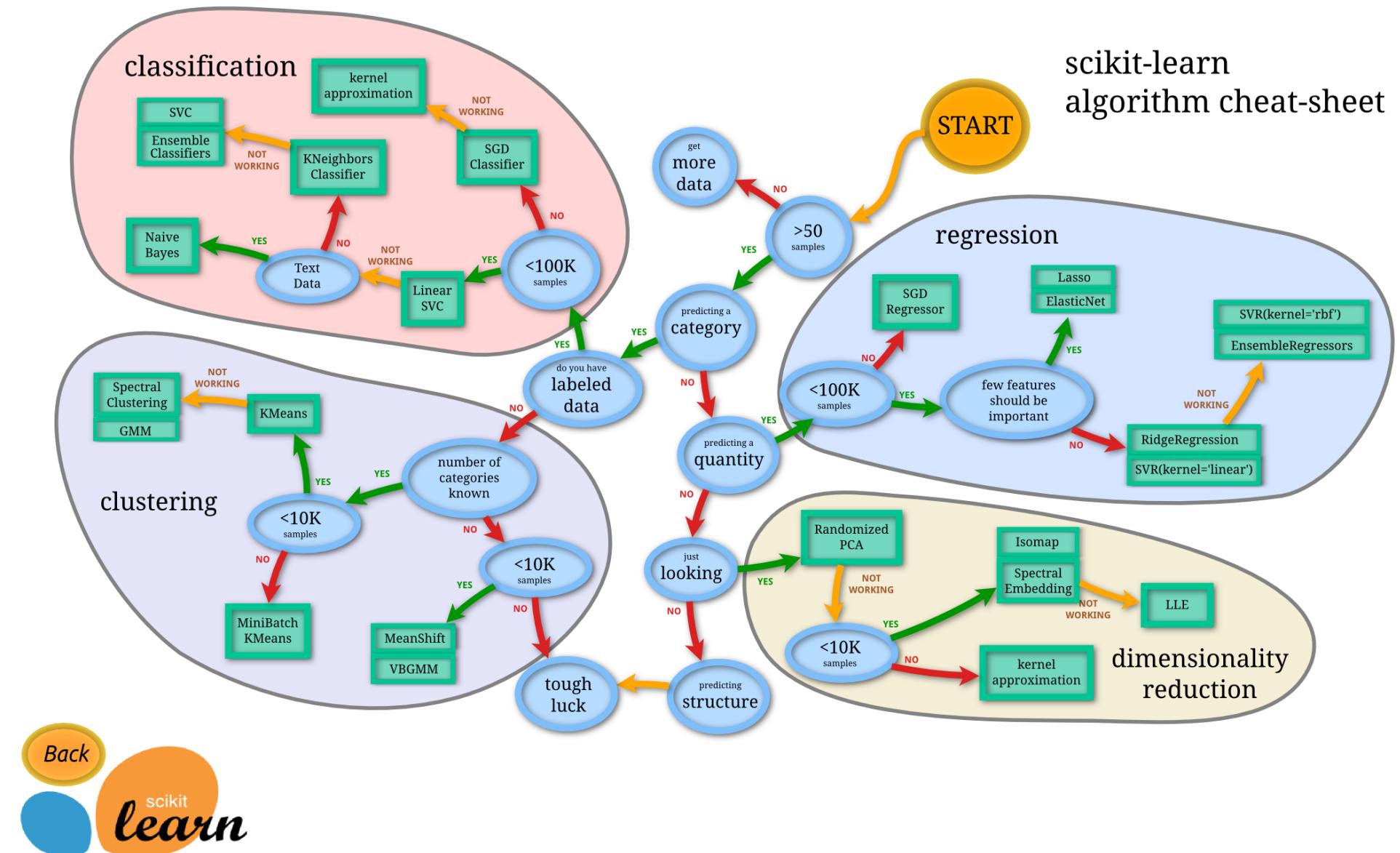
# Apprentissage "out-of-core"



# Exercice

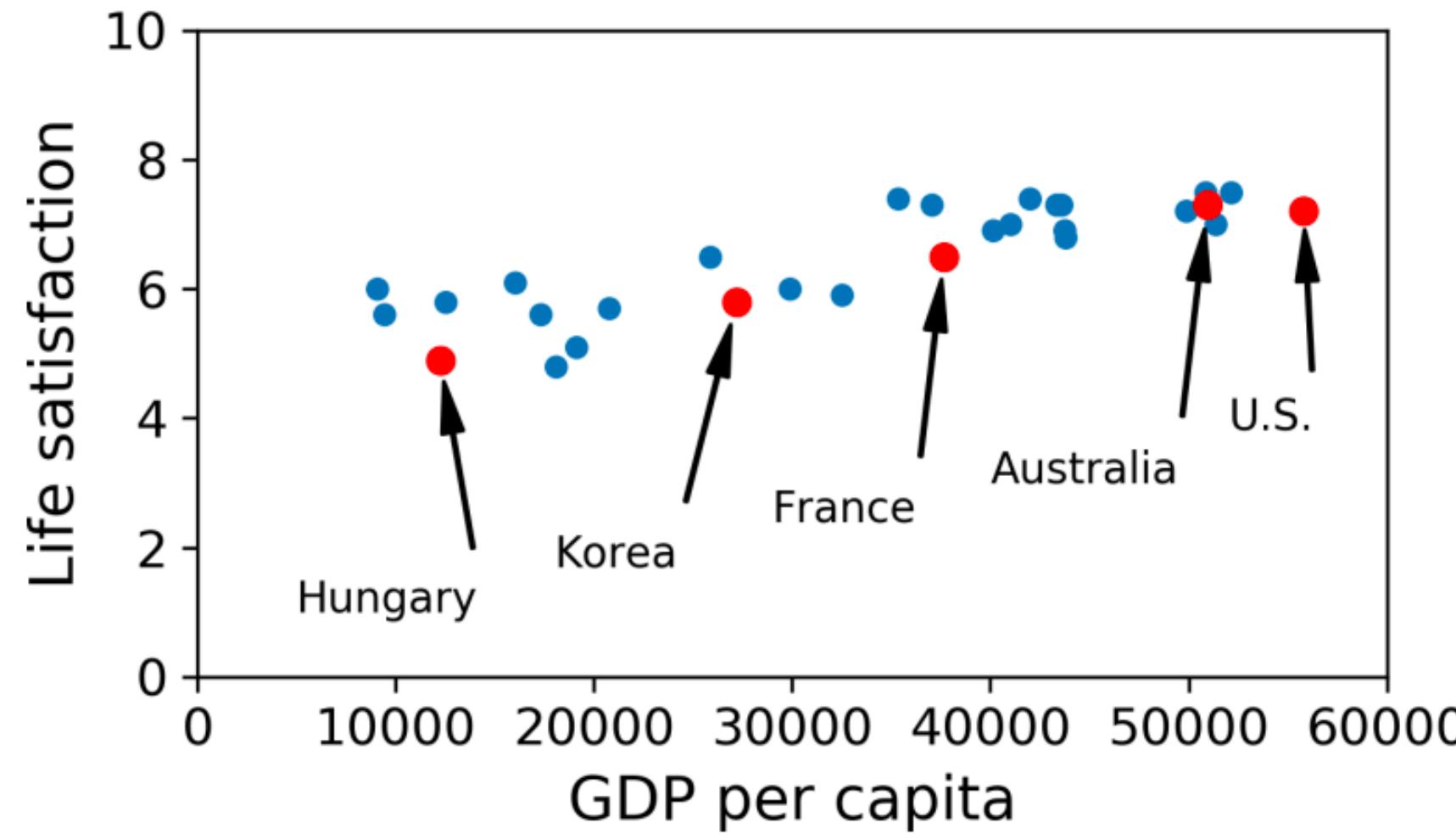
Pour chaque application présentée dans le cours. Identifier le problème ML impliquée.

# Quelques algorithmes par type d'apprentissage



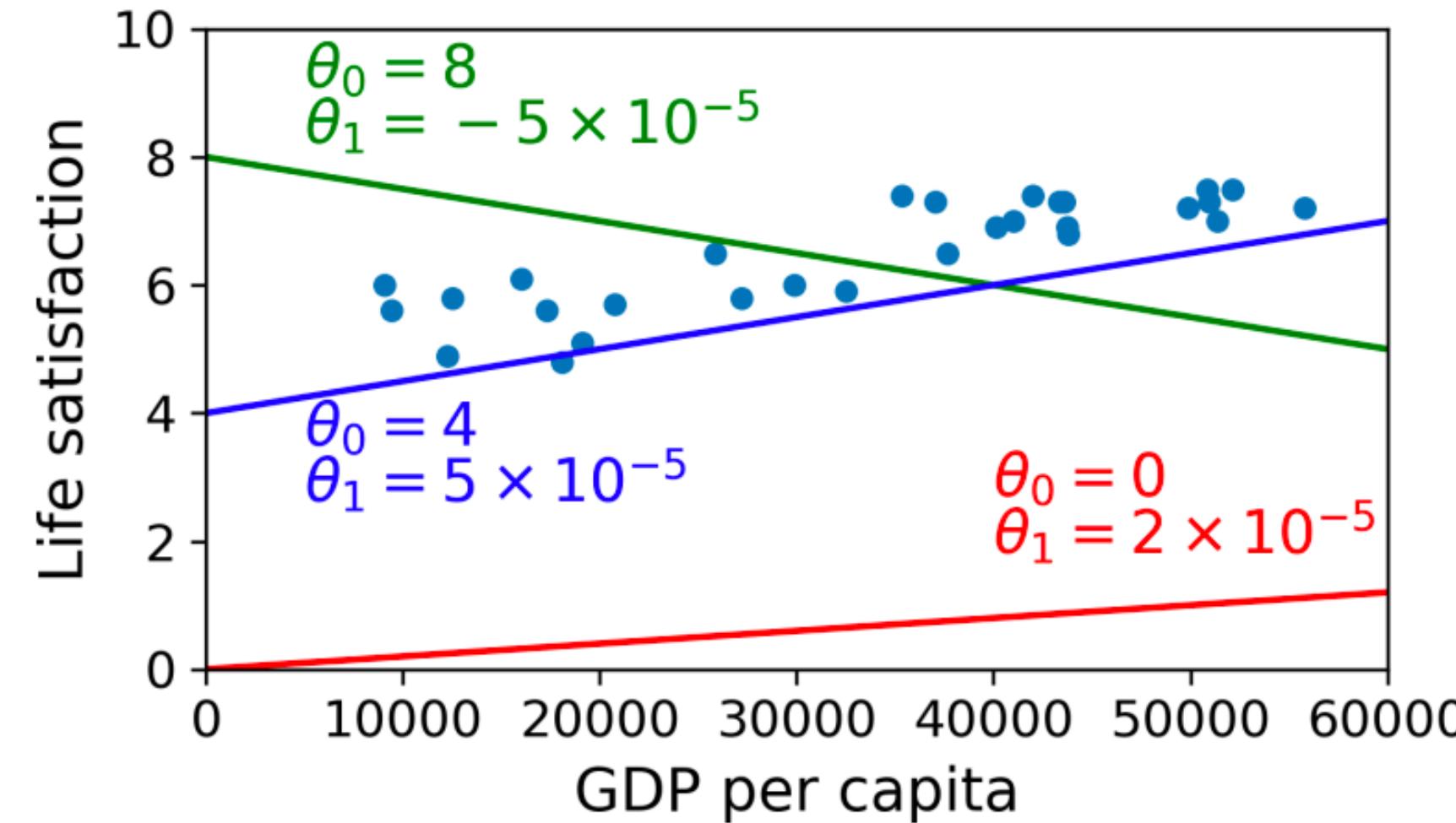
# **Principaux défis du Machine Learning**

# Tendance ?



$$life\_satisfaction = \theta_0 + \theta_1 \times GDP\_per\_capita$$

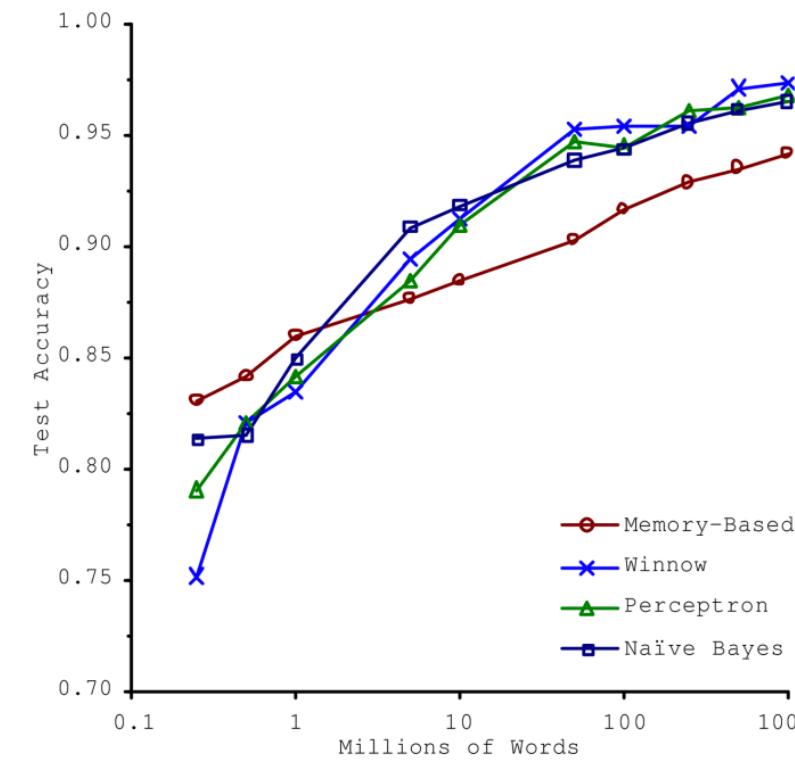
# Modèles simples



$$life\_satisfaction = \theta_0 + \theta_1 \times GDP\_per\_capita$$

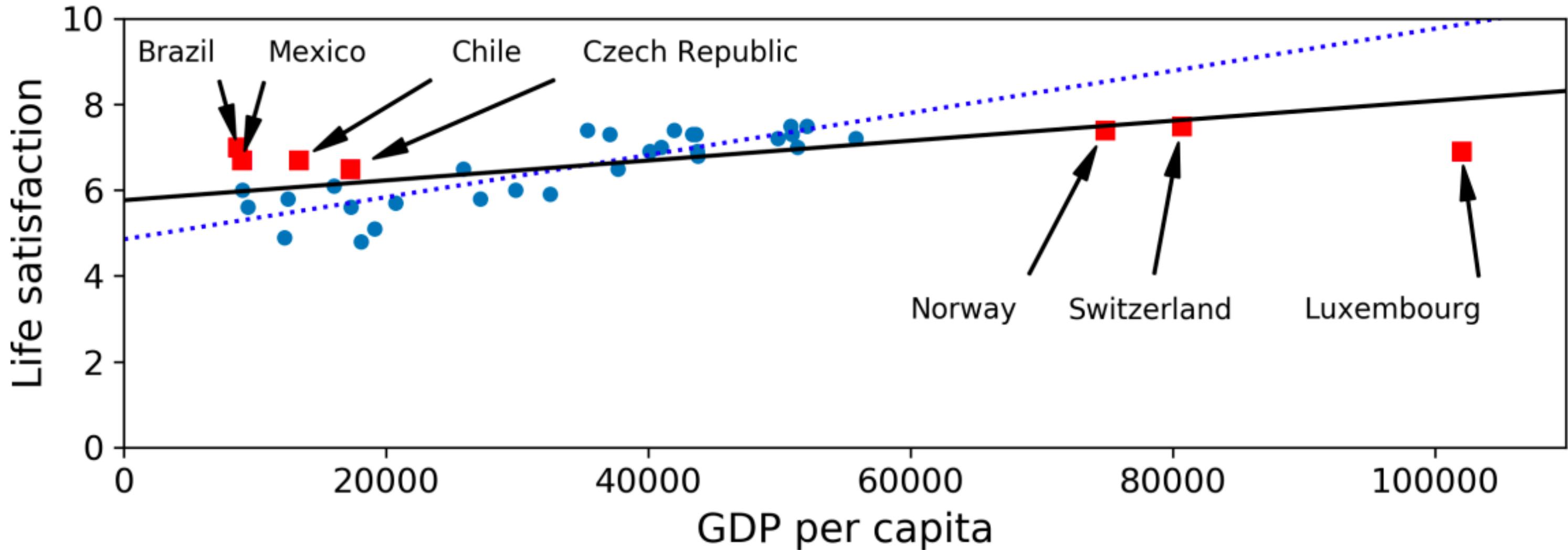
# Quantité insuffisante de données d'entraînement

1



<sup>1</sup> courbe extraite de Banko and Brill (2001), "Learning Curves for Confusion Set Disambiguation."

# Données non-représentatives



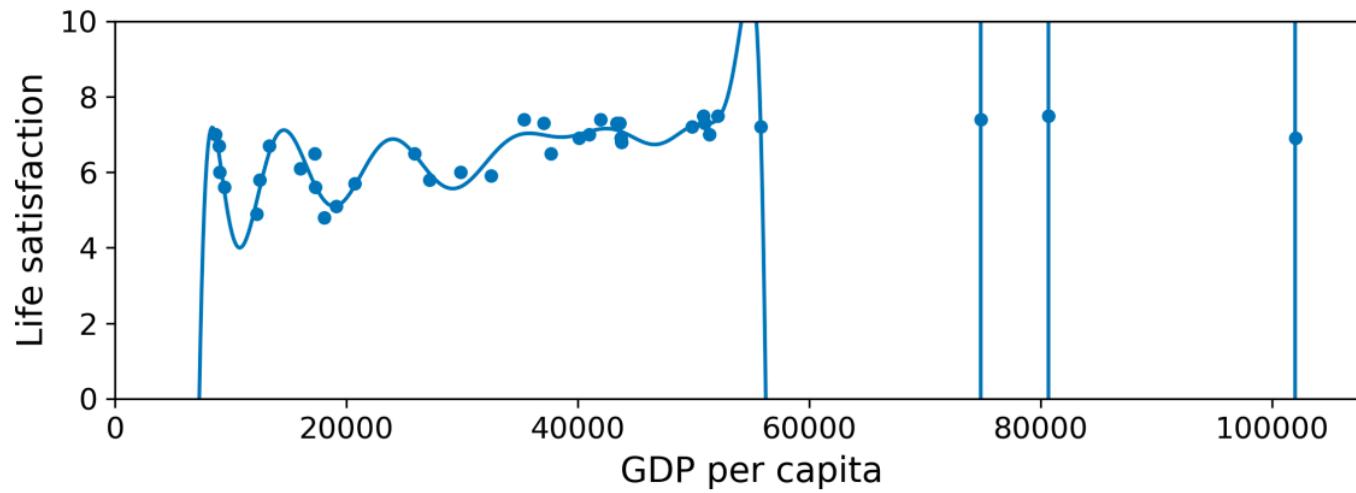
# Données de mauvaise qualité

- Si certaines observations sont clairement aberrantes, il peut être utile de simplement les supprimer ou d'essayer de corriger les erreurs manuellement.
- S'il manque quelques fonctionnalités à certaines instances (par exemple, 5% de vos clients n'ont pas spécifié leur âge), vous devez décider si vous souhaitez ignorer complètement cet attribut, ignorer ces instances, remplir les valeurs manquantes (par exemple, avec la médiane âge), ou entraînez un modèle avec la fonctionnalité et un modèle sans, et ainsi de suite.

# Caractéristiques non pertinentes

- *Feature Selection* (Sélection des caractéristiques) : sélection des caractéristiques les plus utiles pour s'entraîner parmi les fonctionnalités existantes.
- *Feature Extraction* : combiner des caractéristiques existantes pour en produire une plus utile (comme nous l'avons vu précédemment, les algorithmes de réduction de dimensionnalité peuvent aider).
- Création de nouvelles *Features* en collectant de nouvelles données

# Surapprentissage (Overfitting the Training Data)

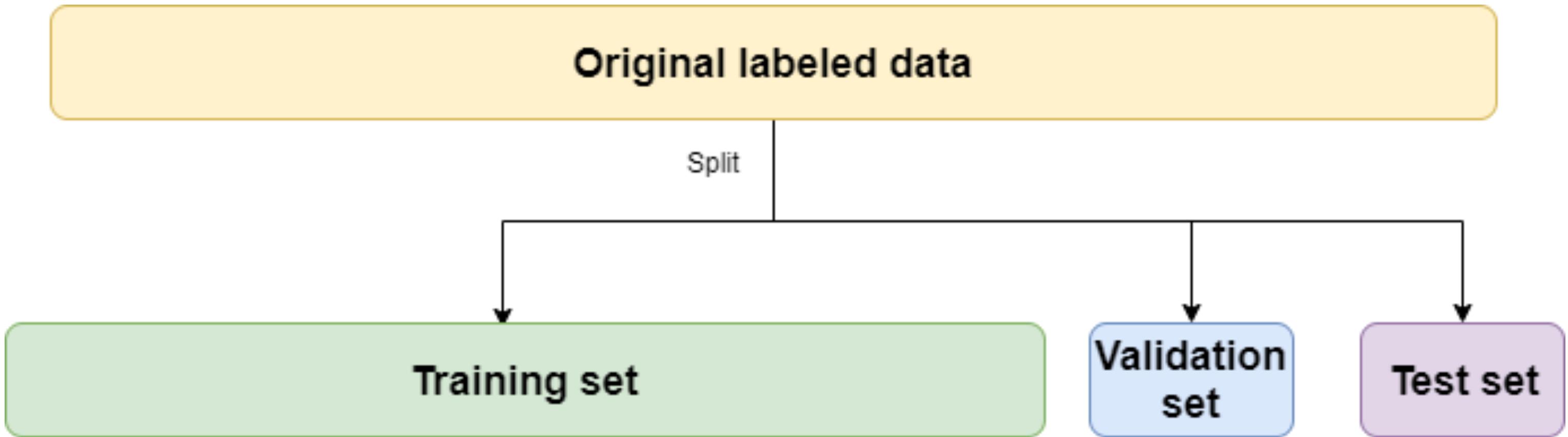


- Simplifier le modèle en sélectionnant un avec moins de paramètres (par exemple, un modèle linéaire plutôt qu'un modèle polynomial à haut degré), en réduisant le nombre de *Features* dans les données d'apprentissage ou en contrignant le modèle
- Recueillir plus de données d'entraînement
- Réduire le bruit dans les données d'apprentissage (par exemple, corriger les erreurs de données et supprimer les valeurs aberrantes)

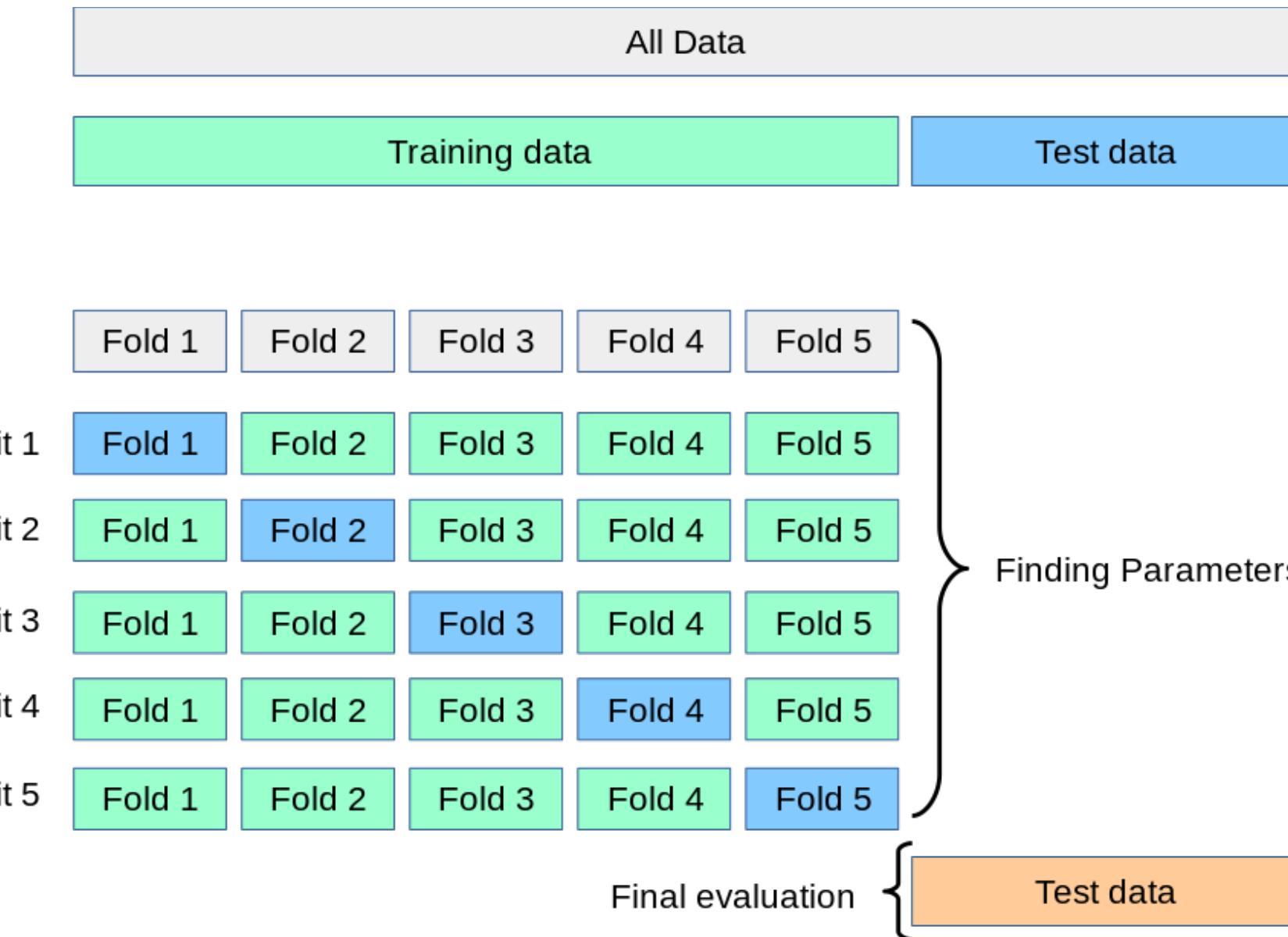
# Sous-apprentissage (Underfitting the Training Data)

- Sélection d'un autre modèle, avec plus de paramètres
- Fournir de meilleures *Features* à l'algorithme d'apprentissage (*Feature Engineering*)
- Réduction des contraintes sur le modèle (par exemple, réduction de l'hyper-paramètre de régularisation)

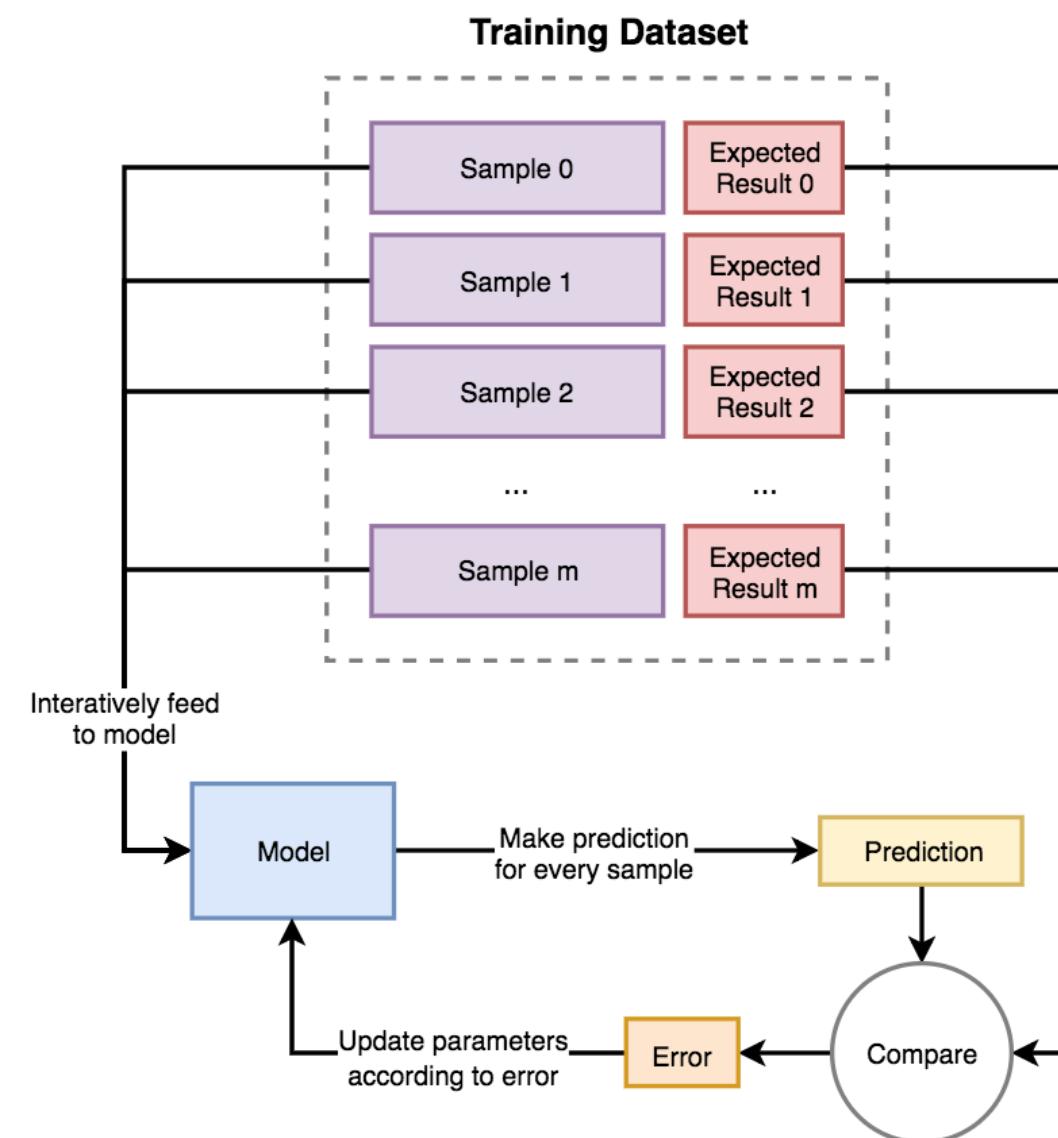
# Test et validation



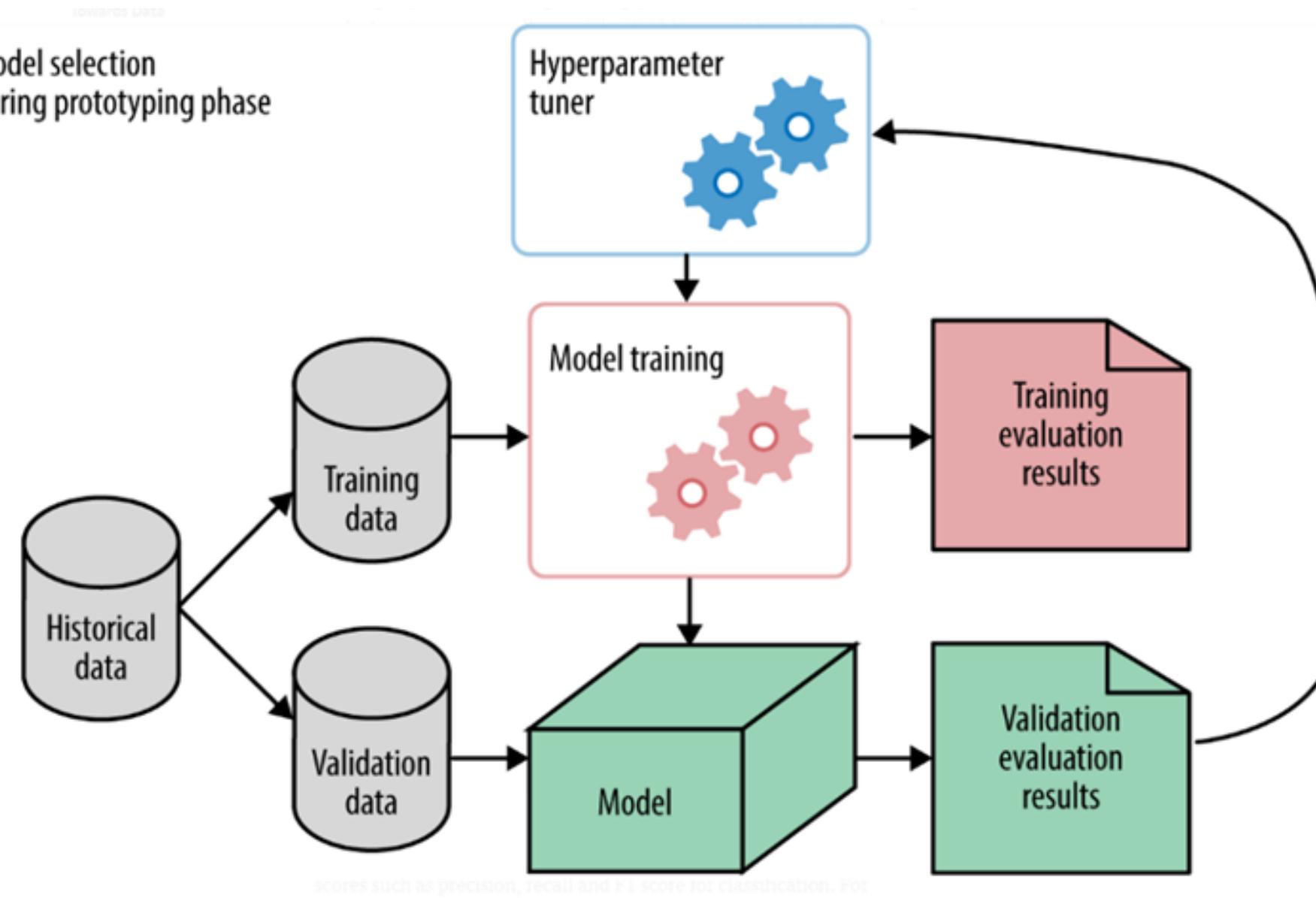
# Test et validation



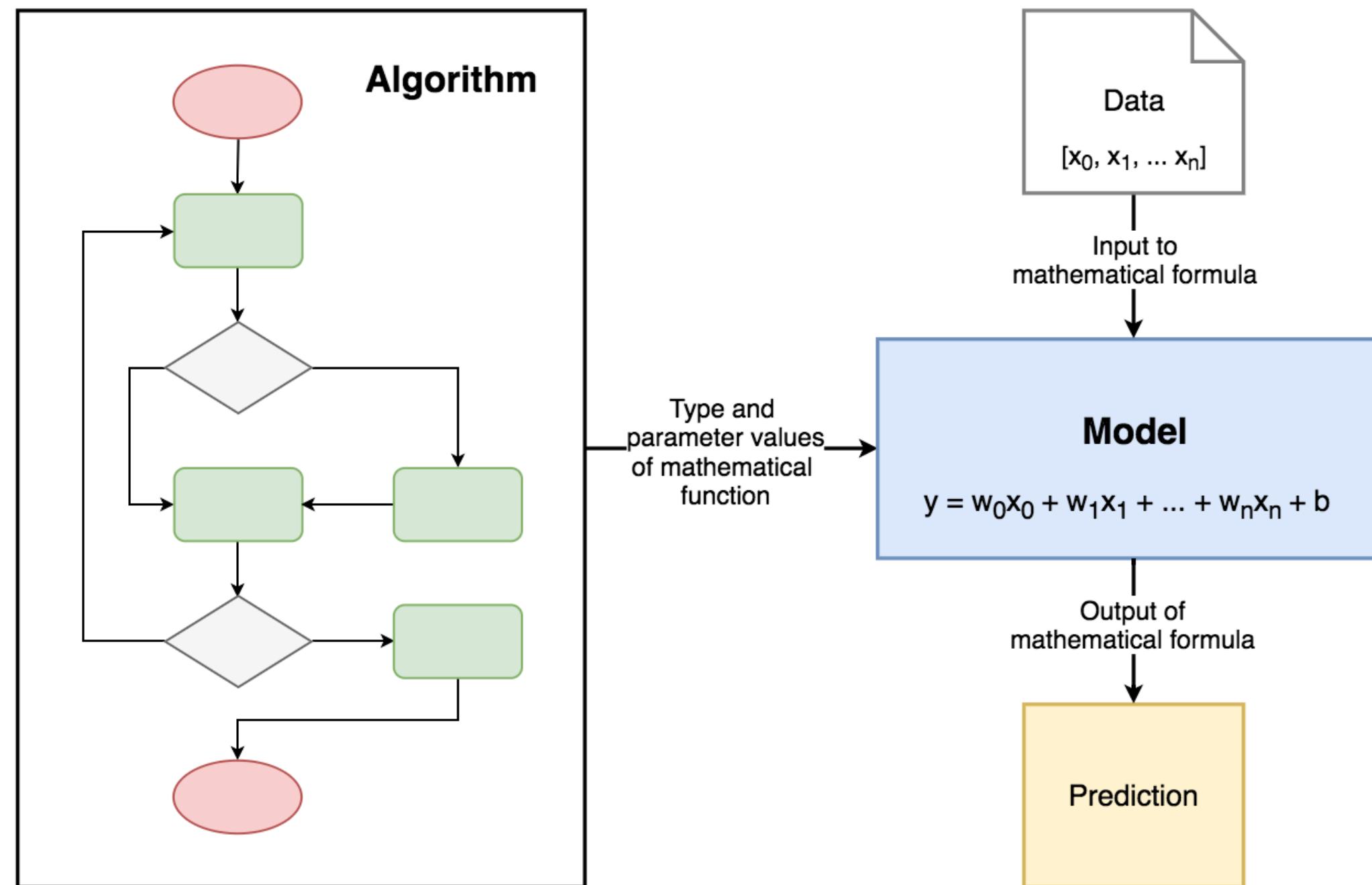
# Réglage des hyperparamètres



# Réglage des hyperparamètres



# Attention ! Algorithme ML $\neq$ Modèle ML



# Evaluation dans un problème de classification

		Positive	Predicted	Negative
		T		
		True Positive (TP) <i>Correct objects</i>		False Negative (FN) <i>Missed objects</i>
		False Positive (FP) <i>Extra objects</i>	True Negative (TN) <i>No objects</i>	
		Precision	Negative predictive value	
		$\frac{TP}{(TP+FP)}$	$\frac{TN}{(TN+TP)}$	

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

# Evaluation d'un problème de régression

$$MAE = \left( \frac{1}{n} \right) \sum_{i=1}^n |y_i - x_i|$$

$$RMSE = \sqrt{\left( \frac{1}{n} \right) \sum_{i=1}^n (y_i - x_i)^2}$$

```
import sklearn.metrics
import math

S1 = [2, 5, 9, 2]
S2 = [6, 3, 6, 1]

mae = sklearn.metrics.mean_absolute_error(S1, S2)

mse = sklearn.metrics.mean_squared_error(S1, S2)
rmse = math.sqrt(mse)
```