Analyse, Classification et Indexation des données (ACID)

Introduction, définitions, ...

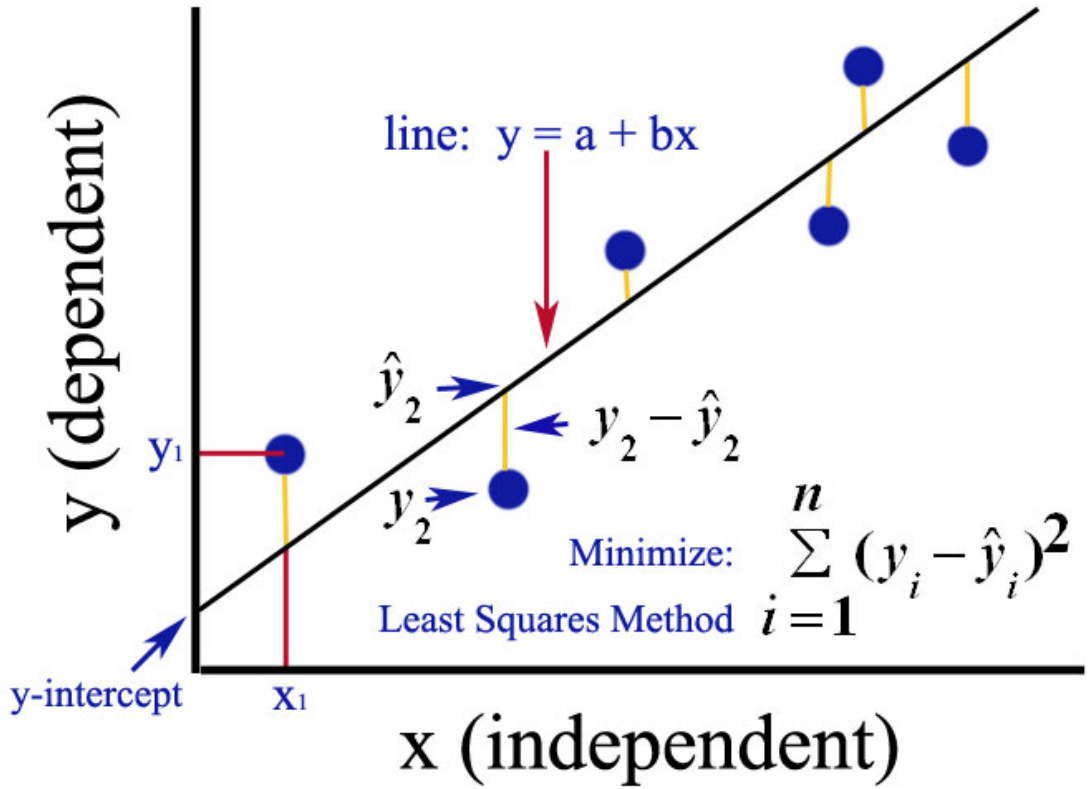
Akka Zemmari

LaBRI, Université de Bordeaux

2023 - 2024

Machine Learning ?

Pour démystifier la chose :



C'est quoi le Machine Learning

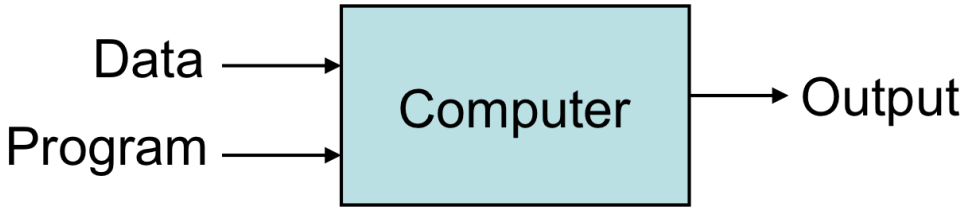
- "The goal of machine learning is to build computer systems that can adapt and learn from their experience." - Tom Dietterich
- To learn: to acquire knowledge by study, experience or being taught.
- A Computer program is said to learn by experience E with respect to class of tasks T performance measure P, if the performance at the task T, measured by performance P, improves by experience E.

C'est quoi le Machine Learning

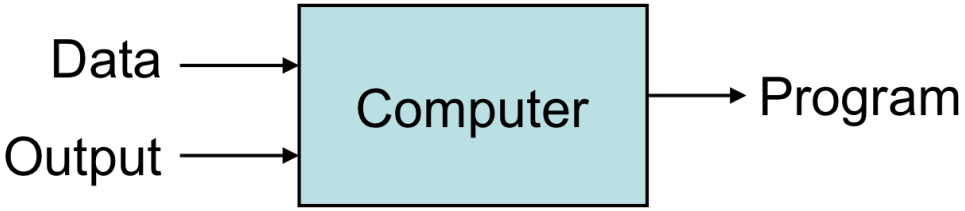
- task to be addressed by the system (e.g. recognizing handwritten characters)
- performance measure to evaluate the learned system (e.g. number of misclassified characters)
- training experience to train the learning system (e.g. labelled handwritten characters)

C'est quoi le Machine Learning

- Traditional Programming

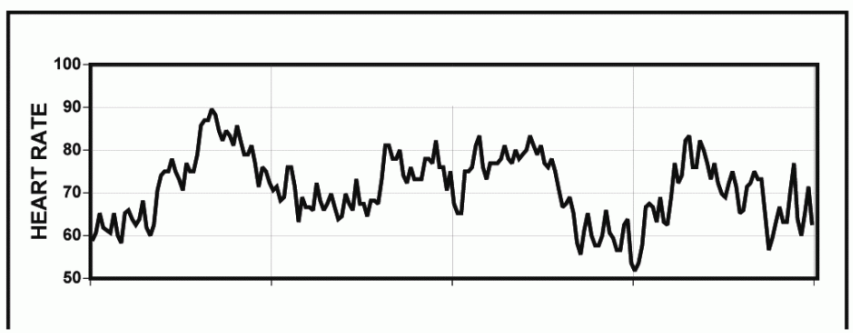


- Machine Learning



Quelques Exemples

- A partir des données médicales (sur 20 ans) sur des patients, peut-on dire si un patient risque d'avoir une crise cardiaque dans les 2,3 prochaines années ?



Quelques Exemples

- On dispose de données cliniques sur 20 ans. Un patient risque-t-il d'avoir une attaque cardiaque dans les 5 prochaines années ?
- Quel sera le prix de ce stock dans 6 mois ?

Quelques Exemples

- On dispose de données cliniques sur 20 ans. Un patient risque-t-il d'avoir une attaque cardiaque dans les 5 prochaines années ?
- Quel sera le prix de ce stock dans 6 mois ?
- Es-ce que ce gribouillage est un 7 ?

Quelques Exemples

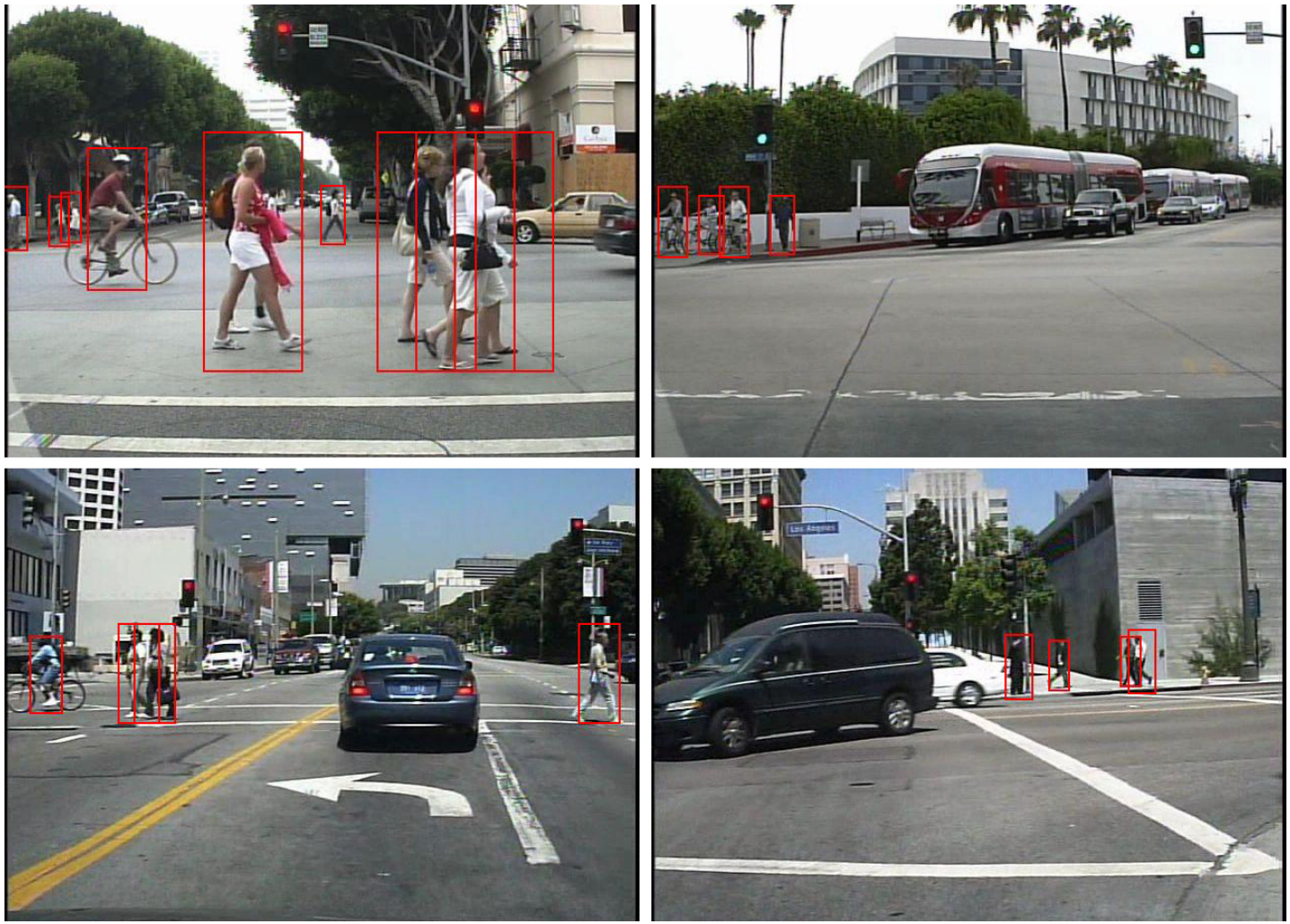
- On dispose de données cliniques sur 20 ans. Un patient risque-t-il d'avoir une attaque cardiaque dans les 5 prochaines années ?
- Quel sera le prix de ce stock dans 6 mois ?
- Es-ce que ce gribouillage est un 7 ?
- Est-ce que le mail que je viens juste de recevoir est un spam ?

Quelques Exemples

- On dispose de données cliniques sur 20 ans. Un patient risque-t-il d'avoir une attaque cardiaque dans les 5 prochaines années ?
- Quel sera le prix de ce stock dans 6 mois ?
- Es-ce que ce gribouillage est un 7 ?
- Est-ce que le mail que je viens juste de recevoir est un spam ?
- Est-ce que je peux regrouper des consommateurs ? Des mots ? Des gènes ?

Et plus récemment ... le *Deep* ...

Reconnaissance vidéo [1]



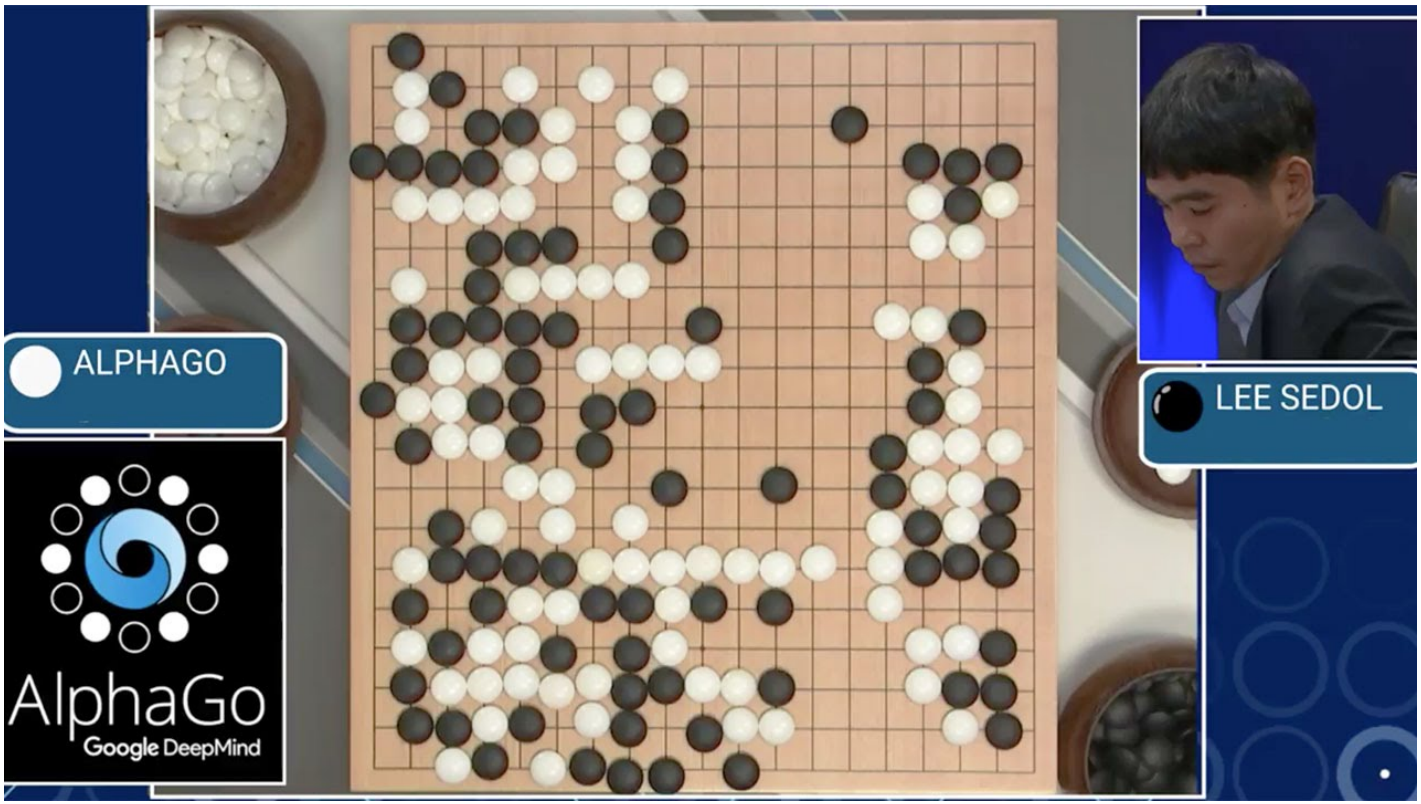
[1] https://3c1703fe8d.site.internapcdn.net/newman/gfx/news/hires/2016/newalgorit

Et plus récemment ... le *Deep* ...

Reconnaissance vocale [2]

[2] https://technologybux.com/urdu-speech-recognition-no-longer-dream/

Et plus récemment ... le *Deep* ...



Compétences nécessaires

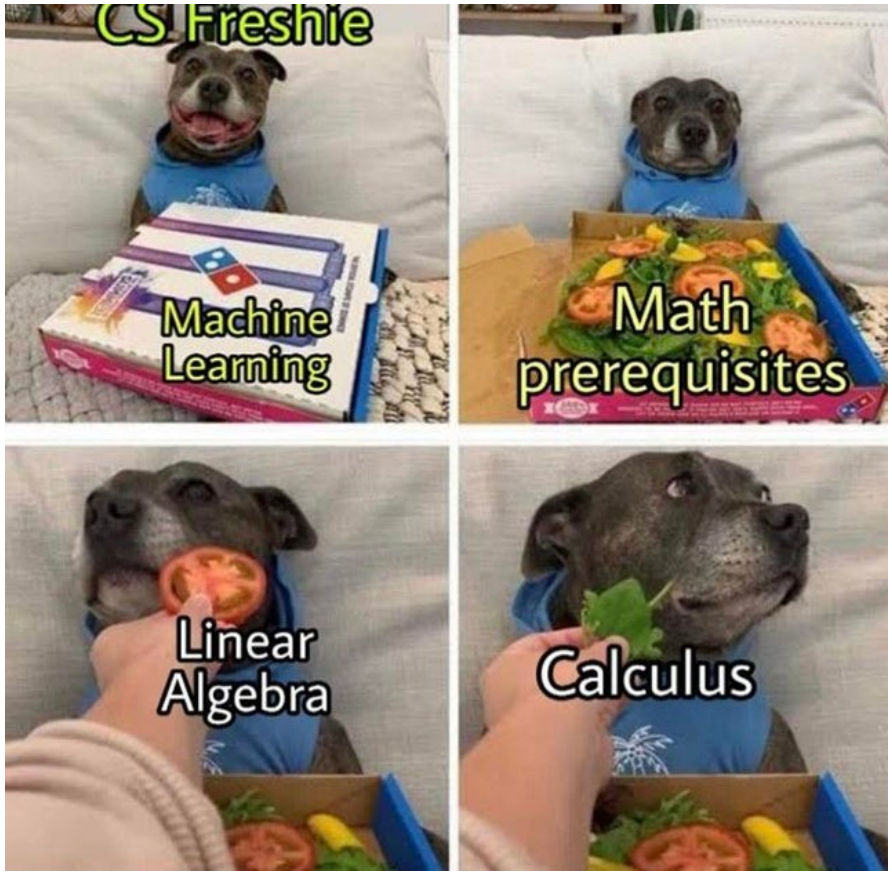
Pour faire du Machine Learning, on a besoin de :

Mathématiques, Algorithmique, Programmation ...

Sinon, on risque d'utiliser le ML comme ...

... une "black box" ...

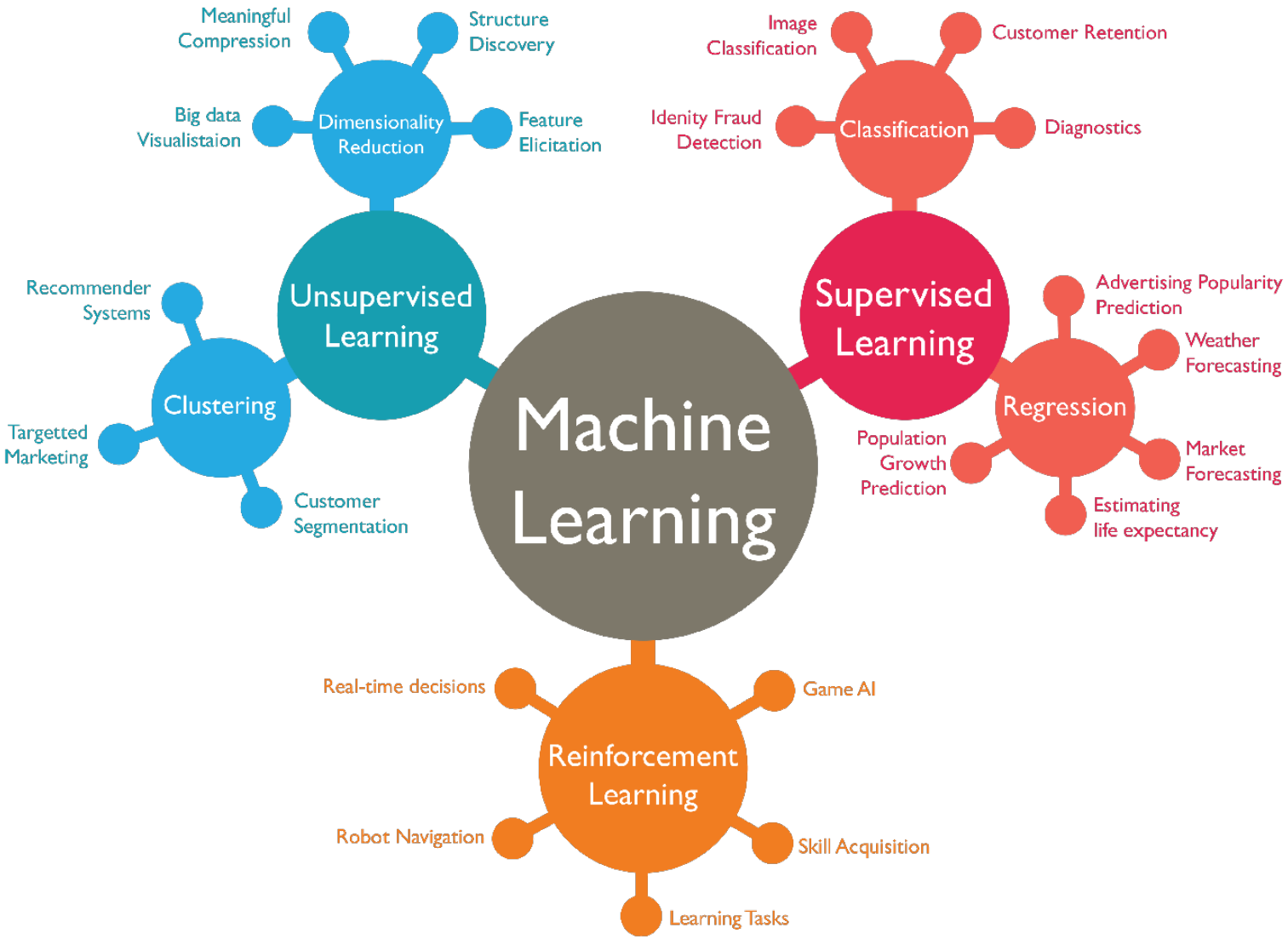
Compétences nécessaires



Étapes à suivre pour concevoir un système ML

1. Formaliser la tâche d'apprentissage.
2. Collecter les données.
3. Extraire les caractéristiques (features).
4. Choisir la bonne classe de modèles d'apprentissage.
5. Entraîner le modèle.
6. Évaluer le modèle.

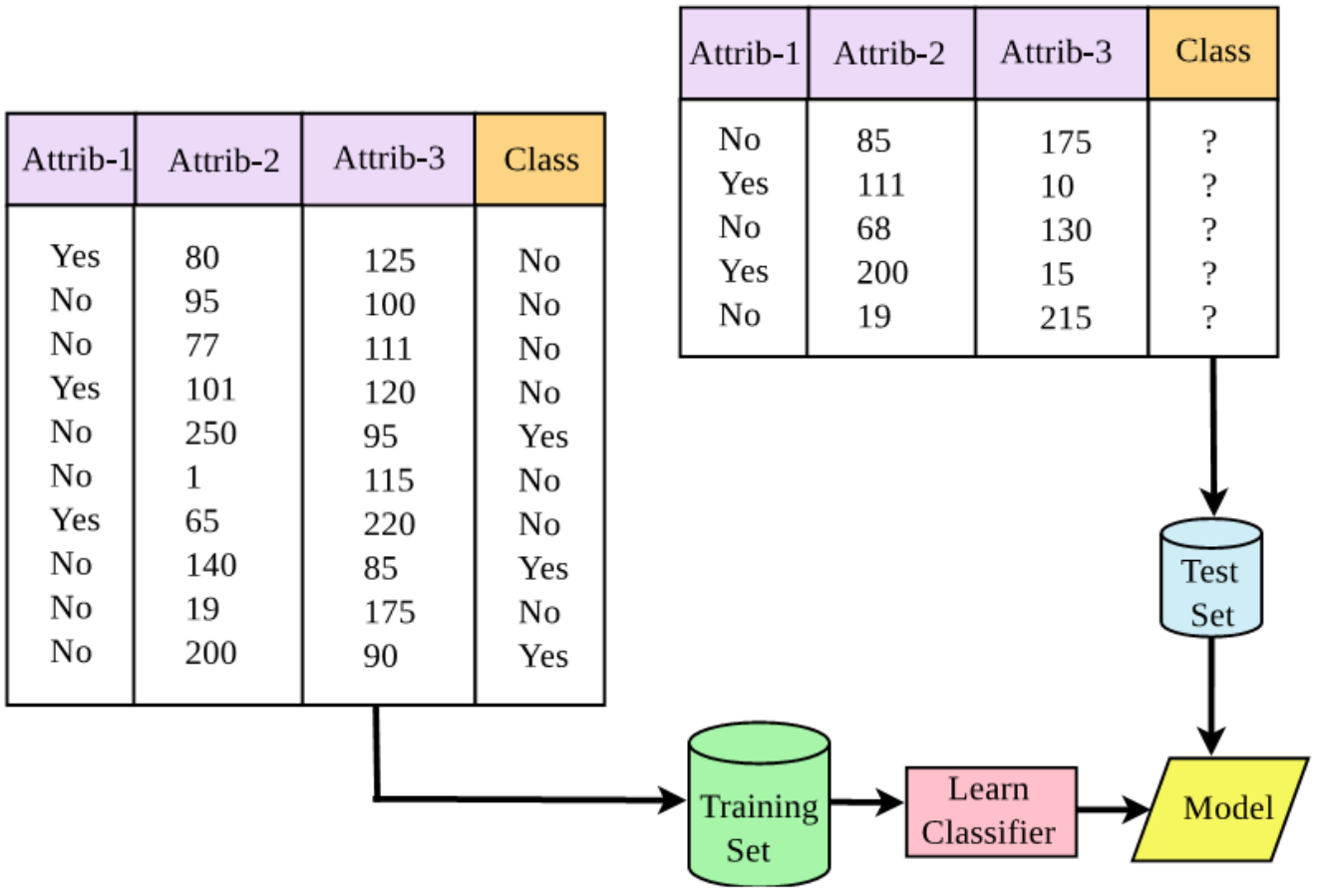
Modèles d'apprentissage



Modèles d'apprentissage

- Apprentissage supervisé :
  - Supervision : les données d'apprentissage (observations) sont accompagnés par les labels indiquant leurs classes.
  - Les nouvelles données sont classifiées en se basant sur le training set.

Apprentissage supervisé



Modèles d'apprentissage

- Apprentissage non supervisé :
  - Le label de classe des éléments observés (training set) n'est pas connu.
  - Le but est de déceler l'existence de classes ou groupes dans les données.

Apprentissage non supervisé

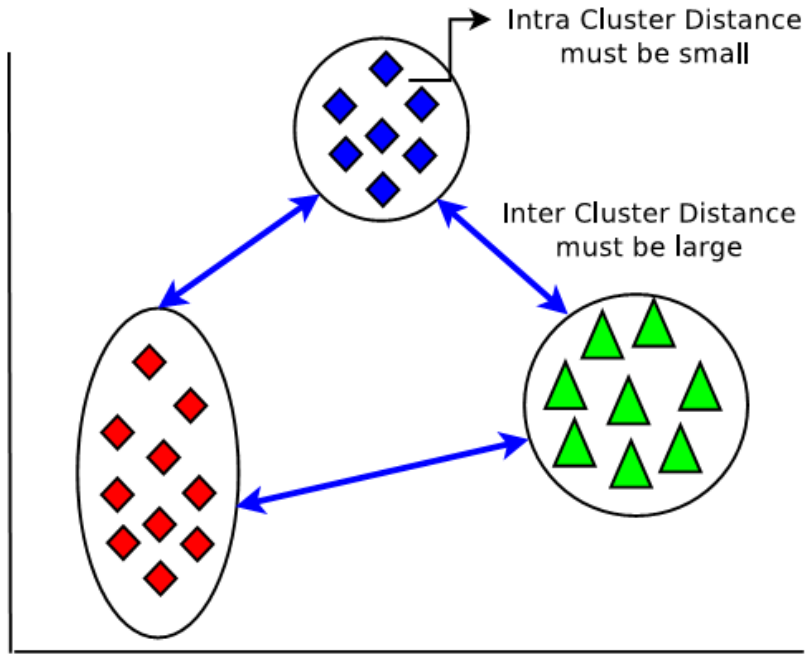


Figure: Unsupervised Learning

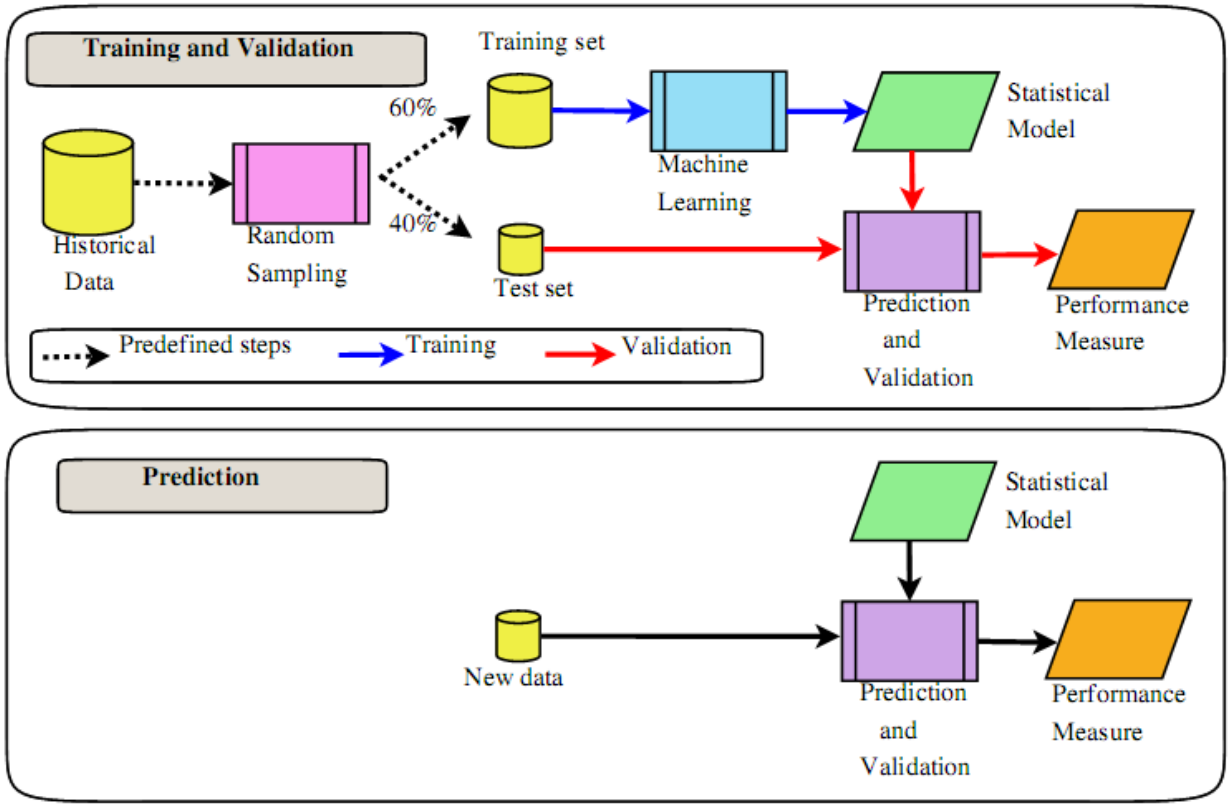
Apprentissage supervisé vs non supervisé

Plus formellement :

Task	Data: based on...	Target: learn...
▸ Supervized	$\mathcal{T} = \{(x_i, y_i)\}_{i=1..n}$	$f(x) = y$
▸ Unsupervized	$\mathcal{T} = \{x_i\}_{i=1..n}$	$x \in X_k$

Evaluation d'un modèle d'apprentissage (classifieur)

Méthode générale :



Evaluation d'un modèle d'apprentissage (classifieur)

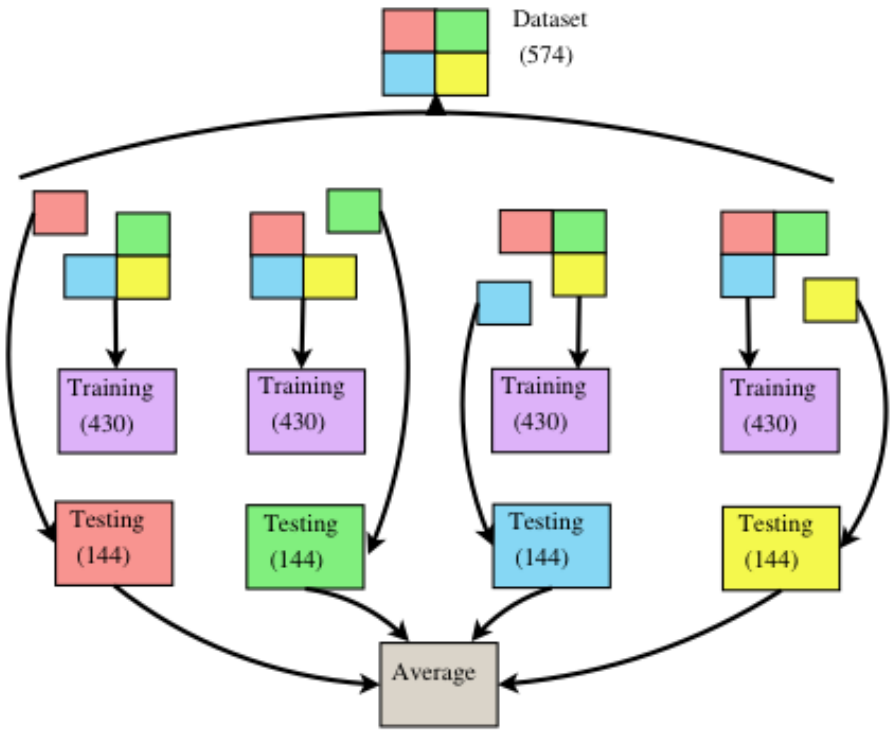
- Si l'ensemble de données est suffisamment grand :
  Découper le dataset en deux sous-ensemble : Training Set ($\sim 80\%$) et Test Set ($\sim 20\%$).

Evaluation d'un modèle d'apprentissage (classifieur)

- Si l'ensemble de données est suffisamment grand : Découper le dataset en deux sous-ensemble : Training Set ($\sim 80\%$) et Test Set ($\sim 20\%$).
- Sinon, utiliser la cross-validation.

Evaluation d'un modèle d'apprentissage (classifieur)

Cross-validation :



Evaluation d'un modèle d'apprentissage (classifieur)

Matrice de confusion (cas binaire) :

	Actual class (Observation)	
Predicted class (Expectation)	TP	FP
	FN	TN

- TP : True Positive
- FP : False Positive
- FN : False Negative
- TN : True Negative

Table: Paramètres d'évaluation.

Métriques d'évaluation

Classification binaire :

Accuracy

Il s'agit de la fraction des exemples bien classés par rapport à toutes les prédictions :

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Métriques d'évaluation

Classification binaire :

Accuracy

Il s'agit de la fraction des exemples bien classés par rapport à toutes les prédictions :

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Problème

- Si les deux classes sont fortement déséquilibrées, alors cette mesure n'apporte pas d'information pertinente ...

Métriques d'évaluation

Classification binaire :

Accuracy

Il s'agit de la fraction des exemples bien classés par rapport à toutes les prédictions :

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Problème

- Si les deux classes sont fortement déséquilibrées, alors cette mesure n'apporte pas d'information pertinente ...
  - Les prédictions sont dominées par la classe dominante
  - Mettre tous les exemples dans la classe dominante permet d'obtenir une grande accuracy

Métriques d'évaluation

Classification binaire :

Precision

C'est la fraction des vrais positifs parmi les exemples prédits comme positifs.

$$Pre = \frac{TP}{TP + FP}$$

Métriques d'évaluation

Classification binaire :

Recall ou Sensibilité

Il s'agit du rapport entre le nombre d'exemples de la classe c correctement prédits et le nombre d'exemples appartenant à la classe c.

$$Rec = \frac{TP}{TP + FN}$$

Métriques d'évaluation

Classification binaire :

F-measure

$$F_\beta = \frac{(1+\beta^2)(Pre * Rec)}{\beta^2 Pre + Rec}$$

- La précision et le recall sont complémentaires, augmenter la précision diminue le recall.
- F-measure combine les deux mesures en équilibrant les deux.
- β est un paramètre de compromis entre la précision et le recall.

Métriques d'évaluation

Classification binaire :

F_1

$$F_1 = \frac{2(Pre * Rec)}{Pre + Rec}$$

- Il s'agit de la F-measure pour $\beta = 1$.
- Cest la moyenne harmonique de la précision et du recall.

Métriques d'évaluation

Voir illustration

Métriques d'évaluation

Classification multiclasses :

$T \setminus P$	y_1	y_2	$\cdots$	y_k
y_1	n_{11}	n_{12}	$\cdots$	n_{1k}
y_2	n_{11}	n_{12}	$\cdots$	n_{1k}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
y_1	n_{k1}	n_{k2}	$\cdots$	n_{kk}

- Il s'agit d'une généralisation de la classification binaire.
- n_{ij} est le nombre d'exemples de la classe y_i prédits comme y_j.
- La diagonale contient les TP de chaque classe.
- La somme des éléments d'une colonne donne les FP pour la classe de la colonne.
- La somme des éléments d'une ligne donne les FN pour la classe de la ligne.

$$FP_i = \sum_{j \neq i} n_{ji} \quad FN_i = \sum_{j \neq i} n_{ij}$$