



Introduction au Machine Learning

UP: GL/BD

Réalisé par : Equipe ML Appliqué

1. Domaines liés à la Data Science
2. Définition de la Data Science
3. Domaines d'application
4. Définition de la Machine Learning
5. Catégories de méthodes Machine Learning
6. Les outils de data science
7. La Science de données est une démarche

- Présenter les bases de l'apprentissage automatique
- Présenter les différentes applications de l'apprentissage automatique
- Présenter les différents types de l'apprentissage automatique

1- Exemple introductif

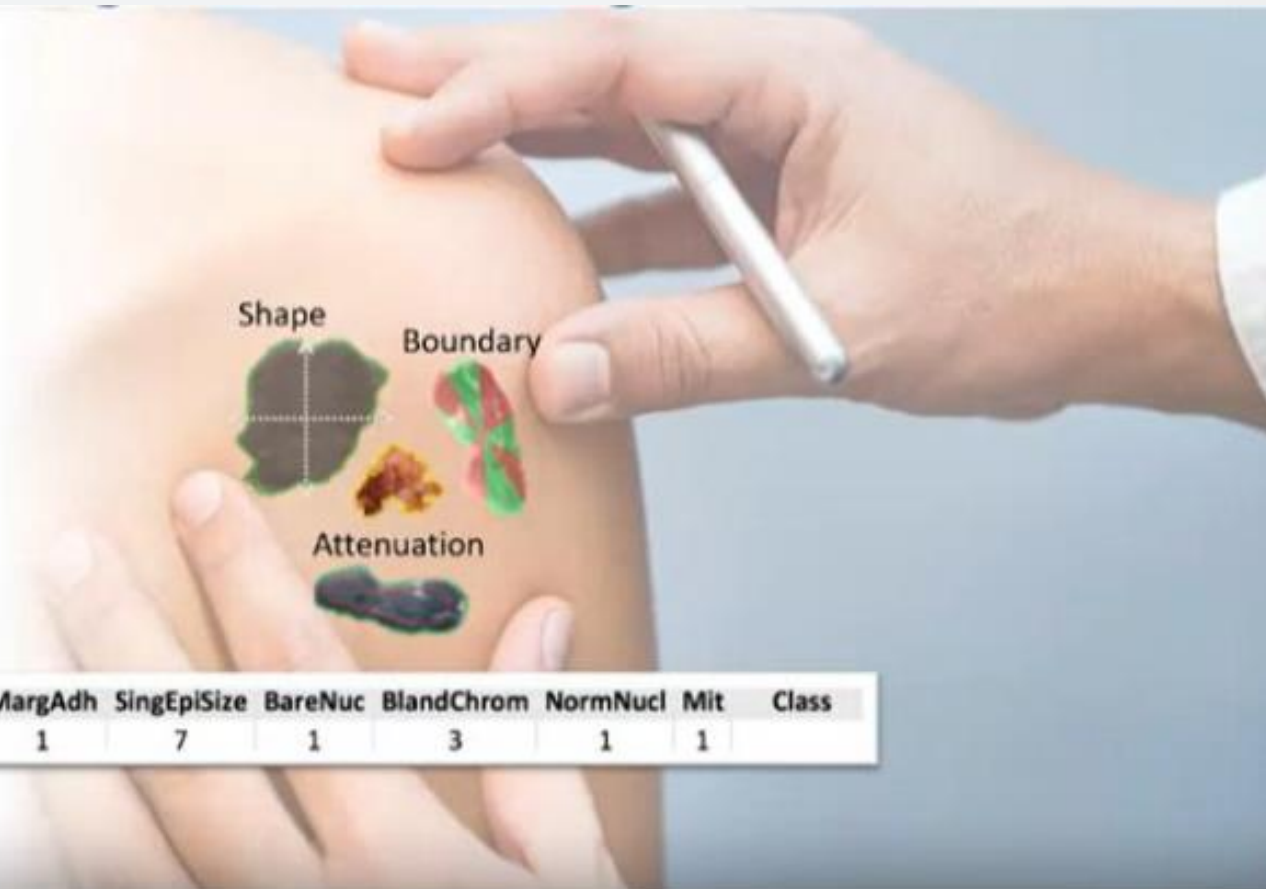
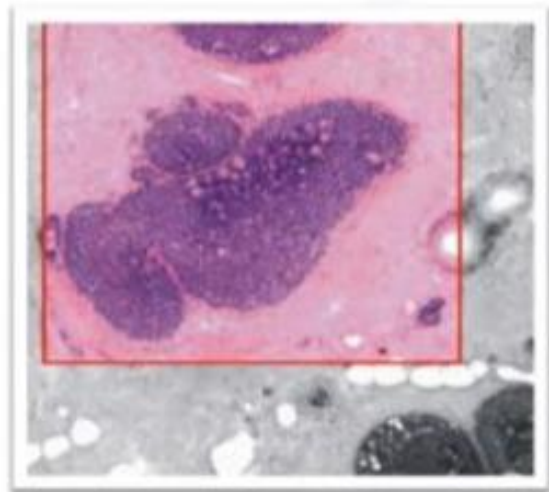
Exemple Introductif

Détection du Tumeur: Bénigne ou Maligne

- Input: Cellule Humaine
- Tache: Trouver la tumeur



Exemple Introductif



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	

Exemple Introductif



Données historiques des patients dont on soupçonne qu'ils sont atteints d'une tumeur.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Source: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%203/data/cell_samples.csv

Exemple Introductif



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Exemple Introductif



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	

=> Détecter si la tumeur est Maligne or Benigne ?

Prédiction



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign


Modeling

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	Benign

Prediction



Accuracy = 89%



2- C'est quoi
l'apprentissage
automatique?



Definition Générale :

*<< L'apprentissage automatique est la discipline donnant aux ordinateurs la capacité d'apprendre **sans qu'ils soient explicitement programmés**. >>*

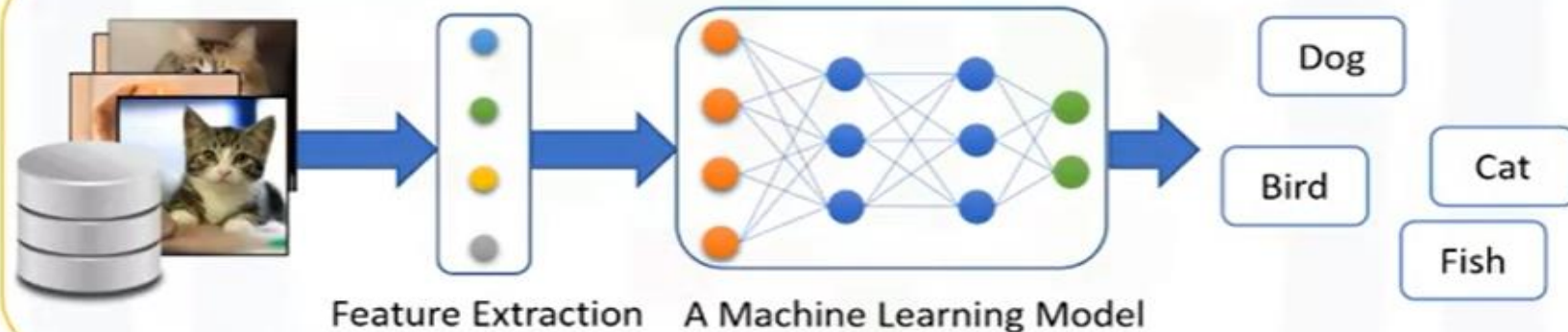
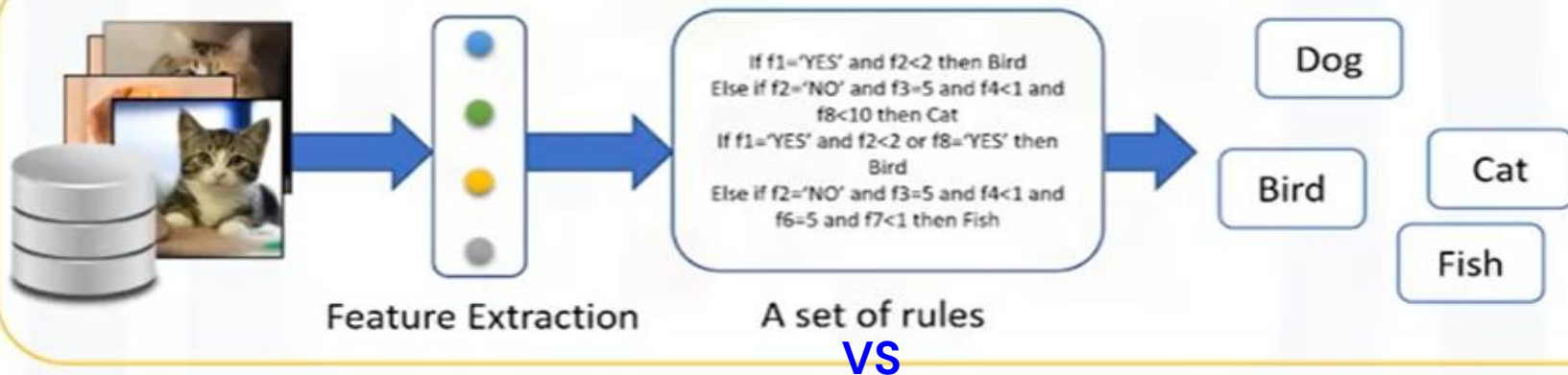
Arthur Samuel, 1959.

Une définition technique :

<< Étant donné une tâche T et une mesure de performance P , on dit qu'un programme informatique apprend à partir d'une expérience E si les résultats obtenus sur T , mesurés par P , s'améliorent avec l'expérience E . >>

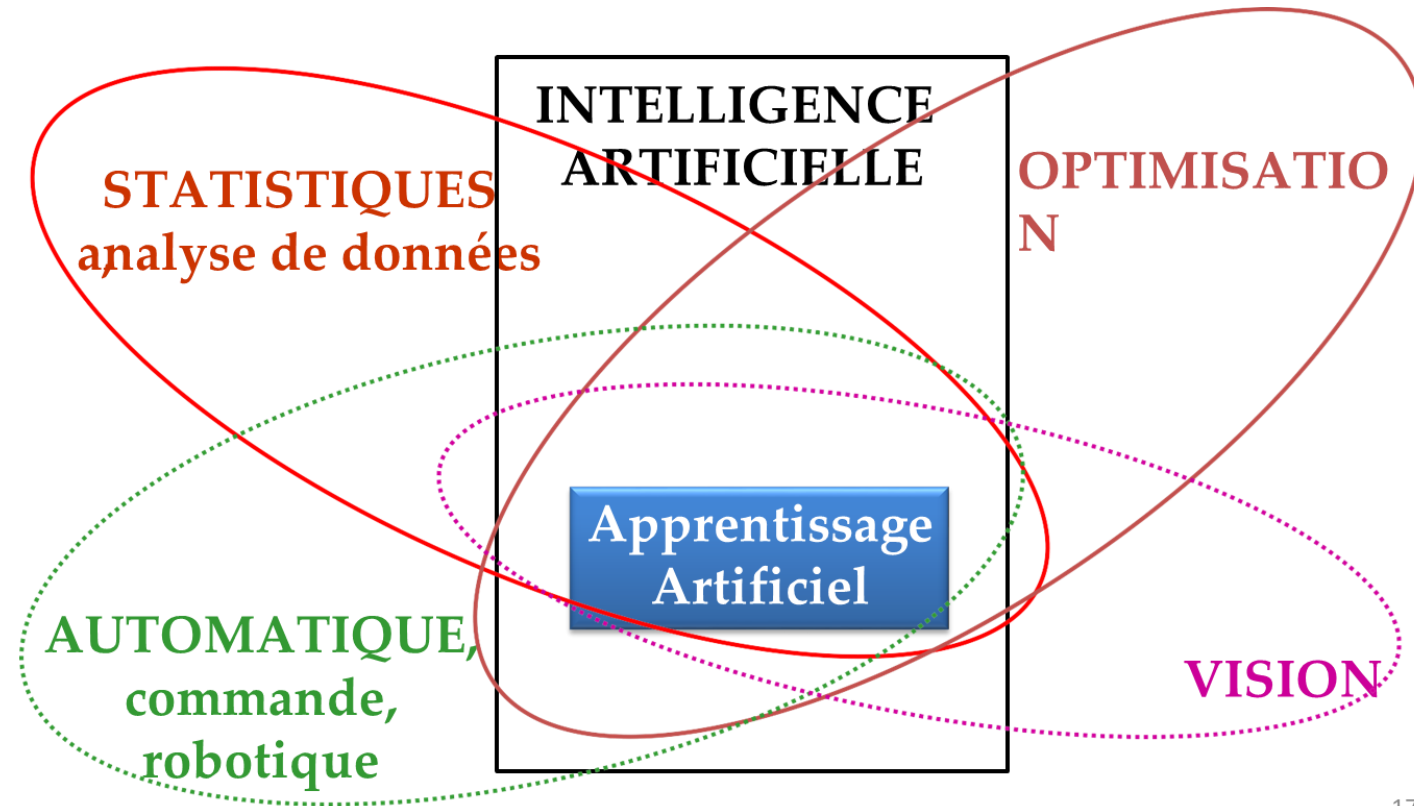
Tom Mitchell, 1997

ML vs Programmation Traditionnelle



ML vs Programmation Traditionnelle





Les applications de ML



Robotiques



Marketing



Detection de Fraude



Les applications de ML



Mail checker



Voiture autonome



Chatbots



➤ **Regression/ Estimation**

Prédire des valeurs continues (numériques)

Exemple: **prédiction du prix** d'une maison

➤ **Classification**

Prédire la classe ou la catégorie d'un cas

Exp: Customer churn/ Employee churn

➤ **Clustering**

Regroupement de groupes de cas similaires

➤ **Association**

Trouver des articles ou des évènements qui se produisent souvent

➤ **Detection des anomalies**

Détecter les cas anormaux et inhabituels

➤ **Sequence mining**

Prédire l'événement suivant(Markov, ...)

➤ **Reduction des dimensions (Feature selection)**

Réduire la taille des dataset (PCA, Selectbest, matrice de corrélation ...)

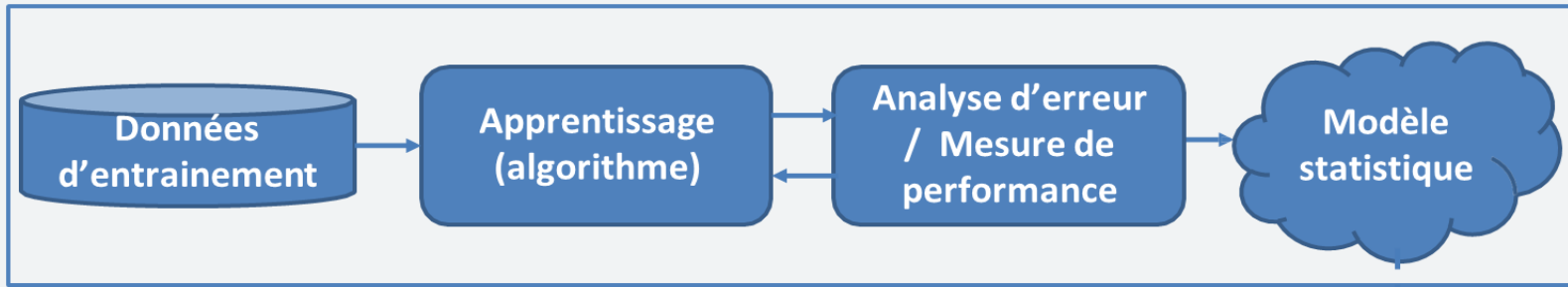
➤ **Systemes de recommandation**

Associer les préférences des gens à d'autres qui ont les goûts similaires

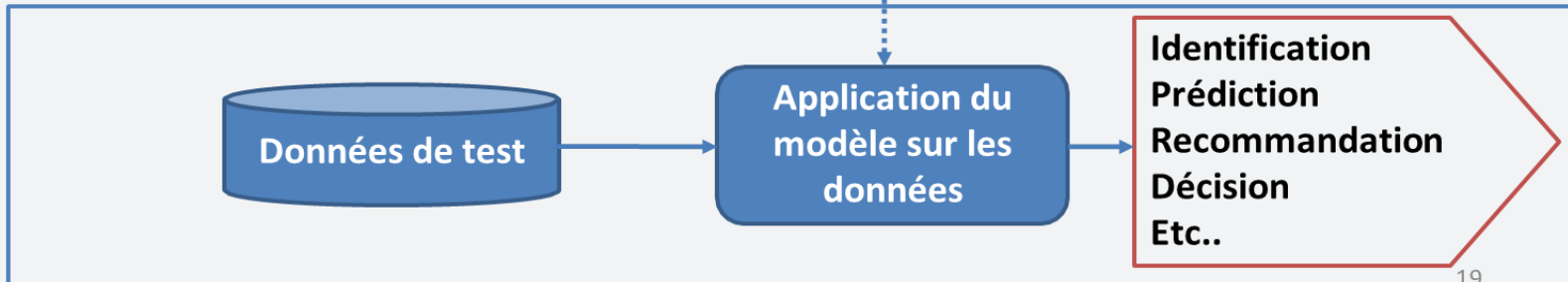
3- Le processus de l'apprentissage automatique

Le processus d'apprentissage

Phase 1 : Apprentissage



Phase 2 : Inférence/Réalisation de tâche



3-Les concepts clés de ML



Il se base sur trois composants essentiels :

- **Les données d'apprentissage** (chaque entrée dans le jeu de données est appelée un exemple, un échantillon, une instance ou une observation)
 - **La tâche à résoudre** (classifier, prédire, ordonner, etc.)
 - **La mesure de performance** (F1-score, jaccard_index,.. etc.)
- Un système apprend s'il améliore ses performances pour résoudre une tâche donnée avec le nombre d'exemples observés de cette tâche.



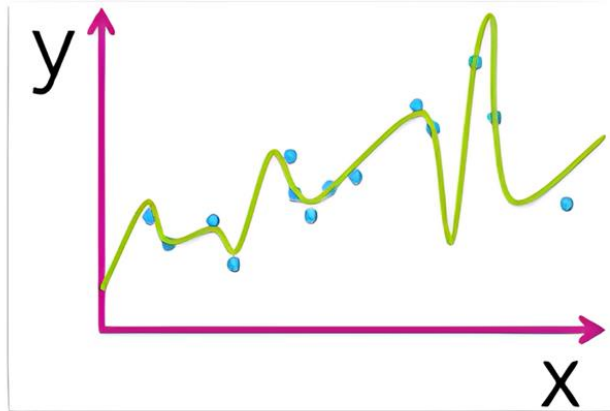
Données d'apprentissage : Réparties en 3 types :

- L'ensemble d'entraînement (training set) : constitue l'ensemble des exemples (images, textes, DB, ...) utilisés pour générer le modèle d'apprentissage.
- L'ensemble de Test (test set) : est constitué des candidats sur lesquels sera appliqué le modèle d'apprentissage (pour tester et réajuster les paramètres du modèle).
- L'ensemble de validation (validation set) : réajuster/valider les hyperparamètres (*les paramètres de l'algorithme d'apprentissage, et non pas du modèle*) (On peut lancer l'algorithme d'apprentissage plusieurs fois, en essayant à chaque fois une valeur différente pour chaque hyperparamètres)



Le sur-apprentissage (l'overfitting)

- Il s'interprète comme un apprentissage «**par cœur** » des données.
- Le modèle s'adapte bien au *TrainingSet* et si précisément qu'il ne généralise pas correctement sur les nouvelles données (capturer le bruit).
- Quand un tel événement se produit, le modèle pourra donner de bonnes prédictions sur les données du Training Set, mais **il prédira mal sur des données qu'il n'a pas encore vues** lors de sa phase d'apprentissage.

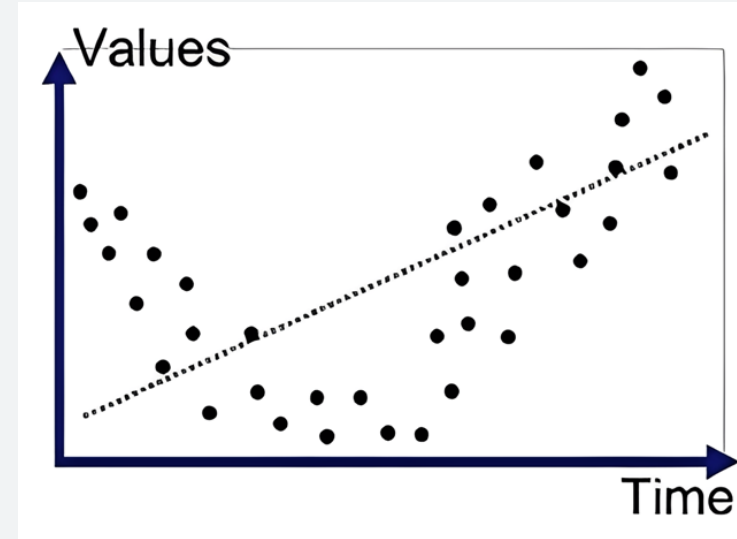




Le sous-apprentissage (L'underfitting)

- Le modèle généré lors de la phase d'apprentissage, s'adapte mal au *Training Set*.
- Quand un tel événement se produit, le modèle n'arrive pas à capturer les corrélations du *Training Set*. Par conséquent, le **coût d'erreur en phase d'apprentissage** reste grand.

✿ le modèle ne se généralise pas sur les données qu'il n'a pas encore vu.





Pour résoudre l'Overfitting :

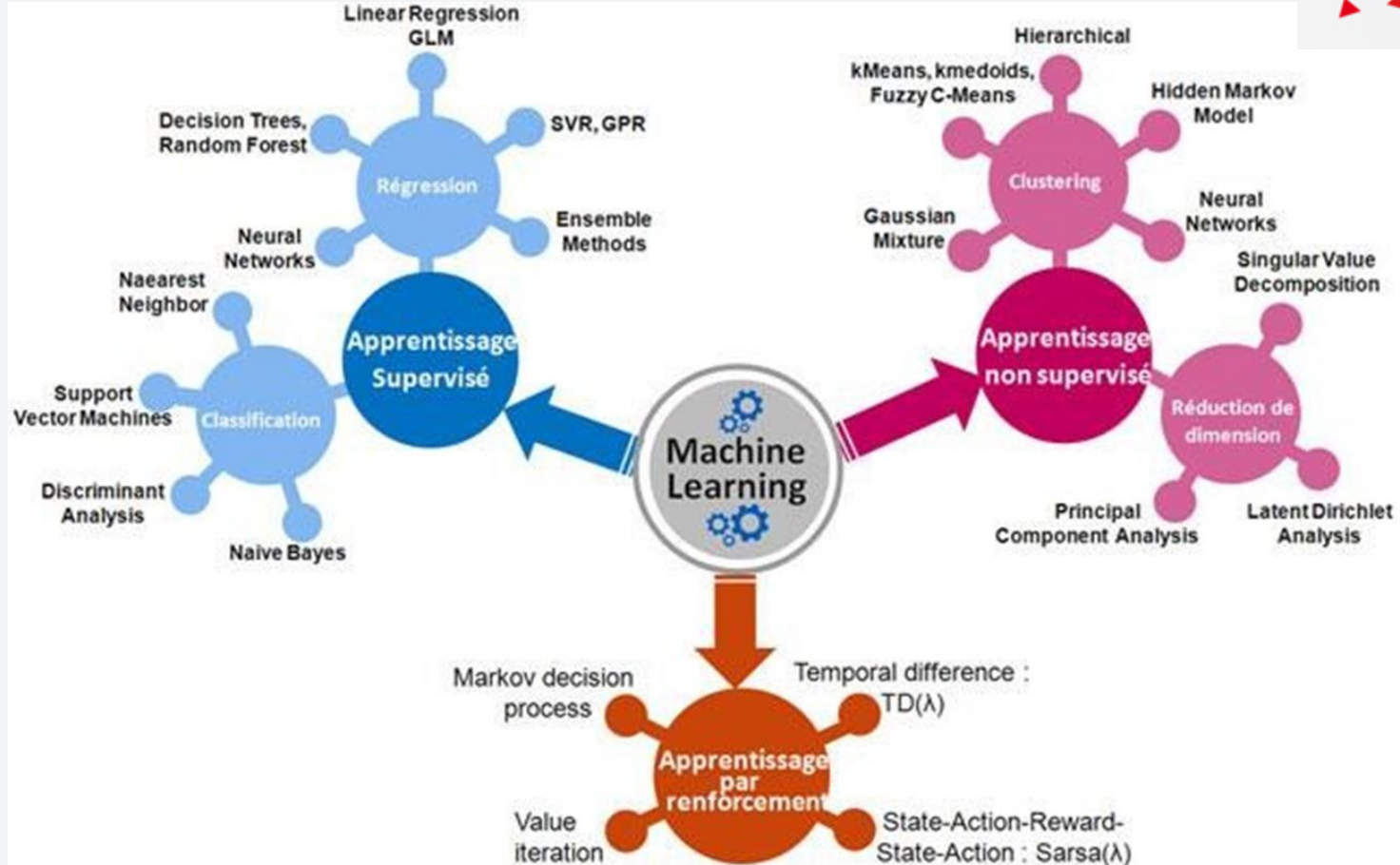
- Régularisation : Utiliser L1 (Lasso) ou L2 (Ridge) pour pénaliser les coefficients élevés.
- Simplifier le modèle : Réduire la complexité en utilisant moins de paramètres ou une architecture plus simple.
- Augmentation des données : Générer plus de données en augmentant les exemples (ex. rotation ou recadrage d'images).
- Validation croisée : Évaluer les performances du modèle sur plusieurs sous-ensembles des données.



Pour résoudre l'Underfitting :

- Modèles plus complexes : Utiliser des algorithmes capables de capturer des relations non linéaires (ex. arbres de décision, réseaux neuronaux).
- Augmenter la durée d'entraînement : Former le modèle pendant plus d'itérations.
- Caractéristiques supplémentaires : Ajouter de nouvelles variables ou utiliser des transformations des données existantes.

Les types de ML





Apprentissage supervisé

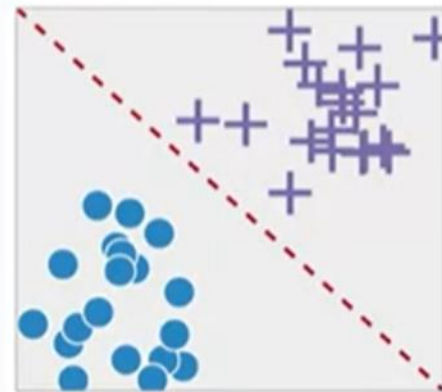
Il permet de construire des modèles à partir des exemples d'apprentissage dont on connaît les **étiquettes** (labels)

- **Classification:**

Classification est la prédiction d'une valeur discrète

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benien

Categorical Values



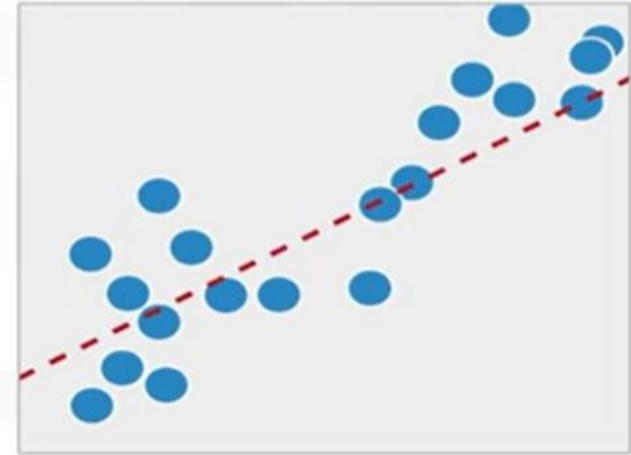
Les types de ML

- Regression

Regression est la prédiction d'une valeur continue

	ENGINE SIZE	CYLINDERS	FUEL CONSUMPTION_COMB	CO2 EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values





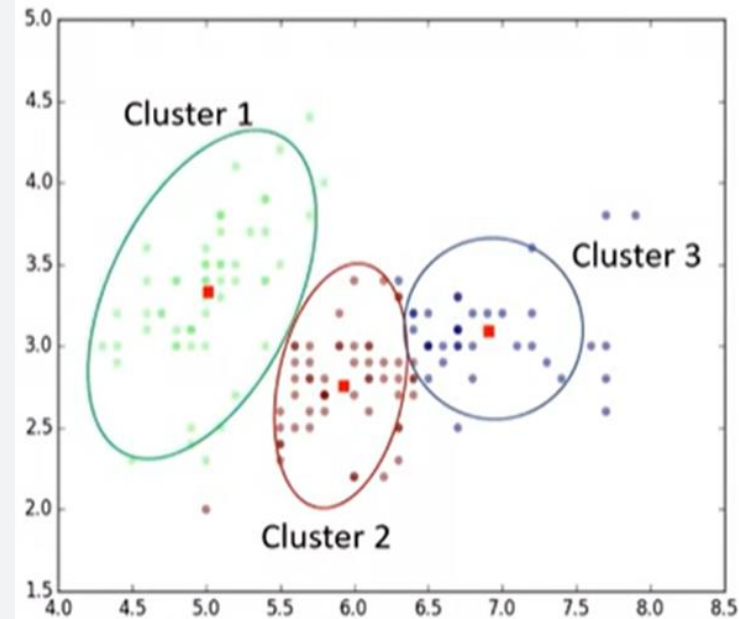
Apprentissage non supervisé

L'algorithme doit découvrir par **lui-même** la structure des données (patterns, informations cachées,...).

L'idée est donc de *découvrir des clusters, des groupes au sein des données,*

Qu'est ce qu'un Cluster?

Un cluster (ou groupe) est un ensemble formé par des données **homogènes** qui "se ressemblent" au sens d'un critère de similarité (distance, densité de probabilité, etc.).



Les types de ML

Apprentissage non supervisé



Unsupervised learning techniques:

- Dimension reduction
- Density estimation
- Market basket analysis
- Clustering



ALL OF THIS DATA
IS UNLABELED

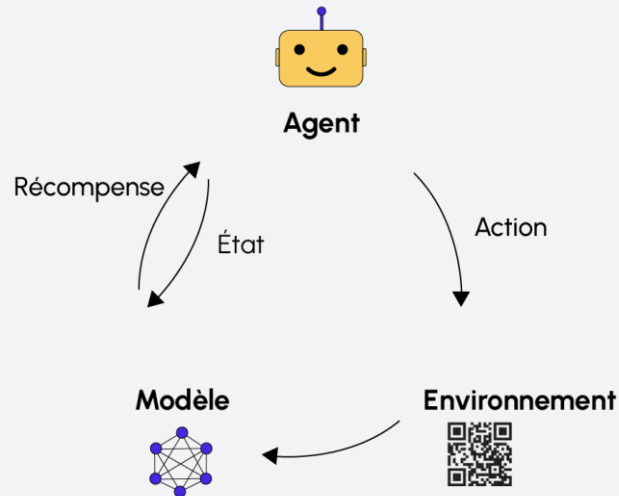
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6



Apprentissage par renforcement

L'apprentissage par renforcement (RL) est une technique de machine learning (ML) qui entraîne les logiciels à **prendre des décisions** en vue d'obtenir les meilleurs résultats.

Les algorithmes de RL utilisent un paradigme de récompense et de punition lorsqu'ils traitent les données.



Apprentissage par renforcement

Principe :

L'agent prend une action dans l'environnement.

L'environnement renvoie une récompense (positive ou négative).

L'objectif est de maximiser les récompenses cumulées sur le long terme.

Exemples d'applications :

- Jeux vidéo (AlphaGo, Dota 2).
- Robots autonomes.
- Optimisation de la gestion de trafic.
- Systèmes de trading automatique.

1. Biais dans les résultats

Défi : Les biais présents dans les données d'entraînement peuvent entraîner des discriminations.

Exemple : Un modèle d'approbation de prêt discrimine selon le genre ou l'origine ethnique.

Solution :

- Analyser et nettoyer les données pour éliminer les biais.
- Tester les modèles sur des ensembles de données diversifiés et représentatifs.

2. Erreurs pouvant causer des dommages

Défi : Les erreurs des systèmes d'IA peuvent avoir des conséquences graves.

Exemple : Un véhicule autonome provoque une collision à cause d'une défaillance du système.

Solution :

- Effectuer des tests rigoureux avant déploiement.
- Implémenter des mécanismes de sécurité pour limiter les impacts en cas de panne.

3. Exposition des données

Défi : Les systèmes d'IA peuvent mettre en danger les données sensibles des utilisateurs.

Exemple : Des diagnostics médicaux utilisent des données de patients stockées de manière peu sécurisée.

Solution :

- Utiliser des techniques de chiffrement pour protéger les données.
- Respecter les réglementations sur la protection des données, comme le RGPD.

4. Solutions non adaptées pour tous

Défi : Certaines solutions d'IA ne prennent pas en compte les besoins spécifiques de tous les utilisateurs.

Exemple : Un assistant domestique sans sortie audio adaptée aux malvoyants.

Solution :

- Inclure des tests d'accessibilité pendant la conception.
- Développer des fonctionnalités inclusives pour divers utilisateurs.



5. Complexité et confiance

Défi : Les utilisateurs doivent faire confiance à des systèmes complexes qu'ils ne comprennent pas.

Exemple : Un outil financier basé sur l'IA propose des recommandations d'investissement sans explication claire.

Solution :

- Améliorer la transparence des systèmes avec des explications claires.
- Fournir des outils pédagogiques pour que les utilisateurs comprennent les recommandations.

6. Responsabilité légale

Défi : Déterminer qui est responsable en cas d'erreur ou d'impact négatif lié à l'IA.

Exemple : Une personne condamnée à tort sur la base d'une reconnaissance faciale incorrecte.

Solution :

- Mettre en place des cadres légaux et éthiques clairs.
- Assurer une supervision humaine pour les décisions critiques.

- ✓ Équité
- ✓ Fiabilité et sécurité
- ✓ Respect de la vie privée
- ✓ Inclusivité
- ✓ Transparence
- ✓ Responsabilité

Thank you for your attendance !

You can find me at: jihene.hlel@esprit.tn