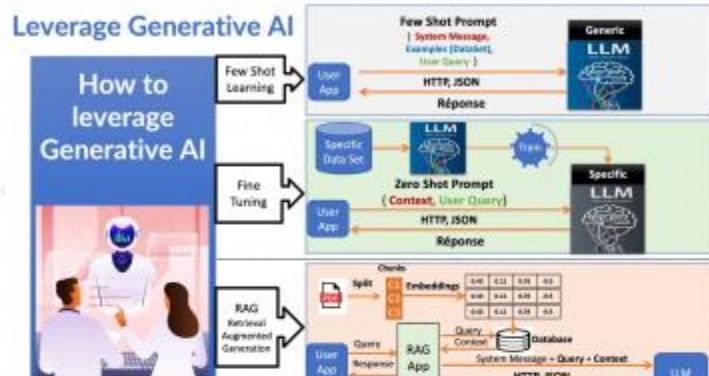


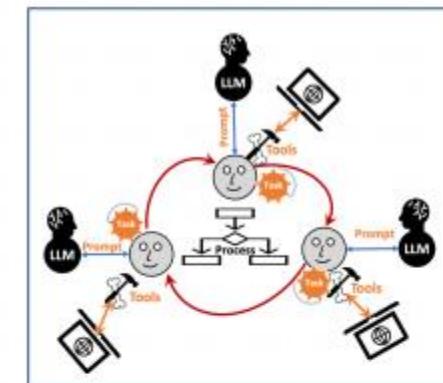
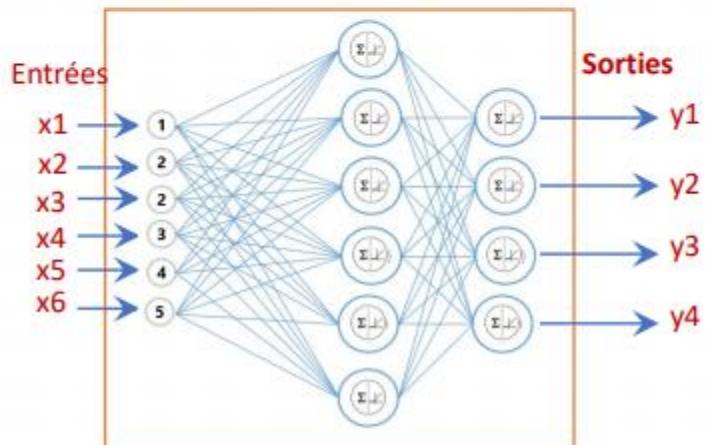
Cours de Machine Learning

BAC4

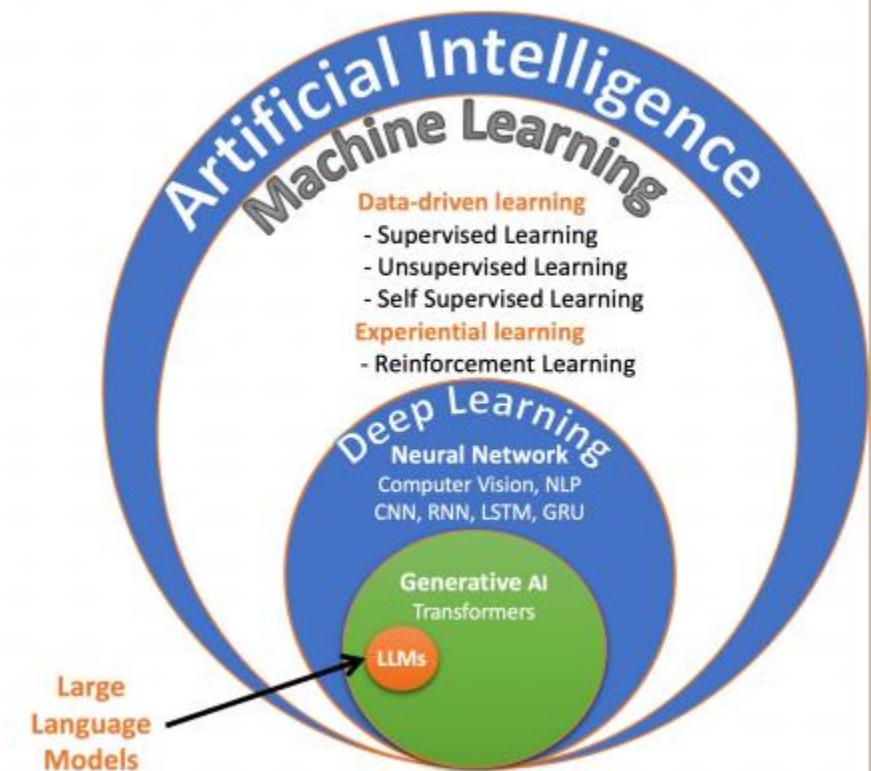
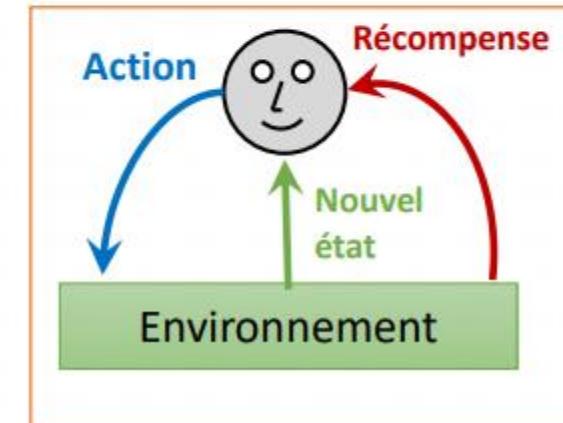
Edouard Ngoy Mushame, Ir. Concepteur des Systèmes d'information
Chercheur en Génie Logiciel, Bases de données et Intelligence Artificielle
Doctorant en Génie Logiciel Industriel



Apprentissage Supervisé



Apprentissage Par Renforcement



Objectifs du cours

Le but de ce cours est de vous fournir des bases solides sur les concepts et les algorithmes de ce domaine en plein essor. Il vous aidera à identifier les problèmes qui peuvent être résolus par une approche machine learning, à les formaliser, à identifier les algorithmes les mieux adaptés à chaque cas, à les mettre en œuvre, et enfin à savoir évaluer les résultats obtenus.

- identifier les problèmes qui peuvent être résolus par des approches de machine learning ;
- formaliser ces problèmes en termes de machine learning ;
- identifier les algorithmes classiques les plus appropriés pour ces problèmes et les mettre en œuvre ;
- implémenter ces algorithmes par vous-même afin d'en comprendre les tenants et aboutissants ;
- évaluer et comparer de la manière la plus objective possible les performances de plusieurs algorithmes de machine learning pour une application particulière.

Pre-requis

- Algèbre linéaire (inversion de matrice, théorème spectral, décomposition en valeurs propres et vecteurs propres).
- Notions de probabilités (variable aléatoire, distributions, théorème de Bayes).

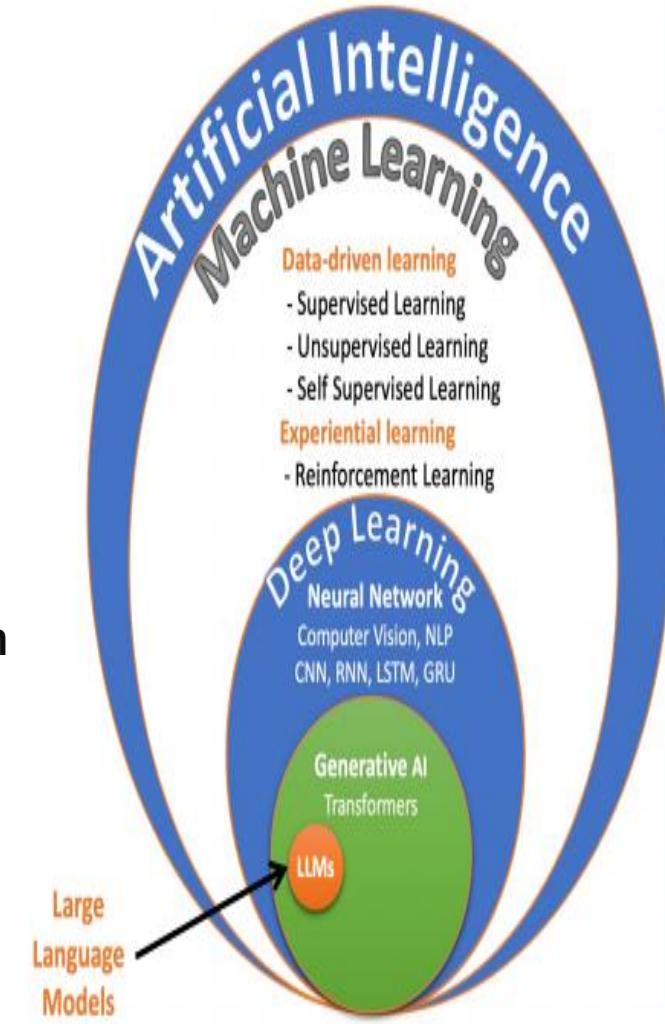
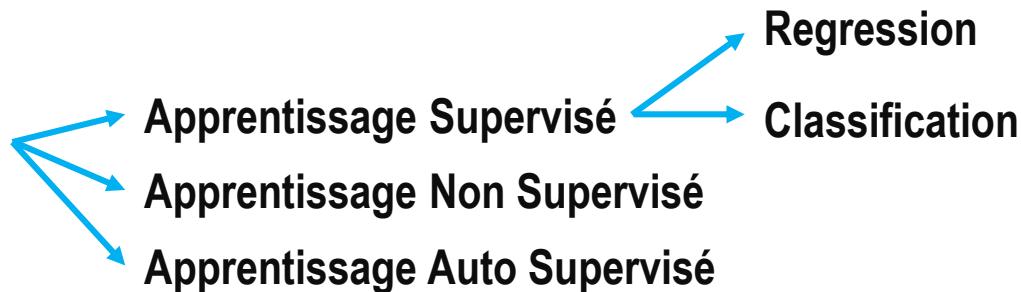
Contenu

- Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning
- Chapitre 2 : Apprentissage supervisé
- Chapitre 3 : Apprentissage non supervisé

Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

Intelligence Artificielle

- L'intelligence artificielle
- Intelligence Artificielle Distribuée =
 - IA : Pour des agents intelligent (Modéliser le savoir et le comportement)
 - + Distribuée : Modéliser leurs interactions => **Intelligence Collective**
- IA = **IA symbolique (5%) + Machine Learning (95%)**
- Techniques d'apprentissage :
 - Piloté par les données
 - Piloté par l'expérience
- Deep Learning
- IA Generative



Le Machine Learning

- **Qu'est ce que le machine learning ?**
 - Une définition qui s'applique à un programme informatique comme à un robot, un animal de compagnie ou un être humain est celle proposée par Fabien Benureau (2015) : « L'apprentissage est une modification d'un comportement sur la base d'une expérience »
- **Pourquoi utiliser les machines learning ?**
 - Le machine learning peut servir à résoudre des problèmes
 - que l'on ne sait pas résoudre (comme dans l'exemple de la prédition d'achats ci-dessus) ;
 - que l'on sait résoudre, mais dont on ne sait formaliser en termes algorithmiques comment nous les résolvons (c'est le cas par exemple de la reconnaissance d'images ou de la compréhension du langage naturel) ;
 - que l'on sait résoudre, mais avec des procédures beaucoup trop gourmandes en ressources informatiques (c'est le cas par exemple de la prédition d'interactions entre molécules de grande taille, pour lesquelles les simulations sont très lourdes). Le machine learning est donc utilisé quand les données sont abondantes (relativement), mais les connaissances peu accessibles ou peu développées.

Fondements du machine learning

- Le machine learning se base fondamentalement sur deux piliers :
 - D'une part les données
 - D'autre part les algorithmes d'apprentissage (**entraînement**)

Ces deux piliers sont aussi importants l'un que l'autre. D'une part, aucun algorithme d'apprentissage ne pourra créer un bon modèle à partir de données qui ne sont pas pertinentes – c'est le concept garbage in, garbage out qui stipule qu'un algorithme d'apprentissage auquel on fournit des données de mauvaise qualité ne pourra rien en faire d'autre que des prédictions de mauvaise qualité. D'autre part, un modèle appris avec un algorithme inadapté sur des données pertinentes ne pourra pas être de bonne qualité.

Cet cours est consacré au deuxième de ces piliers – les algorithmes d'apprentissage. Néanmoins, il ne faut pas négliger qu'une part importante du travail de machine learner ou de data scientist est un travail d'ingénierie consistant à préparer les données afin d'éliminer les données aberrantes, gérer les données manquantes, choisir une représentation pertinente, etc.

Attention !

- Bien que l'usage soit souvent d'appeler les deux du même nom, il faut distinguer **l'algorithme d'apprentissage automatique** du **modèle appris** : le premier utilise les données pour produire le second, qui peut ensuite être appliqué comme un programme classique.

Exemple

- Voici quelques exemples de reformulation de problèmes de machine learning sous la forme d'un problème d'optimisation. La suite de cet ouvrage devrait vous éclairer sur la formalisation mathématique de ces problèmes, formulés ici très librement.
- un vendeur en ligne peut chercher à modéliser des types représentatifs de clientèle, à partir des transactions passées, en maximisant la proximité entre clients et clientes affectés à un même type ;
- une compagnie automobile peut chercher à modéliser la trajectoire d'un véhicule dans son environnement, à partir d'enregistrements vidéo de voitures, en minimisant le nombre d'accidents ;
- des chercheurs en génétique peuvent vouloir modéliser l'impact d'une mutation sur une maladie, à partir de données patient, en maximisant la cohérence de leur modèle avec les connaissances de l'état de l'art ;
- une banque peut vouloir modéliser les comportements à risque, à partir de son historique, en maximisant le taux de détection de non solvabilité.

Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

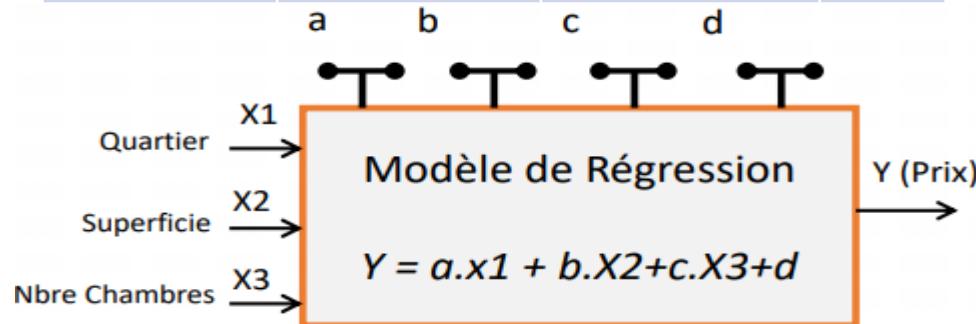
L'apprentissage automatique, ou machine Learning, est une discipline dont les outils puissants permettent aujourd'hui à de nombreux secteurs d'activité de réaliser des progrès spectaculaires grâce à l'exploitation de grands volumes de données.

- Apprentissage Supervisé

- **Régression** : Exemple : Prédiction du prix d'un appartement

Data Set(Donnees etiquetees) Input			Output
Quartier – x1	Superficie – x2	Nbre chambres x3	Prix - y
Biashara	100	3	500
GCM Kolwezi	50	3	150
Industriel	120	4	700
Joli Site	80	3	200
Golf ISTM	200	5	?????

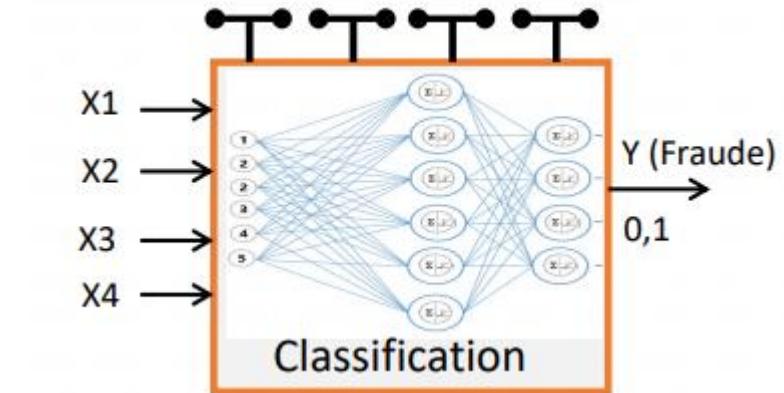
Modèle à 4 Paramètres : a, b, c et d
Algorithmes : Linéaire, Polynomiale, Ridge, Lasso, etc.



Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

- Apprentissage Supervisé
 - Classification
 - Exemple : Prédire si une transaction financière est frauduleuse ou non

Data Set(Donnees etiquetees) Input				Output
Heure (X1)	Montant (X2)	Long (X3)	Lat (X4)	Fraude (Y)
12:20	4500	-1.2	3.2	0
01:45	1.60	0.6	4.3	1
10:08	100	-2.45	1.3	0
11:55	80	-1.2	3.2	1
00:00	3200	-1	3	?????

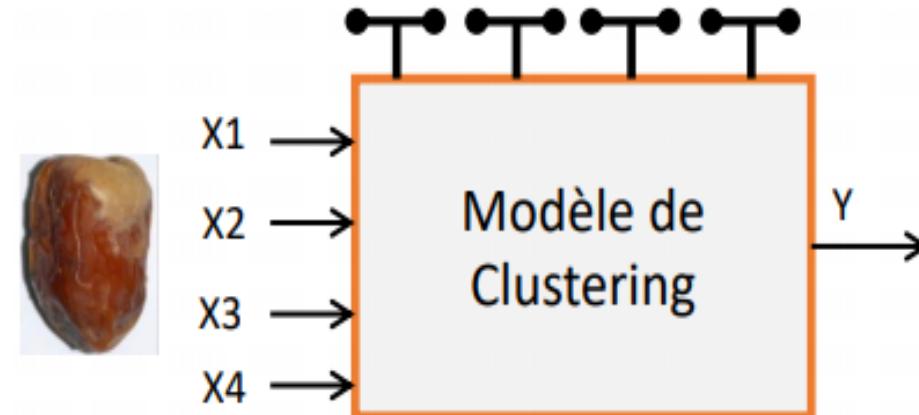


Algorithmes : KNN, SVM, DT, RF, Réseaux de neurons, etc.

Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

- Apprentissage Non Supervisé
 - Clustering

Data Set(Donnees non etiquetees)



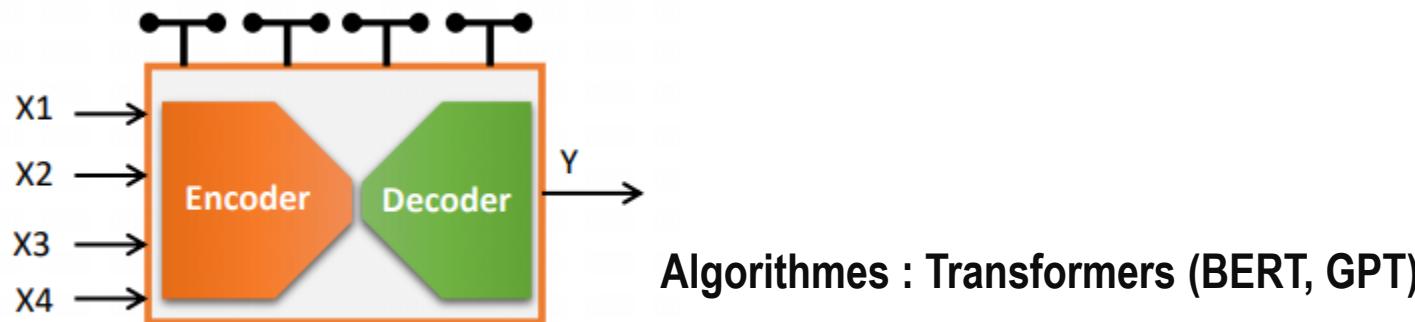
Algorithmes : K-Means

Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

- Apprentissage Auto Supervisé – Self Supervised Learning
 - **Encodeur = Décodeur**

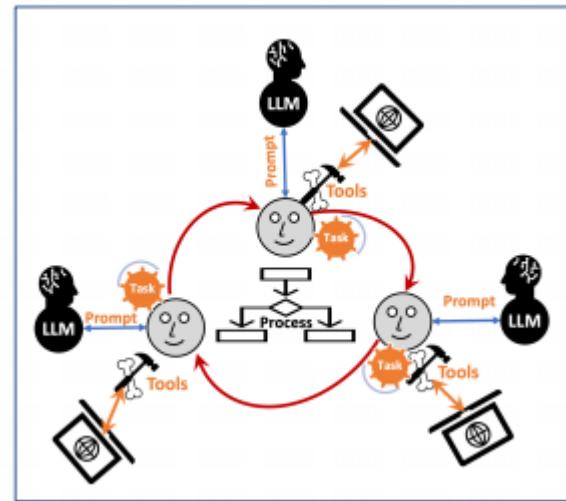
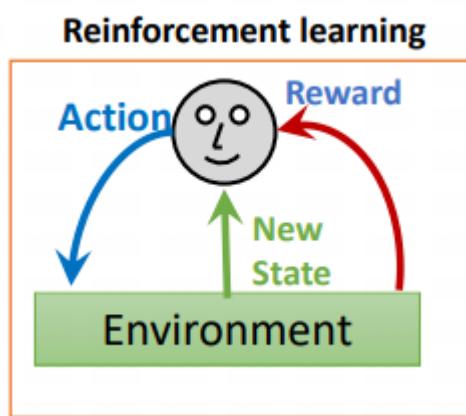
Data Set (Données Non étiquetées)

Cacher une partie des données et entraîner le modèle à prédire ces parties cachées en utilisant le principe des encodeurs et décodeurs Par exemple en NLP, On cache des mots du texte et on entraîne le modèle pour deviner ces mots cachés



Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

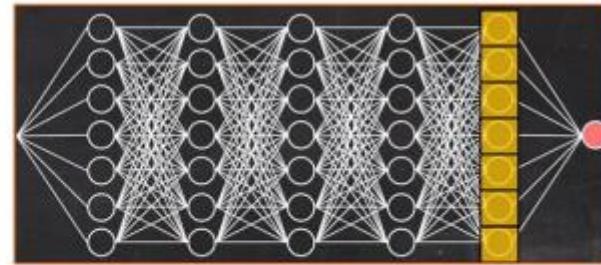
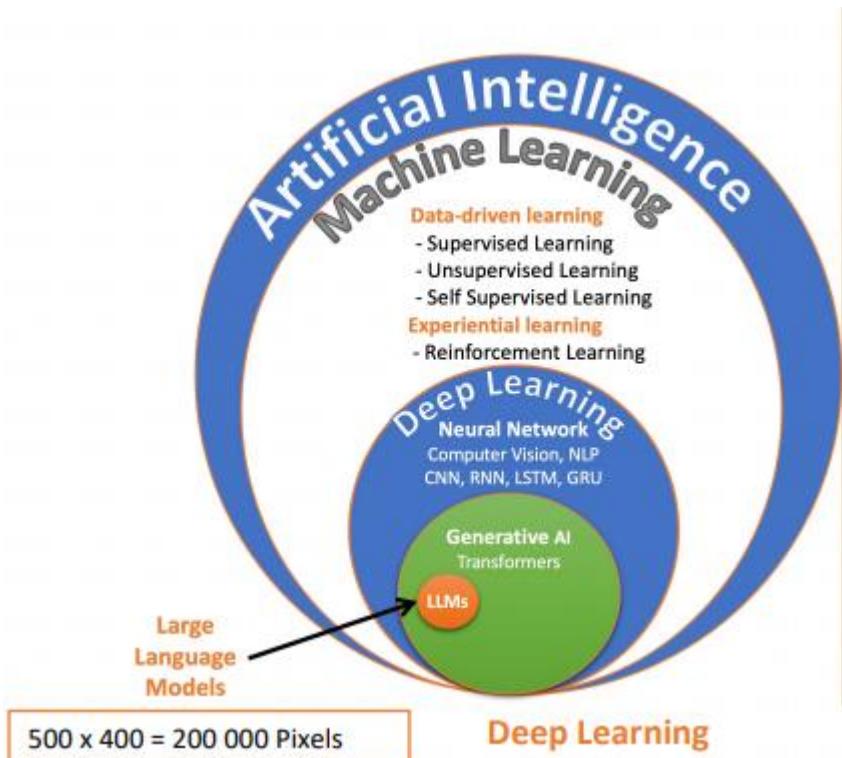
- Apprentissage par Renforcement – Reinforcement Learning
 - Embodiment



Algorithmes : Q-Learning, PPO

Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

- Deep Learning

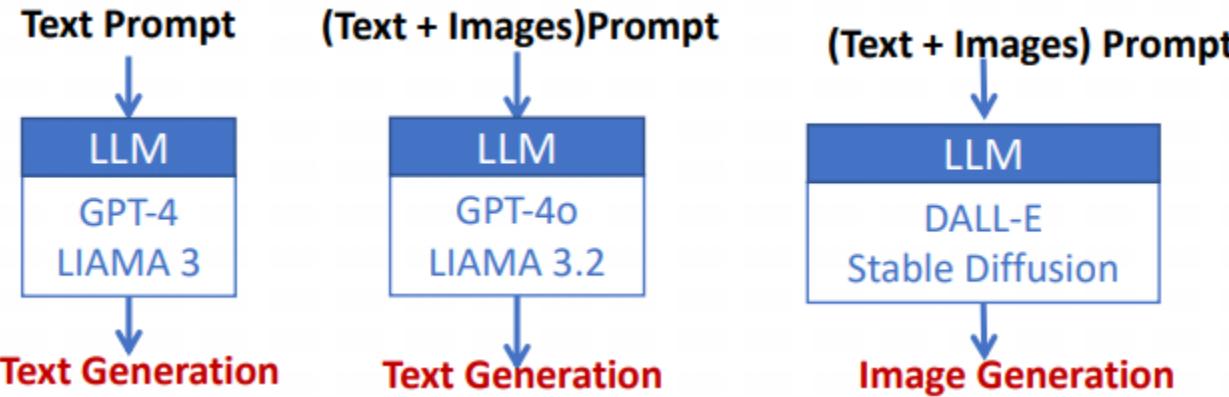


CNN : [Convolution, RELU, MAX PULLING] [Fully Connected]

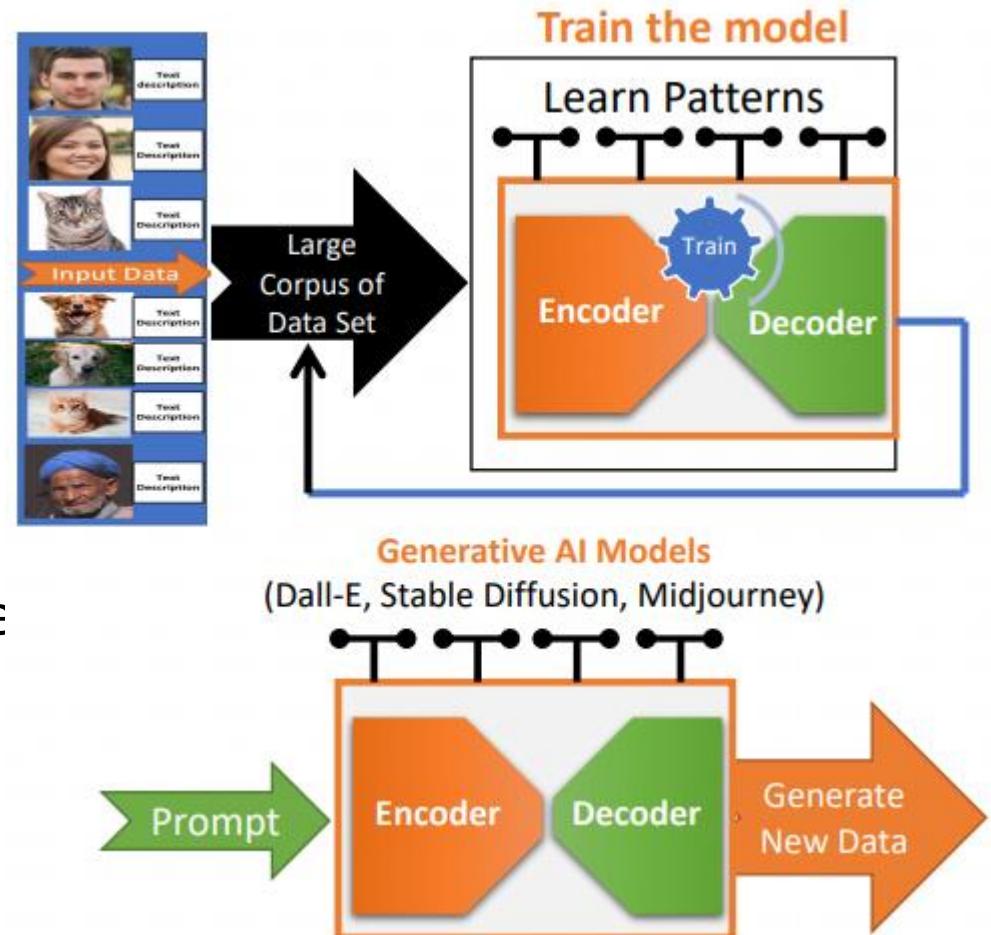
RNN : [Recurrent Neural Network]

Chapitre 1 : Introduction à l'intelligence artificielle et au Machine Learning

- IA Generative



L'IA Générative est une catégorie de l'IA capable de générer des nouvelles données en conséquence d'un raisonnement conditionné par un prompt utilisateur en utilisant les LLMs (Large Language Models).



Concepts fondamentaux de l'IA

- L'intelligence artificielle
- IA = IA symbolique (5%) + Machine Learning (95%)
- Techniques d'apprentissage :
 - Piloté par les données
 - Apprentissage Supervisé
 - Apprentissage Non Supervisé
 - Apprentissage Auto Supervisé
 - Piloté par l'expérience (Apprentissage par Renforcement)
- Intelligence Artificielle Distribuée =
 - IA : Pour des agents intelligent (Modéliser le savoir et le comportement)
 - + Distribuée : Modéliser leurs interactions => Intelligence Collective
- Generative AI
- Large Language Models (LLMs) : Transformers (GPT, BERT)

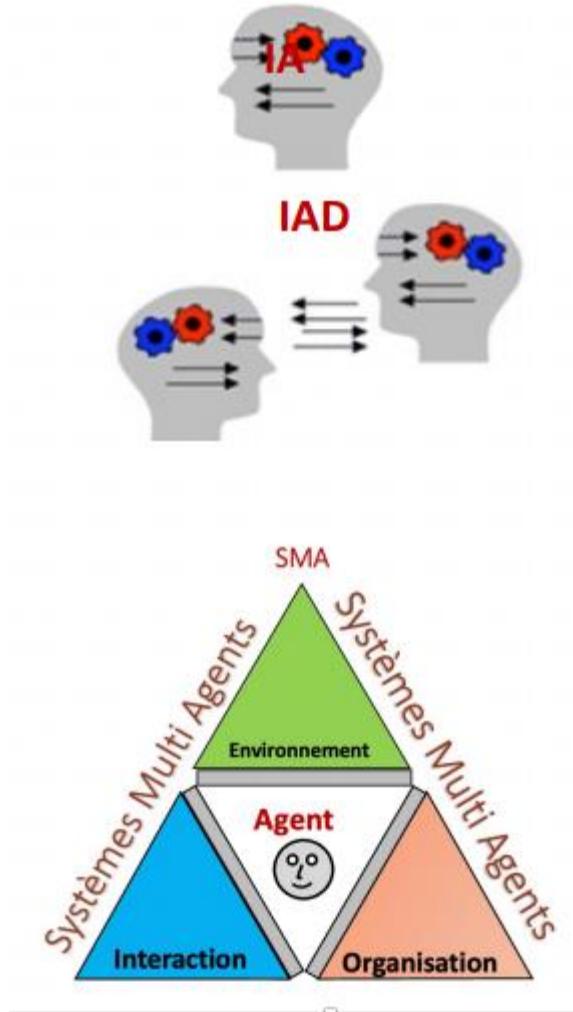
Regression

Classification



Intelligence Artificielle Distribuée & AI Agents

- L'intelligence artificielle est une discipline qui cherche doter les systèmes informatiques avec des capacités intellectuelles semblables à celle des êtres humains et des animaux en utilisant des algorithmes
- IA et IAD :
 - L'Intelligence Artificielle permet de modéliser un penseur isolé en exploitant l'intelligence individuelle d'un agent.
 - L'Intelligence Artificielle Distribuée permet d'exploiter l'intelligence collective de plusieurs agents qui vont collaborer selon une planification et une organisation pour participer ensemble à résoudre des problèmes complexes. Cette discipline est connue par les Systèmes Multi Agents. Un Agent est une entité autonome entraîné pour prendre des décisions pour atteindre un but pour lequel il a été créé.



IA Symbolique et Machine Learning

- Il existe deux **familles d'algorithmes de l'IA** :
 - **I'IA Symbolique**: Des techniques qui n'ont pas pu décoller et donc rarement utilisées aujourd'hui dans les applications de l'IA. Elle couvre moins de 5% des applications industrielles.
 - **Machine Learning** (Apprentissage automatique) (95%): Ce sont les techniques qui sont les plus exploitées aujourd'hui dans les applications de l'IA

Machine Learning

- Dans le domaine du Machine Learning, il existe deux façons pour entraîner les algorithmes :
 - *Apprentissage piloté par les données*
 - *Apprentissage piloté par l'expérience (Apprentissage par renforcement)*
- **Apprentissage piloté par les données :**
 - Consiste à entraîner des algorithmes à effectuer des prédictions en utilisant un ensemble de données collectées du domaine réel étudié.
 - Il existe trois types d'apprentissage pilotés par les données** :
 - *Apprentissage piloté par l'expérience (Apprentissage supervisé)*
 - *Apprentissage non supervisé (Clustering)*
 - *Apprentissage Auto Supervisé (Self Supervised Learning)*
- **Apprentissage par renforcement)**

Apprentissage Supervisé

- Consiste à utiliser un data set étiqueté. C'est-à-dire des données dont on connaît les inputs et les outputs. Les outputs représentent des étiquettes ou des valeurs fournis par les experts du métier.
- Dans l'apprentissage supervisé on distingue deux types de problèmes :
 - **Régression** : Consiste à prédire, en sortie, une valeur continue comme le prix d'un appartement ou la durée de vie d'une pièce mécanique ou la durée de guérison d'un patient : Exemple Régression Linéaire
 - **Classification** : Consiste à prédire des classes d'appartenance parmi un ensemble de classe finies. Par exemple prédire si une transaction est frauduleuse ou encore prédire si un animal est un chien, un chat, un tigre ou un lapin.
 - Exemples d'algorithmes : Régression logistique, Support Vector Machine (SVM), Multi Layer Perceptron (MLP), KNN (K plus proches voisins), Decision Tree, Random Forest, etc.

Apprentissage Non Supervisé

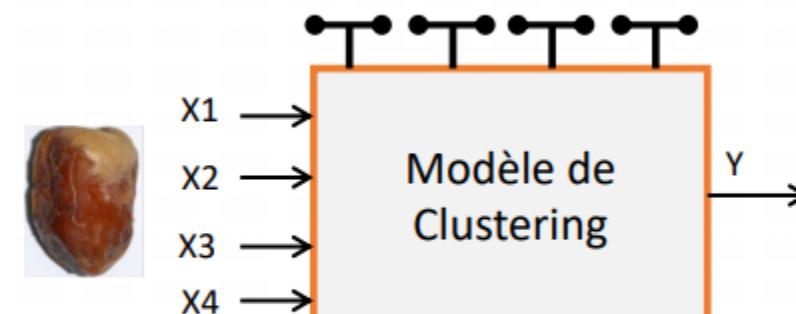
- Apprentissage Non Supervisé :

- Consiste à utiliser des données non étiquetées et utiliser des algorithmes de clustering qui vont fouiller dans les données pour chercher à les segmenter en un ensemble de groupes homogènes en se basant sur des mesures de similarités.
- Exemple : KMeans

Apprentissage Non Supervisé

Clustering

Data Set (Données Non étiquetées)



Algorithmes : K-Means

Apprentissage Auto Supervisé

- Apprentissage Auto Supervisé :
 - Utilisé dans le domaine de NLP (Natural Language Processing),
 - Ce type d'apprentissage consiste à exploiter des données non structurées et non étiquetées comme le texte d'un livre.
 - Pendant l'entraînement d'un algorithme, on cache des parties du texte et puis on entraîner le modèle à prédire les parties cachées du texte en utilisant des encodeurs et des décodeurs qui sont basés sur un mécanisme d'attention qui est la base des Transformers qui sont les plus utilisés dans le domaine de l'IA générative sur les LLMs (Large Language Models) : Exemple BERT, GPT

Apprentissage Par renforcement

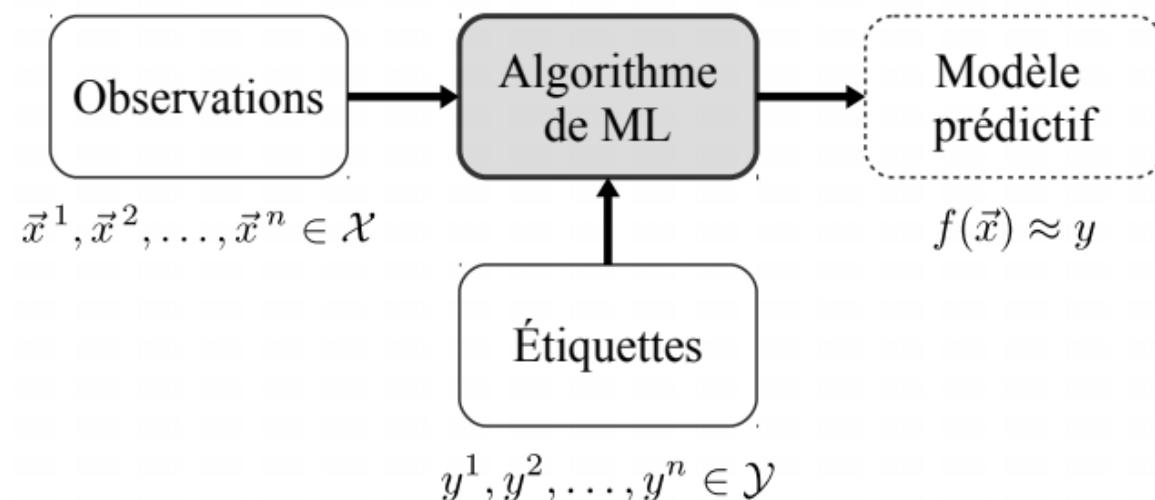
- **Apprentissage piloté par l'expérience (Apprentissage par renforcement) :**
 - Dans l'apprentissage par renforcement, on modélise l'environnement ou un agent peut évoluer en agissant sur l'environnement en utilisant des actions.
 - Pendant la phase d'entraînement, l'agent qui possède un ensemble de Tools (Actions) explore l'environnement en agissant de manière itérative avec des actions disponibles.
 - Ensuite l'environnement lui procure des récompenses ou des observations. Ce qui permet de changer l'état de l'agent.
 - Ce processus répétitif d'exploration permet de construire une table de raisonnement.
 - Une fois que la phase d'entraînement et d'exploration est terminée, l'agent devient capable d'évoluer de manière autonome dans l'environnement en prenant les actions optimales selon son état actuel qui encapsule une abstraction et une représentation de l'environnement. Exemple : QLearning, PPO (Proximal Policy Optimisation)

IA Générative

- **L'IA Générative** est une catégorie de l'IA capable de générer des nouvelles données en conséquence d'un raisonnement conditionné par un prompt utilisateur en utilisant les LLMs (Large Language Models).
- Ces LLMs, basés sur les algorithmes dits « Transformers » sont entraînés en utilisant un large corpus de données.
- Ces LLMs sont des réseaux de neurones de très grandes tailles avec des Billions de paramètres et sont capables de générer des nouvelles données n'ayant jamais été vues de différentes modalités : Texte, Images, Son, Vidéo, Musique. Exemples de LLMs : Gpt GPT-4o, DeepSeek, Claude, Gemini, GPT-O3, Dall-e, CLIP, Stable Diffusion

Apprentissage supervisé

La Régression



Apprentissage Supervisé

- Dans le cas où les étiquettes sont à valeurs réelles, on parle de régression.
- Consiste à utiliser un dataset étiqueté. C'est-à-dire des données dont on connaît les inputs et les outputs. Les outputs représentent des étiquettes ou des valeurs fournies par les experts du métier.

Apprentissage Supervisé

Régression

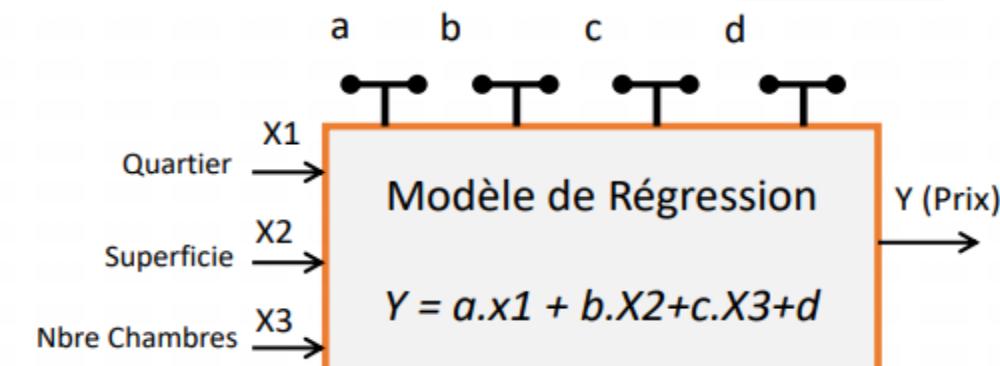
Exemple : Prédiction du prix d'un appartement

Data Set (Données étiquetées)

Input

Output

Quartier X1	Superficie X2	Nbre chambres	Prix (MDH) Y
Maarif	100	3	2
SBATA	120	3	1.1
Maarif	100	4	2.1
Bourgogne	80	3	1.44
Maarif	200	5	???????

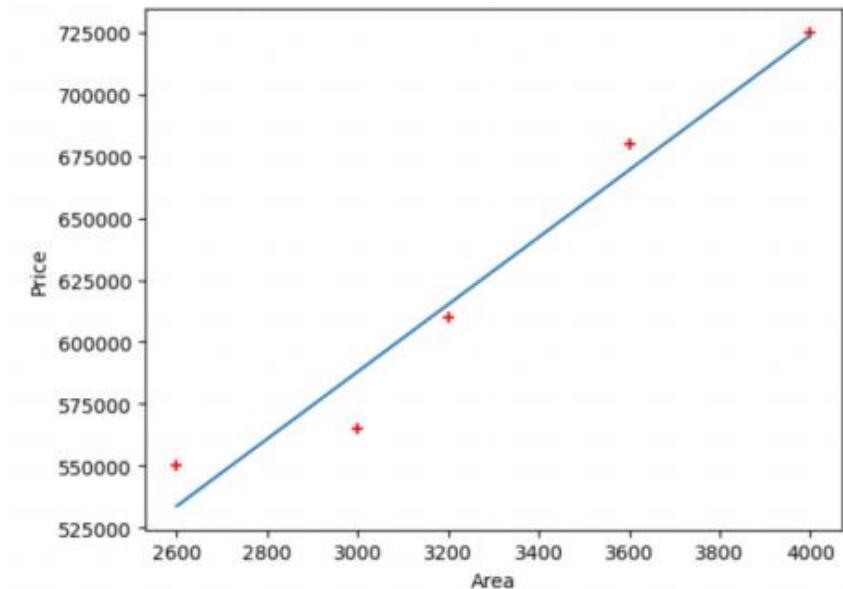


Modèle à 4 Paramètres a, b, c et d

Algorithmes : Linéaire, Polynomiale, Ridge, Lasso, etc..

Apprentissage Supervisé : Régression

- La régression consiste à déterminer une relation entre la ou les variables indépendantes et la variable dépendante.
 - La régression linéaire suppose que la relation entre les variables peut être modélisée par une équation linéaire ou une équation de droite.
 - La variable utilisée pour la prédiction est appelée variable indépendante (Features), tandis que la variable prédictive est appelée variable dépendante (Target or Outcome).
 - Dans le cas d'une régression linéaire avec une seule variable explicative, la combinaison linéaire peut s'exprimer comme suit :
 - **Y = Intercept + (Coefficient * X)**
 - *X est la variable indépendante (Input)*
 - *Y est la variable dépendante (Output)*
- L'objectif est de trouver la droite de régression qui s'ajuste le mieux aux données.
 - Meilleur ajustement => que la ligne sera telle que la distance cumulative de tous les points par rapport à la ligne est minimisée.
 - Mathématiquement, la ligne qui minimise la somme des carrés des erreurs résiduelles est appelée la droite de régression ou la ligne de meilleur ajustement.



Apprentissage Supervisé : Evaluation de la Régression Linéaire

- **R-Squared :**
 - Mesure du % de variance dans la variable cible expliquée par le modèle
 - Généralement la première métrique à examiner pour évaluer la performance d'un modèle linéaire
 - Valeur entre 0 et 1. Plus elle est élevée, mieux c'est
- **Mean Absolute Error:**
 - Métrique la plus simple pour vérifier la précision des prédictions
 - Même unité que la variable dépendante
 - Non sensible aux valeurs aberrantes (outliers), c'est-à-dire que les erreurs n'augmentent pas trop en présence de valeurs aberrantes
 - Difficile à optimiser d'un point de vue mathématique (logique purement mathématique)
 - Plus la valeur est faible, mieux c'est
- **Root Mean Square Error:**
 - Une autre métrique pour mesurer la précision des prédictions
 - Même unité que la variable dépendante
 - Sensible aux valeurs aberrantes (outliers) : les erreurs sont amplifiées en raison de la fonction carrée
 - Mais présente d'autres avantages mathématiques
 - Plus la valeur est faible, mieux c'est

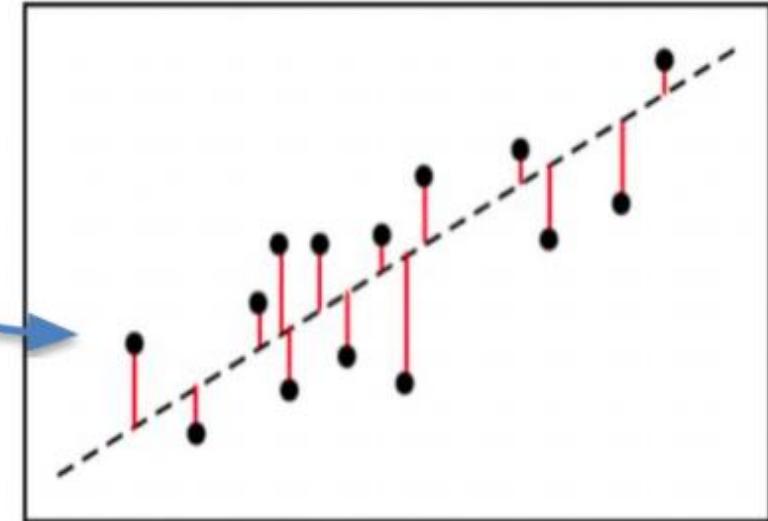
Apprentissage Supervisé : Evaluation de la Régression Linéaire

Obs	Height in Inches, X	Act Weight in Pounds, Y	Predicted Weight \hat{Y}	Residual/Error $e_i = Y_i - \hat{Y}_i$	Residual ² /Error ² $e_i^2 = (Y_i - \hat{Y}_i)^2$
1	63	127	120.1	6.900	47.61
2	64	121	126.3	-5.300	28.09
3	66	142	138.5	3.500	12.25
4	69	157	157.0	0.000	0
5	69	162	157.0	5.000	25
6	71	156	169.2	-13.200	174.24
7	71	169	169.2	-0.200	0.04
8	72	165	175.4	-10.400	108.16
9	73	181	181.5	-0.500	0.25
10	75	208	193.8	14.200	201.64
				0.000	597.28

Sum of Squared Residuals :

$$R^2 = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

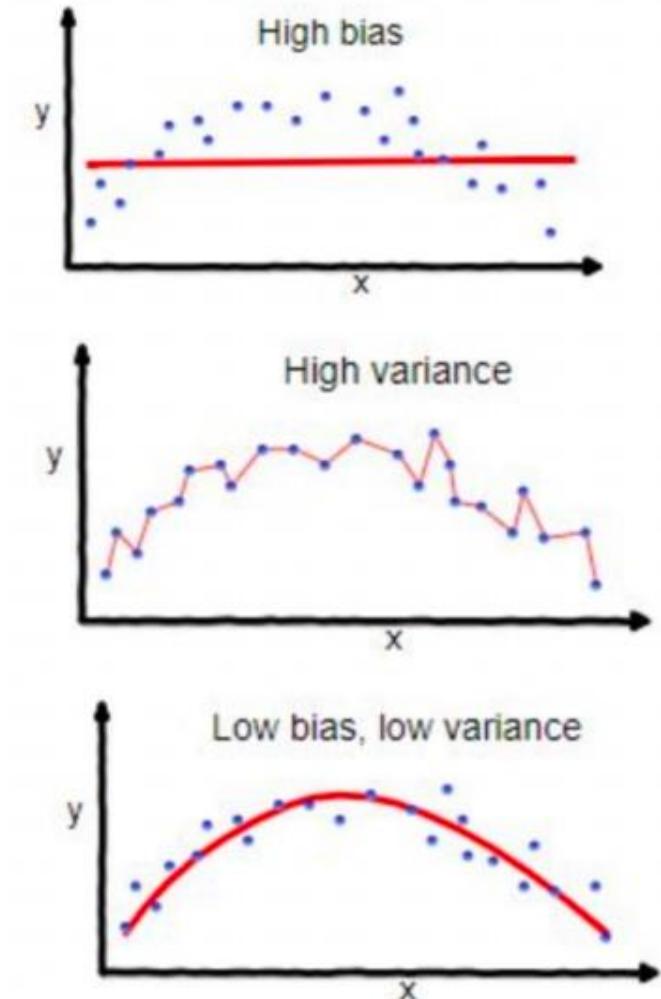
$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$



Bias-Variance: Underfitting and Overfitting

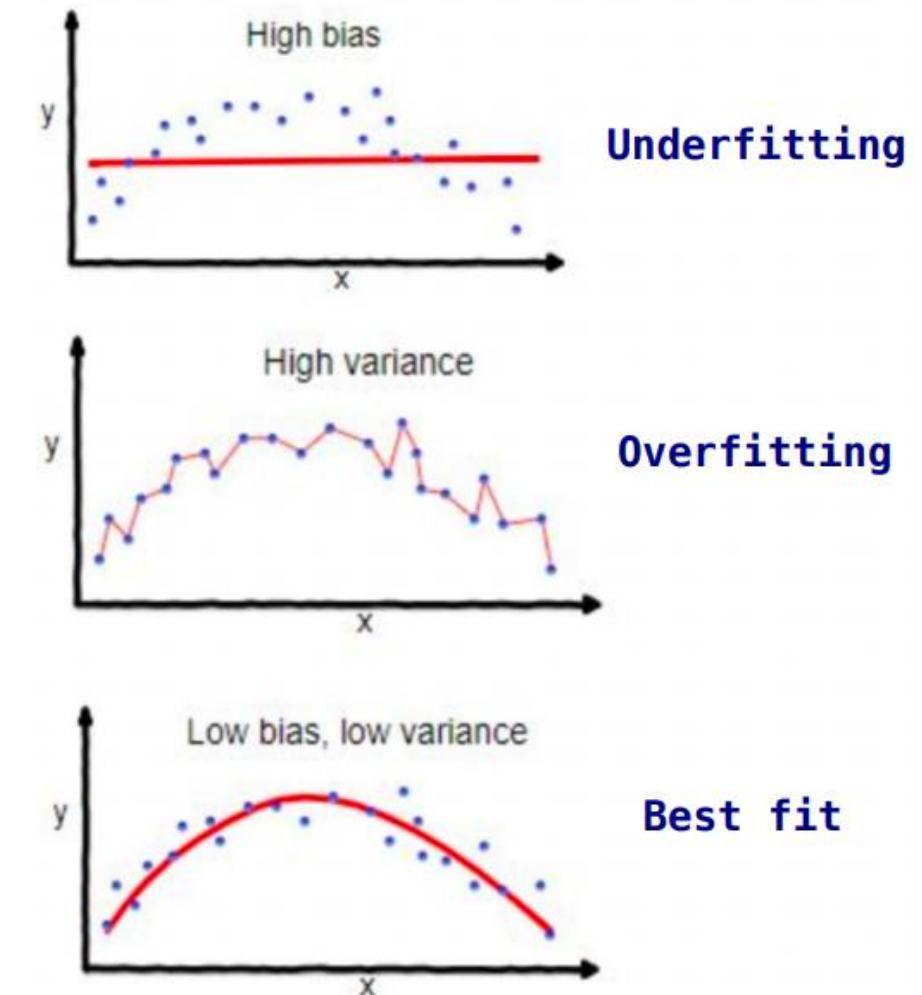
Le biais et la variance sont deux sources d'erreur clés dans les modèles d'apprentissage automatique qui impactent directement leurs performances et leur capacité de généralisation.

- **Biais** : Le biais est la différence entre la prédiction de notre modèle et la valeur correcte que nous essayons de prédire. Un modèle avec un biais élevé accorde moins d'attention aux données d'entraînement et surgénéralise, ce qui entraîne une erreur élevée sur les données d'entraînement et de test.
- **Variance** : La variance est la valeur qui indique la dispersion des données. Un modèle avec une variance élevée accorde beaucoup d'attention aux données d'entraînement en captant aussi le bruit. Ces modèles ne se généralisent pas bien sur les données de test. Par conséquent, ces modèles sont performant très bien sur les données d'entraînement mais ont une erreur élevée sur les données de test.



Bias-Variance: Underfitting and Overfitting

- **Underfitting** (Sous-ajustement) : En apprentissage supervisé, le sous-ajustement se produit lorsqu'un modèle n'est pas capable de capturer la tendance sous-jacente des données. Ces modèles ont généralement un biais élevé et une faible variance.
- **Overfitting** (Sur-ajustement) : se produit lorsque notre modèle capture le bruit (ou les fluctuations aléatoires) en plus de la tendance sous-jacente des données. Ces modèles ont généralement un faible biais et une variance élevée.



Les outils AI



kaggle



Google colab

