MAJEUR PYTHON03 - Introduction au Machine learning

Fièrement dispensé par Ketsia Mulapi Tita

Ensemble, découvrons l'univers de l'intelligence artificielle

1.1 Définitions

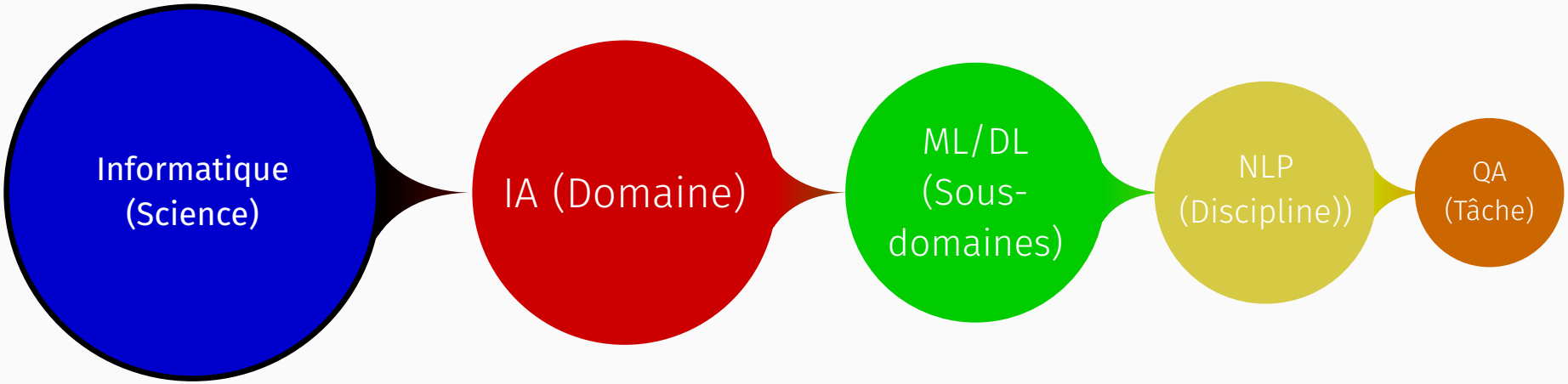
L’informatique

L’informatique *est la science* qui étudie le traitement automatique de l’information. Elle englobe la conception, le développement et l’utilisation de systèmes, de logiciels et de matériels informatiques pour collecter, stocker, traiter et transmettre des données et des connaissances.

Intelligence Artificielle

1. **L'intelligence artificielle**, parfois appelée "IA", est *un domaine de l'informatique* qui vise à créer **des machines capables de penser et d'agir comme des êtres humains.**
2. **Imaginez** un monde où les robots peuvent danser le moonwalk de Michael Jackson et discuter de la philosophie à la vitesse de la lumière !
3. L'IA englobe des **sous-domaines** passionnants tels que **l'apprentissage automatique**, les réseaux de neurones et la robotique.

Structure Hiérarchique : soyons précis, évitons les confusions !



La différence entre un algorithme d'IA, un modèle d'IA et un automate

- **Un algorithme d'IA** : C'est comme une recette pour une machine, indiquant **comment traiter les données pour résoudre un problème**.

- **un modèle d'IA** : C'est comme un artiste qui a appris en étudiant de nombreuses œuvres, la machine elle-même, **formée sur des données pour effectuer des tâches spécifiques**.

- **une automate** : C'est comme un robot préprogrammé pour effectuer des actions selon des règles définies, **sans apprentissage à partir de données**.

1.2 Métiers et disciplines de l’intelligence artificielle

1.2.2 Les disciplines en IA

L’apprentissage automatique (machine learning)

L’apprentissage automatique, c’est comme apprendre à une machine à deviner (décider) correctement en regardant beaucoup de données (d’inforamtions).

C’est comme si nous montrions à un robot comment les super-héros volent en analysant des milliers de vidéos de super-héros en action, afin qu’il puisse prédire qui sauvera la journée !

L’apprentissage automatique avancé (deep learning)

L’apprentissage automatique avancé, ou deep learning, c’est comme apprendre à une machine à devenir un super-héros des données.

C’est comme si nous donnions à une IA un super-cerveau pour qu’elle puisse résoudre des énigmes, détecter des visages dans les foules et comprendre ce que les chats disent quand ils miaulent. C’est l’IA qui devient le véritable super-héros des données !

Le Reinforcement Learning

- **Reinforcement Learning : C’est comme apprendre à jouer à un jeu vidéo en obtenant des récompenses lorsque vous prenez de bonnes décisions.**

  L’IA apprend de ses erreurs en interagissant avec son environnement et en ajustant son comportement pour maximiser les récompenses.

Le Natural Language Processing (NLP)

- **NLP (Traitement du Langage Naturel) : C'est comme apprendre à une machine à comprendre et à parler le langage humain.**

  Cela permet aux ordinateurs de lire, d'interpréter et de générer du texte comme s'ils parlaient notre propre langue.

Machine Learning for Signal Processing

- **Machine Learning for Signal Processing : C’est comme enseigner à un ordinateur à écouter et à comprendre les signaux, comme s’il s’agissait d’une langue secrète.**

  L’IA utilise des modèles pour extraire des informations utiles des signaux, tels que la musique, les images, ou même les signaux électriques.

Computer Vision

- **Computer Vision : C’est comme donner des yeux à un ordinateur pour qu’il puisse ”voir” le monde.**

  L’IA permet aux machines d’interpréter et d’analyser des images et des vidéos, ouvrant la porte à des domaines comme la surveillance vidéo intelligente et la réalité augmentée.

Machine Learning for Image Processing

- **Machine Learning for Image Processing : C'est comme apprendre à une machine à "voir" et à comprendre des images.**

  Les algorithmes d'IA analysent les pixels pour détecter des objets, des visages ou des motifs, rendant possible des applications comme la reconnaissance d'objets et la retouche photo intelligente.

Medical Image Analysis

- **Medical Image Analysis : C’est comme former une IA à lire et à interpréter des images médicales, comme les radiographies ou les IRM.**

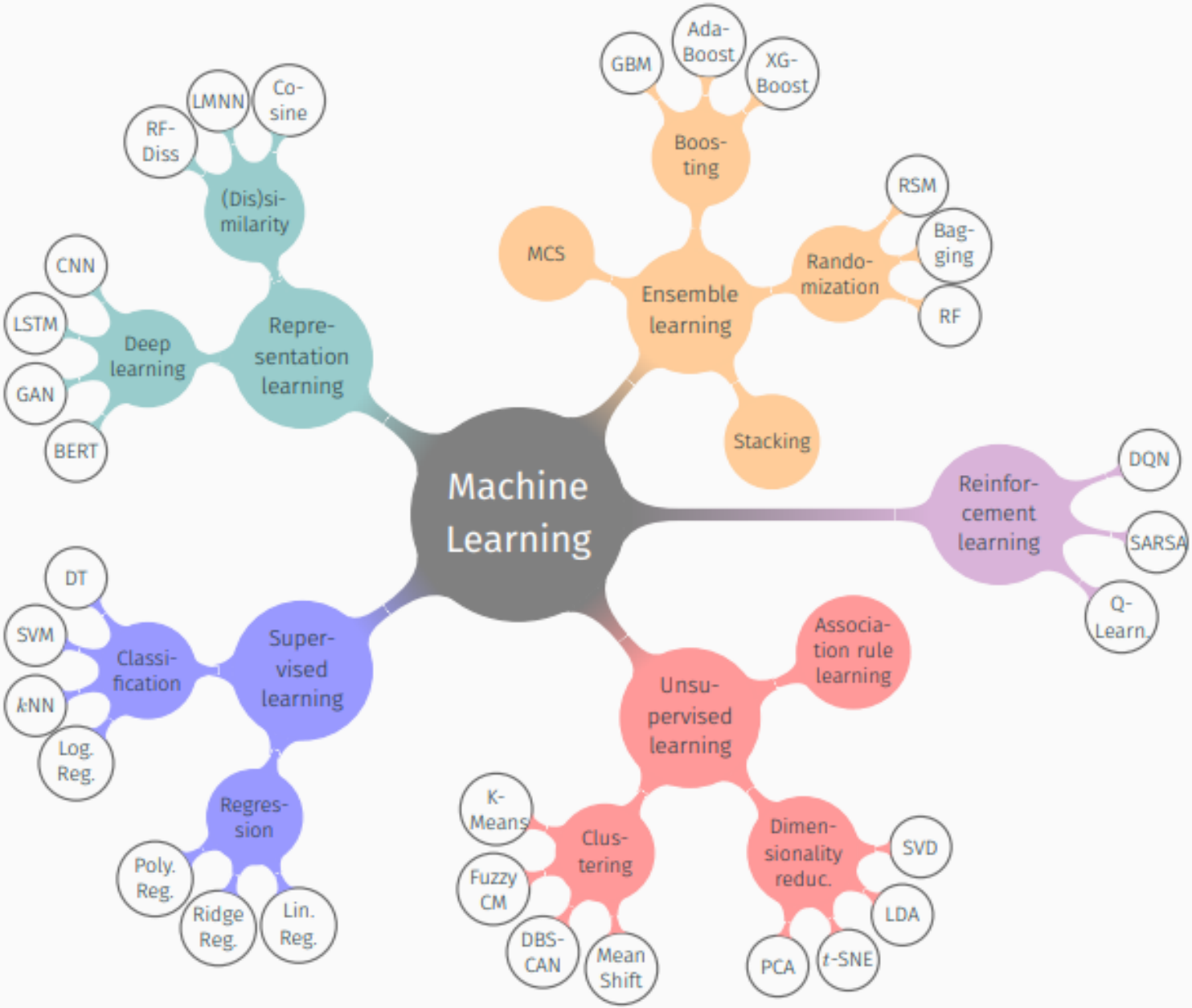
  Cela aide les médecins à diagnostiquer les maladies et à prendre des décisions éclairées pour les soins de santé.

Machine Learning for Graph or Graph Neural Network

- **Machine Learning for Graph ou Graph Neural Network : C'est comme enseigner à un ordinateur à comprendre et à extraire des informations à partir d'un réseau complexe, tel qu'un réseau social ou une structure en arborescence.**

  L'IA apprend à détecter des motifs et à prendre des décisions en utilisant des nœuds et des liens pour résoudre des problèmes liés à ces structures.

Illustration



1.2.3 Les métiers de la Data en IA

L’histoire

Le début des années 2010 marque le boom du phénomène Big Data : C’est l’avènement du Web 3.0, de l’Internet des Objets, un article a été publié dans le magazine Harvard Busness Review en 2012 pour contribué à populariser le rôle des métiers de la data, etc.

(https://source...)

Le Data Analyste et BI

- **Data Analyste et Business Intelligence (BI) : Les Data Analystes sont responsables de la collecte, de l'analyse et de l'interprétation des données pour aider les entreprises à prendre des décisions informées.**

  Ils utilisent des outils de BI pour générer des rapports, des tableaux de bord et des visualisations pour aider les gestionnaires à comprendre les tendances et les performances de l'entreprise.

Le Data Scientist

- **Data Scientist : Les Data Scientists sont des experts en données qui utilisent des techniques avancées de data science (math, etc.) pour extraire des informations significatives à partir de données complexes.**

  Ils développent des modèles statistiques et d'apprentissage automatique pour résoudre des problèmes commerciaux, de la prévision de la demande à l'optimisation des processus.

Le Data Engineer

- **Data Engineer : Les Data Engineers sont responsables de la conception, de la construction et de la maintenance des systèmes de gestion de données.**

  Ils s'assurent que les données sont collectées, stockées et accessibles de manière fiable pour les analyses ultérieures.

  Ils travaillent souvent sur l'intégration de données provenant de multiples sources (processus ETL : extract, transform, load).

Le Machine Learning Engineer

- **Machine Learning Engineer : Les Machine Learning Engineers sont des professionnels qui se concentrent sur le développement et le déploiement de modèles d'apprentissage automatique.**

  Ils travaillent sur la création de systèmes d'IA pour des applications allant de la reconnaissance d'images à la recommandation de produits. Leur rôle est de transformer les modèles en solutions pratiques.

L’ingénieur IA

- **l’Ingénieur en IA : Les Ingénieurs en Intelligence Artificielle sont chargés de concevoir, de développer et de mettre en œuvre des systèmes d’IA.**

  Ils travaillent sur des projets allant des chatbots aux assistants virtuels, en utilisant des techniques avancées d’apprentissage automatique (deep learning) pour rendre les machines ”intelligentes” dans des domaines spécifiques.

Le MLOps Engineer

- le MLOps Engineer : Les MLOps Engineers sont chargés de gérer l'ensemble du cycle de vie des modèles d'apprentissage automatique, de la formation au déploiement en passant par la surveillance continue.

  Ils s'assurent que les modèles restent performants, évolutifs et conformes aux besoins de l'entreprise.

Le Data Steward

- **le Data Steward : Les Data Stewards sont responsables de la gestion des données au sein d'une organisation.**

  Ils veillent à ce que les données soient de haute qualité, sécurisées et conformes aux réglementations. Leur rôle est de garantir que les données sont un actif précieux pour l'entreprise.

1.3 Les tâches des disciplines en IA

Machine Learning et Deep Learning

1. Prédiction - Classification
2. Prédiction - Régression
3. Prédiction - Détection

Natural Language Processing

1. Question Answering
2. Sentiment Analysis
3. Text Summarization
4. Et plus.

Medical Image Analysis

1. Analyse d'images médicales pour le diagnostic
2. Détection de lésions
3. Segmentations d'organes
4. Et plus.

Graph Neural Network

1. Analyse de réseaux complexes
2. Recommandations basées sur des graphes
3. Détection de communautés
4. Et plus.

Machine Learning for Signal Processing

1. Transformation de la parole en texte
2. Analyse de signaux audio
3. Traitement de signaux pour la détection d'événements
4. Et plus.

Reinforcement Learning

1. Apprentissage par renforcement
2. Jeux vidéo et jeux de société
3. Robotique autonome
4. Et plus.

Computer Vision

1. Détection d’objets
2. Reconnaissance faciale
3. Suivi d’objets
4. Et plus.

1.4 Les techniques d’apprentissage en IA

Supervisé

L'apprentissage supervisé utilise des données étiquetées pour entraîner des modèles.Parmi les techniques spécifiques, on trouve :

- Régression linéaire
- Classification par réseaux de neurones
- Arbres de décision
- Machines à vecteurs de support (SVM)

Ces techniques sont largement utilisées dans des domaines tels que la reconnaissance d'images, la prédiction de prix, et la classification de texte.

Non supervisé

L'apprentissage non supervisé explore des données non étiquetées. Parmi les techniques spécifiques, on trouve :

- Clustering (K-means, DBSCAN)
- Réduction de dimensionnalité (ACP, t-SNE)
- Apprentissage profond non supervisé (autoencodeurs, RBM)

Ces techniques aident à découvrir des structures et des schémas cachés dans les données, et sont utilisées dans l'analyse de données, la recommandation de produits, et la segmentation de clients.

Par reinforcement

L'apprentissage par renforcement implique l'apprentissage à partir d'interactions avec un environnement. Parmi les techniques spécifiques, on trouve :

- Q-learning
- Policy gradients
- Deep Q Networks (DQN)

Ces techniques sont couramment utilisées dans des applications telles que les voitures autonomes, les jeux vidéo, et la robotique.

Semi-supervisé

L'apprentissage semi-supervisé combine des données étiquetées et non étiquetées. Les techniques spécifiques peuvent inclure :

- Étiquetage de propagation
- Méthodes d'auto-étiquetage
- Entraînement avec des données générées

L'apprentissage semi-supervisé est utile lorsque l'annotation manuelle de données est coûteuse. Il est couramment utilisé dans la classification de documents, la détection d'anomalies, et la reconnaissance de la parole.

1.5 Les modes d’apprentissage en IA

Entraînement

L'entraînement est le processus d'affectation des valeurs aux paramètres d'un modèle pour qu'il apprenne à partir de données.

Cela implique généralement l'optimisation d'une fonction de perte en ajustant les poids du modèle. L'entraînement peut être supervisé ou non supervisé, etc. en fonction de la disponibilité des étiquettes.

Fine-Tuning (Ajustement fin)

Le fine-tuning est une technique qui consiste à prendre un modèle pré-entraîné sur une tâche et à l'ajuster pour une tâche spécifique.

Par exemple, un modèle de langage pré-entraîné peut être ajusté pour effectuer une tâche de classification de texte spécifique à une entreprise.

1.6 Les prerequis

Prérequis

Besoins identifiés :

- **Algèbre linéaire :**
  1. Scalaires, vecteurs, matrices, ...
  2. Multiplication de matrices et de vecteurs.
- **Statistiques :**
  1. Espérance, variances, médiane, écart type, corrélation.
  2. Tests statistiques (partie feature engineering).
- **Dérivées (gradient, hessienne) et optimisation (Min, Max), conditions-KKT.**

Prérequis et environnement

- le langage de programmation **python**
- **les environnement :** Colab, Anaconda, suite JetBrains pour la data science, VS Code, AWS SageMaker
- Anglais (facultatif, mais n'oubliez pas que la connaissance et les opportunités n'ont pas de limite géographique).
- Vous pouvez également utiliser l'invite de commande (terminal) dans vos environnements et y écrire du code shell, même dans les cellules de vos notebooks. Vous pouvez utiliser des syntaxes shell avec le signe du pourcentage devant, par exemple : `%curl ...` ou `%ls -l`.
- aimer lire
- Un peu de programmation et d'algorithmique

1.7 Mes conseils

Conseils de vie pour un bon début et une bonne performance :

- Ne suivez pas simplement la dernière tendance en intelligence artificielle ou en science des données. Posez d'abord des hypothèses et évaluez si elles sont pertinentes pour votre problème.

- Ne choisissez pas le deep learning par pure hype. Utilisez-le lorsque vous avez de grandes quantités de données ou des problèmes complexes qui le justifient.

- Les modèles d'intelligence artificielle sont souvent des boîtes noires. Assurez-vous de comprendre comment ils prennent leurs décisions, notamment en utilisant des techniques d'explicabilité (XAI).

Conseils de vie pour un bon début et une bonne performance :

- Posez les bonnes questions avant d'adopter une technologie ou une méthodologie. Assurez-vous que cela correspond à vos besoins spécifiques.
- Mettez en place un protocole expérimental scientifiquement rigoureux pour chaque expérience.
- Apprenez à lire des papiers de recherche et pratiquez leur implémentation. C'est ce que l'on appelle la recherche appliquée, essentielle en entreprise.

1.8 Les challenges ou compétitions

Kaggle vs. Zindi

Kaggle : Kaggle.com

1. Prix intéressants et attractifs.
2. Niveau de difficulté élevé.
3. Plateforme américaine.

Zindi : Zindi.Africa

1. Prix moins attractifs.
2. Difficulté relative.
3. Plateforme africaine.

1.9 Les environnements

Où déployer des modèles d'IA

Clouds Publics : Des plateformes de cloud computing telles qu'AWS, Azure, Google Cloud offrent des services pour le déploiement d'applications d'IA. Vous pouvez utiliser des machines virtuelles, des conteneurs ou des services spécifiques d'IA.

Serveurs Locaux : Vous pouvez déployer des modèles sur vos propres serveurs en utilisant des environnements de conteneurisation comme Docker ou en configurant des serveurs dédiés.

Plateformes de Machine Learning en Ligne : Des plateformes telles que Hugging Face, TensorFlow Serving, SageMaker (AWS), et Azure ML permettent de déployer des modèles en quelques clics.

Edge Devices : Pour des applications embarquées ou sur des appareils périphériques, vous pouvez déployer des modèles sur des appareils edge tels que des Raspberry Pi ou des caméras intelligentes.

Comment déployer des modèles d'IA

API (Application Programming Interface) : Exposez votre modèle en tant qu'API web pour permettre aux autres applications d'interagir avec lui. RESTful ou SOAP API.

Notebooks Jupyter : Vous pouvez créer des notebooks Jupyter interactifs qui utilisent le modèle pour effectuer des prédictions.

Applications Web : Intégrez votre modèle dans une application web en utilisant des frameworks comme Django, Flask/Fast (Python), ou Express.js (Node.js).

Applications Mobiles : Utilisez des kits de développement mobile (iOS, Android) pour intégrer des modèles dans des applications mobiles comme Firebase.

Incorporation dans des Services Existant : Intégrez le modèle dans des logiciels existants en utilisant des bibliothèques ou des SDK.

1.10 Les opportunités

Les Salaires

Afrique (Salaire moyen en Afrique du Sud) : 34 000€/an **Europe (France : de Junior à Sénior ou Lead) :** 30K € à >100K € **Amérique (US : de Junior à Sénior ou Lead) :** 61K $>$ 450*K* **Asie (Junior Data Scientist en Chine) :** 30K € à >70K €

Les Missions

Canal + (Offre de Stage en France) :
https://www.glassdoor.fr/Emploi/data-science-emplois-SRCH_KO0,12.htm?jobType=internship&fromAge=3

BNP Paribas (Offre CDI) :
https://www.hellowork.com/fr-fr/emplois/22139019.html

MP Data (CDI) :
https://fr.indeed.com/voir-emploi?cmp=MP-DATA&t=Data+Scientist&jk=73684216569cfc5c&q=data+science&vjs=3

Orange (Offre de thèse CIFRE) :
https://orange.jobs/jobs/v3/offers/112500?lang=fr

Les emplois

- Remote aux États-Unis pour des entreprises de la Silicon Valley : *https://developers.turing.com/dashboard/turing_test?s=outbound*
- Remote en Europe : *https://www.stakha.io/*
- Découvrez des emplois sur Zindi : *https://zindi.africa/*
- Être en haut du classement ou gagner des compétitions peut ouvrir des portes : *https://www.kaggle.com/*
- Possibilités d'immigration au Québec, Canada : *https://www.quebec.ca/immigration/programmes-immigration/intelligence-artificielle*
- Et bien d'autres opportunités à explorer.

Sport Analytics Use Case

- Paris Saint-Germain (PSG) et l'École Polytechnique lancent le Sports Analytics Challenge : *https://www.psg.fr/equipes/the-club/content/le-paris-saint-germain-et-l-ecole-polytechnique-lanc*
- Comment Kevin De Bruyne a utilisé la data science à son avantage : *https://datascientest.com/kevin-de-bruyne-comment-sest-il-servi-de-la-data-sci*
- PSG et l'intelligence artificielle : *https://www.lebigdata.fr/psg-intelligence-artificielle*
- PSG et Polytechnique lancent le premier Sport Analytics Challenge : *https://www.forbes.fr/technologie/le-psg-et-polytechnique-lancent-le-premier-sport-ana*

1.11 initiation au machine learning

Différence entre machine learning et machine learning avancé (deep learning)

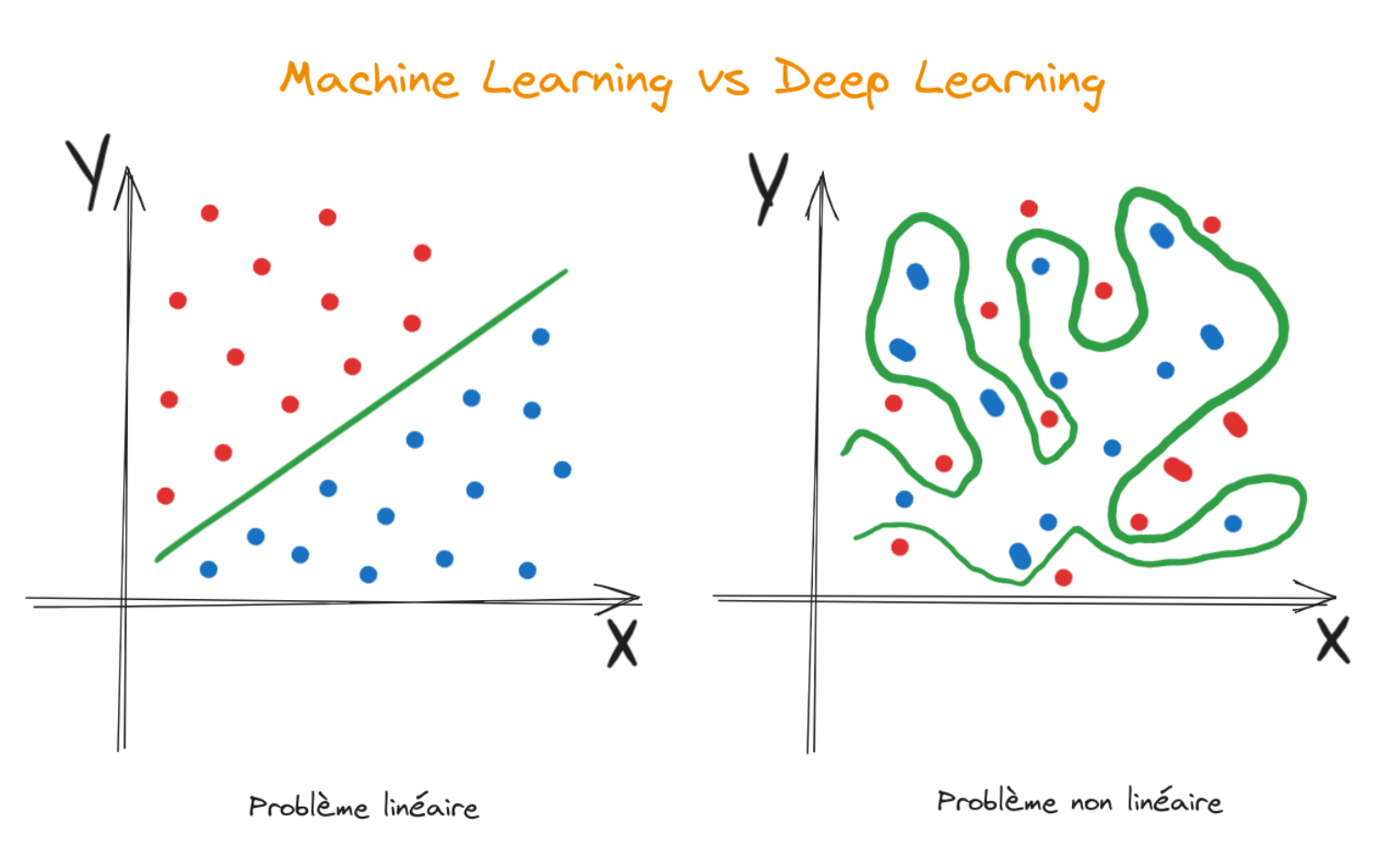


Figure 1: ML vs DL

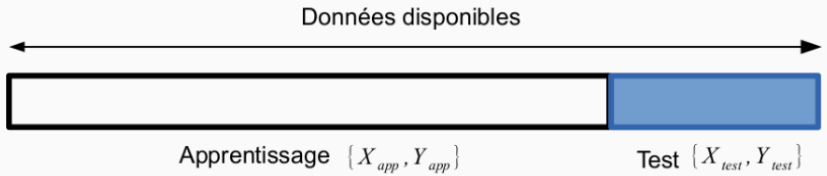
L'apprentissage vise à estimer une relation inconnue

- Les données sont liées à une variable cible que l'on cherche à prédire.
- On suppose l'existence d'une relation (inconnue) : $f : X \to Y$, où Y est le domaine de la variable cible Y.
- L'apprentissage vise à estimer f avec une fonction de prédiction (ou modèle) $h : \hat{y} = h(x), \forall x \in X$, où $\hat{y}$ est la prédiction de Y pour x.
- La fonction h appartient à un espace d'hypothèses H, par exemple, l'ensemble des fonctions polynomiales.
- L'apprentissage supervisé signifie que les données d'apprentissage sont annotées, c'est-à-dire associées à leur "vraie" valeur y. Notation : $D = \{(x_i, y_i) \in X \times Y, i = 1, ..., n\}$.

Découpage des données en train, test

Découper aléatoirement D_n en deux sous-ensembles disjoints D_{app} et D_{test}:

- $D_{\text{app}} = \{(x_i, y_i), i = 1, \ldots, n_{\text{app}}\}$: données servant à l'apprentissage de h.
- $D_{\text{test}} = \{(x_i, y_i), i = 1, \ldots, n_{\text{test}}\}$: données servant à évaluer la capacité de généralisation de h.



Remarques:

- Plus n_{app} est grand, meilleur est l'apprentissage.
- Plus n_{test} est grand, meilleure est l'estimation de la performance en généralisation de h.
- D_{test} n'est utilisé qu'une seule fois!

Par exemple

Voici un problème univarié (problème linéaire), c'est-à-dire à une seule variable (Arrondissement). Considérons que les arrondissement à prédire ne sont pas présents sur ce tableau.

Index \ VAR	X= Arrondissement	Y= Prix (€)
0	1	2000
1	2	4000
2	...	...
3	5	10 000
4	16	32 000

Figure 2: On veut prédire le prix des appartements à Paris .

Par exemple

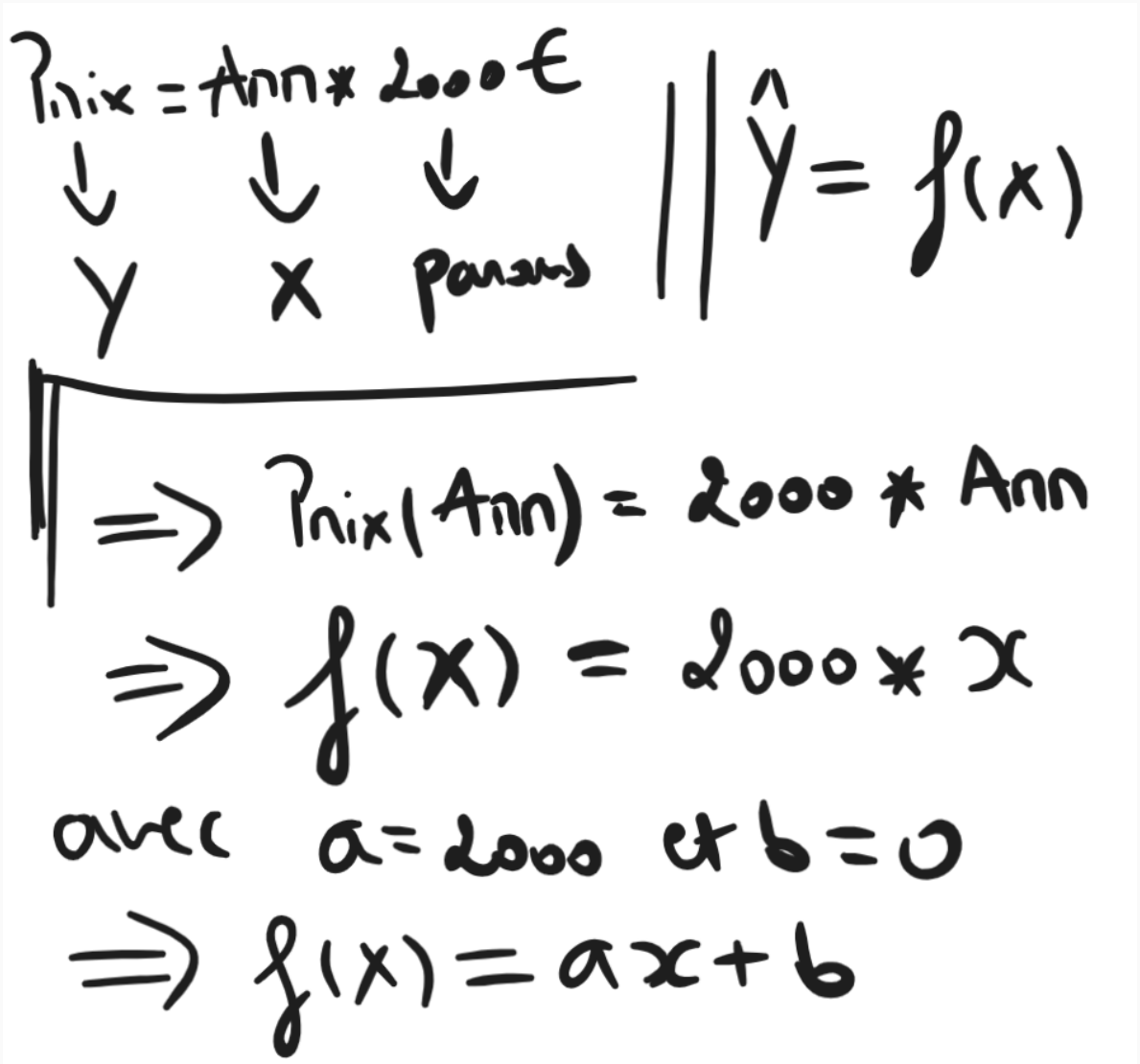


Figure 3: Relation mathématique (l'algorithme)

Par exemple : prédiction 1

On suppose avoir appris sur plusieurs données et que le paramètre **w1** qui correpsond à **a** est égale à **2000** . Le modèle prédit **40000€** pour les appartements du **20e** arrondissement.

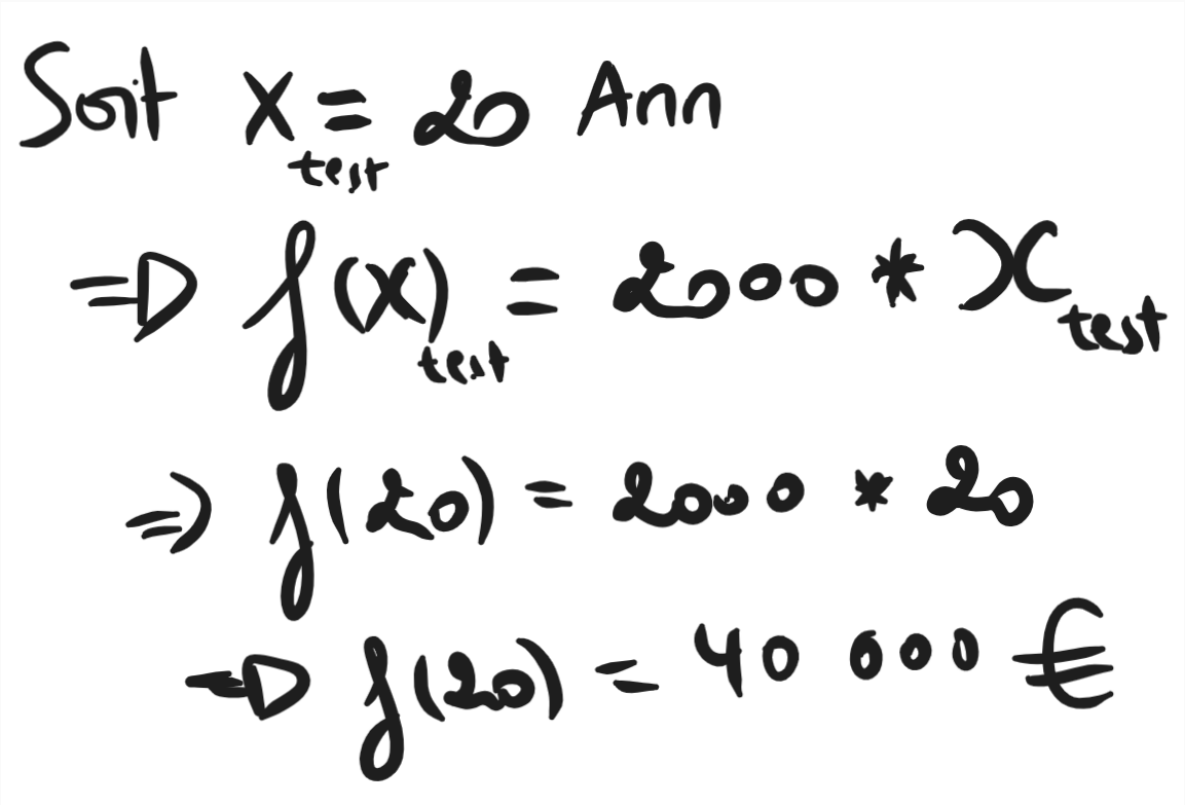


Figure 4: Le modèle d'apprentissage automatique de notre problème

Par exemple : prédiction 2

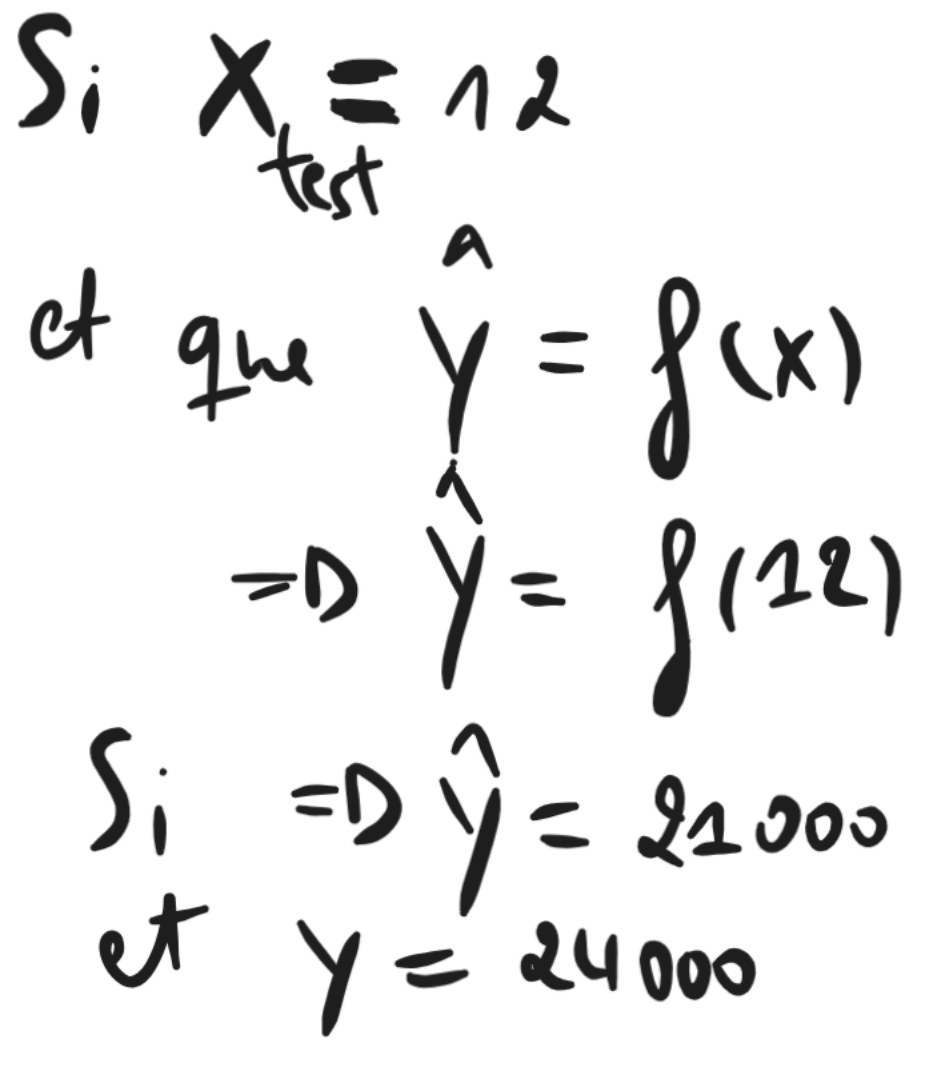


Figure 5: Notre modèle prédit **21000€** , comme prix des appartements du **12e** arrondissement de Paris

Par exemple : évaluation de la prédiction 2

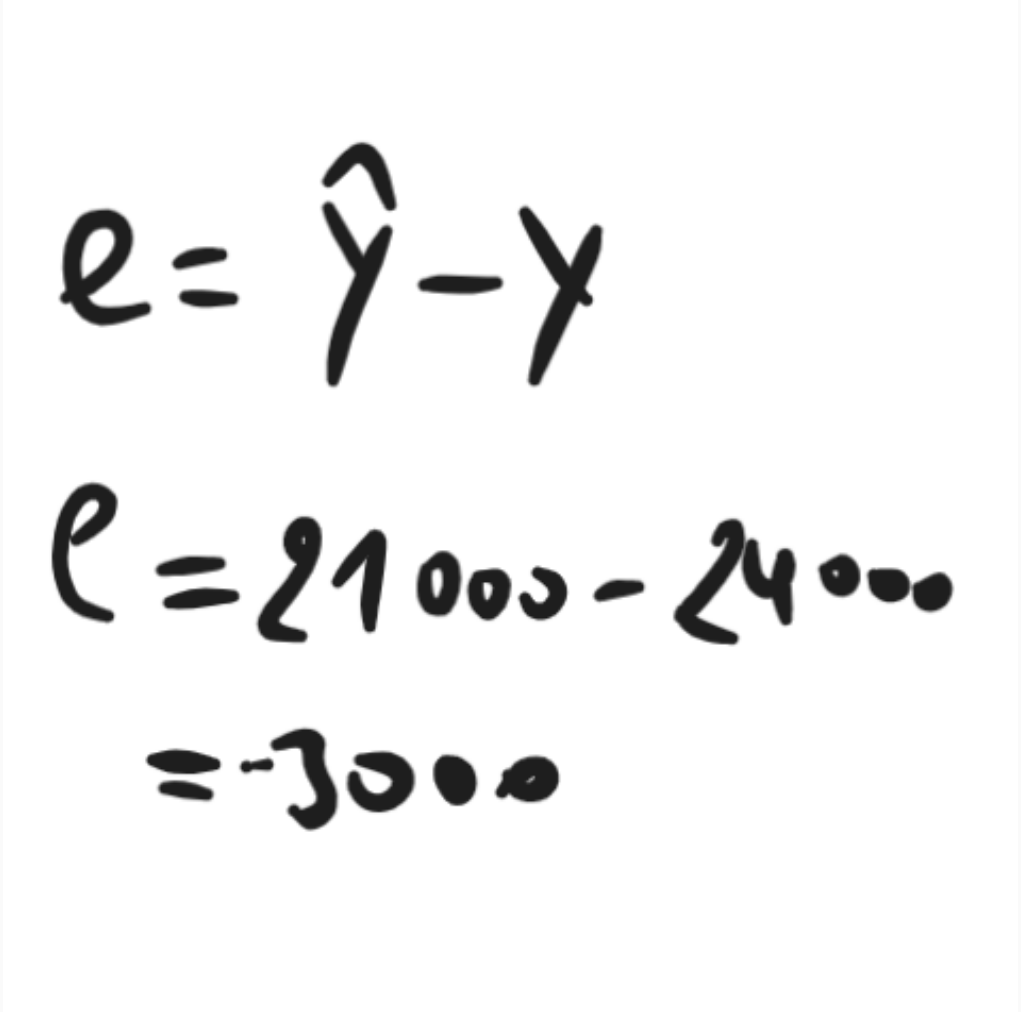


Figure 6: 24000€ , c'est le vrai prix à prédire pour les appartements du **12e** arrodissement de Paris et, **-3000€** c'est l'erreur (le résidu) entre la prédiction et la réalité

Exemple 2

Maintenant voici un problème multivarié (problème linéaire multiple), c'est-à-dire à plusieurs variables (Arrondissement, Mètre carré).

Index \ VAR	X1=Arrondissement	X2= Mètre carré	Y= Prix (€)
0	1	18	2000
1	2	19,5	4000
2	...	...	...
3	5	24	10 000
4	16	40.5	32 000

Figure 7: A l'aide de l'arrondissement et du mètre carré des appartmeents, on veut prédire le prix des appartements des arrondissements de paris

Exemple 2

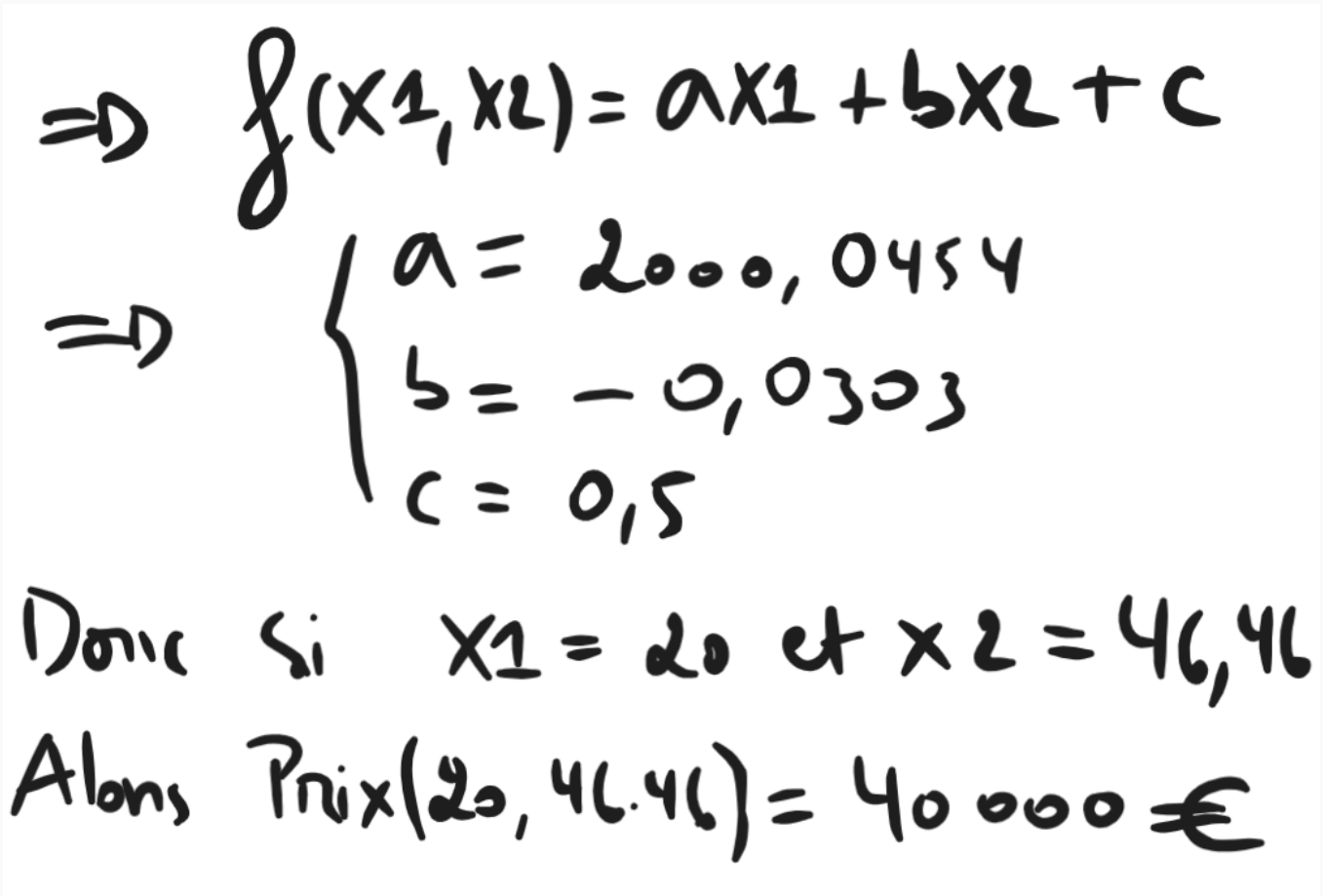


Figure 8: algorithme, modèle et prédiction

Exemple 2

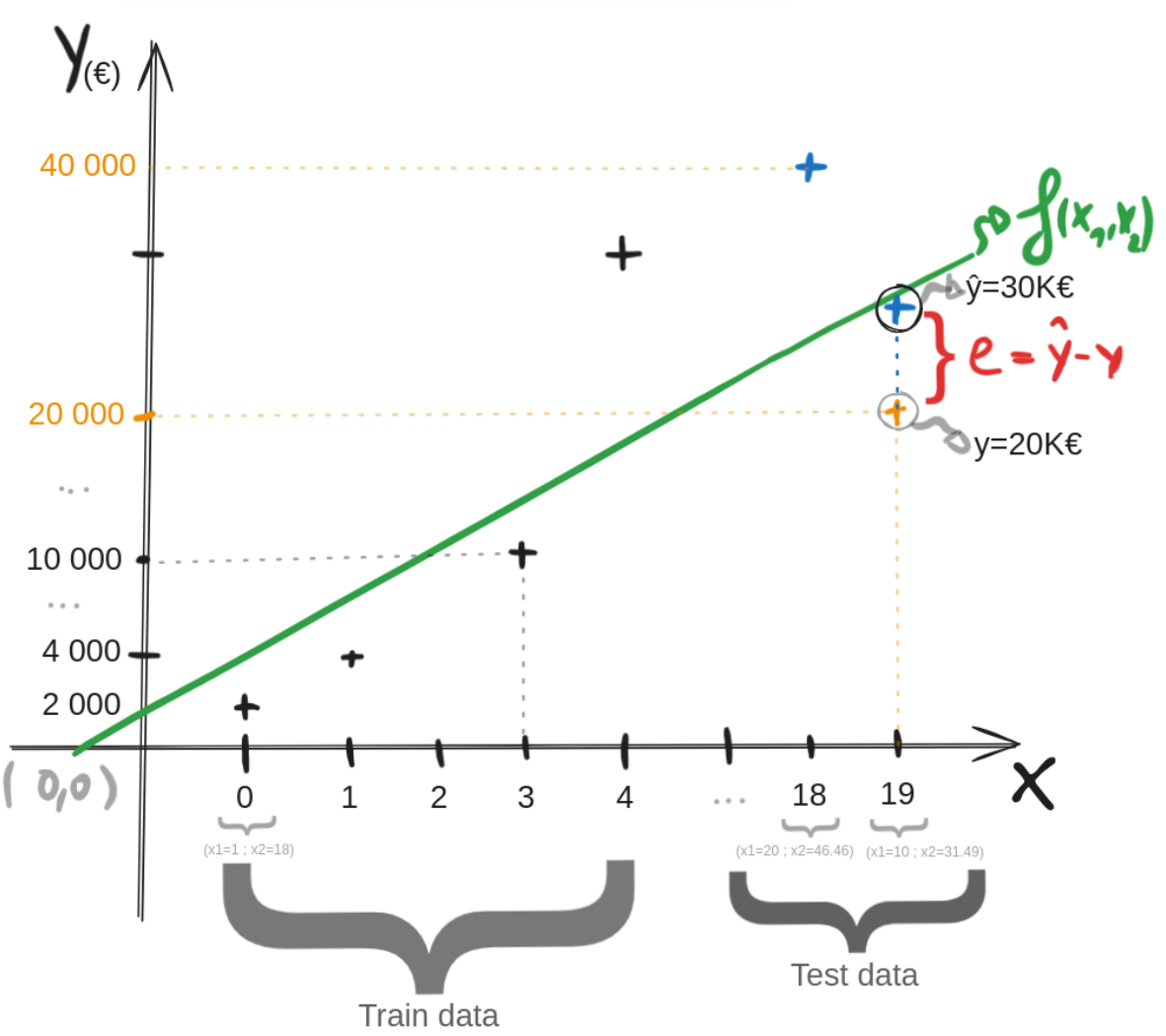


Figure 9: Représentation graphique des réalités, prédictions et de l'erreur, prix Y et index X .

Deux catégories de tâches en apprentissage supervisé

1. **Classification :** La variable cible Y se compose d'un ensemble discret de c valeurs/modalités : $Y = \{\lambda_1, \lambda_2, ..., \lambda_c\}$, avec $c \geq 2$. La fonction h est désignée comme un classifieur **(cas des classes d'appartenances)**.
2. **Régression :** La variable cible Y représente un ensemble continu de nombres réels : $Y \subset \mathbb{R}$. La fonction h est qualifiée de régresseur **(cas des prix, etc.)**.

Evaluation des erreurs : résidu, empirique, réelle

Erreur Résiduelle :

- L'erreur résiduelle c'est la différence entre la valeur prédite et la valeur réelle dans les train data.
- Chaque point de données a une erreur résiduelle associée.

Objectif :

- L'objectif est de minimiser ces erreurs résiduelles, qui servent souvent à évaluer la qualité de l'ajustement du modèle aux données d'entraînement.

Récapitulatif :

- L'erreur réelle mesure la performance réelle du modèle sur de nouvelles données.
- L'erreur empirique évalue la performance sur l'ensemble d'entraînement.
- L'erreur résiduelle quantifie la différence entre les prédictions du modèle et les valeurs réelles dans l'ensemble d'entraînement.

Évaluation des Performances d'un Modèle de Régression (Mean Squared Error

Lorsqu'on veut évaluer la qualité d'un modèle, deux mesures courantes sont utilisées sur des données de test.

Erreur Quadratique Moyenne (MSE):

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

- La MSE est nulle lorsque les prédictions sont parfaites.
- Cependant, elle n'est pas normalisée et dépend de la variance de y.

Évaluation des Performances d'un Modèle de Classification (Matrice de Confusion)

Matrice de Confusion pour la Classification Binaire:

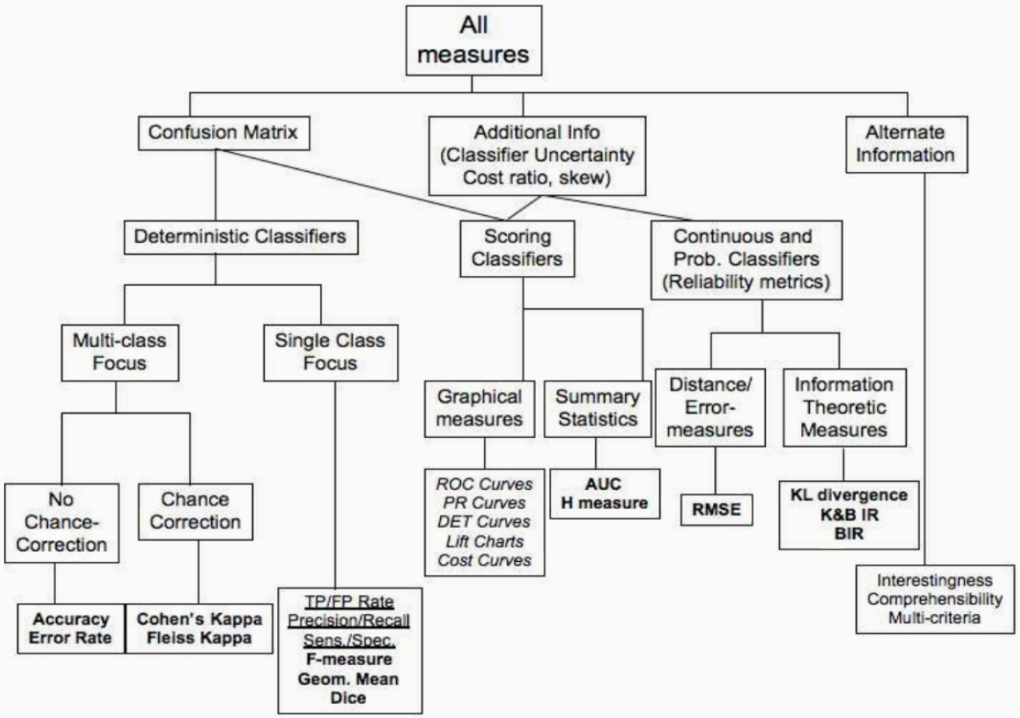
	Vérité Terrain (y)	Positifs	Négatifs
Valeurs Prédites ($\hat{y}$)	Positifs	(TP)	(FP)
	Négatifs	(FN)	(TN)

- TP (True Positive) : nombre de positifs classés positifs (bonnes prédictions).
- FP (False Positive) : nombre de négatifs classés positifs (erreurs).
- FN (False Negative) : nombre de positifs classés négatifs (erreurs).
- TN (True Negative) : nombre de négatifs classés négatifs (bonnes prédictions).

Plusieurs critères peuvent être construits à partir de cette matrice...

Evaluation : Mesures de Performances - Choix de la Mesure

- Il existe de nombreuses autres mesures de performances.
- Chacune d'elles peut désigner un modèle différent comme étant "le meilleur".
- Le choix de la bonne mesure de performances dépend de la tâche d'apprentissage, des modèles étudiés, de la problématique spécifique, etc.



Aller plus loin

Comme vous l'avez peut-être compris, il y a énormément de notions à apprendre en intelligence artificielle :

1. Découpage des données en train/val(dev)/test
2. Recherche du meilleur modèle avec les hyperparamètres
3. Architecture parallèle, séquentielle, cascade
4. etc.

Très souvent les Data Scientists choisissent une discipline de prédilection. **Prenez le temps d'apprendre CORRECTEMENT ce que vous voulez faire**.

Conclusion

En résumé :

1. Comprendre les données
2. Faire le prétraitement des données (aller à la page suivante)
3. Découper les données en train/test (à ce stade nous n'avons abordé que ça!)
4. Séléctionner 2 à 3 modèles (no free lunch theorem)
5. Entrainer les 2 (3) modèles avec les train data
6. Evaluer les perfomances des 2 (3) modèles avec les test data
7. Choisir le meilleur des 2 (3) modèles

ANNEXE : features engineering

Techniques de Prétraitement des Données

1. Normalisation (définir une échelle) des données.
2. Standardisation (réduire et centrer) des données.
3. Gestion des valeurs manquantes (imputation).
4. Encodage des variables catégorielles.
5. Réduction de la dimension (PCA, LDA) si nécessaire.
6. Traitement des données déséquilibrées.
7. Suppression des valeurs aberrantes.
8. Transformation log, exponentielle, etc. (si nécessaire)
9. Extraction des meilleures caractéristiques (variables).
10. Discrétisation des variables continues (si nécessaire : par exemple dans le cas d'un problème de clasification bianire, on va préférer 0 et 1 que 0.0 et 1.0).
11. Création de nouvelles variables (si nécessaire).
12. etc.