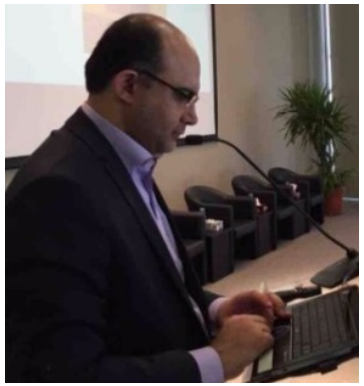
Machine Learning en finance

Votre interlocuteur

Samer EL SAWDA (https://www.linkedin.com/in/ssawda/)

- Docteur en Sécurité de l'information de l'université Pierre et Marie Curie (Paris VI) depuis 01/2011.
- 22 ans d'expérience dans le développement et la gestion des projets informatiques dans plusieurs domaines : télécommunication, industriel et finance.
- Domaines de compétence : Management des projets informatiques, Méthode agile, Leadership et organisation, entrepreneuriat (Instructeur à Society for Effectual Action) et nouvelles technologies (Big Data, AI/ML, BlockChain et Cryptocurrencies),
- Consultant-formateur. Intervention dans plusieurs pays : France, Liban, Turquie, Oman, Dubaï, Egypte et Djibouti.



Depuis Nov. 2015 : Head of Innovative Solutions and projects for Deriveatives, Tresury and Forex – Natixis SA - France

Oct. 2014 – Oct. 2015 : Business Analyst chez BPI - France

Mai 2011 Oct. 2014 : Technical IT manager – Back office chez Natixis - France

Nov. 2006 Mars 2011 : Technical IT manager – Front Office - Risk management chez CACIB - France

Juin 2003 Oct. 2006 : Technical IT manager chez RATP - France

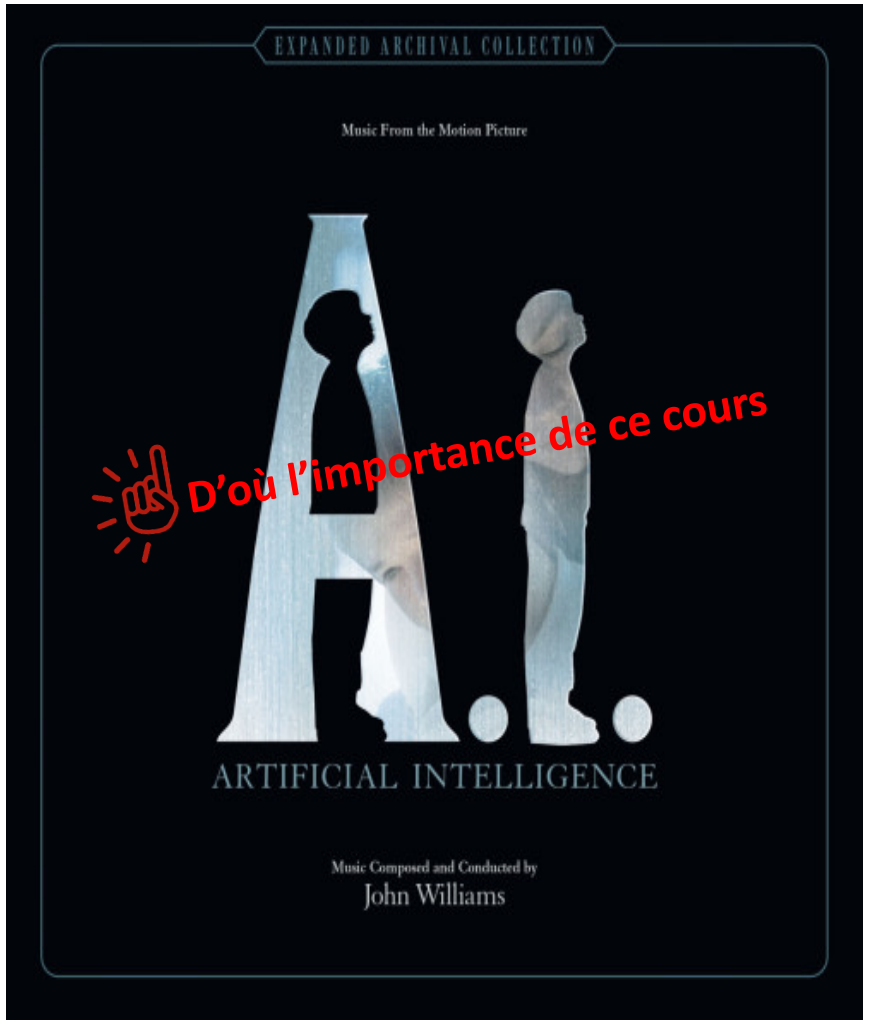
Oct. 2000 Juin 2003 : Chef de projet chez Ericsson / Nortel - France

Pourquoi ce cours

L'intelligence artificielle est une innovation importante qui va changer beaucoup de choses dans le monde de la finance :

- Positivement : augmentation de la productivité, et automatisation des tâches répétitives.
- Négativement : moins d'emplois, disparition de certains métiers.

  Certains métiers dans le domaine de la finance sont en danger face à la concurrence de l'intelligence artificielle : **Comptable**, **banquier**, **employé en assurance**, **gestionnaire de patrimoine**, … ces métiers sont voués à être remplacés petit à petit par l'intelligence artificielle.



D'où l'importance de ce cours

Le but de ce cours

Pour vous initier à un domaine qui est en train de changer votre futur métier

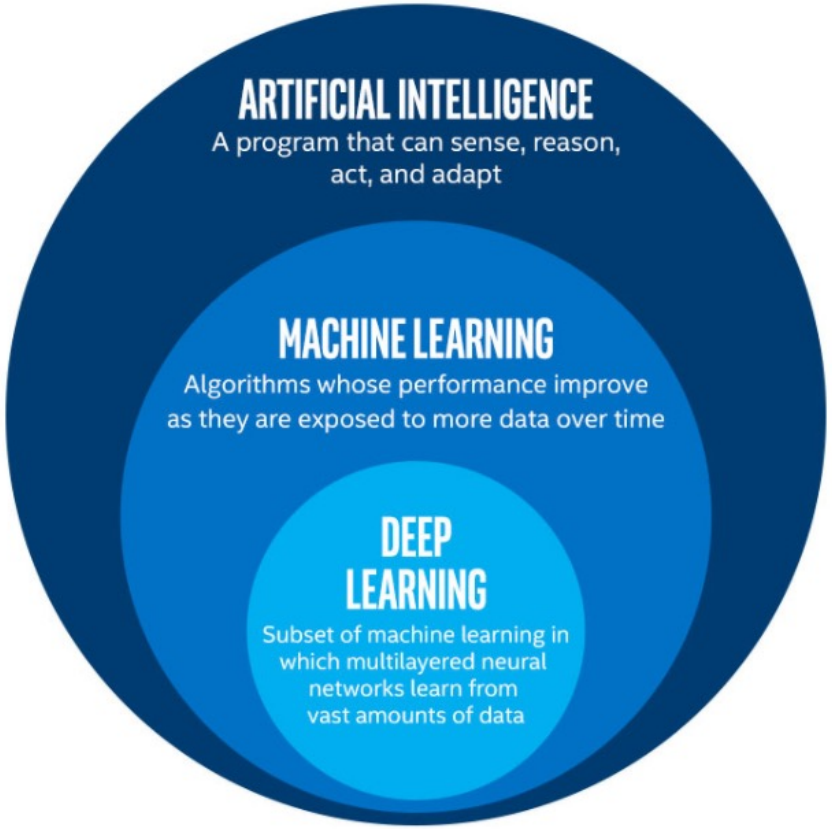
Plan du cours

- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

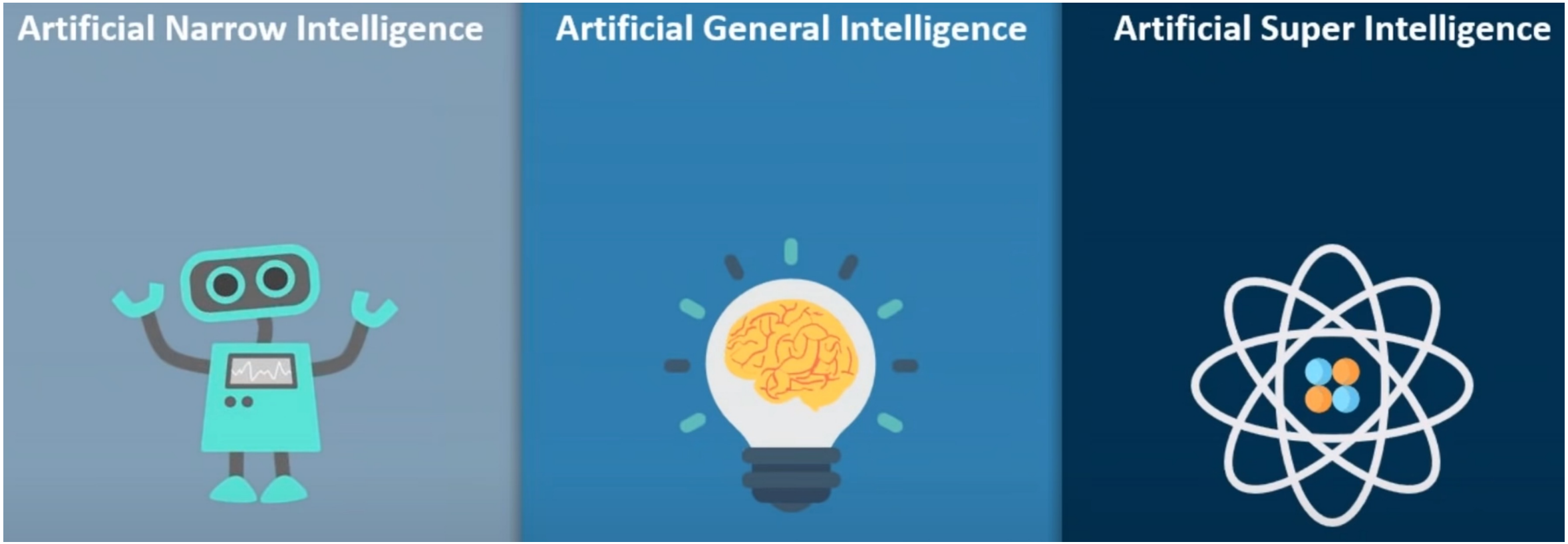
L’intelligence artificielle (AI)

L’intelligence artificielle est définie comme l’ensemble des techniques mises en œuvre afin de construire des machines capables de faire preuve d’un comportement que l’on peut qualifier d’intelligent.

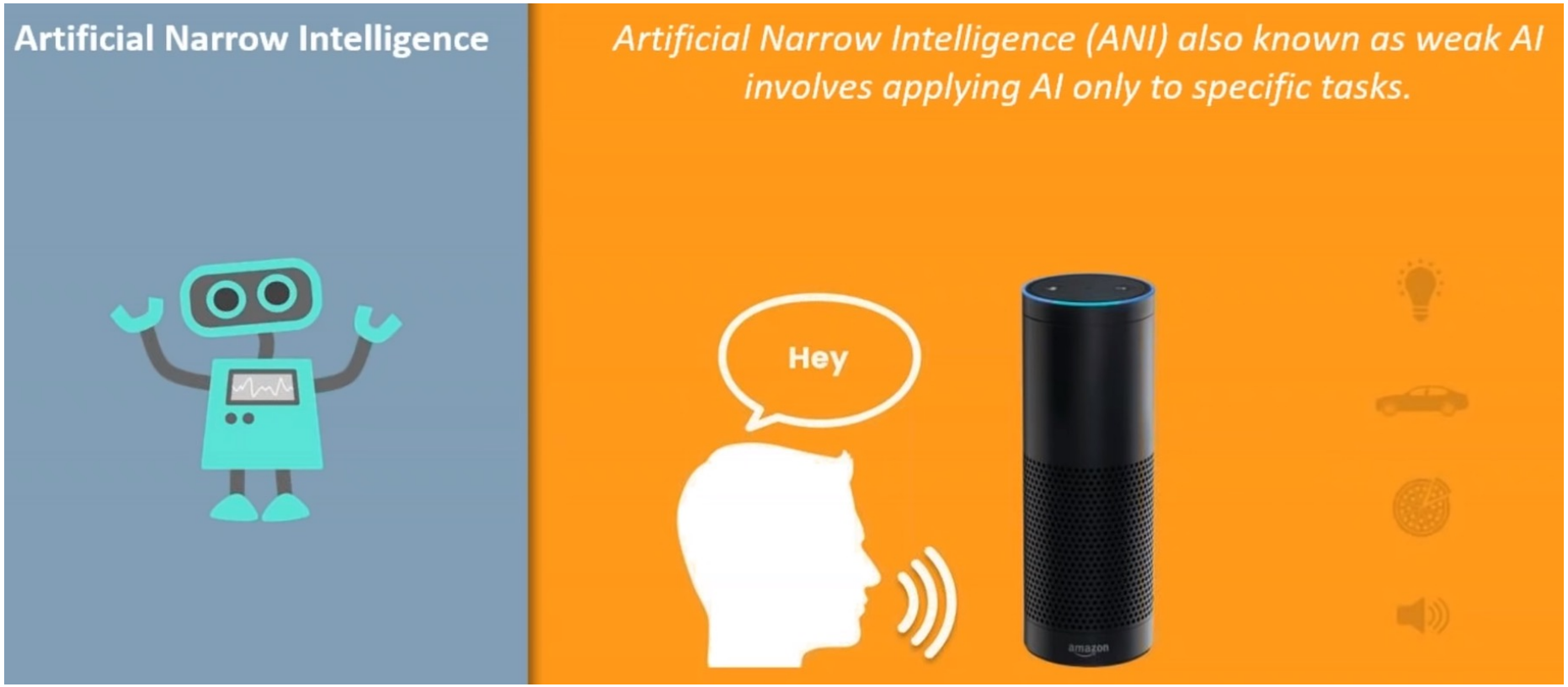
Le ML est une branche de l’intelligence artificielle.



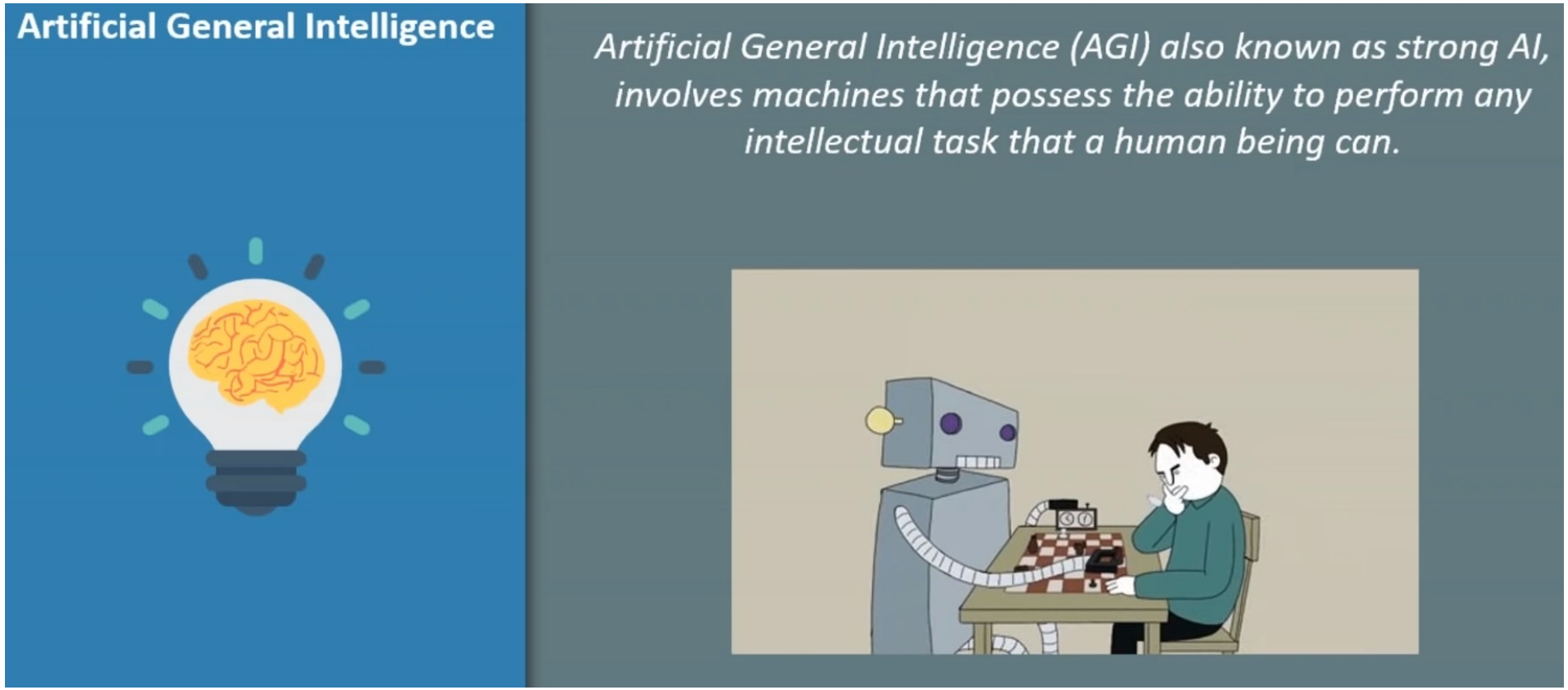
AI : les types


Artificial Narrow Intelligence
Artificial General Intelligence
Artificial Super Intelligence

AI : les types - ANI

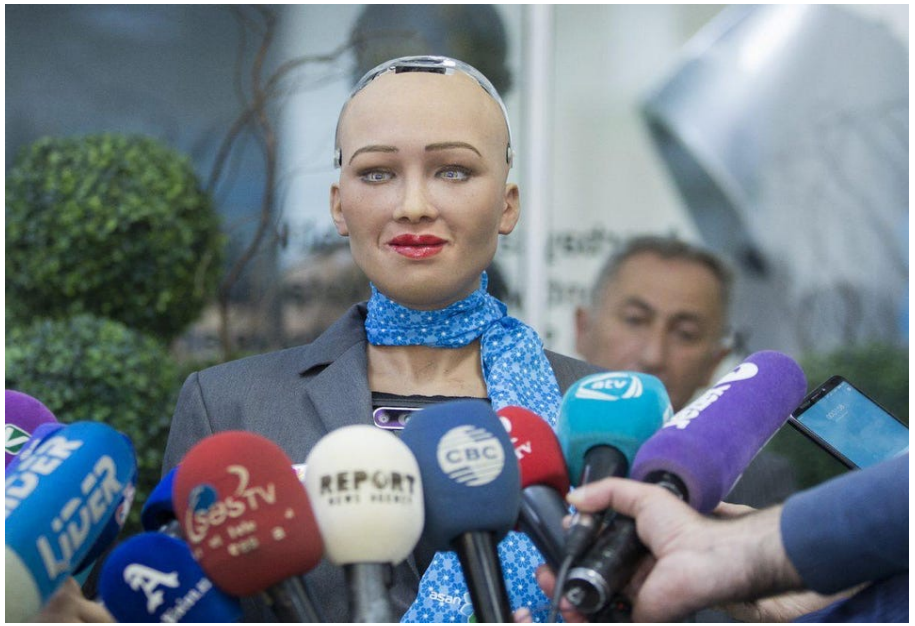


AI : les types - AGI



AI : les types - AGI

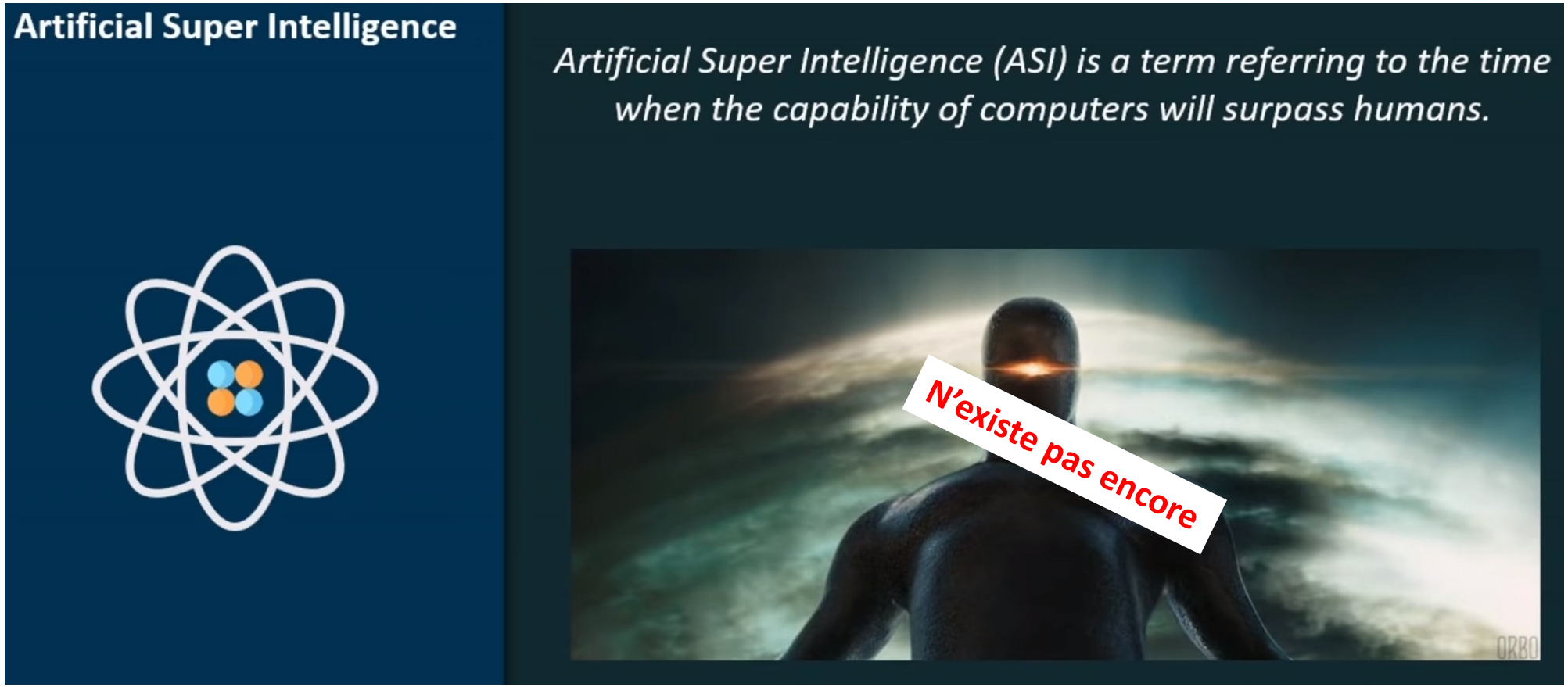
NEO



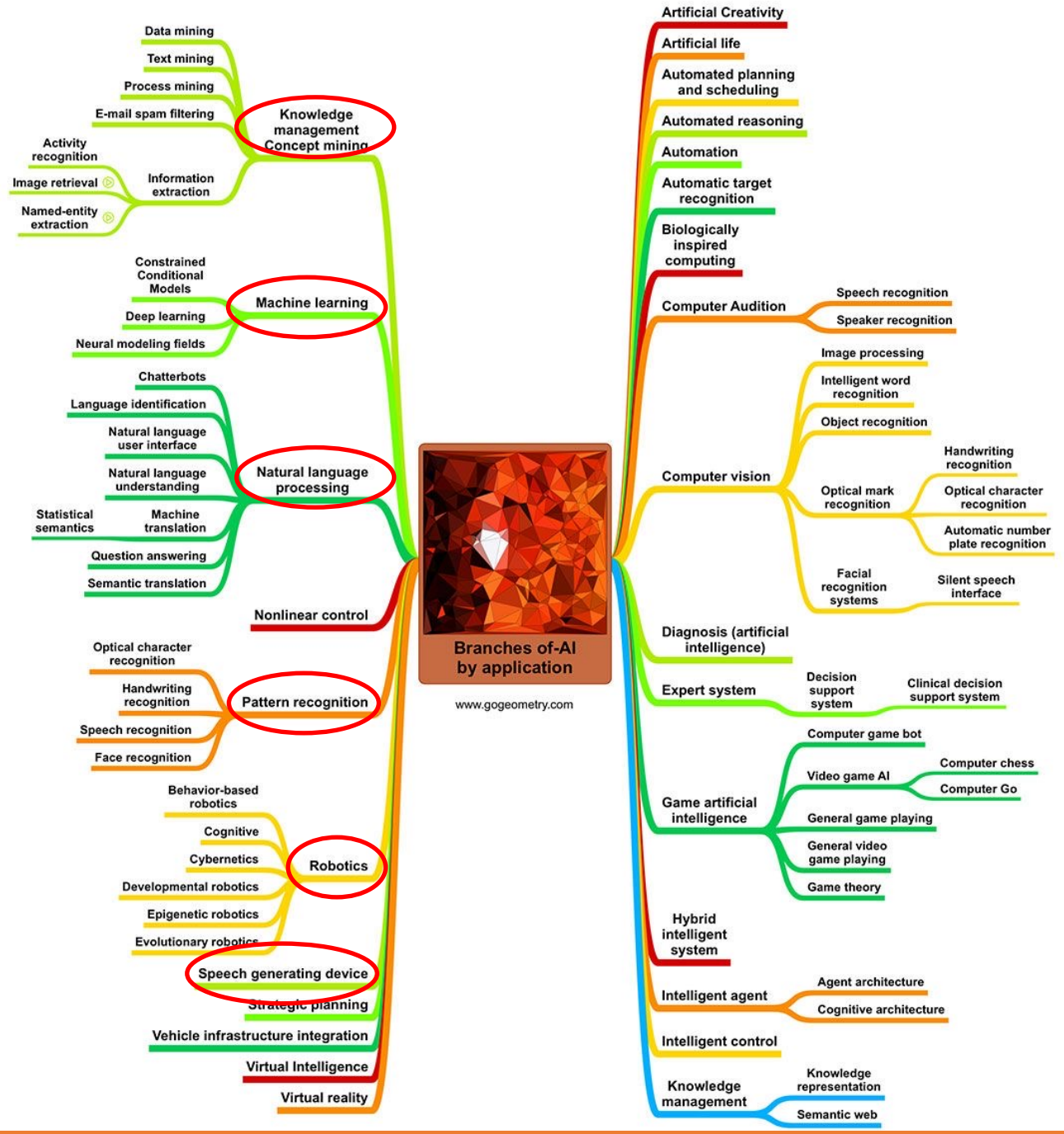
Sophia

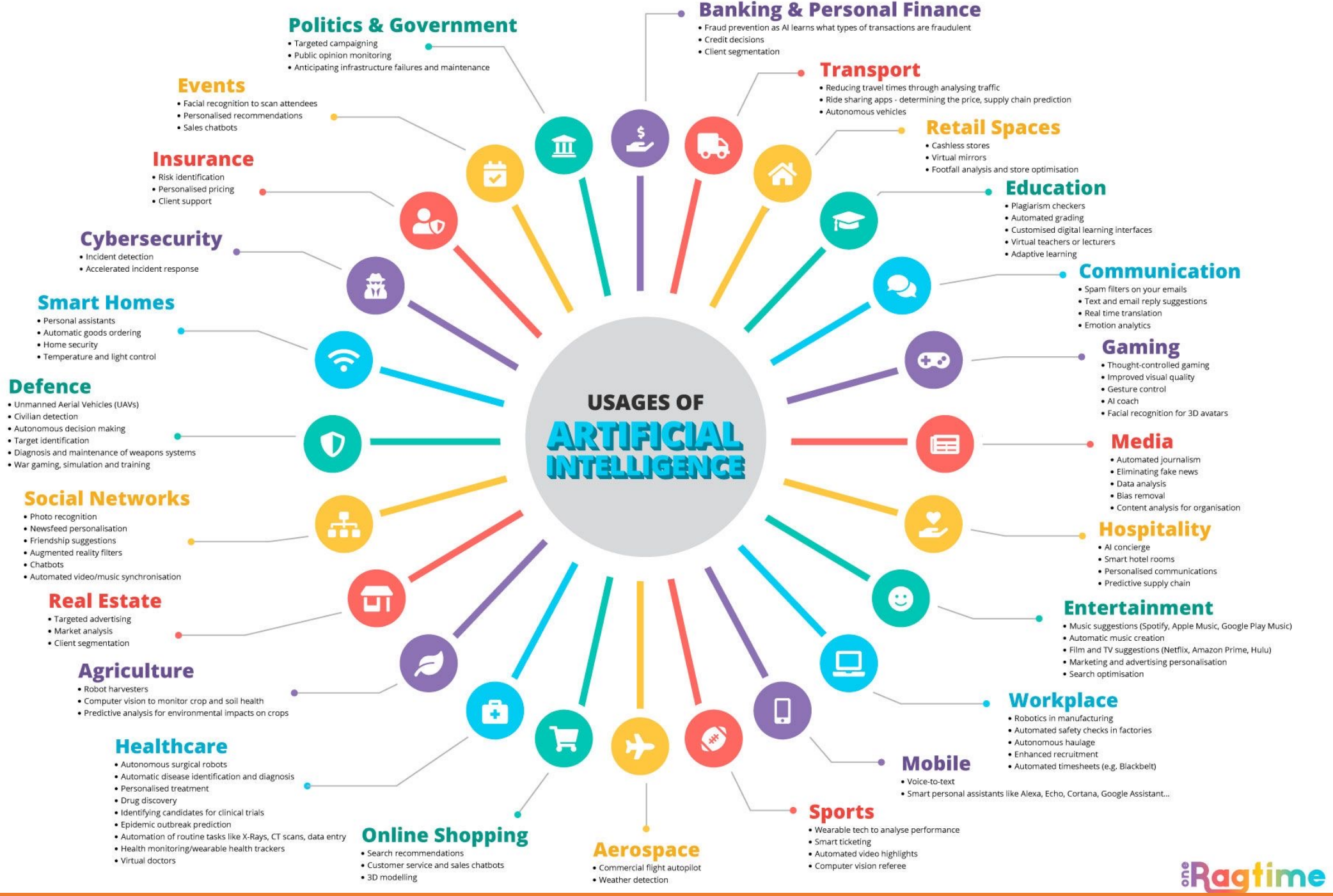
AMECA

AI : les types - ASI



Les Branches de AI




USAGES OF
ARTIFICIAL INTELLIGENCE
Banking & Personal Finance
• Fraud prevention as AI learns what types of transactions are fraudulent
• Credit decisions
• Client segmentation
Transport
• Reducing travel times through analysing traffic
• Ride sharing apps - determining the price, supply chain prediction
• Autonomous vehicles
Retail Spaces
• Cashless stores
• Virtual mirrors
• Footfall analysis and store optimisation
Education
• Plagiarism checkers
• Automated grading
• Customised digital learning interfaces
• Virtual teachers or lecturers
• Adaptive learning
Communication
• Spam filters on your emails
• Text and email reply suggestions
• Real time translation
• Emotion analytics
Gaming
• Thought-controlled gaming
• Improved visual quality
• Gesture control
• AI coach
• Facial recognition for 3D avatars
Media
• Automated journalism
• Eliminating fake news
• Data analysis
• Bias removal
• Content analysis for organisation
Hospitality
• AI concierge
• Smart hotel rooms
• Personalised communications
• Predictive supply chain
Entertainment
• Music suggestions (Spotify, Apple Music, Google Play Music)
• Automatic music creation
• Film and TV suggestions (Netflix, Amazon Prime, Hulu)
• Marketing and advertising personalisation
• Search optimisation
Workplace
• Robotics in manufacturing
• Automated safety checks in factories
• Autonomous haulage
• Enhanced recruitment
• Automated timesheets (e.g. Blackbelt)
Mobile
• Voice-to-text
• Smart personal assistants like Alexa, Echo, Cortana, Google Assistant...
Sports
• Wearable tech to analyse performance
• Smart ticketing
• Automated video highlights
• Computer vision referee
Aerospace
• Commercial flight autopilot
• Weather detection
Online Shopping
• Search recommendations
• Customer service and sales chatbots
• 3D modelling
Healthcare
• Autonomous surgical robots
• Automatic disease identification and diagnosis
• Personalised treatment
• Drug discovery
• Identifying candidates for clinical trials
• Epidemic outbreak prediction
• Automation of routine tasks like X-Rays, CT scans, data entry
• Health monitoring/wearable health trackers
• Virtual doctors
Agriculture
• Robot harvesters
• Computer vision to monitor crop and soil health
• Predictive analysis for environmental impacts on crops
Real Estate
• Targeted advertising
• Market analysis
• Client segmentation
Social Networks
• Photo recognition
• Newsfeed personalisation
• Friendship suggestions
• Augmented reality filters
• Chatbots
• Automated video/music synchronisation
Defence
• Unmanned Aerial Vehicles (UAVs)
• Civilian detection
• Autonomous decision making
• Target identification
• Diagnosis and maintenance of weapons systems
• War gaming, simulation and training
Smart Homes
• Personal assistants
• Automatic goods ordering
• Home security
• Temperature and light control
Cybersecurity
• Incident detection
• Accelerated incident response
Insurance
• Risk identification
• Personalised pricing
• Client support
Events
• Facial recognition to scan attendees
• Personalised recommendations
• Sales chatbots
Politics & Government
• Targeted campaigning
• Public opinion monitoring
• Anticipating infrastructure failures and maintenance

Plan du cours

- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

Définition de l'apprentissage automatique ou Machine Learning (ML)

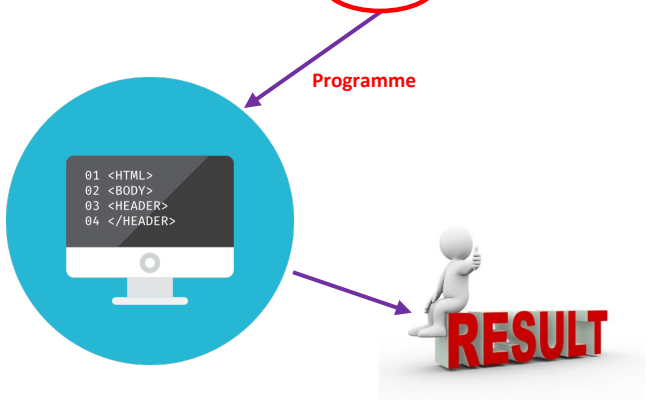
Pourquoi l'homme a inventé l'ordinateur ?

pour l'aider à résoudre en quelques minutes des calculs qui lui prendraient beaucoup d'années.

Donc un ordinateur ne sait faire qu'une seule chose : résoudre les calculs qu'on lui donne.

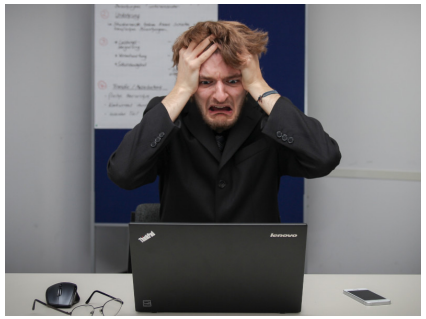
Pour résoudre un problème, 2 situations sont possibles

1. On connait le calcul* qui résout notre problème



Calcul = Formules ou algorithmes

2. On ne connait pas le calcul* qui résout notre problème



Mais si je peux avoir une base de données regroupant de nombreuses instances du problème considéré.

Définition de l'apprentissage automatique ou Machine Learning (ML)

Qu'est-ce qu'apprendre ?

D'après Fabien Benureau (2015) : « *L'apprentissage est une modification d'un comportement sur la base d'une expérience* ». Cette définition s'applique à un programme informatique, un robot, un animal de compagnie ou un être humain.

Qu'est-ce que le ML?

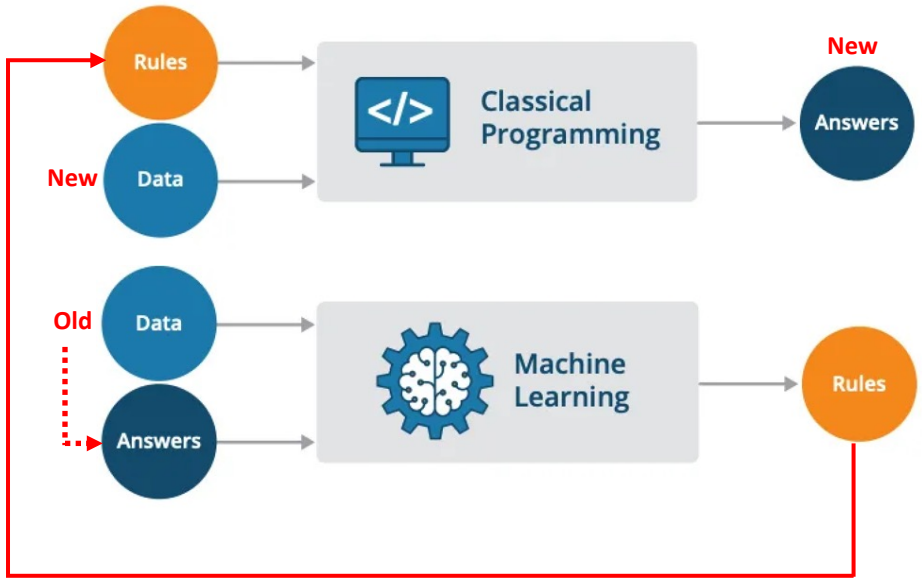
D'après Arthur Samuel, c'est un programme qui a la capacité d'apprendre sans que cette modification d'un comportement ne soit explicitement programmée.

***Arthur Samuel**, un mathématicien américain qui a développé un programme pouvant **apprendre tout seul** comment jouer aux Dames en **1959**.*

Définition de l'apprentissage automatique ou Machine Learning (ML)

Le ML consiste alors à programmer à l'aide des algorithmes permettant d'apprendre automatiquement, des données et des expériences passées, un algorithme cherchant à résoudre au mieux un problème considéré.

Le ML est donc utilisé quand les *données* sont abondantes (relativement), mais les *connaissances* peu accessibles ou peu développées.



Un programme *classique* utilise une procédure et les données qu'il reçoit en entrée pour produire en sortie des réponses.

Un programme *d'apprentissage automatique* utilise les données et leurs réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.

Définition de l’apprentissage automatique ou Machine Learning (ML)

Selon une enquête du Groupe RELX, plus de 91 % des cadres supérieurs interrogés ont compris que l'intelligence artificielle et l'apprentissage automatique étaient importants, mais seulement 56 % les utilisaient et 18 % ignoraient comment ces technologies étaient utilisées dans leur entreprise.

Les principaux cas d'usage de ces technologies sont :

- l'amélioration de la productivité des travailleurs (51 %),
- la prise de meilleures décisions d'affaires (41 %)
- la rationalisation des processus opérationnels (39 %).

les entreprises dégagent des économies en automatisant les décisions (40 %), en conservant leurs clients plus longtemps (36 %) et en détectant plus facilement la fraude et le gaspillage (33 %).

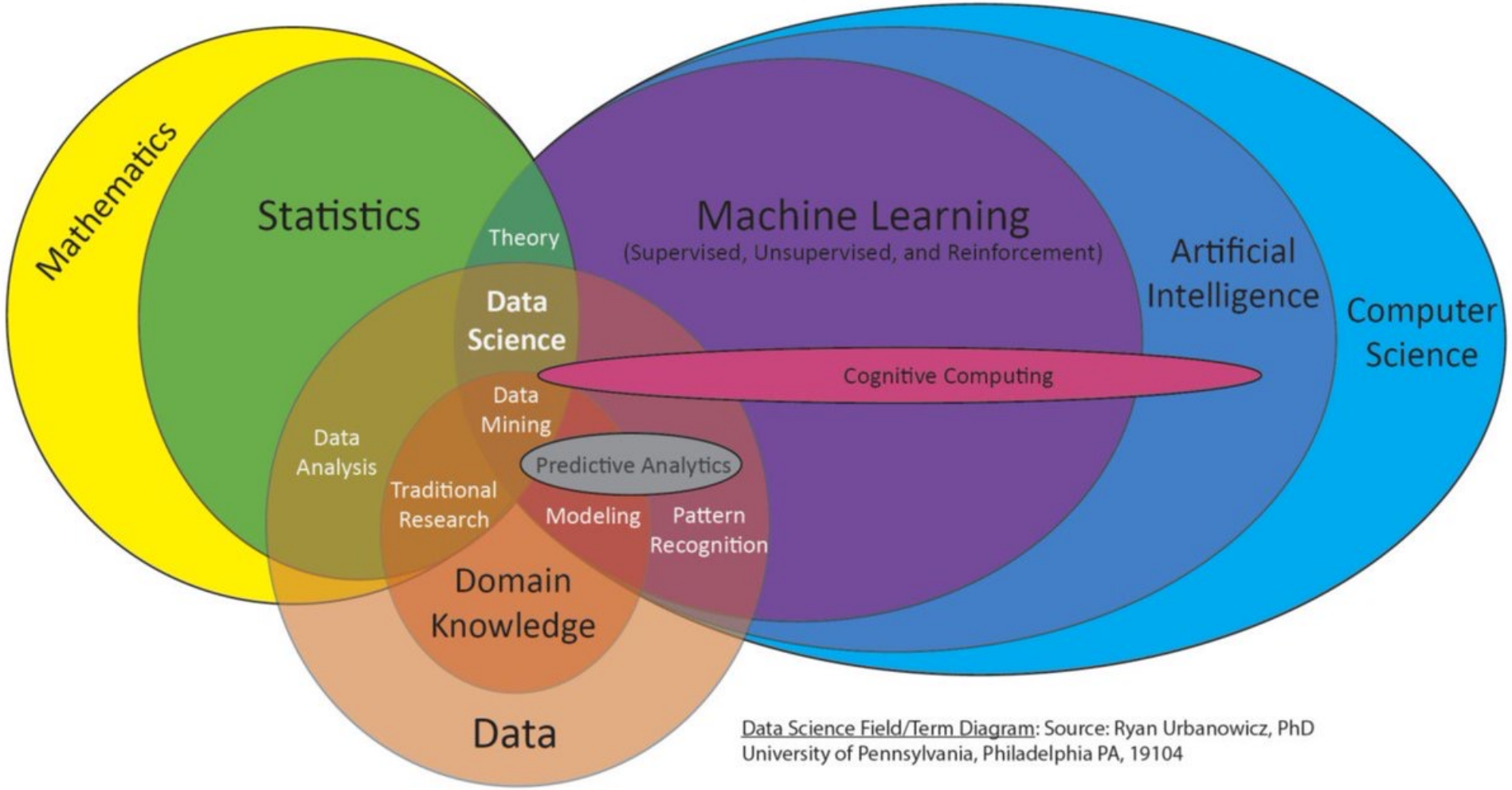
https://www.relx.com/media/press-releases/year-2018/emerging-tech in the US – RELX

ML : Un domaine pluridisciplinaire

Le ML est à la croisée de plusieurs disciplines par ex. :

1. **Les statistiques** : pour ce qui est de la construction de modèles et de leur inférence à partir de données.
2. **Les probabilités** : pour modéliser l'aspect aléatoire inhérent aux données et au problème d'apprentissage.
3. **L'intelligence artificielle :** pour étudier les tâches simples de reconnaissance de formes (comme la reconnaissance des chiffres, des lettres et n'importe quel autre forme), et parce qu'elle fonde une branche du ML dite symbolique qui repose sur la logique et la représentation des connaissances.
4. **L'optimisation** : pour optimiser un critère de performance afin, soit d'estimer des paramètres d'un modèle, soit de déterminer la meilleure décision à prendre, étant donné une instance d'un problème.
5. **L'informatique** : puisqu'il s'agit de programmer des algorithmes qui peuvent être de grande complexité et gourmands en termes de ressources de calcul et de mémoire.

ML : Un domaine pluridisciplinaire



Data Science Field/Term Diagram: Source: Ryan Urbanowicz, PhD
University of Pennsylvania, Philadelphia PA, 19104

ML : A quoi ça sert ?

Le ML peut servir à résoudre des problèmes qu'on :

- ne sait pas résoudre car on n'a pas la connaissance pour le faire (comme prédire le cours de la Bourse, …) ;
- 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) ;
- sait résoudre, mais avec des procédures beaucoup trop gourmandes en ressources informatiques et humaines (c'est le cas par exemple du trading automatique où les algorithmes ML peuvent analyser simultanément des milliers de sources de données, ce que les traders humains ne peuvent pas réaliser).

Le ML peut aussi aider les humains à apprendre : les modèles créés par des algorithmes d'apprentissage peuvent révéler l'importance relative de certaines informations ou la façon dont elles interagissent entre elles pour résoudre un problème particulier. Cet aspect du ML est très utilisé dans la **recherche scientifique**, **le marketing** et **la finance**.

ML : Les techniques et les domaines

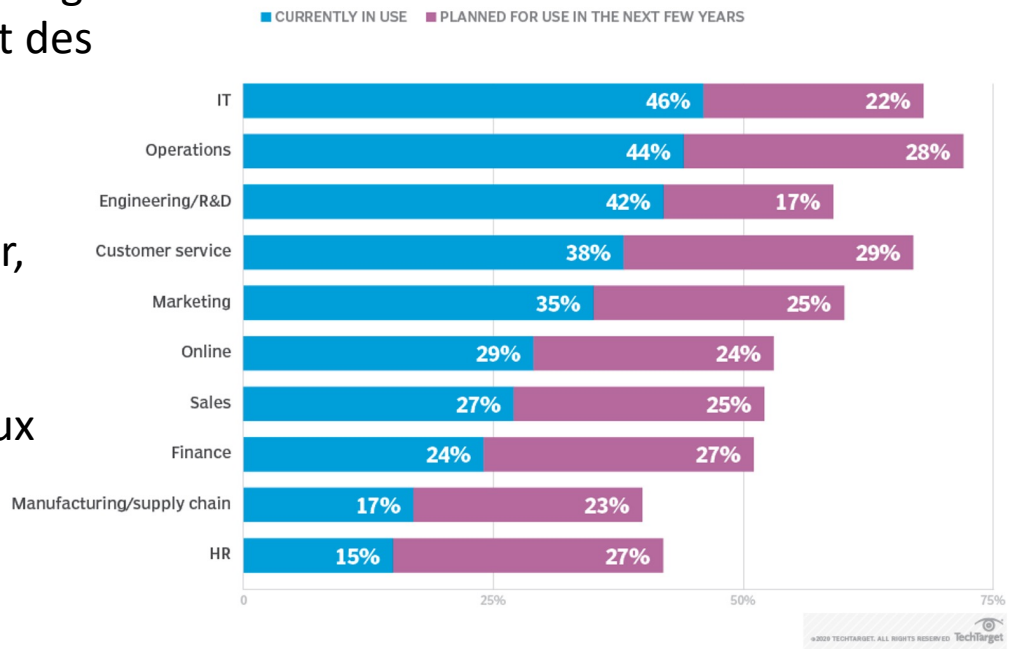
La plupart des techniques utilisées dans le ML reposent sur des théories mathématiques (statistiques avancées, arbres de décision, réseaux bayésiens, réseaux de neurones,…) qui sont connues depuis **50 ans sinon plus**.

Elles deviennent de plus en plus utilisées aujourd'hui du fait de la conjonction de certains facteurs :

- Des coûts de stockage des données en constante diminution
- Une puissance de calcul en constante augmentation
- Une explosion de la quantité d'information disponible sous forme digitale
- Ces informations sont en grande partie non structurée et requiert des techniques d'exploitation différentes des méthodes classiques.

Le ML est présent un peu partout aujourd'hui par ex. :

- les **banques** l'utilisent pour évaluer la solvabilité d'un emprunteur,
- Les **moteurs de recherche** pour présenter à l'internaute des résultats – et des publicités – adaptés à ses attentes.
- **Les applications** : la reconnaissance de textes ou d'images, les jeux de stratégie comme les échecs, ou la robotique et les voitures intelligentes comme dans la Google Car, …



Plan du cours

- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

ML : terminologie de base

Prenons l'exemple d'un ML pour prédire le prix d'un appartement.
Les **caractéristiques** peuvent inclure les éléments suivants :

- Le nom de la ville où se trouve l'appartement;
- L'étage où se trouve l'appartement;
- L'année de construction;
- La surface de l'appartement;
- Le nombre de pièces;
- Le nombre de chambres;
- ~~Les couleurs des murs~~

L'étiquette est le prix de l'appartement.

Exemples étiquetés

Ville	Etage	Année de Const.	surface	nb de pièce	nb de ch.	Prix K€
Créteil	1	2000	60	3	2	200
Noisy le grand	0	2010	120	5	3	300
choisy le roi	2	1988	50	3	2	150
Créteil	3	2001	70	4	3	300
Paris	1	1950	30	2	1	300

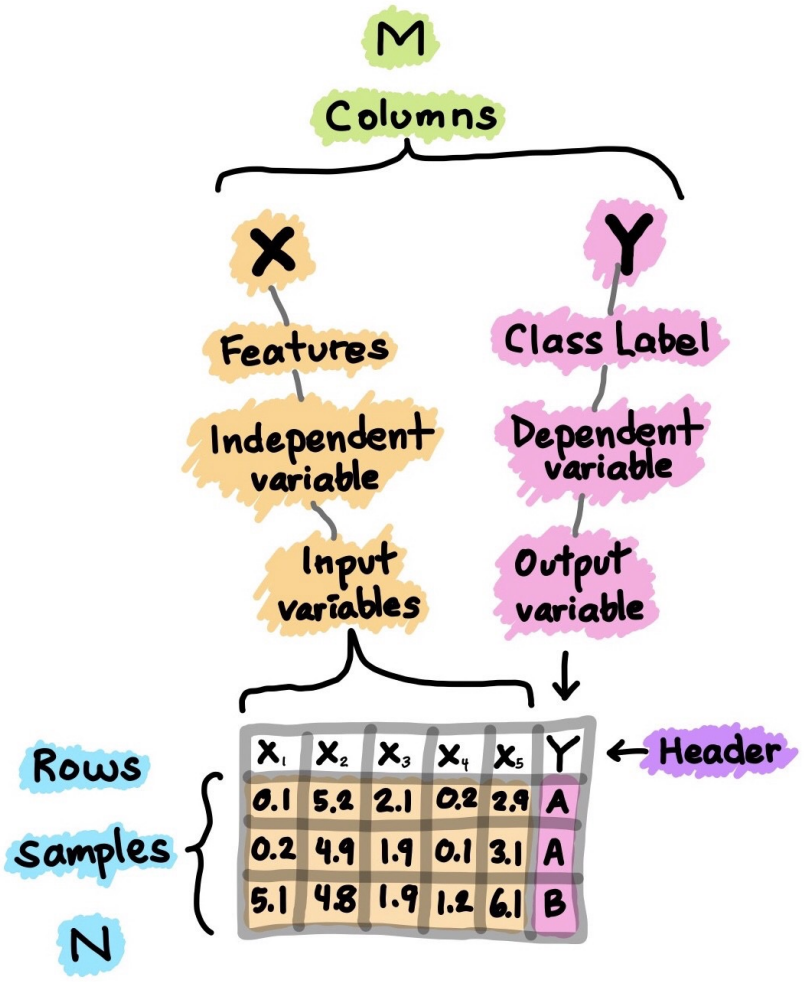
ML : terminologie de base

Etiquette (Label ou Target) : c'est le résultat de la prédiction; Il peut s'agir du cours à venir du blé, de l'espèce animale représentée sur une photo, de la signification d'un extrait audio ou de toute autre chose.

Caractéristique (Feature) : C'est une variable d'entrée; Un projet de ML simple peut utiliser une seule caractéristique, tandis qu'un projet plus sophistiqué, en utilisera plusieurs millions spécifiés sous la forme : x1,x2,...xN.

Un exemple est une instance de donnée particulière, **x**. (**x** est un vecteur.) Les exemples se répartissent dans deux catégories :

- ***Exemples étiquetés*** : comprend une ou plusieurs caractéristiques et l'étiquette : {features, label}: (x, y)
- ***Exemples sans étiquette*** : contient des caractéristiques, mais pas d'étiquette : {features, ?}: (x, ?).



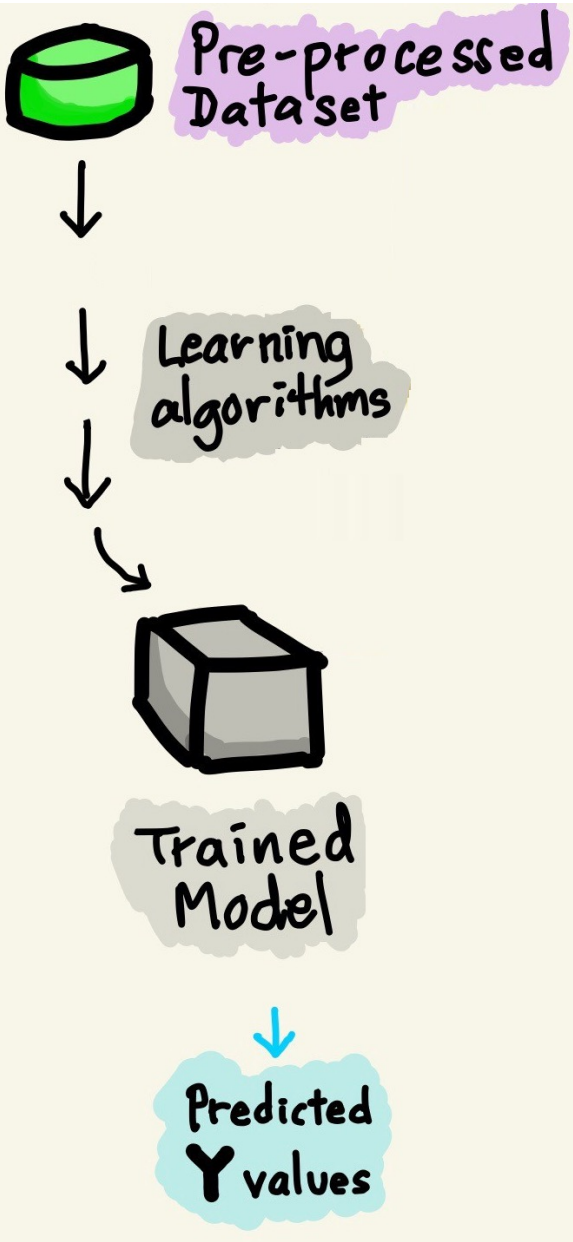
ML : terminologie de base

Modèle (Model) : il définit la relation entre les caractéristiques ou la relation entre les caractéristiques et l'étiquette si elle existe.

L'apprentissage consiste à créer ou à **entraîner** le modèle.

Algorithme d'apprentissage : c'est un algorithme qui apprend un modèle à partir d'exemples, par le biais d'un problème d'optimisation. Les algorithmes de ML peuvent être divisés selon la nature du problème qu'ils cherchent à résoudre, en apprentissage supervisé, non supervisé, semi-supervisé, et par renforcement.

L'inférence consiste à appliquer le modèle entraîné à des exemples sans étiquette. En d'autres termes, vous utilisez le modèle entraîné pour faire des prédictions efficaces.



A ne pas oublier

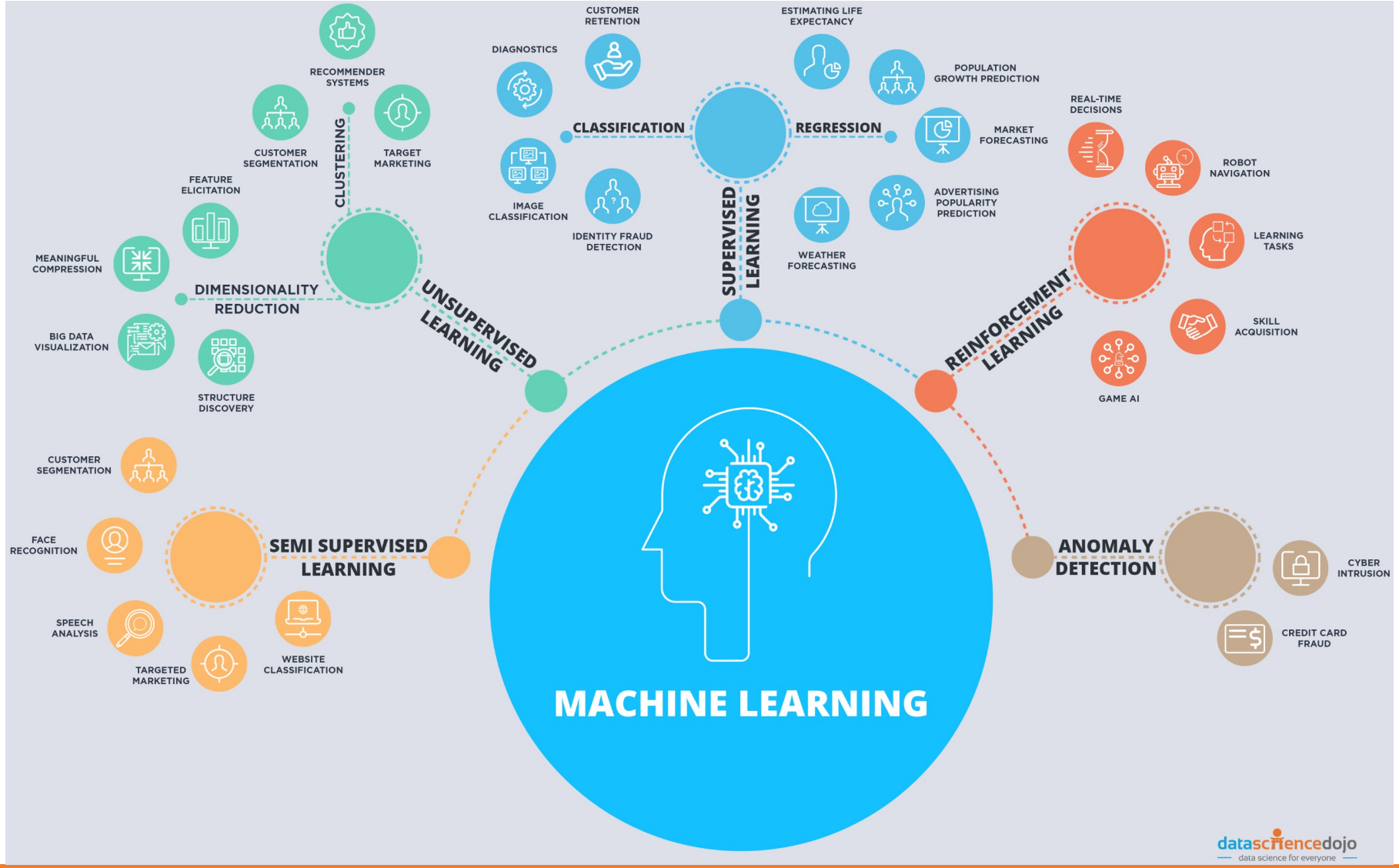
On utilise le ML **seulement** lorsqu'il est **difficile ou impossible** de définir les instructions explicites à donner à un ordinateur pour résoudre un problème, mais que l'on dispose de nombreux exemples illustratifs.

- Aucun algorithme d'apprentissage ne pourra créer un bon modèle à partir de données qui ne sont pas pertinentes : 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.
- Un modèle appris avec un algorithme inadapté sur des données pertinentes ne pourra pas être de bonne qualité.
- Un bon modèle est celui qui a un faible % d'erreurs faites sur les exemples d'apprentissage.

Plan du cours

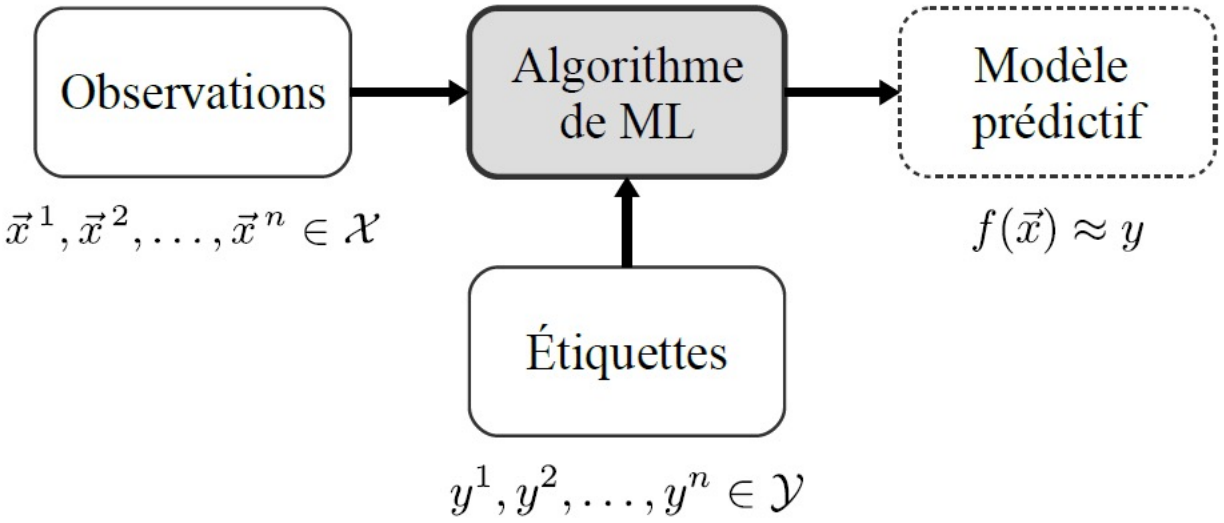
- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

Les Types de Machine Learning



Les Types de Machine Learning : Supervisé

Supervisé : c'est le type de problèmes de ML le plus facile à appréhender. Son but est d'apprendre à faire des *prédictions* à partir d'une liste d'exemples *étiquetés*; c'est-à-dire accompagner de la valeur à prédire. Les étiquettes (*labels*) servent de « professeur » et supervisent l'apprentissage de l'algorithme.



Les Types de Machine Learning : Supervisé

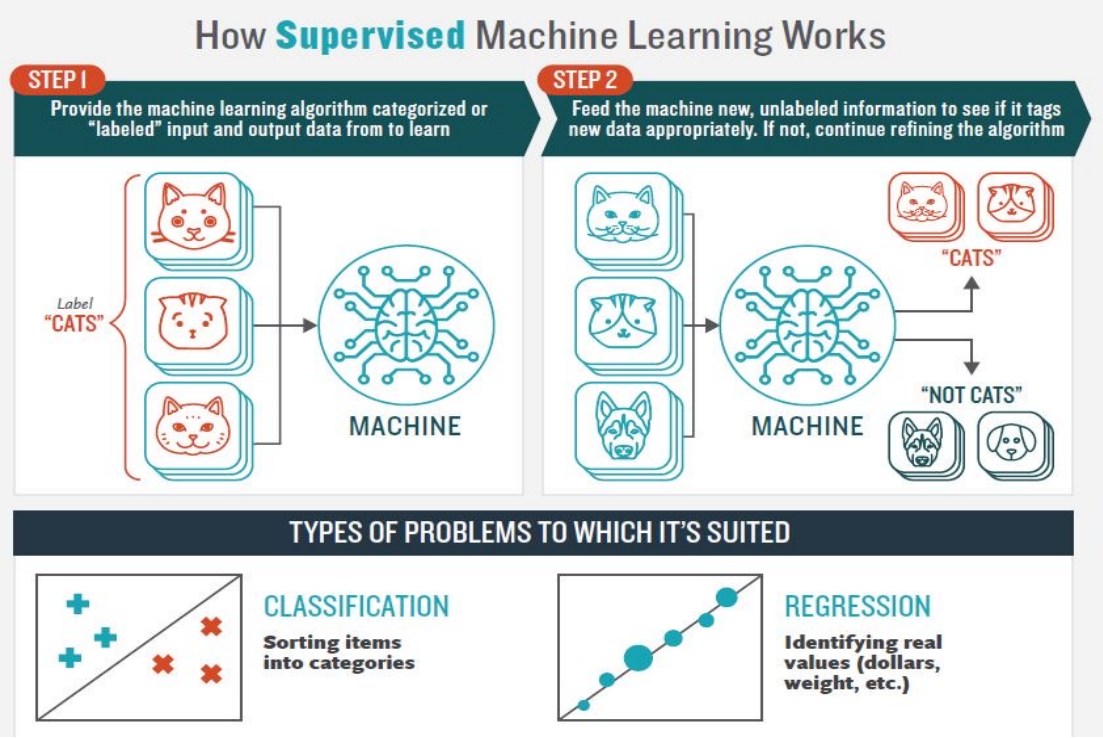
1- Dans le cas où les étiquettes sont *binaires*, elles indiquent l'appartenance à une *classe*. On parle alors de ***classification binaire***.

2- Dans le cas où les étiquettes sont *discrètes*, et correspondent donc à plusieurs *classes*, on parle de ***classification multi-classe***.

3- Dans le cas où les étiquettes sont à valeurs *réelles*, on parle de ***régression***.

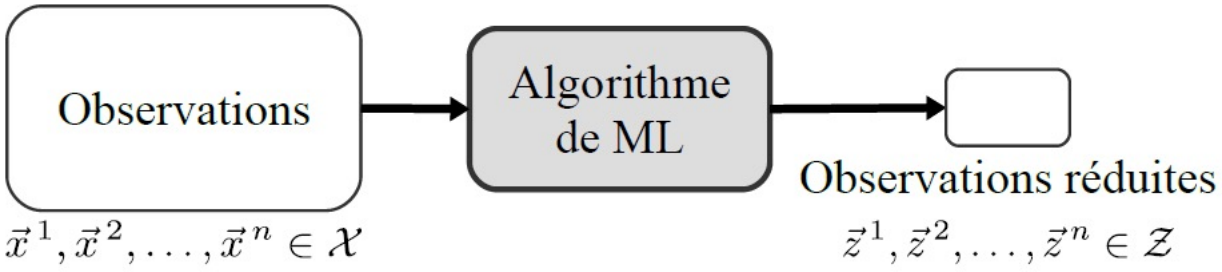
L'apprentissage supervisé tente de répondre à deux questions :

- *Classification : « quelle classe ?";*
- *Régression : « combien ? ».*

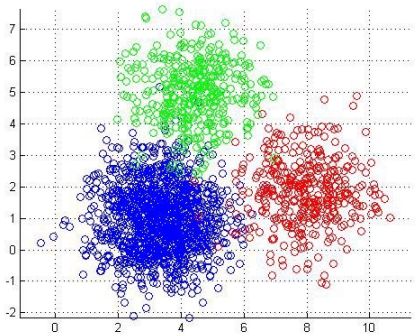


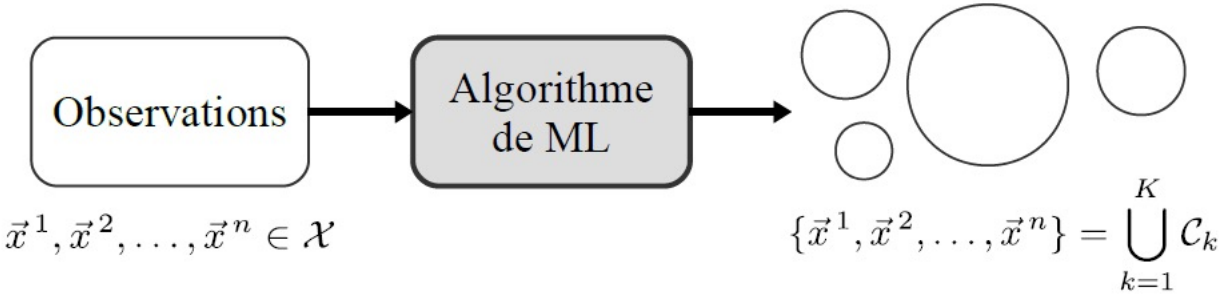
Les Types de Machine Learning: Non supervisé

Non supervisé : on dispose d'un ensemble d'objets sans aucune valeur cible associée ; il faut apprendre un modèle capable d'extraire les régularités présentes au sein des objets pour mieux visualiser ou appréhender la structure de l'ensemble des données.

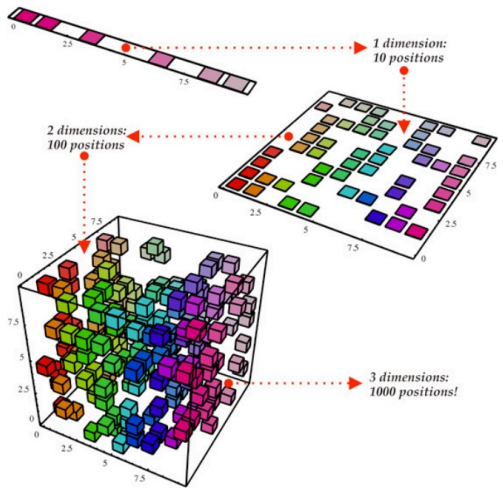


1- Le **clustering** consiste à identifier des groupes dans les données. Cela permet de comprendre leurs caractéristiques générales, et éventuellement d'inférer les propriétés d'une observation en fonction du groupe auquel elle appartient.



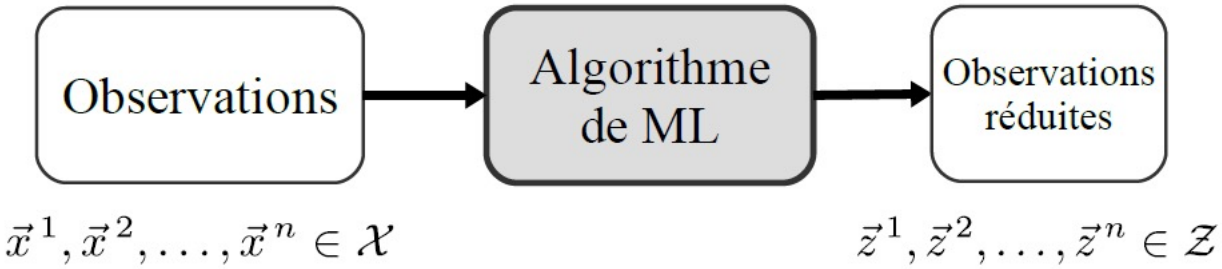


Les Types de Machine Learning: Non supervisé



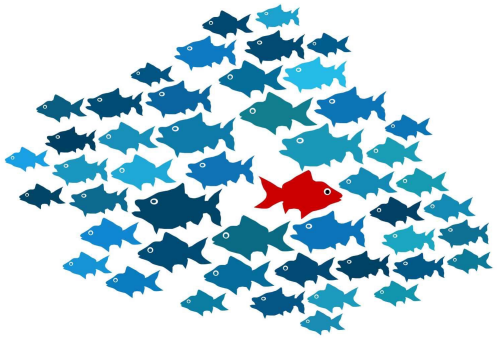
2- La **réduction de dimension** consiste à trouver une représentation des données dans un espace de dimension plus faible que celle de l'espace dans lequel elles sont représentées à l'origine afin de:

- réduire les temps de calcul,
- réduire l'espace mémoire nécessaire au stockage des données,
- améliorer les performances d'un algorithme d'apprentissage supervisé entraîné par la suite sur ces données.

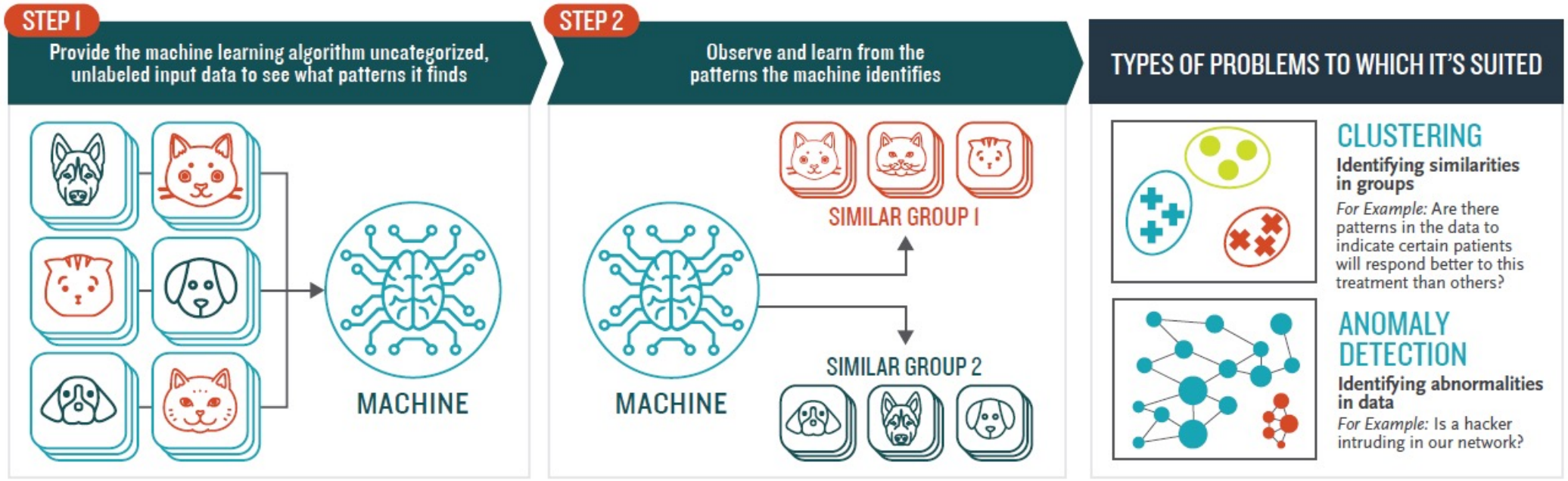


$\{\vec{z}^i\}_{i=1,\ldots,n}$
est la projection de $\{\vec{x}^i\}_{i=1,\ldots,n}$
dans $\mathcal{Z}$

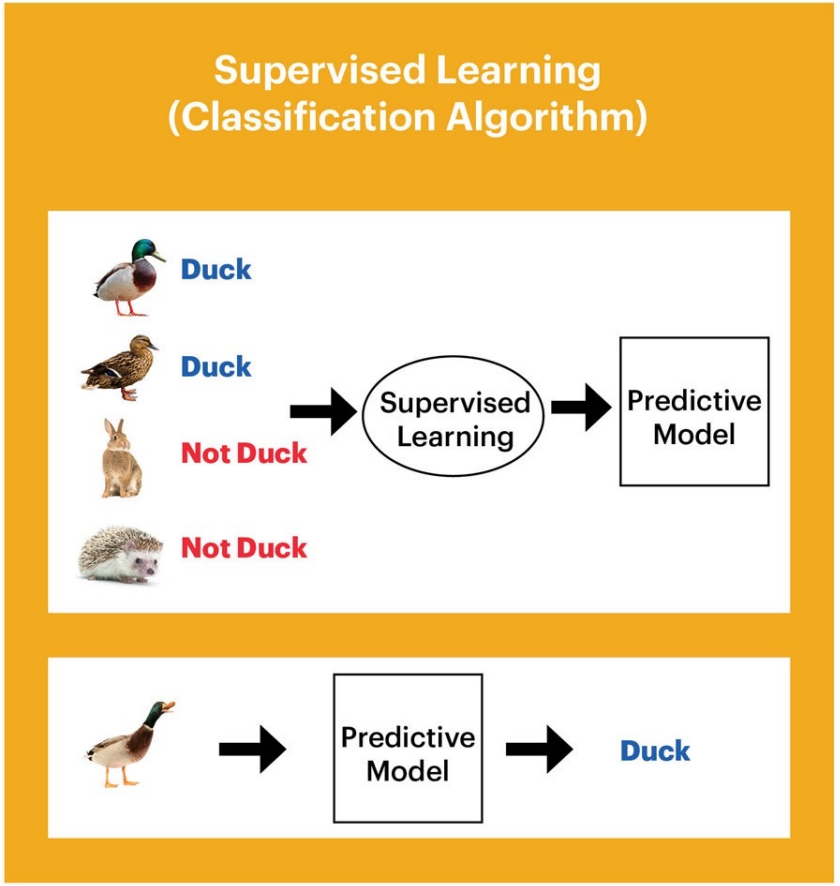
3- La **détection d'anomalie** consiste à repérer des données qui ne sont pas conformes à ce à quoi l'on peut s'attendre par rapport aux autres données. Il s'agit, par exemple, de données qui ne suivent pas le même schéma ou qui sont atypiques pour la distribution de la probabilité observée.

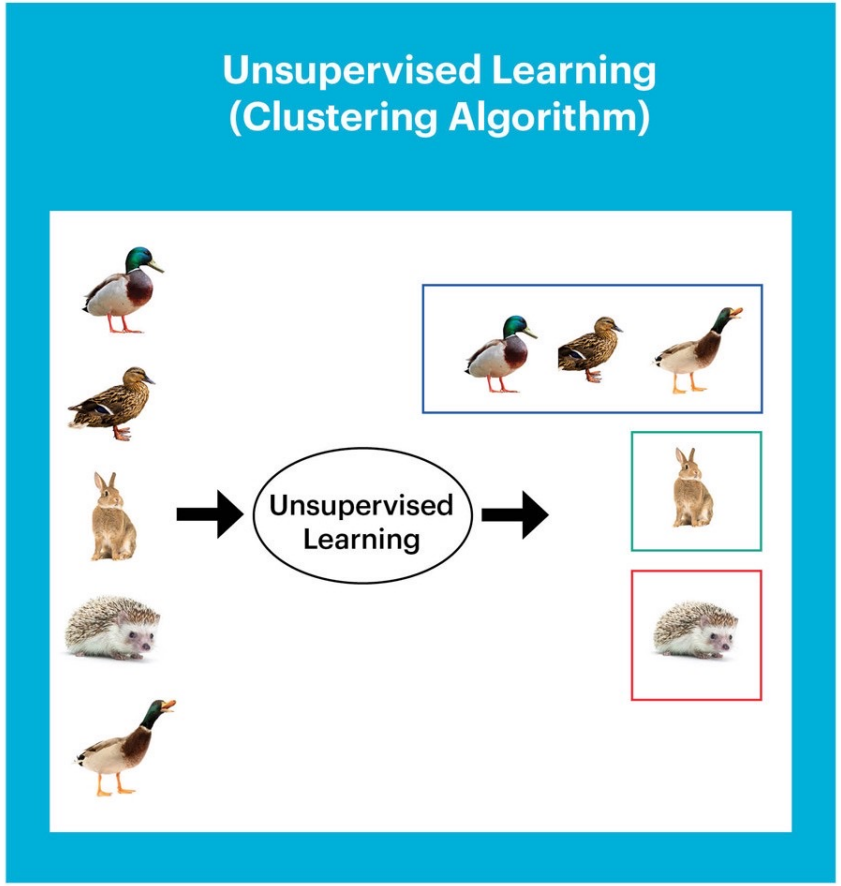


Les Types de Machine Learning: Non supervisé



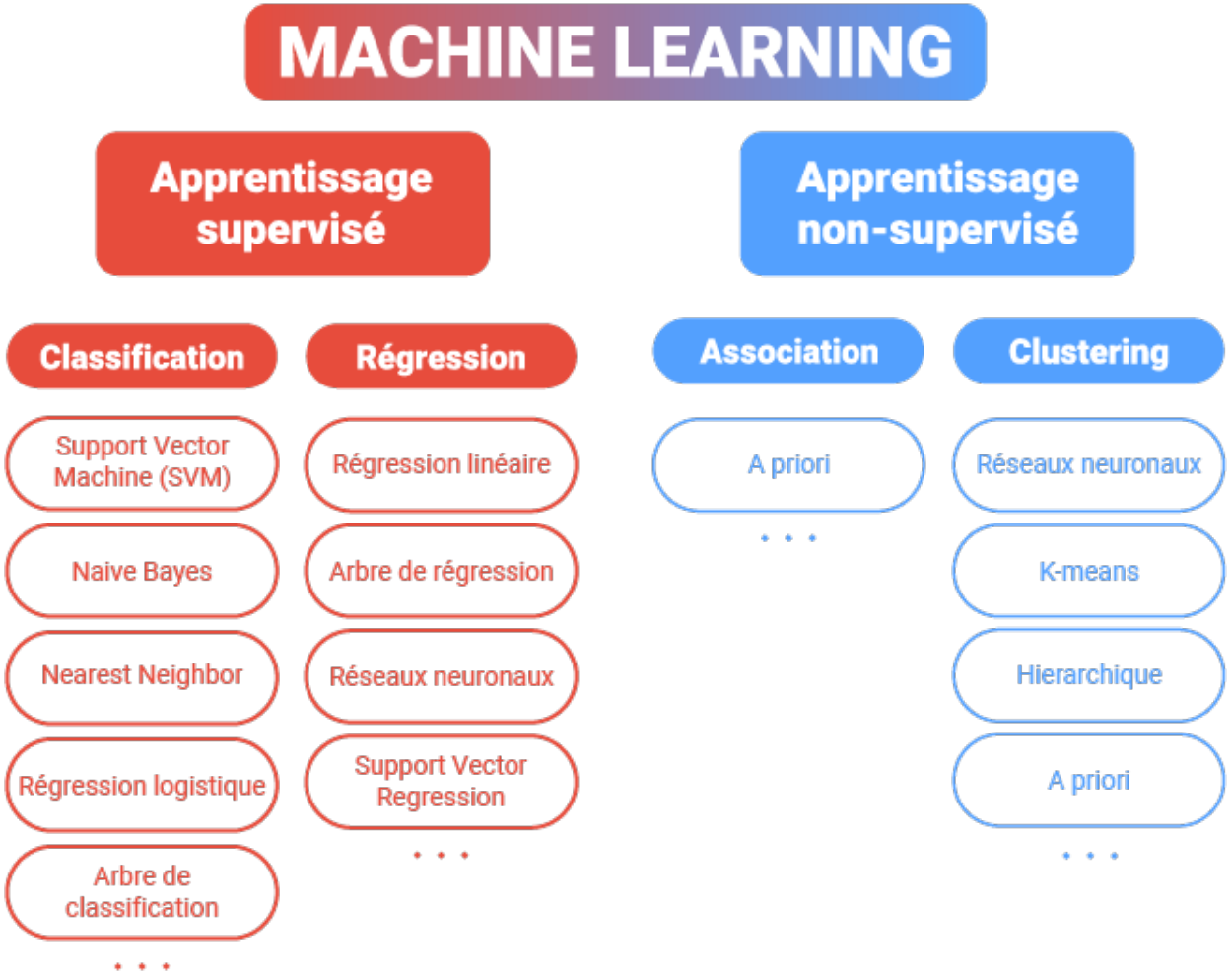
Les Types de Machine Learning : Supervisé vs Non supervisé





Western Digital.

Supervisé et Non supervisé : Principaux algorithmes.

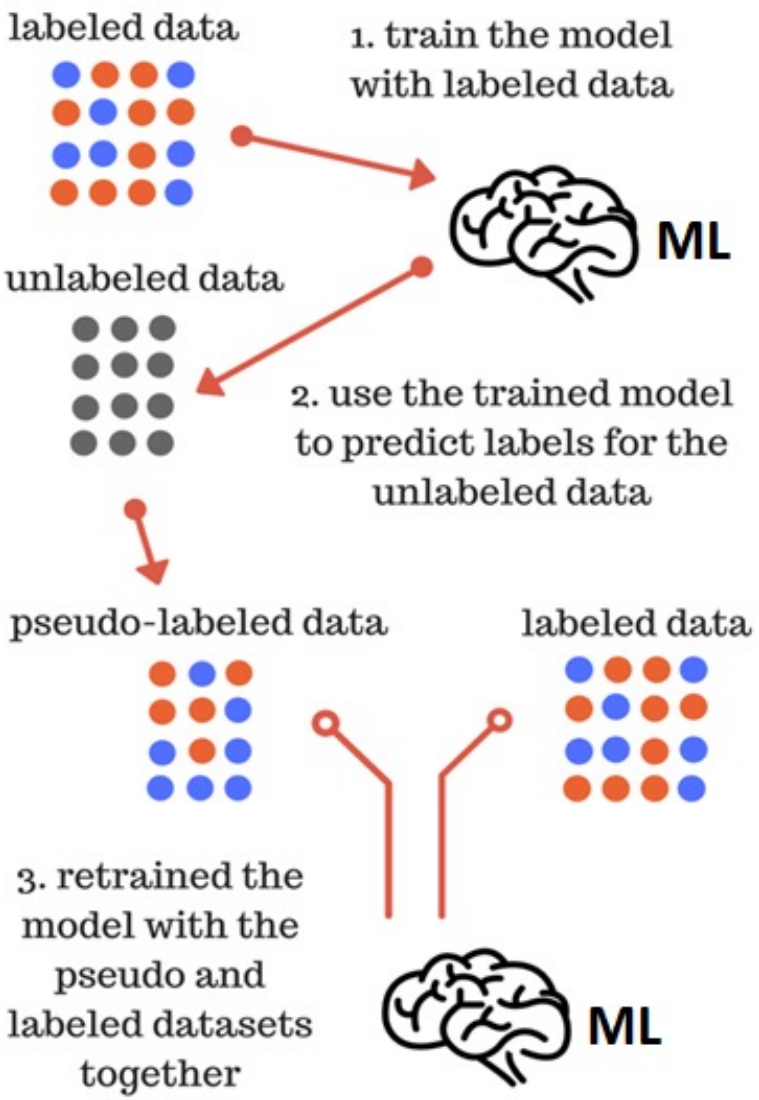


Les Types de Machine Learning : Semi-supervisé

- **Semi-supervisé :** consiste à apprendre des étiquettes à partir d'un jeu de données partiellement étiqueté.

  L'avantage de cette approche est qu'elle permet d'éviter d'avoir à étiqueter l'intégralité des exemples d'apprentissage; ce qui est pertinent, quand il est facile d'accumuler des données mais que leur étiquetage requiert une certaine quantité de travail humain.

  Prenons par exemple la classification d'images : il est facile d'obtenir une banque de données contenant des centaines de milliers d'images, mais avoir pour chacune d'entre elles l'étiquette qui nous intéresse peut requérir énormément de travail.



Les Types de Machine Learning : Par renforcement

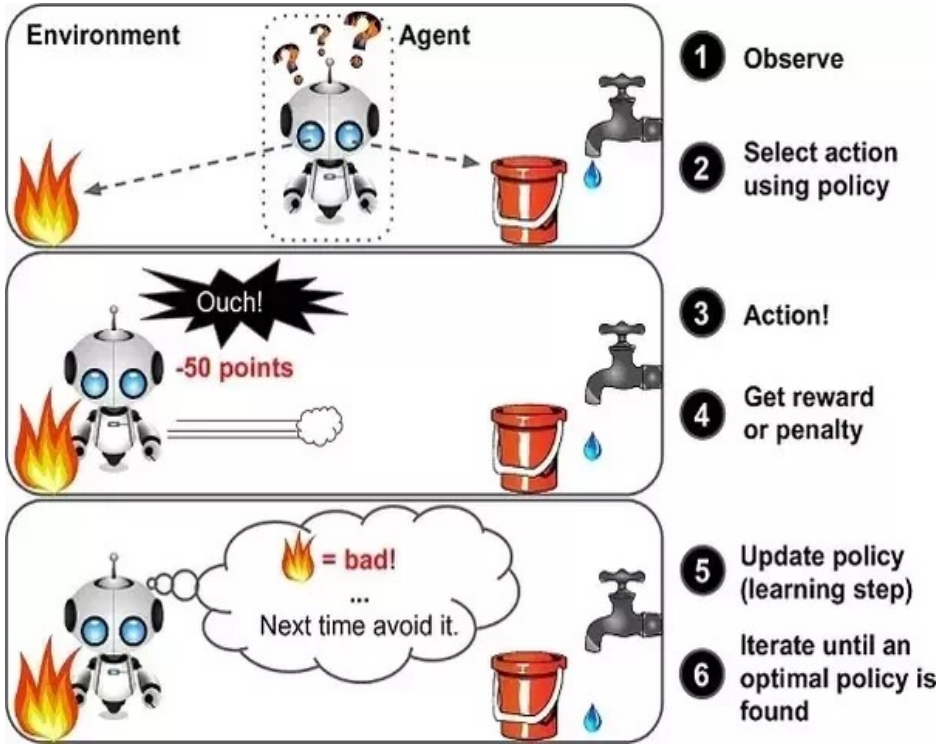
Par renforcement : Dans ce type, le système d'apprentissage peut interagir avec son environnement et accomplir des actions.

En retour de ces actions, il obtient une *récompense*, qui peut être positive si l'action était un bon choix, ou négative dans le cas contraire.

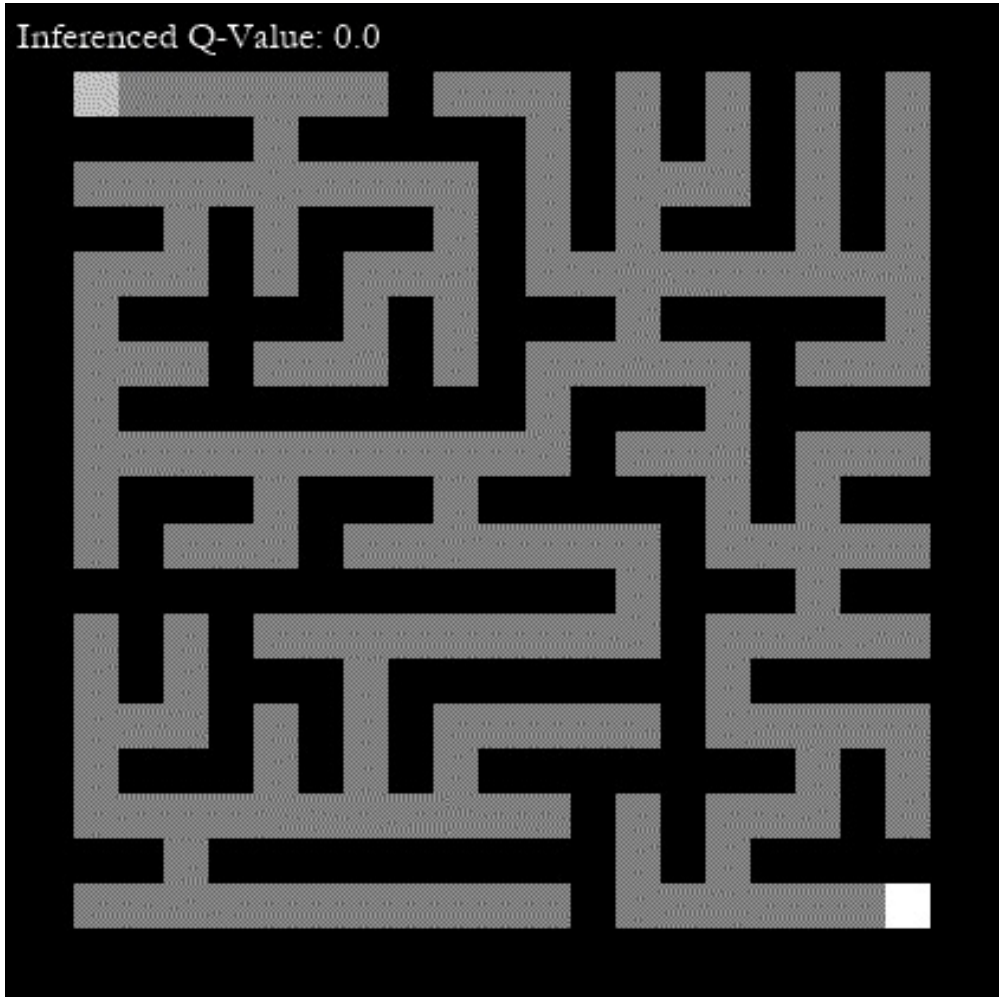
La récompense peut parfois venir après une longue suite d'actions ; c'est le cas par exemple pour un système apprenant à jouer aux échecs.

Ainsi, l'apprentissage consiste dans ce cas à définir une *politique* ou une stratégie permettant d'obtenir systématiquement la meilleure récompense possible.

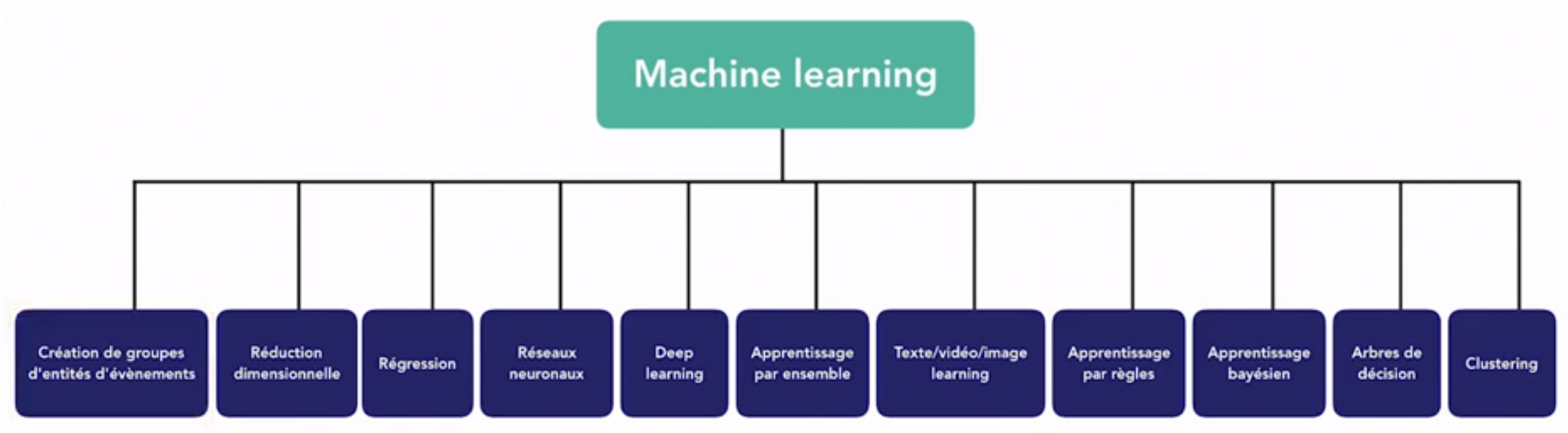
Ses applications principales se trouvent dans les jeux et la robotique des véhicules autonomes.



Les Types de Machine Learning : Par renforcement



Hiérarchie d'algorithmes et méthodes de machine learning



A notre jour il y a plus de 250 méthodes et algorithmes de machine leaning

Quiz

a) Alice veut écrire un programme qui utilise la fréquence des mots « science », « public », « accès », « université », « gouvernement », « financer », « éducation », « budget », « justice » et « loi » pour déterminer si un article traite ou non de la politique scientifique. Elle a commencé par annoter un millier d'articles selon leur sujet. Quel genre de problème d'apprentissage automatique doit-elle résoudre ?

b) Parmi les problèmes suivants, lesquels se prêtent bien à être traités par le Machine Learning ?

   1. Déterminer l'horaire optimal pour poster un contenu sur une page web.
   2. Déterminer le chemin le plus court entre deux nœuds dans un graphe.
   3. Prédire le nombre de vélos à mettre en location à chaque station d'un système de location de vélos citadins.
   4. Évaluer le prix qu'un tableau de maître pourra atteindre lors d'une vente aux enchères.
   5. Débruiter un signal radio.

c) Benjamin dispose de 10 000 articles de journaux qu'il souhaite classer par leur thématique. Doit-il utiliser un algorithme supervisé ou non supervisé ?

Quiz

d) Les données de Cécile sont décrites par 10 variables. Elle aimerait cependant les représenter sur un graphique en deux dimensions. Quel type d'algorithme d'apprentissage doit-elle utiliser ?

e) David gère un outil qui permet d'organiser les liens HTML qui ont été sauvegardés. Il souhaite suggérer des catégories auxquelles affecter un nouveau lien, en fonction des catégories déjà définies par l'ensemble des utilisateurs du service. Quel type d'algorithme d'apprentissage doit-il utiliser ?

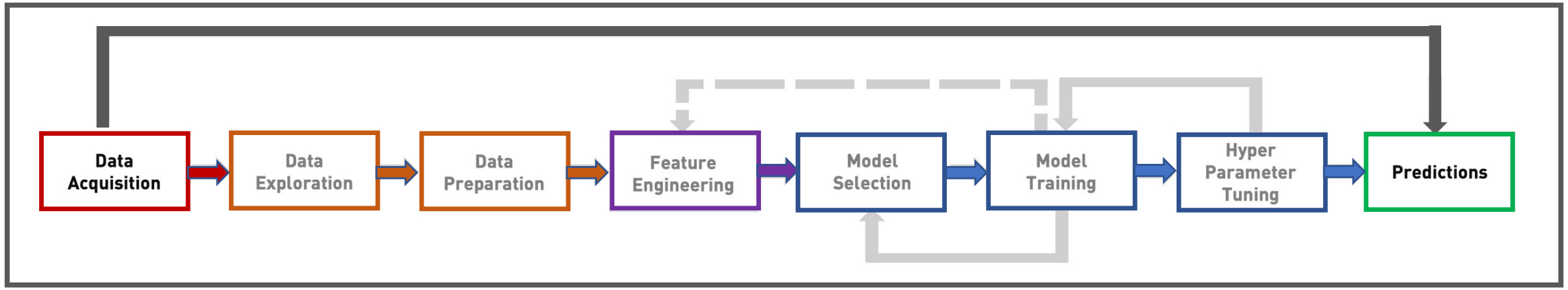
f) Elsa veut examiner ses spams pour déterminer s'il existe des sous-types de spams. Quel type d'algorithme d'apprentissage doit-elle utiliser ?

g) Tom Mitchell définit le ML comme suit : Un programme informatique est dit: « apprendre de l'expérience E pour la tâche T et une mesure de performance P si sa performance sur T, comme mesurée par P, s'améliore avec l'expérience E ». Fred écrit un programme qui utilise des données bancaires dans le but de détecter la fraude bancaire. Que sont E, T, et P ?

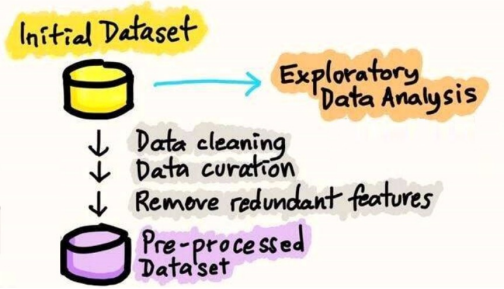
Plan du cours

- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

Machine Learning : Le cycle de vie



Machine Learning - Le cycle de vie : la préparation des données



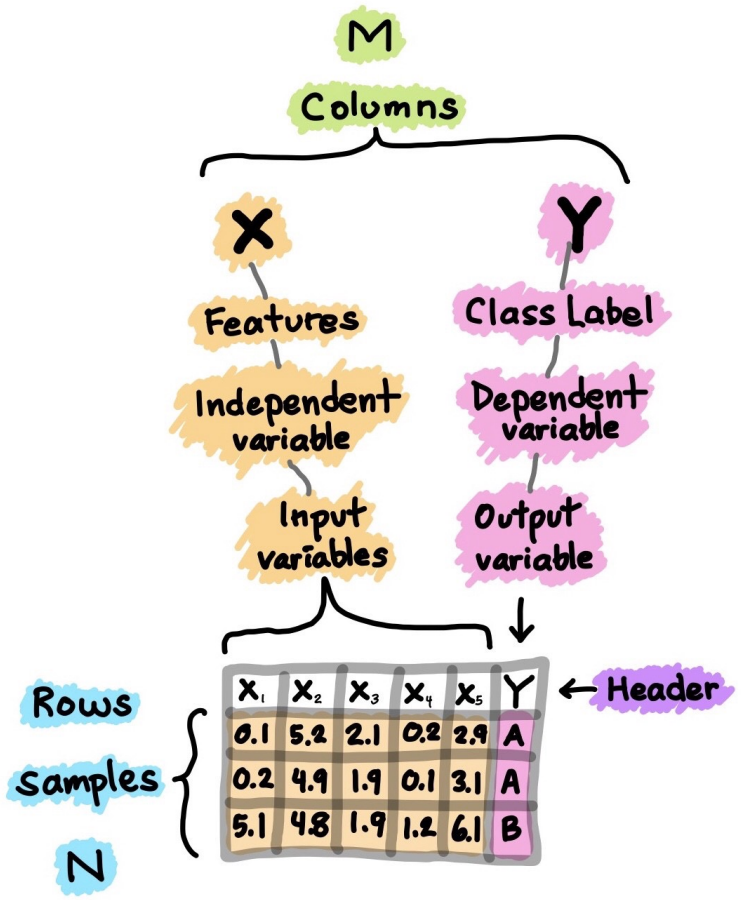
- **Récolte des données** : rassembler les données nécessaires pour le ML sous une forme consolidée dans un seul tableau (Flat Table).

- **Réconciliation (Data Wrangling)** : préparer les données afin de les rendre exploitables par les algorithmes d'apprentissage automatique.
  - *Nettoyage des données* : supprimer les « Null », les valeurs manquantes et les doublons.
  - *Décomposition des données* : diviser les colonnes de texte contenant plus d'une information.
  - *Agrégation de données* : regrouper certaines informations ensemble quand c'est pertinent.
  - *Mise à l'échelle (Data Scaling)* : mettre les données à une échelle commune. La mise à l'échelle des données ne s'applique pas à l'étiquette.
  - *Mise en forme et transformation* (Data Shaping & Transformation) : de catégoriel à numérique.

- **Enrichissement des données** : enrichir les données existantes par des données externes afin de donner à l'algorithme plus d'informations avec lesquelles travailler, ce qui améliore le modèle (par exemple, des données économiques ou météorologiques).

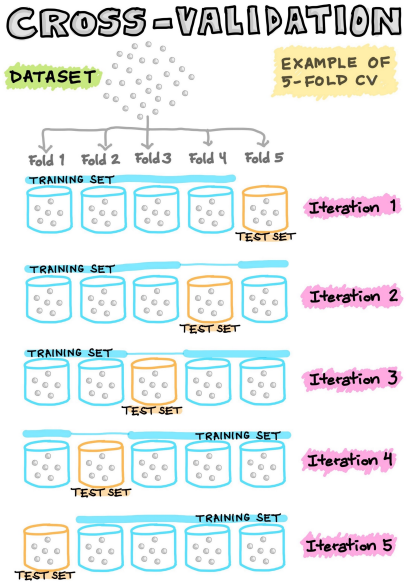
Machine Learning - Le cycle de vie : l'ingénierie des caractéristiques (*Feature Engineering*)

- Visualisez les données dans leur ensemble pour voir s'il existe des liens entre les colonnes (entre les *caractéristiques*, et entre les *caractéristiques* et l'*étiquette)*.
  - si une *caractéristique* est directement dépendante d'une autre => pas besoin des deux *caractéristiques*.
  - les liens entre une *caractéristique* et l'*étiquette* nous permettent de voir si une *caractéristique* aura un fort effet sur le résultat.
- Générer des *caractéristiques* supplémentaires à partir de celles qui existent déjà dans une classification lorsque l'algorithme choisi est incapable de différencier correctement les classes.
- Bien choisir les colonnes à utiliser comme *caractéristiques*.

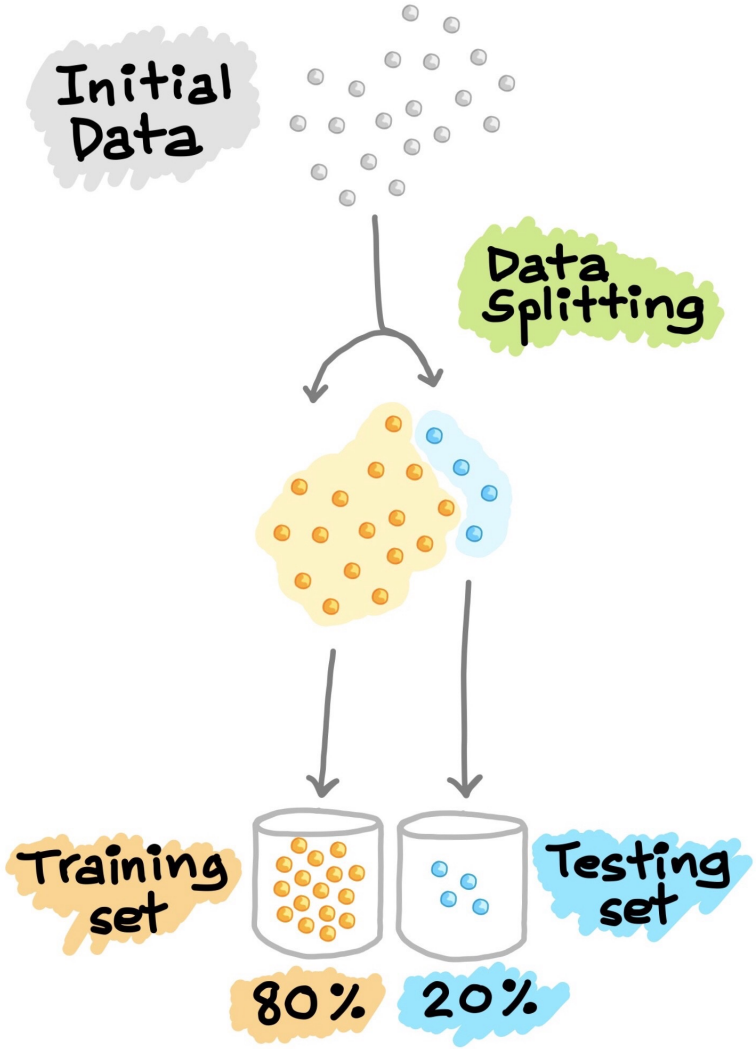


Machine Learning - Le cycle de vie : Découpage des données

- Diviser l'ensemble de données en deux parties avant d'entraîner les algorithmes :
  - Les données d'entraînement serviront à entraîner le ou les algorithmes choisis ;
  - Les données de test seront utilisées pour vérifier la performance du résultat ;

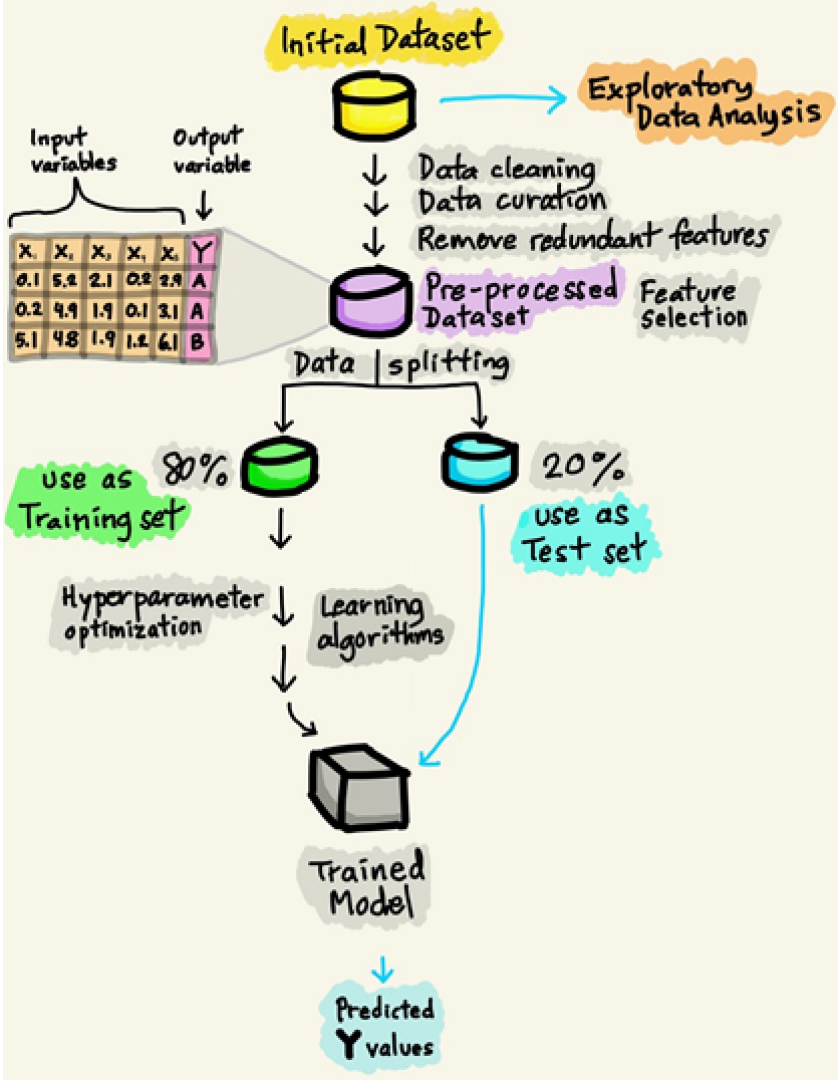


Afin d'utiliser d'une manière économique les données disponibles, nous utilisons **la validation croisée (CV)**. L'ensemble de données est découpé en N partitions (en général 5 ou 10). Avec cette méthode nous effectuons N itérations. Dans chaque itération, une des partitions est utilisée en tant que données de test, tandis que les autres sont utilisées comme données d'apprentissage pour la construction du modèle.



Machine Learning - Le cycle de vie : le choix d'algorithme

- Choisir le ou les algorithmes pertinents.
- Essayer les algorithmes avec différentes combinaisons de paramètres et comparer les performances de leurs résultats.
- Utiliser la procédure des hyperparamètres pour essayer de nombreuses combinaisons, et trouver celle qui donne le meilleur résultat (n'essayer pas de combinaisons manuelles).
- Sauvegarder les modèles satisfaisants même s'ils ne sont pas parfaits, pour éviter le risque de ne plus jamais retrouver la combinaison qui a permis de l'obtenir.
- Continuer les essais après chaque nouveau modèle. Si les résultats ne sont pas satisfaisants, il faut recommencer :
  - Au stade des hyperparamètres, si vous avez de la chance.
  - Recommencer à partir du choix des caractéristiques ou même revoir l'état des données si nécessaire.



Machine Learning - Le cycle de vie : Mesurer la performance d'un modèle

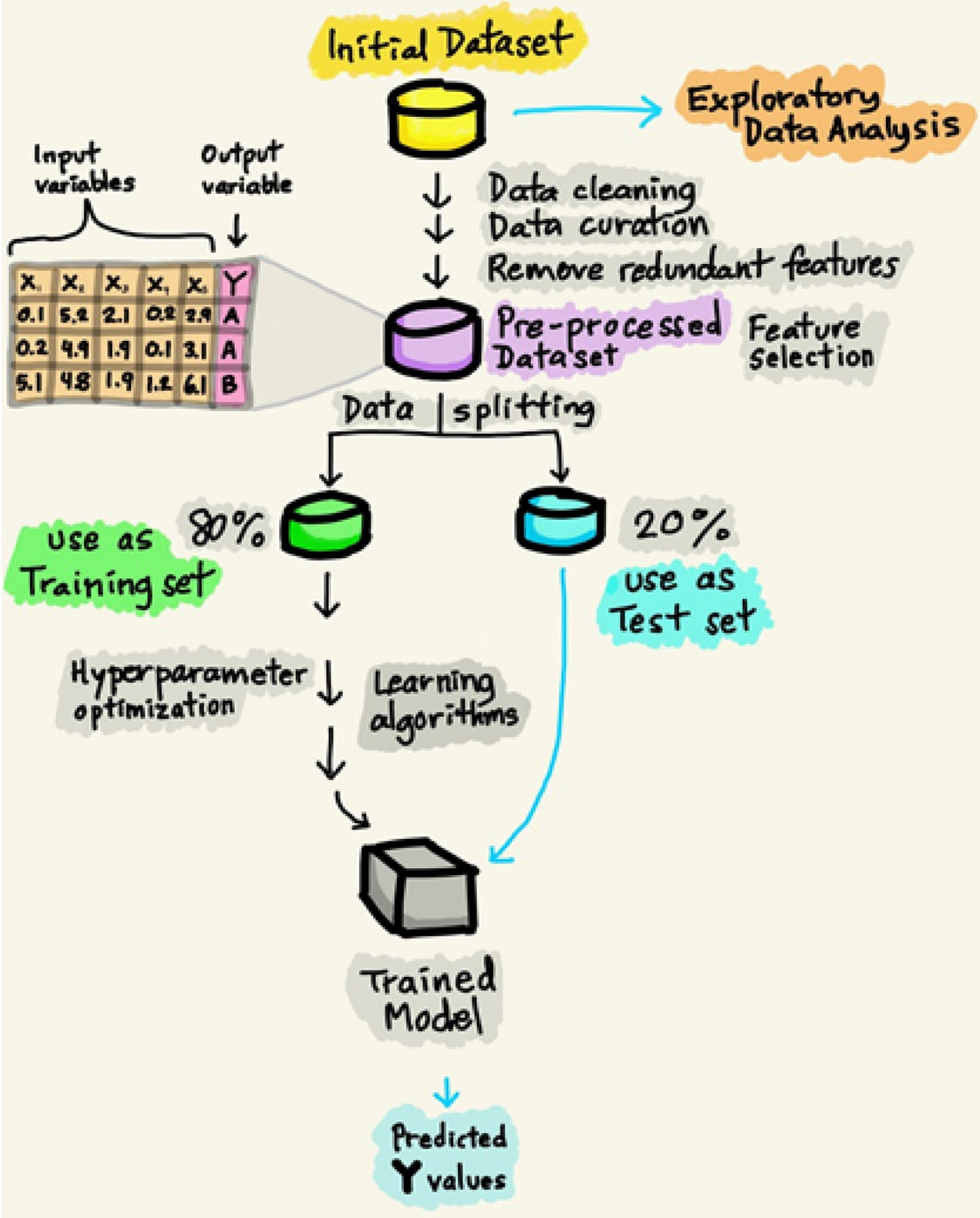
Une matrice de confusion (Confusion Matrix) est un outil permettant de mesurer les performances d'un modèle de Machine Learning en vérifiant notamment à quelle fréquence ses prédictions sont exactes par rapport à la réalité dans des problèmes de classification.

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

- « sensitivity » or « recall » : est le taux de vrais positifs,
- « specificity » : est le taux de vrais négatifs,
- « precision » : est la proportion de prédictions correctes parmi les points que l'on a prédits positifs.
- « accuracy » : est le pourcentage de bonnes prédictions.

L'accuracy, la sensitivity et la precision sont à utiliser ensemble pour donner une vision complète de la performance.
On cherche à avoir des valeurs le plus proche possible de 100% pour les 3 indicateurs.

Machine Learning
Le cycle de vie



Plan du cours

- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

Les défis du Machine Learning : Les problématiques des modèles

Le choix du modèle

Les modèles de Machine Learning sont nombreux et variés, capables de répondre à nos besoins mais ils ne sont pas tous adaptés.

Choisir un mauvais modèle conduira à de mauvais résultats et potentiellement à de mauvaises décisions.

Qu'est ce qu'un modèle adapté ?

- Capable de procéder à des prédictions au travers des données utilisées, mais également avec de nouvelles données pouvant apparaître avec le temps ;
- Permettant d'obtenir le(s) résultat(s) désiré(s) avec le minimum d'écart;
- S'accordant avec la typologie des données observées (textes ou images ou autres formats) ;
- Prenant en compte les contraintes de ressources informatiques (puissance du serveur et taille de la mémoire dédiée) ainsi que le temps de calcul toléré par les utilisateurs (certains modèles vont générer des résultats plus rapidement que d'autres) ;
- Susceptible de fournir le niveau de précision attendue.

Les défis du Machine Learning : Les problématiques des modèles

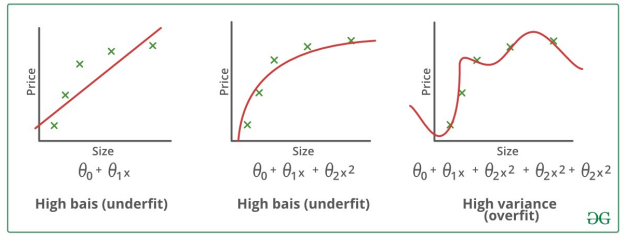
Pour éviter un choix inadapté, une véritable pré-étude est nécessaire.
Il conviendra pour cela de déterminer plusieurs caractéristiques :

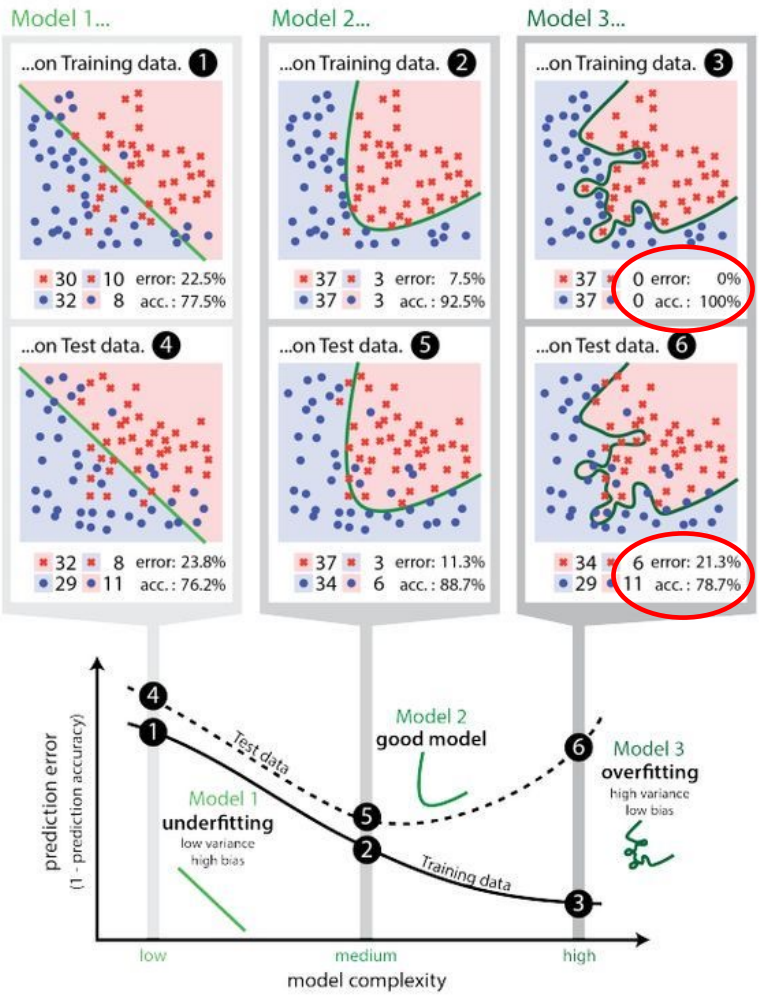
- Le contexte et les motivations relatives au besoin
- L'objectif lié à l'utilisation du Machine Learning
- Le jeu de données mis à disposition et sa forme
- Les limitations (techniques, temporelles, savoir-faire)

Les défis du Machine Learning : Les problématiques des modèles

Le surapprentissage « overfitting »
Il s'agit du cas où le modèle est tellement efficace qu'il ne s'adapte plus à de nouvelles données lui étant inconnues et perd sa précision.

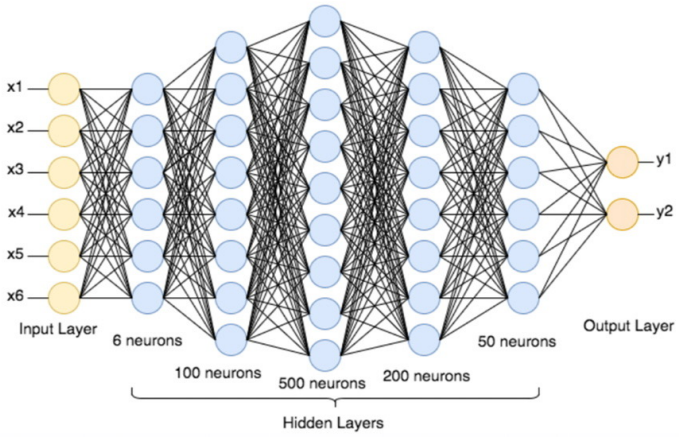
Cette situation apparaît dans certains cas de figure dès lors que l'on entraîne trop notre algorithme.





Un bon modèle est un modèle qui est à la fois adapté au jeu de données et aux objectifs fixés mais également équilibré au travers d'un entrainement adéquat.

Les défis du Machine Learning : Les problématiques des modèles



Le modèle des réseaux de neurones attire par ses nombreux avantages :

- Des résultats de grande qualité,
- Des meilleures performances que d'autres algorithmes,
- Des possibilités d'utilisation multiples : avec des données incomplètes ou bruitées.

Mais ils fonctionnent comme une « boîte noire » car ses résultats sont parfois difficilement explicables.

Les banques, les assureurs, l'entreprise de gestion d'actifs sont tous soumis à des règlementations fortes de la part d'autorités externes. Alors, en utilisant des modèles dits « boites noires », elles pourront se retrouver en situation d'inconfort lorsqu'il lui sera demandé d'en expliquer le fonctionnement et de justifier certaines prises de décision.

Laurence Devillers, chercheuse au CNRS et auteur de « Des robots et des hommes », « s'il est important de rendre ces systèmes plus robustes, il faut aussi expliquer comment ils fonctionnent, et garder à l'esprit qu'ils apprennent sans comprendre, donc qu'ils décident sans comprendre. ».

Les défis du Machine Learning : La qualité des données

Des données de mauvaise qualité auront plusieurs impacts avec des conséquences lourdes pour l'établissement qui les exploite.

Le principal risque est la génération de faux résultats tels que :

- Des mauvaises segmentations ;
- Des estimations erronées ;
- Une tarification inadaptée ;
- Des engagements inappropriés.

Ce qui entraine des décisions pouvant aller à l'encontre de l'entreprise, générateurs de pertes et d'insatisfaction client.

Les défis du Machine Learning : La qualité des données

Comment agir pour éviter ce risque ?

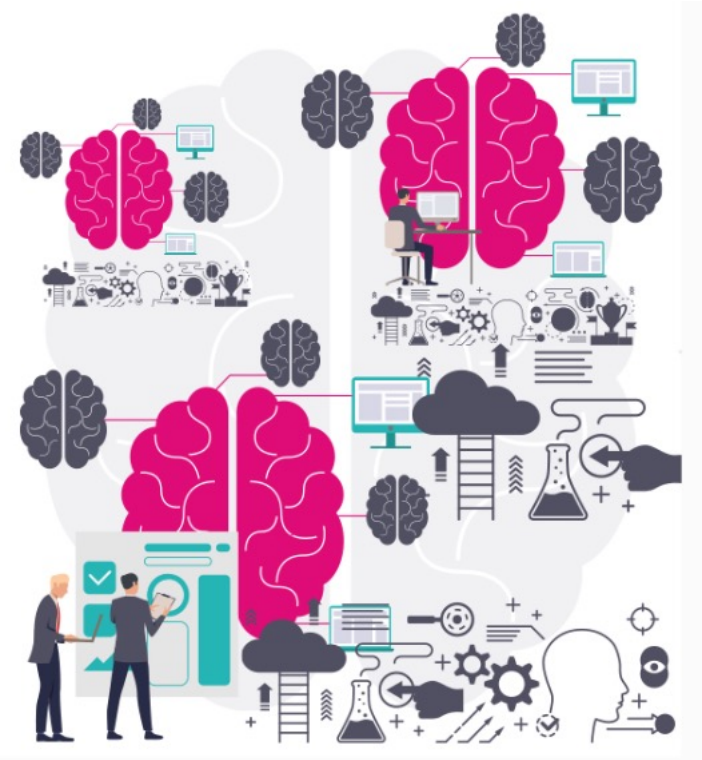
1. Intégrer la qualité des données à la politique interne des risques opérationnels des entreprises : c'est l'une des composantes importantes du patrimoine de l'entreprise et constitue un avantage concurrentiel.
2. Mettre en œuvre une gouvernance SI/Qualité des données avec l'application de plusieurs règles : La qualité de données est un enjeu qu'il ne faut pas ignorer pour le présent et l'avenir des entreprises.

Des technologies de pointes telles que le Machine Learning ne peuvent être efficientes qu'à ces conditions.

Règles prônées par la circulaire BCBS 23911 du comité de Bâle.

Fiabilité : garantir une image fidèle	Lisibilité : disposer d'une information homogène partagée	Pérennité : processus d'amélioration continue de la qualité	Contrôle : mise en place de processus de contrôle renforcés

Les défis du Machine Learning : La mise en production



Passer du modèle PoC à la production implique un effort massif et continu :

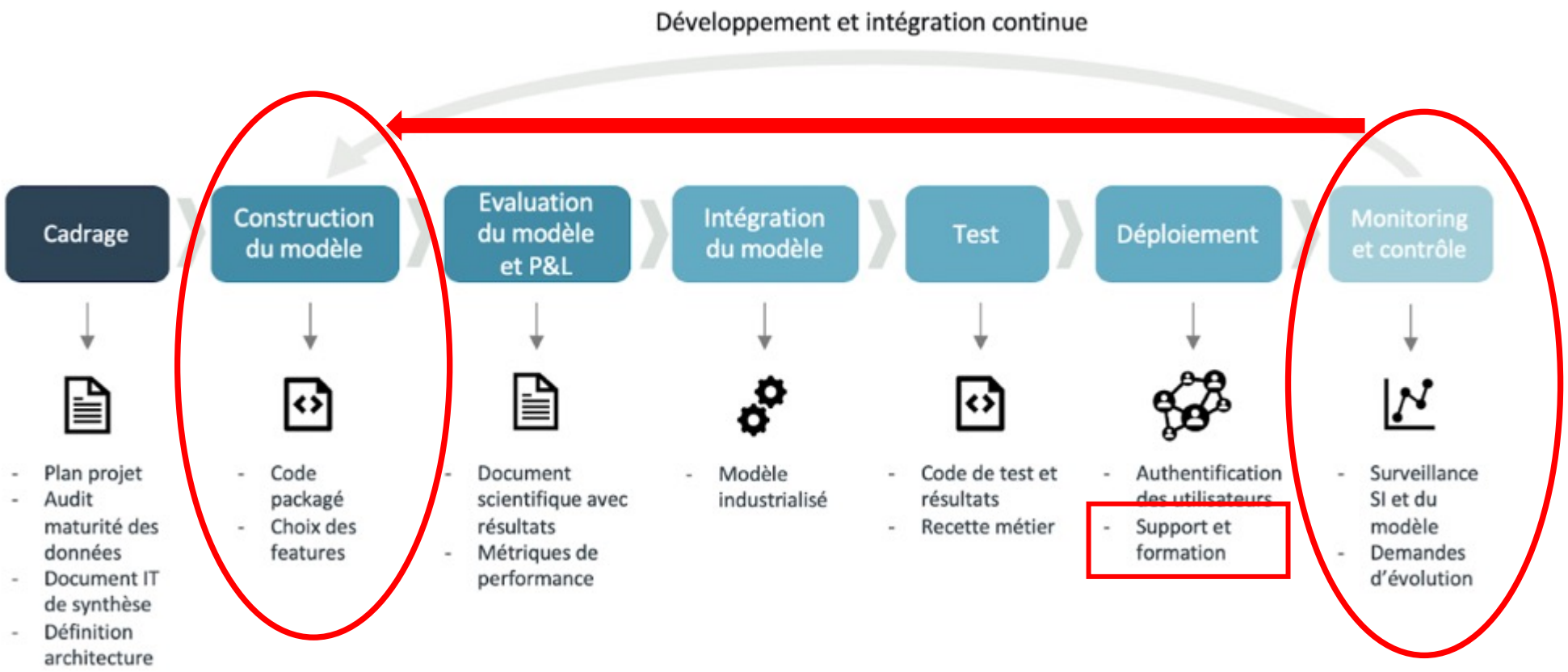
Le ML n’est pas un acte unique, le modèle doit être alimenté en permanence puisqu’il évolue et progresse de lui-même. Il faut lui consacrer beaucoup de temps et d’énergie pour un résultat qui sera toujours insatisfaisant, car c’est un modèle temps réel continu donc sans fin.

« Ce n'est pas comme si vous alliez mettre en place une solution et l’oublier pendant 10 ans. Le ML a besoin d’attention et d’être nourrie en continue ».

La peur de l’inconnu : Beaucoup de ceux qui ont une formation de comptable et de la planification et l'analyse financière, ne comprennent pas forcément comment fonctionne le ML. Pour eux, c'est de la magie. Et parce que c'est [trop] magique, cela devient un frein.

Expliquer et former permet, comme dans beaucoup de domaines, de dissiper le mystère et neutraliser la peur.

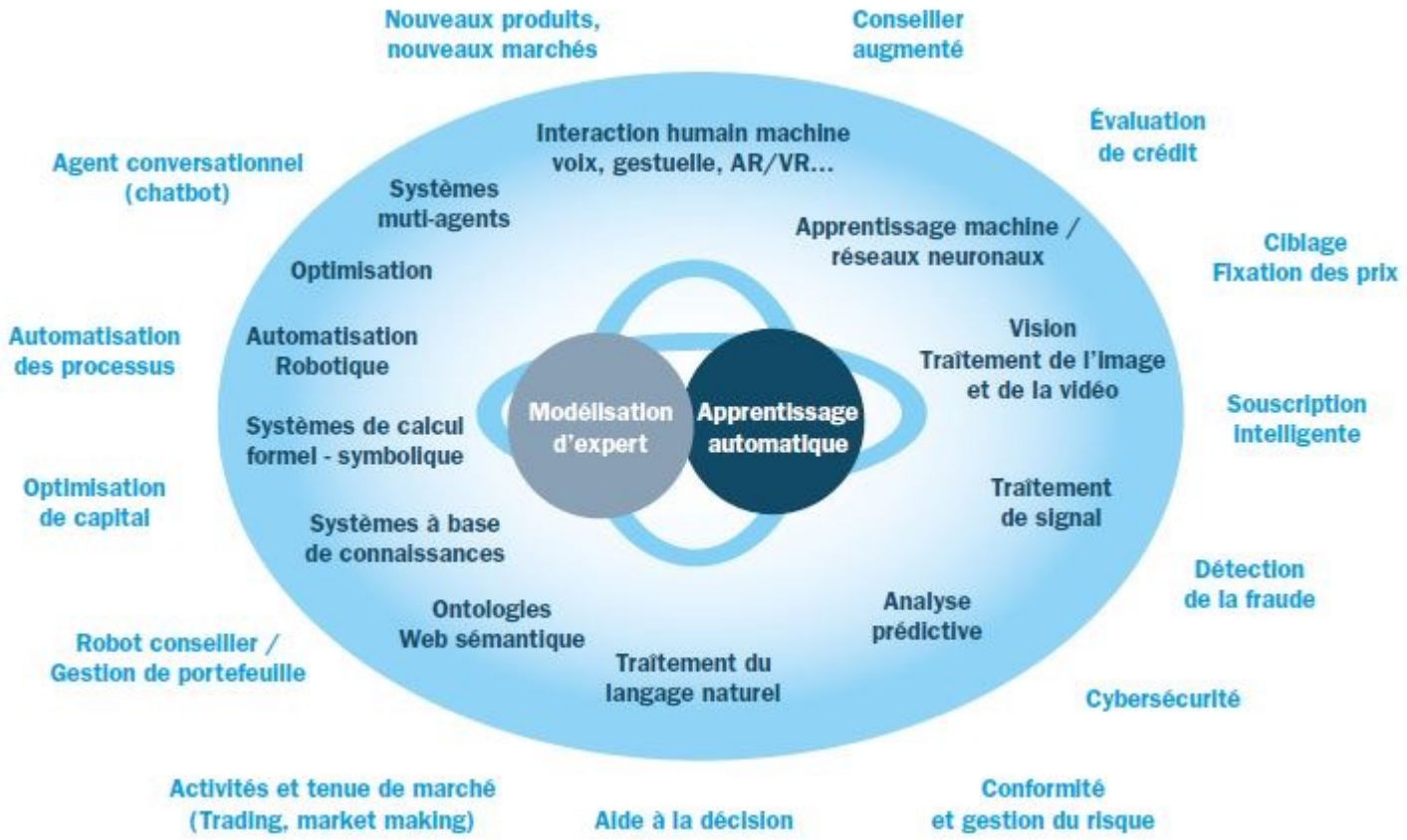
Les défis du Machine Learning : La mise en production



Plan du cours

- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

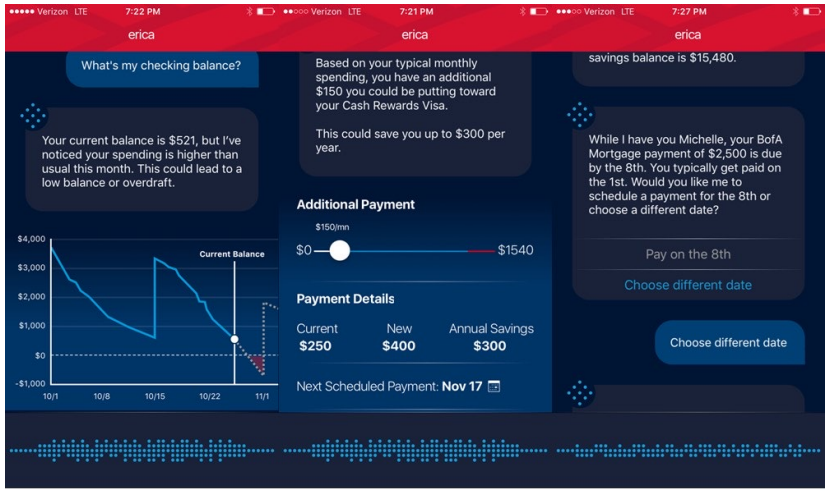
Machine Learning en Finance



Le ML au service des activités de banque de détail

- **Les chatbots** (envoi et réception de messages écrits), voicebots (envoi et réception de messages vocaux);
- **L'analyseur des émails** dont l'objectif est de faciliter la compréhension des offres de produits et de services financiers;
- **L'agent virtuel** sert à répondre aux questions des clients ou de les mettre en relation avec un conseiller;

L'agent virtuel d'Orange Bank, Djingo : traiterait en moyenne 17 000 conversations par semaine, avec un taux de compréhension des questions posées de 85 %



Source: Bank of America

Erica, le chatbot de Bank of America, est en mesure de répondre aux questions les plus souvent posées par les clients, qu'il s'agisse de trouver trace des transactions passées (« Ai-je récemment reçu des remboursements ? »), d'effectuer des demandes de paiements (virements, prélèvements, etc.) ou encore de consulter le solde de leurs comptes.

Le ML au service des activités de banque de détail

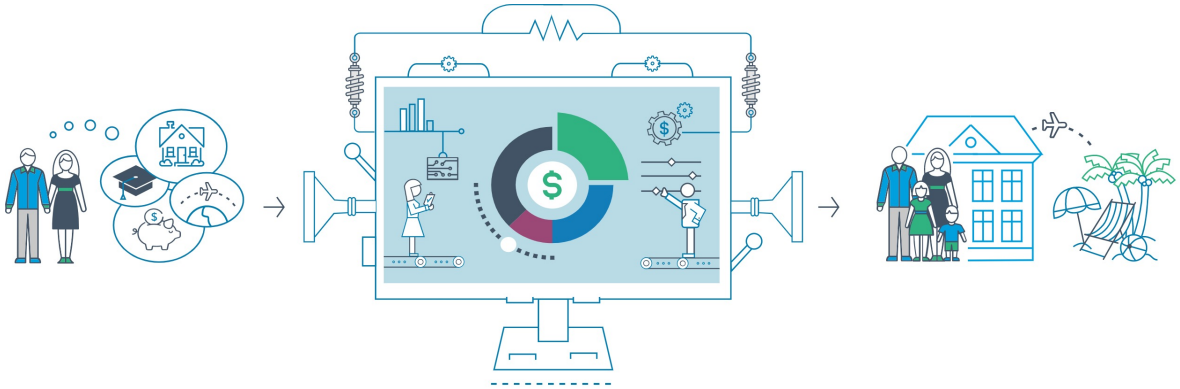
- **L'assistant virtuel:** c'est comme un coach financier virtuel, anticipant les besoins des clients en leur proposant des services de manière proactive :

L'assistant virtuel de Crédit Mutuel-CIC répond aux questions des conseillers pour les guider dans la relation client. Il est aussi un analyseur de courrier électronique pour :

- détecter l'urgence des demandes reçues,
- identifier automatiquement les plus fréquentes d'entre elles (déclaration de fraude, prise de rendez-vous, etc.),
- préremplir la réponse au client si la question posée par ce dernier est suffisamment explicite.

Le ML au service des activités de banque de détail

- **Conseil « automatisé » (ou robo-advice)** : fournissant des services allant des recommandations financières automatisées au courtage en ligne en passant par la gestion de portefeuille (mandat libre ou conseillé) selon une intervention humaine très limitée voire inexistante.



- **Noter (scorer) les clients** à partir d'un ensemble de données (historique des paiements, activité sur les médias sociaux, messages textuels, mails, etc.) dont se déduisent des comportements et des habitudes de consommation, éclairant les chances de remboursement d'un crédit et partant, le risque d'insolvabilité de l'emprunteur; les délais de traitements des demandes de crédit (en particulier de consommation) des clients en sont améliorés, à la grande satisfaction de ceux-ci.

Le ML au service des activités de banque de détail - Avantages

Toutes ces solutions permettent aux banques de :

- Remplacer le travail manuel,
- Automatiser les tâches routinières pour décharger les employés de tâches lourdes et longues à réaliser et les laisser se concentrer sur des missions à plus fortes valeurs ajoutées
- Augmenter la productivité.

ML permet donc aux Banques de détail de :

- Optimiser leurs coûts,
- Améliorer l'expérience de leurs clients
- Développer leurs services.

Plan du cours

- Introduction à l'intelligence artificielle (AI)
- Définition de l'apprentissage automatique ou Machine Learning (ML)
- Machine Learning : terminologies de base
- Les Types de Machine Learning
- Machine Learning : Le cycle de vie
- Les défis du Machine Learning
- Le Machine Learning au service des activités de la banque de détail
- Le Machine Learning au service des activités des banques d'investissement

Le ML au service des activités des banques d'investissement

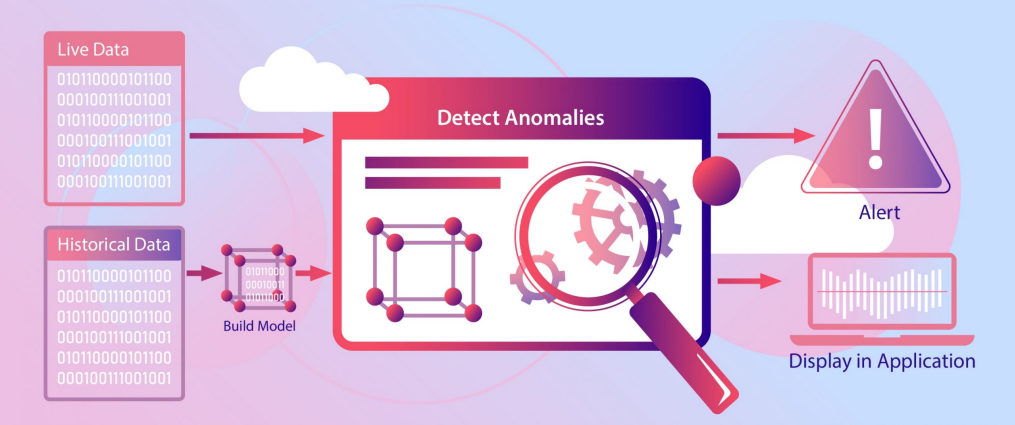
- **Le trading sur les marchés financiers** : ML permet d'identifier les tendances de marché et même de « prédire » ou anticiper le cours des actions.

Si on ajoute aux données du marché des données provenant de plusieurs autres domaines comme les domaines de l'économie, la politique, la santé, la technologie, la sécurité …, on peut obtenir un modèle encore plus puissant…

BlackRock est aujourd'hui la société de gestion la plus importante au monde, avec plus de 5000 Millard € a investi en 2017 30 Milliard € dans le Trading Algorithmique.

Le ML au service des activités des banques d'investissement

- **Détection de fraude** : Le ML permet aux banques ou les « référentiels centraux » d'identifier les incohérences et les anomalies (fraude extérieure, délits d'initiés, erreurs de type « fat finger », etc.) des données qu'ils collectent et de les corriger le cas échéant.



ML est un outil puissant et efficace dans la lutte contre la fraude, comme il permet de :

- Faciliter la prise de décision en temps réel,
- Améliorer la précision,
- Réagir rapidement au changement,
- Réduire les coûts.

Des algorithmes ML développés à la banque Monzo (Royaume-Unis) ont pu détecter des cas de fraudes sur cartes bancaires prépayées et faire passer de 1 % à 0,1 % le nombre de fraudes.

Le ML au service des activités des banques d'investissement

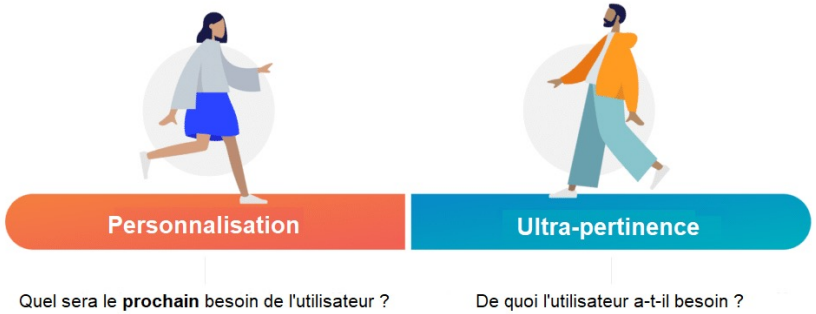
- **Gestion du risque**
  - Les systèmes de réglementation tels que Bâle III exigent que des entreprises gèrent leur risque de liquidité du marché par le biais de tests de résistance.
  - Les entreprises financières gèrent également les risques liés à leurs clients en analysant des portefeuilles complets de clients.
  - Les risques liés au trading algorithmique sont gérés au moyen de stratégies de Backtest sur des données historiques.

L'analyse de données volumineuses avec le ML peut prendre en charge l'alerte en temps réel si un seuil de risque est dépassé.

Le ML au service des activités des banques d'investissement

- **Personnalisation profonde** : Pour être compétitif sur le marché actuel, les entreprises cherchent à susciter l'engagement de leurs clients grâce à des relations personnalisées de haute qualité.

L'idée est d'analyser l'expérience client numérique et de la modifier en tenant compte des intérêts et des préférences du client.



La combinaison d'outils d'analyse prédictive et d'options avancées de transmission numérique peut aider à cette tâche complexe en :

- guidant le client vers la meilleure solution financière,
- suggérant de personnaliser les offres en fonction :
    1. des habitudes de consommation,
    2. des tendances socio-démographiques,
    3. de la localisation,
    4. et d'autres préférences.

Le ML au service des activités des banques d'investissement

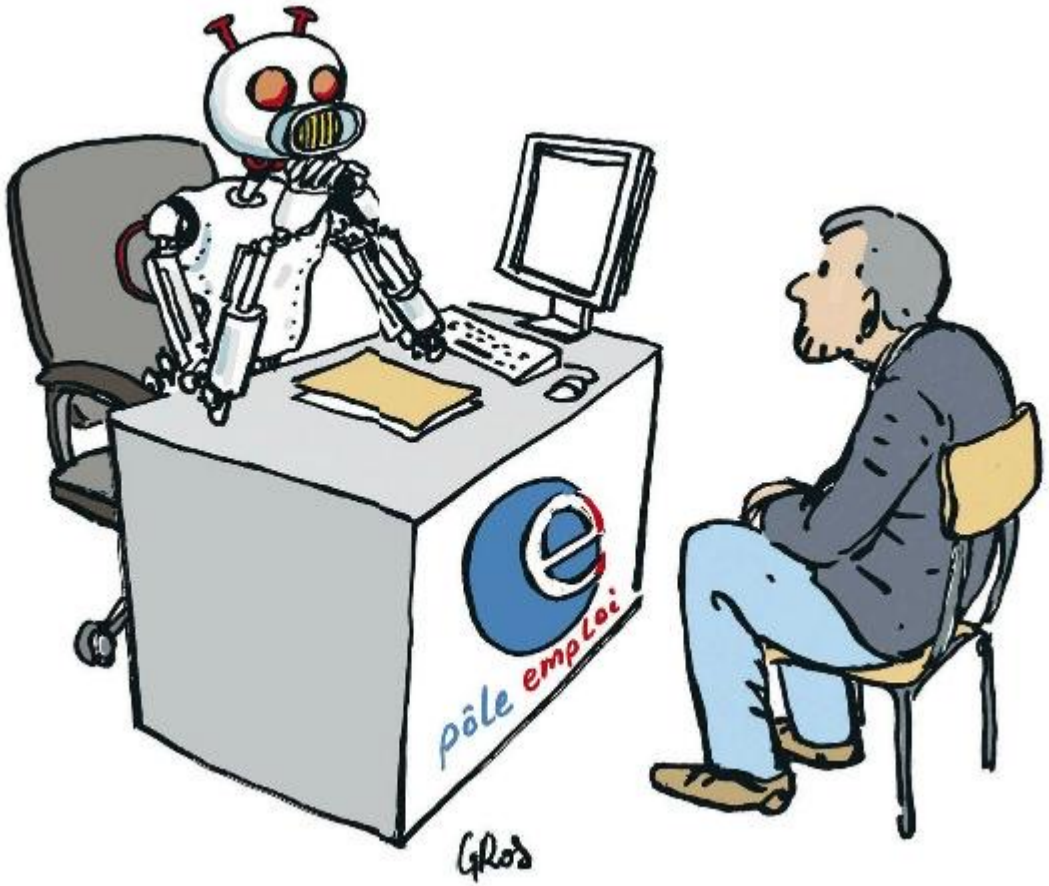
- **Regulations :** Les entreprises financières consacrent des milliards de dollars à la conformité réglementaire (en moyenne 1,34 million de dollars en technologies liées à la conformité en 2017, contre une moyenne de 92 000 dollars en 2011.) Mais de nouvelles règles et réglementations apparaissent toujours.

Les algorithmes de ML peuvent lire et apprendre à partir d'une pile de documents réglementaires, ils peuvent donc détecter les corrélations entre les directives.

Grâce au ML, les établissements bancaires peuvent surveiller les données de transaction pour identifier automatiquement les anomalies.

Le ML peut donc garantir que les transactions des clients sont conformes aux exigences réglementaires.

… 2025 ou 2030 ???!!!



Annexes

ML : Fonction coût

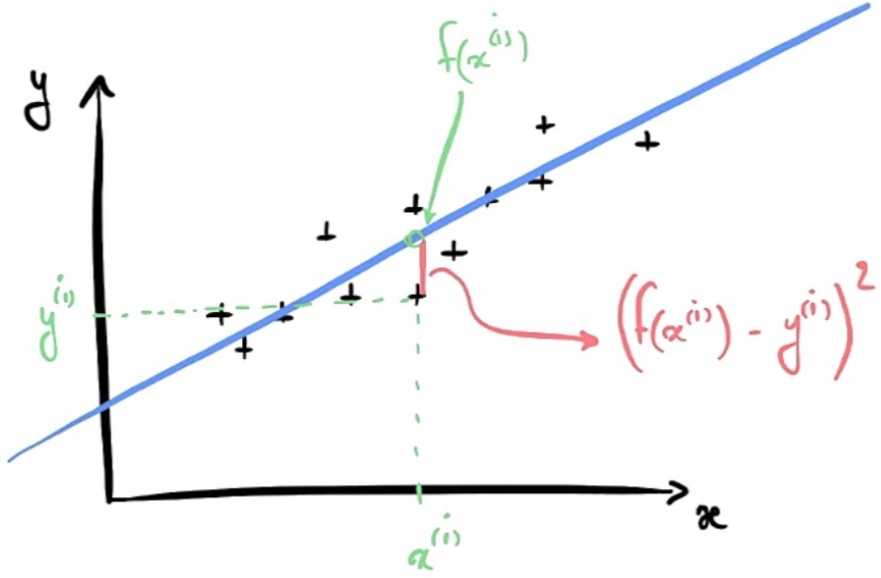
On utilise la norme euclidienne pour mesurer les erreurs entre $\boldsymbol{f}(\boldsymbol{x})$ et $(\boldsymbol{y})$. L'erreur i entre $\boldsymbol{y}$ $(\boldsymbol{i})$ et la prédiction faites en utilisant la surface $\boldsymbol{x}$ $(\boldsymbol{i})$:

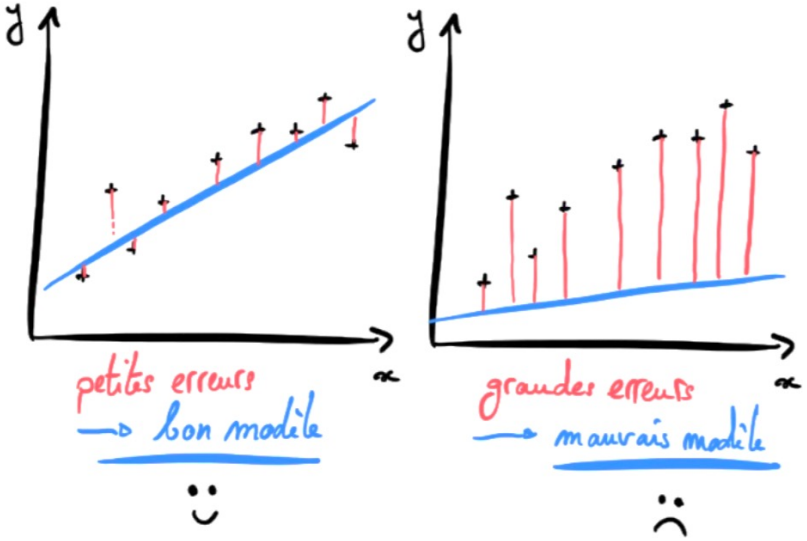
$$erreur(i) = (\ f(x\ (i)\ ) - y\ (i)\ )\ 2$$

La **Fonction Coût** est la moyenne de toutes les erreurs (c'est l'erreur quadratique moyenne (Mean Squared Error)) :

$$J(a,b) = \frac{1}{2m}\sum_{i=1}^{m}\left(\ f\left(x^{(i)}\right) - y^{(i)}\ \right)^2$$

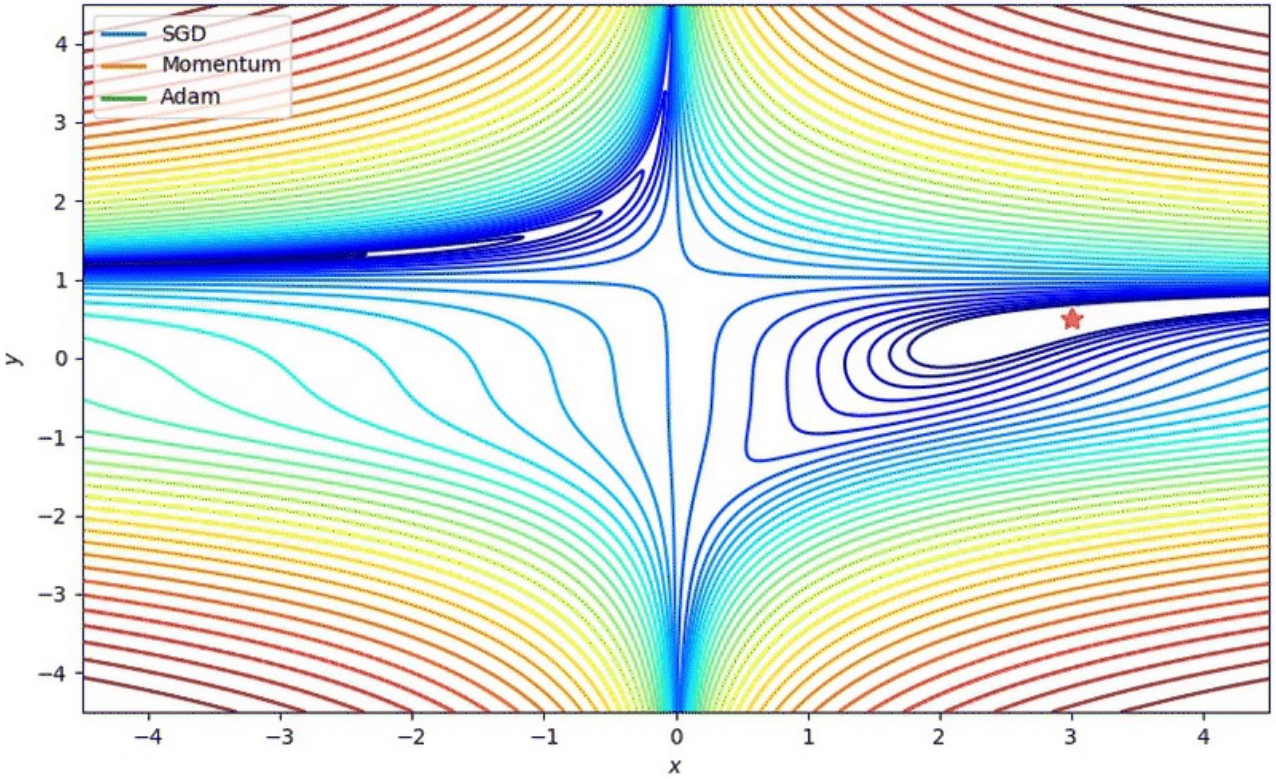
On multiplie l'entrainement dans le but de trouver les **paramètres** de l'algorithme qui **minimisent** la **Fonction Coût**, c'est-à-dire les paramètres qui nous donnent le meilleur modèle.





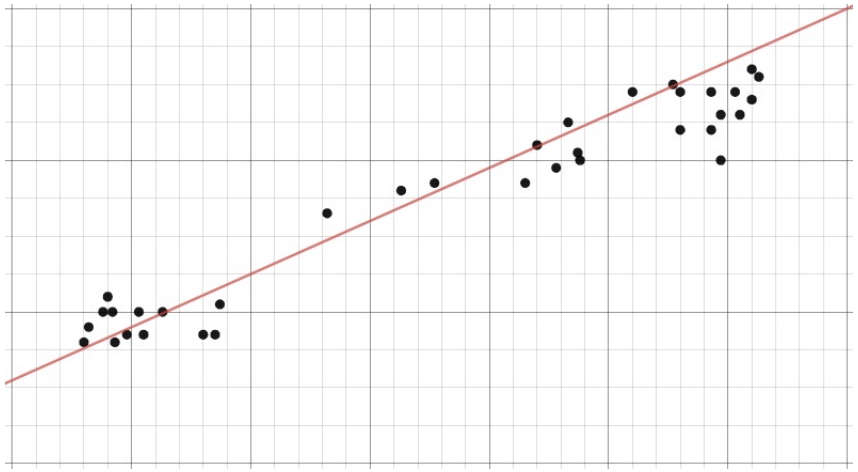
ML : Gradient descent

Les **Gradient descent** sont des algorithmes d'optimisation utilisés pour déterminer avec le moins d'itérations possibles les paramètres optimales d'un modèle de Machine Learning.



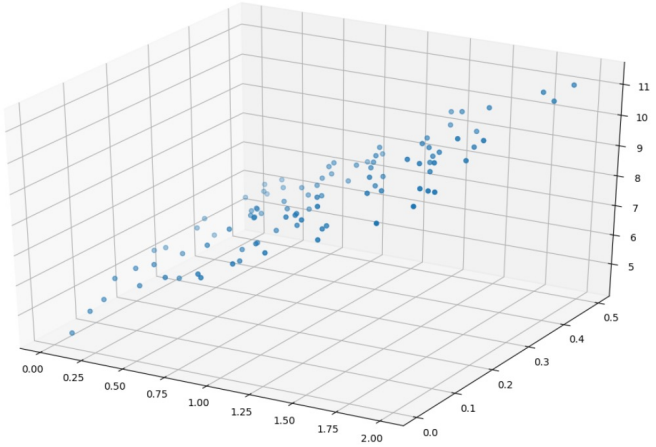
ML Supervisé : Quelques Algorithmes - Régression

Régression linéaire : est de trouver une ligne qui s'adapte (ou correspond) le mieux aux données.

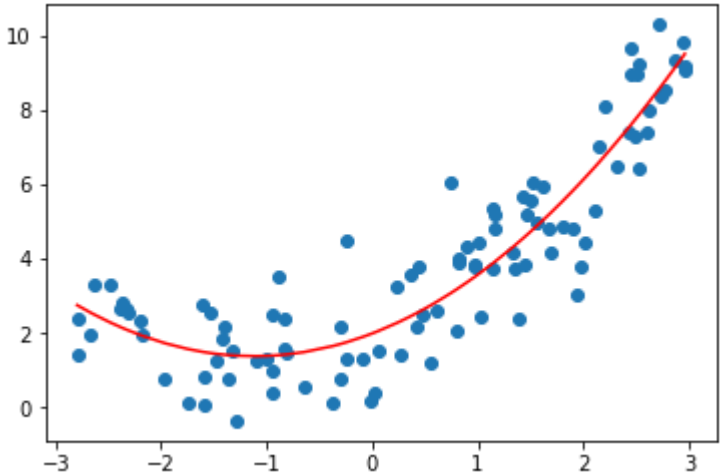


Les extensions de la régression linéaire comprennent :

- la **régression linéaire multiple** (par exemple, trouver un plan qui s'ajuste le mieux)



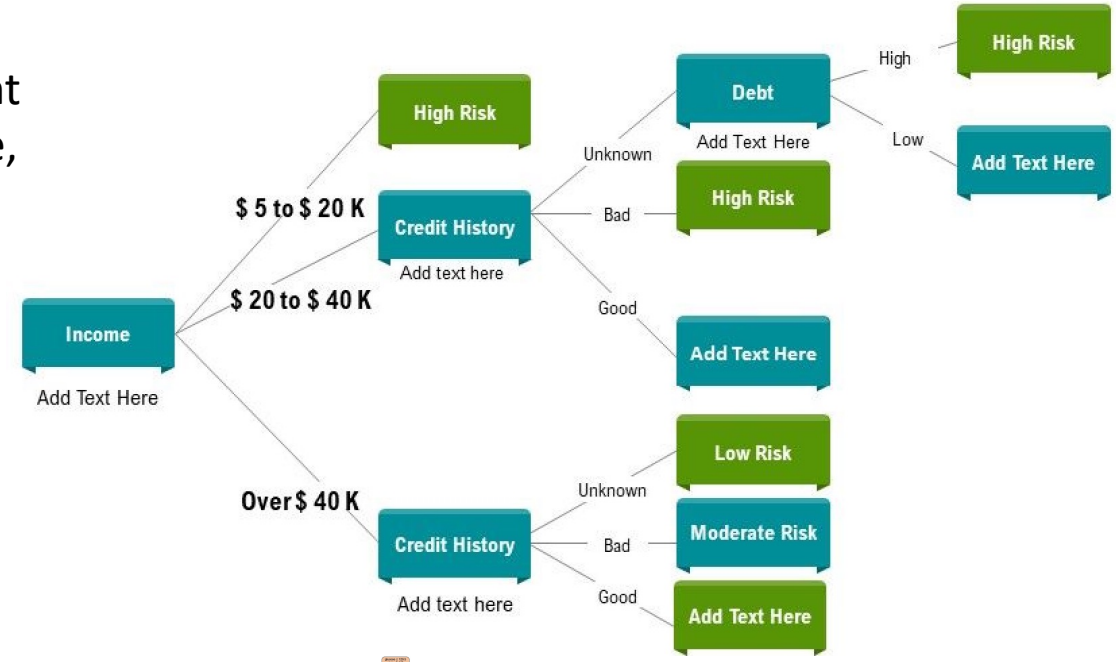
- la **régression polynomiale** (par exemple, trouver une courbe qui s'ajuste le mieux).

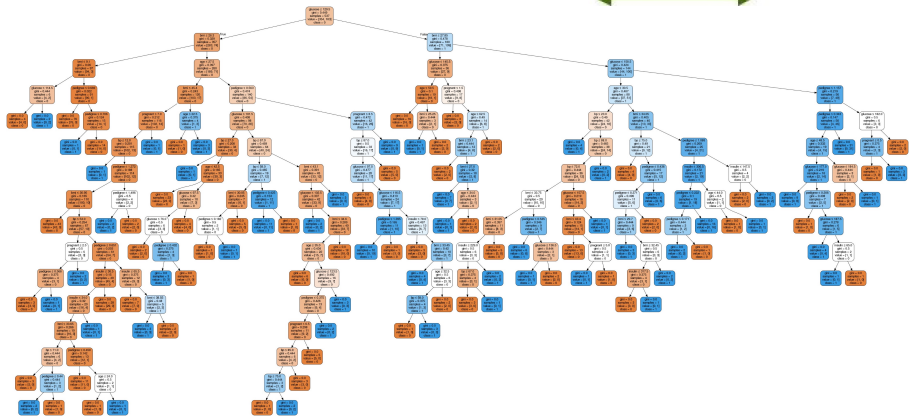


ML Supervisé : Quelques Algorithmes - Régression

Decision Tree (Arbre de décision) : Les arbres de décision sont un modèle populaire, utilisé dans la recherche opérationnelle, la planification stratégique et le **Machine Learning**.

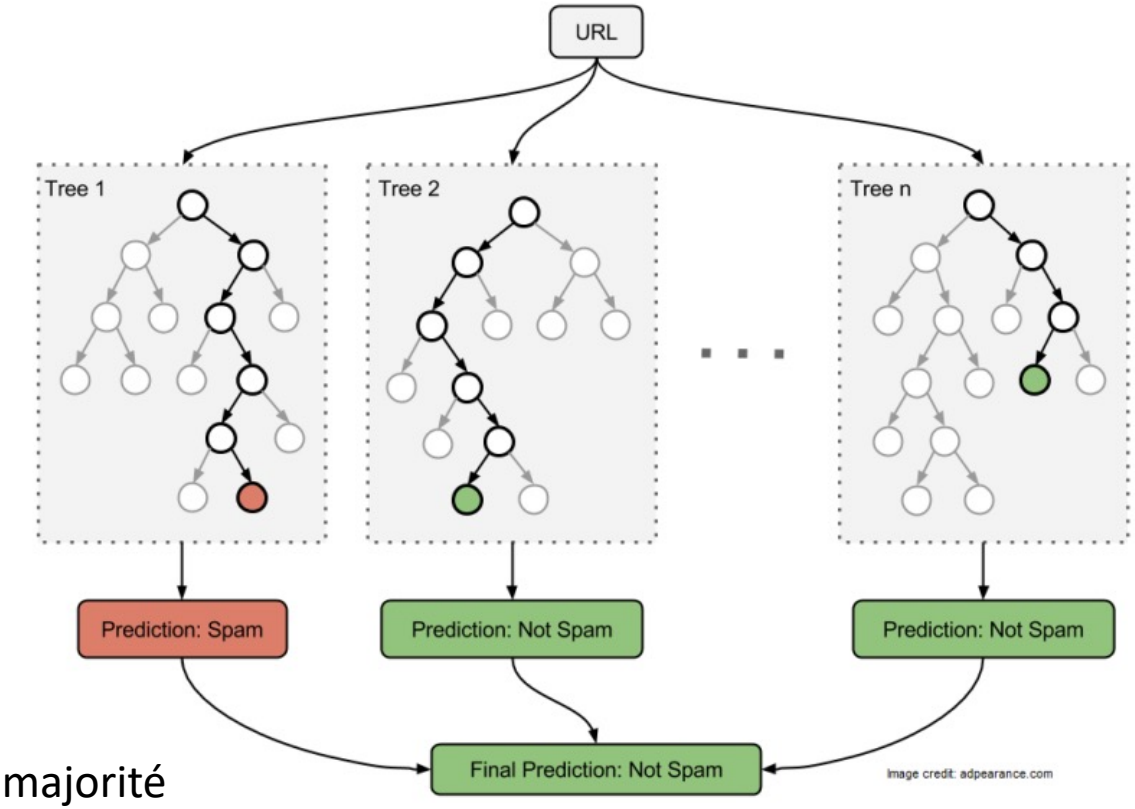
- Chaque rectangle est appelé un **nœud**.
- Plus il y a de nœuds, plus l'arbre décisionnel sera précis (en général).
- Les derniers nœuds de l'arbre décisionnel, où une décision est prise, sont appelés les « **feuilles** » de l'arbre.
- Les arbres décisionnels sont intuitifs et faciles à construire, mais ils font un peu défaut lorsqu'on parle de précision ou d'exactitude.





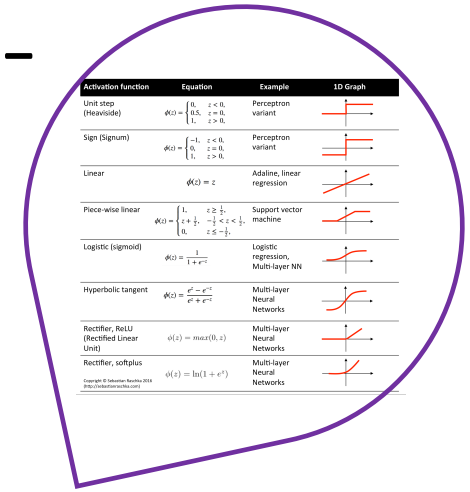
ML Supervisé : Quelques Algorithmes - Régression

Random Forest (Forêt d'arbres décisionnels)
C'est une technique d'**apprentissage ensembliste** qui s'appuie sur des arbres de décision.
Le modèle random forest implique la création d'arbres décisionnels (**decision tree**) multiples en utilisant des ensembles de données fractionnés à partir des données d'origine et en sélectionnant aléatoirement un sous-ensemble de variables à chaque étape de l'arbre décisionnel.
Le modèle sélectionne ensuite le mode de toutes les prédictions de chaque arbre décisionnel.



Quel est l'intérêt de cette méthode ?
En s'appuyant sur un modèle de prévalence de la majorité (c'est-à-dire sur lequel la majorité l'emporte), il réduit le risque d'erreur d'un arbre individuel.

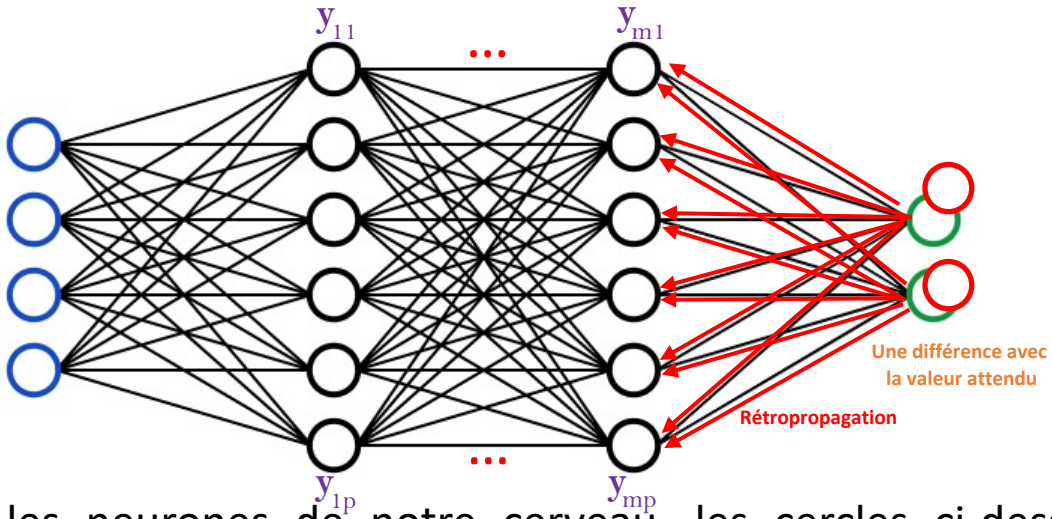
ML Supervisé : Quelques Algorithmes - Régression

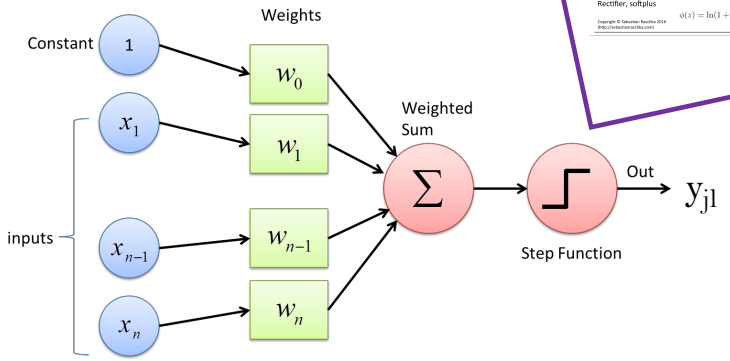


Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^{z})$	Multi-layer Neural Networks	

Copyright © Sebastian Raschka 2016 (http://sebastianraschka.com)

Réseau de neurones : c'est un modèle à plusieurs niveaux inspiré du cerveau humain.





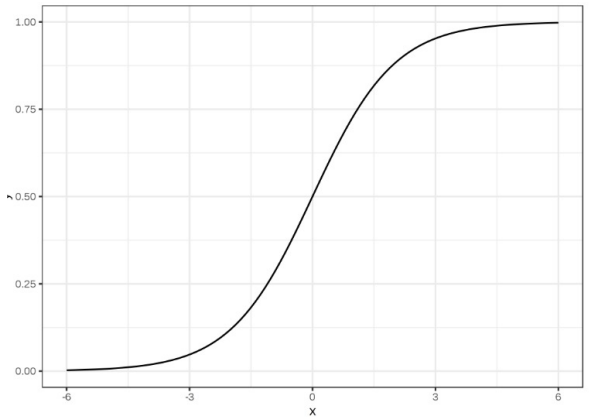
Comme les neurones de notre cerveau, les cercles ci-dessus représentent un nœud. Les cercles bleus représentent la couche d'entrée (**input layers**), les cercles noirs représentent les couches cachées (**hidden layers**) ou intermédiaires et les cercles verts représentent la couche de sortie (**output layers**).

Chaque nœud des couches cachées représente une fonction qui s'applique sur les données d'entrée menant finalement à une sortie des cercles verts.

Les réseaux de neurones sont en fait très complexes et très mathématiques…. Ils sont considéré comme des boites noires.

ML Supervisé : Quelques Algorithmes - Classification

Régression logistique : C'est semblable à la régression linéaire, mais elle est utilisée pour modéliser la probabilité d'un nombre fini de résultats, généralement deux.
Une équation logistique est créée de telle sorte que les valeurs des résultats ne peuvent être qu'entre 0 et 1.



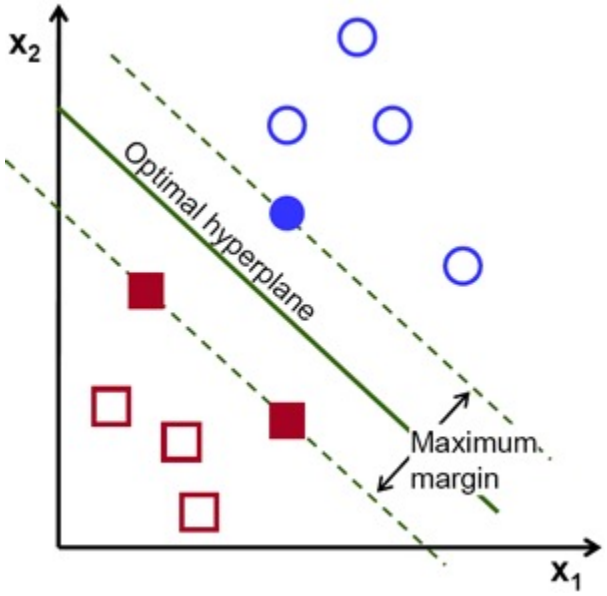
Naive Bayes est un autre classificateur populaire utilisé en Data Science.
L'idée qui se cache derrière est inspirée du théorème de Bayes :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Decision Tree, Random Forest, Neural Network
Ces modèles suivent la même logique que celle expliquée précédemment. La seule différence est que cette sortie est discrète plutôt que continue.

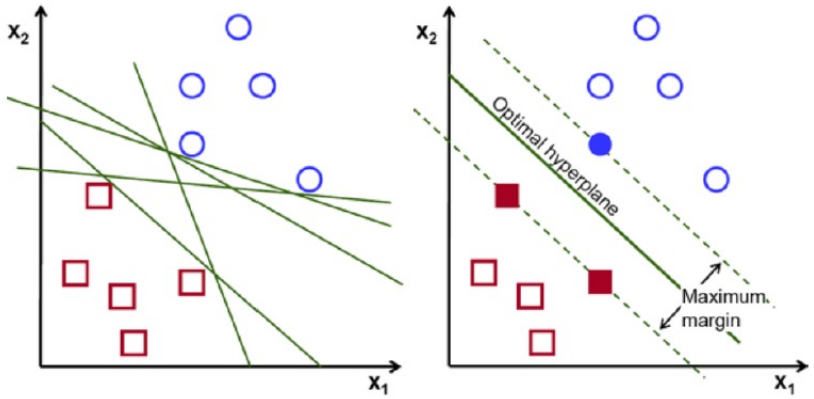
ML Supervisé : Quelques Algorithmes - Classification

Un Support Vector Machine **(SVM)** : Supposons qu'il existe deux classes de données. Une machine à vecteurs de support trouvera un **hyperplan** ou une frontière entre les deux classes de données qui maximisera la marge entre les deux classes. Il y a plusieurs plans qui peuvent séparer les deux classes, mais un seul plan peut maximiser la marge ou la distance entre les classes.



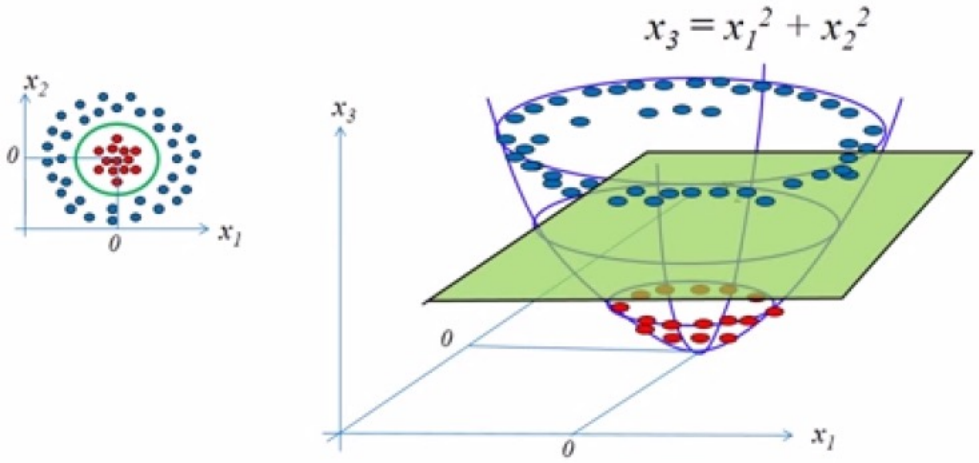
ML Supervisé : Quelques Algorithmes - Classification

Dans le cas où les classes sont **linéairement séparables**, l'objectif est de trouver l'hyperplan $y(x) = w_t x + b$ qui sépare au mieux les deux classes (**marge maximale**)
Le problème d'optimisation permettant de trouver (w, b) se résout à l'aide d'un algorithme d'optimisation numérique.



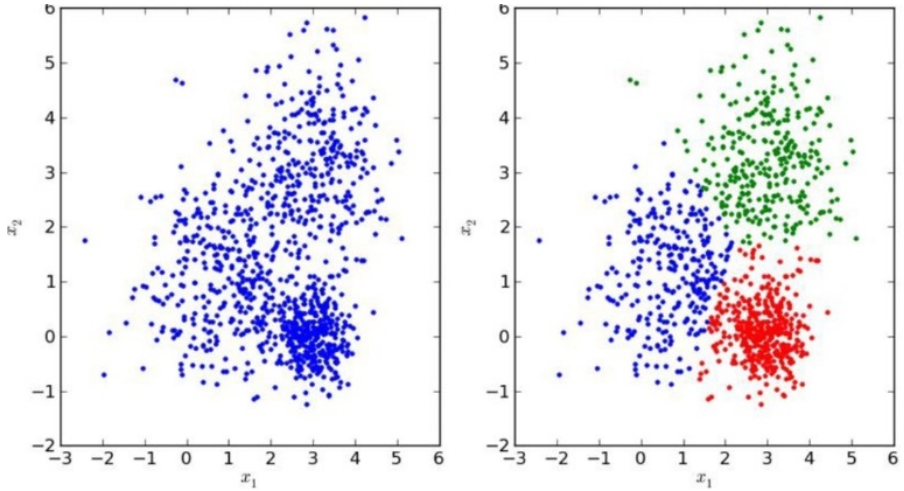
Bien souvent, les classes ne sont pas linéairement séparables, mais on peut changer de variables $x \Rightarrow \phi(x)$.

Le problème d'optimisation consiste alors à trouver les (w, b) de l'hyperplan $y(x) = w_t\, \phi(x) + b$
Mais la recherche du changement de variable peut-être coûteuse.



ML Non Supervisé : Algorithmes pour Clustering

Contrairement à **l'apprentissage supervisé**, l'apprentissage non supervisé est utilisé pour tirer des conclusions et trouver des tendances à partir de données d'entrée sans étiquettes (ou labels).



Les techniques courantes de clustering comprennent la technique des **k-means**, le clustering **hiérarchique**, le clustering par **décalage moyen** et le clustering **basé sur la densité**. Bien que chaque technique ait une méthode différente pour trouver les clusters, elles visent toutes le même objectif.

ML Non Supervisé : Clustering

k-means est un algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïd. Le **choix initial des centroïdes conditionne le résultat final**.

Admettant un nuage d'un ensemble de points, K-Means change les points de chaque cluster jusqu'à ce que la somme ne puisse plus diminuer. Le résultat est un ensemble de clusters compacts et clairement séparés, sous réserve de choisir la bonne valeur **K** du nombre de clusters.

Entrée :

- K le nombre de cluster à former
- Le Training Set (matrice de données)

DEBUT

Choisir aléatoirement K points (une ligne de la matrice de données). Ces points sont les centres des clusters (nommé centroïd).

REPETER

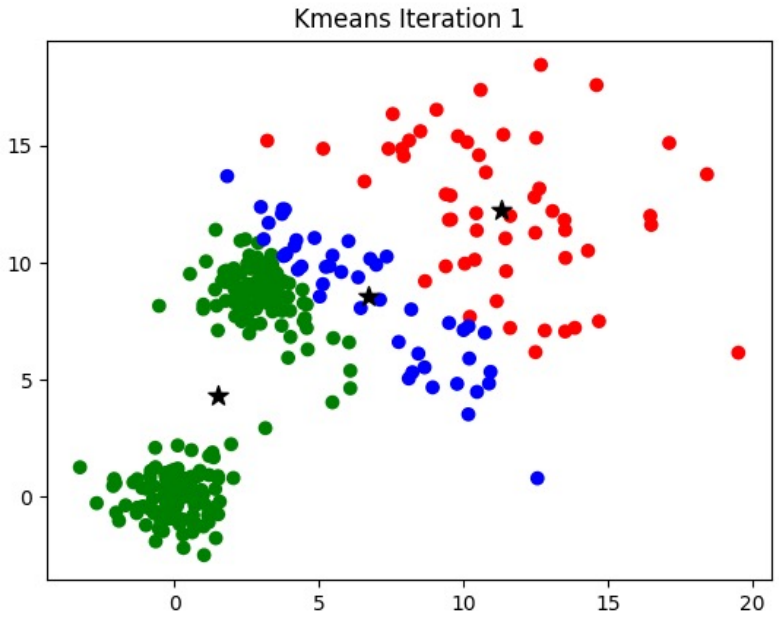
Affecter chaque point (élément de la matrice de donnée) au groupe dont il est le plus proche au son centre

Recalculer le centre de chaque cluster et modifier le centroïde

JUSQU'A CONVERGENCE

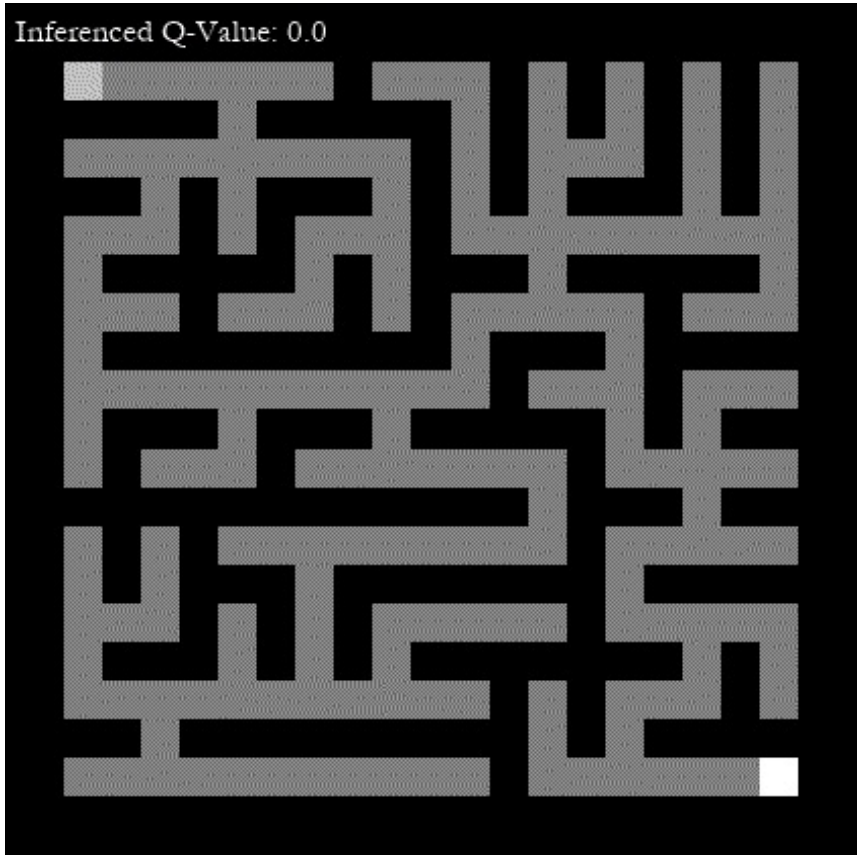
OU (stabilisation de l'**inertie totale** de la population)

FIN ALGORITHME



ML Renformcement : Algorithme Q-Learning

Les algorithmes de **Q-Learning** cherchent à trouver la meilleure méthode (une politique optimale) pour atteindre un objectif défini en cherchant à obtenir un maximum de récompenses.

Ils tentent le plus grand nombre d'actions possibles par état du système sans avoir de connaissance initiale de l'environnement. Un algorithme de ce type peut être construit pour obtenir rapidement des récompenses ou pour atteindre un objectif majeur.

Le Q-Learning est souvent associé à des modèles de Deep Learning dans le cadre de projets de recherche (ex : Google DeepMind).