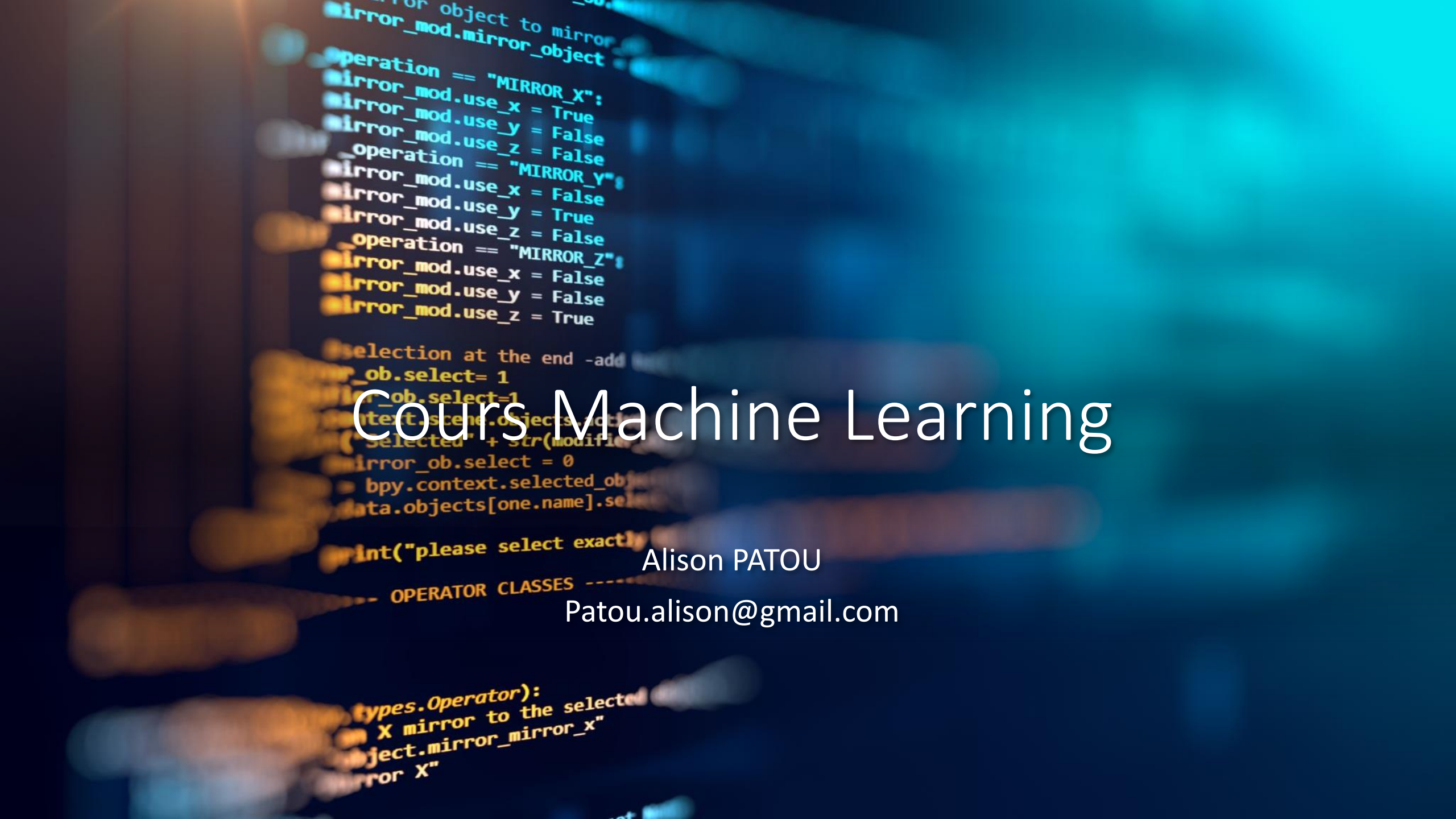


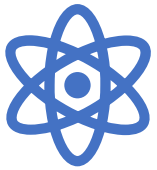
Cours Machine Learning

Alison PATOU

Patou.alison@gmail.com

Programme

Introduction



Jupyter Notebook

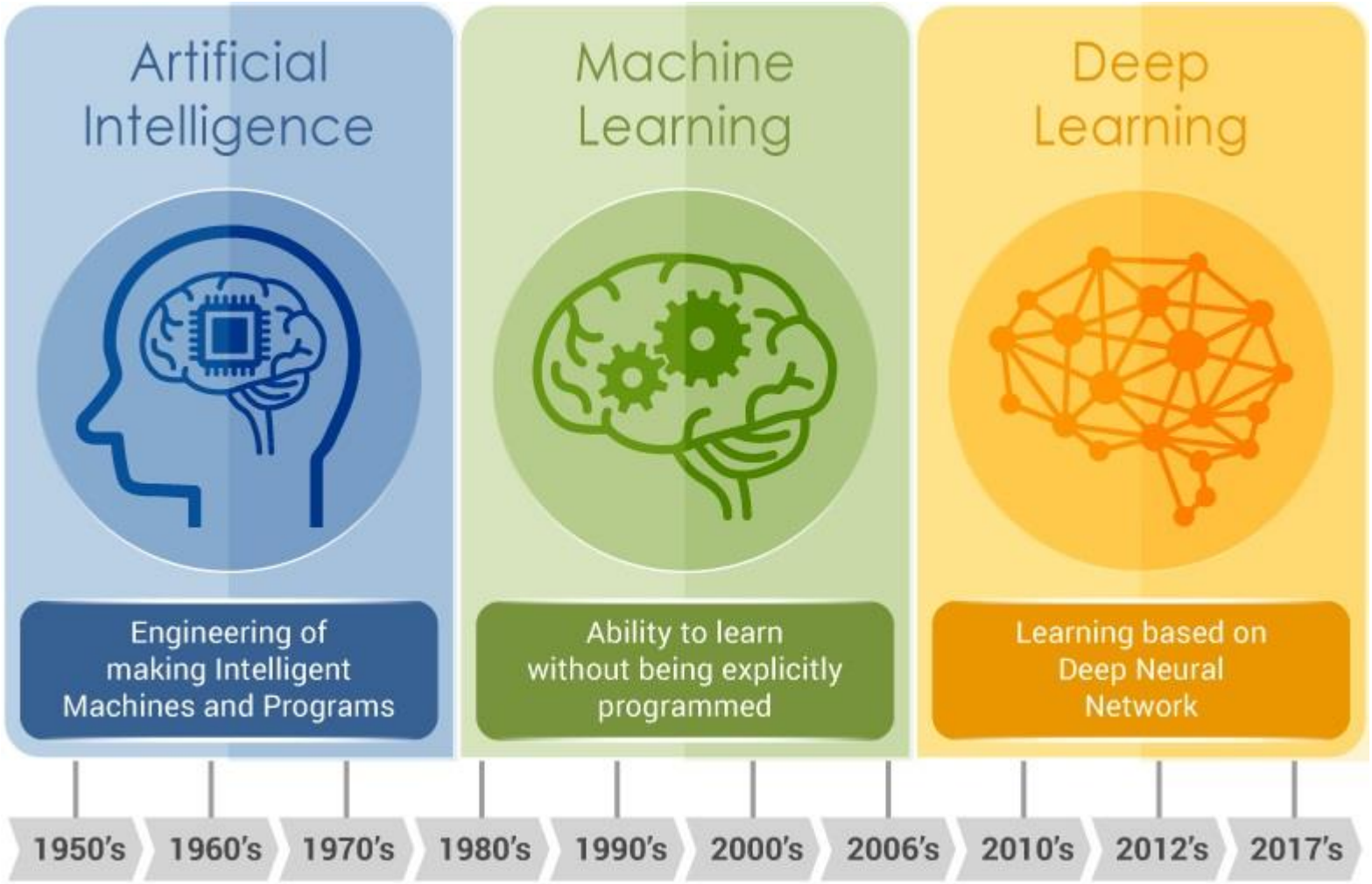
Langage Python

Machine Learning

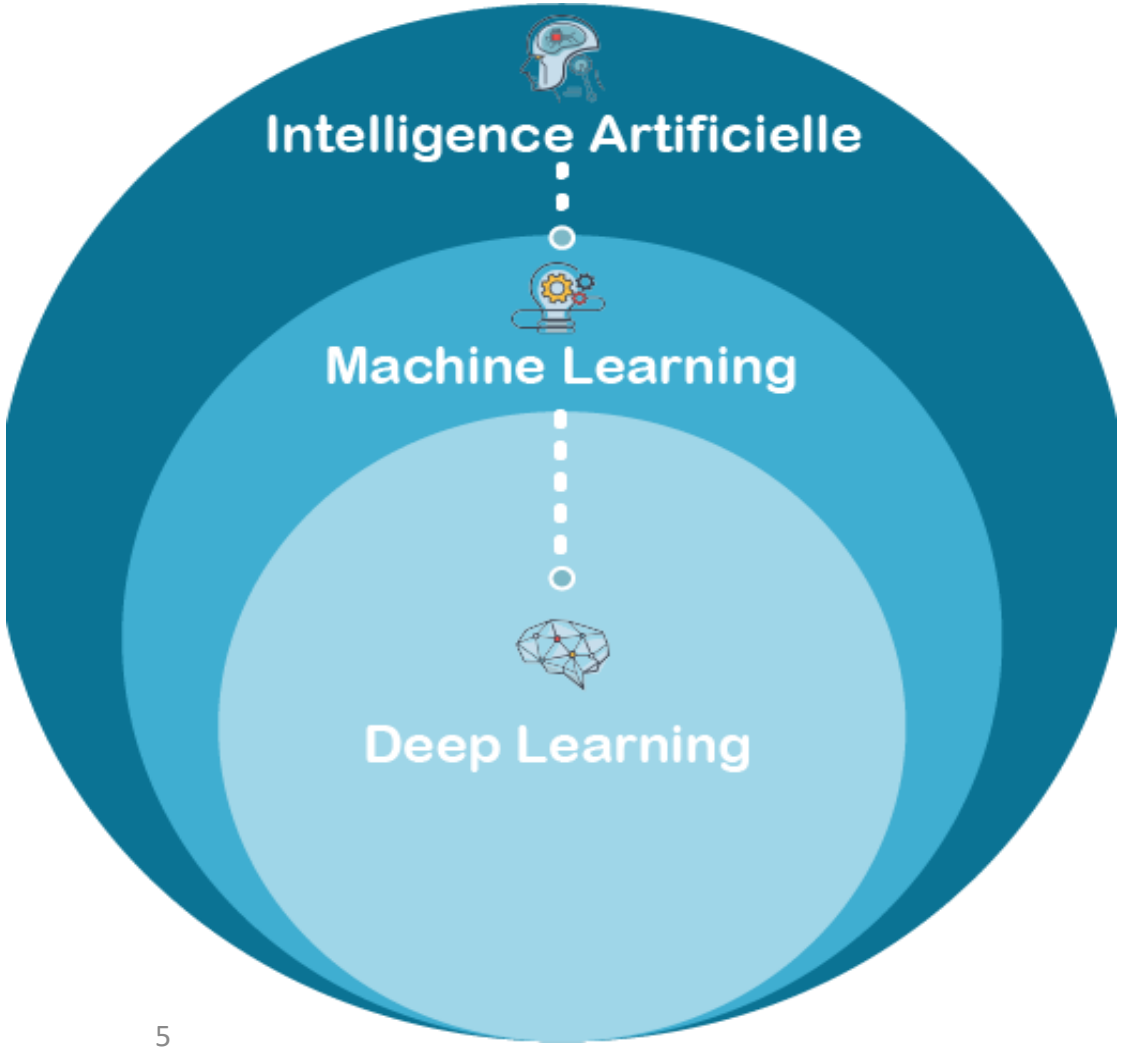
1

Introduction

Histoire du Machine Learning

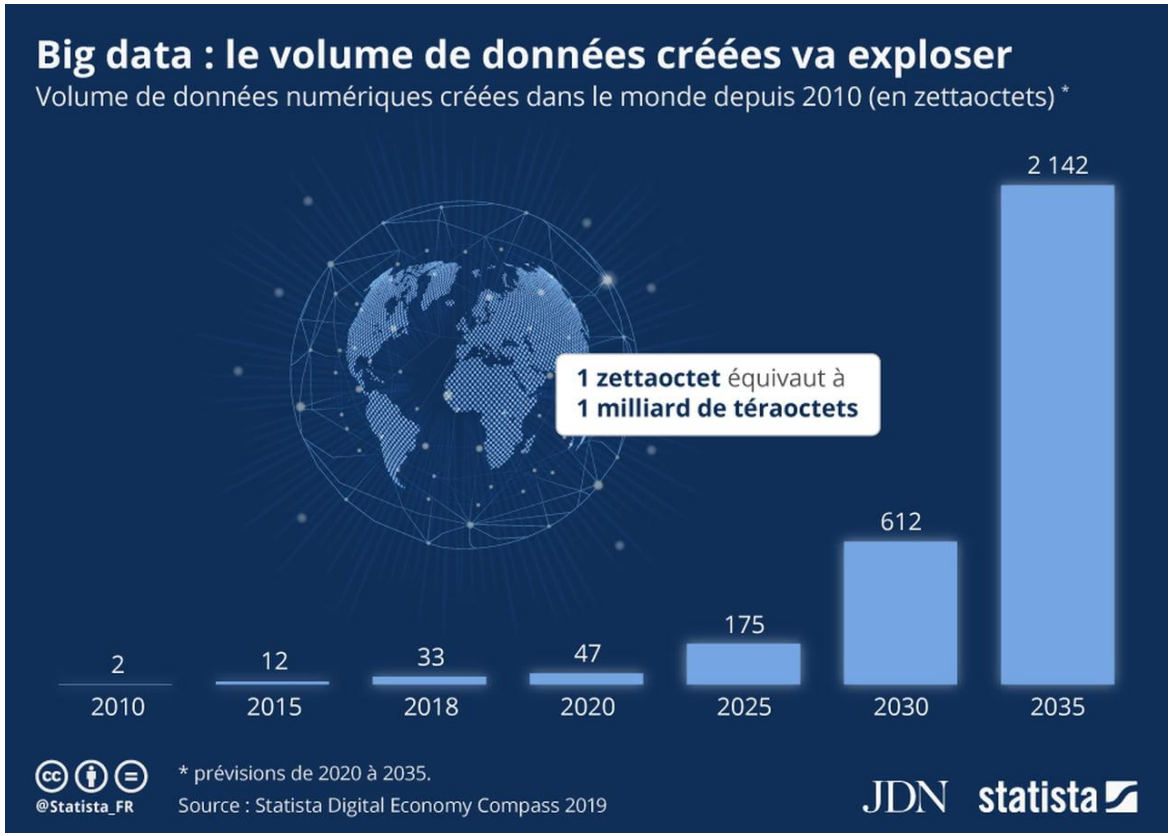


Machine Learning et Deep Learning : Contributeurs de l'Intelligence Artificielle

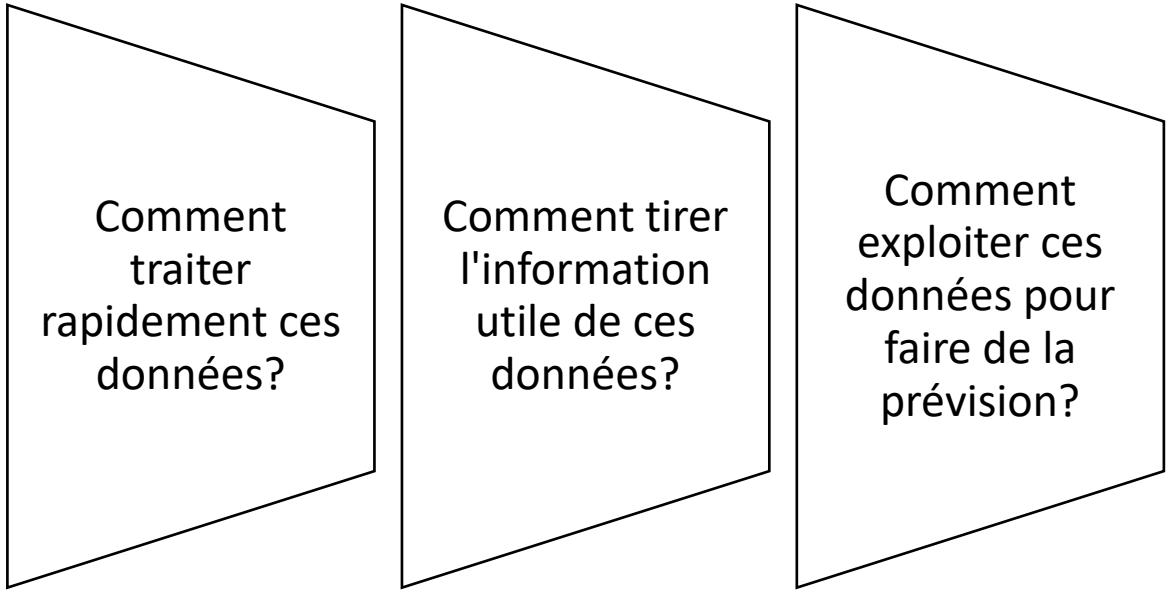


Le Big Data, c'est quoi?

- Données massives
- Développement d'Internet
- Production exponentielle de données
- Infrastructures à capacité élevée de stockage et de traitement de données



Le Big Data et le Machine Learning



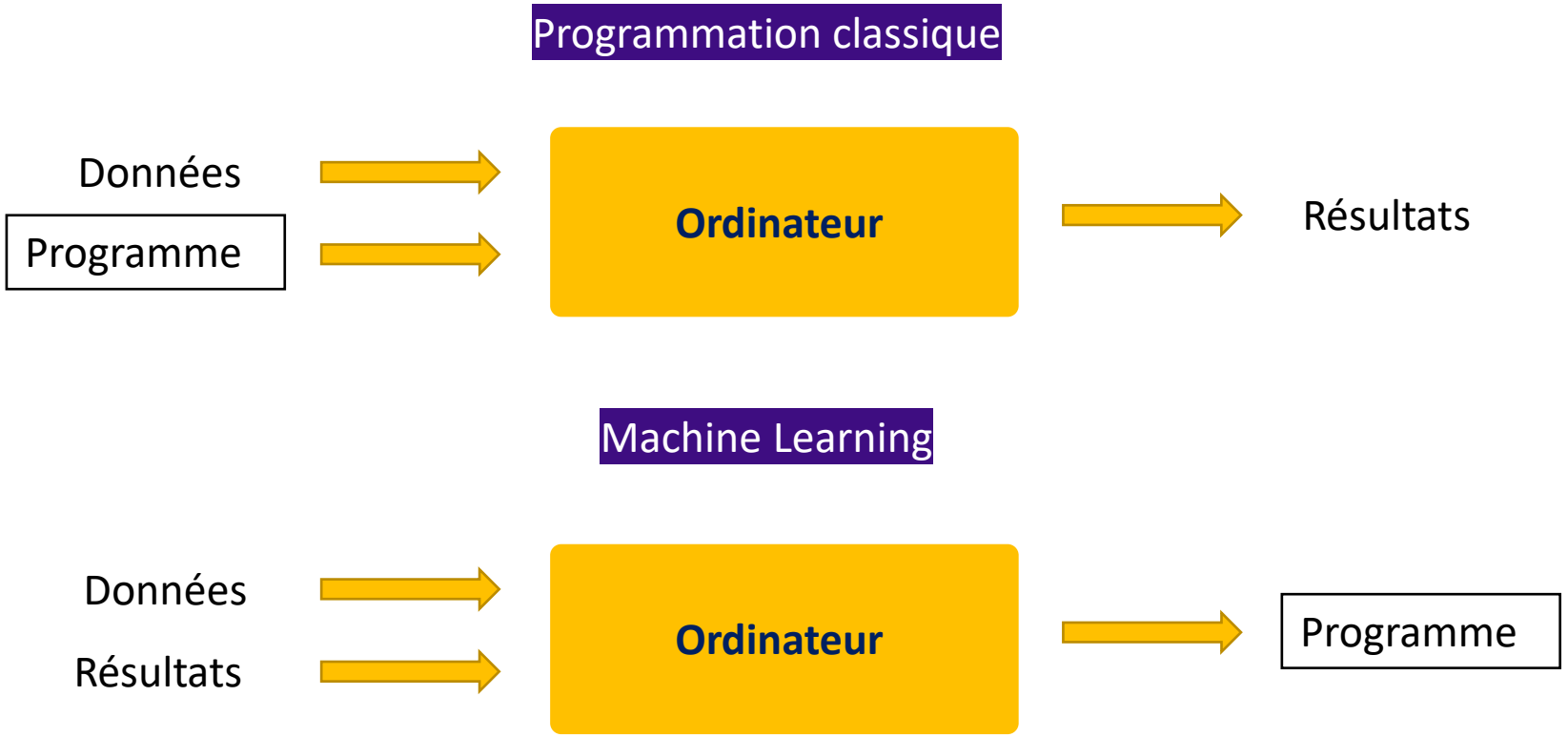
➔ Objectifs du Machine Learning

Le Machine Learning

Arthur Samuel, 1959

« Un champ d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés. »

Machine Learning vs Programmation classique



Le Machine Learning, pourquoi maintenant ?

- De nombreux algorithmes efficaces et efficients sont disponibles
- De grandes quantités de données sont disponibles
- De grandes quantités de ressources de calcul sont disponibles

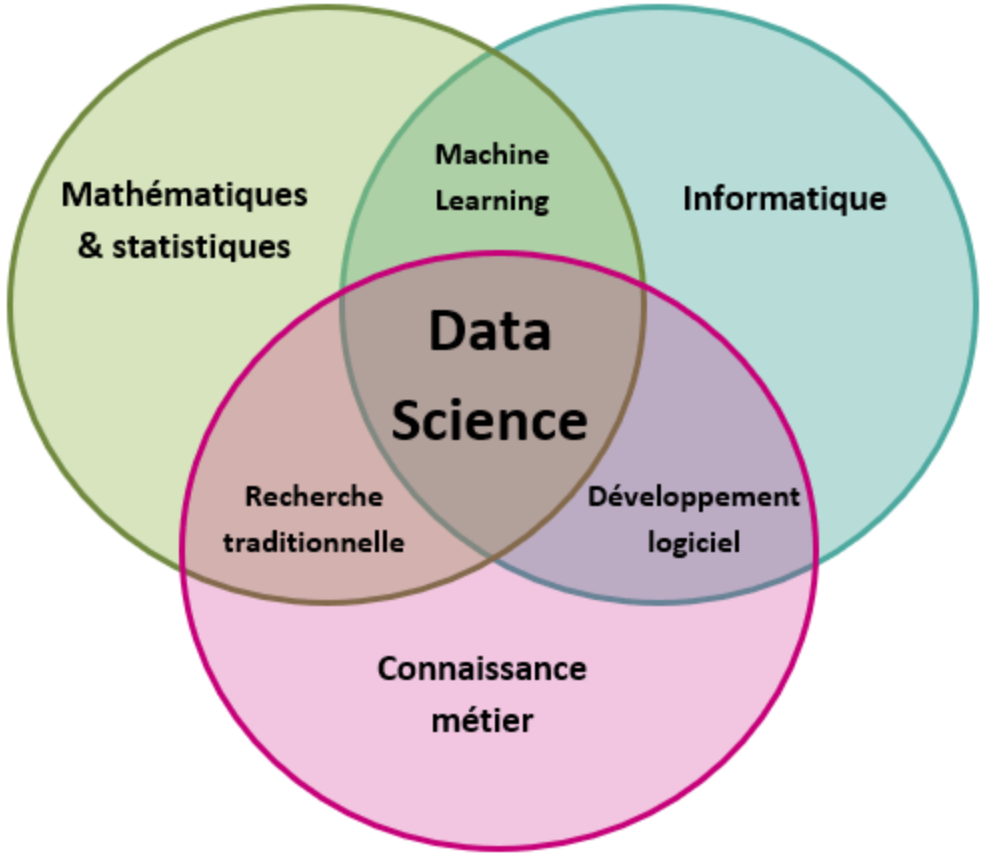
Mathématiques, statistiques, data mining et data science

Patil (LinkedIn) et Hammerbacher (Facebook) :

« Analyste, ça fait trop Wall Street ; statisticien, ça agace les économistes ; chercheur scientifique, ça fait trop académique. Pourquoi pas "data scientist" ? »

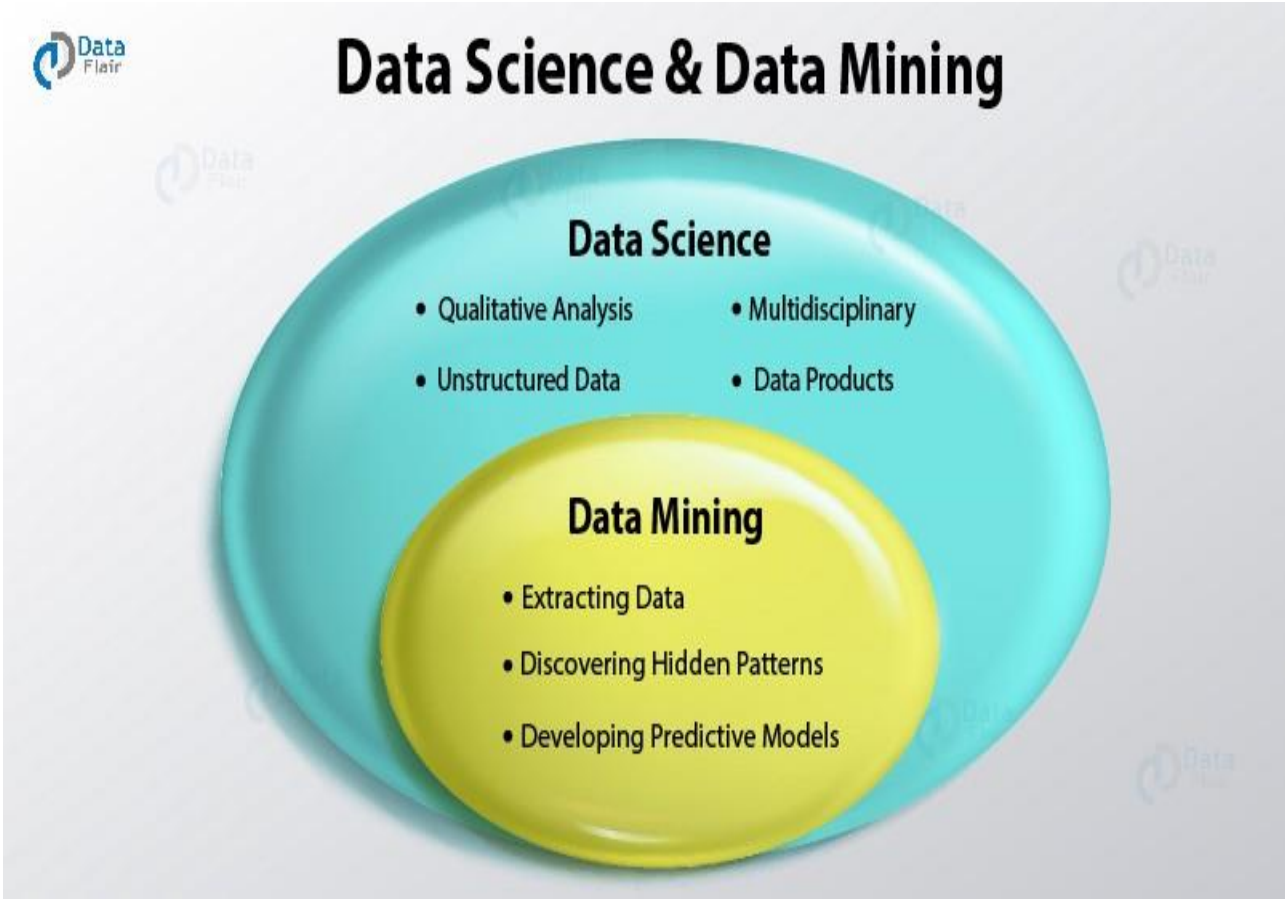
J. Wills (Cloudera):

« Data scientist (n) : Person who is better at statistics than any software engineer and better at software than any statistician »



Data science :

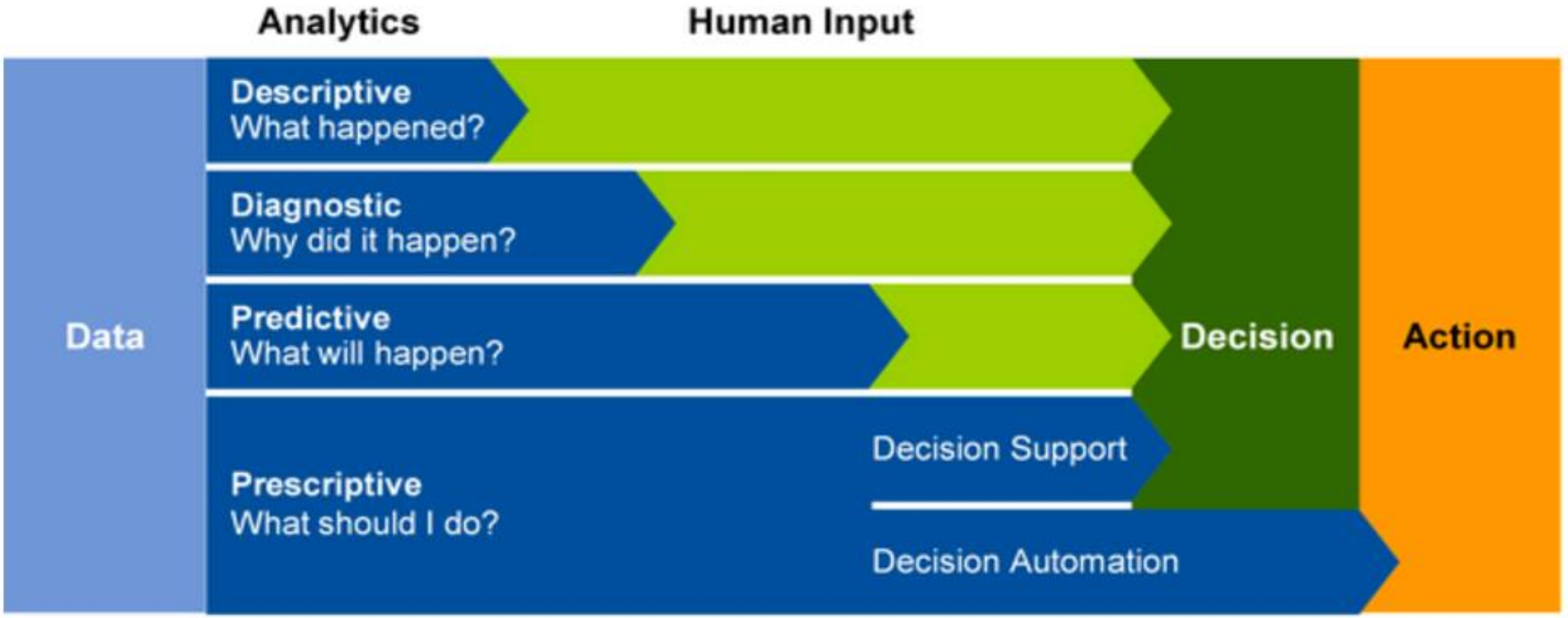
- Existence du concept depuis 1960
- Domaine
- Etude scientifique
- Construction de modèles prédictifs, analyse sociale et découverte de faits inconnus
- Multidisciplinaire
- Données structurées, semi-structrées et non structurées
- Science axée sur les données
- Créer des produits centrés sur les données pour une organisation



Data mining (exploration de données) :

- Apparition du concept dans les années 1990
- Technique
- Processus métier
- Découvrir des faits inconnus ou ignorés
- Sous-branche de la data science
- Données structurées
- Archéologie des données ou extraction de connaissances
- Rendre les données disponibles plus utilisables

Analyse descriptive, prédictive et prescriptive



Analyse descriptive, prédictive et prescriptive

Analyse descriptive :

- Collecter, catégoriser et classer les données
- Identifier et visualiser les modèles

Analyse prédictive :

- Planification et prise de décision à partir de modélisation
- Modélisation du processus de ce qui s'est passé et ce qui se passera

Analyse prescriptive :

- Prescrire la meilleure décision à partir des données décrites et des predictions disponibles
- Evaluer les résultats

Exemples d'applications

- Le diagnostic médical
- La détection de fraude
- Le filtrage de spam dans le courrier électronique
- Les systèmes de recommandations (articles dans un journal, livres, films, musiques, ...)
- Les investissements financiers
- Lecture de chèques
- L'aide à la prise de décision
- Reconnaissance de paroles
- Reconnaissance de caractères manuscrits
- Analyse de satisfaction client

Typologie des algorithmes : exemples

Quatre types d'algorithmes :

- Popularité

Mesure d'audience, décompte de clics des sites web

- Autorité

Classement de l'information, force sociale d'une page à partir du nombre de citations

- Réputation

Influence sur les réseaux sociaux (likes ou autre émoticône)

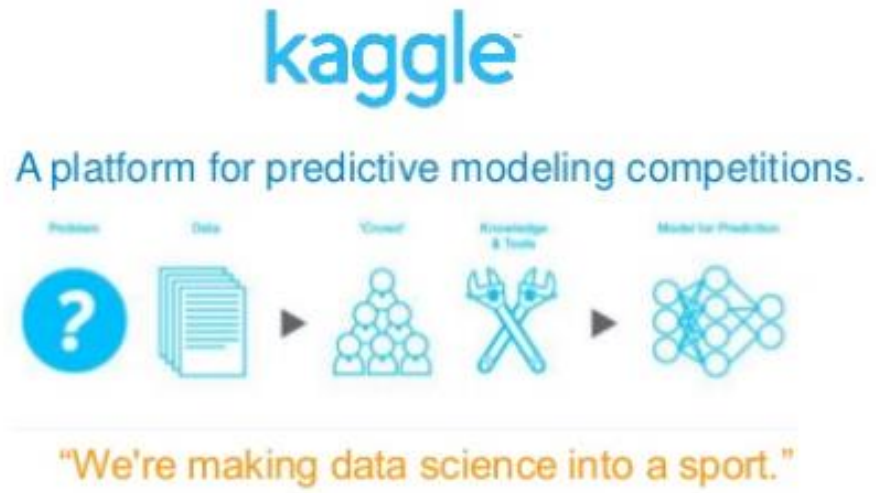
- Prévision

Comportement des internautes, orientation contenu publicitaire (cookies)

Communauté Data Science et Challenges Kaggle

Kaggle, c’est quoi ?

- Une plateforme de competitions de data, de Machine Learning
- L’objectif est de créer le meilleur algorithme possible capable de résoudre une problématique donnée
- Acquisition de nouvelles compétences, reconnaissance de la communauté et gain financier



The Netflix Prize

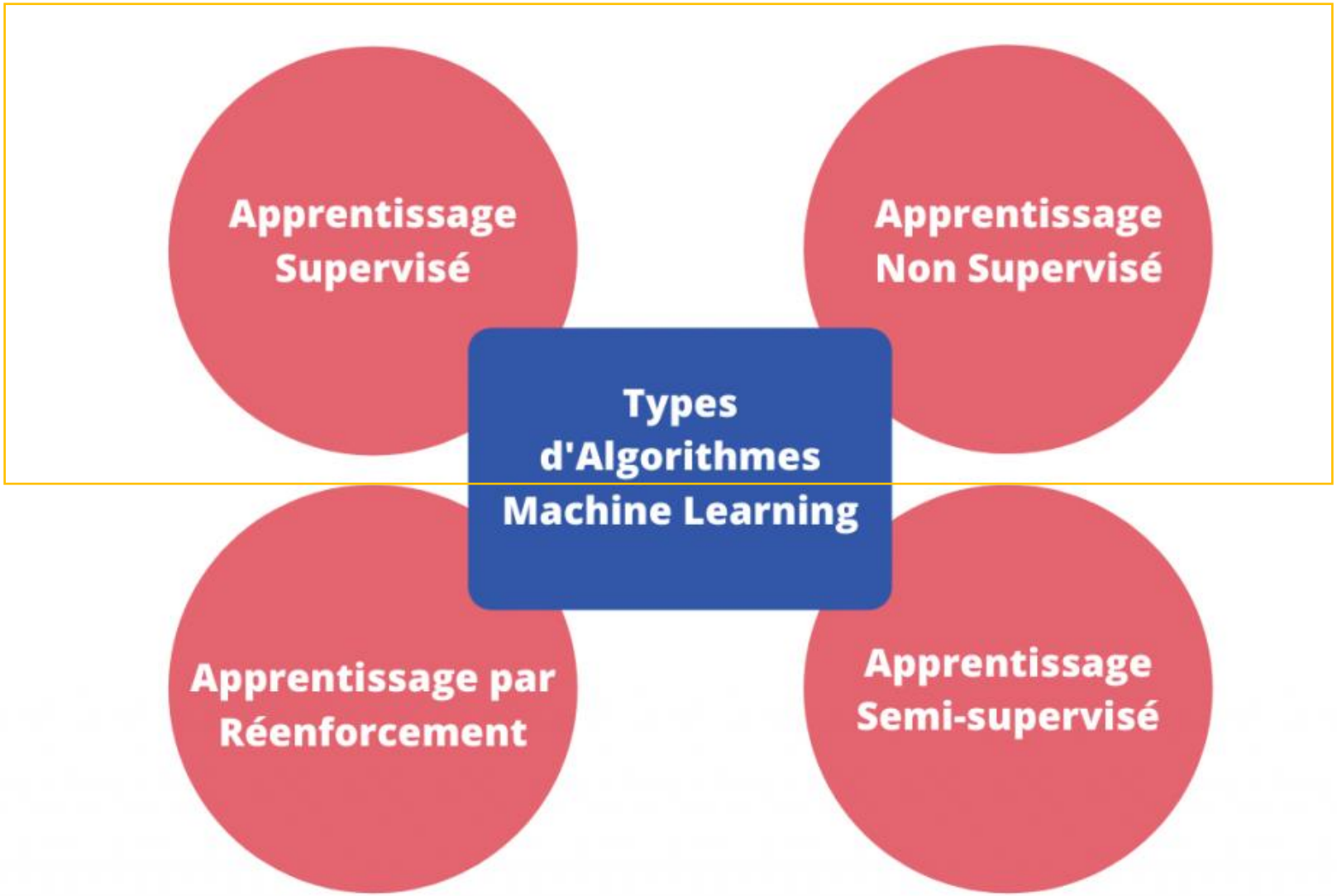
- Concours organisé par Netflix via Kaggle
- Réaliser le meilleur algorithme pour prédire les notes des utilisateurs pour les films
- Récompense de 1 million de $
- Prix remporté en juin 2009 par l'équipe Pragmatic Chaos de Bellkor

The Winners of the Netflix Prize | Getty Images/Jason Kempin/Staff Editoria

2

Les différents types d’apprentissage de ML

Types d’apprentissage en Machine Learning



Que fait chaque type de ML : objectif, types de données, familles de ML dans chacun de ces types etc

30 min

Les différents types d’apprentissage

- **Apprentissage supervisé**
  - Chaque exemple est associé à une étiquette
  - Objectif : prédire l’étiquette de chaque donnée
  - Le système apprend à classer les données

- **Apprentissage non supervisé**
  - Les exemples ne sont pas étiquetés
  - Objectif : trouver une structure aux données
  - Le système apprend un groupement des données

- **Apprentissage semi-supervisé**
  - Certains exemples sont étiquetés et d’autres non étiquetés
  - Objectif : adapter le modèle à la structure du problème
  - Le système entraîne des données étiquetées et exploiter les données non étiquetées (plus nombreuses) en vue d’améliorer les performances

Les différents types d’apprentissage

- **Apprentissage par renforcement**
  - Les exemples sont (parfois) associés à une récompense ou une punition
  - Objectif : trouver les actions qui maximisent les récompenses
  - Le système apprend une politique de décision

Types d’apprentissage en Machine Learning

Apprentissage supervisé

- **Formulation du problème**
  - On dispose d’un certain nombre d’exemples de la réalisation d’une tâche, sous forme de paires (entrée, résultat)
  - On souhaite réaliser un système capable de trouver de façon automatique et relativement fiable les résultats correspondant à toute nouvelle entrée qui pourrait lui être présentée

- **Trois types de tâches de l’apprentissage supervisé**
  - Régression: sortie de type continue
  - Classification: sortie de type nominale
  - Séries temporelles: prédire les valeurs futures d’une certaine quantité connaissant ses valeurs passées ainsi que d’autres informations. Par exemple le rendement d’une action en bourse…

Types d’apprentissage en Machine Learning

Apprentissage non supervisé

- L’objectif est de mettre en évidence des informations présentes mais cachées par le volume des données.
- Il n’y a pas de variable « cible » à prédire :
  - **Analyse factorielle** => Projeter un nuage de points sur un espace de dimension inférieure pour obtenir une visualisation de l’ensemble des liaisons entre les variables tout en minimisant la perte d’information.
  - **Association** => Trouver des événements ayant une forte probabilité de se réaliser ensemble.
  - **Classement (Segmentation ou Clustering)** => Identifier des groupes d’items ayant un comportement similaire. C’est le fait de regrouper des objets en groupes, ou classes, ou familles, ou segments, ou clusters de sorte que :
    - Deux objets d’un même groupe se ressemblent le plus possible
    - Deux objets de groupes distincts diffèrent le plus possible

Types d’apprentissage en Machine Learning

Autres types d’apprentissage

❑ Apprentissage par transfert

Transférer des connaissances acquises en réalisant une tâche antérieure, dans le but de réaliser une nouvelle tâche différente comportant des similitudes.

❑ Apprentissage séquentiel

Apprentissage réalisé lorsque les données sont obtenues ‘à la volée’ et traitées les unes après les autres (« petit à petit »)

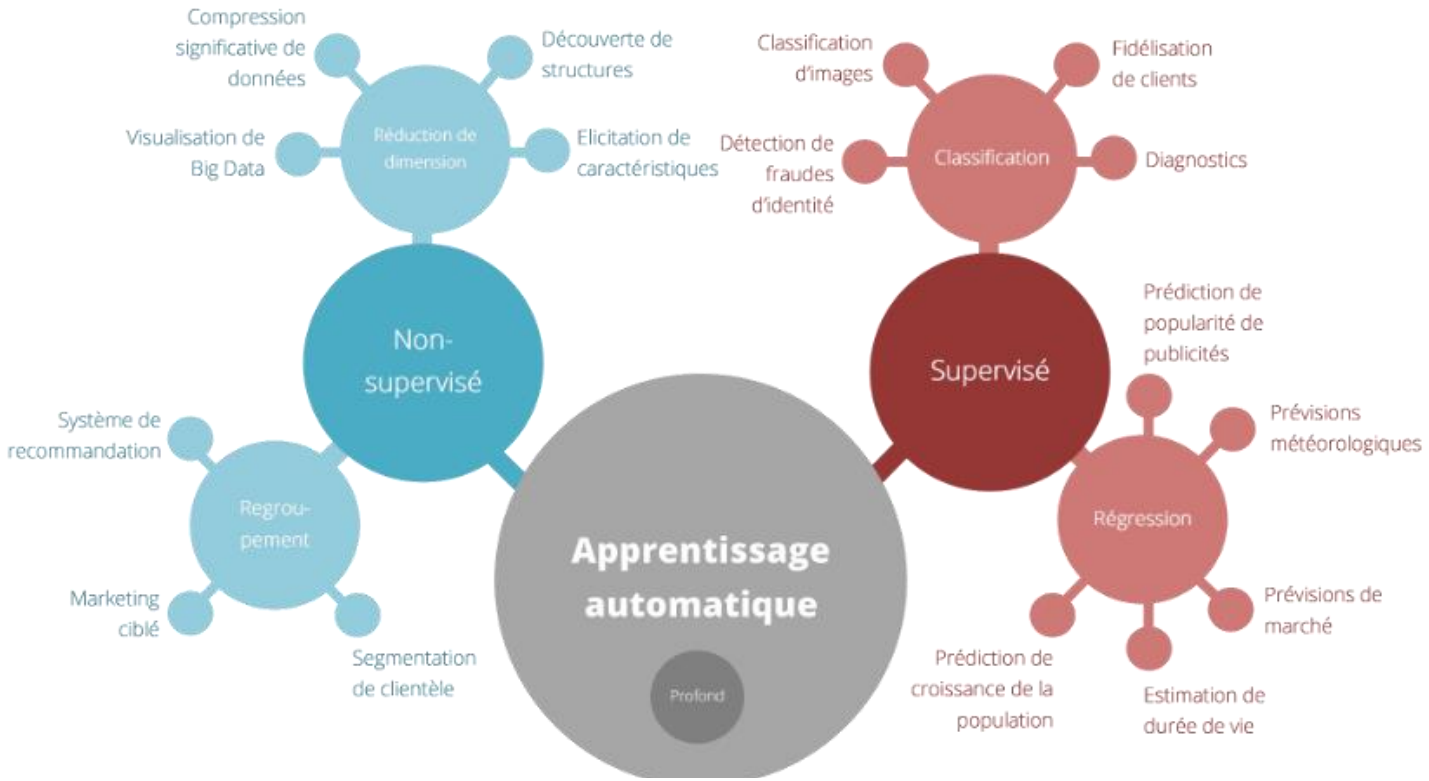
❑ Apprentissage actif

Apprentissage itératif du modèle en interaction avec un expert humain

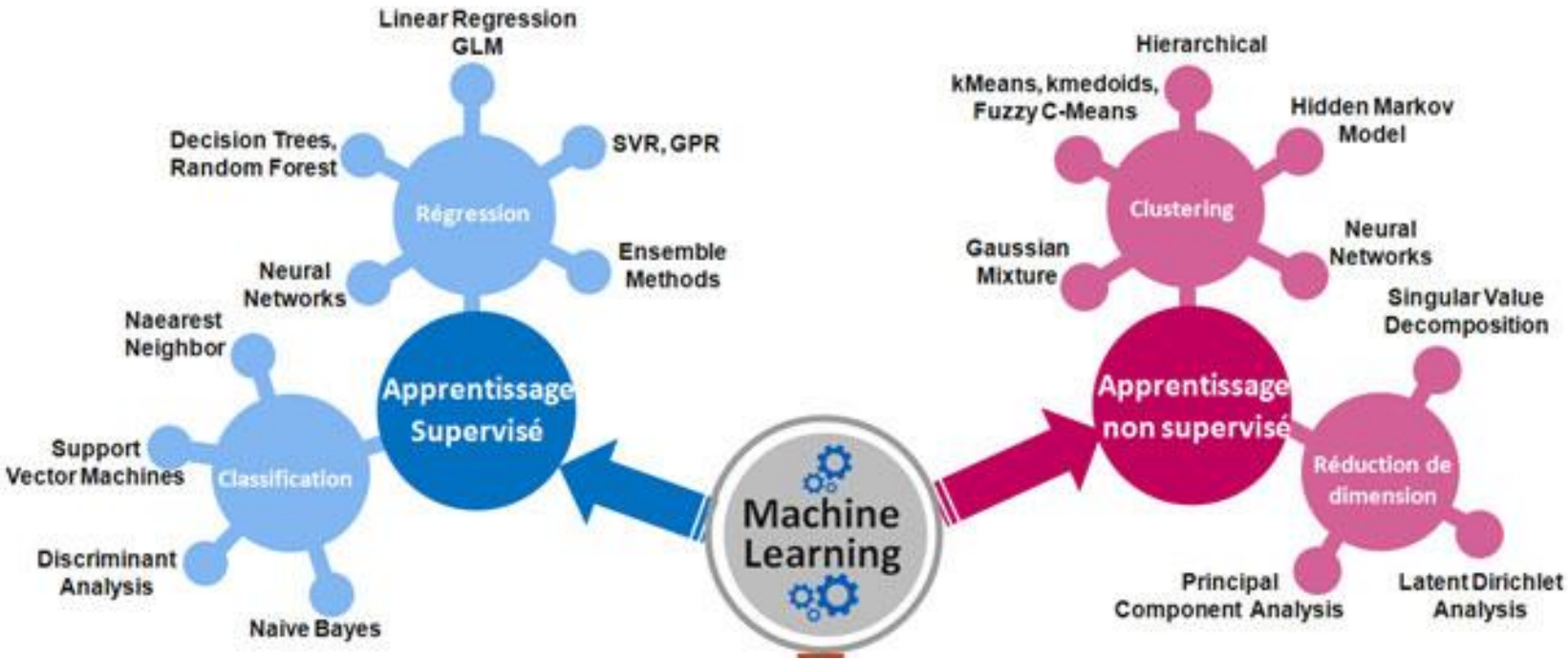
3

Les algorithmes de ML

Algorithmes de Machine Learning



Algorithmes de Machine Learning



Algorithmes de Machine Learning

La régression

- La régression est une technique de modélisation qui permet de mettre en équation une relation entre une variable à expliquer et n variables explicatives.

  Exemple : je souhaite prédire le prix d'un bien immo (variable à expliquer) en fonction de sa superficie, nb de chambre, quartier, … (variables explicatives)

- La régression permet d'analyser la manière dont une variable expliquée est affectée par les valeurs d'une ou plusieurs autres variables explicatives

- Un problème de régression consiste à chercher une fonction f telle que pour tout i, Yi soit approximativement égale à f(Xi).

Algorithmes de Machine Learning

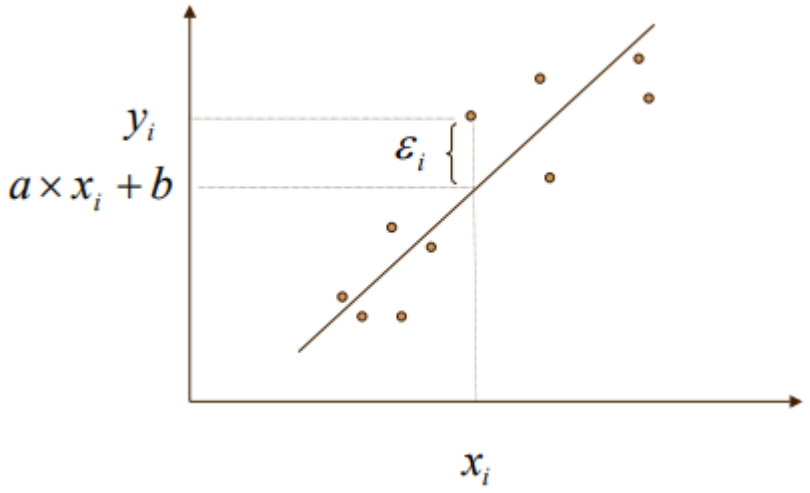
La régression linéaire simple

Principe de la régression linéaire simple

- La relation entre deux variables x et y est décrite par :

$$\boldsymbol{Y_i = ax_i + b + \varepsilon_i}$$

- Où $\boldsymbol{a}$ et $\boldsymbol{b}$ sont deux constantes que l'on cherche à évaluer et ε est un terme aléatoire que l'on appelle erreur
- On estime $\boldsymbol{a}$ et $\boldsymbol{b}$ grâce à l'échantillon $(x_1, y_1), \ldots, (x_n, y_n)$



Algorithmes de Machine Learning

La régression linéaire multiple

- Plus de deux variables
- Une variable expliquée (ou endogène) **y** et plusieurs variables explicatives (ou exogènes) $x_1, x_2, \ldots, x_i$, avec $i = 1,\ldots,n$
- La relation entre les variables est, dans ce cas, décrit par :

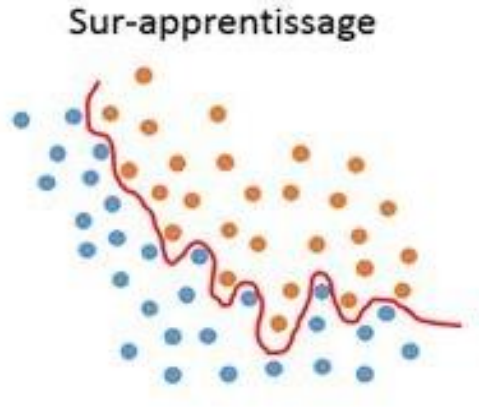
$$Y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + apxip + \varepsilon_i \ , i = 1,\ldots,n$$

  - Où a_0, a_1, …, a_p sont les paramètres du modèle à estimer
  - ε_i est l'erreur du modèle

Algorithmes de Machine Learning

Limites des approches linéaires

- Multicolinéarité
- Nombre de variables plus grand que celui d'observations
- Risque de sur-apprentissage

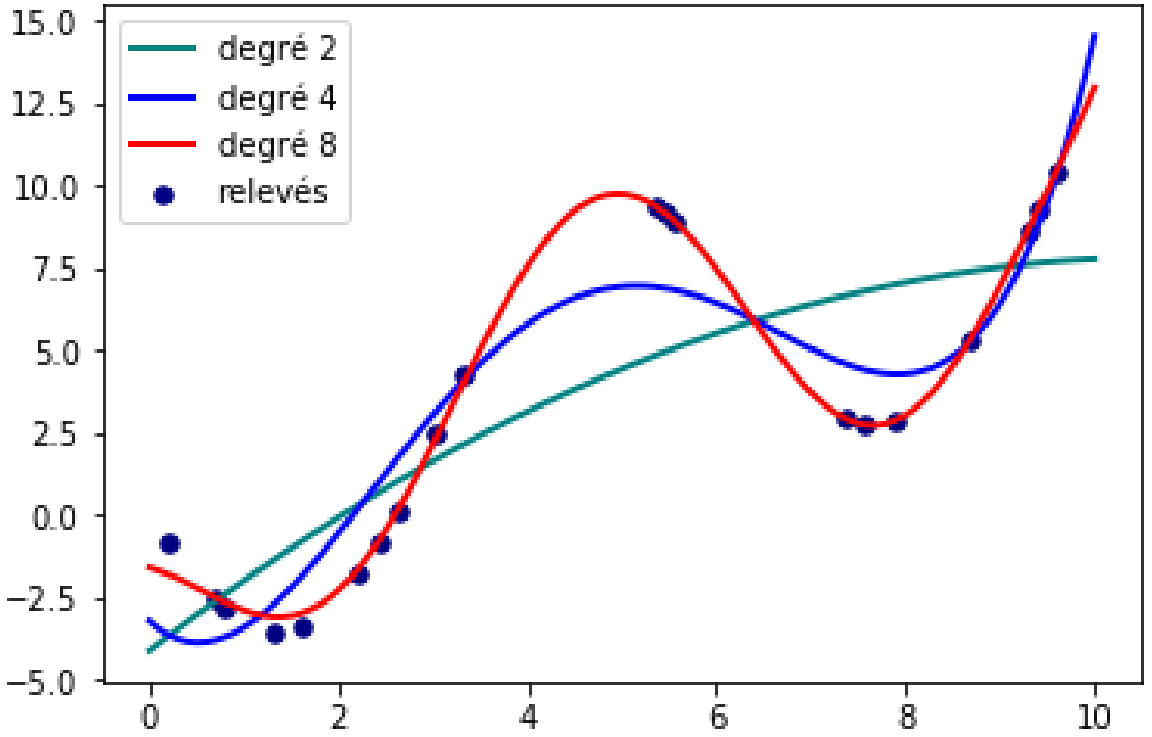


Un modèle trop spécialisé sur les données d'entraînement et qui se généralisera mal

Algorithmes de Machine Learning

La régression polynomiale

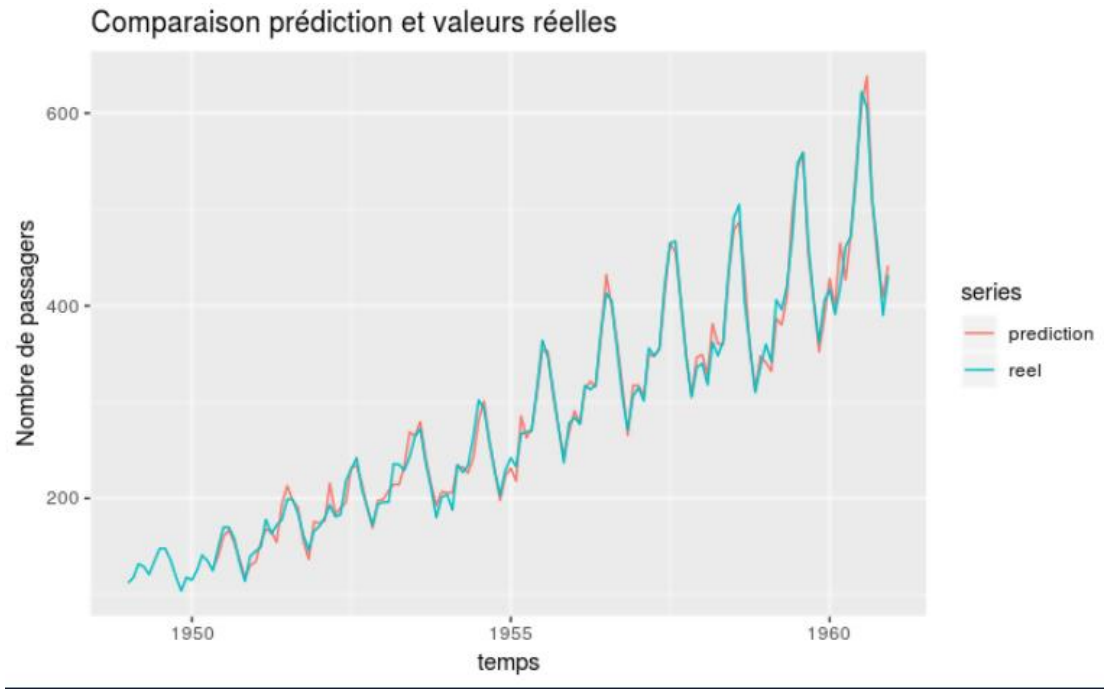
- Généralisation de la régression linéaire multiple
- Introduction de la non-linéarité
- Adaptée pour des données sans relation linéaire apparente



Algorithmes de Machine Learning

Les séries temporelles

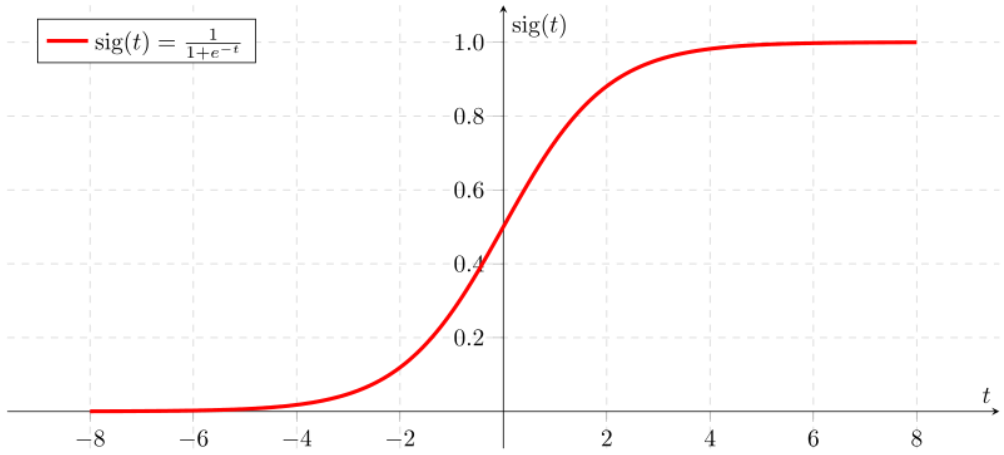
- Suite d'observations d'une même variable dans le temps
- Evolution d'une variable dans le temps
- PIB d'un pays, revenu d'un individu, pluviométrie, nombre de votants



Algorithmes de Machine Learning

La régression logistique

- Modèle de classification linéaire dans lequel Y a deux valeurs possibles 0 ou 1 (non malade/malade ; non solvable/solvable)
- Régression linéaire binomiale



Algorithmes de Machine Learning

Les applications en scoring

- Scoring client (marketing) : Attribuer un score à chaque client afin d'identifier les clients intéressants
- Scoring de crédit (finance) : Evaluer la possibilité d'accorder un crédit à un emprunteur potentiel
- Scoring en assurance : Evaluer les risques d'un assuré potentiel
- Exemple agence de voyage

Algorithmes de Machine Learning

La classification

- La classification est le problème d'identifier à quelle catégorie (sous-population) une nouvelle observation appartient, à partir d'un ensemble de données d'apprentissage contenant des observations (ou instances) dont l'appartenance à une catégorie est connue.
- La classification permet de prédire si un élément est membre d'un groupe ou d'une catégorie donnée.
- Les classes (connues à l'avance) permettent :
  - L'identification de groupes avec des profils particuliers.
  - La possibilité de décider de l'appartenance d'une entité à une classe.

Exemple : je cherche à prédire si un patient est atteint d'une maladie ou non (= la classe) en fonction de son rythme cardiaque, son taux de globule blanc, …

Algorithmes de Machine Learning

La classification hiérarchique

Principe de la classification ascendante hiérarchique (CAH)

1. A l'étape initiale, les n individus constituent des classes à eux seuls.
2. On calcule les distances deux à deux entre individus, et les deux individus les plus proches sont réunis en une classe.
3. La distance entre cette nouvelle classe et les n-2 individus restants est ensuite calculée, et à nouveau les deux éléments (classes ou individus) les plus proches sont réunis.

- Ce processus est réitéré jusqu'à ce qu'il ne reste plus qu'une unique classe constituée de tous les individus. On constate que la nouveauté ici vient de la nécessité de définir deux distances : la distance usuelle entre deux individus, et une distance entre classes.

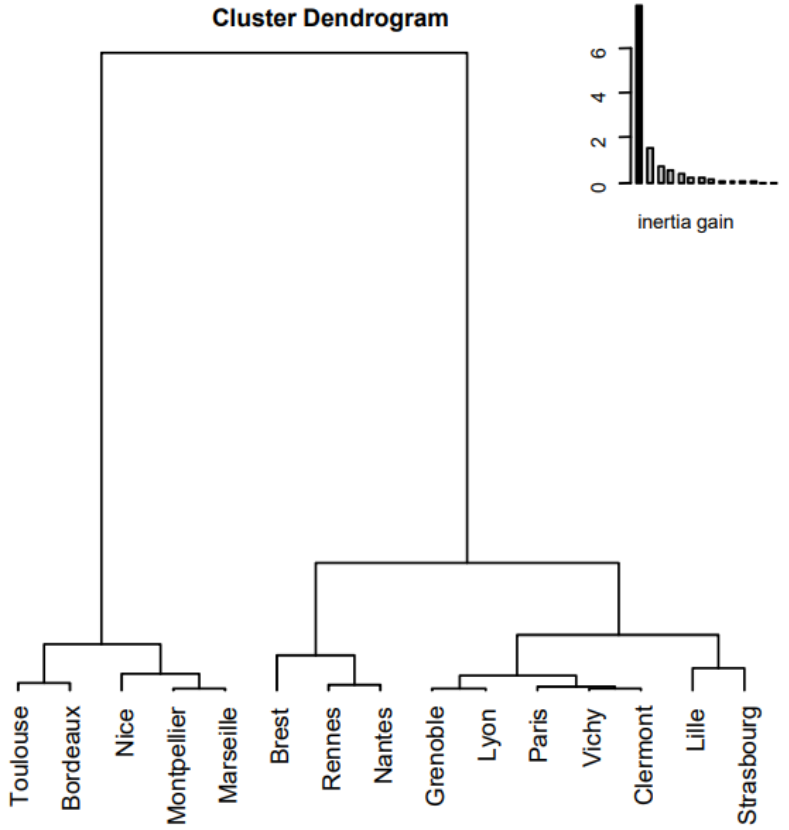
Algorithmes de Machine Learning

La classification hiérarchique

Illustration de la classification ascendante hiérarchique (CAH)

- Quelles villes sont des profils météo similaires?

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4

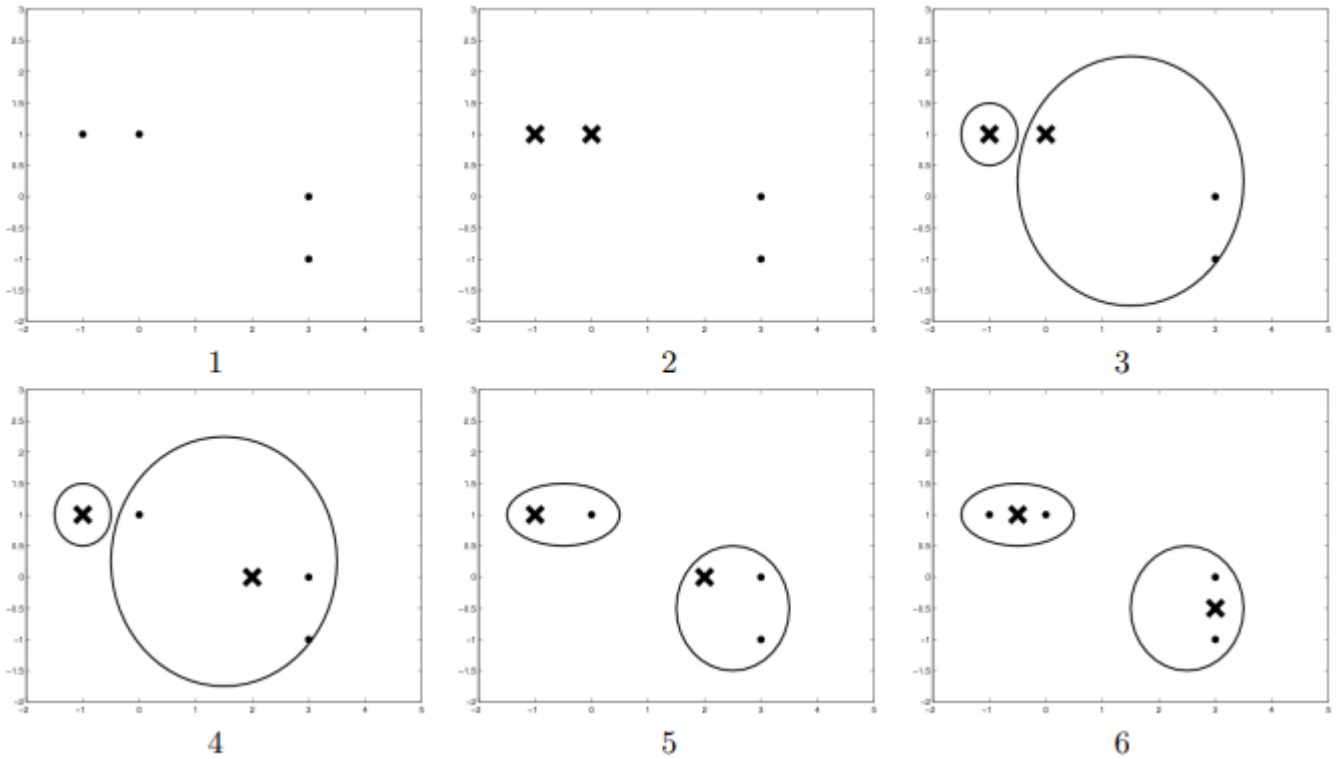


Algorithmes de Machine Learning

La classification non hiérarchique (K-means)

Illustration de la méthode des K-means

- On applique l'algorithme sur un exemple où quatre points a (-1,1), b (0,1), c (3,0) et d (3,-1) doivent être classés en 2 classes.
- On remarque sur cet exemple que bien qu'à l'initialisation les centres de classes sont mal répartis, l'algorithme a convergé en retrouvant les "vraies" classes.



4

L’évaluation des modèles de ML

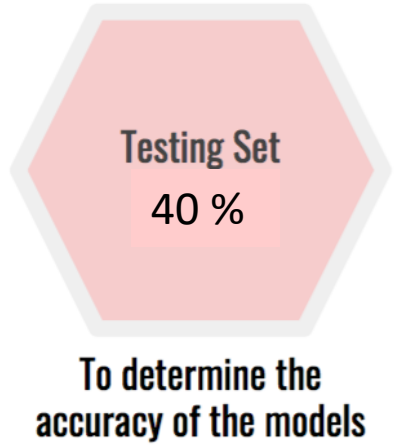
Pour évaluer (= dire si le modèle est performant ou non) un modèle de ML supervise, on a l'habitude de diviser le jeu de données en 2 : jeu de train, jeu de test. Pourquoi?

15 min

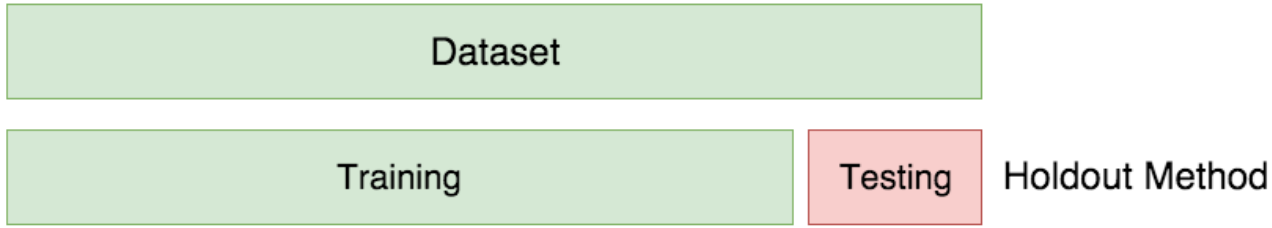
Entraînement et évaluation des algorithmes de régression

Séparation du jeu de données : entraînement, test et validation

- Jeu de données original subdivisé en :
  - ✓ Données d'entraînement : Pour entraîner les modèles
  - ✓ Données de test : Pour déterminer la précision du modèle



Entraînement et évaluation des algorithmes supervisés
Exemple



Jeu de training :

- Y : variable à prédire (= à expliquer)
- X : variables explicatives

Jeu de testing :

- Ypred : ce que l'algorithme va prédire
- Yreel : les valeurs réelles
- X : variables explicatives

Entraînement et évaluation des algorithmes de regression

Définition d'une métrique de performance

Mesures de performance

Régression	Classification
Erreur quadratique (MSE) Erreur absolue (MAE) Coefficient de détermination	Exactitude (accuracy) Précision Rappel (recall) Sensibilité & Spécificité

Entraînement et évaluation des algorithmes de classification

Matrice de confusion et de coût

- **Matrice de confusion**
  - Indiquer la structure de l'erreur du modèle
  - Trouver la manière dont le modèle se trompe

- **Matrice de coût** (à partir de la matrice de confusion)
  - Mesurer les erreurs du modèle
  - Evaluer le coût de chaque type d'erreur

Les deux matrices permettent de comparer les performances des modèles

Le calcul doit être fait sur l'échantillon test

Entraînement et évaluation des algorithmes de classificatior

Qualité d'une classification

- **VP/VN (Le nombre de vrais positifs/négatifs) :** Les exemples de classe positive/négative dont la classe est prédite comme positive/négative.
- **FP/FN (Le nombre de faux positifs/négatifs) :** Les exemples de classe négative/positive dont la classe est prédite comme positive/négative.
- **Matrice de confusion :**

	Classe réelle +	Classe réelle -
Classe prédite +	**VP**	**FP**
Classe prédite -	**FN**	**VN**

Entraînement et évaluation des algorithmes de classification

Qualité d'une classification

- Le rappel (recall) mesure la proportion des exemples positifs trouvés parmi tous les exemples positifs.
- La précision (precision) mesure la proportion des exemples vraiment positifs parmi ceux qui étaient classés positifs.
- L'exactitude (accuracy) mesure la proportion des exemples bien classés parmi tous les exemples.