

Data Science, Machine Learning et IA

De la Théorie aux Applications Avancées

Aziza Merzouki

Université de Genève

28 Novembre, 2025

Programme

■ Jour 1:

- Fondamentaux de la Data Science et du Machine Learning
- Pipeline de l'apprentissage automatique : Des données au modèle
- Travaux Pratiques

■ Jour 2:

- Introduction au Deep Learning
- Modèles de Deep Learning
- Travaux Pratiques

■ Jour 3:

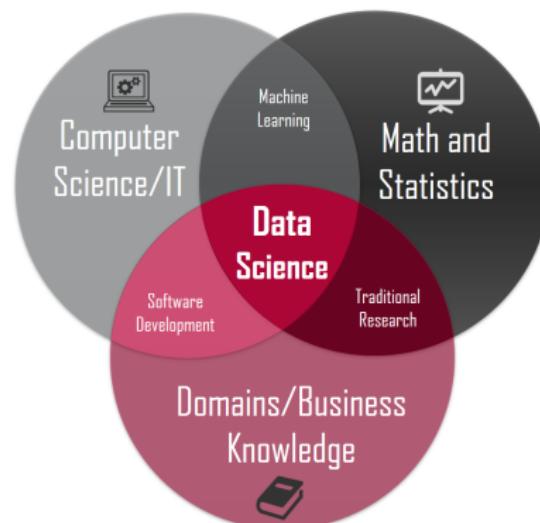
- LLMs et Prompt Engineering
- RAG et Variantes
- Travaux Pratiques

Outline

- 1 Data Science - Introduction
- 2 Data Science, IA, et ML
- 3 Rôle du Data Scientist
- 4 Types d'apprentissage, de problèmes et d'applications abordés par le ML
 - Apprentissage supervisé
 - Apprentissage non-supervisé
 - Autres types d'apprentissage
- 5 Quizz
- 6 Conclusion

Introduction à la Data Science

La Data Science est une discipline qui combine des compétences en **programmation**, en **statistiques** et en **domaines d'expertise** pour explorer, analyser et extraire des **connaissances** et informations utiles à partir de grandes quantités de **données** structurées et non-structurées.



Relations entre Data Science, IA, et ML

Data Science

- Collection, preparation, and analysis of data
- Leverages AI/ML, research, industry expertise, and statistics to make business decisions

Data Science

Machine Learning

- Algorithms that help machines improve through supervised, unsupervised, and reinforcement learning
- Subset of AI and Data Science tool

Machine
Learning

Artificial Intelligence

- Technology for machines to understand/interpret, learn, and make 'intelligent' decisions
- Includes Machine Learning among many other fields

Intelligence Artificielle (IA)

Intelligence Artificielle: Vise à créer des systèmes qui **simulent l'intelligence humaine**, y compris la prise de décision et l'apprentissage.



Machine Learning (ML)

Machine Learning: Sous-domaine de l'IA qui permet aux ordinateurs **d'apprendre des données** et d'améliorer leurs performances **sans être explicitement programmés**.

Programmation implicite : Apprentissage



programme = valeurs
des paramètres



expérience
data !!



programme mis à jour

Machine Learning (ML)

Machine Learning: Sous-domaine de l'IA qui permet aux ordinateurs **d'apprendre des données** et d'améliorer leurs performances **sans être explicitement programmés**.

The traditional approach

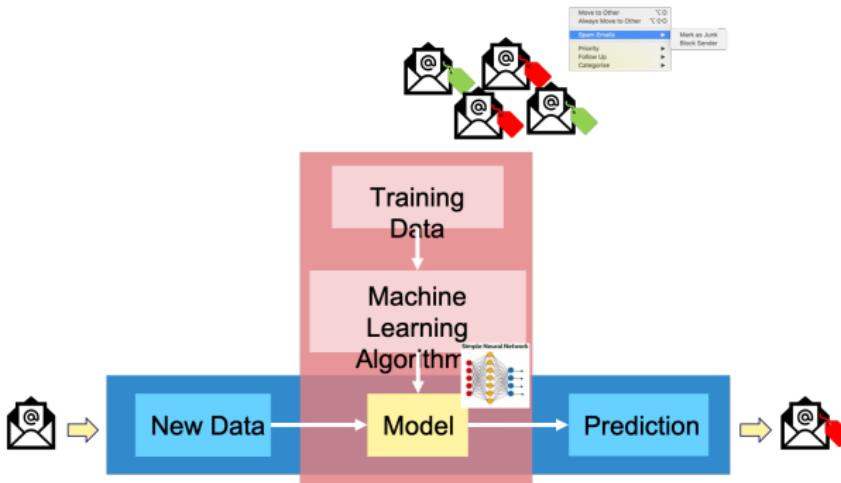
```
if body.contains("Especially 4U") or  
    body.contains("Very important") or ...  
// SPAM
```

Write rules

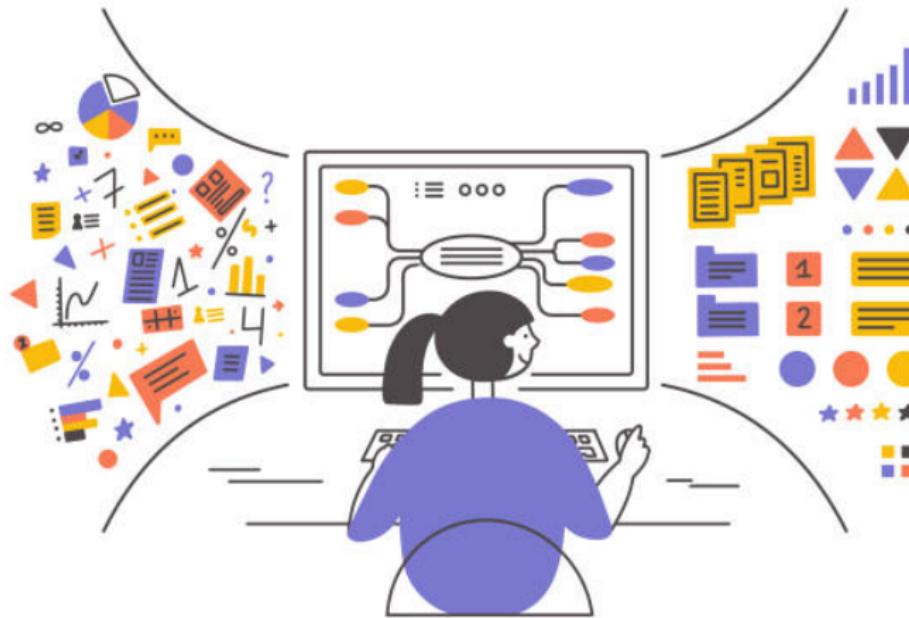


Machine Learning (ML)

Machine Learning: Sous-domaine de l'IA qui permet aux ordinateurs **d'apprendre des données** et d'améliorer leurs performances **sans être explicitement programmés**.



Rôle du Data Scientist



Rôle du Data Scientist - Données

- **Récolter les données** : Diverses sources, e.g., bases de données, fichiers, API.
- **Évaluer la fiabilité et la durabilité des sources** : Qualité, pas d'erreurs, adaptées à l'analyse.
- **Exploration et visualisation** : Identifier les tendances, schémas et anomalies, et rendre les données compréhensibles.
- **Pipeline de données** : Stockage, nettoyage, transformation et distribution.
- **Localiser/définir les "bonnes features"** : Identifier les caractéristiques ou variables pertinentes pour le problème à résoudre.
- **Synthétiser ce qui est important** : Extraire des informations significatives des données.

Rôle du Data Scientist - Modèles

- **Quel problème ?** : Identifier et définir clairement le problème à résoudre.
- **Une IA est-elle adaptée ?** : Est-ce la bonne approche pour résoudre le problème, ou est-ce qu'une autre méthode est plus appropriée?
- **Quelle méthode/modèle pour le résoudre ?** : En fonction de la nature du problème et des données disponibles.
- **Comment évaluer le modèle ?** : Mesurer les performances du modèle.
- **Métriques et objectifs ?** : Choisir des métriques pertinentes pour évaluer le modèle, en fonction des objectifs spécifiques.
- **Comment réviser un modèle ?** : Améliorer et itérer sur les modèles en fonction des résultats de l'évaluation.

Problèmes - Régression et Classification

- **Regression:** Prédit une **variable continue** (un nombre).
- **Classification:** Prédit une **variable discrète** (une classe).

Classification

Predict a **discrete variable** (a class)



Examples:

- {outside, inside}
- {yes, no}
- {cancel, refund, rewards, change}

Regression

Predict a **continuous variable** (a number)



Akama Resort
Urangan

Free Cancellation
Reserve now, pay later

4.7/5 Exceptional (331 reviews)

We have 3 left at
\$169 per night
\$186 total
Includes taxes & fees

Examples:

- room price
- the number of bookings in 5 minutes
- occupancy forecasting

Régression - Applications

- Prédire le prix de location d'un appartement.

Objet	Appartement		
Type d'objet	Appartement		
Pièces	3.5		
Caractéristiques	Animaux domestiques acceptés Balcon / Terrasse / Jardinier Garage / Place de parc		
Disponibilité	de suite		
Type d'annonce	Offre	Loyer, charges comprises	CHF 1'230.-

(a) Input

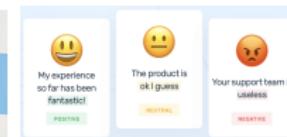
(b) Output

Classification - Applications

- Reconnaissance de **chiffres manuscrits** (tri automatique du courrier en fonction des codes postaux).
- Reconnaissance de **commandes vocales** (assistant virtuel).
- Déterminer si un e-mail est un **spam** ou non.
- Analyse de **sentiments** des critiques de produits.



Hey Siri

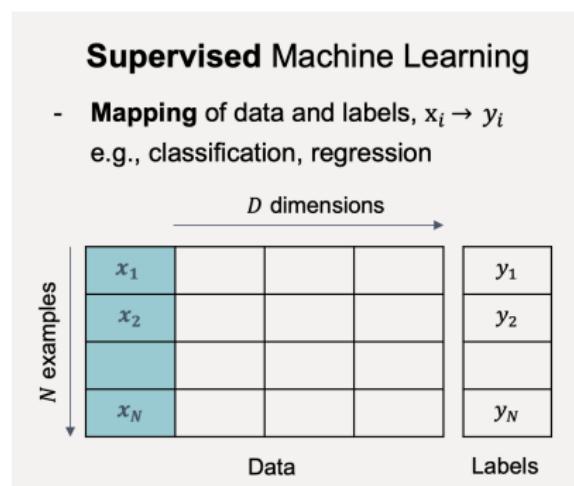


THE MNIST DATABASE of handwritten digits.

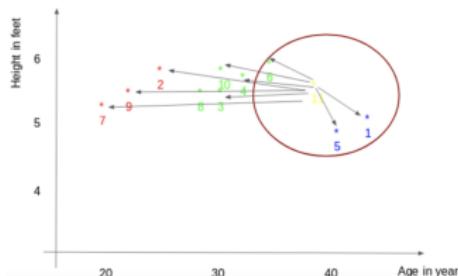
A database for handwritten text recognition research (US Postal Service).

Apprentissage supervisé

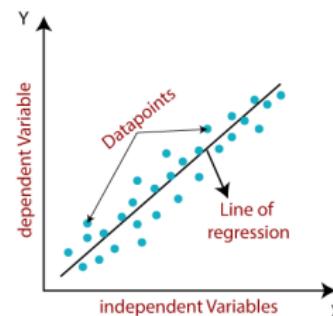
Apprentissage supervisé: Utilise des **données labellisées** pour entraîner le modèle. Le modèle prédit les étiquettes des nouvelles données.



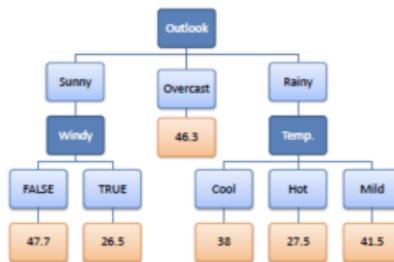
Régression - Exemples de modèles



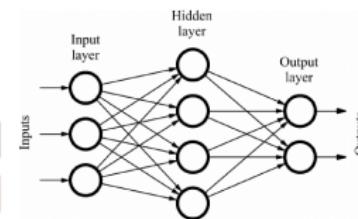
(a) K-NN



(b) Régression linéaire

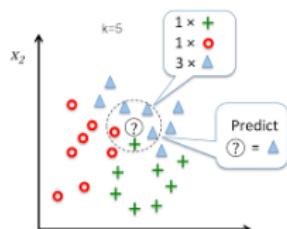


(c) Arbre de décision

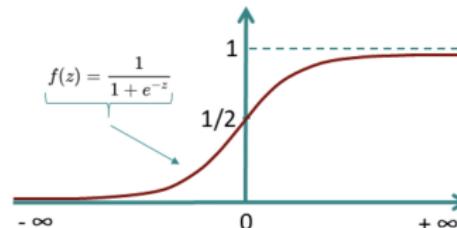


(d) Réseau de neurones

Classification - Exemples de modèles

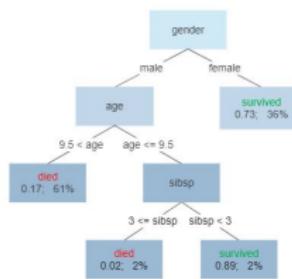


(a) K-NN

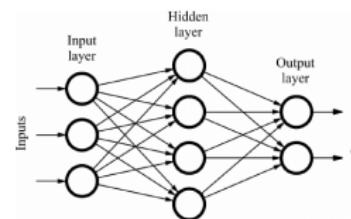


(b) Régression Logistique

Survival of passengers on the Titanic



(c) Arbre de décision

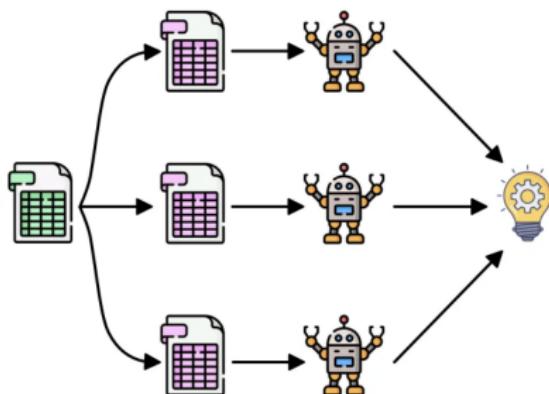


(d) Réseau de neurones

Méthodes d'Ensemble - Bagging et Boosting

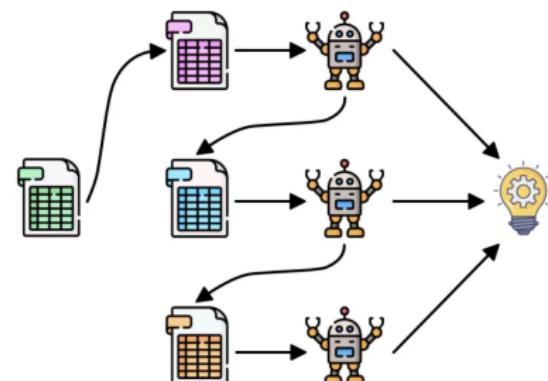
Bagging

- Plusieurs modèles entraînés en parallèle sur différents sous-ensembles de données.
- Prédictions moyennées (régression) ou votées (classification).



Boosting

- Modèles entraînés séquentiellement, chaque nouveau modèle corrige les erreurs des précédents.
- Le boosting se concentre sur les instances difficiles à prédire.

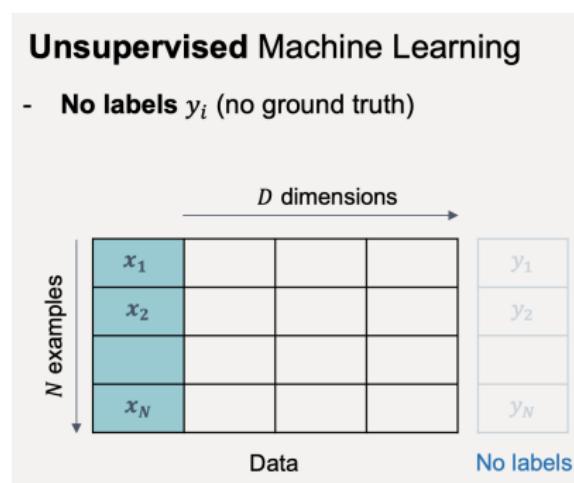


Ensemble Learning: Bagging & Boosting.

Bagging vs. Boosting in Machine Learning: Difference Between Bagging and Boosting.

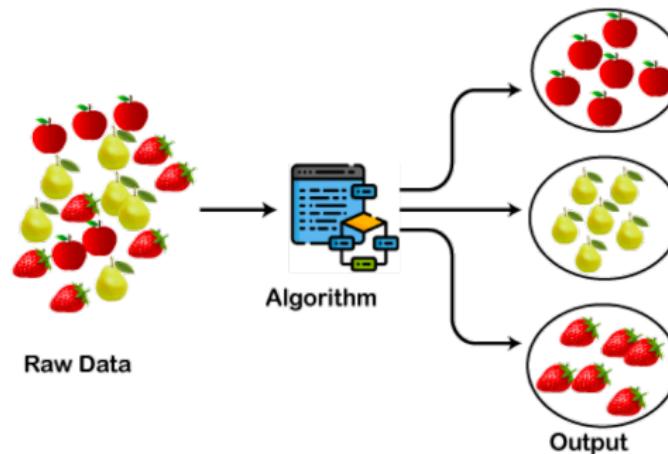
Apprentissage non-supervisé

Apprentissage non-supervisé: Aucune étiquette n'est fournie.
Le modèle identifie des structures et des motifs (**pattern**) dans les données.



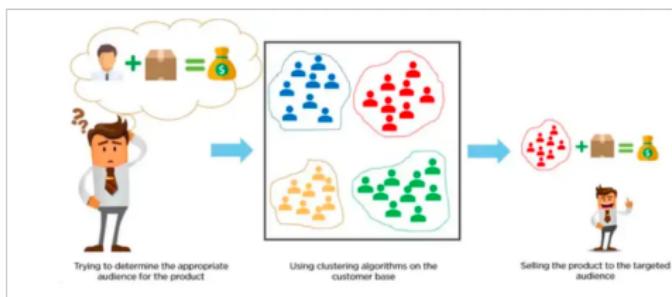
Problèmes - Clustering

- **Clustering:** Regrouper des données **similaires** en **sous-ensembles** homogènes.



Clustering - Applications

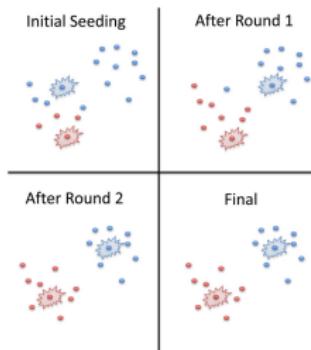
- **Marketing:** Segmentation des clients en fonction de leurs comportements d'achat.
- **Microciblage politique:** Segmentation de l'électorat potentiel pour élaborer des stratégies de communication ciblées.
- **Topic Modeling:** Identifier des thèmes principaux dans un grand corpus.



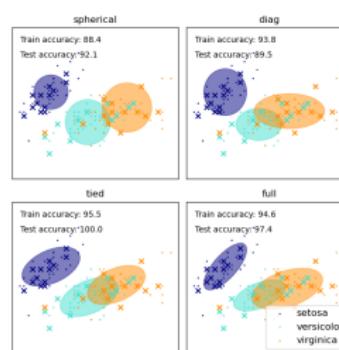
Behind the scenes at Donald Trump's UK digital war room. (Cambridge Analytica)

Google Trends: See what's trending across Google Search, Google News and YouTube.

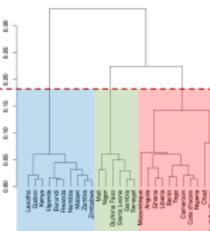
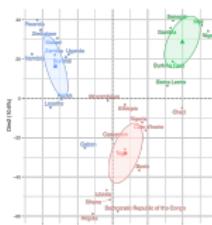
Clustering - Exemples de modèles



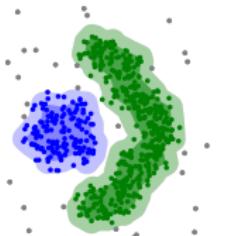
(a) K-means



(b) Gaussian Mixtures



(c) Clustering hiérarchique



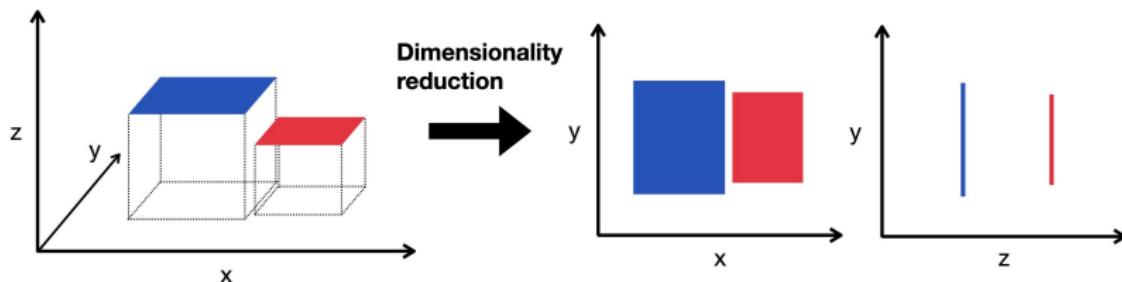
(d) HDBSCAN

Overview of clustering methods.

Types of clustering algorithms and how to select one for your use case.

Problèmes - Réduction de dimensions

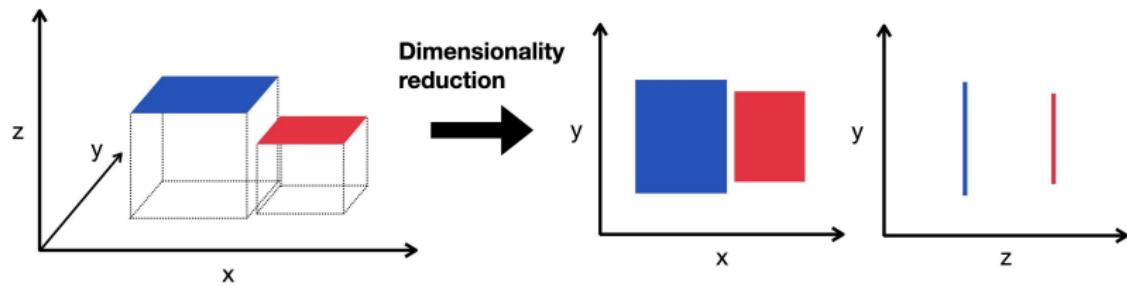
- **Réduction de dimensions:** Transformer des données depuis des **espaces de grande dimension** vers des espaces de **plus petite dimension** sans compromettre les **propriétés significatives** des données d'origine.



What Is the Curse of Dimensionality?.

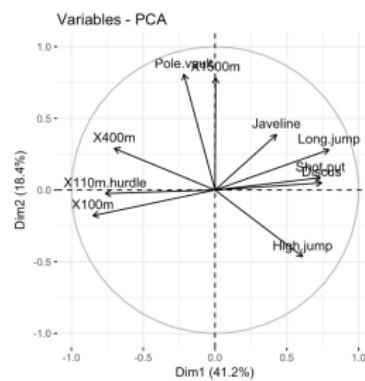
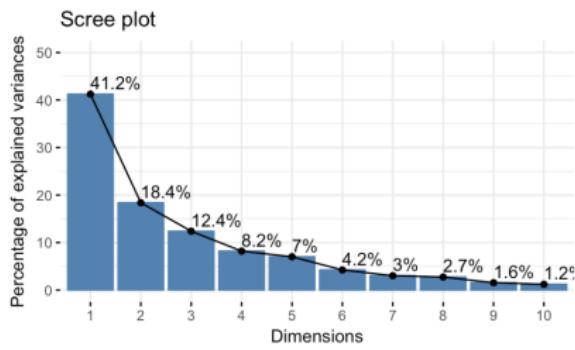
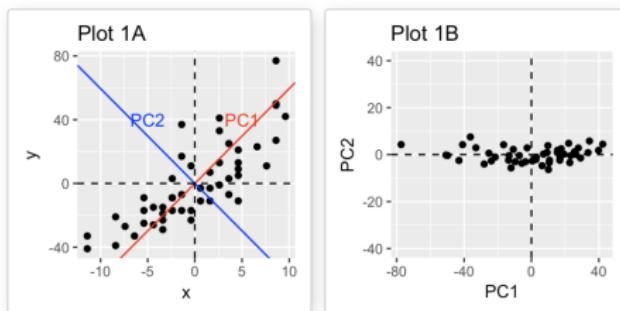
Réduction de dimensions - Applications

- **Visualisation:** Réduire des données depuis des **espaces de grande dimension** vers des espaces en 2D ou 3D à des fins de visualisation.
- **Pré-traitement:** Réduire le nombre de caractéristiques (features) pour éviter le fléau de la dimension (curse of dimensionality).



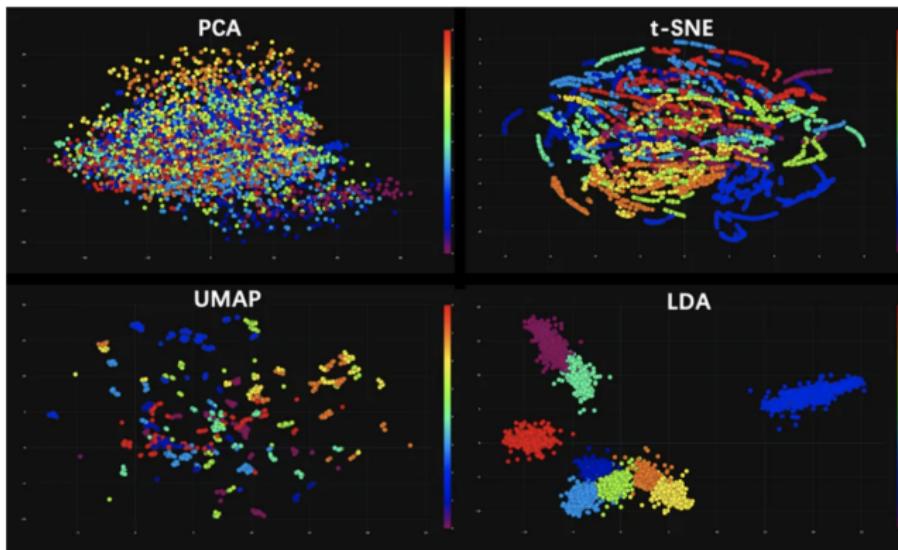
Réduction de dimensions - Exemples de modèles

PCA



Réduction de dimensions - Exemples de modèles

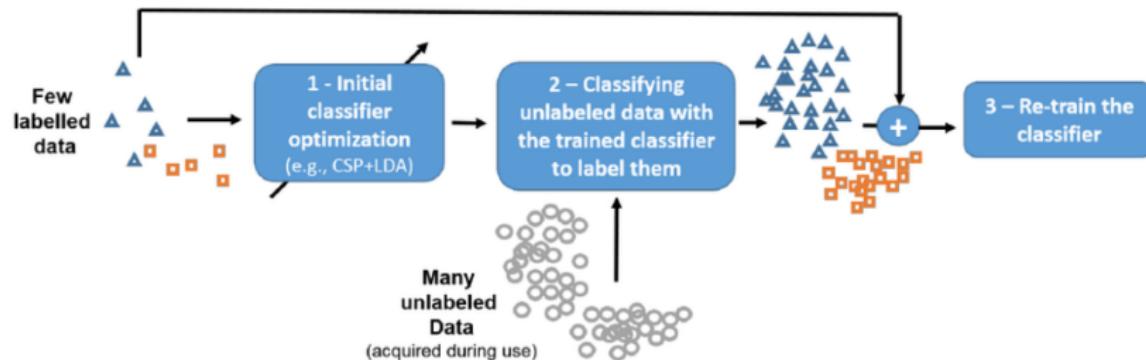
Linéaire vs Non-Linéaires vs Supervisé



Apprentissage semi-supervisé

Apprentissage semi-supervisé: Utilise un **mélange** de données **labellisées** et **non labellisées** pour l'entraînement d'un modèle.

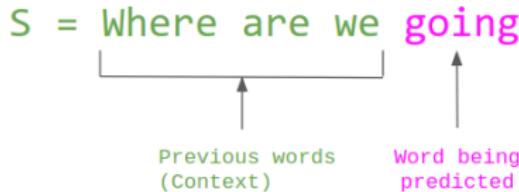
- Améliore la performance de la prédiction.
- Réduit les efforts de collecte d'étiquettes.



Apprentissage auto-supervisé

Apprentissage auto-supervisé: Utilise des données non labellisées où le modèle génère ses propres labels ou signaux d'apprentissage.

- Permet de tirer parti de grandes quantités de données non labellisées.
- Réduit le besoin de données étiquetées manuellement.



Apprentissage par renforcement

- **Applications:** Voitures autonomes, jeux vidéo
- **Objectif:** Maximiser les récompenses cumulées au fil du temps.



[Easy Introduction to Reinforcement Learning](#).

[Reinforcement Learning From Human Feedback \(RLHF\) for LLMs](#).

Apprentissage par renforcement

Apprentissage par renforcement: Le modèle apprend en interagissant avec un **environnement** et en recevant des **récompenses** ou des pénalités en fonction de ses **actions**.

Reinforcement Learning



The machine keeps on taking actions based on each reward it gets. This process keeps on iterating until a desired level of learning is not reached.

[Easy Introduction to Reinforcement Learning.](#)

Reinforcement Learning From Human Feedback (RLHF) for LLMs.

Autres types de problèmes et d'applications

- Détection d'anomalies
 - Détection de fraudes dans les transactions financières
- Recommandation
 - Systèmes de recommandation pour les produits en ligne
- Prévision (Forecasting)
 - Prévision de la demande client
 - CompléTION automatique de texte
- Génération de données (texte, audio, image)
 - Traduction automatique de texte

[Recommendation System.](#)

[Time Series Analysis and Forecasting: Examples, Approaches, and Tools.](#)

[Obama-RNN — Machine generated political speeches.](#)

[Harry Potter: Written by Artificial Intelligence.](#)

Quizz

Définition du problème

Comprendre comment aborder un nouveau problème est une compétence fondamentale. Voici une liste de quatre problèmes différents. Vous pouvez supposer que vous avez accès à autant de données que vous le souhaitez.

Quelles applications pourraient être formulées comme un problème d'apprentissage supervisé ?

- Une application de publicité en ligne qui, étant donné une publicité spécifique et les informations personnelles d'un utilisateur, détermine si l'utilisateur cliquera sur la publicité.
- Un système de conduite autonome qui, étant donné une image de la caméra avant et les informations du radar, détermine la position des autres voitures.
- Une application d'inspection visuelle qui, étant donné la photo d'une carte de circuit imprimé, détermine s'il y a des défauts.
- Une application de messagerie électronique qui, étant donné un nouveau message, détermine s'il s'agit de spam.

Quizz

Modèle de Classification

Noa est une data scientist récemment diplômée qui travaille pour une école. Elle a pour mission de développer un modèle d'apprentissage automatique pour prédire dans quelle université les élèves souhaiteront postuler à la fin de l'année. Noa a accès à toutes les notes de chaque élève précédent, y compris les étiquettes indiquant l'université qu'ils ont choisie. Elle a plusieurs options pour construire un modèle de classification.

Parmi les approches suivantes, laquelle devrait être la meilleure pour construire ce modèle ?

- Noa devrait utiliser un arbre de décision, une technique d'apprentissage supervisé.
- Noa devrait utiliser la régression linéaire, une technique d'apprentissage supervisé.
- Noa devrait utiliser l'apprentissage par renforcement.
- Noa devrait utiliser l'apprentissage non supervisé.

Quizz

Manque de données étiquetées

Milani travaille dans une agence de marketing spécialisée dans les campagnes sur les réseaux sociaux. L'agence a collecté une vaste quantité de données textuelles provenant de divers réseaux sociaux mais manque d'étiquettes. Pour optimiser les futures campagnes, Milani veut classifier le sentiment de chaque échantillon de texte. Cependant, elle sait qu'elle ne peut pas procéder à un apprentissage supervisé sans données étiquetées. Milani doit décider d'une technique pour obtenir des étiquettes pour les données.

Quelle technique pourrait être utilisée par Milani pour étiqueter les données?

- Embaucher une équipe pour examiner et étiqueter chaque échantillon de texte manuellement.
- Utiliser les réactions ou commentaires des utilisateurs sur les échantillons de texte pour générer automatiquement des étiquettes.
- Appliquer une méthode d'apprentissage supervisé pour déduire directement les étiquettes à partir des données existantes.
- Mettre en œuvre un apprentissage semi-supervisé pour répartir les étiquettes sur l'ensemble du jeu de données.

Quizz

Apprentissage auto-supervisé

Makayla est une data scientist expérimentée qui a travaillé avec des modèles d'apprentissage supervisé et non supervisé pendant des années. Récemment, elle a été introduite au concept d'apprentissage auto-supervisé.

Lesquelles des affirmations suivantes concernant l'apprentissage auto-supervisé sont vraies ?

- L'apprentissage auto-supervisé n'est qu'un terme sophistiqué pour l'apprentissage non supervisé.
- Comme l'apprentissage supervisé, les méthodes auto-supervisées peuvent juger si leur prédiction est correcte pendant l'entraînement.
- Les méthodes auto-supervisées ne nécessitent qu'un petit nombre d'échantillons étiquetés.
- Comme l'apprentissage non supervisé, les méthodes auto-supervisées n'ont pas besoin de données d'entraînement étiquetées.

Conclusion

- Introduction à la Data Science
- Relation avec l'IA et le ML
- Types d'apprentissage, de problèmes et d'applications abordés
- Exemples de modèles par type de problème

Prochains Travaux Pratiques :

- Régression Linéaire pour la prédition des prix des maisons
- Clustering de cellules cancéreuses

À venir dans le cours :

- Pipeline de l'apprentissage automatique : Des données au modèle