INTRODUCTION AU MACHINE LEARNING

ESGI – 2024/2025

Duchelle KEPSU – duchellekepsu@yahoo.com

école supérieure de génie informatique

DEROULEMENT DU COURS

Séance 1: Les fondamentaux du Machine Learning et préparation des données (3h) - *25/11/2024*

- Introduction au Machine Learning (ML)
- Préparation des données
- TP: Concepts clés du ML et préparation des données

Séance 2: Les modèles d'apprentissages supervisés (3h) - *28/11/2024*

- Les modèles de Régression
- Les modèles de classification
- TP: Implémentation des modèles de régression et de classification sur un jeu de données

Séance 3: Les modèles d'apprentissages non - supervisés (3h) - *16/12/2024*

- Les modèles de Clustering
- Réduction de dimensionnalité
- TP: réduction de dimensionnalité et implémentation des modèles de Clustering

ESGI
école supérieure de génie informatique

DEROULEMENT DU COURS

Séance 4: Evaluation et amélioration des modèles de ML (3h) - *13/01/2024*

- Evaluation des modèles: mesures de performances
- Amélioration des modèles
- TP: Evaluation des modèles de ML implémentés et amélioration

Séance 5: Impact sociétal du Machine Learning, Application dans le milieu professionnel et préparation à l'examen (3h) - *03/02/2024*

- Application du Machine Learning dans le milieu professionnel et REX
- Impacts sociétaux et réflexions éthiques sur le ML
- Tips pour l'examen

Séance 1: Fondamentaux du Machine Learning et préparation des données

Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

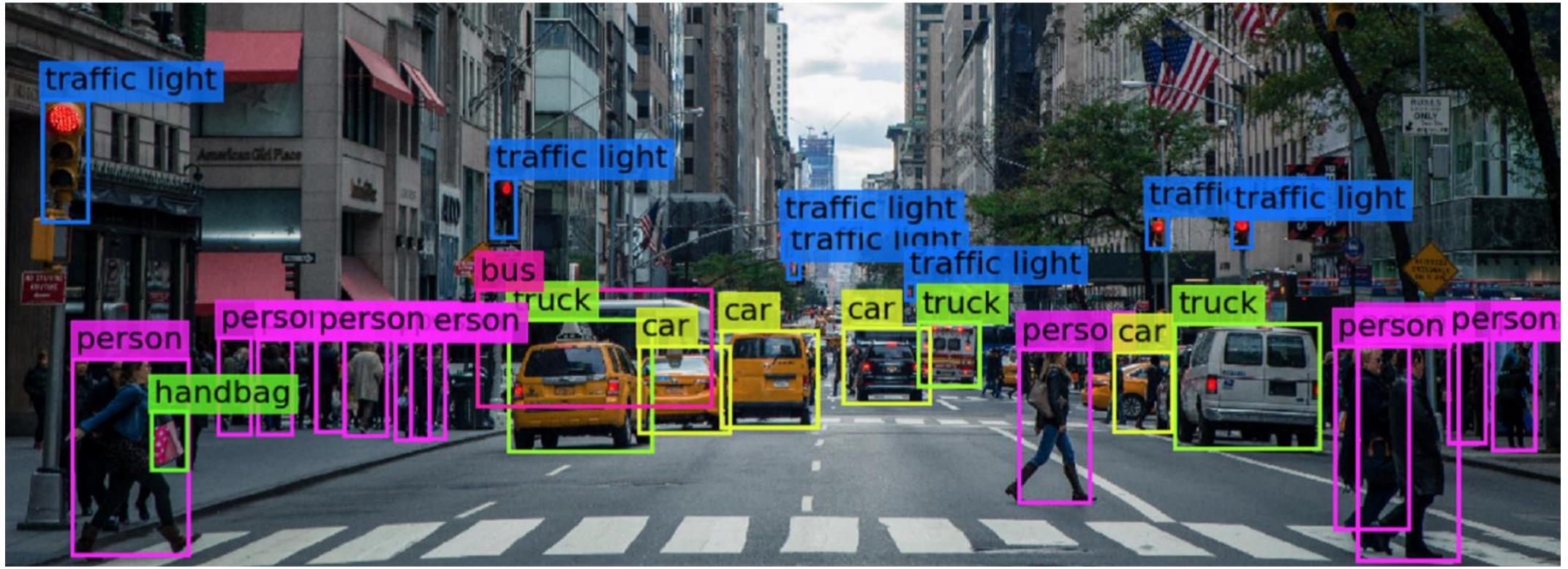
Le machine learning est partout:

- Vision par ordinateur (Computer Vision)
- Robotique
- Véhicule autonome
- Traitement du langage Naturel (NLP) , reconnaissance vocale
- Moteur de recherche (Google, Qwant etc…)
- Système de recommandation
- Diagnostic médical
- Etc.

Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

Computer Vision

- Reconnaissance d’objet | labélisation



Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

Robotique

- Automatisation des tâches

Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

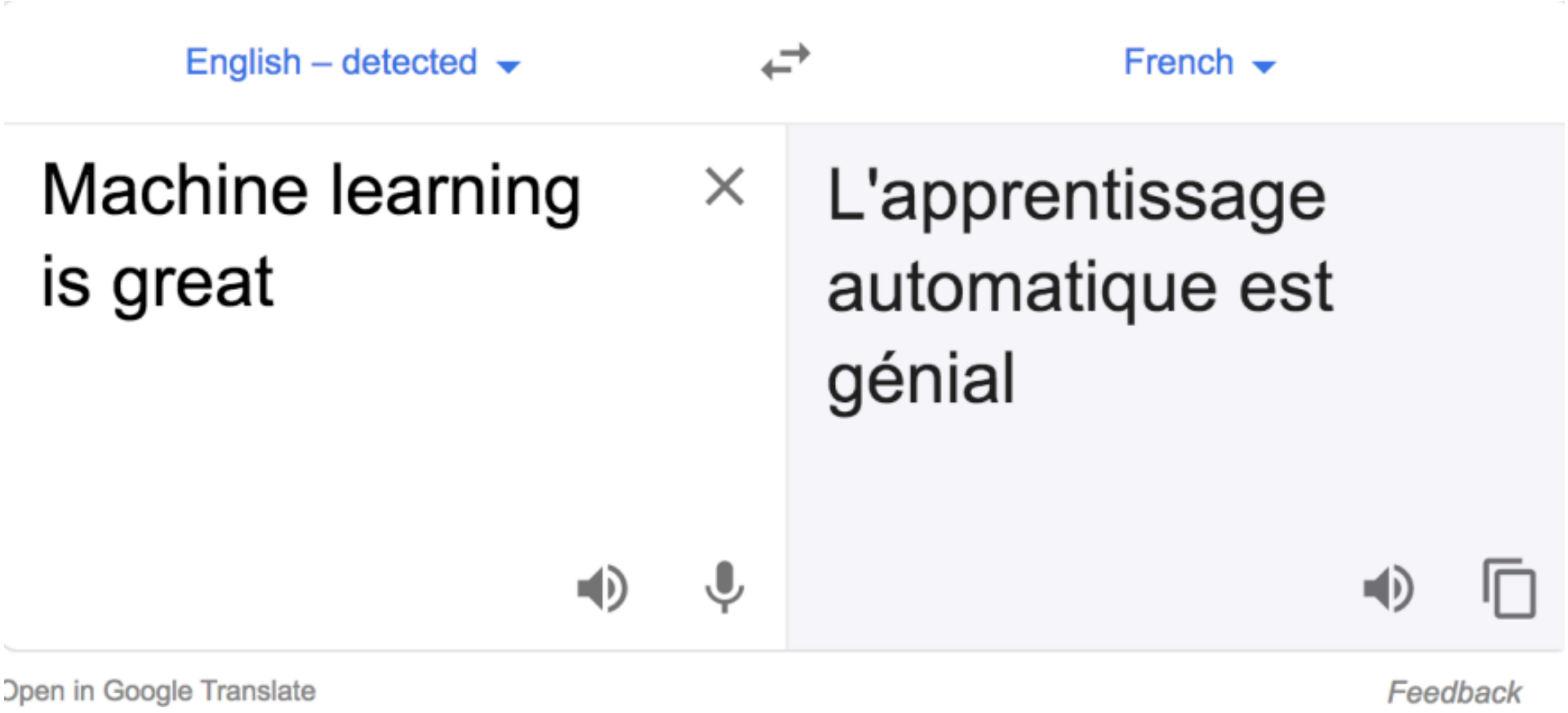
Véhicules autonomes

- Conduite autonome

Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

NLP

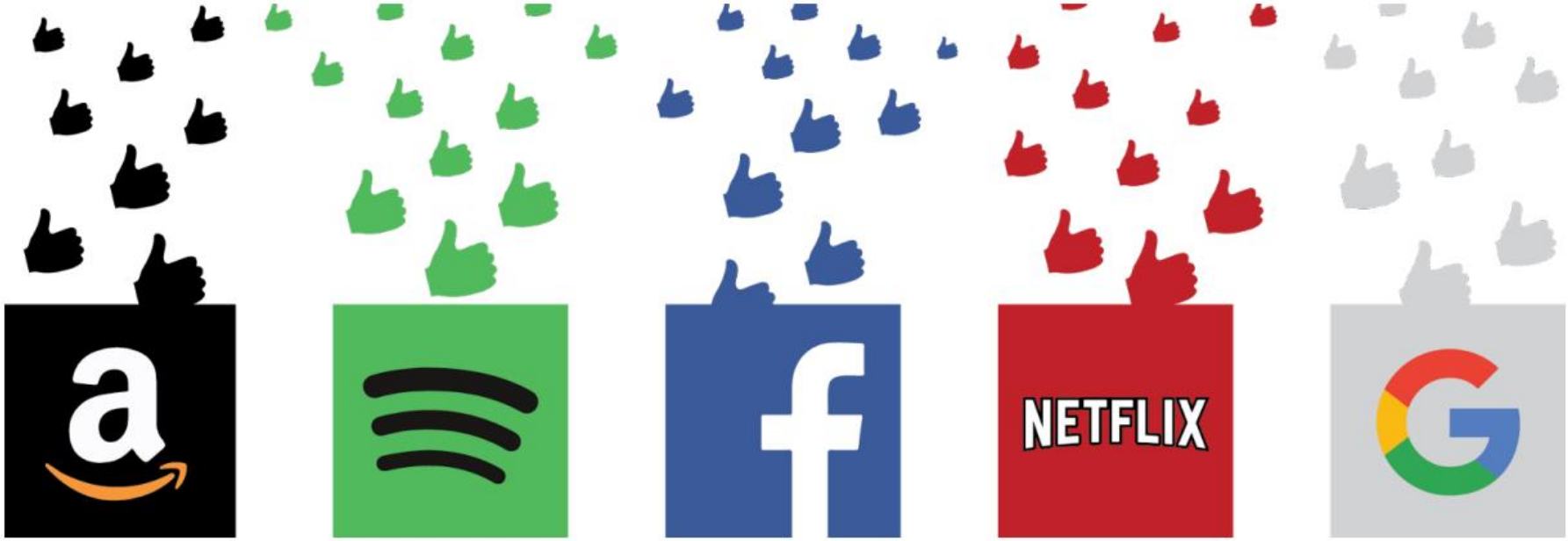
- Traduction automatique



Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

Système de recommandation

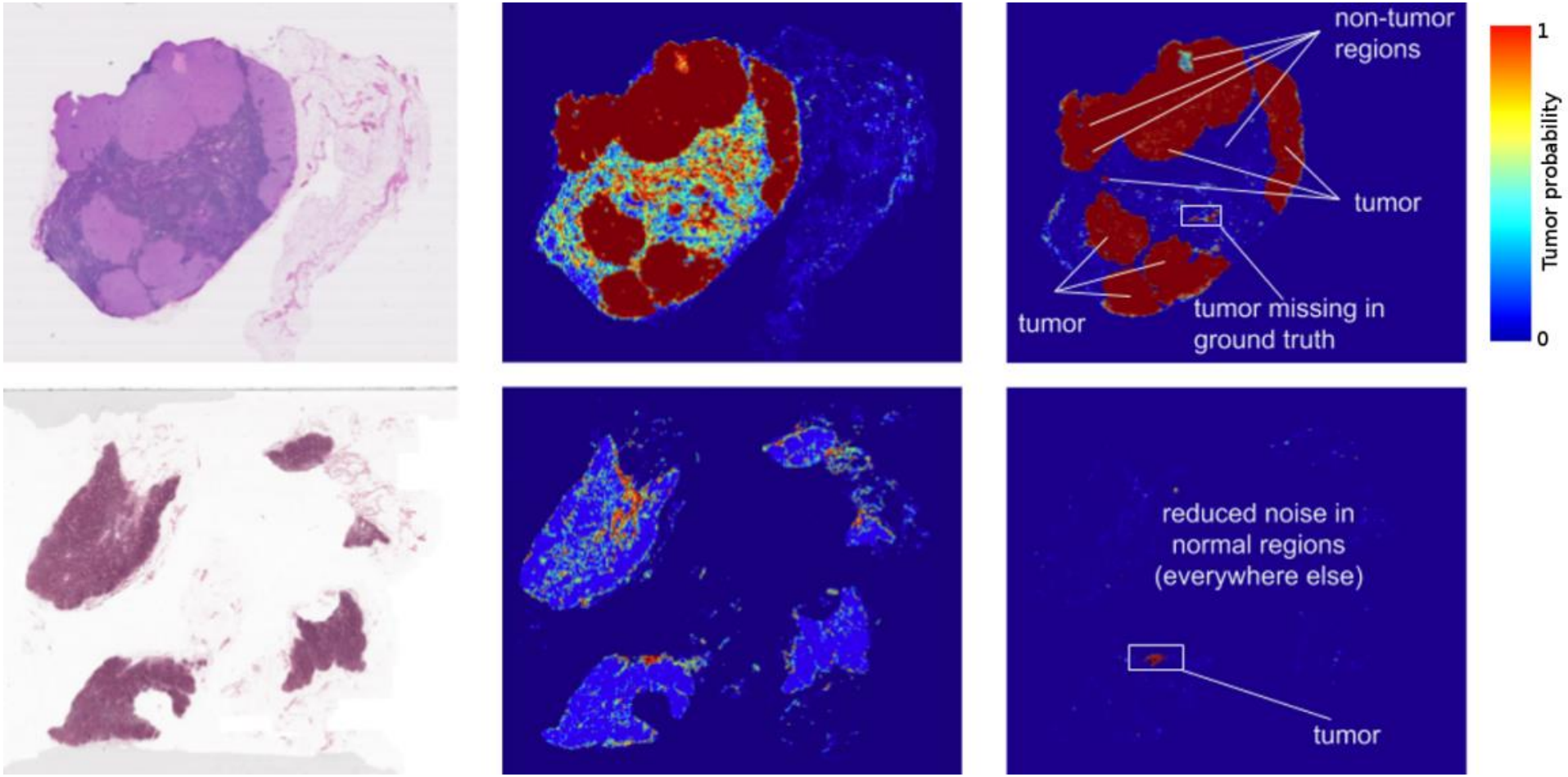
- Optimisation des systèmes de recommandation



Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

Diagnostic médical

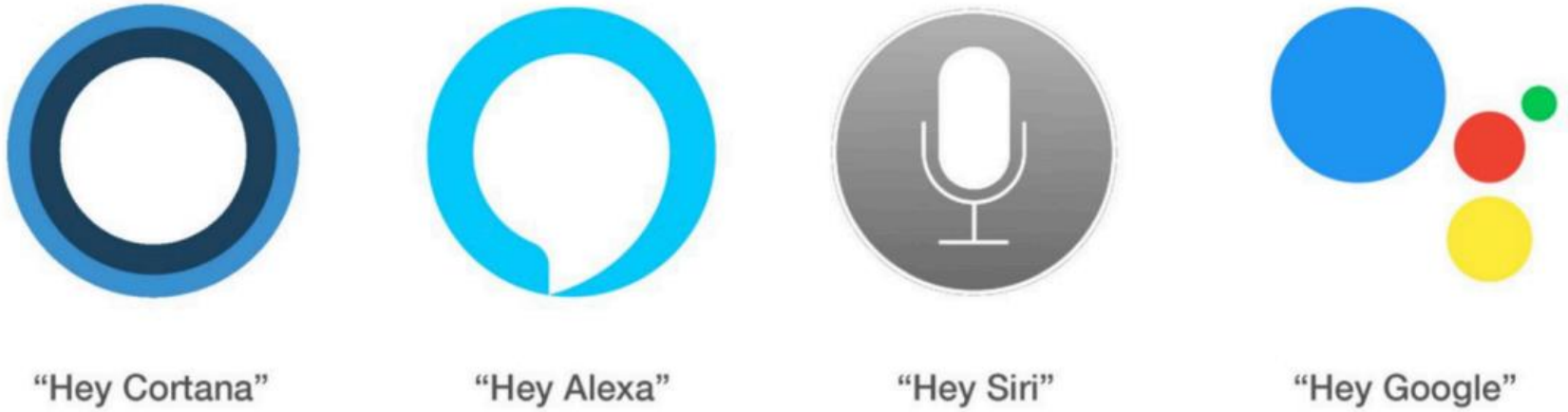
- Détection de cancer



Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

reconnaissance vocale

- Assistants vocaux



Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

Détection de fraude

- Opérations bancaires frauduleux

Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

I.I- Qu'est-ce que le Machine Learning ?

- « **Apprentissage machine** » qu'est-ce que apprendre ? Comment apprend-on ? Et qu'est-ce que cela signifie pour une machine ?

- **Fabien Benureau** : « L'apprentissage est une modification d'un comportement sur la base d'une expérience »

- Et dans le cadre d'un programme informatique, qui est celui qui nous intéresse dans ce cours, on parle d'**apprentissage automatique ou Machine Learning:** c'est une branche de l'IA qui permet à des systèmes informatiques d'apprendre à partir des données, sans être explicitement programmés

Ingrédients du Machine Learning

Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

I.I- Qu'est-ce que le Machine Learning ?

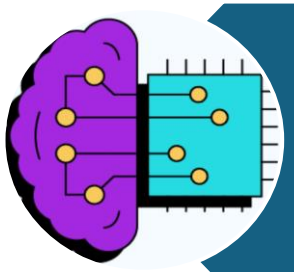
Le machine Learning repose sur:

- ✓ **Les mathématiques**, en particulier les **statistiques**: pour la construction des modèles
- ✓ **L'informatique:** pour la représentation des données et l'implémentation efficace d'algorithme

NB:

- Le ML est utilisé quand les données sont abondantes
- Les deux piliers du ML vus précédemment sont importantes l'un de l'autre:

Aucun algorithme d'apprentissage ne pourra créer un bon modèle à partir des données qui ne sont pas pertinentes



D'autre part, un modèle appris avec un algorithme inadapté sur des données pertinentes ne pourra pas être de bonne qualité

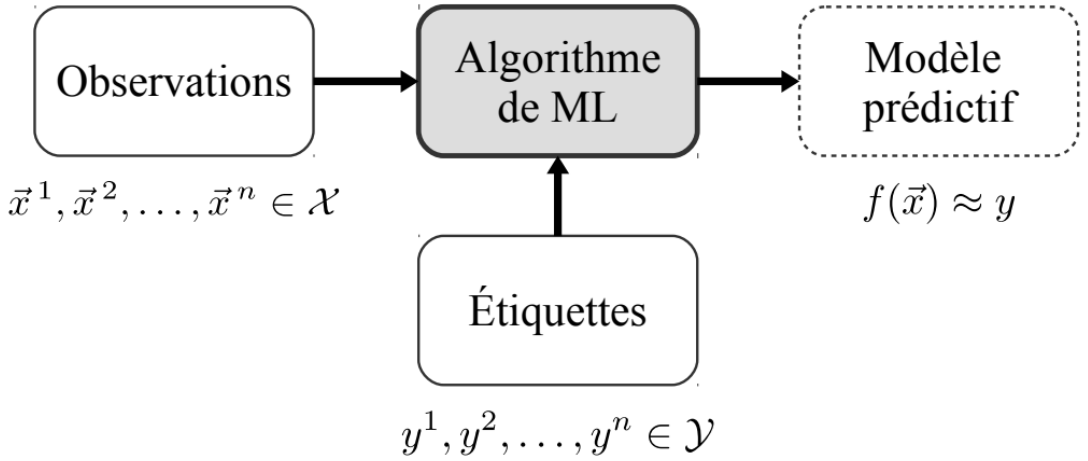
Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

I.2- Les différents types de modèle de Machine Learning

i. Les modèles de Machine Learning supervisés ii. Les modèles de Machine Learning non-supervisés

I.2.i- Apprentissage supervisé:

- Cette forme d'apprentissage consiste à faire des prédictions à partir d'une liste d'exemples étiquetés
- L'algorithme reçoit un ensemble de données d'entraînement étiquetées (x1,y1), (x2,y2), … , (xN,yN), c'est-à-dire des observations (entrée, sortie)



Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

I.2- Les différents types de modèle de Machine Learning

I.2.i- Apprentissage supervisé:

- **Classification binaire:** Les étiquettes Y prennent des valeurs binaires, dans un ensemble fini et non ordonné. Elles indiquent l'appartenance à une classe.

  Exemples:

  - Identifier si un mail est un spam ou non
  - Identifier si une transaction financière est frauduleuse ou non
  - Détecter si un animal est un chien ou un chat etc.

- **Classification multi-classe:** Les étiquettes Y prennent des valeurs discrètes, finies et correspondent à plusieurs classes

  Exemples:

  - Identifier les objets présents sur une photographie
  - Identifier à quelle espèce appartient une plante
  - Identifier en quelle langue un texte est écrit

Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

I.2- Les différents types de modèle de Machine Learning

I.2.i- Apprentissage supervisé:

- **Régression:** Les étiquettes Y prennent des valeurs réelles

  - L'idée de la régression linéaire est simplement de trouver une ligne qui s'adapte le mieux aux données. Les extensions de la régression linéaire comprennent la régression linéaire multiple (trouver un plan qui s'ajuste le mieux) et la régression polynomiale (trouver une courbe qui s'ajuste le mieux)

  - Les arbres de décision font également partis des modèles de régression: la valeur à prédire est une valeur réelle

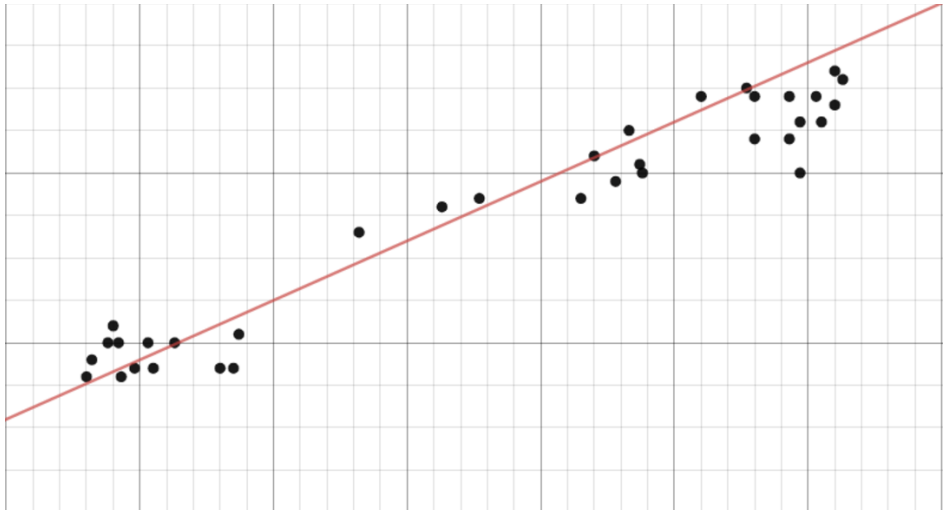


Figure: Régression linéaire

Exemples:

- ❖ Prédire le prix d'un appartement à louer
- ❖ Prédire le nombre d'utilisateurs d'un service en ligne
- ❖ Prédire le chiffre d'affaires qu'une entreprise réalisera le prochain mois

I.2.ii- Apprentissage non- supervisé:

- Les modèles d'apprentissage non supervisés apprennent à partir d'ensembles de données sans étiquettes (données de sortie)

- L'objectif est de trouver des modèles non détectés dans les données, pour mieux les comprendre

  - ✓ Trouver des groupes d'échantillons qui présentent une similitude dans un certain sens
  - ✓ Rechercher des sous-ensembles de fonctionnalités qui se comportent de la même manière
  - ✓ Trouver des combinaisons de fonctionnalités avec le plus de variations (ex: reduction de dimension)

- L'algorithme reçoit en entrées des données (x1,y1), (x2,y2), … , (xN,yN), et apprend une fonction sur ces donn

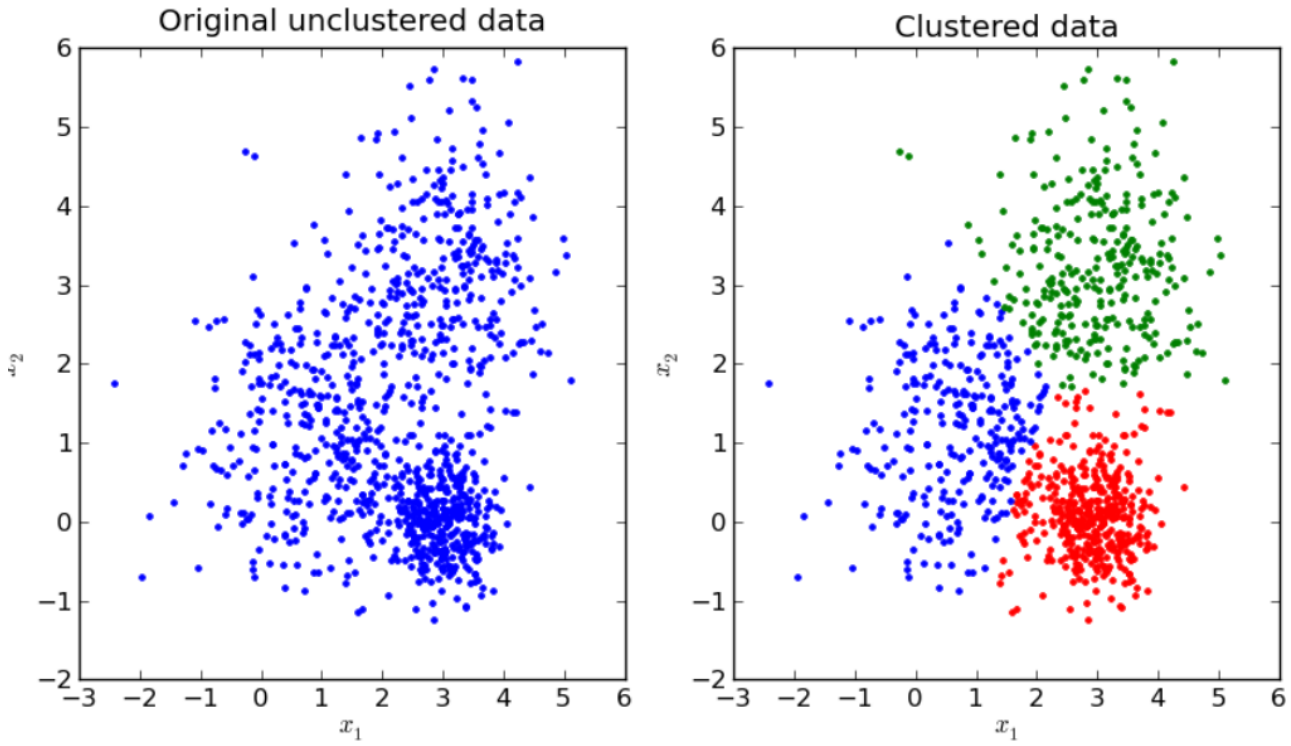
Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

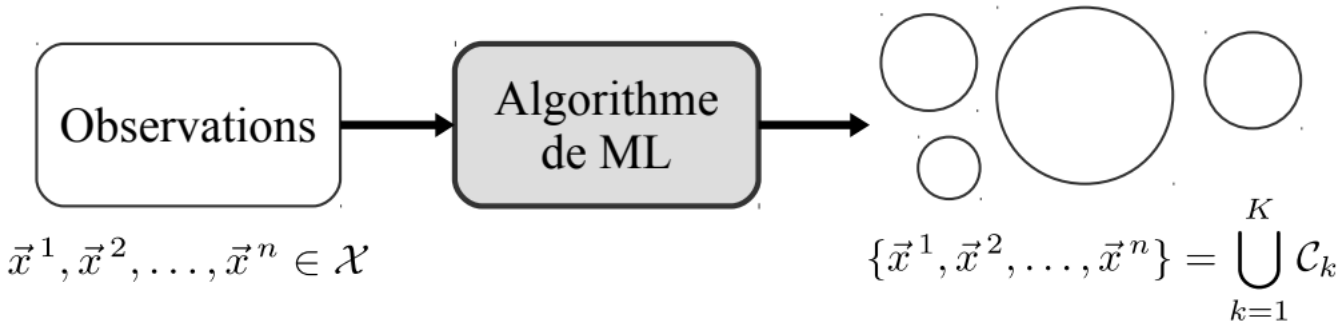
I.2.ii- Apprentissage non- supervisé:

- **Le clustering ou partitionnement:**

  - Problème d'apprentissage non-supervisé pouvant être formalisé comme la recherche de partitions pertinentes dans n observation

  - Identifier les groupes dans les données





Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

I.2.ii- Apprentissage non- supervisé:

- **Le clustering ou partitionnement:**

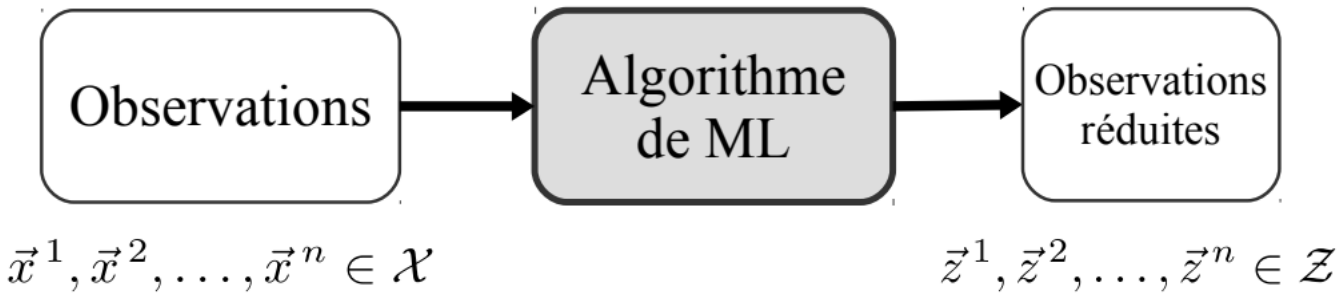
Exemples:

- ❖ La segmentation de marché consiste à identifier des groupes d'usagers ou de clients ayant un comportement similaire. Cela permet de mieux comprendre leur profil, et cibler une campagne de publicité, des contenus ou des actions spécifiquement vers certains groupes.

- ❖ Identifier des groupes parmi les patients présentant les mêmes symptômes permet d'identifier des sous-types d'une maladie, qui pourront être traités différemment.

- **Réduction de dimension:**

Problème d'apprentissage non supervisé pouvant être formalisé comme la recherche d'un espace Z de dimension plus faible que l'espace X dans lequel sont représentées n observations.



Partie I: LES FONDAMENTAUX DU MACHINE LEARNING

I.2.ii- Apprentissage non- supervisé:

- **Réduction de dimension:**

Réduction du temps de calcul

Réduction de l'espace mémoire nécessaire au stockage de données

Amélioration des performances d'un algorithme d'apprentissage supervisé entrainé par la suite sur ces données

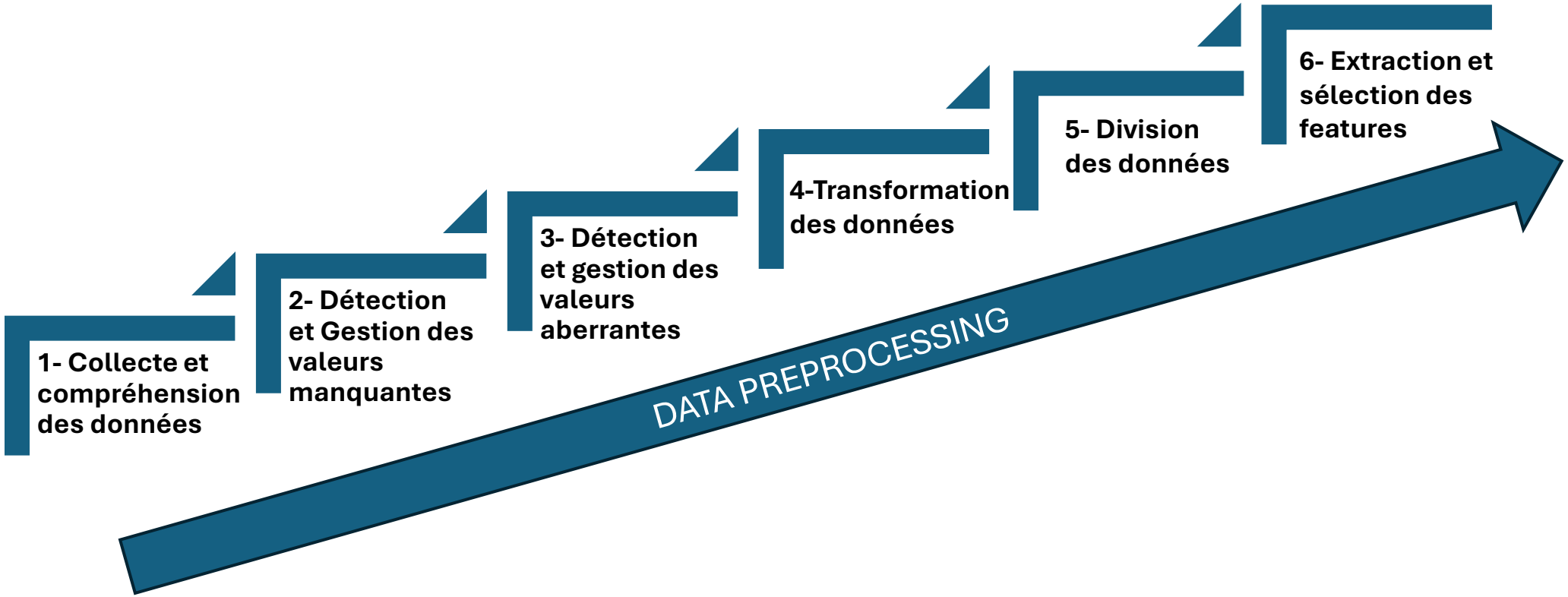
NB: Il existe plusieurs techniques de réduction de dimension:
Nous les verrons dans le chapitre dédié à cela

Partie II: PREPARATION DES DONNES

- Avant la mise en place des modèles de prédiction, une étape importante est menée: le « **Data Preprocessing** » encore appelé « pré-traitement des données » ou Feature Engineering

C’est une étape très importante dans le pipeline de Machine Learning, elle permet au modèle de produire des prédictions de meilleure qualité

- Le Data preprocessing est constitué de plusieurs étapes, listées ci-dessous et que nous allons étudier dans cette partie du cours



Partie II: PREPARATION DES DONNES

I. Collecte et compréhension des données

I.1 – Chargement des données:

Sources courantes:

- Fichiers CSV
- Bases de données SQL
- Les API
- Fichiers Excel
- Données JSON

Exemple: lire un fichier CSV en python

```python
import pandas as pd
data = pd.read_csv('dataset.csv')  # Lire un fichier CSV
print(data.head())  # Aperçu des premières lignes
```

I.2- Compréhension des données

Types de données:

- Variables numériques (continues, discrètes)
- Variables catégorielles (qualitatives)
- Variables textuelles (documents, commentaires)

Exemple: Code python pour vérifier le type de chaque colonne et avoir les statistiques descriptives de chaque colonne (moyenne, écart-type, min, max etc…)

```python
print(data.info())  # Vérifier le type de chaque colonne
```

Outils pour l'analyse:

- Types et format de colonnes
- Statistiques descriptives
- Détection d'anomalies

```python
print(data.describe(include='all'))  # Inclure les colonnes non numér
```

Partie II: PREPARATION DES DONNES

II. Détection et gestion des valeurs manquantes

II.1- Détection des valeurs manquantes:

```python
print(data.isnull().sum() / len(data) * 100)
```

II.2- Identifier la cause des valeurs manquantes:

- Données manquantes aléatoire: erreur de saisie ou de collecte
- Données manquantes non – aléatoires: lié à une condition (ex: absence de revenus pour une personne sans emploi)

II.2- Stratégie de traitement des valeurs manquantes:

- Supprimer la ligne ou les colonnes où il y’a trop de valeurs manquantes (généralement > 50%)

```python
data = data.dropna(axis=0)  # Supprime les lignes avec valeurs nulles
```

Partie II: PREPARATION DES DONNES

II. Détection et gestion des valeurs manquantes

II.2- Stratégie de traitement des valeurs manquantes:

- Imputation des valeurs manquantes: qui consiste à remplir les valeurs nulles avec une estimation

  - Pour les variables numériques: moyenne, médiane

Critère	Moyenne	Médiane
Robustesse aux outliers	Sensible aux outliers	Insensible aux outliers
Type de distribution	Symétrique ou normale	Asymétrique
Maintien de la structure	Représente bien les distributions équilibrées	Convient mieux aux distributions déséquilibrées

- La moyenne est utilisée lorsque les données suivent une distribution **symétrique** et sont proches d'une distribution normale et si les valeurs aberrantes (**outliers**) sont rares ou inexistantes.

- La médiane est utilisée lorsque les données contiennent des **outliers (car elles influencent fortement la moyenne mais pas la médiane)** et si les données sont asymétriques .

Partie II: PREPARATION DES DONNES

II. Détection et gestion des valeurs manquantes

II.2- Stratégie de traitement des valeurs manquantes:

- Pour les variables catégorielles:

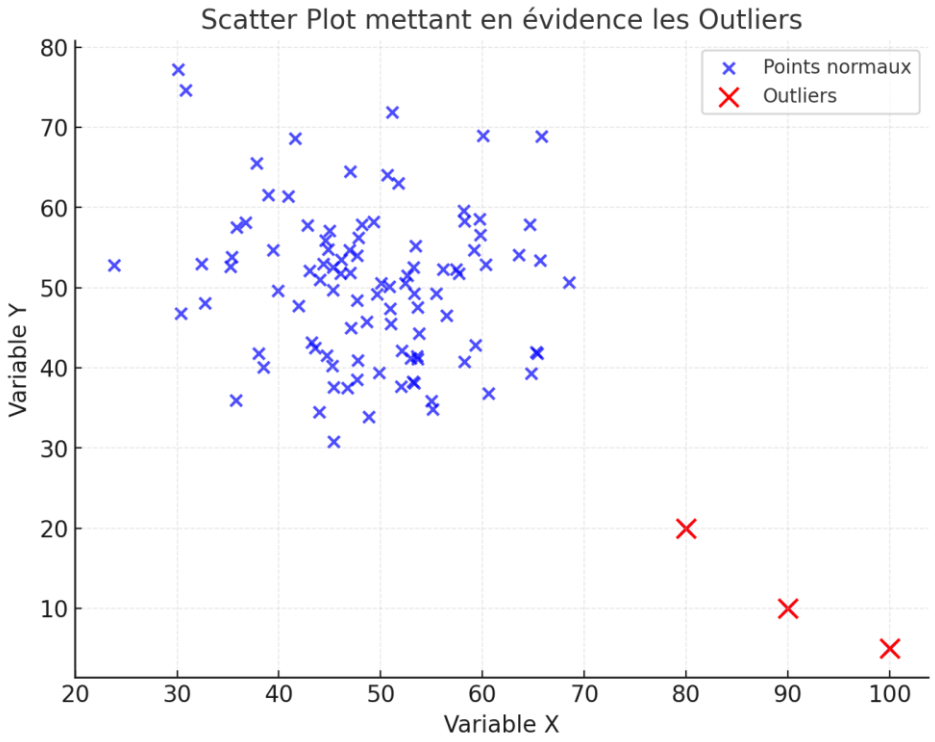
Méthode	Cas d'utilisation	Avantages	Inconvénients
Mode (valeur la plus fréquente)	Peu de catégories, distribution équilibrée	Simple, rapide	Peut biaiser les données si une catégorie domine
Catégorie "Missing" ou "Unknown"	Les données manquantes ont une signification particulière	Préserve la structure des données	Ajoute une nouvelle catégorie
Conditionnelle (groupée)	Forte corrélation entre colonnes	Approche contextuelle	Analyse préalable nécessaire
Par modèles de prédiction	Dataset large, relations complexes entre colonnes	Imputation précise, robuste	Consommation en temps et ressources

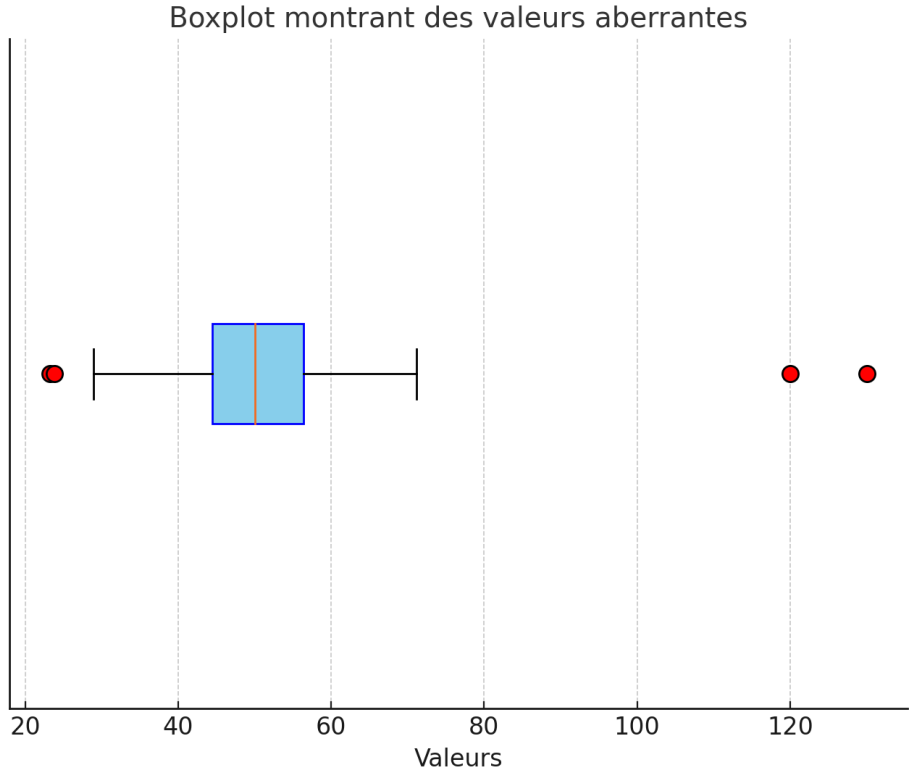
Partie II: PREPARATION DES DONNES

II. Détection et gestion des valeurs abbérantes (outliers)

II.1 - Détection des outliers

- visualisation des données: utilisés le plus souvent pour des variables numériques





- Scatter plot: relation entre deux variables
- Boxplots:

Partie II: PREPARATION DES DONNES

III. Détection et gestion des valeurs aberrantes (outliers)

II.1 - Traitement des outliers

- Suppression directe (si justifié)
- Transformation logarithmique des valeurs: utiliser sur des données très asymétriques et est utile pour compresser de grandes valeurs

IV. Transformation des données

IV.i- Mise à l'échelle:

- Standardisation: Centre les données à 0 avec un écart type de 1

```python
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
```

Généralement utilisé quand les variables ont des échelles différentes et pour des modèles tels que les modèles de régression linéaire qui améliore leur stabilité et leur performance

Partie II: PREPARATION DES DONNES

IV. Transformation des données

IV.i- Mise à l'échelle:

- Normalisation: Met les données dans une plage spécifiques (généralement de 0 à 1)

Généralement utilisé pour des algorithmes qui utilisent les distances: **Ex**: K-Means, K-Nearest Neighbors où les distances entre les points jouent un rôle essentiel dans le calcul de la prédiction

IV.ii- Encodage des variables catégoriques:

- One hot encoding:

Le **One-Hot Encoding** est une technique utilisée pour convertir des variables catégorielles en variables numériques, afin que celles-ci puissent être utilisées par des algorithmes de machine learning ou d'autres modèles qui nécessitent des données numériques en entrée Il transforme chaque valeur d'une variable catégorielle en colonne binaire

Ex

- ❖ Imaginons que nous avons une variable catégorielle "Couleur" avec trois catégories : Rouge, Vert, et Bleu.

Partie II: PREPARATION DES DONNES

IV. Transformation des données

IV.ii- Encodage des variables catégorielles:

Le One-Hot Encoding va transformer cette colonne en trois nouvelles colonnes, une pour chaque catégorie. Chaque ligne sera représentée par un 1 dans la colonne correspondant à sa couleur, et des 0 dans les autres colonnes.

Rouge	Vert	Bleu
1	0	0
0	1	0
0	0	1
0	1	0
0	0	1

- Le label encoding:

Le **Label Encoding** est une technique de prétraitement des données utilisée pour convertir des variables catégorielles en valeurs numériques. Contrairement au One-Hot Encoding qui crée de nouvelles colonnes pour chaque catégorie, le Label Encoding attribue une valeur numérique unique à chaque catégorie de la variable.

Partie II: PREPARATION DES DONNES

IV. Transformation des données

IV.ii- Encodage des variables catégorielles:

Exemple:

Si une variable catégorielle "Couleur" a les catégories "Rouge", "Vert" et "Bleu", le Label Encoding pourrait les transformer en :

- Rouge -> 0
- Vert -> 1
- Bleu -> 2

Le Label Encoding est souvent utilisé lorsque vous avez des variables catégorielles avec une certaine hiérarchie ou un ordre naturel entre les catégories, comme par exemple :

- Taille : "Petit", "Moyenne", "Grand" (avec un ordre implicite)
- Niveau d'éducation : "Baccalauréat", "Licence", "Master", "Doctorat" (également ordonné)

IV. Division des données (Train/Test Split)

Il s'agit ici de diviser les données en jeu d'entrainement et de test. Cette opération sert à évaluer la capacité d'un modèle à généraliser sur de nouvelles données

Partie II: PREPARATION DES DONNES

V. Division des données (Train/Test Split)

1. Jeu d'entraînement (Training set) : C'est la portion des données qui est utilisée pour entraîner le modèle. Le modèle apprend les relations et les patterns entre les variables d'entrée et les résultats à partir de cet ensemble.

2. Jeu de test (Test set) : C'est la portion des données qui est réservée pour évaluer la performance du modèle après l'entraînement. Le but de cet ensemble est de simuler de nouvelles données (inconnues du modèle) pour tester sa capacité à généraliser.

Les proportions typiques pour diviser les données sont :

- 80% pour l'entraînement et 20% pour le test : Cette division est très courante, surtout si vous disposez de beaucoup de données.

- 70% pour l'entraînement et 30% pour le test : Cela peut être utilisé si vous avez une quantité moyenne de données.

- 90% pour l'entraînement et 10% pour le test : Ce cas est souvent utilisé lorsque les données sont rares et que vous voulez maximiser la quantité d'exemples d'entraînement.

Partie II: PREPARATION DES DONNES

Vi. Extraction et sélection des caractéristiques (features):

Ici il s'agit de:

- D'éliminer les colonnes non pertinentes (les colonnes redondantes ou non – informatives)
- Garder les features utiles pour la prédiction

NB: il y'a des modèles comme les arbres de décision, les modèles de régression qui fournissent les informations sur l'importance des caractéristiques

PARTIE III: Mise en pratique des étapes de Data Preprocessing sur des données réelles (cas des données titanic)