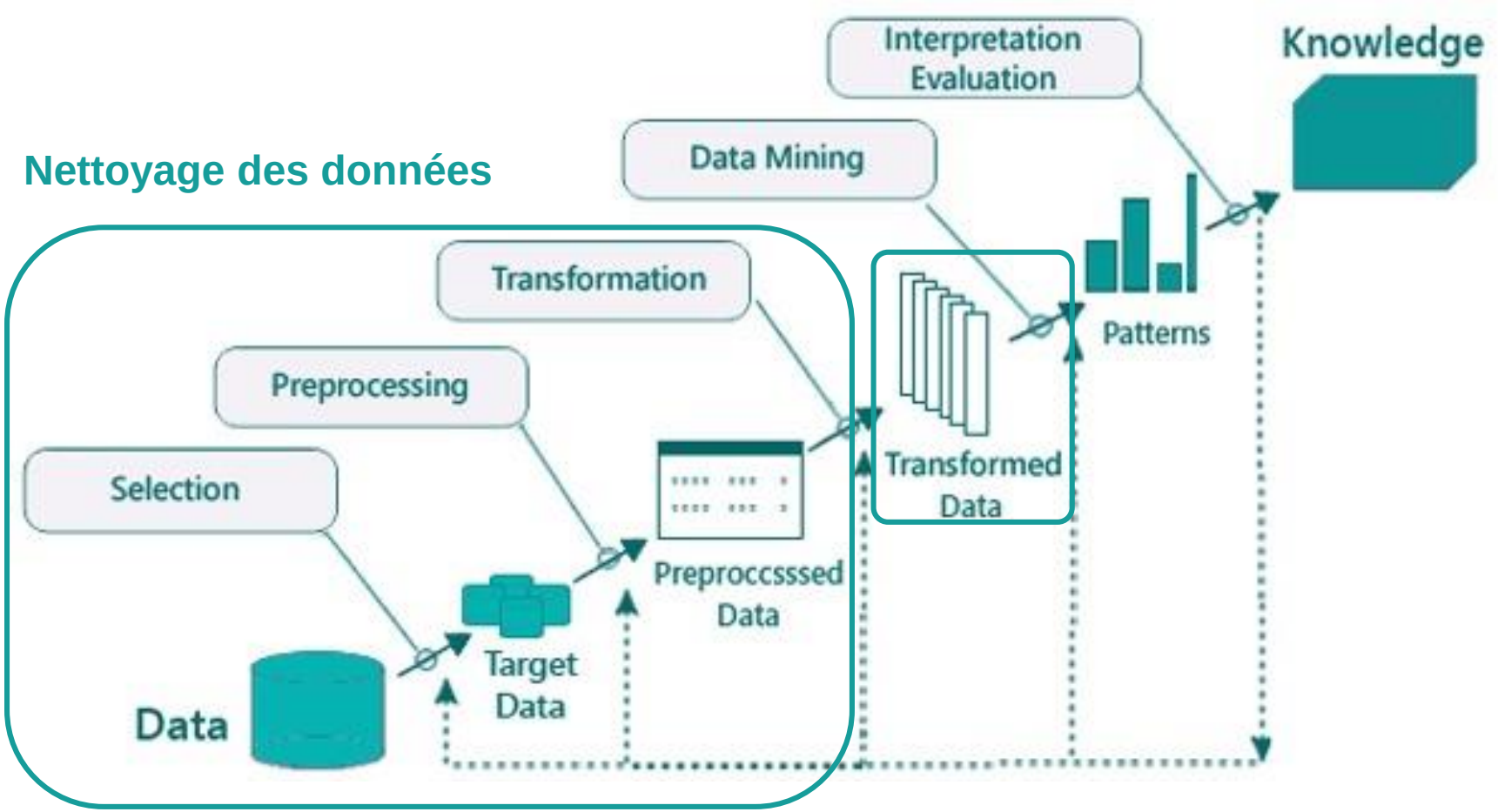
KDD Process Data Mining

Elise Maistre - 24 mars 2025

KDD Process

"From Data Mining to Knowledge Discovery in Databases"



data cleaning, data sampling, dimensionality reduction, data mining algorithms

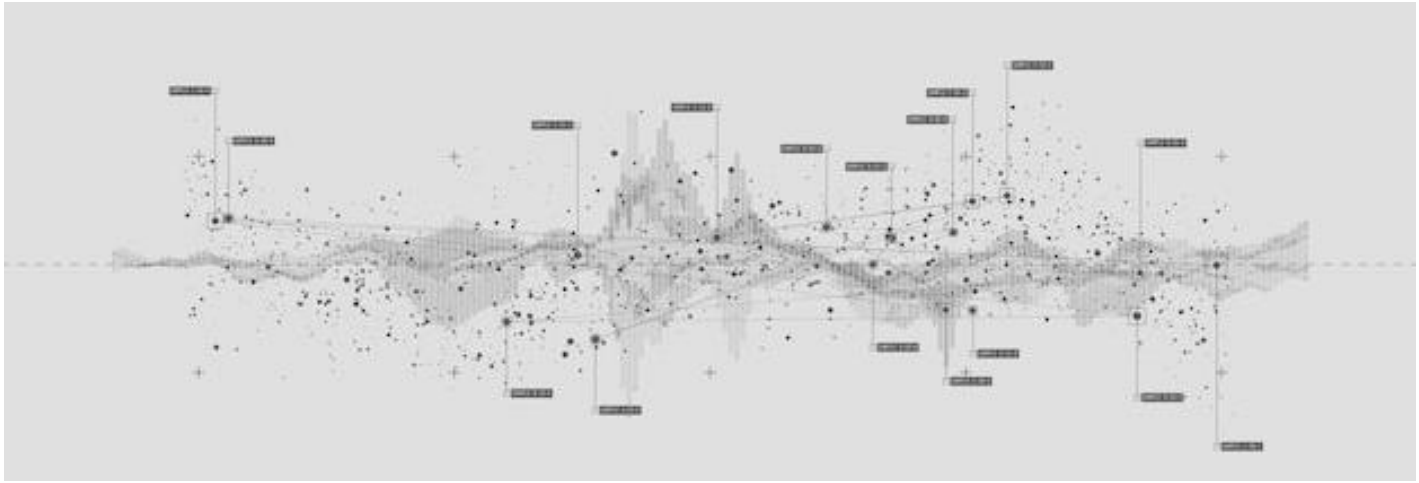
Pourquoi KDD et pas seulement Data Mining ?
Qu'est ce que le KDD ?

Processus entier de découverte de connaissance :

- Apprentissage du domaine (contexte)
- Trouver un dataset et sélectionner les données à utiliser
- Nettoyage des données (environ 60% du travail)
- Réduction des données et transformation (réduction de la dimension, conserver les caractéristiques intéressantes)
- Choix de la méthode de data mining : classification, régression, clustering, association, …
- Choix de l'algorithme de data mining et des hyperparamètres (optimisation : fine tuning)
- Evaluation des résultats
- Déploiement : représentation (visualisation) et utilisation de l'apprentissage

Objectifs du KDD

- Description : Que s'est-il passé ? *(Analyser les données passées pour comprendre les tendances et les comportements)*
  - Ex : "Quels produits ont été les plus vendus le mois dernier ?" "Quelle est la répartition des clients par région ?"
- Diagnostic : Pourquoi cela s'est-il produit ? *(Trouver les causes des tendances observées dans l'analyse descriptive)*
  - Ex : "Pourquoi les ventes ont-elles chuté le mois dernier ?" "Pourquoi certains clients quittent notre plateforme sans acheter ?"
- **Prédiction** : Que pourrait-il se passer ? *(Utiliser les données passées pour prédire des événements futurs)*
  - Ex : "Quel sera le chiffre d’affaires du mois prochain ?" "Quels clients sont susceptibles d’acheter un produit dans les 30 prochains jours ?"



Coding game : Kaggle Titanic

Titanic - Machine Learning from Disaster

Sur Kaggle :
+ New notebook

```
train=pd.read_csv("/kaggle/input/titanic/train.csv")
test=pd.read_csv("/kaggle/input/titanic/test.csv")
```

kaggle

Apprentissage du domaine

First Class Ticket Sceond Class Ticket Third Class Ticket

Contexte Historique du Naufrage du Titanic

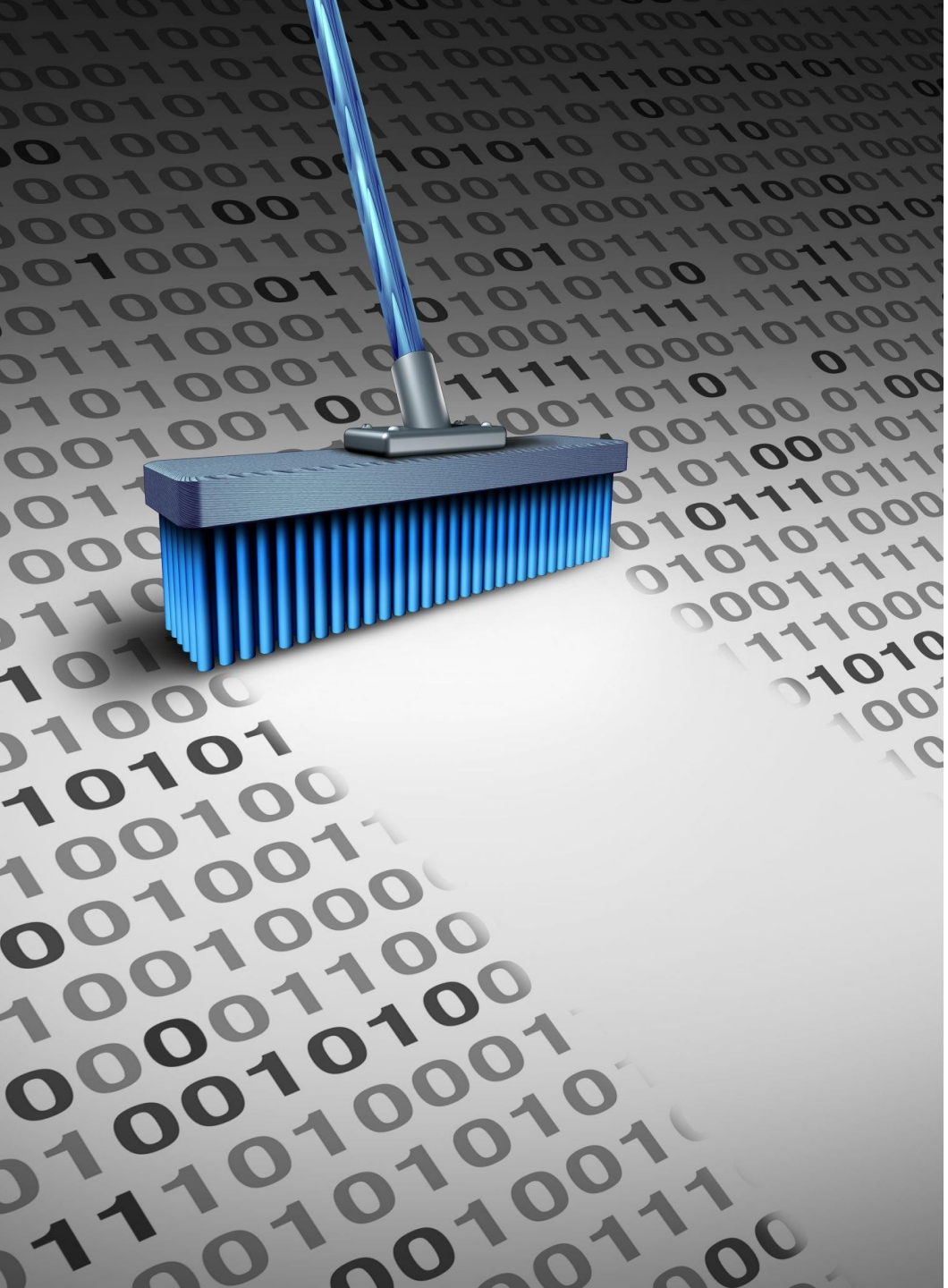
Le Titanic a coulé dans la nuit du 14 au 15 avril 1912 après avoir percuté un iceberg dans l'Atlantique Nord. Sur environ 2 224 personnes à bord, seules 710 ont survécu. Le naufrage a duré environ 2 heures et 40 minutes, avec un nombre de canots de sauvetage insuffisant pour tous les passagers.

Facteurs influençant la survie

• Classe sociale : Les passagers de première classe ont eu plus de chances de survivre que ceux des classes inférieures.
• Genre : La règle "les femmes et les enfants d'abord" a favorisé leur survie, tandis que les hommes ont été majoritairement victimes du naufrage.
• Age : Les enfants ont eu un taux de survie plus élevé que les adultes et les personnes âgées.
• Emplacement sur le navire : Les passagers situés près des canots de sauvetage avaient un avantage décisif.

Sélectionner le dataset et les données

- Possible d’utiliser plusieurs datasets, d’ajouter des données externes
- Dans notre cas : Dataset fourni par kaggle
- Objectif : prédiction, classification binaire selon la survie (Survived)
- Dataset :
  - Train : 891 lignes, Test : 418 lignes
  - Colonnes : PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked



Nettoyage des données

- Explorer le dataset en le visualisant, distributions, fréquence
- Réduire les données (quantité) : retirer les doublons et les erreurs, sélectionner les caractéristiques intéressantes, si dataset trop grand travailler d’abord sur un échantillon
- Améliorer les données (qualité) : enlever le bruit, les données manquantes et les valeurs aberrantes, choisir pour les données redondantes
- Transformation des données : normalisation, données catégoriques, binaires

Data mining : choix de l’algorithme

Méthode de Data Mining

Classification : Prédiction de la survie d’un passager (binaire : 0 = Non, 1 = Oui).

Train/Validation/Test : Diviser le dataset d’entrainement en Train et Validation pour tester avant de soumettre (80%/20%)

Choix des Algorithmes

Régression logistique : Interprétable et efficace pour la classification binaire.

Random Forest : Amélioration des performances grâce à un ensemble d’arbres de décision.

SVM (Support Vector Machine) : Séparation optimale des classes dans des espaces complexes.

XGBoost : Optimisation avancée pour améliorer la précision des prédictions.

Optimisation et Fine Tuning

Recherche des meilleurs hyperparamètres via Grid Search ou Random Search.

Validation croisée pour évaluer la robustesse du modèle.

Feature engineering pour améliorer la qualité des prédictions (extraction de nouvelles variables).

Attention à l’overfitting.

Bon courage !!