Data Mining

Dr. Ilham KADI

Kadi.ilham@gmail.com

2023/2024

Plan

- Introduction aux techniques DM
- Les données
- Processus Data mining
- Prétraitement des données

Exemple introductif : demande de crédit bancaire

- divorcé
- 5 enfants à charge
- chômeur en fin de droit
- compte à découvert

Expérience de l’entreprise : ses clients et leur comportement

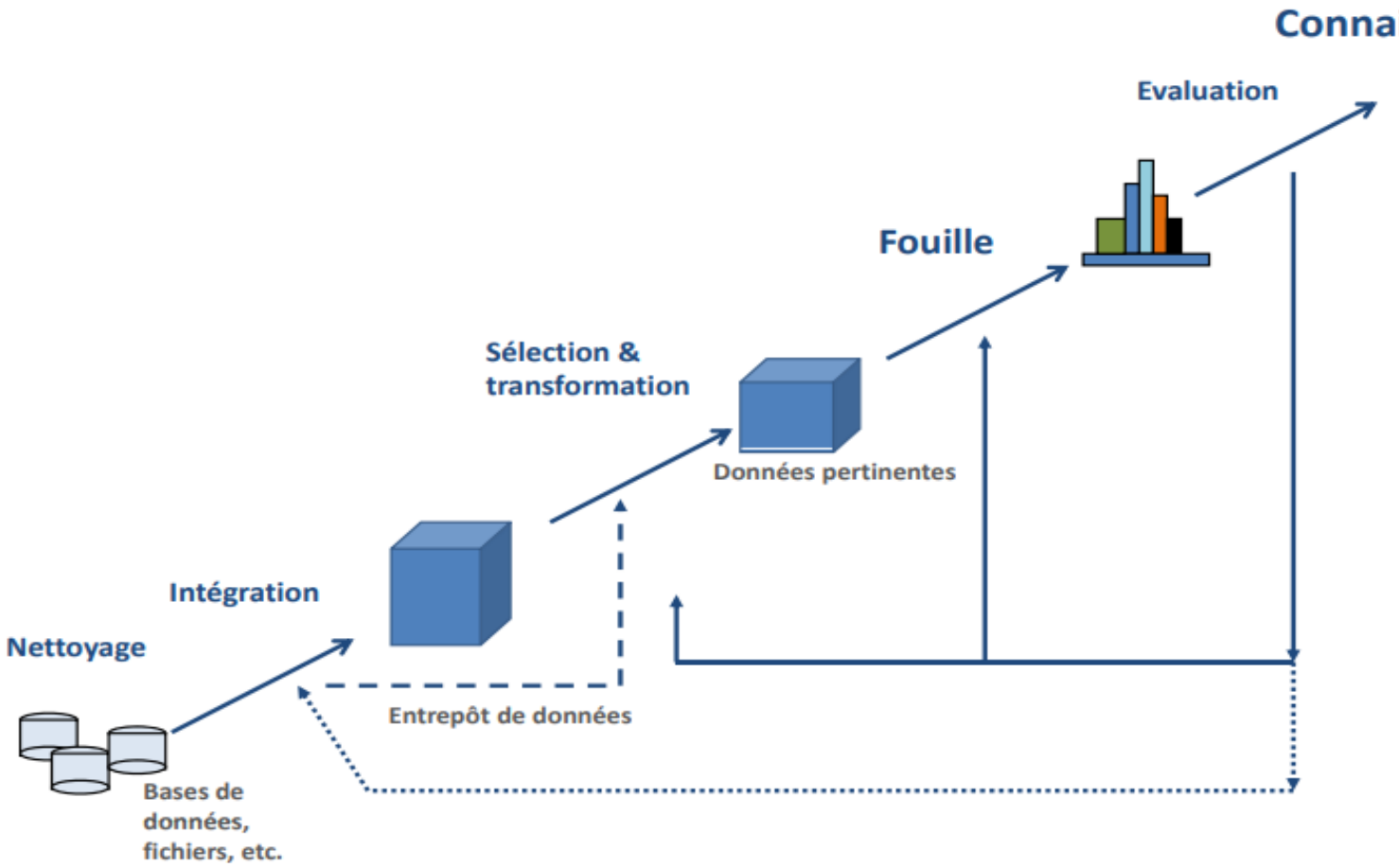
- Coûteuse en stockage
- Inexploitée

Comment et à quelles fins utiliser cette expérience accumulée

KDD: Knowledge Discovery in Databases

KDD (Extraction de Connaissances à partir des Données) est un processus (semi)- automatique d'extraction de connaissances à partir de bases de données où les connaissances sont :

- valides
- non connues a priori
- potentiellement utiles



KDD: Knowledge Discovery in Databases

Motivation (1): Explosion des données

- **Masse importante de données** (millions de milliards d'instances) : elle double tous les 20 mois, BD très larges.
- **Données multi-dimensionnelles** (milliers d'attributs) : BD denses non exploitables par les méthodes d'analyse classiques
- **Collecte de masses** importantes de données (Gbytes/heure)
- **Besoin de traitement** en temps réel de ces données

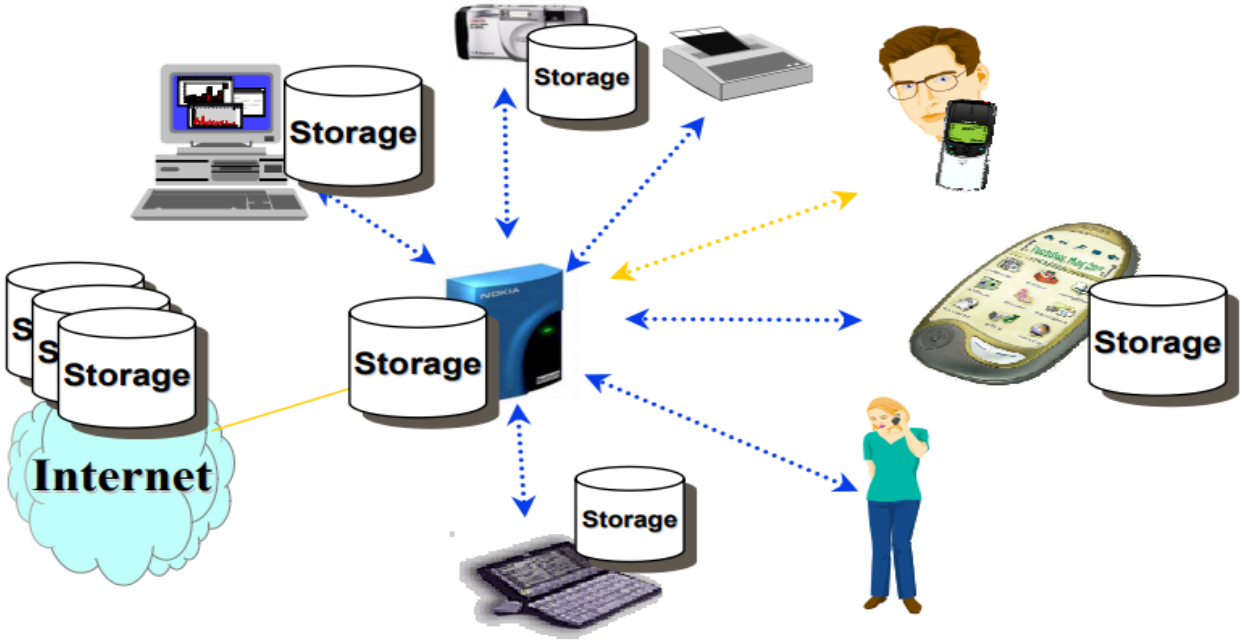
Motivation (2): Améliorer la productivité

- Forte pression due à la concurrence du marché
- Brièveté du cycle de vie des produits
- Besoin de prendre des décisions stratégiques efficaces
  - Exploiter le vécu (données historiques) pour prédire le futur et anticiper le marché

Motivation (3): Croissance en coût

- Croissance en puissance/coût des machines capables
  - de supporter de gros volumes de données
  - d'exécuter le processus intensif d'exploration

Motivation (4): Supports hétérogènes



Data Mining: Définition

Un processus permettant **l'extraction** de connaissances et **découverte** de règle, relations, corrélations et/ou dépendances sous la forme de **modèles** à partir de grandes masses de données

Définition

Ces modèles peuvent être de nature

- **Descriptive** : permettant d'expliquer le comportement actuel des données
- **Prédictive** : comportement futur des données.

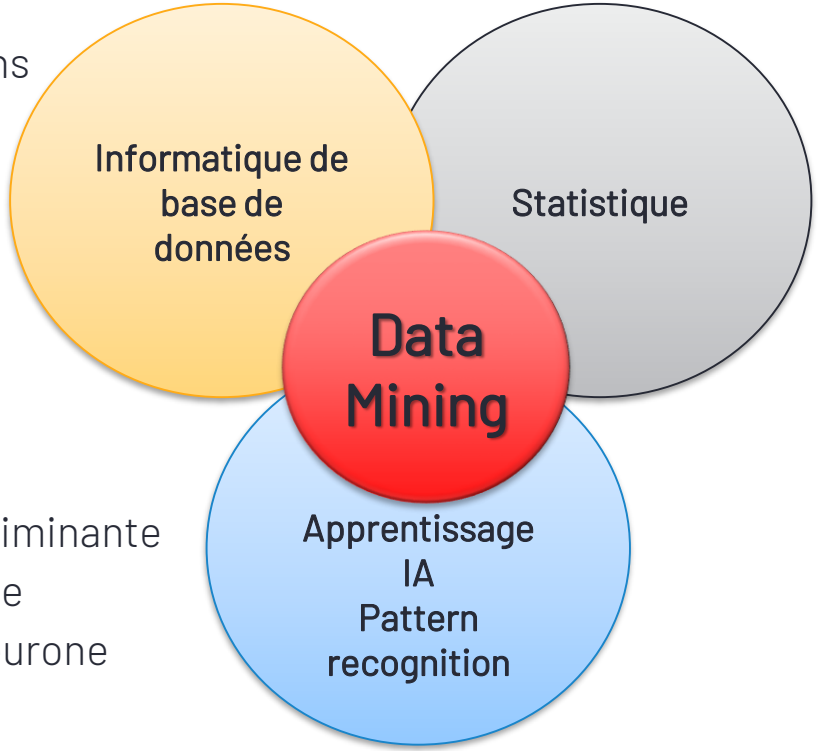
Data Science: vocabulaire

- Reconnaissance des formes (pattern recognition)
- Apprentissage automatique (machine learning)
- Intelligence artificielle
- Fouille de données (data mining)
- Statistique
- ....

Domaines différents avec des intersections plus ou moins grandes

La rencontre de plusieurs disciplines

- Règles d'associations
- ...



- Régression
- Maximum de vraisemblance

- Analyse discriminante
- Apprentissage
- Réseau de neurone
- ...

Data Mining vs Big data

	Big data	Data mining
Volume vs relation	se concentre sur un très grand volume de données non structurées	travaille sur la relation entre les données
Concept vs technique	un concept (voir un environnement)	une méthode scientifique d'extraction
Typologie de données	travaille sur des données complexes et non structurées	des données structurées
Décision vs prédiction	analyse de macro-statistiques qui vont permettre d'établir des prédictions basées sur un grand volume de données	un outil d'aide à la décision précis sur une question

Data Mining vs BI

	BI	Data mining
Volume	Utilisation de grands ensembles de données pour trouver des informations	Utiliser des ensembles de données de plus petite taille
Style	Utilise le suivi des métriques pour obtenir des informations,	Utilise l'intelligence de calcul et des algorithmes pour découvrir des modèles utiles
Résultat	Fournit des informations qui peuvent aider à la prise de décision	Donne des réponses à des questions particulières.

DM: Exemple

Problématique

- Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique, cinéma.
- Il veut étudier ses clients pour découvrir de nouveaux marchés ou vendre plus à ses clients habituels.

Quelques questions

1. Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?
2. A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?
3. Est-ce que les acheteurs de magazines de musique sont aussi amateurs de cinéma ?
4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?
5. Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?

Question	Solution
Q1	Requête SQL à partir des données opérationnelles suffit si les tables concernées ont été suffisamment indexées.
Q2	Nécessite de garder toutes les dates de souscription, même pour les abonnements résiliés: Requêtes multidimensionnelles de type OLAP.
Q3	▪Exemple simplifié de problème où l'on demande si les données vérifient une règle. ▪Réponse formulée par une valeur estimant la probabilité que la règle soit vraie. ▪Utilisation d'outils statistiques.

Question	Solution
Q4	Question plus ouverte, il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser
Q5	Question ouverte : il faut disposer d'indicateurs comme durée d'abonnement, délai de paiement, …

C'est pour ce type de questions que sont mis en œuvre les outils de fouille de données

Domaine d’application de DM

- Entreprise et Relation Clients : création de profils clients, ciblage de clients potentiels et nouveaux marchés
- Bioinformatique : analyse du génome, ADN…
- Médecine: diagnostic, traitement…
- Internet : spam, e-commerce, détection d’intrusion, Sécurité
- Gestion et analyse de risque : Assurances, Banques, Fraud…
- Web mining, text mining…

Exemple 1: E-commerce

Targeting

- Stocker les séquences de clicks des visiteurs, analyser les caractéristiques des acheteurs
- Faire du "targeting" lors de la visite d'un client potentiel

Systèmes de recommandation

- **Opportunité** : les clients notent les produits ! Comment tirer profit de ces données pour proposer des produits à un autre client ?
- **Solutions** : technique de filtrage collaboratif pour regrouper les clients ayant les mêmes "goûts".

Exemple 2: Commerce

Opinion mining

- Exemple : analyser l'opinion des usagers sur les produits d'une entreprise à travers les commentaires sur les réseaux sociaux et les blogs

Les données

Les données

Les données peuvent être vues comme une collection d'objets (enregistrements) et leurs attributs

- **Un attribut** est une propriété et ou une caractéristique de l'objet
- *Exemple*: température, poids...
- L'attribut est également appelé caractéristique, variable, champ

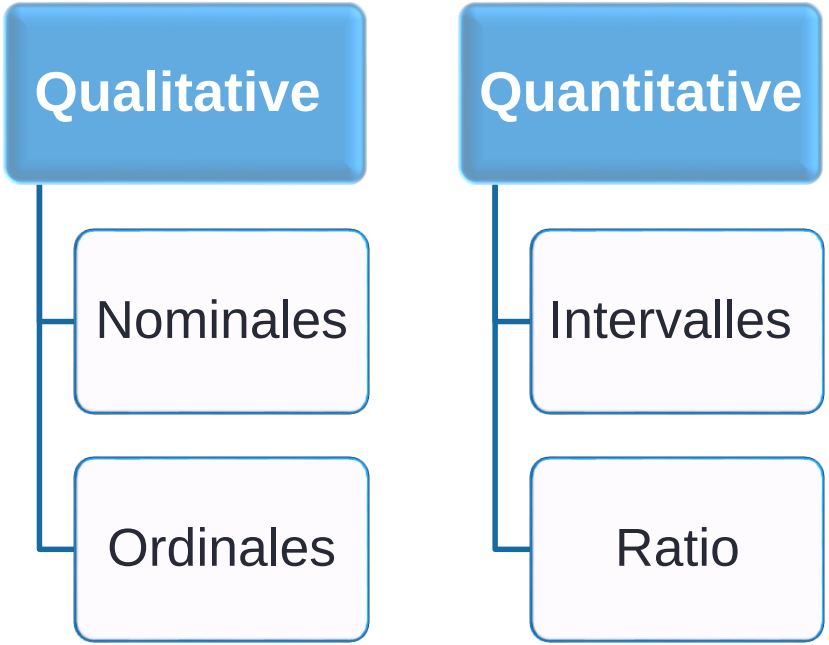
- Un ensemble d'attributs décrit **un objet**
- L'objet est également appelé enregistrement, observation, entité ou instance

Attributs

Objets

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types de données



Types de données: Qualitative

- **Variables qualitatives** ou catégorielles expriment une qualité comme le sexe, le métier ou le nom.
  - Nominales, comme par exemple le nom des journaux, le signe astrologique.
  - Ordinales, désigne le rang : un peu, moyen, beaucoup, énormément

Types de données: Nominal

- Les variables nominales présentent des catégories que l'on nomme avec un nom.
- Le seul calcul faisable sur les variables nominales est le nombre d'éléments ou pourcentage par catégorie.
- **Exemple:**

Quel est votre sexe:

- ○ Homme
- ○ Femme

Votre couleur de cheveux:

- ○ Noir
- ○ Blond
- ○ Châtain
- ○ Autre

Types de données: Ordinal

- Les variables ordinales sont des catégories qui sont naturellement ordonnées.
- Ça peut être le classement à une course, par exemple ou le résultat à questionnaire sur une échelle de Likert.
- **Exemple:**

Types de données: Quantitative

- Variables quantitatives contiennent des valeurs mesurables.
  - De rapports, exemple : distance, durée, valeur;
    - Discrètes exemple : âge, nombre d'habitants;
    - Continues exemple : distance.
  - D'intervalles exemple :date de naissance, heure d'arrivée;
    - Discrètes exemple : date en général;
    - Continues exemple : température.

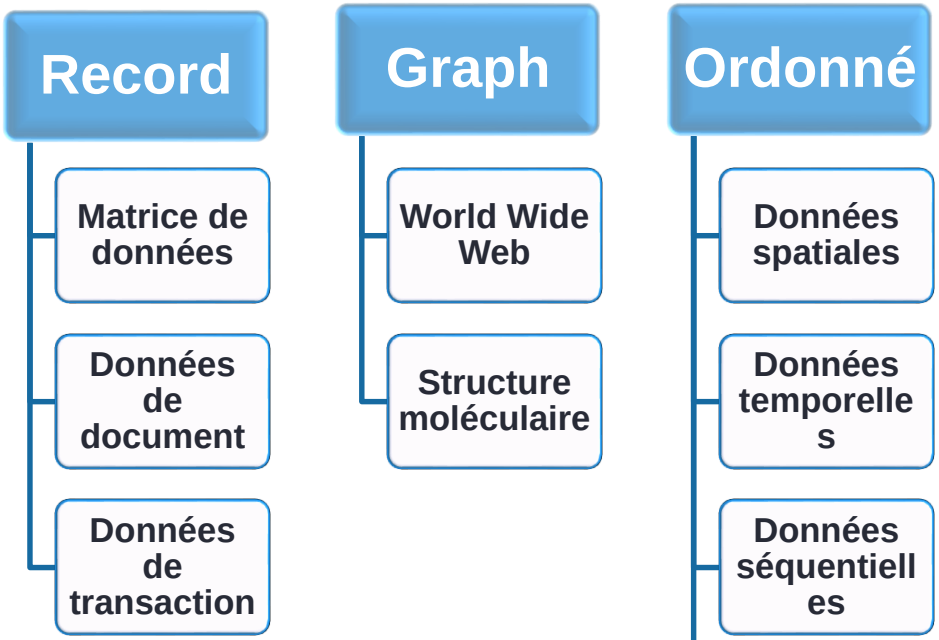
Types de données: Interval

- Nous pouvons, non seulement, ordonner les items qui sont mesurés, mais également mesurer et comparer les tailles des différences entre elles.
- **Exemple** :
  - Nous pouvons dire qu'une température de 40 degrés Celsius est plus haute qu'une température de 30 degrés,
  - et qu'une augmentation de 20 à 40 degrés est deux fois plus qu'une augmentation de 30 à 40 degrés.
  - A 0°c, il y a toujours une température.

Types de données: Ratio ou rapport

- Les variables de ratios sont très semblables à celles d'intervalle avec un point nul absolu identifiable
- Dans une donnée de ratio, le zéro signifie réellement l'absence de quelque chose.
- **Exemple** :
  - Pour la durée d'un test, à 0, il n'y pas de temps
  - Si vous avez zéro produit laitier dans votre panier, c'est qu'il n'y a réellement aucun produit laitier.

TYPES D'ENSEMBLES DE DONNÉES



Données d’enregistrement

- Une collection d'enregistrements avec un ensemble fixe d'attributs

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes

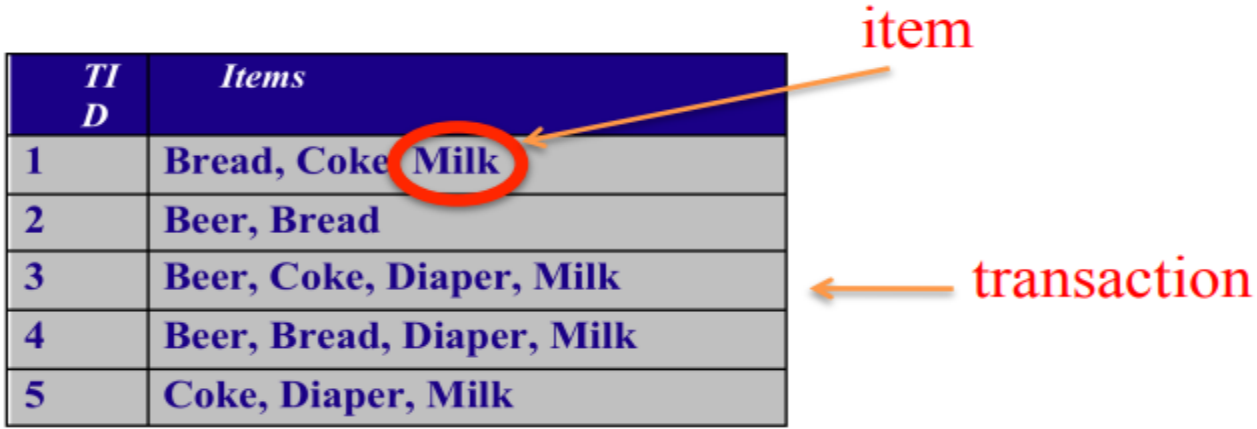
Données de document

- Chaque document devient un vecteur de «terme»,
  - Chaque terme est une composante (attribut) du vecteur,
  - la valeur de chaque composant est le nombre de fois le terme correspondant apparaît dans le document.

	Équipe	Coach	Jouer	Ballon	Score	Jeu
Document 1	3	0	5	0	2	6
Document 2	0	7	0	2	1	0
Document 3	0	1	0	0	1	2

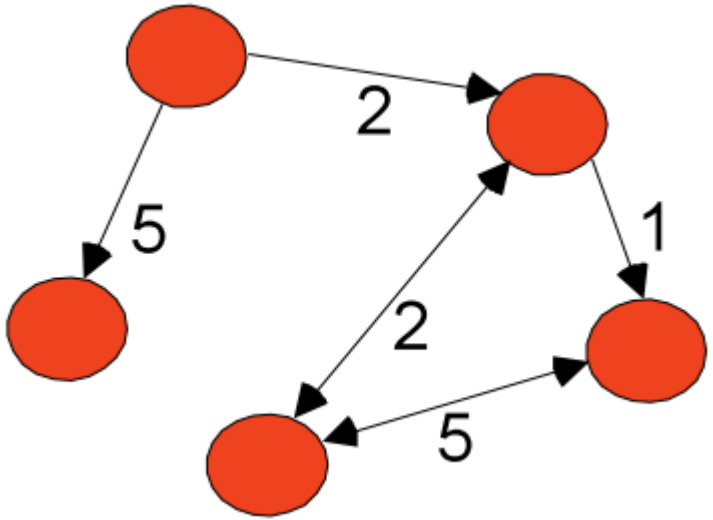
Données de transaction

- Un type spécial d'ensemble de données, où chaque enregistrement (transaction) implique un ensemble d'articles.
- Exemple: dans une épicerie. L'ensemble des produits achetés par un client constitue une transaction, tandis que les produits individuels qui ont été achetés sont les articles (items) .



Graph data

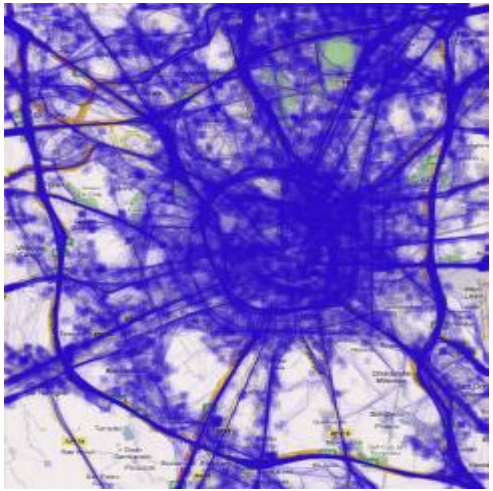
- **Exemple**: graphes génériques et liens HTML



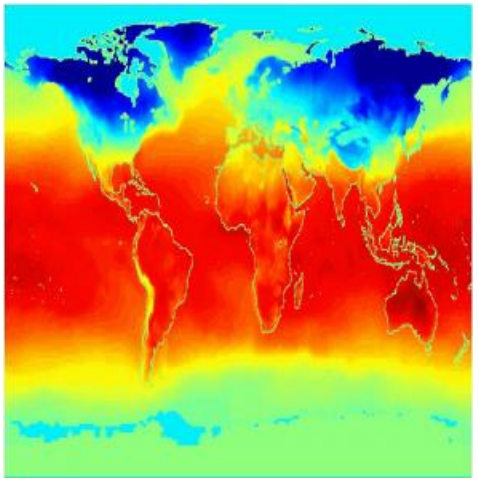
Données de séquence génétique

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Données spatio-temporelles



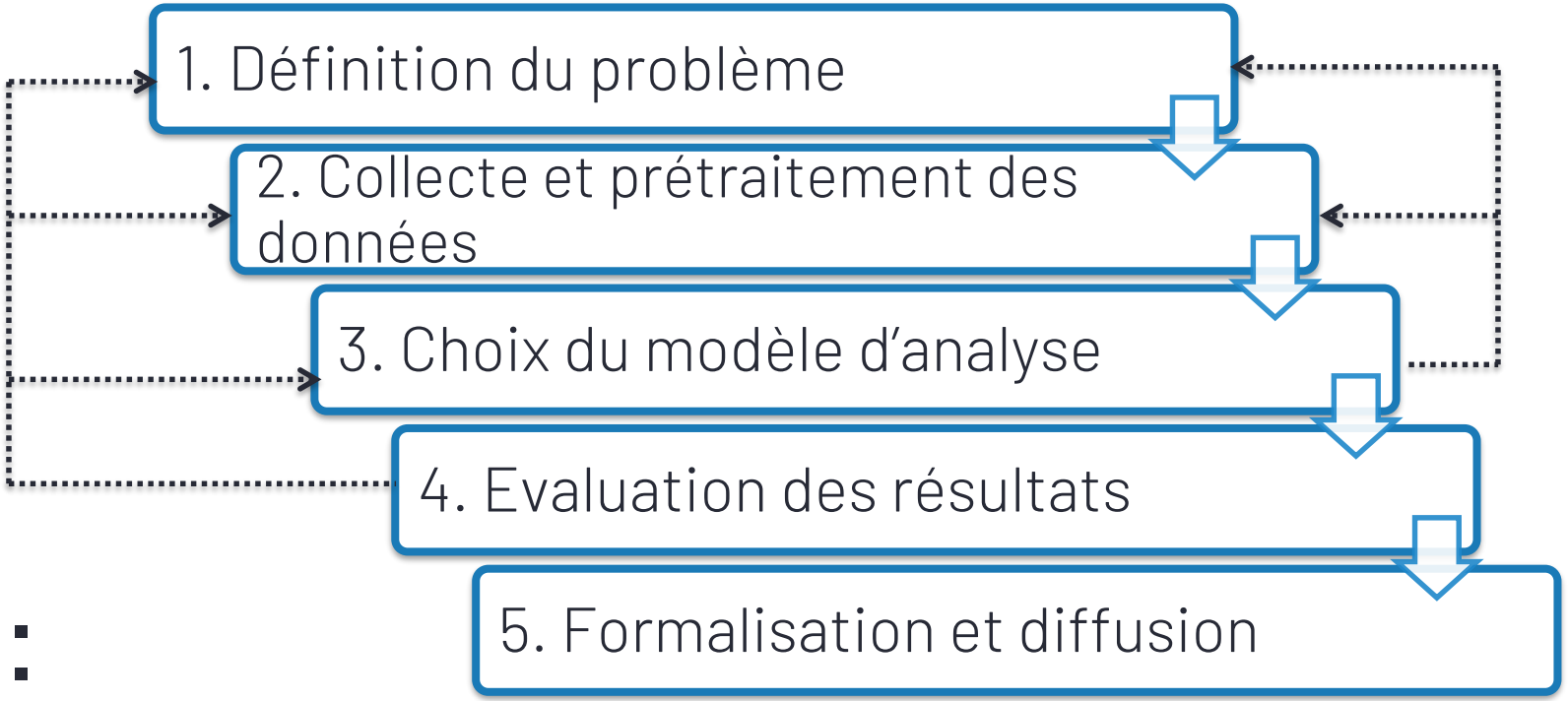
- Trajectoires d'objets en mouvement



- Température mensuelle moyenne de la terre et de l'océan

Processus Data Mining

Processus data mining



1. Définition du problème

- Quel est le but de l'analyse, que recherche-t-on ?
- Quels sont les objectifs ?
- Comment traduire le problème en une question pouvant servir de sujet d'enquête pour cet outil d'analyse bien spécifique ?

Se souvenir que l'on travaille à partir des données existantes, la question doit être ciblée selon les données disponibles.

2. Collecte et prétraitement des données

- N'analyser que des données "propres" et consolidées.
- Extraire de l'analyse les données de qualité douteuse.
- Bien souvent, les données méritent d'être retravaillées.
- S'assurer que la quantité de données soit suffisante pour éviter de fausser les résultats.

La phase de collecte nécessite le plus grand soin

3. Choix du modèle d’analyse

- Choisir l’algorithme d’analyse convenable.
- Valider le choix d'analyse sur plusieurs jeux d'essais en variant les échantillons.
- Une première évaluation peut conduire à reprendre les étapes 1 ou 2.

3. Choix du modèle d'analyse

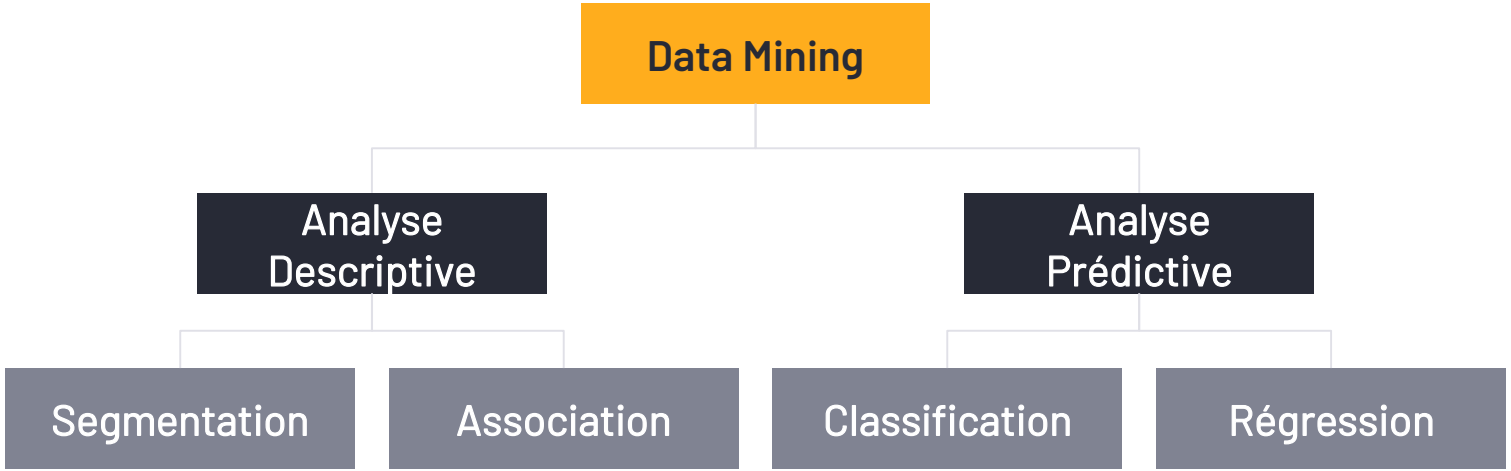
Modèles prédictifs

- Utilisent les données existantes et des résultats connus sur ces données pour développer des modèles capables de prédire les valeurs d'autres données.
- Exp: Prédire les clients qui ne rembourseront pas leur crédit.

Modèles descriptifs

- Proposent des descriptions de données pour aider à la prise de décision.
- Souvent en amont de la construction de modèles prédictifs.

3. Choix du modèle d'analyse



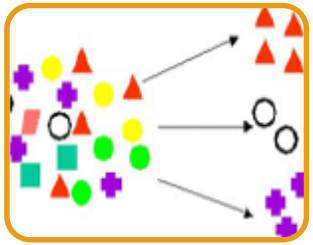
3. Choix du modèle d'analyse

Classification

Régression

Segmentation (Clustering)

Association



La variable décisionnelle est **qualitative**

•Un dossier de crédit peut être classifié : BON ou MAUVAIS

•Un patient peut présenter un fort risque de maladie cardiaque

La Classification a pour objectifs :

•Détecter les variables possédant un lien fort avec la variable décisionnelle

•Construire un modèle de classification liant ces variables à la décision

Plusieurs méthodes et techniques pour classifier :

•Arbre de décision

•Forêts Aléatoires

•K-NN k-nearest neighbor…

Déterminer ce qui ***caractérise*** *un groupe* ***particulier*** *de clients*

3. Choix du modèle d'analyse

Classification

Régression

Segmentation (Clustering)

Association

La variable décisionnelle est **quantitative**

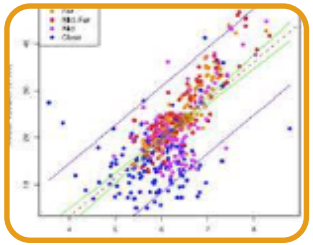
•Prédire les tendances salariales la prochaine année

•Prédire le meilleur pourcentage de réduction de coûts

La régression a pour objectifs :

•Détecter les variables possédant un lien fort avec la variable cible

•Construire un modèle prédictif avec l'ensemble des variables pertinentes afin de prédire la variable d'intérêt



Régression Linéaire

• Méthode des moindres carrés

• Meilleurs prédicteurs

3. Choix du modèle d'analyse

Classification

Régression

Segmentation (Clustering)

Association

Aucune variable décisionnelle

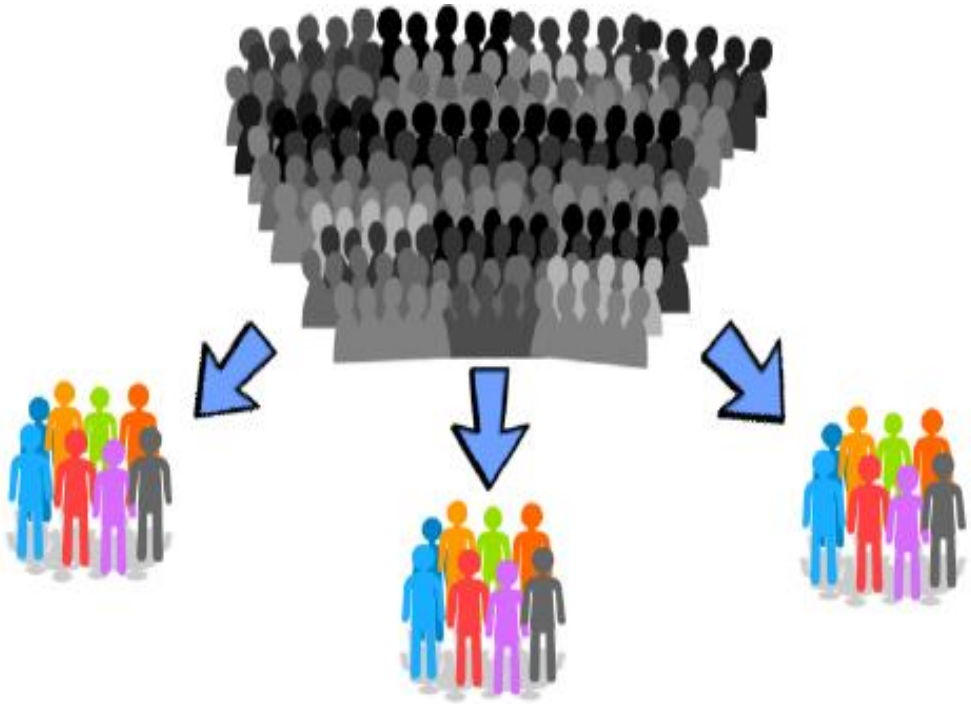
- Les variables d'entrées servent à créer des groupes homogènes
- Les individus de chaque groupe se ressemblent le plus
- Les groupes d'appartenances obtenus se distinguent le plus

La Segmentation a pour objectifs :

- Trouver les variables métiers influençant la répartition en groupes
- Affecter les individus à leurs nouveaux groupes d'appartenance

Plusieurs méthodes et techniques pour segmenter :

- Partionnement : k-means
- Hiérarchique : CAH



Trouver les **comportements typiques** des clients.

SEGMENTATION – TYPOLOGIE

3. Choix du modèle d'analyse

Classification

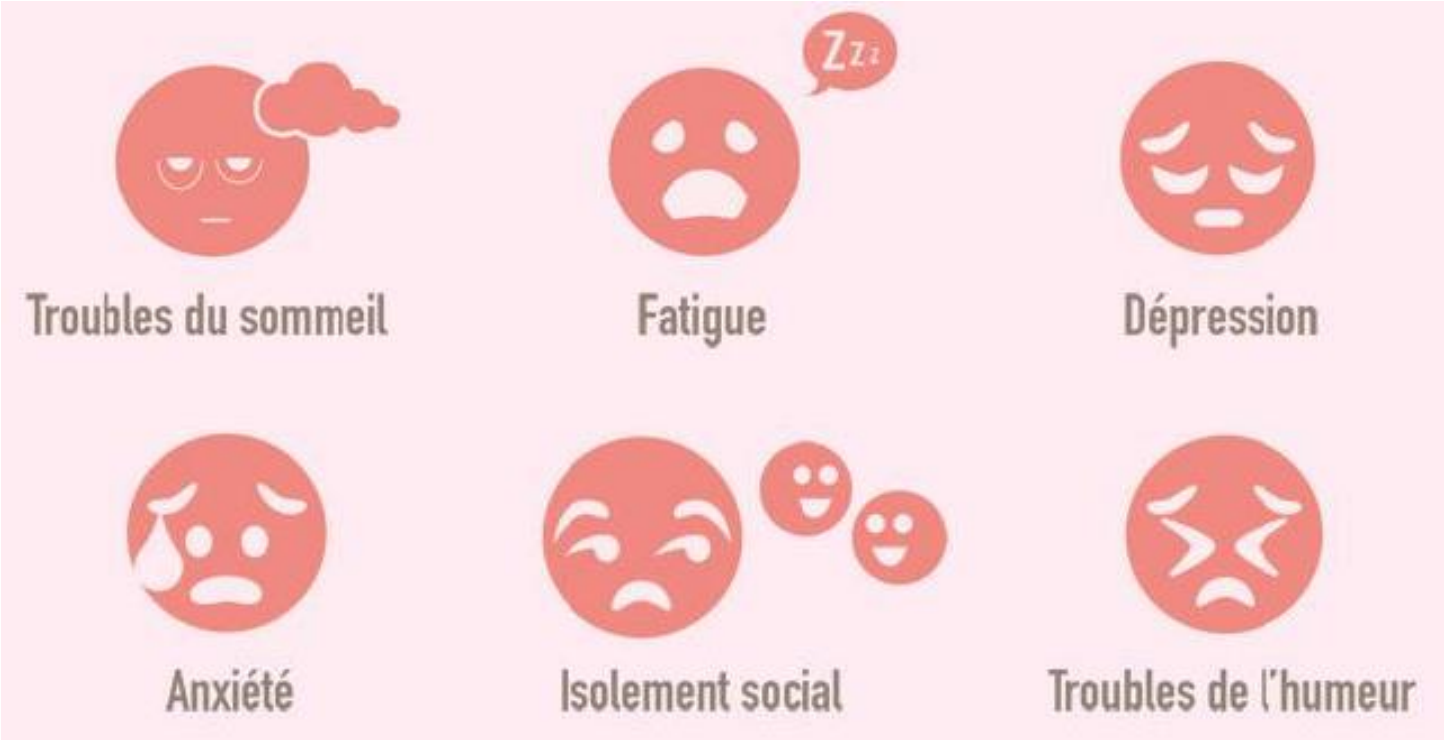
Régression

Segmentation (Clustering)

Association

Recherche des articles les plus/moins **associés**

ANALYSE DU CHARIOT



MÉDECINE: les symptômes associés

4. Evaluation des résultats

- Il est temps d'exploiter les résultats.
- Pour affiner l'analyse, reprendre les étapes 1, 2 ou 3 si les résultats s'avéraient insatisfaisants.
- C'est à dire qu'ils ne seraient pas en phase avec les objectifs fixés à l’étape 1.

5. Formalisation et diffusion

- Les résultats sont formalisés pour être diffuser.
- Incorporation de ces connaissances dans des autres systèmes pour d'autres actions.
- Mesurer l'effet de ces connaissances sur le système, vérifier et résoudre les conflits possibles avec les connaissances antérieures.

Prétraitement des données

Exemple introductif

- Soit l'ensemble de données suivant auquel une technique data mining va être appliqué pour répondre à une question stratégique pour l'entreprise

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

Exemple introductif

- Corrections des doublons, des erreurs de saisie

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

Exemple introductif

- Intégrité de domaine

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémol	Rue du moulin, Paris	11/11/1111	Maison

Exemple introductif

- Information manquante
  - Cas où les champs ne contiennent aucune donnée.
  - Parfois intéressant de conserver ces enregistrements car l'absence d'information peut être informative (e.g. fraude).

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	NULL	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23134	Bémol	Rue du moulin, Paris	NULL	Maison

Exemple introductif

- Représentation horizontale ou éclatée

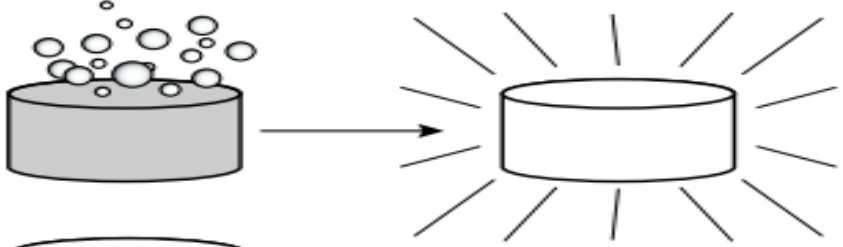
Client	Magazine
23134	Voiture
23134	Musique
23134	BD
31435	BD
43342	Sport
43241	Sport
23134	Maison

Client	Sport	BD	Voiture	Maison	Musique
23134	0	1	1	1	1
31435	0	1	0	0	0
43342	1	0	0	1	0
43241	1	0	0	1	0

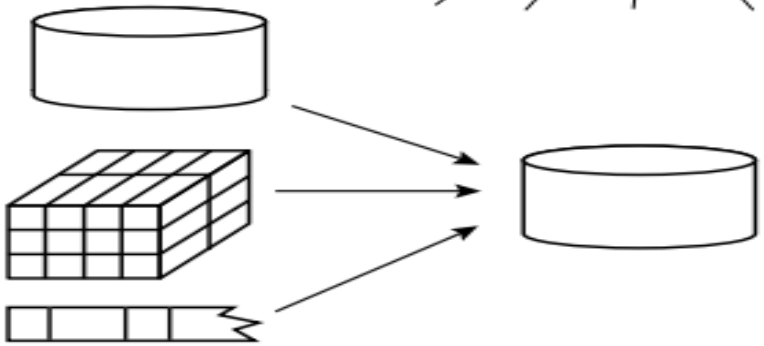
Pourquoi prétraiter les données ?

- Données réelles souvent :
  - Incomplètes : valeurs manquantes, données simplifiées
  - Bruitées : erreurs et exceptions
  - Incohérentes : nommage, codage
- Résultats de la fouille dépendent de la **qualité** des données

Nettoyage de données



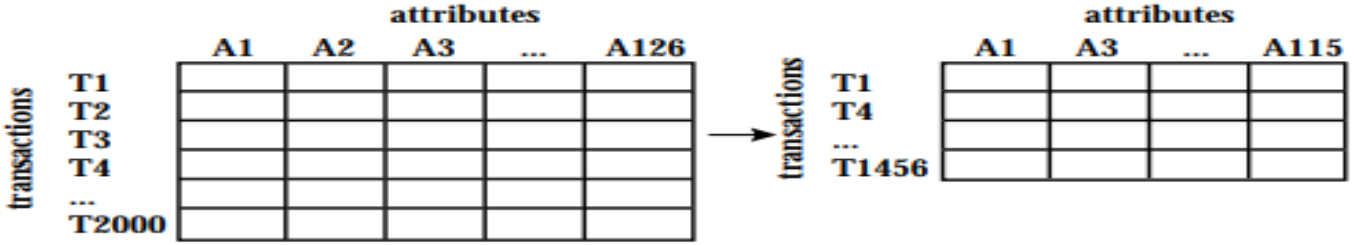
Intégration de données



Transformation

2, 32, 100, 59, 48 ⟶ 0.02, 0.32, 1.00, 0.59, 0.48

Réduction



Principales étapes dans le prétraitement

Nettoyage des données

- Le nettoyage des données est un processus qui vise à identifier et corriger les données altérées, inexactes ou non pertinentes.
- Cette étape est fondamentale dans la **préparation des données.**
- **Objectif:**
  - Garantir que seules des données propres et de haute qualité seront transférées vers les systèmes cibles.
  - Améliorer la cohérence, fiabilité et valeur des données.

Données manquantes

- Données non disponibles
  - Certains attributs n'ont pas de valeur
- **Causes** :
  - Mauvais fonctionnement de l'équipement
  - Incohérences avec d'autres données et donc supprimées
  - Non saisies car non ou mal comprises
  - Considérées peu importantes au moment de la saisie
- Ces données doivent être inférées

Comment remplir les trous ?

Suppression

- Ignorer/supprimer les cas avec des données manquantes
- Peu efficace quand le pourcentage de valeurs manquantes est élevé

Tolérance

- Stratégie de traitement internes dans lesquelles l'analyse est effectuée directement, en utilisant les ensembles de données avec des données manquantes.

Imputation

- Stratégie pour remplir le données manquantes d'un ensemble de données.

Imputation par moyenne / médiane

- Calculer la moyenne / médiane des valeurs non manquantes dans une colonne,
- Remplacer les valeurs manquantes dans chaque colonne séparément et indépendamment des autres.
- Ne peut être utilisé qu'avec des données numériques.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Imputation par (le plus fréquent) ou (zéro / constante):

- **Plus fréquent:**
  - Remplacer les données manquantes par les valeurs les plus fréquentes dans chaque colonne.
  - Fonctionnel pour les données discrète.
- **Zéro/Constante:**
  - Remplace les valeurs manquantes par zéro ou une valeur constante.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Imputation utilisant un algorithme:

- Remplacer les données manquantes par les valeurs la plus probable.
- Utiliser des algorithmes pour estimer la valeur des données manquantes.

- **Exemple:**
  - Imputation par le centre du groupe.
  - Imputation à partir des kk plus proches voisins.
  - Imputation par une moyenne partielle.

Autres problèmes

- Bruit de données
- Enregistrement dupliqués
- Données incomplètes
- Données incohérentes

Intégration des données :

- Intégration des données :
  - Combinaison de différentes sources en une seule
- **Objectif**:
  - Intégrer les métadonnées de différentes sources
  - Identifier les différents noms des mêmes données réelles (Ex : numClient ⇔ clientId)
  - Détecter et résoudre les conflits de valeurs (Ex: Echelle différente)
  - Gestion de la redondance

Transformation (codage et normalisation)

- Une étape très dépendante du choix de l'algorithme DM utilisés.
- **Regroupements**:
  - Cas où les attributs prennent un très grand nombre de valeurs discrètes (e.g. adresses que l'on peut regrouper en 2 régions (Paris - Province))
- **Attributs discrets**:
  - Deux représentations possibles : représentation verticale ou représentation horizontale ou éclatée

Transformation (codage et normalisation)

- Représentation horizontale ou éclatée

Client	Magazine
23134	Voiture
23134	Musique
23134	BD
31435	BD
43342	Sport
43241	Sport
23134	Maison

Client	Sport	BD	Voiture	Maison	Musique
23134	0	1	1	1	1
31435	0	1	0	0	0
43342	1	0	0	1	0
43241	1	0	0	1	0

Transformation (codage et normalisation)

- Changements de types pour permettre certaines manipulations comme par exemple des calculs de distance, de moyenne, date de naissance
- Uniformisation d'échelle.
  - Certains algorithmes sont basés sur des calculs de distance entre enregistrements :
    - Variations d'échelle selon les attributs peuvent perturber ces algorithmes.

Réduction de données

- **Définition**: obtenir une représentation réduite du jeu de données, plus petite en volume, mais qui produit les mêmes (ou presque) résultats analytiques.
- Stratégies
  - Réduction de dimension
  - Réduction de numérosité
  - Discrétisation

Réduction de données

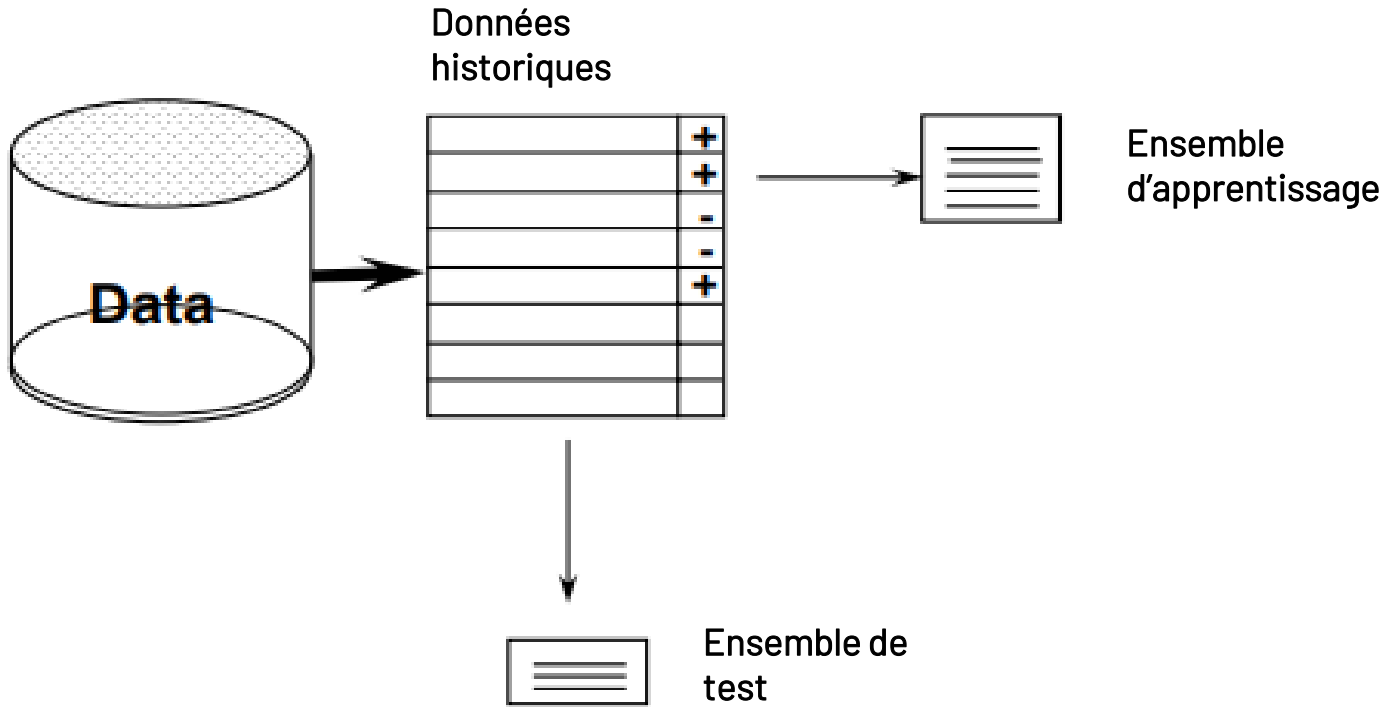
- **Réduction en ligne par échantillonnage** :
  - Pour des raisons de performance.
  - Du fait de la complexité importante des algorithmes d'extraction.
  - Plusieurs méthodes : échantillonnage aléatoire, échantillonnage par clustering.
- **Réduction en colonne** par suppression des attributs redondants:
  - Cas triviaux (âge et date de naissance).
  - Via une analyse des corrélation entre attributs

Mesures d'évaluation

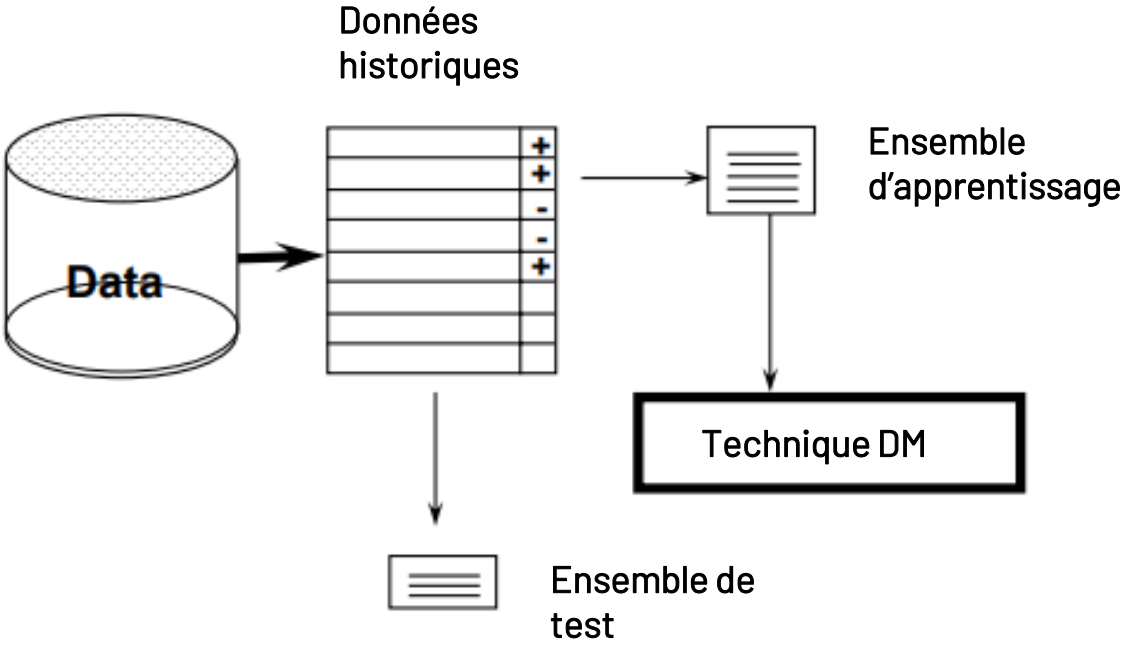
Construction d'un modèle DM

1. Diviser les données en ensemble d'apprentissage et ensembles de test
2. Construire le modèle DM en utilisant l'ensemble d'apprentissage
3. Evaluer le modèle en utilisant l'ensemble de test

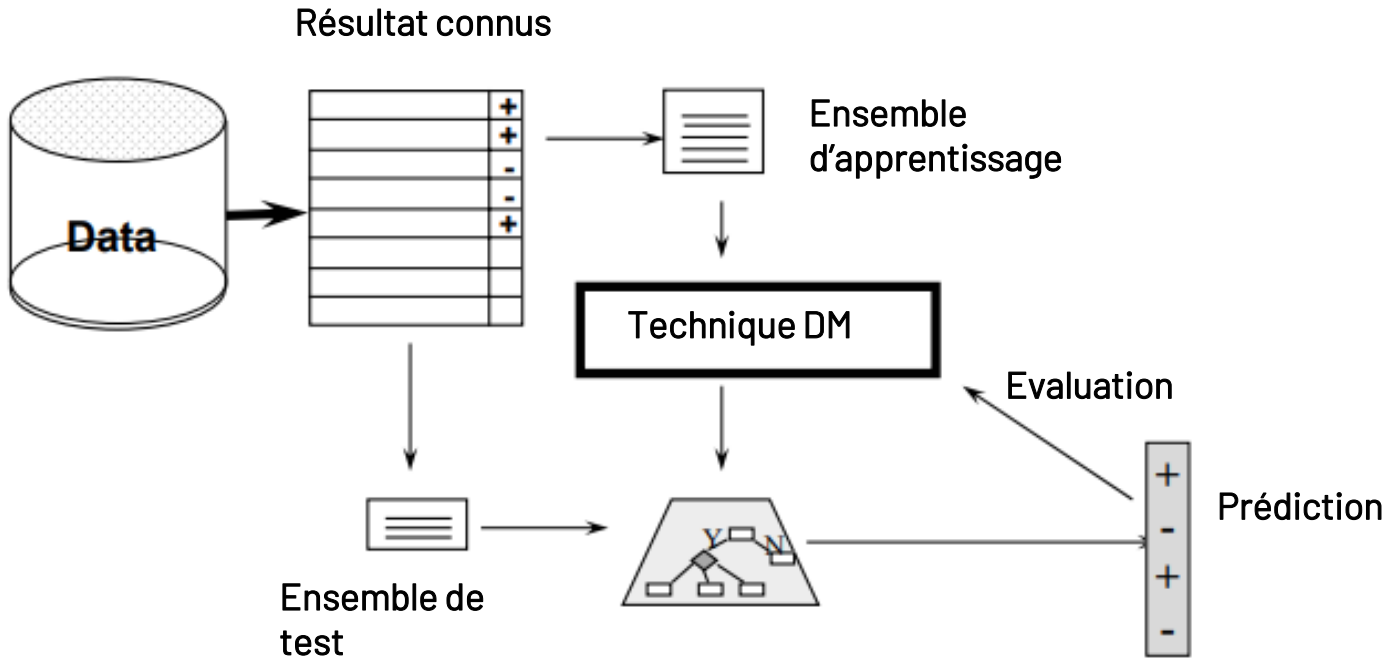
Etape 1


Données
historiques
Data
+
+
-
-
+
Ensemble
d'apprentissage
Ensemble de
test

Etape 2


Données
historiques
Data
+
+
-
-
+
Ensemble
d'apprentissage
Technique DM
Ensemble de
test

Etape 3


Résultat connus
Data
+
+
-
-
+
Ensemble
d'apprentissage
Technique DM
Evaluation
Prédiction
Ensemble de
test
Y
N
+
-
+
-

Echantillonnage de l’ensemble de données

- L’échantillonnage génère différents sous-ensembles de données à partir de l’ensemble initial D.
- Différentes méthodes:
  - Hold-out
  - k-fold cross-validation
  - Leave-one-out

Hold-out

- Utiliser deux ensembles de données indépendants, par exemple, ensemble d'apprentissage (2/3), ensemble de test (1/3);
- Échantillonnage aléatoire
- Il est important que les données de test ne soient en aucun cas utilisées pour créer le modèle DM!
- Une répartition aléatoire est utilisée pour les données très volumineuses

Hold-out

- L'estimation du holdout peut être rendue plus fiable en répétant le processus avec différents sous-échantillons:
  - À chaque itération, une certaine proportion est sélectionnée au hasard pour l'apprentissage
  - Les taux d'erreur (précisions de classification) sur les différentes itérations sont moyennés pour donner un taux d'erreur global
  - Calculez également un écart type!

Pas optimal: les différents ensembles de tests se chevauchent généralement

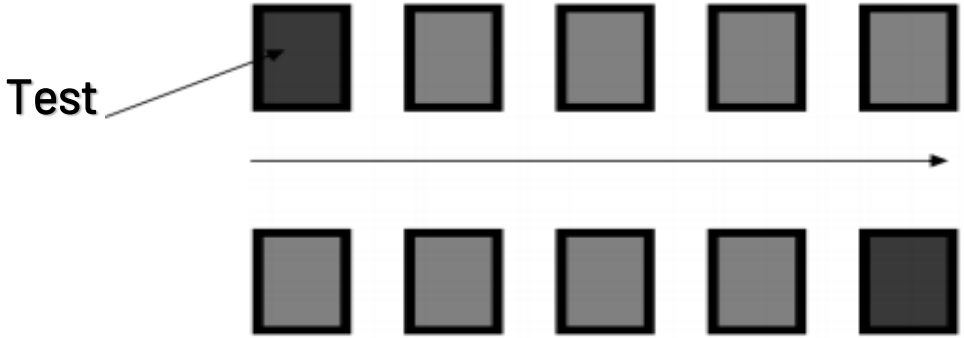
Cross-validation (validation croisée)

- La validation croisée évite les ensembles de tests qui se chevauchent
  - **Première étape:** les données sont divisées en k sous-ensembles de taille égale
  - **Deuxième étape:** utiliser **k-1** sous ensemble comme données d'apprentissage et **un** sous ensemble comme données de test; répéter **k fois.**
- C'est ce qu'on appelle la validation croisée k-fold
- Les estimations d'erreur sont moyennées pour produire une estimation d'erreur globale

Cross-validation: Exemple

1. Divisez les données en groupes de même taille

2. Laisser un groupe pour le test et utilisez le reste pour créer un modèle



3. Répéter

10-fold cross-validation

- Méthode standard d'évaluation
- Pourquoi dix? Des expériences approfondies ont montré que c'est le meilleur choix pour obtenir une estimation précise
- L'écart type est également essentiel pour comparer les algorithmes d'apprentissage.
- Par exemple. La validation croisée décuple est répétée plus de fois et les résultats sont moyennés (réduit la variance)!

Leave-one-out

- Une forme particulière de validation croisée, utilisé pour les données de petite taille.
  - Définir le nombre de sous ensemble en se basant sur le nombre d'instances d'apprentissage.
  - C'est-à-dire, pour **n instances** d'apprentissage, construire le modèle **n fois** mais à partir de n -1 exemples d'apprentissage …
- N'implique aucun sous-échantillonnage aléatoire.
- Assez cher en calcul!

Qualités attendues d'un modèle DM

Précision	Le taux d'erreur, proportion d'individus mal classés doit être le plus bas possible.
Robustesse	Le modèle doit dépendre peu que possible de l'échantillon d'apprentissage et se généraliser à d'autres échantillons.
Concision	Les règles du modèles doivent être aussi simples et aussi peu nombreuses que possible.
Rapidité de calcul	Apprentissage rapide pour affinement du modèle.
Paramétrage	Pouvoir pondérer les erreurs de classement

Méthodes d'évaluation

Classification

- Matrice de confusion
- Taux d'erreur
- Recall / precision
- F-mesure
- Courbe ROC

Association

- Support
- Confidence
- Lift

Clustering

- MDL: longueur minimale de description

Matrice de confusion

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Matrice de confusion

- **TP (True Positives)** : les cas où la prédiction est positive, et où la valeur réelle est effectivement positive.
  - Exemple : le médecin vous annonce que vous êtes malade, et vous êtes vraiment malade.
- **TN (True Negatives)** : les cas où la prédiction est négative, et où la valeur réelle est effectivement négative.
  - Exemple : le médecin vous annonce que vous n'êtes pas malade, et vous n'êtes effectivement pas malade.

Matrice de confusion

- **FP (False Positive)** : les cas où la prédiction est positive, mais où la valeur réelle est négative.
  - Exemple : le médecin vous annonce que vous êtes malade, mais vous n'êtes pas malade.
- **FN (False Negative)** : les cas où la prédiction est négative, mais où la valeur réelle est positive.
  - Exemple : le médecin vous annonce que vous n'êtes pas malade, mais vous êtes malade.

Taux d'erreur: accuracy

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Taux d’erreur: Exemple

- On considère un problème à 2 classes avec : 9990 instances de classe 0 et 10 instances de classe 1.
- Si le modèle prédit que tout instance est de classe 0, on a
  - Accuracy = 9990/10000 = 99, 9 %

Précision

- La **précision** permet de répondre à la question suivante :
  - Quelle proportion d'identifications positives était effectivement correcte ?
- La précision peut être définie comme suit :

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Un modèle ne produisant aucun faux positif a une précision de 1,0.

Précision: Exemple

- Calculons la précision du modèle de DM qui analyse les tumeurs :

Vrais Positifs (TP) : 1	Faux positifs (FP) : 1
Faux négatifs (FN) : 8	Vrais négatifs (TN) : 90

$$\text{Précision} = \frac{TP}{TP+FP} = \frac{1}{1+1} = 0.5$$

- Le modèle a une précision de 0,5. ie, quand il prédit qu'une tumeur est maligne, sa prédiction est juste dans 50 % des cas.

Recall

- Le **rappel** permet de répondre à la question suivante :
  - Quelle proportion de résultats positifs réels a été identifiée correctement ?
- Mathématiquement, le rappel est défini comme suit :

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Un modèle ne produisant aucun faux négatif a un rappel de 1,0.

Recall: Exemple

- Calculons le rappel pour le classificateur de tumeurs :

Vrais Positifs (TP) : 1	Faux positifs (FP) : 1
Faux négatifs (FN) : 8	Vrais négatifs (TN) : 90

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{1}{1+8} = 0.11$$

- Le modèle a un rappel de 0,11. En d'autres termes, il identifie correctement 11 % des tumeurs malignes.

F-mesure

- F-mesure ou F-score est une mesure qui combine la précision et le rappel,
- Moyenne harmonique entre la précision et le rappel :

$$F1 = 2.\frac{precision.recall}{precsion + recall} = \frac{2TP}{2TP + FP + FN}$$

Courbe ROC

- Une **courbe ROC** (receiver operating characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification.
- Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs.
- Le taux de vrais positifs (TPR) est l'équivalent du rappel:

$$TPR = \frac{TP}{TP + FN}$$

Courbe ROC

- Le taux de faux positifs (FPR) est défini comme suit
- $FPR = \frac{FP}{FP+TN}$