Exploration d'informations et Fouille de données

M1 - MER

2015 - 2016

Yoann Pitarch

yoann.pitarch@irit.fr

www.irit.fr/~Yoann.Pitarch

Déroulement du module

Volumes horaires

Cours : 12 heures

TD : 12 heures

TP : 12 heures / utilisation du logiciel **Knime**

Evaluation

Contrôle terminal : 70%

Projet TP : 30%

Contact

pitarch@irit.fr

Partie 1

Fouille de données

Motivations (1)

Explosion des données

- Des données de plus en plus nombreuses et hétérogènes
- Méthodes statistiques inadaptées et inapplicables
- Besoin d'un traitement temps réel

Augmentation de la puissance de calcul

- Stockage moins prohibitif
- Possibilité d'exécuter processus gourmands

Motivations (2)

Amélioration de la productivité

- Concurrence
- Raccourcissement du cycle des produits
- Besoin de prise de décisions stratégiques efficaces
    - Prise en compte de l'historique
    - Centré utilisateur

Le 4ème paradigme de la recherche

1. Expérimentale
2. Théorique
3. Informatique (simulation)
4. Analyse de données

La fouille de données

C’est quoi ?

Processus **inductif**, **itératif** et **interactif**

Dans des bases de données larges

Découverte d’un **modèle** de données **valide**, **nouveau**, **utile** et **compréhensible**

Domaines d'application (1)

- Marketing ciblé

  Population à cibler pour publipostage, leader de communautés

- Gestion et analyse de marchés

  Profils de clients, effet des soldes ou de campagnes publicitaires, …

- Détection de fraudes

  Télécommunications, secteur bancaires

- Gestion de stocks

  Quand commander ? Quelle quantité ?

- Analyse financière

  Maximiser profit d'un portefeuille

Domaines d’application (2)

- Gestion et analyse de risques

  Doit-on accorder un crédit ?

- Bioinformatique et génomique

  Fouille sur séquences ADN

- Médecine et pharmacie

  Aide au diagnostic, choix du médicament

- Fouille de textes

  Analyse d’opinion

- Réseaux sociaux

  Extraction de communautés, de leaders

- …

Cas d'étude (1)

Marketing

Situation

- Vous êtes un gestionnaire marketing d'un opérateur téléphonique
- Contrat d'un an = téléphone gratuit (150 euros)
- Commission de vente = 250 euros
- **Taux de renouvellement = 25%**
- Donner systématiquement un téléphone coûte cher !!
- Tout comme faire revenir un client parti

Cas d'étude (1)

Marketing

Solution

- Utilisation de la fouille de données pour prédire 3 mois avant la fin de contrat si le client a de grandes chances de partir
- Si c'est le cas
  - On offre un nouveau téléphone
- Sinon
  - Ne rien offrir

Cas d'étude (2)

Assurances

Situation

- Vous êtes un agent d'assurance
- Comment évaluer la prime pour un jeune conducteur qui a une voiture sportive ?

Solution

- Analyse de toutes les données de sinistres de la compagnie
- Définir un modèle pour estimer la probabilité d'avoir un accident (à partir du sexe, de l'âge, de l'adresse, …)
- Indexer le montant de la prime en fonction de la probabilité

Cas d'étude (3)

Banque

Situation

- Vous vous rendez compte que des achats frauduleux ont été faits avec votre CB
- Pourquoi la banque vous croirait-elle ?

Solution

- Construction d'un modèle de comportement normal à partir de de l'historique d'utilisation de votre CB
- Etablissement d'un score de similarité entre le profil normal et l'utilisation récente

Induction (1)

[Peirce,1903]

Abduction (diagnostic médical)

- *Tous les humains sont mortels*
- *Je suis mortel*
- *Je suis un humain*

Déduction

A partir de règles et d'hypothèses, on aboutit à une vérité par inférence

- *Tous les humains sont mortels*
- *Je suis un humain*
- *Je suis donc mortel*

Induction (2)

[Peirce,1903]

Induction

- Généralisation à partir d'un ensemble d'évènements singuliers

- Utilisé dans la fouille

- Non fiable à 100%

Exemple

- *Je suis mortel*
- *Vous êtes tous mortels*
- *Il est fort probable que tous les humains soient mortels*

La fouille de données

C’est quoi ?

“Drowning in Data yet Starving for Knowledge” - Anonymous

“Computers have promised us a fountain of wisdom but delivered a flood of data” - William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus

“Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?” - T. S. Eliot

La fouille de données

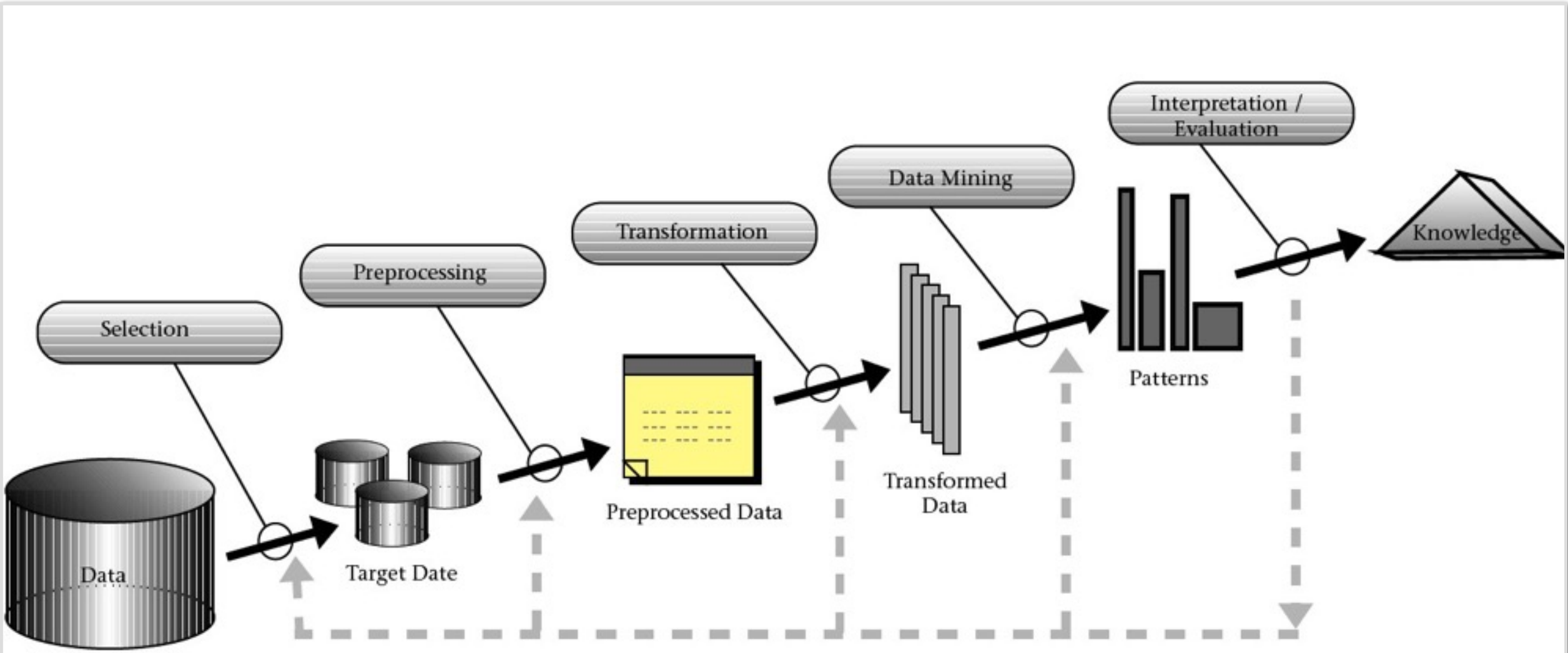
Ce que ce n'est pas

Data Mining, noun: "**Torturing data** until it confesses ... and if you torture it enough, **it will confess to anything**" - *Jeff Jonas, IBM*

"An Unethical Econometric practice of **massaging** and **manipulating** the data to **obtain the desired results**" - *W.S. Brown "Introducing Econometrics"*

Le processus ECD

(Extraction de Connaissances dans les Données)



Source : https://filteralgorithmen.wordpress.com

Méthode (1)

1. Comprendre l’application

2. Sélectionner les données

3. Les nettoyer et les transformer

-

-

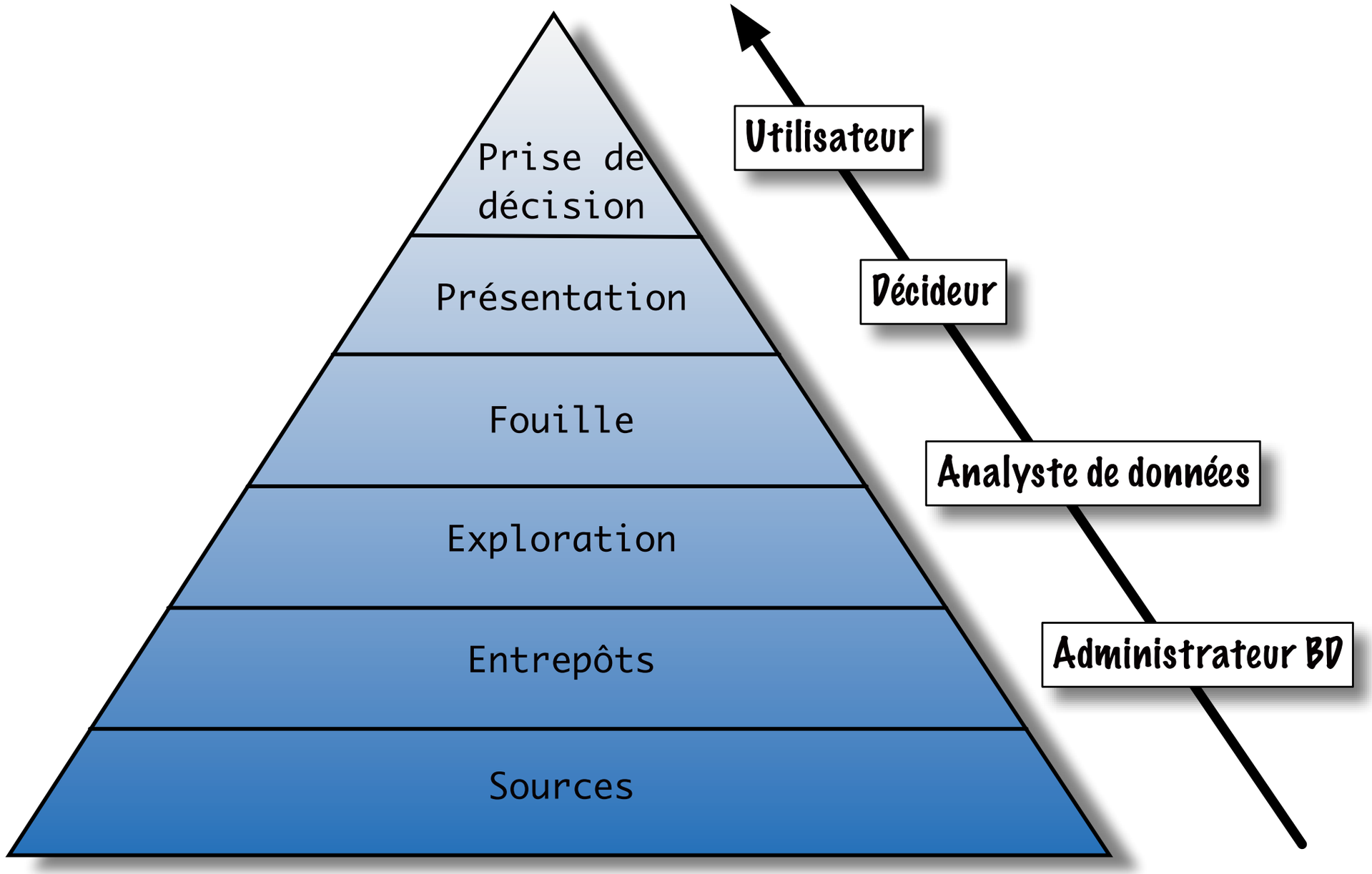
Méthode (2)

4. Appliquer une technique de fouille de données

5. Visualiser, évaluer et interpréter les résultats

- **Gérer la connaissance découverte**

Fouille et prise de décision



Plan du cours

1. Le pré-traitement des données

2. Méthodes non supervisées

3. Méthodes supervisées

4. Méthodes semi-supervisées

5. Fouille de données Web

Plan du cours

1. **Le pré-traitement des données**

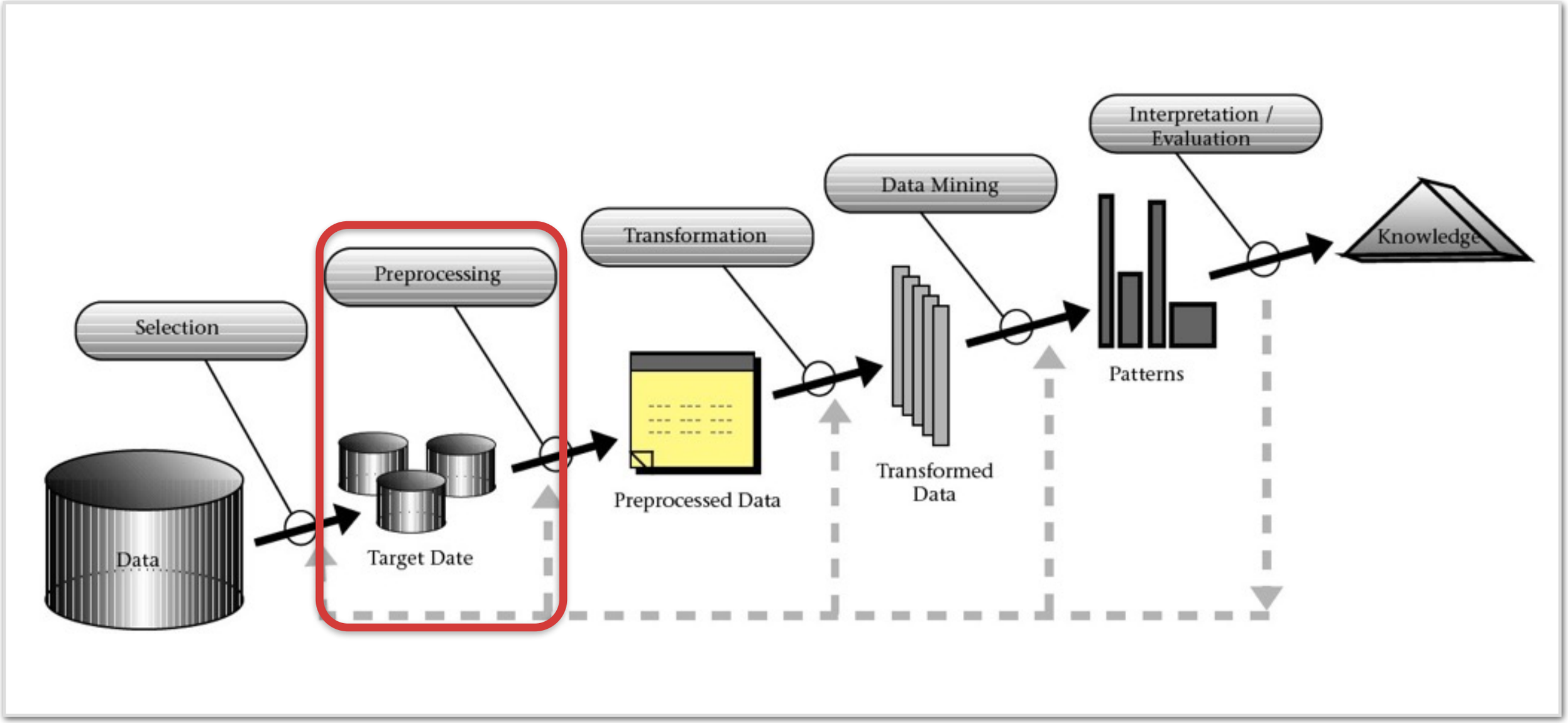
2. Méthodes non supervisées

3. Méthodes supervisées

4. Méthodes semi-supervisées

5. Fouille de données Web

Le pré-traitement



Données et Pré-traitements

- Quelques **définitions** autour des données
- **Explorer** les données pour mieux les comprendre
- Mesurer la **qualité** des données
- **Pré-traitement**
  - Agrégation
  - Echantillonnage
  - Discrétisation
  - Création / transformation / réduction d'attributs

Données

Définitions

- **Essentiel de se poser des question autour du jeu de données manipulé**
  -
  -
  -
  -
  -

Données

Définitions

-
-
-
-
-
-

Temp.	Ciel	Humidité	Vent	Jouable
29	soleil	85	FALSE	Non
26	soleil	90	TRUE	Non
28	couvert	86	FALSE	Oui
21	pluvieux	96	FALSE	Oui
20	pluvieux	80	FALSE	Oui
20	couvert	70	TRUE	Non
19	soleil	64	TRUE	Oui
22	soleil	95	FALSE	Non

Données transactionnelles

- Chaque transaction (enregistrement) fait intervenir un ensemble d'items
- Exemple classique : panier de la ménagère
- Existe dans d'autres situations : visites d'un site web, ensemble d'amis d'un réseau social, les mots d'un documents, …

Tid	Items
1	Bière, Coca, Chips
2	Vin rouge, boeuf, Pain
3	Pain, Bière
4	Lait
5	Lait, Pain

Données transactionnelles

- Les données transactionnelles peuvent être représentées par une **matrice creuse**
- Une ligne = une transaction
- Les attributs sont **binaires** et **asymétriques**

Tid	Bière	Coca	Chips	Vin rouge	Boeuf	Pain	Lait
1	1	1	1	0	0	0	0
2	0	0	0	1	1	1	0
3	1	0	0	0	0	1	0
4	0	0	0	0	0	0	1
5	0	0	0	0	0	1	1

Propriétés des valeurs d'attributs

- Le type d'un attribut dépend des propriétés qu'il possède
  - Distinction
  - Ordre
  - Addition
  - Multiplication
- Attribut
  - **Nominal** :
  - **Ordinal** :
  - **Intervalle** :
  - **Ratio** :

Types d'attributs

Catégoriel

- **Nominal** :
  -
- **Ordinal** :
  -

Numérique

- **Intervalle** :
  -
- **Ratio** :
  -

Attributs discrets,continus,asymétriques

- **Attribut discret**
  - Ensemble dénombrable ou infini dénombrable de valeurs
  - Souvent représenté par des entiers
  - Nominal, ordinal, binaire
- **Attribut continu**
  - Les valeurs sont des nombres réels
  - Intervalle, ratio
- **Attribut asymétrique**
  - Seule la présence importe

Explorer le jeu de données

Mieux appréhender les caractéristiques des données

-
-
-

Que faut-il regarder ?

- Distribution des classes
- Dispersion des valeurs
- Asymétrie, outliers,valeurs manquantes
- Corrélations
- …

Ne pas oublier de visualiser

-
-
-
-

Classe équilibrée

De nombreux jeux de données ont un attribut de classe discret (parfois binaire)

- Quelle est la fréquence de chaque classe ?
- Il y a-t-il un fort déséquilibre ?

A quoi cela sert-il ?

-
-
-

Quelques statistiques utiles

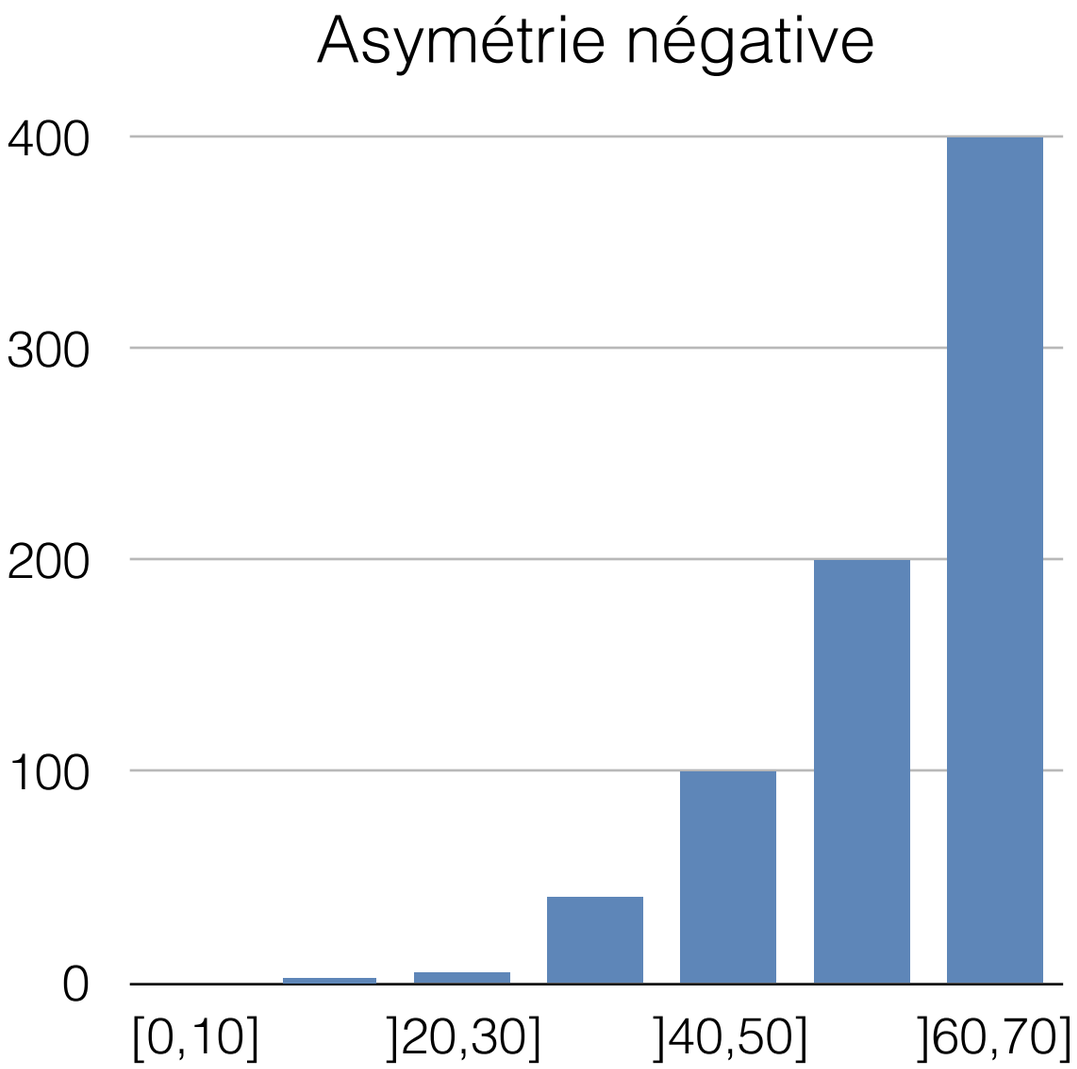
Attribut discret

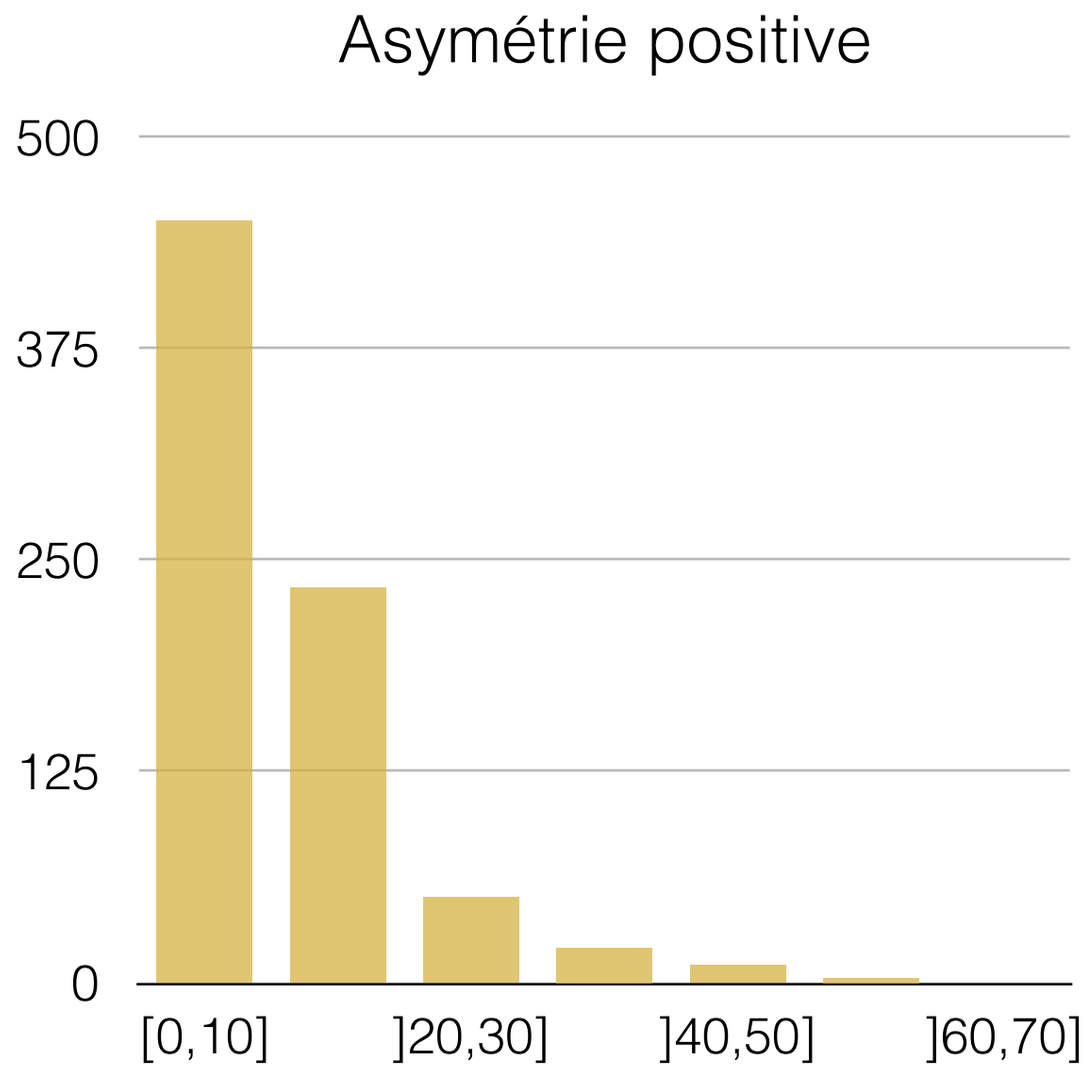
-
-

Attribut continu

- Intervalle de valeur
- Moyenne (sensible aux outliers)
- Médiane
- La distribution sera asymétrique si la médiane et la moyenne sont *assez* différentes

Distribution asymétrique





Résumé en 5 nombres

Uniquement pour les attributs numériques

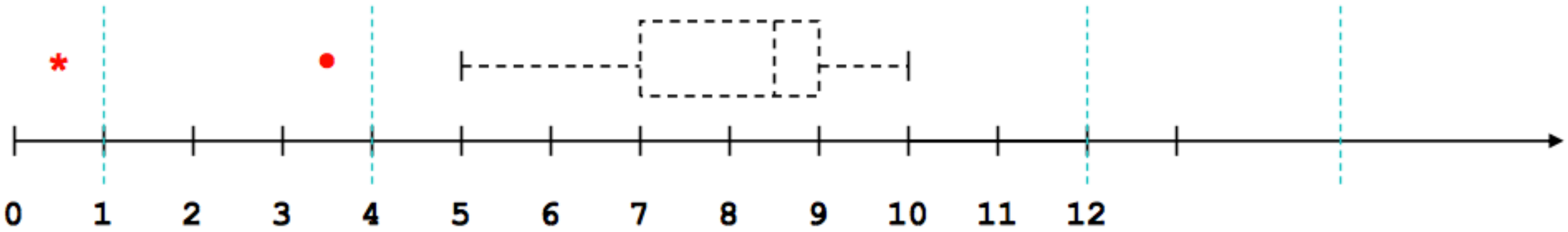
Résumé = (minimum, Q_1, Q_2, Q_3, maximum)

Exemple

-
-
-
-
-
-

Boîtes à moustaches

Prise en compte des outliers



- Q1 = 7 / Q2 = 8,5 / Q3 = 9 / Ecart interquartile (EI) = 2
- Plus grand non-outlier : 10
- Plus petit non-outlier : 5
- Outlier intermédiaire (OI) : 3,5

$$Q_1 - \mathbf{3 \times 1,5}\ EI \leq OI \leq Q_1 - \mathbf{1,5}\ EI$$

$$Q_3 + \mathbf{1,5}\ EI \leq OI \leq Q_3 - \mathbf{3 \times 1,5}\ EI$$

- Outlier extrême (OE) : 0,5

$$EI < Q_1 - \mathbf{3 \times 1,5}\ EI$$

$$EI > Q_3 + \mathbf{3 \times 1,5}\ EI$$

Mesurer la dispersion et les corrélations

Dispersion

- Via le calcul de la variance (sensible aux outliers)
- Via l'écart interquartile
- Si la distribution des valeurs est multimodale, utiliser des outils de visualisation

Corrélations

- Utilisation de scatter plot pour visualiser les données
- Calcul du coefficient de corrélation
- Si corrélation forte, réfléchir à l'utilité de garder les deux attributs

Pourquoi pré-traiter les données ?

- Données réelles souvent :
  - **Incomplètes**

    Valeurs manquantes, manque certains attributs importants, données trop agrégées, …

  - **Bruitées**

    Présence d'erreurs ou d'outliers

  - **Incohérentes**
    - Date de naissance = « 04/11/1990 » et Âge = « 50 »
    - Notes tantôt « 1, 2, 3 » tantôt « A, B, C »

-

- Environ 90% du travail en fouille de données…

Qualité des données

Définition

- **Pas de définition précise** d'une donnée de qualité …
- … mais une **vision multidimensionnelle** qui fait consensus :

Précise

Interprétable

Accessible

Principales étapes du pré-traitement

1. **Nettoyage des données**

2. **Intégration**

3. **Transformation**

   Normalisation, agrégation

4. **Création d'attribut**

5. **Réduction**

   Obtention d'une représentation condensée des données sans dégradation majeure des résultats

6. **Discrétisation**

Nettoyage des données

Pourquoi faire ?

Un des plus gros problème lorsque l'on gère des données (bases de données, entrepôts de données, fouille de données)

En quoi cela consiste ?

-

-

-

-

Nettoyage des données

Données manquantes

Certains attributs n’ont pas de valeurs

Les profils utilisateurs de sites internets sont souvent incomplets

Pourquoi ?

- Equipement défectueux
- Incohérences avec d’autres données
- Données volontairement non saisies ou mal comprises
- …

Nettoyage des données

Données manquantes

Ignorer le n-uplet

Pas plus de 5% des enregistrements

Remplir manuellement (expert du domaine)

Trop fastidieux

Choisir un algorithme de fouille qui fonctionne avec des données manquantes

Arbres de décision par exemple

Exercice nettoyage

Consignes

- Combler les trous via :
  - La moyenne générale
  - La moyenne par classe

Id	A	Classe
1	2	X
2	?	X
3	30	Y
4	?	Y
5	3	X
6	15	Y
7	?	X
8	12	Y
9	4	X

Nettoyage des données

Gestion du bruit

Définition

Erreur aléatoire ou grande variance dans une variable

Pourquoi ?

D’autres problèmes

- Données incomplètes
- Données incohérentes

Nettoyage des données

Gestion du bruit

Partitionnement

Tri puis partition (equi-freq) des données

On les lisse (moyenne, médiane)

Clustering

Inspection humaine et automatique

Coupler analyser automatique et humaine

Détection des valeurs suspectes et vérification humaine

Regression

Lisser les données aberrantes grâce à des fonction de regression

Nettoyage des données

Gestion du bruit - Exercice

Données

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

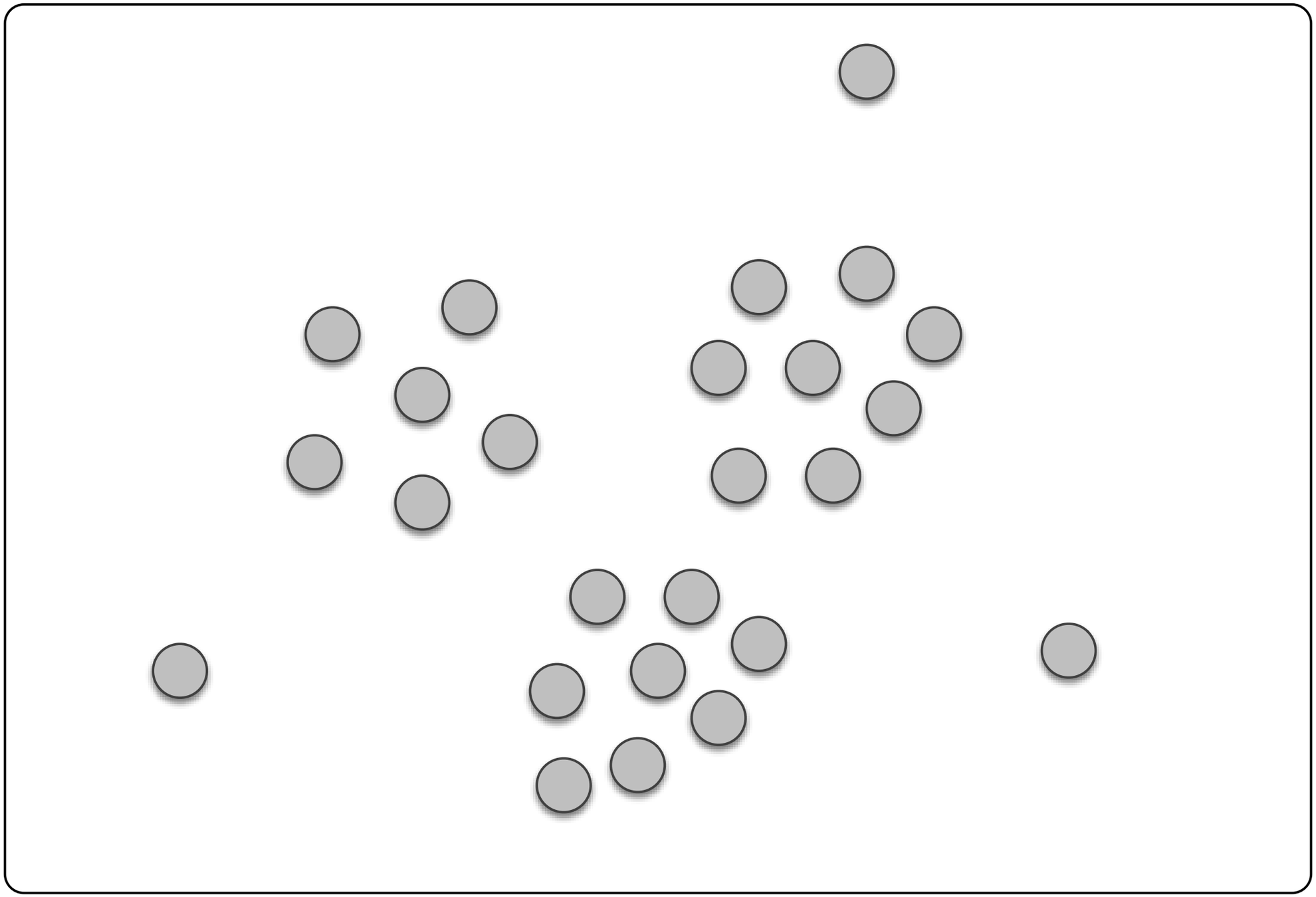
Consignes

Partitionnement en 3 intervalles

Lissage en utilisant la moyenne et la médiane

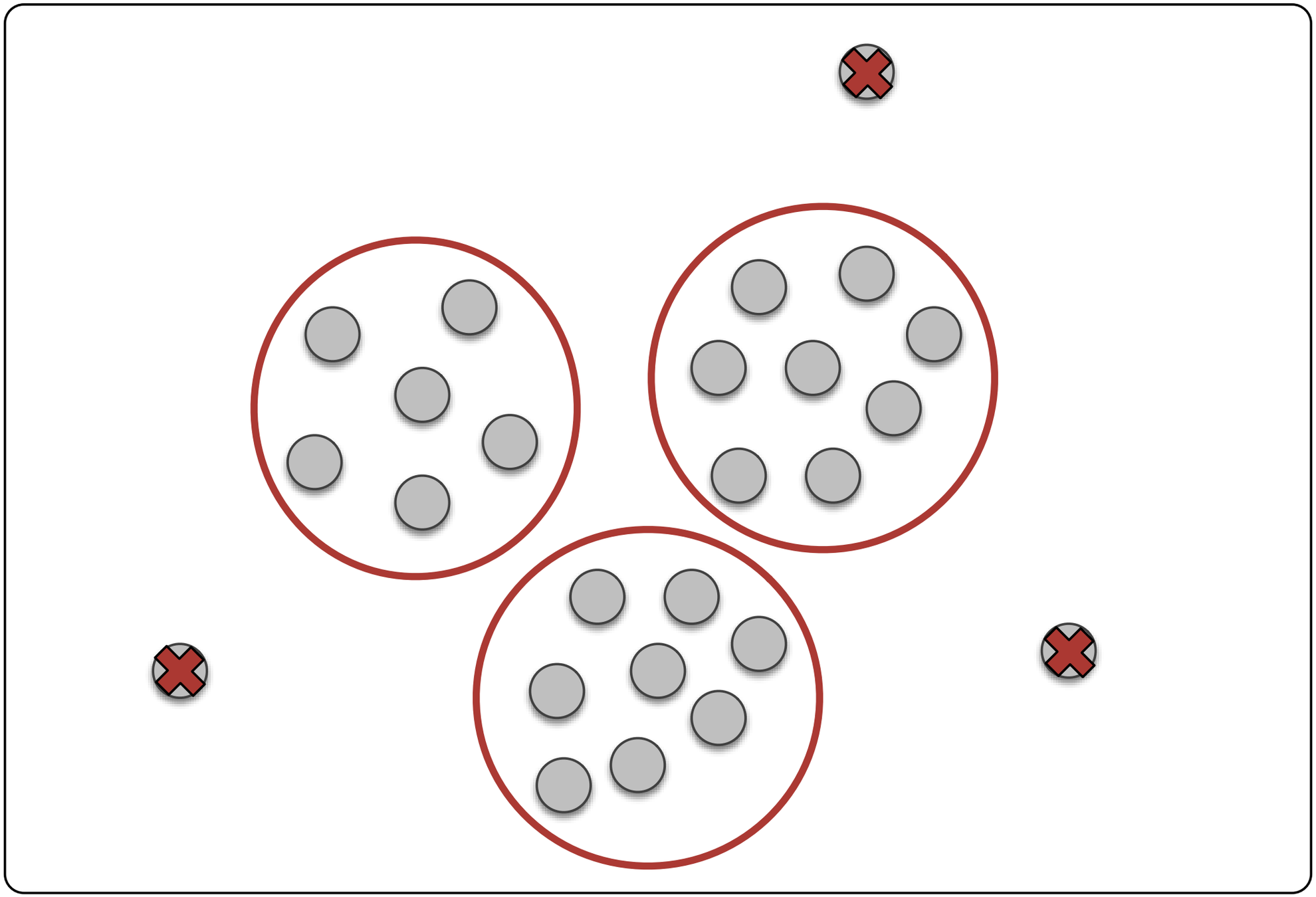
Nettoyage des données

Gestion du bruit - Clustering



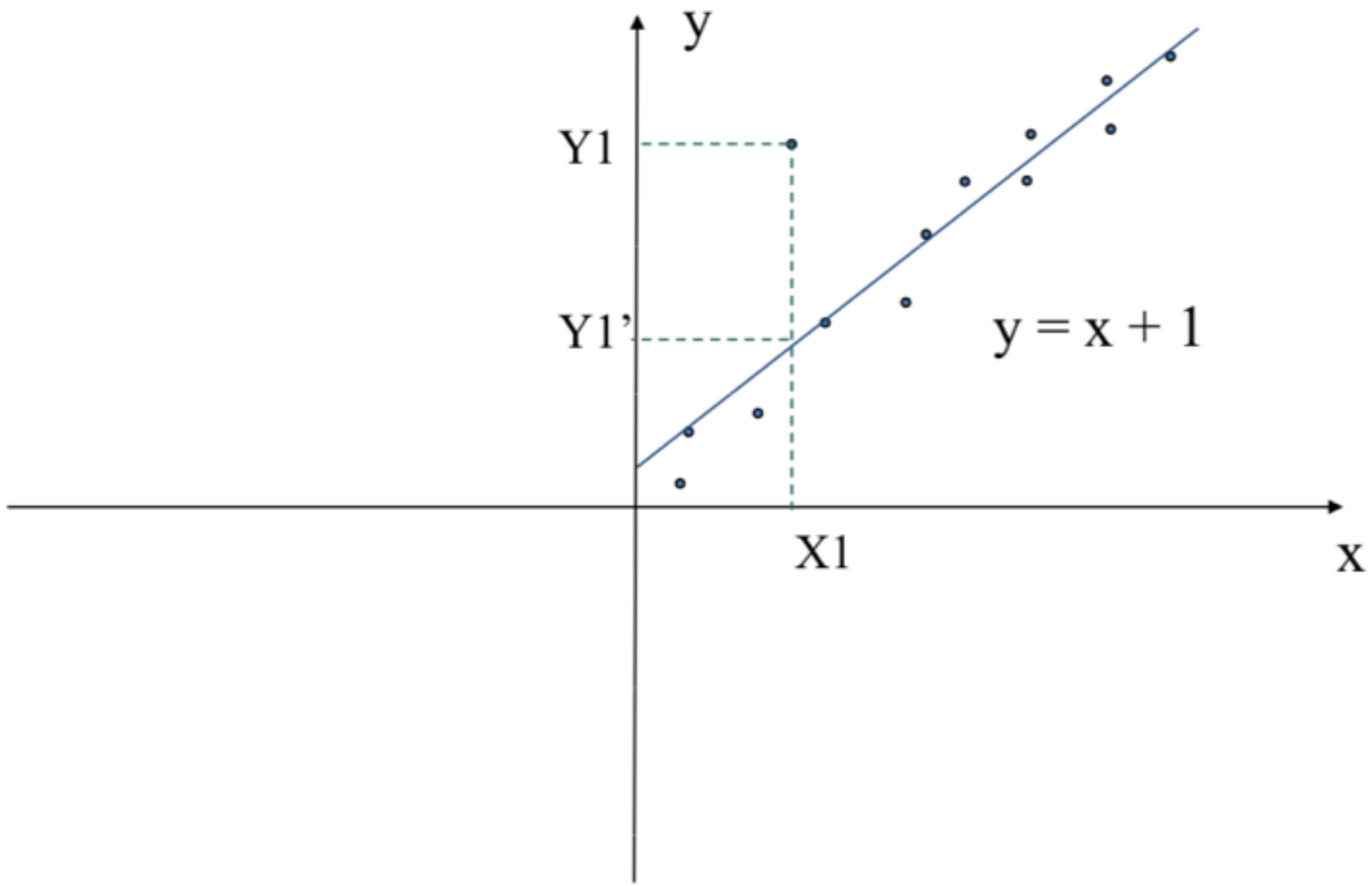
Nettoyage des données

Gestion du bruit - Clustering



Nettoyage des données

Gestion du bruit - Régression



Nettoyage des données

Gestion du bruit - Régression - Exercice

Règle

Modification si |B-f(A)|>3

Consignes

Estimez f(A) visuellement

Corrigez si nécessaire

Id	A	B	B’
1	2	6	?
2	1	-8	?
3	3	6	?
4	8	17	?
5	3	15	?
6	9	18	?
7	3	8	?
8	6	14	?
9	4	2	?

Nettoyage des données

Suppression des outliers

Définition

Différents types d’outliers

- Valide : le salaire d’un chef d’entreprise
- Bruit : quelqu’un qui aurait 200 ans

Techniques d’élimination

-
-
-

Intégration des données

Intégration de données

Combinaison de différentes sources de données en une seule

Intégration de schémas

Détecter et résoudre les conflits de valeurs

Valeur réelle identique mais représentation différente

Cause : échelles différentes, …

Gestion de la redondance

Transformation des données

Lissage

Réduire le bruit (similaire à la phase de nettoyage)

Agrégation

Simplification, construction de cubes de données

Généralisation

Utilisation de hiérarchies de concepts

Normalisation

Permet de « faire rentrer » les valeurs dans un intervalle

- Min-max
- z-score
- mise à l'échelle décimale

Transformation des données

Min-max

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\max_A^{new} - \min_A^{new}) + \min_A^{new}$$

Z-score

$$v' = \frac{v - avg_A}{std_A}$$

Mise à l'échelle décimale

$$v' = \frac{v}{10^j} \; avec \; j \; le \; plus \; petit \; entier \; tq \; \max(\mid v' \mid < 1)$$

Transformation des données

Exercice

Transformez A :

- min = 0 / max = 2
- z-score (avg=4.9 / std = 2,98)
- Mise à l' échelle décimale

Id	A
1	2
2	1
3	3
4	8
5	3
6	9
7	3
8	6
9	4
10	10

Création d'attributs

Pourquoi ?

Mieux capturer les informations essentielles du jeu de données

Comment ?

- Extraction de features
  - Dépendant du domaine
  -
- Construction de features
  - Combinaison de features existants
  -

Réduction de données

Pourquoi ?

Potentiellement très long sur données complètes

Objectif

Obtenir une représentation réduite des données qui garantit à peu près les mêmes résultats

Réduction de dimensions

Pourquoi ?

- De nombreux algorithmes ne passent pas à l'échelle en grandes dimensions
- Résultats plus facilement interprétables et visualisables
- Focus sur les attributs les plus intéressants
- Peut aider à la réduction du bruit

Comment ?

- Considérer chaque attribut indépendamment
- Considérer des sous-ensembles d'attributs

Dimensions non pertinentes et redondantes

Dimensions non pertinentes

- Aucune information utile pour la tâche de fouille de données
- *Le numéro de l'étudiant n'est pas utile pour prédire sa note*

Dimensions redondantes

-

-

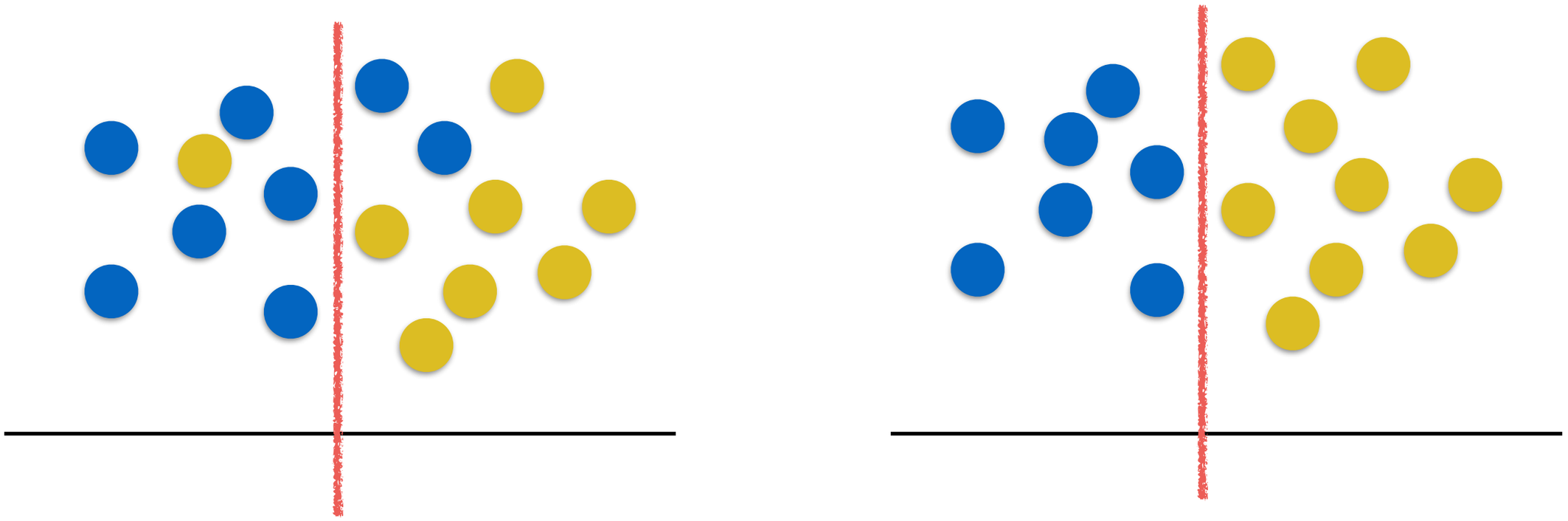
-

Considérer chaque attribut individuellement

1. Déterminer à quel point chaque attribut est discriminant pour la classe
   - Plusieurs mesures sont possibles
   - Nous utiliserons le Gain d'Information (très populaire)
2. Classer les attributs selon cette mesure
3. L'utilisateur peut choisir d'éliminer certains attributs selon cette mesure
   - *Ne retenir que les 10 plus discriminants par exemple*

Gain d'Information

Intuition

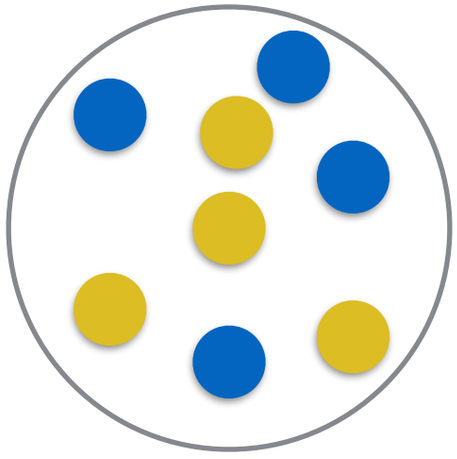


Quel test est le plus informatif ?

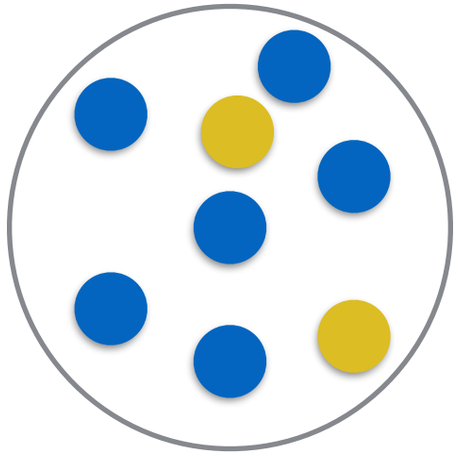
Gain d'Information

Impureté / Entropie

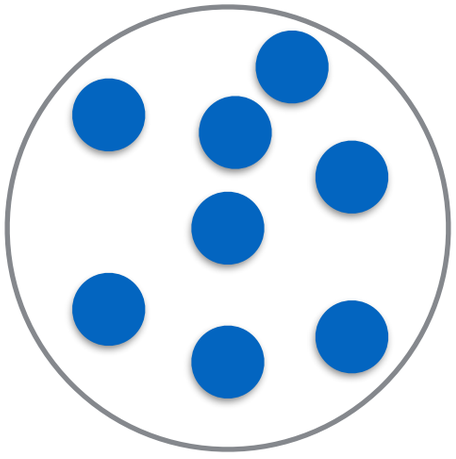
Mesure le niveau d'impureté dans un groupe d'exemples



Très impur



Un peu moins impur



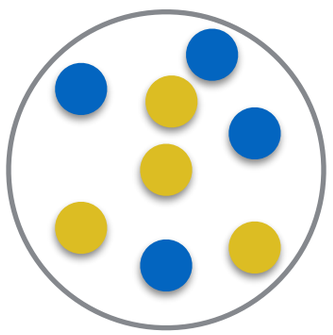
Pur

Gain d'Information

Entropie

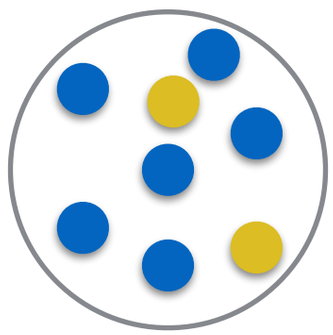
Formellement : $Entropy = \sum_i -p_i \times log_2(p_i)$

Avec p_i la probabilité de la classe i (calculée comme la proportion d'apparition de la classe i dans l'ensemble)



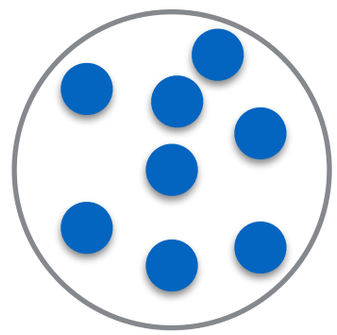
$$Entropy = -\frac{4}{8} \times log_2(\frac{4}{8}) - \frac{4}{8} \times log_2(\frac{4}{8}) = -0,5 \times -1$$

$$-0,5 \times -1 = 1$$



$$Entropy = -\frac{2}{8} \times log_2(\frac{2}{8}) - \frac{6}{8} \times log_2(\frac{6}{8}) = -0,25 \times -2$$

$$-0,75 \times -0,41 = 0,16$$



$$Entropy = -1 \times log_2(1) = 0$$

Gain d'Information

Principes

Objectif

Déterminer quel attribut dans un jeu de données est le plus utile pour séparer les valeurs d'un attribut classe cible

Formule

$$Gain(D, A_i) = Entropy(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Entropy(D_j)$$

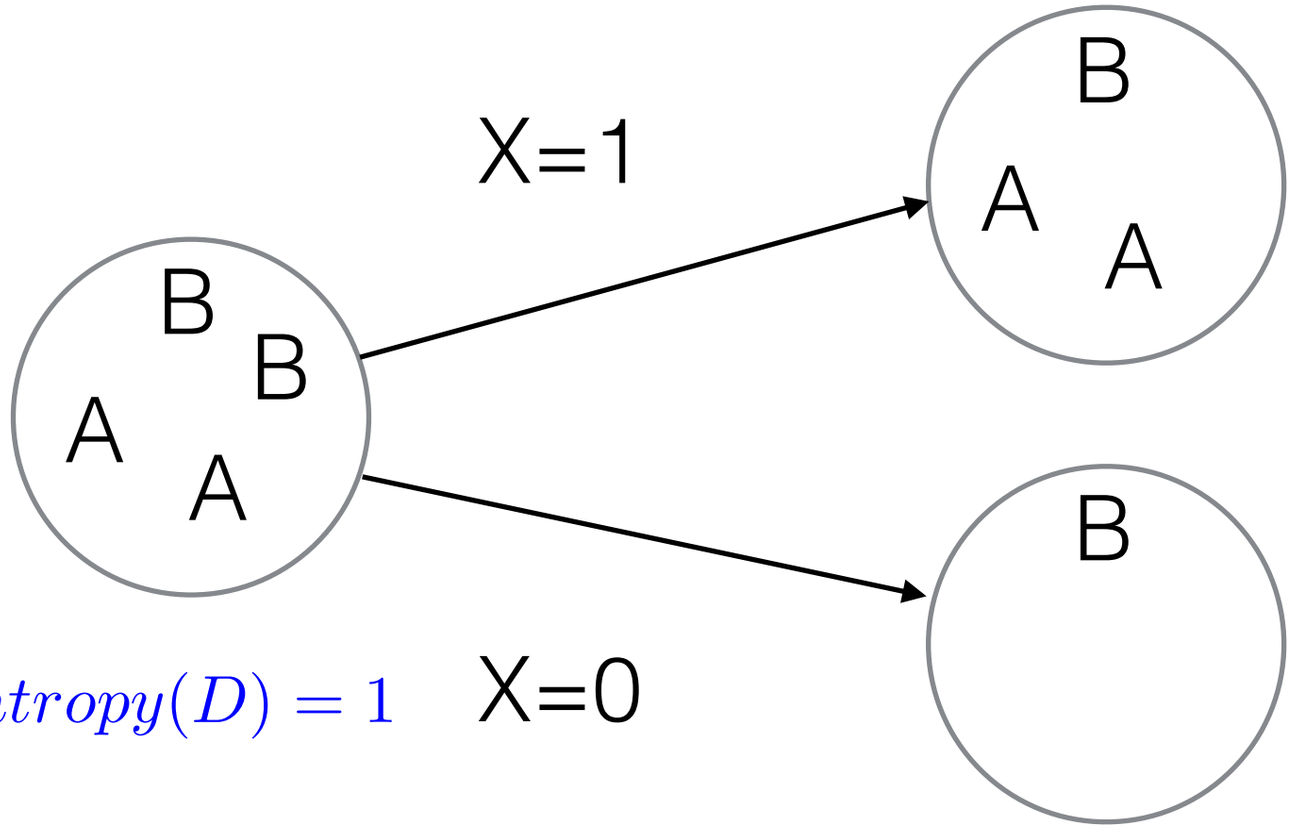
Exemple

X	Y	Z	Classe
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

Gain d'Information

Exemple

Calcul du gain d'information de l'attribut X



$Entropy(D) = 1$

$$Gain(D, X) = 1 - (\frac{3}{4} \times 0,9184 + \frac{1}{4} \times 0) = 0,3112$$

Gain d’Information

Exercice

Calculez le gain d’information pour les attributs Y et Z et concluez.

Réduction de données

Méthodes paramétriques

Suppose que les données suivent un modèle
On estime les paramètres du modèle et on ne stocke qu'eux

Méthodes non paramétriques

-
-
-
-

Réduction de données

Echantillonnage

Principe

- Ne conserver qu'un sous-ensemble des données

Approche naïve

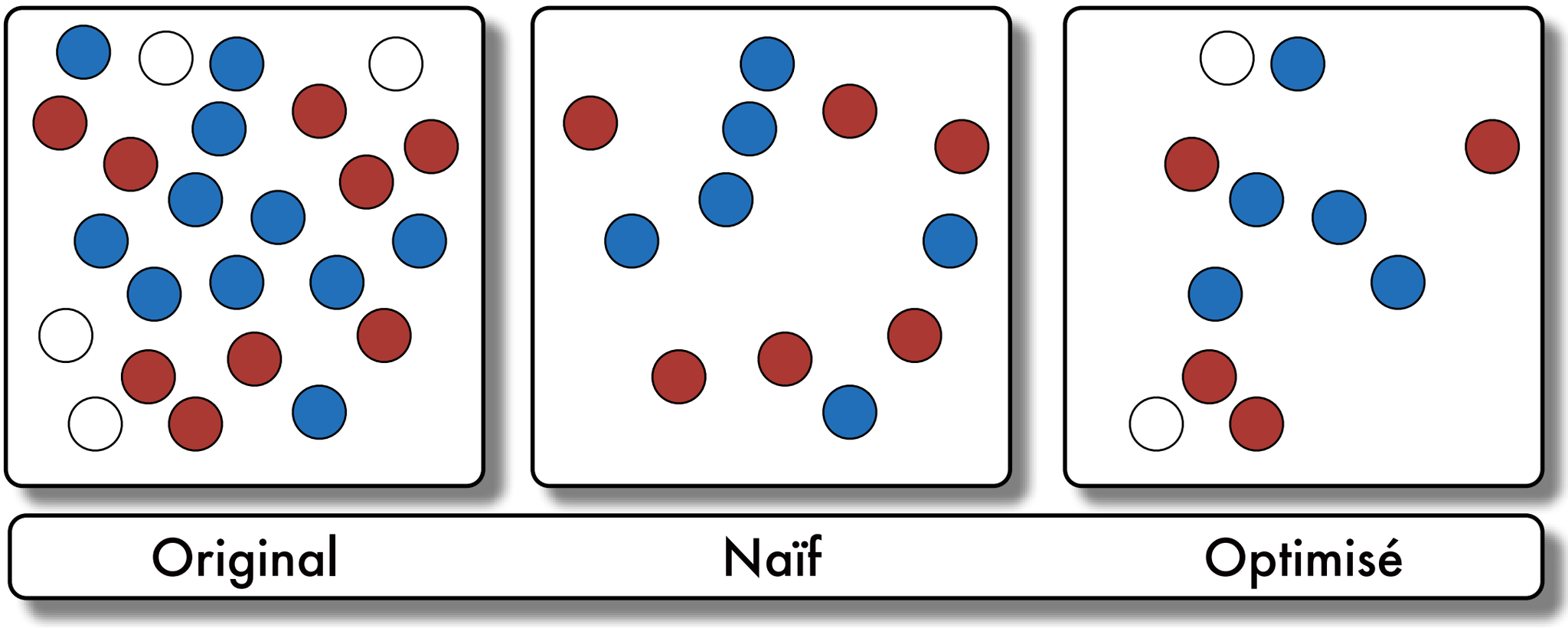
-
-

Approche plus fine

- Estimer la répartition par classe
- Sélection le sous-ensemble en fonction de la répartition

Réduction de données

Echantillonnage



Réduction de données

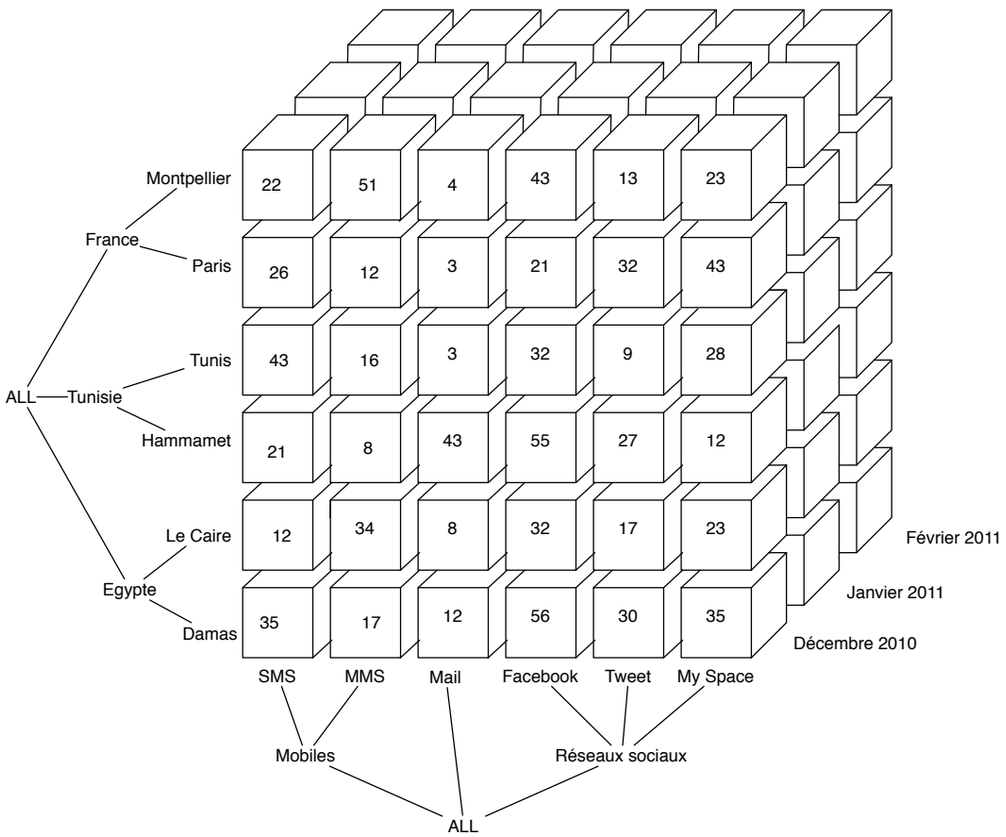
Agrégation

Hiérarchies de concepts

- Remplacer données fines par données de plus haut niveau
- Sur données numériques

  Possibilité de discrétisation multi-couche
- Sur données catégorielles :

Utilisation d'une hiérarchie existante



Du cuboïde (Ville,Média,Mois) au cuboïde (Ville,TypeMédia,Mois)

\+ Réduction de données
\+ Moins sensibles au bruit
\+ Patterns généraux plus identifiables

Réduction de données

Discrétisation

Principe et objectif

Transformer un attribut continu en attribut catégoriel

Certains algorithmes ne fonctionnent qu'avec des données catégorielles

Possibilité d'obtenir de meilleurs résultats avec des attributs discrédités

Approche non supervisée

- Equi-width — On ne tient pas compte de la classe à prédire
- Equi-freq — Le nombre de classes est déterminé expérimentalement

Approche supervisée

- Approche basée sur l'entropie
- Maximisation de la pureté pour chaque intervalle

Réduction de données

Equi-width

Principe

Divise les valeurs de X en k intervalles de taille égale

Approche non supervisée

- Si x_{min} et x_{max} sont les bornes de la variable alors la taille de l'intervalle, notée δ, est définie comme :

$$\delta = \frac{x_{max} - x_{min}}{k}$$

avec les bordes des intervalles définis comme :

$$x_{min} + i \times \delta \; pour \; i \in 1, \ldots, k-1$$

Inconvénient

- Sensible aux outltiers

Réduction de données

Equi-freq

Principe

Chacun des k intervalles contient le même nombre de valeurs

Inconvénient

Si une valeur apparaît de nombreuses fois, elle sera présente dans plusieurs intervalles

Réduction de données

Basée sur l'entropie

Principe

Déterminer chaque intervalle pour qu'il soit le plus pur possible

Algorithme

1. S= Ordonner la variable X
2. Discrétiser(S)

Discrétiser(S)

Tant que critère d'arrêt(S) == Faux

(gauche,droite) = trouverMeilleurPoint(S)

Discrétiser(gauche)

Discrétiser(droite)

Références

Ces ouvrages pointent vers de nombreuses références d'articles scientifiques décrivant les approches vues en cours ou des variantes de celles-ci

- Data Mining - Concepts and Techniques par J. Han et M.Kamber (ed. Morgan Kauffman)
- Web Data Mining - ExploringHyperlink, Contents and Usage Data par B. Liu (ed. Springer)
- Statistiques Exploratoires Multidimensionnelles par L. Lebart et al. (ed. Dunod)