

Introduction aux méthodes pour la science des données

HMIN232M

Pascal Poncelet
LIRMM

Pascal.Poncelet@lirmm.fr
<http://www.lirmm.fr/~poncelet>

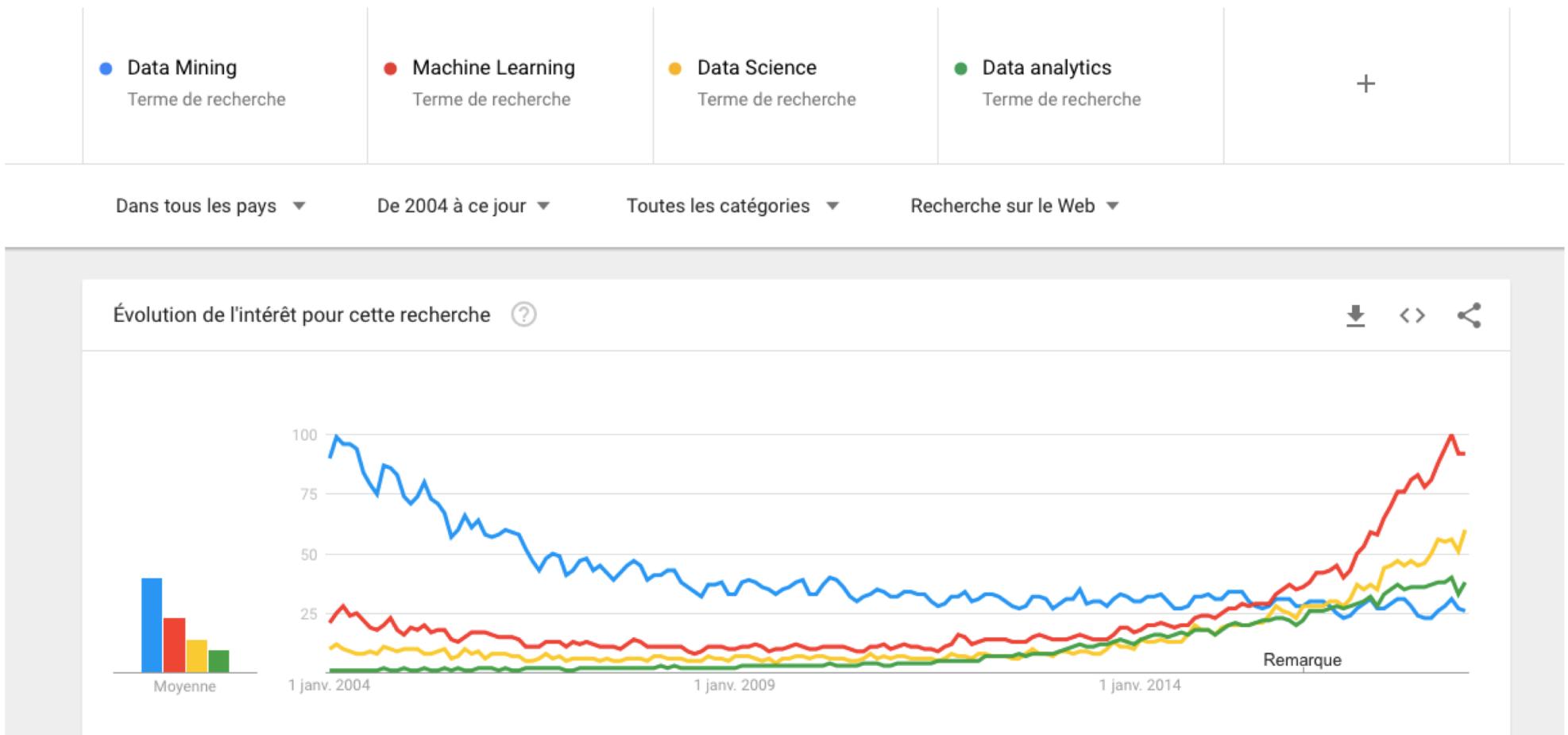


Quel vocabulaire ?

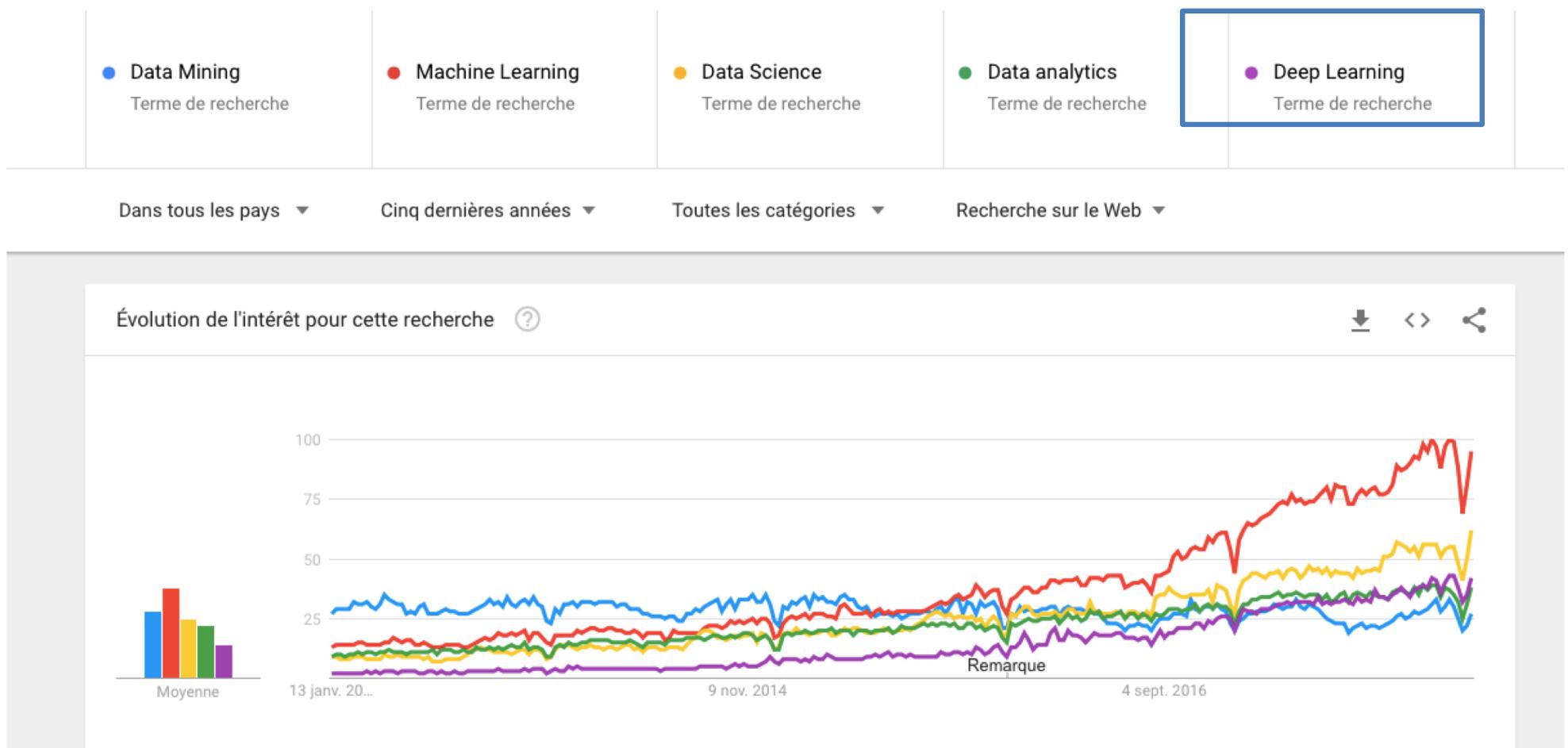
- Historiquement :
 - de nombreux « points de départ »
- Domaine récent dont le vocabulaire n'est pas fixé
 - Data Mining, Machine Learning, KDD, Data Science, Big Data, Deep Learning, ...
- Evolution rapide
- domaine applicatif *versus* domaine de recherche *versus* domaine marketing



Les tendances

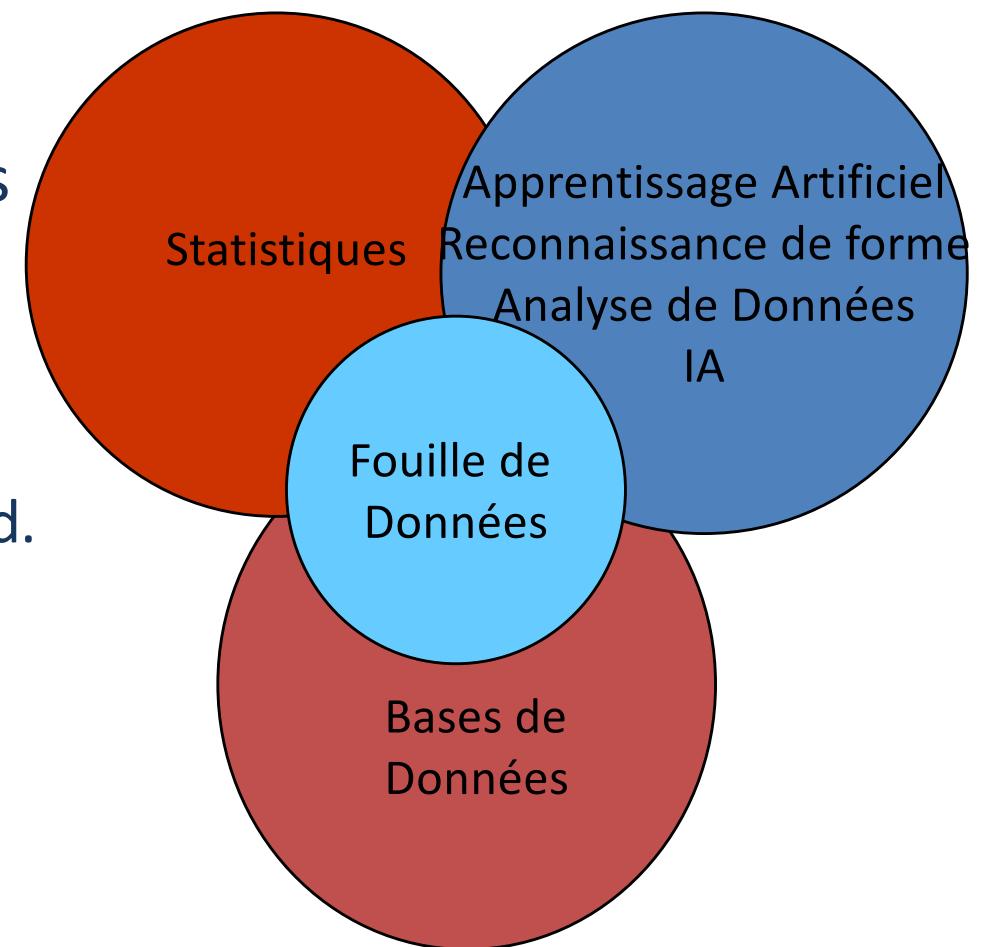


Les tendances (5 dernières années)



L'intersection de nombreux domaines

- Proviennent des techniques de machine learning / Intelligence artificielle, extraction de motifs, bases de données
- Adaptation de ces techniques à :
 - De grandes quantités de données
 - Des données avec de nombreuses dimensions (c.a.d. attributs)
 - Données hétérogènes et distribuées



Plan

Introduction et contexte du processus de KDD

Les données et les pré-traitements

Les tâches de la fouille de données

Conclusions



Pourquoi fouiller les données ?

- De nombreuses données sont collectées et entreposées
 - Données du Web, e-commerce
 - Achats dans les supermarchés
 - Transactions de cartes bancaires
- Les ordinateurs deviennent de moins en moins chers et de plus en plus puissants
- La pression de la compétition est de plus en plus forte
 - Fournir de meilleurs services, s'adapter aux clients (e.g. dans les CRM)



Pourquoi fouiller les données ?

- Les données sont collectées et stockées rapidement (GB/heures)
 - Capteurs : RFID, supervision de procédé
 - Télescopes
 - Puces à ADN générant des expressions de gènes
 - Simulations générant de téraoctets de données

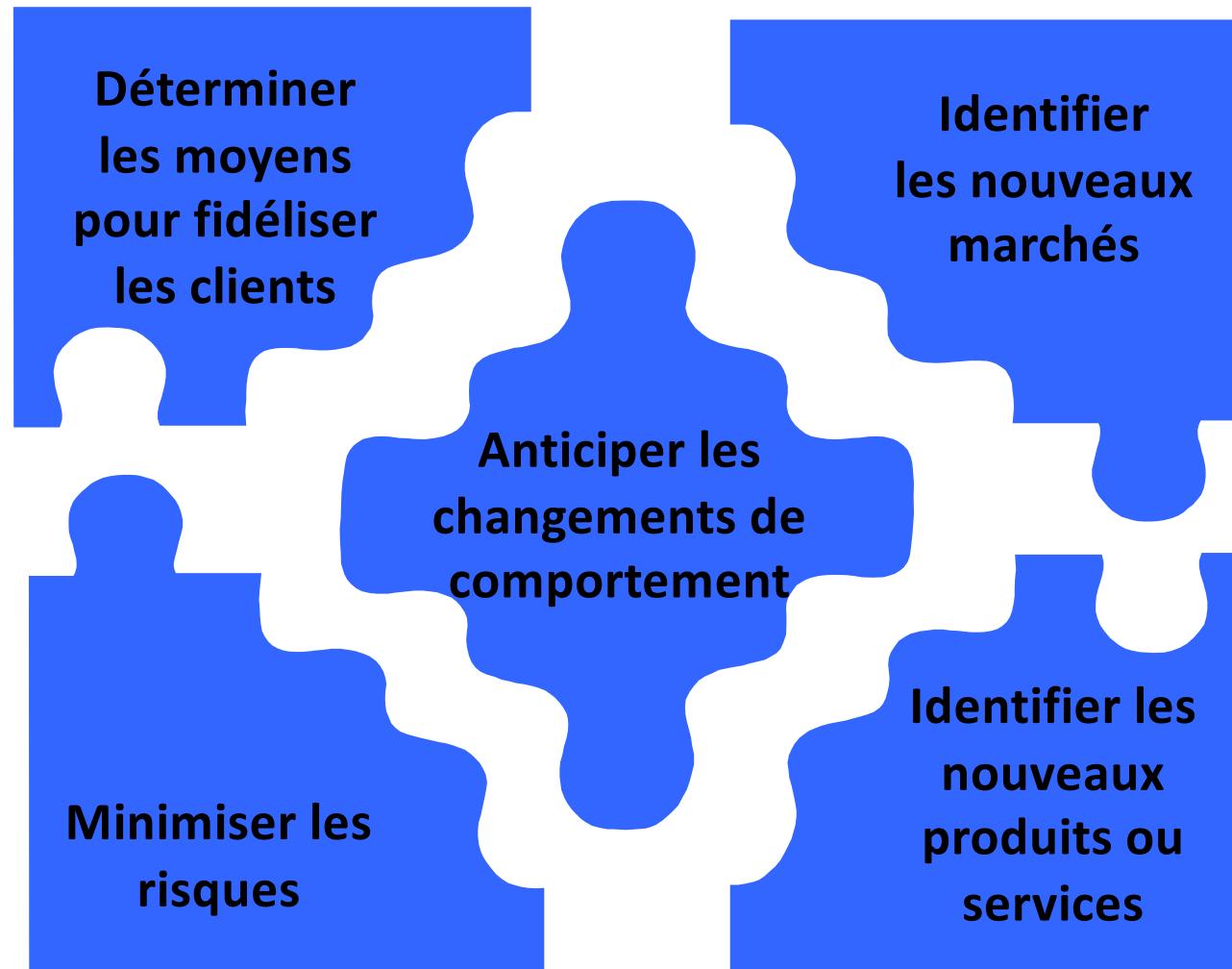


Pourquoi fouiller les données ?

- Les techniques traditionnelles ne sont pas adaptées
- Volume de données trop grands (trop de tuples, trop d'attributs)
Comment explorer des millions d'enregistrements avec des milliers d'attributs ?
- Besoins de répondre rapidement aux opportunités
- Requêtes traditionnelles (SQL) impossibles
« Rechercher tous les enregistrements indiquant une fraude »
- Croyance dans la présence de données importantes



Un enjeu stratégique

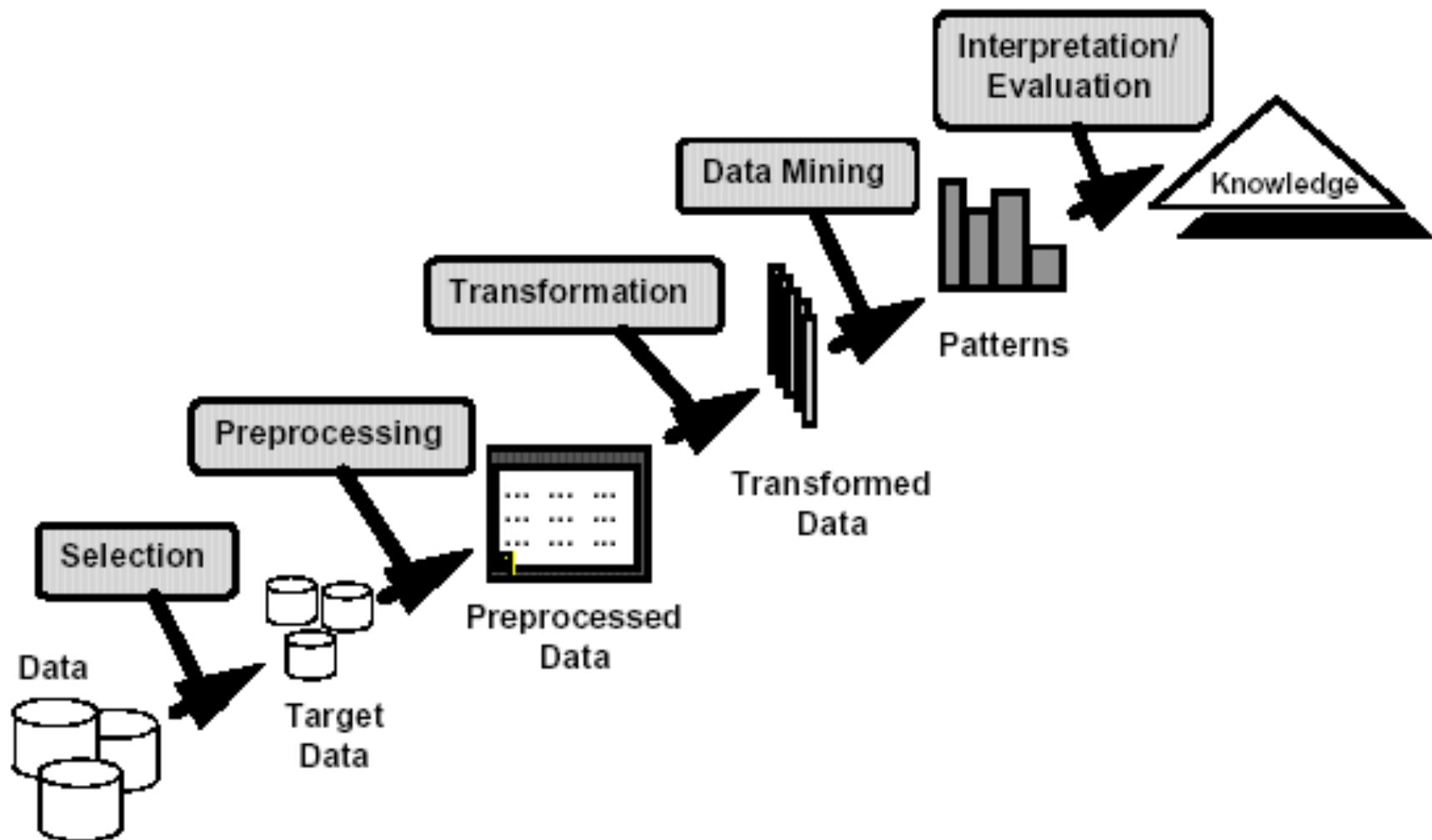


Qu'est ce que le Data Mining ?

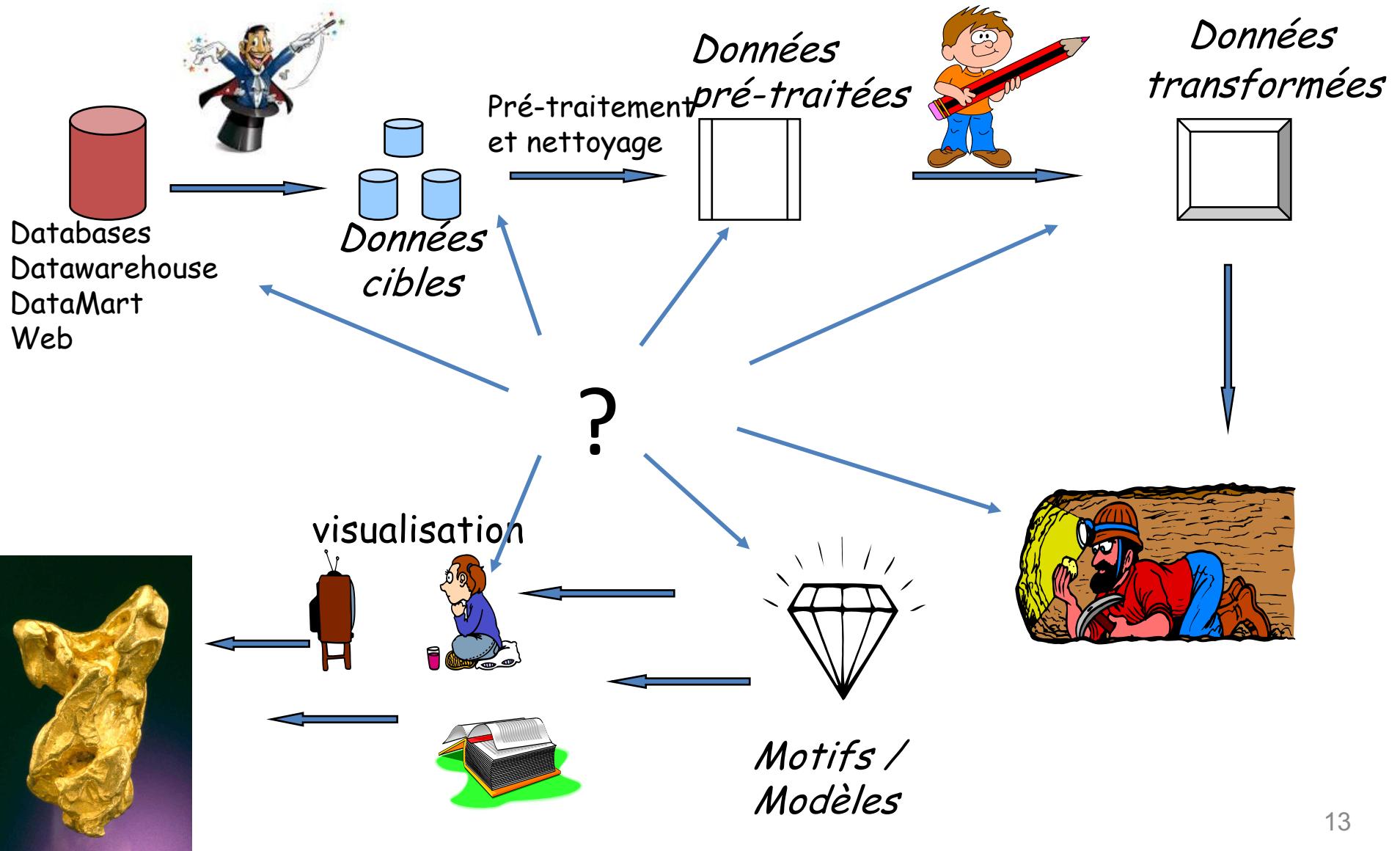
- De nombreuses définitions
 - Processus **non trivial** d'extraction de connaissances d'une base de données pour obtenir de nouvelles données, valides, potentiellement utiles, compréhensibles,
 - Exploration et analyse, **par des moyens automatiques ou semi-automatiques**, de grandes quantités de données en vue d'extraire des motifs intéressants



Le processus de KDD



Le processus de KDD



Données, Informations, Connaissances

Décision

- Promouvoir le produit P dans la région R durant la période N
- Réaliser un mailing sur le produit P aux familles de profil F

Connaissance (data mining)

- Une quantité Q du produit P est vendue en région R
- Les familles de profil F utilisent M% de P durant la période N

Information (requêtes)

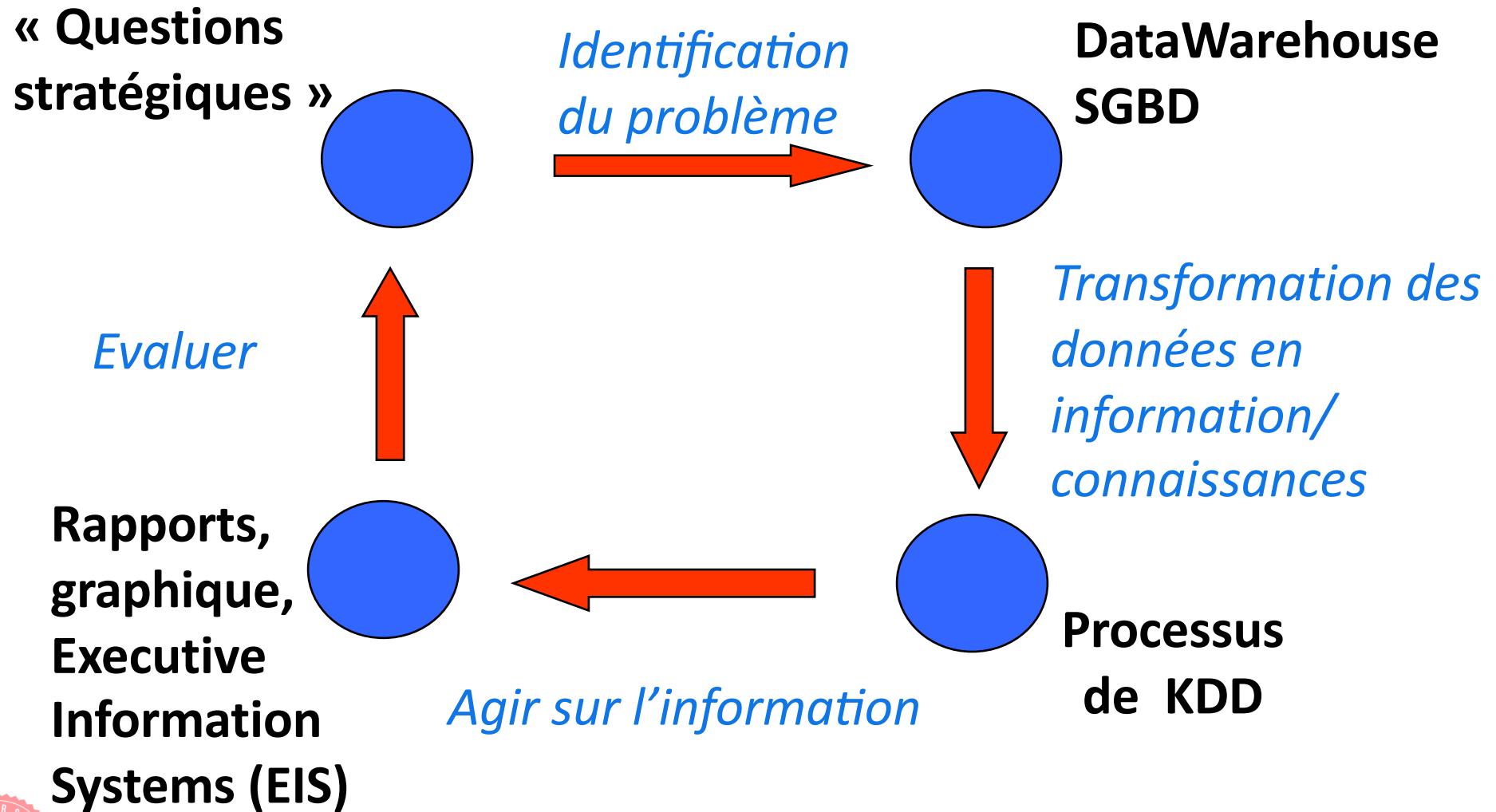
- X habite la région R
- Y a A ans
- Z dépense son argent dans la ville V de la région R

Données

- Consommateurs
- Magasins
- Ventes
- Démographie
- Géographie



Cycle de vie du KDD



Data Mining ou non ?

- **NON**

Rechercher le salaire d'un employé

Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining

- **OUI**

Les supporters achètent de la bière le samedi et de l'aspirine le dimanche

Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu



Un exemple d'analyse et de fouille de données

- 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.
- 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 clients et prévoir des mesures pour les diminuer ?

Source Data Mining. Adrian&Zantig 1996



Un exemple d'analyse et de fouille de données

- Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?
 - Une requête SQL à partir des données suffit
- A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?
 - Nécessite de garder toutes les dates de souscription, même pour les abonnements résiliés. Requêtes multidimensionnelles éventuellement de type OLAP.



Un exemple d'analyse et de fouille de données

- Est-ce que les acheteurs de magazine de musique sont aussi amateurs de cinéma ?
 - Exemple simplifié de problème où l'on demande si les données vérifient une règle : on connaît les acheteurs de magazine de musique on regarde s'ils aiment le cinéma
 - Réponse formulée par une valeur estimant la probabilité que la règle soit vraie. Utilisation d'outils statistiques ou de requêtes sur une base de données



Un exemple d'analyse et de fouille de données

- Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?
 - Question ouverte, il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser.
- Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?
 - Question ouverte : il faut disposer d'indicateurs comme durée d'abonnement, délai de paiement, ...

Il s'agit de tâches de fouille de données



Applications

- Médecine : bio-médecine, drogue, Sida, séquence génétique, gestion hôpitaux, ...
- Finance, assurance : crédit, prédition du marché, détection de fraudes, ...
- Social : données démographiques, votes, résultats des élections,
- Marketing et ventes : comportement des utilisateurs, prédition des ventes, espionnage industriel, ...
- Militaire : fusion de données .. (secret défense)
- Astrophysique : astronomie, « contact » (;-))
- Informatique : agents, règles actives, IHM, réseau, Data-Warehouse, Data Mart, Internet (moteurs intelligent, profiling, text mining, ...)



Plan

Introduction et contexte du processus de KDD

Les données et les pré-traitements

Les tâches de la fouille de données

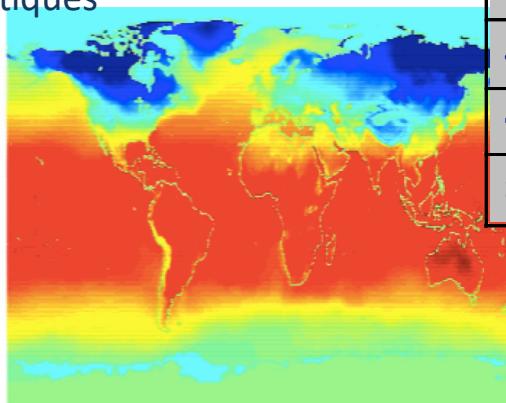
Conclusions



Les types de jeux de données

- Enregistrements
 - Tuples en Relationnel
 - Matrice de données
 - Données document : textes, vecteur de fréquences des termes
 - Données de transactions
- Graphes et réseaux
 - World Wide Web
 - Réseaux sociaux
 - Structures moléculaires
- Ordonnées
 - Données vidéo : séquences d'images
 - Données temporelles : séries temporelles
 - Données séquentielles : séquences de transactions
 - Données de séquences génétiques
- Spatiales, images et multimedia :
 - Données spatiales : cartes
 - Données Image
 - Données vidéo

	team	coach	play	ball	score	game	n	wi	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	0	2
Document 2	0	7	0	2	1	0	0	3	0	0	0
Document 3	0	1	0	0	1	2	2	0	3	0	0



TID	Items
1	Pain, Coca, Lait
2	Bière, Pain
3	Bière, Coca, Couches, Lait
4	Bière, Pain, Couches, Lait
5	Coca, Couches, Lait

Les objets données

- Les ensembles de données sont associés à des objets
- Un objet représente une entité
 - Bases de données de ventes : les clients, les objets vendus,
 - Bases de données médicales : les patients, les traitements
 - Bases de données universitaire : les étudiants, les enseignants, ...
- Aussi appelés : des exemples, des échantillons, des objets, des tuples, des instances
- Les objets sont décrits par des attributs
- Ligne de la base -> objets, colonne de la base -> attributs

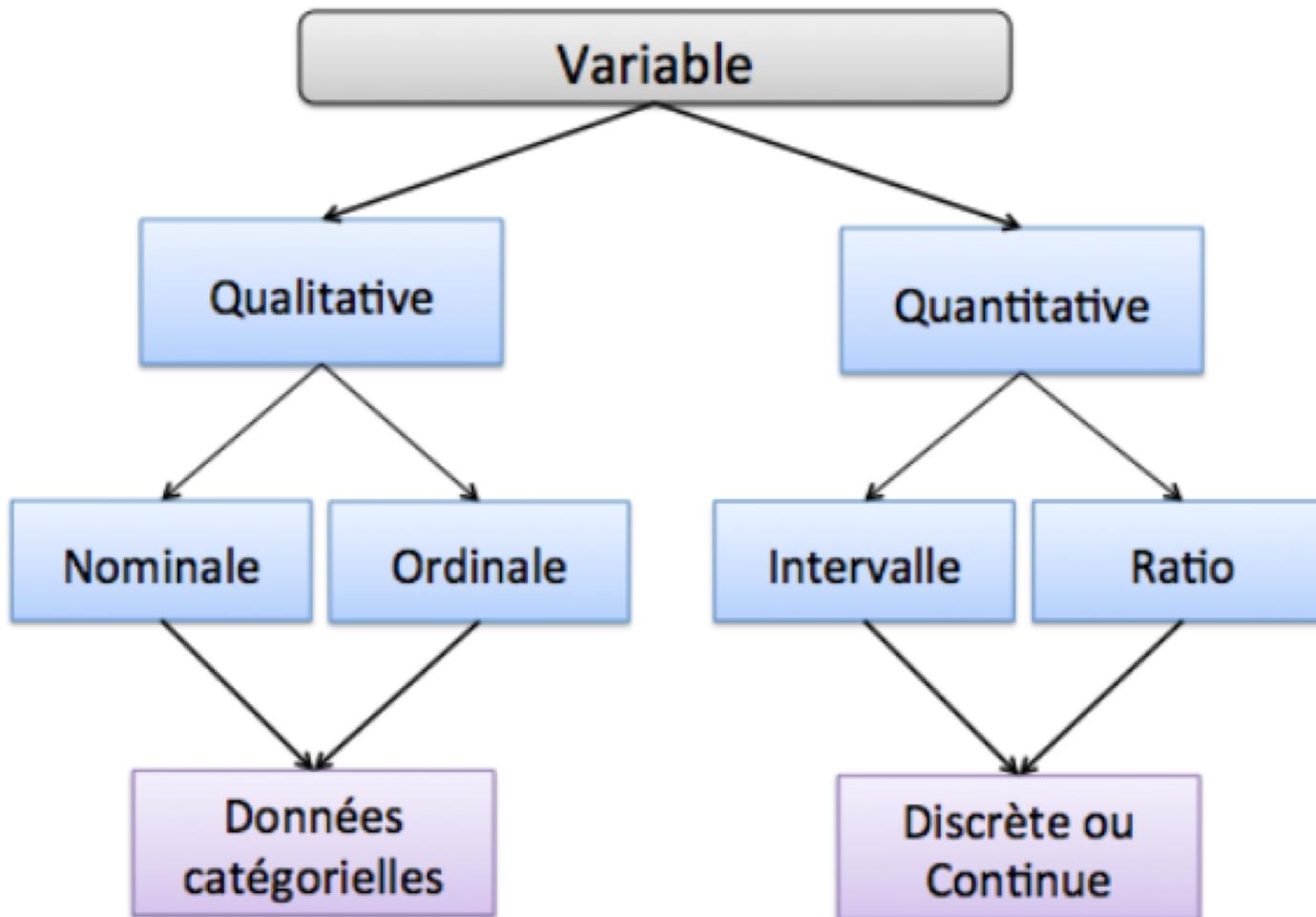


Attributs

- Un attribut (ou *dimensions, features, variable*) a un type et des valeurs contraintes par ce type
- Le type d'un attribut peut être :
 - Ordinal
 - Nominal
 - Intervalle
 - Ratio



Les différents types



Attributs à valeurs nominales

- Les valeurs sont des symboles (des noms)
 - Temps = {Ensoleillé, Pluvieux, Neigeux, Gris}
 - Code Postal, Numéro d'étudiant, couleurs yeux...
- Aucune relation (ordre ou distance) entre les nominaux n'existe
- Seuls des tests d'égalité peuvent être exécutés
- Exemple de règle:
 - If Temps = Pluvieux Then Match = No



Attributs à valeurs nominales

- Cas particulier :
- Valeurs nominales binaires : attribut nominal avec seulement 2 états (0 ou 1)
- Binaire symétrique : les deux résultats ont même importance
 - Genre (Masculin/Féminin)
- Binaire asymétrique : résultat significatif
 - Test médical (positif vs négatif)
 - Traditionnellement 1 quand résultat positif



Attributs à valeur ordinaire

- Une notion d'ordre s'impose sur les ordinaux
- Mais l'amplitude entre les valeurs n'est pas connue
- Les opérations d'addition et de soustraction ne sont pas possibles
- Exemple de règle :
 - Température décrite par les adjectifs {chaud, froid, moyen}, et chaud > moyen > froid
 - If température > froid Then match = Yes



Attributs à valeur ordinaire

- Ils peuvent être convertis en booléen

	FROID	MOYEN	CHAUD
Canada	TRUE	FALSE	FALSE
France	FALSE	TRUE	FALSE
Seychelles	FALSE	TRUE	TRUE



Attributs de type intervalle

- Les intervalles impliquent une notion d'ordre, et les valeurs sont mesurées dans des unités spécifiques et fixées
- Le point zéro n'existe pas où ne correspond en rien à l'absence de phénomène
- Exemples :
 - Le calendrier
 - La température exprimée en degrés Celsius ou Fahrenheit (0 en C -> température de congélation de l'eau, 0 en F -> température de solidification d'un mélange à part égal d'eau et de chlorure d'ammonium)



Attributs de type ratio

- Il existe un point zéro universel
- Toutes les opérations mathématiques sont autorisées sur les attributs de ce type
- Exemple:
 - L'attribut *distance* : on peut comparer, additionner 2 distances, la distance entre un objet et lui-même est zéro
 - Le poids



Attributs discrets et continus

- Une variable discrète prend un nombre fini ou dénombrable de valeurs
 - Nombre de mots dans un document, nombre d'habitants
 - Généralement un entier
- Une variable continue peut prendre un nombre infini ou non dénombrable de valeurs
 - Température, poids, taille
 - Généralement un réel



Pourquoi pré-traiter les données

- Les données du monde réel sont sales :
 - Incomplètes : manque de valeurs d'attributs, manque d'attributs intéressants, ne contenant que des données agrégées
 - métier=“ ”
 - Bruitées : contenant des erreurs ou des outliers
 - Salaire=“-10”
 - Inconsistantes : avec des incohérences dans les codes ou les noms
 - Age=“42” Anniversaire=“11/07/1990”
 - Notation initiale “1,2,3”, notation actuelle “A, B, C”
 - Incohérences entre deux enregistrements similaires



Pourquoi ?

- Incomplètes
 - Valeur pas applicable au moment de la collecte
 - Temps différent entre la collecte et l'analyse
 - Problème techniques/humain
- Bruitées (valeurs incorrectes)
 - Défaut d'instrument
 - Erreur humaine ou de l'ordinateur au moment de l'entrée
 - Erreur de transmission de la donnée
- Inconsistances
 - Différentes sources de données
 - Violation des dépendances fonctionnelles



Pourquoi le pré-traitement est important ?

- Sans données de qualité, il n'y a pas de bons résultats de fouille !
- Toujours regarder les données : 90% des échecs sont liés à la qualité des données !!
- Les étapes de recherche des données, de nettoyage, de transformation correspondent à la phase la plus longue et la plus importante du processus



Pré-traitement des données

- Nettoyer les données
 - Corrections des doublons, des erreurs de saisie
 - Contrôle sur l'intégrité des domaines de valeurs :
 - détection des valeurs aberrantes
 - détection des informations manquantes
- Intégration des données et transformation
- Réduction



Pré-traitement des données

- Correction des doublons et des erreurs de saisie

Client	Nom	Adresse	Position	Date Abonnement	Magazine
2807	Dupond	Av du Palais, Paris	Cadre	12/08/2011	Voiture
2807	Dupond	Av du Palais, Paris	Enseignant	11/07/2014	Musique
2807	Dupond	Av du Palais, Paris	Cadre	09/05/2016	BD
3456	Durand	Av de la mer, Nice	Employe	32/02/2222	BD
4356	Duchemin	Rue Principale, Grenoble	Enseignant	13/06/2015	Sport
5832	Dujardin	Place centrale, Lille	Employe	17/07/2016	NULL
2806	Durant	Rue des Chausseurs, ?	Médecin	14/04/2006	Sport
2807	Dupont	Av du Palais, Paris	Cadre	32/02/2226	Maison



Pré-traitement des données

- Intégrité de domaine ou dépendances fonctionnelles non vérifiées

Client	Nom	Adresse	Position	Date Abonnement	Magazine
2807	Dupond	Av du Palais, Paris	Cadre	12/08/2011	Voiture
2807	Dupond	Av du Palais, Paris	Enseignant	11/07/2014	Musique
2807	Dupond	Av du Palais, Paris	Cadre	09/05/2016	BD
3456	Durand	Av de la mer, Nice	Employe	32/02/2222	BD
4356	Duchemin	Rue Principale, Grenoble	Enseignant	13/06/2015	Sport
5832	Dujardin	Place centrale, Lille	Employe	17/07/2016	NULL
2806	Durant	Rue des Chausseurs, ?	Médecin	14/04/2006	Sport
2807	Dupond	Av du Palais, Paris	Cadre	32/02/2226	Maison



Pré-traitement des données

- Information manquante
- Supprimer l'enregistrement.
 - A faire si la classe est manquante car n'aide pas à la classification
- Remplir manuellement les champs :
 - difficile et long
- Automatiquement :
 - Remplacer un salaire manquant par le salaire médian des clients
 - Prédire les valeurs manquantes, en le déduisant d'autres paramètres (salaire à partir de l'âge et de la profession)
 - Inférer la valeur avec un algorithme de classification (la valeur à prédire devient la classe recherchée)



Notebook

- Ingénierie des données « **Traitement des valeurs manquantes** »



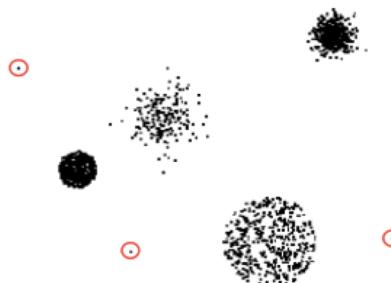
Pré-traitement des données

- Données bruitées : Plusieurs solutions : lissage, segmentation, régression linéaire
- Techniques de lissage (*data smoothing*) :
 1. Trier les différentes valeurs de l'attribut considéré : {4, 8, 15, 21, 21, 24, 25, 28, 34}
 2. Partitionner l'ensemble résultat.
{{4, 8, 15}, {21, 21, 24}, {25, 28, 34}}
 3. Remplacer les valeurs initiales par de nouvelles valeurs en fonction du partitionnement réalisé :
 - par la valeur moyenne des regroupements réalisés {9, 22, 29}
 - par les min et max des regroupements réalisés. {{4, 4, 15}, {21, 21, 24}, {25, 25, 34}}
- Implique une perte de précision ou d'information



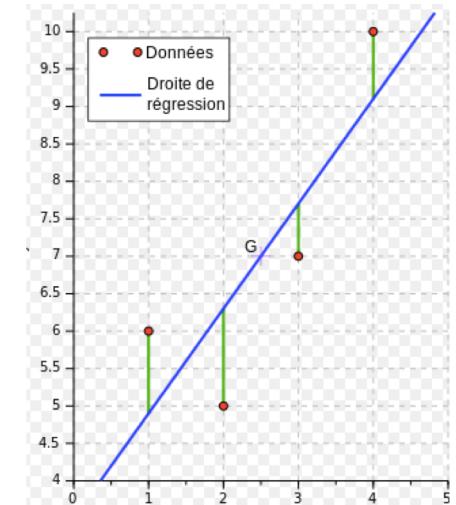
Pré-traitement des données

- Utilisation de la fouille pour aider à pré-traiter les données
- Techniques de segmentation (clustering) :
 - Les valeurs similaires sont placées dans une même classe
 - On ne tient pas compte des valeurs isolées (dans une classe comportant trop peu d'éléments)



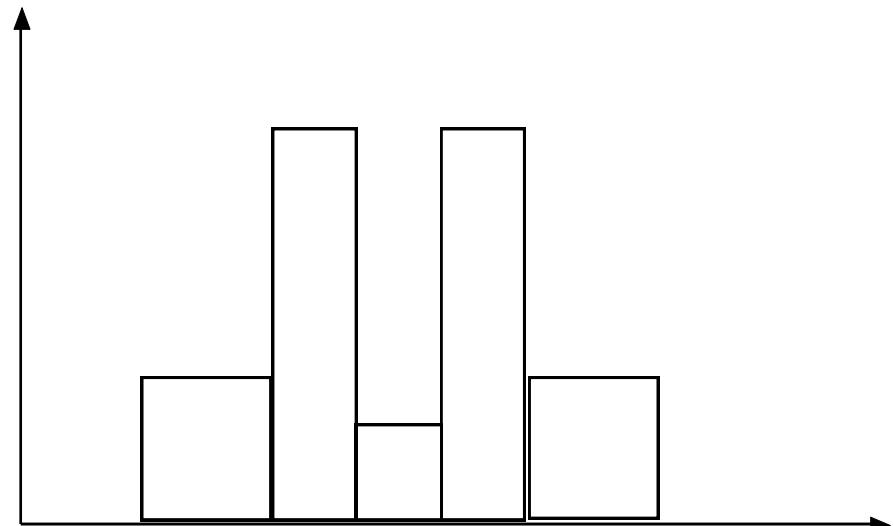
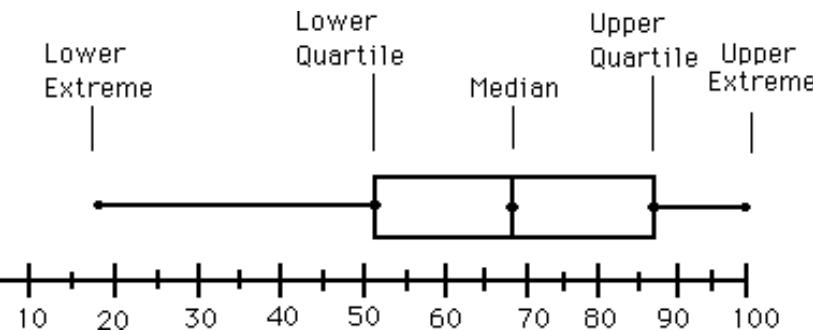
Pré-traitement des données

- Techniques de régression linéaire :
- Hypothèse : un attribut Y dépend linéairement d'un attribut X
 - Années d'expérience X et salaire Y
- Trouver les coefficients a et b tels que $Y = aX + b$
- Remplacer les valeurs de Y par celles prédites
- Données de départ :
 - Un ensemble de couples (X_i, Y_j)
- Détermination des coefficients :
 - Soient \bar{X} et \bar{Y} les valeurs moyennes des attributs X et Y .
 - $a = \text{cov}(x,y)/\text{var}(x)$
 - $b = \bar{Y} - a\bar{X}$



Pré-traitement des données

- Statistiques descriptives sur les données :
- Utiles pour voir la centralité, la dispersion, les variations, les distributions
 - Valeur médiane, moyenne, variance, écart type, mode, quantiles
 - Boxplots, histogrammes,



Attention aux interprétations !



Attention aux interprétations !



Attention aux interprétations!

ON TEENAGERS, ADULTS:

Statistics show that
teen pregnancy
drops off significantly
after age 25.

*Mary Anne Treadie, Republican state senator from Colorado Springs
(convened by Harry F. Pomer)*

MONDAY DECEMBER 1999

Data doesn't create meaning, we do. –Susan Etlinger



Attention aux données !

Quartet d'Ascombe

I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89



Attention aux données !

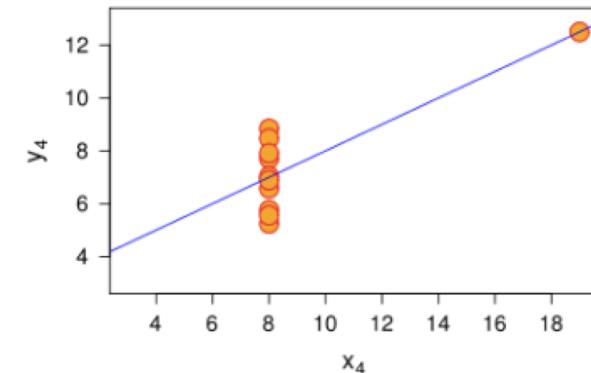
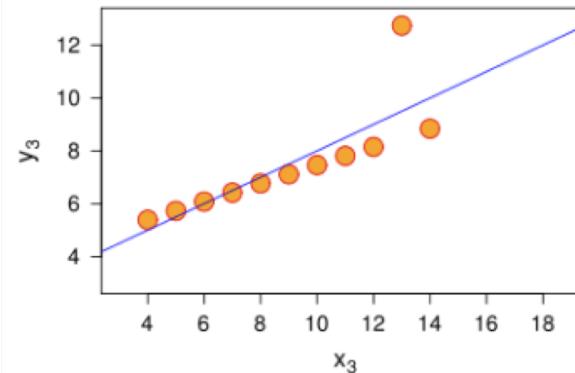
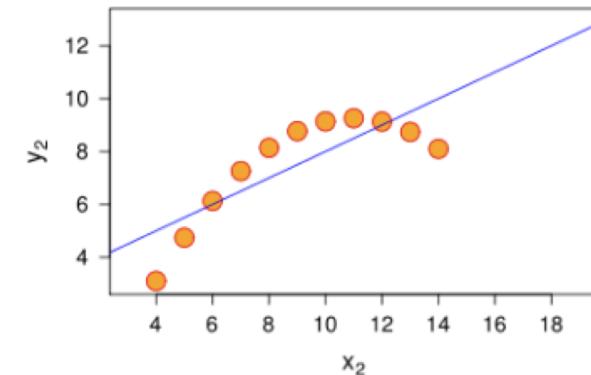
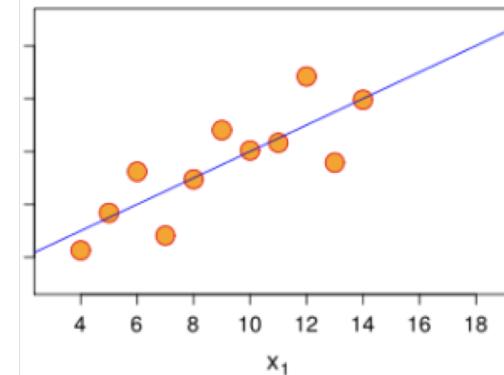
Propriété	Valeur
Moyenne des x	9, 0
Variance des x	10, 0
Moyenne des y	7, 5
Variance des y	3, 75
Corrélation entre les x et les y	0, 816
Équation de la droite de régression linéaire	$y = 3 + 0, 5x$
Somme des carrés des erreurs relativement à la moyenne	110, 0



Importance de la visualisation

Quartet d'Ascombe

I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89



Notebook

- Visualisation des données



Pré-traitement des données

- Intégration de données
 - Combiner des données de différentes sources en un seul lieu (ETL/Entrepôt)
- Intégration de schéma
 - A.cust-id \equiv B.cust-#
- Identification des entités
 - Bill Clinton = William Clinton
- Déetecter et résoudre les conflits de valeurs dans les données
 - Unités différentes (Km \leftrightarrow miles)



Pré-traitement des données

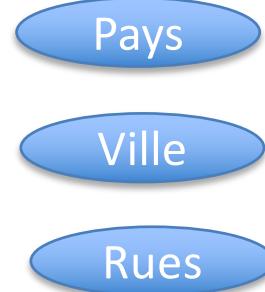
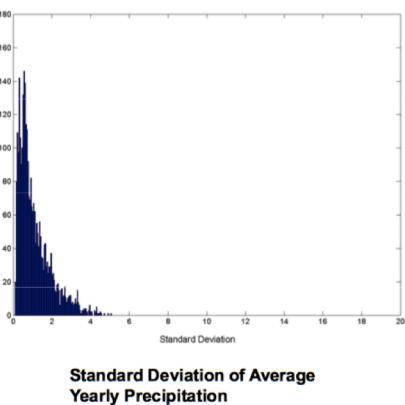
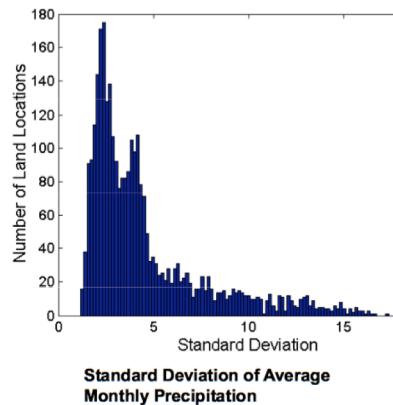
- Possibilités d'avoir de nouvelles données manquantes ou aberrantes

Client	Date Naissance	Salaire	Propriétaire	Voiture
Dupond	05/07/1973	20K	OUI	OUI
Durand	13/11/1995	2K	NON	OUI
Duchemin	12/09/1987	12K	NON	NON
Durant	01/22/1833	32K	NON	NON



Transformation des données

- Dépend de l'algorithme de fouille utilisé
- Regroupements
 - Cas où les attributs prennent un très grand nombre de valeurs discrètes (e.g. adresses que l'on peut regrouper en régions, rapports mensuels en rapports annuels, âge -> jeune, vieux)
 - Agréger des attributs
 - Avantages : les données agrégées ont moins de variations



Transformation des données

- Attributs discrets
 - Les attributs discrets symboliques prennent leurs valeurs dans un ensemble fini donné (e.g. colonne magazine de l'exemple).
 - Deux représentations possibles : représentation verticale ou représentation horizontale ou éclatée (plus adaptée à la fouille de données)



Transformation des données

- Représentation verticale vs éclatée

Client	Magazine
2807	Voiture
2807	Musique
2807	BD
3456	BD
4356	Sport
2806	Sport
2807	Maison

Client	Voiture	Musique	BD	Sport	Maison
2807	1	1	1	0	1
3456	0	0	1	0	0
4356	0	0	0	1	0
2806	0	0	0	1	0

Transformation des données

- Changements de types pour permettre certaines manipulations comme par exemple des calculs de distance, de moyenne (e.g. date de naissance)
- Uniformisation d'échelle
 - Attention certains algorithmes sont basés sur des calculs de distance entre enregistrements. Les variations d'échelles entre ces algorithmes peuvent perturber ces algorithmes



Transformation des données

- Un exemple de transformation

Client	Voiture	Musique	BD	Sport	Maison	DN	REV	Prop	Voiture	PN	DA
2807	1	1	1	0	1	45	20	OUI	OUI	1	7
3456	0	0	1	0	0	23	2	NON	OUI	0	NULL
4356	0	0	0	1	0	31	12	NON	NON	0	3
2806	0	0	0	1	0	35	32	NON	NON	NULL	12

Avec

DN : Date de Naissance -> âge

REV : Revenu

Prop : Propriétaire

PN : Paris/Province

DA : première date d'abonnement



Notebook

- Ingénierie des données « **Traitement des données catégorielles ou qualitatives** »



Similarité ou dissimilarité

- **Similarité**
 - Mesure de la ressemblance de deux objets
 - Plus les objets sont semblables plus grande est la valeur
 - Généralement dans l'intervalle [0,1]
- **Dissimilarité (e.g., distance)**
 - Mesure de la différence entre deux objets
 - Plus la distance est courte plus les objets sont proches
 - La dissimilarité minimale est souvent 0
 - Les limites supérieures sont très variables
- **La proximité fait référence à la similarité ou la dissimilarité**



Matrice et matrice de dissimilarité

- **Matrice de données**

- N points avec n dimensions

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- **Matrice de dissimilarité**

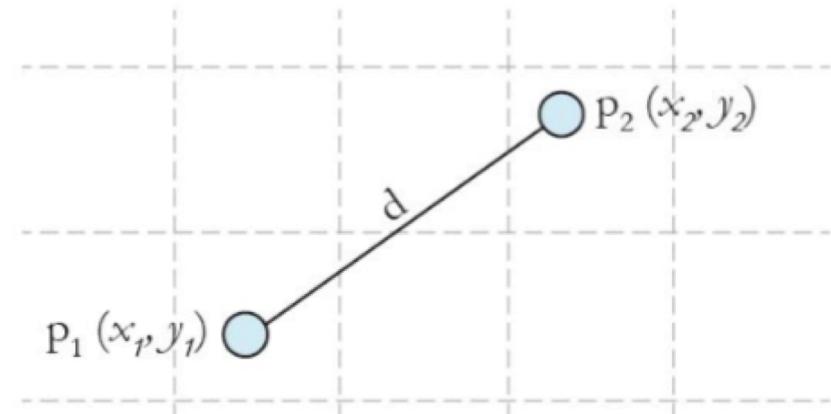
- N points mais seules les distances sont enregistrées
 - Une matrice triangulaire

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$



Attributs numériques

- La distance Euclidienne est la distance « normale » entre deux points



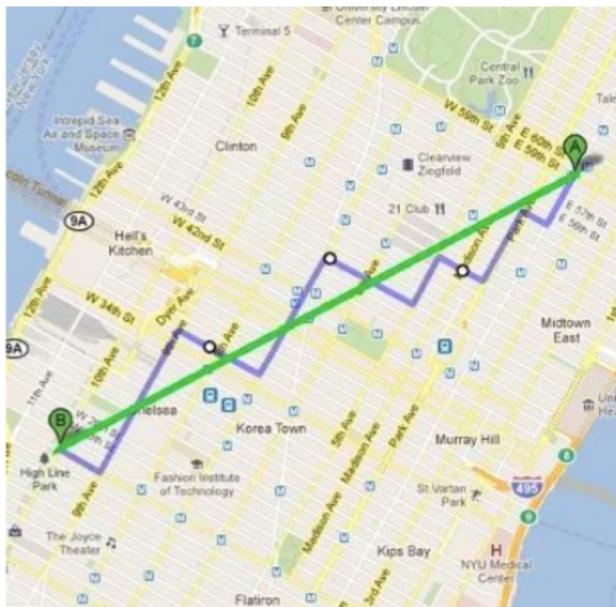
$$\text{Euclidean distance } (\delta) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

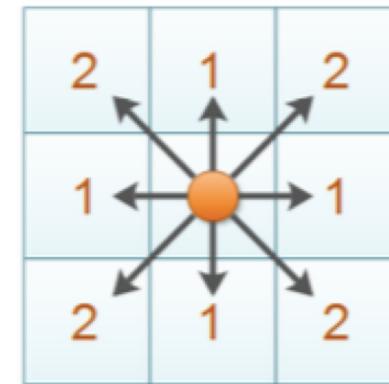
$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$

Attributs numériques

- La distance Manhattan
(inspirée des chauffeurs de taxi à Manhattan)



Manhattan Distance



$$|x_1 - x_2| + |y_1 - y_2|$$

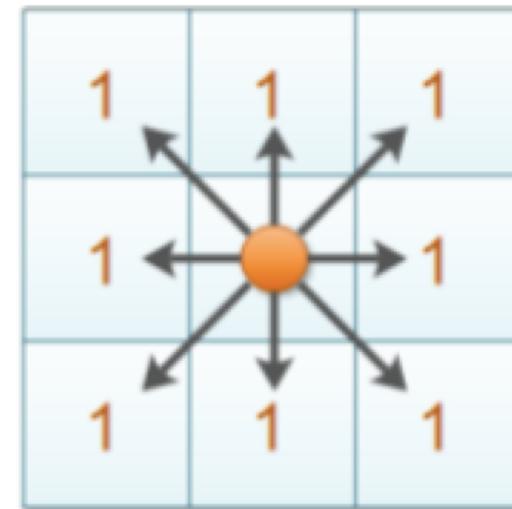
$$d = \sum_{i=1}^n |\mathbf{x}_i - \mathbf{y}_i|$$

Attributs numériques

- La distance de Chebyshev
(``le plus long chemin``)

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	1	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

Chebyshev Distance



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

Attributs numériques

- La distance de Minkowski : une généralisation des distances précédentes :

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

Où $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ et $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ sont des objets de p dimensions (La distance est appelée la L-h norme)

Si $h=1 \rightarrow$ distance Manhattan, $h=2 \rightarrow$ distance Euclidienne, $h \rightarrow \infty \rightarrow$ distance Chebyshev



Attention aux distances

- Rappel : Attention certains algorithmes sont basés sur des calculs de distance entre enregistrements. Les variations d'échelles entre ces algorithmes peuvent perturber ces algorithmes

Nom	Age	Salaire
Clara	50	11000
Marie	70	11100
Léa	60	11122
Lucy	60	11074

De qui Clara est la plus proche :
Marie ou Léa ?



Attention aux distances

- Rappel : Attention certains algorithmes sont basés sur des calculs de distance entre enregistrements. Les variations d'échelles entre ces algorithmes peuvent perturber ces algorithmes

Nom	Age	Salaire
Clara	50	11000
Marie	70	11100
Léa	60	11122
Lucy	60	11074

De qui Clara est la plus proche :
Marie ou Léa ?

De Léa :
 $\text{Diff}(\text{Age}) \text{ Léa} = 10, \text{ Marie } 20$
 $\text{Diff} (\text{Salaire}) \text{ Léa} = 122, \text{ Marie } 100$



Attention aux distances

- Utilisation d'une distance de Manhattan

Nom	Age	Salaire
Clara	50	11000
Marie	70	11100
Léa	60	11122
Lucy	60	11074

$$d(\text{Clara}, \text{Marie}) = 120$$
$$d(\text{Clara}, \text{Léa}) = 132$$

Clara est plus éloignée de Léa !

Problème d'échelle des données



Normalisation

- Normalisation des attributs : valeurs trop grandes qui pénalisent les distances
- Normalisation min-max en [new_min, new_max]

$$v' = \frac{v - \min_A}{\max_A - \min_A} (new_max_A - new_min_A) + new_min_A$$

Si le salaire varie de 11000 à 11122, la valeur 11100 normalisée entre [\emptyset – 1] est :

$$(11100 - 11000) / (11122 - 11000) * (1 - 0) + 0$$

est transformée 0.81



Normalisation

Nom	Age	Salaire
Clara	50	11000
Marie	70	11100
Léa	60	11122
Lucy	60	11074

Min-max normalisation

Nom	Age	Salaire
Clara	50	0
Marie	70	0,81967
Léa	60	1
Lucy	60	0,60656



Normalisation

- Z-score

$$V' = v - m_A / s_A$$

Où m_A est la moyenne pour l'attribut A et s l'écart type pour l'attribut A
Négatif quand V est en dessous de la moyenne positif autrement

- Alternative : calculer l'écart moyen absolu

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

$$V' = v - m_A / s_f$$

- L'écart moyen absolu est plus robuste que l'écart type notamment en cas d'outliers



Normalisation

- Mise à l'échelle décimale

$$v' = \frac{v}{10^j}$$

Nom	Age	Salaire
Clara	50	110
Marie	70	111
Léa	60	111,22
Lucy	60	110,74



Normalisation

- Z-score normalisation

Nom	Age	Salaire
Clara	-2	-0,5
Marie	2	0,18
Léa	0	0,32
Lucy	0	0

$$d(\text{Clara}, \text{Marie}) = 4,67$$
$$d(\text{Clara}, \text{Léa}) = 2,34$$

Clara est plus proche de Léa !

$$\text{Mean}_{\text{age}} = 60 \quad S_{\text{age}} = 5$$

$$\text{Mean}_{\text{Salaire}} = 11074 \quad S_{\text{Salaire}} = 48$$



Attributs tous continu

- Echelles différentes :
 - Il y a des attributs dominants
 - Il faut normaliser avant de calculer des distances
 - Tout ramener entre 0 et 1
- On peut vouloir garder la dissymétrie entre attributs
 - Donner un poids à chaque attribut
 - Calculer la distance en fonction de ce poids

$$\sqrt{w_1(x_1 - y_1)^2 + \dots + w_n(x_n - y_n)^2}$$

- Nécessite une très bonne connaissance du domaine !



Attributs binaires

- Une table de contingence pour données binaires

		Objet <i>j</i>		<i>sum</i>
		1	0	
<i>Objet i</i>	1	<i>a</i>	<i>b</i>	<i>a+b</i>
	0	<i>c</i>	<i>d</i>	<i>c+d</i>
<i>sum</i>		<i>a+c</i>	<i>b+d</i>	<i>p</i>

a : nombre de positions où *i* a 1 et *j* a 1
b : nombre de positions où *i* a 1 et *j* a 0
c : nombre de positions où *i* a 0 et *j* a 1
d : nombre de positions où *i* a 0 et *j* a 0

- Exemple $oi=(1,1,0,1,0)$ et $oj=(1,0,0,0,1)$:
 - $a=1$, $b=2$, $c=1$, $d=2$



Mesures de distances

- Coefficient d'appariement (matching) simple (invariant pour variables symétriques) :

$$d(i, j) = \frac{b + c}{a + b + c + d}$$

- Ex. pour $oi=(1,1,0,1,0)$ et $oj=(1,0,0,0,1)$ $d(oi, oj)=3/5$
- Coefficient de Jaccard

$$d(i, j) = \frac{b + c}{a + b + c}$$

- $d(oi, oj)=3/4$



Rappel codage binaire

- Attribut symétrique :
 - le sexe d'une personne, i.e coder masculin par 1 et féminin par 0 est similaire au codage inverse
- Attribut asymétrique:
 - Test médical. Le test peut être positif ou négatif (0 ou 1) mais il y a une valeur qui sera plus présente que l'autre
 - Généralement, on code par 1 la modalité la moins fréquente
 - 2 personnes ayant la valeur 1 pour le test sont plus similaires que 2 personnes ayant 0 pour le test



Mesures de distances

Nom	Sexe	Fièvre	Tousse	Test-1	Test-2	Test-3	Test-4
Jacques	M	O	N	P	N	N	N
Marie	F	O	N	P	N	P	N
Jean	M	O	P	N	N	N	N

- Sexe est un attribut symétrique. Les autres (fièvre, test-1...) sont asymétriques
- O et P = 1, N = 0, la distance n'est mesurée que sur les asymétriques

$$d(jacques, marie) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jacques, jean) = \frac{1 + 1}{1 + 1 + 1} = 0.66$$

$$d(jean, marie) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

- Les plus similaires sont Jacques et Marie : atteints du même mal



$$d(i, j) = \frac{b+c}{a+b+c}$$

a : nombre de positions où i a 1 et j a 1
 b : nombre de positions où i a 1 et j à 0,
 c nombre de positions où i a 0 et j a 1

Attributs nominaux

- Une généralisation des attributs binaires, ex: rouge, vert et bleu
- Méthode 1: Matching simple
 - m : nombre d'appariements, p : nombre total de variables

$$d(i, j) = \frac{p - m}{p}$$

- Méthode 2 : utiliser un grand nombre d'attributs binaires
 - création d'une variable binaire pour chacun des états d'une variable nominale



Attributs ordinaux

- Un attribut ordinal peut être discret ou continu
- L'ordre peut être important (e.g. froid < tiède < chaud)
- Peuvent être traitées comme les variables intervalles
 - remplacer x_{if} par son rang $r_{if} \in \{1, \dots, M_f\}$
 - Remplacer le rang de chaque variable par une valeur dans $[0, 1]$ en remplaçant la variable f dans l'objet I par

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- Utiliser une distance pour calculer la similarité

Froid => $1-1/3-1=0$

Tiède => $2-1/3-1 = 0.5$

Chaud => $3-1/3-1 = 1$



Attributs chaînes

- La distance de Hamming entre deux chaînes de mêmes longueurs est le nombre de positions où les symboles correspondants sont différents
 - En d'autres termes, elle mesure le nombre minimum de substitutions nécessaires pour changer une chaîne en une autre
- 10**111**01 et 1001001 = 2
"Bon**jour**" et "B**no**jour" = 2
- Utilisé en télécommunications, en bioinformatique, en text mining



Attributs mixtes

- Les objets peuvent être décrits avec tous les types de données
 - binaire symétrique, binaire asymétrique, nominale, ordinale, ...
- Utilisation d'une formule pondérée pour combiner leurs effets

$$d(i, j) = \frac{\sum_{k=1}^p w_k d_k(i, j)}{\sum_{k=1}^p w_k}$$



Attributs mixtes

Nom	Age	Prop	Mensualité
Jean	30	1	1000
Pierre	40	0	2200
Paul	45	1	4000

- $d(x,y) = \sqrt{(10/15)^2 + 1^2 + (1200/3000)^2} = 1.27$
- $d(x,z) = \sqrt{(15/15)^2 + 0^2 + (3000/3000)^2} = 1.41$
- $d(y,z) = \sqrt{(5/15)^2 + 1^2 + (1800/3000)^2} = 1.21$

Le voisin le plus proche de Jean est Pierre
- Distance normalisée et sommation $d(x,y) = d1(x,y) + d2(x,y) ..$



Notebook

- Ingénierie des données « **Mise à l'échelle des valeurs attributs (Feature scaling)** »



Quid du volume de données ?

- Grandes Bases de Données ou non ?
- Faut -il échantillonner ?
 - 100 000 enregistrements, 100 Mo par jour
 - 2 Go par jour, 100 Go par heure
 - *Déjà les petabyte (2^{50}) ...*
- Il est souvent nécessaire d'échantillonner pour une raison simple : temps de traitements trop longs



Plan

Introduction et contexte du processus de KDD

Les données et les pré-traitements

Les tâches de la fouille de données

Conclusions



Les tâches du DM

- Méthodes prédictives
 - Utilisent les données existantes et des résultats connus sur ces données pour développer des modèles capables de prédire les valeurs d'autres données
 - Prédire les clients qui ne rembourseront pas leur crédit
 - Utilisé principalement en classification et prédiction
- Méthodes descriptives
 - Proposent des descriptions de données pour aider à la prise de décision
 - Souvent en amont de la construction de modèles prédictifs
 - Utilisé principalement en segmentation (clustering) et association



Les tâches du DM

- Les principales tâches
 - Classification (prédictive)
 - Groupement/segmentation (*clustering*) (descriptive)
 - Recherche de règles d'association (descriptive)
 - Recherche de motifs (descriptive)
 - Régression (prédictive)
 - Détection d'anomalies (prédictive)
- Pour chacune des tâches n méthodes



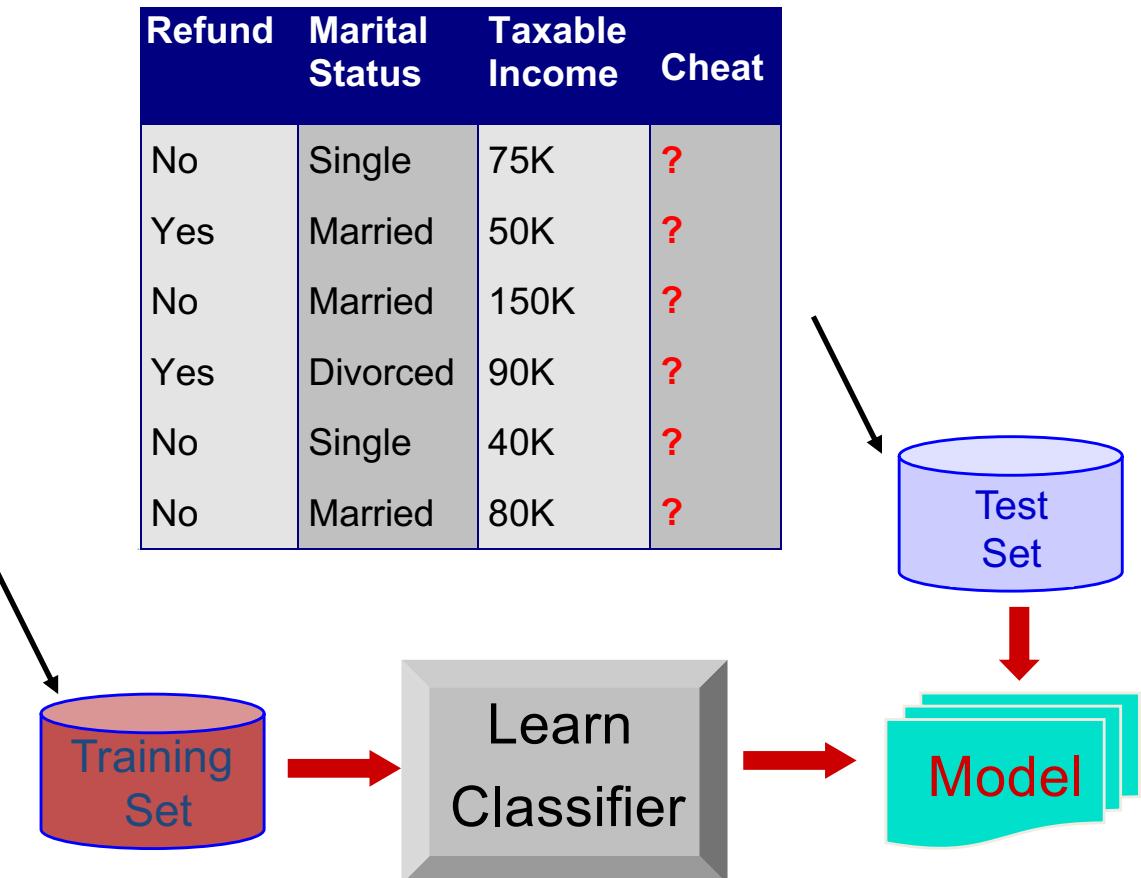
Classification

- Soit une collection d'enregistrements (**ensemble d'apprentissage**)
 - Chaque enregistrement contient un ensemble d'attributs, l'un de ces attributs est la **classe**.
- Rechercher un **modèle** pour l'attribut classe comme une fonction des valeurs des autres attributs
- But : Affecter de la meilleure manière possible les enregistrement non vues dans la classe.
 - Un **jeu de test** est utilisé pour déterminer l'efficacité du modèle. Généralement le jeu de données est divisé en jeu d'entraînement et en jeu de test. Le jeu d'entraînement est utilisé pour apprendre le modèle et le jeu de test pour valider le modèle.



Principe de la classification

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



Classification - Exemples

- Marketing direct
 - But : réduire le coût du mailing en ciblant un ensemble de consommateurs qui achèteront vraisemblablement un nouveau téléphone portable
 - Fonctionnement :
- Utiliser des données pour un produit similaire.
 - On sait quels consommateurs ont acheté. La décision (Achat - Pas achat) est l'attribut classe
 - Collecter diverses informations sur ce type de consommateurs
 - Cette information représente les entrées du classifier.



Notebook

- Premières classifications



Segmentation(Clustering)

- Soit un ensemble d'objets composés d'un ensemble d'attributs, et une mesure de similarité entre eux, rechercher des clusters tels que :
 - Les objets dans un cluster sont les plus similaires les uns des autres
 - Les objets dans des clusters séparés sont les moins similaires entre eux
- Mesures de similarités :
 - La distance Euclidienne si les attributs sont continus
 - D'autres mesures spécifiques au problème



Clustering

- Soit un ensemble d'objets composés d'un ensemble d'attributs, et une mesure de similarité entre eux, rechercher des clusters tels que :
 - Les objets dans un cluster sont les plus similaires les uns des autres
 - Les objets dans des clusters séparés sont les moins similaires entre eux
- Mesures de similarités :
 - La distance Euclidienne si les attributs sont continus
 - D'autres mesures spécifiques au problème



Clustering: Application 1

- Segmentation de marché
- Objectif: sous-diviser un marché en différents ensembles de clients de façon à mieux cibler le marketing
- Approche :
 - Collecter différents attributs sur les consommateurs
 - Trouver des clusters d'utilisateurs similaires
 - Faire une mesure de qualité de clustering en observant les achats des utilisateurs d'un cluster par rapport aux autres



Clustering: Application 2

- Clustering de documents
- Objectif: trouver les groupes de documents qui sont similaires entre eux en fonction des termes qui y apparaissent
- Approche :
 - identifier les termes qui apparaissent fréquemment dans chaque document. Faire une mesure de similarité basée sur la fréquence des termes, et l'utiliser pour faire le clustering



Découverte de règles d'association

- Etant donné un ensemble d'enregistrements qui contiennent des éléments d'une collection
- Générer des règles de dépendance qui prédisent les occurrences d'éléments suivant les occurrences des autres

<i>TID</i>	<i>Items</i>
1	Pain, Coca, Lait
2	Bière, Pain
3	Bière, Coca, Couches, Lait
4	Bière, Pain, Couches, Lait
5	Coca, Couches, Lait

Règles découvertes:
 $\{Lait\} \rightarrow \{Coca\}$
 $\{Couche, Lait\} \rightarrow \{Bière\}$

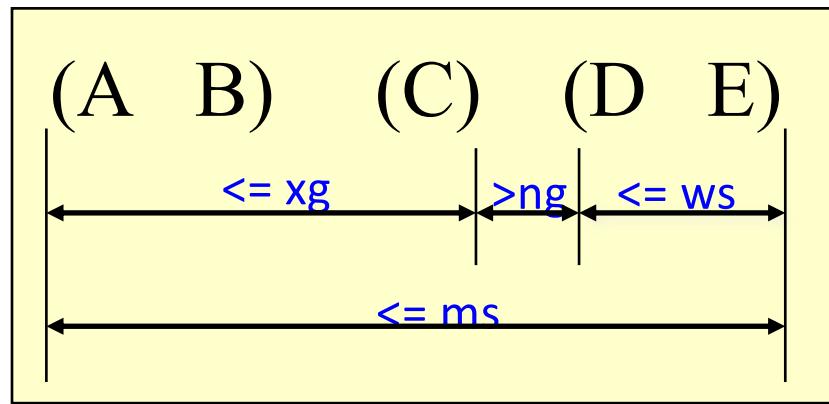


Découverte de motifs séquentiels

- Étant donné un ensemble d'objets, dans lequel chaque objet est associé à une séquence temporelle, trouver des dépendances séquentielles entre les évènements

(A B) (C) (D E)

- Des contraintes temporelles peuvent être prises en compte



Recherche de motifs fréquents

- Analyse des associations
 - Panier de la ménagère, cross marketing, conception de catalogue, analyse de textes
 - Corrélation ou analyse de causalité
- Clustering et Classification
 - Classification basée sur les associations
- Analyse de séquences
 - Web Mining, détection de tendances, analyses ADN
 - Périodicité partielle, associations temporelles/cycliques



Régression

- Prédire la valeur d'une variable connue en utilisant la valeur d'autres variables en supposant une relation (non) linéaire entre elles
- Très utilisé en statistiques
- Exemples:
 - Prédire la quantité de ventes d'un nouveau produit en fonction du budget de publicité
 - Prédire la force du vent en fonction de la température, humidité, pression ...
 - Prédire le cours de la bourse



Détection d'anomalies

- Déte^cter des déviations significatives d'un comportement normal
- Applications :
 - Détection de fraudes (CB)
 - Détection d'attaques sur un réseau

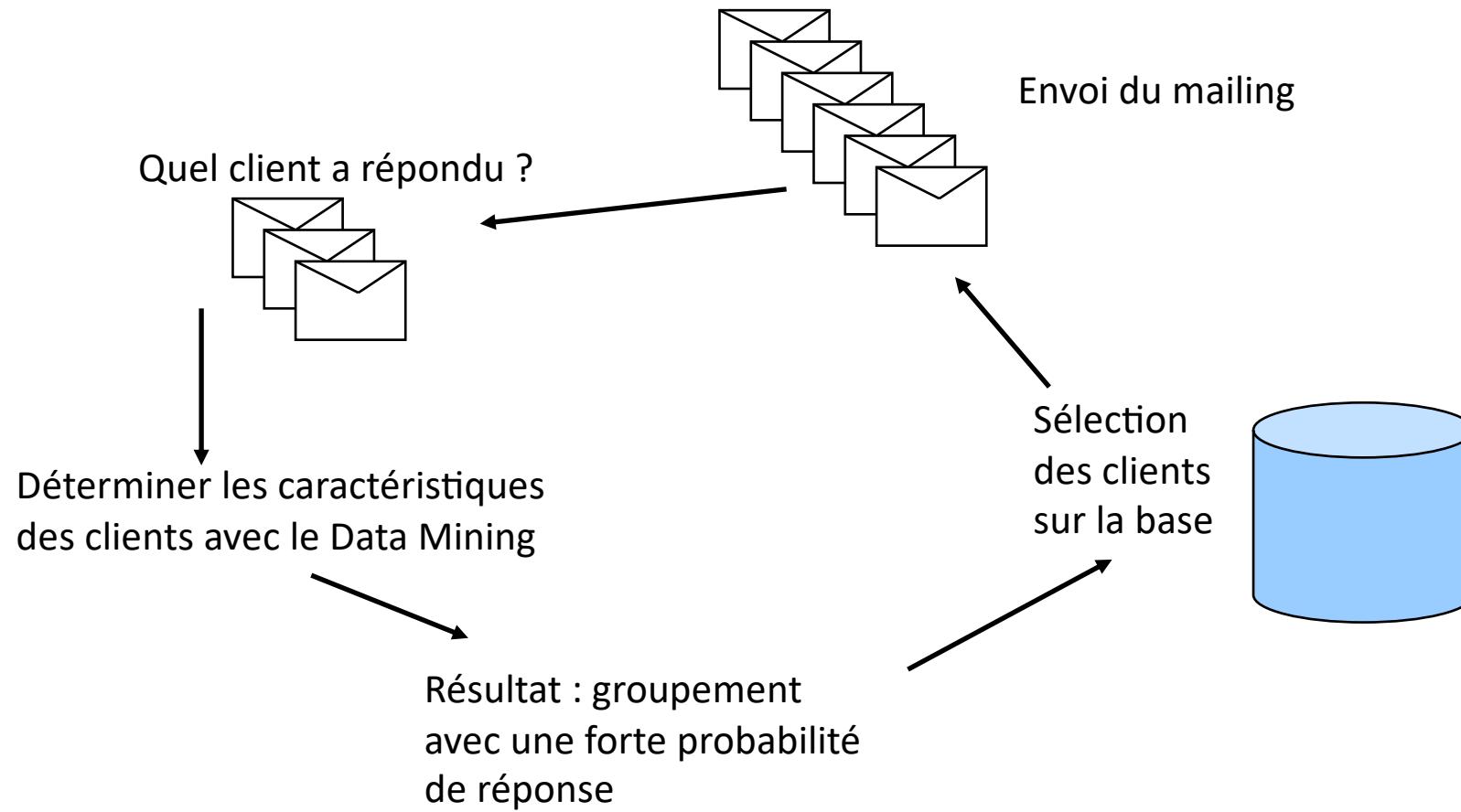


Le mailing

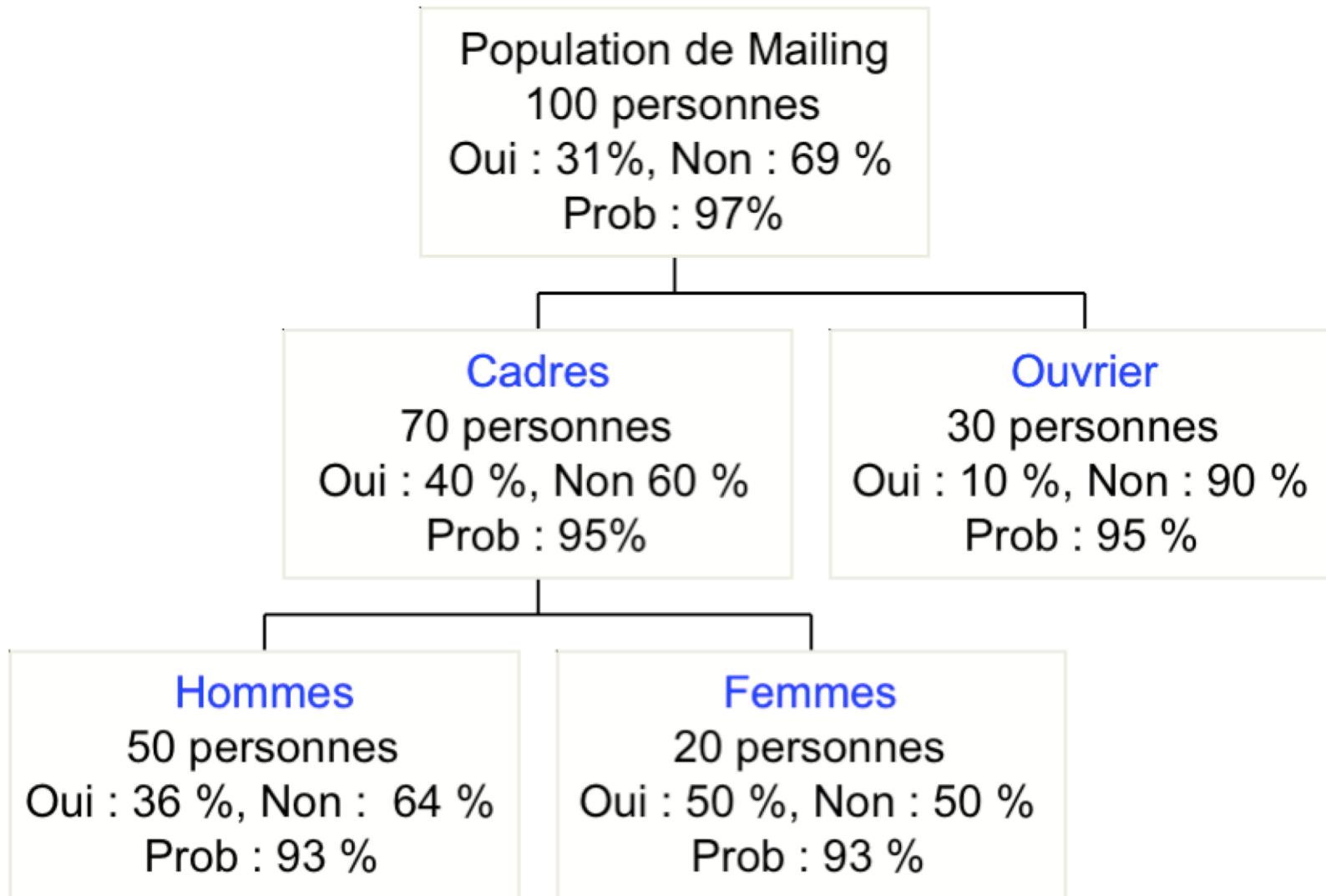
- Classification... un exemple d 'utilisation
 - un cadeau est envoyé par mailing. Un envoi sans réponse coûte 50 € et une réponse assure 100 €.
 - Pas d 'envoi de mailing à un client qui aurait répondu : perte de 100 €.



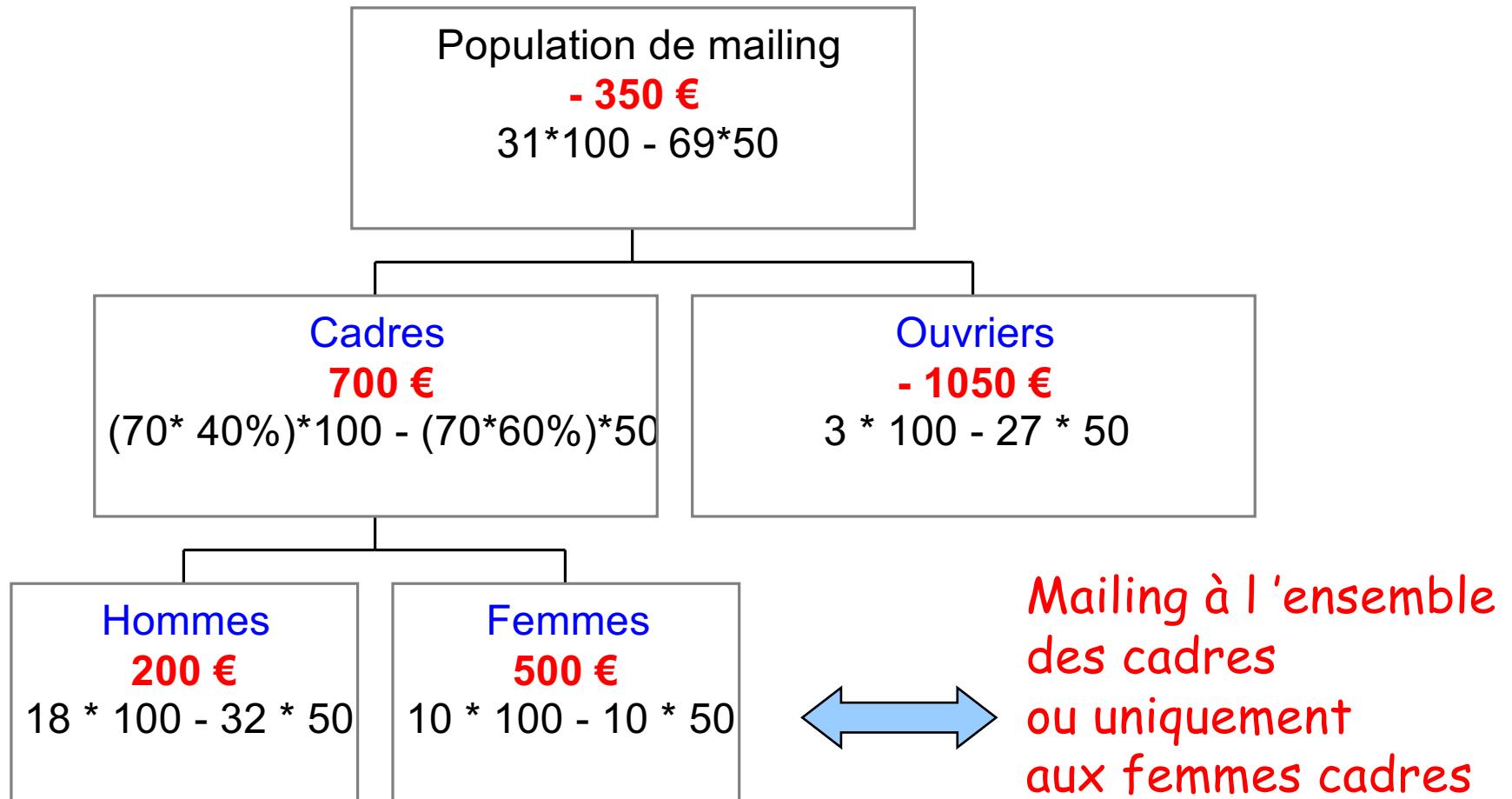
Le mailing



Résultat du mailing



Quantification



Evaluation

		Matrice de coûts			TOTAL	
		OBSERVE				
Prédit	↓	Payé	Retardé	Impayé		
		80	15	5	100	
Payé		1	17	2	20	
Retardé		5	2	23	30	
Impayé		86	34	30	150	
TOTAL						

Validité du modèle : nombre de cas exacts
(=somme de la diagonale) divisé par le nombre total :
 $120/150 = 0.8$

Un exemple de classification binaire

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N



La matrice de confusion

		Réel	
		Positif	Négatif
Prédit	Positif	TP	FP
	Négatif	FN	TN

TP : TRUE POSITIVES

FP : FALSE POSITIVES

FN : FALSE NEGATIVES

TN : TRUE NEGATIVES



De nombreuses informations

		Réel	
		Positif	Négatif
Prédit	Positif	TP	FP
	Négatif	FN	TN

$$\text{ACCURACY : } \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$$

Proportion de prédictions/classifications correctes sur l'ensemble



De nombreuses informations

		Réel	
		Positif	Négatif
Prédit	Positif	TP	FP
	Négatif	FN	TN

RAPPEL :
$$\frac{TP}{TP+FN}$$

Proportion de prédictions/classifications correctes (true positive) de tous les cas qui sont réellement positifs.



De nombreuses informations

		Réel	
		Positif	Négatif
Prédit	Positif	TP	FP
	Négatif	FN	TN

$$\text{PRECISION : } \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Proportion de prédictions/classifications correctes (TP) à partir des cas qui sont considérés comme positifs.



Notebook

- Premières classifications : « Plus loin sur l'évaluation d'un modèle »



Plan

Introduction et contexte du processus de KDD

Les données et les pré-traitements

Les tâches de la fouille de données

Conclusions



Conclusions

- Ces différents éléments vont être vus dans la suite des cours
- Objectifs : être capable de mettre en œuvre les algorithmes, les interpréter et les évaluer
- En M1 : on considère que les données sont propres – peu de prétraitements. L’importance c’est de bien comprendre les fonctionnements et surtout les interpréter



Conclusions

- Fouille de données de très nombreuses perspectives
 - Données de plus en plus hétérogènes
 - Données de plus en plus volumineuses
 - Données de plus en plus rapides
- Attention à ne pas oublier le pourquoi
 - Savoir bien classer est utile
 - Savoir pourquoi un objet a été mis dans une classe est très utile !



Conclusions

- Des questions de droits :
 - Est ce que j'ai le droit d'utiliser les données ?
 - Est ce que je suis propriétaire des données et des connaissances obtenues ?
 - Est ce que mes connaissances préservent la vie privée ?
- Des questions éthiques ...



Conclusions

- Logiciels
 - Il en existe de nombreux !! Sas, Intelligence Miner, Mineset, ...
 - Beaucoup en open source
 - Weka : utile pour une analyse rapide
 - Scikit-learn : machine learning Python
 - R vs Scikit-learn
- Ne pas oublier le plus important est de comprendre les algorithmes et de savoir comment les utiliser
 - K-means est un algorithme de clustering -> il est disponible sur de nombreuses plateformes



Conclusions

- Quelques pointeurs importants :
 - Kdnuggets = une source d'information sur tout ce qui se fait autour de la fouille : tutoriels, stages, jeux de données, offre d'emploi, news (www.kdnuggets.com)
 - Des associations et listes de diffusions : EGC (Extraction et Gestion de connaissances) (egc.assoc.fr), AFIA (Association Française pour l'IA) (afia.asso.fr)
 - De très nombreuses conférences : KDD, ICDM, PKDD, SDM, PAKDD, ...



-
- Des questions ?

