

Apprentissage statistique et Data mining

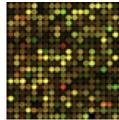
Amira Barhoumi

amira.barhoumi@univ-grenoble-alpes.fr

Organisation du cours

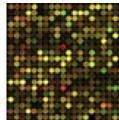
- Déroulement
 - Volume horaire : 30 heures
 - Support sur Chamilio
- Évaluation
 - Note du partiel
- Intervenants
 - Amira BARHOUMI (amira.barhoumi@univ-grenoble-alpes.fr)

- Explosion des données
 - Grande masse de données
 - Données hétérogènes

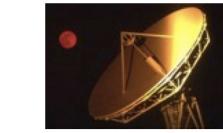


Motivation

- Explosion des données
 - Grande masse de données
 - Données hétérogènes

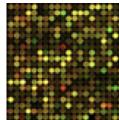


- Analyses manuelles difficiles/impossibles



- Explosion des données

- Grande masse de données
- Données hétérogènes



- Analyses manuelles difficiles/impossibles

- Augmentation de la puissance de calcul

- Grande capacité de stockage
- Possibilité d'exécuter des processus gourmands



Métaphore

- Trop de données
Paradoxe : trop de données mais pas assez d'information
“we are data rich, information poor”



- Difficultés d'accès à l'information



- Trop de pistes à explorer



- ... Pas d'accès facile à l'information



Métaphore

- Besoin d'automatisation



- Extraction des connaissances à partir des bases de données



- Génération d'hypothèses



Solution : Fouille de données

- Par analogie à la recherche de pépites d'or dans un gisement, la fouille de données vise à :
 - 1- Extraire des informations cachées (non connues) par analyse globale
 - 2- Découvrir des modèles/patterns difficiles à percevoir car :
 - le volume de données est très grand
 - le nombre de variables à considérer est important
 - ces patterns sont imprévisibles (même à titre d'hypothèse à vérifier)

Évolution des technologies de bases de données

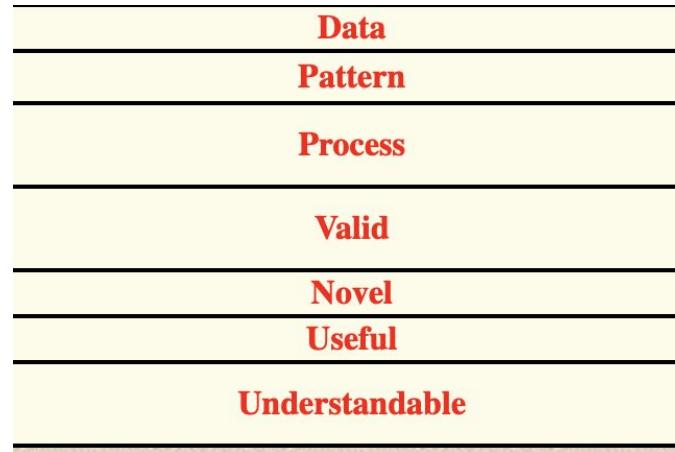
- 1960 : Système de gestion de fichiers, collection de données, bases de données (modèle réseau)
- 1970... : Bases de données relationnelles
- 1980... : Bases de données relationnelles, modèles de données avancés (extension du modèle relationnel, OO, etc), bases de données orienté application (spatiale, scientifique, etc)
- 1990-2000 : Fouille de données, entrepôt de données, bases de données multimédia, bases de données web
- 2010 : Big data, deep learning, cloud computing, data science

- **KDD = Knowledge Discovery in Databases (en)**
= *Extraction de connaissances à partir de données ECD (fr)*
- Processus impliquant l'extraction d'informations utiles, initialement inconnues et potentiellement précieuses à partir de grands ensembles de données.
 - inductif
 - itératif
 - interactif

“*Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*” [Fayyad, Piatetsky-Shapiro and Smyth, 1996]

source: <https://cdn.aaai.org/KDD/1996/KDD96-014.pdf>

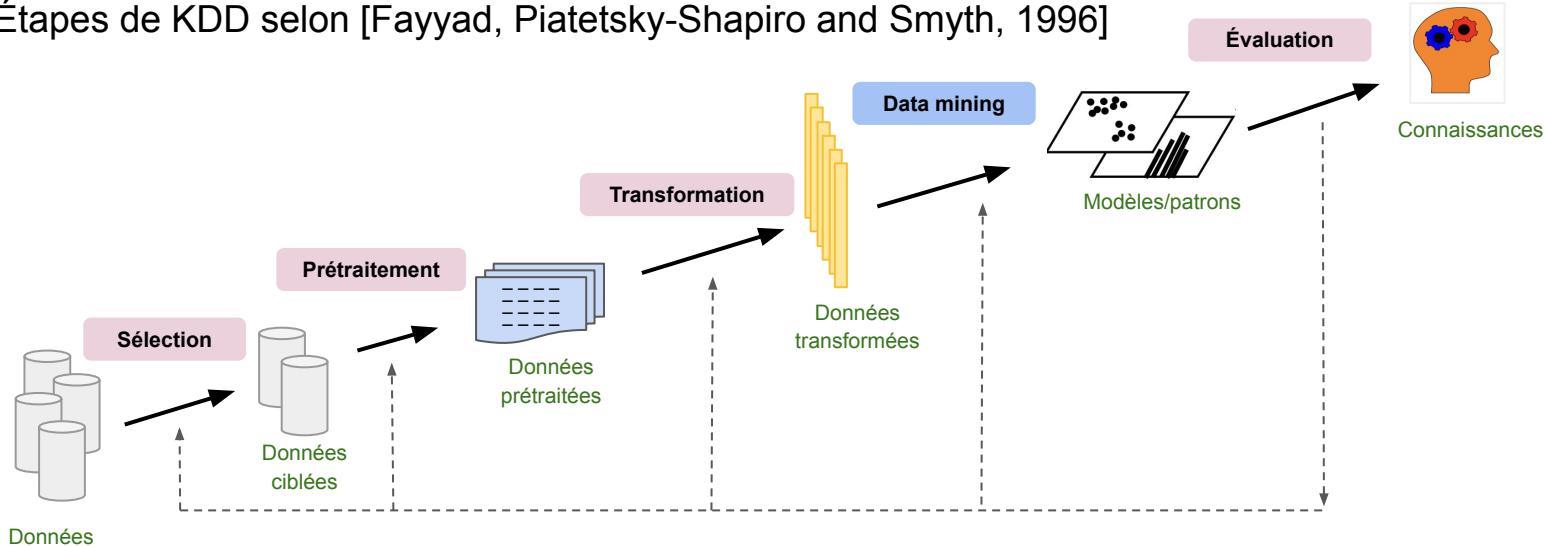
Processus KDD



*“Knowledge Discovery in Databases (KDD) is the **non-trivial process** of identifying **valid, novel, potentially useful, and ultimately understandable patterns** in data.”* [Fayyad, Piatetsky-Shapiro and Smyth, 1996]

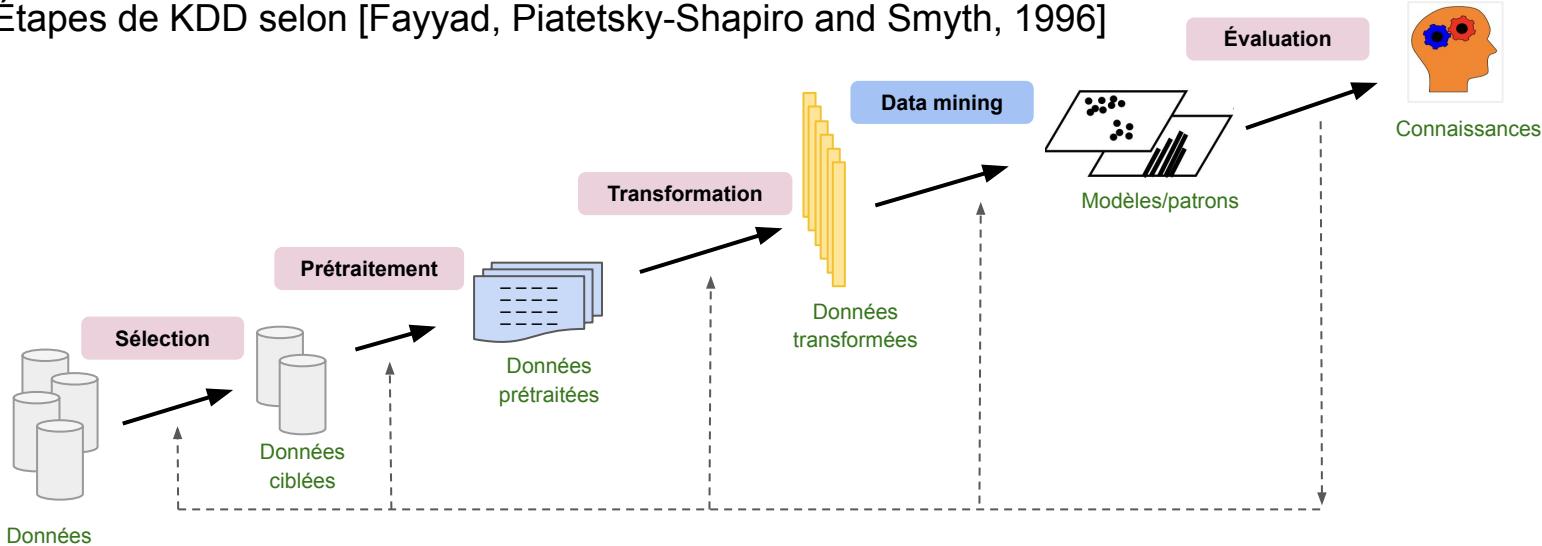
Processus KDD

Étapes de KDD selon [Fayyad, Piatetsky-Shapiro and Smyth, 1996]



Processus KDD

Étapes de KDD selon [Fayyad, Piatetsky-Shapiro and Smyth, 1996]

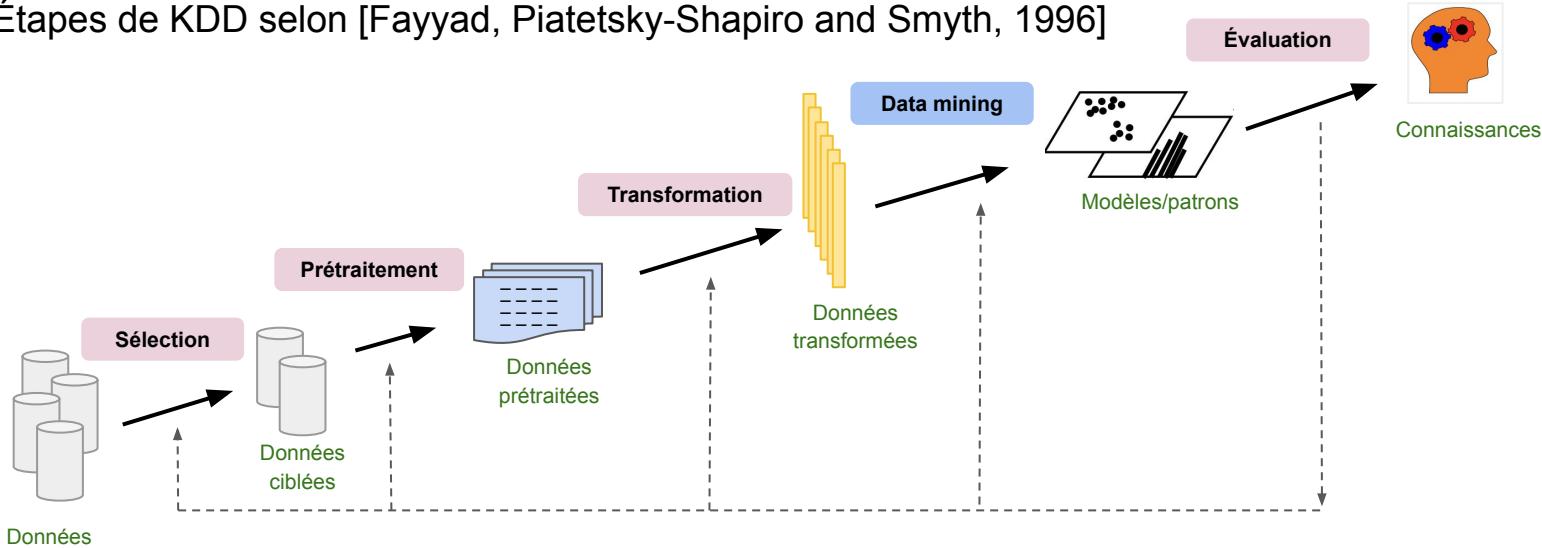


Sélection

- Sélection des données pertinentes
- Focus en une partie
- Fichier / BD

Processus KDD

Étapes de KDD selon [Fayyad, Piatetsky-Shapiro and Smyth, 1996]



Sélection

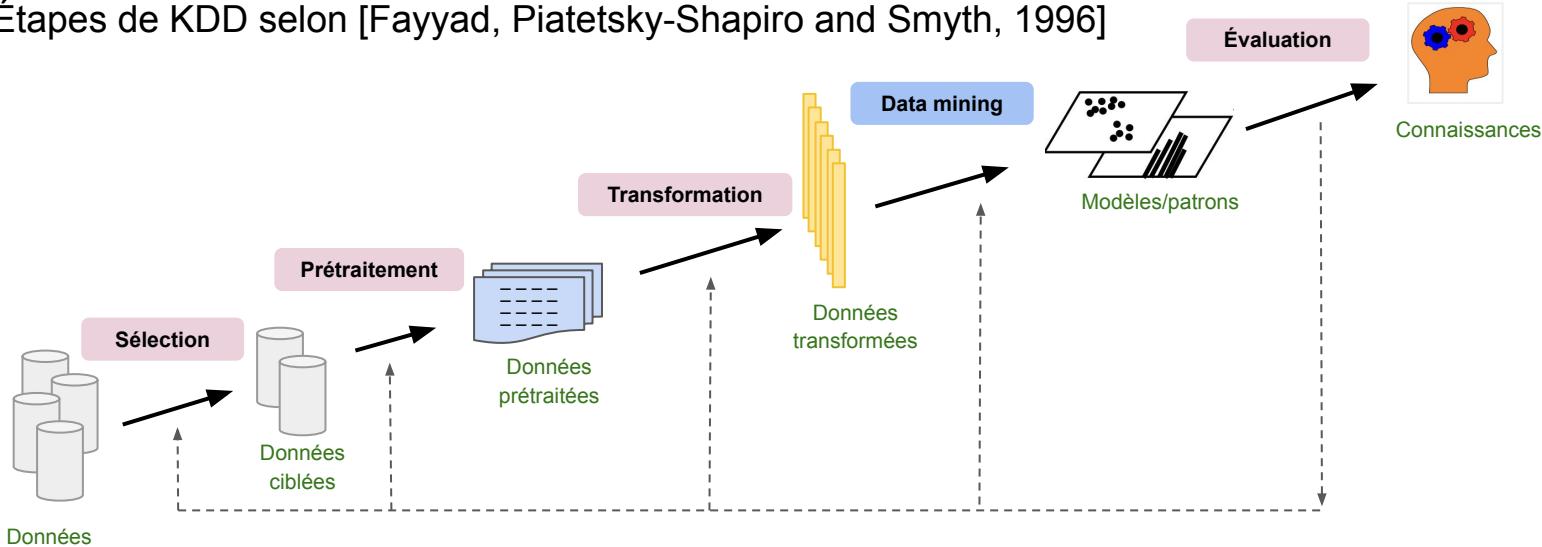
- Sélection des données pertinentes
- Focus en une partie
- Fichier / BD

Prétraitement

- Intégration des données de différentes sources
- Suppression du bruit
- Valeurs manquantes

Processus KDD

Étapes de KDD selon [Fayyad, Piatetsky-Shapiro and Smyth, 1996]



Sélection

- Sélection des données pertinentes
- Focus en une partie
- Fichier / BD

Prétraitement

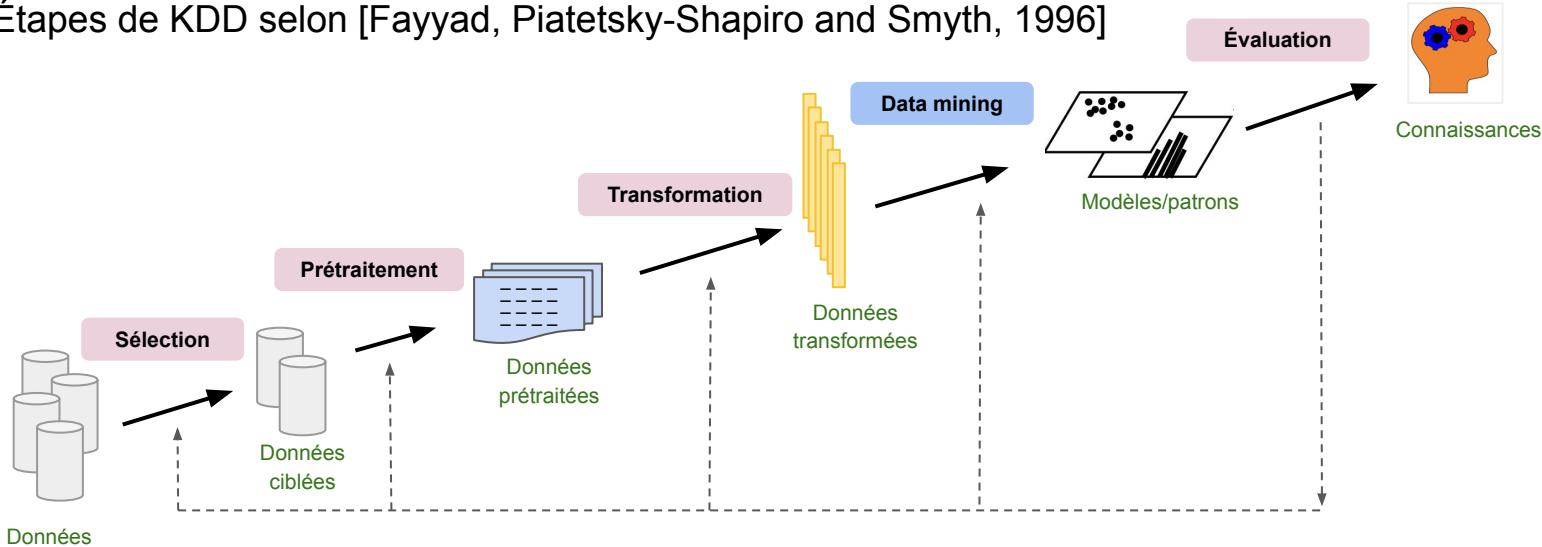
- Intégration des données de différentes sources
- Suppression du bruit
- Valeurs manquantes

Transformation

- Sélection de features utiles/pertinentes
- Transformation des descripteurs
- Réduction de dimension

Processus KDD

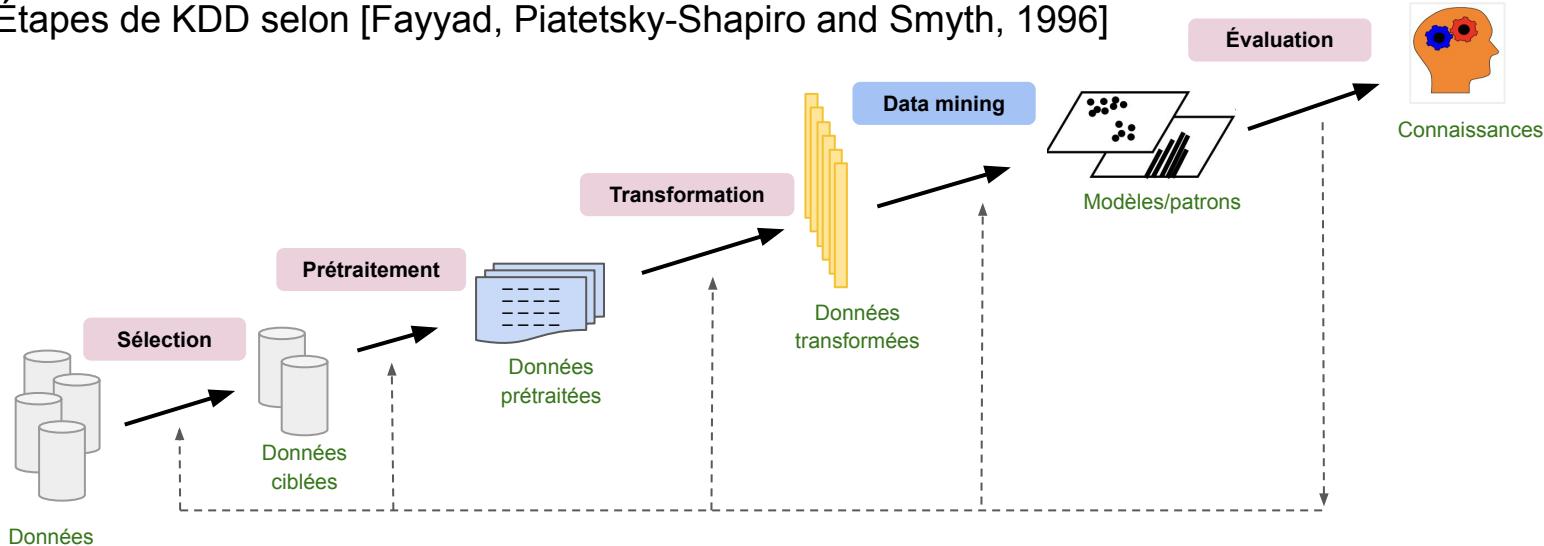
Étapes de KDD selon [Fayyad, Piatetsky-Shapiro and Smyth, 1996]



Sélection	Prétraitement	Transformation	Data mining
<ul style="list-style-type: none">- Sélection des données pertinentes- Focus en une partie- Fichier / BD	<ul style="list-style-type: none">- Intégration des données de différentes sources- Suppression du bruit- Valeurs manquantes	<ul style="list-style-type: none">- Sélection de features utiles/pertinentes- Transformation des descripteurs- Réduction de dimension	<ul style="list-style-type: none">- Application d'algorithme- Construction de modèle(s)

Processus KDD

Étapes de KDD selon [Fayyad, Piatetsky-Shapiro and Smyth, 1996]



Sélection	Prétraitement	Transformation	Data mining	Évaluation
<ul style="list-style-type: none">- Sélection des données pertinentes- Focus en une partie- Fichier / BD	<ul style="list-style-type: none">- Intégration des données de différentes sources- Suppression du bruit- Valeurs manquantes	<ul style="list-style-type: none">- Sélection de features utiles/pertinentes- Transformation des descripteurs- Réduction de dimension	<ul style="list-style-type: none">- Application d'algorithmes- Construction de modèle(s)	<ul style="list-style-type: none">- Évaluation des modèles- Choix des métriques

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

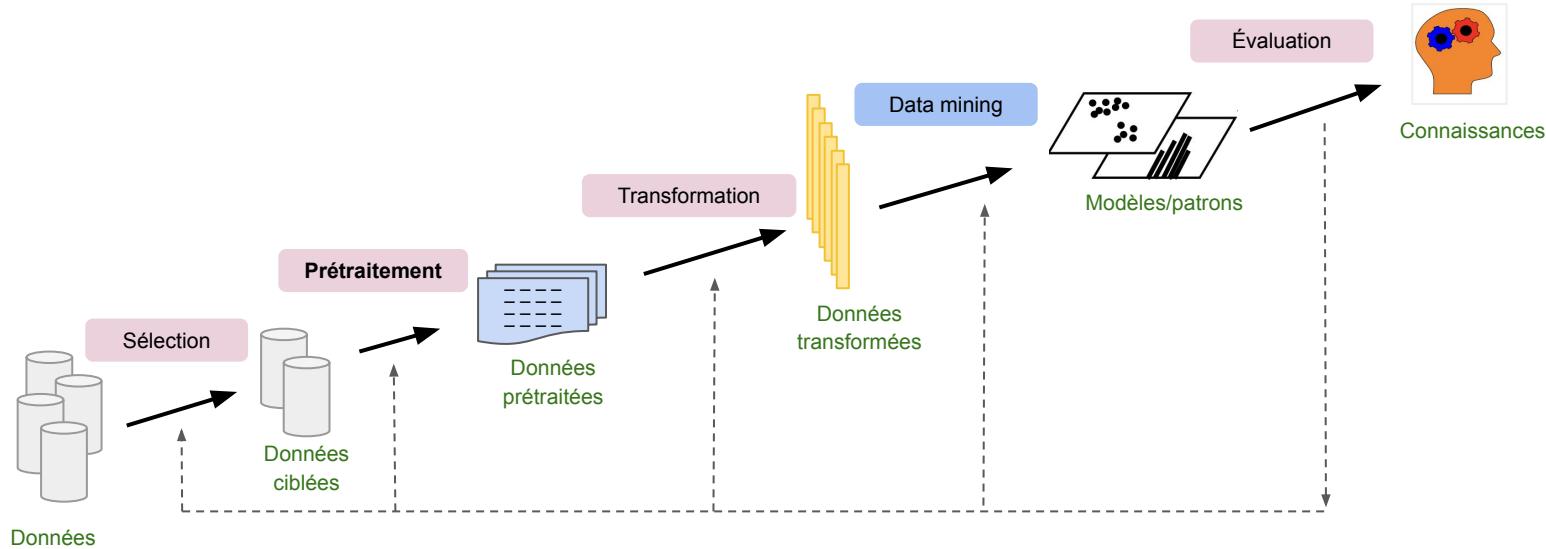
Information (requêtes)

- 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

Processus KDD



Processus KDD : prétraitement

- **Prétraitement des données** (nettoyage) : a pour but de résoudre le problème de la consistance des données.
- Inconsistances de données peuvent être:
 - locales à un enregistrement
Exemples : erreur de frappe, valeur aberrante, valeur manquante
 - locales à une source
Exemples : une même personne a deux adresses différentes
 - Entre sources :
Exemples :
 - Différence de codages pour une même donnée (“M=F” ou “1=2” pour le sexe)
 - Différence d’unités (prix en euro et en dollar)
 - Différence de granularités (nbre d’heures travaillées par semaine et par mois)
 - Différence de plages de valeurs (tranches d’âge : [11-20];[21-30];[31-40] et [15-30];[31-50])
 - Différence de formats (/jj/mm/aa vs. /mm/jj/aa/ ou adresse)
 - Utilisation de synonymes (“sans emploi”=”chômeur”)

Processus KDD : prétraitement

- **Valeur aberrantes :**

- Exclure toutes les valeurs $\notin [\text{Moy} - \text{EcartType}, \text{Moy} + \text{EcartType}]$
- Exclure pour le prix (âge, distance) des valeurs négatives

- **Valeurs manquantes :**

- Exclure les enregistrements incomplets
- Remplacer les données manquantes :
 - Remplacé
 - Calculé
 - Hérité

Processus KDD : transformation

- **Transformation des données** : afin qu'elles soient exploitable par des données de modélisation
 - Modification des types (ancienneté d'un client mieux que la date du premier achat et la date du dernier achat)
 - Normalisation selon une échelle uniforme
- Représentation des données :
 - verticale (pour les attributs discrets)

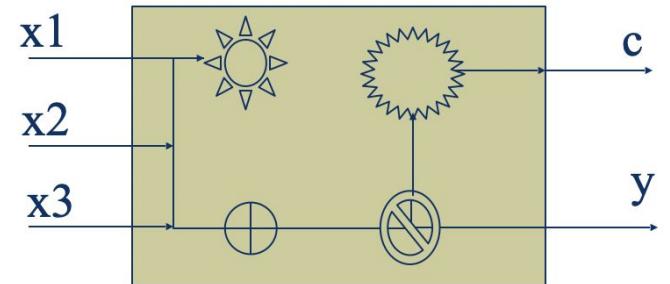
	Attribut
Objet 1	Val1
Objet 2	Val2
Objet 3	Val3

- horizontale

	Val1	Val2	Val3
Objet 1			
Objet 2			
Objet 3			

Processus KDD : data mining

- **Data mining** : consiste à choisir la technique à utiliser permettant d'obtenir un modèle/pattern
 - Dépendance :
 - au type de problème (tâche)
 - au type de données



x_1	x_2	x_3	y
1	10	100	alpha
2	20	200	beta

Processus KDD : data mining

- **Data mining** : consiste à choisir la technique à utiliser permettant d'obtenir un modèle/pattern
 - **Pattern** : une structure **caractéristique** possédée par un **petit nombre d'observations**: niche de clients à forte valeur, ou au contraire des clients à haut risque
Outils: classification, visualisation par réduction de dimension, règles d'association.
 - **Modèle** : Un modèle est un résumé global des relations entre variables, permettant de **comprendre** des phénomènes, et d'émettre des **prévisions**.
«Tous les modèles sont faux, certains sont utiles» selon (Box, G.E.P. and Draper, N.R.: Empirical Model-Building and Response Surfaces, p. 424, Wiley, 1987)

En data mining, les modèles sont :

- linéaires ou non
- explicites ou implicites : réseaux de neurones, arbre de décision, SVM, *Naive bayes*, etc

⇒ *Les modèles/patterns ne sont pas issus d'une théorie mais de l'exploration de données*

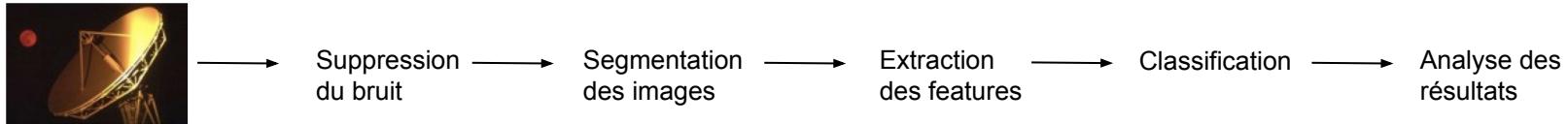
- **Évaluation:**
 - Une fois que les modèles ont été obtenus à partir de diverses méthodes et itérations d'exploration de données
 - Ces modèles doivent être représentés sous des formes discrètes telles que
 - des graphiques à barres,
 - des diagrammes circulaires,
 - des histogrammes,
 - *etc.*
 - pour étudier l'impact des données collectées et transformées au cours des étapes KDD précédentes.
 - Cela aide également à évaluer l'efficacité d'un modèle de données particulier au vu du domaine.

Processus KDD : conclusion

[Fayyad, Piatetsky-Shapiro and Smyth, 1996] (<https://cdn.aaai.org/KDD/1996/KDD96-014.pdf>)

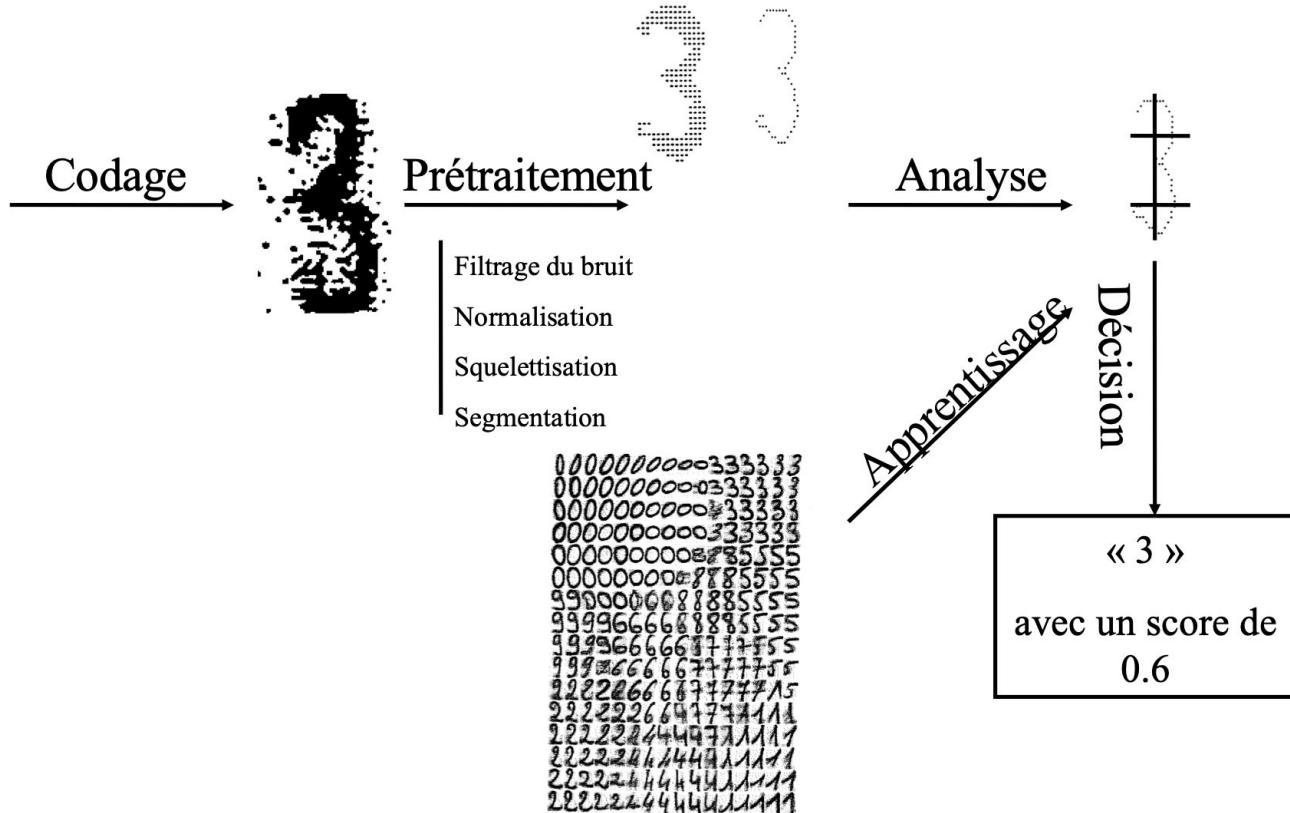
1. **Data cleaning** to remove noise and inconsistent data.
2. **Data integration**, where multiple data sources may be combined.
3. **Data selection**, where data relevant to the analysis task are retrieved from the database.
4. **Data transformation**, where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. **Data mining**, which is an essential process where intelligent methods are applied to extract data patterns.
6. **Pattern evaluation** to identify the truly interesting patterns representing knowledge based on interesting measures.
7. **Knowledge presentation**, where visualization and knowledge representation techniques are used to present mined knowledge to users.

Applications KDD : Astronomie



- Architecture de [Fayyed et al., 1996]
- Méthode de classification : arbre de décision
- Évaluation :
 - plus rapide qu'une classification manuelle
 - classifier des objets célestes très petits

Applications KDD : Reconnaissance des caractères



Applications KDD : Marketing

- Customer segmentation [Piatetsky-Shapiro et al., 2000] :
 - partitionner les consommateurs par rapport à leurs achats
 - établir de nouvelle politique tarifaire
- Évaluation :
 - plus rapide qu'une classification manuelle
 - classifier facilement des nouveaux consommateurs

Produits fréquemment achetés ensemble



Prix éditeur : EUR 63,00
Prix pour les trois: EUR 51,22

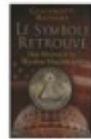
Ajouter ces trois articles au panier

[Afficher la disponibilité du produit et le mode de livraison](#)

- [Cet article : Le Symbole perdu de Dan Brown](#)
- [La première nuit de Marc Levy](#)
- [L'Echappée belle de Anna Gavalda](#)

Les clients ayant acheté cet article ont également acheté

Page 1 sur 17



[Le symbole retrouvé : Dan Brown et le mystère... de Eric Giacometti](#)
 (1)
EUR 15,11



[La première nuit de Marc Levy](#)
 (22)
EUR 19,95



[Le Symbole Perdu Décodé de Alain Bauer](#)
EUR 15,20



[La forêt des Mânes de Jean-Christophe Grangé](#)
 (49)
EUR 21,75



[Le miroir de Cassandre de Bernard Werber](#)
 (20)
EUR 21,75



[La stratégie Bancroft de Robert Ludlum](#)
 (2)
EUR 19,86



Applications KDD

- Marketing ciblé : population à cibler pour publipostage, leader de communautés
- Gestion et analyse de marchés, gestion de stocks : profils de clients, effets des soldes ou de campagnes publicitaires, Quand commander? Quelle quantité?
- Détection de fraudes : télécommunications, secteur bancaire
- Analyse financière et analyse de risques : maximiser le profit d'un portefeuille, accorder un crédit
- Bioinformatique, médecine et pharmacie : analyse des séquences ADN, aide au diagnostic, choix de médicaments
- Analyse de réseaux sociaux et fouille de textes : détermination de la polarité, analyse d'opinions
- ...

Cas d'étude (1)

- **Situation :**
 - Gestion marketing d'un opérateur téléphonique
 - fidélisation des clients
- **Solution :**
 - Offre d'un nouveau téléphone
 - Identification des clients qui ont une grande chance de partir (Par exemple 3 mois avant la fin de leurs contrats)
 - Si c'est le cas
 - On offre le téléphone
 - Sinon
 - Ne rien offrir

Cas d'étude (2)

- **Situation :**
 - Gestion dans une agence d'assurance
 - Évaluation de la prime pour un jeune conducteur
- **Solution :**
 - Analyser les données de l'agence relatives aux accidents
 - Définir un modèle pour estimer la probabilité d'avoir un accident à partir des données
Descripteurs: âge, adresse, voiture, etc
 - Construire le modèle
 - Fixer le montant de la prime en fonction de la probabilité

Cas d'étude (3)

- Situation :
 - Gestion de risques dans une banque
 - Détection des achats frauduleux avec la carte bancaire
- Solution :
 - Analyser les données de la banque relatives aux achats déclarés frauduleux
 - Construire un modèle de fraude
 - Construire un modèle de comportement normal à partir de l'historique de la carte bancaire
 - Établir un score de similarité entre
 - le profil normal
 - l'utilisation récente
 - Si le score > seuil
 - fraude
 - Sinon
 - pas fraude

Cas d'étude (4)

- **Situation :**
 - Gestion des diagnostics médicaux dans les hôpitaux
 - Détection des tumeurs
- **Solution :**
 - Analyser les données de l'hôpital relatives aux tumeurs
 - symptômes
 - résultats d'examen d'imagerie
 - etc
 - Définir un modèle et le construire
 - Détecter la tumeur

Data mining ou non

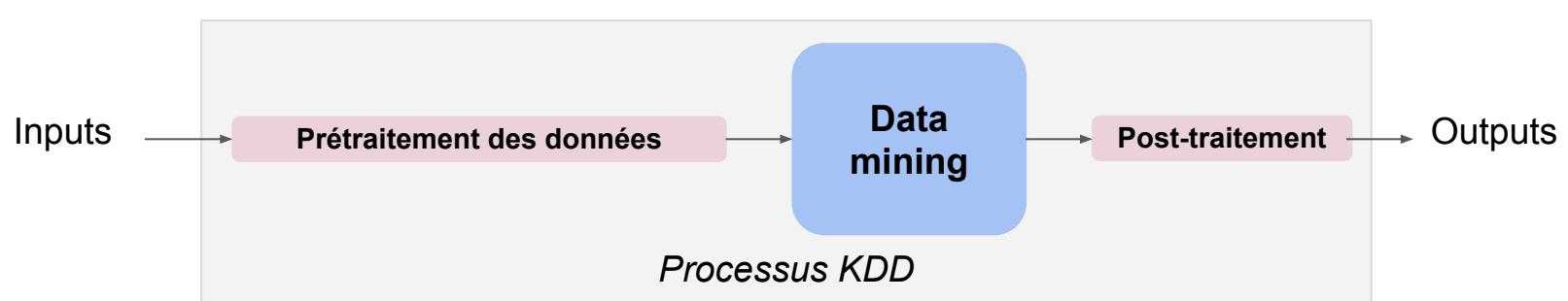
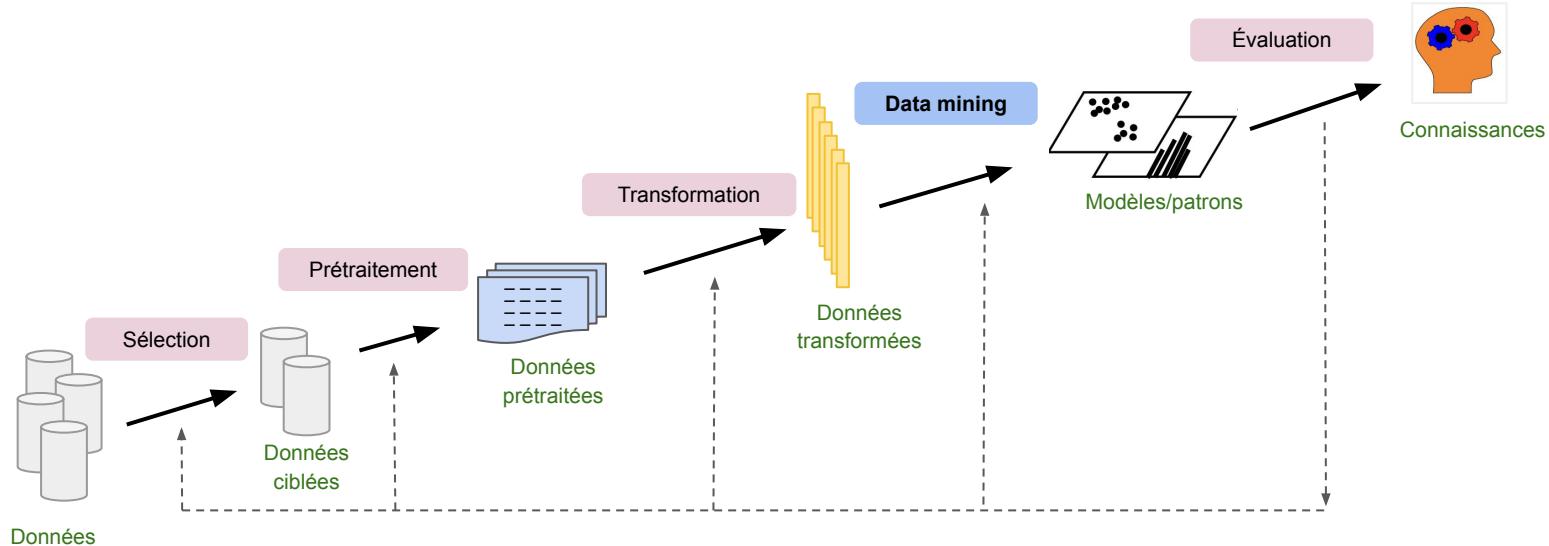
	Oui	Non
Chercher le salaire de l'employé <i>Camelin</i>		
Les supporters achètent de la bière le samedi et de l'aspirine de dimanche		
Interroger un moteur de recherche pour avoir des informations sur la fouille de données		
Regrouper un ensemble de documents en fonction de leurs contenus		
Identifier les clients ayant acheté plus de 500 euros le mois précédent		
Les personnes qui réalisent l'action A réalisent dans le mois qui suit l'action B		
Regrouper les gènes en fonction de leurs caractéristiques		
Trouver tous les clients qui ont acheté du lait		
Identifier les clients qui ont des habitudes d'achat similaires		

Data mining ou non

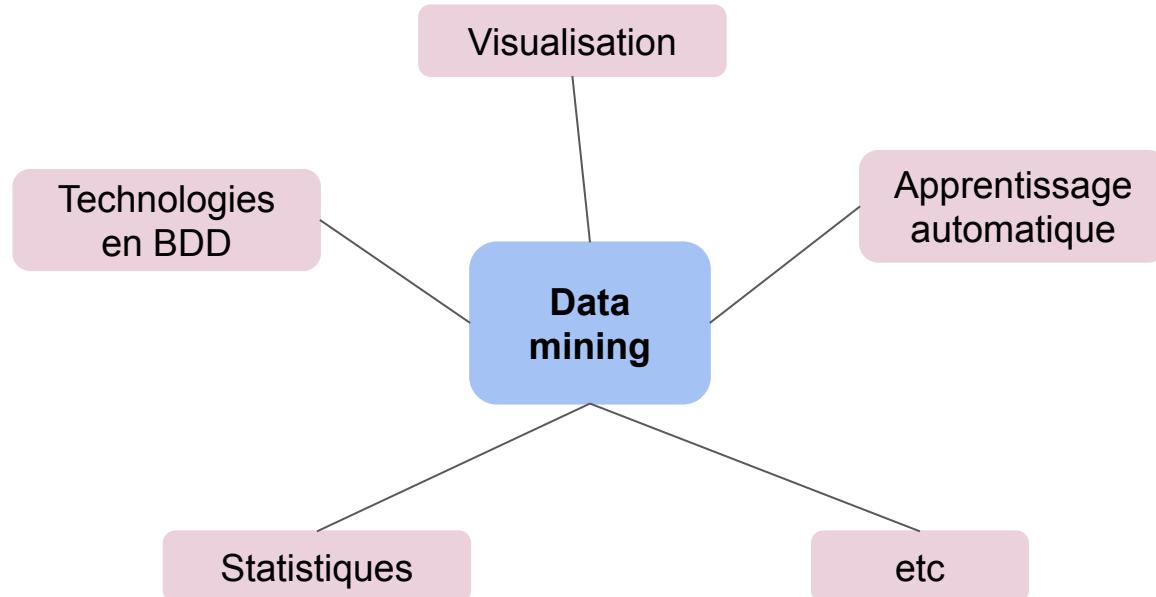
	TEMPS	HUMIDITE	VENT	TENNIS
Ex1	Soleil	Haute	Oui	Oui
Ex2	Soleil	Basse	Non	Non
Ex3	nuageux	Basse	Oui	Oui
Ex4	pluvieux	Haute	Oui	Non
Ex5	pluvieux	Basse	Oui	Non
Ex6	Soleil	Basse	Oui	Oui
Ex7	pluvieux	Basse	Non	Non
	<i>Soleil</i>	<i>haute</i>	<i>Non</i>	<i>?</i>

Va-t-on jouer si il y a du soleil, beaucoup d'humidité et pas de vent ?

Data mining au centre du processus KDD



Data mining au centre du processus KDD



- **Approche prédictive** (modélisation):
 - **Prédire** : vise à extrapoler de nouvelles informations à partir des informations présentes
 - explique les données
 - il y a une variable (qualitative ou quantitative) à expliquer
 - englobe des **techniques supervisées**
 - produit des **modèles de prédiction** (classement ou régression)
 - Exemples d'application:
 - Cibler les potentiels clients d'un produit donnée pour réduire les coûts des démarches téléphoniques
 - Déetecter les transactions frauduleuses de cartes bancaires
 - Analyser les satisfactions des clients

- **Approche descriptive** (recherche de patterns):

- **Décrire** : vise à mettre en évidence des informations présentes mais cachées par le volume de données (c'est l'exemple de segmentation de la clientèle et la recherche d'association de produits sur les tickets de caisse)
- réduit, résume et synthétise les données
- il n'y a pas de variable à expliquer
- englobe des techniques non supervisées
- produit des **modèles de classification** (*clustering*)
- Exemple d'application :
 - Créer automatiquement et sans à priori des profils de clients potentiellement intéressés par les mêmes produits
 - Catégoriser thématiquement les documents en fonction des termes fréquents qui les composent

Conclusion

	KDD	Data mining
Définition		
Étapes		
Techniques utilisées		
Outputs		