Institut Supérieur d'Informatique et de Multimédia de Sfax

Auditoire: 2ème année LSI-ADBD

Cours:
Analyse et Fouille de données

Souhir BOUAZIZ AFFES

souhir.bouaziz@isims.usf.tn

Plan du cours

- **Chapitre 1:** Extraction des Connaissances à partir des Données
- **Chapitre 2:** Introduction aux méthodes multivariées
- **Chapitre 3:** Analyse en composantes principales
- **Chapitre 4:** Méthodes de regroupement: clustering
- **Chapitre 5:** Règles d'association et Fouille de motifs
- **Chapitre 6:** Apprentissage supervisé
- **Chapitre 7:** Analyse discriminante linéaire
- **Chapitre 8:** Régression linéaire multiple
- **Chapitre 9:** Régression logistique

Institut Supérieur d'Informatique et de Multimédia de Sfax

Auditoire: 2ème année LSI-ADBD

Chapitre 1:

Extraction des Connaissances à partir des Données

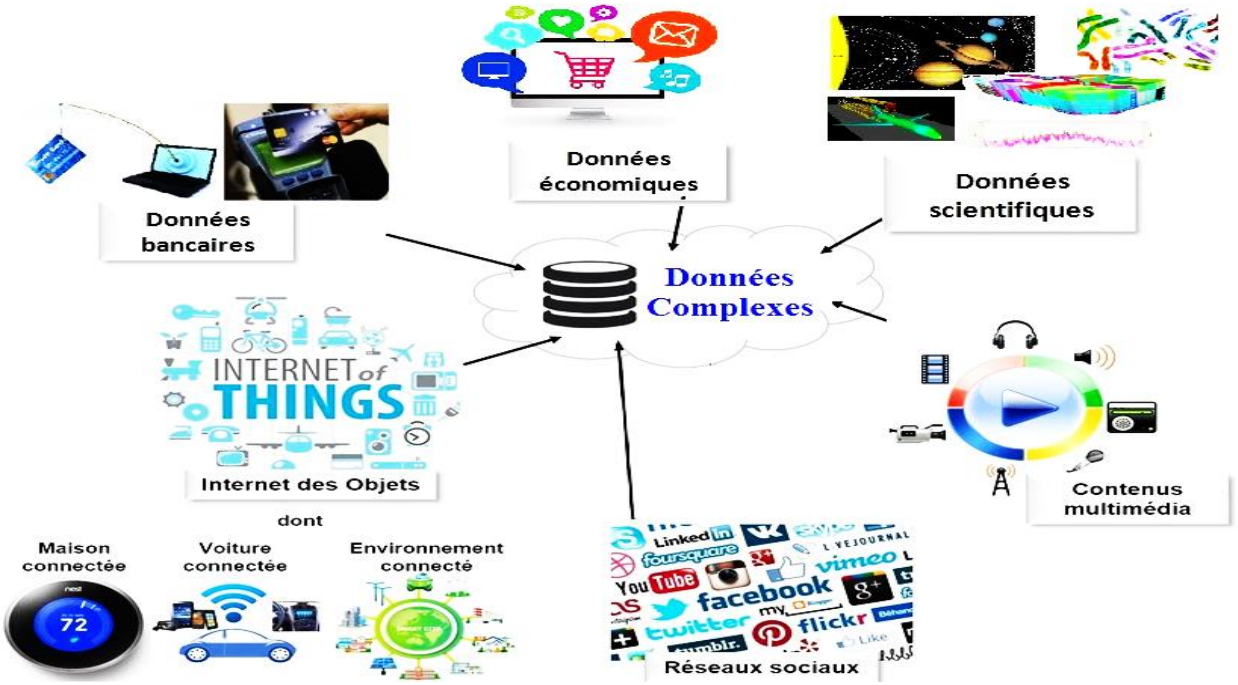
Souhir BOUAZIZ AFFES
souhir.bouaziz@isims.usf.tn

Plan

- Introduction
- Extraction des connaissances à partir des données: ECD
- De données aux connaissances
- Applications de ECD
- Cycle de vie d'un processus ECD
- Exemple de problème
- Étapes du processus ECD

Introduction (1/5)

- Volume de données collectées est en croissance continue

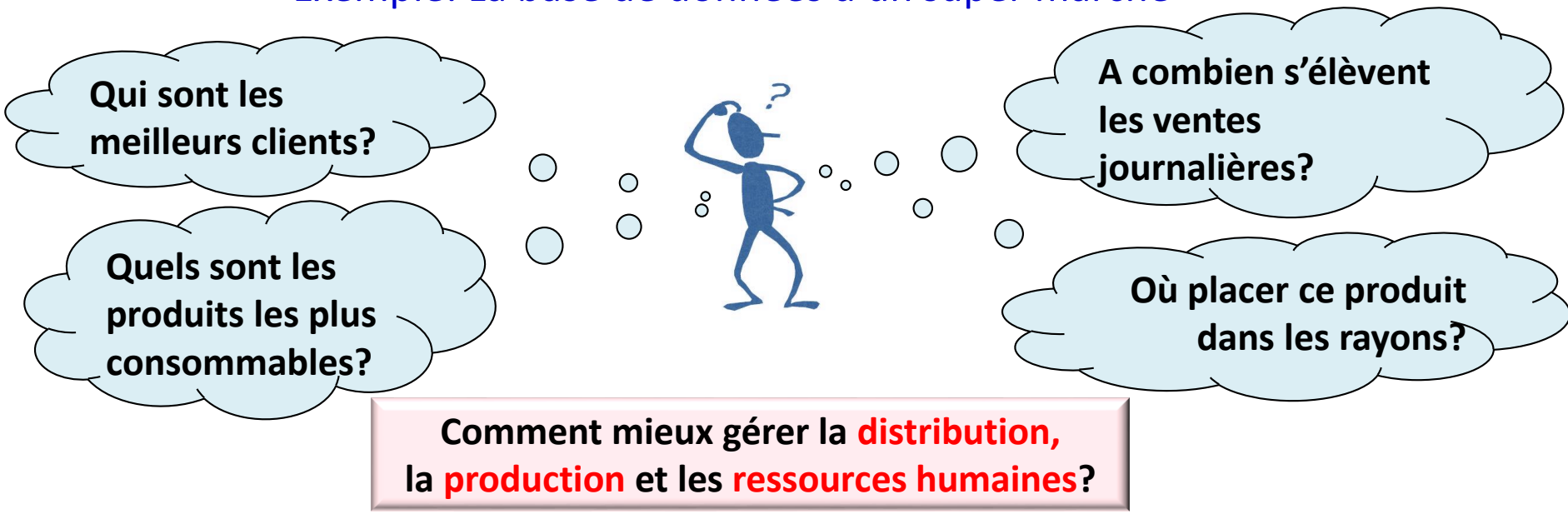


- Experts dépassés par les volumes

Introduction (2/5)

- Problème de l'**explosion de données**
  - Les outils automatiques de collecte de données font que les Bases de Données (BD's) contiennent énormément de données

Exemple: La base de données d'un super marché



- Beaucoup de données mais peu de **connaissances** !

Introduction (3/5)

- Problème de l'**explosion de données**
  - Expérience de l'entreprise : ses clients et leur comportement

- coûteuse en stockage
- inexploitée

Comment et à quelles fins utiliser cette expérience accumulée

Introduction (4/5)

- Comment peut-on prendre de décision ?
  - Où et quand faire la publicité de nos produits?
  - Comment les banques peuvent suspecter un vol de carte avant même que le propriétaire soit au courant?
    - Quels sont les comportements suspicieux? Quand? Comment?

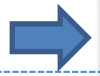
- Faudra-t-il utiliser pour ça des **logiciels d'aide à la décision**?
- Comment fonctionnent ces logiciels?
  - Quels sont leurs entrées?
  - Qu'est ce qu'ils peuvent nous fournir en matière d'aide à la décision?

Introduction (5/5)

- Historique des outils d'aide à la décision:
  - **L'informatique décisionnelle « *Business Intelligence* »:** s'est développé dans les années 70
    - Outils d'édition de rapports, de statistiques, de simulation, d'optimisation, …
  - **Les systèmes experts:** sont apparus au début des années 70
    - les systèmes experts Dendral en chimie organique et les systèmes experts Mycin en médecine avec des bons résultats
    - La formulation sous forme de règles trouve vite ses limites vue la complexité de plusieurs applications
    - **Il nous faut donc des outils d'aide à la décision** (pour aider les experts et non pas prendre leur place)
  - **Les outils de fouille de données « *Data Mining* »:** sont arrivés à maturité vers les années 90
    - méthodes classiques : outils généralistes de l'informatique ou des mathématiques: statistiques descriptives, analyse en composantes principales, etc.
    - méthodes sophistiquées : élaborées pour résoudre des tâches bien définies: règles d'association, arbres de décision, les réseaux de neurones, etc.

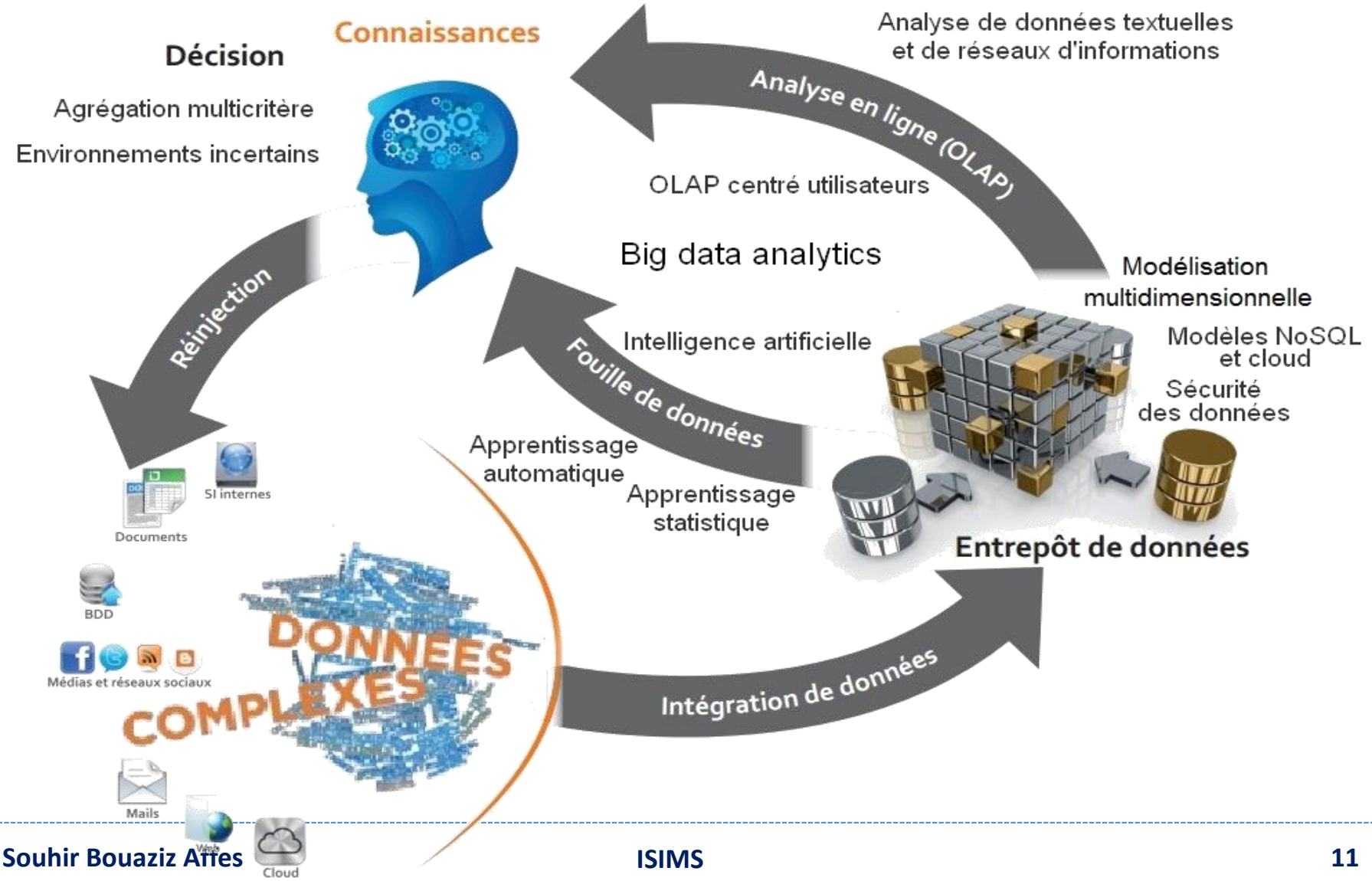
Extraction de Connaissances à partir des Données (1/6)

- **Questions** :
  - Pourquoi nous avons toujours tendance à archiver nos données?
  - Pourquoi il nous est difficile d'effacer nos anciens fichiers ?
  - Mais, est-ce que vraiment nous les utilisons? Est-ce qu'ils contiennent de l'information pertinente ou des connaissances dont nous avons besoin?
  - Pourquoi pas? Qu'est ce qu'il nous manque?
- **Objectifs:**
  - Tirer partie de la complexité des données disponibles
  - Il nous faut des moyens de prévision pour anticiper les changement de comportement
- **Solution:**
  - Extraction des Connaissances à partir des Données (ECD) pour pouvoir exploiter ces données



Processus de l'ECD

Extraction de Connaissances à partir des Données (2/6)



Extraction de Connaissances à partir des Données (3/6)

- Evolution des Bases de Données:
  - 1960s:
    - Collecte des données, création des BD's, IMS et le modèle réseau
  - 1970s:
    - Modèle et SGBD's relationnels, SQL, transactions, OLTP
  - 1980s:
    - Modèles de données et SGBD's avancés (relationnel étendu, OO, déductifs, etc.) et SGBD's dédiés (spatial, génomique, engineering, etc.)
  - 1990s—2000s:
    - Data mining et data warehousing, BD's multimédia, BD's sur le WEB

Extraction de Connaissances à partir des Données (4/6)

- **Extraction de Connaissances à partir des Données** (**ECD**) est un processus **non trivial** d'extraction de connaissances à partir de bases de données pour obtenir de **nouvelles** données :
  - **valides**
  - **potentiellement utiles**
  - et **compréhensibles** [Fayad et al., 96]
- ECD est un processus qui emploi des **techniques d'apprentissage automatique** et **intelligentes** pour **analyser** et **extraire** des connaissances à partir de grandes quantités de données
- Emergence
  - Des machines assez puissantes
  - Des algorithmes efficaces d'analyse et de fouille de données
  - Collections et sauvegardes de données améliorées

Extraction de Connaissances à partir des Données (5/6)

- Terminologie:
  - **Extraction de Connaissances à partir des Données** ou **Découverte de Connaissances dans les Données** (**Knowlege Discovery in Data: KDD**)
    - L'ensemble des processus de découvertes et d'interprétation de régularités dans des données
  - **Fouille de données (Data Mining)**
    - Classe de méthodes et d'algorithmes utilisés pour découvrir des régularités dans les données
  - **Entrepôt de données (Data Warehouse) et magasin de données (Data Mart)**
    - Grandes collections de données générales ou spécialisées
  - **Traitement analytique (Online Analytical Processing: OLAP)**
    - Technologie permettant d'effectuer des analyses de données multidimensionnelles au sein de bases de données

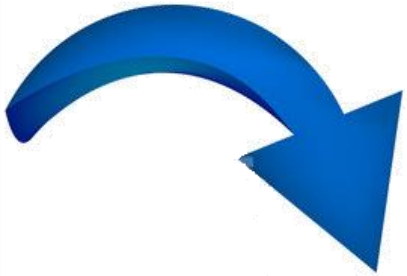
Extraction de Connaissances à partir des Données (6/6)

- Disciplines impliquées:
  - L'ECD fait intervenir des techniques issues d'autres domaines (domaine **pluridisciplinaire**)
    - Statistiques
    - Analyse de données
    - Intelligence artificielle
    - Apprentissage automatique (Machine Learning)
    - Reconnaissance de formes
    - Visualisation des données
    - Big Data
    - Etc.

De données aux connaissances (1/5)

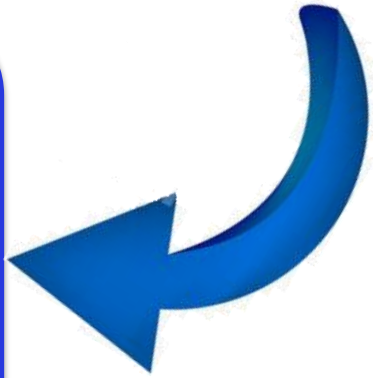
Données

Une donnée est le résultat direct d'une mesure, présentée sous forme conventionnelle



Informations

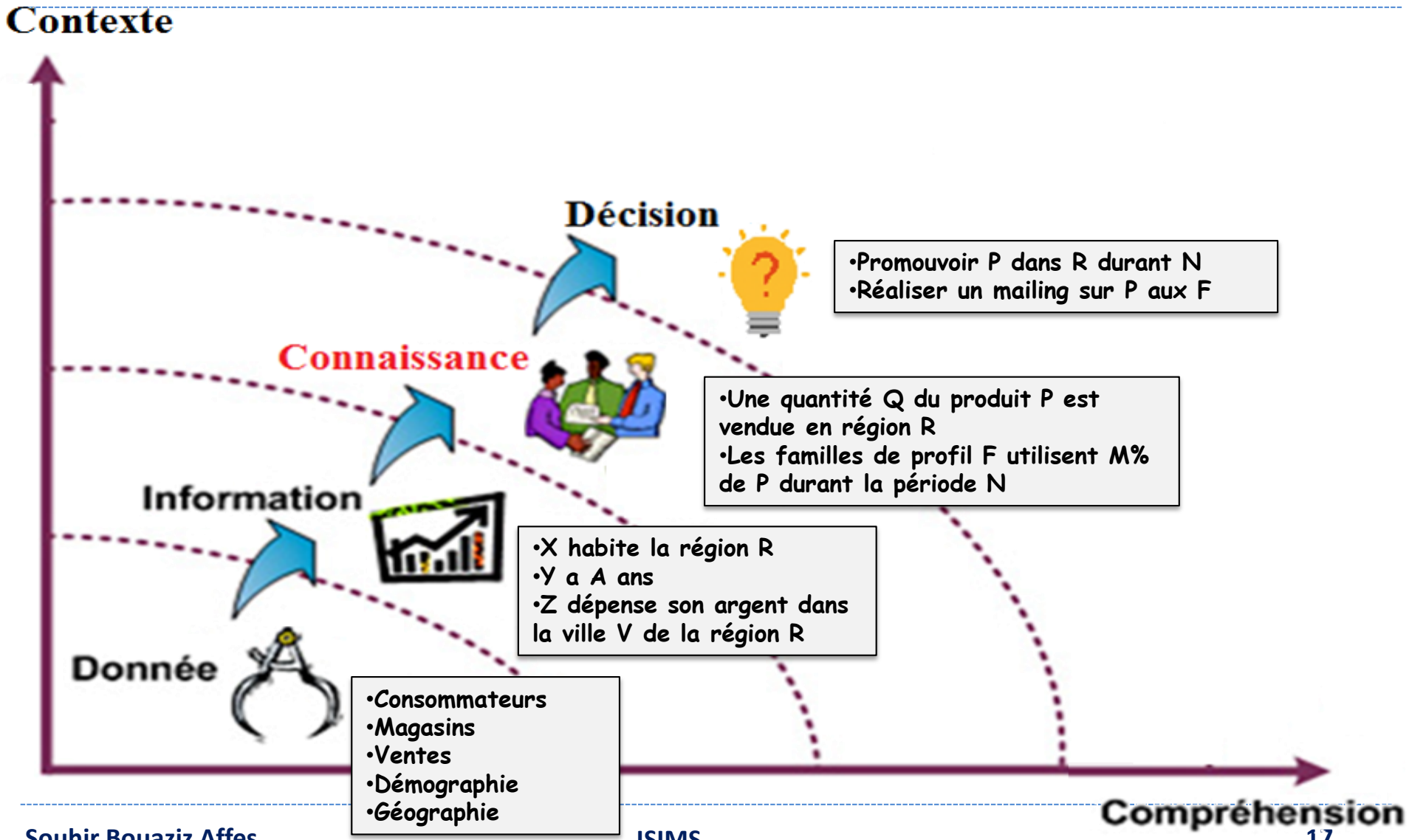
Une information est une donnée à laquelle un sens et une interprétation ont été donnés



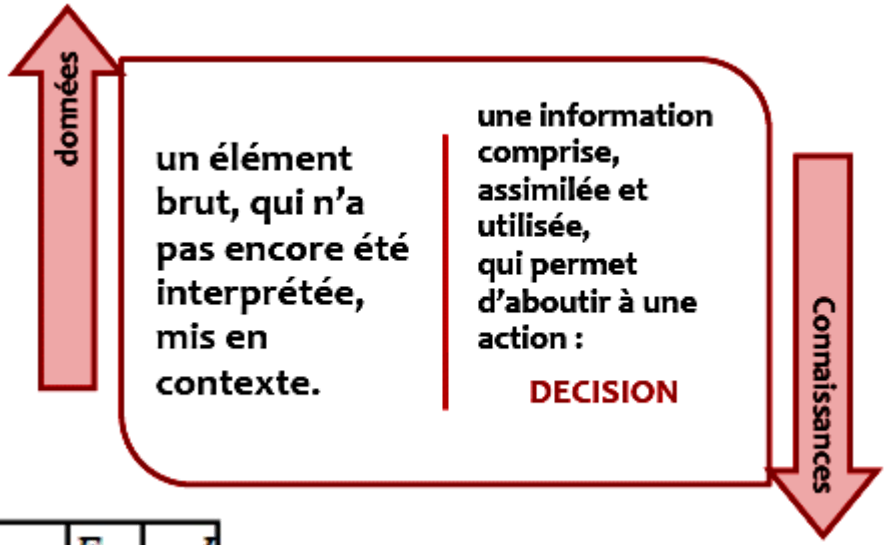
Connaissances

La connaissance est le résultat d'une réflexion sur les informations analysées

De données aux connaissances (2/5)



De données aux connaissances (3/5)



client	*M*	*A*	*R*	*E*	*I*
1	moyen	moyen	village	oui	oui
2	élevé	moyen	bourg	non	non
3	faible	âgé	bourg	non	non
4	faible	moyen	bourg	oui	oui
5	moyen	jeune	ville	oui	oui
6	élevé	âgé	ville	oui	non
7	moyen	âgé	ville	oui	non
8	faible	moyen	village	non	non

Donnée :

Client3 : âge = âgé, Niveau d'études = non

Information :

37,5 % des Clients consultent leurs comptes bancaires sur le Web.

Connaissance :

SI E=non **ALORS** I= non
SI E=oui **ET** A=moyen **ALORS** I=oui
SI E=oui **ET** A=âgé **ALORS** I=non
SI E=oui **ET** A=jeune **ALORS** I=oui

De données aux connaissances (4/5)

- **Exemple de données disponibles:**
  - **Transactions:** tickets de caisse, factures, communications téléphoniques
  - **Bases de données des entreprises :** Factures, Commandes, Suivi
  - **Téléphone portable:** Durée des communication, numéros appelés, abonnement, mobilité
  - **Satellites :** espace (Photos de corps célestes) et la terre (Reconnaissance automatique de formes: cartographie, Type de terrain: cartographie)
  - **Données du web:** récupération facile de pages ou de sites (Contenu des pages, liens entre les pages, historique des connexions)
  - **Données textuelles:** pages Web, fichiers (word, pdf, …), dépêches d'agence, digitalisation de bibliothèques
  - Etc.

De données aux connaissances (5/5)

- Types de connaissances extraites: Connaissances sous la forme de modèles de description permettant de:
  - **décrire le comportement** actuel des données et/ou
  - **prédire le comportement** futur des données.
  - **Analyses**
    - **Exemple:** distribution du trafic routier en fonction de l'heure
  - **Règles**
    - **Exemple:** **si** un client a acheté un produit **alors** il sera intéressé par un autre
  - **Attribution de scores de qualité**
    - **Exemple:** score de fidélité au client
  - **Classification d'entités**
    - **Exemple:** mauvais payeurs

Applications de ECD

- Banque, Finance, Assurances
- Marketing direct
- Détection de fraudes
- Industrie
- Santé, Génétique
- E-Commerce et Web
- Industrie pharmaceutique
- Assurances
- Tourisme, Loisirs
- Sécurité
- Télécommunications
- etc.

Cycle de vie d'un processus ECD

« Questions stratégiques »

Identification du problème

DataWarehouse SGBD

Transformation des données en information

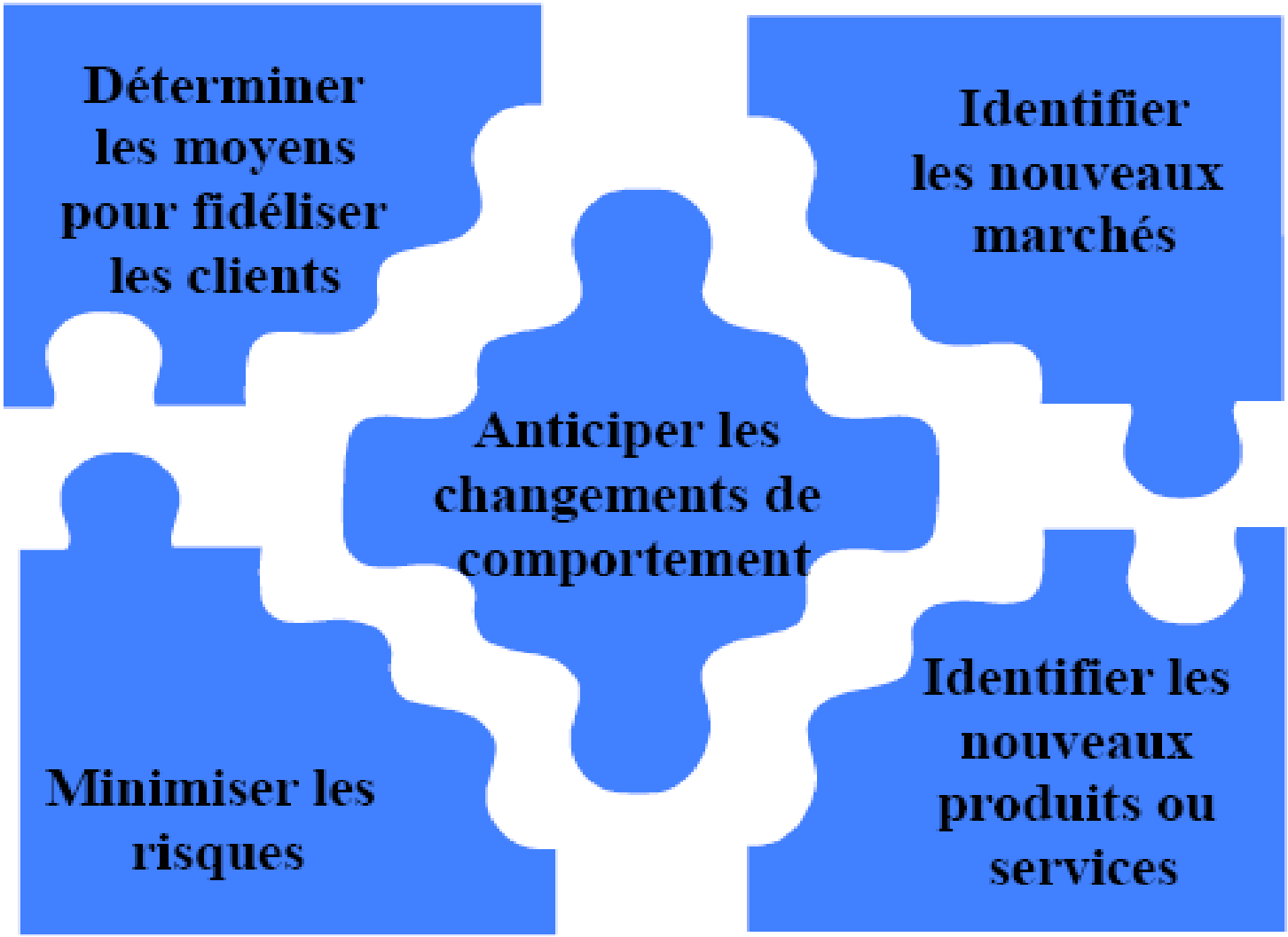
Processus de Data Mining

Agir sur l'information

Rapports, graphique, EIS (Executive Information System)

Evaluer

Cycle de vie d'un processus ECD:
Enjeux stratégiques

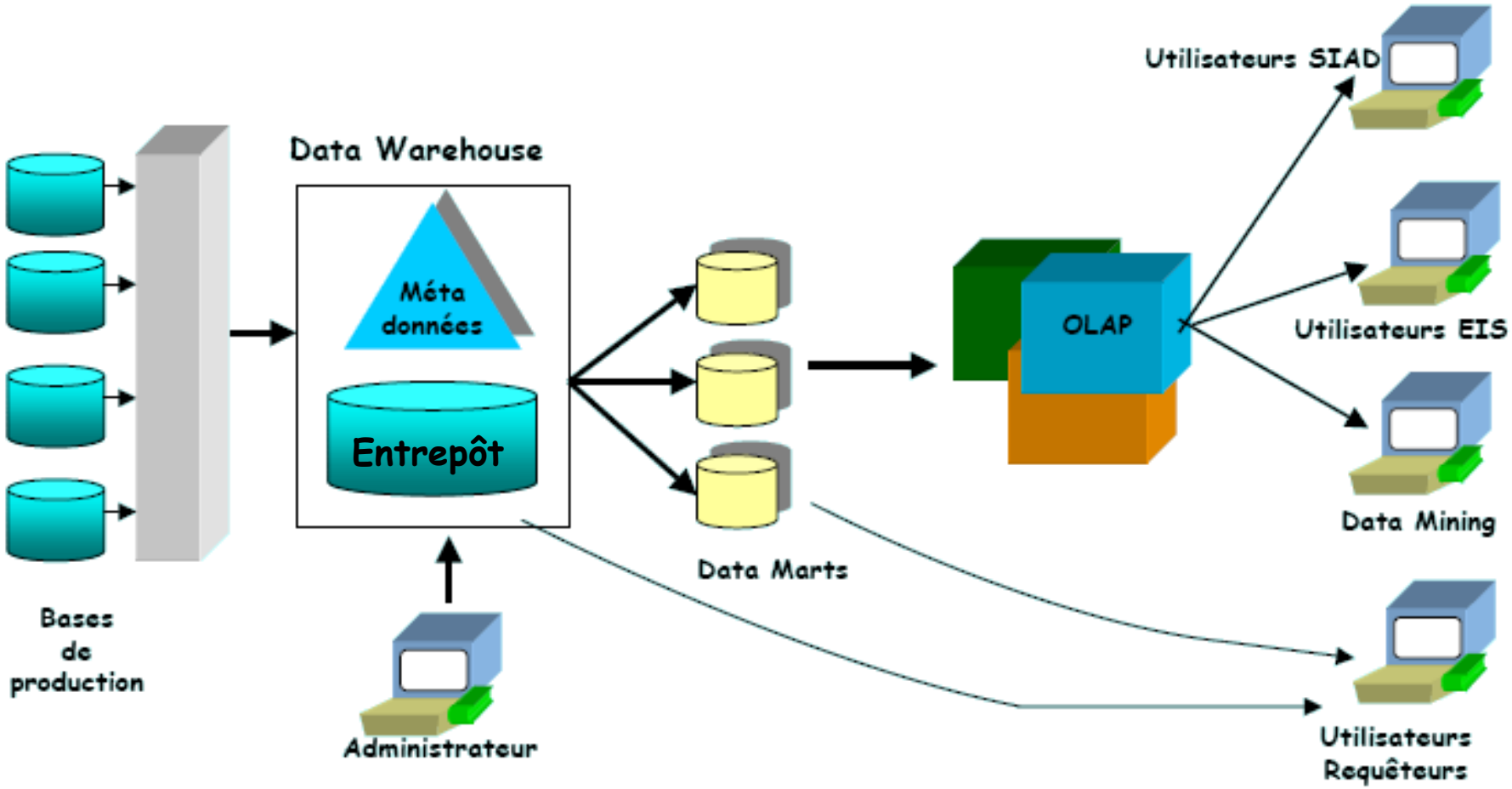


Cycle de vie d'un processus ECD: Data Warehouse (1/6)

- Le concept d'**entrepôt de données** (data warehouse) a été formalisé pour la première fois en 1990 par **Bill Inmon**.
- D'après Bill Inmon :
  - "*Un DW est une collection de données thématiques, intégrées, non volatiles et historisées, organisées pour la prise de décision*."
    - **Thématiques :** thèmes par activités majeures
    - **Intégrées :** divers sources de données (bases de données relationnelles, objets, spatiales, temporelles, textuelles, multimédia, etc.)
    - **Non volatiles :** ne pas supprimer les données du DW
    - **Historisées :** trace des données, suivre l'évolution des indicateurs
- Le data warehouse est une base d'information organisée pour répondre aux besoins spécifiques de l'aide à la décision

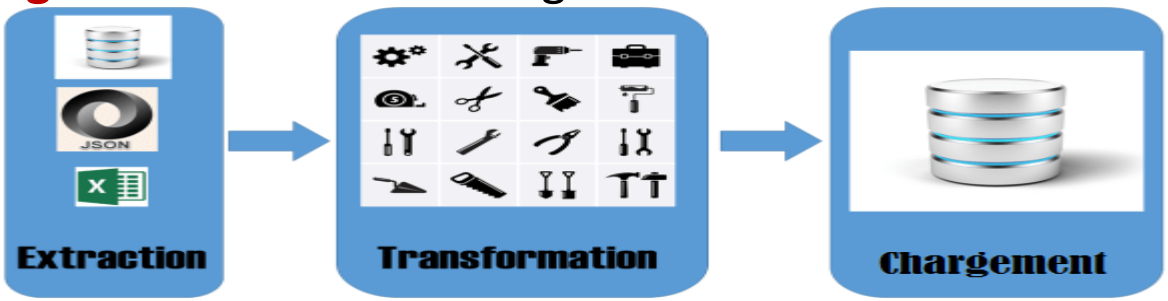
Cycle de vie d'un processus ECD: Data Warehouse (2/6)

- **Architecture et fonctionnement:**



Cycle de vie d'un processus ECD: Data Warehouse (3/6)

- **Les bases de production:**
  - Constituent les systèmes opérants de l'entreprise et
  - Correspondent à l'ensemble des applications informatiques utilisées au quotidien dans l'entreprise
- **Système d'alimentation :**
  - L'alimentation d'un DW est une procédure qui s'effectue en plusieurs étapes :
    - **Extraction:** la sélection de l'information dans la base de données
      - une fonction classique d'interrogation d'une base de données
    - **Transformation:** la modification de la forme des données pour les rendre cohérents et homogènes
    - **Chargement:** consiste à charger les données formatées dans le DW

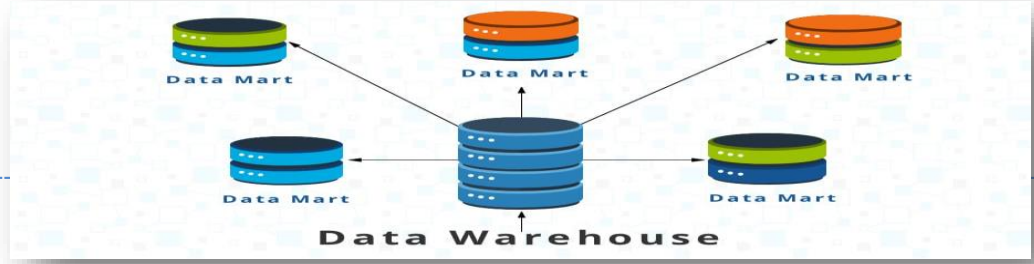


Cycle de vie d'un processus ECD: Data Warehouse (4/6)

- **Administration:** **Elle est constituée de plusieurs tâches pour assurer :**
  - la qualité et la pérennité des données aux différents applicatifs
  - la maintenance
  - la gestion de configuration
  - les mises à jour
  - l'organisation, l'optimisation du SI
  - la mise en sécurité du SI

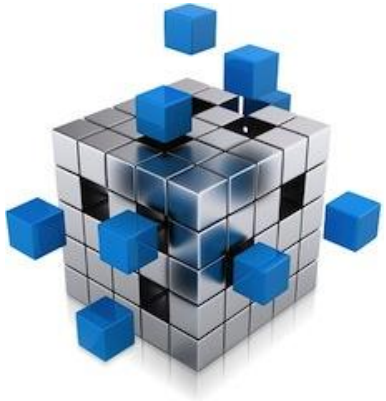
Cycle de vie d'un processus ECD: Data Warehouse (5/6)

- **Utilisation du DW:**
  - **Intégration directe :**
    - Les données de l'entrepôt sont utilisées directement pour l'aide à la décision
  - **Les Data Marts (Magasins de données):**
    - Lieu de stockage des données métier (Ex: Data mart Comptabilité, Data mart RH,...)
    - Duplication de données de l'entrepôt dans des environnements plus restreints
    - Ces mini DW peuvent alors être considérés comme des espaces d'analyse, du fait que les données sont bien moins nombreuses et surtout qu'elles sont thématiques.



Cycle de vie d'un processus ECD: Data Warehouse (6/6)

- **Utilisation du DW …**
  - **OLAP (On-Line Analytical Processing):**
    - Outil d'interrogation et d'analyse de données
    - Utilise une approche multidimensionnelle pour la représentation des données
      - Les données sont représentées sous forme de **cube**
      - Il s'agit d'analyse des faits selon plusieurs dimensions



  - **Les outils d'analyse: SIAD (Système Interactifs d'Aide à la Décision)**
    - Navigation dans des données pour l'accès à l'information
    - Fonctions statistiques pour la description des données
    - Création de cubes multidimensionnels locaux pour alléger le travail des serveurs
  - **Les outils de présentation: EIS (Executive Information System)**
    - Visualisation des résultats de l'analyse sous forme de tableaux et de graphiques

Cycle de vie d'un processus ECD: Data mining

- C'est l'étape de fouille de données:
  - L'étape la plus importante dans un processus ECD
  - Trouver des modèles représentatifs des données observées
- C'est l'**exploration** et l'**analyse** de grandes quantités de données en vue d'en tirer l'information pertinente pour la compréhension du phénomène étudié , la formulation de jugements et la prise de décision. (S. Tuffery)
- Le Data Mining a pour objet **l'extraction d'un savoir ou d'une connaissance** à partir de grandes quantités de données par des **méthodes automatiques**

Cycle de vie d'un processus ECD: Evaluation et validation

- La validation du modèle trouvé est nécessaire
  - Généralement effectuée par un **expert du domaine**
- Une fois évalué, le modèle trouvé devient une **connaissance**
- Application du modèle
- Et ça tourne!

Exemple de problème (1/4)

Exemple issu du livre de P. Adriaans et D. Zantige [Adriaans & Zantige 96]

- Un **éditeur** vend **5 sortes de magazines** : sport, voiture, maison, musique et Bande Dessinée. Il souhaite:
  - mieux étudier ses clients
  - découvrir de nouveaux marchés ou vendre plus de magazines à ses clients habituels
- Quelques questions qu'il peut se poser :
  - **Q1 :** Combien de personnes ont pris un abonnement à un magazine de sport cette année ?
  - **Q2 :** A-t-on vendu plus d'abonnements de magazines de sport cette année que l'année dernière ?
  - **Q3 :** Est-ce que les acheteurs de magazines de BD sont aussi amateurs de sport ?
  - **Q4 :** Quelles sont les caractéristiques principales de mes lecteurs de magazines de voiture ?
  - **Q5 :** Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?

Questions de natures différentes mettant en jeu des processus différents

Exemple de problème (2/4)

- Q1 : Combien de personnes ont pris un abonnement à un magazine de sport cette année ?
  - Requête SQL à partir des données opérationnelles suffit si les tables concernées ont été suffisamment indexées
- Q2 : A-t-on vendu plus d'abonnements 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** de type OLAP

Q1 et Q2

- **Réponse par simples requêtes SQL** : les données recherchées sont que le résultat d'un calcul simple sur un ou des groupes d'enregistrements
- ce qui distingue Q1 et Q2, c'est la notion de **temps** et la **comparaison**

Exemple de problème (3/4)

- Q3 : Est-ce que les acheteurs de magazines de BD sont aussi amateurs de sport ?
  - 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
  - cette question peut être généralisée, on pourrait ainsi :
    - **chercher des associations fréquentes** entre acheteurs de magazine pour effectuer des actions promotionnelles
    - **introduire une composante temporelle** pour chercher si le fait d'être lecteur d'un magazine implique d'être, plus tard, lecteur d'un autre magazine

Exemple de problème (4/4)

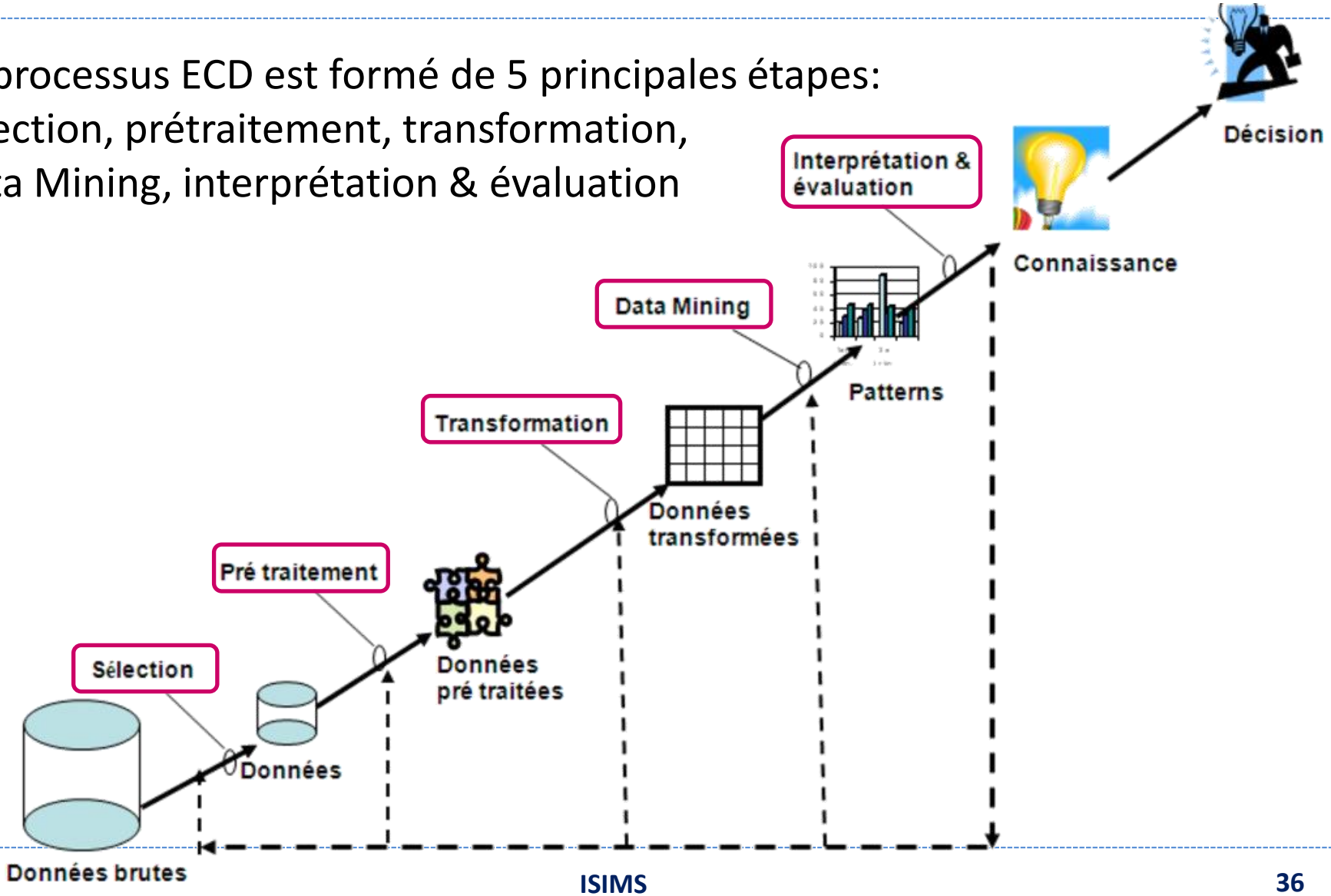
- Q4 : Quelles sont les caractéristiques principales de mes lecteurs de magazines de voiture ?
  - question plus ouverte : il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser

c'est pour ce type de question que sont mis en œuvre des **outils de fouille de données**

- Q5 : Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?
  - question ouverte : Il faut disposer d'**indicateurs** comme : durées d'abonnement, délais de paiement, ...
  - question (classique dans le bancaire) avec une **forte composante temporelle** et **nécessite des données historiques**

Étapes du processus ECD

Le processus ECD est formé de 5 principales étapes: sélection, prétraitement, transformation, Data Mining, interprétation & évaluation



1. Étape de Sélection des données

- L'étape de sélection des données consiste à:
  - Obtenir des données en accord avec les objectifs qu'on s'est fixés
  - Ces données proviennent de bases de production ou d'entrepôts
    - Par l'utilisation d'outils de requêtage (SQL, OLAP, ...)
    - Copie sur une machine adéquate (pour pouvoir les modifier et pour des questions de performance)
  - Structuration des données en champs typés

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

2. Étape de Prétraitement (1/5)

- L'étape de prétraitement des données consiste à:
  - **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
  - **Enrichissement des données**

2. Étape de Prétraitement (2/5)

- **Corrections des doublons, des erreurs de saisie:**
  - Un doublon donne plus d'importance à la donnée répétée
  - Une erreur de saisi peut à l'inverse occulter une répétition

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

2. Étape de Prétraitement (3/5)

- **Intégrité de domaine:**
  - Un contrôle sur les domaines de valeurs peut révéler des valeurs aberrantes

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

2. Étape de Prétraitement (4/5)

- **Information manquante:**
  - Cas où les champs ne contiennent aucune donnée
  - Parfois intéressant de conserver ces enregistrements car l'absence d'information peut être informative (Ex. fraude): le manque d'information est une information
  - Les valeurs des autres champs peuvent être utiles

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

2. Étape de Prétraitement (5/5)

- **Enrichissement :**
  - Recours à d'autres bases de données souvent pour ajouter de nouveaux champs en conservant le même nombre d'enregistrements
  - Plusieurs difficultés:
    - Relier les données, parfois hétérogènes, entre elles
    - Introduction de nouvelles valeurs manquantes et/ou aberrantes

Client	Date naissance	Revenus	Propriétaire	Voiture
Bémol	13/1/50	20 000	Oui	Oui
Bodinoz	21/5/70	12 000	Non	Oui
Airinair	15/06/63	9 000	Non	Non
Manvussa	27/03/47	15 000	Non	Oui

3. Étape de transformation: codage et normalisation (1/3)

- Étape dépendante du choix de l'algorithme de Data Mining utilisé:
  - **Regroupements**
    - Cas où les attributs prennent un très grand nombre de valeurs discrètes (Ex. adresses que l'on peut regrouper en 2 régions (Paris - Province))
  - **Attributs discrets**
    - Les attributs discrets prennent leurs valeurs (souvent textuelles) dans un ensemble fini donné (Ex. colonne *magazine* de l'exemple: 5 valeurs)
    - Deux représentations possibles : représentation verticale ou représentation horizontale (plus adaptée à la fouille de données)
    - Changements de types pour permettre certaines manipulations comme par exemple des calculs de distance, de moyenne (Ex. date de naissance)
  - **Uniformisation d'échelle**
    - Les données sont réparties sur des échelles différentes (dizaines, centaines, …)
    - De telles valeurs perturbent les calculs et les comparaisons entre données
      - Il faut uniformiser les échelles des données

3. Étape de transformation: codage et normalisation (2/3)

- **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

3. Étape de transformation: codage et normalisation (3/3)

- Étape de transformation sur l'exemple

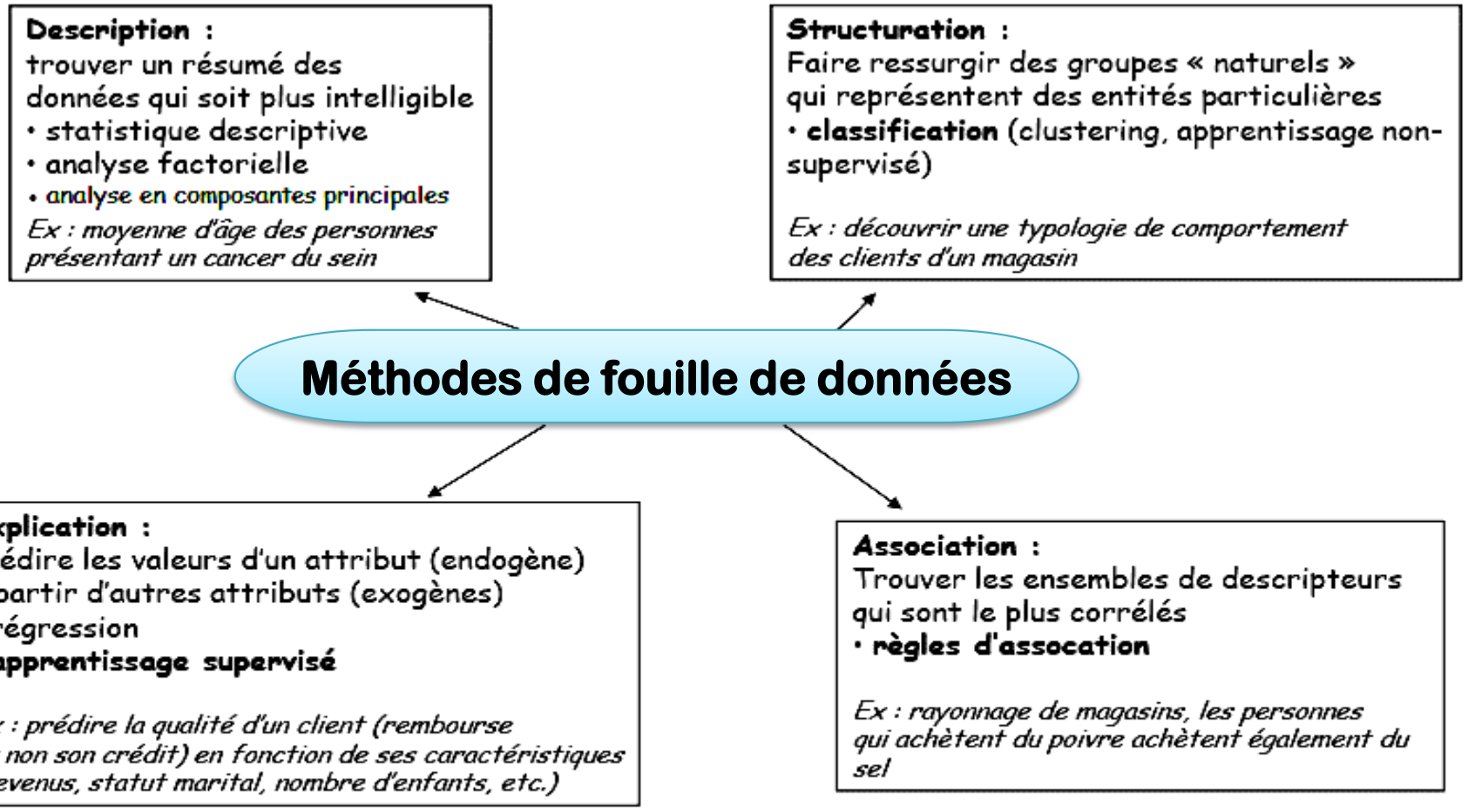
Client	Sport	BD	Voiture	Maison	Musique	DN	Rev	Prop	Voit	PP	DA
23134	0	1	1	1	1	50	20	oui	oui	1	4
31435	0	1	0	0	0	30	12	non	oui	0	null
43342	1	0	0	1	0	37	9	non	non	1	5
43241	1	0	0	1	0	53	15	non	oui	null	4

- **Avec :**
  - DN : date de naissance
  - Rev : revenus
  - Prop : Propriétaire
  - Voit : possède une voiture
  - PP : Paris ou province
  - DA : date d'abonnement

4. Étape de Fouille de données (1/4)

- **L'étape cœur du processus ECD**
  - Souvent difficile et coûteuse à mettre en œuvre
  - Les résultats doivent être interprétés et validés
- Le modèle obtenu est bien évalué s'il vérifie ces critères:
  - Rapidité d'accès
  - Rapidité d'utilisation
  - Compréhension
  - Bonne performances
  - Fiabilité
  - Évolution facile
  - Longue durée (sans dégradation au cours du temps)

4. Étape de Fouille de données (2/4)



les méthodes sont le plus souvent complémentaires !

4. Étape de Fouille de données (3/4)

- **Logiciels de fouille de données: Logiciels commerciaux :**
  - Suites logicielles SAS (SAS Enterprise miner): (http://www.sas.com/offices/europe/france/)
  - SPSS d'IBM: (http://www-01.ibm.com/software/fr/analytics/spss/)
  - Solution Analytics de SAP (http://www.sap.com/pc/analytics/strategy.html)
  - STATISTICA Data Miner (http://www.statsoft.fr/logiciels/data-miner-scoring-segmentation-analyses-predictives.php)
  - RapidMiner: outil Open source à la fois gratuit et commercial (https://rapidminer.com/)
  - Etc.

4. Étape de Fouille de données (4/4)

- **Logiciels de fouille de données: Logiciels universitaires:**
  - Python: https://www.python.org/
    - langage de programmation très puissant utilisé en Data Mining pour faire de l'analyse statistique, la classification, le clustering et l'analyse prédictive.
  - R : https://cran.r-project.org/
    - langage de programmation destiné aux statistiques et à la science des données. Il permet de faire l'analyse statistique, la classification, le clustering et l'analyse prédictive.
  - Weka : http://www.cs.waikato.ac.nz/ml/weka/
    - suite populaire de logiciels d'apprentissage automatique, en Java, développée à l'université de Waikato, Nouvelle-Zélande. Il permet de faire l'analyse statistique, la classification, le clustering et l'analyse prédictive.
  - TANAGRA : http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html
    - logiciel gratuit de Data Mining destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique.
  - Etc.

5. Étape de validation et d'intégration de la connaissance

- **Deux modes de validation:**
  - **Par expertise**
    - Un expert juge de la pertinence du modèle obtenu
  - **Statistique**
- **Intégration de la connaissance :**
  - Il s'agit simplement d'utiliser le modèle validé (comme connaissance) pour prendre la décision dans l'entreprise