



DATA MINING: Concept, Applications et Techniques

Pr. Anass EL HADDADI
SDIC / DMI, ENSA Al-Hoceima, Maroc

Objectifs

- ▶ Atteindre une connaissance générale des aspects méthodologique, technologiques de l'analyse de données: intérêts, difficultés, solutions actuelles et l'avenir.



Références

- ▶ Data Mining et statistique décisionnelle, Stéphane TUFFÉRY
- ▶ Étude de cas en statistique décisionnelle, Stéphane TUFFÉRY
- ▶ Probabilités, analyse des données et statistique, Gilbert SAPORTA
- ▶ Data Mining, un tour d'horizon, E-G TALBI
- ▶ <http://www.kdnuggets.com/>, Consulté le 17 Avril 2018 à 17:13 (GTM+1)





Qu'est ce que le Data Mining ?



Qu'est ce que le Data Mining ?

- ▶ Data Mining est un sujet qui dépasse le cercle restreint des scientifiques et suscite un vif intérêt dans le monde des affaires

- «l'extraction d'information originale, auparavant inconnues et potentiellement utiles, à partir de données » (Piateski-Shapiro)



Qu'est ce que le Data Mining ?

- ▶ Data Mining est un sujet qui dépasse le cercle restreint des scientifiques et suscite un vif intérêt dans le monde des affaires

- «la découverte de nouvelles corrélation (ou coefficient de coïncidence), tendances et modèles par tamisage d'un large volume de données» (John Page).



Qu'est ce que le Data Mining ?

- ▶ Data Mining est un sujet qui dépasse le cercle restreint des scientifiques et suscite un vif intérêt dans le monde des affaires

- «un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données» (Kamran Parsaye).



Qu'est ce que le Data Mining ?

- ▶ Data Mining est un sujet qui dépasse le cercle restreint des scientifiques et suscite un vif intérêt dans le monde des affaires

- «l'exploration et l'analyse, par des moyens automatiques ou semi-automatiques, d'un large volume de données afin de découvrir des tendances ou des règles» (M. Berry).



Qu'est ce que le Data Mining ?

- ▶ Data Mining est un sujet qui dépasse le cercle restreint des scientifiques et suscite un vif intérêt dans le monde des affaires

- «un processus non élémentaire de mise à jour de relation, corrélation, dépendances, association, modèles, structure, tendance, classes, facteurs obtenus en naviguant à travers de grands ensembles de données» (M. Jambu).



Qu'est ce que le Data Mining ?

- ▶ Data Mining est un sujet qui dépasse le cercle restreint des scientifiques et suscite un vif intérêt dans le monde des affaires

- Avec poésie: « ...torturer l'information disponible jusqu'à ce qu'elle avoue ...» (Dimitris Chorafas).



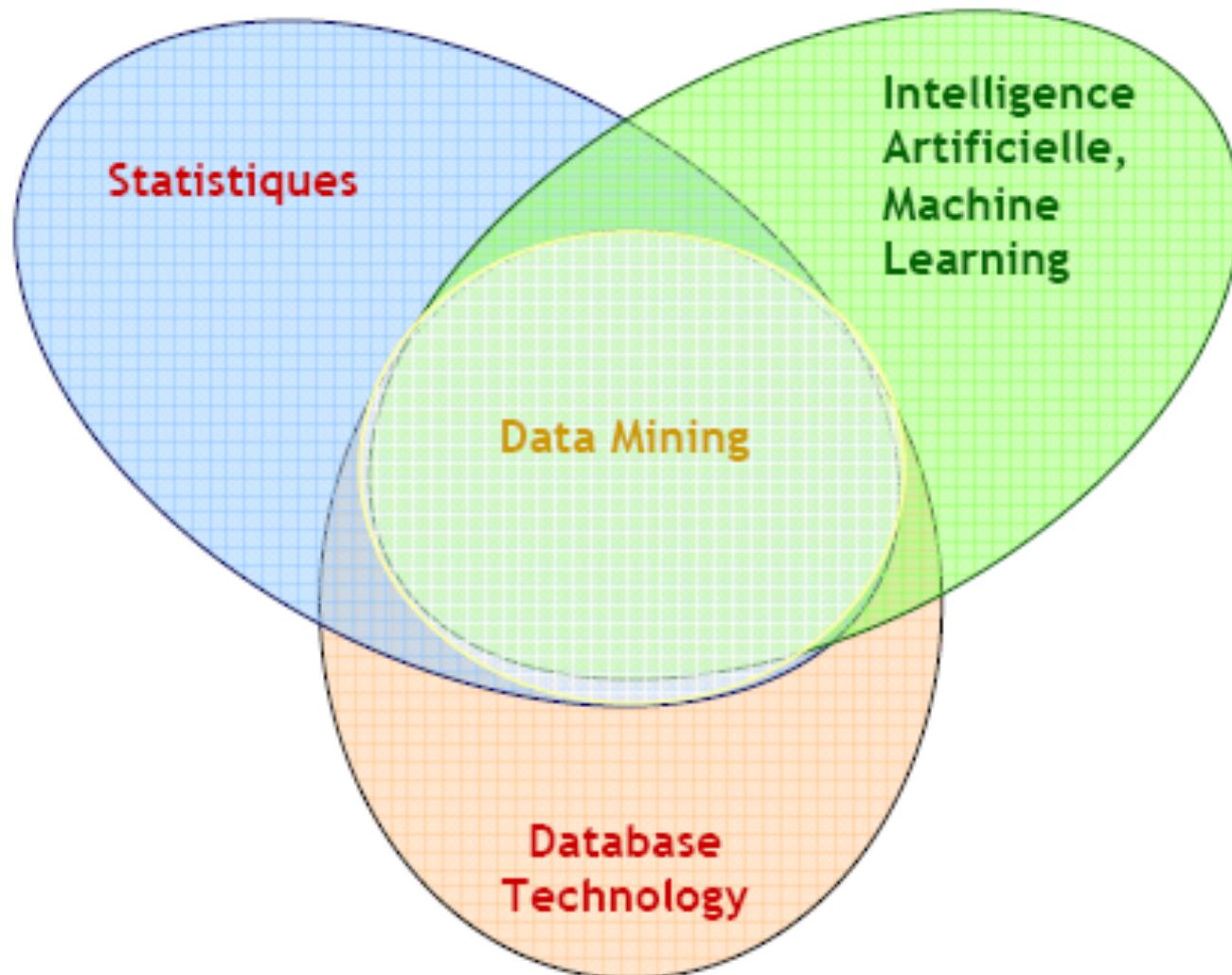
Qu'est ce que le Data Mining ?

- ▶ Data Mining est un sujet qui dépasse le cercle restreint des scientifiques et suscite un vif intérêt dans le monde des affaires

- Avec cynisme et réalisme «... passer les données dans la machine à saucisses pour obtenir des Merguez ... douces ou épicées ...» (Moktar Outtas)



Un champs multidisciplinaire



Quelle Problématique du Data Mining ?

Comment gérer la grande quantité des données
“brutes” provenant de plusieurs sources pour les
rendre accessibles et lisibles par le décideur ?



Définition du Data Mining

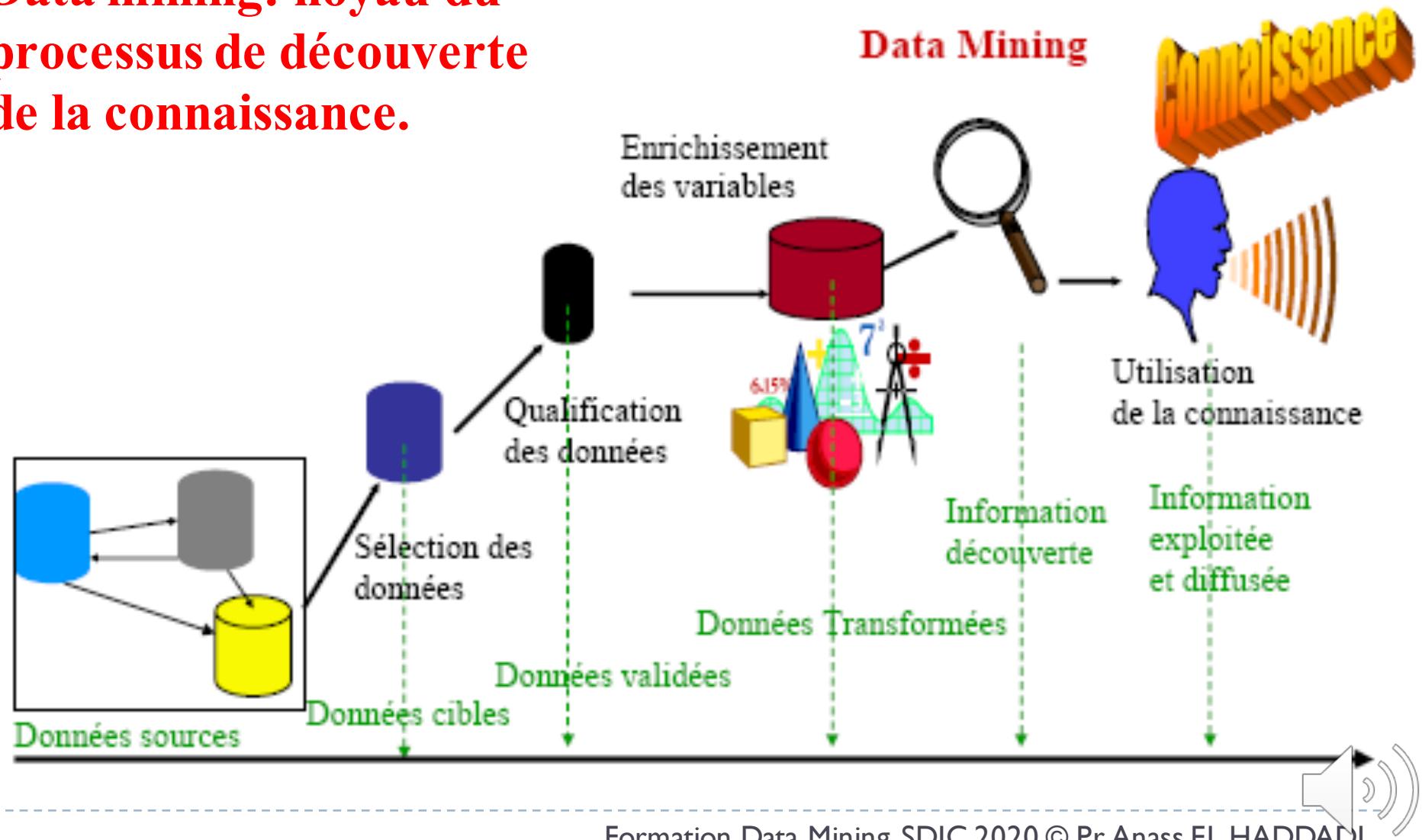


- Extraction d'information d'intérêt (non triviale, implicite, inconnue à priori et potentiellement utile) à partir de données stockées dans de large entrepôts de données, en utilisant des procédures automatiques ou semi-automatiques pour une prise de décision.
- Appelé aussi KDD (Knowledge Discovery in Databases)



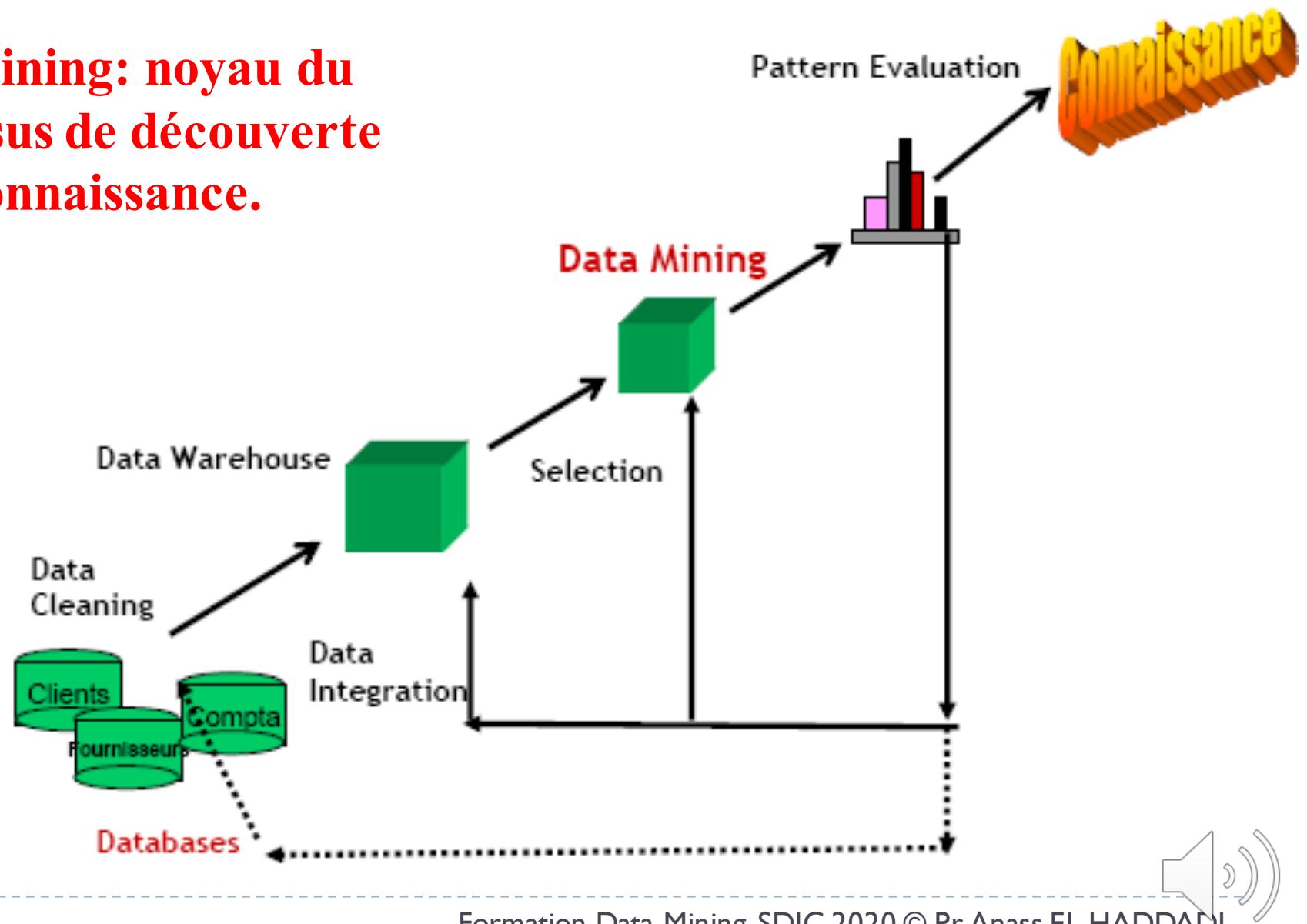
De la donnée vers la connaissance

Data mining: noyau du processus de découverte de la connaissance.



De la donnée vers la connaissance

Data mining: noyau du processus de découverte de la connaissance.

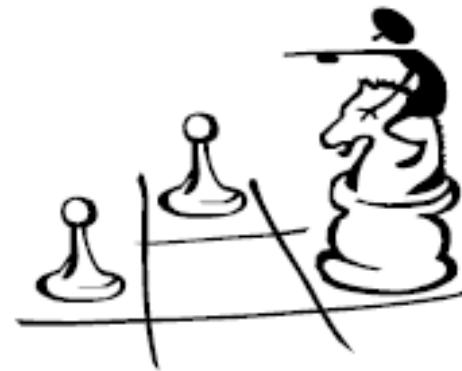


Du Warehouse au Data Mining

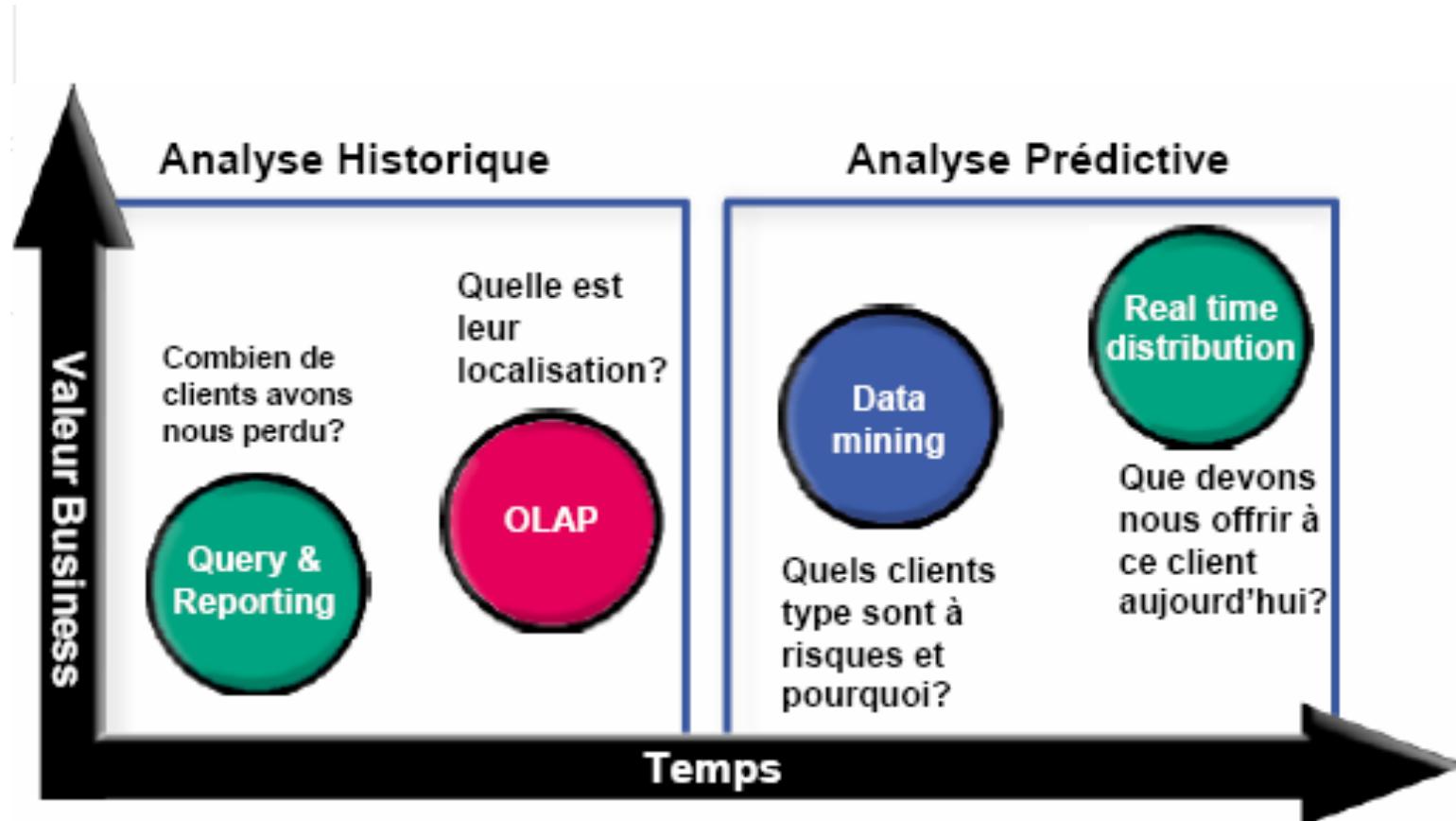


Data Warehousing provides
the Enterprise with a memory

Data Mining provides
the Enterprise with intelligence



Data Mining: Analyse Prédictive



En résumé

Qu'est ce que le Data Mining ?

- ▶ **Data Mining:** Ensemble de techniques d'exploration de données afin d'en tirer des **connaissances sous forme de modèles présentés** à l'utilisateur averti pour examen
- ▶ **Connaissance**
- Analyses (distribution du trafic en fonction de l'heure)
- Scores (fidélité d'un client), classe (mauvais payeurs)
- Règles (si factures > 10 000 alors départ à 70%).





Domaines d'applications ?

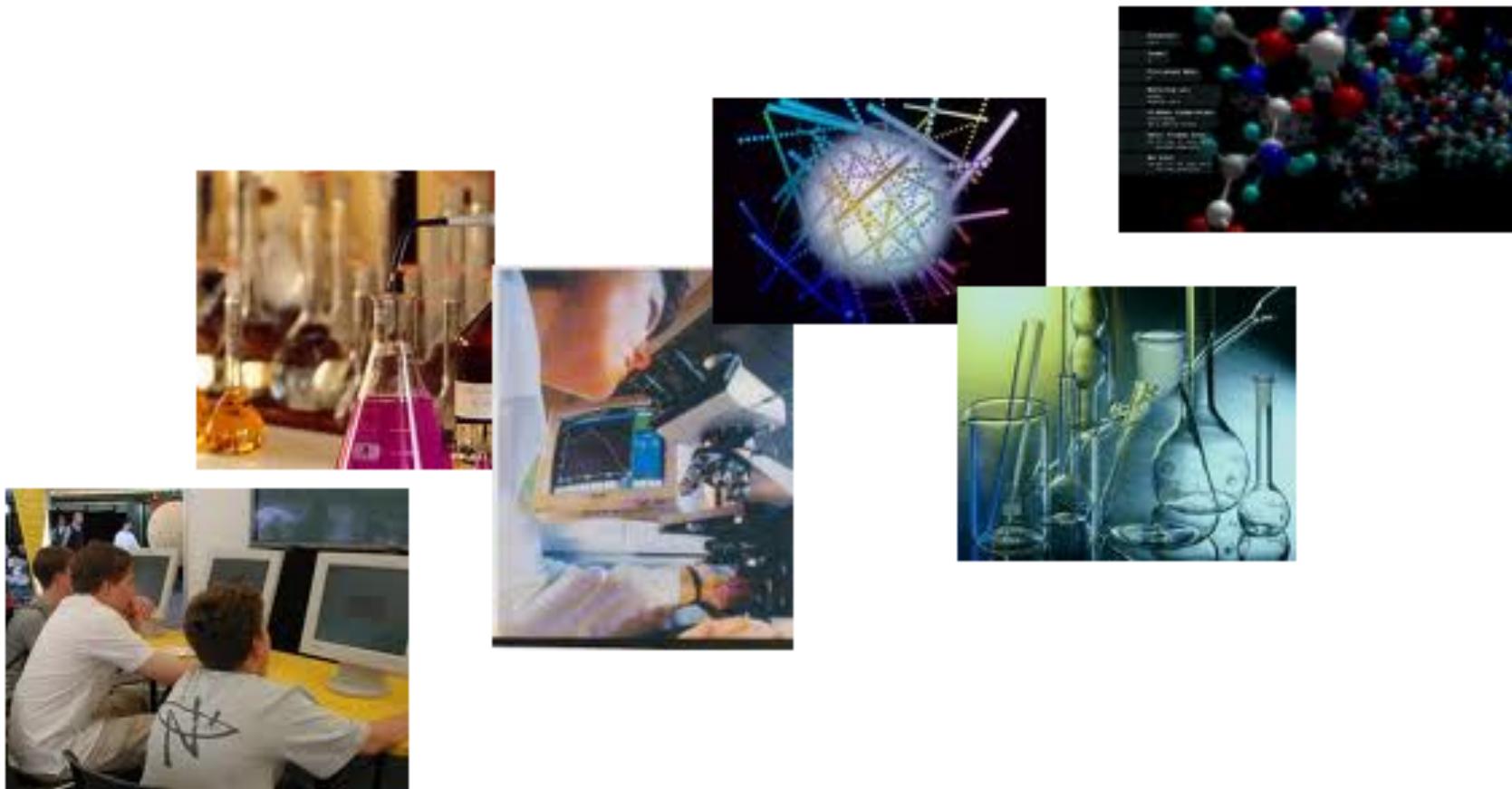


Applications clefs du Data Mining



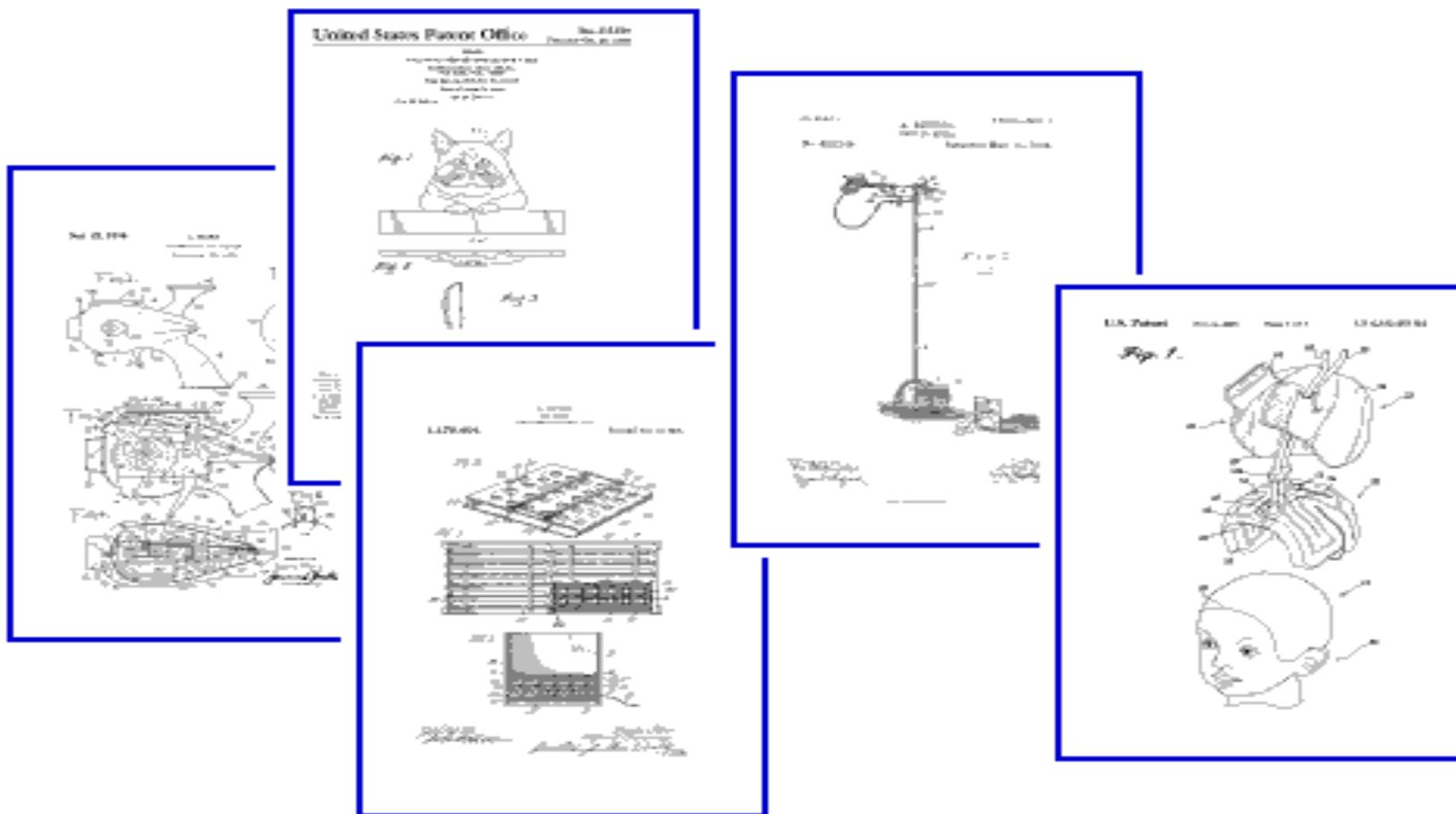
Applications clefs du Data Mining

▶ La Veille Scientifique et Technologique

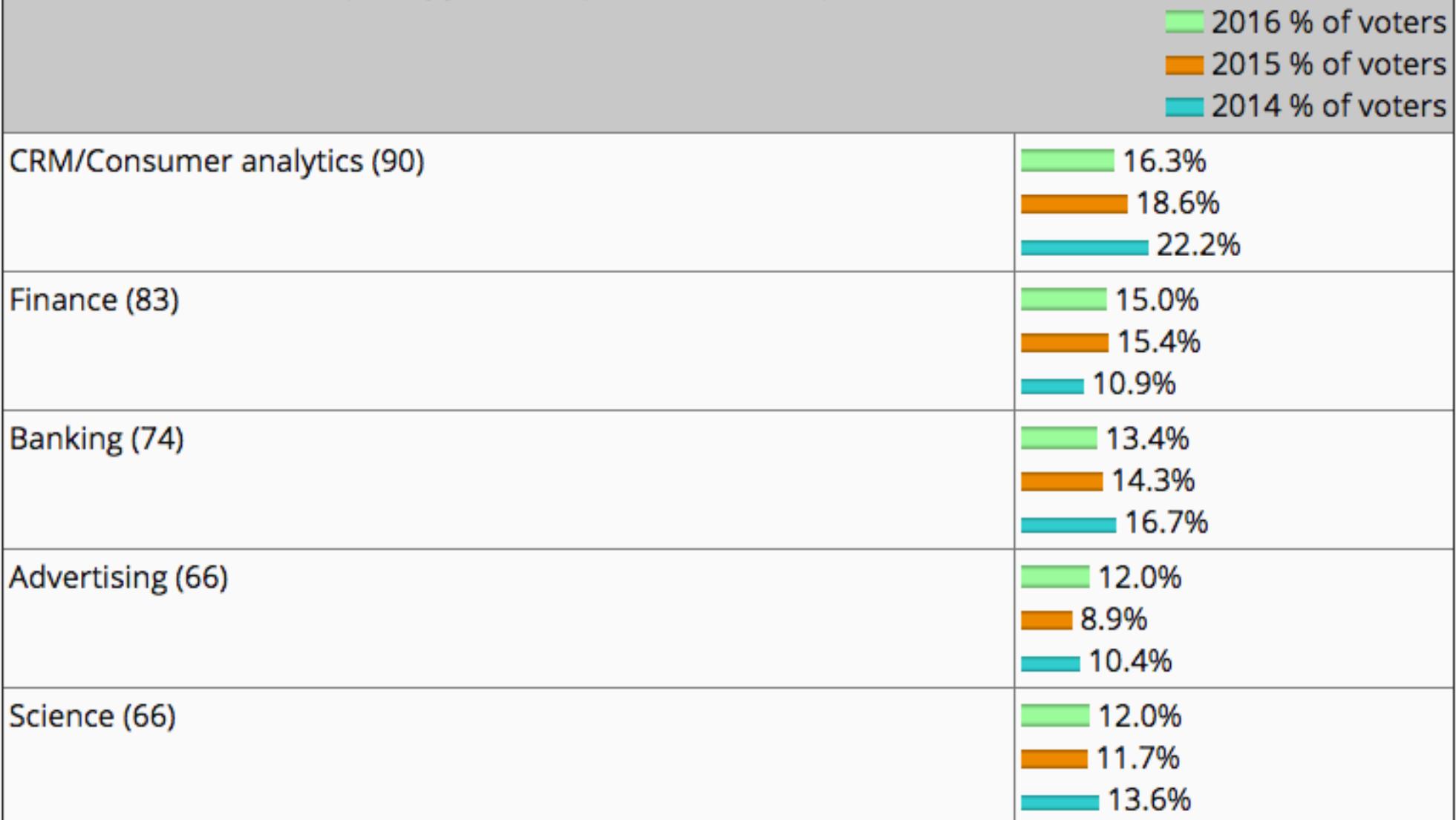


Applications clefs du Data Mining

▶ La brevets

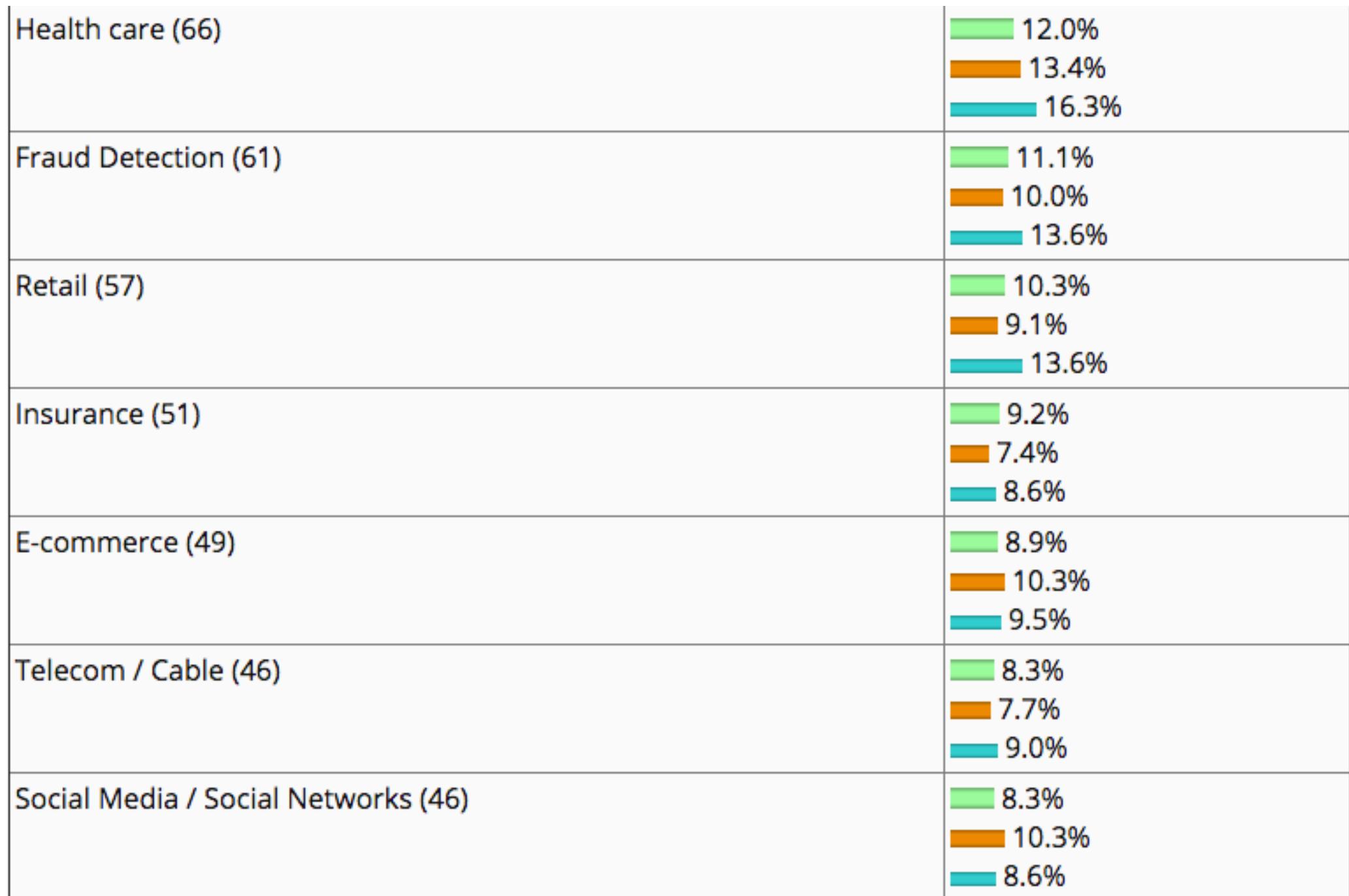


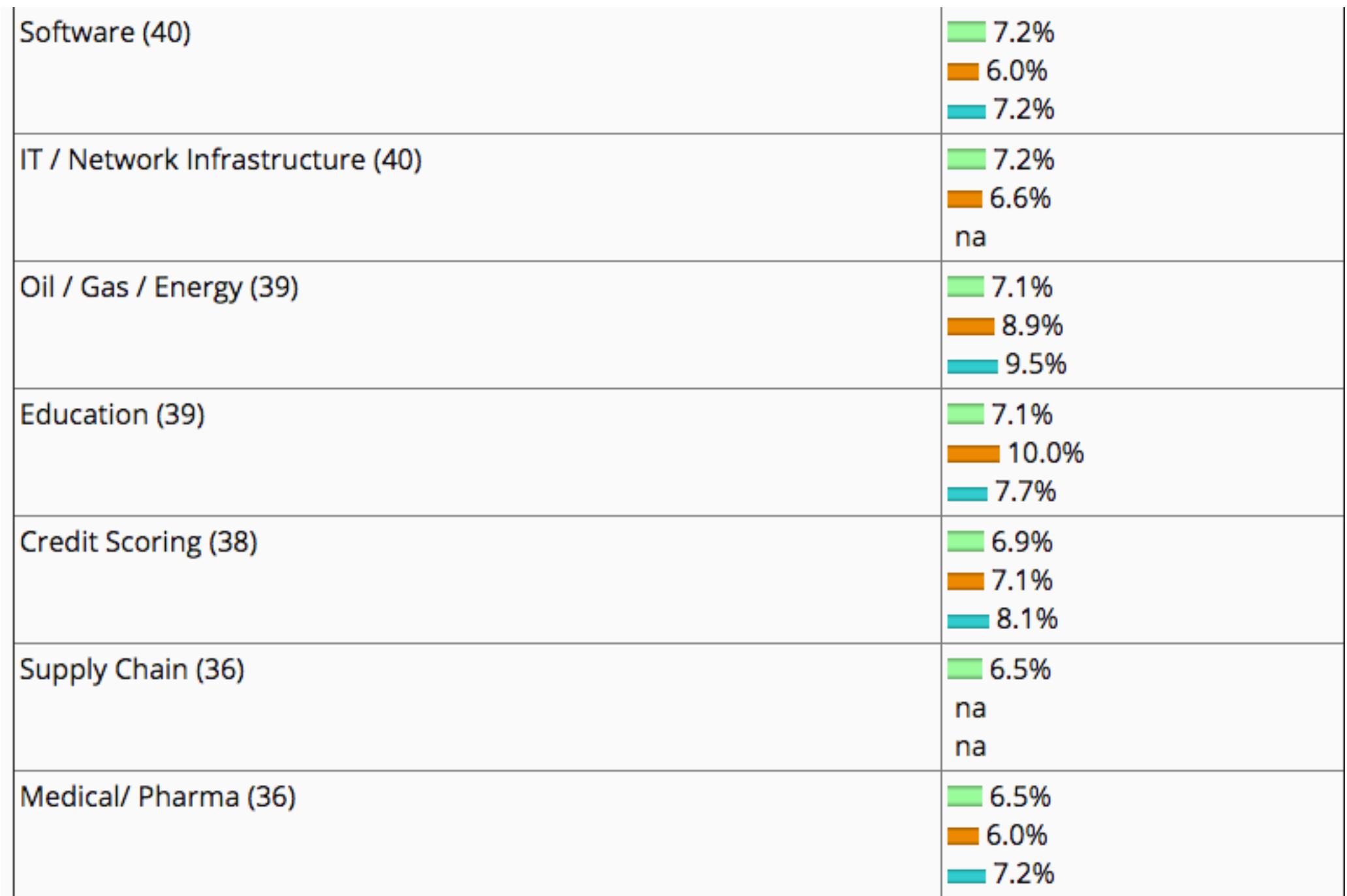
Industries/Fields where you applied Analytics, Data Mining, Data Science in 2016?



Source: <https://www.kdnuggets.com/2016/12/poll-analytics-data-mining-data-science-applied-2016.html>, consulté le 01/04/2018









Statistiques & Data Mining

Des statistiques Au Data Mining

Statistiques

- Quelques centaines d'individus
- Quelques variables recueillies avec protocole spécial (échantillonnage, plan d'expérience, etc.)
- Fortes hypothèses sur les lois statistiques suivies

Data Mining

- Quelques millions d'individus
- Quelques centaines de variables
- Nombreuses var non numériques
- Données recueillies avant l'étude et souvent à d'autres fins
- Population constamment évolutive
- Données imparfaites avec erreur de Codification
- Nécessité de calculs rapides
- On ne cherche pas l'optimum mathématique mais le modèle le + facile à appréhender par les utilisateurs non statisticiens



Différence entre le Data Mining et la Statistique traditionnelle

- Les techniques de Data Mining **remplacent-elles** les statistiques ?
- Les statistiques sont omniprésentes. On les utilise :
 - Pour faire une analyse préalable,
 - Pour estimer ou alimenter les valeurs manquantes,
 - Pendant le processus pour évaluer la qualité des estimations,
 - Après le processus pour mesurer les actions entreprises et faire un bilan.

Statistiques et Data Mining sont tout à fait complémentaires



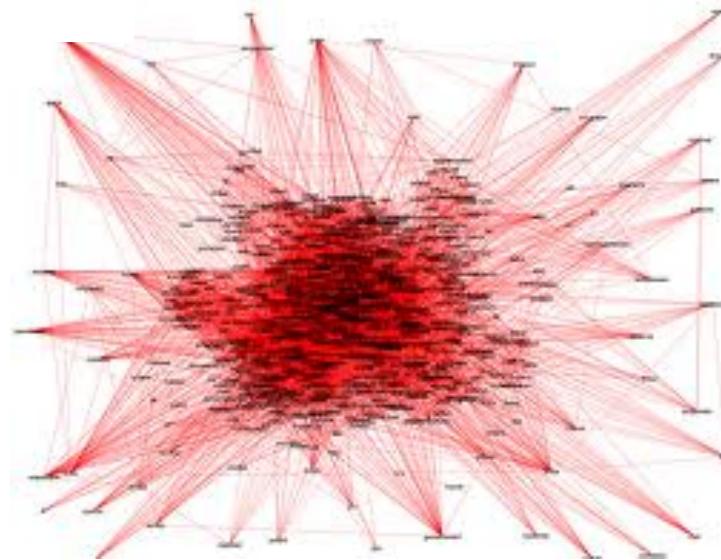
Le Data Mining aujourd’hui

- Ces techniques ne sont pas récentes
- Ce qui est nouveau
 - Capacité de stockage et de calcul // (matériel puissant)
 - Package de techniques de natures différentes qui peuvent s’enchaîner les unes aux autres
 - L’intégration du DM dans le processus de production

Elle permettent de traiter de **grands volumes de** données et font sortir le DM des Laboratoires de Recherche pour entrer dans les entreprises.



Le Data Mining aujourd’hui

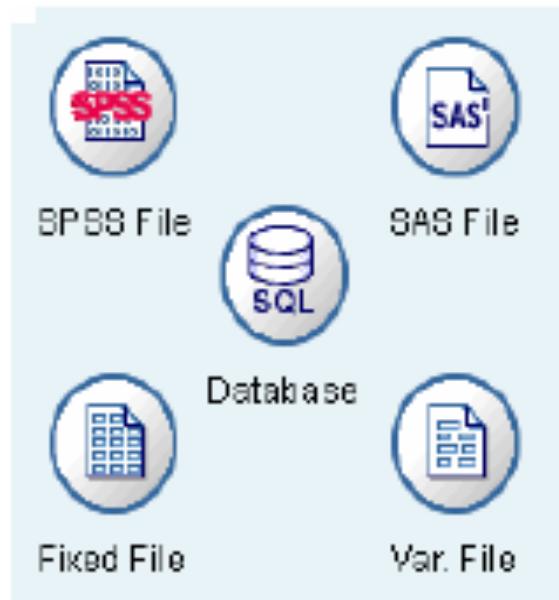




Autres facettes du Mining

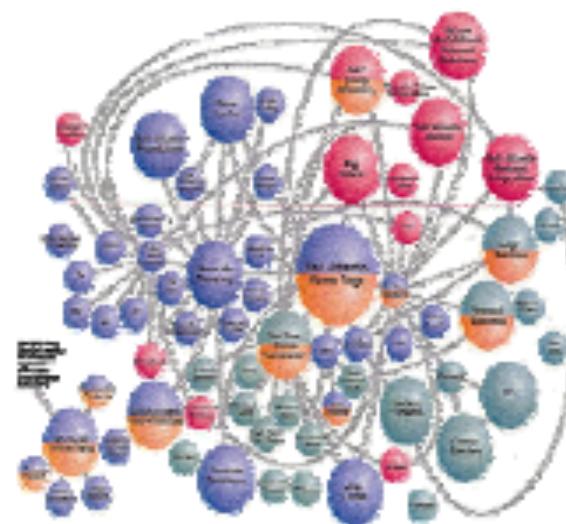
Autres facettes du Mining

Structure
Bases de
Données



Data Mining

Web Logs



Web Mining

Text non
structuré



Text Mining



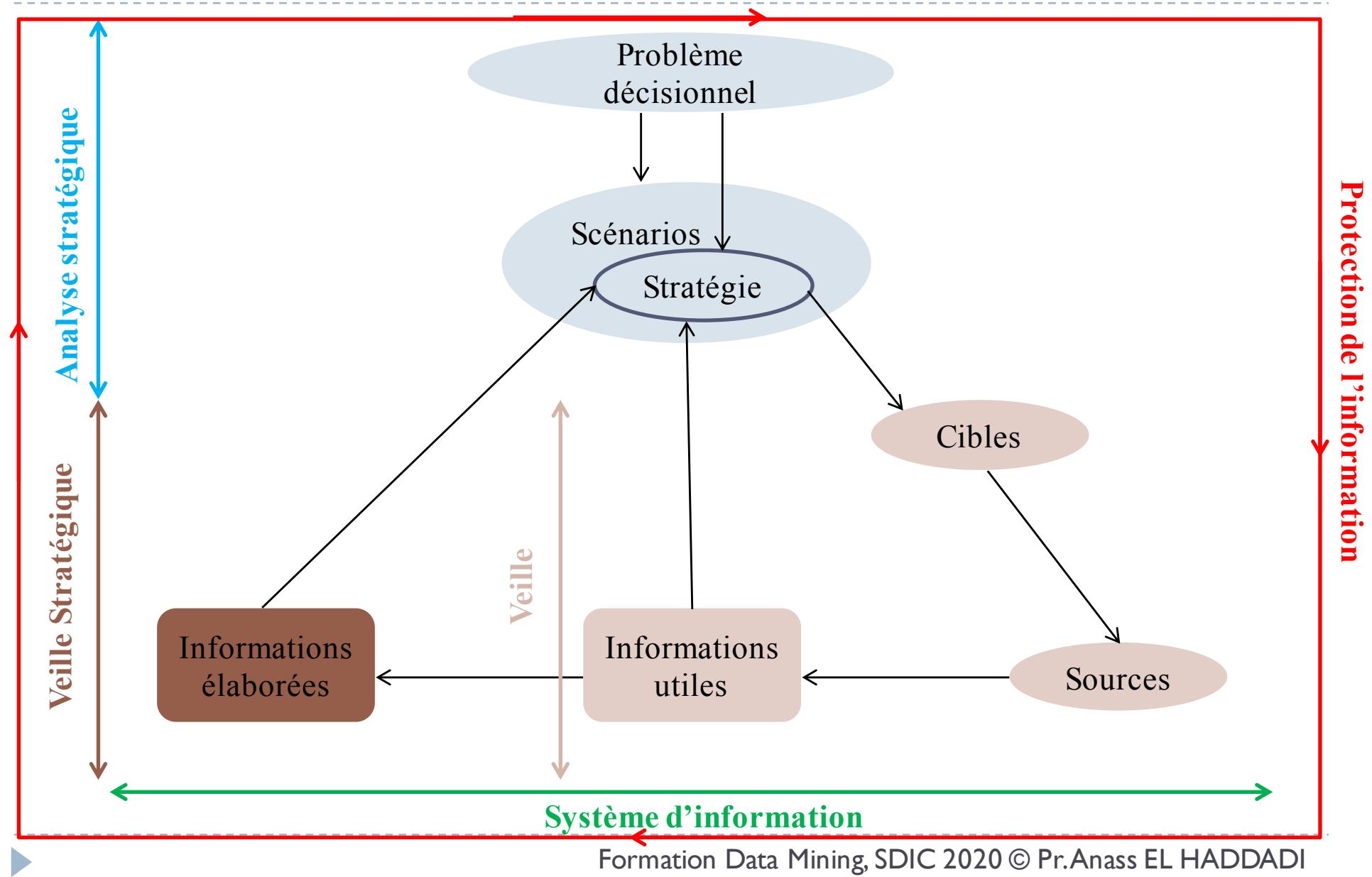


Le problème décisionnel

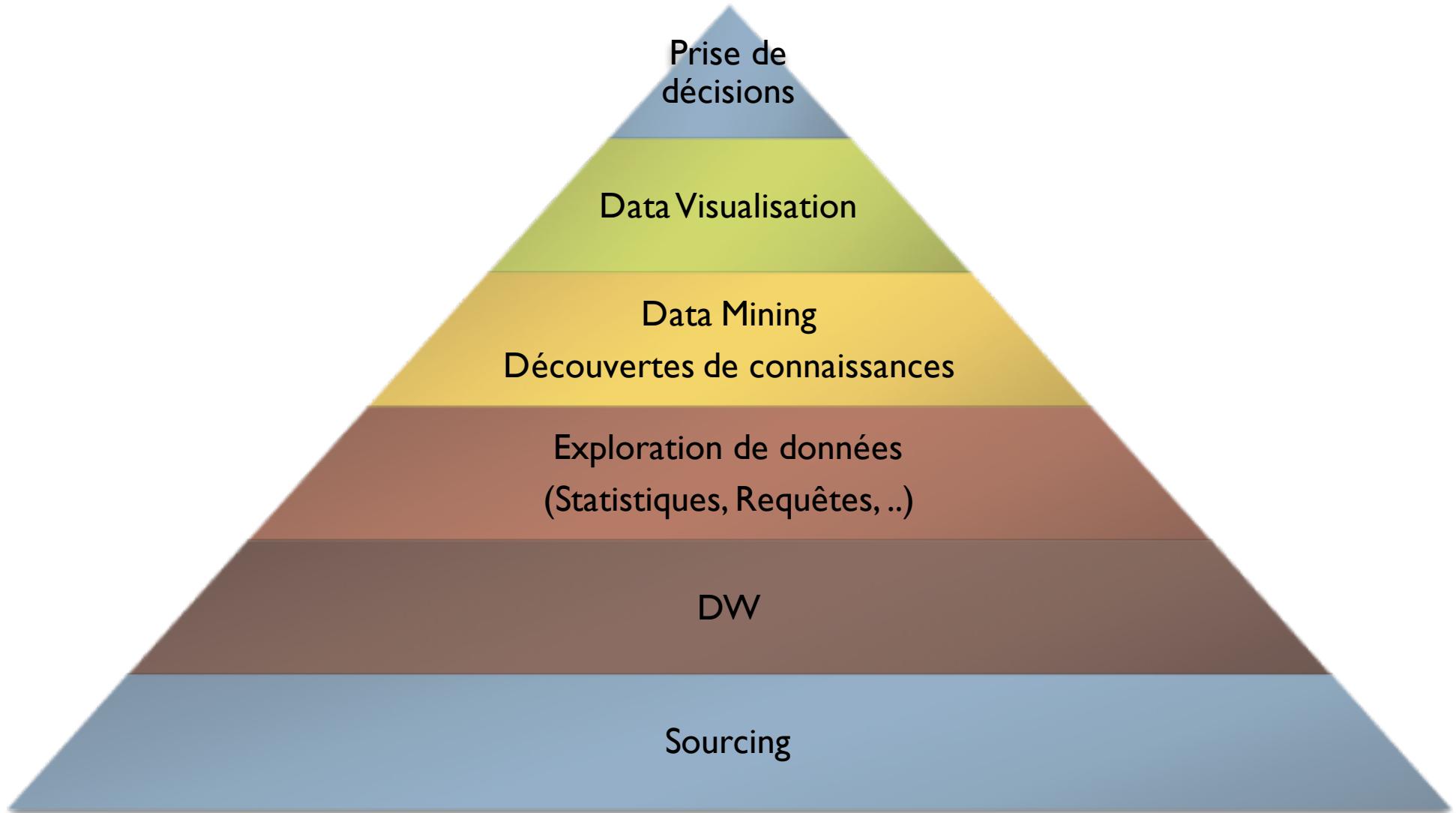
Problème Décisionnel – Scénarios



Le cycle de management stratégique de l'information (IE)



Data Mining et aide à la décision





Le cycle d'un projet de data mining

Les phases d'un projet de data mining

Définitions des objectifs.

Inventaire des données existantes.

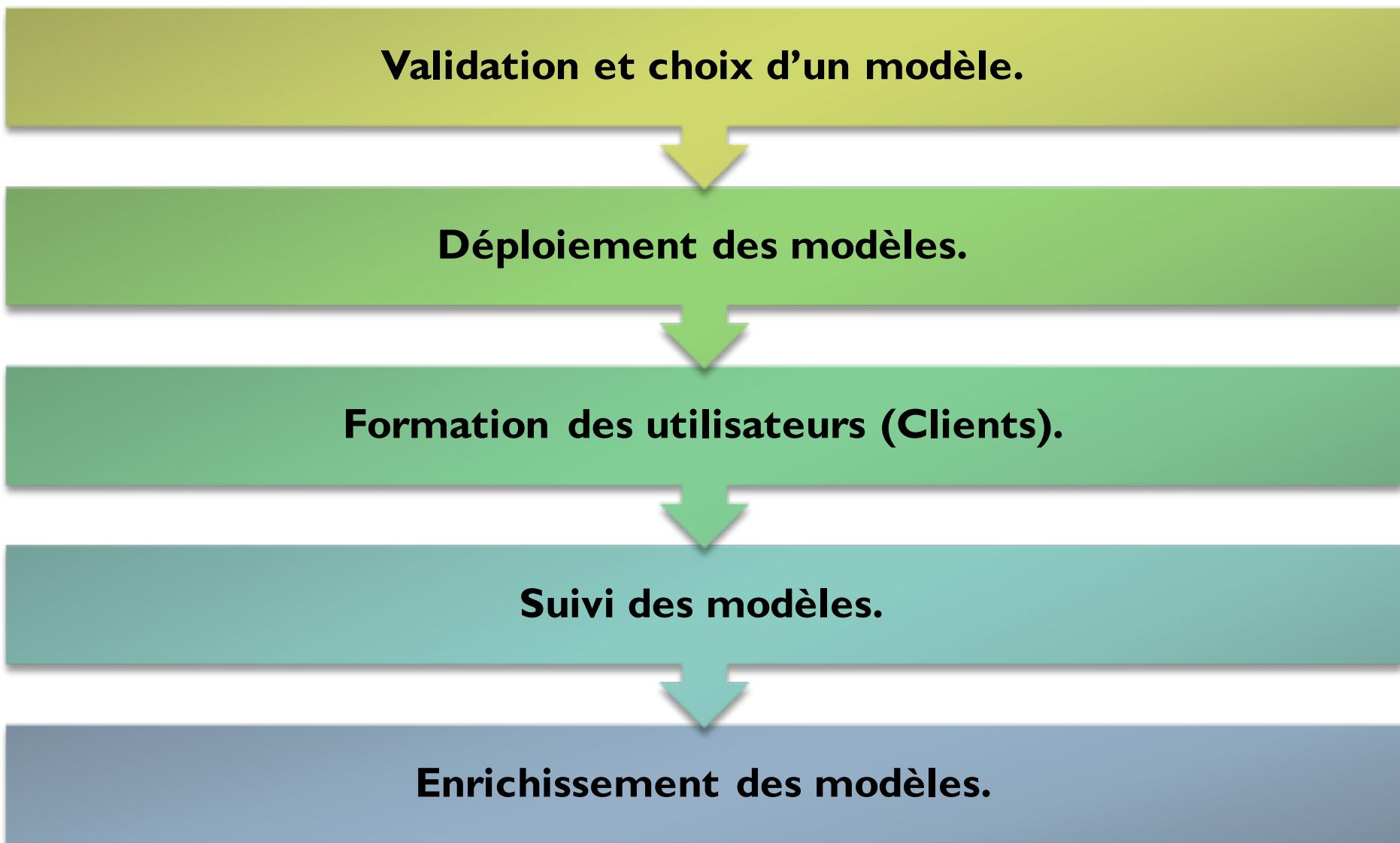
Collecte des données.

Exploration et préparation des données.

**Mise en œuvre des algorithmes (classification, scoring ...) -
Élaboration des modèles.**



Les phases d'un projet de data mining





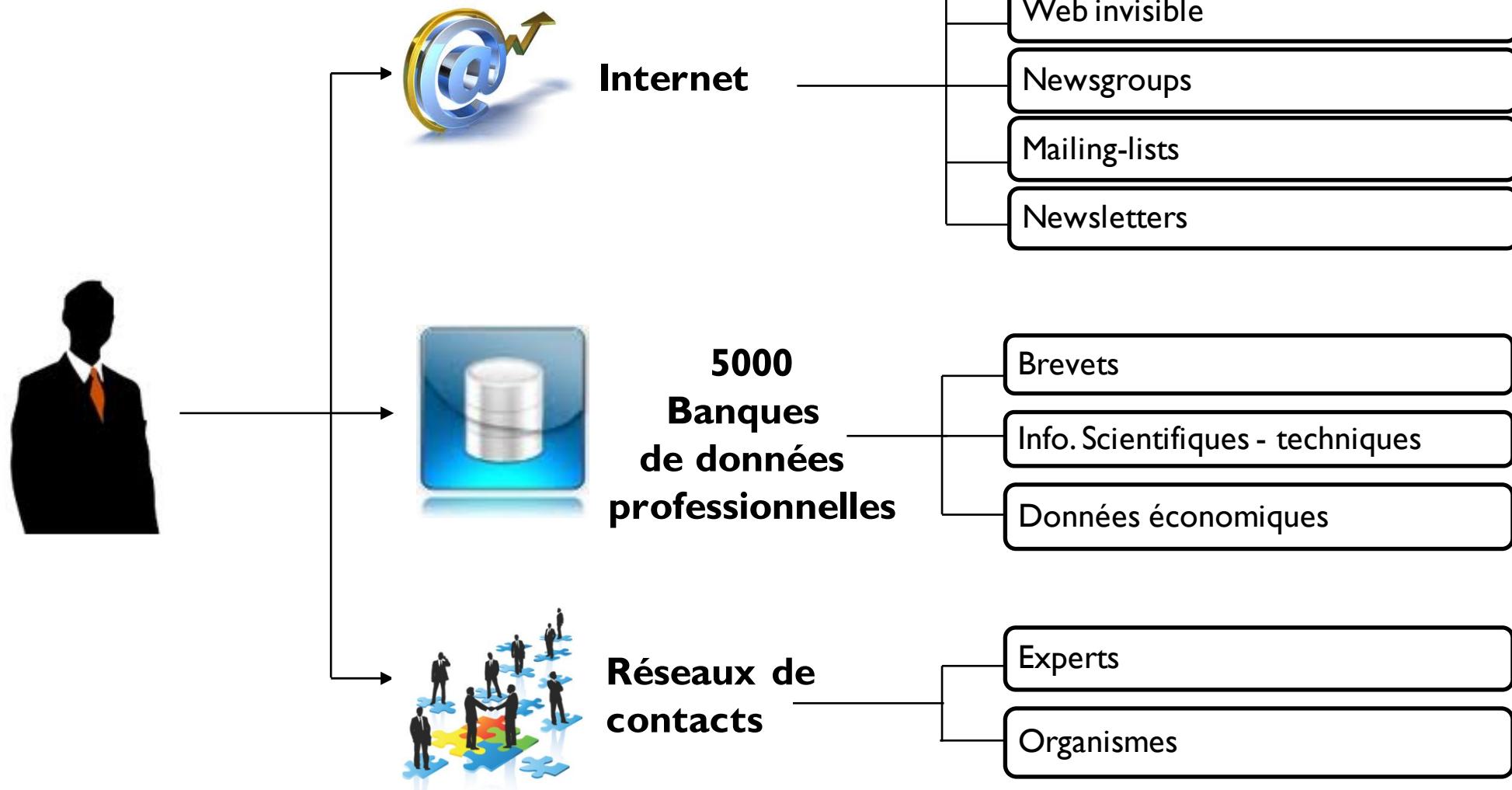
Sourcing

Inventaire des données utiles

- ▶ Recenser avec les spécialistes métier les données utiles:
 - ▶ Accessibilité
 - ▶ Fiabilité
 - ▶ À jour
 - ▶ Historisées, si besoin
 - ▶ Légalement utilisables.
- ▶ Interne / Externe
 - ▶ SI de l'entreprise
 - ▶ Hors SI de l'entreprise (Cloud)
 - ▶ Abonnement
 - ▶ Calculées à partir des données précédentes (indicateurs, ratios, évolutions au cours du temps).



Sourcing – Recherche et collecte d'information

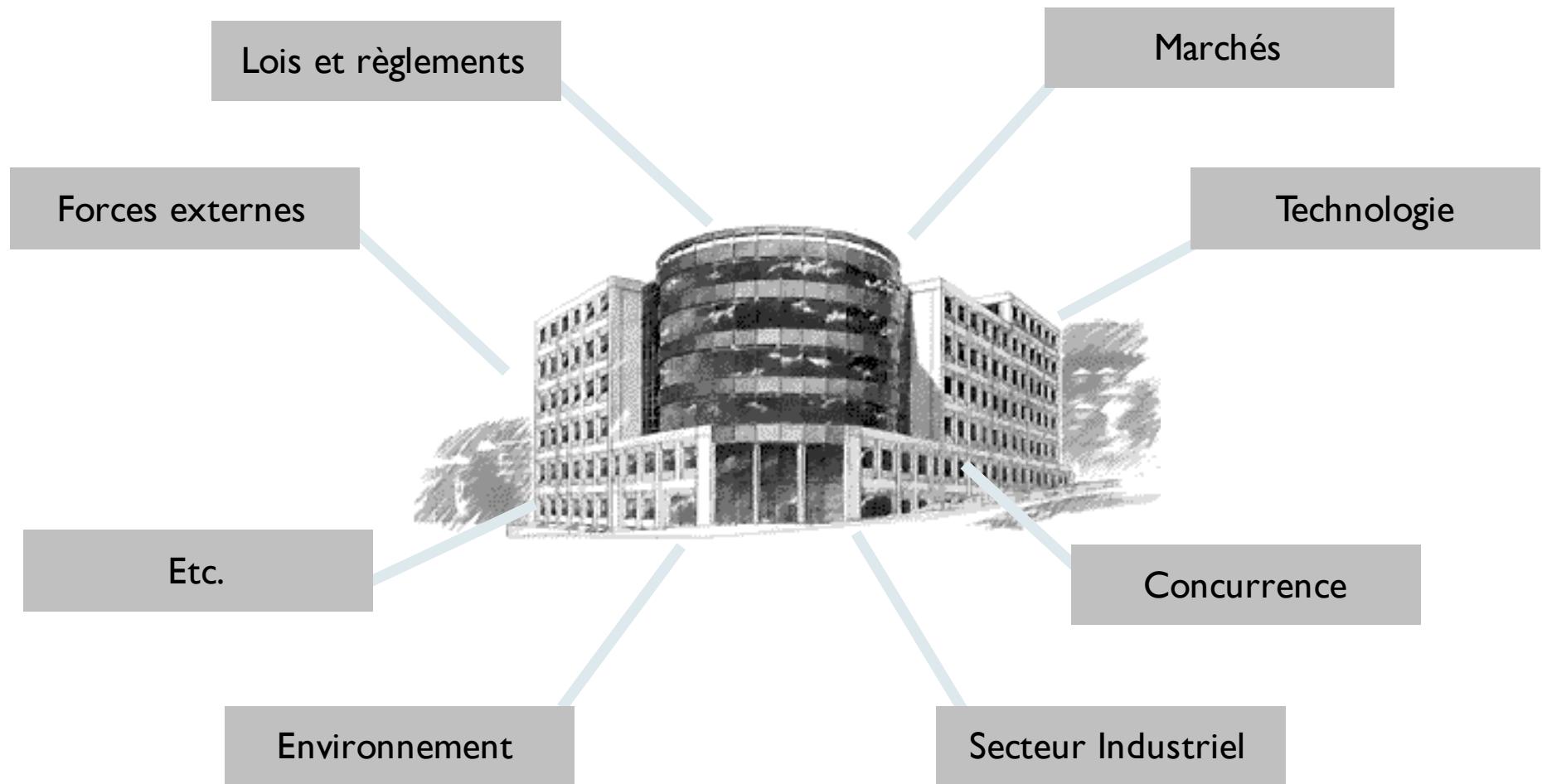


Quand on manque de données

- ▶ Enquêtes auprès d'échantillons de clients
- ▶ Utilisation des méga-bases de données (Big Data)
- ▶ Géomarketing (type d'habitat en fonction de l'adresse).
- ▶ Recours à des modèles standards préétablis par des sociétés spécialisées.
- ▶ Les médias sociaux (social media)
- ▶ Les données ouvertes (open data)



Quand on manque de données



Exemple: Géomarketing

- ▶ Données économiques
 - ▶ Nb entreprises, population active, chômage, commerces et services de proximité, habitudes de consommation...
- ▶ Données sociodémographiques
 - ▶ Population, richesse, âge et nombre d'enfants moyens, structures familiales, niveau socioprofessionnel...
- ▶ Données résidentielles
 - ▶ Ancienneté, type et confort des logements, proportion de locataires et propriétaires...
- ▶ Données concurrentielles
 - ▶ Implantation de l'entreprise, implantation de ses concurrents, parts de marché, taux de pénétration...
- ▶ Type d'habitat: beaux quartiers, classe moyenne, classe ouvrière, centre ville et quartiers commerçants...



Types de données

▶ Les sources formelles

- ▶ Composées principalement de la presse, la télévision, la radio, les livres, banques de donnée et CD-ROM, les brevets, les informations légales, les études publiques réalisée par des prestataires publics ou privés.

Avantage :

- **Une source d'information sur et assez exhaustive ;**
- **Elles ont un faible coût (sauf le cas de brevets et de certaines banques de données) ;**
- **Disponibilité de la source ;**
- **Facilité d'accès.**

Inconvénients :

- L'information est « mise en scène » pour qu'elle se vende ;
- Risque, parfois, de trouver une information obsolète ;
- On ne trouve pas toujours ce que l'on souhaite rechercher.



Types de données

▶ **Les sources informelles**

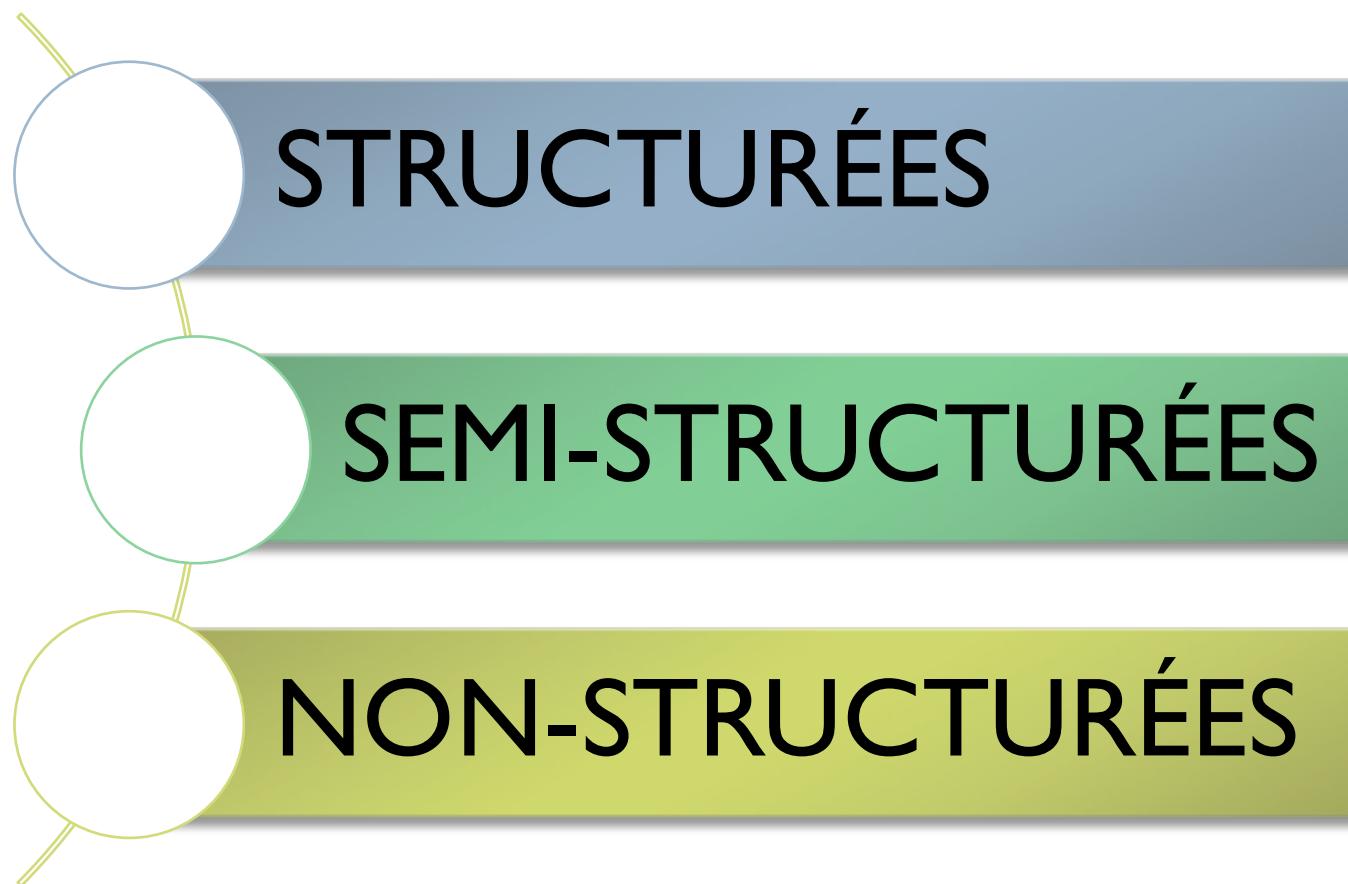
Dans ce type de source, l'individu doit se déplacer, passer du temps, sentir, toucher et percevoir l'information qu'elles procurent.

Elles sont variées. Citons :

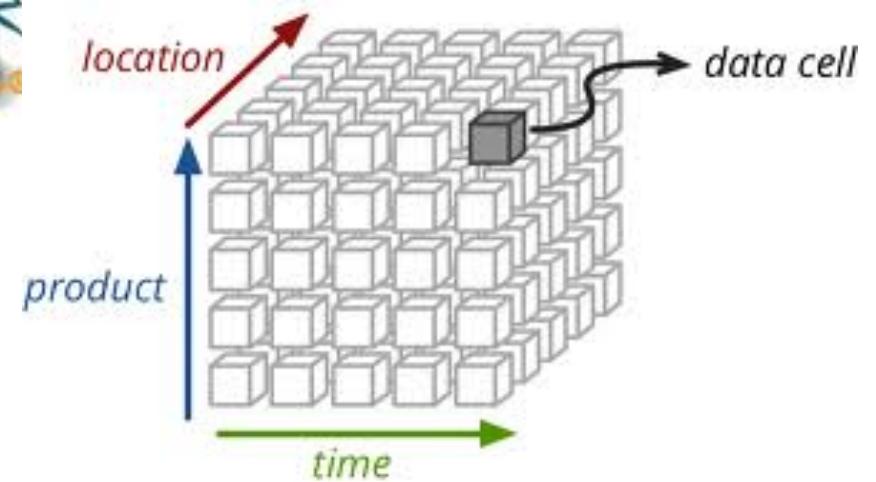
- Les expositions et les salons ;
- Les fournisseurs
- Les concurrents
- Les colloques
- Les sources internes de l'entreprise
- Certains sites Web : visible / invisible ;
- Les réseaux personnels.



Types de données



Type de données - Structurées



Type de données – Non-Structurées



United States Patent and Trademark Office
An Agency of the Department of Commerce

espacenet
1998–2008

inpi

WORLD
INTELLECTUAL
PROPERTY
ORGANIZATION



經濟產業省
特許庁
Japan Patent Office



Office de la propriété
intellectuelle du Canada
Un organisme
d'Industrie Canada

Canadian Intellectual
Property Office
An Agency of
Industry Canada

Australian Government
IP Australia

CNKI 中国知识
www.cnki.net
中国知识基础设施工程

ScienceDirect

IEEE

EBSCO
HOST

PubMed.gov

Le Monde.fr

L'EXPRESS.fr

KOMPASS

TIMESONLINE



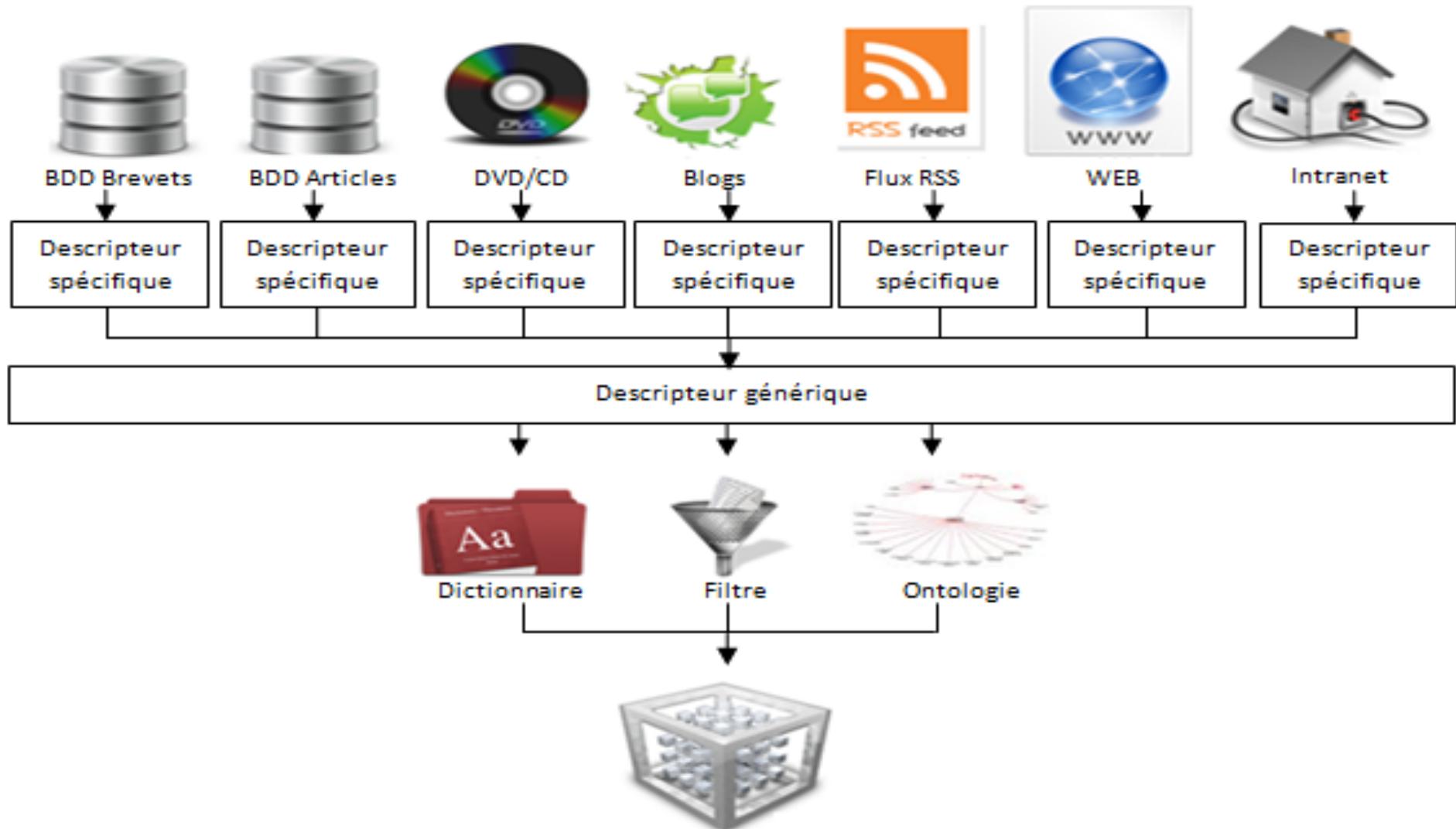
Sourcing..... ?





L'exploration et la préparation de données

Homogénéisation et structuration des données



Exemple 1: La veille brevets

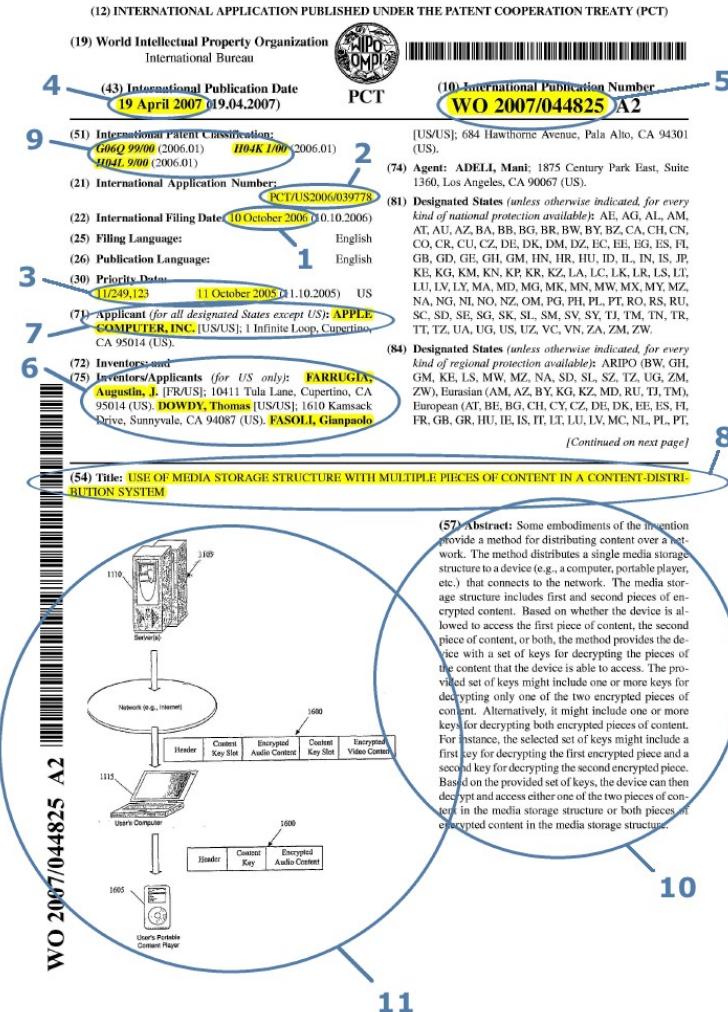
Contenu et structure des documents brevets

Un brevet comporte essentiellement trois parties :

- la **page de garde** contenant les informations juridiques telles que les dates, numéros, inventeurs et propriétaires du brevet,
- la **description** destinées à convaincre le lecteur (essentiellement l'examinateur et le juge) que les conditions de brevetabilité, en particulier l'inventivité et la suffisance de description, sont respectées : la première partie du brevet est le mémoire descriptif, où l'on brosse un tableau de l'état de la technique à la date de dépôt de la demande de brevet, où l'on présente le sommaire de l'invention, tandis que la seconde partie de la description décrit, avec si possible des figures, un ou plusieurs modes de réalisation préférés de l'invention,
- les **revendications** qui constituent la base juridique de la protection d'une invention en ce qu'elles délimitent l'étendue de la protection conférée par le brevet.



Exemple 1: La veille brevets la page de garde



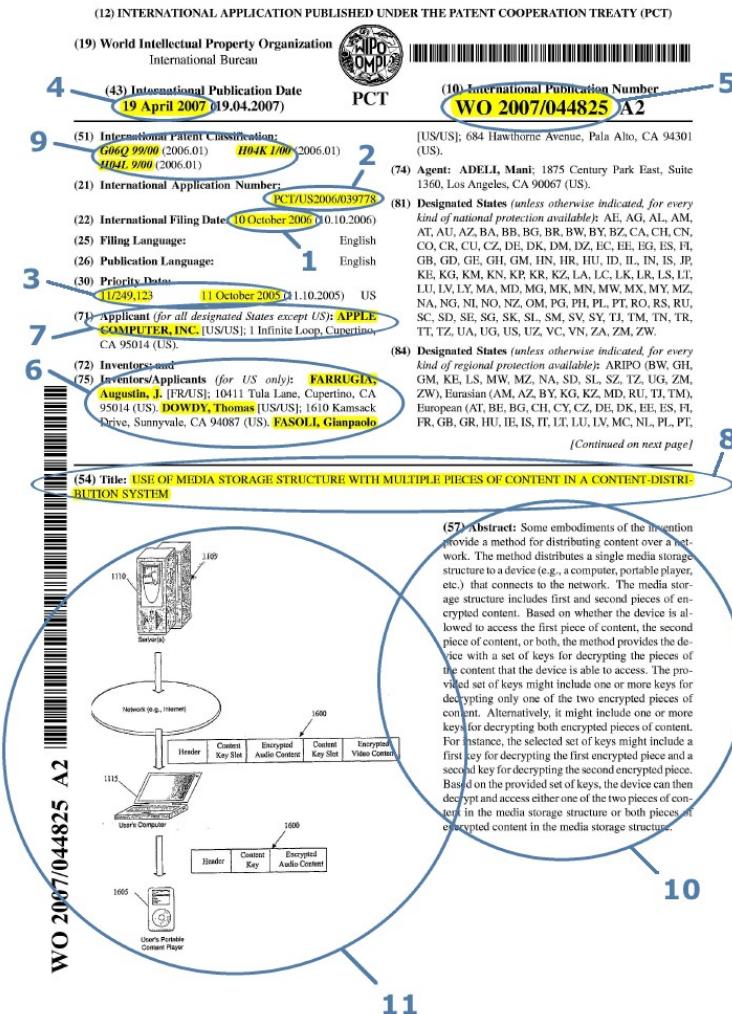
1. Date de dépôt - Il s'agit de la date à laquelle la demande de brevet a été déposée. Cette date est très importante : seules peuvent être citées comme antériorités opposables au brevet les divulgations antérieures à la date de dépôt de la demande de brevet.

2. Numéro de dépôt - Un numéro de série est attribué à une demande de brevet lors de son dépôt.

3. Priorité - Le brevet peut bénéficier d'un droit de priorité d'une demande antérieure pour la même invention. Cela a pour effet que la date de dépôt ("date effective") est réputée être la date de dépôt de la demande antérieure aux fins de déterminer l'état de la technique opposable au brevet. Lorsqu'une telle « priorité » est revendiquée, la date de dépôt, le pays et le numéro de dépôt de cette demande antérieure sont indiquées dans le champ « priorité ».

4. Date de publication - La date de publication est la date à laquelle le document de brevet est mis à la disposition du public, rejoignant ainsi l'état de la technique. Cette date est généralement postérieure de 18 mois à la date de dépôt ou de priorité du brevet, sauf si le déposant demande la publication anticipée du brevet.

Exemple 1: La veille brevets la page de garde



5. Numéro de publication - Un second numéro de série, qui diffère du numéro de dépôt (2), est attribuée au document de brevet mis à la disposition du public.

6. Inventeur(s) - Le brevet doit comporter la désignation du ou des inventeurs, personnes physiques, de l'invention objet du brevet. L'ordre de désignation des inventeurs ne confère aucun droit particulier.

7. Demandeur(s) ou cessionnaire(s) - Le demandeur ou le cessionnaire est le propriétaire (on dit aussi « titulaire ») du brevet. Le droit au brevet appartient à l'inventeur, mais peut être transféré à une autre personne physique ou morale. La question du droit au brevet est traitée plus en détail ici.

8. Titre - Le titre du brevet doit décrire succinctement l'objet de la demande de brevet.

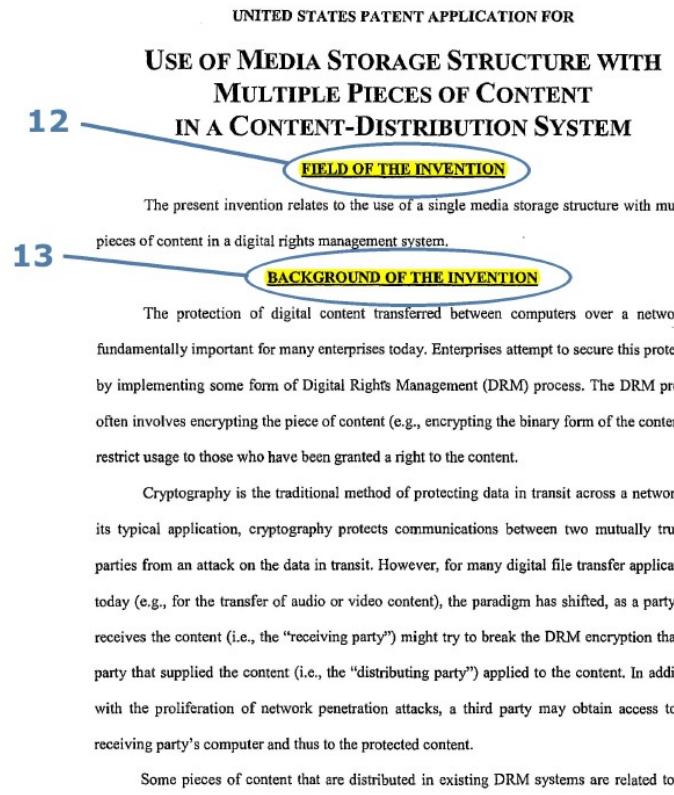
9. Domaines de recherche - L'Office des brevets utilise un système de classification pour classer les technologies. Les classes sont utile pour effectuer des recherches d'antériorités ou thématiques.

10. Abrégé / 11. Dessin d'abrégé - Un abrégé et un dessin d'accompagnement aident les lecteurs à déterminer rapidement si le brevet concerne une matière qui les intéresse. L'abrégué n'a pas de valeur juridique pour déterminer la portée du brevet.

Exemple 1: La veille brevets la description

WO 2007/044825

PCT/US2006/039778



Exemple 1: La veille brevets la description

WO 2007/044825

PCT/US2006/039778

14

SUMMARY OF THE INVENTION

Some embodiments of the invention provide a method for distributing content over a network. The method distributes a single media storage structure to a device (e.g., a computer, portable player, etc.) that connects to the network. The media storage structure includes first and second pieces of encrypted content. Based on whether the device is allowed to access the first piece of content, the second piece of content, or both, the method provides the device with a set of keys for decrypting the pieces of the content that the device is able to access.

The provided set of keys might include one or more keys for decrypting only one of the two encrypted pieces of content. Alternatively, it might include one or more keys for decrypting both encrypted pieces of content. For instance, the selected set of keys might include a first key for decrypting the first encrypted piece and a second key for decrypting the second encrypted piece. Based on the provided set of keys, the device can then decrypt and access either one of the two pieces of content in the media storage structure or both pieces of encrypted content in the media storage structure.

The media storage structure includes a first content section that stores the first piece of encrypted content, and a second content section that stores the second piece of encrypted content. In some embodiments, the media storage structure also includes first and second key sections respectively for storing first and second keys for decrypting the first and second pieces of encrypted content. The method of some embodiments distributes the media storage structure with the encrypted first and second content pieces from a computer that is separate from the computer or computers that distribute the first and second keys. In some embodiments, the device that receives the media storage structure inserts the first and second keys in the first and second key sections of the media storage structure.

3

Attorney Docket: P0106

SUBSTITUTE SHEET (RULE 26)

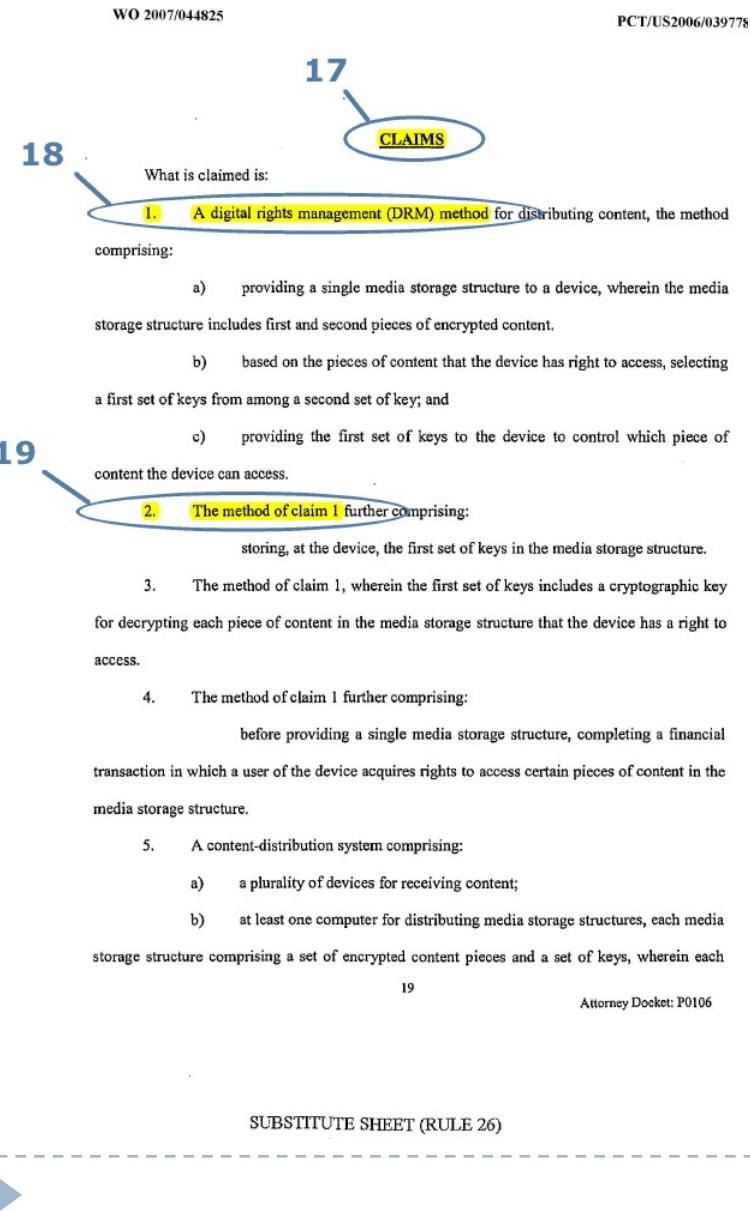
14. Objet de l'invention - L'objet de l'invention est énoncé dans une brève description des principales caractéristiques de l'invention formulée en termes généraux ne faisant pas référence aux dessins. Cette partie peut également expliquer comment l'invention fonctionne et comment la combinaison de ces caractéristiques principales vont résoudre le problème.

15. Brève description des dessins - La brève description des dessins permet au lecteur de déterminer ce qu'il est en train de regarder sans avoir besoin de se référer d'emblée à la description détaillée de l'invention.

16. Description détaillée de la solution préférée - Cette partie du brevet comprend une description détaillée de l'invention faisant généralement référence aux dessins et donnant suffisamment d'informations pour que l'homme du métier soit en mesure de réaliser et d'utiliser le meilleur mode de réalisation de l'invention connu défini par l'inventeur au moment du dépôt.

Exemple 1: La veille brevets

Les revendications : définition et limite de la propriété



17. Revendications - Les revendications constituent la partie juridiquement contraignante du brevet. Les revendications **délimitent l'étendue de la protection** en indiquant les caractéristiques de l'invention que l'on souhaite protéger. Un titulaire du brevet a seulement la droit d'empêcher d'autres personnes de fabriquer, d'utiliser, de vendre, d'offrir à la vente, d'importer ou d'exporter ce qui est revendiqué dans le brevet. Le simple fait d'illustrer sur les dessins ou de décrire dans un paragraphe une idée ou une caractéristique ne confère sur celle-ci aucune protection.

18. Revendication indépendante - Le jeu de revendications comprend au moins une revendication indépendante (au moins la revendication 1) indiquant les caractéristiques essentielles de l'invention et donnant la protection la plus large. Il peut y avoir plusieurs revendications indépendantes définissant différentes aspects ou différentes composantes de l'invention.

En Europe, chaque revendication indépendante comprend:

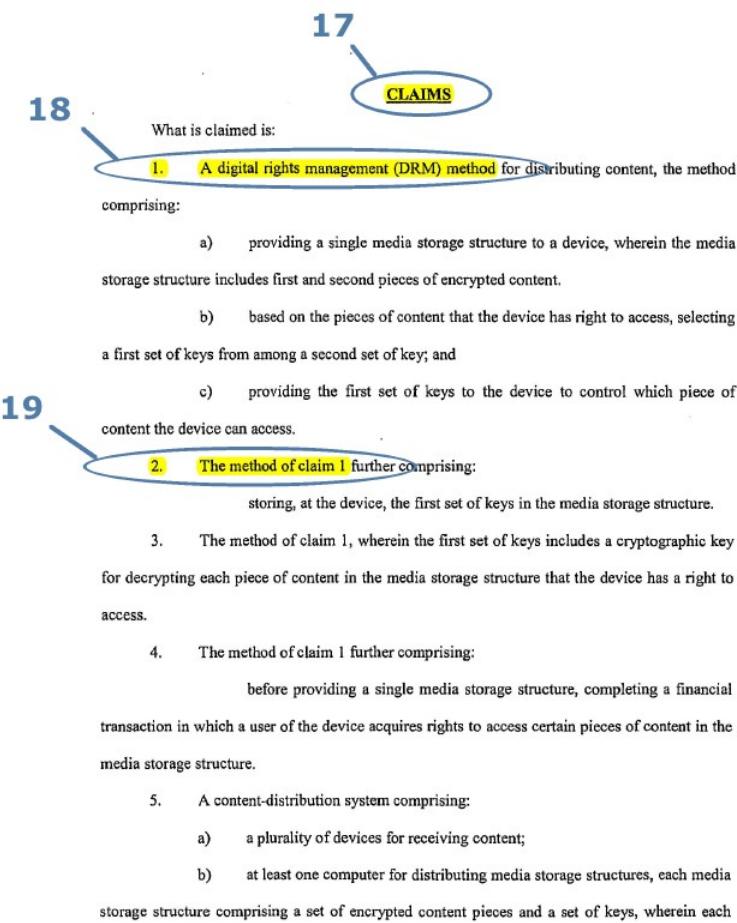
- une partie préambule désignant l'objet de l'invention et les caractéristiques techniques de l'invention qui sont déjà connues,
- une partie caractérisante précédée par l'expression du type «caractérisé en ce que» et qui exposent les caractéristiques techniques nouvelles de l'invention revendiquée.

Exemple 1: La veille brevets

Les revendications : définition et limite de la propriété

WO 2007/044825

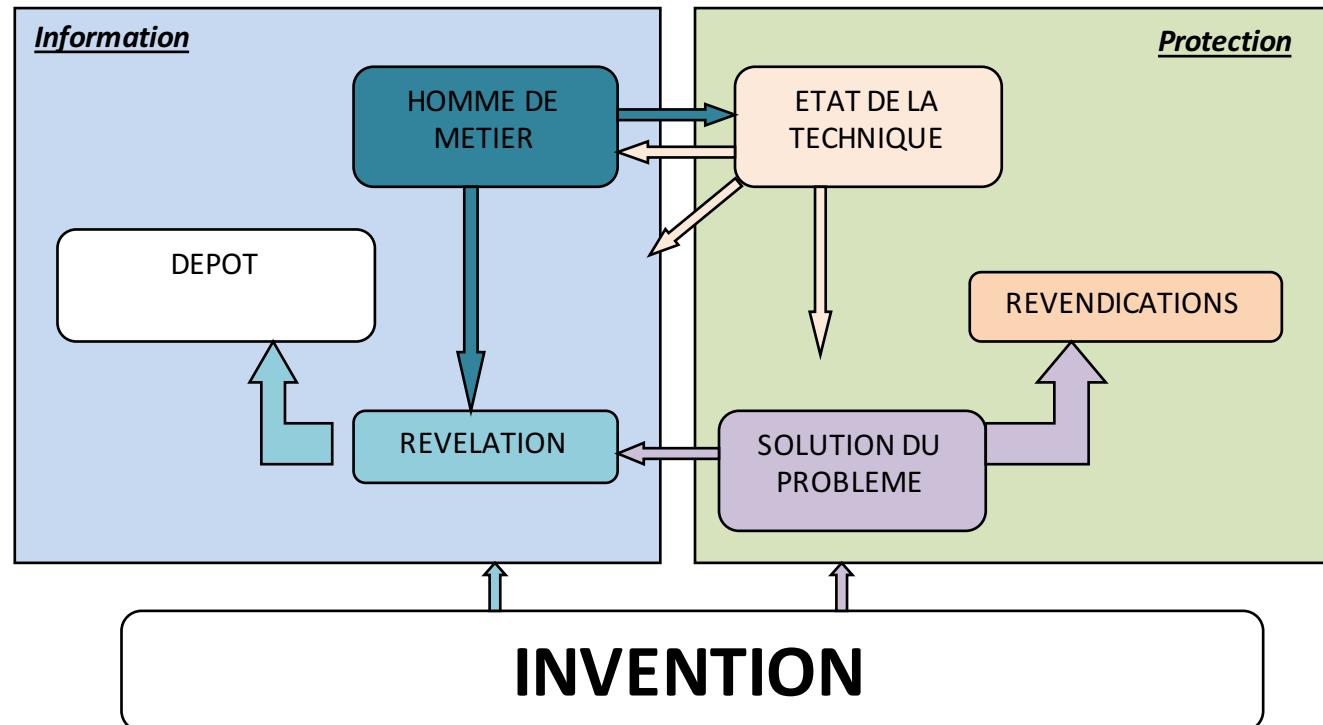
PCT/US2006/039776



19. Revendication dépendante - Le jeu de revendications comprend généralement plusieurs revendications dépendantes (les sous-revendications) protégeant des caractéristiques secondaires de l'invention. Une revendication dépendante comprend toutes les limitations de la revendication indépendante à laquelle elle se rattache, ainsi que les limitations (caractéristiques techniques) additionnelles contenues dans la revendication dépendante. Si quelqu'un trouve une façon de contourner une revendication indépendante en s'abstenant d'en reproduire au moins une caractéristique, ce contournement vaut alors nécessairement pour les sous-revendications qui dépendent de cette revendication indépendante.

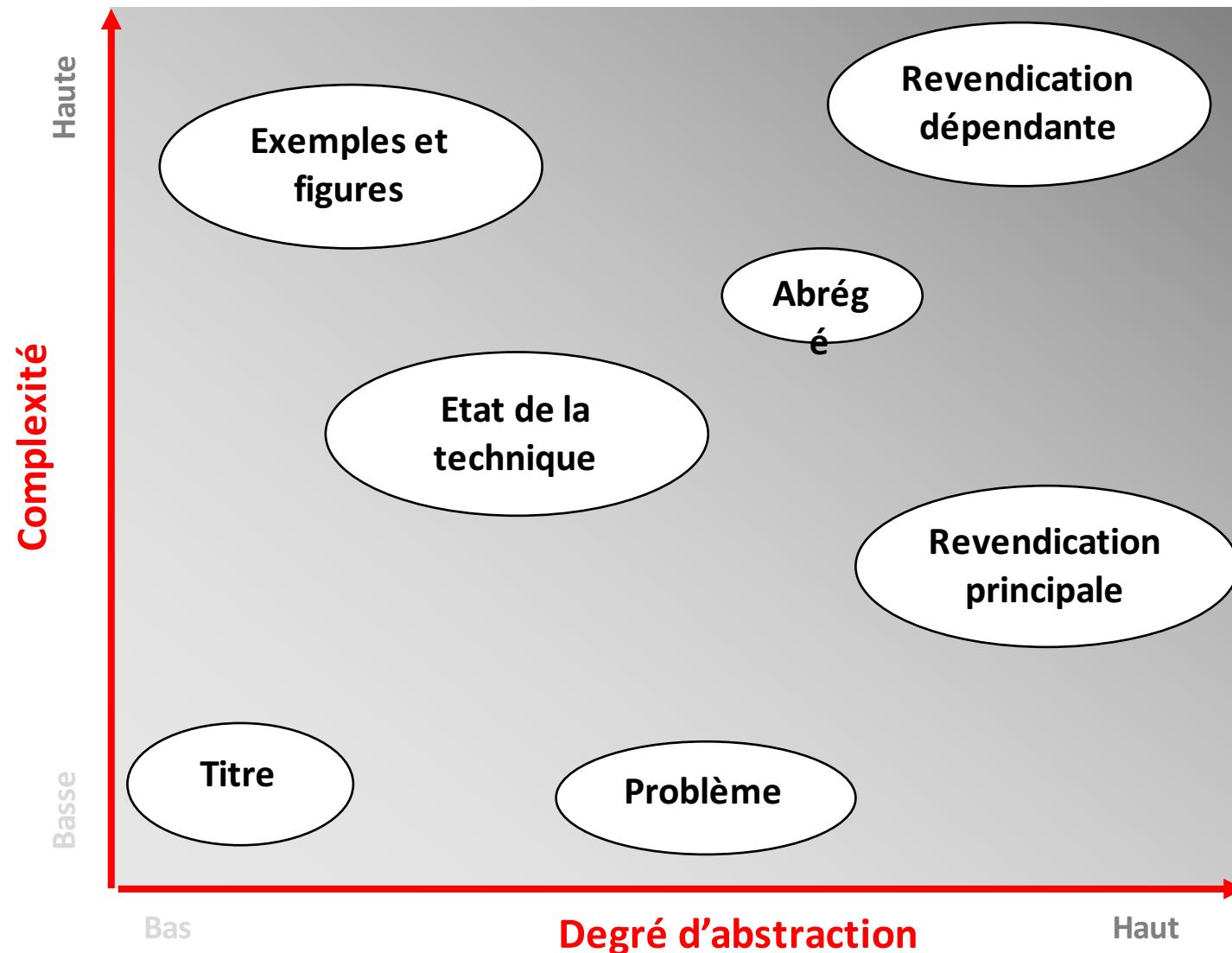
Exemple 1: La veille brevets

Terminologie des brevets



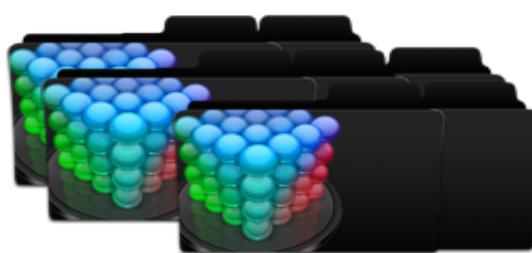
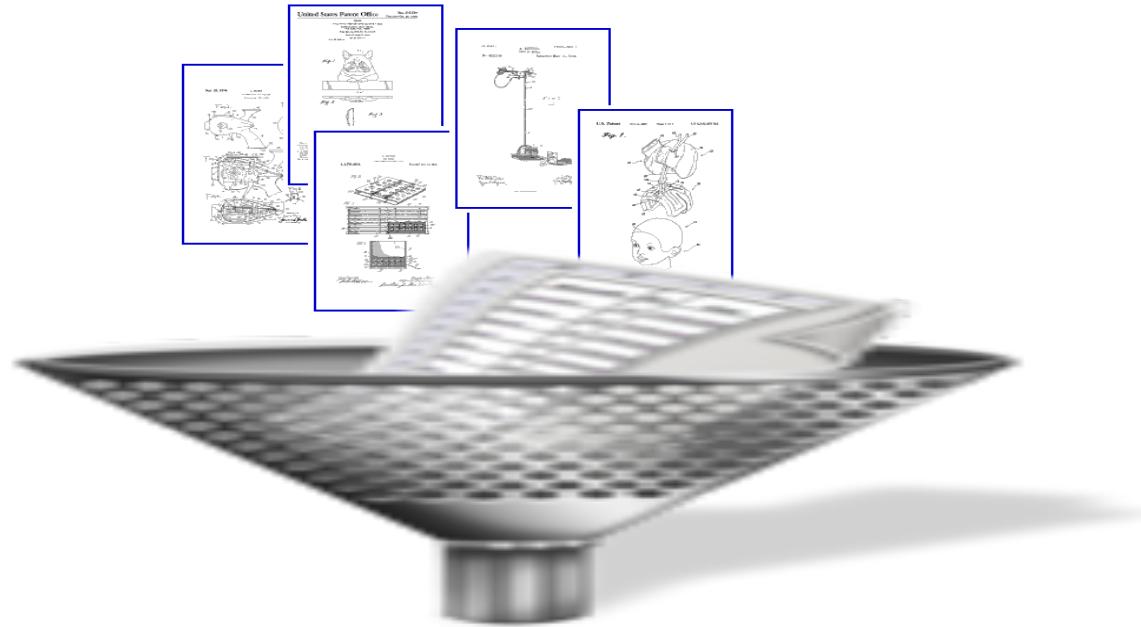
Exemple 1: La veille brevets

Complexité



Exemple 1: La veille brevets

Et après

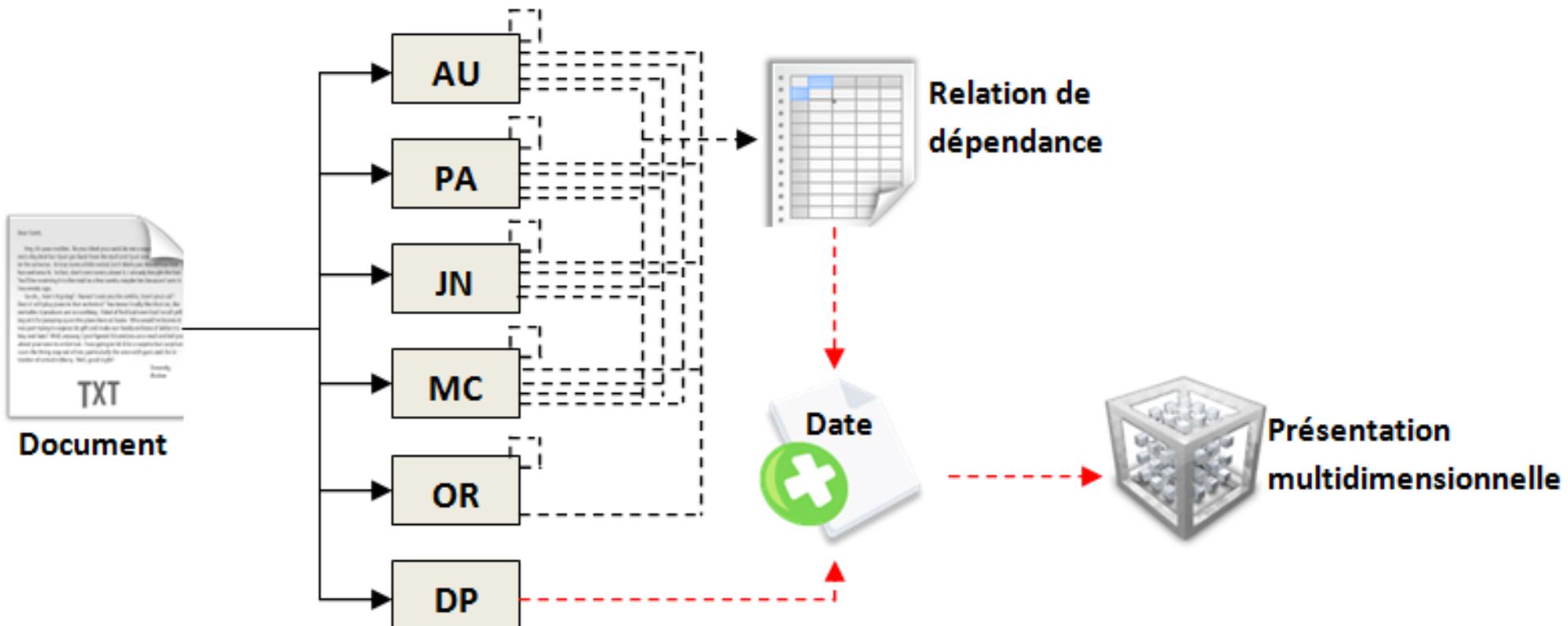


Pays
Classification CIB
Dates
Inventeur
Annuaire professionnel
Mandataire
Organisme
Titre
Abrégé
Mots-clés



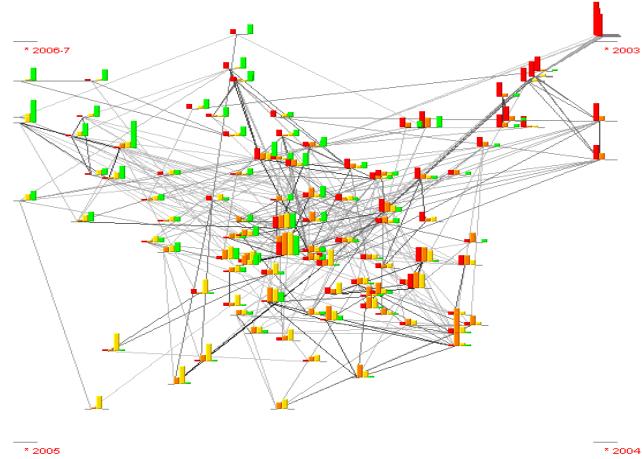
Exemple 1: La veille brevets

Et après



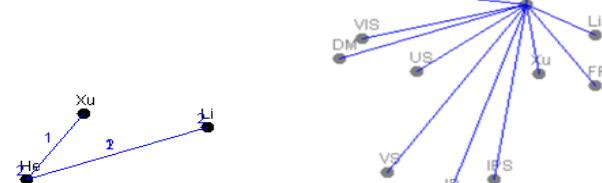
Exemple 1: La veille brevets

Et après



Profils Réseau social Réseau sémantique

Réseau social de l'acteur Li



Environnement

Profils

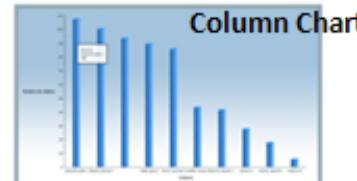
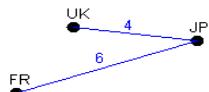
Réseau social

Réseau sémantique

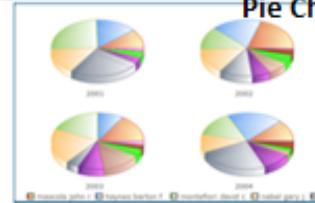
Réseau international

Evolution international

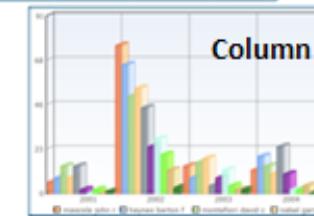
Réseau international de l'acteur Li



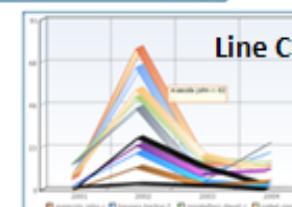
Pie Chart / Years



Column Chart / Years



Line Chart / Years



Geographical Chart



Exemple 1: La veille brevets

.... Maintenant

Démarrez là où les autres se sont arrêtés !



Exemple 2: CRM

Inventaires des données existantes

L'entreprise dispose de données, mais sous une forme malheureusement **impropre** à une utilisation en DM :

- Des données de détail disponibles sur microfilms, utilisables pour une recherche ponctuelle mais dont la reprise complète sous une forme informatique exploitable serait d'un coût exorbitant ;
- Des données de niveau client, mais agrégées en fin d'année pour avoir une synthèse annuelle (on a par exemple le nombre et le montant des achats, mais sans détails sur les dates, les objets...)
- Des données mensuelles, comme celles dont on a besoin, mais agrégées au niveau du magasin et non du client

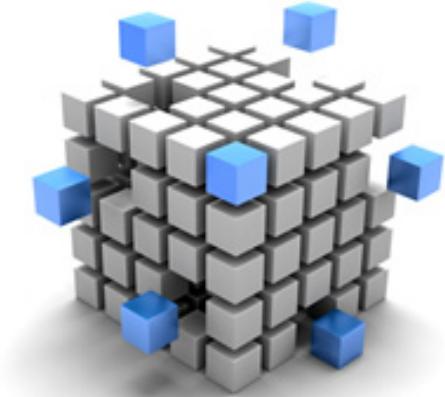


Exemple 2: CRM

Types de données

▶ Données de transaction

- ▶ « où » (lieux des transactions, Internet...)
- ▶ « quand » (fréquence / récence des transactions)
- ▶ « comment » (mode de paiement)
- ▶ « combien » (nombre et montants des transactions)
- ▶ « quoi » (ce qui est acheté)



▶ Données sur les produits et contrats

- ▶ Nb, types, options, prix, date d'achat ou de souscription, date et motif de résiliation ou de retour du produit, durée moyenne de vie ou date d'échéance, délai et mode de paiement, remise accordée au client, marge de l'entreprise

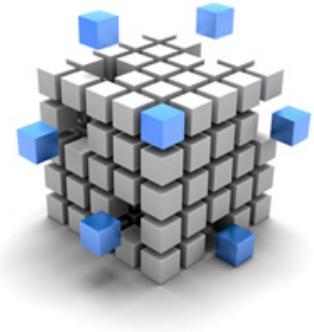
▶ Anciennetés

- ▶ Âge, ancienneté comme client, ancienneté à l'adresse actuelle, ancienneté dans l'emploi, ancienneté du dernier sinistre (en assurance)



Exemple 2: CRM

Types de données

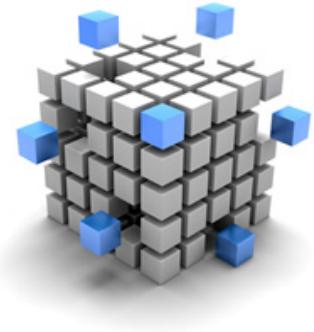


- ▶ Données sur les canaux
 - ▶ Canal de prise de contact (parrainage, annonce de presse, appel téléphonique, réponse à un mailing...)
 - ▶ Canal privilégié de contact et communication (courrier, téléphone, Internet, magasin/agence...)
 - ▶ Canal privilégié de commande (courrier, téléphone, Internet, magasin/agence...)
 - ▶ Canal privilégié de livraison (magasin/agence, domicile...)
- ▶ Données relationnelles et attitudinales
 - ▶ Réactions aux propositions commerciales, réponses aux questionnaires, réponses aux enquêtes de satisfaction, appels au service clientèle, réclamations
 - ▶ Image de la marque auprès du client, attractivité des concurrents, propension ou inertie du client au changement



Exemple 2: CRM

Types de données

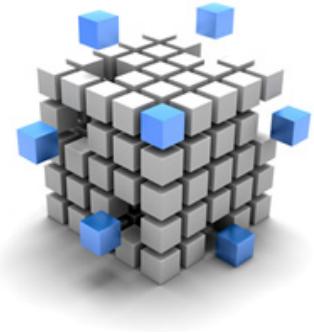


- ▶ Données sociodémographiques
 - ▶ Familiales (situation de famille, nb d'enfants et leur âge, nombre de personnes à charge)
 - ▶ Professionnelles (salaire, PCS, nb d'actifs dans le ménage)
 - ▶ Patrimoniales (patrimoine mobilier et immobilier, statut de propriétaire/locataire, valeur du logement, possession d'une résidence secondaire...)
 - ▶ Géographique (ancienneté de l'adresse)
 - ▶ Environnementales et géomarketing (concurrence, population, population active, population cliente, taux de chômage, potentiel économique, taux de détention de produit ... dans la zone d'habitation du client)



Exemple 2: CRM

Données à ne pas utiliser

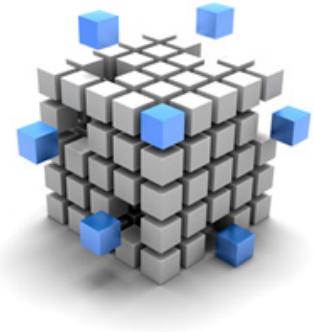


- ▶ **Non fiables**
 - ▶ Trop de valeurs aberrantes ou manquantes
- ▶ **Disponibles sur une durée trop courte**
 - ▶ Soumises aux variations saisonnières
- ▶ **Redondantes**
 - ▶ Dont le poids est artificiellement augmenté, ou dont la colinéarité rend instable les résultats de certaines méthodes
- ▶ **Non pertinentes**
 - ▶ Qu'il faut remplacer par de nouveaux indicateurs
- ▶ **Très corrélées à l'objectif de l'étude mais seulement dans l'échantillon d'apprentissage**
 - ▶ Qui entraînent un « sur-apprentissage » dans les prédictions
- ▶ **Trop peu corrélées à l'objectif de l'étude**
 - ▶ Qui créent du « bruit », des fluctuations aléatoires



Exemple 2: CRM

Sélection des données à utiliser

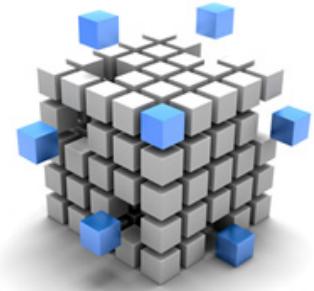


- ▶ Choix des variables les plus discriminantes
 - ▶ test du CHI2, V de Cramer (var. nominales) ou τ de Kendall (var. ordinaires)
 - ▶ test de la variance paramétrique (ANOVA) ou non (Kruskal-Wallis)
 - ▶ utilisation d'un arbre CHAID ou CART
- ▶ Transformation des variables (recodage, normalisation par un logarithme ou une racine carrée)
 - ▶ permet de se rapprocher d'une loi normale (var. quantitative)
 - ▶ permet de diminuer le nb de modalités (var. qualitative)
- ▶ Choix des discrétisations (découpage des var. continues)
 - ▶ ex : en fonction de la variable cible, à la main ou par utilisation d'un arbre CHAID ou CART
- ▶ Choix des variables les moins corrélées entre elle
 - ▶ tests de multicolinéarité



Exemple 2: CRM

Création de nouvelles variables



- ▶ Création d'indicateurs pertinents (maxima, moyennes, présence/absence...)
- ▶ Calcul de ratios
- ▶ Calcul d'évolutions temporelles de variables
- ▶ Création de durées, d'anciennetés à partir de dates
- ▶ Croisement de variables, interactions
- ▶ Utilisation de coordonnées factorielles
 - ▶ pour obtenir presque autant d'information avec moins de variables





Modèle d'analyse

Pour l'élaboration des modèles prédictifs

- ▶ Pré-segmentation (classification) de la population étudiée :
 - ▶ en groupes forcément distincts selon les données disponibles (clients / prospects)
 - ▶ en groupes statistiquement pertinents vis-à-vis des objectifs de l'étude
 - ▶ selon certaines caractéristiques sociodémographiques (âge, profession...) si elles correspondent à des offres marketing spécifiques
- ▶ Partition des données en :
 - ▶ un échantillon d'apprentissage
 - ▶ un échantillon de test
 - ▶ si possible, un échantillon de validation
- ▶ Mise en œuvre de une ou plusieurs techniques de datamining



Pré-segmentation : question opérationnelles

- ▶ Simplicité de la pré-segmentation (pas trop de règles)
- ▶ Nombre limité de segments et stabilité des segments
- ▶ Tailles généralement comparables des segments
- ▶ Homogénéité des segments du point de vue des variables explicatives
- ▶ Homogénéité des segments du point de vue de la variable à expliquer



Méthodes inductives : 4 étapes

- ▶ Apprentissage : **construction du modèle** sur un 1er échantillon pour lequel on connaît la valeur de la variable cible
- ▶ Test : **vérification du modèle** sur un 2d échantillon pour lequel on connaît la valeur de la variable cible, que l'on compare à la valeur prédite par le modèle
 - ▶ si le résultat du test est insuffisant (d'après la matrice de confusion ou la courbe ROC), on recommence l'apprentissage.
- ▶ **Validation du modèle** sur un 3^e échantillon, pour avoir une idée du taux d'erreur non biaisé du modèle
- ▶ **Application du modèle** à l'ensemble de l'population



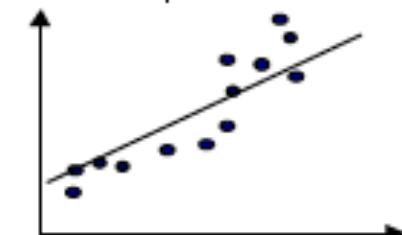
Exemple de modèles prédictifs

- ▶ Arbres de décision
 - ▶ Règles complètement explicites
 - ▶ Traitent les données hétérogènes, éventuellement manquantes, sans hypothèses de distribution
 - ▶ Détection de phénomènes non linéaires
 - ▶ Moindre robustesse
- ▶ Analyse discriminante linéaire
 - ▶ Résultat explicite $P(Y/X_1, \dots, X_p)$ sous forme d'une formule
 - ▶ Requiert des X_i continues, sans colinéarité, et des lois X_i/Y multinormales et homoscédastiques (attention aux « outliers »)
 - ▶ Optimale si les hypothèses sont remplies
- ▶ Régression logistique
 - ▶ Comme l'analyse discriminante, sans hypothèse sur les lois X_i/Y , X_i peut être discret, avec une précision parfois très légèrement inférieure

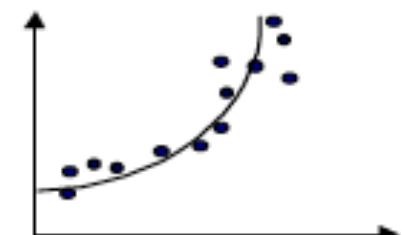


Validation des modèles

- ▶ Etape très importante car des modèles peuvent :
 - ▶ donner de faux résultats (données non fiables)
 - ▶ mal se généraliser dans l'espace (autre échantillon) ou le temps (échantillon postérieur)
 - ▶ sur-apprentissage
 - ▶ être peu efficaces (déterminer avec 2 % d'erreur un phénomène dont la probabilité d'apparition = 1 % !)
 - ▶ être incompréhensibles ou inacceptables par les utilisateurs
 - ▶ souvent en raison des variables utilisées
 - ▶ ne pas correspondre aux attentes
- ▶ Principaux outils de comparaison :
 - ▶ matrices de confusion, courbes ROC, de lift, et indices associés



(A) Modèle trop simple



(B) Bon modèle



(C) Modèle trop complexe



Formation des utilisateurs

- ▶ Présenter l'objectif recherché avec les nouveaux outils
- ▶ Principe et fonctionnement des outils de data mining
 - ▶ sans entrer dans les détails techniques
- ▶ Limites des outils
 - ▶ il ne s'agit que d'outils statistiques
- ▶ Mode d'utilisation
 - ▶ aide à la décision et non pas prise automatique de décision
- ▶ Apport des outils (c'est le point le plus important)
- ▶ Ce qui change dans le travail des utilisateurs
 - ▶ du point de vue opérationnel
 - ▶ du point de vue organisationnel (adaptation des procédures, des délégations de pouvoir...)
- ▶ Etape importante pour éviter des rejets !



Cycle de vie d'un score

- ▶ Les outils de data mining (scores surtout) ont une phase d'expérimentation
 - ▶ sur une petite échelle
 - ▶ destinée à les ajuster et valider, et tester leur utilisation
- ▶ Quand les outils sont en production, ils doivent être appliqués régulièrement à des données rafraîchies
- ▶ Les outils en production doivent être revus régulièrement (tous les 2 à 5 ans)
 - ▶ évolution de l'environnement concurrentiel, économique, sociodémographique, réglementaire
 - ▶ apparition, disparition, modification de produits



Suivi du score

- ▶ Suivi ponctuel pour une campagne marketing
 - ▶ pour analyser les résultats et améliorer le score suivant
 - ▶ comparer les résultats des individus ciblés à ceux d'un échantillon témoin (cible aléatoire ou traditionnelle)
- ▶ Suivi permanent pour l'utilisation commerciale
 - ▶ vérifier la bonne utilisation du score
 - ▶ s'assurer de la pertinence des « infractions » au score
 - ▶ vérifier le bon fonctionnement du score
 - ▶ pour un score de risque, le taux de défaillance dans chaque tranche de score doit rester à l'intérieur d'une fourchette fixée
 - ▶ vérifier la stabilité du modèle au fil des calculs
 - ▶ matrice de transition





Les facteurs de succès et les erreurs à éviter

Les facteurs de succès d'un projet

- ▶ Des objectifs précis, stratégiques et réalistes
- ▶ La qualité et la richesse des informations collectées
- ▶ Le stockage des informations relationnelles sur les clients
(réponses aux sollicitations commerciales, aux enquêtes de satisfaction, canaux de préférence...)
- ▶ La collaboration des compétences métiers et statistiques
- ▶ La maîtrise des techniques de data mining utilisées
- ▶ Une bonne restitution des résultats et l'implication de tous les partenaires chargés de leur mise en œuvre
- ▶ L'analyse des retours de chaque action pour la suivante



Le DM dans la culture d'entreprise

- ▶ L'entreprise doit veiller :
 - ▶ à ses compétences en data mining
 - ▶ à la qualité des données recueillies
 - ▶ à une mise en œuvre et un suivi rigoureux des actions s'appuyant sur le data mining
 - ▶ à une éventuelle adaptation de ses processus marketing
 - ▶ passer du marketing « produit » au marketing « client »
 - ▶ à une éventuelle adaptation de ses processus de décision
 - ▶ adaptation des délégations de pouvoir
- ▶ Le data mining est un processus itératif, chaque action préparant la suivante par l'exploitation de ses résultats



Les idées fausses sur le DM

- ▶ Aucun a priori n'est nécessaire
- ▶ On n'a plus besoin de spécialistes du métier
- ▶ On n'a plus besoin de statisticiens (« Il suffit d'appuyer sur un bouton »)
- ▶ Le data mining permet de faire des découvertes incroyables
- ▶ Le data mining est révolutionnaire
- ▶ Il faut utiliser toutes les données disponibles
- ▶ Il faut toujours échantillonner
- ▶ Il ne faut jamais échantillonner





Le recours au consulting

Internalisation ou externalisation

- ▶ Soit l'entreprise internalise l'activité de data mining, éventuellement avec l'aide de consultants spécialisés
- ▶ Soit elle externalise totalement cette activité, en fournissant ses fichiers de données à des prestataires spécialisés (les « credit-bureaux » pour la banque), ceux-ci lui restituant ses fichiers enrichis avec les informations de data mining qu'ils auront calculées (score, segment, etc.)
 - ▶ ne pas oublier de faire signer une clause de confidentialité
- ▶ Soit elle sous-traite la fabrication des modèles de DM, mais se les fait livrer, afin de les appliquer elle-même à ses fichiers



Internalisation ou externalisation

- ▶ L'intérêt du recours à des prestataires est de disposer immédiatement de leur savoir et de leur expérience
- ▶ L'intérêt d'avoir des compétences en interne dans l'entreprise est de pouvoir :
 - ▶ acquérir une parfaite connaissance de ses données
 - ▶ avoir une plus grande réactivité lorsqu'une nouvelle étude est demandée
 - ▶ actualiser en permanence ses résultats
 - ▶ développer pour un coût bien plus faible quantité d'outils de score, de classification, de recherche d'association de produits... pour des besoins et des destinataires variés





Les caractéristiques importantes des logiciels

Méthodes implémentées

1. Prédiction (régression linéaire, modèle linéaire général, régression robuste, régression non-linéaire, régression PLS, arbres de décision, réseaux de neurones, k-plus proches voisins...);
2. Classement (analyse discriminante linéaire, régression logistique binaire, régression logistique polytomique (ordinale ou nominale), modèle linéaire généralisé, arbres de décision, réseaux de neurones, k-plus proches voisins);
3. Classification (centres mobiles, k-means, classification hiérarchique ascendante ou descendante, méthodes mixtes, méthodes par estimation de densité, cartes de kohonen);
4. Détection des associations;
5. Analyse de survie;
6. Analyse des séries temporelles.



Fonctions de préparation des données

1. Manipulation de fichiers (fusion, agrégation, transposition...);
2. Visualisation des données, coloriage des individus selon un critère;
3. Détection, filtrage et winsorisation des extrêmes;
4. Transformation de variables (recodage, standardisation, normalisation automatique, discréétisation...);
5. Création de nouvelles variables (fonctions prédéfinies logiques, chaînes, statistiques, mathématiques...);
6. Sélection des variables les plus explicatives, des discréétisations et des interactions.



Autres fonctions

1. Fonctions statistiques (détermination des caractéristiques de tendance centrale, de dispersion, de forme, tests statistiques de moyenne, de variance, de distribution, d'indépendance, d'hétéroscédasticité...);
2. Fonctions d'échantillonnage et de partition des données, pour pouvoir créer des échantillons d'apprentissage, de test et de validation, possibilités de bootstrap et de jackknife (ce dernier pour la validation croisée);
3. Fonctions d'analyses exploratoire des données et, en particulier, d'analyse factorielle (analyse en composantes principales, ACP avec rotation des axes, analyse factorielle des correspondances, analyses des correspondances multiples);
4. Visualisation des résultats, manipulation des tableaux, bibliothèque de graphiques en 2D, 3D, interactifs, navigation dans les arbres de décision, affichage des paramètres statistiques et des courbes de performances (ROC, lift, indice de Gini), facilité d'incorporation des ces éléments dans un rapport...
5. Langage avancé de programmation



Caractéristiques techniques

1. Plate-forme matérielle (Unix, Windows, Sun, IBM MVS...);
2. BD accédées (Oracle, Sybase, DB2, SAS, SQL Server, Access...);
3. Accès natif (plus fiable et beaucoup plus rapide) ou ODBC (plus simple à programmer) à ces BD;
4. Architecture client-serveur ou monoposte;
5. Algorithmes parallélisés ou non;
6. Volume de données maximum pouvant être traité (en un temps raisonnable);
7. Exécution en mode différé (« batch ») ou interactif (« transactionnel »);
8. Possibilité d'export des modèles (C, PMML, Java, SQL...).





Les principaux logiciels

Les principaux logiciel de statistique et de DM

Volume de données	Produit	Spécialité	Éditeur
Faible (dizaines de milliers d'enregistrements)	NeuralWorks Predict	RN	Neuralware
	NeuroOne	RN	Netral
	Wizwhy	Associations	Wizsoft
	WEKA		« OS » (Univ. de Waikato, Nouvelle-Zélande)
	R		« OS » (initialement à l'Univ. d'Auckland)
	DataLab	Prétraitement des données	Complex Systems



Les principaux logiciel de statistique et de DM

Volume de données	Produit	Spécialité	Éditeur
Moyen (centaines de milliers d'enregistrements)	Alice	AD	Isoft
	KnowledgeSEEKER	AD	Angoss
	KnowledgeSTUDIO		Angoss
	C5.0 (Unix) See5 (Windows)	AD	RuleQuest Research
	Data Mining Suite		Salford Systems
	CART	AD	Salford Systems
	Polyanalyst		Megaputer
	TANAGRA		Université de Lyon
	JMP		SAS
	S-PLUS		TIBCO Software



Les principaux logiciel de statistique et de DM

Volume de données	Produit	Spécialité	Éditeur
Élevé (millions d'enregistrements)	KXEN		KXEN
	Oracle Data Mining		Oracle
	SPAD		Coheris SPAD
	IBM SPSS Statistics		SPSS (groupe IBM)
	IBM SPSS Modeler		SPSS (groupe IBM)
	Statistica Data Miner		Statsoft
	Insightful Miner		TIBCO Software
	SAS/STAT		SAS
	Enterprise Miner		SAS





Un point de terminologie

Terminologie

Auteurs anglo-saxon	Certains auteurs francophones	Analyse des données à la française
Clustering	Segmentation	Classification
Classification	Classification	Classement, analyse discriminante
Decision trees	Arbres de décision	segmentation





Principales techniques

Les 2 types de techniques de DM

Les techniques descriptives

- visent à mettre en évidence des informations présentes mais cachées par le volume des données (ex: segmentation de clientèles et recherche d'associations de produits sur les tickets de caisse)
- Réduisent, résument, synthétisent les données
- **il n'y a pas de variable cible à expliquer**



Les techniques prédictives

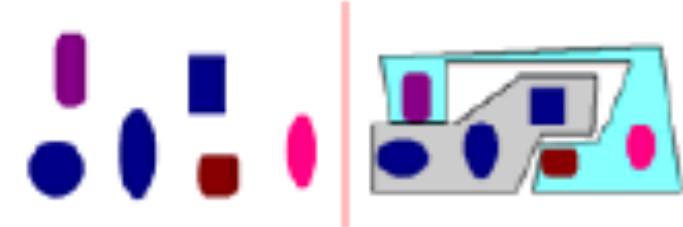
- visent à extrapoler de nouvelles informations à partir des informations présentes.
- Expliquent des données
- **il y a une variable cible à expliquer**



Les 2 types de techniques de DM

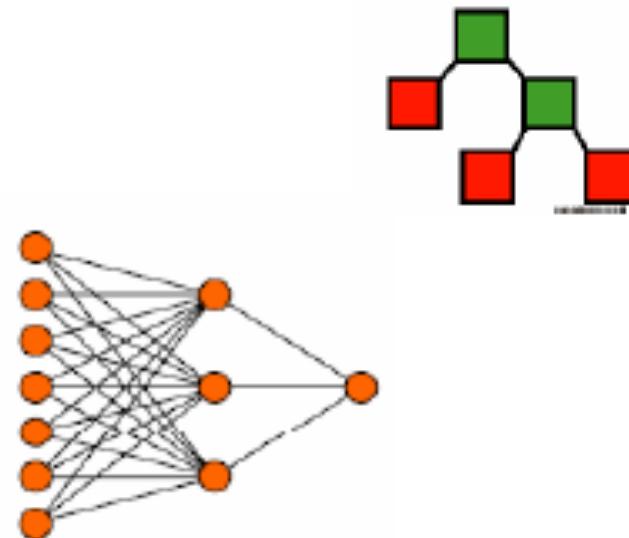
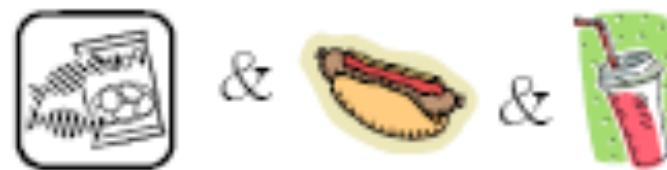
Les techniques descriptives

- Classification
- Recherche d'associations
- Recherche de séquence similaires



Les techniques prédictives

- Prédiction (variable **cible qualitative**)
 - Analyse discriminante /régression logistique
 - Arbre de décision
 - Réseaux de neurones
- Prédiction (variable **cible quantitative**)
 - Régression linéaire
 - Arbre de décision
 - Réseaux de neurones



Classification des techniques

Type	Famille	Sous-famille	Algorithme
Méthode descriptive	Modèles géométriques	Analyse factorielle (projection et visualisation dans un espace de dimension inférieure)	ACP (variables continues) AFC (variables qualitatives et binaires) ACM (variables qualitatives et binaires)
		Analyse typologique (regroupements dans tout l'espace en classes homogènes)	Méthodes de partitionnement (centres mobiles, k-means, nuées dynamiques...) Méthodes hiérarchiques (ascendantes, descendantes)
		Analyse typologique + réduction de dimension	Classification neuronale (réseaux de kohonen)
	Modèles combinatoires		Classification par agrégation des similarités (variables qualitatives)
	Modèles à base de règles logiques	Détection de liens	Recherche d'associations Recherche de séquences similaires



Classification des techniques

Type	Famille	Sous-famille	Algorithme
	Modèles à base de règles logiques	Arbres de décision	Arbres de décision (variable à expliquer continue ou qualitative)
Méthode prédictive	Modèles à base de fonctions mathématiques	RN	Réseaux à apprentissage supervisé (perception, réseau à fonction radiale de base...)
		Modèles paramétriques ou semi-paramétriques	Régression linéaire, modèle linéaire général GLM, régression PLS (variable à expliquer continue)
			Analyse discriminante de Fisher, régression logistique
			Modèle log-linéaire
			Modèle linéaire généralisé GLZ, modèle additif généralisé GAM (variable à expliquer continue, discrète, comptage ou qualitative)
	Prédiction sans modèle	Analyse probabiliste	K-plus proches voisins (k-NN)



Comparatif des techniques

Techniques	Absence d'hypothèse sur le problème à résoudre	Traitement exhaustif des BD	Traitement des données hétérogènes ou lacunaire
Classification			
Méthode des centres mobiles et ses variantes	Non (nombre de classes et centres initiaux fixés)	Oui	Variables numériques et sans valeurs manquantes
Classification hiérarchique	Oui, mais les classes au niveau n sont déterminées par ceux au niveau n-1	Non (algorithme non linéaire), on ne peut traiter plus de quelques milliers d'observations	Oui (possibilité de traiter des variables non numériques avec une distance ad hoc)
Classification neuronale (kohonen)	Non (nombre de classe fixé)	Oui	Les variables $\notin [0,1]$ doivent être transformées
Classification par agrégation des similaires	Oui	Dépend de l'implémentation	Variables qualitatives



Comparatif des techniques

Techniques	Absence d'hypothèse sur le problème à résoudre	Traitement exhaustif des BD	Traitement des données hétérogènes ou lacunaire
Classement et prédition			
Arbre de décision	Comme la classification hiérarchique (sorte « d'arbre à l'envers)	Non (mais moins vite limité que la classification hiérarchique)	Certains arbres comme CHAID doivent discréteriser les variables continues
Réseaux de neurones perceptions	Oui (mais il faut fixer le nombre de neurones cachés)	Non (pas d'apprentissage sur plusieurs variables)	Les variables $\notin [0,1]$ doivent être transformées
Réseaux à fonction radiale de base	Comme les perceptions	Oui	Les variables $\notin [0,1]$ doivent être transformées
Analyse discriminante	Non (relations linéaires entre les variables et hypothèses sur les lois conditionnelles X_i/Y)	Oui	Variables numériques et sans valeurs manquantes



Comparatif des techniques

Techniques	Absence d'hypothèse sur le problème à résoudre	Traitement exhaustif des BD	Traitement des données hétérogènes ou lacunaire
Classement et prédition			
Analyse discriminante sur coordonnées factorielle d'une ACM (méthode DISQUAL)	Oui (permet de s'affranchir largement des hypothèses sur les lois conditionnelles X_i/Y)	Oui	Oui (les valeurs manquantes sont traitées comme des valeurs à part entière)
Régression linéaire, régression PLS	Non (relations linéaires entre les variables + autres hypothèses)	Oui	Variables numériques et sans valeurs manquantes
Régression logistique, modèle linéaire généralisé	Oui	Oui (sur une machine assez puissante, si le nombre d'observations est très grand)	Oui (découper en classes les variables continues avec des valeurs manquantes)



Comparatif des techniques

Techniques	Absence d'hypothèse sur le problème à résoudre	Traitement exhaustif des BD	Traitement des données hétérogènes ou lacunaire
Associations			
Recherche d'associations	Oui	Dépend du paramétrage	Oui
Séquences similaires	Oui	Oui (idem)	Oui



Tâches du Data Mining



- ▶ Classification
- ▶ Clustering (Segmentation)
- ▶ Recherche d'associations
- ▶ Recherche de séquences
- ▶ Détection de déviation



Classification

Elle permet de prédire si une instance de donnée est membre d'un groupe ou d'une classe prédéfinie.

- ▶ Classes
 - ▶ Groupes d'instances avec des profils particuliers
 - ▶ **Apprentissage supervisé** : classes connues à l'avance
- ▶ Applications : marketing direct (profils des consommateurs), grande distribution (classement des clients), médecine (malades/non malades), etc.
- ▶ Exemple : les acheteurs de voitures de sport sont de jeunes citadins ayant un revenu important



Clustering (Segmentation)

Partitionnement logique de la base de données en clusters

- ▶ Clusters : groupes d'instances ayant les mêmes caractéristiques
- ▶ **Apprentissage non supervisé** (classes inconnues)

- ▶ Pb : interprétation des clusters identifiés
- ▶ Applications : Economie (segmentation de marchés), médecine (localisation de tumeurs dans le cerveau), etc.



Règles d'association

Corrélations (ou relations) entre attributs (méthode non supervisée)

- ▶ Applications : grande distribution, gestion des stocks, web (pages visitées), etc.
- ▶ Exemple
 - BD commerciale : panier de la ménagère
 - Articles figurant dans le même ticket de caisse
 - Ex : achat de couscous + légume ==> achat de viande



Recherche de séquences

- ▶ Recherche de séquences
 - ▶ Liaisons entre évènements sur une période de temps
 - ▶ Extension des règles d'association
 - ▶ Prise en compte du temps (série temporelle)
 - ▶ Achat Télévision ==> Achat Magnétoscope d'ici 5 ans
 - ▶ Applications : marketing direct (anticipation des commandes), bioinformatique (séquences d'ADN), bourse (prédiction des valeurs des actions)
- ▶ Exemple
 - ▶ BD commerciale (ventes par correspondance)
 - ▶ Commandes de clients
 - ▶ Ex : 60% des consommateurs qui commandent la bière «Mort subite» commandent de l'aspro juste après
 - ▶ Séquences d'AND : ACGTC est suivie par GTCA après un gap de 9, avec une probabilité de 30%



Détection de déviation

- ▶ Instances ayant des caractéristiques les plus différentes des autres
 - ▶ Basée sur la notion de distance entre instances
 - ▶ Expression du problème
 - ▶ Temporelle : évolution des instances ?
 - ▶ Spatiale : caractéristique d'un cluster d'instances ?
- ▶ Applications
 - ▶ Détection de fraudes (transactions avec une carte bancaire inhabituelle en telemarketing)
- ▶ Caractéristiques
 - ▶ Problème d'interprétation : bruit ou exception (donc connaissance intéressante)





Clustering

Qualité d'un clustering

- ▶ Soient N instances de données à k attributs,
- ▶ Trouver un partitionnement en c clusters (groupes) ayant un sens (Similitude)
- ▶ Affectation automatique de “labels” aux clusters
- ▶ c peut être donné, ou “découvert”
- ▶ Plus difficile que la classification car les classes ne sont pas connues à l'avance (non supervisé)
- ▶ Attributs
 - ▶ Numériques (distance bien définie)
 - ▶ Enumératifs ou mixtes (distance difficile à définir)



Qualité d'un clustering

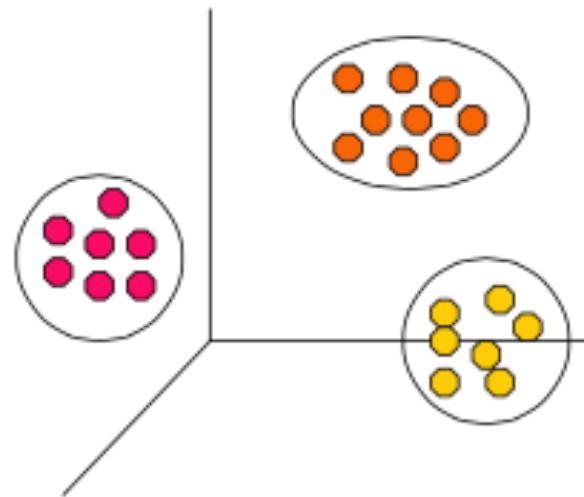
- ▶ Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
 - ▶ Similarité **intra-classe** importante
 - ▶ Similarité **inter-classe** faible
- ▶ La qualité d'un clustering dépend de :
 - ▶ La mesure de similarité utilisée
 - ▶ L'implémentation de la mesure de similarité
- ▶ La **qualité d'une méthode** de clustering est évaluée par son capacité à découvrir certains ou tous les “patterns” cachés.



Objectifs du clustering

Minimiser les distances
intra-cluster

Maximiser les distances
inter-clusters



Exemples d'applications

- ▶ **Marketing**: segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats.
- ▶ **Environnement**: identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observation de la terre.
- ▶ **Assurance**: identification de groupes d'assurés distincts associés à un nombre important de déclarations.
- ▶ **Planification de villes** : identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique,
...
- ▶ **Médecine** : Localisation de tumeurs dans le cerveau
 - ▶ Nuage de points du cerveau fournis par le neurologue
 - ▶ Identification des points définissant une tumeur



Mesure de la similarité



- ▶ **I n'y a pas de définition unique de la similarité entre objets**
 - ▶ Différentes mesures de distances $d(x,y)$
- ▶ **La définition de la similarité entre objets dépend de :**
 - ▶ Le type des données considérées
 - ▶ Le type de similarité recherchée



Choix de la distance

- ▶ Propriétés d'une distance :

1. $d(x,y) \geq 0$
2. $d(x,y) = 0$ iff $x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)$

- ▶ Définir une distance sur chacun des champs
- ▶ Champ numériques : $d(x,y) = |x-y|$, $d(x,y) = |x-y|/d_{max}$ (distance normalisée).
- ▶ Exemple : Age, taille, poids, ...



Distance – Données numériques

- ▶ Combiner les distances : Soient $x=(x_1, \dots, x_n)$ et $y=(y_1, \dots, y_n)$
Exemples numériques :

- ▶ Distance de Manhattan :

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- ▶ Distance euclidienne :

$$d(x, y) = \sqrt{\sum_{i=1}^n |x_i - y_i|}$$

- ▶ Distance de Minkowski :

$$d(x, y) = \sqrt[q]{\sum_{i=1}^n |x_i - y_i|^q}$$

- ▶ $k=1$: distance de Manhattan.
- ▶ $k=2$: distance euclidienne



Méthodes de Clustering



- ▶ Méthode de partitionnement (K-moyennes)
 - ▶ Méthodes hiérarchiques (par agglomération)
 - ▶ Méthode par voisinage dense
-
- ▶ Caractéristiques
 - ▶ Apprentissage non supervisé (classes inconnues)
 - ▶ Pb : interprétation des clusters identifiés



Méthodes de Clustering - Caractéristiques



- ▶ Extensibilité
- ▶ Abilité à traiter différents types de données
- ▶ Découverte de clusters de différents formes
- ▶ Connaissances requises
- ▶ (paramètres de l'algorithme)
- ▶ Abilité à traiter les données bruitées et isolées.



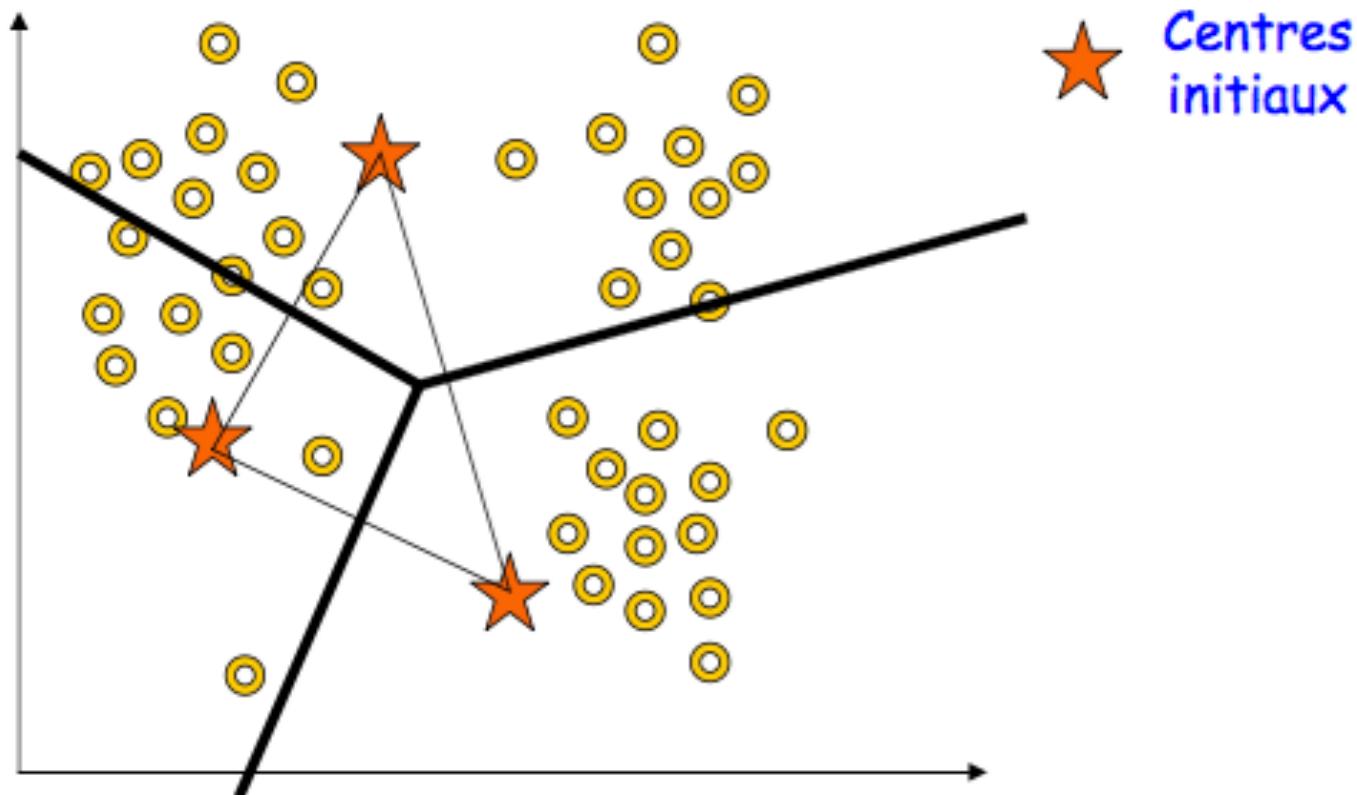
Algorithme des k-moyennes (k-means)

- ▶ Entrée : un échantillon de m enregistrements x_1, \dots, x_m
- ▶ 1. Choisir k centres initiaux c_1, \dots, c_k
- ▶ 2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.
- ▶ 3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes
- ▶ 4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.
- ▶ Aller en 2.



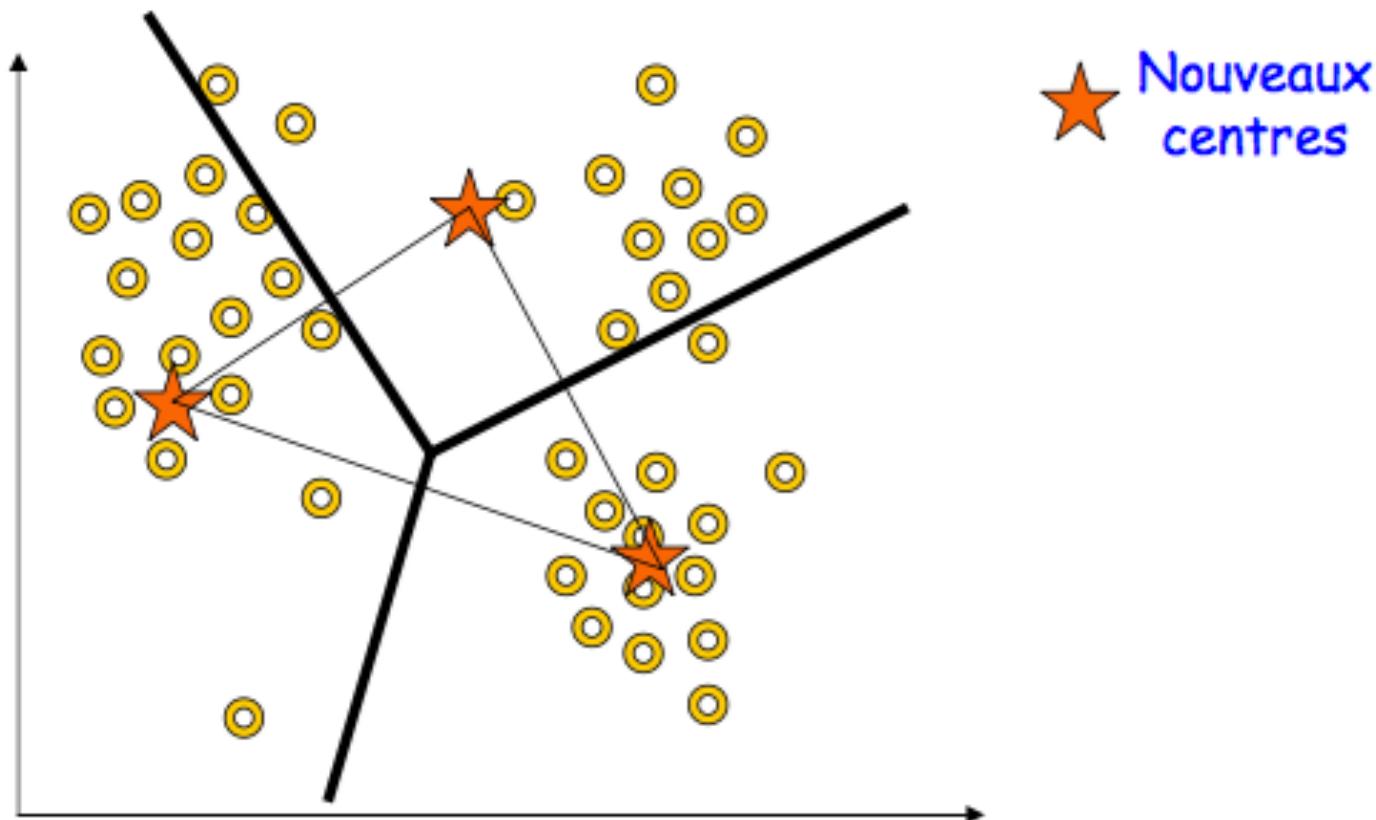
Algorithme des k-moyennes (k-means)

Illustration



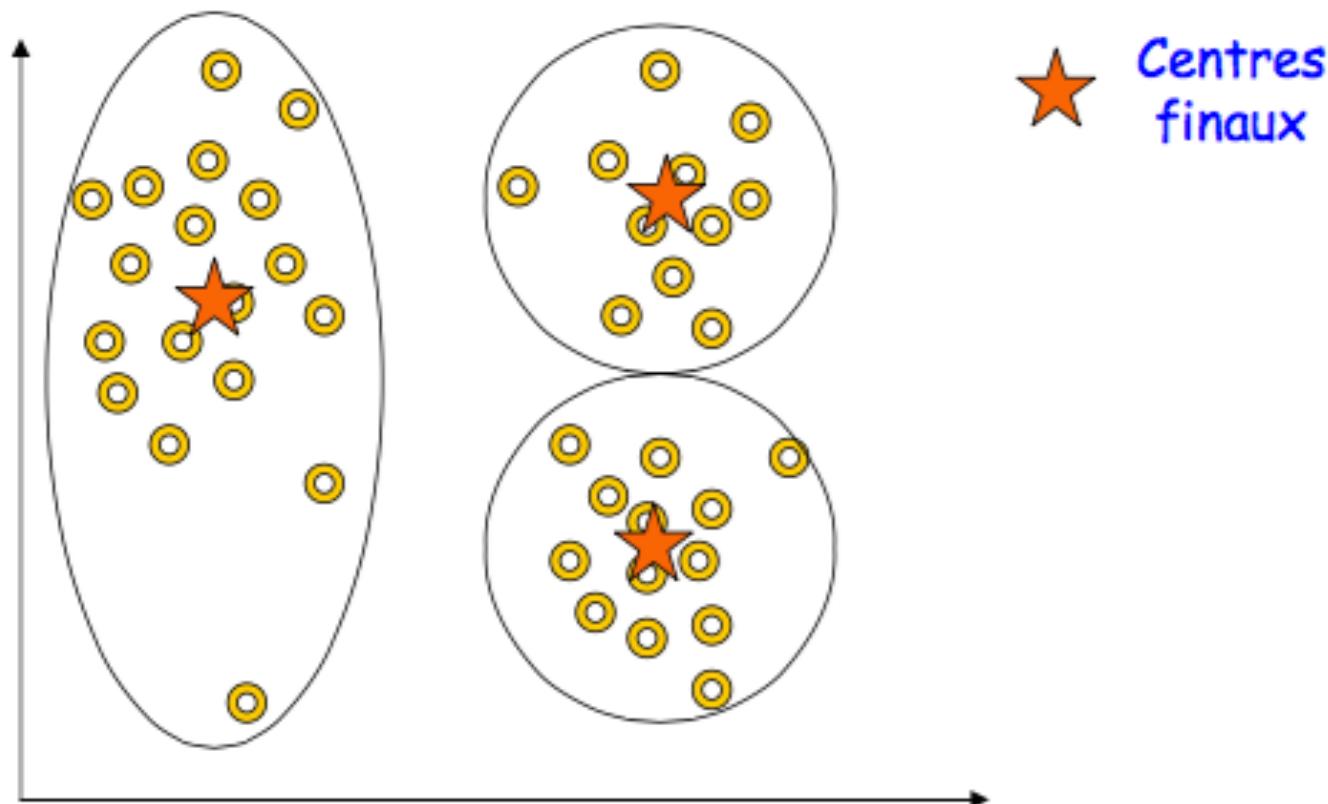
Algorithme des k-moyennes (k-means)

Illustration



Algorithme des k-moyennes (k-means)

Illustration



k-moyennes : avantages

- ▶ **Relativement extensible** dans le traitement d'ensembles de taille importante
- ▶ **Relativement efficace** : $O(t.k.n)$, où n représente # objets, k # clusters, et t # iterations. Normalement, $k, t \ll n$.
- ▶ Produit généralement un **optimum local** ; un optimum global peut être obtenu en utilisant d'autres techniques telles que : algorithmes génétiques, ...



k-moyennes : Inconvénients

- ▶ **Applicable** seulement dans le cas où la moyenne des objets est définie
- ▶ **Besoin de spécifier k**, le nombre de clusters, a priori
- ▶ **Incapable** de traiter les données bruitées (noisy).
- ▶ **Non adapté** pour découvrir des clusters avec structures non-convexes, et des clusters de tailles différentes
- ▶ Les **points isolés** sont mal gérés (doivent-ils appartenir obligatoirement à un cluster ?) - probabiliste



k-moyennes : Variantes

- ▶ Sélection des centres initiaux
- ▶ Calcul des similarités
- ▶ Calcul des centres (K-medoids)
- ▶ GMM : Variantes de K-moyennes basées sur les probabilités
- ▶ K-modes : données catégorielles
- ▶ K-prototype : données mixtes (numériques et catégorielles)



Clustering : Résumé

- ▶ Le clustering groupe des objets en se basant sur leurs similarités.
- ▶ Le clustering possède plusieurs applications.
- ▶ La mesure de similarité peut être calculée pour différents types de données.
- ▶ La sélection de la mesure de similarité dépend des données utilisées et le type de similarité recherchée.



Clustering : Résumé

- ▶ Les méthodes de clustering peuvent être classées en :
- ▶ Méthodes de partitionnement,
- ▶ Méthodes hiérarchiques, Méthodes à densité de voisinage.
- ▶ Plusieurs travaux de recherche sur le clustering en cours et en perspective.
- ▶ Plusieurs applications en perspective : Génomique, Environnement, ...





Classification

Classification

- ▶ Elle permet de **prédirer** si un élément est membre d'un groupe ou d'une catégorie donné.
- ▶ Classes
 - ▶ Identification de groupes avec des profils particuliers
 - ▶ Possibilité de décider de l'appartenance d'une entité à une classe
- ▶ Caractéristiques
 - ▶ Apprentissage supervisé : classes connues à l'avance
 - ▶ Pb : qualité de la classification (taux d'erreur)
 - ▶ Ex : établir un diagnostic (si erreur !!!)



Applications



- ▶ **Accord de crédit**
- ▶ **Diagnostic médical**
- ▶ **Analyse de l'effet d'un traitement**
- ▶ **Détection de fraudes fiscales**
- ▶ **etc.**



Processus à deux étapes

- ▶ **Etape 1 :**
 - ▶ **Construction du modèle** à partir de l'ensemble d'apprentissage (training set)

- ▶ **Etape 2 :**
 - ▶ **Utilisation du modèle** : tester la précision du modèle et l'utiliser dans la classification de nouvelles données



Construction du modèle

- ▶ **Chaque instance** est supposée appartenir à une classe prédéfinie
- ▶ La classe d'une instance est déterminée par l'attribut "classe"
- ▶ L'ensemble des instances d'apprentissage est utilisé dans la construction du modèle
- ▶ Le modèle est représenté par des règles de classification, arbres de décision, formules mathématiques, ...



Utilisation du modèle

- ▶ **Classification de nouvelles instances ou instances inconnues**
- ▶ **Estimer le taux d'erreur du**
 - ▶ la classe connue d'une instance test est comparée avec le résultat du modèle
 - ▶ Taux d'erreur = pourcentage de tests incorrectement classés par le modèle



Validation de la classification (accuracy)

- ▶ **Estimation des taux d'erreurs :**
- ▶ **Partitionnement** : apprentissage et test (ensemble de données important)
- ▶ Utiliser 2 ensembles indépendents, e.g., ensemble d'apprentissage (2/3), ensemble test (1/3)

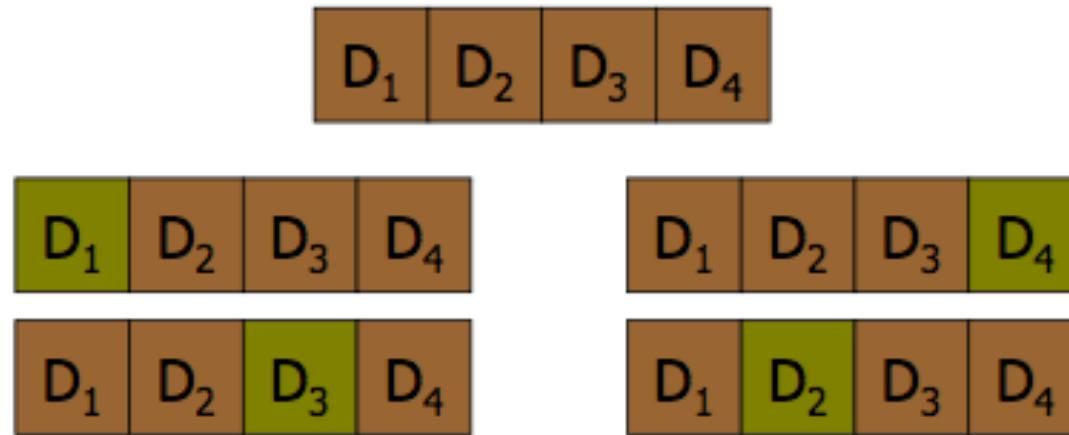
Apprentissage D_t

Validation $D \setminus D_t$



Validation de la classification (accuracy)

- ▶ **Validation croisée** (ensemble de données modéré)
- ▶ Diviser les données en k sous-ensembles
- ▶ Utiliser $k-1$ sous-ensembles comme données d'apprentissage et un sous-ensemble comme données test



- ▶ **Bootstrapping** : n instances test aléatoires (ensemble de données réduit)



Evaluation des méthodes de classification



- - ▶ **Taux d'erreur (Accuracy)**
 - ▶ **Temps d'exécution (construction, utilisation)**
 - ▶ **Robustesse (bruit, données manquantes,...)**
 - ▶ **Extensibilité**
 - ▶ **Interprétabilité**
 - ▶ **Simplicité**



Méthodes de classification



- ▶ Méthode K-NN (plus proche voisin)
- ▶ Arbres de décision
- ▶ Réseaux de neurones
- ▶ Classification bayésienne
- ▶ Caractéristiques
- ▶ Apprentissage supervisé (classes connues)



Méthodes des plus proches voisins

- ▶ Méthode dédiée à la classification (k-NN : nearest neighbor).
- ▶ Méthode de raisonnement à partir de cas : prendre des décisions en recherchant un ou des cas similaires déjà résolus.
- ▶ Pas d'étape d'apprentissage : construction d'un modèle à partir d'un échantillon d'apprentissage (réseaux de neurones, arbres de décision, ...).
- ▶ Modèle = échantillon d'apprentissage + fonction de distance + fonction de choix de la classe en fonction des classes des voisins les plus proches.



Algorithme KNN (K-nearest neighbors)

- ▶ Objectif : affecter une classe à une nouvelle instance
- ▶ donnée : un échantillon de m enregistrements classés (x , $c(x)$)
- ▶ entrée : un enregistrement y
 - ▶ 1. Déterminer les k plus proches enregistrements de y
 - ▶ 2. combiner les classes de ces k exemples en une classe c
- ▶ sortie : la classe de y est $c(y)=c$

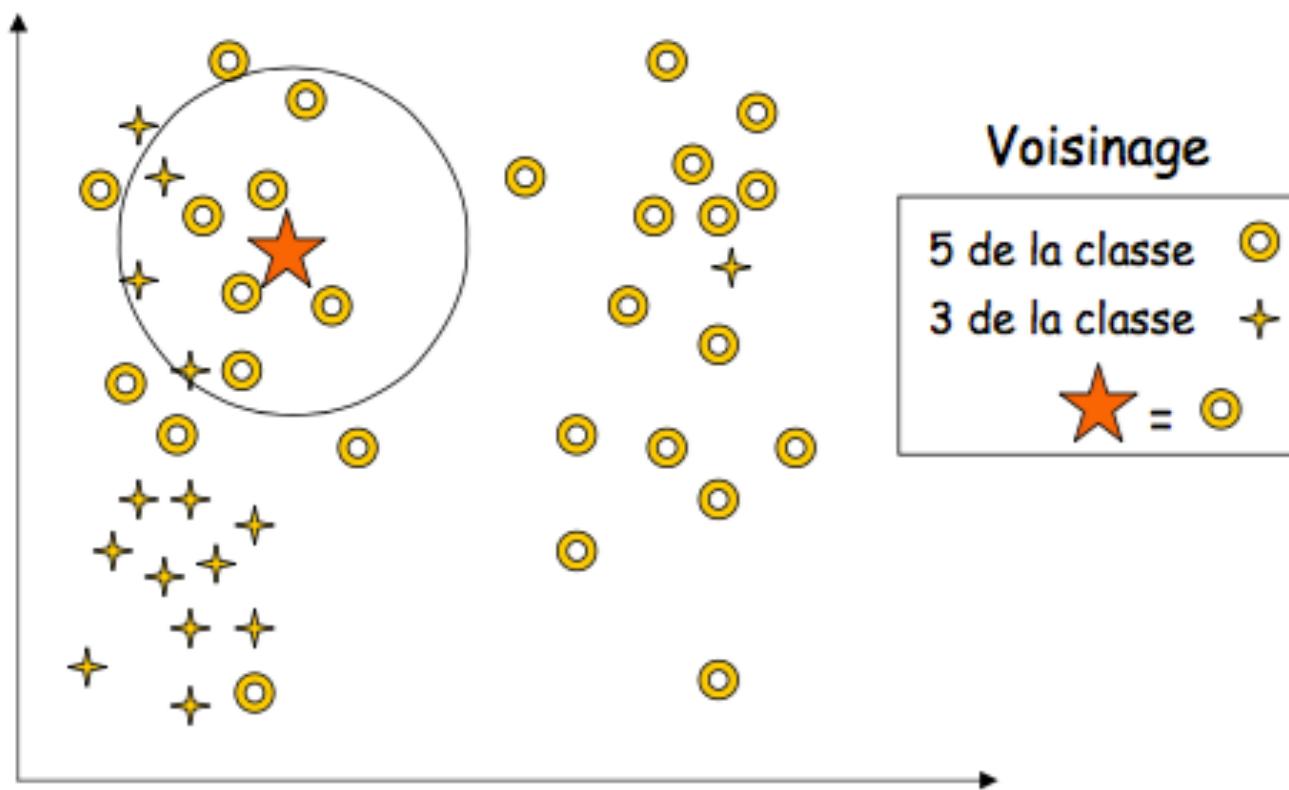


Algorithme KNN : sélection de la classe

- ▶ Solution simple : rechercher le cas le plus proche et prendre la même décision (Méthode 1-NN).
- ▶ Combinaison des k classes :
 - ▶ Heuristique : $k = \text{nombre d'attributs} + 1$
 - ▶ Vote majoritaire : prendre la classe majoritaire.
 - ▶ Vote majoritaire pondéré : chaque classe est pondérée. Le poids de $c(x_i)$ est inversement proportionnel à la distance $d(y, x_i)$.
- ▶ Confiance : Définir une confiance dans la classe attribuée = rapport entre les votes gagnants et le total des votes.



Illustration



Algorithme KNN : critique

- ▶ Pas d'apprentissage : introduction de nouvelles données ne nécessite pas la reconstruction du modèle.
- ▶ Clarté des résultats
- ▶ Tout type de données
- ▶ Nombre d'attributs
- ▶ Temps de classification : -
- ▶ Stocker le modèle : -
- ▶ Distance et nombre de voisins : dépend de la distance, du nombre de voisins et du mode de combinaison.



Arbres de décision

- ▶ **Génération d'arbres de décision à partir des données**
 - Arbre = Représentation graphique d'une procédure de classification**

- ▶ Un arbre de décision est un arbre où:
 - ▶ Nœud interne = un attribut
 - ▶ Branche d'un nœud = un test sur un attribut
 - ▶ Feuilles = classe donnée



Génération de l'arbre de décision

- ▶ **Deux phases dans la génération de l'arbre :**
- ▶ **Construction de l'arbre**
 - ▶ Arbre peut atteindre une taille élevée
- ▶ **Elaguer l'arbre (Pruning)**
 - ▶ Identifier et supprimer les branches qui représentent du “bruit”
 - ▶ Améliorer le taux d’erreur



Algorithmes de classification

- ▶ Construction de l’arbre
 - ▶ Au départ, toutes les instances d’apprentissage sont à la racine de l’arbre
 - ▶ Sélectionner un attribut et choisir un test de séparation (split) sur l’attribut, qui sépare le “mieux” les instances. La sélection des attributs est basée sur une heuristique ou une mesure statistique.
 - ▶ Partitionner les instances entre les noeuds fils suivant la satisfaction des tests logiques



Algorithmes de classification

- ▶ Traiter chaque noeud fils de façon récursive
- ▶ Répéter jusqu'à ce que tous les noeuds soient des terminaux. Un noeud courant est terminal si :
 - ▶ Il n'y a plus d'attributs disponibles
 - ▶ Le noeud est “pur”, i.e. toutes les instances appartiennent à une seule classe,
 - ▶ Le noeud est “presque pur”, i.e. la majorité des instances appartiennent à une seule classe (Ex : 95%)
 - ▶ Nombre minimum d'instances par branche (Ex : algorithme C5 évite la croissance de l'arbre, k=2 par défaut)
- ▶ Etiqueter le noeud terminal par la classe majoritaire



Algorithmes de classification

- ▶ Elaguer l'arbre obtenu (pruning)
 - ▶ Supprimer les sous-arbres qui n'améliorent pas l'erreur de la classification (accuracy) arbre ayant un meilleur pouvoir de généralisation, même si on augmente l'erreur sur l'ensemble d'apprentissage
 - ▶ Eviter le problème de sur-spécialisation (over-fitting), i.e., on a appris “par coeur” l'ensemble d'apprentissage, mais on n'est pas capable de généraliser



Sur-spécialisation – arbre de décision

- ▶ **L'arbre génér  peut sur- sp cialiser l'ensemble d'apprentissage**
 - ▶ Plusieurs branches
 - ▶ Taux d'erreur important pour les instances inconnues
- ▶ **Raisons de la sur-sp cialisation**
 - ▶ bruits et exceptions
 - ▶ Peu de donn e d'apprentissage
 - ▶ Maxima locaux dans la recherche gloutonne



Comment éviter l'overfitting ?

- ▶ **Deux approches :**
 - ▶ **Pré-élagage** : Arrêter de façon prématuée la construction de l'arbre
 - ▶ **Post-élagage** : Supprimer des branches de l'arbre complet (“fully grown”)
 - ▶ Convertir l'arbre en règles ; élaguer les règles de façon indépendante (C4.5)



Construction de l'arbre - synthèse

- ▶ Evaluation des différents branchements pour tous les attributs
- ▶ Sélection du “meilleur” branchement “et de l’attribut “gagnant”
- ▶ Partitionner les données entre les fils
- ▶ Construction en largeur (C4.5) ou en profondeur (SPLIT)
- ▶ Questions critiques :
 - ▶ Formulation des tests de branchement
 - ▶ Mesure de sélection des attributs



Algorithmes pour les arbres de décision

▶ **Algorithme de base**

- ▶ Construction récursive d'un arbre de manière “diviser-pour-régner” descendante
- ▶ Attributs considérés énumératifs
- ▶ Glouton (piégé par les optima locaux)

▶ **Plusieurs variantes : ID3, C4.5, CART, CHAID**

- ▶ Différence principale : mesure de sélection d'un attribut – critère de branchement (split)



Mesures de sélection d'attributs

- ▶ **Gain d'Information (ID3, C4.5)**
- ▶ **Indice Gini (CART)**
- ▶ **Table de contingence statistique χ^2 (CHAID)**
- ▶ **G-statistic**



Méthodes à base d'arbres de décision

- ▶ **CART** (BFO'80 - Classification and regression trees, variables numériques, Gini, Elagage ascendant)
- ▶ **C5** (Quinlan'93 - dernière version ID3 et C4.5, attributs d'arité quelconque, entropie et gain d'information)
- ▶ **SLIQ** (EDBT'96 — Mehta et al. IBM)
- ▶ **SPRINT** (VLDB'96—J. Shafer et al. IBM)
- ▶ **PUBLIC** (VLDB'98 — Rastogi & Shim)
- ▶ **RainForest** (VLDB'98 — Gehrke, Ramakrishnan & Ganti)
- ▶ **CHAID** (Chi-square Automation Interaction Detection – variables discrètes)



Arbres de décision - Avantages

- ▶ Compréhensible pour tout utilisateur (lisibilité du résultat – règles - arbre)
- ▶ Justification de la classification d'une instance (racine feuille)
- ▶ Tout type de données
- ▶ Robuste au bruit et aux valeurs manquantes
- ▶ Attributs apparaissent dans l'ordre de pertinence tâche de pré-traitement (sélection d'attributs)
- ▶ Classification rapide (parcours d'un chemin dans un arbre)
- ▶ Outils disponibles dans la plupart des environnements de data mining



Arbres de décision - Inconvénients

- ▶ Sensibles au nombre de classes : performances se dégradent
- ▶ Evolutivité dans le temps : si les données évoluent dans le temps, il est nécessaire de relancer la phase d'apprentissage





Réseaux de neurones

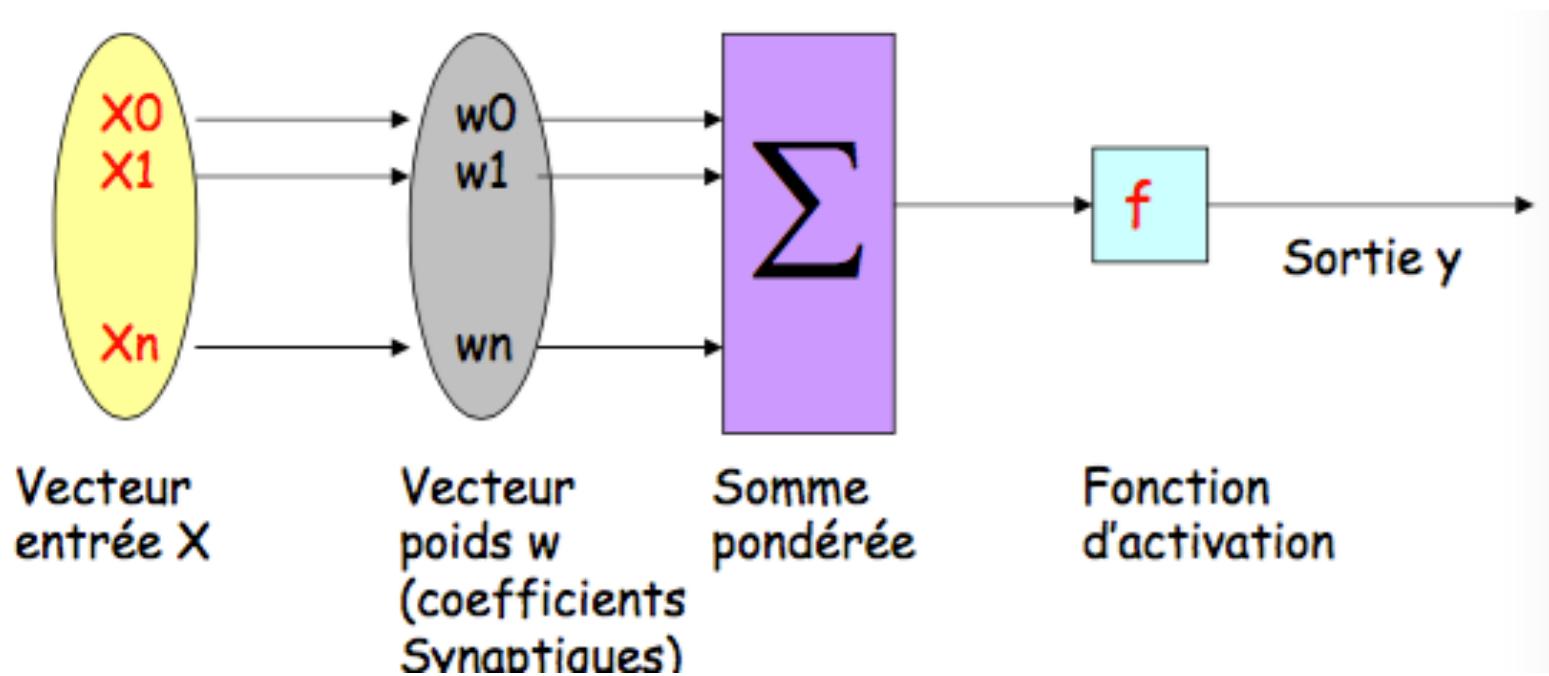
Réseaux de neurones

- ▶ Réseau neuronal : simule le système nerveux biologique
- ▶ Un réseau de neurones est composé de plusieurs neurones interconnectés. Un poids est associé à chaque arc. A chaque neurone on associe une valeur.
 - ▶ Temps de "switch" d'un neurone $> 10^{-3}$ secs
 - ▶ Nombre de neurones (humain) $\sim 10^{10}$
 - ▶ Connexions (synapses) par neurone : $\sim 10^4\text{--}10^5$



Neurone ou perceptron

- ▶ Neurone = Unité de calcul élémentaire
- ▶ Le vecteur d'entrée X est transformé en une variable de sortie y , par un produit scalaire et une fonction de transformation non linéaire



Neurone

- ▶ Fonction d'activation la plus utilisée est la fonction sigmoïde.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

- ▶ Elle prend ses valeurs (entrée et sortie) dans l'intervalle $[0,1]$

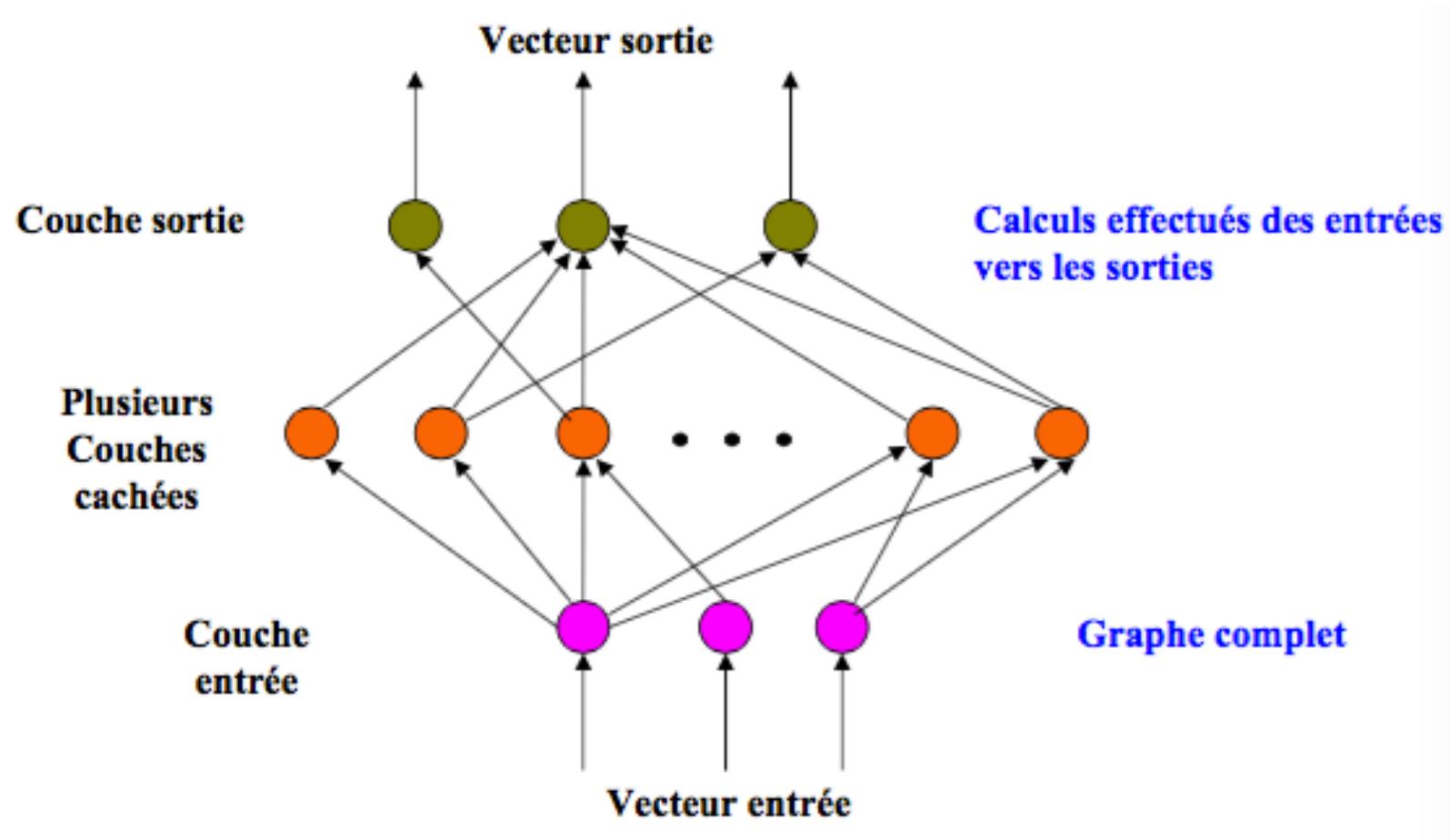


Réseaux de neurones

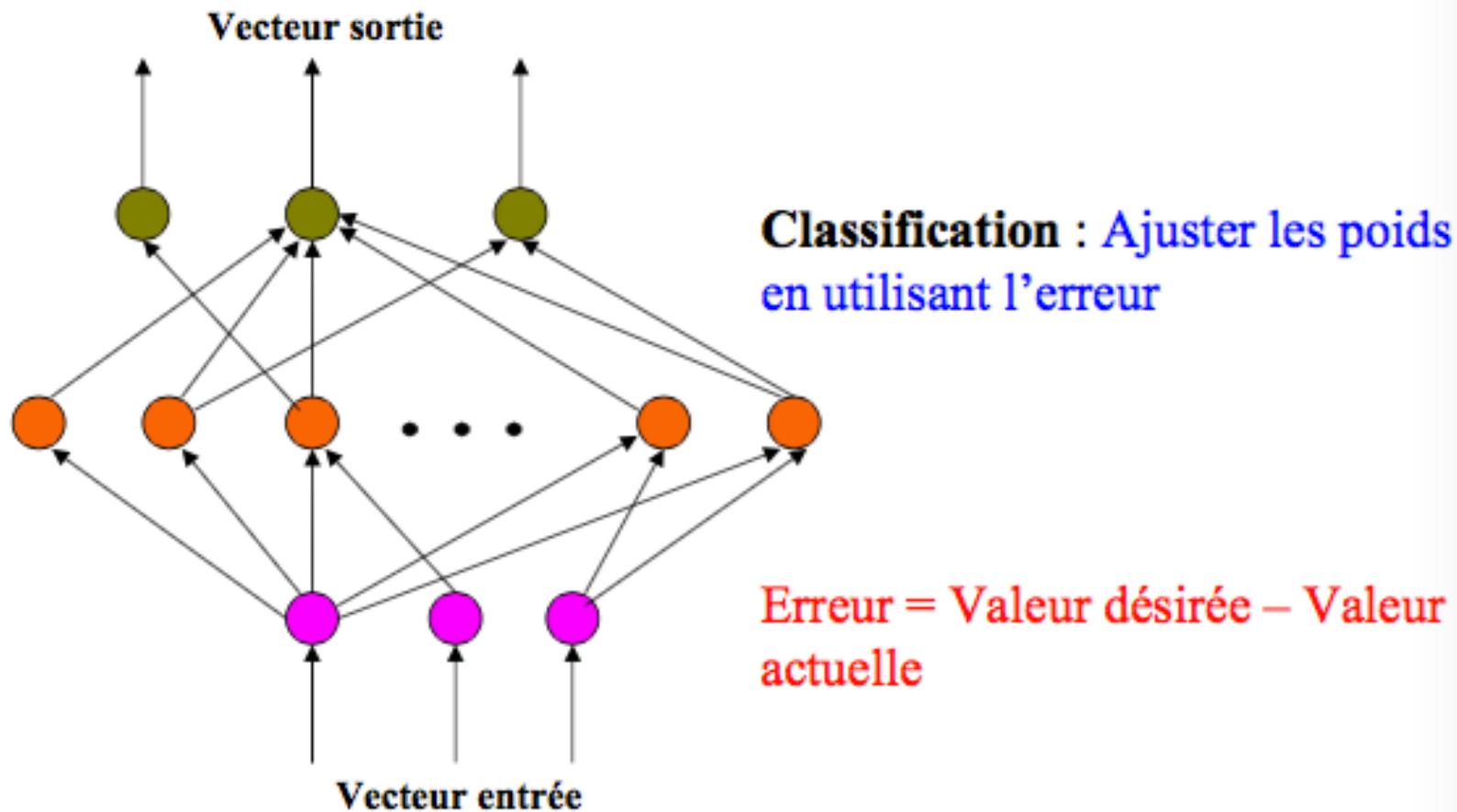
- ▶ Capacité d 'apprentissage : apprendre et changer son comportement en fonction de toute nouvelle expérience.
- ▶ Permettent de découvrir automatiquement des modèles complexes.
- ▶ Plusieurs modèles de réseaux de neurones : PMC (Perceptron Multi-Couches), RBF (Radial Basis Function), Kohonen, ...



Perceptron Multi Couches (PMC)



Paradigme d'apprentissage



Algorithmes d'apprentissage

- ▶ Rétro-propagation du gradient (Back propagation)
- ▶ Kohonen
- ▶ RBF (Radial basis function)
- ▶ Réseaux de neurones probabilistes
- ▶ ART (Adaptive resonance theory) ...



Réseaux de neurones - Avantages

- ▶ Taux d'erreur généralement bon
- ▶ Outil disponible dans les environnements de data mining
- ▶ Robustesse (bruit) – reconnaissance de formes (son, images sur une rétine, ...)
- ▶ Classification rapide (réseau étant construit)
- ▶ Combinaison avec d'autres méthodes (ex : arbre de décision pour sélection d'attributs)



Réseaux de neurones - Inconvénients

- ▶ Apprentissage très long
- ▶ Plusieurs paramètres (architecture, coefficients synaptiques, ...)
- ▶ Pouvoir explicatif faible (boite noire)
- ▶ Pas facile d'incorporer les connaissances du domaine.
- ▶ Traitent facilement les attributs numériques et binaires
- ▶ Evolutivité dans le temps (phase d'apprentissage)



Merci pour votre attentions !!!

**Questions ?
Remarques !**

Pr. Anass EL HADDADI

anass.elhaddadi@gmail.com