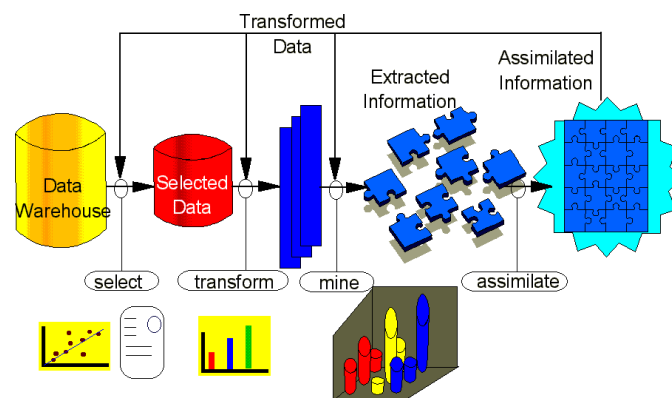


**Travaux de recherches:**

# **LE DATA MINING**



**Par :**

**PERIGNON Xavier**  
**SOH KAMLA Rodrigue**  
Elèves Ingénieur en 2eme année. Groupe 2

**Année 2008**  
**3iIA2**

## Sommaire

Introduction.....	3
I. Datamining : définition .....	4
II.Pratique du data mining.....	7
A Le processus de data mining.....	7
1.Poser le problème .....	7
2.La recherche des données .....	8
3.La sélection des données pertinentes .....	8
4.Le nettoyage des données .....	8
5.Les actions sur les variables .....	9
6.La recherche du modèle.....	9
7.L'évaluation du résultat.....	10
8.L'intégration de la connaissance .....	10
B Techniques de data mining.....	11
1.Le raisonnement à base de cas .....	12
2.Les Knowbots ou agents intelligents .....	12
II. Exploitation du data mining.....	14
a. Utilisations concrètes .....	14
b. Principaux avantages du Data mining.....	16
c) les défauts du data mining .....	16
IV . Cas pratique.....	17
Conclusion.....	23
Glossaire .....	24
Bibliographie.....	25

## Introduction

Dans le cadre de notre deuxième année d'études à l'école 3il, nous avons eu comme mission de développer un sujet en rapport avec l'informatique dans un dossier appelé « Travaux de recherches ». La réalisation d'un tel dossier permet de nous sensibiliser à une nouvelle technique ou technologie, et de développer notre capacité d'apprentissage d'une nouvelle notion.

Pour ce travail de recherche de deuxième année, nous avons choisi d'approfondir la notion de **data mining**. Le data mining est aussi connu sous le nom « d'exploration de données ». Jusqu'ici nous savons que le data mining est souvent utilisé pour définir le comportement type d'un consommateur en supermarché, par exemple. Mais qu'est-ce que c'est exactement ? Comment la met-on en œuvre ?

Afin de présenter au mieux la notion de data mining, nous avons divisé ce dossier en quatre grandes parties ; la première partie approfondira la notion de data mining, à travers quelques petits exemples et schémas. S'en suivra une liste explicative des différentes méthodes exploitées en data mining, avec l'utilité et les difficultés de chacune. Ensuite, nous présenterons concrètement quelles sont les utilisations possibles du data mining, avant enfin de présenter en détail un cas d'étude concret.

## I. Datamining : définition

Le Data mining est un sujet « brûlant ». Il dépasse aujourd'hui le cercle restreint de la communauté scientifique pour susciter un vif intérêt dans le monde des affaires. La littérature spécialisée et la presse ont pris le relais de cet intérêt et proposent de fait une pléthore de définitions. Parmi celles-ci, Dimitri Chorafas parle d'un processus permettant de « torturer l'information disponible jusqu'à ce qu'elle avoue ». Nous avons fait une synthèse de ces définitions, pour permettre d'en tirer une compréhension globale.

Le terme datamining (littéralement, « minage ou extractions de données ») désigne l'ensemble des algorithmes et méthodes destinés à l'exploration et l'analyse de (souvent) grandes bases de données informatiques en vue de détecter dans ces données, des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières, restituant de façon concise, l'essentiel de l'information utile pour l'aide à la décision.

L'analyse utilise des méthodes statistiques avancées, comme le partitionnement de données (rassemblement de données en paquets homogènes), et emploie régulièrement des mécanismes d'intelligence artificielle ou des réseaux neuronaux.

Le but du datamining est de découvrir des relations inconnues dans les données, spécialement quand les données proviennent de bases de données différentes. La découverte de ces relations peut permettre par exemple de réaliser des campagnes de publicité ciblées, ou de prédire comment la production se vendra... Les gouvernements utilisent aussi ces méthodes pour mettre à jour des activités illégales des particuliers, associations, ou autres gouvernements.

Ainsi, à partir de données stockées (le plus souvent stockées dans de grand entrepôts de données encore appelés Datawarehouse), et grâce aux algorithmes issus de domaines divers (bases de données, intelligence artificielle, statistiques), on peut tirer des solutions à des problèmes d'origines diverses. Ces données sont

Après avoir défini le datamining, il convient de préciser ce qui le différencie des domaines d'analyse connexes avec lesquels on pourrait quelques fois le confondre.

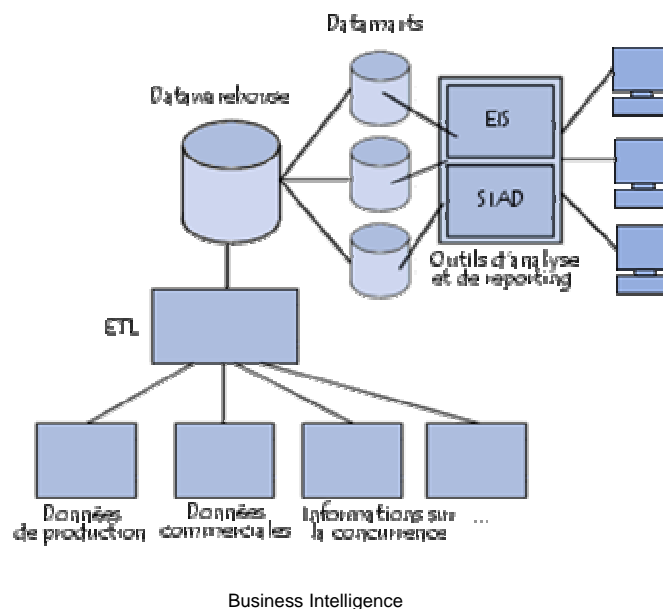
### Datamining vs statistiques

Contrairement à la méthode statistique, le Datamining ne nécessite jamais que l'on établisse une hypothèse de départ qu'il s'agira de vérifier. C'est des données elles-mêmes que se dégageront les corrélations intéressantes, le logiciel n'étant là que pour les découvrir (le Datamining se situe à la croisée des [statistiques](#), de l'[intelligence artificielle](#), des [bases de données](#)). Les programmes d'analyses sont lancés sur la base de données, sans objectifs du genre « trouver la corrélation entre telle et telle données ».

## Datamining vs. Informatique Décisionnelle (Business Intelligence)

L'informatique décisionnelle (... BI pour Business Intelligence) désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données d'une entreprise en vue d'offrir une aide à la décision, et de permettre aux responsables de la stratégie d'une entreprise d'avoir une vue d'ensemble de l'activité traitée :

- Sélectionner les données (par rapport à un sujet et/ou une période)
- Trier, regrouper ou répartir ces données selon certains critères
- Élaborer des calculs récapitulatifs « simples » (totaux, moyennes conditionnelles, etc.)
- Présenter les résultats de manière synthétique (graphique et/ou tableaux de bord)



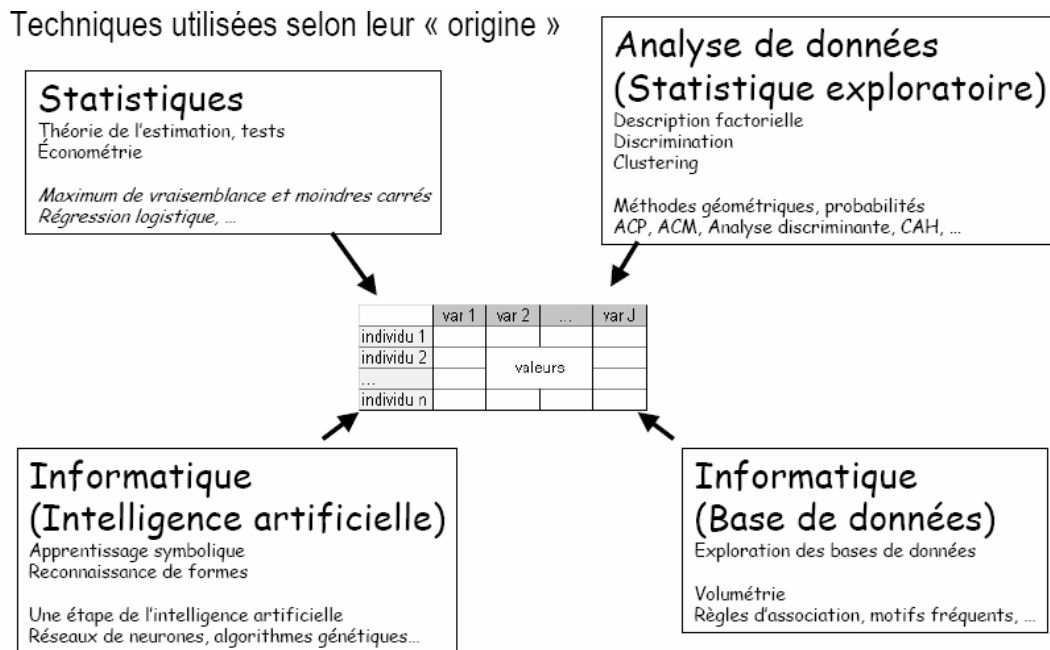
Le Datamining est proche de ce cadre, mais introduit une dimension supplémentaire qui est la modélisation « exploratoire » (détection des liens de cause à effet, validation de leur reproductibilité)

Le datamining transforme en quelques sortes, les données en « connaissances ».

Au dire d'experts (*Michel Bruley*), « ceux qui ont su voir plus loin y ont gagné un formidable avantage concurrentiel en utilisant le data mining pour résoudre des problèmes d'entreprise complexes et voir augmenter leur rentabilité. ». Citons en exemple d'utilisation de datamining, la mise en évidence par les magasins Wal-Mart d'une corrélation très forte entre l'achat de couches pour bébés et de bière le samedi après-midi. Les analystes s'aperçurent alors qu'il s'agissait des messieurs envoyés au magasin par leur dame pour acheter les volumineux paquets de couches pour

bébé. Les rayons furent donc réorganisés pour présenter côte à côte les couches et les packs de bière ... dont les ventes grimperent en flèche.

Comme l'illustre la figure suivante, Le data mining utilise des techniques provenant de disciplines diverses.



Très souvent, ces méthodes reviennent à optimiser les mêmes critères, mais avec des approches / formulations différentes

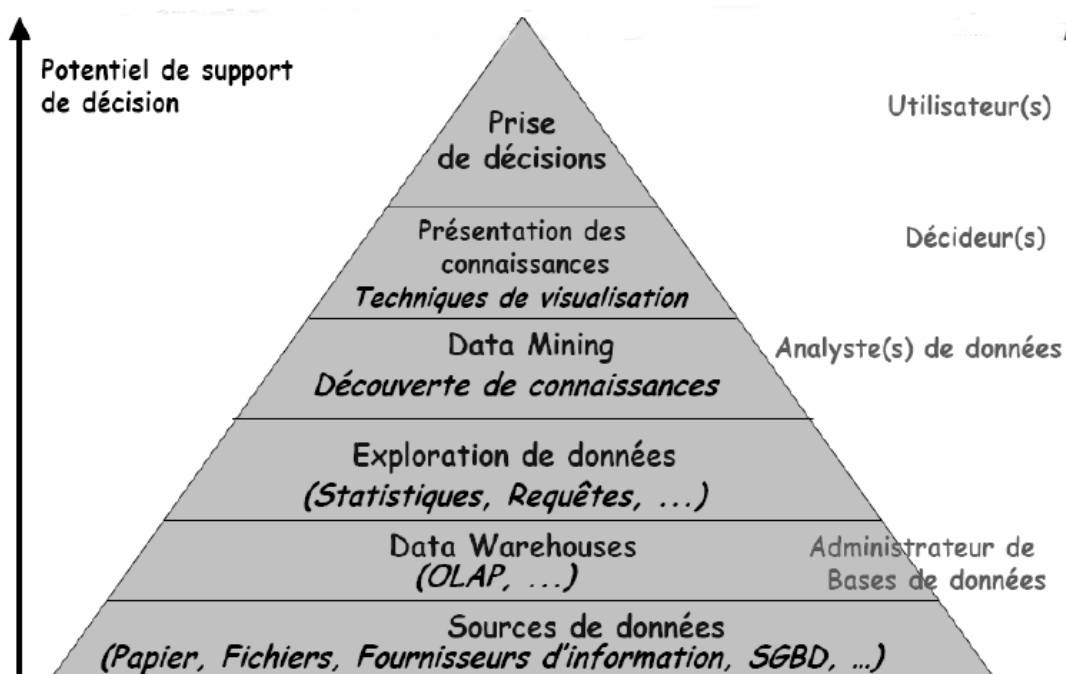
Nous détaillerons dans la suite de ce dossier ce qu'est vraiment ce processus d'extractions de données.

## II.Pratique du data mining

Le data mining utilise plusieurs autres disciplines. A la fin de tout le processus, le but est de prendre des décisions, après une analyse rationnelle. Plusieurs flux de données sont utilisées, et diverses techniques sont appliquées dans le but d'apporter au décideur, en partant de simples données pas très parlantes, des connaissances suffisantes pour effectuer des choix techniques, commerciaux, de gestions, ...

### A Le processus de data mining

L'organisation du flux d'informations et les acteurs



Nous avons regroupé les tâches à effectuer en plusieurs étapes, de la position du problème à l'intégration des connaissances.

### 1.Poser le problème

Dans première phase, on expose le problème et on définit les objectifs. Pour ce faire, on recueille les intuitions et les connaissances existantes des experts du domaine concerné, et on formule le problème à résoudre, tel qu'il sera possible de lui appliquer les techniques et outils de modélisation.

En suite, il faut connaître la typologie du problème (affectation ou structuration). Si à priori, on reconnaît l'appartenance des éléments à une ou plusieurs classes, il

s'agira de définir des facteurs d'affectation. Si l'objectif est plutôt de mettre en évidence des classes ou des facteurs de différenciation, on cherchera à identifier des facteurs de structuration.

Ayant défini le type de problème, on doit bien savoir ce qu'on attend comme résultat et ce l'exploitation qu'on en fera. Ces dernières connaissances faciliteront les choix à effectuer dans les étapes suivantes

## **2.La recherche des données**

Il s'agit dans cette phase de déterminer la structure générale des données ainsi que les règles utilisées pour les constituer. La sélection des données doit être optimale et peut nécessiter la consultation d'experts, afin de déterminer les attributs les aptes à décrire la problématique.

En suite, grâce à des taxinomies, il faudra réduire le nombre des variables obtenues pour faciliter une généralisation du problème. Cette étape peut fortement conditionner la qualité des résultats du processus de datamining.

## **3.La sélection des données pertinentes**

On effectue une collecte et une sélection de données. Il faut constituer une base d'informations qui permet de construire l'apprentissage, c'est à dire la construction de modèles en recherchant dans le passé des évènements similaires. Ce travail peut nécessiter l'intervention de toute une équipe et sera plus ou moins facilité selon les technologies en place dans l'entreprise (base de données ouverte, entrepôt de données exhaustif, ...). La sélection des données peut aboutir sur un échantillon ou une exhaustivité de données qui seront ensuite nettoyées.

## **4.Le nettoyage des données**

Pour pouvoir définir la taille de la base d'exemples, et choisir la manière de la constituer, il faut effectuer un diagnostic de qualité potentielle des données. La phase de nettoyage des données permet d'améliorer la qualité des données afin de minimiser l'effet d'anomalies telles que des erreurs de saisie, des champs nuls, des valeurs aberrantes.

Les modalités de contrôle de l'origine des données dépend de la taille de la base d'exemples (importante ou restreinte) et de son type d'alimentation (automatique ou manuelle).

- La recherche des valeurs aberrantes peut être effectuée en isolant les pics de certaines valeurs dans une distribution statistique, ou en utilisant d'autres méthodes comme la détermination de score.
- Les valeurs manquantes sont gérées, soit en excluant les enregistrements incomplets, en remplaçant les données manquantes, ou en les gérant via des algorithmes précis.
- Une analyse est effectuée pour déceler l'existence d'enregistrements totalement nuls. Elle permet d'en identifier les causes externes possibles (panne de capteurs, saut de lignes par l'agent de saisie, ...)

Pour obtenir un modèle performant et faciliter l'apprentissage, il faut améliorer la qualité des données par l'utilisation de bruits ou de processus flous. Ces opérations, tout comme les précédentes permettent d'obtenir des données fiables.

## 5. Les actions sur les variables

Maintenant que les variables sont pertinentes, et que les données sont fiables, il faut les transformer pour préparer le travail d'analyse. Il s'agit d'intervenir sur des variables pour qu'elles soient mieux exploitables par des outils de modélisation. Ces transformations peuvent être de plusieurs types :

### La transformation monovariable

Lorsqu'on veut améliorer une seule variable, on peut être emmené à modifier une unité de mesure, par normalisation ou transformation logarithmique. Il est aussi important de changer les dates en durées pour faciliter le travail de modélisation. Si on a affaire à des coordonnées géographiques, l'utilisation de géocodage ou de logiciels d'information géographique peut être nécessaire afin de rendre des coordonnées plus significatives.

### La transformation multivariable :

Elle concerne la combinaison de plusieurs variables élémentaires en une nouvelle variable agrégée. En effet, les données brutes sont parfois insuffisantes pour apporter un pouvoir prédictif à un modèle. Les types de transformation sont multiples. On peut utiliser les ratios, la fréquence, des tendances, les combinaisons linéaires, les combinaisons non linéaires, ...

## 6. La recherche du modèle

Après avoir obtenu des variables, on passe à la phase de modélisation. Elle consiste à extraire des données à partir d'un volume de données bruitées et à la présenter sous une forme synthétique. Elle est parfois décrite sous le terme de data mining. Elle repose sur une recherche exploratoire, c'est-à-dire dépourvue de préjugés concernant les relations entre les données.

La recherche du modèle se fait dans la phase d'apprentissage sur une base d'apprentissage qui doit être distincte de la base de tests (dont nous plus loin). La construction de ce modèle peut se faire de manière automatique et interactive. Sa performance dépend du choix d'algorithmes de calculs.

Parmi les techniques de modélisation utilisables, citons trois groupes :

- **La recherche des modèles à base d'équations**, où le décideur s'appuie sur une fonction plus ou moins complexe qui combine les variables ;
- **L'analyse logique** où la décomposition du problème en sous-ensembles successifs permet de construire un raisonnement structuré ;
- **Les techniques de projection** où la complexité initiale du problème est réduite grâce à la mise en évidence des facteurs principaux d'explication.

Quelque soit la précision du modèle, sa précision devra être vérifiée par une évaluation.

## 7.L'évaluation du résultat

L'évaluation du résultat permet d'estimer la qualité du modèle, c'est-à-dire sa capacité à déterminer correctement les valeurs qu'il est censé avoir apprises sur des cas nouveaux. Cette évaluation prend généralement une forme qualitative et une forme quantitative.

**L'évaluation qualitative** : permet d'illustrer le poids ou l'influence d'un facteur. Elle peut se faire sous une forme graphique. Dans ce cas, elle améliore la compréhension des résultats.

**L'évaluation quantitative** : utilise des techniques et notions telles que l'intervalle de confiance, pour fiabiliser les conclusions apportées sur des données futures.

## La validation par des tests

Après avoir construit un modèle, il est possible d'en tester la pertinence sur la base d'apprentissage. Cela étant, il faut éviter d'« apprendre » les données plutôt que le modèle. Par exemple, le fait d'oublier de brasser les données peut conduire à obtenir un modèle qui a appris que les 1000 premiers enregistrements appartiennent à la classe A et les 300 suivants à la classe B. Il faut donc brasser aléatoirement les données avant tout apprentissage, et prévoir une base de test distincte.

Pour valider le modèle, il vaudra mieux constituer une base de test ne servant qu'au test. De cette manière on vérifiera que le modèle est capable classer convenablement les données qu'il n'a jamais rencontrées. La stabilité entre les résultats observés sur le fichier d'apprentissage et le fichier test constitue la **capacité d'apprentissage**.

## 8.L'intégration de la connaissance

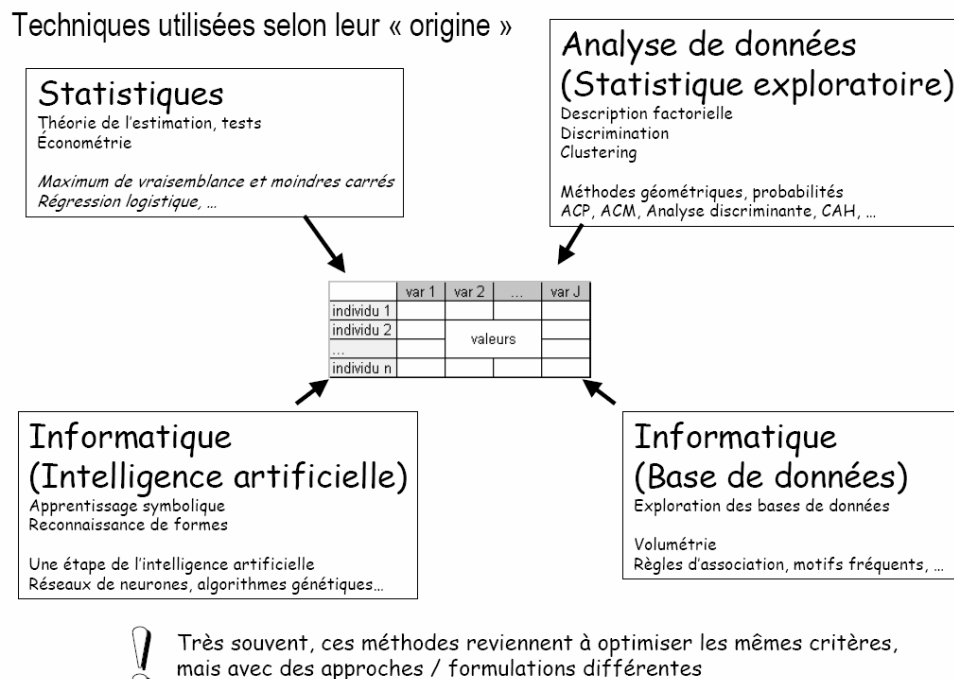
La connaissance n'est rien tant qu'elle n'est pas convertie en décision, puis en action. Il est essentiel d'implanter le modèle et ses résultats dans des systèmes informatiques ou dans les processus de l'entreprise. Cette intégration peut se faire soit sous la forme de données (résultat du modèle) ou sous la forme d'un traitement (algorithme du modèle).

C'est dans cette dernière phase qu'il faut dresser un bilan du déroulement des étapes précédentes. Ce bilan sert à améliorer l'existant en termes de données et de collecte de données.

Les étapes précédentes illustrent, les étapes à suivre pour faire du Data Mining. Ce pendant, comme nous l'avons vu à l'étape 6, pour construire un modèle, des techniques propres à la discipline de Data Mining sont utilisés.

## B Techniques de data mining

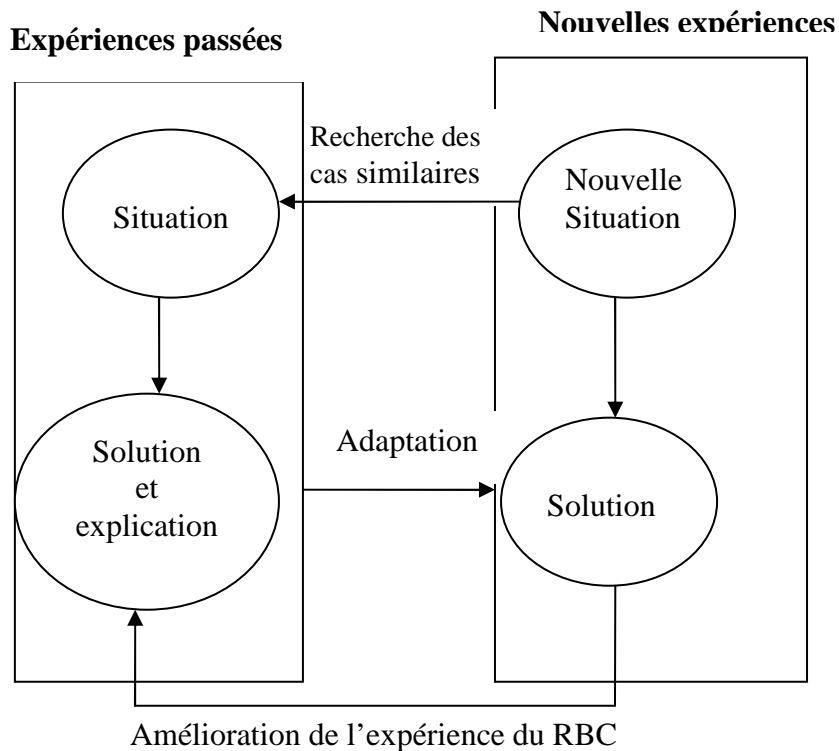
Les outils de Data Mining utilisent les mêmes fondements théoriques que les techniques statistiques traditionnelles. Ils s'appuient sur des techniques relativement similaires, mais représente une remarquable évolution par rapport à ces dernières. En effet, les outils de datamining bénéficient en outre de l'avancée des logiciels de bases de données, et des algorithmes d'apprentissage automatique ( intelligence artificielle)



Les outils de construction de modèles varieront selon la « dose » utilisée de chacune des disciplines ci-dessus. En tant qu'informaticiens, nous nous sommes intéressés à des outils qui s'appuient sur des données stockées.

## 1. Le raisonnement à base de cas

Les systèmes de raisonnement à base de cas (RBC ou CBR pour *Case Based Reasoning*, en anglais) résolvent des problèmes par comparaison d'exemples proches puisés dans un ensemble de cas stockés préalablement. Avec cette méthode de résolution, si une expérience passée et une nouvelle situation sont suffisamment similaires, toutes les conclusions appliquées à l'expérience passée restent valides et peuvent être appliquées à la nouvelle situation.



Les applications des systèmes RBC sont multiples ; la plupart des succès de cette technique concernent le service après vente ou le diagnostic de pannes, notamment sur les centres d'appel et les applications dites « embarquées ».

## 2. Les Knowbots ou agents intelligents

Le terme de Knowbot est une contraction de *Knowledge* et *Robot*. Ils désignent ce qui est connu en français sous le terme d'agents intelligents. Un agent est une entité abstraite qui est capable d'agir sur elle-même et sur son environnement. Il dispose d'une représentation partielle de cet environnement et peut communiquer avec d'autres agents. Avec le principe des agents, il est possible de réaliser des applications distribuées (sur plusieurs agents) afin de répartir un problème de Data Mining complexe en plusieurs objectifs. Pour assurer son fonctionnement, la structure centrale d'un agent contrôle son comportement général. Pour cela elle comporte une zone de contrôle, d'une zone de connaissance, et d'une zone de communication. Les Knowbots sont très utilisés pour la vente et le marketing sur Internet.

Il existe d'autres techniques que nous n'avons pas pu aborder ici. Notamment les techniques d'association et d'arbres de décisions.

## II. Exploitation du data mining

### a. Utilisations concrètes

Dans cette partie nous allons vous présenter quelques cas d'utilisation du data mining, car il est intéressant de bien comprendre dans quels cas le data mining est réellement utile, dans quels cas il peut être appliqué, et dans quels cas il sera inefficace.

#### - Etude du comportement des consommateurs

Afin de maximiser les ventes, beaucoup d'entreprises utilisent des solutions de data mining, afin de déterminer les habitudes des clients pour ensuite mieux cibler leur envies, et surtout en tirer le meilleur profit.

On peut citer l'exemple d'UNILEVER, qui suit à l'aide d'une solution data mining l'utilisation de carnets de coupons de réductions envoyés aux clients. Après cette étude, l'entreprise saura à grande échelle ce qui marche le mieux et le moins bien, et s'adaptera en conséquence.

Les Editions Atlas ont aussi fait appel à une solution de data mining pour analyser le taux d'adhésion suite à une offre du type « 60 fiches à 1,50€ ». En fonction des relances ou des autres offres envoyées aux personnes déjà intéressées par ces fiches, l'entreprise a pu éditer une méthode type permettant de maximiser le nombre de nouveaux abonnés.

#### - Etude de processus et de qualité

Le data mining est utilisé à d'autres fins que de faire du bénéfice, il est parfois utilisé pour étudier certains paramètres sociaux ou médicaux.

Une étude a été réalisée sur une population de femmes de 60 à 82 ans sur le vieillissement de la peau en fonction de plusieurs paramètres : exposition au soleil, tabagisme...) pour mettre en évidence les facteurs principaux de vieillissement et ainsi rendre plus efficace la prévention de ce vieillissement.

Sur le plan médical et psychologique, une étude a été réalisée dans des hôpitaux pour déterminer l'impact de « l'acharnement thérapeutique » comme de « l'arrêt thérapeutique » sur le personnel et ainsi anticiper la réaction de ce même personnel lors d'un cas similaire à l'avenir.

#### - Principaux logiciels de data mining

Nous présentons quelques logiciels d'aide à l'analyse data mining.

*Arbres de décision :*

- **AC2**, de Isoft : ce logiciel possède une interface conviviale, idéale pour débiter en data mining. Il intègre aussi des bibliothèques de programmation pour intégrer son utilisation dans un programme informatique. Une variante plus simple d'AC2 existe : Alice. Prix : environ 4000€.
- **Datamind**, de Datamind : ce logiciel existe en plusieurs versions, différentes pour un usage personnel, professionnel ou serveur. Il utilise un algorithme d'évaluation de probabilités propriétaire. Il existe sur

toutes les plateformes. A savoir que la version UNIX coûte dix fois plus cher que la version Windows.

- **Knowledge Seeker**, de Angoss : il est basé sur des techniques d'induction : il bâtit des arbres automatiquement, selon ses besoins. Il est bien adapté à une utilisation couplée au langage SQL. // coûte 8000€.

#### *Réseaux de neurones*

- **Predict** , de NeuralWare : plusieurs niveaux d'utilisation (débutant, avancé ou expert), il permet de toucher un grand nombre d'utilisateurs. Il s'utilise facilement avec Excel. 10000€.
- **4Thought**, de Cognos : Extrêmement paramétrable, il se présente sous forme de tableur, et peut générer un programme C à partir de ses paramètres pour une utilisation future. 20000€
- **Strata**, de Complex Systems : ce logiciel assemble des réseaux de neurones avec des algorithmes génétiques pour arriver à ses résultats. Il peut valider ses résultats avec des données sensiblement différentes, ce qui le rend particulièrement efficace. 5000€.

On s'aperçoit donc que le data mining est utile et utilisé dans un grand nombre de domaines, pour des problèmes variés, et que les outils pour l'exploiter sont nombreux.

Nous allons essayer de déterminer si c'est une bonne ou une mauvaise chose, à travers la présentation des avantages et inconvénients de cette méthode.

## ***b. Principaux avantages du Data mining***

Nous allons ici présenter quels sont les principaux avantages du data mining, et ainsi mieux cerner l'utilité de cette méthode.

- Le data mining aide à la prise de décision des dirigeants. Par l'analyse des données, la méthode peut « résumer » la situation et alors accélérer la prise de décision des dirigeants à un problème donné. Par contre, le data mining ne remplace pas ces dirigeants.
- Le data mining permet de faire des liens pertinents entre des données qui, à première vue, n'ont aucune corrélation.
- Cette méthode peut améliorer la satisfaction des clients en analysant leur besoins et en proposant des améliorations en fonction des événements passés.
- Permet d'effectuer des profils « type » : des profils de clients comme d'employés en fonction d'un test, pour prévoir l'évolution de ceux-ci aux cotés de l'entreprise.
- Le data mining facilite le développement de nouveaux produits
- Accélère la gestion des stocks, des inventaires, de la logistique
- Peut augmenter les revenus tout en diminuant les coûts.

C'est évident, le data mining a été étudié pour augmenter et optimiser le rendement d'une entreprise ou l'amélioration d'un critère. Cependant cette méthode souffre de quelques défauts qui seront détaillés ci-après.

## ***c) les défauts du data mining***

- Taille de la base : le data mining est totalement dépendant de la base de données qu'il analyse et donc des méthodes et technologies qui permettent d'y accéder ;
  - le stockage des données requiert de très grands espaces. Il doit souvent se faire sur une machine spécifique.
  - le temps de transfert des données entre la base de données et le poste de travail augmente la durée des traitements.
- Sujets d'analyse
  - Dans l'exemple d'analyse d'un site web, la structure de celui-ci est régulièrement modifiée et rend donc une analyse précise difficile.
- L'excès de confiance
  - Nous avons dit dans ce rapport que le data mining permettait de « prévoir l'avenir », d'anticiper la réaction des clients par rapport à une modification, une nouvelle campagne de publicité... c'est vrai, mais il ne faut pas pour autant suivre aveuglément le résultat d'une analyse data mining. C'est pour cela qu'il est prudent de recouper les informations obtenues avec d'autres études, statistiques et autres, et

de disposer d'un statisticien et d'un commercial pour l'analyse des résultats.

- La nécessité de disposer de personnel qualifié
  - Le data mining reste un processus complexe qui nécessite du personnel habitué à l'utiliser. Dans le cas minimum, il faudrait disposer d'un informaticien, d'un commercial connaissant bien la clientèle ainsi que d'un statisticien.

## IV . Cas pratique

Nous allons dans cette partie vous présenter une étude complète qui utilise le data mining. Une telle étude coûte cher, c'est pour cela qu'il est difficile d'en trouver le compte-rendu complet, avec les détails des calculs et autres informations complémentaires.

Nous avons choisi de présenter une étude réalisée par Cisia Ceresta, pour le compte du ministère des transports français. Cette étude a pour but de mettre en évidence les modes de transports des français dans et hors agglomérations, avec toute une panoplie de paramètres.

**Objectif de l'étude** : Etudier les déplacements courtes distances ( <80km à vol d'oiseau) des Français, à partir de chiffres relevés par l'INSEE entre 1982 et 1994.

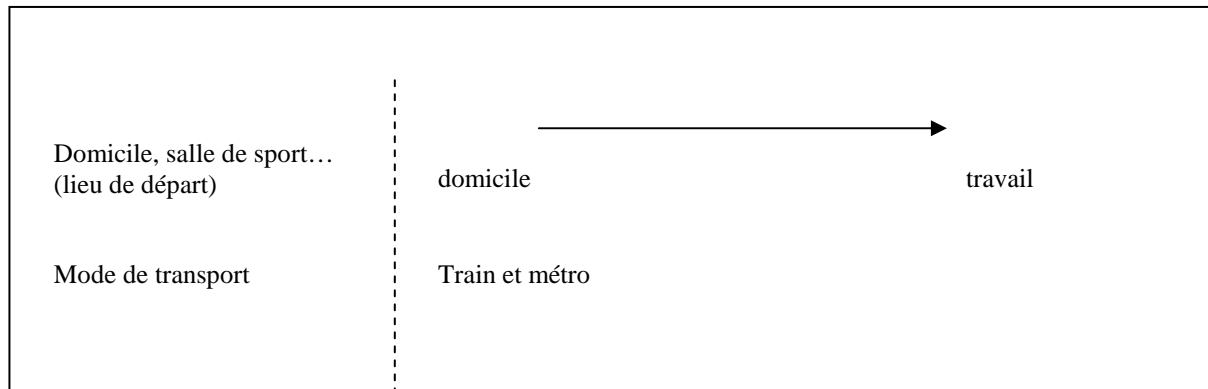
**Méthode** : La base de données utilisée pour l'étude est décomposée en deux grandes parties.

- La base de données qui contient les déplacements des personnes. Elle contient la durée et la distance de déplacement, le lieu (hors ou en agglomération), le type de véhicule emprunté, covoiturage ou non...
- La base de données des personnes, qui contient le sexe, le travail, l'âge, la situation familiale...

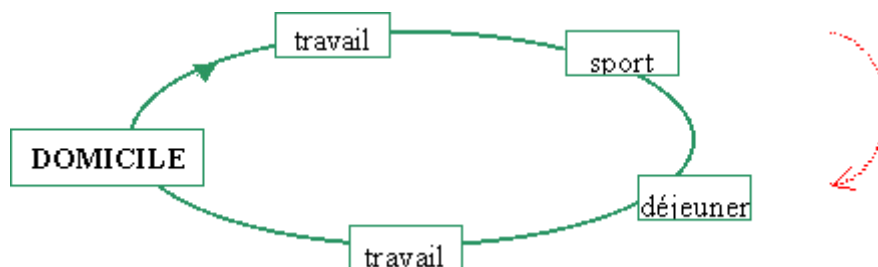
Les années d'étude sont analysées séparément, puis on recoupe les résultats en comparant les boucles de déplacement et le type de personnes qui les effectuent. Les pratiques de déplacements étant assez diverses, on réalise des typologies sur l'ensemble de la population mais aussi sur différentes sous-populations :

- les individus réalisant au moins trois boucles
- les individus très mobiles en terme de distance (+ de 55 km en 82 et + de 70 km en 94)
- les individus réalisant une ou deux boucles
- les individus "peu" mobiles en terme de distance (- de 55 km en 82 et - de 70 km en 94)

Dans une telle étude, il est essentiel de bien définir les éléments sur lesquels vont porter les analyses. Par exemple, quand on parlera de « déplacement », celui-ci devra pouvoir être représenté comme ceci :



De la même façon, on définit une boucle de déplacement comme ceci :



En gros, une boucle de déplacement est une série de déplacements ayant pour départ et fin le domicile. Les informations de la boucle sont : la distance moyenne, le temps moyen de transport, la durée d'absence du domicile, le nombre de fois où chaque mode de transport est utilisé dans la boucle, le nombre de fois où chaque motif de déplacement est utilisé dans la boucle, le nombre de modes différents utilisés, le nombre de motifs différents utilisés, etc....

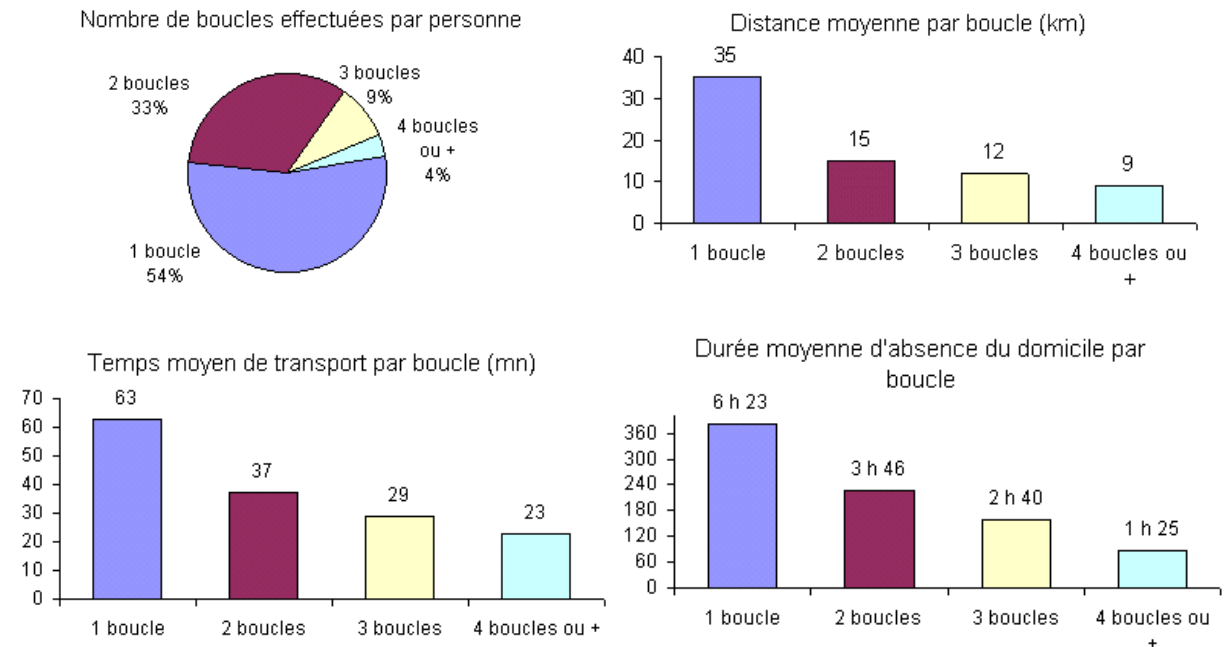
### Exemple de statistiques à traiter :

Nombre de boucles effectuées par jour par personne	de par la	Nombre d'individus concernés	Pourcentage d'individus concernés	Distance moyenne par boucle (en km)	Temps moyen de transport par boucle (en mn)	Durée d'absence du domicile par boucle
Une boucle		22 001 620	54%	35	63	6 h 23
Deux boucles		13 327 242	33%	15	37	3 h 46
Trois boucles		3 624 053	9%	12	29	2 h 40
Quatre boucles et plus		1 431 063	4%	9	23	1 h 25

Ensemble	40 383 977	100%	21	44	4 h 17
----------	------------	------	----	----	--------

Voici une représentation tabulaire de statistiques à traiter. On peut noter, dans le cadre de cette étude précise, l'important nombre de personnes prises en compte.

A partir de ces données, on établit les pré-résultats suivants :



## Résultats :

A la suite des analyses, les individus ont pu être divisés en six groupes :

### 1. LES INCONDITIONNELS DE LA VOITURE (42%)

Ce sont en majorité des hommes ayant une activité professionnelle qui utilisent la voiture en tant que conducteur. Les motifs de déplacements caractéristiques sont le lieu de travail, fixe ou non (on appelle "lieu de travail non fixe" le cas de tournée, de déplacements pour le travail, etc.), l'accompagnement (ex : le père qui dépose ses enfants à l'école avant de se rendre sur son lieu de travail) et les achats. Ce sont les personnes parcourant les plus grandes distances que l'on retrouve dans cette classe : 45,5 km en moyenne. Cette classe est la plus importante : 42% de la population utilise principalement sa voiture en tant que conducteur pour se déplacer.

### 2. LES UTILISATEURS DES TRANSPORTS EN COMMUN (10%)

56% des individus de cette classe résident en Ile-de-France. Les étudiants (ou élèves) et les personnes qui travaillent sont très présents, les motifs de déplacement caractéristiques sont le lieu d'études et le lieu de travail fixe. Ce sont des individus qui restent absents relativement longtemps de leur domicile (près de 9h en moyenne) et dont le temps de déplacement est long (1 h 46).

### 3. LES "SPORTIFS" A VELO (4%)

Les personnes se déplaçant en vélo sont en majorité des hommes, ce sont plutôt des étudiants (ou des élèves). Les motifs de déplacements principaux sont les visites, les loisirs ou le lieu d'études.

#### 4. LES ADEPTES DU DEUX-ROUES MOTORISE (1%)

On trouve principalement dans cette classe des jeunes étudiants ou élèves (plutôt des hommes), les motifs de déplacements dominants sont le lieu d'études et les visites. Seul 1% de la population se trouve dans cette classe.

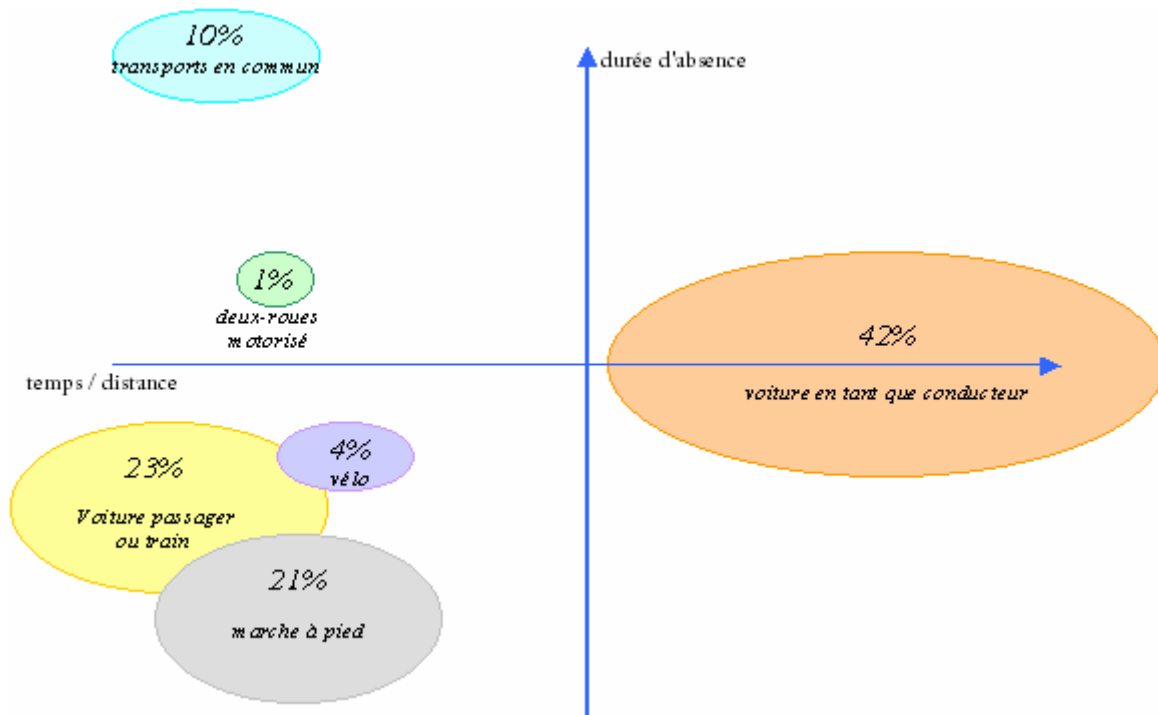
#### 5. LES "COURAGEUX" A PIED (21%)

La marche à pied est un mode de déplacement caractéristique pour 21% de la population : ce sont principalement des retraités, inactifs ou chômeurs et plutôt des femmes et/ou des citadins. La marche à pied est surtout utilisée pour se rendre sur un lieu d'études, faire des achats ou des démarches personnelles. La distance moyenne de l'ensemble des boucles de la journée est plus faible que dans les autres classes (5,2 km) et la durée moyenne d'absence du domicile est également la plus courte (moins de 5 heures).

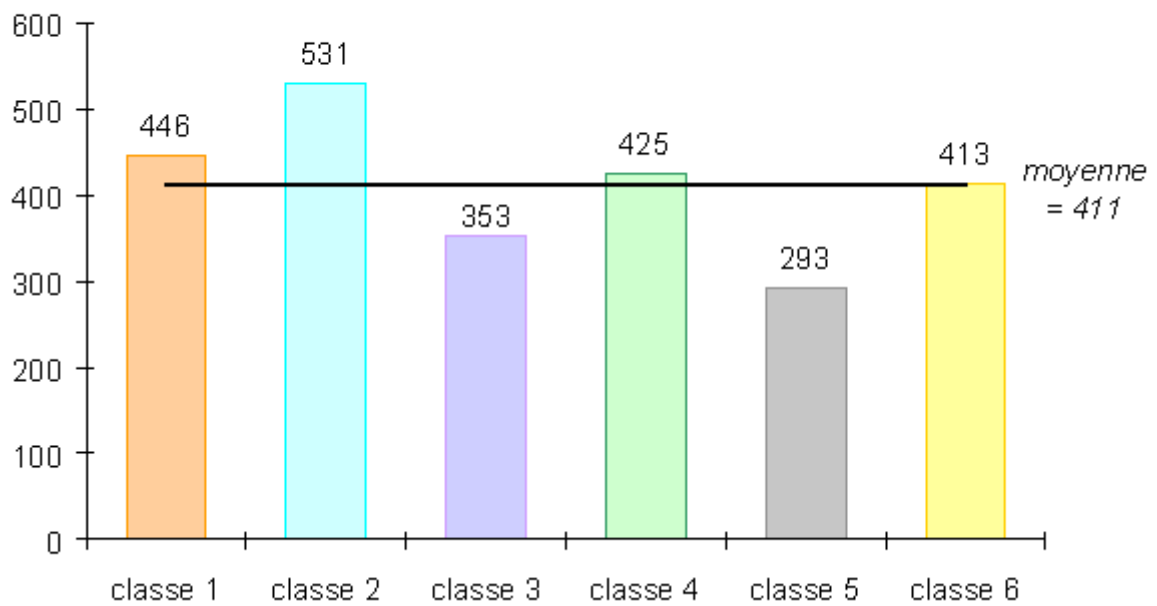
#### 6. LES INDIVIDUS A DOMINANTE "VOITURE PASSAGER" ET/OU "TRAIN" (23%)

Ce sont principalement des femmes, plutôt des étudiants (ou élèves) ou des inactifs. Les motifs de déplacements dominants sont le lieu d'études, les loisirs et les visites.

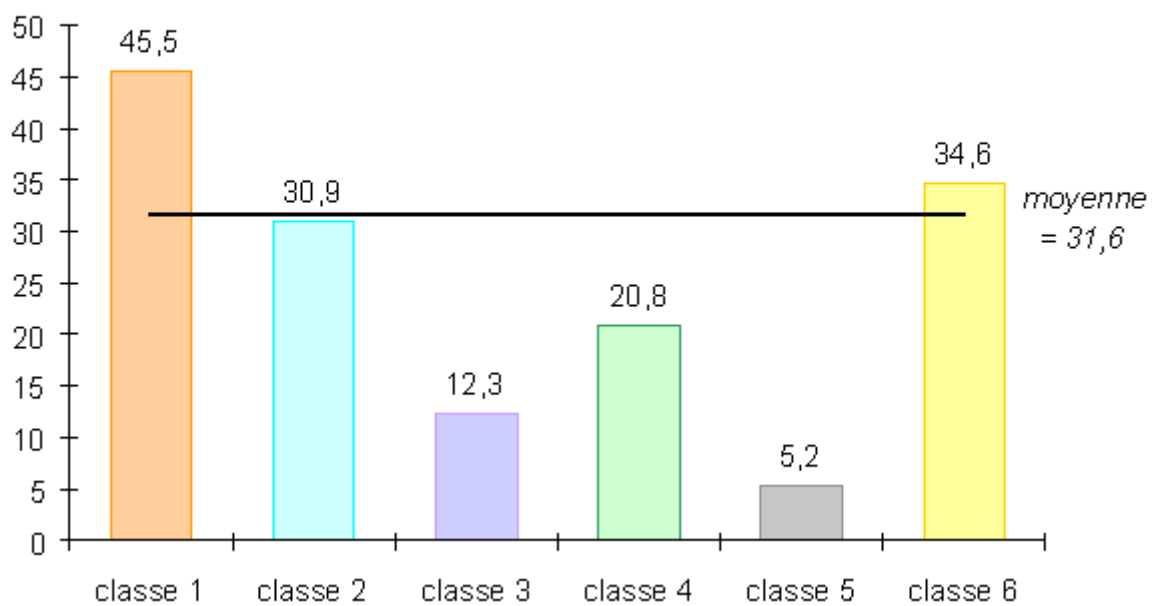
Nous pouvons présenter aussi les graphiques correspondant à cette étude.



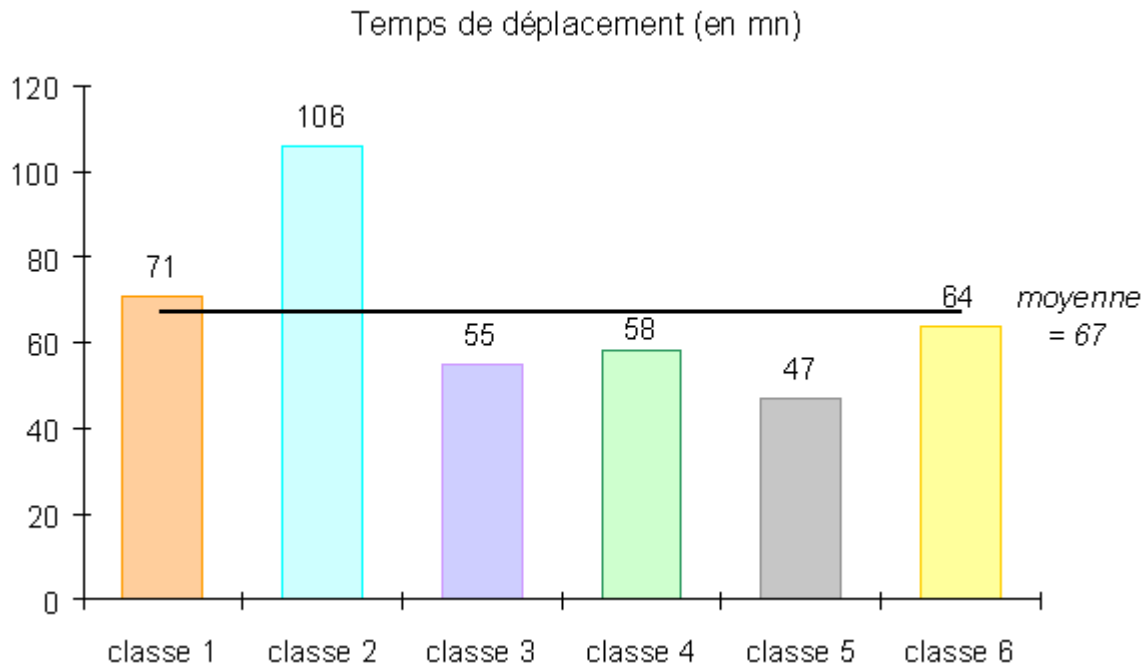
Durée de l'absence au domicile (en mn)



Distance parcourue (en km)



Sources : web-data mining.net



Nous pouvons conclure de cette étude qu'une majorité de personnes préfèrent encore se déplacer en voiture, bien que la distance parcourue soit la plus importante (et donc le coût aussi).

Il ressort aussi que la marche à pied est aussi un des meilleurs moyens de transports : l'absence du domicile est la plus courte, ainsi que le temps de déplacement et la distance parcourue. Il est aussi à noter que c'est le mode de déplacement le plus fatigant...

Une telle étude permet de connaître qui utilise quel transport, et quel sont les stratégies à adopter pour orienter un certain type de personnes vers un certain type de transport. Par exemple, on pourrait sensibiliser les hommes, qui utilisent majoritairement la voiture, et les encourager à se déplacer à pied pour des déplacements courts...

## Conclusion

La réalisation de ce dossier nous a permis de comprendre ce que c'est que le data mining, et surtout d'explorer des utilisations possibles. Nous nous sommes rendu compte que cette discipline est presque omniprésente dans la vie de tous les jours. Les cas les plus flagrants restent dans la grande distribution, où *via* les cartes de fidélités, les consommateurs sont à proprement parler fichés et suivis. Les habitudes sont tellement répertoriées et mises sous formes statistiques, puis utilisées pour amener le consommateur à toujours plus consommer, qu'il est permis de se demander si la spontanéité est encore de mise...

## Glossaire

**Datawarehouse** : Forme de système d'information conçu pour l'aide à la décision. Les données, issues des applications transactionnelles ou d'origine externe, sont mises sous une forme homogène et stockées sur des supports de grande capacité; elles sont organisées par sujet et s'accumulent continuellement (ce sont des données historiques). Le magasin de données joue le rôle d'une véritable mémoire de l'entreprise où les informations relatives aux événements significatifs sont conservées sous forme structurée. Différents outils d'extraction et de traitement permettent ensuite d'utiliser ces informations pour l'aide à la prise de décision. SYN.: entrepôt de données.

**Michel Bruley** est un expert en marketing B2B et a participé à l'installation de nombreux systèmes de CRM. Il a travaillé comme consultant pour plus d'une cinquantaine de groupes importants. Il a publié divers articles et livres blancs. Michel Bruley est le directeur marketing de Teradata France, division de NCR Corporation. Depuis 1975 il a travaillé comme consultant chez Bossard Consultants ou Ciba Geigy. Il est entré chez NCR en 1993 et dirige le Marketing de Teradata depuis 1997.

**Taxinomie** : méthodes de classification des données

**Score** : c'est une note calculable à partir d'une équation : la formule de score. La détermination de l'équation se fait en utilisant des techniques statistiques dites de scoring.

**La normalisation** : Elle permet d'avoir des ordres de grandeur comparables pour chaque variable. Elle consiste à soustraire à chaque valeur la moyenne sur l'échantillon et diviser cette différence par l'écart-type constaté sur chaque échantillon.

**Géocodage** : C'est une technique de géomarketing qui transforme les adresses ou des éléments d'adresse en coordonnées géographiques. Ces coordonnées peuvent servir à positionner des points sur une carte, mais aussi en Data mining, à calculer les distances relatives entre des points comme un magasin et un porteur de carte de fidélité.

**Brassage** : mélange des données de manière aléatoire de façon à faire perdre toute signification à l'ordre dans lequel elles sont présentées aux outils d'apprentissage.

**Capacité d'apprentissage** : c'est une mesure de performance du modèle. Elle est calculée en comparant le modèle à des données nouvelles et en comparant les résultats du modèle à aux valeurs réelles.

## Bibliographie

### Web :

<http://web-datamining.net>

<http://www.wikipedia.org>

<http://data.mining.free.fr/>

<http://www.poste.ch/fr/>

<http://spadsoft.com>

<http://eric.univ-lyon2.fr/~ricco/data-mining/>

<http://creg.ac-versailles.fr/>

<http://www.zdnet.fr/blogs/2005/11/27/data-mining/>

<http://britannica.com>

<http://www.guideinformatique.com/>

<http://www.commentcamarche.net/>

### Livres:

“Le Data mining”, de René Lefébure et Gilles Venturi, Editions Eyrolles