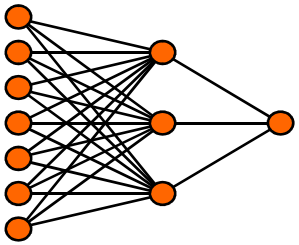
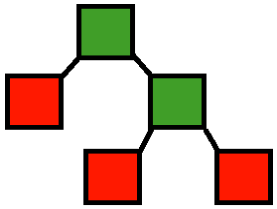
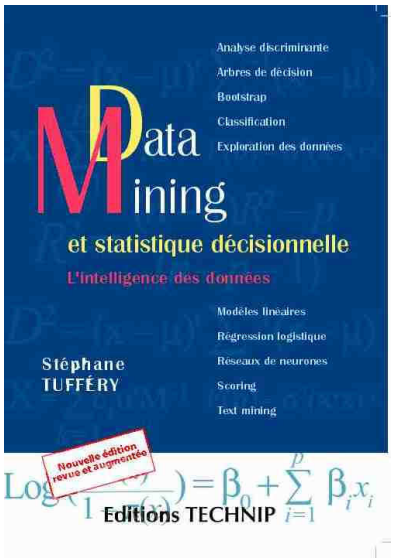
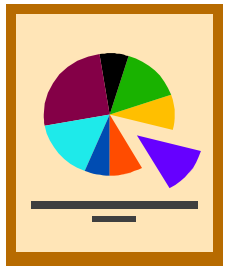
Stéphane Tufféry

DATA MINING
& STATISTIQUE DÉCISIONNELLE









Présentation de l'auteur

- En charge de la statistique et du *data mining* dans un grand groupe bancaire
- Enseigne le *data mining* en Master 2 à l'Université Paris-Dauphine, à l'Université Rennes 1 et à l'ISUP (Université Paris 6)
- Docteur en Mathématiques
- Auteur de :
  - *Data Mining et Scoring* (épuisé), Éditions Dunod, 2002
  - *Data Mining et Statistique Décisionnelle*, Éditions Technip, 2005, 2de édition 2007, préface de Gilbert Saporta

    Ouvrage consacré à l'application en entreprise des techniques et méthodologies de data mining et statistique

Présentation du cours

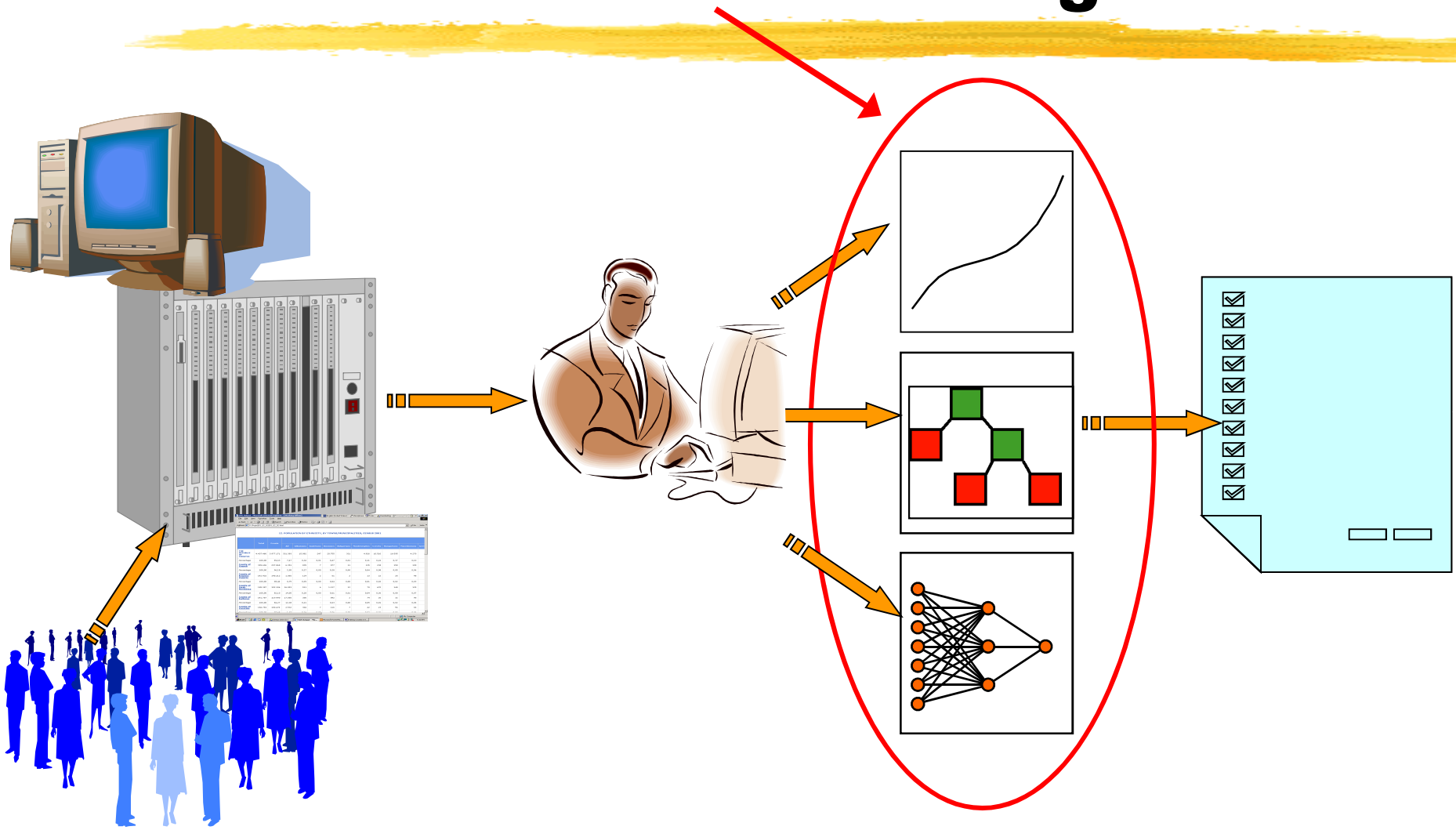
- Cette présentation est issue de cours donnés dans des DESS et Master 2 d'Économétrie et d'Ingénierie Statistique entre 1999 et 2009.
- Ces enseignements ont ensuite trouvé un développement dans des ouvrages publiés chez Dunod puis chez Technip.
- Ces cours sont donc consacrés aux techniques de data mining, de statistique décisionnelle et de scoring, et à leur mise en oeuvre en entreprise. Ils contiennent une introduction, une partie technique (préparation des données, analyse factorielle, régression linéaire, régression logistique, GLM, analyse discriminante, arbres de décision, réseaux de neurones, algorithmes génétiques, SVM, k-means et centres mobiles, CAH…) et une partie méthodologique (conduite de projet, facteurs de succès, RSI, aspects informatiques, CNIL…).

Plan du cours

- ***Qu'est-ce que le data mining ?***
- ***A quoi sert le data mining ?***
- ***Les deux grandes familles de techniques***
- **Le déroulement d'un projet de data mining**
- **Coûts et gains du data mining**
- **Facteurs de succès - Erreurs - Consulting**
- **Informatique décisionnelle et de gestion**
- **La préparation des données**
- **Techniques descriptives de data mining**
- **Techniques prédictives de data mining**
- **Logiciels de statistique et de data mining**
- **CNIL et limites légales du data mining**
- **Le text mining**
- **Le web mining**

Qu’est-ce que le data mining ?

Place du data mining

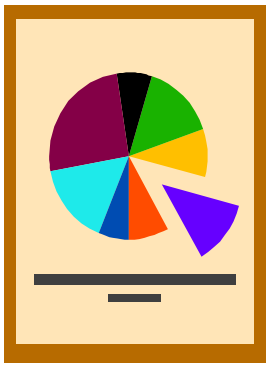
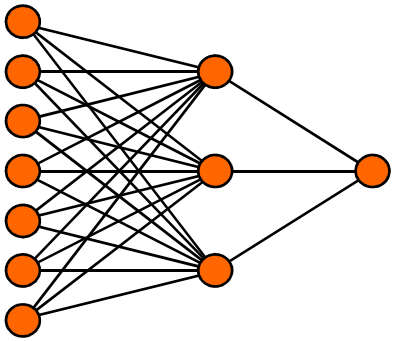
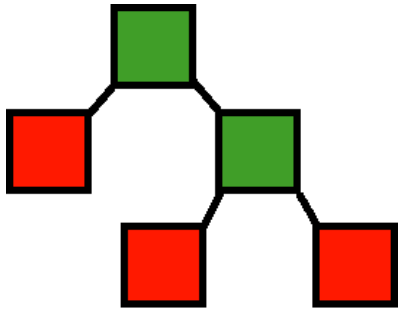


La fouille de données

- Le **data mining** est 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

Data mining ≠ statistiques descriptives

- Les techniques de data mining sont bien sûr plus complexes que de simples statistiques descriptives :
  - outils d'intelligence artificielle (réseaux de neurones)
  - algorithmes sophistiqués (algorithmes génétiques, analyse relationnelle)
  - théorie de l'information (arbres de décision)
  - **beaucoup d'analyse des données « traditionnelle »** (analyse factorielle, classification, analyse discriminante, etc.)

Data mining et statistique 1/2

- Hier :
  - études de laboratoire
  - expérimentations cliniques
  - actuariat
  - analyses de risque - scoring
- Volumes de données limités
- Analyse du réel pour mieux le comprendre :
  - les 1ères observations permettent de formuler des hypothèses théoriques que l'on confirme ou infirme à l'aide de tests statistiques

Data mining et statistique 2/2

- Aujourd'hui :
  - de l'∞ petit (génomique) à l'∞ grand (astrophysique)
  - du plus quotidien (reconnaissance de l'écriture manuscrite sur les enveloppes) au moins quotidien (aide au pilotage aéronautique)
  - du plus ouvert (e-commerce) au plus sécuritaire (détection de la fraude dans la téléphonie mobile ou les cartes bancaires)
  - du plus industriel (contrôle qualité…) au plus théorique (sciences humaines, biologie…)
  - du plus alimentaire (agronomie et agroalimentaire) au plus divertissant (prévisions d'audience TV)
- Volumes de données importants
- Systèmes d'aide à la décision plus ou moins automatiques

Des statistiques ...

- Statistique :
  - quelques centaines d'individus
  - quelques variables recueillies avec un protocole spécial (échantillonnage, plan d'expérience...)
  - fortes hypothèses sur les lois statistiques suivies
  - les modèles sont issus de la théorie et confrontés aux données
  - méthodes probabilistes et statistiques
  - utilisation en laboratoire

- Analyse des données :
  - quelques dizaines de milliers d'individus
  - quelques dizaines de variables
  - construction des tableaux « Individus x Variables »
  - importance du calcul et de la représentation visuelle

… au Data mining

- Data mining :
  - plusieurs millions d'individus
  - plusieurs centaines de variables
  - nombreuses variables non numériques, parfois textuelles
  - données recueillies avant l'étude, et souvent à d'autres fins
  - données imparfaites, avec des erreurs de saisie, de codification, des valeurs manquantes, aberrantes
  - population constamment évolutive (difficulté d'échantillonner)
  - nécessité de calculs rapides, parfois en temps réel
  - on ne recherche pas toujours l'optimum mathématique, mais le modèle le plus facile à appréhender par des utilisateurs non-statisticiens
  - faibles hypothèses sur les lois statistiques suivies
  - les modèles sont issus des données et on en tire des éléments théoriques
  - méthodes statistiques, d'intelligence artificielle et de théorie de l'apprentissage (« machine learning »)
  - utilisation en entreprise

Préhistoire

- 1875 : régression linéaire de Francis Galton
- 1896 : formule du coefficient de corrélation de Karl Pearson
- 1900 : distribution du χ^2 de Karl Pearson
- 1936 : analyse discriminante de Fisher et Mahalanobis
- 1941 : analyse factorielle des correspondances de Guttman
- 1943 : réseaux de neurones de Mc Culloch et Pitts
- 1944 : régression logistique de Joseph Berkson
- 1958 : perceptron de Rosenblatt
- 1962 : analyse des correspondances de J.-P. Benzécri
- 1964 : arbre de décision AID de J.P.Sonquist et J.-A.Morgan
- 1965 : méthode des centres mobiles de E. W. Forgy
- 1967 : méthode des k-means de Mac Queen
- 1972 : modèle linéaire généralisé de Nelder et Wedderburn

Histoire

- 1975 : algorithmes génétiques de Holland
- 1975 : méthode de classement DISQUAL de Gilbert Saporta
- 1980 : arbre de décision CHAID de KASS
- 1983 : régression PLS de Herman et Svante Wold
- 1984 : arbre CART de Breiman, Friedman, Olshen, Stone
- 1986 : perceptron multicouches de Rumelhart et McClelland
- 1989 : réseaux de T. Kohonen (cartes auto-adaptatives)
- vers 1990 : apparition du concept de data mining
- 1993 : arbre C4.5 de J. Ross Quinlan
- 1996 : bagging (Breiman) et boosting (Freund-Shapire)
- 1998 : support vector machines de Vladimir Vapnik
- 2000 : régression logistique PLS de Michel Tenenhaus
- 2001 : forêts aléatoires de L. Breiman

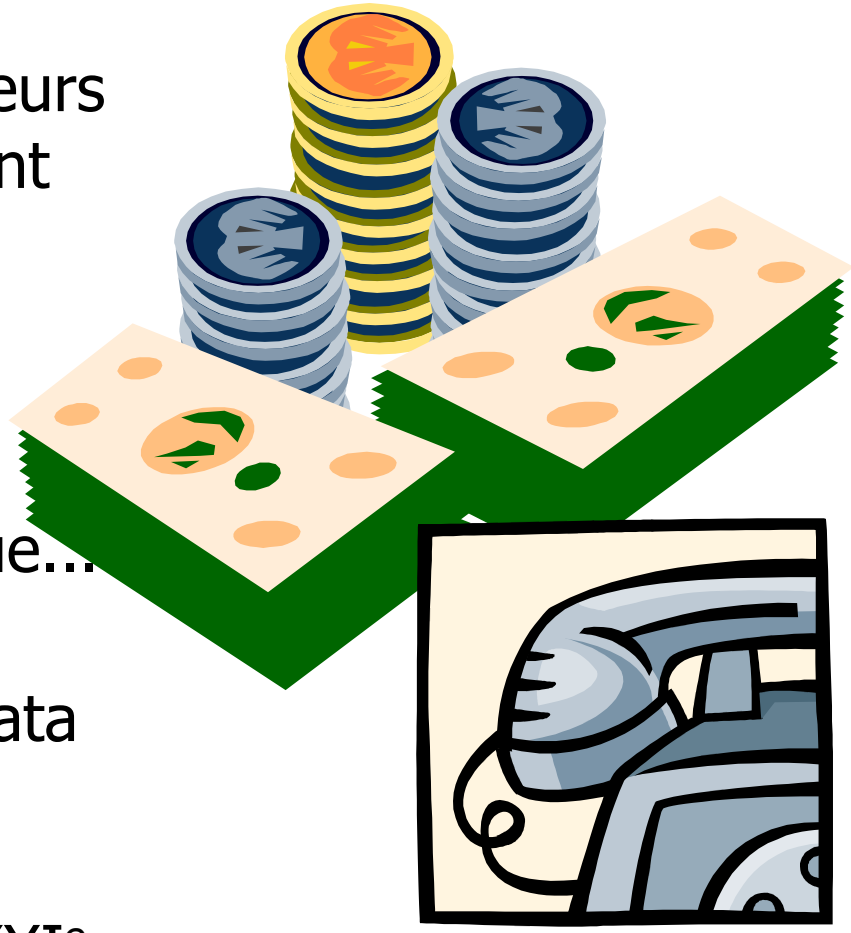
Le data mining aujourd'hui

- Ces techniques ne sont pas toutes récentes
- Ce qui est nouveau, ce sont aussi :
  - la recherche en IA et en théorie de l'apprentissage
  - les capacités de stockage et de calcul offertes par le matériel et les techniques informatiques modernes
  - la constitution de giga-bases de données pour les besoins de gestion des entreprises
  - les logiciels universels, puissants et conviviaux
  - l'intégration du data mining dans les processus de production

➡ qui permettent de traiter de grands volumes de données et font sortir le data mining des laboratoires de recherche pour entrer dans les entreprises

Le data mining aujourd'hui

- Le data mining se répand particulièrement dans les secteurs qui, par leur activité, détiennent de nombreuses informations économiques et comportementales individualisées : VPC, grande distribution, téléphonie, banque...
- Selon le MIT (Massachussets Institute of Technology) : le data mining est l'une des 10 technologies émergentes qui « changeront le monde » au XXI^{e} siècle.

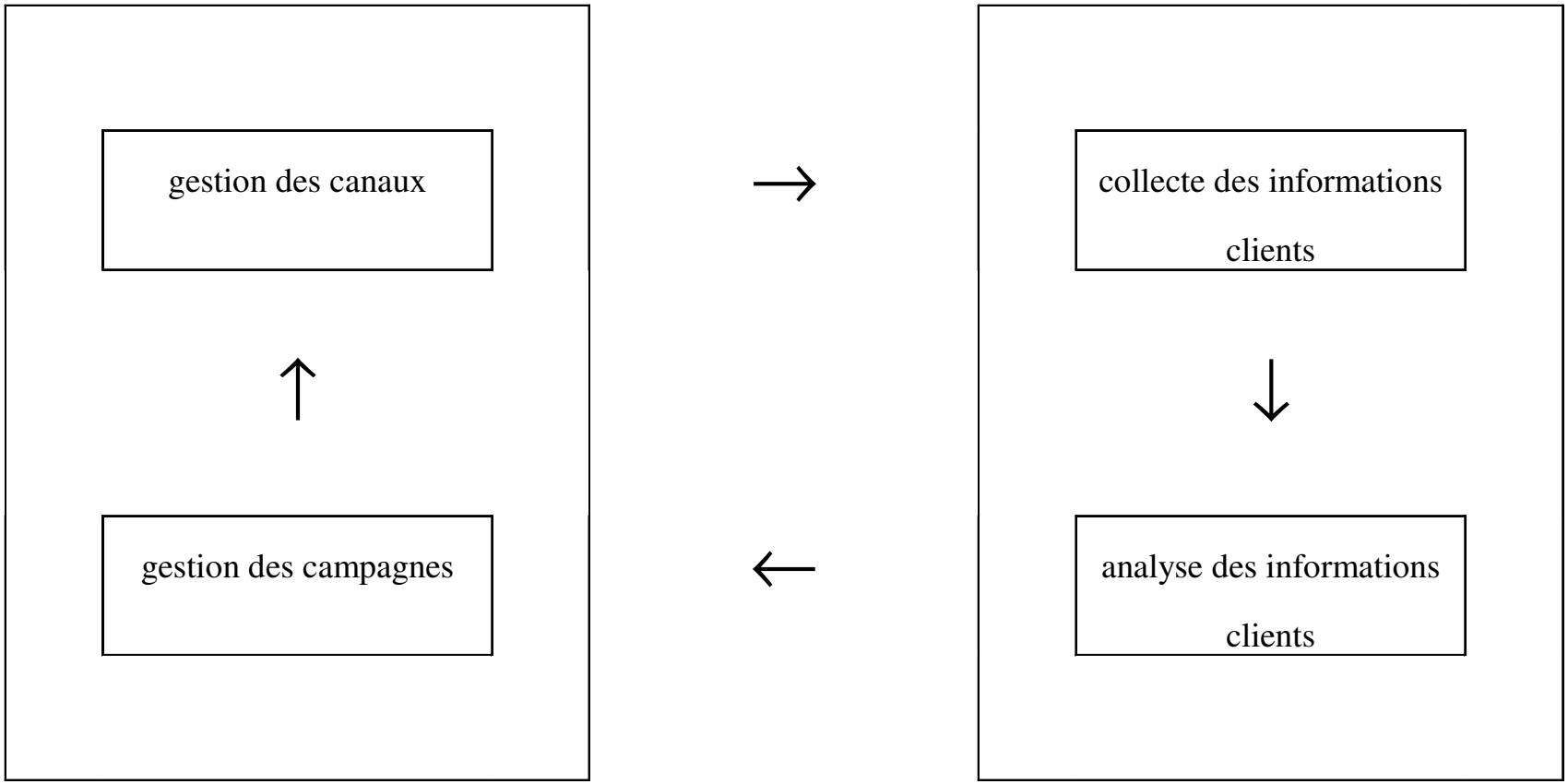


Data mining et CRM

Rappel : Gestion de la relation client

- La richesse des entreprises : leurs clients
- Objectifs des entreprises :
  - augmenter la rentabilité et la fidélité de leurs clients
  - en maîtrisant les risques
  - en utilisant les bons canaux au bon moment pour vendre le bon produit
- Un des moyens d'y parvenir :
  - la Gestion de la Relation Client (GRC)
  - synonyme : *Customer Relationship Management* (CRM)
  - 2 éléments : CRM analytique, CRM opérationnel
- Une matière 1ère précieuse : les données sur les clients

CRM analytique et opérationnel



Le CRM opérationnel

- Objectif
  - mise en œuvre optimale des stratégies identifiées grâce au CRM analytique
- Moyens
  - gestion des différents canaux
    - forces commerciales, centres d'appels téléphoniques, serveurs vocaux, Minitel, centres d'appel web, bornes interactives, téléphonie mobile, TV interactive…
  - gestion des campagnes marketing
- Composants
  - outils interfacés avec les applications de back-office, les progiciels de gestion intégrée (ERP), les outils de work-flow, de gestion des agendas et des alertes commerciales

Le CRM analytique

- Objectif
  - fournir une vision complète et unifiée du client dans l'entreprise et mieux comprendre son profil et ses besoins
- Moyens
  - extraction, stockage, analyse et restitution des informations pertinentes
- Composants
  - data warehouse
  - data mart
  - analyse multidimensionnelle (OLAP)
  - data mining
  - outils de reporting

Ce que l'on veut savoir

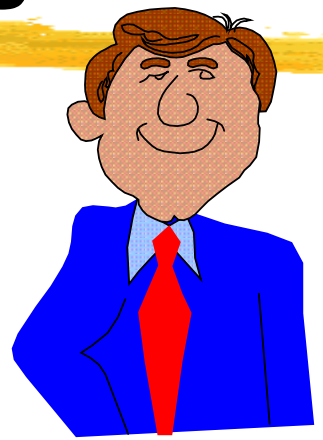
- On ne veut plus seulement savoir :
  - « Combien de clients ont acheté tel produit pendant telle période ? »
- Mais :
  - « Quel est leur profil ? »
  - « Quels autres produits les intéresseront ? »
  - « Quand seront-ils intéressés ? »

Data mining ≠ statistiques descriptives

- Les profils de clientèle à découvrir sont en général des profils complexes : pas seulement des oppositions « jeunes/seniors », « citadins/ruraux »… que l'on pourrait deviner en tâtonnant par des statistiques descriptives, mais des combinaisons plus complexes qui ne pourraient pas être découvertes par hasard.

> **Le data mining fait passer**
  - **d'analyses confirmatoires**
  - **à des analyses exploratoires.**

Utilité du data mining



- Mieux connaître le client
  - ➡ pour mieux le servir
    - ➡ pour augmenter sa satisfaction
      - ➡ pour augmenter sa fidélité

        (+ coûteux d'acquérir un client que le conserver)

- La connaissance du client est encore plus utile dans le secteur tertiaire :
  - les produits se ressemblent entre établissements
  - le prix n'est pas toujours déterminant
  - ce sont surtout le service et la relation avec le client qui font la différence

Applications du data mining au CRM

- Études d'appétence dans les sociétés commerciales
  - pour concentrer les mailings et le phoning sur les clients les plus susceptibles de répondre favorablement
- Prédiction de l'attrition dans la téléphonie mobile
  - attrition = départ d'un client pour un concurrent
- Analyse du ticket de caisse dans les grandes surfaces
  - pour déterminer les produits souvent achetés simultanément, et agencer les rayons et organiser les promotions en conséquence

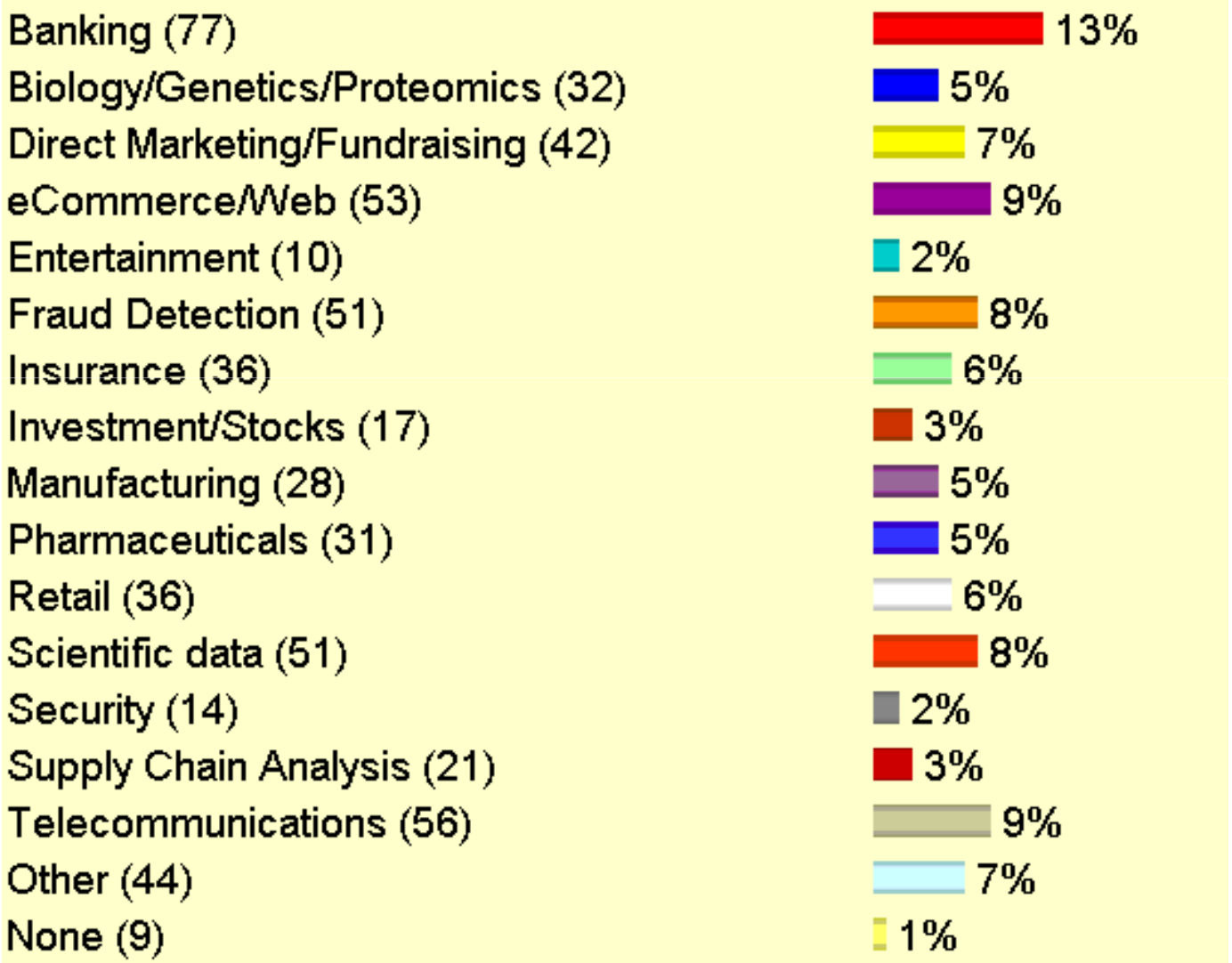
Marketing one-to-one

Marketing traditionnel	**Marketing 1:1**
Client anonyme	Client individualisé
Produit standard	Produit et service personnalisés
Production en série	Production sur mesure
Publicité à large diffusion	Message individuel
Communication unilatérale	Communication interactive
Réalisation d'une vente, fort taux de souscription	Fidélisation du client, faible taux d'attrition
Part de marché	Part de client
Large cible	Niche rentable
Segmentation métier	Segmentation statistique
Canaux de distribution traditionnels, déconnectés	Nouveaux canaux (plates-formes téléphoniques, Internet, mobiles), interconnectés
Marketing orienté « produit »	Marketing orienté « client »

A quoi sert le data mining ?

Sondage sur www.kdnuggets.com

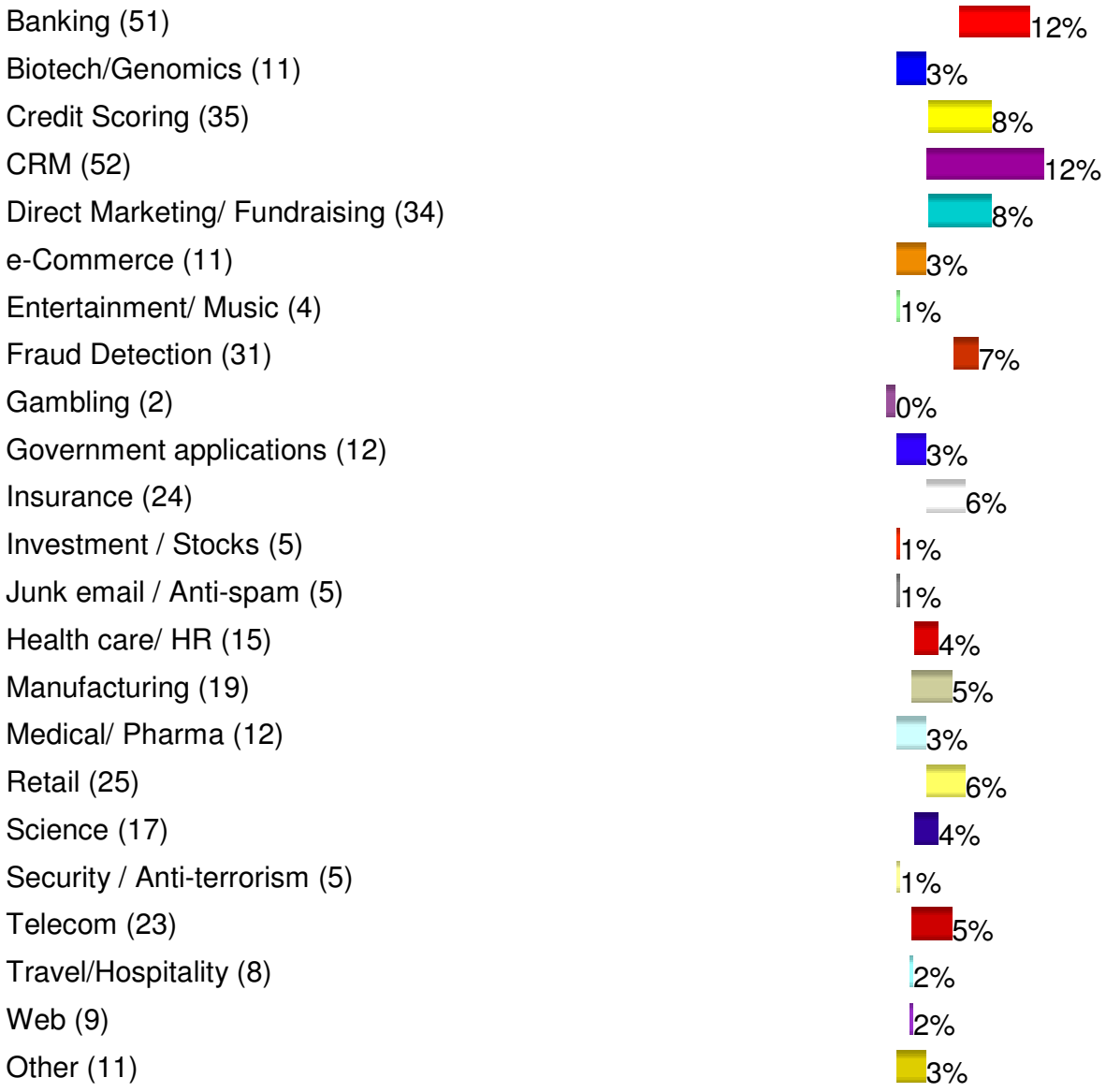
Industries/fields where you currently apply data mining: [608 votes total]



Sondage effectué en juin 2002

Sondage sur www.kdnuggets.com

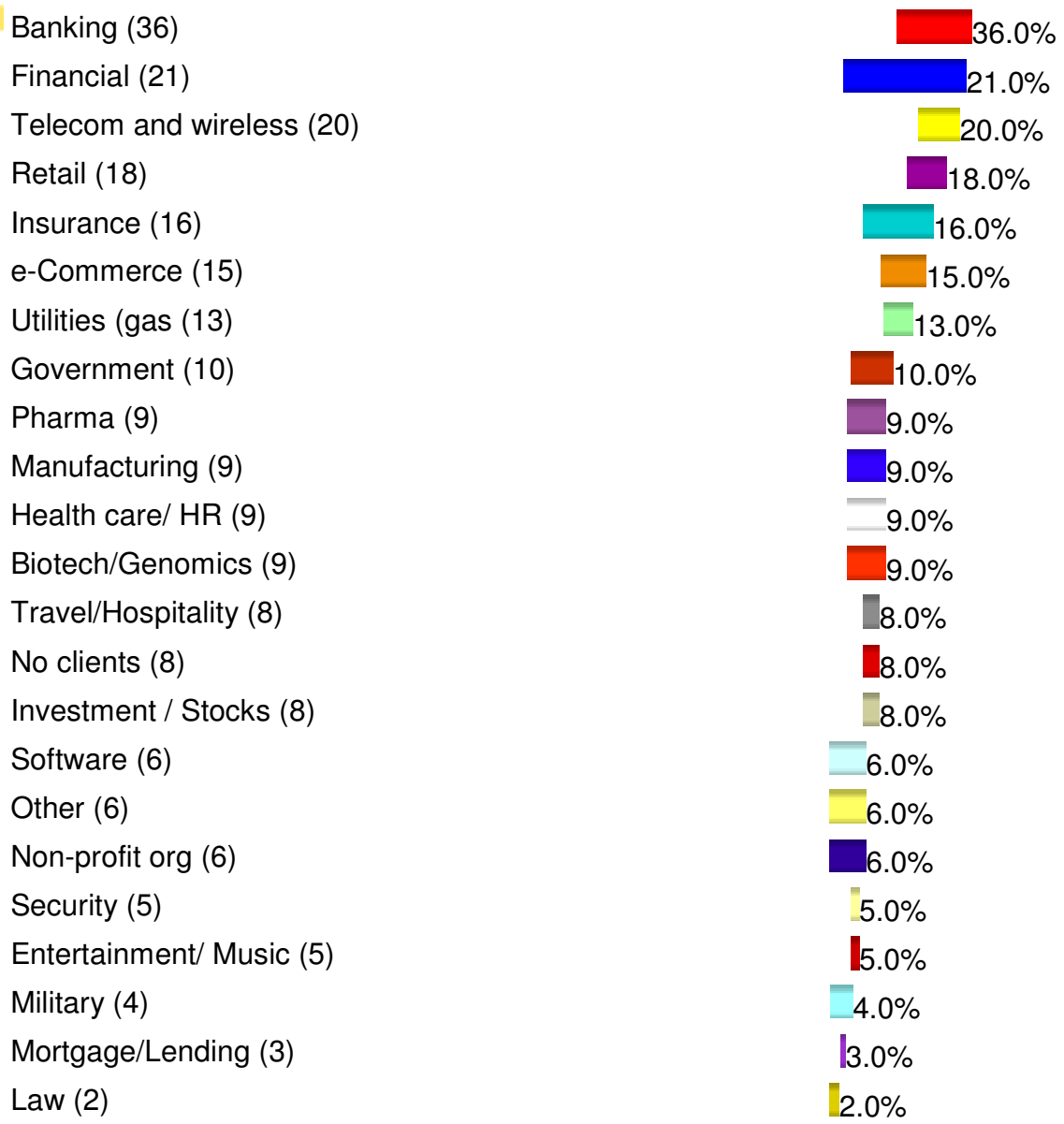
Industries/fields where you *successfully* applied data mining in the past 3 years [149 replies, 421 votes total]



Sondage effectué en juillet 2005

Sondage sur www.kdnuggets.com

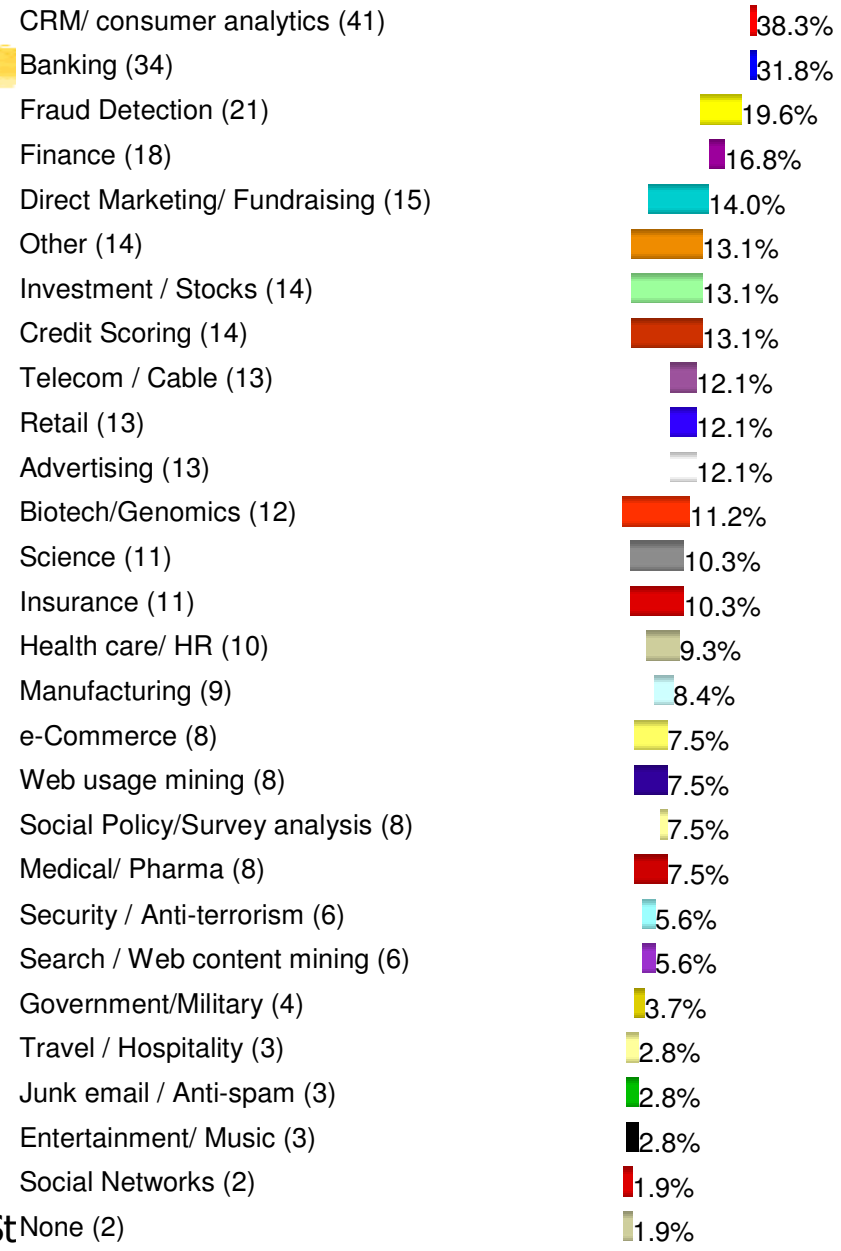
In what industries/sectors were your data mining clients in 2007-2008? [100 voters]



Sondage effectué en mars 2008

Sondage sur www.kdnuggets.com

Industries / Fields where you applied Data Mining in 2008: [107 voters]



Sondage effectué en décembre 2008

Le data mining dans la banque

- Naissance du score de risque en 1941 (David Durand)
- Multiples techniques appliquées à la banque de détail et la banque des entreprises
- Surtout la banque de particuliers :
  - montants unitaires modérés
  - grand nombre de dossiers
  - dossiers relativement standards
- Essor dû à :
  - développement des nouvelles technologies
  - nouvelles attentes de qualité de service des clients
  - concurrence des nouveaux entrants (assureurs, grande distribution) et des sociétés de crédit
  - pression mondiale pour une plus grande rentabilité
  - **surtout :** nouveau ratio de solvabilité Bâle 2

Exemples bancaires

- Utilisation du score de risque pour proposer le montant de crédit le plus adapté à chaque client
- Aide à la décision de paiement
- Meilleur taux de réponse des campagnes marketing
- Découverte de segments de clientèle
- Adaptation de la communication marketing à chaque segment de clientèle
- Choix du meilleur canal de distribution
- Identification des clients susceptibles de partir à la concurrence
- Calcul de la rentabilité et de la *life time value*

Le data mining dans l'assurance IARD

- Des produits obligatoires (automobile, habitation) :
  - soit prendre un client à un concurrent
  - soit faire monter en gamme un client que l'on détient déjà
- D'où les sujets dominants :
  - attrition
  - ventes croisées (*cross-selling*)
  - montées en gamme (*up-selling*)
- Besoin de décisionnel dû à :
  - concurrence des nouveaux entrants (bancassurance)
  - bases clients des assureurs traditionnels mal organisées :
    - compartimentées par agent général
    - ou structurées par contrat et non par client

Le data mining dans la téléphonie

- Deux événements :
  - ouverture du monopole de France Télécom
  - arrivée à saturation du marché de la téléphonie mobile
- D'où les sujets dominants dans la téléphonie :
  - score d'attrition (*churn* = changement d'opérateur)
  - *text mining* (pour analyser les lettres de réclamation)
  - optimisation des campagnes marketing
  - score d'impayés
- Problème du *churn* :
  - coût d'acquisition moyen en téléphonie mobile : 150 euros
  - plus d'un million d'utilisateurs changent chaque d'année d'opérateur

Le data mining dans le commerce

- VPC
  - utilise depuis longtemps des scores d'appétence
  - pour optimiser ses ciblages et en réduire les coûts
  - La Redoute envoie à sa clientèle 250 millions de documents par an
- e-commerce
  - personnalisation des pages d'un site web en fonction du profil de chaque internaute
  - optimisation de la navigation sur un site web
- Distribution
  - détermination des profils de consommateurs, du « ticket de caisse », de l'effet des soldes ou de la publicité
  - détermination des meilleures implantations (géomarketing)

Exemples médicaux

- Déterminer des segments de patients susceptibles d'être soumis à des protocoles thérapeutiques déterminés, chaque segment regroupant tous les patients réagissant identiquement
- Mettre en évidence des facteurs de risque ou de rémission dans certaines maladies. Choisir le traitement le + approprié
- Pronostic des infarctus et des cancers (décès, survie)
- Prédire le temps de rétablissement après une opération, en fonction des données concernant le patient (âge, poids, taille, fumeur, métier, antécédents médicaux, etc.) et le praticien (nb d'opérations pratiquées, nb d'années d'expérience, etc.)
- Décryptage du génome
- Tests de médicaments, de cosmétiques
  - Prédire les effets sur la peau humaine de nouveaux cosmétiques, en limitant le nombre de tests sur les animaux

Exemples divers

- Contrôle qualité
  - recherche des facteurs expliquant les défauts de la production
- Prévisions de trafic routier (Bison futé), recherche des causes des accidents
- Prédiction des parts d'audience pour une nouvelle émission de télévision (BBC)
  - en fonction des caractéristiques de l'émission (genre, horaire, durée, présentateur…), des programmes précédant et suivant cette émission sur la même chaîne, des programmes diffusés simultanément sur les chaînes concurrentes, des conditions météorologiques, de l'époque de l'année et des événements se déroulant simultanément
- Le classement en « étoile » ou « galaxie » d'un nouveau corps céleste découvert au télescope (système SKICAT)

Les deux grandes familles de 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 (c'est le cas des *segmentations* de clientèle et des *recherches d'associations* de produits sur les tickets de caisse)
  - réduisent, résument, synthétisent les données
  - il n'y a pas de variable « cible » à prédire.
- Les techniques prédictives :
  - visent à **extrapoler de nouvelles informations** à partir des informations présentes (c'est le cas du *scoring*)
  - expliquent les données
  - il y a une variable « cible » à prédire.

Les 2 types de techniques de DM

- Les techniques descriptives :
  - analyse factorielle
  - classification automatique (clustering)
  - recherche d'associations (analyse du ticket de caisse)
- Les techniques prédictives :
  - classement/discrimination (variable « cible » qualitative)
    - analyse discriminante / régression logistique
    - arbres de décision
    - réseaux de neurones
  - prédiction (variable « cible » quantitative)
    - régression linéaire (simple et multiple)
    - ANOVA, MANOVA, ANCOVA, MANCOVA (GLM)
    - arbres de décision
    - réseaux de neurones

Méthodes descriptives

type	famille	sous-famille	algorithme
méthodes descriptives	modèles géométriques	analyse factorielle (projection sur un espace de dimension inférieure)	analyse en composantes principales ACP (var. continues)
			analyse factorielle des correspondances AFC (var. qualitativ.)
			analyse des correspondances multiples ACM (var. qualitatives)
		analyse typologique (regroupement en classes homogènes)	méthodes de partitionnement (centres mobiles, *k*-means, nuées dynamiques)
			méthodes hiérarchiques
		analyse typologique + réduction dimens.	classification neuronale (cartes de Kohonen)
	modèles combinatoires		classification relationnelle (var. qualitatives)
	modèles à base de règles logiques	détection de liens	détection d'associations

En grisé : méthodes « classiques »

Méthodes prédictives

type	famille	sous-famille	algorithme
méthodes prédictives	modèles à base de règles logiques	arbres de décision	arbres de décision (variable à expliquer continue ou qualitative)
	modèles à base de fonctions mathématiques	réseaux de neurones	réseaux à apprentissage supervisé : perceptron multicouches, réseau à fonction radiale de base
		modèles paramétriques ou semi-paramétriques	régression linéaire, ANOVA, MANOVA, ANCOVA, MANCOVA, modèle linéaire général GLM, régression PLS (variable à expliquer continue)
			analyse discriminante linéaire, régression logistique, régression logistique PLS (variable à expliquer qualitative)
			modèle log-linéaire, régression de Poisson (variable à expliquer discrète = comptage)
			modèle linéaire généralisé, modèle additif généralisé (variable à expliquer continue, discrète ou qualitative)
	prédiction sans modèle		*k*-plus proches voisins (*k*-NN)

En grisé : méthodes « classiques »