

# Stéphane Tufféry

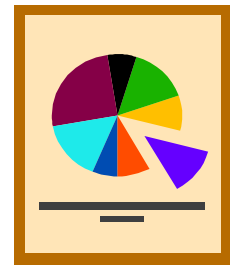
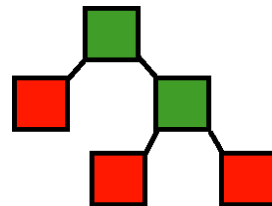
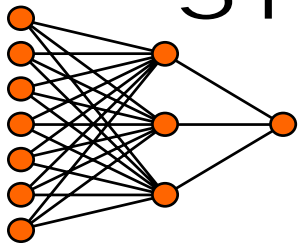
Statisticien - Data Miner - Enseignant



## DATA MINING - SCORING

STATISTIQUE DÉCISIONNELLE

APPLICATION AU CRM



# Présentation de l'auteur

- En charge du *data mining* dans une grande entreprise
  - où il a développé plusieurs systèmes de scoring et de segmentation
- Enseigne le *data mining* en DESS dans les Universités de Rennes et Nantes
- Docteur en Mathématiques
- Auteur de :
  - *Data Mining et Scoring*, Éditions Dunod, 2003  
Ouvrage consacré à l'application des techniques et méthodologies de data mining à la gestion de la relation client - Avec étude de cas
- Contact :
  - [data.mining\[chez\]free.fr](http://data.mining[chez]free.fr)



# Présentation du cours

- Cette présentation est issue de cours donnés dans des cours de DESS d'Économétrie entre 1999 et 2004.
- Ces enseignements ont ensuite trouvé un développement dans un ouvrage publié chez Dunod, dont l'essentiel est résumé ici, exceptée l'étude de cas consacrée à l'élaboration d'un score d'appétence pour le marketing.
- Ce cours est donc consacré aux techniques de data mining et de scoring, et à leur mise en oeuvre en entreprise. Elle contient une introduction, une partie technique (préparation des données, régressions, analyse factorielle, discriminante, arbres de décision, réseaux de neurones, algorithmes génétiques, centres mobiles, CAH...) et une partie méthodologique (conduite de projet, facteurs de succès, aspects informatiques, CNIL...).

# Plan du cours

- *Qu'est-ce que le data mining ?*
- *A quoi sert le data mining ?*
- *Les 2 grandes familles de techniques*
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs à éviter
- Informatique décisionnelle et de gestion
- La préparation des données
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- Logiciels et consultants
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

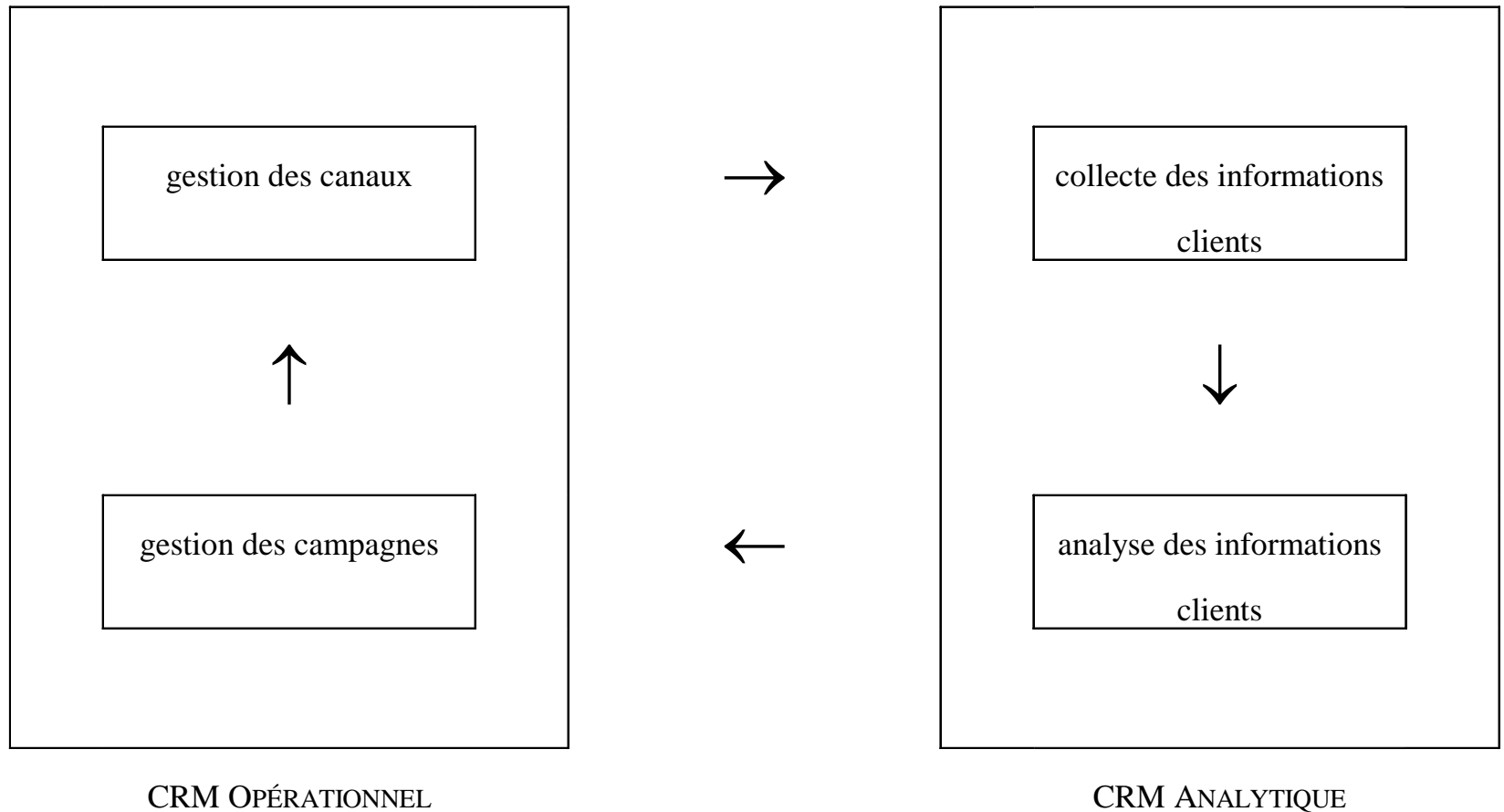


# Qu'est-ce que le data mining ?

# 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 1<sup>ère</sup> précieuse : les données sur les clients

# CRM analytique et opérationnel



# 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

# 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, Internet, 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 workflow, de gestion des agendas et des alertes commerciales

# De plus en plus de données

- L'accroissement des expertises et de la technicité
    - ... font perdre l'approche globale
    - ... obligent à stocker de plus en plus de données pour les besoins opérationnels de la gestion quotidienne
  - Mais : « trop de données tue la donnée »
  - Et nous connaissons de moins en moins nos clients
  - Explosion du nombre de rapports et tableaux de bord
  - Mais : perte du contact avec le client
- > Il faut réussir à tirer partie de cette complexité

# Fouiller dans les données

- Le **data mining** est l'ensemble des :
  - algorithmes et méthodes
  - ... destinés à l'exploration et l'analyse
  - ... de grandes bases de données informatiques
  - ... sans *a priori*
  - ... en vue de détecter dans ces données des règles, des tendances inconnues ou cachées, des structures particulières restituant de façon concise l'essentiel de l'information utile
  - ... pour l'aide à la décision



DATA MINING  
ET SCORING

# 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 $\neq$ 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.

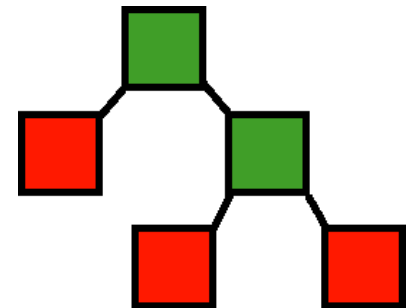
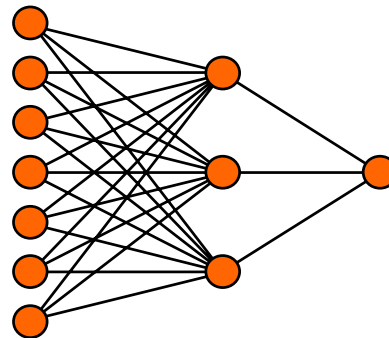
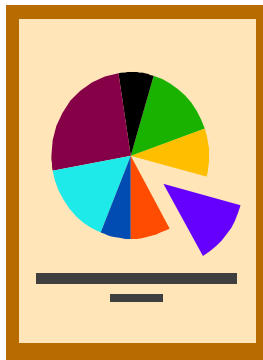
## > Le data mining fait passer

- d'analyses confirmatoires
- à des analyses exploratoires.



# Data mining $\neq$ statistiques descriptives

- Les techniques de data mining sont donc bien sûr + 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.)



# Le data mining aujourd'hui

- Ces techniques ne sont pas toutes récentes
- Ce qui est nouveau, ce sont aussi :
  - les capacités de stockage et de calcul (parallèle) offertes par le matériel et les techniques informatiques modernes (architectures massivement parallèles)
  - le « package » de techniques de natures différentes, qui peuvent s'enchaîner automatiquement les unes aux autres
  - l'intégration du DM 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.

# Historique

- 1896 : régression linéaire et corrélation de Karl Pearson
- 1936 : analyse discriminante de Fisher et Mahalanobis
- 1943 : réseaux de neurones de Mc Culloch et Pitts
- 1962 : analyse des correspondances de J.-P. Benzécri
- 1962 : régression logistique de J. Cornfield
- 1965 : méthode des centres mobiles de E. W. Forgy
- 1965 : arbre de décision AID de J.P.Sonquist et J.-A.Morgan
- 1972 : modèle linéaire général (GLM) de Nelder et Wedderburn
- 1975 : algorithmes génétiques de Holland
- 1984 : arbre CART de Breiman, Friedman, Olshen, Stone
- 1986 : Perceptron multicouches de Rumelhart et McClelland
- 1989 : réseaux de T. Kohonen (cartes auto-adaptatives)
- 1993 : arbre C4.5 de J. Ross Quinlan
- 1996 : bagging (Breiman) et boosting (Freund-Shapire)
- 2001 : forêts aléatoires de Breiman

# Des statistiques ...

- Statistiques
  - quelques centaines d'individus
  - quelques variables recueillies avec un protocole spécial (échantillonnage, plan d'expérience...)
  - fortes hypothèses sur les lois statistiques suivies
- 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
  - quelques millions d'individus
  - quelques centaines de variables
  - nombreuses variables non numériques
  - données recueillies avant l'étude, et souvent à d'autres fins
  - population constamment évolutive (difficulté d'échantillonner)
  - données imparfaites, avec des erreurs de codification
  - nécessité de calculs rapides
  - on ne recherche pas toujours l'optimum mathématique, mais le modèle le plus facile à appréhender par des utilisateurs non-statisticiens

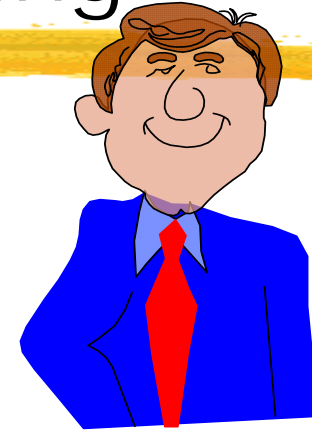
# Le data mining aujourd'hui

- 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...

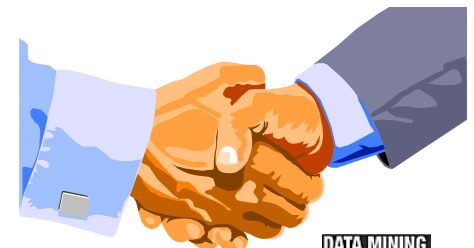


# Utilité du data mining

- 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.



# Marketing one-to-one

<b>Marketing traditionnel</b>	<b>Marketing 1:1</b>
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	Fidélisation du client
Part de marché	Part de client
Large cible	Niche rentable
Canaux de distribution traditionnels	Nouveaux canaux (plates-formes téléphoniques, Internet, mobiles)
Marketing orienté « produit »	Marketing orienté « client »

# Données d'un data warehouse

- **Consolidées** à partir des différents systèmes d'information de production (pour avoir une homogénéité des définitions et codifications)
- **orientées utilisateur** : données structurées par métiers, et non par applications informatiques
- **documentées** : des « métadonnées » indiquent la définition des données du DW, leur provenance, leurs règles et dates de mise à jour...
- **historisées** : avec des dispositifs d'épuration et de récapitulation automatique des données d'une certaine ancienneté
- **agrégées** : toutes les données n'ont pas besoin d'être stockées avec le même niveau de détail que dans les systèmes de production

# En résumé

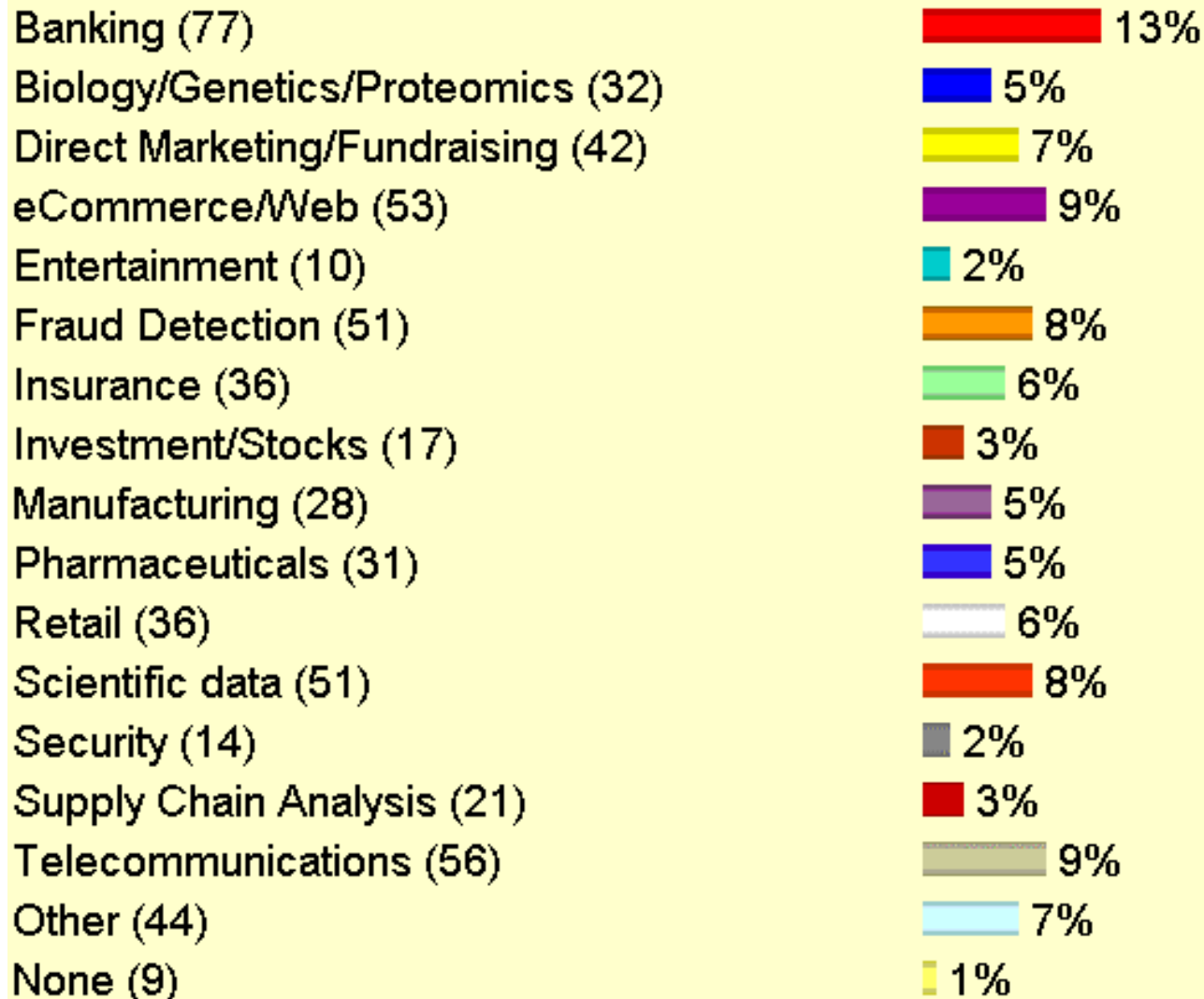
- Le data warehouse **fournit les données brutes** sur les clients, les produits, le réseau de distribution, les transactions (un data mart est un data warehouse, ou un niveau supplémentaire, spécialisé à un domaine)
- Le data mining **raffine ces données** en en **extrayant l'information utile** pour prendre des décisions
  - Comment trouver des diamants dans un tas de charbon ?
- Le reporting **diffuse l'information** décisionnelle.



# A quoi sert le data mining ?

# Sondage sur [www.kdnuggets.com](http://www.kdnuggets.com)

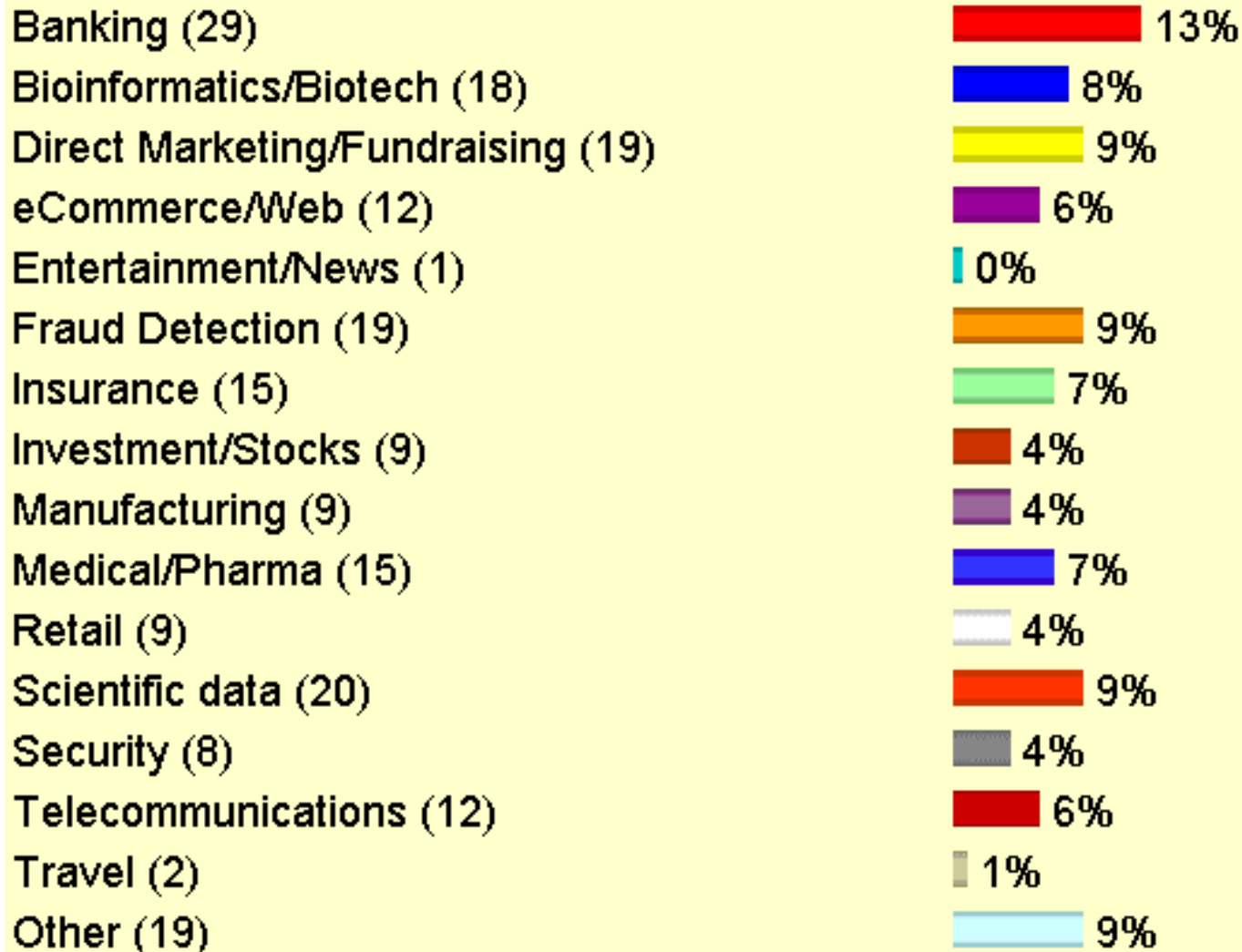
## Industries/fields where you currently apply data mining: [608 votes total]



Sondage  
effectué  
en juin  
2002

# Sondage sur [www.kdnuggets.com](http://www.kdnuggets.com)

## Industries/fields where you currently apply data mining [216 votes total]



Sondage  
effectué  
en août  
2004

# Applications du data mining au CRM

- Analyse du panier de la ménagère dans les grandes surfaces (pour déterminer les produits souvent achetés simultanément, et agencer les rayons et organiser les promotions en conséquence)
- 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 (c'est-à-dire du départ d'un client pour un concurrent) dans la téléphonie mobile

# Autres applications du data mining

- Détection de la fraude (compagnies d'assurance, téléphonie mobile, cartes bancaires)
- Credit scoring
- Contrôle qualité
- Études de marché en agroalimentaire
- Études médicales et pharmaceutiques
- Études biologiques et agronomiques
- Détection de certains risques (épidémiologiques)
  
- Selon le MIT (Massachusetts Institute of Technology) : le data mining est l'une des 10 technologies émergentes qui « changeront le monde » au XXI<sup>e</sup> siècle.



# Le data mining dans la banque

- Naissance du score de risque il y a 50 ans
- 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 + grande rentabilité
  - **surtout** : nouveau ratio de solvabilité Mc Donough

# Exemples bancaires

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

# Apport du scoring pour le crédit

- Les scores d'appétence et de risque permettent :
  - un gain de temps commercial
    - score de risque  $\Rightarrow$  analyse des dossiers plus rapides (pour le client : réponse plus rapide)
    - score d'appétence  $\Rightarrow$  moins de R-V pour vendre + de produits
  - des clients non importunés par des campagnes mal ciblées
  - une plus grande autonomie de négociation et de décision des commerciaux (quand le score de risque est bon)
  - des décisions homogènes selon les différents commerciaux et canaux de distribution
  - des décisions ajustées à la politique de l'entreprise (qui peut déplacer de façon précise les seuils de score de risque et d'appétence selon ses objectifs)
  - une limitation du risque de surendettement et de contentieux.

# 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 : 300 euros
  - + 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 ciblage et en réduire les coûts
  - La Redoute envoie à sa clientèle 250 millions de documents par an
- e-commerce
  - personnalisation des pages du site Web de l'entreprise, en fonction du profil de chaque internaute
- Distribution
  - détermination des profils de consommateurs, le « panier de la ménagère », 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
- 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)
- Prédiction des parts d'audience pour une nouvelle émission de télévision (BBC)
- Projet *Oasys* (« Offenders Assessment System » : système d'évaluation des délinquants) en GB
  - estimer le risque de récidive en cas de libération anticipée
  - afin de normaliser les décisions de libération anticipée
  - et d'augmenter leur nombre



# Les 2 grandes familles d'outils

# 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 :
  - classification (syn : « segmentation », « clustering »)
  - recherche d'associations
  - recherche de séquences similaires.
- 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 des correspondances multiples ACM (var. catégorielles)
		analyse typologique (regroupement en segments homogènes)	centres mobiles, <i>k</i> -means, nuées dynamiques
			classification hiérarchique
			classification neuronale (cartes de Kohonen)
			classification relationnelle
	modèles à base de règles logiques	détection de liens	détection d'associations
			recherche de séries similaires

# Méthodes prédictives

type	famille	sous-famille	algorithme
méthodes prédictives	modèles à base de règles logiques	arbres de décision	arbres de décision (var. à expliquer continue ou catégorielle)
	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 (var. à expliquer continue) modèle linéaire général (var. à expliquer continue)
			régression logistique (var. à expliquer catégorielle) analyse discriminante de Fisher (var. à expliquer catégorielle)
			modèle log-linéaire (var. à expliquer discrète)
			modèle linéaire généralisé (var. à expliquer continue, discrète ou catégorielle) modèle additif généralisé (var. à expliquer continue, discrète ou catégorielle)
	prédiction sans modèle		<i>k</i> -plus proches voisins ( <i>k</i> -NN)

# Utilisation des algorithmes prédictifs

- Plus délicats à mettre en œuvre
  - avoir une méthodologie rigoureuse pour éviter certaines erreurs (sur-apprentissage)
- Nécessitent au moins 1 an d'historique dans les données
- Avantages de cette technique plus complexe :
  - une vision dynamique et non statique du client
    - prendre en compte le passé permet de mieux prédire l'avenir
    - > le pouvoir prédictif d'un score tient + longtemps que celui d'une classification
  - une vision graduelle
    - un client est, ou n'est pas, dans un segment, au lieu d'avoir une note de score plus ou moins élevée

# Utilisation des algorithmes descriptifs

- Plus faciles à mettre en œuvre
- Nécessitent moins d'historique
- La *classification* permet :
  - une vision générale de la clientèle
    - permettant d'alimenter une réflexion stratégique
  - une volumétrie des différents types de client
    - estimation du potentiel commercial selon le type de produit
  - l'affectation de certains types de clients à certains types de commerciaux
  - la détection de certains types de clients
    - tels que les clients multifournisseurs ou à fort potentiel
  - la personnalisation de la communication (mailings ou pages Web) en fonction du segment du client.