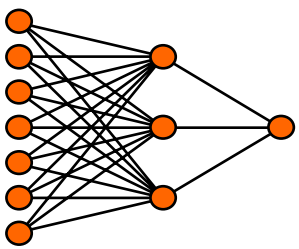
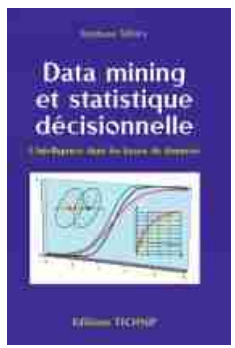
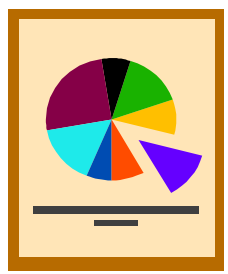
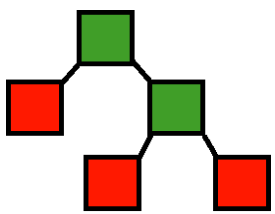
Stéphane Tufféry

DATA MINING
& STATISTIQUE DÉCISIONNELLE









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 - Consulting*
- La préparation des données
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- Logiciels de statistique et de data mining
- Informatique décisionnelle et de gestion
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

Le déroulement d'un projet de data mining

Les 10 étapes d'un projet

- Choix du sujet - Définition des objectifs
- Inventaire des données existantes
- Collecte, nettoyage et mise en forme des données
- Étude statistique de la base d'analyse
- Mise en œuvre des algorithmes (classification, scoring...) - Élaboration des modèles
- Validation et choix d'un modèle
- Déclaration à la CNIL
- Déploiement du modèle
- Formation des utilisateurs
- Suivi des modèles

Définition des objectifs

- Définir précisément le sujet et certains critères essentiels (variable cible)
  - exemple : « client à risque » et « client sans risque »
- Définir la population cible
  - tous les clients, les clients actifs, les prospects aussi...
  - unité statistique : individu, famille, entreprise, groupe...
- Déterminer la période à étudier
- Le sujet doit faire partie des objectifs de l'entreprise et lui apporter un avantage réel
- Les objectifs doivent être réalistes (tenir compte des actions passées et de la saturation du marché)
- Prévoir l'utilisation opérationnelle des modèles produits
  - forme de la restitution, périodicité de mise à jour, suivi...

Inventaire des données utiles

- Recenser avec les spécialistes métier et les informaticiens, les données utiles :
  - accessibles raisonnablement (pas sur microfilms !)
  - fiables
  - suffisamment à jour
  - historisées, si besoin est
  - légalement utilisables
- Il y a les données :
  - du système d'information (SI) de l'entreprise
  - stockées dans l'entreprise, hors du SI (fichiers Excel...)
  - achetées ou récupérées à l'extérieur de l'entreprise
  - calculées à partir des données précédentes (indicateurs, ratios, évolutions au cours du temps)

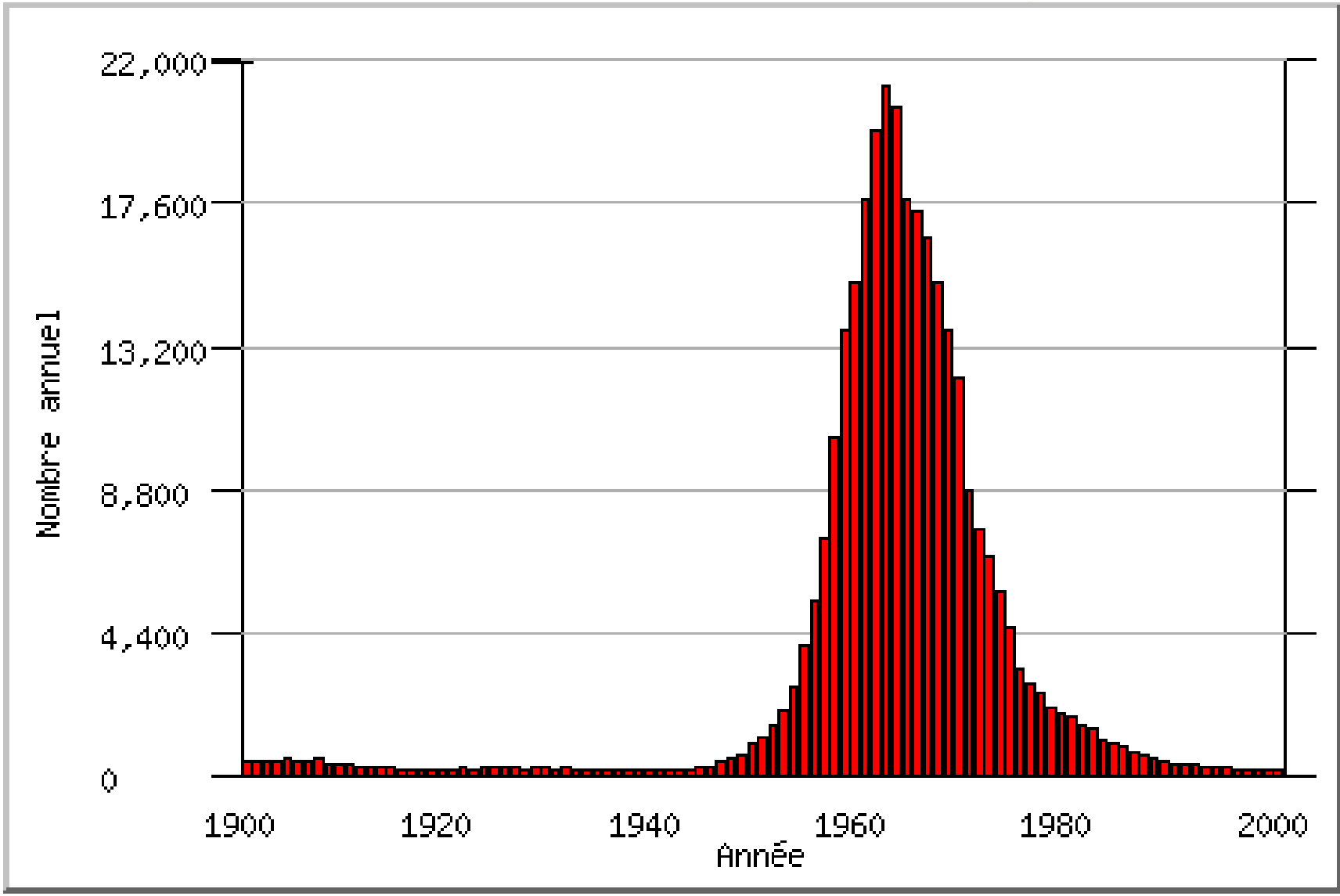
Quand on manque de données

- Enquêtes auprès d'échantillons de clients
  - en les incitant à répondre à des questionnaires en leur proposant des cadeaux
- Utilisation des mégabases de données (Axciom, Wegener Direct Marketing)
- Géomarketing (type d'habitat en fonction de l'adresse)
  - données moins précises que des données nominatives
  - mais disponibles pour des prospects
- « Scoring prénom »
- Recours à des modèles standards pré-établis par des sociétés spécialisées (ex : scores génériques)

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 (™Îlotype) : beaux quartiers, classe moyenne, classe ouvrière, centre ville et quartiers commerçants...

Scoring prénom (ex : Pascal)



Construction de la base d'analyse

n° client	variable cible : acheteur (O/N)	âge	PCS	situation famille	nb achats	montant achats	…	variable explicative *m*	échantillon
1	O	58	cadre	marié	2	40	…	…	apprentissage
2	N	27	ouvrier	célibataire	3	30	…	…	test
…	…	…		…	…	…	…	…	…
…	…	…		…	…	…	…	…	…
k	O	46	technicien	célibataire	3	75	…	…	test
…	…	…		…	…	…	…	…	…
…	…	…		…	…	…	…	…	…
2000	N	32	employé	marié	1	50	…	…	apprentissage
…	…	…		…	…	…		…	…

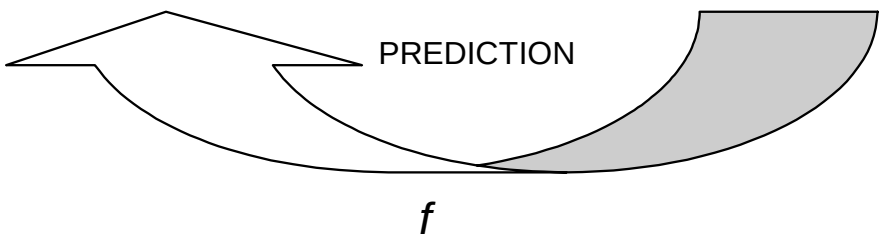
au moins 2000 cas

variable à expliquer
observée année n

variables explicatives
observées année n-1

répartition
aléatoire
des clients
entre les 2
échantillons

O : au moins 1000 clients ciblés dans l'année *n* et acheteurs
N : au moins 1000 clients ciblés dans l'année *n* et non acheteurs



Types de données 1/3

- Données de transaction et RFM
  - « 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)

Types de données 2/3

- Données sur les canaux
  - canal de prise de contact (parrainage, annonce presse, appel téléphonique, réponse à un mailing...)
  - canal privilégié de contact et communication (courrier, téléphone, Internet, magasin/agence...)
  - canal privilégié de commande (courrier, téléphone, Minitel, 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

Types de données 3/3

- 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...)
  - géographiques (ancienneté à l'adresse, code INSEE de la commune, IRIS et îlot INSEE, type d'habitat déduit de l'IRIS ou de l'îlot)
  - 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)

Importance des retours

- Le data mining **ne devine pas** le profil des clients à cibler, il l'**extrapole** à partir des données fournies.

- Pour les études d'appétence, les retours des actions commerciales précédentes (refus d'achat) permettent de dégager les profils positifs et négatifs
  - > **Il est capital de mémoriser cette information.**

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

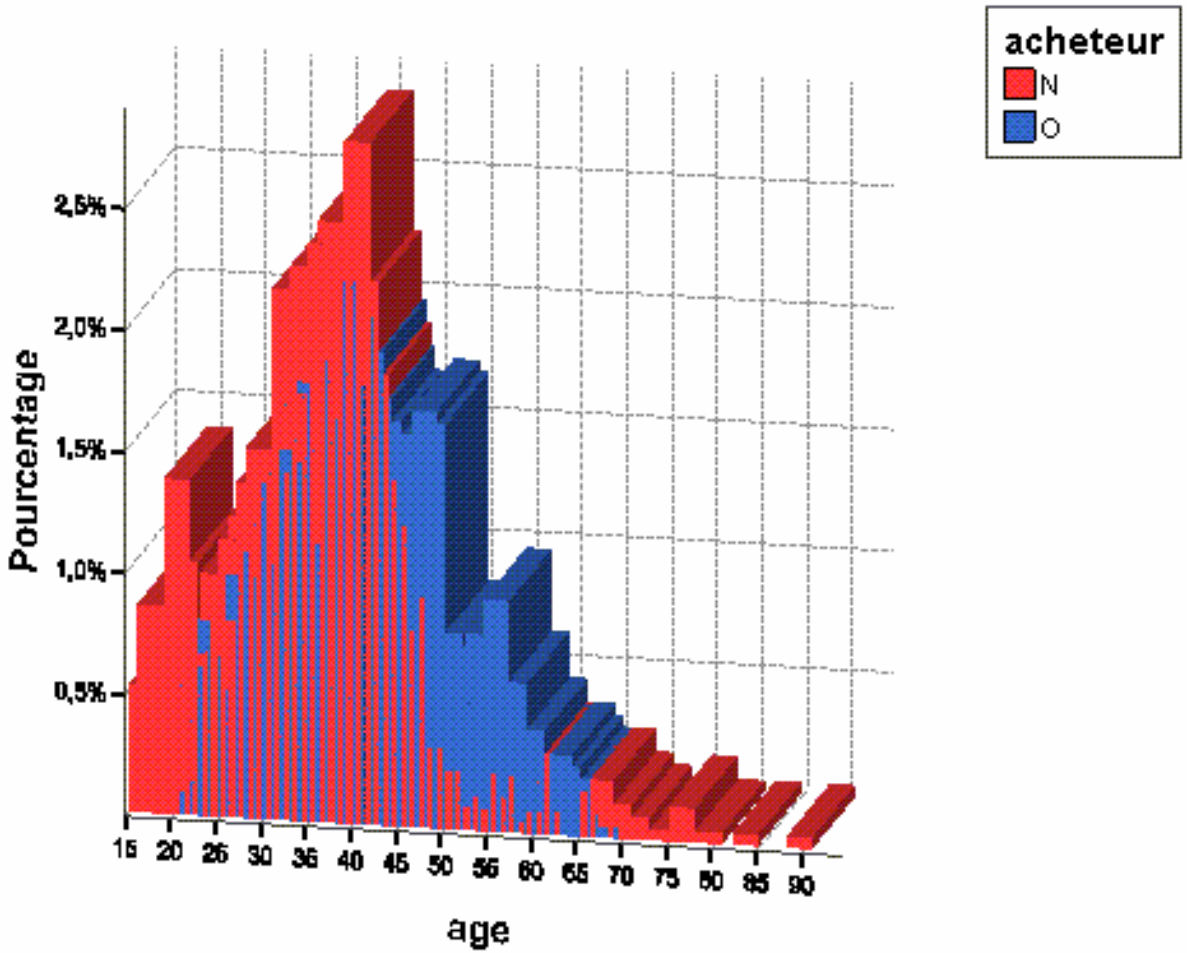
Sélection des données à utiliser

- Choix des variables les plus discriminantes
  - test du χ^2, V de Cramer (var. nominales) ou τ de Kendall (var. ordinales)
  - 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 elles
  - tests de multicolinéarité

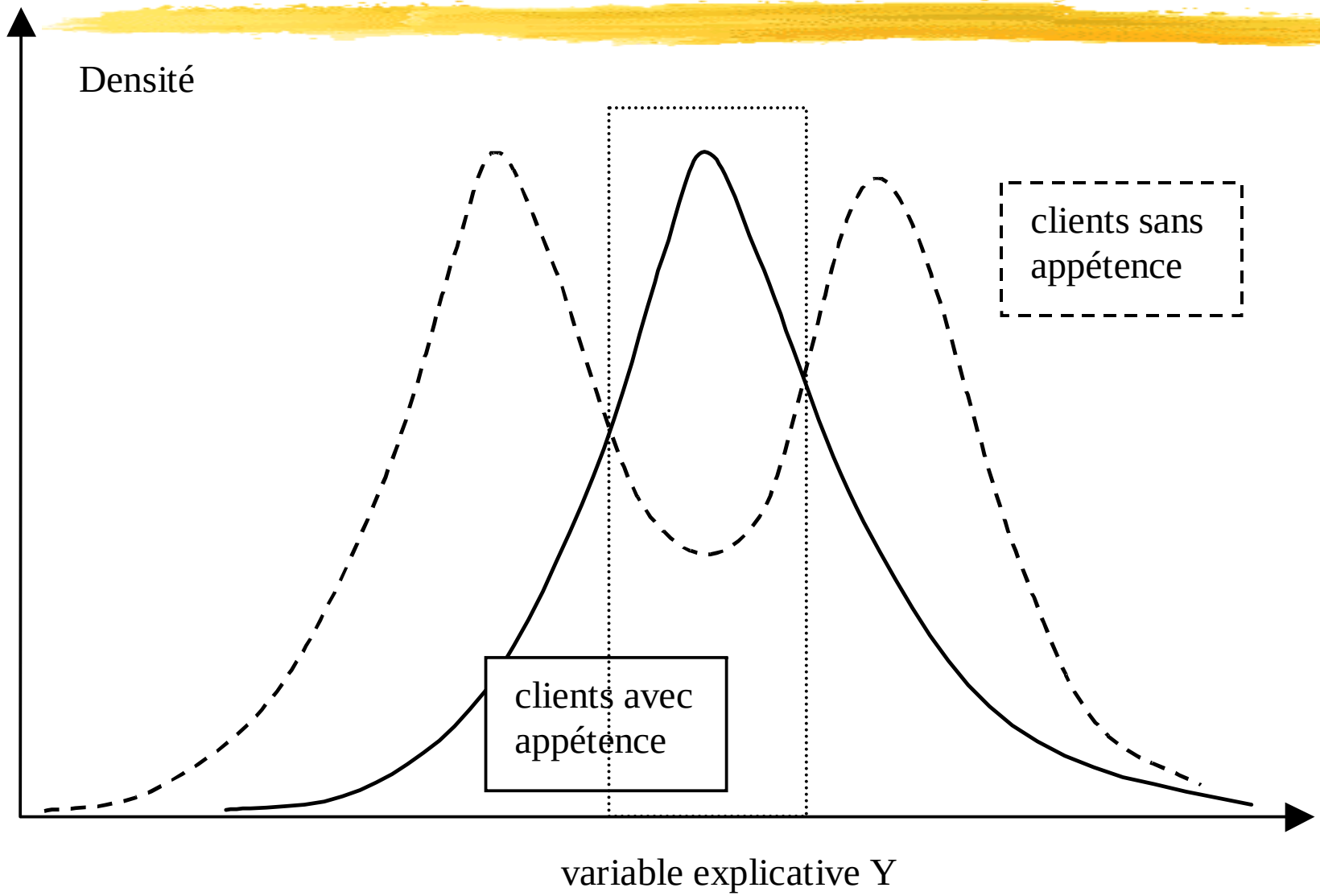
Création de nouvelles variables

- 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

Présence de 3 classes



Discrétisation en tranches naturelles



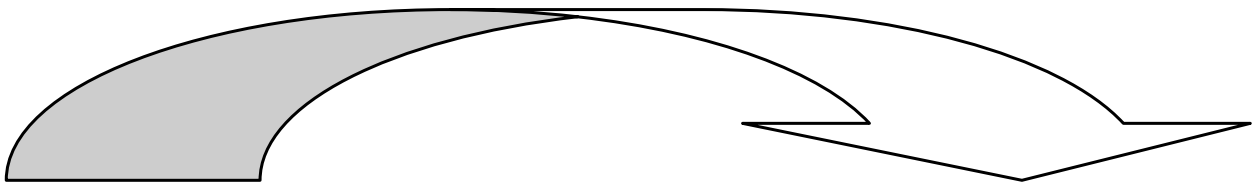
Pour l'élaboration des modèles prédictifs

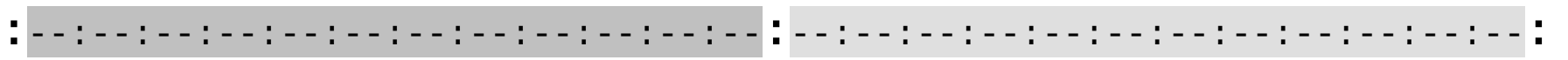
- (Facultatif) Pré-segmentation (classification) de la population étudiée :
  - en groupes forcément distincts selon les données disponibles (clients / prospects)
  - en groupes statistiquement pertinents vis-à-vis des objectifs de l'étude
  - selon certaines caractéristiques sociodémographiques (âge, 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 data mining

Pré-segmentation : questions 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

Analyser le passé pour prédire l'avenir





– 24 mois **– 12 mois** **aujourd'hui**

observation des variables explicatives

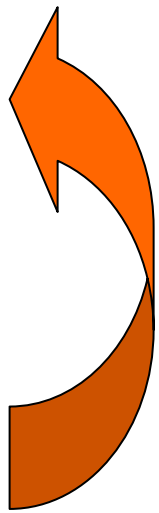
observation de la variable à expliquer

Le **modèle** sera par exemple une fonction *f* telle que :

Probabilité(variable cible = x) = *f*(variables explicatives)

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 3e échantillon, pour avoir une idée du taux d'erreur non biaisé du modèle
- **Application du modèle** à l'ensemble de la population

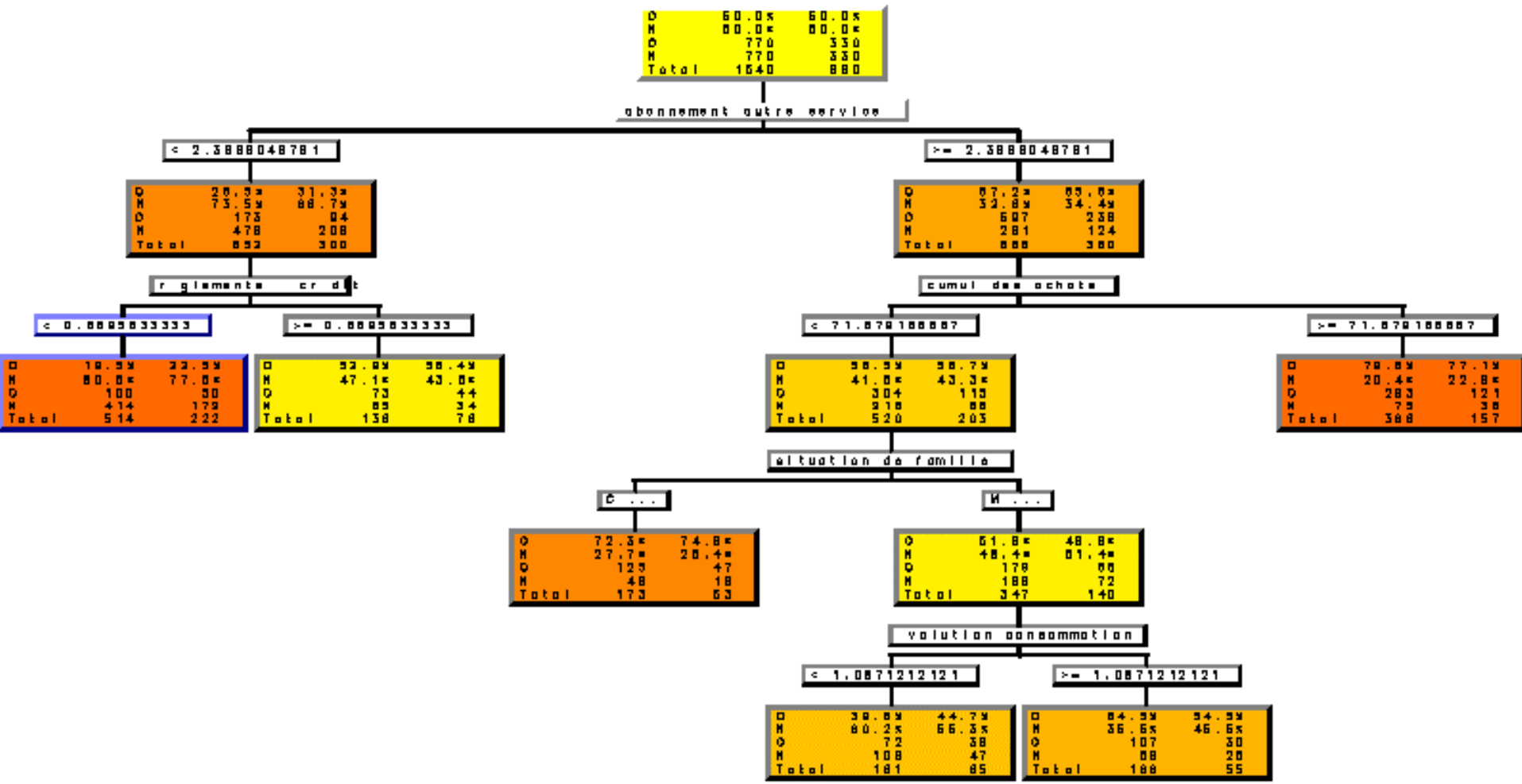


valeur prédite ➔ valeur réelle ➘	A	B	TOTAL
A	1800	200	
B	300	1700	
TOTAL			4000

Exemples de modèles prédictifs

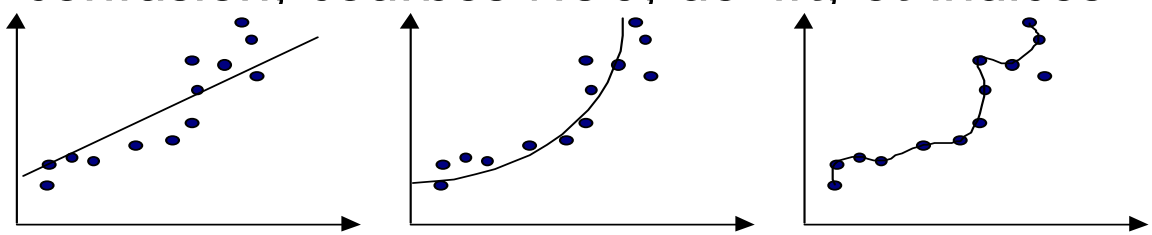
- Arbres de décision
  - Règles complètement explicites
  - Traitent les données hétérogènes, éventuellement manquantes, sans hypothèses de distribution
  - Détection de phénomènes non linéaires
  - Moindre robustesse
- Analyse discriminante linéaire
  - Résultat explicite $P(Y/ X_1, \ldots , X_p)$ sous 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

Algorithme d'arbre de décision

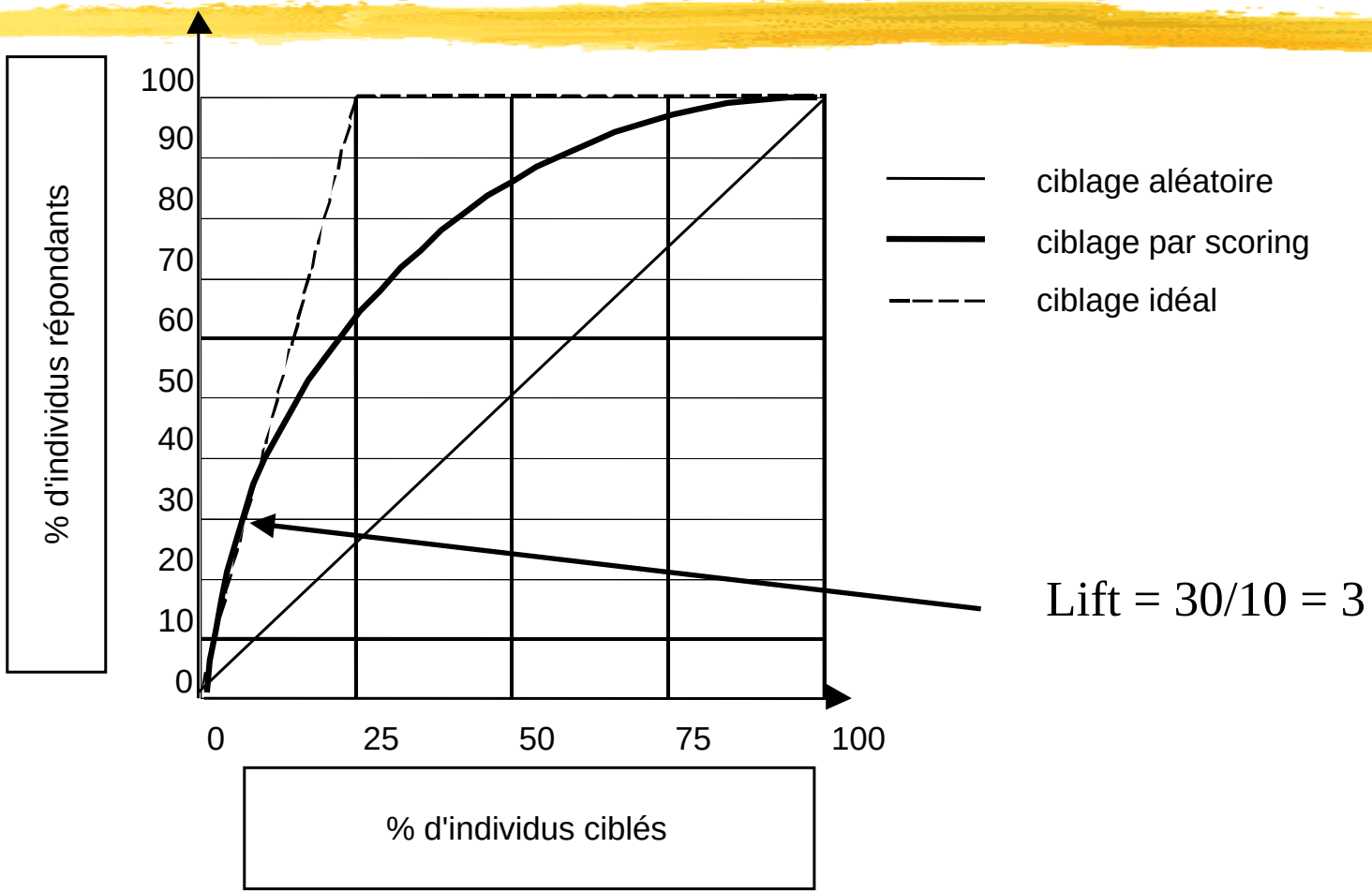


Validation des modèles

- 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



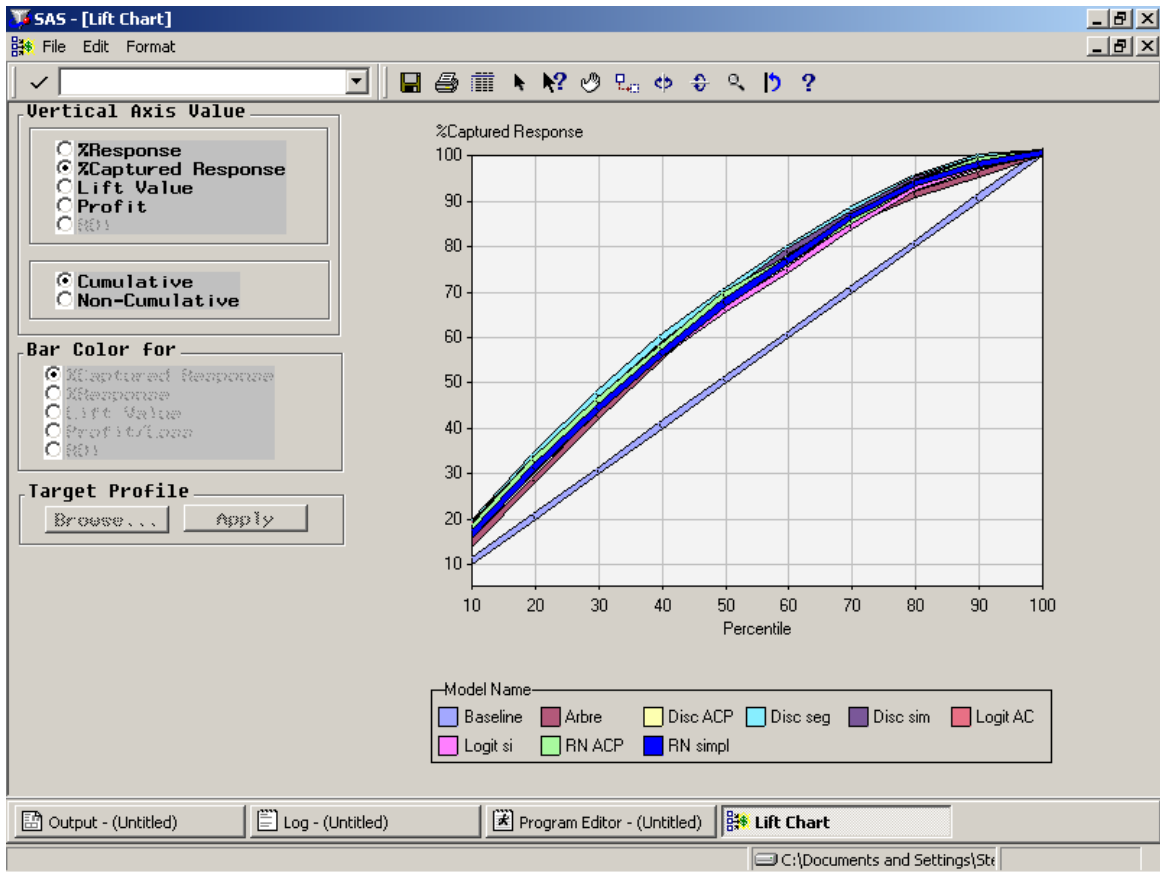
Validation d'un modèle de score



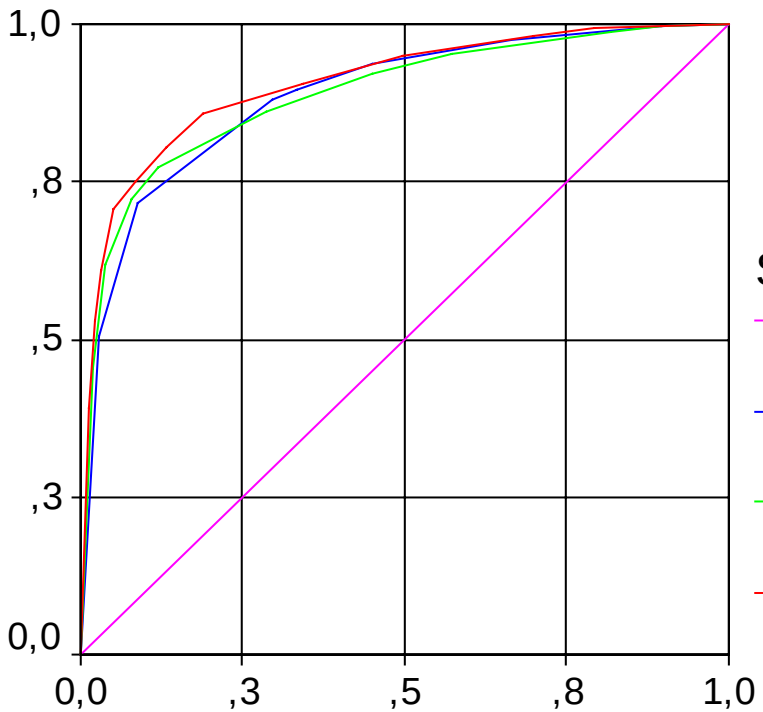
COURBE DE LIFT

Comparaison de modèles de score

- Les indicateurs statistiques de 2 modèles de nature différentes (exemple : R^2 d'une analyse discriminante et d'une régression logistique) sont rarement comparables
- On compare les modèles à l'aide des courbes ROC ou de lift



Comparaison de modèles : courbe ROC



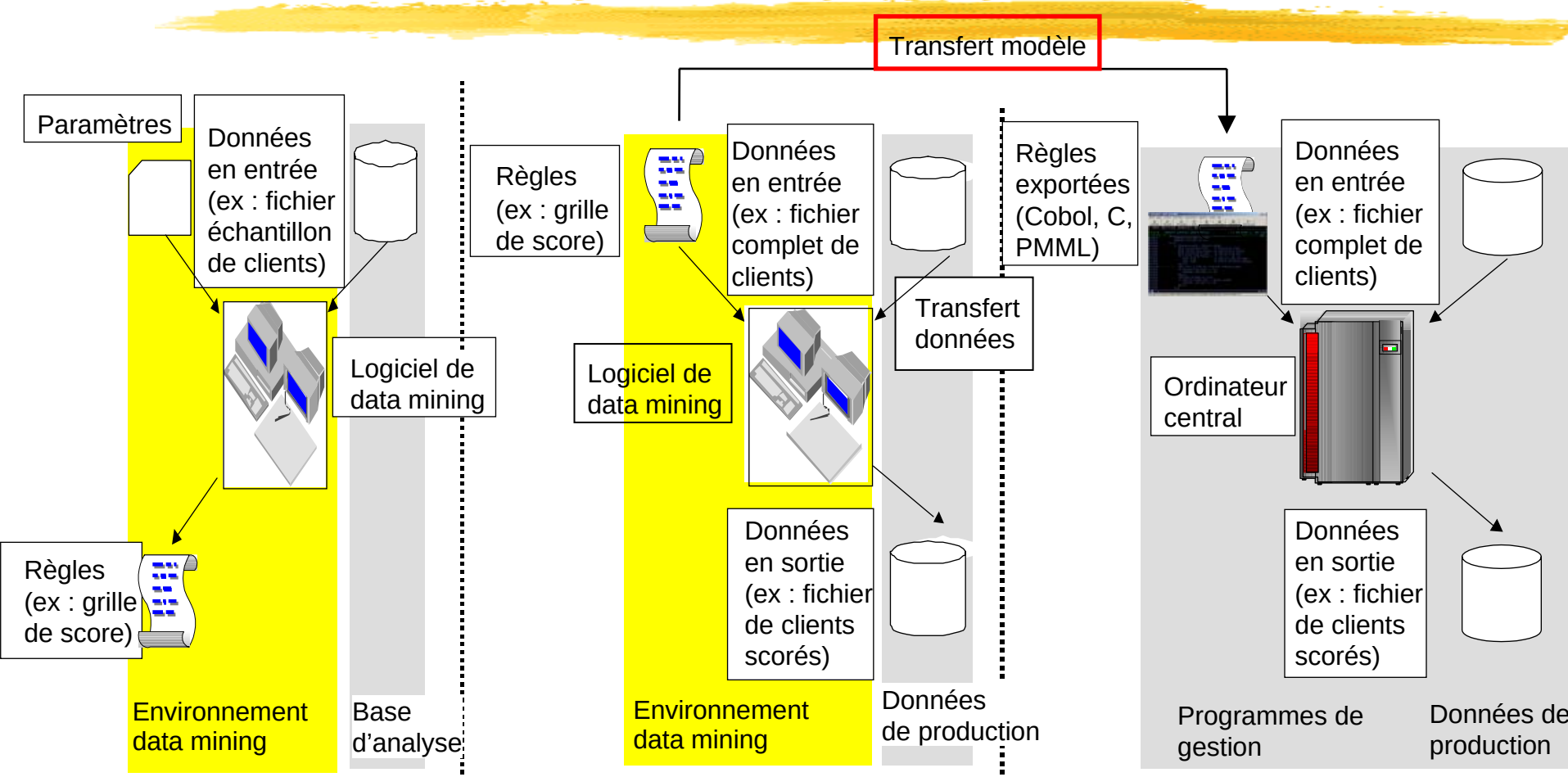
Zone sous la courbe

Variable(s) de	Zone
régression logistique	,906
analyse discriminante	,889
arbre de décision	,887

Déploiement informatique

- Plusieurs possibilités ne s'excluant pas :
  - utilisation d'un tableur sur un PC pour réaliser un publipostage (marketing direct)
  - intégration dans les fichiers clients de production ou d'infocentre (pour des ciblages)
  - intégration dans les fichiers clients de production et sur le poste de travail des commerciaux
- Différents niveaux de finesse de l'information
  - notes fines dans les fichiers (ex : de 1 à 1000)
  - restituées agrégées sur le poste de travail (ex : de 1 à 10)
    - et regroupées en tranches (par ex : faible / moyen / fort)
- Spécifier les habilitations d'accès
- Spécifier la périodicité des mises à jour

Utilisation des logiciels



Construction du modèle

Application du modèle dans l'environnement data mining

Application du modèle après export dans l'environnement de production

Utilisation opérationnelle d'un score

revenus

couleur	**% populat.**	**0 - 1000 €**	**1001 - 2 000 €**	**2 001 - 3 000 €**	**> 3 000 €**
rouge	20 %	à étudier	à étudier	à étudier	à étudier
orange	20 %	à étudier	découvert = 150 € carte = débit imméd. prêt pers. = 0 €	découvert = 450 € carte = débit imméd. prêt pers. = 3000 €	découvert = 750 € carte = débit diff. prêt pers. = 4500 €
vert (« foncé »)	20 %	découvert = 150 € carte = retrait prêt pers. = 0 €	découvert = 450 € carte = débit imméd. prêt pers. = 3000 €	découvert = 1500 € carte = débit diff. prêt pers. = 9000 €	découvert = 1500 € carte = débit diff. prêt pers. = 12 k€
vert (« clair »)	40 %	découvert = 150 € carte = débit immédiat prêt pers. = 0 €	découvert = 750 € carte = débit diff. prêt pers. = 7500 €	découvert = 1500 € carte = débit diff. prêt pers. = 12 k€	découvert = 1500 € carte = *Premium* prêt pers. = 15 k€

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

Suivi du score

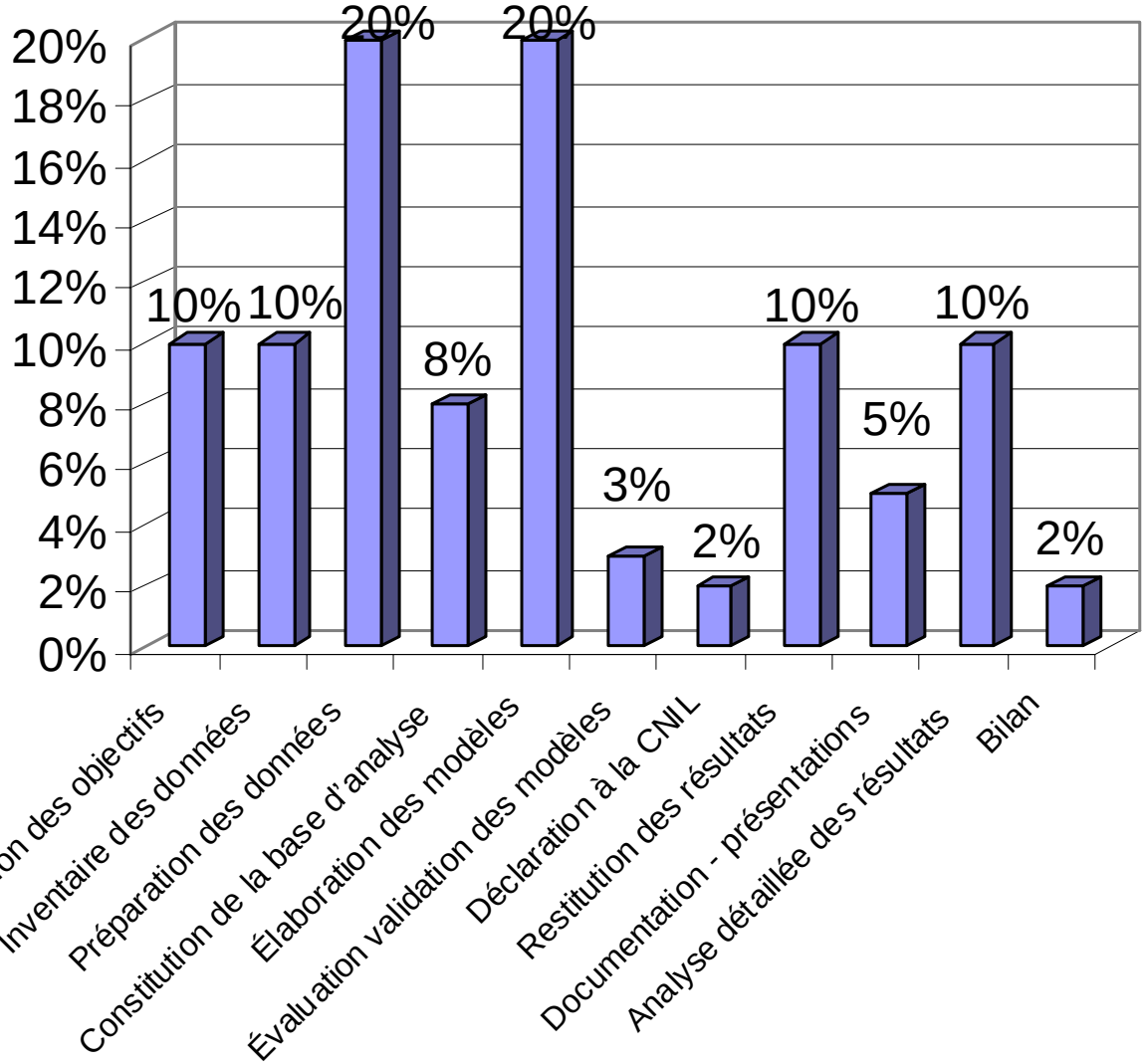
mois score	M – 1	M – 2	M – 3	M – 4	M – 5	M – 6	M – 7	M – 8	M – 9	M – 10	M – 11	M – 12
1												
2												
3												
…												
TOTAL												

Suivi du score : matrice de transition

Permet de contrôler la stabilité du modèle

score M–1 ➔ 🡻 score M–2	1	2	3	…
1				
2				
3				
…				

Répartition de la charge d'étude



Coûts et gains du data mining

Coûts

- Distinguer : coûts du data warehouse et du data mining
- Investissement initial dans un DW : environ 1 M€
- Dépend des choix techniques :
  - machine dédiée au décisionnel ou non
  - progiciels d'alimentation de l'entrepôt ou développements « maison »
  - modèle de données acheté ou élaboré en interne
- Coûts du data mining très inférieurs :
  - coûts humains inférieurs car équipe dédiée au DM + petite
  - coût des logiciels entre 2 k€ (sur PC mais avec plusieurs algorithmes performants) et 200 k€ (sur gros systèmes, avec une architecture informatique de production)
  - existence de logiciels libres : R

Le retour sur investissement

- Le RSI est difficile à évaluer :
  - les gains proviennent du data mining mais aussi d'une bonne communication, d'un marketing efficace, de commerciaux motivés
  - le DM n'est qu'une brique dans le marketing de bases de données (exemple du crédit préaccordé)
- Le RSI vient de :
  - l'augmentation des taux de réponse des actions marketing
  - l'augmentation de la productivité des commerciaux
  - la meilleure utilisation des canaux
  - la fidélisation des clients
  - la réduction des impayés...
- On peut tenter de l'estimer avec un échantillon témoin

Exemple de calcul de RSI

		ciblage tradit.	ciblage DM
A	nombre de clients ciblés	30 000	15 000
B	coût de chaque mailing	1 €	1 €
C	coût de chaque relance téléphonique	5 €	5 €
D	coût total (= A x (B + C))	180 000 €	90 000 €
E	nombre de nouvelles souscriptions	1 000	1 500
F	taux de souscription (= E / A)	3,33 %	10 %
G	coût par souscription (= D / E)	180 €	60 €
H	chiffre d'affaire annuel par souscription	150 €	175 € (montants souscrits supérieurs)
I	CA total annuel (= H x E)	150 000 €	262 500 €
	RSI (= I / D)	83 %	292 %

RSI d'un score d'attrition

A	coût d'acquisition d'un nouveau client	150 €
B	rentabilité annuelle des partants	450 €
C	temps d'activation d'un client	0,5 an
D	perte occasionnée par un départ (= A + (B x C))	375 €
E	coût de fidélisation d'un « partant » détecté	50 €
F	nombre total de clients	1 000 000
G	nombre de partants annuels	80 000
H	taux d'attrition = (G / F)	8 %
I	nombre de « partants » détectés (à tort ou à raison)	40 000
J	coût total de la fidélisation (= E x I)	2 000 000 €
K	nombre de vrais partants retenus	8 000
L	pertes évitées (= D x K)	3 000 000 €
	gain total net (= L – J)	1 000 000 €

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 prédilection...)
- 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

« Vendre » le data mining

- Les commerciaux et analystes peuvent voir une mise en cause de leur savoir-faire
- Il faut les convaincre que le scoring ne fournit qu'**une aide à la décision**, et non la décision elle-même
  - laquelle reste toujours leur prérogative, comme l'exige la loi « Informatique et libertés »
- Il faut aussi les convaincre de **bien alimenter les bases de données marketing**
  - notamment en ce qui concerne les retours des campagnes : refus d'achat
- Ils doivent être sensibilisés au gain de productivité et de sécurité qu'ils peuvent attendre du scoring

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

« Aucun *a priori* n'est nécessaire »

- Les **techniques prédictives** requièrent un *a priori* :
  - puisqu'il faut choisir une variable cible, soigneusement définie
- Certaines **techniques descriptives**, telle la classification, peuvent être mises en œuvre sans savoir quelles seront les classes obtenues, ni même quel est le nombre pertinent de classes
- Mais :
- Le résultat de la classification est influencé par le choix des données et de leur codage en entrée de l'algorithme :
  - > il est donc impossible d'être totalement neutre même dans une classification

« On n'a plus besoin de spécialistes du métier »

- Le concours des spécialistes (métier, marketing) est indispensable dans plusieurs phases :
  - la **définition des objectifs**
    - par exemple, avant l'élaboration d'un score de risque, il convient de s'entendre sur la définition précise d'un *risque*
  - le **recensement des données** utiles et légalement utilisables, données brutes et données composées
    - il est intéressant de connaître les indicateurs considérés comme pertinents par les spécialistes
  - l'**analyse des résultats**
    - le spécialiste métier peut, au vu des 1ers résultats, dire s'ils paraissent triviaux, nouveaux et intéressants à creuser, ou surprenants et très suspects, auquel cas il faudra vérifier la validité des données et de la méthodologie utilisée

« On n'a plus besoin de statisticiens »

- Dans une étude de data mining, la partie la + longue et la + déterminante est le **travail des données**
  - elle ne peut être effectuée qu'au vu d'analyses statistiques permettant de vérifier la fiabilité des données, leurs distributions, leurs corrélations... et de réaliser les mises en forme de données préalables ; ces opérations ne seront pas réalisées à l'identique pour toutes les méthodes de DM
  - certaines méthodes nécessitent un échantillonnage préalable
- Dans les méthodes prédictives, il faut prendre garde de ne pas inclure parmi les variables explicatives des variables corrélées *par définition* à la variable cible. Il faut se méfier du phénomène de **sur-apprentissage**
- Le **paramétrage fin** des algorithmes peut avoir une grande incidence sur les résultats obtenus

« Le data mining permet de faire des découvertes incroyables »

- Les règles mises à jour par le data mining sont rarement incroyables : elles font souvent intervenir des variables considérées comme discriminantes par les spécialistes, d'une façon conforme au bon sens.
- **Où réside donc l'apport du data mining ?**

> Dans le fait qu'il existe des milliers de combinaisons, conformes au bon sens, de variables *a priori* discriminantes dans une problématique donnée…

> … et que le data mining permet de détecter LA meilleure combinaison possible (ou l'une des meilleures), avec, pour chacune de ces variables X, la meilleure valeur précise *n* à tester (« si $X \leq n$, alors… »)

« Le data mining est révolutionnaire »

- Le **data mining englobe la statistique et l'analyse des données traditionnelle**, dont il ne diffère que par les points suivants :
  - certaines techniques de DM n'appartiennent qu'à lui, comme les réseaux de neurones et les arbres de décision
  - le nombre d'individus étudiés est souvent plus important en DM, où l'optimisation des algorithmes est importante
  - le DM fait moins d'hypothèses contraignantes sur les lois statistiques suivies
  - les modèles en DM sont plus souvent des ensembles de règles locales que des modèles globaux
  - le DM recherche parfois plus la compréhensibilité des modèles que leur précision

« Il faut utiliser toutes les données disponibles »

- **Un algorithme de DM est-il d'autant + efficace qu'il a + de données en entrée ? NON !**
- Les données non fiables ou mal renseignées perturbent tous les algorithmes
- La présence d'individus hors-norme (« outliers ») perturbe les modèles linéaires
- Les données avec des modalités aux effectifs irréguliers affectent les analyses factorielles
- Les données peu discriminantes ou colinéaires diminuent le pouvoir prédictif d'une analyse discriminante ou d'une régression logistique
- Les données redondantes peuvent affecter une classification
- Les données trop nombreuses affectent les réseaux de neurones

« Il faut toujours échantillonner »

- Un **bon échantillonnage est toujours délicat à réaliser**, et nécessite une bonne connaissance de la population
  - difficile à avoir, surtout avec les populations instables que sont les clientèles
- Exemple d'inconvénient induit par l'échantillonnage :
  - un écart de distribution d'une variable dans l'échantillon d'apprentissage par rapport à la population totale, peut produire des écarts importants dans les résultats
- Autre contre-indication au recours à l'échantillonnage : la **recherche de phénomènes rares** (typologies de fraude) ou de segments étroits de clientèle

« Il ne faut jamais échantillonner »

- Certaines techniques de data mining, les **techniques de prédiction inductives** (arbres de décision, réseaux de neurones à rétropropagation supervisée...), **imposent le recours à l'échantillonnage**
  - puisqu'elles procèdent par élaboration d'un modèle à partir d'une partie de la population,
  - modèle ensuite testé sur une autre partie de la population
- Il peut aussi être souhaitable de travailler sur un échantillon de la population, si celle-ci est très grande, afin de **limiter des temps de calcul** prohibitifs
- « A powerful computationally intense procedure operating on a subsample of the data may in fact provide superior accuracy than a less sophisticated one using the entire data base ». Jerome H. Friedman (1997)

7 pistes d'amélioration sur le MBDD

- Mémoriser les résultats des campagnes commerciales
- Sensibiliser les commerciaux à l'importance des données saisies
- Acheter des données externes (INSEE, Consodata...)
- Compléter ces données par des enquêtes auprès de clients et de commerciaux
- Préciser la définition et améliorer le calcul de données stratégiques : rentabilité, fidélité...
- Créer de nouvelles variables synthétiques pertinentes
- Augmenter la profondeur de l'historique des données.

Le recours au consulting

Internalisation ou externalisation 1/2

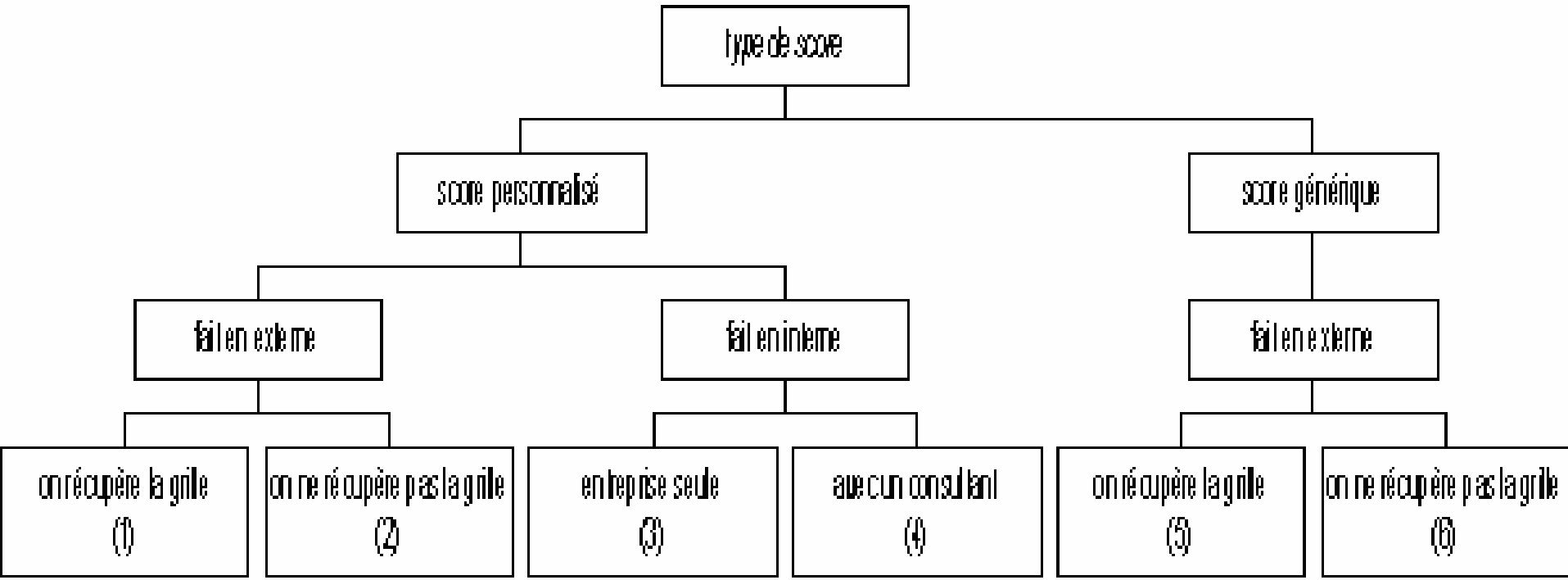
- Soit l'entreprise internalise l'activité de data mining, éventuellement avec l'aide de consultants spécialisés
- Soit elle externalise totalement cette activité, en fournissant ses fichiers de données à des prestataires spécialisés (les « credit-bureaux » pour la banque), ceux-ci lui restituant ses 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 2/2

- 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

Scores personnalisés et génériques



Comparatif des diverses solutions

	Performance du score	Transfert de compétence	Pérennité à court terme du score	Pérennité à long terme du score	Rapidité d'obtention
(1)	+	–	+	–	–
(2)	+	–	– (sert 1 seule fois)	–	–
(3)	+	–	+	+	–
(4)	+	+	+	+	–
(5)	–	–	+	–	+ + (30 jours)
(6)	–	–	– (sert 1 seule fois)	–	+