



Des solutions de data mining pour résoudre des problèmes d'estimation du risque

12 juin 2009

Association des Statisticiennes et Statisticiens du Québec

1 - Data mining vs Statistique

- *Pourquoi les statisticiens nous détestent-ils ?*
- S. Imberman
- *« Ceux qui ignorent la statistique sont condamnés à la réinventer. »*
- B. Efron
- Français: exploration, prospection, découverte ou **forage de données.**



Définition de forage de données

Technique de recherche et d'analyse de données qui permet de dénicher des tendances ou des corrélations cachées parmi des masses de données, ou encore de détecter des informations stratégiques ou de découvrir de nouvelles connaissances en s'appuyant sur des méthodes de traitement statistique.

Source: grand dictionnaire terminologique de l'OLF.



Méthodologies du forage de données (J. Friedman)

Méthodologie	Domaine de développement
Reconnaissance de formes	Informatique, Génie
Gestion de bases de données	Informatique
Réseaux de neurones	Psychologie, Informatique, Génie
Algorithmes d'apprentissage	Informatique, Intelligence artificielle
Modèles graphiques (réseaux bayésiens)	Informatique, Intelligence artificielle
Algorithmes génétiques	Informatique, Génie
<i>Chemometrics</i>	Chimie
Visualisation de données	Informatique, Calcul scientifique (Statistique)

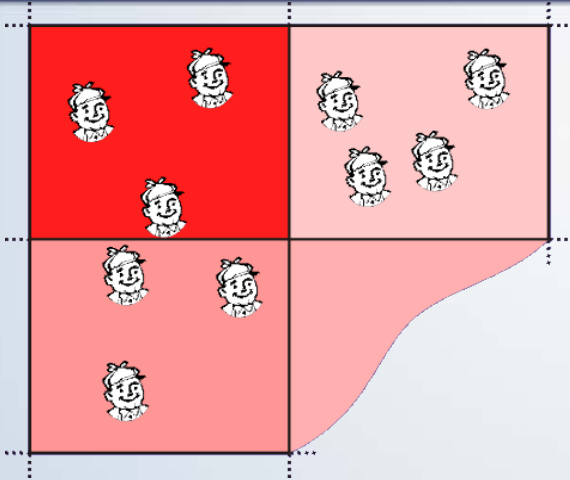


Défis du forage de données

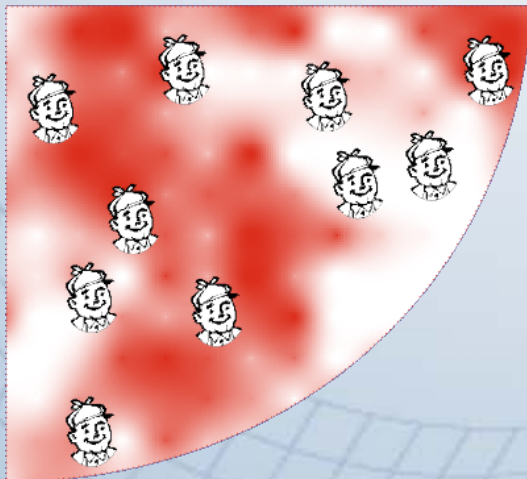
- Calculs hors ligne (*offline*) pour accélérer la performance (*en ligne*)
- Parallélisation
- Gestion de la mémoire vive
- Interfaces entre algorithmes de forage de données et SGBD (entrepôts de données)
- ...
- Signification de $O(n)$:
 - statisticien: vitesse de convergence
 - « foreur » : temps de calcul, espace mémoire



2 - Estimation du risque en assurance automobile



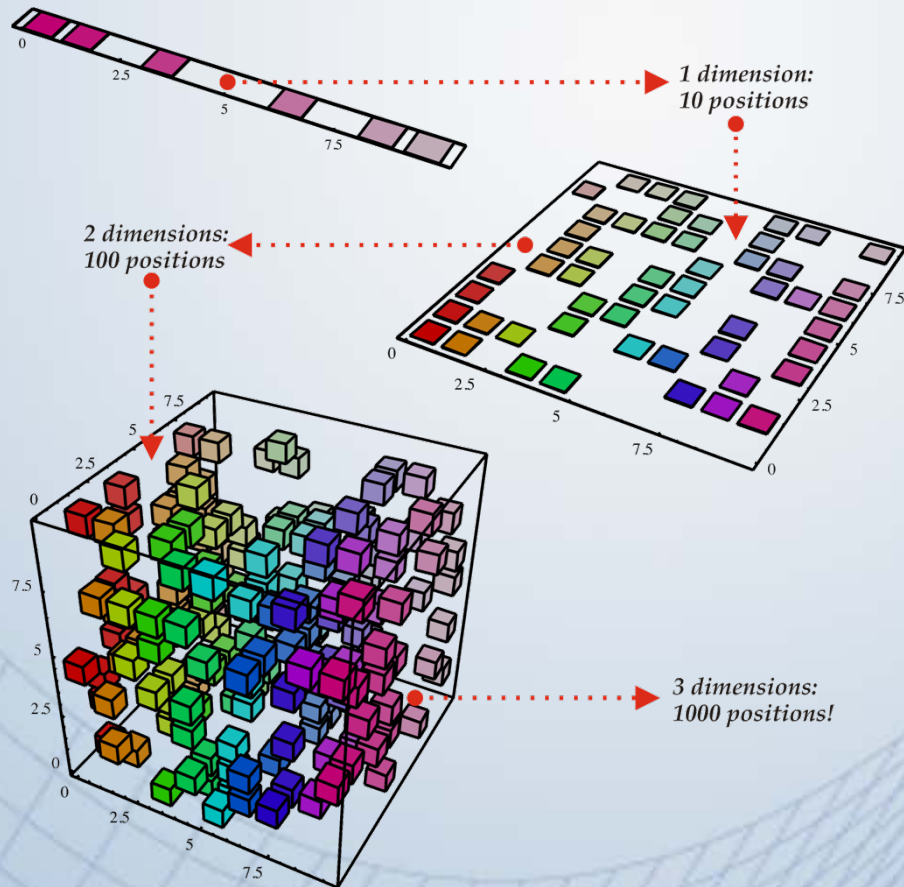
- primes par sous-groupes
- ensemble réduit de critères
- extraction **difficile** des relations de dépendance



- primes individualisées
- nombreux critères
- extraction **automatique** des relations de dépendance



Malédiction de la dimensionalité



Croissance exponentielle du nombre de profils différents, en fonction du nombre de facteurs explicatifs.



Plan de répartition des risques

	Québec	Ontario
Pourcentage	10% volume	5% unités
Dommages corporels	exclus (SAAQ)	inclus
Taux de dépenses	25%	max. 30%
Portion cédée	100%	85%
Cessions profitables	SP > 75%	SP > 70%



Plan de répartition des risques

- Le marketing n'est pas impliqué
 - assurés ne sont pas affectés
 - courtiers ne sont pas affectés
- législation plus souple (IGIF, FSCO)
- aspects stratégiques sans impact
- purement analytique
- *zero-sum game* entre les assureurs
- solution: *utiliser toutes les ressources disponibles pour développer les meilleurs modèles prédictifs*



Images (profils)

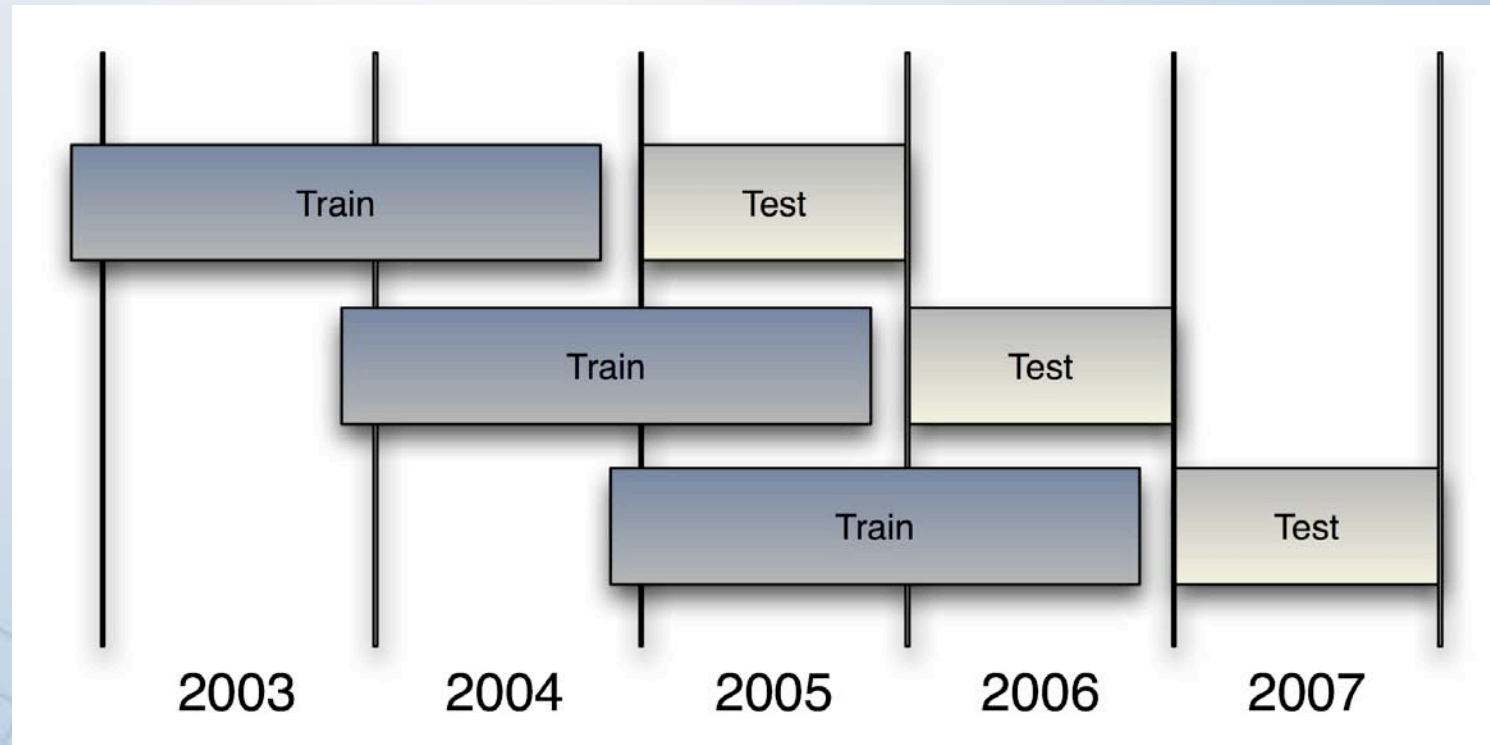
Key	Start Date	Effective End Date	Age	Car Type	Color	Other Data	Claim
ABCD1234	2008/03/15	2008/06/30	23 y/o	Acura TL 2005	Red with Side Flares	<input type="radio"/> <input type="radio"/> <input type="radio"/>	\$ 34,000

- Images:
 - dates de début et de fin
 - valeurs pour toutes les variables explicatives
 - montants des sinistres (s'il y a lieu)
- lien entre données de polices (images) et données de sinistres (variable dépendante)

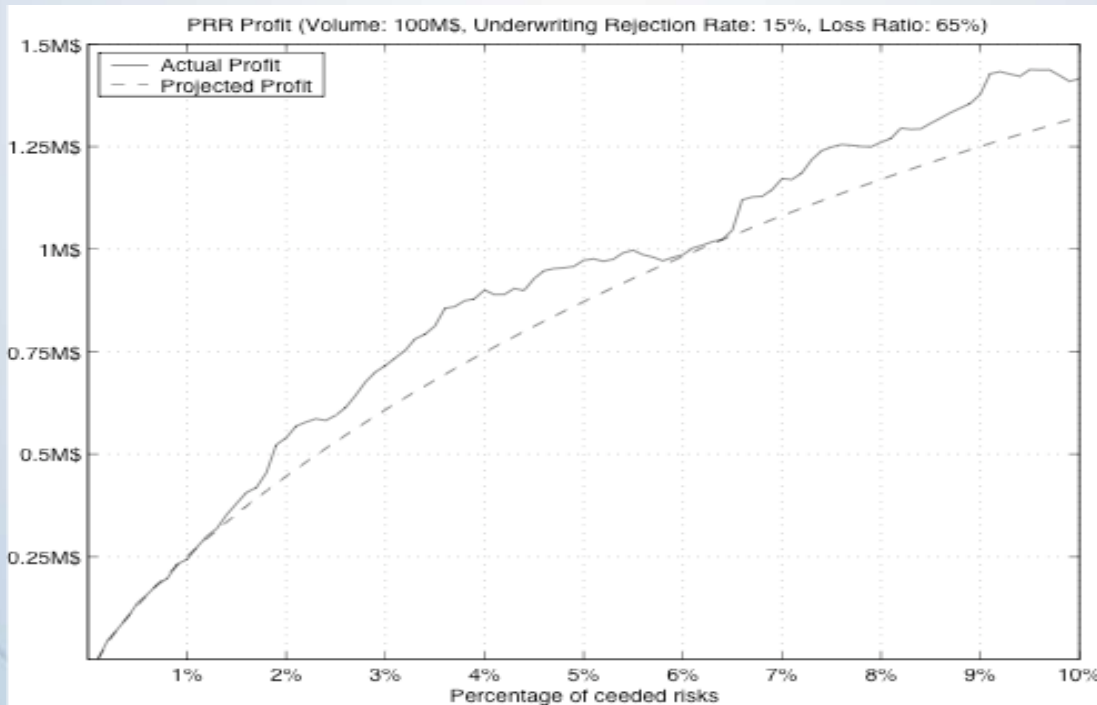


Sélection de modèle: validation séquentielle

Doit être un juste reflet de l'utilisation future



Résultats



Facteurs:

- *volume*
- *ratio SP*
- *modèle en place*
- *tarification*

3 - Valeur Économique du Client

Vue orientée produit

Attirer des clients

Transactions

Produits

Profits des lignes de produits

(product death spiral)

Vue orientée consommateur

Conserver des clients

Relations

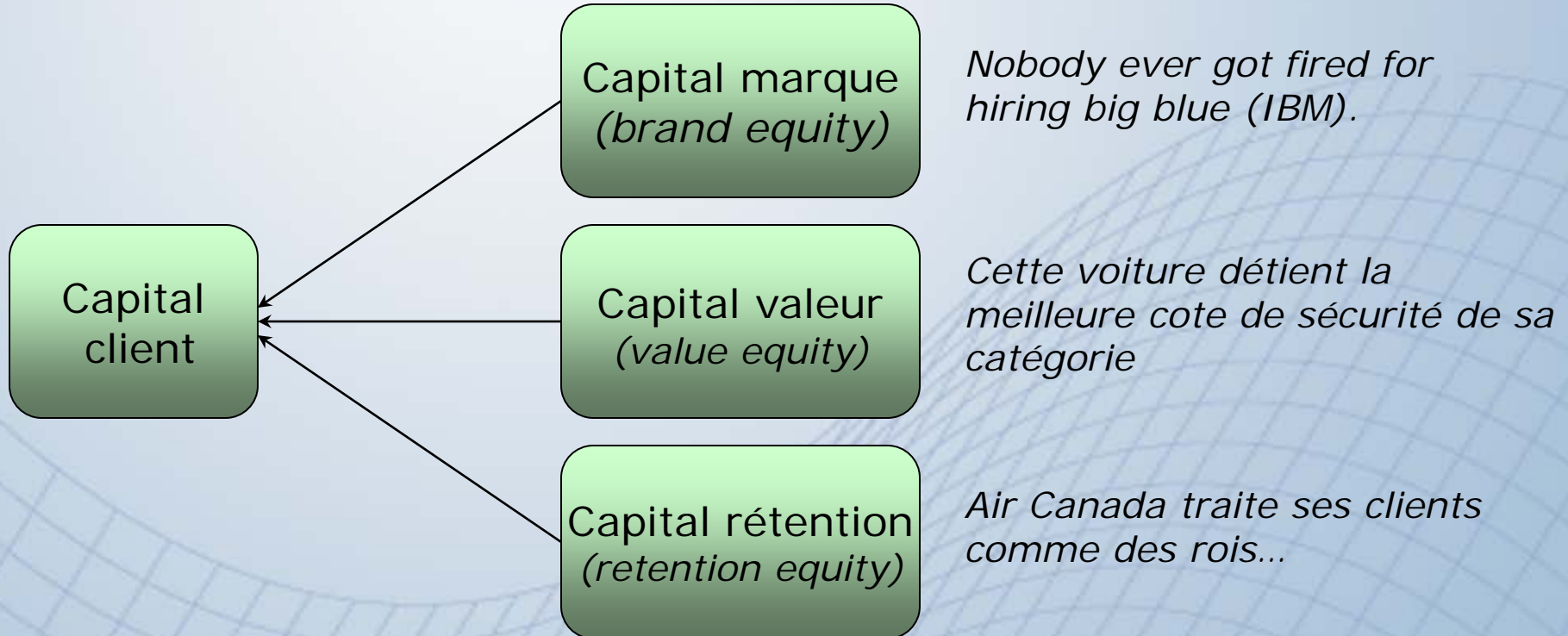
Services

Gestion du capital client



Capital client

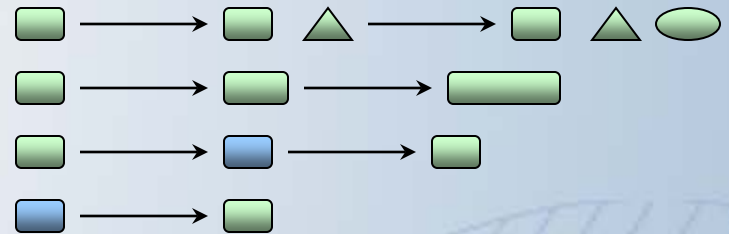
Pour une compagnie, c'est le total des valeurs économiques de ses clients (actuels et futurs)



Définitions du Capital client

Niveau de raffinement:

1. ventes croisées (*cross-selling*)
2. vente incitative (*upselling*)
3. clients peu fidèles (*switchers*)
4. nouveaux entrants



Facteurs difficiles à quantifier:

1. réaction de la compétition
2. bouche à oreille
3. marketing et publicité
4. frais de gestion de la relation
5. frais d'utilisation décroissants



Capital rétention

valeur présente des profits futurs qui découlent des transactions réalisées avec le client

horizon
(e.g. 5ans)

profits dans le temps
(frais fixes?)

$$CLV = \sum_{t=1}^T v^t s(x, t) p(x, t)$$

facteur d'escompte
(taux faible \Rightarrow rétention importante)

taux de rétention
(forage de données)



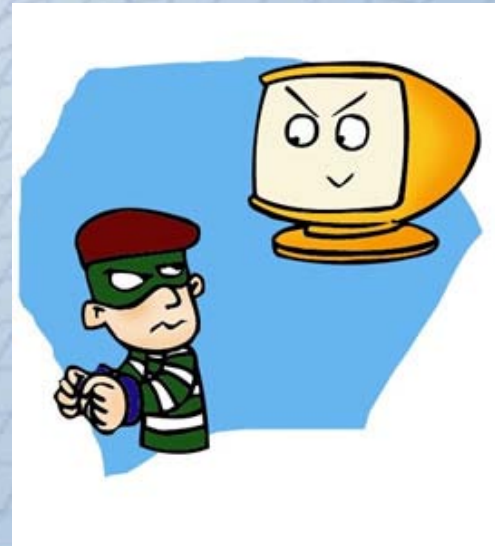
Problèmes ouverts

- Modéliser les variations dans la quantité achetée
- Modéliser les interdépendances entre produits:
 - matrice de transition de l'état du client (quantité = 0/1)
 m produits $\Rightarrow 2^m$ états $\Rightarrow 2^{2^m}$ éléments
 - achats multiples...
- Modéliser les « infidèles »
 - un prospect a-t-il plus/moins de chances d'accepter une offre qu'une personne tirée au hasard dans la population ?
 - les frais d'acquisitions sont-ils les mêmes à chaque retour du client ?



4 – Détection de fraudes de cartes de crédit

- L'une des applications les plus profitables des réseaux de neurones
- 85% des cartes en Amérique protégées par Fair Isaac (HNC)
- Milliards de transactions disponibles pour entraîner les modèles
- Défi d'efficacité en temps réel: 30 sec.



5 – Horaires

- Planifier les horaires des employés en fonction
 - des contraintes individuelles
 - de la convention collective
 - *de l'absentéisme*
 - *de l'achalandage prévu*



Conclusion

- Le forage de données est issu de la statistique
- La statistique pourrait se réapproprier ce domaine:
 - Informatique
 - Mathématiques
 - Mentalité
- Progrès réalisés mais défis encore nombreux



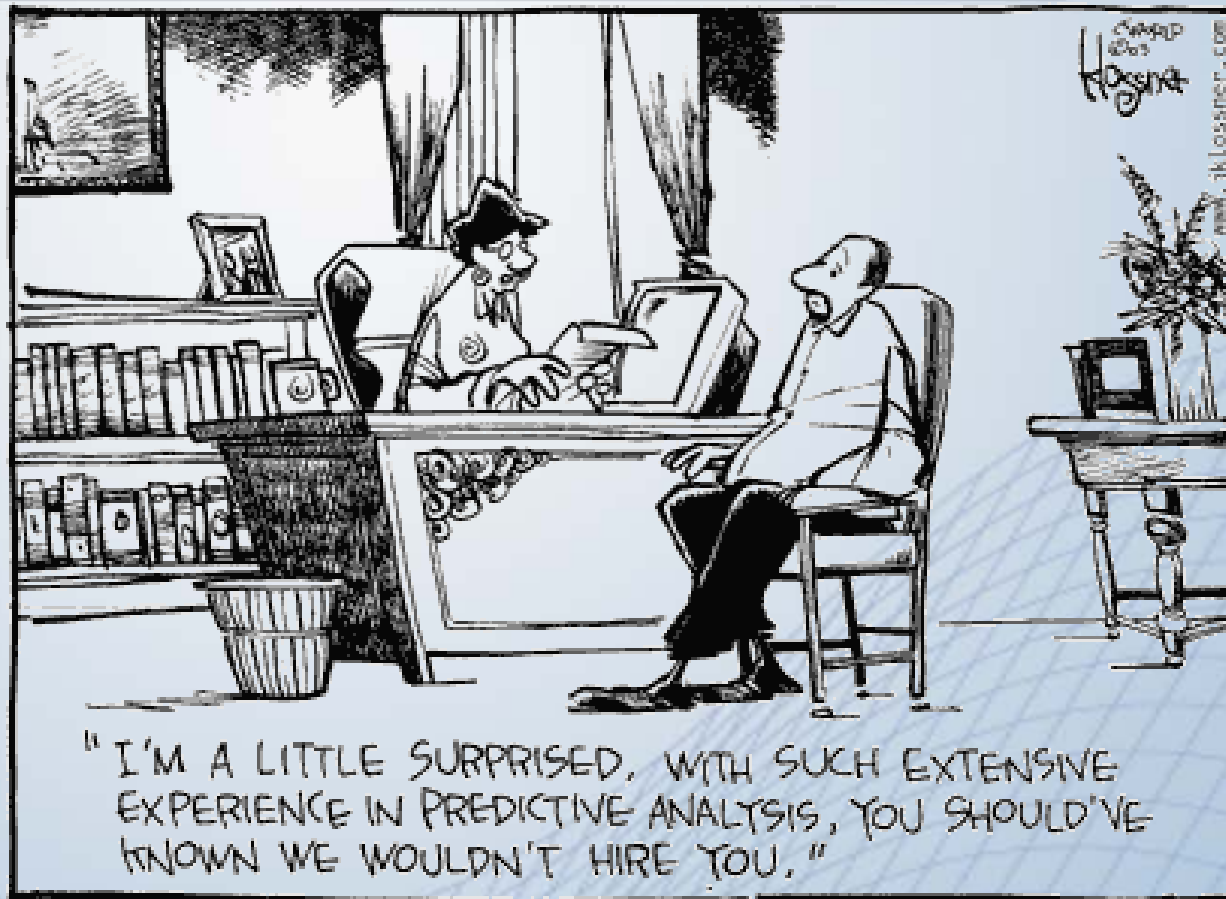
Questions ?

« Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. »

- H.G. Wells



Questions ?



Remember: *All models are wrong, some are useful*



Insurance business

Loss Ratio (LR)

Claims / Premiums (ca. 60%)

Book Volume

Total of all premiums

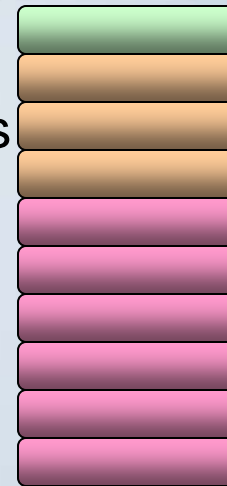
Retention

- opposite of churn (*taux d'attrition*)
- fixed time window
- measured in annual terms (ca. 90%)

profit

expenses

claims



Québec
personal lines
auto insurance

None of these figures is prospective !

Goal

- identify the long term value of the clients: CLV



Steps towards a CLV model

1 Choose definition

- upselling: can you trace a customer's business ?
- cross selling: is there a unique key across products ?
- switchers: can you match profiles through time ?
- new entrants: what do you know of your prospects ?

2 Make assumptions

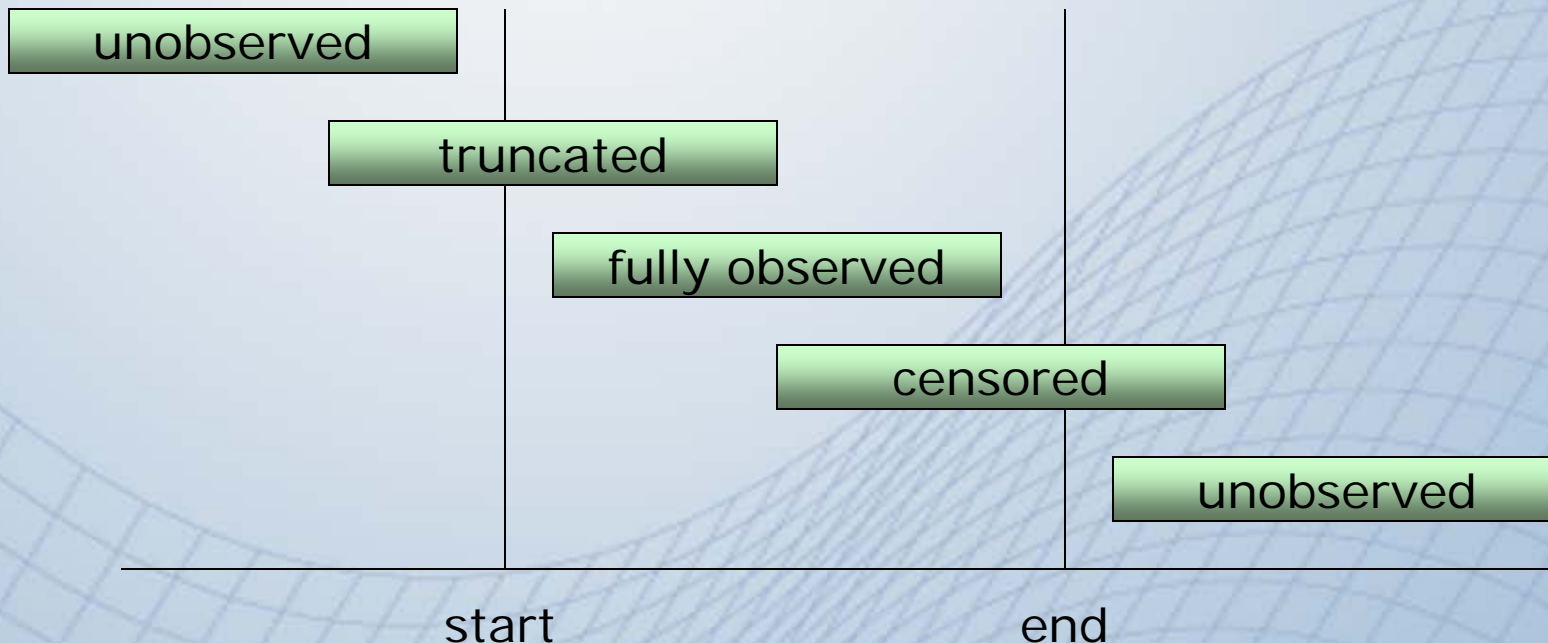
- for each client, project
 - profits or
 - revenues and expenses
- important: recognize acquisition expenses
- discounting factor based on cost of capital
- retention models can be based on survival analysis
- horizon, e.g. 5 years



Steps towards a CLV model

3 Build database for survival analysis

- initial profile with length of stay
(SAS procs: lifetest, lifereg, phreg)
- data "plumbing" : cleaning and encoding



Steps towards a CLV model

4 Modeling

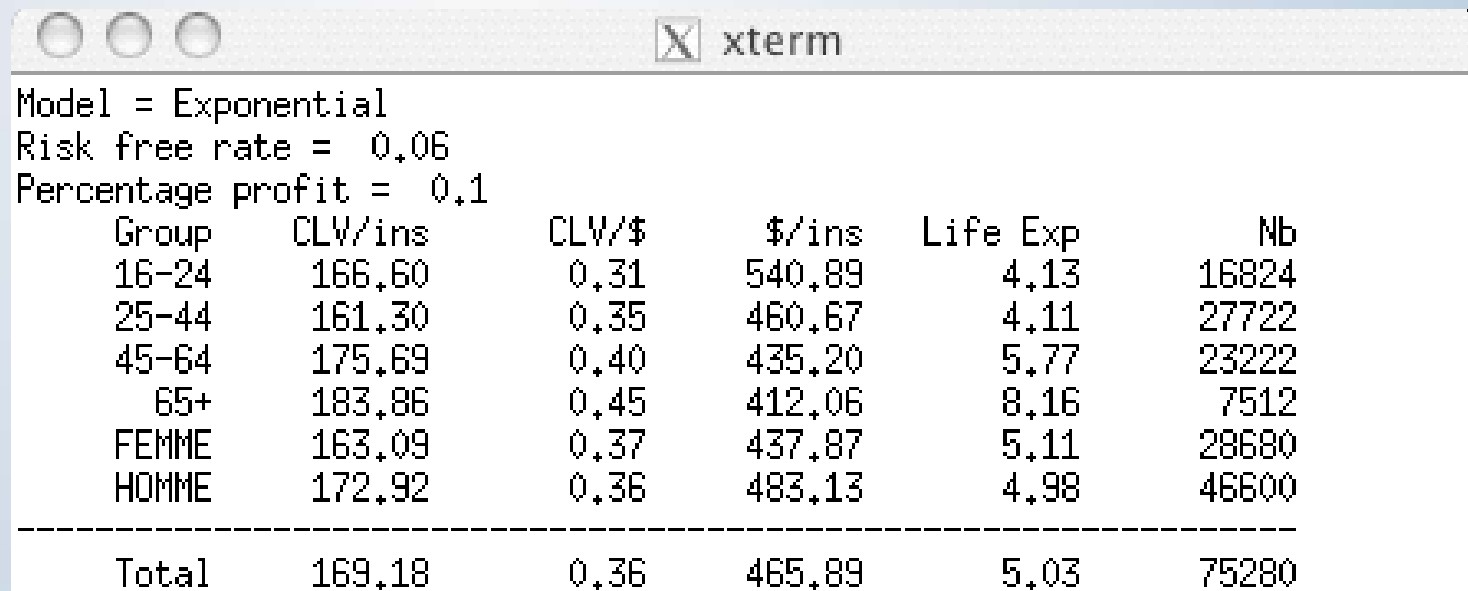
- design modeling experiment
- train models
- select best performing models

5 Enjoy!

- CLV-based comparison of market segments
- CLV-based valuation of retention increases
- Prioritization of outgoing agent calls
- Prioritization of marketing offers



Example of output

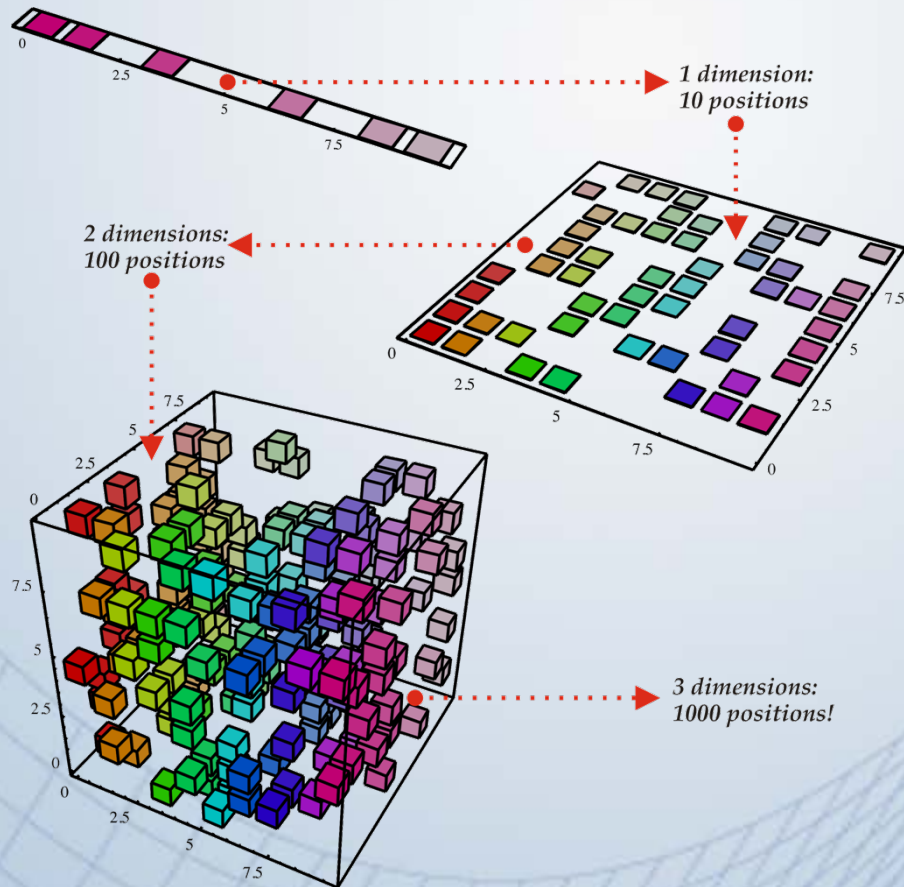


```
Model = Exponential
Risk free rate = 0.06
Percentage profit = 0.1
```

Group	CLV/ins	CLV/\$	\$/ins	Life Exp	Nb
16-24	166.60	0.31	540.89	4.13	16824
25-44	161.30	0.35	460.67	4.11	27722
45-64	175.69	0.40	435.20	5.77	23222
65+	183.86	0.45	412.06	8.16	7512
FEMME	163.09	0.37	437.87	5.11	28680
HOMME	172.92	0.36	483.13	4.98	46600

Total	169.18	0.36	465.89	5.03	75280

Data Mining

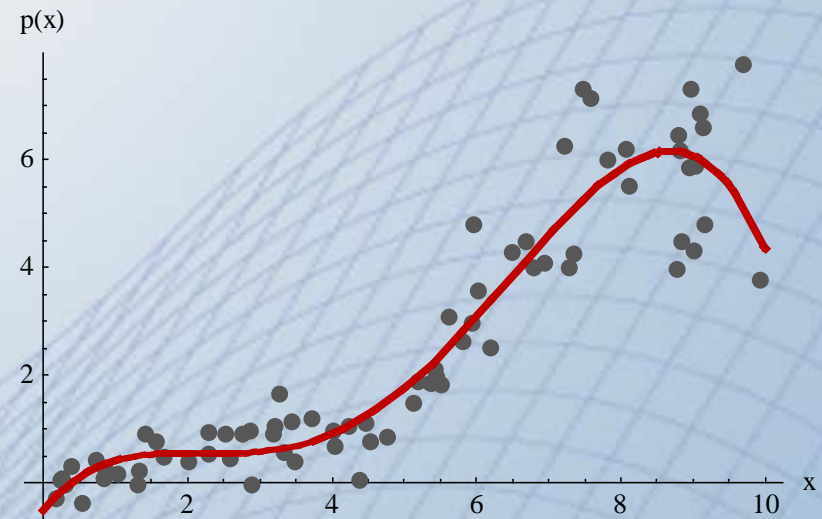
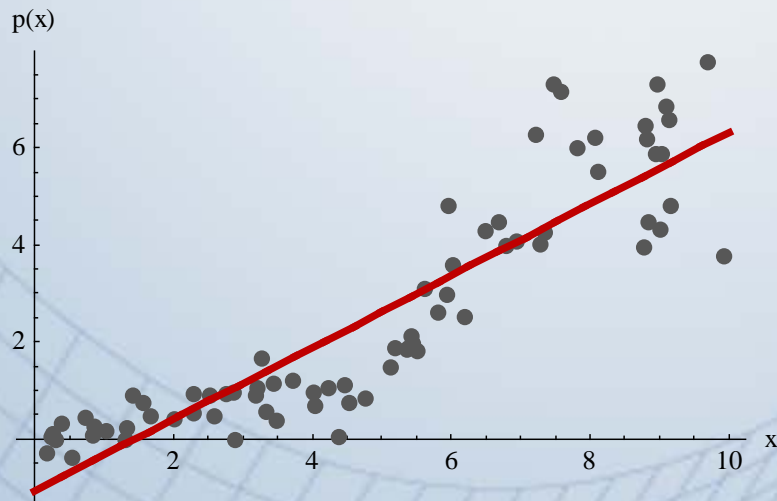


Curse of dimensionality:
The number of potential different customer profiles grows exponentially with the number of factors



Data Mining

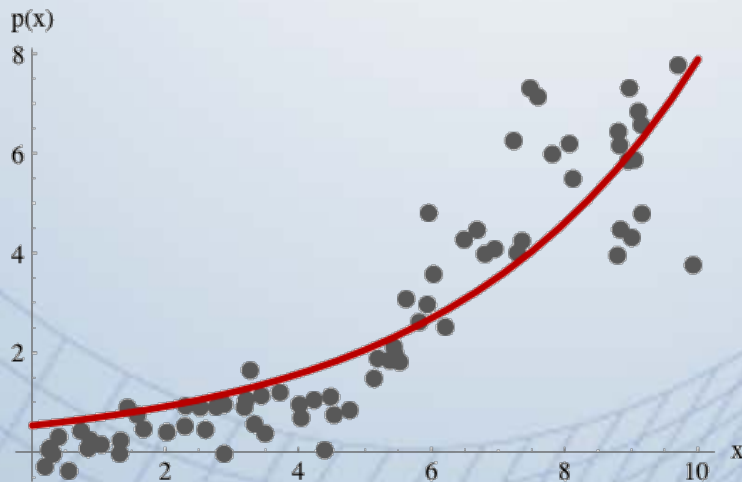
A complex reality commands the use of models with greater capacity, i.e., more flexible models.



Popular models

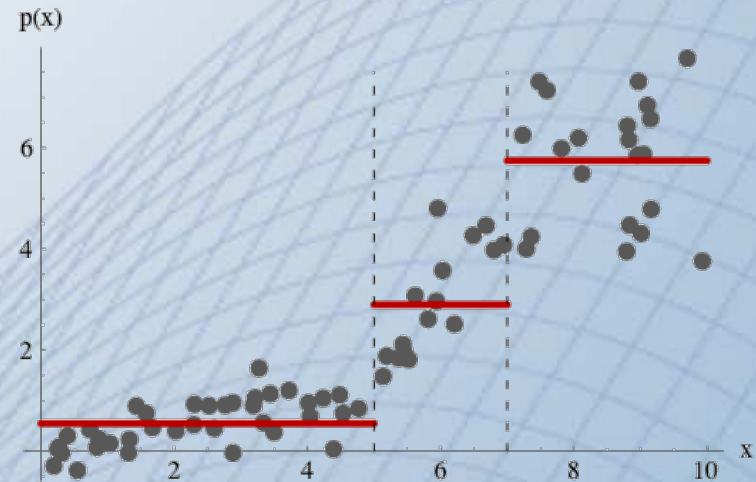
Generalized linear (GLMs)

$$y = f \left(\alpha_0 + \sum_{i=1}^n \alpha_i x_i \right)$$



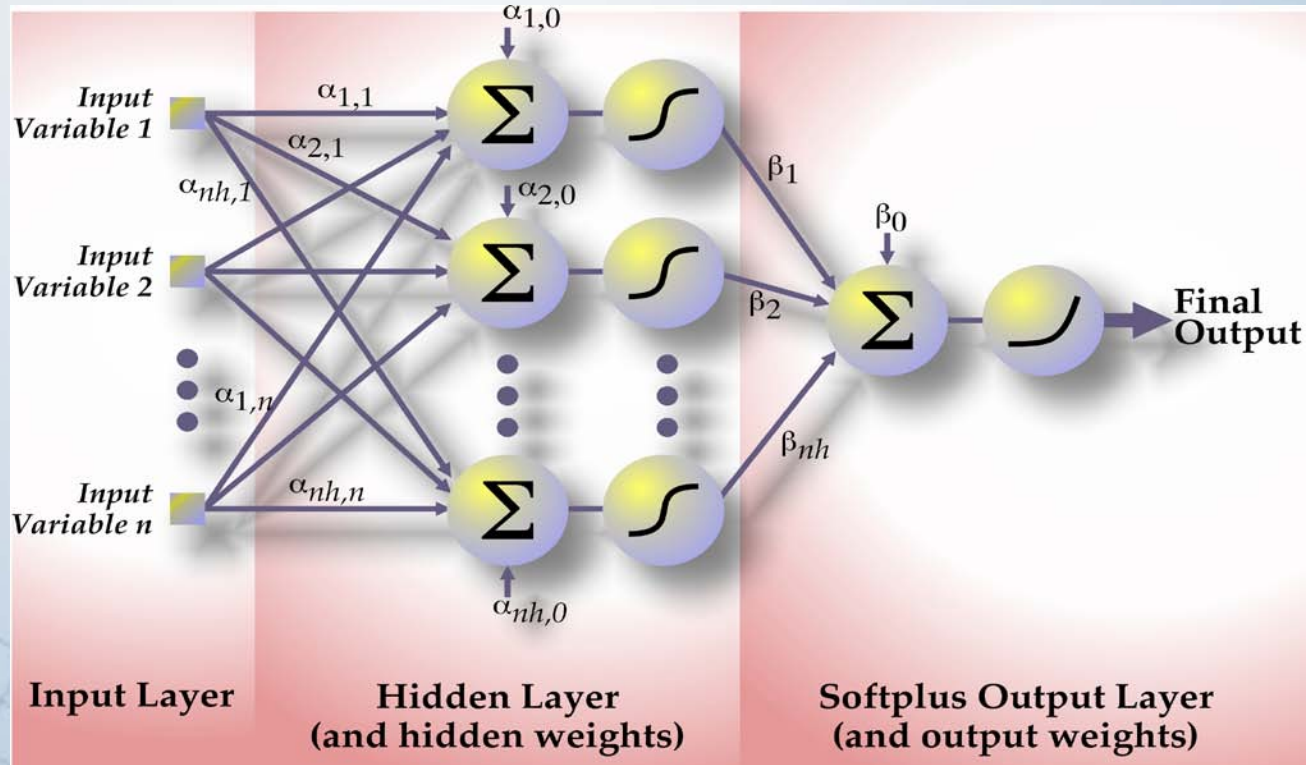
Decision trees

$$y = \sum_{j=1}^{n_l} I_{\{x \in l_j\}} \left(\alpha_{j,0} + \sum_{i=1}^n \alpha_{j,i} x_i \right)$$

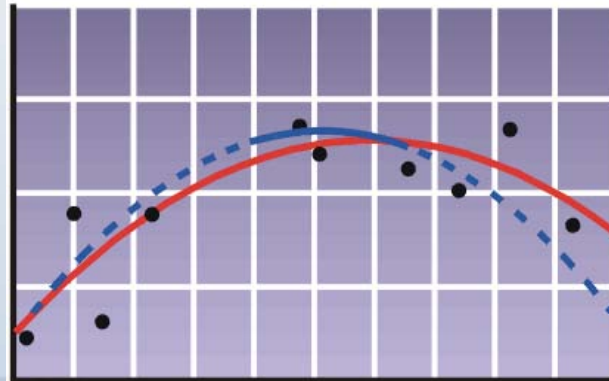
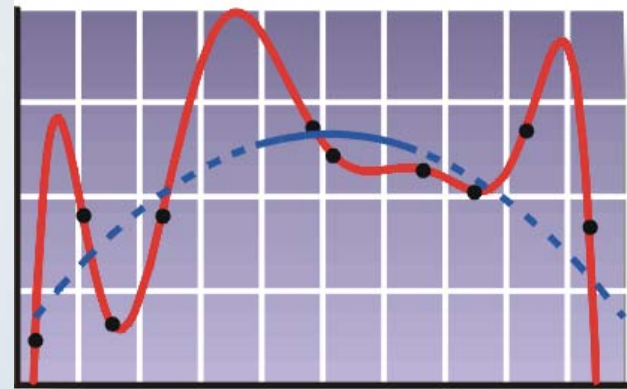
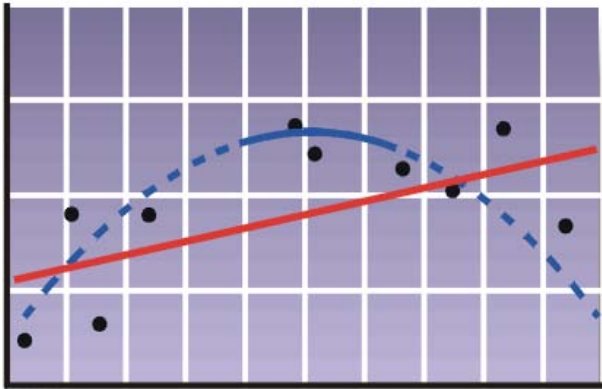


Neural networks

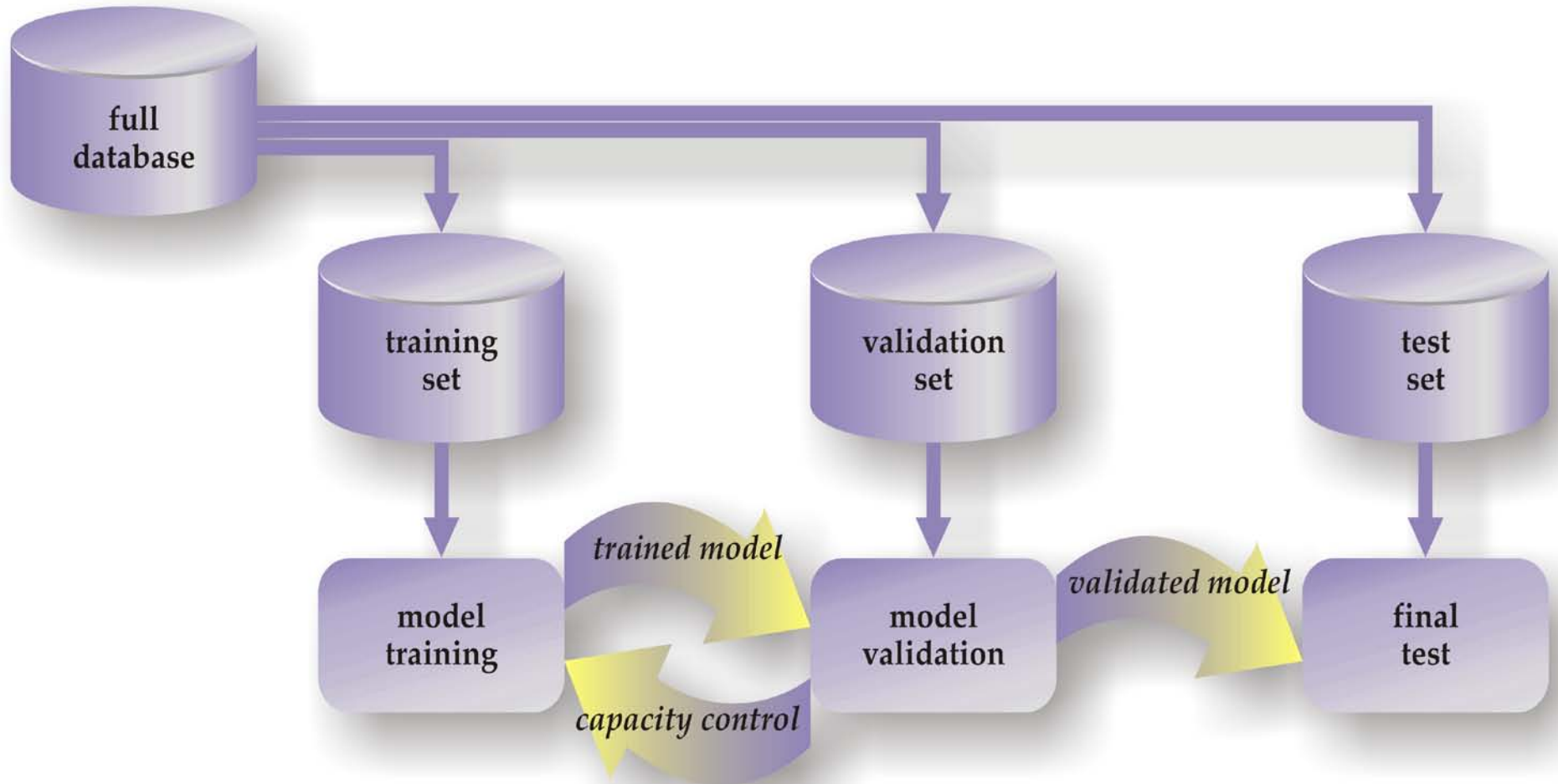
$$h_j = f_1 \left(\alpha_{j,0} + \sum_{i=1}^n \alpha_{j,i} x_i \right) \quad y = f_2 \left(\beta_0 + \sum_{j=1}^{nh} \beta_j h_j \right)$$



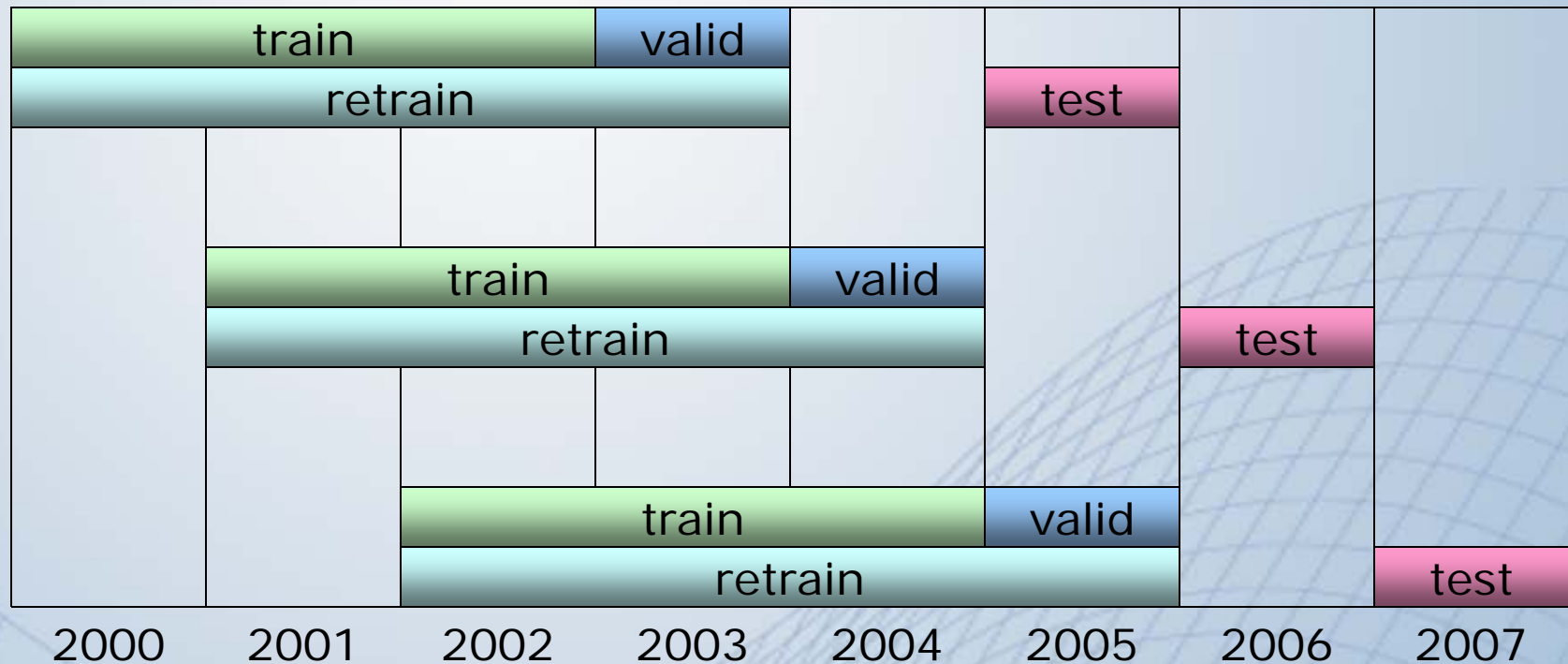
Model Selection



Methodology



Model Selection: Sequential Validation



Data Mining in Insurance

- Data
 - easier if data warehouse
 - otherwise, *projects are costly*
- Models
 - well-known and popular: decision trees, glms
 - little used: neural networks
 - all are available in SAS-EM
- Methodology
 - *very little attention given to model selection !*



Data mining and Insurance Ratemaking



Conclusion

- Importance of CLV
 - part of a shift towards customer-oriented practices
 - can be used internally and for marketing purposes
- Future work
 - Project revenues & expenses rather than profits
 - Out-of-sample tests
 - Work on graphical user interface
 - Develop multiline models
 - Impact of changes on retention

