FORMCO/MAP & INA-PG

Cycle de conférences « Jean-Pierre Fénelon »
sur l'analyse multidimensionnelle des données statistiques

L'industrialisation des analyses Besoins, outils & Applications.

Françoise Soulié Fogelman
VP Strategic Business Development, KXEN

Séminaire Jean-Pierre Fénelon
« Analyse des données »
INA-PG, Paris
Mercredi 29 novembre, 14H00-17H00

Agenda

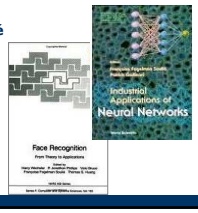
- Le data mining
- Les besoins
- L'usine à modèles
- La mise en œuvre de KXEN
- Quelques exemples
- Conclusion

Un peu d'histoire

- **L'approche statistique**
  - Des statistiques classiques aux statistiques non paramétriques
- **L'approche cybernétique**
  - Des travaux de Wiener & von Neumann au perceptron de Rosenblatt
- **L'approche machine learning**
  - De l'intelligence artificielle, la reconnaissance des formes aux arbres de décision et aux réseaux de neurones
- **Le data mining aujourd'hui**
  - A hérité de tous ces travaux
  - Repose fondamentalement sur le paradigme de l'apprentissage
  - Utilise massivement l'informatique
    - Ordinateurs, bases de données, standards de développement …

Un peu d'histoire

- **Les années 89-95 : les réseaux de neurones**
  - En 1982, Hopfield introduit un modèle de mémoire associative
  - Le réseau multi-couches ou MLP
    - Est introduit en 1985-86 : Rumelhart, LeCun
    - Développement de la technique de « Gradient Back-propagation » et applications à des MLP complexes et efficaces
  - Les cartes topologiques et le Learning Vector Quantization
    - Techniques non supervisées publiées par Kohonen dans les années 80-90
  - Vapnik arrive aux Bell Labs en 1991
    - Ses théories sont appliquées aux réseaux de neurones
    - Il développe ensuite les SVM
  - De nombreux colloques ont présenté
    - Les applications des réseaux de neurones
      - Le colloque ICANN'95 en France
    - Les liens avec la Statistique
      - L'Ecole Modulad de 1996



Le data mining … aujourd'hui

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  using data management and statistics
  - Data captured by instruments
    Or generated by simulator
  - Processed by software
  - Scientist analyzes database / files

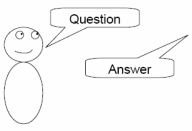
- Historically,
  Computational Science
  = simulation.
- New emphasis on
  informatics:
  - Capturing,
  - Organizing,
  - Summarizing,
  - Analyzing,
  - Visualizing

Courtesy Jim Gray, Microsoft Research.

D'après : Rakesh Agrawal, Microsoft Search Labs, KDD'06, Philadephia, August 2006

Le data mining … hier

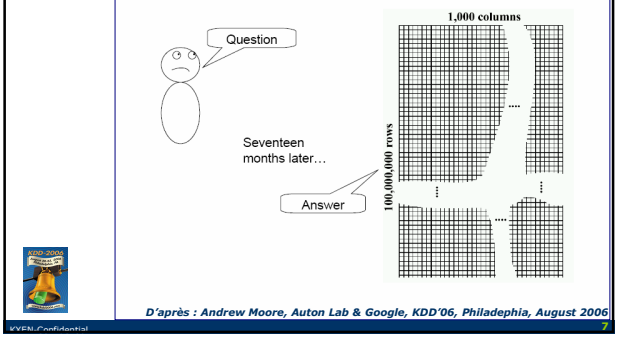
Data Analysis: The old days



Size	Ellipticity	Color
23	0.96	Red
33	0.55	Red
36		Green
40		
20		
48		

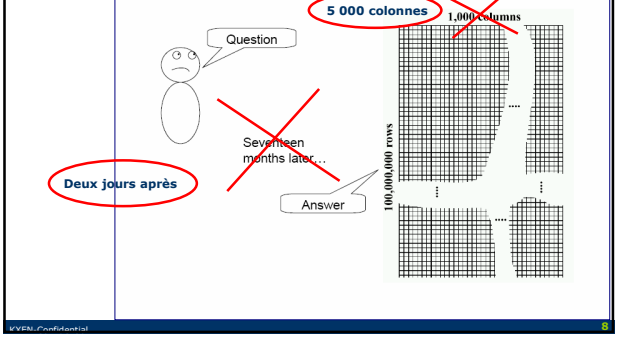
D'après : Andrew Moore, Auton Lab & Google, KDD'06, Philadephia, August 2006

Le data mining … aujourd'hui

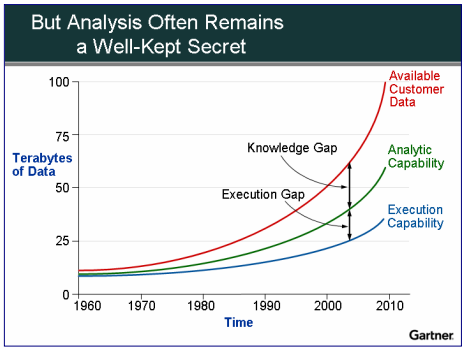


D'après : Andrew Moore, Auton Lab & Google, KDD'06, Philadephia, August 2006

Le data mining … aujourd'hui ?



Le data mining … reste une affaire d'expert



D'après : Gareth Herschel, Gartner, KXEN Users Group, San Francisco, November 2006

Agenda

- Le data mining
- Les besoins
- L'usine à modèles
- La mise en œuvre de KXEN
- Quelques exemples
- Conclusion

Le constat

- **Des sources de données nombreuses**
  - Le volume des données double chaque année (Forrester)…
  - Mais les sources sont nombreuses, non cohérentes
- **Des utilisateurs – et des besoins – nombreux**
  - Les utilisateurs souhaitent répondre par eux-mêmes à leurs questions
    - Sans dépendre d'experts
    - Sans être obligés d'en devenir eux-mêmes.
- **Des décisions de plus en plus nombreuses**
  - Mais la qualité des décisions prises dépend des analyses menées
- **La vitesse est un facteur clé pour la qualité des résultats**
  - Le délai entre la conception & la mise en production d'une analyse doit être aussi réduit que possible

Les données

Les sources de données sont très nombreuses
et doivent intégrer tous les canaux (y compris Internet)

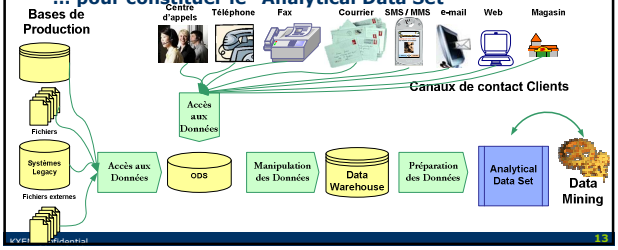
- **On peut commencer à travailler les données**
  - Sans avoir encore consolidé un datawarehouse
  - En constituant des bases thématiques
  - En acquérant éventuellement des données extérieures
    - Données INSEE
    - Données géo-marketing
    - Données comportementales …
- **Les analyses permettent alors de**
  - Obtenir des résultats exploitables rapidement
  - Analyser la qualité des données disponibles
  - Valider la valeur des données externes

… et ainsi de constituer un business case étayé par des premiers bénéfices

Les données

- Les données sont à la base du data mining
  - Pas de données, pas de modèle !
- Le processus de collecte de données est complexe : il faut
  - Identifier l'ensemble des sources de données
  - Mettre en place les mécanismes de collecte
  - Mettre les données en cohérence
  - Manipuler & transformer les données

… pour constituer le "Analytical Data Set"



Les données

- Préparation des données
  - Sélection des variables
    - Choisir les variables utiles
  - Définition de la cible
  - Les transformations "métier"
    - Champs calculés : produire de nouvelles variables à partir de variables existantes
      - Nb de jours entre l'émission de la facture et le paiement
      - Profit : prix d'achat – coût de fabrication
  - Codage : les transformations statistiques nécessaires pour un certain modèle
    - Changer les types de continu à nominal ou ordinal (binning ou regroupement de catégories)
    - Eclater une variable en plusieurs ou Regrouper plusieurs variables en une seule
      - Représentation d'une variable multi-catégorie
  - Évaluer la qualité des données pour déterminer
    - Les valeurs manquantes (blancs, espaces, nuls)
    - Les outliers
    - Les corrélations

Les données

Qualité des données

- Les données doivent être
  - Exactes
    - Valeurs correctes
  - Non redondantes
    - Doublons
  - Complètes : données "manquantes"
    - "missing-rate" d'une variable : combien d'observations ne l'ont pas
    - "filling-rate" d'une observation : combien de variables sont remplies
- Traitement des données "manquantes"
  1. Éliminer toutes les lignes non remplies complètement
     - On risque d'éliminer beaucoup de lignes !
  2. Remplacer les données manquantes par des valeurs calculées
     - Variable nominale : catégorie la plus fréquente,
     - Variable continue : moyenne
  3. Créer une classe spéciale
     - KXEN
- La qualité n'est jamais parfaite !

Les données client

La vue 360° du client

- Propensité d'achat par produit, par canal
- Scores de risque, de churn …
- « Share of wallet »
- Life Time Value

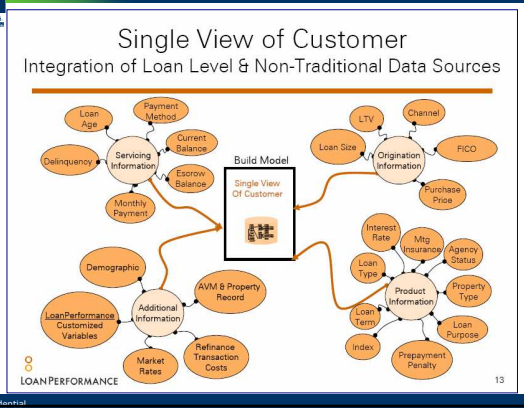
- Aspirations
- Plans futurs
- Attitudes

- Comportement de consommation
- Préférences de canal
- Position dans le cycle de vie
- Comportement de navigation

- Nom, prénom, adresse
- Sexe
- Date de naissance
- Revenu

- Transactions d'achats
- Click-stream
- Réponse aux campagnes
- Appel au centre d'appels
- Rendez-vous commerciaux

Les données client



Les besoins des utilisateurs

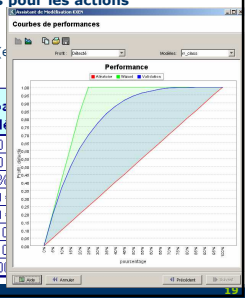
… sont nombreux : par exemple en Assurance

- Appétence
  - Nouveaux contrats
    - Assurance auto, MRH, crédit …
  - Cross-selling & up-selling
- Attrition, résiliation
  - Remboursement anticipé, refinancement, fin de contrat
- Fraude
  - Déclaration de sinistre, souscription
- Performances commerciales
  - Prévision des performances
  - Optimisation des actions de promotion
- Approche qualité globale
  - Amélioration des processus (Six Sigma)
  - Satisfaction client
- Provision pour sinistres …

Les besoins des utilisateurs

- **Qu'apportent les analyses ?**
  - **Connaissance client**
    - Comprendre ce qui différencie les clients
    - Détecter les leviers d'actions
  - **Fournir les informations nécessaires pour les actions**
    - Gagner en efficacité
    - Exemple : campagne ciblée / non ciblée
    - On peut aussi réduire la taille de la cible (e
      maintenant le nombre de réponses

	Campagne non ciblée	Campa cibl
Base Clients	5 000 000	5 000
Nb clients ciblés / an	2 500 000	2 500
Taux de réponse	4%	5%
Coût du contact	10 €	10
Revenu généré / réponse	30 €	30
Nb répondants	100 000	125 0
Retour des campagnes	2 000 000 €	2 500 0
Apport du ciblage		500 0



Les besoins des utilisateurs

- **Chacun de ces besoins doit être satisfait en**
  - **Exploitant les données disponibles**
  - **Produisant un modèle adapté**
- **Le nombre de modèles nécessaires est très grand**
  - **Exemple : scores clients**
  - **20 régions**
  - **10 segments clients**
  - **20 produits**
  - **10 scores**
    - Appétence : nouveaux contrats, Cross-selling & up-selling
    - Attrition : Remboursement anticipé, refinancement, fin de contrat
    - Fraude : Déclaration de sinistre, souscription
    - Performances commerciales : Prévision des performances
    - Satisfaction client
  - **... soit : 20 x 10 x 20 x 10 = 40 000 modèles « fins » !**

Les besoins des utilisateurs

Pourquoi faire des modèles « fins »

- **La performance sur une population homogène est meilleure**
  - **Exemple : Appétence assurance MRH**
  - **à Paris et dans la Creuse**
  - **On doit donc faire un modèle par segment homogène**
- **Le ciblage plus fin permet de**
  - **Être plus pertinent dans le message**
    - Personnalisation
  - **Éviter la sur-pression commerciale**
- **Les volumes sont plus petits, ce qui permet de**
  - **Réduire les temps de calcul**
  - **Restreindre les coûts**
  - **Optimiser la logistique des opérations**

Les besoins des utilisateurs

Exemple : Vodafone

- **Base Teradata**
  - **2500 variables**
- **Environnement analytique**
- **700 modèles / an**
- **Production de très nombreux modèles**
- **Résultats d'analyses disponibles en temps réel**
- **Performance accrue**

So, is this all that is needed? No

- Ideal analytical environment also includes:
  - Automated, over-night, creation of new analytic variables
  - Tighter integration with campaign management systems, profiling and direct link into actions
  - Measurement system to show effectiveness of all scores over time
    - This will be important when there are hundreds of models
- But, what we are going to build is a HUGE improvement versus what Vodafone has now.

experience the POSSIBILITIES 2005

- Consistently high quality models can be produced by less experienced analysts

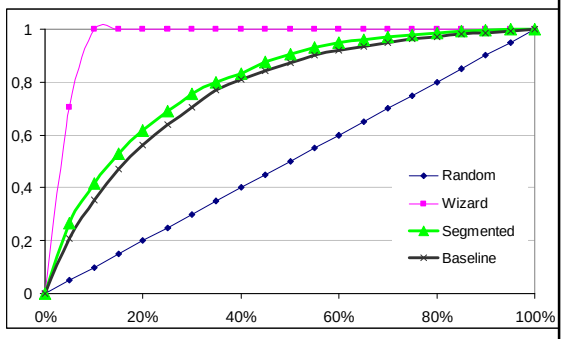
experience the POSSIBILITIES 2005

This means trying to create 716 models per year…

http://www.teradata.com/teradata-partners/conf2005/

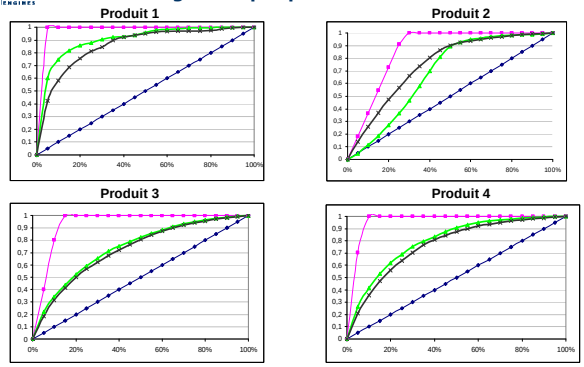
Les besoins des utilisateurs

- **Modèle global / segmenté par produit (4 produits)**



Les besoins des utilisateurs

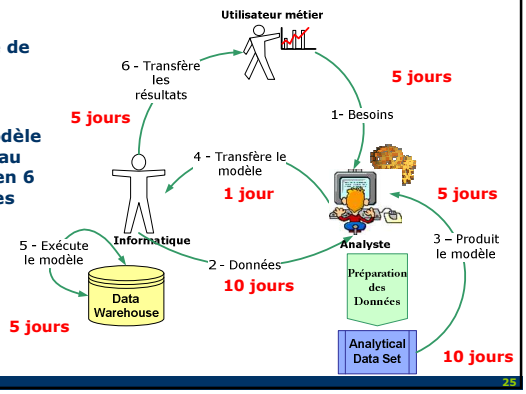
- **Modèle segmenté par produit**



Les besoins des utilisateurs

Le cycle de vie des études

… le modèle répond au besoin en 6 semaines



Les besoins des utilisateurs

- **Comment raccourcir ce cycle ?**
  - **En rapprochant les équipes**
    - Compétences IT & études au sein du département marketing
  - **En industrialisant la mise à disposition des données**
    - Mise en place d'un datawarehouse d'entreprise
    - Accès par les utilisateurs à des « vues » métier
  - **En essayant de « simplifier » le processus études**
    - Automatiser les analyses « simples »
    - Mettre en place des outils orientés « utilisateurs » et pas seulement « statisticien »
- **Que devient alors le rôle du statisticien ?**
  - **Le statisticien devient l'expert de référence**
    - Pour traiter les problèmes durs
    - Pour valider les résultats critiques
  - **L'utilisateur métier exécute les analyses**
    - Dont il a besoin
    - Quand il a besoin
    - Tout seul

La vitesse

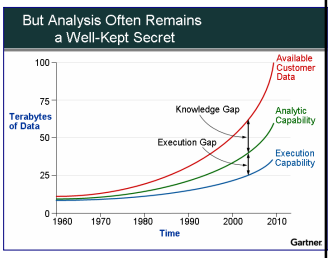
La vitesse est un facteur clé de performance

Un délai réduit pour produire un modèle (depuis la conception à la mise en production) permet de

- **Améliorer la productivité des équipes**
  - **Produire un modèle en 2 jours au lieu de 6 semaines permet de faire plus de modèles**
- **Améliorer les performances**
  - **Les données utilisées pour la modélisation sont récentes, le marché n'a pas changé**
    - La performance du modèle en production est celle attendue par le modèle
- **Améliorer le « time-to-market »**
  - **La réactivité à une offre de la concurrence est plus rapide**

Les besoins

- **Les besoins sont là**
- **Mais la plupart des entreprises n'ont pas réussi à mettre en oeuvre une exploitation industrielle de leurs données**

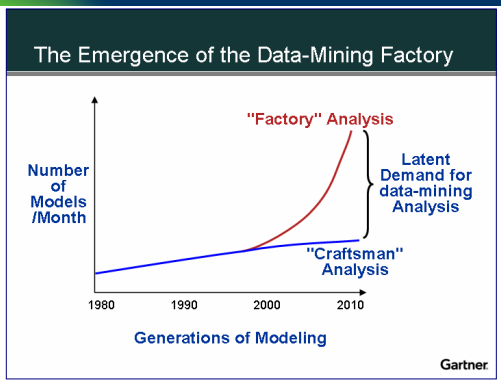


- **« Analytics is typically a low productivity, hand crafting, cottage industry that is inadequate to meets the demands of the modern economy … KXEN is actually an instrument for fundamental business change … »** **David Norris , Associate Analyst, Bloor Research**
  - http://www.it-director.com/technology/applications/content.php?cid=9021

Agenda

- **Le data mining**
- **Les besoins**
- **L'usine à modèles**
- **La mise en œuvre de KXEN**
- **Quelques exemples**
- **Conclusion**

L'usine à modèles



D'après : Gareth Herschel, Gartner, KXEN Users Group, San Francisco, November 2006

L'usine à modèles

L'usine à modèles, c'est la capacité de

Masse de données

- **Traiter des masses de données**
  - **10-100 Millions Clients**
  - **5 000 variables**
  - **… ce qui demande**
  - **Un algorithme linéaire (ou presque)**
  - **Une manipulation des données minimum**
    - Pas de duplication
    - Quelques passes pour lire les données

Masse de projets

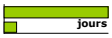
- **Produire des masses de projets**
  - **100-1000 projets / an / semaine / jour**
  - **… ce qui demande**
  - **La possibilité d'automatiser la réalisation du modèle**
  - **La facilité à exporter / intégrer le modèle en production**

L'usine à modèles

L'usine à modèles, c'est la capacité de

Super rapidité

- **Produire les modèles très rapidement**
  - **En quelques jours / heures**
  - **… ce qui demande**
  - **Un outil convivial**
  - **Avec automatisation des tâches lourdes**
    - Codage des données
    - Sélection des algorithmes
    - Exécution du modèle (ex : dans la base de données)



Super auto-matisation

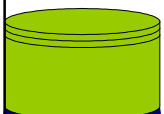
- **Produire des modèles «automatiquement»**
  - **Industrialiser la production du modèle**
  - **Industrialiser l'export du modèle**
  - **Industrialiser l'exécution du modèle**
  - **… ce qui demande**
  - **Codage**
  - **Langage de script**
  - **Export vers tous formats**

L'usine à modèles

L'usine à modèles, c'est la capacité à

Expertise limitée

- **Être utilisable par des utilisateurs métier**
  - **Expertise statistique pas indispensable**
  - **Connaissance du métier & des données suffisante**
  - **… ce qui demande**
  - **Un outil robuste orienté « utilisateurs » et pas seulement « statisticien »**
  - **Une équipe comprend typiquement**
    - 1 expert
    - 10 utilisateurs



Lecture des données

- **Être efficace sur la manipulation des données**
  - **Gros volumes**
  - **Éventuellement dispersés**
  - **… ce qui demande**
  - **Pas de duplication / déplacement des données**
  - **Minimiser les passes sur les données (ex : 3)**

L'usine à modèles

L'usine à modèles c'est

- **L'augmentation de la productivité**
  - **Plus de modèles, produits plus vite par moins de personnes**
  - **Des personnes moins qualifiées**
- **L'augmentation des bénéfices**
  - **Des modèles pour chaque problème**
  - **… même ceux pour lesquels on n'avait pas le temps**
- **L'augmentation de la vitesse**
  - **Un « time-to-market » réduit**
  - **Des modèles sur des données plus récentes**

Ce que nous appelons le « data mining extrême »

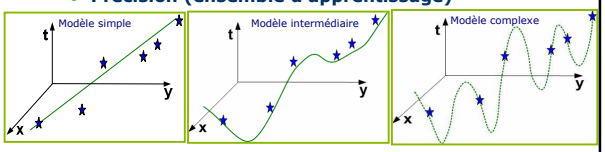
Agenda

- **Le data mining**
- **Les besoins**
- **L'usine à modèles**
- **La mise en œuvre de KXEN**
- **Quelques exemples**
- **Conclusion**

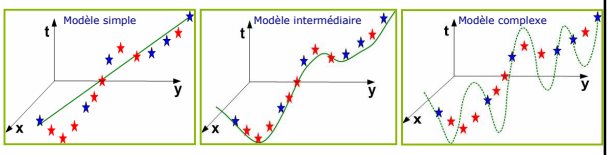
Le cadre mathématique

Qu'attendons-nous d'un modèle

- **Précision (ensemble d'apprentissage)**



- **Robustesse (ensemble de test)**



Le cadre mathématique

- **Données d'apprentissage**
  - La cible *y* peut être continue ou pas
  - $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
  - Dans la « base d'apprentissage », tous les y_i sont connus
- **Une classe de fonctions** $\Phi_\Theta = \{f(.,\theta), \theta \in \Theta\}$
  - Par exemple :
    - La classe des plynômes de degré p
    - La classe des MLP avec p neurones cachés …
- **Un modèle issu de cette classe** $y = f(x, \theta)$
  - Par exemple, le polynôme dont les paramètres sont θ
- **Le « meilleur » modèle** $\hat{y} = f(x, \hat{\theta})$
  - Produit par un certain algorithme ou un principe d'inférence
  - Et qui correspond donc au « meilleur » paramètre $\hat{\theta}$

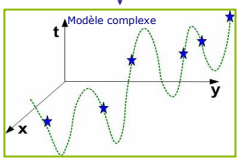
Le cadre mathématique

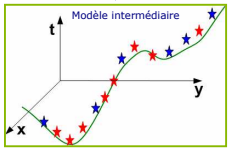
- **Une fonction de coût** $L[y, f(x,\theta)]$
  - Par exemple $L[y, f(x,\theta)] = [y - f(x,\theta)]^2$
    - L'écart quadratique
- **L'erreur en apprentissage ou risque empirique** $R_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L[y_i, f(x_i,\theta)]$
  - Le coût moyen sur l'ensemble d'apprentissage
  - Par exemple l'écart quadratique moyen MSE (Mean Square Error) $R_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n} [y_i - f(x_i,\theta)]^2$
- **L'erreur en généralisation** $R_{Gen}(\theta) = \int L[y, f(x,\theta)] \cdot dP(x,y)$
  - Le coût moyen théorique sur l'ensemble de la population
  - … qui est l'erreur attendue sur de nouvelles données
- **Principe d'inférence** $\hat{\theta} = \arg\min_\theta R_{emp}(\theta)$
  - Minimisation du risque empirique
  - Par exemple : LMSE (Least Mean Square Error) $\hat{\theta} = arg\ min_\theta \frac{1}{n}\sum_{i=1}^{n} [y_i - f(x_i,\theta)]^2$

Le cadre mathématique

Deux notions

- **L'erreur d'apprentissage (précision)** $R_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L[y_i, f(x_i,\theta)]$
- **L'erreur de généralisation (robustesse)** $R_{Gen}(\theta) = \int L[y, f(x,\theta)] \cdot dP(x,y)$





La théorie de Vapnik

- **La « Statistical Learning Theory » de Vapnik est une théorie générale qui repose sur 4 principes**
  1. **Consistence (robustesse)**
     - Capacité à généraliser correctement sur de nouvelles données
  2. **Vitesse de convergence**
     - Capacité à généraliser de mieux en mieux quand le nombre de données d'apprentissage augmente
  3. **Contrôle de la capacité de généralisation**
     - C'est la stratégie qui permet de contrôler la capacité de généralisation à partir des seules données disponibles : les données d'apprentissage
  4. **Stratégie pour obtenir de bons algorithmes**
     - C'est la stratégie qui nous permet de garantir et mesurer la capacité de généralisation du modèle que notre algorithme produit

… et utilise un paramètre la « VC dimension » ou *dimension de Vapnik Chervonenkis*

La théorie de Vapnik

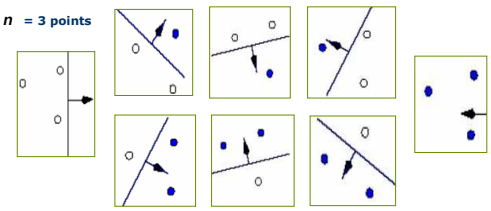
Dimension de Vapnik Chervonenkis

- **Etant donné**
  - Un échantillon de ***n*** observations $(x_1, x_2, \ldots, x_n)$ caractérisées par ***p*** variables : $x_i \in \Re^p$
- **Il y a 2^n façons de séparer ces n observations en 2 classes**
- **On dit que la famille de fonctions $\Phi_\Theta = \{f(.,\theta), \theta \in \Theta\}$ "*pulvérise*" l'échantillon si toutes les 2^n séparations sont réalisables (avec un θ bien choisi)**
- **On dit que la famille Φ_Θ est de *VC dimension* $h \in \mathcal{H}$ si :**
  1. **Tout échantillon de *h* observations de $\Re^p$ peut être pulvérisé par Φ_Θ**
  2. **Il existe au moins un échantillon de *h+1* observations qui ne peut pas être pulvérisé par Φ_Θ**

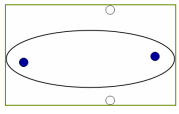
La théorie de Vapnik

Exemple : la famille des droites de $\Re^2$

- ***n*** **= 3 points**



- ***n*** **= 4 points**



- ***h*** **= 3 (=p+1)**

1. Consistence (robustesse)

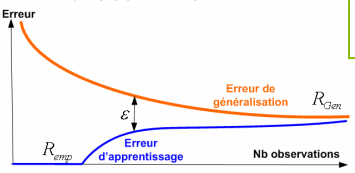
- Capacité à généraliser correctement sur de nouvelles données
- Un modèle $\hat{y} = f(x, \hat{\theta})$ est consistent si et seulement si la famille $\Phi_\Theta = \{f(\,,\theta), \theta \in \Theta\}$ dont il est issu est de VC dimension h finie

2. Vitesse de convergence

- Capacité à généraliser de mieux en mieux quand le nombre de données d'apprentissage augmente

$$R_{Gen}(\theta) \le R_{emp}(\theta) + \varepsilon(n,h)$$

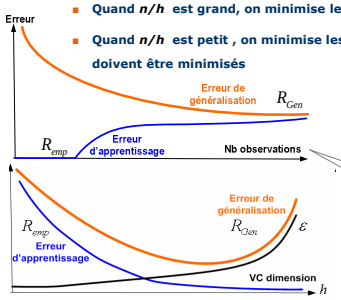
$$\varepsilon(n,h) = \sqrt{\frac{1+\ln(2n/h) - \frac{\ln q}{n}}{n/h}}$$



- Indépendant des distributions de (X,Y)

3. Contrôle de la capacité de généralisation

- C'est la stratégie qui permet de contrôler la capacité de généralisation à partir des seules données disponibles : les données d'apprentissage
- Quand *n/h* est grand, on minimise le risque empirique R_{emp}
- Quand *n/h* est petit , on minimise les deux termes : R_{emp} ET $\varepsilon(n,h)$ doivent être minimisés



$$R_{Gen}(\theta) \le R_{emp}(\theta) + \varepsilon(n,h)$$

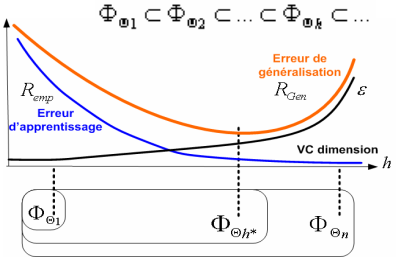
$$\varepsilon(n,h) = \sqrt{\frac{1+\ln(2n/h) - \frac{\ln q}{n}}{n/h}}$$

Statistique classique

4. Stratégie pour obtenir de bons algorithmes

- C'est la stratégie qui permet de garantir et mesurer la capacité de généralisation du modèle que notre algorithme produit
- SRM (Structural Risk Minimization) : on utilise des familles de fonctions emboîtées à VC dimension croissante

$$\Phi_{\Theta 1} \subset \Phi_{\Theta 2} \subset \ldots \subset \Phi_{\Theta k} \subset \ldots$$



- Produire un modèle dans une famille
  - Fit de données
- Choisir le meilleur modèle dans l'ensemble des modèles générés
  - Choix de modèle

Compromis fit-robustesse

- La SRM est une façon d'implémenter le principe du rasoir d'Occam

Occam's Razor
through the ages...

Pluralitas non est ponenda sine necessitate.
(Plurality should not be posited without necessity.)
· William of Ockham

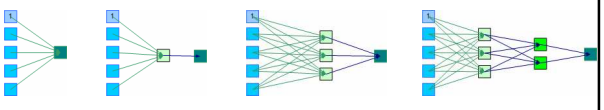
Everything should be made as simple as possible, but not simpler.
- Albert Einstein

Keep
It
Simple,
Stupid !

Moralité

- Ce qu'on ne peut pas contrôler
  - La distribution des données
    - Rarement Gaussiennes …
    - On ne veut pas vraiment connaître la distribution, mais prendre une décision
  - Les approximations de distribution
    - Transformations pour se ramener au cas Gaussien …
  - Principe d'économie
    - Ne pas résoudre un problème complexe (estimer une distribution) pour prendre une décision simple (quelle est la réponse en ce point)
- Ce qu'on peut contrôler
  - La classe de modèles où on recherche la solution Φ_Θ
  - La VC dimension *h* de la classe retenue
- Avec une méthode de contrôle
  - La SRM qui garantit la robustesse

Exemple : Multi Layer Perceptrons
On définit une structure emboîtée
1. Par l'architecture



$$\Phi_{\lambda_1} \subset \Phi_{\lambda_2} \subset \Phi_{\lambda_h} \subset \Phi_{\lambda_n}$$

$$h_1 < h_2 < h_h < h_n$$

La théorie de Vapnik - SRM

Structure emboîtée pour les MLP
2. Par l'algorithme d'apprentissage

- **Prenons la classe** $\Phi_{\lambda_i} = \{F(x; W, \lambda_i), \|W\| \leq \lambda_i\}$
  avec $\lambda_1 < \lambda_2 < \lambda_h < \cdots < \lambda_n$
- **La solution optimale dans** Φ_{λ_i} **est celle qui minimise** $\Re(W, \lambda_i) = \frac{1}{m}\sum_{k=1}^{m}[y^k - F(x^k; W, \lambda_i)]^2 + C_i \sum_j W_j^2$
  avec C_i **qui dépend de** λ_i
- **On retrouve le *weight decay***

La théorie de Vapnik - SRM

Structure emboîtée pour les MLP
3. Par les pré-traitements ACP

- **Prenons la matrice** U_p **dont les colonnes sont les vecteurs propres de** $X \,.\, {}^tX$ **associés aux** p **premières valeurs propres** $\lambda_1 < \lambda_2 < \lambda_h < \cdots < \lambda_n$
- **On projette les exemples** x^k **sur l'espace de l'ACP** E_p **: la projection a pour matrice** $U_p \,.\, {}^tU_p$
- **On obtient la structure en faisant varier p**

$$1 < 2 < \cdots < p < \cdots < n$$

La théorie de Vapnik - SRM

Structure emboîtée pour les MLP
4. Par les pré-traitements

- **Bruitage : on applique un bruit sur les entrées** x^k
- **Smoothing : on applique un pré-processeur** $K(x, \beta)$ **sur les exemples (images)** x_{ij}
  - **Par exemple un noyau exponentiel sur les pixels**

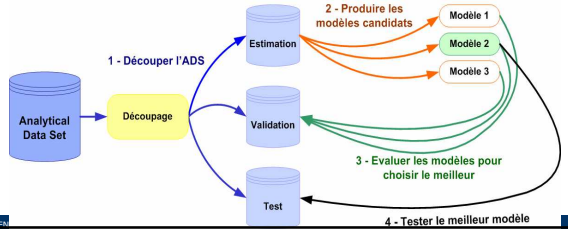
$$z_{ij} = \frac{\sum_k \sum_l x_{i+k\ j+l} exp\left(-\frac{1}{\beta}\sqrt{k^2+l^2}\right)}{\sum_k \sum_l exp\left(-\frac{1}{\beta}\sqrt{k^2+l^2}\right)}$$

La théorie de Vapnik - SRM – Implémentation KXEN

- **La SRM en pratique dans KXEN**

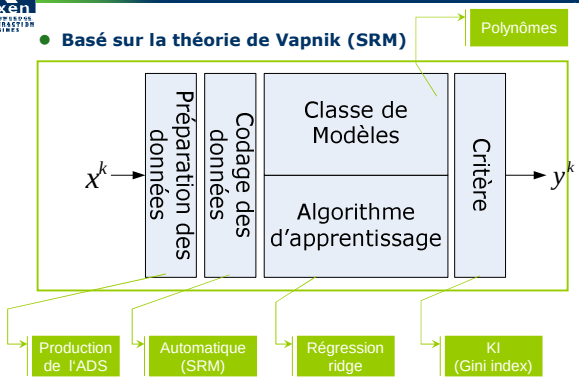
Modélisation
- Choisir la famille emboîtée de fonctions
- Augmenter progressivement la VC dim
- Choisir le modèle qui optimise le compromis précision / robustesse

- **Et deux indicateurs**
  - **Précision : *KI***
  - **Robustesse : *KR***



La théorie de Vapnik - SRM – Implémentation KXEN

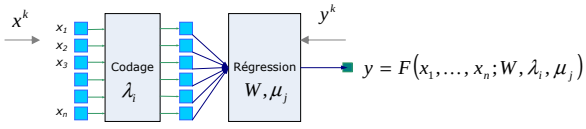
- **Basé sur la théorie de Vapnik (SRM)**



La théorie de Vapnik - SRM – Implémentation KXEN

Régression polynômiale

- **On utilise une structure en deux modules**



qu'on calibre en même temps, en utilisant

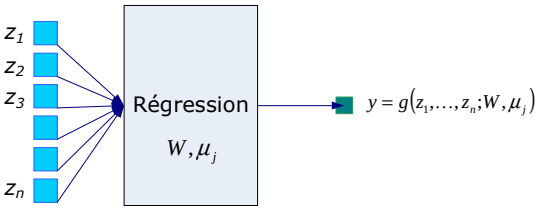
- **Une régression ridge pour le fit des données :** W^*
- **Une SRM pour le choix des modèles :** λ_i, μ_j

La théorie de Vapnik - SRM – Implémentation KXEN

Régression polynômiale

- **On utilise la classe des polynômes**
  - **À degré q donné, famille emboîtée par μ_i croissants**

$$\Phi_{q\,\mu_i} = \{\, g(x;W,\mu_i),\ polynôme\ de\ degré\ q; \|W\| \le \mu_i \,\}$$



La théorie de Vapnik - SRM – Implémentation KXEN

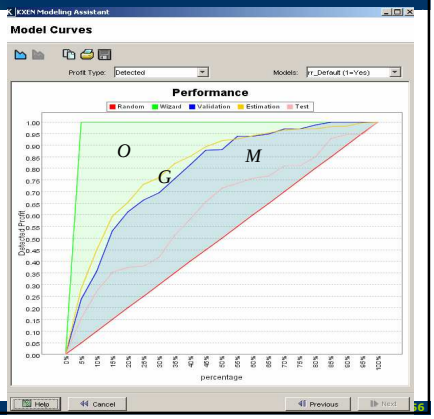
- **Critère (fit)**

$$KI = M/O$$

$$KI = 2AUC - 1$$

- **Robustesse**

$$KR = 1 - G/O$$



La théorie de Vapnik - SRM – Implémentation KXEN

Ce qu'apporte l'exploitation de la SRM

- **Productivité**
  - **Codage automatique**
    - Prise en compte de tous les formats
    - Codage robuste
    - Codage adapté au problème
  - **Automatisation**
    - Un seul algorithme
      - Question : « One size fits all » ?
    - Test & debriefing intégrés
    - Export & industrialisation faciles
- **Efficacité**
  - **Robustesse**
    - La théorie SRM de Vapnik
- **Et une méthodologie performante**
  - **DMAIC hérité de la méthode qualité Six Sigma**

La théorie de Vapnik - SRM – Implémentation KXEN

The Data-Mining Process

Process (Proportion of Effort)	Subprocess	Stakeholder B(usiness)	A(nalyst)	I(T)
Problem Understanding (5%-10%)	Determine objective	B		
	Define success criteria	B		
	Determine data mining goals	B	A	
Data Understanding (10%-15%)	Collect initial data	B	A	
	Explore data		A	
	Verify data quality		A	I
Data Preparation (30%-60%)	Select data		A	I
	Clean data		A	I
	Format data		A	I
Modeling (20%-30%)	Select modeling techniques		A	
	Build models		A	
	Select model		A	
Evaluation of Results (20%-30%)	Validate model	B	A	
	Explain model	B	A	
Deployment (5%-10%)	Deploy model		A	I
	Score deployment			I
	Monitor and maintain	B	A	

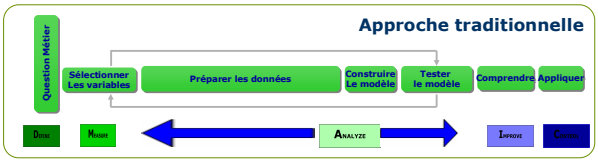
6 semaines

Gartner

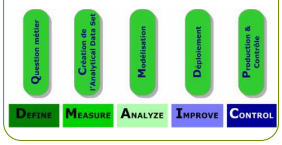
D'après : Gareth Herschel, Gartner, KXEN Users Group, San Francisco, November 2006

La théorie de Vapnik - SRM – Implémentation KXEN

Approche traditionnelle



Approche KXEN



- **Réduction du temps**
  - **Facteur 1 à 10**
- **Industrialisation**
  - **Export tous formats**

La théorie de Vapnik - SRM – Implémentation KXEN

DEMO

Agenda

- Le data mining
- Les besoins
- L'usine à modèles
- La mise en œuvre de KXEN
- Quelques exemples
- Conclusion

Le Crédit Lyonnais

Etat des lieux avant KXEN

- **Offre bancaire de LCL : 400 produits**
  - Produits et services bancaires,
  - Produits de gestion d'actifs et d'assurance,
  - Gestion de patrimoine.
- **Campagnes marketing direct : plus de 130 actions**
  - Équipement en cartes bancaires,
  - Assurance-vie,
  - Fonds commun de placement …

  par emailings, mailings ou SMS
  - soit 10 millions de contacts sur des clients ou des prospects.
- **Les équipes réalisent leurs campagnes à partir d'une dizaine de scores généralistes**
  - « Faire fructifier son capital »,
  - « Percevoir des Revenus »,
  - « S'assurer au quotidien ».

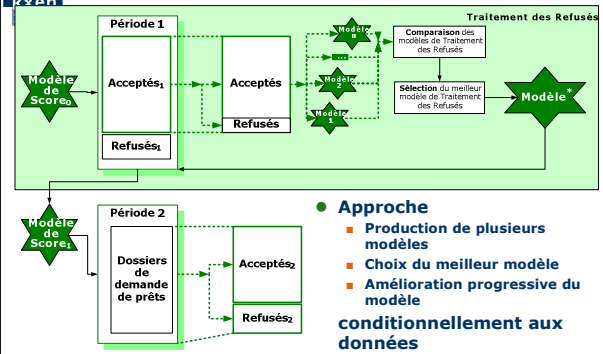
Le Crédit Lyonnais

- **Le département marketing opérationnel souhaite :**
  - Disposer de scores plus précis, facilement évolutifs
  - Spécifiques aux offres intégrées des familles de produits
- **Avec les outils existants**
  - Il faut de 2 à 5 jours pour construire les scores
  - La méthode ne permet pas d'affiner les scores.
- **Projet pilote sur une opération grandeur réelle (assurance MRH)**
  - Deux groupes sont constitués
    - Le premier utilise les scores KXEN,
    - L'autre utilise un score généraliste «S'assurer au quotidien ».
  - Résultats
    - Le taux de retour X 2,5 fois avec KXEN
    - Score KXEN élaboré en une demi-journée au lieu de plusieurs jours.
- **Aujourd'hui**
  - 160 modèles créés par an avec KXEN

Le traitement des refusés

- **Le score d'octroi est un modèle biaisé**
  - Modèle basé sur un historique de données (le comportement des dossiers Acceptés) non représentatif de la population globale (Acceptés + Refusés)
  - Le modèle est utilisé sur tous les dossiers
    - Il produit donc des résultats biaisés
  - Le biais peut être positif ou négatif !
    - Sur ou sous-estimation du risque
- **Le traitement des Refusés (Reject Inference) vise à corriger ce biais**
  - Il existe de nombreuses méthodes de traitement des refusés couramment utilisées dans les organismes de crédit
    1. Extrapolation
    2. Reclassification (Augmented data set)
    3. Augmentation (re-weighting)
    4. Parcelling
  - Aucune technique n'est valide statistiquement !!
    5. Groupe de contrôle
  - La meilleure façon reste de déterminer la meilleure technique pour les données dont on dispose

Le traitement des refusés



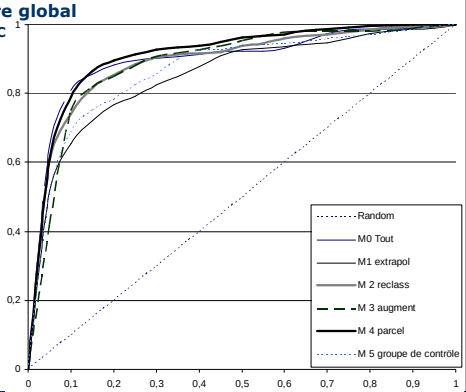
- **Approche**
  - Production de plusieurs modèles
  - Choix du meilleur modèle
  - Amélioration progressive du modèle

  conditionnellement aux données

Cf. E. Viennet, F. Fogelman Soulié. Le Traitement des Refusés dans le Risque Crédit. In «Data Mining et apprentissage statistique : applications en assurance, banque et marketing » F. Béninel et al. eds. RNTI. A paraître

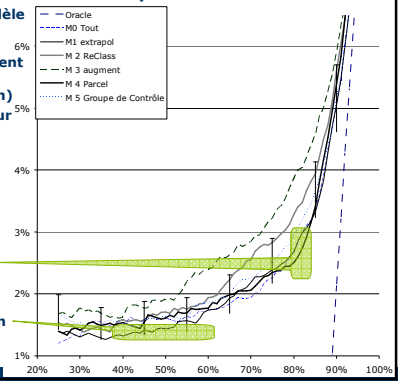
Le traitement des refusés

- **Critère global**
  - ROC



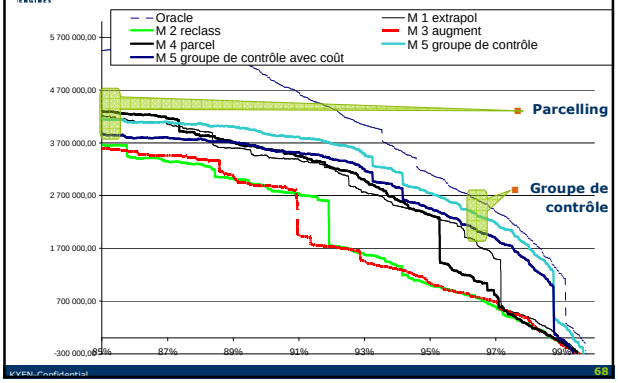
Le traitement des refusés

- **Taux de défaut sur les Acceptés**
  - Meilleur modèle dépend du point de fonctionnement (taux d'acceptation)
  - Barre d'erreur
  - Parcelling
  - Extrapolation

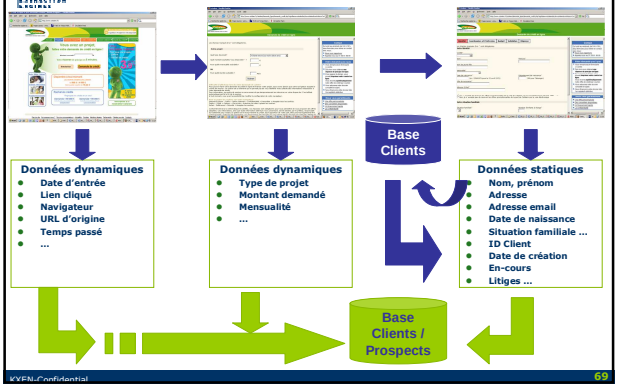


Le traitement des refusés

- **Critère de revenu**

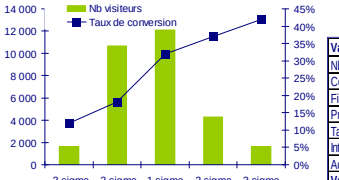


Exploiter les données d'un site



Exploiter les données d'un site : appétence au chat

- **Vente assistée sur le Web**
  - **Score d'appétence**
    - Temps réel
    - Proposition
    - Orientation vers l'agent compétent
  - **Résultats**
    - Taux d'acceptation / taux « naturel »
    - Taux d'achat
    - Business case

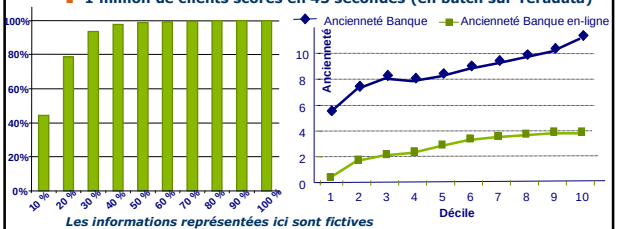


Variables	Exemple	
Nb visites / an		1 000 000
Config PC supportée	80%	800 000
Filtrage	30%	240 000
Proposition (modèle - 3 déciles)	30%	72 000
Taux d'acceptation	17%	12 240
Interaction menée à terme	60%	7 344
Acheteurs	5%	367
Valeur de transaction	**1 000 €**	**367 200 €**

Les informations représentées ici sont fictives

Exploiter les données d'un site : attrition

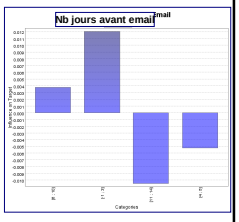
- **Modèle d'attrition des clients de la banque en ligne**
  - **Attrition est défini comme "inactivité de plus de 6 mois"**
  - **Modèle produit en 3 heures**
    - 200 variables réduites à 40
  - **44 % des clients attrités détectés dans le 1er décile**
    - Caractérisation des clients attrités
  - **1 million de clients scorés en 43 secondes (en batch sur Teradata)**



Les informations représentées ici sont fictives

Campagne email

- **Optimisation de campagnes d'un site e-commerce**
  - **Combien de jours après l'achat envoyer l'email**
- **Appétence cross-enseigne pour campagnes d'acquisition**



	Sans ciblage	Avec ciblage
Nb emails envoyés	51 968	29 999
Ouvertures totales		
Nombre	18 890	34 179
Taux	36,35%	113,93%
Ouvertures uniques totales		
Nombre	11 621	16 896
Taux	22,36%	56,32%
Clics totaux		
Nombre	3 222	9 538
Taux	6,20%	31,79%
Clics uniques totaux		
Nombre	2 619	6 879
Taux	5,04%	22,93%

Les informations représentées ici sont fictives

La nouvelle loi de l'e-économie

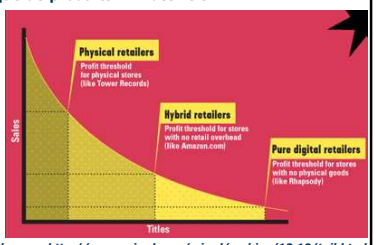
- De la « boutique » à la boutique électronique
  - Brick-and-Mortar
    - Wall Mart, Barnes and Noble, … Carrefour, Relay
  - Boutique électronique de produits matériels
    - Amazon, Netflix, … fnac.com
  - Boutique électronique de produits immatériels
    - Rhapsody, iTunes

Question

- Quel pourcentage de produits est vendu chaque mois ?

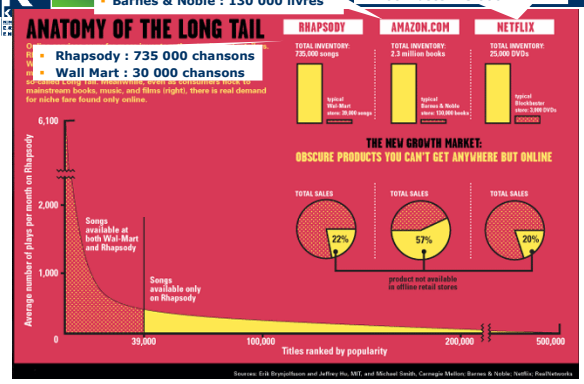
Réponse

- Tous ! (98-99 %)



From Chris Anderson - http://www.wired.com/wired/archive/12.10/tail.html

- Amazon.com : 2.3 M livres
- Barnes & Noble : 130 000 livres
- Netflix : 25 000 DVD
- BlockBuster : 3 000 DVD
- Rhapsody : 735 000 chansons
- Wall Mart : 30 000 chansons



From Chris Anderson - http://www.wired.com/wired/archive/12.10/tail.html

Long Tail

- Offrir « tout »
- Aider le client à trouver ce qu'il recherche

Long Tail

- Le système de recommandations

65 000 films

Netflix and Cinematch Scale

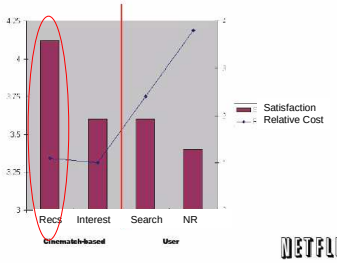
- 5M active customers
  - Ship 1.4M disks per day from 40 locations
- 1.4B ratings since 1997
  - 2M ratings per day
  - 1B predictions per day
- Item-to-item analysis with many data-conditioning heuristics
- 2 days to retrain on new ratings
- Manual item setup for "coldstart" titles
  - Automatically retired

http://blog.recommenders06.com/wp-content/uploads/2006/09/bennett.pdf

Long Tail

- Le système de recommandations

Does It Matter?



http://blog.recommenders06.com/wp-content/uploads/2006/09/bennett.pdf

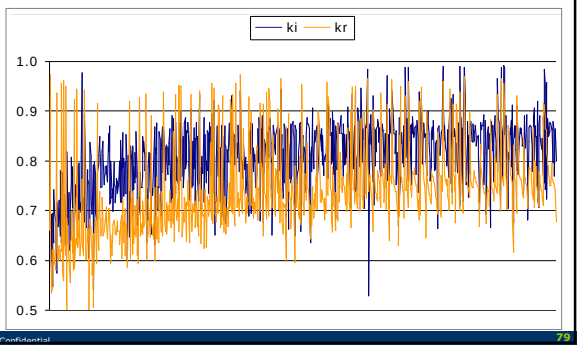
Long Tail

- Utiliser les règles d'association
  - Sur des milliers / millions de produits !

CUST ID	ar_ RuleId	ar_Consequent	ar_KI	ar_ Rules Left	ar_ Length	ar_ Support	ar_ Confidence
1	534	Blueberries	0.431	21	2	297	0.378
1	594	Raspberries	0.423	20	2	205	0.261
1	378	Bounty 2-Ply Paper Towels	0.331	19	2	277	0.581
1	572	Cantaloupe	0.278	18	2	206	0.262
1	456	Green Bell Pepper	0.271	17	2	276	0.352
1	547	Scallions	0.270	16	2	190	0.349
1	416	Red Bell Pepper	0.251	15	2	274	0.349

Long Tail

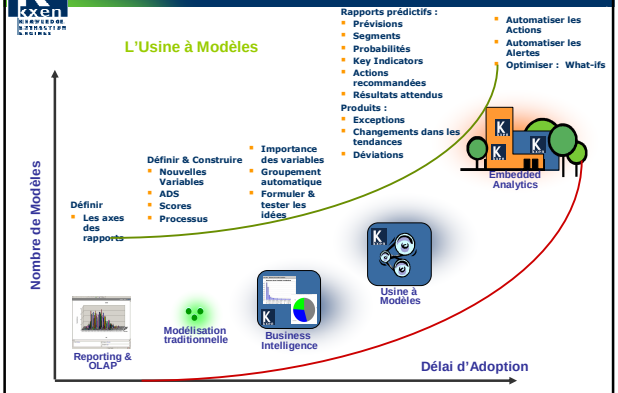
- **Faire des milliers / millions de scores d'appétence**
  - Trouver les niches de produits avec score robuste



Agenda

- Le data mining
- Les besoins
- L'usine à modèles
- La mise en œuvre de KXEN
- Quelques exemples
- Conclusion

Le data mining industriel



QUESTIONS ?