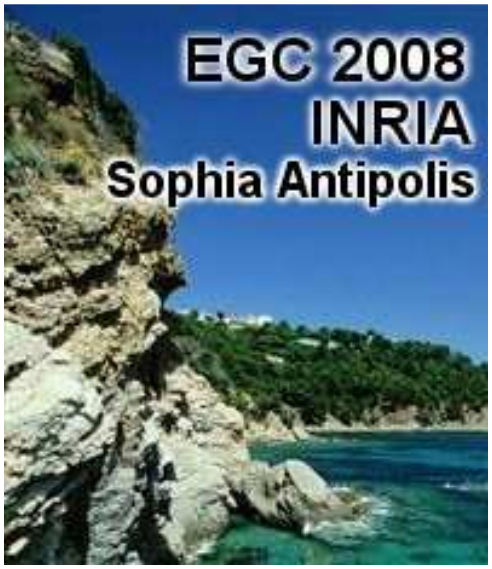


Industrialiser le data mining
Enjeux et perspectives

Françoise Fogelman Soulié

francoise@kxen.com

8èmes journées francophones
Extraction et Gestion des Connaissances
INRIA Sophia Antipolis – Méditerranée
29 janvier - 1er février 2008

Agenda

- Le Data Mining industriel
  - Un peu d'histoire …
  - Les données
  - Les défis
- Quelques exemples

Un peu d'histoire

En 1991, nous bataillions sur les MLP

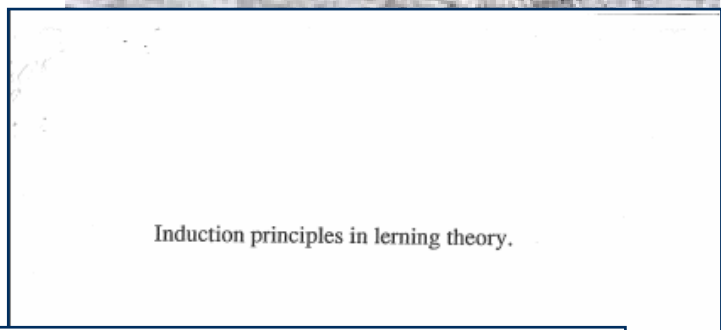
- Early stopping, Optimal Brain Damage, Weight Decay …

Avril 1991 – Snowbird Learning Workshop

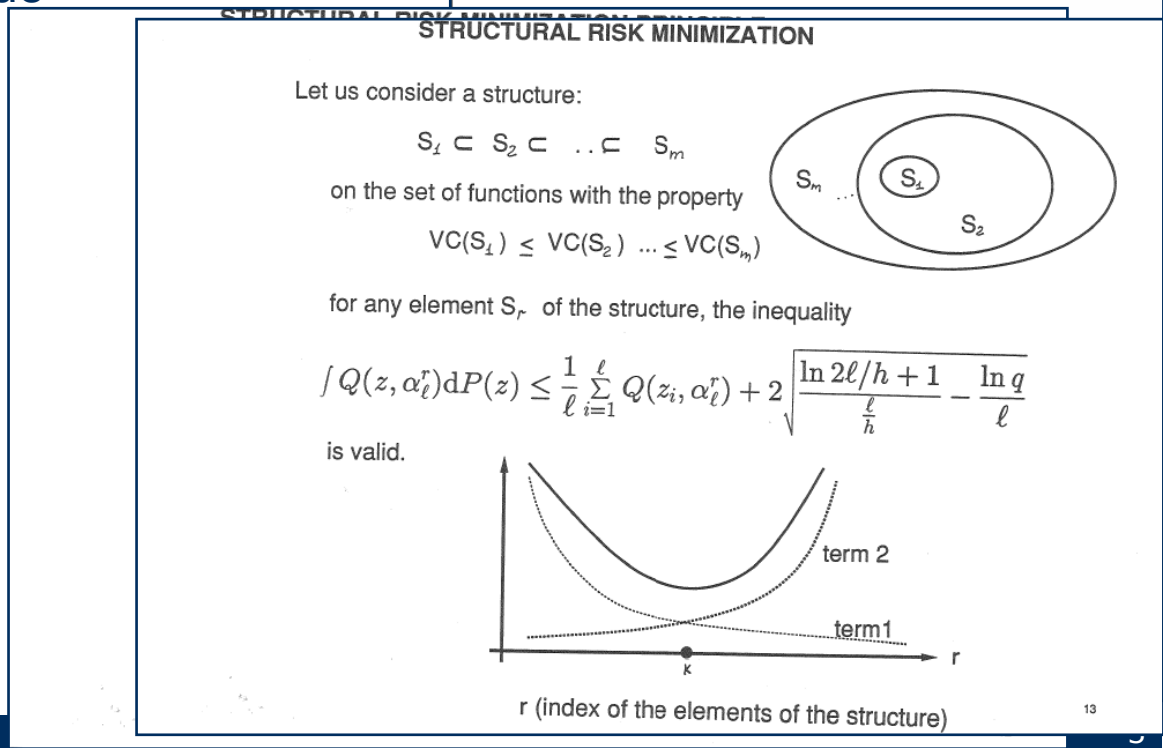
- Un nouveau nom & un titre bizarre

Pour moi, Vapnik a apporté une révolution

- Un beau cadre mathématique



Induction principles in lerning theory.



STRUCTURAL RISK MINIMIZATION

Let us consider a structure:

$S_1 \subset S_2 \subset \ldots \subset S_m$

on the set of functions with the property

$VC(S_1) \leq VC(S_2) \ldots \leq VC(S_m)$

for any element S_r of the structure, the inequality

$$\int Q(z,\alpha_\ell^r)\mathrm{d}P(z) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha_\ell^r) + 2\sqrt{\frac{\ln 2\ell/h + 1}{\frac{\ell}{h}} - \frac{\ln q}{\ell}}$$

is valid.

r (index of the elements of the structure)

13

Un peu d'histoire

En 1991, nous bataillions sur les MLP

- Early stopping, Optimal Brain Damage, Weight Decay …

Avril 1991 – Snowbird Learning Workshop

- Un nouveau nom & un titre bizarre

Pour moi, Vapnik a apporté une révolution

- Un beau cadre mathématique

Mais aussi

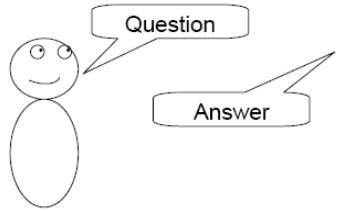
- Un cadre opérationnel
  - Expliquant beaucoup des « trucs » réseaux de neurones
  - Les SVM ont suivi
- Un mécanisme puissant pour contrôler la production de modèles (SRM)

- Nous avons construit notre software data mining KXEN sur la SRM
  - Il a fallu bien sûr beaucoup d'astuces dans la mise en œuvre informatique !

Un peu d'histoire

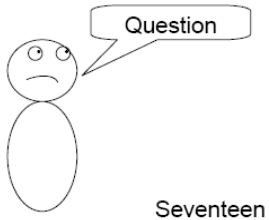
Qu'est ce qui s'est passé d'autre ?

Data Analysis: The old days



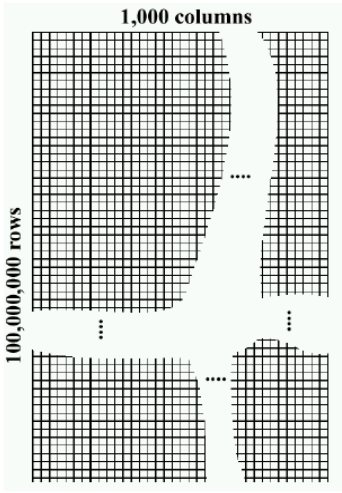
Size	Ellipticity	Color
23	0.96	Red
33	0.55	Red
36		Green
40		
20		
48		

Data Analysis: The new days



Seventeen
months later…

Answer



Andrew Moore, KDD'06

Données

Le volume des données a explosé

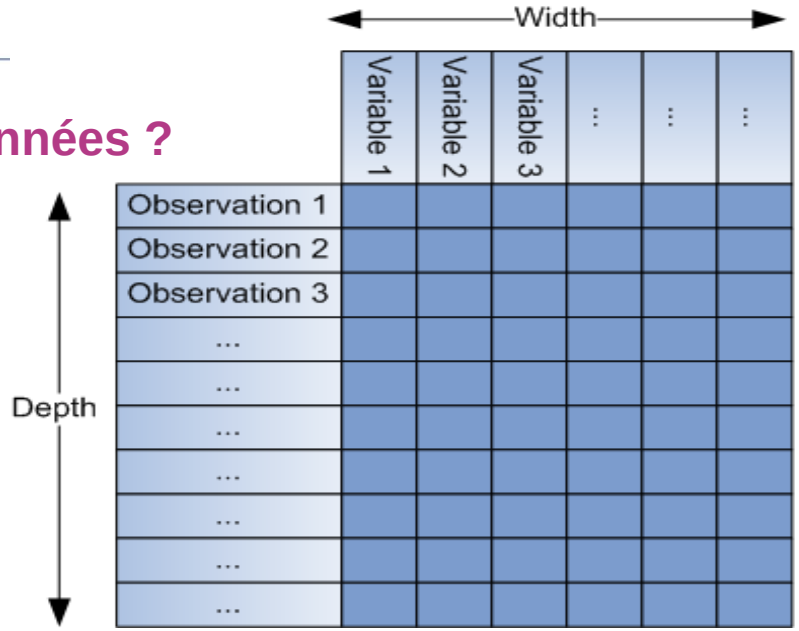
- Dans les années 90
  - A l'Ecole Modulad 1992
- Aujourd'hui
  - Transactions Web **Fayyad, KDD 2007**
    - Yahoo !
      - 16 B événements / jour
      - 425 M visiteurs / mois
      - 10 TO données data / jour
  - RFID **Jiawei,Adma 2006**
    - Un distributeur avec 3 000 magasins, vendant 10 000 items / jour / magasin
      - 300 M événements / jour
  - Réseau social **Kleinberg, KDD'07**
    - Labo de recherche (entreprise) : réseau e-mail de 436 nœuds sur 3 mois
    - Grande université : réseau e-mail de 43 553 nœuds sur 2 ans
    - Communauté blog LiveJournal : réseau d' « amitié » de 4,4 M nœuds
    - Microsoft Instant Messenger : réseau de communications IM de 240 M nœuds sur 1 mois
  - Réseaux télécom mobiles
    - Un opérateur telco génère des 100 M de Call data (CDR) / jour
    - Données techniques générées : 40 M événements / jour dans une grande ville

Réseaux de neurones	Statistiques
apprentissage	estimation
poids	paramètres
connaissance	valeur des paramètres
apprentissage supervisé	régression / classification
classification	discrimination / classement
apprentissage non supervisé	estimation de densité / clustering
clustering	classification / taxinomie
réseau de neurones	modele
grand: 100 000 poids ensemble d'apprentissage	grand: 50 paramètres échantillon
grand: 50 000 exemples	grand: 200 cas

Données

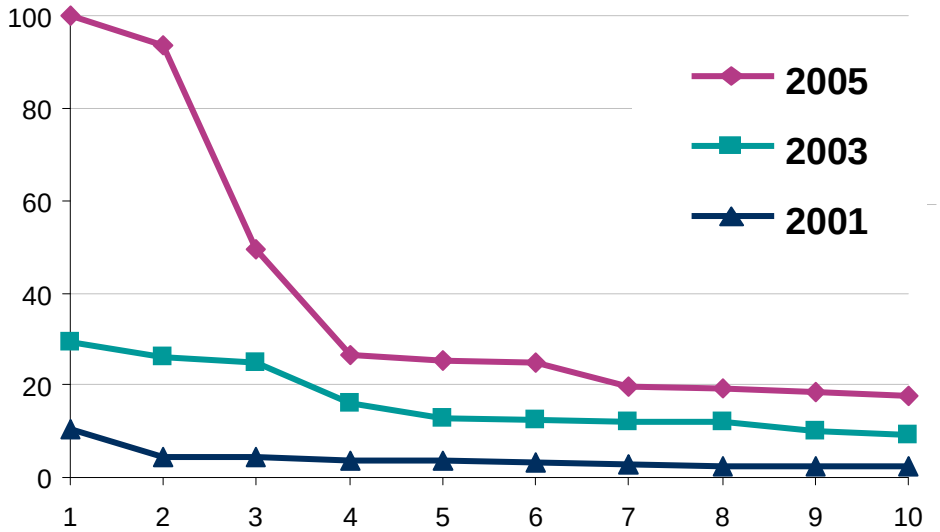
Qu'est ce qu'un « grand » ensemble de données ?

- Profondeur
  - Jusqu'à 100 Millions de lignes
  - Ou quelques milliards ?
- Largeur
  - Des milliers d'attributs
  - Ou quelques Millions ?



Grand aujourd'hui, et demain ?

- Taille des bases de données
  - X2-3 tous les 2 ans
- Part des données hors base
  - X 10 ? X 100 ?



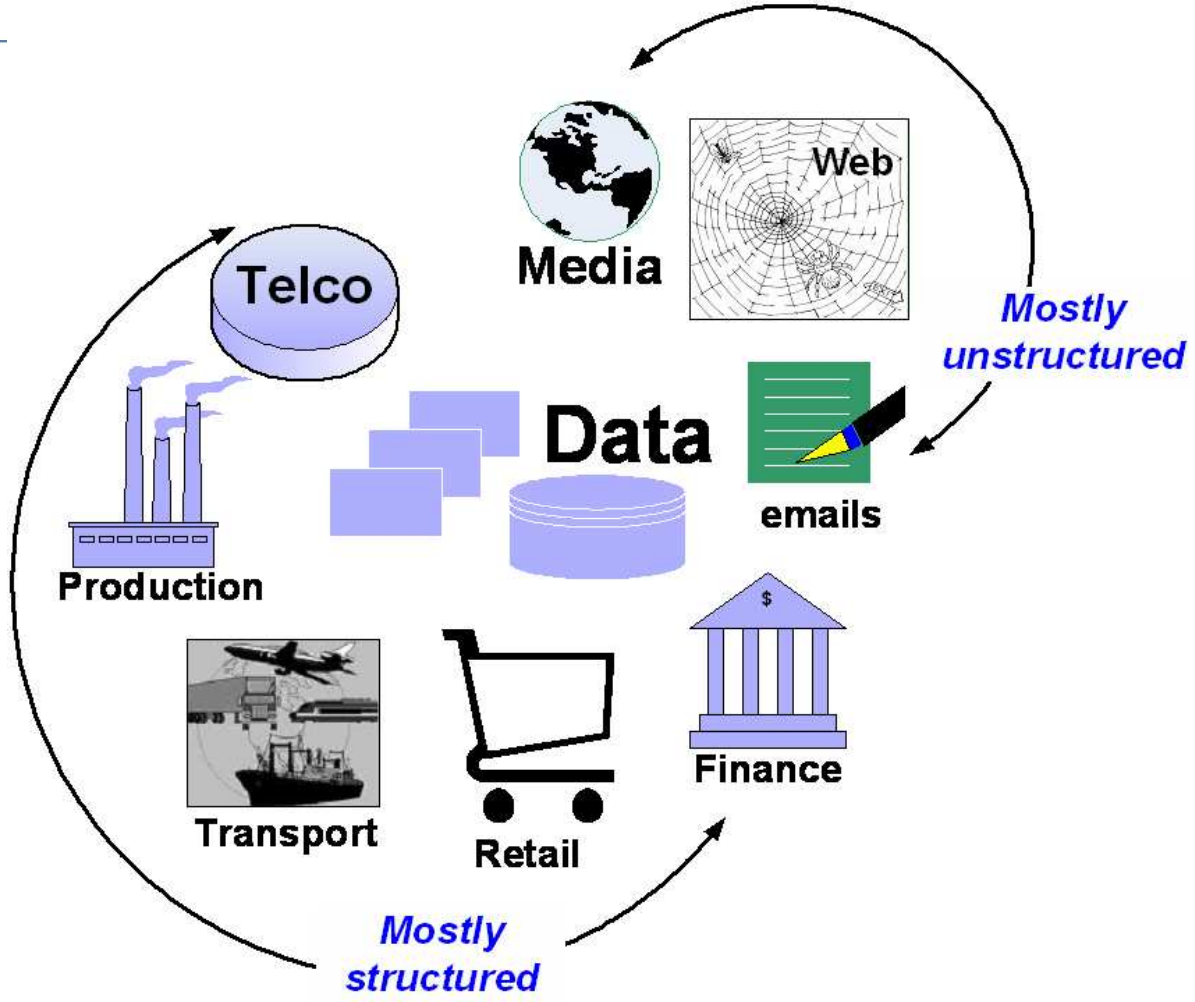
http://www.wintercorp.com

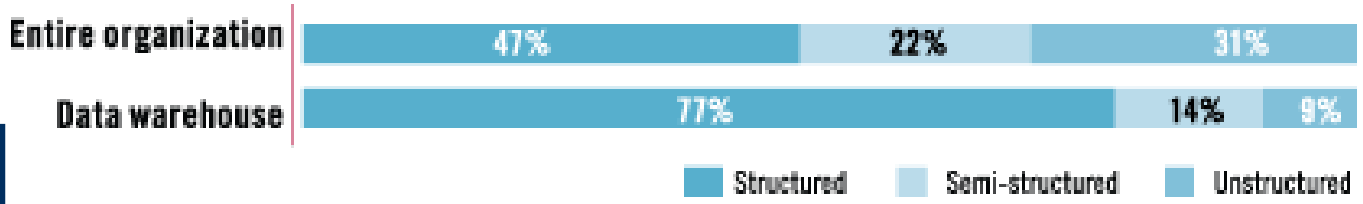
Les « masses de données »

Données

Beaucoup de

- Sources
- Types
  - Structuré
  - Non structuré
    - Texte
    - Image
    - Video
    - Audio
    - …
- Volumes
  - Le Web domine !



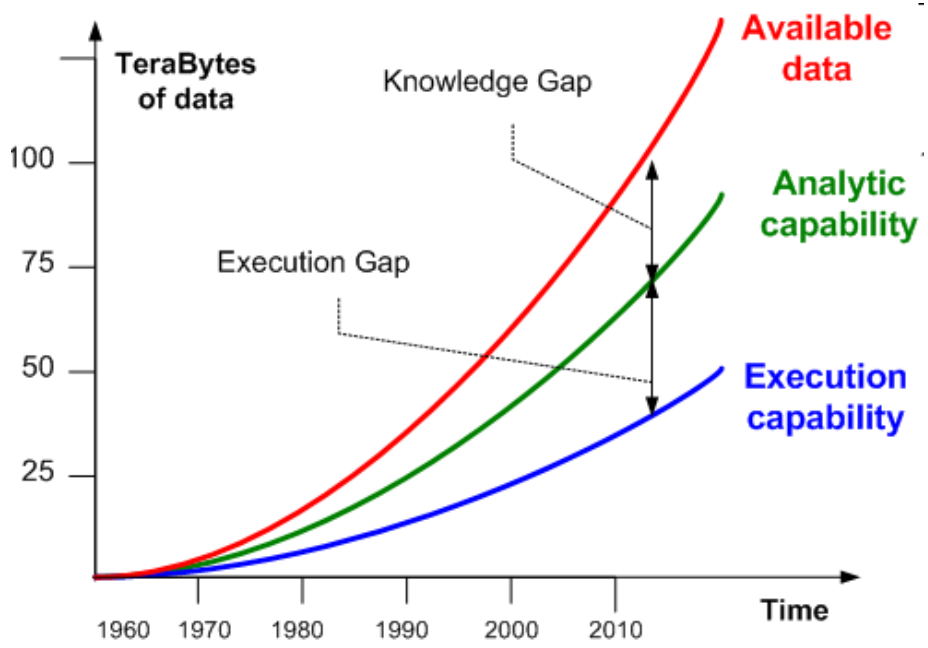


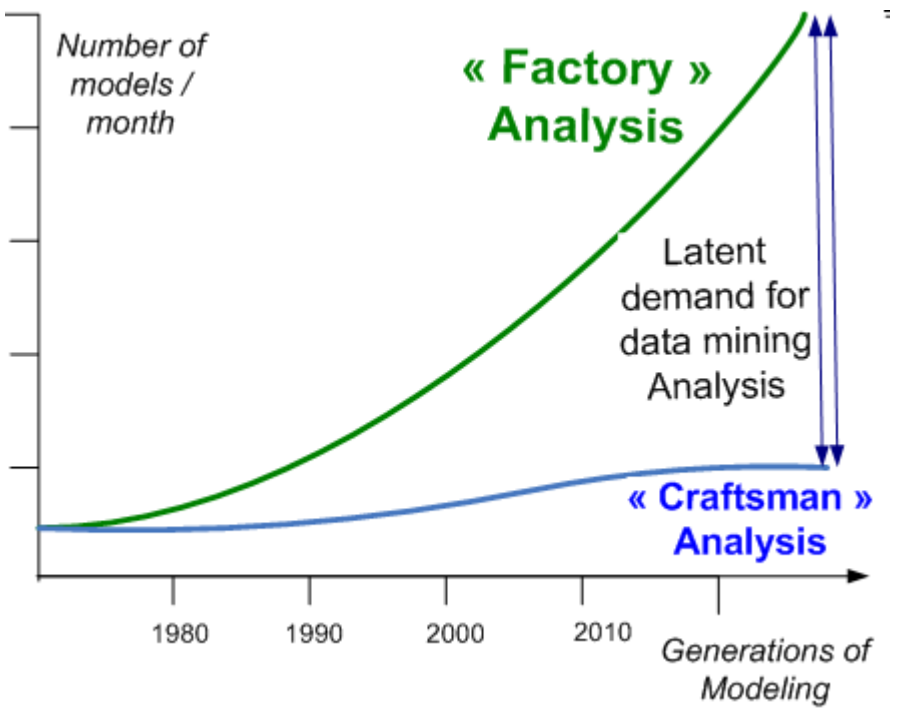
Russom, TDWI 2007

Le contexte industriel

D'après le Cabinet d'analyse Gartner

- Le data mining fournit des moyens de définir des actions
  - Un modèle non utilisé pour une action n'est qu'un coût inutile
- Le volume de données croît exponentiellement
  - Le nombre de modèles doit suivre





Herschel, Gartner 2006

Le contexte industriel

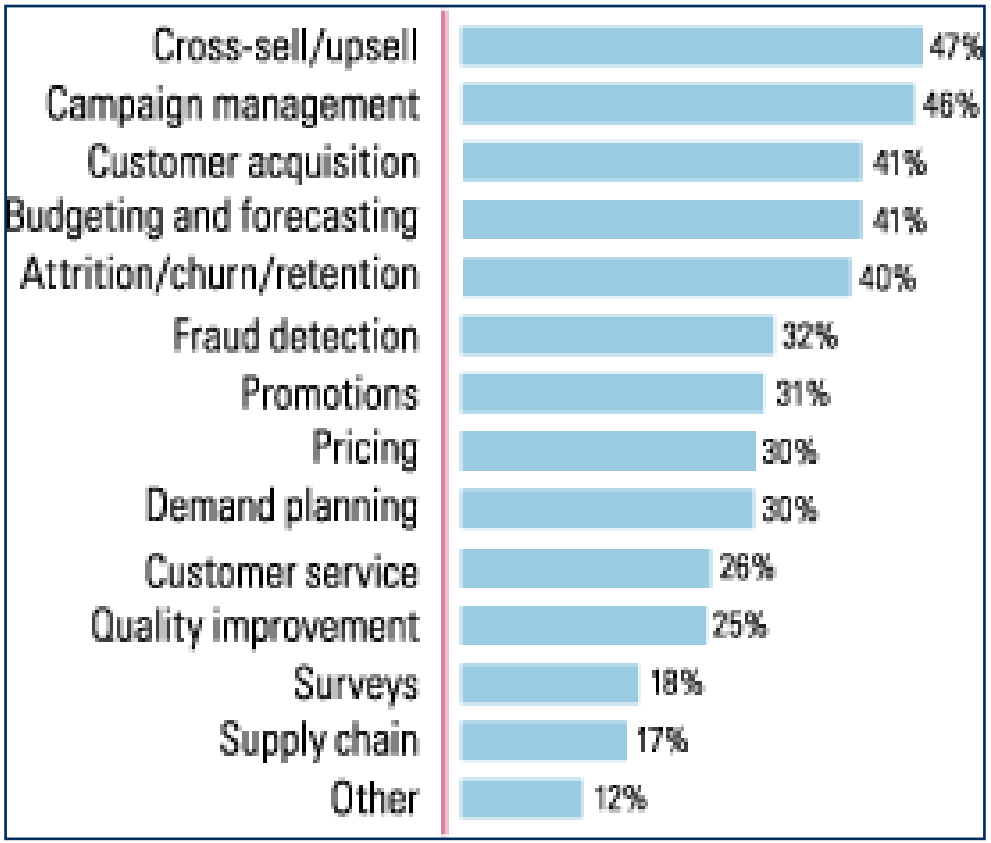
Le Data mining est utilisé surtout dans des applications CRM

- Les utilisateurs ne sont pas des data miners

Pour être productif, il faut être **simple**

- Les utilisateurs doivent pouvoir comprendre les modèles
- …voire les produire ?

Ce n'est pas le cas aujourd'hui

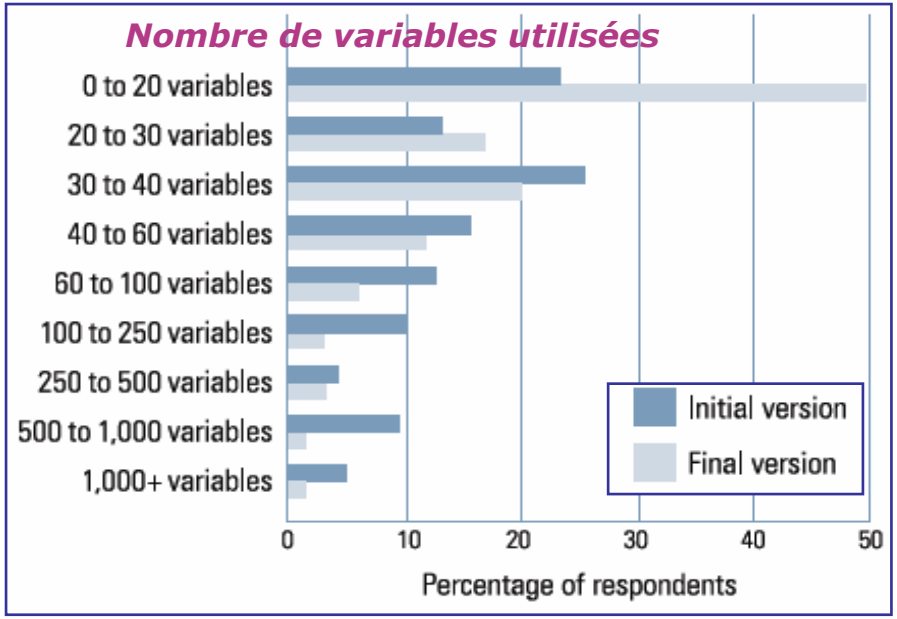


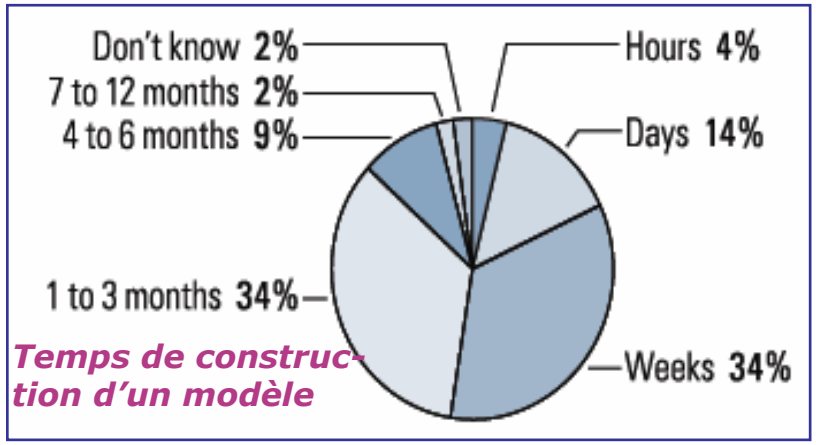
Eckerson, TDWI, 2007

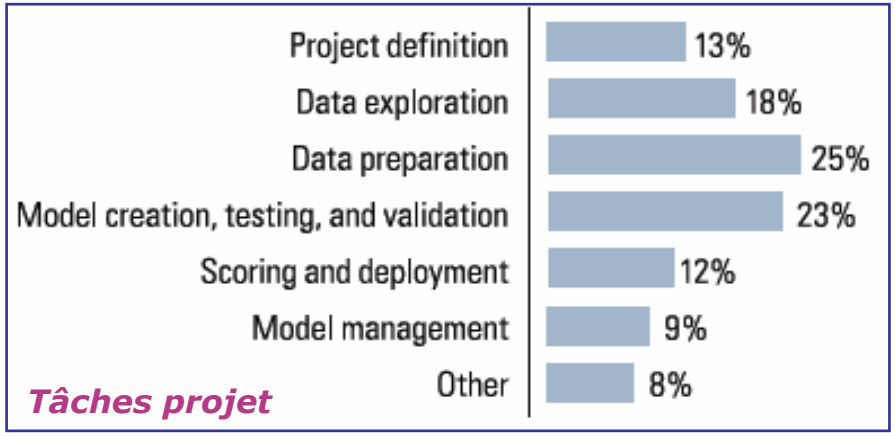
Le contexte industriel

Aujourd'hui, le processus data mining n'est pas efficace

- On n'utilise pas toutes les variables
- La manipulation des données est très lourde
- La construction d'un modèle prend très longtemps
  - Des semaines, voire des mois



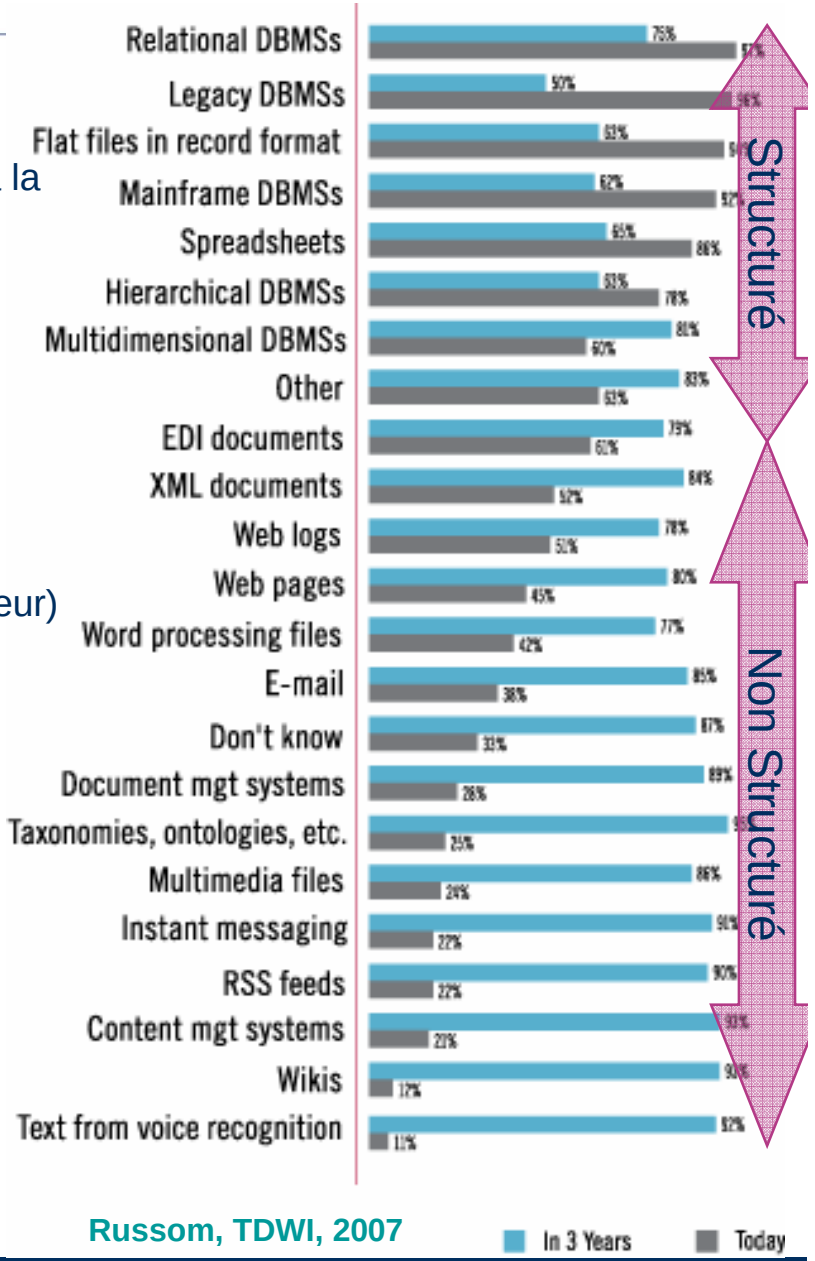




Eckerson, TDWI, 2007

Le contexte industriel – Les défis

- Productivité
  - 40% du temps de production du modèle est passé à la préparation des données
    - Peut-on réduire ?
  - Peut-on utiliser toutes les données ?
    - Volumes ?
    - Structuré / non structuré ?
- Scalabilité
  - Les volumes sont caractérisés par (largeur, profondeur)
  - La modélisation a 2 phases : construire & appliquer
    - Comment le temps varie avec le volume ?
  - Le temps réel est-il possible ?
    - À l'application (intégration, scalabilité)
- Automatisation
  - Le modèle peut-il être construit par
    - Des non experts ?
    - Une machine ?
  - Un modèle peut-il être appliqué et contrôlé par une machine ?



Russom, TDWI, 2007

Le contexte industriel – Les défis

1. **Défi n°1 : Intégration**
   - Le Data mining n'est jamais LA solution : mais une – petite – partie de la solution
     - L'application data mining doit être intégrée dans un système global
     - L'application data mining prend des inputs de & génère des outputs vers le reste du système
   - **Mots clés** : ouverture, standards
2. **Défi n°2 : Productivité**
   - Le Data mining doit apporter de la valeur (= €)
     - Exploiter toutes les données
     - Produire des résultats « actionnables »
     - À coût minimum
     - Simple à utiliser pour des non experts
   - **Mots clés** : Retour Sur Investissement
3. **Défi n°3 : Scalabilité**
   - Le Data mining doit tenir les volumes (données & nombre de modèles)
     - Exploiter dee ensembles de données MASSIFS
     - Produire AUTANT de modèles que nécessaire
   - **Mots clés** : temps pour produire un modèle en fonction de (largeur, profondeur) des données
4. **Défi n°4 : Automatisation**
   - Le Data mining doit faire tout ce qui précède automatiquement (?)
     - Produire les modèles
     - Détecter les problèmes …
   - **Mots clés** : automatisation, contrôle

Le contexte industriel – Les défis

Ce qu'on voit déjà

- Masses de données : Milliers de variables, 10-100 millions de lignes
- Beaucoup de modèles : 100 – 1000 modèles / an / semaine / jour
- Ressources limitées : Quelques utilisateurs (10 – 30 ?)

… généralement dans un secteur de l'entreprise (Marketing, Risque …)

Ce qu'on commence à voir

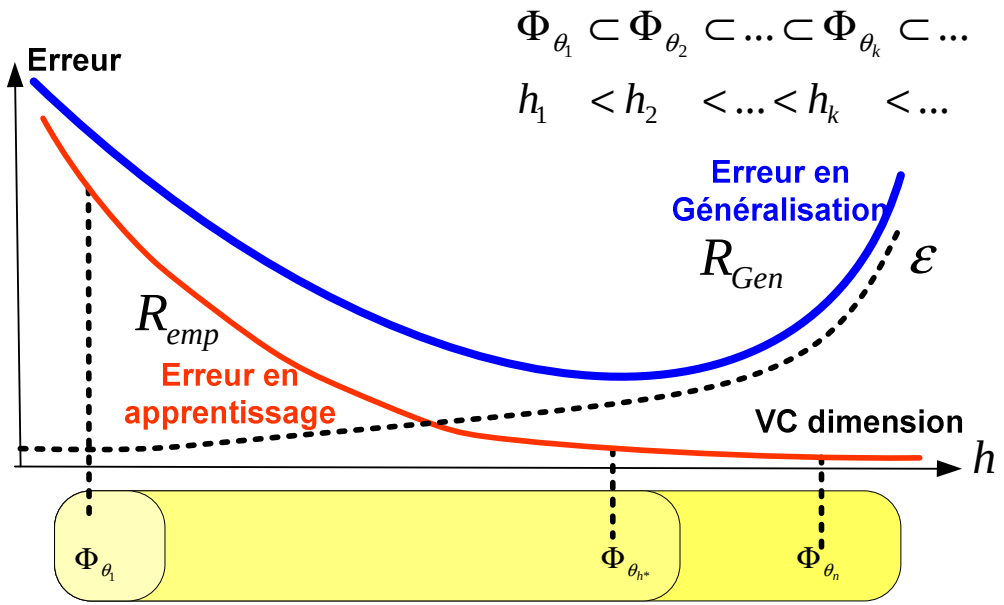
- Des initiatives à l'échelle de l'entreprise
  - Conception, production, vente, maintenance, service client, marketing
- Des ressources à l'échelle de l'entreprise
  - Beaucoup d'utilisateurs (100 – 1000)

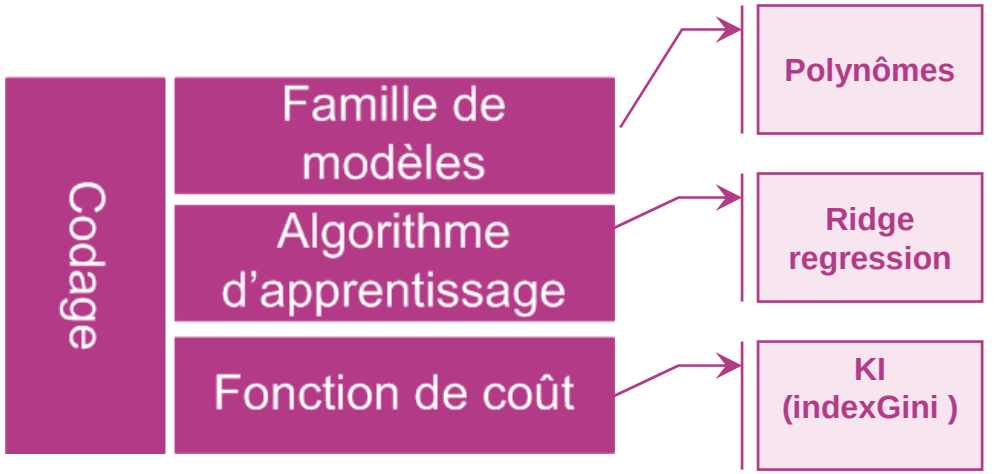
… ce qui va donner une nouvelle dimension au data mining

Je vais illustrer quelques uns des défis précédents en montrant la solution apportée par KXEN

Implémentation KXEN

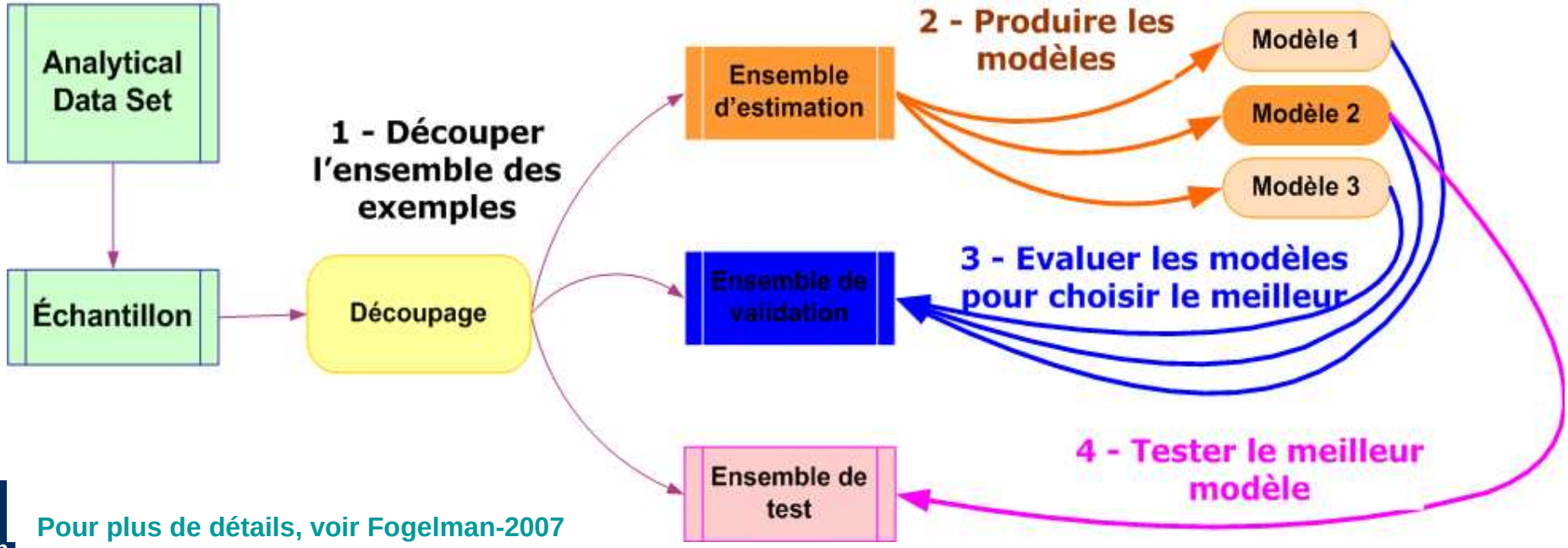
- KXEN a été conçu pour les applications industrielles du data mining
- KXEN est basé sur la **SRM – Structural Risk Minimization** de Vapnik
  - Stratégie pour contrôler le **compromis précision / robustesse**
- KXEN réalise
  - Un codage automatique
    - Non linéaire
  - Puis une régression / classification
    - Polynomiale
- Ce qui permet
  - Intégration
  - **Productivité**
  - **Scalabilité**
  - Automatisation





Implémentation KXEN

- En pratique, pour un modèle final, KXEN en produit beaucoup (SRM)
  - Selon la complexité de la variable, le codage nécessite de 10 à 30 modèles
  - Ensuite environ 100 modèles (pour la régression)
- KXEN utilise des techniques de « data streams »
  - On ne duplique pas les données en mémoire mais on ne fait que les lire
  - Très peu de passes sont nécessaires
- Temps de construction d'un modèle
  - Pratiquement linéaire en largeur & profondeur



Pour plus de détails, voir Fogelman-2007

Agenda

- Le Data Mining industriel
  - Un peu d'histoire …
  - Les données
  - Les défis
- Quelques exemples
  - **Le nombre de variables**
  - Le nombre de modèles

Le nombre de variables

On a vu que les modèles utilisent peu de variables

Y a-t-il un intérêt à en utiliser beaucoup ?

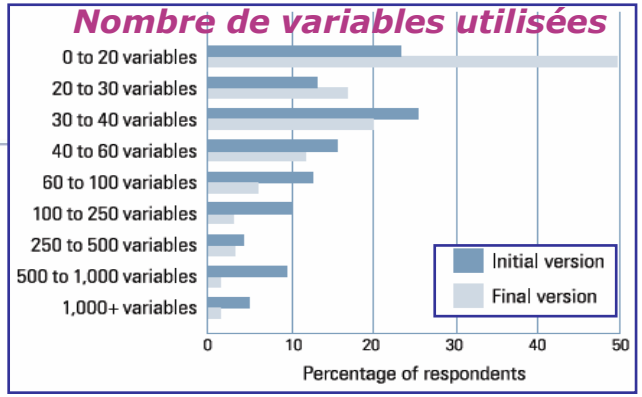
- En exploitant toutes les variables disponibles
  - 3 000, 5 000, 10 000 ?
- En créant de nouvelles variables
  - Agrégats
  - Variables comportementales
  - Variables textuelles
  - Variables « réseaux sociaux » …

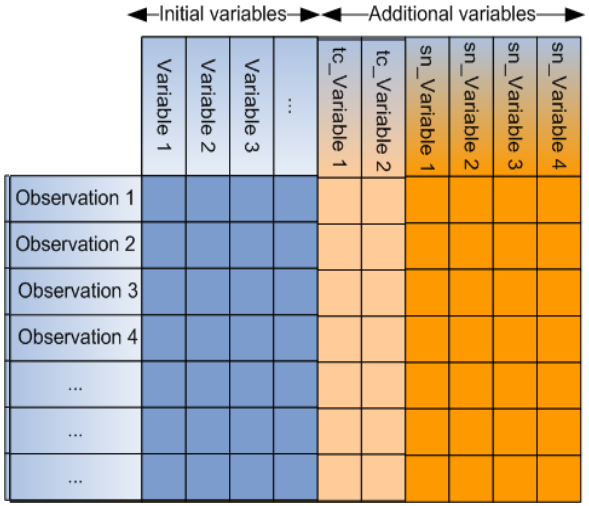
Le but

Améliorer la performance des modèles

Le défi

Faire des modèles avec des milliers de variables





Nombre de variables

Sears	900
Grande banque	1 200
Vodafone D2	2 500
Barclays	2 500
Rogers Wireless	5 800
HSBC	8 000
Credit card	16 000

Le nombre de variables

- Utiliser des données très détaillées
- Produire des agrégats
- Produire des variables calculées

… peut apporter de la valeur

- Mais le nombre de variables croît exponentiellement
  - Ex : 45 M transactions / jour

Comment « choisir » les variables ?

- Il existe beaucoup de méthodes
  - Pas automatique !
  - Les experts se trompent toujours quand il y a des milliers de variables à analyser

… avec un modèle data mining

Extreme Granularity Data

- Numerous sources of granular data (transaction data, payment data, call data, etc.)
- Granularity and detail creates value *if* you can aggregate intelligently
- Number of attributes grows exponentially as you consider time series, interactions, and transformations

Possible approaches to variable selection

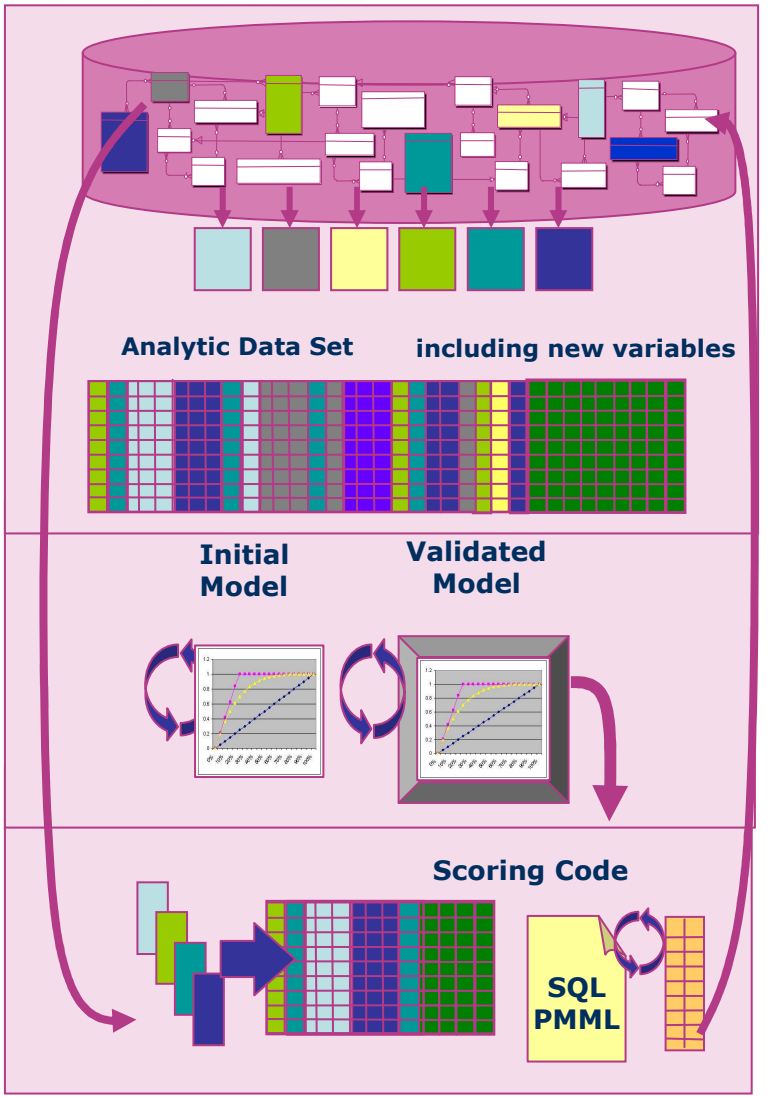
- Use the same variables we used last year
- Based on experience and expertise, select the 500 variables that are most likely to be useful. Then use statistics to pick the 10 best.
- Use all the variables and let the data tell you which are useful

D'après DISCOVER FINANCIAL SERVICES

Exploiter toutes les variables

Le processus d'analyse

- Construire l'ADS (Analytic Data Set)
  - Extraire les données
  - Les transformer, agréger, …
  - Créer l'ADS
- Construire le modèle
  - Produire le modèle initial
  - Affiner, sélectionner les variables
  - Produire le modèle final
- Appliquer le modèle
  - Extraire les données
  - Les transformer, agréger, …
  - Créer l'ADS
  - Appliquer le modèle
  - Exporter les résultats vers la base de données



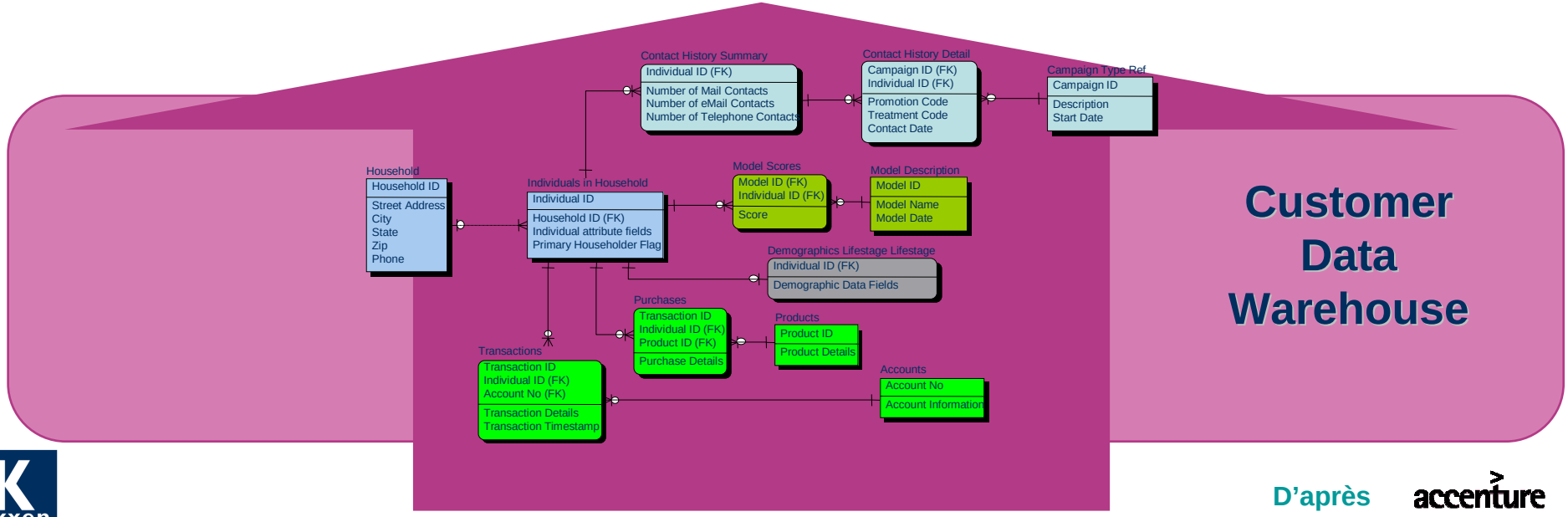
D'après Teradata a division of NCR

Exploiter toutes les variables

- L'ADS contient toutes les variables
  - Exemple de Teradata : c'est une vue. Il n'y a pas de mouvement de données

HH-ID	CUST-ID	NAME	VALUE_SEG	BEHAV_SEG	LIFESTY_SEG	LIFESTG_SEG	EQUITY_12	EQUITY_24	LTV	...	AGE	INCOME_CD	EDUCATION	...
2347387474	4797978698	Gustavo	2	5	8	3	37.22	28.18	49.8	...	28	7	14	...
7879973979	2439970274	Susan	3	3	6	5	18.88	28.97	154.32	...	42	9	18	...
9870908	879979	Andre	1	1	18	4	-1.38	-12.8	-48.76	...	61	5	12	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

ID FIELDS | BEHAVIOR FIELDS | DEMOGRAPHIC FIELDS | MODEL SCORES | CONTACT HISTORY



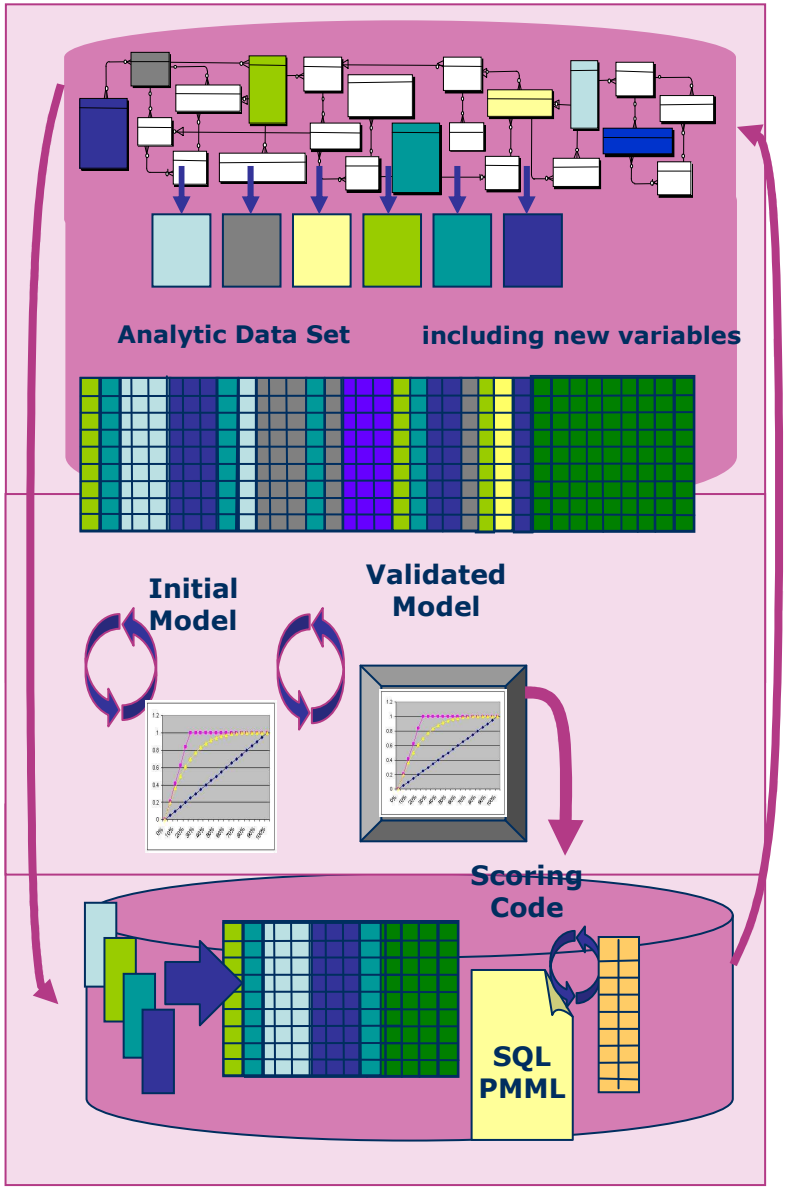
D'après accenture

Exploiter toutes les variables

Le processus d'analyse **< 1 semaine**

- Construire l'ADS (Analytic Data Set)
  - ~~Extraire les données~~ **3 jours**
  - Les transformer, agréger, …
  - Créer l'ADS
- Construire le modèle **< 1 jour**
  - Produire le modèle initial
  - Affiner, sélectionner les variables
  - Produire le modèle final
- Appliquer le modèle **< 1 jour**
  - ~~Extraire les données~~
  - Les transformer, agréger, …
  - Créer l'ADS
  - Appliquer le modèle
  - Exporter les résultats vers la base de données

« In-database Mining »

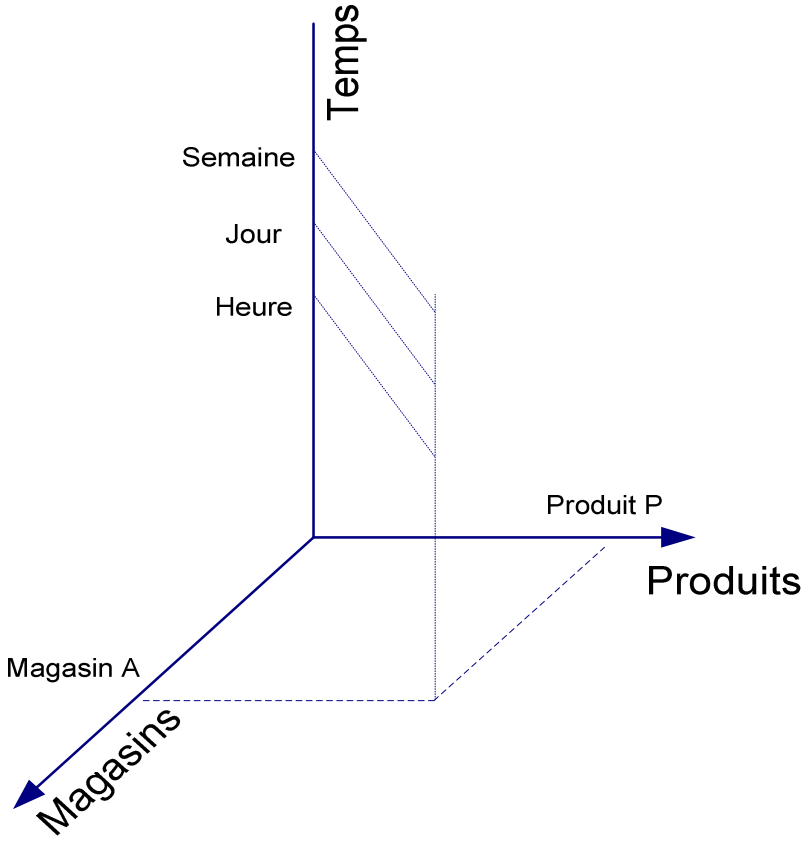


Créer des variables – Agrégats

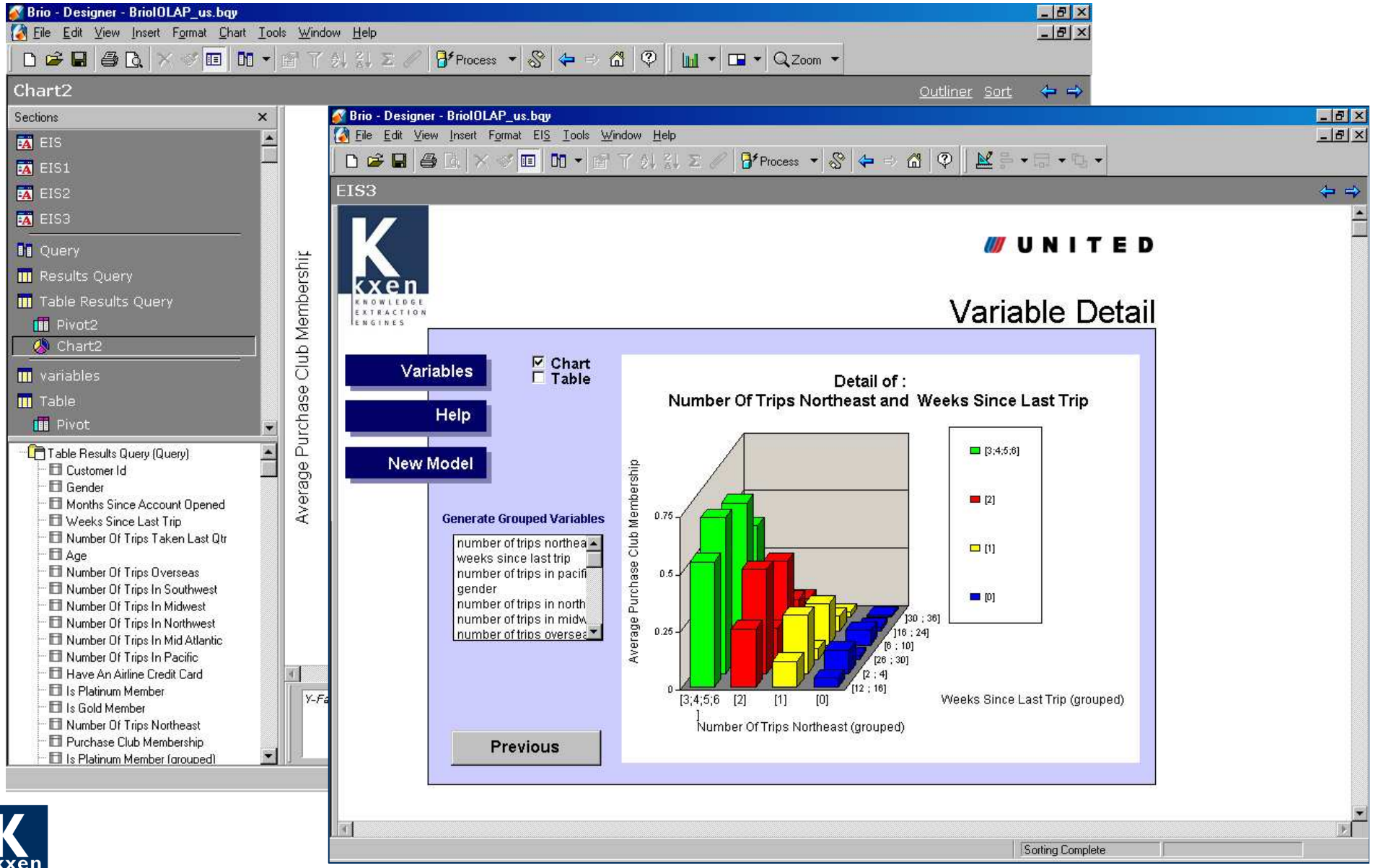
En informatique décisionnelle

- Données de détail
  - Granularité « fine »
- Données agrégées
  - Selon les différents « axes »
  - À quel niveau d'agrégation ?
- Indicateurs
  - KPI (Key Performance Indicators)
- Solution
  - Produire « tous » les agrégats

    Le nombre d'agrégats possibles est très grand
    - Des dizaines de milliers
  - Construire un modèle / KPI
  - Retenir les agrégats les plus significatifs
- Le modèle data mining permet ainsi de produire de meilleurs tableaux de bord

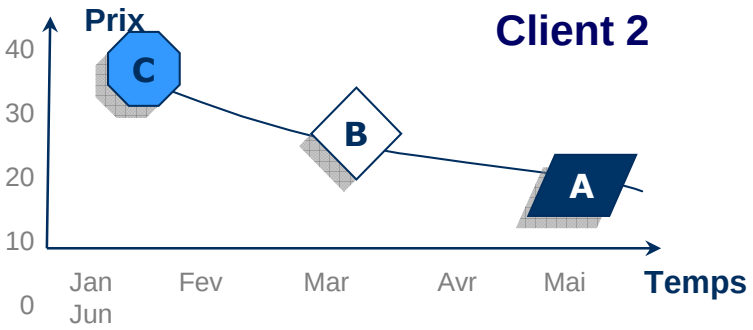


Créer des variables – Agrégats

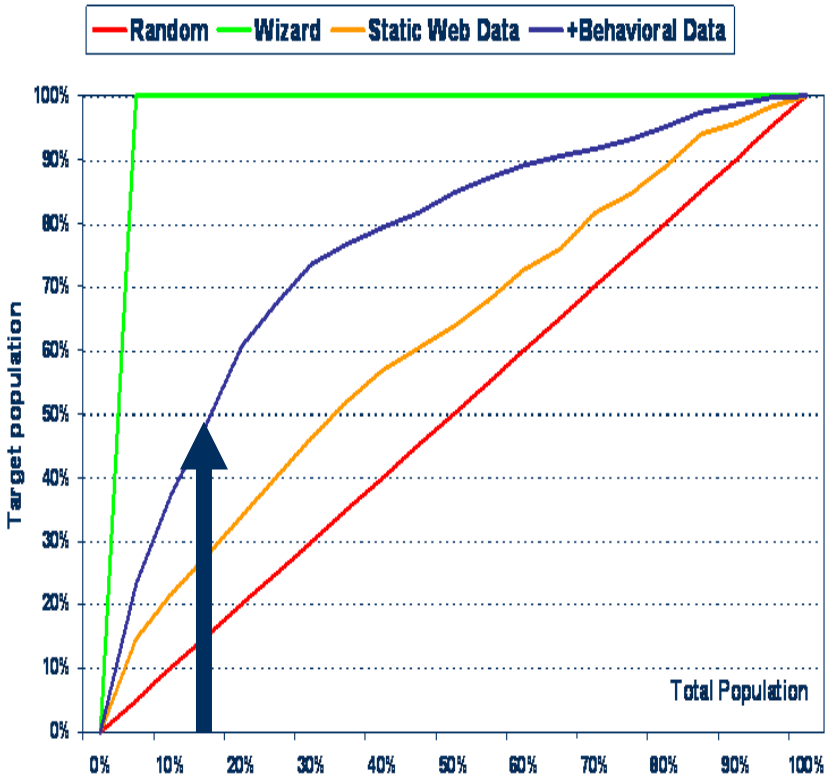


Créer des variables – Variables comportementales

- Beaucoup d'applications comportent des données transactionnelles
  - Achats de produits (tickets de caisse ou achats en ligne)
  - Transactions carte bancaire …
- On crée des « variables comportementales »
  - Transition de transaction A vers transaction B
- On obtient un meilleur modèle
- Le volume généré est énorme !
  - Des milliers de variables supplémentaires

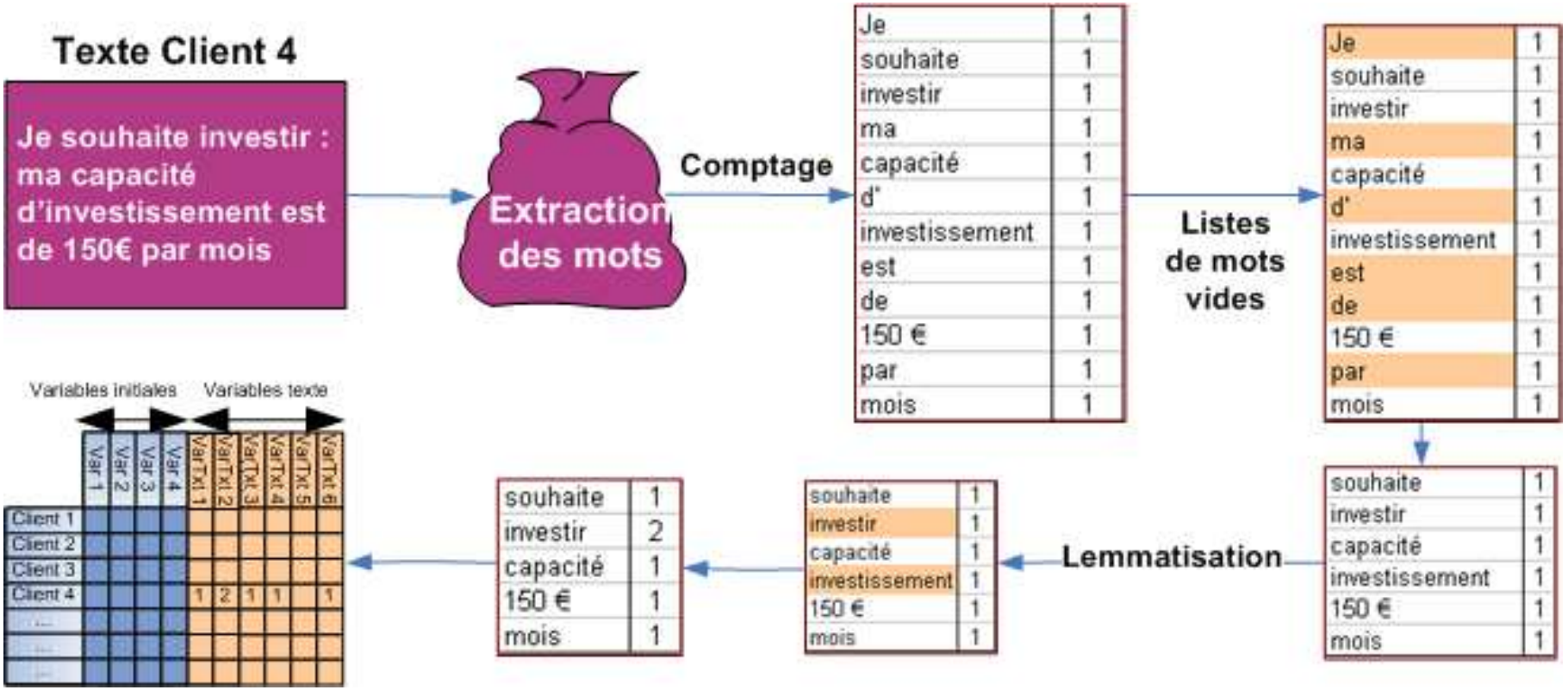


	LastStep	A	B	C	out : A	A : B	B : C	out : C	C : B	B : A	Session Continue?	Next State?
Cust. 2		0	0	1	0	0	0	1	0	0	Y	B
Cust. 2	1	0	1	1	0	0	0	1	1	0	Y	A
Cust. 2	2	1	1	1	0	0	0	1	1	1	N	null



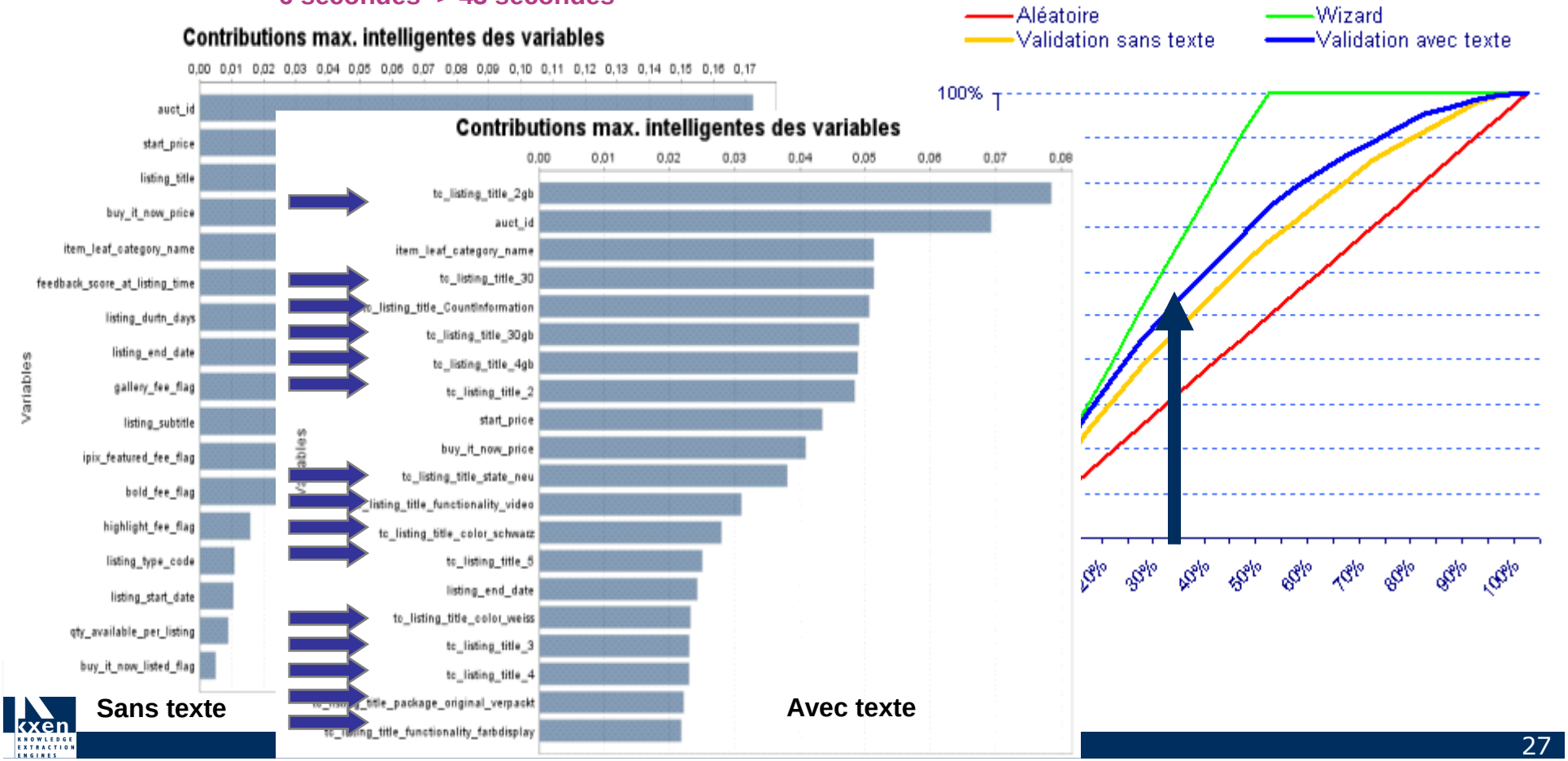
Créer des variables – Variables textuelles

- On extraie les « variables textuelles » des champs texte
  - Des milliers de variables supplémentaires



Créer des variables – Variables textuelles

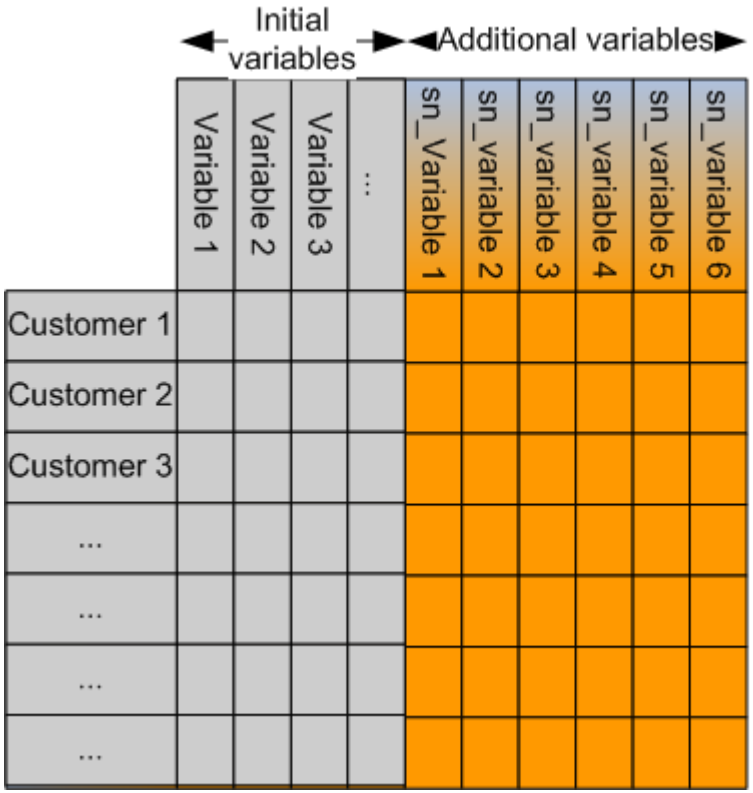
- eBay Germany (Data Mining Cup 2006) http://www.data-mining-cup.com/
  - 8000 enchères sur des produits en vente sur eBay
  - Déterminer un modèle pour prévoir, pour chaque nouvelle enchère, si le prix de vente final sera plus grand que le prix moyen de la catégorie du produit proposé
  - **On ajoute 1000 variables textuelles**
    - **6 secondes -> 43 secondes**

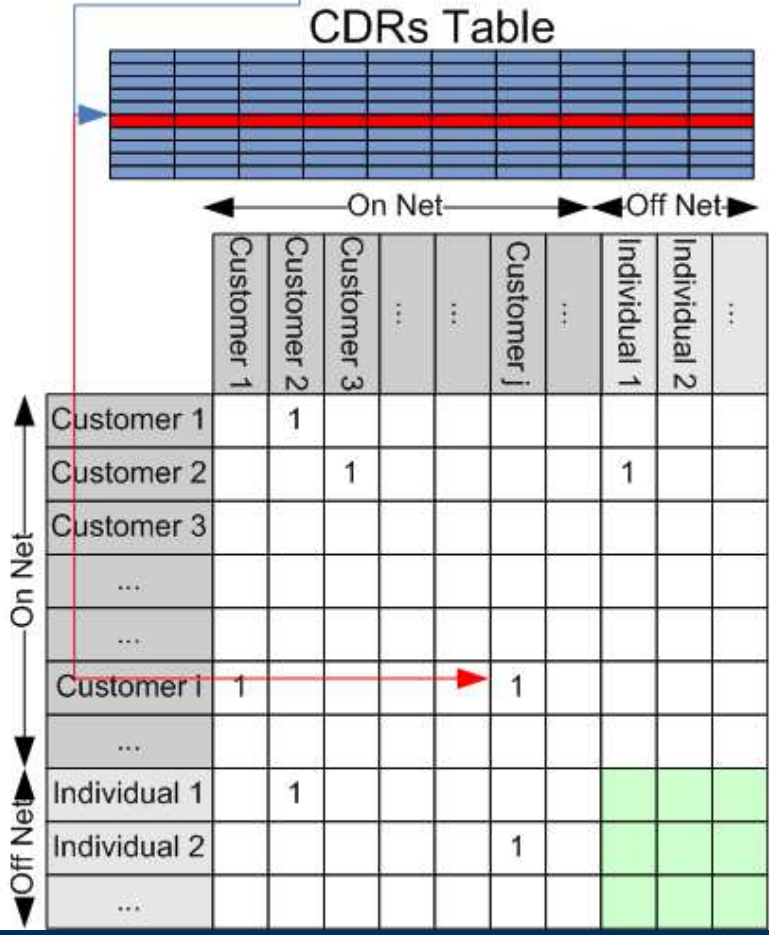


Créer des variables – Variables « réseaux sociaux »

Un exemple dans les telecom

- Construire le réseau social
- Extraire les variables « réseaux sociaux »
  - Quelques dizaines de variables supplémentaires





Créer des variables – Variables « réseaux sociaux »

- 39 variables

IKAN CONSULTING

Circle analysis

- Count the number of contacts
- Rank best contacts

sn_voi_InD
sn_voi_InD5
sn_voi_OutD
sn_voi_OutD5
sn_voi_UndD
sn_voi_UndD5

sn_sms_InD
sn_sms_InD5
sn_sms_OutD
sn_sms_OutD5
sn_sms_UndD
sn_sms_UndD5

sn_mms_InD
sn_mms_InD5
sn_mms_OutD
sn_mms_OutD5
sn_mms_UndD
sn_mms_UndD5

sn_all_InD
sn_all_InD5
sn_all_OutD
sn_all_OutD5
sn_all_UndD
sn_all_UndD5
sn_all_Circle Size

sn_Deg Offnet
sn_Deg Onnet

Connection analysis

- Profile contacts
- Describe customer by his contacts
- Social boundaries

sn_Nb Acquis_After
sn_Nb Acquis_Before
sn_Nb Churn in Circle

Community analysis

- Identify communities
- Add each customer to his community

Social leader analysis

- Identify social leader
- Analyze the impact of a social leader

sn_Centrality_sms_voi
sn_Centrality_sms only
sn_Centrality_voi only

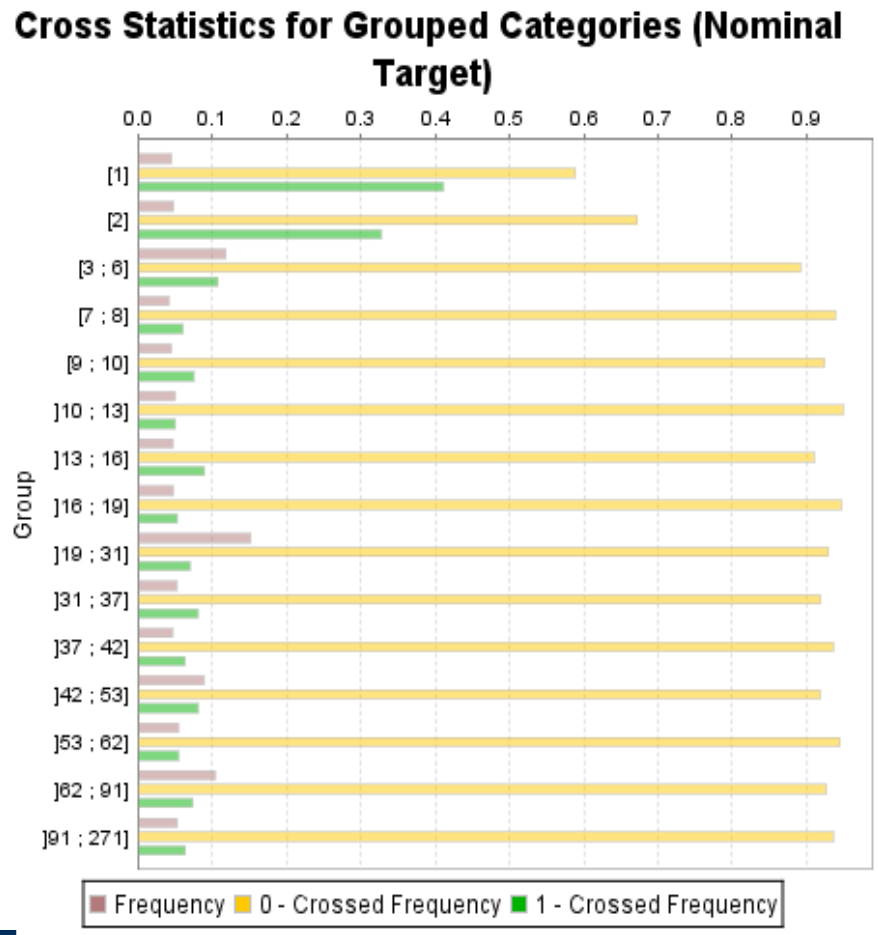
sn_Soc Deg_voi only
sn_Soc Deg_sms only
sn_Soc Deg_sms_voi
sn_Soc Deg Tot_voi only
sn_Soc Deg Tot_sms only
sn_Soc Deg Tot_sms_voi

kxen KNOWLEDGE EXTRACTION ENGINES

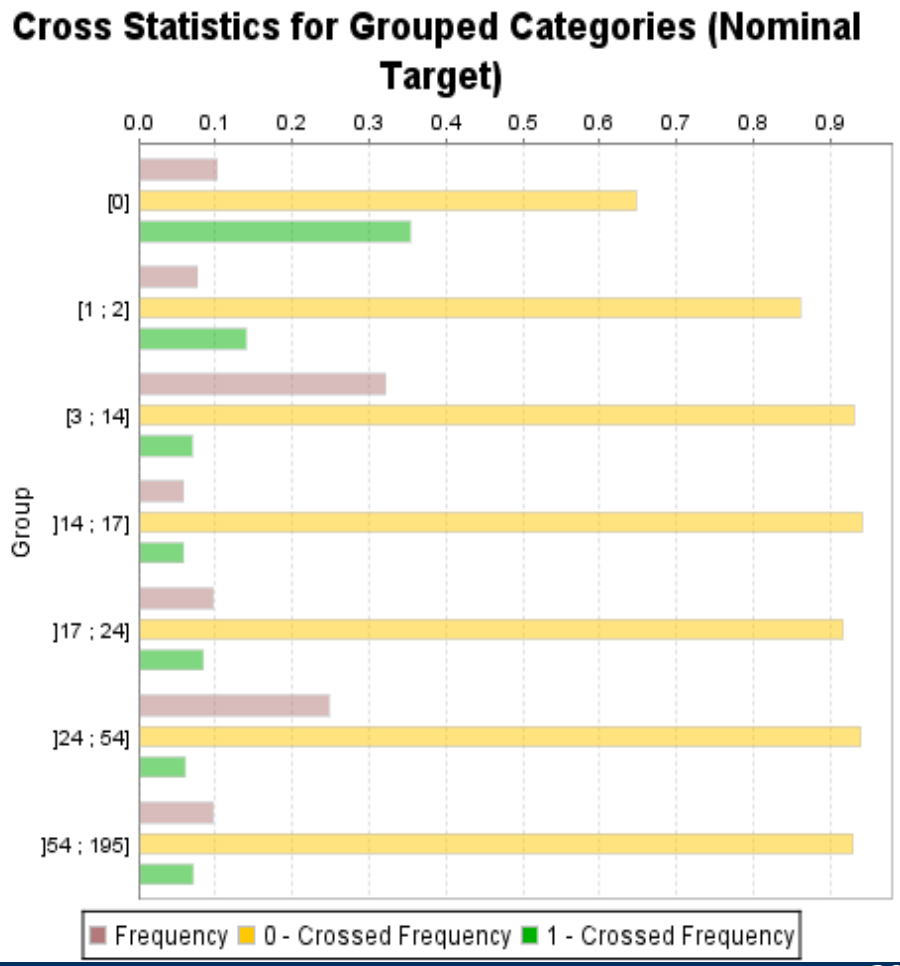
Créer des variables – Variables « réseaux sociaux »

Qui churne ? Les clients peu connectés !

sn_Deg Offnet

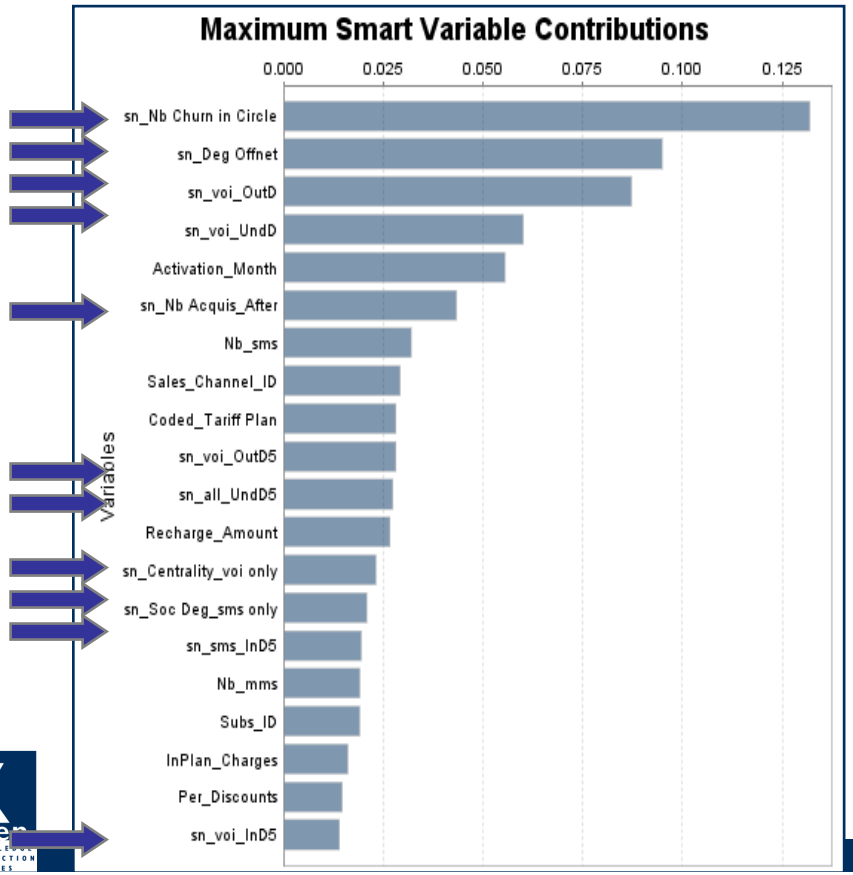


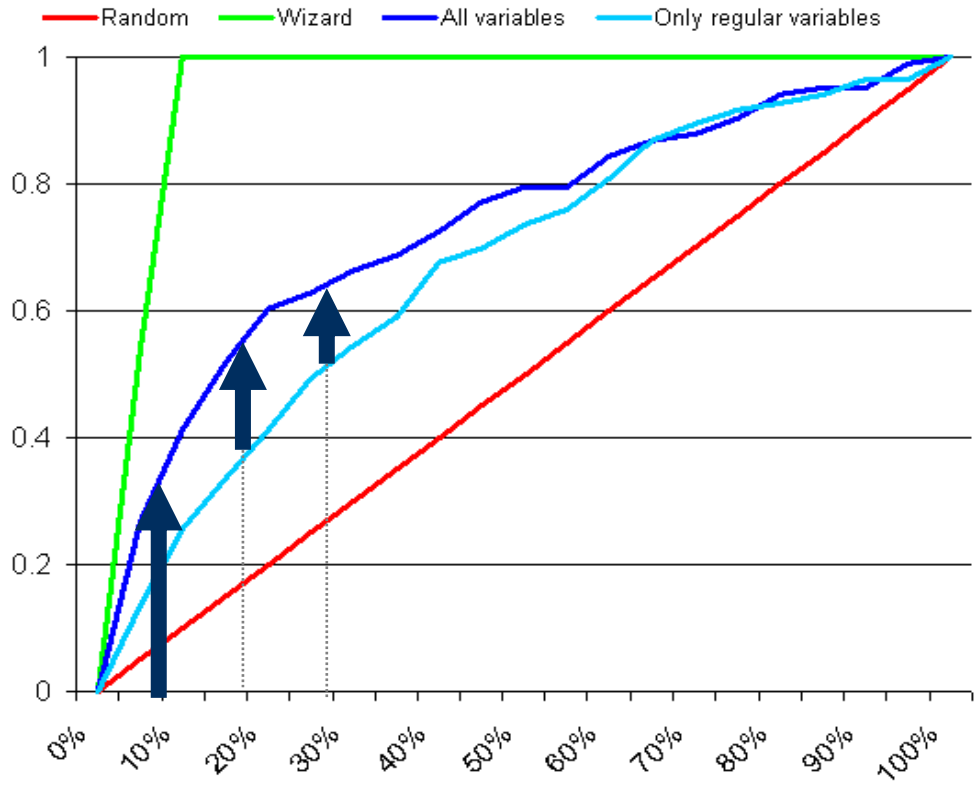
sn_voi_outD

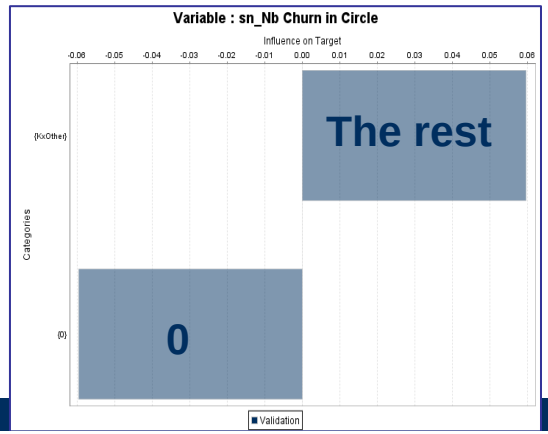


Créer des variables – Variables « réseaux sociaux »

- Les variables « réseaux sociaux » augmentent le lift
  - Globalement : 40%
  - Premier décile : 67%
  - Second décile : 47%



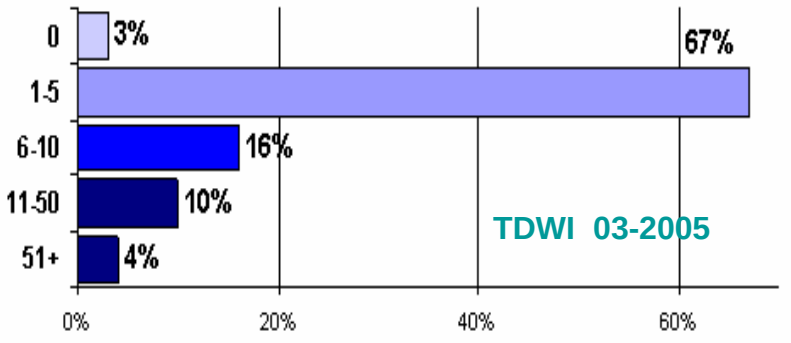




Agenda

- Le Data Mining industriel
  - Un peu d'histoire …
  - Les données
  - Les défis
- Quelques exemples
  - Le nombre de variables
  - **Le nombre de modèles**

Le nombre de modèles



Les entreprises produisent peu de modèles

Y a-t-il un intérêt à en produire beaucoup ?

- Beaucoup de produits / actions
  - Il faut faire – au moins – un modèle par produit par campagne
  - Vente sur le Web & Longue Traîne
- Modèles refaits fréquemment
  - La distribution des données change vite (Web)
- Modèles « fins »
  - La performance sur une population homogène est meilleure
    - Produits /marchés, segments clients, canaux, géographie …

Le but

Améliorer la performance des modèles

Le défi

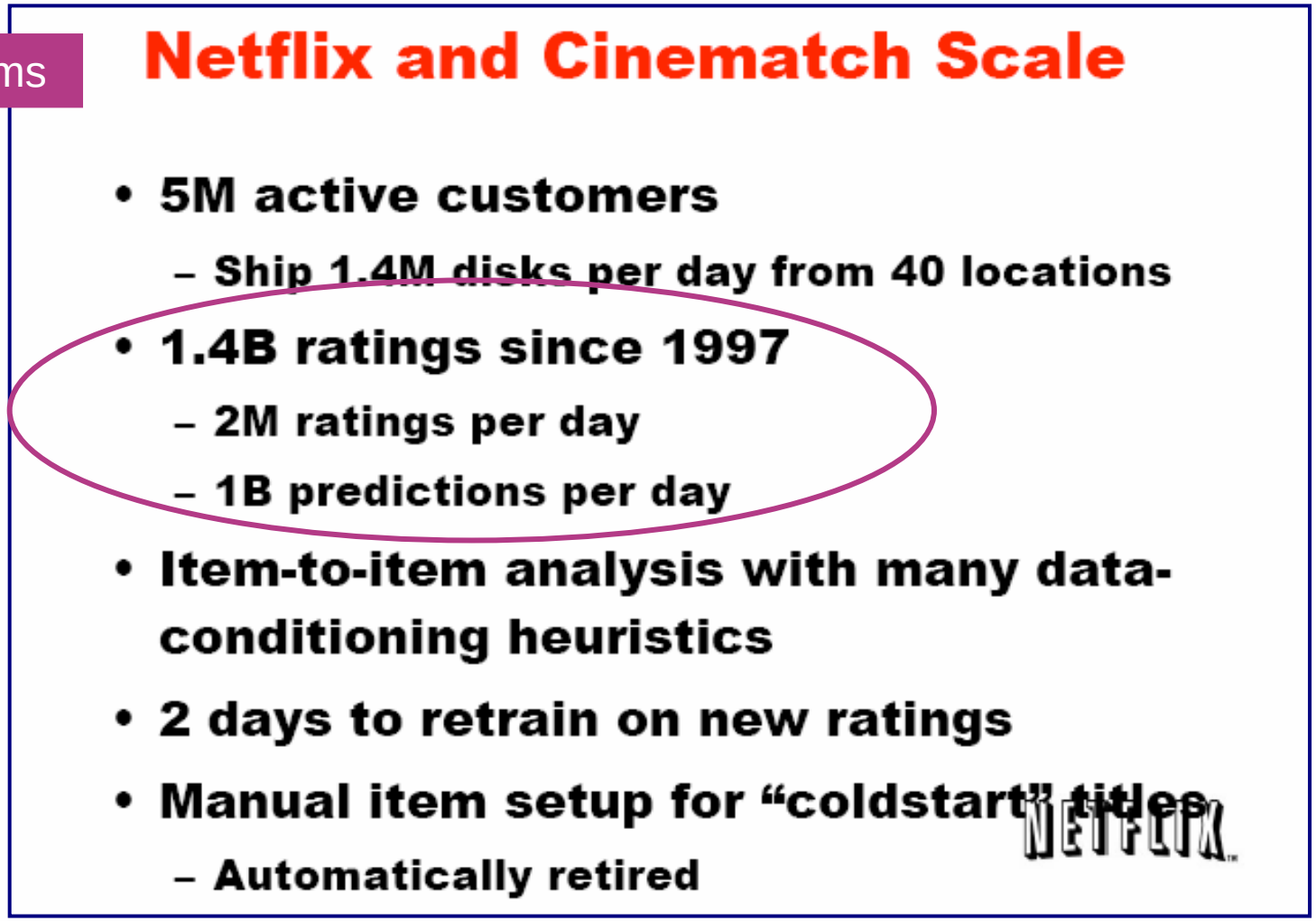
Faire des milliers de modèles

Nombre de Modèles /an	
Vodafone D2	760
Market research	9 600
Cox Comm.	28 800
Real estate	70 000
Lower My Bills	460 000

Le nombre de modèles – Beaucoup de produits

- Un exemple

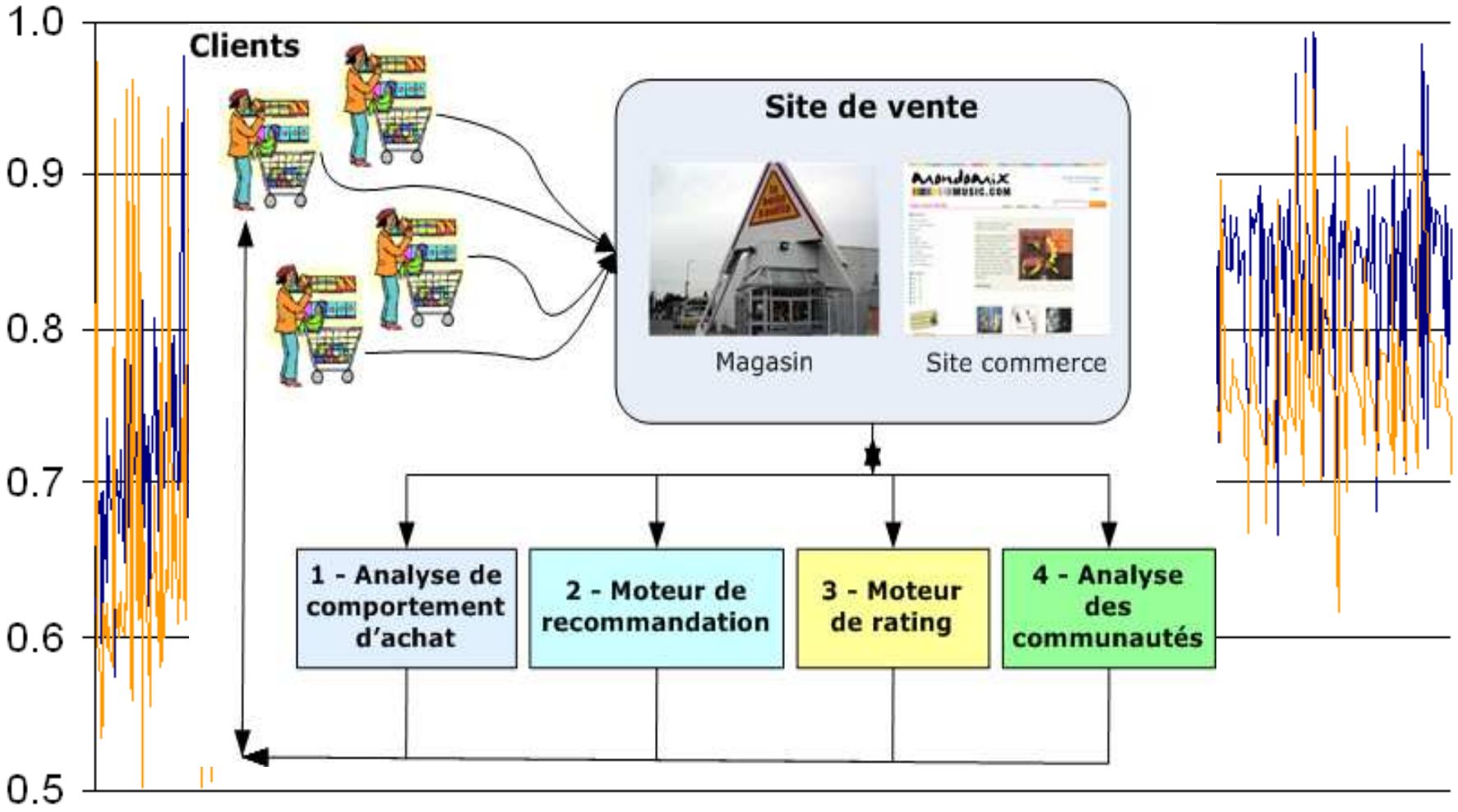
65 000 films



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

Le nombre de modèles – Beaucoup de produits

- CADI – Composants Avancés pour la Distribution » (ANR)



Le nombre de modèles – Beaucoup de campagnes

Les entreprises utilisent des modèles data mining pour cibler leurs campagnes

- La tendance est à des campagnes de plus en plus nombreuses

Exemple – Vodafone : 716 campagnes / an

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

Vodafone needs for Tier 1 Telco

	# Analysis /Year
• **Segmentations** 2*2*10	40
• **Churn in General** 2 * 3 * 2 *3	36
• **Churn per product** 2 * 3 * 2 * 4 * 10	480
• **Cross sell : segments*offers** 2 * 4 * 10	80
• **Acquisition** 2 * 4 * 10	80

This means trying to create 716 models per year…

Le nombre de modèles – Modèles fins

Construire un modèle sur une population homogène

- Géographie
- Détention produits
- Segment clients …

fournit un modèle plus précis (meillerus résultats opérationnels)

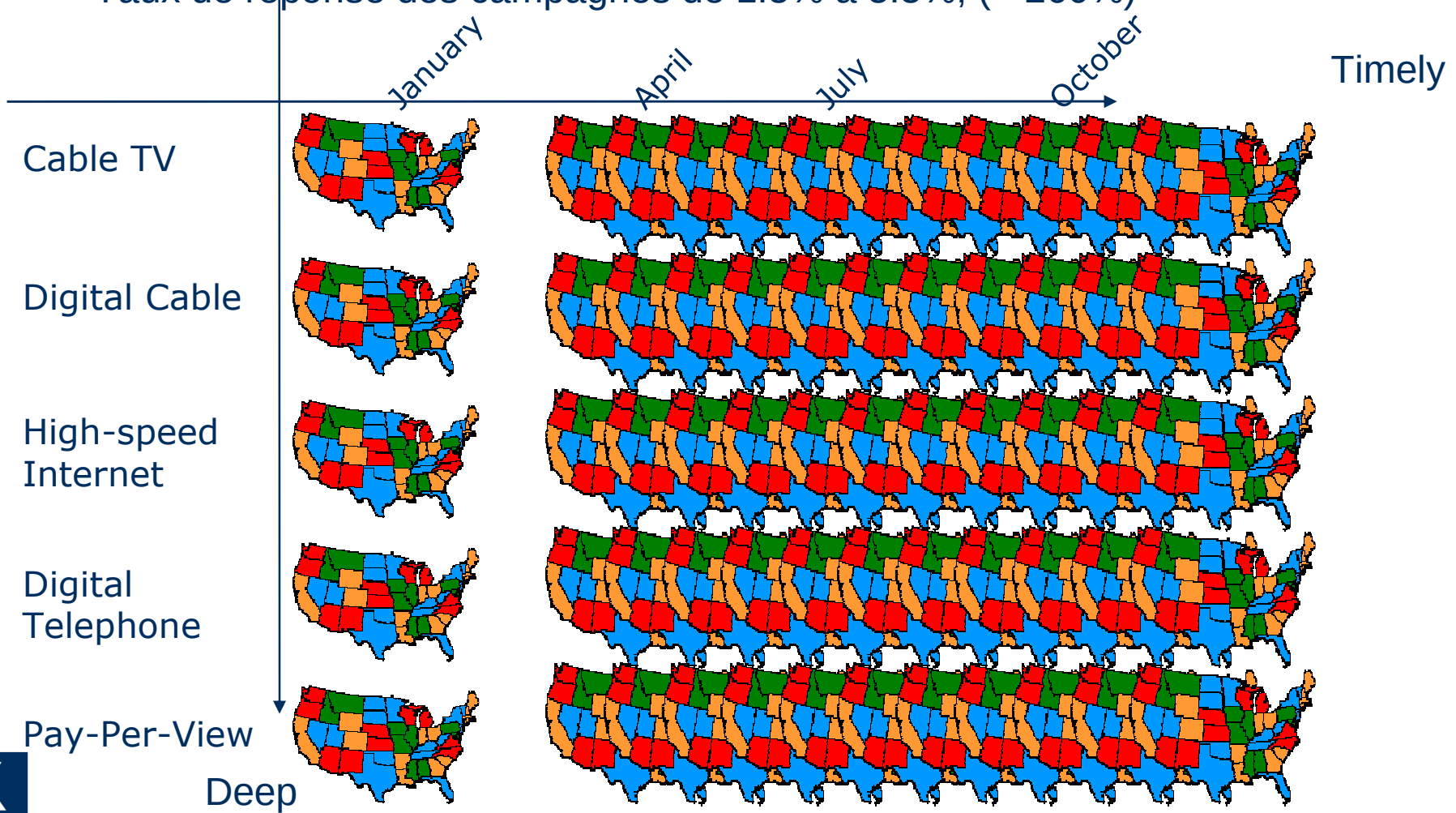
Mais

- Le nombre de modèles produits devient vite très grand
  - De 7 en moyenne / an, on passe à quelques centaines / milliers par an

Le nombre de modèles – Modèles fins

Exemple : Cox Communications

- 28 régions * 5 produits * 12 mois = **1 680 modèles** cross-sell en production / an
- Taux de réponse des campagnes de 1.5% à 5.5%, (+ 260%)



Conclusion

Le data mining peut être utilisé de façon industrielle dans les entreprises

Il doit pour cela répondre à 4 défis majeurs

- Intégration
- Productivité
- Scalabilité
- Automatisation

Gain en productivité	
Rogers Wireless	7x
Vodafone D2	10x
Sears	8x
Belgacom	12x

Manipuler des masses de données,
& produire des masses de modèles impose des contraintes

- Manipulation & codage automatisés des données
- Algorithmes simples & robustes
- Modules ouverts & obéissant aux standards du marché

Conclusion

In fine, **le critère de réussite sera toujours le bilan économique** € $

- Grande Banque : time-to-market ↓ 66%, 100% ROI en 1 modèle
- Bank of Austria : $ 67m new business en 1 trimestre, taux de réponse ↑ 300–500%
- Bell Canada : 100% ROI en 1 modèle
- Sears : coûts opérationnels ↓ 50%, temps de développement ↓ 90%,
- Barclays : courriers ↓70%, 15% taux de conversion, ventes ↑35%, coûts ↓30%, profit ↑ 35%
- Grand opérateur telco : appels sortants ↓ 70%, profit ↑ 20%
- E.ON : ventes ↑ 20%, $5.8M / an ventes supplémentaires en 1 modèle
- Cox : taux de réponse ↑ 260%, produits par foyer ↑ 14%, ROI en 2 mois
- Live Person : ventes en ligne ↑ 200 – 700%

Références

- Wayne W. Eckerson, Predictive Analytics. Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report. Q1 2007. https://www.tdwi.org/Publications/WhatWorks/display.aspx?id=8452
- Françoise Fogelman-Soulié, Erik Marcadé : Mining Massive Data Sets. A paraître dans NATO Workshop, http://videolectures.net/mmdss07_gazzada/
- Gareth Herschel, Gartner Customer Relationship Management Summit 2006.
- Andrew Moore, New Cached-Sufficient Statistics Algorithms for quickly answering statistical questions, KDD'07, http://www.sigkdd.org/kdd/2006/docs/presentations/andrewMoore06Keynote.pdf
- Philip Russom, BI Search & Text Analytics. TDWI Best Practices Report. Q2-2007. https://www.tdwi.org/Publications/WhatWorks/display.aspx?id=8449
- J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli and J. Vandewalle eds : Advances in Learning Theory: Methods, Models and Applications. NATO Science Series, vol 190. http://www.iospress.nl/loadtop/load.php?isbn=ncss

KDD

KDD aux US

KDD'09 à Paris

- 29 juin-1er juillet
- Paris Marriott

Contact Chairman KDD'09

- Françoise Fogelman Soulié
  francoise@kxen.com

Participez