ECML PKDD 2008

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

15 - 19 September 2008, Antwerp, Belgium

Industrializing Data Mining, Challenges and Perspectives

Françoise Fogelman Soulié

francoise@kxen.com

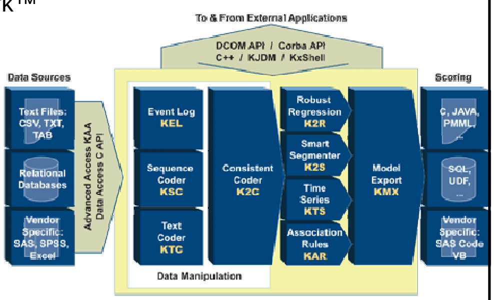
KXEN KNOWLEDGE EXTRACTION ENGINES

THE DATA MINING AUTOMATION COMPANY ™

ECML PKDD 2008
September 17, 2008
Antwerp, Belgium

KXEN

- **KXEN is an editor of a data mining software**
  - The KXEN Analytic Framework™ is a suite of predictive and descriptive modeling engines that create robust analytic models fast and easily
  - KXEN's products are based upon Vladimir Vapnik's Statistical Learning Theory (Structural Risk Minimization)



- **Our vision**
  - KXEN wants to make predictive analytics part of everyday corporate business decisions
- **Our mission**
  - Our mission is to embed advanced analytics into existing enterprise applications and business processes

THE DATA MINING AUTOMATION COMPANY ™

http://www.kxen.com/

2

Data Mining industrial applications

Data Mining industrial applications are in

- Telecommunications
- Bank & Finance
- Retail
- … and increasingly in « Web » companies

For

- Marketing
- Risk, fraud
- Security
- On-line retail and services, advertising, key-word optimization …

- In my talk, I will rely upon examples taken from KXEN customers in 3 sectors

Telecommunications | Banking & Finance | Retail

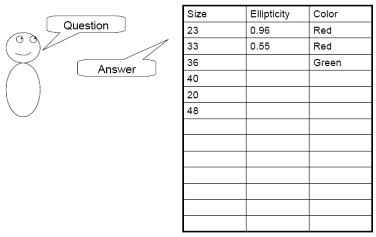
Agenda

- Which world is this ?
- Data Mining in the real world
- Some examples

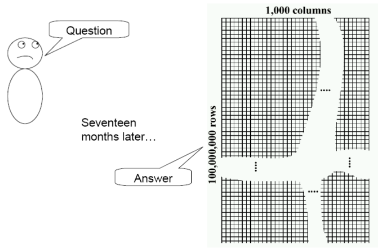
A little bit of history

What has happened ?

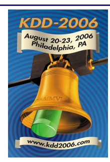
Data Analysis: The old days



Data Analysis: The new days



Andrew Moore



Data

Large in	
Neural Networks	**Statistics**
100,000 Weights	50 parameters
50,000 examples	200 cases

The volume of data has exploded

In the 90s

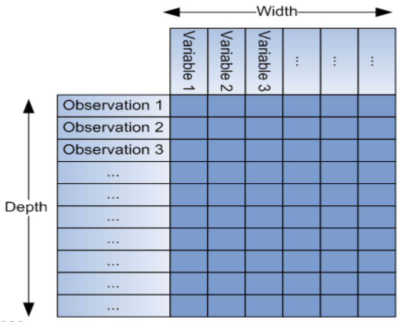
Today

- **Web transactions** Fayyad, KDD 2007
  - At Yahoo !
    - Around **16 B events / day**
    - 425 M visitors / month
    - **10 Tb data / day**
- **RFID** Jiawei, Adma 2006
  - A retailer with 3,000 stores, selling 10,000 items a day per store
    - **300 million events per day** (after redundancy removal)
- **Social network** Kleinberg, KDD'07
  - **4.4-million-node network** of declared friendships on blogging community LiveJournal
  - **240-million-node network** of all IM communication over one month on Microsoft Instant Messenger
- **Cellular networks**
  - A telecom carrier generates **hundreds of millions of CDRs / day**
  - The network generates technical data : **40 M events / day** in a large city

Data

Just how big are big data sets ?

- **Depth**
  - Up to 100 Million lines
  - Or Billion ?
- **Width**
  - Thousands of attributes
  - Or Million ?



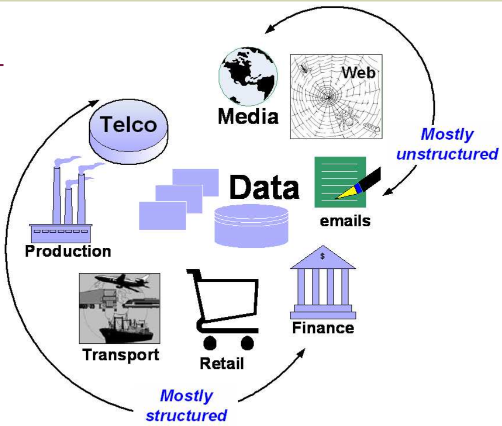
If they're big today, wait for to-morrow

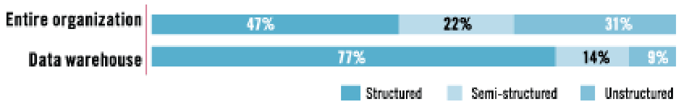
- **Size of databases**
  - X2-3 every 2 years

Data

Many different

- **Sources**
- **Types**
  - Structured
  - Unstructured
    - Text
    - Image
    - Video
    - Audio …
- **Volumes**
  - Web dominates !
- **Lots more data out there**
  - X 10 ? X 100 ?





Russom, TDWI 2007

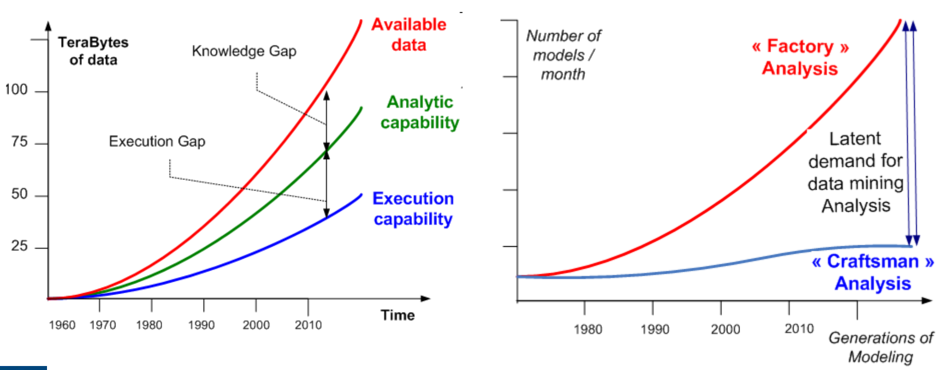
Agenda

- Which world is this ?
- Data Mining in the real world
- Some examples

THE DATA MINING AUTOMATION COMPANY ™

What are the issues in the real-world ?

- Data mining provides ways to define actions
  - A model not used for action is a useless cost
- Data volume grow exponentially : number of models must too



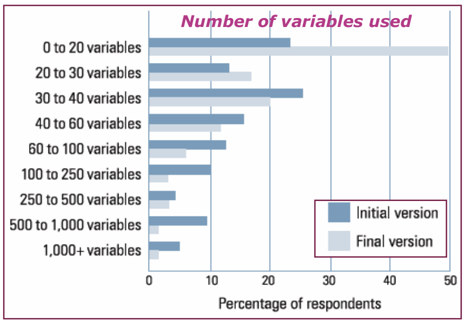
Herschel, Gartner 2006

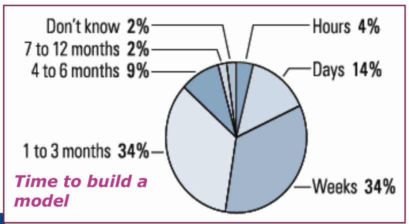
THE DATA MINING AUTOMATION COMPANY ™

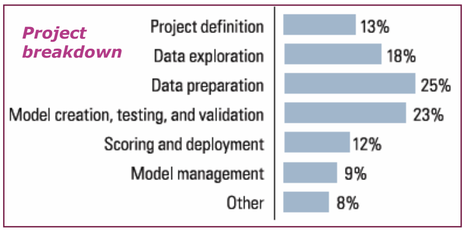
Data mining in practice

The data mining process is not very efficient in practice

- **It does not make use of all variables**
- **It spends a large part of modeling time on data manipulation**
- **And as a result, it still takes a long time to build a model**







Eckerson, TDWI, 2007

THE DATA MINING AUTOMATION COMPANY ™

Challenges for the real-world

1. Challenge n°1 : Integration

Data mining is never THE solution : it only is a – small – part of it

- In the real-world data mining needs to be integrated into a global system
- Data mining needs to take inputs from/generate results to rest-of-the-world
- Key words : openness, standards

2. Challenge n°2 : Productivity

Data mining must bring value

- Exploit all data available & Produce actionable results
- At lowest possible cost
- Be simple to use by non experts
- Key words : Return On Investment

3. Challenge n° 3 : Scalability

Data mining must hold data volumes & number of models

- Handle LARGE data sets
- Produce AS MANY models as needed
- Key words : time to produce a model as function of data set (width, depth)

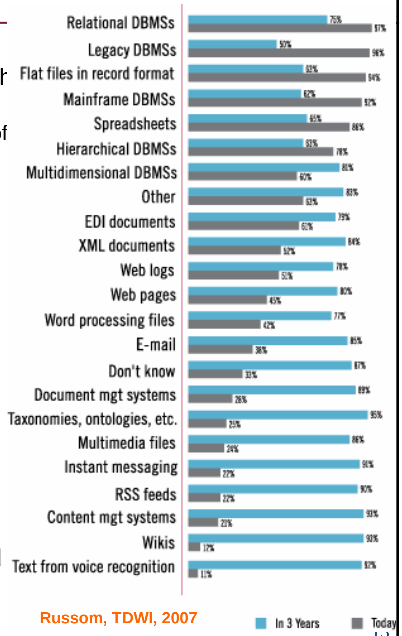
4. Challenge n°4 : Automatisation

Data mining should do all of the above (almost) automatically

- Produce models, detect problems, retrain …
- Key words : automatisation, control

Challenges for the real-world

- **Scalability**
  - Data volume is characterized by (width, deptł
  - Modeling has 2 phases : build & apply
    - How does time scale with volume ? Number of models ?
  - Is real-time possible ?
    - At apply
    - Integration, Productivity, Scalability & Automatisation
- **Productivity**
  - About 40% of modeling time is spent in data preparation, can this be cut ?
  - Can all data be used ?
    - Volumes ?
    - Structured / Unstructured ?
- **Automatisation**
  - Can modeling be done by
    - A machine ?
    - Non-experts ?
  - Can a model be « turned on » and controlled by a machine ?

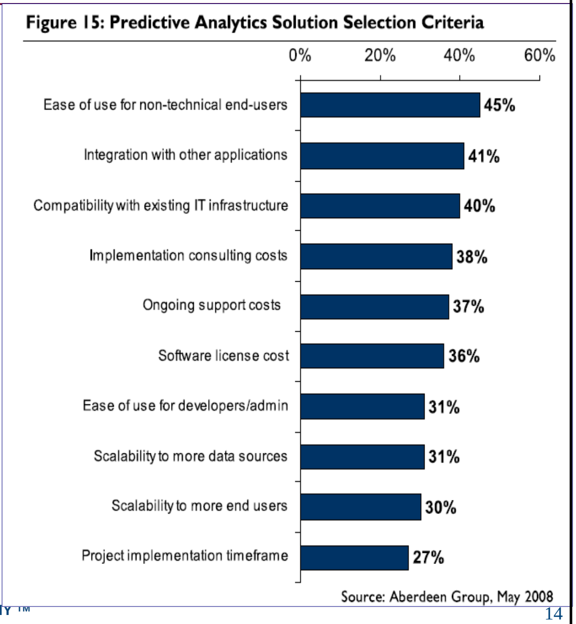


Russom, TDWI, 2007

THE DATA MINING AUTOMATION COMPANY ™

Challenges for the real-world

The ability of Data Mining tools to satisfy the real-world challenges is critical for the wide deployment of data mining applications

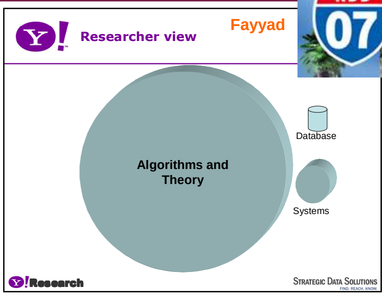


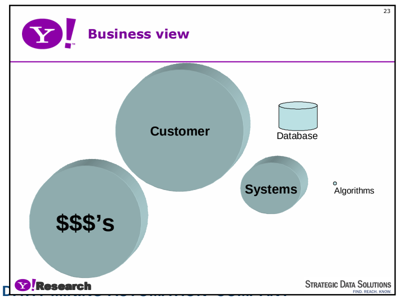
THE DATA MINING AUTOMATION COMPANY ™

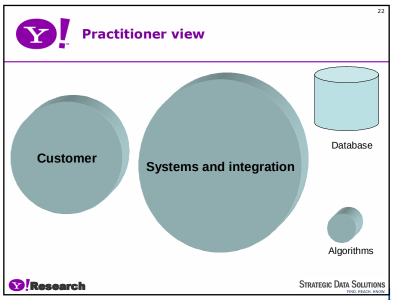
What are the issues in the real-world ?

Algorithms & Theory is NOT central said Fayyad at KDD'07

- **Yes**
  - In real-world, central issue is $
- **But**
  - Only strong algorithms & theory can bring the $
  - Iff they are up to the issues in the real-world







Agenda

- **Which world is this ?**
- **Data Mining in the real world**
- **Some examples**

A large financial institution

The Dilemma

- Efficiency goals
- Limits on resources
- New data sources
- Time-to-market opportunities
- Skill specific dependencies
- Need for competitive advantage
- Unique opportunity to mitigate macroeconomic risks including mortgage and housing troubles

Example n° 1

Extreme Granularity Data

- Competitive advantage through enhanced analytics and unique data sources
- Numerous sources of granular data (transaction data, payment data, call data, etc.)
- Granularity and detail creates value if you can aggregate intelligently and extract knowledge
- Number of attributes grows exponentially as you consider time series, interactions, and transformations

Example n° 1

The Challenge to KXEN

- **Aggregation yields tens of thousands of variables about customers**
- **How do we select the 10 best variables for predicting credit risk, such as, "Will the customer be delinquent in payment 12 months from now?"**
- **How do we develop and deploy models quickly and efficiently?**
- **How can we enable business analysts not statisticians to build predictive models?**
- **How do I regain flexibility as the business owner while ensuring production quality?**

Example n° 1

Possible usage for enhancing variable selection

- **Approach 1: use the same variables we used last year (common in resource constrained environment)**
- **Approach 2: based on experience and expertise, select the 500 variables that are most likely to be useful. Then use statistics to pick the 10 to 20 subset that is best (common in sophisticated analytic shops with heavy analyst presence)**
- **Approach 3: use all the variables and let the data tell you which are useful (rare where attributes >1000)**
- **KXEN POC focused on enabling approach 3 as new modeling process standard**

Example n° 1

The KXEN Fit – Large telco operator

Rapid development and deployment utilizing large volumes of data

- **Data Enhancements**
  - SNA dataset added ~1000 extra variables to our existing analytical datasets (external data plus additional info aggregated from SNA data)
  - Analytical datasets average over 2000 attributes
- **Short time lines**
  - Model development/comparative analysis to pilot execution at times was done over 1-2 week periods
  - KXEN allowed for quick turnaround on model development
  - Built standard models and standard + SNA models for piloting

Example n° 2

THE DATA MINING AUTOMATION COMPANY ™

21

A large american Data provider

Target Source Consumer Database

- **Largest response survey database in North America**
- **Precision data on 2 million Canadian households and 14 million US households**
- **1000 data variables for targeting :**
  - Behavioral, lifestyle, demographics and more
  - Drives solutions for new customer acquisition, modeling, retention / growth and consumer insights

Example n° 3

THE DATA MINING AUTOMATION COMPANY ™

22

A large american Data provider

The Challenge

- **Build 252 models**
- **Score 167 models on 4 million records**
- **Score 85 models on 2 million records**
- **Do it all in a 5 day time period…**

- **And do it with just one analyst…**

Example n° 3

THE DATA MINING AUTOMATION COMPANY ™

23

Why use more data ?

Number of variables	
Sears	900
Large Bank	1 200
Vodafone D2	2 500
Barclays	2 500
Rogers Wireless	5 800
HSBC	8 000
Credit card	16 000

Some companies use lots of data

The goal

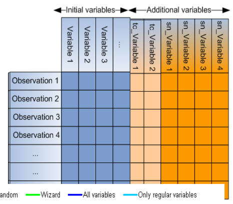
- **Increase the performance of their models**

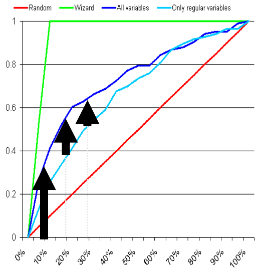
The challenge

- **Produce models with thousands of variables**
- **Exploiting all the available variables**
  - 3 000, 5 000, 10 000 ?
- **Creating new variables, also**
  - Aggregates
  - Behavioral variables
  - Textual variables
  - « Social network » variables …

The results

- More lift, more returns, more €, $





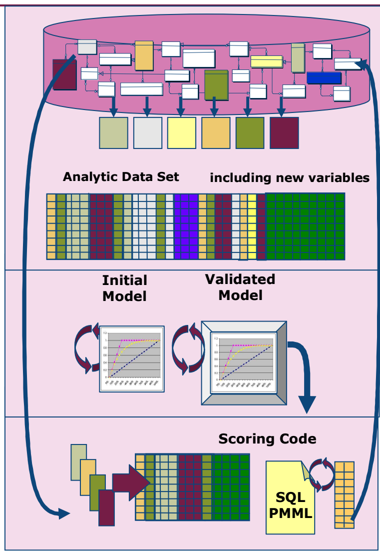
THE DATA MINING AUTOMATION COMPANY ™

24

Exploit all variables

The analysis process

- **Build ADS (Analytic Data Set)**
  - Extract data
  - Transform, aggregate, …
  - Create ADS
- **Build model**
  - Produce initial model
  - Refine, select variables
  - Produce final model
- **Apply model**
  - Extract data
  - Transform, aggregate, …
  - Create ADS
  - Apply model
  - Export results to data base



THE DATA MINING AUTOMATION COMPANY ™

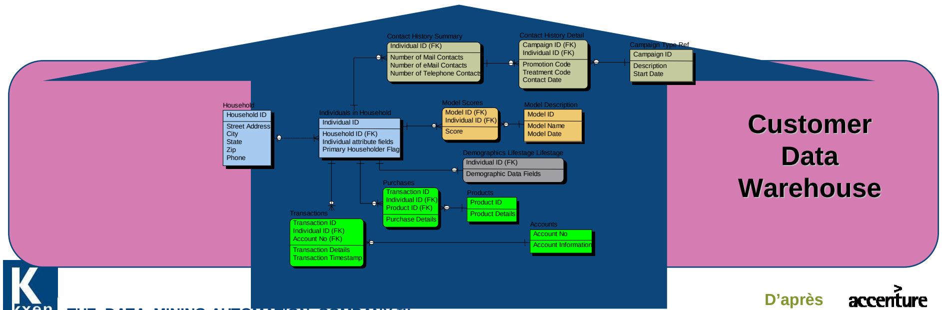
D'après Teradata

Exploit all variables

- **What-if ADS could contain ALL variables**
  - Example : in Teradata ADS is a view. Data are not replicated or moved

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



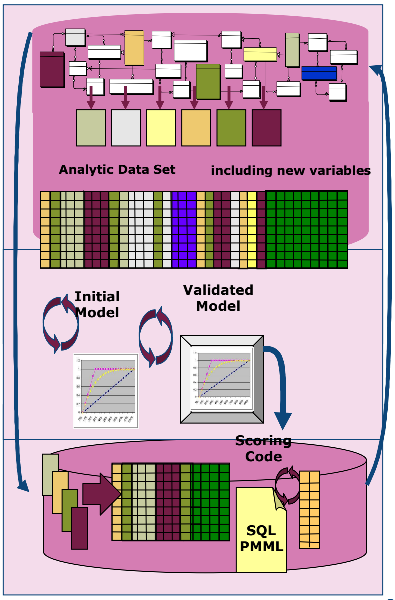
THE DATA MINING AUTOMATION COMPANY ™

D'après accenture

Exploit all variables

The analysis process < 1 week

- **Build ADS (Analytic Data Set)**
  - ~~Extract data~~ 3 days
  - Transform, aggregate, …
  - Create ADS
- **Build model** < 1 day
  - Produce initial model
  - Refine, select variables
  - Produce final model
- **Apply model** < 1 day
  - ~~Extract data~~
  - Transform, aggregate, …
  - Create ADS
  - Apply model
  - Export results to data base



« In-database Mining »

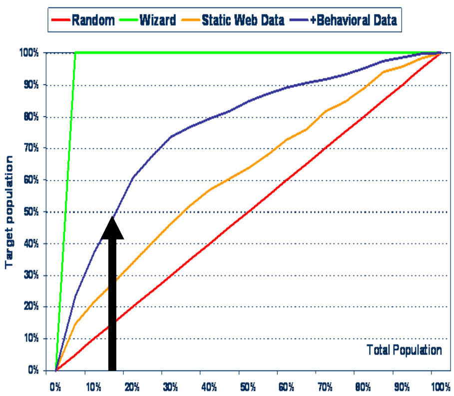
THE DATA MINING AUTOMATION COMPANY ™

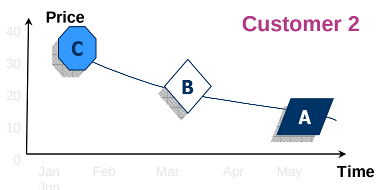
27

Using behavioral data

- **From transactional data produce behavioral data**
  - Transition from transaction A to transaction B
- **Number of variables grows exponentially !**
- **But lift grows !!**

	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

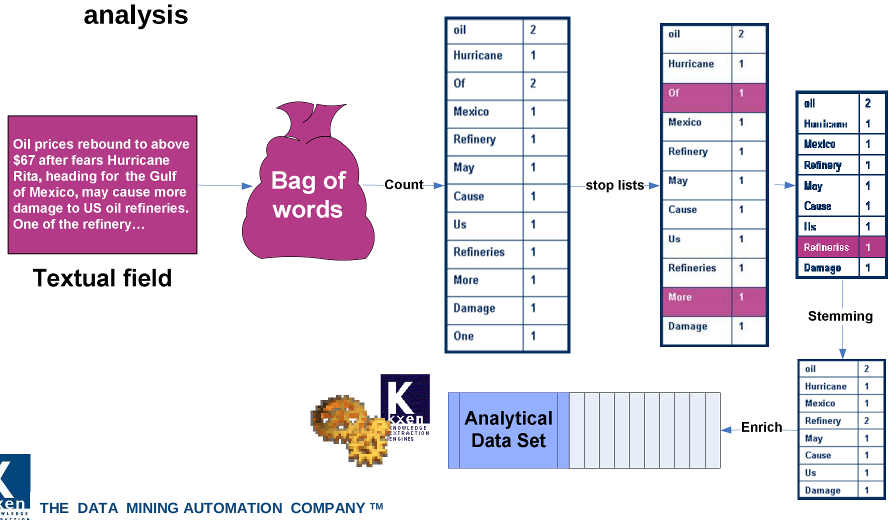




28

Using textual variables

- In many applications (surveys, emails, …) there is a textual field
- These fields can be exploited to enhance results of data mining analysis



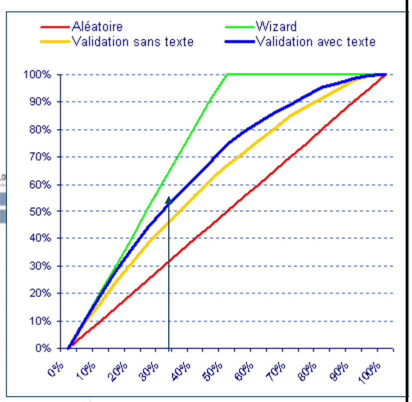
THE DATA MINING AUTOMATION COMPANY ™

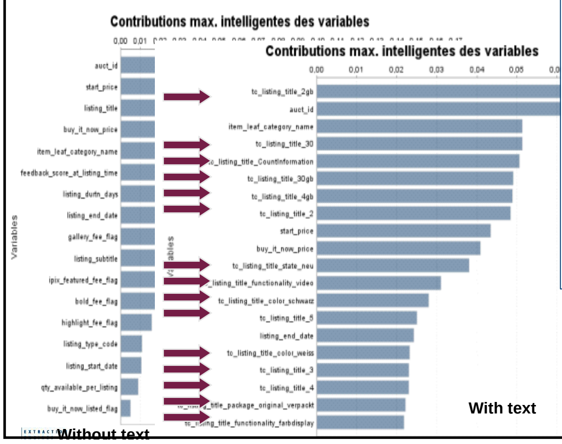
29

Using textual variables – DataMining Cup'06

With 1000 added textual variables

- Computing time : 6 seconds ⇨ 43 seconds
- Most significant variables : textual
- Lift : goes up



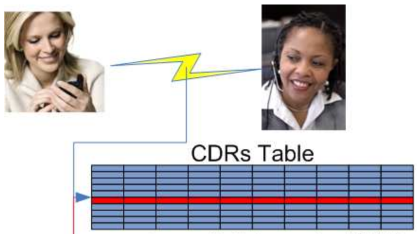


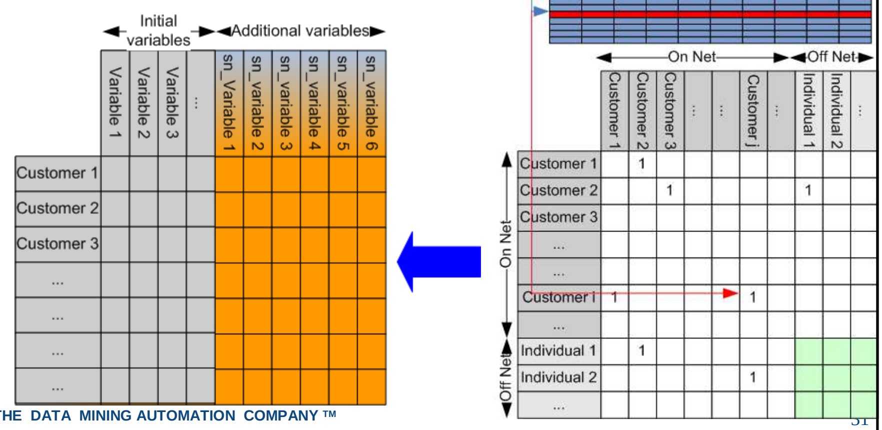
30

Using « social network » variables

An example in telco

- **Build the social network**
- **Extract « social network » variables**
  - A few 10-100 additional variables



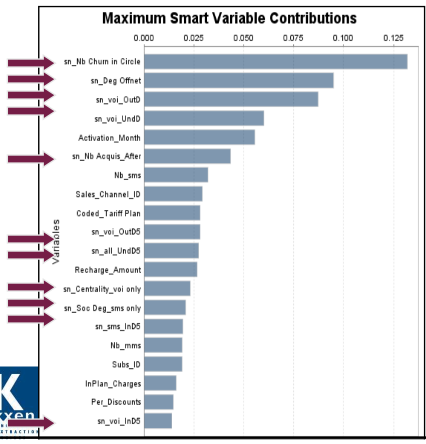


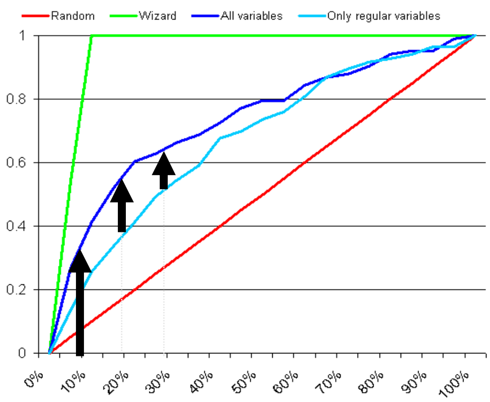
THE DATA MINING AUTOMATION COMPANY ™

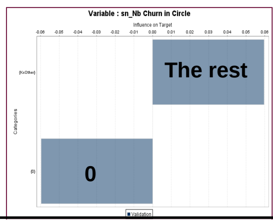
31

Using « social network » variables

- **« social network » variables increase lift**
  - Globally : 40%
  - First decile : 67%
  - Second decile : 47%

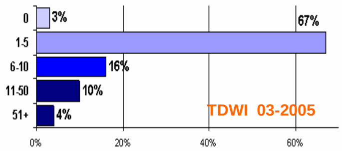






32

Why use more models ?



Some Companies produce lots of models

The goal

- **Increase the performance of their models**

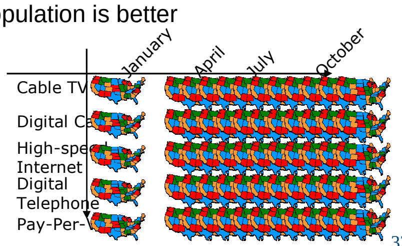
The challenge

- **Produce thousands of models**
- **For every campaign**
- **Refreshed often**
  - Data distribution changes fast (Web)
- **As « fine grain » as possible**
  - Performance on a homogeneous population is better

Number of Models / year	
Vodafone D2	760
Market research	9 600
Cox Comm.	28 800
Real estate	70 000
Lower My Bills	460 000

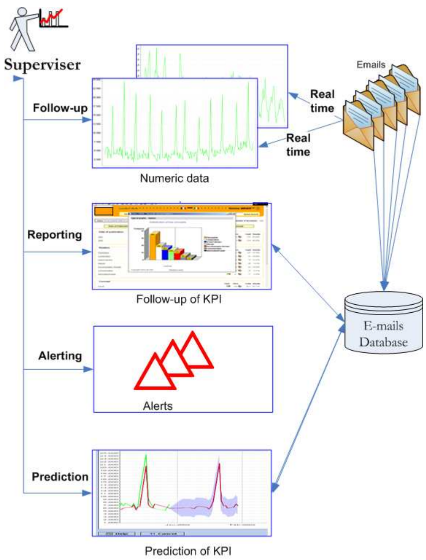
The results

- More lift, more returns, more €, $
- Ex : Response rate + 260%



On-going research projects – Call center supervision

- **Key Performance Indicators**
  - Number of emails received, on-hold, processed
    - Per time period, category, agent …
  - To make sure SLA are met
- **Issues**
  - 100s of KPI
  - Aggregated in hierarchies
- **Predictive models**
  - Alerts
  - Optimize resource (capacity, people)
    - Long term Time Series forecasting on KPI
  - Detect deviations (seasonal effects)
    - Short term Time Series forecasting on KPI
- **"Predictive cubes" for hundreds of indicators**



On-going research projects – Recommendation

Models for recommendation

- **The catalog can have 1 Million products**
- **Hundreds / thousands of models with thousands of attributes**

www.hmv.co.jp

THE DATA MINING AUTOMATION COMPANY ™

Conclusion

Data mining can bring more $

Productivity gains	
Rogers Wireless	7x
Vodafone D2	10x
Sears	8x
Belgacom	12x

But it needs to satisfy constraints

- **Integration**
- **Productivity**
- **Scalability**
- **Automatization**

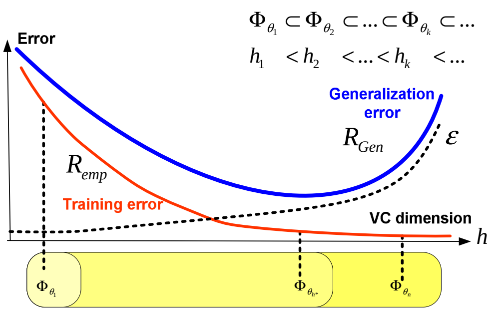
Manipulating MORE data and producing MORE models requires

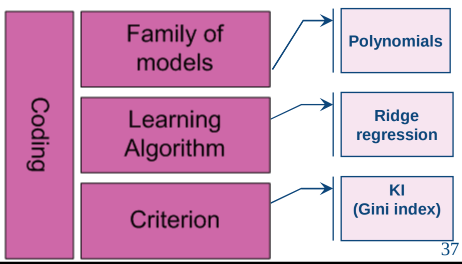
- **Automatized data manipulation & coding**
- **Simple & robust algorithms**
- **Software modules open & in line with market standards**

THE DATA MINING AUTOMATION COMPANY ™

How does KXEN do it

- **KXEN was designed for data miningindustrial applications**
- **KXEN is based upon Vapnik's SRM – Structural Risk Minimization**
  - Strategy to control the trade-off **accuracy / robustness**
- **KXEN produces**
  - An automatic encoding
    - Non linear
  - Then regression / classification
    - Polynomial
- **Which allows**
  - Integration
  - **Productivity**
  - **Scalability**
  - Automatization



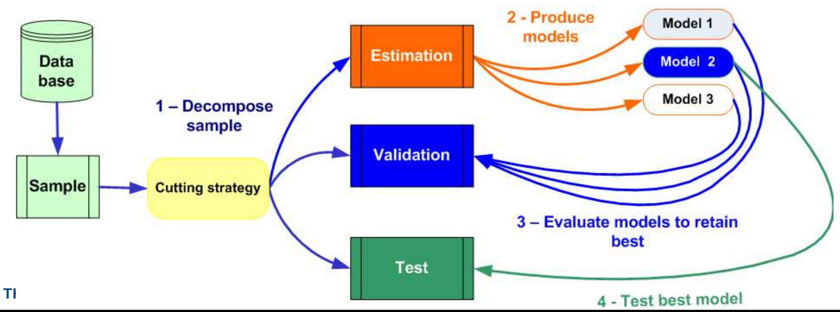


THE DATA MINING AUTOMATION COMPANY ™

37

How does KXEN do it

- **In practice, for one final model, KXEN builds many models (SRM)**
  - Depending upon variables complexity, encoding requires 10 - 30 models
  - Then about 100 models (for regression)
- **KXEN uses « data streams » techniques**
- **There is no data duplication**
  - A few sweeps are necessary
- **Time to build a model**
  - About linear in depth & width



TI

38

Questions ?

- **Location**
  - Marriott Paris Rive Gauche Hotel & Conference Center
    17 Boulevard St Jacques - 75014 Paris, France
- **Dates**
  - June 28 - July 1, 2009
- **Key Submission Dates**
  - Due January 19, 2009 — Workshop Proposals
  - Due February 2, 2009 — Paper Abstracts
  - Due February 6, 2009 — Research/Industrial Track Papers
  - Due February 23, 2009 — Tutorials/Panel Proposals

http://www.kdd.org/kdd2009/index.html

KDD-09 PARIS • June 28th - July 1st 2009
The 15th ACM SIGKDD Conference On Knowledge Discovery and Data Mining

For the first time, in 2009, KDD will leave North America for Europe

KDD'09 will be held in PARIS, France

THE DATA MINING AUTOMATION COMPANY ™ 41

KDD-09 PARIS • June 28th - July 1st 2009
The 15th ACM SIGKDD Conference On Knowledge Discovery and Data Mining

Mark the dates

Conference
- **June 28th – July 1st, 2009**

Key Submission Dates

January 19, 2009	Workshop Proposals
February 2, 2009	Paper Abstracts
February 6, 2009	Research/Industrial Track Papers
February 23, 2009	Tutorials/Panel Proposals
April 10, 2009	Notification
April 27, 2009	Camera Ready

Visit the Conference site

http://www.kdd.org/kdd2009/

THE DATA MINING AUTOMATION COMPANY ™ 42

Contact us

General Chair generalchair@kdd2009.com

- John Elder (Elder Research, Inc.)
- Francoise Fogelman Soulié (KXEN, Inc.)

Program Co-Chair pcchairs@kdd2009.com

- Peter Flach (University of Bristol)
- Mohammad Zaki (Rensselaer Polytechnic Institute

I personnally count on you for sending lots of good papers and show a very strong european attendance at KDD'09

See you there

THE DATA MINING AUTOMATION COMPANY ™ 43