

KDD Challenge 2009

Orange Labs R&D

Vincent Lemaire, Research & Development
03/19/2009, presentation to reading group

<http://perso.rd.francetelecom.fr/lemaire/>

contents

Oranges Labs

CRM at Orange

Problems presentation

Transformation to binary classification problem ?

Data Table size (variables x instances)

Consequences : The proposed challenge

Oranges Labs



a worldwide operator

- One of the **main** telecommunications operators in the world
- Providing services to more than **170 million customers** over five continents
- Including **120 million** under the Orange brand

about Orange and the France Telecom Group



our main activities

on the move

- N°3 in Europe for mobile services with almost **110 million customers**
- 13 million mobile broadband customers with access to the **Orange world portal**



at home

- European leader in broadband Internet (ADSL) with almost **12 million customers**
- 6.1 million **Liveboxes**, the key to high-speed services
- European leader in ADSL television with more than **1.2 million customers**

at work

- with Orange Business Services, the group is one of the **world leaders** supplying telecommunications to more than **3,750 multinationales**

about Orange and the France Telecom Group

Orange Labs: available across the world for all markets





Open positions

- 2 Open Postdoctoral Positions
- See the two subjects (text in French):
 - http://perso.rd.francetelecom.fr/lemaire/Sujet_PostDoc_AssociationRules.pdf
 - http://perso.rd.francetelecom.fr/lemaire/SujetPostDoc2008_2009_FM.pdf
- The English-speaking candidacies are accepted

contents

Oranges Labs

CRM at Orange

Problems presentation

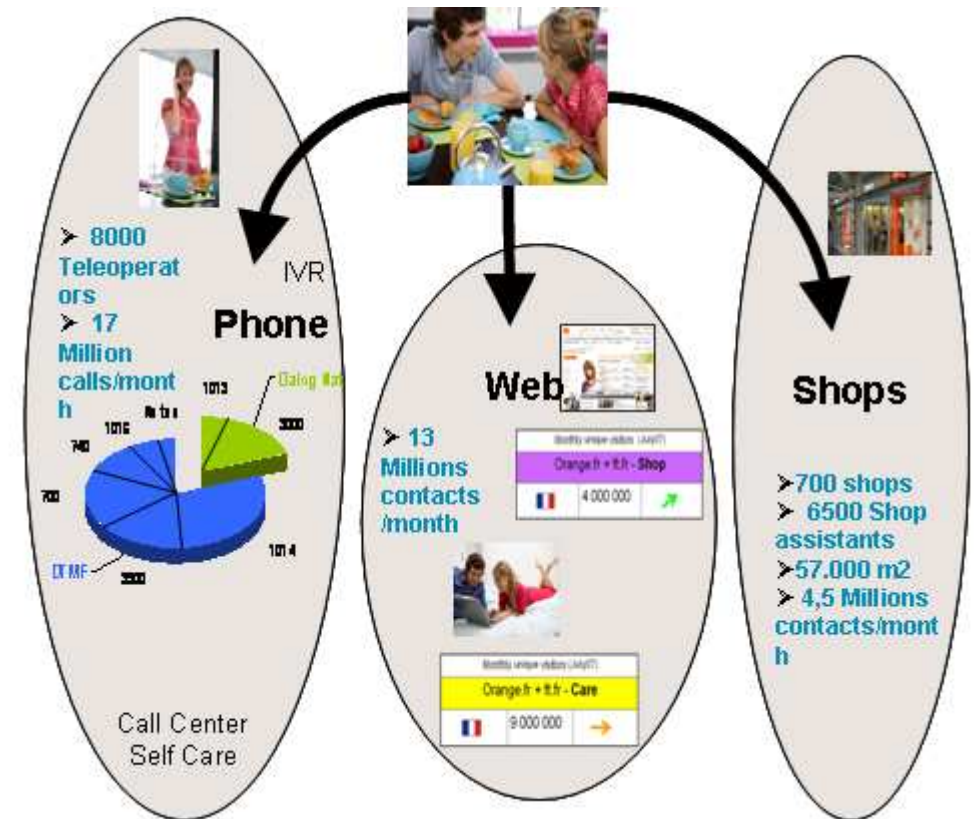
Transformation to binary classification problem ?

Data Table size (variables x instances)

Consequences : The proposed challenge

Interfaces and tools for customer relationship: context

- Customer demand for interaction anywhere, anytime, on any product or service is a strategic issue
- Improve customer experience: leverage each interaction whatever the channel to identify the customer and provide him a **differentiated treatment**.
- **Costs savings**: Maximize customer value and generate cost savings thanks to **increased automation** and better distribution of automatic/non automatic interactions (vocal, web, ...) for sales, after sales, and support.
- Increase revenue: Take advantage of our own experience for developing and selling our CRM offers (Vocal, datamining, ...) to business customers



Customer Relationship (eg in France)

Interfaces and tools for customer relationship: context

- Strong objectives:
 - Custom experience as a major differentiating factor for Orange
 - Commercial costs optimization
- Customer relationship is changing fast due to
 - The evolution of customer usage
 - New technologies
 - Orange entering new territories
- RD focuses on
 - tools for increasing automation and personalization
 - adapted and friendly interfaces
 - customer behavior

contents

Oranges Labs

CRM at Orange

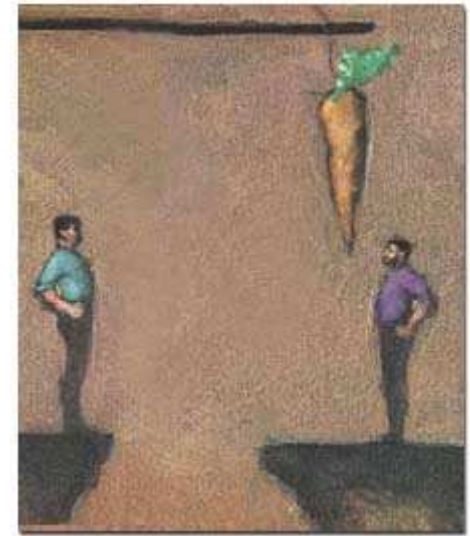
Problems presentation

Transformation to binary classification problem ?

Data Table size (variables x instances)

Consequences : The proposed challenge

Churn

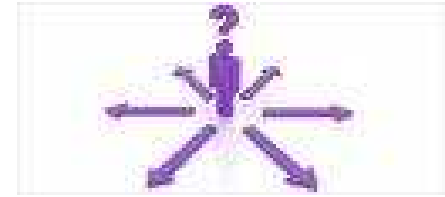


No unique definition - generally, term churn refers to all types of customer attrition whether voluntary or involuntary

In this study:

- moment of churn is the moment when client cancels (“closes”) his product or service in our company
- churner is client having A product or service at time t_n and having no product at time t_{n+1}
- If client still holds the product at time t_{n+1} - non-churner

Cost of Churn



The churn has high cost, to conquer a customer is more expensive than to try to keep customers

Variables:

€X revenue

€Y per customer acquisition cost

\$Z annual revenue/customer

W% per year churn

The cost of churn:

- Assumption : annual income €X million/year
- Only a 1% reduction in churn saves €x million dollars from dropping off the bottom line
- The €x million could be used to recruit new subscribers
- Takes ~ V months to recover the customer acquisition cost

Appetency

In our context, the appetency is the propensity to buy a service or a product.



contents

Oranges Labs

CRM at Orange

Problems presentation

Transformation to binary classification problem ?

Data Table size (variables x instances)

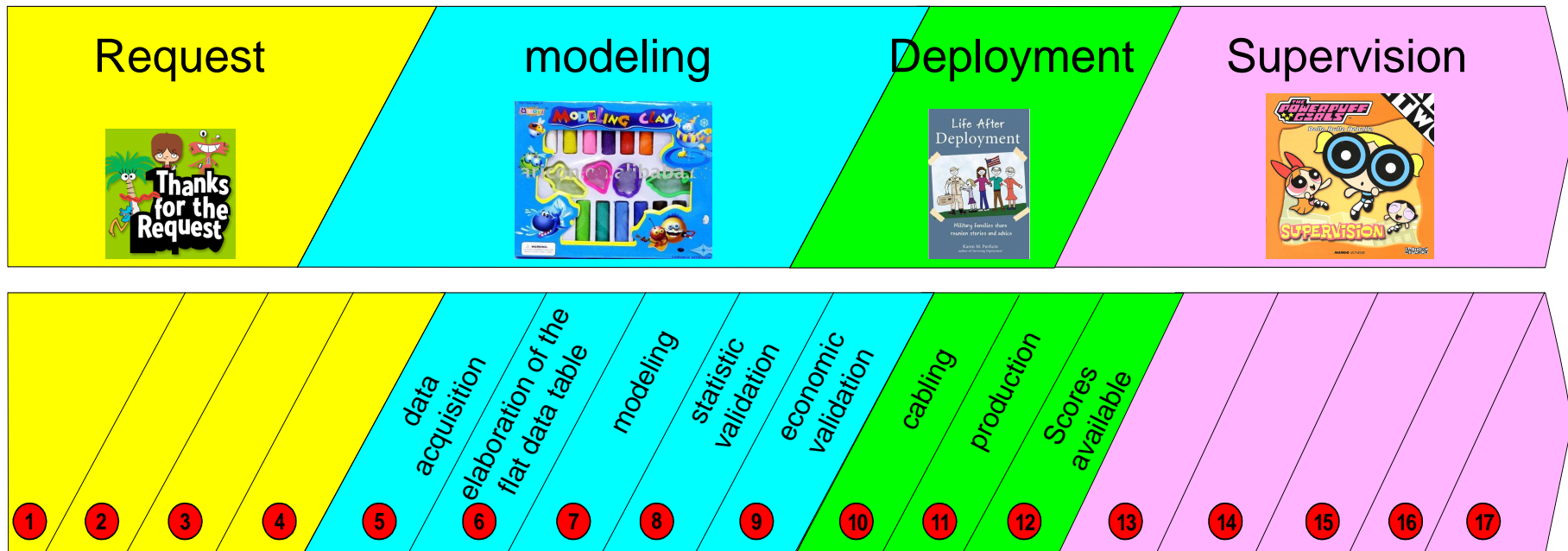
Consequences : The proposed challenge

Score ?

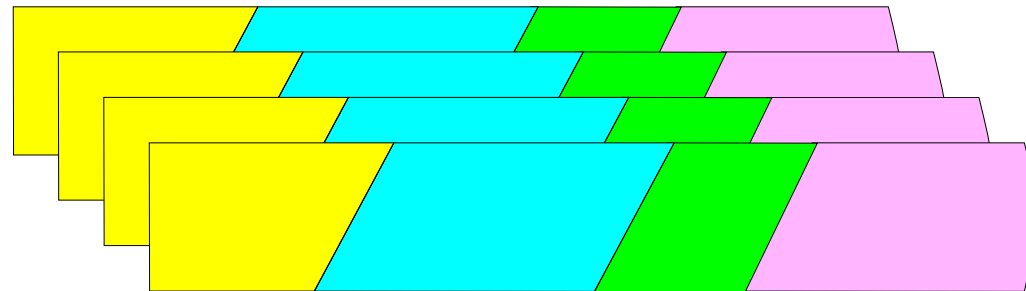
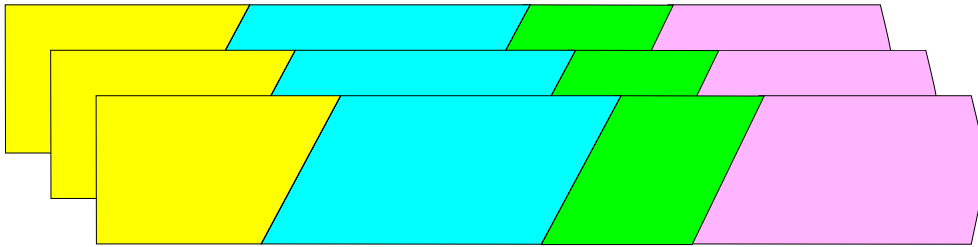
Scoring : Technique of hierarchisation of the users which makes it possible to evaluate by a note the probability that a user answers a request or belongs to the required target within the framework of a direct marketing campaign.

The score is obtained using the quantitative information and qualitative available on the user (given socio-demo, behavior of purchase, preceding answers,...). The scoring makes it possible to optimize the results of campaigns by concentrating the marketing actions on the users having the strongest probability of answer.

Life cycle of a score



Time is money !

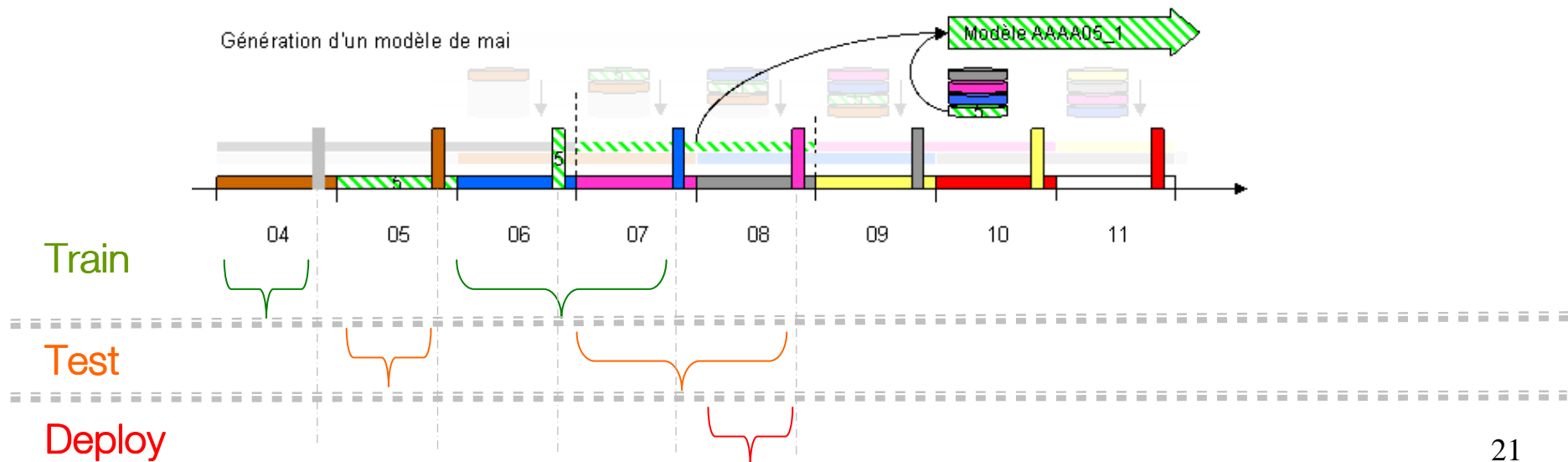


To this day ≈ 60 scores are produced every month

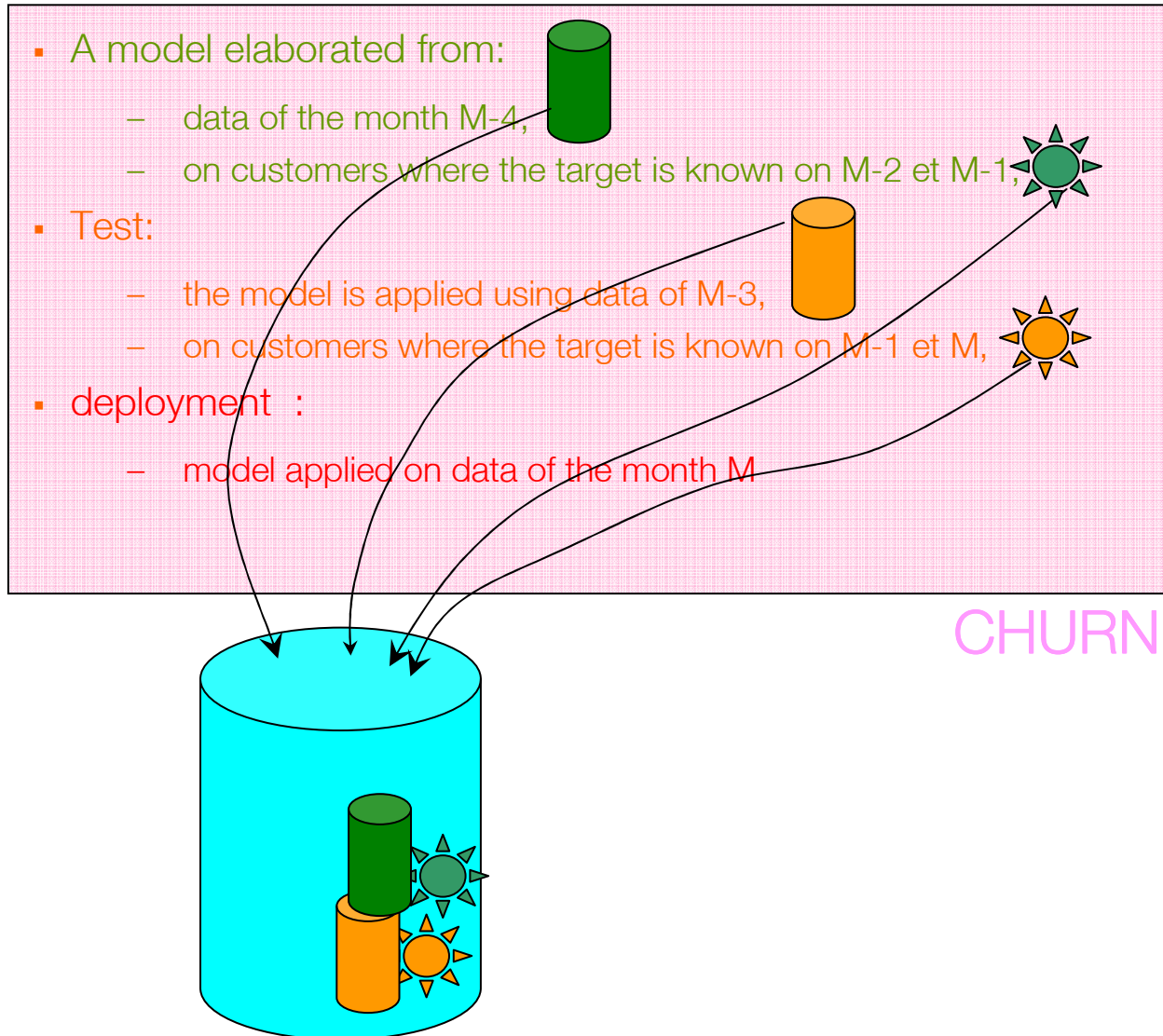
How many scores with a faster process using a faster model ?

Transformation into a binary classification problem (normally)

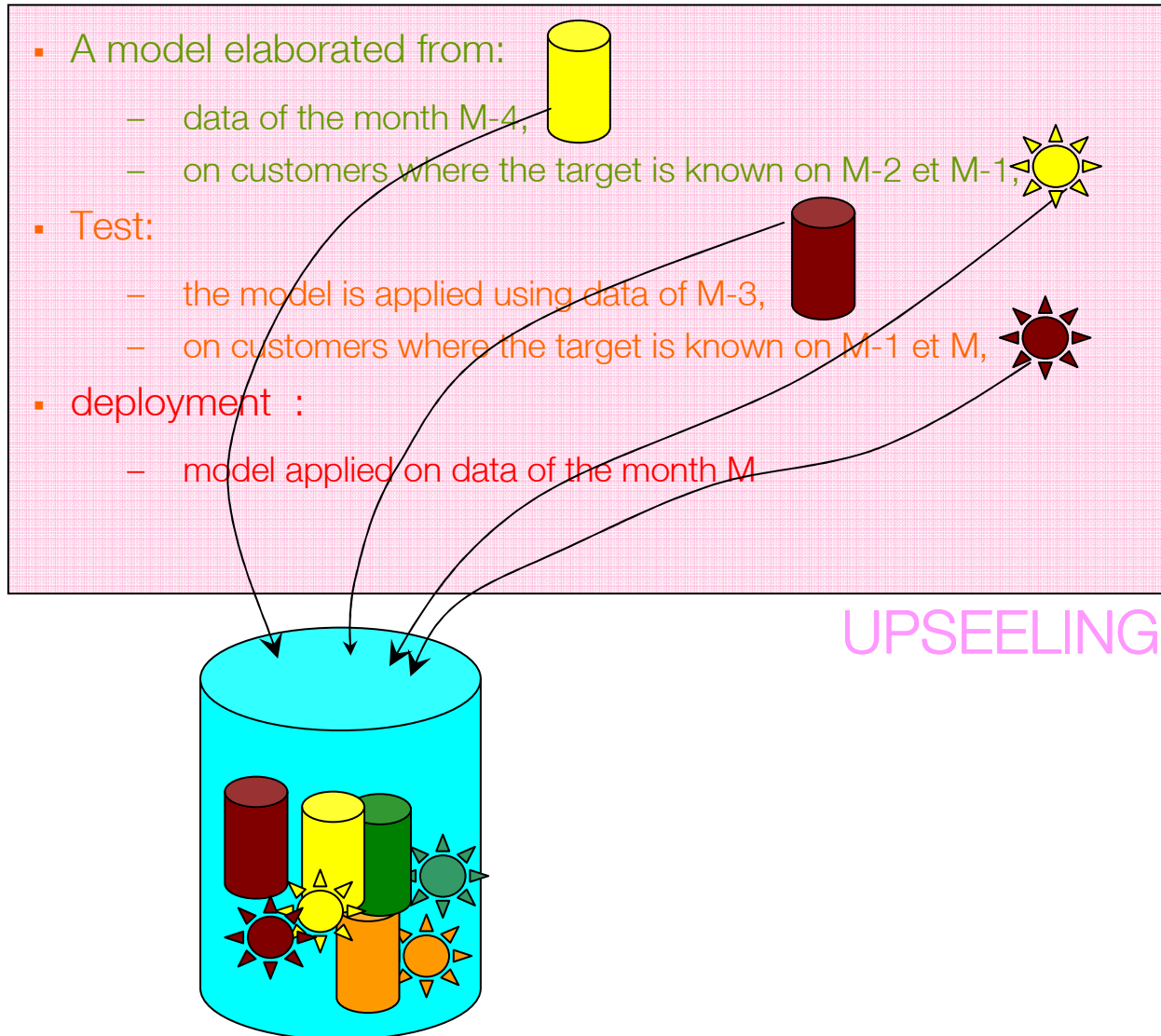
- A model elaborated from:
 - data of the month M-4,
 - on customers where the target is known on M-2 et M-1,
- Test:
 - the model is applied using data of M-3,
 - on customers where the target is known on M-1 et M,
- deployment :
 - model applied on data of the month M



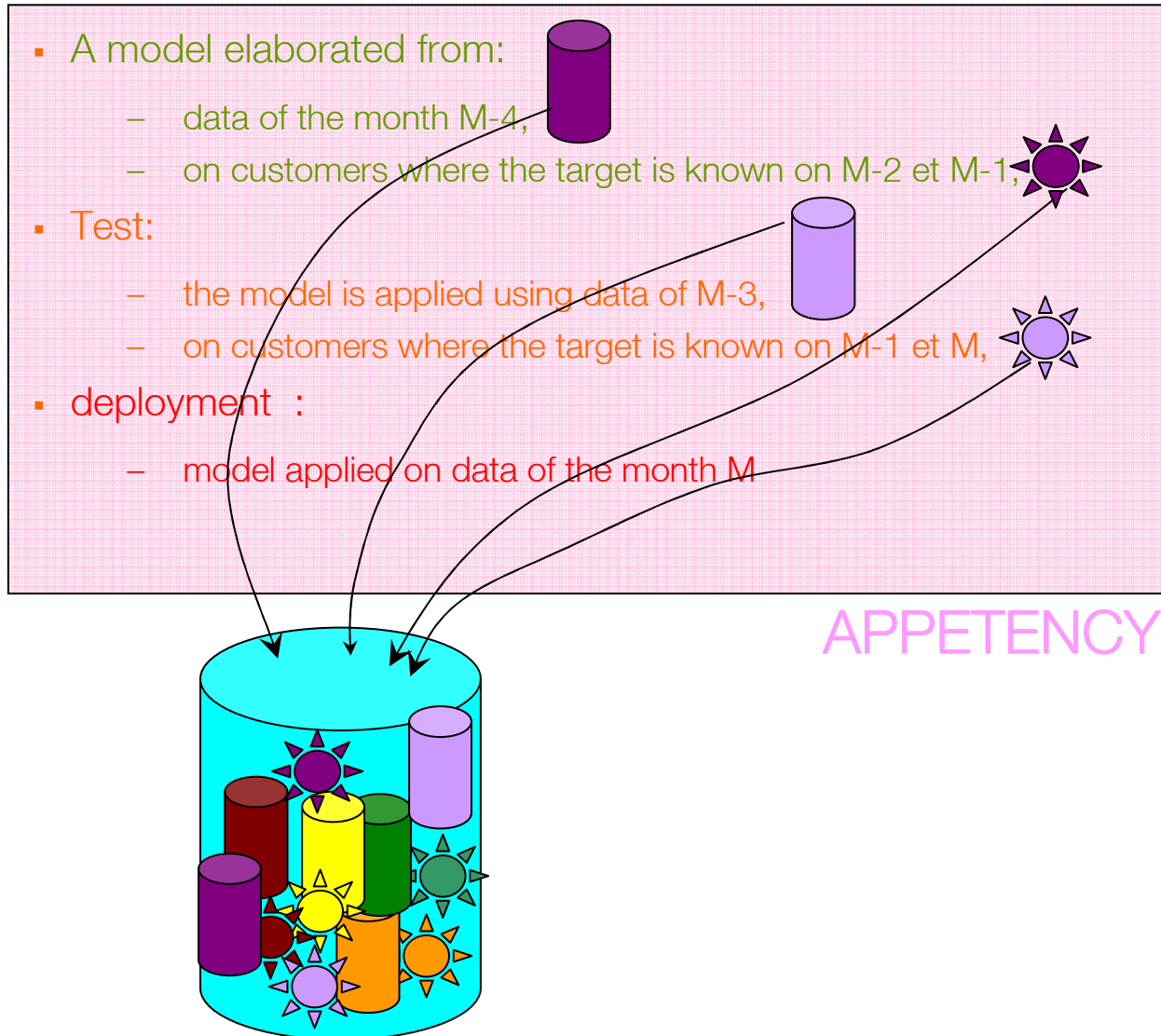
Transformation into a binary classification problem (for the challenge)



Transformation into a binary classification problem (for the challenge)



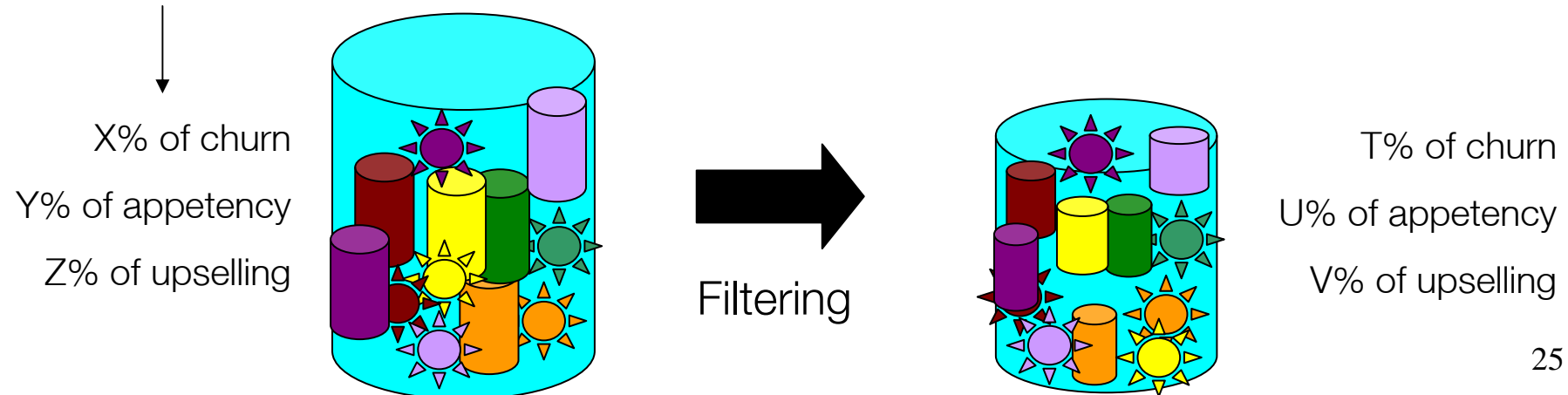
Transformation into a binary classification problem (for the challenge)



Transformation into a binary classification problem (for the challenge)

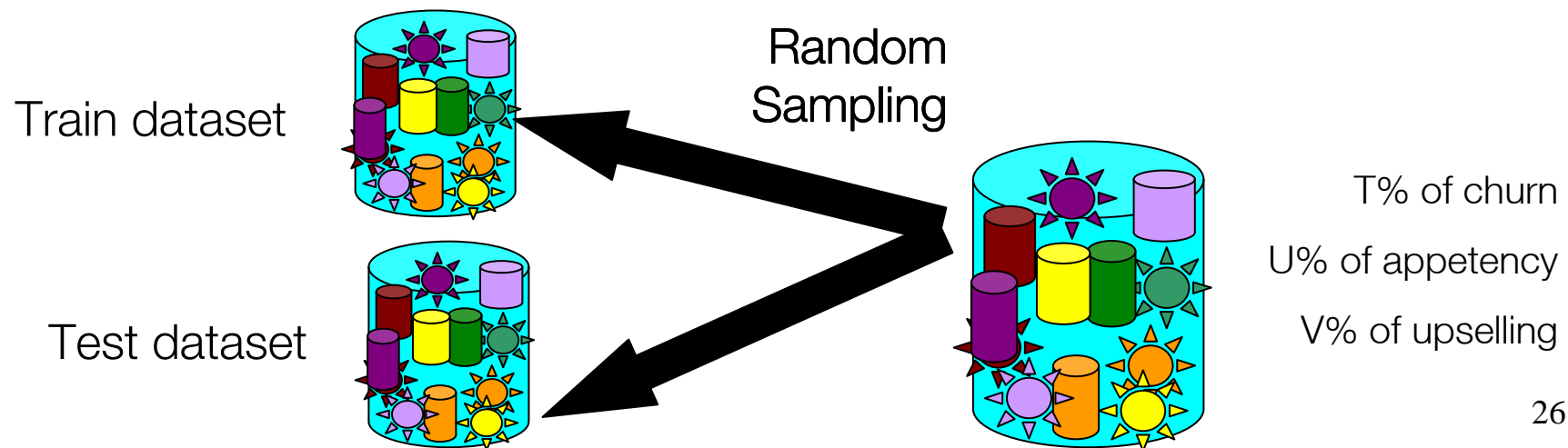
- A model elaborated from:
 - data of the month M-4,
 - on customers where the target is known on M-2 et M-1,
- Test:
 - the model is applied using data of M-3,
 - on customers where the target is known on M-1 et M,
- deployment :
 - model applied on data of the month M

Information not available



Transformation into a binary classification problem (for the challenge)

- A model elaborated from:
 - data of the month M-4,
 - on customers where the target is known on M-2 et M-1,
- Test:
 - the model is applied using data of M-3,
 - on customers where the target is known on M-1 et M,
- deployment :
 - model applied on data of the month M



contents

Oranges Labs

CRM at Orange

Problems presentation

Transformation to binary classification problem ?

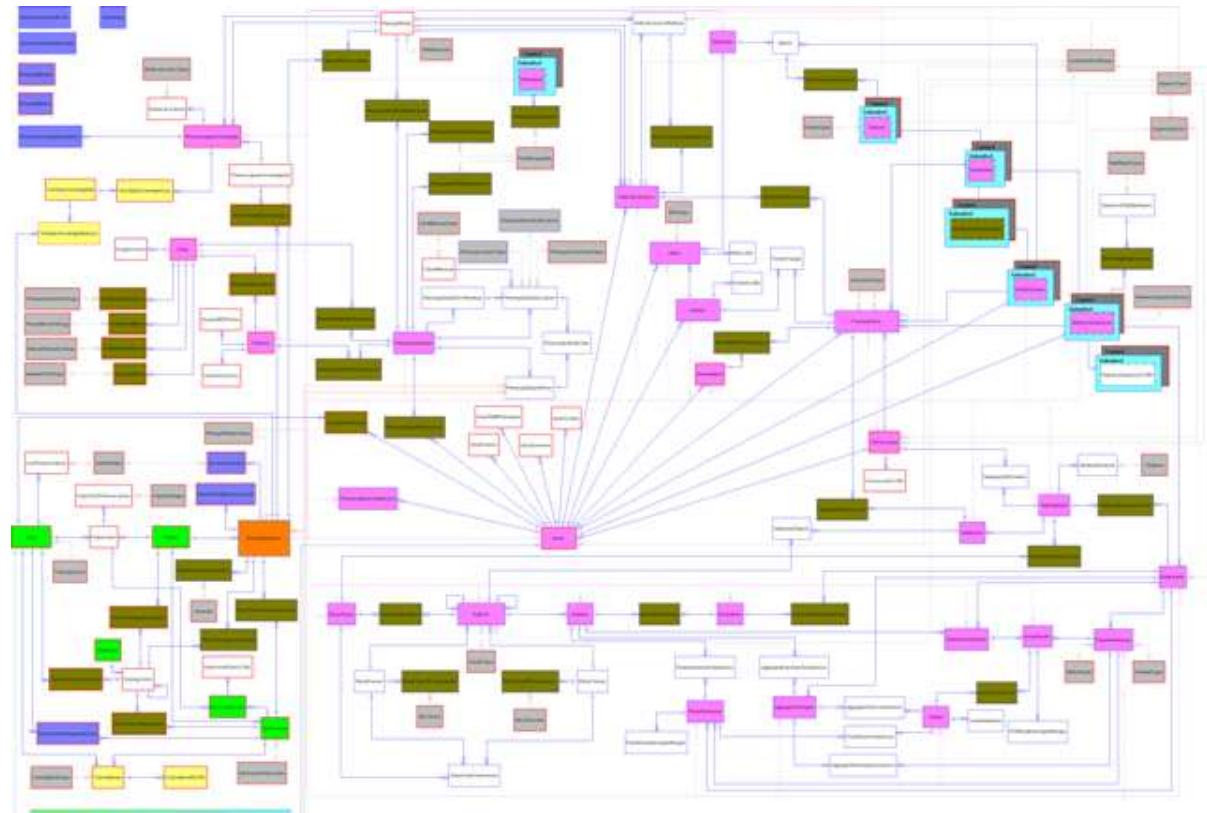
Data

Consequences : The proposed challenge

Data preparation

- Elaboration of the modeling data from data in their native format

- Formatting of a table instances*variables to modeling



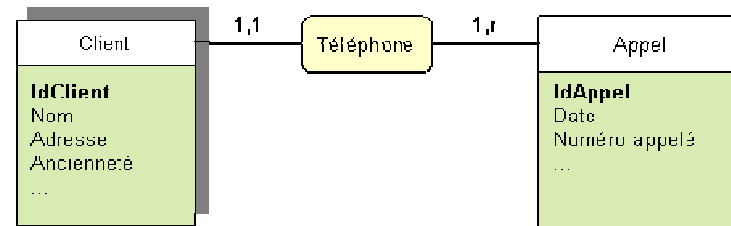
Data preparation

- Elaboration of the modeling data from data in their native format

- Formatting of a table

instances*variables

to modeling



example of the joint of 2 tables

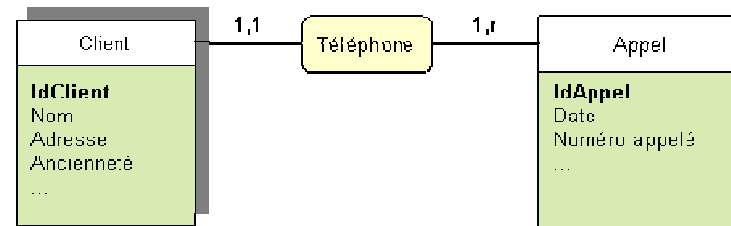
Data preparation

- Elaboration of the modeling data from data in their native format

- Formatting of a table

instances*variables

to modeling



- The column number could be very large (many table and joints are possible)

idClient	Nom	Ancienneté	NbAppels /mois	NbAppels /services	NbAppels / 0-1h	NbAppels / 23-24h	...
32	Dupont	1	23	23	23	23	...
234	Durant	2	25	25	25	25	...
45	Dupond	5	765	765	765	765	...
...



Limitations for data mining

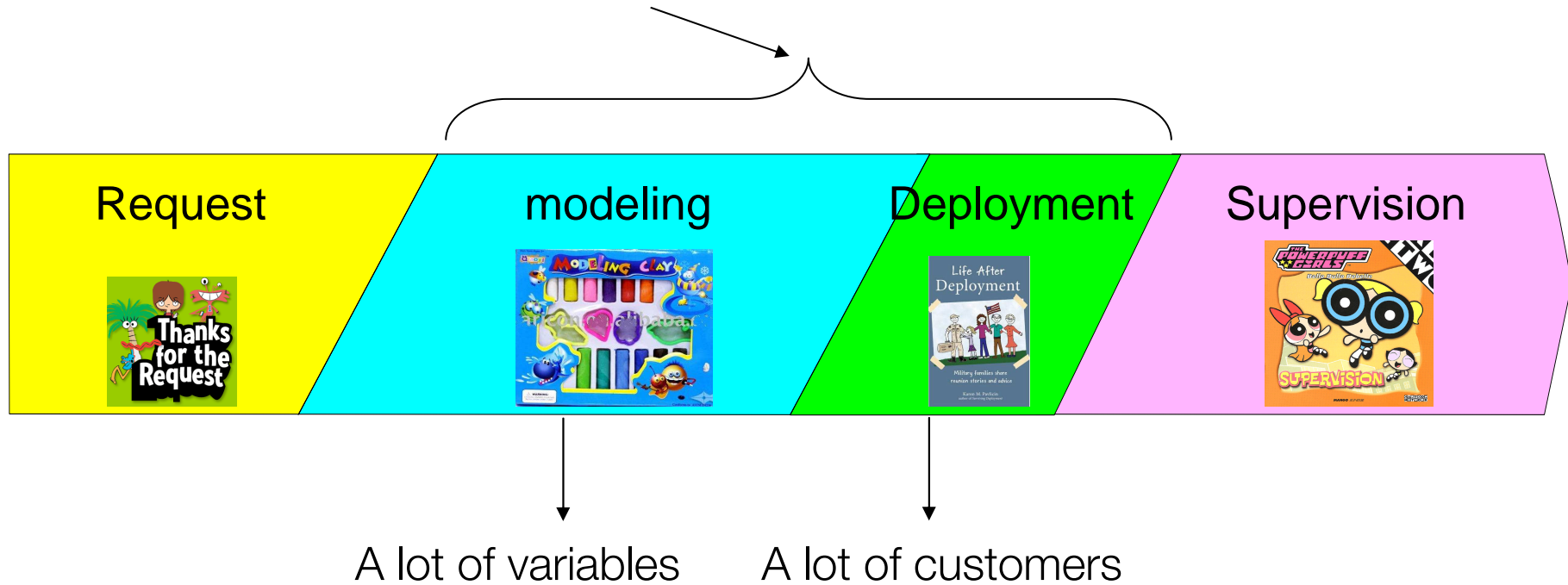
- To analyze the data, it is essential (here) to put them flat.
- It is not possible to know in advance which are the relevant indicators for the study considered (for the moment).
- The number of useful indicators is so potentially very large.

- With current technologies, it is necessary to find a compromise between the performance of the models and their cost of deployment.

BUT

60 scores to produce every month

⇒ ≈ 8 hours to produce a score



contents

Oranges Labs

CRM at Orange

Problems presentation

Transformation to binary classification problem ?

Data Table size (variables x instances)

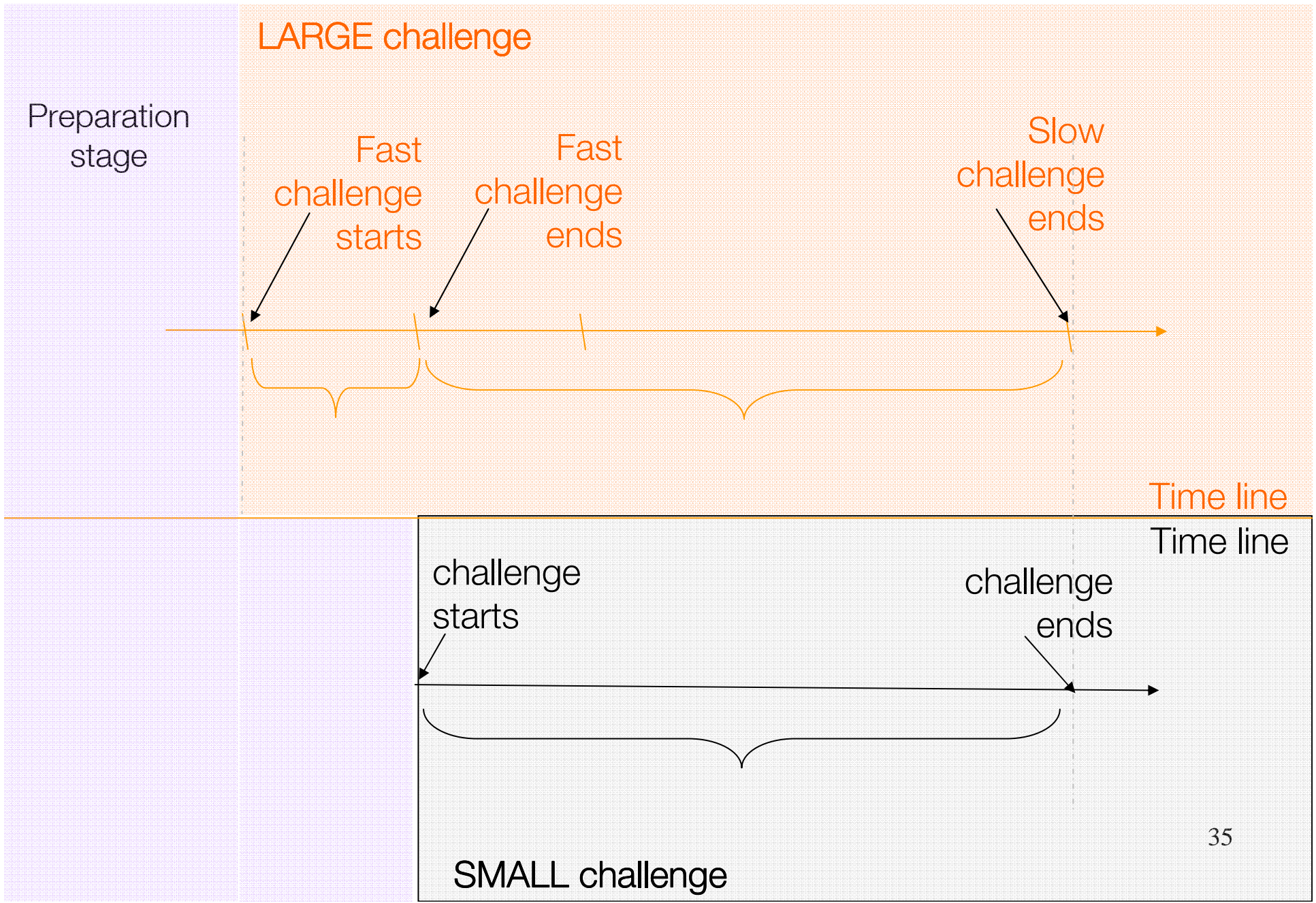
The proposed challenge

The KDD cup 2009

- 10000 Euros in prizes
- Perform at least as well as NB
- Beat the in house system?

- Three targets: churn, appetency, upselling

Challenge protocol: FAST vs. SLOW

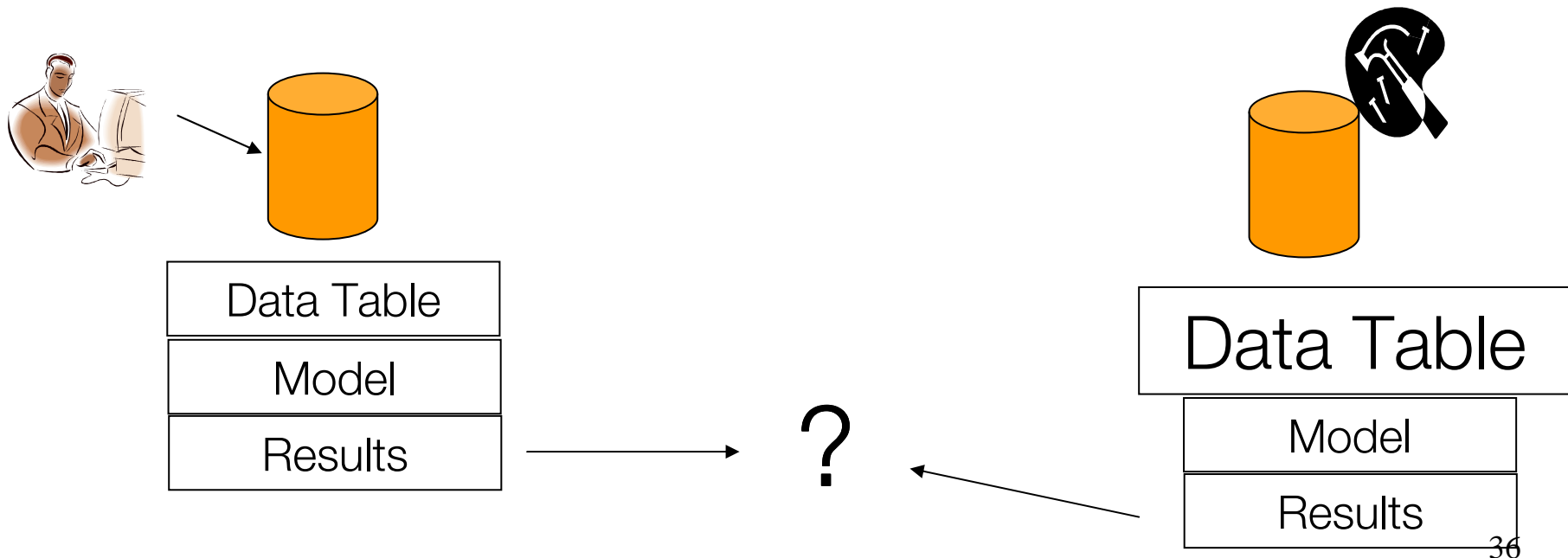


KDD 2009 Challenge

Large versus Small challenge ?

For the small track, we kept a subset of 230 variables, currently used by the marketing teams.

The purpose is to see if a large number of variables (automatically elaborated) allow to obtain a better result than using the 230 variables



KDD 2009 Challenge

Large challenge ?

Primary issue: automation

- Automatic data preparation
- Parameter free methods
- Generic methods

We are very interested by 'unique method' without parameters able to deal with the 3 problems ! and which elaborate a model as quickly as possible

We give 5 days to elaborate 3 models : a compromise between our real need and the purpose to have competitors to the fast challenge !

KDD 2009 Challenge

Datasets...

The large dataset archives are available since the onset of the challenge.

The small dataset will be made available at the end of the fast challenge.

Both training and test sets contain **50,000 examples**.

The data are split similarly for the small and large versions, but the samples are ordered differently within the training and within the test sets.

Both small and large datasets have numerical and categorical variables.

For the large dataset, the first **14,740 variables are numerical** and the last **260 are categorical**.

For the small dataset, the first **190 variables are numerical** and the last **40 are categorical**.

Toy target values are available only for practice purpose.

KDD 2009 Challenge

Data format...

The datasets use a format similar as that of the text export format from relational databases:

- One header lines with the variables names
- One line per instance
- Separator tabulation between the values
- There are missing values (consecutive tabulations)

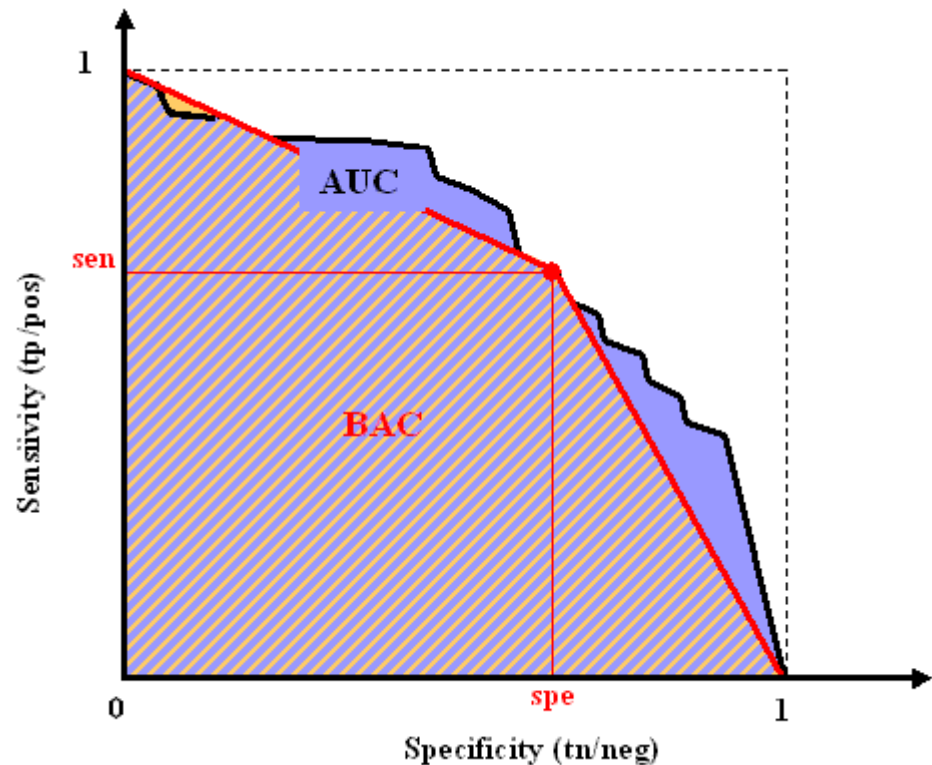
How to submit results?

- Provide a score: any numeric value, which is larger for the positive class. Examples:
 - Binary $\{-1, +1\}$ values indicating class membership.
 - Discriminant values, negative for the negative class and positive for the positive class.
 - A score between 0 and 1 interpretable as the probability of membership of the example to the positive class.
 - A rank, smallest values representing examples classified with highest confidence as members of the negative class.
- Various classification accuracy metrics are obtained by setting a threshold on this score [.../...]

KDD 2009 Challenge

Evaluation ?

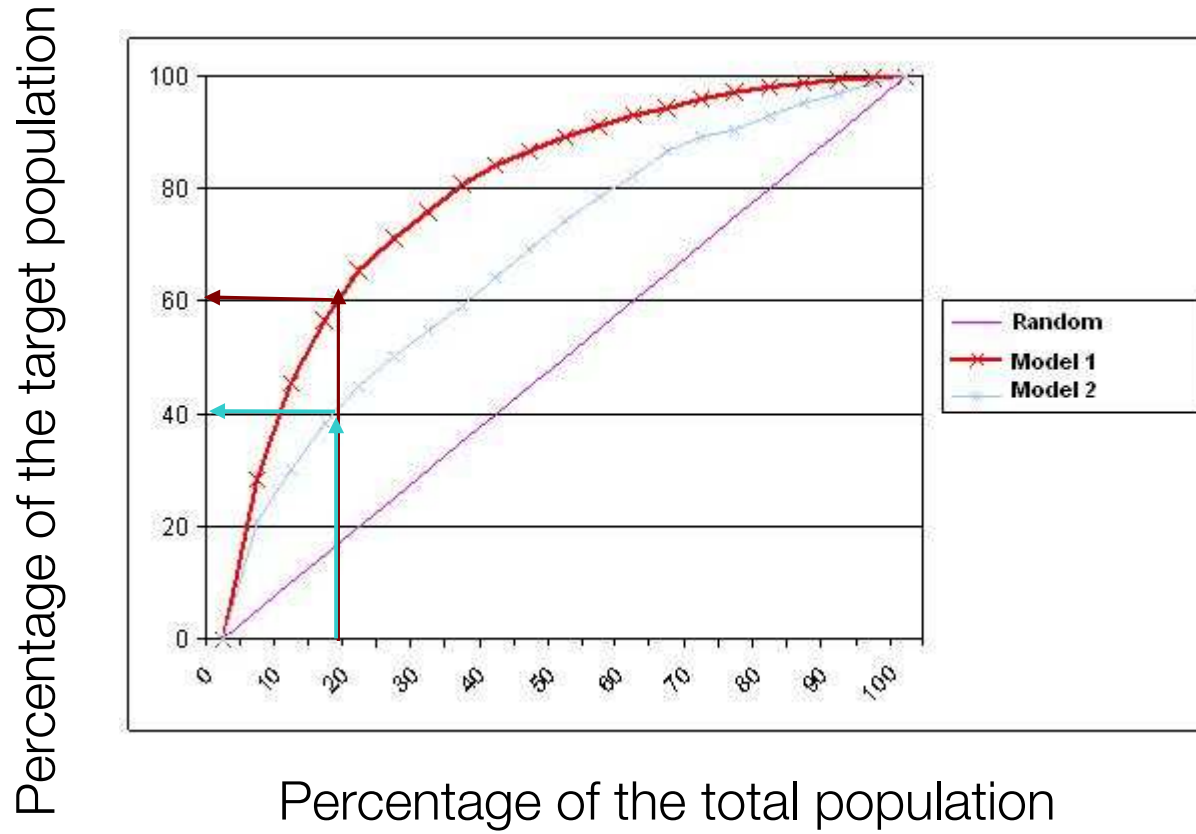
The results will be evaluated with the so-called Area Under Curve (AUC). It corresponds to the area under the curve obtained by plotting sensitivity against specificity by varying a threshold on the prediction values to determine the classification result. The AUC is related to the area under the lift curve and the Gini index used in marketing ($Gini=2 \text{ AUC} - 1$). The AUC is calculated using the trapezoid method.



Feed back using 10% of the test set.

Ranking of competitors using the 90% remaining





The influence of a good classification



KDD 2009 Challenge

a challenge which includes simultaneously key points of last challenges ?

Requirements:

- Fast data preparation and modeling (no parameters ?) 
- A model reliable and accurate 
- Train database with tens to tens of thousands of variables (features selection?) 
- Database which has both categorical and numerical variables 
- Categorical variables have very large number of categories
- Large amount of missing values
- Target variables rare (heavily unbalanced distributions)

KDD 2009 Challenge

a challenge which includes simultaneously key points of last challenges ?

Other interesting points for us not present is the challenge:

- Fast deployment (up to real time classification in network devices?)
- Results understandable (no black box?)
- Test database with tens to millions of instances
- Stationary not always really true (past data to predict future behavior)

KDD 2009 Challenge

Credits

Project team at Orange Labs R&D:



- [Vincent Lemaire](#)
- [Marc Boullé](#)
- Fabrice Clérot
- Raphaël Féraud
- Aurélie Le Cam
- Pascal Gouzien



Beta testing and proceedings editor:

- [Gideon Dror](#)



Web site design:



- [Olivier Guyon](#) (MisterP.net, France), adapted and improved from an earlier design used in previous challenges by [Steve Gunn](#) (University of Southampton, UK).

Coordination (KDD cup co-chairs):



- [Isabelle Guyon](#)
- [David Vogel](#)



KDD 2009 Challenge

Sponsors

We are very grateful to our sponsors who made this project possible:



CLOPINET




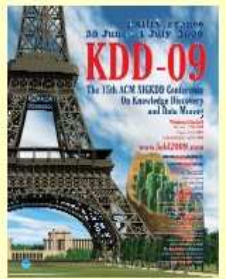
KDD 2009 'Help'

Help - KDD Cup 2009 - Microsoft Internet Explorer

Fichier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris

Adresse <http://www.kddcup-orange.com/help.php> OK Links Google Rechercher Connexion

 **KDD Cup 2009** 

[Index](#) [Data](#) [Instructions](#) [Evaluation](#) [Submit](#) [Results](#) [Prizes](#) [Workshop](#) [Login/Register](#) [Help](#) [Credits](#) [Previous Cups](#)

Frequently Asked Questions

Participation and Registration

What is the goal of the challenge?
The challenge consists of several classification problems. The goal is to make the best possible predictions of a binary target variable from a number of predictive variables.

Can I enter under multiple names?
No, we limit each participant to one final entry, which may contain results on the large dataset only in the fast track and on either or both the small and the large dataset in the slow track. Registering under multiple names would be considered cheating and disqualify you. Your real identity must be known to the organizers. You may hide your identity only to the outside by checking the "Make my profile anonymous" in the registration form.

Can I participate to multiple teams?
No. Each individual is allowed to make only a single final entry into the challenge to compete towards the prizes. During the development period, each team must have a different registered team leader. To be ranked in the challenge and qualify for prizes, each registered participant (individual or team leader) will have to disclose the names of eventual team members, before the final results of the challenge get released. Hence, at the end of the challenge, you will have to choose to which team you want to belong (only one!), before the results are publicly released. After the results are released, no change in team composition will be allowed.

Internet

thank you



Appendix

KDD 2009 Challenge Prizes

There are **10000 Euros** of prizes and travel grants generously donated by Orange, which will be distributed among the cup winners. Only participants having complied with the rules of the challenge and having submitted themselves to eventual post-challenge verifications, as required by the organizers, will be eligible for prizes, or travel awards.

Fast challenge:

- First place: 500 Euros of travel award + 3000 Euros of cash prize + plaque
- Second place: 500 Euros of travel award + 1500 Euros of cash prize + plaque
- Third place: 500 Euros of travel award + award certificate

Slow challenge:

- First place: 500 Euros of travel award + 1500 Euros of cash prize + award certificate
- Second place: 500 Euros of travel award + 1000 Euros of cash prize + award certificate
- Third place: 500 Euros of travel award + award certificate

The travel awards must be used by the awardees to attend the [KDD workshop](#) and will be granted upon delivery of receipts of travel expenses. Prizes and travel awards cannot be cumulated. So if you are among the two first in the fast track, you are eliminated from the slow track ranking (note that the first in slow track wins no more the second in the fast track so it is not unfair to be eliminated from the slow track ranking). If you win a prize or travel award in the slow track and are third in the fast track, you cannot cumulate the travel awards. Any non-attributable or unclaimed travel award or prize will be redistributed as travel grant to other deserving participants.

KDD 2009 Challenge

Other Data description ...

Input data:

The data comes from the marketing information system of France Telecom.

It is first imported into a simplified data warehouse based on a star schema, with the customer at the center of the star, surrounded by secondary tables (telephone calls, products, services...).

This data is then projected on a flat table using our PAC feature construction tool, that produces thousands of indicators.

Instance Selection:

Starting from sample of about X instances, 100000 instances were extracted from the relevant marketing domain, so as to enrich the classes of interest (too rare using a random sampling).

The 100000 are randomly split into 50000 train instances and 50000 test instances, with approximately balanced proportions of each target class (within one standard deviation).

Anonymization

All the variables have been recoded for privacy preservation

Data Mining in France Telecom

- Many domains
 - Marketing
 - Text mining
 - Web mining
 - Traffic classification
 - Sociology
 - Ergonomics
- Many scales
 - Tens to millions of instances
 - Tens to tens of thousands of variables
- Many types of data
 - Numerical
 - Categorical
 - Text
 - Image
 - Relational databases
- Many tasks
 - Data exploration
 - Supervised
 - Unsupervised
- Data constraints
 - Heterogeneous
 - Missing values
 - Multiple classes
 - Heavily unbalanced distributions
- Training requirements
 - Fast data preparation and modeling
- Model requirements
 - Reliable
 - Accurate
 - Parsimonious (few variables)
 - Understandable
- Deployment requirement
 - Fast deployment
 - Up to real time classification in network devices
- Business requirement
 - Return of investment for the whole process