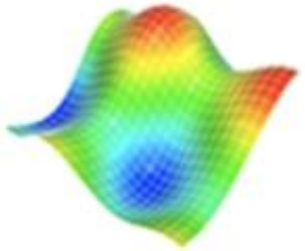
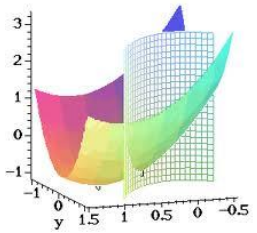
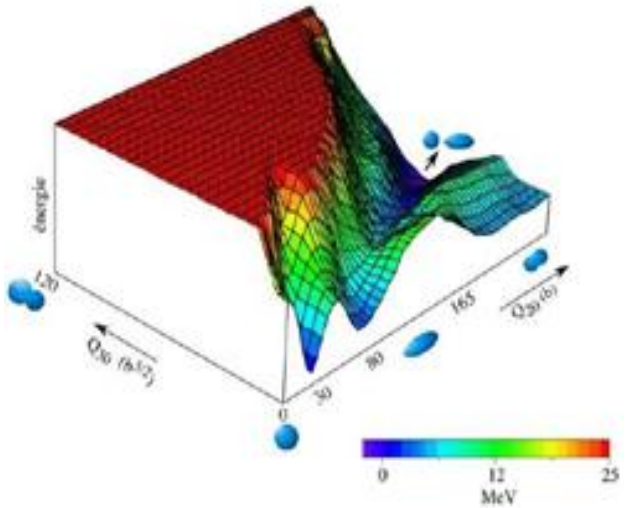
Phénomènes complexes



La méthode est l'art de choisir ce qu'il faut observer ou expérimenter.

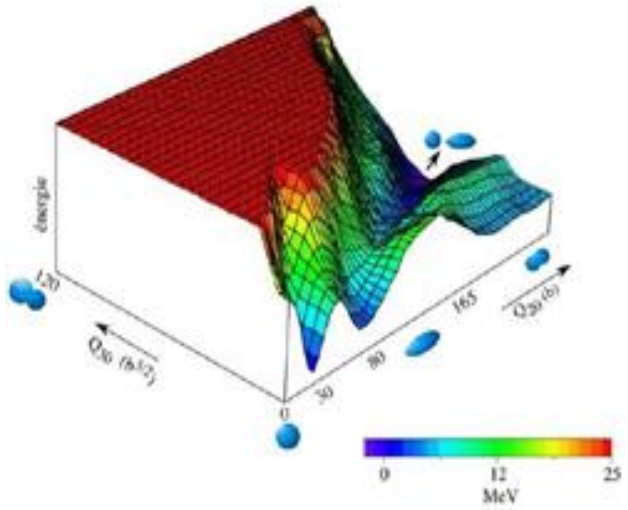


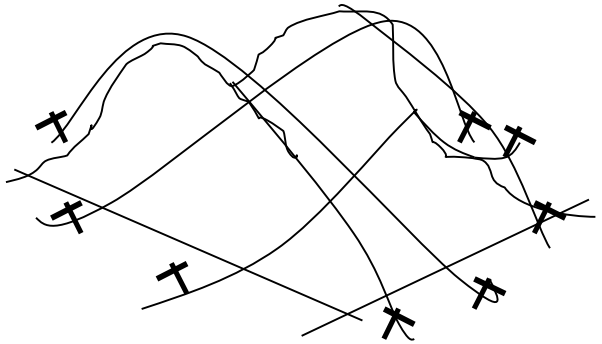


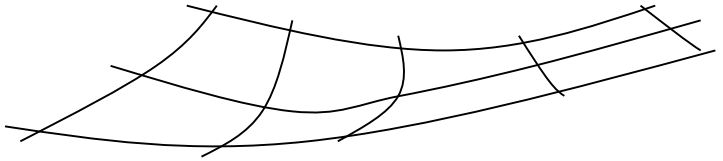
Problématique

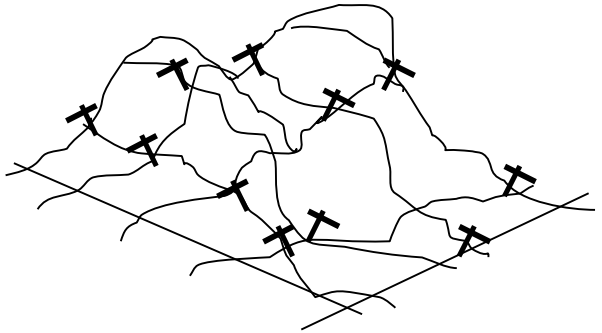
Les plans d'expériences traditionnels consiste de manière générale à choisir des points au bord du domaine expérimental.
Ces types de plans ne permettent pas de capturer les phénomènes symbolisés par le graphique ci-dessous.

Données continu
avec des ruptures



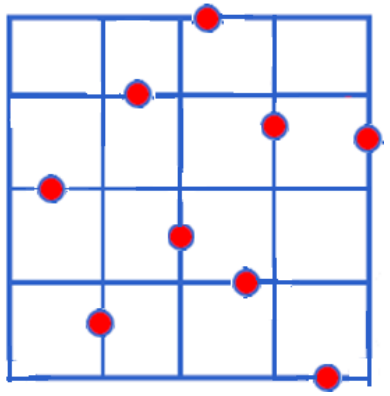






La seule solution est de réaliser des expériences (points) au centre (à l'intérieur du domaine), l'idée étant de remplir au mieux l'espace du domaine.

Hypercubes Latins

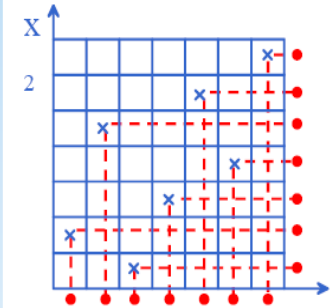


Plan hypercube latin

Plans indépendants d'un modèle statistique
Explorer l'espace pour détecter des irrégularités

- Hypercube Latin, tableau orthogonal

Projections uniformes / beaucoup de niveaux testés

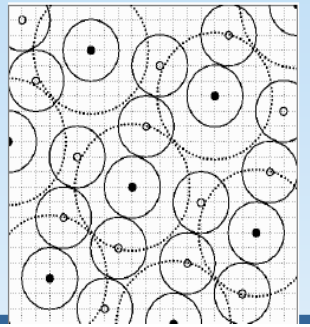


- Plans à remplissage « uniforme »
  - Suites de faible discrépance (Halton, Sobol,…), plans uniformes, etc…

distribution empirique des points ≈ distribution uniforme

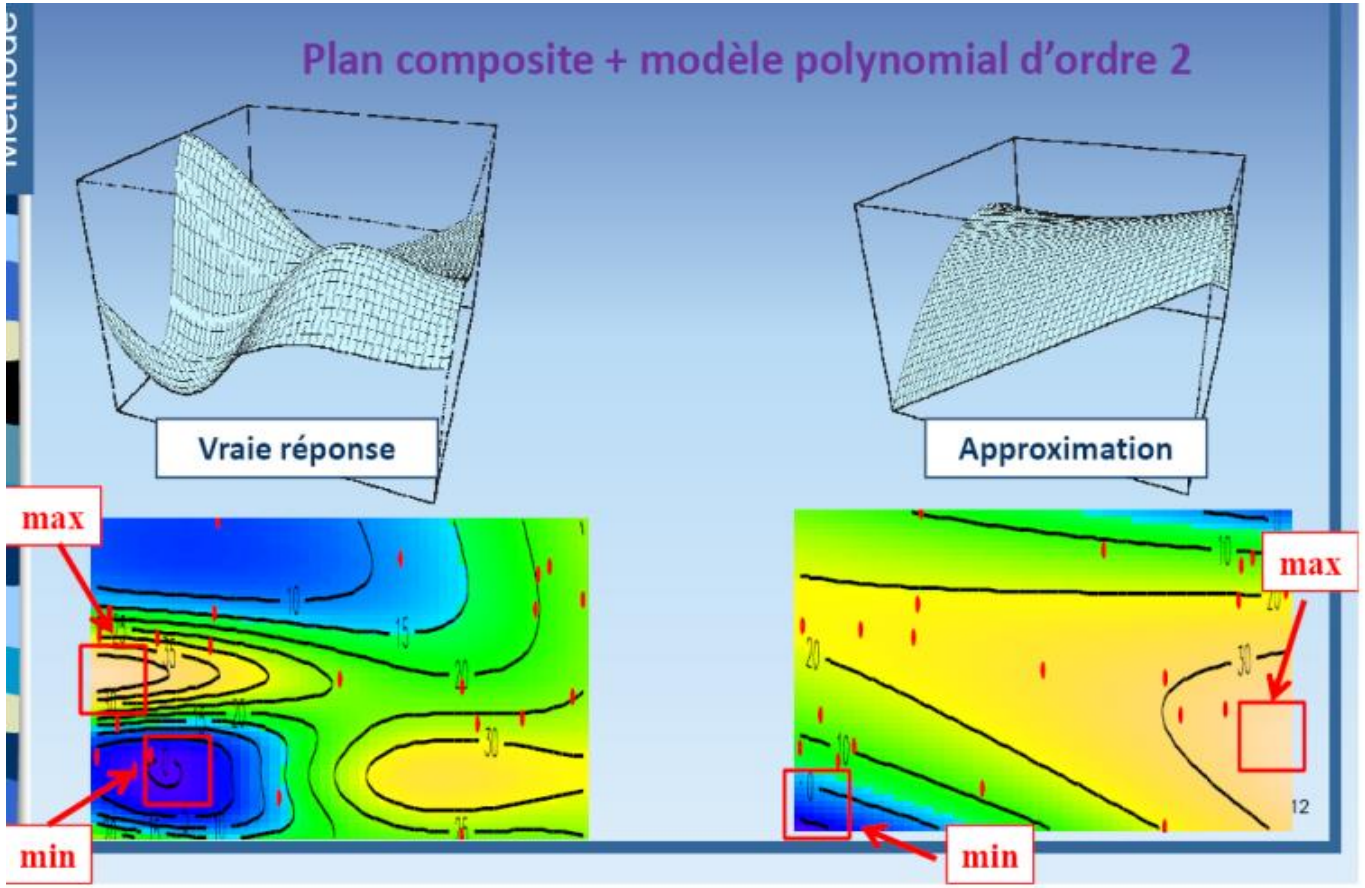
  - Plans Maximin, Plans de Audze-Eglais, etc …

Critères basés sur la distance entre les points



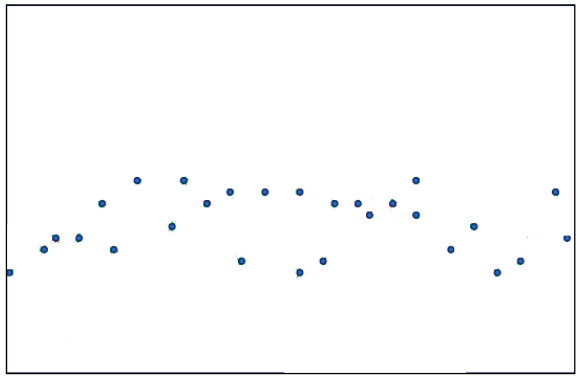
19

Phénomène complexe!

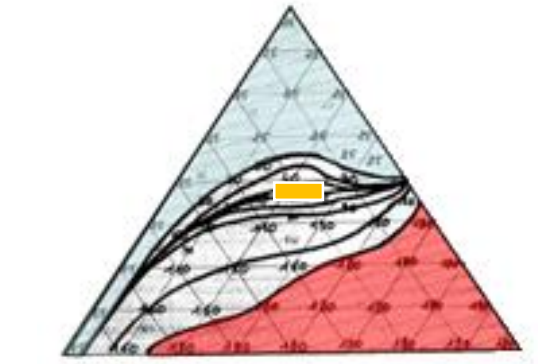


Modèles classiques?

Données expérimentales



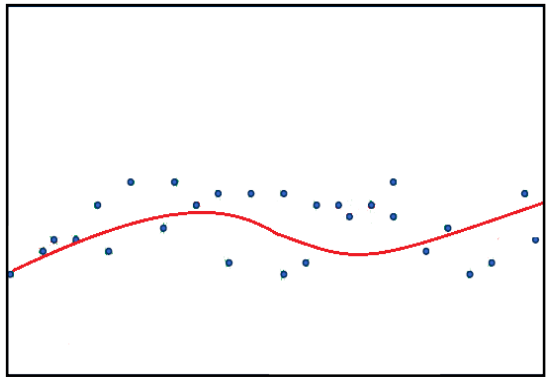
Huile 1 (%)



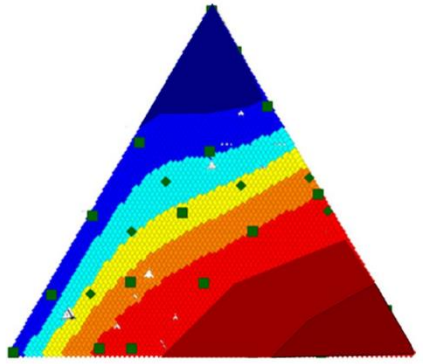
TA 1 (%) TA 2 (%)

Résultat connu

Modèle polynomial (Classique)



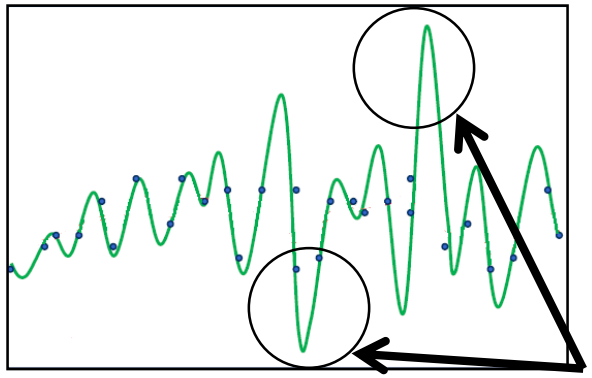
Huile 1 (%)



TA 1 (%) TA 2 (%)

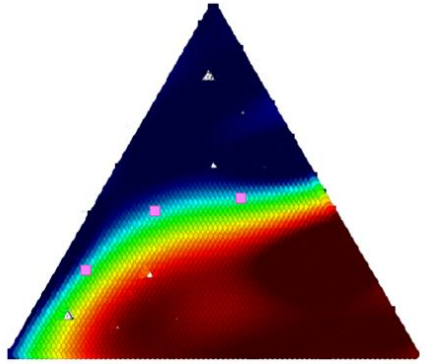
Modèle polynomial (Classique)

Modèle krigeage



Prise en compte des effets locaux

Huile 1 (%)



TA 1 (%) TA 2 (%)

Modèle krigeage

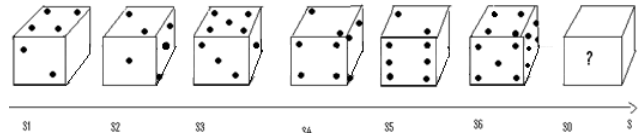
Krigeage

Notions

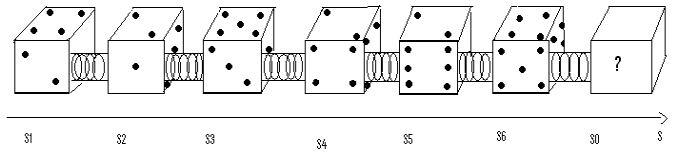
- La géostatistique: Etude des phénomènes repartis dans le temps ou dans l'espace

Variable régionalisée Z

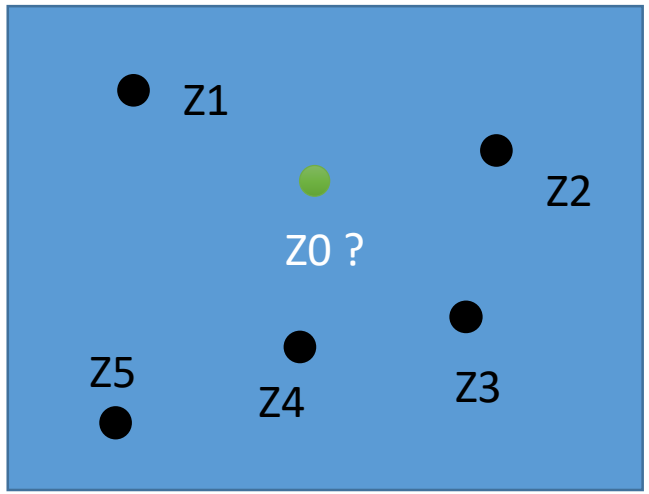
Exemple d'utilisation de fonction aléatoire: **Lancé des dés**



Comment la géostatistique modélise un telle phénomène



Problème caractérisation de la loi de la fonction →
Caractérisation des deux premiers moments



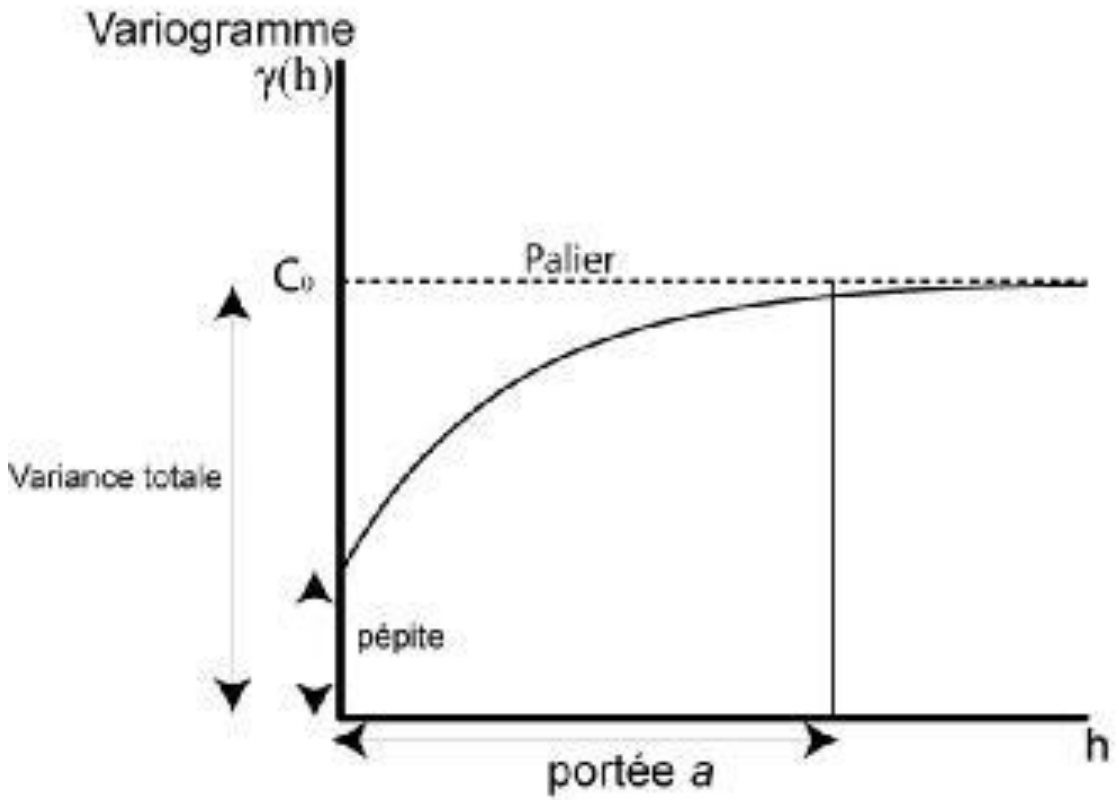
$$\hat{Z} = \sum_{i=1}^{N} W_i Z(s_i)$$

Le but est d'estimer Z0

Le variogramme

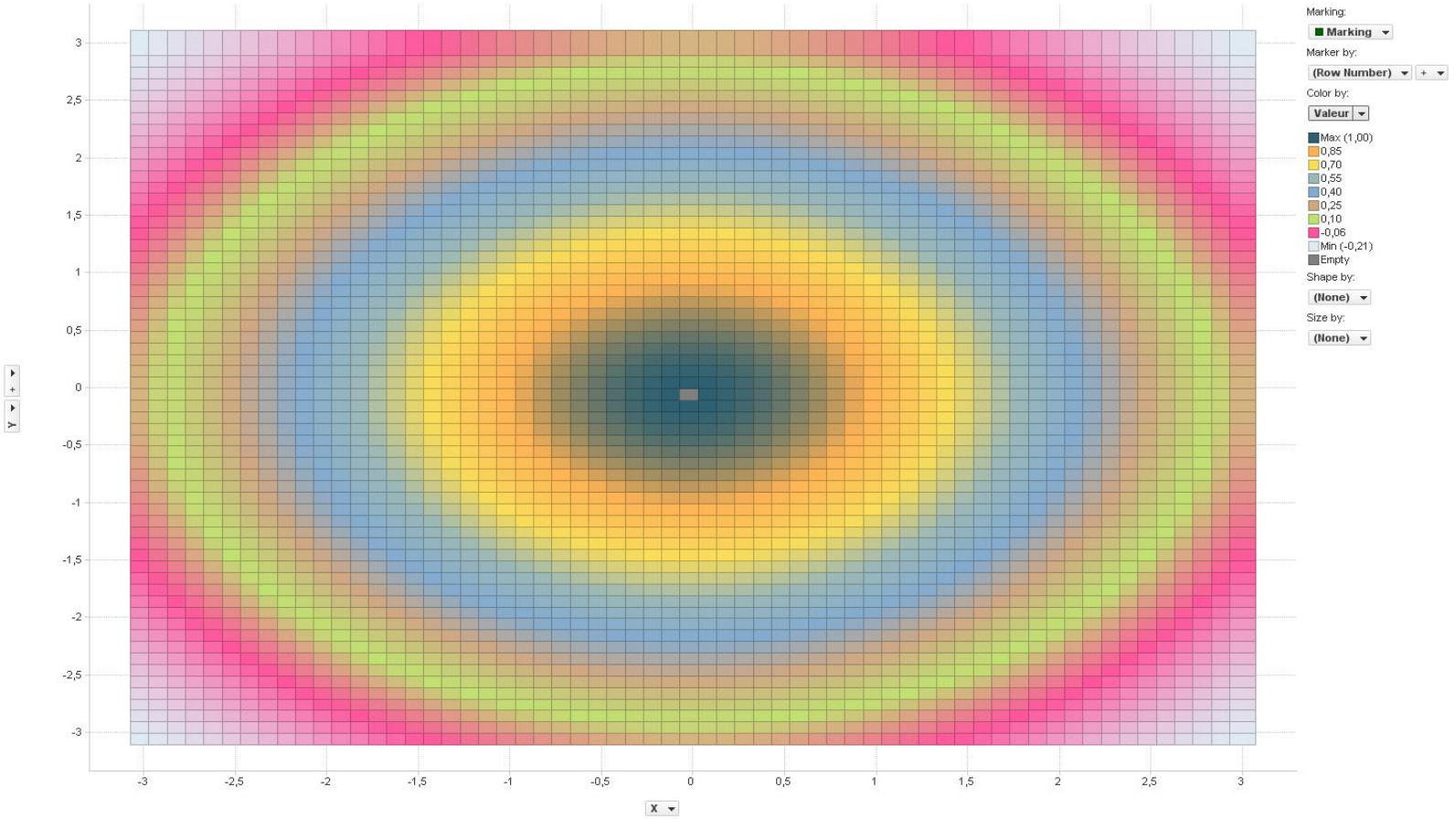
$$2\text{ème } moments = VAR\left(Z(\vec{s}) - Z\left(\overrightarrow{s+h}\right)\right) = E\left([Z(\vec{s}) - Z\left(\overrightarrow{s+h}\right) - m(\vec{h})]^2\right)$$

$$\gamma(h) = \frac{1}{2}Var(Z(s+h) - Z(s))$$

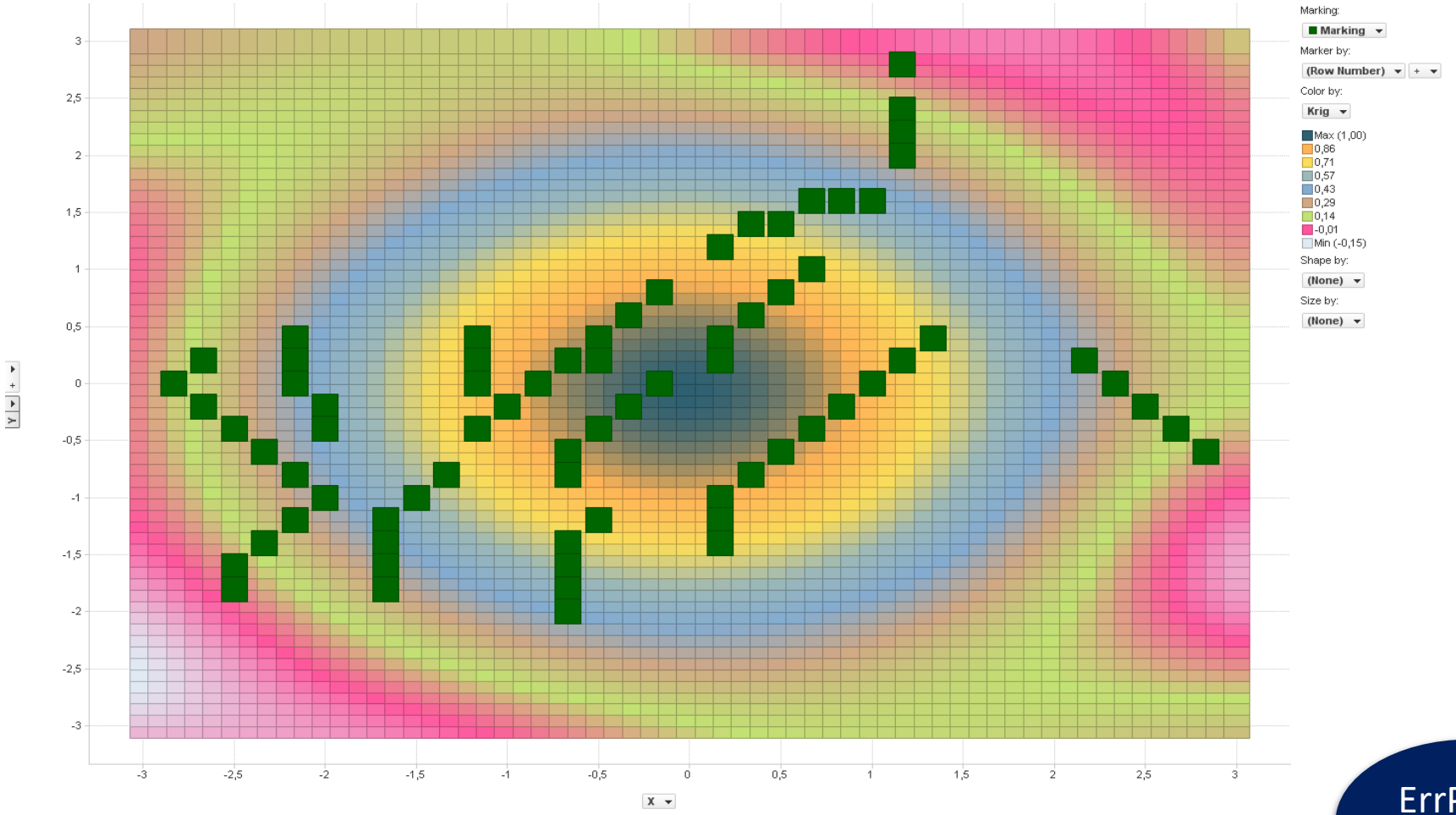


Exemple

La surface étudiée est définie par F(R) =SinR/R avec $R = \sqrt{x^2 + y^2}$ et $(x, y) \in [-3,3]^2$



Exemple (1) : Krigeage ordinaire N=74

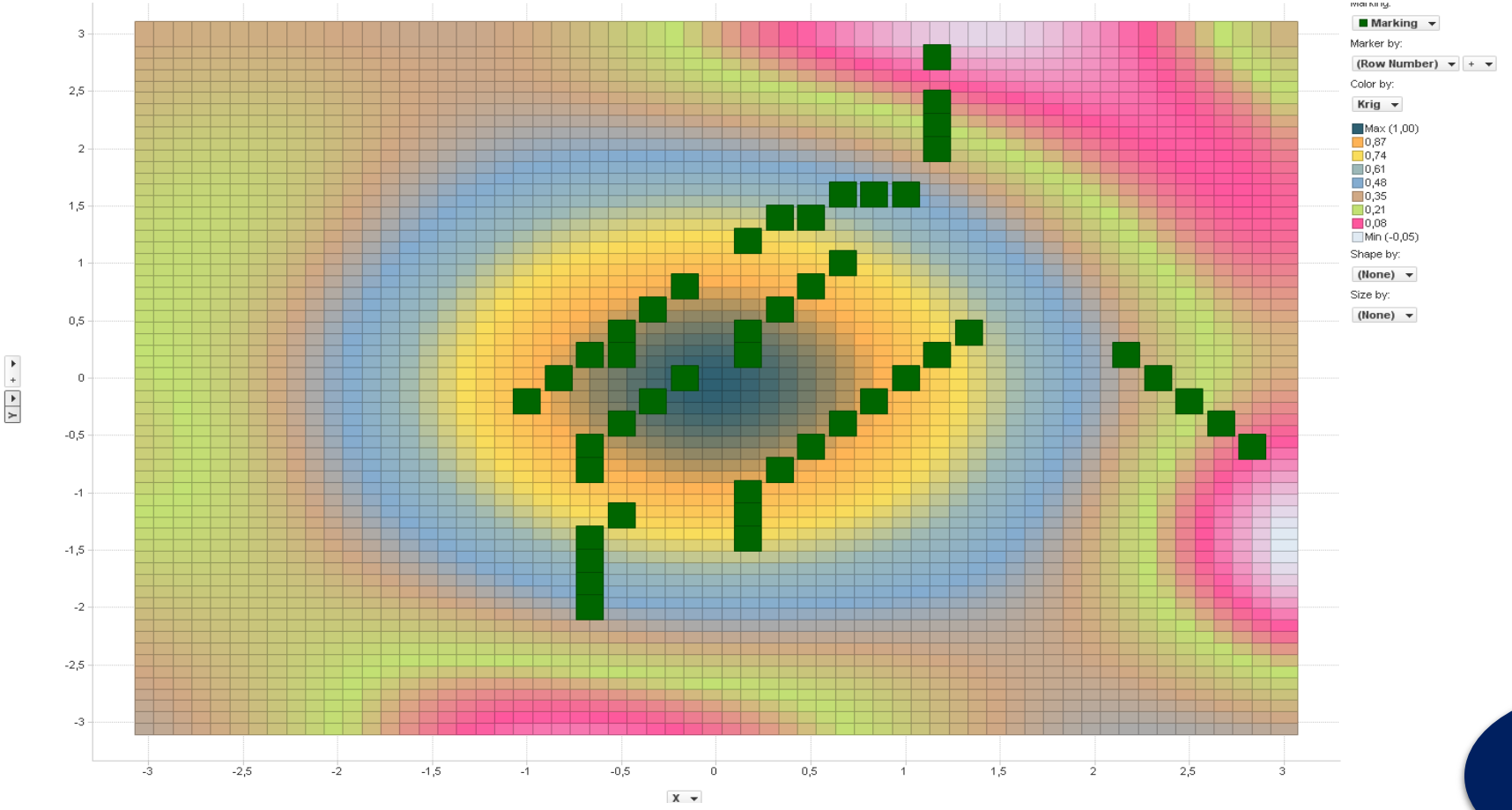


ErrPred=1/(nb de point de grille)*(krig-ValExac)²

Valmoy=moyenne de la variabilité

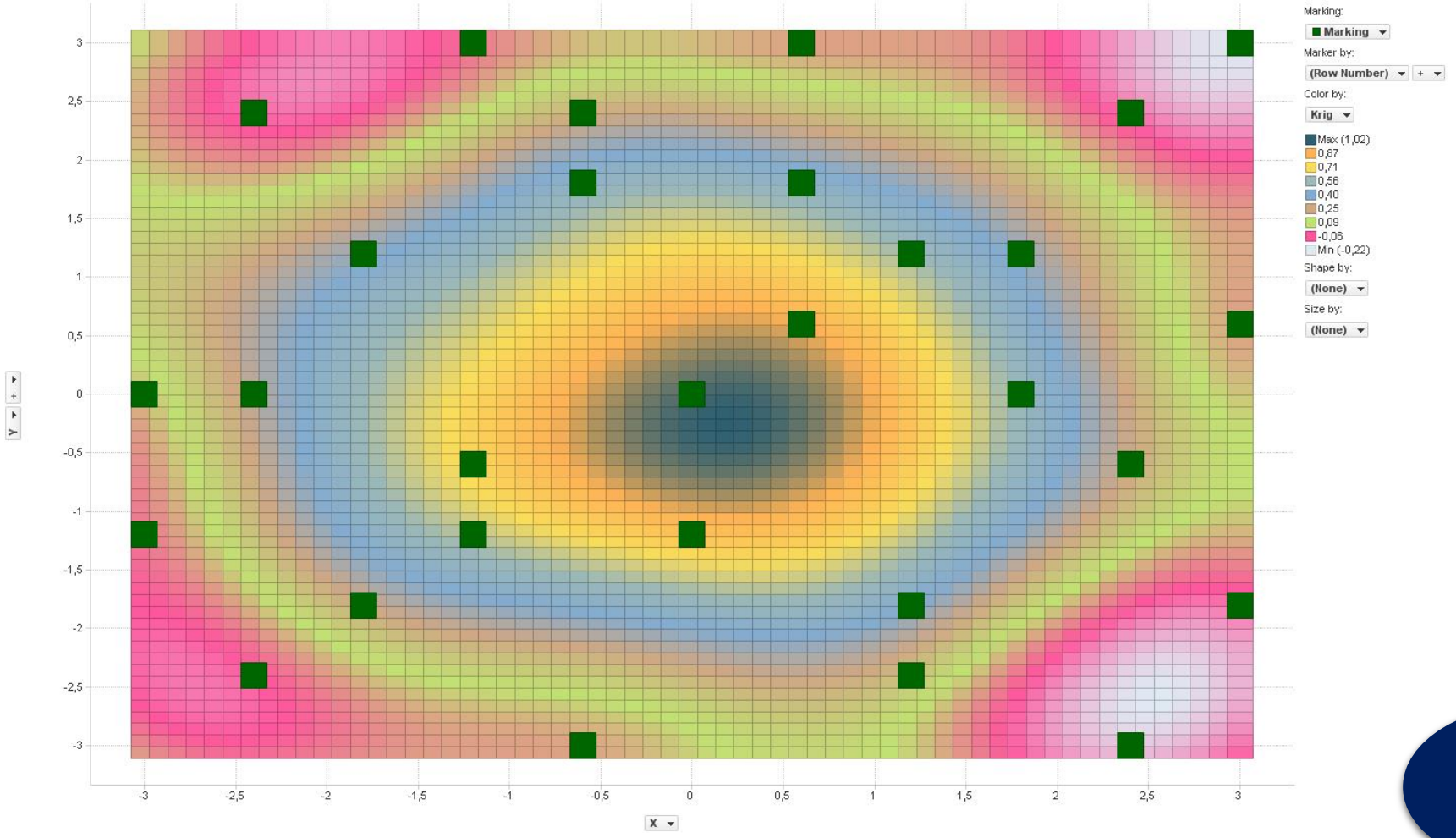
ErrPred= 0.007

Exemple (1): Krigeage ordinaire N=47



ErrPred= 0.028

Exemple (1): Krigeage ordinaire N=29

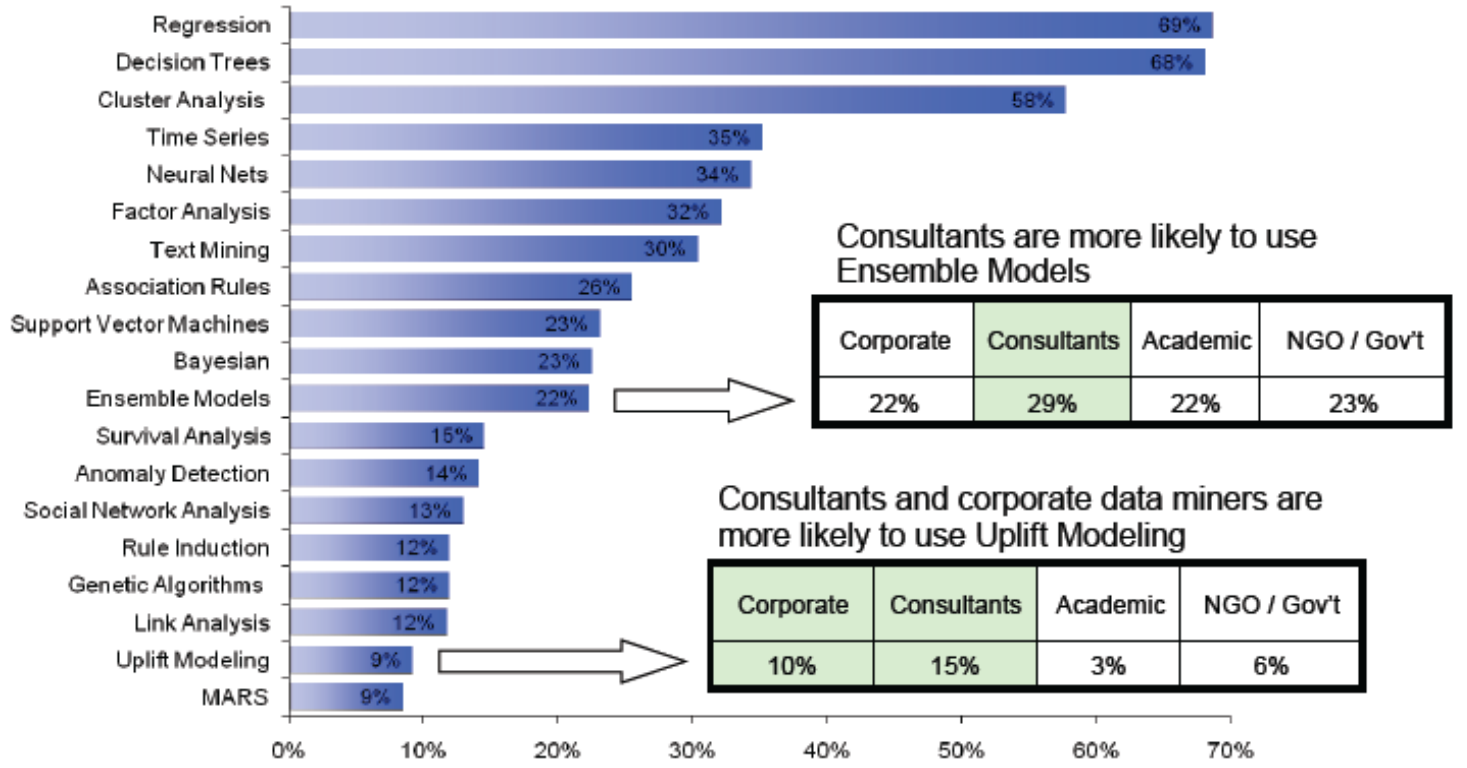


ErrPred= 0.0026

Quels modèles?

The Algorithms Data Miners are Using

- Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent over time.
- However, a wide variety of algorithms are being used.



Consultants are more likely to use Ensemble Models

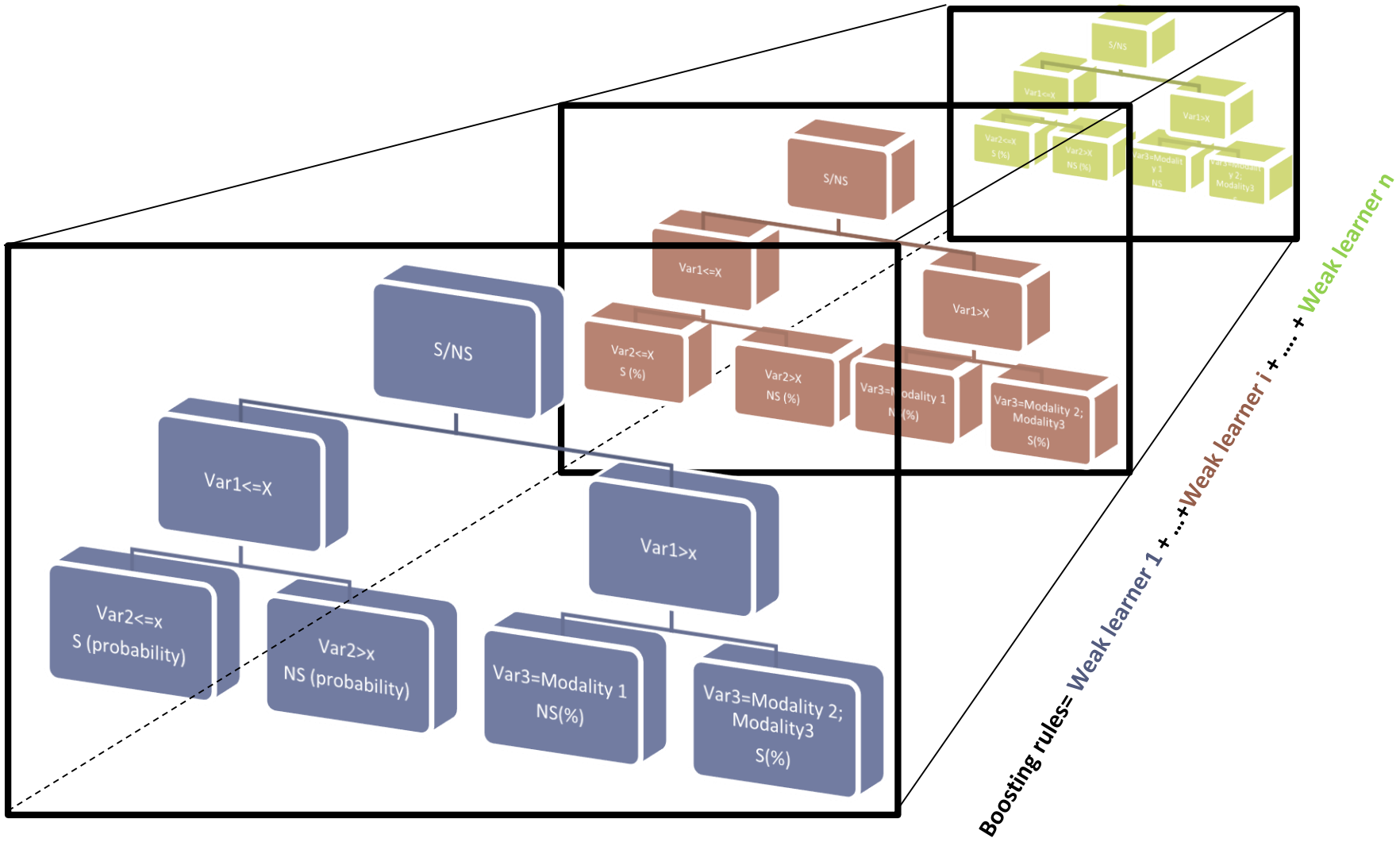
Corporate	Consultants	Academic	NGO / Gov't
22%	29%	22%	23%

Consultants and corporate data miners are more likely to use Uplift Modeling

Corporate	Consultants	Academic	NGO / Gov't
10%	15%	3%	6%

Question: What algorithms/analytic methods do you TYPICALLY use? (Select all that apply)

Ensemble Model :Boosting



Quels modèles?

Objective : Prediction of binary outcome/ Classification/Regression

A large number of supervised classification/regression models have been proposed in the Literature

Which One To Choose?

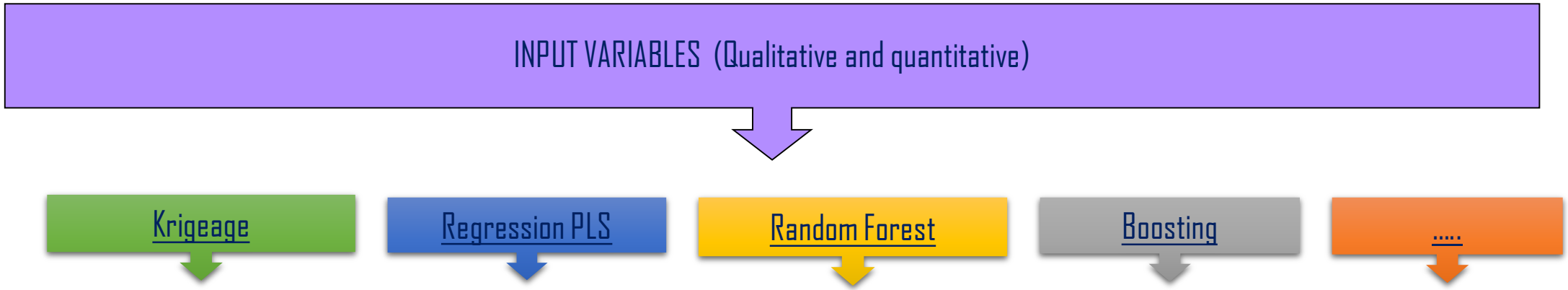
Bias induced by the use of one single statistical approach

Solution “stacking" meta-model.

Wolpert, D.(1992) Stacked Generalization. Neural Networks, 5, 241{259.

Breiman,L. (1996) Stacked Regressions. Machine Learning. 24, 49{64.

Stacking



Prédictif ou supervisé Modèles explicites de type régression, avec régularisation, arbres ..
Modèles de type boîte noire (neurones, SVM ….)

Stacking

is a combination of X supervised regression/classification methods

Le stacking: un cas particulier des méthodes d'ensemble
Bagging, Boosting, Random Forests
…

Cas l’Oréal

Specific Methodology

- "Stacking" meta-model

  Combining models → Logistic PLS-DA instead of OLS

Strong correlation between predictions

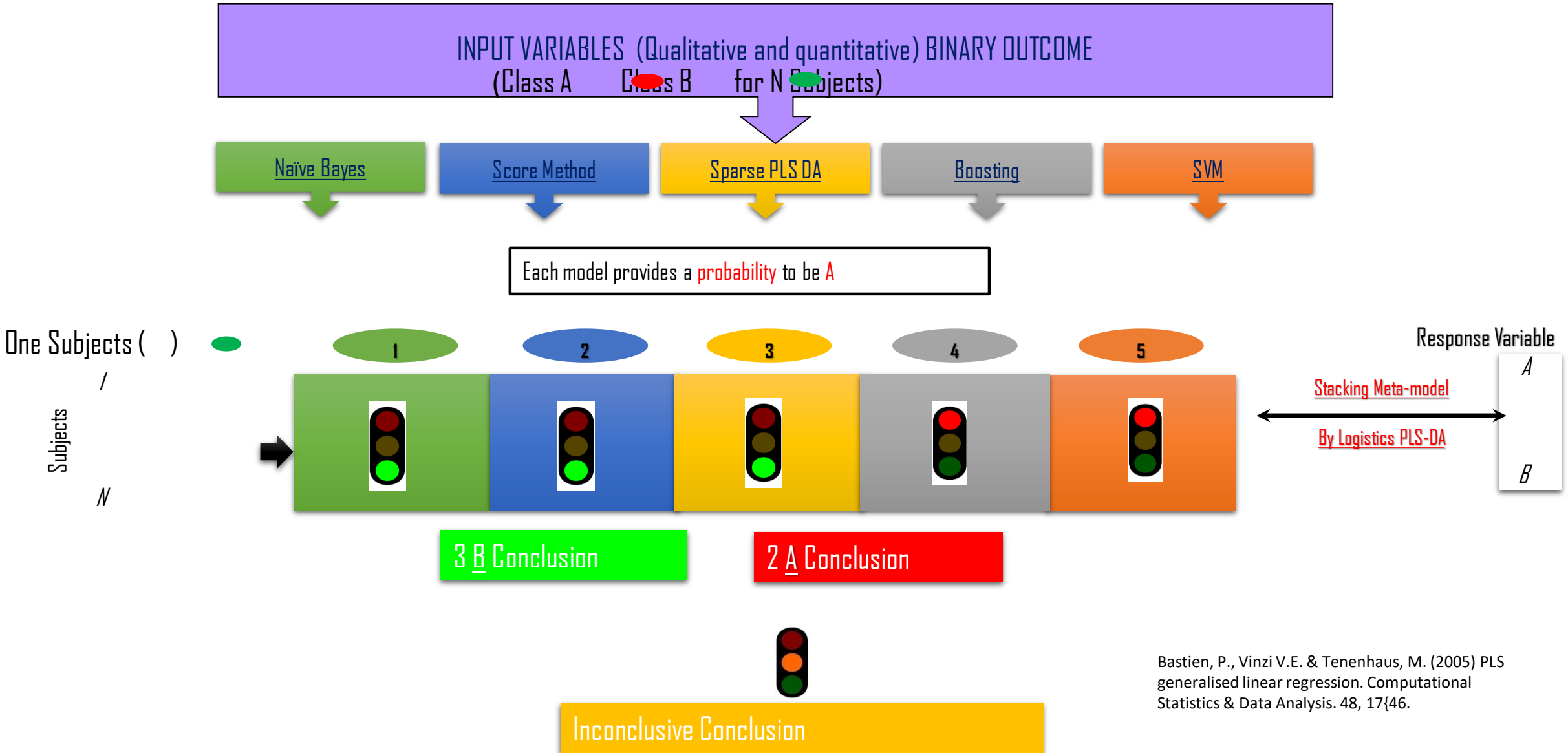
- Choice of different models

  Boosting, Naïve Bayes, SVM, Sparse PLS-DA, and Expert Scoring

- Small number of observations

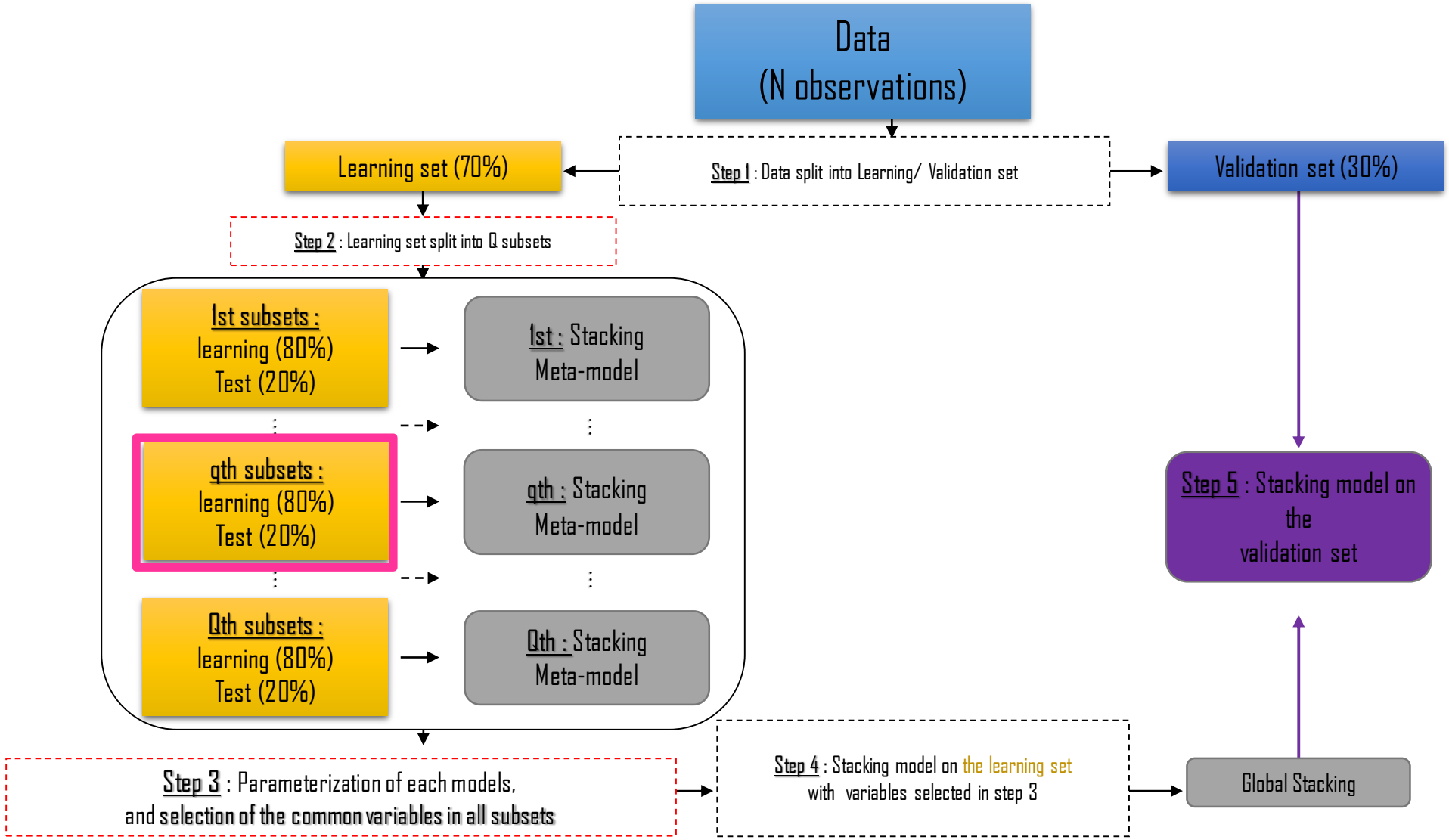
  Repeated sub-sampling for variables selection

Visualization of the methodology



Bastien, P., Vinzi V.E. & Tenenhaus, M. (2005) PLS generalised linear regression. Computational Statistics & Data Analysis. 48, 17{46.

Process of validation rules



Parametrization process for each model

qth subsets :
qth learning (80%)
qth Test (20%)

1

Creation of k-fold cross-validation :
A(1), A(2), ..., A(k) each of the k-fold respect
the proportion of tow groups A and B

2

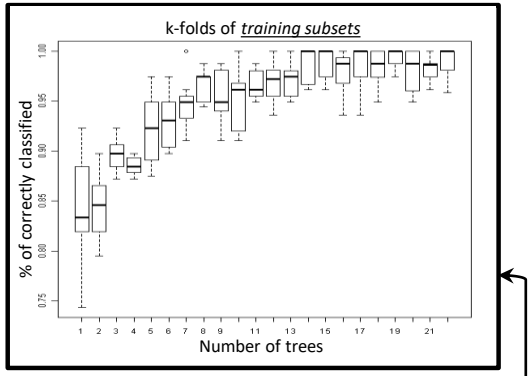
Subsequently, a single subsample is retained as the validation data for testing the model, and the remaining k – 1 subsamples are used as training data

3

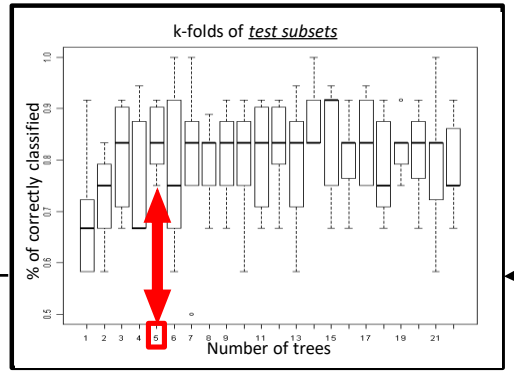
Research of the optimal parametrization
for each model by testing their
performances on the k tests subset (k-fold)

4

Example boosting model :
These performances are represented
on these graphs



5



6

The parameters for each model
are chosen accordingly
to the best results on the
k tests subsets (k-fold)

7

The parameters selected by 5 trees,
are validated by
qth Test

Outils Evaluation Performances

- Comparisons between the five methods and the combined models are done according to two features:
  - global performance with ROC analysis
  - Concordance assessed by Principal component analysis (PCA)
- A decision system with three intervals is finally proposed to the expert, with a no-decision region.

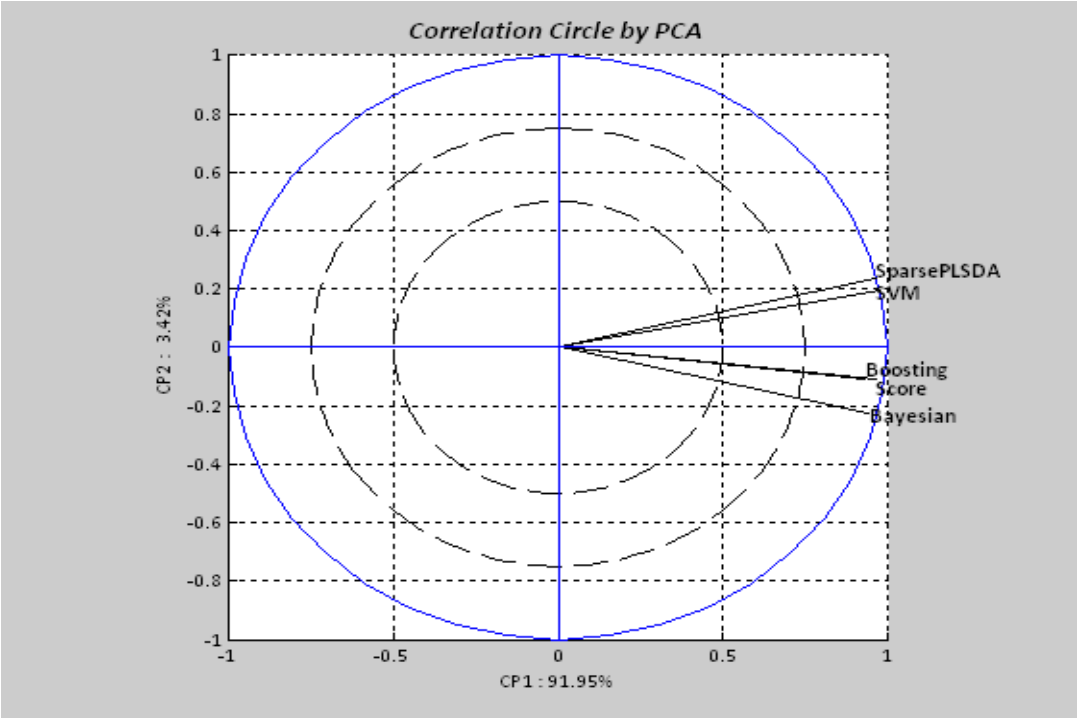
Statutory Context

- L'Oreal is developing approaches for safety evaluation (the evaluation of skin sensitization) of ingredients by combining multiple in vitro and in silico data.

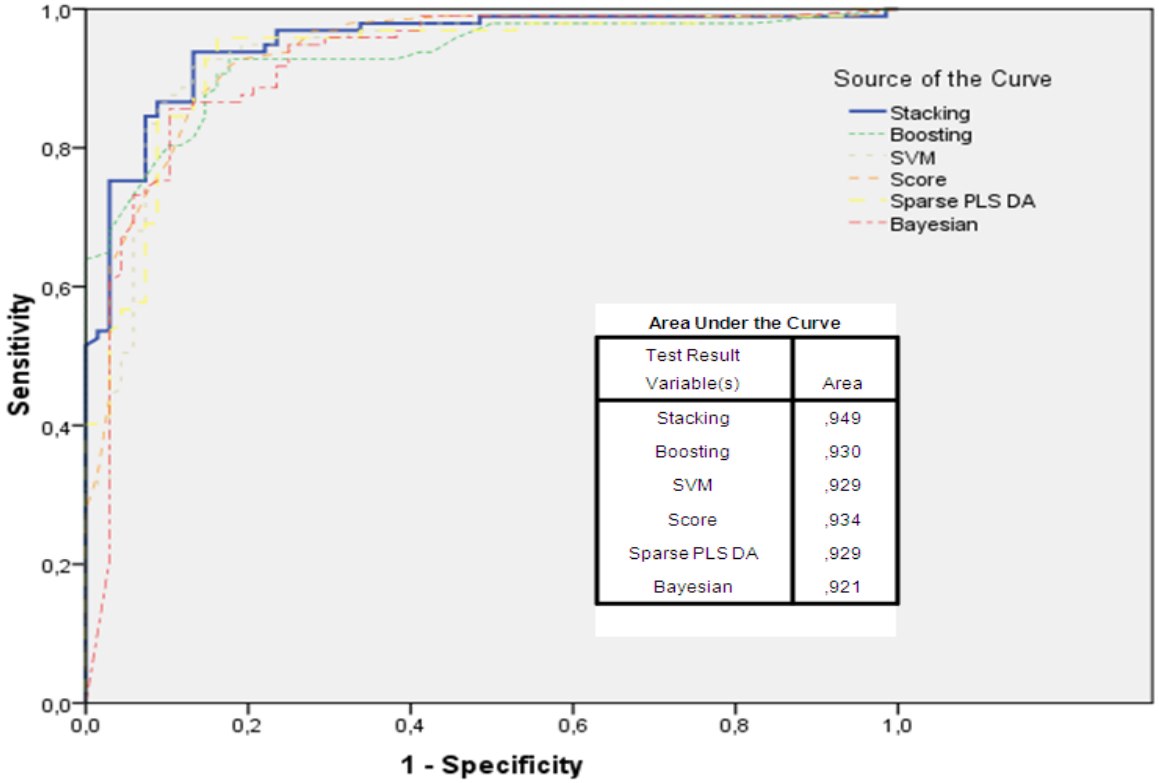
- Data :
  - For this purpose we used a full data set on 165 chemicals composed of 35 different variables, representing
  - the results from *in silico* predictions (Derek, TIMES, Toxtree), from DPRA, MUSST,
  - Nrf-2 and PGE-2 *in vitro* assays as well as numerous physico-chemical experimental or calculated parameters

- In order to predict substances into two groups (Sensitizer/No-Sensitizer)

Experimental result Context

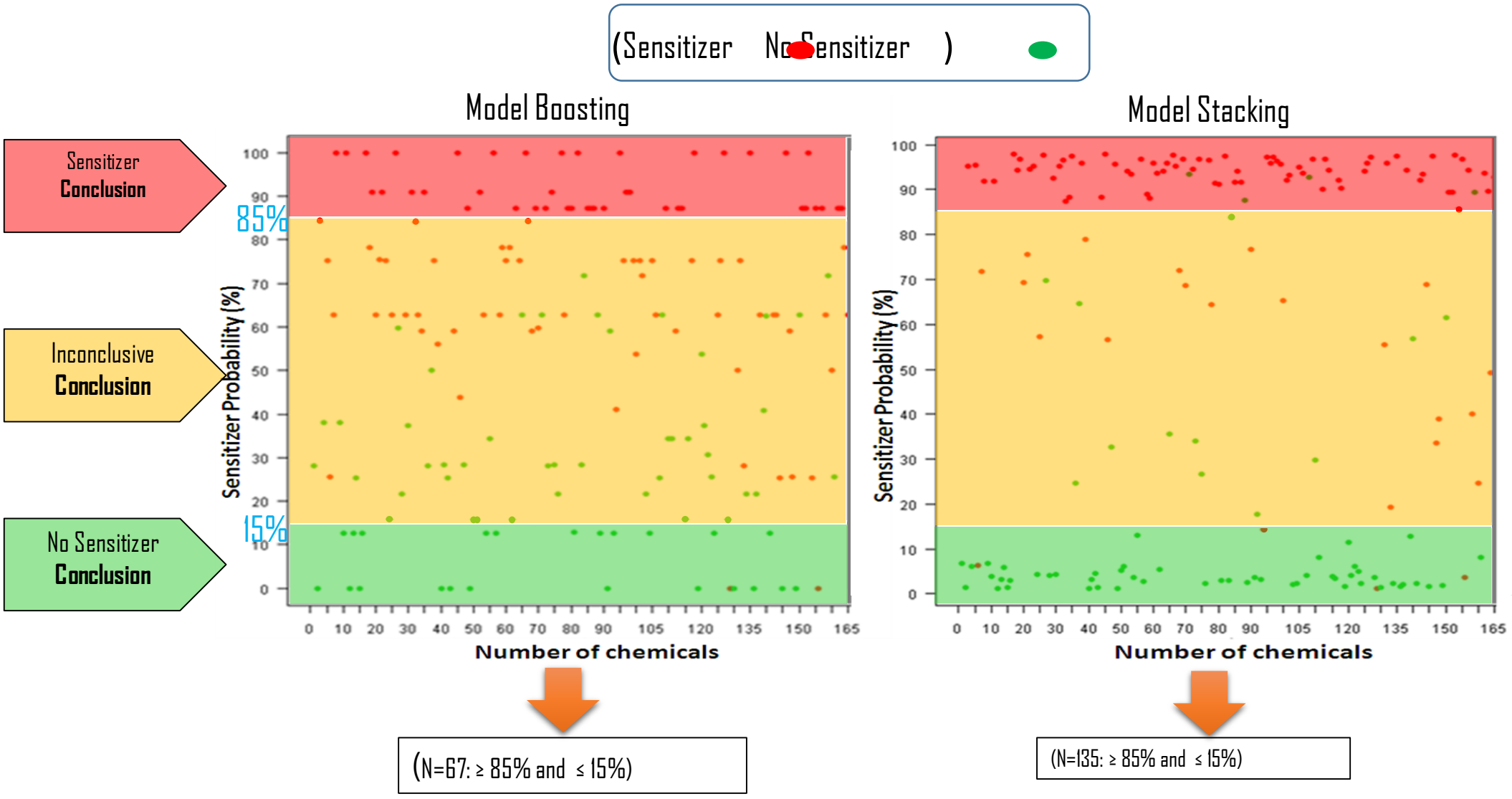


Predictions provided by the five models are obviously highly positively correlated as shows the following PCA analysis



The stacking appears to be the most efficient (blue curve) with the highest area under the curve (0.949).

Confidence area of the boosting model and of the stacking model



Performances on the validation set (N = 50)

- Performance comparisons on a validation set (25 Sensitizer and 25 No Sensitizer) :
  - Take into account only high probabilities (≥ 85% and ≤ 15%) :

Predicted Class	Boosting	Score	Sparse PLSDA	SVM	Naïve Bayes	Stacking
Incoclusive	30	29	15	12	12	10
True Sensitizer	11	7	16	19	19	20
True No Sensitizer	7	13	15	15	16	17
False Predicted	2	1	4	4	3	3
Conclusive	20	21	35	38	38	40

Performances of the prediction	Boosting	Score	Sparse PLSDA	SVM	Naïve Bayes	Stacking
Sensitivity	84.61	87.50	84.21	86.36	95.00	91.00
Specificity	100	100	93.75	93.75	89.00	94.44
Concordance	90.00	95.24	88.57	89.47	92.00	92.50
Kappa	79%	89%	77%	78%	84%	85%

Results show that stacking model has better performance than all the other models taken separately on a larger set

Improving Stacking Methodology for Combining Classifiers: Applications to Cosmetic Industry.
Electronic Journal of Applied Statistical Analysis. 2016. Vol 9, No 2, 340-361

OCDE: POSITIONNEMENT

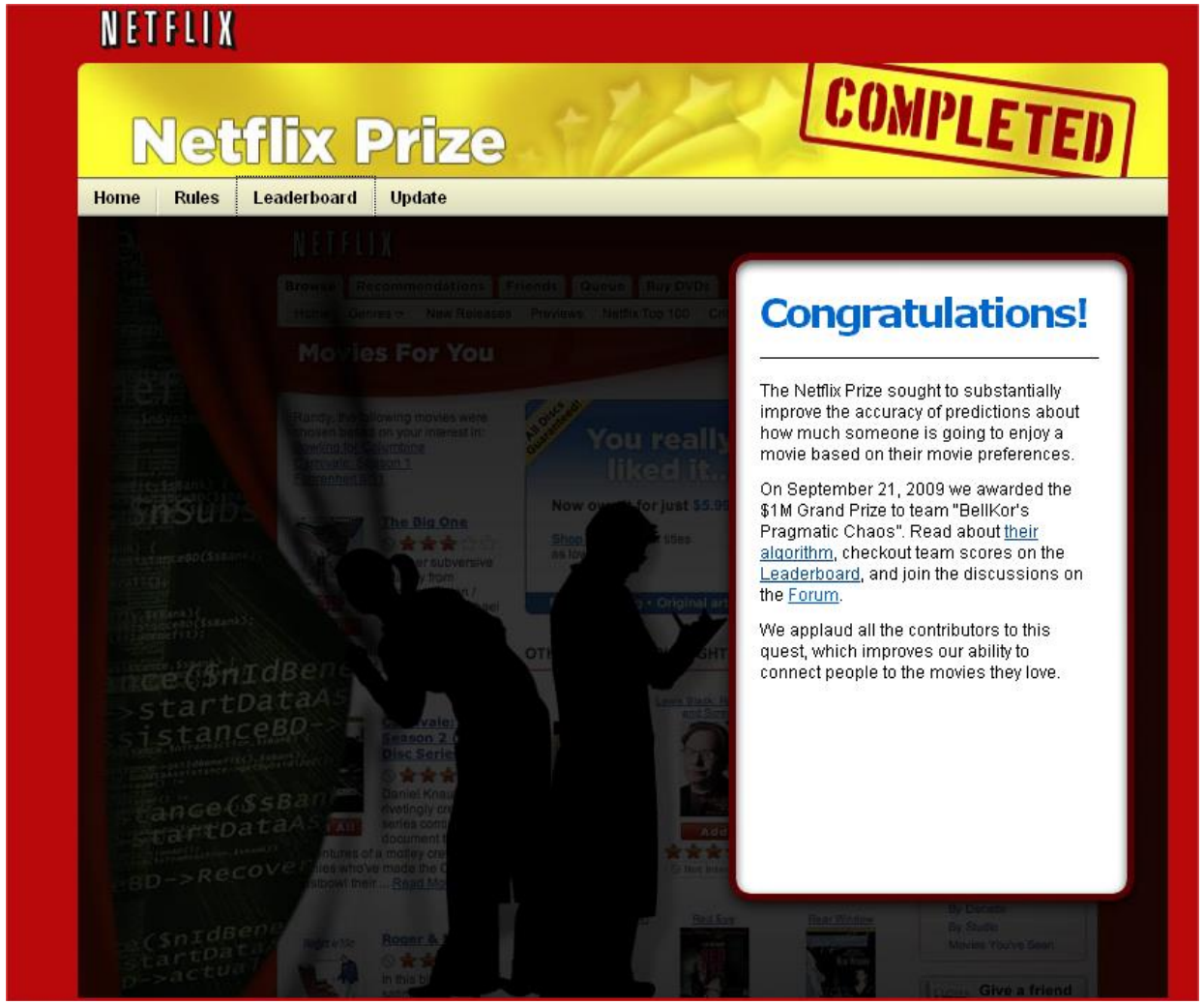
Guidance Document On The Reporting Of Structured Approaches To Data Integration And Individual Information Sources Used Within IATA For Skin Sensitization (final version April 2016) :

1. L'OREAL Stacking Meta-model for Skin Sensitization Hazard Identification.
2. BASF Structured Approaches to Data Integration Reporting Format (RF)
3. KAO Sensitization Potency Prediction Based on Key Event 1 and 3.
4. GIVAUDAN Sensitizer Potency Prediction Based on Key Event 1 + 2
5. RIVM Structured Approaches to Data Integration Reporting Format (RF)
6. UNILEVER DIP For Skin Allergy Risk Assessment (SARA)
7. ICCVAM Integrated Decision Strategy for Skin Sensitization Hazard
8. Procter & Gamble Sensitizer Potency Prediction Based on Key Event 1+2+3
9. SHISEIDO The Artificial Neural Network Model for Predicting LLNA3 EC3

Méthodologies statistiques d'IATA

IATA	Méthodes	type de modèles
L'Oréal	PRISME	combine de 5 modèles différents
Basf	Expert-Scoring	non-linéaire
Kao	SPLS-DA	Linéaire
Givaudan	Régression linéaire	Linéaire
Rivm	Arbre de décision (SAS)	non-linéaire
Unilever	Réseaux bayésiens	non-linéaire
Iccvam	SVM	non-linéaire
Procter Gamble	Approche bayesien	non-linéaire
Shiseido	Réseaux de neurones	non-linéaire

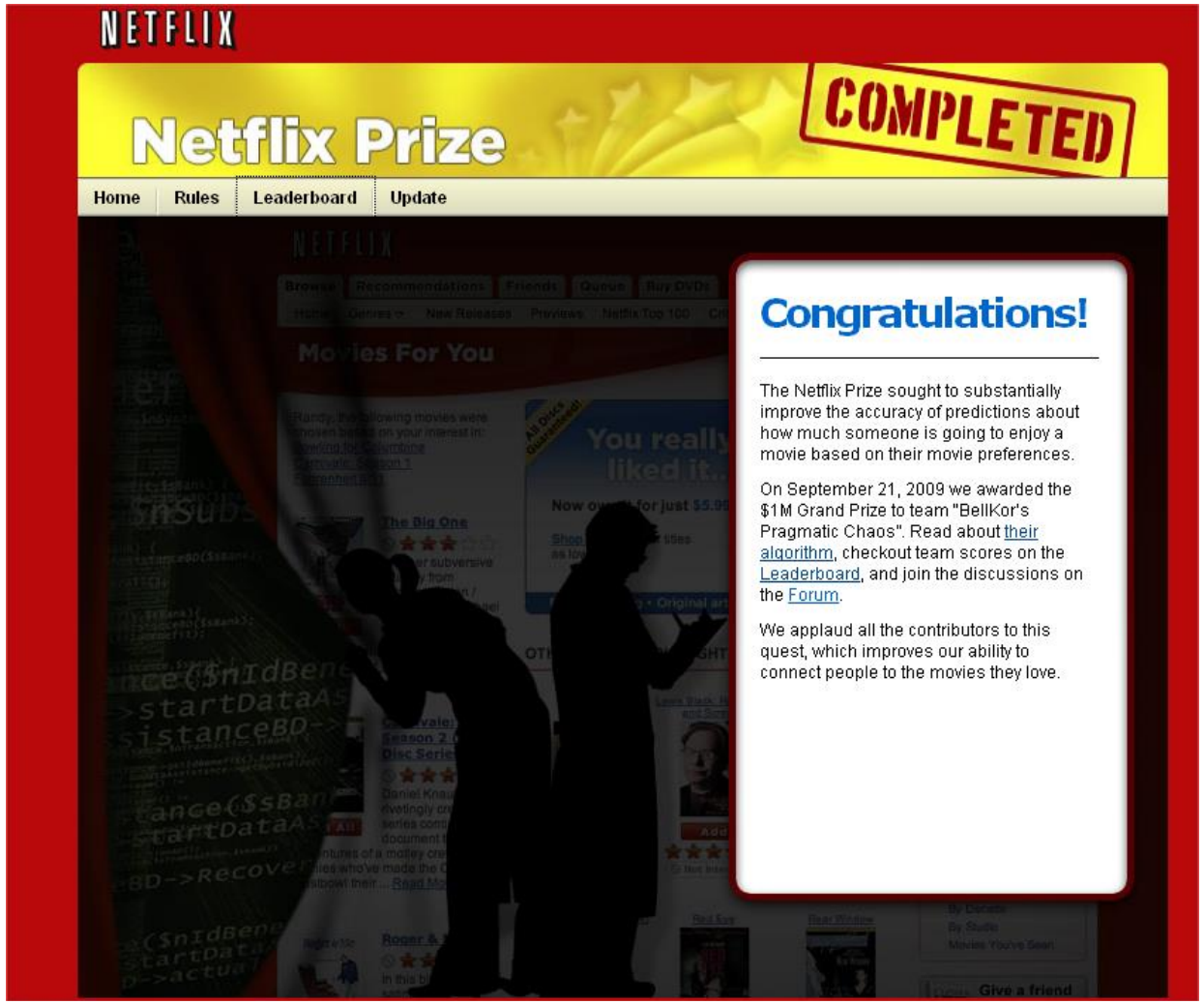
Exemple



The Netflix dataset contains more than 100 million datestamped movie ratings performed by anonymous Netflix customers between Dec 31, 1999 and Dec 31, 2005. This dataset gives ratings about m = 480 189 users and n = 17 770 movies

The contest was designed in a training-test set format. A hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user (or fewer if a user had not rated at least 18 movies over the entire period).The remaining data made up the training set.

Exemple



The winner : « BellKor's Pragmatic Chaos » team.
A blend of hundreds of different models Test RMSE for Bellkor's Pragmatic Chaos: 0.856704 (10.06%)

•“The Ensemble Team”. Blend of 24 predictions
Test RMSE for The Ensemble: 0.856714 (10.06%)

•Bellkor's Pragmatic Chaos defeated The Ensemble by submitting just 20 minutes earlier!

Point d’attention

Modèle pour comprendre

VS

Modèle pour prévoir

Lien

Saporta: Quelles statistique pour les Big Data?

DataRobot

www.**kaggle**.com/competitions