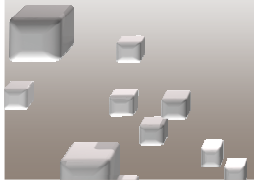
Automatic Modelling Tools for Classification and Prediction

Libei Chen
Managing Director R&D Vadis & Bitong
Invited Professor, STAT/UCL

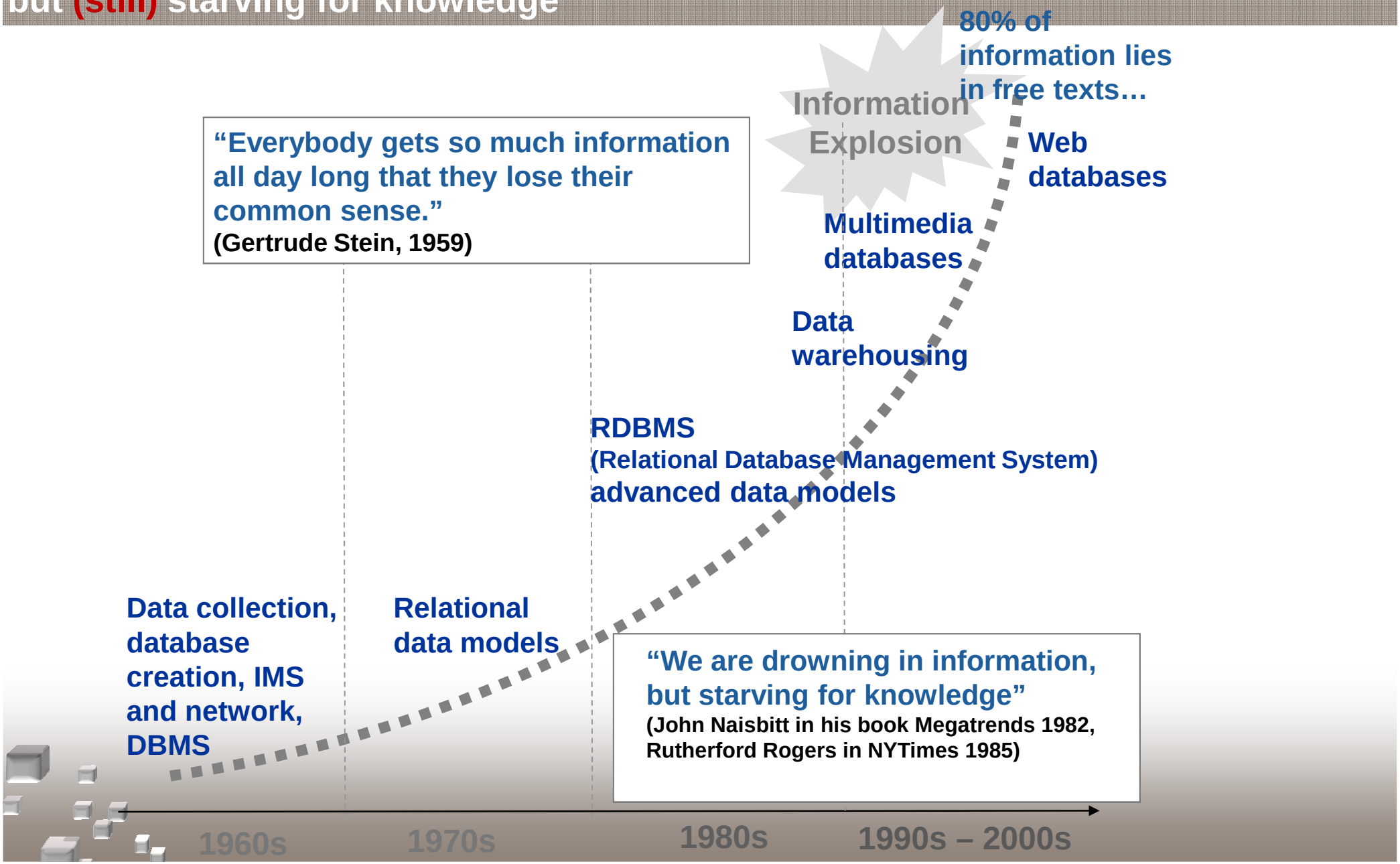
May, 2010

Agenda

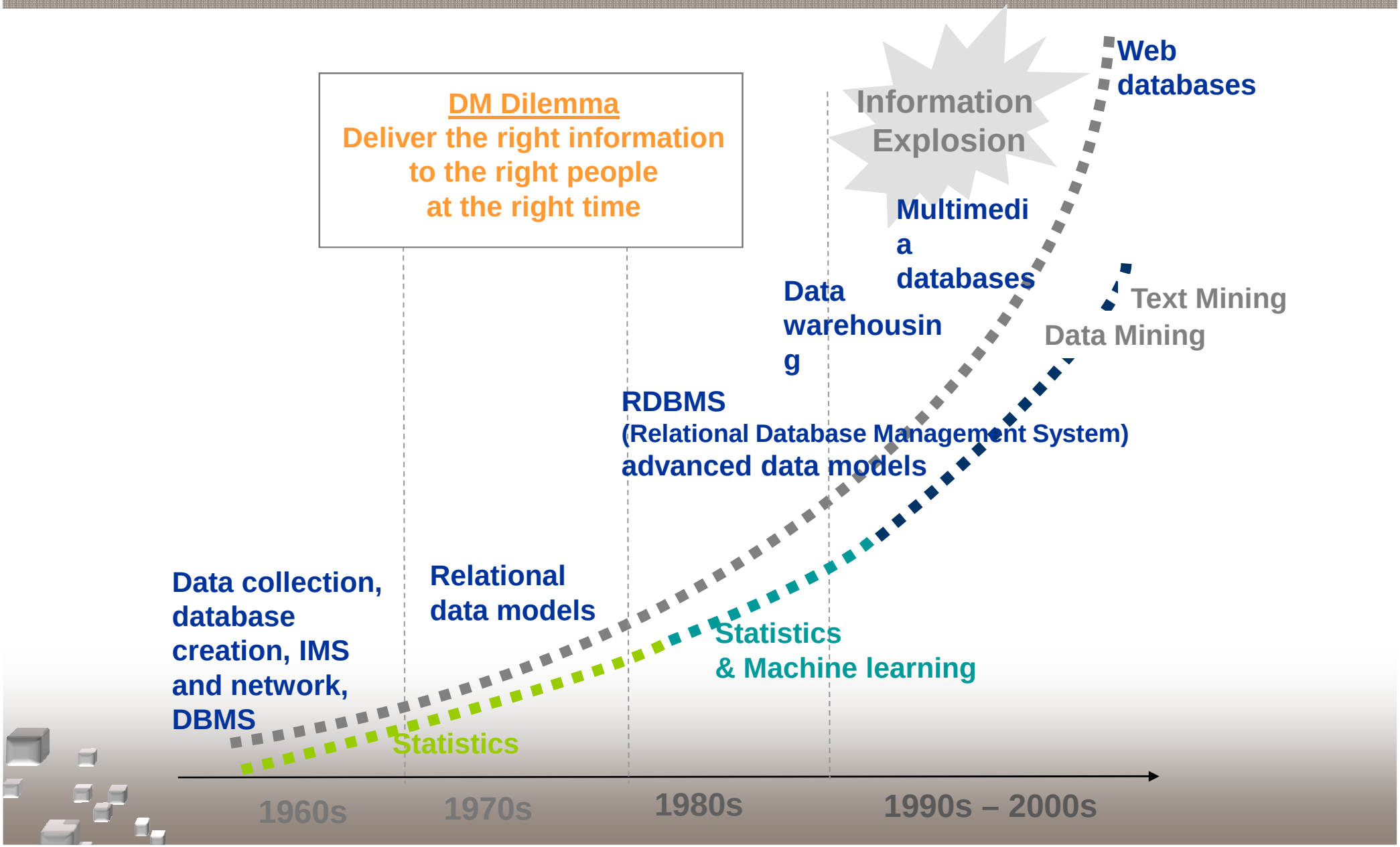
- Motivation
- Methodology
- Data Mining background and Rank
- Massive Modelling framework



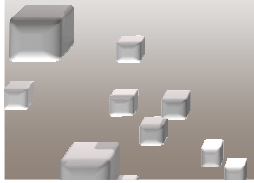
We are (more and more) drowning in information but (still) starving for knowledge



Data mining: a result of the natural evolution of information technology



Data Mining Background



Model classification from analyis point of view

Predictive

- Classification
  - Discrete outcome to classify
- Estimation
  - Continuous outcome, i.e. estimation of values, price, etc.
- Prediction
  - More focus on future prediction based on trained models
  - Including both classification and estimation
- …

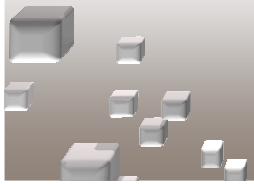
Known outcome (target) from historical data

Build models and test on knowns – to predict the unknown

Descriptive

- Clustering
  - Seek to segment/cluster data based on similarity between records
- Affinity grouping/Association rules
  - Determine what things go together (basket analysis)
- Outlier analysis
  - noise or exception, useful in fraud detection, rare events analysis
- Description and visualization
  - Describe what is going on in a complicated database
  - Increase understanding of people, porducts or processes
- …

No known outcome (target)



Typical input form and predictive model outcome

Customer ID	x1	x2	…	xn	Target (binary decision)	y2 (classes)	y3 (Value)
1					1	C1	1100
2					0	C2	800
3					1	C1	1600
4					1	C3	250
5					0	C4	100
.					.	.	.
.					.	.	.
.					.	.	.
150,000					0	C1	90

Predictive models based on « target » (labelsed classes, values, etc)

Customer ID	Probability to be target =1	Predicted class (estimated y2)	Value estimation (estimated y3)
1	0.9	C1	1000
2	0.7	C2	905
3	0.65	C1	1500
4	0.65	C3	200
5	0.6	C4	100
.	.	.	.
.	.	.	.
.	.	.	.
150,000	0.0001	C1	90

Data Mining Techniques and Algorithms

Predictive == Supervised learning

- **Estimation**:
  - Linear regressions
  - Nearest neighbours
  - Neural Networks (80's)
  - Robust regressions (90's)
  - …
- **Classification**:
  - Logistic regression
  - Decision Trees (80's)
  - Discriminant analysis
  - Support Vector Machines (90's)
  - All regression techniques
  - …

Descriptive == Unsupervise learning

- Clustering:
  - K-means clustering
  - Hierarchical clustering
    - Agglomerative
    - Divisive
  - Self-Organizing-Maps – Kohonen Network
  - …
- Link Analysis:
  - Associative rules basket analysis)
  - Sequences
- Principal Component Analysis
- …

Typical input form and descriptive model outcome

Customer ID	x1	x2	…	xn	Target (binary decision)	y2 (classes)	y3 (Value)
1					1	C1	1100
2					0	C2	800
3					1	C1	1600
4					1	C3	250
5					0	C4	100
.					.	.	.
.					.	.	.
.					.	.	.
150,000					0	C1	90

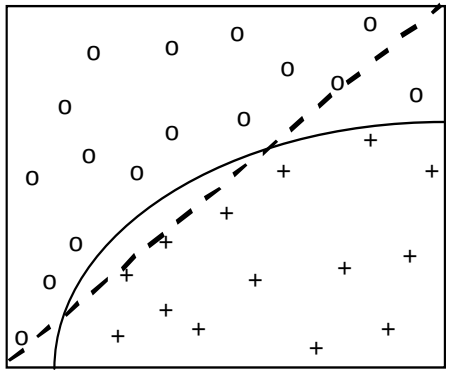
Descriptive models based on input data, often without target

Customer ID	Cluster membership	Cluster center
1	G1	0.55
2	G3	0.3
3	G2	0.01
4	G5	1.6
5	G2	0.01
.	.	
.	.	
.	.	
150,000	G4	-0.01

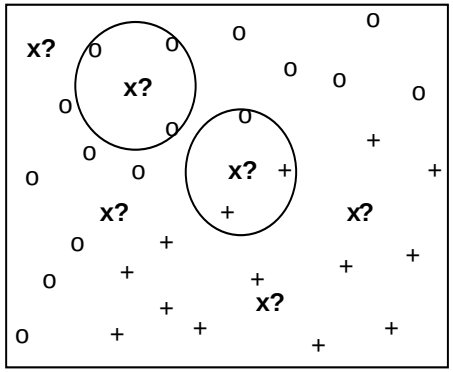
Modelling techniques illustration: classification methods

From different decision boundaries

- **Logistic Regression**
- **Discriminant Analysis**



Linear models

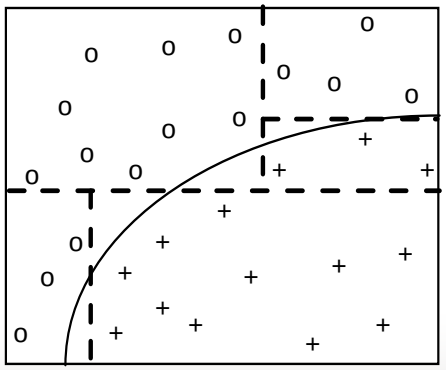


Local models

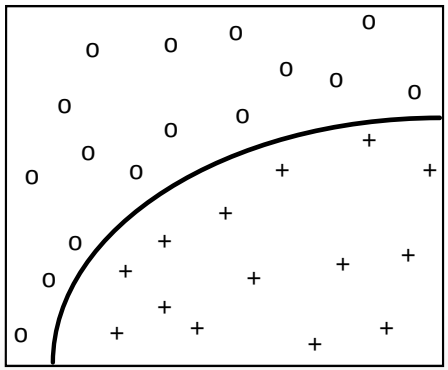
- **k-Nearest Neighbors**
- **Radial Basis Neural Networks**

Nonmetric methods:

- **Decision Trees**
- **Rule-based systems**



Recursive partitions



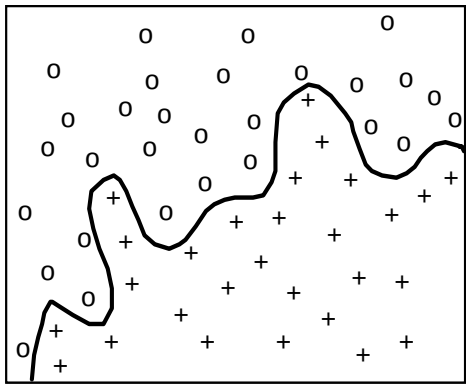
Nonlinear models

- **Multi-layer Neural Networks**
- **Kernel based models (e.g.SVM)**

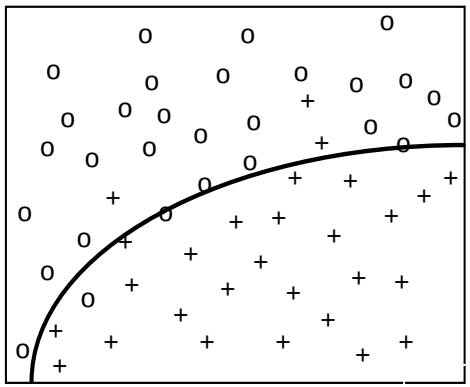
Robustness: trade-off between accuracy and generalisability

- Classification

Accurate but less robust

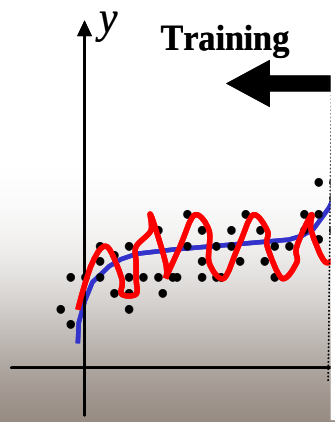


Less accurate but may have better generalisability



Model with intermediate complexity corresponding to a smooth decision boundary, relatively low misclassifications

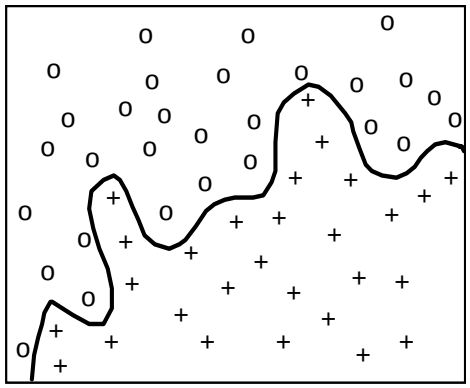
- Regression



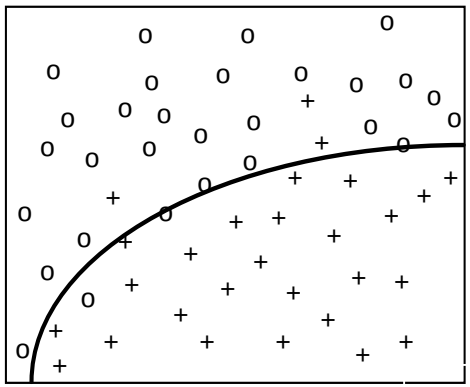
Robustness: trade-off between accuracy and generalisability

- Classification

Accurate but less robust

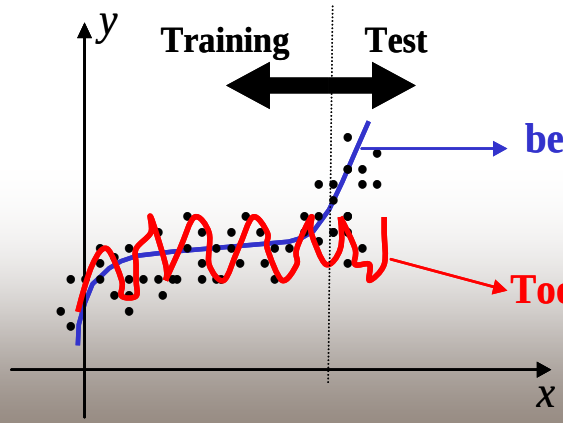


Less accurate but may have better generalisability



Model with intermediate complexity corresponding to a smooth decision boundary, relatively low misclassifications

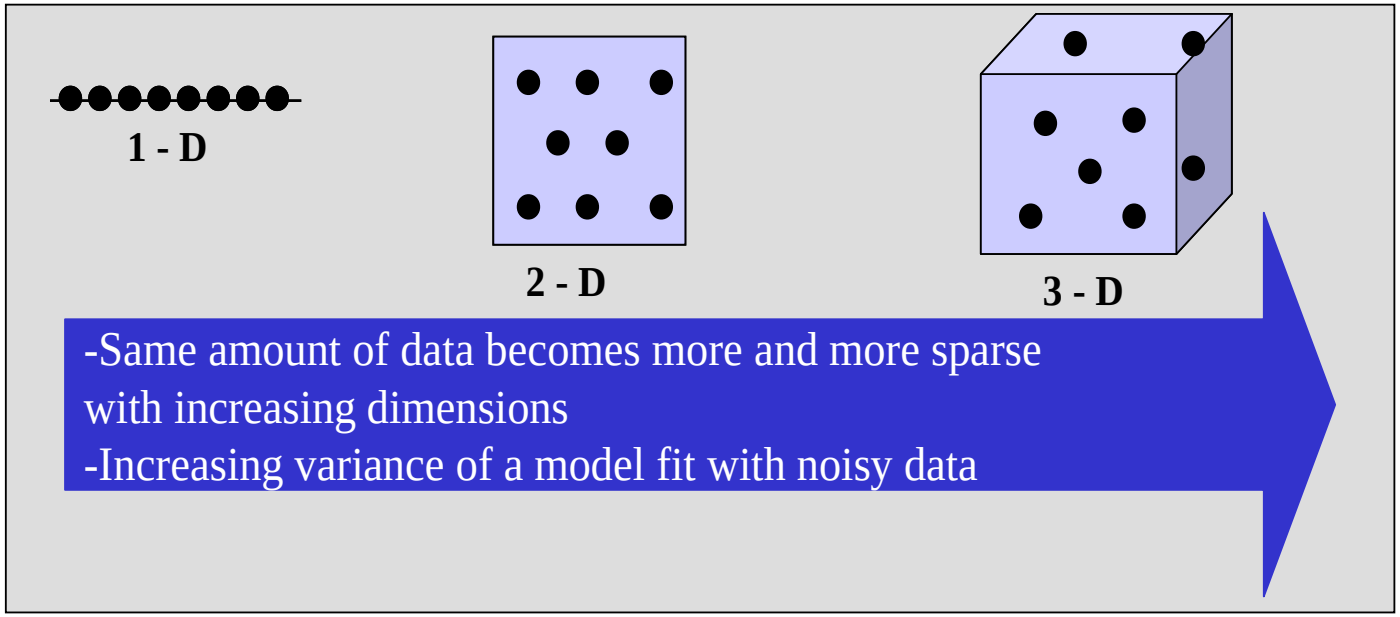
- Regression



Best Approach: Cross-validation to build more robust models

Best in Class Data Mining Challenges:

Curse of dimensionality: Data volume related to problem dimension

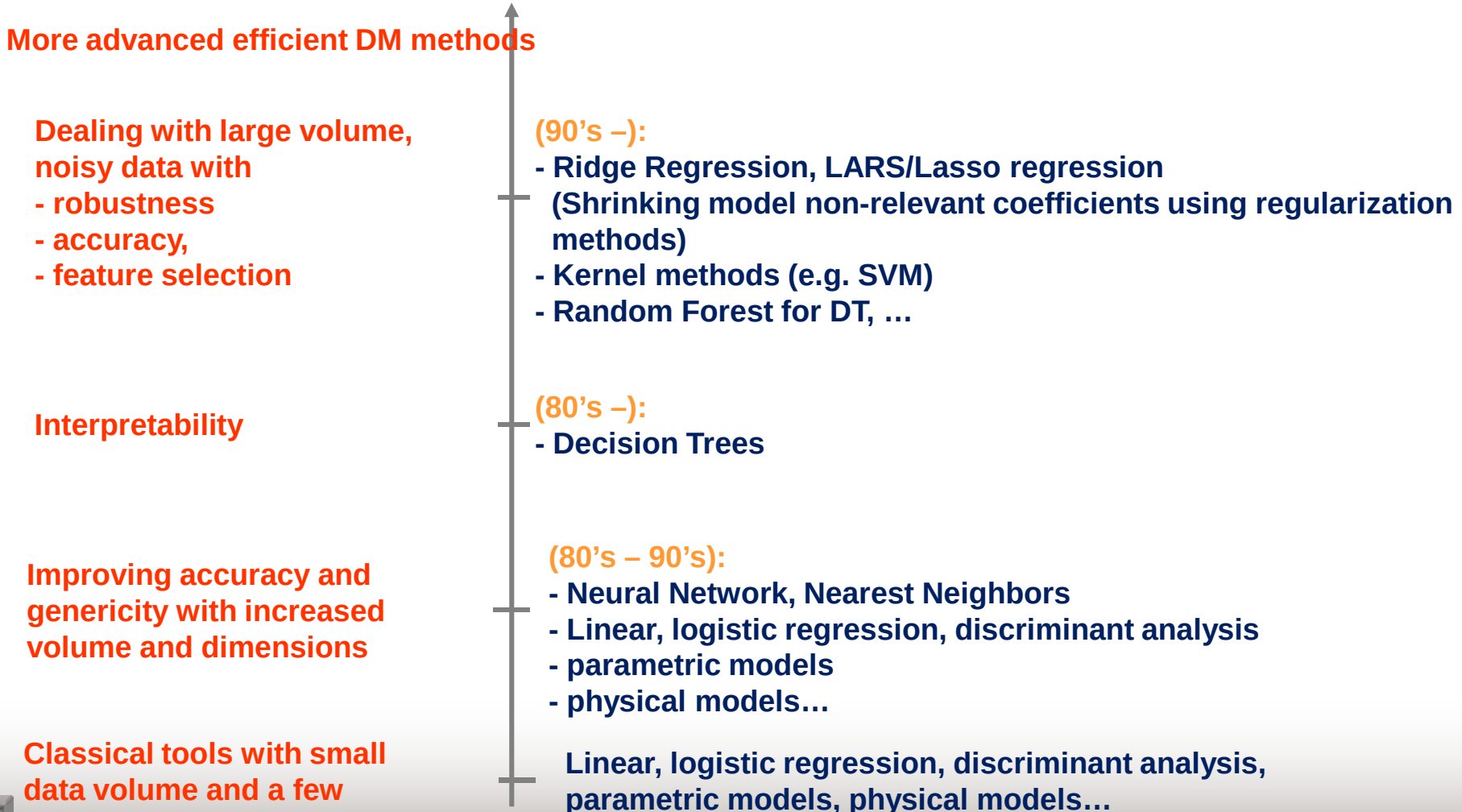


More input variables →

More data points needed to capture the same amount of information, hence to ensure similar modelling accuracy!!

Feature Selection: Identify the right dimensions related to modelling objective

Best in Class Data Mining Challenges: Predictive Modelling Techniques Evolution



Best in Class Data Mining Challenges: Dimensionality reduction and better profiling

Practice I: (Computational) Dimension Reduction

From pure feature selection

Through cross-validation

To more modelling combined feature selection

Degree of automation & computation speed

Feature selection:

- Correlation analysis
- Principal Componen Analysis (PCA)
- Fast Correlation-Based Selection (FCBF)

Wrapper approach

- Forward selection
- Backward selection

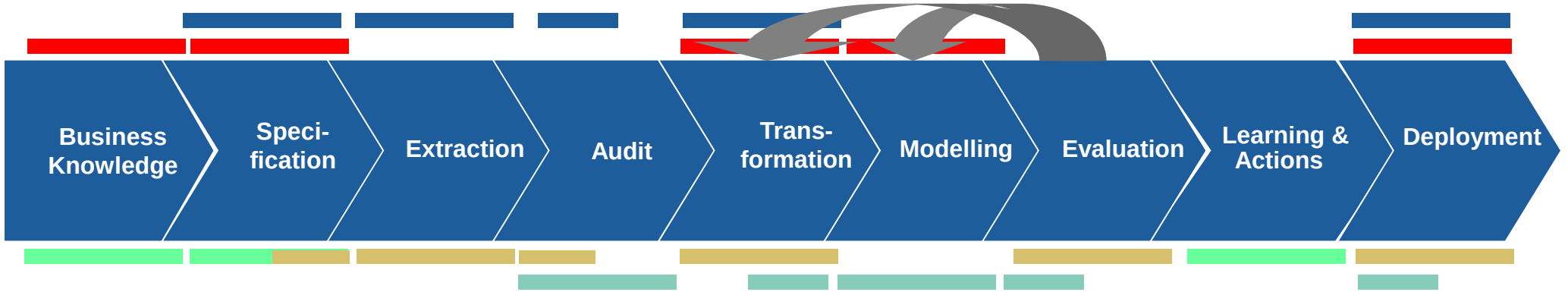
Embedded approach

- LARS/Lasso regression norm L_1
- Ridge Regression for variability penality
- SVM variants

Practice II: Combining Descriptive with predictive

Methodology

Methodology & VADIS tools



Where is time spent?

Data exploration & preparation

Processes (workflow)

Where are the risks?

People

Business experience
- **Objectives**
- **How to achieve**
- **Project mgt**

RANK

Modeling experience
- **Overfitting**
- **Variance**
- **Sampling**
- **Recoding**
- **Variable selection**

Data transformation and mgt systems

Data Workflow maintenance
- **Fast & Reliable update**
- **Documentation**
- **Transferability**

Pourquoi RANK?

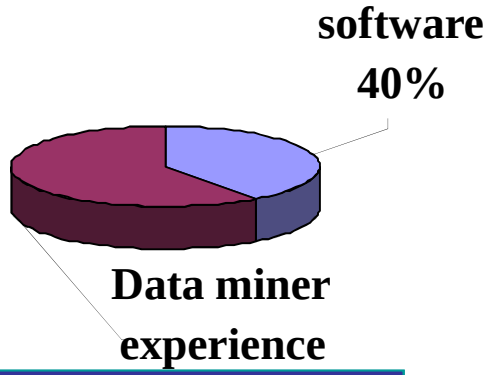
TABLE 2: KDD-CUP-98 Summary of Evaluation Results: Total Profits for Records with Predicted Donation > $0.68

Participant	N*	MIN	MEAN	STD	MAX	SUM**
#1	56,330	-$0.68	$0.26	$5.57	$499.32	$14,712
#2	55,838	-$0.68	$0.26	$5.64	$499.32	$14,662
#3	57,836	-$0.68	$0.24	$5.66	$499.32	$13,954
#4	55,650	-$0.68	$0.25	$5.61	$499.32	$13,825
#5	51,906	-$0.68	$0.27	$5.69	$499.32	$13,794
#6	55,830	-$0.68	$0.24	$5.63	$499.32	$13,598
#7	60,901	-$0.68	$0.21	$5.43	$499.32	$13,040
#8	48,304	-$0.68	$0.25	$5.83	$499.32	$12,298
#9	56,144	-$0.68	$0.20	$5.32	$499.32	$11,423
#10						
#11						
#12						
#13						
#14						
#15						
#16	79,294	-$0.68	$0.12	$4.47	$249.32	$9,464
#17	51,477	-$0.68	$0.11	$4.00	$111.32	$5,683
#18	30,539	-$0.68	$0.18	$5.34	$499.32	$5,484
#19	50,475	-$0.68	$0.04	$3.44	$99.32	$1,925
#20	42,270	-$0.68	$0.04	$3.64	$99.32	$1,706
#21	1,551	-$0.68	-$0.03	$3.60	$53.32	-$54

* N is the number of for which the predicted donation amount> $0.68

** SUM=sum of (Actual Donation-$0.68) for all records with predicted donation > $0.68

The KDD-CUP reflects two aspects:

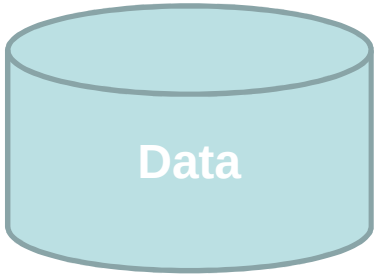


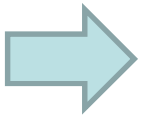
RANK was developed to avoid the “bad luck” effect in modeling. It provides many embedded academic and consulting best practices ensuring optimal performance for any user

- Variable recoding
- Overfitting avoidance
- Model design

Methodology

ToolBox Approach

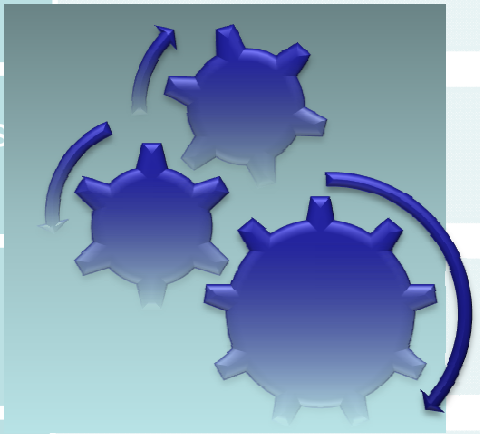




Clustering
- K-means
- Hierarchical clustering

Binary Classification
- SVM
- Decision tree
- Logistic Regression
- …

Recoding
- Normalization
- Binning



Continuous Prediction

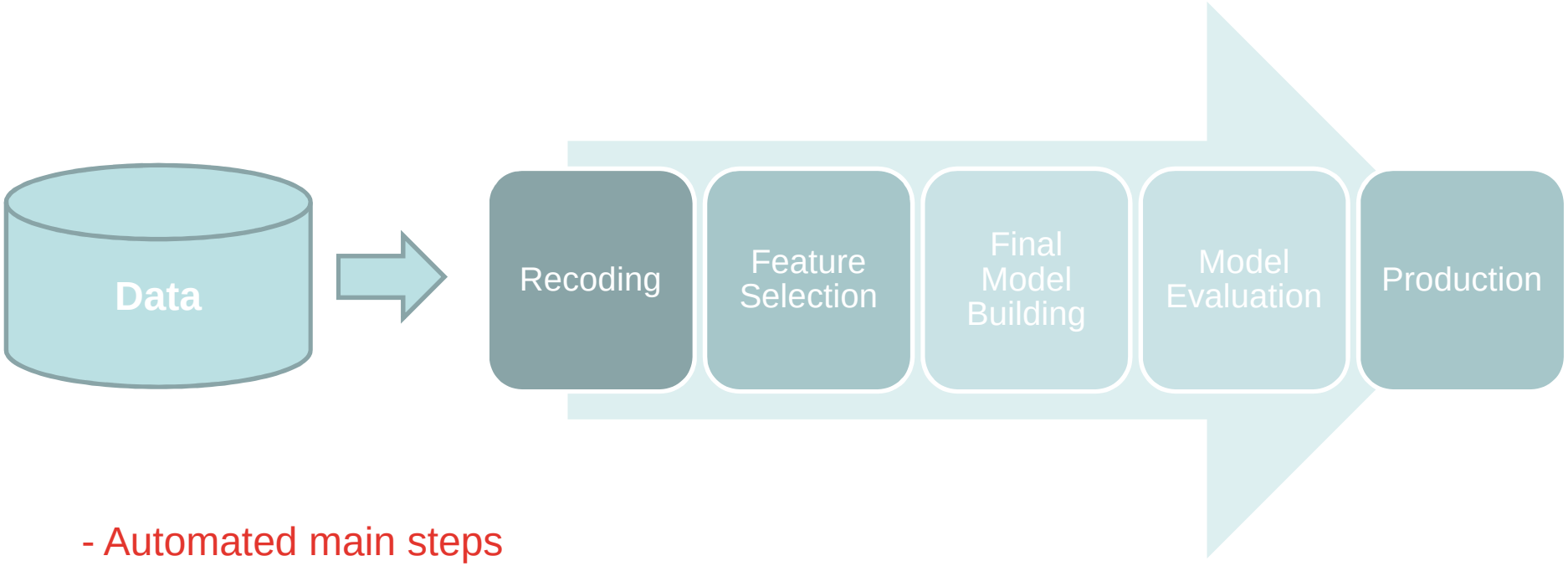
Missing Treatment

Feature Selection
- SVM
- Decision tree
- Logistic Regression
- …

Evaluation
- Graphics
- Test sets
- Statistics
- …

Methodology

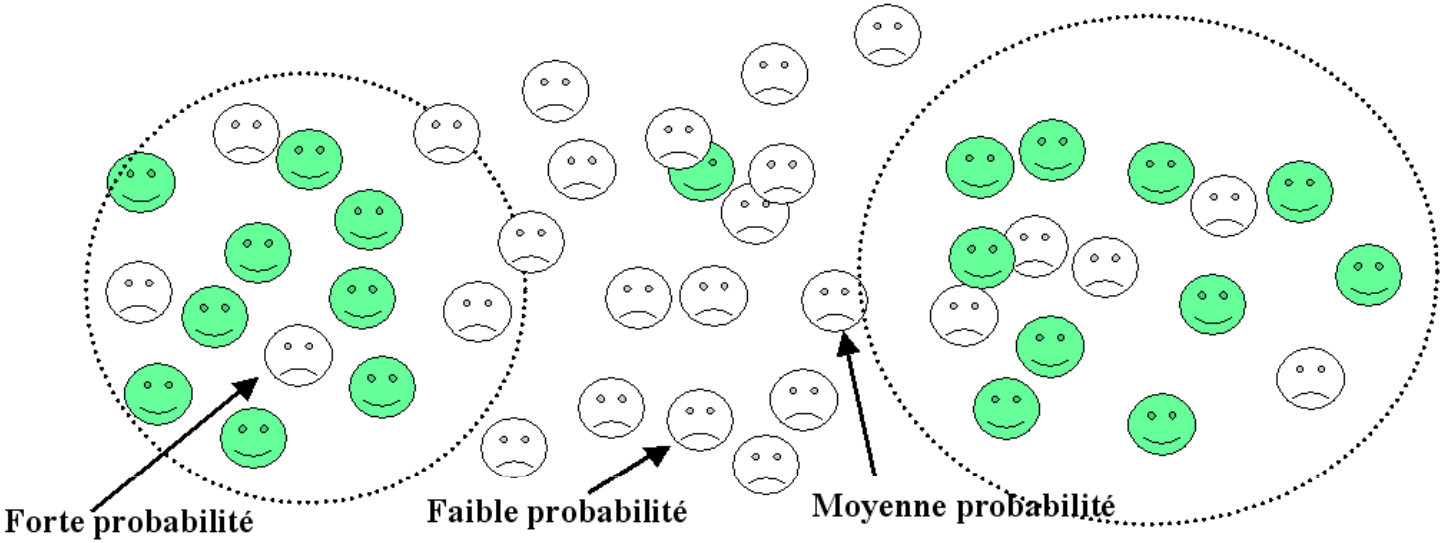
Rank Approach:



- Automated main steps
- Iterations before the best and robust model is chosen
- Focus on results analysis, no implementation, no tedious trial & errors…
- Results easily repeated by non experienced analysts

- Easy production
- Drift control

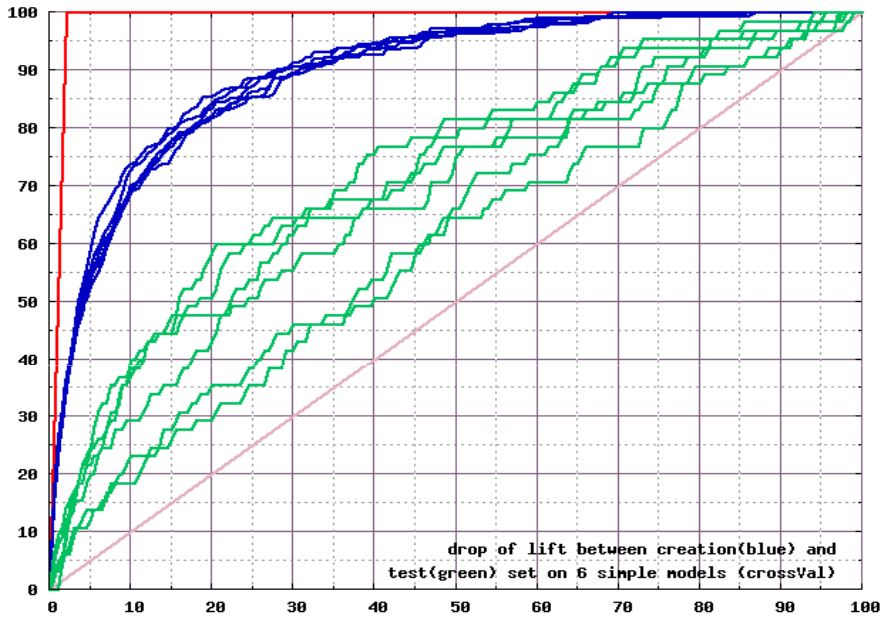
The task

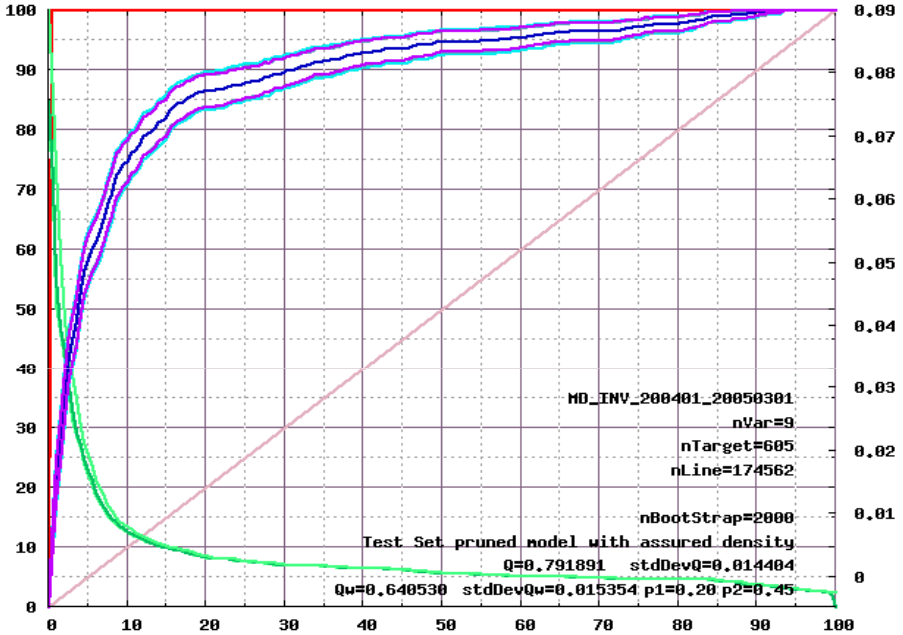


<u>Tasks</u>:
Find a number of typical profiles of green's that allow to recognize them
Use profile proximity for computing probability of being a green…

<u>Problem</u>:
Profile depends a lot of the variables used: how to find the correct ones?
What makes a real (in a statitiscal sense) difference?

Major traps: overfitting & variance





Overfitting has the effect of reducing heavily the capacity of a model to correctly predict on new cases

- You think it will be good and it will not be...

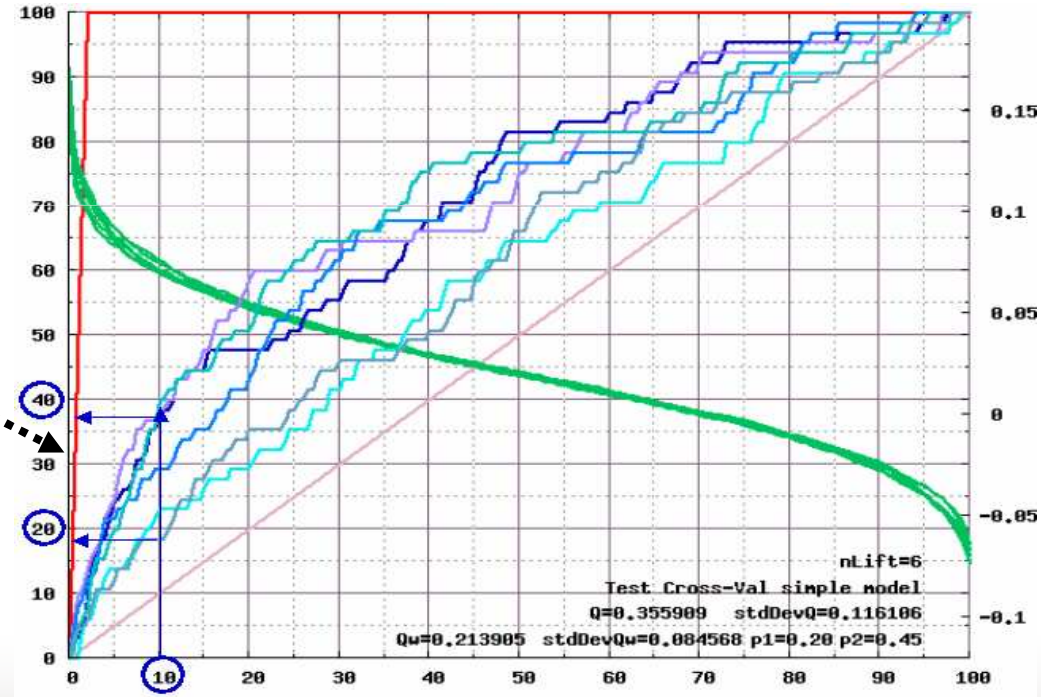
Variance has the effect of reducing the accuracy of the predicted performance

Research & Development - Methodologies

Automated model quality control: Reduce variance

The variance is an alternative way to show the instability of a model, often due to the lack of relevant information in the noisy data. When the variance increases, it means that the model ' hesitates ' between several options, with no information in the data in order to make a strong decision. The following graph shows such a phenomenon.

There is as much chance to have a lift 2 as a lift 4 at 10% selection.



PAKDD 2007
Nanjing, China, 22-25 May 2007

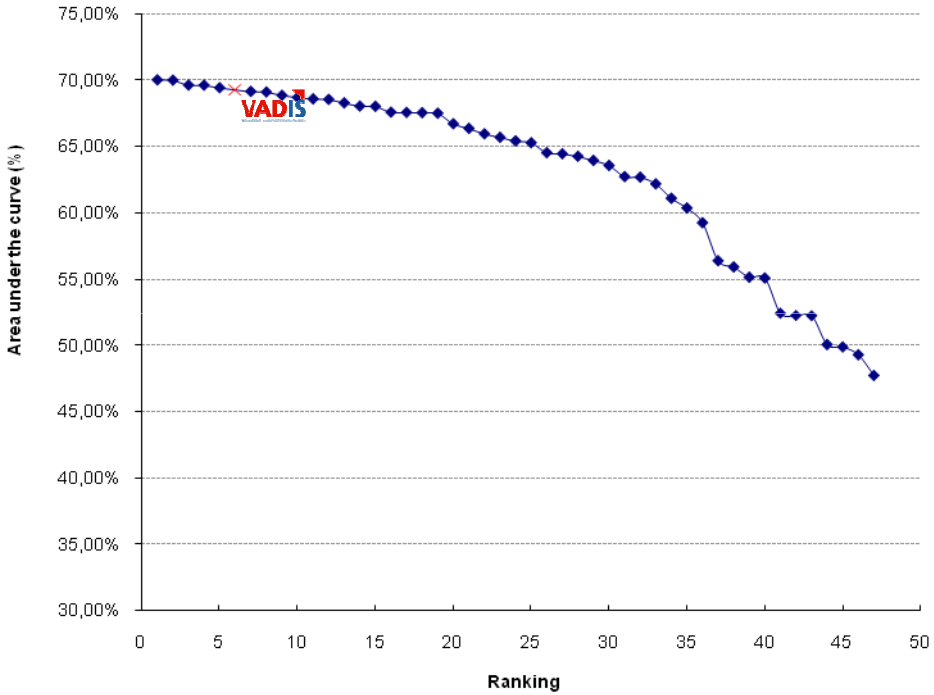
Contest:

- Pacific-Asia conference on Knowledge Discovery and Data Mining
- 250 inscriptions; 47 submissions

Problem

- Cross-sell home loans to credit card customers
- Number of records: 40,700

  Number of vars: 40
- Difficulty: Few targets



Time invested: 2 days

Results: 6th

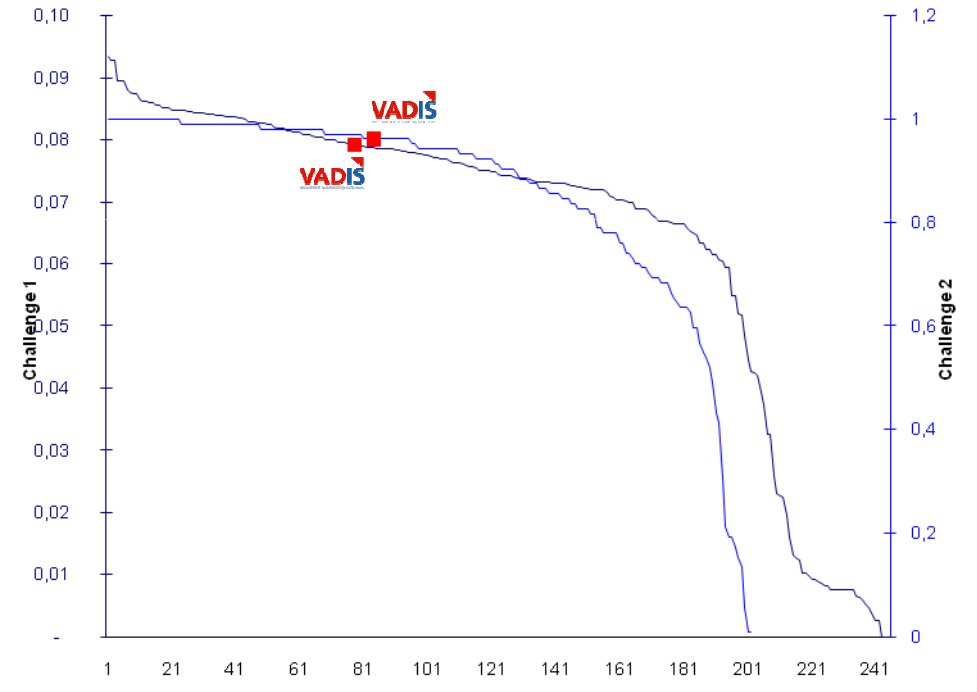
Link: http://lamda.nju.edu.cn/conf/pakdd07/dmc07/finalwinner.htm

Contest:

- First, oldest and most important data mining competition

Problem

- Breast cancer detection (2 challenges)
  - Max recall in a clinical relevant region
  - 100% Sensitivity
- Difficulty:
  - Capture interactions
  - High investment of competitors



Time invested: 19 days

Results:

- 1st challenge: 48th
- 2nd challenge: 53th

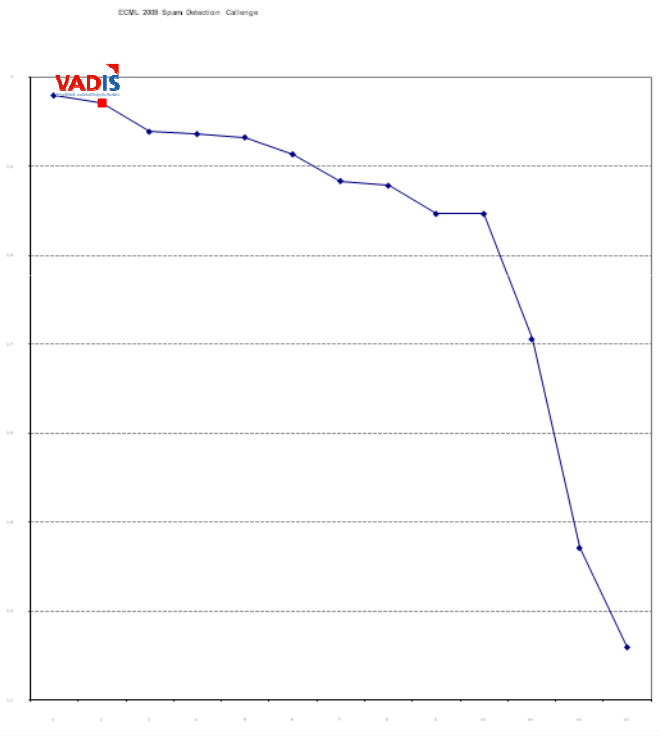
Link: http://www.kddcup2008.com/KDDsite/Challenges.htm

Contest:

- The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- 150 participants, only 13 submissions

Problem

- Spam detection in social bookmarking
- Difficulty:
  - Build variables from raw files (31,715 records and 1,600 variables)
  - Low number of non target



Time invested: 10 days

Results: 2nd

Link: http://www.kde.cs.uni-kassel.de/ws/rsdc08/

Challenge

- Large Set: 50,000 observations, 15,000 variables (260 categorical)
- Small Set: 50,000 observations, 230 variables (40 categorical)
- Targets
  - Churn: 3,672 (7,34%)
  - Appetency: 890 (1,78%)
  - Up-Sell: 3,682 (7,36%)

Timing

March10	Start of the FAST large challenge. Data tables without target values made available for the large dataset. Toy training target values made available for practice purpose.
April 6	Training target values available for the large dataset (churn, appetency, and up-selling). Feed-back: results on 10% of the test set available on-line when submissions are made.
April 10	Deadline for the FAST large challenge. Data tables and training target values made available for the small dataset.
May 11	Deadline for the SLOW challenge (small and large datasets).

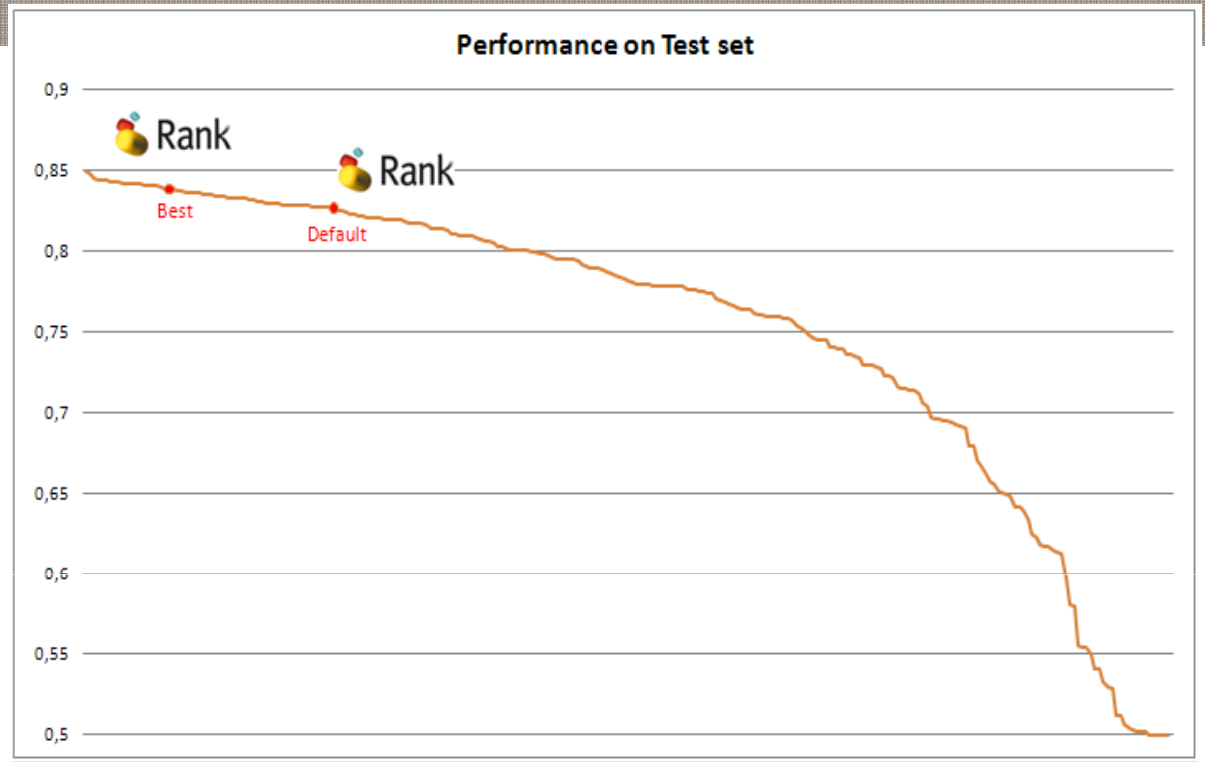
KDD 2009 - Results

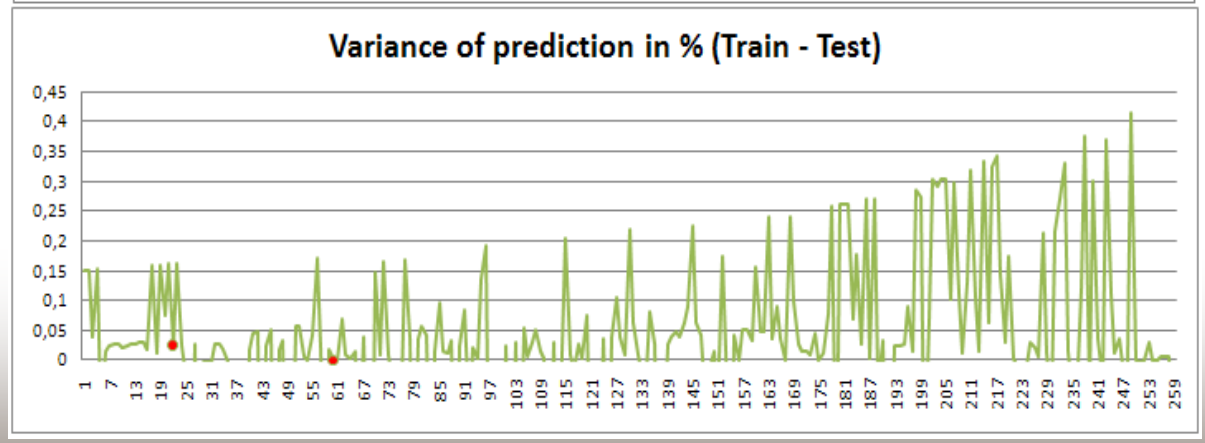
The 1st chart shows the ranking of the models among all participants of the benchmark.

Rank was able to produce results in a few hours after the delivery of the data. The "default" result is the quality of the application of Rank in full automatic mode, using all default parameters values.

The "best" result is after analysis of the data and the setting of best parameters.

The Variance chart reflects the difference between what the expected quality, computed on the training data, and the quality achieved on the test set, computed by the organizers. The variance of Rank is shown by the red balls, showing that Rank indeed is very robust: what you expect is what you'll get.





Link: http://www.kddcup-orange.com/

RANK

RANK Features

Regression

- Least square on linearized space
- Probability mapping

Feature selection:

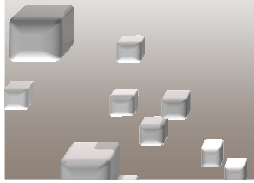
- Least Angle Regression (LARS) with Lasso modification
- Cross-validated Backward

Robustness

- Ridge regression
- Small modalities regrouping
- Missing value treatment

Speed

- Data compressed and stored in RAM
- Automatic sampling



Strong Points Summary

Automatisation → Time spend on analysis not programming

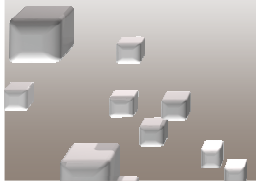
- Variables recoding
- Validation
- Output
  - Results analysis
  - Presentation

Avoid overfitting

No need of data oversampling

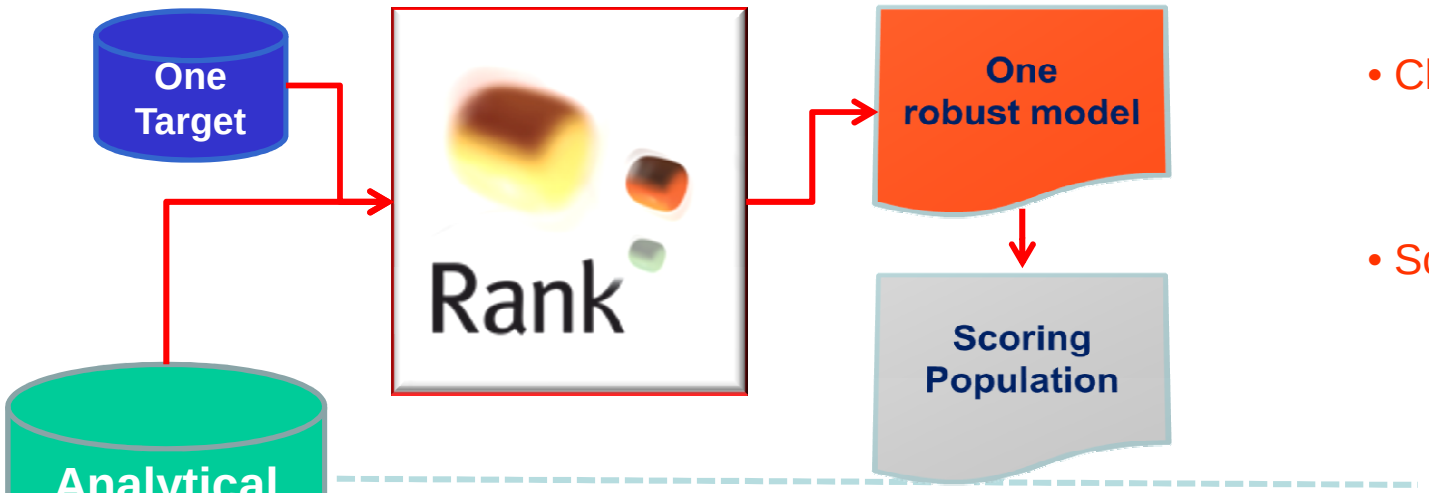
Speed

Large Data



Production framework: Massive Profiling Platform

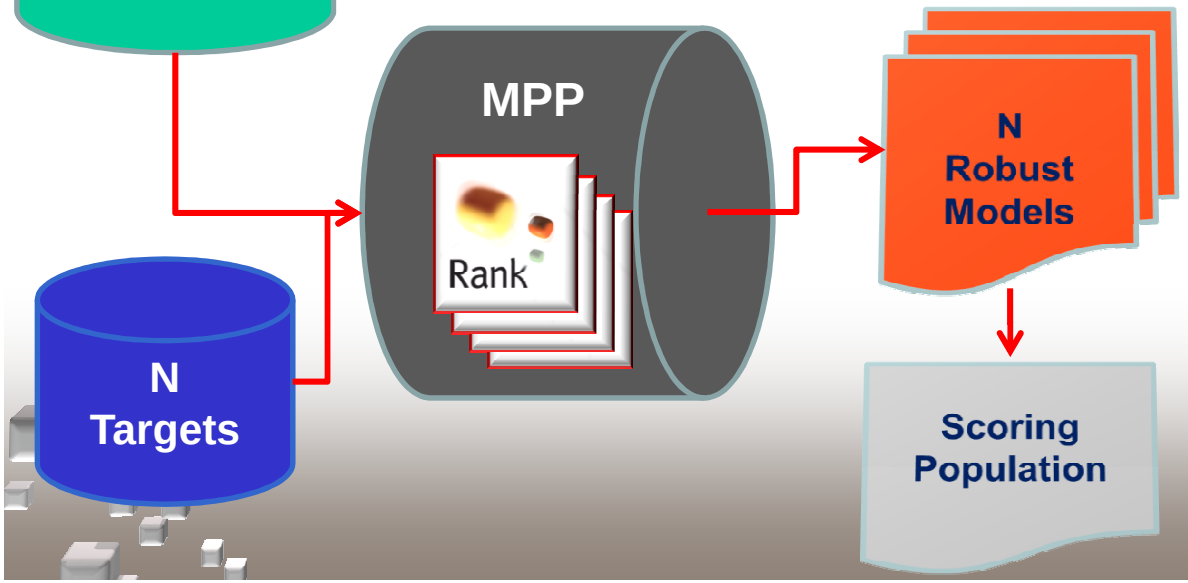
Rank Studio



Functionalities

- Choice of a robust model
  - Parameter setting
  - Feature selection
  - Confidence intervals
- Scoring new population
  - One model scores
  - Drift warning

Massive Production Framework



Functionalities

- Automatic build of N robust models
  - Parameter option set
- Non performance iterations
  - Parameter option re-set
- Automaticly Scoring new population
  - N model scores
  - Drift warning