27th International Congress of Actuaries

ICA CANCUN 2002

A Discussion of Modeling Techniques for Personal Lines Pricing

Cristina Mano, Brazil

Elena Rasa, Italy

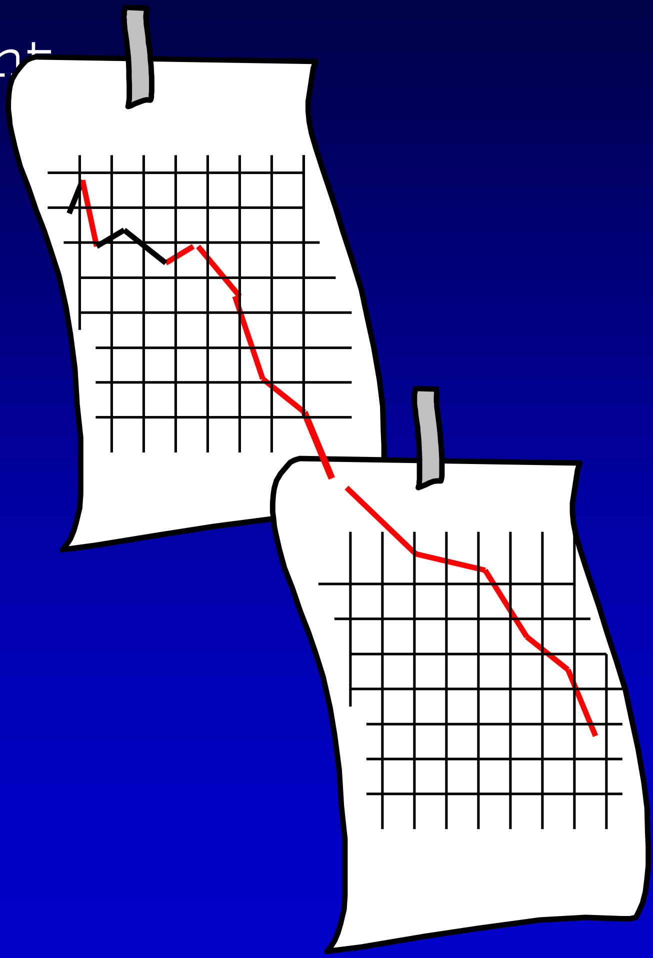
A New Millennium. A New Challenge for Actuaries

Agenda

- Overview of the insurance market
- Pricing approach
- Overview of the methodologies: GLM, Decision Trees and NN
- Comparison among the three techniques
- Conclusions

An overview of the insurance market

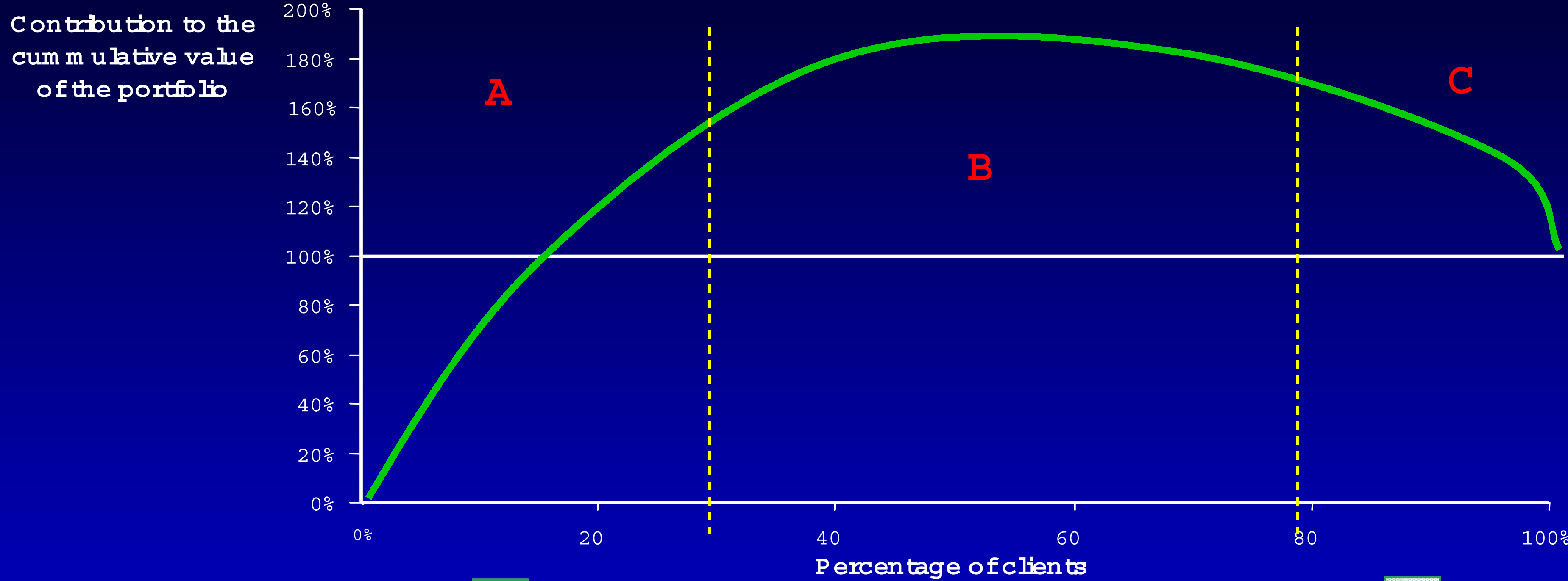
- Presence of inadequate tools for a complete and effective analysis of the client
- Poor knowledge of clients
- Some lines of business are running at a loss in most of the countries (i.e. motor business)
- Low level of sophistication
- Cross-subsidies among clients



Critical areas

- The risk increases with:
  - high exposure on insurance business;
  - ultimate cost not completely recognized ;
  - inadequate selection of underwritten risks ;
  - not effective claim management;
  - low cross-selling level;
  - not enough emphasis on client.
- The necessary elements for reaching the objective "**client**" are:
  - economic cost estimation;
  - lapse probability of the insured and his elasticity of demand.
  - Competitive Market Analysis

The profitability varies with the risk profile



Good clients

- Developing new services
- singling out possible preservation strategies

To actively manage

- Improving the profitability of existing products
- Offering value added services
- Increasing number of products for each client (cross-selling)

"Up or out?"

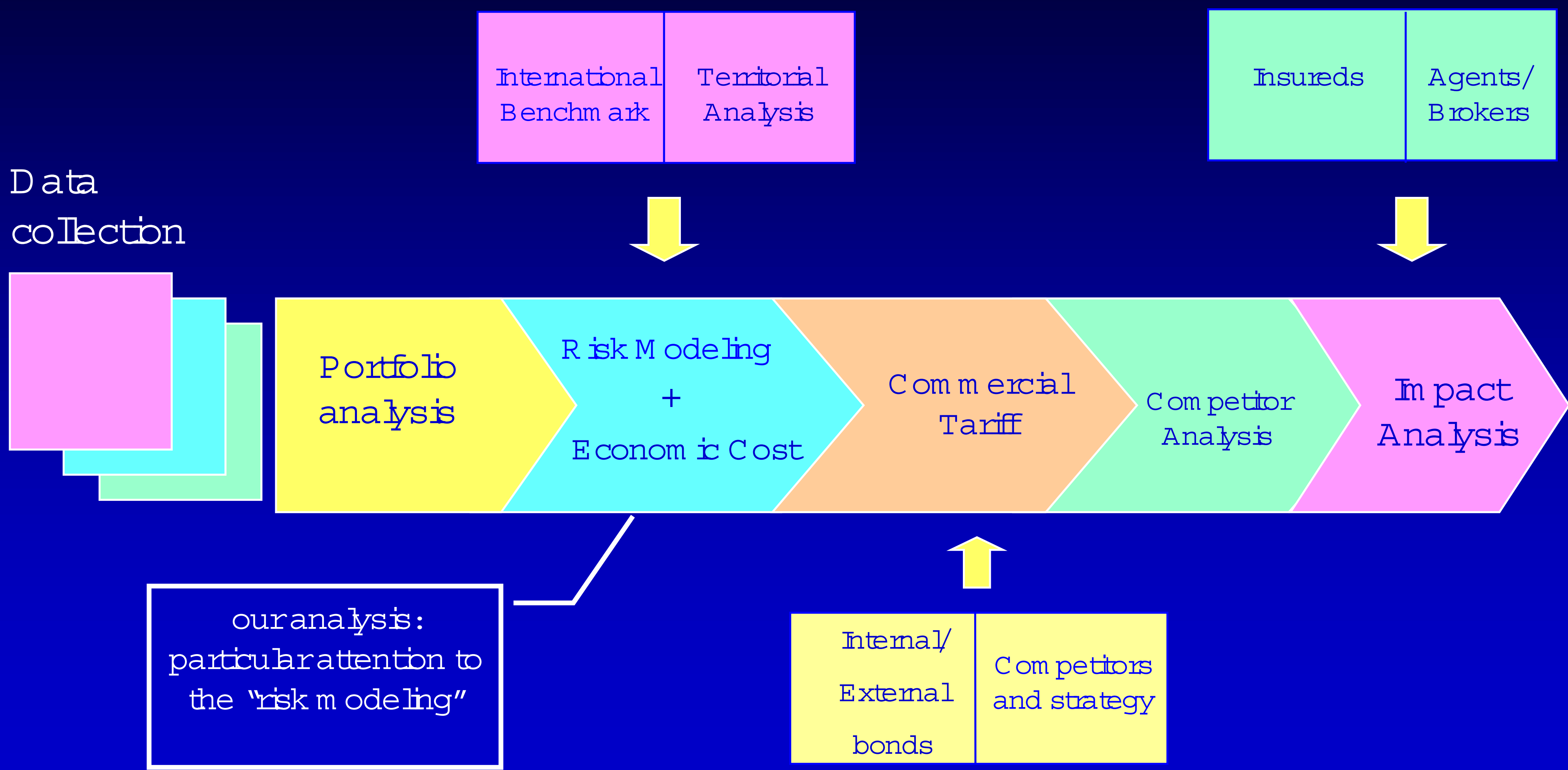
- Is it possible to increase the value?
  - →Decreasing the costs and expenses?
  - →Changing products?
- What clients should we exclude?

Nowadays, too much attention is paid to the cancellation of "bad" risk

… it could be better to find the right strategies so that:

- the premium be correctly calibrated on the risk;
- the economic cost of the client be projected;
- new "ad hoc" products be created;
- the portfolio be segmented in order to define the market niches, which destroy value.

Pricing Strategy



Risk Modelling: Overview of Methodologies

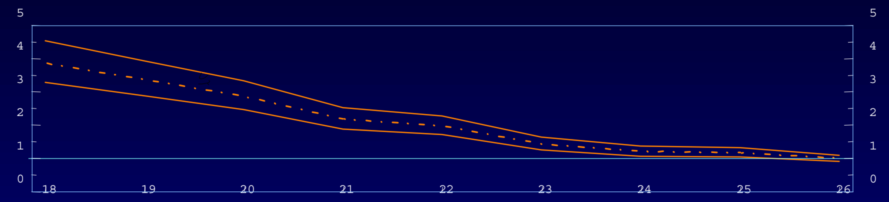
- New multivariate techniques available which are replacing older methodologies;
- These techniques represent advances in:
  - quantifying the TRUE economic cost of writing each policy;
  - measuring the economic impact of adopting any rate plan other than the actuarial model;
  - understanding lapse and renewal experience;
  - market pricing behaviour;
  - in short, managing the business.

Why Multivariate Statistical Techniques?

- Most rating variables are correlated;
- Different variables may be showing the same underlying effect;
- Repeated use of univariate techniques leads to double-counting of the same effects;
- They can capture interactions;
- They provide more than a point estimate and standard errors.

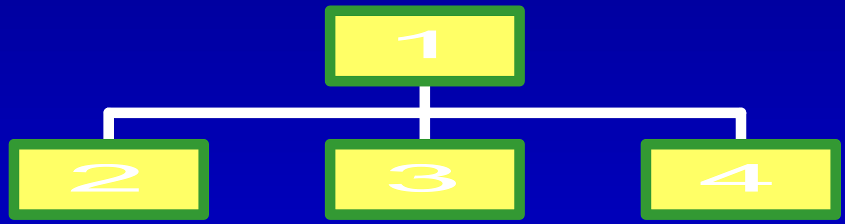
… looking for more suitable techniques

- GLM : Generalized Linear Models

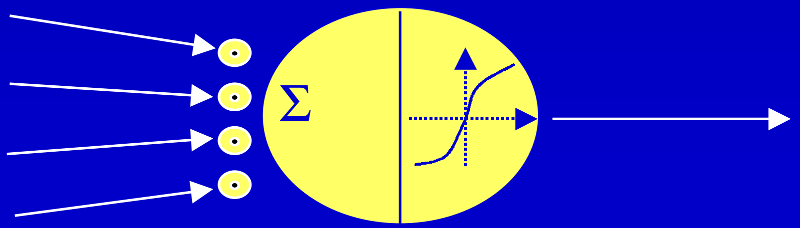


- Geographic Analysis
  - Cluster Analysis
  - GIS

- Decision Trees
  - CHAID
  - CART



- Neural Nets

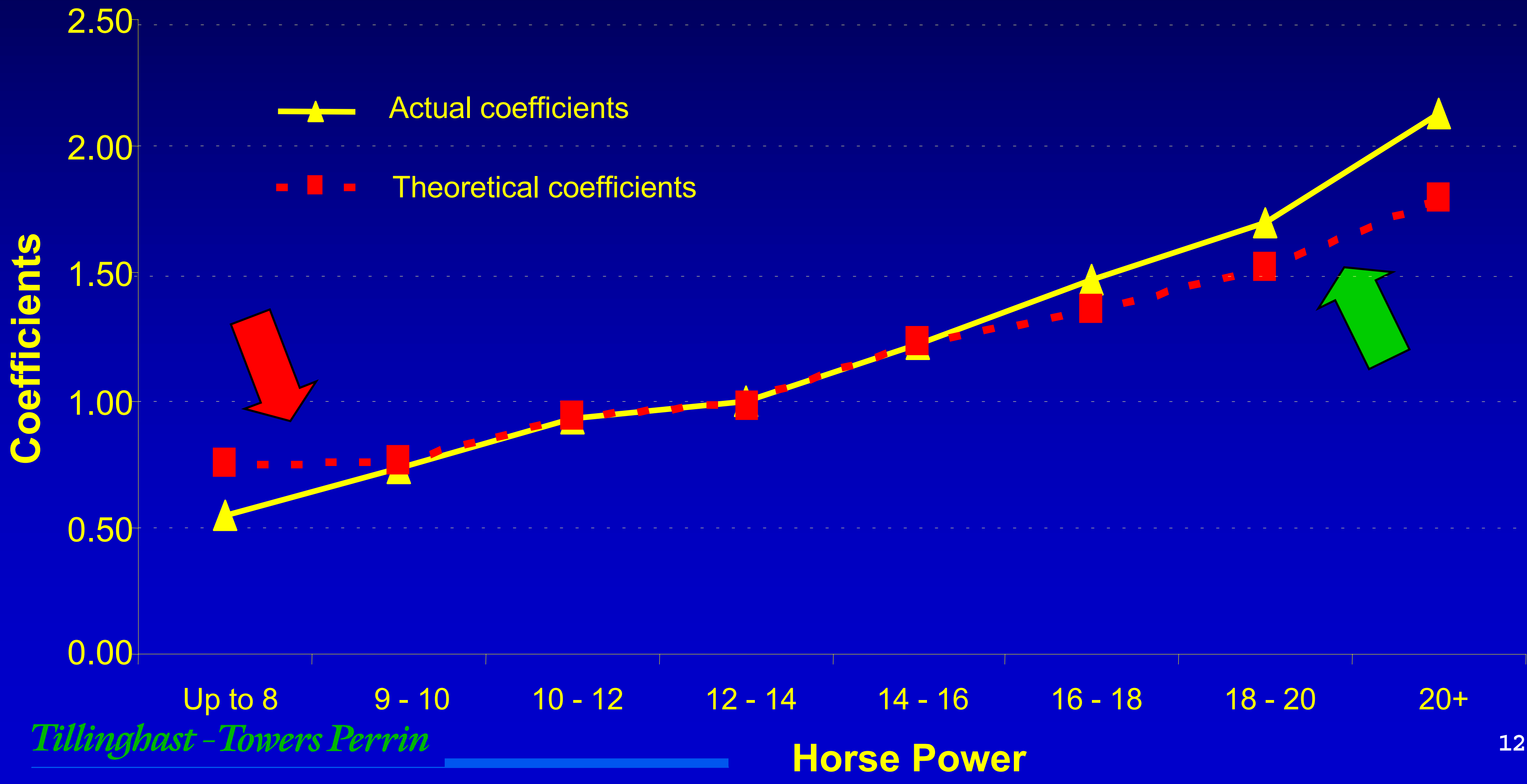


What is GLM?

- It is a statistical procedure for measuring the effect of one or more independent variables upon a dependent variable;
- **Dependent** variable for ratemaking are:
  - frequency
  - severity
  - pure premium
- GLM allows extreme flexibility in model design:
  - multiplicative, additive or mixed plans
  - different error distributions (i.e. Normal, Gamma, etc.)
  - variable interactions (i.e. sex & age)

An example of Generalised Linear Model

This statistical approach allows us to determine the cross-subsides among the clients and to create a theoretical rating structure, which penalizes bad clients and favors good clients

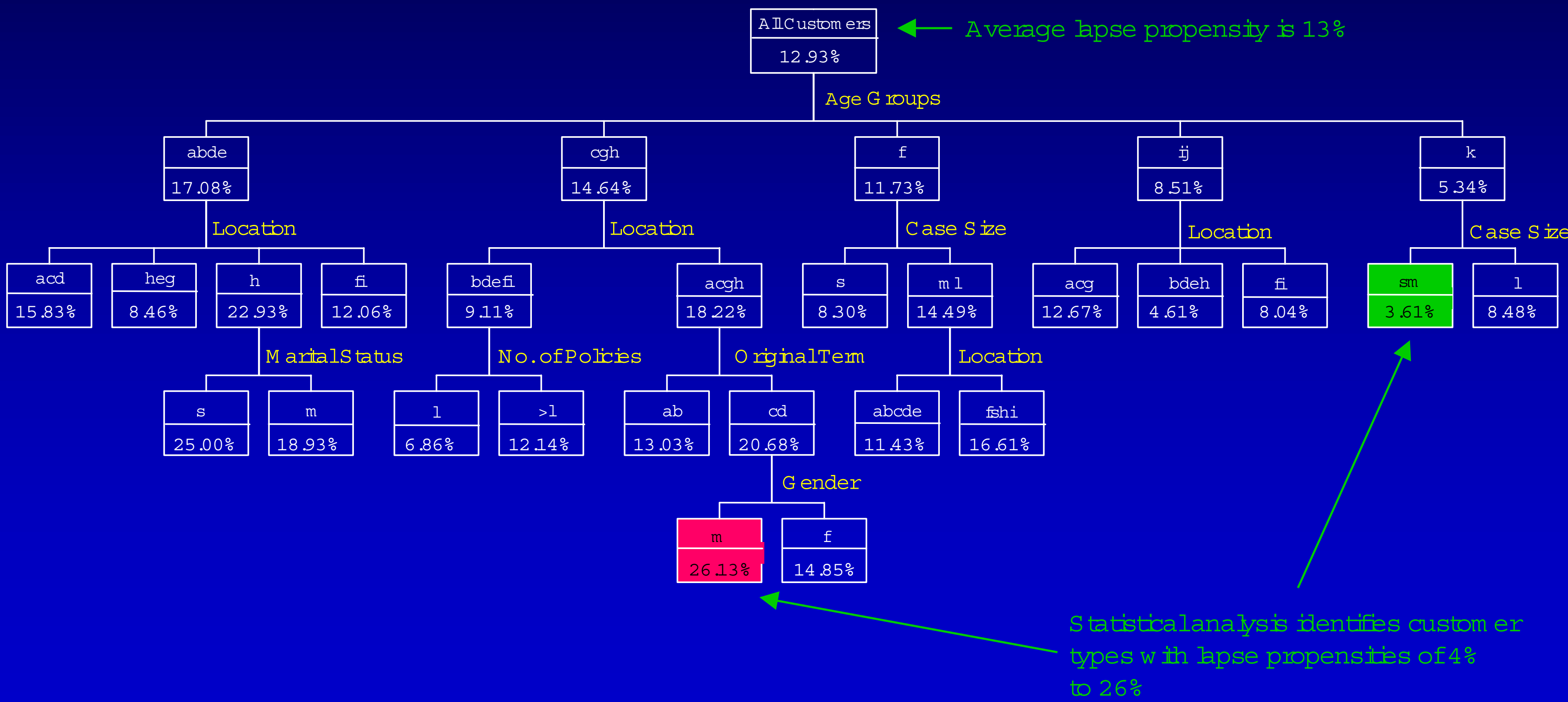


What are decision trees?

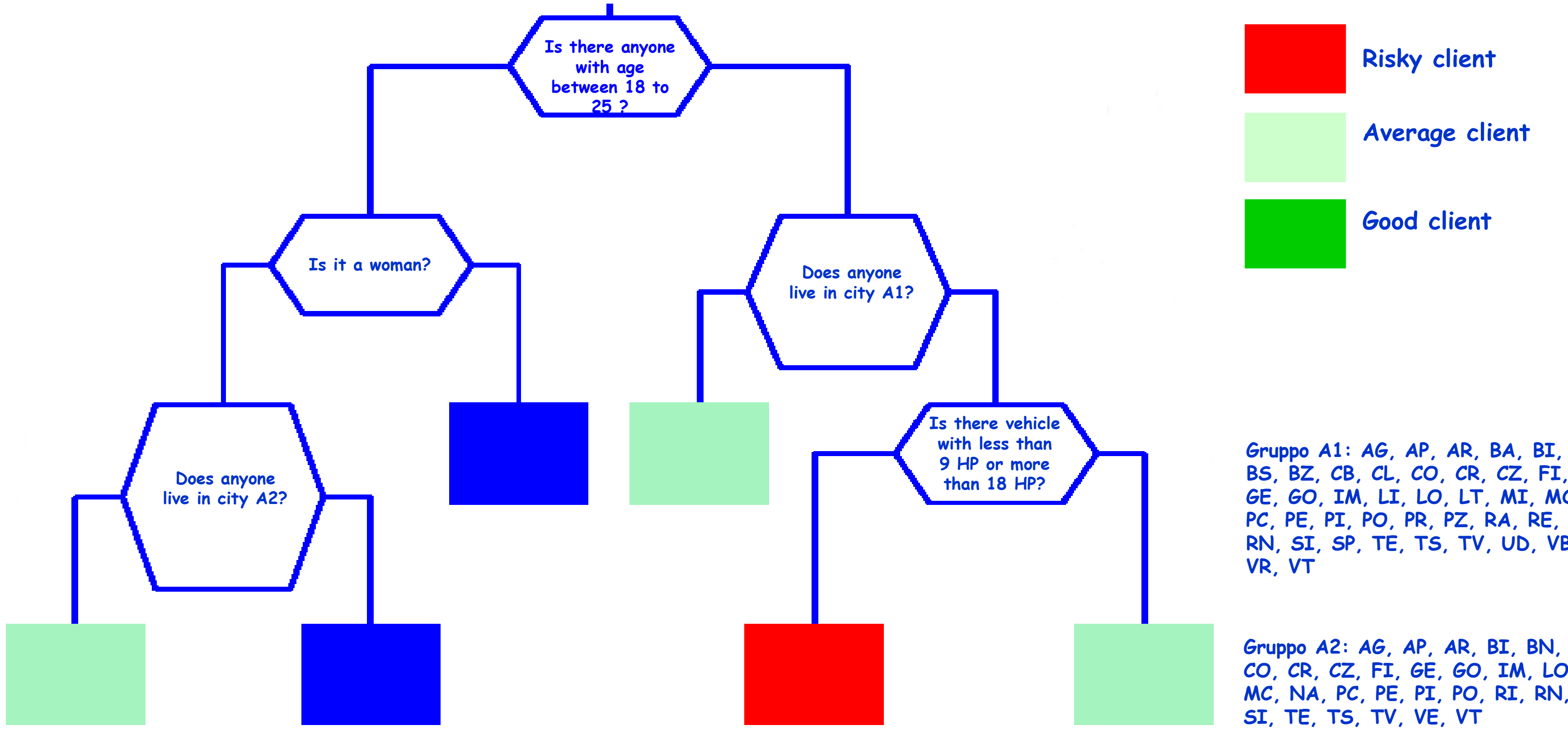
- Procedures for successively subdividing data into homogeneous groups;
- Like GLMs, they use a dependent variable and one or more independent ones;
- Results are not necessarily symmetric;
- Implicitly capture the natural interactions between factors;
- Produces homogeneous groups (i.e., a tree structure), but no rating plan or relativities;
- Possible methodologies, most famous:
  - CHAID: Chi-Square Automated Information Detection
  - CART: Classification and Regression Tree

Decision Trees: 'Divide et impera"

A decision tree is given by a set of decisional rules for predicting a fixed dependent variable (for example, claim severity, frequency or lapse rate)



Decision tree: an example



Gruppo A1: AG, AP, AR, BA, BI, BN, BS, BZ, CB, CL, CO, CR, CZ, FI, FO, GE, GO, IM, LI, LO, LT, MI, MO, PA, PC, PE, PI, PO, PR, PZ, RA, RE, RI, RN, SI, SP, TE, TS, TV, UD, VB, VE, VR, VT

Gruppo A2: AG, AP, AR, BI, BN, BZ, CO, CR, CZ, FI, GE, GO, IM, LO, LT, MC, NA, PC, PE, PI, PO, RI, RN, SA, SI, TE, TS, TV, VE, VT

1) Are you in Bonus classes?

2) Do you live in one of the provinces of zone A

Gruppo A: AG, AP, AR, BA, BI, BN, BS, BZ, CB, CL, CO, CR, CZ, FI, FO, GE, GO, IM, LI, LO, LT, MC, MI, MO, NA, PA, PC, PE, PI, PO, PR, PZ, RA, RE, RI, RN, SA, SI, SP, TE, TS, TV, UD, VB, VE, VR, VT

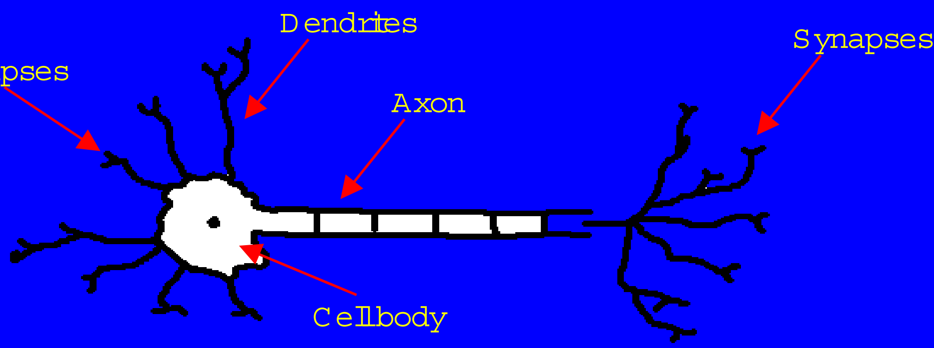
What are Neural Networks?

- **Neural networks**: are non-linear predictive models that learn how to detect a pattern in order to match a particular profile through a training process;
- It's not necessary to split the pure premium into its components:
  - frequency and
  - severity;
- Advantages and disadvantages:

+ Usable even when relationships among variables are unknown
+ Will model non-linearity and interaction well
- The solutions can not be interpreted ("Black box"?)
- Can take tremendous computing power and still not converge to a solution

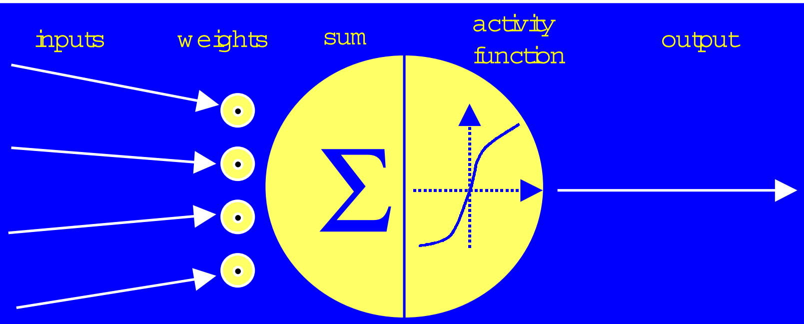
Neural Networks are motivated by a simplified model of biological neurones in the brain

BIOLOGICAL MOTIVATION: PYRAMIDAL CELL

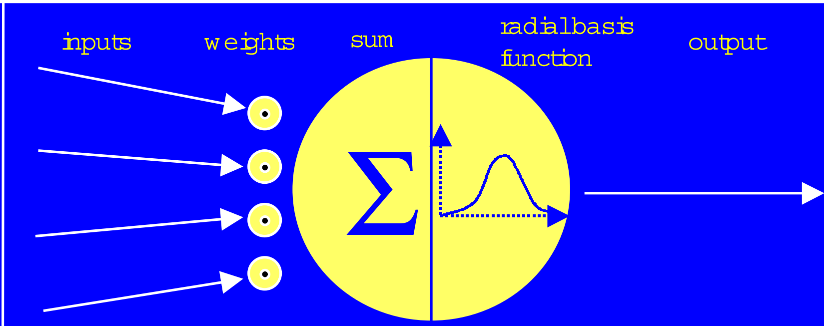


AND THEIR MOST SIMPLIFIED MATHEMATICAL ANALOGY (E.G. TWO IN COMMONLY-USED TYPES OF)

SIGMOID FUNCTION NEURON (-> MLP)

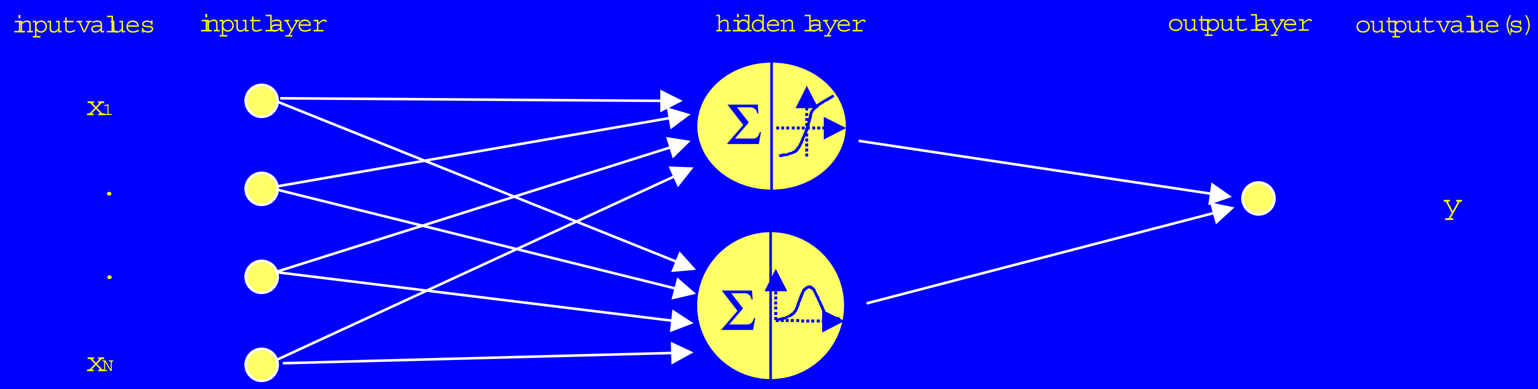


RADIAL BASIS FUNCTION NEURON (-> RBF)



Neural Networks (contd.)

CONNECTED TO A NETWORK



BUILD AN ARTIFICIAL NEURAL NETWORK

Our analysis...

- A simulated portfolio was created using the distributions of the Italian market (public data - 1999) relative to the main rating parameters:

  - Bonus/Malus
  - Horsepower
  - Fuel
  - Sex/Age
  - Territorial zones
  - Claims limit
  - Number of installments

What about the dependent variables?

- Claim frequency
- Severity

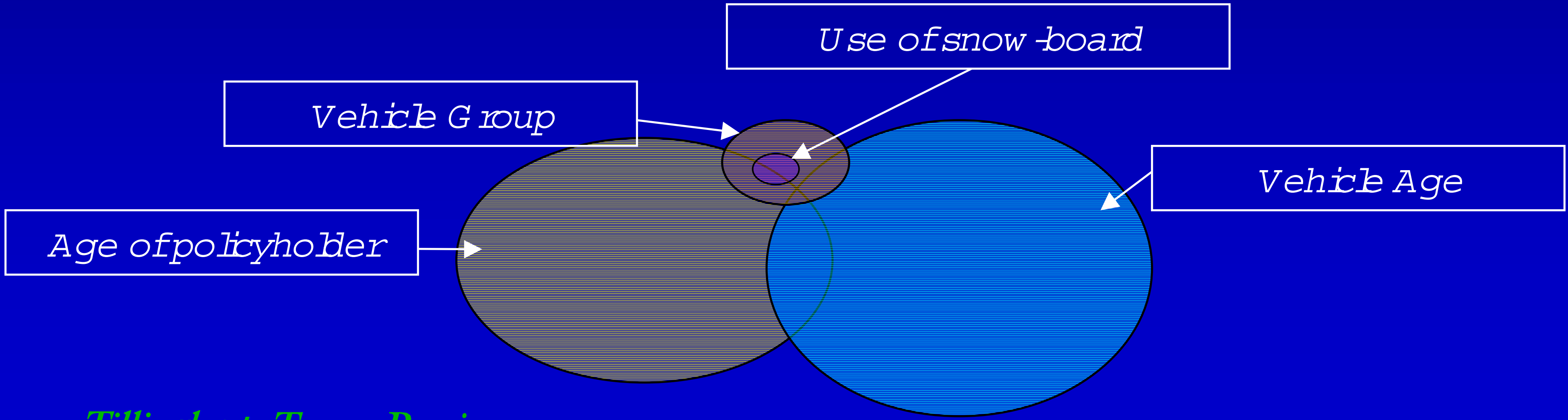
Parametric analysis

- Pure premium

Non-parametric analysis

Choosing the rating parameters

- The selection of inappropriate variables can spoil the final result;
- Including a variable, which does not contribute in any way to the final result, could have the effect of diminishing the model performance;
- This danger is very high in NN, moderate in GLM, totally indifferent in CART/CHAID.



"Over-fitting" or "over-parameterization" danger

- The "over-fitting" (or "over-parameterisation") concept is always there, whatever the statistical method used;
- Using a too many variables in the estimation process, may lead to:
  - memorising also the idiosyncrasy of the training set (in the language of neural nets),
  - incomplete separation of the stochastic part from the deterministic one (in the parametric language);
- In GLM there are specific statistics that report the "over-parameterisation" phenomenon;
- In decision trees analysis, it is necessary to reach maximum depth (challenging the "over-fitting" danger), in order to proceed to the pruning step until the best tree is defined.

Managing missing values

- There are different methods for the missing values management:
  - dropping the record with a missing value - GLM,
  - substituting the missing value with characteristic or typical values (average, quantiles, closest neighbor,…) - CART, NN,
  - estimating the missing value, after having assigned a fixed level ('Errors', 9999) - manipulating the data,
  - building separate models for each set of missing values - manipulating the data,
  - using the non-coded values in the learning phase of the network - NN;
- CART and NN are very robust methods in managing the missing values. The correct way to proceed is not clear…

Managing anomalous values or outliers

- In order to identify every anomalous value, it is a good idea to start with a "data mining" phase using the following:
  - residuals plots,
  - over-dispersion analysis,
  - Cook and Leverage statistics;
- Once the outliers have been identified, it is possible:
  - to assign a low marginal probability,
  - to drop the observation from the data set,
  - to confine them in a separate class and to make an estimate in a successive phase;
- Non-parametric techniques are more robust then the parametric ones in managing the anomalous values and outliers.

Is a Neural Net a "Black box"?

- If such a term refers to a presentational or a synthesis of the results problems, the answer is YES;
- if it refers to the arithmetic of the algorithm, the answer is NO or, at least, not more then other statistical techniques, including GLM and Decision trees;
- In fact, working with a Neural Net:
  - it is difficult to know which are the important variables to be included in the model and how they interact among themselves; and,
  - there is no structure of coefficients (relativities), as it is for the parametric regression and there is not a final model;

Black box = low synthesis, low presentational power

Computing time

- A few years ago, neural networks were used almost exclusively for "pattern recognition" problems, mainly due to the long computing time required;
- A similar problem was true also for decision trees methods. CHAID, in particular, based on a contingency table where a CHI-squared is performed on each cell, could be very heavy from a computing point of view;
- With the advent of new and more powerful computers, even these techniques can be used in the solution of "everyday" problems.

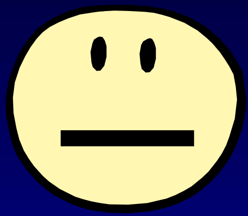
Reading and interpreting the results

- **GLM**
  - The results are directly comparable with the rating coefficients applied by the insurance companies. The reading is easy for specialists;
- **CART/CHAID**
  - It is a method that communicates through images. The results are always in the form of an upside-down tree. The reading is very easy also for lay people;
- **NN**
  - It gives an estimate close to the real observation of the data base in the training and testing set. Once the training set has been memorized, the network can be used in another sample where the observation is missing. The reading is not very easy even by specialists.

Implementation

- **GLM**
  - Fast and easy. The coefficient structure reproduces the rating structure of the company and it is directly comparable and easy to implement;
- **CART/CHAID**
  - The result is very easy and readable. It is composed by a limited number of nodes and to each of them an average premium is associated. The question is, is it acceptable to have a rating structure consisting of 49-50 profiles?
- **NN**
  - It is very useful as discriminate analysis, but it needs a great deal of modification in order to be implemented.

A comparison among the three techniques: a final report card



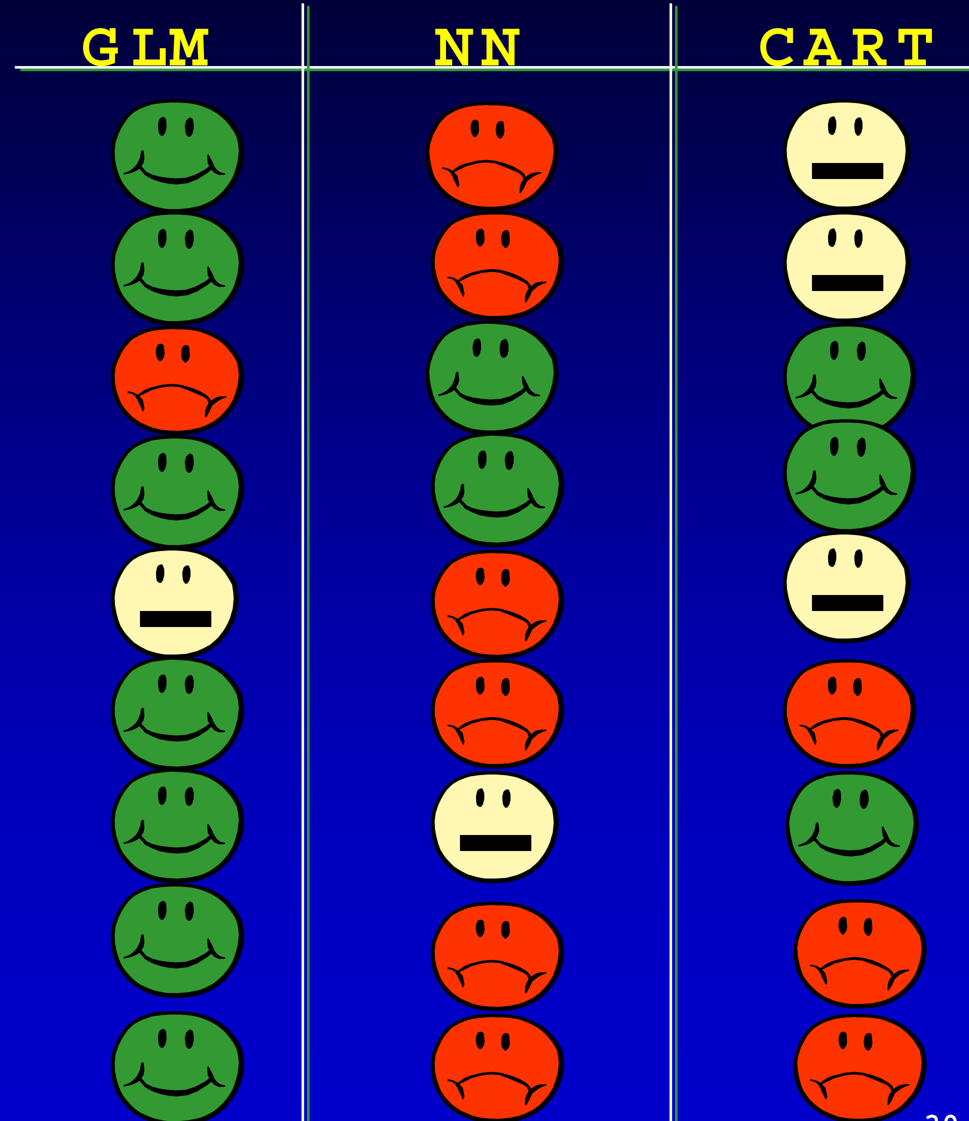
Not applicable/indifferent

Possible but onerous/negative answer

Good results/positive answer

A comparison among the three techniques: a final report card

- Selection of rating parameters
- "Over-fitting" danger
- Missing values management
- Outlier values management
- Black box
- Computing time
- Reading and understanding results
- Implementation
- Overall assessment



Pearls of wisdom ...

- It's important to understand the ideas behind the various techniques, in order to know how and when to use them;
- It's important to accurately assess the performance of a method, to know how well it can be expected to work (... simple methods often perform as well as complex ones!);
- in data mining, understanding the system used is not always a crucial problem. A neural network that produces optimal estimates can be preferable to easier but less efficient models;
- This is an exciting research area, that has applications in science, industry, finance, etc.