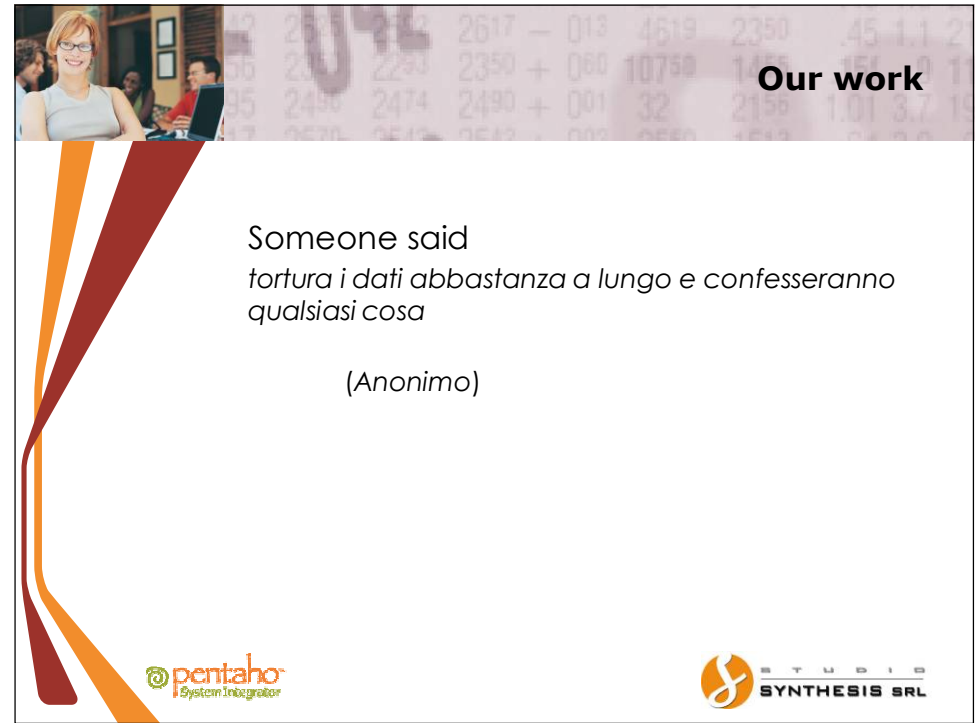




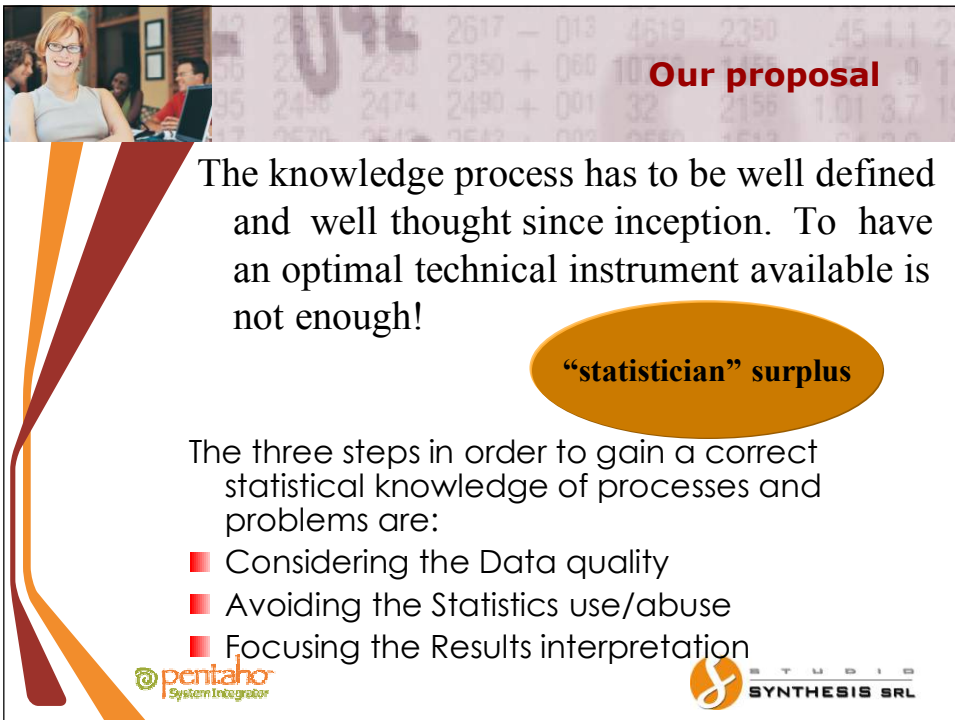
MAINZ 2008 STATISTICIANS AND BI



Our work

Someone said
*tortura i dati abbastanza a lungo e confesseranno
qualsiasi cosa*

(Anonimo)



Our proposal

The knowledge process has to be well defined and well thought since inception. To have an optimal technical instrument available is not enough!

“statistician” surplus

The three steps in order to gain a correct statistical knowledge of processes and problems are:

- Considering the Data quality
- Avoiding the Statistics use/abuse
- Focusing the Results interpretation



Our proposal

1. The Data Quality

Different considerations have to be done in order to define the concept of *data* quality:

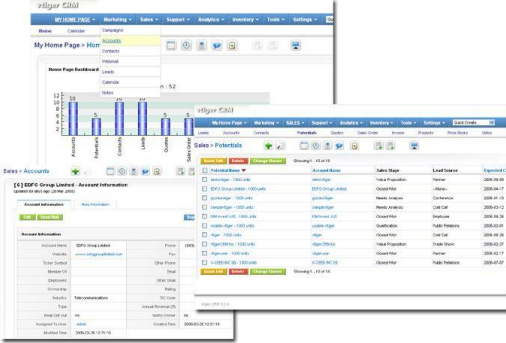
- a) Data and information making up
- b) Data validation



Our proposal **1. The Data Quality**

a. Data and information making up

- CRM
- ✗ Vtiger : modular system



pentaho System Integrator **STUDIO SYNTHESIS SRL**

Our proposal **1. The Data Quality**

a. Data and information making up

- Questionnaires:
- ✗ Collection of subjective perceptions
- ✗ scale validation
- ✗ Indicators and measures construction

Cluster analysis

Regression analysis

Rasch Analysis

PCA/CatPCA

pentaho System Integrator **STUDIO SYNTHESIS SRL**

Our proposal **1. The Data Quality**

b. Data Validation:

- To find problems in the data
- missing data:
- ✗ If a pattern is found they become informative as the data themselves
- ✗ How it is possible work with missing data in the analysis step?

Outliers Detection

Missing values analysis

Listwise deletion vs data imputation

pentaho System Integrator **STUDIO SYNTHESIS SRL**

Our proposal **2. Statistics Use/Abuse**

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write

(H.G. Wells)

Data mining processes are able to select the best model in a set of alternative models. It is not enough! We need to deeply know and understand the methodological aspects of the statistical technique in use.

Which are the statistical techniques suitable for the analysis?

- The variables have to be clearly defined
- The central object of the analysis has to be clearly defined
- Evaluation of the methodological hypothesis of the statistical tools
- Forecasting based on simulated scenarios

pentaho System Integrator **STUDIO SYNTHESIS SRL**

Our proposal **2. Statistics Use/Abuse**



■ **The variables have to be clearly defined**

For example:

Different kinds of variables → Different treatment



Plots and Graphics, descriptive statistics, inferential models

- ✗ Discrete or continuous
- ✗ multivariate or univariate
- ✗ latent or observed
- ✗ Aggregated or simple
- ✗ ...

Our proposal **2. Statistics Use/Abuse**

- ✗ multivariate or bi-univariate variable **High correlation?**
- ✗ continuous or discrete variable **Linear regression vs non linear**
- ✗ latent or observed variable **Structural Equation Models?**
- ✗ aggregated or simple variable **Multicollinearity?**
- ...



Our proposal **2. Statistics Use/Abuse**

■ **The variables have to be clearly defined**

- multivariate or bi-univariate variable **High correlation?**

Obviously, increasing the dimension of a variable (n -dimensional) the analysis becomes more complicated (think about the graphical representation, for example). It could be easier (and sometimes more correct) to reduce the variable's dimension using an appropriate technique (Principal Component Analysis, Categorical Principal Component Analysis, Correspondence Analysis, ...).

It happens when the n variables present an high correlation degree.



 

Our proposal **2. Statistics Use/Abuse**

■ **The variables have to be clearly defined**

- continuous or discrete variable **linear regression vs non linear (probit, logit,...)**

The normal distribution assumption of the response on which a linear regression model is based is not so easy to be satisfied. First of all the variable must be a continuous one (if it is not, we need a model for discrete variables like logit or probit) and then we have to demonstrate the normality, also a posteriori with the residuals diagnostic tools

Our proposal **2. Statistics Use/Abuse**



- **The variables have to be clearly defined**
- Latent or observed variable

Structural Equation Models?

Sometimes we are not interested in an observed response variable, but we want to consider a latent phenomenon working at the basis of the observed values.

Treating latent variables implies the use of techniques that consider the special nature of these kind of variables.

Particularly we are able to test the cause-effect relationships fitting Structural Equation Models



Our proposal **2. Statistics Use/Abuse**

- **The variables have to be clearly defined**
- Aggregated or simple variables

Multicollinearity?

Sometimes we work with variables built for our special purposes in a not simple way making transformations of other variables.

In this case we have to treat carefully such kind of response variables, because they present high correlation with those variables they are derived from. In a regression analysis this means unreliable results and often it implies not convergence of the estimation algorithm.

Our proposal **2. Statistics Use/Abuse**



- **The analysis object has to be clearly defined**

For example:

The analysis is about a "border variable" so the estimated value produce relevant consequences (a little error is not irrelevant: this is the case for example of electoral survey)

Interval estimation Vs point estimation

- ✗ Sample distribution?

Our proposal **2. Statistics Use/Abuse**



- **The statistical techniques must be correctly used**, it also means that their basic hypothesis have to be satisfied

For example, for a regression model we can not use the usual Ordinary Least Squares model (OLS regression) when

- ✗ The residuals are heteroscedastic **GLS regression**
- ✗ Particularly if data have a hierarchical structure **Mixed Effects Models**

They permit:

- ✗ Efficient estimates
- ✗ Context effects evaluation



Our proposal **2. Statistics Use/Abuse**

- Forecasting takes place in creating future scenarios

Initializing parameters are able to incorporate historical informations

i.e. financial markets



Monte Carlo simulation

Our proposal **3. Interpretation**

We can talk a long about the need to avoid mistakes in interpreting the statistical analysis results. Sometimes a mistake in this step produce very dangerous consequences (also from an economical point of view). We only recall

- The need to preserve the multidimensional nature of phenomena: when it is not possible to reduce it, we have not simply to forget one or more variables
- We have to consider the partial correlation to evaluate the cause-effect relations



Conclusions

Knowledge discovery rests on the three balanced legs of computer science, statistics and client knowledge.

It will not stand either on one leg or on two legs, or even on three unbalanced legs.

Successful knowledge discovery needs a substantial to collaboration from all three.

DATA MINING = STATISTICS + CLIENT DOMAIN

Conclusions

There is the opportunity for an immensely rewarding **synergy** between data miners and statisticians. However, most data miners tend to be ignorant of statistics and client's domain; statisticians tend to be ignorant of data mining and client's domain;

and clients tend to be ignorant of data mining and statistics.

