12ª conferenza nazionale di statistica

PIÙ FORZA AI DATI
UN VALORE PER IL PAESE

LABORATORIO NUMERACY

How Statistics changes In the BIG DATA era?

MAURIZIO VICHI | Sapienza Università di Roma
maurizio.vichi@uniroma1.it

Federation of European National Statistical Societies
FENStatS

ESAC
European Statistical Advisory Committee

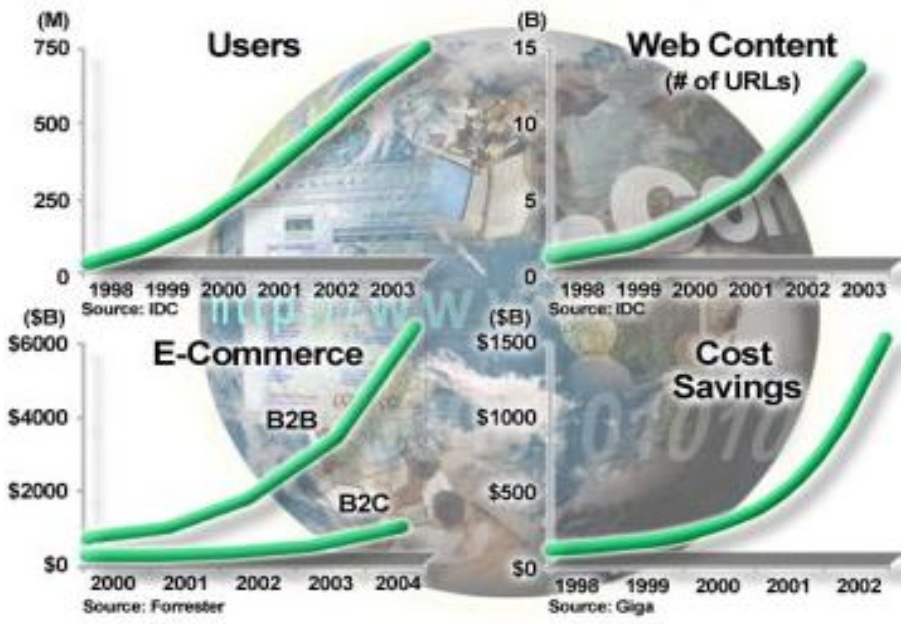
SAPIENZA
UNIVERSITÀ DI ROMA

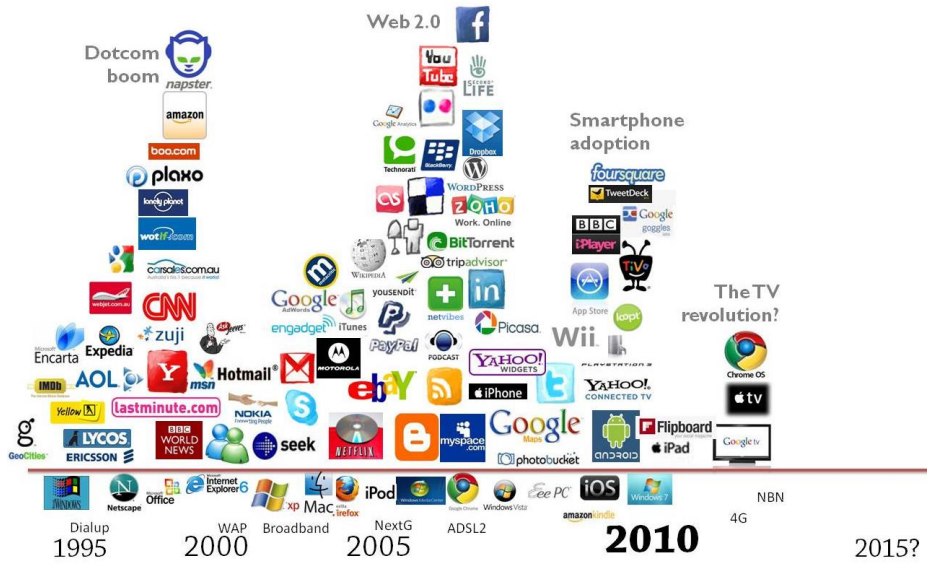
Changes by means of

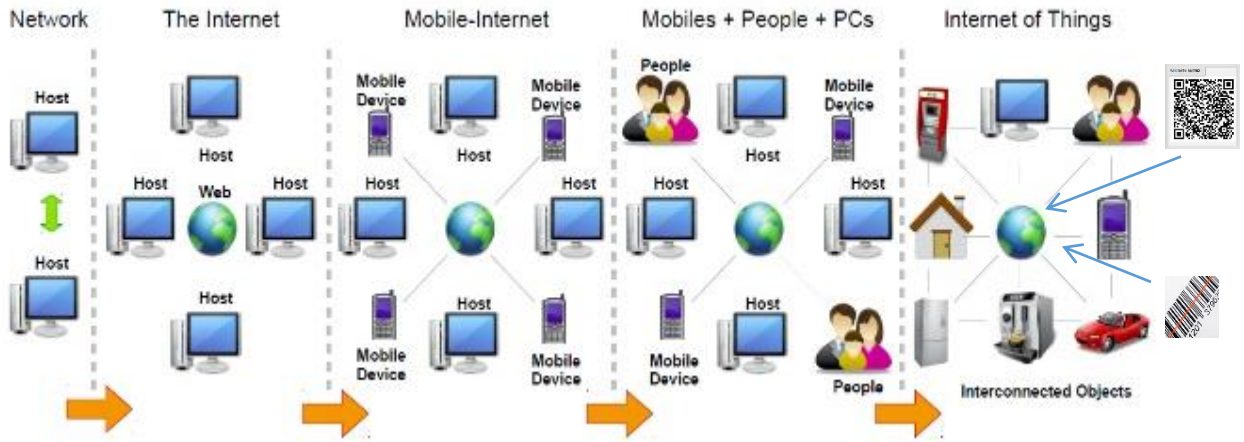
Two Revolutions

The internet Big Data Revolution

Strong Increase
USERS, WEB CONTENT, E-COMMERCE
COST SAVINGS





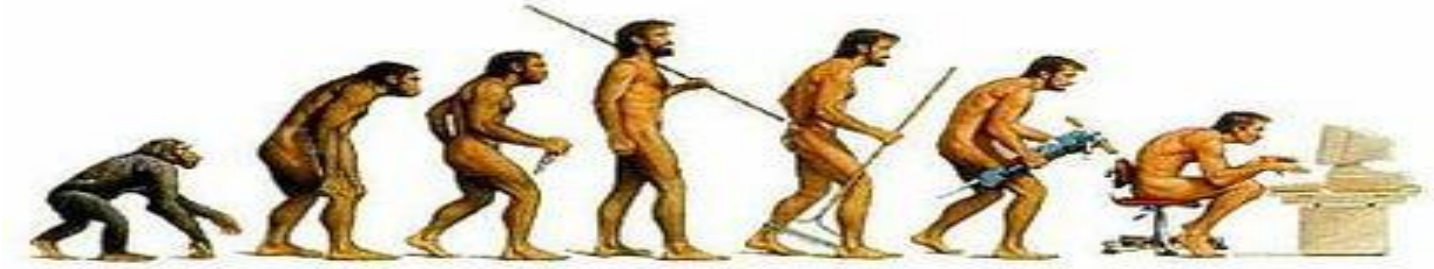


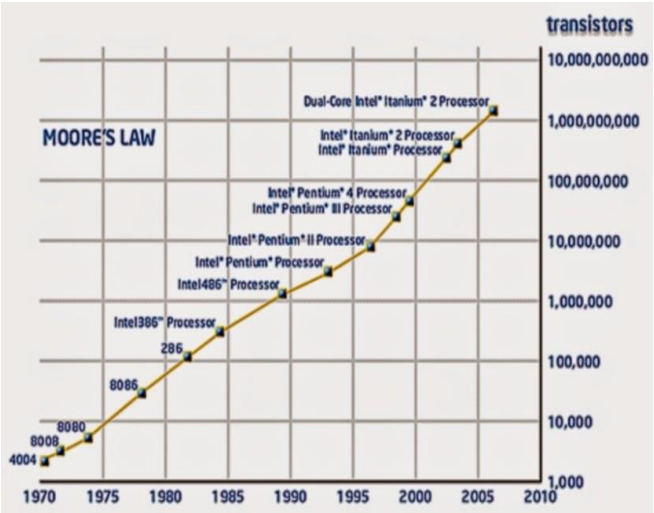
Internet of Things
From connecting Computer
To connect things

Statistics should promote a SEMANTIC WEB

that **provides standards** for **sharing data** and **reuse data** for different applications, for enterprises, for statistics purposes and scientific communities

The Computer and Technology Revolutions



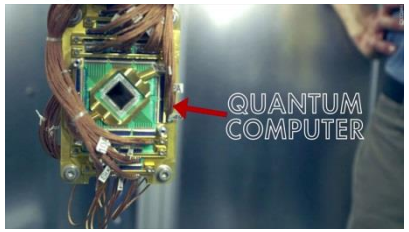


speed of the **computer** would double every 18 months and the costs would decrease by half every 2 years.

Parallel computing directly included in the new CPUs **increases** the possibility to **parallelize** modern **computer intensive** statistical **methodologies** that use independence such as resampling

Quantum computing with molecules that encode bits in **multiple states** . The QC naturally perform myriad operations in parallel, using only a single processing unit.

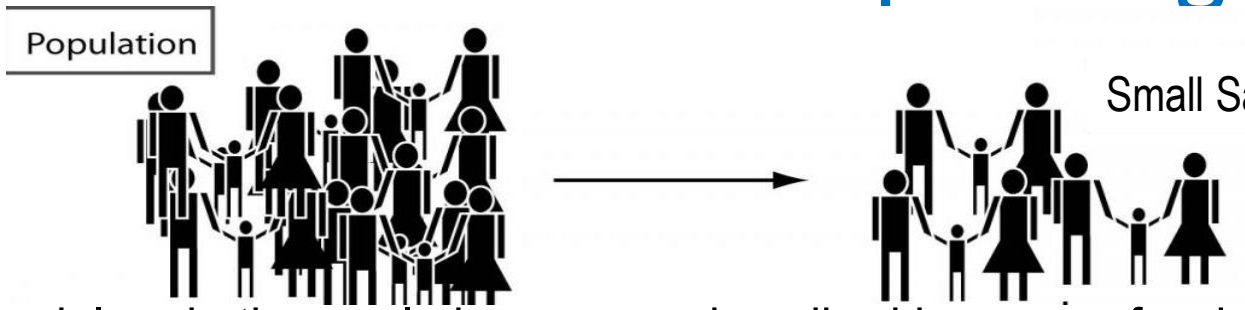
Quantum Computers and **parallel computing** very **helpful** for **computer intensive statistical methods** **(simulation resampling)**



STATISTICS BEFORE AND AFTER REVOLUTIONS

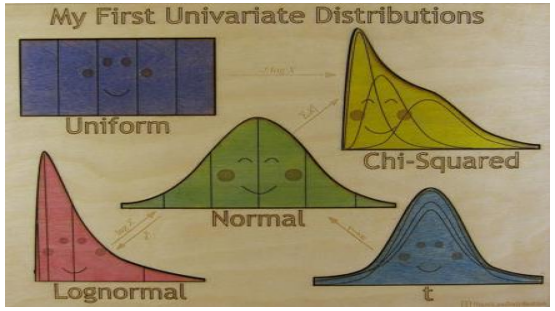
Small samples	vs	Large samples or Populations
Classical statistical Inference	vs	Computer-intensive statistical inference
Phenomena Univariate and bivariate	vs	Multivariate Phenomena in space and time domains

Statistics before computer age

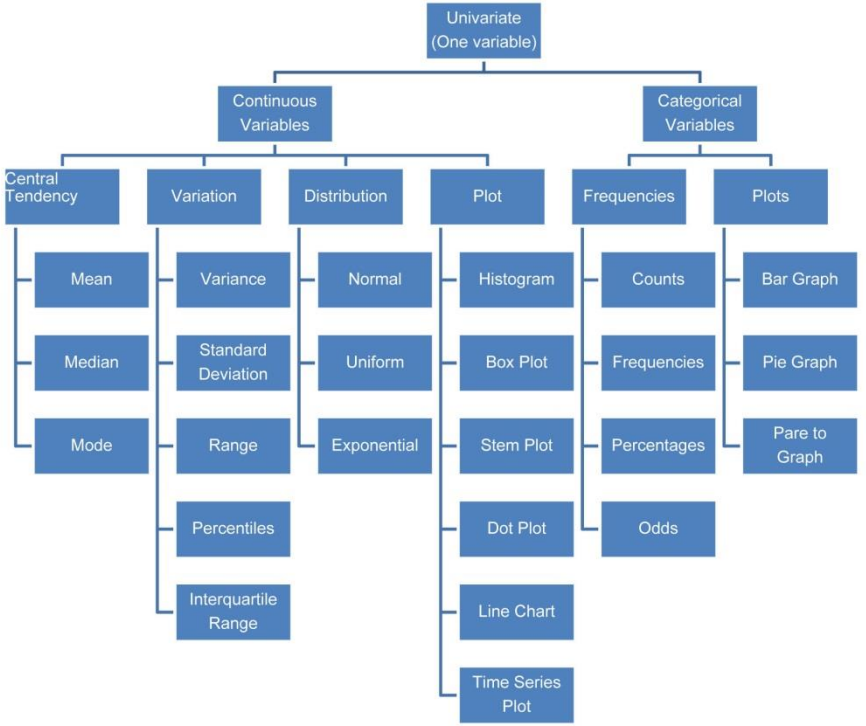


Manual data collection

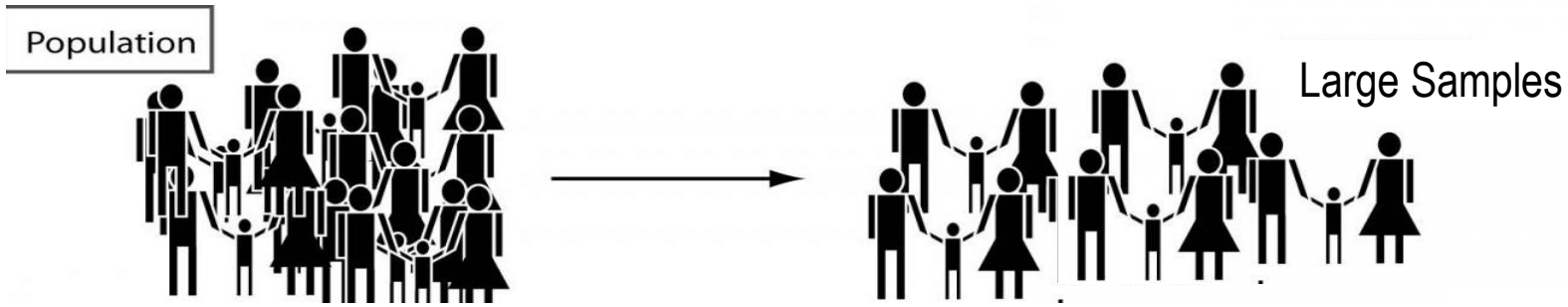
- Economic, social and other real phenomena described by one or few indicators
- Inference mainly on Univariate and Bivariate random variables
- Extensive use of parametric statistics
  especially under normality hypotheses



- Statistics based only on Mathematics and Probability



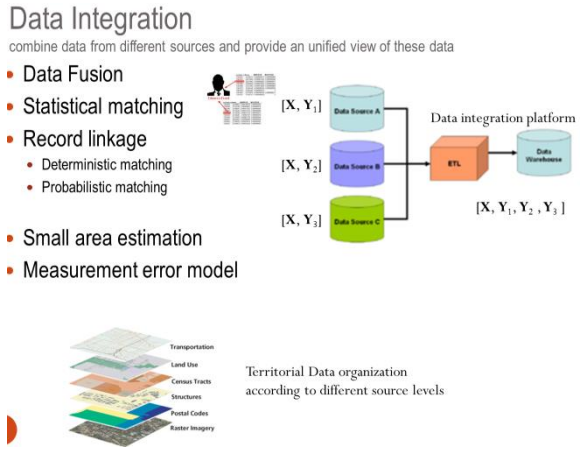
Statistics after computer-age



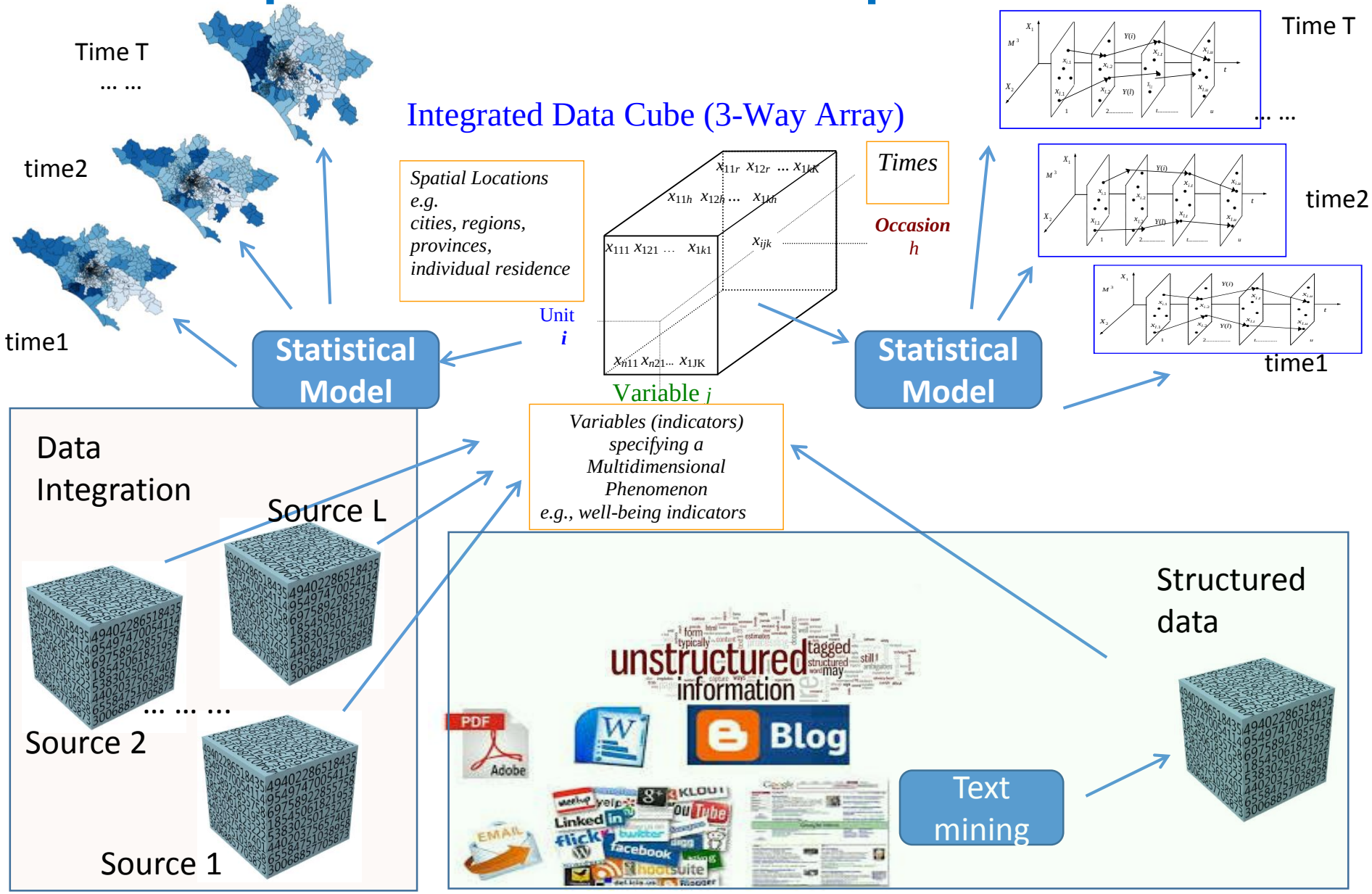
Electronic and Automatic data collection

- Statistics based on Mathematics, Probability and Computer Science (Computational Statistics)
- Computer-intensive statistical Inference (Jackknifing, Bootstrapping, Cross-validation, permutation tests)
- Economic, social and other real phenomena described in their full complexity

- BIG DATA: interconnect data



Big Data Complexity. How Big Data are formed? Multivariate phenomena with their space & time domains



Big Data: information + noise (error)

From data to information

Data Compression (reduce data of a given factor)

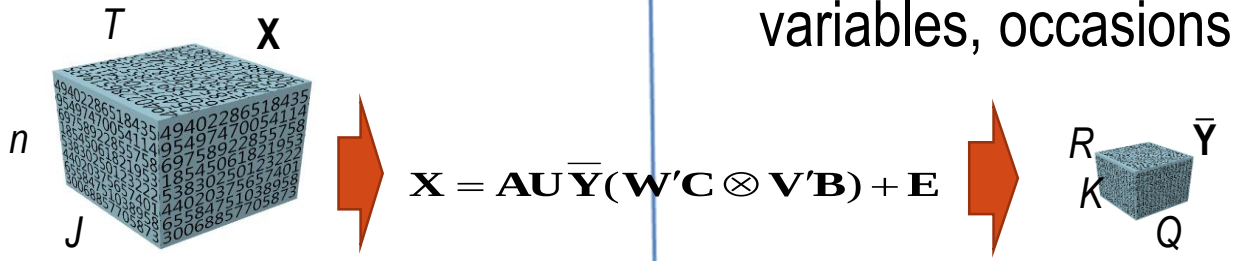
- Soft Data Compression (robustification)
- Data Fusion

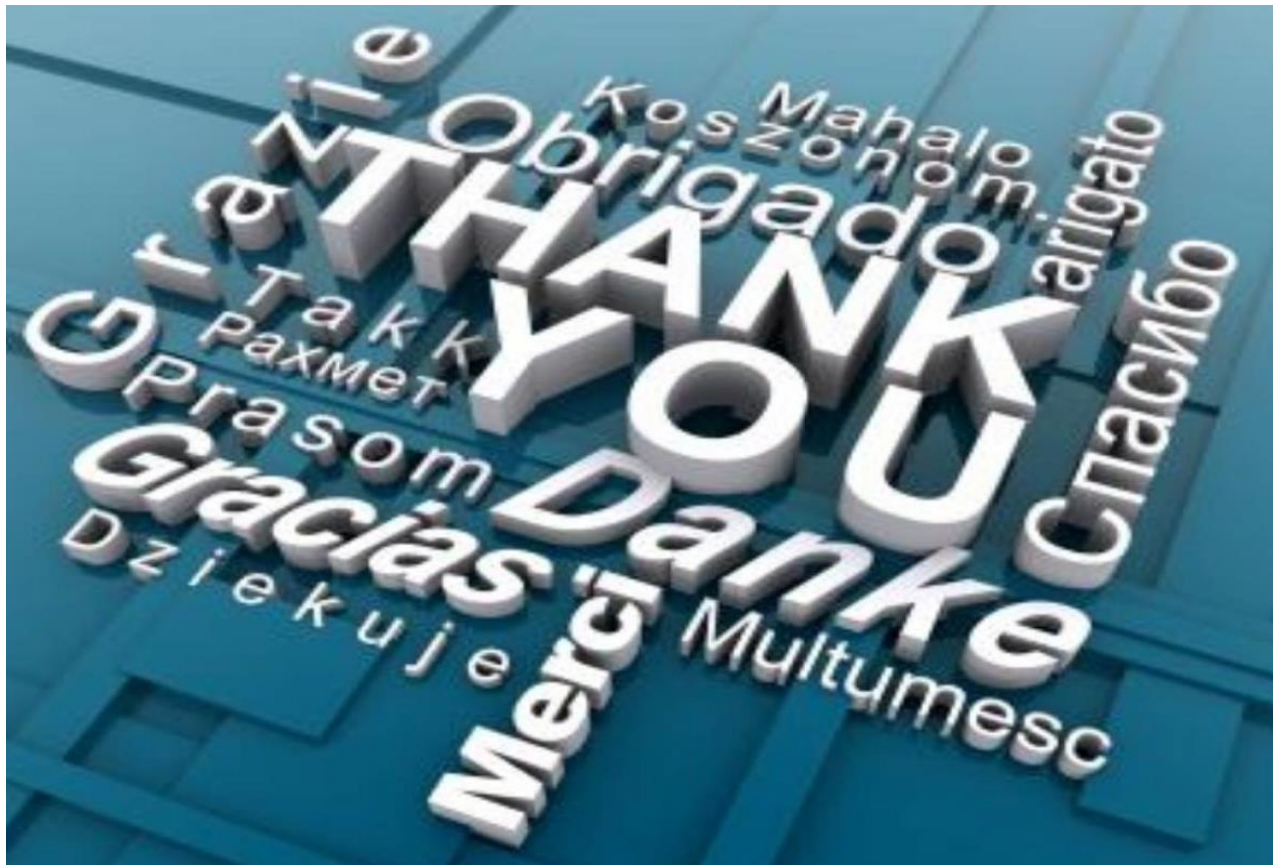
From information to knowledge

Statistical Modelling on compressed data

- Multivariate regression, VAR, SEM
- PCA, Factor Analysis, MCA, Composite Indicators
- Classification (Discriminant analysis, trees, SVM)
- Multidimensional scaling

- Hard Data Compression (Data mining)
- Taxonomy (science of classification)
- Clustering to identify typologies of objects, variables, occasions





L’attenzione è la forma più rara e più pura della generosità
Simone Weil

USERS OF STATISTICS

European Statistical Advisory Committee

☞ **ESAC has suggested a PORTAL for Communication**

PORTAL for USERS

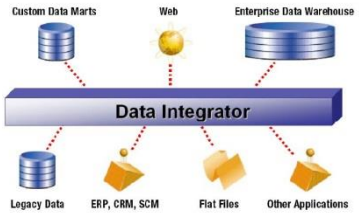
(a) Identify the very broad community of users of statistics;
(b) Segment the users' community into broadly homogeneous groups with similar interests

Groups of users Homogeneous for interest
CLUSTER ANALYSIS

Tools for decision making

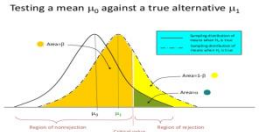
☞ Metadata;

☞ Data integrator



☞ Tools for Scenario Simulation (resampling for different scenarios)

☞ Tools to choose the best scenario (testing, cross-validation,…)



COMPUTER-AGE STATISTICAL INFERENCE

Resampling: Computing random sampling with replacement

Jackknifing: Estimating the precision of the sample **statistics** by using the observed data

Bootstrapping: Estimating the (*empirical*) *sample distribution* of the *statistics*.

Robust alternative to inference based on parametric assumptions when these are in doubt;

Utilization: **compute standard errors**, **confidence interval**, and also for **testing**;

Cross-validation: method for validating a predictive model. The goal is to estimate the expected level of fit of a model to a data set that is independent of the data that were used to train the model.

Divide the sample in validation and training sets

- Exhaustive cross-validation
  - Leave-p-out cross-validation
  - Leave-one-out cross-validation
- Non-exhaustive cross-validation
  - k-fold cross-validation
  - 2-fold cross-validation

Utilization: **compare** performances of different predictive models

variable selection.

Permutation test (exact test):

MCMC methods