The VO-Neural Project:

a GRID based astrophysical data mining environment

M. Brescia[1,2], G. d'Angelo[2,3,4], A. Nocella[3], S. Cavuoti[2,3,4], N.V. Deniskina[2,3,4], M. Garofalo[3], O. Laurino[3] & G. Longo[1,2,3,4]

__1__ - INFN, Napoli Unit; __2__ - INAF, Istituto Nazionale di Astrofisica; __3__ - Università degli Studi di Napoli Federico II; __4__ - PON S.Co.P.E.

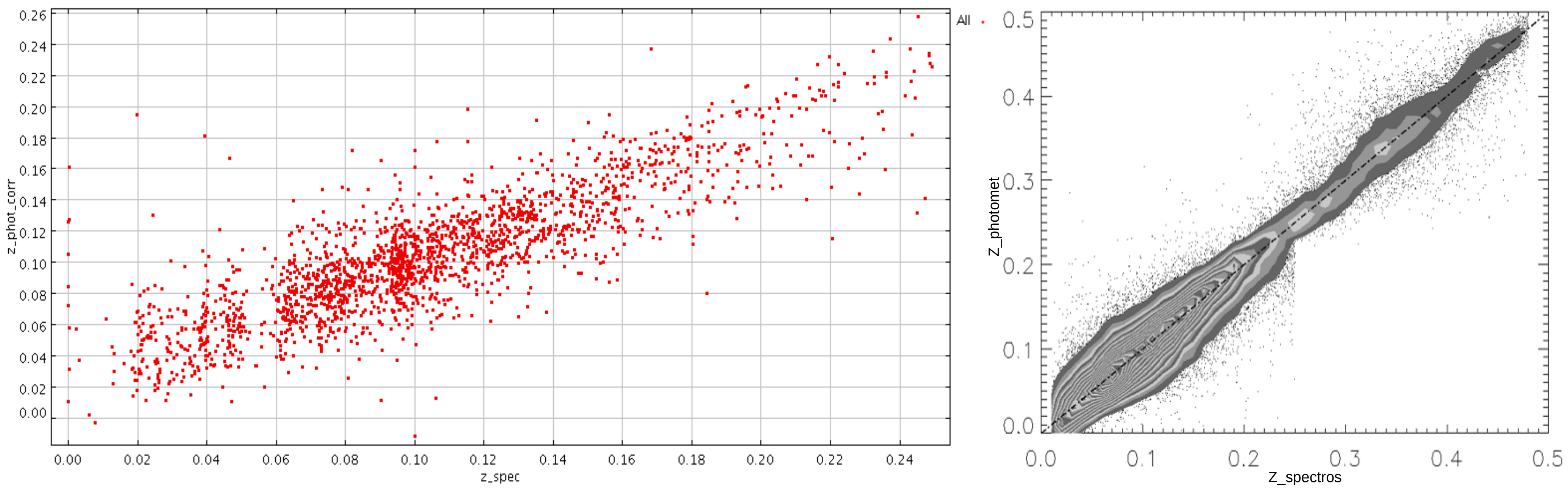
Abstract

The VO-Neural project consists of a data mining framework, whose main goal is to provide the astronomical community with powerful software instruments capable to work on massive (>>1 TB) data sets in a distributed computing environment, matching the IVOA (International Virtual Observatory Alliance) standards and requirements. These tools, being computing intensive, have intrinsic pipeline processing problems, requiring an extensive use of the GRID (PON S.Co.P.E.) infrastructure not only to access the data, but also to perform computations. VO-Neural represents the natural evolution of the Astro-Neural project which was started in 1994, as a collaboration between the Department of Mathematics and Applications at the University of Salerno and the Astronomical Observatory of Capodimonte - INAF, and is actually under continuous evolution, including also the interfacing with the UK - ASTROGRID standards.

A “Template” Scientific case

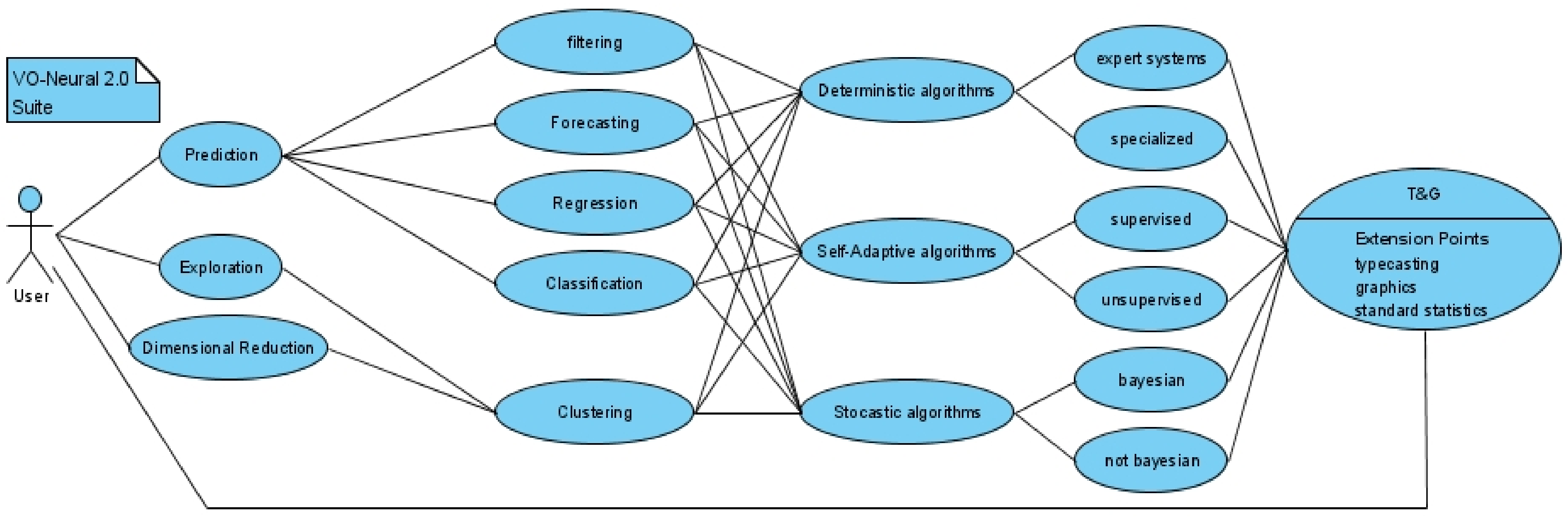
Photometric redshifts of galaxies can be estimated using a neural network trained on a knowledge base made up of a set of sources for which both parameters (in this case measured magnitudes in five optical bands: *u,g,r,i,z*) and target (spectroscopic redshifts) are known (this sample is called the “training sample”). The results of the training of the algorithm depends on the neural net model, the number of hidden layers, input and output neurons and the number of generations used for the training, activation and optimization functions, and other less important parameters.

The first step of the experiment is the training of the neural network (MLP) on the “training” sample of galaxies. After then, the analysis of the results (i.e. the assessment of the resulting redshift estimates) is performed using the trained network obtained in the first step to calculate photometric estimates of the redshift for the objects of the training sample (for which a “reference” spectroscopic measurement of the redshift is already available) and comparing the two different values of redshift for each galaxy in the sample. A qualitative estimate of the agreement between the two values can be obtained scatter plotting photometric redshifts (“photo_z”) against spectroscopic redshifts (“spec_z”).



Z_phot-corr vs z_spec

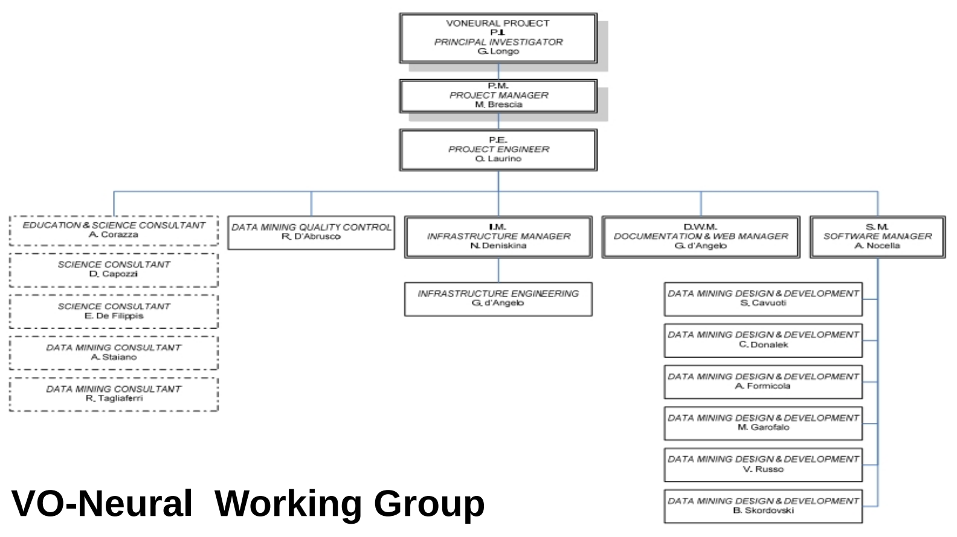
“*Mining the SDSS archive. I. Photometric redshift in the nearby universe*”, R. D'Abrusco, A. Staiano, G. Longo, E. De Filippis, M. Brescia, M. Paolillo: 2007, ApJ, 663, 752.



VO-Neural Suite Use Case Diagram

Project design highlights

VO-Neural is a suite of Data Mining designed primarily for Astrophysics and Astroparticles. It allows to extract from large datasets information useful to determine patterns, relationships, similarities and regularity in the space of parameters and outlayers. It offers main elaborative features like exploratory data analysis, data prediction and ancillary functionality like fine tuning, visual exploration of characteristics of their datasets, etc..


VONEURAL PROJECT
P.I.
PRINCIPAL INVESTIGATOR
G. Longo

P.M.
PROJECT MANAGER
M. Brescia

P.E.
PROJECT ENGINEER
O. Laurino

EDUCATION & SCIENCE CONSULTANT
A. Corazza

SCIENCE CONSULTANT
D. Capozzi

SCIENCE CONSULTANT
E. De Filippis

DATA MINING CONSULTANT
A. Staiano

DATA MINING CONSULTANT
R. Tagliaferri

DATA MINING QUALITY CONTROL
R. D'Abrusco

I.M.
INFRASTRUCTURE MANAGER
N. Deniskina

INFRASTRUCTURE ENGINEERING
G. d'Angelo

D.W.M.
DOCUMENTATION & WEB MANAGER
G. d'Angelo

S.M.
SOFTWARE MANAGER
A. Nocella

DATA MINING DESIGN & DEVELOPMENT
S. Cavuoti

DATA MINING DESIGN & DEVELOPMENT
C. Donalek

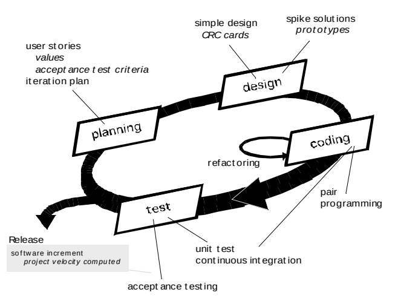
DATA MINING DESIGN & DEVELOPMENT
A. Formicola

DATA MINING DESIGN & DEVELOPMENT
M. Garofalo

DATA MINING DESIGN & DEVELOPMENT
V. Russo

DATA MINING DESIGN & DEVELOPMENT
B. Skordovski

VO-Neural Working Group



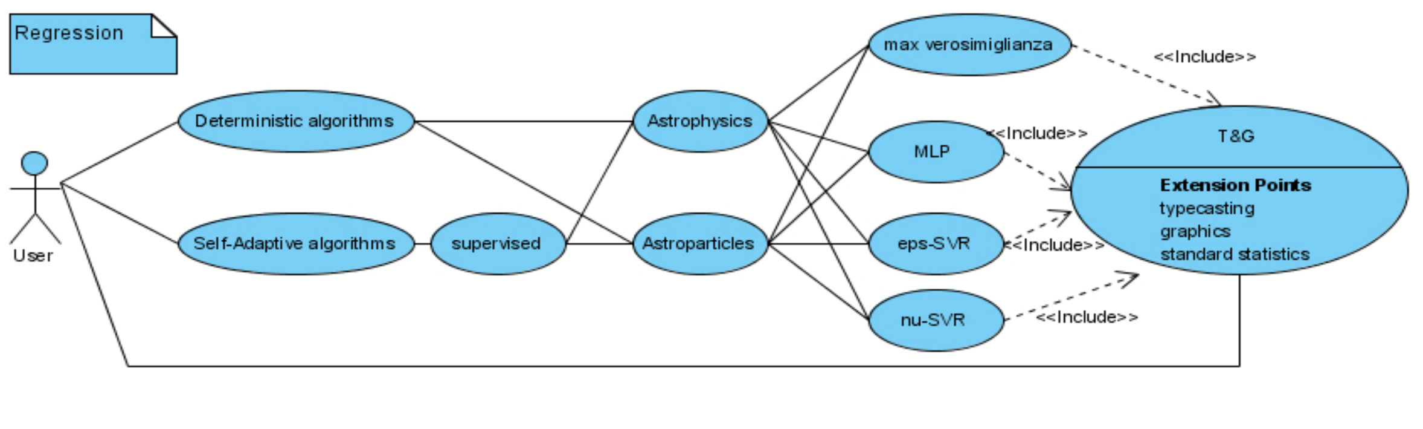
XP-Agile process workflow

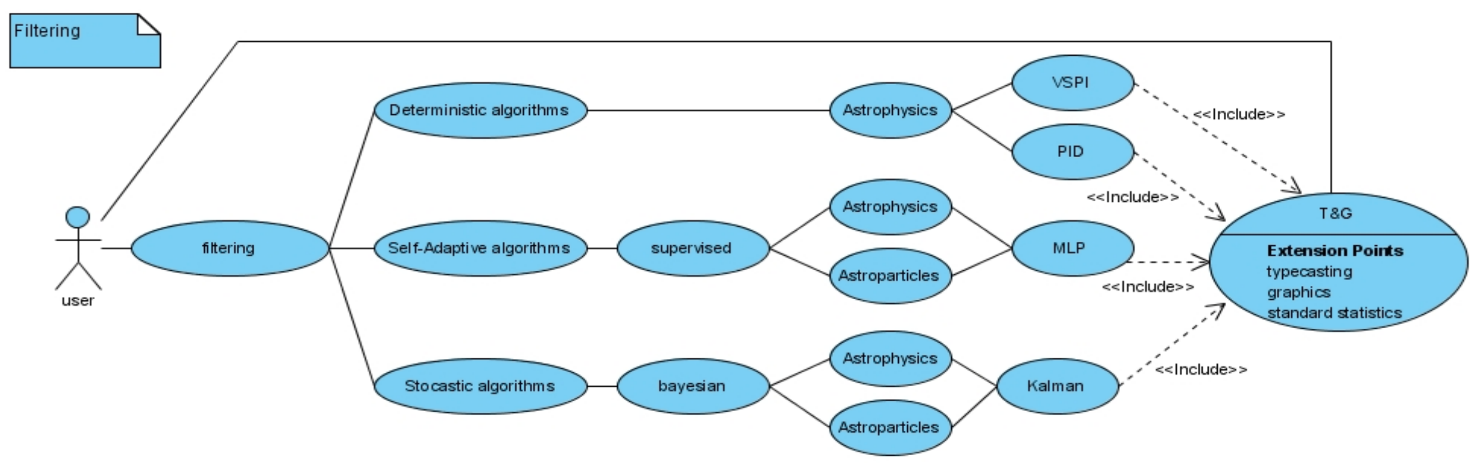
- XP-agile as suite designing method;
- UML (Unified Modelling Language);
- OOP (Object Oriented Programming);
- Interface protocols based on EGEE, VO (Virtual Observatory) & AstroGrid paradigms;
- standard I/O interface methods for software systems integrity;
- SVN (SubVersion) software version control & archiving;
- GRID-based HW Infrastructure;
- technical & scientific documentation standards;
- test & debugging standards;
- webservice-based user interfaces.

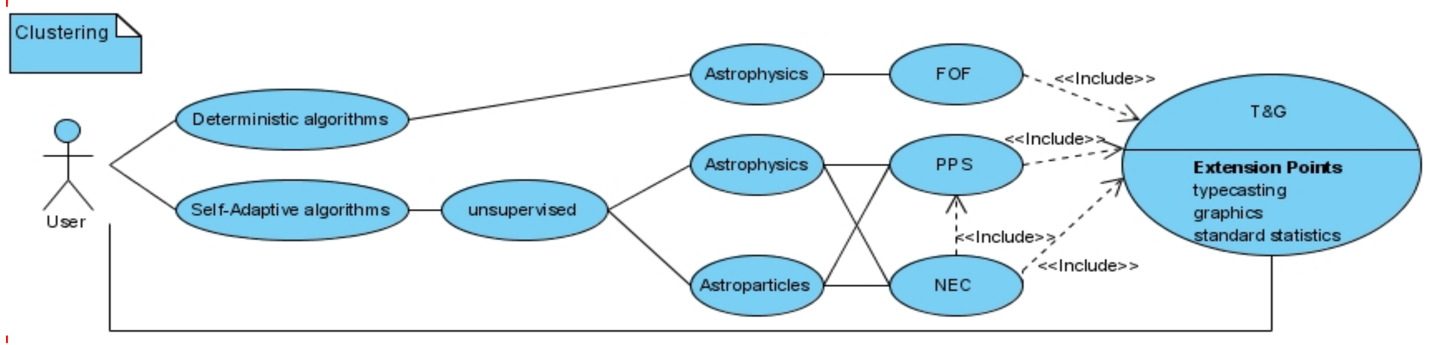
Tools Description and Use Case Diagrams

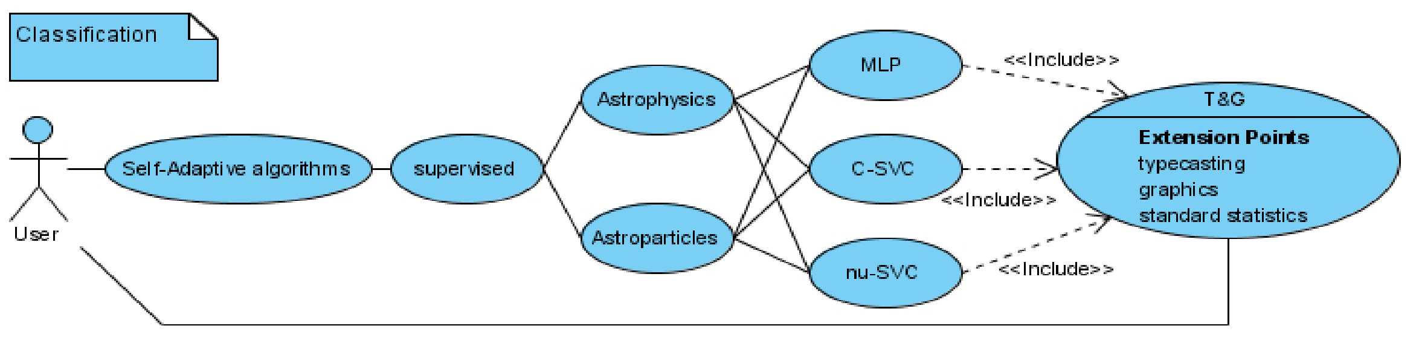
Data exploration means agglomerative clustering and dimensional reduction of parametric space; *data prediction* means: prediction, classification & regression; *fine tuning* means Not a Number (NaN) determination and outlayers, catalogue statistical analysis and data extraction. Moreover a set of graphical analysis tools is available, such as histograms and wisker & bar plot. More in detail *deterministic*, *self-adaptive* and *statistical methods* are implemented to achieve the above functions requirements as embedded in a generic pipeline.

Deterministic models are referred to *trigger* and *data reduction* algorithms. Self-adaptive models are organized in *supervised* and *unsupervised* tools. Finally statistical models are referred to simple statistical functions, Bayesian and not-Bayesian. Dimensional reduction models are referred to clustering methods like PPS, NEC and FOF. Classification provides *self-adaptive* models like supervised neural network (MLP with back-propagation and genetic algorithms, C-SVC and NU-SVC). Finally Regression are referred to MLP, other supervised *self-adaptive* models, like EPSILON-SVR and NU-SVR, and a *data fitting* deterministic algorithms.







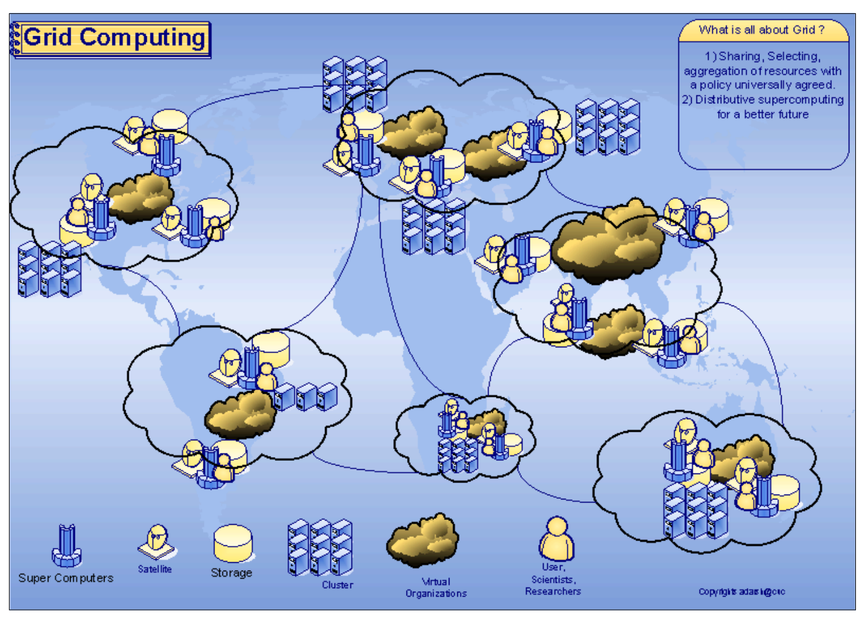


MLP: MultiLayer Perceptron, **SVM**: Support Vector Machine, **PPS** Probabilistic Principal Surfaces, **NEC**: Negative Entropy Clustering, **T&G**: Tipecasting & Graphycs, **VSPI**: Variable Structure Proportional Integral filter, **PID**: Proportional Integral Derivative filter, **FOF**: Friend Of Friend.

GRID Computing

Grid computing is a special type of parallel computing founded on a highly decentralized infrastructure. It allows the use of resources (generally CPU and storage) from different computers interconnected to a network (private, public or the Internet), by a wide number of users.

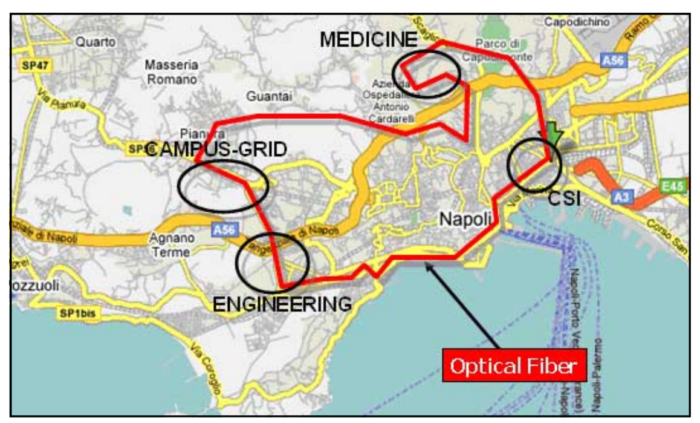
Users from different countries and institutes are organized in groups called Virtual Organizations (VO) and can use computational power and storage elements of a GRID distributed system. The GRID grants a coordinated and ensured access to shared resources as if they were on the same system.


Grid Computing
What is all about Grid ?
1) Sharing, Selecting, aggregation of resources with a policy universally agreed.
2) Distributive supercomputing for a better future
Super Computers
Satellite
Storage
Cluster
Virtual Organizations
User, Scientists, Researchers
Copyright adasi@cdc

Global GRID Computing representation

PON S.Co.P.E. GRID

The S.Co.P.E. (high Performance Cooperative and distributed System for scientific Elaboration) is a research project aimed at developing several applications in the field of fundamental research, which is one of its strategic objectives. The main spin off is the implementation of an open and multidisciplinar GRID infrastructure between different departments of the “Università degli Studi di Napoli Federico II”, distributed on Metropolitan scale.



PON S.Co.P.E. GRID, Napoli (Italy)

One of the most relevant by products of our participation to the project has been the impact that it had on the formation of a new generation of highly motivated, highly skilled young scientist and professionals. Since its start the technological backgrounds and scientific methodology spread out from physics computer science to astrophysics and space physics, in order to explore the state of art in Information Technology as the most powerful instrument for fundamental research in a wide scientific and technology area.

In order to advertise the results of the VO-Neural Project, a website was created (**http://voneural.na.infn.it/**) aiming at facilitating the information exchange among the various talian and european units participating to the project.