



DAME

Astrophysical DAta Mining & Exploration

on SC PE GRID

Progetto
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

M. Brescia – S. G. Djorgovski – G. Longo
&
DAME Working Group

Istituto Nazionale di Astrofisica – Astronomical Observatory of Capodimonte, Napoli
Department of Physics Sciences, Università Federico II, Napoli
California Institute of Technology, Pasadena



The Problem



Astrophysics communities share the same basic requirement: dealing with massive distributed datasets that they want to integrate together with services

In this sense Astrophysics follows same evolution of other scientific disciplines: the growth of data is reaching historic proportions...

“while data doubles every year, useful information seems to be decreasing, creating a growing gap between the generation of data and our understanding of it”

Required understanding include knowing how to access, retrieve, analyze, mine and integrate data from disparate sources

But on the other hand, it is obvious that a scientist could not and does not want to become an expert in its science and in Computer Science or in the fields of algorithms and ICT

In most cases (for mean square astronomers) algorithms for data processing and analysis are already available to the end user (sometimes himself has implemented over the years, private routines/pipelines to solve specific problems).

These tools often are not scalable to distributed computing environments or are too difficult to be migrated on a GRID infrastructure

A Solution



So far, our idea is to provide:

User friendly GRID scientific gateway to easy the access, exploration, processing and understanding of the massive data sets federated under standards according Vobs (Virtual Observatory) rules

There are important reasons why to adopt existing Vobs standards: long-term interoperability of data, available e-infrastructure support for data handling aspect in the future projects

Standards for data representation are not sufficient. This useful feature needs to be extended to data analysis and mining methods and algorithms standardization process. It basically means to define standards in terms of ontologies and well defined taxonomy of functionalities to be applied in the astrophysical use cases

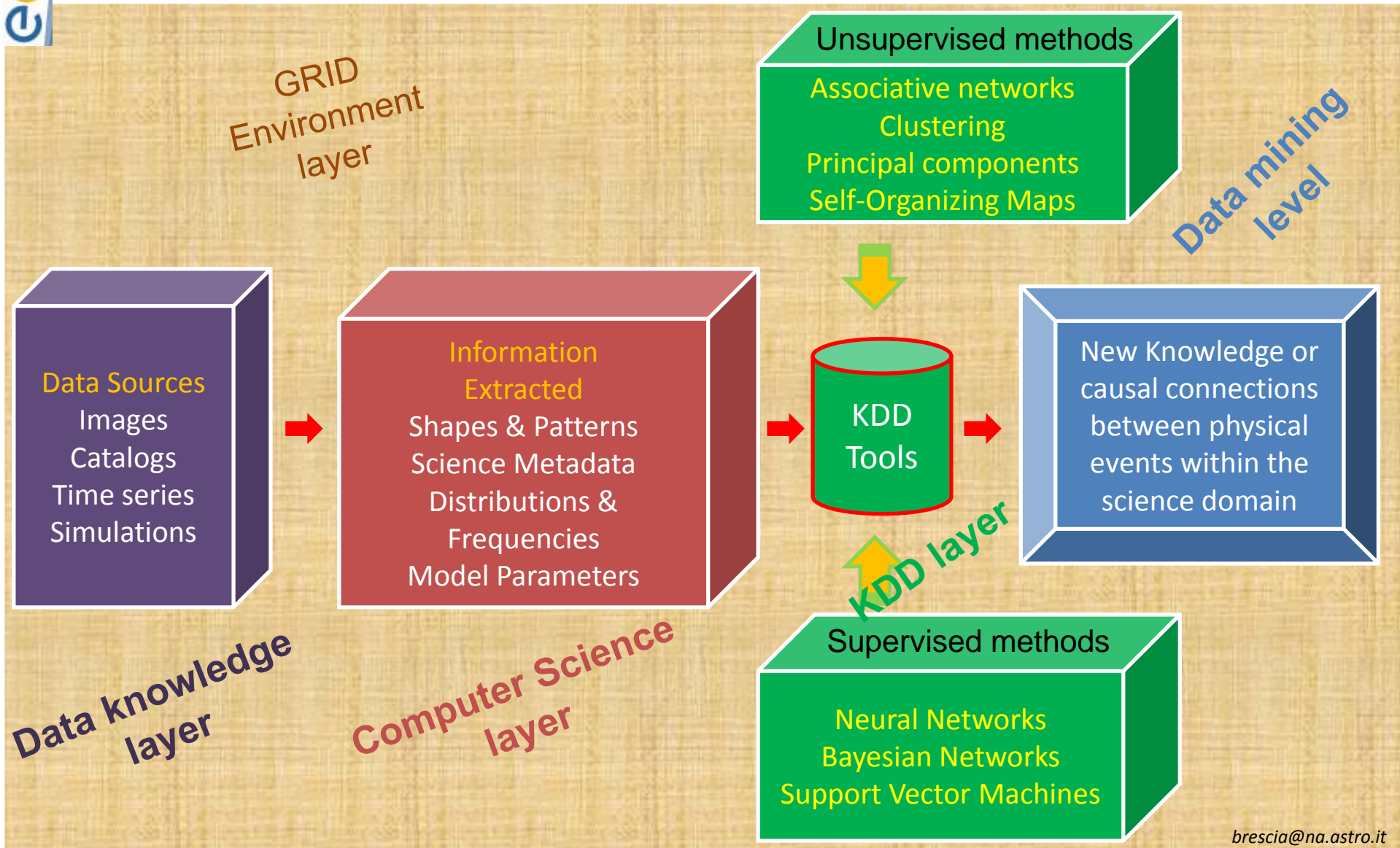
The natural computing environment for the MDS processing is GRID, but again, we need to define standards in the development of higher level interfaces, in order to:

- isolate end user (astronomer) from technical details of VObs and GRID use and configuration;
- make it easier to combine existing services and resources into experiments;

The Required Approach



At the end, to define, design and implement all these standards, a new scientific discipline profile arises: the **ASTROINFORMATICS**, whose paradigm is based on the following scheme



The new science field



Any observed (simulated) datum p defines a point (region) in a subset of R^N

Example:

- experimental setup (spatial and spectral resolution, limiting mag, limiting surface brightness, etc.) parameters
- RA and dec
- λ , time
- fluxes
- polarization

The computational cost of DM:

- N = no. of data vectors,
- D = no. of data dimensions
- K = no. of clusters chosen,
- K_{max} = max no. of clusters tried
- I = no. of iterations,
- M = no. of Monte Carlo trials/partitions

K-means: $K \times N \times I \times D$

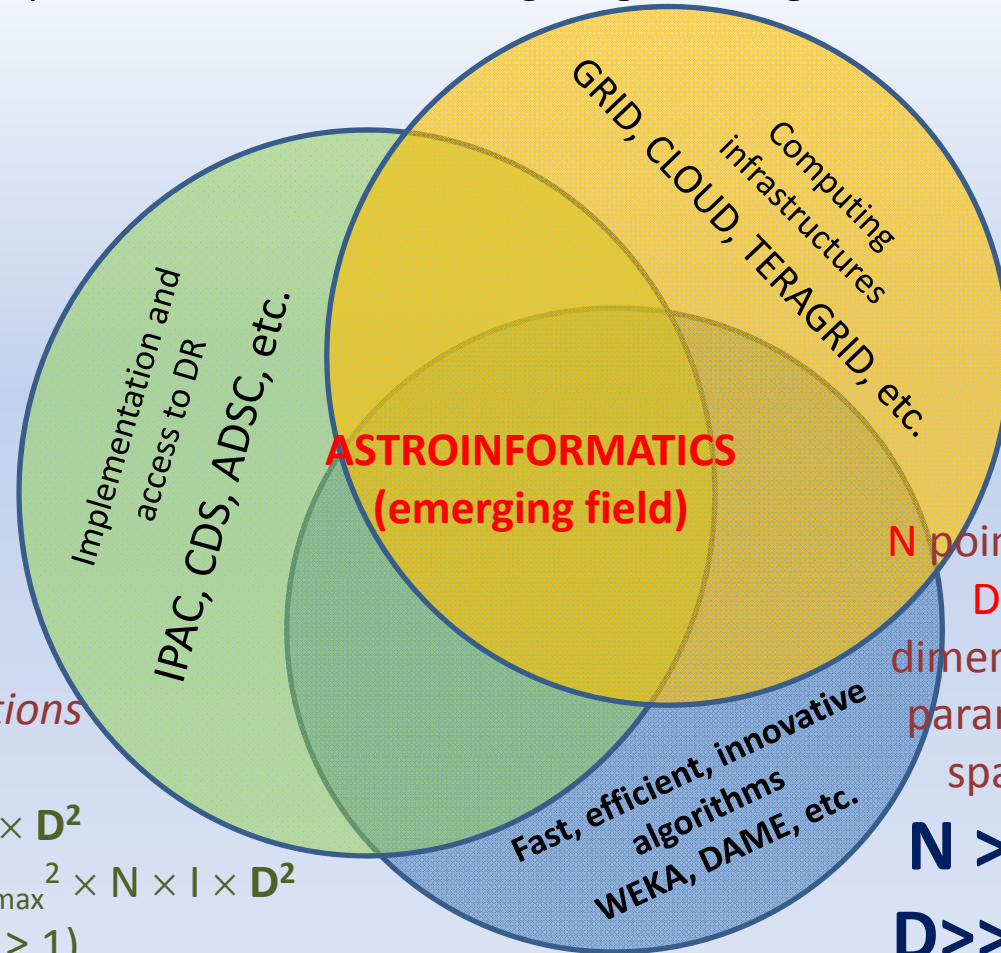
Expectation Maximization: $K \times N \times I \times D^2$

Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times D^2$

Correlations $\sim N \log N$ or N^2 , $\sim D^k$ ($k \geq 1$)

Likelihood, Bayesian $\sim N^m$ ($m \geq 3$), $\sim D^k$ ($k \geq 1$)

SVM $> \sim (N \times D)^3$



N points in a $D \times K$ dimensional parameter space:

- $N > 10^9$
- $D \gg 100$
- $K > 10$

The SCoPE GRID Infrastructure



SCoPE : Sistema Cooperativo distribuito ad alte Prestazioni per Elaborazioni Scientifiche Multidisciplinari (*High Performance Cooperative distributed system for multidisciplinary scientific applications*)

Objectives:

- Innovative and original software for fundamental scientific research
- High performance Data & Computing Center for multidisciplinary applications
- Grid infrastructure and middleware INFGRID LCG/gLite
- Compatibility with EGEE middleware
- Interoperability with the other three PON 1575 projects and SPACI in GRISU'
- Integration in the Italian and European Grid Infrastructure



— Fiber Optic
 — Already Connected
 — Work in Progress



The SCoPE Data Center
 33 Racks (of which 10 for Tier2 ATLAS)
 304 Servers for a total of 2.432 procs
 170 TeraByte storage
 5 remote sites (2 in progress)

What is DAME



DAME is a joint effort between University Federico II, INAF OACN and Caltech aimed at implementing (as web application) a suite (scientific gateway) of data analysis, exploration, mining and visualization tools, on top of virtualized distributed computing environment.

<http://voneural.na.infn.it/>
Technical and management info
Documents
Science cases

DAta Mining & Exploration

Data Mining & Exploration Project

News & Events

- New DAME Prototype released
- DAME Lecture @ IJPC-09
- DAME @ IJCAL-09 Conference
- Workshops and announcements now available
- Past theses

Partners:

- Dipartimento di Fisica (sezione di Astrofisica) - Università degli Studi di Napoli Federico II
- INAF - Osservatorio Astronomico di Capodimonte
- California Institute of Technology, Pasadena - USA

Related links:

- VOTECH (Virtual Observatory Technological Infrastructures)
- S.Co.P.E. (High Performance distributed Cooperative System for scientific Experiment)
- INAF - Osservatorio Astronomico di Trieste (VO-AIDA)
- Dipartimento di Informatica Università degli Studi di Napoli Federico II
- Dipartimento di Ingegneria Informatica Università degli Studi di Napoli Federico II
- MIUR (Italian Ministry of Research)
- EURO-VO (The European Virtual Observatory)
- IVOA (International Virtual Observatory Alliance)

Astronomical Data are collected by means of a large number of different techniques and are stored in very diversified and often incompatible data repositories. Moreover in the e-science environment, it is needed to integrate services distributed across heterogeneous, dynamic "virtual organizations" formed by the different resources within a single enterprise and by external resource sharing and service provider relationships.

The DAME project aims at creating a single distributed e-infrastructure for data exploration, mining and visualization. It provides an integrated access to data collected by very different instruments, experiments and scientific communities in order to be able to correlate them and improve their scientific usability and interoperability.

The project consists in the design and development of a data mining suite which will provide the astronomical community with powerful software instruments able to work on massive data sets in a distributed computing environment, matching the international IVOA standards and requirements.

<http://dame.na.infn.it/>
Web application PROTOTYPE

DAta Mining & Exploration

Giuseppe Iengo
 Last Login: Tue 04 Aug 2009 10:50PM GMT

My Experiments

Home	Experiments List	Name	Science case	Node	Status	Actions
Home	parum	parum	finished	Remove		
Science and Tech	parum	parum	finished	Remove		
MyFavorites	parum	parum	finished	Remove		
MyExperiments	parum	parum	finished	Remove		
Login	parum	parum	finished	Remove		
Help & Tutorials	parum	parum	finished	Remove		
The Team	parum	parum	finished	Remove		
Launch Experiments	parum	parum	finished	Remove		
New MLP	parum	parum	finished	Remove		
New SVM	parum	parum	finished	Remove		
New PhotoZ	parum	parum	finished	Remove		

My Filestore

Dir	Files	Actions
/iengo	No data in this directory!	
/iengo/HAAS	HAAS.jpg	Delete Download
	catalogue_astoria.csv	Delete
	robot_astoria.txt	Delete

What is DAME



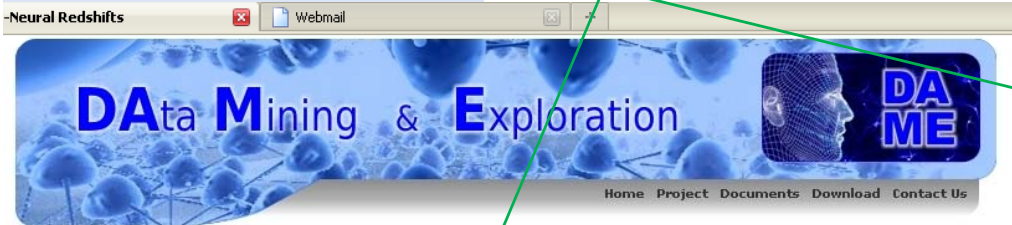
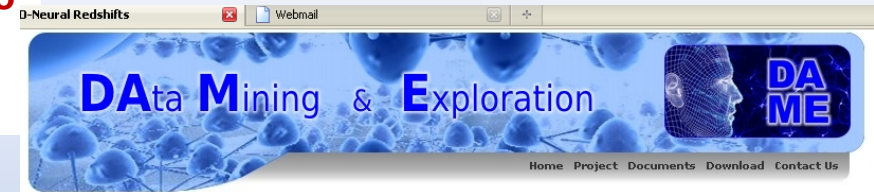
In parallel with the Suite R&D process, all data processing algorithms (foreseen to be plugged in) have been massively tested on real astrophysical cases.

<http://voneural.na.infn.it/>

Technical and management info

Documents

Science cases



Links

- Shakbazian groups in the SDSS
- Photometric redshift for SDSS galaxies
- Documents
- Public Outreach
- Science Papers

A method for the extraction of photometric QSOs candidates

In this page, you will find a description of the method for the extraction of photometric QSOs candidates described in the paper "Quasar candidates selection in the Virtual Observatory era" from D'Abrusco et al. submitted to MNRAS (preprint).

The inspiring principle of this work is the application of statistical and data-mining techniques to obtain a clustering of astronomical sources inside a photometric parameter space and fully characterize the distribution of different types of sources inside this parameter space. This concept has been applied to the problem of the selection of QSOs candidates from broadband photometric data by exploiting the availability of large spectroscopic bases of knowledge (BoK: i.e., samples of sources with a reliable classification).

The procedure for the extraction of candidates can be summarized as follows:

- A BoK consisting of a sample of stellar sources with spectroscopic classification is clustered inside the colour parameter space. This BoK is drawn from the catalogue of photometric sources from where, at the end of the process, the new QSOs candidates will be extracted.
- Several possible partitions of the distribution of sources of the BoK inside the colour space are produced by a combination of two clustering algorithm: PPS and NEC.
- The members of each cluster of each different partition are labelled using the BoK classification.
- Amongst all the possible partitions, the one that best separates the sources is selected.
- The new candidates are extracted from the colour space, to the success of the method.

The details of the method and algorithms can be found in the paper.

The catalogues of QSOs candidates extracted from the SDSS DR7 photometric survey can be downloaded [here](#).

Also, under design a web application for data exploration on globular clusters (VOGCLUSTERS)

Links

- Shakbazian groups in the SDSS
- QSO candidates in the SDSS
- Documents
- Public Outreach
- Science Papers

Evaluation of photometric redshifts using neural networks

Download the catalogues!

The work discussed here represents the natural evolution of a previous attempt described in **these** pages and presented in the **2002** and **2003** papers. The final result, namely the redshifts for a large subsample of the galaxies present in the SDSS are downloadable [here](#). This work was part of the Ph.D. Thesis of **Raffaello D'Abrusco** and has been published in **Ap.J (2007)**.

The main idea behind the work is to exploit the huge data wealth of the SDSS to train a supervised neural network to recognize photometric redshifts. The details of the work can be found in this paper. In short the procedure can be summarized as it follows:

- The training, validation and test sets are built using the SDSS spectroscopic subsample. This sample is almost complete at $m(R) < 17.7$, while for fainter magnitudes it includes mainly Luminous Red Galaxies or LRG's.
- A first MLP is trained at recognizing nearby ($z < 0.25$) objects from distant ($0.25 < z < 0.5$) ones.
- Then two networks are trained in the two different redshift ranges and the optimal architecture is found by varying the NN parameters
- The resulting redshifts show a trend which is corrected by applying an interpolative correction.
- Once the three NN have been trained the photometric data are processed for the whole galaxy sample and the photometric redshifts are derived.

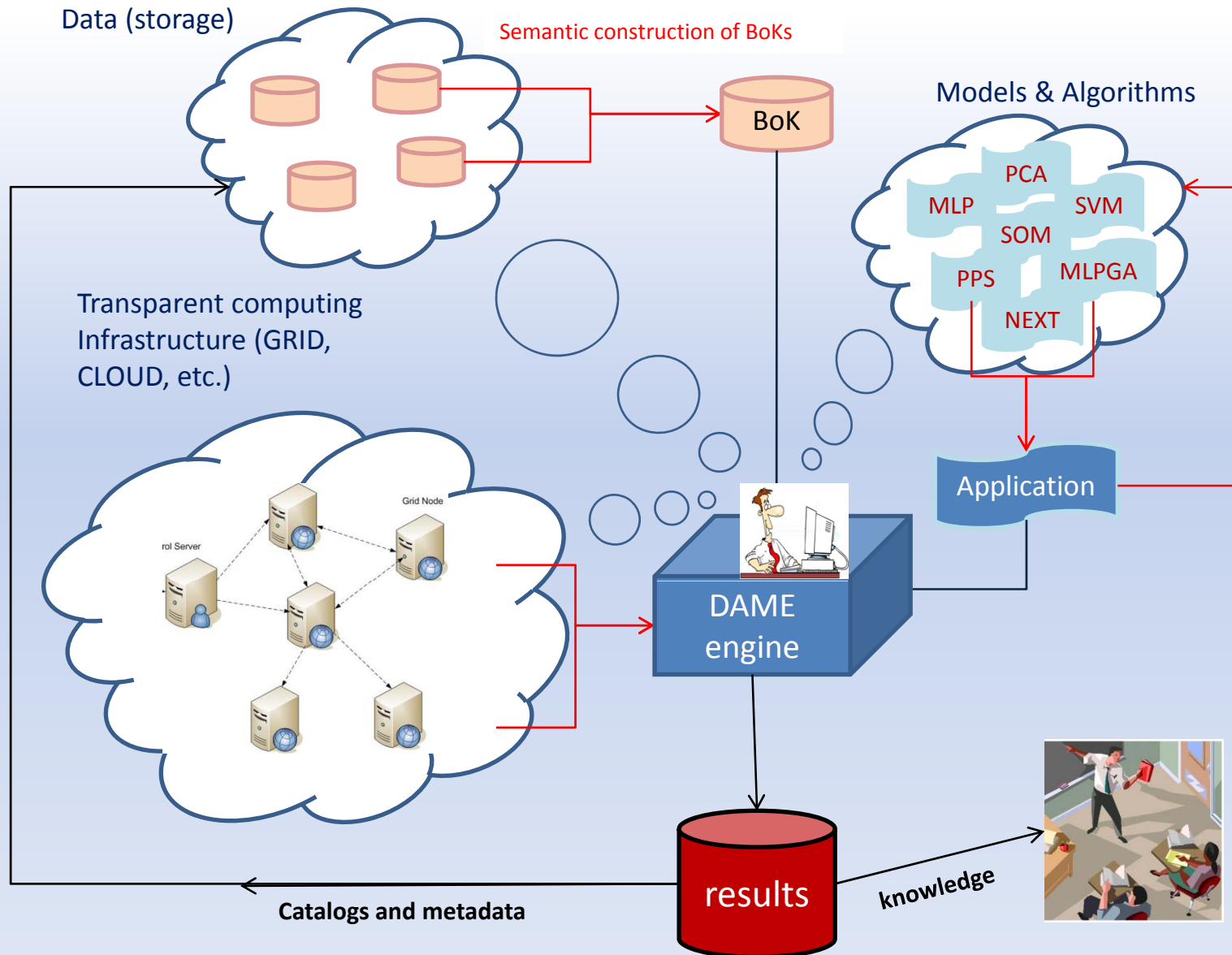
The whole procedure outlined above is repeated independently for all objects in the MAIN GALAXY sample of the SDSS and for the LRG's only. The resulting catalogues can be downloaded [here](#).

The main results can be summarized as it follows.

- 1 The method leads to an r.m.s. error (evaluated on the test set only) better than any other method

Method	Data	Δz	σ	Range
VW	EDR		0.0621	
	EDR		0.0509	
Caabai et al. (2003)	interpolative		0.0451	
	bayesian		0.0402	
Caabai et al. (2003)	empirical, polynomial fit		0.0318	
Caabai et al. (2003)	K-D tree		0.0254	
Suchkov et al. (2005)	Class X	DR-2	0.0340	
Way & Srivastava (2006)*	Gaussian Process	DR-3	0.0230	

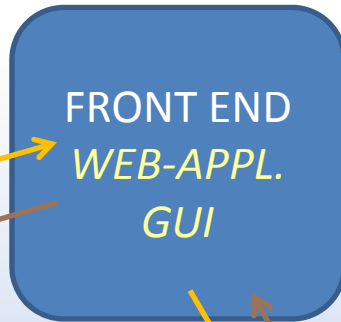
DAME Work breakdown



The DAME architecture



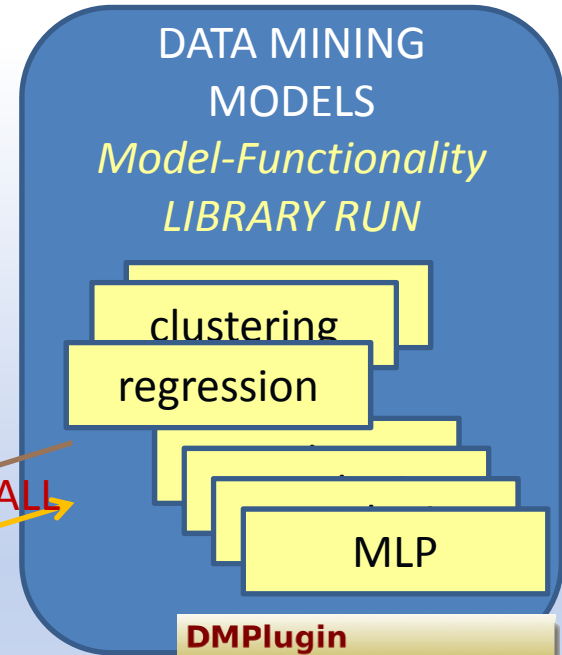
user



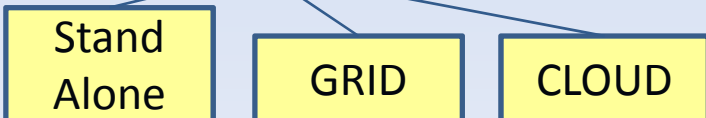
Client-server AJAX (Asynchronous Java-Xml) based; interactive web app based on Javascript (GWT-EXT);



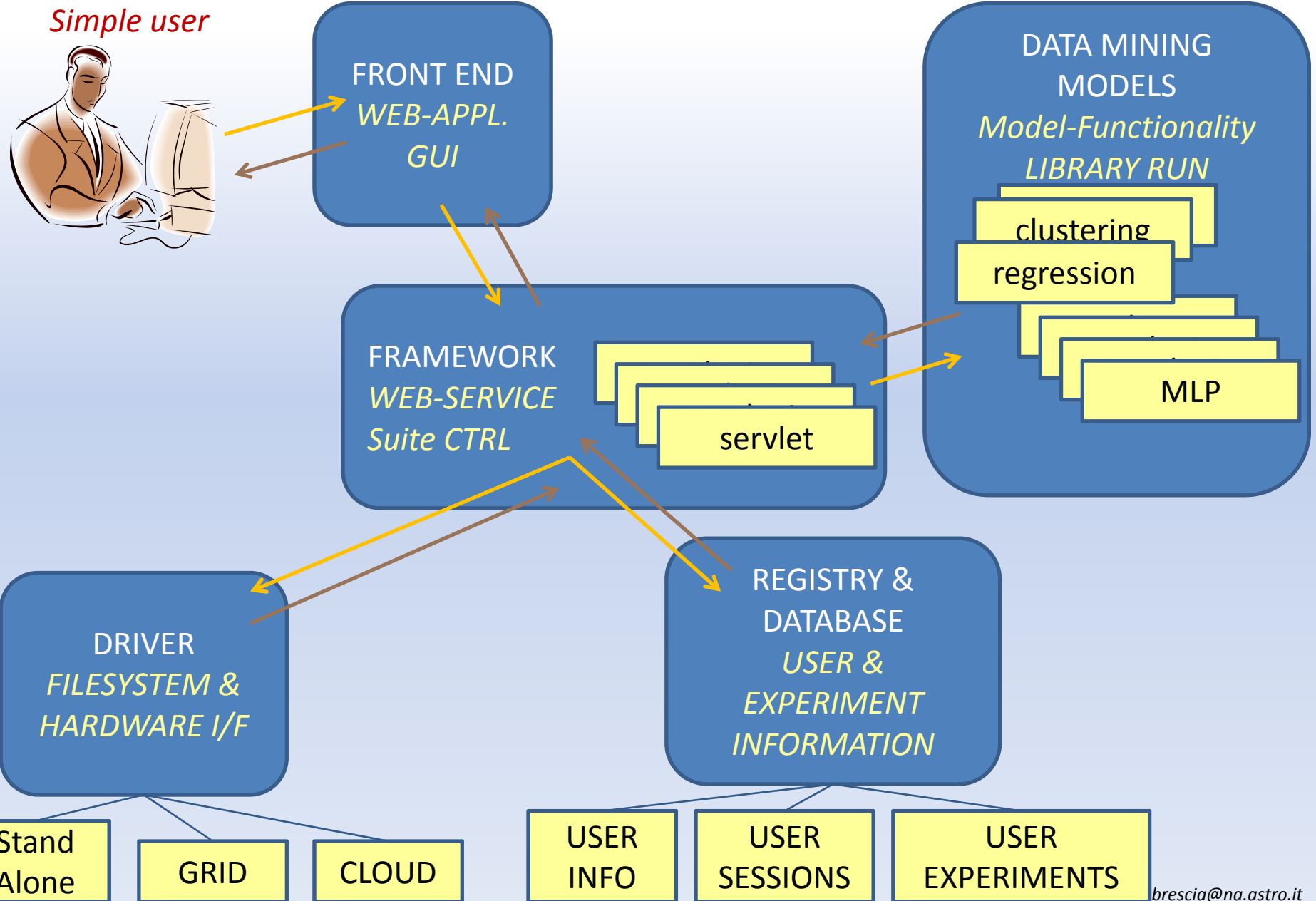
Restful, Stateless Web Service
experiment data, working flow trigger and supervision
Servlets based on XML protocol



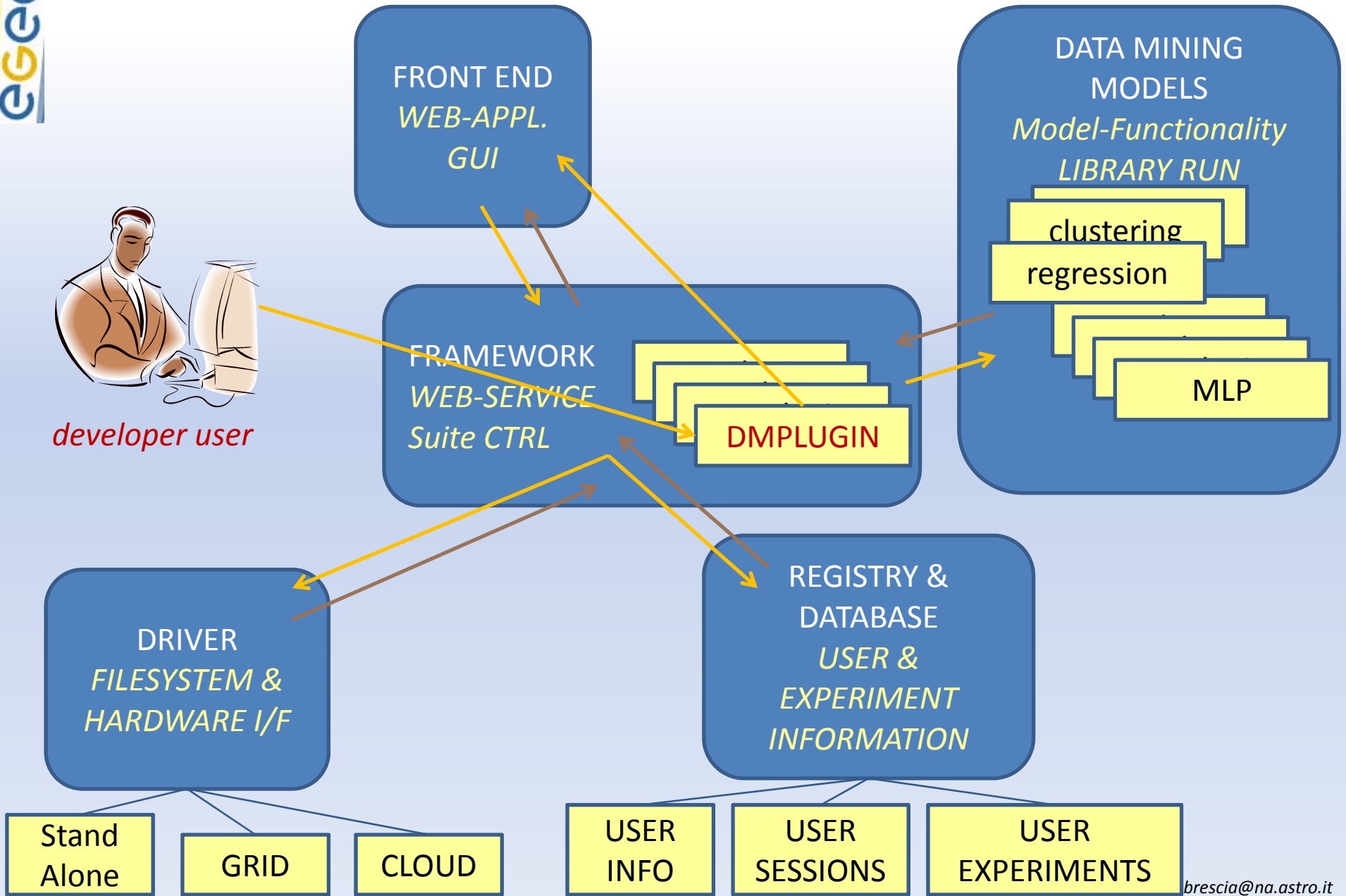
HW env virtualization;
Storage + Execution LIB
Data format conversion



Two ways to use DAME - 1



Two ways to use DAME - 2



Two ways to use DAME - 2

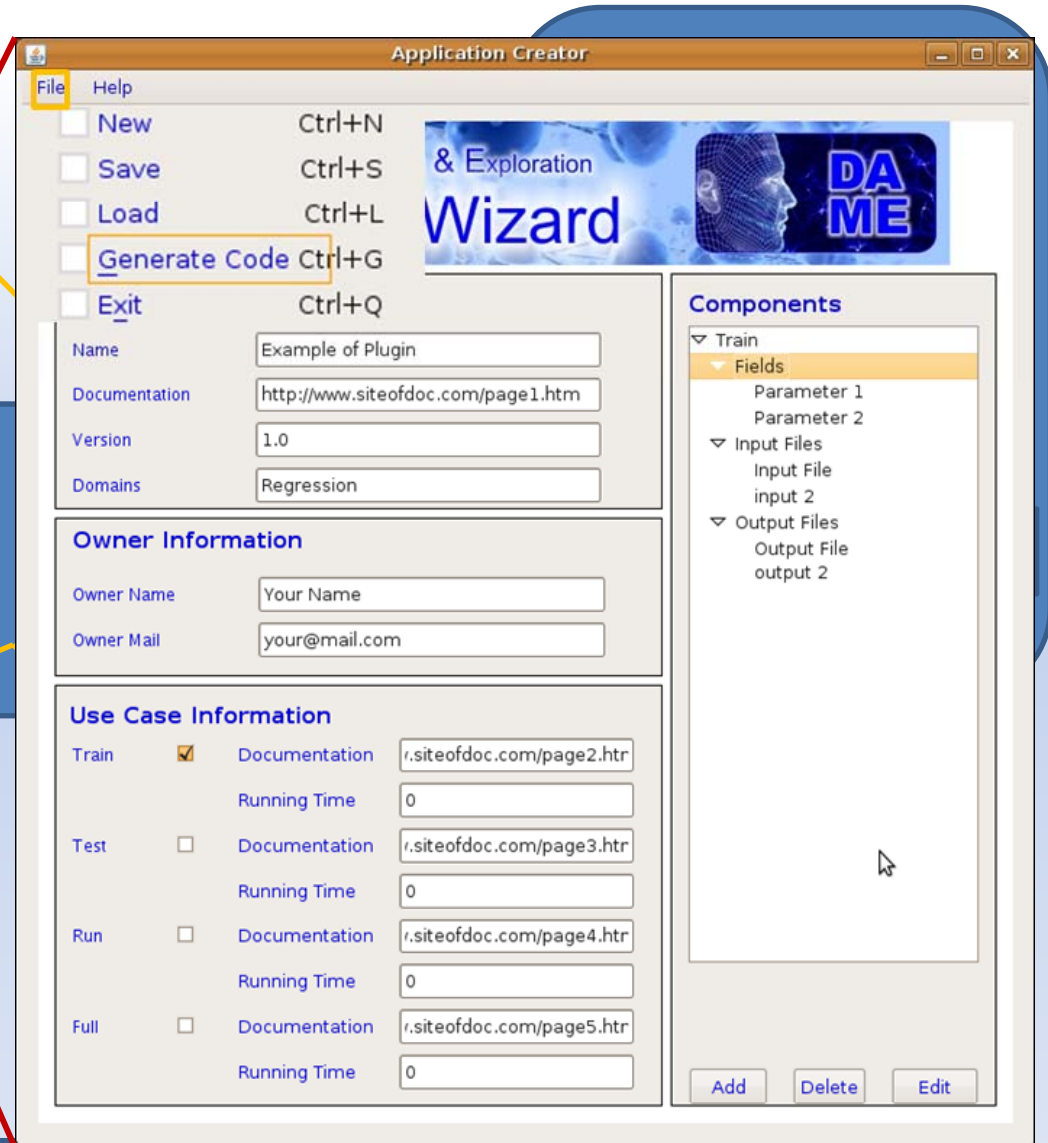


developer user

FRONT END
WEB-APPL.
GUI

FRAMEWORK
DMPlugin
Suite CTRL

DRIVER
FILESYSTEM &
HARDWARE I/F



Stand Alone

GRID

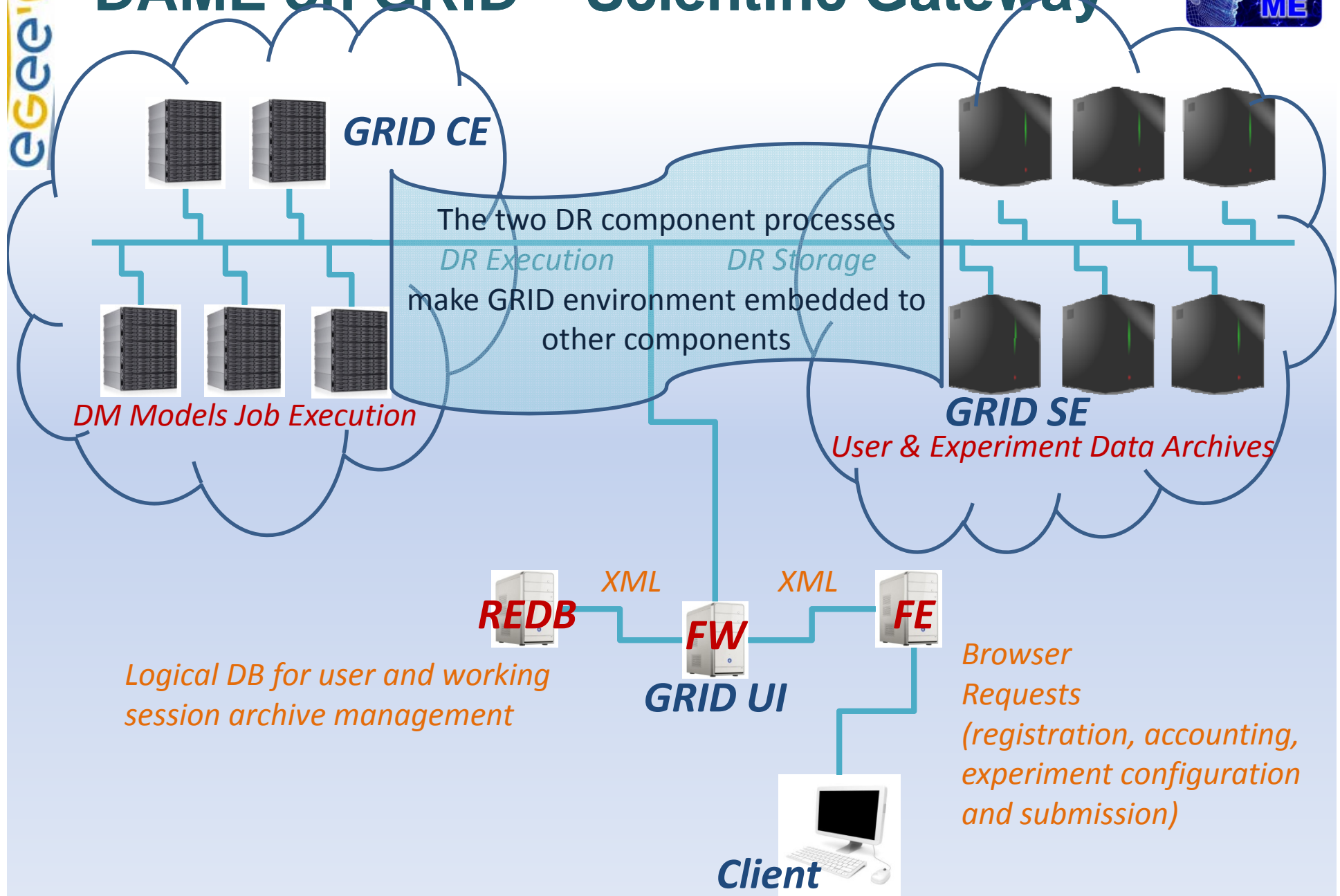
CLOUD

USER
INFO

USER
SESSIONS

USER
EXPERIMENTS

DAME on GRID – Scientific Gateway



Coming soon...



- now: suite deployed on SCoPE GRID, currently under testing;
- DMPlugin package under test (beta SW & Manual already available for download);

One of the main design goals in DAME was to make the user able to deploy any own data mining application in the DAME Suite, representing also an easy way to extend Suite tools and algorithms. To do this, the main component named **FRAMEWORK** (FW) includes a sub-component, named DMPLUGIN, that allows to configure external custom applications and to generate code through an automatic procedure.

- End of October 2009: beta version of Suite and DMPlugin released to the community;

<http://dame.na.infn.it/>

Web application **PROTOTYPE**

<http://voneural.na.infn.it/>

Technical and management info

Documents

Science cases