DAME
Astrophysical DAta Mining & Exploration on GRID

Progetto SCOPE
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

brescia@na.astro.it

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

brescia@na.astro.it

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

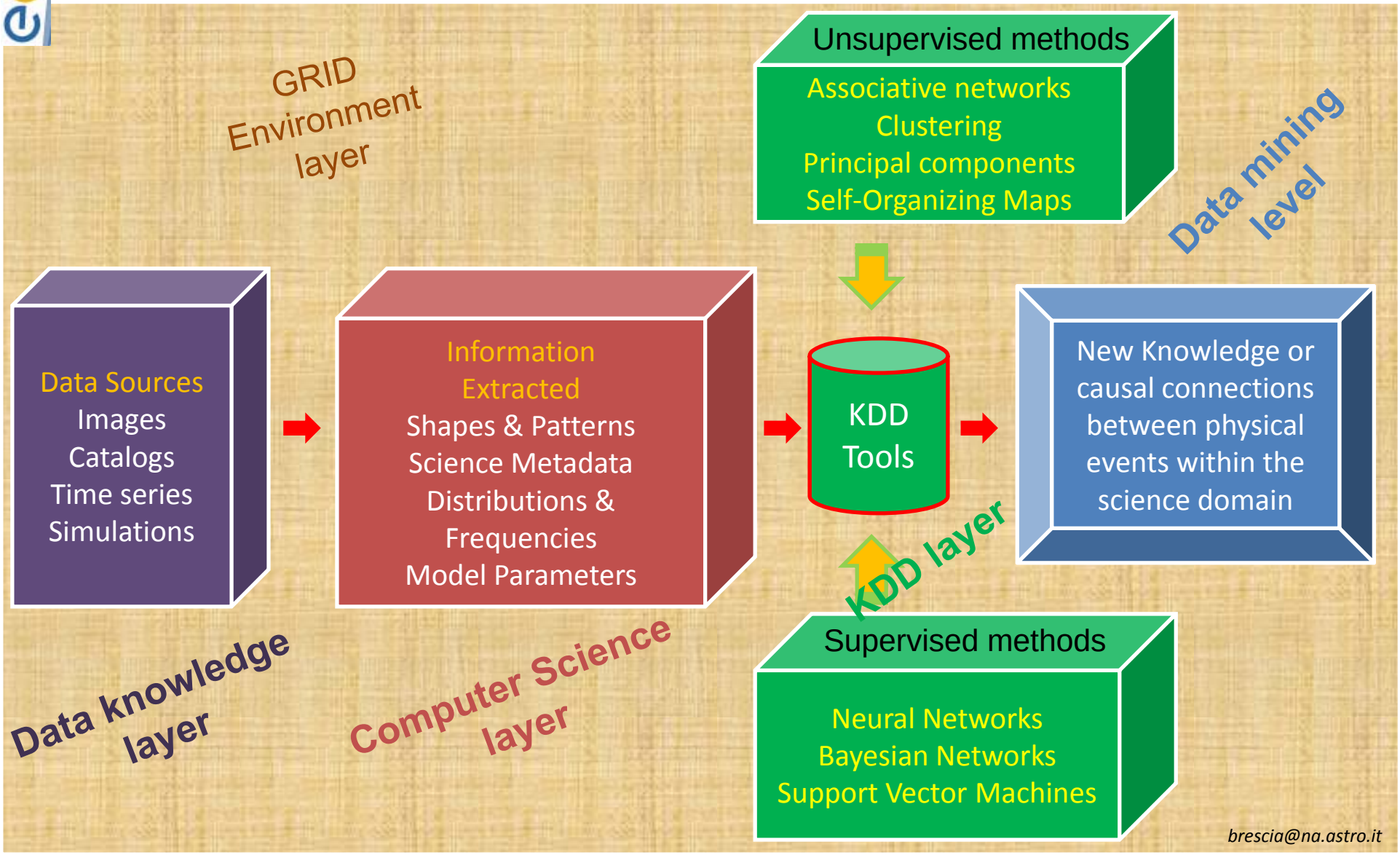
The natural computing environment for the MDS processing is GRID, but again, we need to define standards in the development of higher level interfaces, in order to:

- 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;

brescia@na.astro.it

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



brescia@na.astro.it

The new science field

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

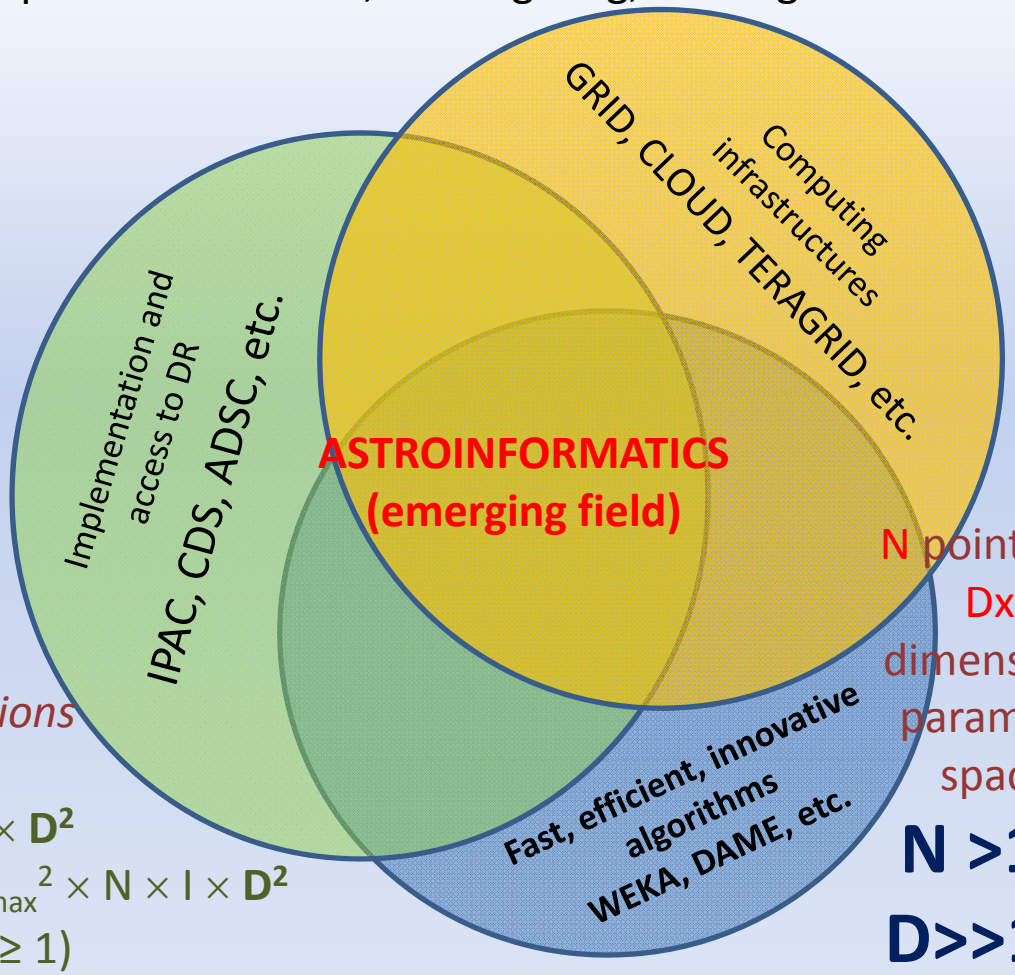
Example:

- 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: $K \times N \times I \times \mathbf{D}$
Expectation Maximization: $K \times N \times I \times \mathbf{D^2}$
Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times \mathbf{D^2}$
Correlations ~ N log N or N^2, ~ D^k ($k \geq 1$)
Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)
SVM > ~ $(NxD)^3$



N points in a DxK dimensional parameter space:

$N > 10^9$

D>>100

K>10

brescia@na.astro.it

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 INFNGRID 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

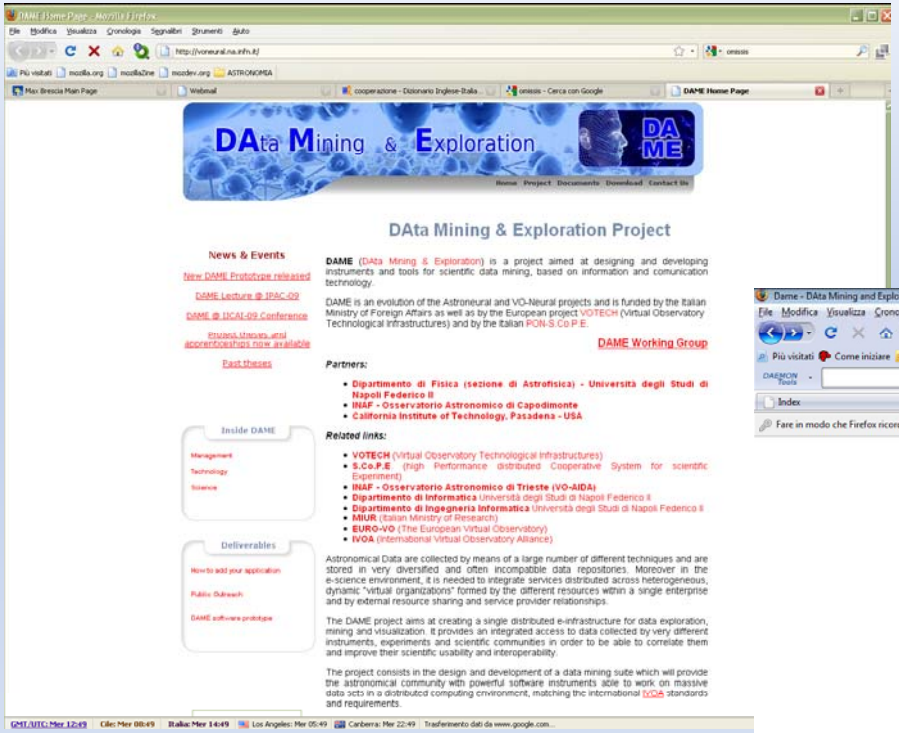
170 TeraByte storage

5 remote sites (2 in progress)

brescia@na.astro.it

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

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

Web application PROTOTYPE

brescia@na.astro.it

What is DAME

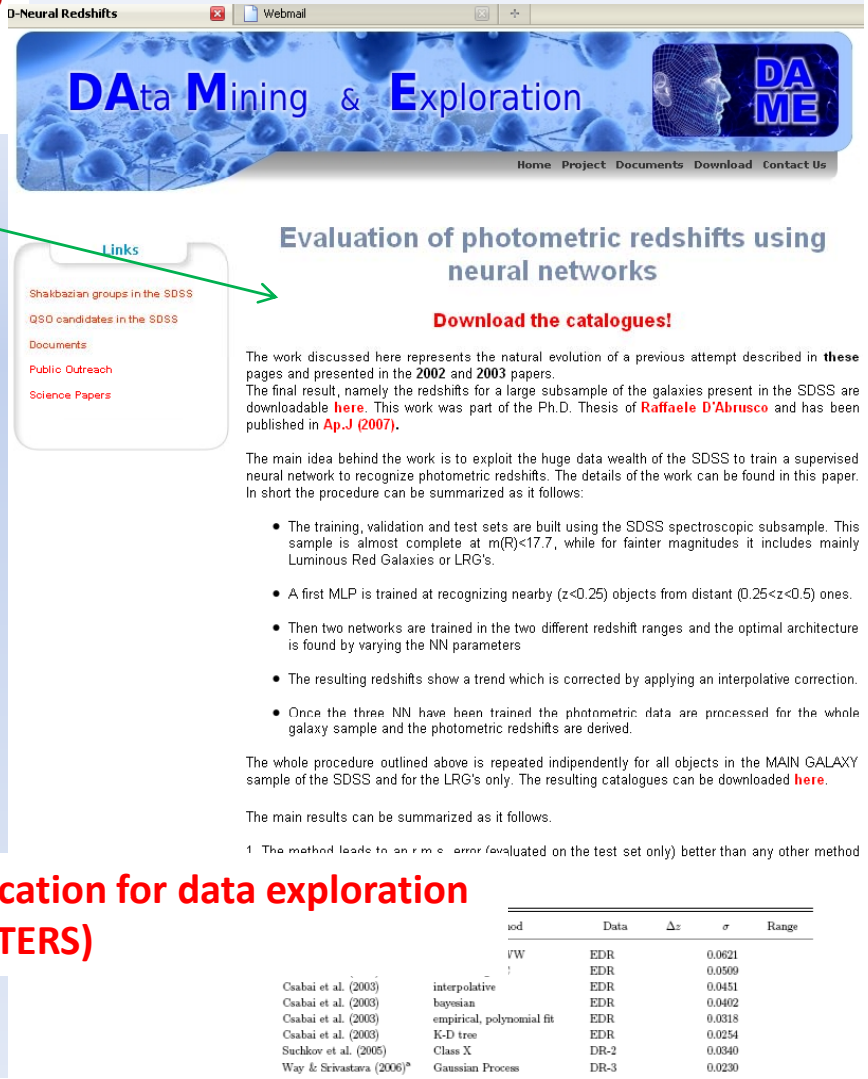
In parallel with the Suite R&D process, all data processing algorithms (foreseen to be plugged in) have been massively tested on real astrophysical cases.

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

Technical and management info

Documents

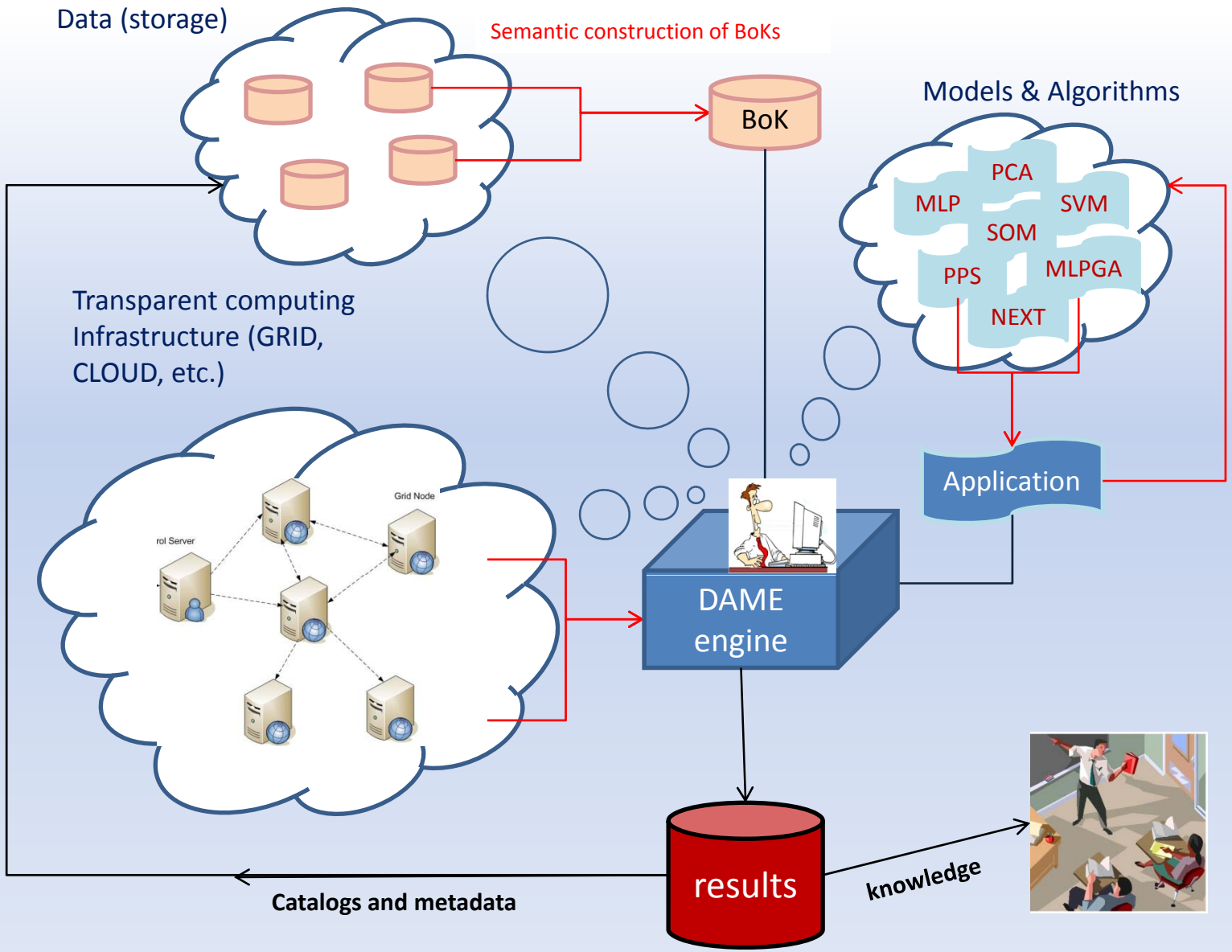
Science cases



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

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

DAME Work breakdown



brescia@na.astro.it

The DAME architecture

FRONT END
WEB-APPL.
GUI

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

XML

DATA MINING MODELS
Model-Functionality
LIBRARY RUN

clustering
regression
MLP

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

FRAMEWORK
WEB-SERVICE
Suite CTRL

servlet

CALL

DMPlugin
DM Functionalities
Classification, Regression, ...
DM Models
SVM, MLP, PPS, ...
DM Library wrappers
JNI, SWIG, ...
DM Libraries
libfann, libsvm, ...
Low Level Libraries
blas, lapack, gsl, ...

CALL

XML

DRIVER
FILESYSTEM &
HARDWARE I/F
Library

HW env virtualization;
Storage + Execution LIB
Data format conversion

REGISTRY &
DATABASE
USER &
EXPERIMENT
INFORMATION

Stand Alone
GRID
CLOUD

USER INFO
USER SESSIONS
USER EXPERIMENTS

brescia@na.astro.it

Two ways to use DAME - 1

FRONT END
WEB-APPL.
GUI

DATA MINING
MODELS
Model-Functionality
LIBRARY RUN
clustering
regression
MLP

FRAMEWORK
WEB-SERVICE
Suite CTRL
servlet

DRIVER
FILESYSTEM &
HARDWARE I/F
Stand Alone
GRID
CLOUD

REGISTRY &
DATABASE
USER &
EXPERIMENT
INFORMATION
USER INFO
USER SESSIONS
USER EXPERIMENTS

brescia@na.astro.it

Two ways to use DAME - 2

developer user

FRONT END
WEB-APPL.
GUI

FRAMEWORK
WEB-SERVICE
Suite CTRL

DMPLUGIN

DATA MINING
MODELS
Model-Functionality
LIBRARY RUN

clustering
regression
MLP

DRIVER
FILESYSTEM &
HARDWARE I/F

Stand Alone
GRID
CLOUD

REGISTRY &
DATABASE
USER &
EXPERIMENT
INFORMATION

USER INFO
USER SESSIONS
USER EXPERIMENTS

brescia@na.astro.it

Two ways to use DAME - 2

FRONT END
WEB-APPL.
GUI

developer user

DMPlugin

FRAMEWORK
ICE
Suite CTRL

DRIVER
FILESYSTEM &
HARDWARE I/F

Stand Alone

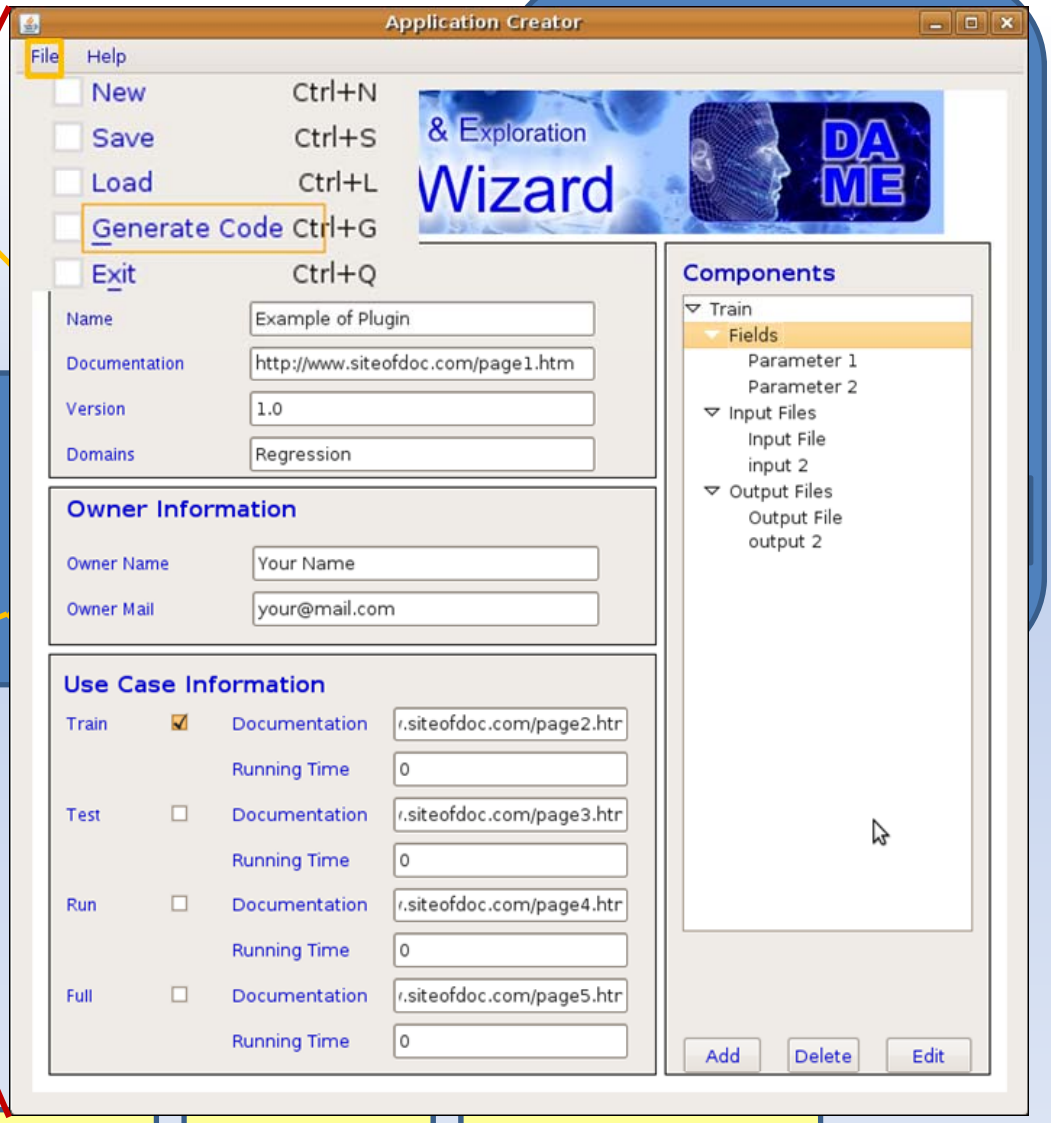
GRID

CLOUD

USER INFO

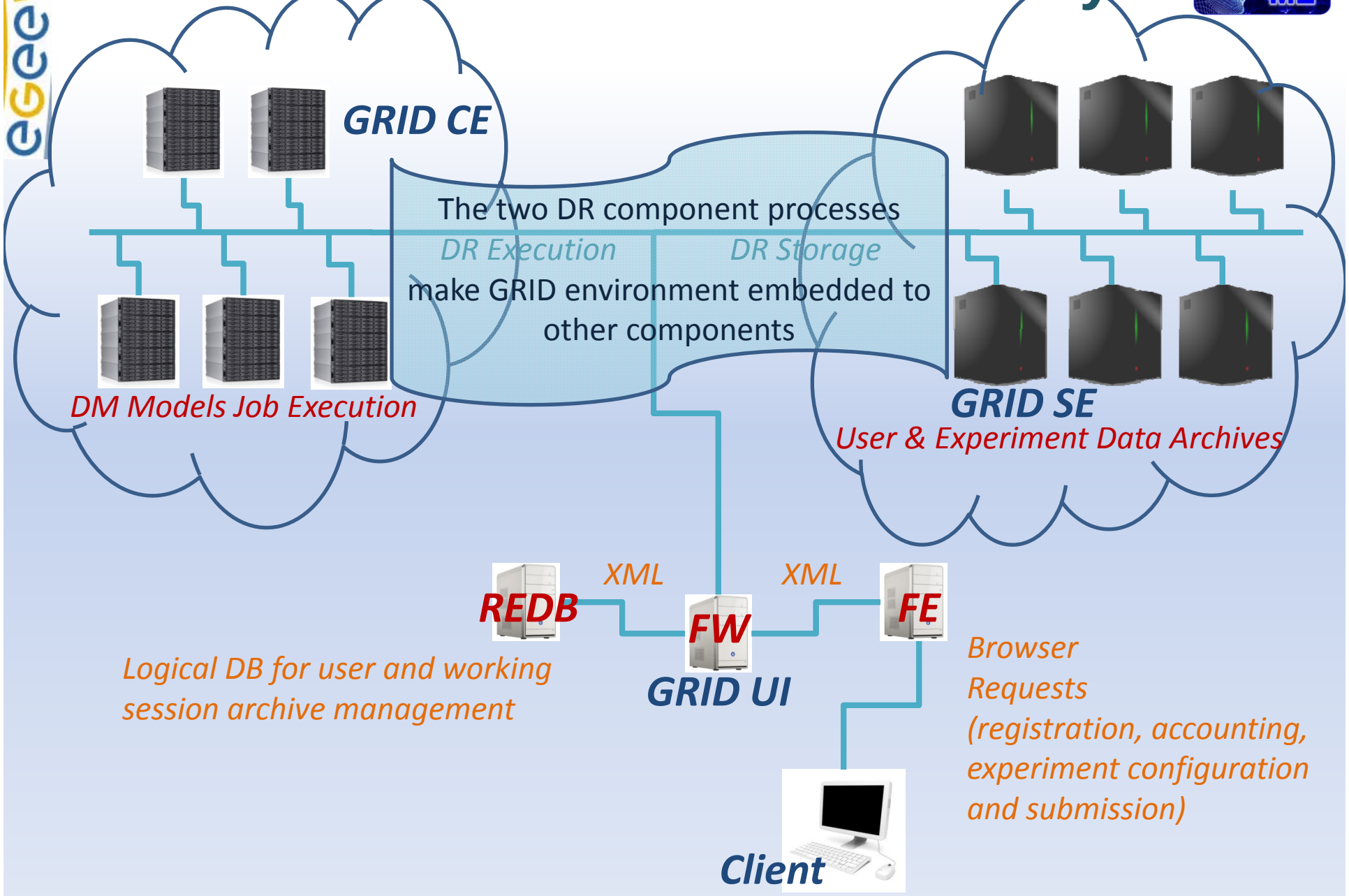
USER SESSIONS

USER EXPERIMENTS



brescia@na.astro.it

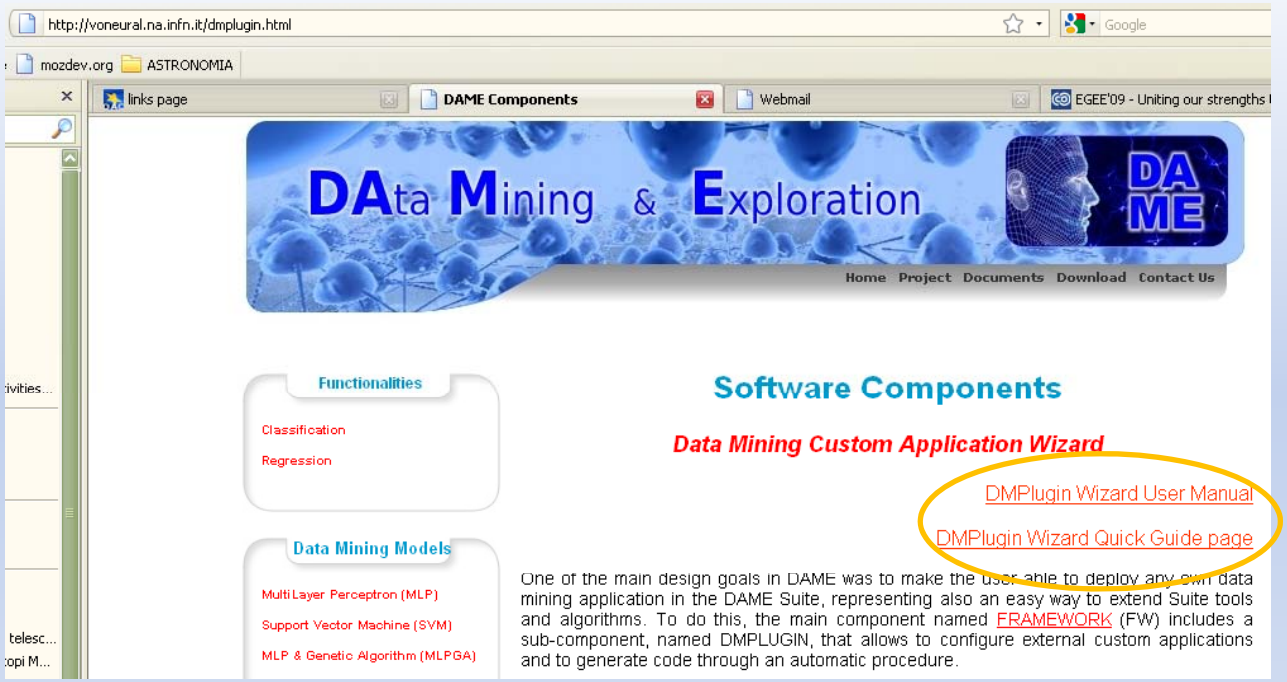
DAME on GRID – Scientific Gateway



brescia@na.astro.it

Coming soon…

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



- 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

brescia@na.astro.it