

**DA**ta **M**ining & **E**xploration



# **DAME: A Web Oriented Infrastructure For Scientific Data Mining And Exploration**

**Stefano Cavuoti<sup>(1)</sup>, Massimo Brescia<sup>(2)</sup>, Giuseppe Longo<sup>(1)</sup>,  
Mauro Garofalo<sup>(1)</sup>, Alfonso Nocella<sup>(1)</sup>.**

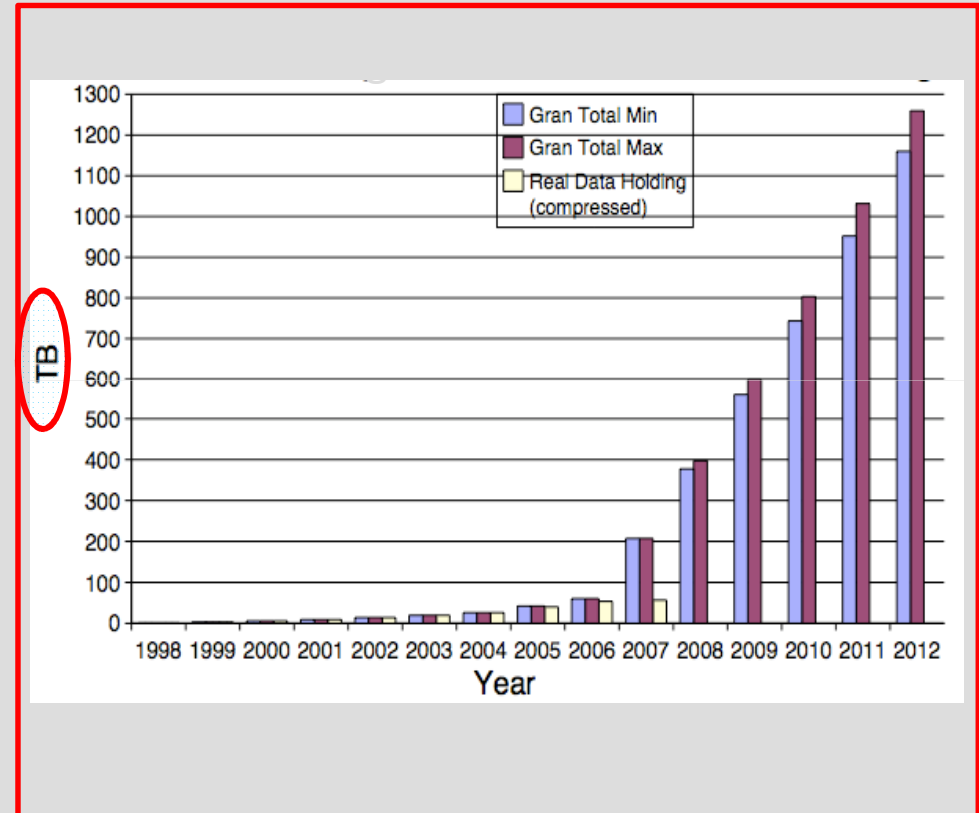
(1) Department of Physics – University Federico II – Napoli

(2) INAF – National Institute of Astrophysics – Capodimonte Astronomical Observatory - Napoli

# Astrophysics as a Data Rich Science



- Telescopes (ground-based and space-based, covering the full electromagnetic spectrum)
- Instruments (telescope/band dependent)
- Large digital sky surveys are becoming the dominant source of data in astronomy: ~ 10-100 TB/survey (soon PB), ~  $10^6$  -  $10^9$  sources/survey, many wavelengths...
- Data sets many orders of magnitude larger, more complex, and more homogeneous than in the past



# The General Astrophysical Problem



Due to new instruments and new diagnostic tools, the information volume grows exponentially

***Most data will never be seen by humans! (BLADE RUNNER)***

➔ The need for data storage, network, database-related technologies, standards, etc.

**Information complexity is also increasing greatly**

***Most knowledge hidden behind data complexity is lost***

➔ Most (all) empirical relationships known so far depend on 3 parameters ....  
Simple universe or rather human bias?

***Most data (and data constructs) cannot be comprehended by humans directly!***

➔ The need for data mining, KDD (Knowledge Discovery in Databases), data understanding technologies, hyper dimensional visualization, AI/Machine-assisted discovery



# Data Mining (KDD) as the Fourth Paradigm Of Science



The old traditional, "Platonic" view:

Pure Theory

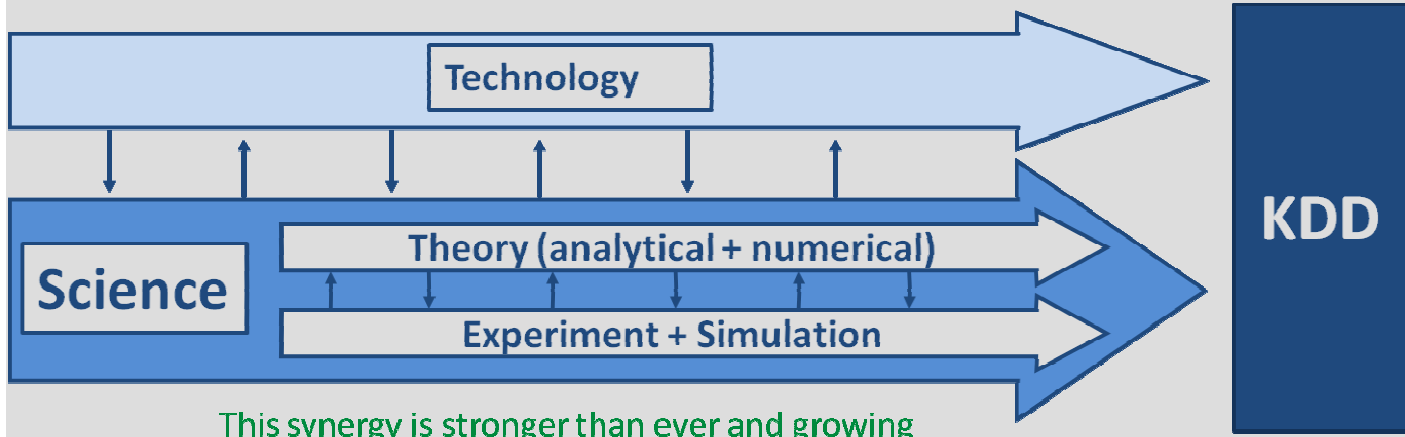


Experiment

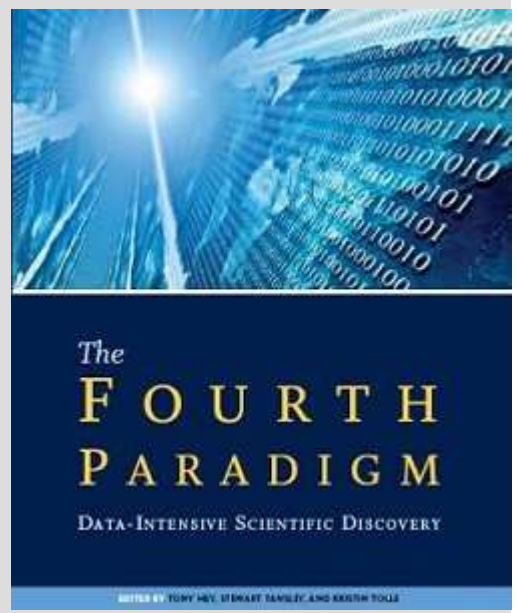


Technology & Practical Applications

The modern and realistic view when dealing with complex data sets:



This synergy is stronger than ever and growing



# The BoK's Problem



Limited number of problems due to limited number of reliable BoKs

## (BoK) Bases of knowledge

*(set of well known templates for supervised (training) or unsupervised (labeling) methods)*

### So far

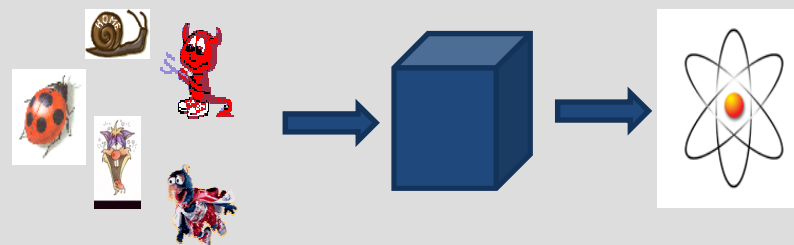
- Limited number of BoK (and of limited scope) available
- Painstaking work for each application (es. spectroscopic redshifts for photometric redshifts training)
- Fine tuning on specific data sets needed (e.g., if you add a band you need to re-train the methods)

**Bases of knowledge need to be built automatically from Vobs Data repositories**

**•There's a need of standardization and interoperability between data together with DM application**

**Community believes AI/DM methods are black boxes**

*You feed in something, and obtain patters, trends, i.e. knowledge...*



# The Choice Problem

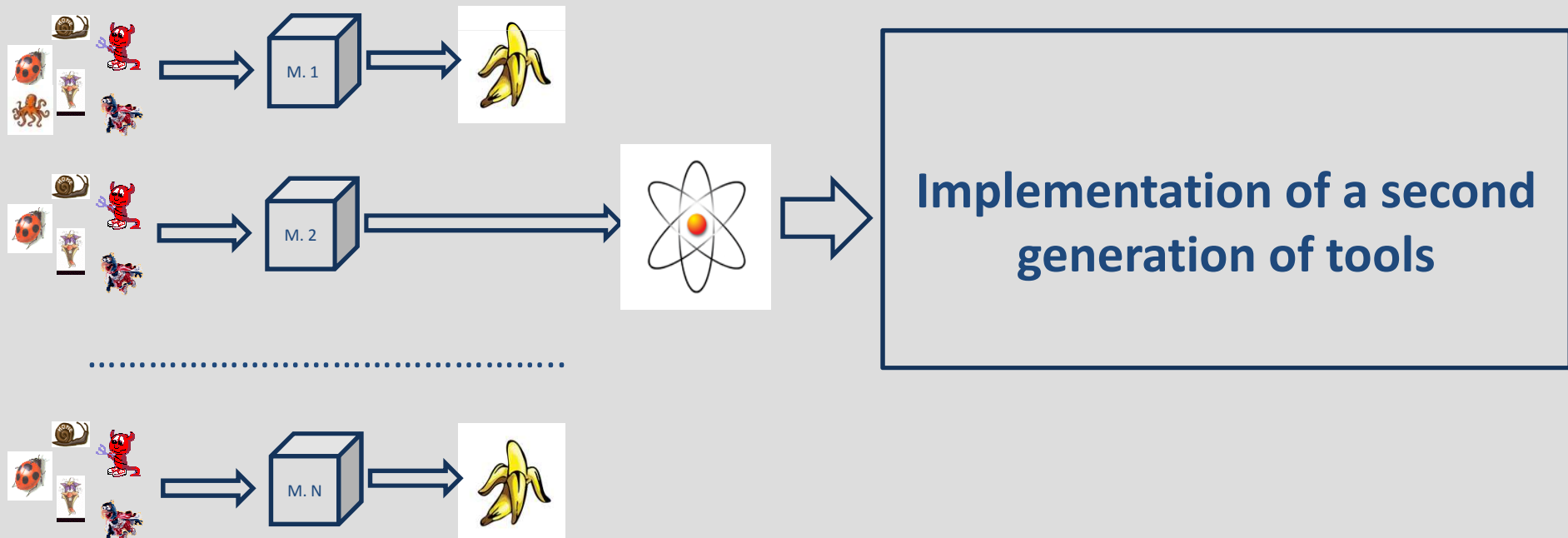


Exposed to a wide choice of algorithms to solve a problem, the r.m.s. astronomer usually panics and is not willing to make an effort to learn them ....

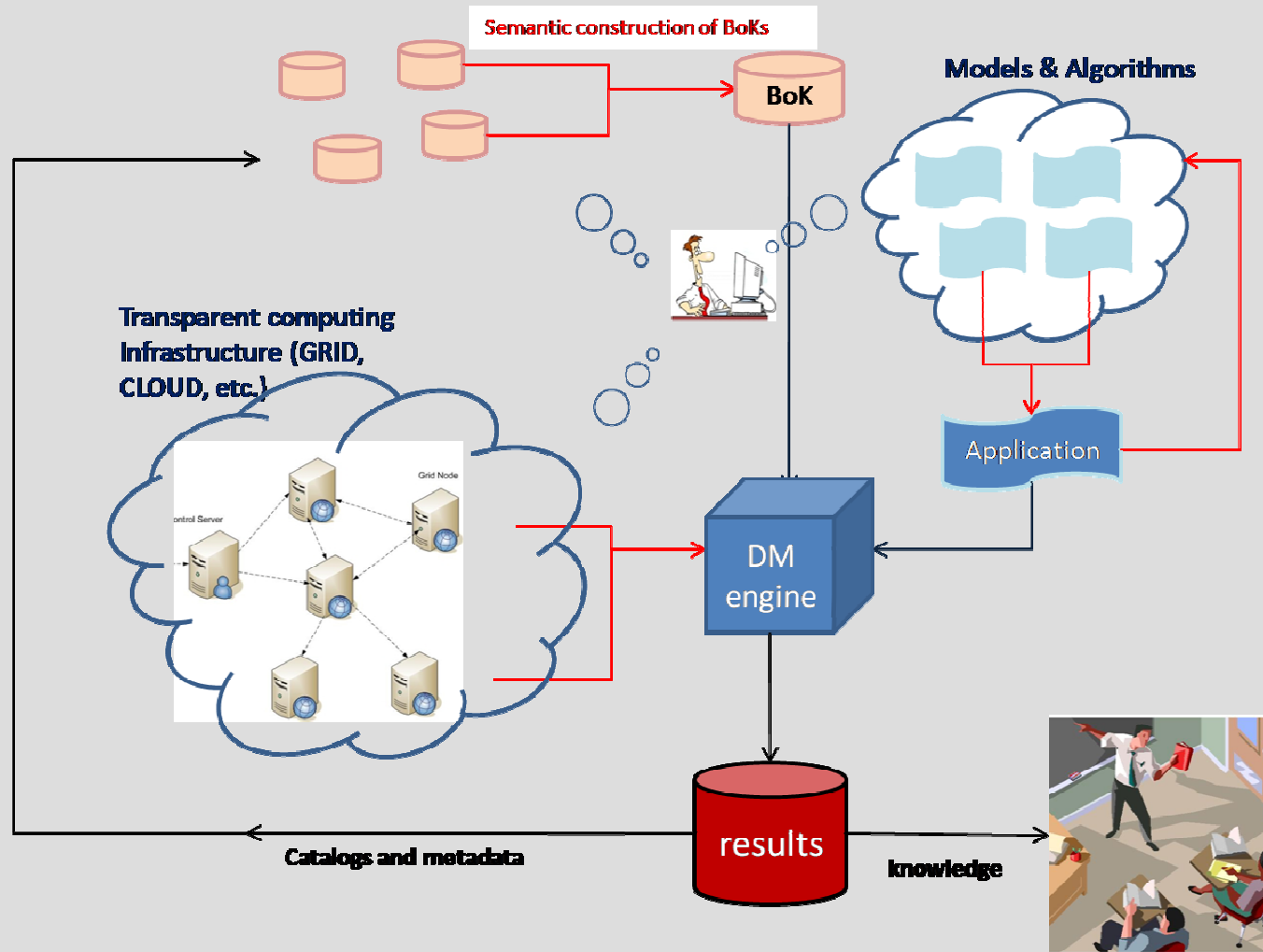
The r.m.s astronomer doesn't want to become a computer scientist or a mathematician  
(large survey projects overcome the problem)

Tools must run without knowledge of GRID/Cloud no personal certificates, no deep understanding of the DM tool etc.

Allow each astronomer to build his black box using own algorithm to run on the infrastructure without knowledge about the infrastructure



# Effective DM process break-down





# The Black box Infrastructure



In this scenario DAME (Data Mining & Exploration) project, starting from astrophysics requirements domain, has investigated the Massive Data Sets (MDS) exploration by producing a taxonomy of data mining applications (hereinafter called functionalities) and collected a set of machine learning algorithms (hereinafter called models).

This association functionality-model represents what we defined as simply "use case", easily configurable by the user through specific tutorials. At low level, any experiment launched on the DAME framework, externally configurable through dynamical interactive web pages, is treated in a standard way, making completely transparent to the user the specific computing infrastructure used and specific data format given as input.

So the user doesn't need to know anything about GRID, Cloud or what else.



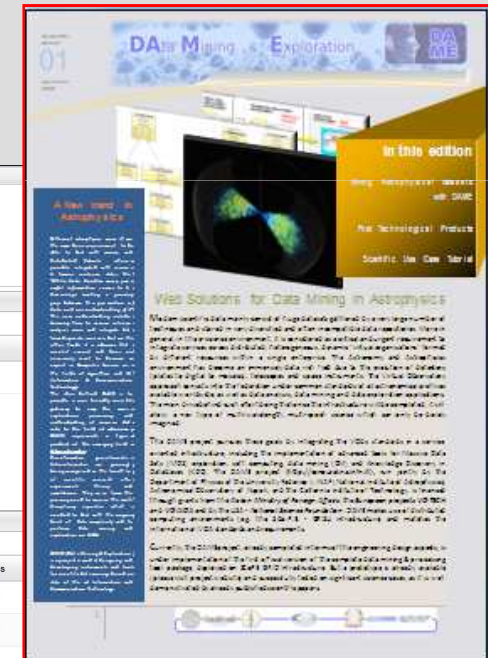


# What DAME is



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

<http://dame.dsf.it/>  
Technical and management info  
Documents  
Science cases  
Newsletter



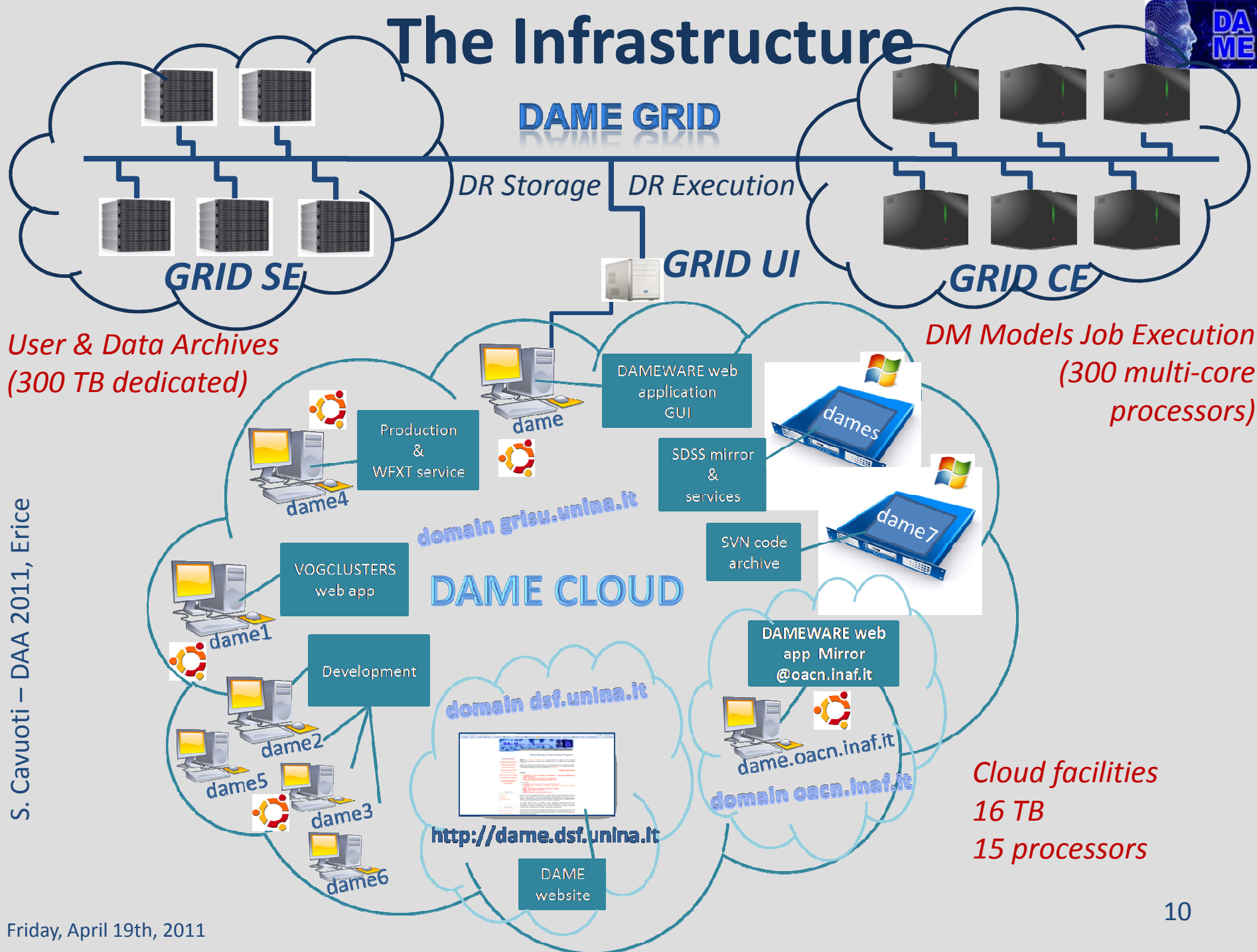
<http://dame.na.astro.it:8080/MyDameFE/>  
Web application Beta Version

Workspace	File	Type	Last Access
experiment	provia.dat	ascii	2010-12-17

Experiment	Status	Last Access
dasdsadas	failed	2011-03-07
dsdsda	failed	2011-03-07
asdsdasda	ended	2011-03-07
asdsasd	ended	2011-03-07
fridf	ended	2011-03-07

# The Infrastructure



S. Cavuoti – DAA 2011, Erice

# 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 INFN GRID 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



**The SCoPE Data Center**  
33 Racks (of which 10 for Tier2 ATLAS)  
304 Servers for a total of 2.432 procs  
300 Terabytes storage (100 for DAME)  
5 remote sites (2 in progress)

# The Available Services



## **DAMEWARE Web Application Resource**

Main service providing via browser a list of algorithms and tools to configure and launch experiments as complete workflows (dataset creation, model setup and run, graphical/text output):

- *Functionalities: Regression, Classification, Image Segmentation, Multi-layer Clustering;*
- *Models: MLP+BP, MLP+GA, SVM, MLP+QNA, K-Means, PPS, SOM, NEXT-II;*

## **SDSS (Sloan Digital Sky Survey)**

Local mirror website hosting a complete SDSS Data Archive and Exploration System;

## **WFXT (Wide Field X-Ray Telescope) Transient Calculator**

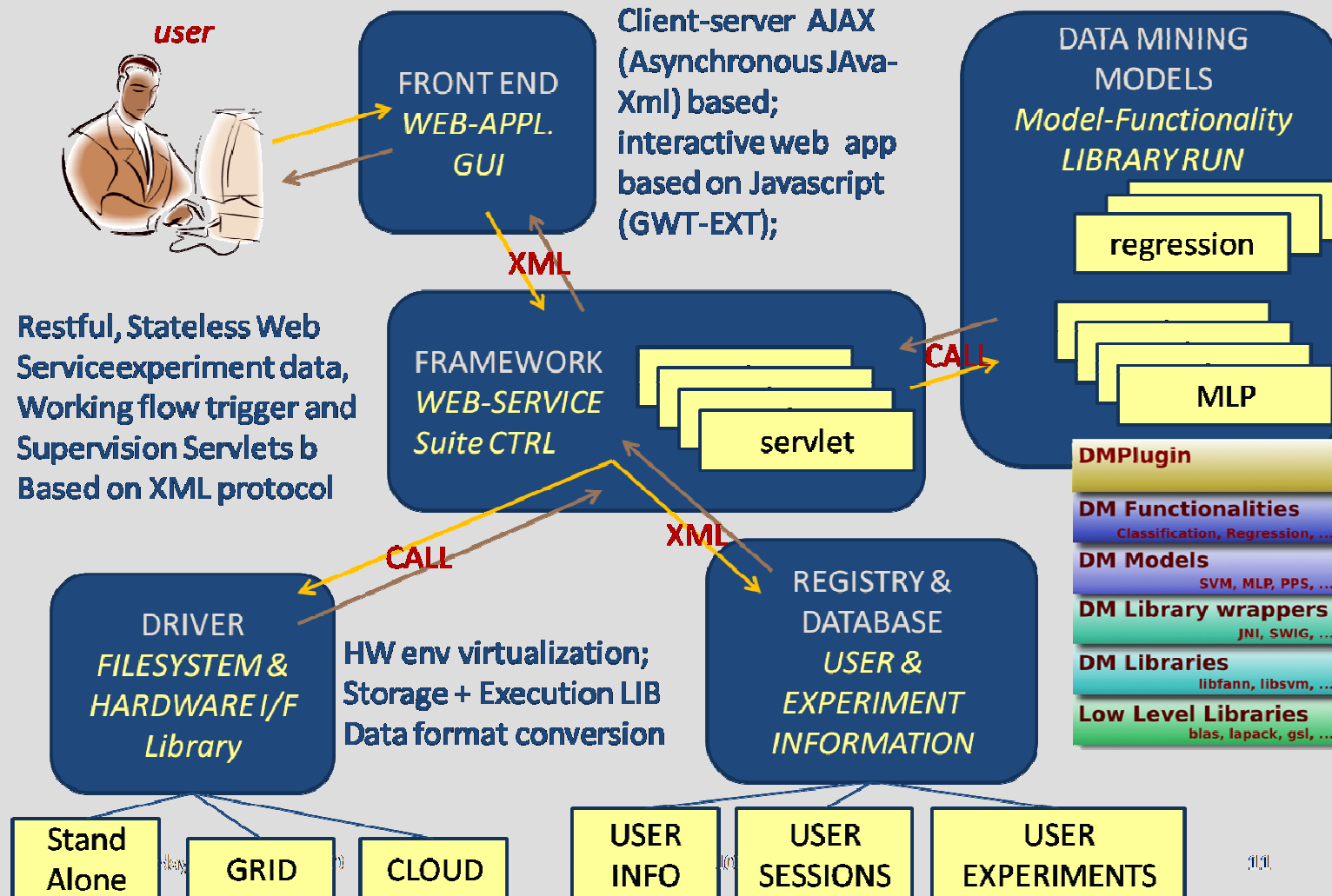
Web application to estimate the number of transient and variable sources that can be detected by WFXT within the 3 main planned extragalactic surveys, with a given significant threshold;

## **VOGCLUSTERS**

VO-compliant Web Application for data and text mining on globular clusters;



# The DAME Architecture



# Framework

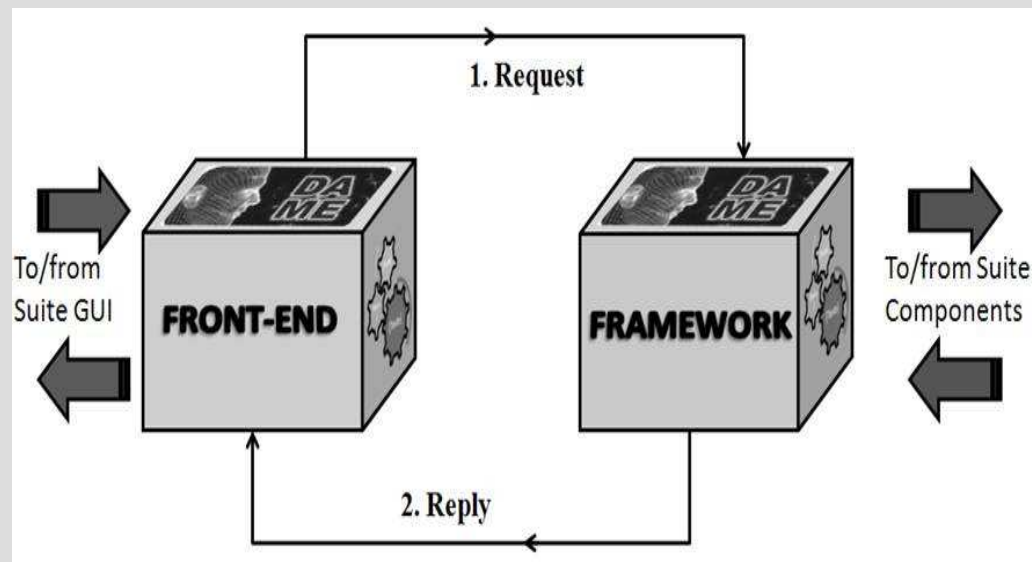


The FW is the core of the Suite. It handles all communication flow from/to FE (i.e. the end user) and the rest of the components, in order to register the user, to show user working session information, to configure and execute all user experiments, to report output and status/log information about the applications running or already finished. One of the most critical factors of the FW component is the interaction of a newly created experiment with the GRID environment. The FW needs to create and configure the plug-in (hereinafter called DMPlugin) associated to the experiment. After the DMPlugin is configured the DR component needs to run the experiment by calling the run method of the plug-in. When executed on the GRID, the process needs to migrate on a Worker Node (WN). To implement this migration we've chosen to serialize the DMPlugin in a file. Serialization is a process of converting an object into a sequence of bits so that it can be stored on a storage medium. Our tests on the GRID environment indicates that this solution works fine and that the jdl file needed to manage the whole process is very simple.

# Front End



The component FE includes the main GUI (Graphical User Interface) of the Suite and it is based on dynamical WEB pages, rendered by the Google Web Toolkit (GWT), able to interface the end users with the applications, models and facilities to launch scientific experiments. The interface foresees an authentication procedure which redirects the user to a personal session environment, collecting uploaded data, check experiment status and driven procedures to configure and execute new scientific experiments, using all available data mining algorithms and tools. From the engineering point of view, the FE is organized by means of a bidirectional information exchange, through XML files, with the component FW, suite engine component.



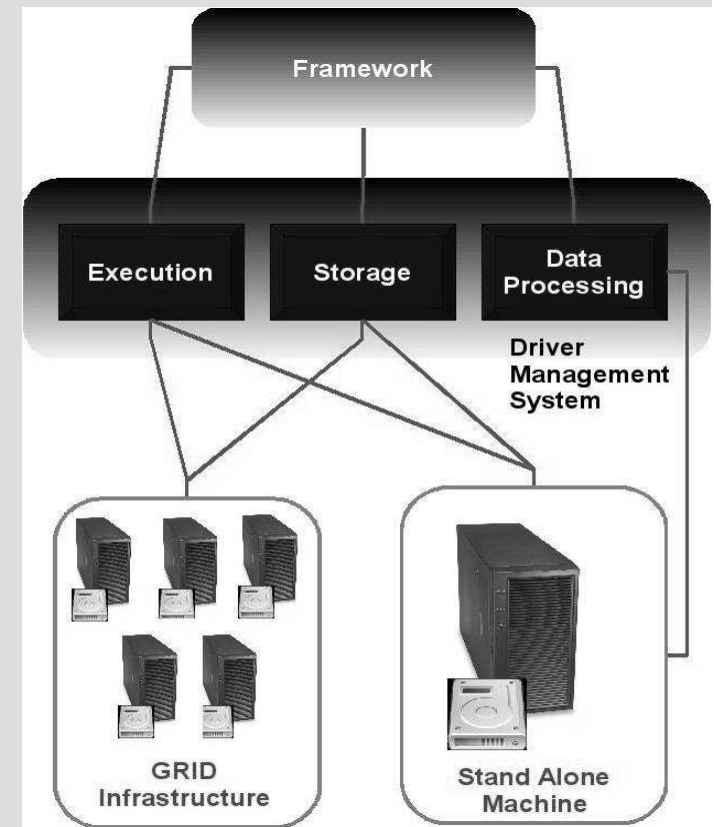


# Driver



The component DR is the package responsible of the physical implementation of the HW resources handled by other components at a virtual level. It permits the abstraction of the real platform (HW environment and related operative system calls) to the rest of Suite software components, including also I/O interface (file loading/storing), user data intermediate formatting and conversions (ASCII, CSV, FITS, VO-TABLE), job scheduler, memory management and process redirection (Fig. 3).

More in detail, a specific sub-system of the DR component, called DRiver Management System (DRMS), has been implemented to delegate at runtime the choice of the computing infrastructure should be selected to launch the experiment.

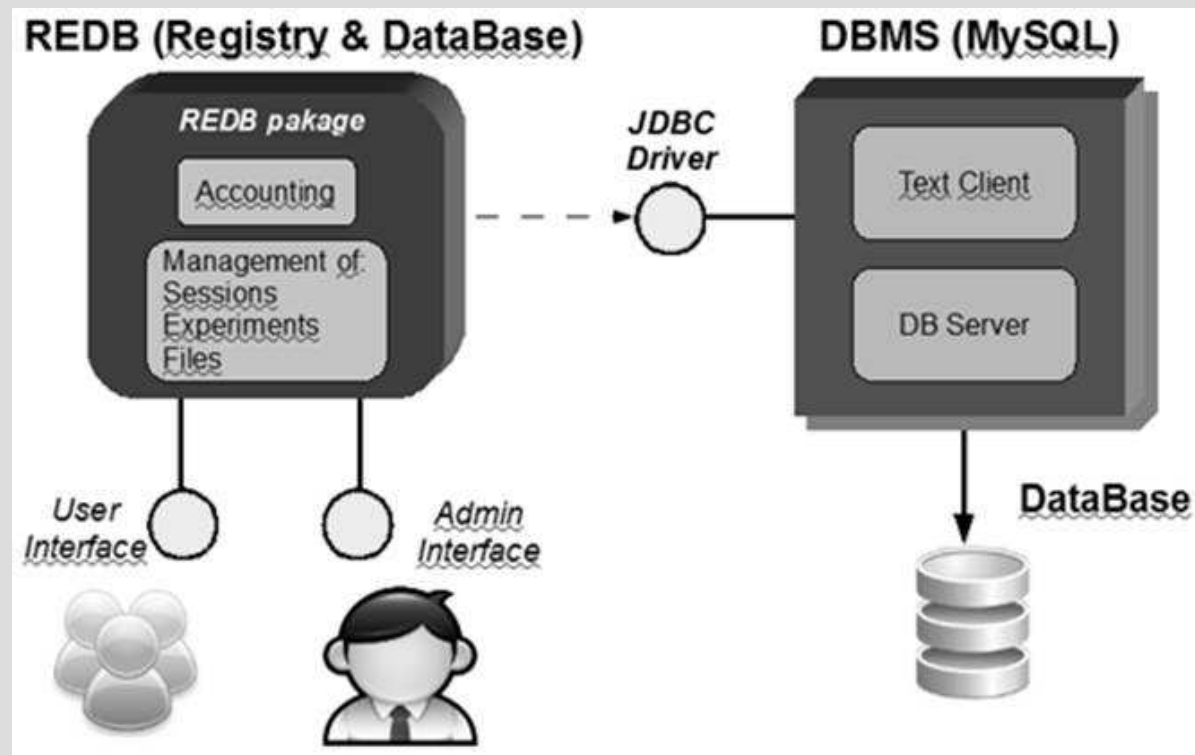


# Registry and Data Base



The component REDB is the base of knowledge repository for the Suite. It is a registry in the sense that contains all information related to user registration and accounts, his working sessions and related experiments.

It is also a Database containing information about experiment input/output data and all temporary/final data coming from user jobs and applications.

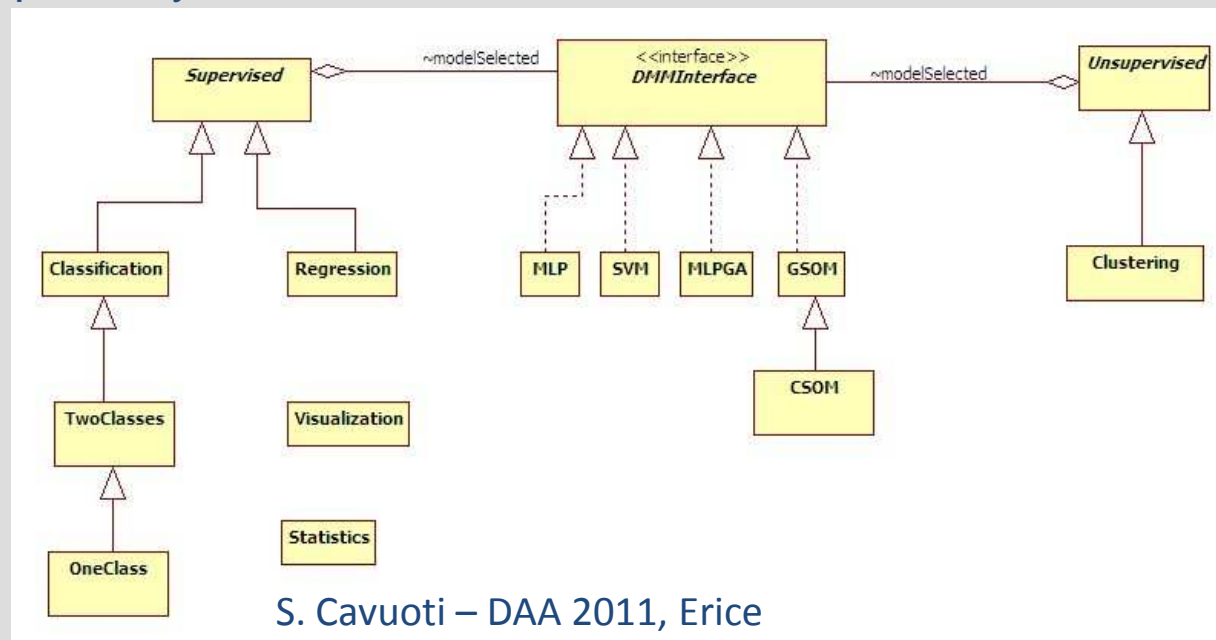


# Data Mining Models



The component DMM is the package implementing all data processing models and algorithms available in the Suite. They are referred to supervised/unsupervised models, coming from Soft Computing, Self-adaptive, Statistical and Deterministic computing environments. It is structured by means of a package of libraries (Java API) referred to the following items:

- Data mining models libraries (Multi Layer Perceptron, Support Vector Machine, Genetic Algorithms, Self Organizing Maps, etc...);
- Visualization and Statistical tools;
- List of functionalities (Classification, Regression, Clustering, etc...);
- Custom libraries required by the user;



# Scientific Results



**Globular Cluster Search:** The study of Globular Clusters populations in external galaxies requires the use of wide-field, multi-band photometry. However to minimize contamination problems and to measure some of the GC properties, such as sizes and structural parameters (core radius, concentration, binary formation rates) high-resolution data are required as well, which are only available through the use of space facilities (i.e. HST). The use of single-band HST data reduces the cost (in terms of observing time) of such studies, and can be eventually integrated with ground-based photometry in other bands to obtain the required color information. In this project we intend to show that even the use of single band photometry can yield very complete datasets with low contamination, through the use of Neural Network (Multi Layer Perceptron trained by a Quasi Newton rule) algorithm. This approach will minimize the observing time requirements, thus allowing to extend such studies to large areas and to the outskirts of nearby galaxies, thus minimizing the observational biases in studies where a very complete dataset is required, such as the study of Low Mass X-ray Binaries in GCs.

*A paper on the results is in preparation*

**Photometric redshifts for the SDSS galaxies:** It makes use of a nested chain of MLP (Multi Layer Perceptron) and allowed to derive the photometric redshifts for ca. 30 million SDSS galaxies with an accuracy of 0.02 in redshift. This result which has appeared in the *Astrophysical Journal*, was also crucial for a further analysis of low multiplicity groups of galaxies (Shakhbazian) in the SDSS sample; *D'Abrusco, R. et al., 2007. Mining the SDSS archive I. Photometric Redshifts in the Nearby Universe. Astrophysical Journal, Vol. 663, pp. 752-764*

**Search for candidate quasars in the SDSS:** The work was performed using the PPS (Probabilistic Principal Surfaces) module applied to the SDSS and SDSS+UKIDS data. It consisted in the search for candidate quasars in absence of a priori constrains and in a high dimensionality photometric parameter space,

*D'Abrusco, R. et al., 2009. Quasar Candidate Selection in the Virtual Observatory era. Under press in MNRAS*

**AGN classification in the SDSS:** Using the GRID-S.Co.P.E. to execute 110 jobs on 110 WN, the SVM model is employed to produce a classification of different types of AGN using the photometric data from the SDSS and the base of knowledge provided by the SDSS spectroscopic subsamples.

*A paper on the results is in preparation*

# Conclusion



- we have designed and provided the DAME infrastructure to empower those who are not machine learning experts to apply these techniques to the problems that arise in daily working life.
- DAME project comes out as an astrophysical data exploration and mining tool, originating from the very simple consideration that, with data obtained by the new generation of instruments, we have reached the physical limit of observations (single photon counting) at almost all wavelengths.
- If extended to other scientific or applied research disciplines, the opportunity to gain new insights on the knowledge will depend mainly on the capability to recognize patterns or trends in the parameter space, which are not limited to the 3-D human visualization, from very large datasets. In this sense DAME approach can be easily and widely applied to other scientific, social, industrial and technological scenarios.
- Our project is in the beta version and we are providing on each release new functionalities and models.
- First scientific test results confirm the goodness of the theoretical approach and technological strategy.