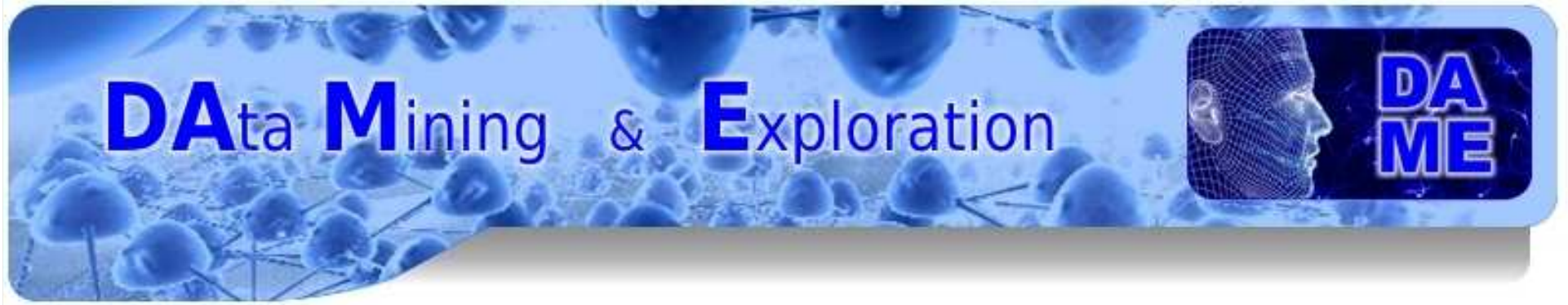


DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases

Massimo Brescia[(1)], Giuseppe Longo[(2)], S. G. Djorgovski[(3)], M. Castellani[(4)], Stefano Cavuoti[(2)], Raffaele D'Abrusco[(5)], Omar Laurino[(6)], Riccardo Smareglia[(6)]

(1) INAF – Astronomical Observatory of Capodimonte - Napoli
(2) Department of Physics – University Federico II – Napoli
(3) CALTECH, Pasadena California, USA
(4) INAF – Astronomical Observatory of Roma
(5) Harvard-Smithsonian Center for Astrophysics, Cambridge MA, USA
(6) INAF – Astronomical Observatory of Trieste

The General Astrophysical Problem

Due to new instruments and new diagnostic tools, the information volume grows exponentially.
We have reached the physical limit of observations (single photon counting) at almost all wavelenght…

➡ ***Most data will never be seen by humans!***
need for data storage, network, database-related technologies, standards, etc.

<u>Information complexity is also increasing greatly</u>

➡ ***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?

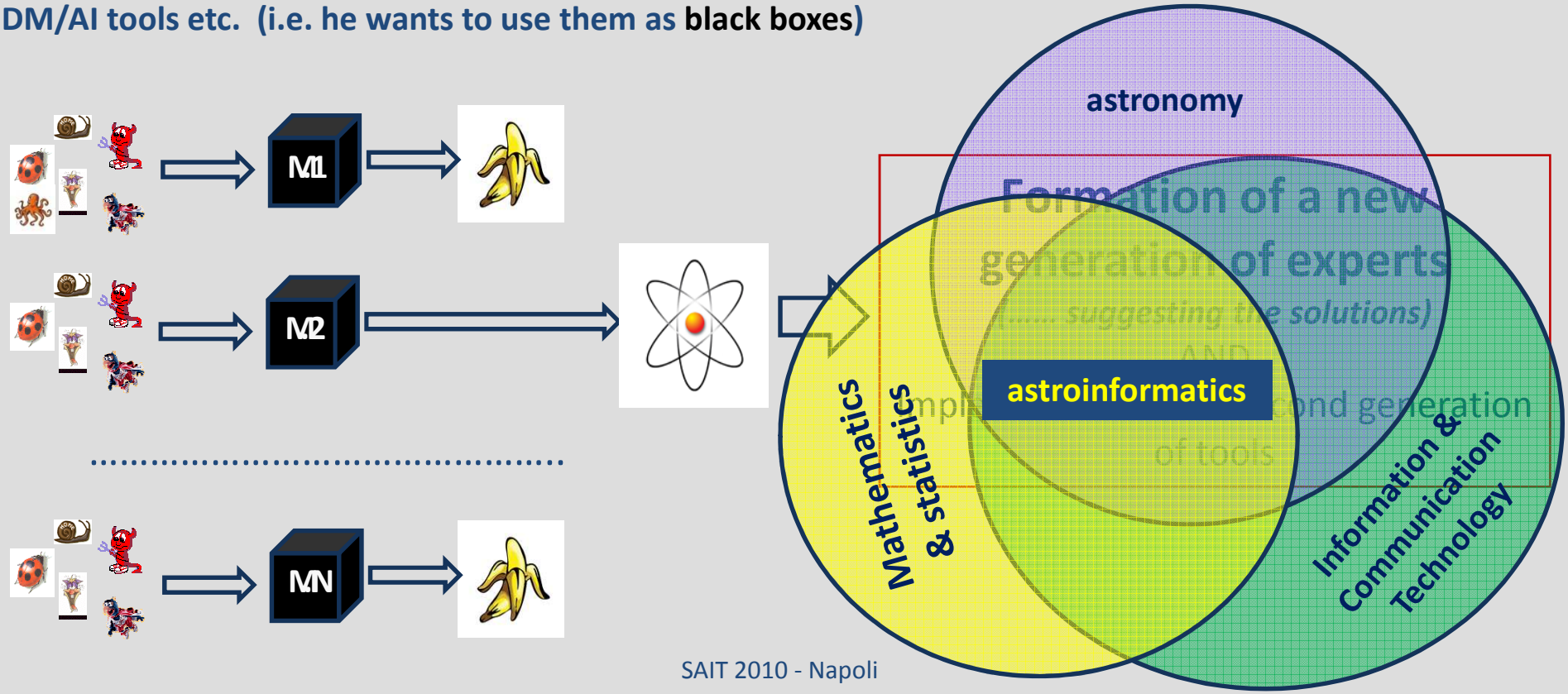
➡ ***Most data (and data constructs) cannot be comprehended by humans directly!***
The need for data mining, KDD (Knowledge Discovery in Databases), data understanding technologies, hyper dimensional visualization, AI/Machine-assisted discovery

The Choice Problem

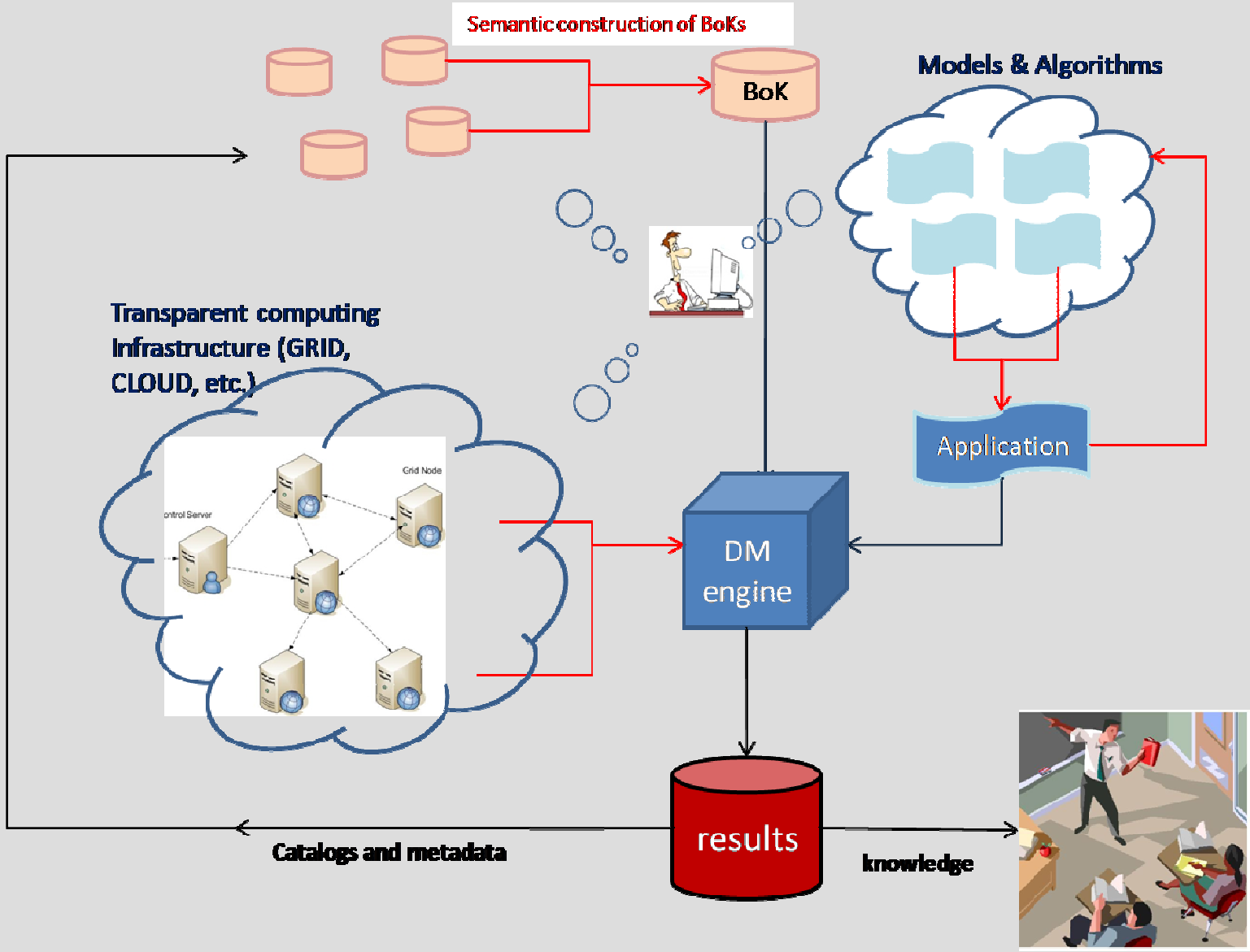
- Limited number of BoK (Bases of Knowledge) templates available
- The interoperability between data was solved by VO. But it remains still open about applications
- Fine tuning on specific data sets needed (e.g., if you add a band you need to re-train the methods)

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
Tools must run without knowledge of GRID/Cloud, authority certificates, no deep understanding of the DM/AI tools etc. (i.e. he wants to use them as **black boxes**)



DM process break-down



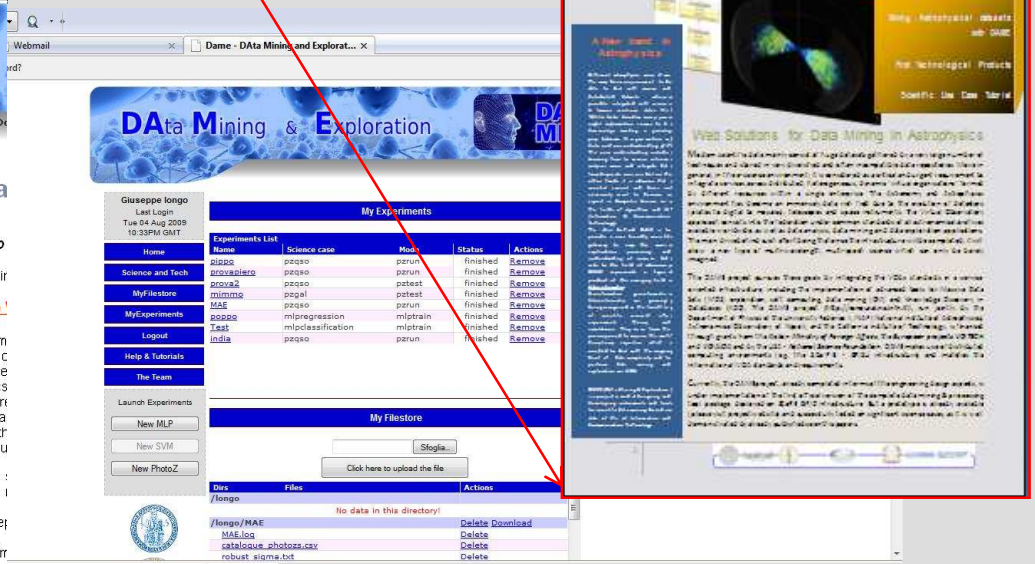
What DAME is

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. It is funded by Italian Foreign Affairs Minister, PON S.Co.P.E. and EURO-VO Consortia.

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

Main user entry point to:

- **Applications**
- **Technical/Management info**
- **Documents**
- **Science cases**
- **Newsletter**



The Available Services

DM WEB Application Prototype

Simplified Web Application providing via browser tools to configure and launch DM experiments

SDSS (Sloan Digital Sky Survey)

Local mirror website hosting a complete SDSS Data Archive and Management 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;

Coming soon (under commissioning):

DM WEB Application Suite (the evolution of prototype)

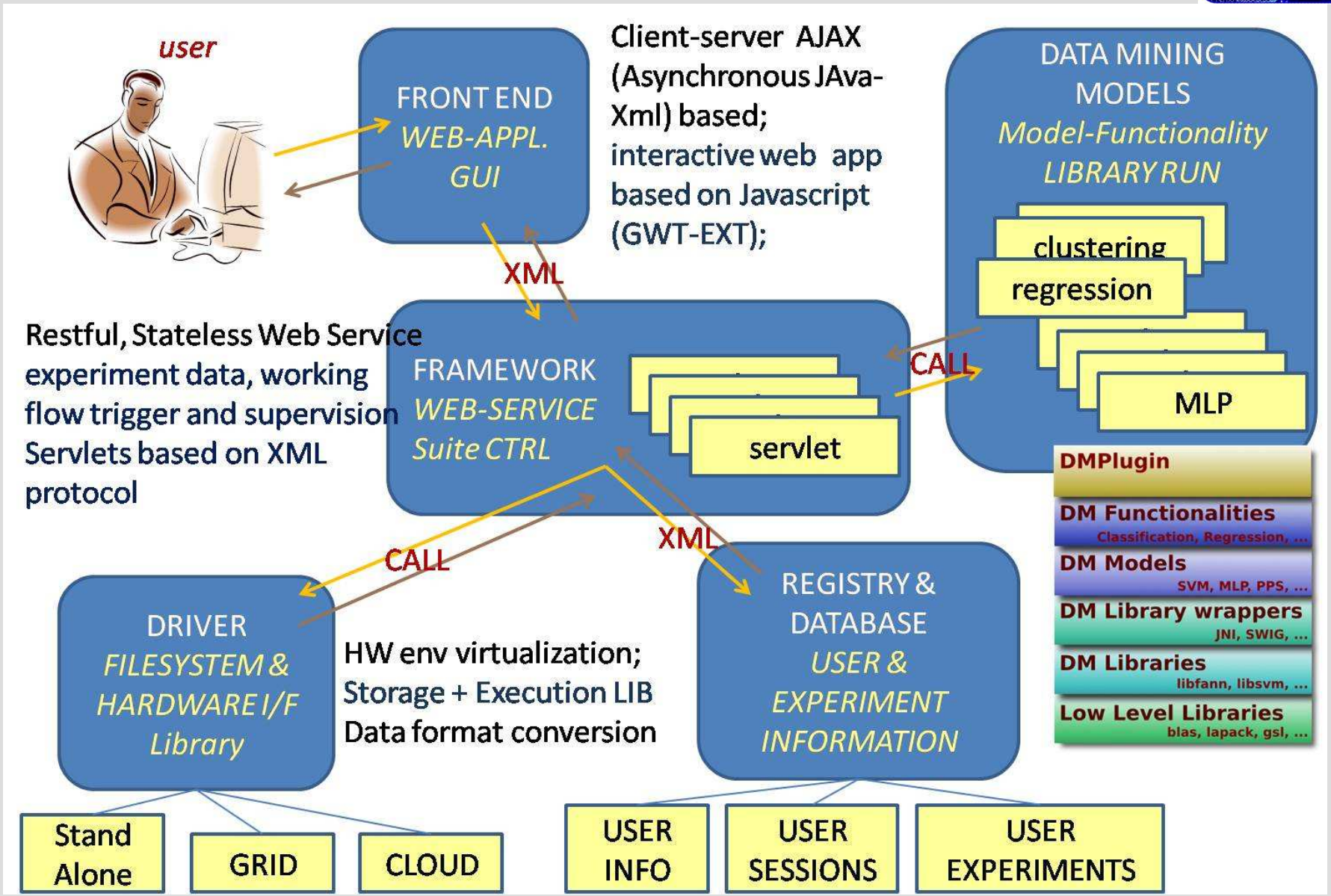
Main service providing via browser a list of algorithms and tools to configure and launch experiments as complete workflows (dataset creation, model setup and run, graphical/text output):

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

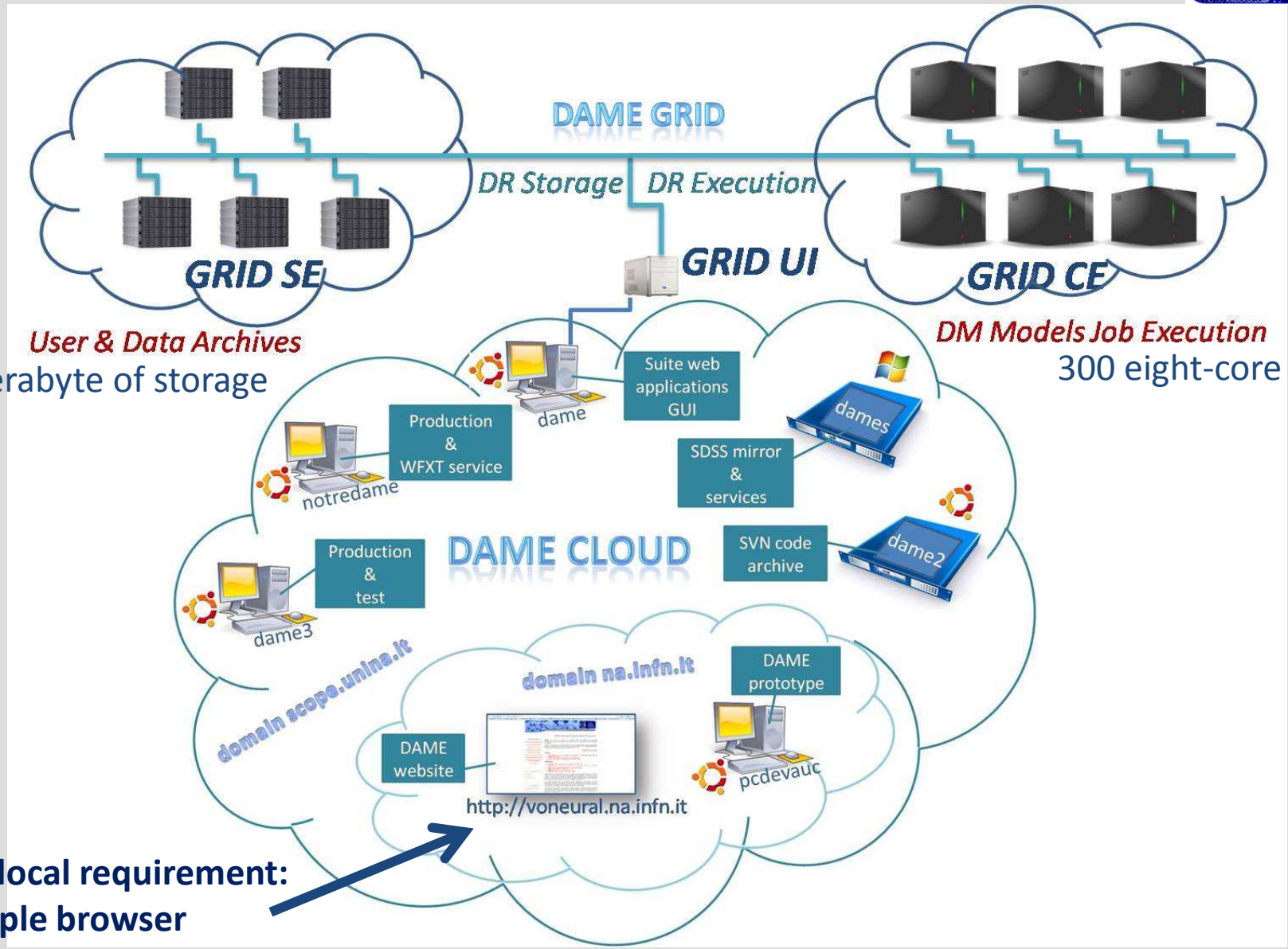
VOGCLUSTERS

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

The DAME SW Architecture



The Available Resources



100 Terabyte of storage

300 eight-core servers

User local requirement:
a simple browser
(NOTHING ELSE!)

First Scientific Results

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

@ VO-Day … in Tour

http://wwwas.oats.inaf.it/voday

From December 2009 to April 2010:

- **12 Sessions + 1 Videoconf with TNG,**
- **Touch all city with INAF structures,**
- **6 tutors for each session (11 people involved)**

First Day:

09:30 - 09:35: Introduction (***R. Smareglia***)
09:35 - 09:50: The Virtual Observatory: an Overview (*R.Smareglia*)
09:50 - 10:15: A Virtual Tour of the Virtual Observatory (*P. Manzato*)
10:15 Use Case 1: Confirmation of a Supernova candidate (*G. Iafrate*)
11:50 Use Case 2: Searching for Data available for the bright galaxy M51 (*M. Molinaro*)
15:00 Use Case 3: Photometric redshifts with DaME (*M. Brescia, O. Laurino, R. D'abrusco*)
17:00 AIDA WP5 outreach dissemination. (*G. Iafrate*)

Second Day):

09:30 Use Case 4: Data Extraction from Multidimensional Dataset (*A. Costa*)
11:30 Discussion / Feedback

EURO VO AIDA Astronomical Infrastructure for Data Access

DA ME

Catania

Teramo

Padova

Napoli

EURO VO AIDA Astronomical Infrastructure for Data Access

DA ME

Poster @ this Congress available

VO — day … in Tour

R. Smareglia , U. Becciani, P. Benvenuti, M. Brescia , A. Costa, B. Garilli, C. Gheller, G. Iafrate, O. Laurino, G. Longo, P. Manzato, M. Molinaro, P. Padovani, F. Pasian, M. Ramella
INAF – OATs, INAF- OACt, ESO, Univ Padova, Univ Napoli, CINECA

Abstract:

A day-and-a-half course devoted to the VO, called VOday, was been therefore organized for the Italian community, in the framework of the Euro VO-AIDA project and of the coordination initiatives at the international (IVOA) and national (VOs.it) levels. The course was held in all INAF cities where there are INAF personnel, with the support of INAF-OATs and INAF-SI, and in collaboration with some the INAF structures, CINECA and some Italian Universities. The goal of the workshop was to expose to the astronomers to the variety of VO tools and services available today so that they can incorporate them efficiently in their everyday research activities

Web Page: http://wwwas.oats.inaf.it/voday

@ VO-Day … in Tour

http://wwwas.oats.inaf.it/voday

Registered: 272

➔more than ¼ of INAF research staff

Attendant: 244

Evaluation Form: 176 (forms are available at VO-day pages)

- About 70% already known VO as name (mainly they know VO tools but without using them)
- Several People request more specific tutorials on VO tools and how to publish own data on VO

Level of interest	no	poor	enough	normal	high
DAME Use Case	1	4	21	60	77

Level of Hardness	easy	normal	hard
DAME Use Case	29	111	22

Conclusion

• We have designed the DAME infrastructure to empower those who are not machine learning experts to apply these techniques and who have not proper resources to make own scientific experiments

• One of the main goals of DAME is to contribute to the full interoperability between data (already obtained within IVOA) and applications (by following the 4th paradigm of Science)

• 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

• By its nature, DAME is an "open and incremental project", offering intrinsic educational and expertise formation opportunities in the Astrophysics and ICT research fields

• Our project has recently passed the R&D phase, *de facto* entering in the commissioning step

• First scientific test results (plus the demonstration to the community with VO-DAYs) confirm the goodness of the theoretical approach and technological strategy

Thanks to all DAME contributors!