

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



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

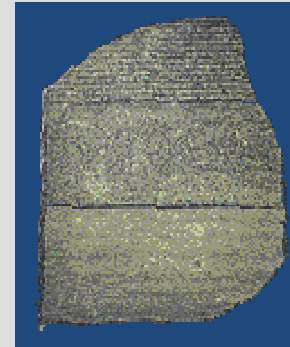
Massimo Brescia<sup>(1)</sup>, Giuseppe Longo<sup>(2)</sup>, S. G. Djorgovski<sup>(3)</sup>, M. Castellani<sup>(4)</sup>,  
Stefano Cavuoti<sup>(2)</sup>, Raffaele D'Abrusco<sup>(5)</sup>, Omar Laurino<sup>(6)</sup>, Riccardo Smareglia<sup>(6)</sup>

- (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 wavelength...



***Most data will never be seen by humans!***

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

# The Choice Problem

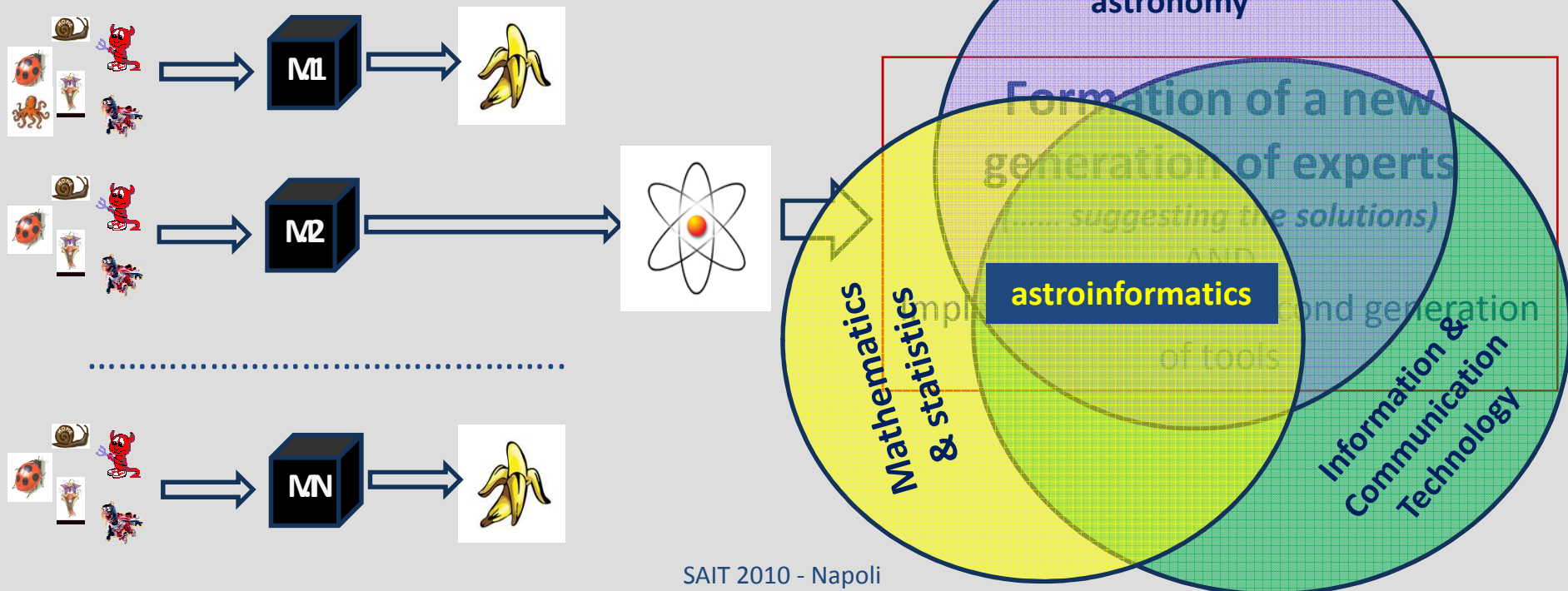


- Limited number of BoK (Bases of Knowledge) templates available
- **The interoperability between data was solved by VO. But it remains still open about applications**
- Fine tuning on specific data sets needed (e.g., if you add a band you need to re-train the methods)

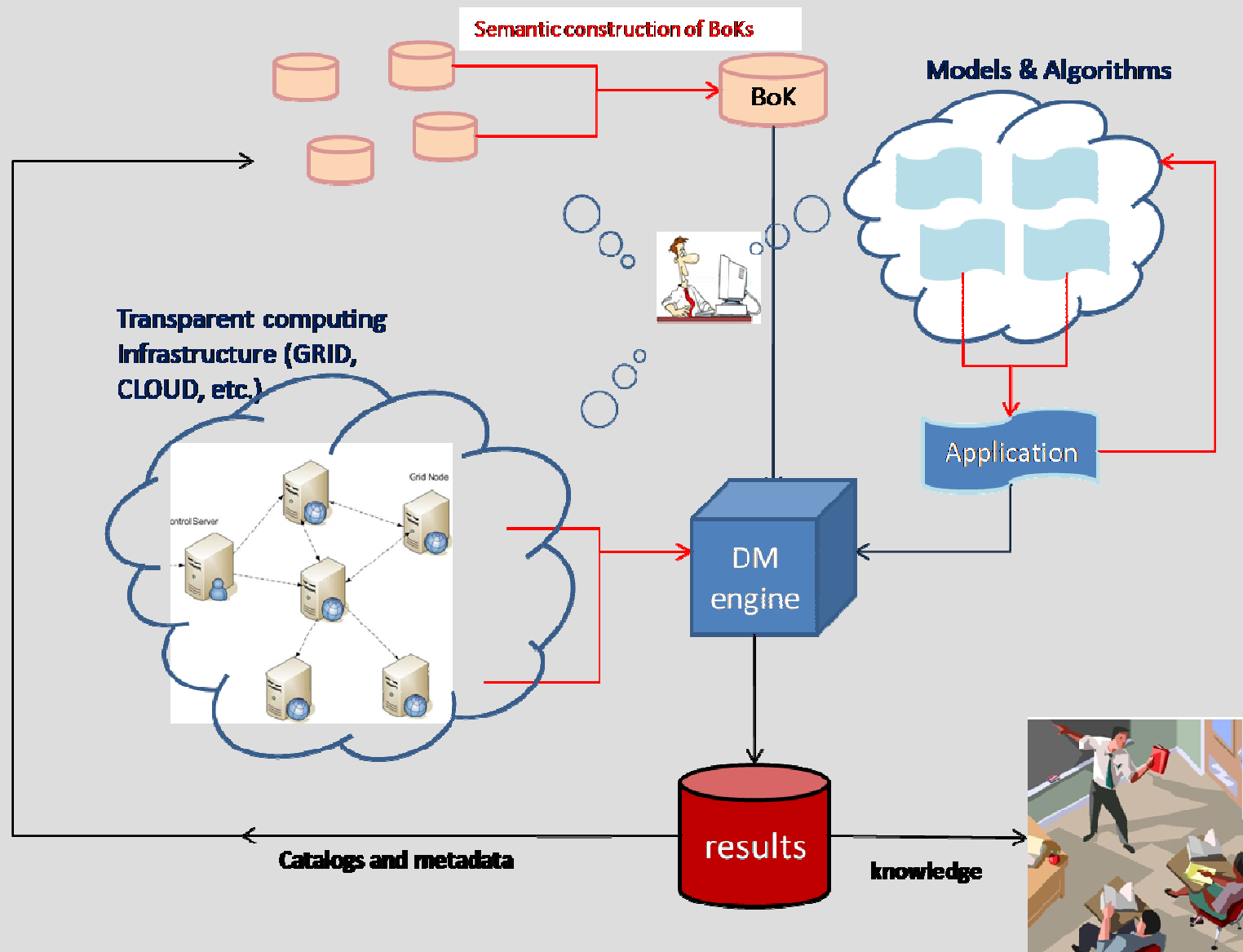
Exposed to a wide choice of algorithms to solve a problem, the r.m.s. astronomer usually panics and is not willing to make an effort to learn them ....

**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

**DAta Mining & Exploration**

Home Project Documents Download Contact Us

**Data Mining & Exploration Project**

News & Events **DAME (Data Mining & Exploration)** is a project aimed at designing and developing instruments and tools for scientific data mining based on information and communication

**DAta Mining & Exploration**

**DAta Mining & Exploration**

**SDSS**

**DAta Mining & Explora**

**Welcome on the WFXT Transient calculator web**

This is an offered service within the DAME Cloud, maintain

[Click here to go directly to the 'Wide Field X-ray Telescope'](#)

**DAME News & Events**

- [DAME Electronic Newsletter](#)
- [DAME @ INAF VO-Days](#)
- [DMPlugin wizard now available](#)
- [New DAME Prototype released](#)
- [theses and apprenticeships offer now available](#)
- [Project theses](#)

**Inside DAME**

- Management
- Technology
- Science

## Welcome on the WFXT Transient calculator web

This is an offered service within the DAME Cloud, maintain

[Click here to go directly to the 'Wide Field X-ray Telescope'](#)

The Wide Field X-Ray Telescope (WFXT) is a medium orders-of-magnitude more sensitive than any previous c area surveys and to match in sensitivity the next generation radio surveys. Using an innovative wide-field X-ray optics view of 1 square degree (10x *Chandra*) with an angular resolution (HEW) nearly constant over the entire field, and a *Chandra* like energy band over the 0.1-7 keV band. WFXT's low-Earth orbit, in five years of operation, WFXT will carry out

1. **WIDE** survey will cover most of the extragalactic : the sensitivity, and twenty times better angular resolution.
2. **MEDIUM** survey will map ~3000 deg<sup>2</sup> to deep sensitivity.
3. **DEEP** survey will probe ~100 deg<sup>2</sup>, or ~1000 sq deg, to the deepest *Chandra* sensitivity.

The feasibility study has been supported through an ASI grant on the Italian side, while the whole project is currently under scrutiny by the 2010 US Decadal Survey, and has received encouraging comments so far.

**My Experiments**

| Experiments List | Name  | Science case | Mode  | Status   | Actions |
|------------------|-------|--------------|-------|----------|---------|
| Home             | ppqso | ppqso        | perun | finished | Remove  |
| Science and Tech | ppqso | ppqso        | perun | finished | Remove  |
| MyFilestore      | ppqso | ppqso        | perun | finished | Remove  |
| MyExperiments    | ppqso | ppqso        | perun | finished | Remove  |
| Logout           | ppqso | ppqso        | perun | finished | Remove  |
| Help & Tutorials | ppqso | ppqso        | perun | finished | Remove  |
| The Team         | ppqso | ppqso        | perun | finished | Remove  |

**My Filestore**

Click here to upload the file

| Dir        | Files               | Actions         |
|------------|---------------------|-----------------|
| /largo     |                     |                 |
| /largo/MAE |                     |                 |
|            | data/               | Delete Download |
|            | data/mae_photos.csv | Delete          |
|            | data/mae_photos.csv | Delete          |

**DAta Mining & Exploration**

**Web Solutions for Data Mining in Astrophysics**

The DAME project involves the implementation of advanced tools for data mining in astrophysics, including the implementation of advanced tools for data mining in astrophysics, including the implementation of advanced tools for data mining in astrophysics...

# 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)

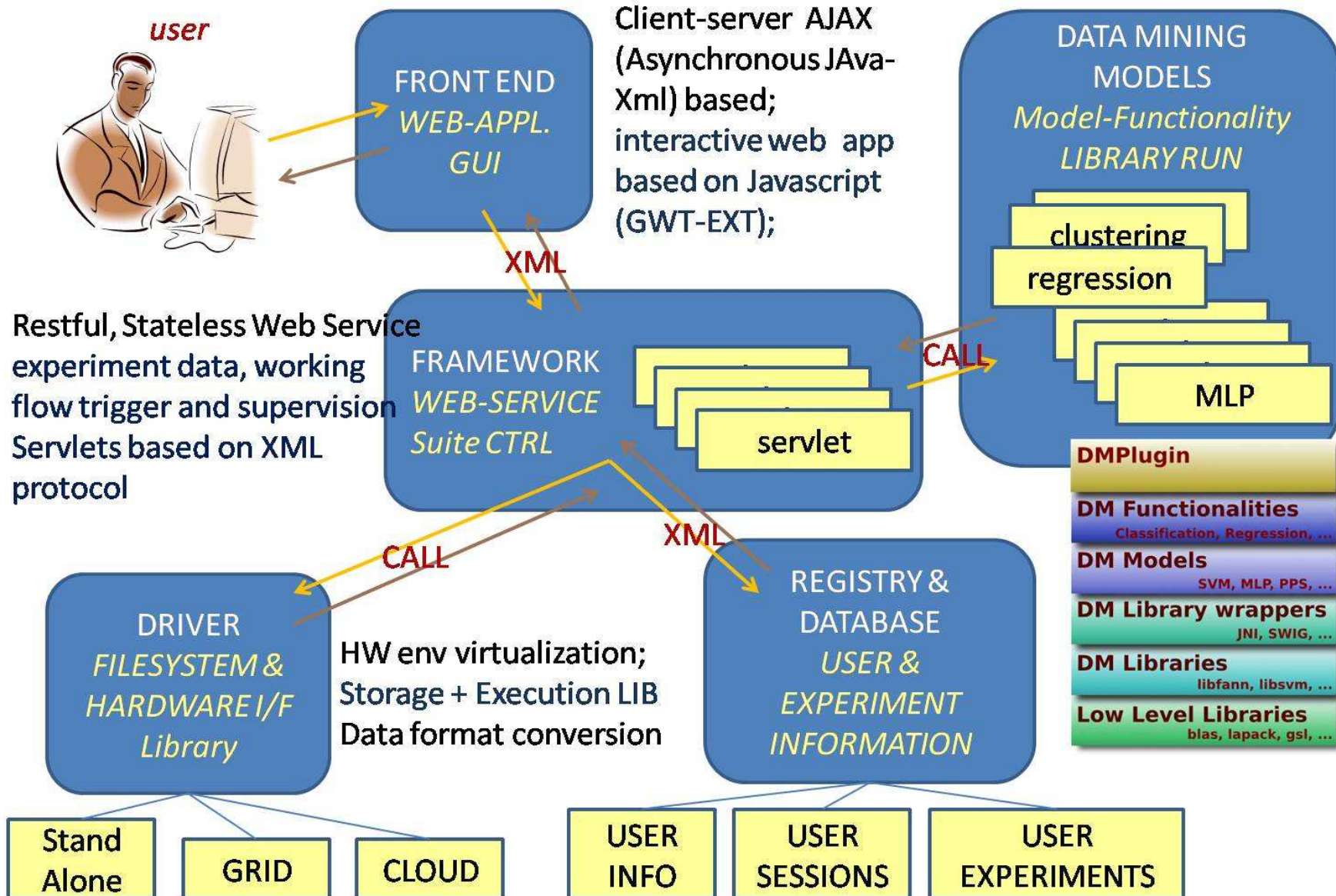
Main service providing via browser a list of algorithms and tools to configure and launch experiments as complete workflows (dataset creation, model setup and run, graphical/text output):

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

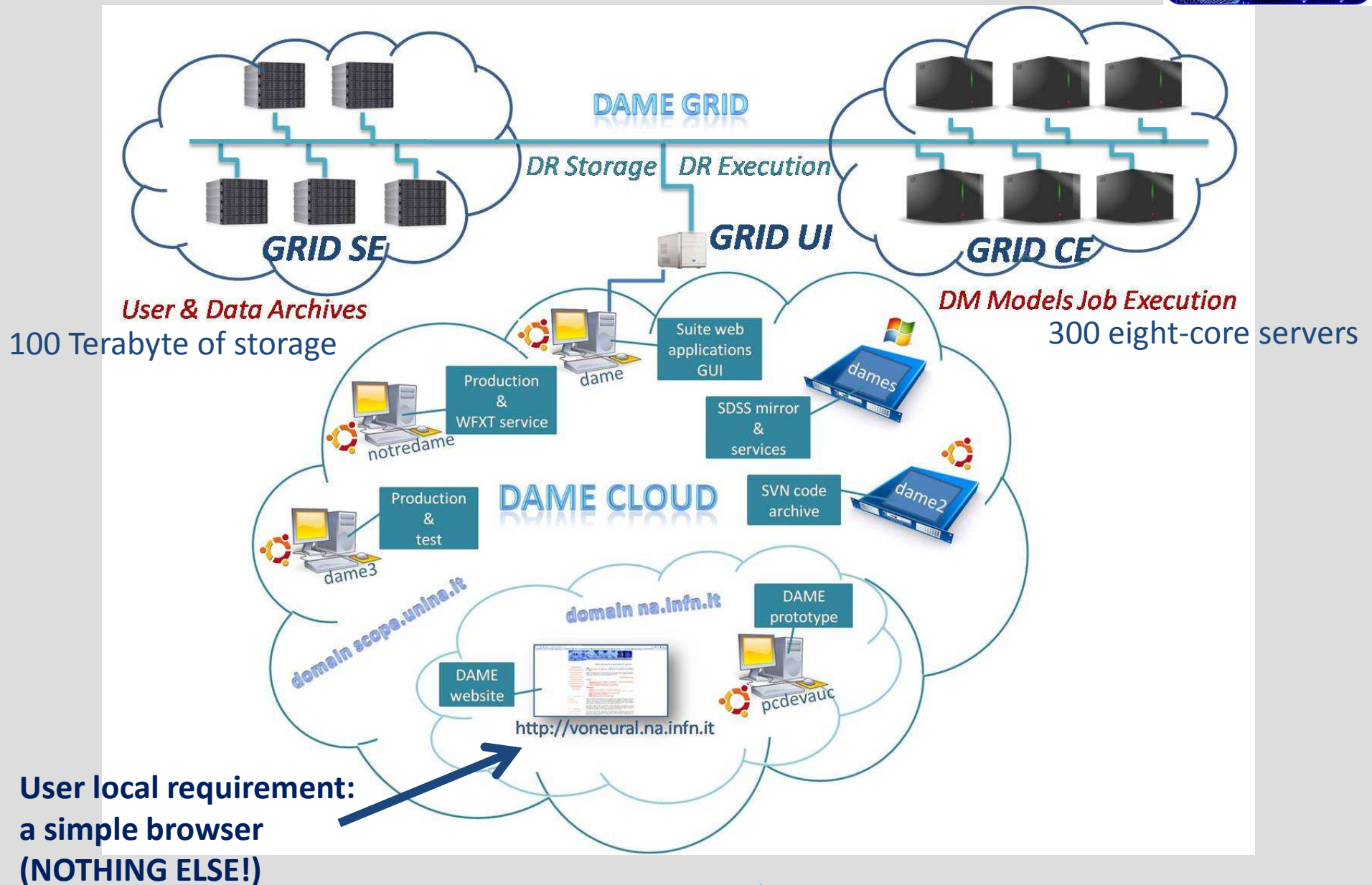
## VOGCLUSTERS

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

# The DAME SW Architecture



# The Available Resources





# 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://www.as.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

Catania



Teramo



Napoli

Padova



Poster @ this Congress available

## VO – day ... in Tour

E. Sciaraglia, U. Becciani, F. Bevrani, M. Brescia, A. Cecc, B. Garilli, C. Gheller, C. Infante, O. Laurina, G. Longo, F. Manzoni, M. Molinari, P. Pedrazzi, E. Parisi, M. Ranzani  
INAF - OAT, INAF - OAC, 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 (VOday.it) levels. The course was held in all INAF sites where there are INAF partners, with the support of INAF-OAT and INAF-SI, and in collaboration with some of 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://www.as.inaf.it/voday>

---

**Vo-Day ... in Tour: Program**

**Methods:**

- A first overview about what is the VO, the IVOA and the tools
- Lecture and tutor the participants on the usage of such tools (with Italian partnership as first)
- give real life examples of scientific applications
- hands-on exercises

**First Day:**

09:30 - 09:55: **Introduction**  
*O. Laurina*

09:55 - 10:15: **The Virtual Observatory: an Overview**  
*A. Cecc, F. Bevrani*

10:15 - 10:45: **A Virtual Tour of the Virtual Observatory**  
*F. Manzoni*

10:45 - 11:15: **Use Case 1: Construction of a Supernova candidate**  
*G. Longo, M. Brescia*

11:15 - 13:00: **Use Case 2: Data available for the nearby galaxy M31**  
*M. Molinari*

13:00 - 14:00: **Lunch**

14:00 - 15:00: **Use Case 2 - (continues)**

15:00 - 17:00: **Use Case 3: Photometric redshifts with PAIR**  
*M. Brescia, O. Laurina, F. Bevrani*

17:00 - 19:00: **Presentazione attività di supporto alla didattica.**  
*G. Longo, M. Brescia*

**Second Day:**

09:30 - 11:30: **Use Case 4: Data Extraction from Multidimensional Databases**  
*U. Becciani, A. Cecc*

11:30 - 12:00: **Feedback**

**Results:**

page)

but without use VO

how publish his/her data

ult.htm

Virtual Observatory; should involved in the project to learn more, enter data and

## VO – day ... in Tour

E. Sciaraglia, U. Becciani, F. Bevrani, M. Brescia, A. Cecc, B. Garilli, C. Gheller, C. Infante, O. Laurina, G. Longo, F. Manzoni, M. Molinari, P. Pedrazzi, E. Parisi, M. Ranzani  
INAF - OAT, INAF - OAC, 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 (VOday.it) levels. The course was held in all INAF sites where there are INAF partners, with the support of INAF-OAT and INAF-SI, and in collaboration with some of 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://www.as.inaf.it/voday>





# @ VO-Day ... in Tour

<http://www.as.oats.inaf.it/voday>

Registered: 272

➔ more than 1/4 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 4<sup>th</sup> 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!

