

# Data Mining & Exploration



*we make science discovery happen*

## DAME: A WEB 2.0 TECHNOLOGY BASED INFRASTRUCTURE FOR DATA EXPLORATION



Dipartimento di Scienze Fisiche  
Università di Napoli "Federico II"

◆ **INAF**

ISTITUTO NAZIONALE  
DI ASTROFISICA

NATIONAL INSTITUTE  
FOR ASTROPHYSICS

George S. Djorgovski, Ciro Donalek, Ashish Mahabal

Giuseppe Longo, Marianna Annunziatella, Stefano Cavuoti,  
Mauro Garofalo, Marisa Guglielmo, Ettore Mancini,  
Francesco Manna, Alfonso Nocella, Luca Pellicchia,  
Sandro Riccardi, Civita Vellucci

Massimo Brescia

# The DAME vision



DAME Program is a joint effort between University Federico II, Caltech and INAF-OACN, aimed at implementing (as web 2.0 applications and services) a scientific gateway for data analysis, exploration and mining, on top of a virtualized distributed computing environment.

**DA**ta Mining & Exploration **ME**

we make science discovery happen

Home Cloud Services Science Technology Machine Learning Publications Education & Lectures The Team Cont...

What is DAME

Nowadays, many scientific areas share the same need of being able to deal with massive and distributed datasets and to perform on them complex knowledge extraction tasks. This simple consideration is behind the international efforts to build virtual organizations such as, for instance, the Virtual Observatory (VOs). DAME (Data Mining & Exploration) is an innovative, general purpose, Web-based, distributed data mining infrastructure specialized in Massive Data Sets exploration with machine learning methods.

Initially fine tuned to deal with astronomical data only, DAME has evolved in a cloud of applications and services useful also in other domains of human endeavor. DAME is an evolving platform and new services as well as additional features are con...

<http://dame.dsf.unina.it/>  
Science and management info  
Documents  
Science cases  
Newsletters

**DA**ta Mining & Exploration **ME**

01

**A New trend in Astrophysics**

**Web Solutions for Data Mining in Astrophysics**

Web Solutions for Data Mining in Astrophysics

The 2011 annual survey has been published by the Virtual Observatory (VO) community. It is a comprehensive report on the state of the art in the field of data mining in astronomy. The report is divided into several sections, including: Introduction, Data Mining in Astronomy, Data Mining in Astrophysics, Data Mining in Cosmology, Data Mining in Planetary Science, Data Mining in Solar System Science, Data Mining in Space Science, Data Mining in Earth and Planetary Science, Data Mining in Environmental Science, Data Mining in Life Sciences, Data Mining in Social Sciences, Data Mining in Humanities, Data Mining in Law, Data Mining in Business, Data Mining in Finance, Data Mining in Marketing, Data Mining in Healthcare, Data Mining in Education, Data Mining in Government, Data Mining in Industry, Data Mining in Energy, Data Mining in Transportation, Data Mining in Defense, Data Mining in Intelligence, Data Mining in Security, Data Mining in Law Enforcement, Data Mining in Public Safety, Data Mining in Homeland Security, Data Mining in Counterterrorism, Data Mining in Counterintelligence, Data Mining in Counterproliferation, Data Mining in Counterterrorism, Data Mining in Counterintelligence, Data Mining in Counterproliferation.

www.youtube.com/user/DAMEmedia

ASTRO MYWORK PROGRAMMING DATA MINING AZIENDE UTILITY PUBBLICAZIONI MEETINGS

**YouTube**

**Data Mining & Exploration**  
Il canale di DAMEmedia

RoundCube Webmail - Posta in arrivo

**WARE**

Data Mining Web

DAME (Data Mining & Exploration) is an innovative, general purpose, Web-based, distributed data mining infrastructure specialized in Massive Data Sets

DAME Application - User: stefano.cavuoti@gmail.com

App Manuals Model Manuals Cloud Services Science Cases

RESOURCE MANAGER Upload in erice Workspace

Workspace

New Workspace

Rename Workspace Upload Experiment Delete

experiment

File Manager

Workspace: experiment

Down Edit File

prova.dat

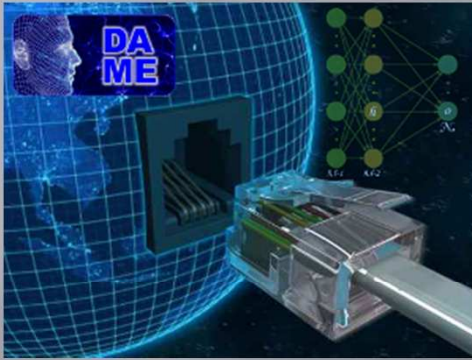
My Experiments

Workspace: experiment

Experiment	Status	Last Access	Delete
dasdsadas	failed	2011-03-07	✖
dsadsa	failed	2011-03-07	✖
asdsadsa	ended	2011-03-07	✖
asdsasd	ended	2011-03-07	✖
frfrf	ended	2011-03-07	✖

<http://www.youtube.com/user/DAMEmedia>  
DAMEWARE Web Application media channel

# DAME projects



Multi-purpose data mining  
with machine learning  
Web App REsource



Extensions

- DAME-KNIME
- ML Model plugin



Specialized web apps for:

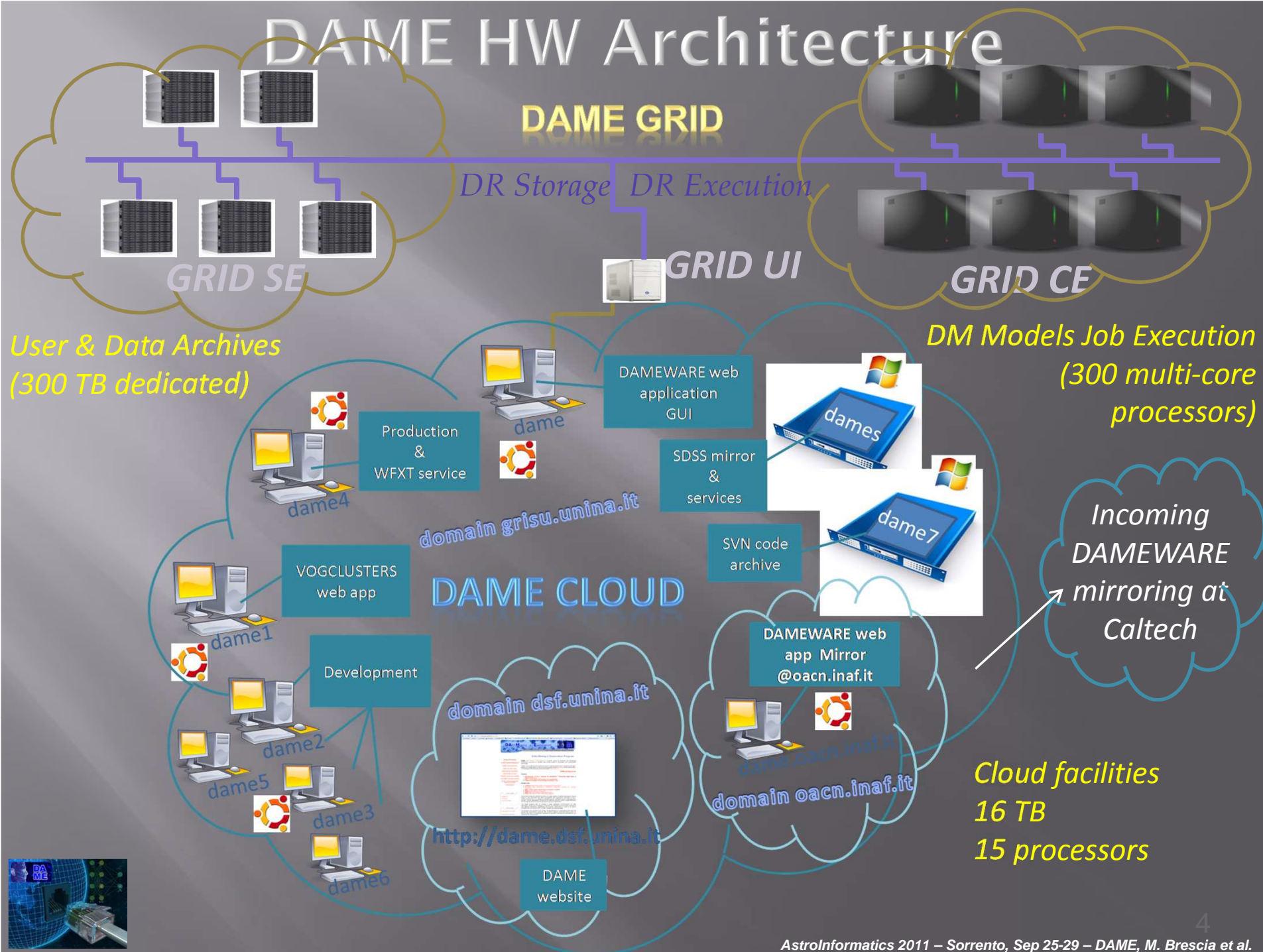
- text mining (VOGCLUSTERS)
- Transient classification (STraDiWA)
- EUCLID Mission Data Quality



Web Services:

- SDSS mirror
- WFXT Time Calculator
- GAME (GPU+CUDA ML model)

# DAME HW Architecture



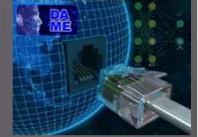
User & Data Archives  
(300 TB dedicated)

DM Models Job Execution  
(300 multi-core processors)

Incoming DAMEWARE mirroring at Caltech

Cloud facilities  
16 TB  
15 processors

# Web 2.0 Features in DAME



*Web 2.0? It is a system that breaks with the old model of centralized Web sites and moves the power of the Web/Internet to the desktop. [J. Robb]*

*the Web becomes a universal, standards-based integration platform. [S. Dietzen]*

software and storage facilities, all through a simple browser

client-side browser with Javascript/Ajax, JDOM and XML standard technologies

Web as a participating and sharing information platform

Machine Learning scalable tools on Massive Datasets

Rich Internet App (RIA)  
Network as a process platform  
Desktop app → Web app

Service Oriented App (SOA)  
Growing functionalities integration via app service interoperability

unification in a single framework of:

- ❖ Client-server structure
- ❖ WYSWYG Dynamical content
- ❖ Network protocols
- ❖ Cloud/Grid virtualized platforms

Machine-based interactions (REST, SOAP) based on standards (PMML, WSDL, XML)



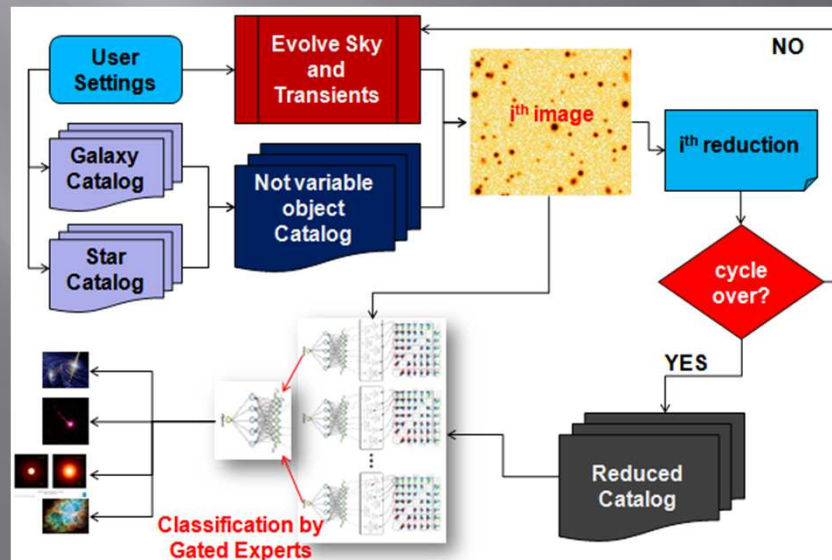
# DAME Projects: STraDiWA



**Sky Transient Discovery Web App** [http://dame.dsf.unina.it/dame\\_td.html](http://dame.dsf.unina.it/dame_td.html)

customizable workflow for real time classification of variable sky objects

- Configurable by user through web I/F;
- Customized mixing of third-part SW (Stuff, Skymaker, SExtractor, PSFEx, Daophot);
- Telescope + instrument signature setup (FOV, pixel scale, gain, readout noise...);
- Setup of Exp. Time, PSF model, filters, magnitude range...;
- Stars+galaxies+background modeling for controlled image simulation;
- transient models to populate simulated images;
- Now available Cepheids (*Sandage et Tammann 2004*), SN Ia (*Contardo et al. 2000*);
- Catalogue extraction to test classifiers based on machine learning models;
- Real images can be used for transient classification with validated models;

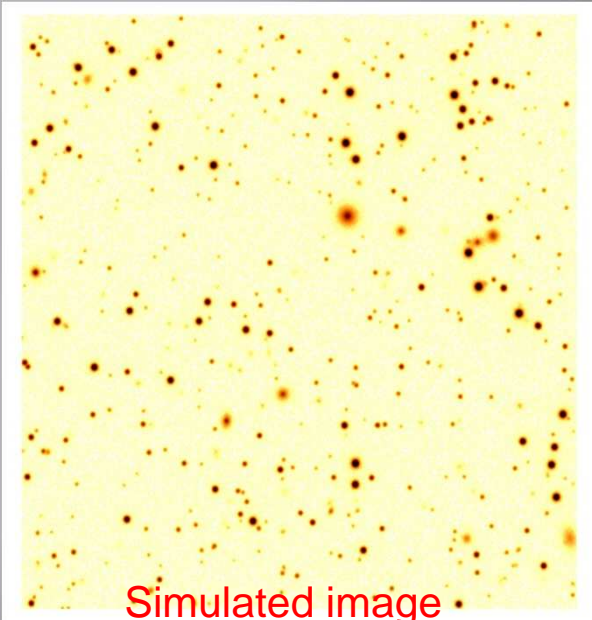


Next steps:

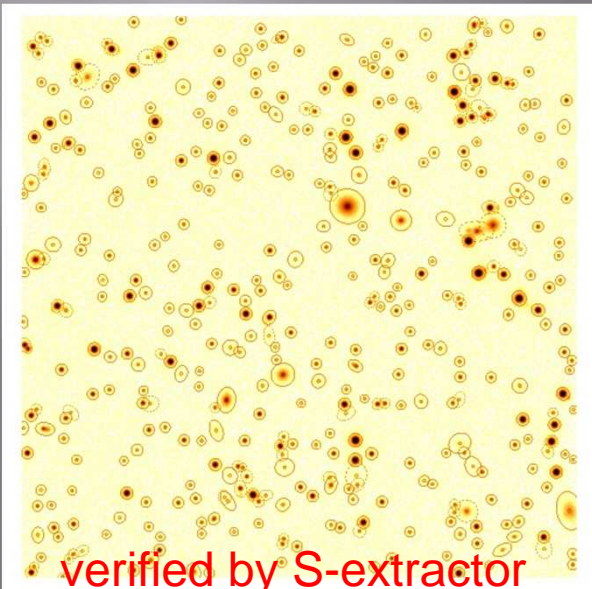
- New transient models;
- Other telescope models;
- ML algorithms test and validation;
- Real data for workflow tuning;

**We are open to collaborations...**

# DAME Projects: STraDiWA



Simulated image



verified by S-extractor

Parameter Source	Image Parameters	Value
SkyMaker	MAG ZEROPOINT	26.0 ADU per sec
	AUREOLE RADIUS	188
Common	GAIN	1.0 e-/ADU
	SEEING	0.7
S-Extractor	THRESHOLD	0.8
	FILTER	gauss_4.0_7x7.conv

VST  
case



```
#STraDiWA v1.0
1# Default configuration file Version 1.0
```

```
#-----Setup Files-----
SETUP_FILES ./default.stuff,./defaultVST.sky,./default.sex
#name and path of configuration files of Stuff, SkyMaker and S-Extractor
```

```
#-----Stuff Parameters-----
STUFF_CATALOG_NAME B.list,V.list,I.list # different for each band
PASSBAND_OBS sandage/B,sandage/V,johnson/I # Observed passbands
in Stuff (sandage, johnson etc are the folders contains filters)
```

```
#-----Variable Objects-----
TRANSIENT 1,20.3,2.5,RANDOM,RANDOM # OBJECT TYPE, INITIAL
MAGNITUDE, PERIOD (days), AMPLITUDE( mag), PHASE(rad) (to have
random values, RANDOM)
```

```
TRANSIENT 1,21.3,1.2,1.8,2Pi # OBJECT TYPE, INITIAL
MAGNITUDE, PERIOD (days), AMPLITUDE( mag), PHASE(rad) (to have
random values, RANDOM)
```

# DAME Projects: STraDiWA



PSFEx does not work directly on images. Instead, it operates on SExtractor catalogues.

PSFEx\_MODEL doesn't work well with the default psf provided by SExtractor. It needs a psf build on the image that we are considering.

We can identify as stars the objects with PSFEx\_MODEL less than a certain threshold and as galaxies the objects with PSFEx\_MODEL greater or equal to that. The best is obtained with value 0.01, although with a little contamination for galaxies.

Bin	$S_{\text{CLASS\_STAR} \geq 0.98} / S_{\text{extracted}}$	$G_{\text{CLASS\_STAR} < 0.98} / G_{\text{extracted}}$
18-19 mag	100,00%	100,00%
19-20 mag	96,43%	100,00%
20-21 mag	100,00%	100,00%
21-22 mag	98,11%	100,00%
22-23 mag	85,39%	100,00%
23-23.5 mag	52,17%	100,00%
23.5-24 mag	18,42%	100,00%
24-24.5 mag	3,41%	100,00%
24.5-25 mag	0,00%	100,00%
25-25.5 mag	0,00%	100,00%
25.5-26 mag	0,00%	100,00%

**SExtractor**

Bin	$S_{\text{PSFEx\_MODEL} < 0.01} / S_{\text{extracted}}$	$G_{\text{PSFEx\_MODEL} \geq 0.01} / G_{\text{extracted}}$
18-19 mag	94,12%	100,00%
19-20 mag	100,00%	100,00%
20-21 mag	100,00%	100,00%
21-22 mag	100,00%	100,00%
22-23 mag	100,00%	100,00%
23-23.5 mag	100,00%	86,96%
23.5-24 mag	100,00%	91,67%
24-24.5 mag	100,00%	74,24%
24.5-25 mag	100,00%	56,10%
25-25.5 mag	98,17%	24,69%
25.5-26 mag	95,04%	14,81%

**SExtractor -> PSFEx -> SExtractor**



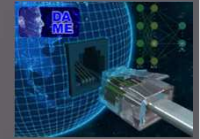
# DAME Projects: VOGCLUSTERS

## Globular Clusters Mining Web App

<http://dame.dsf.unina.it/vogclusters.html>

data and text mining activities for astronomical archives related to globular clusters

- VO and local archives browsing and selection;
- Setup and storage of complex views of data;
- on the fly plots by customized data correlation;
- Export data in multiple formats (EPS, FITS, JPG, PNG...);
- Browse related publications;
- Add new data, imgs, notes and biblio references;
- Update existing data and single parameters;
- Leave comments and research notes;
- Interaction with other users (messages, Facebook, etc.);



**VOGCLUSTERS: a DAME web app on Globular Clusters**  
Marco Castellani<sup>1</sup>, Massimo Brescia<sup>2</sup>, Ettore Mancini<sup>1</sup>,  
Luca Pellecchia<sup>1</sup>, Giuseppe Longo<sup>3</sup>  
<sup>1</sup>INAF - Astronomical Observatory of Roma, <sup>2</sup>INAF - Astronomical Observatory of Capodimonte (Naples),  
<sup>3</sup>Dept. of Physics, University of Salerno, <sup>4</sup>INAF - Astronomical Observatory of Capodimonte II (Naples)

**Summary**  
We present the alpha release of the VOGCLUSTERS web application, a tool for data and text mining on globular clusters. It is one of the web 2.0 technology based projects of the Data Mining & Exploration (DAME) Program, devoted to mine and explore heterogeneous information related to globular clusters data.

Using web 2.0 progress of the Information for Astronomy (Info4A) program, by exploring the web 2.0 technologies and exposing the application and services for a science communities, with a particular emphasis on astrophysical services (Brescia et al., 2010)

**Problem** In data are presently scattered among the various papers and are reported in different (and not homogeneous) web pages. One needs a simple way to have the relevant informations on a given cluster, or a range of clusters in a single source and under a well defined standard. This is a mandatory disclosure a wide range of new information: the more we have detailed information about parameters, the more we are able to investigate upon the possible correlations of those parameters.

**The goal of the project** VOGCLUSTERS is the development of a web application specialized in data and text mining of astronomical archives related to galactic and extragalactic GCs. Main goal is to employ for the simple and quick navigation in the archives and for the manipulation to correlate and integrate internal scientific information. The archives are uniformed under Virtual Observatory standard and constraints, in order to provide an homogeneous and flexible environment, virtually capable of interactions with an ever growing amount of external resources. At variance with its prototype galsters (Galactic Globular Clusters Database, see Castellani 2007), the project has not to be intended as a straightforward website, but as a true web application.

**Among the features of VOGCLUSTERS (not all yet implemented):**

- Browse and make selections in the archives (local and remote ones)
- Save complex views of data
- Make plots on the fly correlating existing and remote data
- Export your work in a variety of formats (EPS, FITS, JPG, PNG, etc...)
- Browse related publications
- Add new data (supervised)
- Update existing data (supervised)
- Leave comments and research note related to a certain cluster
- Interact with other users of the application (messages etc...)
- NEXT GREAT FEATURE! (thanks to the chosen standards & protocols, we can easily implement more...)

**The recently updated home screen of the VOGCLUSTERS web app**

A web application is an application that is accessible via web. They are now widely diffused (well known examples are Gmail and Wikipedia) due to the fact that they can be accessed from everywhere via a standard browser (IE, Firefox, Opera, Chrome, Safari, ...). An evolution of the web app concept is the Rich Internet Application (RIA) which presents the appearance and interactivity typical of a classical desktop application, without the need to install anything on your computer. In fact, in many cases you do not need even a computer – all you need is a device that can connect you to internet: as cheap and small as it can be (a tablet may suffice), it might enable you to do real research, since data and application reside on remote servers. VOGCLUSTERS belongs in this category.

**Further reading & links:**  
VOGCLUSTERS HomePage: <http://dame.dsf.unina.it/vogclusters.html>  
VOGCLUSTERS: <http://galsters.abneta.org>  
Facebook page: <http://www.facebook.com/galsters>  
Harris Catalogue: <http://www.abneta.monasteri.eu/Globular.html>  
ResearchGate Topic: <http://www.researchgate.net/topic/Astronformatics/>  
GWT: <http://code.google.com/webtoolkit/>

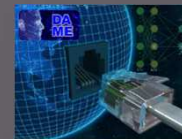
**The globular cluster M15 (ESA, Hubble, NASA)**

**Bibliographic Items:**

- Brescia, M. et al. 2010 in press (arXiv: 1010.4843v2)
- Castellani, M. 2007, MemSAI 79, 2
- Harris, W.E. 1996, AJ 112, 1487

Download this poster at <http://snipurl.com/vogposter>

# DAME Projects: Web Services



Specific services (in collaboration)

- **Sloan Digital Sky Survey Mirror Site**

<http://dames.scope.unina.it/>

complete Sky Survey Database;  
subset of the image archive related to the sky coverage overlapping VST-KIDS survey project area;

- **WFXT Transient Calculator (M. Paolillo)**

[http://dame.dsf.unina.it/dame\\_wfxt.html](http://dame.dsf.unina.it/dame_wfxt.html)

estimation of the number of variable sources that can be detected by WFXT within the 3 main planned extragalactic surveys, with a given significant threshold;

- **GAME (work in progress, M. Sc. Thesis (M. Garofalo), in coll. with Informatics Engineering Faculty of University Federico II, Naples)**

**Genetic Algorithm Modeling Experiment:** a general-purpose GA for supervised classification, implemented on GPU+CUDA parallel computing platform;

- **EUCLID Mission (work in progress, co-head in Data Quality, coordinated by SGS, F. Pasian)**

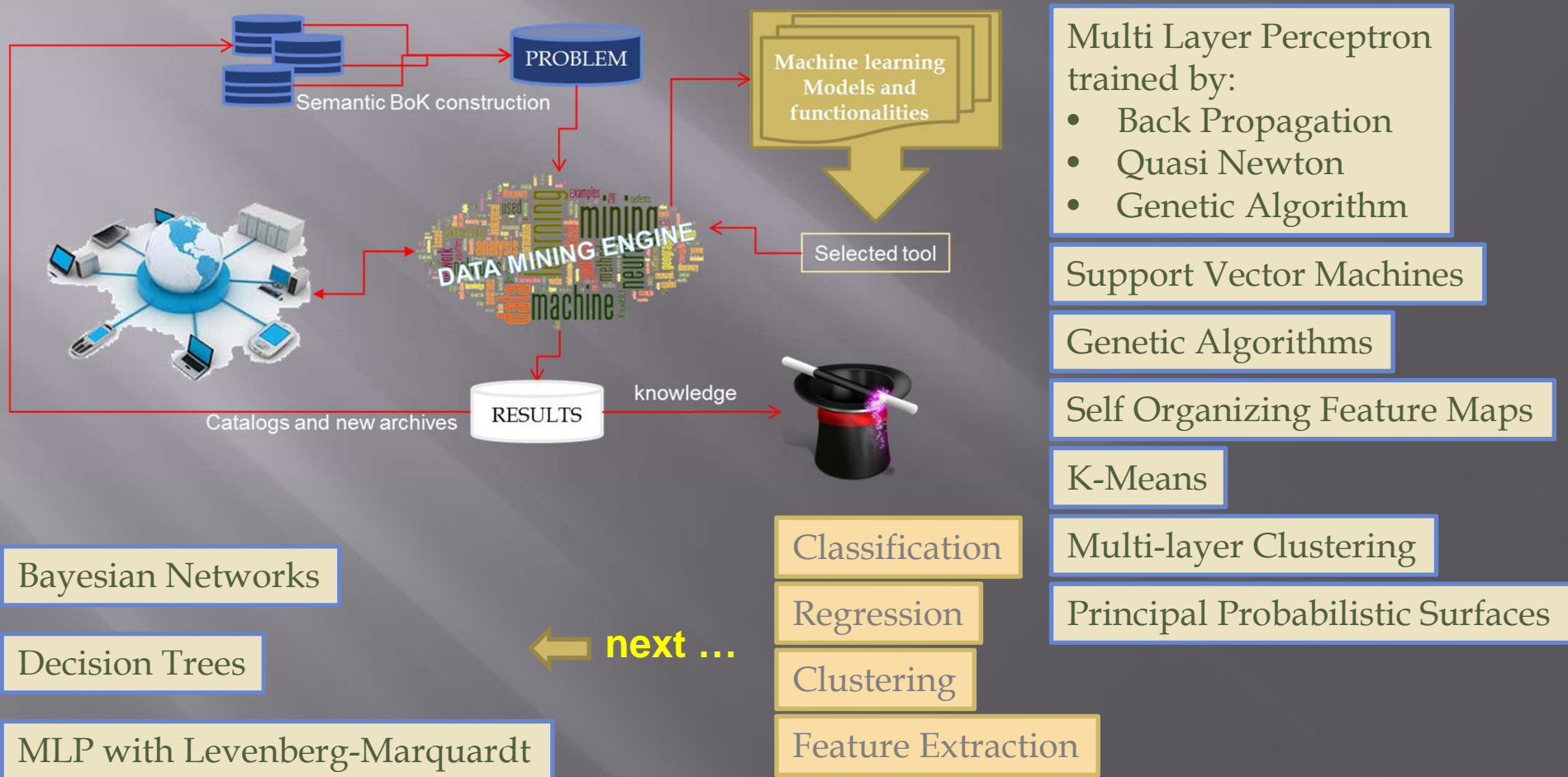
Mission Science Ground Segment (SGS) Data Quality Mining, Science Teams for Photometric Redshifts and Transients;

# DAME Projects: DAMEWARE

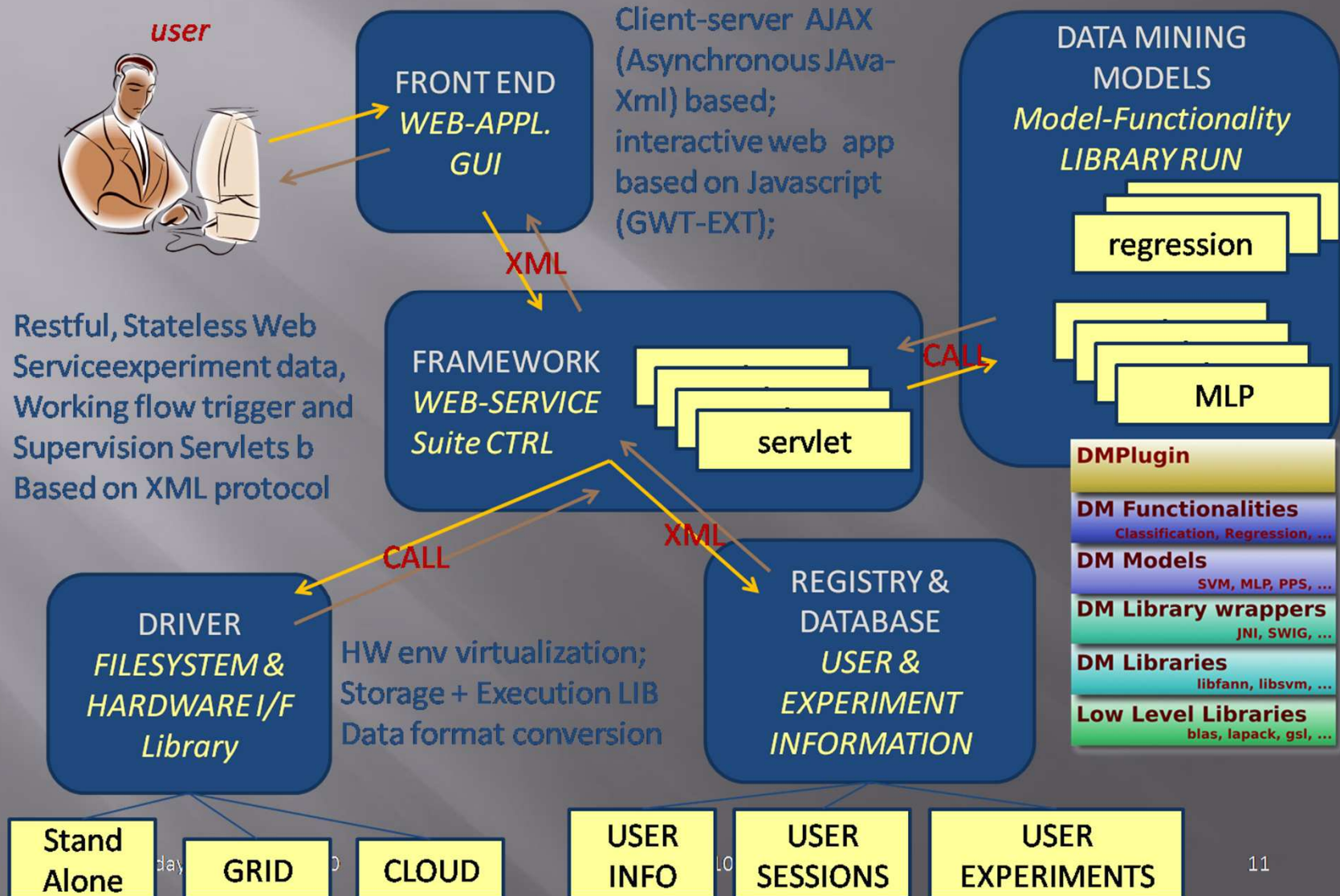


**Data Mining Web Application Resource** [http://dame.dsf.unina.it/beta\\_info.html](http://dame.dsf.unina.it/beta_info.html)

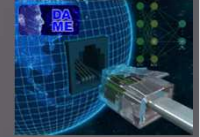
web-based app for massive data mining based on a suite of machine learning methods on top of a virtualized hybrid computing infrastructure



# DAMEWARE SW Architecture



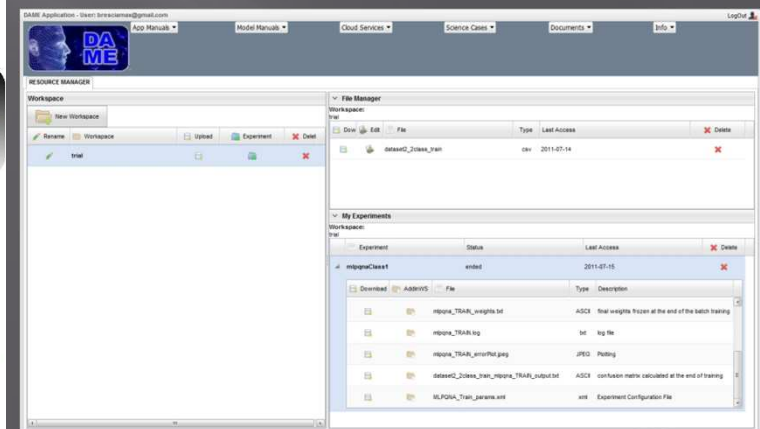
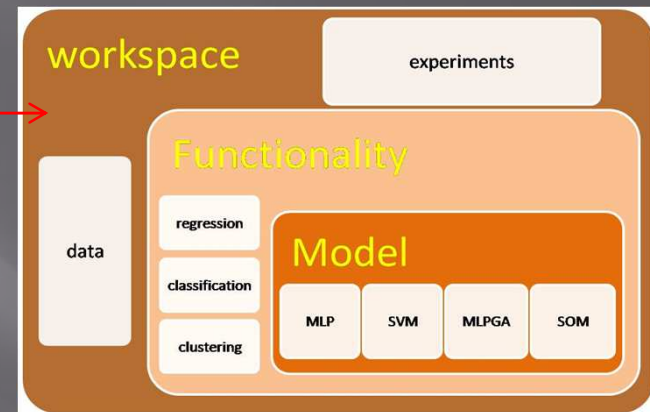
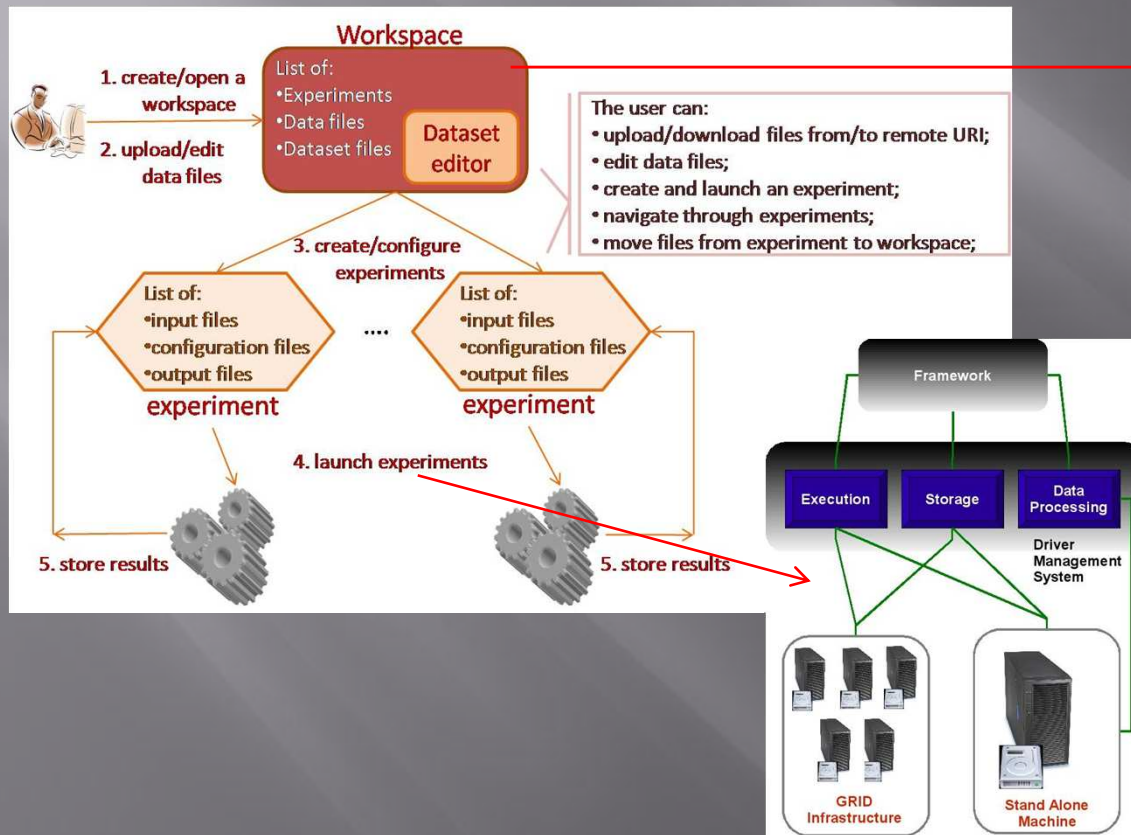
# DAMEWARE fundamentals



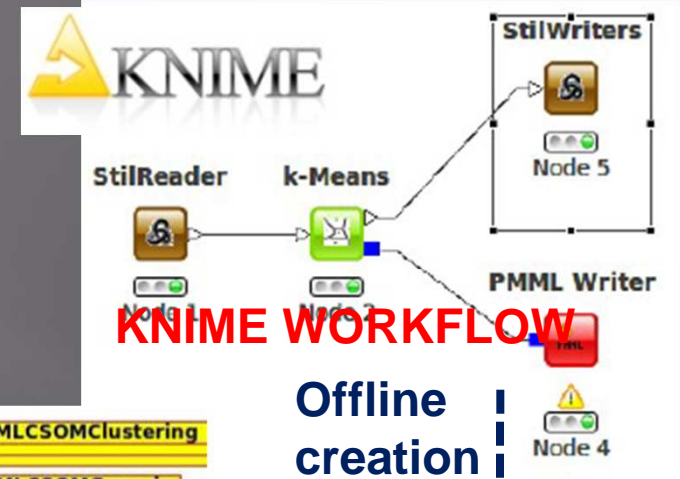
Based on the X-Informatics paradigm, it is multi-disciplinary platform (until now X = Astro)

End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone...)

User can automatically plug-in his own algorithm and launch experiments through the Suite via a simple web browser

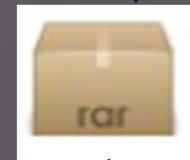


# K-Means (through KNIME)

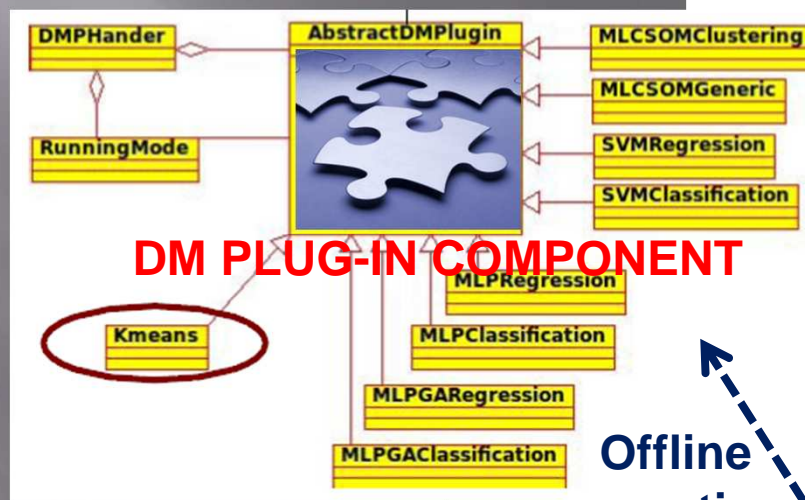


**KNIME WORKFLOW**

Offline creation ↓

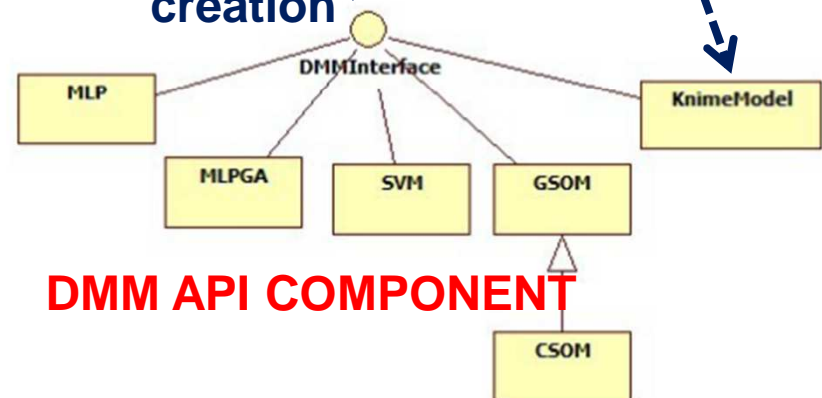


Offline creation ↓



**DM PLUG-IN COMPONENT**

Offline creation ↗

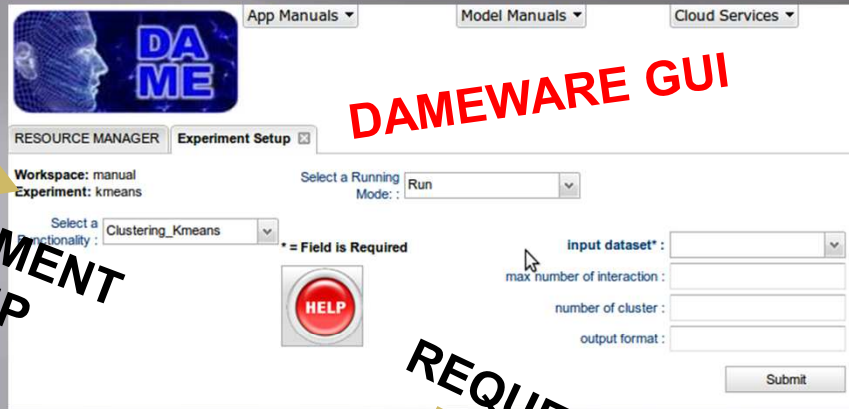


**DMM API COMPONENT**

# K-Means (through KNIME)

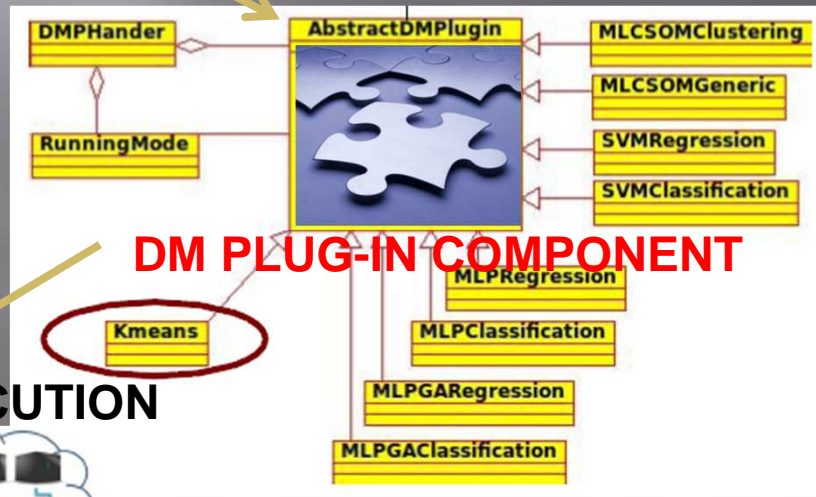


EXPERIMENT  
SETUP



DAMEWARE GUI

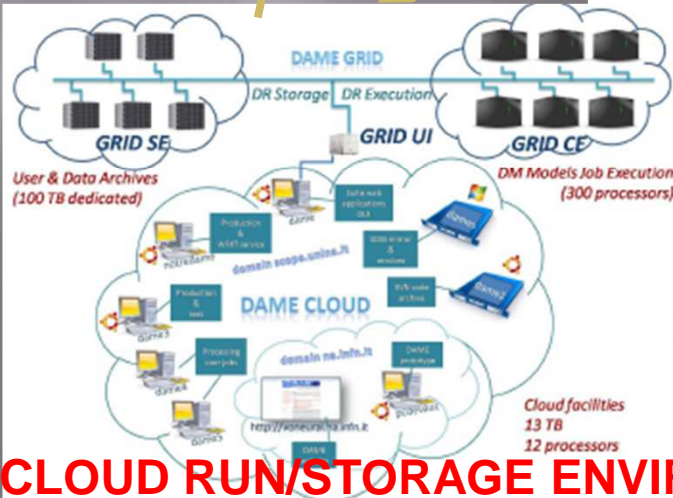
REQUEST



DM PLUG-IN COMPONENT

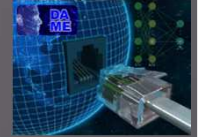


EXECUTION



CLOUD RUN/STORAGE ENVIRONMENT

# DAME Recent Science Cases



## **Global Cluster Search** (classification-MLPQNA + comparison with MLPBP, MLPGA, GAME, SVM)

The use of single band photometry can yield very complete datasets with low contamination, through ANN (MLP trained by Quasi Newton). It will minimize the observing time requirements;

- *Paper under submission to MNRAS;*

## **AGN Classification in the SDSS** (classification-SVM)

Using the GRID to execute 110 jobs on 110 WN, the SVM model produces a classification of different types of AGN using SDSS photometric data and spectroscopic subsamples.

- *Paper in preparation;*

## **Search for Candidate Quasars in the SDSS** (dimension reduction-PPS)

Using PPS applied to SDSS and SDSS+UKIDS data, searching for candidate quasars in absence of a priori constrains and in a high dimensionality photometric parameter space;

- *D'Abrusco et al., 2009. MNRAS;*

## **Photometric Redshifts Evaluation** (regression-MLPBP)

to exploit spectroscopic data wealth of the SDSS to train neural networks to recognize photometric redshifts.

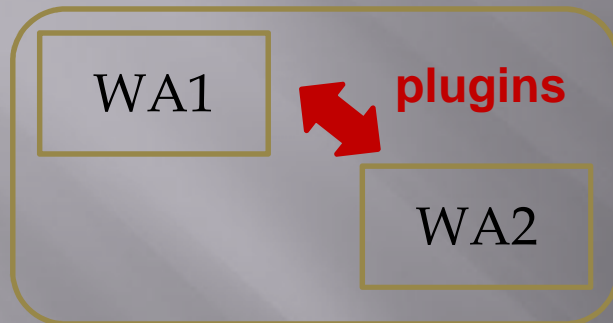
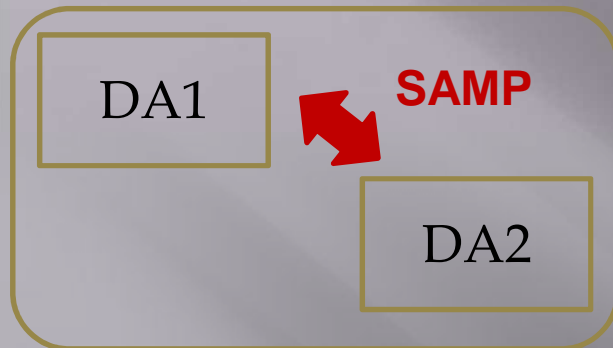
- *D'Abrusco et al., 2007, Ap.J;*



# DAMEWARE improving aspects



During last IVOA Interop we proposed a standardization perspective for KDD apps.



Desktop Apps (DA) has to become Web Apps (WA)

Unique accounting policy (google/Microsoft like)

To overcome MDS flow apps must be plug&play (e.g. any WA1 feature should be pluggable in WA2 on demand)

No local computing power required. Also smartphones can run VO apps

## New Requirements

- **Standard accounting** system and **interoperable** with other data-oriented apps;
- No more MDS moving on the web, but **just moving Apps**, structured **as plugin repositories** and execution environments;
- **standard modeling** of WA and components to obtain the maximum level of granularity;
- Evolution of SAMP architecture to extend web **interoperability** (in particular for the migration of the plugins);
- DAMEWARE also **scriptable** (configurable KDD workflows) for skilled community.

# Conclusions

DAME has been originally conceived as a practical solution to realistic problems that astronomers, like most of us, encountered on exploring massive data sets, coming from new generation of instruments. Nowadays it seems perfectly matching the AstroInformatics perspectives and goals.

Our purpose was:

To propose a sample of what new ICT (Web 2.0) can do for A&A KDD problems.

To propose a new vision of the KDD App approach, that could be extended and adapted, in order to obtain a new generation of instruments, based on the minimization of data transfer and maximization of interoperability (also in the VO community).

If exploited, the new scheme can enlarge the science community, giving the opportunity to share data and apps worldwide, without any particular infrastructure requirements.

Raffaello 1510,  
Athens School,  
Vatican Museums