DAta Mining & Exploration

we make science discovery happen

DAME: A WEB 2.0 TECHNOLOGY BASED INFRASTRUCTURE FOR DATA EXPLORATION

George S. Djorgovski, Ciro Donalek, Ashish Mahabal

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

Massimo Brescia

Dipartimento di Scienze Fisiche
Università di Napoli "Federico II"

INAF
ISTITUTO NAZIONALE DI ASTROFISICA
NATIONAL INSTITUTE FOR ASTROPHYSICS

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.

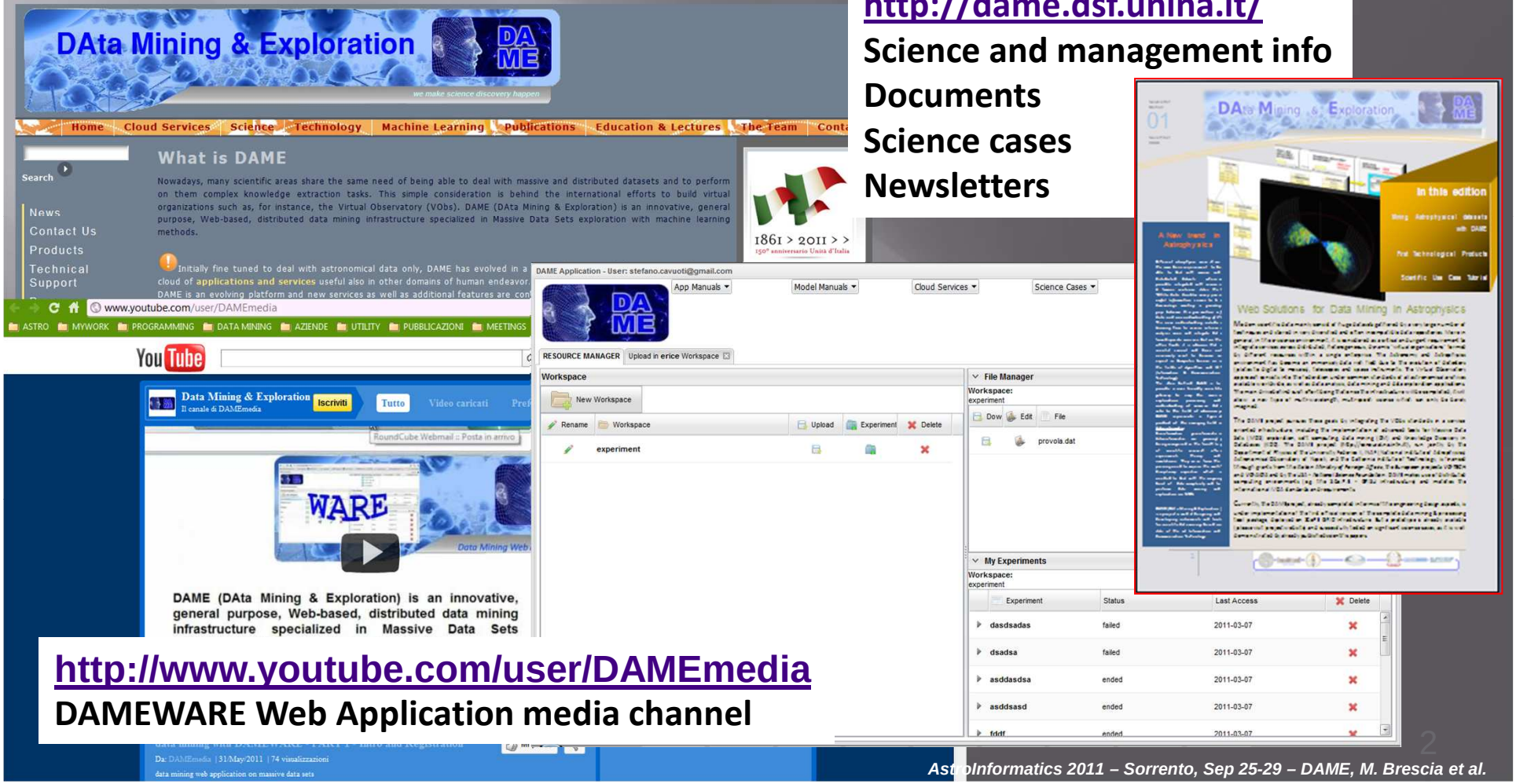
http://dame.dsf.unina.it/

Science and management info

Documents

Science cases

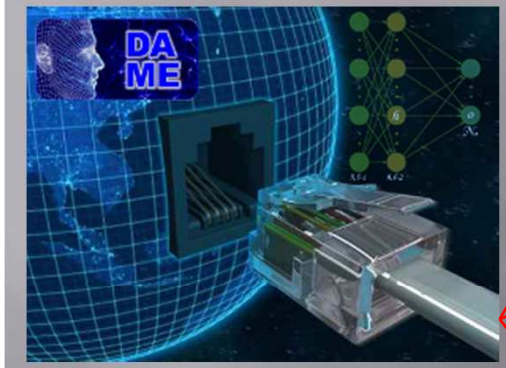
Newsletters



http://www.youtube.com/user/DAMEmedia

DAMEWARE Web Application media channel

DAME projects



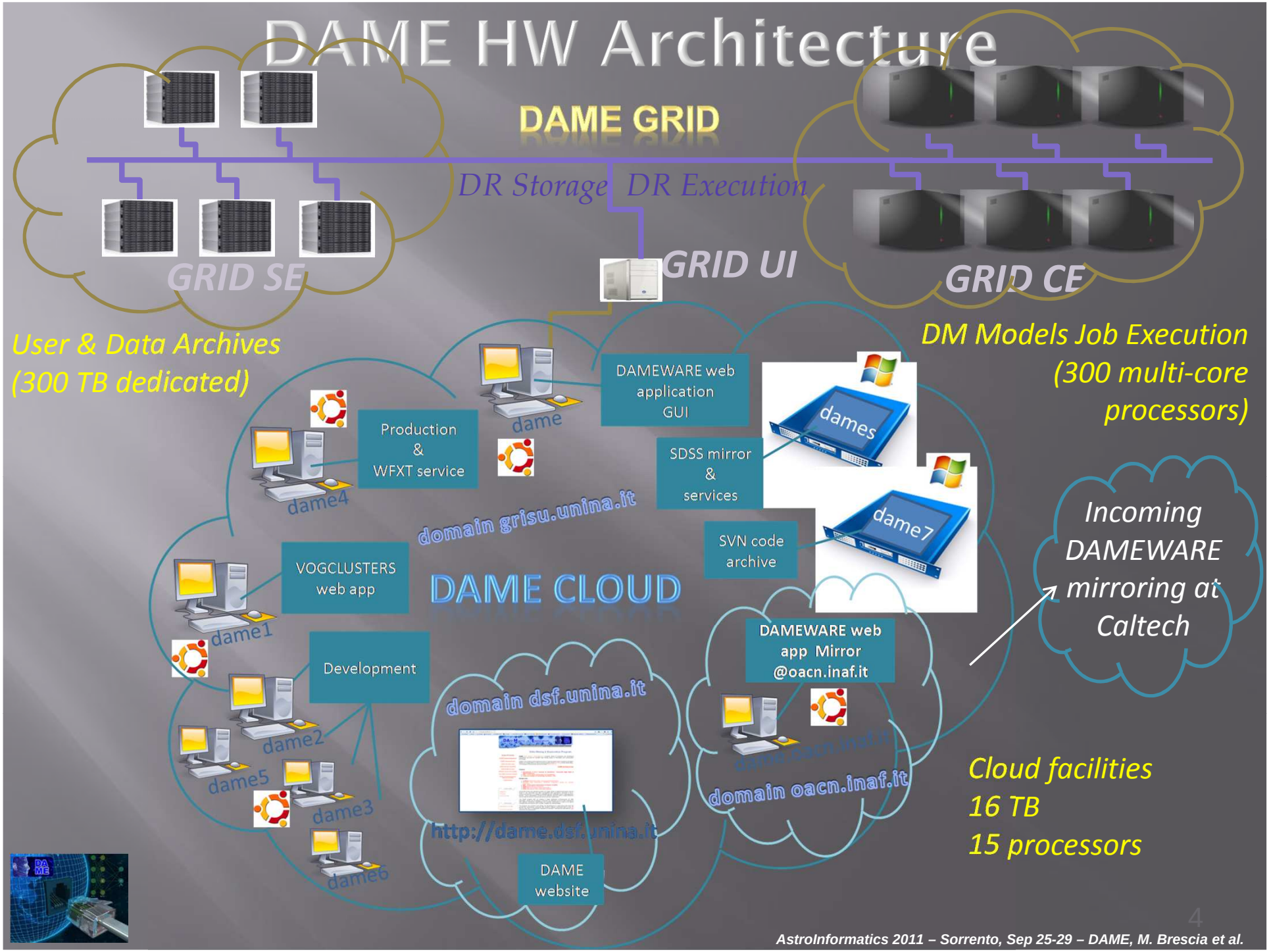
Multi-purpose data mining with machine learning Web App REsource

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



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

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

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

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

unification in a single framework of:

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

Web as a participating and sharing information platform

Machine Learning scalable tools on Massive Datasets

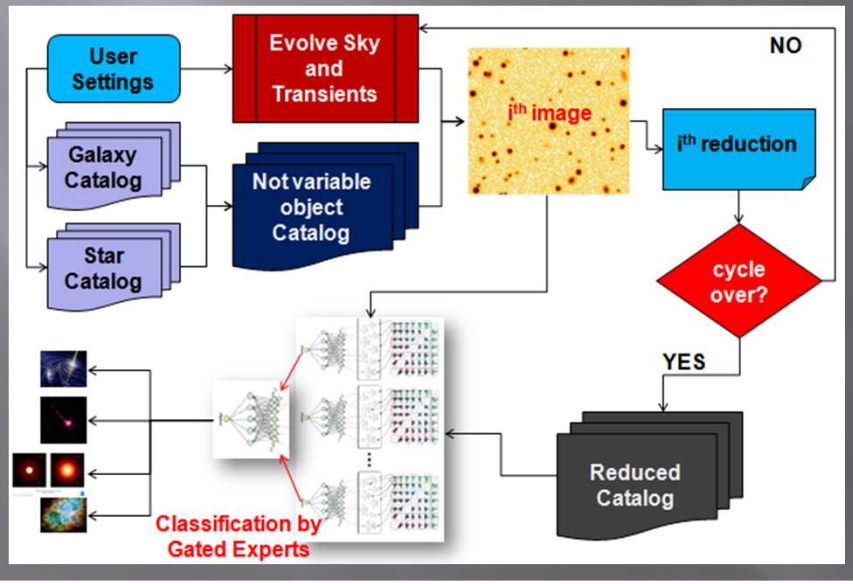
Machine-based interactions (REST, SOAP) based on standards (PMML, WSDL, XML)

DAME Projects: STraDiWA

Sky Transient Discovery Web App http://dame.dsf.unina.it/dame_td.html

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;

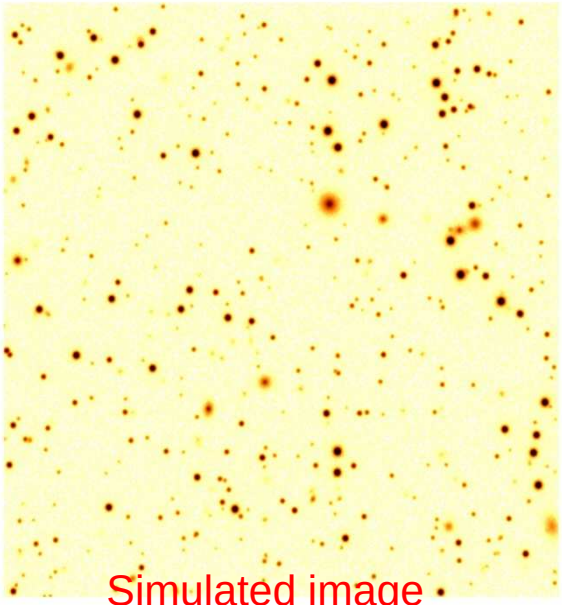


Next steps:

- 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 then a certain threshold and as galaxies the objects with PSFEx_MODEL grater equal that. The best is obtained with value 0.01, although with a little contamination for galaxies.

Bin	$S_{CLASS\_STAR>=0.98}/S_{extracted}$	$G_{CLASS\_STAR<0.98}/G_{extracted}$
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	$S_{PSFEx\_MODEL<0.01}/S_{extracted}$	$G_{PSFEx\_MODEL>=0.01}/G_{extracted}$
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

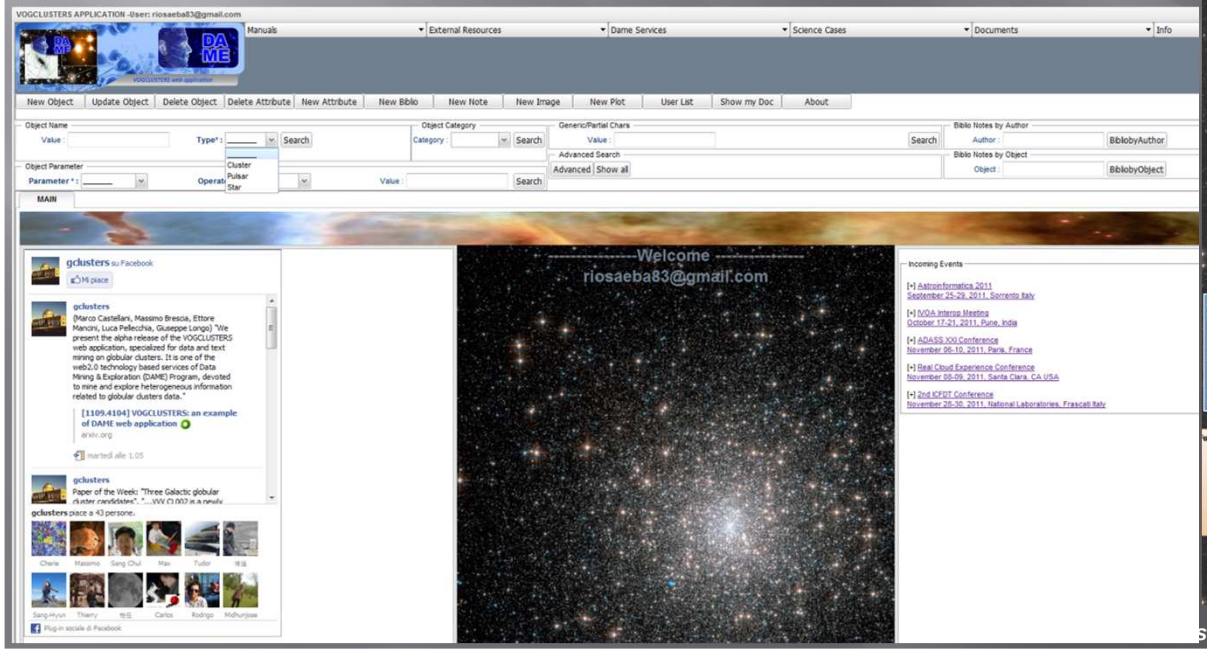
DAME Projects: VOGCLUSTERS

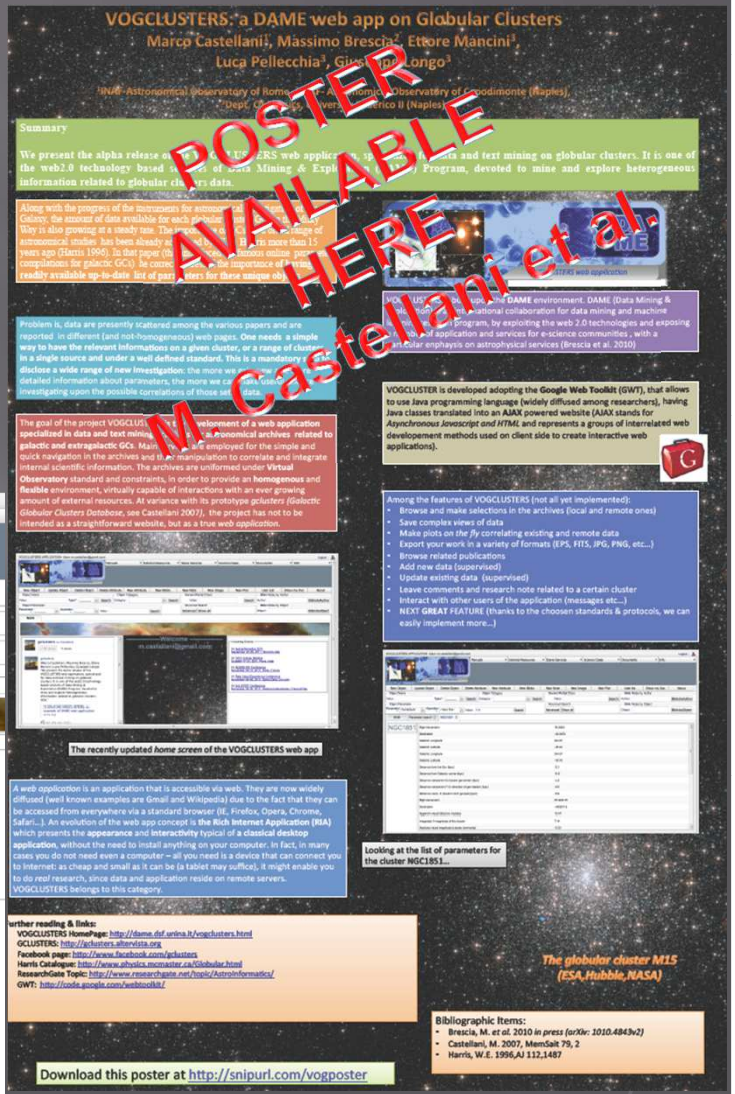
Globular Clusters Mining Web App

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.);





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

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;

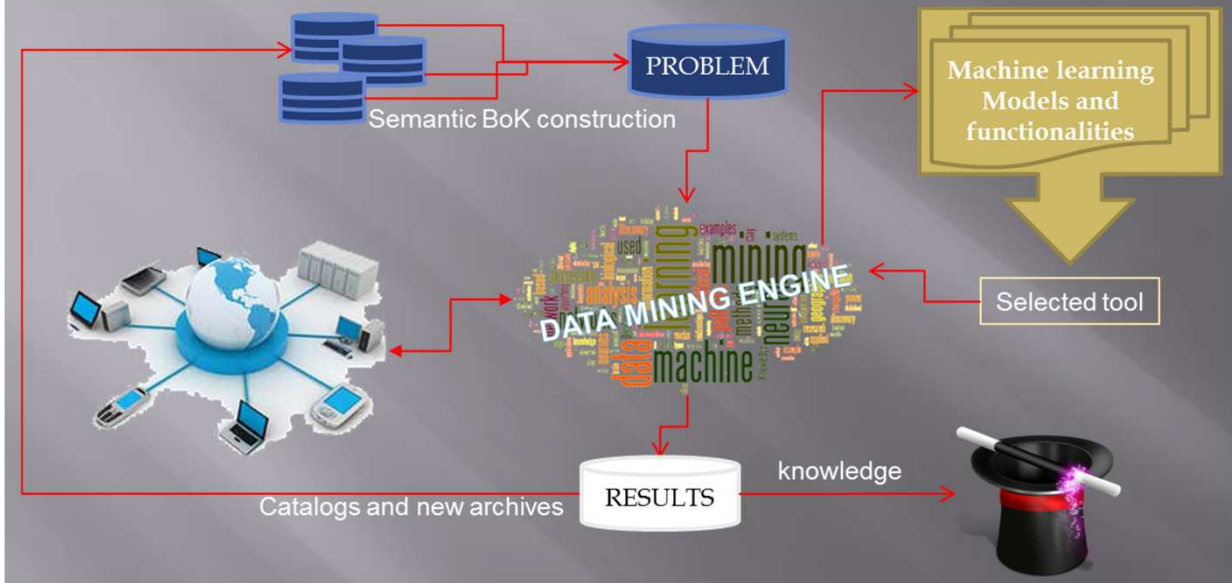
- **EUCLID Mission (*work in progress, co-head in Data Quality, coordinated by SGS, F. Pasian*)**

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

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



Multi Layer Perceptron trained by:
- Back Propagation
- Quasi Newton
- Genetic Algorithm

Support Vector Machines

Genetic Algorithms

Self Organizing Feature Maps

K-Means

Multi-layer Clustering

Principal Probabilistic Surfaces

Classification

Regression

Clustering

Feature Extraction

next …

Bayesian Networks

Decision Trees

MLP with Levenberg-Marquardt

DAMEWARE SW Architecture

FRONT END
WEB-APPL. GUI

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

XML

Restful, Stateless Web Serviceexperiment data, Working flow trigger and Supervision Servlets b Based on XML protocol

FRAMEWORK
WEB-SERVICE Suite CTRL

servlet

CALL

DATA MINING MODELS
Model-Functionality LIBRARY RUN

regression

MLP

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

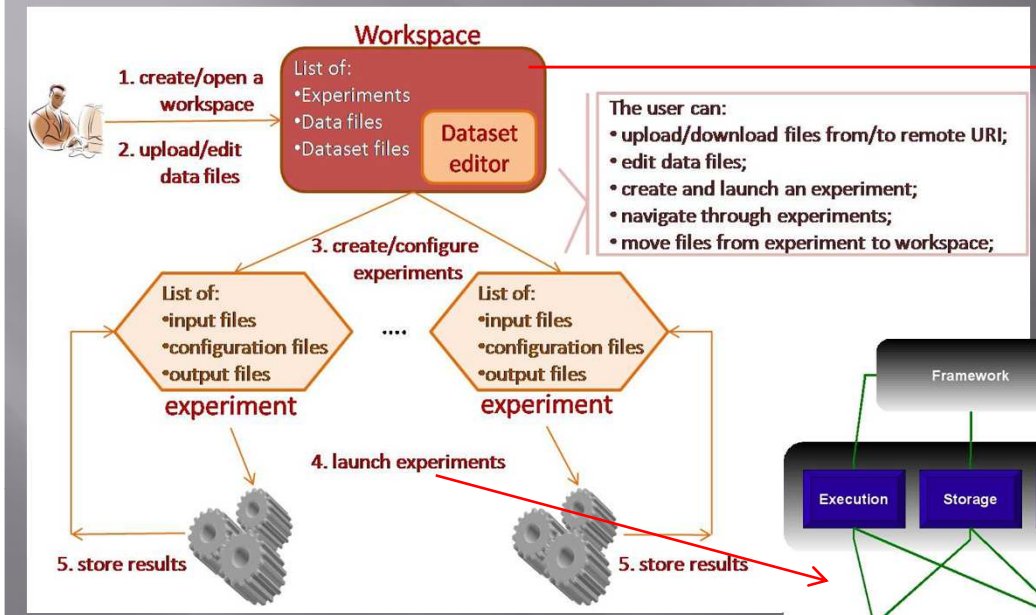
11

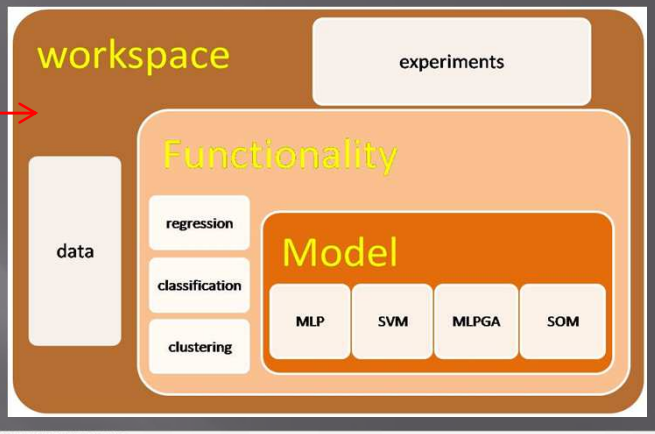
DAMEWARE fundamentals

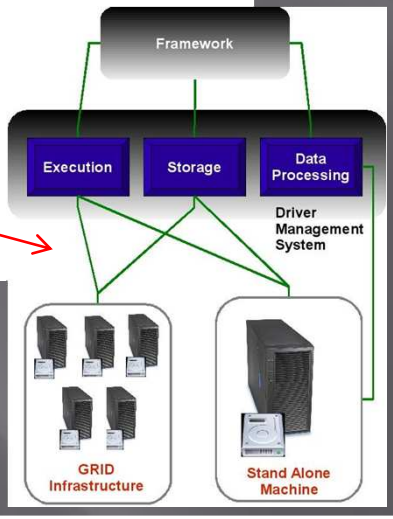
Based on the X-Informatics paradigm, it is multi-disciplinary platform (until now X = Astro)

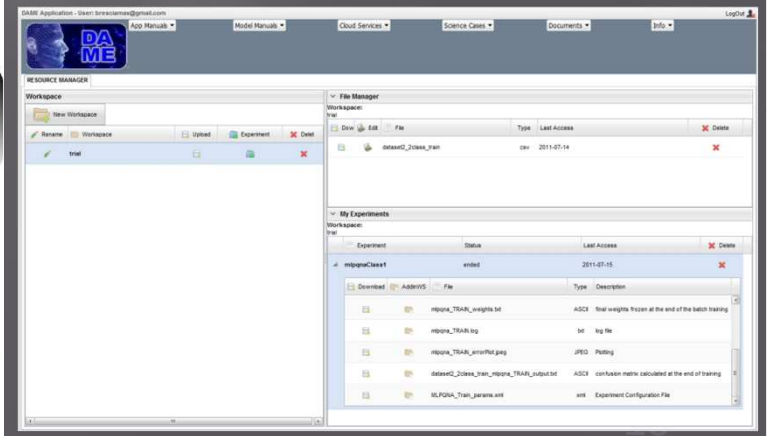
End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

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

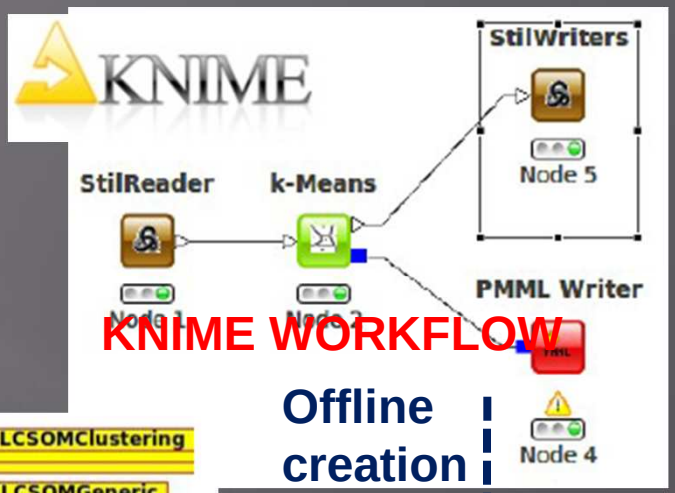






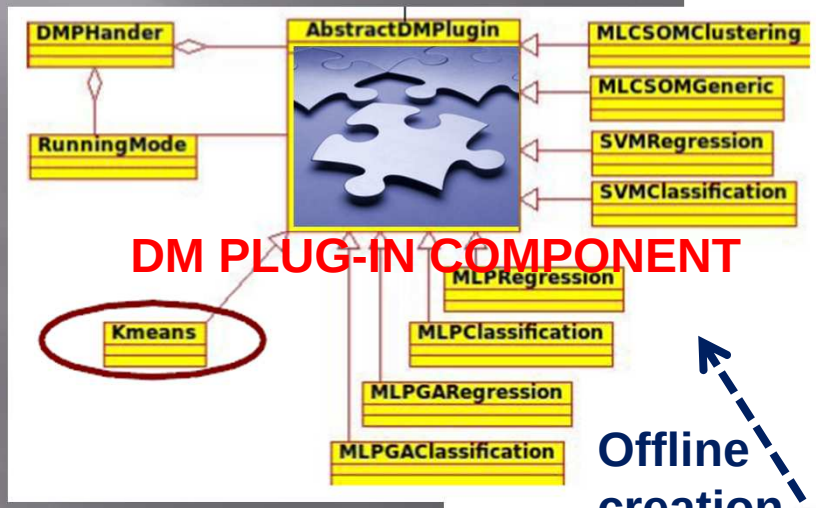


K-Means (through KNIME)



KNIME WORKFLOW

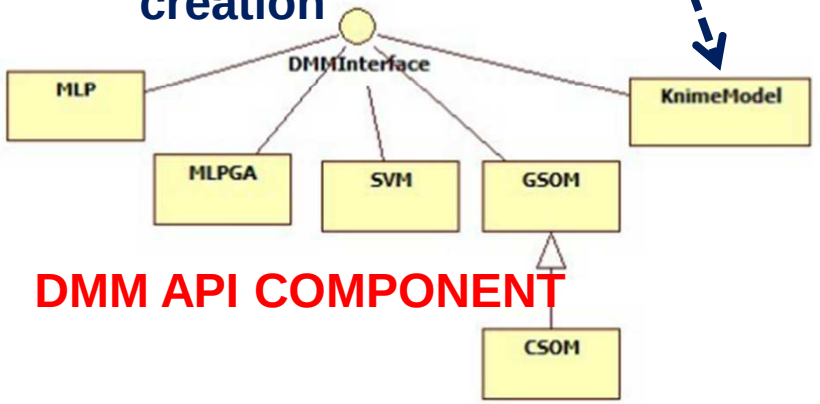
Offline creation



DM PLUG-IN COMPONENT

Offline creation

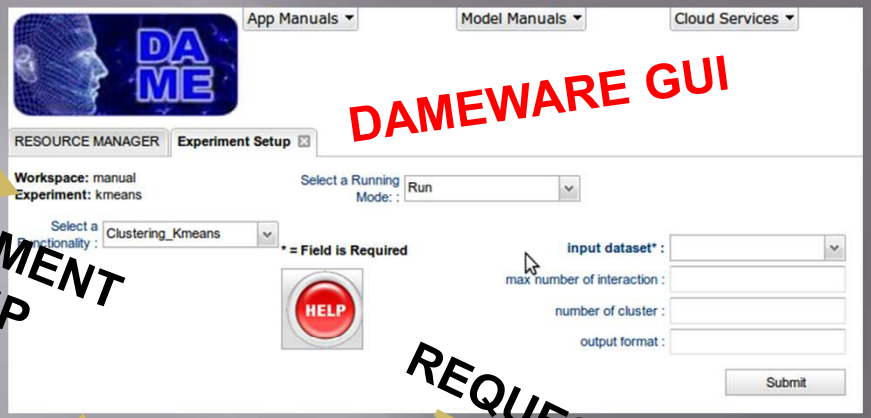
Offline creation



DMM API COMPONENT

K-Means (through KNIME)

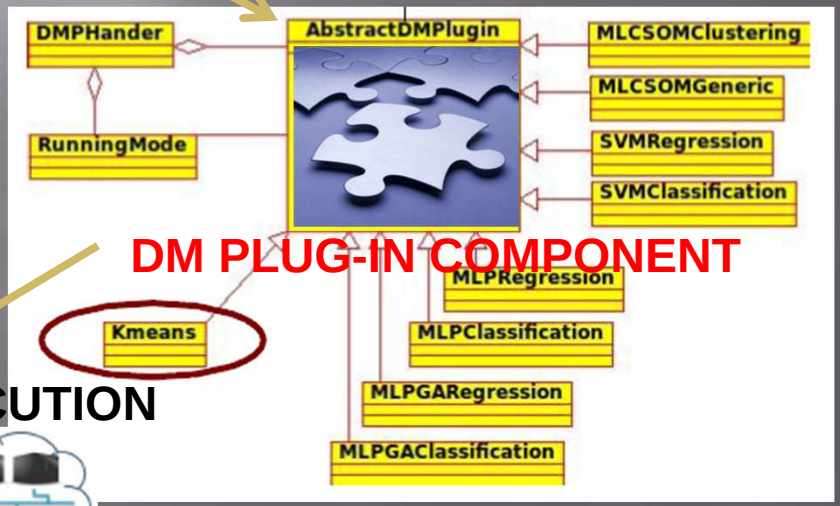
EXPERIMENT SETUP



DAMEWARE GUI

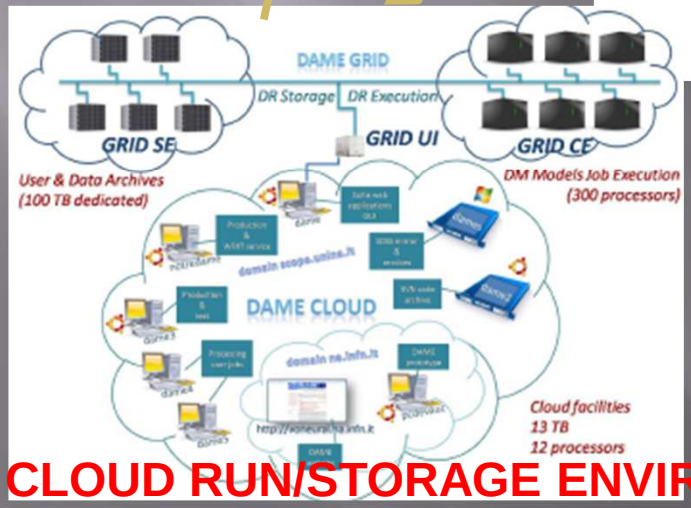
REQUEST

OUTPUT



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;*

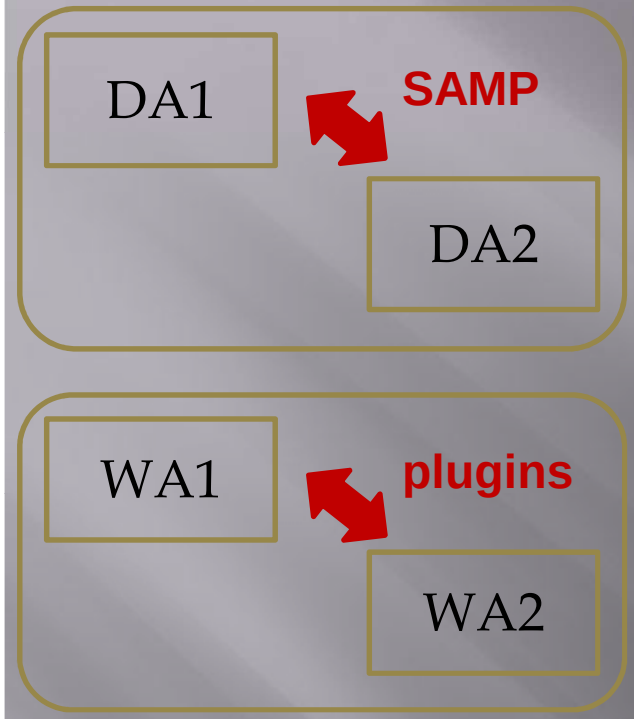
AGN Classification in the SDSS (classification-SVM)
Using the GRID to execute 110 jobs on 110 WN, the SVM model produces a classification of different types of AGN using SDSS photometric data and spectroscopic subsamples
- *Paper in preparation;*

Search for Candidate Quasars in the SDSSS (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