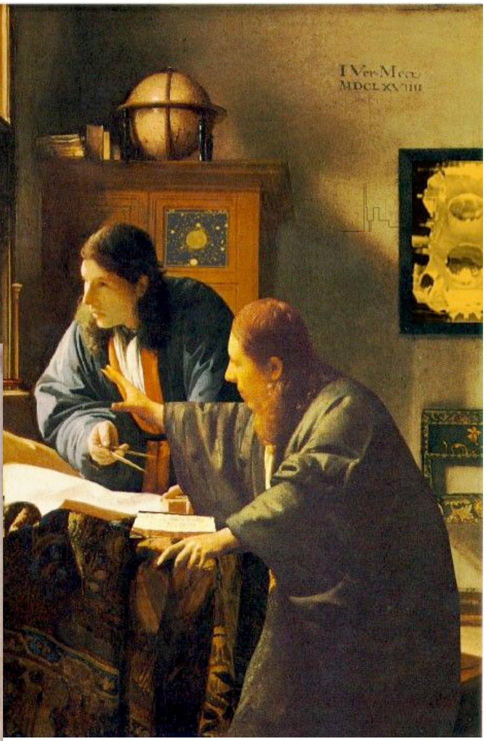
Astronomical Data Mining: Renaissance or dark age?

Giuseppe Longo & the DAME Group
longo@na.infn.it



We know that DM, ML, AI are a must…

Recent past	Now	Near Future ?
Separated archives and data centers *(few TB)*	**Federated archives and data centers** *(10 – 100 Tbyte)*	**Virtual Observatory** *(> 1 Pbyte)*
No common standards (*.fits)	Common standards (*.fits, *.vot, etc.)	Common standards (*.fits, *.vot, etc.)
Little bandwith (10/50 Kb s^{-1})	**Larger bandwith (100-1000 Kb s^{-1})** *(last mile problem)*	Largerbandwith (> 1-10 Gb s^{-1})
Single CPU processing	Still single CPU processing	GRID/Cloud computing

Research praxis

Few objects , few information *(parameter space ~ 10 features)*	**Many objects , much information** *(parameter space > 100 features)*	**Whole sky, multi-λ, multi epoch catalogues** *(parameter space > 100 features)*
Traditional statistics	**Multi variate statistics**	**Statistical Pattern Recognition (DM and ML)**

This is only a part of the game

(*size and not complexity driven)*

Does the community know?

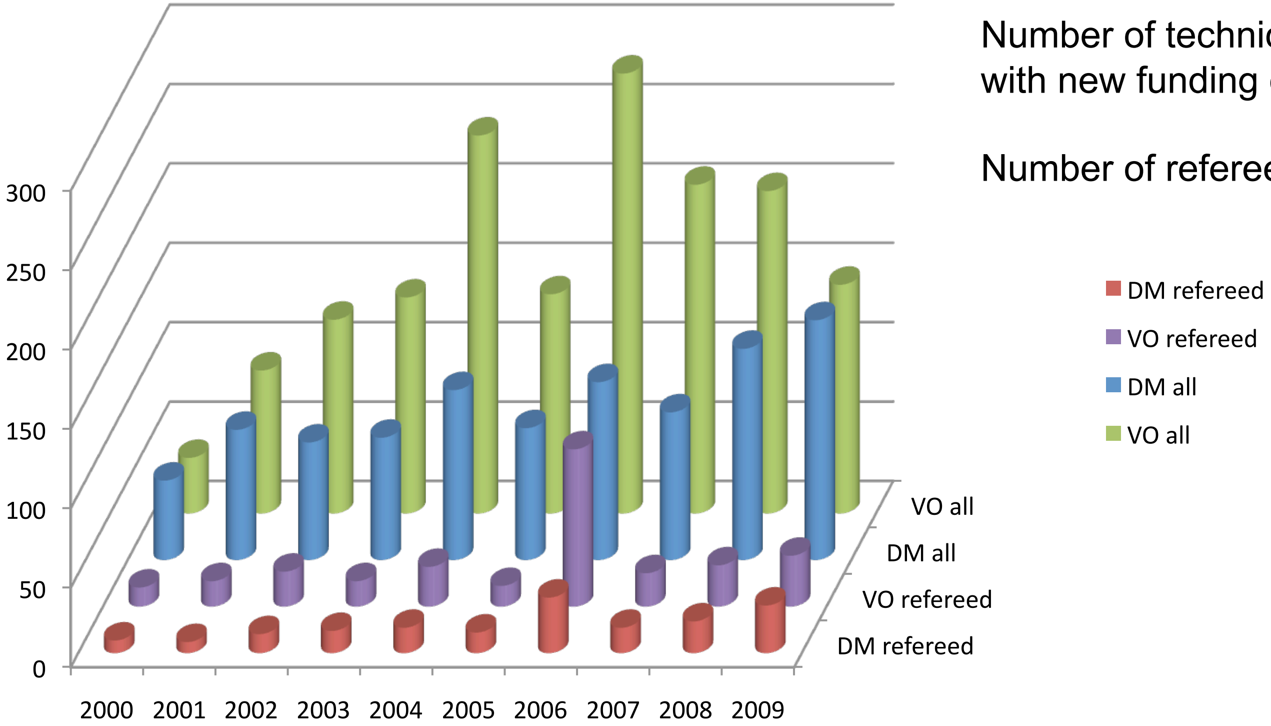
Most people who work in astronomical DM/AI are in this room.

They have implemented methods which are open to the community and have used them to produce science.

BUT

Little use – few citations

(number of citations increases if you avoid the terms AI, ML or DM in writing the paper)

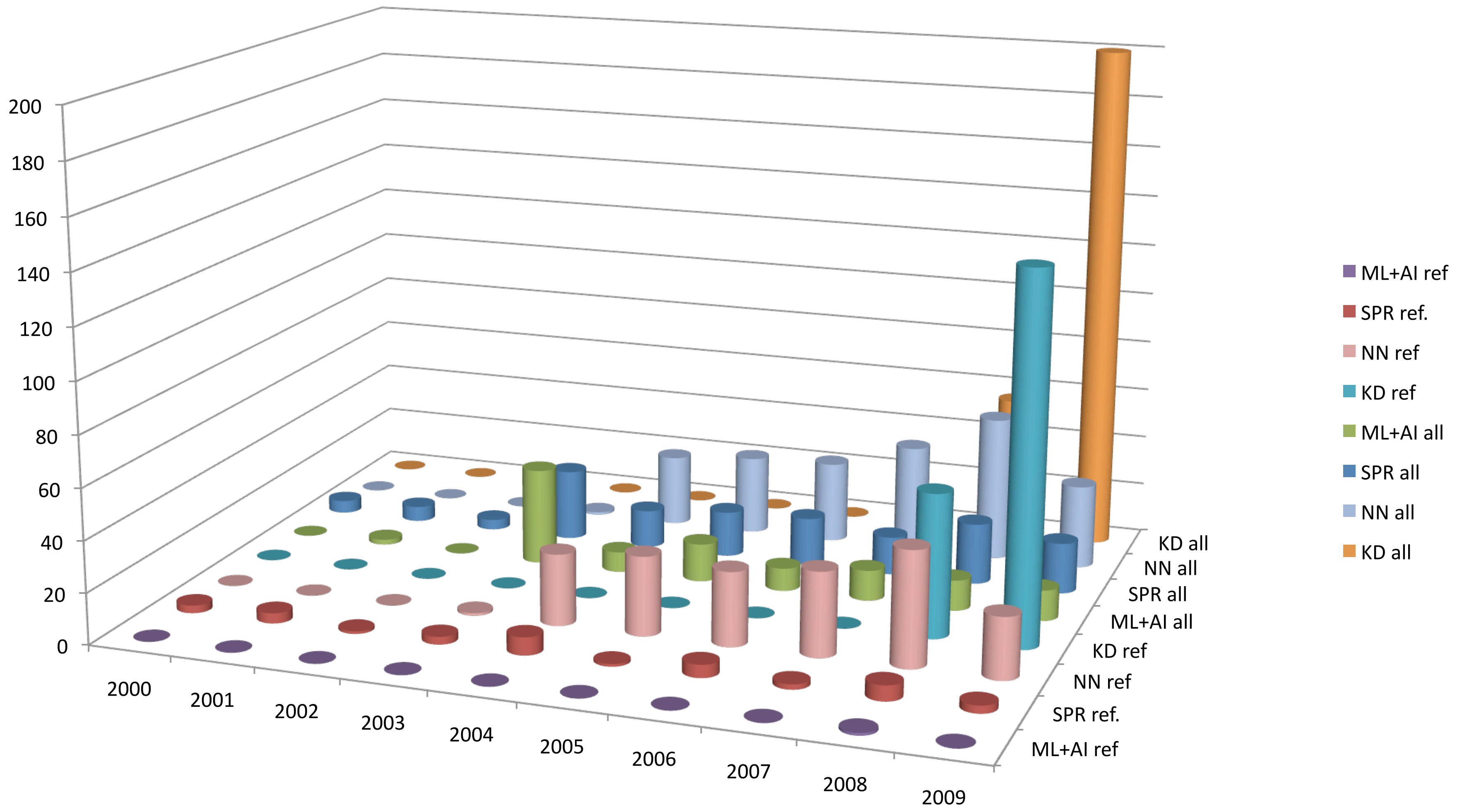


Number of technical/algorithmic papers increases with new funding opportunities

Number of refereed papers remains constant

Most of the work remains at the implementation stage (computer Science and algorithm development) and does not enter the “science production” stage…

Usage of terms is highly fashionable



Out of one thousand papers checked (galaxies, observational cosmology, survey) over the last two years:

DM could be applied or involved in at least 30% of them

Algorithms

Restricted choice of algorithms (MLPs, SVM, Kernel methods, Genetic algoritms (few models), K Means, PPS, SOM …)

Astronomers know little statistics, forget about SPR, DM, etc…

Just a few astronomers go beyond the introductory chapters of the Bishop.

Restricted choice of problems: *the situation is not changed much in the last decade*

Tagliaferri et al. 2003	Ball & Brunner 2009	BoK
S/G separation	S/G separation	Y
Morphological classification of galaxies *(shapes, spectra)*	Morphological classification of galaxies *(shapes, spectra)*	Y
Spectral classification of stars	Spectral classification of stars	Y
Image segmentation	-----	
Noise removal *(grav. waves, pixel lensing, images)*	-----	
Photometric redshifts *(galaxies)*	Photometric redshifts *(galaxies, QSO's)*	Y
Search for AGN	Search for AGN and QSO	Y
Variable objects	Time domain	
Partition of photometric parameter space for specific group of objects	Partition of photometric parameter space for specific group of objects	Y
Planetary studies (asteroids)	Planetary studies (asteroids)	Y
Solar activity	Solar activity	Y
Interstellar magnetic fields	----	
Stellar evolution models	----	

Limited number of problems due to limited number of reliable BoKs

Bases of knowledge

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

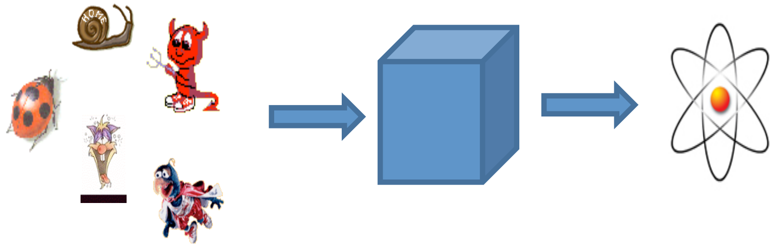
So far

- 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

Community believes AI/DM methods are black boxes

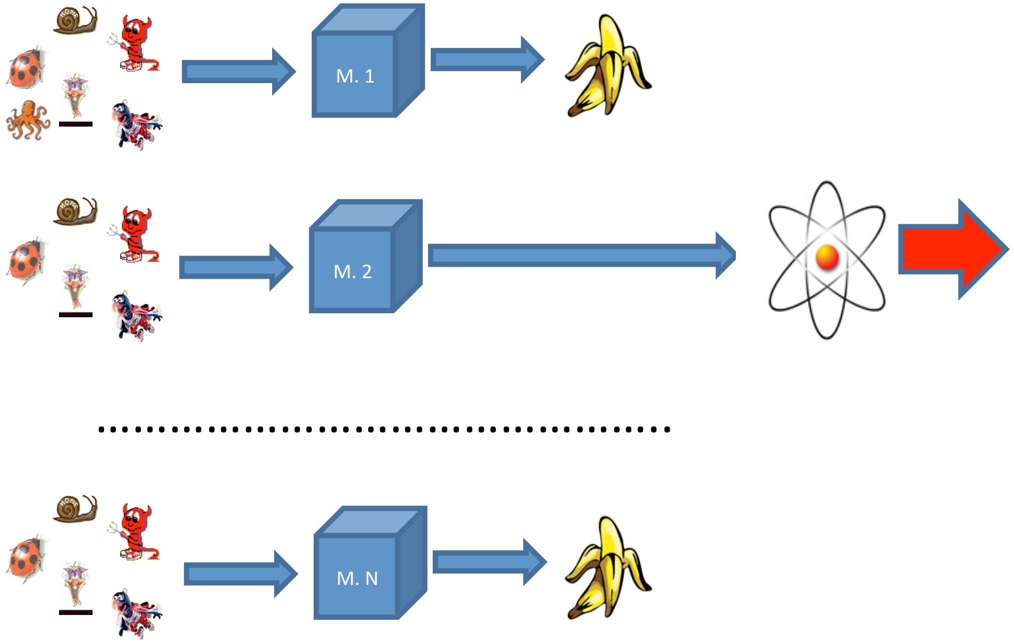
You feed in something, and obtain patters, trends, i.e. knowledge….



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.)



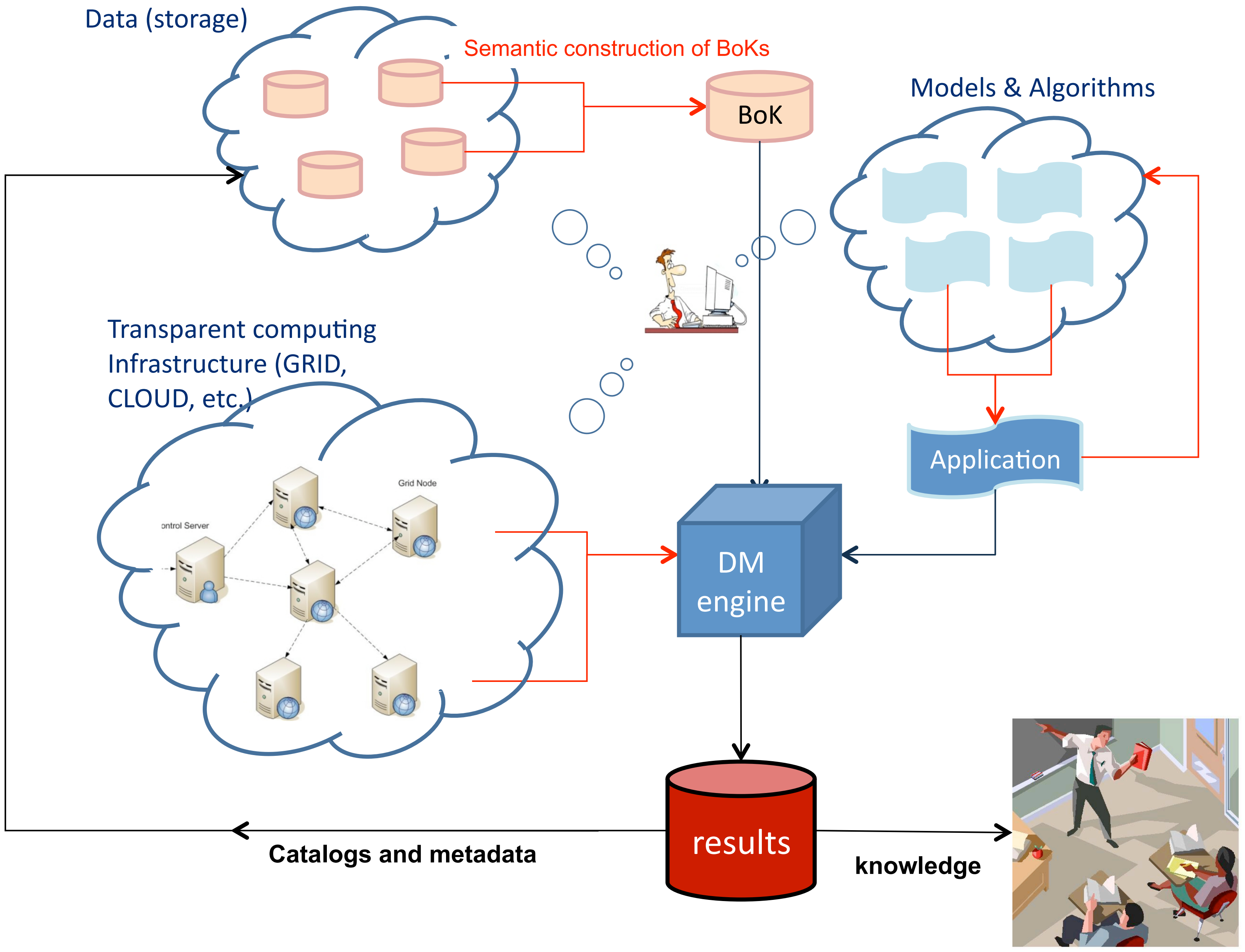
Formation of a new generation of experts

(…… suggesting the solutions)

AND/OR

Implementation of a second generation of tools

A break-down of an effective DM process



Data (storage)

Semantic construction of BoKs

BoK

Models & Algorithms

Transparent computing Infrastructure (GRID, CLOUD, etc.)

Application

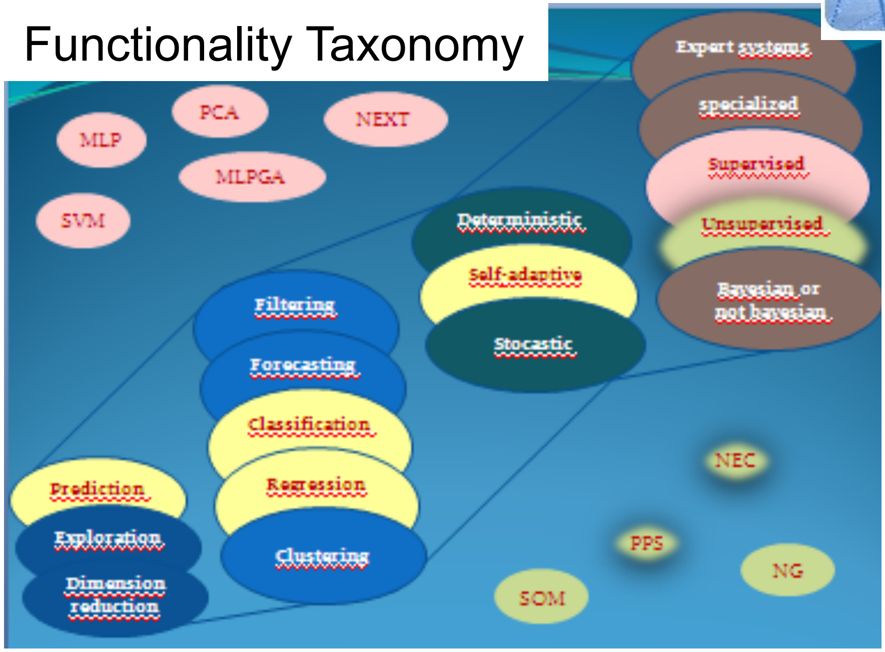
DM engine

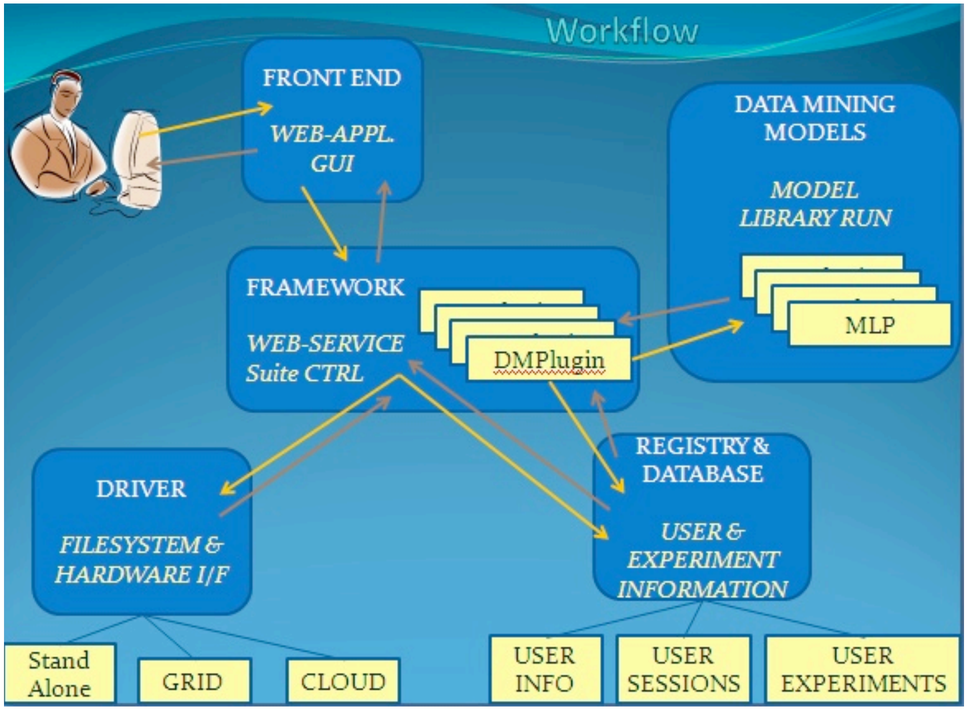
results

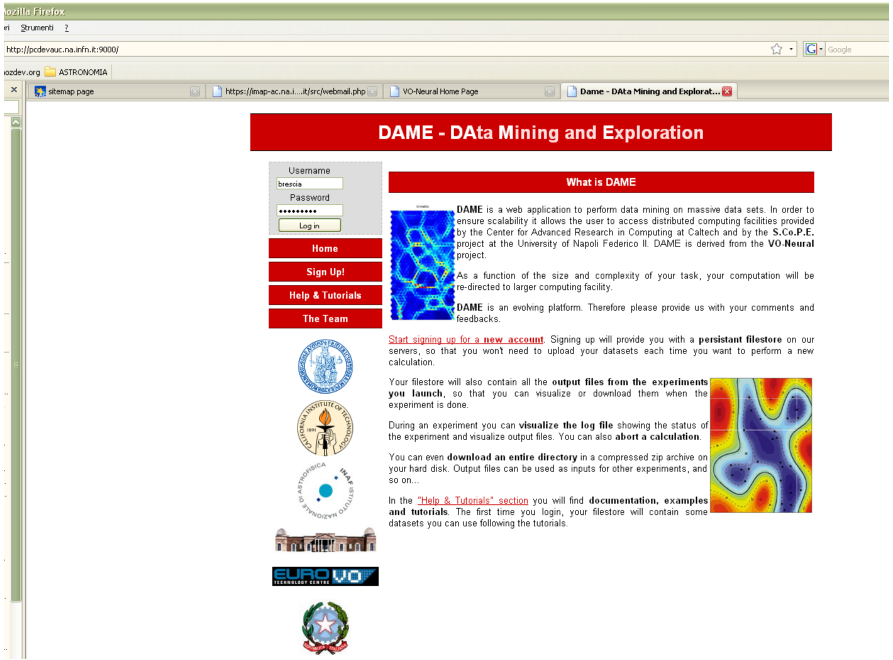
Catalogs and metadata

knowledge

Functionality Taxonomy






ozilla Firefox
ri Strumenti ?
http://pcdevauc.na.infn.it:9000/
ozdev.org ASTRONOMIA
sitemap page
https://imap-ac.na.i....it/src/webmail.php
VO-Neural Home Page
Dame - DAta Mining and Explorat...

DAME - DAta Mining and Exploration

Username: brescia
Password: •••••••••
Log in

What is DAME

DAME is a web application to perform data mining on massive data sets. In order to ensure scalability it allows the user to access distributed computing facilities provided by the Center for Advanced Research in Computing at Caltech and by the **S.Co.P.E.** project at the University of Napoli Federico II. DAME is derived from the **VO-Neural** project.

As a function of the size and complexity of your task, your computation will be re-directed to larger computing facility.

DAME is an evolving platform. Therefore please provide us with your comments and feedbacks.

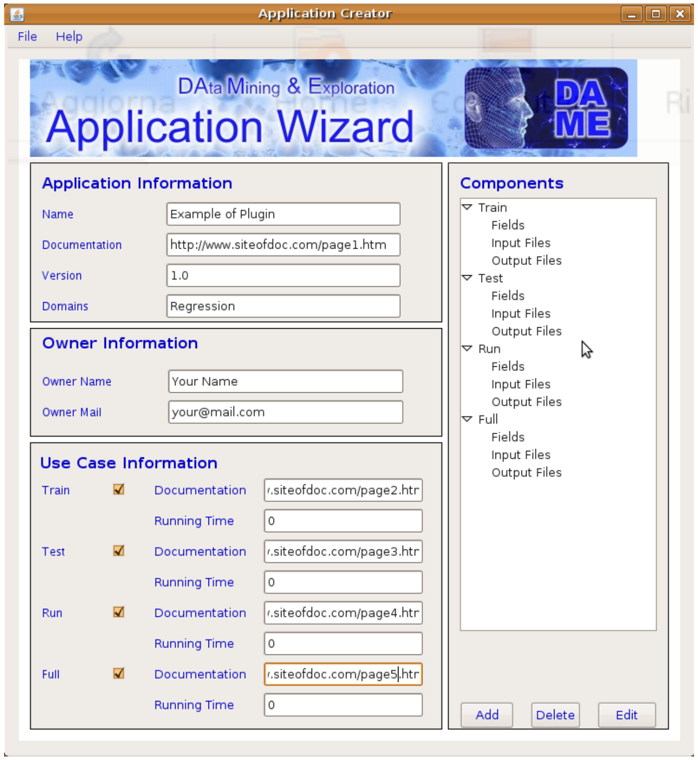
Start signing up for a **new account**. Signing up will provide you with a **persistant filestore** on our servers, so that you won't need to upload your datasets each time you want to perform a new calculation.

Your filestore will also contain all the **output files from the experiments you launch**, so that you can visualize or download them when the experiment is done.

During an experiment you can **visualize the log file** showing the status of the experiment and visualize output files. You can also **abort a calculation**.

You can even **download an entire directory** in a compressed zip archive on your hard disk. Output files can be used as inputs for other experiments, and so on...

In the "Help & Tutorials" section you will find **documentation, examples and tutorials**. The first time you login, your filestore will contain some datasets you can use following the tutorials.



Application Information

Name	Example of Plugin
Documentation	http://www.siteofdoc.com/page1.htm
Version	1.0
Domains	Regression

Owner Information

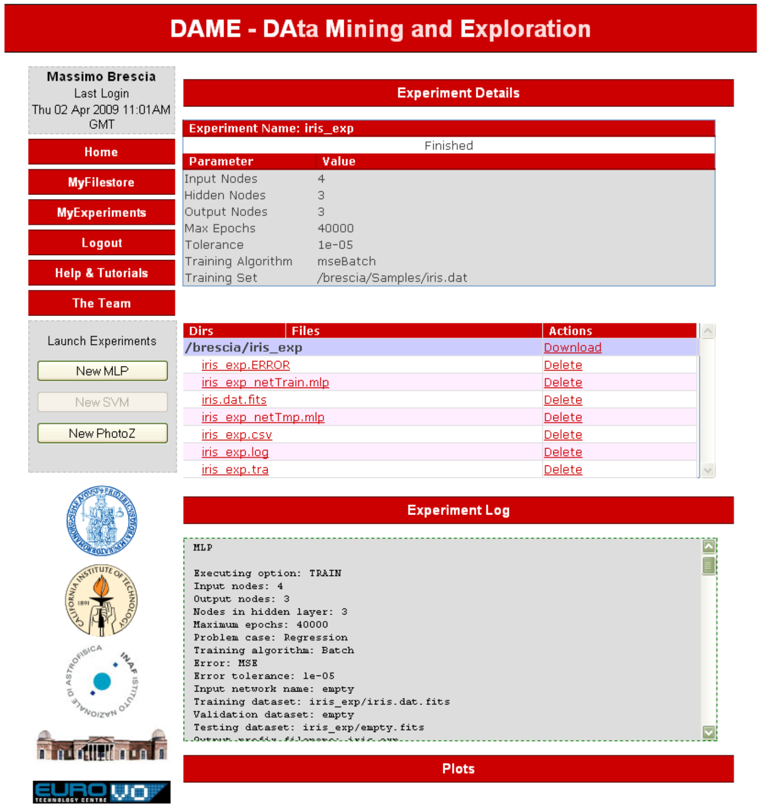
Owner Name	Your Name
Owner Mail	your@mail.com

Use Case Information

Train	☑	Documentation	‹.siteofdoc.com/page2.htr
		Running Time	0
Test	☑	Documentation	‹.siteofdoc.com/page3.htr
		Running Time	0
Run	☑	Documentation	‹.siteofdoc.com/page4.htr
		Running Time	0
Full	☑	Documentation	‹.siteofdoc.com/page5.htr
		Running Time	0

Components

- Train
  - Fields
  - Input Files
  - Output Files
- Test
  - Fields
  - Input Files
  - Output Files
- Run
  - Fields
  - Input Files
  - Output Files
- Full
  - Fields
  - Input Files
  - Output Files



DAME - DAta Mining and Exploration

Massimo Brescia
Last Login
Thu 02 Apr 2009 11:01AM GMT

Experiment Details

Experiment Name: iris_exp

Finished

Parameter	Value
Input Nodes	4
Hidden Nodes	3
Output Nodes	3
Max Epochs	40000
Tolerance	1e-05
Training Algorithm	mseBatch
Training Set	/brescia/Samples/iris.dat

Dirs	Files	Actions
/brescia/iris_exp		Download
	iris_exp.ERROR	Delete
	iris_exp_netTrain.mlp	Delete
	iris.dat.fits	Delete
	iris_exp_netTmp.mlp	Delete
	iris_exp.csv	Delete
	iris_exp.log	Delete
	iris_exp.tra	Delete

Experiment Log

```
MLP

Executing option: TRAIN
Input nodes: 4
Output nodes: 3
Nodes in hidden layer: 3
Maximum epochs: 40000
Problem case: Regression
Training algorithm: Batch
Error: MSE
Error tolerance: 1e-05
Input network name: empty
Training dataset: iris_exp/iris.dat.fits
Validation dataset: empty
Testing dataset: iris_exp/empty.fits
Output prefix filename: iris_exp
```

Plots

Data Mining as a social networking problem ….

During Strasbourg Interop Meeting was started an Interest Group on Data Mining aimed at definining how to effectively bring KDD under the Vobs Umbrella…

M. Brescia, G. Longo (Co-chair), S.G. Djorgovski (co-Chair), K. Borne,
C. Donalek, M. Graham, G. Fabbiano, I. Kilingorov, A. Lawrence, R. Smareglia, &
others…

First step: TO WRITE A 10-12 pages document assessing how to proceed