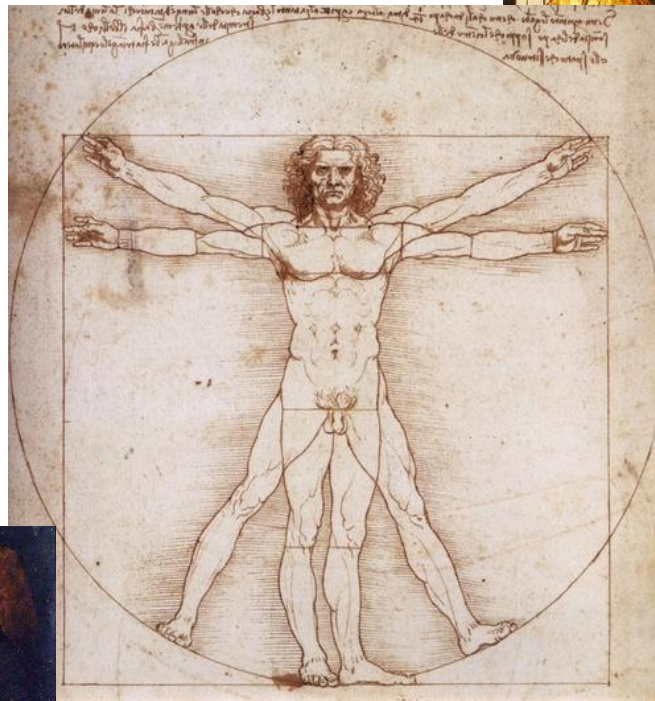
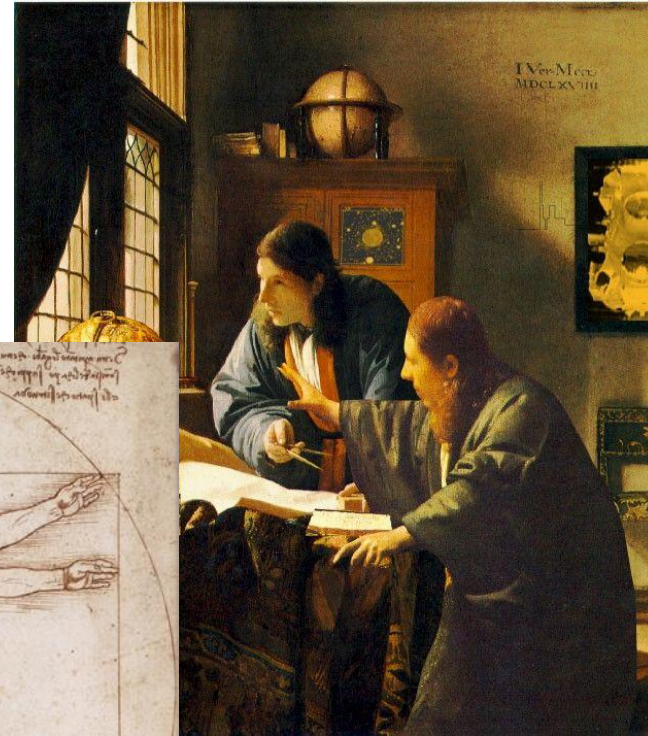


# Astronomical Data Mining: Renaissance or dark age?

Giuseppe Longo & the DAME Group

[longo@na.infn.it](mailto:longo@na.infn.it)



# We know that DM, ML, AI are a must...

Recent past	Now	Near Future ?
<p><b>Separated archives and data centers</b> <i>(few TB)</i></p> <p>No common standards (*.fits)</p> <p>Little bandwidth (10/50 Kb s<sup>-1</sup>)</p> <p>Single CPU processing</p>	<p><b>Federated archives and data centers</b> <i>(10 – 100 Tbyte)</i></p> <p>Common standards (*.fits, *.vot, etc.)</p> <p><b>Larger bandwidth (100-1000 Kb s<sup>-1</sup>)</b> <i>(last mile problem)</i></p> <p>Still single CPU processing</p>	<p><b>Virtual Observatory</b> <i>(&gt; 1 Pbyte)</i></p> <p>Common standards (*.fits, *.vot, etc.)</p> <p>Larger bandwidth (&gt; 1-10 Gb s<sup>-1</sup>)</p> <p>GRID/Cloud computing</p>
<p><b>Few objects , few information</b> <i>(parameter space ~ 10 features)</i></p> <p><b>Traditional statistics</b></p>	<p><b>Many objects , much information</b> <i>(parameter space &gt; 100 features)</i></p> <p><b>Multi variate statistics</b></p>	<p><b>Whole sky, multi-λ, multi epoch catalogues</b> <i>(parameter space &gt; 100 features)</i></p> <p><b>Statistical Pattern Recognition (DM and ML)</b></p> <p><b>This is only a part of the game</b> <i>(size and not complexity driven)</i></p>

## Research praxis

# Does the community know?

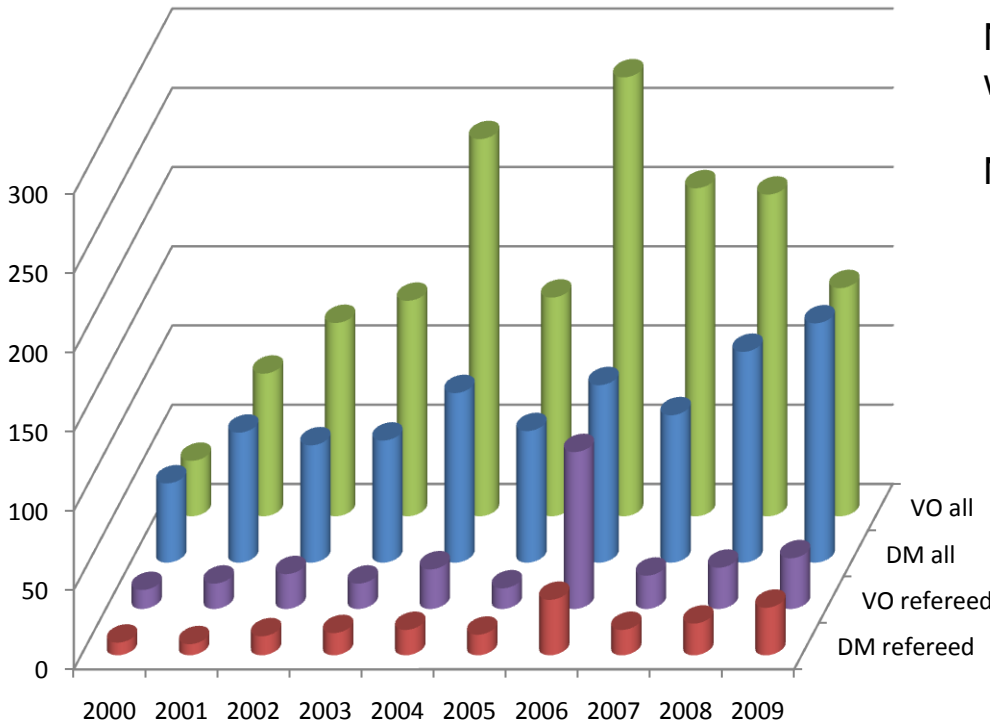
**Most people who work in astronomical DM/AI are in this room.**

*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)*



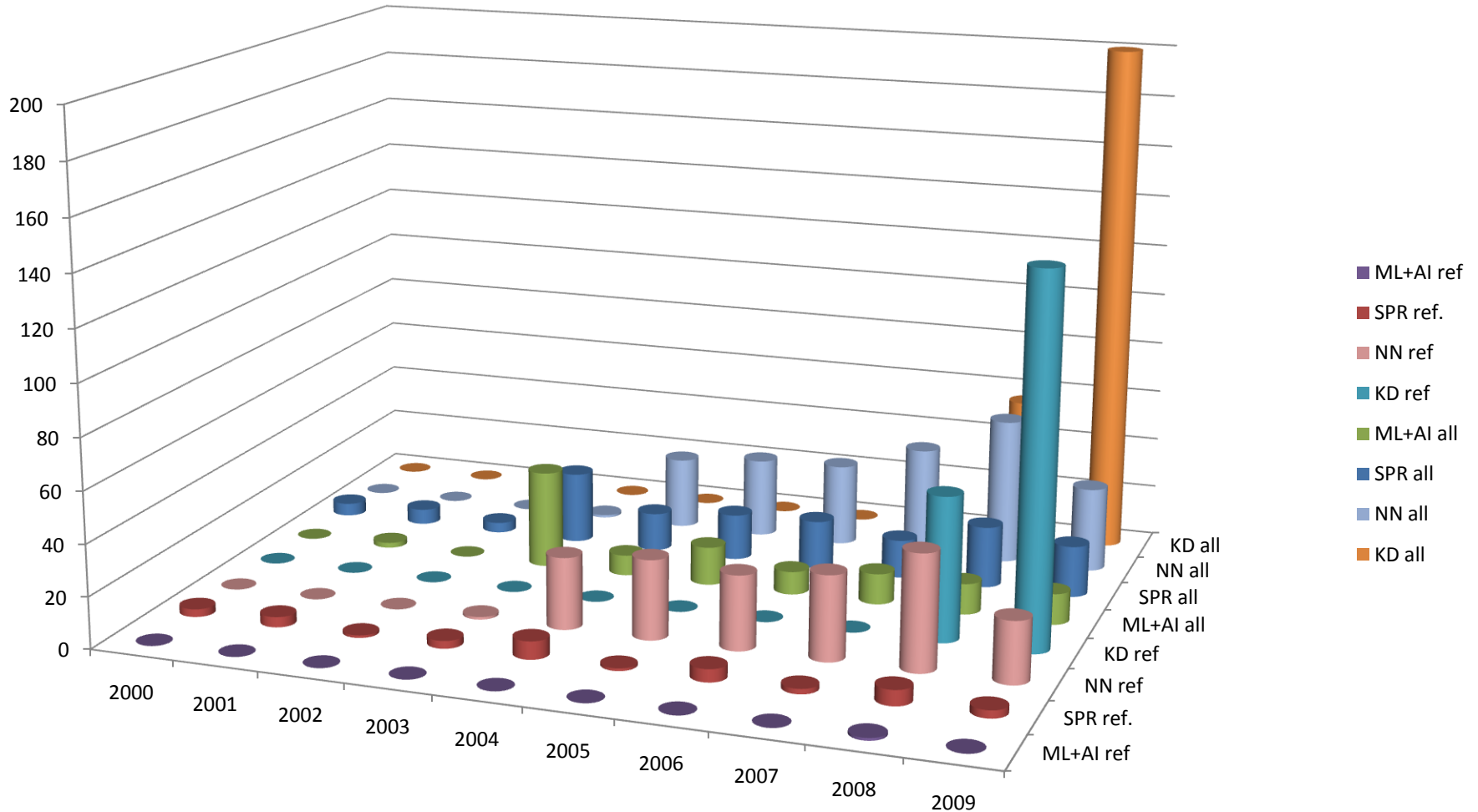
Number of technical/algorithmic papers increases with new funding opportunities

Number of refereed papers remains constant

- DM refereed
- VO refereed
- DM all
- VO all

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 algorithms (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 <i>(shapes, spectra)</i>	Morphological classification of galaxies <i>(shapes, spectra)</i>	Y
Spectral classification of stars	Spectral classification of stars	Y
Image segmentation	-----	
Noise removal <i>(grav. waves, pixel lensing, images)</i>	-----	
Photometric redshifts <i>(galaxies)</i>	Photometric redshifts <i>(galaxies, QSO's)</i>	Y
Search for AGN	Search for AGN and QSO	Y
Variable objects	<b>Time domain</b>	
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
<b>Interstellar magnetic fields</b>	----	
<b>Stellar evolution models</b>	----	

# 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)*

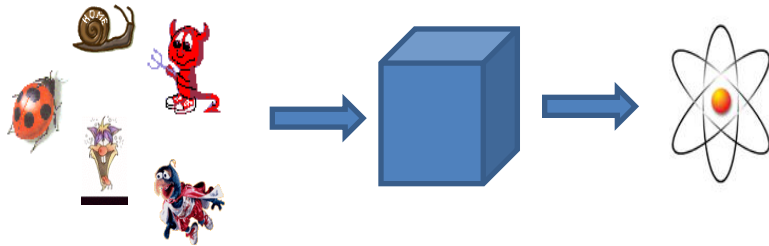
### So far

- 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

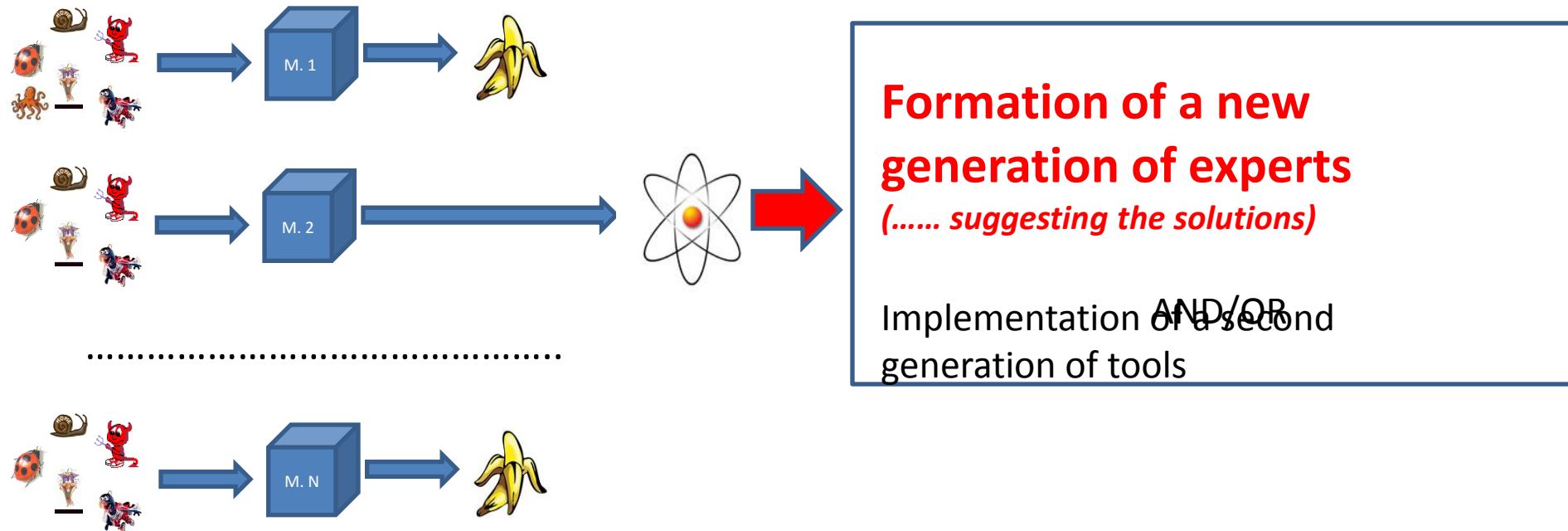
*You feed in something, and obtain patters, trends, i.e. knowledge....*



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 ...

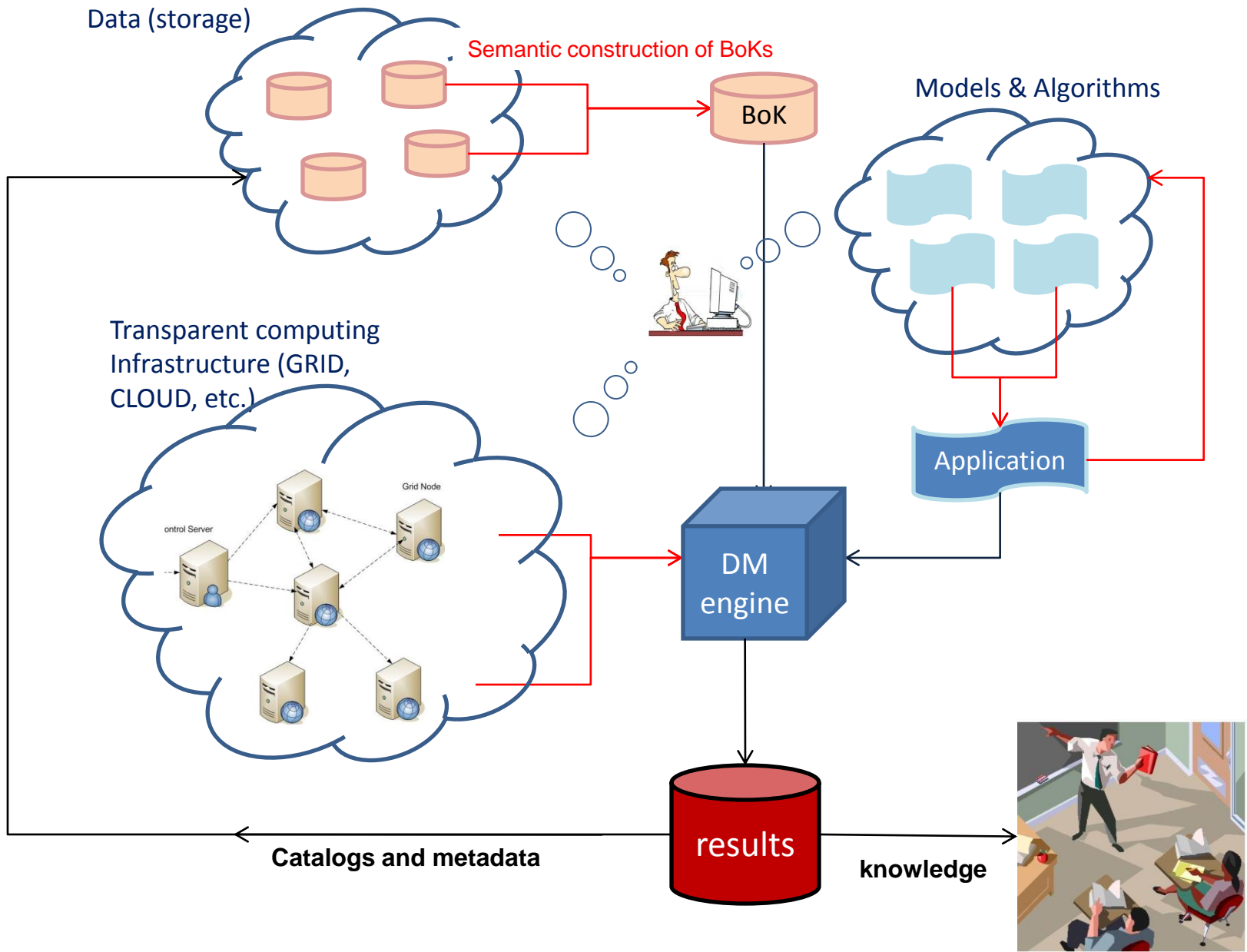
The r.m.s astronomer doesn't want to become a computer scientist or a mathematician (large survey projects overcome the problem)

Tools must run without knowledge of GRID/Cloud no personal certificates, no deep understanding of the DM tool etc. )

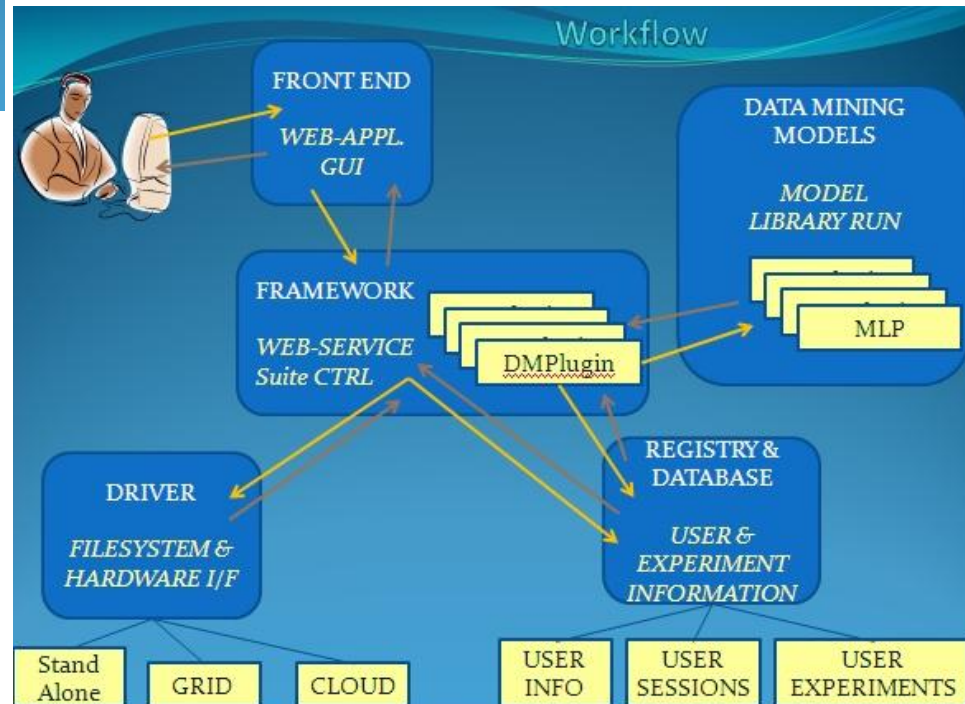
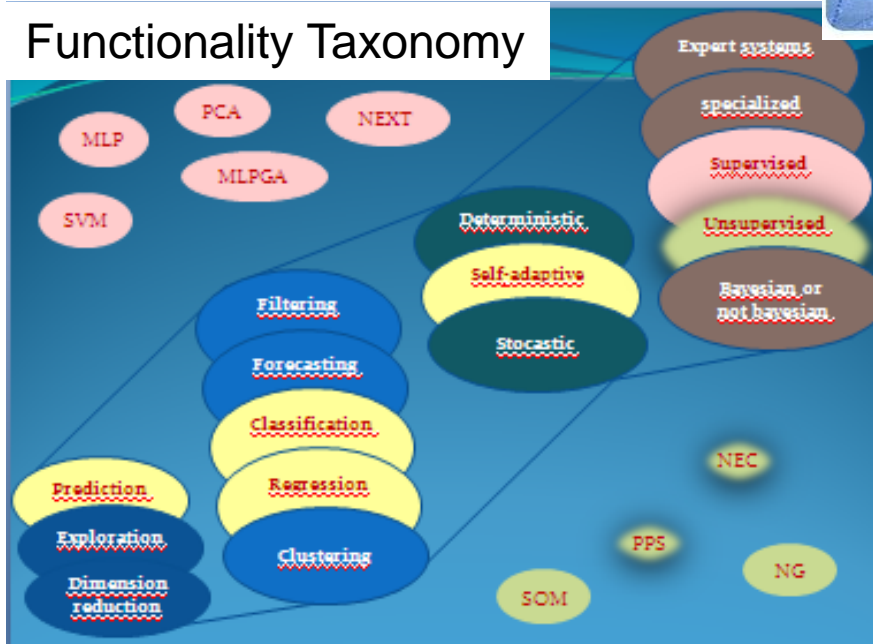


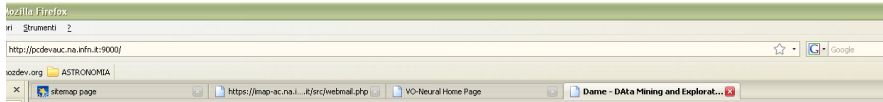


# A break-down of an effective DM process



## Functionality Taxonomy





## DAME - DAta Mining and Exploration

Username  
brescia

Password  
\*\*\*\*\*

Log in

**Home**

Sign Up!

Help & Tutorials

The Team

### 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.

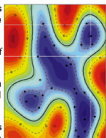
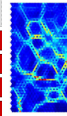
**Start signing up for a new account.** Signing up will provide you with a **persistent filestore** on our servers, so that you won't need to upload your datasets each time you want to perform a new calculation.

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 Creator

DAta Mining & Exploration  
**Application Wizard**

#### Application Information

Name:

Documentation:

Version:

Domains:

#### 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

#### Owner Information

Owner Name:

Owner Mail:

#### Use Case Information

Train	<input checked="" type="checkbox"/>	Documentation	<input type="text" value="http://www.siteofdoc.com/page2.htm"/>	Running Time	<input type="text" value="0"/>
Test	<input checked="" type="checkbox"/>	Documentation	<input type="text" value="http://www.siteofdoc.com/page3.htm"/>	Running Time	<input type="text" value="0"/>
Run	<input checked="" type="checkbox"/>	Documentation	<input type="text" value="http://www.siteofdoc.com/page4.htm"/>	Running Time	<input type="text" value="0"/>
Full	<input checked="" type="checkbox"/>	Documentation	<input type="text" value="http://www.siteofdoc.com/page5.htm"/>	Running Time	<input type="text" value="0"/>

## DAME - DAta Mining and Exploration

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

- Home**
- MyFilestore**
- MyExperiments**
- Logout**
- Help & Tutorials**
- The Team**

- Launch Experiments
- - 
  -

### 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		<a href="#">Download</a>
	<a href="#">iris_exp.ERROR</a>	<a href="#">Delete</a>
	<a href="#">iris_exp_netTrain.mlp</a>	<a href="#">Delete</a>
	<a href="#">iris.dat.fits</a>	<a href="#">Delete</a>
	<a href="#">iris_exp_netTmp.mlp</a>	<a href="#">Delete</a>
	<a href="#">iris_exp.csv</a>	<a href="#">Delete</a>
	<a href="#">iris_exp.log</a>	<a href="#">Delete</a>
	<a href="#">iris_exp.tra</a>	<a href="#">Delete</a>

### 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: iris_exp/iris_exp_train.csv
```

### Plots



Data Mining as a social networking problem ....





During Strasbourg Interop Meeting was started an Interest Group on Data Mining aimed at defining 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