



# The VO-Neural/DAME infrastructure: an integrated data mining system support for the e-science community



G. Longo – M. Brescia  
&  
Project Team



*INAF – Osservatorio Astronomico di Capodimonte*

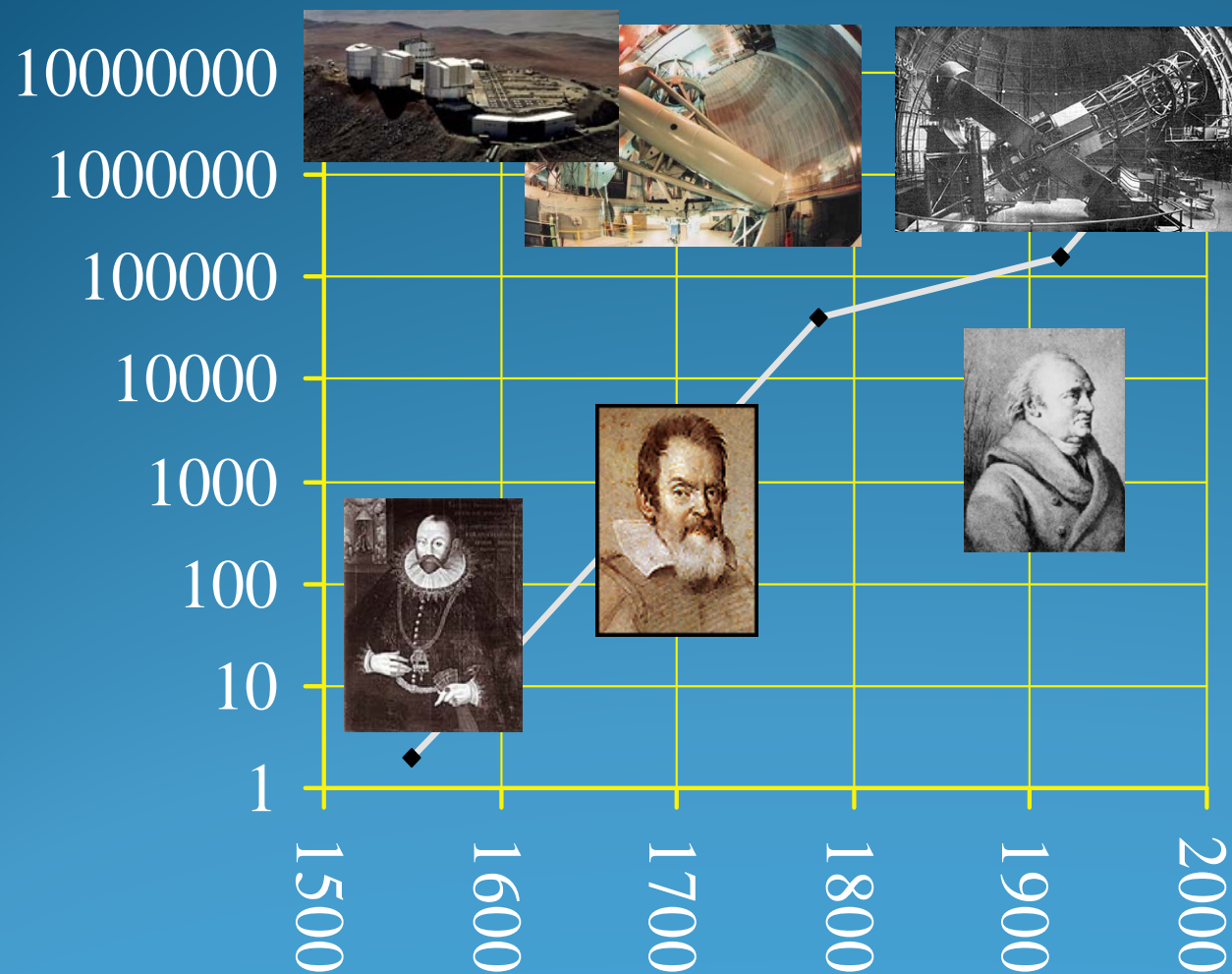
*Dipartimento di Fisica – Università degli Studi di Napoli Federico II*

*California Institute of Technology*



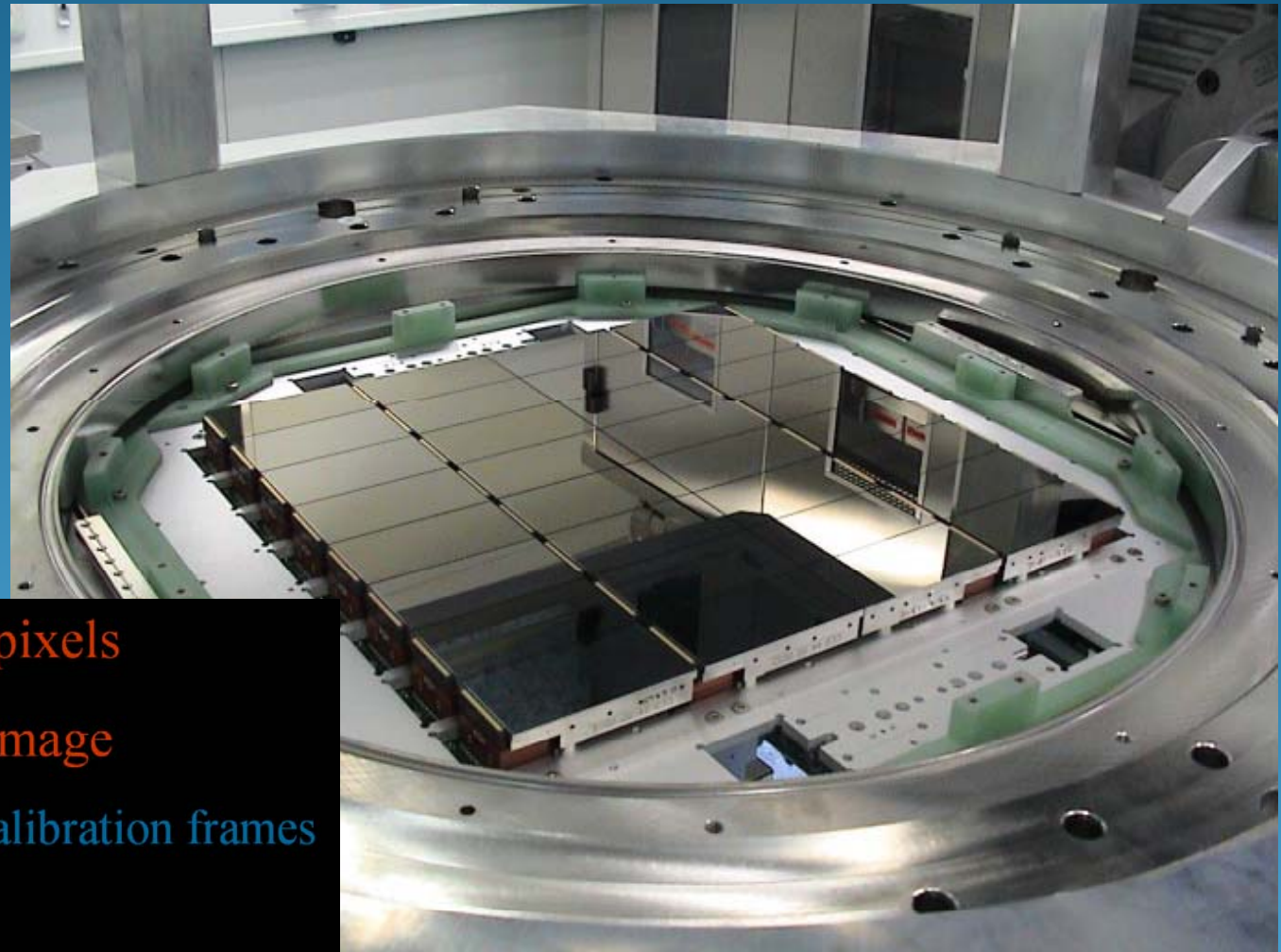
# astrophysical requirements

## *Astronomical data rate*



# astrophysical requirements

## *Astronomical data rate*



268,435,456 pixels

0.5 Gbyte x image

50 science frames + 50 calibration frames

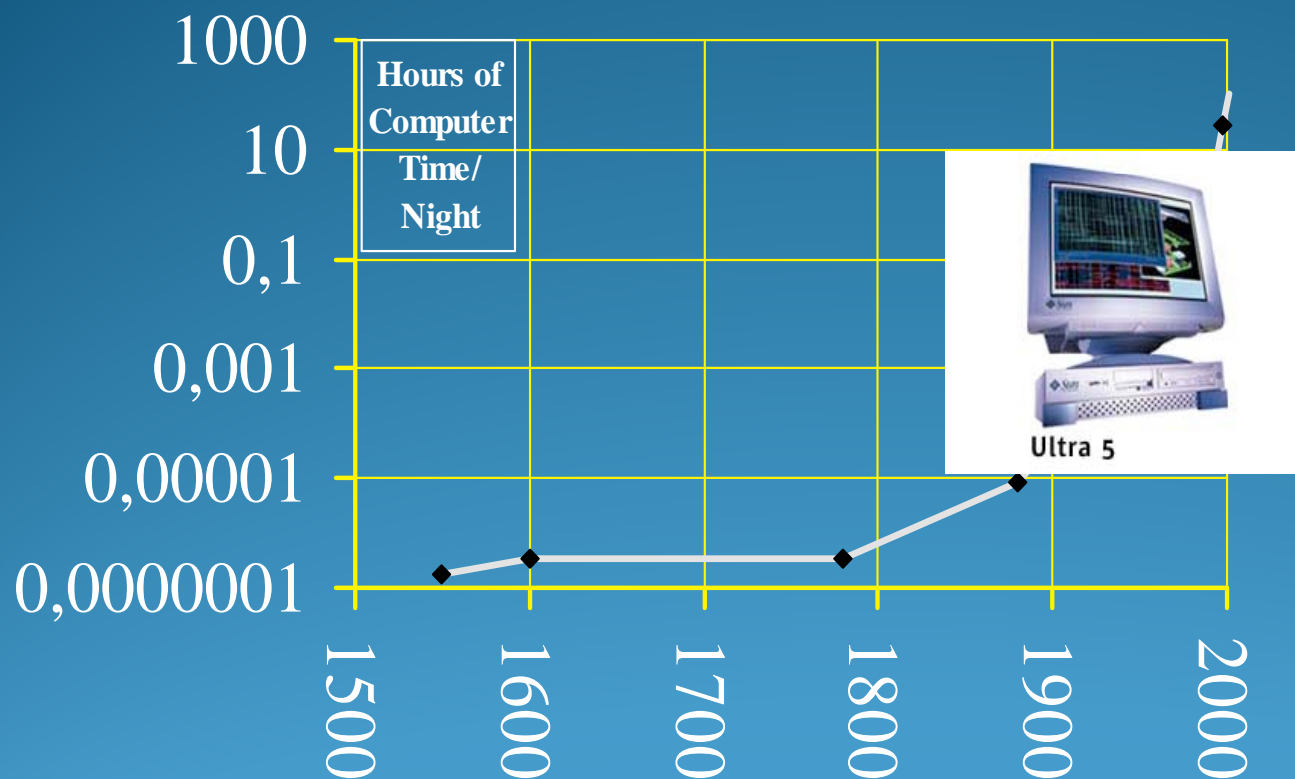


50 Gbyte / Night



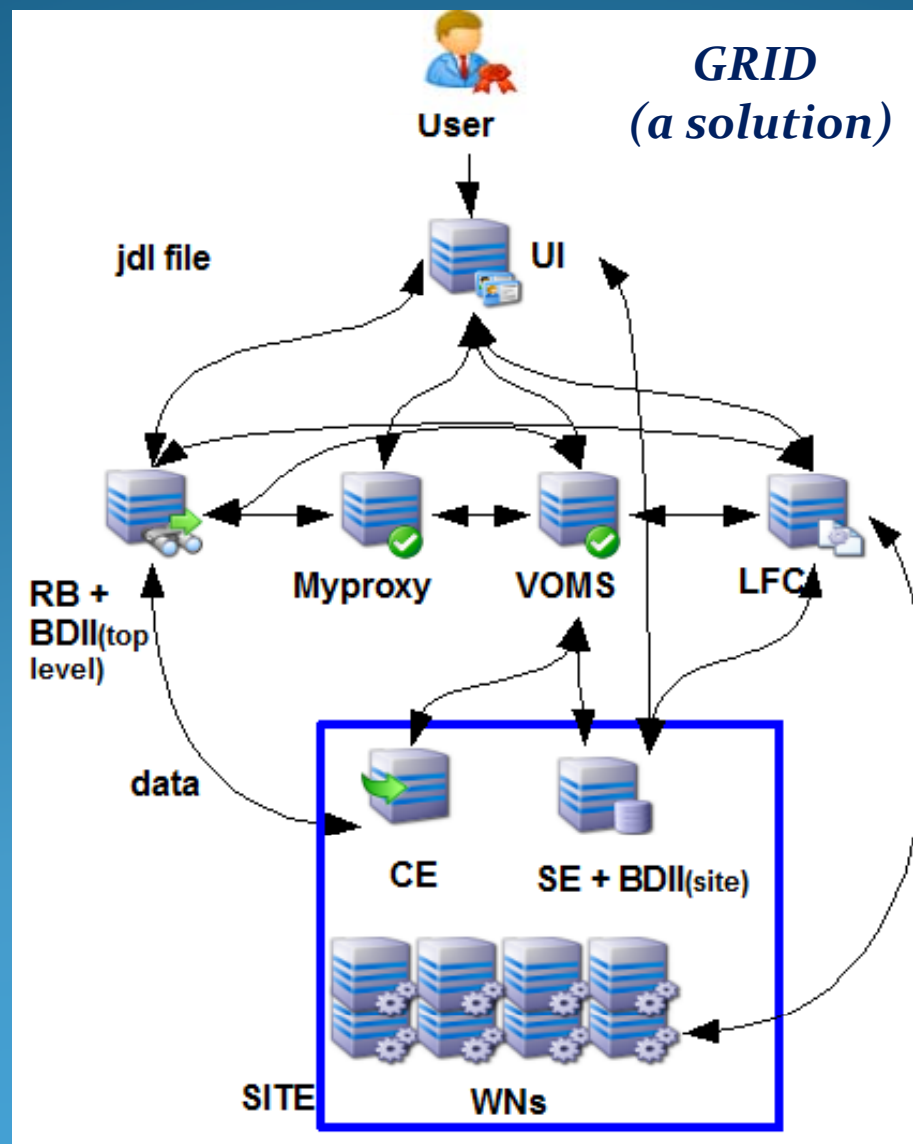
# astrophysical requirements

## *Astronomical computational rate*



# astrophysical requirements

## Astronomical computational rate



## Considerations on the next breakthroughs

- We have reached the physical limit of observations (single photon counting) at almost all wavelength...
- Detectors are linear
- All electromagnetic bands have been opened

### Hence

Our capability to gain new insights on the universe will depend mainly on:

- Capability to recognize patterns or trends in the parameter space (i.e. physical laws) which are not limited to the human 3-D visualization
- Capability to extract patterns from very large multiwavelength, multiepoch, multi-technique parameter spaces

***The answer to these needs is the International Virtual Observatory which (like it or not like it) is bound to be implemented and to change the way astronomers work!***



# VO methodology

**Data Gathering** (e.g., from sensor networks, telescopes...)

→ **Data Farming:**

Storage/Archiving  
Indexing, Searchability  
Data Fusion, interoperability, ontologies, etc.

→ **Data Mining** (or Knowledge Discovery in Databases):

Pattern or correlation search  
Clustering analysis, automated classification  
Outlier / anomaly searches  
Hyperdimensional visualization

→ **Data understanding**

Computer aided understanding  
KDD  
Etc.

→ **New Knowledge**

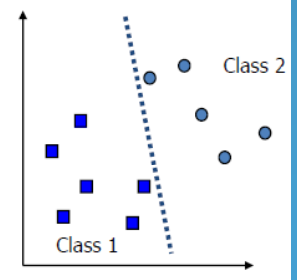
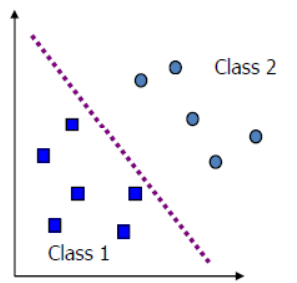
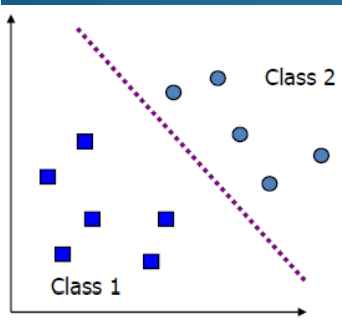
Database technologies

Key mathematical issues

Ongoing research

# Knowledge Discoveries in Databases (KDD) is in practice still unknown to most astronomers

**To implement KDD tools is expensive** (time, computing, need for specialists), requires **coordinated efforts** between astronomers and computer scientists and is aimed to fulfill the needs of **large projects**



## Learning problems as “function approximation”

$\mathbf{X} \equiv \{x_1, x_2, x_3, \dots, x_N\}$  input vectors

$\mathbf{Y} \equiv \{y_1, y_2, y_3, \dots, y_M\}$  target vectors  $M \ll N$

find  $\hat{f}$ :  $\hat{\mathbf{Y}} = \hat{f}(\mathbf{X})$  is a good approximation of  $\mathbf{Y}$

## Exploration on datasets

Dimensional reduction

Classification

Regression

Clustering

Forecasting

Filtering



## Machine learning methods can be broadly grouped in:

### Supervised methods

They learn how to partition the parameter space by means of a training phase based on examples.

Neural Networks such as the Multi Layer Perceptron (MLP), Support Vector Machines (SVM), etc.

### Pro's & Con's

- They are good for interpolation of data, **very bad for extrapolations**
- They **need extensive bases of knowledge** (i.e. uniformly sampling the parameter space) which are difficult to obtain;
- Errors are easy to evaluate
- Relatively easy to use



# Data Mining Exploration with VO-Neural/DAME

## Unsupervised (clustering) methods

They cluster the data relying on their statistical properties only  
Understanding takes place through labeling (very limited BoK).

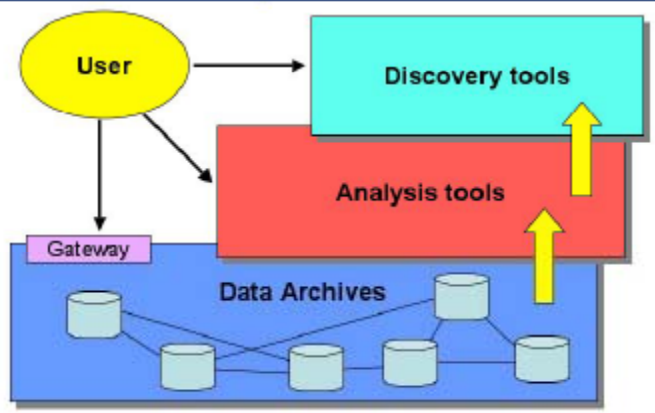
Generative Topographic Mapping (GTM), Self Organizing Maps (SOM), Probabilistic Principal Surfaces (PPS), Support Vector Machines (SVM), etc.

### Pro's & Con's

- In theory they need little or none knowledge a-priori
- Do not reproduce biases present in the BoK
- Evaluation of errors more complex (through complex statistics)
- They are computationally intensive
- They are not user friendly (... more an art than a science; i.e. lot of experience required)

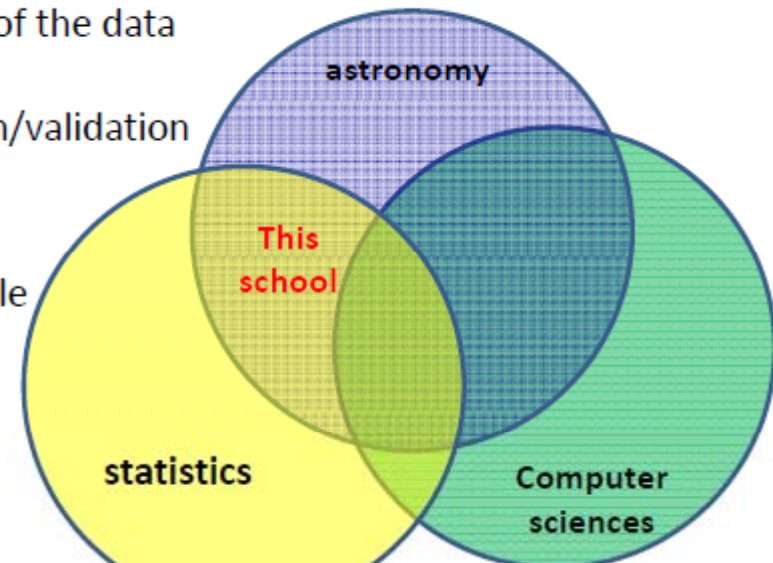


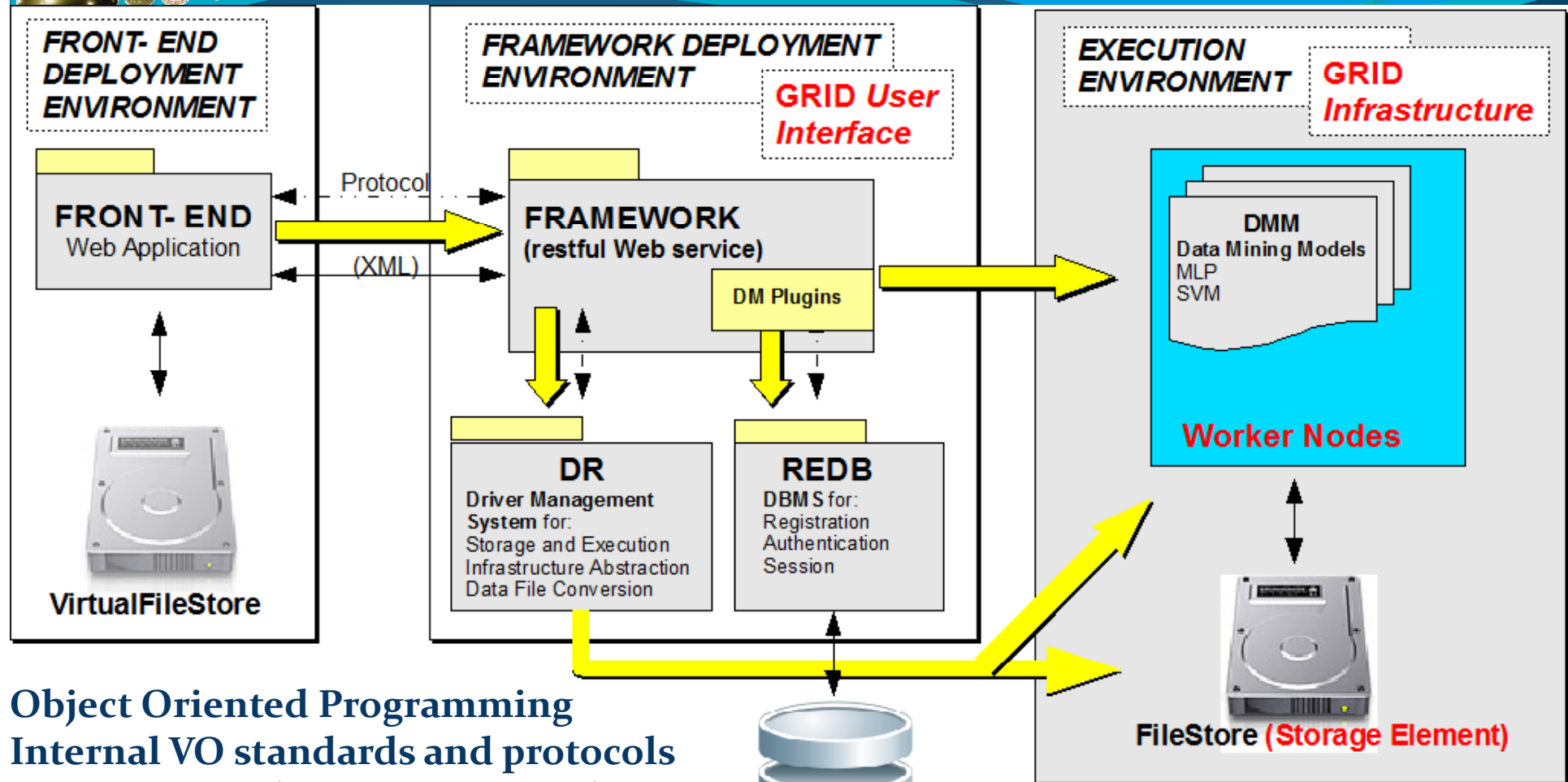
## What we want to do...?



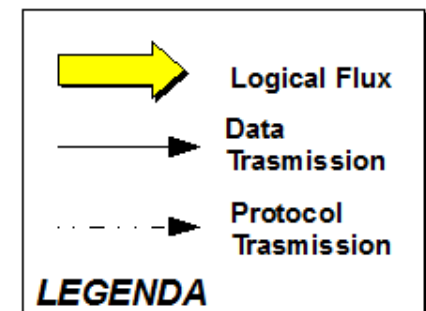
*In 2006, a group of astronomers, computer scientists, engineers and physicians started to explore possible joined effort to create a data mining toolset, based on GRID infrastructure and VO standards, for worldwide users who want to share data, methods and discoveries.*

- **astronomy**: problems, data, understanding of the data structure and biases
- **statistics**: evaluation of the data, falsification/validation of theories/models, etc.
- **computer science**: implementation of infrastructures, databases, middleware, scalable tools, etc.



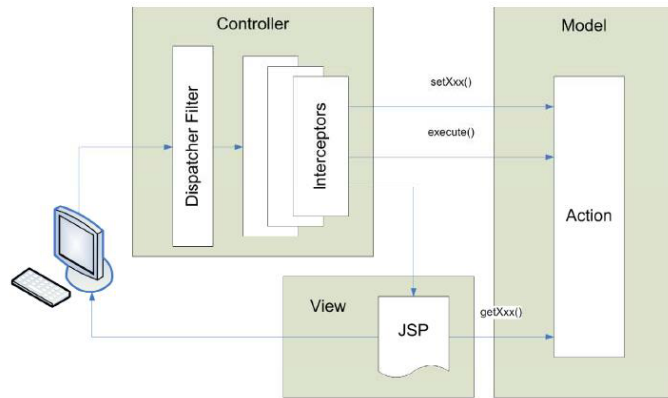


Object Oriented Programming  
 Internal VO standards and protocols  
 Java language (generic for DMM)  
 User/Session Registry DB (MySQL)  
 Web-based User I/O  
 Web Application and Web Service Technology  
 Plugin Modularity (easy to be integrated/modified)  
 Hardware independent through GRID driver  
 Data conversion and manipulation support





## FRONT END Component



## Architecture:

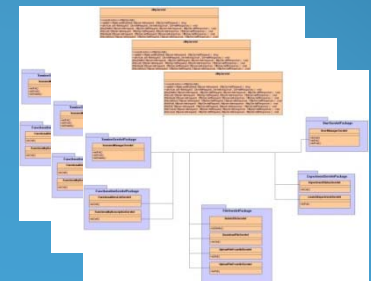
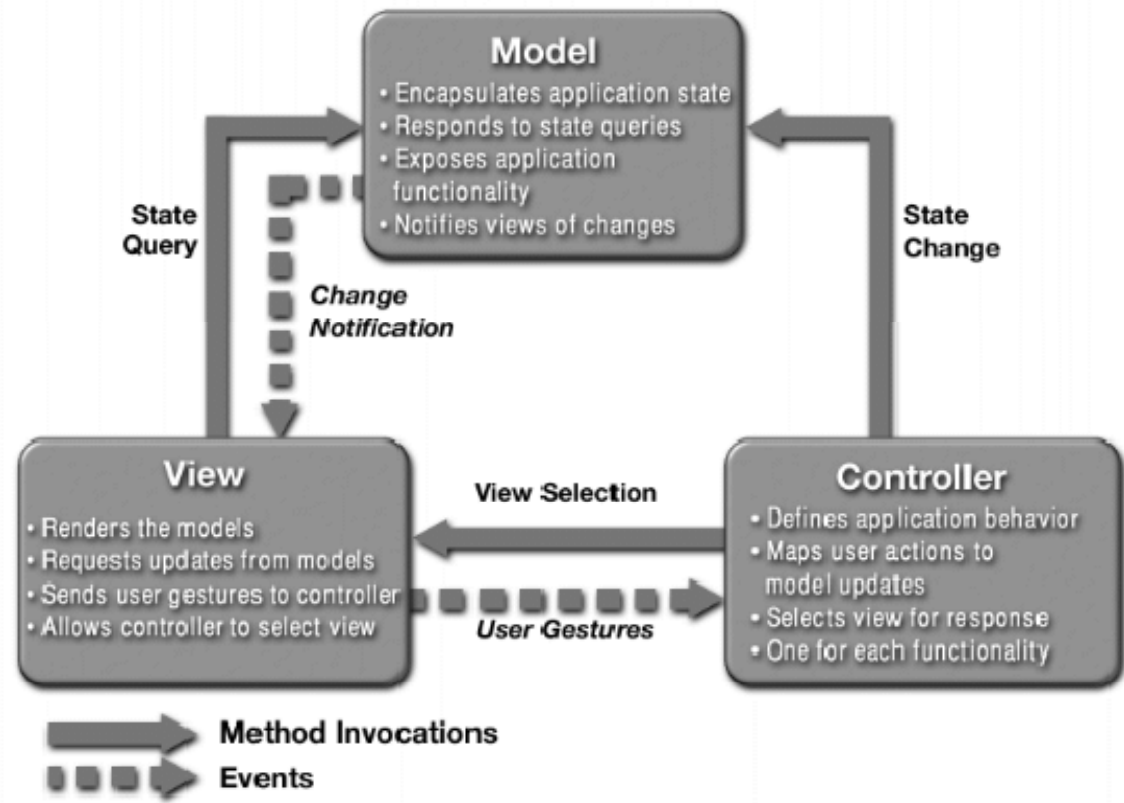
- MVC (Model-View-Controller);

## Technology:

- Struts 2.0 (building infrastructure tool);
- Java Servlet & JSP (dynamic context-dependent web page generation);

## Features:

- User GUI deployment and I/O management;
- interaction with internal components through standard protocol (XML);
- Local User/Session data virtualization through Virtual File Store;



# DR Component

## Architecture:

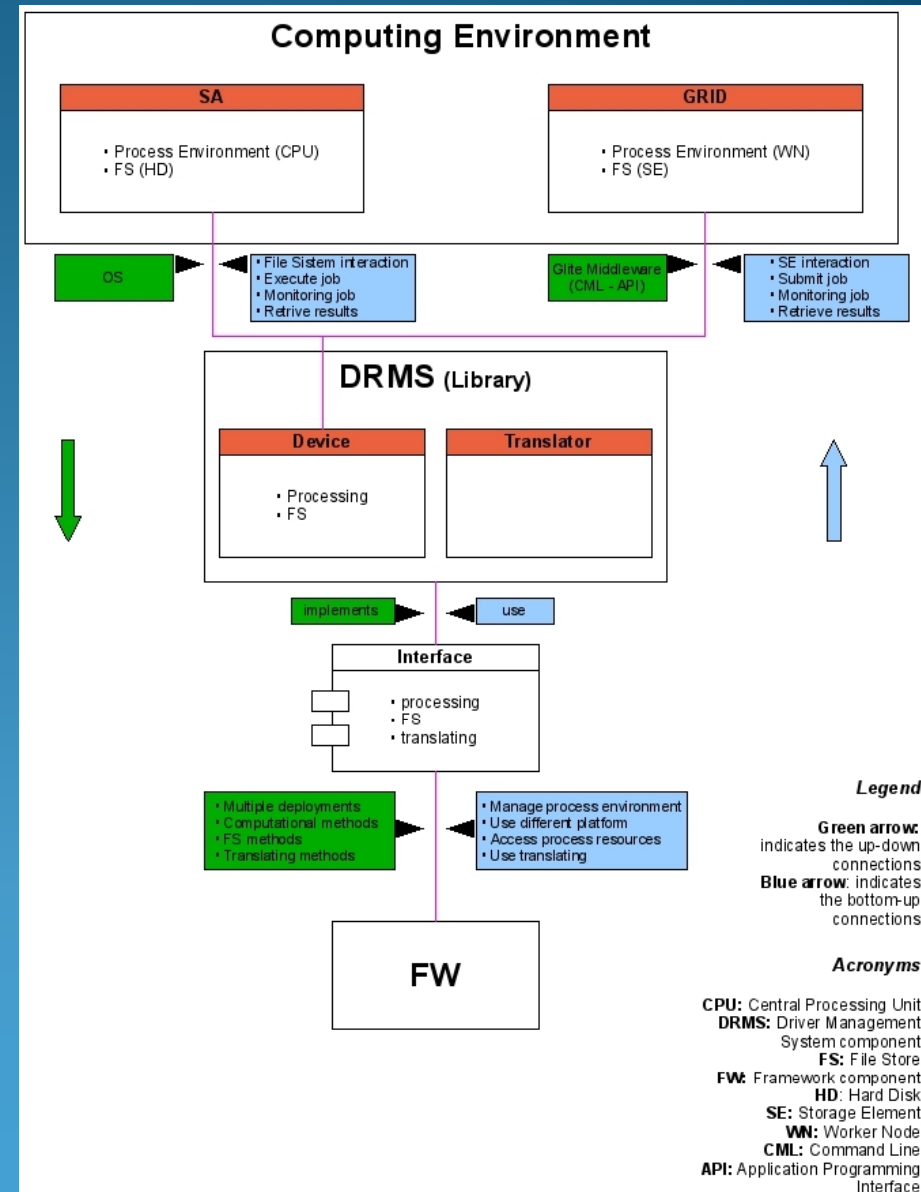
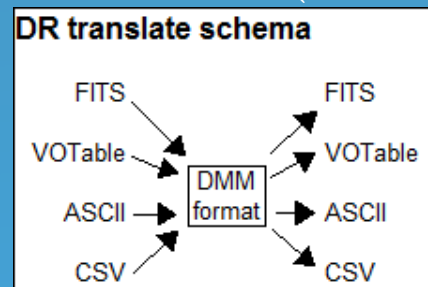
- It depends on the environment choice;
- In S.Co.P.E. DR is a component running on the GRID UI;

## Technology (in S.Co.P.E.):

- GRID Software (middleware gLite);

## Features:

- Storage Device(s) + Execution Environment = Deployment Environment;
- Different Deployment Environments can be more suited for a specific task (e.g. an MLP TEST is unlikely to be a computing intensive task, so GRID latency times are not needed);
- Dynamic Driver Loading => Driver Plugins;
- Drivers are available to the Framework WS and to the Plugins;
- Also used to convert files formats (standard or DMM dependent);



## DMM Component

### Architecture:

- data mining class hierarchy for functionality implementation;

### Technology:

- available model packages and libraries;
- custom ad hoc model design and development;
- custom wrappers for internal standardization;

### Features:

- modularity;
- fast third part application integration;
- functionality specialization;
- multi-language programming support;

**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, ...

# REDB Component

## Architecture:

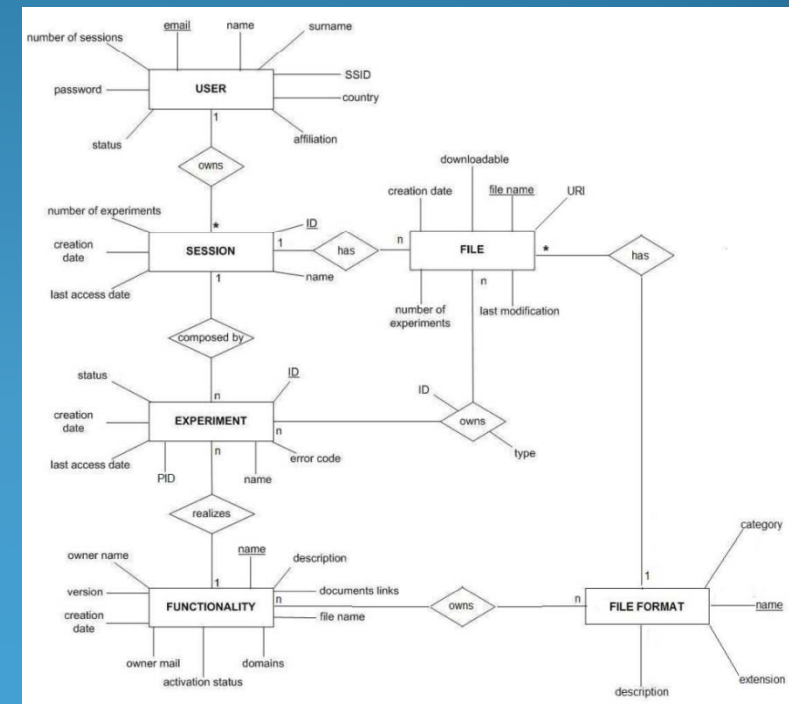
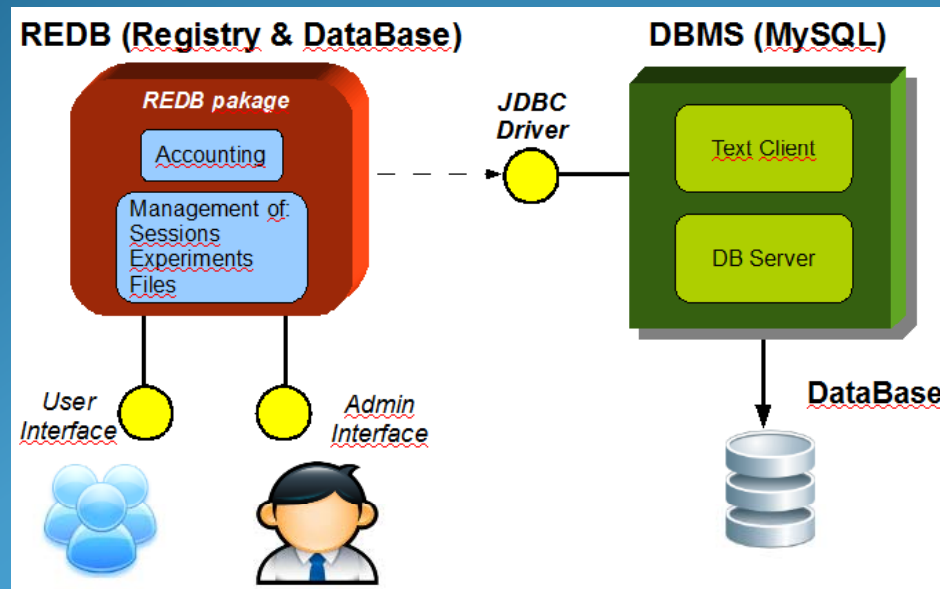
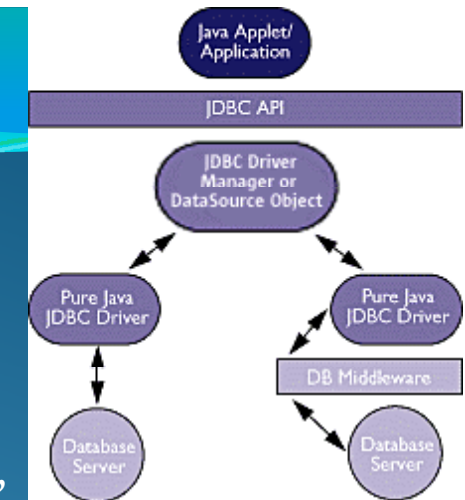
- JDBC;

## Technology (in S.Co.P.E.):

- MySQL and JDBC API;

## Features:

- management of user (registration, authentication, working sessions, experiments and files) information and their relationships;
- store and manage information about three different file's categories: "supported", "exotic" and "custom" (datasets, model configuration and intermediate data);







# FRAMEWORK Component

## Architecture:

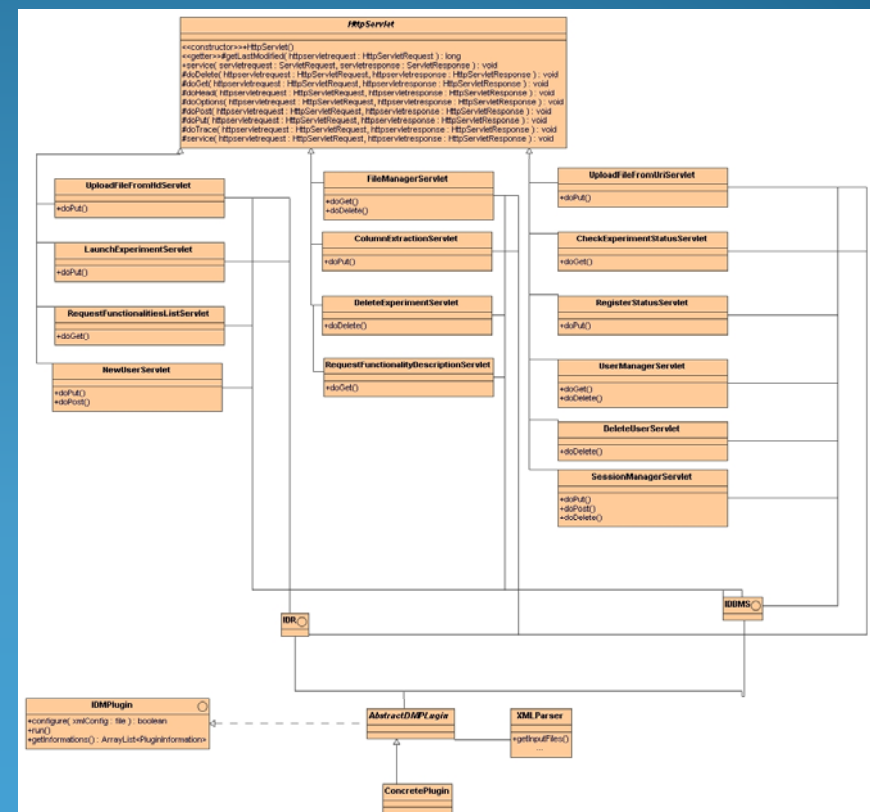
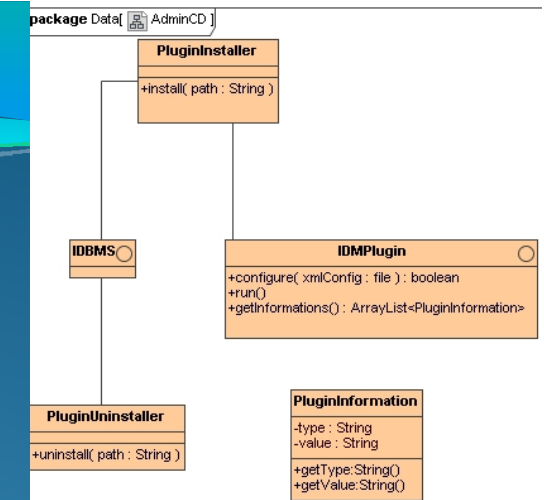
- Restful Web Service (client-server apps with resource addressable with HTTP methods);
- DM models control interface through Plugin SDK;

## Technology:

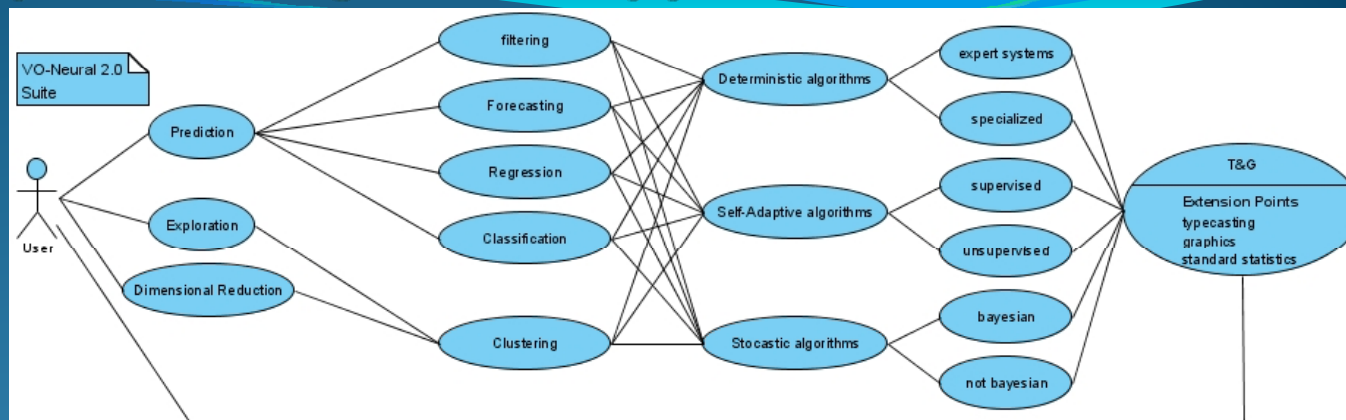
- Web container SUN Apache Tomcat;
- Java Servlet for web service;

## Features:

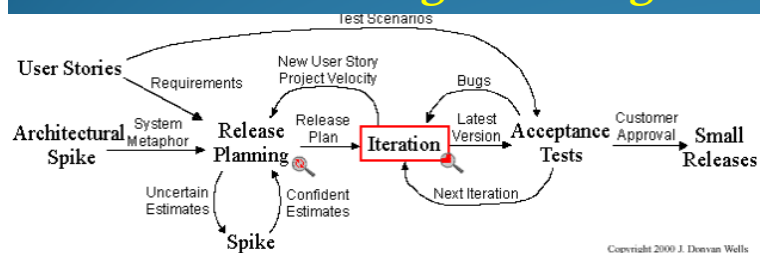
- Internal resource representation through "contextual" VOTables;
- Experiment configuration and execution;
- user authentication and working session management;
- experiment data & working flow trigger and supervision;



## Suite target proposal



## XP – eXtreme Programming



## Design & Documentation process:

1. Statement of work & Project Plan
2. Project Design Description
3. SW Requirement Specifications
4. Software Design Description
5. Implementation
6. Test Procedures
7. Technical Reports
8. Test Reports
9. User & Maintenance Manuals

## VONEURAL/DAME ORGANIZATION CHART



## User/Developer Perspective

A simple user can upload and build his datasets, configure the data mining models available, execute different experiments in service mode, load graphical views of partial/final results.

You are not considering yourself as a simple user? Ok, so you think to be a Developer. Or at least a scientist who wants to upload and use his application (and possibly to share it with others).

Be honest, you don't trust someone else's application.

So You want to extend our framework?

**YES, WE CAN!**

### DM Models Development

Download our DM Models library;

Add new low level/DM shared libraries and related new wrapper;

Extend the DM class hierarchy;

### Model/Driver Plugin Development

Download our SDK;

Implement and test the DMPlugin abstract class;

Provide a method to produce the plugin description and Submit for Registration;

The same if you want to develop a new driver for a specific environment or storage system. Just implement the Driver Plugin Interface and register it;



## Application Prototype

DAME help &amp; tutorials - HowTo - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti ?

[http://pcdevauc.na.infn.it:9000/help/#registration](#)

Più visitati Come iniziare Ultime notizie

Y! Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers Sports Sign In

sitemap page VO-Neural Home Page DAME help &amp; tutorials - HowTo https://imap-ac.n.../src/webmail.php

**DAME – Data Mining and Exploration**  
California Institute of Technology - Università degli Studi Federico II

Username

brescia

Password

\*\*\*\*\*

Log in

Home

Sign Up!

Help &amp; Tutorials

The Team

Links

Ministero degli  
Affari Esteri

## How to use this website/web application

- 1. [Registration](#)
- 2. [Login](#)
- 3. [Filestore](#)
- 4. [Experiments](#)

## 1. Registration

In order to use DAME, you have to request an account. You can freely sign up for a new account and a confirmation message will be sent to your email address.

Signing up you will have access to the tools provided by the webapplication, that is to say:

- [MyFilestore](#)
- [MyExperiments](#)

You can refer to the specific sections for details about this tools, which will allow you to do the following operations:

- Upload files to our servers. This way you will have a persistent repository with all your datasets and you won't need to upload a file each time you need to launch an Experiment.
- View all the output files produced by the experiments you launch.
- Download the results of the experiments to your local hard drive.

To sign up go to the [registration page](#) and fill in the form (all the fields are required).

Choose a password which is not too easy to guess, but we recommend not to set it to a password you use to ssh to other servers.

Once you've successfully submitted the registration form, a confirmation link will be sent to your email address. Click on the link in order to activate your account and start using DAME.

Completato

catania\_febbraio2009 DAME help &amp; tutori... 2009\_tiruvalla\_India... voneural-dame

IT 14.09





## Application Prototype

Dame - DAta Mining and Exploration - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti ?

http://pcdevauc.na.infn.it:9000/

Più visitati Come iniziare Ultime notizie

Y! Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers Sports Sign In

sitemap page VO-Neural Home Page Dame - DAta Mining and Explorat... https://imap-ac.n.../src/webmail.php

**Massimo Brescia**  
Last Login  
Sun 08 Feb 2009  
01:01PM GMT

Home

MyFilestore

MyExperiments

Logout

Help & Tutorials

The Team


Links

Launch Experiments

New MLP

New SVM

New PhotoZ

  
Ministero degli Affari Esteri

**My Experiments**

Experiments List

Name	Science case	Mode	Status	Actions
------	--------------	------	--------	---------

**My Filestore**

Sfoglia...

Click here to upload the file

Dirs	Files	Actions
/brescia		
No data in this directory!		
/brescia/Samples		Delete Download
	iris.dat	Delete
	provapiccolo.csv	Delete

Completato



catania\_febbraio2009

Dame - DAta Minin...

2009\_tiruvalla\_India....

voneural-dame

IT

&lt;

a

a

a

a

14:10



# Application Prototype

Dame - DAta Mining and Exploration - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti ?

http://pcdevauc.na.infn.it:9000/filestore/download/brescia/Samples/provapiccolo.csv/

Più visitati Come iniziare Ultime notizie

Y! Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers Sports Sign In

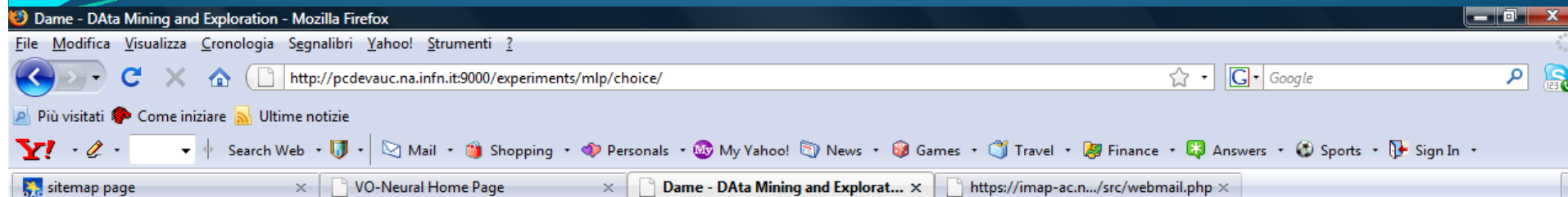
sitemap page VO-Neural Home Page Caricamento in corso... https://imap-ac.n.../src/webmail.php

```
1.692768,0.989927,0.401751,0.315475,0.113712
1.705183,0.972942,0.462513,0.299402,0.104813
1.81883,0.832718,0.387196,0.282825,0.076243
1.754013,0.897587,0.411005,0.295853,0.083096
1.750496,0.987221,0.452448,0.28212,0.135608
2.016077,0.979092,0.447763,0.334382,0.105586
1.725897,1.089231,0.448336,0.378626,0.124194
1.946756,0.970669,0.391537,0.341539,0.105292
1.886108,0.828356,0.398804,0.277565,0.116535
1.963732,0.862139,0.435151,0.337962,0.049855
1.934912,0.979206,0.440548,0.316557,0.114727
1.664152,1.160151,0.517653,0.369753,0.190644
1.809011,1.250761,0.46133,0.358436,0.193634
1.668617,0.915407,0.381332,0.283329,0.143155
1.6555,1.079199,0.450472,0.372208,0.146099
2.03433,0.889687,0.475706,0.338007,0.02892
1.976236,1.048838,0.512547,0.389719,0.12446
2.004997,0.926264,0.402496,0.328529,0.092687
1.87063,0.944035,0.406311,0.32659,0.117726
1.85619,0.996578,0.398401,0.324921,0.117244
2.042259,0.889086,0.421186,0.305461,0.075189
1.971815,0.834901,0.380198,0.290215,0.029776
1.991796,0.981928,0.44222,0.350865,0.103254
1.818487,1.129131,0.449316,0.311396,0.146972
1.797421,0.867332,0.405901,0.326082,0.068398
1.838831,0.746513,0.384384,0.292418,0.068517
1.71731,1.013191,0.414141,0.355978,0.136164
1.750128,0.869583,0.403996,0.245794,0.082915
1.913464,1.039413,0.480566,0.416491,0.118118
2.044985,1.069139,0.447777,0.330549,0.148072
1.825377,0.896421,0.367558,0.299021,0.092071
1.959181,0.947929,0.4189,0.325193,0.09667
1.887531,1.15271,0.464077,0.334187,0.166015
1.66408,0.889212,0.481047,0.372786,0.110893
2.00355,0.890563,0.418589,0.293733,0.062144
1.677416,1.177158,0.472656,0.304817,0.209245
```

Trasferimento dati da pcdevauc.na.infn.it...

catania\_febbraio2009 Dame - DAta Minin... 2009\_tiruvalla\_India... voneural-dame

IT 14.07



**DAME – Data Mining and Exploration**  
California Institute of Technology - Università degli Studi Federico II



**Massimo Brescia**  
Last Login  
Sun 08 Feb 2009  
01:01PM GMT

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

Launch Experiments

- New MLP
- New SVM
- New PhotoZ

## Experiment Configuration

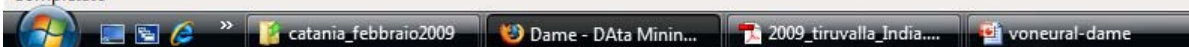
Science case: you can choose whether to classify labeled patterns or to find a regression mapping from examples. You can find documentation [here](#).

Science Case:

Mode: choose the mode you want to run. If you don't know what these modes mean and how they work, refer to [this link](#).

Mode:

Completato



IT < 14.11



## Application Prototype

Dame - DAta Mining and Exploration - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti ?

[http://pcdevauc.na.infn.it:9000/experiments/submission/True/](#)

Più visitati Come iniziare Ultime notizie

Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers Sports Sign In

sitemap page VO-Neural Home Page Dame - DAta Mining and Explorat... https://imap-ac.n.../src/webmail.php

Massimo Brescia

Last Login  
Sun 08 Feb 2009  
01:01PM GMT

Home

MyFilestore

MyExperiments

Logout

Help &amp; Tutorials

The Team

Links

Launch Experiments

New MLP

New SVM

New PhotoZ

Ministero degli  
Affari Esteri

## Experiment Configuration

Name. This is the name that will be associated with the experiment. Be sure the name is meaningful to you. When the experiment is done, you will find your files in a directory in your filestore named after your experiment.

Experiment name: 

Input Nodes. It is the number of input features. If  $N$  is the number of input features and  $M$  the number of target components, then the training set must have exactly  $N+M$  columns.

Input nodes: 

Hidden Nodes Help

Hidden nodes: 

Output Nodes Help

Output nodes: Max epochs: Tolerance: Training algorithm: Resume training: Network: 

CE - BATCH

Training set: Do validation: ☒Validation set: 

Go!

Completato

catania\_febbraio2009 Dame - DAta Minin... 2009\_tiruvalla\_India... voneural-dame

IT 14:13





## Application Prototype

MLP Submission - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti ?

[http://pcdevauc.na.infn.it:9000/experiments/mlptrain/65/](#)

Più visitati Come iniziare Ultime notizie

Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers Sports Sign In

sitemap page VO-Neural Home Page MLP Submission https://imap-ac.n.../src/webmail.php

Last Login  
Sun 08 Feb 2009  
01:01PM GMT

Home

MyFilestore

MyExperiments

Logout

Help &amp; Tutorials

The Team

Links

Launch Experiments

New MLP

New SVM

New PhotoZ

Ministero degli  
Affari Esteri

## Experiment Details

Experiment Name: myExperiment

Finished

Parameter	Value
Input Nodes	4
Hidden Nodes	8
Output Nodes	1
Max Epochs	1000
Tolerance	1e-05
Training Algorithm	mseIncremental
Training Set	/brescia/Samples/provaviccolo.csv
Validation Set	/brescia/Samples/provaviccolo.csv

Dirs	Files	Actions
/brescia/myExperiment		<a href="#">Download</a>
	<a href="#">provaviccolo.csv.fits</a>	<a href="#">Delete</a>
	<a href="#">myExperiment.csv</a>	<a href="#">Delete</a>
	<a href="#">myExperiment.log</a>	<a href="#">Delete</a>
	<a href="#">myExperiment.ERROR</a>	<a href="#">Delete</a>

## Experiment Log

```
Maximum epochs: 1000
Problem case: Classification
Training algorithm: Incremental
Error: MSE
Error tolerance: 1e-05
Input network name: empty
Training dataset: myExperiment/provaviccolo.csv.fits
Validation dataset: myExperiment/provaviccolo.csv.fits
Testing dataset: myExperiment/empty.fits
```

Completato

catania\_febbraio2009 MLP Submission - ... 2009\_tiruvalla\_India... voneural-dame

IT &lt; 14.17



## Our first scientific use cases

### First example

evaluation of SDSS redshift using supervised NN (MLP)

*Mining the SDSS Archive I. Photometric redshifts in the nearby Universe, R. D'Abrusco et al. (The Astrophysical Journal, 663: 752-764, 2007 July 10.*

### Second example

Searching for candidate quasars in the SDSS archive

astro-ph/0805.0156v1; to appear soon in MNRAS (R. D'Abrusco et al.)

### Third example

Classifying AGN in SDSS with SVM

Cavuoti 2008, Thesis (VONeural website, [voneural.na.infn.it](http://voneural.na.infn.it))

## In this Conference poster session:

*A web application for photometric redshifts evaluation*

Omar Laurino et al.