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

G. Longo – M. Brescia

&

Project Team

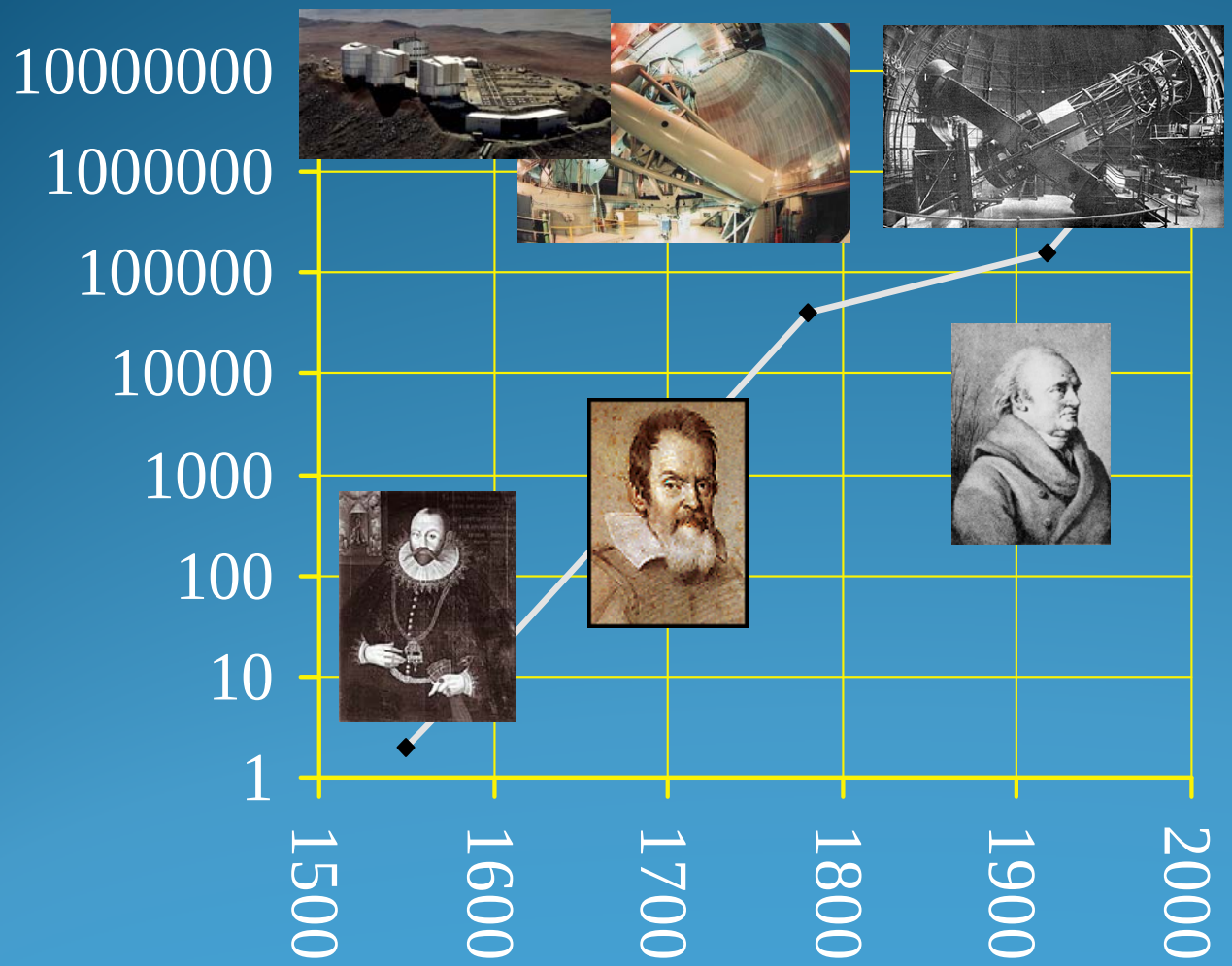
INAF – Osservatorio Astronomico di Capodimonte

Dipartimento di Fisica – Università degli Studi di Napoli Federico II

California Institute of Technology

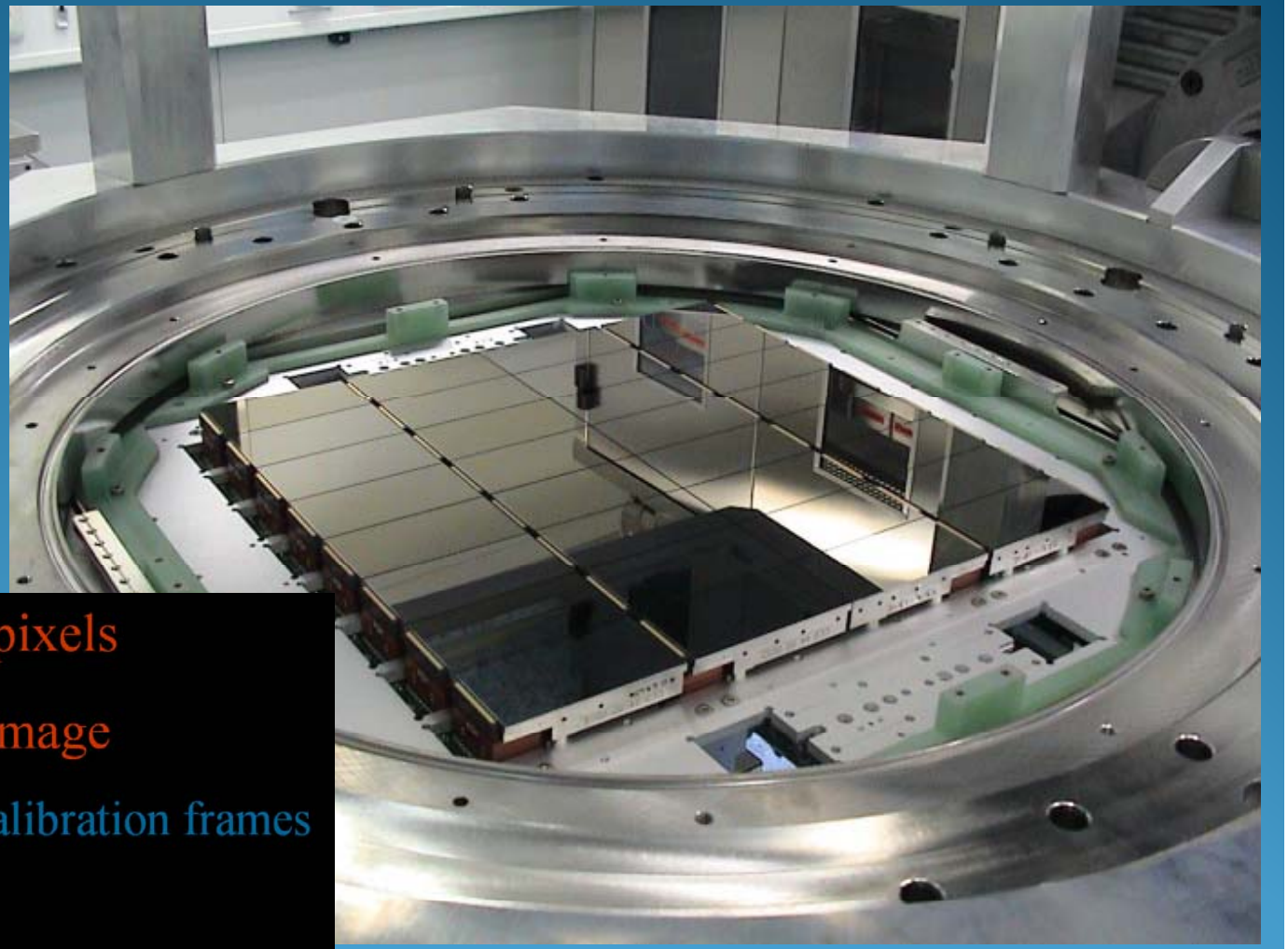
astrophysical requirements

Astronomical data rate



astrophysical requirements

Astronomical data rate



268,435,456 pixels

0.5 Gbyte x image

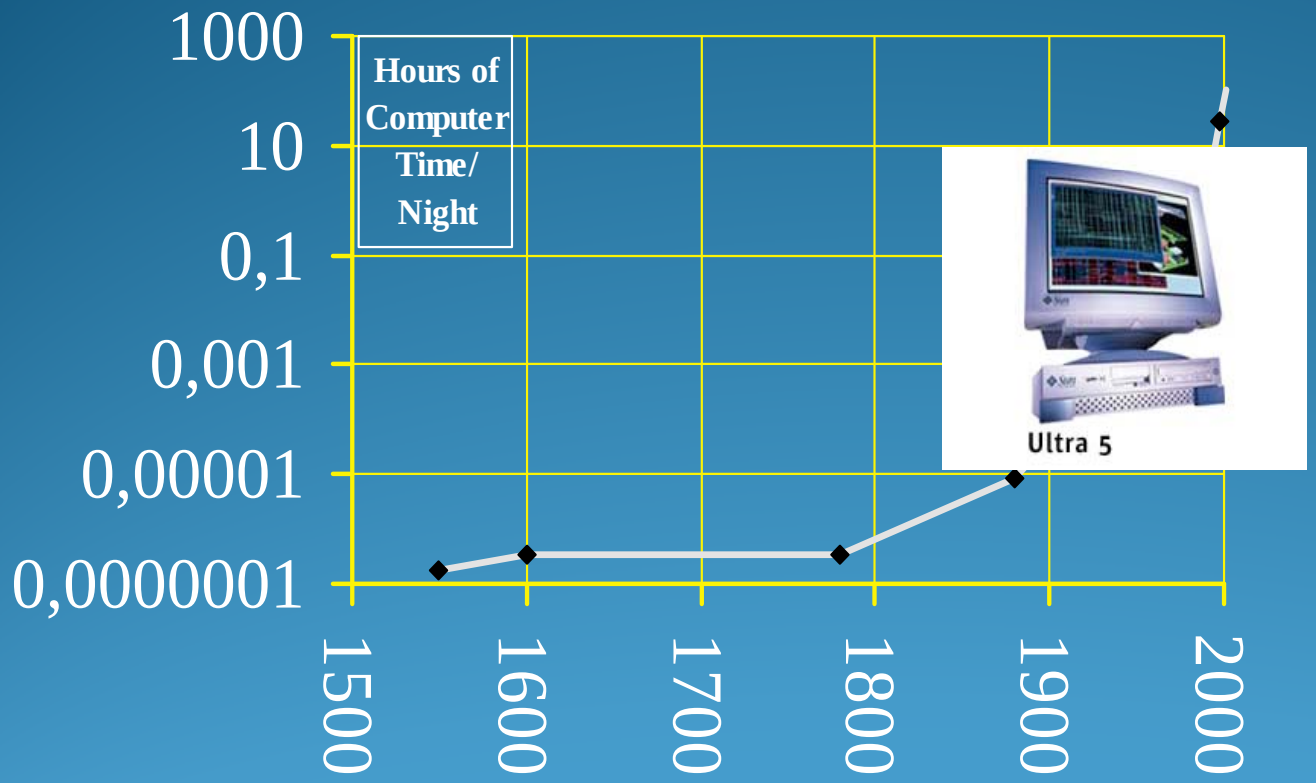
50 science frames + 50 calibration frames

⇩

50 Gbyte / Night

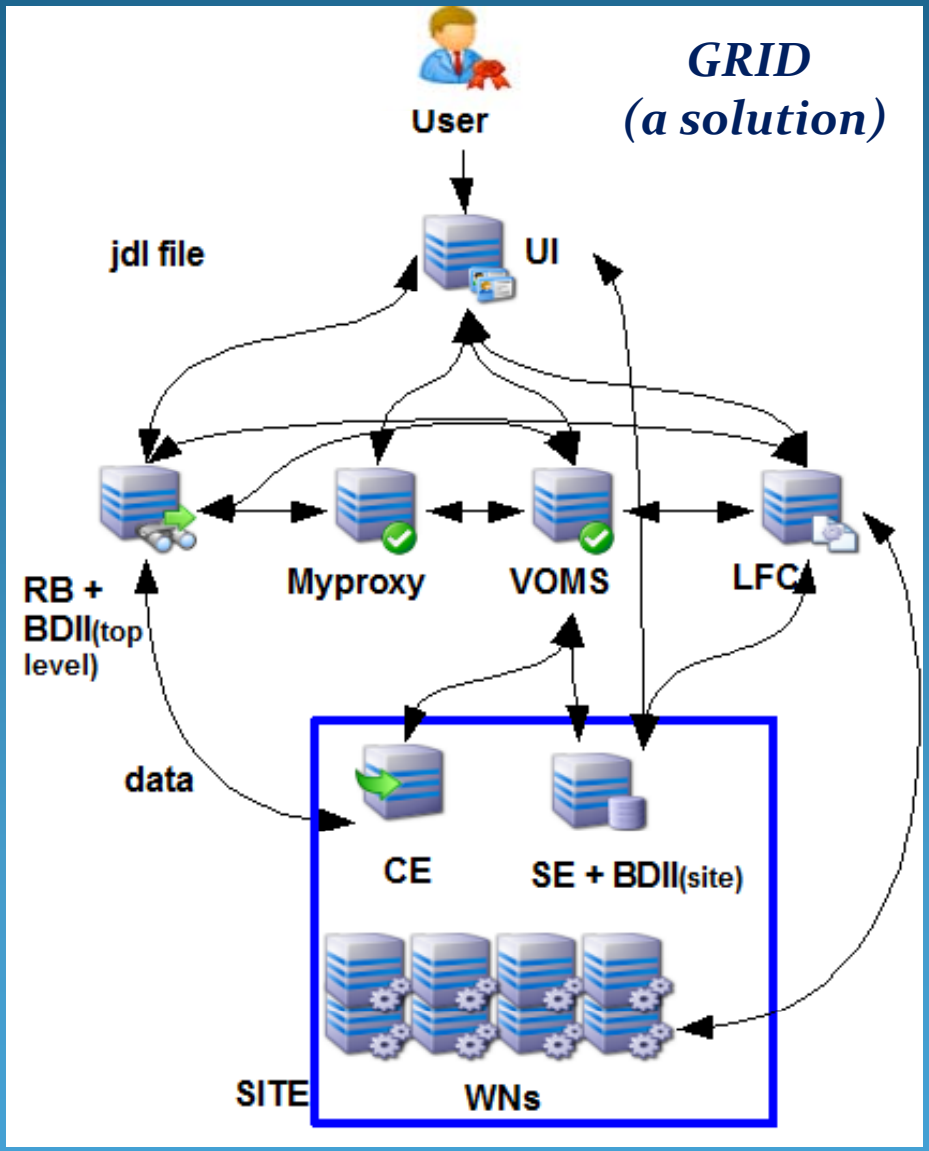
astrophysical requirements

Astronomical computational rate



astrophysical requirements

Astronomical computational rate



astrophysical requirements

Considerations on the next breakthroughs

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

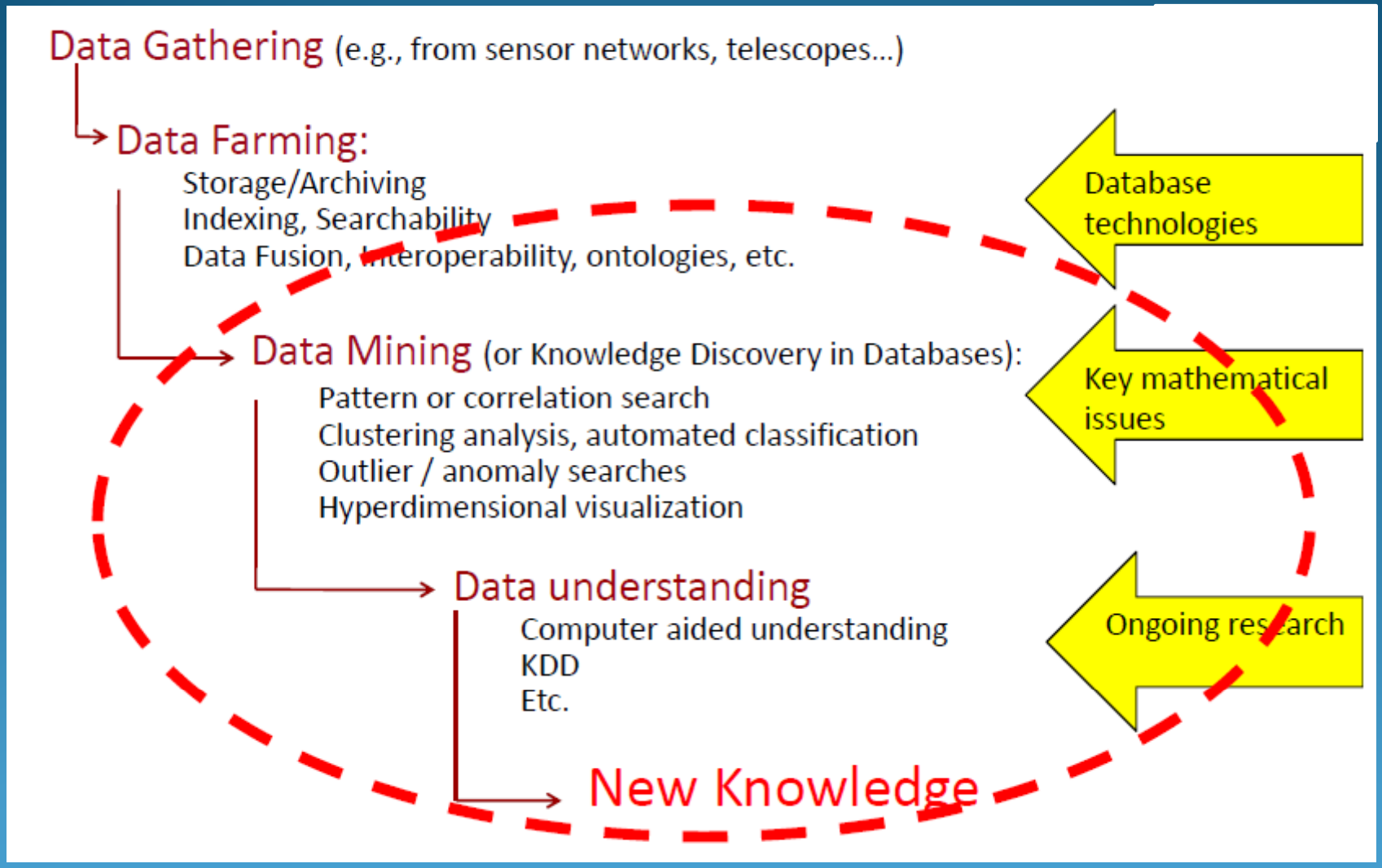
Hence

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 multiwavelenght, multiepoch, multi-technique parameter spaces

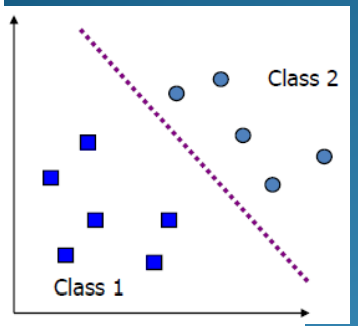
The answer to these needs is the International Virtual Observatory which (like it or not like it) is bound to be implemented and to change the way astronomers work!

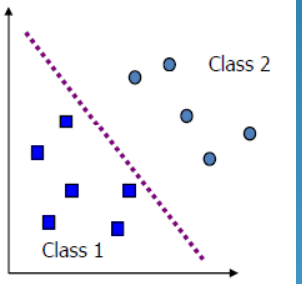
VO methodology

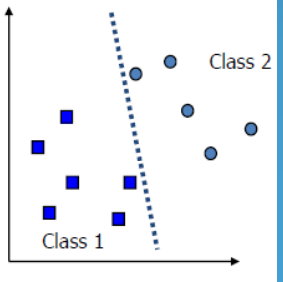


Knowledge Discoveries in Databases (KDD) is in practice still unknown to most astronomers

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"

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

$\mathbf{Y} \equiv \{x_1, x_2, x_3, \ldots x_M\}$ **target vectors** $M << N$

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

Exploration on datasets

Dimensional reduction

Classification

Regression

Clustering

Forecasting

Filtering

Data Mining Exploration with VO-Neural/DAME

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. uniformously 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).

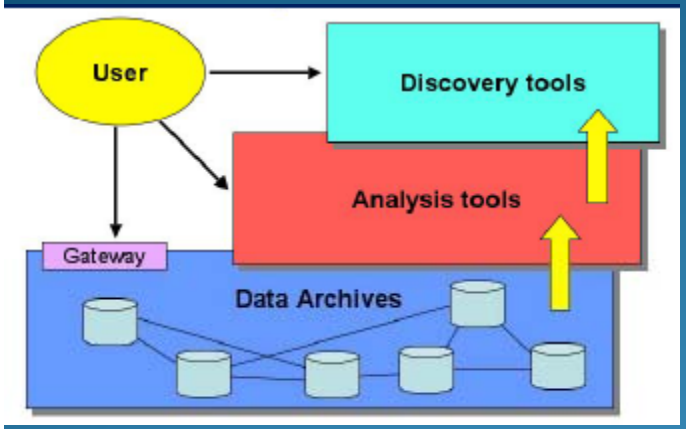
Generative Topographic Mapping (GTM), Self Organizing Maps (SOM), Probabilistic Principal Surfaces (PPS), Support Vector Machines (SVM), etc.

Pro's & Con's

- In theory they need little or none knowledge a-priori
- Do not reproduce biases present in the BoK

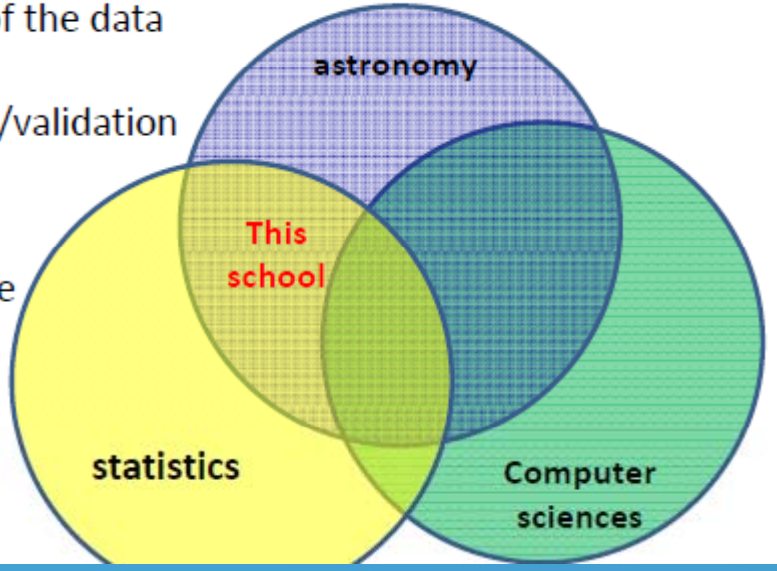
- Evaluation of errors more complex (through complex statistics)
- They are computationally intensive
- They are not user friendly (... more an art than a science; i.e. lot of experience required)

What we want to do…?

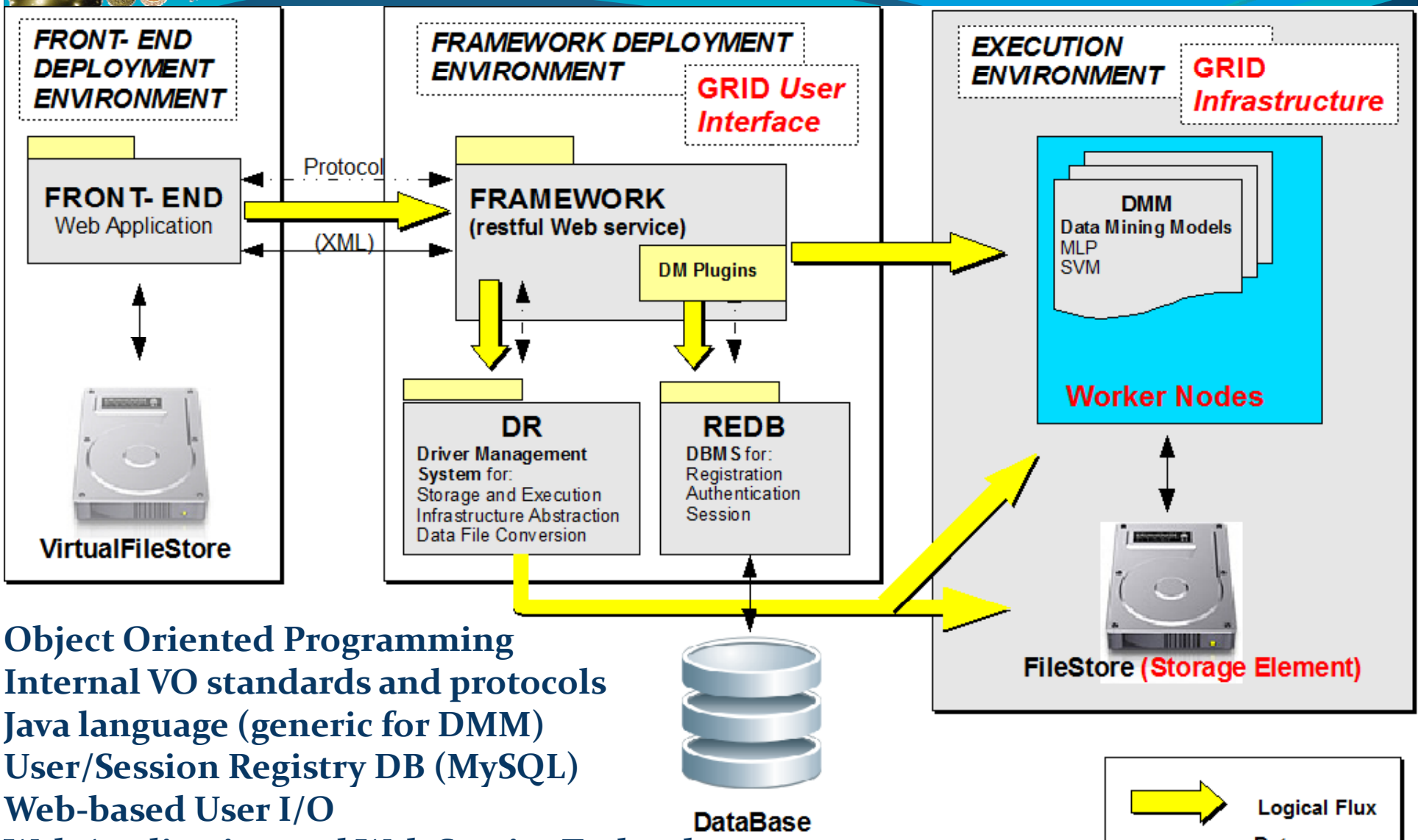


In 2006, a group of astronomers, computer scientists, engineers and physicians started to explore possible joined effort to create a data mining toolset, based on GRID infrastructure and VO standards, for worldwide users who want to share data, methods and discoveries.

- **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.

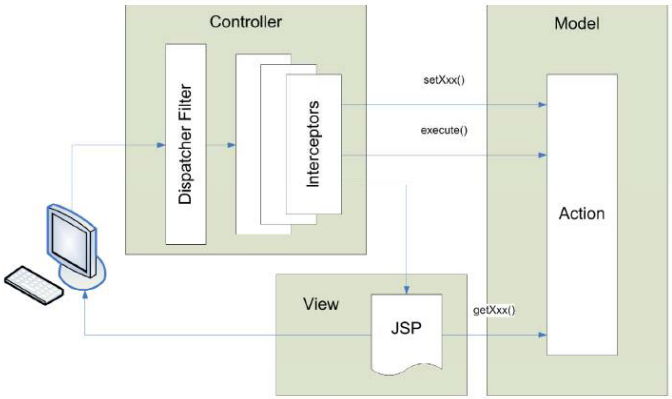


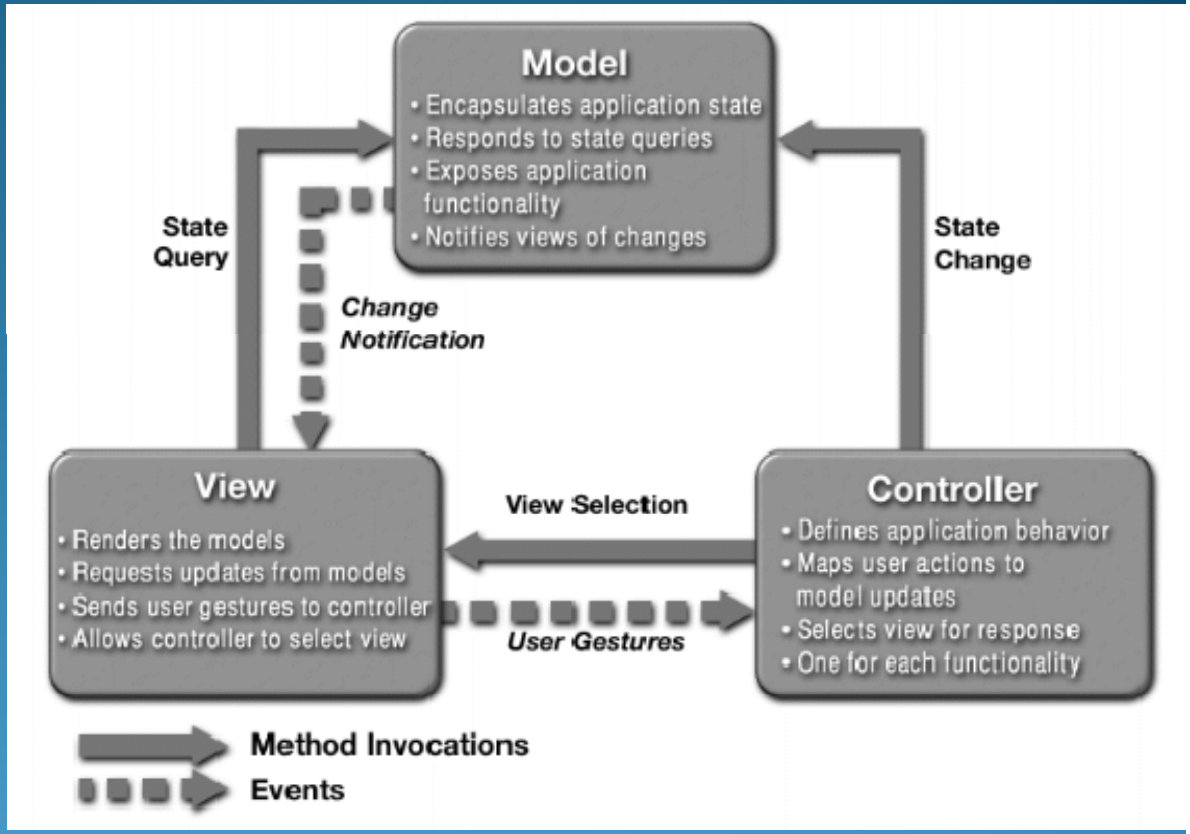
How to do this...?



- 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:

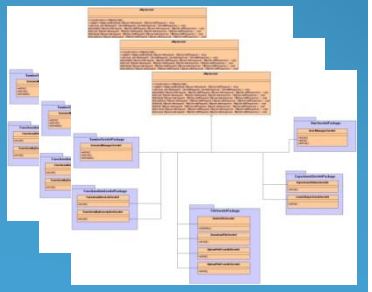
- MVC (Model-View-Controller);

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:

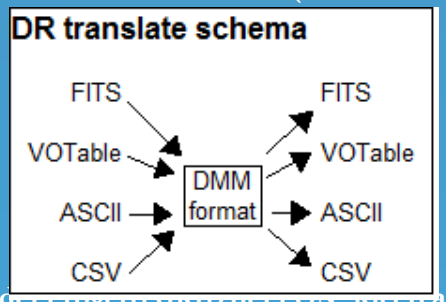
- It depends on the environment choice;
- In S.Co.P.E. DR is a component running on the GRID UI;

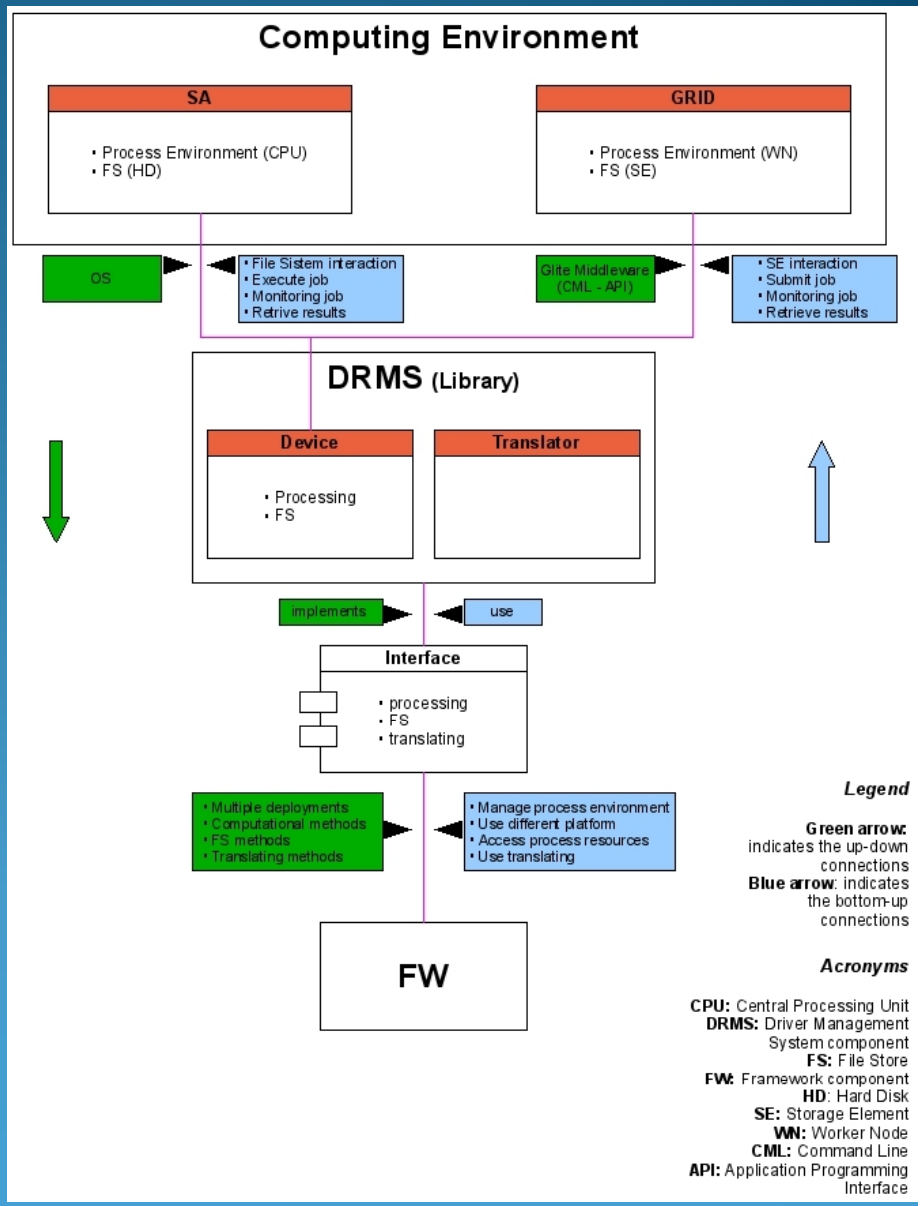
Technology (in S.Co.P.E.):

- 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);





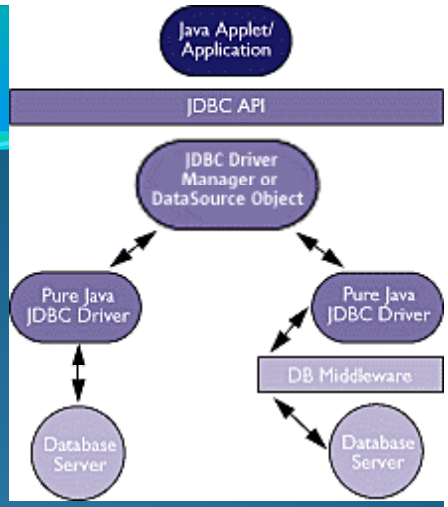
DMM Component

Architecture:
- data mining class hierachy 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;

REDB Component



Architecture:

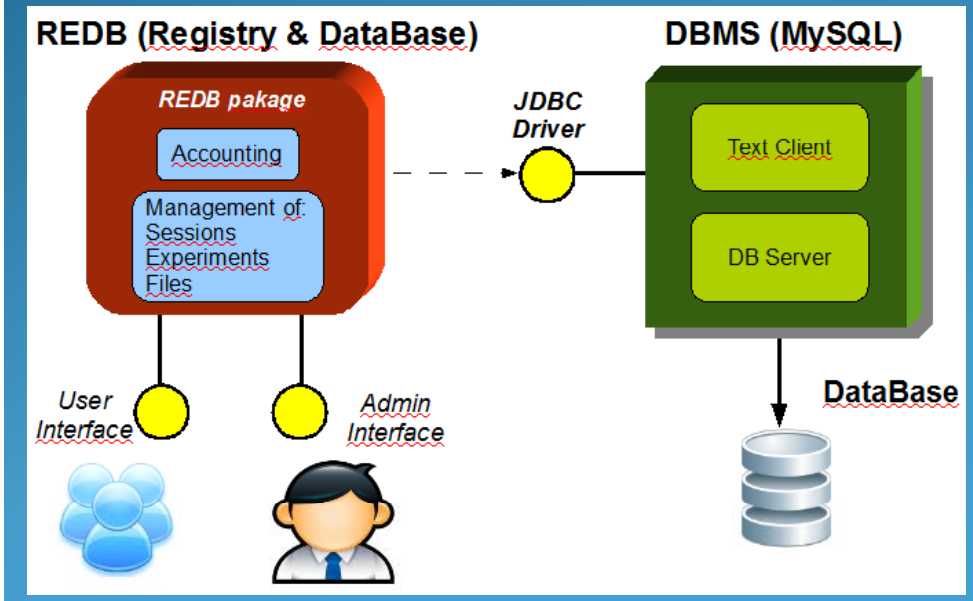
- JDBC;

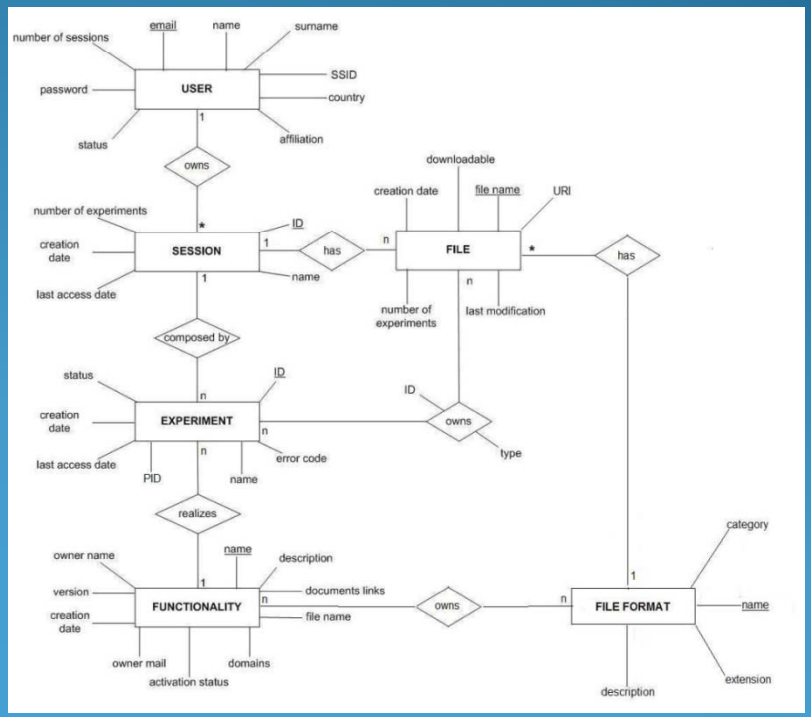
Technology (in S.Co.P.E.):

- MySQL and JDBC API;

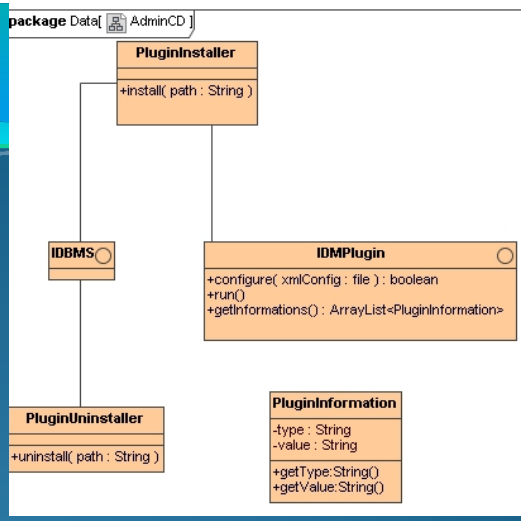
Features:

- 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

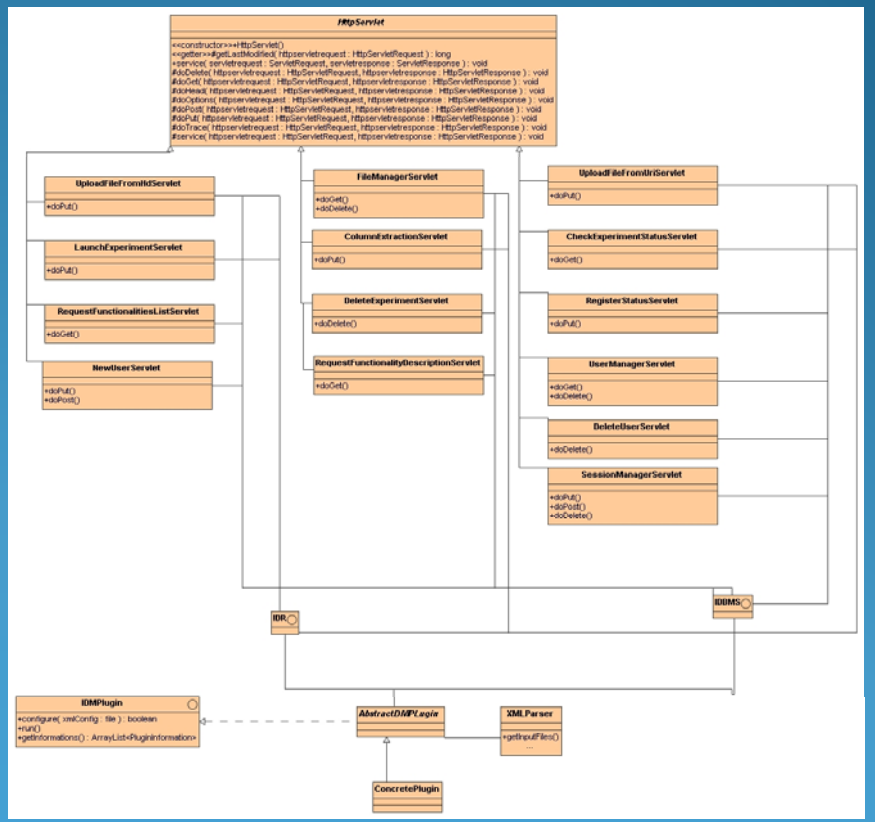


Architecture:

- 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;

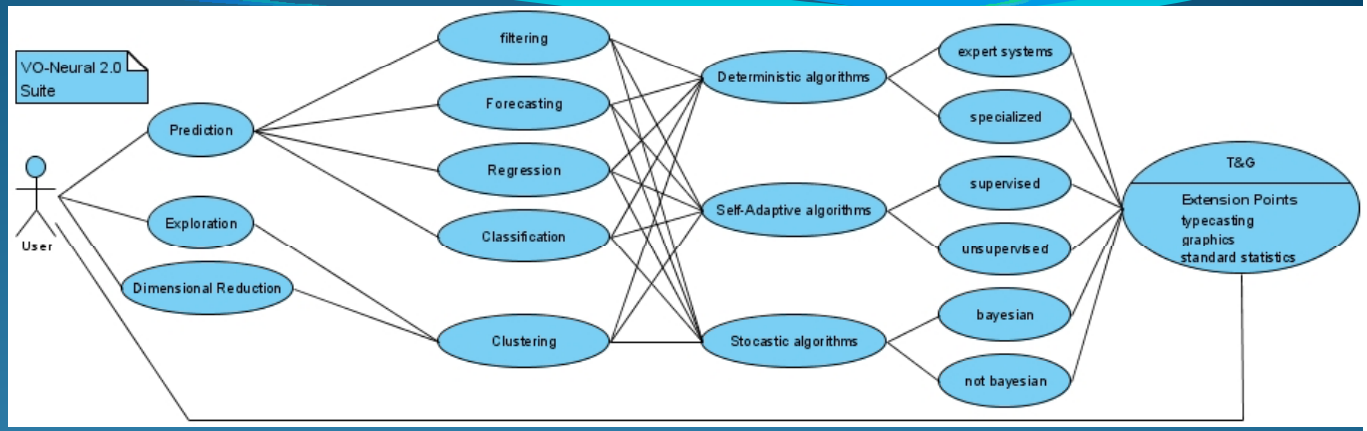


Features:

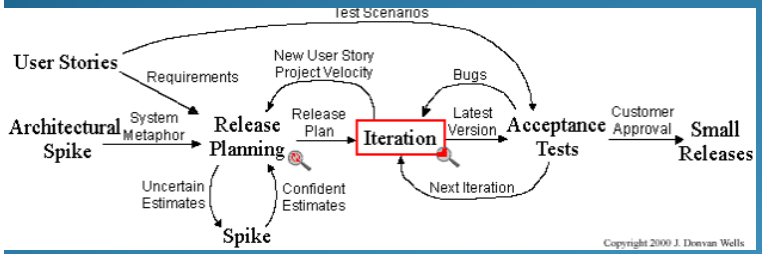
- Internal resource representation through ”contextual” VOTables;
- Experiment configuration and execution;
- user authentication and working session management;
- experiment data & working flow trigger and supervision;

Project Management Approach

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

G. Longo
Principal Investigator

M. Brescia
Project Manager

O. Laurino
Project Engineer

- **Science & Education:** A. Corazza, R. D'Abrusco, D. Capozzi, E. De Filippis, A. Staiano, R. Tagliaferri
- **Data Mining Modeling:** S. Cavuoti, C. Donalek
- **Infrastructure:** G. d'Angelo, N. Deniskina
- **Software Engineering:** A. Nocella, M. Garofalo, F. Manna, M. Fiore, B. Skordovski

precari inside

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

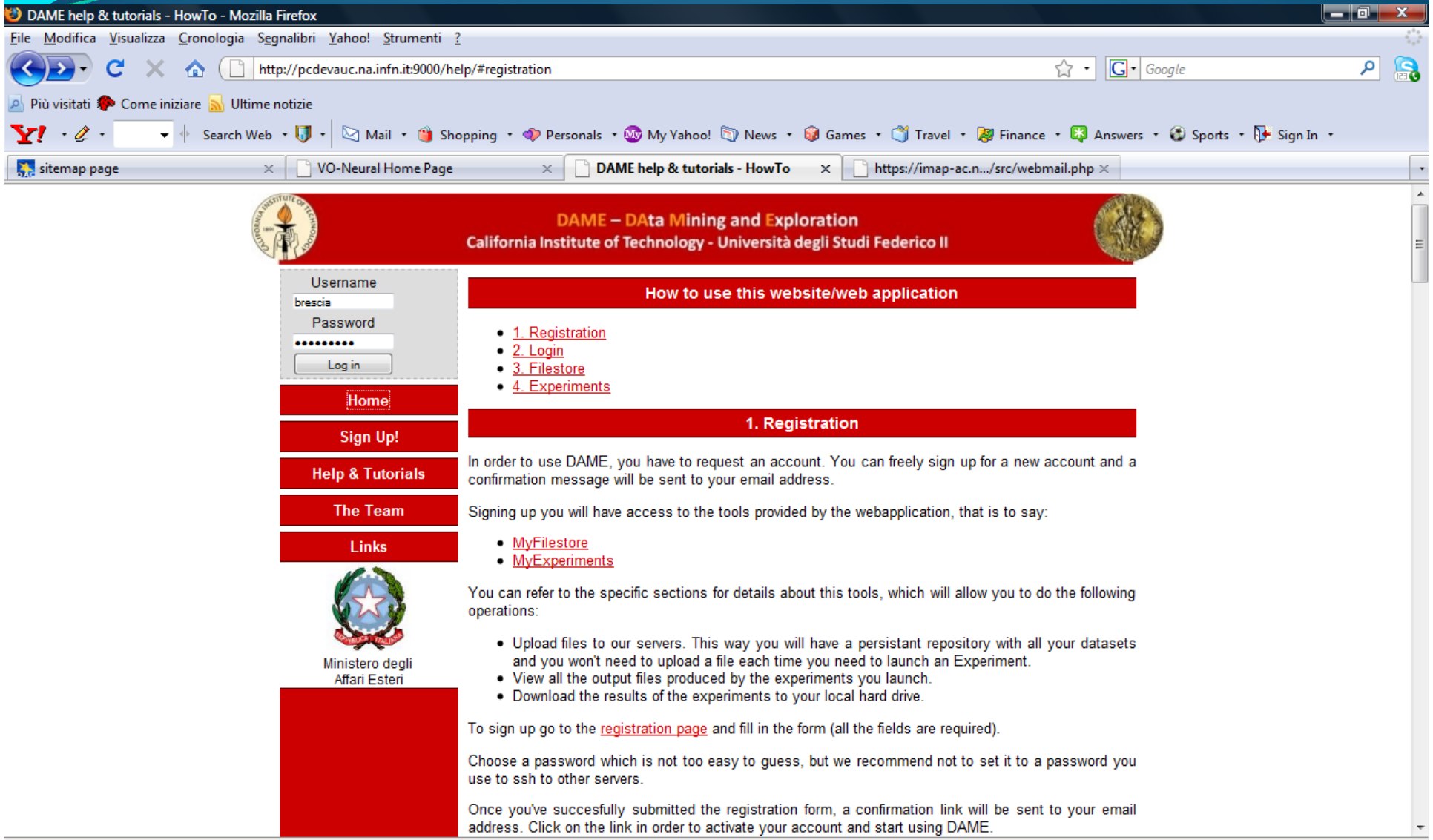
Download our SDK;

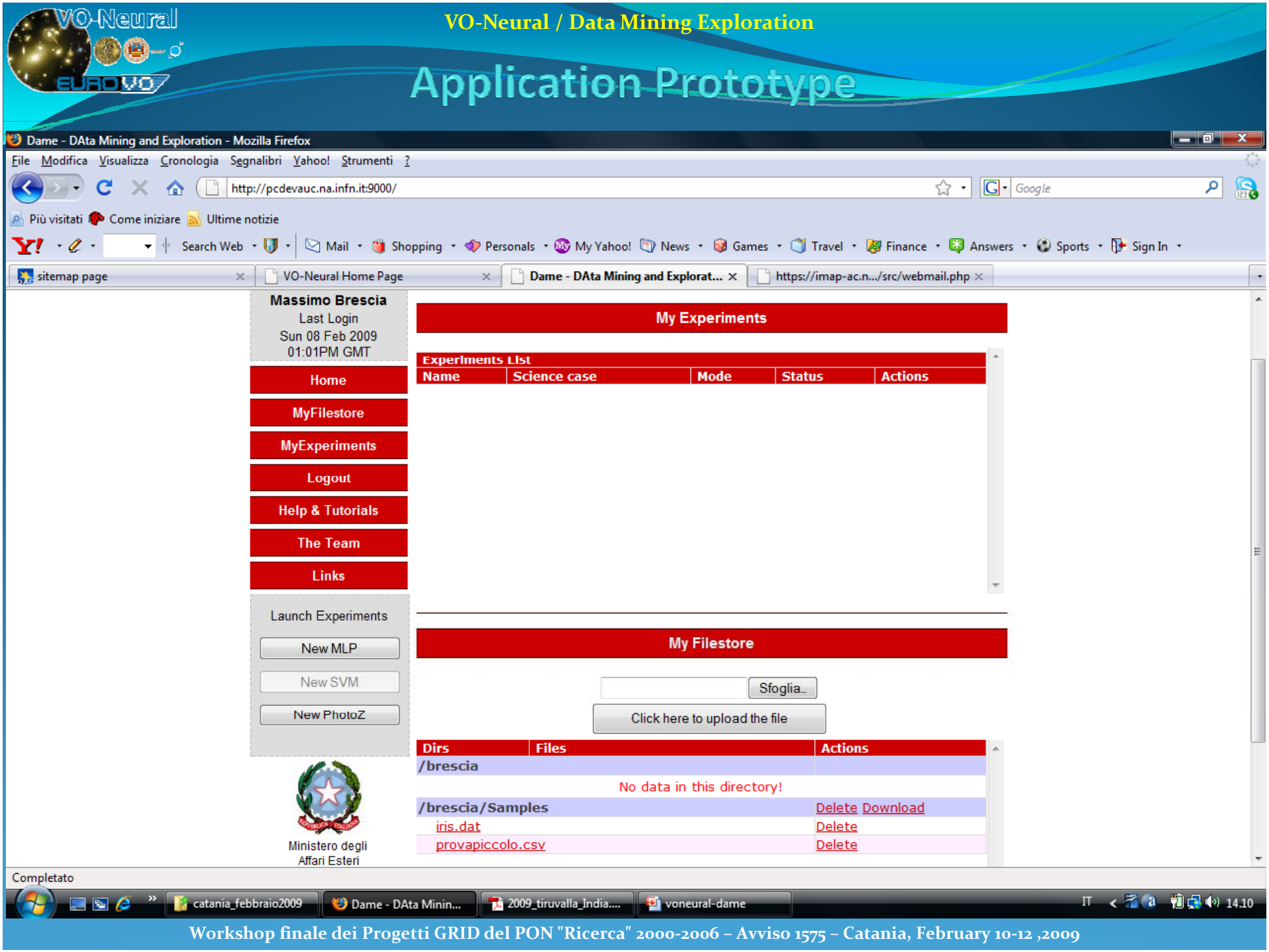
Implement and test the DMPlugin abstract class;

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

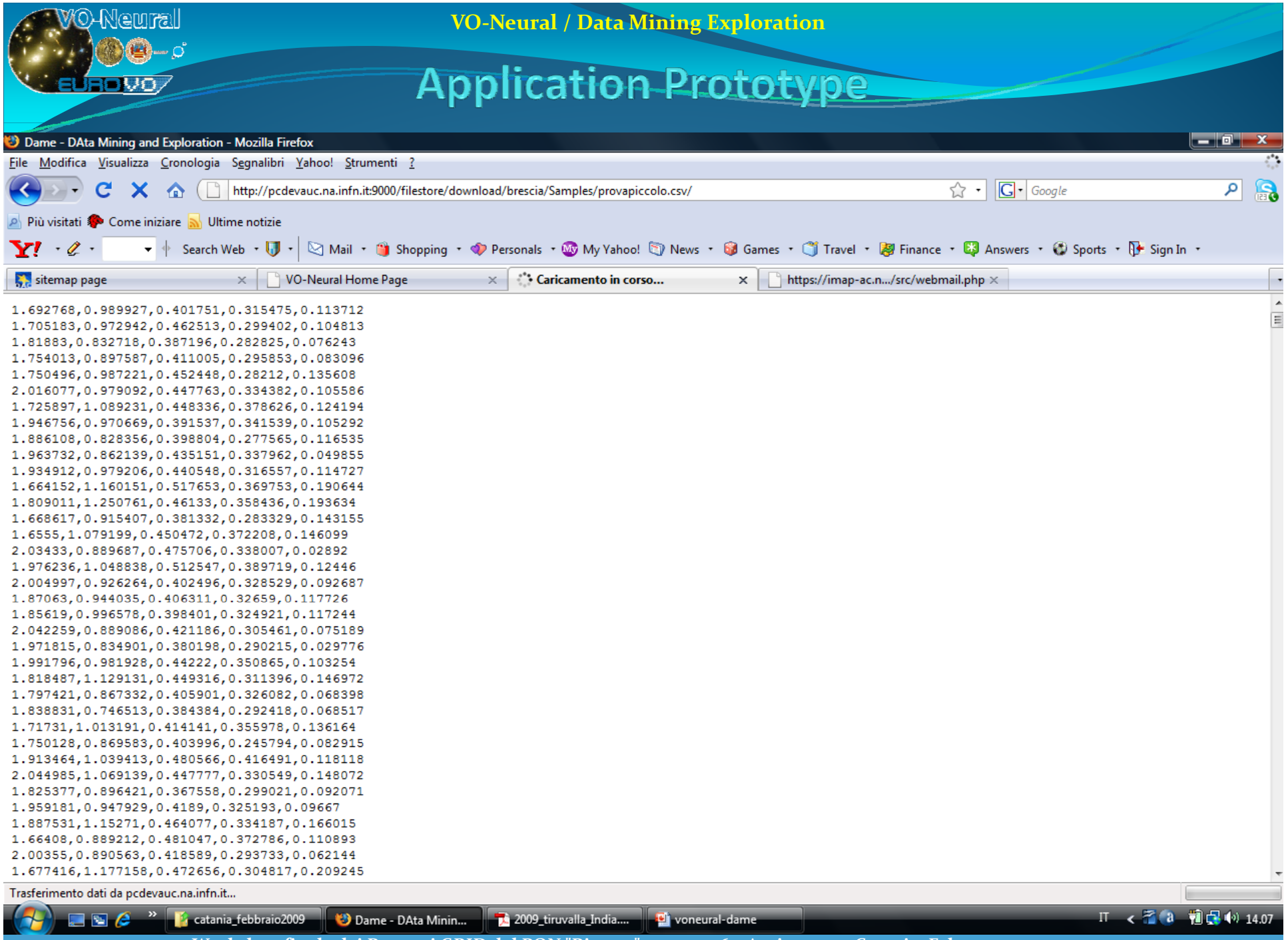
The same if you want to develop a new driver for a specific environment or storage system. Just implement the Driver Plugin Interface and register it;

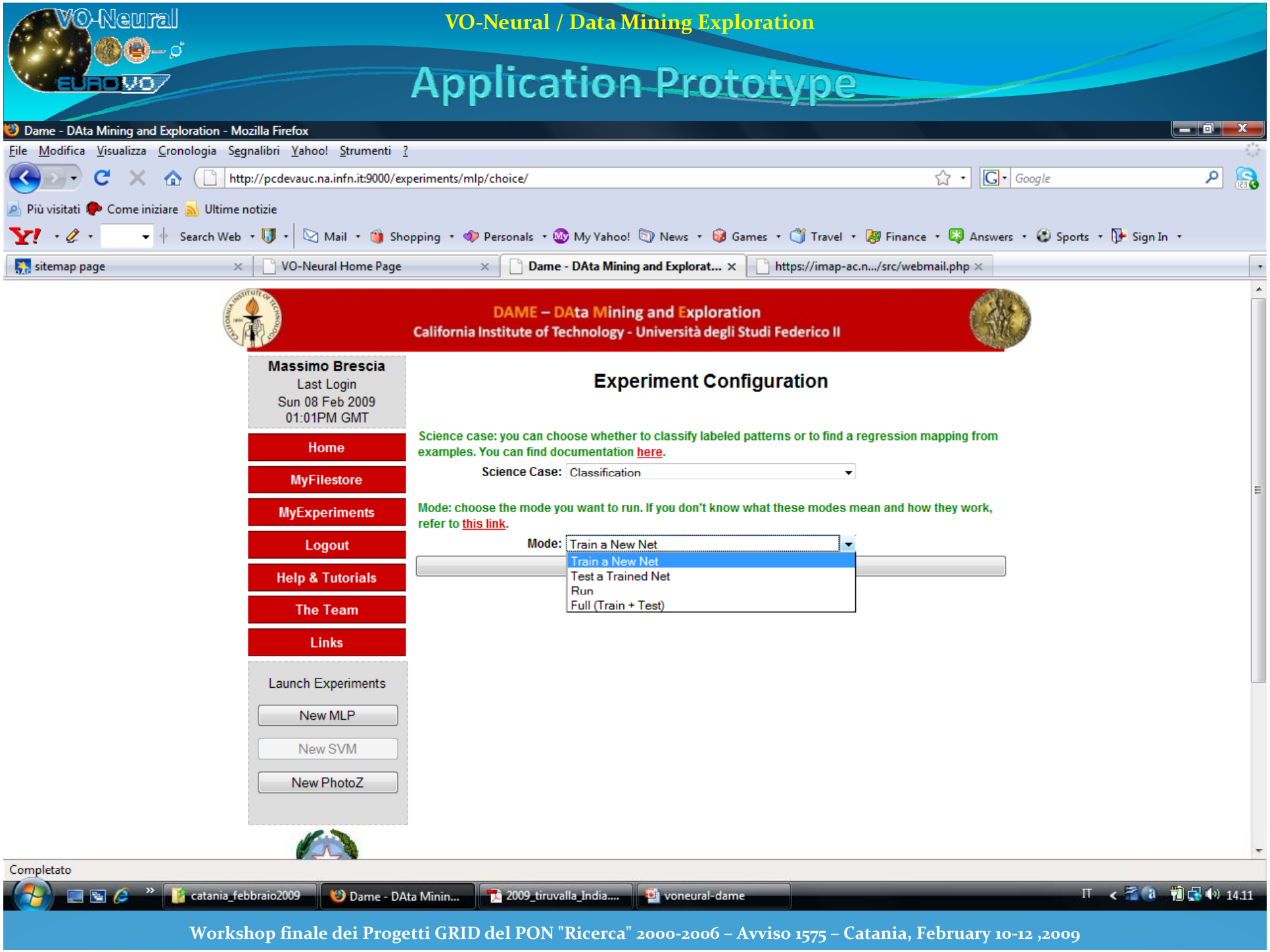
Application Prototype

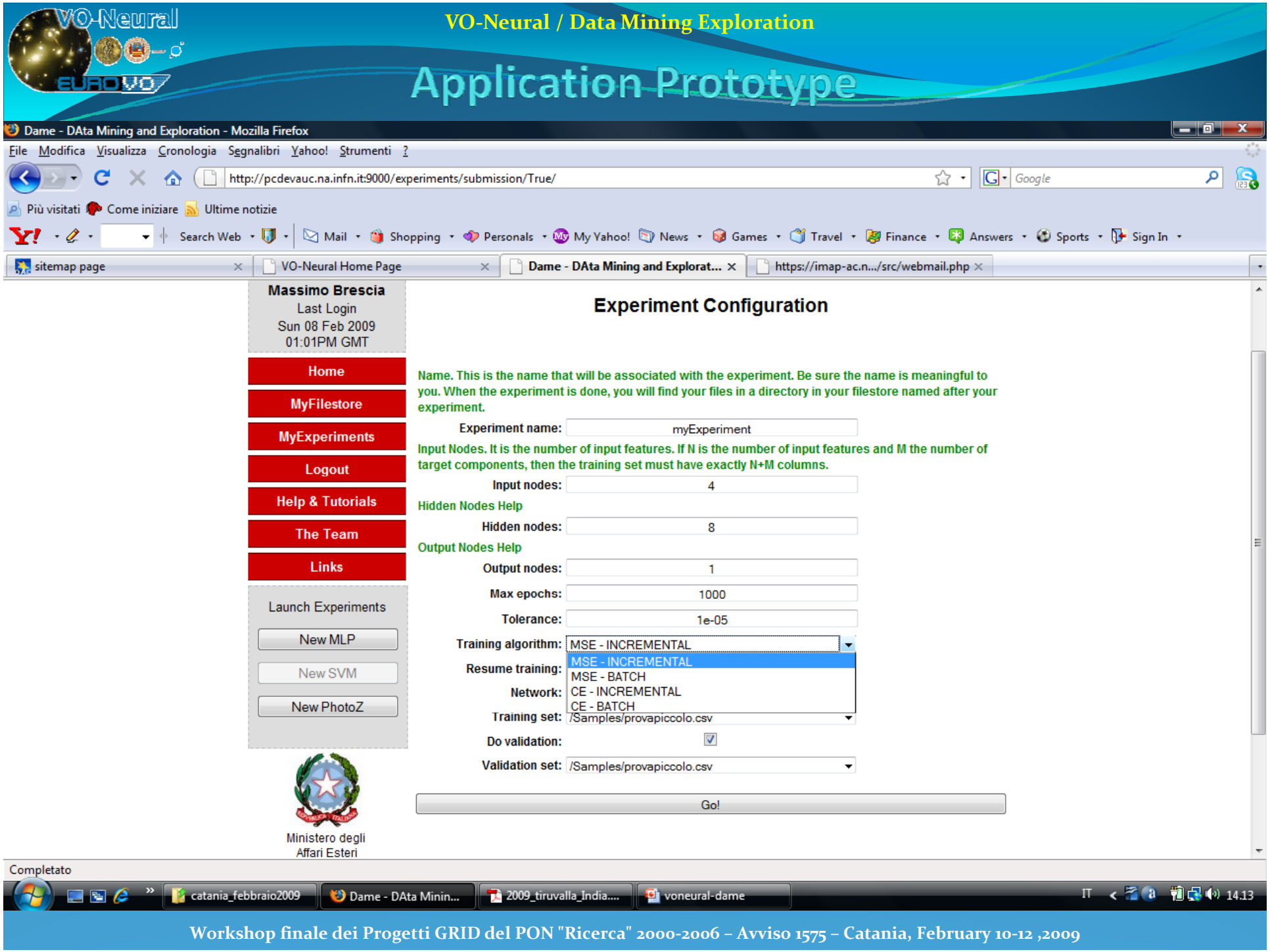


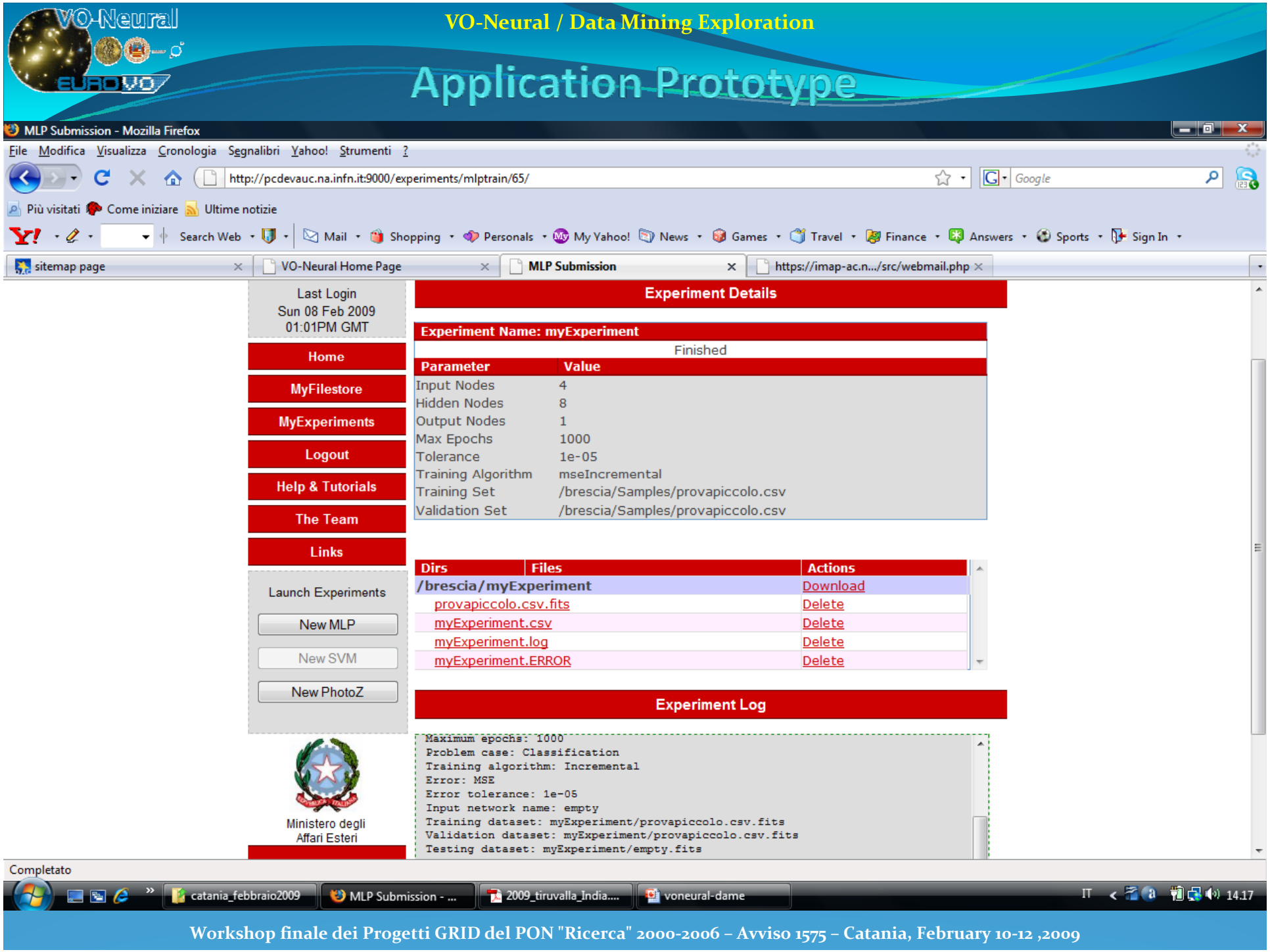

VO-Neural / Data Mining Exploration
Application Prototype
Dame - DAta Mining and Exploration - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti ?
http://pcdevauc.na.infn.it:9000/
Più visitati Come iniziare Ultime notizie
Search Web Mail Shopping Personals My Yahoo! News 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
14.10

Application Prototype




VO-Neural / Data Mining Exploration
Application Prototype
Dame - DAta Mining and Exploration - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Yahoo! Strumenti ?
http://pcdevauc.na.infn.it:9000/experiments/mlp/choice/
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
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: Classification
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: Train a New Net
Train a New Net
Test a Trained Net
Run
Full (Train + Test)
Completato
catania_febbraio2009
Dame - DAta Minin...
2009_tiruvalla_India....
voneural-dame
IT
14.11


VO-Neural / Data Mining Exploration
Application Prototype
Dame - DAta Mining and Exploration - Mozilla Firefox
http://pcdevauc.na.infn.it:9000/experiments/submission/True/
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
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: myExperiment
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: 4
Hidden Nodes Help
Hidden nodes: 8
Output Nodes Help
Output nodes: 1
Max epochs: 1000
Tolerance: 1e-05
Training algorithm: MSE - INCREMENTAL
MSE - INCREMENTAL
MSE - BATCH
CE - INCREMENTAL
CE - BATCH
Resume training:
Network:
Training set: /Samples/provapiccolo.csv
Do validation:
Validation set: /Samples/provapiccolo.csv
Go!
Completato


VO-Neural / Data Mining Exploration
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 & 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/provapiccolo.csv
Validation Set /brescia/Samples/provapiccolo.csv
Dirs
Files
Actions
/brescia/myExperiment
Download
provapiccolo.csv.fits
Delete
myExperiment.csv
Delete
myExperiment.log
Delete
myExperiment.ERROR
Delete
Experiment Log
Maximum epochs: 1000
Problem case: Classification
Training algorithm: Incremental
Error: MSE
Error tolerance: 1e-05
Input network name: empty
Training dataset: myExperiment/provapiccolo.csv.fits
Validation dataset: myExperiment/provapiccolo.csv.fits
Testing dataset: myExperiment/empty.fits
Completato
catania_febbraio2009
MLP Submission - ...
2009_tiruvalla_India....
voneural-dame
IT
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)

In this Conference poster session:

A web application for photometric redshifts evaluation
Omar Laurino et al.