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

Massimo Brescia - Giuseppe Longo

&

Project Team

INAF – Osservatorio Astronomico di Capodimonte

Dipartimento di Fisica – Università degli Studi di Napoli Federico II

California Institute of Technology

Project Highlights

Originally named VO-Neural, recently the project is evolving to DAME (DAta Mining & Exploration), but the final name and logo is still under design.

The project, an evolution of the former AstroNeural Collaboration, is financed through:
E.U. grant VOTECH and VO-AIDA
Italian Ministry of Research in the framework of the PON-S.Co.P.E.
Italian Ministry of Foreign Affairs through a great relevance bilateral project Italy-USA
VO-Neural/DAME main goal is the design and development of scientific data mining tools, based on Information Technology instruments.

Partnership:
Dipartimento di Fisica (sez. di Astrofisica) - Università degli Studi di Napoli Federico II
INAF - Osservatorio Astronomico di Capodimonte
California Institute of Technology, Pasadena - USA

Collaborations:
S.Co.P.E. (high Performance distributed Cooperative System for scientific Experiment)
INAF - Osservatorio Astronomico di Trieste
Dipartimento di Informatica - Università degli Studi di Napoli Federico II
Dipartimento di Ingegneria Informatica - Università degli Studi di Napoli Federico II
EURO-VO The European Virtual Observatory
IVOA (International Virtual Observatory Alliance)

Trend of Information Technology

Cloud / GRID computing

Cloud computing is Internet based development and use of computer technology. The cloud is a metaphor for the Internet and is an abstraction for the complex infrastructure it conceals.

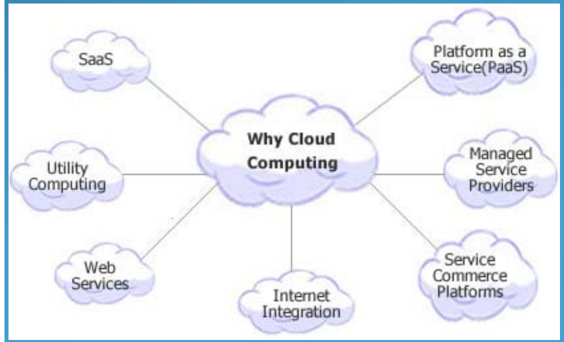
It is a style of computing, provided "*as a service*", to access enabled services from the Internet without knowledge of, expertise with, or control over the technology infrastructure that supports them.

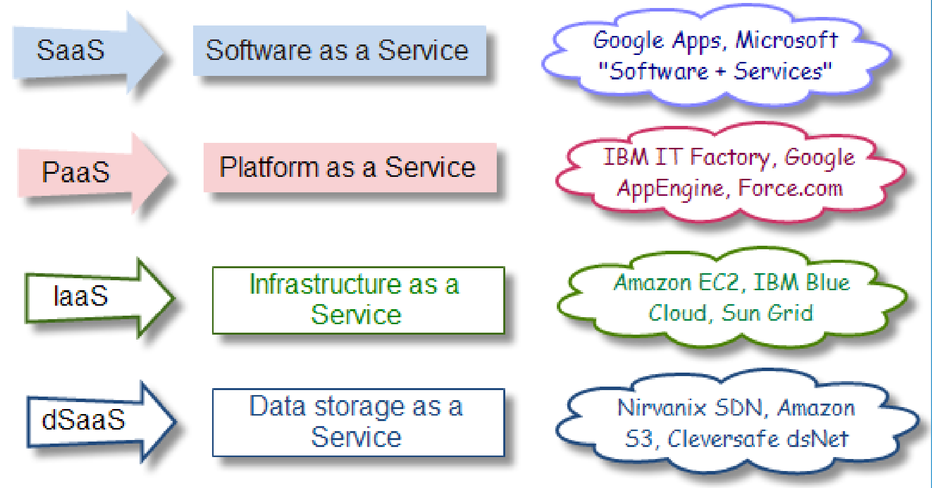
So far, Cloud computing can be considered to implement the following ideas :

Utility computing - which was first suggested by *John McCarthy* in 1961, where computing is viewed as a public utility;

Cluster computing - which views a group of linked computers as a single virtual computer for high-performance computing (HPC);

Grid computing - where the linked computers tend to be organized "*as resources*" to solve a common problem;





Cloud computing landscape

Scientific community requirements

Why to land on an "open" distributed infrastructure

EGEE GRID *PRIVATE GRID*

Enabling GRID for E-SciencE

Target Group	Scientific community	Business
Service	short-lived batch-style processing (job execution)	long-lived services based on hardware virtualization
SLA	Local (between the EGEE project and the resource providers)	Global (between Amazon and users)
User Interface	High-level interfaces	HTTP(S), REST, SOAP, Java API, BitTorrent
Resource-side middleware	Open Source (Apache 2.0)	Proprietary
Ease of Use	Heavy	Light
Ease of Deployment	Heavy	Unknown
Resource Management	probably similar	
Funding Model	Publicly funded	Commercial

A universal research infrastructure:

"*Un ambiente dove le risorse di ricerca (HW, SW e DATI) possano essere condivise rapidamente e a cui si possa accedere da ovunque sia necessario promuovere una ricerca migliore e più efficace*"

Virtualized service/resource oriented Infrastructure

Virtualization brings new standardized capabilities to data centers

Virtualized Data Center

GRID Computing

Resource oriented

Grid computing solutions enable parallel processing of computational tasks, often using idle vs. dedicated capacity.

Goals: "Accelerate throughput from decades to days or from months to minutes."

"Enable deep computations that are otherwise intractable."

Characteristics: Large numbers of work requests run for short periods of time (minutes / hours).

CLOUD Computing

Service oriented

Cloud computing enables self-service provisioning of virtual machines.

Goal: "Simplify deployment of Operating Systems and app servers."

Characteristics: Small numbers of VM allocations held for long periods of time (days/months).

Join data virtualization, resources and services brings to

Virtual Organization of data

HW resource oriented

SW service oriented

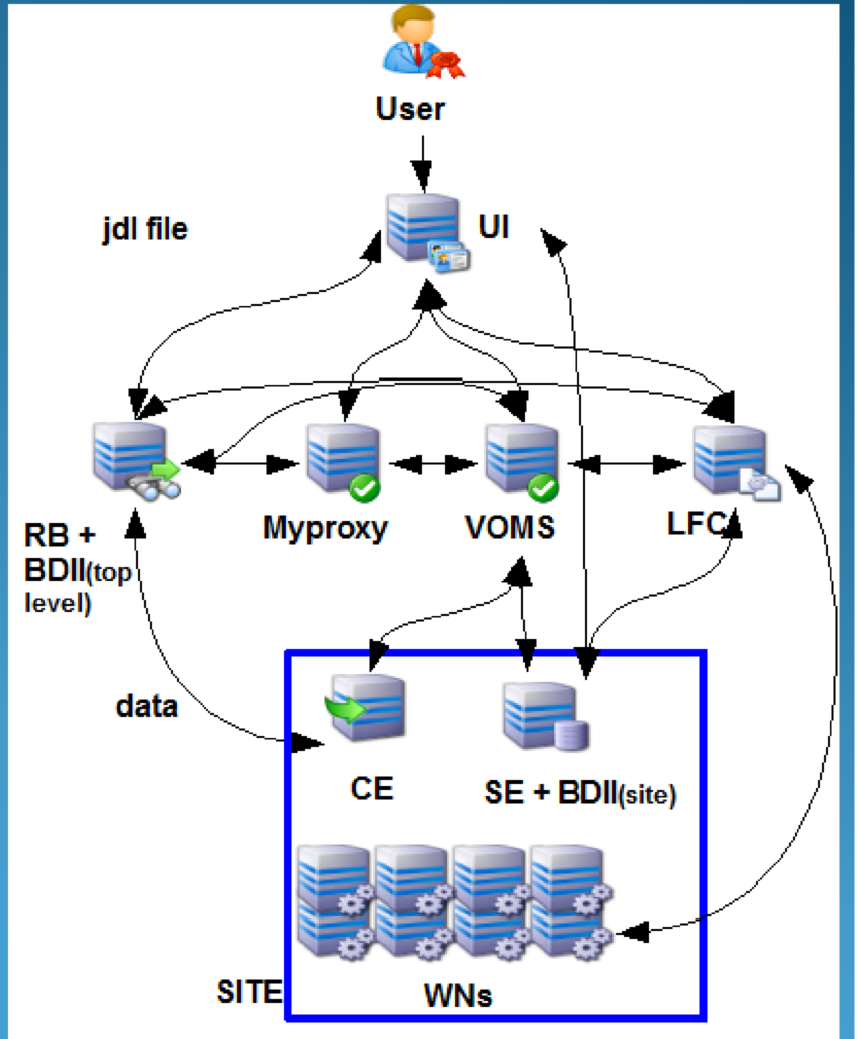
What is S.Co.P.E. GRID

UNIONE EUROPEA

Ministero dell'Università e della Ricerca

PON Ricerca Scientifica Sviluppo Tecnologico Alta Formazione 2000 - 2006

Progetto SCOPE Università degli Studi di Napoli Federico II



INFRASTRUCTURE INTEROPERABILITY

APPLICATION INTEROPERABILITY

Strategic goal:

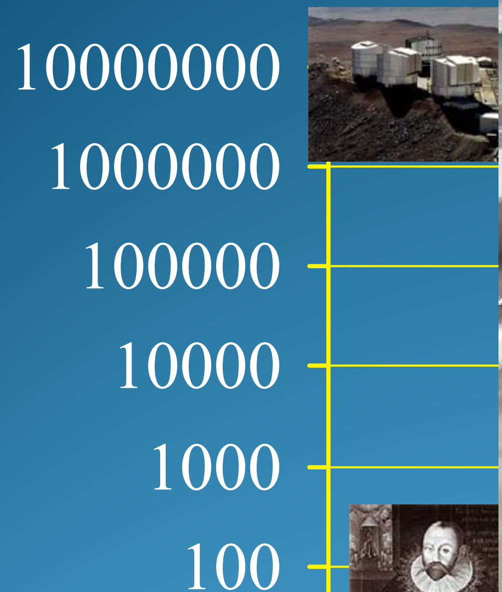
ITALIAN e-INFRASTRUCTURE OPEN TO THE COLLABORATION & INTERACTION BETWEEN RESEARCH AND INDUSTRY

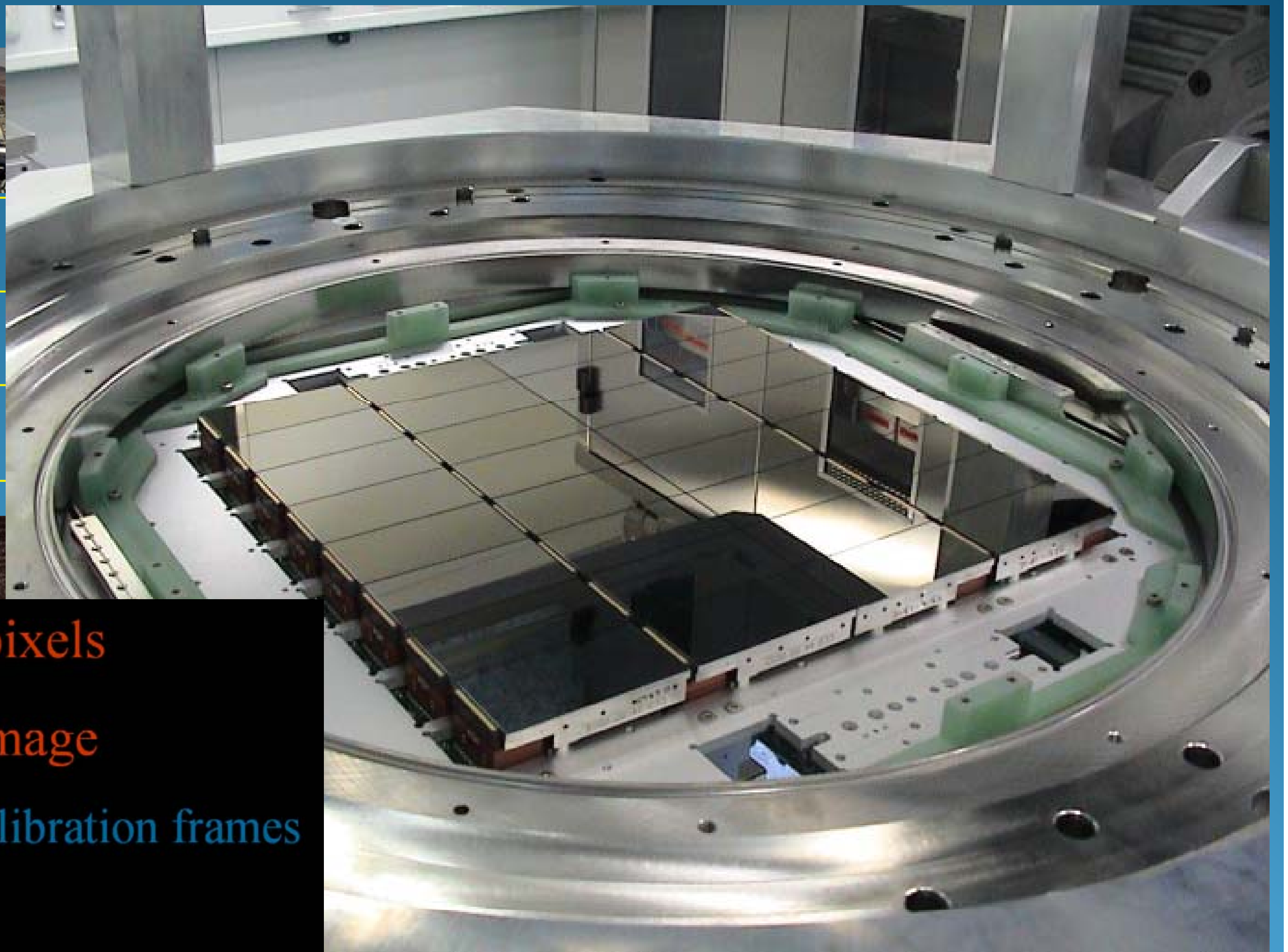
Resources:

- Computing power of some thousands of cores per project
- hundreds of TB per project
- distributed resources for massive computing

The astrophysical problem

Astronomical data rate





268,435,456 pixels

0.5 Gbyte x image

50 science frames + 50 calibration frames

⇩

50 Gbyte / Night

The Astrophysical Application

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

We need:

data archives organized in a unified Virtual Observatory for wide band cross-correlation;
Data mining software tools based on machine learning and self-adaptive mechanisms;
Distributed high performance computing infrastructure able to work on massive datasets;

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

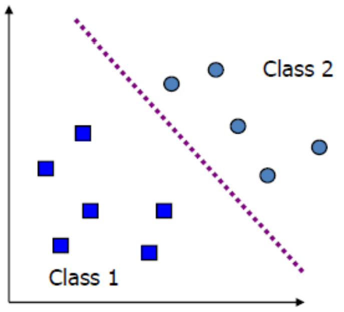
To implement KDD tools is expensive (time, computing, need for specialists), requires **coordinated efforts** between astronomers and computer scientists and is aimed to fulfill the needs of **large projects**

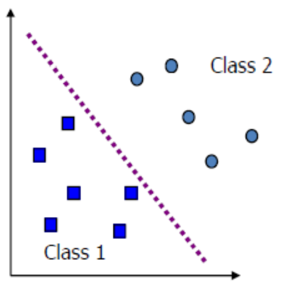
Learning problems as "function approximation"

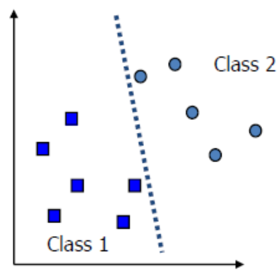
$\mathbf{X} \equiv \{x_1, x_2, x_3, \ldots x_N\}$ **input vectors**

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

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



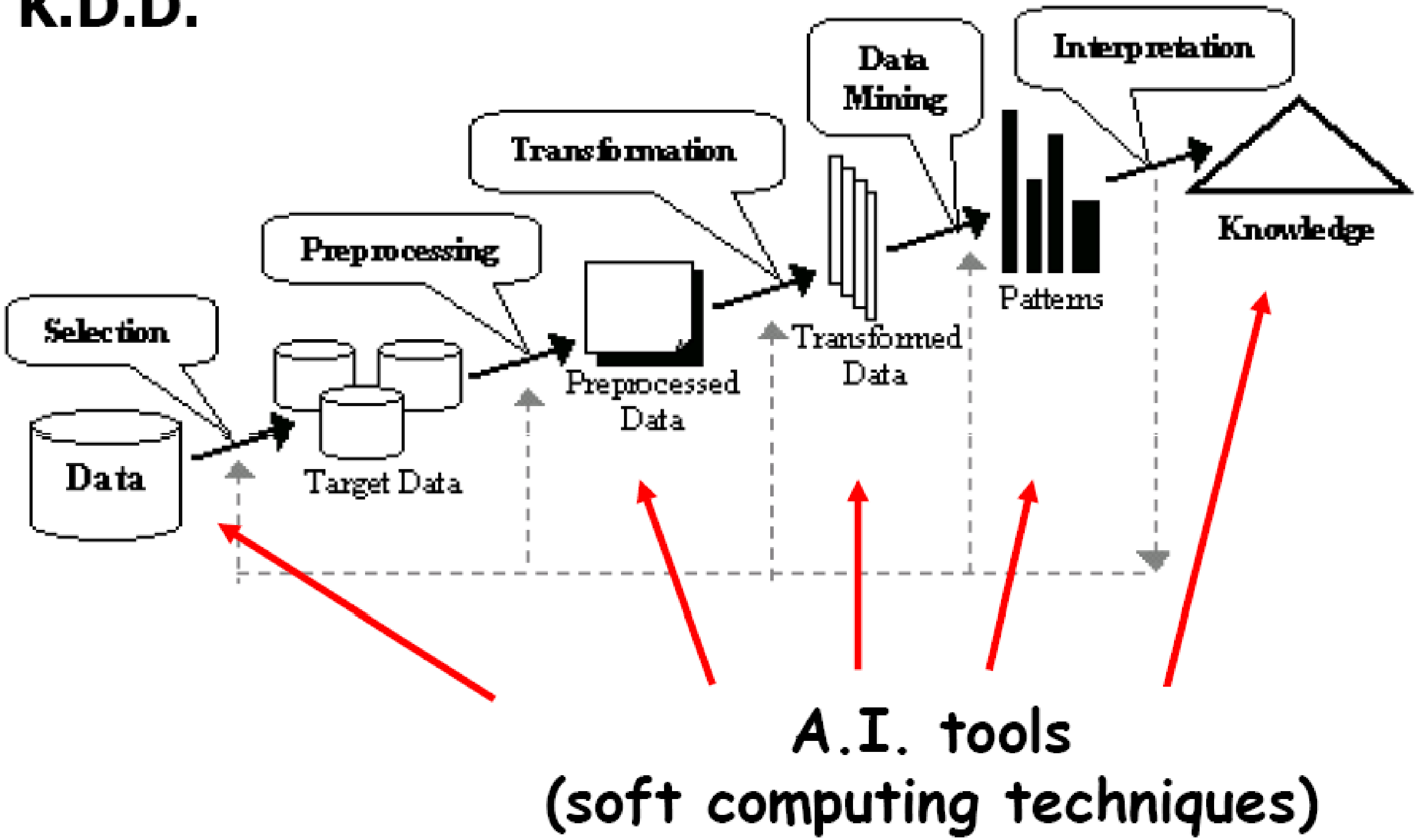




variable	characteristics	Type	operation
Quantitative	Numerical with ordering relationship and possibility to define a metric	Actual measurement	regression
Categorical (non ordered)	Membership into a finite umber of classes. No ordering relationship.	Numerical codes (targets) arbitrarily orderd	Classification
Ordered categorical	Classes orderd by a relationship but there is no metric	Numerical codes n on arbitrarily orderd	Classification

Where do A.I. may fit into

K.D.D.



Data Mining Exploration with VO-Neural/DAME

Machine learning methods can be broadly grouped in:

Supervised methods

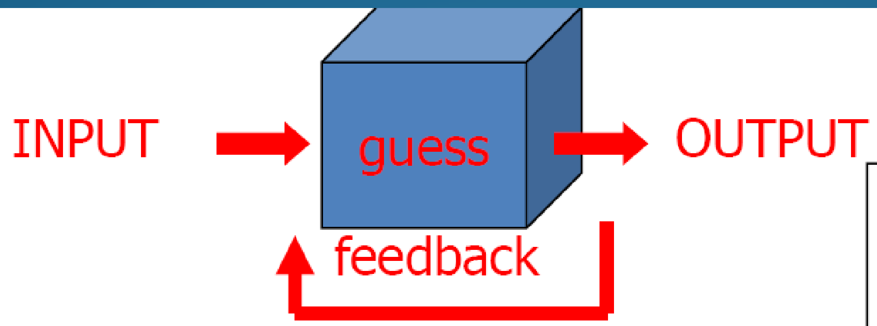
They learn how to partition the parameter space by means of a training phase based on examples.

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

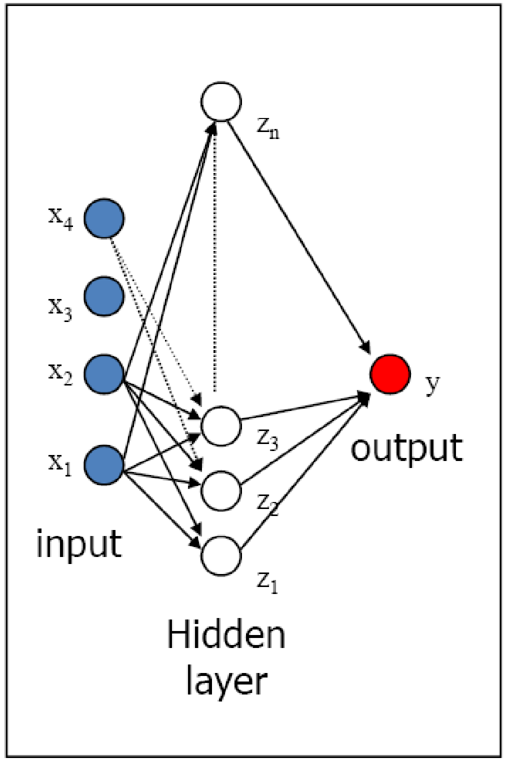
Supervised Models: Multi Layer Perceptron



- input layer (n neurons)
- M hidden layer (1 or 2)
- Output layer (n' <n neurons)

Neurons are connected via activation functions

Different NN's given by different topologies, different activation functions, etc.



Supervised Models: Multi Layer Perceptron

$$N(U,e,P(e)) = \frac{1}{2k}\sum_j \left(Y_j - P(e)\right)^2$$

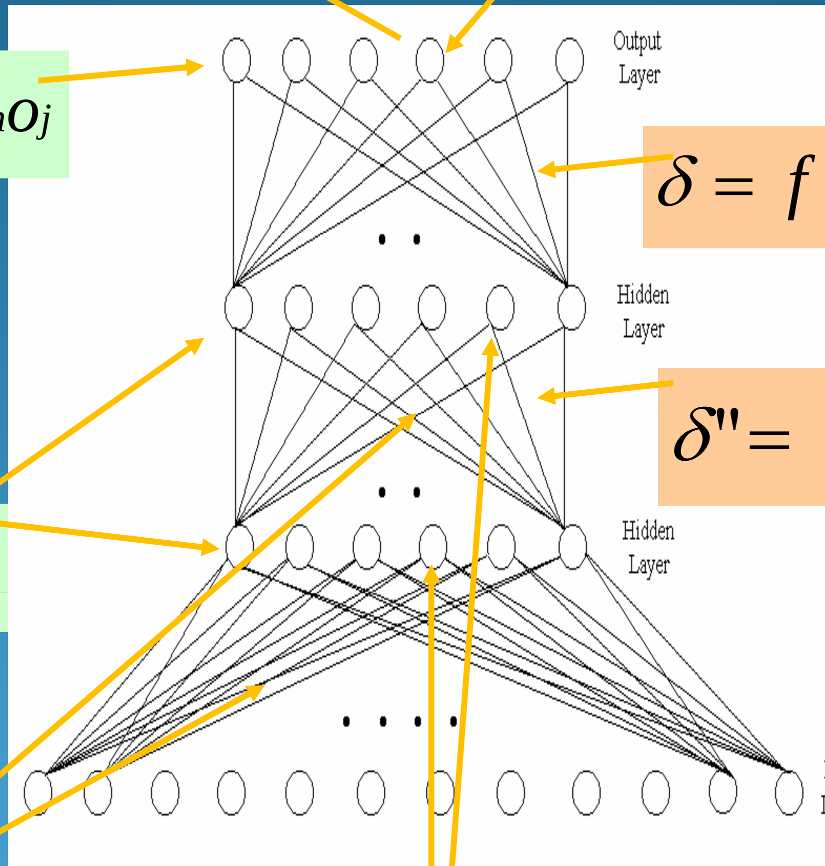
$$\left|Y - P(e)\right| < \varepsilon$$

backward

$$net\_input_j = \sum_h w_{jh} o_j$$

$$\delta = f'(o)\left(Y - P(e)\right)$$

forward



$$\delta'' = f'(o)\sum w\delta$$

$$net\_input_h = \sum_i w_{hi} o_i$$

Back Propagation learning algorithm

$$w_{ji}(new) = w_{ji}(old) + \eta\delta_i o_j + \alpha\Delta w_{ji}(old)$$

$$f(o) = \frac{1}{1+\exp(-o)}$$

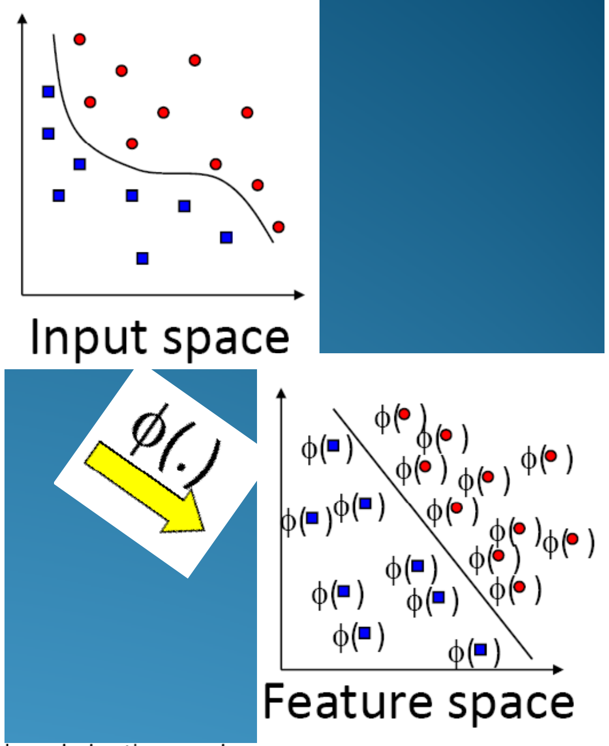
Supervised Models: Support Vector Machine

given a training set formed by pairs [features-label]: (x_i, y_i), i = 1…l
where $x_i \in R^n$ e $y_i \in \{1,-1\}^l$.
Support Vector Machines (SVM) try to solve the following optimization problem:

$$\min_{\omega,b,\xi} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i$$

With the condition:

$$y_i(\omega^T\phi(x_i)+b) \geq 1-\xi_i$$



Vectors x_i are mapped into an higher dimensionality space where the SVM identify an hyper plane which maximizes the distances from the two classes

C > 0 is a classification error correction term

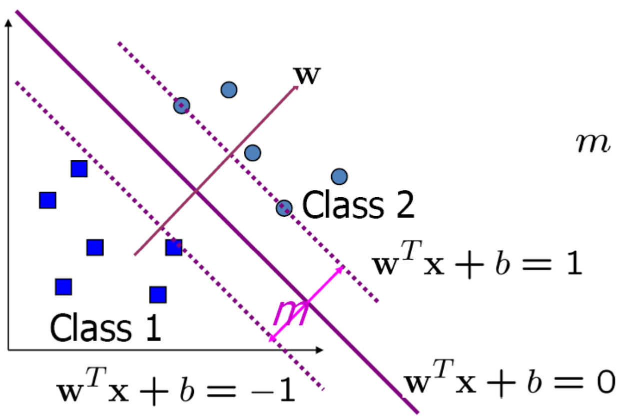
$$K(x_i, x_j) = \phi(x_i)^T(x_j)$$

Is the so called Kernel function

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \qquad \gamma > 0$$

- linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$.
- polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\mathbf{x}_i{}^T\mathbf{x}_j + r)^d$, $\gamma > 0$.
- radial basis function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma > 0$.
- sigmoid: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\mathbf{x}_i{}^T\mathbf{x}_j + r)$.

– We should maximize the margin, *m*

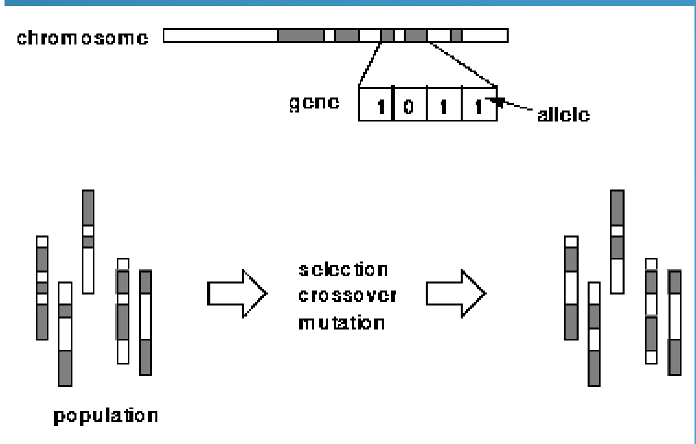


$$m = \frac{2}{||\mathbf{w}||}$$

Supervised Models: MLP & Genetic Algorithms

Genetic algorithms are a part of **evolutionary computing**, which is a rapidly growing area of artificial intelligence. As you can guess, genetic algorithms are inspired by Darwin's theory about evolution. Simply said, solution to a problem solved by genetic algorithms is evolved.

If we are solving some problem, we are usually looking for some solution, which will be the best among others. The space of all feasible solutions is called **search space.** Each point in the search space represent one feasible solution. Each feasible solution can be "marked" by its value or fitness for the problem. We are looking for our solution, which is one point (or more) among feasible solutions - that is one point in the search space. The looking for a solution is then equal to a looking for some extreme (minimum or maximum) in the search space. The search space can be whole known by the time of solving a problem, but usually we know only a few points from it and we are generating other points as the process of finding solution continues.



Chromosomes are strings of DNA and serves as a model for the whole organism. A chromosome consist of **genes**, blocks of DNA. Each gene encodes a **trait**, for example color of eyes. Possible settings for a trait (e.g. blue, brown) are called **alleles**. Each gene has its own position in the chromosome. This position is called **locus**.

Complete set of chromosomes is called **genome**.

Supervised Models: MLP & Genetic Algorithms

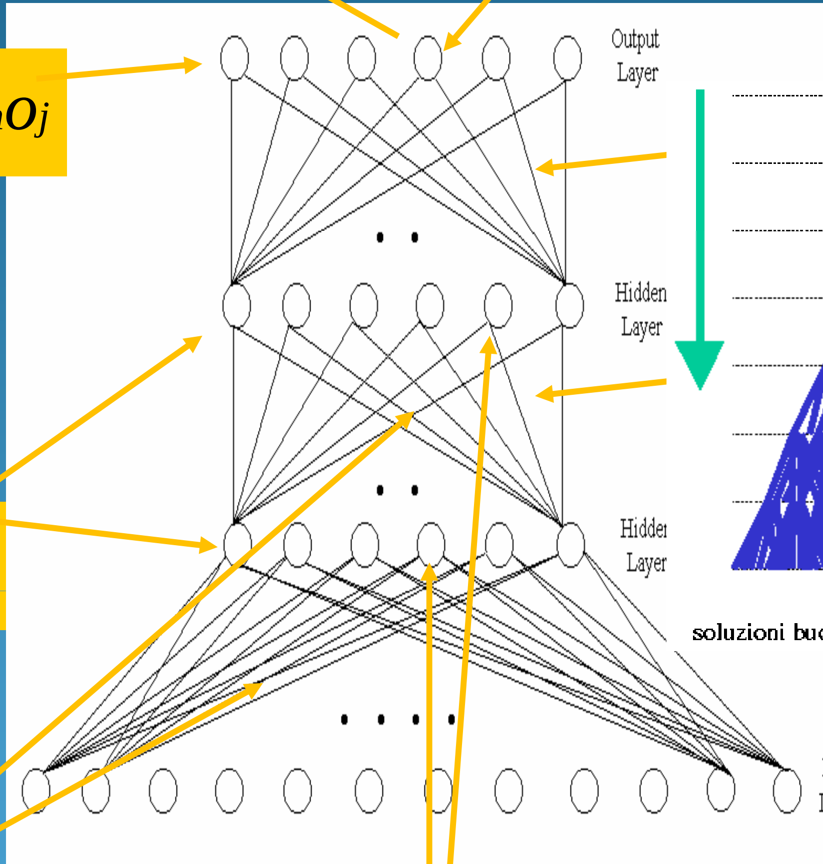
$$N(U,e,P(e)) = \frac{1}{2k}\sum_j (Y_j - P(e))^2$$

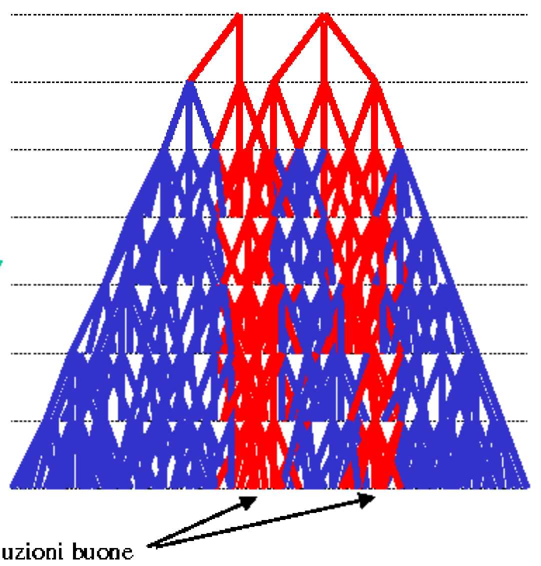
$$|Y - P(e)| < \varepsilon$$

backward

$$net\_input_j = \sum_h w_{jh} o_j$$

forward





$$net\_input_h = \sum_i w_{hi} o_i$$

$$w_{ji}(new) = w_{ji}(old) + \eta \delta_i o_j + \alpha \Delta w_{ji}(old)$$

$$f(o) = \frac{1}{1+\exp(-o)}$$

Due to the complexity and quantity of source code (different languages, input data formats, multi-platform handling, information flow etc.), internal design standard and protocols became fundamental constraints.

Example of source code design standardization:

Supervised Models: MLP & Genetic Algorithms
UML & OOP approach

http://www.na.astro.it/~brescia/mlpga/html/index.html

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)

Unsupervised Models: Self Organizing Maps

The SOM is an algorithm used to visualize and interpret large high-dimensional data sets

The map consists of a regular grid of processing units, "neurons". A vector consisting of features, is associated with each unit. The map attempts to represent all the available observations with optimal accuracy. At the same time vectors become ordered on the grid so that similar vectors are close to each other and dissimilar vectors far from each other.

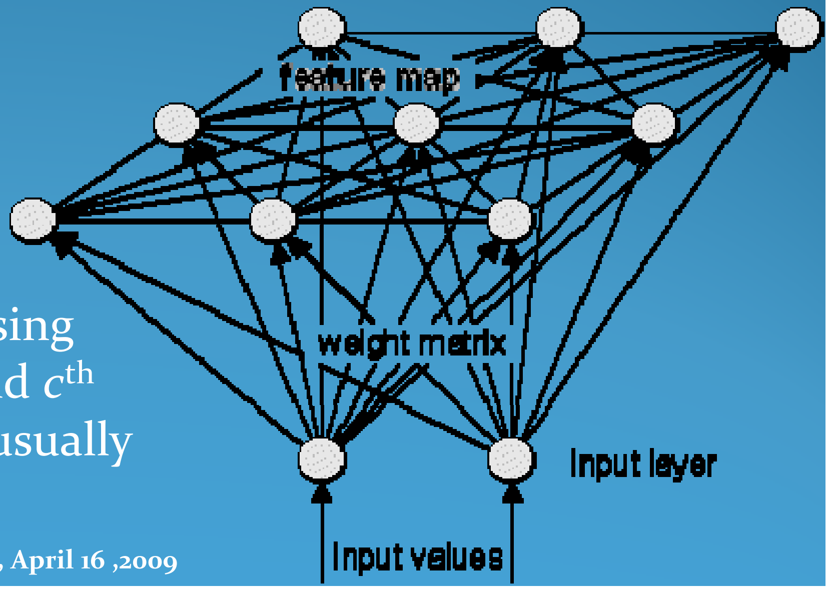
Fitting of the model vectors is usually carried out by a sequential regression process, where t = 1,2,... is the step index: For each sample $\mathbf{x}(t)$, first the winner index c (best match) is identified by the condition

$$\forall i, ||\mathbf{x}(t) - \mathbf{m}_c(t)|| \leq ||\mathbf{x}(t) - \mathbf{m}_i(t)||.$$

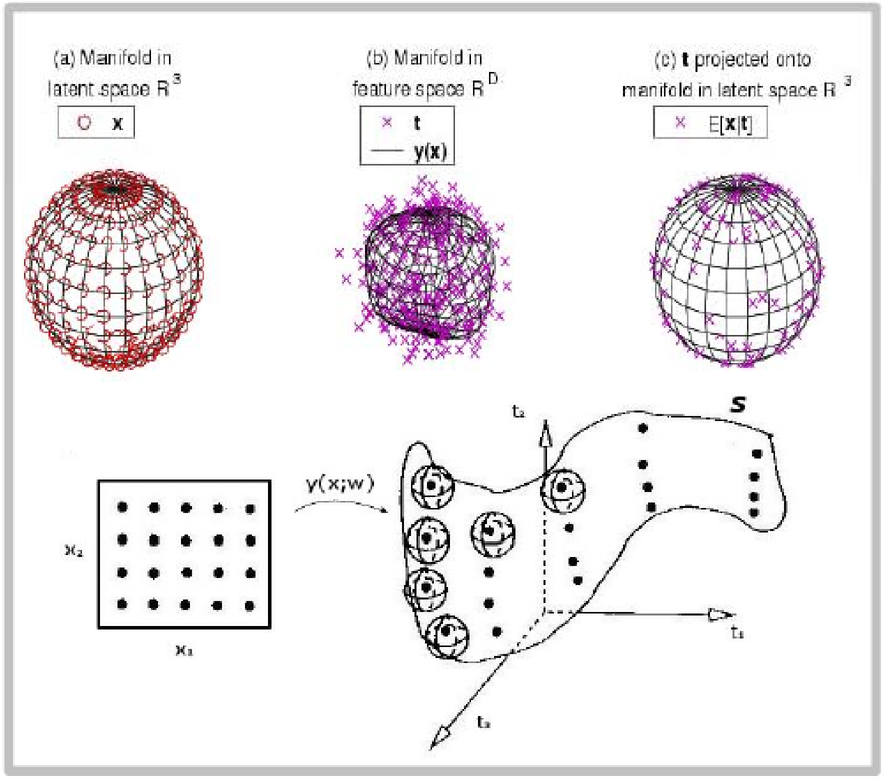
After that, all model vectors or a subset of them that belong to nodes centered around node $c = c(\mathbf{x})$ are updated as

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{c(\mathbf{x}),i}(\mathbf{x}(t) - \mathbf{m}_i(t)).$$

$h_{c(\mathbf{x}),i}$ is the ``neighborhood function'', a decreasing function of the distance between the i^{th} and c^{th} nodes on the map grid. This regression is usually reiterated over the available samples.



Unsupervised Models: PPS & NEC



PPS: the Beauty of Spheres

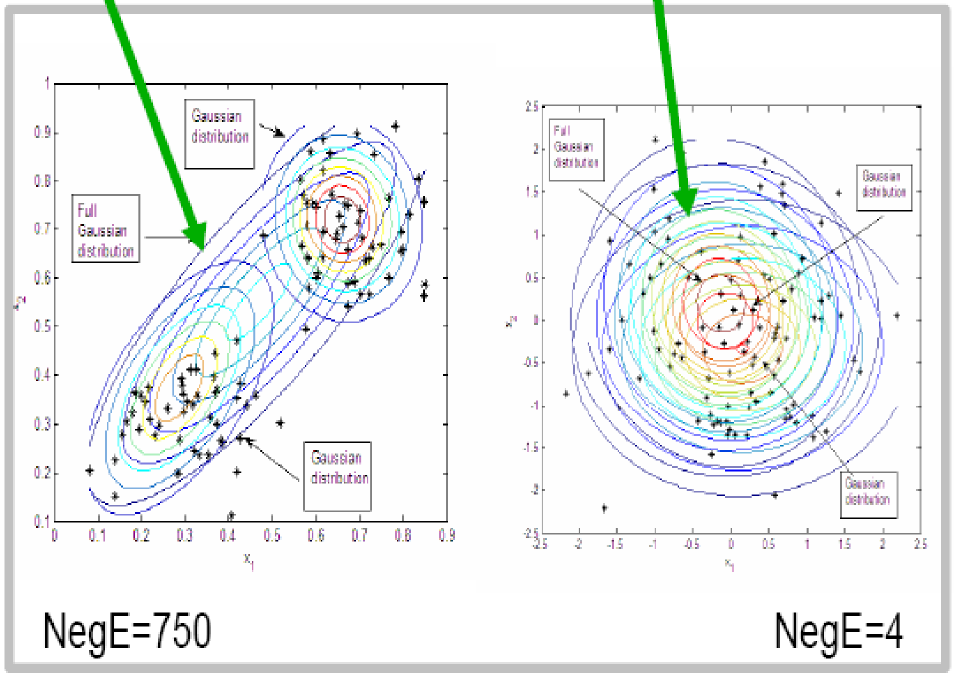
The original *m*-dimensional data space is mapped in a lower *n*-dimensional space, called "latent space". Visualization ease as a spherical manifold is fitted to the data, then projected into the manifold in R^3 and plotted as points on the sphere surface.
Each latent variable on the sphere is responsible for a number of projected points, which form a "cluster".

NEC: a matter of Gaussians

Clustering method based on the "neg-entropy" NegE, a measure of non gaussianity of a variable. If *A* is gaussian, then NegE(*A*) = 0. Given a threshold *d:*

If NegE(*A* U *B*) < *d*, then clusters *A* and *B* are replaced by cluster *A* U *B*

Not replaced! Replaced!



Project Target

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

Data Farming:

- Storage/Archiving
- Indexing, Search ability
- Data Fusion, Interoperability

Data Mining (or Knowledge Discovery in Databases):

- Pattern or correlation search
- Clustering analysis, automated classification
- Outlier / anomaly searches
- Hyper-dimensional visualization

Data understanding

- Computer aided understanding
- KDD
- Etc.

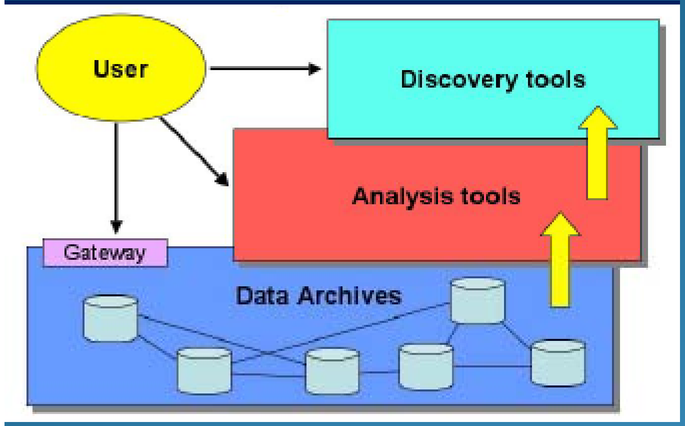
New Knowledge

Database technologies

Key mathematical issues

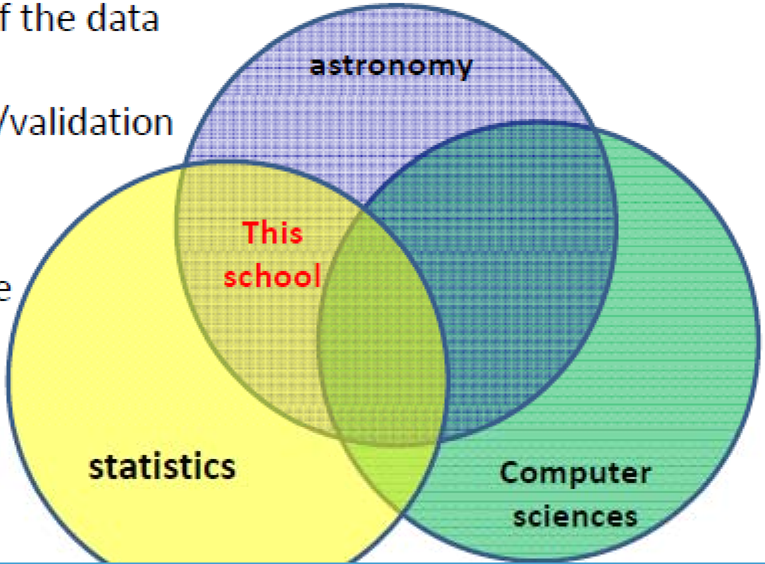
Ongoing research

Project Kick-off



In 2007, a group of astronomers, computer scientists, engineers and physicians started to explore possible joined effort to create a data mining toolset, based on a distributed infrastructure, 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.

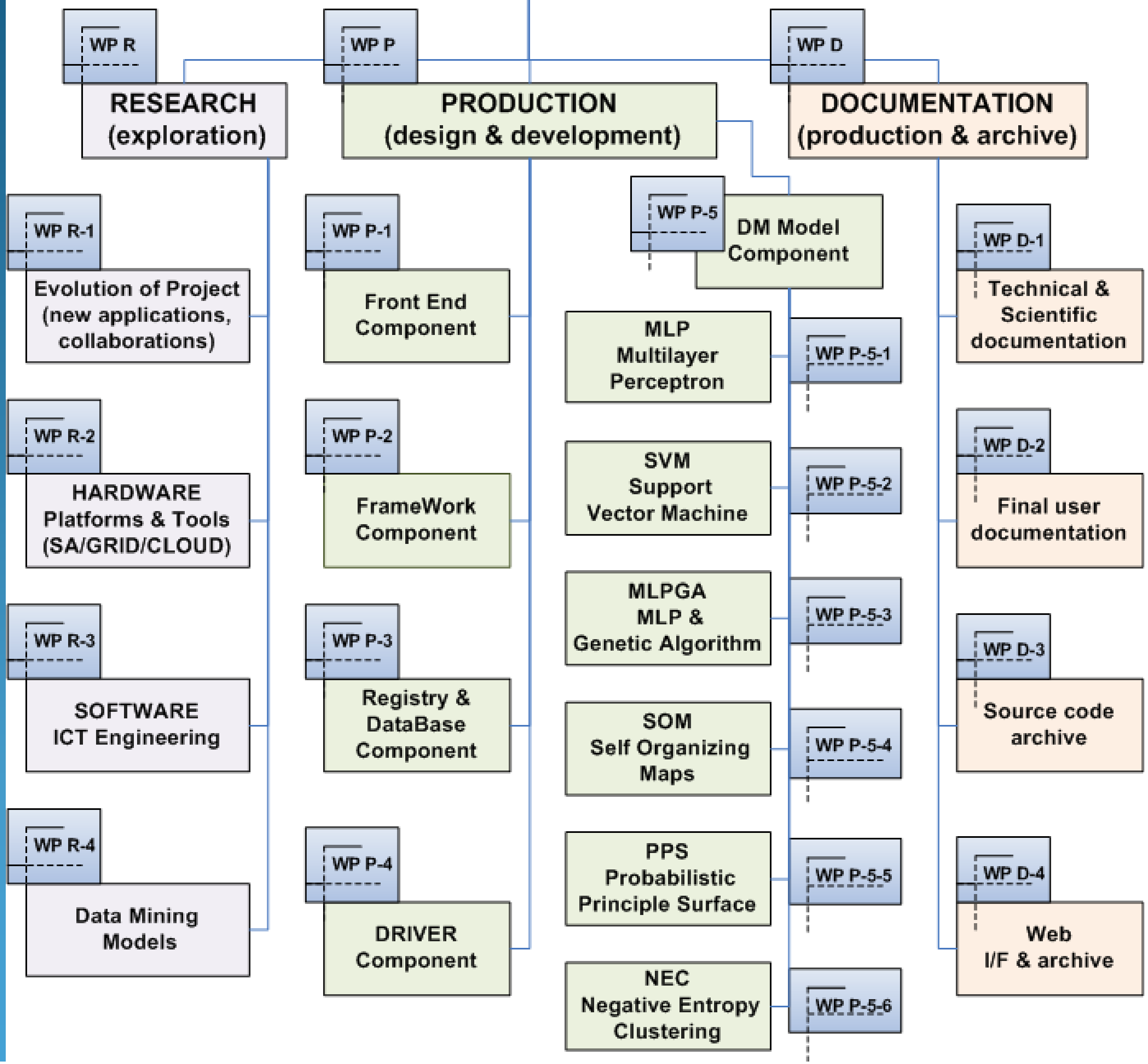


Project Management WBS

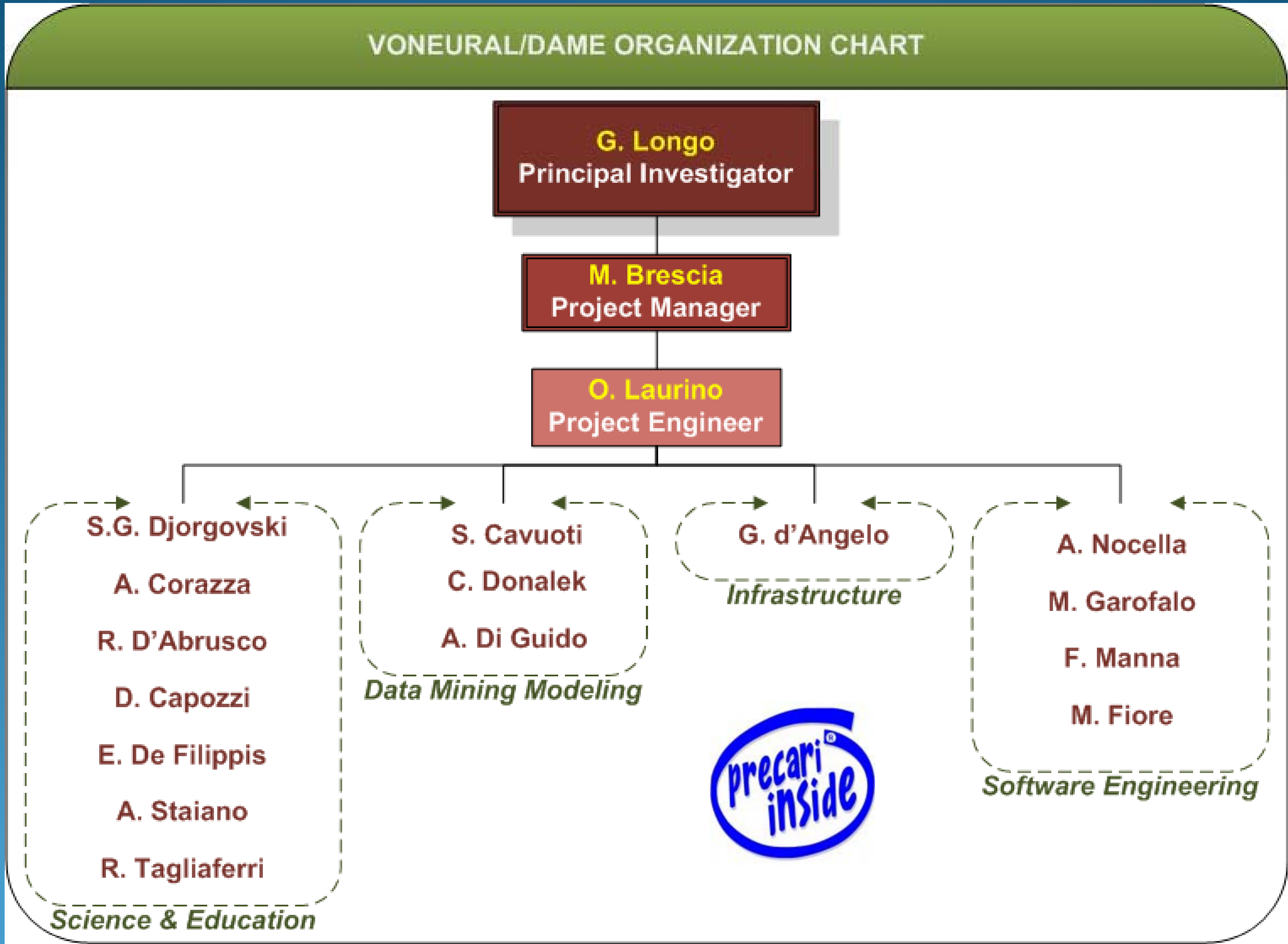
(released on September 2007)

DAME

Work Breakdown Structure



Project Team (now)



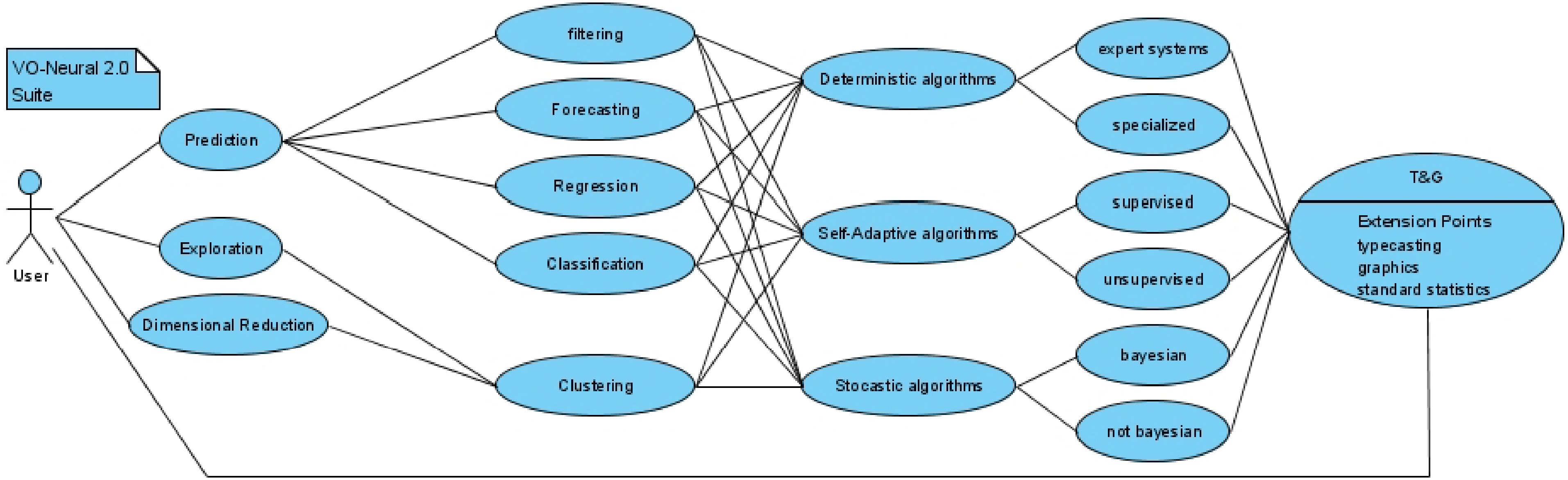
Project team (history)

NOME	RUOLO	TOT MESI	DA	A	POSIZIONE	BACKGROUND
Longo	P.I.	28			staff	astrofisica
Brescia	P.M.	28			staff	informatica/astrofisica
Djorgovski	SCIENCE+ICT	28			staff	astrofisica
Corazza	SCIENCE+EDU	22	13/06/07		staff	informatica
Donalek	AI+SW	28			staff	informatica
D'abrusco	SCIENCE	28			post-doc	fisica
Laurino	P.E.	28			tesi laurea	fisica
Garofalo	SW ENG	28			tesi laurea	ingegneria informatica
Nocella	SW ENG	28			tesi laurea	ingegneria informatica
Cavuoti	SCIENCE+SW	28			contratto	fisica
d'Angelo	WEB+DOC+SW	28			contratto	fisica
Manna	SW ENG	7	24/09/08		tirocinio	informatica
Fiore	SW ENG	6	22/10/08		tirocinio	informatica
Di Guido	SW ENG	2	02/03/09		tirocinio	informatica
Deniskina	SW ENG	12	06/09/07	04/09/08	contratto	informatica
Skordovski	SW	19	13/06/07	31/12/08	tesi laurea	informatica
Russo	SW ENG	10	13/06/07	15/04/08	tesi laurea	informatica
Vaccaro	SW ENG	4	13/06/07	25/09/07	tesi laurea	informatica
Formicola	SW	6	20/02/08	07/07/08	tesi laurea	informatica

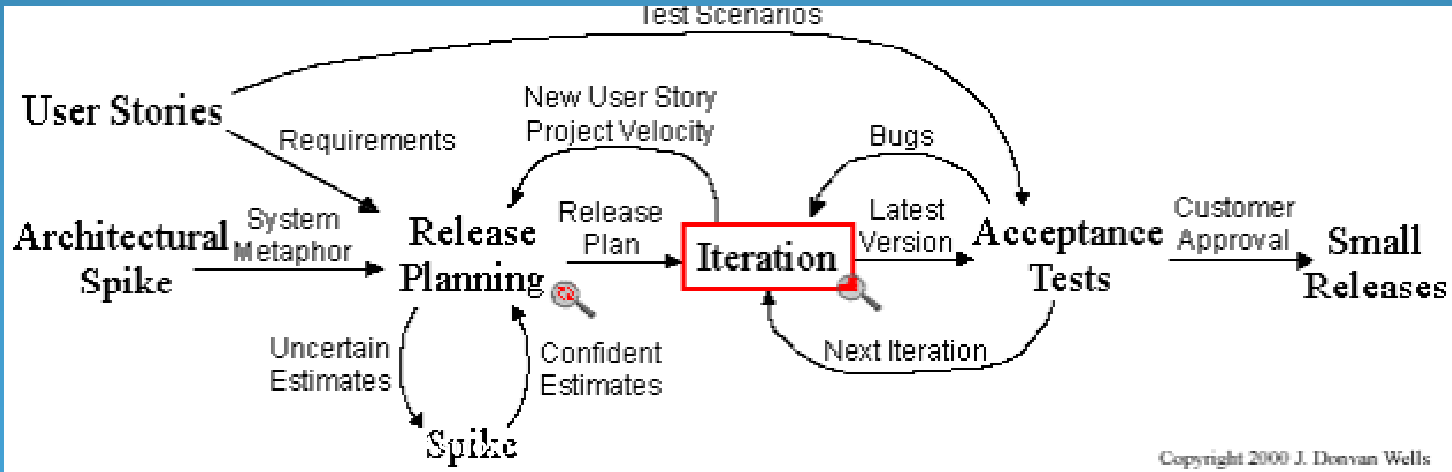
VO-Neural / Data Mining Exploration

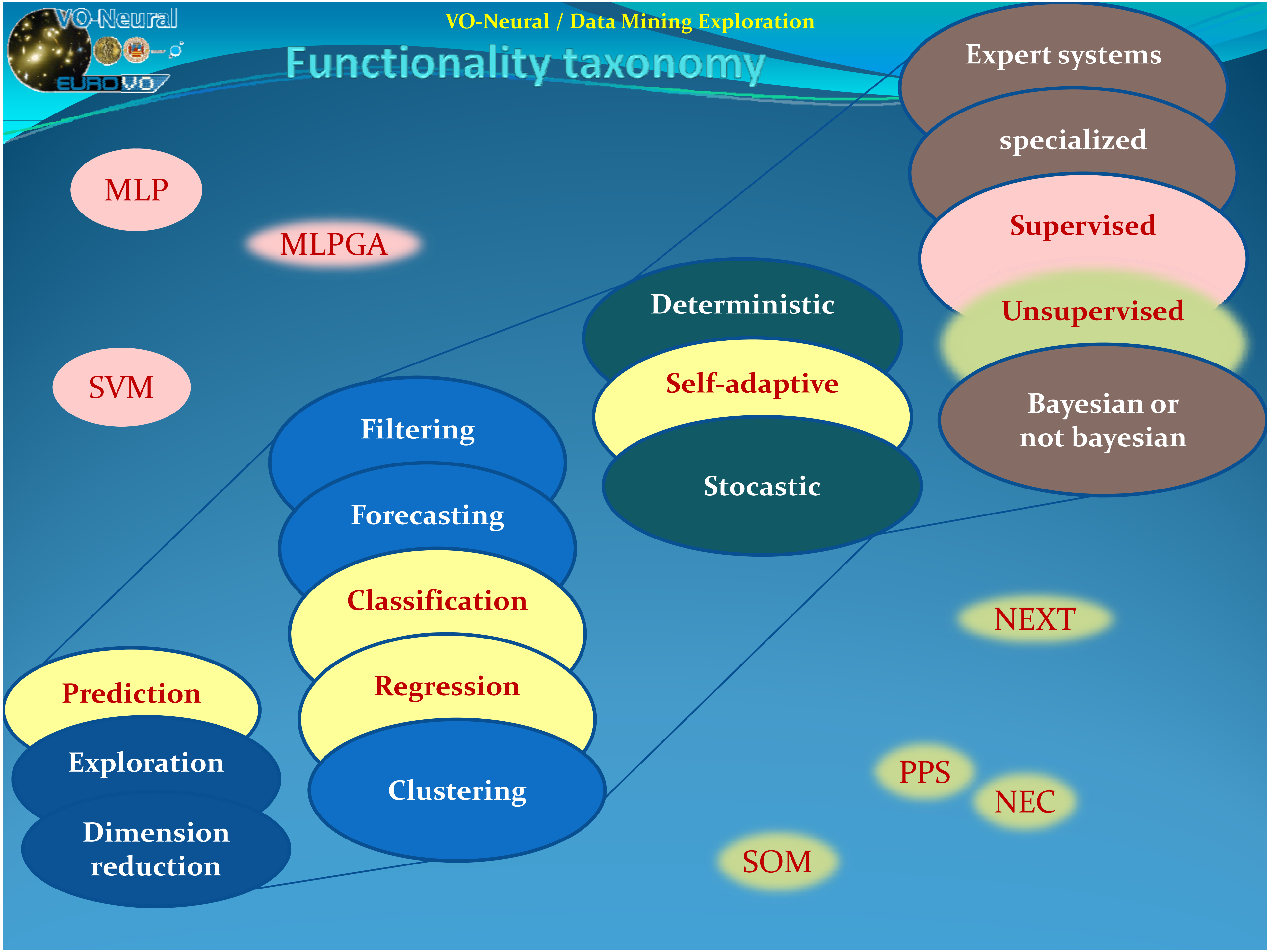
Project Management Highlights

U.M.L. Tools for Suite Functionalities and internal code design



XP – eXtreme Programming as project life cycle




VO-Neural / Data Mining Exploration
Functionality taxonomy
Expert systems
specialized
Supervised
Unsupervised
Bayesian or
not bayesian
MLP
MLPGA
SVM
Deterministic
Self-adaptive
Stocastic
Filtering
Forecasting
Classification
Regression
Clustering
Prediction
Exploration
Dimension
reduction
NEXT
PPS
NEC
SOM

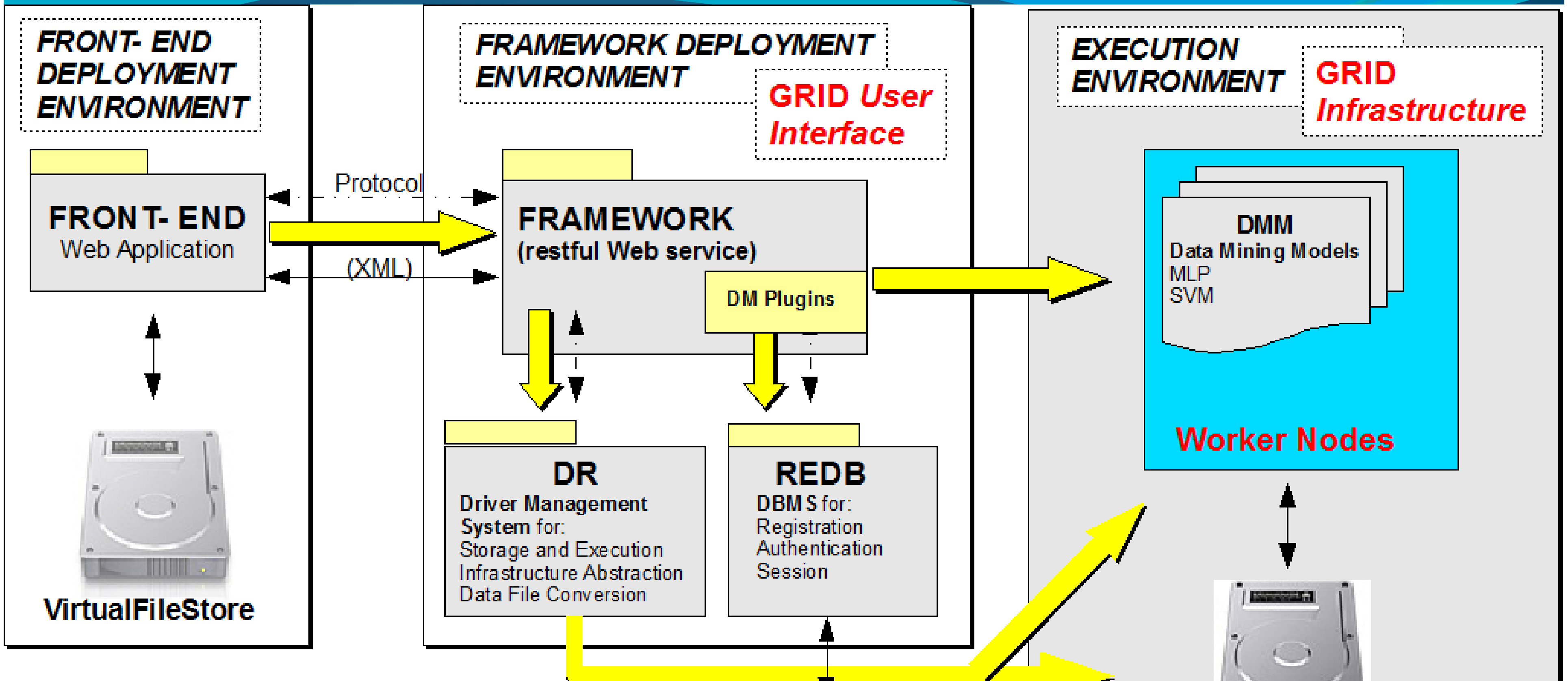
PROJECT DOCUMENTATION PRODUCTION

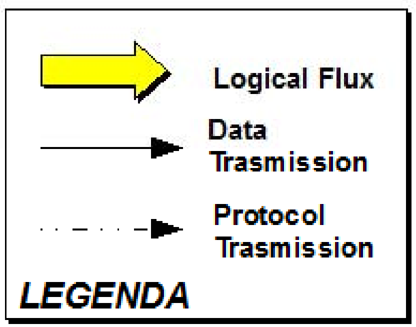
SECTION	CODE	2007	2008	2009	2010	MEANING
MANAGEMENT	SOW	2				Statement Of Work
	PLA	1	1			Project Plan & Work Breakdown Structure
	MIN	12	22	8		Minutes of Meeting
	SCH	1	1			Project schedule
	LIS	12	22	8		Action list
	DOC	8	5	2		Documentation & Website status
DESIGN	SPE	8	1			Technical Specifications
	PDD		1			Project Description Document
	SRS		5			Software Requirement specification
	SDD			5		Software Design Description
DEVELOPMENT & TEST	TRE	5	7			Technical Report
	PRO	3	1	5		Test Procedure
	VER	2				Test Verification & report
RELEASES	MAN		2			User Manual
	IDM					Installation & Deployment Manual
	REN					Release Notes
PUBLICATIONS	TECH		3			Technical papers
	SCIENCE		5	2		Scientific papers
TOTALS		54	76	30		160
PRESENTATIONS	PRE	3	2	4		Meeting talk/poster

Project Milestones

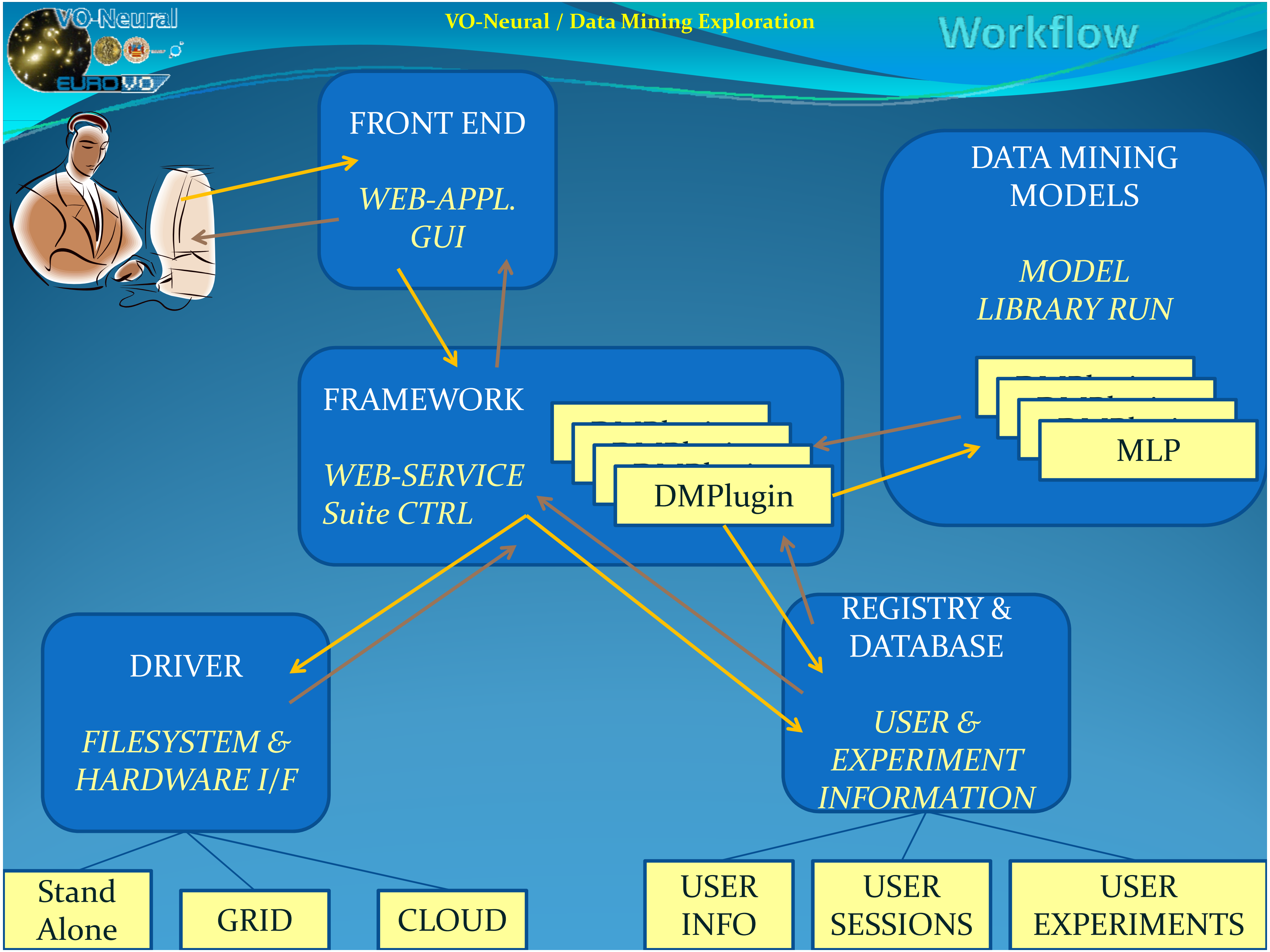
PROJECT MILESTONES	2007	2008	2009	2010
Statement of work				
WBS & Project Plan				
Project Design Description				
Software Requirement Specifications				
DM models Implementation	MLP	SVM	MLPGA NEXT	SOM PPS NEC
DM models scientific validation				
Software Design Description				
Implementation & Test Procedures				
Technical Reports				
Test Reports				
Beta Release Deployment				
User & Maintenance Manuals				
Commissioning of beta release				
Official release 1.0				
New DM models Implementation				
New Functionalities Implementation				

Architecture





- Object Oriented Programming & UML
- Internal standards and protocols (XML)
- 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 platform driver
- Data conversion and manipulation support


VO-Neural
EURO VO
VO-Neural / Data Mining Exploration
Workflow
FRONT END
WEB-APPL. GUI
DATA MINING MODELS
MODEL LIBRARY RUN
FRAMEWORK
WEB-SERVICE Suite CTRL
DMPlugin
MLP
REGISTRY & DATABASE
USER & EXPERIMENT INFORMATION
DRIVER
FILESYSTEM & HARDWARE I/F
Stand Alone
GRID
CLOUD
USER INFO
USER SESSIONS
USER EXPERIMENTS

FRONT END Component

Architecture:
- Client-server AJAX (Asynchronous JAva-Xml) based;

Technology:
- GWT-EXT;

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;

GWT

The **Google Web Toolkit** is an open source toolkit to create client-side applications in Java. GWT compiler translates a Java application into equivalent JavaScript that manipulates a web browser. GWT emphasizes reusable, efficient solutions to asynchronous remote procedure calls, history management and cross-browser portability.

EXT

Ext is an open-source JavaScript library, for building richly interactive web applications using techniques such as AJAX scripting. Ext JS is an excellent framework for building web applications that have desktop-like functionality in a web browser.

GWT-EXT

GWT-EXT is a library integrating GWT and EXT. One of the primary goals is to make the GWT-Ext widgets and API's work seamlessly with the core GWT infrastructure and its API's

DR Component

Architecture:

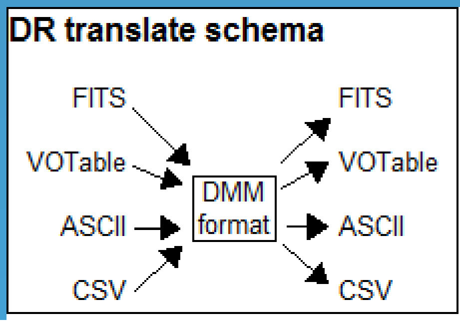
- It depends on the environment choice;
- In S.Co.P.E. DR is a component running on the GRID UI;

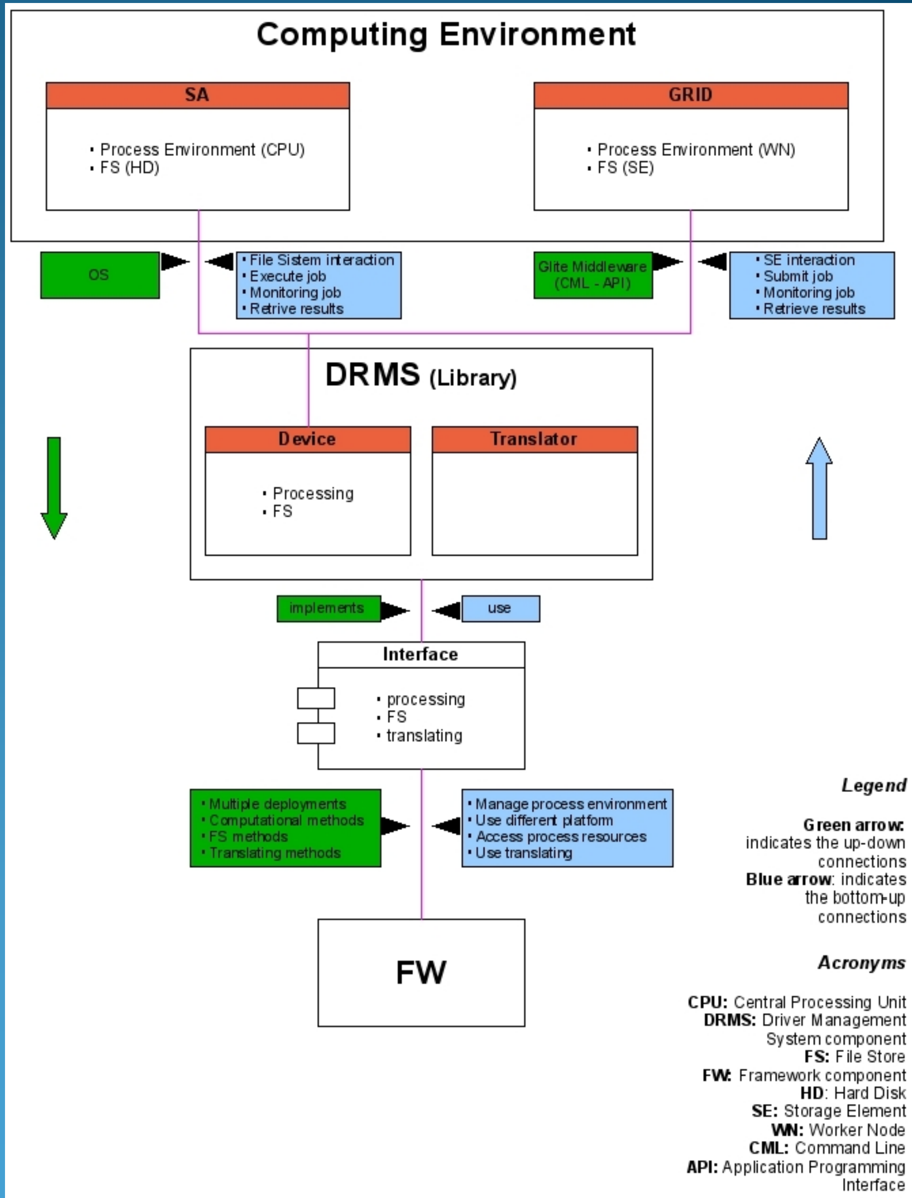
Technology (in S.Co.P.E.):

- 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 functionality class hierachy;

Technology:

• available model packages and libraries;

• custom ad hoc model design and development;

• custom wrappers for internal standardization;

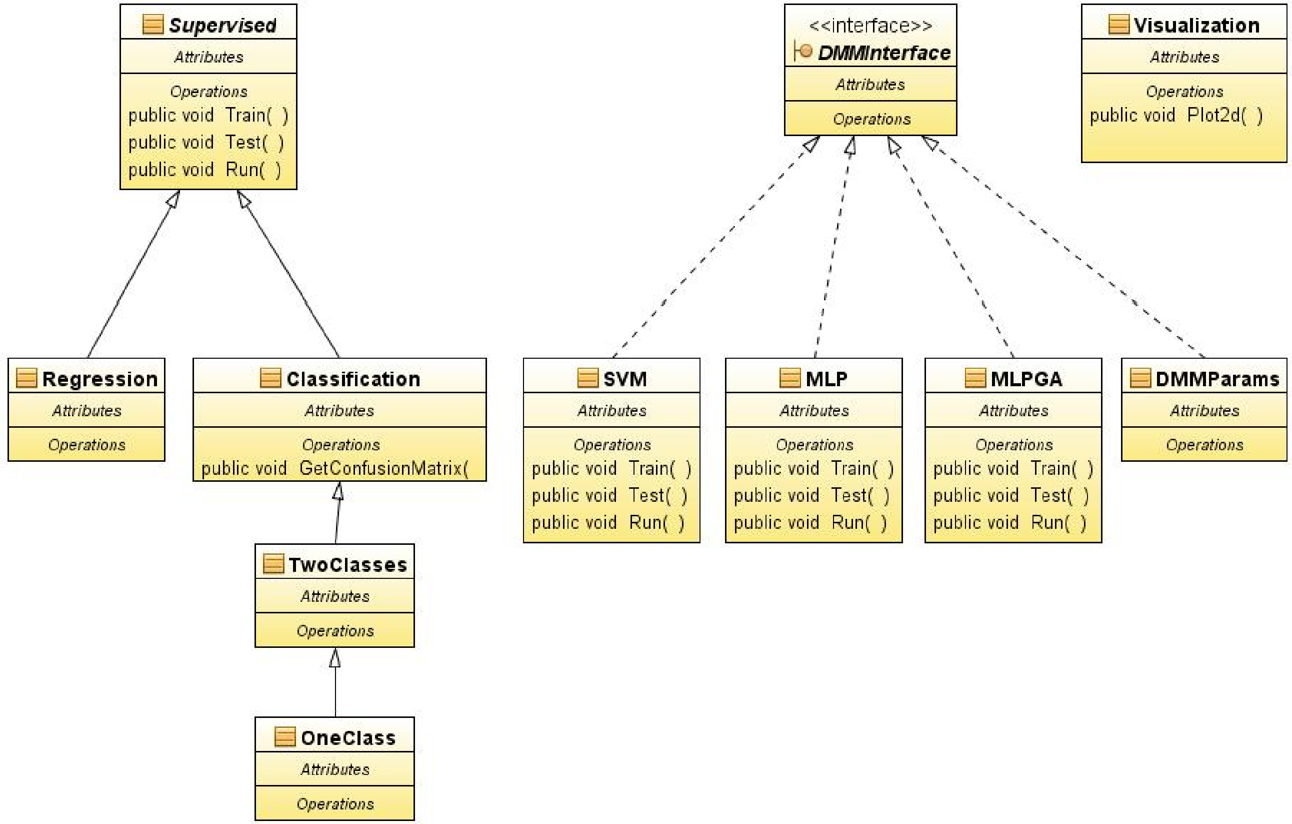
Features:

• modularity;

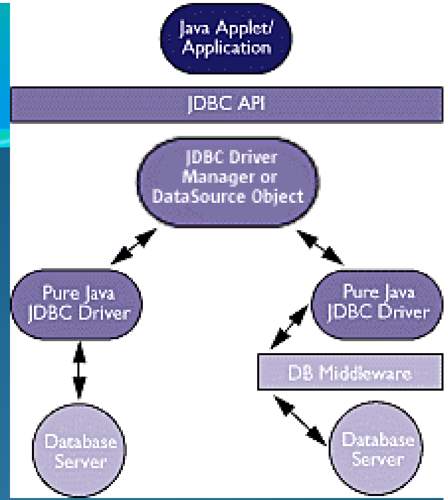
• fast third part application integration;

• functionality specialization;

• multi-language programming support;



REDB Component



Architecture:

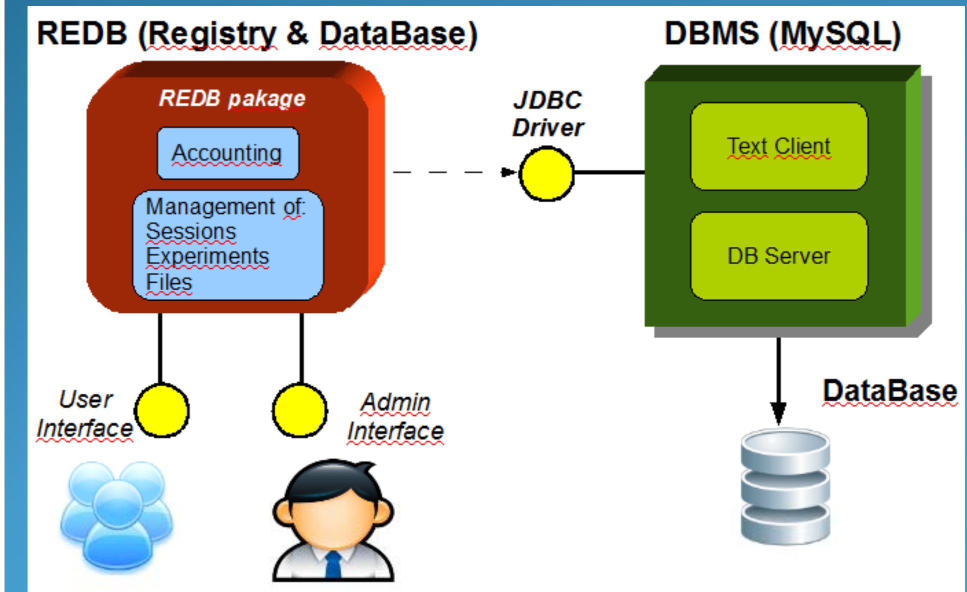
- JDBC;

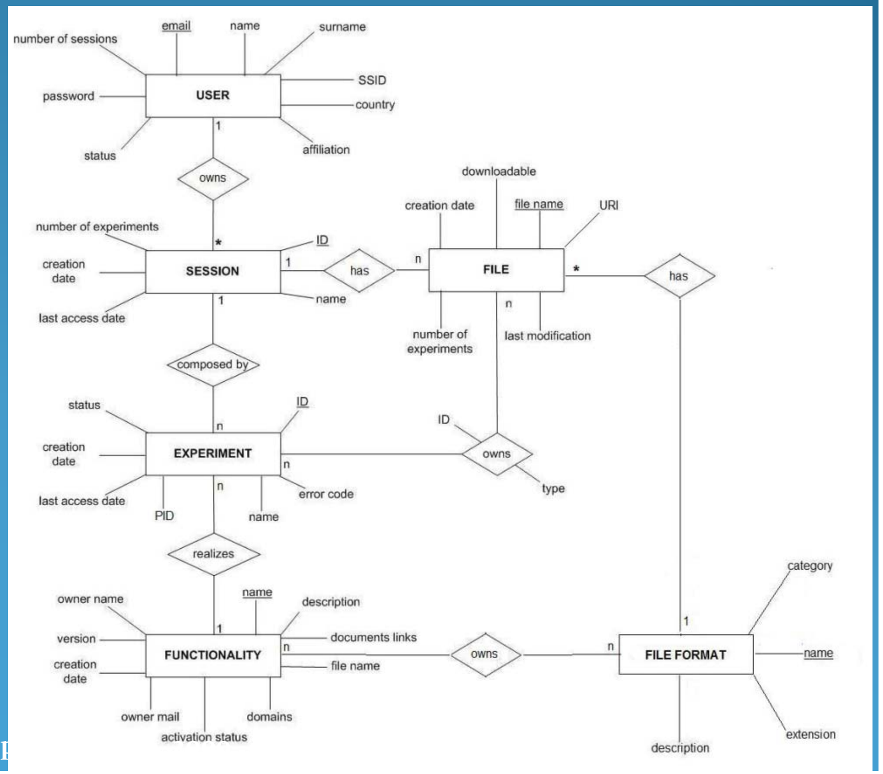
Technology (in S.Co.P.E.):

- MySQL and JDBC API;

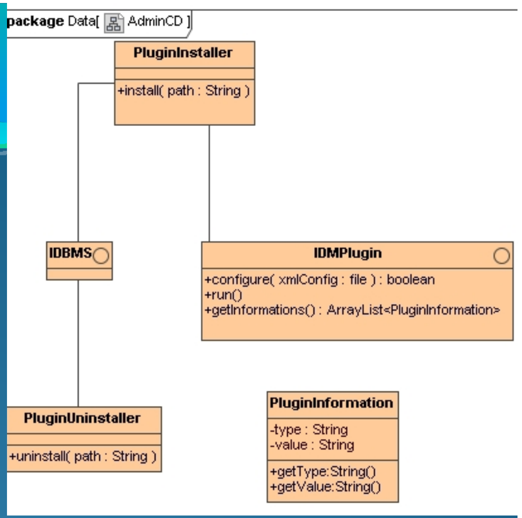
Features:

- 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);



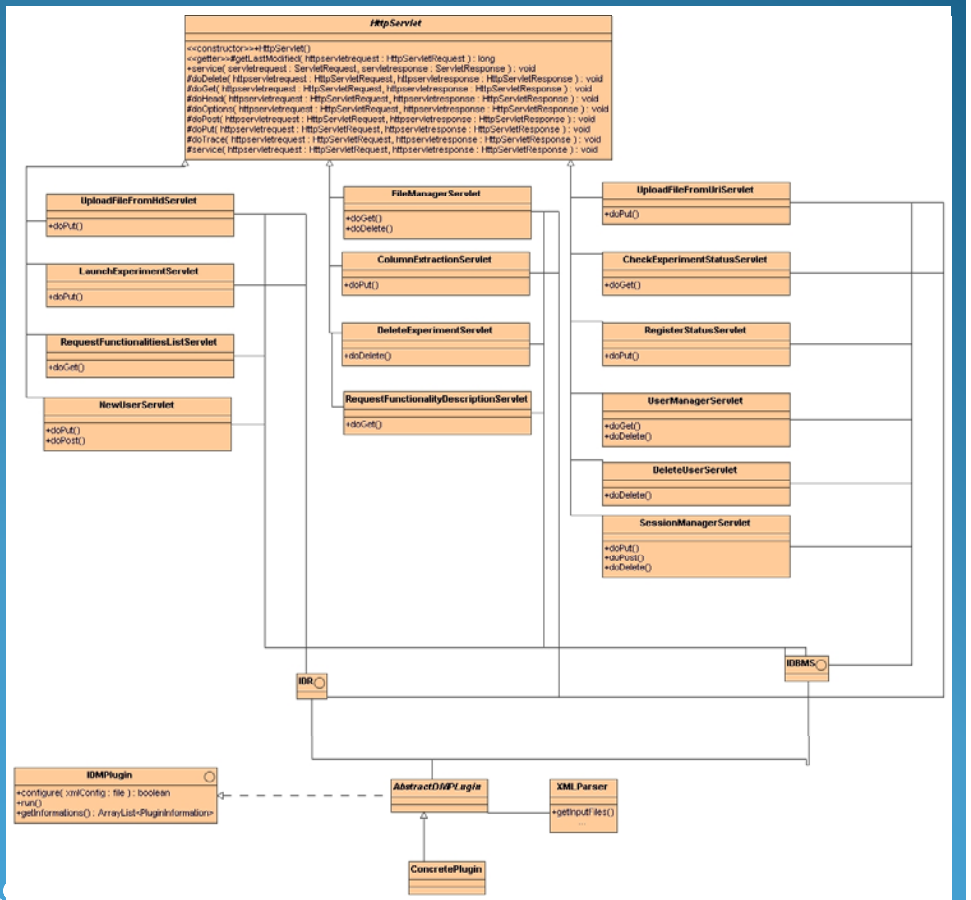


FRAMEWORK Component



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;
- XML based internal communication protocol

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;

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;

VO-Neural / Data Mining Exploration

Application Prototype

voneural.na.infn.it


ri Strumenti ?

http://voneural.na.infn.it/

ozdev.org ASTRONOMIA

sitemap page | https://imap-ac.na.i....it/src/webmail.php | VO-Neural Home Page | Dame - DAta Mining and Exploration

VO-Neural

VO-Neural / DAME

Home | Project | Documents | Download | Contact Us

Deliverables

VO-Neural/DAME software

Science Cases

Photometric redshift for SDSS galaxies

Shakhbazian groups in the SDSS

QSO candidates in the SDSS

Search

Private Area

VO-Neural Project

VO-Neural, an evolution of the former AstroNeural Collaboration, is a part of the European project VOTECH (Virtual Observatory Technological Infrastructures) and of the Italian PON-S.Co.P.E.

VO-Neural/DAME (Virtual Observatory-Neural / DAta Mining and Exploration) consists of an Information Technology project for design and development of instruments and tools for scientific data mining.

VO-Neural/DAME members

Partnership:

- Dipartimento di Fisica (sezione di Astrofisica) - Università degli Studi di Napoli Federico II
- INAF - Osservatorio Astronomico di Capodimonte
- California Institute of Technology, Pasadena - USA

Collaborations:

- VOTECH (Virtual Observatory Technological Infrastructures)
- S.Co.P.E. (high Performance distributed Cooperative System for scientific Experiment)
- INAF - Osservatorio Astronomico di Trieste
- Dipartimento di Informatica - Università degli Studi di Napoli Federico II
- Dipartimento di Ingegneria Informatica - Università degli Studi di Napoli Federico II
- MIUR (Italian Ministry of ResearcH)
- EURO-VO The European Virtual Observatory
- IVOA (International Virtual Observatory Alliance)

Data coming from the astronomical observations of the Universe is gathered by a very large number of techniques and stored in very diversified and often incompatible data repositories. Moreover in the e-science environment, we need to integrate services across distributed, heterogeneous, dynamic "virtual organizations" formed from the different resources within a single enterprise and from external resource sharing and service provider relationships.

The VO-Neural/DAME project aims at creating a single distributed e-infrastructure. It provides integrated access to astrophysical data collected by very different instruments, experiments and scientific communities in order to be able to correlate them and improve their scientific usability.

The project consists of a data mining framework whose main goal is to provide the astronomical community with powerful software instruments to work on massive data sets in a distributed computing environment, matching the international VO standards and requirements. The process of integration needed to achieve a specific quality of the data processing service, when running on top of different native platforms, can be technically challenging.

The VO-Neural/DAME project effort is a service-oriented architecture, by using appropriate standards and incorporating GRID paradigms and restful web-service frameworks where

Application Prototype

New user registration

DAME - DAta Mining and Exploration

- Home
- Sign Up!
- Help & Tutorials
- The Team

Create an account

First name:

Last name:

Username:

Email address:

Password:

Password again

Click when finished: Register →

Fill out the form to the left (all fields are required), and your account will be created; you'll be sent an email with instructions on how to finish your registration.

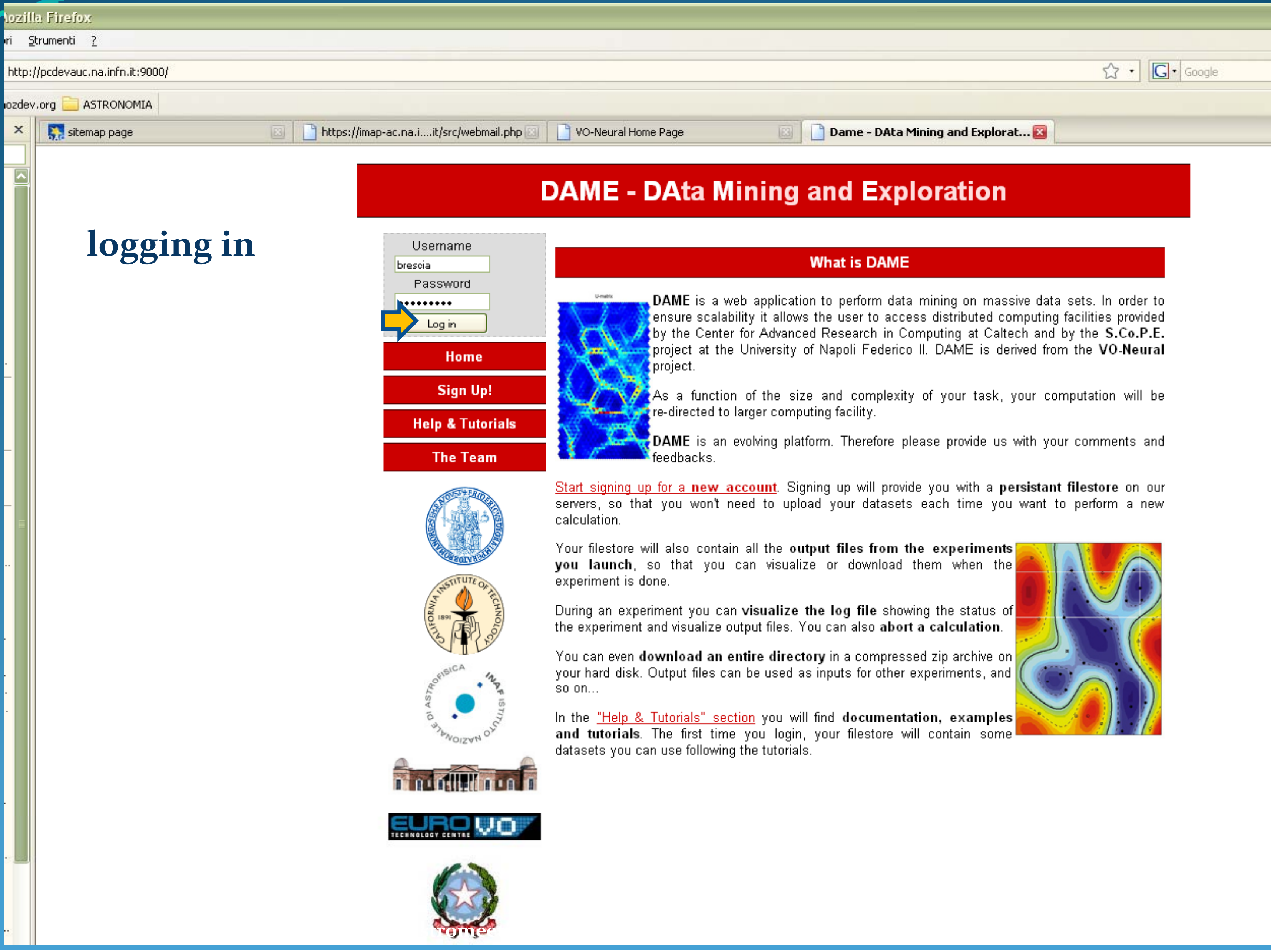
We'll only use your email to send you signup instructions. We hate spam as much as you do.

This account will let you subscribe to event streams for future notifications.

VO-Neural / Data Mining Exploration

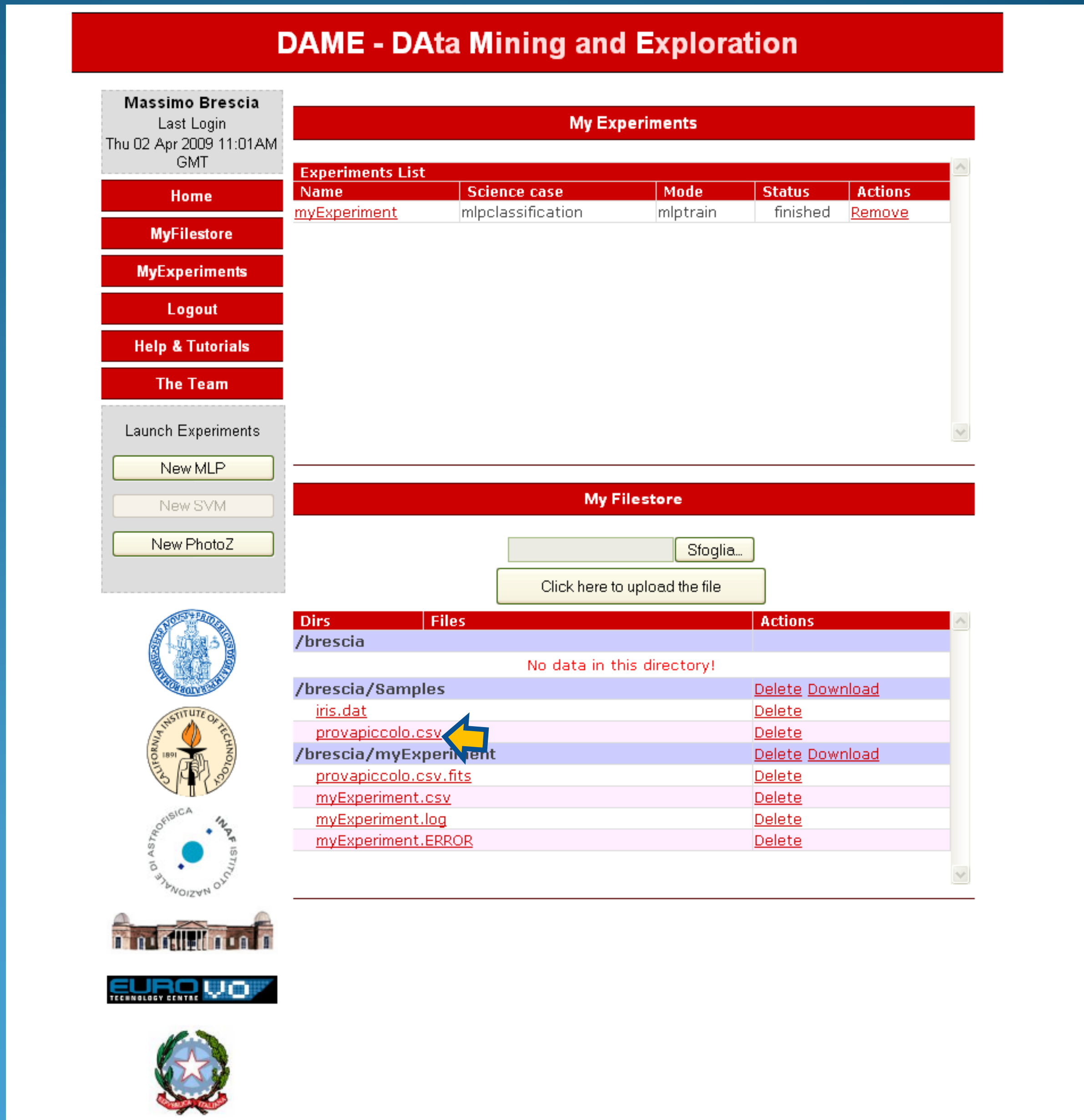
Application Prototype

logging in

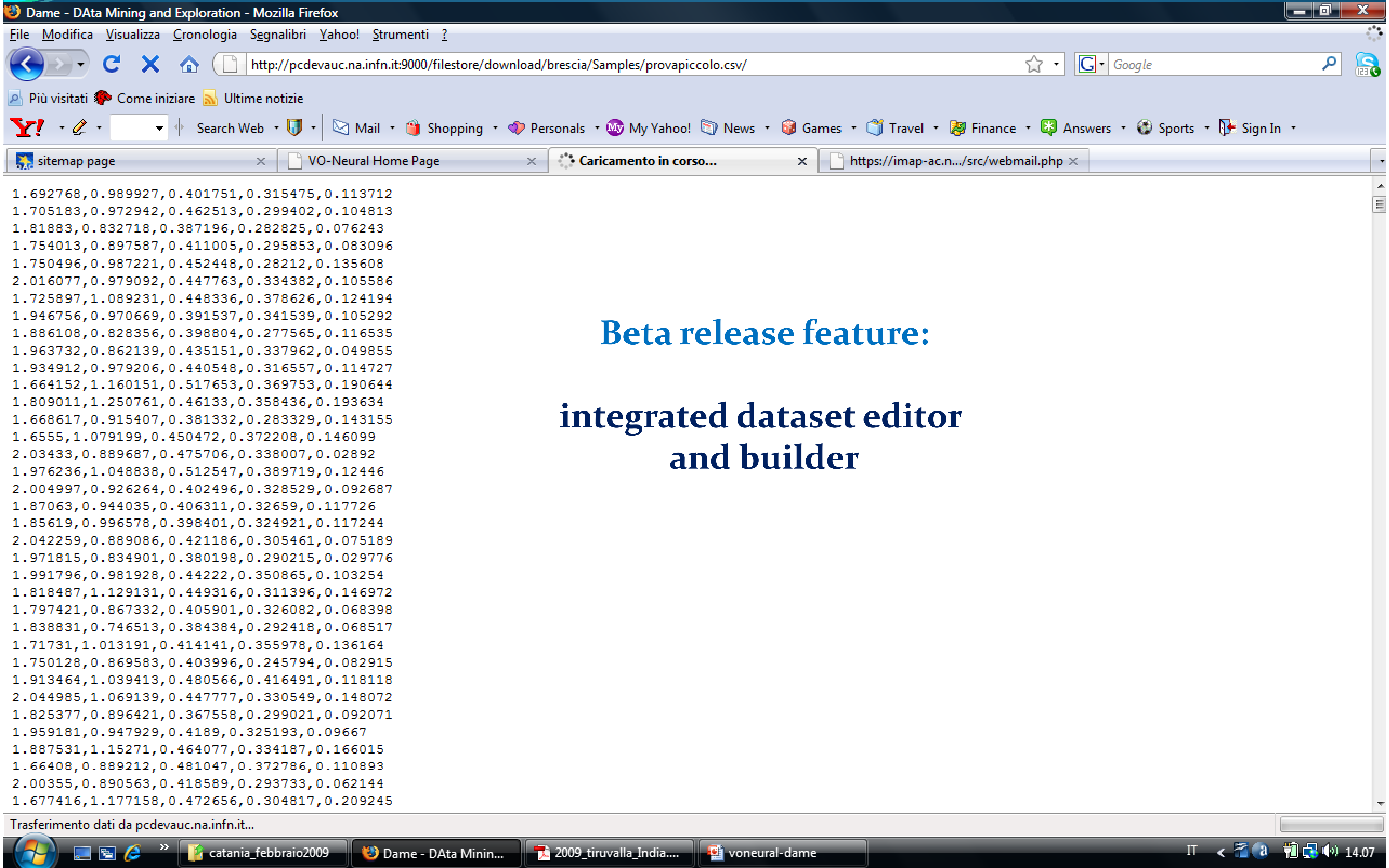


Application Prototype

Show input dataset



Application Prototype

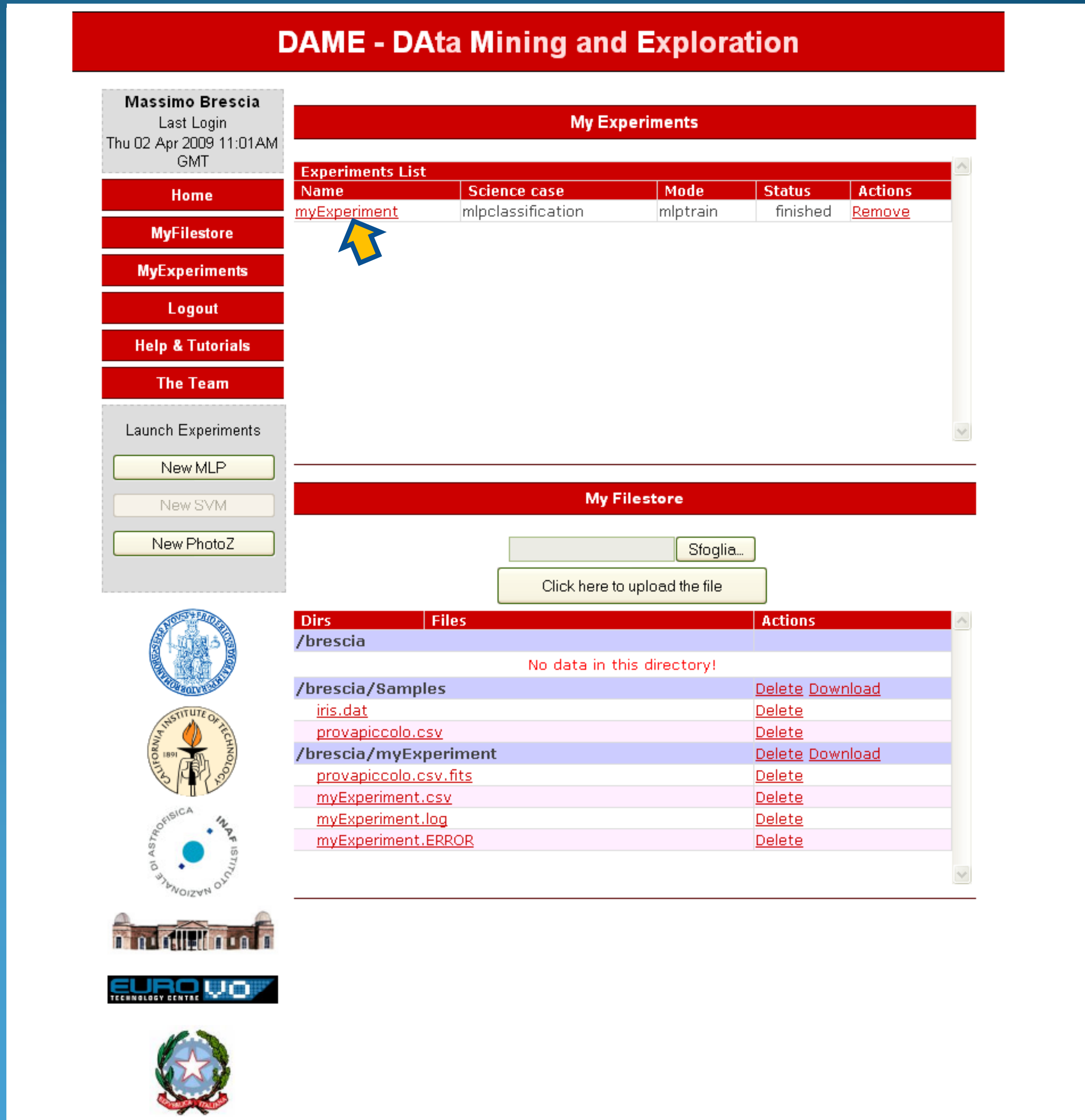


Beta release feature:

integrated dataset editor and builder

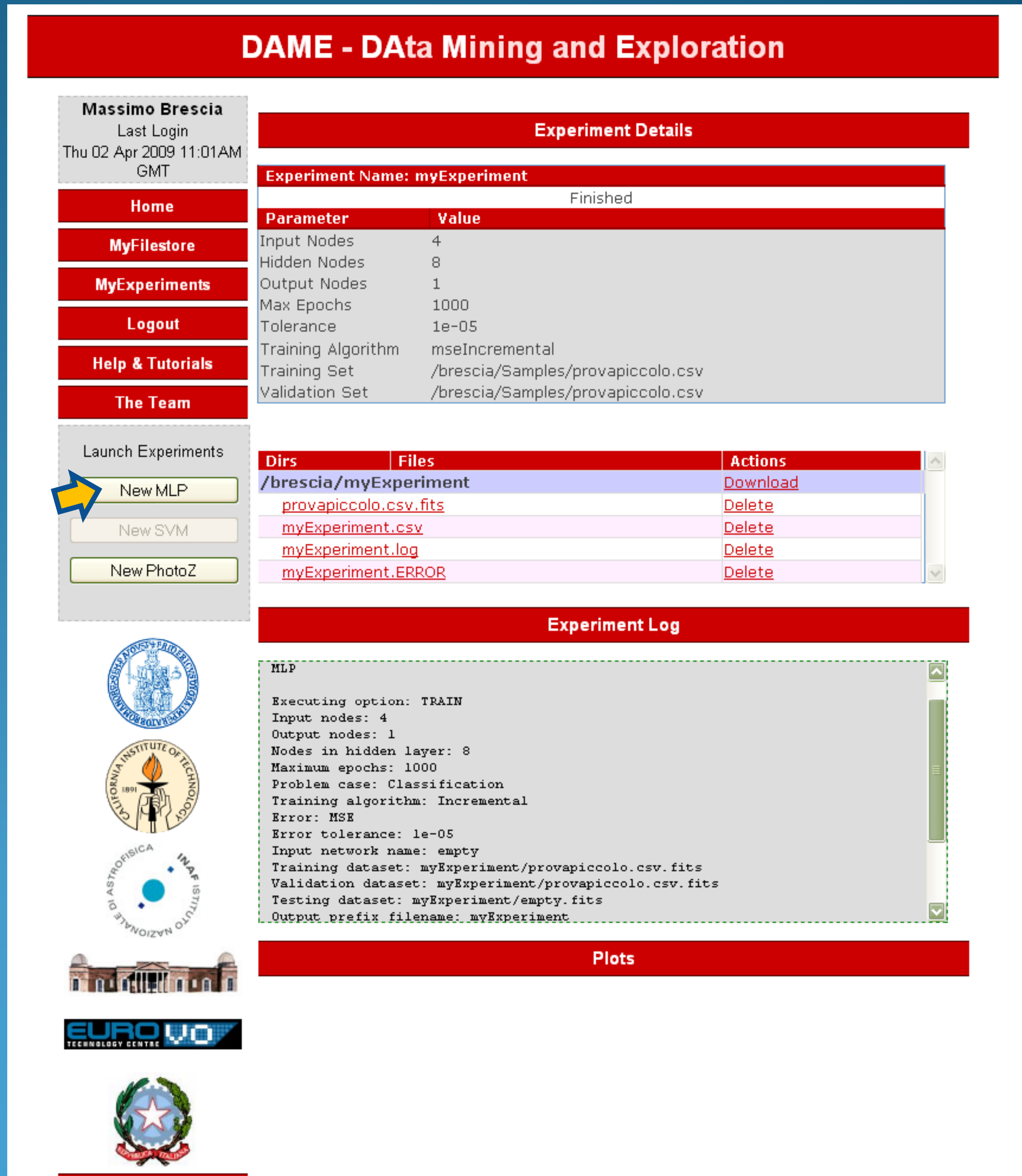
Application Prototype

Check /Edit past experiments

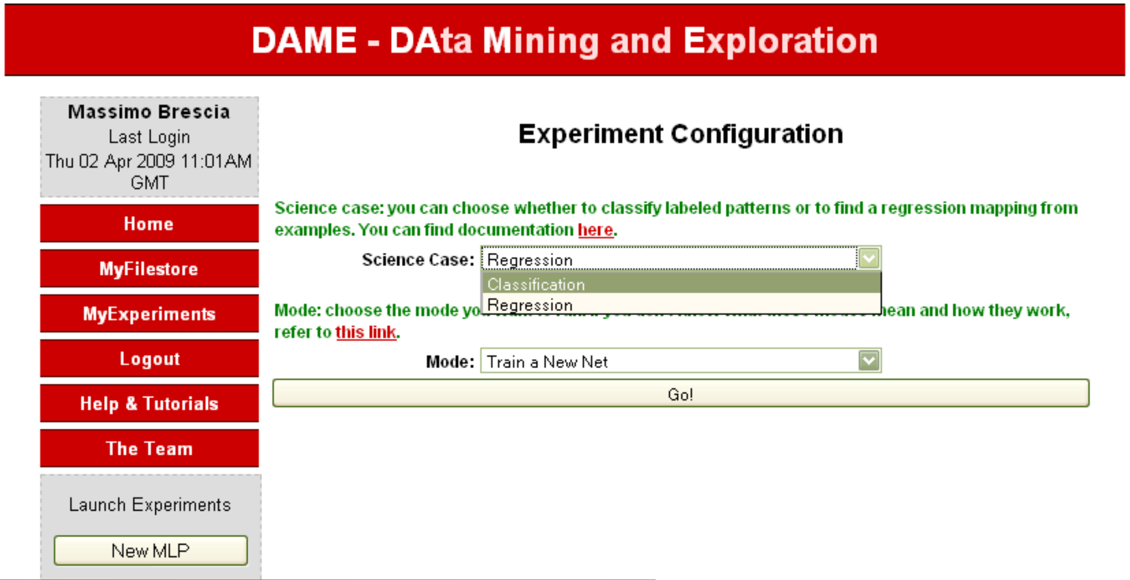


Application Prototype

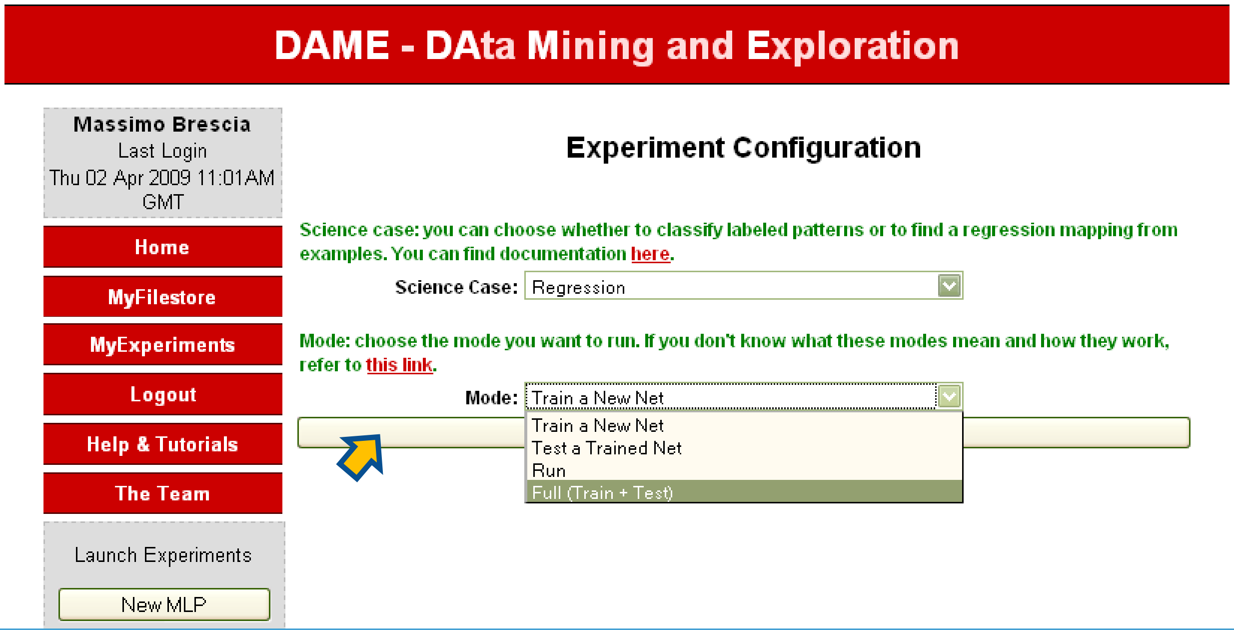
Launch new experiments



Application Prototype



Select functionality



Select use case and launch the job

Application Prototype

Edit and submit model and experiment parameters

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

- New MLP
- New SVM
- New PhotoZ

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: iris_exp

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: 3

Output Nodes Help

Output nodes: 3

Max epochs: 40000

Tolerance: 1e-05

Training algorithm: MSE - BATCH

Resume training: ☐

Network: /Samples/iris.dat

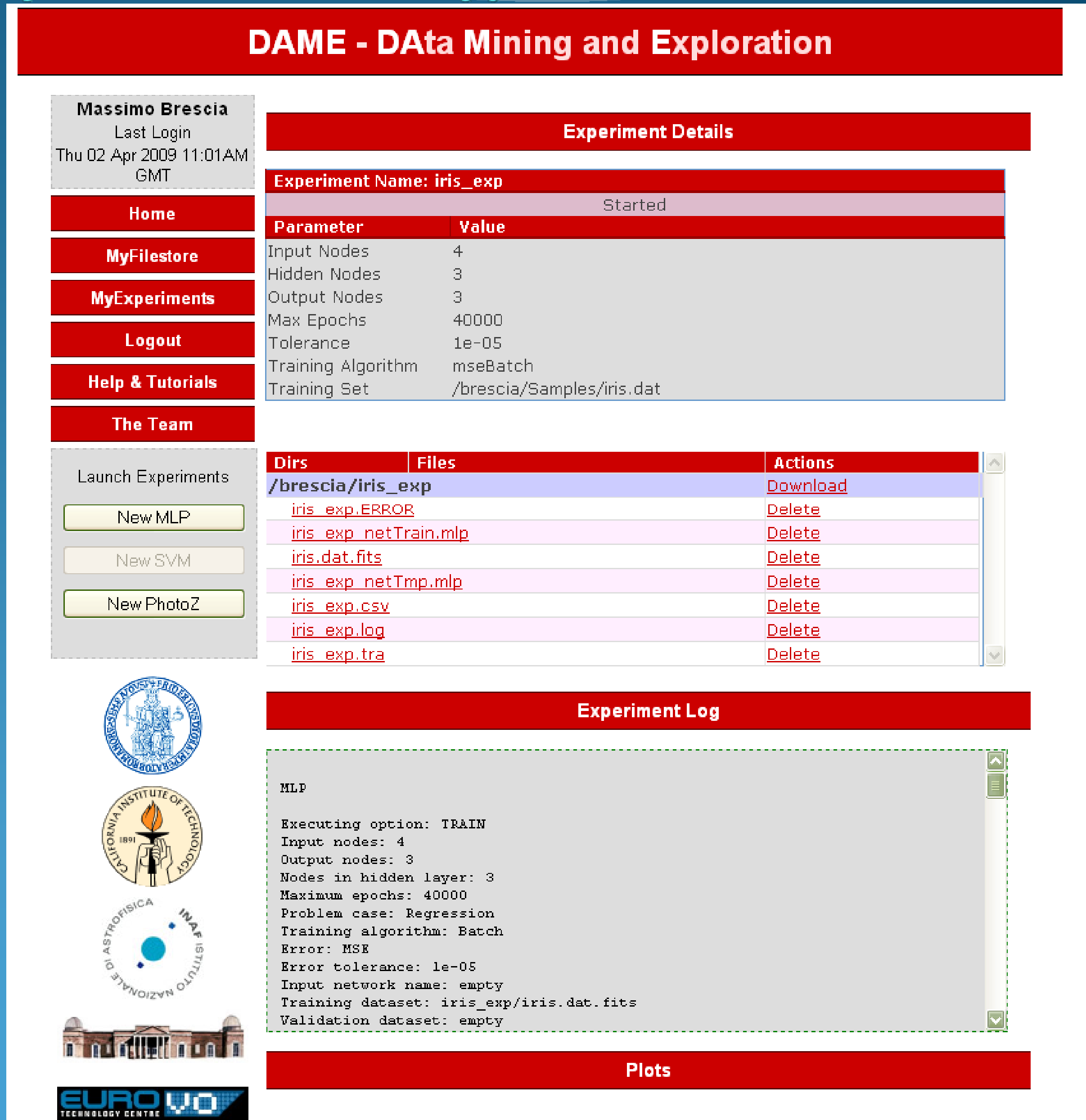
Training set: /Samples/iris.dat

Do validation: ☐

Validation set: /Samples/iris.dat

Go!

Application Prototype

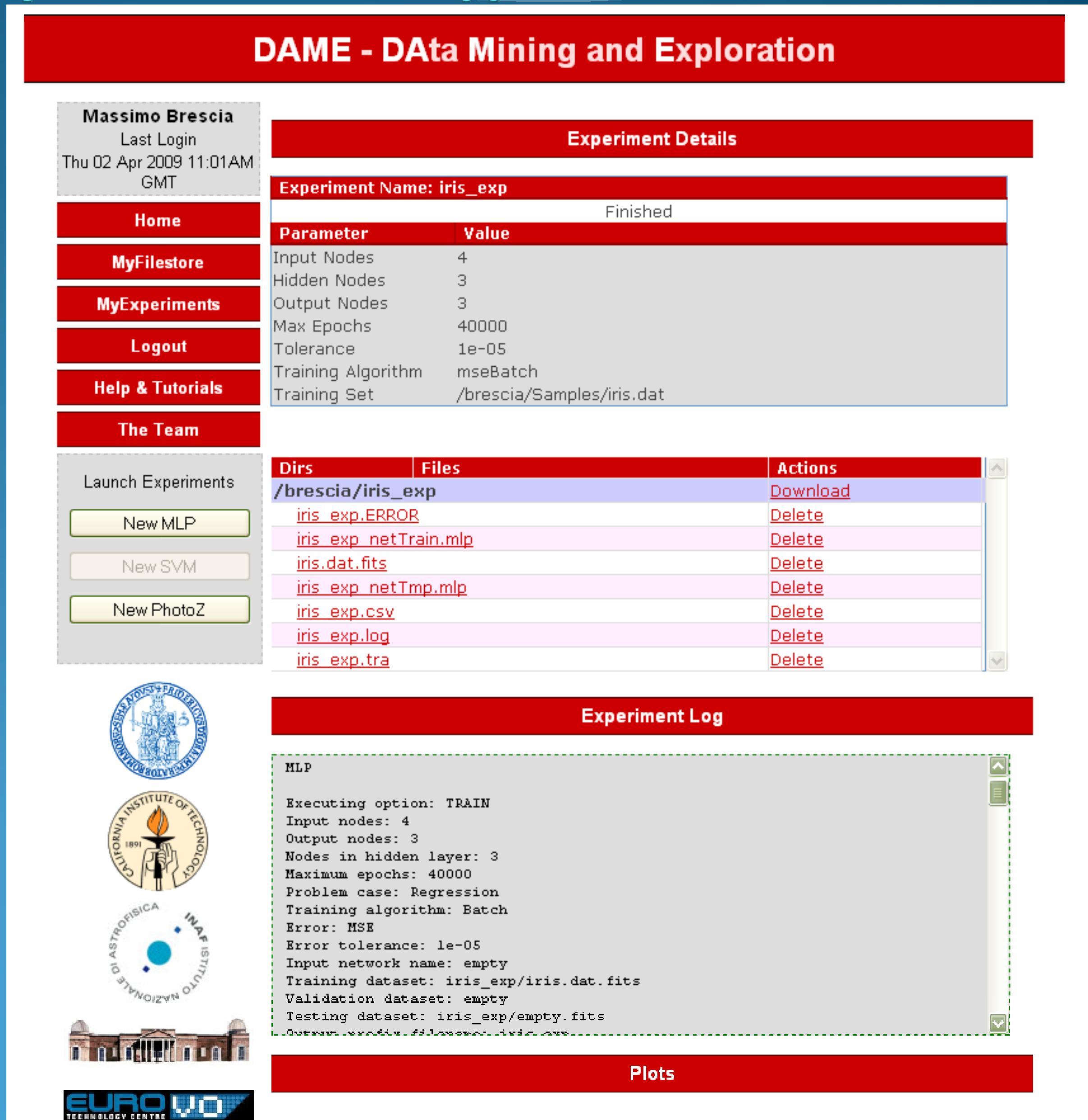


Status during execution

Beta release feature:

Interactive session optional during execution

Application Prototype



Status when finished

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

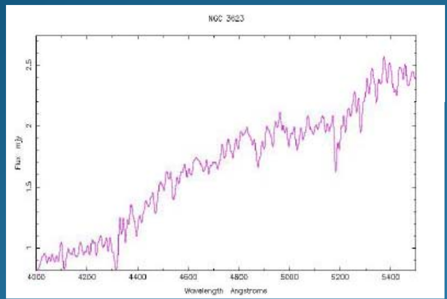
Cavuoti 2008, Thesis (VONeural website, voneural.na.infn.it)

More infos on WEB site documentation page

http://voneural.na.infn.it/documents.html

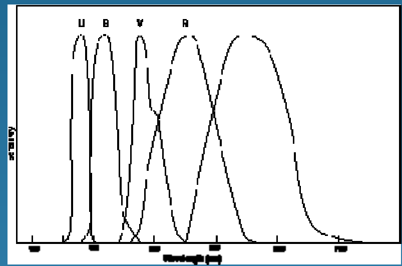
- **Photometric redshifts**
- QSO candidates selection

Science case: Mining the SDSS archive



Galaxy spectrum - F(λ)

X



Photometric system - $S_i(\lambda)$

=

$$m_U = -2.5\log_{10}\frac{\int F(\lambda)S_U(\lambda)d\lambda}{\int S_U(\lambda)d\lambda} + c_u$$

$$m_B = -2.5\log_{10}\frac{\int F(\lambda)S_B(\lambda)d\lambda}{\int S_B(\lambda)d\lambda} + c_B$$

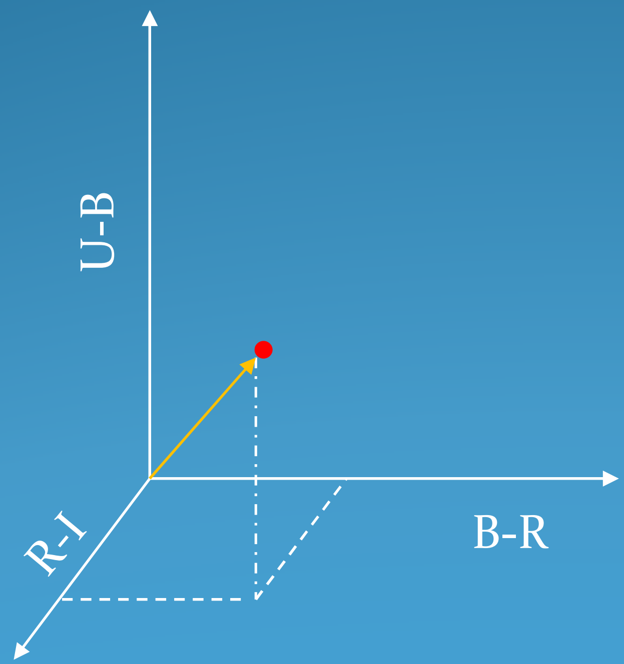
Etc...

Color indexes

$$U - B \equiv m_U - m_B$$

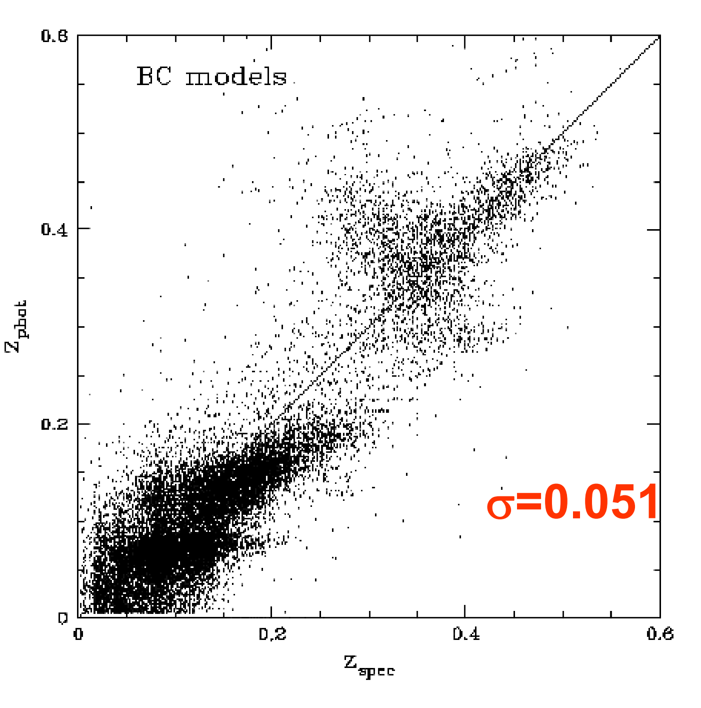
$$B - R \equiv m_B - m_R$$

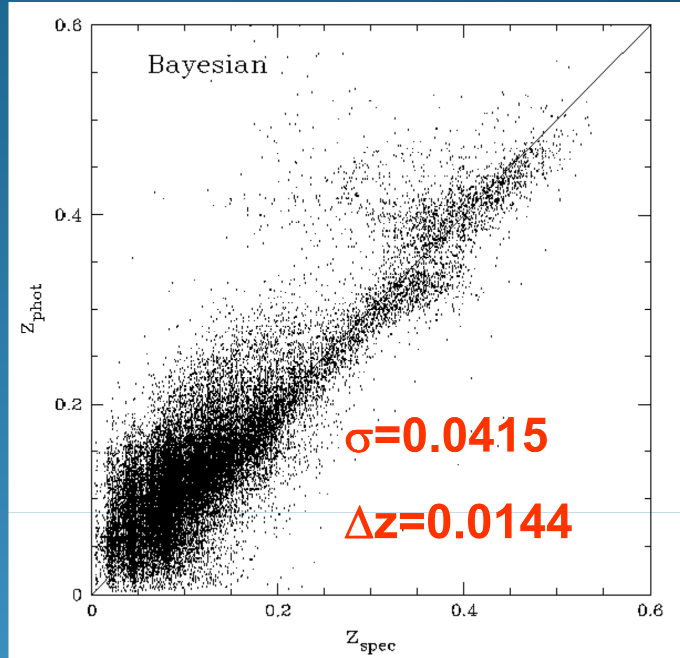
etc.



Photometric Redshifts

- **SED template fitting methods**
- Interpolative methods





type	method	data	Δz_{rms}	Notes	Reference
	CWW	EDR	0.0666		(Csabai et al. 2003)
SEDF	Bruzual-CHarlot	EDR	0.0552		(Csabai et al. 2003)
	Interpolated	EDR	0.0451		(Csabai et al. 2003)
	Polyomial	EDR	0.0318		(Csabai et al. 2003)
	KD-tree	EDR	0.0254		(Csabai et al. 2003)
	ANNz	EDR	0.0229		(Collister & Lahav 2004)
ML	SVM	EDR	0.027		(Wadadekar 2004)
ML	MLP-feed forward	SDSS-DR1 SDSS-RLG	xx.xxx	yes	(Vanzella et al. 2003)

Photometric Redshifts

- **SED template fitting methods**
- Interpolative methods

- the color space is partitioned (KD-tree - a binary search tree) into cells containing the same number of objects from the training set
- In each cell fit a second order polynomial

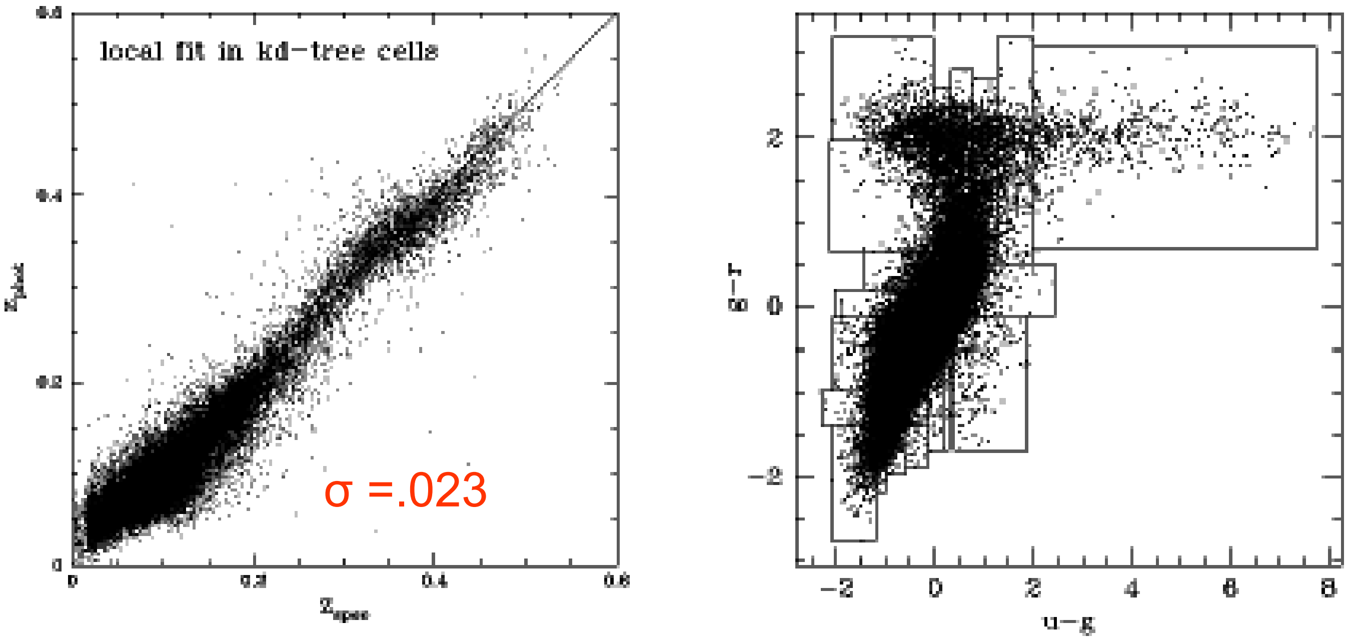
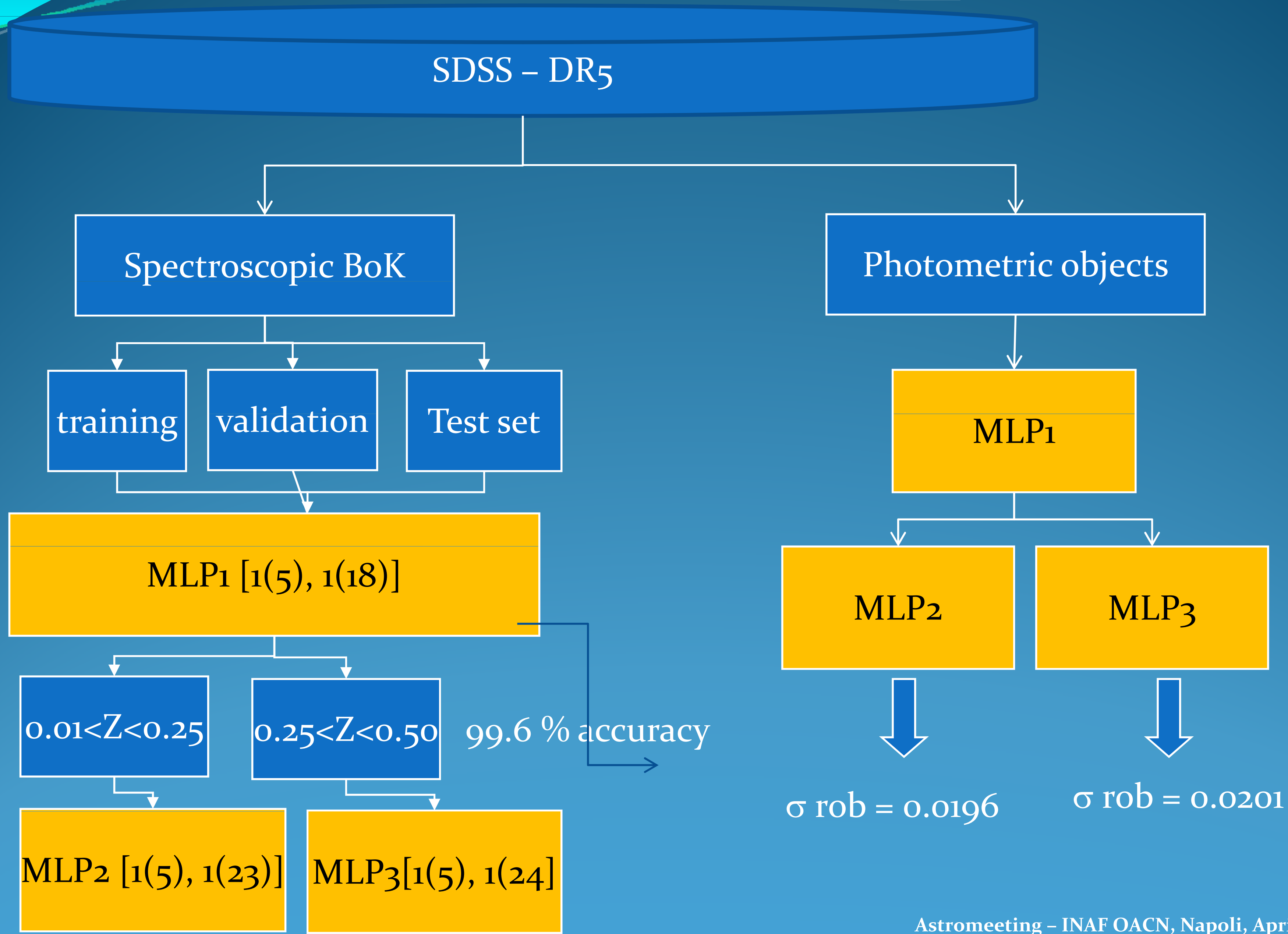
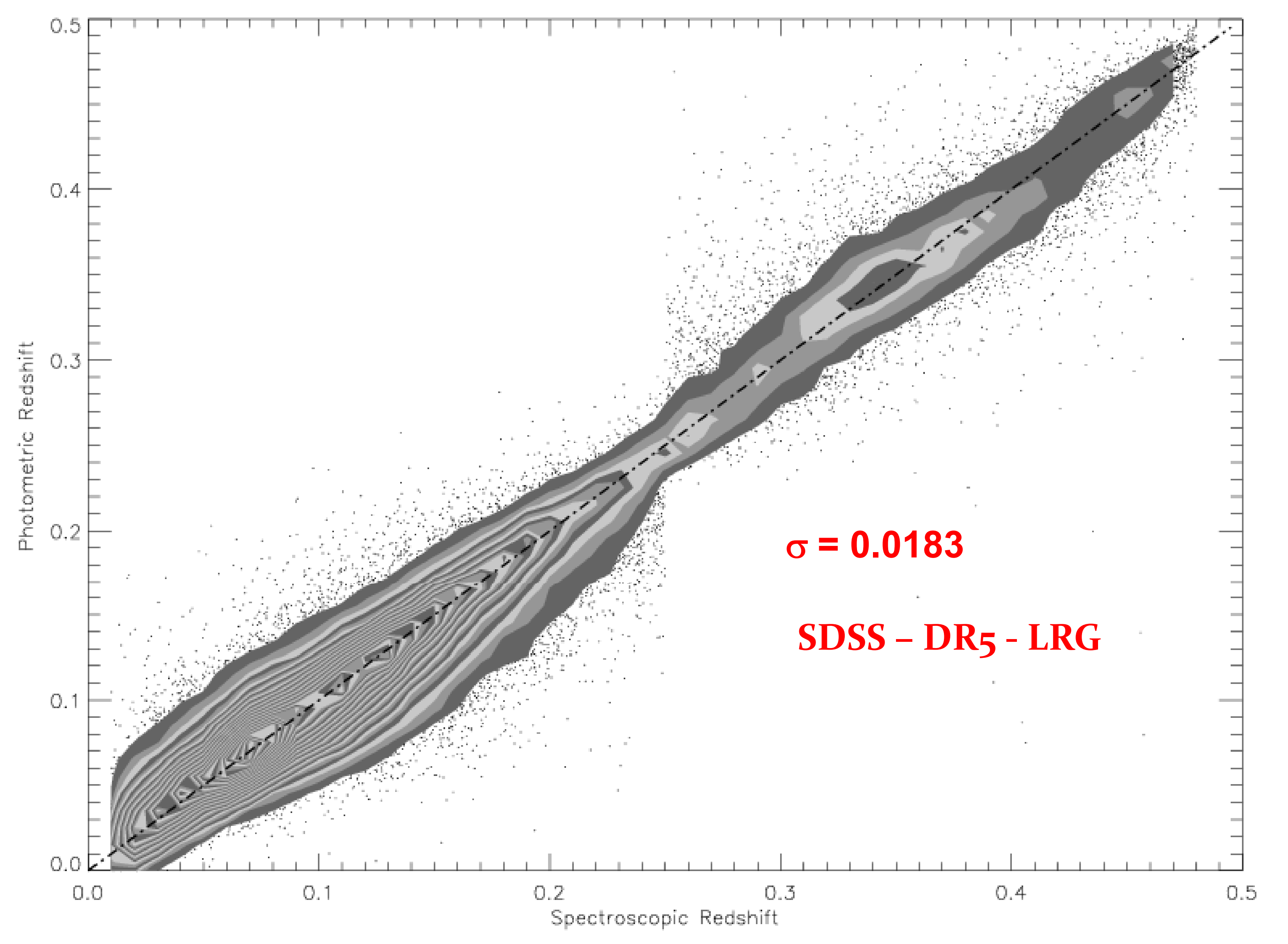


Fig. 4.— On the right we plot a 2 dimensional demonstration of the color space partitioning. In each of these cells we applied the polynomial fitting technique to estimate redshifts. The left figure show the results.

Science case: Photometric Redshifts



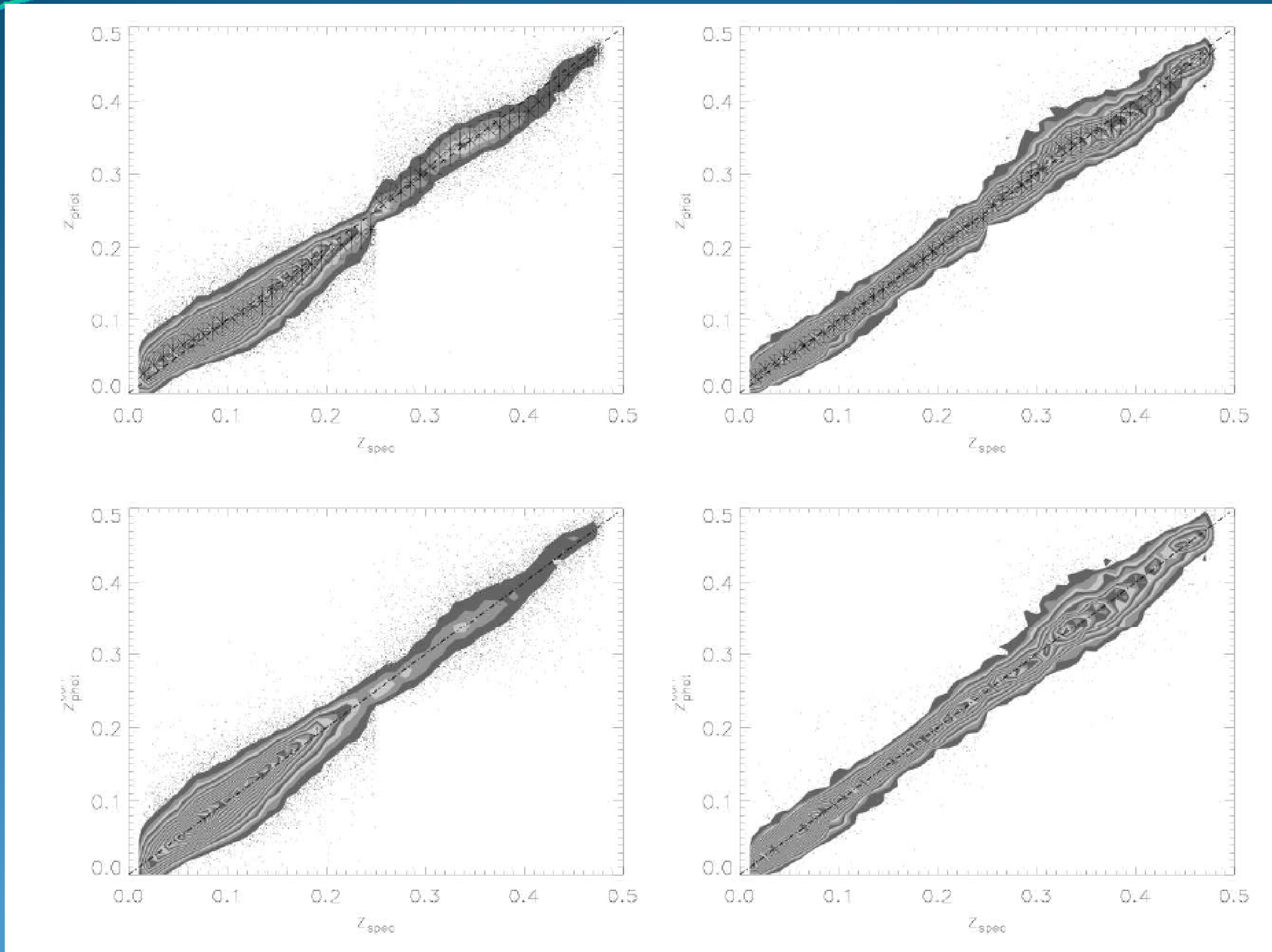
Science case: Photometric Redshifts



Science case: Photometric Redshifts

General galaxy sample

LRG sample



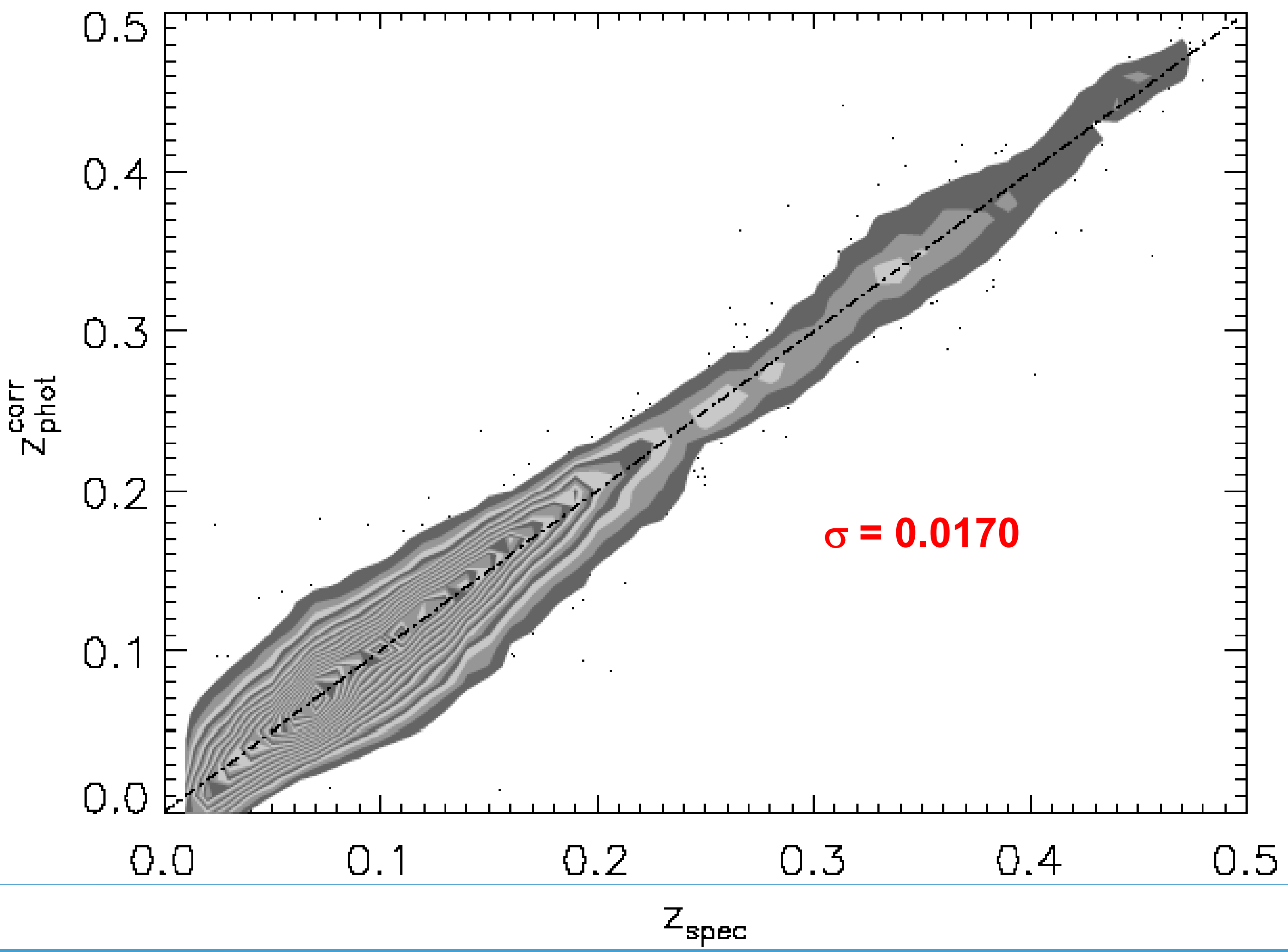
Non LRG only

σ = 0.0363
Δz = -0.0030

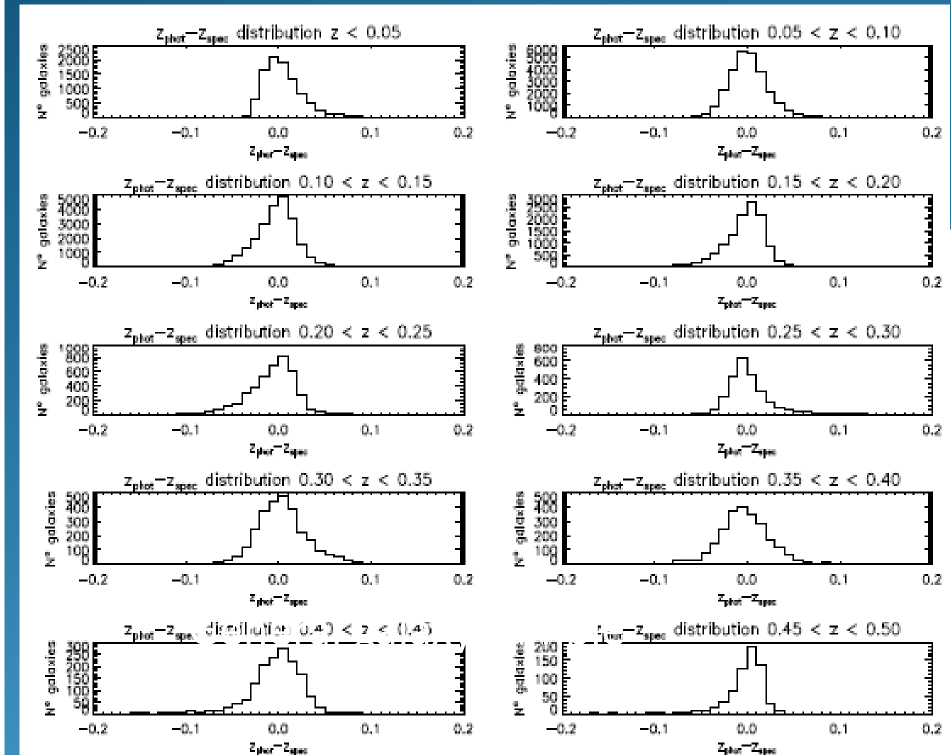
σ = 0.0208
Δz = -0.0029

σ = 0.0178
Δz = -0.0011

Science case: Photometric Redshifts



Science case: Photometric Redshifts



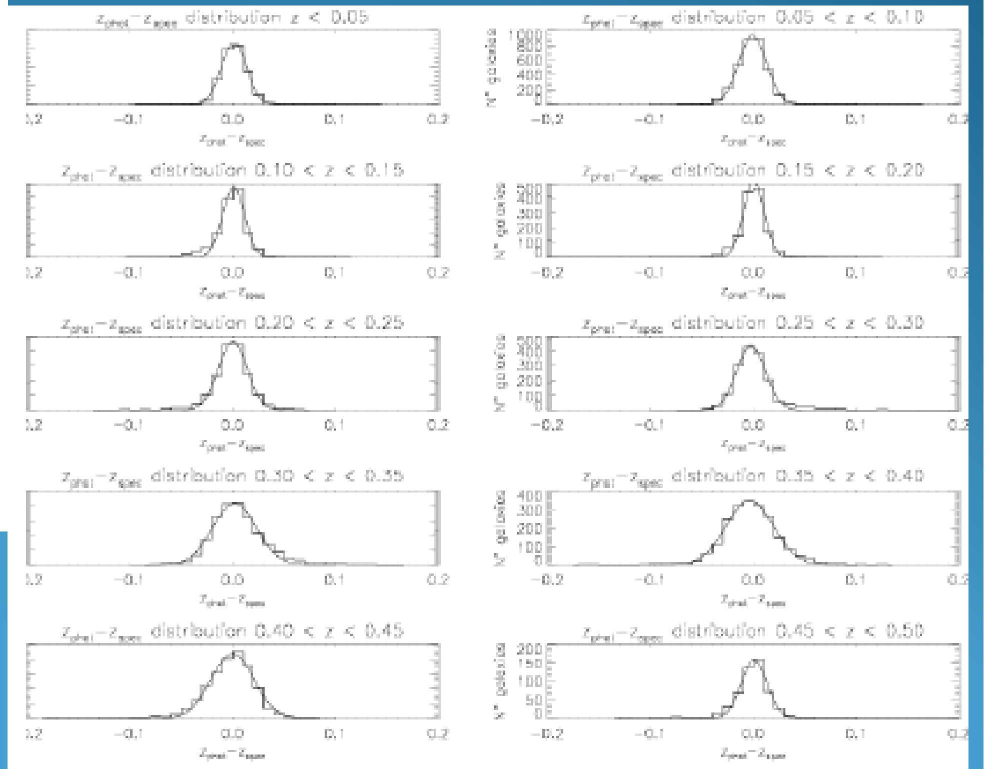


FIG. 9.— Same as in previous figure but for the LRG sample.

SDSS galaxies z_{phot}

- Generalization of the approach described in the previous paper (D'Abrusco et al. 2007) will be presented at AAS 2009
- K-means algorithm for clustering in the photometric parameter space is applied with an optimal number of clusters n_{opt}
- n_{opt} is chosen so that the "weighted accuracy" σ_W of the z_{phot} is maximum. Given N clusters with M_i elements each and rms of the (z_{phot}-z_{spec}) variable σ_i :

$$\sigma_W = M_{tot}^2 \sum_{i=1}^{N} \frac{\sigma_i}{M_i^2}$$

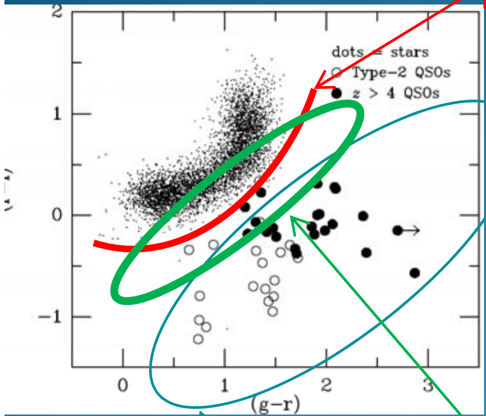
where

$$M_{tot} = \sum_{i=1}^{N} M_i$$

Searching for candidate quasars in the SDSS

Traditional way to look for candidate QSO in 3 band survey

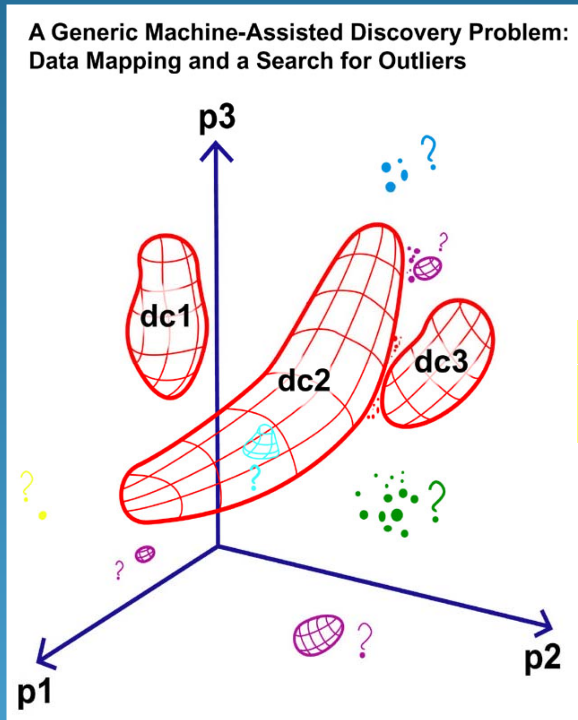
Cutoff line



Candidate QSOs for spectroscopic follow-up's

Ambiguity zone

In 4 bands degeneracy is partially removed



R. D'Abrusco

astro-ph/0805.0156v1

More are the bands the lower is the degeneracy

How to find the interesting regions (clusters)?

- Data Mining is the answer

How to visualize them ?

- Dimensionality reduction

Searching for candidate quasars in the SDSS

Several algorithms for "general purpose" photometric identification of candidate QSOs select sources according to different techniques exist.

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of QSO are radio-loud).
- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.
- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

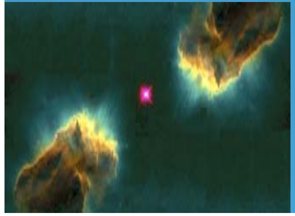
Overall performances of a generic targeting algorithm are usually expressed by two parameters:

Completeness

$$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

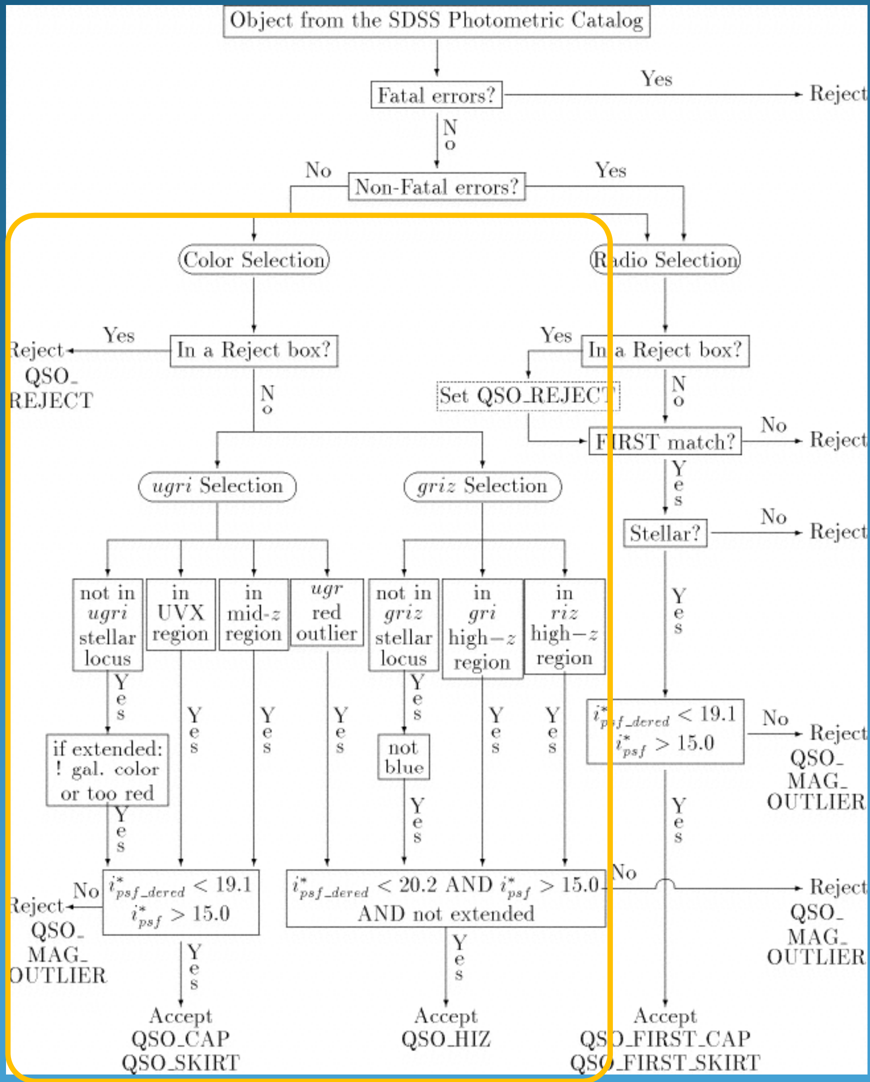
Efficiency

$$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$



SDSS QSOs targeting algorithm (I)

SDSS QSO candidate selection algorithm (Richards et al, 2002) targets star-like objects as QSO candidate according to their position in the SDSS colours space (u-g,g-r,r-i,i-z), if one of these requirements is satisfied:



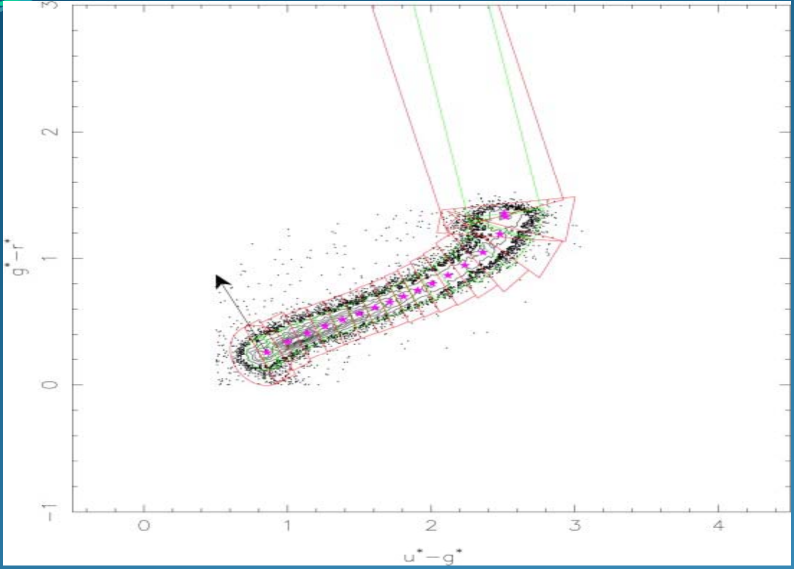
- **QSOs are supposed to be placed >4σ far from a cylindrical region containing the "stellar locus"** (S.L.), where σ depends on photometric errors.

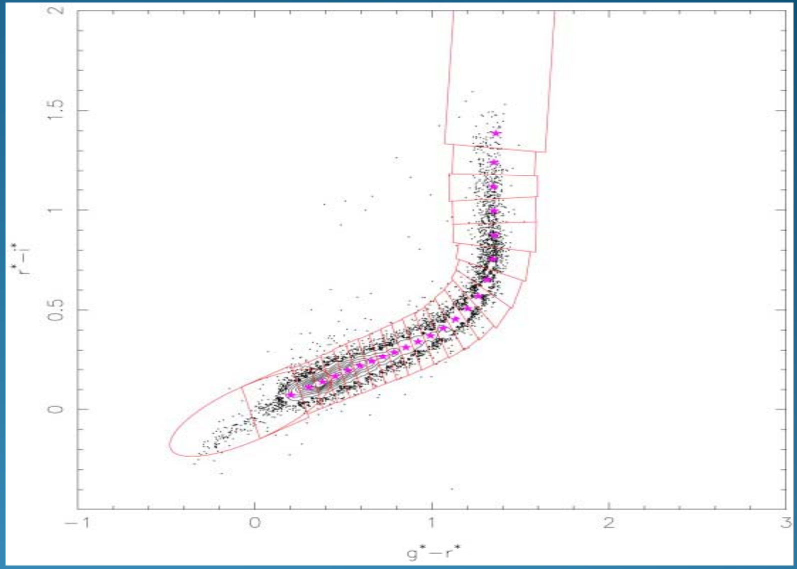
OR

- **QSOs are supposed to be placed inside the inclusion regions**, even if not meeting the previous requirement.

c = 95%, e = 65%
locally less

SDSS QSOs targeting algorithm (II)





1. **inclusion regions** are regions where S.L. meets QSO's area (due to absorption from Lyα forest entering the SDSS filters, which changes continuum power spectrum power law spectral index). All objects in these areas are selected so to sample the [2.2, 3.0] redshift range (where QSO density is also declining), but at the cost of a worse efficiency (Richards et al, 2001).

2. **exclusion regions** are those regions outside the main "stellar locus" clearly populated by stars only (usually WDs). All objects in these regions are discarded.

Overall performance of the algorithm: completeness c = 95%, efficiency e = 65%, but locally (in colours and redshift) much less.

Unsupervised clustering based on latent variable methods

Clustering method PPS

↓

labeling ← Spectroscopic BoK

↓

Cluster aggregation NEC ← Dissimilarity treshold

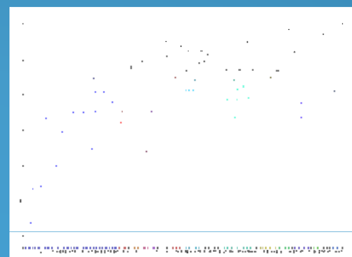
↓

Evaluation of successful clusters

↓

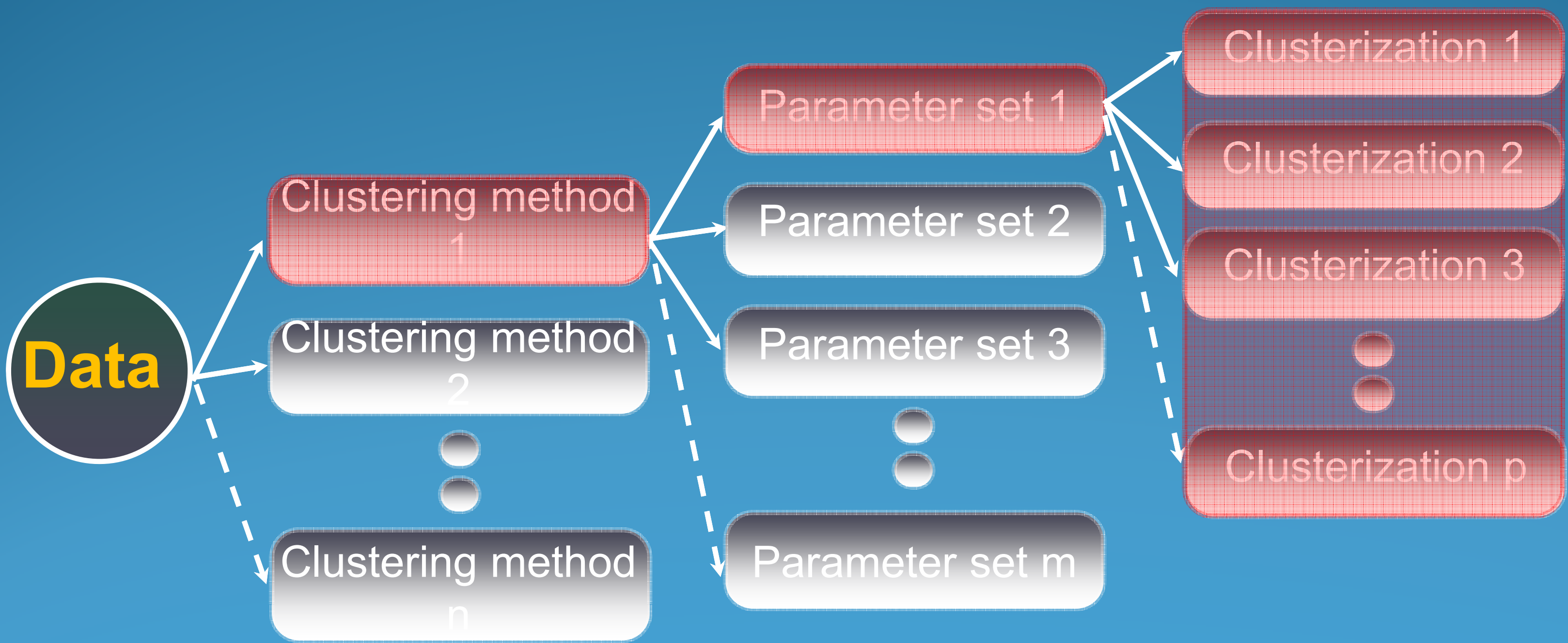
Partition of parameter space

1. **Plateau analysis**: final number of clusters N(D) is calculated over a large interval of D, and critical value(s) D_{th} are those for which a plateau is visible.
2. **Dendrogram analysis**: the stability threshold(s) D_{th} can be determined observing the number of branches at different levels of the graph.



Many experiments are required

1. **Pre-clustering algorithm:** this phase can be accomplished performing a reduction of dimension of the feature space; this reduction via feature extraction/selection can be supervised or unsupervised (our choice in unsupervised).
2. **Agglomerative clustering**: both distance definition and a linkage model (simple, average, complete, Wards, etc.) need to be provided to perform clustering.



Tuning succesfull clusters

Once partition of colours space is completed (as a function of D_{th}), **clusters mainly populated by QSO** (according the knowledge-base at our disposal) **are selected and information about these clusters are exploited for the candidate QSO selection.**

To determine the critical dissimilarity D_{th} threshold we rely not only on a stability requirement. Given the following definition:

$$\left[\text{cluster is “successfull”}\right] \overset{\text{Def}}{\Longleftrightarrow} \left[\begin{array}{c}\text{its fraction of confirmed QSO} \\ \text{is higher then a fixed value}\end{array}\right]$$

we ask D_{th} to maximize the **Normalized Success Ratio** (NSR):

$$NSR = \frac{\text{Number of successful clusters}}{\text{Number of total clusters}}$$

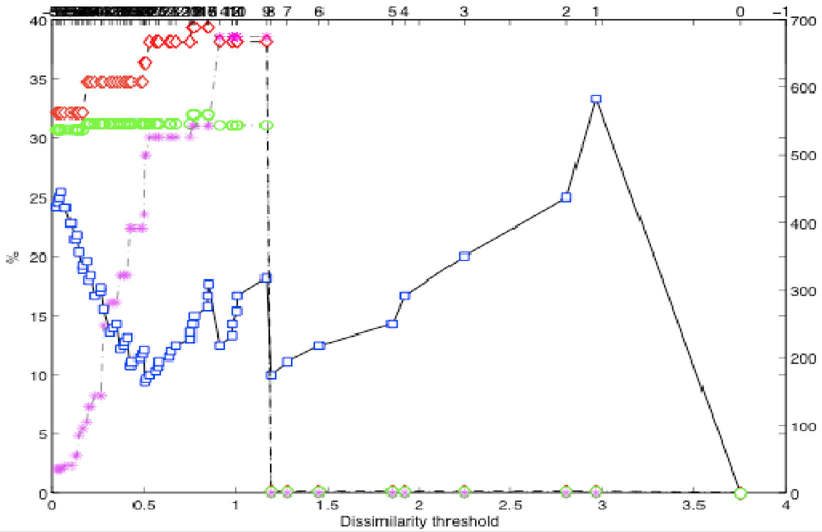
The process is recursive: feeding merged unsuccessful clusters in the clustering pipeline until no other successful clusters are found. The overall efficiency of the process e_{tot} is the sum of weighed efficiencies e_i for each generation:

$$e_{tot} = \sum_{i=1}^{n} e_i$$

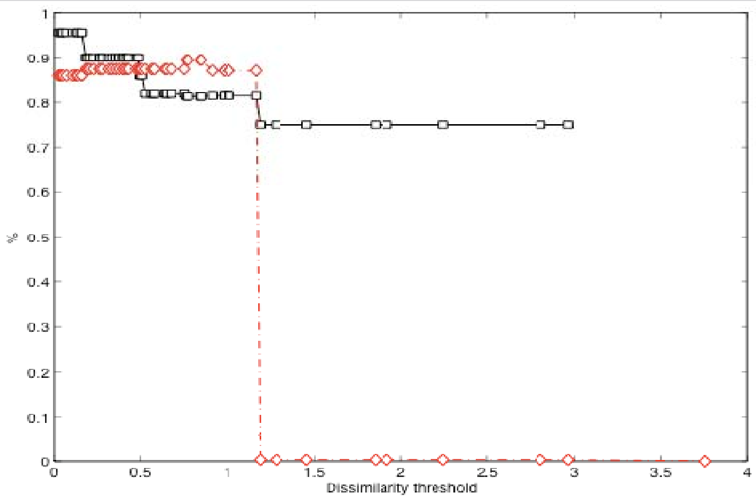
An example of "tuning"

Choice of the clustering

NSR



Efficiency and completeness



e and *c* estimation

To assess the reliability of the algorithm, the same objects used for the "training" phase have been re-processed using photometric informations only. Results have been compared to the BoK.

algorithm \ labels	QSOs	not QSOs
QSOs	759	72
not QSOs	83	1327

Data and experiments

Data samples:

1. **Optical**: sample derived from SDSS database table "Target" queried for QSO candidates, containing ~ $1.11 \cdot 10^5$ records and ~ $5.8 \cdot 10^4$ confirmed QSO ('specClass == 3 OR specClass == 4').

2. **Optical + NIR**: sample derived from positional matching ('best') between SDSS-DR3 database view "Star" queried for all objects with spectroscopic follow-up available and detection in all 5 bands (u,g,r,i,z) with high reliability for redshift estimation and line-fitting classification ('specClass') and high S/N photometry, and UKIDSS-DR1 star-like ('mergedClass == -1') objects fully detected in each of the four Survey bands (Y,J,H,K) and clean photometry

Experiments:

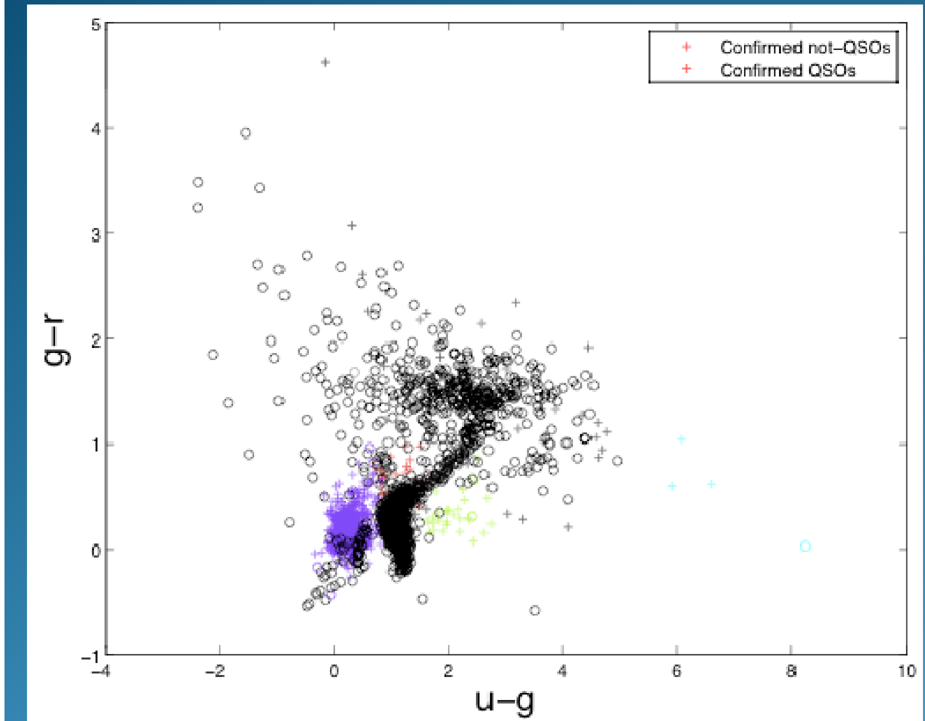
Optical (1)
candidate QSO
4 colours

Optical+NIR (2)
star-like objects
4 + 3 colours

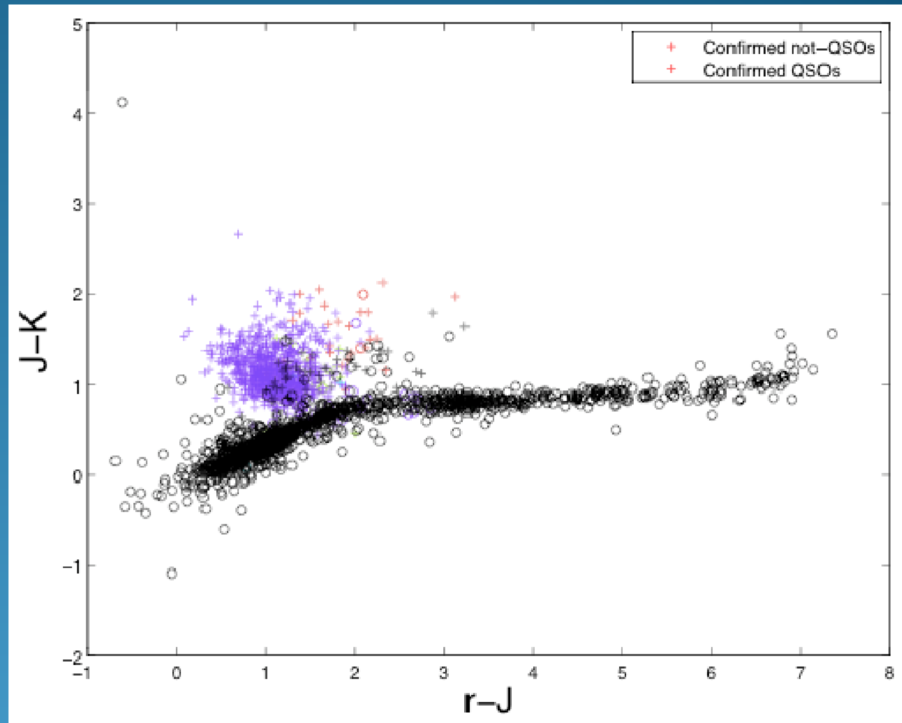
Optical (3)
star-like objects
4 colours

Experiment 2: SDSS ∩ UKIDSS

u - g vs **g - r**



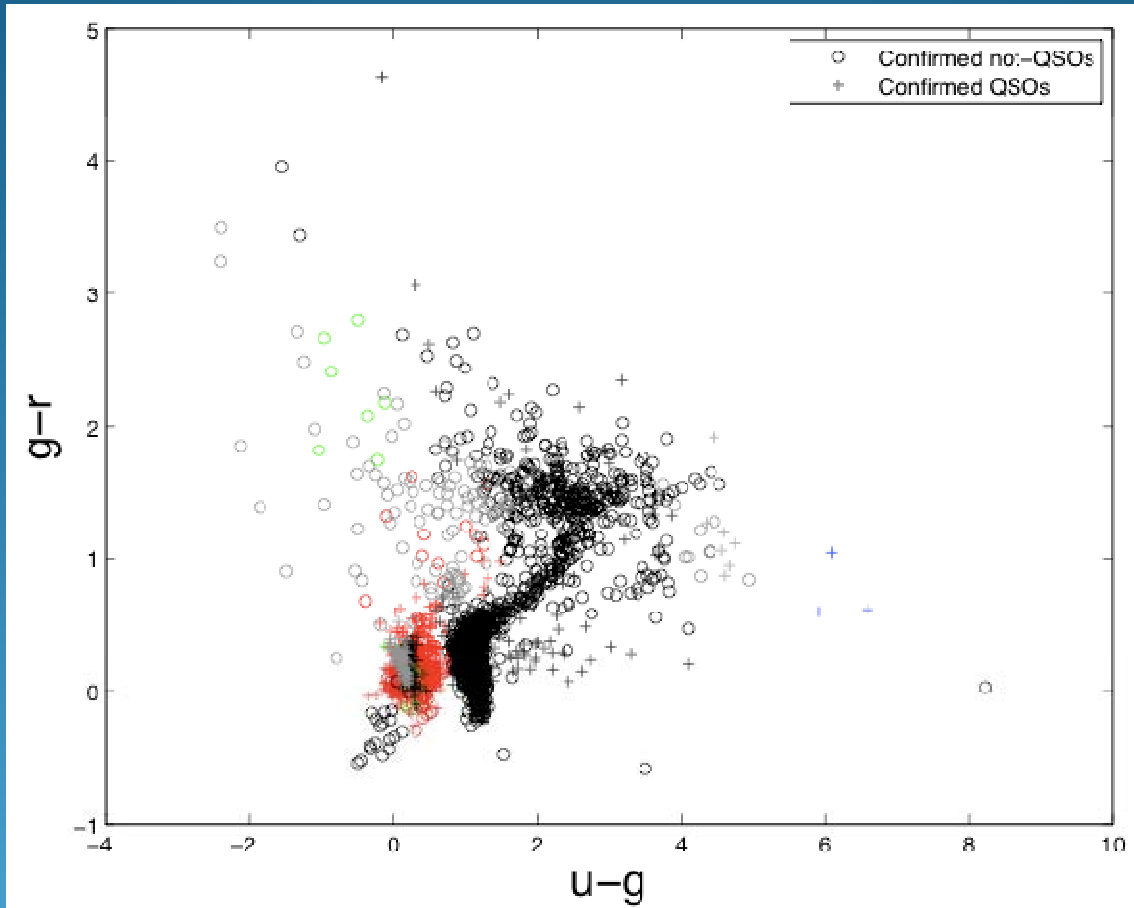
r - J vs **J - K**



Only a fraction (43%) of these objects have been selected as candidate QSO's by SDSS targeting algorithm in first instance: the remaining sources have been included in the spectroscopic program because they have been selected in other spectroscopic programmes (mainly stars).

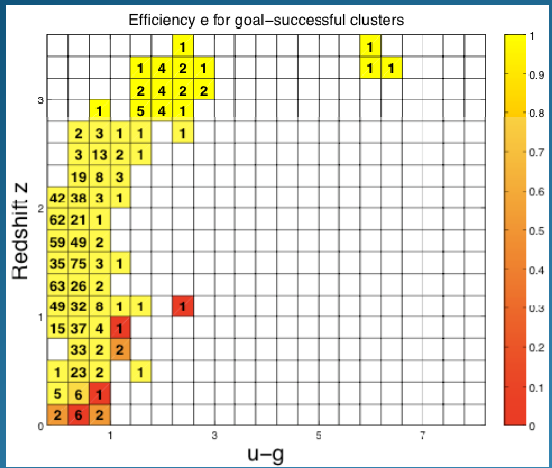
Experiment 3: optical colours

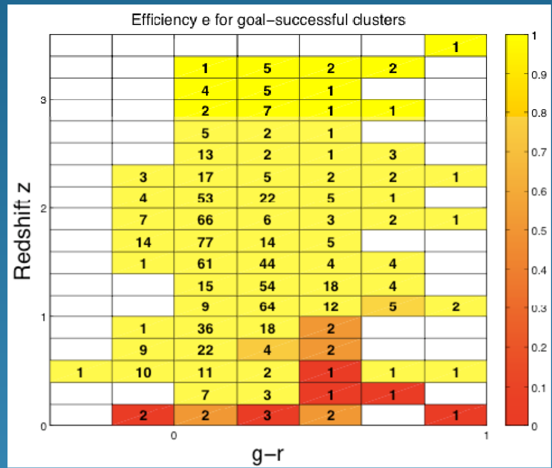
u - g vs g - r

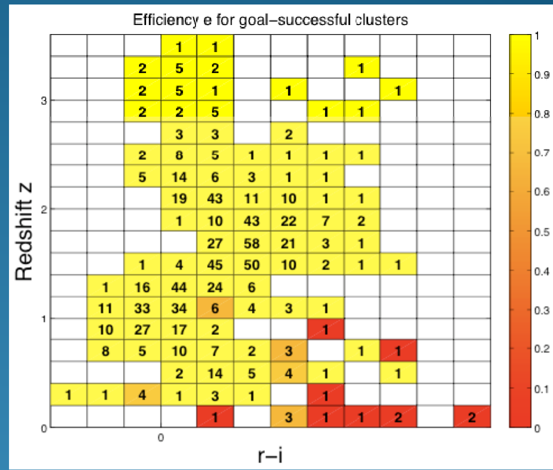


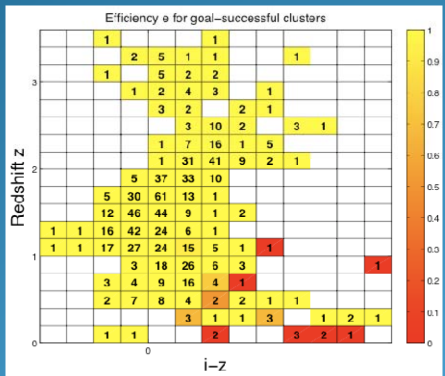
In this experiment the clustering has been performed on the same sample of the previous experiment, using only optical colours.

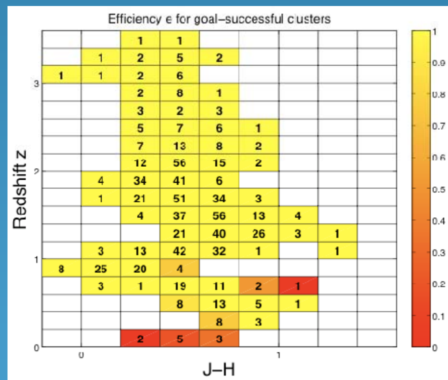
Experiment 2: local values of *e*

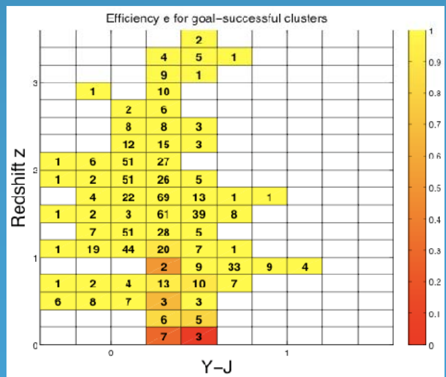


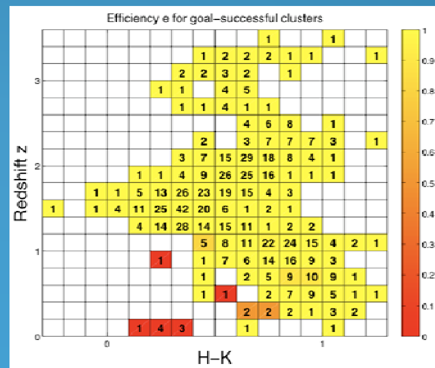




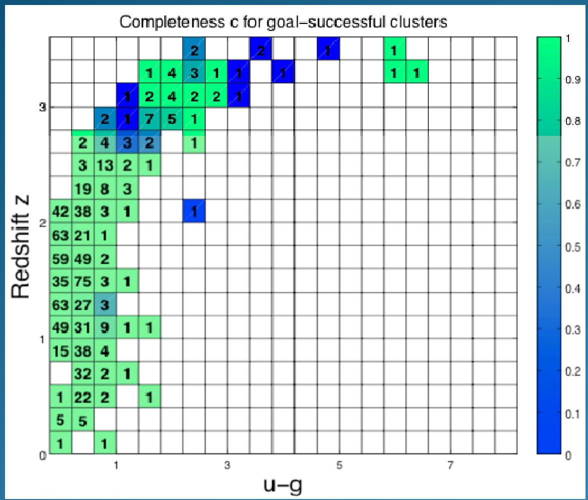


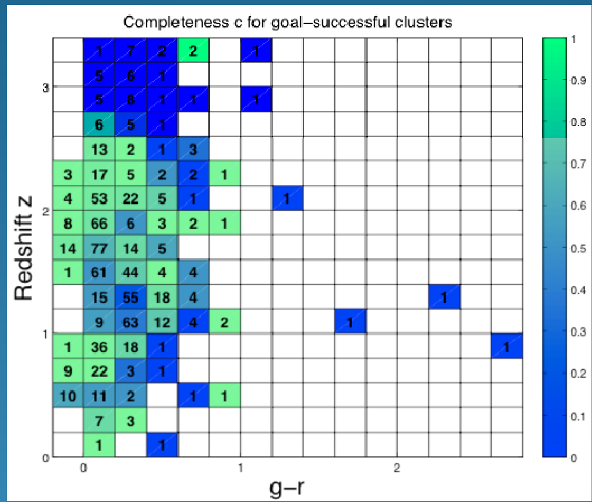


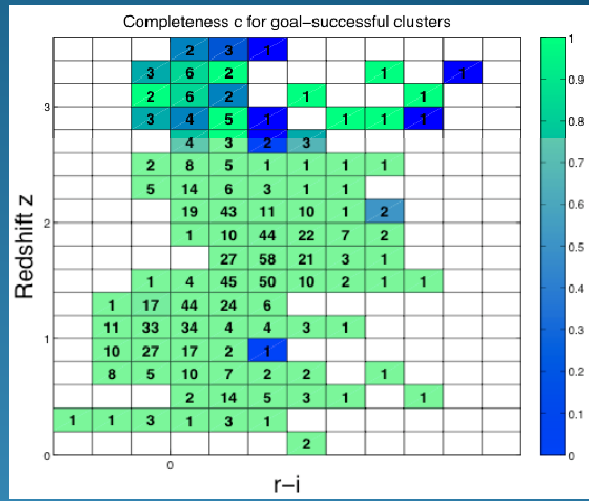


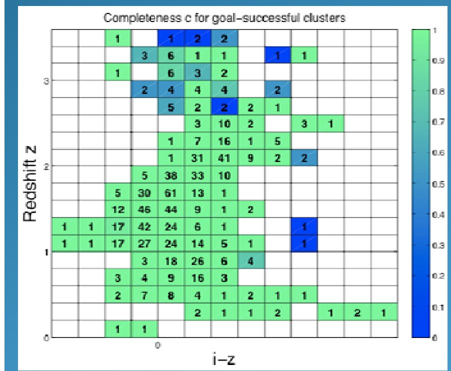


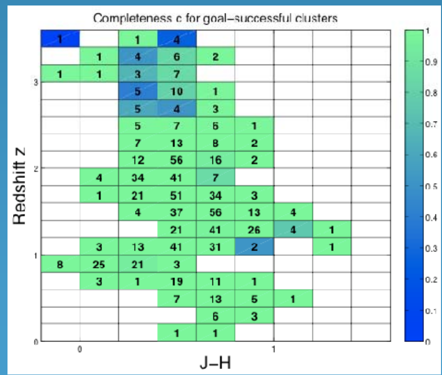
Experiment 2: local values of *c*

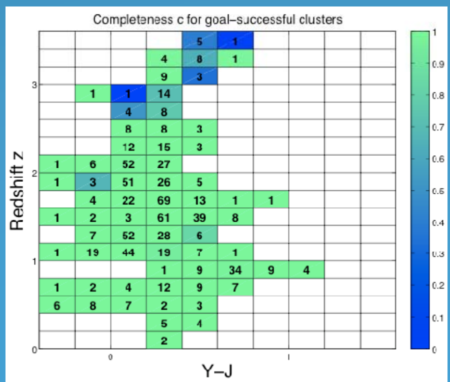


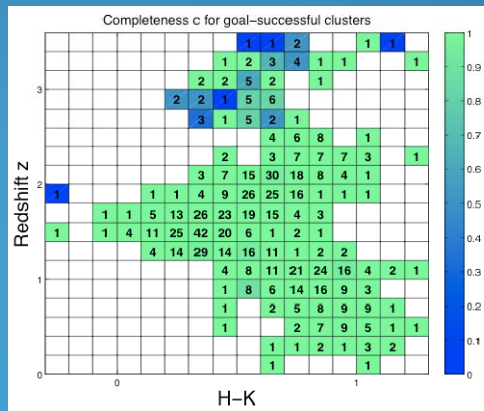






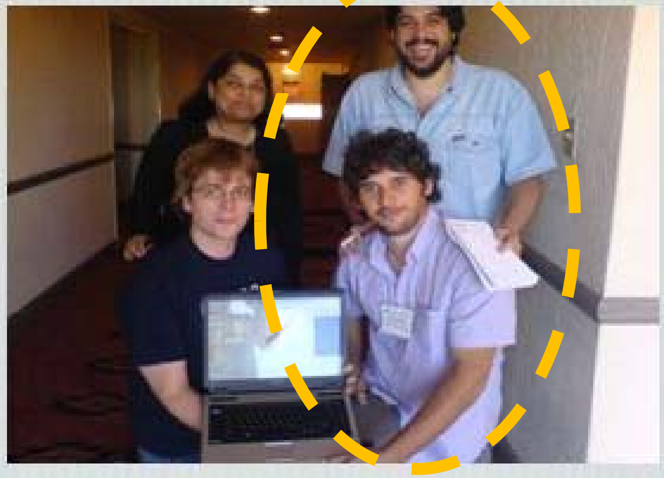






Results (I)

Sample	Parameters	Labels	e_{tot}	c_{tot}	n_{gen}	n_{suc_clus}
Optical QSO candidates (1)	SDSS colours	'specClass'	83.4 % (± 0.3 %)	89.6 % (± 0.6 %)	2	(3,0)
Optical + NIR star-like objects (2)	SDSS colours + UKIDSS colours	'specClass'	91.3 % (± 0.5 %)	90.8 % (± 0.5 %)	3	(3,1,0)
Optical + NIR star-like objects (3)	SDSS colours	'specClass'	92.6 % (± 0.4 %)	91.4 % (± 0.6 %)	3	(3,0,1)



Won one of the 2 prizes
Scientific application within Vobs

Talk at AAS Meeting - 2009

Photometric redshifts estimation for QSOs using Neural Networks

G. Barentsen, R. D'Abrusco, O. Laurino, P. Nayak

NVO Summer School 2008, Santa Fe

Pipeline for Photometric redshift estimation

SDSS

2MASS

VIM

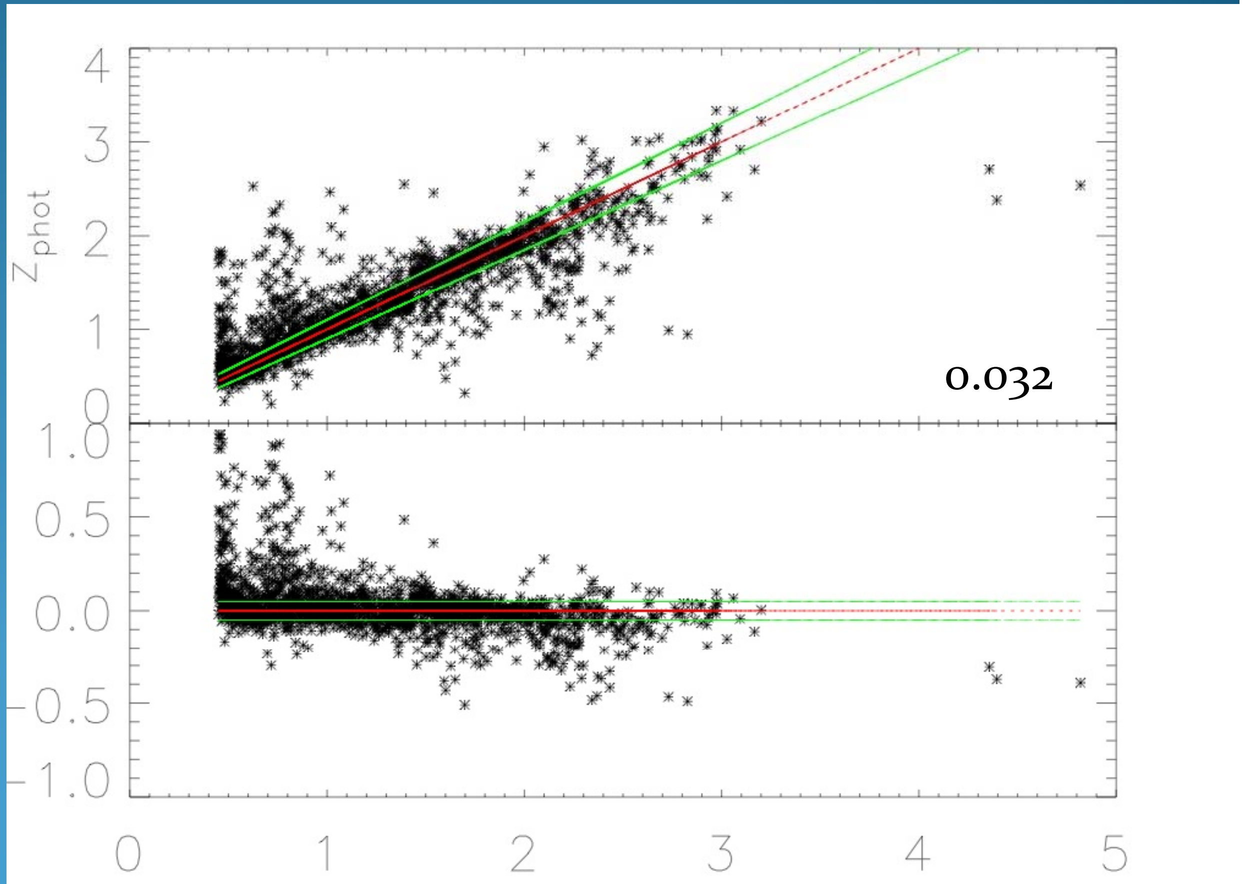
TOPCAT

PCA linear (R)

K-means

MLP

TOPCAT



A unified vision

Galaxy and QSOs photometric redshifts differences depend only on the different sparseness of the data (BoK).

Few points in a high dimensionality space (i.e. spectroscopic **QSOs**).

High sparseness

Many points in a high dimensionality space (i.e. spectroscopic **galaxies**)

Low sparseness

Clustering

Low sparseness

Crispy clustering, disjoint clusters, no redundancy.

High sparseness

Fuzzy clustering, overlapping clusters, redundancy.

Photo sources

N clusters

Disjoint

$C_i \cap C_j = \emptyset$
$\forall i \neq j$

$\mathbf{p} = \begin{cases} 1 & \text{for one clust.} \\ 0 & \text{for others} \end{cases}$

Overlapping

$\exists\, i \neq j:$
$C_i \cap C_j \neq \emptyset$

$\mathbf{p} = (p_i)$ where
$p_i \neq 0$ and $\sum_{i=1} p_i = 1$

Different sparseness of datasets can be taken into account when deriving photometric redshifts, by exploiting <u>redundance</u> between different clusters.

"For usual (crispy) clustering, assigning a photometric source to one of the closest cluster is straightforward (given a distance definition).
For a fuzzy clustering the probabilistic nature of assignment needs to be taken into consideration. This is the reason why the methods for galaxies and QSOs z_{phot} diverge."

Recipes: an outlook

(Low sparseness - galaxies)

- Each photometric source is assigned to one single cluster.
- The z_{phot} is calculated applying the NN trained on the members of that cluster.
- A unique value of z_{phot} with a unique accuracy and likelihood is produced.

(High sparseness - QSOs)

- Each photometric source can have a non-zero probability to belong to every clusters.
- For each source, an estimate of z_{phot} for each cluster is produced.
- A “committee” of NNs is used to determine the most reliable estimate of z_{phot} and the accuracy of the estimate.

Meta Institute for Computational Astrophysics

Conference room

meeting room

Strategy

- To exploit **new communication and interaction tools** (social networks, second life, etc) for teaching and dissemination activities.
- To extend and deepen collaboration with Caltech (and organize a school on e-science (2010) in collaboration with Caltech)
- To extend and deepen collaboration with IUCAA (Inter University Center for Astronomy and Astrophysics, Poona-India)
- To propose a Master in Data Mining and Exploration as joint activity among faculties (Economics, Science and Sociology) and Universities (Federico II, Sannio and Second University)
- To open the use of DAME to new communities (Bioinformatics, economics

Conclusions ?

- Methods are general and have been widely applied also outside of astronomy.
- In order to produce reliable results a large number of experiments is needed (as well as a good understanding of the tools). SUCCESSFUL SCIENCE CASES ARE A MUST
- Fast, optimized algorithms are required. They allow fast processing, with potentially better accuracy and a more detailed tracing of the process (the whole DR6 Galaxy photometric redshift catalogue went from 11 hs to 2.5 min)
- VO (or just the VO tools?) is not yet ready for data mining. But all that is needed is available. Visualization is still an issue
- BoK are the crucial issue for the future (need to bridge onthologies with intelligent BoK engines)

Conclusions ?

Funding

- Italy-USA "great relevance project" financed by MAE has been acknowledged by MAE as best project for 2008 and renewed for 2009.
- Funding pending from MIUR (PRIN) and from EU
- Funding foreseen in the framework of extension of PON-SCOPE

Technical steps

- To add new data mining models (e.g. SOM, PCA and ICA, Bayesian networks, etc)
- To add web applications for specific applications (e.g. NEXT-II + 2D-Phot)
- To integrate advanced visualization capabilities (STILTS, + VO-PLOT)
- To do a feasibility sudy for the automatic extraction of knowledge from VO archives (spectroscopic knowledge in coll. with Padua University and Padua INAF)
- Time series analysis and classification tools

Future science cases

- To integrate radioastronomy and optical data for WIMPS candidates to DM
- To improve on available Star/Galaxy classification using priors
- To improve AGN search and classification using supervised methods and improved spectroscopic base of knowledge.
- To study photometric transient classification and apply it to VST surveys.