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

Astromeeting – INAF OACN, Napoli, April 16, 2009
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)
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;
Why to land on an “open” distributed infrastructure

EGEE GRID  PRIVATE GRID

Enabling GRID for E-SciencE

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

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”
Virtualization brings new standardized capabilities to data centers.

**Virtualized Data Center**

**GRID Computing**
Resource oriented

**Cloud Computing**
Service oriented

Join data virtualization, resources and services brings to Virtual Organization of data

HW resource oriented

SW service oriented

Astromeeting – INAF OACN, Napoli, April 16, 2009
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

Astromeeeting – INAF OACN, Napoli, April 16, 2009
The astrophysical problem

Astronomical data rate

268,435,456 pixels
0.5 Gbyte x image
50 science frames + 50 calibration frames
↓
50 Gbyte / Night
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;

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
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”

\[ \mathbf{X} = \{x_1, x_2, x_3, \ldots, x_N\} \quad \text{input vectors} \]

\[ \mathbf{Y} = \{y_1, y_2, y_3, \ldots, y_M\} \quad \text{target vectors} \quad M \ll N \]

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

<table>
<thead>
<tr>
<th>variable</th>
<th>characteristics</th>
<th>Type</th>
<th>operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
<td>Numerical with ordering relationship and possibility to define a metric</td>
<td>Actual measurement</td>
<td>regression</td>
</tr>
<tr>
<td>Categorical (non ordered)</td>
<td>Membership into a finite number of classes. No ordering relationship.</td>
<td>Numerical codes (targets) arbitrarily ordered</td>
<td>Classification</td>
</tr>
<tr>
<td>Ordered categorical</td>
<td>Classes ordered by a relationship but there is no metric</td>
<td>Numerical codes ordered</td>
<td>Classification</td>
</tr>
</tbody>
</table>
Where do A.I. may fit into K.D.D.

A.I. tools
(soft computing techniques)
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
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

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

\[
\|Y - P(e)\| < \varepsilon
\]

\[
net_{input_j} = \sum_h W_{jh}O_j
\]

\[
net_{input_h} = \sum_i W_{hi}O_i
\]

\[
w_{ji}^{(new)} = w_{ji}^{(old)} + \eta \delta o_j + \alpha \Delta w_{ji}^{(old)}
\]

\[
f(o) = \frac{1}{1 + \exp(-o)}
\]
given a training set formed by pairs [features-label]: \((x_i, y_i), i = 1...l\)

where \(x_i \in \mathbb{R}^n\) e \(y_i \in \{1,-1\}\).

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 hyperplane which maximizes the distances from the two classes

\(C > 0\) is a classification error correction term

\[
K(x_i, x_j) = \phi(x_i)^T \phi(x_j)
\]

Is the so called Kernel function

\[
K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0
\]

- linear: \(K(x_i, x_j) = x_i^T x_j\).
- polynomial: \(K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0\).
- radial basis function (RBF): \(K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0\).
- sigmoid: \(K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)\).

We should maximize the margin, \(m\)

\[
m = \frac{2}{\|w\|}
\]

\[
w^T x + b = 1
\]

\[
w^T x + b = -1
\]

\[
w^T x + b = 0
\]
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 represents 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 consists 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

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

\[ |Y - P(e)| < \varepsilon \]

\[ \text{net\_input}_j = \sum_h W_{jih}O_h \]

\[ \text{net\_input}_h = \sum_i W_{hi}O_i \]

\[ w_{ji}(\text{new}) = w_{ji}(\text{old}) + \eta \delta_{o_j} + \alpha \Delta w_{ji}(\text{old}) \]

\[ f(o) = \frac{1}{1 + \exp(-o)} \]
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**

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)
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 $t = 1, 2, \ldots$ 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 $c = c(x)$ are updated as

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

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

**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 $\text{NegE}(A \cup B) < d$, then clusters $A$ and $B$ are replaced by cluster $A \cup B$.

**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 $\mathbb{R}^3$ 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”.

NegE=750
NegE=4
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**

**Key mathematical issues**

**Database technologies**

**Ongoing research**
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)
VO-Neural / Data Mining Exploration

Project Team (now)

VONEURAL/DAME ORGANIZATION CHART

G. Longo
Principal Investigator

M. Broscia
Project Manager

O. Laurino
Project Engineer

S.G. Djorgovski
A. Corazza
R. D’Abrusco
D. Capozzi
E. De Filippis
A. Staiano
R. Tagliaferri

S. Cauvoti
C. Donalek
A. Di Guido

G. d’Angelo
Infrastructure

A. Nocella
M. Garofalo
F. Manna
M. Fiore

Software Engineering

Data Mining Modeling

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

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

VO-Neural / Data Mining Exploration

XP – eXtreme Programming as project life cycle

Copyright 2000 J. Donvan Wells
Functionality taxonomy

- **MLP**
- **SVM**
- **MLPGA**

**Expert systems**
- Specialized
- Supervised
- Unsupervised
- Bayesian or not bayesian

**Deterministic**
- Self-adaptive
- Stochastic

**Forecasting**
- Classification
- Regression
- Clustering
- Filtering

**Prediction**
- Exploration
- Dimension reduction

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

<table>
<thead>
<tr>
<th>PROJECT MILESTONES</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement of work</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBS &amp; Project Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Design Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Requirement Specifications</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DM models Implementation</td>
<td></td>
<td></td>
<td>SOM</td>
<td></td>
</tr>
<tr>
<td>DM models scientific validation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Design Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation &amp; Test Procedures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical Reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta Release Deployment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User &amp; Maintenance Manuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commissioning of beta release</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Official release 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New DM models Implementation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Functionalities Implementation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*MLP, SVM, MLPGA, PPS, NEC*  
*Astromeeing – INAF OACN, Napoli, April 16, 2009*
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
**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
**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);

*Image of diagram with DRMS (Library) and FW (Framework).*
Architecture:
- data mining functionality class hierarchy;

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

Features:
- modularity;
- fast third party application integration;
- functionality specialization;
- multi-language programming support;
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);
**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
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?

**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;

Astromeeting – INAF OACN, Napoli, April 16, 2009
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-COPE.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.

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

Collaborations:
- VOTECH (Virtual Observatory Technology Infrastructure)
- S.C.O.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
- IFCA (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...
New user registration

DAME - DAta Mining and Exploration

Create an account

First name: [Input field]
Last name: [Input field]
Username: [Input field]
Email address: [Input field]
Password: [Input field]
Password again: [Input field]

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

DAME - DData Mining and Exploration

What is DAME

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

As a function of the size and complexity of your task, your computation will be more devoted to longer computing facility.

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

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

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

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

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

In the "Help & Tutorials" section you will find documentation, examples and tutorials. The first time you log in, your file store will contain some datasets you can use following the tutorials.
Beta release feature:

integrated dataset editor and builder
Launch new experiments

VO-Neural / Data Mining Exploration

DAME - DData Mining and Exploration

Experiment Details

Experiment Name: myExperiment
Progress: Finished

Parameter | Value
--- | ---
Input Nodes | 4
Hidden Nodes | 0
Output Nodes | 1
Max Epochs | 1000
Tolerance | 1E-05
Training Algorithm | momentum
Training Set | /brescia/Samples/provapiccolo.csv
Validation Set | /brescia/Samples/provapiccolo.csv

Exp Log

MIP:
- learning setup: //
- input nodes: 4
- output nodes: 3
- hidden layer: 0
- max epoch: 100
- learning setup: Classification
- training algorithm: Incremental
- error: RSS
- error tolerance: 1E-05
- input net: empty
- output net: empty
- training dataset: /myExperiment/provapiccolo.csv
- validation dataset: /myExperiment/provapiccolo.csv
- testing dataset: /myExperiment/provapiccolo.csv

Plot
DAME - DData Mining and Exploration

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 field]

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: [input field] 4

Hidden Nodes Help

Hidden nodes: [input field] 3

Output Nodes Help

Output nodes: [input field] 2

Max epochs: 40000

Tolerance: 1e-05

Training algorithm: [input field] None - BATCH

Resume training: [input field] None

Network: [SampleStdSvm]

Training set: [SampleStdSvm]

Do validation: [input field] None

Validation set: [SampleStdSvm]

Start
**VO-Neural / Data Mining Exploration**

**Application Prototype**

**DAME - DData Mining and Exploration**

**Status during execution**

**Beta release feature:**

Interactive session optional during execution

Astromeeting - INAF OACN, Napoli, April 16, 2009
Status
when finished

Astromeeting – INAF OACN, Napoli, April 16, 2009
First example
evaluation of SDSS redshift using supervised NN (MLP)

Second example
Searching for candidate quasars in the SDSS archive

Third example
Classifying AGN in SDSS with SVM

More infos on WEB site documentation page
http://voneural.na.infn.it/documents.html


Cavuoti 2008, Thesis (VONeural website, voneural.na.infn.it)
Science case: Mining the SDSS archive

Galaxy spectrum - F(\lambda)  
Photometric system - S_i(\lambda)

\( X = \int \frac{F(\lambda)S_u(\lambda)d\lambda}{\int S_u(\lambda)d\lambda} + c_u \)

\( m_u = -2.5\log_{10} \int \frac{F(\lambda)S_u(\lambda)d\lambda}{\int S_u(\lambda)d\lambda} + c_u \)

\( m_b = -2.5\log_{10} \int \frac{F(\lambda)S_b(\lambda)d\lambda}{\int S_b(\lambda)d\lambda} + c_b \)

Color indexes

\( U - B \equiv m_U - m_B \)

\( B - R \equiv m_B - m_R \)

etc.

Astromeeting – INAF OACN, Napoli, April 16, 2009
Photometric Redshifts

- SED template fitting methods
- Interpolative methods

\[ \sigma = 0.051 \]

\[ \Delta z = 0.0144 \]

\[ \sigma = 0.0415 \]

\[ \Delta z = 0.0144 \]

<table>
<thead>
<tr>
<th>type</th>
<th>method</th>
<th>data</th>
<th>( \Delta z_{\text{rms}} )</th>
<th>Notes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEDF</td>
<td>Bruzual-Charlot</td>
<td>EDR</td>
<td>0.0552</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>Interpolated</td>
<td>EDR</td>
<td>0.0451</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>Polynomial</td>
<td>EDR</td>
<td>0.0318</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>KD-tree</td>
<td>EDR</td>
<td>0.0254</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>ANNz</td>
<td>EDR</td>
<td>0.0229</td>
<td></td>
<td>(Collister &amp; Lahav 2004)</td>
</tr>
<tr>
<td>ML</td>
<td>SVM</td>
<td>EDR</td>
<td>0.027</td>
<td></td>
<td>(Wadadekar 2004)</td>
</tr>
<tr>
<td>ML</td>
<td>MLP-feed forward</td>
<td>SDSS-DR1</td>
<td>xx.xxx xxx</td>
<td>yes</td>
<td>(Vanzella et al. 2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SDSS-RLG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Photometric Redshifts

- 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.

\[ \sigma = 0.023 \]

**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 shows the results.
科学案例：光度学红移

SDSS – DR5

光谱学BoK
- 训练
- 验证
- 测试集

MLP1 [1(5), 1(18)]

0.01<Z<0.25
MLP2 [1(5), 1(23)]

0.25<Z<0.50
MLP3[1(5), 1(24)]

99.6 % 精确度

MLP1

σ rob = 0.0196

MLP2

σ rob = 0.0201

MLP3

光度学对象
VO-Neural / Data Mining Exploration

Science case: Photometric Redshifts

\[ \sigma = 0.0183 \]

SDSS – DR5 - LRG
Science case: Photometric Redshifts

General galaxy sample

LRG sample

Non LRG only

$\sigma = 0.0363$
$\Delta z = -0.0030$

$\sigma = 0.0208$
$\Delta z = -0.0029$

$\sigma = 0.0178$
$\Delta z = -0.0011$

Astromeeting – INAF OACN, Napoli, April 16, 2009
Science case: Photometric Redshifts

\[ \sigma = 0.0170 \]
Science case: Photometric Redshifts

Fig. 9.— Same as in previous figure but for the LRG sample.
SDSS galaxies $z_{\text{phot}}$

- 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 $n_{\text{opt}}$.
- $n_{\text{opt}}$ is chosen so that the “weighted accuracy” $\sigma_w$ of the $z_{\text{phot}}$ is maximum. Given $N$ clusters with $M_i$ elements each and $\text{rms}$ of the $(z_{\text{phot}}-z_{\text{spec}})$ variable $\sigma_i$:

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

where

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

Astromeeting – INAF OACN, Napoli, April 16, 2009
Searching for candidate quasars in the SDSS

Traditional way to look for candidate QSO in 3 band survey

Cutoff line

In 4 bands degeneracy is partially removed

Candidate QSOs for spectroscopic follow-up's

Ambiguity zone

A Generic Machine-Assisted Discovery Problem: Data Mapping and a Search for Outliers

How to find the interesting regions (clusters)?
• Data Mining is the answer

How to visualize them?
• Dimensionality reduction

Astromeeting – INAF OACN, Napoli, April 16, 2009

R. D’Abrusco

astro-ph/0805.0156v1

More are the bands, the lower is the degeneracy
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:

Compleness \( C = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}} \)

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** $>4\sigma$ **far from a cylindrical region containing the “stellar locus” (S.L.), where** $\sigma$ **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
1. **Inclusion regions** are regions where S.L. meets QSO’s area (due to absorption from Lyα forest entering the SDSS filters, which changes the 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

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.

Clustering method

- PPS

  ↓

  labeling

  ↓

  Spectroscopic BoK

  ↓

  Cluster aggregation
  NEC

  ↓

  Dissimilarity treshold

  ↓

  Evaluation of successful clusters

  ↓

  Partition of parameter space
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 successful clusters

Once partition of colors 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:

- A cluster is “successful” if its fraction of confirmed QSO is higher than a fixed value.

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

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

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 $\epsilon_{tot}$ is the sum of weighed efficiencies $\epsilon_i$ for each generation:

$$\epsilon_{tot} = \sum_{i=1}^{n} \epsilon_i$$
An example of “tuning”

**Choice of the clustering**

**NSR**

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.

**Efficiency and completeness**

<table>
<thead>
<tr>
<th>algorithm</th>
<th>QSOs</th>
<th>not QSOs</th>
</tr>
</thead>
<tbody>
<tr>
<td>QSOs</td>
<td>759</td>
<td>83</td>
</tr>
<tr>
<td>not QSOs</td>
<td>72</td>
<td>1327</td>
</tr>
</tbody>
</table>

Astromeeting – INAF OACN, Napoli, April 16, 2009
Data and experiments

Data samples:

1. **Optical**: sample derived from SDSS database table “Target” queried for QSO candidates, containing $\sim 1.11 \cdot 10^5$ records and $\sim 5.8 \cdot 10^4$ 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
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).
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 $\epsilon$
Experiment 2: local values of $c$
### Results (I)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Parameters</th>
<th>Labels</th>
<th>$e_{\text{tot}}$</th>
<th>$c_{\text{tot}}$</th>
<th>$n_{\text{gen}}$</th>
<th>$n_{\text{suc_clus}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optical QSO candidates (1)</strong></td>
<td>SDSS colours</td>
<td><code>specClass</code></td>
<td>83.4 % (± 0.3 %)</td>
<td>89.6 % (± 0.6 %)</td>
<td>2</td>
<td>(3,0)</td>
</tr>
<tr>
<td><strong>Optical + NIR star-like objects (2)</strong></td>
<td>SDSS colours + UKIDSS colours</td>
<td><code>specClass</code></td>
<td>91.3 % (± 0.5 %)</td>
<td>90.8 % (± 0.5 %)</td>
<td>3</td>
<td>(3,1,0)</td>
</tr>
<tr>
<td><strong>Optical + NIR star-like objects (3)</strong></td>
<td>SDSS colours</td>
<td><code>specClass</code></td>
<td>92.6 % (± 0.4 %)</td>
<td>91.4 % (± 0.6 %)</td>
<td>3</td>
<td>(3,0,1)</td>
</tr>
</tbody>
</table>
Photometric redshifts estimation for QSOs using Neural Networks

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

NVO Summer School 2008, Santa Fe

Won one of the 2 prizes
Scientific application within Vobs
Talk at AAS Meeting - 2009
Pipeline for Photometric redshift estimation

- SDSS
- 2MASS

- VIM
- TOPCAT
- PCA linear (R)
- K-means
- MLP
- TOPCAT

Astromeeting – INAF OACN, Napoli, April 16, 2009
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).

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

High sparseness

Low sparseness
Clustering

Low sparseness
Crispy clustering, disjoint clusters, no redundancy.

High sparseness
Fuzzy clustering, overlapping clusters, redundancy.

Different sparseness of datasets can be taken into account when deriving photometric redshifts, by exploiting redundancy 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 $z_{\text{phot}}$ diverge.”

Astromeeting – INAF OACN, Napoli, April 16, 2009
Recipes: an outlook

(Low sparseness - galaxies)

• Each photometric source is assigned to one single cluster.
• The $z_{\text{phot}}$ is calculated applying the NN trained on the members of that cluster.
• A unique value of $z_{\text{phot}}$ 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 $z_{\text{phot}}$ for each cluster is produced.
• A “committee” of NNs is used to determine the most reliable estimate of $z_{\text{phot}}$ and the accuracy of the estimate.

Astromeeting – INAF OACN, Napoli, April 16, 2009
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 ontologies with intelligent BoK engines)
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 study 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.