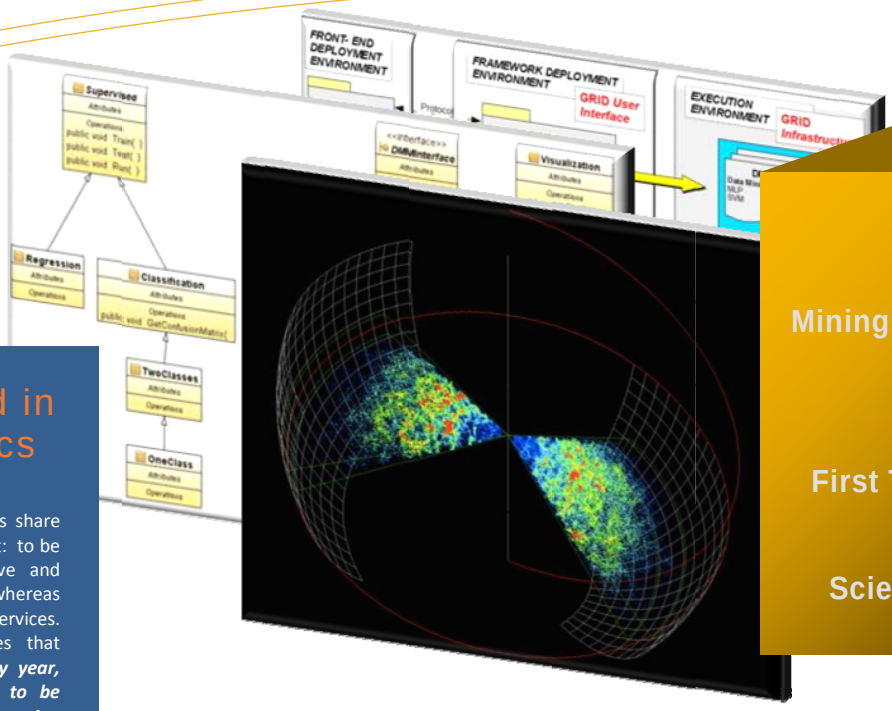
Newsletter Edition
01
November 2009

DAta Mining & Exploration



in this edition

A New trend in Astrophysics

Different astrophysics areas share the same basic requirement: to be able to deal with massive and distributed datasets whereas possible integrated with services. A famous sentence states that ***"While data doubles every year, useful information seems to be decreasing, creating a growing gap between the generation of data and our understanding of it".*** This new understanding includes knowing how to access, retrieve, analyze, mine and integrate data from disparate sources. But on the other hand, it is obvious that a scientist cannot and does not necessarily want to become an expert in Computer Science or in the fields of algorithms and ICT (Information & Communication Technology).

The idea behind DaME is to provide **a user friendly scientific gateway to easy the access, exploration, processing and understanding of massive data sets. In the field of astronomy, DAME represents a typical product of the emerging field of Astroinformatics.**

Bioinformatics, geoinformatics, Astroinformatics are growingly being recognized as the fourth leg of scientific research after experiment, theory and simulations. They arise from the pressing need to acquire the multi-disciplinary expertise which is needed to deal with the ongoing burst of data complexity and to perform data mining and exploration on MDS.

DAME (DAta Mining & Exploration) is a project aimed at designing and developing instruments and tools for scientific data mining, based on state of the art Information and Communication Technology.

Web Solutions for Data Mining in Astrophysics

Modern scientific data mainly consist of huge datasets gathered by a very large number of techniques and stored in very diversified and often incompatible data repositories. More in general, in the e-science environment, it is considered as a critical and urgent requirement to integrate services across distributed, heterogeneous, dynamic "virtual organizations" formed by different resources within a single enterprise. The Astronomy and Astrophysics environment has become an immensely data rich field due to the evolution of detectors (plates to digital to mosaics), telescopes and space instruments. The Virtual Observatory approach consists into the federation under common standards of all astronomical archives available worldwide, as well as data analysis, data mining and data exploration applications. The main drive behind such effort being that once the infrastructure will be completed, it will allow a new type of multi-wavelength, multi-epoch science which can only be barely imagined.

The DAME project pursues these goals by integrating the VObs standards in a service oriented infrastructure, including the implementation of advanced tools for Massive Data Sets (MDS) exploration, soft computing, data mining (DM) and Knowledge Discovery in Databases (KDD). The DAME project **(http://voneural.na.infn.it),** run jointly by the Department of Physics of the University Federico II, INAF (National Institute of Astrophysics) Astronomical Observatory of Napoli, and the California Institute of Technology, is financed through grants from the ***Italian Ministry of Foreign Affairs***, the ***European projects VO-TECH*** and ***VO-AIDA*** and by the USA - ***National Science Foundation***. DAME makes use of distributed computing environments (e.g. the S.Co.P.E. - GRISU infrastructure) and matches the international IVOA standards and requirements.

Currently, the DAME project, already completed in terms of the engineering design aspects, is under implementation of the first official version of the complete data mining & processing tool package, deployed on SCoPE GRID Infrastructure. But a prototype is already available (please visit project website) and successfully tested on significant science cases, as it is well demonstrated by already published scientific papers.

Newsletter
Edition
01
November
2009

Mining Astrophysical datasets with DAME

M. Brescia, G. Longo & the DAME team

The International VObs (Virtual Observatory, cf. the IVO alliance or IVOA at http://ivoa.net) has opened a new frontier to astronomy. In fact, by making available at the click of a mouse an unprecedented wealth of data and by implementing common standards and procedures, the VObs allow a new generation of scientists to tackle complex problems which were almost unthinkable only a decade ago. Astronomers may now access a "virtual" parameter space of increasing complexity (hundreds or thousands features measured per object) and size (billions of objects). The use of the VObs, however, is not yet as widespread in the scientific community as it should be. In our opinion, this is due to several concomitant factors. First of all the fact that, in spite of all the efforts made to render the infrastructure user friendly, the use of the VObs tools still requires a methodological shift that not everyone is willing to do. Users need to become familiar with concepts such as distributed resources, web applications, resource standards, etc. which, at least at the moment, are not treated in any university course. Second and in our opinion more important, in order to convince the community to make the needed effort, the VObs needs to showcase its potentialities by producing significant scientific results.

A survey of the literature listed in the NASA ADS and published in the last decade (i.e. since the meeting "Virtual Observatories of the future" held in Pasadena in 2001) seems to confirm such analysis.

The study was performed by searching the abstracts of the NASA ADS literature database for all papers quoting the words "VObs" or "Data Mining" in all their different formulations. While the total number of papers (refereed + not refereed) is in both cases rather high (with peaks in correspondence of the appearance of specific proceedings), the number of refereed papers (i.e. papers producing scientific results and published on main journals) remains low and practically constant (see Fig. 1 in next page). An inspection of the authors shows that these papers are mainly produced by the same teams and confirms that VObs Technologies are not spreading out of the community at the pace we would have expected.

A more detailed analysis of the papers referring to "data mining" is shown in Fig. 2 where we give the trends for refereed and (refereed + not refereed) papers making use of Neural Networks (NN), Statistical Pattern recognition methods (SPR), Machine Learning or Artificial Intelligence (ML+AI) methods not falling in the previous categories, plus the papers using the words "knowledge discovery" (KD) in the abstract. Some surprising features may be noticed.

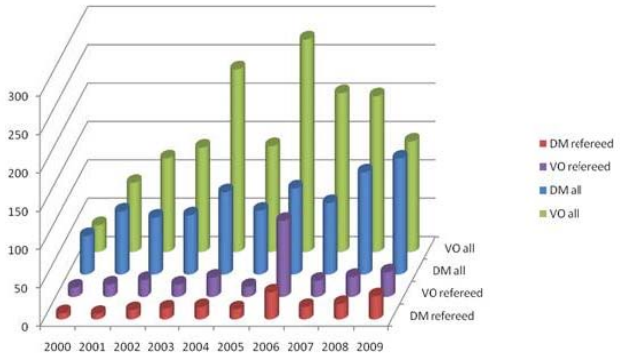


Fig. 1 – NASA ADS database search quoting the words "VObs" or "Data Mining"

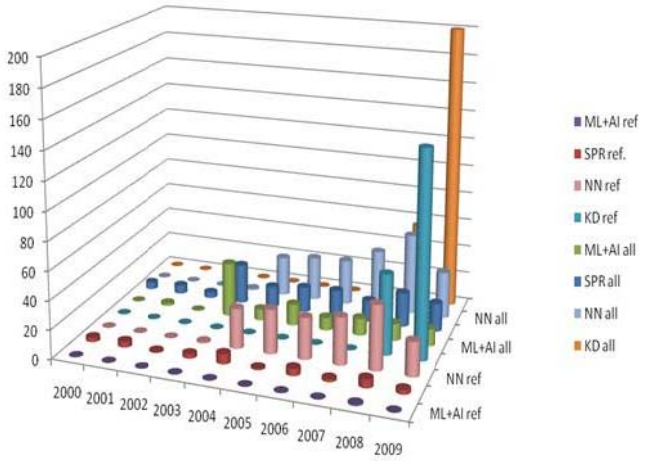


Fig. 2 – NASA ADS database search quoting the words related to Data Mining

The words "knowledge discovery", virtually absent before 2008, suddenly became popular causing a large increase in publications using the terminology. A fact that, at least at the moment, is not easy to understand.

Not surprisingly, Neural Networks, which have been the first machine learning method used in astronomy,

present the largest number of refereed papers. This is mainly due to their application to the problem of photometric redshifts and while the number of methods remains constant, the number of groups using them in their research has increased, thus confirming that once a method is made available to the community in an "user friendly" way, the number of users quickly increases.

These, rather unsatisfactory results, however, could result also from the fact that people using VObs data or tools do not explicitly quote the VObs in their acknowledgements or abstracts.

An additional test was therefore made in the course of a Master Thesis yet to be completed. 1000 papers published in the period 2007-2008 and having among the keywords "observational cosmology" are being surveyed to: i) see whether they had or had not made use of VObs data or methods; ii) see whether they could benefit or not by the use of VObs tools or methods. The latter result will be obviously rather subjective (but still indicative) since the outcome strongly depends on the personal judgement of the reader.

Even so, however, a preliminary analysis of the data shows that almost 12% of the papers made use of VObs technologies without explicitly mentioning the VObs anywhere in the text. This is almost always the result of the fact that the authors did not seem to be aware that some tools they used (such as TOPCAT or ALADIN) are the result of the VObs efforts. Whether this is good (VObs as transparent infrastructure) or bad (not enough return from the investments) remains to be assessed. Even more intriguing is the fact that according to our subjective opinion almost 30% of the papers could have, one way or the other, benefited by the use of VObs methods.

At the present stage it is not possible to draw definitive conclusions but at least two facts are apparent. First of all, the usage of VObs tools and methods (even though still in its infancy) is much more widespread then what can be inferred from a superficial analysis of the literature. Second, the "marketing" strategy of the VObs tools and methods is not optimal: people, either consciously or unconsciously, tend not to acknowledge their usage. The same wording problem exists also for all DM, SPR, KD etc. methods.

Tagliaferri et al. 2003	Ball & Brunner 2009	BoK
S/G separation	S/G separation	yes
Morphological classification of galaxies *(shapes, spectra)*	Morphological classification of galaxies *(shapes, spectra)*	yes
Spectral classification of stars	Spectral classification of stars	yes
----	Image segmentation	no
Noise removal *(grav. waves, pixel lensing, images)*	---	No/yes
Photometric redshifts *(galaxies)*	Photometric redshifts *(galaxies, QSOs)*	yes
Search for AGN	Search for AGN and QSOs	yes
Variable objects	Time domain	Yes
Partition of photometric parameter space for specific group of objects	Partition of photometric parameter space for specific group of objects	no
Planetary studies (asteroids)	Planetary studies (asteroids)	yes
Solar activity	Solar activity	no
Instellar magnetic fields	---	No
Stellar evolution models	---	Yes

Fig. 3 – A cross comparison between two extensive reviews

A cross comparison between two rather extensive reviews published in 2003 and 2009 respectively (Tagliaferri et al 2003, Ball and Brunner 2009) showed that the fields of application of machine learning methods have remained almost unchanged, (Fig. 3). Supervised methods tend to monopolize the game, while unsupervised methods are virtually unused (with only exception of image segmentation and search for QSOs). Not surprisingly, supervised DM (Data Mining) methods tend to emerge only in those fields where reliable and robust basis of knowledge (i.e. training sets based on well defined templates) are either already available or rather easy to build. Typical examples being the spectroscopic redshifts in the SDSS database and/or the morphological types extracted from the Third Reference Catalogue of Bright Galaxies.

In summary, by taking into account all the above, it is quite apparent that in order to increase the usage of DM and ML (Machine Learning) methods within the astronomical community, it is necessary to fulfil at least the following requirements. Methods need to be stable and robust (cf. the photometric redshifts case quoted above) as well as user friendly since astronomers do not seem to be willing to make the effort to understand the details of a method. This implies that supervised methods are necessarily favoured since they allow to achieve a good grasp on the quality of the final results even without a deep understanding of the subtleties of the method itself. This also implies that ML methods tend to be adopted

only for those problems where reliable BoKs are either available or easy to produce.

DAME Generalities

The DAME project aims at creating a distributed e-infrastructure to guarantee integrated and asynchronous access to data collected by very different experiments and scientific communities in order to correlate them and improve their scientific usability. The project consists of a data mining framework with powerful software instruments capable to work on MDS in a distributed computing environment.
One first and important step in the standardization direction has already been undertaken within the Astrophysics community with the set of initiatives flourished under the generic name of Virtual Observatory (VObs). The VObs, organized worldwide by means of the International Virtual Observatory Alliance (IVOA), has defined a set of standards to allow interoperability among different archives and databases in the astrophysics domain, and keeps them updated through the activity of dedicated working groups. The first goal of the VObs is the federation under common standards of all astronomical archives available worldwide. The idea being that having this meta-archive completed, its exploitation will allow a new type of multi-wavelength, multi-epoch science which can only be barely imagined, but also poses unprecedented computing problems. So far, most of the implementation effort for the VO has concerned the storage, standardization and interoperability of the data together with the computational infrastructures. In particular it has focused on the realization of the low-level tools and on the definition of standards.
Our project extends this fundamental target by integrating it in an infrastructure, joining service-oriented software and resource-oriented hardware paradigms, including the implementation of advanced tools for MDS exploration, soft computing, data mining (DM) and Knowledge Discovery in Databases (KDD). Moreover, data mining services often run synchronously. This means basically that they execute jobs during a single HTTP transaction. This might be considered useful and simple, but it does not scale well when it is applied to long-run tasks. Typical long-running activities are the following:

- any archive query traversing a massive DB table;
- a data-mining job running from a batch (sequential) queue;
- a pipeline workflow with several computing-intensive steps;
- a pipeline workflow applied sequentially for many (and massive) data sets.

In any of these cases, the system is stressed if the activity lasts longer than a few minutes and becomes unreasonably fragile if it lasts longer than a few hours. With synchronous operations, all the entities in the chain of command (client, workflow engine, broker, processing services) must remain up for the duration of the activity. If any component goes down or stops then the context of the activity is lost and the must be restarted. To overcome this limitation, one of the main DAME design strategies is to permit asynchronous access to the infrastructure tools, allowing running of activity jobs and processes outside the scope of any particular web-service operation and without depending on the user connection status.
The user, via client web applications, can asynchronously find out the state of the activity, has the possibility to keep track of his jobs by recovering related information (partial/complete results) without having the need to maintain open the communication socket. Moreover, the system is able to automatically perform a sort of garbage collection for cleaning up resources, swap areas and temporary system tools used during the activity run phase.
Furthermore, as it will be discussed in what follows, the DAME design takes into account the fact that the average scientists cannot and/or does not want to become an expert also in Computer Science or in the fields of algorithms and ICT (Information & Communication Technology). In most cases the r.m.s. scientist (our end user) already possesses his own algorithms for data processing and analysis and has implemented private routines/pipelines to solve specific problems. These tools, however, often are not scalable to distributed computing environments or are too difficult to be migrated on a GRID infrastructure.
DAME aims at providing **a user friendly scientific gateway to easy the access, exploration, processing and understanding of the massive data sets federated under standards according Virtual Observatory rules.**

There are important reasons why to adopt existing VO standards: long-term interoperability of data, available e-infrastructure support for data handling aspect in the future projects. We wish to emphasize that standardization needs to be extended to data analysis and mining methods and to algorithms development. It basically means to define standards in terms of ontologies and well defined taxonomy of functionalities to be applied in the astrophysical use cases. The natural computing environment for MDS processing is a distributed infrastructure (GRID/CLOUD), but again, we need to define standards in the development of higher level interfaces, in order to:

- isolate end user (astronomer) from technical details of VO and GRID/CLOUD use and configuration;
- make it easier to combine existing services and resources into experiments;

A new perspective of Data Mining

Data Mining is usually conceived as an application (deterministic/stochastic algorithm) to extract unknown information from noisy data. This is basically true but in some way it is too much reductive with respect to the wide range covered by mining concept domains. More precisely, in DAME, data mining is intended as techniques of exploration on data, based on the combination between parameter space filtering, machine learning, soft computing techniques associated to a functional domain. The functional domain term arises from the conceptual taxonomy of research modes applicable on data. Dimensional reduction, classification, regression, prediction, clustering, filtering are example of functionalities belonging to the data mining conceptual domain, in which the various methods (models and algorithms) can be applied to explore data under a particular aspect, connected to the associated functionality scope.

There is a need to organize in an homogeneous way both syntax and semantics of data manipulation, analysis and reduction, with the aim at the creation of an ontology capable to ease communication between scientists using a distributed computing infrastructure. Despite some advances, the problem of finding suitable methods for the exploration of large datasets remains still formidable. The emphasis is on methods of bridging the gap between accurate representations and mining of the data and the capabilities of current technology and users. The analytical methods based partially on statistical random choices (crossover/mutation) and on knowledge experience acquired (supervised and/or unsupervised adaptive learning) could realistically achieve the discovery of hidden laws behind focused phenomena, often based on nature laws, therefore the simplest.
During the R&D phase of our project, aimed at define and characterize rules, targets, ontology, semantics and syntax standards, the functional breakdown structure outlined in Fig. 4 was derived:

Fig. 4 – The functional/modeling taxonomy for DAME

It provides a taxonomy between possible data exploration modes, made available by our infrastructure as data mining experiment typology (use case). In order to accelerate the infrastructure prototype development, in the first implementation phase we decided to reduce this scenario, by focusing the attention on Classification and Regression functionalities. Classification is a procedure in which individual items are placed into groups based on quantitative information on one or more features inherent to the items (referred to as features) and based on a training set of previously labelled items. A classifier is a system that performs a mapping from a feature space X to a set of labels Y. Basically a classifier assigns a pre-defined class label to a sample. Formally, the problem can be stated as follows: given training data {(x_1,y_1),...,(x_n, y_n)}, (where x_i are vectors), a classifier h:X->Y maps an object x ε X to its classification label y ε Y. Different classification problems could arise:

- crispy classification: given an input pattern x (vector) the classifier returns its computed label y (scalar).
- probabilistic classification: given an input pattern x (vector) the classifier returns a vector y which contains the probability of y_i to be the "right" label for x. In other words in this case we seek, for each input vector, the probability of its membership to the class y_i (for each y_i).

Both cases may be applied to both "two-class" and "multi-class" classification.
Regression is instead the supervised search for a mapping from a domain in R^n to a domain in R^m. One can distinguish between two different types of regression:

- data table statistical correlation: the user tries to find a mapping without any prior assumption on the functional form of the data distribution. Machine learning algorithms are well suited for this kind of regression;
- function fitting: with curve fitting the user tries to validate the hypothesis, suggested by some theoretical framework, that the data distribution follows a well defined, and known, function;

A regression system performs a mapping from a parameter space X to a target space Y. Formally, the problem can be stated as follows: given training data {(x_1,y_1),...,(x_n, y_n)} (where x_i are vectors) a regression operator h:X->Y maps an object x ε X to its value y ε Y.

The project infrastructure

By taking into account both theoretical and conceptual domains, we designed an infrastructure based on the following skill features:

- ✓ DAME Suite composed of several software components which relies on a common infrastructure (Fig. 5);
- ✓ Object Oriented Programming (OOP);
- ✓ Internal standards and protocols (VO, XML);
- ✓ Java language (almost generic for data mining models);
- ✓ User/Session DataBase Management System (MySQL);
- ✓ Web-based User I/O (Google Web Toolkit);
- ✓ Service-Oriented Web Application and Restful Web Service Technology, Servlet (Web Server applets);
- ✓ Plugin-based Modularity (easy to be integrated/modified) for data mining models;
- ✓ Hardware independent;
- ✓ Data conversion and manipulation support (ASCII, FITS, CSV, VOTable);

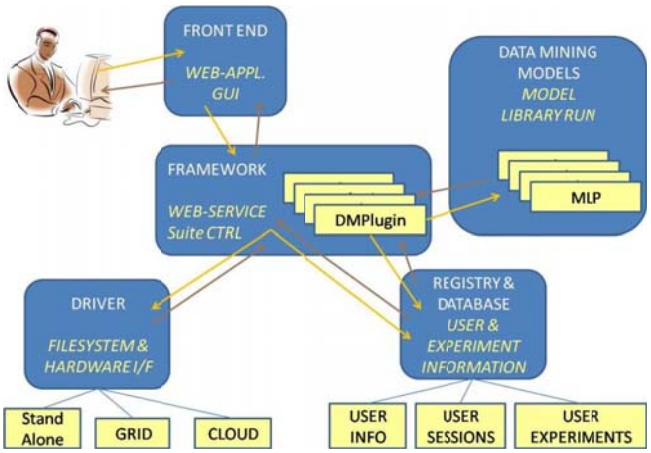


Fig. 5 – DAME architecture layout.

To separate workflow functional requirements from their implementation issues, a library of methods called DRiver (DR) Management System has been provided. The DR is the component used by the Framework (FW), core of the Suite, to manage the processing environment. It implements the low-level interface with computational environment, in order to permit the FW implementation, through the specific drivers, on different platforms (such as Stand-Alone or GRID). In other words, the DR is used to implement the proper access to the platform-dependent resources required by all the specific use cases and functionalities of the Suite (Fig. 6).

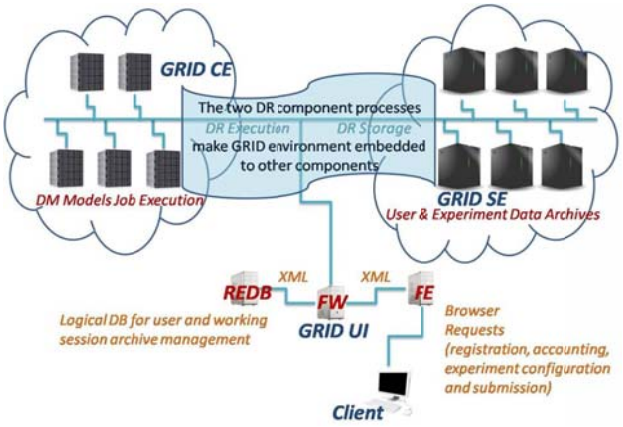


Fig. 6 – The infrastructure abstraction role of DR component in the DAME workflow.

In order to avoid multiple deployment of these data on several components, it was decided to provide a file storage system (SE or Storage Element on GRID), hosting real data files, handled directly by the FW through the DR component methods. This component includes also a library of data file format (FITS, ASCII, CSV and VOTable) translating methods, used by the FW depending on the specific DM model data format supported. The DMM is the component that implements the DM (Data Mining) models and related wrappers code and their use cases. Its main features can be summarized as follows:

1. Implementation oriented to the functionalities (i.e. classes of functionalities, such as Classification and Regression);
2. Possibility of use functionalities with more than one model without duplicate code (Pattern Bridge as standard design pattern);
3. A common interface for all the models (by means of a specific class rendering typical data mining self-adaptive parameters, called DMMParams);

In the first release, the DMM implements MultiLayer Perceptron (MLP) with standard Back Propagation learning rule, MLPGA (MLP & Genetic Algorithms) and Support Vector Machine (SVM) as supervised models. All these models have a common data mining paradigm: the AI (Artificial Intelligence) technique as self-adaptive exploration methodology.
The experiment that user can configure through our suite is in practice made of the choice between available functionalities and related DM model.

Conclusions

DAME project, an evolution of VO-Neural working group, comes out as an astrophysical data exploration tool, originating from the very simple consideration that, with data obtained by the new generation of instruments, we have reached the physical limit of observations (single photon counting) at almost all wavelengths. If extended to other scientific or applied research disciplines, the opportunity to gain new insights on the knowledge will depend mainly on the capability to recognize patterns or trends in the parameter space, which are not limited to the 3-D human visualization, from very large datasets. In this sense DAME approach can be easily and widely applied to other scientific, social, industrial and technological scenarios. Our project has recently passed the R&D phase, de facto entering in the implementation commissioning step and by performing in parallel the scientific testing with first infrastructure prototype, accessible, after a simple authentication procedure, through the official project website address (http://voneural.na.infn.it). First scientific test results confirm the goodness of the theoretical approach and technological strategy. Please consult the science production page (http://voneural.na.infn.it/science_papers.html) for more information.
Next milestone is the release, foreseen at the beginning of next December, of the Web Application (complete Suite), actually under final deployment on the S.Co.P.E. GRID infrastructure and testing.

Who's who in DAME

DAME is funded by the Italian Ministry of Foreign Affairs as well as by the European project VOTECH (Virtual Observatory Technological Infrastructures, http://www.eurovotech.org/) and by the Italian PON-S.Co.P.E. (http://www.scope.unina.it).

Official partners of the project are:

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

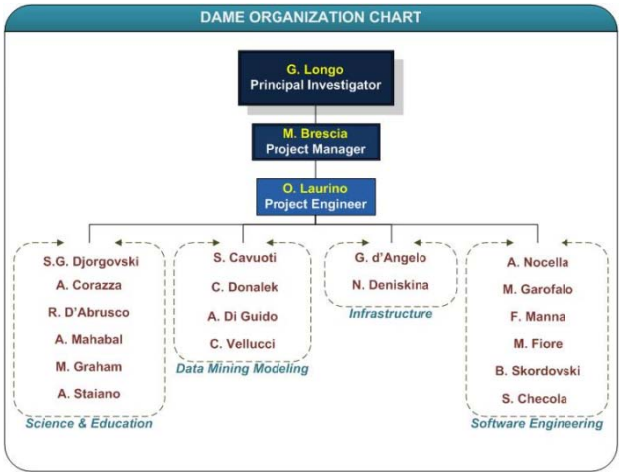


Fig. 7 – DAME team organization chart

First Technological Products

M. Brescia

The official Web Application implementing the complete set of features of DAME Suite is currently under deployment and test on the GRID S.Co.P.E. platform. In the meanwhile, a Java-based plugin wizard for custom experiment setup (hereinafter called DMPlugin) and a Web Application Prototype are already available. They are described in the following.

The DMPlugin Wizard (developed by M. Fiore)

The sub-component DMPlugin has been designed to extend DAME Suite features with user own algorithms to be applied to scientific cases by encapsulating them inside the Suite. The complete software package, together with user manual, can be downloaded from the address http://voneural.na.infn.it/dmplugin.html.
This facility extends the canonical use of the Suite: 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.
If you are not considering yourself as a simple user, 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). So you want to extend our framework? In other words, the user wants to become a DM Models Developer but without any need of programming experience behind our package.
In this case the following is the basic procedure:

- 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. Just implement the Driver Plugin Interface and register it;

How to customize your own application

There is a specific GUI (Fig. 8), called **Application Wizard**, providing a step-by-step wizard able to build and generate user own custom data mining application, without the need to write source code or to know internal mechanisms inside the Suite. After creation, the user can choose if to share or not his generated application with the rest of community through DAME gateway.

Fig. 8 – the java-based DMPlugin Wizard

Classification and/or Regression are the functional use cases that can be chosen by configuring the wizard. The development of a DMPlugin for your application comes in two different steps: developing a constructor and overriding the run method.
In the available user manual there is a step-by-step procedure showing how to create a simple DMPlugin for your application.

The Web Application Prototype (developed by O. Laurino)

The prototype is a Web Application implementing minimal DAME features and requirements, developed in parallel with the project advancement in order to perform a scientific validation of models and algorithms foreseen in the main Suite and to verify all basic project features designed related to the scientific pipeline workflow. It has been implemented as a Python web application and is publicly accessible at http://www.dame.na.infn.it

Fig. 9 – DAME Prototype home page

The prototype, (Fig. 9) implements the basic user interface functionalities: a virtual file store for each registered user is physically allocated on the machine that serves the web application. Users can upload their files, delete them, visualize them: the system tries to recognize the file type and shows images or text contextually. Three kinds of experiments can be launched at the moment: Multi-Layer Perceptron and Support Vector Machines can be used for classification and regression supervised tasks. The third one is a use case specifically developed to serve the results of the current thesis and to show how our method works. So, photometric redshifts for a user provide catalogue can be determined by using the networks trained.
Any astronomical data analysis and/or data mining experiment to be executed on the prototype, can be organized as a data processing pipeline, in which the use of the prototype needs to be integrated with pre and post processing tools, available between Virtual Observatory Web Services. One example is the use of prototype in combination with **Topcat** (http://www.star.bris.ac.uk/~mbt/topcat/) in order to prepare experiment dataset and to obtain plots and diagrams for prototype output evaluation. The next section will introduce you an example of such complete scientific experiment pipelines, organized as a tutorial.

Scientific Use Case Tutorial

R. D'Abrusco, O. Laurino & DAME team

During the project development, some scientific tests have been carried out on the data mining models implemented and provided as internal methods of the DAME Suite package. These tests have been based on astrophysical science cases, chosen between experiments already approached and partially solved by DAME science team in the recent past. These science cases obtained the double positive effect: a scientific validation of the methods and models made available through the Suite and a set of use case templates that can guide, as practical examples, novel users to the intensive and correct use of the Suite application for their own scientific experiments.
In this last context, a series of scientific use cases will be described here and in the next editions of the newsletter. Let's start with the first use case.

Photometric Redshift Estimation

Photometric redshifts have become one of the main tools to investigate the spatial distribution of galaxies, since they are necessary to reconstruct the 3-dimensional position of very large number of sources using only their photometric properties. One application of z-phot's are amazing maps of the Universe like showed in Fig. 10.

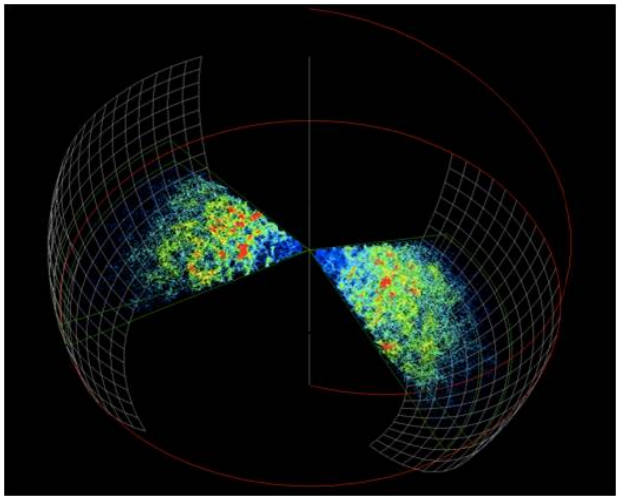


Fig. 10 – zphot map of the Universe

The mechanism responsible of the correlation between the photometric features and the redshift of an astronomical source, is the change in the contribution to the observed fluxes caused by the prominent features of the observed spectrum continuum and line emission components shifting through the different filters of the photometric system as the spectrum of the source is redshifted.
One family of methods for photometric redshift estimation is called "empirical" since these methods can be applied only to "mixed surveys", i.e. to datasets where accurate multiband photometric observations for a large number of source are supplemented by spectroscopic redshifts for a smaller but still significant subsample of the same sources, representative from a statistical point of view of the parent population. These spectroscopic data are used to constrain the fit of an interpolating function mapping the photometric parameter space; different specific methods differ mainly in the way such interpolation is performed. Neural networks (NN), among other machine learning algorithms, are very efficient at recognizing relations between data (see the web page http://voneural.na.infn.it/mlp.html for more details on NN and on the Multi Layer Perceptron, one of the most common models of NN), and in the "training phase" they need a set of "examples" to learn efficiently how to reconstruct the relation between the "parameters" and the "target". In the specific case of photometric redshifts, the parameters are fluxes, magnitudes or colours of the extragalactic sources while the targets, an independent and reliable estimate of the quantity the NN are trained to evaluate, are the redshifts of the sources measured from their observed spectra.
An example of scientific application of the method and algorithms described in this tutorial can be seen here at this link:
http://voneural.na.infn.it/vo_redshifts.html

Let's start. We will consider the general situation where the user (you!) has a file containing a table composed of M columns for R rows (each row represents a different source while each column containing a different quantity). N of such columns contain the photometric data used as "parameters" for the NN (fluxes, magnitudes or colours), while one of the columns contains the spectroscopic redshifts ("target" of the training).
If you don't want to care about the preprocessing or you don't own a suitable file to create the two datasets you need to complete the experiment, skip the "Dataset section" of this tutorial and just download the reference files we have prepared for you from the link
(http://voneural.na.infn.it/voneural_science.html):
These files, named "dataset_train.dat" and "dataset_test.dat", are ASCII files with 5 and 4 columns respectively. The "dataset_train.dat" file will be used in the "training phase" of the experiment: the first 4 columns represent the observed colours of the galaxies while the last contains the spectroscopic redshift, taken from the SDSS DR7 archive. The "dataset_test.dat" file contains instead only the photometric data of a different sample of galaxies (the column containing the spectroscopic redshift is missing) drawn from the same parent sample of "dataset_train.dat", and it will be used in one of the last steps of the experiment to produce a catalogue of photometric redshifts.
Otherwise, if you are brave enough to proceed with your own data, read the following section.

Dataset preparation:

1) Load the file with Topcat
(http://www.star.bris.ac.uk/~mbt/topcat/):
Launch **Topcat** -> **Open a new table** -> Select the format of the file from the **Format** menu and the location of the file clicking on **Filestore browser** -> **Ok**;
2) Inspect the content of the file and select the columns to be used with Topcat: **Display Column Metadata** -> unselect all columns by clicking on

Make all table columns invisible -> select only the columns containing the magnitudes and the spectroscopic redshift by ticking the corresponding checkboxes in the **Visible** column (at this point, this table should contain N + 1 visible columns);

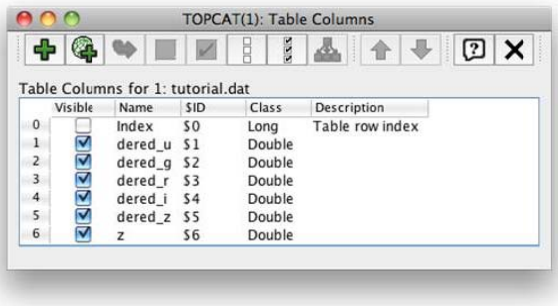


Fig. 11 – Topcat: table columns handling window

3) Close the **Table Columns** window -> **Table Browser**: rearrange the order of the columns so that the last column is the column containing the spectroscopic redshift by dragging and dropping the redshift column in the right place -> close the **Table Browser** window;

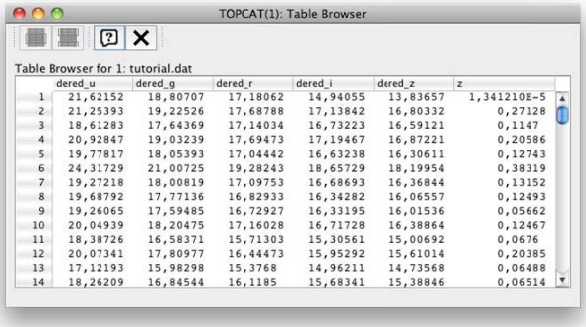


Fig. 12 – Topcat: table browser

4) Open the **Row subsets** window -> split the current table in two different samples by clicking on **New Subset from first Rows** and filling the **Row Count** field with a number approximately equal to the 70% of the total number of rows in your file, then click **OK** -> in the **Row Subsets** windows you will find a table called "head_numrows" where 'numrows' is the number you have used, containing the first 'numrows' sources of the original file. Select it and then click on **Create new subset complementary to selected subset** -> a new subsample of the original file, called "not_head_numrows", the complement to the "head-numrows" sample, will appear -> close the **Row Subsets** window;

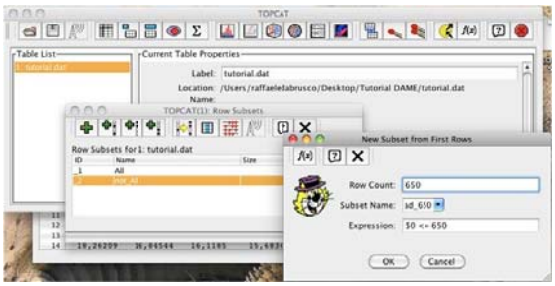


Fig. 13 – Topcat: Row subset window

5) Select the "head_numrows" subsample from the **Row Subsets** menu in the main window -> save the table in a file with **Save Table**: select ASCII as format in the **Output format** menu and choose a location and a name (train.dat is good enough!) for the file by clicking on **Filestore Browser;**

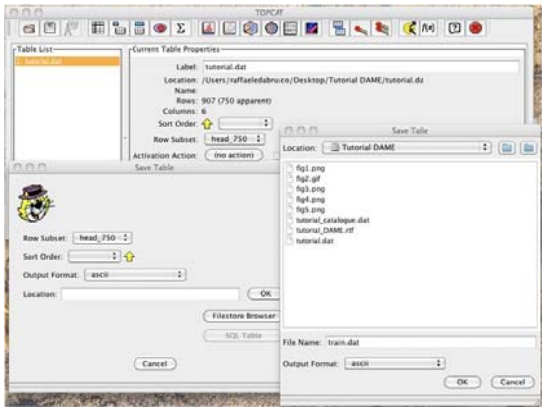


Fig. 14 – Topcat: saving table window

6) This time, select the "not_head_numrows" subsample from the **Row Subsets** menu in the main window -> inspect the content of the file and select the columns to be used with Topcat: **Display Column Metadata** -> unselect the last column by unticking the checkbox of the column containing the the spectroscopic redshifts (at this point, this table should contain with N visible columns) -> save the table in a different file with **Save Table**: select ASCII as format in the **Output format** menu and choose a location and a name (run.dat could be a good choice) for the file by clicking on **Filestore Browser;**

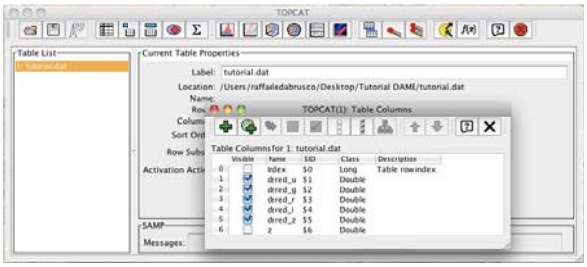


Fig. 15 – Topcat: table column handling window

7) You're done with Topcat (for now) and your input datasets for DAME are ready!

Now you have two different files, either you have chosen to use the ones we prepared for you either you decided to make them by yourself and it's time for you to get acquainted with DAME! DAME will show to you as a webpage whence you can register, upload your files, choose the algorithm and run the experiments, and much more!

Subscription to DAME

1) Go to the webpage http://**dame.na.infn.it** (you can also access the prototype from the DAME project official website (http://voneural.na.infn.it);
2) If you are a new user click on the link **Sign Up!**, otherwise jump directly to point 3 of this section of the tutorial;
    1. Create you account by filling the fields in this webpage: you have to provide your name, choose a username, a valid e-mail address (you can access, of course…) and a password (twice). Click on **Register**;

Fig. 16 – DAME Prototype: registration page

    2. A confirmation email is sent shortly to the address you provided. Click on the link contained in the e-mail in order to activate your DAME account;
    3. Click on the **log in** link that appears in the confirmation page;
3) Provide your **Username** and **Password** in the corresponding fields and click on **Log In**;
4) You have successfully registered as DAME user and this step of the tutorial is over!

Now it's time to launch your experiment! First of all, the NNs needs to learn to recognize the functional relation between the photometric data and spectroscopic redshifts of the sources contained in the "train" input file, in order to be able to approximate it afterward and produce good estimates of the photometric redshift for the parameters values contained in the "run" file. In other words, the NNs need to be trained.
This is what is described in the following section of the tutorial:

Training of the Neural Networks

1) Go to the webpage **http://dame.na.infn.it** and log in, if you have not done it yet;
2) The first page (session page in Fig. 17) you will see contains two sections: **My Experiments**, which will list your experiments as you perform them, and **My Filestore** where you will see the files you upload and/or are produced during the experiment. Upload the files "train.dat" and "run.dat" by selecting them clicking **Scegli Documento** and then clicking on **Click here to upload the file**;

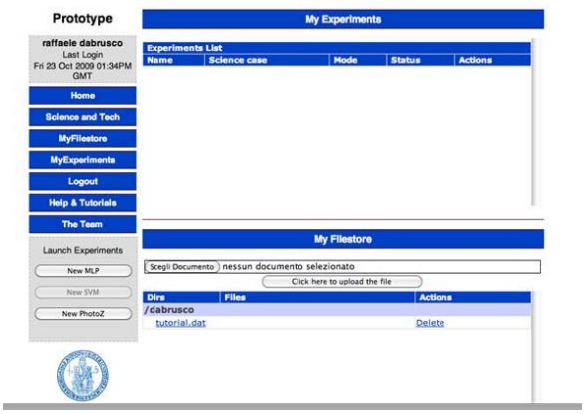


Fig. 17 – DAME Prototype: user session page

3) Click on **New MLP** button on the left of the page -> select **Regression** from the **Science Case** menu -> select **Full (Train + Test)** from the **Mode** menu -> click on **Go!**;

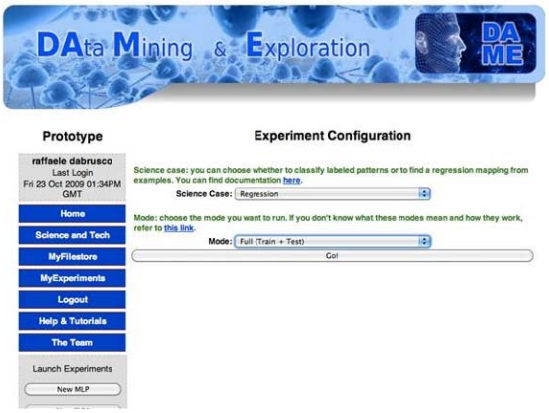


Fig. 18 – DAME Prototype: experiment configuration page

4) In the next page, choose a name and fill the **Experiment name** field -> In the **Input nodes** field is required the number of parameters (photometric information, i.e. magnitudes or colours or fluxes) of the "train" file -> **Hidden nodes**: it depends on the experiment, but usually 20 nodes are fine with almost all kind of experiments -> set 1 as **Output nodes** (the number of output nodes of the NN is always equal to the number of target, in this case the redshift) -> set **Max Epochs** to 500 (but can vary) -> set **Tolerance** to 0.001 (but can vary) -> choose "**MSE-INCREMENTAL**" from the **Training algorithm** menu -> select your "train" file for the following **Training set**, **Validation set** and **Test set** menus -> tick the checkbox **Do validation** and select the same "train" file in the **Validation set** and **Test set** menus-> click the **Go!** button;

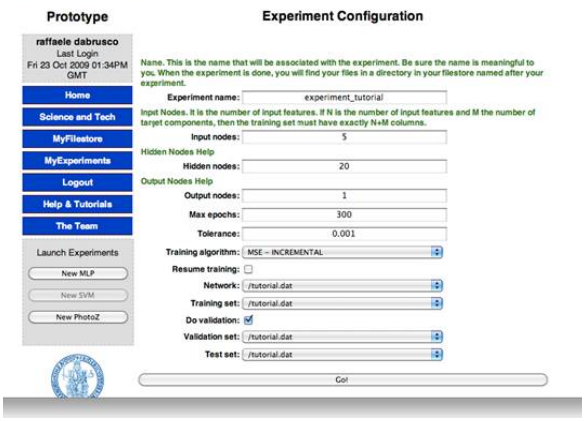


Fig. 19 – DAME Prototype: parameter configuration page

5) The next page will show you in real-time what is happening to your experiment: the status (Launched, Started, Finished), a summary of the details of the experiment, the list of files produced during the experiment, the log and graphics;

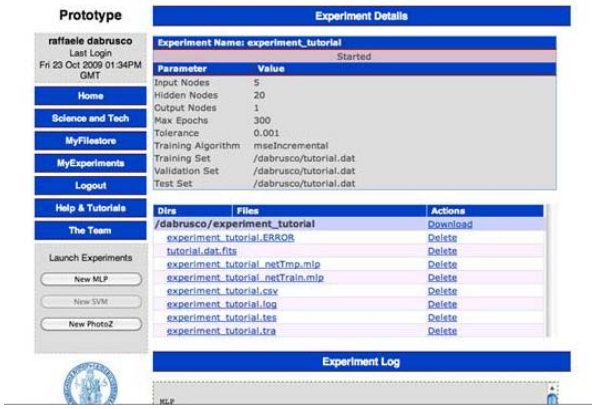


Fig. 20 – DAME Prototype: execution page

6) Now you can evaluate the results of the experiment: click on the link **MyFilestore** on the left of the page -> click on the **Donwload** link on the left of your experiment: a folder containing all files created during the experiment will be downloaded to your local drive;

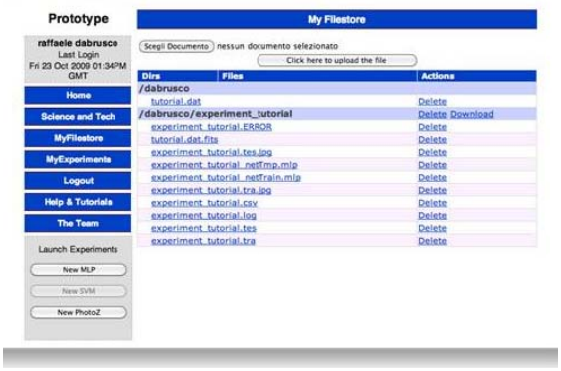


Fig. 21 – DAME Prototype: Experiment filestore page

7) open **Topcat** -> following the same steps described before, load the file ending with ".tes" in csv format (containing two columns: the spectroscopic redshift - target - and the photometric redshifts for a subsample - test set - of the original list of sources) -> click on **Scatter plot**: a scatter-plot of the spectroscopic redshifts vs photometric redshifts for the test set is produced (you may want to estimate the accuracy of the photometric redshifts using Topcat);

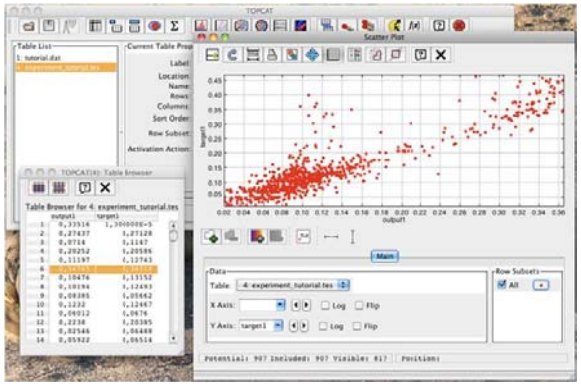


Fig. 22 – Topcat: Experiment output plot page

8) If the reconstruction of the photometric redshifts satisfies you, you have completed the training of the NNs. Compliments!

And now? When you embarked in this experiment, you probably wanted to calculate the photometric redshifts for a sample of galaxies for which you do not have spectroscopic redshift (in this case, our "run.dat" file), i.e. you wanted to create a catalogue of photometric redshifts. That is what our trained NN is for! For the last step of the process, read the next section of the tutorial.

Creating a catalogue of photometric redshifts

1) Go to the webpage http://**dame.na.infn.it** and log in into your account;
2) Click on **New MLP** button on the left of the page -> select **Regression** from the **Science Case** menu -> select **Run** from the **Mode** menu -> click on **Go!**;

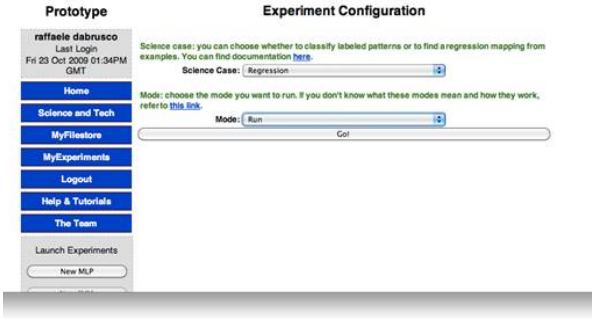


Fig. 23 – DAME Prototype: Experiment Run use case selection page

3) Fill the **Experiment name** field with a suitable name ('catalogue_exp') -> in the **Network** menu, select the file containing saved NN of your experiment (look for the file ending with "_netTrain.mlp" inside folder named as the name of your experiment, this is the saved network) -> select the second file you uploaded (the 'run.dat' file) in the **Data set** menu -> click **Go!**;

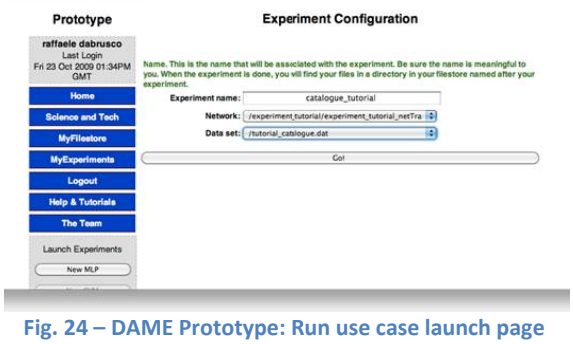


Fig. 24 – DAME Prototype: Run use case launch page

4) Your catalogue is being created. When the status of the experiment is 'Finished', you can visualize on screen the content of each output file just clicking on its name, delete it using the **Delete** button or download them all by clicking on **Download**;

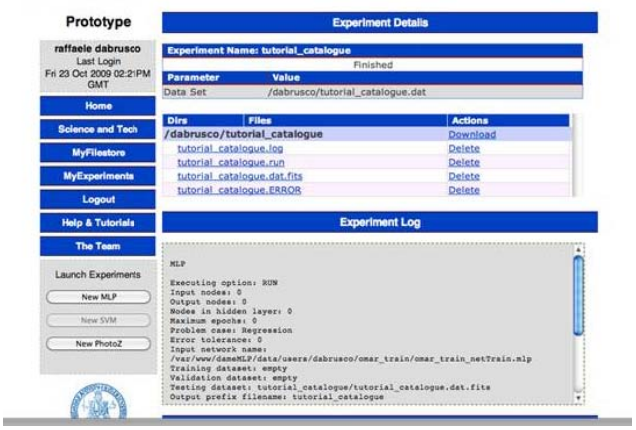


Fig. 25 – DAME Prototype: Experiment result page

5) Download the whole folder created during the experiment, photometric redshifts are contained in the file with extension ".run". Now you can enjoy your brand new photometric redshifts!

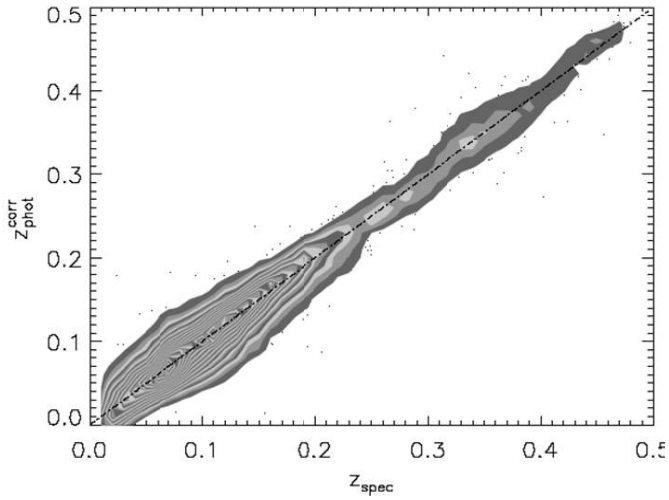


Fig. 26 – Trend of spectroscopic versus photometric redshifts for the Main Galaxy sample

Results on the Science use case

A more complex workflow, based on the application of the NN used in this tutorial, has been employed to obtain a catalogue of photometric redshifts for all photometric galaxies observed in the SDSS DR7. As base of knowledge, i.e. training set of the NNs, a subsample of galaxies for which spectroscopic redshifts from SDSS are available, has been used.

More details on results and techniques employed can be found in papers, like D'Abrusco et al. 2007, at http://voneural.na.infn.it/science_papers.html and at the page http://voneural.na.infn.it/vo_redshifts.html

DAta Mining & Exploration

Technical Support

- helpdame@gmail.com
- Skype helpdesk: help dame

Official Websites

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