Newsletter
Edition
03
April
2012

DAta Mining & Exploration

in this edition

A New trend in E-Science

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 program aimed at designing and developing tools for scientific data mining, based on state of the art of Information and Communication Technology.

Web Solutions for Massive Data Mining

Nowadays, many scientific areas share the same need of being able to deal with massive and distributed datasets and to perform on them complex knowledge extraction tasks. This simple consideration is behind the international efforts to build virtual organizations such as, for instance, the Virtual Observatory (VObs). DAME (DAta Mining & Exploration) is an innovative, general purpose, Web-based, distributed data mining infrastructure specialized in Massive Data Sets exploration with machine learning methods. Initially fine tuned to deal with astronomical data only, DAME has evolved in a general purpose platform program, hosting a cloud of applications and services useful also in other domains of human endeavor. DAME is an evolving platform and new services as well as additional features are continuously added. The modular architecture of DAME can also be exploited to build applications, finely tuned to specific needs. The project represents what is commonly considered an important element of e-science: a stronger multi-disciplinary approach based on the mutual interaction and interoperability between different scientific and technological fields (nowadays defined as X-Informatics, such as Astro-Informatics). Such an approach may have significant implications in the Knowledge Discovery in Databases process, where even near-term developments in the computing infrastructure which links data, knowledge and scientists will lead to a transformation of the scientific communication paradigm and will improve the discovery scenario in all sciences..

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://dame.dsf.unina.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***, the USA - ***National Science Foundation*** and ***PRIN-INAF*** "*Tomography of Galaxy Clusters*". DAME makes use of distributed computing environments (also supported by the S.Co.P.E. - GRISU infrastructure).

Newsletter
Edition
03
April
2012

The DAME Program Overview

In 2007, it was decided to coordinate under a common strategy and design a series of relatively independent scientific and technological experiments in data mining. In the meanwhile, the explosion of technology in the fields of digital processing, computer Science, high performance and distributed computing, astronomical telescopes and focal plane instrumentation, was imposing a new approach to render scientists, capable to explore in an efficient way the incoming "tsunami" of petabytes of data collected in worldwide distributed archives and data centres. A trend from which it has rapidly emerged what is now known as the fourth paradigm of Science: after theory, experimentation and simulations: namely data mining, or equivalently, Knowledge Discovery in Databases (KDD). This scenario convinced three groups (University Federico II in Napoli, the INAF Astronomical Observatory of Capodimonte in Napoli, and the Department of Astronomy at the California Institute of Technology) to pursue their scientific goals from a new, more efficient perspective. The idea was to create an infrastructure capable to merge in an homogeneous way the past scientific products with the state of the art of technology and astrophysics trends, and integrate them with additional features and capabilities in order to realize a product to be shared with the entire scientific community in an "*open and easy way*".

Where "open" means basically easily extendable in terms of functionalities and data mining models to be employed in the general astrophysics research and in data exploration at large.

The term *easy* requires instead a further discussion.

It is clear that modern experimental and observational science requires a good understanding of computer science, network infrastructures, data mining, etc. i.e. of all those techniques which fall into the domain of the so called e-science. Such understanding is almost completely absent in the older generations of scientists and this reflects in the inadequacy of most university programs. A paradigm shift in education is needed: statistical pattern recognition, object oriented programming, distributed computing, parallel programming need to become an essential part of scientific background.

So far, the DAME Program was conceived. DAME offers to a wide variety of e-science communities a large spectrum of computational facilities to exploit the wealth of available massive data sets and powerful machine learning and statistical algorithms. Furthermore, DAME hosts a variety of services developed by the collaboration for the specific needs of astrophysics research.

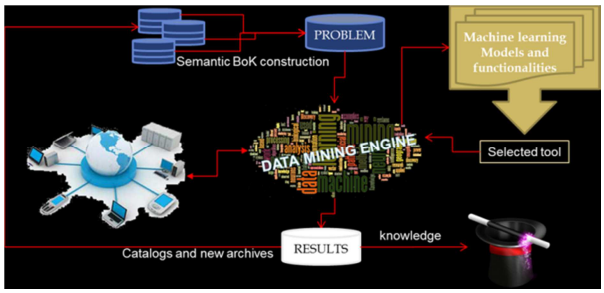


Fig. 1 – DAME data mining concept

DAME can be defined as a "SOUP": **S**ervice **O**riented **U**tilities **P**rogram. While the Service Oriented Utilities may be easily understood the term *Program* derives from the project management context language, and indicates that DAME is a suite (ensemble) of projects, sharing a unique technological platform, based on a Service Oriented Application architecture, and a shared goal, e.g. able to provide a general-purpose and cross-disciplinary computationally distributed environment for knowledge extraction and data mining in massive data sets.

By summarising the main concepts, e-science communities only recently started to face the deluge of data produced by new generation of scientific instruments and by numerical simulations (widely used to model the physical processes and compare them with measured ones). Now data is commonly organized in scientific repositories. Data providers have implemented Web access to their repositories, and http links between them. This *data warehouse network*, which consists of huge volumes of highly distributed, heterogeneous data, opened up many new research possibilities and greatly improved the efficiency of doing science. But it also posed new problems on the cross-correlation capabilities and mining techniques on these massive data sets (MDS) to improve scientific results. The most important advance we expect is a dramatic need of the ease in using distributed e-Infrastructures for the e-science communities. We pursue a scenario where users sit down at their desks and, through a few mouse clicks, select and activate the most suitable *scientific gateway* for their specific applications or gain access to detailed documentation or tutorials.

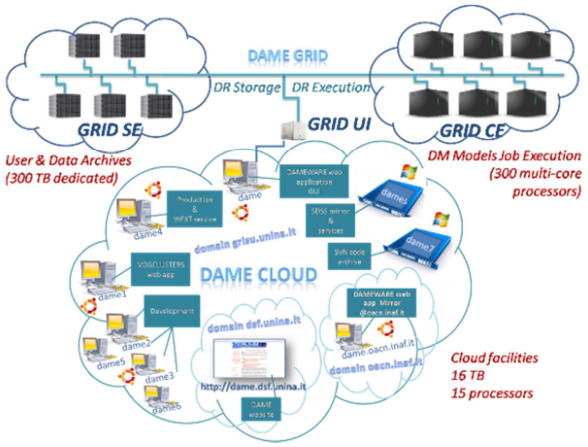


Fig. 2 – DAME distributed computing infrastructure

We call *scientific gateway* an e-Infrastructure which is able to offer remote access and navigation on distributed data repositories together with web services and applications able to explore, analyze and mine data. It doesn't require any software installation or execution on user local PC, it permits asynchronous connection and launch of jobs and embeds to the user any computing infrastructure configuration or management.

In this way the stakeholder communities (both academies and enterprises) will expand their use of the e-Infrastructure and benefit from a fundamental tool to undertake research, develop collaborations, and increase their scientific productivity and the quality of research outputs. Only if this scenario becomes reality, the barrier currently placed between the community of users and technology will disappear

DAME News

Exploratory AstroInformatics Initiative

The emerging field of AstroInformatics, while on the one hand appears crucial to face the technological challenges, on the other is opening new exciting perspectives for new astronomical discoveries through the implementation of advanced data mining procedures. The complexity of astronomical data and the variety of scientific problems, however, call for innovative algorithms and methods as well as for an extreme usage of ICT technologies. The DAME (DAta Mining & Exploration) Program exposes a series of web-based services to perform scientific investigation on astronomical massive data sets. The engineering design and requirements, driving its development since the beginning of the project, are projected towards a new paradigm of Web based resources, which reflect the final goal to become a prototype of an efficient data mining framework in the data-centric era. Under these perspectives there is currently active the ***Exploratory AstroInformatics Initiative***, in collaboration among DAME partners (FEDERICO II, INAF-OACN, CALTECH) and other Institutes and Corporations in USA. Main goal of such inter-disciplinary Initiative is to promote, support and encourage the intensive interoperable exploitation of Web 2.0, Machine Learning, data mining, Cloud and GPU-based parallel computing technologies to provide high qualitative and efficient computing infrastructures and services for a wide variety of scientific and social communities.

The AstroInformatics is one of the novel disciplines under the term X-Informatics (such as Bio, Geo etc.). **The Initiative started within the Astro-Informatics community, but it is extended to all other interested communities and ICT Companies which may contribute to identify common ICT solutions and strategies to perform science in the data-centric era. Therefore this initiative is open to other collaborations and partnerships**.

GPU based Parallel Data Mining

GPGPU is an acronym standing for General Purpose Computing on Graphics Processing Units. It was invented by Mark Harris in 2002, by recognizing the trend to employ GPU technology for not graphic applications. In general the graphic chips, due to their intrinsic nature of multi-core processors (many-core) and being based on hundreds of floating-point specialized processing units, make many algorithms able to obtain higher (one or two orders of magnitude) performances than usual CPUs. They are also cheaper, due to the relatively low price of graphic chip components. Particularly useful for super-computing applications, often requiring several execution days on large computing clusters, the GPGPU paradigm may drastically decrease execution times, by promoting research in a large variety of scientific and social fields (such as, for instance, astrophysics, biology, chemistry, physics, finance, video encoding and so on). Therefore, astronomers should perform a cost-benefit analysis

and some initial development to investigate whether their code could benefit from running on a GPU. Used in the right way and on the right applications, GPU's will be a powerful tool for astronomers processing huge volumes of data.
In this context we designed and developed a multi-purpose genetic algorithm (GAME) implemented with GPGPU and CUDA parallel computing technology. The model derives from the paradigm of supervised machine learning, addressing both the problems of classification and regression applied on massive data sets. Since GAs are embarrassing parallel, the GPU computing paradigm has provided an exploit of the internal training features of the model, permitting a strong optimization in terms of processing performances. The use of CUDA translated into a 75x average speedup, by successfully eliminating the largest bottleneck in the multi-core CPU code and making highly scalable the use of genetic algorithm on massive datasets.
Another important aspect of the work is the current investigation of data mining and statistical applications to evaluate and optimize the Internet network traffic with GPU based technology. **The work is done in collaboration with prof. G. Ventre and his team of the Department of Computer Engineering and Systems of the University Federico II of Naples.**

The Lernaean Hydra idea: Web Application Repositories

There are at least two reasons for not moving data over the network from their original repositories to the user's computing infrastructures. First of all the fact that the transfer could be impossible due to the available network bandwidth and, second, because there could be restrictive policies to data access. In these cases, the problem is to move the data mining toolsets to the data centers. Current strategies, under investigation in some communities such as the Virtual Observatory (VO), are based on implementing web based protocols for application interoperability which are only a partial solution, since they still require to exchange data over the web between application sites. Since the International Virtual Observatory Alliance (IVOA) Interop Meeting, held in Naples in 2011, we proposed a different approach[1]:

1. WA ←→ WA (plugin bi-directional exchange)
    a. All DAs must become Was;
    b. Unique accounting policy (Google and Microsoft like);
    c. To overcome MDS flow, applications must be plug & play (*e.g. any WAx feature should be pluggable in WAy on demand*);

The plugin exchange mechanism foresees a standardized web application repository cloud named "Lernaean Hydra" (from the name of the ancient snake-like monster with many independent but equal heads).
It consists in Web Application Repositories (WAR) of data mining model and tool packages, to be installed and deployed in a generic data warehouse. Different WARs may differ in terms of available models since any hosting data center might require specific kinds of data mining and analysis tools. If the WARs are structured around a pre-designed set of standards which completely describe their interaction with the external environment and application plugin and execution procedures, two generic data warehouses can exchange algorithms and tool packages on demand. On a specific request the mechanism starts a very simple automatic procedure which moves applications, organized under the form of small packages (some MB in the worst case), through the Web from a WAR source to a WAR destination, install them and makes the receiving WAR able to execute the imported model on local data. More refinements of the above mechanism can be introduced at the design phase, such as for instance to expose, by each WAR, a public list of available models, in order to inform other sites about services which could be imported. Such strategy requires a standardized design approach, in order to provide a suitable set of standards and common rules to build and codify the internal structure of WARs and of the data mining applications themselves, such as, for example, any kind of rules like Predictive Model Markup Language (PMML). These standards should be designed to maintain and preserve the compliance with data representation rules and protocols already defined and currently operative in a particular scientific community (such as the VObs in Astronomy).
In case of the above approach, no local computing power is required. ***Also smartphones can run DM***

[1] http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/InterOpMay2011KDD

applications. Then it descends the following series of requirements:

- Standard accounting system;
- No more MDS moving on the web, but just moving applications, structured as plugin repositories and execution environments;
- standard modeling of WA and components to obtain the maximum level of granularity;
- Evolution of existing architectures to extend web interoperability (in particular for the migration of the plugins);

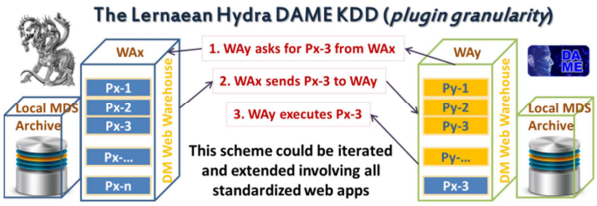


Fig. 3 – The main steps of the application plugins exchange mechanism among data mining web warehouses

After a certain number of such iterations the scenario will become:

- No different WAs, but simply one WA with several sites (eventually with different GUIs and computing environments);
- All WA sites can become a mirror site of all the others;
- The synchronization of plugin releases between WAs is performed at request time;
- Minimization of data exchange flow (just few plugins in case of synchronization between mirrors).

Any data center could implement a suitable computing infrastructure hosting the WAR and thus become a sort of mirror site of a world-wide cross-sharing network of data mining application repository in which it could be engaged a virtuous mechanism of a distributed multi-disciplinary data mining infrastructure, able to deal with heterogeneous or specialized exploration of MDS. **Such approach seems the only effective way to preserve data ownership and privacy policy, to enhance the e-science community interoperability and to overcome the problems posed by the present and future tsunami of data**. By following this approach, the DAMEWARE web application, described in the next section, represents a first prototype towards a WAR mechanism.

DAME Products

From the official website of the program (**http://dame.dsf.unina.it/**) the user has access to all available applications and services. In the following a summary of them is reported. Some of them are under R&D process. Our team, coming from the research world, is an open community, so far, any user interested to collaborate/support our projects is always welcome and encouraged to contact us.

DAMEWARE Web Application Resource

DAMEWARE is a web application, accessible through a simple web browser.
At the moment we have released the β-3 release, available at the following address:
http://dame.dsf.unina.it/beta_info.html

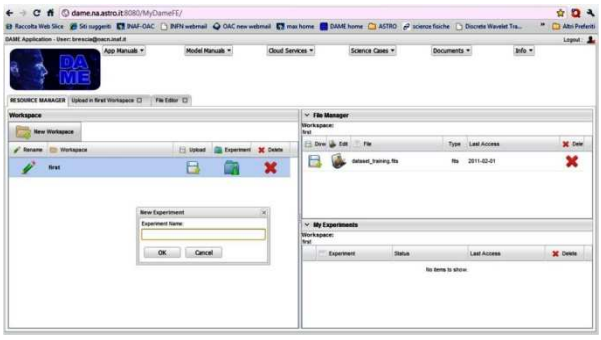


Fig. 4 – DAMEWARE Graphical User Interface

It is structurally organized under the form of working sessions (hereinafter named workspaces) that the user can create, modify and erase. You can imagine the application as a container of services, hierarchically structured as shown in the following figure.

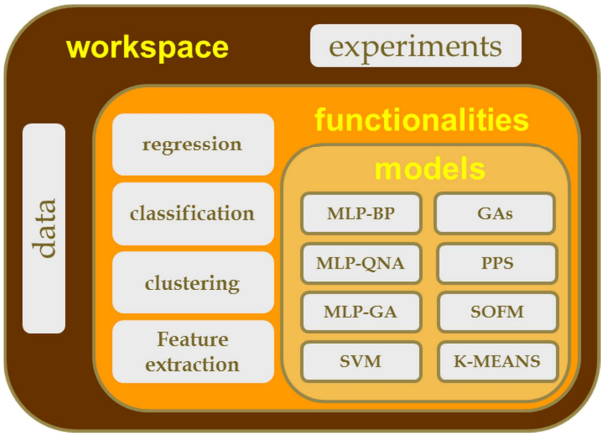


Fig. 5 – DAMEWARE internal architecture

The user can create as many workspaces as desired and populate them with uploaded data files and with experiments (created and configured by using the Suite). Each workspace is enveloping a list of data files and experiments, the latter defined by the combination between a functionality domain and a series (one at least) of data mining models.
The GUI permits to perform a complete workflow, having the following features:

- A workspace to envelope all input/output resources of the workflow;
- A dataset editor, provided with a series of pre-processing functionalities to edit and manipulate the raw data uploaded by the user in the active workspace;
- The possibility to copy output files of an experiment in the workspace to be arranged as input dataset for subsequent execution (the output of training phase should become the input for the validate/run phase of the same experiment);
- An experiment setup toolset, to select functionality domain and machine learning models to be configured and executed;
- Functions to visualize graphics and text results from experiment output;

More information and manuals can be found on the web page at the web address indicated above.
Recently it is included in the EUCLID Mission international program, under design phase, as a candidate toolset for data quality control and management.

VOGCLUSTERS Web Application

The VOGCLUSTERS Project is a web application (α release is available) specialized for data and text mining on globular clusters (GC). It is a toolset of DAME Program to manage and explore GC data in various formats.
http://dame.dsf.unina.it/vogclusters.html

STraDiWA Web-based workflow

The STraDiWA (Sky Transient Discovery Web Application) Project is a web service (at the moment under R&D) specialized for detecting variable objects (Supernovae, variable stars, etc.) from real or simulated images.

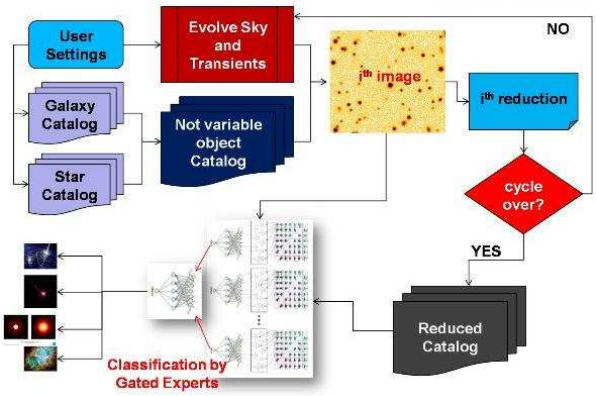
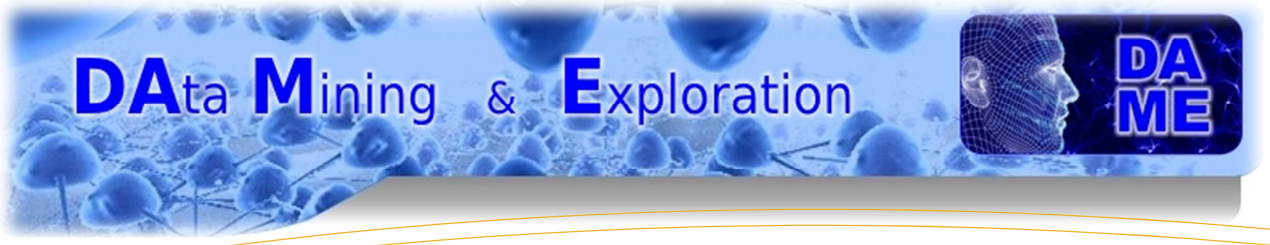


Fig. 6 – STraDiWA workflow

It includes also an automatic workflow to generate images with a user-defined number and type of variable objects, obtained by proper physical models, in order to perform setup of classification machine learning models running on the real images. Recently it is included in the EUCLID Mission international program, under design phase.
http://dame.dsf.unina.it/dame_td.html

DAME Web Services

As discussed before, DAME Program includes not only web applications but also several web services, dedicated to provide a wide range of facilities for different e-science communities. These services include:



- ✓ **SDSS Mirror Site**: The Sloan Digital Sky Survey mirror site, hosted by the GRID S.Co.P.E. Data Center, hosts and offers the complete Sky Survey Database, together with all tools and resources within the official SDSS website to generate scientific queries and to obtain data results. It provides also an image archive, covering a wide portion of southern sky and a subset of the sky object image archive, made available for the sky coverage overlapping KIDS survey project area. It is available at the following address http://dames.scope.unina.it/;

- ✓ **DAME-KappA**: the service DAME-KNIME application Assembly consists of a mechanism embedded and exposed by the DAMEWARE Web Application to make interoperable DAME and KNIME (http://www.knime.org/) platforms, aimed at make easy to build and execute data analysis and/or mining experiment pipelines by gaining in a unique toolset both features of the two projects. At the moment it is under R&D and its first release will be made available with the release 1.0 of DAMEWARE application, foreseen in next March;

- ✓ **DAME-WFXT time calculator**: The Wide Field X-Ray Telescope time calculator is a service allowing, although under simplistic assumptions and a preliminary mission design, to estimate the number of transient and variable sources that can be detected by WFXT within the 3 main planned extragalactic surveys, with a given significant threshold. It is available at the address: http://dame.dsf.unina.it/dame_wfxt.html;

DAME Science

As said in the introduction, DAME Program is specialized in applications and services for the astrophysical community, although data mining resources are in principle to be considered transversal to the particular e-science disciplines and widely portable in different enterprise/academy environments. In this context, current employment samples of our infrastructure covers astrophysical use cases, summarized in the following.

Photometric Redshifts (PHOTO-Z)

The photometric redshifts are essential in constraining dark matter and dark energy studies by means of weak gravitational lensing, for the identification of galaxy clusters and groups, for type Ia supernovae, and to study the mass function of galaxy clusters. The need for fast and reliable methods for photo-z evaluation will become even greater in the near future for the exploitation of ongoing and planned surveys. In fact, future large field public imaging projects, like KiDS (Kilo-Degree Survey[2]), DES (Dark Energy Survey[3]), LSST (Large Synoptic Survey Telescope[4]), and Euclid, require extremely accurate photo-z's to obtain accurate measurements that does not compromise the surveys scientific goals. The accuracy of the photometric redshift reconstruction, in general, is worse than spectroscopic redshifts but provide a more convenient way to estimate them. The physical mechanism responsible of the correlation between the photometric features and the redshift of an astronomical source mechanism implies a non-linear mapping between the photometric parameter space of the galaxies and the redshift values, which can be reconstructed using data mining methods.

PHOTO-Z - SDSS Quasars

In a very challenging task such as the Photometric redshifts for the Quasar Objects (QSO), for which the determination of the distance as a function of the observational parameters is complicated by the existence of higher degree of degeneracy, one of our methods, in the specific the Multi Layer Perceptron trained with Quasi Newton Algorithm (MLPQNA) obtained a very high accuracy. The sample used to train the algorithm started from the list of spectroscopic quasars from the SDSS[5] cross-matched with WISE[6], UKIDSS[7] and GALEX[8], obtaining a dataset

[2] http://www.astro-wise.org/projects/KIDS/
[3] http://www.darkenergysurvey.org/
[4] http://www.lsst.org/lsst/
[5] http://www.sdss.org/
[6] http://wise.ssl.berkeley.edu/
[7] http://www.ukidss.org/
[8] http://www.galex.caltech.edu/

with several bands from the ultra violet to the infrared. Moreover we performed a classification task, with the same MLPQNA model, in order to flag photo-z's with highest accuracy, which have shown an accuracy higher than 30%.

PHOTO-Z - SDSS Galaxies

Our machine learning methods have been used to produce a catalogue[9] of more than 6 million galaxies observed in the Sloan Digital Sky Survey (SDSS), with an unprecedented accuracy. The distinctive feature of this approach has been the application of a classification task before the training of the neural network, in order to separate two distinct classes of galaxies in the parameter space and partly to eliminate the degeneracy in the observational parameters. The classification has been performed using an MLP in a Bayesian framework, with a "logistic" output activation function and using the a-priori spectroscopic classification as targets.

PHOTO-Z - PHAT Contest

A significant advance in comparing different methods was introduced by Hildebrandt and collaborators **Errore. L'origine riferimento non è stata trovata.** with the so called PHAT[10] (PHoto-z Accuracy Testing) contest which adopts a black-box approach which is typical of benchmarking. Instead of insisting on the subtleties of the data structure, they performed a homogeneous comparison of the performances concentrating the analysis on the last link in the chain: the photo-z's methods themselves. However, the subsets used to evaluate the performances are still kept secret in order to provide a more reliable comparison of the various methods. Two different datasets are available. The first one, indicated as PHAT0, is based on a very limited template set and a long wavelength baseline (from UV to mid-IR). It is composed by a noise-free catalogue with accurate synthetic colors and a catalogue with a low level of additional noise. The second one, which is the one used in our work, is the PHAT1 dataset, which is based on real data originating from the Great Observatories Origins Deep Survey Northern field (GOODS-North).

[9] http://dame.dsf.unina.it/dame_photoz.html#sdss
[10] http://www.astro.caltech.edu/twiki_phat/bin/view/Main/GoodsNorth

The PHAT1 dataset covers the full UV-IR range and includes in 18 bands: U (from KPNO), B, V, R, I, Z (from SUBARU), F435W, F606W, F775W, F850LP (from HST-ACS), J, H (from ULBCAM), HK (from QUIRC), K (from WIRC) and 3.6, 4.5, 5.8 and 8.0 m (from IRAC Spitzer). While it is clear that the limited amount of object in the KB is not sufficient to ensure the best performances of most empirical methods, the fact that all methods must cope with similar difficulties makes the comparison very instructive. We performed our experiments with a MLPQNA achieving very competitive results with respect to other empirical methods

Globular Clusters Classification

The study of the GCs populations in external galaxies requires the use of wide-field, multi-band photometry and, in order to minimize contamination from fore/background objects and to measure some of the GC properties (size, core radius, concentration, binary formation rates), high angular resolution data are required. Our supervised learning experiment regarded the attempt to identify GCs in single band wide field images obtained with the Hubble Space Telescope for the galaxy NGC1399, using the base of knowledge ("true" GCs) provided by the multi-wavelength subset. The advantage being that single band data are much less expensive in terms of observing time, and thus easier to obtain than multi-band ones. The machine learning supervised model which obtained the best recognition performances was the MLPQNA. The best results led to a performance of 98.33%. **The experiments are performed through the DAMEWARE Suite and also optimized under the GPU + CUDA parallel computing environment**.
http://dame.dsf.unina.it/dame_gcs.html.

Active Galactic Nuclei Classification

The aim of this work was the selection of Active Galactic Nuclei (AGN), in terms of a minimal set of parameters embodying as closely as possible their physical differences (in this case, whether an AGN is contained or not). The classical methods for classifying galaxies, according to their central activity, involve time-consuming spectroscopic observations, which, as usual, even if very accurate, do not allow an extended exploration of the universe. DAME has developed a

method based on data mining for the classification of AGNs from their photometric data. Our method relies on a training set composed of sources which have been spectroscopically observed by the SDSS and classified according the Kewley, Kauffman and Heckman lines in the Baldwin, Philips and Terlevich (BPT) diagnostic plot, which is based on a set spectroscopic diagnostics derived by the measured intensities of emission lines in the spectra of the galaxies. The goal of the experiments was conservative since we were not interested in completeness, but rather in minimizing the fraction of false positives. Three experiments were performed:

1. Experiment 1: classification AGN/non-AGN;
2. Experiment 2: Type 1 AGN/Type 2 AGN;
3. Experiment 3: Seyfert galaxies/liners.

The results for the Experiment 1, the most important, can be summarized as it follows: we obtain a completeness of about 87% for the non-AGN class (hence, in the worst case, only 13% of the objects contaminate the purity of the AGN list). However, it must be noticed that in the “not-AGN” class, fall both confirmed not-AGN (i.e. laying below the Kauffman line) and objects in the so called mixing zone (above the Kauffman line and below the Kewley line), while sure not-AGNs are all below the Kauffman line.
If we take this distinction into account, the MLP classifies as AGN less than 1% of the confirmed not-AGNs (i.e. false positives). While, if the classifier is fed with only a list of confirmed not-AGNs (i.e. excluding objects in the mixing zone), just 0.8% turns out to be a false positive. These results represent a convincing test of the application of data mining algorithms to the problem of photometric classification of galaxies.

Quasars Classification

One classical method for the selection of QSOs employs spectroscopic observations, which are available in limited number because they are time consuming and are possible only for high signal to noise ratios. Much larger samples of QSOs can be obtained by selecting candidate QSOs directly from photometric data and using a sample of QSOs for which the photometric and spectroscopic features are both measured as BoK. The DAME collaboration has developed an algorithm for the extraction of candidate QSOs from photometric data consisting in unsupervised clustering inside the photometric parameter space of star-like sources and which exploits, as labels of the sources of the base of knowledge (BoK), spectroscopic classification. The overall approach of the method is the following: the distribution of sources belonging to the BoK in the photometric parameter space is partitioned in separate groups of nearby sources by determining a clustering in the photometric parameter space. Such clustering is optimal in the sense that it maximizes the total separation between confirmed QSOs and other sources. Once determined, new candidate QSOs are extracted from a photometric dataset by associating each photometric source to the closest cluster (where distance is calculated by the Mahalanobis' distance, which takes into account the anisotropy of the distribution of members of the clusters along the axis of the parameter space) and by considering as candidates those associated to the clusters containing mostly confirmed QSOs of the BoK. The procedure for the determination of the optimal clustering involves two different algorithms, the Probabilistic Principal Surfaces (PPS) and Negative Entropy Clustering (NEC) methods. Both algorithms do not require any a-priori hypothesis regarding the nature of the underlying distribution of QSOs, except for the initial number of pre-clusters produced by the PPS; it has been empirically proved that the final results do not depend on this parameter though, given it is "large" relative to number of sources in the BoK.

NExt-II Image Segmentation Workflow

NExt-II is a software workflow obtained from the re-engineering of original NEXT package, based on unsupervised neural network employment to extract catalogues of objects through a segmentation process of multi-band astronomical images.

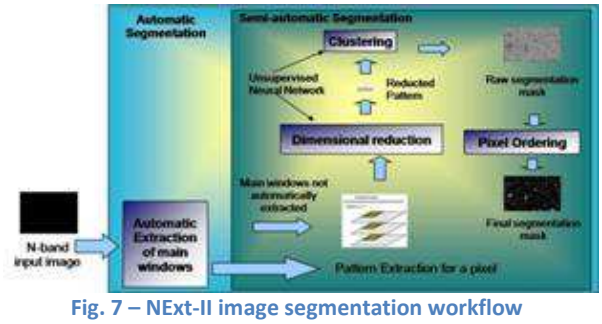


Fig. 7 – NExt-II image segmentation workflow

The work was conceived for the segmentation of astronomical images in signal/background pixels, in order to detect, classify objects and extract their catalogues. At the moment, the re-engineering has been closed and the web GUI interface package is under development.
http://dame.dsf.unina.it/next.html.

Euclid ESA Mission

Euclid is an all-sky space mission designed to map the geometry of the dark Universe. It has been recently approved by the ESA as one of the primary space missions of next decade. Its primary objectives are the study of dark energy, dark matter, modified gravity and cosmic initial conditions. Recently, DAME program has been officially included into the mission consortium, as one of the Science Teams in charge of data quality control, photometric redshift evaluation and variable object (sky transients) detection and classification. http://sci.esa.int/**euclid**/.

Recent Publications in DAME

"Extracting knowledge from massive astronomical data sets", Brescia, M., Cavuoti, S., Djorgovski, G.S., Donalek, C., Longo, G., Paolillo, M., **arXiv:1109.2840v1**, to appear in Astrostatistics and data mining in large astronomical databases, L.M. Barrosaro et al. eds, Springer Series on Astrostatistics, 15 pages, **2011**

"DAME: A distributed data mining and exploration framework within the Virtual Observatory", Brescia M., Cavuoti S., D'Abrusco R., Laurino O., Longo G., in Remote Instrumentation for eScience and Related Aspects, F. Davoli et al. (eds.), Springer Science+Business Media, LLC 2011, DOI 10.1007/978-1-4614-0508-5_17, **2011**

"New Trends in E-Science: Machine Learning and Knowledge Discovery in Databases", Brescia M., invited contribution to the Volume "Horizons in Computer Science Research", Thomas S. Clary (eds.), Series "Horizons in Computer Science", Nova Science Publishers, ISBN: 978-1-61942-774-7, **2011**

"The DAME/VO-Neural infrastructure: an integrated data mining system support for the science community", Brescia, M.; Corazza, A.; Cavuoti, S.; d'Angelo, G.; D'Abrusco, R.; Donalek, C.; Djorgovski, S.G.; Deniskina, N.; Fiore, M.; Garofalo, M.; Laurino, O.; Longo, G.; Mahabal, A.; Manna, F.; Nocella, A.; Skordovski, B., Proceedings of FINAL WORKSHOP OF GRID PROJECTS, PON RICERCA 2000-2006, CALL 1575, Catania, Italy, **2011**

"VOGCLUSTERS: An example of DAME Web Application", Castellani, M., Brescia, M., Mancini, E., Pellecchia, L., Longo, G., Proceedings of Conference "Data Analysis in Astronomy", Erice Sicily, Italy, April 15-22, (http://arxiv.org/abs/1109.4104), **2011**

"DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases", Brescia M., Longo G., Castellani M., Cavuoti S., D'Abrusco R., Laurino O., 54th SAIT Congress "L'astronomia italiana: prospettive per la prossima decade", Astronomical Observatory of Capodimonte, Napoli, May 6, Mem. S.A.It. Suppl. Vol. 19, 324, **2012**

"Astroinformatics, data mining and the future of astronomical research", Brescia, M., Longo, G., To appear in the Proceedings of the 2-nd International Conference on Frontiers in Diagnostic Technologies, Nuclear Instruments and Methods in Physics Research A, NIMA Elsevier Journal, ENEA Laboratori Nazionali di Frascati, Rome, Italy, November 28-30, (http://arxiv.org/abs/1201.1867), **2012**

"Astronomical Images and Data Mining in The International Virtual Observatory Context", Pasian, F., Brescia, M., Longo, G., Astronomical Images and Data Mining in the International Virtual Observatory Context, Science - Image in Action. Edited by Bertrand Zavidovique (Universite' Paris-Sud XI, France) and Giosue' Lo Bosco (University of Palermo, Italy). Published by World Scientific Publishing Co. Pte. Ltd., ISBN 9789814383295, pp. 230-240, **2012**

"DAME: A Web Oriented Infrastructure for Scientific Data Mining and Exploration", Cavuoti, S., Brescia, M., Longo, G., Garofalo, M., Nocella, A., Astronomical Images and Data Mining in the

International Virtual Observatory Context, Science - Image in Action. Edited by Bertrand Zavidovique (Universite' Paris-Sud XI, France) and Giosue' Lo Bosco (University of Palermo, Italy). Published by World Scientific Publishing Co. Pte. Ltd., 2012. ISBN 9789814383295, pp. 241-247, **2012**

"*DAta Mining and Exploration (DAME): New Tools for Knowledge Discovery in Astronomy*", Djorgovski, S. G.; Longo, G., Brescia, M., Donalek, C., Cavuoti, S., Paolillo, M., D'Abrusco, R., Laurino, O., Mahabal, A., Graham, M., American Astronomical Society, AAS Meeting #219, #145.12, Tucson, USA, January 08-12, (http://adsabs.harvard.edu/abs/2012AAS...21914512D, **2012**

"*Genetic Algorithm Modeling with GPU Parallel Computing Technology*". Cavuoti, S.; Garofalo, M.; Brescia, M.; Pescape', A.; Longo, G.; Ventre, G., 22nd WIRN, Italian Workshop on Neural Networks, Vietri sul Mare, Salerno, Italy, May 17-19, **2012**

"*Data mining and Knowledge Discovery Resources for Astronomy in the Web 2.0 Age*". Cavuoti, S.; Brescia, M.; Longo, G., Proceedings of SPIE Astronomical Telescopes and Instrumentation 2012, Software and Cyberinfrastructure for Astronomy II, Volume 8451, RAI Amsterdam, Netherlands, July 1-4, **2012**

"*The detection of Globular Clusters in galaxies as a data mining problem*", Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T., Monthly Notices of the Royal Astronomical Society, Volume 421, Issue 2, pp. 1155-1165, available at http://arxiv.org/abs/1110.2144, **2012**

"*Photometric Redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 Contest*", Cavuoti, S.; Brescia, M.; Longo, G.; Mercurio, A., Submitted to Astronomy & Astrophysics (reference number: AA/2012/19755), available at http://arxiv.org/abs/1206.0876, **2012**

The following is the list of publications, under the DAME umbrella, currently <u>in preparation</u>:

"*High Accuracy Photometric Redshifts for Quasars*", Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A. (MNRAS);

"*Inside Astrophysical Catalogues: A Comparison among Source Extraction Algorithms*", Annunziatella M.; Mercurio A.; Brescia M.; Cavuoti S.; Longo G. (main hypothesis: PASP or A&A);

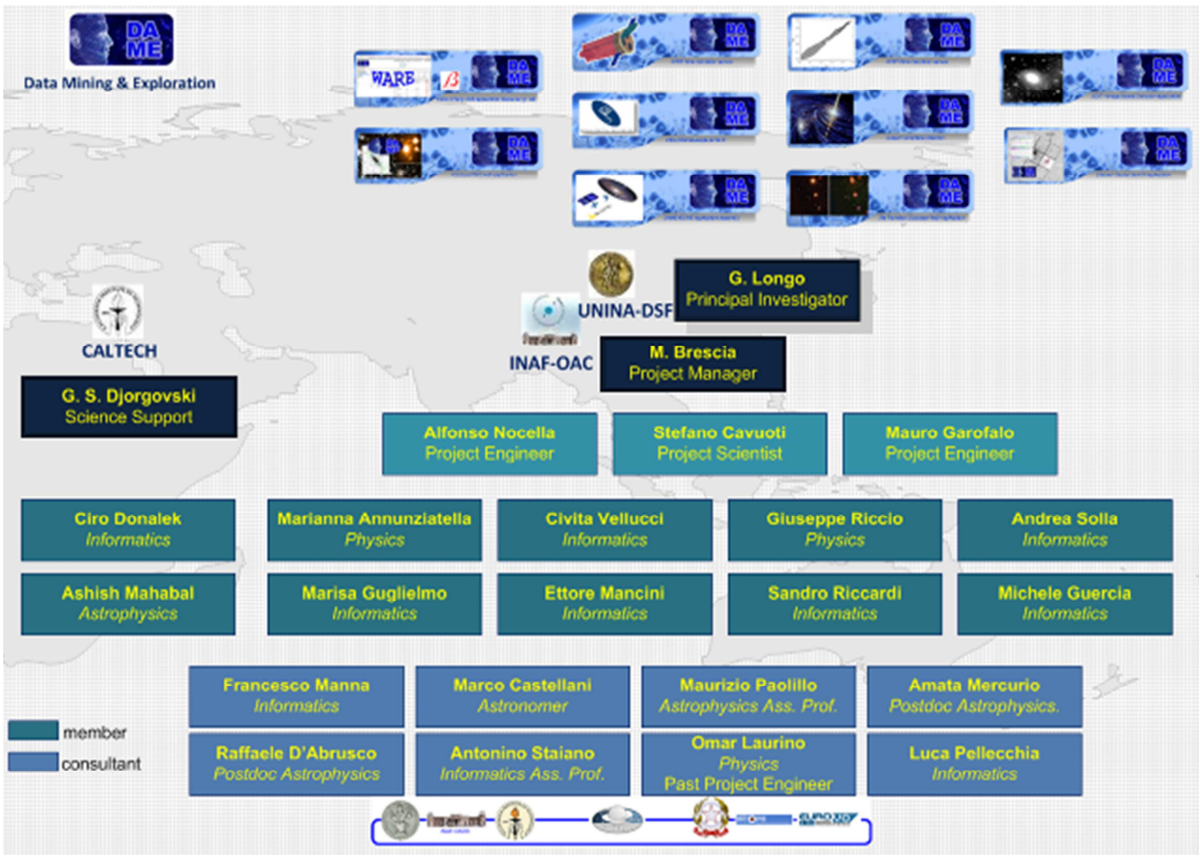
"*Photometric AGN Classification in the SDSS with Machine Learning Methods*", Cavuoti, S.; Brescia, M.; D'Abrusco R., Longo G. (MNRAS);

"*Parallel GPU Based Machine Learning for Classification Problems*", Garofalo, M.; Brescia, M.; Cavuoti, S.; Pescapè, A.; Longo, G.; Ventre G. (main hypothesis: IEEE Transactions or Neural Networks);

"*Data Mining in Astronomy with DAMEWARE. Part I: The Infrastructure*", Brescia, M.; Cavuoti, S.; Donalek, C.; Garofalo, M.; Nocella, A.; Guglielmo, M.; Vellucci, C.; Mahabal, A.; Graham, M.; Djorgovski, G.S.; Longo, G. (main hypothesis: A&A, PASP or Experimental Astronomy);

"*Data Mining in Astronomy with DAMEWARE. Part II: Scientific Applications*", Cavuoti, S.; Brescia, M.; Donalek, C.; D'Abrusco, R.; Mahabal, A.; Graham, M.; Djorgovski, G.S.; Longo, G. (main hypothesis: A&A, PASP or Experimental Astronomy);

"*Data Mining in Astronomy with DAMEWARE. Part III: Modeling Heuristics and Practical Tips*", Brescia, M.; Cavuoti, S.; Donalek, C.; Djorgovski, G.S.; Longo, G. (main hypothesis: A&A, PASP or Experimental Astronomy);


Data Mining & Exploration
CALTECH
G. S. Djorgovski
Science Support
UNINA-DSF
INAF-OAC
G. Longo
Principal Investigator
M. Brescia
Project Manager
Alfonso Nocella
Project Engineer
Stefano Cavuoti
Project Scientist
Mauro Garofalo
Project Engineer
Ciro Donalek
Informatics
Marianna Annunziatella
Physics
Civita Vellucci
Informatics
Giuseppe Riccio
Physics
Andrea Solla
Informatics
Ashish Mahabal
Astrophysics
Marisa Guglielmo
Informatics
Ettore Mancini
Informatics
Sandro Riccardi
Informatics
Michele Guercia
Informatics
Francesco Manna
Informatics
Marco Castellani
Astronomer
Maurizio Paolillo
Astrophysics Ass. Prof.
Amata Mercurio
Postdoc Astrophysics.
Raffaele D'Abrusco
Postdoc Astrophysics
Antonino Staiano
Informatics Ass. Prof.
Omar Laurino
Physics
Past Project Engineer
Luca Pellecchia
Informatics
member
consultant

DAta Mining & Exploration
DAME

Editor in Charge

- M. Brescia (brescia@oacn.inaf.it)

Editorial Board

- R. D'Abrusco (dabrusco@head.cfa.harvard.edu)
- C. Donalek (donalek@astro.caltech.edu)
- S.G. Djorgovski (george@astro.caltech.edu)
- S. Cavuoti (cavuoti@na.infn.it)
- G. Longo (longo@na.infn.it)

Technical Support

- helpdame@gmail.com
- dame@oacn.inaf.it
- Skype helpdesk: help dame

Official Website and MEDIA Resources

- http://dame.dsf.unina.it
- http://www.youtube.com/user/DAMEmedia