Data mining and Knowledge Discovery Resources for Astronomy in the Web 2.0 Age

S. Cavuoti*[a,b], M. Brescia[b,a], G. Longo[a,b,c]
[a] Department of Physics, University Federico II, Naples, Italy;
[b] National Institute for Astrophysics, Capodimonte Astronomical Observatory, Naples, Italy
[c] Visiting associate, Department of Astronomy, Caltech, Pasadena, USA

ABSTRACT

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.

Keywords: data mining, AstroInformatics, web application, virtual observatory

1 INTRODUCTION

There is nowadays one emerging discipline which reflects the need of a new approach to scientific investigation in the modern astrophysical data-centric field: AstroInformatics. It incorporates a new methodological shift within the astronomical community. The new generation of telescopes, space missions and focal plane sensors is going to produce data streams of the order of many terabytes per night. These are quantities of data which cannot be practically explored with traditional software and hardware tools. Such data volumes basically imply two consequences: (i) most processing and analysis tasks have to be performed in an as much automatic as possible fashion, using computing resources which push to the limits all ICT (Information & Communication Technology) technologies; (ii) the huge quantity of data (multi-band and multi-epoch images and/or catalogues) became completely incompatible with the bandwidth of internet resources, de facto making impossible to efficiently move data through the Web. In order to render user friendly and effective data mining on astronomical massive data sets (MDS), there are therefore several types of problems which need to be solved: (i) to provide the users with an easier access to both methods and computing power; (ii) to identify and implement faster algorithms exploiting whereas possible hybrid computing platforms, based on a well-orchestrated mixture of High Performance Computing (HPC), distributed computing (grid/cloud) and parallel computing paradigms; (iii) to minimize or even reduce the data transfer by moving the programs rather than the data. The first problem finds its natural solution in the usage of web applications, where the user interacts with the data mining framework via a browser and the computing infrastructure is completely transparent to him. While the third one needs to completely change the perspective, by trying to exchange well-designed data mining workflows and algorithms over the Web, among standardized data warehouses hosting data repositories. In this context the DAME (Data Mining & Exploration) Collaboration has proposed a solution framework to deal with the above problems. The solution has been moved also towards an extension of the interoperability concept already introduced by Virtual Observatory (VO) Consortium, originally concentrated on the data standard representation and their handling tools, in order to involve and organize also data mining applications in an interoperable context [22].

2 THE DAME PROGRAM

The DAME (Data Mining and Exploration) Program[1] is based on a platform which allows the scientific community to perform data mining and exploration experiments on massive data sets, by using a simple web browser. By means of state of the art Web 2.0 technologies (for instance web applications and services), DAME offers several tools which can be seen as working environments where to choose data analysis functionalities such as clustering, classification, regression, feature extraction etc., together with models and algorithms. All of these methods are derived from the machine learning paradigms, for instance supervised methods, whenever a Base of Knowledge (BoK) is available from data, or unsupervised models, when a BoK is not available and a self-adaptive capability is required to extract knowledge from datasets. The user can setup, configure and execute experiments on his own data on top of a virtualized computing infrastructure, without the need to install any software on his local machine. With DAME applications the user has also the possibility to extend the original library of available tools, by allowing the end users to plug-in and execute their own codes in a simple way, by uploading the programs without any restriction about the native programming language, and automatically installing them through a simple guided interactive procedure. Moreover, the DAME platform offers a variety of computing facilities, organized as a cloud of versatile architectures, from the single multi-core processor to a grid farm, automatically assigned at runtime to the user task, depending on the specific problem, as well as on the computing and storage requirements.

2.1 Applications and Services for Astrophysics

An important part of the computing challenges in astronomy are related to the handling, processing and modeling of large quantities of data. All astrophysical carriers have their peculiarities and weaknesses from the scientific point of view: they sample different energy ranges, different physical phenomena (e.g. thermal, non thermal and stimulated emission mechanisms), and require very different technologies for their detection. So far, the community needs modern infrastructures for the exploitation of the ever increasing amount of data (of the order of PetaByte/year) produced by the new generation of telescopes and space borne instruments, as well as by numerical simulations of exploding complexity. Indeed the basic requirements (extensible also to other application fields) can be summarized in two items: (a) the need of a "federation" of experimental data, by collecting them through several worldwide archives and by defining a series of standards for their formats and access protocols; (b) the implementation of innovative computing tools for Data Mining (DM) and knowledge extraction, user-friendly, scalable and as much as possible based on an asynchronous mechanism.

These topics require powerful, computationally distributed and adaptive tools able to explore, extract and correlate knowledge from multivariate massive datasets in a multi-dimensional parameter space. Every time a new technology enlarges the parameter space or allows a better sampling of it, new discoveries are bound to take place. Hence the scientific exploitation of a multi-band (D bands), multi-epoch (K epochs) universe implies to search for patterns and trends among N points in a DxK dimensional parameter space, where $N > 10^9$, $D >> 100$, $K > 10$.

The problem also requires a multi-disciplinary approach, covering aspects belonging to Astronomy, Physics, Biology, Information Technology, Artificial Intelligence, Engineering and Statistics.

As for data, the concept of "distributed archives" is already familiar to the average astrophysicist. The leap forward in this case is to be able to organize the data repositories to allow efficient, transparent and uniform access: these are the basic goals of the VO [41].

DAME extends this fundamental target by integrating it in an infrastructure, joining CLOUD service-oriented software and GRID resource-oriented hardware paradigm, including the implementation of advanced tools for MDS exploration [5]. In particular, concerning the GRID side, the Suite exploits the S.Co.P.E. GRID infrastructure [6]. The S.Co.P.E. project, aimed at the construction and activation of a Data Center which is now perfectly integrated in the national and international GRID initiatives, hosts 300 eight-core blade servers and 220 Terabyte of storage. The acronym stands for "Cooperative System for Multidisciplinary Scientific Computations", that is a collaborative system for scientific applications in many areas of research [39].

Moreover to overcome the limitation due to synchronous run of services, 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.

[1] http://dame.dsf.unina.it or equivalently http://dame.caltech.edu

The user, via client web applications and through the Ajax (Asynchronous JavaScript and XML) technology, [24], 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.

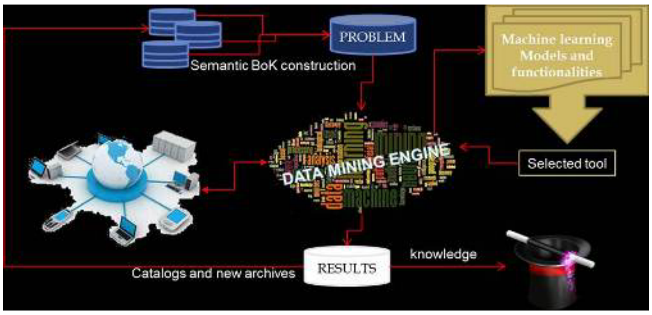

PROBLEM
Semantic BoK construction
Machine learning Models and functionalities
DATA MINING ENGINE
Selected tool
knowledge
Catalogs and new archives
RESULTS

Figure 1. The DAME Virtuous Circle of the Data Mining. We start from a BoK and from a variety of machine learning models, obtaining a result that brings new knowledge and also a new (or better) BoK, which allows to restart the loop.

Furthermore, 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. In most cases a scientist already possesses his own algorithms for data processing and analysis and has implemented private routines/pipelines to solve specific problems. These tools, however, are not scalable to distributed computing environments. DAME aims at providing a user friendly web based tool to encapsulate own algorithm/procedure into the package, automatically formatted to follow internal programming standards [11]. In our project 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, image segmentation are examples of functionalities belonging to the data mining conceptual domain, in which the various methods can be applied to explore data under a particular aspect, connected to the associated functionality scope. Such DM models are mainly derived from Machine Learning and Artificial Intelligence taxonomy: Multi Layer Perceptron (MLP) [20], with classical Back Propagation [4], Quasi Newton [44], or Genetic Algorithms [31], learning paradigms; Support Vector Machine (SVM) [15]; Self-Organizing Feature Maps (SOFM) [36], Principal Probabilistic Surfaces (PPS) [45].

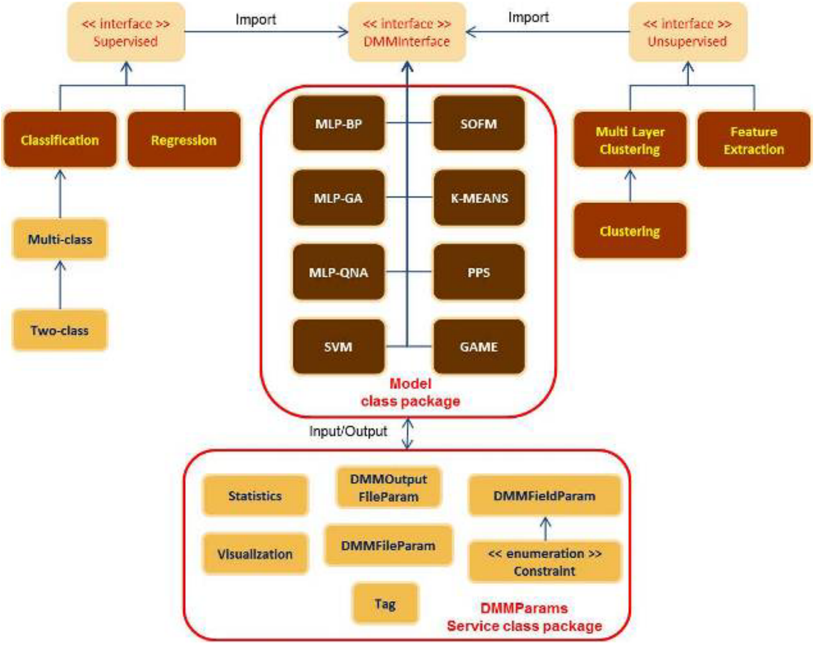
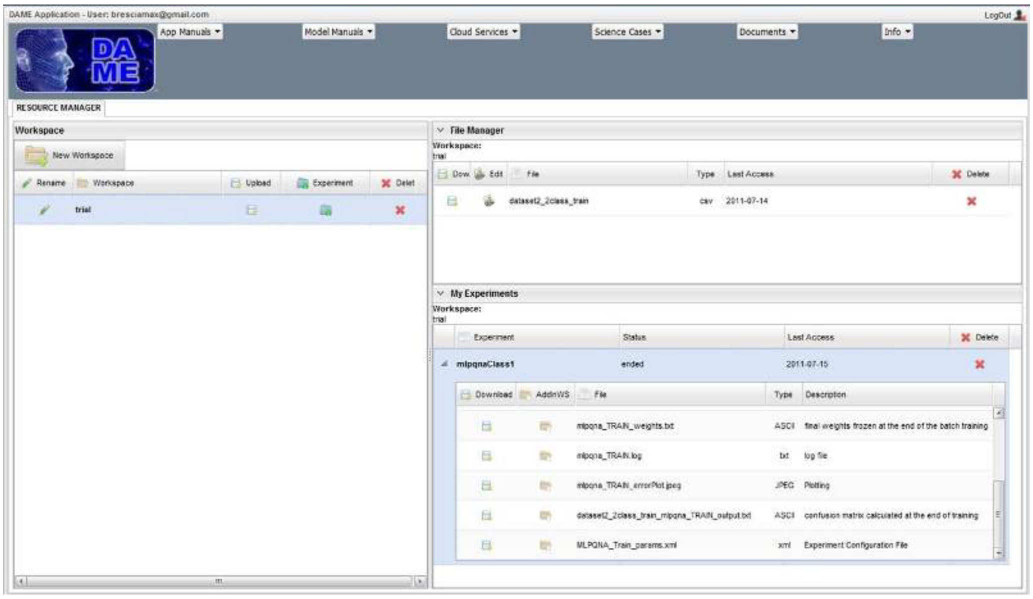


Figure 2. The software organization of data mining algorithms and related experiment functionality domains. Such design pattern is able to expose all kinds of possible experiments to end users through a web interface.

All these 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.



Workspace

Rename	Workspace	Upload	Experiment	Delet
	trial			

File Manager

Workspace: trial

Dow	Edit	File	Type	Last Access	Delete
		dataset2_2class_train	csv	2011-07-14	

My Experiments

Workspace: trial

Experiment	Status	Last Access	Delete
mlpqnaClass1	ended	2011-07-15	

Download	AddinWS	File	Type	Description
		mlpqna_TRAIN_weights.txt	ASCII	final weights frozen at the end of the batch training
		mlpqna_TRAIN.log	txt	log file
		mlpqna_TRAIN_errorPlot.jpeg	JPEG	Plotting
		dataset2_2class_train_mlpqna_TRAIN_output.txt	ASCII	confusion matrix calculated at the end of training
		MLPQNA_Train_params.xml	xml	Experiment Configuration File

Figure 3. The home page of the Graphical User Interface for the DAMEWARE web application, specialized in data mining with machine learning techniques.

As previously remarked, DAME is organized as a CLOUD of web applications and services. The following are the ones already available and accessible from the project website (http://dame.dsf.unina.it):

- **DAMEWARE**: DAME Web Application REsource, the main service of the project, providing via browser a complete DM framework including dataset and experiment configuration, execution and graphical/text tools for outputs. The beta 3 release is available at the following address: http://dame.dsf.unina.it/beta_info.html;
- **VOGCLUSTERS**: a VO-compliant web application for data and text mining on globular clusters, currently available as beta release (http://dame.dsf.unina.it/vogclusters.html);
- **STraDiWA**: Sky Transient Discovery Web Application, a project finalized to the real time classification of photometric transients. The project includes an automatic workflow to generate astronomical images with an user-defined number and type of variable objects, in order to perform setup and calibration of classification models running on the real images coming from observations (http://dame.dsf.unina.it/dame_td.html);
- **SDSS mirror**: local mirror website hosting a complete SDSS (Sloan Digital Sky Survey) data archive and management system, with particular reference to the image archive covering the VST (VLT Survey Project) KIDS survey area (http://dames.scope.unina.it/);
- **WFXT transient calculator**: a web service (http://dame.dsf.unina.it/dame_wfxt.html) for the Wide Field X-ray Telescope project to estimate the number of variable sources detected within the 3 main planned extragalactic surveys, with a given significant threshold;
- **EUCLID Data Quality** Common Tools: the DAME team leads the data quality tools work package for the ESA Euclid space mission[2]. The tools are foreseen to be integrated into the data reduction and analysis pipelines, provided by the Science Ground Segment of the project Consortium[3].

[2] http://sci.esa.int/euclid
[3] http://www.euclid-ec.org/

3 SCIENTIFIC IMPACT AND FIRST RESULTS

The models and algorithms provided with the Suite have been already tested on astrophysical use cases, with successful results, as reported in [8], [9], [13], [14], [16], [17] and [18]. In the following we summarize their topics.

3.1 Regression - Photometric Redshifts

To estimate the distance of an astronomical object, such as a galaxy or a quasar, by using its photometry (that is, the brightness of the object viewed through various standard filters), one can determine the so-called photometric redshifts (photo-z) and by using Hubble's law [32], it is possible to estimate the distance of the sources. This technique relies upon the fact that the spectrum of radiation being emitted by most objects have strong features that can be detected by the broad band filters. Photometric redshifts are one of the main tools to investigate the spatial distribution of galaxies as they are used to reconstruct the 3 dimensional positions of millions of sources. For instance, they are essential in constraining dark matter and dark energy studies by means of weak gravitational lensing, for the identification of galaxy clusters and groups [10], for type Ia supernovae, and to study the mass function of galaxy clusters ([1], [42], [35]). 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[4]), DES (Dark Energy Survey[5]), LSST (Large Synoptic Survey Telescope[6]), and Euclid [21], 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.

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[7] cross-matched with WISE[8], UKIDSS[9] and GALEX[10], obtaining a dataset 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% [8].

SDSS Galaxies

Our machine learning methods have been used to produce a catalogue[11] of more than 6 million galaxies observed in the Sloan Digital Sky Survey (SDSS), with an unprecedented accuracy [16]. 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 [16].

PHAT Contest

A significant advance in comparing different methods was introduced by Hildebrandt and collaborators [30] with the so called PHAT (PHoto-z Accuracy Testing) contest which adopts a black-box approach which is typical of benchmarking.

[4] http://www.astro-wise.org/projects/KIDS/
[5] http://www.darkenergysurvey.org/
[6] http://www.lsst.org/lsst/
[7] http://www.sdss.org/
[8] http://wise.ssl.berkeley.edu/
[9] http://www.ukidss.org/
[10] http://www.galex.caltech.edu/
[11] http://dame.dsf.unina.it/dame_photoz.html#sdss

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 (see [30] for more details).
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; [25]). 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 [13].

3.2 Classification

Statistical 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 labeled 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 in X to its classification label y in Y. Because of the supervised nature of the classification task, the system performance can be measured by means of a test set during the testing procedure, in which unseen data are given to the system to be labeled. The overall error somehow integrates information about the classification goodness. However, when a data set is unbalanced (when the number of samples in different classes varies greatly) the error rate of a classifier is not representative of the true performance of the classifier.

Globular Clusters

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 result led to a performance of 98.33% [9].

Active Galactic Nuclei

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 [34], Kauffman [33] and Heckman [29] lines in the Baldwin, Philips and Terlevich (BPT) diagnostic plot [2], 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 [14].

Quasar

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 [17]. 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 [37], 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 [45]. 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 [17].

4 THE FUTURE OF DATA MINING WAREHOUSES

Computing has rapidly established itself as essential to many branches of science, to the point where *computational science* is a commonly used term. Indeed, the application and importance of computing is set to grow dramatically across almost all the sciences. Computing has started to change how science is done, enabling new scientific advances through enabling new kinds of experiments. These experiments are also generating new kinds of data of increasingly exponential complexity and volume. Achieving the goal of being able to use, exploit and share most effectively these data is a huge challenge. The harder problem for the future is heterogeneity of platforms, data and applications, rather than simply the scale of the deployed resources. The goal should be to allow scientists to explore the data easily, with sufficient processing power for any desired algorithm to process it. Current platforms require the scientists to overcome computing barriers between them and the data [7]. Our convincement is that most aspects of computing will see exponential growth in bandwidth, but sub-linear or no improvements at all in latency. Moore's Law [40] will continue to deliver exponential increases in memory size but the speed with which data can be transferred between memory and CPUs will remain more or less constant and marginal improvements can only be made through advances in caching technology. Certainly Moore's law will allow the creation of parallel computing capabilities on single chips by packing multiple CPU cores onto it, but the clock speed that determines the speed of computation is constrained to remain limited by a *thermal wall* [46]. Thus computing machines will not get much faster. But they will have the parallel computing power and storage capacity that we used to only get from specialist hardware. As a result, smaller numbers of supercomputers will be built but at even higher cost. From an application development point of view, this will require a fundamental paradigm shift

from the currently sequential or parallel programming approach in scientific applications to a mix of parallel and distributed programming that builds programs that exploit low latency in multi core CPUs.
We recall that novel data warehouses should offer techniques to support the new paradigm of data-centric science. We think that in a data-centric environment, it should be as much as possible minimized the massive data flow on the network. It is indeed much more convenient and fast to move applications towards the data centers, especially if they are organized as Knowledge Discovery in Databases (KDD) application warehouses. This of course requires a well-defined standardization process, in order to organize applications in a fully interoperable way:

- For organizational learning to take place, data must be gathered together and organized in a consistent and useful way, hence, Data Warehousing (DW);
- DW should allow an organization to remember what it has noticed about its data;
- Data Mining applications should be interoperable with data organized and shared between DW.

By having the possibility to share on demand applications between standardized and interoperable KDD warehouses, it may engage a virtuous mechanism in which users may operate by remote, through a simple web browser, sharing resource on the network (not data), building flexible computing platforms and orchestrating them in the virtual computing cloud, by interacting with these resource in an asynchronous way (for example by exploiting the web containers based on Ajax technology).

4.1 The Lernaean Hydra proposed solution

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.

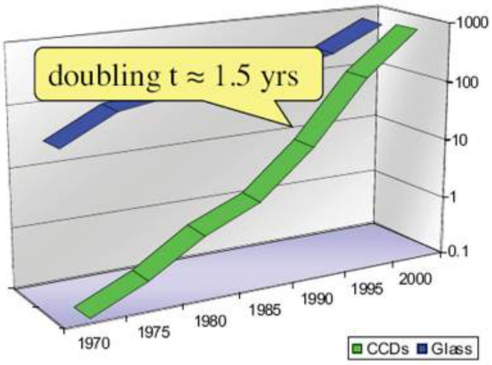


Figure 4. According to the Nielsen’s Law Network bandwidth double every 21 month; instead data in astrophysics are growing exponentially, with a doubling time of 1.5 years (courtesy G. S. Djorgovski).

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 VObs, are based on implementing web based protocols for application interoperability [19] and [26]. Possible interoperability scenarios could be:

1. DA1 ←→ DA2 (data + application bi-directional flow)
    a. Full interoperability between DA (Desktop Applications);
    b. Local user desktop fully involved (requires computing power);

2. DA ←→ WA (data + application bi-directional flow)
    a. Full WA → DA interoperability;
    b. Partial DA → WA (Web Applications) interoperability (such as remote file storing);
    c. MDS must be moved between local and remote apps;
    d. user desktop partially involved (requires minor computing and storage power);

3. WA ←→ WA (data + application bi-directional flow)
   a. Except from URI exchange, no interoperability and different accounting policy;
   b. MDS must be moved between remote apps (but larger bandwidth);
   c. No local computing power required;

All of these mechanisms 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[12]:

4. WA ←→ WA (plugin bi-directional exchange)
   a. All DAs must become Was;
   b. Unique accounting policy (google/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) [27]. 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 scheme 4, no local computing power is required. Also smartphones can run DM 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);

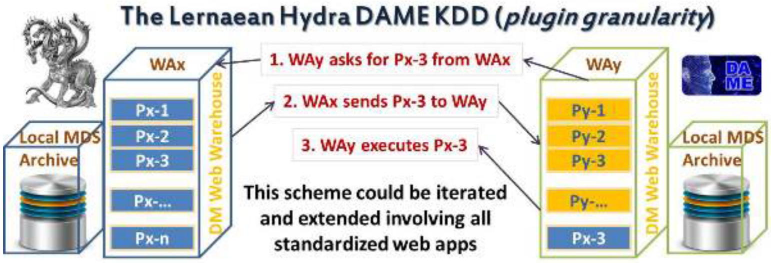


Figure 5. The main steps of the application plugins exchange mechanism among data mining web warehouses.

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

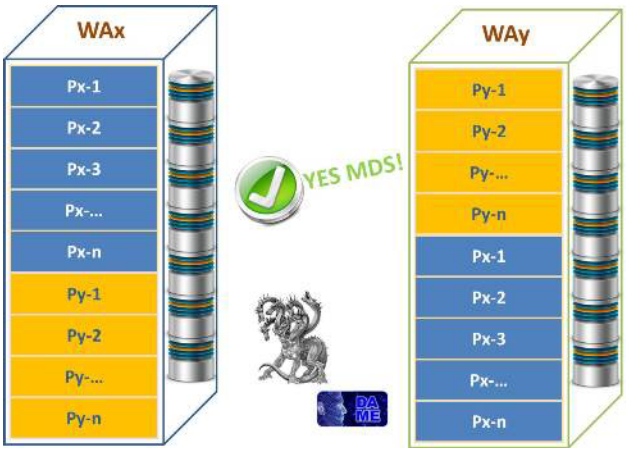


Figure 6. The scenario after several cycles of the application plugins exchange mechanism. At the end, all web repositories are equivalent mirrors, able to perform data mining experiments on local massive data archives.

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 previous chapters, represents a first prototype towards a WAR mechanism.

4.2 GPU Parallel Computing in Astronomy

Whenever there is a large quantity of data, there are two approaches to making learning feasible. The first one is trivial, consisting of applying the training scheme to a decimated data set. Obviously, in this case, the information may be easily lost and there is no guarantee that this loss is negligible in terms of correlation discovery. This approach, however, may turn very useful in the lengthy optimization procedure that is required by many ML methods (such as neural networks or genetic algorithms). The second method relies in splitting the problem in smaller parts (parallelization) sending them to different CPUs (Central Processing Units) and finally combine the results together. However, implementation of parallelized versions of learning algorithms is not always easy [43], and this approach should be followed only when the learning rule, such as in the case of genetic algorithms [38], is intrinsically parallel.

GPGPU is an acronym standing for General Purpose Computing on Graphics Processing Units. It was invented by Mark Harris in 2002, [28], 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).

The critical points for traditional multi-core CPU architecture come out in case of serial programs. In this case, in the absence of the parallel approach, the processes are scheduled in such a way that the full load on the CPU is balanced, by distributing them over the less busy cores each time. However many software products are not designed to fully exploit the multi-core features, so far the micro-processors are designed to optimize the execution speed on sequential programs.

The choice of graphic device manufacturers, like NVIDIA Corp., was the many-core technology (usually many-core is intended for multi-core systems over 32 cores). The many-core paradigm is based on the growth of execution speed for parallel applications. Began with tens of cores smaller than CPU ones, such kind of architectures reached hundreds of core per chip in a few years. Since 2009 the throughput peak ratio between GPU (many-core) and CPU (multi-core) was about 10:1. It must be issued that such values are referred mainly to the theoretical speed supported by such chips, i.e. 1 TeraFLOPS against 100 GFLOPS. Such a large difference has pushed many developers to shift more computing-expensive parts of their programs on the GPUs.

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 a recent paper, Barsdell, Barnes and Fluke [3] have analyzed astronomy algorithms to understand which algorithms can be best engineered to run on GPU's. The authors used algorithm analysis to understand how algorithms can be optimized to run in parallel in a GPU environment (as opposed to implementation optimization).

Broadly speaking, the following are features for high efficiency parallelized algorithms:

- Repetitive operations can be parallelized into many fine-grained elements;
- Standardized mechanisms to access locations in memory;
- Minimize processing threads which execute different instructions;
- Parallelize high intensity arithmetic operations;
- Minimize memory transfers between GPU and host CPU.

In this context we designed and developed a multi-purpose genetic algorithm (GAME) implemented with GPGPU/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 [12]. 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. Although a speedup of up to 200X over a modern CPU is impressive, it ignores the larger picture of use a Genetic Algorithm as a whole. In any real-world the dataset can be very large (those we have previously called Massive Data Sets) and this requires greater attention to GPU memory management, in terms of scheduling and data transfers host-to-device and vice versa. Moreover, the identical results for classification and regression functional cases demonstrated the consistency of the implementation, enhancing the scalability of the proposed GAME model when approaching massive data sets problems.

5 CONCLUSIONS

Scientists in many fields have generally not considered Web 2.0 as a collaborative technology and as a means of sharing data. The main barrier is apparently a lack of understanding of how to get started and what benefits will provide to scientists. On the contrary there are many science projects that could be done effectively with Web 2.0. This is particularly true in astronomy, a data-centric discipline increasingly dominated by a constellation of data centers and multi-institutional research consortia, which in theory should make it an ideal domain for Web 2.0.

Web 2.0 technology could also allow collaborative efforts from surveys. For example, using the virtual data concept [23], different survey projects could be cross-matched, obtaining multi-band and multi-epoch catalogues, which could be make publicly accessible, if respective warehouses are included in a standardized and interoperable virtual organization. Moreover, through the web plugin exchange mechanism, the data could be processed locally at each data center hosting a web data mining application repository, without moving massive data sets over the network. Using such mashup concept, WAR components could be exposed to users with inputs and outputs named to make them clearly understandable, and by having well-standardized interfaces to be plugged in a remote data mining framework. This would require to allow users to compose services as they choose, including mashing them up with other local specific services. This could be achieved by evolving DAMEWARE prototype with VO infrastructures as well.

By the early 2000s, many scientists were designing unique user interfaces and tools to access compute and data resources. The DAME Science Gateways program officially started in 2007 to try to encourage this to continue, with common tools being developed for the scientists and communities who were building the gateways. This was originally done using mostly Web 1.0 technologies, but currently, the DAME Lernaean Hydra Science Gateways program is investigating how to integrate Web 2.0 technologies as well.

ACKNOWLEDGEMENTS

The authors wish to thank all present and former members of the DAME Collaboration, in particular prof. G.S. Djorgovski and the Caltech Team for their unlimited contribution over these years of intensive collaboration. DAME has been funded through several grants which are mentioned on the Collaboration web site. GL wishes to thank also the California Institute of Technology for the kind hospitality. MB wish to thank the financial support of PRIN-INAF 2010, ”Architecture and Tomography of Galaxy Clusters”.

REFERENCES

[1] Albrecht, A., Bernstein, G., Cahn, R. et al., “Report of the Dark Energy Task Force”, (astro-ph/0609591), (2006)

[2] Baldwin, J.A., Phillips M.M. & Terlevich R.: "Classification parameters for the emission-line spectra of extragalactic objects", PASP, 93, 5 (1981)

[3] Barsdell, B.R., Barnes, D.G. Fluke C.J., "Analysing Astronomy Algorithms for GPUs and Beyond", Monthly Notices of the Royal Astronomical Society, Volume 408, Issue 3, pp. 1936-1944 (2010)

[4] Bishop, C. M., [Neural Networks for Pattern Recognition], Oxford Univ. Press, Oxford, (1995)

[5] Brescia, M., Cavuoti, S., d'Angelo, G., D'Abrusco, R., Donalek, C. Deniskina, N., Laurino, O., Longo, G., "Astrophysics in S.Co.P.E.", Mem S.A. It. Suppl, Vol 13, 56, (2009).

[6] Brescia, M. et al., "The DAME/VO-Neural infrastructure: an integrated data mining system support for the science community", Proc. Final Workshop of GRID Projects, PON Ricerca 2000-2006 Avviso 1575, (2009).

[7] Brescia, M., Longo, G., "Astroinformatics, data mining and the future of astronomical research", Proc. The 2-nd International Conference on Frontiers in Diagnostic Technologies, Nuclear Instruments and Methods in Physics Research A, NIMA Elsevier Journal, http://arxiv.org/abs/1201.1867, (2012)

[8] Brescia, M., Cavuoti, S., D’Abrusco, R., Mercurio, A., Longo, G., “Photometric Redshifts from WISE-GALEX-UKIDSS-SDSS Quasars based on Quasi Newton Neural Model (MLPQNA)”, (in preparation), (2012)

[9] Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., Puzia, T., "The detection of Globular Clusters in galaxies as a data mining problem", Monthly Notices of the Royal Astronomical Society, Volume 421, Issue 2, pp. 1155-1165, available at arXiv:1110.2144v1 (2012)

[10] Capozzi, D., De Filippis, E., Paolillo, M., D’Abrusco, R., Longo G., “The properties of the heterogeneous Shakhbazyan groups of galaxies in the SDSS”, MNRAS, 396, 900, (2009)

[11] Cavuoti, S., Brescia, M., Longo, G., Garofalo, M., Nocella, A. "DAME: A Web Oriented Infrastructure for Scientific Data Mining and Exploration", Proc. Astronomical Images and Data Mining in the International Virtual Observatory Context, Science - Image in Action, Edited by Bertrand Zavidovique and Giosue' Lo Bosco. Published by World Scientific Publishing Co. Pte. Ltd., 2012. ISBN 9789814383295, pp. 241-247, (2012)

[12] Cavuoti, S., Garofalo, M., Brescia, M., Pescapè, A., Longo, G. & Ventre G.,"Genetic Algorithm Modeling with GPU Parallel Computing Technology", Proc. 22nd WIRN Italian Workshop on Neural Networks, IIASS Vietri sul Mare, Salerno, May 18, (2012)

[13] Cavuoti S., Brescia, M., Longo, G., Mercurio, A., "Photometric redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 contest.", Submitted to Astronomy & Astrophysics, [arXiv:astro-ph/1206.0876], (2012)

[14] Cavuoti, S., Brescia, M., D’Abrusco, R., Longo, G., “Photometric AGN Classification in the SDSS with Machine Learning Methods”, (in preparation), (2012)

[15] Chang, C.-C., Lin, C.-J., ACM Trans. Intelligent Syst. Technol., 2, 27, (2011)

[16] D'Abrusco, R., Staiano, A., Longo, G., Brescia, M., De Filippis, E., Paolillo, M., Tagliaferri, R., "Mining the SDSS data. I. Photometric redshifts for the nearby universe", The Astrophysical Journal, 663, 752-764, (2007)

[17] D'Abrusco, R., Longo, G., Walton, N.A., "Quasar candidate selection in the Virtual Observatory era", Monthly Notices of the Royal Astronomical Society, (2008).

[18] D'Abrusco, R., Longo, G. & Walton, N. A., "QSO candidates selection in VO era", Monthly Notices of the Royal Astronomical Society, (2009)

[19] Derrierre, S., Boch, T., "Sampling your browser for the semantic web", Proc. Astronomical Data Analysis Software and Systems XIX, Masatoshi Ohishi. Asp conference series, vol. 434. San Francisco: Astronomical Society of the Pacific, P.159, (2010)

[20] Duda R. O., Hart P. D., Storck D. G., [Pattern Classification], 2nd edn. Wiley, NY, (2004)

[21] Euclid Red Book, ESA Technical Document, ESA/SRE(2011)12, Issue 1.1, [arXiv:astro-ph/1110.3193], (2011)

[22] Fabbiano, G., Brogan, C., Calzetti, D. et al. "Recommendations of the Virtual Astronomical Observatory (VAO)", Proc. Science Council for the VAO second year activity 2011, arxiv: 1108.4348 (2011)

[23] Foster, I., Vöckler, J., Wilde, M., Zhao, Y., "Chimera: A Virtual Data System For Representing, Querying, and Automating Data Derivation", Proc. 14th Conference on Scientific and Statistical Database Management (2002)

[24] Garrett, J.J. (18 February 2005). "Ajax: A New Approach to Web Applications", adaptivepath.com, (2008).

[25] Giavalisco, M., Ferguson, H. C., Koekemoer, A. M., et al., ApJ, 600, L93, (2004)

[26] Goodman, A., Fay, J., Muench, A., Pepe, A., Udomprasert, P., Wong, C., "WorldWide Telescope in Research and Education", eprint, arXiv:1201.1285, (2012)

[27] Guazzelli, A., Zeller, M., Chen, W. and Williams, G., "PMML: An Open Standard for Sharing Models", The R Journal, Volume 1/1,(2009)

[28] Harris, M.J., "Real-Time Cloud Simulation and Rendering.", University of North Carolina Technical Report, #TR03-040 (2003)

[29] Heckman, T.M. "An optical and radio survey of the nuclei of bright galaxies - Activity in normal galactic nuclei", Astronomy & Astrophysics, 87, 182, (1980)

[30] Hildebrandt, H., Arnouts, S., Capak, P., Wolf, C. et al., "PHAT: PHoto-z Accuracy Testing", Astronomy & Astrophysics, 523, 31, (2010)

[31] Holland, J. H., [Adaptation in Natural and Artificial Systems], University of Michigan Press, Ann Arbor, (1975)

[32] Hubble E., "A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae", Proc. of the National Academy of Sciences of the United States of America, Volume 15, Issue 3, pp. 168-1731929 (1929)

[33] Kauffman, G. et al.: "The host galaxies of active galactic nuclei", Monthly Notices of the Royal Astronomical Society, 346, 1055, (2003)

[34] Kewley, L.J. et al., "The host galaxies and classification of active galactic nuclei", Monthly Notices of the Royal Astronomical Society, (2006)

[35] Keiichi, U., Medezinski, E., Nonino, M., et al., ApJ Submitted, arXiv:1204.3630, (2012)

[36] Kohonen, T., [Self-Organizing Maps], Springer, 3rd edition, (2000)

[37] Mahalanobis, Prasanta Chandra, "On the generalised distance in statistics" . Proceedings of the National Institute of Sciences of India 2 (1): 49–55, (1936)

[38] Meng Joo, E. and Fan, L., "Genetic algorithms for MLP neural network parameters optimization", Proc. In Control and Decision Conference, Guilin, China, 3653-3658 (2009)

[39] Merola, L., "The SCOPE Project.", Proc. Final Workshop of GRID Projects, PON Ricerca 2000-2006 Avviso 1575, (2008)

[40] Moore, Gordon E., "Cramming more components onto integrated circuits", Electronics Magazine, p. 4, (1965)

[41] Pasian, F., Brescia, M., Longo, G., "Astronomical Images and Data Mining in The International Virtual Observatory Context Astronomical Images and Data Mining in the International Virtual Observatory Context", Science - Image in Action, Edited by Bertrand Zavidovique and Giosue' Lo Bosco, Published by World Scientific Publishing Co. Pte. Ltd., ISBN 9789814383295, pp. 230-240, 20. (2012)

[42] Peacock, J. A., Schneider, P., Efstathiou, G., et al., ESA-ESO Working Group on Fundamental Cosmology, Tech. Rep., (2006)

[43] Rajaraman, A. and Ullmann, J.D.; "Mining of Massive Data Sets", eprint, Available on line at http://infolab.stanford.edu/~ullman/mmds.html (2010)

[44] Shanno, D. F., Math. Comput., 24, 647, (1970)

[45] Staiano, A. et al., “Probabilistic principal surfaces for yeast gene microarray data-mining”, Proc. ICDM'04 - Fourth IEEE International Conference on Data Mining, Brighton, UK, (2004)

[46] Sutter, H., "The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software", Dr. Dobb's Journal, 30, 3, (2005)

*cavuoti@na.infn.it; phone +39 081 5575 553; fax +39 081 456710; http://dame.dsf.unina.it