Newsletter Edition
04
March 2015

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.**

FOLLOW US ON YOUTUBE DAME MEDIA CHANNEL

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.

Official Partners: INAF-OACN, Dept. of Physics UNINA, CALTECH

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. 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 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 Astroinformatics).

DAME partners acknowledge the financial support of the Italian Ministry of Foreign Affairs for the Italy-USA bi-lateral grant Building an e-science Data Mining Infrastructure, the European Union through the projects VO-Tech and VO-AIDA, the Ministry of University and Research through the PON S.Co.P.E. GRID project, the INAF for the partial support through the PRIN 2010 *Tomography of Galaxy Clusters* and the PRIN 2014 *Glittering Kaleidoscopes in the sky, the multifaceted nature and role of galaxy clusters*, the European and Italian Space Agencies for the support in the participation to the ESA EUCLID space mission program and the European Commission for the support through the Program FP7-SPACE-2013-1*ViaLactea*. We acknowledge the support of the EU COST Action TD1403 *Big Data Era in Sky and Earth Observation*. We also acknowledge a crucial support from the Fishbein Family Foundation, and a partial support from the NASA grant 08-AISR08-0085.

Since 2014 INAF-DAME has its proprietary EURO-VO Identification Authority to host VO data services.

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.

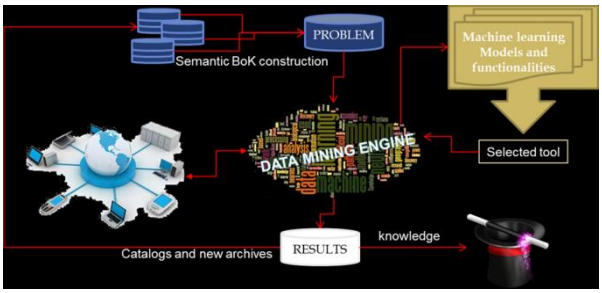

Fig. 1 – DAME data mining concept

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.

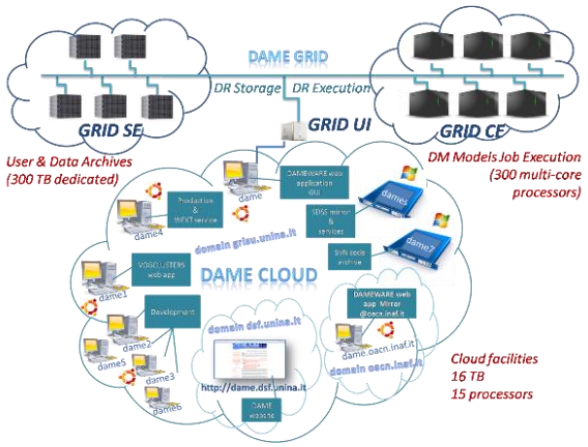

Fig. 2 – DAME distributed computing infrastructure

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 Highlights

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.**

DAME Products

From the indicated page of the DAME portal (http://dame.dsf.unina.it/products.html), 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.
Here some products are briefly presented.

DAMEWARE Web Application Resource

DAMEWARE is a web application, accessible through a simple web browser.
At the moment we have released the 1.0 release, available at the following address:
http://dame.dsf.unina.it/dameware.html

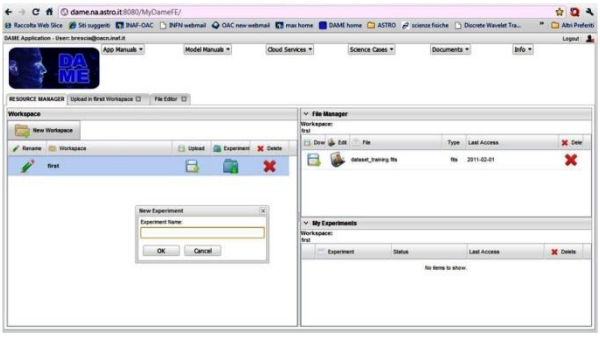

Fig. 3 – DAMEWARE Graphical User Interface

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.

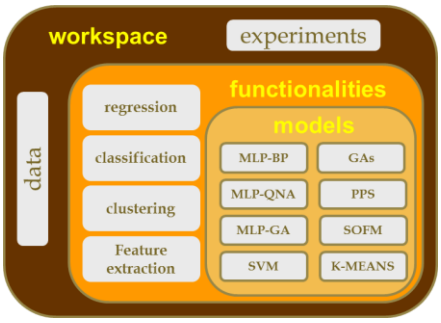

Fig. 4 – DAMEWARE internal architecture

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.

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, some DAME members has been officially included into the mission consortium, as one of the Science Teams in charge of data quality control, photometric redshift evaluation and Supernovae detection and classification. http://sci.esa.int/**euclid/**.

PhotoRApToR

The many ongoing and planned photometric surveys produce huge datasets that would not be analyzed without the data mining methods deriving from the emerging field of Astroinformatics. The problem is that a great part of astronomical data is stored in private archives that are not accessible on line. So, in order to evaluate photo-z it is needed a desktop application that can be used by everyone on its own personal computer. The name chosen for this application is PhotoRApToR, i.e. Photometric Research Application To Redshift. Behind such client desktop application there is a neural network, namely the MLPQNA (Multi Layer Perceptron implemented with a learning rule based on the Quasi Newton Algorithm). The service is able to perform any kind of regression and multi-variate classification experiments on large data.
GOTO the online service.

VST ACCESS – ShaSS Data Release

The Shapley Supercluster Survey (ShaSS, z~0.05) covers a contiguous area of 260 h^-2_70 Mpc^2 including the supercluster core. The project main aim is to quantify the influence of cluster-scale mass assembly on galaxy evolution in one of the most massive structures in the local Universe. The survey includes nine Abell clusters (A3552, A3554, A3556, A3558, A3559, A3560, A3562, AS0724, AS0726) and two poor clusters (SC1327-312, SC1329-313) showing evidence of cluster-cluster interactions. Optical (ugri) and near-infrared (K) imaging acquired with VST and VISTA allow us to study the galaxy population down to m*+6 at the supercluster redshift.
First release of the ShaSS i-band catalogue available.

EU FP7 ViaLactea

VIALACTEA will bring to a common forum the major new-generation surveys of the Galactic Plane from 1um to the radio, both in thermal continuum and in atomic and molecular lines, from Europe-funded space missions and ground-based facilities, to engage one of the fundamental challenges in Galactic astronomy: to quantify Galaxy wide the relationship between the physical agents responsible for the onset and the regulation of star formation in a spiral galaxy and the resulting Rate and Efficiency of star formation, and obtain a "star formation recipe" that will be a cornerstone to trace the star formation history of galaxies back to their formation.
GOTO DAME local service.

DAME SCIENCE

DAME philosophy 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.
GOTO the page of all DAME science cases.

SCIENCE CASES short links:

- PHOTOMETRIC REDSHIFTS
- QSO CLASSIFICATION
- GLOBULAR CLUSTERS
- AGN CLASSIFICATION
- SKY TRANSIENT CLASSIFICATION
- VIA LACTEA
- KiDS SURVEY
- EU COST ACTION ON BIG DATA
- ESA EUCLID MISSION
- HIPPOCAMPAL SEGMENTATION

DAME EDUCATION

DAME project is under continuous evolution. There are graduate (Bachelor and Master of Science) and PhD theses as well as undergraduate apprenticeships available in the following areas:

- ***Astrophysics & Astroinformatics***: mining, exploration, analysis, processing of massive data archives by using Artificial Intelligence, Machine Learning models;
- ***Computer Science & AstroInformatics***: design and development of applications for data processing in distributed infrastructures (GRID/CLOUD) and parallel programming (GPGPU); Design and implementation of data analysis and statistics models;

If you are interested to collaborate, you should contact any of the following people:

- Prof. G. Longo (Principal Investigator), Department of Physics, Univ. Federico II (Tel. +39.081.676.912)
- Dr. M. Brescia (Project Manager), INAF Astronomical Observatory of Capodimonte (Tel. +39.081.5575.553)

Recent Publications in DAME

Girardi M., **Mercurio** A., Balestra I., Nonino M., Biviano A., Grillo C., Rosati P.,**Annunziatella** M., Demarco R., Fritz A., Gobat R., Lemze D., Presotto V., Scodeggio M., Tozzi P., Bartosch Caminha G., **Brescia** M., et al., **2015**. CLASH-VLT: Substructure in the galaxy cluster MACS J1206.2-0847 from kinematics of galaxy populations *. A&A, accepted, 19pp.

Cavuoti S., **2015**. Data-Rich Astronomy: Mining Synoptic Sky Surveys. LAMBERT Academic Publishing, ISBN: 978-3-659-68311-4, 260pp.

Cavuoti S., **Brescia** M., **DeStefano** V., **Longo** G., **2015**. Photometric redshift estimation based on data mining with PhotoRApToR. Experimental Astronomy, Springer, Vol. 39, Issue 1, 45-71. Online article

Tangaro S., Amoroso N., **Brescia** M., **Cavuoti** S., Chincarini A., Errico R., Inglese P., **Longo** G., Maglietta R., Tateo A., **Riccio** G., Bellotti R., **2015**. Feature Selection based on Machine Learning in MRIs for Hippocampal Segmentation. To appear on Computational and Mathematical Methods in Medicine, Hindawi Publishing Corporation, 19 pages, 7 figures (in press)

Astier P., Balland C., **Brescia** M., Cappellaro E., Carlberg R.G., **Cavuoti** S., Della Valle M., Gangler E., Goobar A., Guy J., Hardin D., Hook I.M., Kessler R., Kim A., Linder E., **Longo** G., Maguire K., Mannucci F., Mattila S., Nichol R., Pain R., Regnault N., Spiro S., Sullivan M., Tao C., Turatto M., Wang X.F., Wood-

Vasey W.M., **2014**. Extending the Supernova Hubble diagram to z~1.5 with the Euclid space mission. A&A, Vol. 572, A80, 20pp. (December 2014)

Annunziatella M., A. Biviano, A. **Mercurio**, M. Nonino, P. Rosati, I. Balestra, V. Presotto, M. Girardi, R. Gobat, C. Grillo, E. Medezinski, D. Kelson, M. Postman, M.**Brescia**, B. Sartoris, R. De Marco, A. Koekemoer, D. Lemze, L. Bradley, U. Kuchner, E. Regos, K. Umetsu, B. Ziegler, **2014**. CLASH-VLT: The stellar mass function and stellar mass density profile of the z=0.44 cluster of galaxies MACS J1206.2-0847. A&A, Vol. 571, A80, 15pp. (November 2014)

Brescia, M.; Cavuoti, S.; Longo, G.; De Stefano, V.; **2014**, A catalogue of photometric redshifts for the SDSS-DR9 galaxies, A&A, Vol. 568, A126, 7 pages

Brescia, M.; Cavuoti, S.; Longo, G.; Nocella, A.; Garofalo, M.; Manna, F.; Esposito, F.; Albano, G.; Guglielmo, M.; D'Angelo, G.; Di Guido, A.; Djorgovski, G.S.; Donalek, C.; Mahabal, A.A.; Graham, M.J.; Fiore, M.; D'Abrusco, R.; **2014**,DAMEWARE: A web cyberinfrastructure for astrophysical data mining, PASP, Vol. 126, 942, pp. 783-797

V. Presotto, M. Girardi, M. Nonino, **A. Mercurio**, C. Grillo, P. Rosati, A. Biviano, **M. Annunziatella**, I. Balestra, W. Cui, B. Sartoris, D. Lemze, B. Ascaso, J. Moustakas, H. Ford, O. Czoske, S. Ettori, A. Fritz, U. Kuchner, M. Lombardi, C. Maier, E. Medezinski, M. Scodeggio, V. Strazzullo, P. Tozzi, B. Ziegler, M. Bartelmann, N. Benitez, L. Bradley, **M. Brescia**, et al.; **2014**, Intra Cluster Light properties in the CLASH cluster MACS J1206.2-0847, A&A, Vol. 565, A126, 17 pages, arXiv available

Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Paolillo, M.; **2014**,Photometric classification of emission line galaxies with Machine Learning methods, Monthly Notices of the Royal Astronomical Society, Volume 437, Issue 1, p.968-975

Cavuoti, S.; Garofalo, M.; Brescia, M.; Paolillo, M; Pescape', A.; Longo, G.; Ventre, G., **2014**, Astrophysical data mining with GPU. A case study: genetic classification of globular clusters, New Astronomy, Vol. 26, pp. 12-22, Elsevier

Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A.; **2013**,Photometric redshifts for Quasars in multi band Surveys, ApJ 772, 2, 140

A. Biviano, P. Rosati, I. Balestra, **A. Mercurio**, M. Girardi, M. Nonino, C. Grillo, M. Scodeggio, D. Lemze, D. Kelson, K. Umetsu, M. Postman, A. Zitrin, O. Czoske, S. Ettori, M. Lombardi, E. Medezinski, S. Mei, V. Presotto, P. Tozzi, B. Ziegler, **M. Annunziatella**, M. Bartelmann, N. Benitez, L. Bradley, **M. Brescia**, et al.; **2013**,CLASH-VLT: The mass, velocity-anisotropy, and pseudo-phase-space density profiles of the z = 0.44 galaxy cluster MACS 1206.2-0847, Re-submitted to A&A after first revision

Brescia, M.; Longo, G., **2013**, Astroinformatics, data mining and the future of astronomical research. Nuclear Inst. and Methods in Physics Research, A, Vol. 720, pp. 92-94

Annunziatella, M.; Mercurio, A.; Brescia, M.; Cavuoti, S.; Longo, G.; **2013**,Inside catalogs: a comparison of source extraction software, PASP, Vol. 125, Nr. 923, pp. 68-82

Cavuoti, S.; Brescia, M.; Longo, G.; Mercurio, A.; **2012**, Photometric Redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 Contest, A&A, Vol. 546, A13, pp. 1-8

Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T.; **2012**, 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 atarXiv:1110.2144v1

DAta Mining & Exploration
DAME

Technical Support

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

Official Website and MEDIA Resources

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