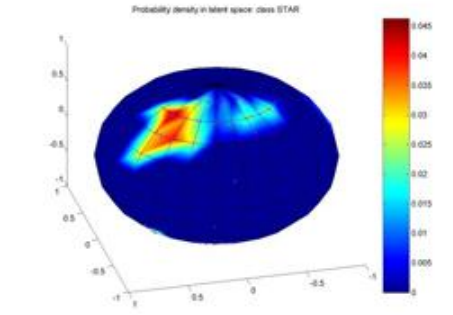
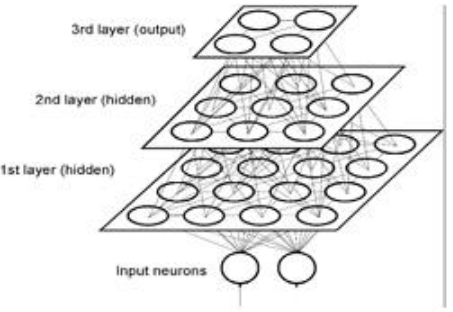
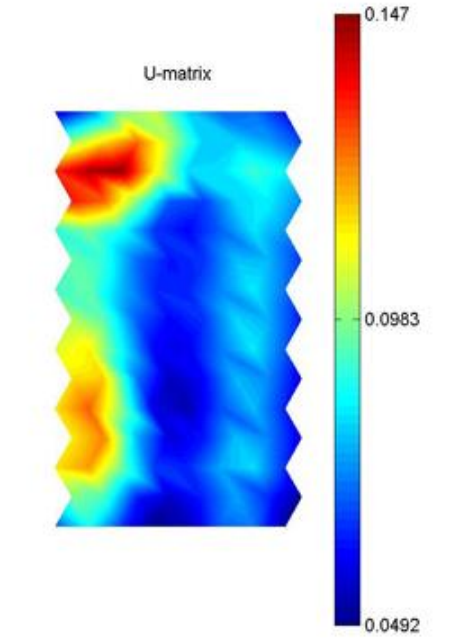






Discussion of the charter for the Data Mining Interest Group DMi - IG

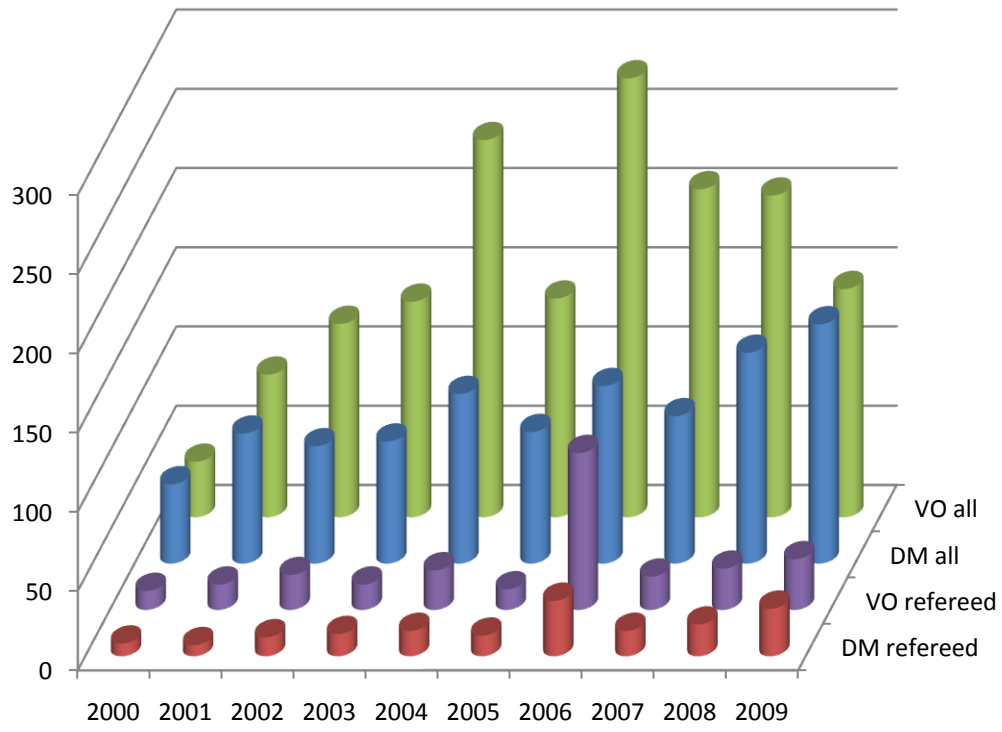
General considerations by
G. Longo and M. Brescia

Draft Charter with inputs from :

K. Borne, C. Donalek, D.de Young, G.S. Djorgovski, M. Graham, O. Laurino, P. Skoda, R. Smareglia, and many others…

Munich, november 2009

Some general considerations on Vobs science – very preliminary (internal circulation only…)



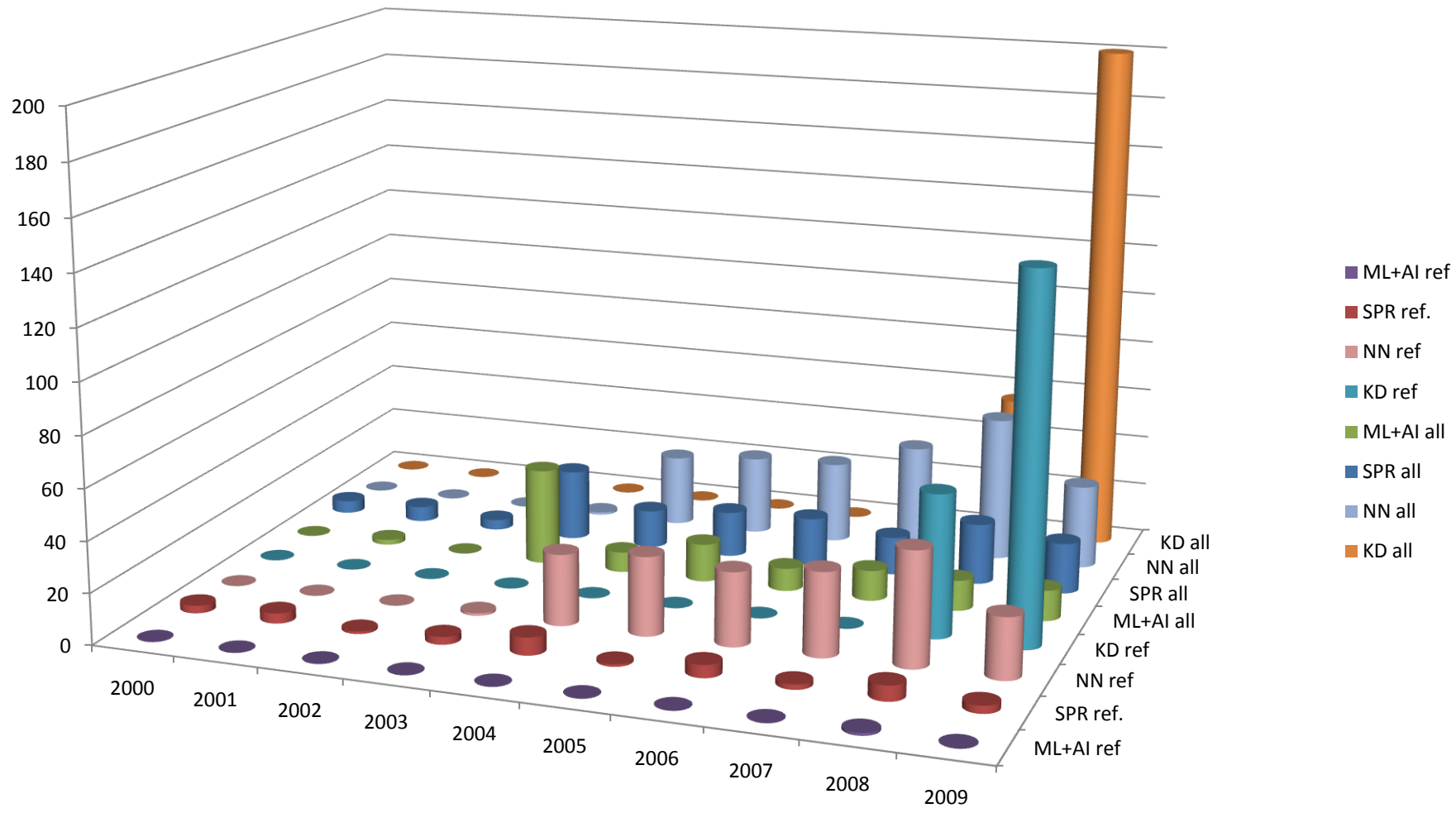
Number of technical/algorithmic papers increases with new funding opportunities

Number of refereed papers remains constant

Overlap between VO and DM refereed > 65%

Little use – few citations

(number of citations increases if you avoid the terms AI, ML or DM in writing the paper)



Out of one thousand papers checked (galaxies, observational cosmology, survey) over the last two years:

DM could be applied or involved in at least 30% of them

Restricted choice of problems: *the situation is not changed much in the last decade*

Tagliaferri et al. 2003	Ball & Brunner 2009	BoK
S/G separation	S/G separation	Y
Morphological classification of galaxies *(shapes, spectra)*	Morphological classification of galaxies *(shapes, spectra)*	Y
Spectral classification of stars	Spectral classification of stars	Y
Image segmentation	-----	
Noise removal *(grav. waves, pixel lensing, images)*	-----	
Photometric redshifts *(galaxies)*	Photometric redshifts *(galaxies, QSO's)*	Y
Search for AGN	Search for AGN and QSO	Y
Variable objects	Time domain	
Partition of photometric parameter space for specific group of objects	Partition of photometric parameter space for specific group of objects	Y
Planetary studies (asteroids)	Planetary studies (asteroids)	Y
Solar activity	Solar activity	Y
Interstellar magnetic fields	----	
Stellar evolution models	----	

Limited number of problems due to limited number of reliable BoKs

Bases of knowledge

(set of well known templates for supervised (training) or unsupervised (labeling) methods

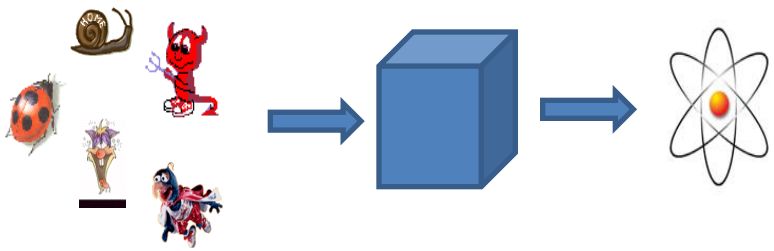
So far

- Limited number of BoK (and of limited scope) available
- Painstaking work for each application (es. spectroscopic redshifts for photometric redshifts training).
- Fine tuning on specific data sets needed (e.g., if you add a band you need to re-train the methods)

Bases of knowledge need to be built automatically from Vobs Data repositories

Community believes AI/DM methods are black boxes

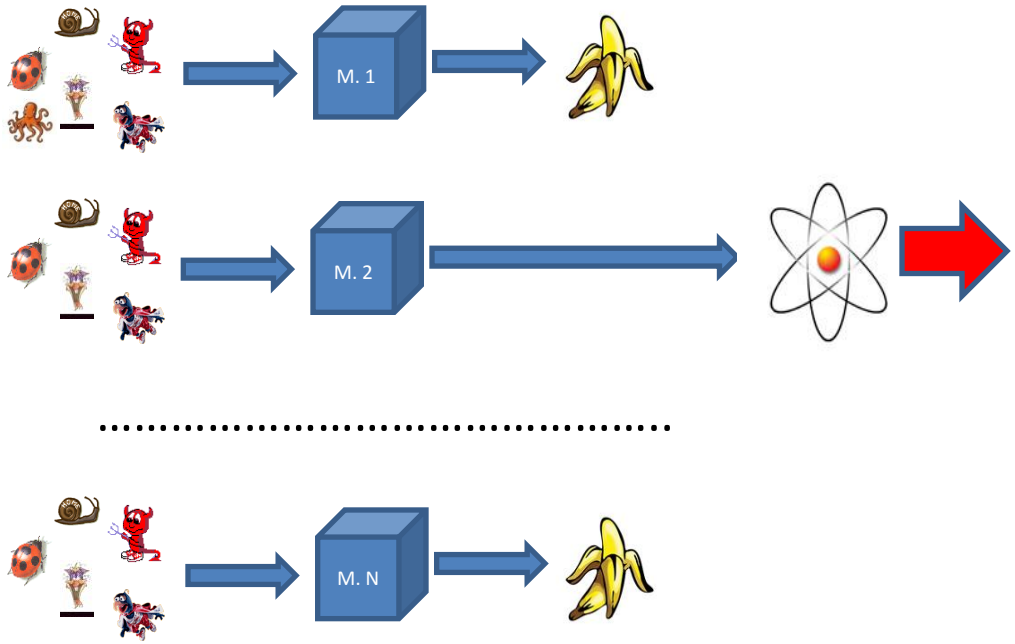
You feed in something, and obtain patters, trends, i.e. knowledge….



Exposed to a wide choice of algorithms to solve a problem, the r.m.s. astronomer usually panics and is (rightly) not willing to make an effort to learn them ….

The r.m.s astronomer doesn't want to become a computer scientist or a mathematician
(large survey projects overcome the problem)

Tools must run without knowledge of GRID/Cloud, no personal certificates, no deep understanding of the DM tool etc.)



Implementation of a second generation of tools

DRAFT FOR A CHARTER
INTEREST GROUP IN DATA MINING (DMI)

During the Strasbourg InterOp Meeting it emerged the need for an Interest Group on Data Mining (KDD-IG) as an indispensable step to bridge the Virtual Observatory Infrastructure with the expected VObs science. In fact, *"…Data mining, or KDD, is the semi-automatic discovery of patterns, associations, changes, anomalies, and statistically significant structures and events in data. In other words, traditional data analysis is assumption driven as a hypothesis is formed and validated against the data. Data mining, in contrast, is discovery driven as the patterns are automatically extracted from data…."*

As such, Data Mining can be considered as the "frontier" of VObs enabled science since it represents the only way to capture and reveal the scientific knowledge (patterns, trends, correlations, etc.) hidden behind the complexity of Massive Data Sets.

Data Mining is a rapidly evolving set of methodologies which needs to be imported under the VObs umbrella and not just another application. As such, DM cannot be just a tool or a suite of tools offered by a group of developers to a "passive community". Data Mining involves a large number of researchers across many domains. The astronomical community, who has only recently entered the MDS era, makes use of just an handful of methods and tools which very often are far from optimal. The synergy of different expertise present in the IVOA makes it the ideal arena for exploring new and more modern approaches.

DMi, more than what happens for other IGs and WGs (more middleware oriented), requires a strong and continuous interaction with the scientific community which, besides testing the proposed solutions, methods and tools, will also provide feedbacks and inputs aiming at extending the scientific capabilities of the VObs.

The DMi-IG will be at the intersection of many other IVOA working and interest groups: applications, semantics, VOEvent, Data Modeling, Grid & Web Services, Resourse registry, semantics. This cross-discipline nature is also the main reason to create a specific IG. Data Mining, in fact, addresses sophisticated and extreme modes of usage which require a careful orchestration and fine tuning of standards, methods and tools provided by the other IVOA WGs and IGs. Typical examples being the automatic extraction of bases of knowledge from VObs archives using VObs ontology; the transparent access to large computational facilities regardless the computational paradigm; the automated switching from asynchronous to synchronous mode of data access; the extreme usage of workflows and advanced visualization methods. Furthermore, effective DM requires the possibility for a non experienced user to contribute, or at least seamlessly use under the VObs infrastructure, his/her own DM routines and methods, an item which puts strong requirements on security issues and opens new problems for ticketing and scheduling.

In other words, the DMi IG will provide feedbacks to the solutions implemented by the WG's and, by posing new operational problems, will stimulate the development and adoption of new solutions and standards.

We also wish to stress that, in ultimate analysis, the goal of the DMi-IG is to allow the VObs to produce new scientific knowledge publishable in astronomical journals. On the one end its activities will contribute to demonstrate to the community the power and necessity of federated access to the vast VObs universe of data and, on the other, DMi will illustrate the power and performance of data mining algorithms to facilitate and accelerate astronomical discovery within this data universe.

2. The charter for the IVOA-DM IG

We will develop and test scalable data mining algorithms and the accompanying new standards for VObs interfaces and protocols, so that these algorithms can be discovered and used transparently within VO science workflows or in standalone data exploration applications.
Therefore the activities of the DMi-IG will be:

- Define an ontology of the DM tasks required by the astronomical community. This ontology will be used to define programming and documentation standards.
- Define an inventory of existing methods relevant for astrophysical applications (more than 100 new DM models and methods appear every month on specialized journals).
- Define reference data sets to be used for comparing, debugging and testing methods and tools.
- Implement –using available VObs standards and methods- general purpose data exploration and data mining methods which will allow the general user to seamless exploit the complex data repositories offered by the VObs.
- Provide/receive feedbacks to/from the other WG in order to improve the usability of VObs tools and standards.
- Provide/receive from the community feedbacks to improve both the usability and the potentialities of the VObs.
- Define and pursue specific science case which will be used to showcase the VObs capabilities to the community

More important than anything else, we wish to use this IG as an arena where different groups can share experiences and plan future developments.

Appendix.
Specific tasks which will be addressed during the first period (t.b.d).

Definition of a taxonomy of DM models. This taxonomy will contribute to the Standard Vocabulary of the Semantics WG.

Definition of the requirement which a DM model needs to match in order to be imported
under the VObs standards.

Inventory of existing DM models of relevant astrophysical interest.

Definition of standard template data sets for DM models test and debugging.
Definition of standard data sets to be used as bases of knowledge for debugging and test of supervised methods.

Definition of procedures to extract and validate robust bases of knowledge from the VObs data archives using the VObs ontology.

Study of the scalability of DM models under different computing infrastructures (definition of best benchmarks).