CHARTER FOR THE KDD-IG
INTEREST GROUP IN KNOWLEDGE DISCOVERY IN DATABASES

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 VO 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…."*[1]

As such, Data Mining (DM) can be considered as the "frontier" of VO 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 VO 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, which has only recently entered the Massive Data Sets era, makes use of just a 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.

KDD-IG requires a strong and continuous interaction with the scientific community which, besides testing the proposed solutions, methods, and tools, will also provide feedback and inputs aiming at extending the scientific capabilities of the VO.

The KDD-IG will interfaces to many other IVOA working and interest groups: Applications, Semantics, VOEvent, Data Models, Grid & Web Services, and Resource Registry. This cross-discipline nature is also a primary 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 are the automatic extraction of bases of knowledge from VO archives using VO ontologies; the transparent access to large computational facilities regardless the computational paradigm; the automated switching from asynchronous to synchronous mode of data access; and the extreme usage of workflows and advanced visualization methods. Furthermore, effective KDD requires the possibility for an inexperienced user to contribute, or at least seamlessly use under the VO infrastructure, his/her own KDD routines and methods. This situation puts strong requirements on security issues and opens new problems for ticketing and scheduling.

In other words, the KDD-IG will provide feedback to the solutions implemented by the WG's

[1] Data Mining Scientific and Engineering Applications, Grossman,Kamath & Kumar http://www-users.cs.umn.edu/~kumar/Presentation/sc2001.html

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 KDD-IG is to allow the VO 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 VO universe of data and, on the other, KDD-IG 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 VO 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 KDD-IG will be:

1. Support the definition of an ontology of the KDD tasks required by the astronomical community. This ontology will be used to define programming and documentation standards.
2. Make an inventory of existing methods relevant for astrophysical applications (more than 100 new KDD models and methods appear every month on specialized journals).
3. Identify reference data sets to be used for comparing, debugging and testing methods and tools.
4. Foster the implementation, using available VO standards and methods, of general purpose data exploration and data mining methods which will allow the general user to seamless exploit the complex data repositories offered by the VO.
5. Provide/receive feedbacks to/from the WGs in order to improve the usability of VO tools and standards.
6. Provide/receive from the community information to improve both the usability and the potentialities of Data Mining tools under the VO.
7. Define and pursue specific science cases which will be used to showcase the VO 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 (12 months from the acceptance).

1.1 Definition of a taxonomy of Data Mining models. This taxonomy will contribute to the Standard Vocabulary of the Semantics WG2.

1.2 Definition of the requirements which a Data Mining model needs to match in order to be imported under the VObs standards.

1.3 Inventory of existing Data Mining models of relevant astrophysical interest.

1.4 Definition of standard template data sets for Data Mining models test and debugging.

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

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

1.7 Study of the scalability of Data Mining models under different computing infrastructures (definition of best benchmarks).