# Distributed Data Mining in the National Virtual Observatory

Kirk D. Borne[a]
[a]Institute for Science & Technology, Raytheon (IST@R)
Raytheon Information Technology and Scientific Services, Lanham, MD 20706
and
Code 630, NASA Goddard Space Flight Center, Greenbelt, MD 20771
mailto:Kirk.Borne@gsfc.nasa.gov

## ABSTRACT

The astronomy research community is about to become the beneficiary of huge multi−terabyte databases from a host of sky surveys. The rich and diverse information content within this "virtual sky" and the array of results to be derived therefrom will far exceed the current capacity of data search and research tools. The new digital surveys have the potential of facilitating a wide range of scientific discoveries about the Universe! To enable this to happen, the astronomical community is embarking on an ambitious endeavor, the creation of a National Virtual Observatory (NVO). This will in fact develop into a Global Virtual Observatory. To facilitate the new type of science enabled by the NVO, new techniques in data mining and knowledge discovery in large databases must be developed and deployed, and the next generation of astronomers must be trained in these techniques. This activity will benefit greatly from developments in the fields of information technology, computer science, and statistics. Aspects of the NVO initiative, including sample science user scenarios and user requirements will be presented. The value of scientific data mining and some early test case results will be discussed in the context of the speaker's research interests in colliding and merging galaxies.

**Keywords**: data mining, scientific databases, distributed computing, interoperability, astronomical data, XML

## 1. INTRODUCTION − THE NATIONAL VIRTUAL OBSERVATORY (NVO)

The astronomical scientific research community is the beneficiary of huge multi−terabyte databases from a host of sky surveys (e.g., Sloan Digital Sky Survey, Two−Micron All−Sky Survey, Digital Palomar Observatory Sky Survey, and more). The rich and diverse information content within this "virtual sky" and the results to be derived therefrom will far exceed the clearly demonstrable results from the first Palomar Observatory Sky Survey (POSS), whose omnipresence in astronomy libraries, departments, and observatories worldwide is undoubtedly one of the great scientific legacies from the mid−20th century. Additional large astronomical sky surveys continue to be carried out for specialized purposes. These new digital surveys have the potential of exceeding the usefulness of the POSS by orders of magnitude! To fulfill this promise, it is imperative that the astronomical and computational sciences research communities work together to mobilize efficient and effective data archiving, access, mining, and distribution resources. This is critical for several reasons: (1) for the continued success and strength of scientific research programs that now routinely handle terabytes of data; (2) to demonstrate and preserve the legacy value of the tremendous investment of resources that have gone into these large scientific data−producing projects; and (3) to reap the maximum scientific benefit from those investments. The opportunities are vast, and the opportunity is now.

This grand vision encompasses more than a collection of projects to develop and enhance infrastructure. The result of these initiatives will be the flowering of the ambitious distributed on−line National Virtual Observatory (NVO, at http://www.us−vo.org/), which will represent a significant evolution in the manner by which astrophysical research is conducted in the new millennium. In fact, the NVO is not national at all, but it is a growing international effort in the astrophysics research community, whose partners are working together toward common goals through the International Virtual Observatory Alliance (IVOA, at http://www.ivoa.net/).

The common feature of the surveys mentioned above is that they will produce massive databases, containing hundreds of gigabytes to tens of terabytes of data products. Future surveys may produce 10's of petabytes of data (e.g., the



Data Mining and Knowledge Discovery: Theory, Tools, and Technology V, Belur V. Dasarathy, Editor, Proceedings of SPIE Vol. 5098 (2003) © 2003 SPIE · 0277-786X/03/$15.00

Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms

Large Synoptic Survey Telescope project, LSST, at http://www.lssto.org/). Enormous scientific databases such as these contain information nuggets that could go undiscovered simply because the databases are too large to be thoroughly investigated even with the support of today's sophisticated database tools. Significant efforts are being made now to extend traditional data access techniques with state−of−the−art research in data mining and visualization. Without this, the scientific research community will be overrun by volumes of data that cannot be assimilated. Scientific discoveries lie hidden among the myriad of entries in not just single databases but most especially among the cross−correlated contents of heterogeneous distributed databases. Techniques are being developed that will be applicable to large, distributed, federated databases within this new distributed data system (NVO) as well as being applicable to scientific knowledge discovery within the large legacy isolated data collections.

Milt Halem *et al*. (2000) argued in their "Data Crisis" white paper that the next evolutionary step in the current data explosion is the transformation of data into knowledge−based information. In other words, the end−goal of data mining and archival research in large databases is not the collection and curation of the data bytes alone but the extraction of the rich information and knowledge content represented therein. Thus, the application of data mining technologies in general, and distributed data mining algorithms in particular, must become an inherent feature of large scientific database research in the 21st century.

## 2. DATA MINING FOR SPACE SCIENCE AND THE NVO

We have compiled a comprehensive data mining resource guide for space science (Borne 2003; Borne and Cheung 2002), which also has application to many other fields of scientific research involving very large databases. We have also developed a series of scientific user requirements and science user scenarios for data mining in an NVO environment (Borne 2000a; Borne 2000b), some of which are presented below. A much more thorough discussion of the challenges, requirements, and goals of a scientific data mining and visualization environment for the NVO can be found in the workshop report by Mann *et al*. (2002): "Scientific Data: Mining, Integration, and Visualisation".

To assist in the design of a data mining and knowledge discovery research environment for astronomical scientific research, we have identified four major categories of system−level user requirements. The data mining system must be able to perform these functions efficiently and effectively for very large databases, which are assumed to be geographically distributed and institutionally federated (i.e., each institution maintains ownership and control over their own metadata, database formats, user interfaces, archive architecture, and services). These basic user requirements are:

1. Object Cross−Identification − It is a well known, notorious "problem" in astronomy that individual objects are often called by many names. It is an even more serious problem now with the advent of all−sky surveys in many wavelength bands − the same object will appear in a multitude of catalogs, observation logs, and databases, and it will almost certainly be catalogued uniquely in each case according to a survey−specific naming convention. Identifying the same object across these data collections is not only crucial to understanding its astrophysical properties and physical nature, it is also one of the most basic inquiries that an astronomer will make of these all−sky surveys. The problem of isolating these object cross−IDs across multiple surveys reduces in its simplest form to finding ***spatial associations*** of given objects among a collection of catalogs. Our system will need to be maximally optimized to do this since this will be the most fundamental query to be addressed to the databases.

2. Object Cross−Correlation − After specific cross−IDs have been established, there is a wealth of astrophysical "What if?" queries that can be applied to the observational parameters in the databases. In the most general sense, these correspond to the application of ***classification***, ***clustering***, and ***regression*** (cross−correlation or anti−correlation) algorithms among the various database parameters. These correlations need not be confined to two parameters; they may be highly dimensional. Correlating the observational properties of a class of objects in one astronomical database with a different set of properties for the same class of objects in another database is the essence of NVO−style "virtual observational astronomy". Our data mining query system must be capable of handling such large−scale distributed "What if?" inquiries.



Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms

3. Nearest−Neighbor Identification − In addition to identifying spatial associations across multiple distributed astronomical data archives, it is also scientifically desirable for the system to have the capability of finding associations (***nearest neighbors***) in the highly dimensional space of complex astronomical observational parameters − to search for dense ***clusterings*** and ***associations*** among observational parameters in order to find new classes of objects or new properties of known classes. These subsets of objects with matching sets of parameters correspond to ***coincidence associations***. They offer a vast potential for scientific exploration and will provide a rich harvest of knowledge discovery. The distributed database indexing and mining system will thus need to include robust clustering (`groupby`) algorithms and aggregation (`aggregate`) schemes with optimal indexing in order to facilitate such nearest−neighbor searches.

4. Systematic Data Exploration − In very large databases, there are likely to be significant subsets of data (regions of observational parameter space) that have gone largely unexplored, even by the originating scientific research teams that produced the data. Archival data researchers accessing the databases will want to apply a wide variety of ***event−based*** and ***relationship−based*** constraints to the data. The exploration process involves numerous iterations on "What if?" queries. Some of these queries will not be well constrained or targeted, and will thus produce gigabyte−sized outputs. For most situations, this will be too large to be useful. So the system will need to allow query previews (to reveal the size of the results prior to retrieval) and to allow iterations on preceding queries (either to modify or to tighten the search constraints). For this type of interactive iterative process to work well, rapid return of results will be required, and this implies a well indexed collection of searchable databases.

As we will describe later in more detail, one of the great unexplored frontiers of astrophysics research is the time domain. We have a good idea of what is varying and what is not, as a result of many centuries of humans staring at the sky, with and without the aid of telescopes. But, there is so much more possibly happening that we are not aware of at the very faintest limits simply because we have not explored the sky systematically night after night on a large scale. When an unusual time−dependent event occurs in the sky (e.g., a gamma−ray burst, supernova, or in−coming asteroid), astronomers (and others) will not only want to examine spatial coincidences of this object within the various surveys, but they will also want to search for other data covering that same region of the sky that were obtained at the same time as this new temporal event. Because of the time−criticality and potential for huge scientific payoff of such "follow−up" observations of transient phenomena, the query−access system must also be able to perform time−based searches very efficiently and very effectively (i.e., to search all of the distributed databases as quickly as possible). One does not necessarily know in advance if such a new discovery will appear in any particular waveband, and so one will want to examine all possible astronomical sky surveys for coincidence events. Most of these "targets of opportunity" will consequently be added immediately to the observing programs of many ground−based and space−based astronomical telescopes and observatories. By providing rapid turn−around on object cross−identifications and perhaps an improved celestial sky coordinate position for the object, our system will provide critical input to and have an immediate impact on on−going astronomical research programs worldwide. We are thus investigating the use of existing time−domain databases for the development of a working distributed data mining system for time series astronomical data.

Another model for data mining that we will apply to the development of the scientific data mining system is one that enables the researcher to carry out data mining thrusts in any one of these four mutually orthogonal directions:

1. Known events / known algorithms − use existing physical models (i.e., descriptive models, algorithms) to locate known phenomena of interest either spatially or temporally within a large database.

2. Known events / unknown algorithms − use pattern recognition and clustering properties of data to discover new observational (in our case, astrophysical) relationships (i.e., unknown algorithms) among known phenomena.

3. Unknown events / known algorithms − use expected physical relationships (predictive models; known algorithms) among observational parameters of astrophysical phenomena to predict the presence of previously unseen events within a database.

4. Unknown events / unknown algorithms − use thresholds or outlier detection techniques to identify new transient or otherwise unique ("one−of−a−kind") events and therefore to discover new astrophysical phenomena.



Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms

## 3. TECHNOLOGIES FOR DISTRIBUTED DATA MINING

Among the technologies that are being explored to facilitate distributed data mining, XML must be considered to be the most versatile and the one being most actively researched. Some call XML "the language of interoperability." The NVO is all about interoperability. In fact, the essence of the NVO projects that are now being funded is in the development of the middleware and metadata standards that will enable this interoperability. The NVO efforts are not focussed (much) on hardware (we are not building a brick−and−mortar archive, but a click−and−order virtual data system), and not focussed (much) on the user interfaces. The Virtual Observatory communities worldwide have created XML languages for distributed astronomical database queries, for query results, for service description registries, for archive content registries, and more (i.e., for the middleware). One of these is the VOTable XML language for distributed database queries and for the query results, which are returned to the querying system in an XML−based tabular form (http://www.us−vo.org/VOTable/). A generalized XML−based extensible data format (XDF) has also been created, for use in any scientific discipline (http://xml.gsfc.nasa.gov/XDF/), not just for the NVO and not just for astronomical data.

Interoperable services for distributed archival access have also been prototyped through the NASA ISAIA (Interoperable Systems for Archival Information Access) project (http://heasarc.gsfc.nasa.gov/isaia/). More recently, the efforts have focussed on the Web Services model, in which "software communicates with software", allowing for resource and data discovery, service descriptions, and user−transparent distributed data system interoperability. We are thereby able to benefit from and take advantage of the huge investment of corporate and federal R&D efforts into developing XML and Web Services standards. These will enable NVO to function in much the same way as they will enable e−commerce and e−government to function, though the scientific research functionality will have some extraordinarily heavy data−throughput requirements.

In addition to the above challenges, it is clear that the volumes of data that we are handling will necessarily prohibit the movement of the data to the query engine, in the distributed environment. It will therefore be necessary to "ship the code, not the data". The MOCHA (Middleware based On a Code−sHipping Architecture) project at the University of Maryland (N.Roussopoulos) is one such implementation that has been considered. However, the overwhelming favorite in this arena is Grid Computing, and the NVO team is collaborating with members of the Global Grid Forum (www.ggf.org) to develop distributed computing environments and virtual data creation for NVO science.

The full implementation of a virtual data system for astronomy (or for any other scientific discipline) must include linkages in an end−to−end manner across the full range of worldwide scientific information systems. One of those components is the scientific literature itself. It is fortunate that most scientific journal publications are produced in SGML format, which is a broader generalization of such meta−languages as XML. Thus, it is at least as easy to integrate the publication information systems into the data mining research environment as any other XML−based science data archive. To this end, we have created an XML−based scientific manuscript markup language (AXML = Article XML Markup Language; Shaya *et al*. 2002). This could replace LaTeX style files and provide inputs to publishers that are almost already in their final preferred format for publication. The XML−based manuscripts can then be "displayed" through any number of stylesheets: for Postscript, for PDF, for HTML, for books, for journals, or back into LaTeX, if desired. We have created simple style sheets for this already. The publication databases can then become part of the end−to−end distributed data archive system that can be subjected to content−based (semantic) data mining queries (e.g., "find me all there is to know about this subject", which will thus include published papers, archival data, metadata collections, or other documentation).

As a meta−language (a language to create new languages), XML thus provides a rich variety of opportunities for distributed data mining within the NVO and other distributed data analysis environments.

## 4. NVO@Home = DATA MINING FOR SCIENCE EDUCATION

Several astronomical research groups have independently and collectively come up with an idea for a distributed data mining system for science education as well as for scientific research. The idea has been simply labeled "NVO@Home" (or VO@Home). It is nothing more than a concept right now, with nothing implemented. But, it is



Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms

patterned in concept after the wildly successful and famous SETI@Home screensaver program, used to search for patterns in radio signal data from space, to look for evidence of extraterrestrial intelligences. The equivalent VO@Home tool would distribute large quantities of astronomical data to users' desktops for analysis in a screensaver mode (i.e., tapping otherwise idle unused computing resources). Most likely these would be data obtained during the previous night (or week) from a large sky survey somewhere in the world. The tool would search for anomalous or time−varying or moving objects, which may be comets, asteroids, variable stars, novae, supernovae, quasars, or "things that go bump in the night". The chances for discovery are very high (unlike SETI@Home), and thus the appeal should be high for school children, science educators, scientists, data mining experts, and the general public.

One project that is in planning right now that will produce raw material for something like VO@Home is the Large Synoptic Survey Telescope project (LSST: http://www.lssto.org/). This project will produce at least 20 Terabytes of data each night for ten years − a net of several tens of petabytes. The goal is to image the full sky (that which is visible to the telescope at that time of year) over and over again, night after night, producing the first−ever long−term time series database of the night sky. This has sometimes been referred to as "Cosmic Cinematography" (Liu *et al*. 2002). The LSST project represents the first attempt to study the entire sky in the time domain, down to the very faintest observable objects in the sky. There is no telling what we might discover. A recent prototype investigation of a very tiny patch of sky revealed a star that increased in brightness by a factor of 300 times in one night, then returned to its "normal" pre−outburst brightness level thereafter (Djorgovski *et al*. 2000). What startled astronomers the most is not there might be such a type of star in the sky. What was most startling is that subsequent spectral analysis of the star revealed that it is in fact a "normal G dwarf" star, just like our Sun! This begs the questions: ***Did our Sun ever do this? Will our Sun ever do this?*** These are questions that not only astronomers would like to have answered, but even school children and Congresspersons would like to know. A distributed data mining system that utilizes the potentially untapped computing resources of millions of CPUs (similar to SETI@Home) may thus enable magnificent and awe−inspiring astronomical discoveries, which may have a lasting impact on the science education and literacy of the next generation of taxpayers!

As mentioned above, there is no implementation yet of this VO@Home concept. We welcome collaborations with data mining and distributed computing experts toward developing such a prototype, and we welcome collaborations with the science education community toward developing a science and math curriculum using such a data mining system in the classroom − we could train pre−college students in science, math, statistics, and computational sciences through such a tool. We are eager to explore such possibilities with interested research groups.

## 5. A SAMPLE ASTRONOMICAL RESEARCH SCENARIO

As a demonstration of some of the key aspects of distributed data mining that can be applied to astrophysics research problems, we present a case study of one particularly interesting class of galaxies that has not been well studied to−date: the Very−Luminous Infra−Red Galaxies (VLIRGs). Data mining in a variety of multi−wavelength multi−project multi−modal (imaging, spectroscopic, catalog) databases will eventually yield interesting new scientific properties, knowledge, and understanding for the VLIRGs as well as for many other objects of astrophysical significance. Very−Luminous IR Galaxies (VLIRGs) are key to the study of galaxy formation and evolution. On one hand, they offer the opportunity to study how the fundamental physical and structural properties of galaxies vary with IR (infrared) luminosity, providing the link between the extremely chaotic and dynamic Ultra−Luminous IR Galaxies (ULIRGs) and the "boring" set of normal galaxies. On the other hand, VLIRGs are believed to be closely related with recently identified cosmological populations in the sense that VLIRGs should be either the low−redshift analogs or the direct result of the evolution of those cosmologically interesting populations of galaxies that appear only at the earliest stages of cosmic time in our Universe.

Several cosmologically interesting populations have been recently identified by astronomers. These include: (1) the galaxies that comprise the Cosmic Infrared Background (CIB); (2) the high−redshift submillimeter sources at high redshift; (3) the infrared−selected quasars; (4) the "Extremely Red Objects" (EROs) found in Hubble Space Telescope images and other deep−sky surveys; and (5) host galaxies for the extremely luminous Gamma−Ray Burst sources. The primary questions pertaining to these cosmological sources are: What are they? Are they dusty quasars? or dusty star−bursting galaxies? or massive old stellar systems? Some of these classes of objects are undoubtedly ULIRGs, while



Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms

the majority are probably related to the significantly more numerous class of galaxies: the VLIRGs. Our NVO data mining project will enable us to identify signatures of VLIRGs that are indicative of this special class, and thus to distinguish star−bursting galaxies, with high star−formation rates, from Quasars, which are otherwise obscured by the characteristic large dust opacities in these galaxies (heated dust being the source of the IR emission). The results will therefore be applicable to understanding and interpreting the properties of the five cosmological samples listed above. If, as believed, both ULIRGs and VLIRGs represent low−redshift analogs to the high−redshift galaxies, a comprehensive analysis of the VLIRGs is urgently needed. This is clear, for instance, when interpreting the diffuse Cosmic IR Background (CIB). Deep submm surveys have confirmed that IR−luminous galaxies are a significant, if not the dominant, contributor to the CIB. However, the most recent submm observations have been able to identify only ULIRGs and the most luminous VLIRGs. Taking into account the significantly larger density of VLIRGs in the Universe, it appears likely that VLIRGs could contribute a significant fraction of the infrared radiation at high redshifts. Further advance depends on acquiring detailed knowledge of the properties of VLIRGs and of their relation to ULIRGs and normal galaxies. A thorough data mining exercise within the NVO across multiple databases will enable a significant research advantage toward resolving these scientific questions.

We initiated a proof−of−concept NVO science search scenario by attempting to identify potential candidate contributors to the CIB (Cosmic Infrared Background). This is significant since it has been shown that fully one−half of all of the radiated energy in the entire Universe comes to us through the CIB! Our approach (Borne 2000a) involved applying the full power of several distributed on−line databases and the linkages between these databases, archives, and published literature. Our search scenario involved finding object cross−identifications among the various distributed source lists and archival data logs. In a very limited sample of targets that we investigated to test our NVO data mining approach to the problem, we did find one object in common among three geographically distributed databases: a known hyperluminous infrared galaxy (HyLIRG) at moderate redshift harboring a Quasar, which was specifically imaged by the Hubble Space Telescope because of its known HyLIRG characteristics. In this extremely limited test scenario, we did in fact find what we were searching for: a distant IR−luminous galaxy that is either a likely contributor to the CIB, or else it is an object similar in characteristics to the more distant objects that likely comprise the CIB.

The success of this limited proof−of−concept study encourages us to pursue even greater database searches and data mining scenarios. This is not possible today. But, with the full growth of scientific data mining technologies and the full implementation of distributed data system interoperability, along with the integration of these two streams of information technology development, wonderful new scientific discoveries are waiting to be made.

## 6. A DISTRIBUTED DATA MINING PROTOTYPE

We are beginning to explore techniques for distributed data mining that will be applicable to the NVO (e.g., using the research results of Kargupta *et al.*, at http://www.csee.umbc.edu/~hillol/ddm.html). In particular, we hope to develop useful functional test results and working algorithms for other scientists to explore the rich distributed astronomical data archives that will comprise the NVO. We anticipate that XML will play a key role in our developments.

A project that is now underway focusses explicitly on the challenges imposed on a distributed scientific data mining system. The project is funded by NASA, and consequently focusses on NASA space science data and NASA data archives. We are searching the huge distributed NASA space science data archives for one of the rare "beasts" of the night sky: the ULIRGs (Ultra−Luminous Infra−Red Galaxies), which are an order of magnitude brighter than the VLIRGs that we discussed earlier and which literally represent one in a million galaxies in the sky (the proverbial "needle in a haystack" − a perfect challenge problem for data mining). Despite their local rarity, the ULIRGs are known to be extremely significant components of the early Universe, and thus greater knowledge of their properties will enable a greater understanding of our Universe.

ULIRGs represent the most violent phase of galaxy formation in the Universe. They are also pivotal objects in the history of the Universe − they are believed to be the sites of giant galaxy formation, Quasar formation, massive star formation, and heavy element production. ULIRGs are rare in today's Universe (only one in a million), but they are increasingly common and important as we look at fainter and fainter galaxies (i.e., further back in cosmic time). For



Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms

this project, we are searching for these "needles in a haystack" by mining the data within multiple distributed space science data archives (including the Hubble Space Telescope data archive in Baltimore, Maryland, and the Two−Micron All−Sky Survey database at Caltech, plus one or two others). In particular, ULIRGs are characterized by a unique clustering of their optical, Near−Infrared, and Far−Infrared emission properties in the corresponding multi−dimensional parameter space. Learning algorithms will be used to enhance the data discovery rate through further classification and association−mining of ULIRG parameters. The main challenge for this and similar NASA research programs is to conduct data mining in a ***distributed*** environment. Our goal is to address this challenge as a pathfinder project for the new National Virtual Observatory (NVO).

In addition to the application of critical data mining technologies, we are investigating the application of several additional leading information technologies as potential solutions to this problem (e.g., XML for distributed data access; Beowulf Linux Clusters for parallel data mining; genetic algorithms for rapid data modeling; Supervised and Unsupervised Learning algorithms for rapid robust object classification; and other data mining methods, such as cluster−finding, association−mining, Bayesian networks, and decision trees).

In general, within the astronomical research community, most data mining algorithms have been tested only on single data sets or within single data archives. Data mining and cross−correlating science parameters across a suite of geographically distributed data collections will lead to: (1) science database interoperability, and thus (2) far greater science knowledge discovery potential. To achieve the goals of such a project, we are focussing initial efforts on developing scientific and technical requirements in designing, implementing, testing, and operating the distributed data mining research environment. Our tools will be made available to future users of the NVO. For the particular science research scenario that we propose (i.e., mining for ULIRGs), we are developing science requirements for the data mining system. We plan to integrate intelligent data understanding algorithms into the data mining system (through Learning Algorithms), for immediate application to our science research problem and, by extension, to the NVO. We will selectively experiment with a distributed data mining environment that explores three of NASA's large distributed Space Science data collections for the discovery, identification, classification, and interpretation of new ULIRGs from multiple distributed data collections. Using ULIRG properties tabulated in existing astronomical data catalogues, we can compile a training set. We will initially use a subset of known ULIRGs to train our data mining algorithm (e.g., Supervised Learning), and then we will attempt to "re−discover" the remaining subset in order to verify our techniques. Cluster−finding and association−mining algorithms will be used to identify additional unique characteristics of ULIRGs that can be mined in larger data sets (present and future). These intelligent data mining techniques can then be deployed as a virtual "science instrument" within the NVO in order to find similar "outliers" within future NVO databases. This project will increase our understanding of the ULIRG class, as a step toward the ultimate goal of understanding the structure and evolution of the Universe.

## 7. DISTRIBUTED DATA MINING WITHIN THE NVO AND BEYOND

The NVO will change the way that astronomical research is conducted in the future. For the first time since astronomers began making observations of the night sky, we have a complete multi−spectral survey database of the entire sky that allows us to conduct this kind of distributed data mining. The volume and complexity of the corresponding data products certainly justify the establishment of the NVO and related data mining technologies. But there is an even more compelling rationale for such a system: even though astronomers have archived the sky, this gigantic "virtual telescope" currently lacks good instrumentation (e.g., tools for distributed data mining). In addition to large all−sky surveys, there are other future astronomy projects that are being designed to do targeted high−resolution observations of specific objects and specific classes of objects. Distributed data mining techniques that operate across a virtual data system of multi−wavelength data collections will be needed in order to support science program definition, target selection, and observation planning for these large complex research projects.

The proven technologies that will result from the NVO distributed data mining projects may eventually be applied to future deep space missions in the exploration, analysis, and interpretation of their voluminous science data products. As future science mission data streams grow exponentially, it may be desired that on−board autonomous data mining and intelligent analysis systems perform preliminary processing steps and thus provide feedback to the science planning system. An "intelligent data understanding" capability may be used in real−time to mine future science data



Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms

streams, for the purposes of autonomous on−board science goal monitoring − for example: "continue this observation", or "stop and look elsewhere", or "stop and send results to another spacecraft" (such as a member of a constellation spacecraft system), or "transmit intermediate results for immediate analysis of an interesting event". In general, we expect these data mining techniques to enhance various space mission data−collecting/mining/analysis modes in the future, including distributed "sciencecraft" (Satellite Constellations), remote sensors (or sensor webs), and geographically distributed science data archives (e.g., NVO). In essence, this is **information and data fusion on a grand scale!**

## 8. SUMMARY

The NVO (National Virtual Observatory) will enable seamless user access to astronomical data from diverse geographically distributed data sources. As the volume of astronomical data resources grows astronomically, one must resort to sophisticated data mining techniques. Data mining is defined as "an information extraction activity whose goal is to discover hidden knowledge in large databases." This is the essence of scientific research and discovery. We envision data mining as a key component of future astronomical research projects involving large distributed databases. To support this "new science" of data−rich astronomy, we began by exploring various data mining techniques for astronomy, and we have compiled a comprehensive Scientific Data Mining Resource Guide (available from http://nvo.gsfc.nasa.gov/nvo_datamining.html). The next step is to apply these techniques to the distributed virtual data system that will be called the ***International Virtual Observatory!***

## 9. ACKNOWLEDGMENT

Support for this work was provided in part by NSF through Cooperative Agreement AST0122449 to the Johns Hopkins University, and in part by research award 749−10−02 from the NASA Code R Computing, Information, and Communications Technology (CICT) Intelligent Systems Program to NASA’s Goddard Space Flight Center.

## 10. REFERENCES

1. K. D. Borne, "Science User Scenarios for a VO Design Reference Mission: Science Requirements for Data Mining", *Virtual Observatories of the Future*, pp. 333−336, http://arXiv.org/abs/astro−ph/0008307, 2000a.
2. K. D. Borne, "Data Mining in Astronomical Databases", *Mining the Sky*, pp. 671−673, http://arXiv.org/abs/astro−ph/0010583, 2000b.
3. K. D. Borne, "Data Mining Resources for Space Science", http://nvo.gsfc.nasa.gov/nvo_datamining.html, 2003.
4. K. D. Borne, and C. Y. Cheung, "Science Data Mining Resources for the National Virtual Observatory (NVO)", *BAAS*, 199, 10.05, 2002.
5. S. G. Djorgovski, et al., "Exploration of Large Digital Sky Surveys", *Mining the Sky*, pp. 305−322.
6. M. Halem et al. "Data Crisis White Paper", http://www−sisn.jpl.nasa.gov/ISSUE51/halem.html, 2000.
7. C. T. Liu, K. Borne, C. Stubbs, J. A. Tyson, & the LSSTO Collaboration, "Cosmic Cinematography with the LSSTO", *BAAS*, 199, 101.09, 2002.
8. R. Mann, R. Williams, M. Atkinson, K. Brodlie, A. Storkey, and C. Williams, "Scientific Data: Mining, Integration, and Visualisation", http://www.anc.ed.ac.uk/sdmiv/, 2002.
9. E. Shaya, K. Borne, B. Thomas, and C. Y. Cheung, "Publishing Scientific Articles in XML", *BAAS*, 199, 10.09 2002.



Downloaded from SPIE Digital Library on 25 Oct 2010 to 129.174.97.34. Terms of Use: http://spiedl.org/terms