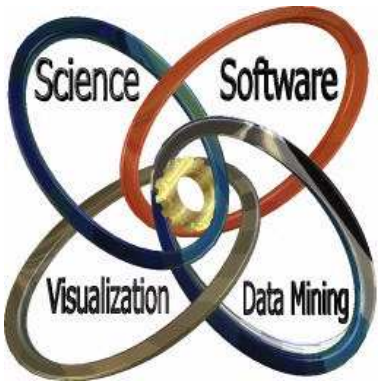
Scientific Data
Mining, Integration, and Visualization



Bob Mann[1,2]
Roy Williams[3]
Malcolm Atkinson[2]
Ken Brodlie[4]
Amos Storkey[1,5]
Chris Williams[5]

[1]Institute for Astronomy, University of Edinburgh, UK
[2]National e-Science Centre, UK
[3]California Institute of Technology, USA
[4]School of Computing, University of Leeds, UK
[5]Division of Informatics, University of Edinburgh, UK

Report of the workshop held at the e-Science Institute, Edinburgh, 24-25 October 2002
This report is at ***http://umbriel.dcs.gla.ac.uk/NeSC/general/talks/sdmiv/report.pdf***

National e-Science Centre

Executive Summary

This report summarises the workshop on *Scientific Data Mining, Integration and Visualization* (SDMIV) held at the e-Science Institute, Edinburgh (eSI[1]) on 24-25 October 2002, and presents a set of recommendations arising from the discussion that took place there. The aims of the workshop were three-fold: (A) To inform researchers in the SDMIV communities of the infrastructural advances being made by computing initiatives, such as the Grid; (B) To feed back requirements from the SDMIV areas to those developing the computational infrastructure; and (C) To foster interaction among all these communities, since the coordinated efforts of all of them will be required to realise the potential for scientific knowledge extraction offered by e-science initiatives worldwide.

The workshop had about fifty participants, ranging from software engineers developing Grid infrastructure software, to computer scientists with expertise in data mining and visualization, to application specialists from a wide range of disciplines, including astronomy, atmospheric science, bioinformatics, chemistry, digital libraries, engineering, environmental science, experimental physics, marine sciences, oceanography, and statistics. It was felt that further meetings should be held, to bring together the SDMIV community, or subsets thereof: the participants felt that the overlapping interests of the communities represented at the workshop made this group more than the sum of its parts, and that future interaction between these communities would be very beneficial.

The workshop produced the following Recommendations, which are detailed in Section 2.7.

R1. **The use of XML for scientific data is recommended, as it aids interoperability and flexibility.** (Section 2.1)

R2. **Research should be undertaken into ways to reference remote data objects that are more flexible and robust than URLs and FTP addresses.** (Section 2.2)

R3. **A new framework of interoperability is emerging from the Grid community, and scientists should build their software to benefit from these standards.** (Section 2.3)

R4. **Libraries of interoperable scientific data mining and visualization services should be built to this Grid standard.** (Section 2.3)

R5. **A mechanism should be sought whereby the peer-reviewed publication of datasets can be made part of the standard scientific process.** (Section 2.4)

R6. **A registry of fiducial datasets should be created and maintained, to be used in the testing of data mining and visualization tools.** (Section 2.4)

R7. **Better methods should be sought for collaborative working in e-science.** (Section 2.5)

R8. **A set of tutorials should be created and maintained, for introducing application scientists to new key concepts in e-science.** (Section 2.6)

R9. **A report should be produced on the data mining requirements of e-science application areas.**

R10. **A report should be produced on the visualization requirements of e-science application areas.**

R11. **A registry of existing data mining resources should be created and maintained.**

R12. **A registry of existing visualization resources should be created and maintained.**

1. Introduction

1.1. Motivation for the Workshop

The importance of data and knowledge extraction in science is growing rapidly. Fields as diverse as bioinformatics, geophysics, astronomy, medicine, engineering, meteorology and particle physics are facing an exponentially increasing volume of available data, as improvements in computational infrastructure make the results from an ever larger number of data and computing resources accessible from the scientist's desktop. A prime challenge for e-science is to enable the effective extraction, integration, exploration, analysis and presentation of knowledge from the data avalanche, so that the scientist can exploit its potential. This requires the coordinated efforts of a number of different communities; from the software engineers building the computational infrastructure that brings together the data from a myriad of sources worldwide, to the computer scientists developing algorithms to aid their integration and exploration, to the application specialists, who will extract scientific knowledge from the data and have to define the metadata standards within each discipline that will make that possible.

1.2. The current status of Grid and e-science initiatives

This is a good time to bring those communities together for a workshop on *Scientific Data Mining, Integration and Visualization* (SDMIV), because of the current status of the UK e-science programme and of Grid computing developments internationally.

Within the past year, the focus in the Grid computing world has shifted from the distributed file manipulation systems that provided the Grid's initial science drivers, to the concept of Grid Services, part of the Open Grid Services Architecture (OGSA[2]). OGSA takes the Web Services[3] model, which is becoming widespread within commercial computing, and supplements it with the "statefulness" needed for the creation of persistent, compound, customized services which can be deployed in a distributed computational environment. The advent of the Grid Services paradigm – and the adoption of its basic principles by many of the major players in commercial computing – marks a significant milestone in the development of the Grid concept, by setting a widely agreed framework for future development.

In the domestic context, work by the various pilot and demonstrator projects funded through the UK e-science programme has started to identify many of the practical problems relating to the production of software systems to facilitate e-science, as well as developing a more detailed understanding of the requirements of scientists in a wide range of application areas. More generally, the concept of e-science is taking hold, as is its emphasis on collaborative working. This is a catalyst to new interactions between communities which have had little contact hitherto, as the commonalities between and differing expertise of various disciplines are becoming recognised.

1.3. SDMIV basics

1.3.1. Scientific Data

Much of the scientific data discussed at the workshop fell into three categories, and, while these do not represent an exhaustive list of scientific data types, much of the technology discussed in the meeting was directed to them. The three categories are:

- The *datacube*, or *array,* class - meaning an annotated block of data in one, two, or more dimensions. This includes time-series and spectra (one dimensional); images, frequency-time spectra, etc (two-dimensional); voxel datasets and hyperspectral images (three-dimensional), and so on. The highly-optimised chips of modern computers handle these data structures well.
- *Records*, or *events*, collected as a *table*. Also known as *multi-parameter* data. These datasets may come directly from an instrument (for example in a particle accelerator) or may be derived by picking features from a datacube (when stars are identified from an astronomical image). Relational databases hold these data effectively.

- Sequences of symbols, for example a biological gene is represented by a sequence of symbols from the set {*ACGT*}. Special sequences are given new names, giving rise to new types of symbols. Tools that operate on such sequences include pattern matching and alignment.

While there has been much emphasis on the sheer bulk of scientific data on the horizon (terabytes, petabytes, ...), there is an orthogonal problem looming: that of data diversity. As archive-based research grows in importance, so each research project creates and publishes the resulting dataset, often using idiosyncratic representations. A future challenge is convincing the community – or communities, as this may only be possible at the level of individual disciplines – of the benefits of standard representations, and of the importance of recording the provenance of data. Often much valuable knowledge exists in the union of several datasets through *data federation,* and for this to be done in a meaningful manner requires adequate *metadata*, describing the individual datasets and the method used to federate them.

1.3.2. Data Mining

Many definitions of data mining exist. Hand, Mannila and Smyth[4] defined it as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner", while Han[5] called it "[the] extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases". Naisbett wrote "We are drowning in data, but starving for knowledge!", which certainly describes the situation so far as scientific data mining is concerned, although the problems caused by data volume should not be emphasized to the exclusion of those arising from the heterogeneous and distributed nature of scientific data.

Hand et al. also distinguished between two of the types of summary relationships that can be derived from data, namely models and patterns: a *model* is a global summary of a data set (e.g. linear regression makes predictions for all input values), while a *pattern* refers only to the restricted regions of space spanned by the variables, as in outlier detection. Both models and patterns are valuable in science, but there may be differences between the "standard view" of data mining, as illustrated in the quotations above, and as arises in the mining of commercial data sources, and scientific data mining, and one goal of this workshop was to address that issue. One notable difference is the level of prior knowledge typically available in scientific problems: the astronomer seeking significant correlations between properties of galaxies observed in the optical and the radio passbands may be looking for hitherto unknown relationships, but has a fairly detailed concept of what a galaxy is; while a supermarket manager analyzing checkout data has little by way of an underlying model for the purchasing behaviour of shoppers. Does this difference limit the usefulness of standard data mining tools for science, and, conversely, how best should the domain knowledge of scientists be incorporated into data mining techniques?

1.3.3. Visualization

Visualization seeks to harness the remarkable capabilities of the human visual system to aid cognition, through the use of computer-generated representations. Visualization techniques find applications in a wide range of areas, from the elucidation of structure in complex datasets to the creation of virtual environments to aid training in dangerous or challenging procedures. In science, visualization can play an important role in exploratory data analysis, where visual representations can help the scientist to build up an understanding of the content of their datasets, while in the specific arena of e-science, much effort is being devoted to the development of collaborative visualization techniques, which aid the interaction between scientists at different locations. Some types of representation used in visualization are computationally expensive to generate, so that one challenge for the visualization community is to respond to the rapidly expanding data volumes seen in many application areas. Another obvious barrier to the ready use of visualization tools in science is the wide variety of data formats used in different scientific disciplines, which often necessitates the translation of data before it can be visualized.

1.4. The Workshop

The workshop organisers sought to bring together participants from the SDMIV communities, to take stock collectively of current practice and problems, to identify common interests between different disciplines and to check both that plans for infrastructure development match the requirements of the SDMIV researchers and, in turn, that these researchers are aware of the framework for the future work of application scientists that is being set by advances in Grid computing.

Forty-seven people attended the workshop, including representatives from all the communities originally sought: for the purposes of organising the first breakout session (see below), participants were grouped into five categories (infrastructure, data mining, visualization, astronomy and bioinformatics), but, in fact, there were representatives from a number of additional fie lds, including digital libraries, engineering, oceanography, marine sciences, statistics, chemistry, experimental physics, and environmental sciences. Most attendees came from academia or governmental research organisations, but there were also several from the industrial and commercial computing sectors. The preliminary programme presented with the workshop announcement provided a lot of time for discussion sessions, in addition to scheduled talks, and attendees were invited to put forward the topics they would most like to see discussed at the meeting, in addition to describing their own area of expertise and current work. The organisers then finalised the workshop programme in the light of the preferences stated by the participants in their applications.

1.5. Report structure

The structure of the remainder of this report is as follows. Section 2 presents a discussion of the main issues arising in the workshop (with some elaboration by the authors) and the recommendations resulting from it. Section 3 summarises all the prepared presentations made at the workshop, together with the reports made as a result of the first breakout session. A set of references is given in Section 4, and this is followed by the Appendices, which reproduce the workshop timetable and list the workshop participants.

2. Issues and Recommendations

In this section we discuss some of the issues arising in the workshop, particularly those from the breakout sessions, where the workshop participants were asked to identify important issues to be studied further over the next 12-18 months. The section concludes with a summary of the recommendations drawn from the discussion of these topics.

2.1. Standards for Data Formats

Data format conversion is the soft underbelly of processing scientific data, consuming immense amounts of time for those who work in a heterogeneous software environment. Sometimes it is easier to rewrite an entire component than it is to convert existing data to and from its input and output formats. There was considerable concern about this in the discussion groups of the meeting, and it was clear that standardization would greatly reduce barriers to software reuse. Combining data to extract new knowledge from the join will be more important in the future, and improved standardization will also enable this kind of federation.

Using standard formats for data clearly makes everything much easier – the output of one module can be the input for another. However, several factors conspire against this. Suppose a user finds an implementation of a published data mining algorithm, and wishes to use it on her own data: the data are already stored in one format, and the new code wants a new format. The format may be too simple or too complicated for the immediate use; if it is too simple, it cannot represent the sophistication of what the user wants, or will want in the future. By contrast, the new format may be quite complex, with different ways to represent many different possible ways of using the algorithm, perhaps because the maker of the data mining code is rewarded for extra sophistication more than for simplicity and robustness.

Standardisation of file formats is encouraged by two major factors: a wide range of high-quality software, and the adoption of the format by the powerful organizations. The latter should be considered seriously by funding bodies, especially those supporting research that crosses the traditional boundaries between disciplines, while the former objective, we feel, is easier to accomplish if file formats are based on XML[6] .

2.1.1. XML Formats

XML (Extensible Markup Language) is now the method of choice for expressing any kind of structured data, having spread throughout the business and academic communities over the last few years. While it is easy to create a simple XML document with a text editor, great complexity and sophistication is possible. XML is used for astronomical tables, for bank transactions, for vector graphics, for remote procedure calls, and many other applications. It is because of the wide variety of tools that are available for XML that it is easier to create a software library than when a format is created from scratch.

An XML document may be *validated* against a *schema*: the document can be thought of as an instance, and the schema as a class definition. The file format is designed by specifying the schema – using an off-the-shelf tool – and documentation created to explain the semantic meaning to a human. The same kind of tools can be used to create instances – files that fit the format – but the tool is intelligent, not allowing invalid data entry, and suggesting that which is valid. (For example, if the schema states that the name of a month is required next, the tool can list the twelve possibilities and not allow anything else).

Programs can read XML through any of a large number of parsers, some already built into programming languages (Java, Python, …), and a variety of parsers available in other languages (C, C++, Perl, …). The parser can provide a view of the document as either a stream of events (SAX[7]), or as a tree-like object (DOM[8]); the former is best for large documents that will not fit into memory, the latter for smaller documents, to enable efficient navigation, manipulation, and querying.

XML documents can be easily translated to other formats using an XSLT[9] translator; as with XML, there is a wide variety of these available. Translations are programmed with the XSL[10] language, which produces output when templates are matched in the input XML document. The latest generation of web browsers is able to make these translations automatically, so that XML documents can be viewed easily by a human, as well as being understandable by machines.

2.1.2. Binary Data

A perceived problem with XML – and one that crops up frequently in scientific circles – is that it is inefficient for large data volumes because of the overhead in storing and processing the language elements in addition to the encoded data. Of course, one can counter by saying that processors and disks are becoming cheaper, and counter that by noting that data size requirements are rising as fast or faster.

A highly efficient solution to this problem is to let the XML represent metadata – typically complex, but small data objects – and have the metadata contain a link to a large, binary file. In this way, the sophisticated XML tools can operate on the metadata, discovering exactly what is needed for processing the big data, then specialized, handmade, efficient code is given the file pointer that allows it to operate on the big data with maximal efficiency. One example of the implementation of this approach is the VOTable [11] format for representing tabular data in astronomy developed by the International Virtual Observatory Alliance[12] . An alternative is to provide an interpolation between these methods by allowing the binary file to be seen in the same way as if it were in pure XML: the user code gets SAX events containing the data, whether or not the document is pure XML or XML with a reference to a binary file. This second approach is being implemented in BinX[13] , a project under development by the Edinburgh Parallel Computing Centre[14] and the National e-Science Centre[15] , originally as part of the OGSA-DAI[16] project. Both of these (and there are probably others being developed by other e-science communities) are XML formats that allow references to remote binary data, and have a rapidly growing collection of support software as well as influential

political support. BinX emphasizes the ability to cover existing binary formats in an XML skin, while VOTable emphasizes complete and standard metadata in the representation of a table of data records. Which of these (or their equivalents in other disciplines) is more suitable in a particular situation depends on the details of the problem, although the BinX approach, by seeking to provide a method for describing any binary data format, does perhaps sit more naturally with the desire to avoid having to reformat data from one discipline before applying to it a data mining or visualization algorithm developed in another discipline.

2.2. Referencing Remote Data

Traditionally, data are brought to a local machine, processed there, and a data file is referenced by its file name on that machine. In the last few years, a new naming system has arisen, the URL, so that data on a distant machine can be referenced, and people now trust the availability enough to exchange URL links to a data object rather than the object itself. In the future, we will carry this further, so that, for example, data-mining services will be able to take a URL as input, or a database view can be defined by the location of the database and the query that made it.

However, we must be careful to build flexibility and robustness into these references. If a URL is not available, the system should look elsewhere for the data object – for example when a machine is connected to the net, it should use the remote version of the data object, but when disconnected, the data may be local. Large datasets may be replicated in several places, and the most convenient location should therefore be used. These technologies are widely used in industry (e.g.[17]), and the OpenURL initiative [18] is a strong standard emerging from the Digital Library community. Another issue of concern here is the provenance of data referenced by a URL: if the data referred to by a URL are modified one night, then different results may be obtained by running the same algorithm using the URL reference on the previous and following days. It is clear that either care must be taken in the design of the URLs (e.g. that they should encode timestamps or version numbers) or the data accessed via the URL should contain adequate metadata, if the basic scientific requirement of reproducibility is to appear to be satisfied.

2.3. Grid Services

2.3.1. Common Open Grid Services for Science

A significant contribution of the workshop was the identification of the requirement for a new, flexible paradigm for making knowledge extraction tools available to the scientists who need them, which we shall call Common Open Grid Services for Science (COGSS), based on OGSA (Open Grid Services Architectures): we refer the reader to [2] for a detailed description of what derivation from OGSA implies for COGSS.

To explain this new way, let us first recall the current situation with scientific software components: a subroutine or class library is developed and tested, then either sold or open-sourced, and clients download, port, compile, and run the component on their own machine. By contrast, the COGSS model provides more flexibility, so that components can be developed and matured on a single machine close to the authors, even though a worldwide collection of clients can be using and beta-testing the code. The component can remain on a server controlled by the authors, selling or giving CPU cycles for the clients, the software always at the latest version because of this control. Alternatively, it may be that users wish to use their own computing resources, perhaps to keep computing and data close together, in which case the software of the COGSS component can be shipped to the client. There is a natural progression from statically linked executables, to shared object libraries, to the applets and jar files of Java, to SOAP[19] web services, to high-performance, dynamically bound OGSA services.

In a third COGSS model, the data input of the component is not shipped directly to the service, but rather a reference to it is sent; this may point to a third location, from which data are imported when the reference is actuated. What is sent to the service is a metadata/control message describing the data and what is to be done to them by whom, plus the reference to the data themselves. A similar story applies to data output: the result of a COGSS service may cause data to be stored – perhaps

temporarily – and a reference returned to the client, along with diagnostics, summary, or other information. Such an output reference – as with the input reference – may have an associated expiry date, so that temporary files can be safely deleted when no longer needed. It may also be that the data reference is not actuated by a component, but only the metadata and control used by the service, for example a service that estimates the time it would take to run another service may only need to know the size of the dataset, not the contents.

Examples of COGSS services that might be derived from algorithms discussed in the workshop might be the *kd*-tree analysis of large multi-dimensional point sets (R. Nichol, see Section 3.2.1), the Hough transform analysis of astronomical images (A. Storkey, see Section 3.2.5), or some gene sequence matching routines (D. Gilbert, see Section 3.2.3). A strong recommendation from the workshop was that there should be a library of standard data mining and visualization tools, which could be used in the construction of more specialised software modules in a wide range of application areas. In the COGSS model, this would be implemented as a library of well documented Grid services. At a wider level, we recommend conversion of the NAG Data Mining Components [20] and similar commercial toolkits into OGSA Grid Services.

2.3.2. Workflow in the Grid Services model

As discussed in the previous subsection, the component parts of a Grid computing application need not be executing within a single machine, but may be geographically distributed, each running in the most suitable place. We can choose to execute close to the data, close to the software developer or on some specialised or economically-chosen computational resource. We think of a human interacting not with a single machine, but with a diverse collection of machines; this leads to requirements for workflow systems for creating compound services and a grid service accounting system which enables users to employ commercial services or have their own services run on commercial hardware.

Once a collection of service definitions has been collected by a user, there should be a way to connect them into a coherent network, so that different implementations can be tried and services debugged in the context of the others: like a *conductor for a symphony of grid services*. The conductor would be close to the human user, perhaps a GUI-based system that is also showing visualization. But it should be emphasized that the services that represent the orchestra would be running on distant machines. One paradigm for such a workflow system would be building a graph of services. Each service takes input from others, and the output of each may go to others. Connections between services would be double-stranded, with both control and data channels. The control channel carries metadata, control, and diagnostics, probably in an XML/SOAP dialect, and may be bi-directional; the latter may carry big data in parallel, asynchronous streams, and is designed for maximum throughput. Control of service execution need not be closely connected to data flow, as in conventional dataflow and visual computing software; we might have a control channel sending a reference to a data object that is passed from service to service, eventually being actuated far from its creation. Elements of this concept are available in systems working today, such as the Chimera Virtual Data System [21] developed by the GriPhyN[22] project.

2.3.3. Accounting in the Grid Services model

As discussed by Rob Baxter in his *Brief History of the Grid* (see Section 3.3.1), the need for a workable accounting system remains the "elephant in the room" in Grid circles – everybody knows it is there, and that it is big, but they still avoid talking about it. If something like the COGSS system develops, and includes the use of commercial services or the running of any services on hardware provided by commercial concerns, then an accounting system is definitely needed. Arguably, such a system might be required even for use within academic science – for example, for regulating the use of national supercomputing or data storage resources – although in this case, it would be possible to use some system of credits, rather than real money. Similarly, new, and more flexible, licensing strategies will have to be developed by software providers: rather than issuing unlimited-use annual licences to university departments, they may have to charge users each time one of their services is used, perhaps as part of some compound workflow, which includes the dynamic discovery of services, so the user may not even know that a commercial service is being used. Such issues may seem well

removed from the usual concerns of the average academic researcher, but their solution will shape the computational environment within which e-science takes place, so they cannot be ignored completely.

2.4. Publishing and Registries

2.4.1. Derived Data and Annotation

The traditional model of scientific data is that a small number of large organizations publish well-crafted, carefully documented datasets, and those in some inner circle know how to access and derive knowledge from them. This is analogous to the era of anonymous FTP in the eighties and early nineties; the era which preceded the web, where thousands of individuals and organizations publish their 'home pages'. We believe that the publishing of derived datasets will flourish in the coming era of e-science and cheap disk, allowing the full harvest of knowledge to be extracted from these terabytes. Scientific knowledge is extended by using the well-documented results of others, and processing trusted data to make new trusted data is an extension of this.

Biologists speak of 'annotation'; for example matching genomic structure with function, identifying mistakes in a sequence, or cross-matching the genome of one species against that of another. Astronomers build 'derived datasets', for example identifying outlier or mistaken sources in a catalogue of stars, or cross-matching one survey catalogue against another. Once such a dataset has been made, we would like it to be published so that others can use it: to save work in reproducing the result, to mine further knowledge in as yet unknown ways, to federate with further datasets, or many other possibilities.

In the workshop, we identified a serious stumbling block to this publication of derived datasets, namely that there is no motivation for a researcher to publish with the sort of care and attention to detail that is necessary. Funding agencies that force the publication of data and metadata often end up with "something that looks like the real thing but isn't", in the words of one workshop participant.

What is needed is to merge the idea of publishing a dataset with the idea of publishing a scientific paper. In the latter case, investigators are very careful, knowing that a good referee report is necessary for publication in a reputable journal, which in turn is necessary for readers, citations, and a successful career. If the publication of a high-quality dataset were accorded the same status as a set of observations or a paper with scientific conclusions, then authors would be motivated to do an excellent job.

One way to create this new paradigm would be for the UK e-Science programme (or some other authority) to publish datasets through an electronic journal, complete with peer-review and editorial board, as well as a list of reliable URLs showing multiple places where the relevant data services are implemented. A secondary aim would be the creation of a 'dataset citation index': when dataset B is created from dataset A, the index would allow B to be found from A.

2.4.2. Registries

The scientific community is becoming much more dispersed: with the Internet and cheap travel it is easy to collaborate with colleagues all over the world. The Grid paradigm encourages this, by dispersing computing and data resources as well as the humans. Because of these effects, it is more and more of a challenge to find resources: to find out if another group is doing a similar project, to find if an algorithm has been implemented, to find where a service class has been implemented, to find what computing resources are available, etc. An emergent theme of the workshop was that of *registries of resources*.

Many application scientists do not seem aware of the rich collections of software that are available for data mining and visualization. We therefore recommend the establishment of registries of data mining and visualization tools, preferably along the sophisticated lines outlined below. The collection of peer-reviewed datasets recommended above could also benefit from a comprehensive registry.

We all use the big search engines such as Google and Yahoo. The former uses keyword searches and is highly effective on resources where the information is in natural language, but is not very effective

where the information is in a database, or if we are searching on, say, numerical attributes. Yahoo is a different model, where humans methodically catalogue items in a hierarchy. Another valuable resource in this regard is when an expert individual has built a web page of resources, for example *Joe's Page of Data Mining Software*. While these kinds of pages often have expertise and specificity, they also tend to be poorly maintained and inflexible in the face of change.

A new type of registry is emerging, based on standards such as Z39.50 (see section 3.4.3, also[23] and [24]), or in a different context, on UDDI (Universal Description, Discovery, and Integration of Web Services). The emphasis is on queries rather than a simple list, and queries are not just keywords, but additionally subject and author searches, as well as numerical attributes. These new types of registries should be designed with the idea of the maintainer as a moderator rather than content provider. The registry is thus a mediator between publisher and client, with the moderator watching. The key to these types of registry is a well-defined metadata schema: for example each book in a library must have a single title and one or more authors, or each Grid service must have a Service Definition Document in a specific language.

We recommend encouraging research and implementation of these kinds of resource registry. One such registry is already running at Daresbury[26] , listing UK e-Science projects. Others can be made for lists of subject-specific datasets or Grid services, or other types of resources, including registries of registries.

2.5. Collaboration and Visualization Technologies

As collaborations disperse, connected only through annual face-to-face meetings and daily emails, we must retain the ability to share results and thereby work together. When it is a formal result, such as a collaborative paper, we have word-processing software with change-tracking, but for informal exploratory work, there are few established mechanisms beyond telephone conferences. However the workshop participants expressed a desire to be able to collaborate in a richer way. Video conferencing allows a speaker to see how many people are listening, and makes the group more coherent.

Personal video-conferencing can be achieved on a workstation using the free VRVS[27] software from CERN/Caltech, or a proprietary alternative such as Microsoft NetMeeting. These systems typically allow sharing of applications within the conferencing system, and so desktops can be shared amongst the collaborating group. For example, VRVS works in conjunction with VNC desktop sharing[29] . Collaborative white board discussions can also be held. At a larger scale, the Access Grid[30] is emerging as a standard way to convert a conference room for distributed large meetings; although currently Access Grid needs a skilled driver for best results, and is therefore not used without a lot of planning. We recommend trying to bridge these technologies, by making an Access Grid usable as simply a higher-resolution version of desktop conferencing. A further recommendation is to investigate how best to integrate visualization and other scientific software into an AccessGrid session, so that scientists can have active *collaborative working* sessions where new studies are undertaken, rather than simply sessions where past individual studies are reported.

2.5.1. Collaborative Visualization

At a higher level of multimedia, we recommend further research in ways to collaboratively visualize complex datasets, so that several dispersed collaborators can see results and discuss significance. Such systems often assume excellent bandwidth, and research is needed in how to be robust against bandwidth reduction. Visualization systems have traditionally been local: dataset, computation, and human all in the same room. But the Grid paradigm changes this; when the computation is distant from the user, we must decide – perhaps dynamically – what processing is done at the remote server, and what is done locally. A visualization of a 3D volume dataset may be done at the server and an image passed to the user; at the next level, the image may have script or links to enhance its flexibility; a collection of triangles representing an isosurface may be sent for rendering on the user's local machine; or the whole 3D voxel block may be sent. As more bandwidth becomes available, more flexibility becomes available. When a number of scientists are collaborating, the distribution problem is accentuated.

2.5.2. Quantitative Measurement, Provenance

Many visualization tools are excellent for giving a qualitative view of a dataset, allowing a researcher to see something interesting and thereby form a hypothesis. Generally, the researcher must then build a program to test and measure that hypothesis in a quantitative manner before any strong conclusion can be made. We suggest that visualization tools would be more useful to scientists if there were a greater emphasis on quantitative measurement tools rather than more sophisticated lighting and texture mapping. Such tools should include point measurements, as well as histograms, line graphs, scatter plots, and other simple plotting devices, with axes properly and automatically labelled.

Visualization tools should also be able to retain the provenance information of a dataset: who made it, where was it published, how the data was processed – the information that allows a coloured picture to be viable as scientific evidence.

2.5.3. Scalable Visualization

A great deal of effort has been expended to make visualisation hardware fast, with large numbers quoted for polygons per second and so on. This allows sophisticated software to provide magnificent and subtle images of scientific data sets. Unfortunately many of these systems require complex processing environments, a special room for viewing, and headsets or stereo glasses. Because of this, the system is often underutilized. There was a feeling at the workshop that such systems would be more useful if they were part of a continuum from the portable, perhaps slow desktop environment upwards, so that learning at the desktop still works on the high-end equipment.

2.6. Tutorials and Documentation

Developments in computational infrastructure, such as the Grid, are driven by the IT community, and application scientists only get involved once it is clear that they can gain some significant advantage by doing so. To aid the uptake of new techniques in e-science it would be very useful for there to be available tutorials and sample code describing and illustrating the use of the these techniques at a level amenable to application scientists, rather than IT specialists. Specific topics cited in this workshop include:

- X.509 Digital Certificates[31] : what are they and what can I do with them, how do I get one, how do I recognize one, how and why is this identity document different from a login/password or a passport?
- What is a SOAP web service, and where are examples that will benefit my work? Who can help me convert my utility software to this form?
- What is OGSA and how is it more than SOAP web services? How does all this IT infrastructure get my work done more effectively?
- Where can I find software to let me start on these ideas? – software that is simple enough to be transparent, yet illustrates the basic ideas. How can I publish and subscribe to a "Hello World" OGSA service?

2.7. Summary of recommendations

- **R1: The use of XML for scientific data is recommended, as it aids interoperability and flexibility.**
  Standardising data formats for scientific data is an old challenge, but the participants agreed that XML technologies provide an excellent start for creating a new format or rebuilding an existing format. Data represented with XML combines rich structure, rich choice of tools, and wide acceptance.
- **R2: Research should be undertaken into ways to reference remote data objects that are more flexible and robust than URLs and FTP addresses.**
  A full reference can be a collection of possibilities and replicated instances, rather than a single fallible URL or FTP that causes drop-dead failure if unavailable.

- **R3*:* A new framework of interoperability is emerging from the Grid community, and scientists should build their software to benefit from these standards.**
  The business world is moving to SOAP-based web services, in the expectation of profit from a richer, more integrated Internet. In academic science, the promise of equivalent technologies should be leveraged and exploited.
- **R4: Libraries of interoperable scientific data mining and visualization services should be built to this Grid standard.**
  Many data mining and visualization algorithms are likely to be of use in a number of e-science application areas, so that it would be efficient to develop a public library of such routines. In the Grid Services paradigm (see 2.3), this library would be implemented as a set of services, and, conceptually, would sit between the basic infrastructure delivered by OGSA and the application layer, where individual scientists or projects would develop services to meet their specific needs.
- **R5: A mechanism should be sought whereby the peer-reviewed publication of datasets can be made part of the standard scientific process.**
  This is already the case in some communities, but not all. For example, some journals require the submission of datasets along with papers, so that others can reproduce the results published in the papers, while others collaborate with data centres in making available datasets (e.g. tables) published in papers. This might profitably be extended to include further disciplines, as well as generalised, so that a valuable stream of citations can be obtained by the scientist who publishes a dataset derived from some set of original sources through the application of an analysis procedure.
- **R6: A registry of fiducial datasets should be created and maintained, to be used in the testing of data mining and visualization tools.**
  These datasets should be representative of the sorts of data types, formats and volumes that application scientists wish to analyse using data mining or visualization tools, and it should be easy for scientists to add further datasets to the registry.
- **R7: Better methods should be sought for collaborative working in e-science**
  A key feature of e-science is its collaborative nature and the interdisciplinary interactions that it fosters could lead to very beneficial collaborations, but for them to bear fruit there need to be better ways for scientists in distributed teams to work together. One particular aspect of this is the development of collaborative visualization tools.
- **R8: A set of tutorials should be created and maintained, for introducing application scientists to new key concepts in e-science.**
  These should be written at a level assuming no specialist IT knowledge, and should be designed with the aim of increasing the uptake of new techniques (web services, digital certification, etc) across all scientific disciplines.
- **R9: A report should be produced on the data mining requirements of e-science application areas.**
  This report, detailing the data mining requirements of as wide a range of e-science application areas as possible, would be used in setting the scope of the standard libraries of R4, by helping to identify just what are the tools that would have wide applicability.
- **R10: A report should be produced on the visualization requirements of e-science application areas.**
  Similarly, this report, detailing visualization requirements, would be used in setting the scope of the standard libraries.
- **R11: A registry of existing data mining resources should be created and maintained.**
  Initially a WWW page, but eventually a machine-interpretable registry, this will aid application scientists seeking data mining algorithms for their work.
- **R12: A registry of existing visualization resources should be created and maintained.**
  Similarly, this will aid scientists wishing to use visualization in their work.

3. Summary of Workshop Presentations

In what follows we present only a brief summary of each of the presentations made at the workshop. Copies of the slides from these are available from the SDMIV workshop WWW site, at the following URL: `http://umbriel.dcs.gla.ac.uk/NeSC/action/esi/contribution.cfm?Title=114`.

3.1. Overview Talks

The workshop opened with a set of Overview Talks presenting an introduction to the issues to be addressed during the meeting.

3.1.1. Scientific Motivation Overview – Roy Williams

Knowledge extraction from scientific data can often be viewed as the act of concentration that takes an experimental/observational *Datacube* and derives from it an *Event Set*, which is often represented as a table of attributes characterising the event and which contains the scientific knowledge distilled from the original data. Further knowledge may be extracted by the integration of event sets – e.g. cross-matching observations of astronomical sources made in different passbands and stored in databases distributed around the world – and many practical problems arise in doing this, some of which the Grid may solve, if/when databases are properly included in the Grid (e.g. by the OGSA-DAI project [16]). Scientists in many disciplines will require a grid of services to make use of the data available to them and the provision of many of them poses computational challenges, given the size of some of the datasets: how can one visualize, classify, and find outliers in distributions of 10^{10} points?; how can one perform joins on such large tables, especially when they are contained in distributed database? Standards are needed, to allow referencing of data and resources in the Grid, and to turn it into a problem-solving environment, with both plumbing (bulk data transport) and electrical (control and metadata) subsystems. Web services and workflow provide the means of doing that, but, ultimately, semantic information will be required for the location of classes and implementations of services.

3.1.2. Data Mining Overview – Chris Williams

The essence of data mining is the finding of structure in data. A large number of different tasks fall under the heading of data mining – such as exploratory data analysis, both descriptive and predictive modelling, as well as the discovery of association rules and outliers – and many practical problems arise from their application to complex types of data. Predictive modelling centres on the learning from existing input/output pairs, so that the output(s) can be predicted given further input sets, and this can comprise use of a number of techniques, such as neural networks, decision trees, nearest neighbour methods and Support Vector Machines. All such supervised learning is inherently inductive in nature, and its key issue is generalisation – how to make predictions for new inputs based on previous knowledge. Descriptive modelling seeks to find significant patterns in data with no external guidance, and this is simply done using techniques such as clustering and reducing the dimensionality of the dataset by fitting it to a lower dimensional manifold. All data mining is greatly aided by a suitable computational environment, in which operations can be pipelined, data visualized and results evaluated. Several of these exist in the data mining community, and it is important to assess how applicable these are to scientific data mining problems, where a great deal of prior domain knowledge may be available and can be factored into the data mining procedure. Probabilistic modelling may provide the correct framework for modelling complex networks of non-deterministic relationships, and probabilistic expert systems have been created in many areas (e.g. medicine).

3.1.3. Visualization Overview – Ken Brodlie

Visualization – "*Use of computer-supported, interactive, visual representations of data to amplify cognition*" (Card, Mackinlay, Shneiderman[31]) – has its origins, as a discipline, in an influential 1987 NSF report, 'Visualization in Scientific Computing' by McCormack, de Fanti and Brown[33] , and is now widely used in computational science and engineering. It is useful to consider it as a pair of topics: *scientific visualization* (the visualization of physical data) – e.g. plotting the properties of the Earth's ozone layer on a 3D representation of the globe; and *information visualization* (the

visualization of abstract data – e.g. presenting a network of links on an automobile web site). Scientific visualization is largely concerned with visualizing datacubes, where there is a clean separation between dependent and independent variables. The independent variables give the dimension of the space being visualized, often 1D, 2D or 3D but occasionally greater. Much scientific visualization follows a standard procedure – read in the data, construct a model of the underlying entity, construct a visualization in terms of geometry, render the visualization as an image – and there exist a number of modular software systems for implementing that procedure. This dataflow model has proven very successful, but the impetus now is to extend it, for example, to allow collaborative visualization and computational steering within the dataflow. The aim of information visualization is to display and reveal relationships within multivariate datasets, which are usually represented as tables; its focus is the dataset itself, not the physical entity underlying it, reflecting the fact that the relationships between the variables are usually not well understood. Techniques commonly used in information visualization include parallel coordinates, scatter plot matrices, and pixel-based and glyph techniques. In many practical situations, screen space becomes the limiting factor in the visualization of large, complex datasets. In such cases, it is necessary to reduce the scale of the problem, either by restricting attention to subsets of variables, or employing some sort of descriptive modelling technique (e.g. clustering), which makes use of the structure of the data themselves to reduce the dimensionality of the problem.

3.1.4. An example of data mining and visualization in practice – Jeremy Walton

The NAG [30] Data Mining Components and IRIS Explorer[35] visualization tool comprise one of the available commercial frameworks for mining and visualizing scientific data. As an example of using such a system in practice, consider the following analysis of some image data from the Landsat[36] Multi-Spectral Scanner. The dataset comprises images of different regions, with 36 independent variables per region – a 3x3 array of pixels, and four spectral bands per pixel. Each pixel is to be assigned to one of six classes of land use, and extrapolation is to be made on the basis of these results, so that land use classification can be performed using the multi-band pixel values. As a first step, principal component analysis is used to reduce 36 dimensions to two that explain 85% of the variance in the data. Three land use classes are chosen – cotton crop, damp grey soil and soil with vegetation stubble – and a decision tree is used to model the assignment of the pixels to these land use classes. On the basis of this model, boundaries can be defined in the data space for the three land use classes, and the classification of the further data points is made. The distribution of the original and predicted class values can then be visualized, to help in the assessment of this classification procedure.

3.2. Application Talks

After the Overview Talks came a series of Application Talks, in which researchers from different parts of the SDMIV community described their own work and identified some of the challenges that they had faced in undertaking it.

3.2.1. Computational astrostatistics – Bob Nichol

The Pittsburgh Computational Astrostatistics (PiCA[37]) Group brings together statisticians, computer scientists and astronomers from Carnegie Mellon University and the University of Pittsburgh, to develop new statistical tools for the analysis of large astronomical datasets, notably the Sloan Digital Sky Survey (SDSS[38]). This collaboration works well because it is of benefit to all concerned: the astronomers want to exploit the large, rich SDSS dataset to the full scientifically, and that requires expertise in algorithms for knowledge extraction, while that same size and richness challenges and stimulates the computer scientists and statisticians who work on such algorithms. The key to the success of the PiCA group's algorithms is the use of multi-resolution k-d trees for data storage. As well as partitioning the data effectively, these trees also store basic statistical information about the objects stored in each node. This representation of the data is usually sufficiently condensed that many operations – such as the calculation of the *N*-pt spatial correlation functions, which characterise the clustering of the galaxies in the SDSS – can be performed in memory, with the tree structure avoiding many needless computations made by traditional techniques. Another application

of the k-d trees is in a fast mixture model code, which has been used to select rare objects from large, multivariate catalogues of SDSS object attributes.

3.2.2. Distributed Aircraft Maintenance Environment (DAME) – Tom Jackson

The work of the DAME[39] project is based on the AURA[40] data mining system. AURA is a set of tools to build fast pattern recognition systems, using neural network based associative storage. It is aimed at unstructured data, and is designed to be scalable and readily applicable to large data volumes. The AURA storage system uses Correlation Matrix Memory (CMM), and exploits threshold logic methods and the distributed encoding of information. Data typically have to undergo pre-processing to make them amenable to AURA analysis. Efficient implementation in software and hardware is made possible by use of a set of binary "weights", and this system is ideal for use on bit vector machines, although the AURA C++ library has been run on a variety of systems, from individual PCs or workstations, through Beowulf clusters, to an Origin 2000 supercomputer, as well as bespoke hardware. The DAME project is designed to demonstrate the diagnostic capability of the Grid, using Rolls Royce aeroengines as its application. The key requirement is that the system must analyse and report any "novel" engine behaviour and identify its cause very quickly, despite dealing with many TB of data, and a very distributed network, involving aircraft, airline offices, data centres and maintenance depots in different countries. As part of the DAME project, a Globus[41] -enabled version of AURA (called AURA-G) has been developed, and this will be available in late 2002 for use in other projects. AURA-G is designed for scalable pattern matching, using multiple CMMs at different sites, and is OGSA-compliant. The DAME project highlights some of the issues facing data miners in the Grid environment, such as provenance and standards to maintain data transparency independent of location and to manage the database/data mining link in a distributed environment.

3.2.3. Bioinformatics – David Gilbert

Bioinformatics is the application of molecular biology, computer science, artificial intelligence, statistics and mathematics to model, organise, understand and discover interesting knowledge associated with large-scale molecular biology databases. This combination of expertise is required not only because of the rapid increase in the volume of molecular biology data, but also because of how it used; the life sciences are characterized by coordinated study at many different levels of granularity – from a single nucleotide sequence, through protein structure to a cell, to an organ, all the way up to the physiology of a whole organism. Classification is a major part of biology, so classification techniques feature strongly in bioinformatics, often using similarities of structure (found through pattern-matching – e.g. in gene sequences) to infer similarity of function. A variety of such techniques are used, both deterministic and probabilistic. It is becoming common to combine multiple, complementary techniques in analyses, with the goal of increasing the power for discovering useful knowledge, but the lack of coherence in these sets of methods can make their results difficult to combine in a meaningful way. As in most other applications of data mining techniques, a great deal of the effort in bioinformatics is devoted to the preparation of data, before the particular machine learning algorithm, or whatever, can be applied to it. This includes the identification of training and test sets, as well as more basic operations like dealing with missing data and transforming to the data format required by the algorithm. More fundamental problems arise from the "dirty" nature of biological databases, which tend to contain experimental errors, erroneous annotations and interpretations, and data biased by selection and taken using non-standard experimental procedures.

3.2.4. Potential SDMIV applications in CLRC/RAL collaborations – Julian Gallop

CLRC[42] holds, or provides access to, significant data resources across a wide range of fields, notably space, earth observation, particle physics, microstructures, and synchrotron radiation, and is a partner in a number of e-science projects concerned with different aspects of the SDMIV problem. Within many of these, such as the Data Portal[43] and the NERC Data Grid[44] , the development of metadata models has proven to be very important for the discovery of data sources, while another – climateprediction.net[45] – is an example of a "cycle-scavenging" application, as it uses spare CPUs on home PCs to run climate prediction models, and therefore encounters made Grid security issues.

3.2.5. Scientific data mining: applications to astronomical data – Amos Storkey

Problems in astronomy increasingly require use of machine learning, data mining and informatics techniques, giving rising to a new field of astroinformatics. Examples of astroinformatics topics are: detection of "junk" objects in sky survey datasets, record linkage between astronomical databases, object classification and clustering, and data compression to aid analysis and storage. One example is the detection of spurious objects in the SuperCOSMOS Sky Survey (SSS[46]): this is based upon a collection of several thousand photographic plates taken by the UK Schmidt Telescope[47] , in Australia, which have been digitized by the SuperCOSMOS plate-scanning machine[48] in Edinburgh. This process yields a pixel image for each plate, over which an analysis algorithm is run, to detect and characterize astronomical sources in the image. The object catalogues so produced contain a variety of classes of "junk" caused by scratches in the photographic emulsion, fibres that remain on the plate despite clean-room conditions, as well as "real" emission from undesirable sources, such as aeroplanes and artificial satellites, and artefacts like diffraction patterns around bright stars, caused by the optics of the UK Schmidt. An analysis of the problem indicated that these spurious objects could only be identified on the basis of their displaying statistically unlikely linear or circular configurations, and a number of machine learning techniques – such as the Hough and circular Hough transform, and a new approach, based on Hidden-Markov renewal processes – have been employed, with varying success, to detect and label them in the SSS database. This project exemplifies that, while machine learning and data mining techniques are, and will continue to be, very useful in astronomy, they do not always work automatically, and success may require significant interaction between the domain specialist and the informatics researcher.

3.3. Computational Infrastructure Talks

The final pair of formal presentations covered the computational infrastructure which provides the framework within which scientific data mining, integration and visualization takes place. Rob Baxter presented a chronological account of the development of Grid computing from its earliest precursors to the current day, while Malcolm Atkinson described the Grid and e-science concepts in more detail, and discussed where the SDMIV communities should fit into that picture.

3.3.1. A complete history of the Grid (abridged) – Rob Baxter

The Grid vision centres on the realisation that the world contains vast numbers of (generally under utilised) computational resources connected by high-performance networks (albeit using a variety of access modes and protocols), coupled with the desire to use them effectively, whether for science, commerce or anything else. Grid software today comprises toolkits, schedulers, data management systems, portals and web services, all of which are starting to work together harmoniously. The development of the Grid may be traced back to the world's first packet-switched network, at the NPL, in 1967, and, although, the course of that development has not run straight over the past thirty-five years, it is possible, with hindsight, to identify some of the crucial milestones along the way. The first twenty years of that history saw a general development in the power of computers and network systems, but nothing really "Griddy" appeared until the Condor[49] project started, in 1988; that Condor remains a key part of the Grid concept now attests to its great importance in this history. Condor is a job management system designed to make effective use of spare CPU cycles on under-used PCs. It matches job requirements with computer capabilities, and is robust enough to handle jobs that do not complete: with the later addition of elements of the Globus Toolkit[50] (to produce Condor-G), it became possible to seek out available and suitable machines for a given job, rather than simply replying on knowledge of the machines present on a LAN. The next key date in the history of the Grid was 1993, when the Legion[51] project was launched. Legion took an object model approach, in which the Grid is viewed as a single virtual machine: this is attractive conceptually, but it proved difficult to implement efficiently. The following year, 1994, saw the arrival of Nimrod[52] , a system designed to automate the management of task farms, which was notable for having a concept of a charge per CPU cycle, so the master process can schedule jobs on the basis of cost as well as time. In 1997 a unified environment for German HPC users was delivered by UNICORE[53] , which is a combination of a toolkit and a portal: it provides middleware, so that users can create and run

their own jobs, not just precompiled applications, but it hides all of the details of doing so behind a single GUI. Later in the same year the Storage Resource Broker (SRB[54]) was launched, offering a means of accessing heterogeneous storage resources, including replicated datasets. The Globus project, producers of the Grid's leading toolkit, started life in 1998, and it remains the basis for the majority of Grid systems now, as well as one of the driving forces behind the OGSA. In 2001, the specification for the Web Services Description Language (WSDL[55]) was submitted, starting the remarkably rapid rise of web services in commercial computing, leading to the convergence with Globus development in the launch of OGSA in 2002.

3.3.2. Future infrastructure for SDMIV – Malcolm Atkinson

E-Science is fundamentally about collaboration. The sharing – of ideas, of resources, of data, etc – that it envisages requires trust, but can also change the way that science is done, if sufficiently developed and resourced. The Grid should make e-science much easier, by providing a common, supported high level of software and organisational infrastructure, and the initiatives centred on the Global Grid Forum[56] and OGSA development are "the only game in town" so far as delivering that infrastructure is concerned. For the SDMIV community, a key part of OGSA is the OGSA-DAI (Data Access and Integration) project, which is led by the UK and which is developing the services (specifications and reference implementations) required to integrate data resources (both relational databases and XML repositories) into OGSA. The key components of OGSA-DAI are: a Grid Data Service (GDS), which provides access to data and database operations; a Grid Data Service Factory (GDSF), which makes GDSs and GSSFs; a Grid Data Service Registry (GDSR), which effects the discovery of GDSs and GDSFs; a Grid Data Translation Service, which translates or translates data; and a Grid Data Transport Deport (GDTD), which provides data transport with persistence. OGSA-DAI will implement a role-based authorization scheme. The first OGSA-DAI product release is scheduled for December 2002, to coincide with the release of Globus Toolkit 3. The SDMIV community should think how it fits into the OGSA picture. The standard picture sees scientific applications built upon the OGSA infrastructure, but maybe it is better to envisage a layer between these two – an SDMIV (Grid) Application Component Library, which interfaces most directly with the DAI portion of OGSA, and which delivers standard SDMIV services which can form the basis of scientific applications developed by domain experts for their particular discipline.

3.4. Breakout Session 1

In their application material, many of the workshop participants indicated (unsurprisingly) that what they would particularly like to see discussed at the workshop was the overlap between their own area of expertise and one of the other SDMIV disciplines. So it was decided that the first Breakout Session should address the challenges arising in the areas of overlap between pairs of SDMIV communities, and to identify what requirements they placed on computational infrastructure. For this purpose, each workshop attendee was assigned to one of five categories (infrastructure, data mining, visualization, astronomy and bioinformatics) marking the largest groups of people present, and this list was used to select discussion groups on the following topics: data mining and astronomy; data mining and bioinformatics; data mining and visualization; visualization and astronomy; visualization and bioinformatics. The members of the infrastructure category were distributed evenly throughout the groups, and, inevitably, those who did not fit naturally into one of the five categories were assigned to a group not directly relevant to their own work, but they were exhorted to bring up any particular requirements of that work during the session. Each group selected a "scribe" to record its discussions, and a chair to summarise them in a session on the second day.

3.4.1. Data mining and astronomy

What causes problems currently is the time taken to convert between data formats, recode algorithms to match local environments and/or user expertise, and master local computational environments (ensure the presence of the correct versions of compilers, libraries, etc). Data volume is not always an issue – other problems can make the mining of smallish datasets complicated. There are also issues concerned with the management of change – both of the algorithms themselves and the computational environment – and with the control that the user has over them. Data mining tools are available, but it

is often difficult to discover what tools are available and suitable for a given analysis, as well as which of them are available in the desirable language and/or which can be modified to accept the desired data format. In practice, mutually-beneficial collaboration between the creator and user of the algorithm is more likely to lead to success than just having the user take an algorithm off the web and apply it with no guidance. Also there is a missing middle ground, for effecting that collaboration and translating standard algorithms for use within a particular domain. There is a problem that such activity might not be rewarded in either community: it won't look novel to the informatics researcher, if it is just an application of an existing technique, while the domain scientist is less likely to get credit for modifying a data mining tools than for use in his/her field than for spending the same amount of time doing research within that field.

3.4.2. Data mining and bioinformatics

Data mining in bioinformatics is hampered by many facets of biological databases, including their size, their number, their diversity and the lack of a standard ontology to aid the querying of them, as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels and domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanisms appropriate to all. The integration of biological databases is also lacking, so it can be very difficult to query more than one database at once. Finally, the possible financial value of, and the ethical considerations connected with, some biological data means that the data mining of biological databases is not always as easy to perform as is the case in some other areas. One development within the bioinformatics knowledge extraction field that may be of wider utility is the Distributed Annotation Service (e.g. [39]), which allows for a distributed set of annotations to a database, without the modification of the original database itself.

3.4.3. Data mining and visualization

The overlap between data mining and visualization is very large – indeed, it can be difficult to say what is data mining and what is visualization. One possible distinction is that data mining is machine-oriented, while visualization is human-oriented, however this distinction becomes blurred in exploratory data analysis, which may well be interactive and use visualization, but which may be reliant on statistical techniques from data mining. Given this blurred boundary, it is surprising that, in general, the standards do not exist that enable the ready combination of data mining and visualization tools into pipelines. One concern for the near future is the place of Z39.50 in the era of OGSA-DAI. It is not obvious what OGSA-DAI provides that Z39.50 does not, so far as the querying and retrieval of data from remote, heterogeneous databases is concerned, and it is to be hoped that the practical experience built up within the Z39.50 community (e.g. in the digital library world) is not being forgotten during the design and implementation of OGSA-DAI. The Open Archives Forum [58] has an alternative approach, centred on the harvesting of metadata, and, again, interaction between that group and the OGSA-DAI developers might be beneficial. The major challenge for the future in the data mining/visualization overlap region is the development of standards that will enable users to select and deploy appropriate tools. Standards are required for the description of the inputs and outputs from tools, as well as what they do to get from one to the other, and, ultimately, these descriptions should be machine-readable, so that they can be used in the Grid resource discovery process. There is scope for further interaction between data miners and visualization researchers to use data mining techniques to speed up visualization algorithms: this is already happening to some extent (e.g. using clustering techniques to circumvent the need to visualize very large numbers of data points), but more could be done.

3.4.4. Visualization and astronomy

Astronomy data are largely static, which simplifies their visualization, but they tend to be manipulated in large volumes, which introduces practical problems. In particular, it is not clear where to strike the balance between the client and server when visualizing large amounts of remote data: it is clear that low latency is a very important requirements of a successful visualization tool. Astronomers want to know more about what visualization tools are available for use, particularly those which are modifiable to meet particular requirements and those which can be integrated with statistical analysis

tools for data exploration. One concern is whether visualization tools retain enough of the metadata describing the original data to allow quantitative manipulation of visualized data – i.e. it is not enough to present an image which aids understanding of a dataset, the user wants to be able to make measurements directly on that rendered image.

3.4.5. Visualization and bioinformatics

Visualization is used in many areas within bioinformatics, with varying success: for some topics (e.g. the mapping of a single chromosome) good tools already exist, while for others (e.g. visualization to aid the comparison of genes from different species) they do not, often both because of the screen space problems common to many visualization problems and because sufficient thought has not gone into how best to visualize the data to aid comprehension. In many area of bioinformatics (e.g. the viewing of phylogenetic trees) it is important to be able to view information at several levels of detail and shift between them readily, which can be challenging for software. Some bioinfomatics visualizations are also computationally challenging: for example, in polymer docking studies for drug design, one wants to be able to manipulate fairly complex molecules, but this involves the recalculation of intermolecular forces every time a movement is made. A general complaint is that visualization tools are not sufficiently interactive to allow effective exploration of data: often one has to wait for a server to generate a new GIF image in response to a CGI request with new parameter values, rather than being able to change the rendered image more interactively. The visualization of biological data is also hampered by the wide range of data types and exponentially increasing volume of data available, and by the lack of interoperability of existing tools. To overcome this problem clearly requires the development of policies on data sharing and standards (best enforced by funding agencies?), and it is hoped that XML might form the technical basis of solutions to problems of data format and the description and interoperability of tools.

References

[1] The e-Science Institute, Edinburgh:
http://umbriel.dcs.gla.ac.uk/NeSC/general/esi

[2] I. Foster, C. Kesselman, J. Nick, S. Tuecke, *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*
http://www.globus.org/research/papers/ogsa.pdf

[3] W3C Web Services Activity:
http://www.w3.org/2002/ws/

[4] D.J. Hand, H. Mannilla, P. Smyth, *Principles of Data Mining*, (MIT Press)
http://mitpress.mit.edu/

[5] Home page of Jiawei Han
http://www-faculty.cs.uiuc.edu/~hanj

[6] Extensible Markup Language (XML)
http://www.w3.org/XML

[7] Simple API for XML (SAX)
http://www.saxproject.org

[8] Document Object Model (DOM)
http://www.w3.org/DOM

[9] Extensible Stylesheet Language Transformations (XSLT)
http://www.w3.org/TR/xslt

[10] Extensible Stylesheet Language (XSL)
http://www.w3.org/Style/XSL

[11] International Virtual Observatory Alliance, *VOTable: A Proposed XML Format for Astronomical Tables*
http://www.us-vo.org/VOTable/

[12] International Virtual Observatory Alliance
http://www.ivoa.net

[13] Martin Westhead, Mark Bull , *Representing Scientific Data on the Grid with BinX – Binary XML Description Language*
http://www.epcc.ed.ac.uk/~gridserve/WP5/Binx/sci-data-with-binx.pdf

[14] Edinburgh Parallel Computing Centre (EPCC)
http://www.epcc.ed.ac.uk

[15] National e-Science Centre (NeSC)
http://www.nesc.ac.uk

[16] OGSA-DAI GridServe Project
http://www.epcc.ed.ac.uk/~gridserve/

[17] Exodus Content Delivery Network Service
http://www.exodus.net/solutions/cdns/content_delivery.html

[18] OpenURL: A Transport Mechanism for Context Objects
http://www.niso.org/committees/committee_ax.html

[19] Simple Object Access Protocol (SOAP)
http://www.w3.org/TR/SOAP

[20] Data Mining Components from Numerical Algorithms Group
http://www.nag.co.uk/numeric/DR/drdescription.asp

[21] Chimera Virtual Data System
http://www-unix.griphyn.org/chimera

[22] Grid Physics Network (GriPhyN)
http://www.griphyn.org/

[23] Z39.50 Maintenance Agency Page
http://lcweb.loc.gov/z3950/agency

[24] Registry of Z39.50 registries in the UK
http://www.ukoln.ac.uk/distributed-systems/zdir/

[25] Universal Description, Discovery and Integration of Web Services (UDDI)
http://www.uddi.org

[26] R. J. Allen et. al., *UDDI and WSIL for e-Science*
http://esc.dl.ac.uk/Papers/UDDI/uddi/

[27] VRVS: Virtual Rooms Videoconferencing System
http://www.vrvs.org/

[28] Microsoft NetMeeting
http://www.microsoft.com/windows/netmeeting

[29] Virtual Network Computing (VNC)
http://www.uk.research.att.com/vnc or http://www.tightvnc.org/

[30] AccessGrid
http://www-fp.mcs.anl.gov/fl/accessgrid

[31] International Telecommunication Union Recommendation X.509
http://www.itu.int/rec/recommendation.asp?type=folders&lang=e&parent=T-REC-X.509

[32] S.K. Card, J.D. Mackinlay, B. Shneiderman, *Readings in Information Visualization: Using Vision to Think* (Morgan Kaufmann)
http://www.mkp.com/books_catalog/catalog.asp?ISBM=1-55860-533-9

[33] B. H. McCormick , T. A. De Fanti, M. D. Brown, *Visualization in Scientific Computing*, ACM SIGGRAPH Computer Graphics Volume 21 , Issue 6 (November 1987)

[34] Numerical Algorithms Group
http://www.nag.co.uk/

[35] IRIS Explorer
http://www.nag.co.uk/Welcome_IEC.html

[36] Landsat Program
http://geo.arc.nasa.gov/sge/landsat/landsat.html

[37] Pittsburgh Computational Astrostatistics Group:
http://www.picagroup.org

[38] Sloan Digital Sky Survey:
http://www.sdss.org

[39] Distributed Aircraft Maintenance Environment (DAME)
http://www.cs.york.ac.uk/dame

[40] AURA Home Page
http://www.cs.york.ac.uk/arch/nn/aura.html

[41] Globus Project
http://www.globus.org

[42] Council for the Central Laboratory of the Research Councils (CLRC)
http://www.clrc.ac.uk

[43] Data Portal
http://www.e-science.clrc.ac.uk/Activity/JS=FALSE;ACTIVITY=DataPortal;

[44] NERC Data Grid
http://ndg.badc.rl.ac.uk

[45] Climateprediction.net
http://www.climateprediction.rl.ac.uk

[46] SuperCOSMOS Sky Survey (SSS)
http://www-wfau.roe.ac.uk/sss

[47] UK Schmidt Telescope Unit
http://www.roe.ac.uk/ukstu
[48] SuperCOSMOS plate-measuring machine
http://www.roe.ac.uk/cosmos/scosmos.html
[49] Condor Project
http://www.cs.wisc.edu/condor
[50] Globus Toolkit
http://www.globus.org/toolkit
[51] Legion Project
http://legion.virginia.edu/
[52] Nimrod
http://www.csse.monash.edu.au/~davida/nimrod.html
[53] UNICORE
http://www.unicore.de
[54] Storage Resource Broker (SRB)
http://www.npaci.edu/DICE/SRB
[55] Web Services Description Language (WSDL)
http://www.w3.org/TR/wsdl.html
[56] Global Grid Forum (GGF)
http://www.gridforum.org
[57] BioDas.org: open source development of distributed annotation services in bioinformatics
http://www.biodas.org
[58] Open Archives Forum:
http://www.oaforum.org/overview/

Appendix A: Workshop Programme

Day One: Thursday, October 24th

10.00 Welcome & introduction to NeSC/eSI (Malcolm Atkinson)
10.15 Introduction to workshop and its aims (Bob Mann)
10.30 Scientific Motivation overview (Roy Williams)
11.00 Coffee
11.30 Data Mining overview (Chris Williams)
12.00 Data Visualization overview (Ken Brodlie)
12.30 Charge to groups in Breakout Session 1*
13.00 Lunch
13.30 Breakout Session 1
15.00 Tea
15.30 SDMIV Application Talk 1 (Bob Nichol)
16.15 SDMIV Application Talk 2 (Tom Jackson)
17.00 SDMIV Application Talk 3 (David Gilbert)
17.45 Formal close of Day One
19.30 Dinner

Day Two: Friday, October 25th

09.00 SDMIV Application Talk 4 (Julian Gallop)
09.20 SDMIV Application Talk 5 (Amos Storkey)
09.45 Current computational infrastructure (Rob Baxter)
10.30 Coffee
11.00 Future computational infrastructure (Malcolm Atkinson)
11.45 Report back from Breakout Session 1 and discussion
13.00 Lunch
13.45 Choice of topics for Breakout Session 2.**
14.00 Breakout Session 2
15.00 Tea
15.30 Report back from Breakout Session 2 and panel discussion of future actions.
17.00 Close of workshop

Notes on Breakout Sessions:
* Breakout Session 1: Identify challenges in overlap areas between principal communities represented at workshop, and list requirements that they place on computational infrastructure. Breakout groups: data mining and astronomy; data mining and bioinformatics; data mining and visualization; visualization and astronomy; visualization and bioinformatics.
** Breakout Session 2: Randomly-selected groups, each to identify up to six important issues to be studied further over the next 12-18 months, particularly where these impact on infrastructure development, and/or suggest components for the *SDMIV Grid Application Component Library* envisaged by Malcolm Atkinson's *Future Computational Infrastructure* talk.

Appendix B: Workshop Participants

Malcolm Atkinson	National e-Science Centre
Elizabeth Auden	Mullard Space Science Lab, UCL
Rob Baxter	Edinburgh Parallel Computing Centre
Lisa Blanshard	CLRC - Daresbury Laboratory
Thomas Boch	Centre de Données astronomiques de Strasbourg
François Bonnarel	Centre de Données astronomiques de Strasbourg
Ken Brodlie	School of Computing, University of Leeds
Simon Clifford	Dept of Chemistry, Queen Mary, University of London
John Collins	Institute of Cell & Molecular Biology, University of Edinburgh
Clive Davenhall	Institute for Astronomy, University of Edinburgh
Julian Gallop	CLRC – RAL
Bob Gibbins	National e-Science Centre
Nicolas Gilardi	Dalle Molle Institute for Perceptual Artificial Intelligence
David Gilbert	Bioinformatics Research Centre, University of Glasgow
Steven Gontarek	Scottish Association for Marine Science
Neil Hanlon	Bioinformatics Research Centre, University of Glasgow
Martin Hill	Institute for Astronomy, University of Edinburgh
Carren Holden	BAE Systems
Richenda Houseago-Stokes	Southampton Oceanography Centre
Ela Hunt	Department of Computing Science, University of Glasgow
Thomas Jackson	Department of Computer Science, University of York
Andy Knox	IBM
Andy Lawrence	Institute for Astronomy, University of Edinburgh
Bryan Lawrence	CLRC – RAL
Siu-wai Leung	Division of Informatics, University of Edinburgh
Tony Linde	Dept of Physics and Astronomy, University of Leicester
Elizabeth Lyon	UKOLN, University of Bath
Bob Mann	Institute for Astronomy, University of Edinburgh & NeSC
Ralph Mansson	Department of Mathematics, University of Southampton
Philip McNeil	European Bioinformatics Institute
Tara Murphy	Institute for Astronomy, University of Edinburgh
Matt Neilson	Institute of Biomedical & Life Sciences, University of Glasgow
Bob Nichol	Dept of Physics, Carnegie Mellon University
Patricio Ortiz	Dept of Physics & Astronomy, University of Leicester
James Osborne	Department of Computer Science, University of Hull
Stephen Pickles	Computer Services for Academic Research, Univ of Manchester
Greg Riccardi	Computer Science, Florida State University
Anita Richards	Jodrell Bank Observatory, University of Manchester
Jonathan Roberts	Computing Laboratory, University of Kent at Canterbury
Amos Storkey	Division of Informatics, University of Edinburgh
Aik Choon Tan	Bioinformatics Research Centre, University of Glasgow
Colin Venters	Manchester Visualization Centre, University of Manchester
Jeremy Walton	The Numerical Algorithms Group
Alan Welsh	Department of Mathematics, University of Southampton
Ted Qingying Wen	National e-Science Centre
Chris Williams	Division of Informatics, University of Edinburgh
Roy Williams	Centre for Advanced Computing Research, Caltech