KDD for Science Data Analysis: Issues and Examples

Usama Fayyad
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
fayyad@microsoft.com

David Haussler
University of California, Santa Cruz
Santa Cruz, CA 95064, USA
haussler@cse.ucsc.edu

Paul Stolorz
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109, USA
pauls@aig.jpl.nasa.gov

Abstract

The analysis of the massive data sets collected by scientific instruments demands automation as a pre-requisite to analysis. There is an urgent need to create an intermediate level at which scientists can operate effectively; isolating them from the massive sizes and harnessing human analysis capabilities to focus on tasks in which machines do not even remotely approach humans—namely, creative data analysis, theory and hypothesis formation, and drawing insights into underlying phenomena. We give an overview of the main issues in the exploitation of scientific datasets, present five case studies where KDD tools play important and enabling roles, and conclude with future challenges for data mining and KDD techniques in science data analysis.

keywords: Applications in Science, Data Analysis, overview article, large databases, automated analysis, scietific data sets, scientific discovery.

To appear: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, August 2-4, 1996, AAAI Press. http://www-aig.jpl.nasa.gov/kdd96

KDD for Science Data Analysis: Issues and Examples

Usama Fayyad
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
fayyad@microsoft.com

David Haussler
Computer Science Dept.
University of California, Santa Cruz
Santa Cruz, CA 95064, USA
haussler@cse.ucsc.edu

Paul Stolorz
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109, USA
pauls@aig.jpl.nasa.gov

Abstract

The analysis of the massive data sets collected by scientific instruments demands automation as a prerequisite to analysis. There is an urgent need to create an intermediate level at which scientists can operate effectively; isolating them from the massive sizes and harnessing human analysis capabilities to focus on tasks in which machines do not even remotely approach humans—namely, creative data analysis, theory and hypothesis formation, and drawing insights into underlying phenomena. We give an overview of the main issues in the exploitation of scientific datasets, present five case studies where KDD tools play important and enabling roles, and conclude with future challenges for data mining and KDD techniques in science data analysis.

1 Introduction

Scientists in a variety of fields are experiencing a data glut problem. Modern scientific instruments can collect data at rates that, less than a decade ago, were considered unimaginable. Scientific instruments, coupled with data acquisition systems, can easily generate terabytes and petabytes of data at rates as high as gigabytes per hour. Be it a satellite collecting data from a remote sensing platform, a telescope scanning the skies, or a microscope probing the minute details of a cell, the scientist at the other end of this data collection machinery is typically faced with the same problem: what do I do with all this data?

There is a rapidly widening gap between data collection capabilities and the ability of scientists to analyze the data. The traditional approach of a lone investigator, or even a team of scientists, staring at data in pursuit of (often hypothesized) phenomena or in search of some underlying structure is quickly becoming infeasible. The root of the problem is fairly simple: the data is increasing dramatically in size and dimensionality. While it is reasonable to assume that a scientist can work effectively with a few thousand observations, each having a small number of measurements, say 5, associated with it, it is not at all acceptable to assume that a scientist can effectively "digest" millions of data points, each with tens or hundreds of measurements.

Note that large data sets with high dimensionality can be effectively exploited when a problem is fully understood and the scientist knows what to look for in the data via well-defined procedures. In fact, the term used in many scientific disciplines is *data reduction*. By *reducing* data, a scientist is effectively bringing it down in size to a range that is "analyzable". However, where does this leave us if the phenomena are not completely understood? Since in scientific investigation we are often interested in new knowledge, the problem of effective manipulation and exploratory data analysis is looming as one of the biggest hurdles standing in the way of exploiting the data. If left unanswered, a scientist would have little choice but to use only parts of the collected data and simply "ignore" the rest of it. Since data collection is typically expensive, this would be a clear waste of resources, not to mentioned the missed opportunity for new knowledge and understanding.

2 Data Reduction Versus Automated Analysis

We believe that data mining and knowledge discovery in databases (KDD) techniques for automated data analysis have an important role to play as an interface between scientists and large data sets. Machines are still far from approaching human abilities in the areas of synthesis of new knowledge, hypothesis formation, and creative modelling. The processes of drawing insight and conducting investigative analyses are still clearly in the realm of tasks best left to humans. However, automating the data reduction procedure is a significant niche suitable for computers. Data reduction involves cataloging, classification, segmentation, partitioning of data, and so forth. This stage of the analysis process is well-suited for automation for the following reasons:

1. It requires dealing with the raw data and performing passes over large data sets.
2. Typical data reduction operations are fairly tedious and hence scientists are eager to cooperate in automating them.
3. Data reduction is usually decomposable into simpler independent tasks, hence one only needs to consider

solving the easier subproblems individually.

4. Humans reason about the underlying phenomena on levels higher than the low-level data. Sections 3.1, 3.2, and 3.3 provide good examples of how KDD can cover this gap.

Once a data set is reduced (say to a catalog or other appropriate form), the scientist can proceed to analyze it using more traditional (manual), statistical, or visualization techniques. For example, in the case of an astronomy sky survey, astronomers want to analyze catalogs (recognized, cataloged, and classified sky objects) rather than images. Reduction is equally important in time-series data (extracting features measured over sequences), for measurements obtained from spatially separated sensors, and for mapping raw sensor readings (e.g. multi-spectral) to a convenient feature space.

Higher-level "creative" analysis capabilities of humans, which machines are currently notably lacking, are put to better use if the lower level work is automated. The "higher" levels of analysis include theory formation, hypothesis of new laws and phenomena, filtering what is useful from background, and searching for hypotheses that require a large amount of highly specialized domain knowledge.

2.1 Data Considerations

Data comes in many forms: from measurements in flat files to mixed (multi-spectral/multi-modal) data that include time-series (e.g. sonar signatures or DNA sequences), images, and structured attributes. Most data mining algorithms in statistics [10] and KDD [12] are designed to work with data in flat files of feature vectors.

Image Data: common in science applications, it offers unique advantages in that it is relatively easy for humans to explore. Since the display format is predetermined, it is also fairly easy to display results (e.g. detections, classes). A user interface involving interactive and incremental analysis is also feasible since humans can "digest" a large number of values when represented as an image. On the other hand, image data poses serious challenges on the data mining side. Feature extraction becomes the dominant problem. Using individual pixels as features is typically problematic since a small subarea of an image easily turns into a high-dimensional vector (e.g. a 30×30 pixel region contains 900 individual values), and thus much more training data would be required to perform recognition or classification (see section 3.2). This is compounded by the fact that often the mapping from pixels to meaningful features is quite complex and noisy.

Time-series and sequence data: while it is easy to visualize for a single variable, time series data of multiple measurements are difficult to deal with, especially if the variables are collected at different rates (time scales). Time-series of continuous values are typically fairly non-smooth with random spiking and dipping. A discrete sequence of a single variable, such as a DNA molecule, can be quite complex and difficult to analyse due to its nonstationary behaviour, and the sometimes subtle signals associated with change of underlying hidden state variables that govern the process. Challenges include extracting stationary characteristics of an entire series, if it is stationary, and if not, segmentation to identify and extract non-stationary behavior and transitions between quantitatively and qualitatively different regimes in the series. Transition probabilities between process state variables must be inferred from the observed data. In many application areas, these problems have been attacked using Hidden Markov Models (HMMs) (see section 3.3).

Numerical measurements vs. categorical values: While a majority of measurements (pixels or sensors) are numeric, some notable examples (e.g. protein sequences, section 3.3) consist of categorical measurements. The advantage of dealing with numerical data is that the notion of "distance" between any two data points (feature vectors) is easier to define. Many classification and clustering algorithms rely fundamentally on the existence of a metric distance and ability to define means and centroids.

Structured and sparse data: In some problems variables may have some structure to them (e.g. hierarchical attributes or conditional variables that have different meanings under different circumstances). In other cases different variables are measured for different observations. Turning these data sets into standard flat file (feature vector) form is unlikely to be useful since it results in high dimensional sparse data sets. Unfortunately, there are few algorithms that are capable of dealing with structured data (e.g. [2, 6]).

Reliability of data (sensor vs. model data): Often, raw sensor-derived data is "assimilated" to provide a smooth homogeneous data product. For example regular gridded data is often required in climate studies, even when data points are collected haphazardly. This raises the question of data reliability - some data points need to be dealt with especially carefully, as they may not correspond to direct sensor-derived information.

3 Brief Case Studies

We shall briefly review five case studies in order to illustrate the contribution and potential of KDD for science data analysis. For each case, the focus will primarily be on impact of application, reasons why KDD systems suceeded, and limitations/future challenges.

3.1 Sky Survey Cataloging

The 2nd Palomar Observatory Sky Survey is a major undertaking that took over six years to complete [21]. The survey consist of 3 terabytes of image data containing an estimated 2 billion sky objects. The 3,000 photographic images are scanned into 16-bit/pixel resolution digital images at $23{,}040\times23{,}040$ pixels per image. The basic problem is to generate a survey catalog which records the attributes of each object along with its class: star or galaxy. The attributes are defined by the astronomers. Once basic image segmentation is performed, 40 attributes per object are measured. The problem is identifying the class of each object. Once the class is known, astronomers can conduct all sorts of scientific analyses like probing Galactic structure from star/galaxy counts, modelling evolution of galaxies, and studying the formation of large structure in the universe [28]. To achieve these goals we developed the SKICAT system (Sky Image Cataloging and Analysis Tool) [27].

Determining the classes (star vs. galaxy) for faint objects in the survery is a difficult problem. The majority of objects in each image are faint objects whose class cannot be determined by visual inspection or classical computational approaches in astronomy. Our goal was to classify objects that are at least one isophotal magnitude fainter than objects classified in previous comparable surveys. We tackled the problem using decision tree learning algorithms [11] to accurately predict the classes of objects. Accuracy of the procedure was verified by using a very limited set of high-resolution CCD images as ground truth.

By extracting rules via statistical optimization over multiple trees [11] we were able to achieve 94% accuracy on predicting sky object classes. Reliable classification of faint objects increased the the size of data that is classified (usable for analysis) by 300%. Hence astronomers were able to extract much more out of the data in terms of new scientific results [27]. In fact, recently this helped a team of astronomers discover 16 new high red-shift quasars in the universe in at least one order of magnitude less observation time [7]. These objects are extremely difficult to find, and are some of the farthest (hence oldest) objects in the universe. They provide valuable and rare clues about the early history of our universe.

SKICAT was successful for the following reasons:

1. The astronomers solved the feature extraction problem: the proper transformation from pixel space to feature space. This transformation implicitly encodes a significant amount of prior knowledge.
2. Within the 40 dimensional feature space, we believe at least 8 dimensions are needed for accurate classification. Hence it was difficult for humans to discover which 8 of the 40 to use, let alone how to use them in classification. Data mining methods contributed by solving the difficult classification problem.
3. Manual approaches to classification were simply not feasible. Astronomers needed an automated classifier to make the most out of the data.
4. Decision tree methods, although involving blind greedy search, proved to be an effective tool for finding the important dimensions for this problem.

Directions being pursued now involve the unsupervised learning (clustering) version of the problem. Unusual or unexpected clusters in the data might be indicative of new phenomena, perhaps even a new discovery. In a database of hundreds of millions of objects, automated analysis techniques are a necessity since browsing the feature vectors manually would only be possible for a small fraction of the survey. The idea is to pick out subsets of the data that look interesting, and ask the astronomers to focus their attention on those, perhaps perform further observations, and explain why these objects are different. A difficulty here is that new classes are likely to be a rare in the data, so algorithms need to be tuned to looking for small interesting clusters rather than ignoring them (see Section 4).

3.2 Finding Volcanoes on Venus

The Magellan spacecraft orbited the planet Venus for over five years and used synthetic aperture radar (SAR) to map the surface of the planet penetrating the gas and cloud cover that permanently obscures the surface in the optical range. The resulting data set is a unique high-resolution global map of an entire planet. In fact, we have more of the planet Venus mapped at the 75m/pixel resolution than we do of our own planet Earth's surface (since most of Earth's surface is covered by water). This data set is uniquely valuable because of its completeness and because Venus is most similar to Earth in size. Learning about the geological evolution of Venus could offer valuable lessons about Earth and its history.

The sheer size of the data set prevents planetary geologists from effectively exploiting its content. The first pass of Venus using the left-looking radar resulted in over 30,000 1000×1000 pixel images. The data set was released on 100 CD-ROMs and is available to anyone who is interested. Lacking the proper tools to analyze this data, geologists did something very predictable: they simply examined browse images, looked for large features or gross structure, and cataloged/mapped the large-scale features of the planet. This means that the scientist operated at a much lower resolution, ignoring the potentially valuable high resolution data actually collected. Given that it took billions of dollars to design, launch, and operate the sensing instruments, it was a priority for NASA to insure that the data is exploited properly.

To help a group of geologists at Brown University analyze this data set [1], the JPL Adaptive Recognition Tool (JARtool) was developed [4]. The idea behind this system is to automate the search for an important feature on the planet, small volcanoes, by training the system via examples. The geologists would label volcanoes on a few (say 30-40) images, and the system would automatically construct a classifier that would then proceed to scan the rest of the image database and attempt to locate and measure the estimated 1 million small volcanoes. Note the wide gap between the raw collected data (pixels) and the level at which scientists operate (catalogs of objects). In this case, unlike in SKICAT, the mapping from pixels to features would have to be done by the system. Hence little prior knowledge is provided to the data mining system.

Using an approach based on matched filtering for focus of attention (triggering on any candidates that vaguely resemble volcanoes; and with a high false detection rate), followed by feature extraction based on projecting the data onto the dominant eigenvecotrs in the training data, and then classification learning to distinguish true detections from false alarms, JARtool can match scientist performance for certain classes of volcanoes (high probability volcanoes versus ones which scientists are not sure about) [4]. Limitations of the approach include sensitivity to variances in illumination, scale, and rotation.

The use of data mining methods here was well-motivated because:

1. Scientists did not know much about image processing or about the SAR properties. Hence they could easily label images but not design recognizers; making the training-by-example framework natural and justified.
2. Fortunately, as is often the case with cataloging tasks, there was little variation in illumination and orientation of objects of interest. Hence the mapping from pixels to features can be performed automatically.
3. The geologists did not have any other easy means for finding the small volcanoes, hence they were motivated to cooperate by providing training data and other help.
4. The result is to extract valuable data from an expensive data set. Also, the adaptive approach (training by example) is flexible and would in principle allow us to reuse the basic approach on other problems.

With the proliferation of image databases and digital libraries, data mining systems that are capable of searching for content are becoming a necessity. In dealing with images, the train-by-example approach, i.e. querying for "things that look like this" is a natural interface since humans can visually recognize items of interest, but translating those visual intuitions into pixel-level algorithmic contraints is difficult to do. Future work on JARtool is proceeding to extend it to other applications like classification and cataloging of sun spots.

3.3 Biosequence Databases

In simplest computer form the human genome is a string of about three billion letters. The letters are A, C, G, and T, representing the four nucleic acids, the constituents of DNA, which are strung together to make the chromosomes in our cells. When combined into one string, the chromosomes contain our genetic heritage, a blueprint for a human being. A large international effort is currently underway to obtain this string. This project may be complete in as little as five years. However, obtaining the string is not enough. It has to be interpreted.

According to the central dogma of molecular biology, DNA is transcribed into RNA, and RNA is translated into protein by the molecular machinery within the cell. A piece of DNA that serves as a template for a protein in this fashion is called a gene. It is the proteins that do most of the work within the cell, and each of the approximately 100,000 different kinds of protein in a human cell has a unique structure and function. Certain RNA molecules, called structural RNA molecules, also have key roles other than producing proteins, and each of these also has a unique structure and function. Elucidating the structure and function of proteins and structural RNA molecules, for humans and for other organisms, is the central task of molecular biology.

There are several international databases of genetic sequences that coordinate, to a certain extent, the archiving of biosequences. The largest DNA database is GENBANK, maintained by the National Center for Biotechnology Information (NCBI) in Bethesda, with a database of about 400 million letters of DNA from a variety of organisms, and growing very rapidly. Two prominent protein databases are PIR and SWISSPROT. After the redundancies are removed from these protein databases, they contain about 200,000 different protein sequences.

The most pressing data mining tasks for biosequence databases are:

1. Find the genes in the DNA sequences of various organisms. It turns out that the genes are interspersed with DNA that has other functions, such as gene regulation, and it is difficult to locate the exact boundaries of the genes themselves, so that they may be extracted from the DNA database. Gene-finding programs such as GRAIL [29], GeneID [16], GeneParser [24], GenLang [3], FGENEH [23], Genie [19] and EcoParse [18] use neural nets and other AI or statistical methods (discussed further below) to locate genes in DNA sequences. Looking for ways to improve the accuracy of these methods is a major thrust of current research in this area.

2. Once a gene has been correctly extracted from the DNA, it is straightforward to determine the protein that it codes for, using the well known genetic code. Proteins can be represented as sequences over a 20 letter alphabet of amino acids. This is referred to as the primary structure of the protein. Each three consecutive letters of DNA code for one letter of protein according to the genetic code. While it is easy to determine the primary structure of a protein, in the cell the protein sequence folds up on itself in a fashion that is unique to each protein, giving it a higher order structure. Understanding this higher order structure is critical to understanding the protein's function. The situation is similar for structural RNA molecules. The second pressing task for biosequence database mining is to develop methods to search the database for sequences that will have similar higher order structure and/or function to the query sequence, rather than doing a more naive string matching, which only pays attention

to matches in the primary structure.

One statistical method that has shown promise in biosequence database mining is the use of Hidden Markov Models (HMMs) [17]. Two popular systems that use this method are HMMer [8] and SAM [15]. A hidden Markov model [20] describes a series of observations by a "hidden" stochastic process—a Markov process. In speech recognition, where HMMs have been used extensively, the observations are sounds forming a word, and a model is one that by its "hidden" random process generates certain sequences of sounds, constituting variant pronunciations of a single word, with high probability. In modeling proteins, a word corresponds to a protein sequence, and a family of proteins with similar structure and/or function, such as the globin proteins, which include the oxygen-carrying protein hemoglobin found in red blood cells, can be viewed as a set of variant pronunciations of a word. Hence, here the observations are the amino acids, and a model of a protein family such as the globin family is one that generates sequences of amino acids forming globins with high probability. In this way, the model describes not just one particular globin sequence, but the general structure of a globin sequence, explicitly modeling the possibility that in some globins, extra amino acids may be inserted in some places in the primary structure and deleted in other places.

It has been conjectured that there are only a few thousand different protein families in biology [5]. Once an HMM has been built for each of these families, or for the different protein domains within the sequences in these families, then it may be possible to assign tentative structure and function to newly discovered protein sequences by evaluating their likelihood under each of the HMMs in this model library, again, in analogy with the way that isolated words are recognized by HMM-based speech recognition systems. One difference is that in biology, the dictionary of fundamentally different protein structures/families is not simply provided to the designer of such a system, but must to a certain extent itself be discovered as part of the modeling process. This leads to a third data mining task, that of clustering protein sequences into families of related sequences to be modeled by a common HMM.

HMMs and variants of HMMs have also been applied to the gene-finding problem [19, 18], and to the problem of modeling structural RNA [9, 22]. The gene-finding methods GeneParser, Genie, and EcoParse mentioned above are examples of this. RNA analysis uses an extension of HMMs known as stochastic context-free grammars. This extension permits one to model certain types of interactions between the letters of the sequence that are distant in the primary structure but adjacent in the folded RNA structure, without incurring the huge computational overhead of the general protein threading models. However, there is still some significant overhead, making large database searches quite slow. On the other hand, using these models, one is able to do search based directly on high order structural similarity between molecules, which gives much better discrimination.

Computer-based analysis of biosequences is having an increasing impact on the field of biology. Computational biosequence analysis and database searching tools are now an integrated and essential part of the field, and have lead to numerous important scientific discoveries in the last few years. Most of these have resulted from database searches that revealed unexpected similarities between molecules that were previously not known to be related. However, these methods are increasingly important in the direct determination of structure and function of biomolecules as well. Usually this process relies heavily on the human application of biological knowledge and laboratory experiment, in conjunction with the results from the application of several different fairly simple programs that do statistical analysis of the data and/or apply simple combinatorial methods. HMMs and related models have been more successful in helping scientists with this task because they provide a solid statistical model that is flexible enough to incorporate important biological knowledge, Such knowledge is incorporated the form of hidden state structure and *a priori* parameter estimates. The key challenge for the future is to build computer methods that can interpret biosequences using a still more complete integration of biological knowledge and statistical methods at the outset, allowing the biologist to operate at a higher level in the interpretation process where his or her creativity and insight can be of maximal value.

3.4 Earth Geophysics - Earthquake Photography from Space

Important signals about temporal processes are often buried within noisy image streams, requiring the application of systematic statistical inference techniques. Consider for example the case of two images taken before and after an earthquake, at a pixel resolution of say 10 meters. If the earthquake fault motions are only up to 5 or 6 meters in magnitude, a relatively common scenario, then it is essentially impossible to describe and measure the fault motion by simply comparing the two images manually (or even by naive differencing by computer). However, by repeatedly registering different local regions of the two images, a task that is known to be do-able to subpixel precision, it is possible to infer the direction and magnitude of ground motion due to the earthquake. This fundamental concept is broadly applicable to many data mining situations in the geosciences and other fields, including earthquake detection, continuous monitoring of crustal dynamics and natural hazards, target identification in noisy images and so on.

Data mining algorithms of this kind need to simultaneously address three distinct problems in order to be successful, namely 1)design of a statistical inference engine that can reliably infer the fundamental processes to acceptable precision, 2) development and implementation of scalable algorithms on scalable platforms suitable for massive datasets, and 3) construction of automatic and reasonably seamless systems that can be used by domain scientists on a large number of datasets.

One example of such a geoscientific data mining system is Quakefinder [25], which automatically detects and measures tectonic activity in the Earth's crust by examination of satellite data. Quakefinder has been used to automatically map the direction and magnitude of ground displacements due to the 1992 Landers earthquake in Southern California, over a spatial region of several hundred square kilometers, at a resolution of 10 meters, to a (sub-pixel) precision of 1 meter. It is implemented on a 256-node Cray T3D parallel supercomputer to ensure rapid turn-around of scientific results. The issues of scalable algorithm development and their implementation on scalable platforms are quite general with serious impact to data mining with genuinely massive datasets.

The system addressed a definite scientific need, as there was previously no area-mapped information about 2D tectonic processes available at this level of detail. In addition to automatically measuring known faults, the system also enabled a form of automatic knowledge discovery by indicating novel unexplained tectonic activity away from the primary Landers faults that had never before been observed.

Quakefinder was successful for the following reasons:

1. It was based upon an integrated combination of techniques drawn from statistical inference, massively parallel computing and global optimization.
2. Scientists were able to provide a concise description of the fundamental signal recovery problem.
3. Portions of the task based upon statistical inference were straightforward to automate and parallelize, while still ensuring accuracy.
4. The relatively small portions of the task not so easily automated, such as careful measurement of fault location based on a computer-generated displacement map, are accomplished very quickly and accurately by humans in an interactive environment.

The limitations of the approach include the fact that it relies upon successive images being "similar enough" to each other to allow inference based upon cross-correlation measures. This is not always the case in regions where, for example, vegetation growth is vigorous. The method also requires reasonably cohesive ground motions over a number of pixels. It does not, however, require co-registered images, in contrast to many satellite image applications. Nevertheless, the overall system provides a fast, reliable, high-precision change analyzer able to measure earthquake fault activity to high resolution. The field of remote sensing is likely to become increasingly populated with data mining systems of this type in the future, in which dynamic phenomena are extracted directly from raw data, in addition to successful classification systems that deal with static imagery. One of the primary challenges for remote sensing will be generalization and extension of systems such as Quakefinder to deal with spatio-temporal information in an efficient, accessible and understandable form.

3.5 Atmospheric Science

Analysis of atmospheric data is a another classic area where processing and data collection power has far outstripped our ability to interpret the results. The mismatch between pixel-level data and scientific language that understands spatio-temporal patterns such as cyclones and tornadoes is huge. A collaboration between scientists at JPL and UCLA, has developed CONQUEST (COncurrent QUErying Space and Time) [26], a scientific information system implemented on parallel supercomputers, to bridge this gap.

Parallel testbeds (MPP's) were employed by Conquest to enable rapid extraction of spatio-temporal features for content-based access. In some cases, the features are known beforehand, e.g. detection of cyclone tracks. Other times indexable features are hidden in the enormous mass of data. Hence one of the goals here has been the development of "learning" algorithms on MPPs which look for novel patterns, event clusters or correlations on a number of different spatial and temporal scales. MPPs are also used by CONQUEST to service user queries requiring complex and costly computations on large datasets.

An atmospheric model can generate gigabytes of data covering several years of simulated time on a $4^\circ \times 5^\circ$ resolution grid. We have implemented parallel queries concerning the presence, duration and strength of extratropical cyclones and distinctive "blocking features" in the atmosphere, which can scan through this dataset in minutes. Other features of interest are being added, including the detection and analysis of ocean currents and eddies. Upon extraction, the features are stored in a relational database (Postgres). This content-based indexing dramatically reduces the time required to search the raw datasets of atmospheric variables when further queries are formulated. Also featured are parallel implentations of singular value decomposition and neural network pattern recognition algorithms, in order to identify spatio-temporal features as a whole, in contrast to the separate treatment of spatial and temporal information components that has often been used in the past to study atmospheric data.

The long term goal of projects such as Conquest is the development of flexible, extensible, and seamless environments for scientific data analysis, which can be applied ultimately to a number of entirely different scientific domains. Challenges here include the ability to formulate compound queries spread across several loosely federated databases, and the construction and integration of high-bandwidth I/O channels to deal with the massive sizes of datasets involved. Alghough these ideas and systems are still in their infancy, their potential impact on fields that are currently overwhelmed by the sheer volume of high-resolution spatial and temporal imagery cannot be overestimated.

4 Issues and Challenges

Several issues need to be considered when contemplating a KDD application in science data sets. We summarize some of these below.

Feature extraction: Can the scientist provide transformations from low-level data to features? While some classification problems might be too difficult for humans to perform, it is often possible for the user to provide significant amounts of domain knowledge by stating key attributes to measure. Often, sufficient information is contained in the attributes, but the scientist does not know how to use the high-dimensional feature space to perform classification (e.g. in the SKICAT and the gene-finding problems).

Minority (low probability) classes: in problems of automated discovery where algorithms are being used to sift through large amounts of data, the new class of interest may occur only with very low probablity (e.g. one case per million). Traditional clustering techniques would ignore such cases as "noise". Random sampling would fail by definition. Specialized algorithms or biased sampling schemes are needed.

High degree of confidence: a dimension along which science applications of data mining differ from their commercial or financial counterparts is that high acuracy and precision in prediction and description are required (e.g. in SKICAT, a 90% or better confidence level was required, otherwise results of cataloging cannot be used to test or refute competing theories). Similar high accuracies are required in gene-finding.

Data mining task: The choice of task (see [13] for a list of tasks) is important. For example, supervised classification is generally easier to perform than unsupervised learning (clustering). Rather than simply discriminating between given classes, a clustering algorithm must "guess" what the key (hidden) variable is. Regression (where the class variable is continuous) can be easier to do than classification, hence it may be better to map a classification problem into a regression problem where one is attempting to predict the *probability* of a class or some related smooth quantity.

Understandability of derived models: is an important factor if ultimately the findings need to be interpreted as knowledge or explained. In cases where certain steps are being automated in pre-processing (e.g. JARtool), understandability may not be an issue.

Relevant domain knowledge: unfortunately, other than at the stage of feature definition, most current data mining methods do not make use of domain knowledge. Such knowledge can be critical in reducing the search space an algorithm has to explore. In science applications a large body of knowledge on the topic at hand is typically available.

Scalable machines and algorithms: The sheer scale of modern-day datasets require the highest level of computational resources to enable analysis within reasonable time scales. Apart from the issue of raw CPU power, many data mining applications require fast I/O as the fundamental resource, while others rely on large internal memory. Scalable I/O and scalable computing platforms, together with suitably crafted scalable algorithms, are crucial ingredients.

In conclusion, we point out that KDD applications in science may in general be easier than applications in business, finance, or other areas. This is due mainly to the fact that the science end-users typically know the data in intimate detail. This allows them to intuitively guess the important transformations. Scientists are trained to formalize intuitions into procedures/equations making migration to computers an easier matter. Background knowledge is usually available in well-documented form (papers and books) providing backup resources when the initial data mining attempts fail. This luxury (sometimes a burden) is not usually available in fields outside of science. Finally, the fact that scientists typically use high-tech instruments and equipment in their daily chores biases them (as a community) to look favourably upon new techniques for analysis that in other communities may be shunned as "experimental".

5 Acknowledgements

The authors are grateful to all their collaborators on the projects described within this summary paper. They are too numerous to list here. The work described in this paper was performed in part at Jet Propulsion Laboratory, California Institute of Technology under a contract from the National Aeronautics and Space Administration.

References

[1] J. Aubele, L. Crumpler, U. Fayyad, P. Smyth, M. burl, and P. Perona (1995), In *Proc. 26th Lunar and Planetary Science Conference*, 1458, Houston, TX: LPI/USRA.

[2] Auriol, Manago, Althoff, Wess, and Dittrich (1995) "Integrating Induction and Case-Based Reasoning : Methodological Approach and First Evaluations" in *Advances in Case-Based Reasoning*, Haton J.P., Keane M. & Manago M. (Eds.) pp. 18-32, Springer Verlag, 1995.

[3] S. Dong and D. B. Searls (1994). "Gene Structure Prediction by Linguistic Methods", *Genomics*, 162:705–708.

[4] M.C. Burl, U. Fayyad, P. Perona, P. Smyth, and M.P. Burl (1994). "Automating the Hunt for Volcanoes on Venus", in *proc. of Computer Vision and Pattern Recognition Conference (CVPR-94)*, pp. 302-308, IEEE CS Press.

[5] C. Chothia (1992). "One thousand families for the molecular biologist", *Nature*, 357:543–544.

[6] S. Djoko, D. Cook, and L. Holder (1995). "Analyzing the Benefits of Domain Knowledge in Substructure Discovery", in *Proc. of KDD-95: First International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: The AAAI Press.

[7] J.D. Kennefick, R.R. De Carvalho, S.G. Djorgovski, M.M. wilber, E.S. Dickinson, N. Weir, U. Fayad, and J. Roden (1995). *Astronomical Journal*, 110-1:78-86.

[8] S. Eddy (1995). "Multiple alignment using hidden Markov models", *Proc. Conf. on Intelligent Systems in Molecular Biology*, AAAI/MIT Press.

[9] S. Eddy and R. Durbin (1994). "RNA sequence analysis using covariance models", *Nulceic Acids Research*, 22:2079-2088.

[10] J. Elder and D. Pregibon (1996). "Statistical Perspectives on KDD", in *Advances in Knowledge Discovery in Databases*, U. Fayyad et al (Eds.). Cambridge, MA: MIT Press.

[11] U.M. Fayyad, N. Weir, and S. Djorgovski (1993) Skicat: A machine learning system for the automated cataloging of large-scale sky surveys. In *Proc. of 10th International Conference on Machine Learning*, pp 112–119.

[12] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (1996). *Advances in Knowledge Discovery in Databases*, Cambridge, MA: MIT Press.

[13] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth (1996). "From Data Mining to Knowledge Discovery: An Overview", in *Advances in Knowledge Discovery in Databases*, U. Fayyad et al (Eds.). Cambridge, MA: MIT Press.

[14] J. W. Head et al. (1992) Venus volcanism: classification of volcanic features and structures, associations, and global distribution from magellan data. *Journal Geophysical Res.*, 97(E8):13153–13197.

[15] R. Hughey and A. Krogh (1995). "SAM: Sequence alignment and modeling software system", *tech. Rep. UCSC-CRL-95-7*, University of California, Santa Cruz.

[16] R. Guigo, S. Knudsen, N. Drake, and T. Smith (1992) Prediction of Gene Structure. *J. Mol. Biol.* 226:141–157",

[17] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler (1994). "Hidden Markov models in computational biology: Applications to protein modeling", *J. Mol. Biol.*, 235:1501–1531

[18] A. Krogh, I. S. Mian and D. Haussler (1994) "A Hidden Markov Model that finds genes in *E. coli* DNA", *Nucleic Acids Research*, 22:4768-4778.

[19] D. Kulp, D. Haussler, M. Reese, and F. Eeckman (1996). "A generalized hidden Markov model for the recognition of human genes in DNA", *Proc. Conf. on Intelligent Systems in Molecular Biology* AAAI Press.

[20] L. R. Rabiner (1989) "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE* vol. 77:257–286.

[21] I. Reid et al D., (1991) "The Second Palomar Sky Survey". *Publications of the Astronomical Society of the Pacific*, vol. 103, no. 665.

[22] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R.C. Underwood, and D. Haussler (1994). "Stochastic Context-Free Grammars for tRNA modeling", *Nulceic Acids Research*, 22:5112-5120.

[23] V. Solovyev, A. Salamov, and C. Lawrence (1994). "Predicting internal exons by oligonucleotide composition and discriminant analysis of splicable open reading frames", *Nucl. Acids Res.* 22:5156–5163.

[24] E.E. Snyder and G.D. Stormo (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks, *Nucl. Acids Res.* 21:607–613.

[25] P. Stolorz, C. Dean, R. Crippen, and R. Blom (1995), "Photographing Earthquakes from Space", in *Concurrent Supercomputing Consortium Ann. Rep.*, ed. T. Pauna, 20-22.

[26] P. Stolorz et al (1995) "Fast Spatio-Temporal Data Mining of Large Geophysical Datasets", in *Proc. 1st International Conf. on Knowledge Discovery and Data Mining*, pp. 300-305, AAAI Press.

[27] N. Weir, U.M. Fayyad, and S.G. Djorgovski (1995) Automated Star/Galaxy Classification for Digitized POSS-II. *The Astronomical Journal*, 109-6:2401-2412.

[28] N. Weir, S.G. Djorgovski, and U.M. Fayyad (1995) Initial Galaxy Counts From Digitized POSS-II. *Astronomical Journal*, 110-1:1-20.

[29] Y. Xu, J.R. Einstein, M. Shah, and E.C. Uberbacher (1994). "An improved system for exon recognition and gene modeling in human DNA sequences.", *Proc. Conf. on Intelligent Systems in Molecular Biology*, Menlo Park, CA: AAAI/MIT Press.