NASA Workshop on Issues in the Application of Data Mining to Scientific Data

Jeanne Behnke
NASA Goddard Space Flight Center
Earth Science Data and Information System
Project/Code 423
Greenbelt, MD 20771
jeanne.behnke@gsfc.nasa.gov

Elaine Dobinson
Jet Propulsion Laboratory
Pasadena, CA 91109
elaine.r.dobinson@jpl.nasa.gov

ABSTRACT

In this paper, we describe the NASA sponsored workshop on Issues in the Application of Data Mining to Scientific Data. The workshop was held at the University of Alabama in Huntsville on October 19-21, 1999. The full text of the report can be found in PDF and MSWord format at the following website: http://www.cs.uah.edu/NASA_Mining/

Keywords

Data Mining, NASA, earth science, statistics.

1. EXECUTIVE SUMMARY

On October 19-21, 1999, NASA sponsored a workshop on **Issues in the Application of Data Mining to Scientific Data** at the University of Alabama in Huntsville (UAH). A team from the Goddard Space Flight Center, the Jet Propulsion Laboratory, UAH, and the Ames Research Center organized the workshop. The objectives of the workshop were to bring together computer scientists and physical scientists to assess the state of the practice in data mining for scientific applications; share experiences in applying this technology; and to help determine the direction of future work in data mining. A web page describing the workshop can be found at: http://www.cs.uah.edu/NASA_Mining/

The workshop agenda was organized around a mixture of presentations and disciplinary theme groups. Presentations provided examples of data mining applications as well as discussions of certain analytical techniques. Dr. Ghassem Asrar, the Associate Administrator (AA) for the NASA Office of Earth Science, discussed future directions for NASA Earth Science. He stressed the need to go beyond the delivery of data to the delivery of information and knowledge derived from NASA's remote sensing missions. Theme groups addressed a set of questions, in the context of three different earth science themes. The groups were 1) solid earth and natural hazards; 2) global climate change; and 3) terrestrial ecology. The questions that each group was asked to address were:

- What data mining techniques have been successfully applied?
- What are the key characteristics of the theme area that need to be understood?
- What is the next set of science questions to which data mining could be applied?
- What are the scientific grand challenges to which data mining could contribute?
- What data mining research and development objectives follow from these scientific questions?

1.1 Key Findings

1.1.1 Data volume and resolution are key drivers for data mining

A major theme that emerged consistently throughout the meeting was the clear need to respond to new data analysis challenges posed by the overwhelming volume and high resolution of datasets generated by modern-day satellites and other data acquisition systems. This was seen universally as the fundamental driver for data mining techniques and systems. Without a systematic effort to generate data mining solutions, it was agreed that much of the data acquired at great expense by the Earth Science Enterprise will be greatly under-exploited, and our efforts to understand the Earth will be considerably diminished.

1.1.2 Component disciplines

A number of important technology disciplines were identified as being fundamentally important to the overall data mining enterprise. They include:

- Machine learning
- Statistics
- Database technology
- High performance computing
- Data visualization
- Image processing

It was generally agreed that, typically, *combinations* of several methods drawn from each discipline were needed to make significant progress on the fundamental issues associated with analyzing large-scale data.

1.1.3 Earth Science data mining

A number of possible definitions of data mining were discussed, and the needs of "scientific data mining" were compared and contrasted with broader data mining activities in the commercial sector. It was difficult to arrive at a consensus for the definition of data mining, apart from the clear importance of scalability as an underlying theme. In particular, it was hard to isolate a core set of fundamental techniques that clearly distinguish data mining from any single component discipline: in some way it is a uniquely powerful combination of individual techniques that characterizes

the field. One participant offered this definition of earth science data mining as a result of the meeting:

> The science, tools, environment, and facilities to scale up and/or automate scientific analysis of large-scale data streams, consisting of:
>
> - *Exploration of anomalies in geophysical data, where the detection of an anomaly may initiate an 'alert' requiring further human-in-the-loop analysis (e.g. using statistical or other methods);*
>
> - *Scaling up of current analysis techniques that detect known phenomena such that large-scale data product streams may be automatically analyzed;*
>
> and characterized by:
>
> - *Critical partnerships between physical scientists, computer scientists, and statisticians for the effective integration of analysis processes, scientific algorithms, statistical approaches, and enabling computer architectures.*

Viewing data mining as a general approach that looks in depth at each of its component disciplines "raises the profile" of the techniques within each discipline that best address the scaling problems introduced by high data volume and high resolution. As computing resources have grown, the value of these methods as broadly useful ways to bring structure to large data sets has increased substantially. By building on these ideas, extending them in new ways, and applying the methods to ever larger and more diverse databases, in conjunction with scalable methods from other fields, they can be used in data mining.

1.1.4 Motivation for teaming

It was consistently stressed that successful data mining can only come about through the formation of coherent interdisciplinary teams involving end users, systems implementers and algorithm developers. Teaming in a wide range of technology development and implementation activities is crucial to success, and must be closely coordinated with any sponsoring activities by NASA and other agencies that invest in data mining.

1.1.5 Importance of Major Challenges

A number of major challenges were identified, along with several past successes, that can serve to generate impetus within the scientific and technology communities for aggressive development of data mining methods and applications. Several of these challenges emerged within the context of the three theme groups. Although the burden of meeting these challenges is great, there is clearly enormous benefit to the public and the scientific community by rising to them.

2. RECOMMENDATIONS

Data mining techniques are key elements in making use of the prodigious volume of data that NASA and other earth remote sensing organizations will be capturing in the coming millennium. As such, it is important that investments be made in developing advanced data mining tools and technologies. The following recommendations were made:

- Consider developing NASA funding opportunities that are focused on data mining. Examples include funding opportunities based on an ESE challenge or prototyping.
- Since data mining includes techniques designed to address the discovery of features in large data sets that might not otherwise be detectable, proposals containing data mining elements should be seriously considered on their data mining merits as well as other factors that may apply.
- Characterize and track data mining technologies in the Earth Science Technology Office (ESTO) database.
- Develop a data mining journal that would collect contributions from projects that involve science data mining.
- Establish a regular conference on data mining that includes other related disciplines.
- Consider ways of facilitating teaming between data mining practitioners and NASA scientists on science data projects.

It was also recommended that a follow-on forum on data mining be organized. Such a forum would: 1) be open to all interested participants; 2) be conducted as a focused symposium consisting of both invited and contributed papers; 3) include scientific data mining efforts in disciplines other than earth science; and 4) include a tutorial session prior to the symposium for those scientists that would like to develop a more in-depth understanding of data mining techniques. It was proposed that such a symposium might be sponsored jointly with the National Science Foundation.

3. OVERVIEW OF THE WORKSHOP

Scientific data collections, including those from NASA and non-NASA satellites, are growing at an increasing rate. If they can be processed and analyzed, the data hold great potential for the discovery of new knowledge. A significant problem is how to provide users with the means to effectively extract useful information from this growing volume of data. In the last few years, a new discipline called data mining has emerged that enables knowledge extraction from this large volume of data.

Data mining consists of an evolving set of techniques that can be used to extract valuable information and knowledge from massive volumes of data. Up to this time, data mining research and tools have primarily focused on commercial sector applications. Only a limited amount of data mining research has focused on scientific data and remotely-sensed satellite data. While a number of conferences on various aspects of data mining have included some discussion of scientific data mining, and data mining has been considered in the context of scientific data conferences, a focused interchange of ideas between domain scientists and the data mining community on issues important to mining scientific data has been lacking.

It was determined that a workshop could serve as a vehicle to bring together these two communities so they could begin to identify issues and formulate mutually beneficial data mining objectives. The organizing committee decided that the initial workshop should be a small, focused session to facilitate a useful interchange of ideas. The initial workshop, sponsored by NASA, was held over two and a half days, October 19-21, 1999, at the University of Alabama in Huntsville. For this first workshop, the focus was limited to data mining in the earth science domain.

The organizing committee consisted of the following members:

Dr. Sara Graves	Univ. of Alabama in Huntsville
Thomas Hinke	NASA Ames Research Center
Elaine Dobinson	Jet Propulsion Laboratory
David A. Nichols	Jet Propulsion Laboratory
Paul E. Stolorz	Jet Propulsion Laboratory
Karen L. Moe	NASA Goddard Space Flight Ctr
Jeanne Behnke	NASA Goddard Space Flight Ctr

3.1 Objectives

The general objectives of the data mining workshop were oriented toward identifying issues relevant to data mining as applied to scientific data. The objectives included:

- Bring together representatives of the data mining community and the domain science community so that they can begin to understand the current capabilities and research objectives of each other's communities related to data mining.
- Identify a set of research objectives from the domain science community that would be facilitated by current or anticipated data mining techniques.
- Identify a set of research objectives for the data mining community that could support the research objectives of the domain science community.
- Identify any requirements for additional national infrastructure to support data mining.

Another objective of this workshop was to create a report for NASA that would discuss the application of data mining to the earth science domain. To achieve this, group participation was deemed to be critical. The organizing committee decided to create breakout groups in three theme areas to focus discussion on data mining and the earth science domain. The three theme groups included solid earth and natural hazards, global climate change, and terrestrial ecology. The theme groups were integral to the overall success of the workshop.

Each theme group was to address the following questions:

- What data mining techniques have been successfully applied?
- What are the key characteristics of the theme area that need to be understood?
- What is the next set of science questions to which data mining could be applied?
- What are the scientific grand challenges to which data mining could contribute?
- What data mining R&D objectives follow from these scientific questions?

3.2 Attendance

Participation in the workshop was by invitation in order to facilitate informal interaction and discussion between earth scientists and computer scientists in order to understand what the science community would like to see from the data mining community. The list of attendees can be found in the original report at the website.

3.3 Agenda

The workshop began with keynote presentations from a data mining authority, Dr. Paul Stolorz, and a domain science authority, Dr. Joseph Coughlan. These two keynote speakers respectively characterized the current state of data mining and the state of earth science problems that are currently challenging existing technology or are anticipated to challenge it in the future. Key to the workshop was the formal presentation of three case studies in which data mining has been applied to a science domain: a natural hazards case study, an atmospheric science case study and an environmental epidemiology case study. Synopses of the case studies can be found in the original report at the website. A Mining Technology/Science Snap Shots session highlighted various projects relevant to data mining. These smaller, invited talks provided more examples of science themes and data mining. A copy of the agenda is contained can be found in the original report at the website. Some presentations are available through links from the agenda of the conference on the web page: http://www.cs.uah.edu/NASA_Mining/. A highlight of the workshop was the discussion with Dr. Ghassem Asrar, the AA for the Office of Earth Science at NASA Headquarters, which served to reinforce topics discussed at the workshop. The key point of Dr. Asrar's discussion was a description of his priorities for earth science. He mentioned that he had three priorities for NASA programs: 1) doing first rate science, 2) building a comprehensive information system and 3) having a good plan of action for the next decade.

The presentations served as catalysts for discussions that were held within smaller theme groups, each of which consisted of an interdisciplinary group of scientists and data mining experts. These theme groups addressed a number of questions that formed the basis for presentations at the conclusion of the workshop. Organizing committee members served as Recorders for a group, allowing the Chairs and participants to concentrate on the discussion flow. There were group discussions and reports on all 2.5 days of the workshop.

3.4 Theme Group Discussion

The responses of each theme group to the five discussion questions are provided in Sections 3.4.1-3.4.5. While each group's responses varied due to their differing earth science perspectives, a number of common themes could be identified for each of the questions posed. Some general points that emerged from the discussions are:

- Data mining techniques that have been applied successfully among the various groups include feature extraction, correlation analysis, anomaly detection, pattern recognition and filtering.
- All of the groups emphasized that data should be readily available for data mining applications, i.e. online rather than in archives.
- Tele-connection was an important theme with respect to the next set of science questions to which data mining could be applied. Two such questions are: 1) How do ocean/atmosphere/land phenomena occurring in one geographic location affect the ocean/atmosphere/land environments proximate to and geographically distant from that location; and 2) How are ocean, atmosphere and land processes coupled?
- The greatest variability in responses was in the major scientific challenges identified by each group. The solid earth and natural hazards group focused on predictive

statistical models, physical/mechanistic models, and event-driven data acquisition. The global climate change and terrestrial ecology groups both identified the importance of understanding how changes in terrestrial ecology relate to/serve as an indicator of global climate change. The so-called "missing carbon" question was identified as both a scientific question to be answered and a major scientific challenge.

- Common data mining research & development objectives included:
  - promoting the interoperability of disparate data sets in multiple formats through commensurate measurements, metadata standards and retention of error bars, confidence limits and uncertainty;
  - enhancing visualization techniques; and
  - automating query expression/optimization and the use of appropriate metadata and correlation detection.
- An on-line repository of data mining tools and the education of the science community in the use of those tools were also identified as significant goals.

3.4.1 Theme Group Responses for Question 1

Question 1 was **What data mining techniques have been successfully applied?**

The Global Change group submitted this list:

- Subsetting and subsampling via classification
- Detection of eddies and fronts using Sobel filters
- European Center for Medium Range Weather Forecasts (ECMWF) cyclone identifier from GSFC using neural net classification
- Thresholding (e.g. to remove clouds)
- Coupled pattern analysis
- Low pass filters (e.g. to look for mesoscale features)
- Principal component analysis – Empirical Orthogonal Functions (EOFs) – Latitudinal variation in Rossby radius of deformation
- Spectral analysis for time series data
- Are we smoothing away important variations in the data—how can we preserve these? What technologies look for unusual features? –Data mining is one of these--develop loss functions for unusual features and look for data mining methodologies that look for features based on the loss function. This is implicitly Bayesian.
- Troubleshooting for outliers using statistics
- Wavelets to look for trends and surprises
- Correlation analysis and then look at residuals
- Visualization to study 8-dimensional space of aerosol characteristics for satellite data retrieval

The Natural Hazards group developed this list:

- Feature extraction: cyclone detection
  -- Supports clustering of storms, spatial and temporal
  -- Supports searches for periodicities
- Feature extraction: surface change detection
  -- Sub-pixel surface earthquake rupture
  -- Visualization
- Fires: pattern recognition for decision support
  -- Scaling relations between fire sizes and distribution
  -- Correlations between fuel type/moisture, recent history, topography, hot spots, wind speed direction/variability, equipment access
  -- Early detection and prediction via remote sensing
- Disease surveillance – correlation between the Advanced Very High-Resolution Radiometer/Sea Surface Temperature(AVHRR/SST) and timing and location of Rift Valley Fever outbreaks

The Terrestrial Ecology group developed this list:

- Supervised and unsupervised land cover mapping
- Analyzing growing season anomalies in AVHRR-Normalized Difference Vegetation Index (NDVI); found anomaly in Northwest Territory and Alaska
- Vegetation habitat analysis using clustering and ordination techniques

3.4.2 Theme Group Responses for Question 2

Question 2 was **What are the key characteristics (questions) of the theme area that need to be understood?**

The Global Change group submitted:

- Is the climate changing?
- What are the signatures of climate change? – Climate change (e.g. global warming) fingerprinting
- If it is changing, why is it?
  -- Anthropogenic vs. natural
  -- Feedback processes affected
- How can we use this to predict future climate, given different policy choices?

The Natural Hazards group submitted:

- Automate analysis functions (e.g., recognition of volcanoes)
- Anomaly detection
- On-line models and on-line databases
  -- Open architecture, object models
  -- XML DTD's (eXtensible Markup Language Document Type Definition) for models and data
- Pattern recognition
  -- Dense: earthquake time series data
  -- Sparse: pattern recognition for wild fires, etc.
- Event-driven data acquisition/flagging
  -- e.g., volcanic ash clouds
- False alarm problem is a major issue for natural hazards warning in general: cost, trust
- Human impact: population density, building inventory

The Terrestrial Ecology group submitted:

- Datasets are both large and small, with a lot of variability in the source and format of datasets
- Heterogeneous datasets (variable by spectral, spatial, and temporal characteristics)
- Real-time is not important for research, but highly important for applications
- Human-influenced (driven) variabilities, factors often of interest
- Metadata and documentation is very critical, standards are important here
- Clouds are noise in this theme area

3.4.3 *Theme Group Responses for Question 3*

Question 3 was **What is the next set of science questions to which data mining could be applied?**

The Global Change group submitted:

- Mine radar data for storms
- Missing carbon question—what is its distribution in the terrestrial ecosystem, and what is the distribution in the ocean. What is the net flux?
- Look for tele-connections (e.g. El Nino and more rain in midwest) among the ocean, atmosphere and land environments
- Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) – mine for Primary Production
- Use ocean data to study hydrologic cycle
- Mining partial clouds to understand their effects better
- Look for unexpected patterns and features in existing long-term data: 3 to 5 yr spatial/temporal patterns from missions such as Earth Radiation Budget Experiment (ERBE), AVHRR, Landsat, Total Ozone Mapping Spectrometer (TOMS), Stratospheric Aerosols and Gas Experiment (SAGE), Geostationary Operational Environmental Satellite (GOES), Topography (ocean) Experiment (TOPEX)/Poseidon
- Comment: Exploratory data analysis part of data mining needs to receive enhanced support

The Natural Hazards group submitted:

- Earthquakes
  -- Discriminate between young and old faults, then survey
  -- Mine time series data for properties of the crust: viscosity, thickness, rigidity, stress, strain, fault geometry (Global Positioning System (GPS), Interferometric Synthetic Aperture Radar (InSAR), seismic, strainmeters, magnetic, topographic)
  -- Space-time correlations: triggered slip, foreshocks & aftershocks
  -- Identify spatial patterns (feature extraction): blocks, strain accumulation zones
- Cyclones
  -- Predicting storm tracks
  -- Predicting changes in intensity as storm approaches land
  -- Need more data to understand microphysics of storms => data fusion

The Terrestrial Ecology group submitted:

- Tele-connection analysis is important (manifestation of problems introduced by geographic distances): applies to drought, seasonal changes, space-time data analysis
- Quantifying global-scale disturbance rates (don't have automated techniques to do this): fire deforestation, urbanization. E.g., has land clearing perturbed hydrologic runoff and flood frequency?
- Global-scale estimates/rates/interpretation of terrestrial ecological attributes (biomes, Net Primary Production (NPP), Leaf Area Index (LAI)) similar to weather. Inferences based on satellite data
- Predicting conditions that result in droughts, floods, fire potential, etc. using time series analysis; anomaly detection and causality
- What are the most logical and efficient couplings of terrestrial ecosystem modeling for integration with ocean and atmospheres for Earth System Models?

3.4.4 *Theme Group Responses for Question 4*

Question 4 was **What are the scientific grand challenges to which data mining could contribute?**

The Global Change group submitted:

- Is the climate changing? -- What are the signatures of climate change? – Climate change (e.g. global warming) fingerprinting
- Investigate linkages between hydro cycle with radiation balance with ocean circulation dynamics with bio-geochemistry of the oceans with terrestrial ecology--use data mining to find surprises in the data to help with understanding these: exploratory data analysis, testing of hypotheses
- Analysis of synoptic events having regional climate impact (e.g., volcanic eruptions, desert flash floods)
- Issue of downscaling from General Circulation Models – can data mining help reduce the output?

The Natural Hazards group submitted:

- Spatial-temporal-spectral pattern recognition
  --Unsupervised learning, SVD, clustering, etc.
- Predictive statistical models: seismic activity, fire spreading
  --Regression/classification
  --Multiple time series
  --multiple sensor/instrument modalities ("fusion")
- Physical/mechanistic models: fault modeling, interactions between crust/mantle, stress transfer
  --Collaborative distributed modeling
  --Distributed heterogeneous data
  --Distributed models
  --Model cross-validation
- Rapid systematic dissemination of results and warnings
  Event-driven data acquisition
  -- Event recognition
  --- Ground-based: GPS, seismic
  --- Onboard (real-time): volcanic ash plumes, fires, floods, etc.
  -- Adaptive data acquisition
  --- Closed loop autonomy for directed acquisition of data, early warning, etc.
  – Prioritize data downlink
  – Retarget instruments/sampling rates

The Terrestrial Ecology group submitted:

- Evaluating and predicting global sustainability
- Where does the missing carbon go?
- When and where will the global availability of fresh water become a problem? Precipitation balance – missing part of the Earth where we don't know the precipitation
- What agricultural areas will be the winners and losers in future global change trends?
- Accurate global land cover characterization. Current maps are not accurate.
- Understanding how ocean, atmosphere and land processes are coupled, and the role of human influences. E.g., where

and when does El Nino/La Nina influence land productivity?
- Understanding interannual variability and decadal trends in climate and ecosystem response, and land surface feedbacks to climate.
- Identifying terrestrial change detection as indicator of global climate change

3.4.5 Theme Group Responses for Question 5

Question 5 was **What data mining R&D objectives follow from these scientific questions?**

- The Global Change group submitted:
- Data mining environments – flexible, reusable
- Visualization of high dimensions – grand tour, linked views
- Data Integration Problem: building the "mine": finding and getting the data, dealing with multiple formats, making the measurements commensurate, dealing effectively with computational intensity
- Data mining tools for satellite data validation – statistics research coupled with Earth science—comparing data sets at different spatial scales and temporal interpolations
- Education of the science community in the use of statistics and data mining—Statistics in Geophysics program at NCAR is a good example. Need to do lots of examples and publish them

The Natural Hazards group submitted:

- Everything on-line: rapid access
- End products should be web sites/code/etc., not simply research papers
- Open reference architectures: object models, metadata standards
- Fusion
  -- Automated tools for expressing and optimizing federated queries across disparate datasets: interoperability, extensibility
  -- Combination of evidence
- Scalability
  -- Parallel/distributed
  -- Algorithm scaling with dimensionality
- Algorithm engineering (automation) and development
- Special hardware (onboard)
- Shared software model
- Reproducibility of models

The Terrestrial Ecology group submitted:

- Interoperability of disparate datasets provided by an array of data providers
- Systems that retain error bars, confidence limits, uncertainty; standards for expressing error
- Automated use of appropriate metadata
- Taxonomy of data mining techniques
- Integrated visualization and analysis tools
- Web crawlers, aiding in resource discovery
- Automated correlation detection tools
- How to reduce search space or data set through feature selection
- Data mining that principal investigator has total control over, data mining of NASA data, going beyond NASA to other users
- Need semantic model for each of the databases that are used
- What are the additional expenses on the data provider to support data mining?

4. ANALYSIS OF RESULTS

In this section, the results from each of the theme groups will be analyzed as a whole. The objective of this section is to be able to draw some conclusions about where data mining should go in the future in order to satisfy the needs identified by each of the theme groups. This analysis will not be in terms of specific techniques that are required, since the hope is that data mining practitioners will develop new techniques to answer the needs identified by this workshop. The intent is not to bias this development process by listing specific techniques. The objective is to extract the data mining application areas that were identified by the workshop, some of the constraints that data mining systems must address and some general research and development objectives for pursuing the development of data mining systems. The desire is to stimulate work to develop new data mining techniques and systems that can make a contribution in these application areas.

The words "data mining" mean different things to different people. During the workshop, the theme groups spent some time attempting to come to grips with what data mining meant in the context of the particular theme group. While the workshop as a whole did not arrive at a consensus as to the meaning of data mining, it is useful to have a definition when considering the requirements for data mining in the future. For the purposes of this analysis, the following operational definition will be used:

> *Data mining is the process by which information and knowledge are extracted from a potentially large volume of data using techniques that go beyond a simple search through the data.*

In order to provide some structure for this analysis, data mining is viewed as techniques that process data in one or more dimensions. The spatial dimension is important for earth science data since satellite data has latitude and longitude and, in some cases, altitude. Within the spatial dimension, mining involves properties of data at various locations. An example of this type, presented as part of the atmospheric science case study, is the identification of mesoscale-convective systems (severe storms) using passive microwave data.

In many cases the mining systems need to consider data over time, as the temporal dimension is also important for the earth science domain. Within the temporal dimension, mining involves the processing of data captured at different times. An example of this type of mining, presented as part of the natural hazards case study, is the identification of earthquake faults though the sensing of sub-pixel movement.

In the most general case, mining involves both the spatial and temporal dimensions. An example of this type of mining, provided by the environmental epidemiology case study, is disease surveillance to determine the timing and location of Rift Valley Fever outbreaks.

The remainder of this section is organized into two major subsections. The first presents the data mining application areas, organized in terms of the spatial-temporal mining model. The second major subsection presents the overall data mining research

and development objectives and the associated constraints that such systems must satisfy to address the various application areas.

4.1 Data Mining Application Areas

This section presents application areas where researchers felt that data mining could potentially make a contribution. As the reader will observe, in some cases the precise nature of the data mining contribution was not identified. In these cases, the general science area is included so that data mining practitioners can interact with the science community to determine the specifics of where mining technology could contribute. The first subsection looks at pure spatial mining, the second pure temporal mining, and the third subsection includes mining in both dimensions. Each subsection first provides a brief overview of what has been done to date, and then describes where attendees would like to see data mining contributions in the future.

4.1.1 Mine in the spatial dimension, with time fixed

In the spatial dimension, data mining techniques have been applied to perform land cover mapping, to find eddies and fronts, and to find latitudinal variation in Rossby radius of deformation in ocean data. It has also been used to find various events such as cyclones, fires and mesoscale convective systems. Data mining techniques have also been applied to identify clouds in order to remove their influence, since they are considered to be noise in some research areas.

In the future, data mining could be extended to mine radar data for storms or to mine SeaWiFS data for primary production. Within this dimension, a major future objective to which data mining could contribute is an accurate global characterization of land cover, since current maps are not accurate. An even more complex question to which data mining could contribute is the missing carbon question -- what is its distribution in the terrestrial ecosystem, and what is the distribution in the ocean. What is the net flux?

4.1.2 Mine in the temporal dimension, with a fixed spatial dimension

In the temporal dimension, data mining techniques have been applied to mine for surface changes over time (e.g., earthquake rupture) and used to identify growing season anomalies.

In the future, data mining techniques could be extended to allow for the discrimination between new and old faults and these results used as the basis for a general survey of faults. Also in the earthquake domain, there is the desire to mine time series data for properties of the crust: viscosity, thickness, rigidity, stress, strain and fault geometry using various data sources including GPS, InSAR, seismic, strainmeter, magnetic and topographic analysis.

4.1.3 Mine in the spatial and temporal dimension

This is the most complex mining arena since it involves both the spatial and temporal dimensions. Current mining in this area reported upon at the workshop has been limited to a few examples including disease correlation between SST/AVHRR data and timing and location of Rift Valley Fever outbreaks. There have also been some uses of wavelets to look for trends and surprises in data. The future desires in this area are, however, quite extensive and represent the bulk of the data mining application areas for the future. The various data mining applications have been subdivided into a number of sections that have similar requirements.

4.1.3.1 Exploratory Pattern Mining

In exploratory pattern mining, there is a desire to look for unexpected spatial/temporal patterns and features in existing long-term data that cover a three to five year period. Examples of data that could be mined include ERBE, AVHRR, Landsat, TOMS, SAGE, TOVS, GOES and TOPEX/Poseidon.

4.1.3.2 Relationship Mining

The mining for various types of relationships covers the bulk of the data mining requirements in the spatial/temporal dimension. For presentation purposes, these have been organized into various types of relationships.

Cause or Indicator Relationships:

Cause or indicator relationships include the application of data mining techniques to find the actual causes of an "event", as well as mining to find leading indicators that can be used to predict that an "event" will occur. An example in the cause category is the desire to apply data mining techniques to determine if land clearing has perturbed hydrologic runoff and/or flood frequency. In the indicator category is the desire to apply data mining techniques to identify whether terrestrial change detection can be an indicator of global climate change. At a more general level is the desire to use data mining to identify the signatures or fingerprints that indicate that climate change (e.g., global warming) is occurring.

Effect Relationships:

Effect relationships include the application of data mining techniques to the analysis of synoptic events having regional climate impact (e.g., volcanic eruptions, desert flash floods).

Correlation Relationships:

Research personnel would like to apply data mining techniques in the correlation domain to the study of earthquakes. Of particular interest are the space-time correlations involving triggered slips, foreshocks and aftershocks.

Linkage Relationships:

In linkage relationships, there is the desire to apply data mining techniques to various types of tele-connections in which an event at one particular time and place is related to an event at another time and place. This includes the tele-connection between special events (e.g., El Nino) and various related events, such as an increase of rain in the Midwest or drought or other seasonal changes in particular locations around the world. In the general case, this involves mining for tele-connections among the ocean, atmosphere and land environments. Linkage relationships also include using data mining to investigate linkages among the hydrologic cycle, radiation balance, ocean circulation dynamics, bio-geochemistry of the ocean, and terrestrial ecology. There is a desire to use data mining to find surprises and anomalies in the data that may help with understanding these linkages.

Prediction Relationships:

Prediction relationships include the application of data mining techniques to help researchers predict storm tracts and changes in intensity as storms approach land. Data fusion will be required to understand the microphysics of storms. Prediction relationships also include the use of data mining to identify conditions that will result in droughts, floods or fire potential, to name but three events of interest. At a more complex level is the desire to apply data mining techniques to evaluating and predicting global sustainability.

4.1.3.3 Complex Process Characterization

Researchers would like to perform some complex process characterizations and feel that data mining can make some valuable contributions in this area. The workshop participants did not identify the precise areas in which they felt that data mining could contribute. It is left as an exercise for the data mining community, working with the science community, to determine these precise areas. In complex process characterization, there is a desire:

- to produce global-scale estimates, rates and interpretation of terrestrial ecological attributes (e.g., biomes, NPP, LAI), similar to what has been done with weather, with the ultimate desire to be able to make inferences in this area based on the satellite data. Researchers would like to understand how ocean, atmosphere and land processes are coupled, such as where and when El Nino/La Nina influences land productivity and the role of human influences.
- to develop an Earth System Model by coupling various other models such as terrestrial ecosystem models, ocean models, and atmospheric models. To accomplish this, one must identify the most logical and efficient couplings between these various global-scale models.
- to develop predictive statistical models that can be applied to areas such as seismic activity or the spreading of fire.
- for physical/mechanistic models that can be applied to areas such as earthquake fault modeling, interactions between crust/mantle and stress transfer.
- to be able to perform cross-validation among the models that have been produced.
- for data mining systems to be able to deal with multiple time series from multiple sensors/instruments as these data streams are fused for a single purpose.
- to see if data mining can be used to reduce the output from the models that are constructed, such as the General Circulation Models. This would permit the researcher to concentrate on the important aspects of the model and not be overwhelmed by the huge volume of data produced. In essence, the data mining system would be mining for what is interesting within the massive volume of data produced by these models.

4.1.3.4 Additional Challenge Questions

Some of the data mining application areas did not fall within a specific area and are presented as additional challenges. The following questions represent areas in which data mining techniques may be applicable:

- Is the climate changing?
- What are the signatures of climate change?
- Where does missing carbon go?
- What agricultural areas will be the winners and losers in future global change trends?
- When and where will the global availability of fresh water become a problem?

4.2 Research and Development

This section begins with the constraints that apply to the development of data mining systems and techniques and then outlines the data mining research objectives identified by the various workshop theme groups.

4.2.1 Constraints on Data Mining Techniques

This section looks at the constraints that will affect the development of data mining systems and techniques.

4.2.1.1 Constraints based on the nature of the data

One of the present and future challenges that data mining systems face is the very large size of the data sets that are available to be mined. Newer sensors have higher resolution than previous sensors and the volume of data captured per orbit can be quite large. However, even though the large data set provides a major challenge for data mining system, workshop participants noted that there still exist small data sets that need to be mined as well. The second major challenge is in the diversity of data. Data come from a diverse set of sources and are stored in a wide variety of formats. Unfortunately (for data mining systems), there is no universally accepted data format. In addition, current and anticipated data are very heterogeneous in nature, having a wide variety of different spectral, spatial and temporal characteristics.

There are a number of generic solutions for addressing the problem of mining this diverse collection of data. One approach would use a different mining system, or at least a different reader, for each type of data. The Algorithm Development and Mining (AdaM) system, presented as part of the atmospheric science case study, supports a different type of reader for each type of data. Alternatively, if all data were stored in the same universally accepted data format then a single data mining system could be used. However, at present data are stored in many different formats.

An approach between these two extremes would require machine-readable metadata for each type of data. The data mining system could use this metadata to determine how to mine the data. The utility of this approach would be enhanced by the existence of a universally accepted metadata standard. If all metadata were written to this standard, then the data mining system would have to address only the metadata standard in order to support a universal mining capability.

Until machine-readable metadata are developed to a sufficiently high level to support data mining, it is critical that data have adequate documentation for those intending to mine it. Without adequate documentation, assumptions might be made that could lead to erroneous mining results.

4.2.1.2 Constraints based on search space

Data mining algorithms can involve a large search space as they attempt to derive hidden information from massive volumes of data. Finding ways to reduce the search space would reduce processing time and provide more efficient data mining algorithms and systems.

4.2.1.3 Constraints based on timeliness

Workshop participants noted that while real-time data mining is not important for research, it is highly important for applications. Thus, the amount of effort that one puts into providing a very fast data mining system depends upon one's ultimate objective.

4.2.1.4 Constraints based on need for human interaction

While some data mining can proceed without human involvement, workshop participants noted that there are times in which human influence is desired. In some cases, a human may drive the direction of the data mining.

4.2.1.5 Constraints based on density of relevant data

The data to be mined can be characterized in terms of the density of the data of interest within the whole data set. For example, if one is mining for wild-fires from global-coverage satellite data, the volume of data relevant to the phenomenon of interest is only a very small percentage of the total volume of data being mined due to the spatial sparseness of wild-fires. In contrast, if one is mining time series earthquake data to extract some new rules relating features of the time series, then all of the data being mined are relevant to the phenomena under investigation. In the first case, we have a sparse pattern recognition mining task and in the second case a dense pattern recognition task. These factors may be relevant to the design of a data mining algorithm or system, if for no other reason than in the first case, non-fire data can be discarded immediately when it does not match the fire pattern; in the second case, this cannot be done. The density of the relevant data will have an impact on the amount of state information with which a data mining system must contend.

4.2.1.6 Error Bars

It is important for the mining systems to retain error bars, confidence limits and uncertainty indications applicable to the source data. Further, error of confidence information should be available for the various mining algorithms, so that the compounded error for the resulting mining products can be calculated. There also needs to be a standard for expressing error.

4.2.1.7 Constraints based on false alarm minimization

Data mining systems, as well as any other systems, which could result in an alert of imminent danger, must be concerned about not generating false alerts.

4.2.1.8 Other Issues

It is interesting to note that the mining focus of attention of one group may represent noise to be eliminated by another group. For example, while clouds may be of interest to atmospheric scientists, they are considered noise to those involved in terrestrial ecology research. In this case, the mining results from one group could be used to eliminate noise from the processing of another group.

4.2.2 Data Mining R&D objectives

This section considers data mining R&D objectives in terms of data, mining infrastructure, data mining techniques and visualization associated with data mining.

4.2.2.1 Data: Discovery, Accessibility, Interoperability, Characterization

Data mining R&D objectives begin with the desire to discover relevant data that can be mined for particular purposes. While much of the NASA data are available through their own search portals, there is no central repository for data that could be amenable to mining. Thus, a search engine (Web crawler) that is able to ferret out data that could be of use to those who desire to find the most useful data for their particular purposes. Once the data are found, it would be useful for that data to have associated metadata that could automate the ability of data mining systems to begin to mine the data. It would be desirable for this metadata to include a semantic model oriented toward data mining. Such a semantic model would provide sufficient information such that subtle differences could be detected among similar data. The goal is to permit the data mining system to detect differences relevant to the actual mining of the data or to the interpretation of the results from the mining.

The ultimate goal is to provide mining interoperability of disparate data sets provided by an array of data providers/sensors/instruments. All data should be virtually online to allow the data mining system rapid access to the data. This may require the development of an open reference architecture consisting of object models for the data and associated metadata standards. A major concern is to limit the additional expenses on the data provider to support data mining possibly though extensive use of automation. The nature of that automation is a significant research and development challenge.

4.2.2.2 Mining Infrastructure

The goal of a data-mining infrastructure is to support data integration to facilitate the building of the set of data to be mined using data relevant to the user's objectives. As has been noted previously, this infrastructure must include the necessary tools for finding and obtaining the data, dealing with multiple formats, making the measurements commensurate and dealing effectively with the computational intensity inherent in data mining.

The infrastructure must include the necessary means for algorithm engineering and development. This work includes the development of scalable algorithms for particular mining objectives. Scalability is required due to the large volumes of data that need to be mined and, in some cases, to address the high dimensionality of the data (e.g., hyper-spectral data). The development of appropriate parallel and/or distributed systems and algorithms will be required to provide the desired performance. While the sharing of mining operation software across mining systems is not possible today, it would be desirable for future systems to share software related to specific algorithms, much as class libraries for various programming languages currently permit the sharing of common, useful algorithms.

An important component of data mining is the ability to fuse data from multiple mining activities. This can be supported by the development of automated tools for expressing and optimizing federated queries across disparate data sets. This will require interoperability among data sets, which could be facilitated

through metadata, and the ability to extend the fusion capability to new mining systems and algorithms. This will also require the ability to combine evidence from different sources (data and mining activities) in order to provide a more accurate picture of the results than could be obtained from a single source.

All of these factors point to a need for a data-mining infrastructure that is both flexible and reusable. Also of interest is an environment that allows a user to mine data that may be proprietary to that user in combination with data from the various NASA archives, as well as other data sources and perhaps other mining results. This might even include data that result from on-board processors on satellites. For certain results, the infrastructure needs to support the rapid systematic dissemination of results and warnings, subject of course to the concern previously expressed about the danger of false warnings.

4.2.2.3 Data Mining Techniques

Many of the required data mining techniques must be derived by analyzing the application areas previously described. The identification of application areas to which data mining practitioners could target their research was a primary goal of the workshop. In this section are described a few of the specific issues or techniques that came out of the workshop theme groups.

At the highest level of these issues is the desire for a taxonomy of data mining techniques to help catalog the current state-of-the-art in data mining techniques and indicate where holes need to be filled. At the technology requirement level, there is the desire to couple statistical research and earth science research, much as bio-informatics has coupled statistics and biology. This includes tools for comparing data sets at different spatial scales and temporal interpolations and automated correlation detection tools.

In general, representatives of the science community expressed a desire for the data mining community to provide web sites with mining systems, tools and services that the science community can use, rather than just lists of papers, as is the current situation. On the education front, a desire was expressed to provide the science community with training in the effective use of statistics in data mining. The *Geophysical Statistics Project* at NCAR (http://www.cgd.ucar.edu/stats/) was cited as a good example of what was desired. In addition, there is the desire for lots of published examples of the use of data mining tools and techniques in real science applications. From a programmatic perspective, there is a desire for NASA to structure its NASA Research Announcements to require teaming between science and technology practitioners. This has been done to some extent with the Earth Science Information Partners, but more of this is needed with respect to data mining and science.

4.2.2.4 Visualization and Data Mining

There is a desire to see the integration of visualization and analysis tools as they relate to data mining. This includes the ability to visualize high dimensionality data as well as to provide a grand tour view of the relevant data and to link various views of the data in an effective way.

5. ACKNOWLEDGMENTS

The authors would like the opportunity to thank the UAH staff, the attendees, and NASA for all of their support for this conference. We would also like to thank SIGKDD for the opportunity to discuss these issues in a broader venue.

About the authors:

The report was written by several committee members and collated by Jeanne Behnke. The authors include:

Dr. Sara Graves is Director of the Information Technology and Systems Center and Professor of Computer Science at the University of Alabama in Huntsville. She can be reached at sgraves@itsc.uah.edu.

Dr. Thomas H. Hinke is a Professor of Computer Science at the University of Alabama in Huntsville and is currently working at NASA Ames Research Center on data mining projects while on sabbatical. He can be reached at thinke@mail.arc.nasa.gov.

Elaine Dobinson and **David A. Nichols** are with the Jet Propulsion Laboratory. Their email addresses are elaine.r.dobinson@jpl.nasa.gov and david.a.nichols@jpl.nasa.gov

Jeanne Behnke is with the Earth Science Data Information System Project at NASA Goddard Space Flight Center. She can be reached at jeanne.behnke@gsfc.nasa.gov.