Data Mining Research: Opportunities and Challenges

A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data*

Robert Grossman, Simon Kasif, Reagan Moore, David Rocke, and Jeff Ullman**

January 1999

*These workshops were supported in part by the National Science Foundation under grant IRI-9802160 (PI: Robert Grossman) from the Information and Data Management Program (*http://www.cise.nsf.gov/iis/idm*), grant DMS-9714104 from the Algebra and Number Theory Program (*http://www.nsf.gov/mps/dms/dmsprogs.htm*), and grant DMS-9705599 (PI: David Rocke) from the Statistics and Probability Program (*http://www.nsf.gov/mps/dms/dmsprogs.htm*). All opinions, findings, conclusions and recommendations in any material resulting from these workshops are those of the workshops' participants, and do not necessarily reflect the views of the National Science Foundation. **Robert Grossman (*grossman@uic.edu*) is the editor of this report. Comments for future versions are welcome.

1. Executive Summary

A group of researchers met in Chicago in July, 1997 and in La Jolla in March, 1997 and February, 1998 to discuss the current state of the art of data mining and data intensive computing and the opportunities and challenges for the future. The focus of the discussions was on mining large, massive, and distributed data sets. Here are the main conclusions of the workshops:

- The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications to science, engineering, medicine, business, and education. Data mining attempts to formulate, analyze and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data. Data mining extracts patterns, changes, associations and anomalies from large data sets. Work in data mining ranges from theoretical work on the principles of learning and mathematical representations of data to building advanced engineering systems that perform information filtering on the web, find genes in DNA sequences, help understand trends and anomalies in economics and education, and detect network intrusion. Data mining is also a promising computational paradigm that enhances traditional approaches to discovery and increases the opportunities for breakthroughs in the understanding of complex physical and biological systems. Researchers from many intellectual communities have much to contribute to this field. These include the communities of machine learning, statistics, databases, visualization and graphics, optimization, computational mathematics, and the theory of algorithms.

- The amount of digital data has been exploding during the past decade, while the number of scientists, engineers, and analysts available to analyze the data has been static. To bridge this gap requires the solution of fundamentally new research problems, which can be grouped into the following broad challenges: A) developing algorithms and systems to mine large, massive and high dimensional data sets; B) developing algorithms and systems to mine new types of data; C) developing algorithms, protocols, and other infrastructure to mine distributed data; D) improving the ease of use of data mining systems; and E) developing appropriate privacy and security models for data mining.

- There is an important need for support by government and business of basic, applied, multidisciplinary and interdisciplinary research in data mining and knowledge discovery in order to respond to these challenges.

- There is an important experimental component to data mining and knowledge discovery which requires the creation and maintenance of appropriate systems, computational infrastructures and test beds.

For more information. For more information, please see the M3D2 website at *http://www.ncdm.uic.edu/m3d2.htm*.

2. What is Data Mining?

Data mining is the semi-automatic discovery of patterns, associations, changes, anomalies, rules, and statistically significant structures and events in data. That is, data mining attempts to extract knowledge from data.

Data mining differs from traditional statistics in several ways: formal statistical inference is assumption driven in the sense that a hypothesis is formed and validated against the data. Data mining in contrast is discovery driven in the sense that patterns and hypothesis are automatically extracted from data. Said another way, data mining is data driven, while statistics is human driven. The branch of statistics that data mining resembles most is exploratory data analysis, although this field, like most of the rest of statistics, has been focused on data sets far smaller than most that are the target of data mining researchers.

Data mining also differs from traditional statistics in that sometimes the goal is to extract qualitative models which can easily be translated into logical rules or visual representations; in this sense data mining is human centered and is sometimes coupled with human-computer interfaces research.

Data mining is a step in the data mining process, which is an interactive, semi-automated process which begins with raw data. Results of the data mining process may be insights, rules, or predictive models.

The field of data mining draws upon several roots, including statistics, machine learning, databases, and high performance computing.

In this report, we are primarily concerned with large data sets, massive data sets, and distributed data sets. By large, we mean data sets which are too large to fit into the memory of a single workstation. By massive, we mean data sets which are too large to fit onto the disks of a single workstation or a small cluster of workstations. Instead, massive clusters or tertiary storage such as tape are required. By distributed, we mean data sets which are geographically distributed.

The focus on large data sets is not an just an engineering challenge; it is an essential feature of induction of expressive representations from raw data. It is only by analyzing large data sets that we can produce accurate logical descriptions that can be translated automatically into powerful predictive mechanisms. Otherwise, statistical and machine learning principles suggest the need for substantial user input (specifying meta-knowledge necessary to acquire highly predictive models from small data sets).

3. Recent Research Achievements

The opportunities today in data mining rest solidly on a variety of research achievements, the majority of which were the result of government sponsored research. In this section, we mention a few of the more important ones. Note that several of them are interdisciplinary in nature, resting on discoveries made by researchers from different disciplines working together collaboratively.

Neural Networks. Neural networks are systems inspired by the human brain. A basic example is provided by a back propagation network which consists of input nodes, output nodes, and intermediate nodes called hidden nodes. Initially, the nodes are connected with random weights. During the training, a gradient descent algorithm is used to adjust the weights so that the output nodes correctly classify data presented to the input nodes. The algorithm was invented independently by several groups of researchers.

Tree-based Classifiers. A tree is a convenient way to break a large data sets into smaller ones. By presenting a learning set to the root and asking questions at each interior node, the data at the leaves can often be analyzed very simply. For example, a classifier to predict the likelihood that a credit card transaction is fraudulent may use an interior node to divide a training data set into two sets, depending upon whether or not five or fewer transactions were processed during the previous hour. After a series of such questions, each leaf can be labeled fraud/no-fraud by using a simple majority vote. Tree based classifiers were independently invented in information theory, statistics, pattern recognition and machine learning.

Graphical Models and Hierarchical Probabilistic Representations.
A directed graph is a good means of organizing information about qualitative knowledge about conditional independence and causality gleamed from domain experts. Graphical models generalize Markov models and hidden

Markov models, which have proved themselves to be a powerful modeling tool. Graphical models were independently invented by computational probabilists and artificial intelligence researchers studying uncertainty.

Ensemble Learning. Rather than use data mining to build a single predictive model, it is often better to build a collection or ensemble of models and to combine them, say with a simple, efficient voting strategy. This simple idea has now been applied in a wide variety of contexts and applications. In some circumstances, this technique is known to reduce variance of the predictions and therefore to decrease the overall error of the model.

Linear Algebra. Scaling data mining algorithms often depends critically upon scaling underlying computations in linear algebra. Recent work in parallel algorithms for solving linear system and algorithms for solving sparse linear systems in high dimensions are important for a variety of data mining applications, ranging from text mining to detecting network intrusions.

Large Scale Optimization. Some data mining algorithms can be expressed as large-scale, often non-convex, optimization problems. Recent work has provided parallel and distributed methods for large-scale continuous and discrete optimization problems, including heuristic search methods for problems too large to be solved exactly.

High Performance Computing and Communication. Data mining requires statistically intensive operations on large data sets. These types of computations would not be practical without the emergence of powerful SMP workstations and high performance clusters of workstations supporting protocols for high performance computing such as MPI and MPIO. Distributed data mining can require moving large amounts of data between geographically separated sites, something which is now possible with the emergence of wide area high performance networks.

Databases, Data Warehouses, and Digital Libraries. The most time consuming part of the data mining process is preparing data for data mining. This step can be stream-lined in part if the data is already in a database, data warehouse, or digital library, although mining data across different databases, for example, is still a challenge. Some algorithms, such as association algorithms, are closely connected to databases, while some of the primitive operations being built into tomorrow's data warehouses should prove useful for some data mining applications.

Visualization of Massive Data Sets. Massive data sets, often generated by complex simulation programs, required graphical visualization methods for best comprehension. Recent advances in multi-scale visualization allow the rendering to be done far more quickly and in parallel, making these visualization tasks practical.

4. New Applications

The discipline of data mining is driven in part by new applications which require new capabilities not currently being supplied by today's technology. These new applications can be naturally divided into three broad categories.

a. *Business & E-commerce Data*. Back-office, front-office, and network applications produce large amounts of data about business processes. Using this data for effective decision making remains a fundamental challenge.

b. *Scientific, Engineering & Health Care Data*. Scientific data and meta-data tend to be more complex in structure than business data. In addition, scientists and engineers are making increasing use of simulation and of systems with application domain knowledge.

c. *Web Data*. The data on the web is growing not only in volume but also in complexity. Web data now includes not only text and image, but also streaming data and numerical data.

In this section, we describe several such applications from each category.

Business Transactions. Today, businesses are consolidating and more and more businesses have millions of customers and billions of their transactions. They need to understand risks (Is this transaction fraudulent? Will this customer pay their bills?) and opportunities (What is the expected profit of this customer? What product is this customer most likely to buy next?).

Electronic Commerce. Not only does electronic commerce produce large data sets in which the analysis of marketing patterns and risk patterns is critical, but unlike some of the applications above, it is also important to do this in real or near-real time, in order to meet the demands of on-line transactions.

Genomic Data. Genomic sequencing and mapping efforts have produced a number of databases which are accessible over the web. In addition, there are also a wide variety of other on-line databases, including those containing information about diseases, cellular function, and drugs. Finding relationships between these data sources, which are largely unexplored, is another fundamental data mining challenge. Recently, scalable techniques have been developed for comparing whole genomes.

Sensor Data. Satellites, buoys, balloons, and a variety of other sensors produce voluminous amounts of data about the earth's atmosphere, oceans, and lands. A fundamental challenge is to understand the relationships, including causal relationships amongst this data. For example, do industrial pollutants affect global warming? There are also large terabyte to petabyte data sets being produced by sensors and instruments in other disciplines, such as astronomy, high energy physics, and nuclear physics.

Simulation Data. Simulation is now accepted as a third mode of science, supplementing theory and experiment. Today, not only do experiments produce huge data sets, but so do simulations. Data mining, and more generally data intensive computing, is proving to be a critical link between theory simulation, and experiment.

Health Care Data. Health care has been the most rapidly growing segment of the nation's GDP for some time. Hospitals, health care organizations, insurance companies, and the federal government have large collections of data about patients, their health care problems, the clinical procedures used, their costs, and the outcomes. Understanding relationships in this data is critical for a wide variety of problems, ranging from determining what procedures and clinical protocols are most effective to how best to deliver health care to the most people in an era of diminishing resources.

Multi-media Documents. Few people are satisfied with today's technology for retrieving documents on the web, yet the number of documents and the number of people accessing these documents is growing explosively. In addition, it is becoming easier and easier to archive multi-media data, including audio, images, and video data, but harder and harder to extract meaningful information from the archives as the volume grows.

The Data Web. Today the web is primarily oriented toward documents and their multi-media extensions. HTML has proved itself to be a simple, yet powerful language for supporting this. Tomorrow the potential exists for the web to prove equally important for working with data. The Extensible Markup Language (XML) is an emerging language for working with data in networked environments. As this infrastructure grows, data mining is expected to be a critical enabling technology for the emerging data web.

5. Success Stories

In this section, we briefly describe some success stories involving data mining and knowledge discovery.

Association Rules. Suppose we have a collection of items. The data for many applications consists of multiple transactions, where each transaction consists of one or more items. A basic example is provided by a supermarket, where the items are the products offered for sale and the transactions are purchases, consisting of one or more products purchased by an individual at a given time. A fundamental problem is to uncover associations: which products tend to be purchased together. There has been a lot of recent work on this problem and a variety of algorithms have been developed which can discover associations, even in very large data sets, with just a few passes over the data. A variety of commercial data mining systems support association rules and they are now routinely applied to a range of problems from database marketing to product placement for supermarkets. In addition, association rules algorithms have spurred new research in a variety of areas from databases to complexity theory.

Fraud Detection. Although relatively few credit card transactions are fraudulent, the sheer volume of transactions means that over $500M are lost each year in this way. A variety of data mining techniques have been used to develop fraud systems which can detect fraudulent credit card transactions in near-real time. This problem is challenging due to the size of the data sets, the rarity of the events of interest, and the performance requirements for near-real time detection. Data mining has also improved fraud detection in other application areas, including telecom fraud and insurance fraud.

Astronomical Data. Traditionally, the search for new galaxies, stars, and quasars has primarily been done by astronomers visually examining individual photographic plates. Classification algorithms from data mining have recently been used to automate this process yielding new astronomical discoveries. The classification algorithms are applied to derived attributes produced by image processing, such as the brightness, area, and morphology of sky objects. The approach has also proved useful for detecting new objects too faint to be observed by a manual analysis or traditional computational techniques. For the 2nd Palomar Observatory Sky Survey, this approach resulted in over a three-fold increase in the size of the catalog.

Genomic Data. Genomic data is stored all over the world, in a variety of formats and managed by a variety of applications and systems. Recently, systems have been developed which allow discoveries to be made involving information distributed over several systems. In particular, the new systems have enabled for the first time whole genome comparison, gene identification, and whole genome functional interpretation and analysis. The techniques developed for analyzing genomic and other types of scientific data can be expected to play a role in analyzing a broad range of biological data.

Distributed Data Mining. Traditionally, data mining has required that the relevant data be warehoused in a single location. Recently, distributed data mining systems have exploited wide area, high performance next networks, such as the NSF vBNS network, to mine large amounts of distributed scientific and health care data. Recently, these systems have set records for the sustained movement of very large amounts of data over wide area networks. Separately, a prototype has been developed exploiting distributed data mining to improve the detection of credit card fraud.

Text Mining. Recently, data mining has been combined with algorithms from information retrieval to improve the precision and recall of queries on very large collections of documents. In particular, some of these algorithms have proved useful on multi-lingual collections and others have shown their worth on querying using concepts instead of key words.

6. Trends that Effect Data Mining

In this section, we describe five external trends which promise to have a fundamental impact on data mining.

Data Trends. Perhaps the most fundamental external trend is the explosion of digital data during the past two decades. During this period, the amount of data probably has grown between six to ten orders of magnitude. Much of this data is accessible via networks. On the other hand, during this same period the number of scientists, engineers, and other analysts available to analyze this data has remained relatively constant. For example, the number of new Ph.D.'s in statistics graduating each year has remained relatively constant during this period. Only one conclusion is possible: either most of the data is destined to be write-only, or techniques, such as data mining, must be developed, which can automate, in part, the analysis of this data, filter irrelevant information, and extract meaningful knowledge.

Hardware Trends. Data mining requires numerically and statistically intensive computations on large data sets. The increasing memory and processing speed of workstations enables the mining of data sets using current algorithms and techniques that were too large to be mined just a few years ago. In addition, the commoditization of high performance computing through SMP workstations and high performance workstation clusters enables attacking data mining problems that were accessible using only the largest supercomputers a few years ago.

Network Trends. The next generation internet (NGI) will connect sites at OC-3 (155 MBits/sec) speeds and higher. This is over 100 times faster than the connectivity provided by current networks. With this type of connectivity, it becomes possible to correlate distributed data sets using current algorithms and techniques. In addition, new protocols, algorithms, and languages are being developed to facilitate distributed data mining using current and next generation networks.

Scientific Computing Trends. As mentioned above, scientists and engineers today view simulation as a third mode of science. Data mining and knowledge discovery serves an important role linking the three modes of science: theory, experiment and simulation, especially for those cases in which the experiment or simulation results in large data sets.

Business Trends. Today businesses must be more profitable, react quicker, and offer higher quality services than ever before, and do it all using fewer people and at lower cost. With these types of expectations and constraints, data mining becomes a fundamental technology, enabling businesses to more accurately predict opportunities and risks generated by their customers and their customers' transactions.

7. Research Challenges

In this section, we describe some of the major research challenges identified by the three workshops. The research challenges are arranged into five broad areas: A) improving the scalability of data mining algorithms, B) mining non-vector data, C) mining distributed data, D) improving the ease of use of data mining systems and environments, and E) privacy and security issues for data mining.

A. Scaling data mining algorithms. Most data mining algorithms today assume that the data fits into memory. Although success on large data sets is often claimed, usually this is the result of sampling large data sets until they fit into memory. A fundamental challenge is to scale data mining algorithms as

1. the number of records or observations increases;
2. the number of attributes per observation increases;
3. the number of predictive models or rule sets used to analyze a collection of observations increases;
4. and, as the demand for interactivity and real-time response increases.

Not only must distributed, parallel, and out-of-memory versions of current data mining algorithms be developed, but genuinely new algorithms are required. For example, association algorithms today can analyze out-of-memory data with one or two passes, while requiring only some auxiliary data be kept in memory.

B. Extending data mining algorithms to new data types. Today, most data mining algorithms work with vector-valued data. It is an important challenge to extend data mining algorithms to work with other data types, including 1) time series and process data, 2) unstructured data, such as text, 3) semi-structured data, such as HTML and XML documents, 4) multi-media and collaborative data, 5) hierarchical and multi-scale data, and 6) and collection-valued data.

C. Developing distributed data mining algorithms. Today most data mining algorithms require bringing together all data to be mined in a single, centralized data warehouse. A fundamental challenge is to develop distributed versions of data mining algorithms so that data mining can be done while leaving some of the data in place. In addition, appropriate protocols, languages, and network services are required for mining distributed data to handle the meta-data and mappings required for mining distributed data. As wireless and pervasive computing environments become more common, algorithms and systems for mining the data produced by these types of systems must also be developed.

D. Ease of Use. Data mining today is at best a semi-automated process and perhaps destined to always remain so. On the other hand, a fundamental challenge is to develop data mining systems which are easier to use, even by casual users. Relevant techniques include improving user interface, supporting casual browsing and visualization of massive and distributed data sets, developing techniques and systems to manage the meta-data required for data mining, and developing appropriate languages and protocols for providing casual access to data. In addition, the development of data mining and knowledge discovery *environments* which address the *process* of collecting, processing, mining, and visualizing data, as well as the collaborative and reporting aspects necessary when working with data and information derived from it, is another important fundamental challenge.

E. Privacy and Security. Data mining can be a powerful means of extracting useful information from data. As more and more digital data becomes available, the potential for misuse of data mining grows. A fundamental challenge is to develop privacy and security models and protocols appropriate for data mining and to ensure that next generation data mining systems are designed from the ground up to employ these models and protocols.

8. Testbeds and Infrastructure

Experimental studies will play a critical role in advancing the field of data mining. Developed testbeds for high performance and distributed data mining is essential for progress in the field.

The requirements for data mining testbeds are different than those for general purpose high performance computing testbeds. For example, the computing resources for data mining testbeds are as much disk-oriented as processor-oriented; the network resources must be able move data sets and data elements between geographically distributed sites with guaranteed quality of service; and a variety of general purpose and specialized data mining software must be available.

Perhaps the two most difficult challenges in creating data mining testbeds and national resources in data mining are assembling a) the appropriate data sets and b) the required interdisciplinary and multidisciplinary teams.

9. Findings and Recommendations

In this section, we list some of the major findings of these three workshops.

For all interested parties:

Data mining and knowledge discovery is a new emerging discipline with both a scientific and engineering component that is of strategic importance for the U.S. and of critical importance to future information access technologies. All interested parties are encouraged to work towards the maturation of data mining and knowledge discovery, towards its establishment as a scientific and engineering discipline in its own right, and towards the evolution of a community that includes the relevant traditions and disciplines and puts them together in the proper context.

For the federal government:

1. Create programs which encourage the emergence of data mining and knowledge discovery as an independent discipline.

2. Support interdisciplinary and multidisciplinary research projects. Many advances in data mining require teams of mathematicians and statisticians, computer scientists, and application domain scientists working together to create the appropriate data sets and the required algorithms and software to analyze them.

3. Support basic research in computer and information sciences that underlies data mining, including machine learning, knowledge systems, data bases, high performance computing, high performance networking, and digital libraries.

4. Support basic research in mathematics and statistics that underlies data mining, including statistics, probability, applied mathematics, logic, discrete mathematics, analysis and dynamical systems, linear algebra, and computational geometry and algebra.

5. Support data mining testbeds. The hardware, software, data and consulting requirements for data mining often out-strip the resources of individual scientists and small research groups. Supporting national resources and testbeds for data mining is important in order to provide the proper experimental infrastructure required for next generation data mining experiments.

For companies:

1. Support applied research in data mining.
2. Work to develop, implement and support appropriate privacy and security models for data mining systems.
3. Create sanitized versions of real data sets for use by data mining researchers.
4. Support joint research projects between industry and universities. Support collaborative testbeds and demonstration projects.

For scientists and engineers:

1. As new data is collected and archived, support emerging protocols, languages, and standards to facilitate the future analysis and mining of the data, especially by scientists and engineers from other disciplines.

2. As new data is collected and as new systems are built to manage it, ensure that the best available privacy and security models are used to protect inadvertent disclosures of private information.

3. Provide long-term maintenance and access to data sets created by scientists and engineers, as well as to the knowledge and information extracted from them.

10. Conclusions

Data mining and knowledge discovery are emerging as a new discipline with important applications to science, engineering, health care, education, and business. Data mining rests firmly on 1) research advances obtained during the past two decades in a variety of areas and 2) more recent technological advances in computing, networking and sensors. Data mining is driven by the explosion of digital data and the scarcity of scientists, engineers, and domain experts available to analyze it.

Data mining is beginning to contribute research advances of its own, by providing scalable extensions and advances to work in associations, ensemble learning, graphical models, techniques for on-line discovery, and algorithms for the exploration of massive and distributed data sets.

Advances in data mining requires a) supporting single investigators working in data mining and the underlying research domains supporting data mining; b) supporting inter-disciplinary and multi-disciplinary research groups working on important basic and applied data mining problems; and c) supporting the appropriate testbeds for mining large, massive and distributed data sets.

Appropriate privacy and security models for data mining must be developed and implemented.

11. References

References for the material above can be found in the Supplement to this report.

12. Acknowledgements

The editors would like to thank Usama Fayyad for his comments on an earlier draft of this report.

Robert Grossman is at the National Center for Data Mining at the University of Illinois at Chicago and Magnify, Inc. Simon Kasif is at the National Center for Data Mining and the Department of Electrical Engineering and Computer Science at the University of Illinois at Chicago. Reagan Moore is at the San Diego Supercomputer Center. David Rocke is at the Center for Image Processing and Integrated Computing at the University of California at Davis. Jeff Ullman is at the Department of Computer Science at Stanford University.

Appendix A. M3D-97, Chicago

This M3D-97 Workshop took place on July 12-15, 1997 in Chicago and was supported by NSF grant DMS-9714104 from the Algebra and Number Theory Program. Listed below are the speakers and the titles of their talks.

1. **Michael Berry**, University of Tennessee, "Dynamic Information Management using Latent Semantic Indexing"
2. **Herbert Edelsbrunner**, University of Illinois at Urbana Champaign, "Constructing Shapes from Points in Dimension Three and Beyond"
3. **John Elder**, ELDER Research, "Fusing Diverse Algorithms"
4. **Christos Faloutsos**, University of Maryland at College Park, "Applications, Requirements and Database Tools for Massive Data Mining"
5. **Usama Fayyad**, Microsoft, "Data Mining and KDD: So What's New?"
6. **Robert Grossman**, University of Illinois at Chicago &amp; Magnify, Inc. "Dynamic Similarity: Mining Collections of Trajectories"
7. **Peter Jones**, Yale University, "On the Structure of Low Dimensional Sets"
8. **Michael Jordan**, MIT, "Graphical models and variational approximation"
9. **Simon Kasif**, University of Illinois at Chicago, "A Computational Framework for Data Mining: Data Structures and Algorithms for Efficient Probabilistic Inference"
10. **Heikki Mannila**, University of Helsinki, "Association rules, episode rules and frequent sets: algorithms and applications"
11. **Vince Poor**, Princeton University, "Quickest Detection: Time Optimal Methods for Statistical Change Detection"
12. **J. Ross Quinlan**, University of Sydney, "Tree-Based Classifiers and Extensions"
13. **Eric Ristad**, Princeton University, Maximum Entropy Modeling for Discrete Domains"
14. **Dan Roth**, Weizman Institute of Science, "Learning and Managing Knowledge in Large Scale Natural Language Inferences"
15. **Stuart Russell**, University of California at Berkeley, "Adaptive Probabilistic Networks"
16. **Fred Warner**, Yale University, "Adapted Waveform Analysis as a Tool for Data Transcription, Rudimentary Modeling and Feature Extraction"

Appendix B. M3D2-98, La Jolla

The M3D2-98 workshop took place on February 5-6, 1998 in La Jolla, California and was supported by NSF grant IRI-9802160 from the Information and Data Management IDM) Program. The individuals listed below participated in workshop. Discussions took place in two break out groups: 1) Research Issues and Fundamental Challenges, and 2) Testbeds and Infrastructure.

Stuart Bailey, University of Illinois at Chicago
Scott Baden, University of California, San Diego
Chaitanya Baru, San Diego Supercomputing Center
Don Benton, University of Pennsylvania
Peter Buneman, University of Pennsylvania
Alok Choudhary, Northwestern University
Thomas Dietterich, Oregon State University
Marcio Faerman, University of California, San Diego
Robert Grossman, University of Illinois at Chicago and Magnify
Bob Hollebeek, University of Pennsylvania
Chandrik Kamath, Lawrence Livermore National Laboratory
Carl Kesselman, University of Southern California
Reagan Moore, San Diego Supercomputing Center
Ron Musick, Lawrence Livermore National Laboratory
Pavlos Protopapas, University of Pennsylvania
Arcot Rajasekar, San Diego Supercomputing Center
David Rocke, University of California, Davis
Joel Saltz, University of Maryland
Vivek Sharma, University of California, San Diego
Terry Smith, University of California, Santa Barbara
Padhraic Smyth, University of California, Irvine
Paul Stolorz, Jet Propulsion Laboratory
Jeff Ullman, Stanford University
Roy Williams, Caltech
Maria Zemankova, NSF

The following talks were given:
1. **Robert Grossman**, University of Illinois at Chicago and Magnify, Inc. & Reagan Moore, San Diego Supercomputing Center, "Managing and Mining Massive Data Sets: Introduction."
2. **Thomas Dietterich**, Oregon State University, "Scaling Up Machine Learning for Data Mining Applications."
3. **Carl Kesselman**, University of Southern California, "An Overview of Infrastructures for Wide Area High Performance Computation."
4. **Jeff Ullman**, Stanford University, "Association-Rule Mining."
5. **Padhraic Smyth**, University of California -Irvine &amp; Jet Propulsion Laboratory, "Fundamental Challenges in Data Mining."

Appendix C. Approaches to the Analysis and Visualization of Massive Data Sets, La Jolla

The Data Mining and Visualization Workshop was held on March 14-15, 1997 in La Jolla and was supported by NSF grant DMS-9705599 from the Statistics and Probability Program. Listed below are the speakers and the titles of their talks.

1. **William Eddy**, Carnegie Mellon University, "Interaction with Massive Data Sets via an Index"
2. **Dan Gusfield**, University of California at Davis, "Extracting the Essence From Collections of Molecular Sequences; a Problem Statement"
3. **Bernd Hamann**, University of California at Davis, "Issues Regarding the Hierarchical Representation of Large Data Sets for Analysis and Visualization"
4. **Ken Joy**, University of California at Davis, "Hierarchical Reduction Methods for Massive Data Sets"
5. **Giovanni Marchisio**, Mathsoft, Inc., "Document-term Matrix Decomposition for Intelligent Information Retrieval"
6. **Nelson Max**, Lawrence Livermore National Laboratory, "Visualizing Global Climate and Finite Element Simulations"
7. **Reagan Moore**, San Diego Supercomputer Center, "Information Based Computing"
8. **Robert Moorhead**, Mississippi State University, "Visualization of Air/Sea Model Data"
9. **Emanuel Parzen**, Texas A&M University, "Comparison Density and Quantile Statistical Methods and Massive Data Set Analysis"
10. **Tom Prince**, California Institute of Technology, "Digital Sky"
11. **Jim Quinn**, University of California at Davis, "Strategies for Integrating Interagency Data on Biodiversity and Water Issues"
12. **David Rocke**, University of California at Davis, "Partitioning and Subsampling to Uncover Subtle Structure in Massive Data Sets"
13. **Hanan Samet**, University of Maryland, "Sorting in Space"
14. **David Scott**, Rice University, "Statistics and Massive Data Sets: A New Look at the Method of Moments and Maximum Likelihood"
15. **David Shanno**, Rutgers University, "Topics in Very Large Scale Optimization"
16. **Elizabeth Thompson**, University of Washington, "Monte Carlo Likelihood in Some Problems of Genetic Analysis"
17. **Ed Wegman**, George Mason University, "Thoughts on Statistical Data Mining"
18. **David Woodruff**, University of California at Davis, "Heuristic Search Applications"