An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications

Gregory Piatetsky-Shapiro
GTE Laboratories
40 Sylvan Road
Waltham, MA 02154
gps@gte.com

Ron Brachman
AT&T Research
600 Mountain Avenue
Murray Hill, NJ 07974
rjb@research.att.com

Tom Khabaza
ISL
Berk House, Basing View
Basingstoke RG21 4RG
UK, tomk@isl.co.uk

Willi Kloesgen
GMD, D-53757
Sankt Augustin, Germany
kloesgen@gmd.de

Evangelos Simoudis
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
simoudis@almaden.ibm.com

Abstract

This paper surveys the growing number of industrial applications of data mining and knowledge discovery. We look at the existing tools, describe some representative applications, and discuss the major issues and problems for building and deploying successful applications and their adoption by business users. Finally, we examine how to assess the potential of a knowledge discovery application.

1 Introduction

A significant need exists for a new generation of techniques and tools with the ability to *intelligently* and *automatically* assist humans in analyzing the mountains of data for nuggets of useful knowledge. These techniques and tools are the subject of the emerging field of data mining and knowledge discovery in databases (Fayyad et al. 1996). This paper examines a growing number of industrial applications in this field (see Fayyad, Haussler, and Stolorz 1996 for a survey of scientific applications). We survey the existing data mining tools, describe some representative applications, and discuss the major issues and problems for building and deploying successful applications.

Knowledge Discovery in Databases (KDD) is an umbrella term used to describe a large variety of activities for making sense of data. We will use the term *knowledge discovery* to describe the overall process of finding useful patterns in data, which includes not only the data mining step of running the discovery algorithms, but also pre- and post-processing, and various other activities (see Section 3).

The knowledge discovery *goals* are defined by the intended use of the system. We can distinguish two types of goals: **verification**, where the system is limited to verifying the user's hypothesis (the most prevalent type of data analysis to date), and **discovery**, where the system autonomously finds new patterns. We further subdivide the Discovery goal into **prediction**, where the system finds patterns for the purpose of predicting the future behaviour of some entities; and **description**, where the system finds patterns for the purpose of presenting them to a user in a human-understandable form. Although the boundaries between prediction and description are not sharp – some of the predictive models can be descriptive (to the degree that they are understandable), and some of the descriptive models could be used for prediction – this distinction is useful for understanding the discovery goal.

The framework for knowledge discovery and the data mining tasks of classification, regression, clustering, summarization, dependency modeling, and deviation detection are discussed in more detail in (Fayyad, Piatetsky-Shapiro, & Smyth 1996).

2 Data Mining Tools

Data mining in the industry today is still primarily verification-oriented and performed mainly by analysts whose primary training and professional duties are in statistics or data analysis. By and large, the tools used are not expressly "data mining tools", but rather statistical analysis tools like S and SAS, graph and chart-drawing software and spreadsheets, and database query engines. Direct programming in languages like C and awk is typically used for complex analyses, usually on data selected from a database, but dumped into a flat file for further manipulation.

Among the tools that support data mining[1] – see `<http://info.gte.com/~kdd/siftware.html>` for a catalog, the first group is **generic, single-task tools**. There are many dozens of such tools available, especially for classification, using primarily decision trees, neural networks, example-based, and rule-discovery approaches. Such tools mainly

[1] sometimes called siftware because this is software for *sifting* through the data

support only the data mining step in the knowledge discovery process and require significant pre- and post-processing. The target user of such tools is typically a consultant or a developer who would integrate them with other modules as part of a complete application.

The next group is **generic, multi-task tools.** These tools perform a variety of discovery tasks, typically combining classification (perhaps using more than one approach), visualization, query/retrieval, clustering, and more. They include Clementine, Darwin, IBM Intelligent Miner, IMACS, MLC++, MOBAL, and SGI MineSet.

These tools support more of the KDD Process and simplify embedding of discovered knowledge in the application. Usually, the target user of such tools is a "power" analyst who understands data manipulation. Generally such tools require appropriate training and some tool customization before being used by domain experts, but there are some exceptions.

Clementine, in particular, is reported to have been widely used without customization by a variety of end-users ranging from business analysts to biochemists. This is made possible by Clementine's highly graphical user interface to data mining functions.

Another example is the IMACS system (Brachman et al. 1993), which used a knowledge representation system to represent domain and task objects and integrate various aspects of the discovery process. IMACS allowed the user to create new, object-centered views over data that was stored in arbitrary ways (relational database, flat files). Data could be segmented in a simple way using these views, and new segments could be defined easily from old ones. This was one big step towards allowing a business user to interact with data in terms s/he was familiar with, since the views (or "concepts") were expressed entirely in the user's terms rather than being slaved to database schemas created for other purposes.

Finally, the last group is **domain-specific tools.** These tools support discovery only in a specific domain and already talk the language of the end user, who needs to know very little about the analysis process. Examples of such tools include Opportunity Explorer (Anand 1995) which generates reports on changes in retail sales, IBM Advanced Scout `<http://www.research.ibm.com/xw-scout>` which analyzes basketball game statistics and finds patterns of play that coaches can use, and HNC Falcon `<http://www.hnc.com/>`, a neural network-based system for credit-card fraud detection.

These systems and others represent a growing trend in moving data mining technology into the business world. The key elements that help make the core statistical, machine learning, and other data mining technologies accessible to a mainstream user include

- putting the problem in the business user's terms, including viewing the data from a business model perspective (both concepts and rules);
- support for specific key business analyses like segmentation;
- presentation of results in a form geared to the business problem being solved; and
- support for an iterative exploratory process protracted in time, as examined in the next section.

3 Knowledge Discovery Process

The core of the knowledge discovery process is the set of data mining tasks, used to extract and verify patterns in data. However, we should emphasize that this core takes only a small part (estimated as 15 to 25% of the overall effort) of the entire process of knowledge discovery. No complete methodology for this process yet exists, but knowledge discovery takes place in a number of stages (more about this in Brachman & Anand, 1996):

- Data and task discovery - the process of becoming familiar with both the data that will be analyzed and the task that the business user needs to accomplish. This is more significant than it may sound, especially when the data is to be pulled from multiple sources, and when the analysis will not be done by the business user;
- Acquisition - bringing the data into the appropriate environment for analysis;
- Integration and checking - confirming the expected form and broad contents of the data, and integrating into tools as required;
- Data cleaning - removing records with errors or outliers (if considered insignificant), etc.; looking for obvious flaws in the data and removing them;
- Model and hypothesis development - simple exploration of the data by passive techniques, and elaboration by deriving new data attributes where necessary; selection of an appropriate model in which to do analysis; development of initial hypotheses to test;
- Data mining step - application of the core discovery procedures to discover patterns and new knowledge, or to verify hypotheses developed prior to this step;
- Testing and verification - assessing the discovered knowledge: testing predictive models on test sets, analyzing segmentation etc.;
- Interpretation and use - integration with existing domain knowledge, which may confirm, deny or challenge the newly discovered patterns; if predictive, subsequent use on novel data sets.

Throughout the process, we also have presentation or visualization of results as an integral activity. A key thing to note about this process is that it is not simple and linear, but thoroughly iterative and interactive, the results of analysis being fed back into the modeling and hypothesis derivation process to produce improved results on subsequent iterations. This activity takes time, and if it is applied to data generated on a regular basis (e.g., quarterly or yearly results) can have a very long lifespan.

4 Representative Applications

Numerous knowledge discovery applications and prototypes have been developed for a wide variety of domains including marketing, finance, manufacturing, banking, and telecommunications. A majority of the applications have used predictive modeling approach, but there were also a few notable applications using other methods. Here we describe some of the representative examples.

4.1 Marketing

Leading market research companies such as A.C. Nielsen and Information Resources in USA, GfK and Infratest Burke in Europe apply KDD tools to the rapidly growing sales and marketing databases. Because of a strong competitive pressure, the often saturated market potential and maturity of products, there is a shift from a quality to an information competition where detailed and comprehensive knowledge on the behavior of customers and competitors is crucial.

Market research companies collect data on special markets, analyze this data and sell data and analyses to their clients. The clients add their own data for further analyses. Medium sized datasets are captured when market research companies perform surveys (e.g. 2000 persons interviewed each month) or organize test samples of households. BehaviorScan approaches provide test households with special alternative TV commercials. Much larger data sets are available in the form of point of sale data, when e.g. purchases in supermarkets are captured by scanners.

Marketing, which has been a long time user of statistical and other quantitative methods, has been in the forefront of adopting new knowledge discovery techniques. Most marketing applications fall into the broad area called "Database Marketing" (or "mailshot response" in Europe). This is an approach which relies on analysis of customer databases, using a number of techniques including interactive querying, market segmentation to identify different customer groups, and predictive modeling to forecast their behaviour. Business Week (Berry 1994) has estimated that over half of all retailers are using or planning to use database marketing, and those who do use it have good results such as 10-15% increase in credit card use reported by American Express.

An interesting application to predict television audiences using neural networks and rule induction was developed by Integral Solutions for the BBC. Rule induction was used to examine which factors play the most important role in relating the size of a program's audience to its scheduling slot. The final models were equivalent to the best performance of human experts, but highly robust against change, because the models could be retrained from up-to-date data (Fitzsimons, Khabaza, & Shearer 1993).

Other applications are more descriptive – their focus is to find patterns that will help market analysts make better decisions. Among the first systems developed and deployed in this area were Coverstory (Schmitz, Armstrong, & Little 1990) and Spotlight (Anand & Kahn 1992), which analyzed supermarket sales data and produced reports, using natural language and business graphics, on the most significant changes in a particular product volume and share broken down by region, product type, and other dimensions. In addition, causal factors such as distribution channels and price changes were examined and related to changes in volume and share. These systems were quite successful – Spotlight was reported to be among the best selling products of A.C. Nielsen.

Spotlight was later extended into Opportunity Explorer system (Anand 1995), which supports the sales representative of a consumer packaged good company in examining the business with a single retailer. This is accomplished by presentations that highlight the advantages for the retailer if additional products are stocked or special promotions are performed. A new feature of Opportunity Explorer was generation of interactive reports with hyperlinks (even before the Web!), which allowed easy navigation between different report sections.

The MDT (Management Discovery Tool) system, a product under development at AT&T and NCR, incorporates several other innovative ideas to allow a business person to directly interface with data. MDT incorporates a set of business rules (encoded as metadata) that make it easy to set up monitors that detect significant deviations in key business indicators. To accommodate the mainstream business user, MDT provides a limited set of analysis types, including summarization, trend analysis, and measure and segment comparison.

Another marketing area is Market basket analysis, which looks at associations between different products bought by the customer. These methods are generally based on the association discovery algorithms (Agrawal et al. 1996). A number of companies, including IBM and SGI offers tools for Market basket analysis.

4.2 Investment

Many financial analysis applications employ predictive modeling techniques, such as statistical regression or neural networks, for tasks like portfolio creation and optimization and trading model creation. Such applications have been in use for several years. To maintain a competitive advantage, the users and developers of such applications rarely publicize their exact details and effectiveness.

We can, however, mention a few examples. Fidelity Stock Selector fund is using neural network models to select investments and has performed quite well until recently. However the output of those models is evaluated by the fund manager Brad Lewis before the action is taken, so it is not clear how to divide the credit between man and machine.

Morgan Stanley and Co. has developed AI (Automated Investor) system which identifies good trading opportunities by using clustering, visualization, and prediction. The system has been deployed and is being evaluated.

Daiwa Securities used MATLAB toolkit to build a portfolio management tool which analyzes a large number of stocks and selects an optimal portfolio based on the stock risk and expected rate of return (Pittaras 96).

LBS Capital Management uses expert systems, neural nets and genetic algorithms to manage portfolios totalling $600 million and since its start in 1993, their system has outperformed the broad stock market (Hall, Mani, & Barr 1996).

Carlberg & Associates have developed a neural network model for predicting S&P 500 Index, <http://carlassoc.com/> using interest rates, earnings, dividends, the dollar index, and oil prices. The model was surprisingly successful and explained 96% of the variation in the S&P 500 index from 1986 to 1995.

In these applications, predictive accuracy is paramount compared to the ability to use the extracted knowledge to explain a recommended action. Thus, the main focus is ensuring that modeling methods do not overfit the data.

4.3 Fraud Detection

Not all the systems developed for this have been publicized, for obvious reasons, but several are worth mentioning.

The HNC Falcon credit risk assessment system, developed using a neural network shell, is used by a large percentage of retail banks to detect suspicious credit card transactions. Falcon deployment was facilitated by the fact that credit card transaction data is captured by just a few companies. Even though each such company uses its own data format, every bank issuing credit cards uses one of these few formats. Therefore, an application that can work with even one format effectively can easily be adopted by a large number of banks.

The FAIS system (Senator et al. 1995) from US Treasury's Financial Crimes Enforcement Network, is used to identify financial transactions that may be indicative of money laundering activity. FAIS uses data from common government forms and consists of a combination of off-the-shelf and custom built components. Its use is expected to expand to a variety of government agencies that are concerned with the detection of suspicious financial transactions indicative of money laundering operations. FAIS has the hardest data quality problem because much of its data comes from poorly handwritten notes.

AT&T has developed a system for detecting the international calling fraud by displaying the calling activity in a way that lets the user quickly see the unusual patterns (Eick & Fyock 1996).

The Clonedetector system, developed by GTE (Davis & Goyal 1993), is using customer profiles to detect the cellular cloning fraud. If a particular customer suddenly starts calling in a very different way, fraud alert automatically kicks in.

Another cellular fraud detection system is under development at NYNEX. The developers first mine the data to discover indicators of fraudulent usage. Subsequently, they automatically generate detection systems by feeding these indicators into a detector constructor program, which uses the indicators to instantiate detector templates. Finally, the system learns how to combine the detectors for optimal performance.

5 Manufacturing and Production

Controlling and scheduling technical production processes is a an application of KDD with a high potential profit. The goal is to discover process conditions that lead to good quality products. At present, large volumes of data generated during a production process are often only poorly exploited. Also, the relations between the control, process, and quality variables are not completely understood by the engineers. In addition, time and space constraints, which play an especially important role in manufacturing, are not well handled by most data mining tools.

A typical example is a project which is run in a large chemical company in Europe to analyze a production process in a plant for polymeric plastics. Data includes control variables (e.g. quantities of raw material, the heating parameters), the process variables (temperatures, pressures, and chemical reaction times), and quality variables measured in a laboratory. Quality variables are determined several times a day, process and control variables nearly continuously. Even simple approach of introducing separate variables for distinct time points (process variables measured every hour), and applying rule-inducing discovery methods to the resulting data can lead to valuable insights into the manufacturing process.

Another example is the CASSIOPEE troubleshooting system, developed by a joint venture of General Electric and SNECMA using the KATE discovery tool. The system is applied by three major European airlines to diagnose and predict problems for BOEING 737. To derive families of faults, clustering methods are used. CASSIOPEE received the European first prize for innovative applications (Manago and Auriol 96).

5.1 Telecommunication

Another application area involving a strong time component is the management of telecommunication networks. These large and complex networks produce large amounts of alarms daily. The sequence of alarms contains valuable knowledge about the behavior of the network. Regularities in the alarms can be used in fault management systems for filtering redundant alarms, locating problems in the network, and predicting severe faults.

At the University of Helsinki, the Telecommunication Alarm Sequence Analyzer (TASA) was built in cooperation with a manufacturer of telecommunication equipment and three telephone networks (Mannila, Toivonen, & Verkamo 1995). The system uses a novel framework for locating frequently occurring alarm episodes from the alarm stream and presenting them as rules. Large sets of discovered rules can be explored with flexible information retrieval tools supporting interactivity and iteration. In this way, TASA offers pruning, grouping, and ordering tools to refine the results of a basic brute force search for rules. The system has discovered rules that have been integrated into the alarm handling software of the telephone networks.

5.2 Other Areas

Health care is an information-rich and high payoff area, ripe for data mining. One of the first applications in this area is KEFIR (Matheus, Piatetsky-Shapiro, & McNeill 1996). The system performs an automatic drill-down through data along multiple dimensions to determine the most interesting deviations of specific quantitative measures relative to their previous and expected values. It explains "key" deviations through their relationship to other deviations in the data, and, where appropriate, generates recommendations for actions in response to these deviations. KEFIR uses a Web browser to present its findings in a hypertext report, using natural language and business graphics.

Improving data quality is another important application area. One aspect of data quality is the automatic detection of errors. A number of applications were developed for checking data (in particular financial trading data), detecting errors currently impossible to detect by conventional means. Another aspect of data quality is the identification of related and duplicate entities – an especially acute problem for database marketers and catalog senders. Identification of duplicate claims was performed by Merge/Purge system (Hernandez & Stolfo 1995), successfully used on data from Welfare Department of the State of Washington.

Basketball statistics are also plentiful, and IBM Advanced Scout helps NBA coaches and league officials organize and interpret the data amassed at every game. A sample finding from a Jan 6, 1995 game between Knicks and Cavaliers is that when Mark Price played the 1Guard position, John Williams attempted four jump shots and made each one. Advanced Scout not only finds this pattern, but explains that it is interesting because it differs considerably from the average shooting percentage of 49.30% for the Cavaliers during that game. This is the kind of pattern that coaches might not ordinarily detect, yet it conveys valuable information about possible improvements in their strategy. Scout has been used by several NBA teams (US News 95).

5.3 Discovery Agents

Finally, a novel and very important type of discovery system has appeared recently – Discovery Agents. Although the idea of active triggers has long been analyzed in the database field, really successful applications of this idea appeared only with the advent of the Internet. These systems ask the user to specify a profile of interest and search for related information among a wide variety of public domain and proprietary sources.

To mention a few examples, the Firefly is personal music recommendation agent – asks user their opinion of several music pieces and then suggests other music that the user may like `<http://www.ffly.com/>`.

Crayon `<http://crayon.net/>` allows users to create their own free newspaper (supported by ads). Farcast `<http://www.farcast.com/>` and NewsHound `<http://www.sjmercury.com/hound/>` from San Jose Mercury automatically search information from a wide variety of sources, including newspapers and wire services, and email relevant documents directly to the user.

6 Application Development Issues

While the data mining and knowledge discovery technology is quite well developed, its practical application in industry is hampered by a number of difficulties, reviewed below.

Insufficient training: graduates of business schools will be familiar with verification-driven analysis techniques, occasionally with predictive modeling, and very seldom with other discovery techniques. Extending the training of business analysts to acknowledge the full range of techniques available can alleviate this problem; it can also be addressed by making the discovery techniques easily available to business users (see sec. 2).

Inadequate tool support: most available data mining tools support at most one of the core discovery techniques, typically only prediction. Other methods, such as clustering, deviation detection, visualization, summarization are also needed, as well as methods for dealing with exceptions (rare cases) which may be significant in some applications. The tools must also support the complete knowledge discovery process (section 3) and provide a user interface suitable for business users rather than technologists. Some integrated toolkits are now emerging which satisfy these requirements.

Data inaccessibility: for a given business problem, the required data is often distributed across the organization in a variety of different formats, and is often poorly organized or maintained. For this reason, data acquisition and pre-processing usually play a very significant part in any knowledge discovery project. Data warehousing is now becoming widespread, and can potentially alleviate such problems. Both warehousing and data mining often serve to highlight the problems of data quality in an organization, but can also help to solve them.

Overabundance of patterns: when search for patterns has a wide scope, a very large number of patterns can be discovered. Proper statistical controls are needed to avoid discoveries due to chance, while domain knowledge can help to focus on the interesting findings. Rule refinement (Major & Mangano 1995) and other generalization methods could be used to further compress findings.

Changing and time-oriented data: many applications deal with behaviour that changes significantly over time (e.g. stock market or fashions). Such applications are on one hand more challenging because common algorithms suitable for flat tables do not work well with sequential and other time-oriented patterns. On the other hand, such applications can become especially successful, since it is easier to retrain a system than to retrain a human. The improvement in decision-making due to regular updating of decision-making tools is referred to as "volatility benefit". A few recent data mining methods are designed for handling deviation detection and time-oriented data (Agrawal & Psaila 1995, Berndt & Clifford 1996, Mannila, Toivonen, & Verkamo 1995).

Space oriented data: other applications, especially in manufacturing, biology, and geographically-oriented systems, are dealing with spatial data. Here again there are special types of patterns which require special algorithms (Stolorz et al. 1995). Geographical information systems have been very successful in helping to find some types of spatial patterns visually.

Complex data: Other types of information, including text, images, audio, video, and anything related to the Web present an even grander challenge with potentially great rewards.

Scalability: Although many papers (including this one) talk about needing to mine vast amounts of data, none of the tools can do that today. Data warehouses that start at 200GB are not infrequent today, yet current tools can at best deal with 1GB at a time. Progress, however, is being made towards using massively parallel and high performance computing which will help to deal with large databases.

Privacy: When dealing with databases of personal information, governments and businesses have to be careful to adequately address the legal and ethical issues of invasion of privacy (see Piatetsky-Shapiro 1995).

7 Assessing Benefits Of KDD Applications

The domains suitable for data mining are those that are information rich, have a changing environment, do not already have existing models, require knowledge-based decisions, and provide high payoff for the right decisions. Given a suitable domain, we examine costs and benefits of a potential application by looking at the following factors.

- Alternatives: there should be no simpler alternative solutions.
- Relevance: relevant factors should be included.
- Volume: there should be a sufficient number of cases (several thousands at least). Extremely large databases may be a problem when the results are needed quickly.
- Complexity: the more variables (fields) there are the more complex is the application. Complexity is also increased for time-series data.
- Quality: Error rate should be relatively low.
- Accessibility: data should be easily accessible – accessing data or merging data from different sources increases the cost of an application.
- Change: although dealing with change is more difficult, it can also be more rewarding (the volatility benefit) since the application can be automatically and regularly "re-trained" on up-to-date data.
- Expertise: The more expertise available, the easier is the project. It should be emphasized that expertise on the form and meaning of the data is just as important as knowledge of problem-solving in the domain.

Although the challenges are many and the difficulties are substantial, the future of data mining applications looks bright. There is a widespread realization of the potential value of data mining and a growing number of researchers and developers are working on the topic. However, data mining by itself is only a part of the overall application and all other components, as described in Section 3 need to be addressed for a successful application.

Acknowledgments: We thank Usama Fayyad and Sam Uthurusamy for encouraging us to put together this survey, a version of which will also appear in Communications of ACM. Robert Golan helped with information on financial data mining. Colin Shearer (ISL) has contributed much useful material to this article.

8 References

Anand, T. and Kahn, G. 1992. SPOTLIGHT: A Data Explanation System. In *Proceedings Eighth IEEE Conference on Applied AI*, 2–8. Washington, D.C.: IEEE Press.

Anand, T. 1995. Opportunity Explorer: Navigating Large Databases Using Knowledge Discovery Templates. *Journal of Intelligent Information Systems* 4(1):27–38.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. 1996. Fast Discovery of Association Rules. In *AKDDM*, Cambridge, MA: AAAI/MIT Press.

Agrawal, R., and Psaila, G. 1995. Active Data Mining. In *Proceedings of KDD-95*, 3–8, Menlo Park, CA: AAAI Press.

Brachman, R., et al. 1993. Integrated Support for Data Archaeology. In *Proceedings of KDD-93 Workshop*, Menlo Park, CA: AAAI Press.

Brachman, R. and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human Centered Approach. In *AKDDM*, Cambridge, MA: AAAI/MIT Press.

Berndt, D. and Clifford, J. 1996. Finding Patterns in Time Series: A Dynamic Programming Approach. In *AKDDM*, Cambridge, MA: AAAI/MIT Press.

Berry, J. 1994. Database Marketing. *Business Week*, 56–62, Sep 5.

Codd, E.F. 1993. Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate. E.F. Codd and Associates.

Davis, A. and Goyal, S. 1993. Management of Cellular Fraud: Knowledge- Based Detection, Classification and Prevention. In *Proceedings of 13th Int. Conf. on AI, Expert Systems and Natural Language*, Avignon, France, Vol. 2, pp. 155-164.

Eick, S. and Fyock, D. 1996. Visualizing Corporate Data, *AT&T Technical Journal*, January/February, pp. 74–86.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. eds. 1996. *Advances in Knowledge Discovery and Data Mining* (AKDDM). Cambridge, MA: AAAI/MIT Press.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Proceedings of KDD-96*, Menlo Park, CA: AAAI Press.

Fayyad, U., Haussler, D., and Stolorz, P. 1996. KDD for Science Data Analysis: Issues and Examples. In *Proceedings of KDD-96*, Menlo Park, CA: AAAI Press.

Fitzsimons, M., Khabaza, T., and Shearer, C. 1993. The Application of Rule Induction and Neural Networks for Television Audience Prediction. In *Proc. of ESOMAR/EMAC/AFM Symposium on Information Based Decision Making in Marketing*, Paris, November, pp 69-82.

Hall, J., Mani, G., and Barr, D. 1996. Applying Computational Intelligence to the Investment Process. In *Proc. of CIFER-96: Computational Intelligence in Financial Engineering*. Piscataway, NJ: IEEE Press.

Hernandez, M. and Stolfo, S. 1995. The Merge/Purge Problem for Large Databases. In *Proc. of the 1995 ACM-SIGMOD Conference*, 127–138. NY: ACM Press.

Major, J., and Mangano, J. 1995. Selecting among Rules Induced from a Hurricance Database. *Journal of Intelligent Information Systems* 4(1):39–52.

Manago, M. and Auriol, M. 1996. Mining for OR. *ORMS Today*, February, Special issue on Data Mining, 28–32.

Mannila, H., Toivonen, H., and Verkamo, A. 1995. Discovering Frequent Episodes in Sequences, In *Proceedings of KDD-95*, 210–215. Menlo Park, CA: AAAI Press.

Matheus, C., Piatetsky-Shapiro, G., and McNeill, D. 1996. Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data. In *AKDDM*, Cambridge, MA: AAAI/MIT Press, 495–516.

Piatetsky-Shapiro, G. 1995. Knowledge Discovery in Personal Data vs. Privacy – a Minisymposium. *IEEE Expert*, April 1995.

Schmitz, J., Armstrong, G. and Little, J. D. C. 1990. CoverStory – Automated News Finding in Marketing. In *DSS Transactions*, ed. L. Volino, 46–54. Providence, R.I.: Institute of Management Sciences.

Senator, T. et al. 1995. The Financial Crimes Enforcement Network AI System (FAIS), *AI Magazine*, Winter 1995, 21–39.

Stolorz, P. et al. 1995. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets. In *Proceedings of KDD-95*, 300–305, Menlo Park, CA: AAAI Press.

US News & World Report, December 11, 1995. "Basketball's new high-tech guru: IBM software is changing coaches' game plans."