APPENDIX B
DATA-MINING APPLICATIONS

Many businesses and scientific communities are currently employing data-mining technology. Their number continues to grow, as more and more data-mining success stories become known. Here we present a small collection of real-life examples of data-mining implementations from the business and scientific world. We also present some pitfalls of data mining to make readers aware that this process needs to be applied with care and knowledge (both, about the application domain and about the methodology) to obtain useful results.

In the previous chapters of this book, we have studied the principles and methods of data mining. Since data mining is a young discipline with wide and diverse applications, there is a still a serious gap between the general principles of data mining and the domain-specific knowledge required to apply it effectively. In this appendix, we examine a few application domains illustrated by the results of data-mining systems that have been implemented.

B.1 DATA MINING FOR FINANCIAL DATA ANALYSIS

Most banks and financial institutions offer a wide variety of banking services such as checking, savings, business and individual customer transactions, investment services, credits, and loans. Financial data, collected in the banking and financial industry, are often relatively complete, reliable, and of a high quality, which facilitates systematic data analysis and data mining to improve a company's competitiveness.

In the banking industry, data mining is used heavily in the areas of modeling and predicting credit fraud, in evaluating risk, in performing trend analyses, in analyzing profitability, as well as in helping with direct-marketing campaigns. In the financial markets, neural networks have been used in forecasting stock prices, options trading, rating bonds, portfolio management, commodity-price prediction, and mergers and acquisitions analyses; it has also been used in forecasting financial disasters. Daiwa Securities, NEC Corporation, Carl & Associates, LBS Capital Management, Walkrich

Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition. Mehmed Kantardzic.

Investment Advisors, and O'Sallivan Brothers Investments are only a few of the financial companies who use neural-network technology for data mining. A wide range of successful business applications has been reported, although the retrieval of technical details is not always easy. The number of investment companies and banks that mine data is far more extensive than the list mentioned earlier, but you will not often find them willing to be referenced. Usually, they have policies not to discuss it. Therefore, finding articles about banking companies who use data mining is not an easy task, unless you look at the SEC reports of some of the data-mining companies who sell their tools and services. There, you will find customers such as Bank of America, First USA Bank, Wells Fargo Bank, and U.S. Bancorp.

The widespread use of data mining in banking has not been unnoticed. *Bank Systems & Technology* commented that data mining was the most important application in financial services in 1996. For example, fraud costs industries billions of dollars, so it is not surprising to see that systems have been developed to combat fraudulent activities in such areas as credit card, stock market, and other financial transactions. Fraud is an extremely serious problem for credit-card companies. For example, Visa and MasterCard lost over $700 million in 1995 from fraud. A neural network-based credit card fraud-detection system implemented in Capital One has been able to cut the company's losses from fraud by more than 50%. Several successful data-mining systems are explained here to support the importance of data-mining technology in financial institutions.

U.S. Treasury Department

> *Worth particular mention is a system developed by the Financial Crimes Enforcement Network (FINCEN) of the U.S. Treasury Department called "FAIS." FAIS detects potential money-laundering activities from a large number of big cash transactions. The Bank Secrecy Act of 1971 required the reporting of all cash transactions greater than $10,000, and these transactions, of about 14 million a year, are the basis for detecting suspicious financial activities. By combining user expertise with the system's rule-based reasoner, visualization facilities, and association-analysis module, FIAS uncovers previously unknown and potentially high-value leads for possible investigation. The reports generated by the FIAS application have helped FINCEN uncover more than 400 cases of money-laundering activities, involving more than $1 billion in potentially laundered funds. In addition, FAIS is reported to be able to discover criminal activities that law enforcement in the field would otherwise miss, for example, connections in cases involving nearly 300 individuals, more than 80 front operations, and thousands of cash transactions.*

Mellon Bank, USA

> *Mellon Bank has used the data on existing credit-card customers to characterize their behavior and they try to predict what they will do next. Using IBM Intelligent Miner, Mellon developed a credit card-attrition model to predict which customers will stop using Mellon's credit card in the next few months. Based on the prediction results, the bank can take marketing actions to retain these customers' loyalty.*

Capital One Financial Group

Financial companies are one of the biggest users of data-mining technology. One such user is Capital One Financial Corp., one of the nation's largest credit-card issuers. It offers 3000 financial products, including secured, joint, co-branded, and college-student cards. Using data-mining techniques, the company tries to help market and sell the most appropriate financial product to 150 million potential prospects residing in its over 2-terabyte Oracle-based data warehouse. Even after a customer has signed up, Capital One continues to use data mining for tracking the ongoing profitability and other characteristics of each of its customers. The use of data mining and other strategies has helped Capital One expand from $1 billion to $12.8 billion in managed loans over 8 years. An additional successful data-mining application at Capital One is fraud detection.

American Express

Another example of data mining is at American Express, where data warehousing and data mining are being used to cut spending. American Express has created a single Microsoft SQL Server database by merging its worldwide purchasing system, corporate purchasing card, and corporate-card databases. This allows American Express to find exceptions and patterns to target for cost cutting. One of the main applications is loan application screening. American Express used statistical methods to divide loan applications into three categories: those that should definitely be accepted, those that should definitely be rejected, and those which required a human expert to judge. The human experts could correctly predict if an applicant would, or would not, default on the loan in only about 50% of the cases. Machine learning produced rules that were much more accurate—correctly predicting default in 70% of the cases—and that were immediately put into use.

MetLife, Inc.

MetLife's Intelligent Text Analyzer has been developed to help automate the underwriting of 260,000 life insurance applications received by the company every year. Automation is difficult because the applications include many free-form text fields. The use of keywords or simple parsing techniques to understand the text fields has proven to be inadequate, while the application of full semantic natural-language processing was perceived to be too complex and unnecessary. As a compromise solution, the "information-extraction" approach was used in which the input text is skimmed for specific information relevant to the particular application. The system currently processes 20,000 life-insurance applications a month and it is reported that 89% of the text fields processed by the system exceed the established confidence-level threshold.

Bank of America (USA)

Bank of America is one of the world's largest financial institutions. With approximately 59 million consumer and small business relationships, 6,000 retail banking offices and more than 18,000 ATMs, Bank of America is among the world's leading wealth management companies and is a global leader in corporate and investment banking and *trading across a broad range of asset classes. Bank of America identified savings of $4.8 million in2 years (a 400% return on investment) from use of a credit risk management system provided by SAS institute consultants and based on statistical and data-mining analytics ["Predicting Returns from the Use of Data Mining to Support CRM,"* http://insight.nau.edu/WhitePapers.asp*]. They have also developed profiles of most valuable accounts, with relationship managers being assigned to the top 10% of the bank's customers in order to identify opportunities to sell them additional services ["Using Data Mining on the Road to Successful BI, Part 3," Information Management Special Reports, Oct. 2004]. Recently, to retain deposits, the Global Wealth and Investment Management division has used KXEN Analytic Framework in identifying clients likely to move assets and then creating offers conducive to retention ["KXEN Analytic Framework," Information Management Magazine, July/Aug 2009].*

B.2 DATA MINING FOR THE TELECOMUNICATIONS INDUSTRY

The telecommunication industry has quickly evolved from offering local and long-distance telephone services to providing many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer, and Web-data transmission, and other data traffic. The integration of telecommunications, computer networks, Internet, and numerous others means of communication and computing is under way. The U.S. Telecommunication Act of 1996 allowed Regional Bell Operating Companies to enter the long-distance market as well as offer "cable-like" services. The European Liberalization of Telecommunications Services has been effective from the beginning of 1998. Besides deregulation, there has been a sale by the FCC of airwaves to companies pioneering new ways to communicate. The cellular industry is rapidly taking on a life of its own. With all this deregulation of the telecommunication industry, the market is expanding rapidly and becoming highly competitive.

The hypercompetitive nature of the industry has created a need to understand customers, to keep them, and to model effective ways to market new products. This creates a great demand for data mining to help understand the new business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of services. In general, the telecommunications industry is interested in answering some strategic questions through data-mining applications such as:

- How does one retain customers and keep them loyal as competitors offer special offers and reduced rates?
- Which customers are most likely to churn?
- What characteristics indicate high-risk investments, such as investing in new fiber-optic lines?
- How does one predict whether customers will buy additional products like cellular services, call waiting, or basic services?
- What characteristics differentiate our products from those of our competitors?

Companies like AT&T, AirTouch Communications, and AMS Mobile Communication Industry Group have announced the use of data mining to improve their marketing activities. There are several companies including Lightbridge and Verizon that use data-mining technology to look at cellular fraud for the telecommunications industry. Another trend has been to use advanced visualization techniques to model and analyze wireless-telecommunication networks. Selected examples of data-mining applications in the telecommunication industry follow.

Cablevision Systems, Inc.

Cablevision Systems Inc., a cable TV provider from New York, was concerned about its competitiveness after deregulation allowed telecom companies into the cable industry. As a consequence, it decided that it needed a central data repository so that its marketing people could have faster and more accurate access to data. Using data mining, the marketing people at Cablevision were able to identify nine primary customer segments among the company's 2.8 million customers. This included customers in the segment that are likely to "switch" to another provider. Cablevision also focused on those segments most likely to buy its offerings for new services. The company has used data mining to compare the profiles of two sets of targeted customers—those who bought new services and those who did not. This has led the company to make some changes in its messages to customers, which, in turn, has led to a 30% increase in targeted customers signing up for new services

Worldcom

Worldcom is another company that has found great value in data mining. By mining databases of its customer-service and telemarketing data, Worldcom has discovered new ways to sell voice and data services. For example, it has found that people who buy two or more services were likely to be relatively loyal customers. It also found that people were willing to buy packages of products such as long-distance, cellular-phone, Internet, and other services. Consequently, Worldcom started to offer more such packages.

BBC TV

TV-program schedulers would like to know the likely audience for a proposed program and the best time to show it. The data for audience prediction are fairly complex. Factors, which determine the audience share gained by a particular program, include not only the characteristics of the program itself and the time at which is shown, but also the nature of the competing programs in other channels. Using Clementine, Integral Solutions Limited developed a system to predict television audiences for the BBC. The prediction accuracy was reported to be the same as that achieved by the best performance of BBC's planners.

Bell Atlantic

Bell Atlantic developed telephone technician dispatch system. When a customer reports a telephone problem to Bell Atlantic, the company must decide what type of technician *to dispatch to resolve the issue. Starting in 1991, this decision was made using a handcrafted expert system, but in 1999 it was replaced by another set of rules created with machine learning. The learned rules save Bell Atlantic more than 10 million dollars per year because they make fewer erroneous decisions. In addition, the original expert system had reached a stage in its evolution where it could not be maintained cost-effectively. Because the learned system was built by training it on examples, it is easy to maintain and to adapt to regional differences and changing cost structures.*

B.3 DATA MINING FOR THE RETAIL INDUSTRY

Slim margins have pushed retailers into data warehousing earlier than other industries. Retailers have seen improved decision-support processes leading directly to improved efficiency in inventory management and financial forecasting. The early adoption of data warehousing by retailers has allowed them a better opportunity to take advantage of data mining. The retail industry is a major application area for data mining since it collects huge amounts of data on sales, customer-shopping history, goods transportation, consumption patterns, and service records, and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing availability and popularity of business conducted on the Web, or e-commerce. Today, many stores also have Web sites where customers can make purchases online. A variety of sources and types of retail data provide a rich source for data mining.

Retail data mining can help identify customer-buying behaviors, discover customer-shopping patterns and trends, improve the quality of customer services, achieve better customer retention and satisfaction, enhance goods consumption, design more effective goods transportation and distribution policies, and, in general, reduce the cost of business and increase profitability. In the forefront of applications that have been adopted by the retail industry are direct-marketing applications. The direct-mailing industry is an area where data mining is widely used. Almost every type of retailer uses direct marketing, including catalogers, consumer retail chains, grocers, publishers, B2B marketers, and packaged goods manufacturers. The claim could be made that every Fortune 500 company has used some level of data mining in their direct-marketing campaigns. Large retail chains and groceries stores use vast amounts of sale data that are "information-rich." Direct marketers are mainly concerned about customer segmentation, which is a clustering or classification problem.

Retailers are interested in creating data-mining models to answer questions such as:

- What are the best types of advertisements to reach certain segments of customers?
- What is the optimal timing at which to send mailers?
- What is the latest product trend?
- What types of products can be sold together?
- How does one retain profitable customers?
- What are the significant customer segments that buy products?

Data mining helps to model and identify the traits of profitable customers, and it also helps to reveal the "hidden relationship" in data that standard-query processes have not found. IBM has used data mining for several retailers to analyze shopping patterns within stores based on point-of-sale (POS) information. For example, one retail company with $2 billion in revenue, 300,000 UPC codes, and 129 stores in 15 states found some interesting results: ". . . we found that people who were coming into the shop gravitated to the left-hand side of the store for promotional items, and they were not necessarily shopping the whole store." Such information is used to change promotional activities and provide a better understanding of how to lay out a store in order to optimize sales. Additional real-world examples of data-mining systems in retail industry follow.

Safeway, UK

Grocery chains have been another big user of data-mining technology. Safeway is one such grocery chain with more than $10 billion in sales. It uses Intelligent Miner from IBM to continually extract business knowledge from its product-transaction data. For example, the data-mining system found that the top-spending 25% customers very often purchased a particular cheese product ranked below 200 in sales. Normally, without the data-mining results, the product would have been discontinued. But the extracted rule showed that discontinuation would disappoint the best customers, and Safeway continues to order this cheese, although it is ranked low in sales. Thanks to data mining, Safeway is also able to generate customized mailing to its customers by applying the sequence-discovery function of Intelligent Miner, allowing the company to maintain its competitive edge.

RS Components, UK

RS Components, a UK-based distributor of technical products such as electronic and electrical components and instrumentation, has used the IBM Intelligent Miner to develop a system to do cross selling (suggested related products on the phone when customers ask for one set of products), and in warehouse product allocation. The company had one warehouse in Corby before 1995 and decided to open another in the Midlands to expand its business. The problem was how to split the products into these two warehouses so that the number of partial orders and split shipments could be minimized. Remarkably, the percentage of split orders is just about 6% after using the patterns found by the system, much better than expected.

Kroger Co. (USA)

The Kroger is the largest grocery store chain in the United States. Forty percent of all U.S households have one of Kroger's loyalty cards. The Kroger is trying to drive loyalty for life with their customers. In particular, their customers are rewarded with offers on what they buy instead of trying to be sold something else. In other words, each of them could receive coupons different from each other, not the same coupons. In order to match the best customers with the right coupons, the Kroger analyses customers' behavior using the data-mining techniques. For instance, one recent mailing was customized to 95% of the intended recipients. Such business strategy for looking at customers to win customers for life makes the Kroger beat their largest competitor,

> *Walmart, for the last 6 years largely. [http://www.kypost.com/dpp/news/region_central_cincinnati/downtown/data-mining-is-big-business-for-kroger-%26-getting-bigger-all-the-time]*

Korea Customs Service (South Korea)

> *The Korea Customs Service (KCS) is a government agency established to secure national revenues by controlling imports and exports for the economic development of South Korea and to protect domestic industry through contraband control. It is responsible for the customs clearance of imported goods as well as tax collection at the customs border. For detecting illegal cargo, they implemented a system using SAS for fraud detection, based on its widespread use and trustworthy reputation in the data-mining field. This system enabled more specific and accurate sorting of illegal cargo. For instance, the number of potentially illegal factors increased from 77 to 163. As a result, the detection rate for important items, as well as the total rate, increased by more than 20% [http://www.sas.com/success/kcs.html].*

B.4 DATA MINING IN HEALTH CARE AND BIOMEDICAL RESEARCH

With the amount of information and issues in the health-care industry, not to mention the pharmaceutical industry and biomedical research, opportunities for data-mining applications are extremely widespread, and benefits from the results are enormous. Storing patients' records in electronic format and the development in medical-information systems cause a large amount of clinical data to be available online. Regularities, trends, and surprising events extracted from these data by data-mining methods are important in assisting clinicians to make informed decisions, thereby improving health services.

Clinicians evaluate a patient's condition over time. The analysis of large quantities of time-stamped data will provide doctors with important information regarding the progress of the disease. Therefore, systems capable of performing temporal abstraction and reasoning become crucial in this context. Although the use of temporal-reasoning methods requires an intensive knowledge-acquisition effort, data mining has been used in many successful medical applications, including data validation in intensive care, the monitoring of children's growth, analysis of a diabetic patient's data, the monitoring of heart-transplant patients, and intelligent anesthesia monitoring.

Data mining has been used extensively in the medical industry. Data visualization and artificial neural networks are especially important areas of data mining applicable in the medical field. For example, NeuroMedicalSystems used neural networks to perform a pap smear diagnostic aid. Vysis Company uses neural networks to perform protein analyses for drug development. The University of Rochester Cancer Center and the Oxford Transplant Center use KnowledgeSeeker, a decision tree-based technology, to help with their research in oncology.

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and advances in cancer therapies to the identification and study of the human genome. The logic behind investigating the genetic causes of diseases is that once the molecular bases of diseases are known, precisely targeted medical interventions for diagnostics, prevention, and treatment of the disease themselves can be developed. Much of the work occurs in the context of the development of new pharmaceutical products that can be used to fight a host of diseases ranging from various cancers to degenerative disorders such as Alzheimer's Disease.

A great deal of biomedical research has focused on DNA-data analysis, and the results have led to the discovery of genetic causes for many diseases and disabilities. An important focus in genome research is the study of DNA sequences since such sequences form the foundation of the genetic codes of all living organisms. What is DNA? Deoxyribonucleic acid, or DNA, forms the foundation for all living organisms. DNA contains the instructions that tell cells how to behave and is the primary mechanism that permits us to transfer our genes to our offspring. DNA is built in sequences that form the foundations of our genetic codes, and that are critical for understanding how our genes behave. Each gene comprises a series of building blocks called nucleotides. When these nucleotides are combined, they form long, twisted, and paired DNA sequences or chains. Unraveling these sequences has become a challenge since the 1950s when the structure of the DNA was first understood. If we understand DNA sequences, theoretically, we will be able to identify and predict faults, weaknesses, or other factors in our genes that can affect our lives. Getting a better grasp of DNA sequences could potentially lead to improved procedures to treat cancer, birth defects, and other pathological processes. Data-mining technologies are only one weapon in the arsenal used to understand these types of data, and the use of visualization and classification techniques is playing a crucial role in these activities.

It is estimated that humans have around 100,000 genes, each one having DNA that encodes a unique protein specialized for a function or a set of functions. Genes controlling production of hemoglobin, regulation of insulin, and susceptibility to Huntington's chorea are among those that have been isolated in recent years. There are seemingly endless varieties of ways in which nucleotides can be ordered and sequenced to form distinct genes. Any one gene might comprise a sequence containing hundreds of thousands of individual nucleotides arranged in a particular order. Furthermore, the process of DNA sequencing used to extract genetic information from cells and tissues usually produces only fragments of genes. It has been difficult to tell using traditional methods where these fragments fit into the overall complete sequence from which they are drawn. Genetic scientists face the difficult task of trying to interpret these sequences and form hypotheses about which genes they might belong to, and the disease processes that they may control. The task of identifying good candidate gene sequences for further research and development is like finding a needle in a haystack. There can be hundreds of candidates for any given disease being studied. Therefore, companies must decide which sequences are the most promising ones to pursue for further development. How do they determine which ones would make good therapeutic targets? Historically, this has been a process based largely on trial and error. For every lead that eventually turns into a successful pharmaceutical intervention that is effective in clinical settings, there are dozens of others that do not produce the anticipated results. This is a research area that is crying out for innovations that can help to make these analytical processes more efficient. Since pattern analysis, data visualization, and similarity-search techniques have been developed in data mining, this field has become a powerful infrastructure for further research and discovery in DNA sequences. We will describe one attempt to innovate the process of mapping human genomes that has been undertaken by Incyte Pharmaceuticals, Inc. in cooperation with Silicon Graphics.

Incyte Pharmaceuticals, Inc.

> *Incyte Pharmaceuticals is a publicly held company founded in 1991, and it is involved in high-throughput DNA sequencing and development of software, databases, and other products to support the analysis of genetic information. The first component of their activities is a large database called LiveSeq that contains more than 3 million human-gene sequences and expression records. Clients of the company buy a subscription to the database and receive monthly updates that include all of the new sequences identified since the last update. All of these sequences can be considered as candidate genes that might be important for future genome mapping. This information has been derived from DNA sequencing and bioanalysis of gene fragments extracted from cell and tissue samples. The tissue libraries contain different types of tissues including normal and diseased tissues, which are very important for comparison and analyses.*
>
> *To help impose a conceptual structure of the massive amount of information contained in LifeSeq, the data has been coded and linked to several levels. Therefore, DNA sequences can be grouped into many different categories, depending on the level of generalization. LifeSeq has been organized to permit comparisons of classes of sequence information within a hypothesis-testing mode. For example, a researcher could compare gene sequences isolated from diseased and non-diseased tissue from an organ. One of the most important tools that are provided in LifeSeq is a measure of similarity among sequences that are derived from specific sources. If there is a difference between two tissue groups for any available sequences, this might indicate that these sequences should be explored more fully. Sequences occurring more frequently in the diseased sample might reflect genetic factors in the disease process. On the other hand, sequences occurring more frequently in the non-diseased sample might indicate mechanisms that protect the body from the disease.*
>
> *Although it has proved invaluable to the company and their clients in its current incarnation, additional features are being planned and implemented to extend the LifeSeq functionality into research areas such as*
>
> - *identifying co-occurring gene sequences,*
> - *tying genes to disease stage, and*
> - *using LifeSeq to predict molecular toxicology.*
>
> *Although the LifeSeq database is an invaluable research resource, queries to the database often produce very large data sets that are difficult to analyze in text format. For this reason, Incyte developed the LifeSeq 3-D application that provides visualization of data sets, and also allows users to cluster or classify and display information about genes. The 3-D version has been developed using the Silicon Graphics MineSet tool. This version has customized functions that let researchers explore data from LifeSeq and discover novel genes within the context of targeted protein functions and tissue types.*

Maine Medical Center (USA)

> *Maine Medical Center—a teaching hospital and the major community hospital for the Portland, Maine, area—has been named in the U.S. News and World Report Best Hospitals list twice in orthopedics and heart care. In order to improve quality of patient care in measurable ways, Maine Medical Center has used scorecards as key performance indicators. Using SAS, the hospital creates balanced scorecards that measure everything from staff hand washing compliance to whether a congestive heart patient is actually offered a flu vaccination. One hundred percent of heart failure patients are getting quality care as benchmarked by national organizations, and a medication error reduction process has improved by 35%.*
>
> *http://www.sas.com/success/mainemedicalcenter.html*
>
> *In November 2009, the Central Maine Medical Group (CMMG) announced the launch of a prevention and screening campaign called "Saving Lives Through Evidence-Based Medicine." The new initiative is employed to redesign the ways that it works as a team of providers to make certain that each of our patients undergoes the necessary screening tests identified by the current medical literature using data-mining techniques. In particular, data-mining process identifies someone at risk for an undetected health problem http://www.cmmc.org/news.taf].*

B.5 DATA MINING IN SCIENCE AND ENGINEERING

Enormous amounts of data have been generated in science and engineering, for example, in cosmology, molecular biology, and chemical engineering. In cosmology, advanced computational tools are needed to help astronomers understand the origin of large-scale cosmological structures as well as the formation and evolution of their astrophysical components (galaxies, quasars, and clusters). Over 3 terabytes of image data have been collected by the Digital Palomar Observatory Sky Survey, which contain on the order of 2 billion sky objects. It has been a challenging task for astronomers to catalog the entire data set, that is, a record of the sky location of each object and its corresponding classification such as a star or a galaxy. The Sky Image Cataloguing and Analysis Tool (SKICAT) has been developed to automate this task. The SKICAT system integrates methods from machine learning, image processing, classification, and databases, and it is reported to be able to classify objects, replacing visual classification, with high accuracy.

In molecular biology, recent technological advances are applied in such areas as molecular genetics, protein sequencing, and macro-molecular structure determination as was mentioned earlier. Artificial neural networks and some advanced statistical methods have shown particular promise in these applications. In chemical engineering, advanced models have been used to describe the interaction among various chemical processes, and also new tools have been developed to obtain a visualization of these structures and processes. Let us have a brief look at a few important cases of data-mining applications in engineering problems. Pavilion Technologies' Process Insights, an application-development tool that combines neural networks, fuzzy logic, and statistical methods has been successfully used by Eastman Kodak and other companies to develop chemical manufacturing and control applications to reduce waste, improve product quality, and increase plant throughput. Historical process data is used to build a predictive model of plant behavior and this model is then used to change the control set points in the plant for optimization.

DataEnginee is another data-mining tool that has been used in a wide range of engineering applications, especially in the process industry. The basic components of the tool are neural networks, fuzzy logic, and advanced graphical user interfaces. The tool has been applied to process analysis in the chemical, steel, and rubber industries, resulting in a saving in input materials and improvements in quality and productivity. Successful data-mining applications in some industrial complexes and engineering environments follow.

Boeing

> *To improve its manufacturing process, Boeing has successfully applied machine-learning algorithms to the discovery of informative and useful rules from its plant data. In particular, it has been found that it is more beneficial to seek concise predictive rules that cover small subsets of the data, rather than generate general decision trees. A variety of rules were extracted to predict such events as when a manufactured part is likely to fail inspection or when a delay will occur at a particular machine. These rules have been found to facilitate the identification of relatively rare but potentially important anomalies.*

R.R. Donnelly

> *This is an interesting application of data-mining technology in printing press control. During rotogravure printing, grooves sometimes develop on the printing cylinder, ruining the final product. This phenomenon is known as banding. The printing company R.R. Donnelly hired a consultant for advice on how to reduce its banding problems, and at the same time used machine learning to create rules for determining the process parameters (e.g., the viscosity of the ink) to reduce banding. The learned rules were superior to the consultant's advice in that they were more specific to the plant where the training data was collected and they filled gaps in the consultant's advice and thus were more complete. In fact, one learned rule contradicted the consultant's advice and proved to be correct. The learned rules have been in everyday use in the Donnelly plant in Gallatin, Tennessee, for over a decade and have reduced the number of banding occurrences from 538 to 26.*

Southern California Gas Company

> *The Southern California Gas Company is using SAS software as a strategic marketing tool. The company maintains a data mart called the Customer Marketing Information Database that contains internal billing and order data along with external demographic data. According to the company, it has saved hundreds of thousands of dollars by identifying and discarding ineffective marketing practices.*

WebWatcher

> *Despite the best effort of Web designers, we all have had the experience of not being able to find a certain Web page we want. A bad design for a commercial Web site*

obviously means the loss of customers. One challenge for the data-mining community has been the creation of "adaptive Web sites"; Web sites that automatically improve their organization and presentation by learning from user-access patterns. One early attempt is WebWatcher, an operational tour guide for the WWW. It learns to predict what links users will follow on a particular page, highlight the links along the way, and learn from experience to improve its advice-giving skills. The prediction is based on many previous access patterns and the current user's stated interests. It has also been reported that Microsoft is to include in its electronic-commerce system a feature called Intelligent Cross Sell that can be used to analyze the activity of shoppers on a Web site and automatically adapt the site to that user's preferences.

AbitibiBowater Inc. (Canada)

AbitibiBowater Inc. is a pulp and paper manufacturer headquartered in Montreal, Quebec, Canada. The pulp and paper, a key component of the forest products industry, is a major contributor to Canada's economy. In addition to market pulp, the sector produces newsprint, specialty papers, paperboard, building board and other paper products. It is the largest industrial energy consumer, representing 23% of industrial energy consumption in Canada. AbitibiBowater Inc. used data-mining techniques to detect a period of high performance and reduce energy consumption in the paper making process, so that they recognized that lower temporary consumption is caused by the reduced set point for chip preheating and cleaning of the heating tower on the reject refiners. AbitibiBowater Inc. was able to reproduce the process conditions required to maintain steam recovery. This has saved AbitibiBowater 200 gigajoules[1] *daily—the equivalent of $600,000 a year. [Head Up CIPEC (Canadian Industry Program for Energy Conservation) new letter: Aug. 15, 2009 Vol. XIII, No.15]*

eHarmony

The eHarmony dating service, which rather than matching prospective partners on the basis of their stated preferences, uses statistical analysis to match prospective partners, based on a 29-parameter model derived from 5000 successful marriages. Its competitors such as Perfectmatch use different models, such as the Jungian Meyers-Briggs personality typing technique to parameterize individuals entered into their database. It is worth observing that while the process of matching partners may amount to little more than data retrieval using some complex set of rules, the process of determining what these rules need to be involves often complex knowledge discovery and mining techniques.

The maintenance of military platforms

Another area where data-mining techniques offer promising gains in efficiency is in the maintenance of military platforms. Good and analytically based maintenance programs, with the Amberley Ageing Aircraft Program for the F-111 a good example,

[1] A gigajoule (GJ) is a metric term used for measuring energy use. For example, 1 GJ is equivalent to the amount of energy available from either: 277.8 kWh of electricity, or 26.1 m^3 of natural gas, or 25.8 L of heating oil.

systematically analyze component failure statistics to identify components with wear out or other failure rate problems. They can then be removed from the fleet by replacement with new or reengineered and thus more reliable components. This type of analysis is a simple rule-based approach, where the rule is simply the frequency of faults in specific components.

B.6 PITFALLS OF DATA MINING

Despite the above and many other success stories often presented by vendors and consultants to show the benefits that data mining provides, this technology has several pitfalls. When used improperly, data mining can generate lots of "garbage." As one professor from MIT pointed out: "Given enough time, enough attempts, and enough imagination, almost any set of data can be teased out of any conclusion." David J. Lainweber, managing director of First Quadrant Corp. in Pasadena, California, gives an example of the pitfalls of data mining. Working with a United Nations data set, he found that historically, butter production in Bangladesh is the single best predictor of the Standard & Poor's 500-stock index. This example is similar to another absurd correlation that is heard yearly around Super Bowl time—a win by the NFC team implies a rise in stock prices. Peter Coy, Business Week's associate economics editor, warns of four pitfalls in data mining:

1. It is tempting to develop a theory to fit an oddity found in the data.
2. One can find evidence to support any preconception if you let the computer churn long enough.
3. A finding makes more sense if there is a plausible theory for it. But a beguiling story can disguise weaknesses in the data.
4. The more factors or features in a data set the computer considers, the more likely the program will find a relationship, valid or not.

It is crucial to realize that data mining can involve a great deal of planning and preparation. Just having a large amount of data alone is no guarantee of the success of a data-mining project. In the words of one senior product manager from Oracle: "Be prepared to generate a lot of garbage until you hit something that is actionable and meaningful for your business."

This appendix is certainly not an inclusive list of all data-mining activities, but it does provide examples of how data-mining technology is employed today. We expect that new generations of data-mining tools and methodologies will increase and extend the spectrum of application domains.