Technical White Paper

Introducing BUSINESSMINER

Business Mining for Decision Support Insights

BUSINESS OBJECTS®

Author: Ceferino Lamb.

Contributors: Virginia Kunze, Pierre Chanty and Anne-Marie Burgeat.

Audience: This paper is intended for BusinessMiner customers and prospects worldwide.

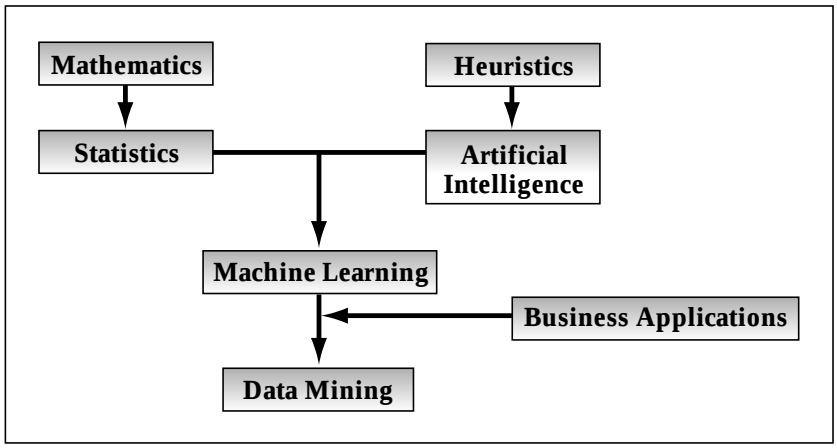
Contents

Introduction

Data mining represents an exciting new opportunity for mainstream businesses to benefit from advanced high technology. Business Objects believes that data mining has a well-earned place on mainstream business users' desktops. This paper introduces BUSINESSMINER, discusses the reasons why data mining offers additional analysis power to business users, summarizes the history of data mining, and describes some useful examples of applications for data mining. We also discuss some common data mining terms and technologies.

A Thumbnail History of Data Mining

Recently, software and business magazines have carried numerous articles on the subject of data mining. However, just a few years ago, very few people had even heard the term. Even though data mining is an evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s. This section reviews the history of data mining.



■ ***Figure 1 Data Mining family tree.***

► Statistics

Data mining traces its roots back along three family lines. The longest of these three lines is classical statistics. Without statistics, there could be no data mining, as statistics is the foundation of most technologies on which data mining is built. Classical statistics embrace concepts such as standard distribution, standard variance, regression analysis, standard deviation, cluster analysis, discriminant analysis, and confidence intervals, all of which are used to study data and data relationships. These are the building blocks with which more advanced statistical analyses are underpinned. Indeed, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

► Artificial Intelligence

Data mining's second family line is artificial intelligence, or AI. This discipline, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Because this approach requires overwhelming computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at moderate prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of everyone else. The notable exceptions were certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS).

► Machine Learning

The third and final family line of data mining is called machine learning, which is more accurately described as the marriage of statistics with AI. While AI was not a commercial success, its techniques were largely co-opted by machine learning. Machine learning was able to take advantage of the ever-improving price / performance ratios offered by computers of the 80s and 90s, and it found more applications because the entry price was lower than AI. Machine learning can be considered a further evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs "learn" about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.

► Machine Learning Meets Businesses

In many ways, data mining is fundamentally the adaptation of machine learning techniques to business applications. So we see that data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns therein. Today, data mining is finding increasing acceptance in sciences and businesses which need to analyze large amounts of data to discover trends which they could not otherwise find there.

Several main branches of data mining techniques exist:

- Decision trees
- Neural nets
- Statistics / time series analysis
- Visualization
- Rule induction

These techniques are described in the next section. Of the data mining techniques in use today, decision trees are of special interest, due to their ease of use, ability to decide between competing causal factors, and intuitive nature. The advantages of decision trees are discussed at length in the following section. But what are the origins of decision trees?

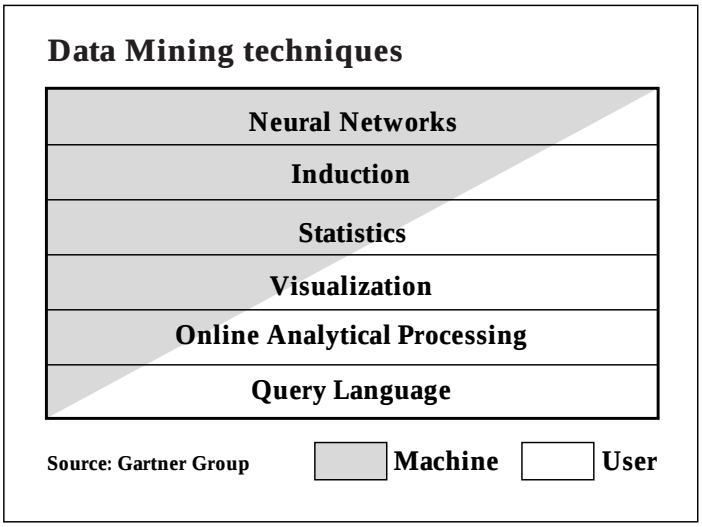
► History of Decision Trees

Decision trees are an evolution of techniques that arose during the development of machine learning disciplines. Decision trees grew from an approach to analysis called Automatic Interaction Detection (AID), developed at the University of Michigan. AID works by automatically testing all values in the data to identify those values which are strongly associated with the output item selected for examination. The values which are found to have a strong association are the key predictors or explanatory factors, usually called rules about the data. Another early algorithm named CHAID (Chi squared + AID) was developed by extending the capabilities of AID somewhat through the addition of a Chi squared statistical formula.

But it was a professor in Australia who developed the technology that allowed decision trees to sprout up. Many people in the data mining industry consider Ross Quinlan from the University of Sydney, Australia, as the "father of decision trees." Quinlan's contribution was a new algorithm called ID3, which he developed in 1983. ID3 and its later evolutionary siblings (ID4, ID6, C4.5, See 5) are well suited to use in conjunction with decision trees, as they produce rules ordered by their importance. These rules are then used to produce a decision tree model of the factors affecting the output item. Newer decision trees formulas such as Gini, a widely-used computational index invented by Ron Bryman, are also well-suited to decision trees, and offer increased computational speed plus broader abilities to process numbers concurrently with text.

Overview of Data Mining Technologies

Data mining is a field currently comprised of several main technology branches. Each type of technology has its own advantages and drawbacks, such that no one tool can meet all needs in all applications. So choosing the best technology for most of the anticipated uses is the best way to proceed in selecting a data mining technology. In this section we will discuss the various merits and applications of each main type of data mining, and show the business advantages of a decision trees solution.



■ *Figure 2 Data Mining harnesses more machine power.*

► Neural Nets

This technology is quixotic in that it offers perhaps the deepest data mining power, but is the most difficult to understand. Neural nets attempt to build internal representations of patterns found in data, but these representations are not presented to the user. With neural nets, the pattern discovery process is handled by the data mining program within a "black box" procedure. Software tools must then be built to make the decisions visible to the user. The problem with this approach is that decisions are made in a black box, which are inexplicable. While neural nets arguably offer the most advanced data mining power, many businesses cannot make use of it because the end results are not explained. Even experienced neural net engineers cannot fathom all of the workings of the system. Those businesses which must justify decisions cannot rely on neural nets to help make decisions. More broadly, business users are reluctant to trust black box tools which work in a non-understandable manner.

► Rule Induction

Rule induction refers to detection of trends within data sets, or "rules" about the data. The rules are then presented to users as an unordered list. Various algorithms and indices are available to accomplish this process, including Gini, C4.5, and CHAID. In rule induction, the vast majority of the discovery process is done by the machine, but a smaller part is done by the user. For example, translating the rules into a usable model must be done by the user, or by a decision trees interface. From a business user's point of view, the major problem with rules is that the data mining program does not rank the rules by importance. The business user is therefore forced to undertake a manual analysis of all reported rules in order to determine those which are most important in the data mining model, and to the business issue involved. This can be a tedious process.

► Decision Trees

Decision trees are a means of representing data mining results in the form of a tree, which resembles a horizontal organizational chart. Given a set of data with numerous columns and rows, a decision tree asks the user to pick one of the columns as the output object, then shows the single most important factor correlated with that output object as the first branch (nodes) of the decision tree. The remaining factors are subsequently classified as nodes under the master node(s). This means that the user can quickly see at a glance what factor most drives their output object, and that the user can understand why that factor was chosen. A well designed decision trees tool will also let users explore the tree at will, such that they can find target groups which interest them most, then zoom in on the exact data associated with their target group. Users can also select the underlying data at any node of the tree, moving it into a spreadsheet or other tool for further study.

► Statistics/Time Series Analysis

Statistics are the oldest technology in data mining, and part of the basic foundation of all other techniques. Statistics incorporate heavy user involvement, requiring skilled engineers, to build models that describe the behavior of data by way of classical mathematical methods. Interpreting the resulting models requires specialized expertise. In using statistical techniques, heavy machine/engineer collaboration is required. Time series analysis is of note because it is frequently confused with a simpler genre of data mining called forecasting. While time series analysis is a highly-specialized branch of statistics, forecasting is in fact a much less rigorous discipline which can be accomplished, albeit with less reliability, via most other data mining technologies.

► Visualization

Visualization techniques are somewhat difficult to define, because many people apply the definition to complex data visualization tools, while others use it for simple data graphing capabilities. In either case, visualization maps the data being mined according to specified dimensions. No analysis is performed by the data mining program beyond underlying clustering or basic statistical manipulation. A business user or data mining analyst then interprets the data while looking at the display. The analyst can then query the tool further to get different views or other dimensions.

► Advantages of Decision Trees

Decision trees are almost always used in conjunction with rule a induction engine. Decision trees are unique in that they present the results of rule induction in a prioritized format. So, the most important rule is presented in the tree as the first node. Less-relevant rules are presented as subsequent nodes of the tree. The primary advantages of decision trees are that they make decisions regarding which rules are most relevant, and they are understandable to most business persons. By choosing and presenting the rules in order of importance, decision trees allow business users to see at a glance which factors most drive their business. By their understandable and intuitive format, decision trees offer a means for easy data mining by average business users.

How Data Mining is Becoming Part of DSS

Decision Support Systems (DSS) have become a critical piece of many corporate Information Technology (IT) structures worldwide. A good DSS solution, in partnership with DBMS and data warehouse systems, allows everyday business users to have access to corporate information on their desktop, when they want it, either via ad hoc queries or custom reports generated by IS. As DSS tools become increasingly popular, they bring with them increased accessibility and ability for everyday users to bring corporate data to their desktop.

With this increased accessibility to data come greater challenges in using it all. When most mainstream business users can easily access large data warehouses, the DSS tool they use needs to enable more than just getting the data. Access to larger amounts of data means increased user needs for powerful, intuitive data analysis functions. To meet this need, DSS tools have evolved to include sophisticated analysis functions such as online analytical processing (OLAP), cross-tab multi-block reporting, 3D data visualization, and integration of data from several sources into one user document. All of these features help users master the greater amounts of data to which they have easy access.

► Manual Data Exploration

Of these functions, OLAP is arguably the most sophisticated, as it allows users to study data in a multidimensional manner, such that they can drill down to details or view summary slices of data, as they wish, while pursuing answers. OLAP falls solidly into the genre of a DSS analysis tool, or more precisely, a tool for user-guided manual data exploration and analysis. OLAP lets the user view data from numerous perspectives, and at numerous levels of detail or aggregation. This helps the user see items that either interest them directly, or lead them in the direction of data that interests them more. As the user navigates the multidimensional “cube” of data, they follow a manual, iterative process of data exploration and discovery.

► Automatic Detection of Trends in Data

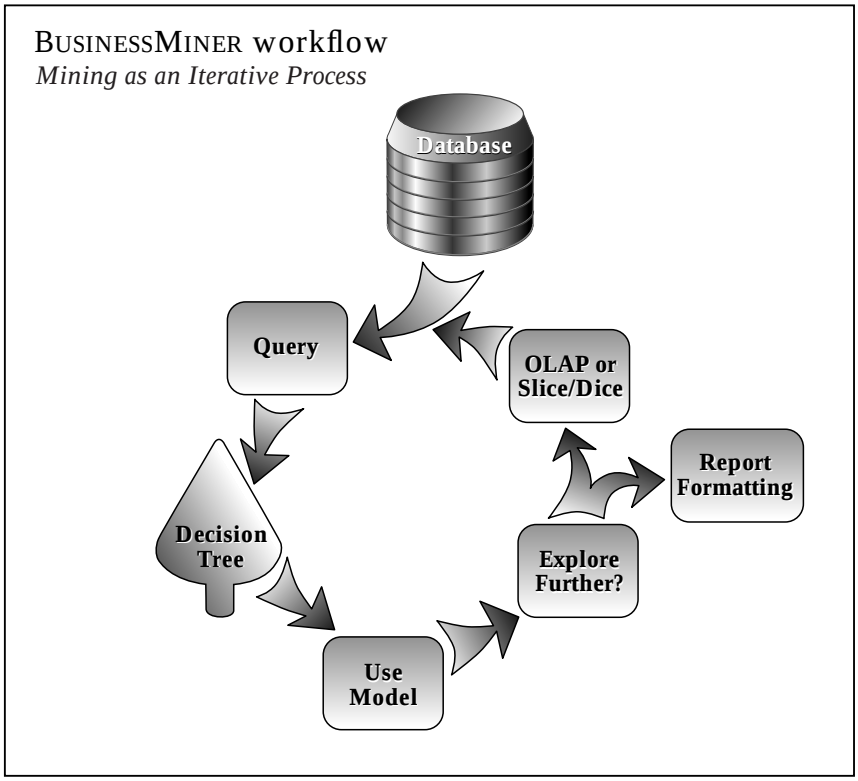
In comparison, data mining presents an alternative, automatic method of discovering patterns in data. Data mining is alternative in its approach, because it runs directly against all data in the data set, rather than following a given path through some of the data and drilling down to details. Data mining is automatic in its execution, because the tool itself studies the data and presents its findings to the user. While the user must still exercise reasonable care to provide useful data to the data mining tool, once that has been done the tool takes over and crunches through the data set on its own.

Because of this alternative approach and automatic execution, data mining is suitable to analyze data sets which would be difficult to fathom using OLAP - data sets which are too large to easily navigate manually, or which contain data that is too dense or non-intuitive to easily comprehend. Data mining does not require the user to steer the tool through the data set, or provide input to the process while underway. Given suitable data input, data mining can make sense of an overwhelming data set by finding hidden trends and presenting them to the user in an understandable format. This capability clearly adds value to a DSS solution.

► OLAP Meets Data Mining

In many cases, the results presented by data mining will raise interesting questions about the original data or related data in the warehouse. This is another example of when data mining adds value to a DSS solution. When the results of data mining pose additional questions, users can follow up in real time by simply running another query against the database, then mining again. Or, they may wish to use the results of data mining as a guide to help them navigate more data using OLAP. Often, users will find that they follow an iterative process of querying, data mining, and OLAP, until they get to the information that interests them most, at which time they can take advantage of the powerful reporting functions in a complete DSS solution.

So we see that data mining is becoming an essential new component for analysis within DSS, complementing existing abilities to perform querying, OLAP, charting, and advanced reporting. As users demand more ease of use and power from DSS solutions, data mining steps up to the task by providing an alternative, automatic tool for finding information that would be difficult to find using a less-complete DSS solution.



■ ***Figure 3 Data Mining with query and reporting and OLAP.***

Yesterday's Barriers to Effective Data Mining

Although data mining can now add value directly to DSS solutions, this was not true until very recently. In fact, there were several imposing barriers to effective data mining in the context of DSS. The most important have been overcome while others remain, albeit to a lesser degree than previously. Fundamentally, the most important historical barriers were high cost, requirements for large amounts of data housed on powerful servers, and the poor comprehensibility of existing data mining tools to anyone other than highly-skilled doctorate engineers. Other historical barriers included the challenge of preparing data for mining, difficulties in making an educated cost/benefit analysis before beginning a data mining project, and concern about the viability of many current data mining tool vendors.

► High Costs

The high cost of most data mining tools makes it difficult for enterprises to consider deploying them widely. Simple math shows that a $20,000 per user solution cannot be deployed to more than a small handful of selected business users. Certain data mining tool vendors have recently introduced products which cost less than $5000, but even this price still limits wide acceptance by most business users. Clearly, the per-seat costs need to be reduced before the benefits of this technology can reach mainstream business users.

► Too Much Data Needed

A more difficult obstacle to data mining was the need to house and service huge mountains of data on powerful and/or proprietary back office servers. This alone put a very high price tag on mere entry into the data mining market. Most vendors still promote an approach to data mining that requires terabytes of data on big servers, but more accessible solutions have recently appeared on the market, creating more affordable entry points into data mining. As most data mining analysts will agree, 80% of the value in a given data set can be found in 20% of that data, so it is logical that vendors will move increasingly to find that 20% and mine it. The trend toward personally-manageable data sets allows users to mine personal slices of data on their desktop, effectively skirting the large-database barrier. Although not intended to replace applications which require mining large data sets, desktop data mining provides an accessible alternative which can also be used in conjunction with back office data mining tools.

► Difficult to Understand

But even if newer tools allow fruitful mining on moderate amounts of data, a third serious barrier remained. Most current data mining tools are nearly incomprehensible to all but a small elite of highly-skilled engineers. In fact, many tools do their work in a black box, so that even the expert does not know how the tool reached its results. This situation meant that data mining was necessarily done in the context of an IT backlog, where business users had to submit requests, wait for days or weeks while the expert processed the data, then receive and review the summarized output. If the results did not satisfy the business user, the whole process had to be restarted. Fortunately, decision trees offer an understandable approach to data mining which lets anyone with basic office computer skills use and comprehend the process. This allows businesses to eliminate the IT backlog, putting data mining power directly on the business user's desktop. It also means that users can be confident that their results are justifiable, which is often a requirement for business decisions.

► Preparing Data to Mine

While the above three barriers have now been overcome, some others remain for almost all data mining tools. Preparing data for mining is widely considered to be 80% of the work of data mining. The data must be relevant to business needs, clean (free of egregious logical or data entry errors), consistent, and free of excessive nulls. The corporate move toward data warehouses has helped enable data mining, because warehouse data is normally clean and centrally located. But even clean data still needs preparation to mine, as choosing the right data to mine is critical. This is an area where a business mining tool offers a unique advantage, because it can harness DSS query power and its business-aware semantic layer. Users can therefore use familiar business terms to easily select exactly the data that interests them most, then start the data mining module, effectively mining a slice of pre-qualified data on the desktop. This process of pre-selecting only the most useful data is called partitioning in the data mining field. The process of selecting a smaller set of data which accurately represents a much larger set is called sampling. By using query and semantic layer powers to partition and/or sample raw data, business mining overcomes the data preparation barrier.

► Estimating Project Costs and ROI

Deciding to embark on a data mining project can be risky if one uses traditional data mining tools. Because of the high entry costs explained above, it is difficult to forecast a reasonable return on investment (ROI). A mid-size IT shop could easily invest hundreds of thousands of dollars just to get a traditional data mining project into early stages. Estimating ROI is complicated due to the fact that data mining tools are meant to discover trends in data which are not otherwise visible. It is virtually impossible to estimate the ROI of something unknown. That is why a reduced entry price is critical to the conception of any data mining project. A business mining solution lets users get started in data mining for a very low price. As users become familiar with data mining concepts and benefits, this means that the enterprise can make a much more informed cost/benefit decision before proceeding with wider deployment.

► Viability of Data Mining Vendors

Finally, the market viability of many current data mining vendors is a concern for enterprises seeking a reliable tool for today and the near future. The data mining market is fully populated with dozens of firms ranging from tiny startups who do nothing but one tool, to large firms for whom data mining is only a sidebar. As with any nascent technology, choosing a vendor is as important as choosing a tool. Today's leading-edge technology could be tomorrow's struggling acquisition target, or worse, business failure. It is important to choose a vendor that is stable, established, and has a clear focus and vision on data mining in the DSS context.

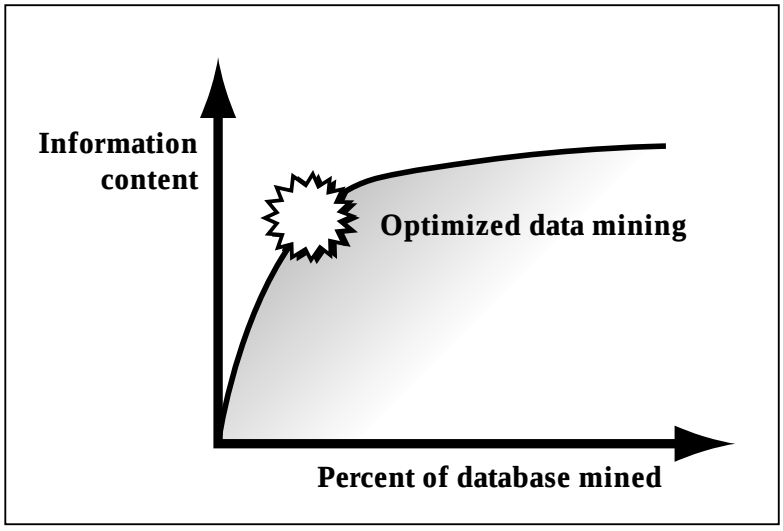
How Much Data is Needed to Perform Data Mining?

While classical data mining tools require huge amounts of data, business mining can be performed usefully on much smaller sets of data. Keeping in mind minimum statistical viability, data sets for data mining can be substantially reduced due to the 80/20 rule, good understanding of the specific business need to address, and the desired scope of analysis.

► The 80/20 Rule

Analysts agree that within any large data set, 80% of the information can be found within 20% of the data. This rule augers to reduce the needed size of the data set analyzed. To further reduce the needed data set size, partitioning of the data should be performed. This means that only more relevant data is selected for mining, further concentrating the useful data into the 20% selected for analysis. If the partitioned data set is still very large, it can be sampled to take a representative sample from the larger set.

■ ***Figure 4 80% of information can be found in 20% of data.***



► Minimum Statistical Viability

When attempting to minimize needed data sets, it is important to keep in mind that a minimum number of records is needed to have statistical viability. In general, a minimum of two hundred records should be analyzed in order to have statistically-viable results. This is a record size well within the scope of business data. Larger data sets can be assumed to have statistical relevance - but this can also be verified by analysis functions offered by a qualified business mining solution.

► Understanding Specific Business Needs

In the business context, data mining finds suitable applications in studying specific aspects of a business. We refer to this specific business aspect as the scope of analysis. For example, a bank branch manager is almost certainly more concerned with the customers of his or her branch than with the nationwide or statewide customer base. So the branch manager's scope of analysis would be the local branch, or possibly the local region, in order to compare local results with regional results. A local or regional scope of analysis lends itself perfectly to study by a business mining solution. Large amounts of data are not required for significant results.

Why Business Data is Easier to Mine

Business data presents unique opportunities for data mining. Compared to other types of data such as scientific, actuarial, or statistical data, business data is more homogeneous and more intuitive. Business data also lends itself more easily to aggregation, which can often reduce the amount of raw data needed for a given operation. Finally, business data is owned and maintained by business people who understand its significance. For modern computerized businesses, their data is the heart of their business. Non-business data is more often compiled by a remote process or persons, then transferred to other analysts for further processing, reducing their ability to understand the meaning of each aspect of the data.

► Business Data is Predictable

Predictability is a key factor in rendering business data more mine-able. Business data is collected within the framework of a particular business, describing, for example, the customers of that business. Clean business data will tend to contain values which fall within certain reasonable ranges. It is highly unlikely, for example, that a car dealer would sell a new car for two hundred dollars, or two million dollars. Car prices will tend to fall into a fairly predictable range. Likewise, it is unlikely that the same car dealer will sell cars to persons who reside in remote countries or pay in foreign currencies. Business data about car sales transactions will tend to describe sales to local residents, who paid in the national currency. Because business data has less exceptions, it is easier to use data mining in order to spot trends. Exceptional values in a data set make it more difficult to mine. In data mining terms, exceptional values are called noise. Business data benefits from reduced noise.

► Business Data is Intuitive

The intuitive nature of business data is another enabler for data mining. While scientific data is likely to contain inscrutable values and great minutia, business data is at the other end of the scale. Business data describes businesses, and is named and compiled by and for business persons. Terms such as revenue, expenses, response rate, inventory level, and credit limit are intuitive terms, and the data contained for those terms makes sense in a business context. Business persons would know intuitively that a credit limit value is likely to be an amount of money, and that a response rate value is likely to be a percentage. The fact that business data is intuitively understandable to business persons is a great advantage for data mining, leveraging the amount of knowledge that can be gleaned from a small or moderately-sized data set. In data mining terms, intuitive data enjoys what is called native data cognizance.

► Aggregation can Speed Data Mining

Business data is often stored in aggregated formats, such as revenue by quarter, sales per region, or promotion responses per zip code. These formats can be much easier to mine than raw data. On the other hand, non-business data is much less likely to support summarization of data. One useful way to think of these business aggregation formats is as a lever, which allows useful data mining on a much smaller data set than would be possible with scientific data. By mining the aggregates, business users can discover trends in their business at any level they wish. Mining on regionally aggregated revenue per advertising dollar spent in various media would describe what type of advertising is most cost effective in each region. There would be no need in this case to mine every sale nationwide. A modern DSS tool set lets businesses easily build aggregated objects for general use, and a business mining tool can use those objects very effectively to leverage data mining.

► Data Users are Data Owners

Another significant advantage of business data is that it is almost always collected, maintained and owned by the same people who use it, i.e., business people. In contrast, scientific data is often made up of stratified-source data, such as samples collected in field research, then sent to headquarters for compilation by another team of analysts. In simplest terms, this means that one person's output becomes another person's input, and the second person is less likely to fully understand the data. Business data is much less likely to suffer this loss of understanding. For example, a marketing manager is certain to understand the meaning of sales figures by product in their region. An analyst mining customer data to find out what drives customer loyalty is sure to grasp the meaning of each item of data about customers, because that analyst is involved and familiar with the customer data.

Introducing BUSINESSMINER

BUSINESSMINER is the world's first data mining tool which brings advanced mining technology to the desktops of mainstream business users. Available as an option to BUSINESSOBJECTS 4.0, BUSINESSMINER is a powerful new analysis technique that is called directly from the BUSINESSOBJECTS toolbar. Users typically use BUSINESSOBJECTS to select objects which interest them, run the query, then study the results or distribute the report to other users. If users want greater analysis power, OLAP is available for manual data exploration. If users find that the report is difficult to understand even with slice and dice and/or drill, BUSINESSMINER is available for automatic data exploration. BUSINESSMINER will be able to find trends in data which are buried too deeply to be found with OLAP, then display these trends as a decision tree. For greatly improved performance, BUSINESSMINER is tightly integrated with BUSINESSOBJECTS, such that no duplicated data sets or temporary files are used - which also saves system resources. BUSINESSMINER is designed to offer intuitive ease of use, giving users a comfortable, familiar BUSINESSOBJECTS look and feel which is Microsoft Office95 compliant.

► Automated Analysis Power

BUSINESSMINER introduces an automated analysis tool to the DSS desktop. When users are faced with slices of data which do not lend themselves well to OLAP manual data exploration, BUSINESSMINER offers automated exploration to find trends in the data. BUSINESSMINER delivers accurate results automatically, using the full power of its modern rule induction engine. The results are presented to users as a decision tree, which can be built automatically or interactively. BUSINESSMINER brings DSS the needed next logical step beyond OLAP; automated analysis of data.

► Easy and Understandable

BUSINESSMINER offers decision trees, a data mining technology which is easily understood by mainstream business users. Decision trees are the most intuitive data mining technology available, because they present the results of data mining in the form of a simple visual tree. In addition to this inherently understandable paradigm, the BUSINESSMINER user interface was conceived and designed for maximum ease of use. To achieve this, BUSINESSMINER has a familiar Office95 look and feel, wizards which let users feel at ease mastering multi-step procedures, and alerters to clearly highlight any mining results which interest the user. Business users do not need specialized experience or a Ph.D. to use BUSINESSMINER, because its reliable algorithms require no tuning yet still provide full data mining functionality.

► Integrated with BUSINESSOBJECTS

Because DSS users need additional, automatic analysis power on the desktop, BUSINESSMINER is tightly integrated with BUSINESSOBJECTS, and features a familiar BUSINESSOBJECTS interface. This integration allows users to take advantage of the unique BUSINESSOBJECTS patented semantic layer that represents complex database concepts as familiar business terms. In addition to removing the complexity from query and reporting, the semantic layer offers leveraged data mining power, allowing users to data mine rich slices of business data which were selected easily for analysis via BUSINESSOBJECTS. Users can easily and directly mine business aggregations and complex objects which are specific to their business, thereby avoiding hours, days or weeks of tedious data preparation. BUSINESSMINER sidesteps the 80/20 rule which we discussed earlier, letting users get directly to most of the information in their data without spending eighty percent of their time getting and preparing data to mine.

► Workflow Using BUSINESSMINER

BUSINESSMINER is an additional analysis option available to BUSINESSOBJECTS users. As such, users can easily interact with BUSINESSMINER while doing data analysis with BUSINESSOBJECTS. Typical workflow is that the users select objects to study using the BUSINESSOBJECTS query panel, then runs the query. At that point users can choose from several analysis options, which are slice and dice, OLAP, and BUSINESSMINER. If users wish to use automated analysis to study the data, they simply call BUSINESSMINER via a button in the BUSINESSOBJECTS toolbar. All data is then passed directly to BUSINESSMINER, which automatically launches its rule induction engine and analyzes the data. At this point users are free to explore the data mining results via intuitive decision trees. Users may then elect to either display the decision tree node by node, following the data discovery path that interests them most, or let BUSINESSMINER display a complete decision tree automatically. At this point all basic data mining features are available.

► BUSINESSMINER Features

• **Modeling:** BUSINESSMINER performs this task automatically as soon as it receives the data to analyze. No user intervention is required to build the model. Once the model is built, BUSINESSMINER has discovered the rules describing trends it found in the data. BUSINESSMINER will use the model for all further operations.

• **Discover Rules:** BUSINESSMINER lets users easily discover rules about their data. This is a simple way of listing target groups within the data which are of particular interest to users. A good example of this would be letting BUSINESSMINER display an ordered description of all groups of customers who are likely to generate high profitability.

• **Visualization:** this is a quick way to profile groups of data as a chart. Numerous types of charts are available, including bar charts, scatter charts, and line charts. Visualization can be a fast and powerful tool to determine how predictable the data is. If there are wide ranges of values in the data, or exceptional values which might warrant further investigation, visualization will make this immediately apparent. On the other hand, if the data falls within a predictable range, that will also be easy to see by using visualization.

• **What-If?:** this useful feature lets users predict results based on actual data in the model. A practical example of this would be to use What-If? in a telesales office, allowing staff to quickly predict the creditworthiness of prospective new customers before closing the sale or delivering the product.

• **Segmentation:** segmentation allows users to list an exact description of a target group found in the decision tree. This means that once an interesting group of data has been found in the model, segmentation can be used to display all known characteristics about that group. A business example of this would be to locate a group of customers which BUSINESSMINER has determined are eighty percent likely to defect to a competing business. Segmentation would then provide a description of that customer group, so that a list of similar customers could be purchased from a mailing list company. Having found a profile of customers who are likely to defect, and purchased a list of names and addresses of similar individuals, a business could target market to similar current customers of a competitor, knowing that they are more likely to defect from that competitor. A complementary feature allows actual records to be copied from BUSINESSMINER into another tool such as Excel. This could be used to generate a list of existing customers who are likely to defect, so that they could be offered inducements to stay.

► The BUSINESSMINER Engine

At the heart of BUSINESSMINER lies a robust rule induction engine. The engine examines the data carefully via an extremely fast search, induces rules about the data which describe trends found therein, then stores these rules internally as an ordered index which will be used to build decision trees. The BUSINESSMINER engine uses the proven Gini index (also used by CART), then employs additional algorithms designed to improve performance and capabilities. This means that BUSINESSMINER can build decision trees as the impressive rate of one thousand rows of data per second. Another benefit is the capability to analyze both text and numbers with the same power, which is important for business applications. Finally, unique formulas also allow BUSINESSMINER to perform intelligent binning. This means that data mining results are more reliable and meaningful than they would be with mere manual or automatic binning, no matter what type of data is analyzed.

How can Mainstream Businesses use BUSINESSMINER?

We have showed how business data lends itself to data mining, but we have not yet discussed specific examples of how mainstream businesses can use BUSINESSMINER. We will examine here the cases of managing customer relationships, targeted marketing, prioritizing spending, and credit risk analysis. While this is a short list, it should give good ideas which can be applied to other mainstream uses for BUSINESSMINER.

► Managing Customer Relationships

For the example of managing customer relationships, let's agree that our goal is to determine the profiles of loyal customers, and those who are most likely to switch to a competitor. This is a very important issue for many businesses - losing customers to competitors is sometimes called churn. We could profile loyal customers, or those likely to churn, by launching BUSINESSMINER on the output of a BUSINESSOBJECTS 4.0 query which contained numerous facts about our customers. Then we would set an object describing customer loyalty as the output object, to find out which related customer facts are most influential. BUSINESSMINER would build a decision tree which shows the most important factor associated with customer loyalty. From the same tree, we could determine the least loyal customers, and use that profile to find similar customers who have not gone to the competition yet. From that point, we might decide to target loyal and churn risk customers for entirely different marketing approaches.

► Targeted Marketing

Using targeted marketing, businesses can cut costs and boost results. As businesses grow, they find that marketing costs escalate as they try to reach more prospects. Targeted marketing attempts to reach out only to those prospects who are more likely to respond to an offer. In this example, we could find the best new prospects by data mining on earlier marketing campaign results. We could do this by launching BUSINESSMINER on the output of a BUSINESSOBJECTS 4.0 query which contained numerous facts about our last marketing campaign results. Here we would simply set BUSINESSMINER to find which customer facts most influenced response. We would then use the Discover option to find types of customers that interest us most, or use the Segment option to exactly describe the types of customers which BUSINESSMINER found to respond best. Using this knowledge, we would target our next marketing campaign carefully, saving on costs and improving results.

► Prioritized Spending

Another real world example of BUSINESSMINER uses is to prioritize spending. In this case, we assume that a business wishes to determine the most important factor in success, such as the telemarketing closing rate, then invest more where it matters most. This could be done by letting BUSINESSMINER study BUSINESSOBJECTS 4.0 query results which lists average telemarketing revenue by telemarketing representative, by sales region, by lead generation type, and related customer facts. Then we would set BUSINESSMINER to find out what factors most drive revenue. We would be able to determine these factors by representative, by sales region, and by lead generation type. We might find that direct mail works much better in some regions than others, or that targeted cold calls to a certain purchased list are markedly more effective than targeted cold calls to other purchased lists.

► Analyzing Credit Risk

For the example of credit risk analysis, we agree to find out what types of loan applicants pay fully and promptly, or which applicants are more likely to default. This could be done by running BUSINESSOBJECTS 4.0 to find numerous facts describing your customers, such as their existing credit limit, their marital status, whether or not they own a home, how many children they have, their monthly disposable income, how reliably they have repaid their loan(s), and other related facts. Next we would call BUSINESSMINER and set loan repayment reliability as the output object. BUSINESSMINER would then build a decision tree which showed the customer facts that are most closely associated with repaying loans on time. From the same decision tree, we could also see which types of customers are likely to fail to repay their loans. BUSINESSMINER offers many tools to further zoom in on these two types of customers.

Appendix A: Glossary of Data Mining Terms

AID: an early data mining algorithm developed at the University of Michigan. AID means Automatic Interaction Detection. AID handled symbolic values only. See CHAID, ID3, symbolics.

Algorithm: complex mathematical formulae at the heart of all data mining tools. See AID, C4.5, CART, CHAID, Gini, ID3, ID4, ID6, see 5.

Artificial Intelligence (AI): a forefather of data mining. AI is based on heuristics, as opposed to statistics. Some AI techniques were adopted in RDBMS, military, and scientific applications. As AI matured and expanded to embrace statistics, it became known as machine learning. See machine learning, heuristics, statistics, data mining.

Association: when one data item is found to be closely related to another data item, or cause another data item, we say that they are associated. Association refers to finding those associated data items. Note that association does not necessarily mean that one data item causes the other data item.

Automatic Binning: binning which sets the number of bins based on the range of a numeric value. Therefore, the user is not required to specify the number of bins. However, certain values may be "lost" from the decision tree because of automatic binning, which is not the case with intelligent binning. See binning, intelligent binning.

Binning: choosing the number of bins into which a numeric range is split. For example, if salaries range from $20,000 to $100,000, the values must be binned into some number of groups, probably between eight and twenty. Many data mining products require the user to manually set binning. See intelligent binning, automatic binning.

Black Box: any technology which does not explain its results. Users cannot find out how the answer was determined. This renders some data mining technologies unsuitable for many business applications. See neural nets.

C4.5: a data mining algorithm which was developed from ID3, ID4, and ID6. C4.5 handles both numerics and symbolics well. See ID3, Gini, numerics, symbolics, see 5.

CART: a chi squared statistical regression algorithm used for classical statistical analysis. CART stands for classification and regression trees. CART can be used to build decision trees, in which case it can also use the Gini index. CART can only process numeric values effectively. See statistics, CHAID, Gini, numerics, symbolics.

Causal Factor: any data item which drives, influences, or causes another data item. For example, if customer credit limit drives how profitable a customer is likely to be, it is called a causal factor. See discriminating factor.

CHAID: a hybrid algorithm which grafts a chi squared statistics formula onto AID (heuristics), in an attempt to handle both numerics and symbolics. While CHAID is reliable, it is slow and limited in power. See AID, CART, Gini, statistics, heuristics, numerics, symbolics.

Confidence Window or Level: a statistical measurement of how sure one can be that a certain result is true. The window or level describes how close the value is likely to be to the exact result. See statistical significance.

Data Mining: the automatic detection of trends and associations hidden in data. DM is part of a larger process called knowledge discovery. Data mining can also be described as the application of machine learning techniques to business applications. See machine learning, association.

Decision Trees: a data mining technology which determines causal factors ranked by their importance, and presents them in the form of a tree, made up of a root, branches and leaves. Decision trees are similar to organization charts, with statistical information presented at each node.

Diapers and Beer: an popular anecdote which describes the power of data mining. The anecdote (probably apocryphal) recounts that a large supermarket chain used data mining to discover that customers often bought diapers and beer at the same time. This encouraged the retailer to display the two items together, increasing sales of both.

Discovery: finding trends and associations hidden in data. See modeling, associations, rule induction.

Discriminating Factor: a measure of how important a causal factor is, used by decision trees to build the tree. See decision trees, causal factor.

Forecasting: adapting data mining techniques to forecast future trends with statistical reliability. Forecasting is often confused with prediction, but is usually much more complex. See time series analysis / forecasting, what-if analysis, neural nets.

Gini: a modern decision tree index algorithm which was developed by Ron Bryman. Gini handles both numbers and text, and offers good processing speed. See C4.5, CHAID, ID3, numerics, symbolics.

Grouping: similar to binning, but for symbolics. The main difference from binning is that the user can manually ungroup values. In a decision tree, grouping is done based on the discriminating factor. See binning.

Heuristics: formulae which are based on artificial intelligence principles such as entropy theory, rather than statistical principles. Heuristics were the first algorithms which successfully process text values (also called symbolics). See statistics, symbolics.

ID3: the first algorithm which was designed to build decision trees. ID3 was invented by Ross Quinlan at the University of Sidney Australia. ID3 was followed by ID4, ID6 and see 5. See C4.5, Gini, CHAID, CART.

Intelligent Binning: a unique BUSINESSMINER feature which automatically and intelligently bins numeric values based on the range, values and distribution of the data. Intelligent binning addresses a historical criticism of decision trees; certain values may be lost from the tree due to binning. Note that simple automatic binning does not read values and bin accordingly.

Machine Learning: as AI progressed, it incorporated technologies from classical statistics. This marriage produced useful technology advances, and came to be known as machine learning.

Market Basket Analysis: a technique, used in large retail chains, which studies every purchase made by customers to find out which sales are most commonly made together. This is the basis of the (possibly false) anecdote about diapers and beer. See diapers and beer.

Modeling: building a model which describes the trends and associations discovered. This model lets users explore the trends and associations to understand them better. See data mining, associations.

Neural Nets: a very powerful but complicated data mining technology, which attempts to mimic the complex reasoning functions of the brain. The main problem with neural nets is that the tools do not explain how they determined their results. Another limitation is that only skilled professionals can successfully use them. See black box.

Numerics: data in number format, i.e., numbers. BUSINESSMINER handles numbers as well as symbolics, but some other data mining tools do not. See symbolics, CHAID, Gini.

Overfitting: the tendency to mistake noise in data for trends. For example, a certain typographical error which is frequently made during data entry may be modeled by the data mining tool.

Partitioning: choosing data which is most interesting for mining. This is typically at least eighty percent of the work of data mining. BUSINESSMINER offers business-aware partitioning due to the power of the BUSINESSOBJECTS 4.0 semantic layer. See sampling.

Prediction: using existing data to predict how other factors will behave, assuming that some facts are known about the new factor. Making a credit check of new customers by using data on existing customers is a good example of prediction. See What-If? analysis, time series analysis / forecasting, forecasting.

Regression Analysis: a statistical method of doing time series analysis / forecasting and some aspects of data mining. See time series analysis / forecasting.

Rule Induction: a method of performing discovery by inducing rules about data. Rule induction tests given values in the data set to see which other data are its strongest associated factors. See decision trees, discovery, causal factor.

Sampling: taking a random sample of data in order to reduce the number of records to mine. Sampling is statistically complicated, but can be done in an RDBMS by use of a simple random number generator and column in the database. See partitioning.

See 5: the most recent of the Quinlan series of decision tree algorithms.See Algorithm C4.5, ID3.

Segmentation: an exact description of a target group found by data mining. BUSINESSMINER offers segmentation as one of its basic features. Segmentation is extremely useful for targeted marketing and other applications which require a precise description of certain client groups.

Simple Forecasting: see prediction.

Statistical significance: a measure the statistical likeliness that a given numerical value is true. See confidence window or level.

Statistics: based in advanced mathematics, statistics are one of the basic building blocks of data mining. Statistics incorporate formulae developed over the centuries, adapted to modern needs. Not to be confused with heuristics, which study non-mathematical formulae. See heuristics.

Symbolics: data in text format, such as ASCII or varchar. BUSINESSMINER handles symbolics as well as numbers. See numerics, CHAID, Gini.

Time Series Analysis/Forecasting: a complicated technology which is used to give statistically accurate forecasting. This is often confused with prediction or simple forecasting, but time series analysis / forecasting is much more difficult, and mathematically based. See forecasting.

Verification: the use of an alternate technology or tool to verify the results of another data mining technology or tool. For example, OLAP may be used to verify the results of data mining.

Visualization: presenting the results or intermediate steps of data mining in visual formats such as charts and graphs so that users can see patterns. BUSINESSMINER offers visualization as one of its basic features.

What-if analysis: a method of doing prediction or simple forecasting, based on variable input from the user. BUSINESSMINER offers what-if analysis as one of its basic features. See prediction, forecasting, time series analysis / forecasting.

Americas
Business Objects Americas
2870 Zanker Road
San Jose, CA 95134
USA
Tel: +1 408 953 6000
+1 800 527 0580
Fax: +1 408 953 6001

Europe
Business Objects S.A.
1, square Chaptal
92309 Levallois-Perret cedex
France
Tel: +33 1 41 25 21 21
Fax: +33 1 41 25 31 00

Australia
Business Objects Australia Pty Ltd.
Suite 210, 283 Alfred Street North
North Sydney, NSW 2060
Tel: +61 2 9922 3049
Fax: +61 2 9922 3069

Belgium
6, Minervastraat
1930 Zaventem
Tel: +32 2 720 0085
Fax: +32 2 720 7121

Canada
Business Objects Canada Inc.
3 Robert Speck Parkway
Suite 900
Missisauga, Ontario L4Z G5
Tel: +1 905 306 7575
Fax: +1 905 306 7825

France
Business Objects S.A.
Tour Chantecoq
5, rue Chantecoq
92808 Puteaux cedex
Tel: +33 1 41 25 21 21
Fax: +33 1 41 25 21 20

Germany
Business Objects GmbH
Kölner Straße 259
D-51149 Köln - Porz
Tel: +49 2203 9152-0
Fax: +49 2203 9152-100

Italy
Business Objects Italia
Via Laurentina, 756
00143 Roma
Tel: +39 650 1 59 83
Fax: +39 650 1 59 79

Japan
Business Objects Nihon BV
5F Ryoshin Ginza Building
3-4-15 Ginza, Chuou-ku
Tokyo 104
Tel: +81 3 3561 2350
Fax: +81 3 3561 2360

Netherlands
Business Objects Nederland BV
“KCN Tower”
Nevelgaarde 47
3436 ZZ Nieuwegein
Tel: +31 30 60 22 934
Fax: +31 30 60 22 963

Singapore
Business Objects Asia Pacific Pte. Ltd
8 Robinson Road #08-00
Cosco Building
Singapore 048544
Tel: +65 270 4596
Fax: +65 276 5410

Spain
Business Objects España
Avenida de Burgos 12 - 4a planta
28036 - Madrid
Tel: +34 1 766 87 43
Fax: +34 1 766 87 78

Sweden
Business Objects Nordic AB
Kungsgatan 59, 4 floor
111 22 Stockholm
Tel: +46 8 545 120 30
Fax: +46 8 545 120 40

Switzerland
Business Objects Switzerland SA
16, Chemin des Coquelicots - Tour A
1214 Vernier - Genève
Tel: +41 22 306 09 00
Fax: +41 22 306 09 09

United Kingdom
Business Objects (UK) Ltd.
Objects Court
29-41 Moorbridge Road
Maidenhead
Berkshire, SL6 8LT
Tel: +44 1628 764600
Fax: +44 1628 764601

Distributed in:
Argentina
Australia
Austria
Barbados
Bahrain
Belgium
Brazil
Chile
China
Colombia
Denmark
Finland
Greece
Hong Kong
Iceland
India
Japan
Kuwait
Lebanon
Luxembourg
Malaysia
Mexico
Morocco
Netherland Antilles
New Zealand
Norway
Peru
Poland
Portugal
Puerto Rico
Russia
Saudi Arabia
Singapore
South Africa
Sweden
UAE
Venezuela

World Wide Web
http://www.businessobjects.com