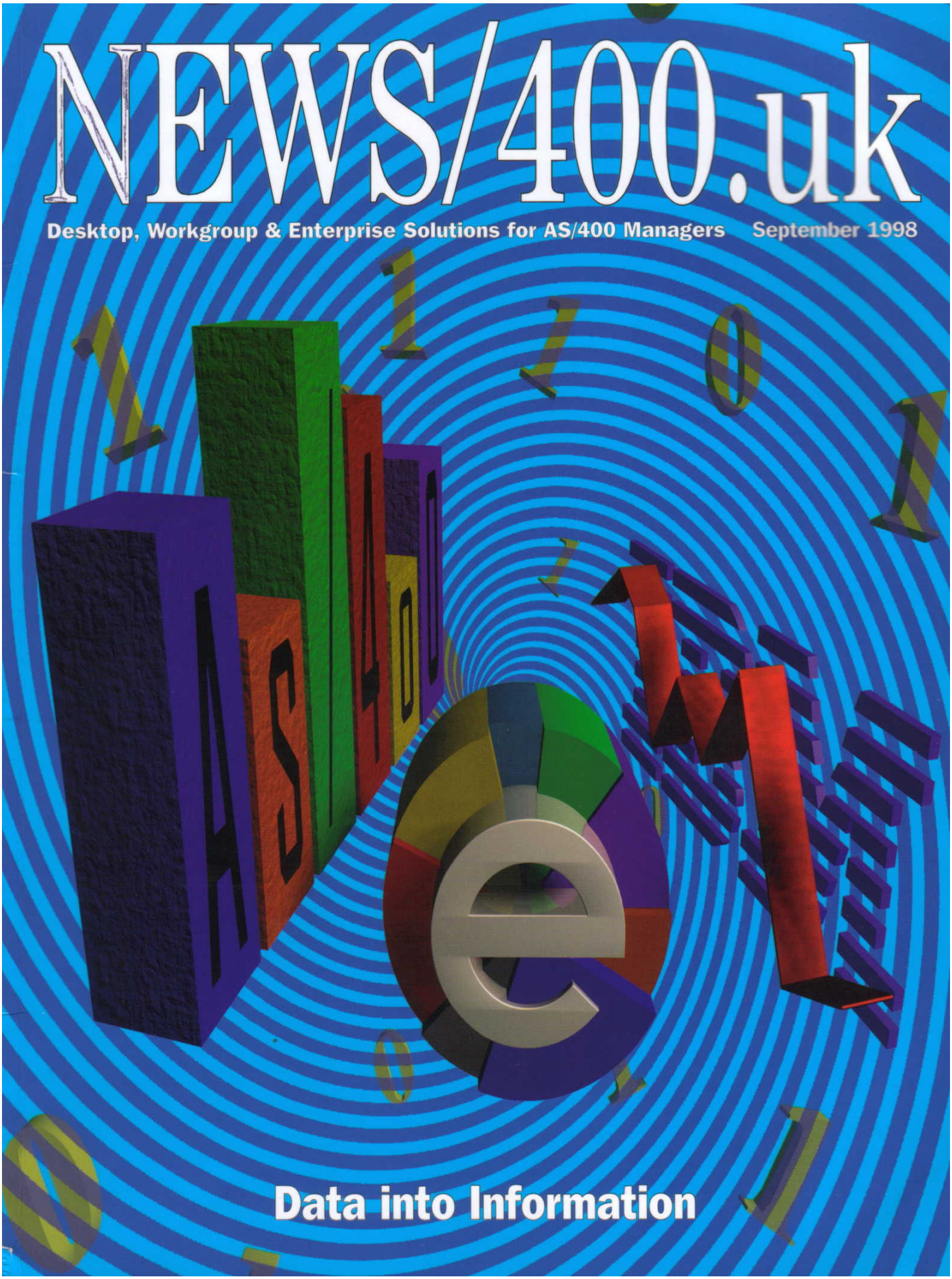
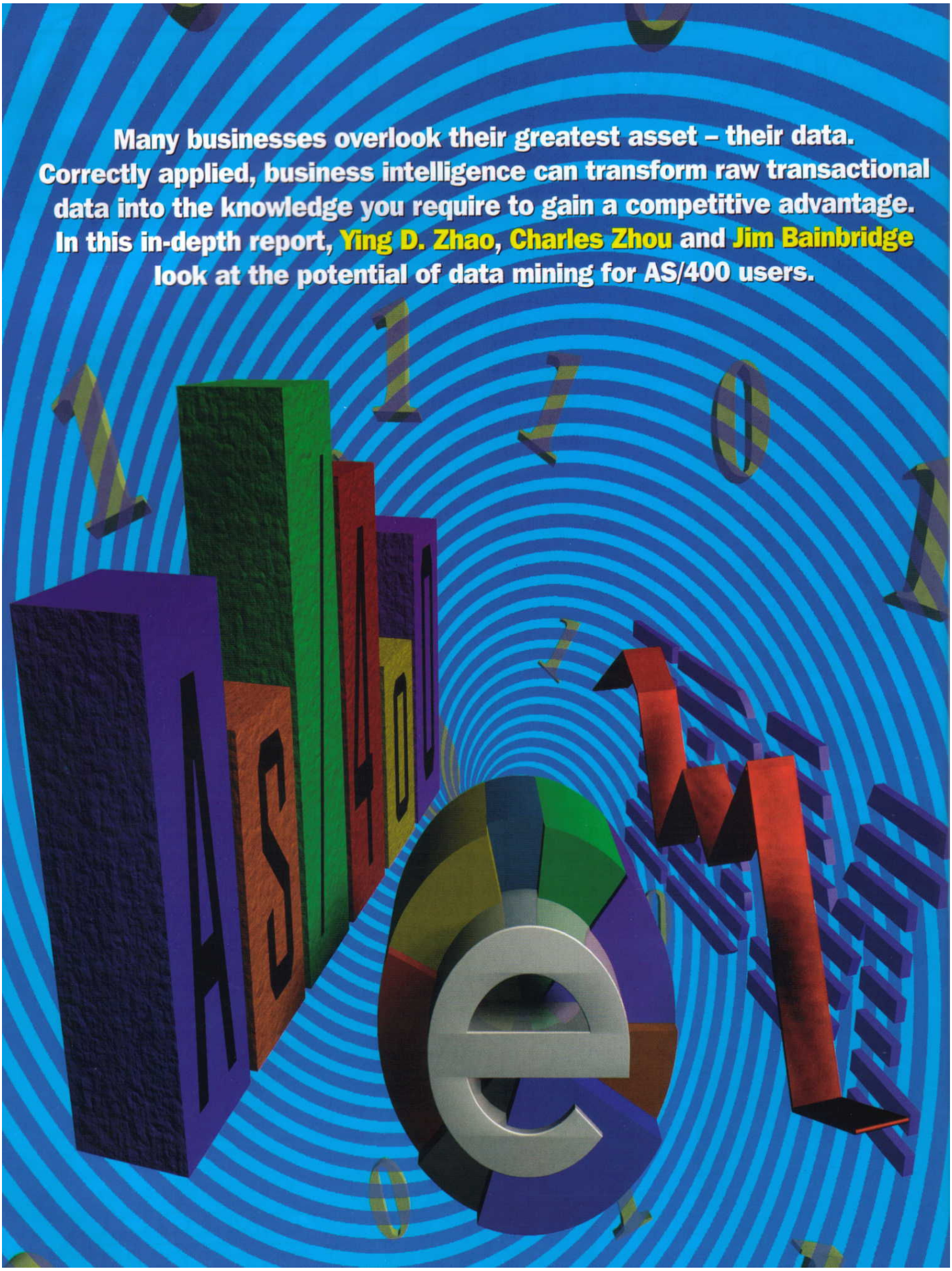
NEWS/400.uk

Desktop, Workgroup & Enterprise Solutions for AS/400 Managers September 1998



Data into Information

Many businesses overlook their greatest asset – their data. Correctly applied, business intelligence can transform raw transactional data into the knowledge you require to gain a competitive advantage. In this in-depth report, Ying D. Zhao, Charles Zhou and Jim Bainbridge look at the potential of data mining for AS/400 users.



Data into Information

Applied intelligence

Businesses are constantly and aggressively seeking new ways to differentiate themselves and to grow market share in an increasingly competitive, global and deregulated environment. At the same time, companies are generating enormous amounts of data as they conduct their day-to-day business. Business Intelligence is the process of transforming this data into a useful format that can provide a competitive advantage. It is a process of automatically capturing and analysing business data, discovering hidden patterns, and then going to market with this new-found information. This process allows companies to make strategic decisions about which markets or segments of a market to enter, which customers to target, or which products to promote.

In *Discovering Data Mining: From Concept to Implementation* (an IBM data mining red book by Cabena, Hadjinian, Stadler, Verhees, and Zanasi, Prentice Hall PTR, 1997), data mining is defined as "the process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions." This process is recognised as a key component of a successful business intelligence endeavour.

The AS/400's inherent features and its positioning as a widely used midrange server create a unique opportunity in data mining for the platform. Significant characteristics of the AS/400 are its integrated DB2/400 database and its high degree of robustness and reliability. In addition, the AS/400 is found in every size company in every industry. Of special note are the retail and distribution industries, segments that require intensive and intelligent data management capabilities. Currently, there are over 550,000 AS/400s with millions of users attached, and all these systems have a known entity – DB2/400.

Known quantities

Many AS/400 customers are middle-sized companies with under 500 employees and revenues under £500 million. Due to the integrated, self-managing nature of OS/400 and the large selection of off-the-shelf applications available from ISVs, many of these companies do not have large IT departments. These companies are constantly looking for ways to gain competitive advantages, and often turn to new technologies to address this problem.

Many of these customers initially purchased their AS/400s to run the production side of their businesses, relying on the system for things like order entry, scheduling, general ledger, and customer billing. Gradually, many are incorporating data marts or enterprise data warehouse applications to better understand and leverage the information contained in their data assets.

Several factors are encouraging existing customers to extend the function of their AS/400s by investing in AS/400 business intelligence capabilities. First of all, many AS/400 customers are highly motivated to apply business intelligence to gain competitive advantages and grow their businesses. In contrast to larger companies, the marketing and IT professionals of smaller companies are relatively well connected. In many cases, marketing and sales departments in these companies find it easier to work together to determine their business intelligence needs. As a result, the whole process from concept to implementation can be realised within a much shorter cycle. In other words, these companies foster a business environment that is conducive to the successful achievement of business intelligence objectives.

Furthermore, many AS/400 users have very large databases to draw upon. The AS/400 is a very database-oriented platform, and analytical algorithms can be run directly against the DB2/400 data residing on the AS/400 database server. Integration of analysis and database technology on the AS/400 is much more feasible than for some other platforms.

A third reason for investing in the data mining capabilities of the AS/400 is that IBM's AS/400 Division has made significant commitments and efforts to push the advancement of business intelligence capabilities on the AS/400 platform.

Business values

There are tremendous benefits for companies who invest in the cutting edge of business intelligence to transform data into useful information. For example, data mining is increasingly being deployed to address business problems in the area of database

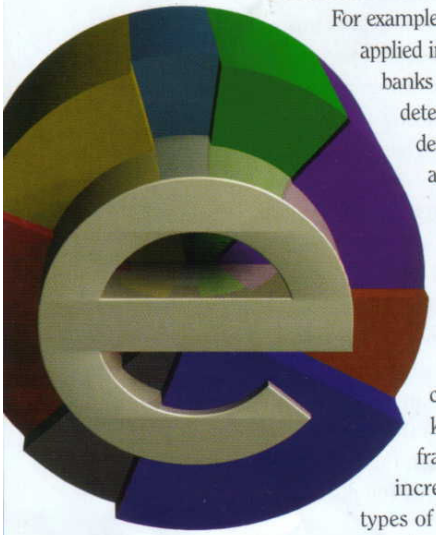
marketing. The typical business need in this area is to understand customers' buying behaviours and associated characteristics.

One way that data mining can be used to achieve this goal is by assisting in customer segmentation. The objective of customer segmentation is to divide customers into characteristic, like-minded groups. Because each segment is subsequently homogenous, every segment can be treated as if it were a single person. Customised marketing strategies can therefore be applied specifically to each individual segment. A typical data mining operation such as 'clustering' can be used in this area.

Other applications for the data mining process are target marketing and cross-selling. The goal of target marketing is to identify potential customers who are most likely to respond to a new product, a catalogue mailing, a marketing campaign or a special marketing programme. The emphasis here is to generate a target mailing list of highly responsive people. Business problems, such as Customer Retention/Attrition/Acquisition, can be solved using a similar process. Various predictive modelling techniques can be applied here, for instance, linear/logistical regression, neural networks, decision trees, or radial basis functions.

The objective of cross-selling is to generate intelligent product portfolios that include new products being marketed to new or existing customers. Using current purchase transactions as input, discovering product associations is an important step for cross-selling applications. Association rules algorithms can be applied in this area.

Data mining techniques are useful not only for marketing purposes, but also for discovering information about potential customers who may be dangerous to deal with. For example, risk analysis is typically applied in financial institutions such as banks or insurance companies to determine the root cause of loan defaults and other risk-related activities. Risk analysis enables companies to take proactive actions to prevent future financial losses. Predictive data mining models are typically applied in this area.



Likewise, fraud detection can help companies detect the kind of abnormal and fraudulent behaviours that are increasingly problematic for all types of business. Removing fraud in an organisation can provide significant return on investment; many companies have reduced costs by more than 30% as the result of employing fraud detection applications. Predictive models and sequential pattern analysis can be combined to solve the business problems.

In addition to the areas outlined above, data mining applications can be found in the fields of segment migration, quality control, mail stream optimisation and promotion effectiveness.

Technology, challenge and risk

The process of migrating from data to business intelligence is an integrated process that involves not only new disciplines and technologies but also tremendous human interaction and expertise. Successfully extracting unknown, comprehensible and actionable information is not a trivial process. Before a company invests in this area, it must understand the challenges and risks that are inherent in a data mining project. The bottom line is that people who really know and understand the business are best able to take advantage of data mining results. Therefore, human interaction is an important factor in the success of the data mining project.

The disciplines involved in the acquisition of business intelligence through data mining include database technology, statistics and pattern recognition. Statistical data analysis (statistical hypothesis testing, linear regression, etc.) has been applied to business problems for years. Database technologies such as database software, data warehouse applications and OLAP are becoming more and more important for businesses. Rooted in statistics, in the 1980s, pattern recognition evolved into a new field that typically integrates standard statistical techniques with non-linear data modelling and optimisation techniques. The success of transforming data into business intelligence relies on the **integration** of database technology, statistics, and pattern recognition.

In order to be successful, data mining projects must address three main functions. Firstly, it is essential to extract previously unknown information. This requires that the data mining results produce new information and insight which previous analytical approaches failed to uncover. Business owners, domain experts and company analysts must be comfortable that data mining results are valid and contain valuable new business information.

Secondly, the project should produce information that is comprehensible. This may seem obvious; however, if the presentation requirements are not considered at the outset, the entire success of the project may be jeopardised. The data mining tool must present the results of a data mining run in an understandable, business-oriented format. The form and quality of data visualisation plays an important role here. Data visualisation can affect the data mining algorithm selected for a particular task. For example, neural networks have been used in many engineering fields to generate accurate non-linear predictive models. However, when they are applied to the business world, modelling techniques such as decision trees and radial basis functions are often more appropriate than neural networks because of their greater comprehensibility.

Lastly, it is essential that the data mining tools extract information which can be acted on. This requirement affects not only the choice of data mining algorithms but also the infrastructure of the data mining process. For example, suppose that a mining tool finds a set of rules (e.g., customers who spend more than £1000 per month, live in either Chelsea or Kensington, and own a BMW) which will target the most profitable customers. The rules generated are easy to understand, and are also easy to implement because a simple SQL statement can quickly extract those prospective customers who satisfy the criteria. Since the findings are actionable and can be measured, the company's decision-maker will usually require some level of confidence in these findings before broad-based action is taken. A verification process using traditional statistics or an OLAP tool is often necessary to support the findings produced by the data mining process.

Business values

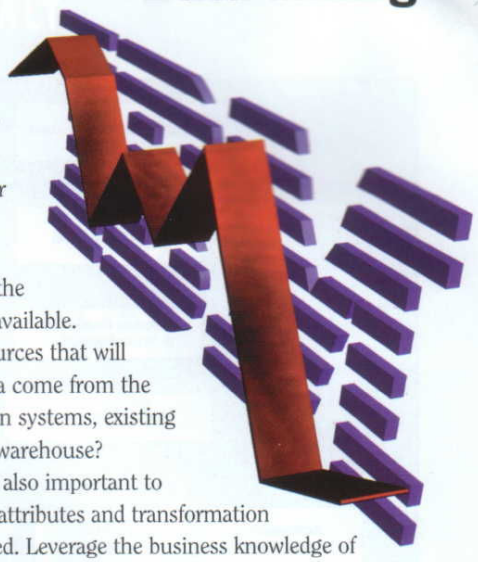
Unfortunately, no data mining tool by itself can address all these requirements. Although many tools include powerful and comprehensible mining algorithms, significant effort and special expertise is required to apply these algorithms correctly in order to discover previously unknown and actionable business results.

Companies investing in data mining can greatly reduce their risk and increase their success rate if they follow business assessment and data cleansing procedures before beginning the data mining process.

Business assessment involves assessing the business problems that can be addressed through the use of data mining technology. It is also an opportunity to educate employees and system users on the complete data mining process.

Business assessment typically involves compiling a list of business problems that can be addressed using the data mining process, then selecting the problem that has the highest probability of being solved within a few weeks as the pilot project. Each item on the list is scored using the following criteria:

1. **Interests of potential key sponsors**
2. **Level of difficulty, measured in terms of available data and techniques**
3. **Potential benefits to the corporation (e.g., savings, increased market share)**
4. **Potential for producing an actionable implementation**



When defining the scope of the work for the pilot project, it is important to consider the question of whether the data necessary to answer the business problem is available. What are the data sources that will be used? Will the data come from the company's transaction systems, existing data marts or a data warehouse?

At this stage, it is also important to identify the business attributes and transformation rules which are needed. Leverage the business knowledge of the audience to derive new business attributes and transformation rules that will be used in the analysis.

Once these decisions have been made, you can then develop an analysis plan and choose the proper tools and algorithms needed to answer the business question. This is also the time to define the action plan for implementation of the results. Work with the business experts to define the actions that can be taken based on the results of the data mining exercise.

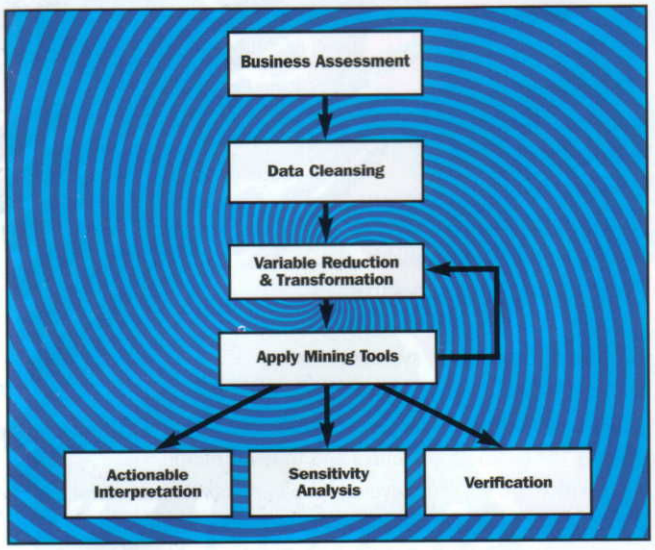
Refined gold

Business assessment is followed by the **data cleansing** stage, in which raw business data must be collected from the various transaction-oriented data systems that are being used. This data may exist in accumulated form in a company's data warehouse or data marts, or may only exist in its raw form in legacy databases. In either case, data integrity needs to be verified before any analysis can begin. The data cleansing and verification process typically involves steps such as examining simple statistics, replacing missing values, and getting rid of outliers that may skew results.

However, data cleansing is only one step towards generating good data mining results. If data is forwarded to a data-mining tool directly after the data cleansing step, the output results are typically demonstration results which may not useful for advancing business. To obtain actionable business results, an iterative process is required which includes important steps such as intelligent variable transformation and reduction, and repeated use of one or more mining tools. The goal of variable transformation and reduction is to leverage traditional statistical analysis to remove linear relationships (collinearity) in the data. Once these linear relationships are removed, you can focus on the remaining data elements that may harbour hidden non-linear relationships. It is these non-linear relationships which will be the focus of the data mining application; they provide the nuggets of information which traditional statistics, human experts, or OLAP tools have difficulty in discovering.

Statistics functions are used at the beginning of the mining process to perform variable reduction and transformation.

Data Mining



Statistics and OLAP tools can also be used at the end of the process to verify the mining results. In the data mining process, database technology serves as the interface between data and analysis. The core data mining systems are used as non-linear pattern discovery and recognition tools in this process, with verification and interpretation of the results playing an equally significant role. You can also use statistical tests such as t-test, F-test and chi-square to further verify the statistical significance of the data mining results. The full data mining life cycle is shown in Figure 1 (*left*).

Real world

The results of a successful data mining run can be used to determine the course of action to be taken by the organisation. Because the data mining process can have such powerful implications for a business, the skills involved in taking decisions like these reach far beyond the selection and use of data mining tools. Equally important are the analytical skills and business knowledge required to identify the information that needs to be extracted, and to interpret and communicate the results.

Ying Debbie Zhao is a business intelligence consultant at IBM. Charles Zhou is one of the principal founders of Intelligent Data Management Group, a consulting firm specialising in business intelligence for the AS/400. Jim Bainbridge is AS/400 Partners In Development Business Intelligence Solutions team leader at IBM.