Audit management staff members are constantly challenged to cut time in completing testing. They evaluate automated controls constantly by reviewing system options, edit logs, etc. They ask themselves: Are information technology (IT) auditors getting the most out of the available technology that can enable financial/operational auditors to effectively perform their duties to detect inefficient and ineffective processes including identifying fraud, waste and abuse of company resources? The answer can be at management’s fingertips if it uses well-known data mining techniques, which are part of continuous auditing.¹

The automated tools available today, when compared with 20 years ago, are well beyond sorting techniques. These tools are capable of analyzing terabytes of information and searching for patterns that may not be identified easily by manual means. In addition, over the past five years, numerous articles and large consulting practices have been created to assist companies in understanding their data, so they can make the most use of data mining. The largest selling point to obtain resources for establishing a data mining process is that organizations can increase their profitability by identifying process improvements, detecting fraud and improving risk management. Furthermore, the patterns uncovered using data mining help organizations make better, timelier and more profitable decisions. An ancillary benefit to data mining is to assist in identifying data that can be sold to other organizations for profit, assuming that the information cannot be traced back to a single individual and/or is not in conflict with a law or government regulation.

The tools used for data mining can range from simplistic and inexpensive tools, such as Microsoft Access and Excel, to standard industry tools that can be costly, such as Audit Control Language (ACL) and Interactive Data Extraction and Analysis (IDEA). This article is not focused on the automated tools, but the conceptual process of understanding the importance of extracting and analyzing data to assist the auditor in reporting potential inefficient and ineffective processes, including potential fraud, waste and abuse of corporate resources. In addition, this article provides a limited view of statistical methods used to identify unusual behavior or anomalies indicating a need for follow-up by the auditor.

Data mining may be different from the approach of computer-assisted audit techniques common 20 years ago. Specifically, the fundamental differences may include the following:

- Auditors to automatically extract data on a scheduled basis and analyze this data without the need for audit queries to be embedded in the program source code.
- Audit tools have become more powerful and easier to use.
- As a result of the previous two points, auditors can not only analyze single data sets, but can also cross-match data sets to perform much richer analysis, which was difficult to do previously.

### Definition of Data Mining

Data mining is a technique that provides specific information that can detect weaknesses in controls. Furthermore, an objective of data mining techniques is to uncover patterns indicating a broken process and/or develop predictive patterns in business information. The first objective is for the auditors to know the purpose of each data element, including how collective data patterns play a role in business decision making. Typically, there may be hundreds or thousands of data elements or variations that require a great deal of auditors’ time in developing an understanding through a partnership with the business owner.

### Potential Financial Benefits of Using Data Mining Techniques

Depending upon the organization, there are numerous methods that can be used to reduce the cost of external and internal audits. There are significant benefits to all parties impacted by audit.

For example, to reduce external audit fees, the IT internal auditor may use data mining to validate interface software that performs data transfers between systems. The successful comparison of data extraction from each system used in data transfer can validate balancing routines that can occur between systems. Using data mining techniques is especially important when validating data transfers between noncore systems, which are created internally, and an enterprise resource planning (ERP) system (e.g., SAP) used to record financial statement journal entries. In addition, data mining can validate the data transfer between the ERP (e.g., Lawson, Oracle) and a financial statement reporting package (e.g., ESSBASE), which is essential to financial statement integrity.

At the request of management, data mining can be used to validate a known control such as a preventive and detective duplicate payment control within the accounts payable system (disbursement process). If properly established, the use of data mining appears to be limitless.

Finally, the use of data mining can reduce the need for auditors to travel to a work site, thus reducing travel expenses for the company. In addition, time is saved by not requesting business management to supply unnecessary supporting
documentation when the process is efficient and effective based upon the values noted from performing a data mining analysis. There is greater precision when using data mining techniques to evaluate the most critical processes, which will result in a greater return on the auditor’s time and expense by the company. Truly, there is no downside, except for the risk of not properly establishing a baseline expectation to measure trends. Ultimately, the real value of data mining is educating the business process owner on the means and methods of identifying fraud, waste and abuse, so it can be embedded within the organization’s management controls. In the end, management must accept responsibility for using these means and methods to control the business environment.

Rules for Data Extraction

First, the IT audit must ensure that the source of the data is extracted as early as possible in the data creation process. Auditors should understand that there is a risk that data are scrubbed or altered, which could impact the level of integrity (and therefore reliability) required for detailed analysis by the auditor. Specifically, the auditors should typically request that the IT group embed a software algorithm to tap the data from point of origin. While use of a data warehouse is an acceptable practice by business users, the integrity of the data becomes crucial when this information is used to ascertain which process elements need further audit review.

Second, the auditor must fully understand all the data elements. The auditor should consult with business analysts to document each data element, including its significance to the critical success factors of the enterprise.

Initiating Data Mining Methodology

The methods employed by the IT audit group to initiate a data mining exercise could result in a full-fledged continuous auditing process requiring scheduled hours. Furthermore, data mining may lead to a separate continuous assurance process to oversee management data analysis, which requires additional audit resources that audit management may not be able to fulfill. Therefore, audit management should properly plan and have a reasonable perspective before embarking on a data mining exercise.

As with all projects, adequate controls should be established, including project management and system development life cycle controls. However, development methodology, such as agile software development, can be employed to shorten the time frame for developing the necessary queries for data sorting and analysis.

Audit management should ensure that all auditors utilize data mining to better understand potential risks within the various financial and operational processes.

In addition, there are other methods presented in this article, including direct analysis in search of questionable occurrences of values within the data and a statistical method that may be used to identify variations of predictive values within the data.

Standard Data Analysis

As noted previously, direct analysis in search of questionable occurrences of values within the data is the most common data analysis method employed in data mining. Specifically, data analysis usually begins by searching the data files for specific occurrences of data indicating potential fraud, waste and abuse. Examples of questionable practices, typically revealed by data mining within a sample of financial transaction processes, include:

- Risk associated with revenue:
  - Sales to customers in the last month before the end of an accounting period with terms more favorable than previous months to make sales targets and receive bonus pay without approval of management
  - Sales with affiliates and related parties
  - Abnormal number of order cancellations by specific salespeople after the end of an accounting period
  - Recording fictitious sales to nonexistent customers and recording phony sales to legitimate customers
  - Billings to customers that do not equate to customer contracts (within the contract management system)
  - Excessive number of credit memo and other credit adjustments to accounts receivable after the end of the account period
  - Unusual entries to the accounts receivable subledger or sales journal
  - Unusual reconciling differences between sales journals and the general ledger
  - Journal entries made directly to the sales or revenue account

- Risk associated with inventory:
  - Unusual, excessive inventory adjustment amounts that appear repetitive from cycle counts
  - Significant changes in gross profit percentages
  - Large increases in inventory balances without corresponding increases in purchases
  - Journal entries made directly to the inventory account and not through the purchases journal
  - Increases in certain types of inventory or in branches or other locations not examined by the auditors
  - Slow inventory turnover compared to the past

- Risk associated with disbursement (accounts payable):
  - Invoices from companies with a P.O. box address and/or no phone number
  - Invoices from companies with the same address and/or phone number as employees
  - Multiple companies with different names with the same address and phone number
  - The amount of each invoice from a vendor falls just below the threshold for review
  - Check presented for payment that the company did not issue (two checks deposited with same check number)

Data Mining Using Statistical Modeling

Aside from the simplistic analytical review noted previously, there may be a need for a more detailed analysis that requires a statistical understanding of the data to ascertain predictive patterns, especially if there are voluminous amounts of data. Overall, the auditor will strive to know the following:

- Patterns in the database and which ones are critical
- Likelihood that an event will occur
- What the summary of the database tells the customers
In conjunction with the previous list, the auditor should identify key indicators. In addition, the auditor should research what values impact (drive or are predictors of) other values (predictive value). The IT audit must first ascertain the following key indicators to get a general understanding of the key values (e.g., categories with the greatest impact on the company’s critical success factors):

- **Max**—The maximum value based upon a driver (predictor)
- **Min**—The minimum value based upon a predictor
- **Mean**—The average value based upon a predictor
- **Mode**—The most common value based upon a predictor
- **Median**—The value based upon a predictor that separates the database into two parts containing an equal number of records
- **Variance**—The measure of how spread out the values are from the average value

Typically, the next step is to create some form of regression analysis that can be used as a predictor. There is at least one predictor that drives the critical predictive value up or down. For example, direct labor time (predictor value) for a construction project drives the variable overhead cost (predictive value) up or down, since it is associated with management of the direct labor time. Another possible example is that reduction in “inventory on hand” over a sizeable time period may increase cash flow. There are numerous predictors that, when combined with a value, provide a predictive pattern that the auditor can use to evaluate the efficiency and effectiveness of a process or determine if there is potential fraud, waste and abuse of corporate resources. This allows the auditor to compare future actual values to determine if the behavior is consistent, identify anomalies and respond sooner.

The relationship between the values (predictive and predictor) can be mapped onto a two-dimensional graph. A common method of mapping using linear regression attempts to explain this relationship with a straight line fit to the data:

\[ Y = a + bX + e. \]

The “residual” \( e \) is a random variable with a mean of zero. The coefficients \( a \) and \( b \) are determined by the condition that the sum of the square residuals is as small as possible. There is an intuitive assumption that the data are linear and, therefore, it is possible to find the slope and intercept that make a straight line and best fit the data.

As can be seen, this is the simplest form of regression, which seeks to build a predictive model that is a line that maps between each predictor value to a prediction value (see figure 1). Of the many possible lines that could be drawn through the data, the one that minimizes the distance between the line and the data points is the one chosen for the predictive model.

Using a regression model can provide the auditor an easier method of identifying unusual occurrences in the critical values. In addition, adding more predictors (or a variation or multiplication of them) to create the linear equation can produce more complicated lines that take more information into account and, hence, make a better prediction. This is called multiple linear regression, which is beyond the scope of this article, but the auditor should evaluate all aspects when using any model. However, it is up to the auditors to ascertain which values can be used as predictors of critical predictive values. This can be achieved only by understanding the business process and all of the data elements, which may not be a small task. The most important takeaway from this is that the auditor can use a regression model to predict values and, therefore, identify values that indicate potential of an inefficient and/or ineffective process.

**Conclusion**

As noted, audit departments can improve their efficiencies greatly by creating or expanding their data mining efforts. In addition, the real value of data mining by auditors is in educating the business owners on using data mining means and methods, including available technologies, to better manage their financial and operational processes. From that point, the hope is that the business owner will take on a continuous monitoring approach. In addition, the auditor must become a strong partner with the business process owners to fully understand the data elements captured within the IT systems that denote an inefficient and ineffective process.

**Endnotes**

1. For the purpose of this article, continuous auditing defines the technologies and processes that allow an ongoing review and analysis of business information on a real-time basis. Continuous monitoring is the process and technology used by management, which could be the result of an audit recommendation, to detect compliance and risk issues associated with an organization’s financial and operational environment. Continuous assurance is the audit process that verifies that management’s continuous monitoring is operating effectively.

2. Benford’s law states that the leading digit

\[ d \in \{1, \ldots, b - 1\} \]

in base \( b \) (\( b \geq 2 \)) occurs with probability proportional to \( \log_b(d + 1) - \log_b(d) = \log_b((d + 1)/d) \). This quantity is exactly the space between \( d \) and \( d + 1 \) in a log scale. In base 10, the leading digits have the following distribution in Benford’s law, where \( d \) is the leading digit and \( p \) the probability:

<table>
<thead>
<tr>
<th>Leading Digit</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.1%</td>
</tr>
<tr>
<td>2</td>
<td>17.6%</td>
</tr>
<tr>
<td>3</td>
<td>12.5%</td>
</tr>
<tr>
<td>4</td>
<td>9.7%</td>
</tr>
<tr>
<td>5</td>
<td>7.9%</td>
</tr>
<tr>
<td>6</td>
<td>6.7%</td>
</tr>
<tr>
<td>7</td>
<td>5.8%</td>
</tr>
<tr>
<td>8</td>
<td>5.1%</td>
</tr>
<tr>
<td>9</td>
<td>4.6%</td>
</tr>
</tbody>
</table>

3. The mode values can be utilized via a segmentation method called clustering, which is a method by which like values
(records) are grouped together. Clustering may provide the business owner a top-level view of what to expect from similar types of data categories (e.g., customers with purchasing habits). The true value of clustering is the ability to more easily identify changes to critical values that previously behaved similarly to the values in the cluster. This method can be used to identify changes in disbursements and revenue based upon a specific predictor value within the cluster.

Depending upon the size of the variance, this in itself may be a strong indicator of problems requiring auditor follow-up.

Statistical prediction is usually synonymous with regression of some form. The line takes a given value for a predictor and maps it to a given value for a prediction. For example, a mutual fund company managing an employee retirement savings account with predicted average yearly retirement savings (in the US) for employees making over US $100,000 might equal US $1,000 plus 0.15 multiplied by the employee’s annual income.

The goal with predictive modeling is to define values that best minimize the error of not equaling \( Y \) over various values of \( X \). The most common method to calculate the error is the square of the difference between the predicted and actual values. Calculated this way, points that are farthest from the line have a great effect on moving the choice of line toward themselves to reduce the error. The values of \( a \) and \( b \) in the regression equation minimize this error, which may be calculated directly from the data.

**John Ott, CISA, CPA**

has more than 15 years of experience in several *Fortune* 100 companies. He has specialized in technical IT audits ranging from mainframe and midrange infrastructures to application audits and systems development life cycle reviews. He was a member of the CISA Test Enhancement Committee and is a member of the ISACA Standards Board. He is currently the director of IT audit at AmerisourceBergen, a *Fortune* 29 company.

**Andrew MacLeod, CISA, CIA, FCPA, MACS, PCP**

is the chief internal auditor at the Brisbane City Council in Australia, the largest municipal government in the Southern Hemisphere. Macleod has more than 30 years of experience in information systems and internal audit in Australia and Hong Kong, working with large multinational and major companies in the airline, banking and government sectors. He is a member of the ISACA Standards Board and the Institute of Internal Auditors Professional Issues Committee.

**Kevin Mar Fan, CISA, CA**

is a data analysis and continuous assurance manager at the Brisbane City Council, Queensland, Australia. The Brisbane City Council is well progressed in the implementation of continuous assurance. He has more than 14 years of experience in financial and IS audit, working with major multinational companies in Australia, Europe and the US.