BEYOND RULES-BASED TESTS: THE “SEVEN WONDERS” COUNTER-FRAUD STATISTICAL MODELING TESTS

Learn how leading anti-fraud analytics are no longer just asking questions of the data based on rules, but are letting the data speak for itself using predictive models, anomaly detection, pattern recognition, and text-mining techniques. This session will cover seven practical counter-fraud techniques that rely on statistical algorithms to uncover hidden patterns, improper payments, and rogue employee activities.

You will learn how to:

- Uncover hidden patterns in your data using practical forensic data analytics techniques.
- Implement predictive-modeling use-cases to detect potentially improper payments.
- Apply seven core statistics-based counter-fraud techniques to your financial accounting data.

VINCENT M. WALDEN, CFE, CPA, CITP
Partner
Fraud Investigation & Dispute Services, Ernst & Young LLP
New York, NY

Vincent Walden is a partner at EY, specializing in forensic technology, business intelligence, and fraud detection analytics. He leads a national team of skilled forensic technology and data mining professionals. With a focus on anti-fraud analytics, forensic data mining, third-party due diligence, and electronic discovery services, he has more than 17 years of experience handling the information management, forensic analysis, and electronic discovery needs for large-scale, complex litigations; investigations; and proactive anti-fraud and compliance programs. He has been featured in many publications including Fraud Magazine, Internal Auditor, Compliance Week, Forbes, The Economist, The FCPA Report, ABC News Online, CNBC, and other leading media outlets.

“Association of Certified Fraud Examiners,” “Certified Fraud Examiner,” “CFE,” “ACFE,” and the ACFE Logo are trademarks owned by the Association of Certified Fraud Examiners, Inc. The contents of this paper may not be transmitted, re-published, modified, reproduced, distributed, copied, or sold without the prior consent of the author.
Introduction
Executives across multiple business functions, industries, and geographies have made significant advancements to solve business problems more effectively and efficiently using predictive modeling and statistical techniques. For those charged with preventing, detecting, and investigating fraud, corruption, and other non-compliant behavior, predictive modeling can be a particularly powerful tool in compliance and anti-fraud management efforts.

As companies explore how to mine information across ever-expanding, more complex, and disparate reporting systems, the challenge of analyzing this data becomes increasingly important. Regulatory inquiries, investigations, and in some cases litigation can burden an organization when faced with analyzing an issue that cuts across these information technology systems. Data analysis is equally important when an organization is working to assess non-compliance with risks posed by legal and regulatory standards that vary by industry and geographic footprints. How does an organization effectively interrogate data when evaluating a problem in order to increase their confidence assessing the root cause? It’s critical that an organization has the right people and technology to understand the systems—and, most importantly, the experience to identify fraud and corruption risks.

Fortunately, technology has evolved to where individuals don’t need to have a PhD or Master’s degree in statistics to benefit from the use of predictive modeling. From simple utilities in MS Excel to advanced features provided in software product offerings from IBM, SAP, or SAS, the graphical user interface, work flow, and modeling techniques have become more simplified so that an anti-fraud professional with a base understanding of data
management, technology, statistics, and fraud can effectively run predictive modeling.

With ever-mounting data and legislative requirements, companies are looking to incorporate analytics to sift through the mountains of data that are collected during the normal course of business. Incorporating analytics allows investigators to query 100 percent of the data through a series of test or filters. These tests and filters aid in sifting out the non-relevant transactions and classifying or ranking the remaining transactions for further human review. The analytics are a combination of rules-based tests, text mining, data mining, clustering, and visual analytics.

The purpose of this course is to provide you with a base understanding of the capabilities that predictive modeling and statistical analysis can provide. We summarize some of the more useful statistical techniques for fraud detection as the “Seven Wonders Tests” as they tend to provide some of the best results in an anti-fraud or anti-corruption context.

The goal is to arm you with “key questions to ask” around practical use case applications. The course will not make you a statistician, nor will it dive into large complex formulas of statistical analysis. Rather, the intent is to keep this practical and give you appropriate business contexts to deploy these advanced anti-fraud analytics.

The Forensic Data Analytics Landscape
Incorporating the concepts set forth in the introduction, there are three key themes that summarize the current business landscape when it comes to making the business case for incorporating forensic data analytics into your anti-fraud program.
The regulators are upping their game.
- Be ready—the regulators are investing in advanced monitoring technology. In fact, Mary Jo White, the current SEC Chair, mentioned in her May 2014 Congressional Testimony that the SEC is “committed to using technology to streamline operations and increase the effectiveness of their operations.”

Big risks require “big data” thinking.
- New approaches to counter fraud and compliance monitoring, beyond simple rules-based tests, are incorporating larger data sets—particularly with combining multiple data sources to identify hidden patterns and relationships that could simply not be identified if one were only using a single data source like accounts payable or travel and entertainment. These approaches combine multiple sources, including third-party data sets such as sanctions or watchlist databases or social media, to paint a broader picture and reduce false positives. This often requires statistical tools to sift through.

Compliance fatigue? Analytics can help.
- Often we hear that the businesses can get frustrated with too much compliance. Adding layer upon layer of manual controls or checks can cripple the business operations. Analytics can help! Statistical or advanced analytics techniques can help reduce the manual checks and tedious controls by looking for outliers and evidence of rogue activities, thus improving the efficiency of the program—saving both time and valuable resources.
Big Risks Require Big Data Thinking: 2014 Survey Highlights
In March of 2014, EY release the first annual Global Forensic Data Analytics Survey, entitled “Big Risks Require Big Data Thinking” (www.ey.com/fdasurvey).

The first of its kind, the survey focused on interviewing anti-fraud professionals, specifically those people responsible for the fraud prevention and detection activities in their organization. The survey covered multiple countries and almost 500 recipients, and focused on the organization’s use of forensic data analytics and related activities. Some of the key findings, which demonstrate the need for more advanced approaches, include:

- 63 percent of companies agree they need to do more to improve their anti-fraud/anti-bribery procedures, including the use of forensic data analytics (FDA).
- 68 percent of companies responded that the board of directors is a top beneficiary of FDA.
- 72 percent of companies say that big data technologies can play a key role in fraud prevention and detection, yet very few are deploying it for FDA.
- Only 11 percent indicated the use of statistical analysis and data-mining packages. Clearly, there is more to do in increasing the use of statistical analysis—especially as technology helps make the deployment easier and the data volumes demand better approaches.

What Is the Difference Between Rules-Based Tests and Statistical Analytics Techniques?

Gartner’s Data Analytics Maturity Framework
The level of analytics sophistication can range from simple descriptive statistics that described what happened on a retrospective basis all the way to prescriptive analytics that advise as to what should be done. Rules-based tests are typically associated with
“descriptive analytics” and “diagnostic analytics,” which ask what happened with the data or why certain things happened. A common element of rules-based tests is that they are inherently retrospective in nature.

A key advantage of statistical-based tests is that they can be both retrospective (or backwards looking) in nature as well as proactive and forward-looking. Statistical-based tests include techniques like predictive modeling, which asks “what will happen,” or even prescriptive analytics, which use data to suggest “what should be done.”

The Counter Fraud “Seven Wonders” Statistical Tests
We will discuss the following advanced analytics tests in detail:
- Predictive modeling
- Text analytics/concept clustering
- K-means clustering
- Entity resolution
- Pattern matching and social network analysis
- Statistical anomaly detection
- Jackknife resampling
Predictive Modeling
When it comes to fraud prevention and detection, one of the best use cases for predictive is in the context of “find me more like this.” By taking known suspect payments, predictive modeling analyzes the independent variables to determine what the drivers are for the dependent variable—suspicious payments.

In the predictive modeling case example, we analyzed 400,000 transactions for suspected bribery payments per DOJ subpoena. The team reviewed 2,000 transactions from ledger data (text comments, amounts, dates, etc.) and tagged 400 suspicious payments. Note that the 1,600 non-suspicious entries were also important to tell the model to “not” look for attributes in this population.

The predictive model (we used a C5 decision tree using IBM’s Counter Fraud software) created a statistical model: Is Suspicious/Is Not Suspicious that was applied to the remaining 398,000 additional transactions. The result in the model is that, using the 400 “tagged potentially improper payments,” the model identified 14,000 new suspicious transactions that totaled approximately $10 million in payments. This number was far less than what the DOJ was alleging, which helped the client in settlement negotiations.

Text Analytics
Text analytics uses natural language processing rules and statistical techniques to reveal conceptual meaning and sentiments in text. Text mining is important as it helps identify “corrupt intent” or a perpetrator’s “rationalization” in the context of the fraud triangle. Whether in email, user documents, social media data, or even in the free text payment descriptions of payables or travel and entertainment, understanding the meaning of key events or
transactions through text analytics can be a powerful tool in fraud detection and prevention.

Concept extraction is a text analytics technique that analyzes the top noun or noun phrases in a dataset to identify suspicious words used in a specified dataset. In this case, the free text descriptions of T&E data were analyzed, which identified suspicious descriptions such as “errors,” “mistake,” “seat upgrade,” and other terms perceived as higher risk or out of policy. Note, this is not keyword searching. This is an unsupervised learning technique that looks for the frequency of terms.

Other examples of text analytics include sentiment analysis, which seeks to associate an emotive tone related to certain words or word strings in a selected data set.

### Keyword “Hits” in E-Mail

<table>
<thead>
<tr>
<th></th>
<th>Joe</th>
<th>Frank</th>
<th>Tom</th>
<th>Jill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Override</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
</tr>
<tr>
<td>Write-Off</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
</tr>
<tr>
<td>Delay w/5 ship*</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
</tr>
<tr>
<td>Side Agreement</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
</tr>
<tr>
<td>Recognize w/5 revenue</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
</tr>
<tr>
<td>Side w/5 letter</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
</tr>
<tr>
<td>Illegal</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
<td>☀</td>
</tr>
</tbody>
</table>

### K-Means Clustering: Categorical Analysis

K-means clustering is a popular statistical method for partitioning data into groups that share a common set of quantitative attributes or categorical variables. This is a method of organizing data in a sensible, objective way for
subsequent analysis. Re-clustering is often performed to look at the same set of data in different ways, based on minor tweaks to the algorithm parameters to identify hidden categories of data or suspicious categories of like data types.

**Entity Analytics**
An entity could be an individual, vehicle, vessel, etc. Entity analytics enables an organization to resolve like entities, even when the entities do not share key values (e.g., ID number).

The result is more accurate analytics, based on correctly resolved entities. Entity analytics are especially important with Know Your Customer analytics, third party due diligence, and any dataset where individuals are trying to resolve multiple entities into a correct single entity, such as in a Vendor Master or Employee Master context.

- Named-entity recognition (NER) seeks to locate and classify elements in text into pre-defined categories, such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
- Example: “Jim bought 300 shares of Acme Corp. in 2006.”
  - Person, organization, and time are now fields that are populated into a structured database table for analysis.

**Pattern Matching and Social Network Analysis**
Pattern matching and social network analysis provides a better way to analyze groupings as compared to simple database matches, fuzzy matches, or comparing rows and columns together. This form of visual analytics identifies...
individual groups and relationships between individuals to help identify hidden patterns or relationships. Further, it helps identify core groups and leaders within them based on the number of connections. For example, using social network analysis, we can analyze a simple matching scenario that links single bank accounts to multiple vendor or individual names and displays them in a highly visual manner showing who is linked to whom. While this can certainly be done in a spreadsheet with records being compared, viewing the data in a social network display makes a significant difference in spotting trends.

Another use case of social network analysis is in email and communications data—or, who is talking to whom? For example, live server logs can be pulled for early case assessment in an e-discovery matter to see who is talking to whom around a specific topic or date range.

**Statistical Anomaly Detection**

Perhaps one of the most powerful statistical tests in fraud detection is statistical anomaly detection. This technique looks at statistical variance across all variables (not just the ones you selected) to identify hidden anomalies. In this case, the client paid for a significant hotel expense for a lavish conference. Further, the company had a “no-first-class” policy, and the anomaly detection found a very expensive ticket (around USD 20,000) that had the description of first-class that was overlooked by the traditional rules-based tests.

**Jackknife Resampling**

Jackknife resampling systematically re-computes the statistic estimate, leaving out one or more observations at a time from the sample set. This technique is commonly used to quality assure the model, removing one transaction at a time to see if it makes a meaningful difference to the
Beyond Rules-Based Tests: The “Seven Wonders” Counter-Fraud Statistical Modeling Tests

Model. However, in fraud detection, it can also be a technique for finding anomalous transactions that stand out from their peer group.

According to Wikipedia: “In statistics, the jackknife is a resampling technique especially useful for variance and bias estimation. The jackknife predates other common resampling methods such as the bootstrap. The jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations. Given a sample of size N, the jackknife estimate is found by aggregating the estimates of each N−1 estimate in the sample.”

The jackknife technique was developed by Maurice Quenouille (1949, 1956). John W. Tukey (1958) expanded on the technique and proposed the name jackknife since, like a Boy Scout's jackknife, the technique is a “rough and ready” tool that can solve a variety of problems even though specific problems may be more efficiently solved with a purpose-designed tool.

Effective Deployment Considerations

Understanding Your Data

Before one jumps into using statistical analytics as part of an anti-fraud program, it is critical to first understand the nature of the data and inspect it for completeness and accuracy. The following are some key questions to help you scope your work.

- Is the data available? This may seem like an obvious question, but be aware that, although data might be available, it may not be in a form that can be easily used. Data can be imported from databases (via ODBC) or from files, social media, news feeds, or other systems. Further, the data could be held in

NOTES
some other form that cannot be directly accessed due to privacy, security, or technology resources.

- Does the data cover the relevant attributes? The object of data mining is to identify relevant attributes, so this may seem like an odd question. It is very useful, however, to look at what data is available and to try to identify the likely relevant factors that are not recorded. In trying to predict ice cream sales, for example, you may have a lot of information about retail outlets or sales history, but you may not have weather and temperature information, which is likely to play a significant role. This questioning also follows for anti-fraud analytics. Missing attributes don’t necessarily mean that data mining will not produce useful results, but they can limit the accuracy of resulting predictions. A quick way of assessing the situation is to perform a comprehensive audit of your data.

- Is there enough data? For statistical analysis, it is not necessarily the size of a dataset that is important. The representativeness of the dataset is far more significant, together with its coverage of possible outcomes and combinations of variables. Typically, the more attributes that are considered, the more records will be needed to give representative coverage. If the data is representative and there are general underlying rules, it may well be that a data sample of a few thousand (or even a few hundred) records will give equally good results as a million—and you will get the results more quickly.

- Is expertise on the data available? In many cases, you will be working on your own data and will therefore be highly familiar with its content and meaning. However, if you are working on data for another department of your organization or for a
client, it is highly desirable that you have access to experts who know the data. They can guide you in the identification of relevant attributes and help interpret the results of data mining, distinguishing the true nuggets of information from “fool’s gold,” or artifacts caused by anomalies in the datasets

Selecting the Best Predictive Modeling Projects

The graphic below is a helpful framework for prioritizing projects based on the (1) business value and (2) availability of the data. As mentioned, the key is to make sure you are asking the right business questions and evaluating their respective ROI or business impact. That said, important “cost” factors to consider in determining the feasibility of the project are the risk of cost overruns or the difficulty of obtaining the necessary data for the analysis.

Therefore, a helpful exercise is to evaluate the ROI on the x-axis and the availability of the data on the y-axis. Those projects that are both high-ROI and involve highly accessible data would be considered your best project (or “lowest hanging fruit”) as depicted with the green dots. Your second tier would be high-ROI, with lower data accessibility (yellow dots).
No One Tool Does It All

The four components of a comprehensive enterprise counter-fraud platform include analytics activities that help detect, respond, discover, and investigate. In order to effectively conduct predictive modeling, you first have to get your observations from an ongoing detection process where transactions can be scored or predicted for effective and timely response. Those anomalies then need to be analyzed further to validate the issue, which can be done in the discover phase. When validated, the anomalies can then move into a case management platform where they can be properly investigated and dealt with accordingly. The following graph depicts this process:
**Counter-Fraud Framework**

Looking at the same process in a linear data flow model, you can see how multiple data sources can be loaded for analysis and each phase of the analytics platform performs its function in either a monitoring or investigative context.
Five Success Factors in Deploying FDA

In conclusion, we will share five common success factors we have seen for successful counter-fraud analytics using statistical and predictive modeling.

- Focus on the low-hanging fruit; the priority of the first project matters.
- Go beyond traditional “rules-based” tests; incorporate big data thinking and some, if not all, of the tests we discussed today.
- Communicate: Share information on early successes across departments/business units to gain broad support.
- Leadership gets it funded, but interpretation of the results by experienced or trained professionals makes the program successful.
- Enterprise-wide deployment takes time; don’t expect overnight adoption. Integrating advanced anti-fraud analytics capabilities should not be viewed as a one-time project. It is a way of doing business that requires
constant diligence and innovation as fraud risks constantly change.

**Compliance Analytics Vision Recap**
As we recap these seven tests, it is important to realize that no one test will “change your game” in fraud detection. The key is to align the right fraud risk questions to the right data sources (both internal and external) and provide federated access to a variety of analytics tools that are easy to use and do not require a PhD in statistics to operate.

**What’s Ahead—Cognitive Computing**

*About Cognitive Computing: Definition and Discussion*
According to Wikipedia, *cognitive computing* makes a new class of problems computable. It addresses complex situations characterized by ambiguity and uncertainty; in other words, it handles human kinds of problems. To respond to the fluid nature of users’ understanding of their problems, the cognitive computing system offers a synthesis not just of information sources, but also of influences, contexts, and insights. To do this, systems often need to weigh conflicting evidence and suggest an answer that is “best” rather than “right.”

**Conclusion and Discussion**

**Contacts**
Vincent Walden, CFE, CPA
Ernst & Young LLP
Partner, Assurance Services
Fraud Investigation & Dispute Services
New York, NY
(212) 773-3643
vincent.walden@ey.com