ALL THE DATA: INTEGRATING DATA ANALYTICS AND DIGITAL FORENSICS INTO FRAUD EXAMINATIONS

While examiners are quite familiar with the use of structured data in fraud examinations, it is important to consider all available data. Much of the remaining data, perhaps as much as 80 percent, is unstructured data. Commonly comprising email communications, social media, text messaging, computer activities, and deleted files, this data can be incredibly valuable. Leveraging the knowledge and skills of data analytics experts and digital forensics experts provides insight and information that can increase both the efficiency and effectiveness of an investigation. This session will discuss how you can leverage all available data, both structured and unstructured, in your next examination.

You will learn how to:

- Differentiate between structured and unstructured data.
- Use data analytics related to structured data.
- Identify the six common components of text mining.
- Describe and define predictive coding.
- Identify the relationship between digital forensics and data analytics, and how they impact an examination.

JEREMY CLOPTON, CFE, CPA, ACDA
Senior Managing Consultant
BKD LLP
Springfield, MO

As a member of BKD’s Forensics & Valuation Services division, Jeremy Clopton specializes in data analytics with applications in forensic investigations, fraud prevention and detection, litigation support, continuous auditing, internal control assessment, and risk assessment. His use of data analytics is built on a solid foundation of fraud examination experience, including investigating allegations of misappropriation through interviewing, document examination, and data analysis. Clopton’s project experience includes the development and implementation of continuous auditing programs for Fortune 500 companies, development of analytics for compliance with anti-bribery and corruption regulations, and investigative experience working with the FBI and U.S. Department of Justice.
Building a Foundation
Before beginning a discussion on how the worlds of data analytics and digital forensics can intersect to enhance fraud examinations, a discussion of the basics is required. First, we will start with a discussion of the differences between structured and unstructured data.

**Structured**
The world of structured data is one with which we are all quite familiar. Spreadsheets, databases, and other data nicely organized in columns and rows—many times with headers—are the most common examples of structured data. Nearly every examiner has worked with structured data in some capacity, and will continue to do so in nearly every new examination.

**Unstructured**
Unstructured data, on the other hand, is essentially all of the other data in the world. Gartner Research, an information technology research and advisory company, indicates that unstructured data now accounts for about 80 percent of all available data in an organization. Factor in the growing number of data connected devices—phones, tablets, thermostats, refrigerators, etc.—thanks to the Internet of things, and 80 percent will soon be a low estimate. Also, technologies in both digital forensics and data mining have reached the stage where they can work together. (See Gartner’s May 16, 2005, report, “Introducing the High-Performance Workplace: Improving Competitive Advantage and Employee Impact” at https://tinyurl.com/oqzc738.)

Data Analytics Basics
With a solid foundational understanding of the two main types of data, a foundational understanding of both data
analytics and digital forensics is required. Without properly defining these terms, many examiners find the concepts quite intimidating.

**Definitions**

**BIG DATA**

Do a quick online search for the definition of *big data*, and you will get millions of results (my Google search resulted in 30 million results in 0.55 seconds). It is no wonder the term big data is one that intimidates many examiners. However, if we use the definition of big data from Gartner, big data is nothing more than “information of extreme size, diversity and complexity” ([www.gartner.com/technology/topics/big-data.jsp](http://www.gartner.com/technology/topics/big-data.jsp)).

Put another way, big data is data that is big. Although big data may differ from examiner to examiner, all examiners will encounter data they feel is big at some point in their career.

**DATA ANALYTICS**

Data analytics is another term with varied definitions (6.6 million search hits in 0.46 seconds). The definition that resonates with most examiners is “the procedures and methods used to extract useful information from data sets and answer strategic questions.” In the case of fraud examinations, the strategic question typically involves determining where fraud is occurring or has occurred. However, it is this broad definition that allows us to change the question and apply data analytics to virtually any aspect of an examination or organization.
### Examples of Common Uses in Examinations

Applying the concepts of data analytics and big data to examinations is quite common. Vendor management, accounts payable, payroll, and many other areas of an organization generate enormous amounts of structured data each day. The days of relying on a sample of 30 documents are passing. The examiner of tomorrow will be using variations of the procedures described below in nearly every examination.

#### ACCOUNTS PAYABLE

Fraud schemes related to cash disbursements are some of the most common, according to the Association of Certified Examiners’ 2014 *Report to the Nations*. For this reason, accounts payable is one of the most common areas of application for data analytics. Some of the procedures in accounts payable include:

- Identification of employee/vendor relationships
- Analysis of cash disbursement sequences
- Analysis of payment issuance patterns
- Identification of fictitious vendor indicators
- Geographic proximity analysis
- Trend analysis of vendor payments

#### PAYROLL

One of the other more common areas of data analytics application is payroll. Between the employee file, payroll detail, paycheck detail, timekeeping systems, and other human resources data, there is a large amount of data available for analysis. Some common procedures include:

- Identification of ghost employee indicators
- Trend analysis of payroll hours and earnings
- Analysis of paycheck issuance
PURCHASING CARDS
Likely the fastest growing area of data analytics application is related to purchasing cards (also known as corporate credit cards, travel cards, procurement cards, p-cards, etc.). While these are generally considered high-risk by investigators, purchasing cards have some of the best data available in an organization. Not only is the data generally in an easy-to-access format, but it is provided by the banks with little-to-no human data entry. The result is high-quality data in a high-risk area, something that works well for examiners.

Some of the common procedures include:
- Analysis of purchasing card activity for indications of personal use
- Identification of purchasing card activity during personal leave
- Analysis of activity in contradiction to corporate policies and procedures

Digital Forensics Basics
The term *digital forensics* is an update to the term *computer forensics*. For most organizations, computers contain only a fraction of the data available for analysis. Phones, tablets, flash drives, personal data assistants, and other devices generate much of the unstructured data available for analysis. A definition of digital forensics will provide a solid foundation for the remainder of this paper.

**Definition of Digital Forensics**
Digital forensics is the collection, preservation, analysis, and reporting of digital evidence in an examination. Prominent forms of digital evidence include computers, email, servers, mobile devices (such
as smartphones and tablets), cloud storage, and external storage devices.

**Historic Uses in Examinations**

The sources of information available through forensic analysis are rich. On many forms of digital media, examiners can recover deleted activity and find communications never intentionally saved by the user in areas such as the computer’s memory. For example, on smartphones, recovered photos may include GPS coordinate information from the time the photo was taken. Other traditional sources include:

- Internet history and related artifacts
- Chat/instant message history
- iPhone backups
- Chronological timeline of events on devices
- Financial records and software

Forensic software allows a fraud examiner to search digital media in a relatively efficient manner—usually beginning with simple keyword searches and analysis of log files generated by the digital device. The software can also present historic activity in a chronological timeline to assist in corroborating digital activity with other events.

**Integration of Data Analytics and Digital Forensics**

Historically, digital forensics and data mining have existed separately in the fraud examination arena. Investigators used digital forensics to analyze unstructured data and used data mining for structured data. However, mining structured data has become an established component of fraud examinations and forensic accounting examinations. Fraud examiners may now extract the results of digital forensics analyses and export them into formats readable by most data mining software. Data mining software also has
the ability now to read much more than just the traditional structured data format. This results in an abundance of opportunities to integrate these technologies for an even more powerful examination. Taken this way, digital forensics tools are in effect an enabling technology for more advanced forms of analysis.

**Synergies by Integration**

Combining data analytics and digital forensics enables fraud examiners to use more exotic forms of analysis, such as:

- Artificial intelligence-assisted searches for relevant content (often called *predictive coding* or *technology-assisted review* in legal and investigative circles)
- Detection of emotional tone of communications
- Collection of names, places, events, and dates to construct “relationship maps” of possible related parties, which fraud examiners couldn’t otherwise detect because they might be too obscure or separated by too many degrees

The data and information gathered, analyzed, and produced using text analytics provides even more value to an investigator when used in combination with other procedures related to data analytics involving structured data. Named entities, email recipients/senders, and relationships provide further insights into how employees, customers, and vendors interact. The relationship map becomes even more robust by adding the structured data relationships (employee-vendor, employee-customer, vendor-customer, etc.) identified based on common attributes such as name, address, phone number, or tax identification number. In some instances, analyzing structured data and unstructured data independently provides two interesting, but
possibly incomplete, results. Combining the results of both into a single analysis provides a more complete picture.

Incorporating dates/times of email, document creation, social media postings, and computer-based activities (downloads, uploads, deletions, etc.) provides chronological events that can be useful in further analyzing transactions to identify possible correlations and/or causations. For example, analyzing communications from a purchasing director involving a request for proposal (RFP) process can be used to find indications of potential bid rigging. Red flags may include email or phone communications to the winning vendor minutes before the submission deadline, or a specific vendor winning each time a particular individual is on the RFP evaluation committee.

**Text Analytics**
The most common method of combining data analytics and digital forensics is text analytics. One of the most frequent applications of this integrated approach is email analytics. Email, as a source of evidence, not only contains word-for-word communications but also possesses a date/time element, metadata, and emotional tone as expressed through various idioms, phrases, and adjectives. Therefore, email analysis shouldn’t be limited to keyword searches alone. It also should include extraction of meaning and topics; emotional tone of conversations; the creation of relationship networks to visualize how key players and topics interact, influence, and evolve over time; and integration into conflict-of-interest and related-party testing based on new relationship information learned.
COMPONENTS OF TEXT ANALYTICS

- Predictive coding incorporates artificial intelligence (AI) to assist analysts in finding related and similar documents in a massive collection of text. AI is capable of determining the underlying concepts in documents or emails, so analysts can perform predictive coding independently of traditional methods that rely on keyword searches. Predictive coding can find highly relevant content rapidly, but, more importantly, it can reduce the volume of material that has to be manually reviewed by as much as 95 percent in an examination. The AI and human analyst leveraging each other’s strengths to achieve augmented intelligence makes this possible.

- In addition, the technology can dissect communications into their grammatical subcomponents, creating topic maps. Many people refer to these as word clouds, which in general include words of varying sizes and colors that bear specific meaning. Topic mapping tools are also adept at bundling terms that relate to the same topic in one category, so true underlying topics can easily be visualized.

- Part of speech tagging is the process of a computer program breaking text into its grammatical parts. Leveraging this function, one of the more useful and exciting types of analysis—tone detection—is possible.

- Tone detection uses adjectives, idioms, and phrases found in communications to assess the emotional tone of the communication. This ability has powerful implications—an investigator can quickly hone in on red flags without even having an initial theory or starting
point in an examination. Common tones to search for include tense, vague, nervous, low-esteem, and conspiratorial, among others.

- **Named Entity Extraction** (NEE) provides a particularly powerful analysis to the investigator. Text-mining tools can identify grammatical components, so they’re very adept at identifying proper names, places, and events. Because names and events can be pulled from email communications, NEE is useful in *relationship mapping*, the process of graphically representing relationships among the various subjects of an examination.

### Specific Components
Some of the components of text analytics warrant a more detailed discussion. Those components—predictive coding and network relationship mapping—are discussed in more detail below.

#### Predictive Coding
The legal profession uses the term predictive coding to describe a technology that can search through vast collections of documents and retrieve matches by both keyword and concept. Outside of the legal world, this might be called *concept searching* or something very similar to Google’s “find more like this.” Here’s how it works in a non-litigation examination setting:

- The examination team identifies an initial set of relevant documents or emails.
- The team passes the relevant set through artificial intelligence-based machine learning algorithms, which employ natural language processing (NLP) functions to search for conceptually similar documents.
• The team reviews the resulting set of potentially relevant documents and again assesses their relevance.
• The iterations continue until the team reaches a reasonable comfort level. Each pass returns more relevant material and also uses the “not relevant” material to validate the results.

The purpose of predictive coding in a fraud examination is to find relevant content faster and with greater accuracy than traditional methods such as keyword searches or manual review. Often an examiner will use it early in an examination to help “widen the net” more efficiently or near the end of an examination as a “final sweep.” The fraud examination use of this technology differs markedly from the more restrictive, formal approach that’s necessary in a litigation environment.

While the hype around this technology focuses on finding the “smoking gun” evidence, the primary benefit is minimizing the documents you review. Conservative estimates put the reduction in amount of text items to review (and possibly even more importantly, cost) at 80 percent. (See “E-Discovery and the Rise of Predictive Coding” by Ben Kershberg, Forbes, March 23, 2011, http://tinyurl.com/m3hbgg3.)

**Network Relationship Analysis**

At the human level, the structure of networks is quite complex because our relationships involve social and communicative elements—i.e., social networks. These networks form because of our choices of whom we interact with—based on ideology, religion, social status, personality types, or our surrounding environment.
Social Network Analysis (SNA) is the science of studying social networks, sometimes referred to as graph database analysis or relationship mapping. The proliferation of email, messaging, electronically stored documents, and social media not only provides a rich repository of data for social scientists but also provides a platform for analysis of relationships in fraud examinations.

In the context of occupational fraud, sources of data from which networks are constructed might include, among others:
- Email
- Structured data
- Phone records
- Depositions
- Documents
- Human resources data
- Open source data
- Social media
- Medical records
- Loan documents
- Interview notes
- Insurance claims data
- Online news

Within these data sources are the individual data elements for identifying relationships among various individuals and entities. If we collectively analyze these individual elements, we’re able to construct network maps that provide the foundation for SNA.

SNA provides a quick means to “see” complex social networks using data visualization and can often lead to immediate and useful insights. However, more powerful than the visual component of SNA is graph
metrics—mathematically derived data about connections and groups.

The key graph metrics include:
- Degree—count of connections for an entity
- Eigenvector centrality—connection quality measurement
- Betweenness centrality — measurement of influence in a network

Degree is the simplest of graph metrics and represents the number of connections for a given entity. In examinations, people with a high degree are of interest because they might be a rich source of information about other members of the network.

Eigenvector centrality (EC) is a measure of overall potential influence, factoring in an entity’s degree and also the degree of an entity’s connections. It’s important in fraud examinations to consider both the degree and EC metrics to help identify key players in a large network.

The key metric in FCPA and corruption examinations is the betweenness centrality (BC). In an examination, this helps identify the individual acting as a bridge between the organization or municipality and a group of outside influencers that are involved in the corrupt activities. Another important feature of SNA is community detection, which statistically clusters and identifies subgroups within the larger network. Community detection is important in examinations to identify cliques or buddy networks—the networks within the networks—that exist, which may be indicative of a collusive or conspiratorial situation.
These metrics are only a subset of possible metrics using SNA; however, they’re some of the most useful in a fraud examination.

While network relationship maps are effective in a static view, the ability to watch relationships evolve over time is more effective. Watching the maps change—both figuratively and literally—relative to key dates and events might help further identify the importance of entities within a network.

Another way to enhance SNA is by combining this approach with natural language processing, named entity extraction, tone detection, and part of speech tagging in email communications. For example, importing email data into a network analysis tool will allow the creation of a simple map based on to/from fields. Taking this a step further, named entity extraction allows the incorporation of actual names, dates, and places into a network map. Include the date and time stamps, and you now have a time series analysis of the network map.

**Impact on an Examination**

The growth of unstructured data is a key driver in the need for the collaboration of data analytics and digital forensics. Initially, the incorporation of unstructured data into examinations is a daunting task. The volume and complexity of the data to be analyzed is a challenge, a challenge best conquered by the collaboration of data analytics and digital forensics. The result of this collaboration is a comprehensive analysis—fully integrating both structured and unstructured data. It’s a process well worth the effort. For example, while corruption is the second-most common categorical type of fraud according to the
ACFE’s 2014 Report to the Nations on Occupational Fraud and Abuse, it’s the most common fraud scheme in most industries studied. (See www.ACFE.com/rttn.aspx, Figure 24.) Corruption schemes are based on the use of influence to manipulate and deceive. Merely analyzing accounting records and common financial data sets isn’t effective enough to identify influence. Leveraging structured and unstructured data together allows us to perform SNA—a method designed to identify and measure influence.

Framework for Application
The challenge in applying the principles in this paper is determining how to apply them to your examinations. For that, a six-phase framework provides the structure necessary to better incorporate data analytics and digital forensics in your next examination.

- **Phase 1:** Assess risk. Before any types of analytics or digital forensics can be performed, it is necessary to first assess the fraud risk and identify the specific risks to be addressed. Examples of specific risks include the risk of kickbacks, corruption, ghost employees, or personal use of purchasing cards. Once the specific risk has been identified, we move to Phase 2.

- **Phase 2:** Identify objectives. Once the risk is identified, define the objectives needed to properly address the risk. These objectives should be specific. In most scenarios, it will require more than one objective to properly address a risk identified in Phase 1.

- **Phase 3:** Obtain data. The third phase is traditionally one of the most difficult phases. At this point, the examiner is ready to obtain the various data required to meet the defined objectives. Care should be taken to ensure two types of data are obtained. First, it is important to obtain all data elements necessary for a complete analysis. Second, it is equally important to
obtain the data elements needed to complete a thorough follow-up on results.

- **Phase 4:** Develop and apply procedures. When developing procedures, it is important to keep a focus on the objectives defined in Phase 2. All procedures should be designed with the goal of meeting those objectives. Just as numerous objectives are required to address a risk, it is likely a number of complex procedures will be required to meet an objective. To most effectively develop procedures, each procedure should be broken down to its most basic components. Each of the components should be developed individually and compiled once successfully developed.

- **Phase 5:** Analyze results. This phase is likely already familiar to examiners, because analyzing results has been a part of examinations since the beginning. As it relates to analytics and digital forensics, the key is determining if the procedures developed in Phase 4 are meeting the objectives previously defined, and if false positives are present in the results. If false positives are present, it may be necessary to modify procedures.

- **Phase 6:** Manage results. Once the procedures are meeting the objectives and have an acceptable number of false positives, the results must be managed. Determining how to use the results of the procedures in the context of the examination may be a challenge. Some results may contain blatantly improper activity, while others may just be an indicator of potential improprieties. Successful application requires proper management of results within the context of the examination.

**Conclusion**

With the prevalence of data—all data, not just tabular data—it is critical that fraud examiners find a successful way to leverage this data within examinations. Individually,
data analytics and digital forensics are both effective means of analyzing data. However, integrating the two technologies provides examiners with an incredibly powerful set of tools to increase both the effectiveness and efficiency of an examination.