

The Larger Picture: Moving Beyond Predictive Coding for Document Productions to Predictive Analytics for Information Governance

By Sandra Serkes, President & CEO of Valora Technologies, Inc.

Predictive Coding (really, data analytics) is a means for harnessing (or suppressing) the potential information locked in large data sets – aka Big Data. Whether the data set is a collection of a litigant’s corporate emails, a call log of customer complaints at a retail establishment, or an entire state’s tax forms, the starting point is the same: a big, ol’ collection of stuff. And once there is a document population, there is information contained within. The debate begins with whether that hidden information is helpful (an asset), or harmful (a liability), or perhaps both. It progresses with whether or not it is worth the cost, time and effort to find out; and concludes with what to do about it once the status is known (or could reasonably become so). This last point is essentially Information Governance, and the path from technology-optimized litigation document review to full-on information management and control is a short one. The techniques used in predictive analytics for document review are essentially the same as those used in much broader application of the same capabilities. This chapter explores the use of data analytics for understanding, diagnosing, organizing, managing, mining, forecasting and reporting on all manner of document data well beyond litigation and eDiscovery purposes.

For everyone’s sanity, let’s start with a little terminology. The whole area of **analytics** is newly popular, but actually quite mature and well understood by its practitioners: statisticians, data miners, computational linguists, and the like. So, what are analytics? In the case of document and/or content assessment, analytics are pattern-matching software algorithms that “parse” (a kind of digital machine-reading) text. The algorithms are typically modified matter to matter to best optimize precision and recall, two inter-related measurements surely discussed earlier in this book. An iterative process ultimately runs over the entire document **population** (also interchangeably called corpus, collection, or source content). Populations often consist of documents, but can be broadened to include any matter of quasi-organized content. For our case, we use the term document very loosely. A **document** can be a once-physical piece of paper, such as a letter or fax cover sheet, that is ultimately scanned to digital image and enters the analytics realm once there is a text rendition of it, resulting from OCR. A document can also be a natively-borne digital document, such as an email or a webpage. Finally, as far as predictive analytics are concerned a document can also include things we don’t typically describe this way, such as a voicemail, tweet or text message, video or audio, transactional or measurement data and much more. As a general rule, if it can be captured in some way as content, then it’s a document.

Litigation matters are often very concerned with who knew what when and thus focus on the types of documents that often convey such information. Email, in particular, is the litigator’s friend (or foe) in this regard, as email not only *contains* content (the what), it also *transmits* it to others (the who) by its express design. Email documents even conveniently carry a timestamp of all their actions; one stamp for submission, one for each transit hop, one for receipt, and so on (the when). Email also has the blessed intrinsic value of being electronically generated, meaning a) its textual contents are clean and easily obtained for further analysis and b) its structural nature comes with built-in metadata, such as

authorship, creation date and transmission time. Thus email is the number one document of choice (or damage) in litigation matters. But as any good document student knows, email comes with its albatross – attachments! Attachments also have the dual benefit/pain of content + transmission, however, they do not follow a simple structured protocol with metadata or easily extractable content. In fact, an email attachment can ultimately be anything, particularly if it is something that has been scanned in to digital format. Take a file with a .pdf extension. Is it a simple PDF rendition of a Word document? Perhaps a digital drawing or photo? Maybe it is one of those fancy PDFs that let you sign and edit portions of content with custom PDF tools? On the surface, it is impossible to tell without further analytics of its contents.

A fundamental difference between analyzing documents for litigation and for IG is the underlying purpose for doing so in the first place. In litigation, the majority of document review (whether automated or otherwise) is to safely rule out or dismiss the majority of documents (often called “culling”), so that only the most important or critical documents remain. In particular, litigation documents are usually being prepared for production to other parties, and the emphasis is on safely eliminating documents wherever possible, using various withholding options, such as invoking different types of content- or source-based privileges, eliminating duplicates and creating restrictions based on time windows, key words and custodial sources. By contrast, the goals in IG are often to specifically *preserve* documents, and to catalog them as fully and well-informedly as possible. There is little need to cull out documents (other than for obsolescence or retention/deletion purposes), and the goal is to make the contents and analytics as useable as possible for future purposes. In fact, a hallmark of IG uses of predictive analytics is to move well *beyond* simple culling, into areas such as classification, organization, trendlining and forecasting, and modeling past or future behaviors.

It is true that litigation is becoming more cooperative, particularly regarding discovery, document review and productions. In that sense, it is becoming a bit more like Information Governance. While IG can certainly benefit from the utilizing litigation’s tools for predictive document review, litigation can, in turn, learn from IG’s expansion of the tools and capabilities, as well as its overarching view of information as both asset and liability – something to be intelligently, actively and purposefully managed all the time, not just during a litigation crisis.

Information: Asset or Liability?

Just as the world contains both optimists and pessimists, so too are the bi-polar prevailing views on information stored within large document sets. Ask someone from the knowledge management, records management or line of business side of the house, and you will hear how stored information is a tremendous asset. A tool for forecasting, and predicting consumer behavior. A rich vein of pattern recognition and statistical correlation for analyzing business trends, and a stellar means for organizing disparate, unstructured information for later reporting and retrieval. If only these vast data stratagems could be unlocked and unleashed from their currently unclassified, disorganized locations! After all, it is surely more cost-effective to have software algorithms do the work all at once rather than either expensive, manual retrieval from Iron Mountain or the digital equivalent of endless, fruitless shared drive searches each and every time!

In fact, this view of data as an asset is so prevalent on the optimists' side of the spectrum that it is a constant refrain from trend-setting organizations, such as ARMA, ILTA and the Big Data community. Consider these headlines and links from well-respected organizations and media sources:

- [Data Management Problems: Organizations should regard data as their greatest asset \(emphasis added\) - and invest in data management accordingly.](#) (ARMA)
- [Treat Data as an Institute Asset: MIT Information Technology Guiding Principles \(MIT\)](#)
- [CIO's Consider Putting a Price Tag on Data \(CIO Magazine 6/23/2014\)](#)
- [The Big Mystery: What's Big Data Really Worth?](#) (Wall Street Journal 10/12/2014)

Clearly, there is a large contingent arguing for the capture, analysis, mining and accounting of data to be treated as an important corporate asset.

However, just as loudly, there is a contingent arguing against the value of such data. In fact, to many corporate legal and compliance groups, information stored in corporate documents is a vast cesspool of exposure and liability. A sea of personally identifying information (PII), rampant personal health information (PHI), internal trade secrets and evidence of improper behavior of all types. These groups are particularly fearful of email collection and analysis –as email represents not just a data store, but an explicit communication trail of all that ugly underbelly information. (See earlier discussion.)

The fact that software algorithms can race through vast document stores and lay bare all that is contained within, including patterns of behavior over time, is terribly dangerous. The onus to *act* on any information found is overwhelming and a harkening back to the days of “don't ask (or look), don't tell” sounds downright safer. With company after company being subjected to data hacks and breaches at all levels, the media constantly reminds all of us just how exposed we all are.

In fact, this view of data as liability is so prevalent on the pessimists' side of the spectrum that it is a constant refrain from trend-setting organizations, such as the ACC, and corporate and outside counsel communities. Consider these headlines and links from well-respected organizations and media sources:

- [A definition of Electronic Data Liability](#) – as provided by the IRMI, International Risk Management Institute
- [How to Create a Moore's Law for Data](#) – (Forbes, 12/12/2013)
- [How to avoid becoming a big data liability](#) – (Information Age, 3/14/2014)
- [Target Says Data Breach Bigger Than Previously Thought](#) – (CBS News, 1/10/2014)

So, who's right? They both are, of course. Yes, data can surely be used as an asset to help steer decisions and budget allocations in the corporation. But, also yes, information is dangerous and can fall into the wrong hands, and so it needs to be understood, managed and controlled properly – the very cornerstones of Information Governance.

How & Why to Understand What Data Your Documents Hold

Whether you subscribe to the data-is-an-asset or the data-is-a-liability camp, or perhaps somewhere in between, it is critical either way for your organization to understand what information it holds in the form of stored documents and files. The most common types of documents in a modern corporation in the mid-2010's are:

- Email and attachments
- Spreadsheets of tables, lists, financial information, and forecasts
- Contracts, agreements, HR documentation, and other legal documents
- Simple, unstructured documents (think: MS Word)
- Presentations, marketing collateral, and other more formalized information
- Databases and SaaS applications
- Websites and social media

Nearly every one of these document types contains identifiable content – thus information. Each unique file can be mined using relatively straightforward analytics, and analyzed using well-established correlation mechanisms. If you are reading this book, then you are or are close to litigation practitioners. You already understand the value of reading through discovery materials to understand each single document's value to your case. The Information Governance application is similar, except it is applied to a much broader spectrum of materials and purposes. Consider the list above, and now multiply that by organizational divisions, locations, and staff count and you will get a sense of the magnitude of information involved, and why IG is often intertwined with discussions about Big Data.

A case in point comes from one of Valora's customers, a large, international law firm. They have well over 3,000 practicing attorneys, with a similar number of non-attorney staff. One email server storing approximately 10 years' worth of email communications holds 32 TB of data, which in turn represents approximately 120 million *non-duplicative* email messages and files. This scope is simply too large to evaluate, classify and handle manually. The costs to do so would be astronomical. (See below.) Even the largest, most bet-the-farm litigation matters are on par with this figure and they, too, would not sustain the manual cost to evaluate each of the collected documents. Using the techniques of Predictive Coding or Predictive Analytics applied to the next level (the entire department or organization) is the answer.

Historically, the idea that any centralized group inside the organization should and would know what all the organization's document contained would have been ludicrous. In a paper-oriented document content world, it would have been impossible to keep up. But, with the advent of electronic document storage systems (commonly called DMS's) and multi-variate software algorithms, it is very possible to achieve that kind of corporate information omniscience. And therein lies the rub. Should an organization, particularly the group tasked with corporate, ethical and risk oversight (i.e., Corporate Legal) seek out such knowledge given that it is relatively easy and cost-effective to do so? Might they, perhaps, be *obligated* to do so as stewards of the company's behavior, actions and ethical obligations?

Ease of Use

Let's break down those two assumptions: easy and cost-effective. Today's predictive analytics are relatively easy for non-statisticians to use. There are many consulting experts in the field who can do the heavy statistical or programming work for you, and there are simplified software interfaces to step a knowledgeable person through the tasks themselves. (Predictive Coding applications for litigation document review are a good example of the latter.) Today, predictive analytics are about as easy to use as any other higher-order analysis task, such as calculating financial markets or economic behavior, or correlating buying behavior with price sensitivity. In other words, if you can handle simple calculus and probability as concepts, you can handle predictive analytics, which puts you on par with a high school honors student.

The complexity arises in the "edge" situations of document or content decisioning. For example, having to think about and then make distinctions between document content that might be important vs. urgent, or not-quite-a-problem, but not quite scot-free either. The ease of routine, mathematical analysis gives rise to a secondary issue: nuance. Back in the don't look – don't tell days, we did not worry much about content nuance, because we didn't worry much about content at all! Now, that it is easy enough to evaluate document content on many levels, we burst open the wide grey area between black & white on many levels.

This gives rise to a discussion on Context. Context is the setting around individual elements of content. Context involves questions such as: **who** said or wrote the information? **For whom** was it intended? Did the intended party **receive** or read it? **Why** was this information communicated? What was the **intent** or purpose of this information? Was it **appropriate** to do so? And so on. Context is important in predictive analytics because it moves things from purely mathematical correlation to indications of behavior, intent, and proof of actions. Context analysis is the way in which grey area nuance is handled in predictive analytics. Consider the cartoon below for a great example of the importance of context.



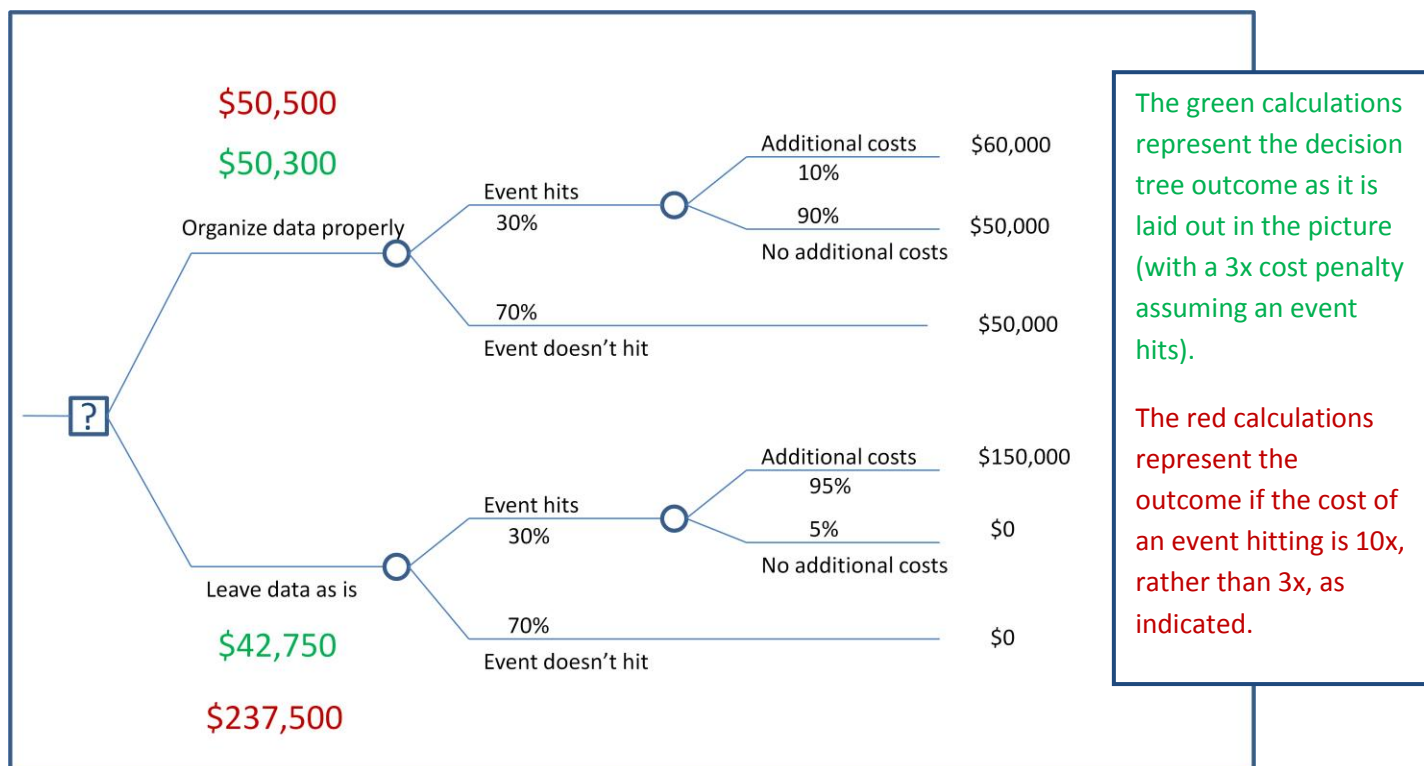
Context matters. It makes a big difference who says these words and why.

Not all predictive techniques make use of context, which is why the results can sometimes seem childishly wrong, even startling. Context adds the color commentary around algorithmic correlation and should be incorporated by predictive analytics rules engines as much as possible. To do this, it is necessary for the correlations to incorporate *indexing*, a method of tagging document metadata and utilizing those tags in the analysis. Context analysis provides a much more sophisticated and robust outcome than content analysis alone.

Cost-effectiveness

Now, let's look at cost-effectiveness. Which costs more? Storing paper documents away in an offsite storage facility or going through the effort to scan them to digital format, OCR and analyze them properly. There has been study upon study validating the image & analyze approach, though typically prepared and funded by imaging providers¹. It's easy enough to do your own ROI analysis using a [decision tree approach](#). See the example below.

To Scan, Image & Analyze Documents (“Store Properly”) or to Leave “As Is”?



An ROI decision tree works just as well for ESI documents housed in servers, hard drives, backup tapes, etc., rather than in boxed storage as paper files. Such similarly “blind” or unaware storage is less cost-effective than the effort to create analytics- and content-aware storage. A new wrinkle, however, is that there are different corporate responsibility requirements for ESI than there are for paper. Paper enjoys a naïve and convenient responsibility dodge due to its “out of sight, out of mind” nature. No one expects that offsite, aging paper to really be analyzed on a regular, ongoing basis. But ESI is readily


¹ [Here is a great example](#) of a well-executed vendor study on scanning vs. storage.

available, being generated anew every day, and frighteningly accessible to anyone with lawful (or unlawful) access. There is a societal bias and expectation that ESI be properly and appropriately managed, in a way that paper never was. Financial markets and investors, consumers, industry watchdog groups, government agencies and the media are clamoring daily about the inappropriate and untamed Big Data beast, with each subsequent data breach fueling their fire. In 2015 alone², there have been over 80 million data breaches across every sector of business, government, health care, the military and more. [Experts](#) estimate that the Target data breach alone (more than 70 million records breached in early 2014) caused more than \$148 million in damage to Target, mostly from lost sales, due to consumers' lack of confidence in the company and its management of their data. So, to accurately assess the cost efficiency of predictive analytics efforts, it is essential to also account for the opportunity cost in not doing so. In other words, what would it ultimately cost to *not* assess and control the vast information stored in corporate documents?


4 Real-World Examples Making the Transition to Predictive Analytics for Information Governance

Fortunately, there is an easy on-ramp to progress from using predictive analytics to analyze discovery documents to utilizing the techniques for full-on Information Governance: Email. Email has many virtues as both a communications mechanism, as well as a default document storage system. Most of us today use foldering, address books, search tools and more to manage our incoming and outgoing email as both a communications record and information storage area. Email has additional virtues for document processing as well. Emails themselves are electronic in nature and origin, which makes their text very strong and reliable for analytics. Further, emails contain useful metadata regarding senders, recipients, subject matter, timestamps, and more. They neatly divide themselves into structured (metadata) and unstructured content (body text and attachments). Most emails are typically easily extractable for processing due to their quasi-storage nature in our inboxes and folders³. Finally, and perhaps most importantly, email is well-understood by the legal community - IG folks and attorneys. All of us use email all the time, and we are also drowning in it, so why not start there?

Enterprise Email Management is a very good Case In Point



- ✓ Universal Issue
- ✓ Involves several key IG problems:
 - Storage/hosting
 - Content analysis & classification
 - Context – correspondence, notification & record, date/time/file signatures, transmission & attachments, custodianship, etc.
 - Administration, management & maintenance
- ✓ Elements of Backfile and Day Forward records management
- ✓ ESI is generally easier & lower cost to tackle than paper files
- ✓ Because of Context, EEM is a hot button issue with real budgets available
 - Investor & media attention
 - Customer concerns
 - Risk & compliance danger zone
- ✓ Predecessor to managing social media



Typical Steps in a Corporate Email Classification Project

² Data breaches numbers are reflective of incidents year-to-date, as of the writing of this chapter in October, 2015, For an up to date incident account and details, visit the [Identity Theft Resource Center \(ITRC\) website](#), and read their weekly published report.

³ For those who don't know, a PST is a compressed archive of stored emails and attachments from a particular folder or custodian.

To begin an enterprise classification project, you need several things. Of course, you'll need the extracted emails themselves (don't forget the attachments!). But, more importantly, you will need a set of specifications of a) what you are looking for and b) how you wish to classify the information. It is sometimes helpful to clients to realize that these specifications don't need to be set in stone at the outset of the process. As has surely been described earlier in this book, predictive analytics are an iterative technology, and there will be many rounds for modification, improvement and refinement of your requirements over time. Almost all predictive analytics techniques require some sort of generalized starting point. Whether a seed set of exemplar documents (as used in predictive coding), or a set of requirements rules (as used in rules-based, pattern-matching algorithms), you will need to do some up front work to codify what will constitute success. An experienced provider of these services will be able to help you with checklists, foldering taxonomies, typical pitfalls, and the like.

Once your specifications, guidance and input files are gathered, the iterative rules-creation and testing (or seed set coding) process begins. One way or another, several rounds of coding, testing and optimization will take place and a point of diminishing returns will be reached. At that mark, it is time to move into the high-volume scenario of the stored backlog of email files (or whatever is being analyzed). If properly tuned, this process should move quickly, sometimes on the order of hundreds of thousands of files per day.

It is important to note that software algorithms for classifying data based on content, metadata and other contextual clues are not perfect. There is typically a 3-10% error rate, which should be accounted for in ROI analysis and other project expectations. There is almost always some mechanism for "Exception Handling," in which problematic, error-prone or non-analyzable documents are handled. The typical exception handling is done by people with explicit knowledge about the materials as well as the project specifications. Exception handling can be performed by an outside vendor, contract labor, outside counsel, inside staff or any combination therein.

In many enterprise email classification projects, there is both a backlog of stored email as well as a "Day Forward" need for continuous assessment and classification on new emails incoming and being created by the organization. For such scenarios, there is an additional step in rolling forward from historical backlog processing (essentially large-scale batch processing) to Day Forward processing happening in real-time. This likely requires integration with the email server, and possibly other document management systems as well.

Finally, it is important to understand that email classification projects (or frankly, any predictive analytics work) is not a "one-and-done" situation. Particularly with Day Forward scenarios, such systems need to be monitored and adjusted over time. Similar to a piano, which goes out of tune over time and requires slight re-wiring to come back into tune, predictive analytics also go out of tune over time. It is not that

Key Steps in a Predictive Analytics Project

1. Gather specification requirements & input data
2. Iterate writing & refining rules on a target set of data
3. Run high-volume backlog through systems
4. Manage exceptions through manual exception handling workflows
5. Roll forward into Day Forward, real-time processing of new incoming & outgoing email
6. Periodic monitoring of performance & needs

the algorithms “go bad,” but rather that the documents themselves change over time. We change our turns of phrase, how we lay things out on a page, who the key contacts are and so on. When such evolutionary changes begin to pile up, the rules (or prior coding) will seem slightly off, and system confidence measures will go down. In a similar manner, priorities and circumstances change within the organization and the rules may need to be re-calibrated for new realities. Fortunately, the solution is simple. A periodic analysis & maintenance of the analytics will suffice, unless there are drastic changes. A safe assumption is a maintenance window of 3-6 months for an ongoing system.

Further Examples of Predictive Analytics Being Used Beyond Litigation Document Review and Enterprise Email Classification

1. Case Management – savvy litigants and their counsel are harnessing the power of predictive analytics to manage the other side of the litigation document world – that is, the work product documents they themselves create to conduct the litigation. Such documents consist of the pleadings, motions, transcripts, depositions, exhibits, other court filings and communications that are *about* the litigation matter(s). Such populations contain a wealth of information about litigants, witnesses, experts, judges, dockets, the exhibits themselves and so on. For example, which exhibits seem to recur from matter to matter, or litigant to litigant? Which answers to interrogatories have we already created for past matters and could re-use for new ones? Which responses yielded the least pushback, or the best results? Which judges tended to rule in our favor and under what circumstances? What did the expert say last time? The technological data-crunching for such answers is fairly simple. Lots of parsing, cross-reference and probability correlation, in fact. What’s difficult, typically, is gathering the documents from multiple matters and sources in the first place, and asking the right questions in the second. To do this successfully requires coordination and compliance of outside counsel to submit their documents to a centralized clearinghouse, and then rigorous application of data mining that can be easily re-configured as new needs and questions arise. Frequent litigants and their counsel realize the power of data mining their own work product and that of opposing parties, particularly when they are involved in serial litigation activities due to their business activities, and the exposure around products and services they sell.
2. PII/PHI Detection & Redaction – the same techniques that are used to determine a document’s author, recipient or parties mentioned within its contents are utilized in different applications for documents containing sensitive information. The identification and subsequent redaction of Private Identification Information (PII) or Private Health Information (PHI) is a major source of manual effort and expense on the part of hospitals, insurers, pharmaceutical companies and any other organization that routinely collects patient, customer, account, or employee information. Still other organizations have trade secret information or other intellectual property in their documents’ contents. They often have need for classification of documents based on file content, secrecy or security level, or retention stage. Typically in such predictive analytics scenarios, the detection of sensitive information is not enough; it must also be properly redacted to prevent accidental or further disclosure. In such cases, predictive analytics have

proven exceptionally valuable at performing redactions on image, in text and into metadata – a task historically performed painstakingly and at great expense by hand.

3. Media Monitoring & Topical Content – an interesting and novel use of predictive analytics occurs around the real-time digestion and categorization of public media information – press articles, releases, news coverage, industry and government reports, announcements and notices, and more. Such push-style information sources yield a plethora of documents that are often rich with thematic content that is of particular interest to specialized groups. Consider organizational groups whose responsibility it is to track customer sentiment, supply chain activity, misuse or illicit activity around their product, or geographic trend data. There is so much information coming at us from sources all over the world 24x7. Tracking this information with a particular agenda can be an impossible task, which is where highly tuned custom analytics come in. It is a simple matter for such rules-based systems to monitor and “parse” the incoming firehose of digital content, all the while sniffing for the relevant terms, sources, date ranges and other contextual clues that identify an item of interest. These technologies routinely handle information originating from several thousand sources each day. Once obtained, the same engines easily perform trend and forecasting analysis, serving as a basis for decision-making and resource allocation.

The Future of Information Governance in a Predictive Analytics World

As more and more people become comfortable with the notion of Big Data, and the organizations they are part of become comfortable with predictive analytics, there will be ever more inroads into how and why we manage information. We will routinely be asked to address questions such as:

What kind of information is (or should be) contained in the typical course of business activity? What is possible to learn or understand? What assets might be present? What dangers, exposure or liabilities might be lurking?

And when the analytical mechanics, the financial risks and rewards, and the assets vs. liabilities arguments are all resolved, the real questions will surface:

What does it mean when every unit of written, spoken or visual content is known and/or can be easily gotten or retrieved? As Information Governance and risk management professionals, what exactly are we responsible for? What would, could or should we be expected know and how will we use that information and prescience? What does proper IG assume about our abilities to understand, manage, manipulate, organize & control our information?