Turning Back Time: The Application of Predictive Technology to Big Data

Deborah Baron
Nuix North America Inc.
660 York Street, Suite 102
San Francisco, CA 94110
+1 877 470 6849
deborah.baron@nuix.com

Angela Bunting
Nuix Pty Ltd
Level 23, 1 Market St
Sydney NSW 2000, Australia
+61 2 9280 0699
angela.bunting@nuix.com

Brian J. Krupczak
Nuix North America Inc.
660 York Street, Suite 102
San Francisco, CA 94110
+1 877 470 6849
brian.krupczak@nuix.com

ABSTRACT
This paper examines new and conventional applications of predictive technologies in the electronic discovery process. The conventional use case applies 'predictive coding' during document review – near the end of the process.. This paper proposes a model, technology-assisted linguistic analytics, that is applied earlier in the discovery process to address the rapidly growing size of data collections. The methodology combines predictive coding with expert knowledge and complementary analytical techniques to address big data and reduce the volume of irrelevant data ahead of the document review phase.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing – text analysis. (In process)

General Terms
Algorithms, Management, Economics, Human Factors, Standardization, Languages, Theory, Legal Aspects. (in process)

Keywords
Xxxxxxxxxxx (in process)

1. INTRODUCTION
The legal industry in recent years has looked to an application of machine-learning technology known as "predictive coding" to lower the cost of discovery in litigation [1]. This technology aims to reduce costs in a number of legal contexts including disputes, investigations and regulatory inquiries by minimizing the number of documents that must be reviewed by human beings [1].

The technologies that underpin predictive coding may include statistical analysis, machine learning, auto-classification and pattern identification.. These technologies have been in use for over a decade in information management software products. For example, in an IT security context, combinations of these technologies form the basis for spam filters in use by nearly every organization on their PCs and networks.

The emerging use of predictive coding in litigation review is positive in that it has increased the interest of the legal community in machine learning technologies and their application. However, many practitioners have developed misconceptions about predictive coding and how effective it can be in reducing litigation costs. For example that it is a very expensive technology useful only in the largest cases, and that it must be specifically utilized in the latter stage of discovery during document review.

This paper proposes a new usage model, technology-assisted linguistic analytics, that begins early in the discovery lifecycle. This model combines predictive coding with attorneys' expert knowledge and complementary technologies—including deduplication and near-deduplication, text and metadata extraction, content indexing and entity analysis—applied at various stages of litigation discovery. This includes before a case begins, at the outset of a case and throughout the discovery process. By applying this methodology, legal practitioners can greatly reduce the volume of irrelevant data and focus on the most relevant and critical evidence.

The paper will draw upon real world experiences across the disciplines of information governance, electronic discovery and internal investigation.

It will also examine a range of further applications for auto-classification technologies, including information management and privacy.

2. DATA VOLUME DRIVES THE NEED TO REDUCE LITIGATION COSTS
"Certainty? In this world nothing is certain but death and taxes."

— Benjamin Franklin

In the modern world a new certainty has taken its place alongside death and taxes – *ever growing volumes of data*.

By any metric, in the age of petabytes and exabytes the burden of electronic discovery is increasing dramatically. In *Pension Committee v. Banc of America*, Judge Shira Scheindlin said all lawyers must now work "[i]n an era where vast amounts of electronic information is available for review," and that "discovery in certain cases has become increasingly complex and expensive." [2] Surveys of corporate counsel and business executives consistently show they expect to face greater litigation and regulatory scrutiny [3] [4]. Discovery production timeframes remain fixed, but the volume of data involved doubles every 18-24 months. [5]

Digital technologies have made it easy to store and retrieve millions of pages of text in seconds. But when litigation requires a human being to review each of those pages, for example to code documents for responsiveness or to identify and redact privileged information, the costs quickly become enormous.

A study by Pace and Zakaras estimated review costs were an average $18,000 per gigabyte of data and up to $30,000 per gigabytes in some circumstances [6] To put this in perspective, a single high-end smartphone or tablet device today has up to 64 gigabytes of internal storage, while many personal computers have a terabyte or more of disk storage.

Corporate data sets are orders of magnitude larger again. For example, in the Lehman Brothers Holdings Chapter 11 case in Bankruptcy Court, the Examiner had to contend with a data collection of three petabytes, or 350 billion pages, in size [7]. By restricting requests to 281 custodians and using dozens of complex searches, the Examiner narrowed this to five million documents, which were manually reviewed by a team of more than 70 attorneys.

3. IS PREDICTIVE CODING A QUICK WIN?

A 2009 paper by The Sedona Conference aptly summarized the dilemma the legal industry faces from this avalanche of data. "The legal profession is at a crossroads: the choice is between continuing to conduct discovery as it has 'always been practiced' in a paper world — before the advent of computers, the Internet, and the exponential growth of electronically stored information (ESI) — or, alternatively, embracing new ways of thinking in today's digital world." [8]

This paper stressed the need to address the scale of data involved in legal discovery while retaining the comprehensiveness and quality of the pre-digital era.

The legal industry has in recent years focused considerable attention on a machine learning technology, predictive coding, as a potential way to balance these needs [1]. Predictive coding is most commonly used in the later stages of litigation as a method of minimizing the number of pages human beings must review before a legal team produces evidence to court [1].

This was a logical place to start. Pace and Zakaras found that manual document reviews accounted for almost three-quarters of eDiscovery production costs [6, p. xiv]. By applying machine learning to automate the most expensive part of the process, the legal industry could achieve a "quick win."

3.1 What is Predictive Coding?

Predictive coding uses statistical analysis and machine learning techniques to automatically classify documents, for example as relevant, responsive or privileged.

To use a predictive coding engine, an attorney with in-depth knowledge of the case and the legal principles involved would manually select two sets of "training documents"—for example, a set of responsive documents and a set of random documents that have no relevance to the case. The engine would analyze these training documents and build a model to differentiate between the two.

The engine would then apply its model to the original training documents and attempt to categorize them according to model it developed. Comparing the difference between the way the original human reviewer categorized the documents and the way the predictive coding model did so allows the engine to gauge its accuracy. Depending on the results, the trainer may need to refine the seed set of documents multiple times until the model is sufficiently accurate.

Once the model is accurate enough, it can be applied to an entire data set, classifying each document as A or B, for example responsive or not responsive. The engine provides a "confidence score" for each document it classifies, indicating how closely the document fits the model. The human reviewer can then code the documents based on the model's recommendations.

Typically, attorneys would conduct quality checks on the results of a predictive coding model. For example, they might statistically sample the results and manually review them to ensure the engine categorized them correctly. A more conservative approach would be to rely on the model's predictions for those documents with a high confidence score but to manually review those documents about which the model was less confident.

3.2 Naïve Bayes or Automated Language Analysis

More than 20 software vendors offer one form or another of predictive coding technology [9]. Predictive coding technologies essentially boil down to two approaches:

- Naïve Bayes classifier
- Automated language analysis (also called "language modeling" or "latent semantic analysis")

Both techniques are combined with workflows to improve the accuracy of the model over several iterations.

A naïve Bayes classifier uses the frequency of words within a document to make predictions about its content. It is a probability model based on Bayes's theorem where a dependent variable C is conditional on a number of variables F_1 through F_n.

$$p(C|F_1 \dots, F_n) = \frac{p(C)p(F_1 \dots, F_n|C)}{p(F_1 \dots, F_n)}$$

It seeks to calculate the probability that, for example, a document is responsive based its containing a number of words or phrases in common with manually selected responsive documents.

The classifier is "naïve" because it is based on the assumption that each word is independent of all the other words. Although this assumption is unrealistic, a naïve Bayes classifier is highly successful in practice and on par with more sophisticated techniques [10].

Automated language analysis techniques take the opposite approach, using complex algorithms to find the connections and meanings between words.

3.3 More Accurate Than Humans?

Both Bayesian and linguistic analysis approaches have been demonstrated to categorize documents at least as accurately as human reviewers. This is because the technology is advanced and because human reviewers are not as infallible as many in the industry believe.

In a previous paper at this conference, Thomas Barnett and Svetlana Godjevac detailed the results of an experiment where they have the same set of 28,000 documents to seven sets of reviewers and asked them to code each one for responsiveness. Examining the tags afterward revealed an inter-reviewer agreement rate of 43% for either responsive or non-responsive determinations [11]. This was "much lower than might be suspected based on the general level of confidence on the part of the legal profession in the accuracy and consistency of document review by humans," the authors said [11].

The Electronic Discovery Institute's Herbert Roitblat, Anne Kershaw and Patrick Oot concluded in a 2009 study that "machine categorization is no less accurate at identifying

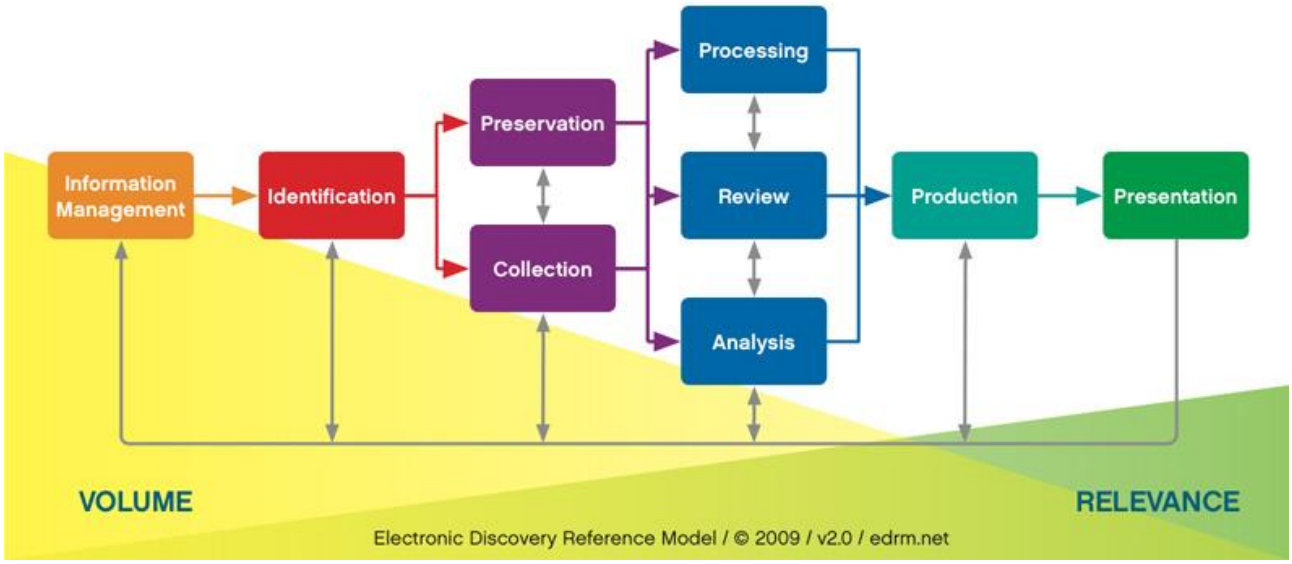

Figure 1: The Electronic Discovery Reference Model. [23]

relevant/responsive documents than employing a team of reviewers." [12]

Maura Grossman and Gordon Cormack compared the accuracy of several predictive coding approaches to a team of human reviewers. They found that "technology-assisted processes, while indeed more efficient, can also yield results superior to those of exhaustive manual review, as measured by recall and precision." [13]

3.4 Judicial Acceptance

The legal fraternity is not alone in grappling with the choice between human and automated decision making. Statistician and author Nate Silver argues society as a whole must change the way it thinks about ideas and how we test them. "We must become more comfortable with probability and uncertainty," he writes. We must accept that our judgments are imperfect and use probabilistic models such as Bayes's theorem to "learn about [the universe] through approximation, getting closer and closer to the truth as we gather more evidence." [14]

In an October 2011 article in *Law Technology News*, Judge Andrew Peck of the U.S. District Court for the Southern District of New York attempted to correct the misconception that "the judiciary has signed off on keywords, but has not on computer-assisted coding." [15] Rather, Peck argued that many judges were highly critical of the keyword approach and that he saw no reason why computer-assisted coding could not be used "in those cases where it will help 'secure the just, speedy, and inexpensive' determination of cases in our e-discovery world."

Judge Peck made good this promise in June 2012 in *Da Silva Moore v. Publicis Groupe* [16]. However, the litigants could not agree on how to decide which documents were responsive. In several other recent cases, including *Global Aerospace Inc. v. Landow Aviation, L.P.* [17] and *Kleen Products LLC v. Packaging Corporation of America* [18], judges approved or even ordered the use of predictive coding as a way to reduce review costs.

4. BROADER LEGAL USE OF MACHINE LEARNING TECHNOLOGIES

The machine-learning technologies behind predictive coding have been used for decades in information management. For example, naïve Bayes classification is an open and well-known technology that organizations have used for more than 10 years in applications such as spam filtering and data mining.

Similar to predictive coding, a spam filter builds up a word frequency model based on email users' manual selection of "spam" and "legitimate" emails. This is why some spam messages sometimes contain large sections of unrelated text, such as a quote from a book, in an attempt to skew the word frequencies so that the contents look—to the computer—to be legitimate.

Predictive coding has other uses within organizational data sets. For example, records managers use "auto-classification" to identify company records that have been stored in emails and on network file shares and not correctly categorized.

Organizations might also use predictive coding in areas where keyword searches do not locate all the relevant information such as:

- Generating a complete record of an organization's knowledge about an issue relating to a regulatory dispute or a merger
- Mining unstructured data (documents, emails and other communications) for previously undiscovered intellectual property and other forms of business value.

5. TECHNOLOGY-ASSISTED LINGUISTIC ANALYTICS FOR EDISCOVERY ACCELERATION

There is strong evidence to support the proposition that predictive coding technology can significantly reduce review costs. Researchers such as Pace and Zakaras have touted the benefits of predictive coding ahead of technologies such as clustering, near-duplicate detection and email threading, concluding "it is unlikely that these techniques would foster sufficiently dramatic improvements in review speed for most large-scale reviews." [6, p. xvii]

However, Pace and Zarakas analyzed these technologies as alternative methods to bulk coding. In the authors' experience of large-scale litigation and discovery matters, this gravely underestimates the value of such technologies when viewed in context of the entire discovery process.

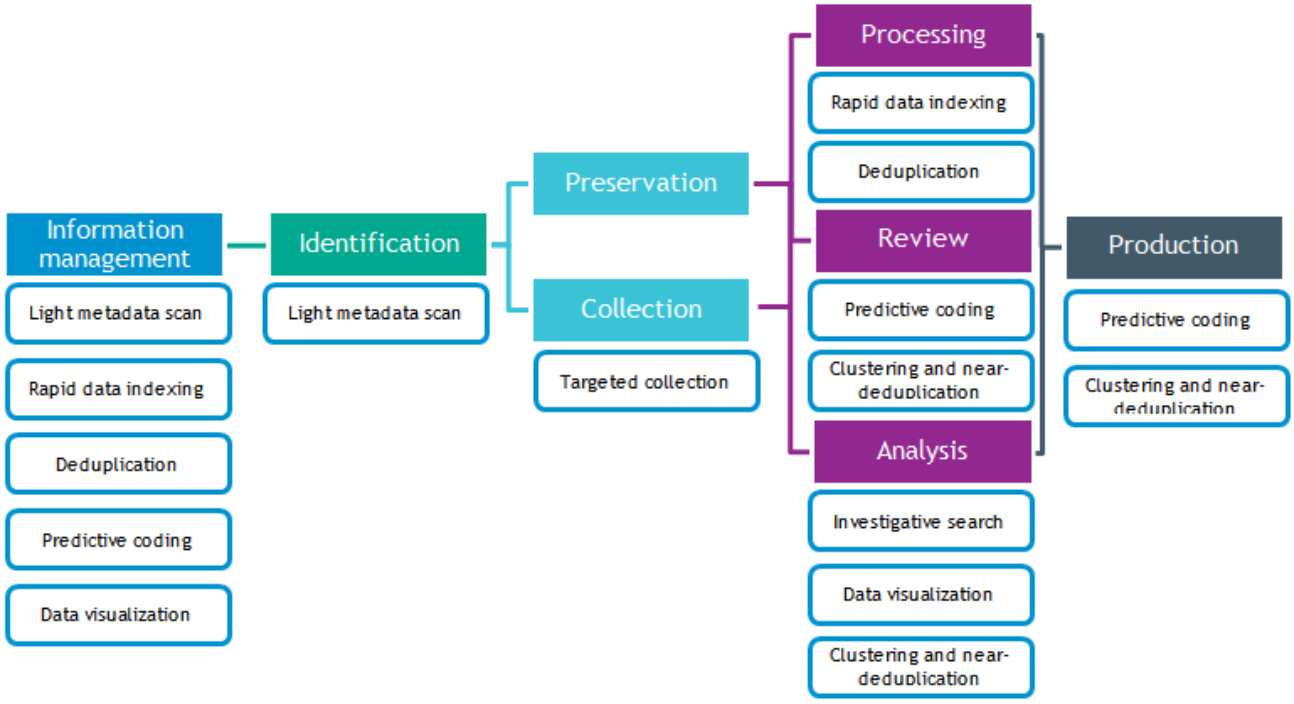

Figure 2: Useful technologies across various stages of the Electronic Discovery Reference Model. [21]

We believe predictive coding alone will not greatly reduce discovery costs because it occurs in the final review before production (see Figure 1). By this point, legal teams should only have to deal with a small number of highly relevant documents.

Rather than applying predictive coding at the eleventh hour, discovery practitioners have a large number of opportunities to make litigation more effective and affordable well before they have reached the review stage. They have a potential arsenal of analytical tools and techniques at their disposal, not least of which, their own expertise and understanding of language.

Several recent papers and articles have discussed the value of language analysis, supplemented by technology, in streamlining data sets for eDiscovery. A white paper by Katey Wood and Brian Babineau from Enterprise Strategy Group examines the utility of language analysis "to understand what each document is about in a large collection, mine its contents for topics of interest, and intelligently set aside irrelevant data." [19]

An article in Forbes by Amanda Jones and Ben Kerschberg offers that "Statistical algorithms for text classification are capable of amazing feats when it comes to detecting and quantifying meaningful patterns amongst large data sets, but they are not capable of making the type of subjective qualitative assessments that constitute the art of discovery." [5]

The technology-assisted linguistic analytics approach we are proposing in this paper combines multiple techniques with attorneys' expertise to reduce the volume and increase the relevance of the documents fed through the litigation process.

6. INCREASING RELEVANCE FROM COLLECTION TO COURT

At each stage of the Electronic Discovery Reference Model (EDRM) process, attorneys can apply linguistic and metadata analyses, using words and their context to cull irrelevant material and focus on the most important documents (see Figure 2).

The ultimate aim of this process is to minimize the number of documents handed over to legal advisors. However, a vital first step is to start with all the facts of the case, gathered from all available custodians and data sources.

6.1 Start With All the Facts

Conducting an extremely thorough investigative workflow can unearth the custodians, documents and facts that traditional approaches would miss, and avoid nasty surprises further down the track.

For example, foreign-language documents often remain hidden in data sets until very late in the review process. At this point, an organization must pay large amounts of money for expedited translation services. By identifying foreign-language documents at the start of the process, the legal team can take a more strategic approach, such as finding staff members who speak the relevant languages. I have seen this save hundreds of thousands of dollars in translation costs.

Digital evidence is typically stored in many different devices and formats. These may include hard drives, file shares, email and collaboration servers, smartphones, tablet devices, flash memory, cloud services, archives, compliance storage repositories and legacy platforms.

Without the ability to collect from and index the contents of all these formats, an organization may face an unknown "smoking gun" document. For example, one party to litigation may only have a copy of an incriminating email in an obsolete email system or archive, while the other party may have kept the same message in an easily readable format.

Speed is also essential. At every stage of the process, organizations must avoid bottlenecks such as technologies that cannot quickly map, collect and analyze data.

6.2 Light Metadata Scan

Starting at the Identification stage of the EDRM model, a light metadata scan involves indexing the contents of each storage repository and extracting file-level metadata such as date, size, file name or subject and owner of each item. It does not analyze the text within each file.

A light metadata scan is many times faster than full text indexing: a single server can process tens of terabytes of data per day. This provides enough information to identify documents an organization must place under legal hold (the Preservation stage of the EDRM) and to conduct a targeted collection processes (Collection).

6.3 Targeted Collection from All Data Sources

Having conducted a light metadata scan, an organization can "pre-filter" the data it collects to the most relevant custodians, dates and document types.

6.4 Rapid Data Indexing

Electronic evidence is almost entirely made up of unstructured data—chiefly words and pictures, but also numbers, dates and facts. Because this data is not stored in neat rows and columns such as a spreadsheet, database or business application, it is much harder for machines to analyze.

Gaining timely insights from huge volumes of unstructured data requires an indexing engine with massively parallel processing capabilities. The engine must be able to deal with the unpredictable and "lumpy" nature of unstructured data, while making the best use of the available processing power. It must also have forensic precision, ensuring every single item fed into it is processed or, if it fails, accounted for.

Indexing all available text and metadata in the Processing stage enables legal practitioners to perform simple and complex searches, clustering, deduplication and near-deduplication, and predictive coding.

6.5 Deduplication

In the average organization, 50–70% of the data they hold is ROT: redundant, obsolete or trivial [20]. Effective deduplication across an entire data set eliminates documents that would otherwise clog up subsequent parts of the eDiscovery process.

6.6 Predictive Coding

As previously discussed, predictive coding is a very useful technology for classifying documents as responsive or unresponsive and for locating privileged documents in production sets without having to review each one manually. Alternatively, attorneys can use predictive coding to prioritize their review strategy. For example, they can rely on the predictive coding engine's decisions for documents with a high confidence score but use humans to review the documents about which the model is less certain.

6.7 Clustering and Near-Deduplication

Near-duplication analyzes the similarity of documents by tallying short phrases or groups of words. This can help identify documents that contain identical text but are in different formats—for example, a Microsoft Word document that had been converted to an Adobe Acrobat PDF file.

This can be helpful in identifying multiple revisions of the same document and placing them on a timeline. It can also show how blocks of text and ideas move across an organization over time.

These techniques are a useful supplement to keyword searches and predictive coding, because they can locate related documents that either technique may have missed. They are also useful in locating similar documents to help fine-tune a predictive coding model.

6.8 Investigative Search

The limitations of keyword searches have been widely discussed. However, attorneys can supplement keyword searches with techniques borrowed from law enforcement and corporate investigators

For example, investigators examine connections between suspects by extracting and cross-referencing intelligence items such as credit card numbers, Social Security numbers, email addresses and IP addresses. eDiscovery practitioners can use similar techniques when looking for information gaps and relationships between custodians.

Law enforcement investigators have also used near-duplicate analysis to uncover criminal networks and scams. For example, investigators looking into a company fraudulently selling aircraft that didn't exist used near-duplicate analysis to locate similar documents. This unearthed several related companies, previously unknown to the investigators, conducting fraudulent transactions for aircraft parts, boats and other high-value products.

Investigators have very recently begun using near-duplication phrase lists as a sophisticated supplement for basic keyword searches.

For example, the list of search terms in the Lehman Bros insolvency case takes up 113 pages. Rather than searching for a term such as "*solven* w/20 (transfer* or mov* or pledg*)", the Examiner could have consulted a list of phrases that contained the words "solvency" or "insolvency" and quickly narrowed this down to a few key phrases that would have yielded a much more targeted group of documents with far fewer false positives.

6.9 Data Visualization

Visualization tools can help lawyers examine relationships

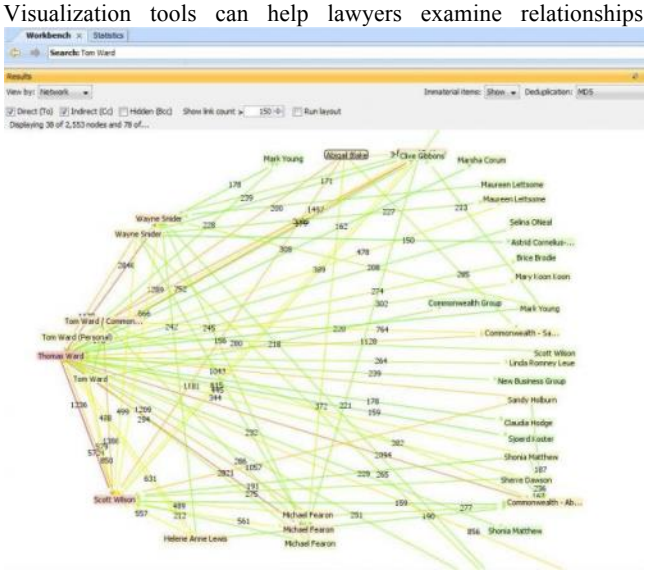

Figure 3: A network diagram showing connections between custodians. [22]

between custodians and the context of important documents.

These include:

- Network diagrams to visualize email trails and other connections between custodians (see Figure 3)
- Timelines to see where each document fits in the context of external events
- Date-trend charts to quickly identify the relevant period of time when the actions in question took place.

7. DO I ALSO NEED TO BE A STATISTICIAN AND A LINGUIST?

The articles by Wood and Babineau [19], and Jones and Kerschberg [5] call for the expertise of linguists and statisticians who can apply their skills toward optimizing the use of these analytical technologies. However, as this paper has sought to demonstrate, the techniques required to streamline eDiscovery using the technology-assisted linguistic analytics approach are easily within the grasp of attorneys and discovery professionals.

8. A FINAL WORD ON INFORMATION GOVERNANCE

Readers will note on the far left of Figure 2 is the Information Management stage of the EDRM, which organizations often neglect. In the tiny gap between Information Management and Identification lies a trigger event such as a summons or regulatory notice. Many organizations take no steps to understand the contents of their information stores until after this trigger event.

However, just as predictive coding for review comes too late in the process to effectively minimize data volumes, trying to remedy information management shortcomings after a trigger event is like putting a band-aid on a gaping wound.

Organizations can also apply many of the techniques we have discussed in this paper proactively, including:

- Regular metadata scans
- Frequently updated "living indexes" of important data stores
- Scheduled and defensible deletion of redundant, obsolete and trivial data
- Searching and remediating legacy storage systems such as archives.

These form part of a proactive information governance regime that enables organizations to discover and address risks in their data before they ever reach a court.

9. ABOUT THE AUTHORS

9.1 Deborah Baron, MBA, BEc

Deborah Baron is Chief Marketing Officer of Nuix. Deborah actively participates in industry forums and conferences, serving on the Advisory Board of the Electronic Discovery Reference Model (EDRM) and as a member of The Sedona Conference Working Groups 1 and 6. She also writes articles for US and international publications; conducts Minimum Continuing Legal Education courses nationally; and regularly meets with clients and industry thought leaders. She recently wrote a chapter for the legal textbook Dispute Resolution and e-Discovery published by Thomson Reuters in 2012. Deborah holds an MBA from the Kellogg Graduate School of Management at Northwestern University and a BA Economics from Occidental College.

9.2 Angela Bunting, [qualifications]

[bio]

9.3 Brian Krupczak, [qualifications]

[bio]

10. REFERENCES

[1] J. R. Baron, "Law in the Age Of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search," *Richmond Journal of Law and Technology,* pp. 5-6, 2011.

[2] *Pension Comm. of Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC,* 2010.

[3] Fulbright & Jaworski LLP, "8th Annual Litigation Trends Report," 2011.

[4] The Cowen Group, "Q2, 2012 Quarterly Critical Trends Corporate Market Snapshot," 2012.

[5] A. Jones and B. Kerschberg, "What Technology-Assisted Electronic Discovery Teaches Us About the Role of Humans in Technology," *Forbes.com,* 9 January 2012.

[6] N. Pace and L. Zakaras, "Where the Money Goes: understanding litigant expenditures for producing electronic discovery.," 2012.

[7] A. Valukas, "Report of Anton R. Valukas, Examiner In re Lehman Brothers Holdings Inc., No. 08-13555 (JMP) (Bankr. SDNY., Mar. 11 2010)," 2010.

[8] The Sedona Conference, "Commentary on Achieving Quality in E-Discovery," 2009.

[9] complexdiscovery.com, "Predictive Coding One-Question Provider Implementation Survey," 2013. [Online]. Available: http://www.complexdiscovery.com/info/2013/03/05/running-results-predictive-coding-one-question-provider-implementation-survey/.

[10] I. Rish, "An empirical study of the naive Bayes classifier," 2001.

[11] T. I. Barnett and S. Godjevac, "Faster, Better, Cheaper Legal Document Review, Pipe Dream or Reality?," in *ICAIL 2011/DESI IV: Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, 2011.

[12] H. Roitblat, A. Kershaw and P. Oot, "Document Categorization in Legal Discovery: Computer Classification vs. Manual Review," *Journal of the American Society for Information Science and Technology,* vol. 61, no. 1, 2009.

[13] M. Grossman and G. Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," *Richmond Journal of Law and Technology,* vol. XVII, no. 3, 2011.

[14] N. Silver, The Signal and the Noise: why most predictions fail but some don't, The Penguin Press, 2012.

[15] A. Peck, "Search, Forward," *Law Technology News,* October 2011.

[16] *Da Silva Moore v. Publicis Groupe,* 2012.

[17] *Global Aerospace Inc. v. Landow Aviation, L.P.,* 2012.

[18] *Kleen Products, LLC v. Packaging Corporation of America,* 2011.

[19] K. Wood and B. Babineau, "Review Acceleration: Leveraging Language," Enterprise Strategy Group, 2012.

[20] K. Wood, "Defensible Deletion: Quantifying the Benefits.," Enterprise Strategy Group, 2012.

[21] Nuix, "Reducing the Costs of eDiscovery from Collection to Court," 2013.

[22] International Consortium of Investigative Journalists, "How ICIJ's Project Team Analyzed the Offshore Files," 3 April 2013. [Online]. Available: http://www.icij.org/offshore/how-icijs-project-team-analyzed-offshore-files.

[23] "The Electronic Discovery Reference Model," [Online]. Available: http://www.edrm.net/.