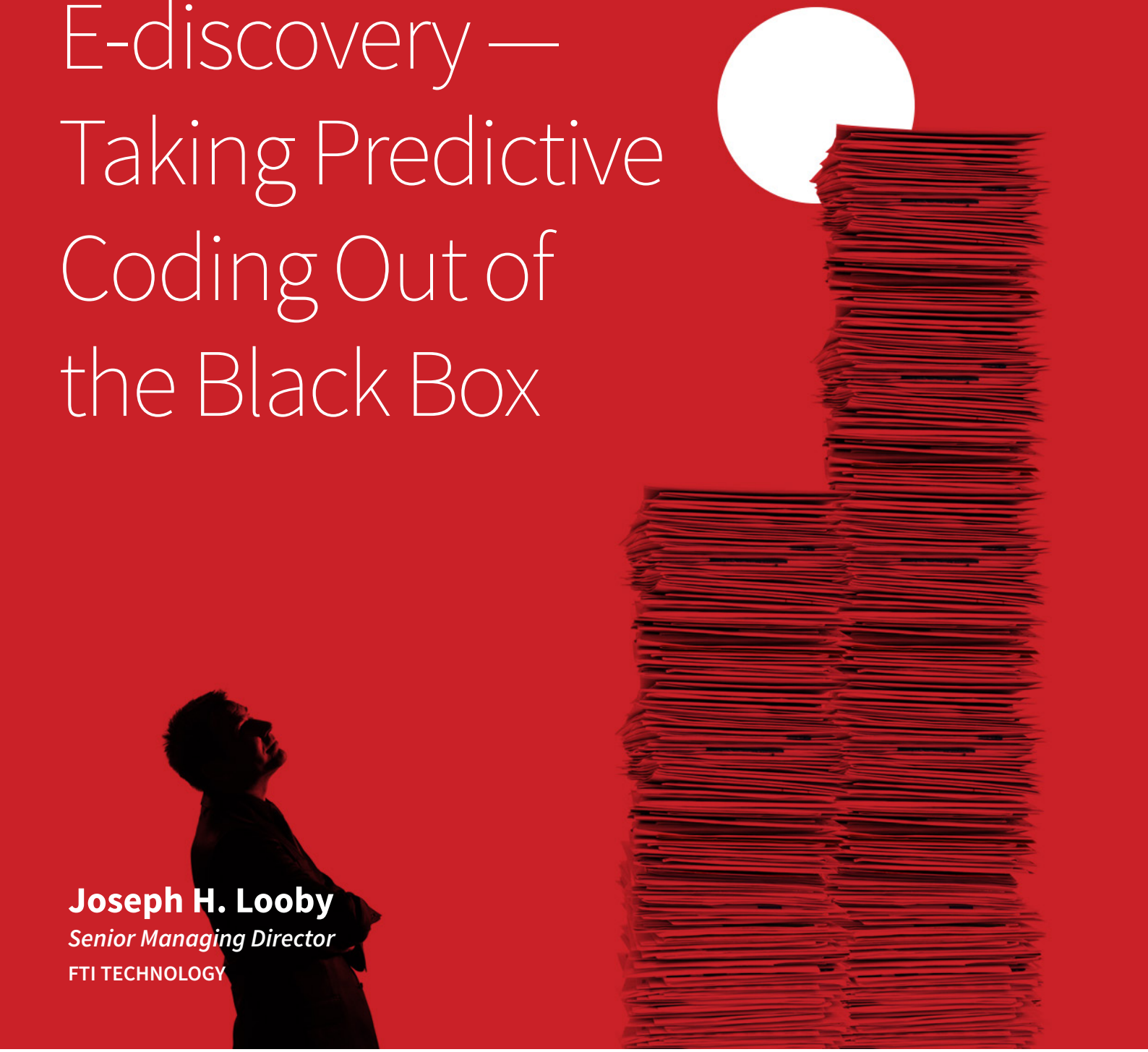
E-discovery — Taking Predictive Coding Out of the Black Box



Joseph H. Looby
Senior Managing Director
FTI TECHNOLOGY

IN CASES OF COMMERCIAL LITIGATION, *the process of discovery can place a huge burden on defendants. Often, they have to review millions or even tens of millions of documents to find those that are relevant to a case. These costs can spiral into millions of dollars quickly.*

anual reviews are time consuming and very expensive. And the computerized alternate – keyword search – is too imprecise a tool for extracting relevant documents from a huge pool. As the number of documents to be reviewed for a typical lawsuit increases, the legal profession needs a better way.

Illustration by Big Human

Predictive discovery is that approach. Predictive discovery is a machine-based method that rapidly can review a large pool of documents to determine which are relevant to a lawsuit. It takes attorneys' judgments about the relevance of a sample set of documents and builds a model that accurately extrapolates that expertise to the entire population of potential documents. Predictive discovery is faster, cheaper and more accurate than traditional discovery approaches, and this year courts in the United States have begun to endorse the use of this method. But according to a survey FTI Consulting recently conducted, a significant obstacle stands in the way of widespread adoption: Most attorneys do not understand how predictive discovery works and are apprehensive about using it.

In this article, we explain how predictive discovery functions and why attorneys should familiarize themselves with it.

Companies defending themselves against litigation frequently must produce documents that may be relevant to a case by searching through archives. The same is true for companies being investigated by a regulatory body. The process of plumbing one's own archives to produce relevant documents is known as discovery. When the documents are in electronic form, the process is called e-discovery. Large companies today often must review millions or even tens of millions of letters, e-mails, records, manuscripts, transcripts and other documents for this purpose.

The discovery process once was entirely manual. In the 1980s, the typical commercial lawsuit might have entailed searching through tens of thousands of documents. A team of attorneys would pore through all the paper, with perhaps one or two subject matter experts to define the search parameters and oversee the procedure. A larger number of more junior lawyers frequently would do the bulk of the work. This process (known in the trade as linear human or manual review) would cost a few thousand dollars (today, it runs about $1 per document).

But now, the large number of documents to be reviewed for a single case can make manual searches a hugely expensive and time-consuming solution. For a case with 10 million documents, the bill could run about $10 million — a staggering sum. Conducting keyword searches of online documents, of course, is much faster. But in certain circumstances, such searches have been shown to produce as little as 20 percent of the relevant documents, along with many that are irrelevant.

With the number of documents in large cases soaring (we currently have a case involving 50 million documents), costs are getting out of hand. Yet the price of failing to produce relevant documents can be extraordinarily high in big-stakes litigation. The legal profession needs a new approach that improves reliability and minimizes costs.

There now is a third approach. Predictive discovery takes expert judgment from a sample set of documents relevant (or responsive) to a case and uses computer modeling to extrapolate that judgment to the remainder of the pool. In a typical discovery process using predictive coding, an expert reviews a sample of the full set of documents, say 10,000-20,000 of them, and ranks them as responsive (R) or non-responsive (NR) to the case. A computer model then builds a set of rules that reflects the attorney's judgment on the sample set. Working on the sample set and comparing its verdict on each document with the expert's, the software continually improves its model until the results meet the standards agreed to by the parties and sometimes a court. Then the software applies its rules to the entire population of documents to identify the full responsive set.

Studies show that predictive coding usually finds a greater proportion — typically in the range of 75 percent (a measure known as recall) — of the responsive documents than other approaches. A seminal 1985 study found keyword searches yielded average recall rates of about 20 percent.[1] A 2011 study showed recall rates for linear human review to be around 60 percent.[2] And, according to these same studies, the proportion of relevant documents in the pool of documents identified by predictive coding (a measure known as precision) is better, too.

The cost of predictive discovery is a fraction of that for linear manual review or keyword search, especially for large document sets, because most of the expense is incurred to establish the upfront search rules. Afterwards, the cost increases only slightly as the size of the collection grows.

In a recent engagement, we were asked to review 5.4 million documents. Using predictive discovery, we easily met the deadline and identified 480,000 relevant documents — around 9 percent of the total — for about $1 million less than linear manual review or keyword search would have cost.

Definition: **Trial and Error**

The process of experimenting with various methods of doing something until one finds the most successful.

So how does predictive discovery work? As suggested above, there are five key steps:

1. Experts code a representative sample of the document set as responsive or non-responsive.
2. Predictive coding software reviews the sample and assigns weights to features of the documents.
3. The software refines the weights by testing them against every document in the sample set.
4. The effectiveness of the model is measured statistically to assess whether it is producing acceptable results.
5. The software is run on all the documents to identify the responsive set.

1. "An Evaluation Of Retrieval Effectiveness For A Full-Text Document Retrieval System," by David C. Blair of the University of Michigan and M.E. Maron of the University of California at Berkeley, 1985
2. "Technology-Assisted Review in E-discovery Can Be More Effective and More Efficient than Exhaustive Manual Review," by Maura R. Grossman and Gordon V. Cormack, *Richmond Journal of Law and Technology*, Vol. XVII, Issue 3, 2011.

Let's look at these steps in more detail.

1 *Experts code a representative sample of the document set as responsive or non-responsive.*

In a typical case, a plaintiff might ask a company to produce the universe of documents regarding some specifics about Project X. The company prepares a list of its employees involved in the project. From all the places where teams of employees working on Project X stored documents, the firm secures every one of the potentially relevant e-mails and documents. These can easily total 1 million or more items.

These documents are broken into two subsets: the sample or training set and the prediction set (i.e., the remainder). To have a high level of confidence in the sample, the training set must be around 17,000 documents. One or more expert attorneys codes the training set's documents as R or NR with respect to the litigation.

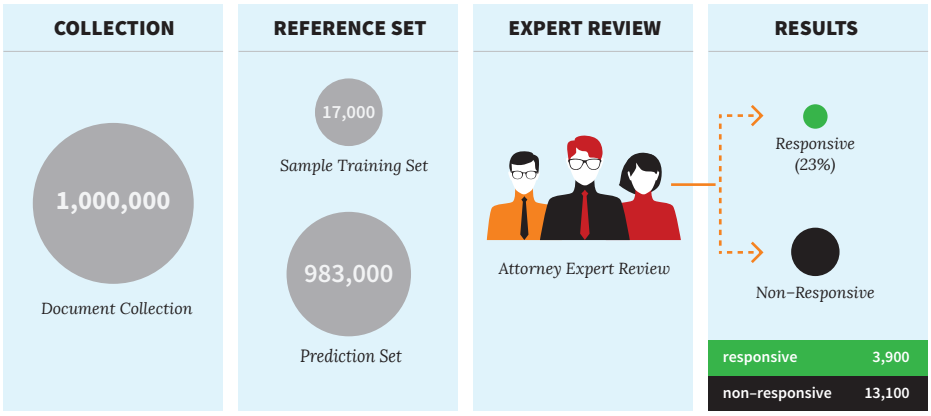
The results of this exercise will predict with reasonable accuracy the total number of Rs in the full set. A statistically valid random sample of 17,000 documents generally is sufficient to ensure that our answer will be accurate to +/-1 percent at the 99 percent confidence level. For example, if the number of R documents in the sample set is 3,900 (23 percent of 17,000), it is 99 percent certain that between 22 to 24 percent of the 1 million documents are responsive.

2 *Predictive coding software reviews the sample and assigns weights to features of the documents.*

For the sample set of documents, the software creates a list of all the features of every document in the sample set. Features are single words or strings of consecutive words; for example, up to three words long. The sentence "Skilling's abrupt departure will raise suspicions of accounting improprieties and valuation issues …." for instance, yields 33 features as shown below.

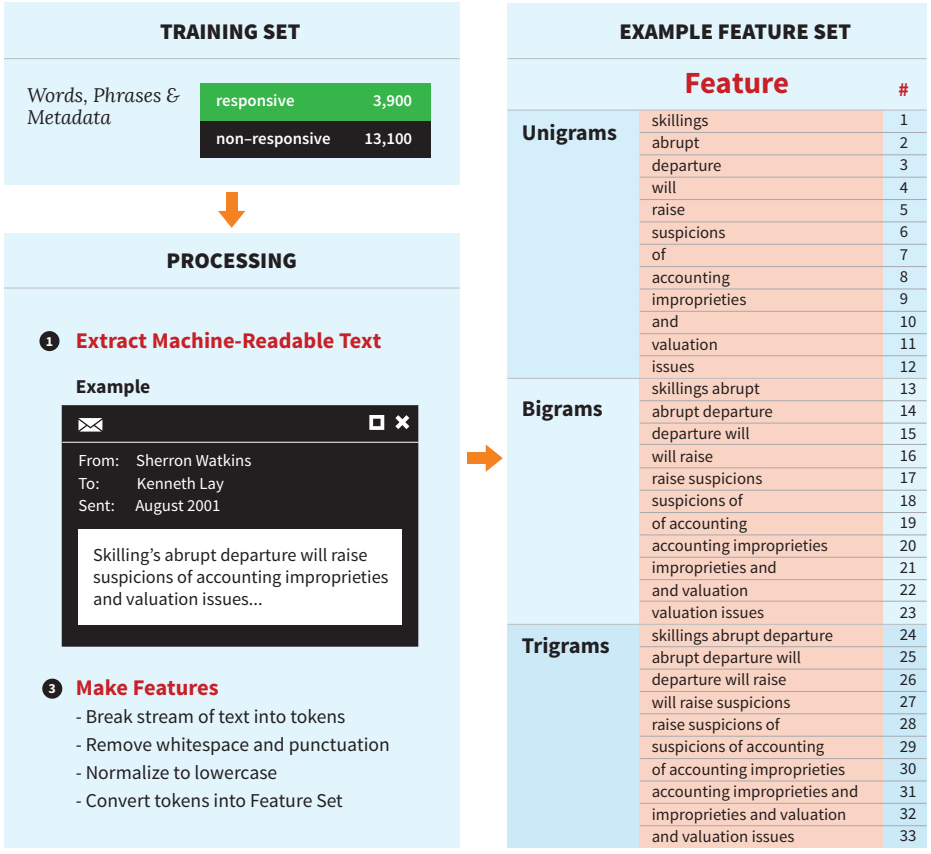
The total number of features in a set of 1 million documents is huge, but even large numbers of features easily can be processed today with relatively inexpensive technology.

Each feature is given a weight. If the feature tends to indicate a document is responsive, the weight will be positive. If the feature tends to indicate a document is non-responsive, the weight will be negative. At the start, all the weights are zero.



Estimate Of Goal

We are 99% confident that 23% (+/- 1%) of the 1,000,000 documents are responsive.



3 *The software refines the weights by testing them against every document in the sample set.*

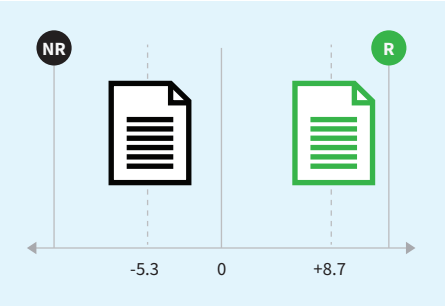
The next step is one of trial and error. The model picks a document from the training set and tries to guess if the document is responsive. To do this, the software sums the importance weights for the words in the document to calculate a score. If the score is positive (+), the software guesses responsive, or R — this is the trial.

If the model guesses wrong (i.e., the model and attorney disagree), the software adjusts the importance weights of the words recorded in the model's weight table. This, of course, is the error. Gradually, through trial and error, the importance weights become better indicators of responsiveness. Pretty soon, the computer will have created a weight table with scores of importance and unimportance for the words and phrases in the training set, an excerpt of which might look like the illustration at the right.

As the software reviews more documents, the weight table becomes an increasingly better reflection of the attorney expert's judgment. This process is called machine learning. The computer can take several passes through the sample set until the weight table is as good as it can be based on the sample.

WEIGHT TABLE

Feature	Weight
will raise	-0.6
suspicions	0.7
of	-
improprieties and valuation	1.8



Once the weight table is ready, the program is run on the full sample to generate a score for every document according to the final weight table. The machine score for a document is the sum of the weights of all its features. In the illustration to the left, we show two documents: a document with a score of -5.3 (probably non-responsive) and a document with a score of +8.7 (probably responsive).

Then, if predictive coding is to be used to select the highly responsive documents from the collection of 1 million and to discard the highly non-responsive documents, a line has to be drawn. The line is the minimum score for documents that will be considered responsive. Everything with a higher score (i.e., above the line), subject to review, will comprise the responsive set.

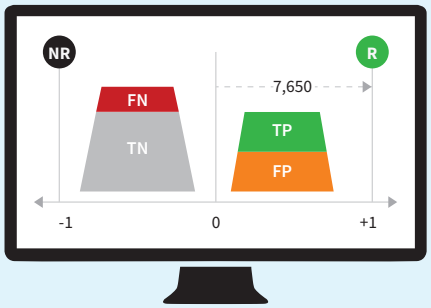
In our 17,000 document sample, of which the expert found 3,900 to be R, there might be 7,650 documents that scored more than zero. Of these, 3,519 are documents the expert coded R, and 4,131 are not. Therefore, with the line drawn at 0, we will achieve a recall of 90 percent (3,519 divided by 3,900) and a precision of 46 percent (3,519 divided by 7,650).

ATTORNEY EXPERT REVIEW

responsive	3,900
non–responsive	13,100
	17,000

Vs

COMPUTER MODEL (AT >0)



Found

True +	3,519	90% Recall
False +	4,131	46% Precision
	7,650	

FALSE POSITIVES
TRUE POSITIVES
FALSE NEGATIVES

4 *The effectiveness of the model is measured statistically to assess whether it is producing acceptable results.*

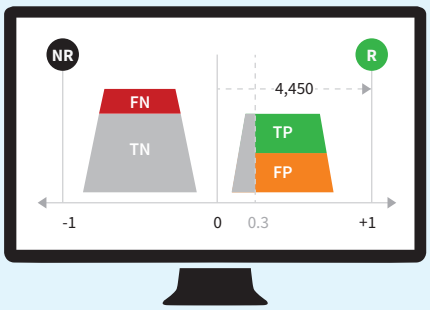
By drawing the line higher, one can increase the ratio of recall to precision. For example, if the line is drawn at +0.3, the precision may be significantly higher (say 70 percent) but with lower recall (i.e., 80 percent because some of the Rs between 0 and +0.3 are missed). On the basis of the sample set, the software can determine recall and precision for any line. Typically, there is a trade-off; as precision goes up (and the program delivers fewer NR documents), recall declines (and more of the R documents are left behind). However, not only are recall and precision scores higher than for other techniques, the trade-off can be managed much more explicitly and precisely.

ATTORNEY EXPERT REVIEW

responsive	3,900
non-responsive	13,100
	17,000

Vs.

COMPUTER MODEL (AT >0.3)



Found

True +	3,125	80% Recall
False +	1,325	70% Precision
	4,450	

FALSE POSITIVES | TRUE POSITIVES | FALSE NEGATIVES

Recall measures how well a process retrieves relevant documents. For example, 80% recall means a process is returning 80% of the estimated responsive documents in the collection.

80% = 3,125 / 3,125 + 775

Recall | True Positives (TP) | True Positives (TP) | False Negatives (FN)

Precision measures how well a process retrieves only relevant documents. For example, 70% precision means for every 70 responsive documents found, 30 non-responsive documents also were returned.

70% = 3,125 / 3,125 + 1,325

Precision | True Positives (TP) | True Positives (TP) | False Positives (FP)

5 *The software is run on all the documents to identify the responsive set.*

Once the trade-off between recall and precision has been made and the line determined, the full set of documents can be scored. The outcomes generally will be close to those predicted, although there usually is some degradation because the sample set never perfectly represents the whole.

As a final quality check, samples can be taken from both above and below the line to make sure results meet expectations. In most cases, samples of 2,000 are adequate, giving a +/-3 percent margin of error (at a 99% confidence interval). Experts then review these random samples against the computer rating. They should confirm that approximately as many Rs occur above the line and as few beneath it as predicted.

Why This is Better than Keyword Search

Attorneys have been using search terms for e-discovery for many years. But the biggest drawback of search is that people determine the terms on which to search. A set of 1 million documents may have 100,000 distinct words. People easily can come up with some likely terms — for example, the name of the project under investigation and the key players. But this is a primitive way of mining a complex set of documents. Consequently, recall and precision typically are low.

With predictive coding, attorneys begin with no preconceptions about the most relevant words. They let a computer figure that out, considering each word and every combination of two and three consecutive words without constraint. Further, predictive coding can use the fact that certain words will indicate against responsiveness. Processing this volume of possibilities is much too big a manual task. But, of course, the procedure can be done easily (and cheaply) by a computer. And it produces far better results than keyword search.

Is The World Ready For Predictive Discovery?

Predictive discovery is not new. With patents dating back to 2005, FTI Consulting was the first firm to use machine learning for commercial e-discovery. Yet adoption has been slow. In a survey conducted this year, we found only about half of corporate in-house and external lawyers are using predictive discovery. But even companies using the technology are largely experimenting with it to cull and prioritize documents for manual review. [3]

Why has the uptake been slow? Our study and experience suggest two primary reasons: 1) a reluctance to invest in something the courts might not support and 2) a widespread lack of understanding of how predictive coding works. Two recent court cases are helping eliminate the first concern:

- Da Silva Moore: In February, both parties in this case agreed to e-discovery by predictive coding. This was the first validation of the use of predictive coding by a court.
- Landow Aviation: In April, a Virginia state court judge permitted the defendant (over plaintiff's objection) to use predictive coding to search an estimated 2 million electronic documents.

The second concern, though, has yet to be overcome. Our recent survey found that one of the biggest impediments to the adoption of predictive coding is that many lawyers do not know how it works.

As we have tried to show here, the principles are basic; and we intentionally keep the computer model's math simple, too. For example, the machine score for a document is the sum of the weights of all its features. There are no more straightforward arithmetic functions than addition and subtraction.

As courts begin to endorse predictive coding and judges articulate its advantages, attorneys will not want to be left behind.

The Way Forward

As we mentioned, we found when we surveyed in-house counsel and outside lawyers that most are using predictive coding to cull the document set to reduce the task for the follow-on manual review. The need for some human review never will go away — with predictive coding, manual review is required for coding the sample set, learning the facts of the case, and for identifying privileged documents, for example. But predictive coding can take a higher proportion of the load in e-discovery than most are allowing it to do today.

To be sure, predictive discovery isn't right for every e-discovery process. Cases with many documents in which only a handful are expected to be responsive — searches for needles in a haystack — can be approached differently. Smart people, process, and a variety of technologies can address these varying and different research goals.

But for most companies that have millions of documents to review in e-discovery, predictive discovery is a lower-cost solution with greater reliability and a higher level of transparency that should be used more frequently than it is today. ■

3 Advice from counsel: Can predictive coding deliver on its promise? by Ari Kaplan of Ari Kaplan Advisors and Joe Looby of FTI Consulting, 2012.

Joe Looby
Senior Managing Director, FTI Technology
joe.looby@fticonsulting.com

For more information and an online version of this article, visit ftijournal.com.