Measurement in eDiscovery

A Technical White Paper

Herbert Roitblat, Ph.D.

CTO, Chief Scientist

Measurement in eDiscovery

From an information-science perspective, eDiscovery is about separating the responsive documents from the nonresponsive ones. For the attorneys, of course, eDiscovery involves a lot more, but this information perspective is a natural precursor to the legal analysis. This separation, sometimes called first-pass review or even early case assessment (again focusing on the information-science perspective) has traditionally been done by having people review each document—also called linear review.

In recent years technological tools have emerged to make the process of eDiscovery, particularly first pass review, easier and more automated. These tools are often collectively referred to as Computer Assisted Review, Technology Assisted Review, Assisted Review Technology, or just Predictive Coding. Along with these tools, and a healthy skepticism about their effectiveness, has come an interest in measuring the effectiveness of the eDiscovery process, whether conducted using these technology tools or by the more traditional approach of having teams of people read every document.

As the use of these technological tools becomes more widespread, a number of protocols have evolved to guide their use in eDiscovery. Measurement is a key feature of these protocols, so a thorough understanding of measurement in eDiscovery can be a substantial advantage in negotiation over these protocols. Some of these protocols are unnecessarily complex. Others focus on assessments that may not actually provide much information. Understanding the nature of measurement in eDiscovery will facilitate the effective use of technological tools and help to moderate the costs of eDiscovery.

FRCP rule 26(g) requires the party producing data to sign a declaration in which "an attorney or party certifies that to the best of the person's knowledge, information, and belief formed after a reasonable inquiry (A) with respect to a disclosure, it is complete and correct as of the time it is made…" Part of assessing whether an inquiry has been reasonable is the degree to which it produces complete and accurate information. Until recently, however, the completeness of disclosure has rarely been measured explicitly. The advent of these technological tools has changed that. As a result, we have found that effort is only a weak indicator of the success of the inquiry and more explicit measurement has become more widespread.

Measures

The results of first pass review can be categorized into a small table, called a contingency table. The rows represent the true way to categorize documents as responsive or not responsive and the columns represent the output of the process we are trying to measure. Most of the measures used in information retrieval are derived from this table.

Table 1. Contingency Table for Information Retrieval

		Predicted: Responsive	Predicted: Non-responsive	
True	Responsive	A	B	E
	Non-responsive	C	D	F
		G	H	J

This table shows all combinations of predicted and true categorization of the documents in a collection. Cell A, for example, shows how often the process correctly predicted that a document was responsive. Cell D shows how often the process correctly predicted that a document was non-responsive. Cell B shows how often the process missed a responsive document and Cell C shows how often the process called a document responsive when it was not truly responsive.

E represents the number of documents that were truly responsive (A + B). F represents the number of documents that were truly non-responsive (C + D). G shows the number of documents that were classified as responsive by the process (A + C) and H refers to the number of documents that were classified as non-responsive(B + D). J refers to the total number of documents in the collection (A + B + C + D).

Prevalence or richness is the proportion of responsive documents relative to all documents. It is computed as E / J.

The simplest measure of correctness is the proportion correct often called Accuracy, (A + D) / J, the number of correctly classified documents divided by the total number of documents.

The two most commonly used measures of correctness and completeness are Precision and Recall. Precision is the proportion of the retrieved documents (those classified as responsive) that are, in fact responsive. Precision indicates the specificity of the process. In this table it is computed as A / G (where G is the total of A + C). Recall is the proportion of the truly responsive documents that were identified by the process. Recall indicates the completeness of the process in finding all of the responsive documents. Recall is computed as A/E (where E is the total of A + B).

Precision and Recall focus on the number of responsive documents that are correctly classified. No attention is given to correct classification of the non-responsive documents. eDiscovery, however, is symmetrical, especially first-pass review. Success can come from either correctly selecting the responsive documents or from correctly excluding the non-responsive ones. A spam filter, for example, is successful when it correctly identifies and rejects junk emails. Two other measures that could be used would be to compute Precision and Recall for recognizing the non-responsive documents. These can be called Inverse Precision and Inverse Recall. Which document class is called positive and which is negative is a conventional choice, it could be done the other way around.

In eDiscovery, the proportion of non-responsive documents is typically much higher than the proportion of responsive documents. Non-responsive documents are easier to find and so there may be some

advantages to using Inverse Recall and Precision from a sampling perspective. I'll return to that idea in a bit.

Elusion is another measure that has received some attention in eDiscovery. Elusion is the proportion of rejected documents that are truly responsive (B / H). Of the documents that are designated non-responsive by the system, what proportion of those are actually responsive? Better processes will erroneously reject few if any responsive documents.

There are other measures that can be computed from these tables as well. Table 2 shows some of the direct measures of efficacy that can be derived from the contingency table. The direction column shows whether better performance is a higher number (+) or a lower number (-).

Table 2. Information Retrieval Measures.

Measure	Formula	Direction	Description
Accuracy	(A+D)/J	+	Overall agreement between authority and prediction
Precision	A/(A+C)	+	Specificity of prediction
Recall	A/(A+B)	+	Completeness of prediction
Inverse Precision	D/(D+B)	+	Specificity of prediction
Inverse Recall	D/(D+C)	+	Completeness of prediction
Dice	A/(A+(B+C)/2)	+	Overlap of positive predictions
Jaccard	A/(A+B+C)	+	Overlap of positive predictions
Fallout	C/(C+D)	-	Proportion of true negatives that are incorrect
Correct Rejection Rate	D/(C+D)	+	Proportion of the true negatives that are correct, the degree to which non-responsive documents have been rejected
Elusion	B/(B+D)	-	Proportion of predicted negatives that are incorrect
Miss rate	B/(A+B)	-	Proportion of true positives that are missed

There are other measures, such as F1, d', AUC, Phi, and Matthews Correlation Coefficient, that can be further derived from the contingency table. F1 is a kind of average of Precision and Recall, which weights each of them equally. D' and AUC are measures derived from optimal decision theory that measure the ability of a system to distinguish positive from negative items, independent of bias. Phi and Mathews Correlation Coefficient measure the degree to which the prediction accuracy departs from chance. These derived measures can be very useful, but they will not be considered further here.

Accuracy is limited by the fact that in eDiscovery, the non-responsive documents are much more common than the responsive ones. Although, this measure is easy to compute, it has to be interpreted carefully because of the typically high volume of nonresponsive documents in an eDiscovery collection. If 90% of the documents in a collection are non-responsive, then a process that called every document non-responsive would be 90% correct. When interpreting this measure of accuracy, one needs to keep this baseline in mind. It is difficult to compare accuracy obtained with one collection with the accuracy obtained with another when they different proportions of responsive documents.

Precision is easy to compute, but it does not tell us about the completeness of our search. Recall is a challenge to compute because you need to know the number of correct documents in a collection in order to know how many of these were retrieved. To compute A / E you need to know A and E. Inverse Precision and Inverse Recall have many of the same limitations, plus, they emphasize the negative aspects of discovery, rather than the positive.

Dice and Jaccard measure the overlap between two sets of judgments or between one set of judgments and the correct choice. These measures can be strongly affected by the bias of the judge to call a document responsive. If one judge calls more documents responsive than the other, their overlap will be smaller and so they will appear to perform more poorly.

Fallout is, in some sense, the opposite of recall. It expresses the relative number of truly nonresponsive documents that have been falsely classified as responsive. Fallout is also sometimes called the False-Alarm Rate. Fallout tells us about one kind of error, but by itself, it is incomplete for measuring accuracy. A fallout rate of 0 can be trivially achieved by never categorizing any documents as responsive.

Correct Rejection Rate is the complement of Fallout. It is the number of truly non-responsive documents that are classified as non-responsive. It, too is an incomplete measure. A high correct rejection rate can be trivially achieved by categorizing all documents as non-responsive.

Elusion is a measure of the responsive documents that have been missed. Ideally, elusion should be zero. Of the documents classified as non-responsive, ideally, none of them should truly be responsive. Elusion, in combination with an estimate of prevalence, can be used to estimate recall with very little effort. Instead of reviewing enough documents to have a reasonable sample of responsive ones for computing Recall directly, a relatively small number of documents categorized as non-responsive can be reviewed and an estimate of Recall can be computed from this sample.

Elusion can also easily be used as the basis for a quality control measure. "Accept-on-zero" quality control tests are widely used in industrial settings. They are called "Accept-on-zero" tests because they can be passed only if there are no responsive documents in a small sample. To perform this test, one must choose a criterion level of Elusion, say 1.25%. This value is the maximum acceptable percentage of responsive documents left behind in those that have been classified as non-responsive. It should be a small fraction of the prevalence identified for that collection.

See http://alias-i.com/lingpipe/docs/api/com/aliasi/classify/PrecisionRecallEvaluation.html for more measures.

How to choose a measure

All of these measures are roughly equivalent in that they all measure some aspect of efficacy. Except for Elusion, Fallout, and Miss rate, they all increase when the number of correct decisions increases (all other things being equal). Elusion, fallout, and Miss rate decrease as the number of correct predictions increases. Each measure has a specific emphasis and specific biases.

As in all things statistical, the right measure to use depends on the question that you are trying to answer. The question, in turn, depends on the argument that you are trying to make and on your goals. In addition, the right measure depends on the resources you have available for answering that question. The measure you select has to be not only theoretically valid, but practically attainable. The best measure is useless if you don't have the wherewithal to collect it.

The eDiscovery goal for this process is to separate the responsive from the non-responsive documents and we would like to know how successful we were at doing that. Furthermore, we would like to do this assessment with as little effort as possible.

Information science has focused largely on Precision and Recall for a number of reasons. Psychologically, it seems more satisfying to focus on the kind of documents that we do want, rather than on the ones that we do not want (e.g., Inverse Recall). The proportion of relevant documents in a collection (prevalence or richness) is typically much smaller than the proportion of non-relevant ones. If only a certain number of documents could be assessed for true relevance, then that effort would have to be focused on the documents that might be relevant—those retrieved by the information retrieval system. With the advent of the web, the number of non-relevant documents is, for practical purposes, infinite, so there is no way to measure the non-relevant documents in web searches.

Many of the TREC tracks (not just the Legal Track) have focused on Precision and to a lesser degree Recall because that fit in very well with their goals and questions. These measures, or variations of them, were the right measures for the purposes for which they were used. As the size of the collections in the various TREC tracks grew, the organizers recognized that they could not validate all of the documents. They focused their effort on those that were retrieved by one or more systems, and only peripherally sampled those that were not selected as relevant by any system. This approach allowed Precision to be estimated quite well, but the true level of recall was not so accurately calculated because any document that was not retrieved by any of the systems being evaluated had only a disproportionately tiny chance of being sampled. For TREC purposes this was fine because they were interested in the relative performance of systems, not their absolute performance. The TREC Legal Track handled this problem a little differently, especially during the last few years. I do not want to imply any criticism of TREC or its measures, they were the right measures for what they were trying to accomplish. But, they are not the only measures or the only practical measures for eDiscovery.

eDiscovery protocols often call for the use of sampling because it would be prohibitively expensive to review all of the documents in a collection to determine which are truly responsive and which are not. With the ever-increasing size of document collections, it is simply not cost-effective to read every document. Sampling allows us to assess a subset of documents and on the basis of that subset make inferences about the collection as a whole. A random sample from a population provides the most representative sample of the whole collection and, therefore, the soundest inferences from sample to population. The reliability of the estimate depends on the sample size. The larger the sample size, the more reliable is the estimate.

First pass review returns a group of documents that are identified by the process (whether computer or human readers) as responsive. Our goal is to evaluate the claim that all (actually as many as is reasonably possible) and only (actually as exclusively as is reasonably possible) responsive documents have been identified.

FRCP Rule 26(g) would seem to be tailor made for Recall to be measured, but there is a problem. Remember that Recall is the proportion of responsive documents that were identified by the system as responsive. To measure that, we need to know which documents were, in fact, truly responsive, but we do not really know that up front. If we did, we would not have to do the review. This is a dilemma.

One way out of this dilemma is to draw a random sample of documents without regard to the classification assigned by the process and select from this sample those that are truly responsive. In eDiscovery, the proportion of responsive documents is typically fairly low. We tend to see that around 5 to 10% of the documents in an average collection are responsive, though we have seen higher (up to around 50%) and lower (around 0.5%) prevalence on occasion. Others report even lower prevalence. Prevalence depends, among other things, on how the data were collected prior to analysis.

Keeping it simple, let's assume that 10% of the documents are responsive—prevalence is 10%. In order to evaluate the success of the system, we would need a sample of, say 400 responsive documents (to achieve 95% confidence and a 5% margin of error. To compute Recall, we would find the proportion of these 400 documents that were correctly predicted by the system. That would tell us how complete our process was.

In order to find 400 responsive documents at 10% richness or prevalence, we would have to sample approximately 4,000 documents, randomly chosen from the collection as a whole, without regard to whether they were predicted to be responsive or not (10% of 4,000 is 400). That's a lot of work, and it may be more than the number of documents needed to train the process in the first place (if we are using predictive coding). If prevalence is lower, if only a small percentage of documents is actually, responsive, measuring recall directly can be even more costly, because a still large sample of random documents would have to be examined. Calculating Recall directly can be done, but it takes a very substantial amount of work just to find the responsive documents to measure.

There is an added problem in trying to assess a large sample of documents. This sample may be too large for a single authoritative reviewer to assess. When we include multiple reviewers to find our sample of responsive documents we run into the fact that reviewers differ in their criteria for calling a document responsive. Professional eDiscovery reviewers agree with one another on relevance calls only about 50% of the time (Roitblat, Kershaw, & Oot, 2010). So in addition to the variability inherent in sampling from a large population, we also have the variability due to differing judgments of responsiveness. Some of the documents that we intend to be truly responsive will actually be false positives—this is a shaky standard against which to compare the performance of our process.

Fortunately, there are other ways to assess whether we conducted a reasonable inquiry with far less effort. One of these can be called "differential prevalence," which is just Elusion.

Differential Prevalence—another view of Elusion

One way to think about differential prevalence is to assess the degree to which we removed the responsive documents from the collection as a whole. This approach involves a variation on the contingency table shown above. If we start with the knowledge that, say, 10% of the documents are responsive and then find after our removal process that few if any responsive ones remain, then we have been successful at separating the responsive from the non-responsive documents.

Instead of counting the responsive documents that we found, we count the ones that we left behind. We measure the prevalence of responsive documents in the collection as a whole and then we measure the prevalence of responsive documents after we have applied our process to identify and remove the responsive documents. We are successful to the extent that the post-selection prevalence is close to zero.

To perform this analysis, we assess a random sample of documents before doing any first-pass review to determine the prevalence, and then assess a random sample of the documents that are rejected by the process to determine their prevalence in the remainder set. The difference between the two reflects the number of responsive documents that have been removed. Simplifying a little, if our initial prevalence was 10% and our final prevalence after removing the responsive documents identified by the system is 1%, then we must have correctly identified roughly 90% of the responsive documents--Recall.

In order to compute recall more precisely, we need to know four things: Prevalence, Elusion, the total number of documents, and the total number of documents designated responsive by process. Two of these (Prevalence and Elusion) need to be estimated from samples. The other two (the number of documents and the number predicted responsive) are known exactly. Knowing these values, we can compute all of the entries in the contingency matrix and can get any of the measures we want with a relatively small investment in documents reviewed.

If we have an estimate of prevalence, either from a random sample drawn before first-pass review or one drawn in the process of the first-pass review (OrcaTec Predictive Coding draws completely random samples of documents during training), then we have enough information to fill out the contingency table.

If you are familiar with the game Sudoku, then you know about the idea of constraints in a table. In Sudoku, the rows have to add up to a certain number, the columns have to add up, the squares have to add up, etc. Let's start filling out a contingency table in the same way.

First, we know the total number of documents in the collection. Let's say that there are 100,000 documents in the collection (J). We know the total number of documents predicted by the system to be responsive (G); that is the output of our first-pass review. If we know G and J, then we automatically know H, because H is simply J – G. We now have the column totals filled out.

From Prevalence, we have an estimate the proportion of documents in the top row, those that are truly responsive. That proportion times the total number of documents is the number of responsive documents estimated in the collection. Again, if we know the total number of documents and the total

number of responsive documents, then we know the total number of non-responsive documents. F = J – E. So now, we have the row totals filled out.

Elusion is used to calculate the number of documents in cells B and D. Cell B is Elusion times the number of documents in the predicted non-responsive column (H). Cell D is the difference between H and B. We now have the second column of the contingency table filled out. Cell A is calculated from the difference between the first row total (E) and cell B, which we just calculated. The final cell, (C) is then the difference between the first column total, G and A. We have now filled out the entire contingency table from knowledge of Elusion, Prevalence, the number of documents emitted by first-pass review, and the number of documents in the collection. We can compute Recall and Precision without a highly burdensome evaluation effort, by sampling a relative small number of documents.

Rather than exhaustively assessing a large random sample of thousands of documents, with the attendant variability of using multiple reviewers, we can obtain similar results by taking advantage of the fact that we have identified putatively responsive and putatively non-responsive documents. We use that information and the constraints inherent in the contingency table to evaluate the effectiveness of our process. Estimating Recall from Elusion can be called eRecall.

On Confidence

Two of the key ideas in sampling statistics are *confidence level* and *confidence interval*. These are not ideas that are difficult to understand, but they are also ideas that are easy to misunderstand. We would love to have confidence in our process. Unfortunately, the statistical notions of confidence level and confidence interval do not refer to our confidence in our process, but to the representativeness of a sample used to measures our process. Higher confidence does not mean that we have done a better job. We could have a very high confidence level and do very poorly at separating responsive from non-responsive documents.

Confidence level refers to the repeatability of our sample. For example, when estimating prevalence, we draw a random sample of documents and count which of these is responsive or not. Depending on our sample size, we can then say that with a certain level of confidence, the true prevalence in the population will be within a certain range (the confidence interval) of our estimated prevalence.

If we adopt a 95% confidence level, which is a widely used standard level, then it means that if we drew 100 samples, then about 95% of those samples would find that the true prevalence is within the specified range. The confidence interval is that range. Unless the sample involves a variable with extremely high risk (such as specifying the lifetime of a key component in a space vehicle), a 95% confidence level should ordinarily be enough. Remember, a higher confidence level does not mean better performance.

The confidence interval is the range around our estimate that is supposed to contain the true value. All other things being equal, the larger the sample size, the narrower this range. Roughly, a sample of 400 objects will give a confidence interval (at a 95% confidence level) of ± 5% no matter how large the population is. In election polling, for example, a poll of 400 random voters will give an estimate of ± 5%

whether the poll is for mayor of a city, governor of a state, or President of the United States. A sample size of 600 will yield about a 4% confidence interval and 1,000 will yield approximately 3%. Reducing the size of the confidence interval further requires considerably more effort, which may not be justifiable under the circumstances. We'll return to the question of appropriate effort shortly.

Although the confidence interval is often described as a margin of error, not all values within that range are equally likely. If our sample yields a prevalence of, say 9.7%, then our best estimate of the prevalence of responsive documents in the whole collection is 9.7%. The greater the difference between our best estimate and other values, the lower the probability that those other values will be correct, yielding the familiar bell-curve.

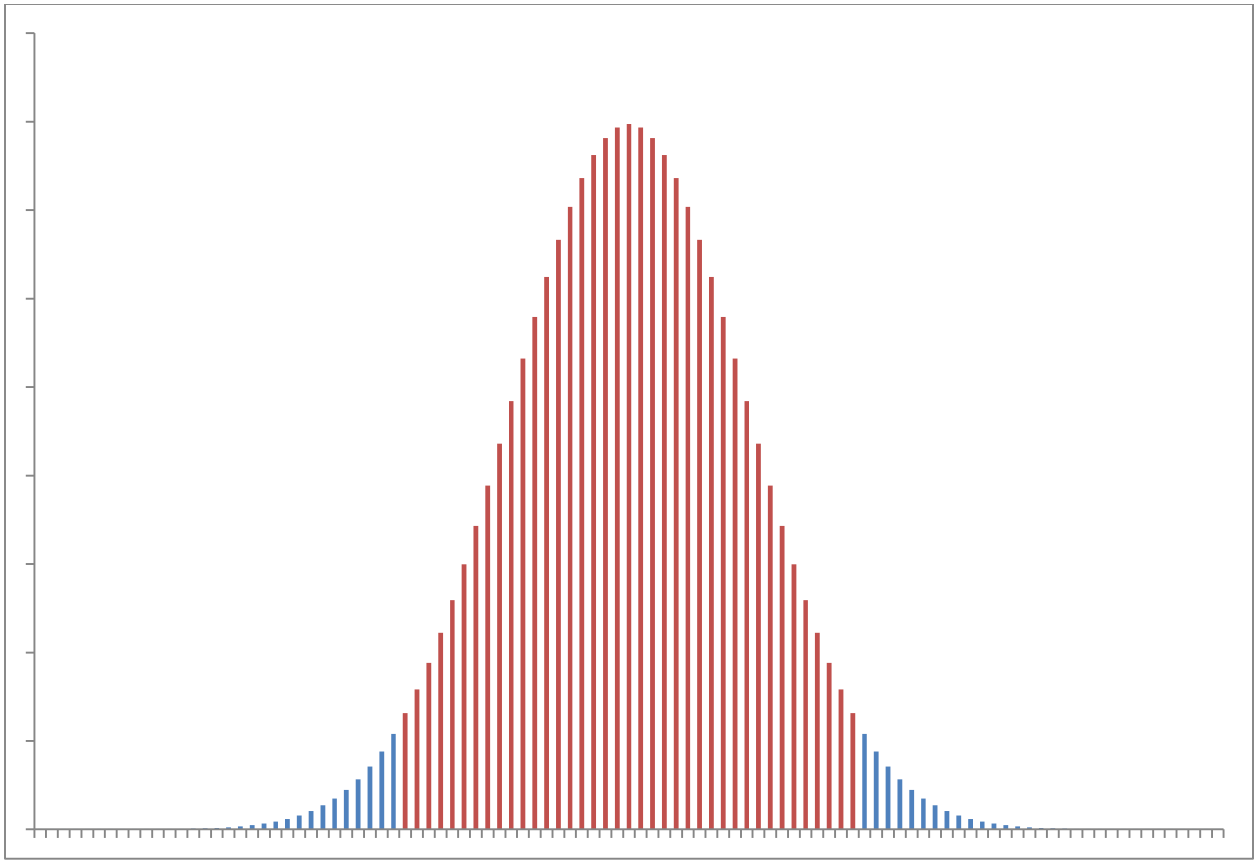

Figure 1. Confidence interval for discrete values on either side of an estimate. The red bars are within the 95% confidence interval.

In fact, the true value is many times more likely to be (shown by the height of the bar) in the center of the range than toward one of the endpoints or other.

If we take a sample of 400 documents with a prevalence of 50% (the worst case), then the 95% of the time, the sample will be within the range of 200 (the population value) ± 19 documents. About 80% of the time, it will be within a range of ± 12, and 60% of the time it will be in the rage of ± 8.

A confidence level of 95% is conventional, but not all values within the resulting confidence interval are equally likely. Even if a 95% confidence interval is overly broad, there is still information within that range that helps us to predict the population value.

Choosing a confidence interval

For a given confidence level, the size of confidence interval we select determines the sample size required. The larger the sample size, the narrower the confidence interval. There is a balance between the amount of effort needed to collect and assess a sample and size of the confidence intervals. A sample of 400 documents, for example, will yield a confidence interval of ± 5%, but cutting this range in half, requires quadrupling the number of documents to read.

Because the size of the confidence interval tells us so little about the quality of results from our review process, there may not be a lot of value derived from selecting narrower confidence intervals. One can have very narrow confidence intervals around very mediocre performance. Small improvements in the confidence interval require large increases in effort. In my opinion, the size of the confidence interval should be set to the minimum necessary to make consequential decisions.

One of the factors to consider in specifying the accuracy of our estimate is the consequences of that measurement. What are we going to do with that information? Is there some value that we expect a process to exceed in order to be acceptable? Just learning that some process was accurate to within a certain range tells you nothing about the acceptability of that level of accuracy, that is a separate discussion.

When we measure the accuracy of discovery, we measure it with some purpose in mind. For example, we are trying to make a reasonableness decision. Did the producing party meet its obligations, for example, under FRCP Rule 26(g)? But what level of accuracy do we require to establish "a reasonable inquiry?" Like other reasonableness judgments it will depend on the stakes in the case, the level of effort required, and the alternative methods available for finding responsive documents. Similarly, there is no universal standard for setting the confidence interval (or even whether you have to set one). In my opinion, we should set the confidence interval to allow reasonable decisions, recognizing that narrower confidence intervals require more effort, but may not provide much more useful information about the accuracy of the process.

We may be satisfied if recall (or whatever measure we are using) is more than some certain value, say 75%, and be dissatisfied if we fail to reach that criterion. There is no principled method for selecting that criterion, however. Our measurements are going to depend on exactly how our analysis was done and on the difficulty of distinguishing responsive from non-responsive documents, and to some degree on the prevalence of responsive documents. We cannot pick any specific number and claim that all successful eDiscovery projects must exceed that number. There cannot be any industry-wide black line for success. Our assessment of success must depend on reasonableness judgments that take into account the factors that affect our measurement and the values that could be expected from other approaches to document decisioning.

Even if we cannot establish a universal standard for accuracy in all matters, we may be able to come up with one for each particular collection of documents and the issues that surround them. Setting a target or criterion level of accuracy may be part of the meet and confer process, but it may require some further discussion after we know better the characteristics of the data (for example, with very low prevalence, or very subtle distinctions between responsive and non-responsive documents, it may be difficult to achieve high levels of accuracy). We may consider the accuracy that can be expected from using alternative methods (for example, Roitblat et al., 2010, report Recall around 50% for two human review teams compared with an original review). We may consider the amount at risk and the cost of extended effort. Whether implicit or explicit, we probably have some notion of what level of accuracy we will consider reasonable. Without such a criterion, there is not much point in measuring performance in eDiscovery and there is certainly no point in insisting on a narrow confidence interval when doing it.

One final point on confidence intervals. In statistics, the way we collect and analyze our data depends strongly on the precise question that we are asking. Ordinary confidence intervals are written as a two-sided range above and below our estimated value. But what if we only care about one side of that comparison. We may not care that our observed Recall, for example, could be 5% better than we observed in our sample, because that has no consequences. If we have a minimum score in mind, then it may not be consequential by how much we exceed that minimum. Rather than estimating our measure to plus or minus some range, we may be interested only in the minus or only the plus side. Does our observed Recall, for example, exceed our implicit criterion? When we are interested in only one side, we can "reallocate" the uncertainty from the side that we don't care about and focus our effort only on the side that we do care about.

One way to think about the confidence interval is in the context of hypothesis testing. For example, if we define reasonable performance in eDiscovery as Recall that exceeds some criterion (and we pick the criterion in some reasonable way), then that amounts to a statistical hypothesis test. We form two hypotheses. In the jargon of statistical testing, the null hypothesis supposes that any difference we might observe is due only to chance (within the confidence interval). The motivated hypothesis is that the difference is not likely to have emerged by chance (is beyond the confidence interval).

We need to determine whether the difference between two values is within the confidence interval of the difference or outside of it. If it is within the confidence interval, then it could have occurred by chance, because 95% (our confidence level) of the time the difference will be within the confidence interval. If the difference is greater than the confidence interval, then we say that that difference is unlikely (less than 5% likely) to have occurred by chance.

In a situation like discovery, we count the process as a success only if the difference between our observed Recall and our criterion is in one direction, not the other. If our observed recall is below the criterion, we would not call that a success. So, we are interested in the difference and the direction of the difference. Our null hypothesis is that the observed Recall is less than or equal to the criterion. Our motivated hypothesis is that the observed Recall is greater than the criterion. We are interested in only one side of the difference, so we can use a one sided test. Our 95% confidence interval includes all of

the potential Recall scores below the criterion and some amount above the criterion. Because we care about only one side of the comparison, the difference can be smaller and still be outside of our confidence interval relative to when we care about scores that are both greater or less than our criterion.

Conclusion

eDiscovery efforts can be evaluated with a reasonable amount of effort. Technological solutions are important because they make the process of eDiscovery less burdensome, allowing cases to be resolved based on their merits, rather than on the size of the parties' pockets. Assessment can play an important role in eDiscovery, but if the assessment is too burdensome, then much of the value of the technology is lost to assessment or to disputes about how the assessment is to be conducted. If the effort required to assess the effectiveness of predictive coding exceeds the effort needed to do the predictive coding, then one advantage of these technologies is lost.

The analysis above demonstrates that these assessments do not have to be overly burdensome. The completeness of eDiscovery document selection can be measured without having to review large collections of random documents in order to find a relatively small set of responsive ones. Moreover, the documents that are produced are, by definition, available to both parties. Each party is free to make whatever assessment of them is appropriate. What the other parties typically do not receive is the set of documents that are not designated responsive and are not produced. Elusion or differential prevalence evaluates precisely those documents and provides a powerful window through which to assess the efficacy of any eDiscovery process.

Both the producing and receiving parties gain from the use of these technologies because the producing side can greatly reduce the burden of eDiscovery while the receiving party gets a more complete and accurate set of documents in a shorter period of time. Justice is then served when both parties achieve their eDiscovery goals without undue burden or risk.

