PREDICTIVE CODING: TURNING KNOWLEDGE INTO POWER

A strategic approach to negotiating the scope of discovery.

Thomas I. Barnett, Managing Director, eDiscovery Practice Leader
Michael Sperling, Managing Director, Chief Data Scientist
STROZ FRIEDBERG

No self-respecting attorney would consider going into a hearing without knowledge of the basic facts of the case and the issues under consideration. Yet when it comes to making required disclosures about the nature and location of discoverable electronically stored information (ESI), negotiating the scope of discovery and developing a discovery plan under Rule 26 of the Federal Rules of Civil Procedure, many lawyers repeatedly find themselves flying blind. This lack of knowledge can be costly. Discovery can be one of the most expensive aspects of litigation. Document review is understood to account for as much as 80% of the cost of discovery. As a result, failing to understand how much discoverable data there is, what effort will be required to review it, and what the information is likely to reveal is a recipe for significant risk, increased cost, and in the worst case, outcome determinative failure.

Predictive coding offers an unprecedented opportunity to bridge this knowledge gap. It can deliver detailed, verifiable information about the nature of and the cost of reviewing discoverable information, and provide a meaningful strategic advantage to parties that choose to take advantage of the opportunity in negotiating the scope of discovery.

In a relatively short period of time, predictive coding has gone from an obscure technological curiosity in litigation circles to the most widely discussed and debated approach to managing the challenge of discovery of ESI. But while predictive coding is new to the legal profession, the underlying technology (*supervised machine learning*) is anything but new in the business, scientific and academic communities. Based on technological processes going back decades and mathematical principles going back centuries, predictive coding offers a highly quantitative and verifiable approach to analyzing and classifying textual information. And when executed properly based on review of sample documents by attorneys knowledgeable about the underlying matter, predictive coding has been demonstrated to be superior to unassisted manual document review.

One of the factors associated with this rise in notoriety and attention, presumably partially causative, has been the handful of legal opinions and rulings that have involved the use of the predictive coding process. Unpredictably, the manner in which the use of predictive coding was put forward in these cases has varied widely: from both parties agreeing to use predictive coding (*Da Silva* Moore), to plaintiffs seeking to require defendants to use it (*Kleen Products*), to the court suggesting that the parties use predictive coding (*EORHB V. HOA Holdings*), to defendants seeking to use it over plaintiffs' objections *(Global Aerospace*).

In the history of using technology to search, organize and categorize ESI in litigation discovery no other technology based approach has been thrust into the center of such heated debate and judicial attention. Why is that? What is it about predictive coding that arouses such strong feelings and passionate debate?

There is probably no definitive answer to that question, but certain unique characteristics, or at least characterizations, may provide a clue. The term *predictive coding* itself suggests that the process replaces a task traditionally and familiarly performed exclusively by human beings—deciding how to classify (*code*) documents potentially subject to discovery. This is viewed as a quintessentially, and until now, exclusively, human task: applying judgment, experience, understanding of context and specific knowledge about a case to make what amounts to a legal judgment about potential evidence in a case.

Other technologies used in discovery of ESI fall short of that. No such passionate or heated debate occurs about whether to use processes such as keyword searching, latent semantic analysis (i.e., concept organization, clustering and searching), data extraction, email thread and social network analysis, near duplicate identification, de-duplication and date and file type filtering (e.g., exclusion of system files). These technologies are viewed as efficient tools for doing brute force tasks that humans don't want to or can't perform efficiently.

Predictive coding, by contrast, can evoke fear and apprehension. It is often viewed skeptically as a potential replacement for legal judgment and decision making, uniquely cognitive, human tasks—and perhaps a threat to lawyer employment. But given the potential risk and the corresponding strategic opportunity, lawyers shouldn't fear predictive coding, they should fear failing to take advantage of the available information and fear that their adversary is. In addition to a strategic edge, predictive coding offers a uniquely powerful means of achieving transparency and cooperation—frequently discussed, but rarely achieved, aspirational values in negotiating the scope of discovery.

The underlying issue in most negotiations about the efficacy of technological processes used to identify potentially relevant material boils down to two factors: *accuracy* and *completeness*. In the language of probability and statistics: *precision* and *recall*. When implemented correctly, predictive coding, unlike other technology based approaches to document analysis, allows for a high degree of transparency of these measures, and, uniquely, offers the ability to adjust and calibrate them. Accordingly, it affords the promise of facilitating more substantive and factually grounded discussion and negotiation about the scope of discovery than does, for example, debating key words.

Following is a hypothetical case study demonstrating how predictive coding can estimate precision and recall over a document population and thus be used as a basis for substantive, content-based rather than simply data volume-based negotiations about the scope and cost of discovery:

1) Scenario

a. The parties to a lawsuit are attempting to negotiate the scope of discovery of ESI. Realizing that it is impossible to identify and produce the exact set of responsive documents, they agree that the production will include at least X% of the total responsive document population (*recall* X) and that at least Y% of the documents in the population will be responsive (*precision* Y).

2) Process

a. A model is trained and run on a holdout sample to create an estimated recall/precision curve (figure 1). The model assigns a probability score of responsiveness to each document. The graph shows the estimated precision for each standard recall point (10%, 20%, 30%, …) and the associated probability threshold. For example, assume that all documents with a probability score of 34% or higher are responsive. In that case, recall would be 80% and precision would be approximately 73%.

b. In order to improve the recall/precision curve, active learning techniques are applied. In this case, 3 rounds of active learning are applied (figure 2). The first 2 rounds of active learning achieves sizable gains in performance, while the third round achieves a marginal gain. The precision for 80% recall improves to 85%, a full 12% higher.

c. The final recall/precision curve and document counts at each recall point (2nd line under X axis with "Docs:" labels) for an unclassified population of 20,000 documents are shown in figure 3. The parties agree to produce a responsive population with 90% recall and 85% precision. By producing all 5400 documents with a probability above 54%, a precision of 85% is achieved, but recall is only 80%. However, the graph shows that 90% recall can be achieved by producing all documents with probability above 9% and that there are 1800 documents (7200-5400) with responsive probability between 54% and 9%. Therefore, by manually coding those 1800 documents and producing the responsive population in these 1800 documents together with all documents with probability greater than 54%, the production will meet the 90% recall and 85% precision targets.

3) Accuracy Estimation

a. Producing a precision/recall curve that estimates performance for the entire unclassified corpus is not trivial. The estimate must *underestimate* true recall and precision to ensure delivering a production that fulfills the recall/precision requirement. In figure 4, the blue line is the recall/precision curve computed from the test set. The computed precision is always greater than or equal to the estimated precision. Also, the last line under the X axis indicates the recall computed for that point from the test set. For example, the estimated 80% recall point has a recall of 84% computed from the test set. Again, the computed recall is always greater than or equal to the estimated recall.

Conclusion

Predictive coding has been increasingly considered as a viable means of reducing the time and cost (i.e., burden) of litigation document review while simultaneously improving the accuracy of purely manual review. A significant but largely untapped opportunity exists to use predictive coding in the negotiation phase for determining the scope of discovery. Predictive coding's ability to estimate the amount of supplemental manual review required to meet recall/precision goals offers the potential for far more substantive and quantitatively grounded negotiations and cooperation between parties, better informed strategic decision making and unprecedented accuracy in estimating discovery cost and burden.

FIGURES

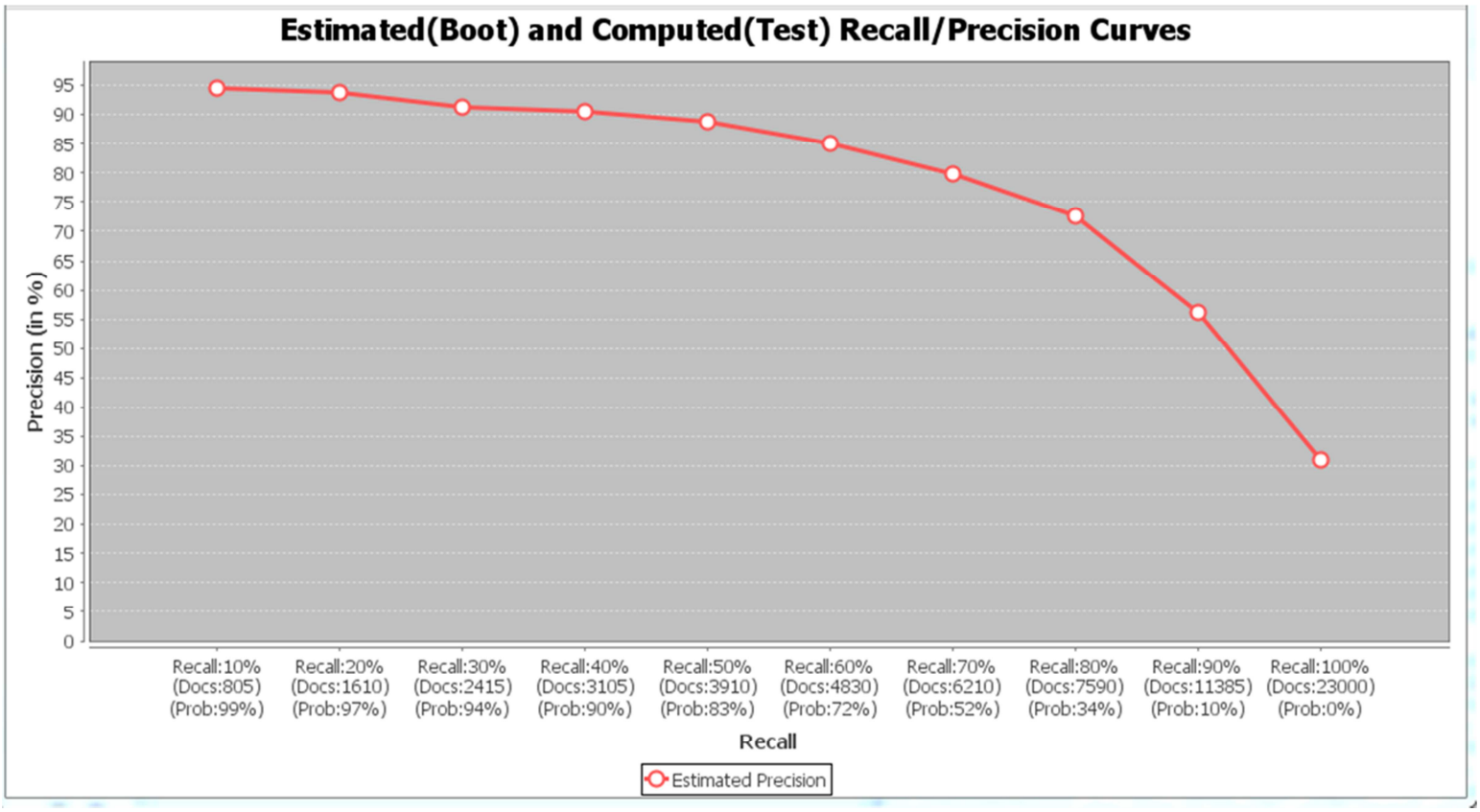

Figure 1. Initial estimated precision/recall curve

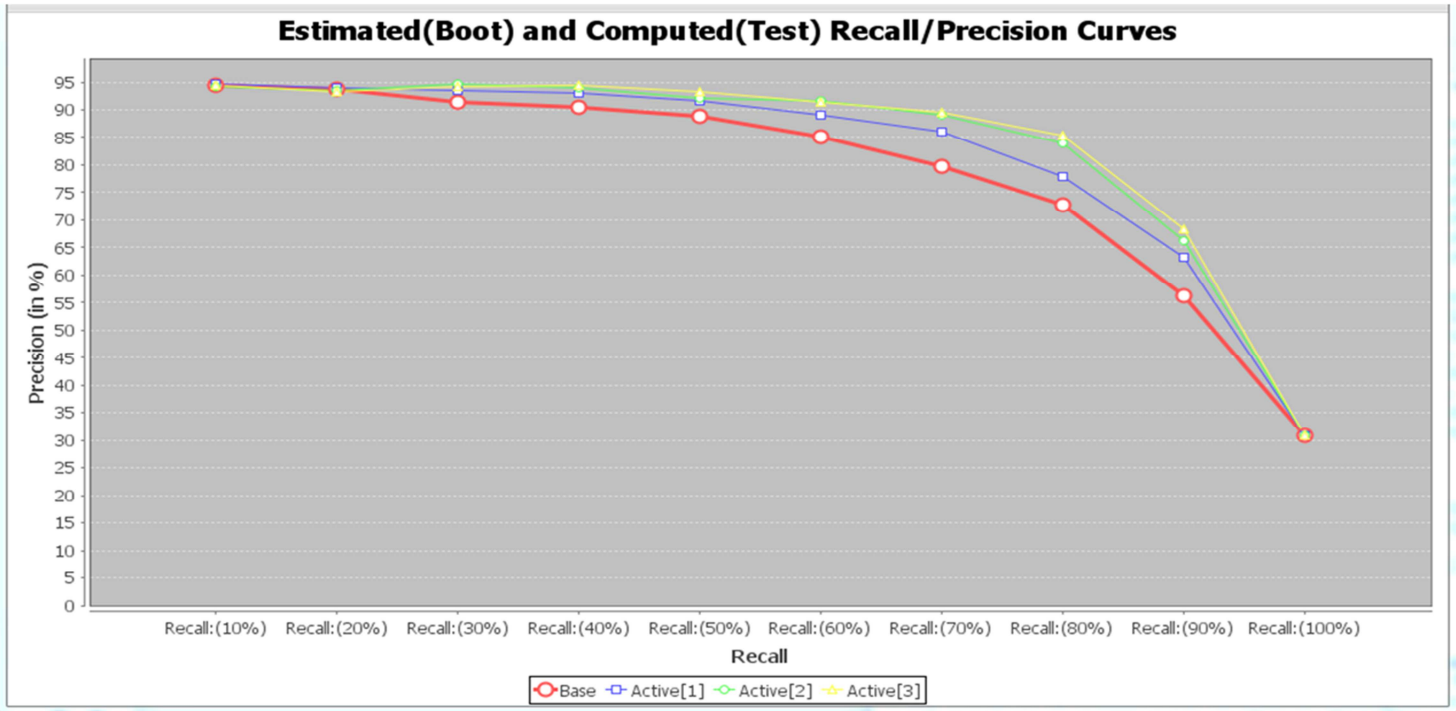

Figure 2. Active learning iterations leading to final curve.

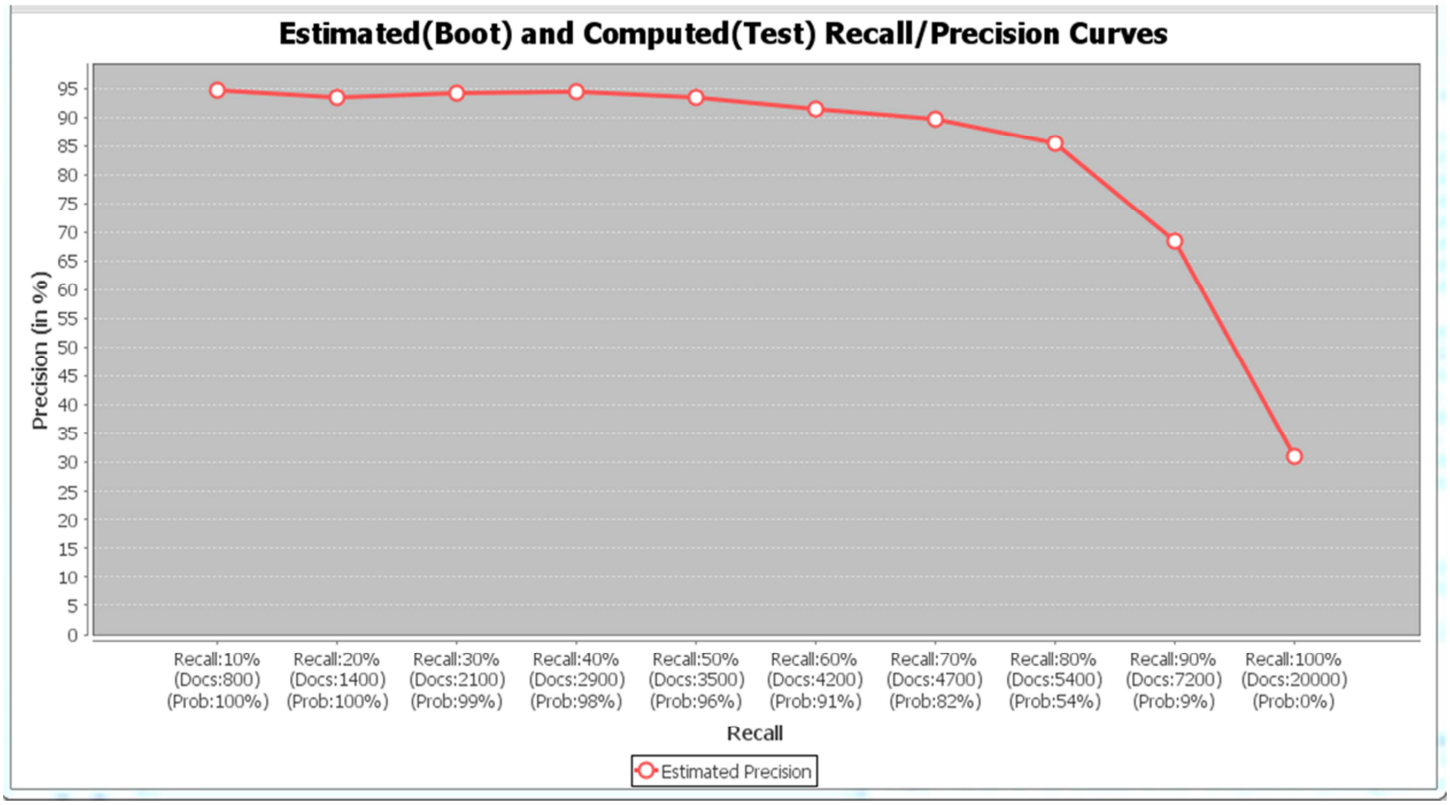

Figure 3. Final estimated precision/recall curve

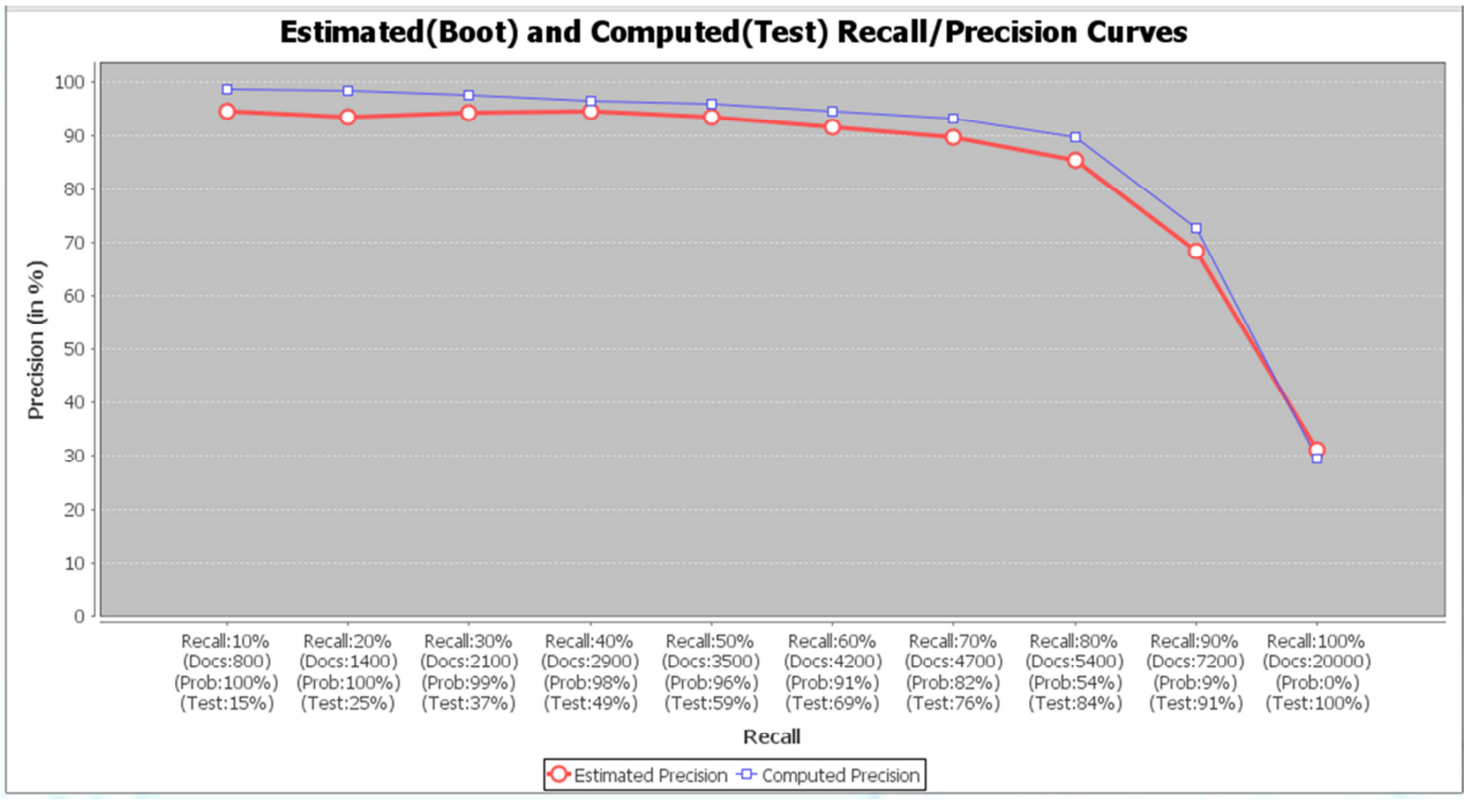

Figure 4 Final estimated precision/recall curve with computed precision overlay