The Metropolitan
Corporate Counsel®
National Edition
www.metrocorpcounsel.com
Volume 22, No. 6 June 2014

Predictive Coding In Practice: Success Stories From Three Years Of Predictive Coding

Barclay T. Blair

ViaLumina LLC

This article was published as part of Equivio's "Predictive Coding Minus the Hype" educational series.

Introduction

The e-discovery community is buzzing about predictive coding, and with good reason. The volume of electronically stored information (ESI) is expanding at a breakneck pace. Every two years the amount of digital data is expected to almost double.[1] Predictably, e-discovery costs are rising.

Barclay T. Blair

Predictive coding is one method of gaining control over the expanding universe of data. By reducing the volume of data requiring attorney review, predictive coding offers the promise of significant cost savings where expenditures are at their highest.

Although the number of documents contained in a gigabyte of ESI varies significantly by file type, even at 5,000 documents per gigabyte, review costs add up fast. At $2.50 per document for the complete review process (initial review, quality control, pre-production, and so on), it would cost about $12,500 just to review one gigabyte. A terabyte would cost $12.8 million. One study of e-discovery costs found that in a majority of the cases it examined, "review consumed at least 70 percent of the total costs of document production . . ."[2] Review costs are significant, but the inability to quickly identify key documents may also lead to *strategic* disadvantages.

Predictive coding is one method of gaining control over the expanding universe of data. By reducing the volume of data requiring attorney review, predictive coding offers the promise of significant cost savings where expenditures are at their highest (i.e., review). By getting to key information faster, predictive coding also offers the promise of helping attorneys litigate and advise their clients *better*. Clearly, practitioners need to include predictive coding and other forms of technology assisted review (TAR) in their toolkit.

E-discovery practitioners own the challenge of translating the theoretical promise of technological advances into real-world improvements *in practice*. Understanding where and how it can most effectively be applied is essential. Real-world examples of successful application of predictive coding can be invaluable in translating theory into practice.

D4 is a national provider of e-discovery, digital investigations, information management and security solutions to law firms and corporations. D4 has been implementing predictive coding workflows since 2009. Equivio's Zoom® predictive coding suite is at the core of D4's predictive coding workflow. Each of the cases described in this paper were completed using Zoom. Over the past three years, D4 has rigorously tracked data from its predictive coding cases and has constantly refined its predictive coding workflow. The goal of this paper is to share insights from D4's experiences, including a close look at four specific cases.

These cases include:

Case 1: "The Document Dump" (A large-volume, incoming production)

Case 2: "A Merger at Risk – When Speed Matters" (A large-volume, limited-time, second request)

Case 3: "High Cost/Low Merits" (Potentially high discovery costs and low merits)

Case 4: "The Pressure Cooker" (A discovery scope expansion with no change in production deadline)

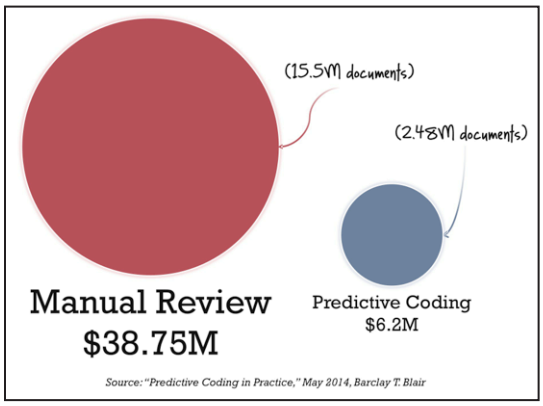
These case studies not only tell the story of cost savings, but also the strategic value of getting to key data faster.

Comparing Costs Over Three Years of Predictive Coding

Over the past three years, D4 has conducted dozens of predictive coding projects for a variety of clients, with more than 15.5 million documents in play. Using traditional linear review methods, the cost to review each of these documents would have been nearly $40 million (applying a fully loaded cost of $2.50 per document).

***Barclay T. Blair** is an advisor to Fortune 500 companies, software and hardware vendors, and government institutions, and is an author, speaker, and internationally recognized authority on information governance. He has led several high-profile consulting engagements at the world's leading institutions to help them globally transform the way they manage information. He is the president and founder of ViaLumina.*

On the other hand, (see figure below) when using predictive coding, D4's average case had 10 percent richness, 80 percent recall, and 50 percent precision. At these rates, predictive coding reduces a corpus of 15.5 million documents to 2.48 million, which means that only 16 percent of the total corpus would need to be reviewed by attorneys prior to production. Applying the same per document review cost ($2.50), the total review cost when using predictive coding would be $6.2 million. This translates to a potential savings of $32.55 million (84 percent cost savings). To be clear, predictive coding at D4's average case parameters would enable an organization to retrieve 80 percent of the relevant documents for 16 percent of the cost of fully manual review.



By comparison, D4 has learned that traditional techniques like keyword searches are much less effective, with recall rates as low as 20 to 30 percent. D4's experience is consistent with widely cited studies like Blair and Maron, which found, for example, that similar methods "retrieve[d] only 20 percent of the relevant documents, whereas the lawyers using the system believed they were retrieving a much higher percentage (i.e., over 75 percent)."[3]

Case 1: "The Document Dump"

1. The Case

In this case, opposing counsel produced more than 800,000 documents (267GB) to D4's client. (see figure at right) Their client suspected that this was a "document dump," i.e., a production purposely designed to make it harder to find relevant information. The collection produced was based on keyword searches that both parties had agreed upon, but the client suspected the opposing party had not reviewed the production for relevancy. Receiving counsel needed to find out which documents were most important to support their side of the dispute, but they were faced with a gargantuan task.

2. The Strategy

Rather than review the entire production, D4 and their client decided to use predictive coding to focus and narrow their review. This decision was rewarded. Predictive coding revealed that the team could defensibly focus their review on just 30 percent of the documents. Here's why:

More than half of the most relevant information was found in just 30 percent of the corpus (240,000 documents).

The "bottom" 30 percent of the corpus contained information with such low relevance scores that it could be safely ignored.

The team decided that the remaining 40 percent would not be reviewed unless they failed to find what they were looking for in the targeted 240,000 documents.

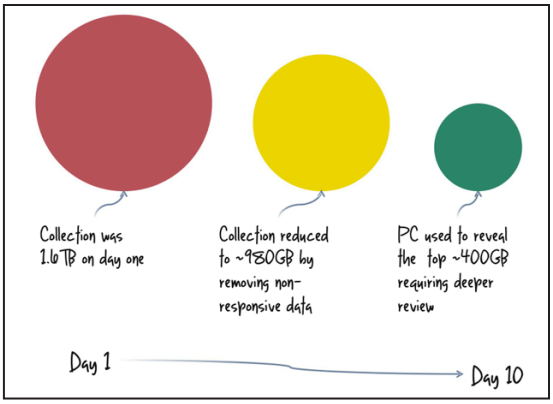
3. The Results

In the end, receiving counsel was able to fully avoid reviewing approximately 70 percent (560,000 documents) of what opposing counsel had submitted – with very little risk. Keep in mind that this was not a review conducted *for* production, but rather a review of documents produced by the other side. Thus, there was no risk on the client's side for failing to produce a relevant document. Sufficient documents to support the case were found in the top 30 percent. As such, D4's client saved approximately $1.4 million in review costs, while focusing their attention on the most important documents earlier in the process. The eventual outcome was that D4's client won the case. This customer now uses predictive coding as a standard practice for incoming productions.

Case 2: "A Merger at Risk – When Speed Matters"

1. The Case

A Fortune 500 multi-national company was in the process of a merger when the U.S. Department of Justice (DOJ) notified them of a second request under the Hart-Scott-Rodino Antitrust Improvements Act. In scope were 5 million documents (1.6 terabytes), some of which were in foreign languages. (See Figure below) The company had only 10 days until production, or the merger might not be approved.

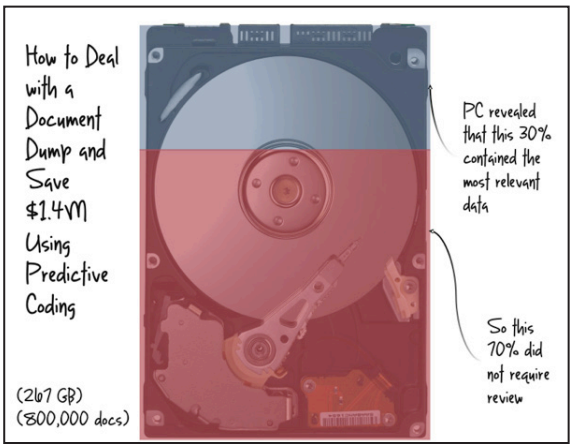


2. The Strategy

The company retained an AMLAW100 firm, who in turn engaged D4 to assist with the e-discovery process. D4 worked with the firm's staff to identify large chunks of data that were clearly not responsive and to thus remove 40 percent of the data from the review set (avoiding $5 million in potential review costs). Predictive coding was then used to identify the documents that satisfied an agreed upon recall threshold, comprising the top 25 percent of the collection. D4 was then able to further isolate the potentially responsive data for deeper review. D4 deployed the necessary operational and project management resources that were required to support such a large data set with a 10-day turnaround.

3. The Results

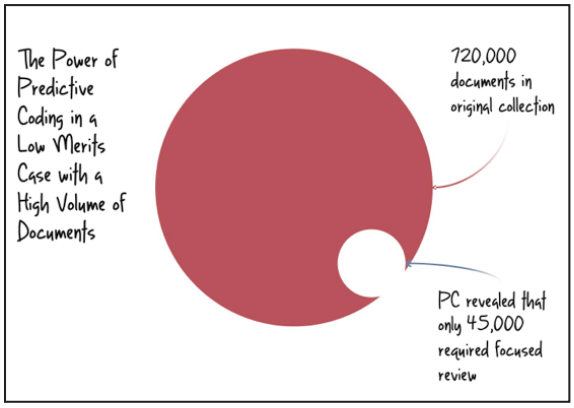
Using this workflow enabled the law firm to meet the DOJ production specifications, complete the production in 10 days, and save the client significant time and money given the significant reduction in review volumes. The DOJ made no material objections to the ultimate document production, and the transaction successfully closed as planned. Without the use of predictive coding, the merger may have been at risk, or at the very least, the transaction costs would have been significantly higher.



Case 3: "High Cost/Low Merits"

1. The Case

A major healthcare company with operations in more than 100 countries was involved in a contract dispute. (See Figure on following page) Their gen-



eral counsel of litigation characterized the dispute as "high cost/low merits" because discovery costs were liable to become disproportionately high relative to the legal risks. The initial collection contained 238GB of ESI (720,000 documents) from various sources including file servers, email servers, client machines, and archives. An efficient method of culling, analytics, and reviewing was needed to bring discovery costs more in line with the stakes.

2. The Strategy

D4 first culled 540,000 non-responsive documents from the corpus using preliminary processes and tools, thus avoiding more than $1.3 million in potential review costs upfront.

Predictive coding was then used to evaluate and rank the remaining 180,000 documents. This revealed that 45,000 of those documents (the top-scoring 25 percent of the collection) contained the vast majority of the relevant material (80 percent). As such, D4's client was able to focus their attention and review on these documents, representing only 6 percent of the original collection of 720,000 documents. In other words, the client was able to review only 6 percent of the overall original body of documents and still capture 80 percent of the most relevant documents.

D4 established defensibility by statistically verifying that the remaining 75 percent of the post-culling collection (documents with low relevance scores, below the cut-off) contained no more than 20 percent of the relevant documents.

3. The Results

By strategically employing various data reduction and analytic techniques, including predictive coding, a company brought its discovery costs closer in line to the actual value of its case.

Case 4: The Pressure Cooker

1. The Case

D4 was hired by a large consumer manufacturing company that was defending itself from a competitor's false advertising and intellectual property claims. ESI for five primary custodians was collected and loaded into the review platform. The collection was approximately 200,000 documents. A tight production schedule was established by both parties and the magistrate judge. Review was in progress when opposing counsel successfully argued a scope expansion with the magistrate, resulting in a 10-fold increase in the number of documents to be reviewed and produced.

D4's client felt that it would provide a strategic advantage to their defense if they could stick to the agreed upon schedule, despite the massive scope expansion. The thinking was that opposing counsel was playing "dirty pool" and wouldn't be able to comply with the production deadline. As such, D4 found itself dealing with 10 times the ESI – now approximately 2,000,000 documents – and neither the timeline nor the budget was expanded.

2. The Strategy

At D4's recommendation, a parallel approach was employed. This approach leveraged the work already done in reviewing the ESI of the five primary custodians, as well as engaging a subject matter expert (SME) to train the predictive coding engine for the new data. Documents with higher scores were pushed to the front of the line for human review, and nothing would be lost from the previous work.

After the SME spent a few days training the system, the top-scoring 27 percent of the collection (containing 80 percent of the relevant documents) was targeted for review. A quick check was performed to confirm the system's coding designation for the initial five custodians. The team found that over 90 percent of the initial human calls were consistent with the predictive coding system. This provided an additional degree of confidence in the process. The team then reviewed the top 27 percent as well as confirmed that the bottom 73 percent contained no more than 20 percent of the relevant material. It actually tested to be less than 10 percent, meaning the top 27 percent actually contained more than 90 percent of the relevant documents.

3. The Results

By leveraging the relevance scores from the predictive coding engine, the review team was able to meet the original timeline and budget, saving the customer significant costs. As it turned out, opposing counsel was obliged to ask for an extension, which placed the defendant (D4's customer) in an advantageous position with the magistrate judge.

Final Thoughts

There is no doubt that predictive coding can provide significant benefits when used in an appropriate context. Further, as D4's extensive experience over a three-year period illustrates, predictive coding can confer both strategic advantage and direct cost savings by reducing overall ESI volume.

Because the case studies provide us a clearer understanding of the scenarios in which predictive coding was used, the strategic advantage it can confer is more obvious in these examples than might be shown by looking at aggregate or cost savings data alone. For example, in Case 1, the potentially devastating consequences of a document dump were avoided. In Case 2, a merger that was jeopardized went forward. In Case 3, litigation costs were brought in line with the true value of the case, allowing a clearer strategic assessment of best next steps rather than letting costs drive decisions. In Case 4, the potential catastrophe of an expanding discovery scope was averted and meeting a tough deadline (when opposing counsel did not) created a favorable impression with the court.

Predictive coding, especially when deployed by experienced practitioners, consistently yields significant cost savings. Given the sheer magnitude of ESI many will need to process, those savings can translate into millions of dollars. However, even these savings, significant as they are, are overshadowed by the strategic advantages that predictive coding yields for litigators.

Note that figures provided in this paper for the sizes of document collections are estimates that have been rounded up or down for simplicity and readability.