The Cutting Edge

Predictive Coding

College Sexual Assault Disciplinary Proceedings

Burying the Ten Commandments of Cross-Examination
Litigators bemoan the fact that the volume of documents and data produced and requested during discovery continues to increase year on year—as does the volume of electronically stored information (ESI) that clients must preserve. There is a glut of ESI arising as a consequence of the exponential growth in the creation and exchange of electronic files and email traffic. The result is that even in cases with a relatively low monetary value, there are often still hundreds of thousands or even millions of files relevant to the case, and a significant portion of the cost of litigation is driven by the obligation to preserve, search, review, and produce this ESI. However, the same computing power that has provided this overabundance of data may also provide the solution for reviewing, coding, and producing it in litigation with minimal human review.

Computer-assisted review, technology-assisted review, and “predictive coding” all refer to the use of computer-generated algorithms—based on human modeling—to code documents without the need for document-by-document human review. Predictive coding “predicts the relevance of discovery documents [or relative match of any document to the model set] based on the prior coding of a small sample of discovery documents by an attorney.” Nicholas Barry, Man Versus Machine Review: The Showdown Between Hordes of Discovery Lawyers and a Computer-Utilizing Predictive-Coding Technology, 15 VAND. J. ENT. & TECH. L. 343, 344 (2013). In other words, predictive coding is an iterative process because it depends on the coding first applied by attorneys familiar with the case and responds as the attorneys continue to interact with the documents. The computer becomes an extension of the attorney reviewer—coding documents in the review universe based on the algorithms derived from a model created by the attorney.

Starting Computer-Assisted Review
It begins with the collection and culling of client files and data, which can quickly become a daunting task without the support of a technical specialist. It was not long ago that only the most significant cases would encompass large amounts of hard-copy documents and electronic files. In fact, for many years, there was an inherent proportionality between the complexity of the case and the volume of documents expected to be at issue. Today, however, because each of us and the many devices we use are creating data at a near constant rate, even small-dollar cases can have considerable ESI subject to a duty to preserve and discovery. Accordingly, it has become commonplace for relevant custodians to possess hundreds of gigabytes, or even terabytes, of data sourced solely from emails requiring preservation, search, and review.

AARON GOODMAN
The author is with DLA Piper LLP (US), Phoenix.
The growing flurry of spoliation claims and requests for sanctions seen over the past several years—many arising out of allegations that one side or the other failed to properly preserve email—has resulted in companies preserving data out of an abundance of caution even though everyone knows that most of what is being collected and preserved will be irrelevant. This is how we arrive at big data: Enterprises with hundreds or thousands of employees each sending and receiving as many as 100 or more emails per day, day after day, month after month, for years until litigation is commenced, after which all of those emails are frozen for use in the litigation.

After this large amount of data is collected, the data will be sorted through the use of discovery tools. These tools will remove unnecessary junk files (a process called de-NISTing), will search for logical files, and will de-duplicate files so that only one copy of each document remains. Once processed, the data will be loaded into an analytical database platform and search terms can be run. It is at this point in the process that predictive coding will take over. Generally, if a normal human-based review is to occur, the data hitting on search terms will be identified and the documents will be loaded into a document review platform for human review. With predictive coding, rather than commencing a wholesale human review of all documents, software will be run over the data to categorize the documents and to create a review sample. This review sample will then be manually reviewed by an attorney, who codes the sample of documents for responsiveness, privilege, and any other desired categories.

Coding a review sample for the purpose of generating a model set differs from a typical document review, however, because documents are evaluated on an individual basis. In a typical document review, attorneys group documents together in families (i.e., a “parent” email is reviewed for responsiveness together with its attachments, or “children”) and assess them on a family basis. With predictive coding, because the computer will use the model set to identify and code specific related documents on an individual basis, attorneys coding the review sample must also look to the content of each individual document and disregard family relationships in order to generate a useful model set. For
example, coding a blank or generally irrelevant cover email as responsive (and thus pulling it into the model set with its relevant attachments) could skew the model because it would include those irrelevant pages in the creation of the search algorithms. It is only after the predictive coding model has been completed and all relevant documents identified that these documents can then be reunited with their families.

After the model set has been created, the predictive coding software will generate algorithms based on that model that can be applied to the remaining documents in the review universe. However, the entire universe of documents is unlikely to be coded in the first application of the model set to the database. The attorney who coded the model set will need to repeat the coding process after each application—adding additional documents to the model that will be used by the computer to revise and expand its algorithms. This cycle can go on a number of times before the majority of the review universe is coded by the computer. For example, if the review universe comprises 1,000,000 documents, an attorney may undertake four to six (or more) rounds of model coding and a manual review of 30,000–60,000 documents, the application of which may result in computer-generated predictive coding of 925,000–975,000 of the remaining documents in the review universe. The documents that remain uncoded after each coding round are referred to as “uncategorized”—either because they could not be coded based on the model set or because they are unique within the document universe. The process can be continued until the model effectively codes all but the unique documents. (The review team may also determine at some point to review the remaining uncategorized documents, if the number is small enough, through manual review.)

Model sets can also be set up to code for issues other than just responsiveness and privilege. For example, predictive coding can be used to search for documents according to importance or usefulness. That means if the model set comprises a group of particularly interesting or hot documents (also referred to as “Rank A” documents), you can use predictive coding to search the document universe for other Rank A documents. The same is true for documents on specific topics. It can even be applied to search for work-product documents that may not be picked up in a privilege screen. The key is the model set—if it is narrow, the search will be more targeted; if it is broader, there is likely to be a larger, more general result.

It is important to understand that the process is dynamic—meaning that because the model is changing as additional documents are added, earlier computer-generated coding may also change. The attorney reviewer may also change how she is coding certain documents, whether intentionally or unintentionally, which can change the model and create conflicts that will need to be resolved. Such changes and conflicts are almost always inevitable (depending on the complexity of the issues) because as the attorney reviewer becomes educated on the document universe, her understanding will naturally broaden and her knowledge of the documents will likely change how she is coding the documents. This is also why the process is dynamic and the coding on the entire model set will adjust over the course of the review cycles and, the hope is, ultimately result in a good model.

Additional reviewers can be included in coding the model set (if more speed is required), but that will introduce additional conflicts in the coding—because each reviewer will necessarily code some documents differently. The more reviewers, the more conflicts that will be created. Each of the conflicts will need to be resolved (taking more time) before the model can be applied to the review universe. Whatever the approach taken, it is important to remember that predictive coding is not a magic wand, and it will only be as good as the human reviewer creating the model. And, frankly, it can be a grueling process with little reward until it is completed.

### The Important Questions

From a litigation standpoint, the two important questions about predictive coding are (1) how do the courts view predictive coding, and (2) can it be used to satisfy party discovery obligations? One of the first published cases addressing the use of predictive coding in litigation was written by federal District Court Judge Andrew Peck in 2012, not long after he had written an article on the very same topic. Judge Peck’s expository opinion in Moore v. Publicis Groupe & MSL Group, 287 F.R.D. 182, 183 (S.D.N.Y. 2012), quickly became a seminal case on predictive coding. In Moore, Judge Peck provides a primer on predictive coding and announces that it “is an acceptable way to search for relevant ESI in appropriate cases.” Id. at 183–84 (quoting Andrew Peck, Search, Forward, L. TECH. NEWS, Oct. 2011, at 25, 29). Writing with the understanding that his opinion in Moore could create precedent, Judge Peck provided the following advice:

What the Bar should take away from this Opinion is that computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review. Counsel no longer have to worry about being the “first” or “guinea pig” for judicial acceptance of computer-assisted review. As with keywords or any other technological solution to e-discovery, counsel must design an appropriate process, including use of available technology, with appropriate quality control testing, to review and produce relevant ESI while adhering to Rule 1 and Rule 26(b)(2)(C) proportionality. Computer-assisted review now can be considered judicially-approved for use in appropriate cases.

*Id.* at 193.
Since Moore, case law has “developed to the point that it is now black letter law that where the producing party wants to utilize predictive coding for document review, courts will permit it.” Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 127 (S.D.N.Y. 2015). A tax court even stated that “the technology industry now considers predictive coding to be widely accepted for limiting e-discovery to relevant documents and effecting discovery of ESI without an undue burden.” Dynamo Holdings Ltd. P’ship v. C.I.R., 143 T.C. 183, 191–92 (2014).

Despite these strong judicial proclamations, predictive coding has yet to achieve general acceptance or come into popular use. It appears from the prevailing cases that courts will not, sua sponte, require parties to use predictive coding or other analytics. There are no reported cases of parties seeking to compel, over a court’s denial, an obligation to use predictive coding. Nonetheless, courts will permit predictive coding to be used, if there is an agreement among the parties, and may even allow it over the objections of other parties—if the proposal is reasonable under the circumstances of the case. The latter is particularly true under the 2015 amendment to Rule 26 of the Federal Rules of Civil Procedure.

The 2015 amendment to Rule 26—driven in part by the recognition that the cost of discovery was becoming prohibitive in that it often would exceed the damages at stake in litigation—shifted the standard for discoverability from “reasonably likely to lead to the discovery of admissible evidence” to “relevant . . . and proportional to the needs of the case.” The Rule 26 amendment reads a reasonableness element into whether or not documents need to be produced, regardless of whether or not they could be relevant. That should mean that the courts will sit in judgment on whether or not the scope of discovery sought is reasonably proportional to the case. It may also mean that parties will be forced to agree that predictive coding may be used to identify responsive documents for production. Notably, the 2015 Advisory Committee on the Rules of Civil Procedure specifically contemplated that predictive coding could be used to reduce the burden and expense of voluminous discovery—striving for a proportional balance between cost and disclosure.

The burden or expense of proposed discovery should be determined in a realistic way. This includes the burden or expense of producing electronically stored information. Computer-based methods of searching such information continue to develop, particularly for cases involving large volumes of electronically stored information. Courts and parties should be willing to consider the opportunities for reducing the burden or expense of discovery as reliable means of searching electronically stored information become available.

If predictive coding appears to be the best way to approach discovery, parties should agree at the inception of the case on a protocol to be followed. That protocol should then be submitted as a stipulated order for review and entry by the court. Failing to secure such agreements can and will lead to disputes. For example, once documents have been produced, disputes will inevitably arise over whether there are discovery deficiencies. As a result, the substantial time and expense incurred to identify and produce the documents through predictive coding could be wasted without such a stipulation.

All of this invites a showdown between predictive coding and human review and raises a critical question: Which is better? Currently, courts, clients, and attorneys each tend to favor a document-by-document human review over any computer-assisted technology or analytics. It makes sense—people trust what they can see with their own eyes, and the use of computer-based review technology is seen as the exception (or an addition) to the presumption that an attorney will review each and every page of every document. But this bias toward human review rests on the false premise that “manual review by humans of large amounts of information is as accurate and complete as possible—perhaps even perfect—and constitutes the gold standard by which all searches should be measured.” The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, 8 SEDONA CONF. J. 189, 198–99 (Aug. 2007). If ever that was the case, it is no longer.

Eyes-Only No Longer Workable

The Sedona Conference working group, a nonprofit institute that focuses on the advanced study of the law in complex litigation, among other areas, explains that “with increasingly complex computer networks, and the exponential increase in the volume of information existing in the digital realm, the venerated process of ‘eyes only’ review has become neither workable nor economically feasible.” Id. The Sedona Conference points to one study that found that human reviewers are only 20–25 percent accurate and complete in conducting such document searches and reviews. See David C. Blair & M.E. Maron, An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System, 28 COMMUNICATIONS OF THE ACM 289 (Mar. 1985). The unfortunate truth is that humans are fallible, even the very best of us, and it is difficult to consistently code thousands upon thousands (and hundreds of thousands) of documents using increasingly complex and laborious document-coding protocols. In addition, despite its complexity, this task is often relegated to more junior attorneys, contract lawyers, and paralegals. The potential for errors and internal conflicts is multiplied even further when the number of reviewers is increased to groups of reviewers, larger teams of reviewers, or even small armies of reviewers.
The preference may always be that, rather than using predictive coding or a large number of junior reviewers, one or a few key attorneys should review all of the documents because those attorneys have been immersed in the case and will pick up on nuances that no algorithm could track and that junior reviewers less familiar with the case will be unable to spot. Even if this is the preference, the growth of ESI means that such a review is often not feasible. For that reason, clients and their counsel should consider—collectively looking at factors such as accuracy, efficiency, and cost—whether predictive coding is a workable alternative to a document-by-document review. In fact, multiple courts have found that “[p]redictive coding has emerged as a far more accurate means of producing responsive ESI in discovery . . . than human review or keyword searches which have their own limitations.” Progressive Cas. Ins. Co. v. Delaney, No. 2:11-CV-00678-LRH, 2014 U.S. Dist. LEXIS 69166, at *8–9 (D. Nev. July 18, 2014) (citations omitted); see also Fed. Hous. Fin. Agency v. HSBC N. Am. Holdings Inc., No. 11 CIV. 6189 DLC, 2014 U.S. Dist. LEXIS 19156, at *3 (S.D.N.Y. Feb. 14, 2014) (recognizing that literature showed that “predictive coding had a better track record in the production of responsive documents than human review”); Malone v. Kantner Ingredients, Inc., No. 4:12CV3190, 2015 U.S. Dist. LEXIS 41951, at *3 n.7 (D. Neb. Mar. 31, 2015) (“Predictive coding is now promoted (and gaining acceptance) as not only a more efficient and cost effective method of ESI review, but a more accurate one.”). There will be some who believe there can never be a substitute for the human review of a document in hand (or on screen). However, with the increase of litigation in which ESI runs into the millions of files, integrating predictive coding and analytics into modern legal practice may become necessary or even required as the legal industry and applicable rules change.

What about common client concerns? Predictive coding is not necessarily a cheap solution. Rather, it is a more efficient and often less expensive way to search, code, and produce documents. It will take many hours of focused work from a senior attorney to correctly code each model set. The person coding the model set must be consistent and have a good working knowledge of the case. Clients may understandably fret about the cost to have an experienced attorney reviewing documents. Another downside to predictive coding is that there are no results (i.e., nothing to show for the time spent by a senior attorney) until the process is complete or substantially complete. The process could take weeks, perhaps months, depending on the size of the document universe being coded. It is important that clients understand and approve both the process—including when there will be results—and the cost.

Clients are also justifiably concerned about the inadvertent production of documents. This is a more difficult prospect because there is no guarantee that a privileged or protected document will not be produced. It is arguably more likely that a human review will result in an inadvertent production, but that is difficult for any client to stomach. Three steps may help alleviate this common client concern:

- First, the document universe should be screened for privileged documents at the outset. That will eliminate most of the obviously privileged documents.

- Second, as documents are coded into the model set, a tag should be created for any documents identified as work product or privileged and not identified in the privilege screen. Predictive coding can then be used to search for these documents.

- Third, there should be a strong claw back provision incorporated into an order entered by the court (e.g., a protective order, agreement on ESI, or the Rule 26(f) discovery plan) providing for the automatic return of any such documents.

If a client remains nervous, the client can always elect the greater expense of a manual document-by-document review. However, the client may change his mind after reviewing the proposed budget. This decision point poses the question of, between two paths, each of which is imperfect, what can a law firm offer to its clients that has the least impact on staff hours and the most prospect for finding the needles in the haystack? Computers and predictive coding appear to be the answer.