small_frog/E+/Getty Images

Long Live Predictive Coding

E-discovery experts and other observers expected that counsel and clients would embrace predictive coding programs as a standard practice in document-heavy litigation. While use of this technology has not become commonplace as quickly as many anticipated, courts increasingly recognize the value of predictive coding. With this uptick in court endorsement, counsel should look for ways to incorporate predictive coding technology in furtherance of a cost-effective and efficient discovery process.

WALLIS M. HAMPTON
COUNSEL
SKADDEN, ARPS, SLATE, MEAGHER & FLOM LLP

Wallis focuses his practice on business litigation. He defends companies and their directors and officers in securities and fiduciary duty cases, class actions, government and regulatory investigations, and complex commercial disputes. Wallis has published extensively on best practices for e-discovery.

GIYOUNG SONG
DISCOVERY COUNSEL
SKADDEN, ARPS, SLATE, MEAGHER & FLOM LLP

Giyoung dedicates her practice primarily to discovery-focused litigation, advice, and counsel in federal and state court-based class actions, multidistrict litigation, and other disputes. She has extensive experience litigating and managing complex and sophisticated discovery issues, and regularly advises clients regarding discovery law, best practices, and practical solutions.

Traditionally, keyword search terms that use Boolean logic (for example, "stock /2 option") were the gold standard for identifying potentially responsive electronically stored information (ESI) during discovery. With keyword searching, counsel can electronically apply search terms to select ESI and review documents containing the search terms to determine if they are relevant or privileged. Because the legal industry is not an early adopter of technological advances, this traditional review method has lingered even in the face of staggering volumes of ESI.

Several companies, however, developed advanced algorithms to electronically identify potentially relevant ESI. These more sophisticated methods of technology-assisted review (TAR), including predictive coding, were not immediately adopted by litigants. In recent years, that hesitance has started to dissolve as courts have blessed the use of predictive coding and counsel increasingly have publicized its advantages.

Computer analytics like predictive coding programs are poised to become a standard practice in e-discovery. Some courts already encourage, and may eventually require, parties to use these tools for large cases. The real uncertainty is not whether parties will use predictive coding, but how and when they should use it.

This article examines:

- The basic technology behind predictive coding.
- The ways in which counsel can incorporate predictive coding into case preparation.
- How courts have ruled on the use of predictive coding.
- The advantages and disadvantages of using predictive coding.

UNDERSTANDING PREDICTIVE CODING

Broadly speaking, predictive coding refers to the use of a software program to identify documents that are relevant to a particular case or issue. Predictive coding involves a counsel-guided, machine learning process and a combination of different algorithmic tools.

THE MACHINE LEARNING PROCESS

In general, counsel "train" the program by identifying a set of relevant ESI (seed set) from a broader set of potentially relevant ESI. Experienced counsel who are intimately familiar with the case then individually code each document in the seed set for relevance. The predictive coding program analyzes the seed set ESI and coding to understand the types of ESI that are relevant to the case. The program then uses this information to project whether each remaining document (the ESI not in the seed set) is likely to be relevant or irrelevant.

Through an iterative process, counsel continue to train the program by reviewing the program's coding decisions and accepting or rejecting the projected relevance classifications. The program incorporates this feedback into its coding decisions. Just as human reviewers may reach different decisions on the relevance of the same document, a predictive coding program may not agree with experienced counsel's decision in every instance. Total congruity is not expected. Instead, the goal is to have the predictive coding program agree with counsel's coding for a predetermined percentage of the documents (for example, 95%). The iterative process typically repeats for several cycles until the program's predictive coding is sufficiently accurate when compared to counsel's coding.

After counsel complete the iterative training process, the predictive coding program analyzes and codes each document in the entire data set for relevance. What counsel do with the documents the program identifies as relevant depends on the review team's particular workflow. Initially, counsel should perform a quality control review to ensure that the program did not misclassify a significant number of relevant documents as irrelevant, or vice versa.

Once counsel are confident in the program's relevance classifications, they may either:

- Proceed with a privilege review of the ESI that the program identified as relevant.
- Use the program's relevance calculations to prioritize the order in which document reviewers manually review the ESI for relevance or privilege.

Some e-discovery vendors also offer continuous active learning (CAL) tools. CAL software uses reviewers' real-time feedback about what documents are relevant to continuously refine its predictions about the relevance of all documents in the review set. Because CAL tools learn from the reviewers' coding as they work through the review set, they generally do not require or use a seed set.

Search Continuous Active Learning for TAR for more on how counsel can implement a CAL protocol in the discovery process.

COMMON TOOLS

The range of available predictive coding tools and the various methodologies and algorithms that the tools use for training and coding are sweeping. Most tools, however, use similar methodologies, such as:

- **Concept searching.** Instead of searching one particular word, a concept searching algorithm considers the meaning of a word to identify potentially relevant documents. It relies on different sources to provide the context in which the word appears, including dictionaries, thesauruses, taxonomies (an organization scheme that looks for similar concepts), ontologies (an organization scheme that looks for related concepts), or mathematical formulas that consider the context in which the word appears. For example, if one of the original search terms was "car":
  - a taxonomy-based algorithm might look for automobiles, trucks, and pickups; and
  - an ontology-based algorithm might look for items related to cars, such as drivers and service stations.
- **Contextual searching.** These algorithms consider how and where specified search terms appear in the document, rather than focusing exclusively on search term matches. For instance, if two of the original search terms were "car" and "insurance," the algorithm might focus on whether these related concepts appeared repeatedly in the same discussion.
- **Metadata searching.** Some algorithms focus on certain metadata fields, such as the author, recipient, and date fields, to identify relevant materials. For example, if a certain communication between John Smith and Jane Doe on January 23, 2018 is relevant, a metadata searching algorithm might assign a higher priority to other communications between those people during the same time period.

The program can also organize ESI using one or more of the following tools:

- **Probability theory.** An algorithm based on probability theory makes decisions about how likely a document is to be relevant. For example, a probability algorithm using concept searching might conclude that a document containing 15 relevant search terms or phrases is more likely to be relevant than a document containing only one relevant term.

- **Relevance ranking.** Many predictive coding programs use their algorithms to rank how likely a document is to be relevant. To illustrate, a document containing 15 relevant search terms might have a ranking of "85," while a document with one search term might have a ranking of only "20."
- **Clustering.** This method groups documents with similar content (as determined by the algorithm), permitting a reviewer to view all documents that appear related to a single concept. For example, a clustering algorithm might group all the emails that appear to relate to the same topic, even if they came from different email threads.
- **Sorting documents by issue.** Documents can be sorted and ranked by issues identified by the human reviewers during the training process. This approach can be particularly helpful in identifying the key documents on specific topics at an early stage, or before the start, of the litigation.

Even within these categories, each algorithm is unique. Vendors develop their own proprietary programs and, for obvious reasons, do not share all of the details on how their algorithms work.

Search E-Discovery Glossary for a list of terms commonly used in the e-discovery context.

USING PREDICTIVE CODING IN LITIGATION

Some counsel are reluctant to use predictive coding as the primary review tool until the law and practice around it develop further. Others face resistance from opposing counsel, or even their own clients. But counsel can use predictive coding in a variety of ways, even if not as part of the formal methodology to identify responsive documents. For instance, counsel can use it to:

- Identify key strengths and weaknesses in a client's case during early case assessments and preliminary investigations.
- Streamline aspects of document review when responding to document requests.
- Analyze a document production received from an opposing party or a third party.
- Prepare for depositions, expert discovery, summary judgment motions, and trial.

EARLY CASE ASSESSMENT

Counsel may use predictive coding during their early case assessment to sort through the client's ESI and assess the strengths and weaknesses of the case.

As noted above, many predictive coding programs can rank and sort documents by likely relevance. Review teams can initially focus on the documents identified as most likely to be relevant, which often will contain many of the key documents that form the backbone of the case. Among other benefits, this early case assessment:

- Permits counsel to conduct an early risk analysis.
- Identifies key witnesses.
- Facilitates more efficient resource allocation.

Search Case Assessment and Evaluation and The Advantages of Early Data Assessment for more on conducting effective early case assessments.

REVIEWING CLIENT DOCUMENTS

The value of predictive coding is clear for pre-production search and review. Nonetheless, counsel should consider seeking the opposing party's consent (and, if necessary, the court's approval) before using predictive coding as part of a formal pre-production review process and incurring the associated costs. As a practical matter, counsel may more easily "sell" predictive coding to an opposing party or counsel who have experience with it, or at least have substantial experience with e-discovery generally. Similarly, a judge with substantial e-discovery experience is likely to be more receptive to predictive coding.

Search Rule 26(f) Conference Checklist for more on how to approach the use of predictive coding or other TAR tools for pre-production review with opposing counsel at the initial discovery planning conference.

Even when the parties do not agree to use predictive coding in lieu of traditional keyword searches, counsel can incorporate predictive coding tools into their internal workflows to make the review more efficient and effective. For example, counsel may use predictive coding to:

- **Prioritize pre-production review.** Counsel may use traditional keyword searches to identify the universe of potentially relevant documents and then use predictive coding to organize and prioritize their review of those documents. In this situation, predictive coding would not change the universe of documents set for review. Instead, it would assist counsel in implementing a prioritized search and review method. For example, a party may instruct:
  - its primary law firm to review the documents that are most likely to be relevant (for example, documents that the predictive coding program scored between 80 and 100); and
  - contract attorneys to review the documents that are less likely to be relevant (for example, documents that the predictive coding program scored below 80).
- **Sort documents by potential privilege.** While predictive coding has not proven particularly reliable at identifying privileged information, counsel may use it to rank the likelihood that certain documents are privileged. As with relevance predictions, counsel may allocate the potentially privileged documents to different reviewers based on the likelihood of the document being privileged. Moreover, clustering and email threading can help reviewing attorneys ensure consistency on privilege calls across similar documents (for more information, search E-Discovery: Processing Electronically Stored Information and Considerations for Using Email Threading in Discovery on Practical Law).
- **Quality control a traditional review before production.** Counsel can compare the results of a linear document review (in which an attorney manually reviews documents one after

Even when the parties do not agree to use predictive coding in lieu of traditional keyword searches, counsel can incorporate predictive coding tools into their internal workflows to make the review more efficient and effective.

another) with the predictive coding results on the same set of documents to assess whether reviewers should revisit any decisions on relevance or privilege.

REVIEWING OTHER PRODUCTIONS

Counsel can use predictive coding to review document productions received from opposing parties and third parties. Because counsel often do not know the content and organization of these productions, the ability to quickly rank these documents by potential relevance is extremely valuable, particularly in a fast-moving case.

As with review of client documents, counsel can use one or more of the following tools to organize and understand documents received from an opposing party or a third party:

- Concept and metadata searching.
- Relevance ranking.
- Clustering.
- Sorting documents by issue.

Additionally, counsel may use predictive coding to identify gaps in other parties' productions (in other words, the absence of documents that counsel expected to receive in the production).

Search Discovery Deficiency Letter or see page 22 in this issue for a sample letter notifying opposing counsel of perceived deficiencies in their production and requesting additional discovery materials to remedy those deficiencies, with explanatory notes and drafting tips.

OTHER STAGES OF LITIGATION

Predictive coding has potential uses at other stages of litigation as well, including for:

- Deposition preparation (for example, to assemble deponent-specific materials with high relevance rankings).
- Expert report and deposition preparation (for example, to identify documents concerning the subject of the expert's report and testimony).
- Preparing or responding to summary judgment motions.
- Trial.

LEARNING FROM THE COURTS

For years, predictive coding was mired in a state of limbo. Most practitioners continued to use a combination of traditional keyword searches and linear review while courts largely ignored the issue. That has started to change, as federal and state courts render more decisions addressing whether predictive coding can and should be used. These decisions typically focus on issues concerning:

- The defensibility of a party's use of predictive coding in searching for documents responsive to subpoenas or document requests.
- The level of cooperation between opposing counsel when using predictive coding in litigation.

While disputes on these issues are increasingly making their way into court, predictive coding has not become as common as many experts and observers predicted several years ago. Counsel must understand that judicial treatment of predictive coding remains an evolving area of the law, and courts will undoubtedly continue to refine their views as they are presented with additional factual scenarios and the technology develops.

SEARCH, REVIEW, AND PRODUCTION

Beginning with the landmark decision in *Da Silva Moore v. Publicis Groupe*, courts have approved a party's use of predictive coding to search for responsive documents.

In *Da Silva Moore*, both sides agreed to use predictive coding, but disagreed on the details. The court concluded that predictive coding "now can be considered judicially-approved for use in appropriate cases," but cautioned that its holding did not require parties to use predictive coding in all cases and did not endorse the protocol used in that case as appropriate in other cases. (287 F.R.D. 182, 193 (S.D.N.Y. 2012), adopted, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).)

In the wake of *Da Silva Moore*, courts have variously:

- Approved of or encouraged a party's voluntary use of predictive coding (see, for example, *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125, 127 (S.D.N.Y. 2015) (approving a party's proposed predictive coding protocol and noting that it is black letter law that courts will permit a willing producing party to use TAR); *Malone v. Kantner Ingredients, Inc.*, 2015

WL 1470334, at *3 n.7 (D. Neb. Mar. 31, 2015) (noting the promotion and acceptance of predictive coding as an efficient and cost-effective review tool); *Nat'l Day Laborer Organizing Network v. U.S. Immigration & Customs Enf't Agency*, 877 F. Supp. 2d 87, 109, 111-12 (S.D.N.Y. 2012) (questioning the effectiveness of keyword searches and encouraging a government agency to use predictive coding when responding to a Freedom of Information Act request)).

- Required parties to consider using predictive coding, whether in response to a producing party's claim that a document review is unduly burdensome or otherwise (see, for example, *Johnson v. Ford Motor Co.*, 2015 WL 4137707, at *11 (S.D. W. Va. July 8, 2015) (ordering the parties to consider alternative methods of searching the defendant's ESI, "such as predictive coding")).
- Declined to compel an unwilling party to incorporate predictive coding into its document review and production workflow (see, for example, *City of Rockford v. Mallinckrodt ARD Inc.*, 2018 WL 3766673, at *3 (N.D. Ill. Aug. 7, 2018) (stating that the court "will not micromanage the litigation and force TAR onto the parties"); *In re Viagra Prods. Liab. Litig.*, 2016 WL 7336411, at *2 (N.D. Cal. Oct. 14, 2016) (holding that the court could not compel the producing party to use predictive coding when it preferred to use search terms); *Hyles v. New York City*, 2016 WL 4077114, at *1 (S.D.N.Y. Aug. 1, 2016) (declining to force the defendant to use predictive coding as part of its review process)). Notably, in their reasoning, these courts sometimes invoke The Sedona Conference Principle 6, which acknowledges that a producing party is best positioned to identify effective search and review processes for its own ESI (see *Kleen Prods. LLC v. Packaging Corp. of Am.*, 2012 WL 4498465, at *5 (N.D. Ill. Sept. 28, 2012)).

Additionally, courts have disagreed on whether a party may use a hybrid review method in which the party first employs traditional search techniques like keyword searches and de-duplication to limit the full data set and later applies a predictive coding program to the filtered data (compare *Bridgestone Ams., Inc. v. Int'l Bus. Machs. Corp.*, 2014 WL 4923014, at *1 (M.D. Tenn. July 22, 2014) and *In re Biomet M2a Magnum Hip Implant Prods. Liab. Litig*, 2013 WL 1729682, at *3 (N.D. Ind. Apr. 18, 2013) (permitting keyword filtering before predictive coding) with *FCA US LLC v. Cummins, Inc.*, 2017 WL 2806896, at *1 (E.D. Mich. Mar. 28, 2017) and *Progressive Cas. Ins. Co. v. Delaney*, 2014 WL 3563467, at *9-11 (D. Nev. July 18, 2014) (prohibiting keyword filtering before predictive coding)).

COOPERATION AND TRANSPARENCY

Courts have varying perspectives on the requisite levels of transparency and cooperation required when a party uses predictive coding. Disputes between parties have largely focused on the extent of disclosure a producing party should make about its seed set and training and validation processes.

Some courts have encouraged and recognized the benefits of a producing party's willingness to share its seed set with the opposing party, noting that this level of transparency heightens the opposing party's and the court's comfort with the process (see, for example, *Rio Tinto*, 306 F.R.D. at 128-29; *Bridgestone*, 2014 WL 4923014, at *1; *In re Biomet M2A Magnum Hip Implant Prods. Liab. Litig.*, 2013 WL 6405156, at *2 (N.D. Ind. Aug. 21, 2013); *Da Silva Moore*, 287 F.R.D. at 192).

Relatedly, at least one court ordered a party to permit the opposing party to "play an active role" in the predictive coding training process to foster transparency (*Indep. Living Ctr. v. City of Los Angeles*, No. 12-551, slip op. at 1-2 (C.D. Cal. June 26, 2014); see also *Progressive*, 2014 WL 3563467, at *10-11 (prohibiting a party from using predictive coding when the party refused to disclose training documents to the opposing party)).

Other courts, however, have acknowledged that the Federal Rules of Civil Procedure do not require a party to disclose information that is not relevant to a claim or defense, such as information about a party's predictive coding process, and therefore have declined to compel a producing party to disclose its seed set (see, for example, *In re Biomet*, 2013 WL 1729682, at *2 and 2013 WL 6405156, at *1-2).

Some courts have also suggested that seed sets and information on training and validation processes implicate the work product doctrine. For example, in one widely publicized decision, a court found that details surrounding a producing party's predictive coding process were protected from disclosure as attorney work product. After an *in camera* review, the court also found that the producing party was not negligent in its training process. Yet despite these findings, the court concluded that the requesting party had elicited sufficient information suggesting flaws in the training process to justify having the producing party disclose, on an attorneys'-eyes-only basis, a limited sample of documents that the predictive coding tool had designated as non-responsive. (*Winfield v. City of New York*, 2017 WL 5664852, at *11-12 (S.D.N.Y. Nov. 27, 2017).)

Predictive coding requires significant attention from experienced counsel during the machine learning process. A flawed seed set or training process will cascade those flaws throughout a production.

Parties evaluating predictive coding in a particular case should consider any relevant case law in their jurisdiction regarding the required levels of disclosure and the parties' willingness to be transparent about their process.

DECIDING TO USE PREDICTIVE CODING

When deciding whether to use predictive coding, counsel should first identify and consider any available case law or other relevant authority regarding review technologies because some jurisdictions have developed jurisprudence or rules on this issue.

Some federal courts have incorporated references to using predictive coding in model conference orders (see, for example, Fed. Jud. Ctr., Mandatory Initial Discovery Pilot Project Standing Order at C(2)(a)(ii), available at *fjc.gov*). Additionally, at least one state forum for complex commercial disputes has proactively amended its rules to encourage parties to use the most efficient way to review documents. The amended rules specifically reference the appropriateness of using predictive coding during document review and production. (See N.Y. State Supreme Court, Commercial Division Rule 11-e(f) (22 NYCRR § 202.70(g)), effective Oct. 1, 2018.)

Many jurisdictions have not yet weighed in on when a party may or should use predictive coding. In these jurisdictions, counsel and litigants must consider whether the circumstances of the case favor the use of predictive coding.

Search Technology-Assisted Review: Advice for Requesting Parties for more on issues to consider when deciding whether to use predictive coding or other TAR processes in discovery.

ADVANTAGES

The greatest advantage of predictive coding is the potential to dramatically reduce the number of documents that counsel must review, which ultimately saves time and money (although some people have questioned the true scale of these savings). Predictive coding also can:

- Foster a higher level of consistency in the review process by minimizing the inconsistent production and privilege calls that plague every large document review.
- Identify more relevant documents than a traditional linear review.
- Substantially reduce the risk of being accused of deliberately hiding relevant documents, because it is easier to justify the nonproduction of an important document when the predictive coding program coded it as nonresponsive.

DISADVANTAGES

Predictive coding has its downsides and limitations. Most significantly, it is not yet a standard practice so there is uncertainty about how a court or opposing counsel might view it. Not all predictive coding programs (or vendors) are created equal, and deciding which ones are best for a particular case can be challenging.

Further, many algorithms cannot effectively evaluate spreadsheets or documents without searchable text. Similarly, most commonly used predictive coding programs cannot yet reliably analyze other file types, such as videos, graphics, and audio files, which may be critical in certain types of cases. Therefore, counsel will need a good vendor and a strong project manager to tailor the predictive coding program to meet the specific challenges in the case.

Search Choosing Outside E-Discovery Service Providers and Questions to Ask a Prospective E-Discovery Vendor Checklist for more on selecting an e-discovery vendor.

Counsel must also consider whether the opposing party is willing to use predictive coding itself or objects to its use in the case by any party. The most common scenario for party agreement on predictive coding involves litigation in which both sides face substantial production obligations. In those cases, the parties' interests in making discovery as efficient as possible are more likely to be aligned. Conversely, in asymmetric cases, the party with fewer documents to produce may be less open to more novel or unfamiliar approaches to document review because they do not otherwise face an expensive and time-consuming review process.

In both symmetric and asymmetric cases, opposing counsel may press to be actively involved in developing a precise predictive coding protocol. For example, opposing counsel may seek to:

- Help select the seed set.
- Participate in coding the seed set.
- Review the predictive coding program's initial relevance classifications, including documents that the program classified as irrelevant.

Depending on the forum court, opposing counsel may gain access to review irrelevant but still sensitive or damaging documents included in the seed set that would otherwise be outside the scope of discovery.

Finally, predictive coding requires significant attention from experienced counsel during the machine learning process. A flawed seed set or training process will cascade those flaws throughout a production. To guard against this risk, counsel must commit substantial time and financial resources at the start of a case. For this reason, many have questioned whether predictive coding is more cost effective than traditional attorney review, particularly in smaller cases.

Search Reducing E-Discovery Costs: Applying an Analytical Approach for more on strategies to minimize e-discovery costs.

The views expressed in this resource are those of the authors and not necessarily those of Skadden, Arps, Slate, Meagher & Flom LLP or its clients.