KEYWORD VS. PREDICTIVE CODING:
THE SHIFT AWAY FROM BASIC SEARCHING

Increasingly, modern litigation practice requires the review and production of electronically stored information, or "ESI." In many cases, ESI can involve document volumes that make traditional, page-by-page reviews prohibitively expensive and time consuming. When faced with litigation involving prohibitively large volumes of potentially responsive documents, practitioners must select a method of culling the universe of documents down to a manageable size.

Practitioners have traditionally relied upon keyword searches to identify responsive documents located within the mass of ESI in the custody and control of document custodians. A "keyword search" is one in which the full text and metadata of ESI is reviewed via computer program searching for single words or simple word combinations. The search can be a simple search or an enhanced "Boolean" search, which utilizes Boolean logic to search for words in different relationships with one another. Keyword searches are useful in locating specific documents or where the searched-for language is predictable. Keyword searching is relatively easy to understand and implement. It also has enjoyed an extended period of acceptance in the federal courts. *See Spannaus v. C.I.A.*, 841 F.Supp. 14, 17 (D. D.C. 1993) (approving CIA's use of keyword search to narrow production); *Venturedyne, Ltd. v. Carbonyx, Inc.*, No. 2:14-CV-351-RL-JEM, 2016 WL 6694946, at *1 (N.D. Ind. Nov. 15, 2016) ("Keyword searches 'have long been recognized as appropriate and helpful for ESI search and retrieval…")(citing *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 250 F.R.D. 251, 260 (D. Md. 2008)).

Courts, however, have increasingly noted the limitations inherent in the use of simple keyword searching. "[T]here are well-known limitations and risks associated with [keyword searches]…'" *Carbonyx, Inc.*, No. 2:14-CV-351-RL-JEM, 2016 WL 6694946, at *1. "Chief among [those limitations] is that such a search necessarily results in false positives (irrelevant documents flagged because they contain a search term) and false negatives (relevant documents *not* flagged since they do not contain a search term)." *Makowski v. SmithAmundsen LLC*, No. 08-C-6912, 2012 WL 1634832, at *1 (N.D. Ill. May 9, 2012). As a result, "[e]lectronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI." *William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co.*, 256 F.R.D. 134, 136 (S.D.N.Y. 2009).

Given the limitations and risks inherent in keyword searching, courts have begun expressing a preference for technology assisted review programs ("TAR"), also sometimes referred to as predictive coding. Generally, this approach involves a program running an algorithm to produce a small sample of documents the program believes to be relevant and responsive. A human reviewer then reviews the sample and tags the documents within the sample as "highly relevant," "relevant," and "not relevant." Those tags are then fed back into the program, which then "learns" from the tags and refines the program to produce a new, hopefully better, sample. A human reviewer then reviews the second sample and tags that sample again to refine relevancy. This process continues through several iterations until the algorithm reaches the desired confidence level (e.g. 95% of the documents reviewed in a given sample a correctly identified by the algorithm as "relevant" or "not relevant.")

Courts have increasingly begun expressing a preference for the use of TAR over simple keyword searches. *See Moore v. Publicis Groupe*, 287 F.R.D. 182, 185 (S.D.N.Y. 2012) ("Key words, certainly unless they are well done and tested, are not overly useful. Key words along with predictive coding and other methodology, can be very instructive."); *F.D.I.C. v. Bowden*, 2914 WL 2548137 at *12 (S.D. Ga. June 6, 2014) (*sua sponte* ordering the parties to consider the use of predictive coding); *Green v. American Modern Home Ins. Co.*, 2014 WL 6668422 at * 1 (W.D. Ark. Nov. 24, 2014) (approving use of predictive coding a part of discovery protocol). One court has gone so far as to state that "the case law has developed to the point that it is now black letter law that where the producing party wants to utilize TAR 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).

Even where TAR is used, it is highly recommended that other techniques, including sampling, be used to verify the set of data identified as non-responsive. *Moore* provides an instructive guide to this process. There, the parties agreed upon a 95% confidence level of the predictive coding [e.g. those documents that the program determined to be responsive according to the coding instructions with a 95% or higher confidence level] to generate a sample of documents to be reviewed by the parties to ensure the predictive coding was generating accurate results. *Id.* at 186-87. The requesting party then reviewed this sample – 2,399 documents – and suggested adding two new issue tags to the predictive coding algorithm to be used in the final review. The parties further agreed on additional coding that would be used to generate the production. *Id.* The court made a point to note that the TAR protocol was created by "senior attorneys (not paralegals, staff attorneys, or junior associates).

Despite the cooperation between the parties to develop the TAR protocol – and the hands-on assistance of a federal magistrate judge – the requesting party in *Moore* objected to the use of TAR, primarily on the grounds that there was no way to assess the reliability of the coding used. *Id.* at 188-89. The court disagreed, finding that the requesting party's fears that the coding would miss relevant documents were premature, and could be addressed further along in the process.

Perhaps of greater interest, however, was the court's *dicta* consideration of the appropriateness of TAR in the face of an objection by a party. The court set forth its suggested criteria in making such a decision:

> The slightly more difficult case would be where the producing party wants to use computer-assisted review and the requesting party objects. The question to ask in that situation is what methodology would the requesting party suggest instead? Linear manual review is simply too expensive where, as here, there are over three million emails to review. Moreover, while some lawyers still consider manual review to be the "gold standard," that is a myth, as statistics clearly show that computerized searches are at least as accurate, if not more so, than manual review.

Id. at 188-89.

Courts have not insisted on the use of TAR, however. In one recent case in the Northern District of Indiana, the court refused the plaintiffs' steering committee's request that the defendant be ordered to re-start the discovery process using predictive coding where the defendant had unilaterally decided to use a keyword search to narrow the universe of responsive documents

before using predictive coding to search the documents first identified by the keyword search. See *In re Bioment M2a Magnum Hip Implant Products Liability Litigation*, 2013 WL 1729682 (N.D. Ind. April 18, 2013) (holding that defendant's use of both keyword searches and predictive coding was not a violation of the federal rules on discovery).

TAR will generally be more effective than a simple key word search. It is also often more intensive and costly. Selecting between TAR and key word searching will therefore require an assessment by counsel of the appropriateness of TAR based upon the stakes of the litigation and the volume of ESI that will need to be reviewed. Nevertheless, based up court's increasingly stating a preference for the use of TAR in complex, document intensive litigation, TAR should be at the forefront of a practitioner's mind when selecting the appropriate ESI review process.