New York Law Journal

NEW YORK LAW JOURNAL SPECIAL REPORT

An ALM Publication

Litigation

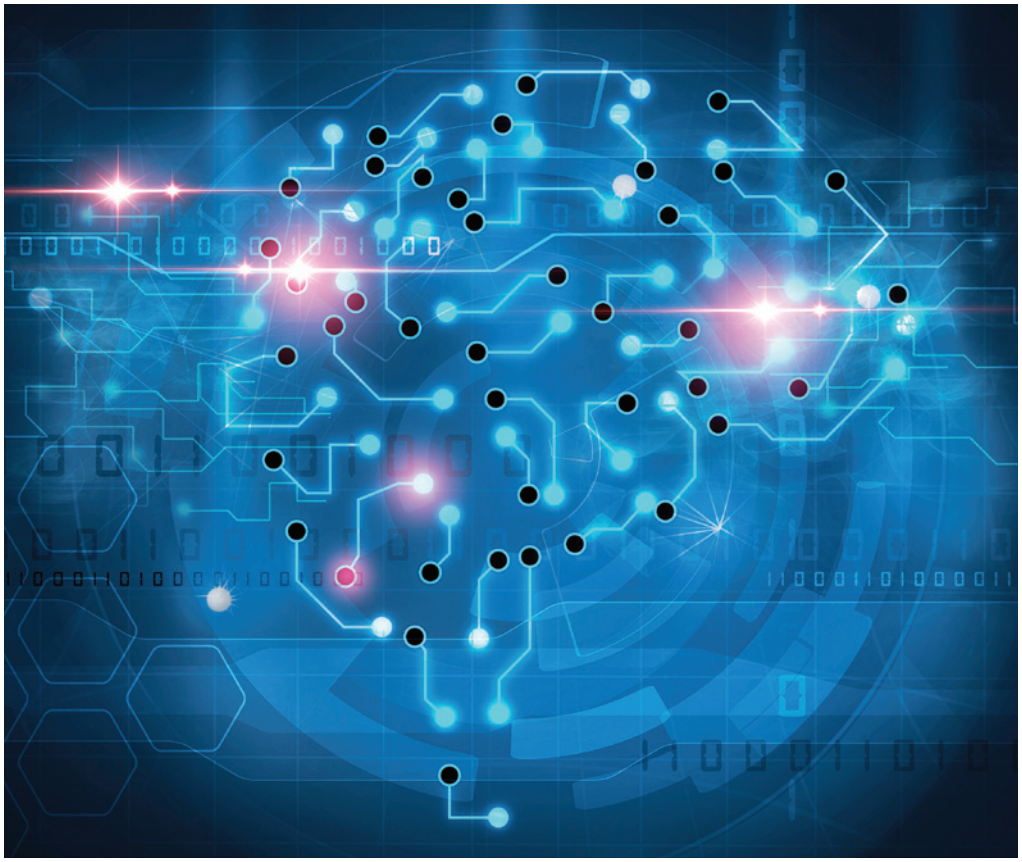
WWW.NYLJ.COM

VOLUME 258—NO. 92

MONDAY, NOVEMBER 13, 2017

AI in Discovery: The Future Is Now

BY ROBERT J. BURNS, BENJAMIN R. WILSON AND JOAN M. WASHBURN



SHUTTERSTOCK

Few current topics in legal practice generate the extremity that artificial intelligence does. Utopians promise a radical transformation in litigation, with machines doing our dirty work and leaving us to higher endeavors. Dystopians counter that AI will soon make us all superfluous and unemployed.

This article will not adjudicate the AI dispute. Nor will it assure the reader that he or she may continue litigating in the old familiar ways, blissfully unaware of AI It may be viewed as good news, or bad news, but it is a fact: This technology is here, it is transformative, and it is gaining judicial traction. If an adversary, a judge, or a client hasn't already asked you whether AI should be employed in your next discovery process, they will soon. Be prepared.

By now, everyone knows we are surrounded by an ocean of data. Clients generate new data at unprecedented speeds and volumes. But when that data becomes discoverable, prior discovery technologies have only increased the risks of drowning. Studies have shown that keyword searching coupled with linear review is ineffective and extremely costly. Practice has shown that first-generation computer-assisted methodologies (TAR), while more precise, are challenging to implement and still too costly. As a result, despite the initial transformative promise of TAR, e-discovery remains the most dreaded phase of litigation—for counsel doing the work, for judges resolving the disputes, and for clients paying the bills.

ROBERT J. BURNS *is a partner and* BENJAMIN R. WILSON *is an associate in Holland & Knight's New York office.* JOAN M. WASHBURN *is the firm's director of litigation e-discovery services.*

But AI technologies now in active deployment—and enjoying the first blush of judicial endorsement—are likely to transform this process in the near future.

The Past

Since the early 2000s, litigators have already endured several great leaps in discovery practice. In the first stage of evolution, we emerged from dusty file warehouses, and we (or armies of contract lawyers) conducted linear reviews of electronic files from the comfort of our screens. But we soon realized that our clients were generating data faster than our ability to review it meaningfully.

In the second stage, we employed filtering methodologies and keyword searches to cut through the data volume. But those methodologies, too, required linear review of the substantial balance of documents identified. These methodologies also lacked precision, generating numerous false hits. And they triggered competing tensions: plaintiffs' motivation to exhaust all potential sources of discoverable information and defendants' desire to minimize burdens and expenses. The net result was more, not fewer, discovery disputes.

The third stage—TAR, or "predictive coding"—initially seemed the solution. In 2012, the first federal judge endorsed TAR as a viable e-discovery tool, and by 2015, it was "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). TAR was enthusiastically embraced by the litigation community as an emerging best practice: TAR, we thought, promised improved effectiveness and efficiency in e-discovery, and a greater understanding of the client's data sets.

But over time, it has become clear that TAR is not the hoped-for panacea. Traditional TAR protocols require a substantial amount of expensive, upfront work. That work should be done by senior lawyers who understand the client's data and the factual and legal issues in the case. Such lawyers have high price tags and severe time constraints, but TAR needs them to review rounds of documents to complete the control and seed set, and to repeat this process until stabilization occurs.

TAR presents other issues, too. Its efficacy depends, in significant measure, on cooperation and transparency among counsel. But courts have been reluctant to compel that transparency, and in its absence uncertainty reigns: The propounding party fears that the algorithm was (intentionally or negligently) trained to miss relevant documents, the responding party fears costly "do-overs," and both parties face costly motion practice to sort this out. These shortcomings have caused some litigants to revert to the old-fashioned techniques that most had thought abandoned.

The Future (Is Now)

The fourth evolutionary stage is at hand. Continuous Active Learning technology (CAL) allows parties to identify relevant documents with far less effort and cost. Its functionality is simple: Using either key documents or keywords, the technology retrieves ESI, presenting first the documents most likely to be of interest, followed by those less likely to be of interest. The technology continually improves its understanding of what is likely to be relevant. And as that understanding strengthens, the database re-ranks each document in the collection by its likely relevance.

What does this mean in practice? In short, an e-discovery process that more closely parallels the fact-development process we undertake in our cases. Consider this scenario:

A dispute has ripened into litigation, or the threat of it. Your client presents you with a stack of documents that, based on initial conversations with the key players, seem to matter most to the dispute: contracts, demand letters, correspondence, witness statements, and the like. In the old days, you might review those key documents to extract keywords most likely to unearth similar documents. Or, if using first-generation TAR, your review team might use these documents to guide their manual efforts to unearth similar

documents and, thereby, train the algorithm. In each instance, humans needed to figure out the story of the case, and then figure out how to tell that story in a language a machine would understand.

But using CAL technology, the machine ingests the fruits of your initial investigation and takes it from there. The database's algorithm performs contextual analysis of your "hot documents" and identifies relevant authors, recipients, and custodians. It identifies the types of documents most likely to be relevant. It searches these materials for words in contextual combinations similar to those in the ingested key documents. And it grows more intelligent as it continues these analyses, with minimal human guidance, ultimately returning a discrete subset of relevant ESI.

This technology requires minimal attorney oversight and minimal refinement. It has no need for recall rates or reliance on seed sets determined through labor-intensive rounds of review. Research has shown that, effectively used, CAL culls higher levels of relevant documents more quickly and with less effort (and therefore cost) than prior e-discovery tools. And, most importantly, CAL comports with the methods litigators actually use to litigate cases: Skilled human lawyers do the initial investigation to determine the story of the case and to identify the key documents telling that story, and machines then dig through data sets to identify all other documents that bear on that story, refining their (and, by extension, the lawyers') "learning" about the case along the way.

This technology now exists and is available for use. And CAL has already received some initial judicial support. Magistrate Judge Andrew Peck, recognized within the Southern District of New York as an expert in e-discovery issues, has twice expressed favorable views regarding CAL as an efficient and effective tool. See *Rio Tinto*, 306 F.R.D. at 128; *Hyles v. New York City*, 10 Civ. 3119, 2016 WL 4077114, at *3 (S.D.N.Y. Aug. 1, 2016). We have every confidence that, as CAL technology enters wider use, other courts will join Judge Peck in support.

For early adopters, the scarcity of case law means there is not (yet) a well-defined judicial roadmap for a defensible CAL-based methodology. In the meantime, we offer the following suggestions to maximize your prospects for success:

First, CAL is science, not magic. Upfront issues of preservation and collection remain, and are critical to an effective process. A law firm and its clients must identify, preserve, and collect data sets likely to be relevant, and that data must be converted to structured form in a database. Look to collect and convert a data set with maximum potential to contain relevant information. The more comprehensive the materials are, the smarter the algorithm will be from the outset, netting responsive documents more quickly. Further, a complete and diverse document set will mitigate preservation issues and will ease concerns that key documents were missed.

Second, work closely with technical support staff to develop a CAL methodology that is fully documented, technically sound, and consistent with your client's data systems and architecture. Look also to the well-developed body of case law on TAR defensibility; defensibility concerns in those cases will be relevant to courts in assessing CAL methodologies.

Third, embrace cooperation and transparency. The best defense to any e-discovery methodology is that both parties understand and have agreed to it. Ideally, both parties would compile sets of key documents, and confer to compile the fullest and fairest set of key documents for initial ingestion into the CAL system. In any event, helping your adversary understand the technology, and the manner in which you will implement it, will minimize the risks of motion practice and potentially costly do-overs down the road.

In sum, a well-designed and well-executed process utilizing CAL technology might be what you—and your client, your adversary, and your judge—are seeking to minimize costs, maximize efficiency, and fast-track your case to its more fruitful, and more enjoyable, stages.