Big Data and Artificial Intelligence: Implications for E-Discovery's Future

Mark S. Sidoti and Luis J. Diaz

New York Law Journal, February 6, 2017

'Big Data"—extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions—is raising new challenges and ethical concerns in litigation. Many corporate legal departments are exploring new technologies, including sophisticated data processing and technology-assisted review tools. As this technology evolves, all signs point to the continued emergence of data evaluation systems based on artificial intelligence (AI) that will have the ability to analyze massive data stores, and even deduce patterns of human behavior, at a fraction of the current cost.

Big Data is increasingly prevalent in the age of the "Internet of Things," with information that is potentially relevant to litigation, and thus targeted in e-discovery, increasing exponentially. Electronically stored information (ESI), which overtook paper discovery years ago, now involves much more than the emails and documents found on enterprise-connected computers. It also derives from social media sites like Twitter, Facebook and LinkedIn, stand-alone data storage and Internet-connected devices such as smartphones and automobile "black boxes," and new-to-market "smart" household devices like thermostats, security systems and AI-driven home personal assistance devices like Amazon Echo, all of which store data locally and/or on the cloud.

The sheer volume of this data is overwhelming, with worldwide storage of digital information now estimated to be around three zettabytes (one zettabyte is equivalent to 152 million years of high-definition video). With mountains of data accumulating each minute, it is not uncommon to see a complicated litigation matter that involves a terabyte or more of data. In short, Big Data is here to stay, making it difficult, if not impossible, in many cases to conduct cost-effective searches to identify and segregate relevant data for e-discovery purposes without the use of AI-enabled technology. The legal industry has responded by implementing a set of increasingly sophisticated software tools. This trend is expected to continue as software continues to incorporate AI to deal with Big Data in e-discovery.

AI-Enhanced Review

Not so long ago, e-discovery involved primarily the use of keywords that attorneys and consultants would utilize to run Boolean searches against potentially discoverable data sets and newer, "passive" analytic technologies like concept searching and clustering. However, given the exponential increase in data volume, human analysis (linear review) and even so-called "unsupervised machine learning" alone is no longer practical for many lawsuits and investigations. Today, the state-of-the-art in e-discovery is AI-assisted predictive coding. This technology generally uses small "seed sets" of pre-selected relevant documents and the judgment of "subject matter expert" (SME) reviewers to "teach" the AI system to recognize patterns of relevance in the larger document set and rank documents accordingly. The goal of training the system is to ultimately allow the computer—with varying degrees of continuous human input—to accurately predict relevance for the remaining documents in the larger data set. This "active machine learning" process is interactive and iterative, with the most recent studies showing that cycles of continuous reviewer and SME feedback trains the system to more accurately and efficiently select relevant documents.

The research from several studies confirms that predictive coding through a continuous, active machine learning process is indeed faster and less expensive and, most importantly, more accurate than earlier, more labor-intensive methods of identifying relevant evidence in large data sets. See Gordon V. Cormack and Maura R. Grossman, "Evaluation of machine-learning protocols for technology-assisted review in electronic discovery," Proceedings of the 37th International ACM SIGIR Conference on Research & development in information retrieval. ACM, 2014. In an earlier study by RAND Corporation titled "Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery," the authors concluded that "predictive coding has the potential to lower the cost of unwieldy e-discovery processes by reducing the number of documents requiring human review." That study was based on a review of 57 case studies from eight large corporations and an analysis of the relevant research on electronic discovery review processes, including attendant costs.

The courts have generally welcomed predictive coding. For example, in *Global Aerospace v. Landow Aviation*, No. 61040 (Loudoun County, Va. Cir. Ct. April 23, 2012), the trial judge allowed the use of predictive coding by defendants to review over two million documents over the requesting plaintiff's objections. In an exhaustive memorandum, the defendants successfully argued that predictive coding was "capable of locating upwards of seventy-five percent of the

potentially relevant documents and can be effectively implemented at a fraction of the cost and in a fraction of the time of linear review and keyword searching." Judicial acceptance of predictive coding will likely expand now that the issue of proportionality in discovery and the attendant costs have been brought in to the spotlight by the recent amendments to Federal Rule of Civil Procedure 26.

And already, divergent approaches to the application of AI to e-discovery are emerging. In his series of related blog posts collectively entitled, "Using Hybrid-Multimodal Methods—Predictive Coding 4.0 and Intelligent Spaced Training (IST)," Ralph Losey argues convincingly that the most effective use of predictive coding through AI-based review systems requires constant guidance and ultimate decision making authority by SMEs throughout the process. Losey describes a workflow in which the AI system is continuously trained by the reviewers through a "positive feedback loop" that "continues until the computers predictions are accurate enough to satisfy the proportional needs of the case." A similar but somewhat competing vision is presented by Maura Grossman, whose research with Gordon Cormack is at the forefront of this field. Grossman posits the application of AI based on a "continuous active learning" methodology, which relies more heavily on the computer's initial relevancy ranking determinations.

There are several companies today offering AI-enabled ESI searching capabilities. Products like Kroll Ontrack's EDR software, Catalyst's Insight Predict, BlackStone's Discovery IQ, and Cognitive Analytics NexLP Story Engine incorporate AI into their platforms to not only assist with standard ranking and coding of large document sets, but also develop searches that will identify each category of documents and use AI-driven concept search, enhanced threading, in-depth filtering, visual analytics and sentient analysis. Even so, these systems represent merely the tip of the iceberg in terms or harnessing the potential of AI in the legal field.

The Next Generation

Just as robotic surgical systems have revolutionized the way doctors approach minimally invasive surgery, AI technology is changing the practice of law, even beyond the ESI search and review process. Future systems will continue to leverage increasingly complex predictive coding algorithms for ESI review and beyond. Within the next few years, AI-enabled systems will include natural language interfaces to improve usability. Thus, a lawyer working on a complex breach of contract matter will be able to call his robotic "assistant" via any telephony

device and request in plain English that it find all communications and interactions between two key witnesses occurring within a specified date range, prepare a summary report of the substance, and update the same on a weekly basis as additional data is processed. Similarly, systems like ROSS, which is built on IBM's cognitive computer Watson, is designed to read and understand natural language, develop theories of a case when asked questions and conduct legal research. ROSS is on the market and has already been deployed by several major law firms. With appropriate human input, AI-based legal assistants will continuously "learn" from handling various assignments and gain greater efficiency and accuracy in assisting the lawyer.

Other subject-matter applications that perform automated tasks—called "bots" in the AI world—are being developed to assist with specific legal issues. Capable of checking files, linking related data, and producing results on a 24/7 basis and at a fixed cost, bots can be expected to emerge in the near future to aid lawyers with investigation, discovery, complex fact analysis and research. For example, in a pharmaceutical products liability cases, a bot could be used that mines big data in social media on a global basis to find evidence of otherwise obscure adverse events. In a complex breach of contract matter, a bot could assist with locating discovery and conducting legal research regarding performance obligations by the various parties. Eventually, some predict, bots will do much more than locate and rank relevant documents at computer speeds or conduct simple research. They will be able to identify and extract pertinent issues in the case utilizing a natural language interface that will not depend on specific terms, but rather on patterns of behavior, or concepts, buried in terabytes of data.

Conclusion

AI will revolutionize the field of e-discovery, and to some degree the practice of law, in the years to come. The obvious promise of ROSS and other bots developed using similar systems is that they can provide greater efficiency and cost-savings for common legal tasks, including e-discovery. These AI-based legal assistants are also expected to make e-discovery more accessible and affordable to individuals and mid-sized companies in litigation, thus enhancing access to justice. However, unlike medicine, the law will always remain a social science. While AI can assist practitioners in identifying relevant evidence in a case, it cannot do this without constant training and refinement by competent lawyers, and it will never replace the skill and judgment of a lawyer in utilizing that evidence in advocacy.