ARTIFICIAL INTELLIGENCE CREATES E-DISCOVERY EFFICIENCIES, CONTROLS COSTS

Artificial intelligence is gaining greater acceptance in the legal profession, especially in relation to analyzing large sets of electronically stored information, because it can solve budget and staffing constraints. The use of AI lets legal departments do more with less and frees attorneys for more legal-knowledge-driven work. Many attorneys are familiar with basic e-discovery tools such as simple word searches to find key documents, eliminate duplicates, and connect conversations, but AI has moved beyond these basics.

One of the most powerful e-discovery technologies is predictive coding, which searches documents for context, concepts, and tone. It greatly increases accuracy and relevance in document review, and completes tasks in minutes, not days or months. "The demand for computer-assisted techniques to do large document reviews is so great that predictive coding has become the one mainstream application of machine learning techniques in the legal industry," said Warren Agin, principal at Analytic Law and founding chair of the American Bar Association's Legal Analytics Committee.

"AI can help to find the story in the data."

At its most basic, AI refers to computers learning to perform tasks usually done by people.[1] Using AI in discovery often involves technology such as predictive coding, a form of technology-assisted review (also known as computer-assisted review) and a category of machine learning.[2]

MACHINE LEARNING AND DEEP LEARNING LEVERAGE ANALYTICS

"AI can help to find the story in the data and evaluate electronically stored information to suggest key issues, confidentiality, and overall case relevance," said George Socha, cofounder of EDRM and managing director in BDO's technology and business transformation services. To do that, AI uses analytics tools and methods, may use machine learning (a subcategory of AI involving computers learning with experience), and often involves statistics.[3] Among the analytics tools are algorithms, which provide instructions for categorizing documents.

Algorithms apply computational methods and statistics, given factual inputs and subsequent results, to weight key pieces of information about documents and to develop rules for their categorization. The computer learns how to apply those rules to new information, deciding whether a document meets the selected criteria and even prioritizing

documents it identifies as potentially being more relevant or privileged according to the rules it has developed for quicker manual review.[4] "Machine learning can efficiently find relationships using inductive reasoning," Agin said.

Deep learning is a subset of machine learning that increases accuracy compared with basic machine learning. Basic machine learning requires perfect matches to identify patterns, while deep learning recognizes and categorizes information with only pieces of a whole.

Deep learning allows recognition of patterns when most of the parts are present, but the system isn't thrown off track by a small deviation from its prior experience. Deep learning systems may have different layers, with each layer comprising different algorithms.

STATISTICAL SAMPLING VALIDATES RESULTS

Attorneys may be troubled by the possibility that predictive coding could overlook relevant documents. One way to address this concern is through data sampling, a process that involves reviewing a set of sample documents and extrapolating the results to the entire population of a document production.

"Machine learning can efficiently find relationships using inductive reasoning."

To check for reliability in document production, reviewers select a statistically valid random sample of documents to review manually for responsive documents. Human reviewers look at the selected sample to determine whether the software accurately identified the documents. If the sample contains too many irrelevant or fewer than expected relevant documents, the algorithm is run again (after being fed additional examples of relevant documents), and another sample set is generated.[5] The process continues until the human reviewers are satisfied with the accuracy of the results.

Statistical measurements are indispensable for ensuring that the technology produces an accurate, defensible result. While no means of screening for privilege is perfect, statistical testing of the screen can bring greater confidence in the accuracy of the outcome.

QUALITY SEED SETS ARE KEY TO QUALITY RESULTS

"Garbage in, garbage out" has long been an axiom in computer science. For machine learning and deep learning to fulfill their potential, the training set used to teach the computer must be carefully considered. To create the training materials, known as a "seed set," expert reviewers select a representative cross section of documents from the full population that needs to be reviewed.[6]

The reviewers then code, or label, each document in the seed set as responsive or unresponsive and input those results into the predictive coding software. The AI or

machine learning software analyzes the seed set and creates an algorithm for predicting the responsiveness of future documents. Reviewers then test as described above to verify accuracy and refine the algorithm until the desired results are achieved.[7]

When seed sets are done well, predictive coding greatly reduces the volume of data requiring manual review and increases accuracy. "Machine learning systems are limited only by the quality of the data and the power of the computers running them," Agin said.

BIAS IN DATA AND ALGORITHMS AFFECTS MACHINE LEARNING

Artificial intelligence is as prone to bias as humans are. Bias can seep into algorithms and the data machine learning uses to train, influencing the results.[8] Bias can arise when data used to calibrate machine-learning algorithms is insufficient or the algorithms themselves are poorly designed. The data picked by the trainers is subject to the trainers' biases, so they must be vigilant to avoid bias in assembling seed sets.[9]

Bias in assessing electronic data can be a significant problem if, for instance, only text-search software is used to select files and only text analytics is used to evaluate the files. This is true because sometimes electronically stored information is stored as images and is therefore invisible to text-only searches unless the images are converted to text.[10]

In addition, if the person training the AI lacks sufficient knowledge to accurately gauge the difference between responsive and unresponsive documents, the engine will learn incorrectly.[11] The key to combating bias in machine learning is to assume it exists and work to remove it.

"Don't ignore the human element when utilizing artificial intelligence."

Sterling Miller, general counsel of Marketo, a Silicon Valley—based marketing technology company, said, "Getting the full benefit of AI plus e-discovery requires more than just the technology. It requires a team of people who understand the technology and the litigation process. AI is just a tool-it takes savvy lawyers to make it work. Don't ignore the human element when utilizing artificial intelligence."

CONCLUSION, RECOMMENDATIONS, AND COST EFFICIENCIES

When machine learning analytics technology is applied to electronically stored information, a highly navigable framework can be created that enables lawyers to see connections they might not have considered at the beginning of the discovery process. Machine learning leverages sampling techniques and advanced algorithms to predict whether documents are responsive to criteria established by the e-discovery team.[12]

As artificial intelligence is refined to produce increasingly accurate and complex results, costs related to e-discovery will be reduced and confidence regarding responsiveness increased (provided the underlying algorithms and seed sets are sound). As these advances occur, lawyers will need to keep pace to use them to the fullest. "The typical lawyer isn't going to be able to learn how to build these tools, but they should develop an understanding of the different types of machine learning tools, and understand their capabilities and limitations," said Analytic Law's Agin.

These increasing efficiencies will allow attorneys to reduce time spent reviewing document collections and instead allow them to focus legal knowledge on the complexities of their work. "Lawyers need to understand that not only will the marriage of artificial intelligence and e-discovery save you money, it will allow you to make much better strategic decisions about litigation because of the way it will analyze documents and bring back deep insights that you and your outside counsel might have otherwise missed," Miller said. "AI will do the grunt work and help you do the heavy thinking, and that is the best use of lawyers."

Costs related to e-discovery will be reduced.

The benefits of using predictive coding in e-discovery are that it:

- Prioritizes or eliminates documents based on relevancy scores
- Lowers costs by reducing the number of documents requiring manual review
- Sorts documents so they can more easily be assigned to specific reviewers
- Enables strategic decision-making earlier in a case

Predictive coding increases accuracy while reducing the time and expense of document review by using machine learning technology combined with expert reviewers.

ABOUT CANON DISCOVERY SERVICES

Canon Discovery Services has been helping organizations deal with the scope of discovery and evolving regulatory requirements for over thirty years. One example is Canon's CaseData® web-based document review system, which offers SmartReview predictive coding. With this feature, CaseData can help organizations increase accuracy while reducing the time and expense of document review. SmartReview uses machine learning technology combined with expert reviewers. Canon Discovery Services is part of Canon Business Process Services, a leading provider of managed services and technology. Visit the Legal Services page of Canon's website to learn more.

[1] Lauri Donahue, "A Primer on Using Artificial Intelligence in the Legal Profession," http://jolt.law.harvard.edu/digest/a-primer-on-using-artificial-intelligence-in-the-legal-profession

[2] George Socha, "What Will AI Mean for You?" Vol. 101, No. 3 at 6, 8, Judicature, Autumn 2017, https://judicialstudies.duke.edu/wp-content/uploads/2017/09/Judicature-Fall2017-Socha.pdf

[3] Lauri Donahue, "A Primer on Using Artificial Intelligence in the Legal Profession"

[4] Warren E. Agin, "A Simple Guide to Machine Learning," *Business Law Today*, American Bar Association, Feb. 7, 2017, https://www.americanbar.org/publications/blt/2017/02/07_agin.html

[5] Lauri Donahue, "A Primer on Using Artificial Intelligence in the Legal Profession"

[6] Ibid

[7] Ibid

[8] Hope Reese, "Bias in Machine Learning, and How to Stop It," *Tech Republic*, Nov. 18, 2016, https://www.techrepublic.com/article/bias-in-machine-learning-and-how-to-stop-it/

[9] Jesse Emspak, "How a Machine Learns Prejudice," *Scientific American*, Dec. 29, 2016, https://www.scientificamerican.com/article/how-a-machine-learns-prejudice/

[10] Lawrence Hart, "What Data Will You Feed Your Artificial Intelligence?" CMS Wire, Feb. 28, 2018, https://www.cmswire.com/information-management/what-data-will-you-feed-your-artificial-intelligence/

[11] Ibid

[12] Joe Forward, "What Solo and Small Firms Should Know About Artificial Intelligence," State Bar of Wisconsin, *Inside Track*, Vol. 9, No. 3, Feb. 2017, https://www.wisbar.org/NewsPublications/InsideTrack/Pages/Article.aspx?ArticleID=25356&Issue=3&Volume=9