

Discovery Insights

10 questions for demystifying predictive coding



An interview with Jack Walker, Deloitte Advisory, Deloitte Transactions and Business Analytics LLP.

Many attorneys are testing the waters of analytics-based predictive coding, also referred to as technology-assisted review or document categorization. As with any new and potentially disruptive technology, predictive coding has its skeptics — perhaps more so than normal because of the potential stakes involved in the litigation it supports.

Here, Jack Walker answers the questions that many attorneys are asking about predictive coding, then offers his perspective based on recent client engagements and more than 20 years of experience in legal discovery.

Question	Jack Walker's take
Does predictive coding attempt to replace attorney review?	Not at all — attorneys are still very much involved. But it's no secret that discovery is often the most expensive part of litigation and document review is the most expensive part of discovery ¹ . If predictive coding can accelerate document review, at a fraction of the cost, and you can demonstrate statistically that it matches or possibly exceeds the quality of human review, why wouldn't you use it?
How are attorneys still involved in predictive coding?	The machine presents a series of documents to an attorney who makes a coding decision: for instance, "relevant" or "not relevant." This type of iterative attorney supervision enables continuous improvement of the predictive coding scores and results. Once the designated level of accuracy is achieved, the machine can score the remaining population.
Are off-the-shelf predictive coding products effective?	They can be. But more relevant is this question: which machine-learning approach is best suited to your current case and the unique characteristics of the documents in the case? For example, there are close to a dozen publicly available algorithms suitable for document categorization, each containing many customizable settings that can affect the accuracy of results. An off-the-shelf product typically uses one standard approach to machine learning without any customization capabilities. This is acceptable for some cases, but not for others, so using the same package for all cases can create risks. No single approach works best for all possible scenarios.
What alternative is there to off-the-shelf packages?	Datasets from different businesses require different machine-learning techniques. The complexity of the document language, along with other characteristics of the document population, determine the approach that should be used. You can't decide in advance which approach applies best to each situation, much less the fine-tuning of algorithms and other options and variables. Instead, qualified scientists and statisticians, working with attorneys and other specialists, can sample test data to determine an appropriate and defensible approach.
If predictive coding is challenged in court, how can we defend?	The approach described above involves a team of lawyers who are highly experienced in legal discovery, along with specialists in machine learning and statistics. We have also built a history of cases in which the approach has been used and are, therefore, able to continually enhance and improve our processes and technology.
Are there suggested protocols or best practices that would help us defend our processes?	Yes. There are several cases that outline a workflow that the parties used in their particular matter. In several instances, the courts accepted these protocols and therefore, can serve as a model to benchmark the processes and procedures for your case.

Question

Jack Walker's take

What does the case law say about predictive coding?

The case law is relatively new; however, it does provide insight into the issue. Judge Andrew Peck of the United States District Court for the Southern District of New York has stated that computer assisted review is now judicially approved for use in appropriate cases². Other courts have approved predictive coding for a party's own use and have asked the parties to cooperate to formulate a predictive coding protocol³.

How big does the training set of documents need to be to ensure a defensible result?

Many vendors suggest that 2,000 to 3,000 documents is an appropriate sample size, and yes, that sample size supports a typical process that many vendors follow. However, in most cases, data sets will be different from matter to matter—taking a “one size fits all” approach won't handle realities of any specific case, including the human learning that goes on over the course of a case or changes in case issues. Generally, larger sample sizes are associated with better classification, but another appropriate strategy may be start to with fewer documents and to anticipate iterations as the case develops. It's all about reviewing the right documents and anticipating risks where even a properly drawn sample may not yield results of sufficient accuracy.

How long does predictive coding take?

While traditional human document reviews can take many months to complete, a striking advantage to the predictive coding process is the small amount of time required to obtain results. The process of attorneys reviewing the training set of documents — the iterative process to improve results — and the scoring of a few million documents can typically be performed within a month.

Are particular types of datasets problematic for predictive coding?

Predictive coding in a vacuum may not be the most appropriate option for documents consisting largely of numeric data — spreadsheets, for example — image files, and short text, such as instant messages or certain social media messages. More text typically leads to greater accuracy in predictive coding. However, there are a wide variety of supplemental analytics that can be performed to accelerate review through these data sets, and predictive coding can inform those analytics.

My point of view: It's important to understand what predictive coding is — and isn't.

It's very scary for lawyers — frankly, for any professional — to load data into a black box and then have it spit out results you don't understand. Predictive coding does not have to be that way. Done correctly, lawyers are involved in various review and sampling processes, both in the initial phases of the predictive coding process, but also in later stages of evaluating the results and subsequent review decisions based on those results.

One of the greatest benefits of predictive coding is being able to place your most important discovery documents in the hands of the appropriate lawyers in the earliest stages of a case, enabling decisions that may inform litigation or settlement strategies before extensive document review, with its resulting costs, is performed.

You will still want to use many other technologies as part of the discovery process. They include such things as advanced search, near duplicate identification, email threading, social network analysis, and many others. Predictive coding is not a substitute for these technologies, but used correctly it can decrease the cost and time required for document review.

Bottom line, how would you like to be able to save your clients significant amounts of money while still producing superior results? Predictive coding performed effectively has the potential to do that.

¹ *Rand Institute for Civil Justice*, <http://www.rand.org/pubs/monographs/MG200/>

² *Da Silva Moore v. Publicis Groupe*, 2012 U.S. Dist. LEXIS 23350 (S.D.N.Y. Feb. 24, 2012).

³ *Global Aerospace Inc., et al. v. Lindow Aviation, L.P. dba Dulles Jet Center, et al*, Consolidated Case No. CL 00061040-00, (Circuit Court Loudoun Cty., Apr. 23, 2012); *In Re: Actos (Pioglitazone) Products Liability Litigation*, MDL No. 6:11-md-2299 (W.D.LA July 27, 2012); *Kleen Prods. LLC v. Packaging Corp. of Am.*, 2012 U.S. Dist. LEXIS 139632 (N.D. Ill. Sept. 28, 2012).

For more information, please contact:

John “Jack” Walker

Principal, Deloitte Advisory

Deloitte Transactions and Business Analytics LLP

johnwalker@deloitte.com

+1 312 486 3149

As used in this document, “Deloitte” means Deloitte & Touche LLP, which provides audit and enterprise risk services; Deloitte Financial Advisory Services LLP, which provides forensic, dispute, and other consulting services; and its affiliate, Deloitte Transactions and Business Analytics LLP, which provides a wide range of advisory and analytics services. Deloitte Transactions and Business Analytics LLP is not a certified public accounting firm. These entities are separate subsidiaries of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte LLP and its subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.

As a nonlegal service provider, Deloitte's services are provided under the direction and supervision of the client's legal counsel.

Copyright © 2015 Deloitte Development LLC. All rights reserved.

Member of Deloitte Touche Tohmatsu Limited