Predictive Coding: A Rose by any Other Name

by Sharon D. Nelson, Esq. and John W. Simek
© 2012 Sensei Enterprises, Inc.

Is there general agreement about what predictive coding is? No.

Is there general agreement about what to call it? No.

Is it the biggest and most talked about development in e-discovery today? Yes.

There is a great quote from a *Forbes* article which indicates why all lawyers need to understand a bit about predictive coding:

> "A lawsuit can really knock a company for a loop. Imagine being sued and asked to produce all responsive information, only to find that means sifting through 10 TB of emails. The process is complicated and it can be very costly. After all, the company must somehow determine with confidence whether each and every one of those emails is relevant to the lawsuit and/or subject to attorney-client privilege. This process has become much more manageable using technology to assist the review process."

Here is one definition of technology-assisted review offered by e-discovery expert Craig Ball: It is the use of more sophisticated algorithms – math – and advanced analytics to replace or supplement the individualized judgment of lawyers respecting the responsiveness, non-responsiveness and privilege of documents and data sets. We would add the critical human element - humans help machines to understand what documents are relevant to a case. With enough iterations of the sampling process, computers can learn enough so that they are fairly reliable in being able to judge the responsiveness of a document.

Why all the fuss over the name? Vendors are waging a very public and somewhat silly war about what to call predictive coding. We are using that term in our title (tipping our hat to Mr. Shakespeare) simply because the e-discovery firm Recommind was the first on the beach and that's what they called it. This is no doubt one reason why other vendors are taking pains to distance themselves from a competitor's term. Even the non-vendor experts don't agree on what to call it. Craig Ball likes "enhanced search." Ralph Losey likes "computer-assisted review." We've also seen technology-assisted review, automated document review, adaptive coding, predictive priority, meaning-based coding and many more.

Because Ralph and Craig are such wags, we thought readers would enjoy reading a portion of their e-mail answers to our "what do we call it?" query.

Ralph said, “I like CAR better, computer assisted review. This ties in well with a key metric of reviews, files per hour. How many files per hour can your CAR go? Also fits with my Hybrid model, where computers and human reviewers work together, as opposed to all manual review, or all computer review where humans don't double check the computer's prediction. So, I will not say Tar, I will say Car. What kind of search engine does your car have?”

Craig’s answer, which made us both laugh out loud, was “The folks at LegalTech gravitated to Technology Assisted Review (TAR) and seemed unreceptive to my suggestion of Super Human Information Technology. Wonder why? Must we have an acronym for everything? Are we at the point where you're not a real technology unless you have a catchy acronym? Nobody uses PC for predictive coding. I don't much care for CAR or TAR, as they are both FAR below PAR and make me want to go to a BAR.

I'd probably be content with ES for enhanced search.”

When we talked to Eric Seggebruch of Recommind, he said that “Predictive Coding” was being used and at times misused. He also stated what we have heard from others - that some companies say they do predictive coding when in fact they are just using automated review or perhaps just predicting a percentage of relevant documents without identifying which ones are relevant. In true predictive coding, he says the machines will tell you what documents you should be looking at next, identifying which ones are the most relevant and important.

Some of Eric’s thoughts, responding in part to Ralph and Craig, are below:

“The correct analysis here is to differentiate (borrowing from Ralph) what engine is inside your CAR (Eric is not a fan of TAR, is it sounds too much like being “stuck” as in “stuck in the TAR”). Analytics tools that use word frequency analysis (for example, word counts) rather than more sophisticated engines such as PLSA (Probabilistic Latent Semantic Analysis) are causing people to incorrectly put all these dissimilar approaches in the same category of TAR or CAR. It is going to take the market some time to sift through the noise and differentiate those vendors and those technologies that make a measurable difference from those merely ‘Buzzword Compliant’.

In my view, "Predictive Coding” properly implemented (or CSR “Computer Suggested Review” for those vendors not wanting to send folks to http://www.predictivecoding.com) will be the clear winner in the TAR, CAR (and in deference to Craig) BAR wars (I have recently taken over some projects where it looks like the coding was done in a BAR). There are measurable differences between these approaches and anyone evaluating the technologies available needs to understand this:

What exactly is Predictive Coding?

A process employing people, workflow and technology - MUST have all three.

Predictive Coding requires ALL of the following – anything else is simply “Technology Assisted Review” - TAR or CAR

- Integrated, keyword-agnostic analytics to quickly generate accurate seed sets
- Language and keyword-agnostic machine-learning technology to accurately find relevant documents during the “training” process
- A sound and well-documented workflow
- Integrated sampling to verify results to a statistical certainty before, during and after review
- A completely integrated, purpose-built system to ensure results are consistent throughout the entire process, every time

What is “Technology-Assisted Review”? Or CAR?

- Anything using advanced technology that does NOT meet the above predictive coding criteria.”

We have no doubt that some vendors might take issue with Eric’s definitions, but that is part of the current hurly burly surrounding this new technology – there is little consensus about definitions.

Eric related that in a case involving 1,000,000 documents approximately 15% might end up being reviewed, with 850,000 not being reviewed. The workflow is to use the documents that you identify as responsive to “find more like this”. This workflow is repeated for multiple iterations until the system, while still returning documents, no longer returns a significant percentage of relevant documents. Once the review is complete, the documents not reviewed by humans are sampled again and the response rate for the non-reviewed documents is compared to the response rate from a random sample (it should be less). As Eric said wryly, “It’s pretty much wash, rinse, repeat until the engine is no longer suggesting documents like the ones you care about."

Craig is reminded of the ECA (early case assessment) hoopla and says that predictive coding is being oversold and overheated by marketers selling something they can dress up like something more than a keyword search. As he notes, many of these technologies don’t assess the meaning of a document as a human would. Some only look at the frequency and geometric juxtaposition of words in a way that might be described as “I don’t know what it says, but it uses the same sorts of words with about the same incidence and arrangement, so it’s likely to be saying much the same thing.”

In our conversations with other experts, we all agree that some sort of technology-assisted review is here to stay – it helps trim time and e-discovery costs – and is probably more accurate than human reviewers who get tired or have headaches. This doesn't take humans out of the equation since senior lawyers are critical to the process of teaching the machines, but it certainly means less human hours overall, something which contract attorneys who do document review have bemoaned.

How much money can it save? According to a survey by the Electronic Discovery Institute, it can save an average of 45% with some respondents reporting savings of up to 70%. How much does it cost? Ever tried nailing Jell-O to a tree? You will feel precisely like you are doing that when you attempt to question vendors about specific costs. Everyone seems to agree that it is an appropriate solution for large volume cases, which means we know it is expensive. There is a great disagreement about whether it is appropriate for smaller cases – our own suspicion is that it is not appropriate for garden variety cases – but once again, it depends on the definition being used by vendors – you may be getting something called technology-assisted review which is a far cry from the technology Craig, Ralph and Eric have described.

Recently, the e-discovery world was rocked by an order issued by Magistrate Judge Andrew J. Peck in the *Da Silva Moore et al v. Publicis Groupe & MSL Groupe* employment discrimination case in the Southern District of N.Y. Judge Peck has been a vocal advocate of predictive coding and actually wrote an article about it called "Search, Forward" for *Law Technology News*.

He quoted from his own article in his February 2012 opinion, where he not only ordered the use of predictive coding, but seemingly endorsed it, noting that it may save costs, stating that it was superior to available alternatives such as keyword or manual searches, pronouncing it better than "unattainable perfection" and saying that he was less interested in the science of the vendor's software (in this case Recommind) than in "whether it produced responsive documents with reasonably high recall and high precision." He did note that technology alone would not suffice, commenting that it was pivotal to design an appropriate process and to have quality control testing.

One line in the opinion certainly caught everyone's eye: "Computer-assisted review now can be considered judicially-approved for use in appropriate cases." The issuance of this opinion elicited an online round of high-fiving among predictive coding vendors that was unprecedented, although they seemed to miss the fact that Judge Peck used the CAR word.

The high-fiving grew silent a few weeks later. On March 13th, U.S. District Judge Andrew L. Carter granted the Plaintiffs' request to file a brief documenting their objections to Judge Peck's rulings. The plaintiffs objected to the fact that Judge Peck relied on outside sources not in evidence and that the judge had ordered the use of this new predictive coding tool as the sole determinate of what e-discovery should be produced. They protested the judge's wholesale adoption of the defendants' e-discovery protocol and said Recommind's software has no generally accepted reliability standards. Further, they said the use of predictive coding was inappropriate where the vast majority of the relevant evidence was held by the defendant-employer.

The Objection to Judge Peck's order was filed under Rule 72 of the Federal Rules of Civil Procure which allows a district judge to "modify or set aside any part of the order that is clearly erroneous or contrary to law."

Judge Peck appears to have foreseen possible objections to his opinion and proactively attempted to fend some of them off. He specifically concluded that:

1. The technology was being employed effectively under the circumstances
2. Senior lawyers were making sampling decisions
3. The training documents made the process transparent
4. Some non-corresponding "comparator" e-mail sets were excluded from the coding; and
5. Documents which were found to be irrelevant would be sampled later in the process to determine the effectiveness of the training.

Although the authors believe that technology-assisted review will march on, right now its judicial status appears to be somewhat uncertain. We also believe that all predictive coding software needs to be independently peer-reviewed. Right now, most experts believe that there are at least some examples of technology-assisted review software that would not survive a Daubert challenge.

We write this column in March 2012, so we are uncertain if there will be updates to this case by the time it is published. If there are, we'll try to update the online version of this article.

In the meantime, like everyone else in the e-discovery world, we are waiting for the next act of technology-assisted review (or whatever you call it) to unfold.

Thanks to Craig, Ralph and Eric for their substantial contributions to this important topic.

The authors are the President and Vice President of Sensei Enterprises, Inc., a legal technology, information security and computer forensics firm based in Fairfax, VA. 703-359-0700 (phone) www.senseient.com