

WHY I LOVE PREDICTIVE CODING

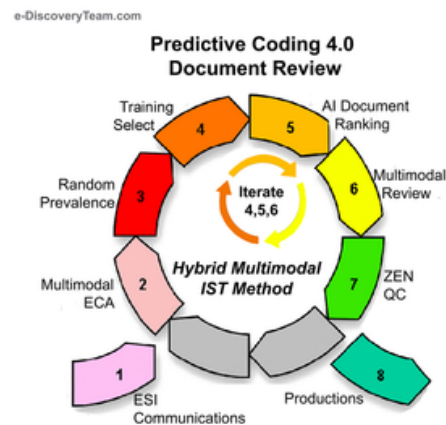
Making document review fun again with Mr. EDR and Predictive Coding 4.0

Ralph Losey^{*}
e-Discovery Team

Many lawyers and technologists like predictive coding and recommend it to their colleagues. They have good reasons. It has worked for them. It has allowed them to do e-discovery reviews in an effective, cost efficient manner, especially the big projects. That is true for me too, but that is not why I *love* predictive coding. My feelings come from the excitement, fun, and amazement that often arise from seeing it in action, first hand. I love watching the predictive coding features in my software find documents that I could never have found on my own. I love the way the AI in the software helps me to do the impossible. I really love how it makes me far smarter and skilled than I really am.

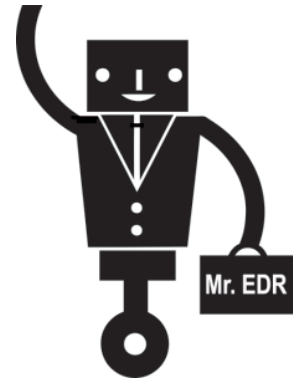


I have been getting those kinds of positive feelings consistently by using the latest Predictive Coding 4.0 methodology (shown right) and KrollDiscovery's latest [eDiscovery.com Review](http://eDiscovery.com) software ("EDR"). So too have my e-Discovery Team members who helped me to participate in TREC 2015 and 2016 (the great science experiment for the latest text search techniques sponsored by the [National Institute of Standards and Technology](http://NationalInstituteofStandardsandTechnology.gov)). During our grueling forty-five days of experiments in 2015, and again for sixty days in 2016, we came to admire the intelligence of the new EDR software so much that we decided to personalize the AI as a robot. We named him Mr. EDR out of respect. He even has his own website now, MrEDR.com, where he explains how he helped my e-Discovery Team in the 2015 and 2015 TREC Total Recall Track experiments.



^{*}This is an edited reprint of the author's personal blog, e-discoveryteam.com, and contains his personal opinions and not those of his law firm or its clients. Copyright Ralph Losey 2015, 2017. Reference to any products should not be construed as a commercial endorsement.

Bottom line for us from this research was to prove and improve our methods. Our latest version 4.0 of Predictive Coding, Hybrid Multimodal IST Method is the result. We have even open-sourced this method, well most of it, and teach it in a free seventeen-class online program: [TARcourse.com](https://www.tarcourse.com). Aside from testing and improving our methods, another, perhaps even more important result of TREC for us was our rediscovery that with good teamwork, and good software like Mr. EDR at your side, **document review need never be boring again**. The documents themselves may well be boring as hell, that's another matter, but the search for them need not be.



How and Why Predictive Coding is Fun

Steps Four, Five and Six of the standard eight-step workflow for Predictive Coding 4.0 is where we work with the active machine-learning features of Mr. EDR. These are its predictive coding features, a type of artificial intelligence. We train the computer on our conception of relevance by showing it relevant and irrelevant documents that we have found. The software is designed to then go out and find all other relevant documents in the total dataset. One of the skills we learn is when we have taught enough and can stop the training and complete the document review. At TREC we call that the Stop decision. It is important to keep down the costs of document review.

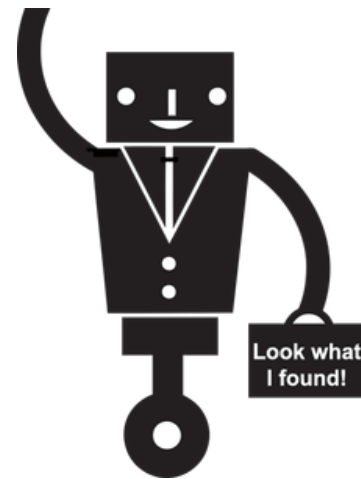


We use a multimodal approach to find training documents, meaning we use all of the other search features of Mr. EDR to find relevant ESI, such as keyword searches, similarity and concept. We iterate the training by sample documents, both relevant and irrelevant, until the computer starts to understand the scope of relevance we have in mind. It is a training exercise to make our AI smart, to get it to understand the basic ideas of relevance for that case. It usually takes multiple rounds of training for Mr. EDR to understand what we have in mind. But he is a fast learner, and by using the latest hybrid multimodal IST ("intelligently spaced learning") techniques, we can usually complete his training in a few days. At TREC, where we were moving fast after hours with the A-Team, we completed some of the training experiments in just a few hours.

After a while Mr. EDR starts to "get it," he starts to really understand what we are after, what we think is relevant in the case. That is when a happy shock and awe type moment can happen. That is when Mr. EDR's intelligence and search abilities start to exceed our own. Yes. It happens. The pupil then starts to evolve beyond his teachers. The smart algorithms start to see patterns and find evidence invisible to us. At that point we sometimes even let him train himself by automatically accepting his top-ranked predicted relevant documents without even looking at them. Our main role then is to determine a good range for the automatic acceptance and do some spot-checking. We are, in effect, allowing Mr. EDR to take over the review. Oh what a feeling to then watch what happens, to see him keep finding new relevant documents and keep

getting smarter and smarter by his own self-programming. That is the special AI-high that makes it so much fun to work with Predictive Coding 4.0 and Mr. EDR.

It does not happen in every project, but with the new Predictive Coding 4.0 methods and the latest Mr. EDR, we are seeing this kind of transformation happen more and more often. It is a tipping point in the review when we see Mr. EDR go beyond us. He starts to unearth relevant documents that my team would never even have thought to look for. The relevant documents he finds are sometimes completely dissimilar to any others we found before. They do not have the same keywords, or even the same known concepts. Still, Mr. EDR sees patterns in these documents that we do not. He can find the hidden gems of relevance, even outliers and black swans, if they exist. When he starts to train himself, that is the point in the review when we think of Mr. EDR as going into superhero mode. At least, that is the way my young e-Discovery Team members likes to talk about him.



By the end of many projects the algorithmic functions of Mr. EDR have attained a higher intelligence and skill level than our own (at least on the task of finding the relevant evidence in the document collection). He is always lightning fast and inexhaustible, even untrained, but by the end of his training, he becomes a search genius. Watching Mr. EDR in that kind of superhero mode is what makes Predictive Coding 4.0 a pleasure.

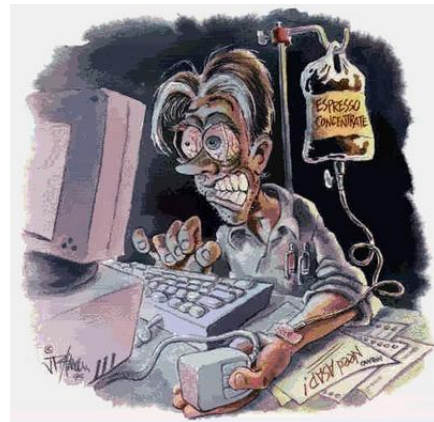
The Empowerment of AI Augmented Search

It is hard to describe the combination of pride and excitement you feel when Mr. EDR, your student, takes your training and then goes beyond you. More than that, the super-AI you created then **empowers you** to do things that would have been impossible before, absurd even. That feels pretty good too. You may not be Iron Man, or look like Robert Downey, but you will be capable of remarkable feats of legal search strength.



For instance, using Mr. EDR as our Iron Man-like suits, my e-discovery ã-Team of three attorneys was able to do thirty different review projects and classify 17,014,085 documents in 45 days. See 2015 TREC experiment summary at [Mr. EDR](#). We did these projects mostly at nights, and on weekends, while holding down our regular jobs. What makes this crazy impossible, is that we were able to accomplish this by only personally reviewing 32,916 documents. That is less than 0.2% of the total collection. That means we relied on predictive coding to do 99.8% of our review work. Incredible, but true.

Using traditional linear review methods it would have taken us 45 years to review that many documents! Instead, we did it in 45 days. Plus our recall and precision rates were insanely good. We even scored 100% precision and 100% recall in one TREC project in 2015 and two more in 2016. You read that right. Perfection. Many of our other projects attained scores in the high and mid nineties. We are not saying you will get results like that. Every project is different, and some are much more difficult than others. But we are saying that this kind of AI-enhanced review is not only fast and efficient, it is effective.

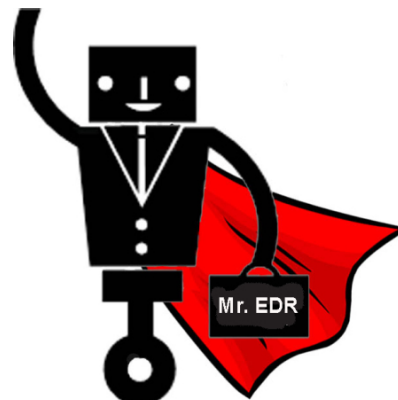


Yes, it's pretty cool when your little AI creation does all the work for you and makes you look good. Still, no robot could do this without your training and supervision. We are a team, which is why we call it hybrid multimodal, man and machine.

Having Fun with Scientific Research at TREC 2015 and 2016

During the 2015 TREC Total Recall Track experiments my team would sometimes get totally lost on a few of the really hard Topics. We were not given legal issues to search, as usual. They were arcane technical hacker issues, political issues, or local news stories. Not only were we in new fields, the scope of relevance of the thirty Topics was never really explained. (We were given one to three word explanations in 2015, in 2016 we got a whole sentence!) We had to figure out intended relevance during the project based on feedback from the automated TREC document adjudication system. We would have some limited understanding of relevance based on our suppositions of the initial keyword hints, and so we could begin to train Mr. EDR with that. But, in several Topics, we never had any real understanding of exactly what TREC thought was relevant.

This was a very frustrating situation at first, but, and here is the cool thing, even though we did not know, Mr. EDR knew. That's right. He saw the TREC patterns of relevance hidden to us mere mortals. In many of the thirty Topics we would just sit back and let him do all of the driving, like a Google car. We would often just cheer him on (and each other) as the TREC systems kept saying Mr. EDR was right, the documents he selected were relevant. The truth is, during much of the 45 days of TREC we were like kids in a candy store having a great



time. That is when we decided to give Mr. EDR a cape and superhero status. He never let us down. It is a great feeling to create an AI with greater intelligence than your own and then see it augment and improve your legal work. It is truly a hybrid human-machine partnership at its best.

I hope you get the opportunity to experience this for yourself someday. The TREC experiments in 2015 and 2016 on recall in predictive coding are over, but the search for truth and justice goes on in lawsuits across the country. Try it on your next document review project.

Do What You Love and Love What You Do

Mr. EDR, and other good predictive coding software like it, can augment our own abilities and make us incredibly productive. This is why I love predictive coding and would not trade it for any other legal activity I have ever done (although I have had similar highs from oral arguments that went great, or the rush that comes from winning a big case).

The excitement of predictive coding comes through clearly when Mr. EDR is fully trained and able to carry on without you. It is a kind of Kurzweilian [mini-singularity](#) event. It usually happens near the end of the project, but can happen earlier when your computer catches on to what you want and starts to find the hidden gems you missed. I suggest you give Predictive Coding 4.0 and Mr. EDR a try. To make it easier I open-sourced our latest method and created an online course. TARcourse.com. It will teach anyone our method, if they have the right software. Learn the method, get the software and then you too can have fun with evidence search. You too can love what you do. ***Document review need never be boring again.***



Caution

One note of caution: most e-discovery vendors, including the largest, do not have active machine learning features built into their document review software. Even the few that have active machine learning do not necessarily follow the Hybrid Multimodal IST Predictive Coding 4.0 approach that we used to attain these results. They instead rely entirely on machine-selected documents for training, or even worse, rely entirely on random selected documents to train the software, or have elaborate unnecessary secret control sets.

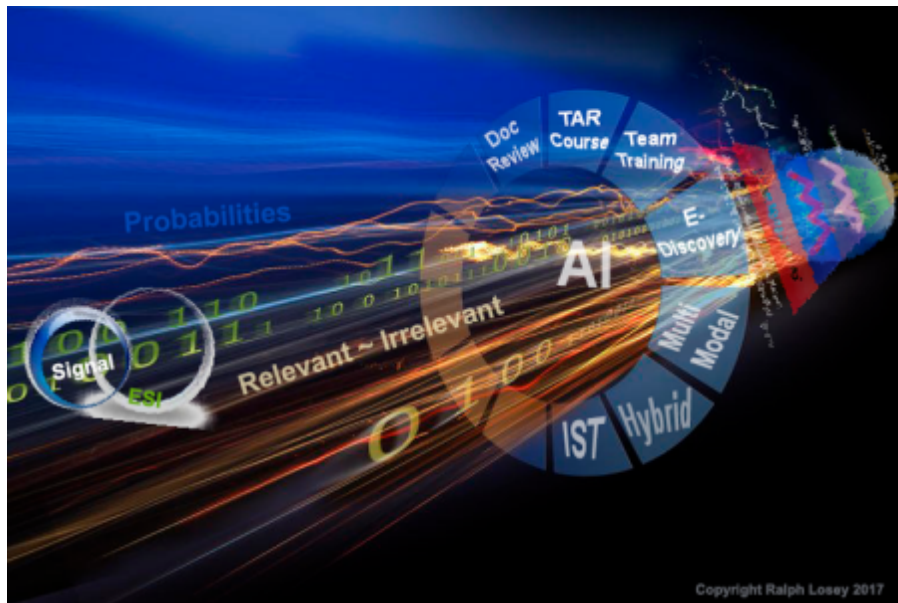


The algorithms used by some vendors who say they have "predictive coding" or "artificial intelligence" are not very good. Scientists tell me that some are only dressed-up concept search

or unsupervised document clustering. Only bona fide active machine learning algorithms create the kind of AI experience that I am talking about. Software for document review that does not have any active machine learning features may be cheap, and may be popular, but they lack the power that I love. Without active machine learning, which is fundamentally different from just "analytics," it is not possible to boost your intelligence with AI. So beware of software that just says it has advanced analytics. Ask if it has "active machine learning"?

It is impossible to do the things described in this essay unless the software you are using has active machine learning features. This is clearly the way of the future. It is what makes document review enjoyable and why I love to do big projects. ***It turns scary to fun.***

So, if you tried "predictive coding" or "advanced analytics" before, and it did not work for you, it could well be the software's fault, not yours. Or it could be the poor method you were following. The method that we developed in *Da Silva Moore*, where my firm represented the defense, was a version 1.0 method. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 183 (S.D.N.Y. 2012). We have come a long way since then. We have eliminated unnecessary random control sets and gone to continuous training, instead of train then review. This is spelled out in the TARcourse.com that teaches our latest version 4.0 techniques.

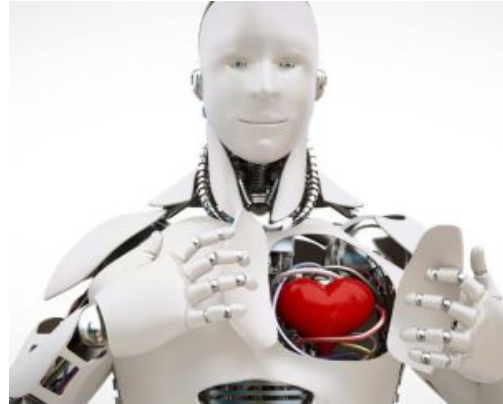


The new 4.0 methods are not hard to follow. The TARcourse.com puts our methods online and even teaches the theory and practice. And the 4.0 methods certainly will work. We have proven that at TREC, but only if you have good software. With just a little training, and some help at first from consultants (most vendors with bona fide active machine learning features will have good ones to help), you can have the kind of success and excitement that I am talking about.

Do not give up if it does not work for you the first time, especially in a complex project. Try another vendor instead, one that may have better software and better consultants. Also, be sure that your consultants are Predictive Coding 4.0 experts, and that you follow their advice. Finally, remember that the cheapest software is almost never the best, and, in the long run will cost you a small fortune in wasted time and frustration.

Conclusion

Love what you do. It is a great feeling and sure fire way to job satisfaction and success. With these new predictive coding technologies it is easier than ever to love e-discovery. Try them out. Treat yourself to the AI high that comes from using smart machine learning software and fast computers. There is nothing else like it. If you switch to the 4.0 methods and software, you too can know that thrill. You can watch an advanced intelligence, which you helped create, exceed your own abilities, exceed anyone's abilities. You can sit back and watch Mr. EDR complete your search for you. You can watch him do so in record time and with record results. It is amazing to see good software find documents that you know you would never have found on your own.



Predictive coding AI in superhero mode can be exciting to watch. Why deprive yourself of that? Who says document review has to be slow and boring? ***Start making the practice of law fun again.***

The author can be reached at Ralph.Losey@gmail.com or at work at Ralph.Losey@JacksonLewis.com. Consultations by the author related to predictive coding, e-discovery or any other for-pay services are provided exclusively to current clients of the author's law firm, Jackson Lewis P.C.