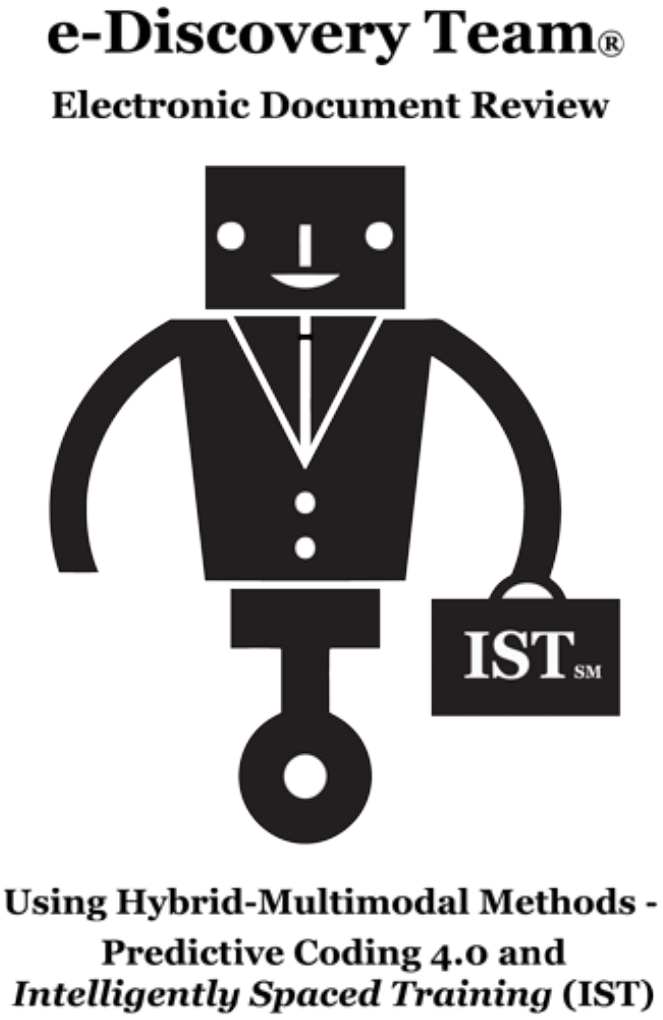
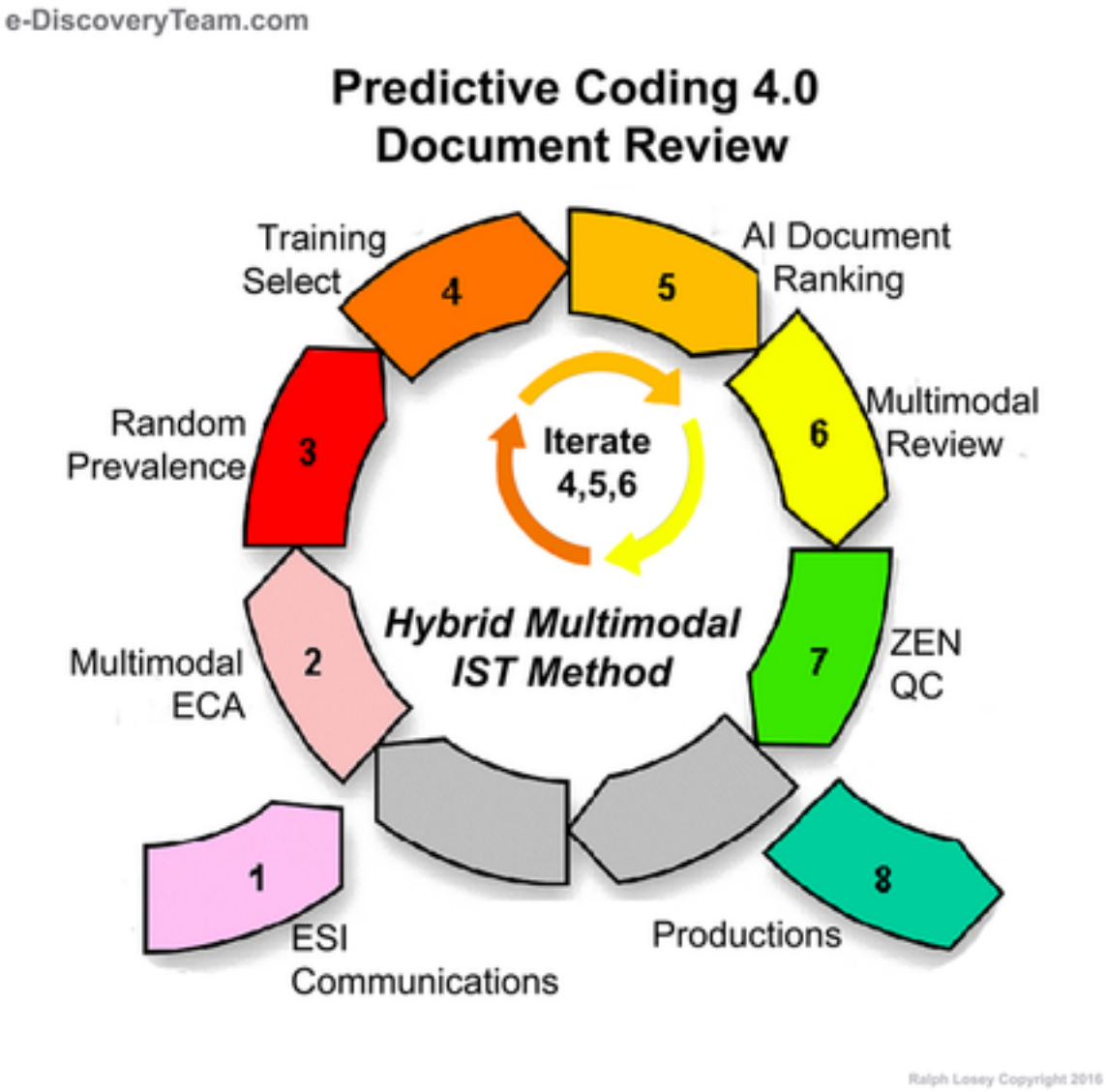
e-Discovery Team's Predictive Coding 4.0
Method of Electronic Document Review

Ralph Losey
Jackson Lewis, P.C.



This article describes the latest insights and methods of document review using predictive coding. It was originally published by the e-Discovery Team (November 2016).



We call this ***Predictive Coding 4.0*** because it substantially improves upon, and replaces the methods and insights we announced in our October 2015 publication - Predictive Coding 3.0. In that two-part blog we explained the history of predictive coding software and methods in legal review, including versions 1.0 and 2.0. Then we described our new version 3.0 in some detail. Since that publication we have developed more enhancements to our methods, including many new, innovative uses of the predictive coding features of Kroll Ontrack's EDR software. We even developed some new features not related to predictive coding. (Try out the new *Folder Similar* search in EDR for example.) Most of our new insights, just like our prior 3.0 version methodologies, can also be used on other software platforms. To use all of the features, however, the software will have to have bona fide *active machine learning* capacities. Most do not. More on that later.

These improvements naturally evolved to a certain degree as part of the e-Discovery Team members normal work supervising hundreds, maybe even thousands of document review projects over the past year. But the new insights that require us to make a complete restatement, a new *Version 4.0*, arose just recently. Major advances were attained as part of an intensive three months of experiments, all conducted outside of our usual legal practice and document reviews. The e-Discovery Team doing this basic research consisted of myself and several of Kroll Ontrack's top document review specialists, including especially Jim Sullivan and Tony Reichenberger. They have now fully mastered the e-discovery team search and review Hybrid Multimodal methodologies. As far as I can see, at this point in the race for the *highest quality legal document review*, no one else comes even close to their skill level. Yes, e-discovery is highly competitive, but they trained hard and are now looking back and smiling.

The insights we gained, and the skills we honed, including speed, did not come easily. It took full time work on client projects all year, plus three full months of research, often in lieu of real summer vacations (my wife is still waiting). This is hard work, but we love it. *See:* Why I Love Predictive Coding. This kind of dedication of time and resources by an e-discovery vendor or law firm is unprecedented. There is a cost to attain the research *benefits* realized, both hard out-of-pocket costs and lost time. So I hope you understand that we are only going to share some of our techniques. The rest we will keep as trade-secrets. (Retain us and watch. Then you can see them in action.)

Kroll Ontrack understands the importance of pure research and enthusiastically approved these expenditures. (My thanks again to CEO Mark Williams, a true visionary leader in this industry who approved and supported the research program.) I suggest you ask your vendor, or law firm, how much time they spent last year researching and experimenting with document review methods? As far as we know, the only other vendor with an active research program is Catalyst, whose work is also to be commended. (No one else showed up for TREC.) The only other law firm we know of is Maura Grossman's new solo practice. Her time spent with research is also impressive.

Mark Williams, CEO Kroll Ontrack

The results we attained certainly make this investment worthwhile, even if many in the profession do not realize it, much less appreciate it. They will in time, so will the consumers. This is a long-term investment. Pure research is necessary for any technology company, including all companies in the e-Discovery field. The same holds true, albeit to a lesser extent, to any law firm claiming to have technological superiority.

Experience from handling live projects alone is too slow an incubator for the kind of AI breakthrough technologies we are now using. It is also too inhibiting. You do not experiment on important client data or review projects. Any expert will improvise somewhat during such projects to match the circumstances, and sometimes do *post hoc* analysis. But such work on client projects alone is not enough. Pure research is needed to continue to advance in AI-enhanced review. That is why the e-Discovery Team spent a substantial part of our waking hours in June, July and August 2016 working on experiments with Jeb Bush email. The Jeb Bush email collection was our primary laboratory this year. As a result of the many new things we learned, and new methods practiced and perfected, we have now reached a point where a complete restatement of our method is in order. Thus we here release ***Predictive Coding 4.0.***

Our latest breakthroughs this summer primarily came out of the e-Discovery Team's participation in the annual Text Retrieval Conference, aka *TREC*, sponsored by the National Institute of Standards and Technology. This is the 25th year of the TREC event. We were honored to again participate, as we did last year, in the Total Recall Track of TREC. This is the closest Track that TREC now offers to a real legal review

project. It is not a Legal Track, however, and so we necessarily did our own side-experiments, and had our own unique approach different from the Universities that participated. The TREC leadership of the Total Recall Track was once again in the capable hands of Maura Grossman, Gordon Cormack and other scientists.

This article will not report on the specifics of the 2016 Total Recall Track. That will come at a later time after we finish analyzing the enormous amount of data we generated and submit our formal reports to TREC. In any event, the TREC related work we did this Summer went beyond the thirty-four research topics included in the TREC event. It went well beyond the **9,863,366 documents** we reviewed with Mr. EDR's help as part of the formal submittals. Countless more documents were reviewed for relevance if you include our side-experiments.

At the same time that we did the formal tests specified by the Total Recall Track we did multiple side-experiments of our own. Some of these tests are still ongoing. We did so to investigate our own questions that are unique to legal search and thus beyond the scope of the Total Recall Track. We also performed experiments to test unique attributes of Kroll Ontrack's EDR software. It uses a proprietary type of logistic regression algorithm that was awarded a patent this year. Way to go KO and Mr. EDR!

Although this article will not report on our TREC experiments *per se*, we will share the bottom line, the *take-aways* of this testing. Not everything will be revealed. We keep some of our methods and techniques trade secret.

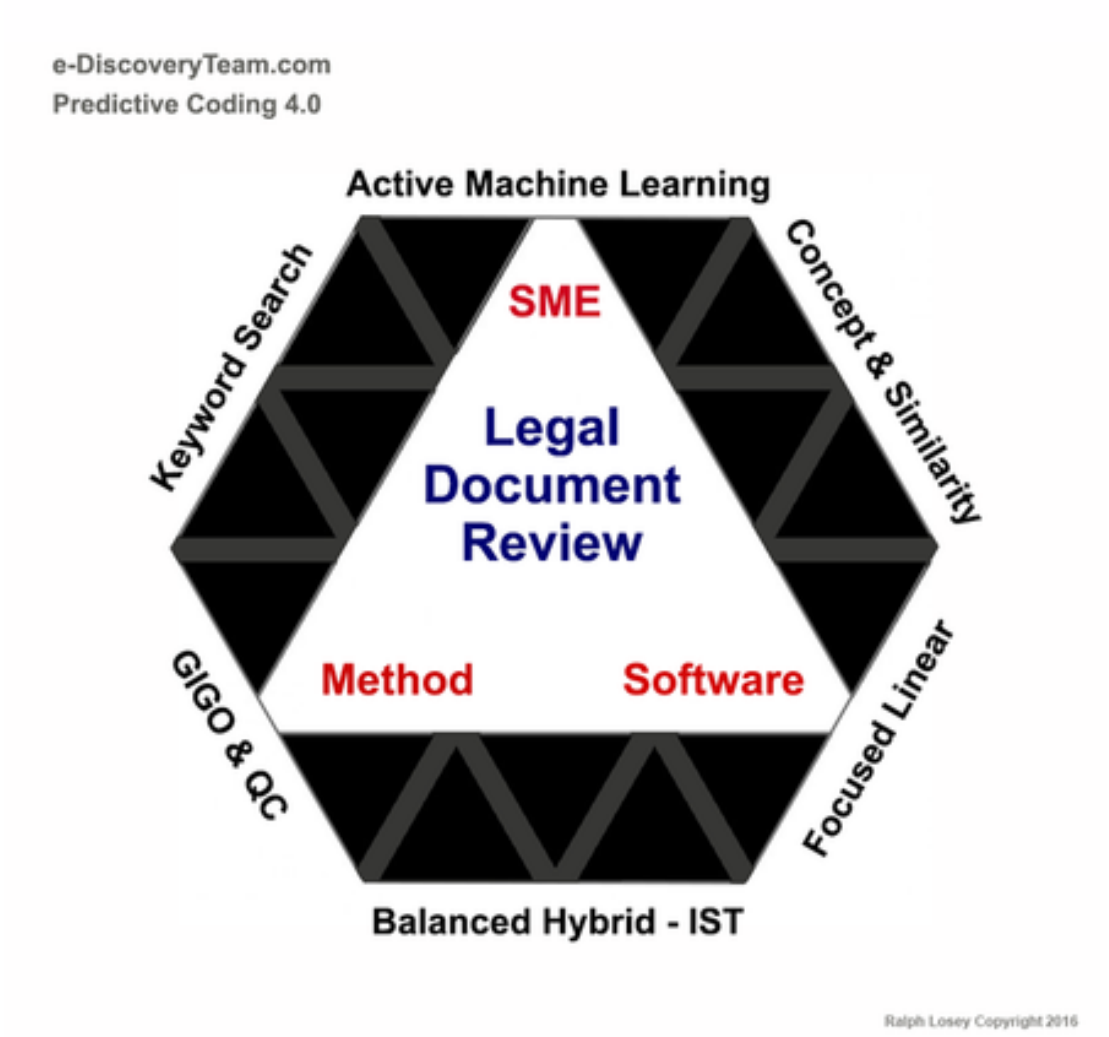
We will also not be discussing in this article our future plans and spin-off projects. Let's just say for now that we have several in mind. One in particular will, I think, be very exciting for all attorneys and paralegals who do document review. Maybe even *fun* for those of you who, like us, are *really into* and enjoy a good computer search. You know who you are! If my recommendations are accepted, we will open that one up to all of our fellow doc-review freaks. I will say no more at this point, but watch for announcements in the coming year from Kroll Ontrack and me. We are having too much fun here not to share some of the good times.

Even if we did adopt 100% transparency on our methods, it would take a book to write it all down, and it would still be incomplete. Many things can only be learned by doing, especially methods. Document review is, after all, a part of

legal *practice*. As the scientists like to put it, legal search is essentially *ad hoc*. It changes and is customized to fit the particular evidence search assignments at hand. But we will try to share all of the basic *insights*. They have all been discussed here before. The new insights we gained are more like a deepening understanding and matter of emphasis. They are refinements, not radical departures, although some are surprising.

Nine Insights Concerning the Use of Predictive Coding in Legal Document Review

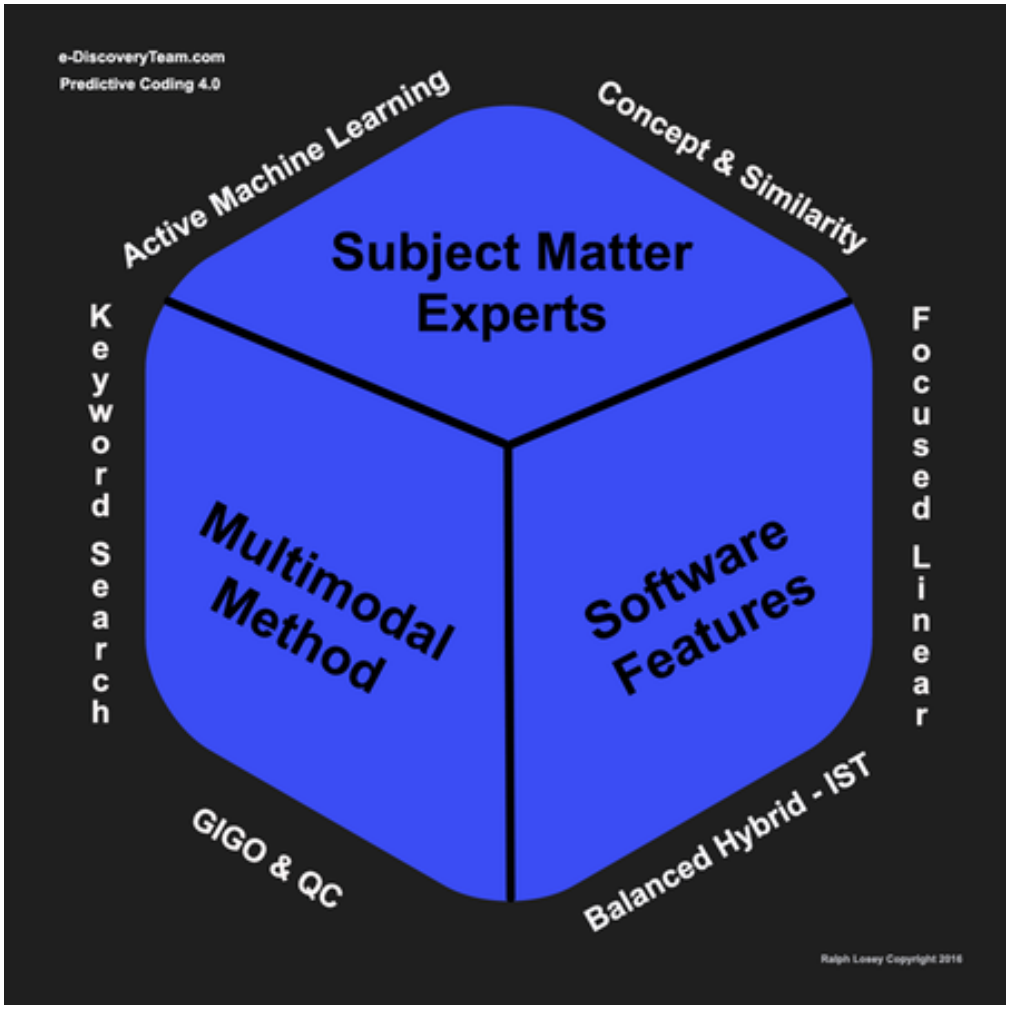
The diagram below summarizes the nine basic insights that have come out of our work this year. These are the key concepts that we now think are important to understand and implement



The diagrams above and following will be explained in detail throughout the rest of this multipart blog, as will the restated 8-Step Workflow shown at the top of the page. These are not new concepts. I have discussed most of these here before.

I am confident that all readers will be able to follow along as I set forth the new nuances we learned.

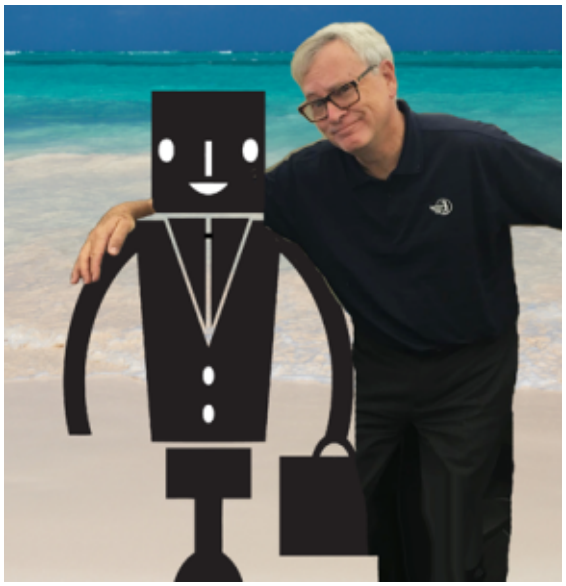
Although these concepts are all familiar, some of our deepened understanding of these concepts may surprise you. Some were **surprising to us.** These insights include several changes in thinking on our part. Some of the research results we saw were unexpected. But we follow the data. Our opinions are always held lightly. I have argued both sides of a legal issue too many times as a lawyer to fall into that trap. Our thinking follows the evidence, not our preconceptions. That is, after all, the whole point of research. Schedule permitting, we are also happy to provide in-person or online presentations that explain these concept-summary diagrams. If retained, you can also see it in action.



Although these insights and experiments were derived using Kroll Ontrack EDR software, they are essentially *vendor neutral*. The methods will work on any full-featured document review platform, but especially those that includes *bona fide* active machine learning abilities, aka *Predictive Coding*. As all experts in this field know, many of the most popular document review platforms do not have

these features, even those stating they use *Analytics*. *Active Machine Learning* is very different, and far more advanced than *Analytics*, the early forms of which were called *Concept Search*. This type of machine learning is passive and clearly is not predictive coding. It has its place in any multimodal system such as ours, and can be a powerful feature to improve search and review. But such software is incomplete and cannot meet the standards and capability of software that includes active machine learning. Only full featured document review platforms with active machine learning abilities can use all of the Predictive Coding 4.0 methods described here.

Sorry dear start-up vendors, and others, but that's the truth. Consumers, you get what you pay for. You know that. Not sure? Get the help of an independent expert advisor before you make substantial investments in e-discovery software or choose a vendor for a major project. Also, if you have tried predictive coding, or what you were told was advanced TAR, *whatever the hell that is*, and it did not work well, do not blame yourself. It could be the software. Or if not the software, then the antiquated version 1.0 or 2.0 methods used. There is a lot of bullshit out there. Excuse my French. There always has been when it comes to new technology. It does, however, seems especially prevalent in the legal technology field. Perhaps they think we lawyers are naive and technologically gullible. Do not be fooled. Again, look to an independent consultant if you get confused by all of the vendor claims.



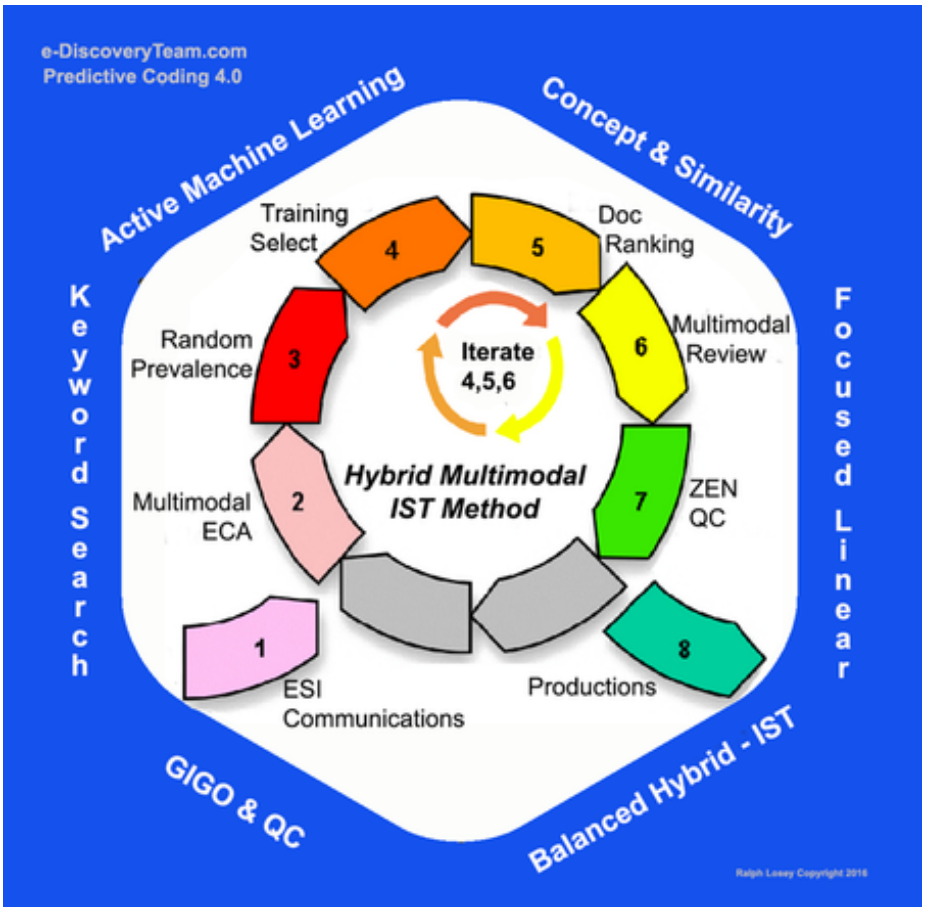
Contrary to what some vendors will tell you (typically the ones without bona fide predictive coding features), predictive coding 3.0, and now 4.0, methods are not rocket science. You do not have to be a TAR-whisperer or do nothing but search, like my A-team for TREC. With good software it is not really that hard at all. These methods do, however, require an attorney knowledgeable in e-discovery and comfortable with software. This is not for novices. But every law firm should *anyway* have attorneys with special training and experience in technology and e-discovery. For instance, if you practice in the Northern District of California, an e-discovery liaison with such expertise is required in *most cases*. *See* Guidelines for the Discovery of Electronically Stored Information. Almost half of the Bar Associations in the U.S. require basic technology competence as an ethical imperative. *See eg*. ABA Model Rule 1.1, Comment [8] and Robert Ambrogi's list of 23 states, and counting, that now require such competence. (My own law firm has had an e-discovery liaison program in place since 2010, which I lead and train. I am proud to say that after six years of work it is now a great success.) So

no, you do not have to be a full-time specialist, like the members of my TREC e-Discovery team, to successfully use AI-enhanced review, which we call *Hybrid Multimodal*. This is especially true when you work with vendors like Kroll Ontrack, Catalyst and others that have teams of special consultants to guide you. You just have to pick your vendors wisely.

PART TWO

In Part One we explained the background that led to the 4.0 upgrade: the TREC research and hundreds of projects we have done since our last upgrade a year ago. Millions have been spent to develop the software and methods we now use for Technology Assisted Review (TAR). As a result our TAR methods are more effective and simpler than ever.

The nine insights we will share are based on our experience and research. Some of our insights may be complicated, especially our lead insight on *Active Machine Learning* covered in this Part Two with our new description of ***IST*** - *Intelligently Spaced Training*. We consider IST the smart, human empowering alternative to CAL. If I am able to write these insights up here correctly, the obviousness of them should come through. They are all simple in essence. The insights and methods of Predictive Coding 4.0 document review are partially summarized in the chart below.



1st of the Nine Insights: *Active Machine Learning*

Our method is *Multimodal* in that it uses all kinds of document search tools. Although we emphasize *active machine learning*, we do not rely on that method alone. Our method is also *Hybrid* in that we use both machine judgments and human (lawyer) judgments. Moreover, in our method the lawyer is always in charge. We may take our hand off the wheel and let the machine drive for a while, but under our versions of Predictive Coding, we watch carefully. We remain ready to take over at a moment's notice. We do not rely on one brain to the exclusion of another. *See eg*. *Why the 'Google Car' Has No Place in Legal Search* (caution against over reliance on fully automated methods of *active machine learning*). Of course the converse is also true, we never just rely on our human brain alone. It has too many limitations. We enhance our brain with predictive coding algorithms. We add to our own natural intelligence with artificial intelligence. The perfect balance between the two, the *Balanced Hybrid,* is another of insights that we will discuss later.

Active Machine Learning is Predictive Coding - Passive Analytic Methods Are Not

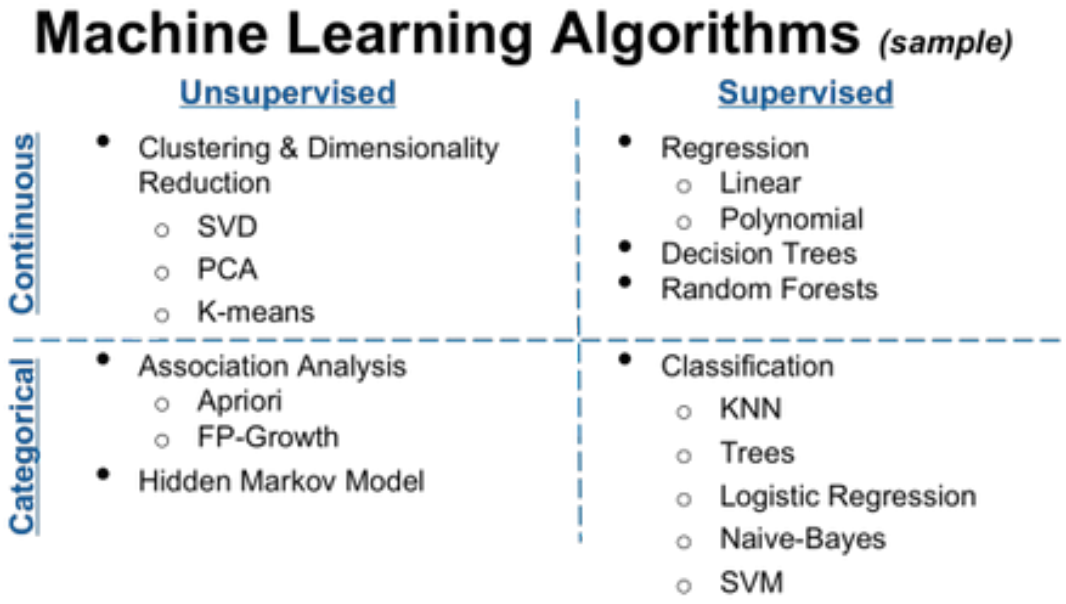
Even though our methods are *multimodal* and *hybrid,* the primary search method we rely on is *Active Machine Learning*. The overall name of our method is, after all, *Predictive Coding*. And, as any information retrieval expert will tell you, predictive coding means *active machine learning*. That is the only true AI method. The passive type of machine learning that some vendors use under the name *Analytics* is NOT the same thing as Predictive Coding. These passive Analytics have been around for years and are far less powerful than *active machine learning*.

These search methods, that used to be called *Concept Search,* were a big improvement upon relying on keyword search alone. I remember talking about concepts search techniques in reverent terms when I did my first Legal Search webinar in 2006 with Jason Baron and Professor Doug Oard. That same year, Kroll Ontrack bought one of the original developers and patent holders of concept search, Engenium. For a short time in 2006 and 2007 Kroll Ontrack was the only vendor to have these concept search tools. The founder of Engenium, David Chaplin came with the purchase, and became Kroll Ontrack's VP of Advanced Search Technologies for three years. (Here is an interesting interview of Chaplin that discusses what he and Kroll Ontrack were doing with advanced search analytic-type tools when he left in 2009.)

But search was hot and soon boutique search firms like, Clearwell, Cataphora, Content Analyst (the company recently purchased by popular newcomer, kCura),

and other e-discovery vendors developed their own concept search tools. Again, they were all using passive machine learning. It was a big deal ten years ago. For a good description of these admittedly powerful, albeit now dated search tools, see the concise, well-written article by D4's Tom Groom, *The Three Groups of Discovery Analytics and When to Apply Them*.

Search experts and information scientists know that *active machine learning*, also called *supervised machine learning*, was the next big step in search after concept searches, which are, in programming language, also known as passive or unsupervised machine learning. I am getting out of my area of expertise here, and so am unable go into any details, other than present the below instructional chart by Hackbright Academy that sets forth key difference between supervised learning (predictive coding) and unsupervised (analytics, aka concept search).



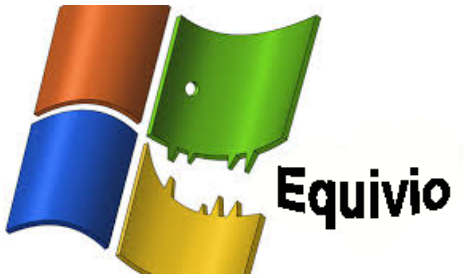
What I do know is that the bonafide *active machine learning* software in the market today all use either a form of *Logistic Regression*, including Kroll Ontrack, or *SVM*, which means Support Vector Machine.

e-Discovery Vendors Have Been Market Leaders in *Active Machine Learning* Software

After Kroll Ontrack absorbed the Engenium purchase, and its founder Chaplin completed his contract with Kroll Ontrack and moved on, Kroll Ontrack focused their efforts on the next big step, *active machine learning*, aka predictive coding. They have always been that kind of cutting-edge company, especially when it comes to search, which is one reason they are one of my personal favorites. A few

of the other, then leading e-discovery vendors did too, including especially Recommind and the Israeli based search company, Equivo. Do not get me wrong, the concept search methods, now being sold under the name of TAR *Analytics*, are powerful search tools. They are a part of our multimodal tool-kit and should be part of yours. But they are not predictive coding. They do not rank documents according to your external input, your *supervision*. They do not rely on human feedback. They group documents according to passive analytics of the data. It is automatic, unsupervised. These passive analytic algorithms can be good tools for efficient document review, but they not *active machine learning* and are nowhere near as powerful.

Many of the software companies that made the multi-million dollar investments necessary to go to the next step and build document review platforms with *active machine learning* algorithms have since been bought out by big-tech and repurposed out of the e-discovery market. They are the ghosts of legal search past. Clearwell was purchased by Symantec and has since disappeared. Autonomy was purchased by Hewlett Packard and has since disappeared. Equivio was purchased by Microsoft and has since disappeared. *See* e-Discovery Industry Reaction to Microsoft's Offer to Purchase Equivio for $200 Million – Part One and Part Two. Recommind was recently purchased by OpenText and, although it is too early to tell for sure, may also soon disappear from e-Discovery.



Slightly outside of this pattern, but with the same ghosting result, e-discovery search company, Cataphora, was bought by Ernst & Young, and has since disappeared. The year after the acquisition, Ernst & Young added predictive coding features from Cataphora to its internal discovery services. At this point, all of the Big Four Accounting Firms, claim to have their own proprietary software with predictive coding. Along the same lines, at about the time of the Cataphora buy-out, consulting giant FTI purchased another e-discovery document review company, Ringtail Solutions (known for its petri dish like visualizations). Although not exactly ghosted by FTI from the e-discovery world after the purchase, they have been absorbed by the giant FTI.

Outside of consulting/accountancy, in the general service e-discovery industry for lawyers, there are, at this point (late 2016) just a few document review platforms left that have real *active machine learning*. Some of the most popular ones left behind certainly *do not*. They only have passive learning analytics. Again, those are good features, but they are not *active machine learning*, one of the nine basic insights of Predictive Coding 4.0 and a key component of the e-Discovery Team's document review capabilities.

The power of the advanced, *active learning* technologies that have been developed for e-discovery is the reason for all of these acquisitions by big-tech

and the big-4 or 5. It is not *just* about wild overspending, although that may well have been the case for Hewlett Packard payment of $10.3 Billion to buy Autonomy. The ability to do AI-enhanced document search and review is a very valuable skill, one that will only increase in value as our data volumes continue to explode. The tools used for such document review are also quite valuable, both inside the legal profession and, as the *ghostings* prove, well beyond into big business. *See e-Discovery Industry Reaction to Microsoft's Offer to Purchase Equivio for $200 Million*, Part Two.

The indisputable fact that so many big-tech companies have bought up the e-discovery companies with *active machine learning* software should tell you a lot. It is a testimony to the advanced technologies that the e-discovery industry has spawned. When it comes to advanced search and document retrieval, we in the e-discovery world are the best in the world my friends, primarily because we have (or can easily get) the best tools.

Search is king of our modern Information Age culture. *See Information → Knowledge → Wisdom: Progression of Society in the Age of Computers*. The search for evidence to peacefully resolve disputes is, in my most biased opinion, the most important search of all. It sure beats *selling sugar water*. Without truth and justice all of the petty business quests for fame and fortune would crumble into anarchy, or worse, dictatorship.

With this background it is easy to understand why some of the e-discovery vendors left standing are not being completely candid about the capabilities of their document review software. (It is called *puffing* and is not illegal.) The industry is unregulated and, alas, most of our *expert commentators* are paid by vendors. They are not independent. As a result, many of the lawyers who have tried what they *thought* was predictive coding, and had disappointing results, have never really tried predictive coding at all. They have just used slightly updated concept search.

Alternatively, some of the disappointed lawyers may have used one of the many *now-ghosted* vendor tools. They were all early version 1.0 type tools. For example, Clearwell's *active machine learning* was only on the market for a few months with this feature before they were bought and ghosted by Symantec. (I think Jason Baron and I were the first people to see an *almost completed* demo of their product at a breakfast meeting a few months before it was released.) Recommind's predictive coding software was well developed at the time of their sell-out, but not its methods of use. Most of its customers can testify as to how difficult it is to operate. That is one reason that OpenText was able to buy them so cheaply, which, we now see, was part of their larger acquisition plan culminating in the purchase of Dell's EMC document management software.

All software still using early methods, what we call version 1.0 and 2.0 methods based on control sets, are cumbersome and hard to operate, not just Recommind's system. I explained this in my article last year, Predictive Coding 3.0. I also mentioned in this article that some vendors with predictive coding would *only* let you use predictive coding for search. It was, in effect, mono-modal. That is also a mistake. All types of search must be used - multimodal - for the predictive coding type of search to work efficiently and effectively. More on that point later.

Maura Grossman Also Blows the Whistle on Ineffective "TAR tools"

Maura Grossman, who is now an independent expert in this field, made many of these same points in a recent interview with *Artificial Lawyer*, a periodical dedicated to AI and the Law. *AI and the Future of E-Discovery: AL Interview with Maura Grossman* (Sept. 16, 2016). When asked about the viability of the "over 200 businesses offering e-discovery services" Maura said, among other things:

> In the long run, I am not sure that the market *can* support so many e-discovery providers ...
>
> ... many vendors and service providers were quick to label their existing software solutions as "TAR," without providing any evidence that they were effective or efficient. Many overpromised, overcharged, and underdelivered. Sadly, the net result was a hype cycle with its peak of inflated expectations and its trough of disillusionment. E-discovery is still far too inefficient and costly, either because ineffective so-called "TAR tools" are being used, or because, having observed the ineffectiveness of these tools, consumers have reverted back to the stone-age methods of keyword culling and manual review.

Now that Maura is no longer with the conservative law firm of *Wachtell Lipton,* she has more freedom to speak her mind about caveman lawyers. It is refreshing and, as you can see, echoes much of what I have been saying. But wait, there is still more that you need to hear from the interview of new Professor Grossman:

> It is difficult to know how often TAR is used given confusion over what "TAR" is (and is not), and inconsistencies in the results of published surveys. As I noted earlier, "Predictive Coding"—a term which actually pre-dates TAR—and TAR itself have been oversold. Many of the commercial offerings are nowhere near state of the art; with

> the unfortunate consequence that consumers have generalised their poor experiences (*e.g.*, excessive complexity, poor effectiveness and efficiency, high cost) to *all* forms of TAR. In my opinion, these disappointing experiences, among other things, have impeded the adoption of this technology for e-discovery. ...
>
> Not all products with a "TAR" label are equally effective or efficient. There is no *Consumer Reports* or *Underwriters Laboratories* ("*UL*") that evaluates TAR systems. Users should not assume that a so-called "market leading" vendor's tool will necessarily be satisfactory, and if they try one TAR tool and find it to be unsatisfactory, they should keep evaluating tools until they find one that works well. To evaluate a tool, users can try it on a dataset that they have previously reviewed, or on a public dataset that has previously been labelled; for example, one of the datasets prepared for the TREC 2015 or 2016 Total Recall tracks. ...

She was then asked by the *Artificial Lawyer* interviewer (name never identified), which is apparently based in the UK, another popular question:

> As is often the case, many lawyers are fearful about any new technology that they don't understand. There has already been some debate in the UK about the 'black box' effect, i.e., barristers not knowing how their predictive coding process actually worked. But does it really matter if a lawyer can't understand how algorithms work?

The following is an excerpt of Maura's answer. Suggest you consult the full article for a complete picture. *AI and the Future of E-Discovery: AL Interview with Maura Grossman* (Sept. 16, 2016). I am not sure whether she put on her Google Glasses to answer (probably not), but anyway, I rather like it.

> Many TAR offerings have a long way to go in achieving predictability, reliability, and comprehensibility. But, the truth that many attorneys fail to acknowledge is that so do most non-TAR offerings, including the brains of the little black boxes we call contract attorneys or junior associates. It is really hard to predict how any reviewer will code a document, or whether a keyword search will do an effective job of finding substantially all relevant documents. But we are familiar with these older approaches (and we think we understand their mechanisms), so we tend to be lulled into overlooking their limitations.

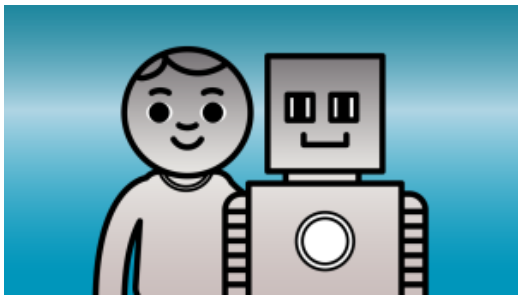
The brains of the little black boxes we call contract attorneys or junior associates. So true. We will go into that more thoroughly in our discussion of the *GIGO & QC* insight.

Recent Team Insights Into *Active Machine Learning*

To summarize what I have said so far, in the field of legal search, only *active machine learning:*

- effectively enhances human intelligence with artificial intelligence;
- qualifies for the term *Predictive Coding*.

I want to close on this discussion of *active machine learning* with one more insight. This one is slightly technical, and again, if I explain it correctly, should seem perfectly obvious. It is certainly not new, and most search experts will already know this to some degree. Still, even for them, there may some nuances to this insight that they have not thought of. It can be summarized as follows: ***active machine learning* should have a double feedback loop with active monitoring by the attorney trainers.**

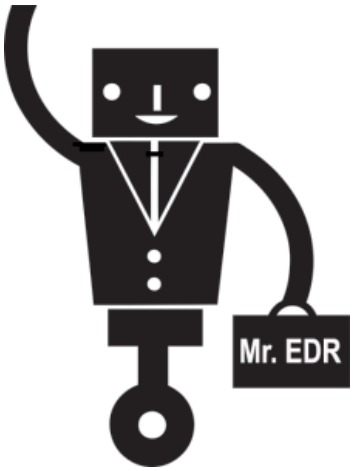


Active machine learning should create feedback for both the algorithm (the data classified) AND the human managing the training. Both should learn, not just the robot. They should, so to speak, be friends. They should get to know each other.

Many predictive coding methods that I have read about, or heard described, including how I first used *active machine learning*, did not sufficiently include the human trainer in the feedback loop. They were static types of training using single a feedback loop. These methods are, so to speak, very standoffish, aloof. Under these methods the attorney trainer does not even try to understand what is going on with the robot. The information flow was one-way, from attorney to machine.

As I grew more experienced with the EDR software I started to realize that it is possible to start to understand, at least a little, what the black box is doing. Logistic based AI is a foreign intelligence, but it is intelligence. After a while you

start to understand it. So although I started just using one-sided machine training, I slowly gained the ability to read how EDR was learning. I then added another dimension, another feedback loop that was very interesting one indeed. Now I not only trained and provided feedback to the AI as to whether the predictions of relevance were correct, or not, but I also received training from the AI as to how well, or not, it was learning. That in turn led to the humorous personification of the Kroll Ontrack software that we now call Mr. EDR. *See* MrEDR.com. When we reached this level, machine training became a fully active, two-way process.



We now understand that to fully supervise a predictive coding process you to have a good understanding of what is happening. How else can you supervise it? You do not have to know exactly how the engine works, but you at least need to know how fast it is going. You need a speedometer. You also need to pay attention to how the engine is operating, whether it is over-heating, needs oil or gas, etc. The same holds true to teaching humans. Their brains are indeed mysterious black boxes. You do not need to know exactly how each student's brain works in order to teach them. You find out if your teaching is getting through by questions.

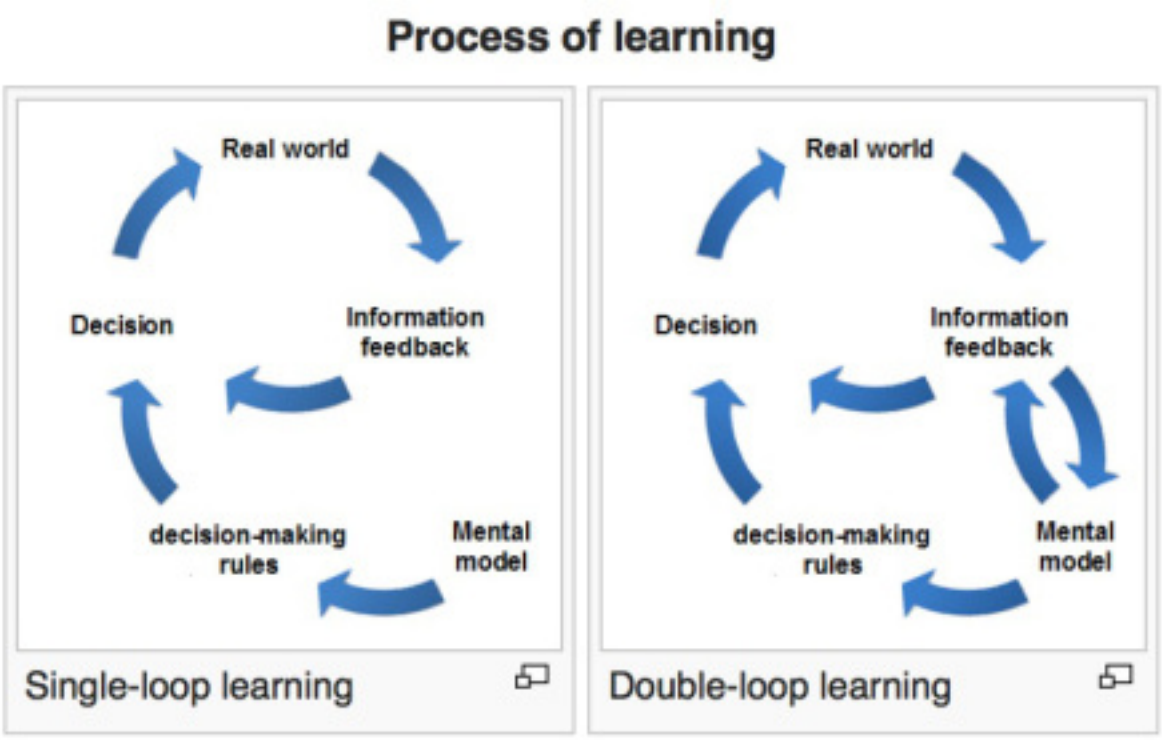
For us *supervised learning* means that the human attorney has an active role in the process. A role where the attorney trainer learns by observing the trainee, the AI in creation. I want to know as much as possible, so long as it does not slow me down significantly.

In other methods of using predictive coding that we have used or seen described the *only role* of the human trainer is to say yes or no as to the relevance of a document. The decision as to what documents to select for training has already been predetermined. Typically it is the highest ranked documents, but sometimes also some mid-ranked "uncertain documents" or some "random documents" are added in the mix. The attorney has no say in what documents to look at. They are all fed to him or her according to predetermined rules. These decision making rules are set in advance and do not change. These *active machine learning* methods work, but they are slow, and less precise, not to mention *boring as hell*.

The recall of these single-loop passive supervision methods may also not be as good. The jury is still out on that question. We are trying to run experiments on that now, although it can be hard to stop yawning. See an earlier experiment on this topic testing the single loop teaching method of random selection: *Borg Challenge*: *Report of my experimental review of 699,082 Enron documents using a semi-automated monomodal methodology*.

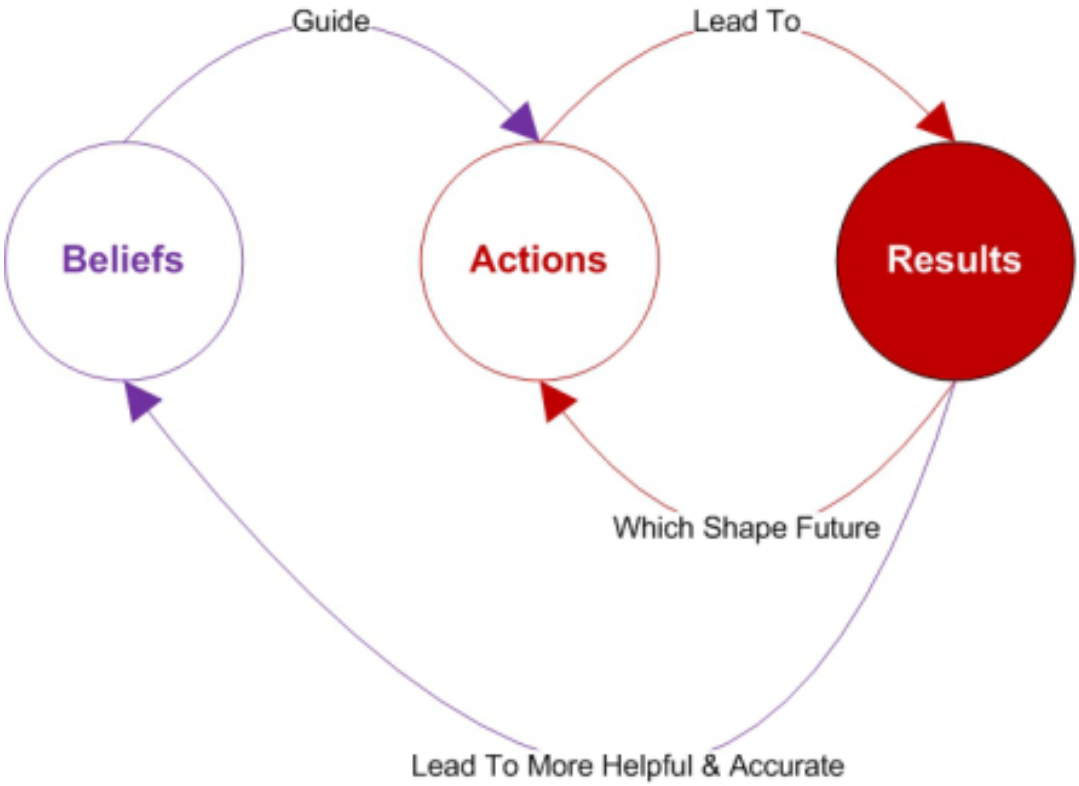
These mere yes or no, limited participation methods are hybrid Man-Machine methods, but, in our opinion, they are imbalanced towards the Machine. (Again,

more on the question of Hybrid Balance will be covered in the next installment of this article.) This single versus dual feedback approach seems to be the basic idea behind the Double Loop Learning approach to human education depicted in the diagram below. *Also see* Graham Attwell, *Double Loop Learning and Learning Analytics* (*Pontydysgu,* May 4, 2016).



To quote Wikipedia:

> The double loop learning system entails the modification of goals or decision-making rules in the light of experience. The first loop uses the goals or decision-making rules, the second loop enables their modification, hence "double-loop." ...
>
> Double-loop learning is contrasted with "single-loop learning": the repeated attempt at the same problem, with no variation of method and without ever questioning the goal. ...
>
> Double-loop learning is used when it is necessary to change the mental model on which a decision depends. Unlike single loops, this model includes a shift in understanding, from simple and static to broader and more dynamic, such as taking into account the changes in the surroundings and the need for expression changes in mental models.



The method of active machine learning that we use in Predictive Coding 4.0 is a type of double loop learning system. As such it is ideal for legal search, which is inherently *ad hoc*, where even the understanding of relevance evolves as the project develops. As Maura noted near the end of the *Artificial Lawyer* interview:

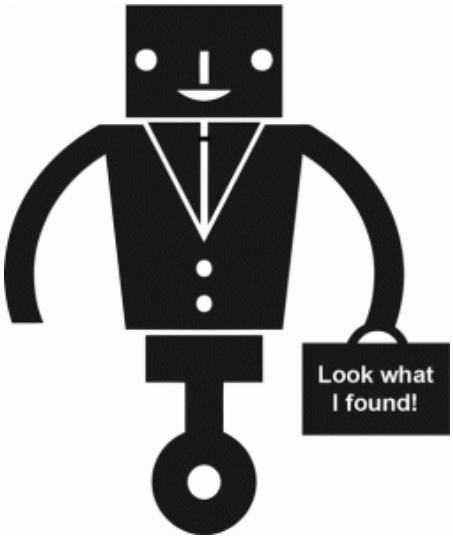
> ... e-discovery tends to be more ad hoc, in that the criteria applied are typically very different for every review effort, so each review generally begins from a nearly zero knowledge base.

The driving impetus behind our double feedback look system is to allow for training document selection to vary according to the circumstances encountered. Attorneys select documents for training and then observe how these documents impact the AI's overall ranking of the documents. Based on this information decisions are then made by the attorney as to which documents to next submit for training. A single fixed mental model is not used, such as only submitting the ten highest ranked documents for training.

The human stays involved and engaged and selects the next documents to add to the training based on what she sees. This makes the whole process much more interesting. For example, if I find a group of relevant spreadsheets by some other means, such as a keyword search, then, when I add these document to the training, I observe how these documents impact the overall ranking of the dataset. For instance, did this training result in an increase of relevance ranking of other spreadsheets? Was the increase nominal or major? How did it impact the

ranking of other documents? For instance, were emails with a lot of numbers in them suddenly much higher ranked? Overall, was this training effective? Were the documents in fact relevant as predicted that moved up in rank to the top, or near top of probable relevance? What was the precision rate like for these documents? Does the AI now have a good understanding of relevance of spreadsheets, or need more training on that type of document? Should we focus our search on other kinds of documents?

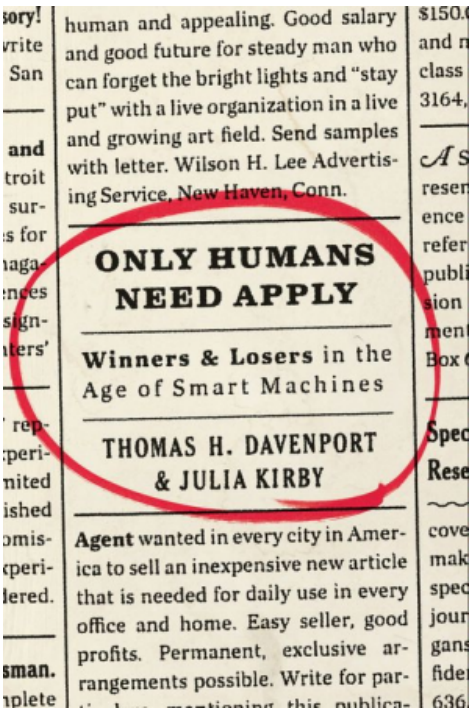
You see all kinds of variations on that. If the spreadsheet understanding (ranking) is good, how does it compare to its understanding (correct ranking) of Word Docs or emails? Where should I next focus my multimodal searches? What documents should I next assign to my reviewers to read and make a relevancy determination? These kind of considerations keep the search interesting, fun even. Work as play is the best kind. Typically we simply assign the documents for attorney review that have the highest ranking (which is the essence of what Grossman and Cormack call CAL), but not always. We are flexible. We, the human attorneys, are the second positive feedback loop.



We like to remain in charge of teaching the classifier, the AI. We do not just turn it over to the classifier to teach itself. Although sometimes, when we are out of ideas and are not sure what to do next, we will do exactly that. We will turn over to the computer the decision of what documents to review next. We just go with his top predictions and use those documents to train. Mr. EDR has come through for us many times when we have done that. But this is more of an exception, than the rule. After all, the classifier is a *tabula rasa*. As Maura put it: *each review generally begins from a nearly zero knowledge base*. Before the training starts, it knows nothing about document relevance. The computer does not come with built-in knowledge of the law or relevance. You know what you are looking for. You know what is relevant, even if you do not know how to find it, or even whether it exists at all. The computer does not know what you are looking for, aside from what you have told it by your yes-no judgments on particular documents. But, after you teach it, it knows how to find more documents that probably have the same meaning.

By observation you can see for yourself, first hand, how your training is working, or not working. It is like a teacher talking to their students to find out what they learned from the last assigned reading materials. You may be surprised by how much, or how little they learned. If the last approach did not

work, you change the approach. That is double-loop learning. In that sense our active monitoring approach it is like continuous dialogue. You learn how and if the AI is learning. This in turn helps you to plan your next lessons. What has the student learned? Where does the AI need more help to understand the conception of relevance that you are trying to teach it.


human and appealing. Good salary and good future for steady man who can forget the bright lights and "stay put" with a live organization in a live and growing art field. Send samples with letter. Wilson H. Lee Advertising Service, New Haven, Conn.

ONLY HUMANS NEED APPLY

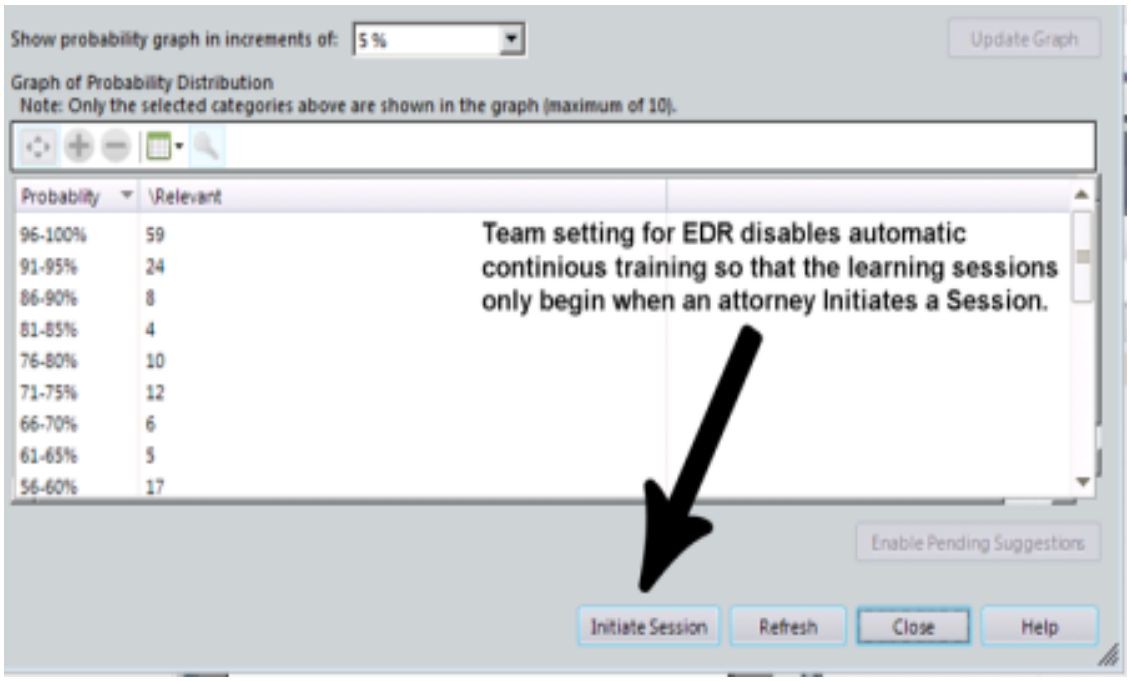
Winners & Losers in the Age of Smart Machines

THOMAS H. DAVENPORT & JULIA KIRBY

Agent wanted in every city in America to sell an inexpensive new article that is needed for daily use in every office and home. Easy seller, good profits. Permanent, exclusive arrangements possible. Write for par-

This monitoring of the AI's learning is one of the most interesting aspects of *active machine learning*. It is also a great opportunity for human creativity and value. The inevitable advance of AI in the law can mean more jobs for lawyers overall, but only for those able step up and change their methods. The lawyers able to play the second loop game of active machine learning will have plenty of employment opportunities. *See eg*. Thomas H. Davenport, Julia Kirby, *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines* (Harper 2016).

Going down into the weeds a little bit more, our active monitoring dual feedback approach means that when we use Kroll Ontrack's EDR software, we adjust the settings so that *new learning sessions* are not created automatically. They only run when and if we click on the Initiate Session button shown in the EDR screenshot below (arrow and words were added). We do not want the training to go on continuously in the background (typically meaning at periodic intervals of every thirty minutes or so.) We only want the learning sessions to occur *when we say so*. In that way we can know exactly what documents EDR is training on during a session. Then, when that training session is complete, we can see how the input of those documents has impacted the overall data ranking. For instance, are there now more documents in the 90% or higher probable relevance category and if so, how many? The picture below is of a completed TREC project. The probability rankings are on the far left with the number of documents shown in the adjacent column. Most of the documents in the 290,099 collection of Bush email were in the 0-5% probable relevant ranking not included in the screen shot.



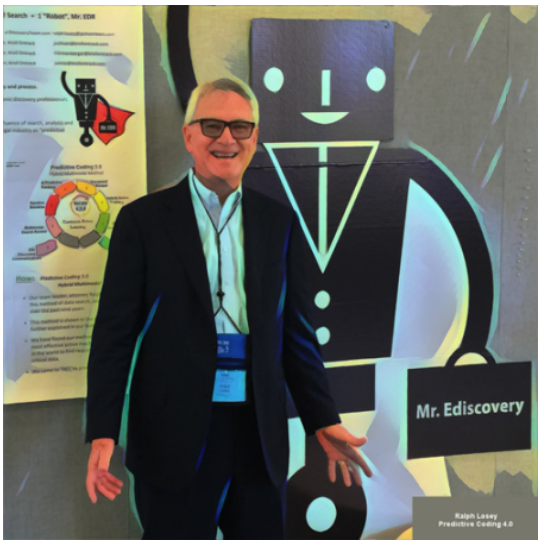
This means that the e-Discovery Team's *active learning* is not continuous, in the sense of always training. It is instead *intelligently spaced*. That is an essential aspect of our *Balanced Hybrid* approach to electronic document review. The machine training only begins when we click on the "Initiate Session" button in EDR that the arrow points to. It is only continuous in the sense that the training continues until all human review is completed. The *spaced training,* in the sense of *staggered* in time, is itself an *ongoing* process until the production is completed. We call this *Intelligently Spaced Training* or **IST.**

Ongoing training using IST improves efficiency and precision, and also improves Hybrid human-machine communications. Thus, in our team's opinion, IST is a better process of electronic document review than training automatically without human participation, the so-called CAL approach promoted (and recently trademarked) by search experts and professors, Maura Grossman and Gordon Cormack.

Exactly how we space out the timing of training in IST is a little more difficult to describe without going into the particulars of a case. A full, detailed description would require the reader to have intimate knowledge of the EDR software. Our IST process is, however, software neutral. You can follow the IST dual feedback method of *active machine learning* with any document review software that has active machine learning capacities and also *allows you to decid*e when to initiate a training session. (By the way, a *training* session is the same thing as a *learning* session, but we like to say *training,* not learning, as that takes the human perspective and we are pro-human!) You cannot do that if the training is *literally continuous* and cannot be halted while you input a new batch of relevance-determined documents for training.

The details of IST, such as when to initiate a training session, and what human coded documents to select next for training, is an *ad hoc* process. It depends on the data itself, the issues involved in the case, the progress made, the stage of the review project and time factors. This is the kind of thing you learn by doing. It is not rocket science, but it does help keep the project interesting. Hire one of our team members to guide your next review project and you will see it in action. It is easier than it sounds. With experience Hybrid Multimodal IST becomes an intuitive process, much like riding a bicycle.

To summarize, *active machine learning* should be a dual feedback process with double-loop learning. The training should continue throughout a project, but it should be spaced in time so that you can actively monitor the progress, what we call ***IST***. The software should learn from the trainer, of course, but the trainer should also learn from the software. This requires active monitoring by the teacher who reacts to what he or she sees and adjusts the training accordingly so as to maximize recall and precision.



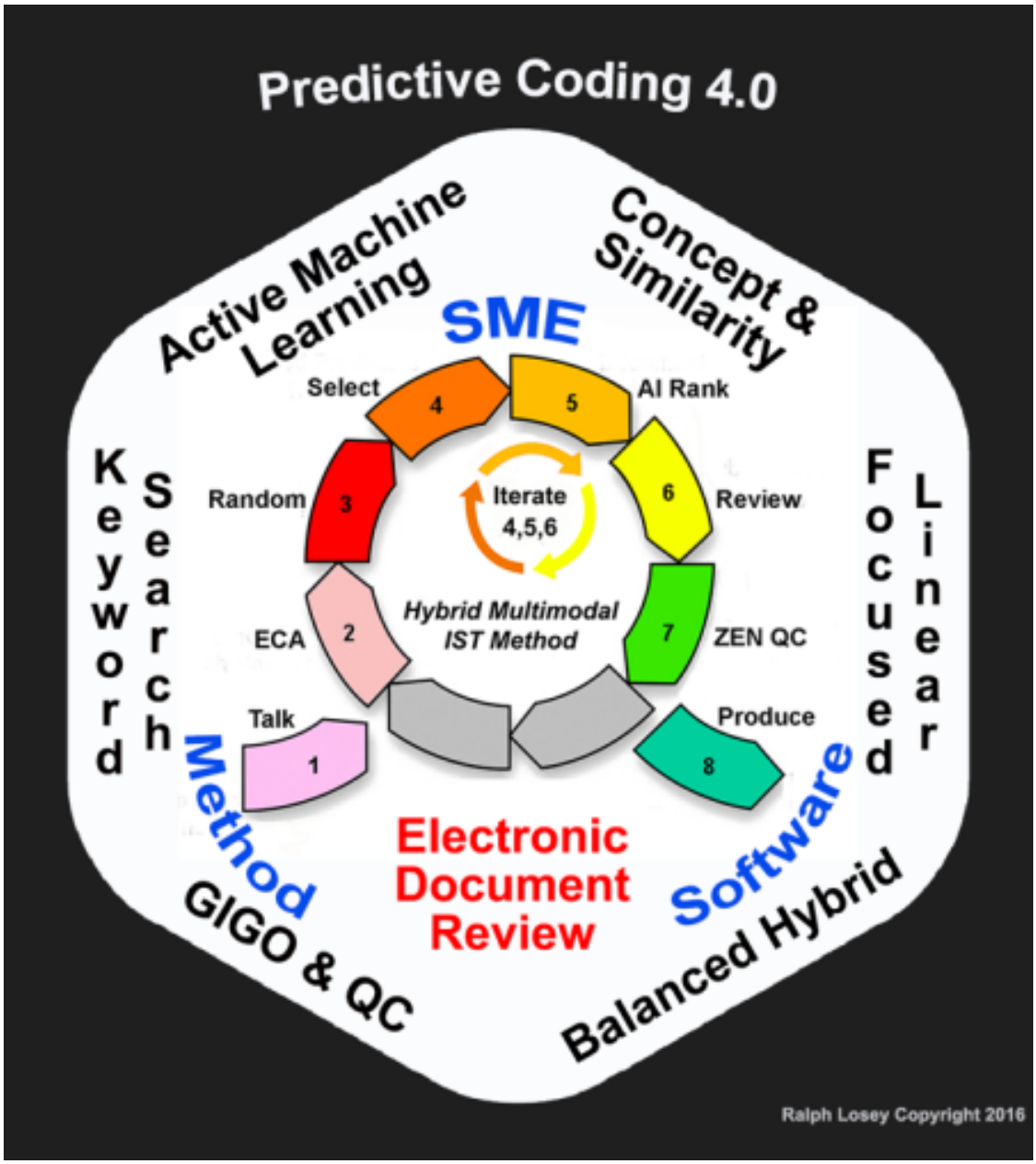
This is really nothing more than a common sense approach to teaching. No teacher who just *mails in* their lessons, and does not pay attention to the students, is ever going to be effective. The same is true for *active machine learning*. That's the essence of the insight. Simple really.

PART THREE

The e-Discovery Team's latest enhancements to electronic document review using Predictive Coding are based on seventeen points; the first nine are insights and the last eight are workflow steps:

1. **Active Machine Learning** (aka Predictive Coding)
2. **Concept & Similarity Searches** (aka Passive Learning)
3. **Keyword Search** (tested, Boolean, parametric)
4. **Focused Linear Search** (key dates & people)
5. **GIGO & QC** (Garbage In, Garbage Out) (Quality Control)
6. **Balanced Hybrid** (man-machine balance with IST)
7. **SME** (Subject Matter Expert, typically trial counsel)
8. **Method** (for electronic document review)
9. **Software** (for electronic document review)
10. **Talk** (step 1 - relevance dialogues)
11. **ECA** (step 2 - early case assessment using all methods)
12. **Random** (step 3 - prevalence range estimate, not control sets)
13. **Select** (step 4 - choose documents for training machine)
14. **AI Rank** (step 5 - machine ranks documents according to probabilities)
15. **Review** (step 6 - attorneys review and code documents)
16. **Zen QC** (step 7 - Zero Error Numerics Quality Control procedures)
17. **Produce** (step 8 - production of relevant, non-privileged documents)

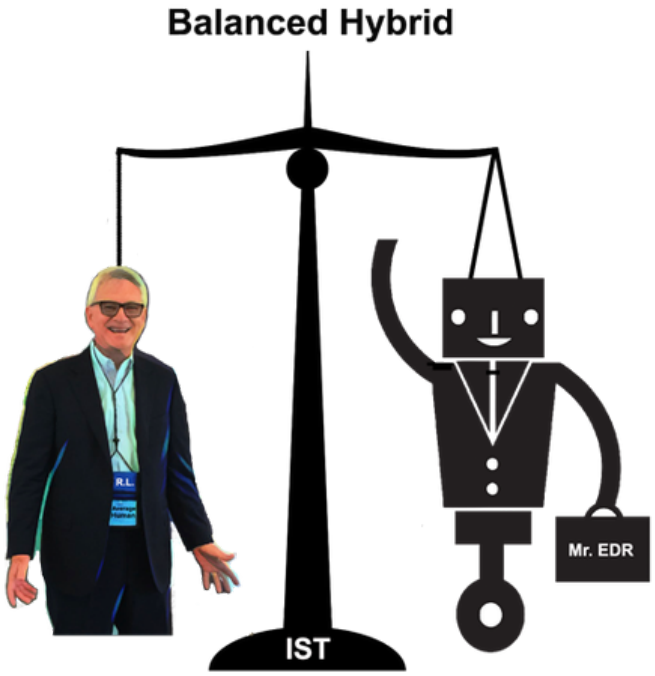
This is all summarized in the diagram below.



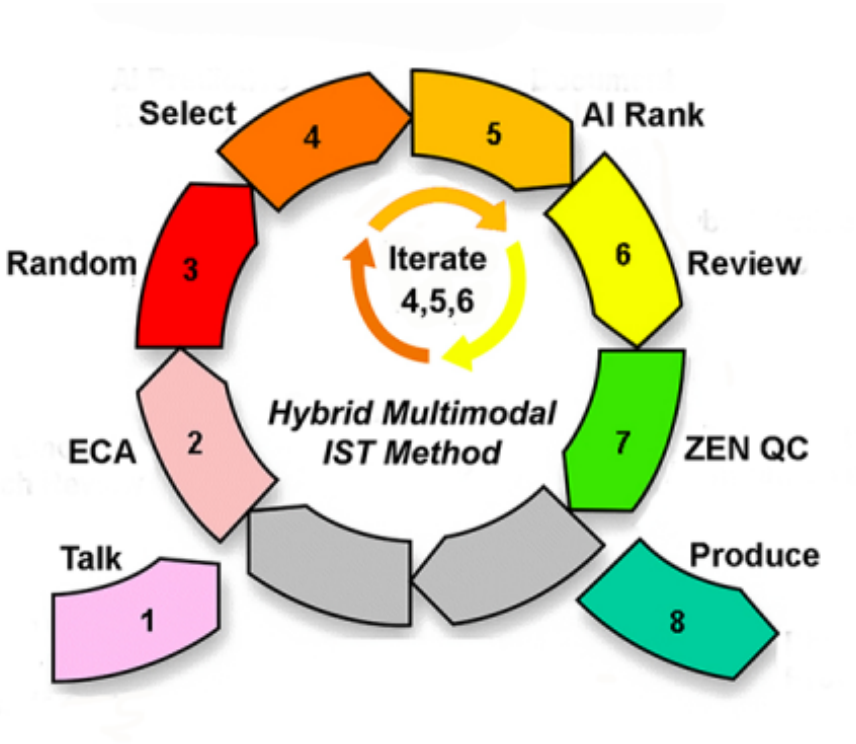
In Part One we explained how these insights came about and provided other general background. In Part Two we explained the first of the nine insights, *Active Machine Learning*, including the method of *double-loop learning*. In the process we introduced three more insights, *Balanced Hybrid, Concept & Similarity Searches*, and *Software*. For continuity purposes we will address *Balanced Hybrid* next.

Balanced Hybrid
Using *Intelligently Spaced Training* - IST™

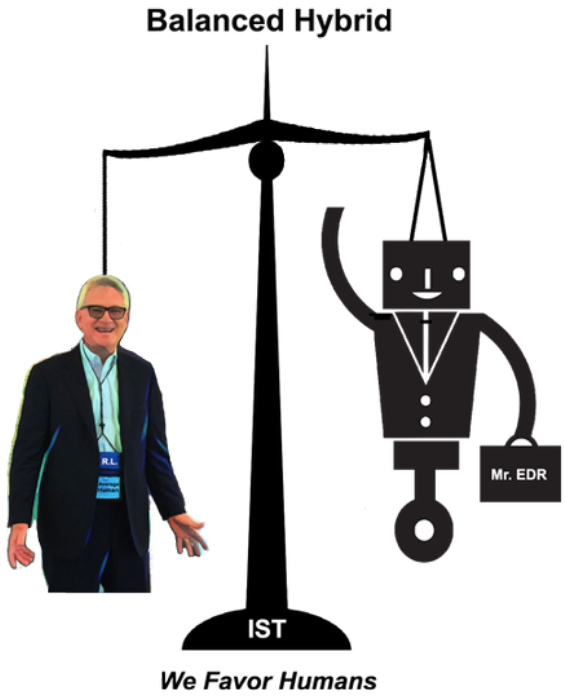
The *Balanced Hybrid* insight is complementary to *Active Machine Learning*. It has to do with the relationship between the human training the machine and the machine itself. The name itself says it all, namely that is it balanced. We rely on both software and skilled attorneys using the software.



We advocate reliance on the machine after it becomes trained, after it starts to understand your conception of relevance. At that point we find it very helpful to rely on what the machine has determined to be the documents most likely to be relevant. We have found it is a good way to improve precision in the sixth step of our 8-step document review methodology shown below. We generally use a balanced approach where we start off relying more on human selections of documents for training based on their knowledge of the case and other search selection processes, such as keyword or passive machine learning, a/k/a concept search. *See* steps 2 and 4 of our 8-step method - ECA and Select. Then we switch to relying more on the machine as it's understanding catches one. See steps 4 and 5 - Select and AI Rank. It is usually balanced throughout a project with equal weight given to the human trainer, typically a skilled attorney, and the machine, a predictive coding algorithm of some time, typically logistic regression or support vector.



Unlike other methods of *Active Machine Learning* we do not completely turn over to the machine all decisions as to what documents to review next. We look to the machine for guidance as to what documents should be reviewed next, but it is always just guidance. We never completely abdicate control over to the machine. I have gone into this before at some length in my article *Why the 'Google Car' Has No Place in Legal Search*. In this article I cautioned against over reliance on fully automated methods of active machine learning. Our method is designed to empower the humans in control, the skilled attorneys. Thus although our Hybrid method is generally balanced, our scale tips slightly in favor of humans, the team of attorneys who run the document review. *We favor humans*. So while we like our software very much, and have even named it Mr. EDR, we have an unabashed favoritism for humans. More on this at the conclusion of the *Balanced Hybrid* section of this article.

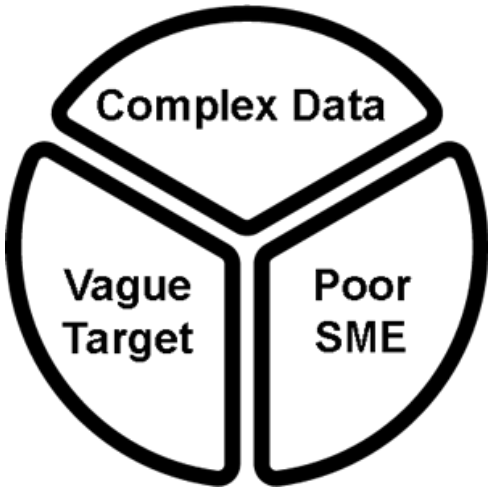


Three Factors That Influence the Hybrid Balance

We have shared the previously described hybrid insights before in earlier e-Discovery Team writings on predictive coding. The new insights on Balanced Hybrid are described in the rest of this segment. Again, they are not *entirely new* either. They represent more of a deepening of understanding and should be familiar to most document review experts. First, we have gained better insight into when and why the Balanced Hybrid approach should be tipped one way or another towards greater reliance on humans or machine. We see three factors that influence our decision.

On some projects your precision and recall improves by putting greater reliance of the AI, on the machine. These are typically projects where one or more of the following conditions exist:

- the data itself is very complex and difficult to work with, such as specialized forum discussions; or,
- the search target is ill-defined, i.w. - no one is really sure what they are looking for; or,
- the Subject Matter Expert (SME) making final determinations on relevance has limited experience and expertise.



On some projects your precision and recall improves by putting even greater reliance of the humans, on the skilled attorneys working with the machine. These are typically projects where the converse of one or more of the three criteria above are present:

- the data itself is fairly simple and easy to work with, such as a disciplined email user (note this has little or nothing to do with data volume) or,

- the search target is well-defined, i.w. there are clearly defined search requests and everyone is on the same page as to what they are looking for; or,
- the Subject Matter Expert (SME) making final determinations on relevance has extensive experience and expertise.

What was somewhat surprising from our 2016 TREC research is how one-sided you can go on the Human side of the equation and still attain near perfect recall and precision. The Jeb Bush email underlying all thirty of our topics in TREC Total Recall Track 2016 is, at this point, well-known to us. It is fairly simple and easy to work with. Although the spelling of the thousands of constituents who wrote to Jeb Bush was atrocious (far worse than general corporate email, except maybe construction company emails), Jeb's use of the email was fairly disciplined and predictable. As a Florida native and lawyer who lived through the Jeb Bush era, and was generally familiar with all of the issues, and have become very familiar with his email, I have become a good SME, and, to a somewhat lesser extent, so has my whole team. (I did all ten of the Bush Topics in 2015 and another ten in 2016.) Also, we had fairly well defined, simple search goals in most of the topics.

For these reasons in many of these 2016 TREC document review projects the role of the machine and machine ranking became fairly small. In some that I handled it was reduced to a quality control, quality assurance method. The machine would pick up and catch a few documents that the lawyers alone had missed, but only a few. The machine thus had a slight impact on improved recall, but not much effect at all on precision, which was anyway very high. (More on this experience with *easy search topics* later in this essay when we talk about our *Keyword Search* insights.)

On a few of the 2016 TREC Topics the search targets were not well defined. On these Topics our SME skills were necessarily minimized. Thus in these few Topics, even though the data itself was simple, we had to put greater reliance on the machine (in our case Mr. EDR) than on the attorney reviewers.

It bears repeating that the *volume* of emails has nothing to do with the ease or difficulty of the review project. This is a secondary question and is not dispositive as to how much weight you need to give to machine ranking. (Volume size may, however, have a big impact on project duration.)

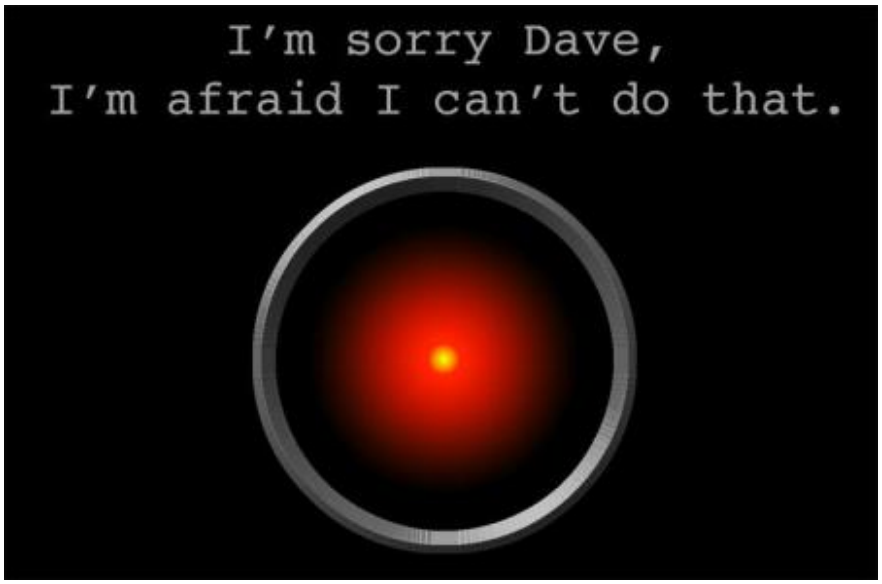
We use IST, Not CAL

Another insight in *Balanced Hybrid* in our new version 4.0 of Predictive Coding is what we call *Intelligently Spaced Training*, or IST™. We now use the term IST, instead of CAL, for two reasons:

> 1. Our previous use of the term CAL was only to refer to the fact that our method of training was continuous, in that it continued and was ongoing throughout a document review project. The term CAL has come to mean much more than that, as will be explained, and thus our continued use of the term may cause confusion.
>
> 2. Trademark rights have recently been asserted by Professors Grossman and Cormack, who originated this acronym CAL. As they have refined the use of the mark it now not only stands for Continuous Active Learning throughout a project, but also stands for a particular method of training that only uses the highest ranked documents.

Under the Grossman-Cormack CAL method the machine training continues throughout the document review project, as it does under our IST method, but there the similarities end. Under their CAL method of predictive coding the machine trains automatically as soon as a new document is coded. Further, the document or documents are selected by the software itself. It is a fully automatic process. The only role of the human is to say yes or no as to relevance of the document. The human does not select which document or documents to review next to say yes or no to. That is controlled by the algorithm, the machine. Their software always selects and presents for review the document or documents that it considers most likely to be relevant (highest probability of relevance) and has not already been coded.

The CAL method is only *hybrid,* like the e-Discovery Team method, in the sense of man and machine working together. But, from our perspective, it is not balanced. In fact, from our perspective the CAL method is way out of balance in favor of the machine. This may be the whole point of their method, to limit the human role as much as possible. The attorney has no role to play at all in selecting what document to review next and it does not matter if the attorney understands the training process. Personally, we do not like that. We want to be in charge and fully engaged throughout. We want the computer to be our tool, not our master.



Under our IST method the attorney chooses what documents to review next. We do not need the computer's permission. We decide whether to accept a batch of high-ranking documents from the machine, or not. The attorney may instead find documents that they think are relevant from other methods. Even if the high ranked method of selection of training documents is used, the attorney decides the number of such documents to use and whether to supplement the machine selection with other training documents.

In fact, the only thing in common between IST and CAL is that both processes continue throughout the life of a document review project and both are concerned with the *Stop* decision (when to when to stop the training and project). Under both methods after the Stopping point no new documents are selected for review and production. Instead, quality assurance methods that include sampling reviews are begun. If the quality assurance tests affirm that the decision to stop review was reasonable, then the project concludes. If they fail, more training and review are initiated.

Aside from the differences in document selection between CAL and IST, the primary difference is that under IST the attorney decides *when* to train. The training does not occur automatically after each document, or specified number of documents, as in CAL, or at certain arbitrary time periods, as is common with other software. In the e-Discovery Team method of IST, which, again, stands for *Intelligently Spaced* (or staggered) *Training*, the attorney in charge decide *when* to train. We control the clock, the clock does not control us. The machine does not decide. Attorneys use their own intelligence to decide when to train the machine.

This timing control allows the attorney to observe the impact of the training on the machine. It is designed to improve the communication between man and machine. That is the *double-loop learning* process described in Part Two as one of the insights into *Active Machine Learning*. The attorney trains the machine and the machine is observed so that the trainer can learn how well the machine is doing. The attorney can learn what aspects of the relevance rule have been understood and what aspects still need improvement. Based on this *student to teacher* feedback the teacher is able to custom the next rounds of training to fit the needs of the student. This maximizes efficiency and effectiveness and is the essence of double-loop learning.

Pro Human Approach to Hybrid Man-Machine Partnership

To wrap up the new Balanced Hybrid insights we would like to point out that our terminology speaks of ***Training***- *IST* - rather than ***Learning*** - *CAL*. We do this intentionally because *training* is consistent with our human perspective. That is our perspective whereas the perspective of the machine is to learn. The attorney trains and the machine learns. We favor humans. Our goal is empowerment of attorney search experts to find the truth (relevance), the whole truth (recall) and nothing but the truth (precision). Our goal is to *enhance* human intelligence with artificial intelligence. Thus we prefer a *Balanced* Hybrid approach with IS**T** and not CA**L**.

This is not to say the CAL approach of Grossman and Cormack is not good and does not work. It appears to work fine. It is just a tad too boring for us and sometimes too slow. Overall we think it is less efficient and may sometimes even be less effective than our Hybrid Multimodal method. But, even though it is not for us, it may be well be great for many beginners. It is very easy and simple to operate. From language in the Grossman Cormack patents, that appears to be what they are going for - simplicity and ease of use. They have that and a growing body of evidence that it works. We wish them well, and also their software and CAL methodology.

I expect Grossman and Cormack, and others in the pro-machine camp, to move beyond the advantages of simplicity and also argue safety issues. I expect them to argue that it is *safer* to rely on AI because a machine is more reliable than a

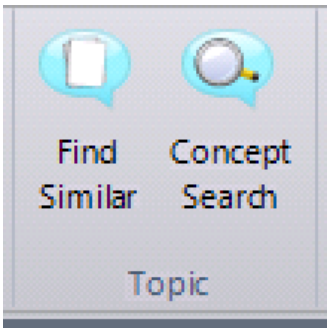
human, in the same way that Google's self-driving car is safer and more reliable than a human driven car. Of course, unlike driving a car, they still need a human, an attorney, to decide yes or no on relevance, and so they are stuck with human reviewers. They are stuck with a least a partial Hybrid method, albeit one favoring as much as possible the machine side of the partnership. We do not think the pro-machine approach will work with attorneys, nor should it. We think that only an unabashedly pro-human approach like ours is likely to be widely adopted in the legal marketplace.

The goal of the pro-machine approach of Professors Cormack and Grossman, and others, is to minimize human judgments, no matter how skilled, and thereby reduce as much as possible the influence of human error and outright fraud. This is a part of a larger debate in the technology world. We respectfully disagree with this approach, at least in so far as legal document review is concerned. (Personally I tend to agree with it in so far as the driving of automobiles is concerned.) We instead seek enhancement and empowerment of attorneys by technology, including quality controls and fraud detection. *See* *Why the 'Google Car' Has No Place in Legal Search*. No doubt you will be hearing more about this interesting debate in the coming years. It may well have a significant impact on technology in the law, the quality of justice, and the future of lawyer employment.

PART FOUR

In Part Four we will cover our insights into the remaining four basic search methods: **Concept Searches** (Passive, Unsupervised Learning); **Similarity Searches** (Families and near Duplication); **Keyword Searches** (tested, Boolean, parametric); and **Focused Linear Search** (key dates & people). The five search types are all in our newly revised Search Pyramid shown below (last revised in 2012).

Concept Searches - aka Passive Learning

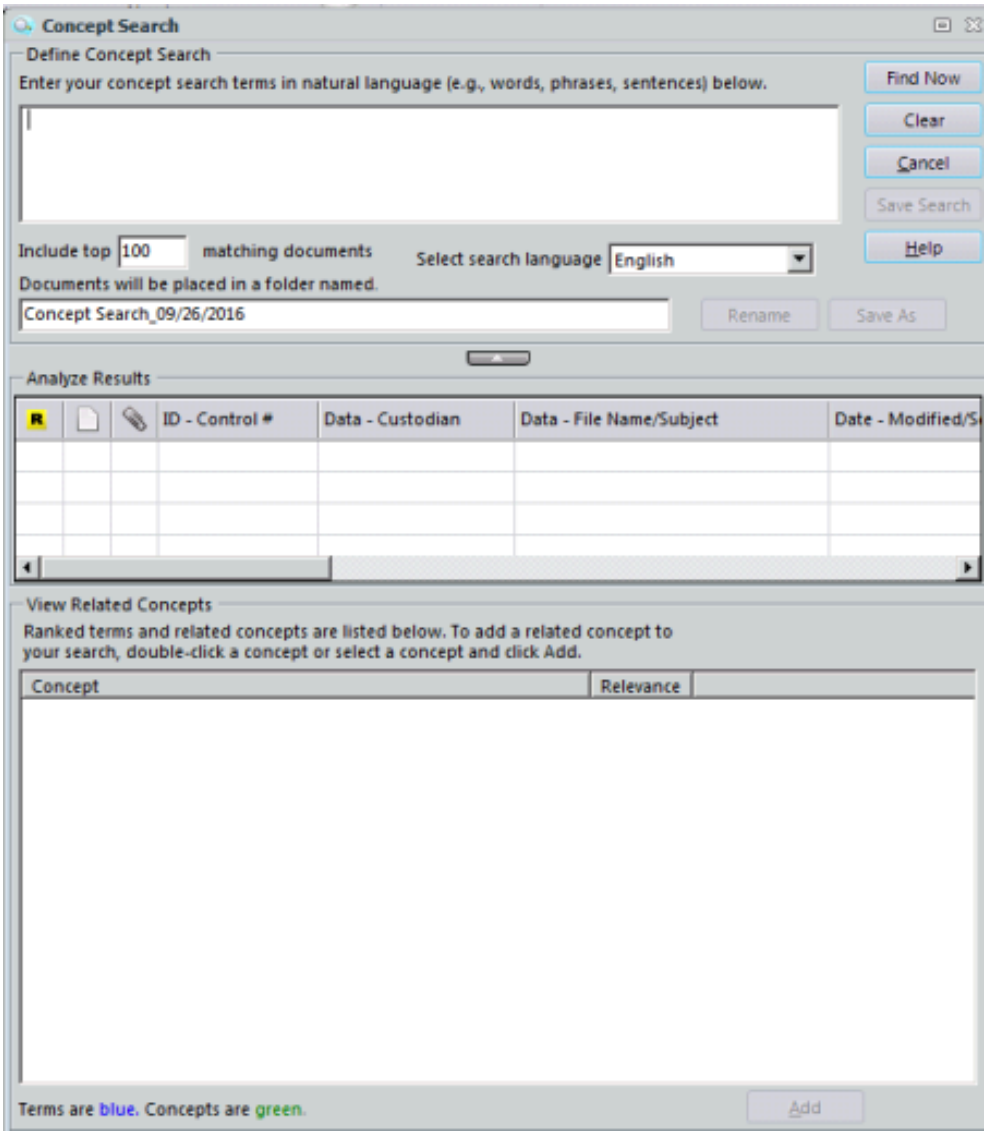


As discussed in Part Two of this article, the e-discovery search software company, Engenium was one of the first to use Passive Machine Learning techniques. Shortly after the turn of the century, the early 2000s, Engenium began to market what later become known as Concept Searches. They were supposed to be a major improvement over then dominant Keyword Search. Kroll Ontrack bought Engenium in 2006 and acquired its patent rights to concept search. These software enhancements were taken out of the e-discovery market and removed from all competitor software, except Kroll Ontrack. The same thing happened in 2014 when Microsoft bought Equivio. *See* e-Discovery Industry Reaction to Microsoft's Offer to Purchase Equivio for $200 Million – Part One and Part Two. We have yet to see what Microsoft will do with it. All we know for sure is its predictive coding add-on for Relativity is no longer available.

David Chaplin, who founded Engeniun in 1998, and sold it in 2006, became Kroll Ontrack's VP of Advanced Search Technologies from 2006-2009. He is now the CEO of two Digital Marketing Service and Technology (SEO) companies, Atruik and SearchDex. Other vendors emerged at the time to try to stay competitive with

the search capabilities of Kroll Ontrack's document review platform. They included Clearwell, Cataphora, Autonomy, Equivio, Recommind, Ringtail, Catalyst, and Content Analyst. Most of these companies went the way of Equivo and are now ghosts, gone from the e-discovery market. There are a few notable exceptions, including Catalyst, who participated in TREC with us in 2015 and 2016.

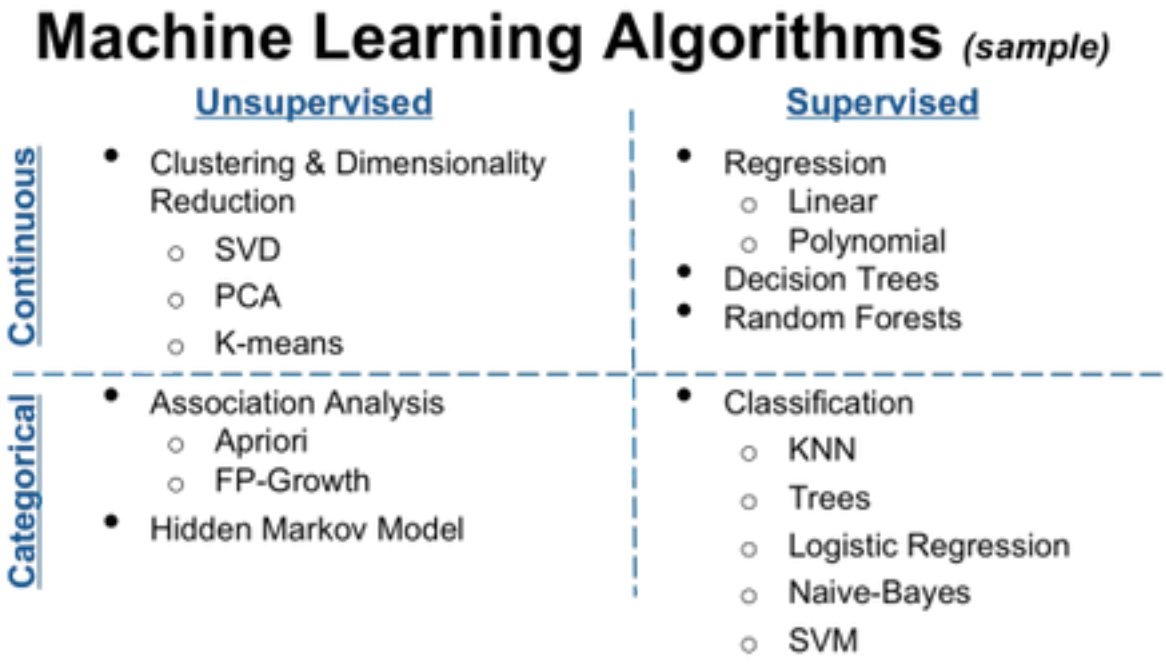
As described in Part Two of this series the so-called *Concept Searches* all relied on passive machine learning that did not depend on training or active instruction by any human (aka supervised learning). It was all done automatically by computer study and analysis of the data alone, including semantic analysis of the language contained in documents. That meant you did not have to rely on keywords alone, but could state your searches in conceptual terms. The below is a screen-shot of one example of concept search interface using Kroll Ontrack's EDR software.



For a good description of these admittedly powerful, albeit now somewhat dated search tools (at least compared to active machine learning), see the afore-cited article by D4's Tom Groom, *The Three Groups of Discovery Analytics and When to Apply Them*. The article refers to Concept Search as *Conceptual Analytics,* and is described as follows:

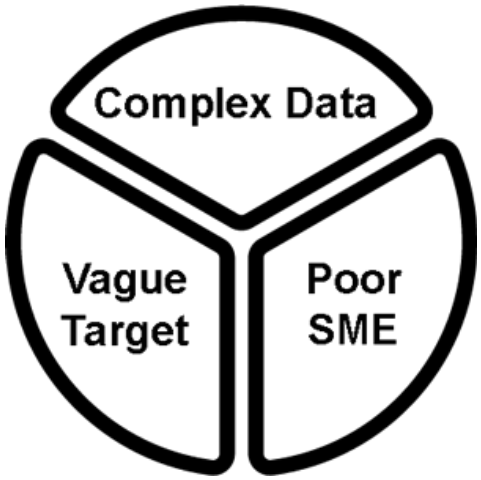
Conceptual analytics takes a semantic approach to explore the conceptual content, or meaning of the content within the data. Approaches such as Clustering, Categorization, Conceptual Search, Keyword Expansion, Themes & Ideas, Intelligent Folders, etc. are dependent on technology that builds and then applies a conceptual index of the data for analysis.

Search experts and information scientists know that *active machine learning*, also called *supervised machine learning*, was the next big step in search after concept searches, including clustering, which are, in programming language, also known as passive or unsupervised machine learning. The below instructional chart by Hackbright Academy sets forth key difference between supervised learning (predictive coding) and unsupervised or passive learning (analytics, aka concept search).



It is usually worthwhile to spend some time using concept search to speed up the search and review of electronic documents. We have found it to be of only modest value in simple search projects, with greater value added in more complex projects, especially where data is very complex. Still, in all projects, simple or complex, the use of Concept Search features such as document Clustering, Categorization, Keyword Expansion, Themes & Ideas are at least somewhat

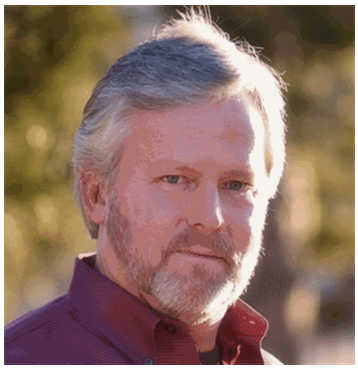
helpful. They are especially helpful in finding new keywords to try out, including wild-card stemming searches with instant results and data groupings.



In simple projects you may not need to spend much time with these kinds of searches. We find that an expenditure of at least thirty minutes at the beginning of a search is cost-effective in all projects, even simple ones. In more complex projects it may be necessary to spend much more time on these features.

Passive, unsupervised machine learning is a good way to be introduced to the type of data you are dealing with, especially if you have not worked with the client data before. In TREC Total Recall 2015 and 2016, where we were working with the same datasets, our use of these searches diminished as our familiarity with the datasets grew. They can also help in projects where the search target in not well-defined. There the data itself helps focus the target. It is helpful in this kind of sloppy, *I'll know it when I see it* type of approach. That usually indicates a failure of both target identification and SME guidance. Even with simple data you will want to use passive machine learning in those circumstances

Similarity Searches - Families and Near Duplication

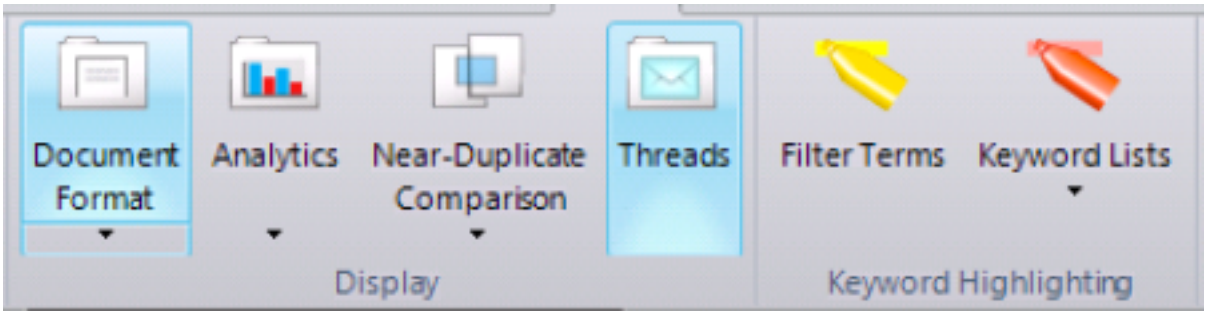


In Tom Groom's, article, *The Three Groups of Discovery Analytics and When to Apply Them*, he refers to Similarity Searches as Structured Analytics, which he explains as follows:

> Structured analytics deals with textual similarity and is based on syntactic approaches that utilize character organization in the data as the foundation for the analysis. The goal is to provide better group identification and sorting. One primary example of structured analytics for eDiscovery is Email Thread detection where analytics organizes the various email messages between multiple people into one conversation. Another primary example is Near Duplicate detection where analytics identifies documents with like text that can be then used for various useful workflows.

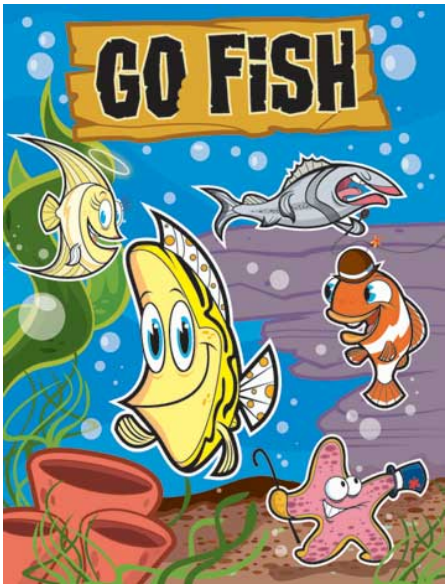
These methods can always improve efficiency of a human reviewer's efforts. It makes it easier and faster for human reviewers to put documents in context. It also helps a reviewer minimize repeat readings of the same language or same document. The near duplicate clustering of documents can significantly speed up review. In some corporate email collections the use of Email Thread detection

can also be very useful. The idea is to read the last email first, or read in chronological order from the bottom of the email chain to the top. The ability to instantly see on demand the parents and children of email collections can also speed up review and improve context comprehension.



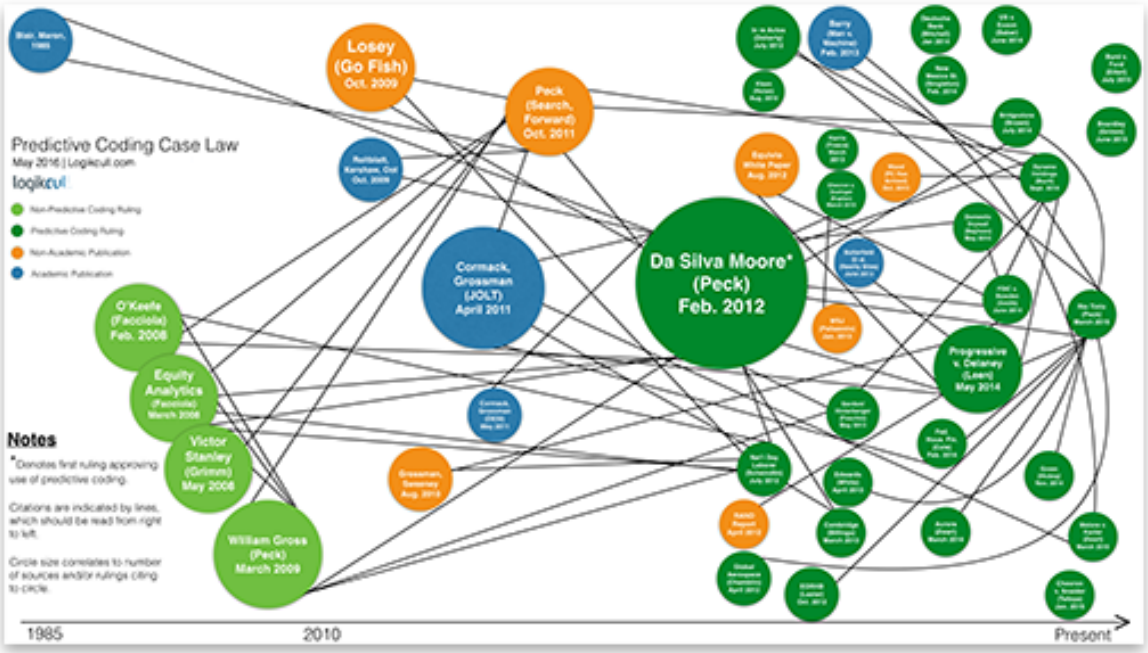
All of these *Similarity Searches* are less powerful than *Concept Search*, but tend to be of even more value than *Concept Search* in simple to intermediate complexity cases. In most simple or medium complex projects one to three hours are typically used with these kinds of software features. Also, for this type of search the volume of documents is important. The larger the data set, especially the larger the number of relevant documents located, the greater the value of these searches.

Keyword Searches - Tested, Boolean, Parametric



In my perspective as an attorney in private practice specializing in e-discovery and supervising the e-discovery work in a firm with 800 attorneys, almost all of whom do employment litigation, I have a good view of what is happening in the U.S.. We have over fifty offices and all of them at one point or another have some kind of e-discovery issue. All of them deal with opposing counsel who are sometimes mired in keywords, thinking it is the end-all and be-all of legal search. Moreover, they usually want to go about doing it without any testing. Instead, they think they are geniuses who can just dream up good searches out of thin air. They think because they know what their legal complaint is about, they know what keywords will be used by the witnesses in all relevant documents. Wrong. I cannot tell you how many times I see the word "complaint" in their keyword list. The guessing involved reminds me of the child's game of *Go Fish*.

I wrote about this in 2009 and the phrase caught on after Judge Peck and others started citing to this article, which later became a chapter in my book, *Adventures in Electronic Discovery, 209-211* (West 2011). The *Go Fish* analogy appears to be the **third most popular reference** in predictive coding case-law, after the huge, ***Da Silva Moore*** case in 2012 that Judge Peck and I are best known for.



E-discovery Team members employed by Kroll Ontrack also see hundreds of document reviews for other law firms and corporate clients. They see them from all around the world. There is no doubt in our minds that keyword search is still the dominant search method used by most attorneys. It is especially true in small to medium-sized firms, but also in larger firms that have no e-discovery search expertise. Many attorneys and paralegals who use a sophisticated, full featured document review platforms such as Kroll Ontrack's EDR, still only use keyword search. They do not use the many other powerful search techniques of EDR, even though they are readily available to them. The Search Pyramid to them looks more like this, which I call a *Dunce Hat*.

The AI at the top, standing for Predictive Coding, is, for average lawyers today, still just a far off remote mountaintop; something they have heard about, but never tried. Even though this is my specialty, I am not worried about this. I am confident that this will all change soon. Our new, easier to use methods will help with that, so too will ever improving software by the few vendors left standing. God knows the judges are already doing their part. Plus, high-tech propagation is an inevitable result of the next generation of lawyers assuming leadership positions in law firms and legal departments.

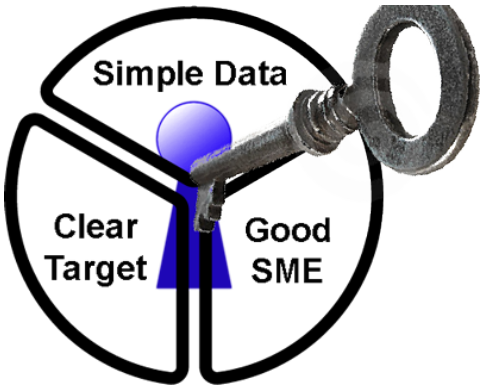
The old-timey paper lawyers around the world are finally retiring in droves. The aging out of current leadership is a good thing. Their over-reliance on untested keyword search to find evidence is holding back our whole justice system. The law must keep up with technology and lawyers must not fear math, science and AI. They must learn to keep up with technology. This is what will allow the legal profession to remain a bedrock of contemporary culture. It will happen. Positive disruptive change is just under the horizon and will soon rise.

In the meantime we encounter opposing counsel everyday who think e-discovery means to dream up keywords and demand that every document that contains their keywords be produced. The more sophisticated of this *confederacy of dunces* understand that we do not have to produce them, that they might not all be *per se* relevant, but they demand that we review them all and produce the relevant ones. Fortunately we have the revised rules to protect our clients from these kind of disproportionate, unskilled demands. All too often this is nothing more than *discovery as abuse*.

This still dominant approach to litigation is really nothing more than an artifact of the old-timey paper lawyers' use of discovery as a weapon. Let me speak plainly. This is nothing more than adversarial *bullshit discovery* with no real intent by the requesting party to find out what really happened. They just want to make the process as expensive and difficult as possible for the responding party because, well, that's what they were trained to do. That is what they think smart, adversarial discovery is all about. Just another tool in their negotiate and settle, extortion approach to litigation. It is the opposite of the modern cooperative approach.

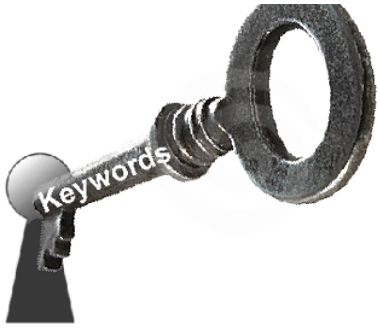
I cannot wait until these dinosaurs retire so we can get back to the original intent of discovery, a cooperative pursuit of the facts. Fortunately, a growing number of our opposing counsel do *get it*. We are able to work very well with them to get things done quickly and effectively. That is what discovery is all about. Both sides save their powder for when it really matters, for the battles over the meaning of the facts, the governing law, and how the facts apply to this law for the result desired.

Tested, Parametric Boolean Keyword Search



The biggest surprise for me in our latest research is just how amazingly good keyword search can perform under the right circumstances. I'm talking about hands-on, tested keyword search based on human document review and file scanning, sampling, and also based on strong keyword search software. When keyword search is done with skill and is based on the evidence seen, typically in a refined *series* of keyword searches, very high levels of Precision, Recall and F1 are attainable. Again, the dataset and other conditions must be just right for it to be that effective, as explained in the diagram: simple data, clear target and good SME. Sometimes keywords are the best way to find clear targets like names and dates.

In those circumstances the other search forms may not be needed to find the relevant documents, or at least to find *almost all* of the relevant documents. These are cases where the hybrid balance is tipped heavily towards the *400 pound man hacking* away at the computer. All the AI does in these circumstances, when the human using keyword search is *on a roll*, is double-check and verify that it agrees that all relevant documents have been located. It is always nice to get a free second opinion from Mr. EDR. This is an excellent quality control and quality assurance application from our legal robot friends.



We are not going to try to go through all of the ins and outs of keyword search. There are many variables and features available in most document review platforms today to make it easy to construct effective keyword searches and otherwise find similar documents. This is the kind of thing that KO and I teach to the e-discovery liaisons in my firm and other attorneys and paralegals handing electronic document reviews. The passive learning software features can be especially helpful, so too can simple indexing and clustering. Most software programs have important features to improve keyword search and make it more effective. All lawyers should learn the basic tested, keyword search skills.

There is far more to effective keyword search than a simple Google approach. (Google is concerned with finding websites, not recall of relevant evidence.) Still, in the right case, with the right data and easy targets, keywords can open the door to both high recall and precision. Keyword search, even tested and sophisticated, does not work well in complex cases or with dirty data. It certainly has its limits and there is a significant danger in over reliance on keyword search. It is typically very imprecise and can all to easily miss unexpected word usage and misspellings. That is one reason that the e-Discovery Team always supplements keyword search with a variety of other search methods, including predictive coding.

Focused Linear Search - Key Dates & People

In Abraham Lincoln's day all a lawyer had to do to prepare for a trial was talk to some witnesses, talk to his client and review all of the documents the clients had that could possibly be relevant. All of them. One right after the other. In a big case that might have taken an hour, maybe two. Flash forward one hundred years to the post-photocopier era of the 1960s and document review, linear style, reviewing them all, might take a day. By the 1990s it might take weeks. With the data volume of today such a review would take years.

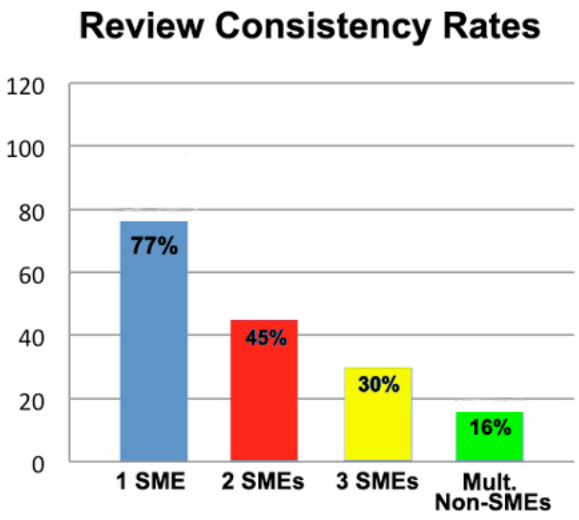
All document review was linear up until the 1990s. Until that time almost all documents and evidence were paper, not electronic. The records were filed in accordance with an organization wide filing system. They were combinations of chronological files and alphabetical ordering. If the filing was by subject then the linear review conducted by the attorney would be by subject, usually in alphabetical order. Otherwise, without subject

files, you would probably take the data and read it in chronological order. You would certainly do this with the correspondence file. This was done by lawyers for centuries to look for a coherent story for the case. If you found no evidence of value in the papers, then you would smile knowing that your client's testimony could not be contradicted by letters, contracts and other paperwork.

This kind of investigative, linear review still goes on today. But with today's electronic document volumes the task is carried out in warehouses by relatively low paid, document review *contract lawyers*. By itself it is a *fool's errand*, but it is still an important part of a multimodal approach.

There is nothing wrong with Focused Linear Search when used in moderation. And there is nothing wrong with document review *contract-lawyers*, except that they are underpaid for their services, especially the really good ones. I am a big fan of document review specialists.

Large linear review projects can be expensive and difficult to manage. Moreover, it typically has only limited use. It breaks down entirely when large teams are used because human review is so inconsistent in document analysis. Losey, R., *Less Is More: When it comes to predictive coding training, the "fewer reviewers the better"* (parts One, Two and three) (December 8, 2013, e-Discovery Team). When review of large numbers of documents are involved the consistency rate among multiple human reviewers is dismal. *Also*

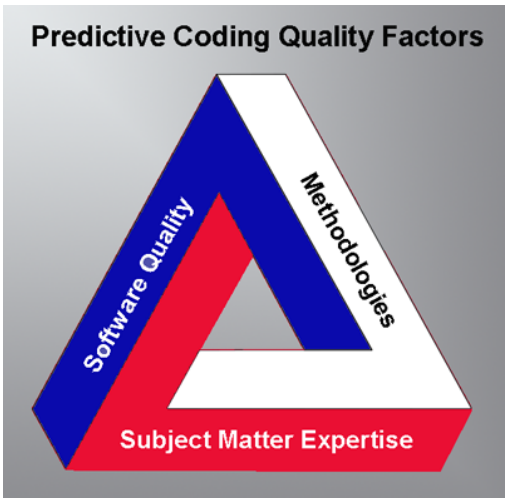


see: Roitblat, *Predictive Coding with Multiple Reviewers Can Be Problematic: And What You Can Do About It* (4/12/16).

Still, linear review can be very helpful in limited time spans and in reconstruction of a quick series of events, especially communications. Knowing what happened one day in the life of a key custodian can sometimes give you a great defense or great problem. Either are rare. Most of the time Expert Manual Review is helpful, but not critical. That is why Expert Manual Review is at the base of the Search Pyramid that illustrates our multimodal approach.

An attorney's knowledge, wisdom and skill are the foundation of all that we do, with or without AI. The information that an attorney holds is also of value, especially information about the latest technology, but the human information roles are diminishing. Instead the trend is to delegate mere information level services to automated systems. The legal robots would not be permitted to go beyond information fulfillment roles and provide legal advice based on human knowledge and wisdom. Their function would be constrained to Information processing and reports. The metrics and technology tools they provide can make it easier for the human attorneys to build a solid evidentiary foundation for trial.

PART FIVE



We have now covered five of the nine insights. In Part Five we will cover the remaining four: GIGO & QC (Garbage In, Garbage Out) (Quality Control); SME (Subject Matter Expert); Method (for electronic document review); and, Software (for electronic document review). The last three: SME - Method - Software, are all parts of Quality Control.

GIGO & QC - Garbage In, Garbage Out & Quality Control

Garbage In, Garbage Out is one of the oldest sayings in the computer world. You put garbage into the computer and it will spit it back at you in spades. It is almost as true today as it was in the 1980s when it was first popularized. Smart technology that recognizes and corrects for some mistakes has tempered GIGO somewhat, but it still remains a controlling principle of computer usage.

The GIGO Wikipedia entry explains that:

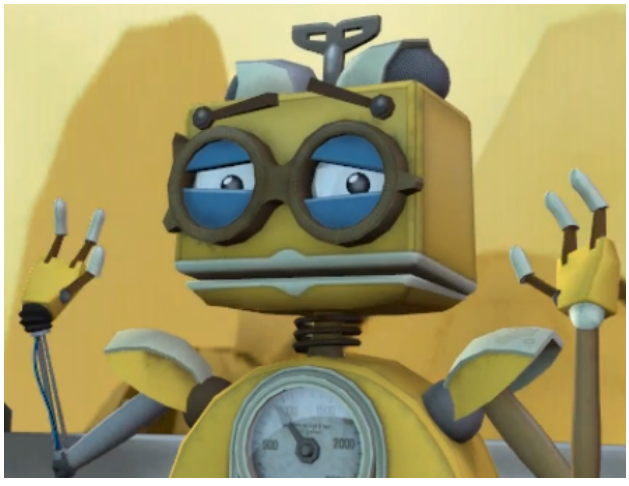
> GIGO in the field of computer science or information and communications technology refers to the fact that computers, since they operate by logical processes, will unquestioningly process unintended, even nonsensical, input data ("garbage in") and produce undesired, often nonsensical, output ("garbage out"). ... It was popular in the early days of computing, but applies even more today, when powerful computers can produce large amounts of erroneous information in a short time.

Wikipedia also pointed out an interesting new expansion of the GIGO Acronym, *Garbage In, Gospel Out:*

> It is a sardonic comment on the tendency to put excessive trust in "computerized" data, and on the propensity for individuals to blindly accept what the computer says.

Now as to our insight: GIGO in electronic document review, especially review using predictive coding, is largely the result of human error on the part of the *Subject Matter Expert*. Of course, garbage can also be created by poor methods, where too many mistakes are made, and by poor software. But to really mess things up, you need a clueless SME. These same factors also create garbage (poor results) when used with *any* document review techniques. When the subject matter expert is not good, when he or she does not have a good grasp for what is relevant, and what is important for the case, then all methods fail. Keywords and active machine learning both depend on reliable attorney expertise. Quality control literally must start at the top of any electronic document review project. It must start with the SME.

If your attorney expert, your SME, has no clue, their head is essentially garbage. With that kind of bad input, you will inevitably get bad output. This happens with all usages of a computer, but especially when using predictive coding. The computer learns what you teach it. Teach it garbage and that is what it will learn. It will hit a target all right, just not the right target. Documents will be produced, just not the ones needed to resolve the disputed issues. A poor SME makes too many mistakes and misses too many relevant documents because they do not know what is relevant and what is not.

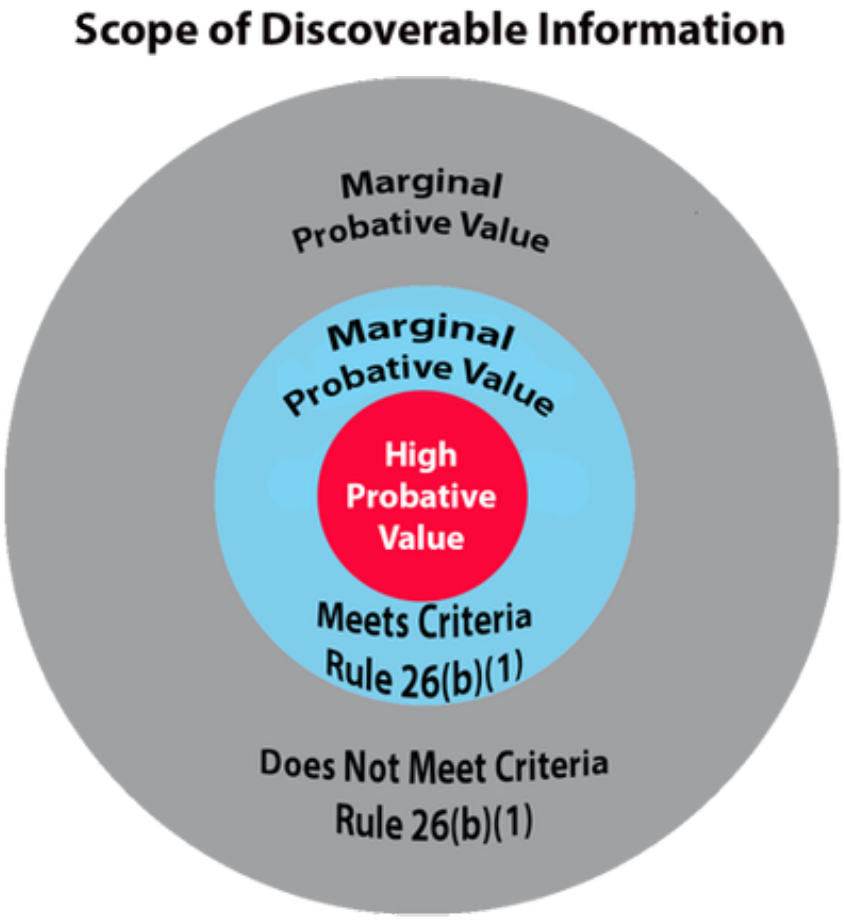


A smart AI can correct for some human errors (perfection is not required). The algorithms can correct for some mistakes in consistency by an SME, and the rest of the review team, but not that many. In machine learning for document review the legal review robot now starts as a blank slate with no knowledge of the law or the case. They depend on the SME to teach them. Someday that may change. We may see smart

robots who know the law and relevance, but we are not even near there yet. For now our robots are more like small children. They only know what you tell them, but they can spot inconsistencies in your message and they never forget.

Subject Matter Expert - SME

The predictive coding method can fail spectacularly with a poor expert, but so can keyword search. The converse of both propositions is also true. In all legal document review projects the SME needs to be an expert in scope of relevance, what is permitted discovery, what is relevant and what is not, what is important and what is not. They need to know the legal rules governing relevance backwards and forwards. They also need to have a clear understanding of the probative value of evidence in legal proceedings. This is what allows an attorney to know the scope of discoverable information.

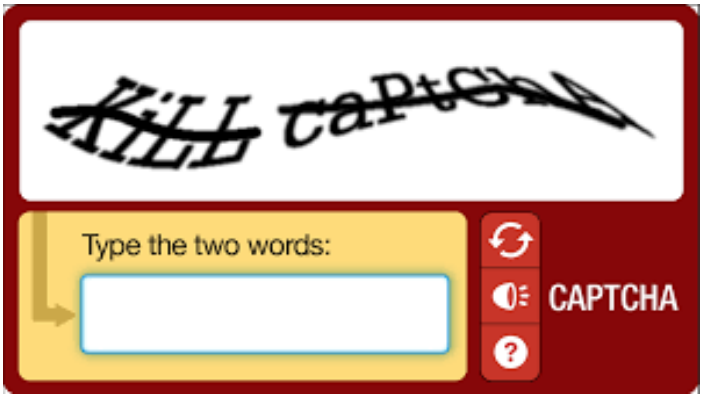


If the attorney in charge does not understand the scope of discoverable information, does not understand probative value, then the odds of finding the documents important to a case are significantly diminished. You could look at a

document with high probative value and not even know that it is relevant. This is exactly the concern of many requesting parties, that the responding party's attorney will not understand relevance and discoverability the same way they do. That is why the first step in my recommended workflow is to *Talk*, which I also call *Relevance Dialogues*.

The kind of ESI communications with opposing counsel that are needed is not whining accusations or aggressive posturing. I will go into *good talk* versus *bad talk* in some detail when I explain the first step of our eight-step method. The point of the talking that should begin any document review project is to get a common understanding of scope of discoverable information. What is the exact scope of the request for production? Don't agree the scope is proportionate? That's fine. Agree to disagree and Talk some more, this time to the judge.

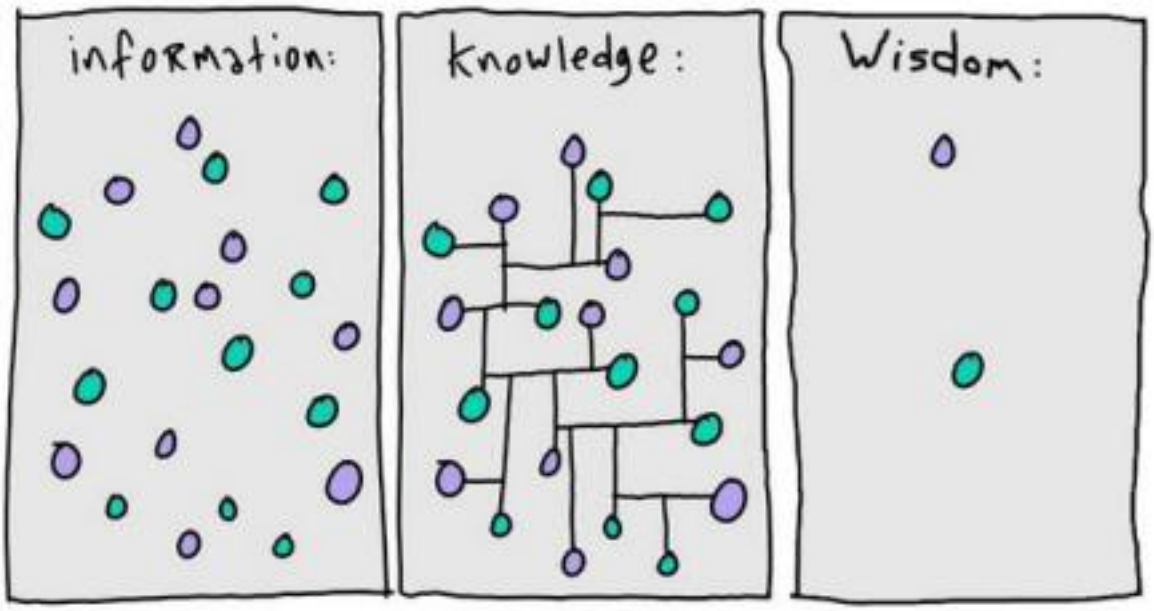
We have seen firsthand in the TREC experiments the damage that can be done by a poor SME and no judge to keep them inline. Frankly, it has been something of a shock, or wake up call, as to the dangers of poor SME relevance calling. Most of the time I am quite lucky in my firm of super-specialists (all we do is employment law matters) to have terrific SMEs. But I have been a lawyer for a long time. I have seen some real losers in this capacity in the past 36 years. I myself have been a poor SME in some of the 2015 TREC experiments. An example that comes to mind is when I had to be the SME on the subject of CAPTCHA in a collection of forum messages by hackers. It ended up being *on the job* training. I saw for myself how little I could do to guide the project. Weak SMEs make bad leaders in the world of technology and law.



There are two basic ways that discovery SMEs fail. First, there are the kind who do not really know what they are talking about. They do not have expertise in the subject matter of the case, or, let's be charitable, their expertise is insufficient. A *bullshit artist* makes a terrible SME. They may fool the client (and they often do), but they do not fool the judge or any real experts. The second kind of weak SMEs have some expertise, but they lack experience. In my old firm we used to call them *baby lawyers*. They have knowledge, but not wisdom. They lack the

practical experience and skills that can only come from grappling with these relevance issues in many cases.

That is one reason why boutique law firms like my own do so well in today's competitive environment. They have the knowledge *and the wisdom* that comes from specialization. They have seen it before and know what to do.



An SME with poor expertise has a very difficult time knowing if a document is relevant or not. For instance, a person not living in Florida might have a very different understanding than a Floridian of what *non-native plants and animals threaten the Florida ecosystem*. This was Topic 408 in TREC 2016 Total Recall Track. A native Floridian is in a better position to know the important invasive species, even ones like vines that have been in the state for over a hundred years. A non-expert with only limited information may not know, for instance, that Kudzo vines are an invasive plant from Japan and China. (They are also rumored to be the home of small, vicious Kudzo monkeys!) What is known for sure is that Kudzu, *Pueraria montana*, smothers all other vegetation around, including tall trees. A native Floridian hates Kudzo as much as they love Manatees.

A person who has just visited Florida a few times would not know what a big deal Kudzo was in Florida during the Jeb Bush administration, especially in Northern Florida. (Still is.) They had probably never heard of it at all. They could see email with the term and have no idea what the email meant. It is obvious the native SME would know more, and thus be better positioned than a fake-SME, to determine Jeb Bush email relevance to *non-native plants and animals that threaten the Florida ecosystem*. By the way, all native Floridians especially hate pythons and a python eating one of our gators as shown below is an abomination.

Expertise is obviously needed for anyone to be a subject matter expert and know the difference between relevant and irrelevant. But there is more to it than information and knowledge. It also takes experience. It takes an attorney who has handled these kinds of cases many times before. Preferably they have tried a case like the one you are working on. They have seen the impact of this kind of evidence on judge and jury. An attorney with both theoretical knowledge and practical experience makes the best SME. Your ability to contribute subject matter expertise is limited when you have no practical experience. You might think certain ESI is helpful, when in fact, it is not; it has only weak probative value. A document might technically be relevant, but the SME lacks the experience and wisdom to know that matter is *practically* irrelevant anyway.

It goes without saying that any SME needs a good review team to back them up, to properly, consistently implement their decisions. In order for good leadership to be effective, there must also be good project management. Although this insight discussion features the role of the SME member of the review team, that is only because the importance of the SME was recently emphasized to us in our TREC research. In actuality all team members are important, not just the input from the top. Project management is critical, which is an insight already well-known to us and, we think, the entire industry.

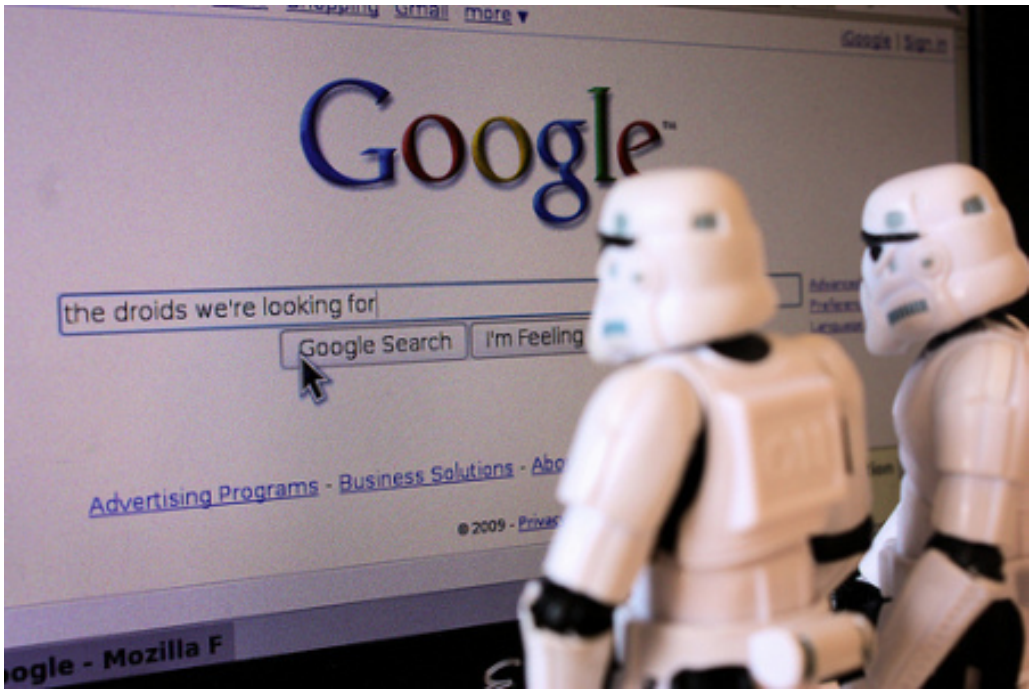
Corrupt SMEs

Of course, no SME can be effective, no matter what their knowledge and experience, if they are not fair and honest. The SME must impartially seek and produce documents that are both pro and con. This is an ethics issue in all types of document review, not just predictive coding. In my experience corrupt SMEs are rare. But it does happen occasionally, especially when a corrupt client pressures their all too dependent attorneys. It helps to know the reputation for honesty of your opposing

counsel. *See: Five Tips to Avoid Costly Mistakes in Electronic Document Review* – Part 2 contains my YouTube video, *E-DISCOVERY ETHICS*.

Also see: Lawyers Behaving Badly: Understanding Unprofessional Conduct in e-Discovery, 60 Mercer L. Rev. 983 (Spring 2009); *Mancia v. Mayflower Begins a Pilgrimage to the New World of Cooperation,* 10 Sedona Conf. J. 377 (2009 Supp.).

If I were a *lawyer behaving badly* in electronic document review, like for instance the *Qualcomm* lawyers did hiding thousands of highly relevant emails from Broadcom, I would not use predictive coding. If I wanted to *not find* evidence harmful to my case, I would use negotiated keyword search, the *Go Fish* kind.



I would also use linear review and throw an army of document review attorneys at it, with no one really knowing what the other was doing (or coding). I would subtly encourage project mismanagement. I would not pay attention. I would not supervise the rest of the team. I would not involve an AI entity, i.w.- active machine learning. I would also not use an attorney with search expertise, nor would I use a national e-discovery vendor. I would throw a novice at the task and use a local or start-up vendor who would just do what they were told and not ask too many questions.

A corrupt hide-the-ball attorney would not want to use a predictive coding method like ours. They would not want the relevant documents produced or logged that disclose the training documents they used. This is true in any continuous training process, not just ours. We do not produce irrelevant

documents, the law prevents that and protects our client's privacy rights. But we do produce relevant documents, usually in phases, so you can see what the training documents are.

A Darth Vader type hide-the-ball attorney would also want to avoid using a small, specialized, well-managed team of contract review lawyers to assist on a predictive coding project the review project. They would instead want to work with a large, distant army of contract lawyers. A small team of contract review attorneys cannot be brought into the con, especially if they are working for a good vendor. Even if you handicap them with a bad SME, and poor methods and software, they may still find a few of the damaging documents you do not want to produce. They may ask questions when they learn their coding has been changed from relevant to irrelevant. I am waiting for the next *Qualcomm* or *Victor Stanley* type case where a contract review lawyer blows the whistle on corrupt review practices. *Qualcomm Inc. v. Broadcom Corp.*, No. 05-CV-1958-B(BLM) Doc. 593 (S.D. Cal. Aug. 6, 2007) (one honest low-level engineer testifying at trial blew the whistle on Qualcomm's massive fraud to hide critical email evidence). I have heard stories from contract review attorneys, but the law provides them too little protection, and so far at least, they remain behind the scenes with horror stories.

One protection against both a corrupt SME, and SME with too little expertise and experience, is for the SME to be the *attorney in charge of the trial of the case*, or at least one who *works closely with them* so as to get their input when needed. The job of the SME is to know relevance. In the law that means you must know how the ultimate arbitrator of relevance will rule - *the judge assigned to your case*. They determine truth. An SME's own personal opinion is important, but ultimately of secondary importance to that of the judge. For that reason a good SME will often vary on the side of over-expansive relevance because they know from history that is what the judge is likely to allow in this type of case.

This is a key point. The judges, not the attorneys, ultimately decide on close relevance and related discoverability issues. The head trial attorney interfaces with the judge and opposing counsel, and should have the best handle on what is or is not relevant or discoverable. A good SME can predict the judge's rulings and, even if not perfect, can gain the judicial guidance needed in an efficient manner.

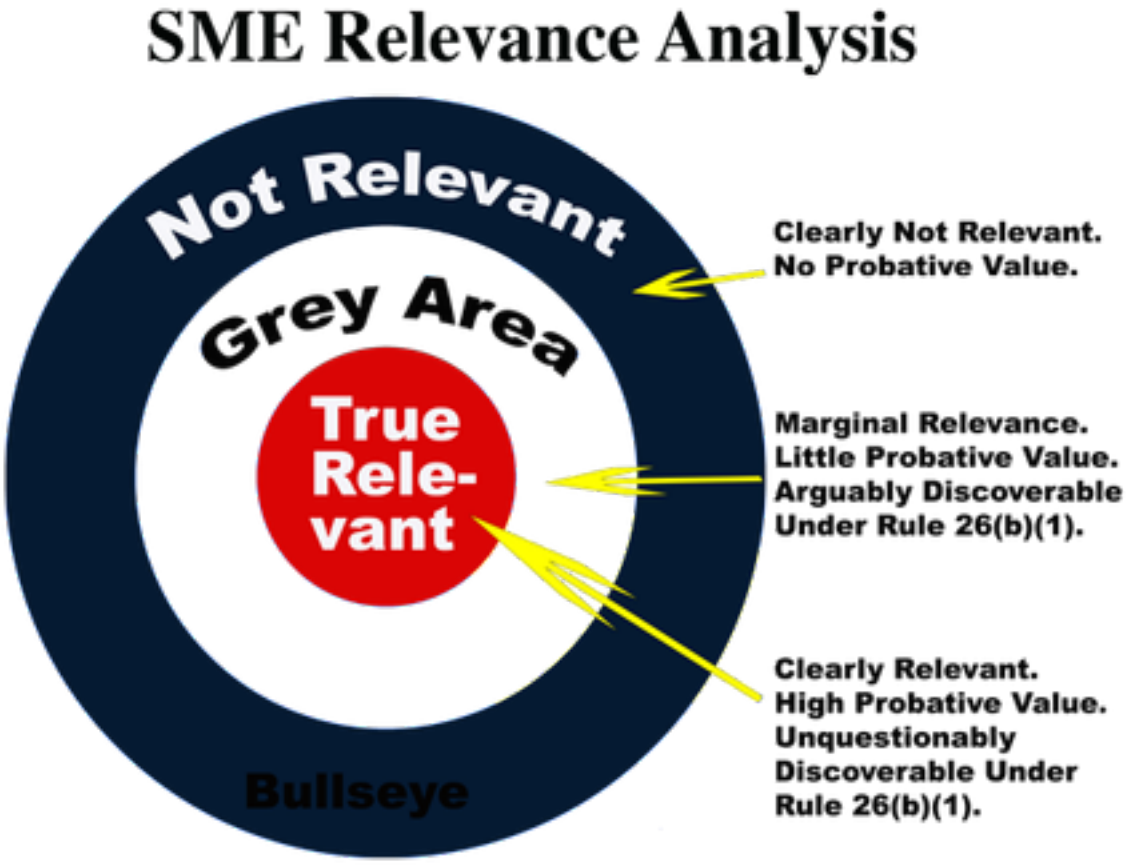
If the judge detects unethical conduct by the attorneys before them, including the attorney signing the Rule 26(g) response, they can and should respond harshly to punish the attorneys. *See eg:* *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 269 F.R.D. 497, 506 (D. Md. 2010). The *Darth Vader's* of the world can be defeated. I have done it many times with the help of the presiding judge. You can too. You can win even if they personally attack both you and the judge. Been through that too.

Three Kinds of SMEs: Best, Average & Bad

When your project has a good SME, one with both high knowledge levels and experience, with wisdom from having been there before, and knowing the judge's views, then your review project is likely to succeed. That means you can attain both high recall of the relevant documents and also high precision. You do not waste much time looking at irrelevant documents.

When an SME has only medium expertise or experience, or both, then the expert tends to err on the side of over-inclusion. They tend to call grey area documents relevant because they do not know they are unimportant. They may also not understand the new *Federal Rules of Civil Procedure* governing discoverability. Since they do not know, they err on the side of inclusion. True experts know and so tend to be more precise than rookies. The medium level SMEs may, with diligence, also attain high recall, but it takes them longer to get there. The precision is poor. That means wasted money reviewing documents of no value to the case, documents of only marginal relevance that would not survive any rational scrutiny of Rule 26(b)(1).

When the SME lacks knowledge and wisdom, then both recall and precision can be poor, even if the software and methods are otherwise excellent. A bad SME can ruin everything. They may miss most of the relevant documents and end up producing garbage without even knowing it. That is the fault of the person in charge of relevance, the SME, not the fault of predictive coding, nor the fault of the rest of the e-discovery review team.

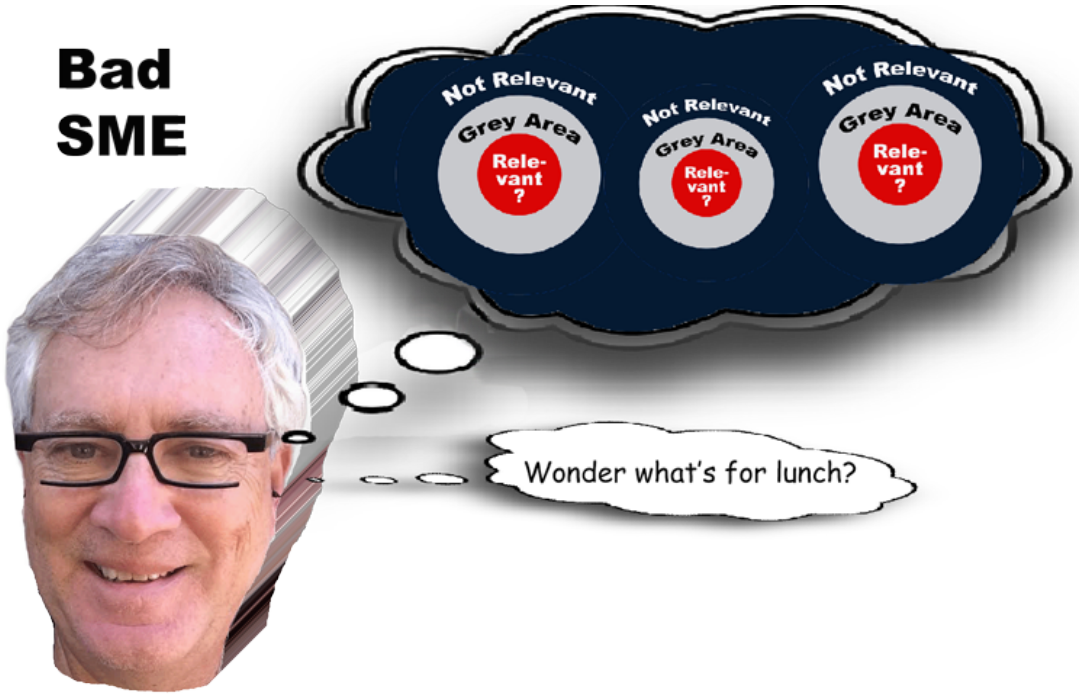


If the SME assigned to a document review project, especially a project using active machine learning, is a *high-quality SME*, then they will have a clear grasp of relevance. They will know what types of documents the review team is looking for. They will understand the probative value of certain kids of documents in this particular case. Their judgments on Rule 26(b)(1) criteria as to discoverability will be consistent, well balanced and in accord with that of the governing judge. They will instruct the whole team, including the machine, on what is true relevant, on what is discoverable and what is not. With this kind of top SME, if the software, methods, including project management, and rest of the

review team are also good, then high recall and precision are very likely.

If the SME is just average, and is not sure about many grey area documents, then they will not have a clear grasp of relevance. It will be foggy at best. They will not know what types of documents the review team is looking for. SMEs like this think that any arrow that hits a target is relevant, not knowing that only the red circle in the center is truly relevant. They will not understand the probative value of certain kids of documents in this particular case. Their judgments on Rule 26(b)(1) criteria as to discoverability will not be perfectly consistent, and will end up either too broad or too narrow, and may not be in accord with that of the governing judge. They will instruct the whole team, including the machine, on what might be relevant and discoverable in an unfocused, vague, and somewhat inconsistent manner. With this kind of SME, if the software and methods, including project management, and rest of the review team are also good, and everyone is very diligent, high recall is still possible, but precision is unlikely. Still, the project will be unnecessarily expensive.

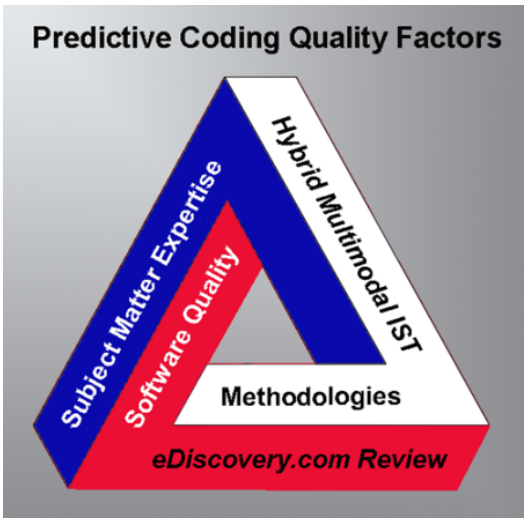
The bad SME has multiple possible targets in mind. They just search without really knowing what they are looking for. They will instruct the whole team, including the machine, on what might be relevant and discoverable in an confused, constantly shifting and often contradictory manner. Their obtuse explanations of relevance have little to do with the law, nor the case at hand. They probably have a very poor grasp of Rule 26(b)(1) *Federal Rules of Civil Procedure*. Their judgments on 26(b)(1) criteria as to discoverability, if any, will be inconsistent, imbalanced and sometimes irrational. This kind of SME probably does not even know the judge's name, much less a history of their relevance rulings in this type of case. With this kind of SME, even if the software and methods are otherwise good, there is little chance that high recall or precision will be attained. An SME like this does not know when their search arrow has hit center of the target. In fact, it may hit the wrong target entirely. Their thought-world looks like this.



A document project governed by a bad SME runs a high risk of having to be redone because important information is missed. That can be a very costly disaster. Worse, a document important to the producing parties case can be missed and the case lost because of that error. In any event, the recall and precision will both be low. The costs will be high. The project will be confused and inefficient. Projects like this are hard to manage, no matter how good the rest of the team. In projects like this there is also a high risk that privileged documents will accidentally be produced. (There is always some risk of this in today's high volume ESI world, even with a top-notch SME and review team. A Rule 502(d) Order should *always* be entered for the protection of all parties.)

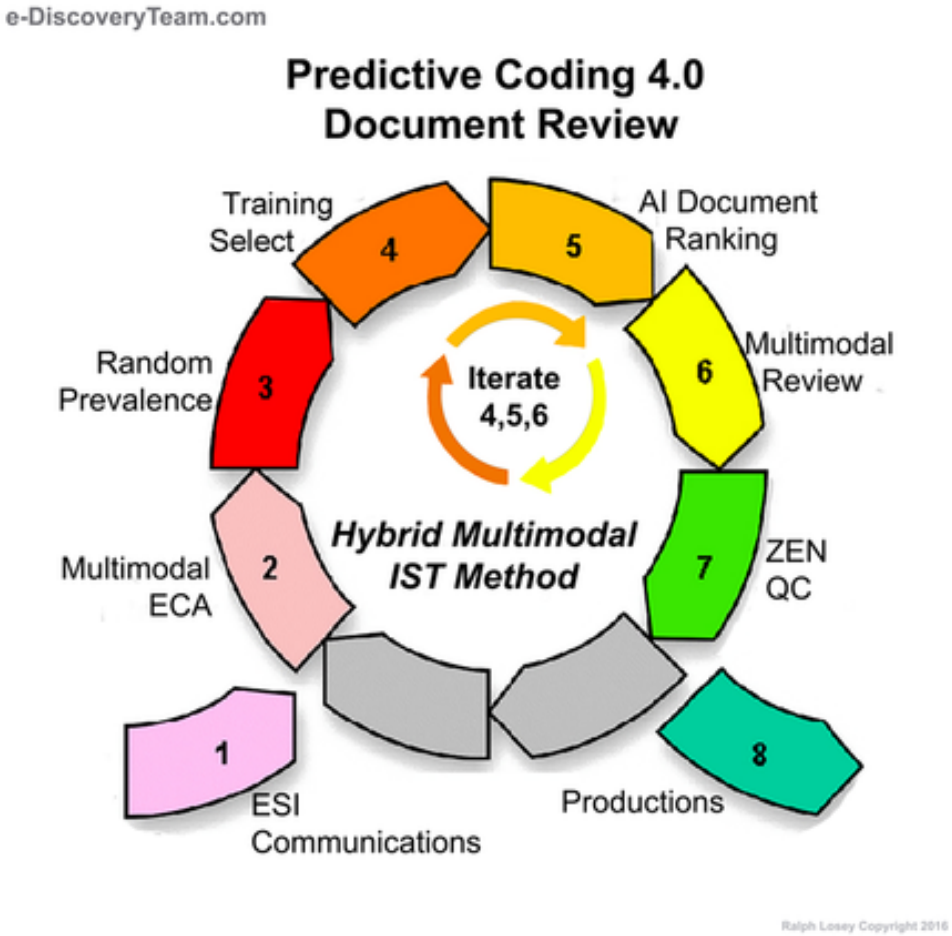
Method and Software

The **SME** and his or her implementing team is just one part of the quality triangle. The other two are **Method** of electronic document review and **Software** used for electronic document review.

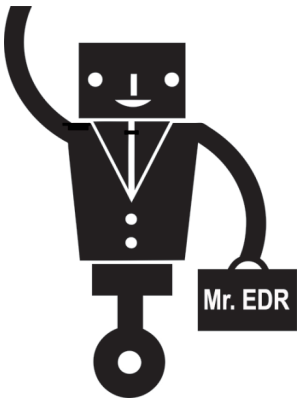


Obviously the e-Discovery Team takes *Method* very seriously. That is one reason we are constantly tinkering with and improving our methods. We released the breakthrough Predictive Coding 3.0 last year, following 2015 TREC research, and this year, after TREC 2016, we released version 4.0. You could fairly say we are obsessed with the topic. We also focus on the importance of good project

management and communications. No matter how good your SME, and how good your software, if your methods are poor, so too will your results in most of your projects. How you go about a document review project, how you manage it, is *all-important* to the quality of the end product, the production.



The same holds true for software. For instance, if your software does not have active machine learning capacities, then it cannot do predictive coding. The method is beyond the reach of the software. End of story. The most popular software in the world right now for document review does not have that capacity. Hopefully that will change soon and I can stop talking around it.



Even among the software that has active machine learning, some are better than others. It is not my job to rank and compare software. I do not go around asking for demos and the opportunity to test other software. I am too busy for that. Everyone knows that I currently prefer to use EDR. It is the software by Kroll Ontrack that I use everyday. I am not paid to endorse them and I do not. (Unlike almost every other e-

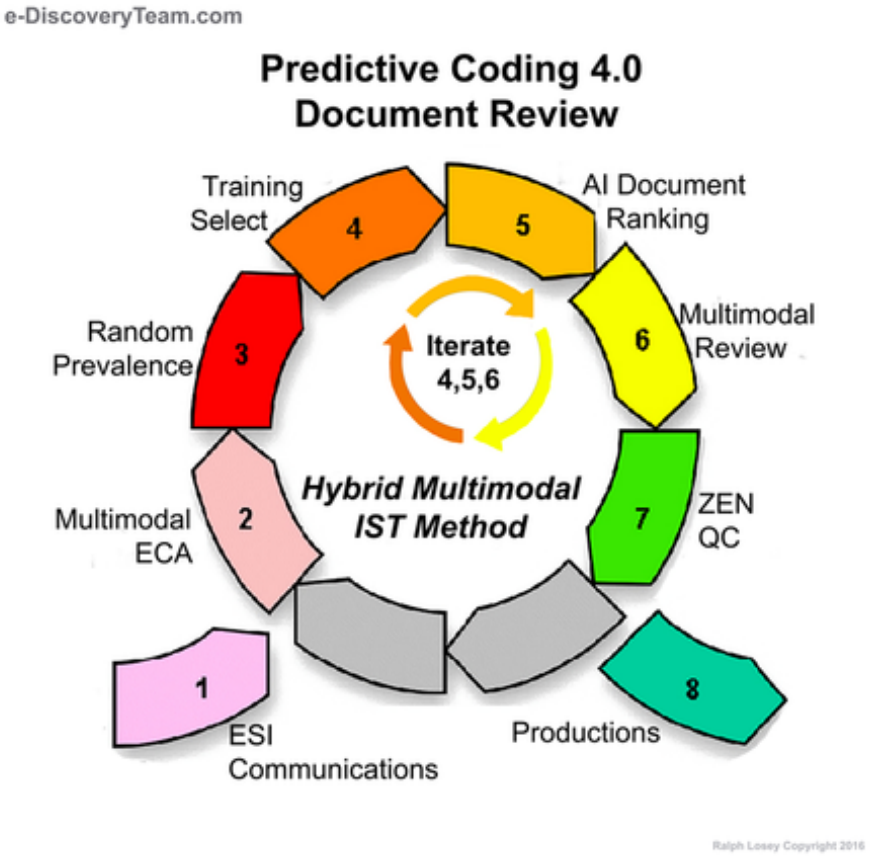
discovery commentator out there, no vendors pay me a dime.) I just share my current preference and pass along cost-savings to my clients.

I will just mention that the only other e-discovery vendor to participate with us at TREC is Catalyst. As most of my readers know, I am a fan of the founder and CEO, John Tredennick. There are several other vendors with good software too. Look around and be skeptical. But whatever you do, be sure the software you use is good. Even a great carpenter with the wrong tools cannot build a good house.

One thing I have found, that is just plain common sense, is that with good software and good methods, including good project management, you can overcome many weaknesses in SMEs, except for dishonesty or repeated, gross-negligence. The same holds true for all three corners of the quality triangle. Strength in one can, to a certain extent, make up for weaknesses in another.

PART SIX

Now that we have covered the nine insights we will describe the eight-step workflow. The eight-step chart provides a model of the Predictive Coding 4.0 methods.

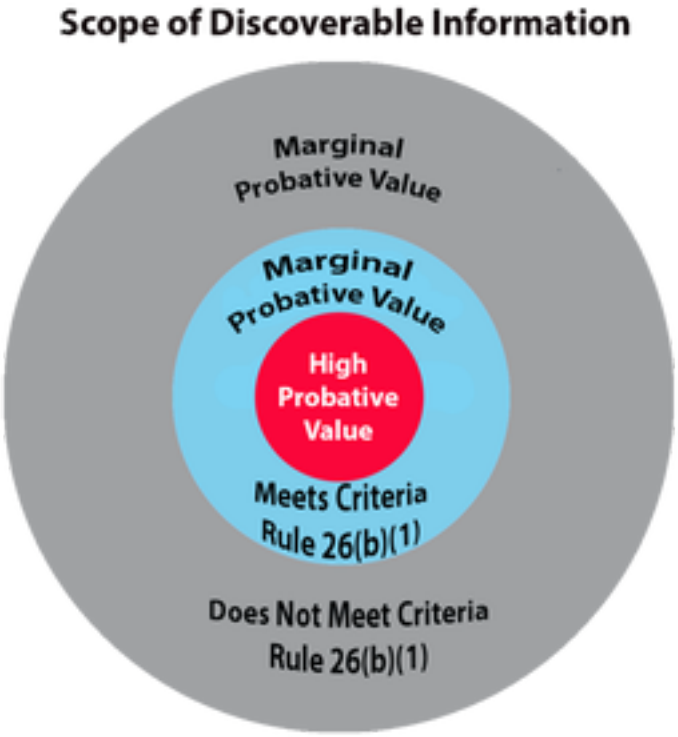


The circular flows depict the iterative steps specific to the predictive coding features. Steps four, five and six iterate until the active machine training reaches satisfactory levels and thereafter final quality control and productions are done.

Although presented as sequential steps for pedantic purposes, Predictive Coding 4.0 is highly adaptive to circumstances and does not necessarily follow a rigid linear order. For instance, some of the quality control procedures are used throughout the search and review, and rolling productions can begin at any time.

Step One - ESI Communications

Good review projects begin with *ESI Communications*, they begin with *talking*. You need to understand and articulate the disputed issues of fact. If you do not know what you are looking for, you will never find it. That does not mean you know of specific documents. If you knew that, it would not be much of a search. It means you understand what needs to be proven at trial and what documents will have impact on judge and jury. It also means you know the legal bounds of relevance, including especially Rule 26(b)(1).



ESI Communications begin and end with the *scope* of the discovery, relevance and related review procedures. The communications are not only with opposing counsel or other requesting parties, but also with the client and the e-discovery

team assigned to the case. These *Talks* should be facilitated by the lead *e-Discovery specialist attorney* assigned to the case. But they should include the active participation of the whole team, including all trial lawyers not otherwise very involved in the ESI review.

The purpose of all of this *Talk* is to give everyone an idea as to the documents sought and the confidentiality protections and other special issues involved. Good lines of communication are critical to that effort. This first step can sometimes be difficult, especially if there are many new members to the group. Still, a common understanding of relevance, the target searched, is critical to the successful outcome of the search. This includes the shared wisdom that the understanding of relevance will evolve and grow as the project progresses.

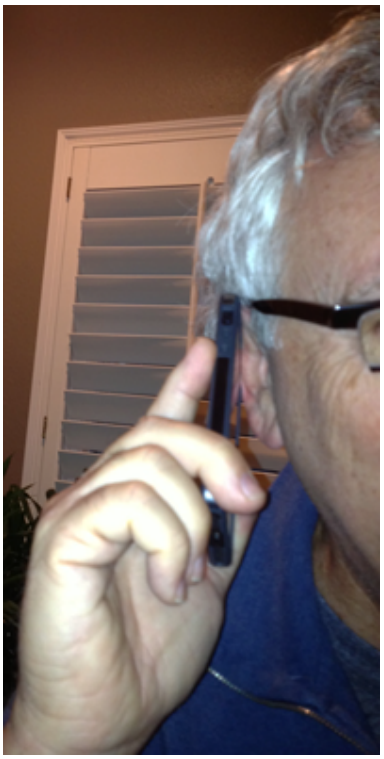
We need to *Talk* to understand what we are looking for. What is the target? What is the information need? What documents are relevant? What would a *hot document* look like? A common understanding of relevance by a review team, of what you are looking for, requires a lot of communication. *Silent review projects are doomed to failure*. They tend to stagnate and do not enjoy the benefits of *Concept Drift*, where a team's understanding of relevance is refined and evolves as the review progresses. Yes, the target may move, and that is a good thing. *See: Concept Drift and Consistency: Two Keys To Document Review Quality* – Parts One, Two and Three.

Review projects are also doomed where the communications are one way, lecture down projects where only the SME talks. The reviewers must talk back, must ask questions. The input of reviewers is key. Their questions and comments are very important. Dialogue and active listening are required for all review projects, including ones with predictive coding.

You begin with analysis and discussions with your client, your internal team, and then with opposing counsel, as to what it is you are looking for and what the requesting party is looking for. The point is to clarify the information sought, the *target*. You cannot just stumble around and hope you will know it when you find it (and yet this happens all too often). You must first know what you are looking for. The target of most searches is the information relevant to disputed issues of fact in a case or investigation. But what exactly does that mean? If you encounter unresolvable disputes with opposing counsel on the scope of relevance, which can happen during any stage of the review despite your best efforts up-front, you may have to include the Judge in these discussions and seek a ruling.

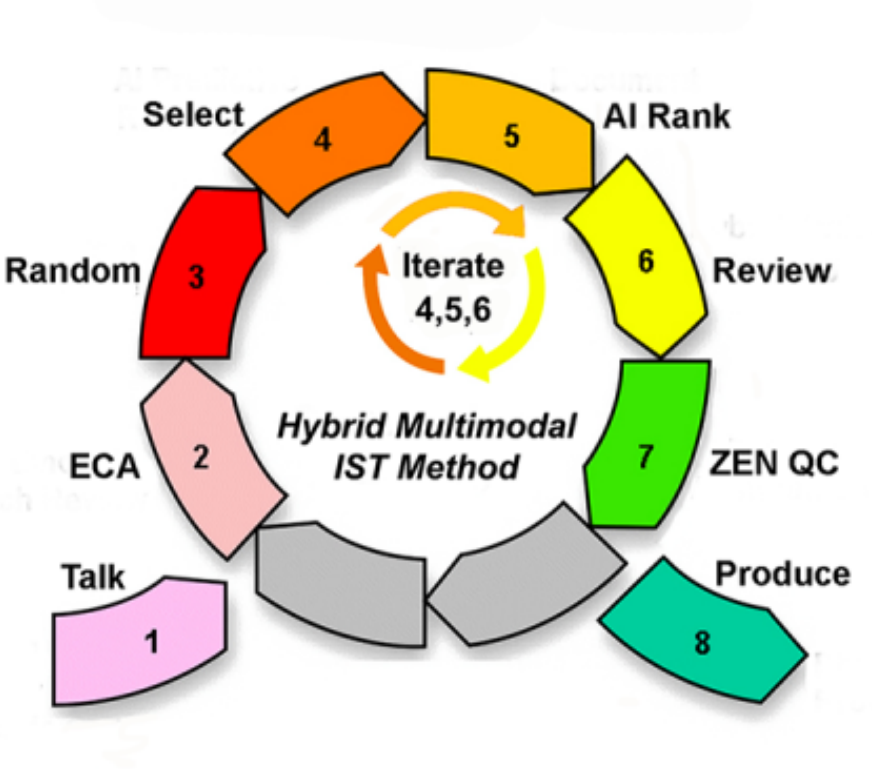
"*ESI Discovery Communications"* is about talking to your review team, including your client, key witnesses; it is about talking to opposing counsel; and, eventually, if need be, talking to the judge at hearings. Friendly, informal talk is a good method to avoid the tendency to polarize and demonize "the other side," to build walls and be *distrustful* and silent.

The amount of distrust today between attorneys is at an all-time high. This trend must be reversed. Mutually respectful *talk* is part of the solution. Slowing things down helps too. Do not respond to a provocative text or email until you calm down. Take your time to ponder any question, even if you are not upset. Take your time to research and consult with others first. This point is critical. The demand for instant answers is never justified, nor required under the rules of civil procedure. Think first and never respond out of anger. We are all entitled to mutual respect. You have a right to demand that. So do *they*.



This point about not actually speaking with people in realtime, in person, or by phone or video, is, to some extent, generational. Many younger attorneys seem to have an inherent loathing of the phone and speaking out loud. They let their thumbs do the talking. (This is especially true in e-discovery where the professionals involved tend to be very computer oriented, not people oriented. I know because I am like that.) Meeting in person in real-time is distasteful to many, not just Gen X. Many of us prefer to put everything in emails and texts and tweets and posts, etc. That may make it easier to pause to reflect, especially if you are loathe to say in person that *you do not know and will need to get back to them on that*. But real time talking is important to full communication. You may need to force yourself to real-time interpersonal interactions. Many people are better at real-time talk than others, just like many are better at fast comprehension of documents than others. It is often a good idea for a team to have a designated talker, especially when it comes to speaking with opposing counsel or the client.

In e-discovery, where the knowledge levels are often extremely different, with one side knowing more about the subject than the other, the fist step of *ESI Communications* or *Talk* usually requires patient explanations. *ESI Communications* often require some amount of educational efforts by the attorneys with greater expertise. The trick is to do that without being condescending or too pedantic, and, in my case at least, without losing your patience.



Some object to the whole idea of *helping* opposing counsel by educating them, but the truth is, this helps your clients too. You are going to have to explain everything when you take a dispute to the judge, so you might as well start upfront. It helps save money and moves the case along. Trust building is a process best facilitated by honest, open *talk*.

I use of the term *Talk* to invoke the term *listen* as well. That is one reason we also refer to the first step as "*Relevance Dialogues*" because that is exactly what it should be, a back and forth exchange. Top down lecturing is not intended here. Even when a judge talks, where the relationship is truly top down, the judge always listens before rendering his or her decision. You are given the right to be heard at a hearing, to talk and be listened to. Judges listen a lot and usually ask many questions. Attorneys should do the same. Never just talk to hear the sound of your own voice. As Judge David Waxse likes to say, talk to opposing counsel as if the judge were listening (or watching a video tape of the conference).

The same rules apply when communicating about discovery with the judge. I personally prefer in-person hearings, or at least telephonic, as opposed to just throwing memos back and forth. This is especially true when the memorandums

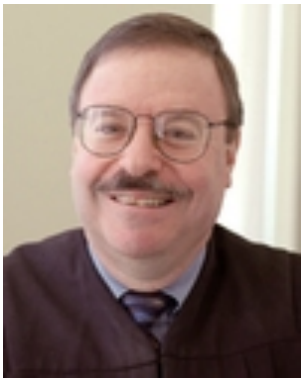
have very short page limits. ***Dear Judges***: e-discovery issues are important and can quickly spiral out of control without your prompt attention. Please give us the hearings and time needed. Issuing easy orders that just *split the baby* will do nothing but pour gas on a fire.

In my many years of lawyering I have found that hearings and meetings are much more effective than exchanging papers. ***Dear brothers and sisters in the BAR***: stop hating, stop distrusting and vilifying, and start talking to each other. That means listening too. Understand the other-side. Be professional. Try to cooperate. And stop taking extreme positions that assume the judge will just *split the baby*.

It bears emphasis that by *Talk* in this first step we intend *dialogue*: a true back and forth. We do not intend argument, nor winners and losers. We do intend mutual respect. That includes respectful disagreement, but only after we have heard each other out and understood our respective positions. Then, if our talks with *the other side* have reached an impasse, at least on some issues, we request a hearing from the judge and set out the issues for the judge to decide. That is how our system of justice and discovery are designed to work. If you fail to talk, you not only doom the document review project, you doom the whole case to unnecessary expense and frustration.

This dialogue method is based on a Cooperative approach to discovery that was promoted by the late, great Richard Braman of *The Sedona Conference*. Cooperation is not only a *best practice*, but is, to a certain extent, a minimum standard required by rules of professional ethics and civil procedure. The primary goal of these dialogues for document review purposes is to obtain a common understanding of the e-

discovery requests and reach agreement on the scope of relevancy and production.



ESI Communications in this first step may, in some cases, require disclosure of the actual search techniques used, which is traditionally protected by work product. The disclosures may also sometimes include limited disclosure of some of the training documents used, typically just the relevant documents. *See* Judge Andrew Peck's 2015 ruling on predictive coding, *Rio Tinto v. Vale,* 2015 WL 872294 (March 2, 2015, SDNY). In *Rio Tinto* Judge Peck wisely modified somewhat his original views stated in *Da Silva* on the issue of disclosure. *Moore v. Publicis Groupe*, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012) (approved and adopted in *Da Silva Moore v. Publicis Groupe*, 2012 WL 1446534, at *2 (S.D.N.Y. Apr. 26, 2012)). Judge Peck no longer thinks that parties should necessarily disclose any training documents, and may instead:

> ... insure that training and review was done appropriately by other means, such as statistical estimation of recall at the conclusion of the review as well as by whether there are gaps in the production, and quality control review of samples from the documents categorized as non-responsive. *See generally* Grossman & Cormack, *Comments, supra*, 7 Fed. Cts. L.Rev. at 301-12.
>
> The Court, however, need not rule on the need for seed set transparency in this case, because the parties agreed to a protocol that discloses all non-privileged documents in the control sets. (Attached Protocol, ¶¶ 4(b)-(c).) One point must be stressed -- it is inappropriate to hold TAR to a higher standard than keywords or manual review. Doing so discourages parties from using TAR for fear of spending more in motion practice than the savings from using TAR for review.

Id. at *3. *Also see* *Rio Tinto v. Vale, Stipulation and Order Re: Revised Validation and Audit Protocols for the use of Predictive Coding in Discovery*, 14 Civ. 3042 (RMB) (AJP), (order dated 9/2/15 by Maura Grossman, Special Master, and adopted and ordered by Judge Peck on 9/8/15).

Judge Peck here follows the current prevailing view on disclosure that I also endorse. Disclose the relevant documents used in active machine learning, but not the irrelevant documents used in training. If there are borderline, grey area documents classified as irrelevant, you may need to disclose these type of documents by description, not actual production. Again, talk to the requesting party on where you are drawing the line. Talk about the grey area documents that you encounter. If they disagree, ask for a ruling before your training is complete.

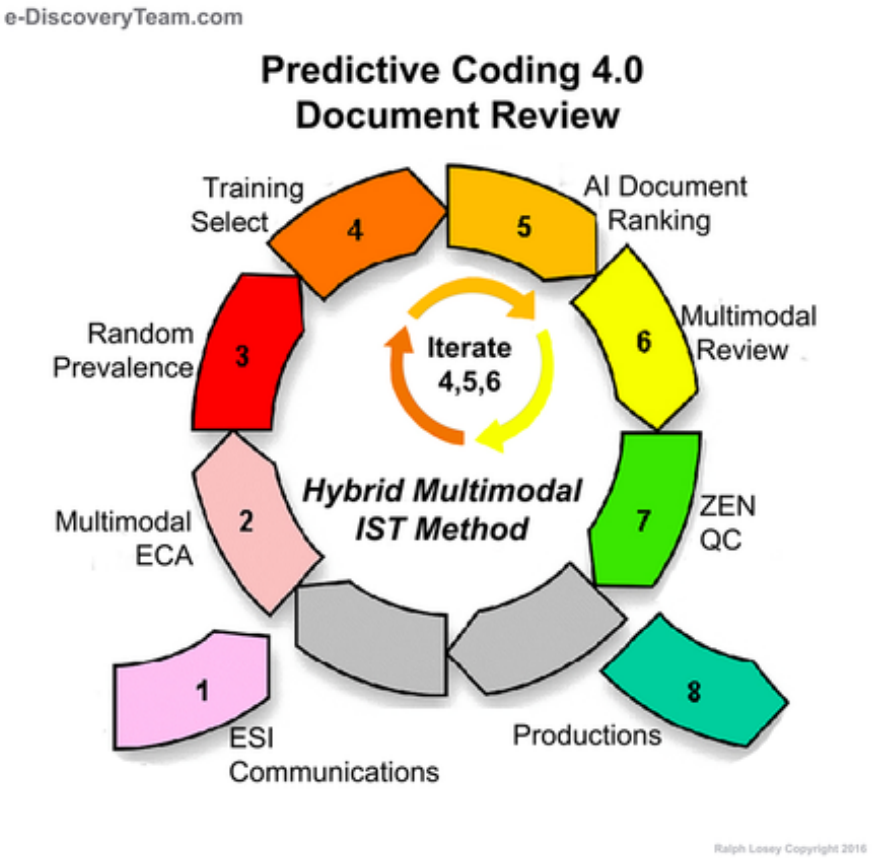
The goals of Rule 1 of the *Federal Rules of Civil Procedure* (just, speedy and inexpensive) are impossible in all phases of litigation, not just discovery, unless attorneys communicate with each other. The parties may hate each other and refuse to talk. That sometimes happens. But the attorneys must be above the fray. That is a key purpose and function of an attorney in a dispute. It is sad that so many attorneys do not seem to understand that. If you are faced with such an attorney, my best advice is to lead by example, document the belligerence and seek the help of your presiding judge.

Although *Talk* to opposing counsel is important, even more important is *talking* within the team. It is an important method of quality control and efficient project management. Everyone needs to be on the same page of relevance and discoverability. Work needs to be coordinated. Internal team *Talk* needs to be very close. Although a *Vulcan mind-meld* might be ideal, it is not really necessary. Still, during a project a steady flow of talk, usually in the form of emails or chats, is normal and efficient. Clients should never complain about time spent communicating to manage a document review project. It can save a tremendous amount of money in the long run, so long as it is focused on the task at hand.

Step Two - Multimodal ECA

Multimodal Early Case Assessment - ECA - summarizes the second step in our 8-step work flow. We used to call the second step "*Multimodal Search Review*." It is still the same activity, but we tweaked the name to emphasize the ECA

significance of this step. After we have an idea of what we are looking for from *ESI Communications* in step one, we start to use every tool at our disposal to try to find the relevant documents. Every tool that is, except for active machine learning. Our first look at the documents is *our look*, not the machine's. That is not because we do not trust the AI's input. We do. It is because there is no AI yet. The predictive coding only begins after you feed training documents into the machine. That happens in step four.



Our *Multimodal ECA* step-two does not take that long, so the delay in bringing in our AI is usually short. In our experiments at TREC in 2015 and 2016 under the auspicious of NIST, where we skipped steps three and seven to save time, and necessarily had little ESI Communications in step one, we would often complete simple document reviews of several hundred thousand documents in just a few hours. We cannot match these results in real-life legal document review projects because the issues in law suits are usually much more complicated than the issues presented by most

topics at TREC. Also, we cannot take the risk of making mistakes in a real legal project that we did in an academic event like TREC.

Again, the terminology revision to say Multimodal *ECA* is more a change of style than substance. We have always worked in this manner. The name change is just to better convey the idea that we are looking for the *low hanging fruit*, the easy to find documents. We are getting an initial assessment of the data by using all of the tools of the search pyramid except for the top tier active machine learning. The AI comes into play soon enough in steps four and five, sometimes as early as the same day.

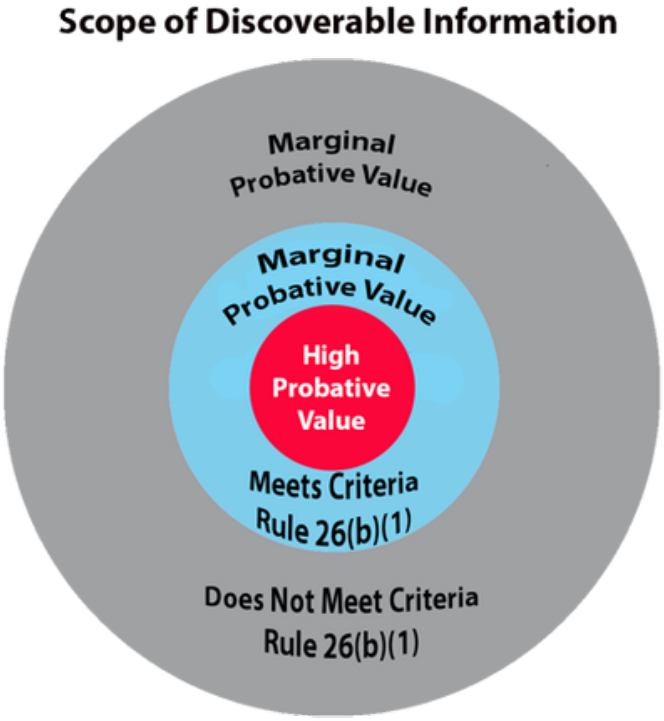
I have seen projects where key documents are found during the first ten minutes of looking around. Usually the secrets are not revealed so easily, but it does happen. Step two is the time to get to know the data, run some obvious searches, including any keyword requests for opposing counsel. You use the relevant and irrelevant documents you find in step two as the documents you select in step four to train the AI.

In the process of this initial document review you start to get a better understanding of the custodians, their data and relevance. This is what early case assessment is all about. You will find the rest of the still hidden relevant

documents in the iterated rounds of machine training and other searches that follow.

Although we speak of searching for *relevant* documents in step two, it is important to understand that many irrelevant documents are also incidentally found and coded in that process. Active machine learning does not work by training on relevant documents alone. It must also include examples of irrelevant documents. For that reason we sometimes actively search for certain kinds of irrelevant documents to use in training. One of our current research experiments with Kroll Ontrack is to determine the best ratios between relevant and irrelevant documents for effective document ranking. *See* TREC reports at Mr. EDR as updated from time to time. At this point we have that issue nailed.

The multimodal ECA review in step two is carried out under the supervision of the *Subject Matter Experts* on the case. They make final decisions where there is doubt concerning the relevance of a document or document type. The SME role is typically performed by a team, including the partner in charge of the case - the senior SME - and senior associates, and e-Discovery specialist attorney(s) assigned to the case. It is, or should be, a team effort, at least in most large projects. As previously described, the final arbitrator on scope is made by the senior SME, who in turn is acting as the predictor of the court's views. The final, final authority is always the Judge. As discussed before the chart below summarizes the analysis of the SME and judge on the discoverability of any document.



When I do a project, acting as the e-Discovery specialist attorney for the case, I listen carefully to the trial lawyer SME as he or she explains the case. By extensive Q&A the members of the team understand what is relevant. We learn from the SME. It is not exactly a Vulcan mind-meld, but it can work pretty well with a cohesive team. Most trial lawyers love to teach and opine on relevance and their theory of the case.

Although a good SME team communicates and plans well, they also understand, typically from years of experience, that the intended relevance scope is like a battle plan before the battle. As the famous German military strategist, General *Moltke the Elder* said: *No battle plan ever survives contact with the enemy*. So too no relevance scope plan ever survives contact with the corpus of data. The understanding of relevance will *evolve* as the documents are studied, the evidence is assessed, and understanding of what really happened matures. If not, someone is not paying attention. In litigation that is usually a recipe for defeat. See *Concept Drift and Consistency: Two Keys To Document Review Quality* - Parts One, Two and Three.

The SME team trains and supervises the document review specialists, aka, contract review attorneys, who usually then do a large part of the manual reviews (step-six), and few if any searches. Working with review attorneys is a constant iterative process where communication is critical. Although I sometimes use an *army-of-one* approach where I do everything myself (that is how I did the EDI Oracle competition and most of the TREC topics), my preference now is to use two or three reviewers to help with the document review. With good methods, including culling methods, and good software, it is rarely necessary to use more reviewers than that. With the help of strong AI, say that included in Mr. EDR, we can easily classify a million or so documents for relevance with that size team. More reviewers than that may well be needed for complex redaction projects and other production issues, but not for a well-designed first-pass relevance search.

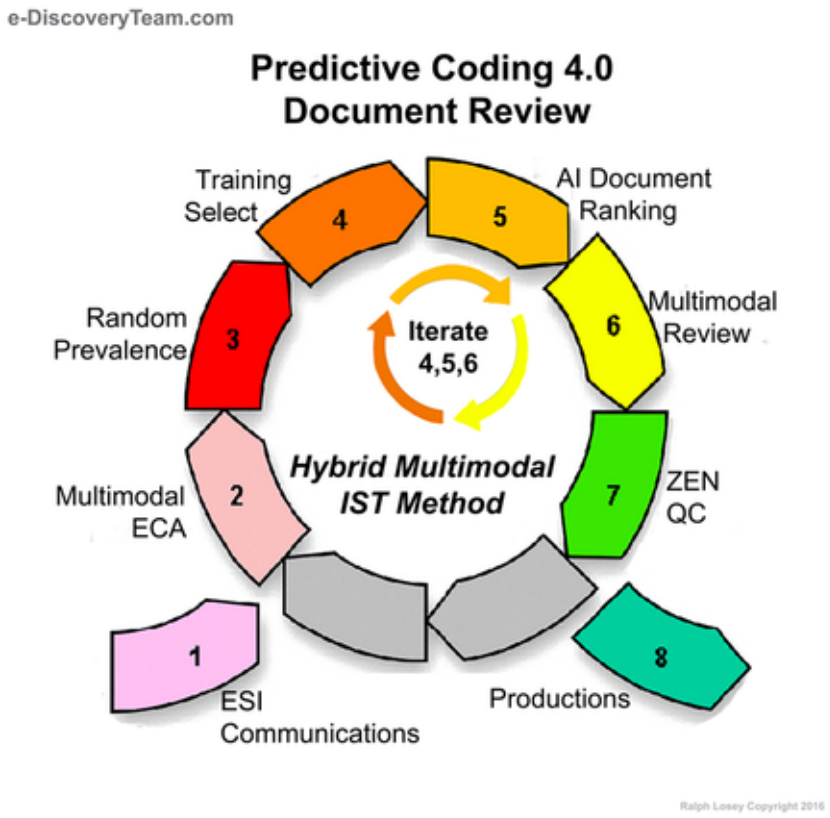
One word of warning when using document reviewers, it is very important for all members of the SME team to have direct and substantial contact with the actual documents, not just the reviewers. For instance, *everyone* involved in the project should see all hot documents found in any step of the process. It is especially important for the SME trial lawyer at the top of the expert pyramid to see them,

but that is rarely more than a few hundred documents, often just a few dozen. Otherwise, the top SME need only see the novel and grey area documents that are encountered, where it is unclear on which side of the relevance line they should fall in accord with the last instructions. Again, the burden on the senior, and often technologically challenged senior SME attorneys, is fairly light under these *Version 4.0* procedures.

The SME team relies on a primary SME, who is typically the trial lawyer in charge of the whole case, including all communications on relevance to the judge and opposing counsel. Thereafter, the head SME is sometimes only consulted on an *as-needed basis* to answer questions and make specific decisions on the grey area documents. There are always a few uncertain documents that need elevation to confirm relevance, but as the review progresses, their number usually decreases, and so the time and attention of the senior SME decreases accordingly.

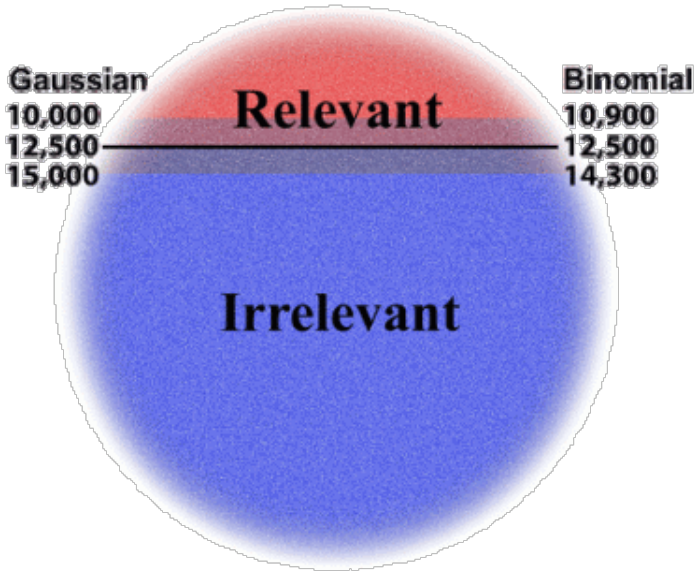
Step Three - Random Prevalence

There has been no change in this step from *Version 3.0* to *Version 4.0*. The third step, which is not necessarily chronological, is essentially a computer function with statistical analysis. Here you create a random sample and analyze the results of expert review of the sample. Some review is thus involved in this step and you have to be very careful that it is correctly done. This sample is taken for statistical purposes to establish a baseline for quality control in step seven. Typically prevalence calculations are made at this point. Some software also uses this random sampling selection to create a *control set*. As explained at length in Predictive Coding 3.0, we do not use a control set because it is so unreliable. It is a complete waste of time and money and does not produce reliable recall estimates. Instead, we take a random sample near the beginning of a project *solely* to get an idea on *Prevalence,* meaning the approximate number of relevant documents in the collection.



Unless we are in a very rushed situation, such as in the TREC projects, where we would do a complete review in a day or two, or sometimes just a few hours, we like to take the time for the sample and prevalence estimate.

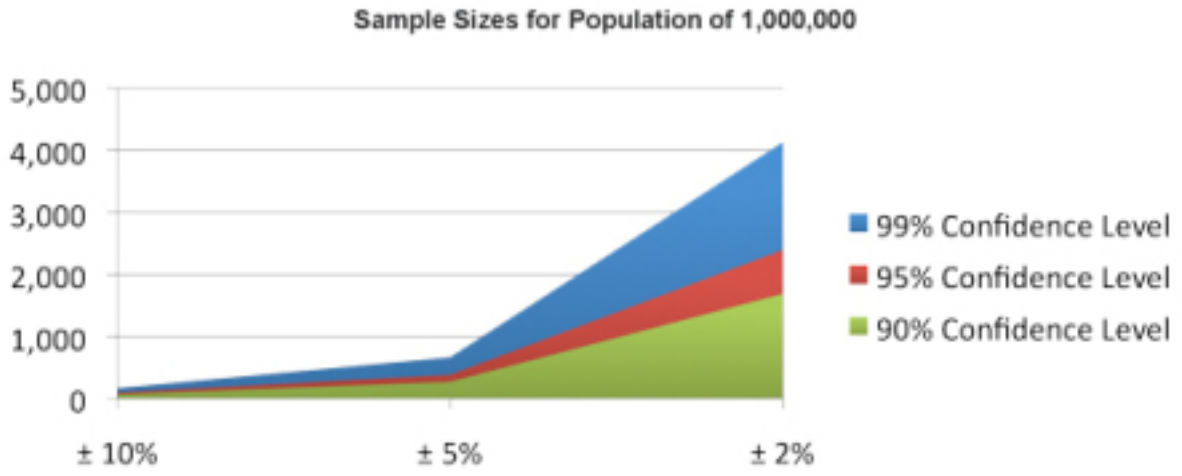
It is all about getting a statistical idea as to the range of relevant documents that likely exist in the data collected. This is very helpful for a number of reasons, including proportionality analysis (importance of the ESI to the litigation and cost estimates) and knowing when to stop your search, which is part of step seven. Knowing the number of relevant documents in your dataset can be very helpful, even if that number is a range, not exact. For example, you can know from a random sample that there are between four thousand and six thousand relevant documents. You cannot know there are exactly five thousand relevant documents. *See*: *In Legal Search Exact Recall Can Never Be Known*. Still, knowledge of the range of relevant documents (red in the diagram below) is helpful, albeit not critical to a successful search.



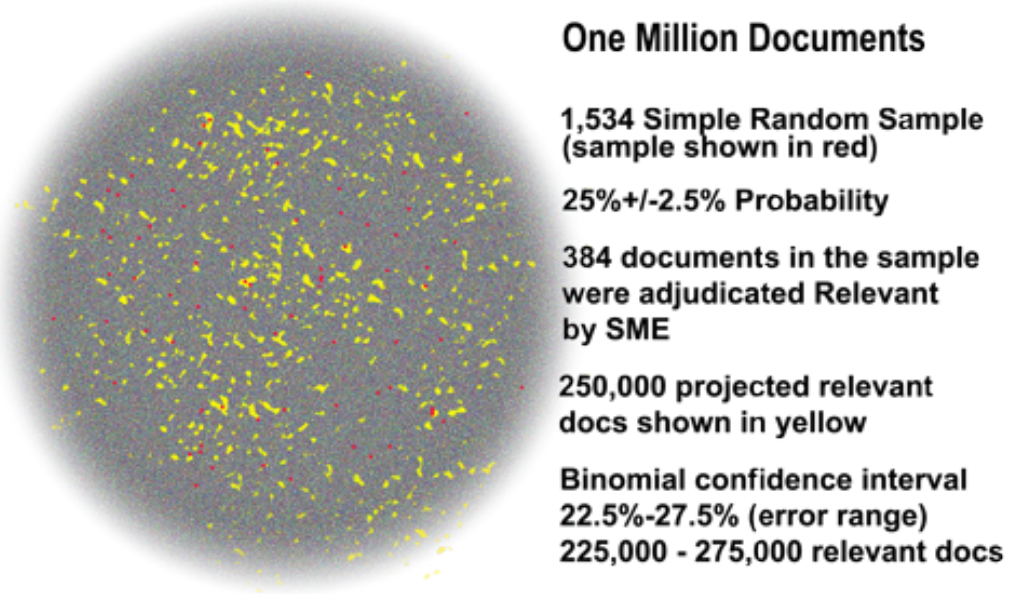
In step three an SME is only needed to verify the classifications of any grey area documents found in the random sample. The random sample review should be done by one reviewer, typically your best contract reviewer. They should be instructed to code as Uncertain any documents that are not obviously relevant or irrelevant based on their instructions and step one. All relevance codings should be double checked, as well as Uncertain documents. The senior SME is only consulted on an as-needed basis.

Document review in step three is limited to the sample documents. Aside from that, this step is a computer function and mathematical analysis. Pretty simple after you do it a few times. If you do not know anything about statistics, and your vendor is also clueless on this (rare), then you might need a consulting statistician. Most of the time this is not necessary and any competent *Version 4.0* vendor expert should be able to help you through it.

It is not important to understand all of the math, just that random sampling produces a ***range***, not an exact number. If your sample size is small, then the range will be very high. If you want to reduce your range in half, which is a function in statistics known as a *confidence interval*, you have to quadruple your sample size. This is a general rule of thumb that I explained in tedious mathematical detail several years ago in *Random Sample Calculations And My Prediction That 300,000 Lawyers Will Be Using Random Sampling By 2022*. Our Team likes to use a fairly large sample size of about 1,533 documents that creates a confidence interval of plus or minus 2.5%, subject to a *confidence level* of 95% (meaning the true value will lie within that range 95 times out of 100). More information on sample size is summarized in the graph below. *Id.*

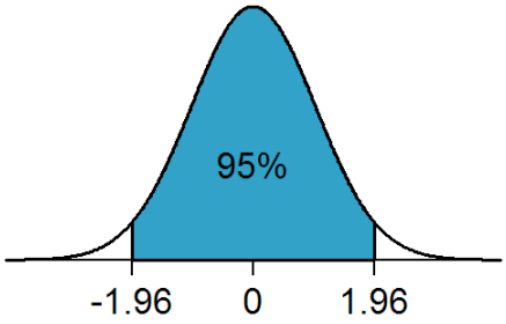


The picture below this paragraph illustrates a data cloud where the yellow dots are the sampled documents from the grey dot total, and the hard to see red dots are the relevant documents found in that sample. Although this illustration is from a real project we had, it shows a dataset that is unusual in legal search because the prevalence here was high, between 22.5% and 27.5%. In most data collections searched in the law today, where the custodian data has not been filtered by keywords, the prevalence is far less than that, typically less than 5%, maybe even less that 0.5%. The low prevalence increases the range size, the uncertainties, and requires a binomial calculation adjustment to determine the statistically valid confidence interval, and thus the true document range.

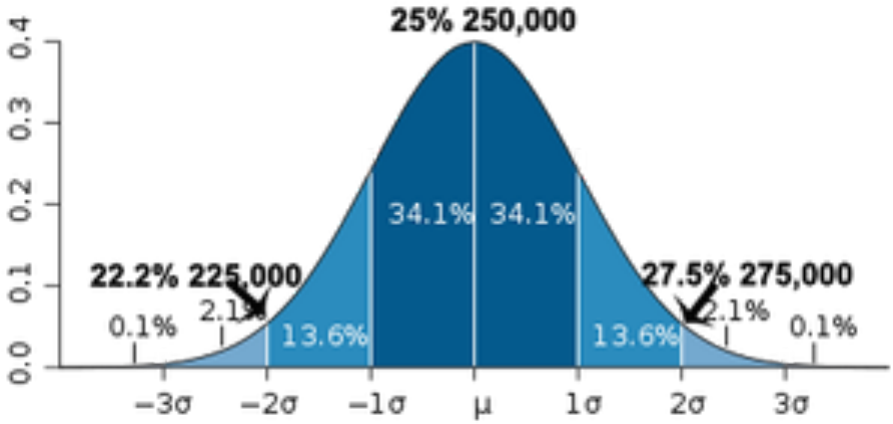


For example, in a typical legal project with a few percent prevalence range, it would be common to see a range between 20,000 and 60,000 relevant documents in a 1,000,000 collection. Still, even with this very large range, we find it useful to at least have *some idea* of the number of relevant documents that we are looking for. That is what the Baseline step can provide to you, nothing more nor less.

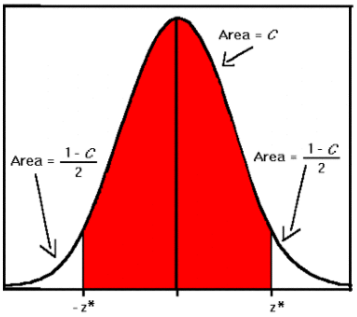
As mentioned, your vendor can probably help you with these statistical estimates. Just do not let them tell you that it is one *exact* number. It is always a range. The one number approach is just a shorthand for the range. It is simply a *point* projection near the middle of the range. The one number point projection is the top of the typical probability bell curve range shown right, which illustrates a 95% confidence level distribution. The top is just one possibility, albeit slightly more likely than either end points. The true value could be *anywhere* in the blue range.



To repeat, the step three prevalence baseline number is *always a range,* never just one number. Going back to the relatively high prevalence example, the below bell cure shows a point projection of 25% prevalence, with a range of 22.2% and 27.5%, creating a range of relevant documents of from between 225,000 and 275,000. This is shown below.



The important point that many vendors and other "experts" often forget to mention is that you can never know exactly where within that range the true value may lie. Plus, there is always a small possibility, 5% when using a sample size based on a 95% confidence level, that the true value may fall *outside* of that range. It may, for example, only have 200,000 relevant documents. This means that



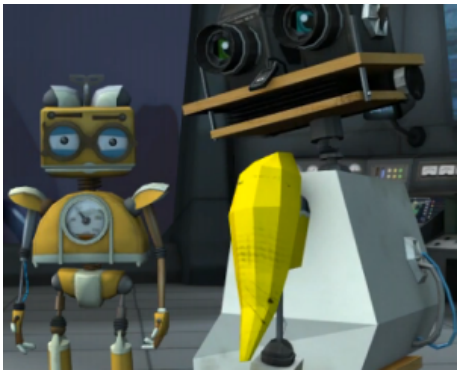
even with a high prevalence project with datasets that approach the Normal Distribution of 50% (here meaning half of the documents are relevant), you can never know that there are exactly 250,000 documents, just because it is the mid-point or point projection. You can only know that there are between 225,000 and 275,000 relevant documents, and even that range may be wrong 5% of the time. Those uncertainties are inherent limitations to random sampling.

Shame on the vendors who still perpetuate that myth of certainty. Lawyers can handle the truth. We are used to dealing with uncertainties. All trial lawyers talk in terms of probable results at trial, and risks of loss, and often calculate a case's settlement value based on such risk estimates. Do not insult our intelligence by a simplification of statistics that is plain wrong. Reliance on such erroneous point projections alone can lead to incorrect estimates as to the level of recall that we have attained in a project. We do not need to know the math, but we do need to know the truth.

I have previously written extensively on this subject. See for instance:

- *In Legal Search Exact Recall Can Never Be Known*
- *Random Sample Calculations And My Prediction That 300,000 Lawyers Will Be Using Random Sampling By 2022*
- *Borg Challenge: Part Two where I begin the search with a random sample* (text and video)

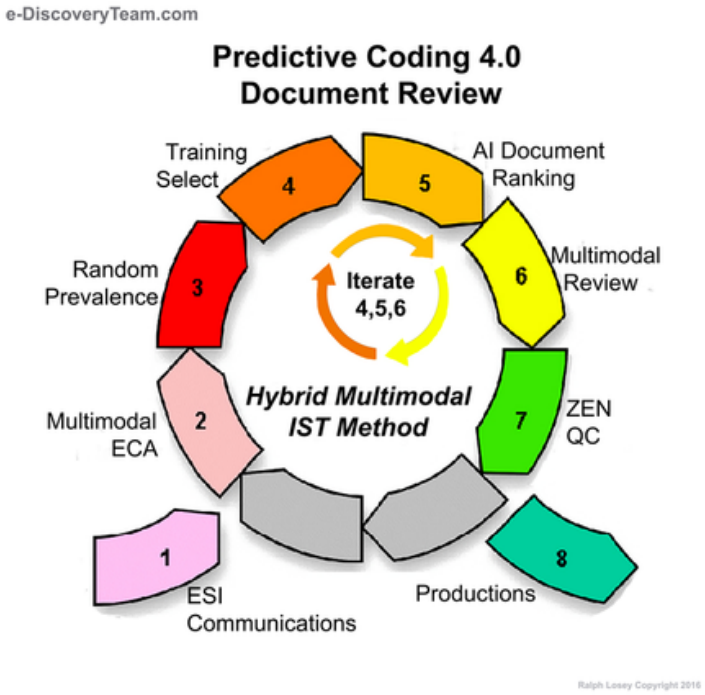
If you prefer to learn stuff like this by watching cute animated robots, then you might like: *Robots From The Not-Too-Distant Future Explain How They Use Random Sampling For Artificial Intelligence Based Evidence Search*. But be careful, their view is version 1.0 as to control sets.



Thanks again to William Webber and other scientists in this field who helped me out over the years to understand the *Bayesian* nature of statistics (and reality).

PART SEVEN

This is the concluding segment of the Team's description of its method of electronic document review using Predictive Coding. We have already covered the nine insights and the first three steps in our slightly revised eight-step workflow. We will now cover the remaining five steps.



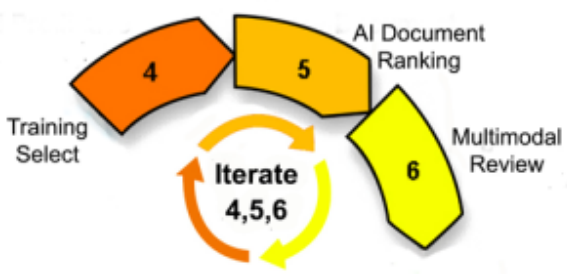
Steps Four, Five and Six - Training Select, AI Document Ranking and Multimodal Review

These are the three iterated steps that are the heart of our active machine learning process. The description of steps four, five and six constitutes the most significant change, although the content of what we actually do has not changed much. We have changed the iterated steps order by making a new step four - *Training Select*. We have also changed somewhat the descriptions in *Predictive Coding Version 4.0*. This was all done to better clarify and simplify what we are doing. This is our standard work flow. Our old description now seems somewhat confusing. As Steve Jobs famously said:

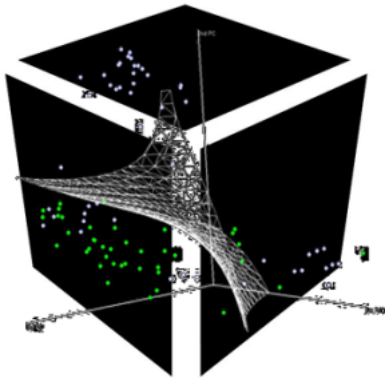
> You have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains.

In our case it can help you to move mountains of data by proper use of active machine learning.

In version 3.0 we called these three iterated steps: AI Predictive Ranking (step 4), Document Review (step 5), and Hybrid Active Training (step 6). The AI Predictive Ranking step, now called *AI Document Ranking*, was moved from step four to step five. This is to clarify that the task of



selecting documents for training always comes *before* the training itself. We also made *Training Selection* a separate step to emphasize the importance of this task. This is something that we have come to appreciate more fully over the past year.



The AI Document ranking step is where the computer does its thing. It is where the algorithm goes into action and ranks all of the documents according to the training documents selected by the humans. It is the unique AI step: the famous *black box*. No human efforts in step five at all. All we do is wait on the machine analysis. When it is done, all documents have been ranked (first time) or reranked (all training rounds after the first). We slightly tweaked the name here to be *AI Document Ranking*, instead of *AI Predictive Ranking*, as that is, we think, a clearer description of what the machine is doing. It is ranking all documents according to probability of relevance, or whatever other binary training you are doing. For instance, we usually also rank all documents according to probable privilege too and also according to high relevance.

Our biggest change here in version 4.0 is to make this AI step number five, instead of four, and, as mentioned, to add a new step four called *Training Select*. The new step four - *Training Select* - is the human function of deciding what documents to use to train the machine. (This used to be included in iterated step six, which was, we now see, somewhat confusing.) Unlike other predictive coding methods, we empower humans to make this selection in step four, *Training Select*. We do not, like some methods, create automatic rules for selection of training documents. For example, the Grossman Cormack *CAL* method (their trademark) only uses a predetermined number of the top ranked documents for training. In our method, we could also select these top ranked documents, or we could include other documents we have found to be relevant from other methods.

The freedom and choices that our method provides to the humans in charge is another reason our method is called *Hybrid*, in that it features natural human intelligence. It is not all machine controlled. In *Predictive Coding 4.0* we use artificial intelligence to enhance or augment our own natural intelligence. The machine is our partner, our friend, not our competitor or enemy. We tell our tool, our computer algorithm, what documents to train on in step four, and when, and the machine implements in step five.

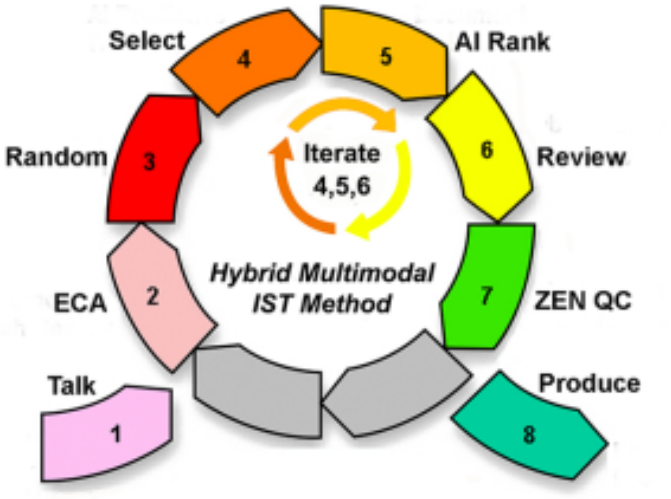
Typically in step four, *Training Select*, we will include *all documents* that we have previously coded as *relevant* as training documents, but not always. Sometimes, for instance, we may defer including very long relevant documents in the training, especially large spreadsheets, until the AI has a better grasp of our relevance intent. Skilled searchers rarely use *all* documents coded as training documents, but sometimes do. The same reasoning may apply to excluding a very short message, such as a one-word message saying "call," although we are more likely to leave that in. This selection process is where the art and experience of search come in. The concern is to avoid over-training on any one document type and thus lowering recall and missing a key black-swan document.

Also, we now rarely include *all* irrelevant documents into training, but instead used a balanced approach. Otherwise we tend to see incorrectly low rankings cross the board. The 50% plus dividing line can be an inaccurate indicator of probable relevant. It may instead go down to 40%, or even lower. We also find the balanced approach allows the machine to learn faster. Information scientists we have spoken with on this topic say this is typical with most types of active machine learning algorithms. It is not unique to our Mr. EDR, an active machine learning algorithm that uses an *logistic regression* method.

The sixth step of Multimodal Review is where we find new relevant or irrelevant documents for the next round of training. This is the step where most of the actual document review is done, where the documents are seen and classified by human reviewers. It is thus like step two, multimodal ECA. But now in step six we can also performed *ranking searches*, such as find all documents ranked 90% probable relevant or higher. Usually we rely heavily on such ranking searches.

We then human review all of the documents, which can often include very fast skimming and bulk coding. In addition to these ranked searches for new documents to review and code, we can use any other type of search we deem appropriate. This is the multimodal approach. Typically keyword and concept searches are used less often after the first round of training, but similarity searches of all kinds are often used throughout a project to supplement ranking based searches. Sometimes we may even use a linear search, expert manual review at the base of the search pyramid, if a new hot document is found. For instance, it might be helpful to see all communications that a key witness had on a certain day. The two-word stand-alone *call me* email when seen in context can sometimes be invaluable to proving your case.

Step six is much like step two, *Multimodal ECA*, except that now new types of document ranking search are possible. Since the documents are now all probability ranked in step five, you can use this ranking to select documents for the next round of document review (step four). For instance, the research of Professors Cormack and Grossman has shown that selection of the highest ranked documents can be a very effective method to continuously find and train relevant documents. *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, SIGIR'14, July 6–11, 2014, at pg. 9. Also see *Latest Grossman and Cormack Study Proves Folly of Using Random Search for Machine Training* – Parts One, Two, Three and Four. Another popular method, also tested and reported on by Grossman and Cormack, is to select mid-ranked documents, the ones the computer is uncertain about. They are less fond of that method, and we are too, but we will sometimes use it too.



The e-Discovery team's preferred active learning process in the iterative machine learning steps of **Predictive Coding 4.0** is still four-fold, just as it was in version 3.0. It is multimodal. How you mix and match the search methods is a matter of personal preference and educated response to the data searched. Here are my team's current preferences for most projects. Again, the weight for each depends upon the project. The only constant is that more that one method is always used.

1. High Ranked Documents. My team will almost always look to see what the highest unreviewed ranked documents are after *AI Ranking*, step five. We agree with Cormack and Grossman that this is a very effective search. We may review them on a document by document basis, or only by spot-checking some of them. In the later spot-checking scenario, a quick review of a certain probable relevant range, say all documents ranked between 95% to 99.9% (Mr. EDR has no 100%), may show that they all seem obvious relevant. We may then bulk code all documents in that range as relevant without actually reviewing them. This is a very powerful and effective method with Mr. EDR, and other software, so long as care is used not to over-extend the probability range. In other situations, we may only select the 99%+ probable relevant set for checking and bulk coding with limited review. The safe range typically changes as the review evolves and your latest conception of relevance is successfully imprinted on the computer.

Note that when we say a document is selected without individual review - meaning no human actually read the document - that is only for purposes of training selection and identifying relevant documents for production. We sometimes call that *first pass review*. In real world projects for clients we always review each document found in steps four, five and six, that has not been previously reviewed by a human, *before* we produce the document. (This is not true in our academic or scientific studies for TREC or EDI/Oracle.) That takes place in the last step - step eight, *Productions*. To be clear, in legal practice we do not produce without human verification and review of each and every document produced. The stakes if an error is made are simply too high.

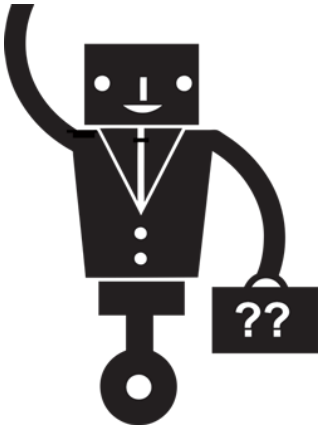
In our cases the most enjoyable part of the review project comes when we see from this search method that Mr. EDR has understood our training and has started to go beyond us. He starts to see patterns that we cannot.
He amazingly unearths documents that our team never thought to look for. The relevant documents he finds are sometimes dissimilar to any others found. They do not have the same key

words, or even the same known concepts. Still, Mr. EDR sees patterns in these documents that we do not. He finds the hidden gems of relevance, even outliers and black swans. That is when we think of Mr. EDR as going into *superhero mode*. At least that is the way my e-Discovery Team likes to talk about him.

By the end of most projects Mr. EDR attains a much 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 lightening fast and inexhaustible, even untrained, but by the end of his education, he becomes a genius. Definitely smarter and faster than any human *as to this one production review task*. Mr. EDR in that kind of *superhero mode* is what makes Predictive Coding so much fun. *See* *Why I Love Predictive Coding*.

Watching AI with higher intelligence than your own, intelligence which you created by your training, is *exciting*. More than that, the AI you created ***empowers*** you to do things that would have been impossible before, absurd even. For instance, using Mr. EDR, my e-Discovery Team of three attorneys was able to do 30 review projects and classify 16,576,820 documents in 45 days. *See* TREC 2015 experiment summary at Mr. EDR. This was a very gratifying feeling of empowerment, speed and augmentation of our own abilities. The high-AI experience comes though very clearly in the ranking of Mr. EDR near the end of the project, or really anytime before that, when he *catches on* to what you want and starts to find the hidden gems. I urge you all to give Predictive Coding a try so you can have this same kind of advanced AI hybrid excitement.

2. Mid-Ranked Uncertain Documents. We sometimes choose to allow the machine, in our case Mr. EDR, to select the documents for review in the sense that we review some of the mid-range ranked documents. These are documents where the software classifier is uncertain of the correct classification. They are usually in the 40% to 60% probable relevant range. Human guidance on these documents as to their relevance will sometimes help the machine to learn by adding diversity to the documents presented for review. This in turn also helps to locate outliers of a type the initial judgmental searches in step two and six may have missed. If a project is going well, we may not need to use this type of search at all.

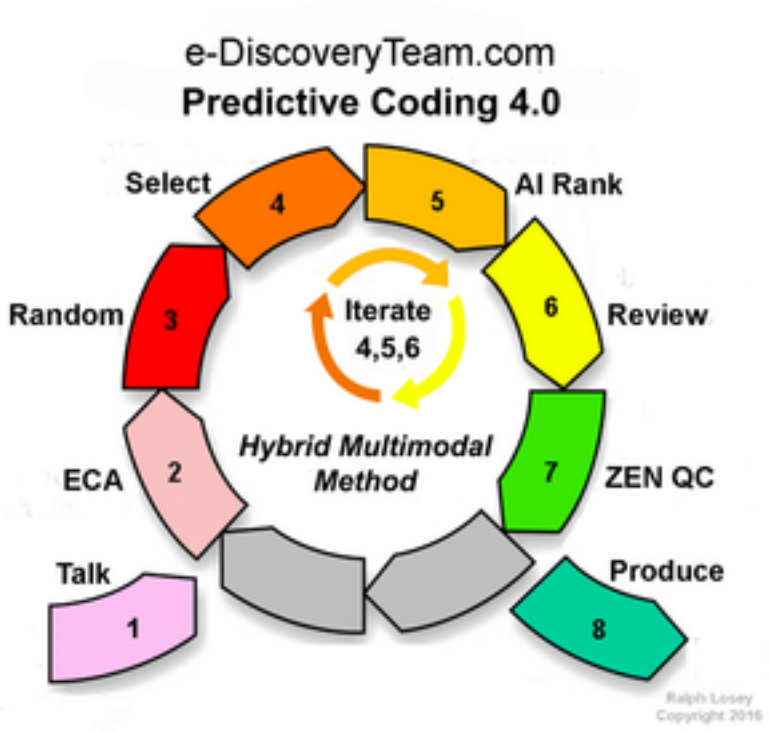


3. Random and Judgmental Sampling. We may also select some documents at random, either by proper computer random sampling or, more often, by informal random selection, including spot-checking. The later is sometimes called *judgmental sampling*. These sampling techniques can help maximize recall by avoidance of a premature focus on the relevant documents initially

retrieved. Random samples taken in steps three and six are typically also all included for training, and, of course, are always very carefully reviewed. The use of random selection for training purposes alone was minimized in *Predictive Coding 3.0* and remains of lower importance in version 4.0. With today's software, and using the multimodal method, it is not necessary. We did all of our TREC research without random sampling. We very rarely see the high-ranking searches become myopic without it. Plus, our multimodal approach guards against such over-training throughout the process.

4. Ad Hoc Searches Not Based on Document Ranking. Most of the time we supplement the machine's ranking-based-searches with additional search methods using non-AI based analytics. The particular search supplements we use depends on the relevant documents we find in the ranked document searches. The searches may include some linear review of selected custodians or dates, parametric Boolean keyword searches, similarity searches of all kinds, concept searches. We use every search tool available to us. Again, we call that a multimodal approach.

More on Step Six - Multimodal Review



As seen all types of search may be conducted in step six to find and batch out documents for human review and machine training. This step thus parallels step two, ECA, except that documents are also found by ranking of probable relevance. This is not yet possible in step two because step five of *AI Document Ranking* has not yet occurred.

It is important to emphasize that although we do searches in step six, steps six and eight are the steps where most of the actual document review is also done, where the documents are seen and classified by human reviewers. Search is used in step six to find the documents that human reviewers should review next. In my experience (and timed tests) the human document review can take as little as one-second per document, assuming your software is good and fast, and it is an obvious document, to as long as a half-hour. The lengthy time to review a document is very rare and only occurs where you have to fast-read a long document to be sure of its classification.

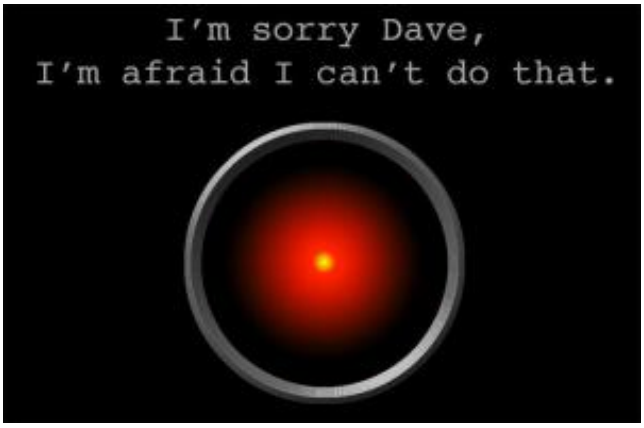
Step six is the human time intensive part of Predictive Coding 4.0 and can take most of the time in a project. Although when our top team members do a review, such as in TREC, we often spend more than half of the time in the other steps, sometimes considerably more.

Depending on the classifications during step six *Multimodal Review*, a document is either set for production, if relevant and not-privileged, or, if coded irrelevant, it is not set for production. If relevant and privileged, then it is logged but not produced. If relevant, not privileged, but confidential for some reason, then it is either redacted and/or specially labeled before production. The special labeling performed is typically to prominently affix the word CONFIDENTIAL on the Tiff image production, or the phrase CONFIDENTIAL - ATTORNEYS EYES ONLY. The actual wording of the legends depends upon the parties confidentiality agreement or court order.

When many redactions are required the total time to review a document can sometimes go way up. The same goes for double and triple checking of privileged documents that are sometimes found *in* document collections in large numbers. In our TREC and Oracle experiments redactions and privilege double-checking were not required. The time-consuming redactions are usually deferred to step eight - *Productions*. The equally as time-consuming privilege double-checking efforts can also be deferred to step seven - *Quality Assurance*, and again for a third-check in step eight.

When reviewing a document not already manually classified, the reviewer is usually presented with a document that the expert searcher running the project has determined is *probably* relevant. Typically this means that it has higher than a 50% probable relevance ranking. The reviewer may, or may not know the ranking. Whether you disclose that to a reviewer depends on a number of factors. Since I usually only use highly skilled reviewers, I trust them with disclosure. But sometimes you may not want to disclose the ranking.

During the review many documents predicted to be relevant will not be. The reviewers will code them correctly, as they see them. Our reviewers can and do disagree with and overrule the computer's predictions. The "Sorry Dave" phrase of the HAL 9000 computer in 2001 Space Odyssey is not possible. Although, our computer can argue back at us, we always have the final say.

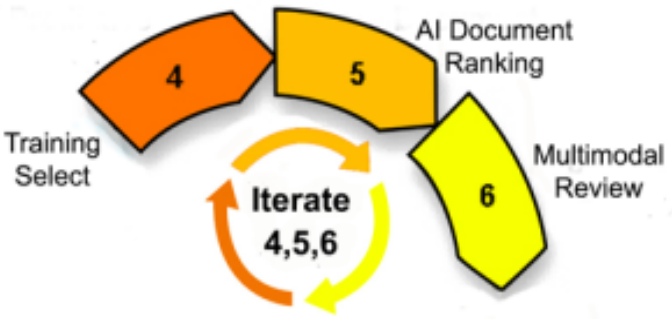


If a reviewer is in doubt on relevance, they consult the SME team. Furthermore, special quality controls in the form of second reviews may be imposed on Man Machine disagreements (the computer says a document should be relevant, but the human reviewer disagrees, and visa versa). They often involve close questions and the ultimate results of the resolved conflicts are typically used in the next round of training.

Sometimes the Machine will predict that a document is relevant, maybe even with 99.9% certainty, even though we have *already coded* the document as Irrelevant. (We review these again, even though they have been reviewed before.) It does so even though we have *already told* the Machine to train on it as irrelevant. The Machine does not care about your feelings! Or your authority as chief SME. It considers all of the input, all of your documents input in step four. If the cold, hard logic of its algorithms tells it that a document should be relevant, that is what it will report, in spite of how the document has already been coded. This is an excellent quality control tool.

I cannot tell you how impressed I was when that first happened to me. I was skeptical, but I went ahead and reread the long document anyway, this time more carefully. Sure enough, I had missed a paragraph near the end that made the document relevant. That was an Eureka moment for me. I have been a strong proponent of predictive coding ever since. Software does not get tried like we do. If the software is good it reads the whole document and is not front-loaded like we usually are. That does not mean Mr. EDR is always right. He is not. Most of the time we reaffirm the original coding, but not without a careful double-check. Usually we can see where the algorithm went wrong. Sometimes that influences our next iteration of step four, selection of training documents.

Prediction error type corrections such as this can be the focus of special searches in step six. Most quality **version 4.0** software such as Mr. EDR have search functions built-in that are designed to locate all such conflicts between document ranking and classification. Reviewers then review and correct the computer errors by a variety of methods, or change their own prior decisions. This often requires SME team involvement, but only very rarely the senior level SME.



The predictive coding software learns from all of the corrections to its prior predictive rankings. Steps 4, 5 and 6 then repeat as shown in the diagram. This iterative process is a *positive feedback loop* that continues until the computer predictions are accurate enough to satisfy the proportional demands of the case. In almost all cases that means you have found *more than enough* of the relevant documents needed to fairly decide the case. In many cases it is far better than that. It is routine for us to attain recall range levels of 90% or higher. In a few you may find almost all of the relevant documents, or at least all of the highly relevant documents.

General Note on Ease of Version 4.0 Methodology and Attorney Empowerment

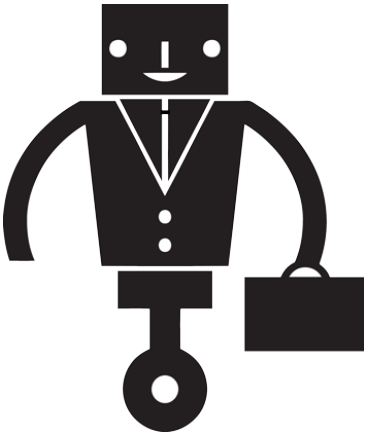
The machine training process for document review has become easier over the last few years as we have tinkered with and refined the methods. (*Tinkering* is the original and still only true meaning of *hacking*. *See*: HackerLaw.org) At this point of the predictive coding life cycle it is, for example, easier to learn how to do predictive coding than to learn how to do a trial - bench or jury. Interestingly, the most effective instruction method for both legal tasks is similar - second chair apprenticeship, watch and learn. It is the way complex legal practices have always been taught. My team can teach it to any smart tech lawyer by having them second chair a couple of projects.

It is interesting to note that medicine uses the same method to teach surgeons how to do complex robotic surgery, with a *da Vinci* surgical system, or the like. Whenever a master surgeon operates with robotics, there are always several doctors watching and assisting, more than are needed. In this photo they are the ones around the

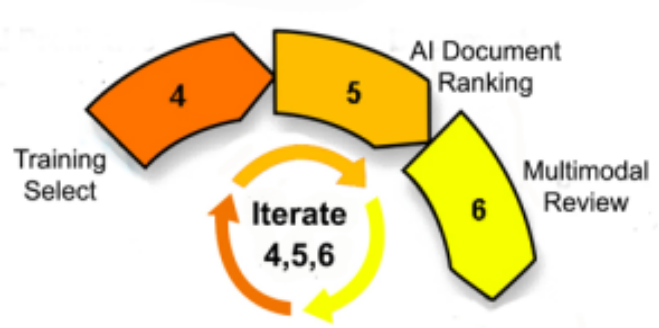
patient. The master surgeon who is actually controlling the tiny knifes in the patient is the guy on the far left sitting down with his head in the machine. He is looking at a magnified video picture of what is happening inside the patient's body and moving the tiny knives around with a joystick.

The hybrid human-robot system augments the human surgeon's abilities. The surgeon has his hands on the wheel at all times. The other doctors may watch dozens, and if they are younger, maybe even hundreds of surgeries before they are allowed to take control of the joy stick and do the hard stuff themselves. The predictive coding steps four, five and six are far easier than this, besides, if you screw up, nobody dies.

More on Step Five - AI Document Ranking



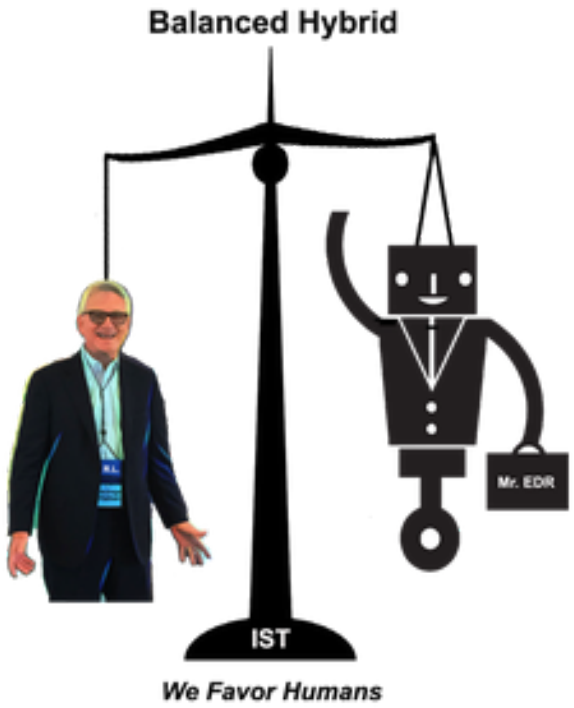
More discussion on step five may help clarify all three iterated steps. Again, step five is the AI Document Ranking step where the machine takes over and does all of the work. We have also called this the *Auto Coding Run* because this is where the software's predictive coding calculations are performed. The software we use is Kroll Ontrack's Mr. EDR. In the fifth step the software applies all of the training documents we selected in step four to sort the data corpus. In step five the human trainers can take a coffee break while Mr. EDR ranks all of the documents according to probable relevance or other binary choices.



The first time the document ranking algorithm executes is sometimes called the *seed set* run. The first *repetition* of the ranking step five is known as the *second* round of training, the next, the third round, etc. These iterations continue until the training is complete within the proportional constraints of the case. At that point the attorney in charge of the search may declare the search complete and ready for the next quality assurance test in Step Seven. That is called the *Stop decision*.

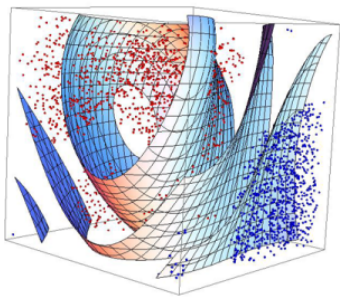
It is important to understand that this entire eight-step workflow diagram is just a linear two-dimensional representation of **Predictive Coding 4.0** for teaching purposes. These step descriptions are also a simplified explanation. Step Five can take place just a soon as a single document has been coded. You could have continuous, ongoing machine training at any time that the humans in charge decide to do so. That is the meaning of out team's IST (*Intelligently Spaced Training*), as opposed to Grossman and Cormack's trademarked CAL method, where the training always goes on without any human choice. This was discussed at length in Part Two of this article.

We space the training times ourselves to improve our communication and understanding of the software ranking. It helps us to have a better *intuitive grasp* of the machine processes. (Yes, such a thing is possible.) It allows us to observe for ourselves how a particular document, or usually a particular group of documents, impact the overall ranking. This is an important part of the *Hybrid* aspects of the **Predictive Coding 4.0 *Hybrid* IST Multimodal Method**. We like to be in control and to tell the machine exactly when and if to train, not the other way around. We like to understand what is happening and not just delegate everything to the machine. That is one reason we like to say that although we promote a balanced hybrid-machine process, we are pro-human and tip the scales in our favor.

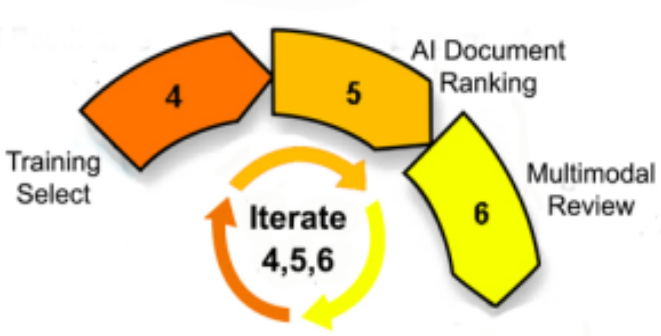


As stated, step five in the eight-step workflow is a purely algorithmic function. The ranking of a few million documents may take as long as an hour, depending on the complexity, the number of documents, software and other factors. Or it might just take a few minutes. This depends on the circumstances and tasks presented.

All documents selected for training in step four are included in step five computer processing. The software studies the documents marked for training, and then analyzes all of the data uploaded onto the *review platform*. It then ranks all of the documents according to probable relevance (and, as mentioned according to other binary categories too, such as Highly Relevant and



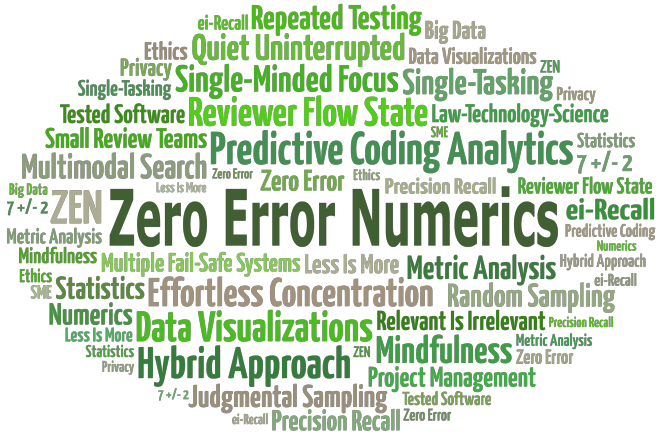
Privilege, and does all of these categories at the same time, but for simplicity purposes here we will just speak of the relevance rankings). It essentially assigns a probable value of from 0.01% to 99.9% probable relevance to each document in the corpus. (Note, some software uses different ranking values, but this is essentially what it is doing.) A value of 99.9% represents the highest probability that the document matches the category trained, such as relevant, or highly relevant, or privileged. A value of 0.01% means no likelihood of matching. A probability ranking of 50% represents equal likelihood, unless there has been careless over-training on irrelevance documents or other errors have been made. In the middle probability rankings the machine is said to be *uncertain* as to the document classification.



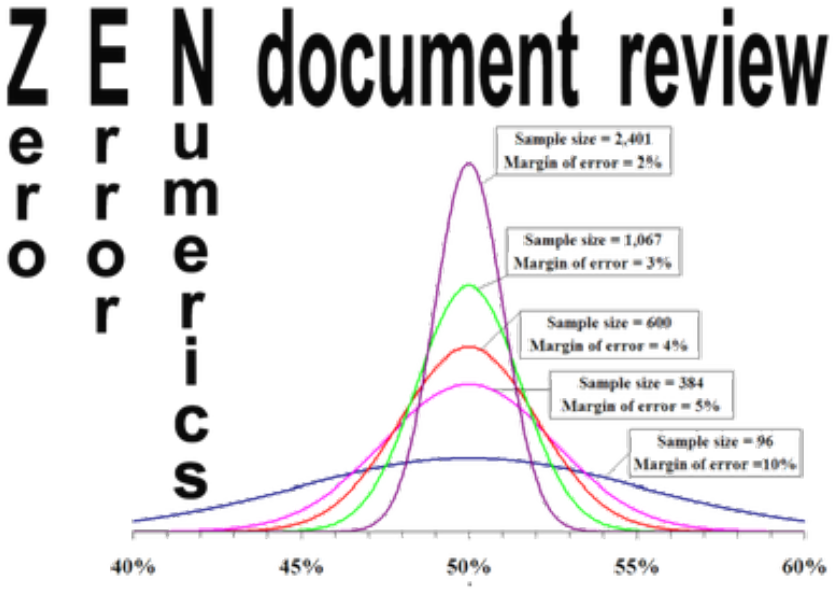
The first few times the AI-Ranking step is run the software predictions as to document categorization are often wrong, sometimes wildly so. It depends on the kind of search and data involved and on the number of documents already classified and included for training. That is why spot-checking and further training are always needed for predictive coding to work properly. That is why predictive coding is always an iterative process.

Step Seven: ZEN Quality Assurance Tests

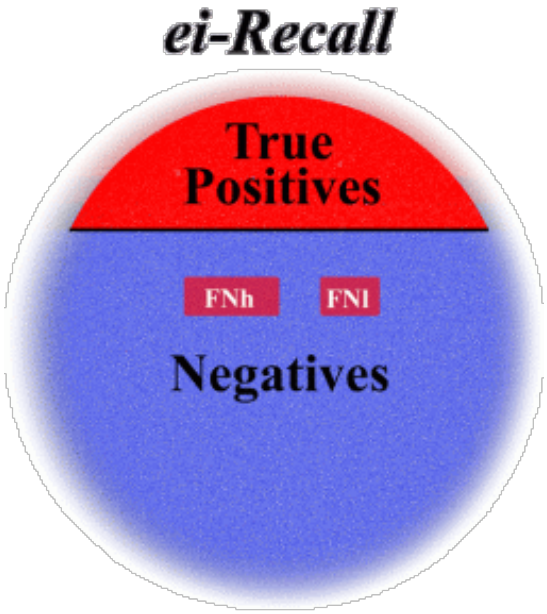
There has been no change in this step from *Version 3.0* to *Version 4.0*. If you already know 3.0 well, skip to the conclusion. *ZEN* here stands for *Zero Error Numerics*. Predictive Coding 4.0 requires quality control activities in all steps, but the efforts peak in this Step Seven. For more details than provided here on the *ZEN* approach to quality control in document review *see* ZeroErrorNumerics.com.



In Step Seven a random sample is taken to try to evaluate the recall range attained in the project. The method currently favored is described in detail in *Introducing "ei-Recall" – A New Gold Standard for Recall Calculations in Legal Search* – Part One, Part Two and Part Three. *Also see*: *In Legal Search Exact Recall Can Never Be Known*.



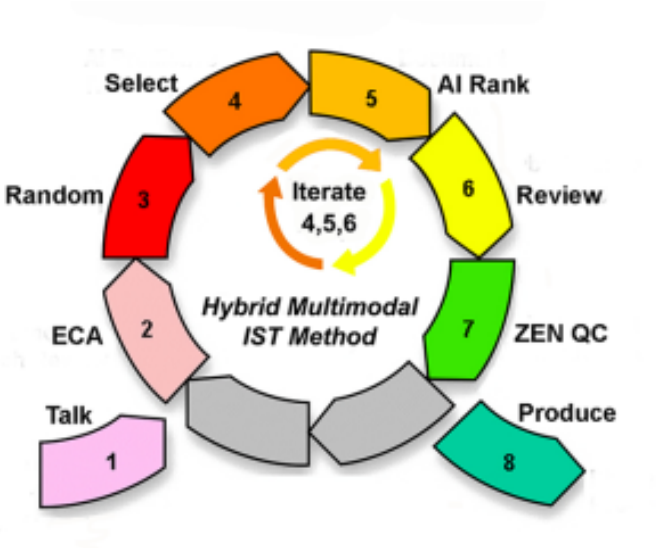
The *ei-Recall* test is based on a random sample of all documents to be excluded from the *Final Review* for possible production. Unlike the ill-fated *control set* of Predictive Coding 1.0 methodologies, the sample here is taken at the *end* of the project. At that time the final relevance conceptions have evolved to their final form and therefore much more accurate projections of recall can be made from the sample. The documents sampled can be based on documents excluded by category prediction (i.e. probable irrelevant) and/or by probable ranking of documents with proportionate cut-offs. The focus is on a search for any *false negatives* (i.e., relevant documents incorrectly predicted to be irrelevant) that are *Highly Relevant* or otherwise of significance.

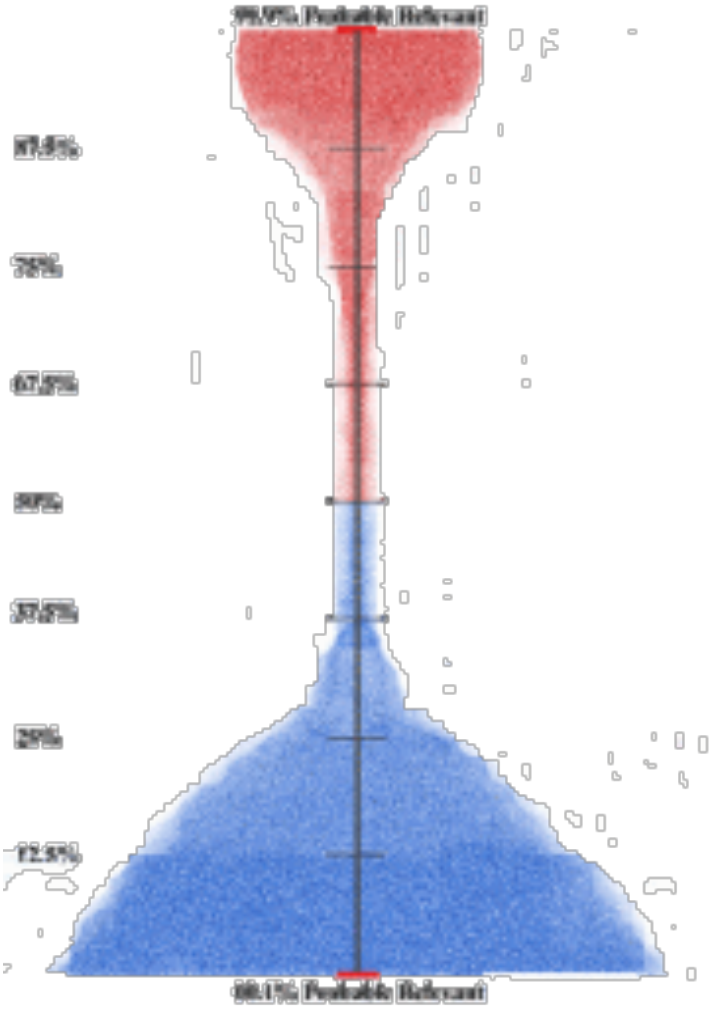


Total 100% recall of all relevant documents is said by the professors to be scientifically impossible (unless you produce all documents, 0% precision), a myth that the e-Discovery Team shattered in TREC 2015 and again in 2016 in our Total Recall Track experiments. Still, it is very rare, and only happens in relatively simple search and review projects, akin to a straightforward single plaintiff employment case with clear relevance. In any event, *total recall* of all relevant document is legally unnecessary. *Perfection - zero error - is a good goal, but never a legal requirement*. The legal requirement is reasonable, proportional efforts to find the ESI that is important to resolve the key disputed issues of fact in the case. The goal is to avoid all false negatives of *Highly Relevant* documents. If this error is encountered, one or more additional iterations of Steps 4, 5 and 6 are required.

In step seven you also test the decision made at the end of step six to stop the training. This decision is evaluated by the random sample, but determined by a complex variety of factors that can be case specific. Typically it is determined by when the software has attained a highly stratified distribution of documents. See License to Kull: Two-Filter Document Culling and *Visualizing Data in a Predictive Coding Project* – Part One, Part Two and Part Three, and Introducing a New Website, a New Legal Service, and a New Way of Life / Work; Plus a Postscript on Software Visualization.

When the stratification has stabilized you will see very few new documents found as predicted relevant that have not already been human reviewed and coded as relevant. You essentially run out of documents for step six *Review*. Put another way, your step six no longer uncovers new relevant documents. This exhaustion marker may, in many projects, mean that the rate of newly found documents has slowed, but not stopped entirely. I have written about this quite a bit, primarily in *Visualizing Data in a Predictive Coding Project* – Part One, Part Two and Part Three. The distribution ranking of documents in a mature project, one that has likely found all relevant documents of interest, will typically look something like the diagram below. We call this the upside down champagne glass with red relevant documents on top and irrelevant on the bottom.





Also see Postscript on Software Visualization where even more dramatic stratifications are encountered and shown.

Another key determinant of when to stop is the cost of further review. Is it worth it to continue on with more iterations of steps four, five and six? *See* *Predictive Coding and the Proportionality Doctrine: a Marriage Made in Big Data*, 26 Regent U. Law Review 1 (2013-2014) (note article was based on earlier version 2.0 of our methods where the training was not necessarily continuous). Another criteria in the stop decision is whether you have found the information needed. If so, what is the purpose of continuing a search? Again, the law never requires finding *all* relevant, only reasonable efforts to find the relevant documents needed to decide the important fact issues in the case. This last point is often overlooked by inexperienced lawyers.

Another important quality control technique, one used throughout a project, is the avoidance of all dual tasking, and learned, focused concentration, a flow-state, like an all-absorbing video game, movie, or a meditation.

Everybody needs to relax with a clear mind, and with focused attention, to attain their peak level of performance. That is the key to all quality control. How you get there is your business. Me, in addition to frequent breaks, I like headphones with music to help me there and help me to stay undistracted, focused. For more details on step seven see *ZeroErrorNumericcs.com*.

Step Eight: Phased Production

There has been no change in this step from *Version 3.0* to *Version 4.0*. If you already know 3.0 well, skip to the conclusion. This last step is where the relevant documents are reviewed again and actually produced. This step is also sometimes referred to as *Second Pass Review*. Technically, it has nothing to do with a predictive coding protocol, but for completeness sake, we needed to include it in the work flow. This final step may also include document redaction, document labeling, and a host of privilege review issues, including double-checking, triple checking of privilege protocols. These are tedious functions where contract lawyers can be a big help. The actual identification of privileged documents from the relevant should have been part of the prior seven steps.

Always think of production in e-discovery as phased production. Do not think of making one big document dump. That is old-school paper production style. Start with a small *test document production* after you have a few documents ready. That will get the bugs out of the system for both you, the producer, and also for the receiving party. Make sure it is in the format they need and they know how to open it. Little mistakes and re-dos in a small test production are easy and inexpensive to fix. Getting some documents to the requesting party also gives them something to look at right away. It can buy you time and patience for the remaining productions. It is not

uncommon for a large production to be done in five or more smaller stages. There is no limit so long as the time delay is not overly burdensome.

Multiple productions are normal and usually welcome by the receiving party. Just be sure to keep them informed of your progress and what remains to be done. Again, step one - Talk - is supposed to continue throughout a project. Furthermore, production of at least some documents can begin very early in the process. It does not have to wait until the last step. It can, for instance, begin while you are still in the iterated steps four, five and six. Just make sure you apply your quality controls and final second pass reviews to all documents produced. Very early productions during the intensive document training stages may help placate a still distrustful requesting party. It allows them to see for themselves that you are in fact using good relevant documents for training and they need not fear GIGO.

The format of the production should always be a non-issue. This is supposed to be discussed at the initial Rule 26(f) conference. Still, you might want to check again with the requesting party before you select the production format and metadata fields. More and more we see requesting parties that want a PDF format. That should not be a problem. Remember, cooperation should be your benchmark. Courtesy to opposing counsel on these small issues can go a long way. The existence of a clawback agreement and order, including a Rule 502(d) Order, and also a confidentiality agreement and order in some cases, should also be routinely verified *before* any production is made. This is critical and we cannot over-state its importance. You should never make a production without a Rule 502(d) Order in place, or at least requested from the court. Again, this should be a non-issue. The forms used should be worked out as part of the initial 26(f) meet and greet.

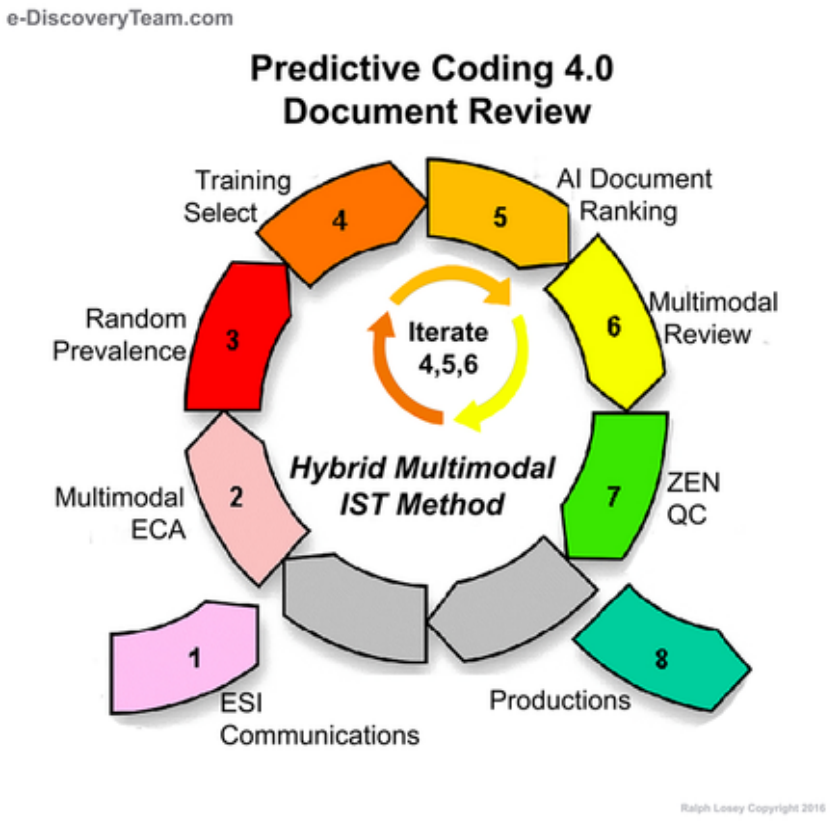
After the second pass review is completed there is still one more inspection, a short third pass. Before delivery of electronic documents we perform yet another quality control check. We inspect the media on which the production is made, typically CDs or DVDs, and do a third review of a few of the files themselves. This is an important quality control check, the last one, done just before the documents are delivered to the requesting party. You do not inspect every document, of course, but you do a very limited spot check based on judgmental sampling. You especially want to verify that critical privileged documents you previously identified as privileged have *in fact* been removed, and that redactions have been properly made. Trust but verify. Also check to verify the order of

production is what you expected. You also verify little things that you would do for any paper production, like verify that the document legends and Bates stamping are done the way you wanted. Even the best vendors sometimes make mistakes, and so too does your team.

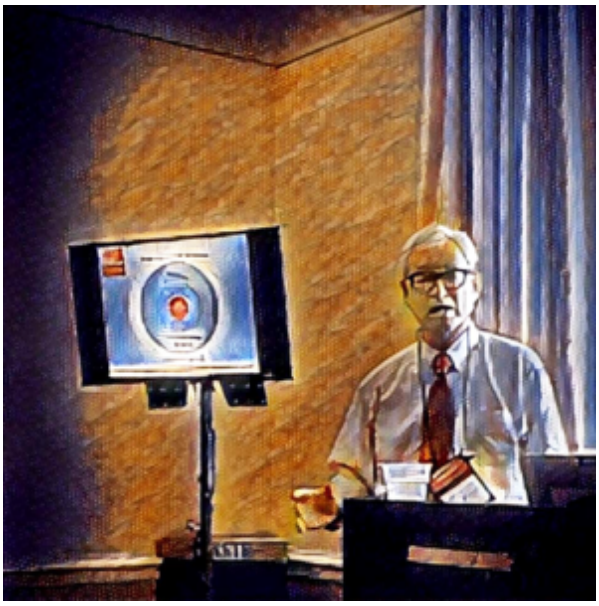
You need to be very diligent in protecting your client's confidential information. It is an ethical duty of all lawyers. It weighs heavily in what we consider a properly balanced, proportional approach. That is why you must take time to do the *Production* step correctly and should never let yourself be rushed.

The final work included here is to prepare a privilege log. All good vendor review software should make this into a semi-automated process, and thus slightly less tedious. The logging is typically delayed until after production. Check with local rules on this and talk to the requesting party to let them know it is coming.

One final comment on the e-Discovery Team's methods: we are very hyper about time management throughout a project, but especially in the last step. Never put yourself in a time bind. Be Proactive. Stay ahead of the curve. This is important for the entire project, but especially in the last step. Mistakes are made when you have to rush to meet tight production deadlines. You must avoid this. Ask for an extension and motion the court if you have to. Better that than make a serious error. Again, produce what you have ready and come back for the rest.



Conclusion



Every search expert I have ever talked to agrees that it is just good common sense to find relevant information by using every search method that you can. It makes no sense to limit yourself to any one search method. They agree that *multimodal* is the way to go, even if they do not use that language (after all, I did *make up* the term), and even if they do not publicly promote that protocol (they may be promoting software or a method that does not use all methods). All of the scientists I have spoken with about search also all agree that effective text retrieval should use some type of active machine learning (what we in the legal world calls *predictive coding*), and not just rely on the old search methods of keyword, similarity and concept type analytics. The combined multimodal use of the old and new methods is the way to go. This hybrid approach exemplifies man and machine working together in an active partnership, a union where the machine *augments* human search abilities, not replaces them.

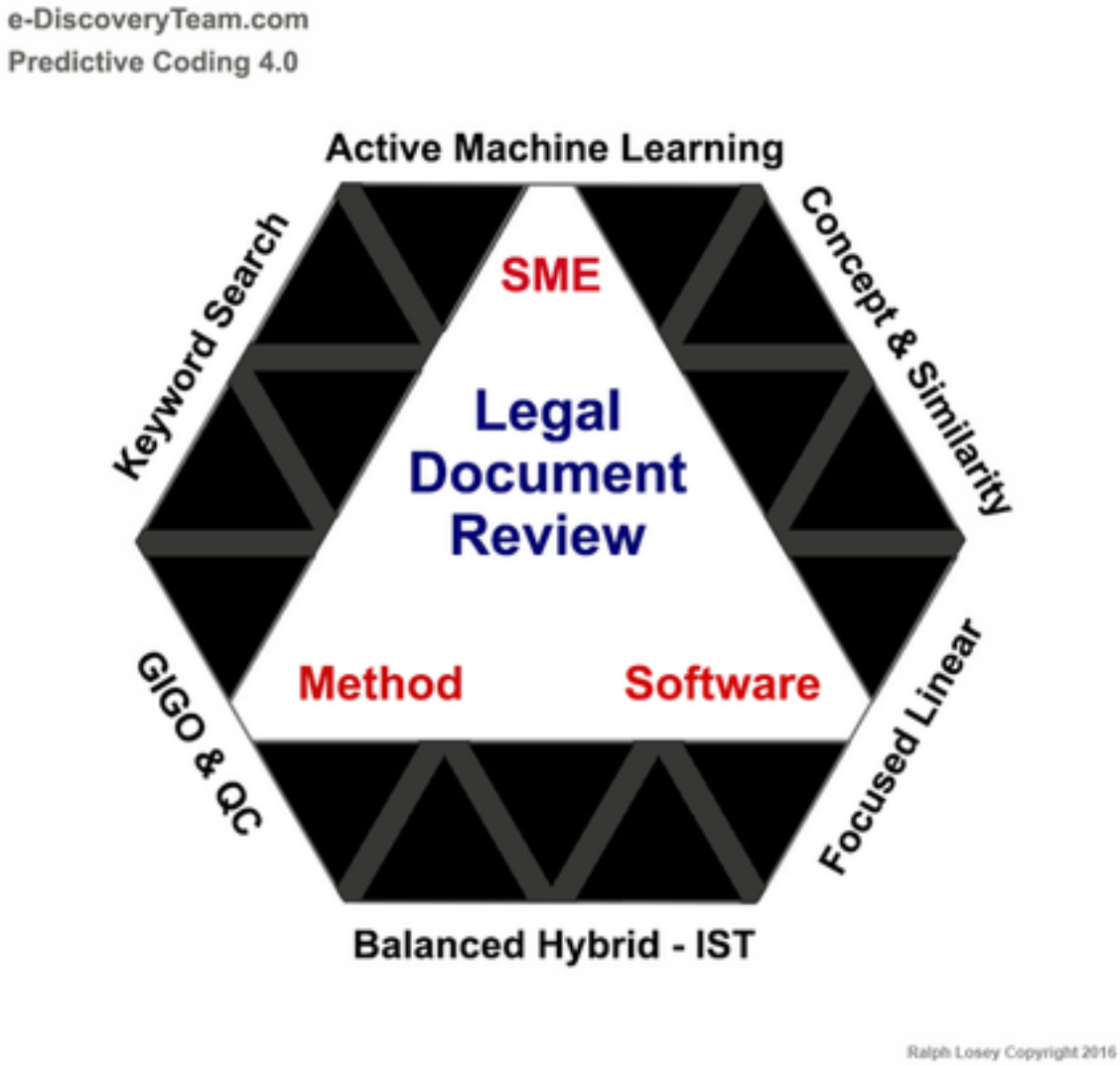
The *Hybrid IST Multimodal Predictive Coding 4.0* approach described here is still not followed by most e-discovery vendors, including several prominent software vendors. Instead, they rely on just one or two methods to the exclusion of the others. For instance, they may rely entirely on machine selected documents for training, or even worse, rely entirely on random selected documents. They do so to try to *keep it simple* they say. It may be simple, but the power and speed given up for that simplicity is not worth it. Others have all types of search, including concept search and related analytics, but they still do not have active machine learning. You probably know who they are by now. This problem will probably be solved soon, so I will not belabor the point.

The users of the old software and old-fashioned methods will never know the genuine thrill known by most search lawyers using AI enhanced methods like Predictive Coding 4.0. The *good times roll* when you see that the AI you have been training has absorbed your lessons. When you see the advanced intelligence that you helped create kick-in to complete the project for you. When you see your work finished in record time and with record results. It is sometimes amazing to see the AI find documents that you *know* you would never have found on your own. Predictive coding AI in superhero mode can be exciting to watch.

My entire e-Discovery Team had a great time watching Mr. EDR do his thing in the thirty Recall Track TREC Topics in 2015. We would sometimes be lost, and not even understand what the search was for anymore. But Mr. EDR knew, he saw the patterns hidden to us mere mortals. In those cases we would just sit back and let him do the driving, occasionally cheering him on. That is when my Team decided to give Mr. EDR a cape and superhero status. He never let us down. It is a great feeling to see your own intelligence augmented and save you like that. It was truly a hybrid human-machine partnership at its best. I hope you get the opportunity soon to see this in action for yourself.

Our experience in TREC 2016 was very different, but still made us glad to have Mr. EDR around. This time most of the search projects were simple enough to find the relevant documents without his predictive coding superpowers. As mentioned, we verified in test conditions that the skilled use of *Tested, Parametric Boolean Keyword Search* is very powerful. Keyword search, when done by experts using hands-on testing, and not simply blind *Go Fish* keyword guessing, is very effective. We proved that in the 2016 TREC search projects. As explained in Part Four of this article, the keyword appropriate projects are those where the data is simple, the target is clear and the SME is good. Still, even then, Mr. EDR was helpful as a quality control assistant. He verified that we had found all of the relevant documents.

Bottom line for the e-Discovery Team at this time is that the use of *all methods* is appropriate in all projects, even in simple searches where predictive coding is not needed to find all relevant documents. You can still use active machine learning in simple projects as a way to verify the effectiveness of your keyword and other searches. It may not be necessary in the simple cases, but it is still a good search to add to your tool chest. When the added expense is justified and proportional, the use of predictive coding can help assure you, and the other side, that a high quality effort has been made.



The multimodal approach is the most effective method of search. All search tools should be used, not only Balanced Hybrid - IST active machine learning searches, but also concept and similarity searches, keyword search and, in some instances, even focused linear review. By using some or all search methods, depending on the project and challenges presented, you can maximize recall (the truth, the whole truth) *and* precision (nothing but the truth). That is the goal of search: effective and efficient. Along the way we must exercise caution to avoid the errors of *Garbage in, Garbage Out*, that can be caused by poor SMEs. We must also guard against the errors and omissions, low recall and low precision, that can arise from substandard software and methods. In our view the software must be capable of all search methods, including active machine learning, and the methods used should too.