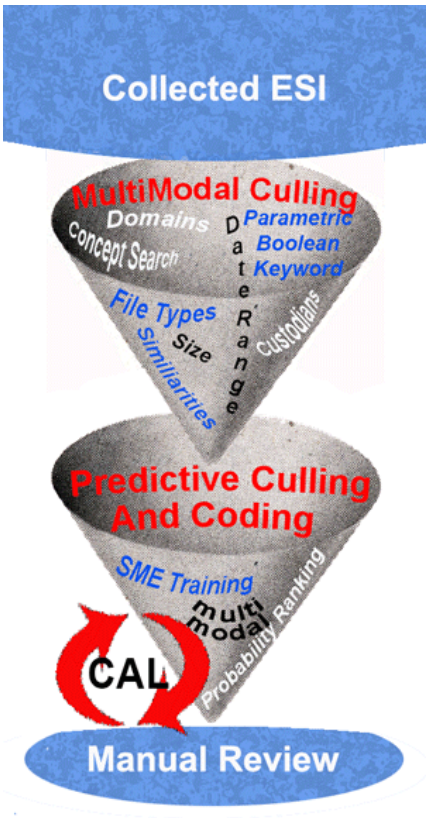
License to Cull

The Two-Filter Document Culling Method

Ralph C. Losey[·]
National e-Discovery Counsel
Jackson Lewis P.C.

Every attorney has a *license to cull* irrelevant data before beginning expensive linear review. It is part of their duty to protect their clients and country from waste and abuse. This article describes the two-filter culling method I've devised over the years to identify and bulk-code irrelevant documents. The method is designed for use *before* commencing a detailed attorney review. The efficacy of any large-scale document review project can be enhanced by this double-cull method. In my experience, it not only helps to reduce costs, it also maximizes recall, allowing an attorney to find all of the documents needed for a case quickly and efficiently.

I briefly introduced this method, and the diagram shown right illustrating it, at the conclusion of a lengthy article on document review quality control: *Introducing "ei-Recall" – A New Gold Standard for Recall Calculations in Legal Search – Part Three* (e-DiscoveryTeam, 2015). I use the two-filter method in most large projects as part of my *multimodal, bottom line driven, AI-Enhanced* (i.w. – predictive coding) method of review. I have described segments of this method, including especially predictive coding, in prior articles on document review. They are listed at the bottom of the Legal Search Science website. I also described this process as part of the *Electronic Discovery Best Practices* website, found at EDBP.com, which outlines my views on the best practices for lawyers doing e-discovery. (Please note that all views expressed here, and my other writings, are my own personal opinions, and not necessarily those of my law firm or clients.)



[·] Publisher, principal author of e-DiscoveryTeam.com, *also see* RalphLosey.com.

The two-filter culling method includes the well-known technology processes of *deduplication* and *deNisting* in the first filter. (Note: I always do full *horizontal* deduplication across all custodians.) Deduplication and deNisting are, however, just technical engineering filters, not based on legal analysis or judgment. They are well-established industry standards and so I will not discuss them further in this article.

Many e-discovery beginners think that deNisting and deduplication are the *end-all* of ESI culling, but that is far from true. They are just the beginning. The other methods described here all require legal judgment, and so you cannot just hire an e-discovery vendor to do it, as you can with deduplication and deNisting. Legal judgment is critical to all effective document review, including culling of irrelevant documents before lawyers spend their valuable time in linear review. In my opinion, all legal review teams should employ some type of two-filter culling component.

My thirty-five plus years of experience as a practicing lawyer have shown me that the most reliable way for the *magic of justice* to happen is by finding the key documents. You find *the truth, the whole truth, and nothing but the truth,* when you find the *key* documents needed to complete the picture of what happened and keep witnesses honest. In today's information flooded world, that can only happen if you use technology in a strategic manner to find relevant evidence quickly and inexpensively. The two-filter method makes it easier to do that. This almost 10,000 word article provides an explanation of how to do it that is accessible to beginners and *eLeet* alike.

I have been working to refine this irrelevant culling method since 2006. At that time I limited my practice to e-discovery and put aside my commercial and employment litigation practice. For more background on my personal views and opinions on e-discovery, and for a description of many other document review methods that I have developed, not just two-filer culling, see my independent blog, e-DiscoveryTeam.com, especially the About Page. I have written over a million words on e-Discovery, including five books, but this article is the first full description of document culling.

This article contains a lengthy description of document culling, but still is not complete. My methods vary to adapt to the data and changing technologies. I share these methods to try to help all attorneys control the costs of document review and find the information needed to do justice. All too often these costs spiral out of control, or the review is done so poorly that key documents are not found. Both scenarios are obviously bad for our system of justice. We need cases to be decided on the merits, on the facts.

Hopefully my writings can help make that happen in some small way. Hopefully a more tech-savvy Bar can stem the tide of over-settlement that we have seen in the profession since the explosion of data began in the nineties. All too often cases are now decided on the basis of settlement value, not merits. As it now stands, way too many frivolous cases are filed hoping there will be some kind of payout. These cases

tend to drown out the few with merit. Judges are overwhelmed and often do not have the time needed to get down to the nitty-gritty details of the truth.

Most of the time judges and juries are never given the chance to do their job. The cases all settle out instead. As a result **only one percent** of federal civil cases actually go to trial. This is a big loss for society, and for the "trial lawyers" in our profession, a group I once prided myself to be a part. Now I just focus on getting the facts from big data, to help keep the witnesses honest, and cases decided on the true facts, the evidence. Then I turn it over to the trial lawyers in my firm. They are then armed with the truth, the key documents, good or bad. The trial lawyers then put the best face possible on these facts, which hopefully is handsome to begin with. They argue how the law applies to these facts to seek a fair and just result for our clients. The disputed issues of fact are also argued, but based on the evaluation of the meaning of the key documents and the witness testimony.

That is, in my opinion, how our system of justice is supposed to operate. It is certainly the way our legal system functioned when I learned to practice law and had my first trials back in 1980. Back then we only had a few thousand files to cull through to find the key documents, perhaps tens of thousands in a big case. Now we have hundreds of thousands of documents to cull through, millions in a big case. Still, even though the data volumes are far greater today, with the two-filter method described here, the few key documents needed to decide a case can be found.

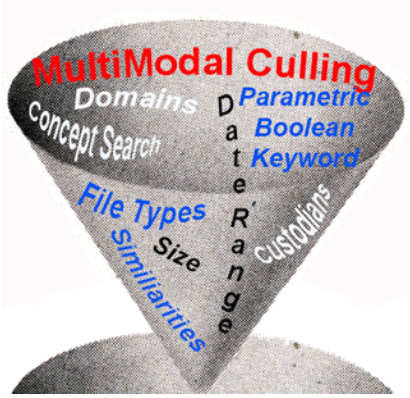
Big Data today presents an opportunity for lawyers. There are electronic writings everywhere and can be hard to destroy. The large amount of ESI floating in cyberspace means that the truth is almost always out there. You just have to find it.

There is so much data that it is much more likely for key documents to exist than ever before. The *digital trails* that people leave today are much bigger than the *paper trails* of old.

The fact that more truth is out there than ever before gives tech-savvy lawyers a great advantage. They have a much better chance than lawyers in the past ever did to find the documents needed to keep witnesses honest, or put more politely, to help refresh their memory. The flood of information can in this way improve the quality of justice. It all depends on our ability to find the truth from the massive quantities of irrelevant information available.

The more advanced culling methods described here, primarily the ones in the second filter that use predictive coding - AI-enhanced document ranking methods - are especially effective in culling the chaff from the wheat in big data cases. I expect this kind of predictive analytics software to keep on improving. For that reason I am confident that we will continue to be able to find the core kernels of truth needed to do justice, no matter how much data we generate and save.

Some Software is Far Better than Others



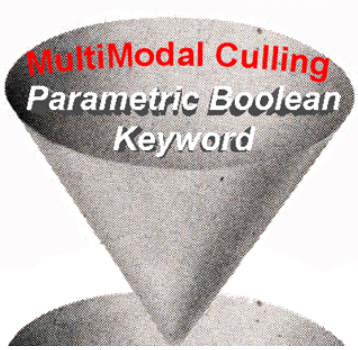
One word of warning, although this method is software agnostic, in order to emulate the two-filter method, your document review software must have certain basic capabilities. That includes effective, and easy, bulk coding features for the first filter. This is the multimodal broad-based culling. Some of the multiple methods do not require software features, just attorney judgment, such as excluding custodians, but other do require software features, like domain searches or similarity searches. If your software does not have the features that will be discussed here for the first filter, then you probably should switch right away, but, for most, that will not be a problem. The multimodal culling methods used in the first filter are, for the most part, pretty basic.

Some of the software features needed to implement the second filter, are, however, more advanced. The second filter works best when using predictive coding and probability ranking. You review the various strata of the ranked documents. The second filter can still be used with other, less advanced multimodal methods, i.e. keywords. Moreover, even when you use bona fide active machine learning software features, you continue to use a smattering of other multimodal search methods in the second filter. But now you do so not to cull, but to help find relevant and highly relevant documents to improve training. I do not rely on probability searches alone, although sometimes in the second filter I rely almost entirely on predictive coding based searches to continue the training.

If you are using software without AI-enhanced active learning features, then you are forced to only use other multimodal methods in the second filter, such as keywords. Warning, true active learning features are *not* present in most review software, or are very weak. That is true even with software that claims to have predictive coding features, but really just has dressed-up *passive learning*, i.e. concept searches with latent semantic indexing. You handicap yourself, and your client, by continuing to use such less expensive programs. Good software, like everything else, does not come cheap, but should pay for itself many times over if used correctly. The same comment goes for lawyers too.

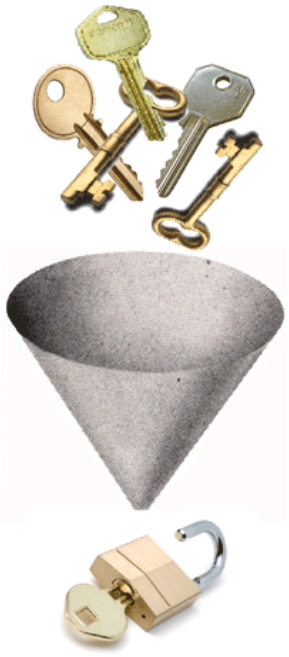
First Filter – Keyword Collection Culling

Some first stage filtering takes place as part of the ESI collection process. The documents are preserved, but not collected nor ingested into the review database. The most popular collection filter as of 2015 is still keyword, even though this is very risky in some cases and inappropriate in many. Typically such keyword filtering is driven by vendor costs to avoid processing and hosting charges.



Some types of collection filtering are appropriate and necessary, for instance, in the case of custodian filters, where you broadly preserve the ESI of many custodians, just in case, but only collect and review a few of them. It is, however, often *inappropriate* to use keywords to filter out the collection of ESI from admittedly key custodians. This is a situation where an attorney determines that a custodian's data needs to be reviewed for relevant evidence, but does not want to incur the expense to have all of their ESI ingested into the review database. For that reason they decide to only review data that contains certain keywords.

I am not a fan of keyword filtered collections. The obvious danger of keyword filtering is that important documents may not have the keywords. Since they will not even be placed in the review platform, you will never know that the relevant ESI was missed. You have no chance of finding them.



See eg, William Webber's analysis of the *Biomet* case where this kind of keyword filtering was use before predictive coding began. *What is the maximum recall in re Biomet?*, Evaluating e-Discovery (4/24/13). Webber shows that in *Biomet* this method first-filtered out over 40% of the relevant documents. This doomed the second filter predictive coding review to a maximum possible recall of 60%, even if was perfect, meaning it would otherwise have attained 100% recall, which never happens. The *Biomet* case very clearly shows the dangers of over-reliance on keyword filtering.

Nevertheless, *sometimes* keyword collection may work, and may be appropriate. In some simple disputes, and with some data collections, obvious keywords may work just fine to unlock the truth. For instance, sometimes the use of names is an effective method to identify all, or almost all, documents that may be relevant. This is especially true in smaller and simpler cases. This method can, for instance, often work in employment cases, especially where unusual names are involved. It becomes an even more effective method when the

keywords have been tested. I just love it, for instance, when the plaintiff's name is something like the famous Mister Mxyzptlk.

In some cases keyword collections may be as risky as in the complex *Biomet* case, but may still be necessary because of the proportionality constraints of the case. The law does not require unreasonably excessive search and review, and what is reasonable in a particular case depends on the facts of the case, including its value. See my many writings on proportionality, including my law review article *Predictive Coding and Proportionality: A Marriage Made In Heaven*, 26 Regent U. Law Review 1 (2013-2014). Sometimes you have to try for rough justice with the facts that you can afford to find given the budgetary constraints of the case.

The danger of missing evidence is magnified when the keywords are selected on the basis of educated guesses or just limited research. This *technique*, if you can call it that, is, sadly, still the dominant method used by lawyers today to come up with keywords. I have long thought it is equivalent to a child's game of *Go Fish*. If keywords are dreamed up like that, as mere educated guesses, then keyword filtering is a high-risk method of culling out irrelevant data. There is a significant danger that it will exclude many important documents that do not happen to contain the selected keywords. No matter how good your predictive coding may be after that, you will never find these key documents.

If the keywords are not based on a mere guessing, but are instead *tested,* then it becomes a real technique that is less risky for culling. But how do you test possible keywords without first collecting and ingesting all of the documents to determine which are effective? It is the old *cart before the horse* problem.

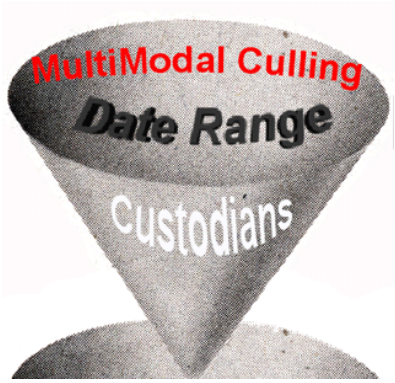
One partial answer is that you could ask the witnesses, and do some *partial reviews* before collection. Testing and witness interviews is required by Judge Andrew Peck's famous *wake up call* case. *William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co.*, 256 F.R.D. 134, 134, 136 (S.D.N.Y. 2009). I recommend that opinion often, as many attorneys still need to *wake up* about how to do e-discovery. They need to add ESI use, storage, and keyword questions to their usual new case witness interviews.

Interviews do help, but there is nothing better than actual hands on reading and testing of the documents. This is what I like to call *getting your hands dirty in the digital mud* of the actual ESI collected. Only then will you know for sure the best way to mass-filter out documents. For that reason my strong preference in all significant size cases is to collect in bulk, and *not* filter out by keywords. Once

you have documents in the database, *then* you can then effectively screen them *out* by using *parametric Boolean keyword* techniques. See your particular vendor for various ways on how to do that.

By the way, *parametric* is just a reference to the various parameters of a computer file that all good software allows you to search. You could search the text and all metadata fields, the entire document. Or you could limit your search to various metadata fields, such as date, prepared by, or the *to and from* in an email. Everyone knows what *Boolean* means, but you may not know all of the many variations that your particular software offers to create highly customized searches. While predictive coding is beyond the grasp of most vendors and case managers, the intricacies of keyword search are not. They can be a good source of information on keyword methods.

First Filter – Date Range and Custodian Culling

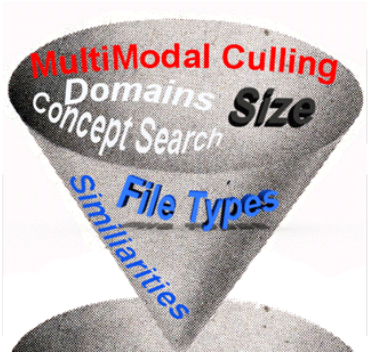


Even when you collect in bulk, and do not *keyword filter* before you put custodian ESI in the review database, in most cases you should filter for date range and custodian. It is often possible for an attorney to know, for instance, that no emails before or after a certain date could possibly be relevant. That is often not a highly speculative guessing game. It is reasonable to filter on this time-line basis before the ESI goes in the database. Whenever possible, try to get agreement on date range screening from the requesting party. You may have to widen it a little, but it is worth the effort to establish a line of communication and begin a cooperative dialogue.

The second thing to talk about is *which custodians* you are going to include in the database. You may put 50 custodians on hold, and actually collect the ESI of 25, but that does not mean you have to load all 25 into the database for review. Here your interviews and knowledge of the case should allow you to know who the key, key custodians are. You rank them by your evaluation of the likely importance of the data they hold to the facts disputed in the case. Maybe, for instance, in your evaluation you only need to review the mailboxes of 10 of the 25 collected.

Again, disclose and try to work that out. The requesting party can reserve rights to ask for more, that is fine. They rarely do after production has been made, especially if you were careful and picked the right 10 to start with, and if you were careful during review to drop and add custodians based on what you see. If you are using predictive coding in the second filter stage, the addition or deletion of data mid-course is still possible with most software. It should be robust enough to handle such mid-course corrections. It may just slow down the ranking for a few iterations, that's all.

First Filter – Other MultiModal Culling



There are many other bulk coding techniques that can be used in the first filter stage. This is not intended to be an exhaustive search. Like all complex tasks in the law, simple *black letter* rules are for amateurs. The law, which mirrors the real world, does not work like that. The same holds true for legal search. There may be many *Gilbert's* for search books and articles, but they are just 1L types guides. For true legal search professionals they are mere starting points. Use my culling advice here in the same manner. Use your own judgment to mix and match the right kind of culling tools for the particular case and data encountered. Every project is slightly different, even in the world of repeat litigation, like employment law disputes where I currently spend much of my time.

Legal search is at core a heuristic activity, but one that should be informed by science and technology. The knowledge triangle is a key concept for today's effective e-Discovery Team. Although e-Discovery Teams should be led by attorneys skilled in evidence discovery, they should include scientists and engineers in some way. Effective team leaders should be able to understand and communicate with technology experts and information scientists. That does not mean all e-discovery lawyers need to become engineers and scientists too. That effort would likely diminish your legal skills based on the time demands involved. It just means you should know enough to work with these experts. That includes the ability to see through the vendor sales propaganda, and to incorporate the knowledge of the *bona fide* experts into your legal work.

One culling method that many overlook is *file size*. Some collections have thousands of very small files, just a few bits, that are nothing but backgrounds, tiny images, or just plain empty space. They are too small to have any relevant information. Still, you need to be cautious and look out for very small emails, for instance, ones that just says "yes." Depending on context it could be relevant and important. But for most other types of very small files, there is little risk. You can go ahead a bulk code them irrelevant and filter them out.

Even more subtle is filtering out files based on their being very large. Sort your files by size, and then look at both ends, small and big. They may reveal certain files and file types that could not possibly be relevant. There is one more characteristic of big files that you should consider. Many of them have millions of lines of text. Big files

are confusing to machine learning when, as typical, only a few lines of the text are relevant, and the rest are just noise. That is another reason to filter them out, perhaps not entirely, but for special treatment and review outside of predictive coding. In other projects where you have many large files like that, and you need the help of AI ranking, you may want to hold them in reserve. You may only want to throw them into the ranking mix after your AI algorithms have acquired a pretty good idea of what you are looking for. A maturely trained system is better able to handle big noisy files.

File type is a well-known and often highly effective method to exclude large numbers of files of a same type after only looking at a few of them. For instance, there may be database files automatically generated, all of the same type. You look at a few to verify these databases could not possibly be relevant to your case, and then you bulk code them all irrelevant. There are many types of files like that in some data sets. The first filter is all about being a smart gatekeeper.

File type is also used to eliminate, or at least divert, non-text files, such as audio files or most graphics. Since most second filter culling is going to be based on text analytics of some kind, there is no point for anything other than files with text to go into that filter. In some cases, and some datasets, this may mean bulk coding them all irrelevant. This might happen, for instance, where you know that no music or other audio files, including voice messages, could possibly be relevant. We also see this commonly where we know that photographs and other images could not possibly be relevant. Exclude them from the review database.

You must, however, be careful with all such gatekeeper activities, and never do bulk coding without some judgmental sampling first. Large unknown data collections can always contain a few unexpected surprises, no matter how many document reviews you have done before. Be cautious. Look before you leap. Skim a few of the ESI file types you are about to bulk code as irrelevant.

This directive applies to all *first filter* activities. Never do it blind on just logic or principle alone. Get you hands in the digital mud. Do not over-delegate all of the *dirty work* to others. Do not rely too much on your contract review lawyers and vendors, especially when it comes to search. Look at the documents yourself and do not just rely on high-level summaries. Every real trial lawyer knows the importance of that. *The devil is always in the details.* This is especially true when you are doing judgmental search. The client wants your judgment, not that of a less qualified associate, paralegal, or minimum wage contract review lawyer. Good lawyers remain hands-on, to some extent. They know the details, but are also comfortable with appropriate delegation to trained team members.

There is a constant danger of too much delegation in big data review. The lawyer signing the Rule 26(g) statement has a legal and ethical duty to closely supervise document review done in response to a request for production. That means you cannot just hire a vendor to do that, although you can hire outside counsel with special expertise in the field.

Some non-text file types will need to be diverted for different treatment than the rest of your text-based dataset. For instance, some of the best review software allows you to keyword search audio files. It is based on phonetics and wave forms. At least one company I know has had that feature since 2007. In some cases you will have to carefully review the image files, or at least certain kinds of them. Sorting based on file size and custodian can often speed up that exercise.

Remember the goal is always efficiency, and caution, but not over cautious. The more experienced you get the better you become at evaluating risks and knowing where you can safely take chances to bulk code, and where you cannot. Another thing to remember is that many image files have text in them too, such as in the metadata, or in ASCII transmissions. They are usually not important and do not provide good training for second stage predictive coding.

Text can also be hidden in dead Tiff files, if they have not been OCR'ed. Scanned documents Tiffs, for instance, may very well be relevant and deserve special treatment, including full manual review, but they may not show in your review tool as text, because they have never been OCR text recognized.

Concept searches have only rarely been of great value to me, but should still be tried out. Some software has better capacities with concepts and latent semantic indexing than others. You may find it to be a helpful way to find groupings of obviously irrelevant, or relevant documents. If nothing else, you can always learn something about your dataset from these kind of searches.

Similarity searches of all kinds are among my favorite. If you find some files groups that cannot be relevant, find more like that. They are probably bulk irrelevant (or relevant) too. A similarity search, such as find every document that is 80% or more the same as this one, is often a good way to enlarge your carve outs and thus safely improve your efficiency.

Another favorite of mine is *domain culling* of email. It is kind of like a spam filter. That is a great way to catch the junk mail, newsletters, and other purveyors of general mail that cannot possibly be relevant to your case. I have never seen a mail collection that did not have dozens of domains that could be eliminated. You can sometimes cull-out as much as 10% of your collection that way,

sometimes more when you start diving down into senders with otherwise safe domains. A good example of this is the IT department with their constant mass mailings, reminders and warnings. Many departments are guilty of this, and after examining a few, it is usually safe to bulk code them all irrelevant.

Second Filter – Predictive Culling and Coding

The second filter begins where the first leaves off. The ESI has already been purged of unwanted custodians, date ranges, spam, and other obvious irrelevant files and file types. Think of the first filter as a rough, coarse filter, and the second filter as fine-grained. The second filter requires a much deeper dive into file contents to cull out irrelevance. The most effective way to do that is to use predictive coding, by which I mean active machine learning, supplemented somewhat by using a variety of methods to find good training documents. That is what I call a multimodal approach that places primary reliance on the Artificial Intelligence at the top of the search pyramid. If you do not have active machine learning type of predictive coding with ranking abilities, you can still do fine grained Second Level filtering, but it will be harder, and probably less effective and more expensive.

All kinds of second filter search methods should be used to find highly relevant and relevant documents for AI training. Stay away from any process that uses just one search method, even if the one method is predictive ranking. Stay *far away* if the *one*

method is rolling dice. Reliance on random chance alone has been proven to be an inefficient and ineffective way to select training documents. *Latest Grossman and Cormack Study Proves Folly of Using Random Search For Machine Training* – Part One, Two, Three and Four. No one should be surprised by that.

The first round of training begins with the documents reviewed and coded relevant incidental to the first filter coding. You may also want to defer the first round until you have done more active searches for relevant and highly relevant from the pool remaining *after* first filter culling. In that case you also include *irrelevant* in the first training round, which is also important. Note that even though the first round of training is the only round of training that has a special name – *seed set* – there is nothing all that important or special about it. All rounds of training are important.

There is so much misunderstanding about that, and *seed sets*, that I no longer like to even use the term. The only thing special in my mind about the first round of training is that it is often a very *large* training set. That happens when the first filter turns up a large amount of relevant files, or they are otherwise known and coded before the second filter training begins. The sheer volume of training documents in many first rounds thus makes it special, not the fact that it came first.

No good predictive coding software is going to give special significance to a training document just because it came first in time. (It might if it uses a *control set*, but that is a different story.) The software I use has no trouble at all disregarding any early training if it later finds that it is inconsistent with the total training input. It is, admittedly, somewhat aggravating to have a machine tell you that your earlier coding was wrong. But I would rather have an emotionless machine tell me that, than another gloating attorney (or judge), especially when the computer is correct, which is often (not always) the case.

That is, after all, the whole point of using good software with artificial intelligence. You do that to enhance your own abilities. There is no way I could attain the level of recall I have been able to manage lately in large document review projects by reliance on my own, limited intelligence alone. That is another one of my search and review secrets. Get help from a higher intelligence, even if you have to create it yourself by following proper training protocols.

Privacy Issues

Maybe someday the AI will come prepackaged, and not require training, or at least very little training. I know it can be done, especially if other data analytics techniques are used. I am working on this project now. *See* PreSuit.com. In addition to technical issues, there are serious ethical concerns as well, including especially employee privacy concerns. *Should Lawyers Be Big Data Cops?* The implications on the law of *predictive misconduct* are tremendous. I am now focusing my time and resources accordingly.

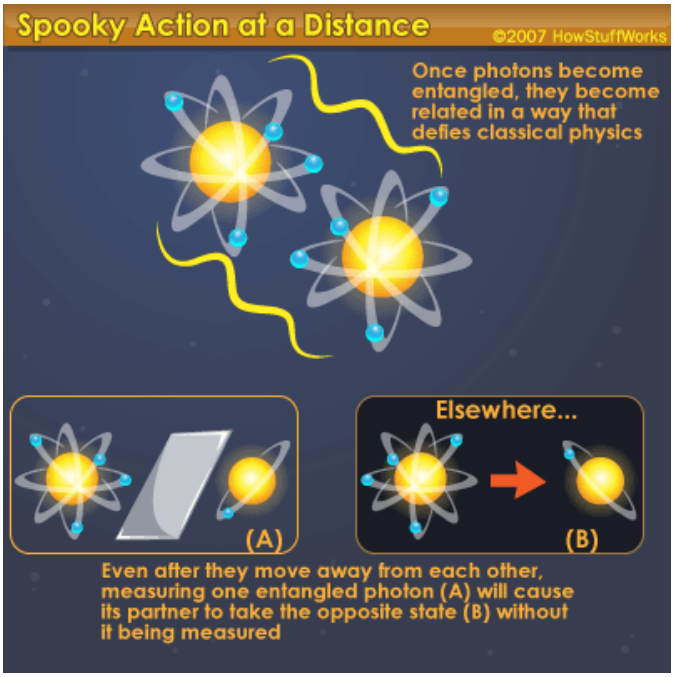
Information governance in general is something that concerns me, and is another reason I hold back on Presuit. *Hadoop, Data Lakes, Predictive Analytics and the Ultimate Demise of Information Governance* – Part One and Part Two. *Also see: e-Discovery Industry Reaction to Microsoft's Offer to Purchase Equivio for $200 Million* – Part Two. I do not want my information *governed*, even assuming that's possible. I want it secured, protected, and findable, but only by me, unless I give my express written assent (no contracts of adhesion permitted). By the way, even though I am cautious, I see no problem in requiring that consent as a condition of employment, so long as it is reasonable in scope and limited to only business communications.

I am wary of Big Brother emerging from Big Data. You should be too. I want AIs under our own individual control where they each have a real big off switch. That is the way it is now with legal search and I want it to stay that way. I want the AIs to remain under my control, not visa versa. Not only that, like all Europeans, I want a *right to be forgotten* by AIs and humans alike.

But wait, there's still more to my vision of a free future, one where the *ideals* of freedom and liberty triumph. I want AIs smart enough to *protect individuals* from governments, all governments, including the Obama administration. His DOJ has continued the disgraceful acts of the Bush Administration to ignore the Constitutional prohibition against *General Warrants*. *See*: Fourth Amendment to the U.S. Constitution. Now that Judge Facciola has retired, who on the federal D.C. bench is brave enough to protect us? *See*: Judge John Facciola Exposes Justice Department's Unconstitutional Search and Seizure of Personal Email.

Perhaps quantum entanglement encryption is the ultimate solution? *See eg.* Entangled Photons on Silicon Chip: Secure Communications & Ultrafast Computers, *The Hacker News*, 1/27/15. Truth is far stranger than fiction. Quantum Physics may seem irrational, but it has repeatedly been proven true. The

fact that it may seem irrational for two electrons to interact instantly over any distance just means that our sense of reason is not keeping up. There may soon be spooky ways for private communications to be forever private.



At the same time that I want unentangled freedom and privacy, I want a government that can protect us from crooks, crazies, foreign governments, and black hats. I just do not want to give up my Constitutional rights to receive that protection. We should not have to trade privacy for security. That is a false choice. Once we lay down our Constitutional rights in the name of security, the terrorists have already won.

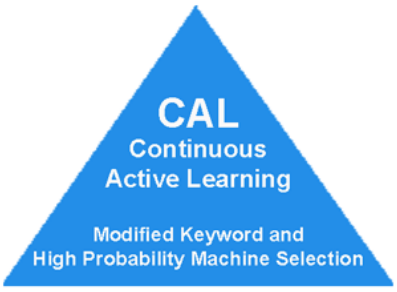
Getting back to legal search, and how to find out what you need to know inside the law by using the latest AI-enhanced search methods, there are three kinds of probability ranked search engines now in use for predictive coding.

Three Kinds of Second Filter Probability Based Search Engines

After the first round of training, you can begin to harness the AI features in your software. You can begin to use its probability ranking to find relevant documents. There are currently three kinds of ranking search and review strategies in use: uncertainty, high probability, and random. The uncertainty search, sometimes called SAL for *Simple Active Learning*, looks at middle ranking documents where the code is unsure of relevance, typically the 40%-60% range. The high probability search looks

at documents where the AI thinks it knows about whether documents are relevant or irrelevant. You can also use some random searches, if you want, both simple and judgmental, just be careful not to rely too much on chance.

The 2014 Cormack Grossman comparative study of various methods has shown that the high probability search, which they called CAL, for *Continuous Active Learning* using high ranking documents, is very effective. Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery, SIGIR'14, July 6–11, 2014. *Also see:* Latest Grossman and Cormack Study Proves Folly of Using Random Search For Machine Training – Part Two.



My own experience also confirms their experiments. High probability searches usually involve SME training and review of the upper strata, the documents with a 90% or higher probability of relevance. The exact percentage depends on the number of documents involved. I may also check out the low strata, but will not spend very much time on that end. I like to use both uncertainty and high probability searches, but typically with a strong emphasis on the high probability searches. And again, I supplement these ranking searches with other multimodal methods, especially when I encounter strong, new, or highly relevant type documents.

Sometimes I will even use a little random sampling, but the mentioned Cormack Grossman study shows that it is not effective, especially on its own. They call such chance-based search *Simple Passive Learning,* or *SPL.* Ever since reading the Cormack Grossman study I have cut back on my reliance on any random searches. You should too. It was small before, it is even smaller now. This does not mean sampling does not still have a place in documents review. It does, but in quality control, not in selection of training documents. *See eg*. ZeroErrorNumerics.com and *Introducing "ei-Recall" – A New Gold Standard for Recall Calculations in Legal Search*.

Irrelevant Training Documents Are Important Too

In the second filer you are on a search for the gold, the highly relevant, and, to a lesser extent, the strong and merely relevant. As part of this second filter search you will naturally come upon many irrelevant documents too. Some of these documents should also be added to the training. In fact, is not uncommon to have more irrelevant documents in training than relevant, especially with low prevalence collections. If you judge a document, then go ahead and code it and let the computer know your judgment. That is how it learns. There are some documents that you

judge that you may not want to train on – such as the very large, or very odd – but they are few and far between.

Of course, if you have culled out a document altogether in the first filter, you do not need to code it, because these documents will not be part of the documents included in the second filter. In other words, they will not be among the documents ranked in predictive coding. The will either be excluded from possible production altogether as irrelevant, or will be diverted to a non-predictive coding track for final determinations. The later is the case for non-text file types like graphics and audio in cases where they might have relevant information.

How To Do Second Filter Culling Without Predictive Ranking

When you have software with active machine learning features that allow you to do predictive ranking, then you find documents for training, and from that point forward you incorporate ranking searches into your review. If you do not have such features, you still sort out documents in the second filter for manual review, you just do not use ranking with SAL and CAL to do so. Instead, you rely on keyword selections, enhanced with concept searches and similarity searches.

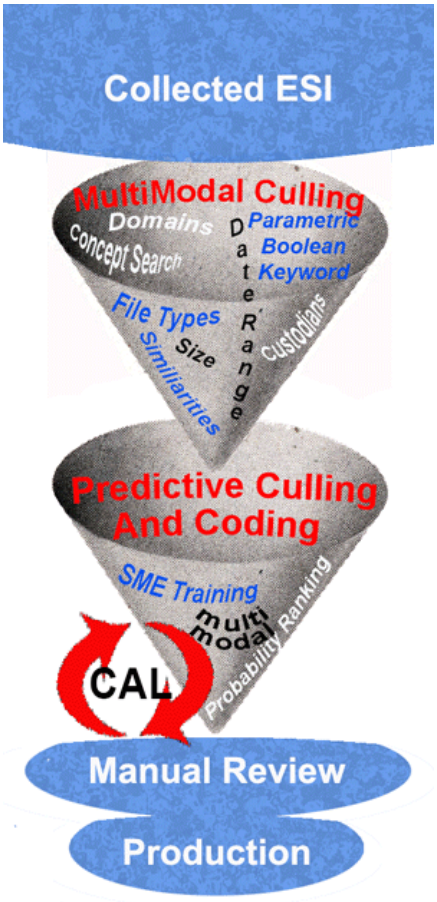
When you find an effective parametric Boolean keyword combination, which is done by a process of party negotiation, then testing, educated guessing, trial and error, and judgmental sampling, then you submit the documents containing proven hits to full manual review. Ranking by keywords can also be tried for document batching, but be careful of large files having many keyword hits just on the basis of file size, not relevance. Some software compensates for that, but most do not. So ranking by keywords can be a risky process.

I am not going to go into detail on the old fashioned ways of batching out documents for manual review. Most e-discovery lawyers already have a good idea of how to do that. So too do most vendors. Just one word of advice, when you start the manual review based on keyword or other non-predictive coding processes, check in daily with the contract reviewer work and calculate what kind of precision the various keyword and other assignment folders are creating. If it is terrible, which I would say is less than 50% precision, then I suggest you try to improve the selection matrix. Change the Boolean, or key words, or something. Do not just keep plodding ahead and wasting client money.

I once took over a review project that was using negotiated, then tested and modified keywords. After two days of manual review we realized that only 2% of the documents selected for review by this method were relevant. After I came in and spent three days with training to add predictive ranking we were able to increase

that to 80% precision. If you use these multimodal methods, you can expect similar results.

Review of Basic Idea of Two-Filter Search and Review

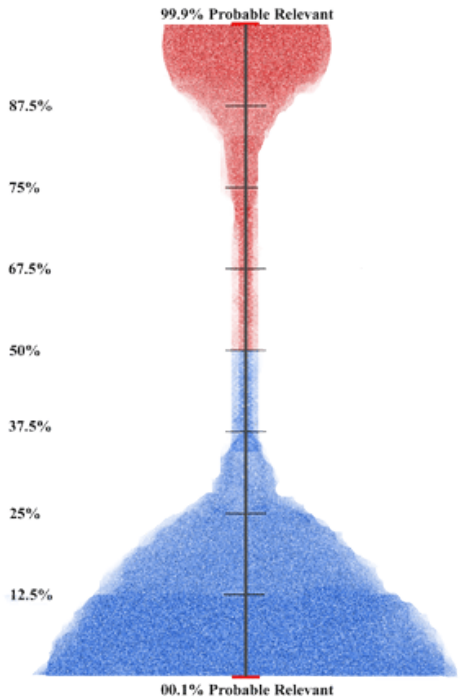


Whether you use predictive ranking or not, the basic idea behind the two-filter method is to start with a very large pool of documents, reduce the size by a coarse first filter, then reduce it again by a much finer second gilter. The result should be a much, much small pool that is human reviewed, and an even smaller pool that is actually produced or logged. Of course, some of the documents subject to the final human review may be overturned, that is, found to be irrelevant, False Positives. That means they will not make it to the very bottom production pool after manual review in the diagram right.

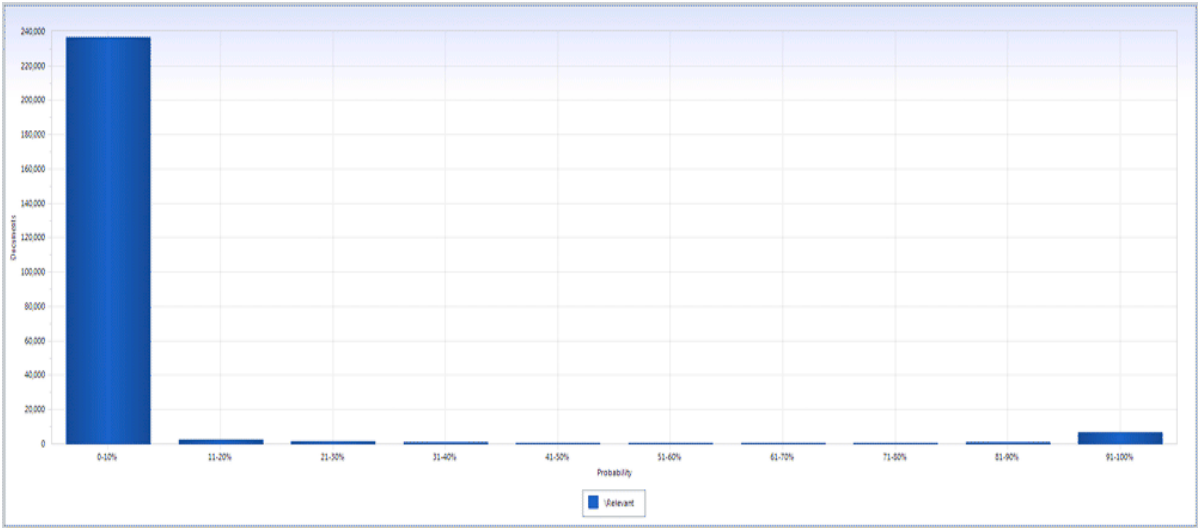
In multimodal projects where predictive coding is used the precision rates can often be very high. Lately I have been seeing that the second pool of documents, subject to the manual review has precision rates of at least 80%, sometimes even as high as 95% near the end of a CAL project. That means the final pool of documents produced is almost as large as the pool after the second filter.

Please remember that almost every document that is manually reviewed and coded after the Second Filter gets recycled back into the machine training process. This is known as *Continuous Active Learning* or **CAL,** and in my version of it at least, is multimodal and not limited to only high probability ranking searches. *See: Latest Grossman and Cormack Study Proves Folly of Using Random Search For Machine Training*– Part Two. In some projects you may just train for multiple iterations and then stop training and transition to pure manual review, but in most you will want to continue training as you do manual review. Thus you set up a CAL constant feedback loop until you are done, or nearly done, with manual review.

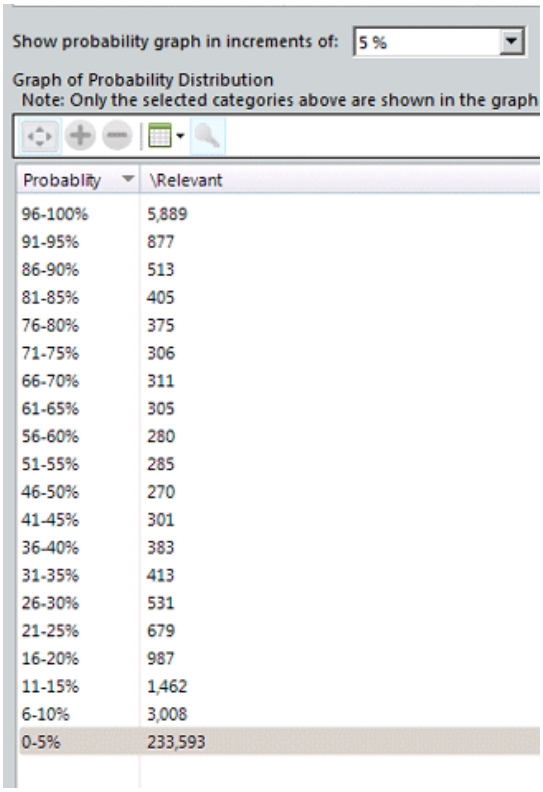
As mentioned, active machine learning trains on both relevance and irrelevance. Although, in my opinion, the documents found that are Highly Relevant, the hot documents, are the most important of all for training purposes. The idea is to use predictive coding to segregate your data into two separate camps, relevant and irrelevant. You not only separate them, but you also rank them according to probable relevance. The software I normally use, Kroll Ontrack's EDR, has a percentage system from .01% to 99.9% probable relevant and visa versa. A very good segregation-ranking project should end up looking like an upside down champagne glass.



A near perfect segregation-ranking project will end up looking like an upside down T with even fewer documents in the unsure middle section. If you turn the graphic so that the lowest probability relevant ranked documents are on the left, and the highest probable relevant on the right, a near perfect project ranking looks like this standard bar graph.



The above is a screen shot from a recent project I did after training was complete. This project had about a 4% prevalence of relevant documents, so it made sense for the relevant half to be far smaller. But what is striking about the data stratification is how polarized the groupings are. This means the ranking distribution separation, relevant and irrelevant, is very well formed. There are an extremely small number of documents where the AI is unsure of classification. The slow curving shape of irrelevant probability on the left (or the bottom of my upside down champagne glass) is gone.



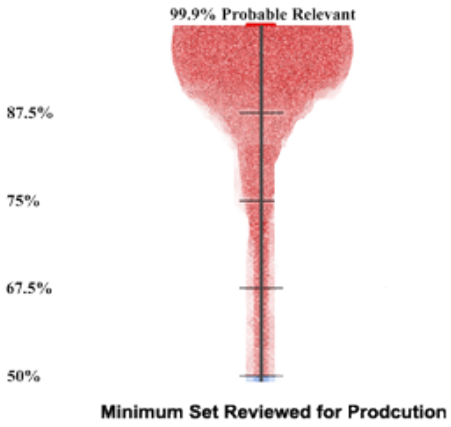
Probablity	\Relevant
96-100%	5,889
91-95%	877
86-90%	513
81-85%	405
76-80%	375
71-75%	306
66-70%	311
61-65%	305
56-60%	280
51-55%	285
46-50%	270
41-45%	301
36-40%	383
31-35%	413
26-30%	531
21-25%	679
16-20%	987
11-15%	1,462
6-10%	3,008
0-5%	233,593

The visualization shows a much clearer and complete ranking at work. The AI is much more certain about what documents are irrelevant. To the right is a screenshot of the table form display of this same project in 5% increments. It shows the exact *numerics* of the probability distribution in place when the machine training was completed. This

is the most pronounced polar separation I have ever seen, which shows that my training on relevancy has been well understood by the machine.

After you have segregated the document collection into two groups, and gone as far as you can, or as far as your budget allows, then you cull out the probable irrelevant. The most logical place for the *second filter* cut-off point in most projects in the 49.9% and less probable relevant. They are the documents that are more likely than not to be irrelevant. But do not take the 50% plus dividing line as an absolute rule in every case. There are no hard and fast rules to predictive culling. In some cases you may have to cut off at 90% probable relevant. Much depends on the overall distribution of the rankings and the proportionality constraints of the case. Like I said before, if you are looking for *Gilbert's* black-letter law solutions to legal search, you are in the wrong type of law.

Almost all of the documents in the production set (the red top half of the diagram) will be reviewed by a lawyer or paralegal. Of course, there are shortcuts to that too, like duplicate and near-duplicate syncing. Some of the documents in the irrelevant low ranked documents will have been reviewed too. That is all part of the CAL process where both relevant and irrelevant documents are used in training. If all goes well, however, only a few of the very low percentage probable relevant documents will be reviewed.



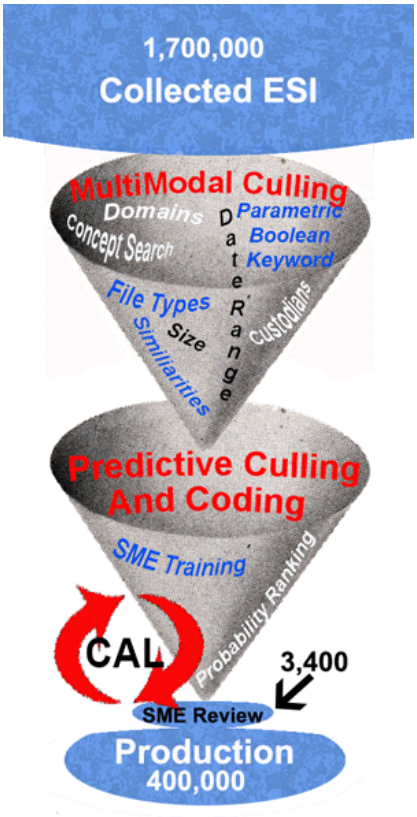
Limiting Final Manual Review

In some cases you can, with client permission (often *insistence*), dispense with attorney review of *all or near all* of the documents in the upper half. You might, for instance, stop after the manual review has attained a well-defined and stable ranking structure. You might, for instance, only have reviewed 10% of the probable relevant documents (top half of the diagram), but decide to produce the other 90% of the probable relevant documents without attorney eyes ever looking at them. There are, of course, obvious problems with privilege and confidentiality to such a strategy. Still, in some cases, where appropriate clawback and other confidentiality orders are in place, the client may want to risk disclosure of secrets to save the costs of final manual review.

In such productions there are also dangers of imprecision where a significant percentage of irrelevant documents are included. This in turn raises concerns that an adversarial view of the other documents could engender other suits, even if there is some agreement for return of irrelevant. Once the bell has been rung, privileged or hot, it cannot be un-rung.

Case Example of Production With No Final Manual Review

In spite of the dangers of the *unringable bell*, the allure of *extreme cost savings* can be strong to some clients in some cases. For instance, I did one experiment using multimodal CAL with no final review at all, where I still attained fairly high recall, and the cost per document was ***only seven cents***. I did all of the review myself acting as the sole SME. The visualization of this project would look like the below figure.



Note that if the SME review pool were drawn to scale according to number of documents read, then, in most cases, it would be much smaller than shown. In the review where I brought the cost down to $0.07 per document I started with a document pool of about 1.7 Million, and ended with a production of about 400,000. The SME review pool in the middle was only 3,400 documents.

As far as legal search projects go it was an unusually high prevalence, and thus the production of 400,000 documents was very large. Four hundred thousand was the number of documents ranked with a 50% or higher probable prevalence when I stopped the training. I only personally reviewed about 3,400 documents during the SME review. I then went on to review another 1,745 documents after I decided to stop training, but did so only for quality assurance purposes and using a random

sample. To be clear, I worked alone, and no one other than me reviewed any documents. This was an *Army of One* type project.

Although I only personally reviewed 3,400 documents for training, I actually instructed the machine to train on many more documents than that. I just selected them for training without actually reviewing them first. I did so on the basis of ranking and judgmental sampling of the ranked categories. It was somewhat risky, but it did speed up the process considerably, and in the end worked out very well. I later found out that other information scientists often use this technique as well. *See eg.* *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery, SIGIR'14, July 6–11, 2014, at pg. 9.*

My goal in this project was recall, not precision, nor even F1, and I was careful not to over-train on irrelevance. The requesting party was much more concerned with recall than precision, especially since the relevancy standard here was so loose. (Precision was still important, and was attained too. Indeed, there were no complaints about that.) In situations like that the slight over-inclusion of relevant training documents is not terribly risky, especially if you check out your decisions with careful judgmental sampling, and quasi-random sampling.

I accomplished this review in two weeks, spending 65 hours on the project. Interestingly, my time broke down into 46 hours of actual document review time, plus another 19 hours of analysis. Yes, about one hour of *thinking and measuring* for every two and a half hours of review. If you want the *secret* of my success, that is it.

I stopped after 65 hours, and two weeks of calendar time, primarily because I ran out of time. I had a deadline to meet and I met it. I am not sure how much longer I would have had to continue the training before the training fully stabilized in the traditional sense. I doubt it would have been more than another two or three rounds; four or five more rounds at most.

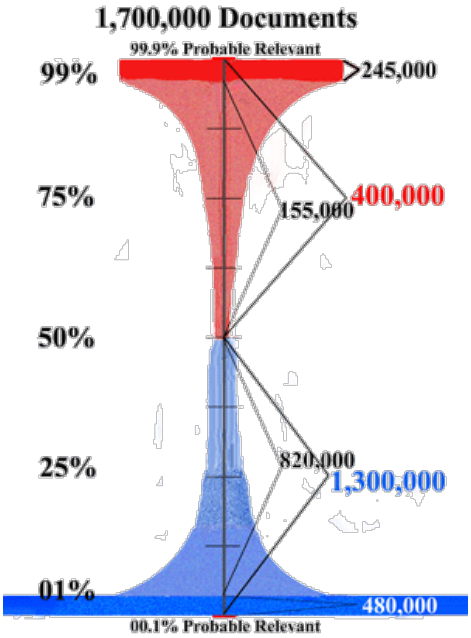
Typically I have the luxury to keep training in a large project like this until I no longer find any significant new relevant document types, and do not see any significant changes in document rankings. I did not think at the time that my culling out of irrelevant documents had been ideal, but I was confident it was good, and certainly reasonable. (I had not yet uncovered my ideal upside down champagne glass shape visualization.) I saw a slow down in probability shifts, and thought I was *close* to the end.

I had completed a total of sixteen rounds of training by that time. I think I could have improved the recall somewhat had I done a few more rounds of training, and spent more time looking at the mid-ranked documents (40%-60% probable relevant). The precision would have improved somewhat too, but I did not have the time. I am also sure I could have improved the identification of privileged documents, as I had only trained for that in the last three rounds. (It would have been a partial waste of time to do that training from the beginning.)

The sampling I did after the decision to stop suggested that I had exceeded my recall goals, but still, the project was much more rushed than I would have liked. I was also comforted by the fact that the elusion sample test at the end passed my *accept on zero error* quality assurance test. I did not find any hot documents. For those reasons (plus great weariness with the whole project), I decided *not* to pull some all-nighters to run a few more rounds of training. Instead, I went ahead and completed my report, added graphics and more analysis, and made my production with a few hours to spare.

A scientist hired after the production did some post-hoc testing that confirmed an approximate 95% confidence level recall achievement of between 83% to 94%. My work also confirmed all subsequent challenges. I am not at liberty to disclose further details.

In post hoc analysis I found that the probability distribution was close to the ideal shape that I now know to look for. The below diagram represents an approximate



depiction of the ranking distribution of the 1.7 Million documents at the end of the project. The 400,000 documents produced (obviously I am rounding off all these numbers) were 50% plus, and 1,300,000 not produced were less than 50%. Of the 1,300,000 Negatives, 480,000 documents were ranked with only 1% or less probable relevance. On the other end, the high side, 245,000 documents had a probable relevance ranking of 99% or more. There were another 155,000 documents with a ranking between 99% and 50% probable relevant. Finally, there were 820,000 documents ranked between 49% and 01% probable relevant.

The file review speed here realized of about 35,000 files per hour, and extremely low cost of about $0.07 per document, would not have been possible without the

client's agreement to forgo full document review of the 400,000 documents produced. A group of contract lawyers could have been brought in for second pass review, but that would have greatly increased the cost, even assuming a billing rate for them of only $50 per hour, which was 1/10th my rate at the time (it is now much higher.)

The client here was comfortable with reliance on confidentiality agreements for reasons that I cannot disclose. In most cases litigants are not, and insist on *eyes on review* of every document produced. I well understand this, and in today's harsh world of *hard ball litigation* it is usually prudent to do so, clawback or no.

Another reason the review was so cheap and fast in this project is because there were very little opposing counsel transactional costs involved, and everyone was hands off. I just *did my thing*, on my own, and with no interference. I did not have to talk to anybody; just read a few guidance memorandums. My task was to find the relevant documents, make the production, and prepare a detailed report – 41 pages, including diagrams – that described my review. Someone else prepared a privilege log for the 2,500 documents withheld on the basis of privilege.

I am proud of what I was able to accomplish with the two-filter multimodal methods, especially as it was subject to the mentioned post-review analysis and recall validation. But, as mentioned, I would not want to do it again. Working alone like that was very challenging and demanding. Further, it was only possible at all because I happened to be a subject matter expert of the type of legal dispute involved. There are only a few fields where I am competent to act alone as an SME. Moreover, virtually no legal SMEs are also experienced ESI searchers and software power users. In fact, most legal SMEs are *technophobes*. I have even had to print out key documents to paper to work with some of them.

Even if I have adequate SME abilities on a legal dispute, I now prefer to do a small team approach, rather than a solo approach. I now prefer to have one of two attorneys assisting me on the document reading, and a couple more assisting me as SMEs. In fact, I can act as the conductor of a predictive coding project where I have very little or no subject matter expertise at all. That is not uncommon. I just work as the software and methodology expert; the *Experienced Searcher*.

Recently I worked on a project where I did not even speak the language used in most of the documents. I could not read most of them, even if I tried. I just worked on procedure and numbers alone. Others on the team got their *hands in the digital mud* and reported to me and the SMEs. This works fine if you have good bilingual SMEs and contract reviewers doing most of the hands-on work.

Conclusion

There is much more to efficient, effective review than just using software with predictive coding features. The methodology of *how* you do the review is critical. The two-filter method described here has been used for years to cull away irrelevant documents before manual review, but it has typically just been used with keywords. I have show in this article how this method can be employed in a multimodal manner that includes predictive coding in the second filter.

Keywords can be an effective method to both cull out presumptively irrelevant files, and cull in presumptively relevant, but keywords are only one method among many. In most projects it is not even the most effective method. AI-enhanced review with predictive coding is usually a much more powerful method to cull out the irrelevant and cull in the relevant and highly relevant.

If you are using a one-filter method, where you just do a rough cut and filter out by keywords, date, and custodians, and then manually review the rest, you are reviewing too much. It is especially ineffective when you collect based on keywords. As shown in *Biome*t, that can doom you to low recall, no matter how good your later predictive coding may be.

If you are using a two-filter method, but are not using predictive coding in the second filter, you are still reviewing too much. The two-filter method is far more effective when you use relevance probability ranking to cull out documents from final manual review.