Predictive Coding Narrative: Searching for Relevance in the Ashes of Enron

Ralph C. Losey
Jackson Lewis, LLP

This is a narrative description of a legal search project using predictive coding. Follow along while I search for evidence of involuntary employee terminations in a haystack of 699,082 Enron emails and attachments.

Joys and Risks of Being First

To the best of my knowledge, this writing project is the first time anyone has written a blow-by-blow, detailed description of a large legal search and review project of any kind, much less a predictive coding project. Many experts on predictive coding speak only from a mile high perspective; never from the trenches (you can speculate why). That has been my practice too, until now, and also my practice when speaking about predictive coding on panels or in various types of conferences, workshops, and classes.

There are many good reasons for this, including the main one that lawyers cannot talk about their client's business or information. That is why in order to do this I had to run an academic project and search and review the Enron data. Many people *could* do the same. In fact, each year the TREC Legal Track participants do similar search projects of Enron data. But still, no one has taken the time to describe the details of their search, not even the spacey *TRECkies*.

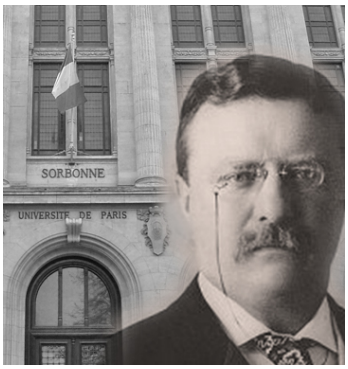
A search project like this takes an enormous amount of time. In my narrative I will report the amount of time that I put into the project on a day-by-day basis, and

---

Ralph C. Losey is a partner and the National e-Discovery Counsel of Jackson Lewis, LLP, where he lead's the firm's Electronic Discovery practice group. Ralph supports the 750 attorneys in his law firm located in 51 offices around the world. The opinions expressed in this article are his own, and not necessarily those of his law firm, clients, or University. Ralph can be reached at Ralph.Losey@Gmail.com. Ralph has a long history in general commercial litigation and employment litigation, but has limited his practice to electronic discovery since 2006. Ralph is the author of five books on electronic discovery published by West Thomson and the ABA, with his latest being an *iBook* by e-Discovery Team. Ralph is the publisher and principle author of the *e-Discovery Team Blog* at http://e-discoveryteam.com, on e-discovery opinion and analysis. Ralph is also the founder and principle editor of *Electronic Discovery Best Practices*, EDBP.com. Ralph is an adjunct law professor at the University of Florida on e-discovery subjects. The online classes he created for the University in 2011 are now available to the general public available at e-DiscoveryTeamTraining.com. Ralph has been involved with computers and the law since he began private practice in 1980. His full biography may be found at RalphLosey.com.

also, sometimes, on a per task basis. I am a lawyer. I live by the clock and have done so for thirty-two years now. Time is important to me, even non-money time like this. There is also a not-insignificant amount of time it takes to write it up a narrative like this. I did not attempt to record that.

There is one final reason this has never been attempted before, and it is not trivial: the risks involved. Any narrator who publicly describes their search efforts assumes the risk of criticism from *Monday morning quarterbacks* about *how the sausage was made*. I get that. I think I can handle the inevitable criticism. A quote that Jason R. Baron turned me on to a couple of years ago helps, the famous line from Theodore Roosevelt in his *Man in the Arena* speech at the Sorbonne:



> It is not the critic who counts: not the man who points out how the strong man stumbles or where the doer of deeds could have done better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood, who strives valiantly, who errs and comes up short again and again, because there is no effort without error or shortcoming, but who knows the great enthusiasms, the great devotions, who spends himself for a worthy cause; who, at the best, knows, in the end, the triumph of high achievement, and who, at the worst, if he fails, at least he fails while daring greatly, so that his place shall never be with those cold and timid souls who knew neither victory nor defeat.

I know this narrative is no *high achievement*, but we all do what we can, and this seems within my marginal capacities.

Desired Impact

I took the time to do this in the hope that such a narrative will encourage more attorneys and litigants to use predictive coding technology. Everyone who tries this new technology agrees it is the best way yet to find evidence in an economical manner. It is the best way to counter all those who would use *discovery as abuse*, and not a tool for truth. *See eg.:*

- *Discovery As Abuse*
- *E-Discovery Gamers: Join Me In Stopping Them*
- *Judge David Waxse on Cooperation and Lawyers Who Act Like Spoiled Children*

Predictive coding can finally put an end to this abuse. We can use these methods to search large volumes of ESI in a fast, efficient, and economical manner. We have to do this. It is imperative because the volumes of ESI continue to grow, and, along with this flood, the costs of discovery continue to spiral out of control. Despite all of our efforts at cooperation and professionalism, there are still far too many attorneys out there who take advantage of this situation and use discovery as a weapon to try to force defendants to settle meritless cases. *See, e.g.:*

- *Bondi v. Capital & Fin. Asset Mgmt. S.A.*, 535 F.3d 87, 97 (2d Cir. 2008)"This Court . . . has taken note of the pressures upon corporate defendants to settle securities fraud 'strike suits' when those settlements are driven, not by the merits of plaintiffs' claims, but by defendants' fears of potentially astronomical attorneys' fees arising from lengthy discovery.")
- *Spielman v. Merrill Lynch, Pierce, Fenner & Smith, Inc.*, 332 F.3d 116, 122-23 (2d Cir. 2003) ("The PSLRA afforded district courts the opportunity in the early stages of litigation to make an initial assessment of the legal sufficiency of any claims before defendants were forced to incur considerable legal fees or, worse, settle claims regardless of their merit in order to avoid the risk of expensive, protracted securities litigation.")
- *Lander v. Hartford Life & Annuity Ins. Co.*, 251 F.3d 101, 107 (2d Cir. 2001) ("Because of the expense of defending such suits, issuers were often forced to settle, regardless of the merits of the action. PSLRA addressed these concerns by instituting . . . a mandatory stay of discovery so that district courts could first determine the legal sufficiency of the claims in all securities class actions." (citations omitted))
- *Kassover v. UBS A.G.*, 08 Civ. 2753, 2008 WL 5395942 at *3 (S.D.N.Y. Dec. 19, 2008) ("PSLRA's discovery stay provision was promulgated to prevent conduct such as: (a) filing frivolous securities fraud claims, with an expectation that the high cost of responding to discovery demands will coerce defendants to settle; and (b) embarking on a 'fishing expedition' or 'abusive strike suit' litigation.")

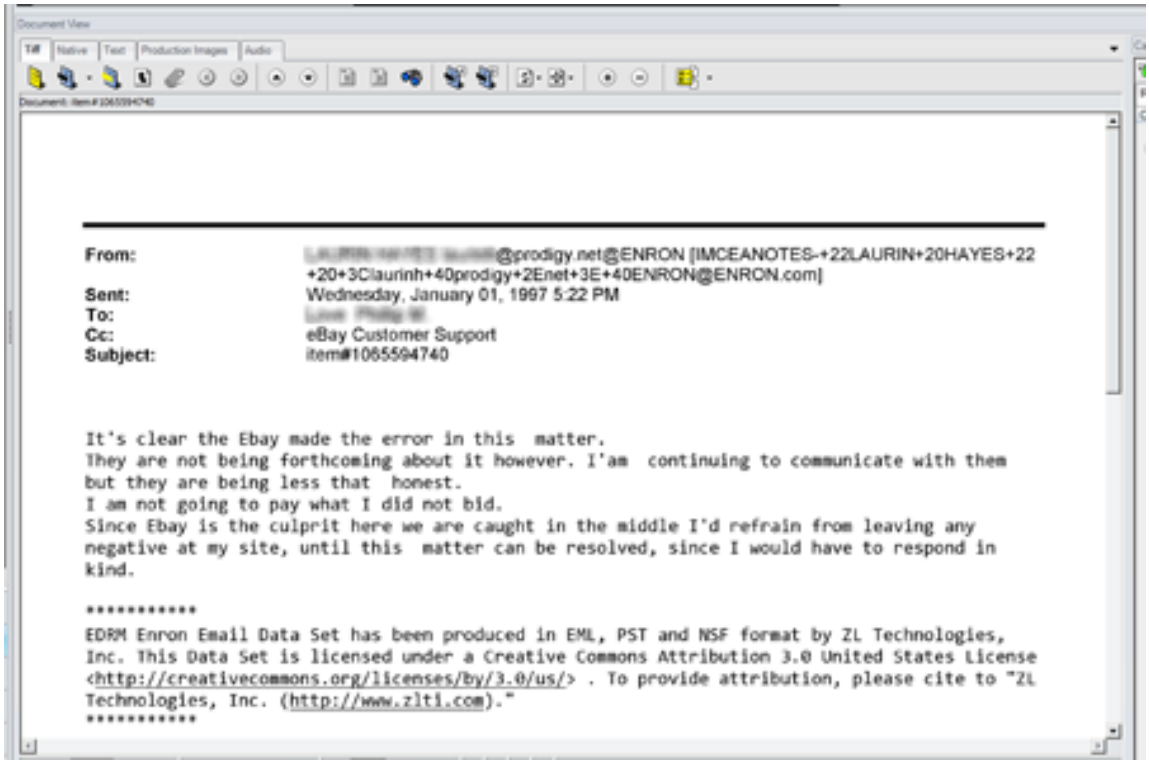
Follow me now while I search for relevance in the *ashes of Enron*.

699,082 Enron Documents: the Ashes of a Once-Great Empire

My search was of a special slice of the Enron database made by EDRM and Kroll Ontrack (“KO”). I conducted the search using the KO *Inview* software. This was a somewhat random selection of emails and attachments, not unlike what you would find by review of the PST files from a number of key custodians. It went in time from the late 90s, when the company was doing great, and involuntary terminations were rare, to its

eventual dissolution. In the early emails life for the 20,000 Enron employees was good and their email reflected that. Enron was growing and hiring. It was one of the hottest companies in America with revenues of over $100 Billion. But all of that changed rapidly near the end of the company, when it fell into bankruptcy in late 2001 and the emails slowly came to an end.

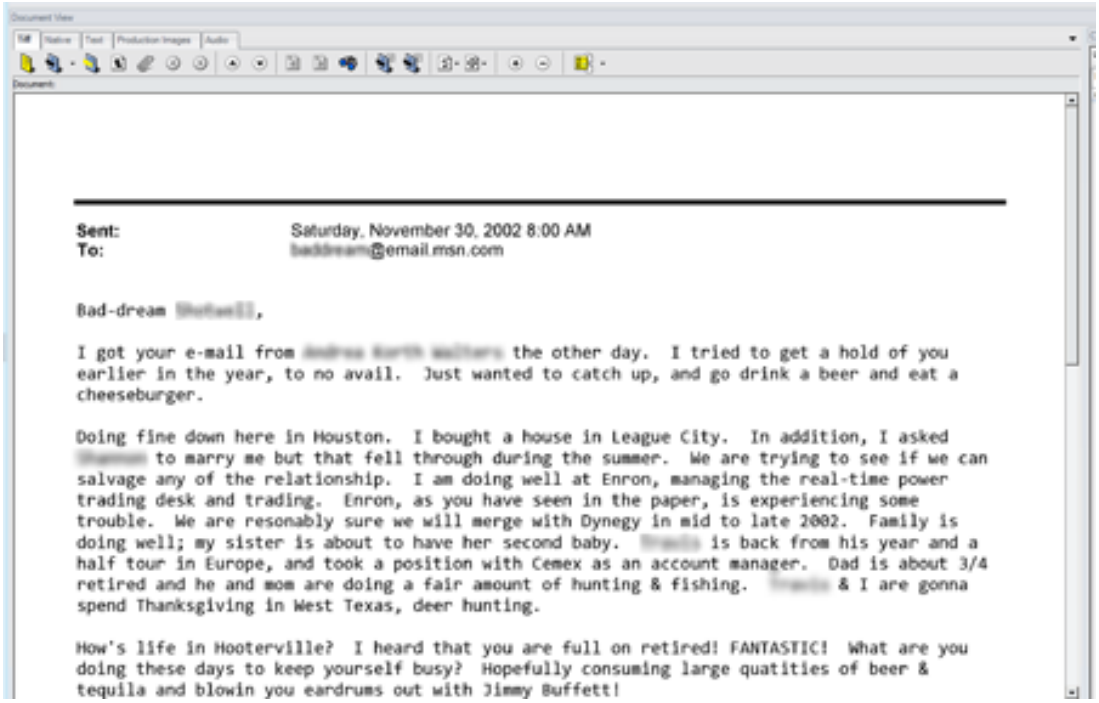
The date range in this Enron data collection, excluding a few outliers, ranged from January 1, 1997 to November 30, 2002. Here is a screen shot of one of the oldest and newest emails in the collection.



From: [redacted]@prodigy.net@ENRON [IMCEANOTES-+22LAURIN+20HAYES+22+20+3Claurinh+40prodigy+2Enet+3E+40ENRON@ENRON.com]
Sent: Wednesday, January 01, 1997 5:22 PM
To: [redacted]
Cc: eBay Customer Support
Subject: item#1065594740

It's clear the Ebay made the error in this matter.
They are not being forthcoming about it however. I'am continuing to communicate with them but they are being less that honest.
I am not going to pay what I did not bid.
Since Ebay is the culprit here we are caught in the middle I'd refrain from leaving any negative at my site, until this matter can be resolved, since I would have to respond in kind.

****
EDRM Enron Email Data Set has been produced in EML, PST and NSF format by ZL Technologies, Inc. This Data Set is licensed under a Creative Commons Attribution 3.0 United States License <http://creativecommons.org/licenses/by/3.0/us/> . To provide attribution, please cite to "ZL Technologies, Inc. (http://www.zlti.com)."
****



Sent: Saturday, November 30, 2002 8:00 AM
To: [redacted]@email.msn.com

Bad-dream [redacted],

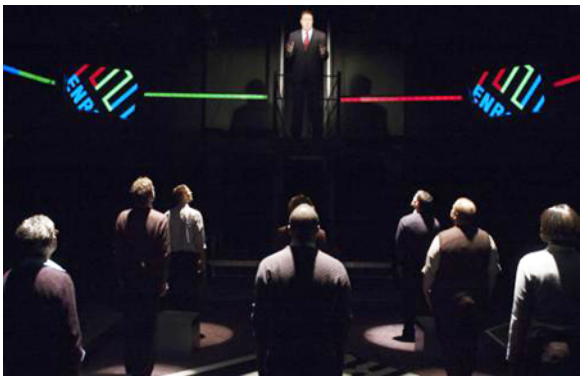
I got your e-mail from [redacted] the other day. I tried to get a hold of you earlier in the year, to no avail. Just wanted to catch up, and go drink a beer and eat a cheeseburger.

Doing fine down here in Houston. I bought a house in League City. In addition, I asked [redacted] to marry me but that fell through during the summer. We are trying to see if we can salvage any of the relationship. I am doing well at Enron, managing the real-time power trading desk and trading. Enron, as you have seen in the paper, is experiencing some trouble. We are resonably sure we will merge with Dynegy in mid to late 2002. Family is doing well; my sister is about to have her second baby. [redacted] is back from his year and a half tour in Europe, and took a position with Cemex as an account manager. Dad is about 3/4 retired and he and mom are doing a fair amount of hunting & fishing. [redacted] & I are gonna spend Thanksgiving in West Texas, deer hunting.

How's life in Hooterville? I heard that you are full on retired! FANTASTIC! What are you doing these days to keep yourself busy? Hopefully consuming large quatities of beer & tequila and blowin you eardrums out with Jimmy Buffett!

My search for involuntary employee termination related documents led me to focus on the final sad months when the empire fell apart, and ultimately the company itself was dissolved. The search involved fine relevancy distinctions between voluntary and involuntary terminations, with only involuntary being relevant. Such distinctions are common to all search projects, and this was no exception.

You would have to be a cold human indeed not to feel some of the pain of the thousands of people who ended up losing their jobs, their incomes, their retirement savings, because of the dastardly behavior of a few bad apples at the top. I have read their Enron email, which included many personal notes to family, friends, and even lovers. I have seen family photos, read their jokes and inspirational messages. I have even seen their porn and their cursing in anger. It is all there in the email, their life.



I have intruded into their privacy, uninvited, and unwelcome. For that I almost feel a need to apologize, but this is now public data, and my purpose was an academic study, not commercial or personal exploitation. Still, out of respect for the hundreds of people whose privacy I have necessarily invaded by the search of 699,082 emails and attachments, I will try not to

include any specific information in this narrative about these people and their lives. I owe them that much, and anyway, it seems like the decent thing to do. I do not think the omission in any way detracts from the value of the narrative.

Learn By Watching, Then Doing

The original point of the exercise was to provide training to a group of my firm's e-discovery liaisons who attended a KO training session in Minnesota. We trained by a demonstration of the use of the *Inview* software to respond to a hypothetical RFP. During the exercise emphasis was placed on use and description of the predictive coding features.

The feedback from my liaisons was that this was a good way to learn. This is not surprising because lawyers typically learn best by doing, and before we *do* something for the first time*,* in an ideal situation at least, we usually *observe* someone else who already knows how to do it. Any good law firm will, for instance, have a new associate watch an experienced partner take a few depositions before they let them take a deposition on their own. It is part of the legal apprenticeship program and one reason we call it the *practice* of law.

Overview of Efforts

I conducted this search over the course of eight workdays in May and June 2012. At the end of each day I sent out a description of what I had done. All of the lawyers in the training were invited to log onto the database and follow along. I have since edited these daily reports into a single narrative, all for general instruction purposes.

This Search is Just One Example Among Many

This narrative shows one example of the use of predictive coding in a typical legal setting. It is just one illustration, and many alternative approaches could have been followed. Indeed, if I were to do this over again, I would do many things different now that I have the benefit of 20/20 hindsight. Also, my knowledge of this particular software, Inview, has improved since this relatively early experiment, especially the ins and outs of how its predictive coding features work. It was, however, *not* my first such experiment with KO's *Inview*, not to mention my work with other vendors' software, each of which works slightly differently. As they say in Texas (and central Florida), *this was not my first rodeo.* Still, if I rode this particular bull again, I would do it differently. And, on any one day, I am sure that there are any number of people who could do it better, including many of my readers.

Best Practice

I do not contend that the particular search efforts here described were the *best* possible way the search of this data for this purpose could have been performed. Moreover, I readily admit that it was *not even close* to a perfect process. Perfection in legal search is never possible by anyone with any software. Perfection is never required by the law, in search and review, or anything else. I do contend, however, that the efforts here described constituted a *reasonable* search effort. It should, therefore, withstand any legal challenge as to its adequacy, since the law only requires reasonable search efforts.

Having said all that, to be honest, I think that the search here described was *fairly well done*. Otherwise I would not waste the reader's time with the description, nor use this narrative for instruction. Since I specialize in this stuff, and am considered an expert in legal search, particularly predictive coding, I would go a bit further and claim that my efforts were more than just adequate. (A quick footnote on my qualifications: I have over 30 years of experience searching for ESI on computers, a pending patent on one legal search method, and I am a published author and frequent speaker on the subject.) Right now predictive coding, and related legal doctrine and methods such as proportionality and *bottom line driven review*, are my primary interest in e-discovery. You could say I am obsessed and literally talk about it all of the time. *See eg.:* http://www.youtube.com/watch?v=4YFbXkAid6Q

Based on my background and experience, I think it is fair to contend that the search conducted was more than a mere legally adequate effort, more than just a reasonable effort. I would argue that it constitutes an example of a *best practice* of search and review. It qualifies as a best *practice* (as opposed to best possible) for two reasons: (1) advanced, predictive coding based software was used; and, (2) the search was conducted by a qualified expert. Still, the particular details and methods used in the search described in the narrative are *just one example* of a best practice; one among many possible approaches. Also, it is certainly *not a standard* to be followed in every search. It is important not to confuse those two things. Standards are more general. They are never single-case specific. They are never reviewer specific. So, after all of this long introduction, we finally come to the search narrative itself.

Come, Watson, come! The game is afoot.

First Day of Review (8.5 Hours)

The review began by judgmental sampling. I just looked around the 700-thousand documents to get a general idea of the types of documents in the dataset, the people involved, the date ranges, and the kinds of subjects their email addressed. This could have been done with reports generated by *Inview*, but I choose to do so by displaying all of them, and sorting the display in various

ways, including one of my favorites, display by file types. That allowed very easy viewing of the underlying documents whenever I wanted.

Another good method I could have used was the Analytics view with graphics displays providing visual information about the data. This includes a pull down menu where you can review certain files types. Or you can review by custodian with a visual display of who is emailing who. Or you can view by graphical displays of date ranges. These kind of visual displays of ESI contours are now common in most advanced software.

I also ran a few obvious, and some not so obvious, keyword searches pertaining to employee termination and other things.

By just looking around as I did, and running a few easy keyword searches, I found some relevant documents, and many more irrelevant ones. Although I was just beginning to familiarize myself with the data, I went ahead and coded some documents where it was obvious they were either relevant or irrelevant; the lowest hanging fruit, if you will. I coded 412 documents in this manner.

I call this *judgmental sampling* because I was using my own judgment to select and review small samples of the overall data. Before I began the predictive coding search process, I would also include random sampling, as this is core to all predictive coding methods. But, as is usual for me, I started here with judgmental sampling.

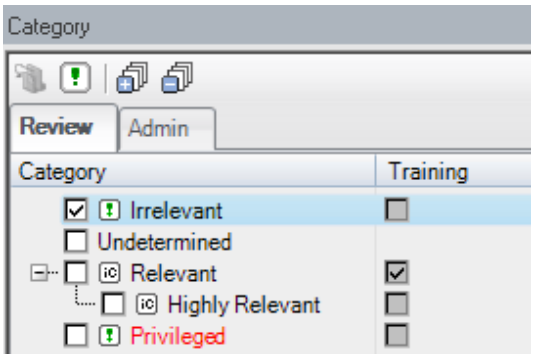
By the way, although I will be very transparent here, I am not going to tell you *exactly everything* I did. I am going to save a few trade secrets, a little bit of *secret sauce*, such as exactly what keyword searches I ran at the very beginning. As Maura Grossman likes to call such disclosure, its *translucent*, not *transparent*. I hope you understand.

Category Coding

I designed only five coding choices in this exercise:

- Irrelevant,
- Undetermined (relevancy),
- Relevant,
- Highly relevant (a sub-category of Relevant),
- Privileged

My categorization screen shown here included these categories, plus a box to check to tell the computer to train on the document. This training box is always optional. This will be explained further along in this narrative as we dive deeper into the predictive coding aspects of the search. The training button should only be checked on a category chosen for the document.



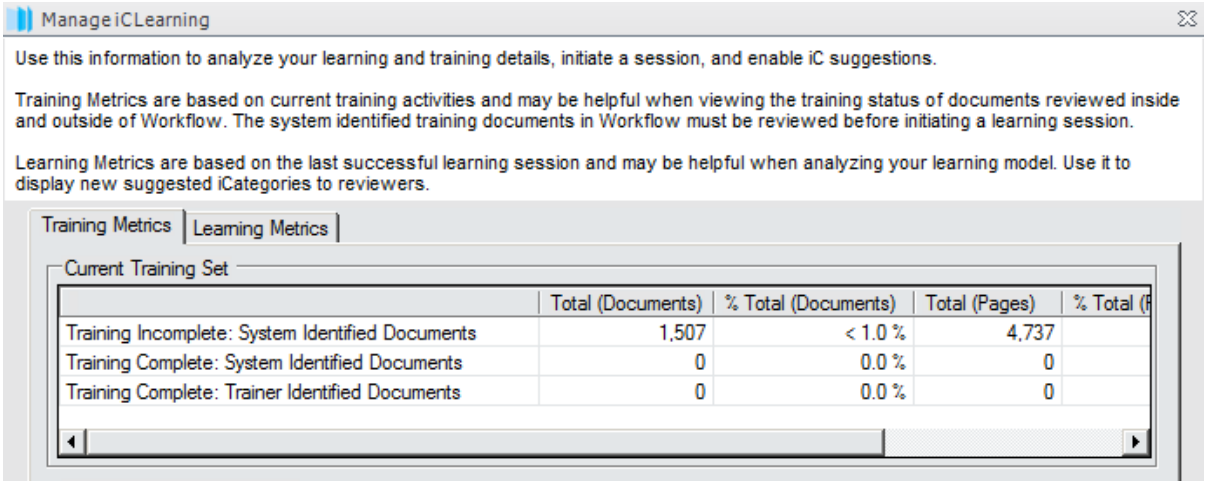
The screen shot here shows a cross-categorization error; the wrong Training box has been checked. The computer will not allow you to proceed with the error. Here you would either have to uncheck the Training button on Relevant, and check it instead on Irrelevant, or not at all. Alternatively, you could change the category check box to Relevant, and leave the Training box checked. This kind of consistency safeguard is present in all software systems that I looked at. Ask your vendor to confirm that they have similar consistency safeguards in the coding.

Regarding the Privileged category, I only ran into a few privileged documents that were relevant. When this happened, I of course marked them as such. But they were so rare as to not be valuable to describe here. The narrative will instead focus entirely on my search for relevancy.

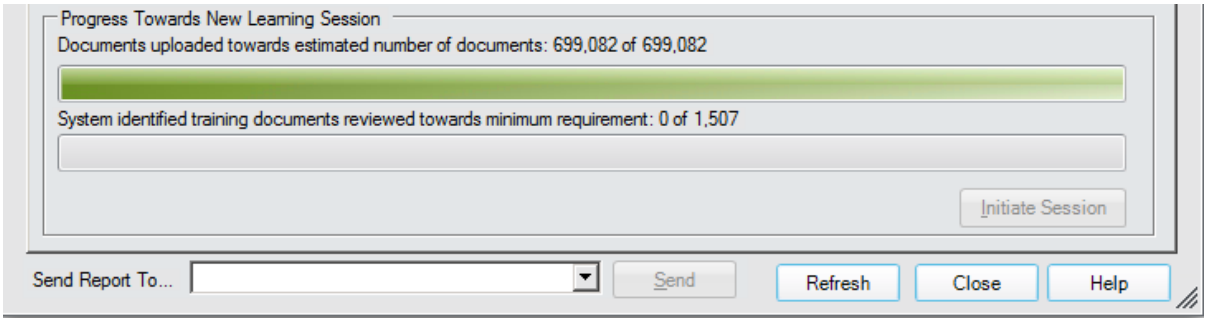
This initial orientation period lasted about **three hours.** (Whenever I report time herein, that is *billable type time*. I'm not including breaks or significant interruptions.)

First Predictive Coding Run

After this orientation I began the search and coding project in earnest by starting the predictive coding procedures of Inview. I began by generating the first random sample of the data. I used a 95% confidence level and a +/- 3% confidence interval. Based on these specifications the software randomly selected 1,507 documents. I'll explain that number soon, as observant readers will note it seems too high.



My first actual systematic coding was done by review of each of the 1,507 randomly selected documents. In the language of KO and Inview, these are called the *machine-selected documents* that will be used to train the system; a/k/a, the first seed set. They also served as the initial baseline for my Quality Control calculations, as will be explained later also.



I completed this review of the 1,507 documents in **5.5 hours.** After I completed that review the Initiate Session button shown above became active. At that point I could start a machine learning session, but not before.

I made an effort during the review to monitor my review speeds. I started this review at a review speed of about 200 documents per hour. Gradually as I got better with using the controls on my MacPro (first time I had used it for Inview review (loved it)), and as I gained closer familiarity with the stupid Enron documents, my speed went up to about 300 files per hour. I made liberal use of the bulk coding capabilities to attain these speeds, but was still careful. On a dual screen monitor, knowing what I now know about the kind of random docs I'm likely to see, and how to use the software and keyboard shortcuts, I think I could attain a speed of 400 files per hour, maybe even 500. Remember, this is only possible (for me at least) in review of null-set type collections, i.w. documents where almost all are irrelevant. It is much slower to review culled sets where there are 10% or more relevant documents. There you will be lucky to see 100 to

200 files per hour, even from top reviewers using clever sorting tricks and bulk coding.

Out of 1,507 items I reviewed, only 2 documents were identified as relevant. None was identified as highly relevant. Remember the goal is to find documents about employee termination (not contract termination, and not employee's voluntary termination, or retiring, etc.). Moreover, the ultimate goal is to find the few highly relevant documents about involuntary employee termination that might be used at trial.

Based on this first review of the 1,507 random documents (also called by Inview *System Identified documents*), plus my earlier casual review at the beginning of the day of 412 documents (called by Inview *Trainer Identified documents*), Inview went to work. I called it a night and let the computer take over.

The computers in the clouds (well actually they are in Eden Prairie, Minnesota, at KO's secure facility) then churned away for a couple of hours. (No, I do not record this as billable time!) The computer studied my *hopefully expert* input on relevance or irrelevance. While I slept, Inview analyzed the input, analyzed all of the 699,082 documents, and applied the input to the documents. It then ranked the likely relevance and irrelevance of all 699,082 documents from almost 0% to 100%. The first predictive coding seed set training then completed. All documents were now ranked and ready for me to review when I next logged on.

Day Two of a Predictive Coding Narrative: More Than A Random Stroll Down Memory Lane

Day One of the search project ended when I completed review of the initial 1,507 machine-selected documents and initiated the machine learning. I mentioned in the above Day One narrative that I would explain why the sample size was that high. I will begin with that explanation. On the original blog article of this narrative, I went even deeper into statistical sampling with the help of information scientist, William Webber. I omit that here, but if you are interested, see my original blog for the full technical details.

Before I begin the narrative proper for the second day I will also give you the big picture of my

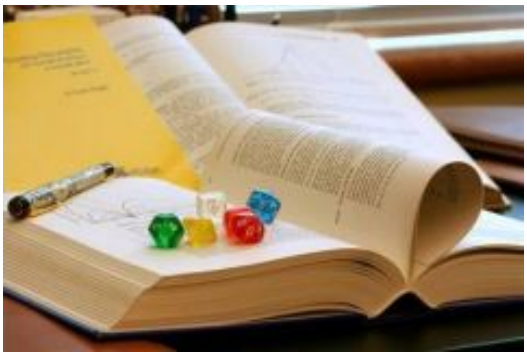
review plan and search philosophy: its *hybrid* and *multimodal*. Some search experts disagree with my philosophy. They think I do not go far enough to fully embrace machine coding. They are wrong. I will explain why and rant on in defense of humanity. Only then will I conclude with the Day Two narrative.

Why the 1,507 Random Sample Size to Start Inview's Predictive Coding

A pure random sample using 95% +/-3% and a 50% prevalence (the most conservative prevalence estimate) would require a sample of 1,065 documents. But Inview generates a larger sample of 1,507. This is because it uses what KO calls a *conservative* approach to sampling that has been reviewed and approved by several experts, including KO's outside consulting expert on predictive coding, David Lewis (an authority on information science and a co-founder of TREC Legal Track). In fact, this particular feature is under constant review and revisions are expected in future software releases.

Inview's uses a so-called *simple random sample* method in which each member of the population has an equal chance of being observed and sampled. But KO uses a larger than required minimum sample size because it uses a kind of continuous stream sampling where data is sampled at the time of input. That and other technical reasons explain the approximate 40% over-sampling in Inview, i.w., the use of 1,507 samples, instead of 1,065 samples, for a 95% +/-3% probability calculation.

This is typical of KO's conservative approach to predictive coding in general. The over-sampling adds slightly to the cost of review of the random samples (you must review 1,507 documents, instead of 1,065 documents). But this does not add that much to the cost. That is because the review of these sample sets goes fast, since almost all of them in most cases will be irrelevant. Review of irrelevant documents takes far less time on average than review of relevant documents. So I am convinced that this extra cost is really negligible, as compared to the increased defensibility of the sampling.



Since this approximate 40% larger than normal sample size is standard in Inview, even though the confidence level is supposedly only 3%, you can argue that in most datasets it represents an even smaller margin of error. A random sample of 1,507 documents in a dataset of this size would normally represent a 95% confidence interval with a margin of error (confidence interval) of only 2.52%, not 3%. *See* my prior blog on random sample calculations: *Random Sample Calculations And My Prediction That 300,000 Lawyers Will Be Using Random Sampling By 2022*.

Baseline Quality Control Sample Calculation

At the beginning of every predictive coding project I like to have an idea as to how many relevant documents there may be. For that reason I use the random sample that Inview generates for predictive coding training purposes, for another purpose entirely, for quality control purposes. I use the random sample to calculate the probable number of relevant documents in the whole dataset. Only simple math is required for this standard baseline calculation. For this particular search, where I found 2 relevant documents in the sample of 1,507 documents, it is: 2/1507=.00132714. We'll call that 0.13%. That is the percentage of relevant documents found in the whole, which is called the *prevalence,* a/k/a *density rate* or *yield.*

Based on this random sample percentage, my projection of the likely total number of relevant documents in the total database (aka *yield*) is **928** (.13%*699,082=928). So my general goal was to find 928 documents. That is called the *spot projection* or *point projection*. It represents a loose target or general goal for the search, a bulls-eye of sorts. It is not meant to be a *recall* calculation, of F1 measure, or anything like that. It is just a standard baseline for quality control purposes that many legal searchers use, not just me. It is, however, *not* part of the standard KO software or predictive coding design. I just use the random sample they generate for that secondary purpose.

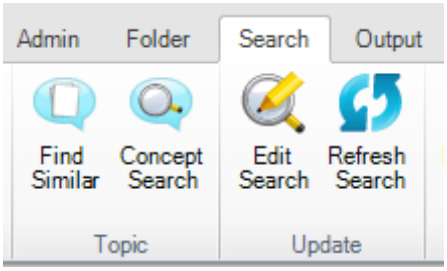
The KO random sampling is for an entirely different purpose of creating a machine training set for the predictive coding type algorithms to work. This is an important distinction to understand that many people miss. David Lewis had to explain that basic distinction to me many times before I finally got it. This distinction in the use of random samples is basic to all information science search, and is not at all unique to KO's Inview.

You need to be aware that there may well be *more or less* than the spot projection number of relevant documents in the collection (here 928). This is because of the limitations inherent in all random sampling statistics; the confidence intervals and levels. Here we used a confidence level (95%) and the confidence interval (+/- 3%). With a 3% confidence interval, there could be as many as 3,345 relevant documents or

as few as 112. See my original blog for the full mathematical explanation of that. For now, just know that it involves something in statistics called a *binomial confidence interval*.

Generating the Seed Set for Next Predictive Coding Session Using a Hybrid Multimodal Approach

I began day two with a plan to use any reviewer's most powerful tool, their brain, to find and identify additional documents to train Inview. My standard Search and Review plan is *multimodal*. By this I mean my standard is to use all kinds of search methods, in addition to predictive coding. The other methods include expert human review, the wetware of our own brains, and our unique knowledge of the case as lawyers who understand the legal issues, understand relevancy, and the parties, witnesses, custodian language, timeline, opposing counsel, deciding judge, appeals court, and all the rest of the many complexities that go into legal search.



I also include Parametric Boolean Keyword search, which is a standard type of search built into Inview and most other modern review software. This allows keyword search with Boolean logic, plus searches delimited to certain document fields and metadata.

I also include Similarity type searches using near duplication technology. For instance, if you find a relevant document, you can then search for documents similar to it. In Inview this is called *Find Similar*. You can even dial in a percentage of similarity. You can also do *Add Associated type* search methods which finds and includes all associated documents, like email family members and email threads. Again, these Similarity type search features are found in most modern review software today, not just Inview, and can be very powerful tools.

Finally, I used the Concept search methods to locate good training documents. Concept searches used to be the most advanced feature for software review tools, and is present in many good review platforms by now. This is a great way to harness the ability of the computer to know about linguistic patterns in documents and related keywords that you would never think of on your own.

Under a multimodal approach all of the search methods are used between rounds to improve the seed set, and predictive coding is not used as a stand-alone feature.

My plan for this review project is to limit the input of each seed set, of course, but to be flexible on the numbers and search time required between rounds, depending upon what conditions I actually encounter. In the first few rounds I plan to use keyword searches, and concept searches, and searches on high probability rank and mid-probability rank (the software's grey area) searches. I may use other methods depending again on how the search develops. My reviews will focus on the documents found by these searches. The data itself will dictate the exact methods and tools employed.

This multimodal, multi-search-methods approach to search is shown in the diagram below. Note IR stands for Intelligent Review, which is the KO language for predictive coding, a/k/a probabilistic coding. It stands at the top, but incorporates and includes all of the rest.

Some Vendors and Experts Disagree with Hybrid Multimodal

The multimodal approach is also encouraged by KO, which is one reason we selected KO as our preferred vendor. But not all software vendors and

experts agree with the multimodal approach. Some advocate use of pure predictive coding methods alone, and do not espouse the need or desirability of using other search methods to generate seed sets. In fact, some experts and vendors even oppose the Hybrid approach, which means equal collaboration between Man and Machine. They do so because *they favor of the Machine!* (Unlike some lawyers who go to the other extreme and distrust the machine entirely and want to manually review everything.)

The anti-hybrid, anti-multimodal type experts would, in this search scenario and others like it, proceed directly to another machine selected set of documents. They would rely entirely on the computer judgment and computer selection of documents. The human reviewers would only be used to decide on the coding of the documents that the computer finds and instructs them to review.

That is a mere *random stroll down memory lane*. It is not a bona fide Hybrid approach, any more than is linear review where the humans do not rely on the computers to do anything but serve as a display vehicle. That is the style of old-fashioned e-discovery where lawyers and paralegals simply do a manual linear review on a computer, but without any real computer assistance.

***Hybrid* for me means use of both the natural intelligence of humans, namely skilled attorneys with knowledge of the law, and the artificial intelligence of computers, namely advanced software with ability to learn from and leverage the human instructions and review tirelessly and consistently**.

Fighting for the Rights of Human Lawyers

I was frankly surprised to find in my due diligence investigation of predictive coding type software that there are several experts who have, in my view at least, a distinct *anti-human, anti-lawyer* bent. They have an anti-hybrid prejudice in *favor* of the computer. As a result, they have designed software that minimizes the input of lawyers. By doing so they have, in *their opinion*, created a pure system with better quality controls and less likelihood of human error and prejudice. Humans are weak-minded and tire easily. They are inconsistent and make mistakes. They go on and on about how their software prevents a lawyer from gaming the system, either intentionally and unintentionally. Usually they are careful in how they say that, but I have become sensitized after many such conversations and learned to read between the lines and call them on it.

These software designers want to take lawyers and other mere humans out of the picture as much as possible. They think in that way they will insulate their predictive model from bias. For instance, they want to prevent untrustworthy

humans, especially tricky lawyer types, from causing the system to focus on one aspect of the relevancy topic to the detriment of another. They claim their software has no bias and will look for all aspects of relevancy in this manner. (They try to sweep under the carpet the fact, which they dislike, that it is the human lawyers who train the system to begin with in what is or is not relevant.) These software designers put a new spin on an old phrase, and say *trust me, I'm a computer*.

You usually run into this kind of attitude when talking to software designers and asking them questions about the software, and pressing for a real answer, instead of the bs they often throw out. They are pretty careful about what they put into writing, as they realize lawyers are their customers, and it is never a good idea to directly insult your customer, or their competence, and especially not their honesty. I happened upon an example of this in an otherwise good publication by the EDRM on search, a *collaborative publication* (so we do not know who wrote this particular paragraph among the thousands in the publication) *EDRM Search Guide,* May 7, 2009, DRAFT v. 1.17, at page 80 of 83:

> In the realm of e-discovery, measurement bias could occur if the content of the sample is known before the sampling is done. As an example, if one were to sample for responsive documents and during the sampling stage, content is reviewed, there is potential for higher-level litigation strategy to impact the responsive documents. If a project manager has communicated the cost of reviewing responsive documents, and it is understood that responsive documents should somehow be as small as possible, that could impact your sample selection. To overcome this, the person implementing the sample selection should not be provided access to the content.

See what I am talking about? Yes, it is true lawyers *could* lie and cheat. But it is also true that the vast majority do not. They are honest. They are careful. They do not allow *higher-level litigation strategy to impact the responsive documents.* They do their best to *find* the evidence, not hide the evidence. Any software design built on the premise of the inherent dishonesty and frailty of mind of the users is inherently flawed. It takes human intelligence out of the picture based on an excessive disdain for human competence and honesty. It also ignores the undeniable fact that the few dishonest persons in any population, be it lawyers, scientists, techs, or software designers, will always find a way to lie, cheat, and steal. Barriers in software will not stop them.

In my experience with a few information scientists, and many technology experts, many of them distrust the abilities of all human reviewers, but especially lawyers, to contribute much to the search process. I speculate they are like this because: (a) so many of the lawyers and lit-support people they interact with tend to be relatively unsophisticated when it comes to legal search and technology; or, (b) they are just crazy in love with computers and their own software and don't

particularly like people, especially *lawyer people*. I suppose they think the Borg Queen is quite attractive too. Whatever the reason, several of the predictive coding software programs on the market today that they have designed rely *too much* on computer guidance and random sampling to the neglect of lawyer expertise. (Yes. That is what I *really* think. And no, I will not name names.)

After enduring many such *experts* and their pitches, I find their anti-lawyer, anti-human intelligence attitude offensive. I for one will *not* be assimilated into the Borg hive-mind. I will fight for the rights of human lawyers. I will oppose the borg-like software. Resistance is not futile!

The Borg-like experts design fully automated software for drones. Their software trivializes user expertise and judgment. The single-modal software search systems they promote underestimate the abilities (and honesty) of trained attorneys. They also underestimate the abilities of other kinds of search methods to find evidence, i.e., concept, similarity, and keyword searches.

I promote diversity of search methods and intelligence, but they do not. They rely too much on the computer, on random sampling, and on this one style of search. As a result, they do not properly leverage the skills of a trained attorney, nor take advantage of all types of programming.

In spite of their essentially hostile attitude to lawyers, I will try to keep an open mind. It is possible that a pure computer, pure probabilistic coding method may *someday* surpass my multimodal hybrid approach that still keeps humans in charge. Someday a *random stroll down memory lane* may be the way to go. But I doubt it.

In my opinion, legal search is different from other kinds of search. The goal of relevant evidence is inherently fuzzy. The ***7±2 Rule*** reigns supreme in the court room, a place where most such computer geeks have never even been, much less understand. Legal search for possible evidence to use at trial will, in my opinion, always require trained attorneys to do correctly. It is a mistake to try to replace them entirely with machines. Hybrid is the only way to go.

So, after this long random introduction, and *rant in favor of humanity*, I finally come to the narrative itself about Day Two.

Second Day of Review (3.5 Hours)

I was disappointed at the end of the first day that I had not found more relevant documents in the first random sample. I knew this would make the search more difficult. But I wanted to stick with this hypothetical of involuntary terminations and run through multiple seed sets to see what happens. Still, when I do this again with this same data slice, and that is the current plan for the next set of trainees, I will use another hypothetical, one where I know I will find more hits (higher prevalence), namely a search for privileged documents.

I started my second day by reviewing all of the 711 documents containing the term "firing." I had high hopes I would find emails about firing employees. I did find a couple of relevant emails, but not many. Turns out an energy company like Enron often used the term *firin g*to refer to starting up coal furnaces and the like. Who knew? That was a good example of the flexibility of language and the limitations of keyword search.

I had better luck with "terminat*" within 10 words of "employment." I sped through the search results by ignoring most of the irrelevant, and not taking time to mark them (although I did mark a few for training purposes). I found several relevant documents, and even found one I considered Highly Relevant. I marked them all and included them for training.

Next I used the "find similar" searches to expand upon the documents already located and marked as relevant documents. This proved to be a successful strategy, but I still had only found 26 relevant documents. It was late, so I called it a night. (It is never good to do this kind of work without rest, unless absolutely required.) I estimate my time on this second day of the project at three and a half hours.

Days Three and Four of a Predictive Coding Narrative: Where I find that the computer is free to disagree

The description of day-two was short, but it was preceded by a long explanation of my review plan and search philosophy, along with a rant in favor of humanity and against over-dependence on computer intelligence. Here I will just stick to

the facts of what I did in days three and four of my search.

Third Day of Review (4 Hours)

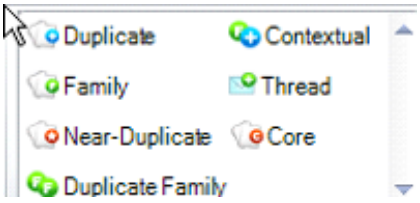
I continued to search for more relevant documents, or irrelevant documents, that would be useful in training for the next round of predictive coding. I started the day by running concept searches based on the first twenty-six documents that I had already identified as relevant. I also reviewed from 20 to 50 of the most similar documents per search (one search was especially good and I reviewed the top 100 similar). I only bothered to mark the relevant ones, most of which I also instructed to train. I only took time to mark a few irrelevant documents and marked them to train. These documents were close to relevant and I wanted to try to teach Inview the distinctions.

This exercise of reviewing about 1,500 docs took about **four hours**. It is faster when you don't have to mark (code) a document. I attained an average speed of 375 files per hour, even though I had a few documents that I really had to think and look carefully at to determine whether the termination was voluntary or not, or otherwise met the relevancy scope for this assignment. As a result of this exercise I have now found a total of 55 relevant docs, plus 8 more highly relevant docs (total 63).

I did not do any of the IRT ("Intelligent Review Technology") ranking based reviews at this time because I wanted to train the system more before investing time in that. I did not think *Inview* had enough relevant documents to train on from the first session, only 2. So I did not want to waste my time on ranking-driven based reviews yet.

The *Inview* system allows you the flexibility to do that. Some other predictive coding software do not. Still, for academic, record-keeping purposes, I did a search of all docs ranked 50% or higher probable relevant. It is interesting to observe that InView agreed with me because it did not rank any docs (-0-) as probable relevant (51% or more).

Fourth Day of Review (8 Hours)



I began the fourth day by attempting to expand upon the Highly Relevant docs found yesterday, and went from 8 to 14 hot documents. I did this my right-clicking on each of the six hot

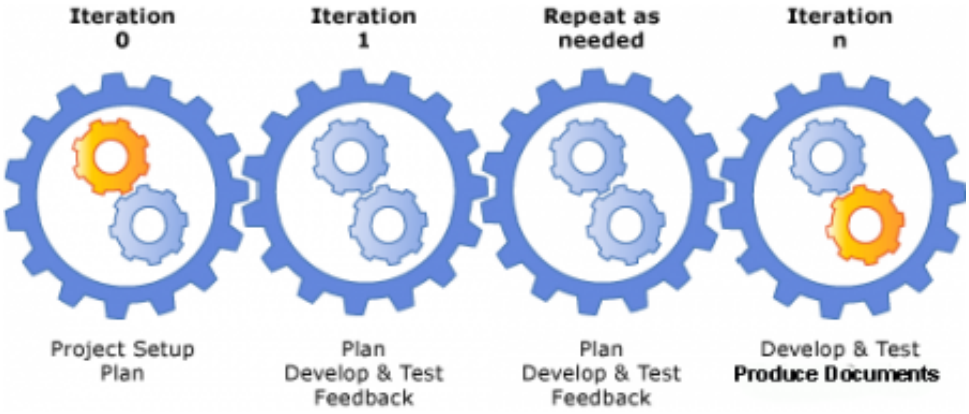
documents (yes, you can do that on a Mac too) that I started with. I then selected the *Add Associated Documents* from the drop down menu. That opens up seven more menu selections (Family Members, Threads and Attachments, Duplicates, Near-Duplicates, Core Near-Duplicates, Duplicate Family Members, and Contextual Duplicates).

I ran through most of them for all six docs (and the 8 additional this exercise found) to find the additional highly relevant docs. This took about ***three hours.*** I am sure I could do that faster the next time. I learned-by-doing, and by later instructions from KO experts, several new tricks to do this kind of expansion work quicker. Like anything, the more you work with new software, the faster you can go. My thousands of hours playing video games helped too. The next generations of gamers will be able to go even faster, I'm sure.

At this point I initiated another *learning session* (KO's language), and so we are now finally starting round two of predictive coding. I figure this might take an hour or more for the KO mainframe computers to run the *learning session*, so I signed off and did something else for a few hours. You do not have to stay connected for your commands to execute on the mainframes.

Second Round of Predictive Coding

The machine learning, a/k/a *auto coding* completed, and so at this point I had now run two seed set trainings, two auto coding iterations. Note this iterative design follows standard project management protocol for creative IT processes. See the diagrams below adapted from standard Microsoft project management illustrations.



At the end of the second iteration of seed set builds, the total number of training documents, by coincidence, was exactly 2,000 documents. 1,507 of the documents were selected at random by the computer (a/k/a “system identified”) in the first session, and another 493 were selected by me (“trainer identified”) and marked for training before the second session.

I then ran what Inview calls a *iC Relationships* search, looking for *iCategory Probability of Relevance* 51% or higher. It had 162 docs. Recall that after the first round, based only on 2 relevant documents, and the rest marked irrelevant, there were no -0- docs with probable (51% +) relevance predictions. Now after the second round, where we had 55 marked relevant in the seed set, the computer returns 162 as probable relevant, and marks the degree of likely relevance for each. In other words, it found 111 new docs that it thought were likely relevant. They had a probability range from 99.9% probable relevance to 54.3%.

The computer keeps going and marks below 51% probable relevance too, all the way down to 1%, but I did not include them in this search and review. I only wanted to see the documents the computer predicted would “more likely than not” be relevant. You could also just see the 90% plus docs, or whatever probability-range you wanted, depending on your purposes, including proportionality, i.w. the number of docs you could afford to have reviewed. You would just adjust the search parameters.

Inview gives you a lot of flexibility in the ways you can look at the probability sets of the total collection. Still, I was not satisfied, and complained about the work flow involved. (Personal note: I have never met a piece of software that I did not think needed improvement, nor a program I could not crash.) In a future version of Inview (coming soon), I have been promised that the display ranking will be

easier to determine. They will include a new column in the general display with a probability percentage ranking. Then all you will have to do is click on the probability % column to arrange the total documents display in either descending or ascending order. You can do something like that now by display of the IP Relevance score, which is too complicated to explain, and anyway is awkward and, to me at least, does not work as well.

Training Inview on Close Distinctions

One document is interesting to point out, control number 12005925, and is representative of many others. It is predicted to be 64.8% likely relevant. It is a one-page employee memo agreement having to do with payment of an end-of-year performance bonus. It mentions termination of employment as grounds for forfeiture of the bonus.

Document: 6151emailfromdrewmhc.doc

Confidential
Interoffice
Memorandum

Expires if not accepted by close of business on the November 30, 2001

To: «LAST_NAME», «FIRST_NAME» («GIS_ID») **Department:** «DEPT_NAME»

Subject: Performance Bonuses **Date:** November 29, 2001

This memorandum describes Performance Bonuses to be provided to you under the Enron Corp. Bonus Plan for calendar year 2001 performance, subject to the terms and conditions of this memorandum. You understand and agree that you will not receive any other cash performance bonus for calendar year 2001, whether described in a Company incentive plan, contract, letter, or otherwise, other than the Performance Bonus described in this memorandum.

I am pleased to inform you that you shall receive a cash Performance Bonus in the amount of $«M_2001_BONUS», less applicable taxes, as soon as practicable after you have accepted the terms of this Performance Bonus memorandum by signing below. To accept the terms of this Performance Bonus memorandum, you must sign this memorandum by the close of business on November 30, 2001; after that time, this memorandum expires and the offer to pay a Performance Bonus is revoked.

You agree to repay 125% of the Performance Bonus in the event you voluntarily terminate your employment with Company within ninety (90) days after receipt of the Performance Bonus, or if you disclose the terms of the Performance Bonus to any other person or entity, except your spouse, attorney, or financial advisor; such repayment shall be made within thirty (30) days after your last date of employment at Company. By accepting the payment of the Performance Bonus, you also authorize the Company to deduct from any wages or other amounts owed to you by the Company such amounts as may be necessary to satisfy your obligation to make repayment hereunder.

YOU UNDERSTAND AND AGREE THAT YOUR RECEIPT OF THIS CASH PERFORMANCE BONUS IS CONFIDENTIAL. ANY DISCLOSURE OF THE TERMS OR CONDITIONS OF THIS MEMORANDUM WILL RESULT IN CORRECTIVE ACTION, INCLUDING THE FORFEITURE OF THE PERFORMANCE BONUS PREVIOUSLY PAID; THE FORFEITED PERFORMANCE BONUS MUST BE REPAID WITHIN THIRTY (30) DAYS OF COMPANY'S REQUEST. Any bonus received is neither intended nor should be construed as being an addition to base salary or included in calculations of benefits or salary increases. This agreement does not provide you with any rights to continued employment of any specified duration.

The parties acknowledge their agreement and acceptance of the terms of this memorandum by signing below.

Enron Corp. **«LAST_NAME», «FIRST_NAME»**

By: _____ _____

Name: [redacted] This ___ day of November, 2001

Title: Vice President

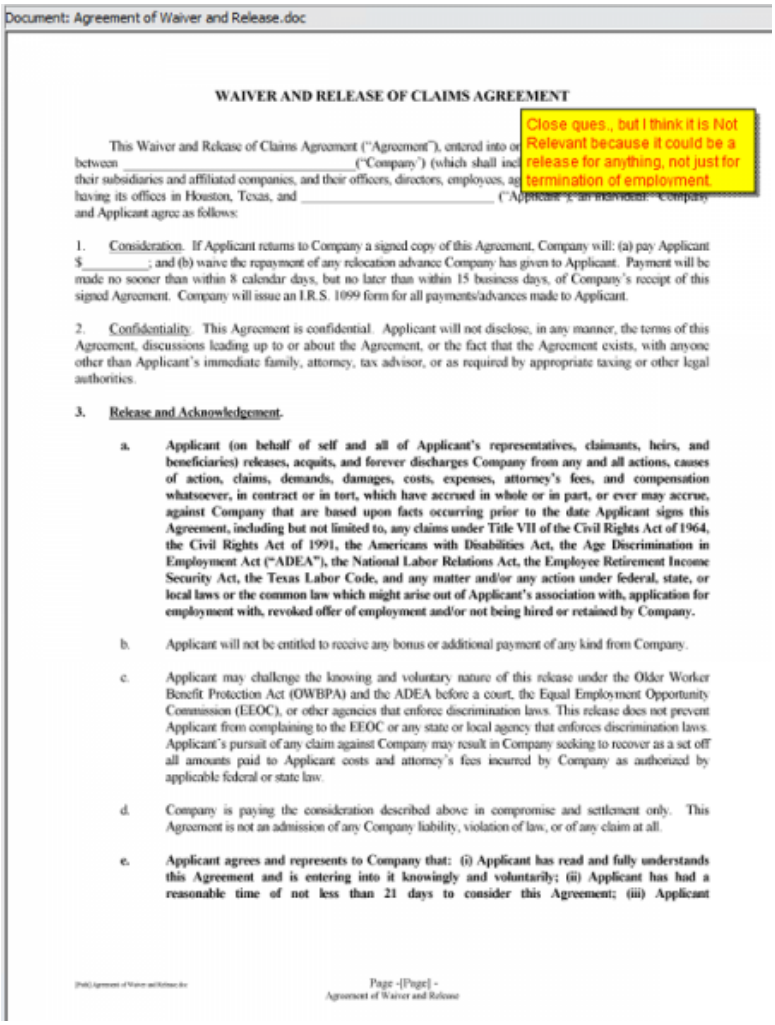
This ___ day of November, 2001

[Path] 6151emailfromdrewmhc.doc

I know that it looks and reads a lot like similar documents used for payouts when an employee is terminated, as in a short release. I suppose that is why the computer thought is was probably relevant, but *I knew* that it is not. So I marked it as irrelevant, and marked it to train, thus continuing the effort to try to teach the computer to distinguish between this document, that is irrelevant, and others similar to it that are relevant. In my experience with this and other predictive coding software, the training on such fine distinctions may take several rounds of instruction. (Same holds true with humans, by the way).

Computer is Free to Disagree with Me

Another interesting document to consider is control #12007393. I had already marked this as **irrelevant** after the first iteration. I took the time to add an electronic yellow sticky note with the comment that it was a close question, but I decided to call it irrelevant.



Document: Agreement of Waiver and Release.doc

WAIVER AND RELEASE OF CLAIMS AGREEMENT

Close ques., but I think it is Not Relevant because it could be a release for anything, not just for termination of employment.

This Waiver and Release of Claims Agreement ("Agreement"), entered into or
between _____("Company") (which shall inc
their subsidiaries and affiliated companies, and their officers, directors, employees, ag
having its offices in Houston, Texas, and _____ ("Applicant"), an individual. Company
and Applicant agree as follows:

1. Consideration. If Applicant returns to Company a signed copy of this Agreement, Company will: (a) pay Applicant $_____; and (b) waive the repayment of any relocation advance Company has given to Applicant. Payment will be made no sooner than within 8 calendar days, but no later than within 15 business days, of Company's receipt of this signed Agreement. Company will issue an I.R.S. 1099 form for all payments/advances made to Applicant.

2. Confidentiality. This Agreement is confidential. Applicant will not disclose, in any manner, the terms of this Agreement, discussions leading up to or about the Agreement, or the fact that the Agreement exists, with anyone other than Applicant's immediate family, attorney, tax advisor, or as required by appropriate taxing or other legal authorities.

3. **Release and Acknowledgement.**

a. **Applicant (on behalf of self and all of Applicant's representatives, claimants, heirs, and beneficiaries) releases, acquits, and forever discharges Company from any and all actions, causes of action, claims, demands, damages, costs, expenses, attorney's fees, and compensation whatsoever, in contract or in tort, which have accrued in whole or in part, or ever may accrue, against Company that are based upon facts occurring prior to the date Applicant signs this Agreement, including but not limited to, any claims under Title VII of the Civil Rights Act of 1964, the Civil Rights Act of 1991, the Americans with Disabilities Act, the Age Discrimination in Employment Act ("ADEA"), the National Labor Relations Act, the Employee Retirement Income Security Act, the Texas Labor Code, and any matter and/or any action under federal, state, or local laws or the common law which might arise out of Applicant's association with, application for employment with, revoked offer of employment and/or not being hired or retained by Company.**

b. Applicant will not be entitled to receive any bonus or additional payment of any kind from Company.

c. Applicant may challenge the knowing and voluntary nature of this release under the Older Worker Benefit Protection Act (OWBPA) and the ADEA before a court, the Equal Employment Opportunity Commission (EEOC), or other agencies that enforce discrimination laws. This release does not prevent Applicant from complaining to the EEOC or any state or local agency that enforces discrimination laws. Applicant's pursuit of any claim against Company may result in Company seeking to recover as a set off all amounts paid to Applicant costs and attorney's fees incurred by Company as authorized by applicable federal or state law.

d. Company is paying the consideration described above in compromise and settlement only. This Agreement is not an admission of any Company liability, violation of law, or of any claim at all.

e. **Applicant agrees and represents to Company that: (i) Applicant has read and fully understands this Agreement and is entering into it knowingly and voluntarily; (ii) Applicant has had a reasonable time of not less than 21 days to consider this Agreement; (iii) Applicant**

[Init.] Agreement of Waiver and Release.doc

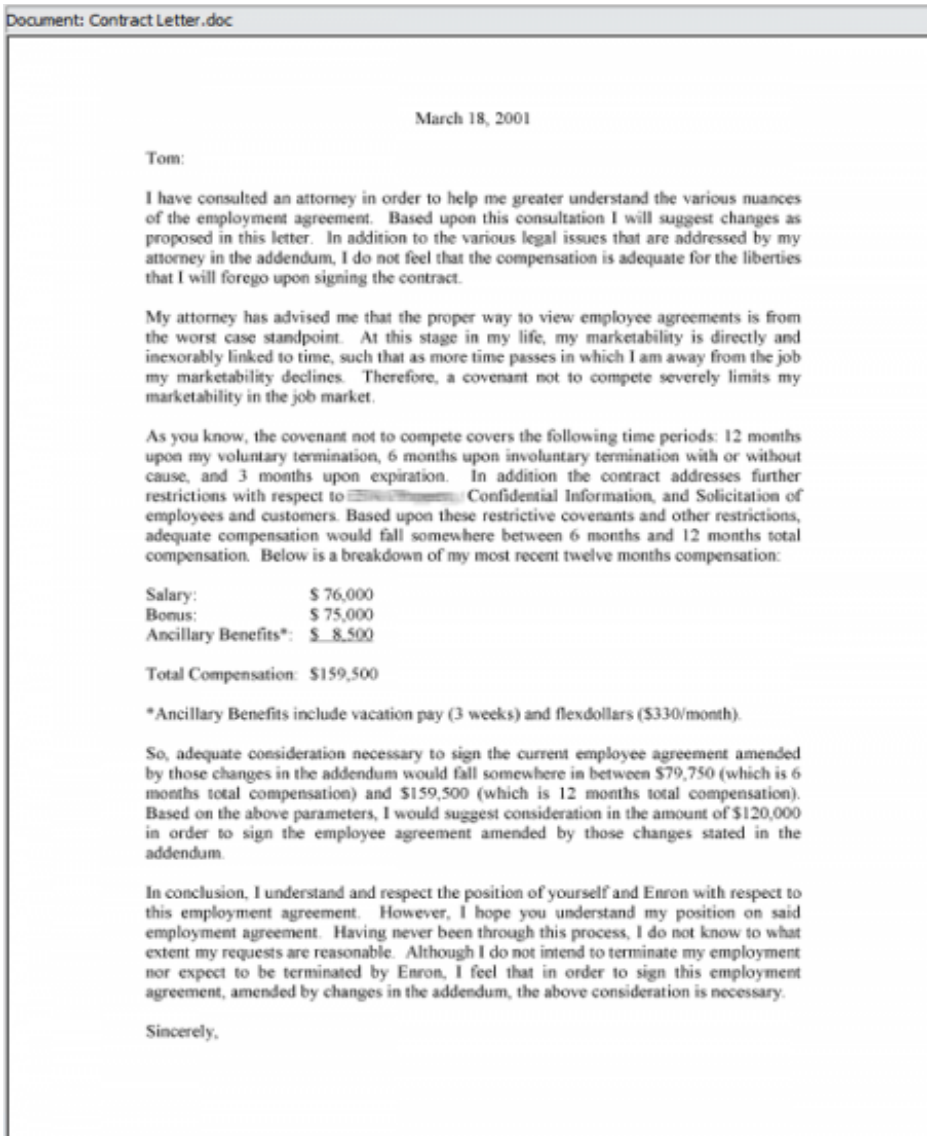
Page -[Page] -
Agreement of Waiver and Release

It is a 2-page form waiver and release of claims agreement. It could be used in the case of a termination, or maybe not. It is a form. **The computer's analysis of this document is that it was 77.5% likely relevant.** It did not change my coding of the document as irrelevant, but it did not let my prior marking of the document as irrelevant stand in its way of analysis that it probably was relevant. Interesting, eh? The computer is free to disagree with me.

This experience contradicts, or rather refines, the whole *GIGO* theory (garbage in, garbage out). You can make mistakes and the computer will correct, and suggest corrections. I looked carefully at the document again based on its input, but I stuck by my guns on this. I stayed with the irrelevant call. But perhaps with

another document I might be persuaded. Once you get used to the computer disagreeing with you, you start to realize what a cool feature it is. That is especially true when you consider the fuzziness of relevance and how it can change over time during a review and recall the complete consistency of computer code. Yes. I am a believer in hybrid, but not Borg where the computer makes the final decisions.

For another example of the computer disagreeing, slightly, see document with control number 8400149. I had marked this letter irrelevant, but the computer gave it an 54.5% chance of relevancy. It was a letter referencing consultation with a lawyer about an employment agreement, and what happens if employment is terminated. So I can understand the confusion, but again I am pretty certain I am right about that one.



Document: Contract Letter.doc

March 18, 2001

Tom:

I have consulted an attorney in order to help me greater understand the various nuances of the employment agreement. Based upon this consultation I will suggest changes as proposed in this letter. In addition to the various legal issues that are addressed by my attorney in the addendum, I do not feel that the compensation is adequate for the liberties that I will forego upon signing the contract.

My attorney has advised me that the proper way to view employee agreements is from the worst case standpoint. At this stage in my life, my marketability is directly and inexorably linked to time, such that as more time passes in which I am away from the job my marketability declines. Therefore, a covenant not to compete severely limits my marketability in the job market.

As you know, the covenant not to compete covers the following time periods: 12 months upon my voluntary termination, 6 months upon involuntary termination with or without cause, and 3 months upon expiration. In addition the contract addresses further restrictions with respect to [redacted] Confidential Information, and Solicitation of employees and customers. Based upon these restrictive covenants and other restrictions, adequate compensation would fall somewhere between 6 months and 12 months total compensation. Below is a breakdown of my most recent twelve months compensation:

Salary:	$ 76,000
Bonus:	$ 75,000
Ancillary Benefits*:	$ 8,500

Total Compensation: $159,500

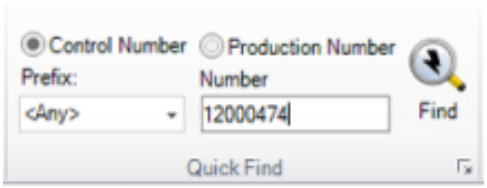
*Ancillary Benefits include vacation pay (3 weeks) and flexdollars ($330/month).

So, adequate consideration necessary to sign the current employee agreement amended by those changes in the addendum would fall somewhere in between $79,750 (which is 6 months total compensation) and $159,500 (which is 12 months total compensation). Based on the above parameters, I would suggest consideration in the amount of $120,000 in order to sign the employee agreement amended by those changes stated in the addendum.

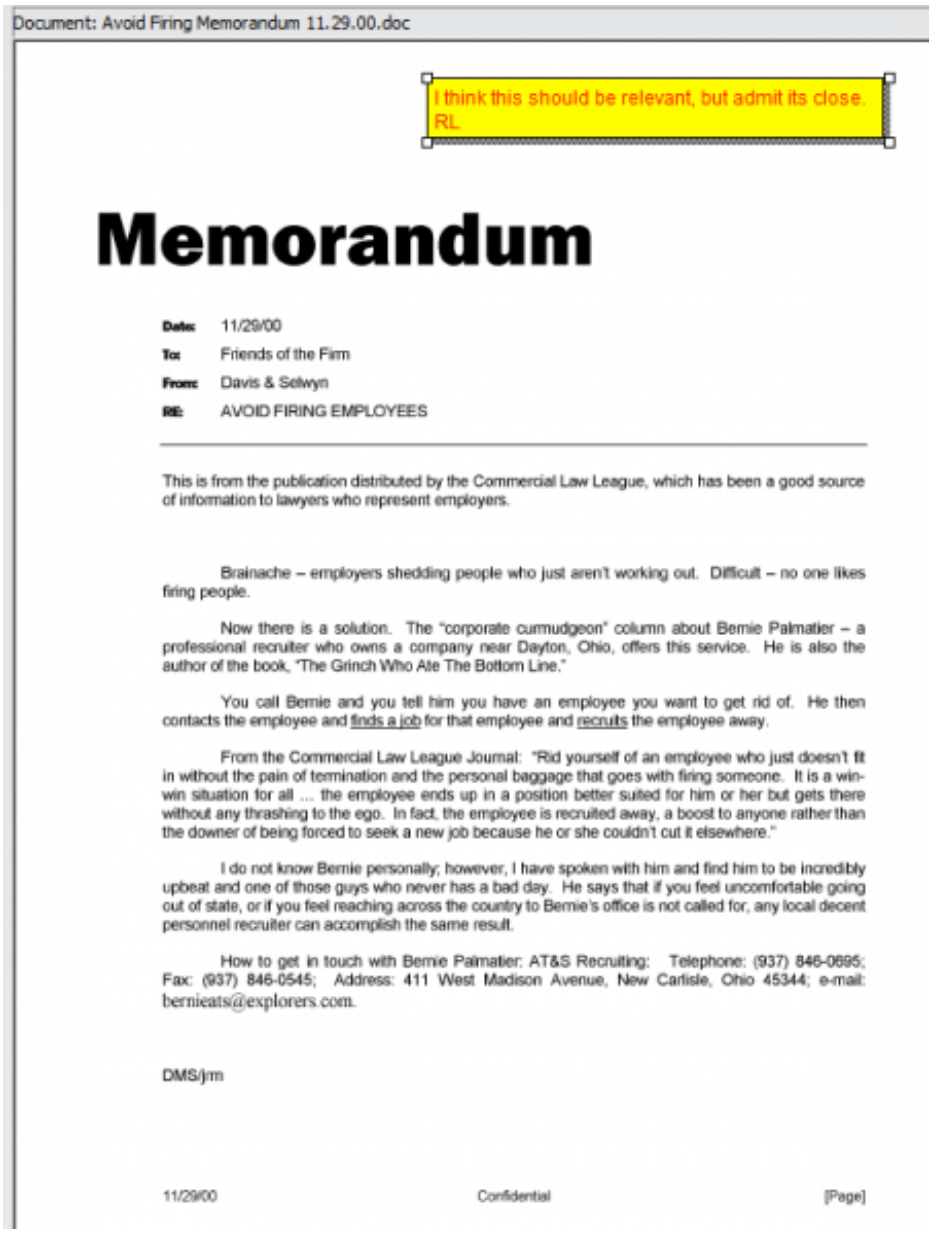
In conclusion, I understand and respect the position of yourself and Enron with respect to this employment agreement. However, I hope you understand my position on said employment agreement. Having never been through this process, I do not know to what extent my requests are reasonable. Although I do not intend to terminate my employment nor expect to be terminated by Enron, I feel that in order to sign this employment agreement, amended by changes in the addendum, the above consideration is necessary.

Sincerely,

For one more like this, see 12000474. By the way, to the right is a screen shot of how to quickly find any documents using the Inview control numbers. Each document has it own number, and if it is multiple pages, it has a beginning and end control number.



Again I had placed a sticky note in my prior review that this document was a close question. The computer agreed with me that it was relevant, but marked it as only 74.7% likely. It seems to understand that it is a close question too, even though I cannot share with the computer any more that yes or no. (In future versions of IRT we hope to see ranked relevancy, but that may take some time to perfect, and no one has that level of sophistication, yet.)



Document: Avoid Firing Memorandum 11.29.00.doc

I think this should be relevant, but admit its close. RL

Memorandum

Date: 11/29/00
To: Friends of the Firm
From: Davis & Selwyn
RE: AVOID FIRING EMPLOYEES

This is from the publication distributed by the Commercial Law League, which has been a good source of information to lawyers who represent employers.

Brainache – employers shedding people who just aren't working out. Difficult – no one likes firing people.

Now there is a solution. The "corporate curmudgeon" column about Bernie Palmatier – a professional recruiter who owns a company near Dayton, Ohio, offers this service. He is also the author of the book, "The Grinch Who Ate The Bottom Line."

You call Bernie and you tell him you have an employee you want to get rid of. He then contacts the employee and finds a job for that employee and recruits the employee away.

From the Commercial Law League Journal: "Rid yourself of an employee who just doesn't fit in without the pain of termination and the personal baggage that goes with firing someone. It is a win-win situation for all ... the employee ends up in a position better suited for him or her but gets there without any thrashing to the ego. In fact, the employee is recruited away, a boost to anyone rather than the downer of being forced to seek a new job because he or she couldn't cut it elsewhere."

I do not know Bernie personally; however, I have spoken with him and find him to be incredibly upbeat and one of those guys who never has a bad day. He says that if you feel uncomfortable going out of state, or if you feel reaching across the country to Bernie's office is not called for, any local decent personnel recruiter can accomplish the same result.

How to get in touch with Bernie Palmatier: AT&S Recruiting: Telephone: (937) 846-0695; Fax: (937) 846-0545; Address: 411 West Madison Avenue, New Carlisle, Ohio 45344; e-mail: bernieats@explorers.com.

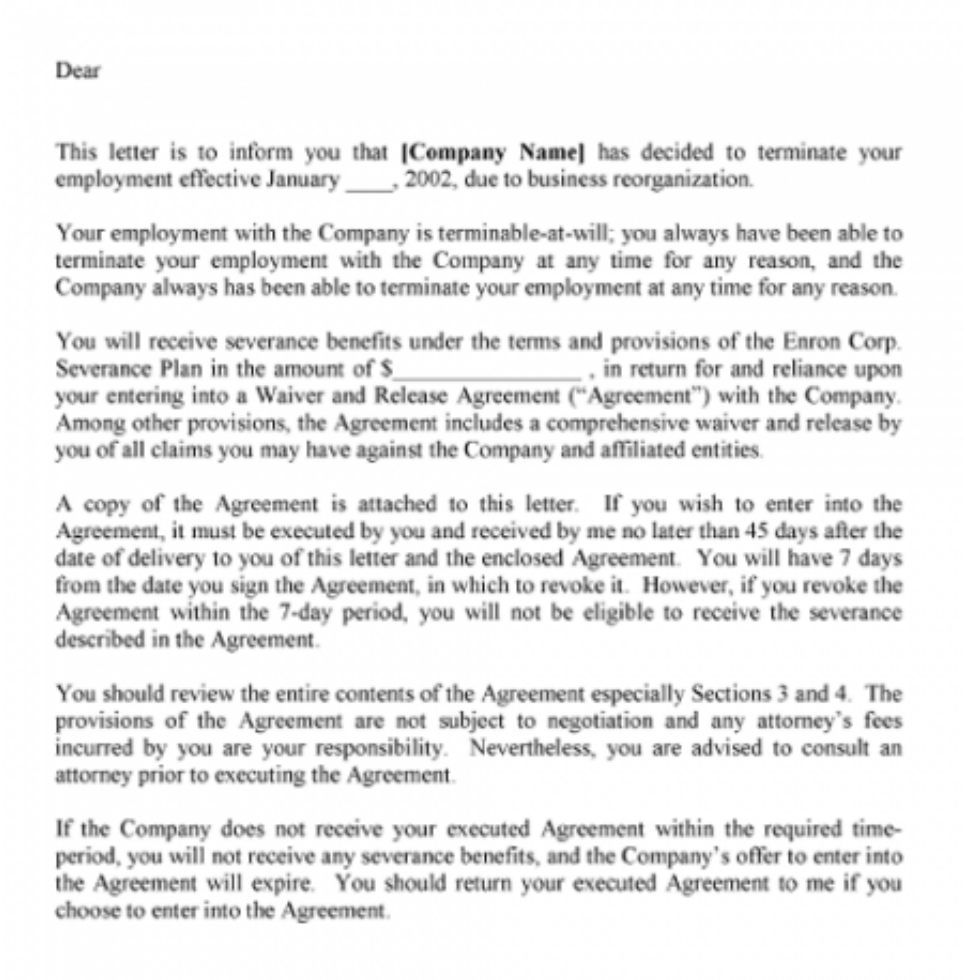
DMS/jrm

11/29/00 Confidential [Page]

I also searched for 51% probable Highly Relevant. I found 12 docs, only one of which was new, control number 12005880. It was a near duplicate of another email already marked highly relevant, which I had not seen before, or possible had seen but missed the connection (as it is not obvious). Recall we had found only 14 Highly Relevant documents before and trained on 13. So it's not surprising it only returned one new one, as it had so little to go on.

What is somewhat surprising is that the computer essentially disagreed with me on two documents that I had marked as hot (Highly Relevant). On control number 12006691 Inview gave only a 30.4% probability of hot, but did agree that it was relevant with a probability score of 80.3%. This was a form letter for termination only, and included payment of a severance. It looks very similar to forms used for

all-purpose employee departures. But recall we are going to define relevance as only *involuntary* terminations, not voluntary. This can be a subtle distinction when it comes to documentation. Do you agree with Inview or me on this? Hot or Not? (Hmm, might be a good name for a website?)



Dear

This letter is to inform you that **[Company Name]** has decided to terminate your employment effective January ____, 2002, due to business reorganization.

Your employment with the Company is terminable-at-will; you always have been able to terminate your employment with the Company at any time for any reason, and the Company always has been able to terminate your employment at any time for any reason.

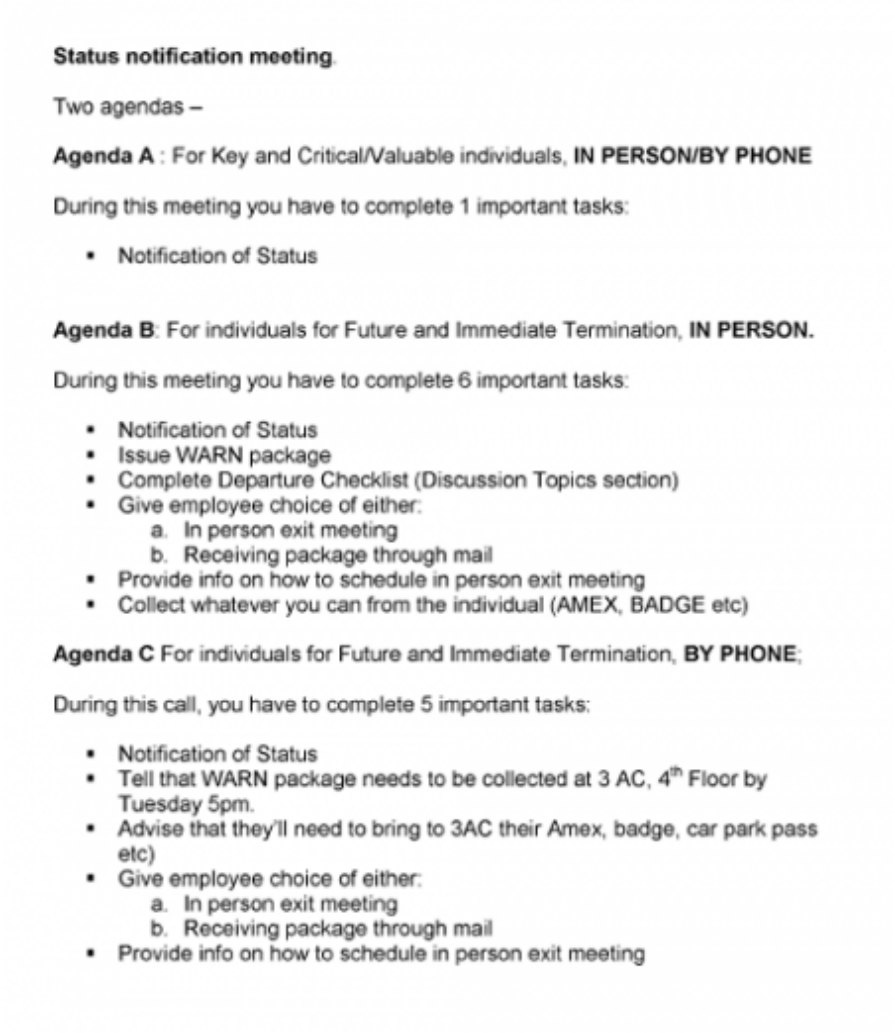
You will receive severance benefits under the terms and provisions of the Enron Corp. Severance Plan in the amount of $_____ , in return for and reliance upon your entering into a Waiver and Release Agreement ("Agreement") with the Company. Among other provisions, the Agreement includes a comprehensive waiver and release by you of all claims you may have against the Company and affiliated entities.

A copy of the Agreement is attached to this letter. If you wish to enter into the Agreement, it must be executed by you and received by me no later than 45 days after the date of delivery to you of this letter and the enclosed Agreement. You will have 7 days from the date you sign the Agreement, in which to revoke it. However, if you revoke the Agreement within the 7-day period, you will not be eligible to receive the severance described in the Agreement.

You should review the entire contents of the Agreement especially Sections 3 and 4. The provisions of the Agreement are not subject to negotiation and any attorney's fees incurred by you are your responsibility. Nevertheless, you are advised to consult an attorney prior to executing the Agreement.

If the Company does not receive your executed Agreement within the required time-period, you will not receive any severance benefits, and the Company's offer to enter into the Agreement will expire. You should return your executed Agreement to me if you choose to enter into the Agreement.

The other *Inview* disagreement with me on hot docs is on control # 12005730. *Inview* had only a 35.5% probability of Hot, and just 50.1% that it was even relevant. Jeesh! This is a short agenda documents with three scenarios. *Agenda A* employees are told they are important and are not fired. *Agenda B* employees get fired in person. Poor *Agenda C* employees get fired by phone. (Wonder if they left a message if no one answered?) Clearly *Inview* has not yet been trained as to the emotional impact (probative value) of certain kinds of documents.



Status notification meeting.

Two agendas –

Agenda A : For Key and Critical/Valuable individuals, **IN PERSON/BY PHONE**

During this meeting you have to complete 1 important tasks:

- Notification of Status

Agenda B: For individuals for Future and Immediate Termination, **IN PERSON.**

During this meeting you have to complete 6 important tasks:

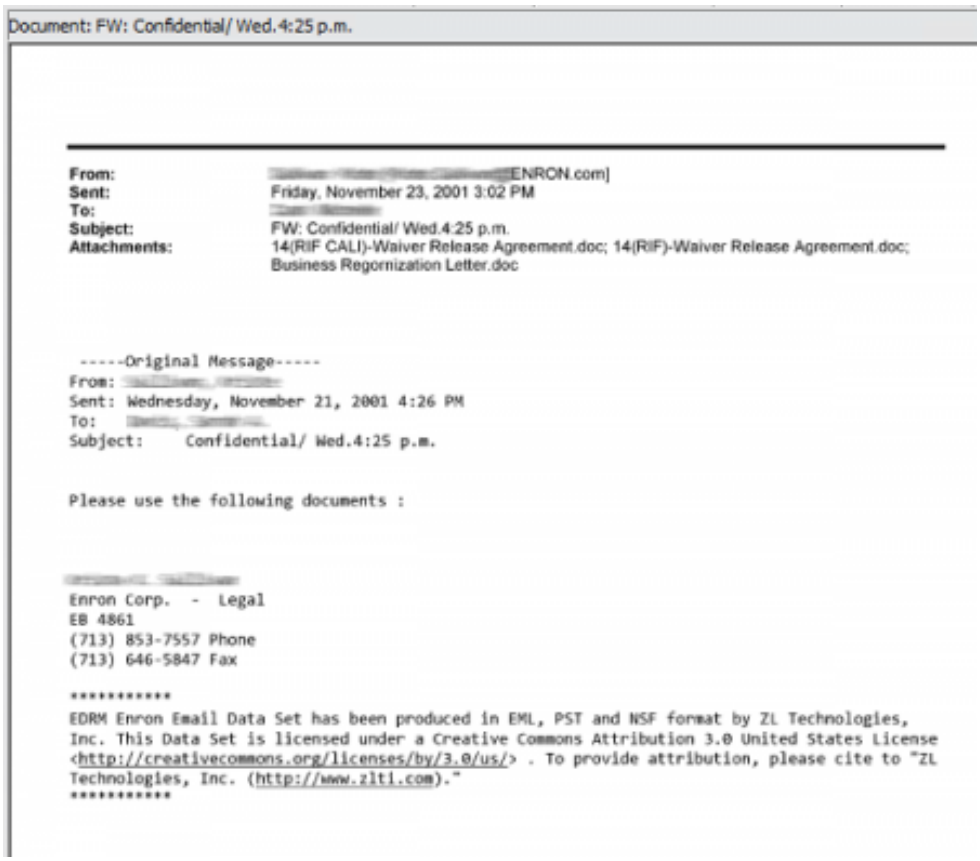
- Notification of Status
- Issue WARN package
- Complete Departure Checklist (Discussion Topics section)
- Give employee choice of either:
  a. In person exit meeting
  b. Receiving package through mail
- Provide info on how to schedule in person exit meeting
- Collect whatever you can from the individual (AMEX, BADGE etc)

Agenda C For individuals for Future and Immediate Termination, **BY PHONE**;

During this call, you have to complete 5 important tasks:

- Notification of Status
- Tell that WARN package needs to be collected at 3 AC, 4th Floor by Tuesday 5pm.
- Advise that they'll need to bring to 3AC their Amex, badge, car park pass etc)
- Give employee choice of either:
  a. In person exit meeting
  b. Receiving package through mail
- Provide info on how to schedule in person exit meeting

Finally, let's look at the one hot doc that I marked as hot, but decided not to train on it, because I thought it might throw *Inview* off the track. See control number 12006686. It is just an email transmittal that says "please use the following documents." The only thing hot about it is the attachment.



Document: FW: Confidential/ Wed.4:25 p.m.

From: [redacted]ENRON.com]
Sent: Friday, November 23, 2001 3:02 PM
To: [redacted]
Subject: FW: Confidential/ Wed.4:25 p.m.
Attachments: 14(RIF CALI)-Waiver Release Agreement.doc; 14(RIF)-Waiver Release Agreement.doc; Business Regornization Letter.doc

-----Original Message-----
From: [redacted]
Sent: Wednesday, November 21, 2001 4:26 PM
To: [redacted]
Subject: Confidential/ Wed.4:25 p.m.

Please use the following documents :

[redacted]
Enron Corp. - Legal
EB 4861
(713) 853-7557 Phone
(713) 646-5847 Fax

****
EDRM Enron Email Data Set has been produced in EML, PST and NSF format by ZL Technologies, Inc. This Data Set is licensed under a Creative Commons Attribution 3.0 United States License <http://creativecommons.org/licenses/by/3.0/us/> . To provide attribution, please cite to "ZL Technologies, Inc. (http://www.zlti.com)."
****

Inview said it was 97.1% sure it was *irrelevant*, with a 1% chance it was relevant, and a 1% chance it was hot. (It does not have to add up to 100%.) This is what I expected and it confirms my decision not to mark this stand-alone email as a Trainer in the first place. In fact, if Inview served this document up to me to review in the initial random sample, the so-called “System Identified” documents, where everything you code is automatically a trainer document for Inview auditing purposes, I would have marked it as irrelevant, even if I knew it to be hot due to its association with the attachment. It relevance was purely derivative.

Do not forget, there is no chance for under-production by doing this because we produce the underlying email if one of the attachments is marked for production. No orphan productions. My general rule is that the parent must always accompany the child, unless the requesting party or court does not want them. That is my standard, and I demand the same from parties producing to us. No orphans, please. Think of the poor little email children! But, as you will see later, I am not adverse to separating siblings. However, I am having second thoughts about that too.

This was a long day with a total of 8-hours of search and review work. This does not include the analysis time, nor time it took me to write this up. (The first draft of this narrative was written contemporaneously and shared daily with trainees who were invited to follow along.)

Days Five and Six of a Predictive Coding Narrative: Deep into the weeds and a computer mind-meld moment

In this fourth installment I continue to describe what I did in days five and six of the project. In this narrative I go deep into the weeds and describe the details of multimodal search. Near the end of day six I have an affirming *hybrid multimodal mind-meld moment,* which I try to describe. I conclude by sharing some helpful advice I received from Joseph White, one of Kroll Ontrack's (KO) experts on predictive coding and KO's Inview software. Before I launch into the narrative, a brief word about vendor experts. Don't worry, it is not going to be a commercial for my favorite vendors; more like a warning based on hard experience.

Vendor Experts

As part of your due diligence when selecting a vendor for any significant predictive coding project, I suggest that you interview the vendor *experts* that will be available to assist you, especially on the predictive coding aspects. They should have good knowledge of the software and the theory. They should also be able to explain everything to you clearly and patiently. They should not just be parrots of company white papers, or even worse, of sales materials and software manuals.

If a vendor expert truly understands, they can transcend the company jargon; they can rephrase so that you can understand. They can adapt to changing circumstances. The advice of a good vendor expert, one that not only understands the software, but also the law and the practical issues of lawyers, is invaluable. Periodic consults during a project can save you time and money, and improve the overall effectiveness of your search.

When talking to the experts, be sure that you understand what they say to you, and never just *nod in agreement* when you do not really *get it*. I have been learning and working with new computer software of all kinds for over thirty years, and am not at all afraid to say that I do not understand or follow something.

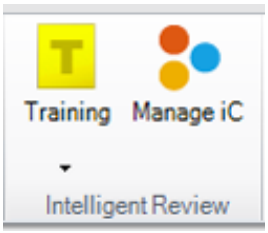
Often you cannot follow because the explanation is so poor. For instance, often the words I hear from vendor tech experts are too filled with company specific jargon. If what you are being told makes no sense to you, then say so. Keep asking questions until it does. Do not be afraid of looking foolish. You need to be able to explain this. Repeat back to them what you do understand in your own words until they agree that you have got it right. Do not just be a parrot. Take the time to understand. The vendor experts will respect you for the questions, and so will your clients. It is a great way to learn, especially when it is coupled with hands-on experience.

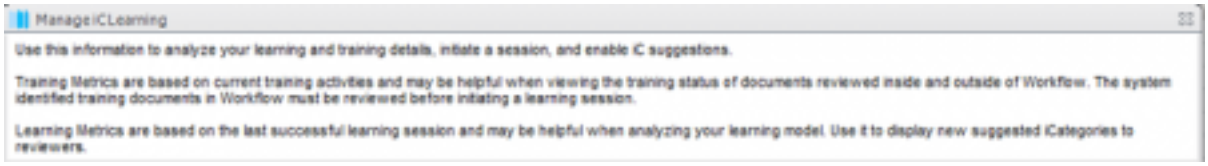
Fifth Day of Review (4 Hours)

I began the fifth day of work on the project by reviewing the 161 documents that I had found on my last day of working on this project. They were all predicted to be relevant (51% +). I had not finished reviewing them in the fourth day (which is reality was three-weeks ago). This first task took about one hour. Note that I elected not to train on all of them. This is an important degree of flexibility that Inview software provides, which others that I have seen do not.

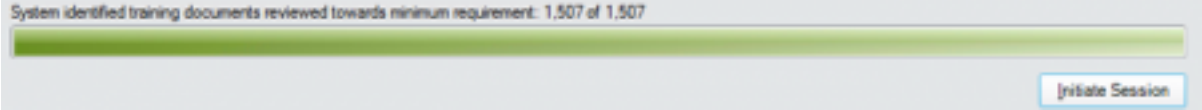
Third Round of Predictive Coding

Next, I ran the third iteration of predictive coding analysis by the software. That is called "*initiating a Session*" in Inview, a new learning session. The menu screen for this is found in *Workflow / Manage iC* (the three colored dots logo). Click on that *Manage iC* button and you open the *Manage iC Learning* page.





On this screen, below the opening splash shown above, you can initiate a training session. A partial screen shot showing the *Initiate Session* button is shown below.



After you click on the *Initiate Session* button, a message appears in red font saying: "*Learning session currently in progress*." This learning session can take over an hour or more, but that's computer time. It only took five minutes of my actual, billable time. The computer during this time is analyzing every document based on the new matrix. Put another way, the learning session is based on the new information, the new coding of documents that I marked for Training, that I provided today and in Day Four. All other trained documents are also considered.

Basically the only new coding I had provided to Inview between rounds two and three were my coding of the 162 docs predicted to be relevant (51% +), 111 of which were new, and the 12 predicted to be Highly Relevant, only 1 of which Hot documents was new. Again, I did not train on any grey areas, as I thought it was too early to look at that. Depending on what results we get from this third round, I may include that in the next training.

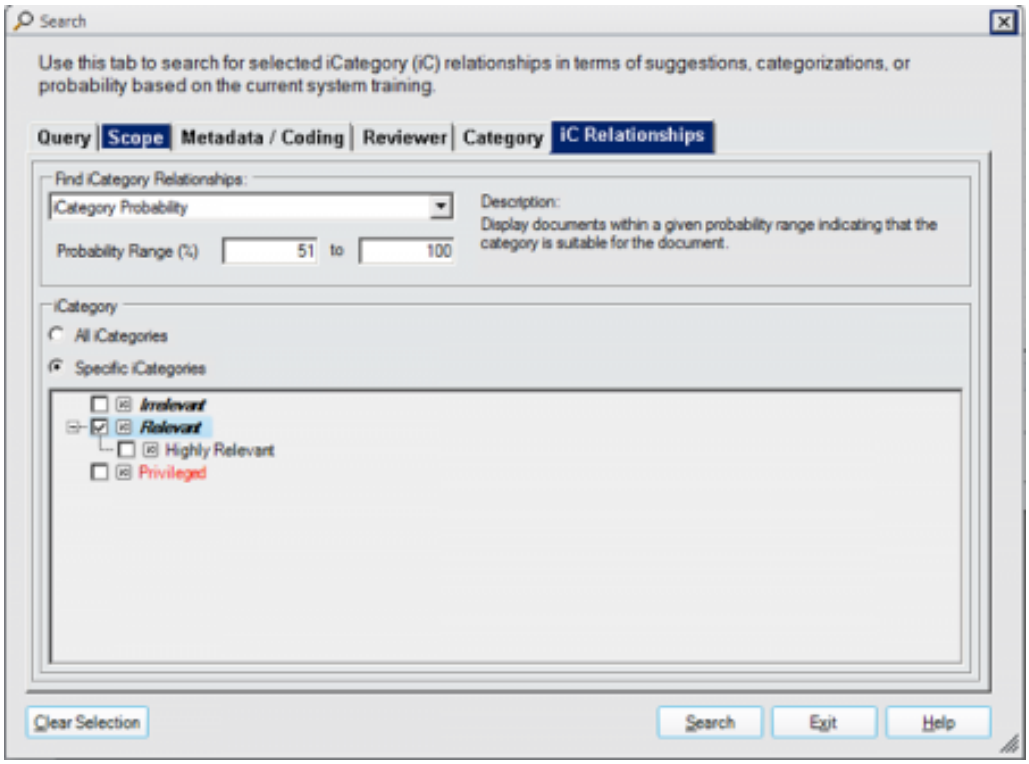
Please note that KO's Inview gives the Trainers (me and/or any other attorney with authority to manage the IRT process) the ability to pick and choose how we train. Other software is much more rigid and controlled, i.e. – they require review of grey area documents before each training, plus top ranked documents. I like the flexibility in KO's software. It gives some credit to the ability of lawyers as expert searchers, at least when it comes to evidence and legal classifications. For a fuller explanation of my preferred *hybrid* approach, where computers and lawyers work together, and my opposition to a total computer-controlled approach, which I have called *Borg-like*, *see Day Two of a Predictive Coding Narrative: More Than A Random Stroll Down Memory Lane,* (subsections *Some Vendors and Experts Disagree with Hybrid Multimodal* and *Fighting for the Rights of Human Lawyers*).

The third training session completed and the report stated that 534 *Trainer identified Docs* were used in the training, and that again there were 1,507 *System Identified Documents*. Recall that after the second training session completed there were 1,507 documents selected at random by the computer (a/k/a *system identified*) and 493 more documents selected by me (*trainer identified*) and marked for training, for a total of exactly 2,000. This meant I had only added 41 new documents for training since in the last session.

More Searching After the Training

I then ran a search to find the 51% probable relevant documents after the third training. Below is a screen shot of how that search is run. You could plug in whatever probability range you wanted. You could also include a variety of other

search components at the same time. Inview, like most state-of-the-art review platforms, has all kinds of very powerful search functions.

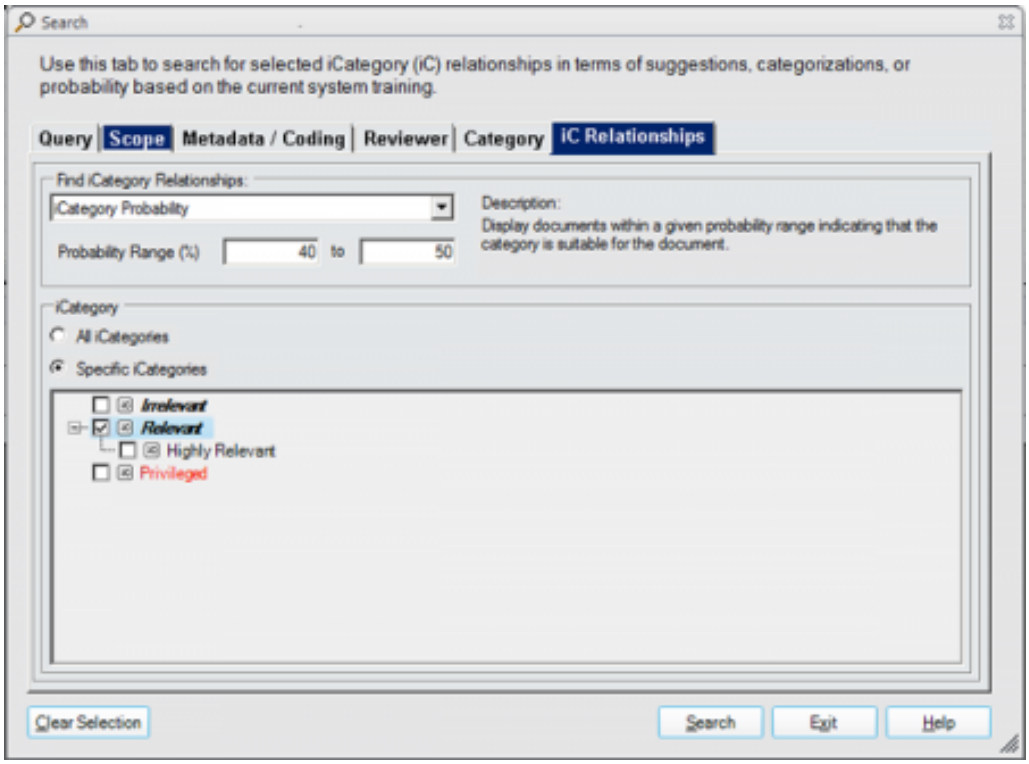


This search found 132 predicted relevant documents, instead of 161 docs classified as probable relevant in the prior round. I saved those in my search folder that I named “51% probable relevant after 3rd round – 132 documents” with date. Twenty-six (26) of the documents were new, but most of them were dupes and near dupes, so there were actually less than 10 brand new docs. Also, interestingly, I only disagreed with one of the predictions, whereas in the prior round I had disagreed with several.

Next I did the same kind of *iCategory probability* search, 51% plus, but for Highly Relevant only. Recall that last time *Inview* returned 12, this time it again returned 12. There were all the same documents. (Note, this was not really necessary because Highly Relevant documents are included in the Relevant category, but I wanted to make sure on the count.)

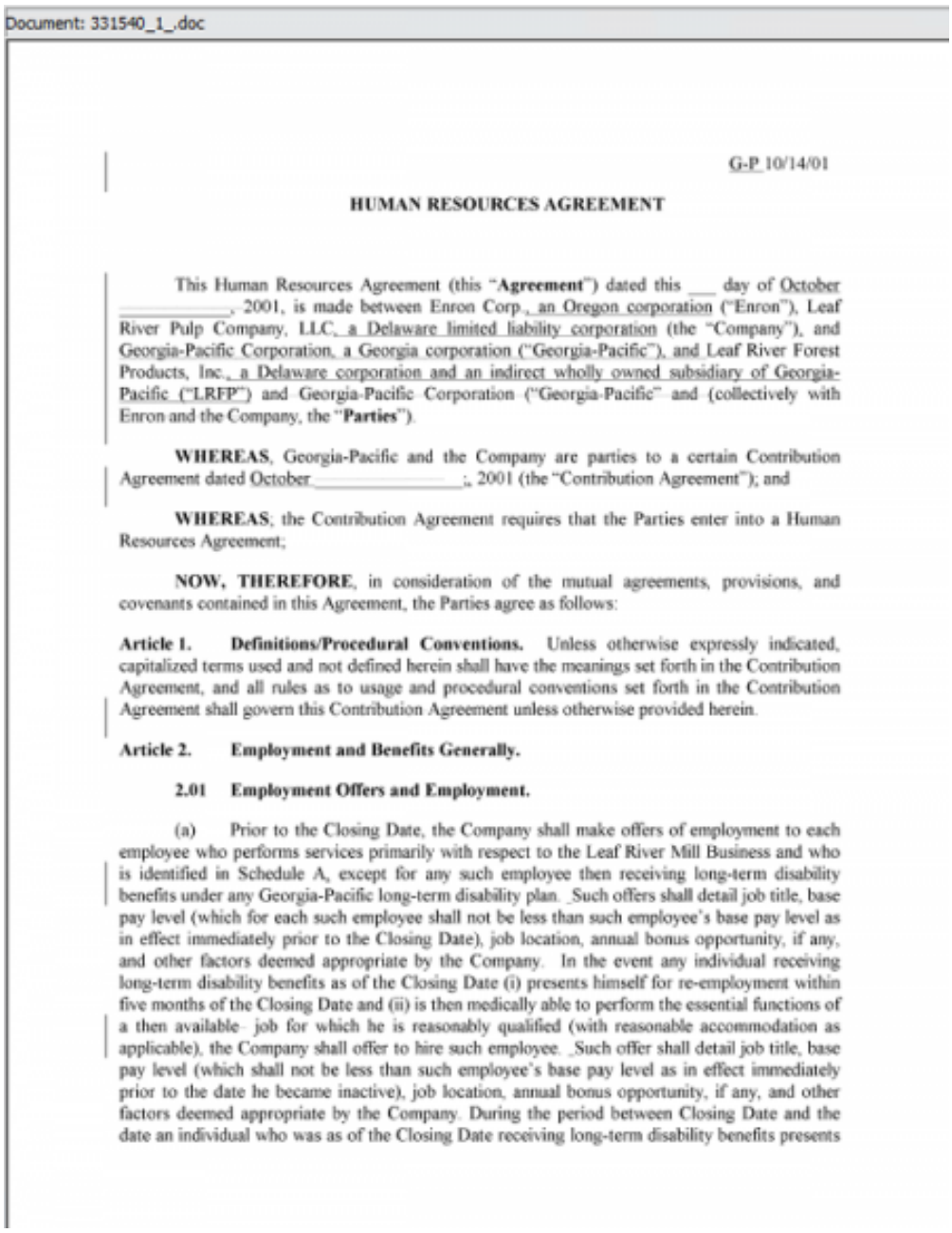
My initial analysis, more like speculation, is that I am either: (1) stuck in *too narrow* an instruction, and the training needs to be expanded so that our recall is better; or, (2) almost done.

With that in mind I did the first search of the mid-range, the grey area, and searched the 40%-50% probability. (I could have done this another way, and let the machine select the mid-range, but this method allowed greater control.) This returned 29 documents.



I had reviewed 7 of these documents before and marked 5 of them as relevant, 1 as undetermined, and 1 as irrelevant. All 5 of the documents I had marked as relevant I had decided *not* to Train on. I had thought all 5 were irregular for some reason and it would not be good to use them for training. So now I marked all 5 of them for training because the computer was not certain about them. That is why they were in the 40-50% probability range.

It is interesting to note that I had marked one of these seven documents before as Undetermined. This is a complex legal document, an eight-page contract entitled *Human Resources Agreement* by and between Enron and several entities, including Georgia Pacific, dated October 2011. It is assigned control number 12009960. The first page of this document is shown below.



Document: 331540_1_.doc

G-P 10/14/01

HUMAN RESOURCES AGREEMENT

This Human Resources Agreement (this "**Agreement**") dated this ___ day of October _____, 2001, is made between Enron Corp., an Oregon corporation ("Enron"), Leaf River Pulp Company, LLC, a Delaware limited liability corporation (the "Company"), and Georgia-Pacific Corporation, a Georgia corporation ("Georgia-Pacific"), and Leaf River Forest Products, Inc., a Delaware corporation and an indirect wholly owned subsidiary of Georgia-Pacific ("LRFP") and Georgia-Pacific Corporation ("Georgia-Pacific" and (collectively with Enron and the Company, the "**Parties**").

WHEREAS, Georgia-Pacific and the Company are parties to a certain Contribution Agreement dated October_____; 2001 (the "Contribution Agreement"); and

WHEREAS, the Contribution Agreement requires that the Parties enter into a Human Resources Agreement;

NOW, THEREFORE, in consideration of the mutual agreements, provisions, and covenants contained in this Agreement, the Parties agree as follows:

Article 1. Definitions/Procedural Conventions. Unless otherwise expressly indicated, capitalized terms used and not defined herein shall have the meanings set forth in the Contribution Agreement, and all rules as to usage and procedural conventions set forth in the Contribution Agreement shall govern this Contribution Agreement unless otherwise provided herein.

Article 2. Employment and Benefits Generally.

2.01 Employment Offers and Employment.

(a) Prior to the Closing Date, the Company shall make offers of employment to each employee who performs services primarily with respect to the Leaf River Mill Business and who is identified in Schedule A, except for any such employee then receiving long-term disability benefits under any Georgia-Pacific long-term disability plan. Such offers shall detail job title, base pay level (which for each such employee shall not be less than such employee's base pay level as in effect immediately prior to the Closing Date), job location, annual bonus opportunity, if any, and other factors deemed appropriate by the Company. In the event any individual receiving long-term disability benefits as of the Closing Date (i) presents himself for re-employment within five months of the Closing Date and (ii) is then medically able to perform the essential functions of a then available job for which he is reasonably qualified (with reasonable accommodation as applicable), the Company shall offer to hire such employee. Such offer shall detail job title, base pay level (which shall not be less than such employee's base pay level as in effect immediately prior to the date he became inactive), job location, annual bonus opportunity, if any, and other factors deemed appropriate by the Company. During the period between Closing Date and the date an individual who was as of the Closing Date receiving long-term disability benefits presents

I was not sure when I first started the project whether it was relevant or not, so I sort of passed, and marked it as Undetermined. I like to do that at the beginning of projects. This document had many provisions on employment termination, but I was not sure if it really pertained to involuntary terminations or not, plus it looked like this was just a draft document, not a final contract. The computer was also unsure, like I used to be. But now with my greater experience of the *relevancy border* I was defining, and especially because of my now greater familiarity of the types of legal documents that Enron was using, I was able to make a decision. I considered the document to be irrelevant. The secondary references to involuntary termination were trumped by the primary intent to deal

with voluntary departures as part of a merger. Plus, this was just a draft legal document. I was unsure of that before, but not now, not after I had seen dozens of documents by Enron lawyers. For all of these reasons, and more, I marked it as irrelevant and marked it for training.

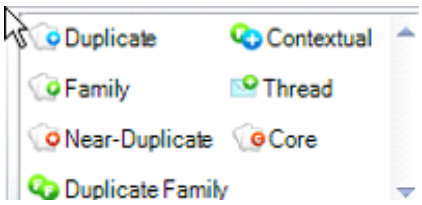
I also marked for training the one document that I had marked before as irrelevant, but had not marked for training. I was hoping this would clear up some obvious confusion in my prior training.

I next reviewed the 22 docs in the grey area that I had not reviewed before. Most of them were dupes and near dupes. There were really only 10 new documents not reviewed before. I disagreed with about half of the predictions and only marked 8 out of the 22 as relevant (but they were somewhat close questions). This is all to be expected for grey area documents. This kind of close-call review is rather slow, and took almost three more hours for the post training tasks, for a total billable time today of four hours.

Sixth Day of Review (4 Hours)

I started with a search to confirm the total number of docs we have now marked as relevant and put them in a folder labeled “All docs Marked Relevant after third run – 137 docs” with date. That was just for housekeeping metrics. Careful labeling of the search folders that Inview generates automatically of each search is very important. It takes a little time to do, but can save you a lot of time later.

Add Associated Searches



Next I tried to expand on the 137 docs by using the *Add Associated* series of commands in the *Home* tab. This is a kind of similarity search function described before.

I started with “*Duplicate*.” This did not add any new files. My prior duplicate exercise had already caught them.

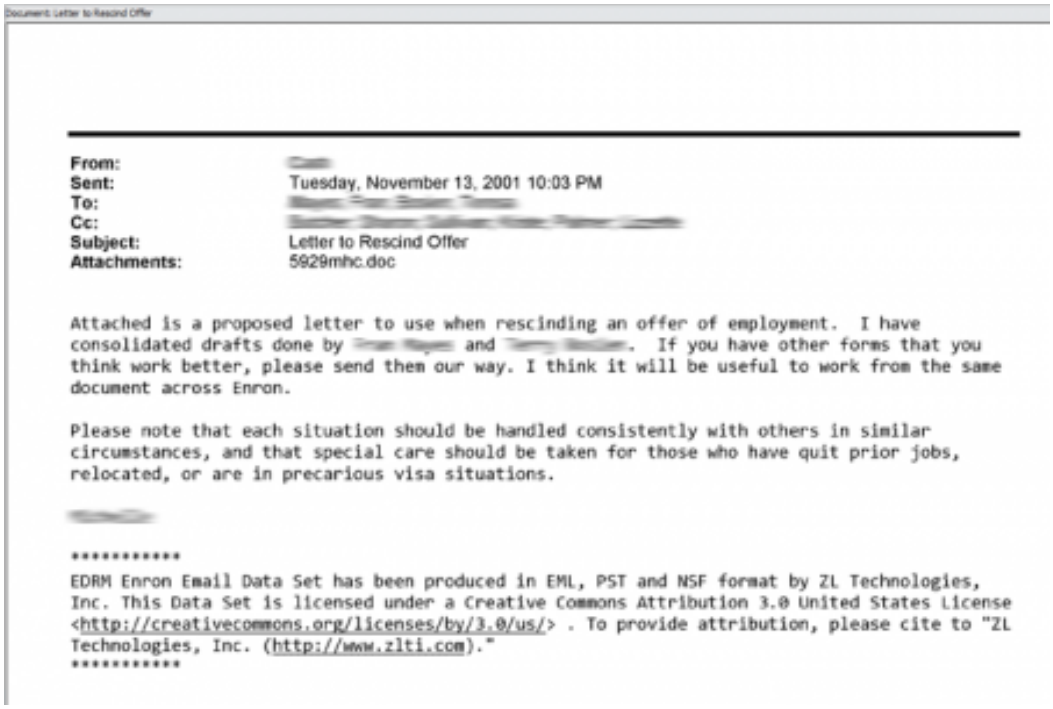
Then I used “*Family*” this added one new email, which transmitted a relevant Q&A document as an attachment. According to our protocol both the attachment and email would be produced, so I marked the email relevant (although nothing on the face of the email alone would be relevant). That was document control number 3600805. But I did not mark the email itself for Training, as I assumed that would not be helpful. We are now up to 138 documents categorized as relevant.

Next I pressed the “*Near-Duplicate*” button and this added no new documents. So then I activated the “*Contextual*” duplicates command. Again, nothing new. I also tried the *Core* near duplicates, again nothing added.

Thread Search

Then I activated the "*Thread*" command and this time it expanded the folder to 306 documents, an increase of 168 documents (it was 138). So this *add associated Thread* function more than doubled the size of the folder. But I had to review all of the 168 new documents from *Thread* to see whether in fact I considered them to be relevant or not. I thought that they probably all would be, or at least might be, because they were part of an email chain that was relevant, but maybe not, as least on their own. This proved to be a very time-consuming task, which I here describe in some detail. In fact, I found 162 out of the 168 to be relevant for various reasons described below and only disallowed 6 thread documents.

I found that many of the new 168 documents were emails that were transmittals of attachments that I had marked relevant, so again in accord with my protocol to always produce email parents of relevant attachments, I marked them all as Relevant, but did not tell to train. If the email has some content that was in itself relevant, than I also marked the document to Train, *see eg* – control number 12010704 shown below.



Document: Letter to Rescind Offer

From:
Sent: Tuesday, November 13, 2001 10:03 PM
To:
Cc:
Subject: Letter to Rescind Offer
Attachments: 5929mhc.doc

Attached is a proposed letter to use when rescinding an offer of employment. I have consolidated drafts done by and . If you have other forms that you think work better, please send them our way. I think it will be useful to work from the same document across Enron.

Please note that each situation should be handled consistently with others in similar circumstances, and that special care should be taken for those who have quit prior jobs, relocated, or are in precarious visa situations.

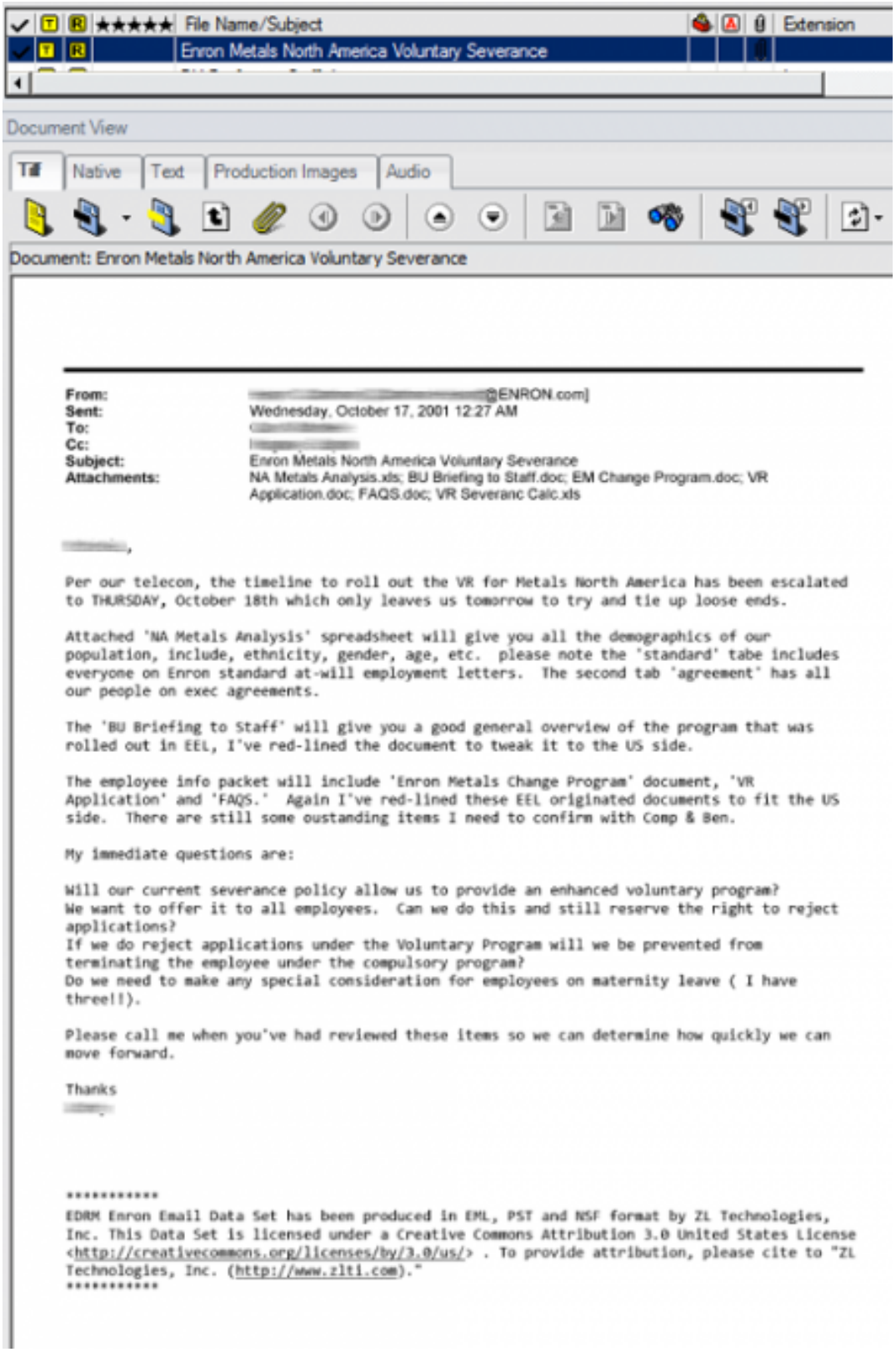
****
EDRM Enron Email Data Set has been produced in EML, PST and NSF format by ZL Technologies, Inc. This Data Set is licensed under a Creative Commons Attribution 3.0 United States License <http://creativecommons.org/licenses/by/3.0/us/> . To provide attribution, please cite to "ZL Technologies, Inc. (http://www.zlti.com)."
****

I also found siblings that were not relevant, and so marked them as Irrelevant in accord with my standard protocol for this project. My standard existing protocol was only to produce relevant attachments (but I am having second thoughts on this, as I will explain below). If one email has two attachments, one relevant, and one irrelevant, under this protocol I would only produce the relevant attachments and the email (parent). I would not produce the irrelevant attachment (sibling).

That is what email families are all about. I would love to hear readers thoughts about that?

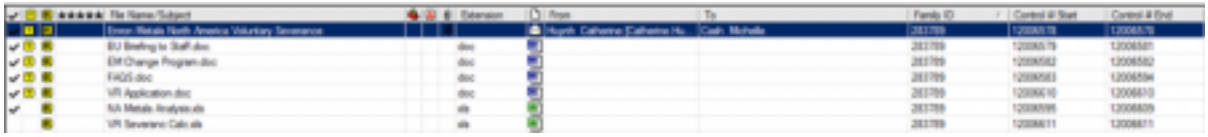
I found the most efficient way to search these new Thread documents was to sort using the *Family ID* column, which I dragged to the left side for easy viewing. To sort you just click on the column in the *Document View*.

For a good example of the kind of parent-child emails I am speaking about, look at the parent email named *Enron Metals North America Voluntary Severance*, control number 12006578.



This email is the parent of *Family ID # 283789*. There are 7 docs in this big family. Three out of the six children had already been marked as Relevant, but

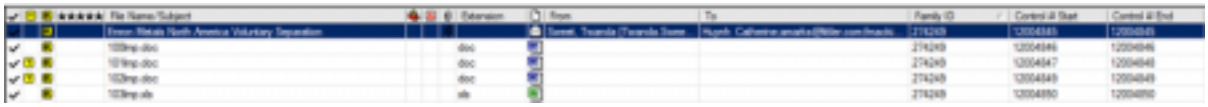
the parent email had not been reviewed, and neither had the three siblings, the other attachments.



Per protocol I marked the parent email as relevant. But in fact, when I read it carefully, I saw that it was relevant on its own. I noted some language in the body of the email itself talking about termination of employees (… w*ill we be prevented from terminating the employee under the compulsory program?*"), so I also marked the parent to Train.

One new sibling, a Word attachment, was reviewed and found to be relevant, so I marked it as Relevant and to Train. I had to give some thought to the two spreadsheets attached to this family, as they were lists of employees. But taking the email and other attachments into consideration, I decided these were likely employees identified for this "voluntary severance" program, which could in these circumstances amount to involuntary termination. The spreadsheet included ethnicity and age, and it is interesting to note that almost all of them were 50 years of age or older. Hmmm. I marked the two spreadsheets as relevant, but did not mark them for Training. This kind of analysis was fairly time-consuming.

For another close family question that I spent time analyzing, see *Family ID # 274249*.

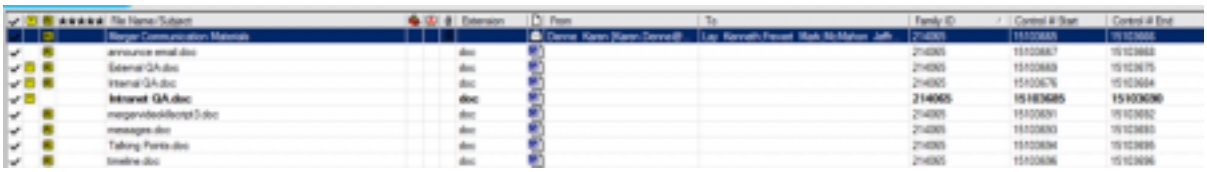


I had previously reviewed and marked as relevant a word document called *101Imp.doc*, control number 12004847. Although much of the document talks about "voluntary separation," some of it talks about termination if an employee does not elect to voluntarily quit. Thus again it looked relevant to me. (Remember I decided that bona fide resignations were not relevant, but forced terminations were.) Do you see a hint in the screen shot of the parent email that suggest this email and attachments may also have been privileged?

This family has two other word docs and an excel spreadsheet. The other word docs were just limited to voluntary separations and so I marked them as Irrelevant, but not for training, as they were a close call. The spreadsheet calculated a "separation payment" if you elected to quit, so I considered that irrelevant too, but again did not Train on it.

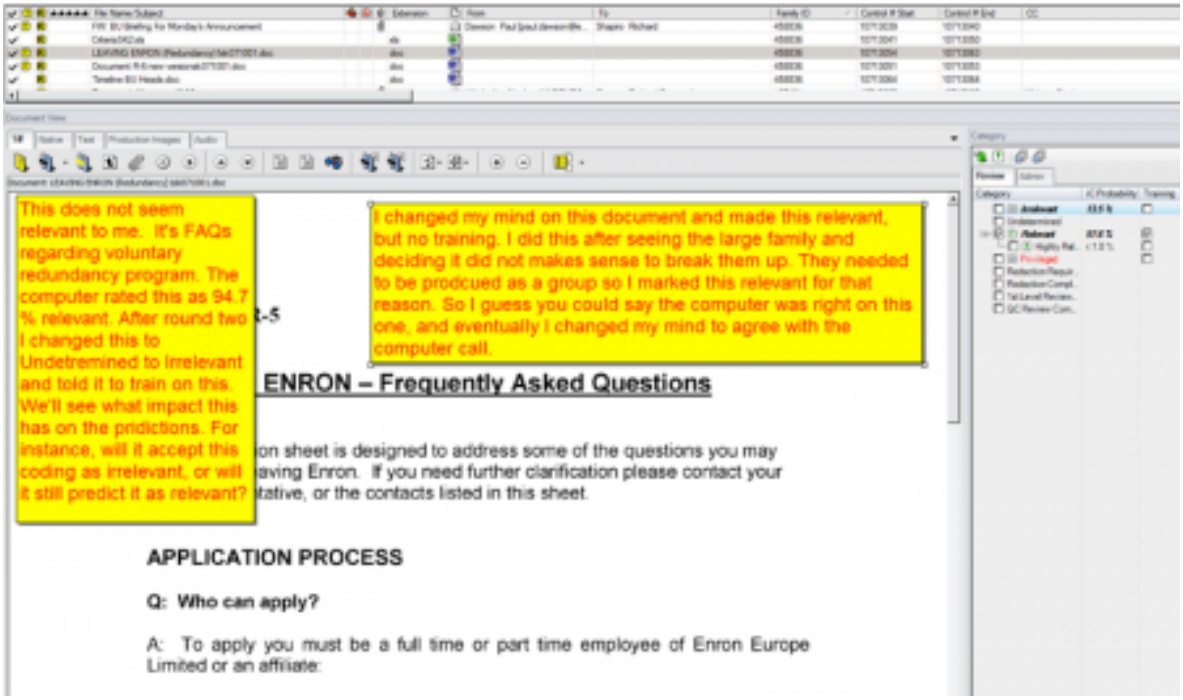
Sometimes it does not make sense to separate the children because they are all so close and interconnected that you could not fully understand the relevant

attachment without also considering the other attachments, which, on their own, might not be considered relevant. The *Family # 214065* is an example of this.



It consists of an email and eight attachments. It concerns Ken Lay's announcement of the purchase of what is left of Enron by Dynegy. Two of the attachments talked about employee terminations, but the others talked about other aspects of the deal. I thought you needed to see them all to understand the ones mentioning layoffs, so I marked them all as relevant. I did the same for *Family # 564604* concerning the same event. I did the same for Family 648122 concerning the Dynegy merger.

I also did the same thing for *Family # 458836*. This last family caused me to change one document that I had previously called Irrelevant and Trained on, and made a sticky note about. I changed the coding to Relevant, but said no for Training. See doc control number 10713054.



As you can see, it is a FAQ document about leaving Enron in the context of voluntary departure, but it was part of a larger package concerning the massive 50% layoffs in October 2001. I left a new sticky note on the document explaining my flip-flop. I started off not knowing if it was relevant or not and so marked it Undetermined (essentially put off for later determination). In round two the

computer rated this FAQ document as 94.7% likely relevant. I was convinced *at that time* that the computer was wrong, that the document was irrelevant because it only pertained to voluntary terminations.

Now, in this third round, I changed my mind again and agreed with the computer, and thought that this FAQ document was relevant. But I thought it was relevant for a new reason, one that I had not even considered before, namely its email family context.

Hybrid Multimodal Mind-Meld Search

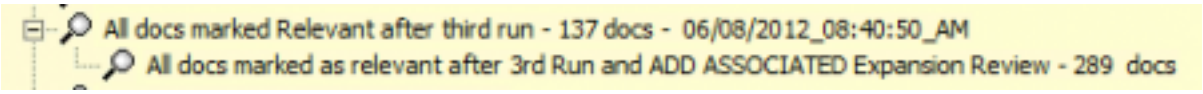
The computer has decided that this FAQ document was 57.6% probable relevant. It did so, instead of a higher relevancy prediction, as you might expect, since I had marked it as relevant and told it to train. Although I did not like my own flip-flopping and agreeing with the computer, I was gratified by this low percentage. It was just 57.6% probable relevant. That indicated, as I thought it should, that Inview still considered the document something of a close call. So did I. To me this was yet another piece of evidence that the procedures were working, that the AI and human minds were melding. It was a hybrid computer-human process, yet I was still in control. That is what I mean by *hybrid* in my catch phrase *hybrid multimodal* search.

How Big Should Your Families Be?

This Family Analysis is a slow process and took me over three hours to complete. It might actually save substantial time to have a more expansive family protocol, one where all attachments are auto marked as relevant if the email is relevant or any of the attachments. But then you end up disclosing more, and possibly triggering more redaction work too. I wanted input on this, especially since exports from *AdvanceView* always include all families, like it or not. So I asked KO's experts on this and they indicated that most people produce entire families without dropping any members, but there is some variation in this practice. Again, I would welcome reader comments on this full family production issue.

Concern Regarding Scope of Relevance

At the end of this exercise to *Add Associated* documents based on the 137 previously categorized as Relevant, I had 289 Relevant documents, an increase of 162 documents (118%). See the screen shot below of the search folders where I stored these documents.



So this proved to be an effective way to increase my relevance count, my recall, and to do so with very good (96%) precision (162 out of the 168 added as *Thread* members were marked by me as relevant). But it was not that helpful in Training, I didn't think, because very few of the 162 newly classified relevant documents were worthy of training status. Most were just technically relevant, for example, because they were an email parent transmitting a relevant child.

For that reason, I still wanted to make at least one more effort to *reach for outliers*, relevant documents on employee termination that had not yet discovered. I was concerned that there might be relevant documents in the collection of a completely different type that I had not found before. I was concerned that my training might have been too narrow. Either that, or perhaps I was near the end. The only way to know for sure was to make special efforts to broaden the search. I decided to broaden the scope of training documents before I ran another Training Session.

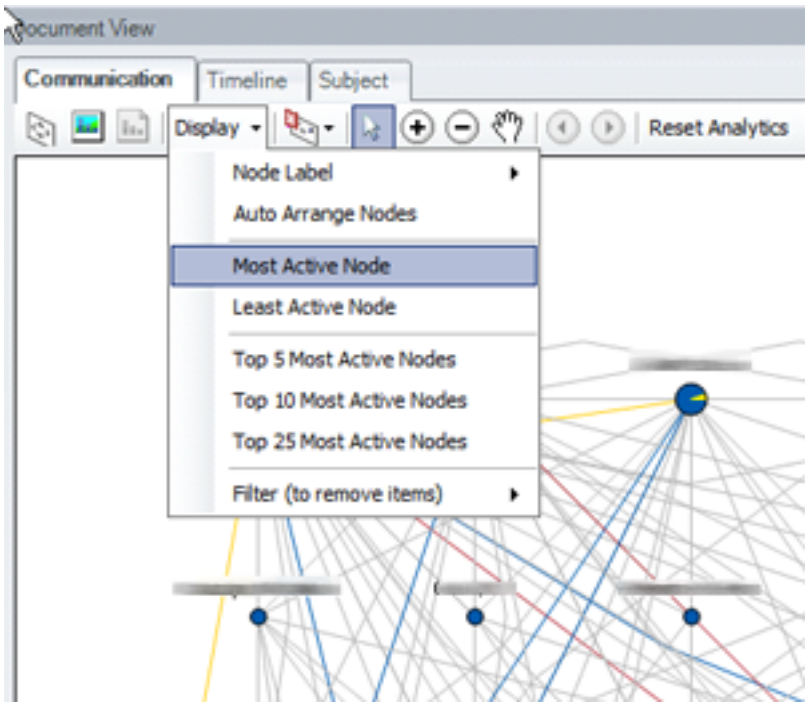
Input from KO's Joe White

To double-check my analysis and plan, I consulted with the KO IRT search expert helping me, Joseph White. He basically agreed with my analysis of the results to date and provided several good suggestions. He agreed that we were at a *tipping point between continuing to search for examples vs. considering the system trained.* He also agreed with my decision to keep going and make more efforts to broaden the search for new training documents.

Joe advised me to run a new learning suggestion at this time to be sure that my *suggestion status* was current. Then he suggested I run another Focus document training session as this would, in his words, help the Inview classifiers. He described the process in shorthand as "*Run Learning Session, optionally enable new suggestions, pull new Focus documents and train them, repeat.*" That is the essence of the predictive coding part of multimodal search. Joe explained that these training sessions can happen many times across all categories, or just the ones you are most concerned about and want further clarity for the system. Joe observed that in my search project to date there were relatively few documents in gray areas, as opposed to other projects he had seen. That meant my project might not need much more iterative training.

In Joe's experience in situations like this one, where *fewer than 2-3% of the corpus is presenting as relevant*, it is generally more difficult to determine how well the system is doing. He suggested that when facing such low prevalence rates, which I know from experience is typical in employment cases, that I should continue to use other search techniques, as I had already been doing, to try to locate additional relevant documents *to feed into training.* In other words, he was recommending the multimodal approach.

For instance Joe suggested use of other Inview search features, including: Associated Documents; Topic Grouping; Concept Searching *to look for terms that may help you find other terms/documents that will yield relevant content*; Find Similar; and Keyword Searching using special Inview capacities such as the Data Dictionary function to view term variants/counts. He also suggested that I continue to engage in general *analysis of date ranges, custodians/people, and metadata patterns related to documents you have found (to help expand on the story)*. Joe suggested I use the Analytics view to help do this. This graphics display of data and data relationships allows for *visual navigation* and *selection of communication patterns, date ranges and subject lines*. Below is a screen shot of one example of the many Analytic views possible.



Most good software today has similar visual representations, including the one shown above of email communication patterns between custodians.

As to the pure predictive coding search methods, which Joe refers to as *Active Learning*, he suggested I continue to use the Focus document system he

described before. He noted that if the gray area count diminishes, and few new relevant documents turn up, then I will know that, in his words, I'm *in a good place*.

Joe applauded my efforts in nuanced selection of documents for training to date. He suggested that I continue to look for new types of relevant documents to include in the IRT training. (The KO people rarely say*predictive coding*, which is a habit I'm trying to break them of. (The term *predictive coding* is descriptive, has been around a long time, and cannot be trademarked.))

Joe said I was correct to not *focus inwardly* on the already-trained documents. But he pointed out that such an inward focus might be appropriate in other projects where there is a concern with prior training quality, such as where there is a change mid-course in relevancy or where mistakes were made in initial coding. Since this had not happened here, he said I was on the right track to focus my search instead on outliers, new types of relevant documents, using the *multimodal* approach, which, by the way is my words, not Joe's. Like most vendor experts, he tends to use proper corporate speak, and is slow to be indoctrinated into my vocabulary. Still, progress is being made on language, and Joe is never hesitant to respond to my questions. Joe's near 24/7 access is also a treat.

My total time estimate for this sixth day of four hours did not include my time to study Joe White's input or write up my work.

Days Seven and Eight of a Predictive Coding Narrative: Where I have another hybrid mind-meld and discover that the computer does not know God

In this fifth installment I continue my description, this time covering days seven and eight of the project. As the title indicates, progress continues and I have another hybrid mind-meld moment. I also discover that the computer does not recognize the significance of references to God in an email. This makes sense logically, but is unexpected and kind of funny when encountered in a document review.

Seventh Day of Review (7 Hours)

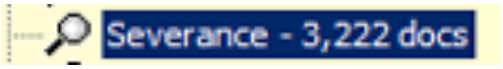
This seventh day I followed Joe White's advice as described at the end of the last narrative. It was essentially a three-step process:

One: I ran another learning session for the dozen or so I'd marked since the last one to be sure I was caught up, and then made sure all of the prior Training documents were checked back in. This only took a few minutes.

Two: I ran two more focus document trainings of 100 docs each, total 200. The focus documents are generated automatically by the computer. It only took about an hour to review these 200 documents because most were obviously irrelevant *to me*, even if the computer was somewhat confused.

I received more of an explanation from Joe White on the *focus documents*, as Inview calls them. He explains that, at the current time at least (KO is coming out with a new version of the Inview software soon, and they are in a state of constant analysis and improvement), 90% of each focus group consists of grey area type documents, and 10% are pure random under IRT ranking. For documents drawn via workflow (in the demo database they are drawn from the System Trainers group in the Document Training Stage) they are selected as 90% focus and 10% random; where the 90% focus selection is drawn evenly across each category set for iC training.

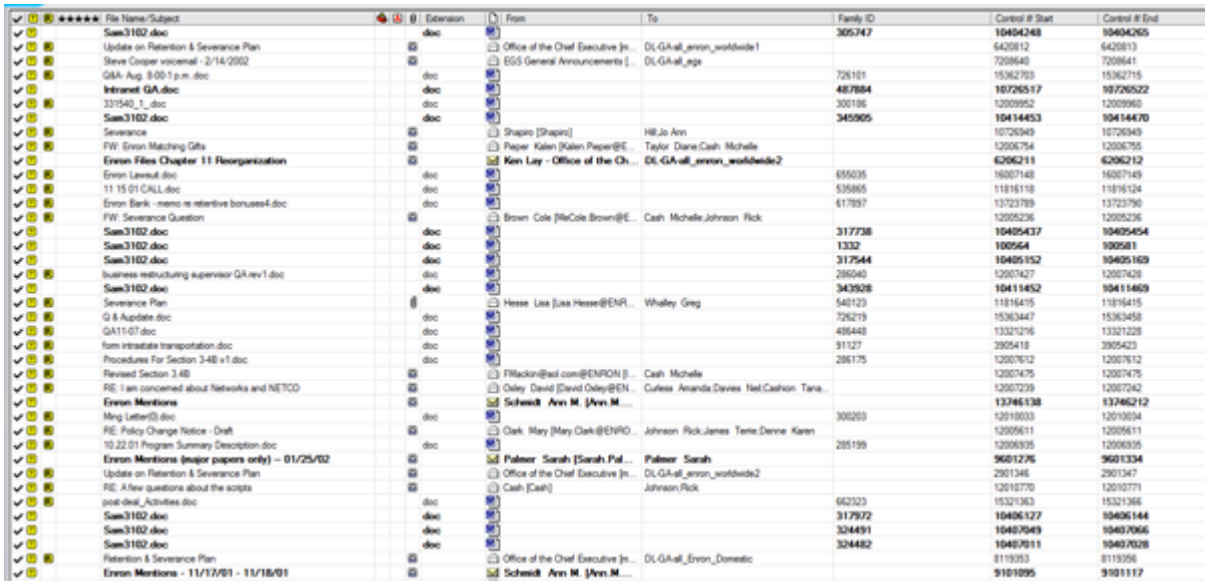
The focus documents come from the areas of least certainty for the algorithm. A similar effect can be achieved by searching for a given iC category for documents between 49 – 51%, etc., as I had done before for relevance. But the automated focus document system makes it a little easier because it knows when you do not have enough documents in the 49 – 51% probability range and then increases the draw to reach your specified number, here 100, to the next least-certain documents. This reduces the manual work in finding the grey area documents for review and training.



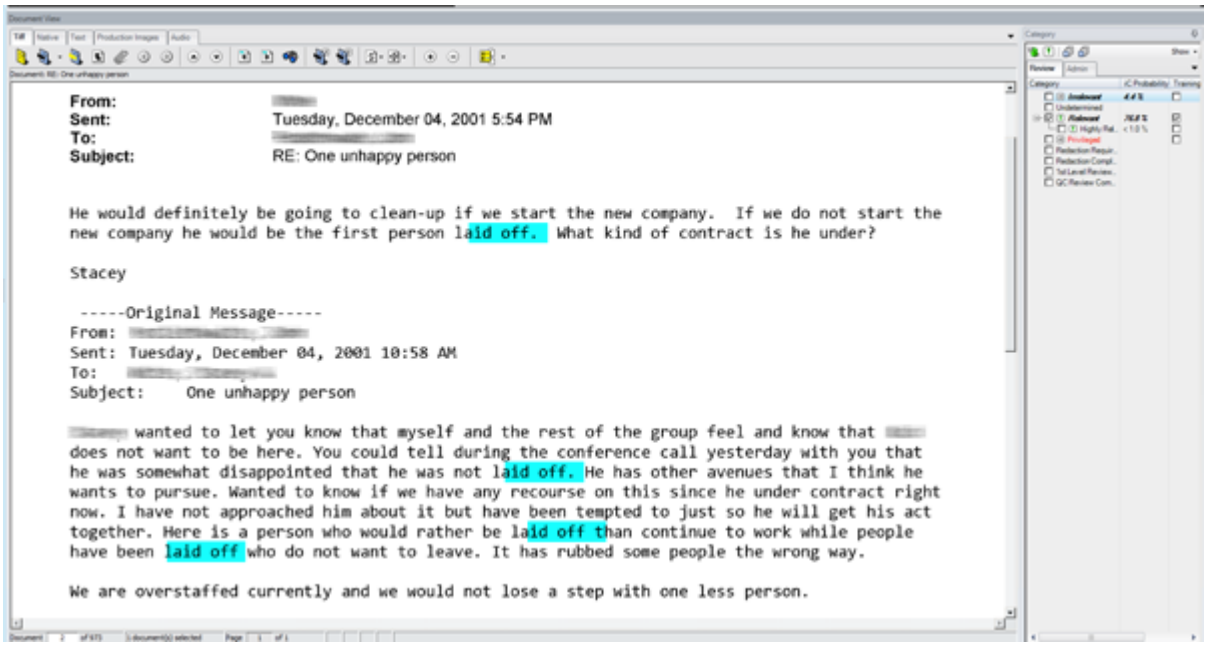
Three: I looked for more documents to evaluate/train the system. I had noticed that "severance" was a key word in relevant documents, and so went back and ran a search for this term for the first time. There were 3,222 hits, so, as per my standard procedure, I added this document count to name of the folder that automatically saved the search.

I found many more relevant documents that way. Some were of a new type I had not seen before (having to do with the mass lay-offs when Enron was going under), so I knew I was expanding the scope of relevancy training, as was my intent. I did the judgmental review by using various sort-type judgment searches in that folder, i.e. by ordering the documents by subject line, file type, search terms hits (the star symbols), etc., and did not review all 3,222 docs. I did not find

that necessary. Instead, I honed in on the relevant docs, but also marked some irrelevant ones here that were close. Below is a screen shot of the first page of the documents sorted by putting those selected for training at the top.



I had also noticed that “lay off” “lay offs” and “laid off” were common terms found in relevant docs, and I had not searched for those particular terms before either. There were 973 documents with hits with one of these search terms. I did the same kind of judgmental search of the folder I created with these documents and found more relevant documents to train. Again, I was finding new documents and knew that I was expanding the scope of relevancy. Below is one new relevant document found in this selection; note how the search terms are highlighted for easy location.



From: [redacted]
Sent: Tuesday, December 04, 2001 5:54 PM
To: [redacted]
Subject: RE: One unhappy person

He would definitely be going to clean-up if we start the new company. If we do not start the new company he would be the first person laid off. What kind of contract is he under?

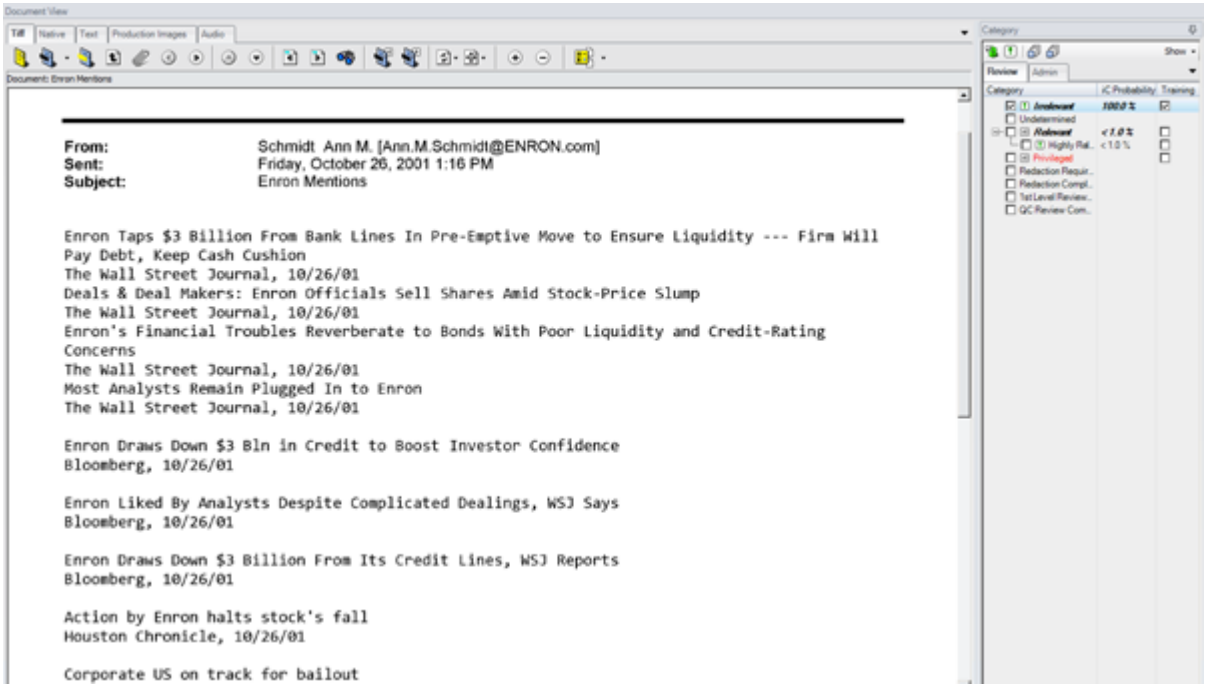
Stacey

-----Original Message-----
From: [redacted]
Sent: Tuesday, December 04, 2001 10:58 AM
To: [redacted]
Subject: One unhappy person

[redacted] wanted to let you know that myself and the rest of the group feel and know that [redacted] does not want to be here. You could tell during the conference call yesterday with you that he was somewhat disappointed that he was not laid off. He has other avenues that I think he wants to pursue. Wanted to know if we have any recourse on this since he under contract right now. I have not approached him about it but have been tempted to just so he will get his act together. Here is a person who would rather be laid off than continue to work while people have been laid off who do not want to leave. It has rubbed some people the wrong way.

We are overstaffed currently and we would not lose a step with one less person.

I also took the time to mark some irrelevant documents in these new search folders, especially the documents in the last folder, and told them to train too, since they were otherwise close from a similar keywords perspective. So I thought I should go ahead and train them to try to teach the fine distinctions.



From: Schmidt Ann M. [Ann.M.Schmidt@ENRON.com]
Sent: Friday, October 26, 2001 1:16 PM
Subject: Enron Mentions

Enron Taps $3 Billion From Bank Lines In Pre-Emptive Move to Ensure Liquidity --- Firm Will Pay Debt, Keep Cash Cushion
The Wall Street Journal, 10/26/01
Deals & Deal Makers: Enron Officials Sell Shares Amid Stock-Price Slump
The Wall Street Journal, 10/26/01
Enron's Financial Troubles Reverberate to Bonds With Poor Liquidity and Credit-Rating Concerns
The Wall Street Journal, 10/26/01
Most Analysts Remain Plugged In to Enron
The Wall Street Journal, 10/26/01

Enron Draws Down $3 Bln in Credit to Boost Investor Confidence
Bloomberg, 10/26/01

Enron Liked By Analysts Despite Complicated Dealings, WSJ Says
Bloomberg, 10/26/01

Enron Draws Down $3 Billion From Its Credit Lines, WSJ Reports
Bloomberg, 10/26/01

Action by Enron halts stock's fall
Houston Chronicle, 10/26/01

Corporate US on track for bailout

The above third step took another five hours (six hours total). I knew I had added hundreds of new docs for training in the past five hours, both relevant and irrelevant.

Fourth Round

I decided it was time to run a training session again and force the software to analyze and rank all of the documents again. This was essentially the Fourth Round (not counting the little training I did at the beginning today to make sure I was served with the right (updated) Focus documents).

After the Training session completed, I asked for a report. It showed that 2,663 total documents (19,731 pages) have now been categorized and marked for Training in this last session. There were now 1,156 Trainer (me) identified documents, plus the original 1,507 System ID'ed docs. (Previously, in Round 3, there were the same 1,507 System ID'ed docs, and only 534 Trainer ID'ed docs.)

Then I ran a report to see how many docs had been categorized by me as Relevant (whether also marked for Training or not). Note I could have done this before the training session too, and it would not make any difference in results. All the training session does is change the predictions on coding, not the actual prior human coding. This relevancy search was saved in another search folder called "All Docs Marked Relevant after 4th Round – 355 Docs." After the third round I had only ID'ed 137 relevant documents. So progress in recall was being made.

Prevalence Quality Control Check

As explained in detail in *Day Two of a Predictive Coding Narrative: More Than A Random Stroll Down Memory Lane,* my first random sample search allowed me to determine prevalence and get an idea of the total number of relevant document likely contained in the database. The number was 928 documents. That was the spot or point projection of the total *yield* in the corpus. (*Yield* is another information science and statistics term that is useful to know. It means in this context the expected number of relevant documents in the total database. *See eg.* Webber, W., *Approximate Recall Confidence Intervals,* ACM Transactions on Information Systems, Vol. V, No. N, Article A (2012 draft) at A2.)

My *yield* calculation here of 928 is based on my earlier finding of 2 relevant documents in the initial 1,507 random sample. (2/1507=.00132714) (.13*699,082=928 relevant documents). So based on this I knew that I was correct to have gone ahead with the fourth round, and would next check to see how many documents the IRT now predicted would be relevant. My hope was the number would now be closer to the 928 goal of the projected yield of the 699,082 document corpus.

This last part had taken another hour, so I'll end Day Seven with a total of 7 hours of search and review work.

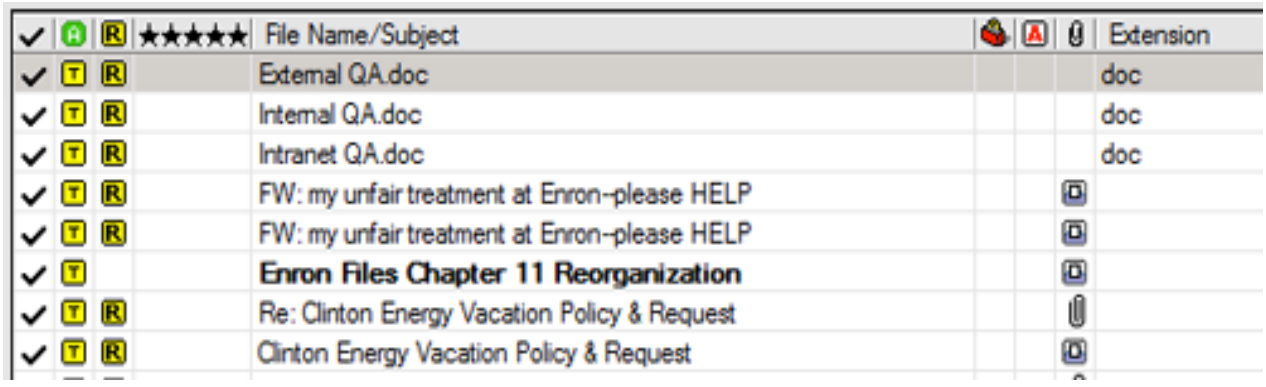
Eighth Day of Review (9 Hours)

First I ran a probability search as before for all 51%+ probable relevant docs and saved them in a folder by that name. After the fourth round the IRC now predicted a total of 423 relevant documents. Remember I had already actually reviewed and categorized 355 docs as relevant, so this was only a potential max net gain of 68 docs. As it turned out, I disagreed with 8 of the predictions, so the actual net gain was only 60 docs, for a total of 415 confirmed relevant documents.

I had hoped for more after broadening the scope of documents marked relevant in the last seeding. So I was a little disappointed that my last seed set had not led to more predicted relevant. Since the "recall goal" for this project was 928 documents, I knew I still had some work to do to expand the scope. Either that or the confidence interval was at work, and there were actually fewer relevant documents in this collection than the random sample predicted as a point projection. The probability statistics showed that the actual range was between 112 documents 3,345 documents, due to the 95% confidence level and +/-3% confidence interval.

51%+ Probable Relevant Documents

Next I looked at the 51%+ probable relevant docs folder and sorted by whether the documents had been categorized on not. You do that by clicking on the symbol for categorization, a check, which is by default located in the upper left. That puts all of the categorized docs together, either on top or bottom. Then I reviewed the 68 new documents, the ones the computer predicted to be relevant that I had not previously marked relevant.



										File Name/Subject				Extension
✓	A	R	★	★	★	★	★							
✓	T	R								External QA.doc				doc
✓	T	R								Internal QA.doc				doc
✓	T	R								Intranet QA.doc				doc
✓	T	R								FW: my unfair treatment at Enron--please HELP				
✓	T	R								FW: my unfair treatment at Enron--please HELP				
✓	T									**Enron Files Chapter 11 Reorganization**				
✓	T	R								Re: Clinton Energy Vacation Policy & Request				
✓	T	R								Clinton Energy Vacation Policy & Request				

This is always the part of the review that is the most informative for me as to whether the computer is actually "getting-it" or not. You look to see what documents it gets wrong, in other words, makes a wrong prediction of probable

relevance, and try to determine why. In this way you can be alert for additional documents to try to correct the error in future seeds. You learn from the computer's mistakes where additional training is required.

I then had some moderately good news in my review. I only disagreed with eight of the 68 new predictions. One of these documents only had a 52.6% probability for relevance, another 53.6%, another 54.5%, another 54%, another 57.9%, and another other only 61%. Another two were 79.2% and 76.7% having to do with "voluntary" severance again, a mistake I had seen before. So even when the computer and I disagreed, it was not by much.

Computer Finds New Hard-to-Detect Relevant Documents

A couple of the documents that Inview predicted to be relevant were long, many pages, so my study and analysis of them took a while. Even though these long documents at first seemed irrelevant to me, as I kept reading and analyzing them, I ultimately agreed with the computer on *all* of them. A careful reading of the documents showed that they did in fact include discussion related to termination and terminated employees. I was surprised to see that, but pleased, as it showed the *software mojo* was kicking in. The predictive coding training was allowing the computer to find documents I would likely never have caught on my own. The mind-meld was working and hybrid power was again manifest.

These hard to detect issues (for me) mainly arose from the unusual situation of the mass terminations that came at the end of Enron, especially at the time of its bankruptcy. To be honest, I had forgotten about those events. My recollection of Enron history was pretty rusty when I started this project. I had not been searching for bankruptcy related terminations before. That was entirely the computer's contribution and it was a good one.

From this study of the 68 new docs I realized that although there were still some issues with the software making an accurate distinction between voluntary and involuntary severance, overall, I felt pretty confident that*Inview* was now pretty well-trained. I based that on the 60 other predictions that were spot on.

Note that I marked most of the newly confirmed relevant documents for training, but not all. I did not want to excessively weight the training with some that were redundant, or odd for one reason or another, and thus not particularly instructive.

This work was fairly time-consuming. It took three long hours on a Sunday to complete.

Fifth Round

Returning to work in the evening I started another training session, the Fifth. This would allow the new teaching (document training instructions) to take effect.

My plan was to then have the computer serve me up the 100 close calls (Focus Documents) by using the document training Checkout feature. Remember this feature selects and serves up for review the grey area docs designed to improve the IRT training, plus random samples.

But before I reviewed the next training set, I did a quick search to see how many *new* relevant documents (51%+) the last training (fifth round) has predicted. I found a total of 545 documents 51%+ predicted relevant. Remember I left the last session with 415 relevant docs (goal is 928). So progress was still being made. The computer had added 130 documents.

Review of Focus Documents

Before I looked at these new ones to see how many I agreed with, I stuck to my plan, and took a Checkout feed of 100 Focus documents. My guess is that most of the newly predicted 51%+ relevant docs would be in the grey area anyway, and so I'll be reviewing some of them when I reviewed the Focus documents.

First, I noticed right away that it served up 35 irrelevant junk files that were obviously irrelevant and previously marked as such, such as PST placeholder files, and a few others like that, which clutter this ENRON dataset. Obviously, they were part of the random selection part of the Focus document selections. I told them all to train in one bulk command, hit the completed review button for them, and then focused on the remaining 65 documents. None had been reviewed before. Next I found some more obviously irrelevant docs, which were not close at all, i.e. 91% irrelevant and only 1% likely relevant. I suspect this is part of the general database random selection that makes up 10% of the Focus documents (the other 90% are close calls).

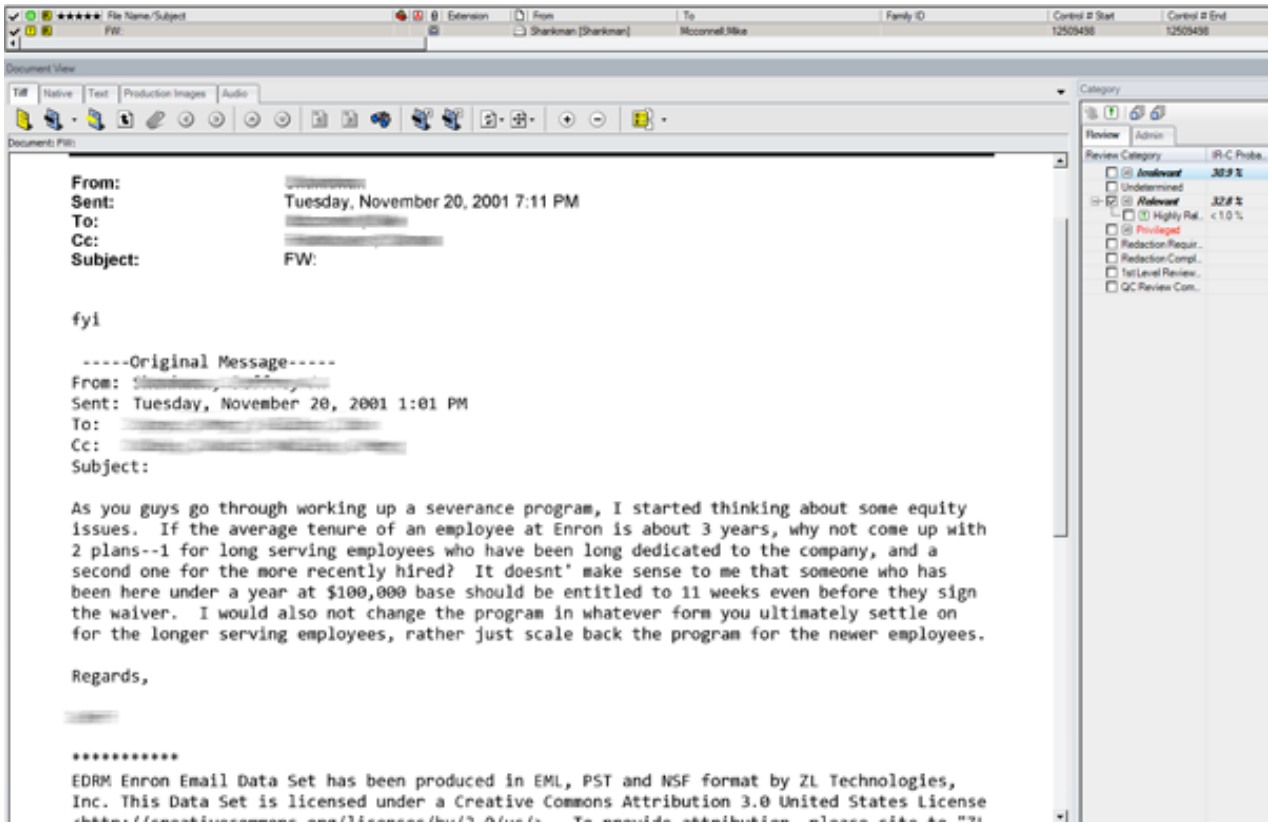
Next I did a file type sort to see if any more of the unreviewed documents in this batch of 100 were obviously irrelevant based on file type. I found 8 more such files, mass categorized them, mass trained them and quickly completed review for these 8.

Now there were 57 docs left, 9 of which were Word docs, and the rest emails. So I checked the 9 word docs next. Six of these were essentially the same document called "11 15 01 CALL.doc." The computer gave each approximately a 32.3% probability of irrelevance and a 33.7% probability of relevance. Very close indeed. Some of the other docs had very slight prediction numbers (less than 1%). The documents proved to be very close calls. Most of them I found to be irrelevant. But in one document I found a comment about mass employee layoffs, so I decided to call it relevant to our issue of employee terminations. I trained those eight and checked them back in. I then reviewed the remaining word docs, found that they were also very close, but marked these as irrelevant and checked them in, leaving 48 docs left to review in the Training set of 100.

Next I noticed a junk kind of mass email from a sender called “Black.” I sorted by “From” found six by Black, and a quick look showed they were all irrelevant, as the computer had predicted for each. Not sure why they were picked as focus docs, but regardless, I trained them and checked them back in, now leaving 42 docs to review.

Next I sorted the remaining by “Subject” to look for some more that I might be able to quickly bulk code (mass categorize). It did not help much as there were only a couple of strings with the same subject. But I kept that subject order and sloughed through the remaining 42 docs.

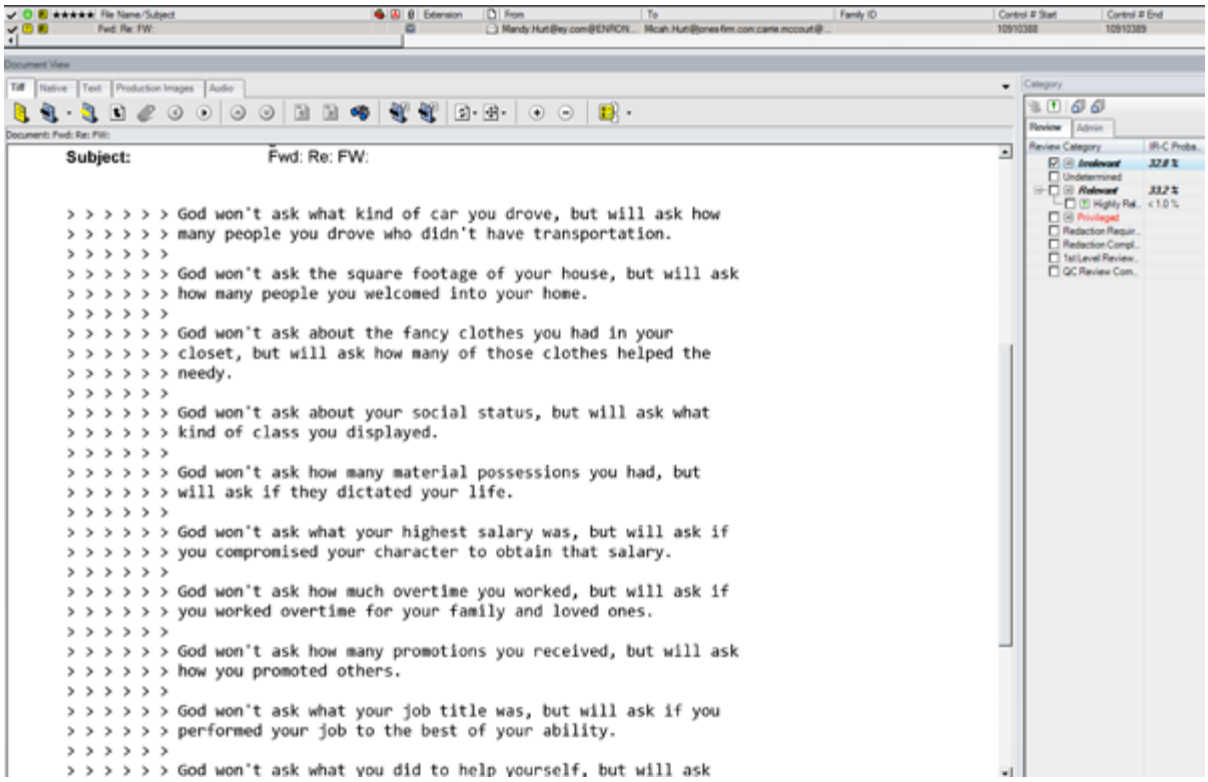
I found most of the remaining docs were very close calls, all in the 30% range for *both* relevant and irrelevant. So they were all uncertain, i.w. a split choice, but none were actually predicted relevant, that is, none were in the over 50% likely relevant range. I found that most of them were indeed irrelevant, but not all. A few in this uncertain range were relevant. They were barely relevant, but of the *new type* recently marked having to do with the bankruptcy. Others that I found relevant were of a type I had seen before, yet the computer was still unsure with basically an even split of prediction in the 30% range. They were apparently different from the obviously relevant documents, but in a subtle way. I was not sure why. *See Eg*: control number 12509498.



It was 32.8% relevant and 30.9% irrelevant, even though I had marked an identical version of this email before as relevant in the last training. The computer was apparently suspicious of my prior call and was making sure. I know I'm anthropomorphizing a machine, but I don't know how else to describe it.

Computer's Focus Was Too Myopic To See God

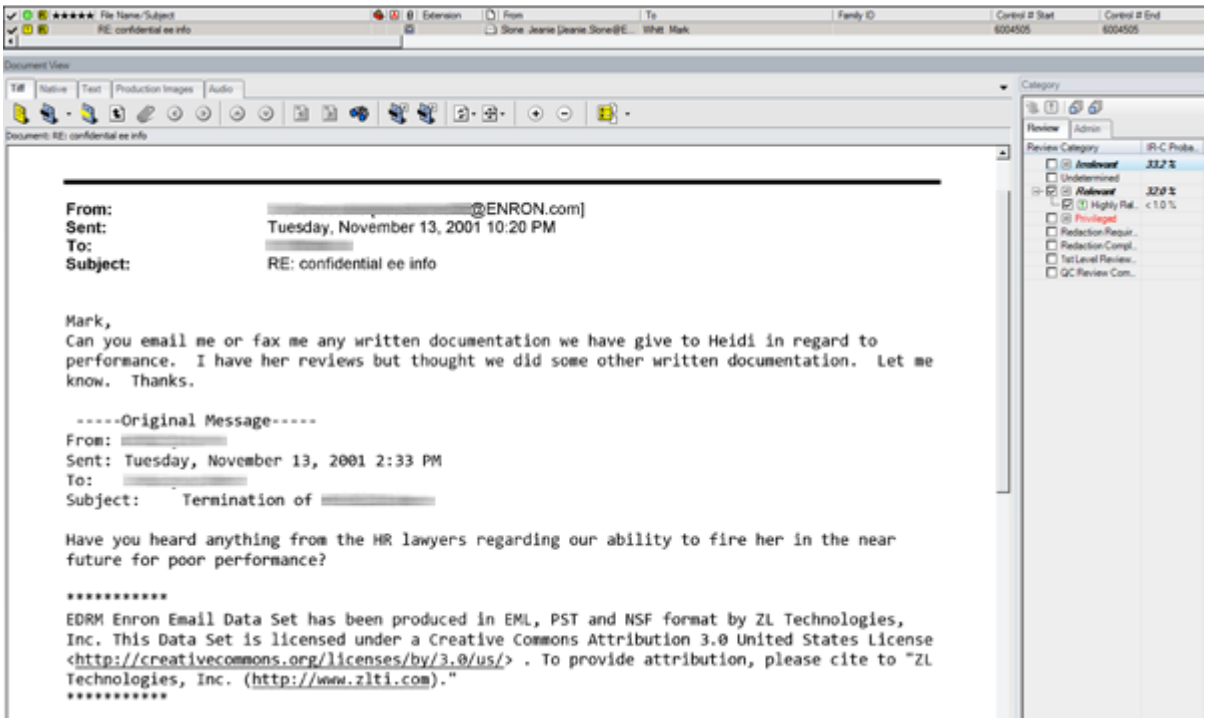
One of the focus documents that the computer found a close call in the 30% range was email with control number 10910388. It was obviously just an inspirational message being forwarded around about God. You know the type I'm sure.



It was kind of funny to see that this email confused the computer, whereas any human could immediately recognize that this was a message about God, not employee terminations. It was obvious that the computer did not know God.

Suddenly My Prayers Are Answered

Right after the funny God mistake email, I reviewed another email with control number 6004505. It was about wanting to fire a particular employee. Although the computer was uncertain about the relevancy of this document, I knew right away that *it rocked*. It was just the kind of evidence I had been looking for. I marked it as Highly Relevant, the first hot document found in several sessions. Here is the email.



From: [redacted]@ENRON.com]
Sent: Tuesday, November 13, 2001 10:20 PM
To: [redacted]
Subject: RE: confidential ee info

Mark,
Can you email me or fax me any written documentation we have give to Heidi in regard to performance. I have her reviews but thought we did some other written documentation. Let me know. Thanks.

-----Original Message-----
From: [redacted]
Sent: Tuesday, November 13, 2001 2:33 PM
To: [redacted]
Subject: Termination of [redacted]

Have you heard anything from the HR lawyers regarding our ability to fire her in the near future for poor performance?

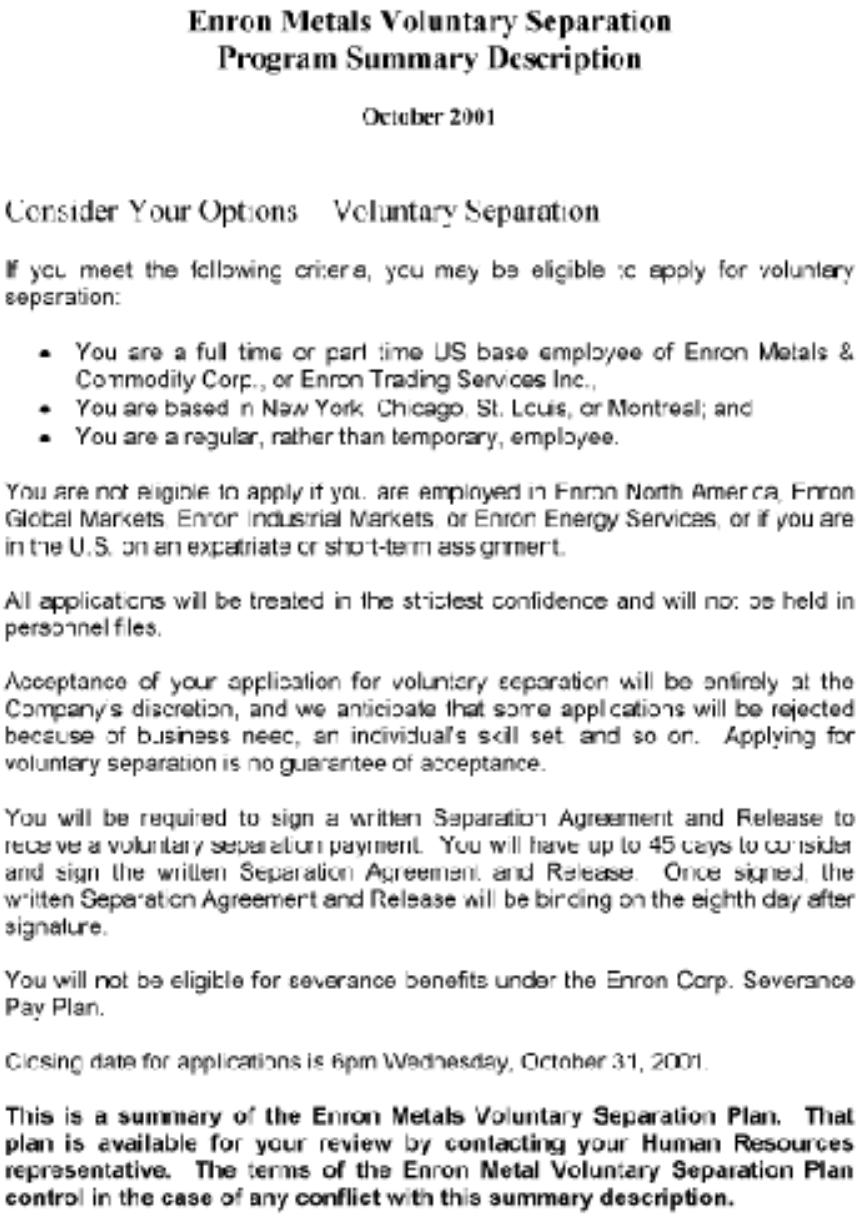
****
EDRM Enron Email Data Set has been produced in EML, PST and NSF format by ZL Technologies, Inc. This Data Set is licensed under a Creative Commons Attribution 3.0 United States License <http://creativecommons.org/licenses/by/3.0/us/> . To provide attribution, please cite to "ZL Technologies, Inc. (http://www.zlti.com)."
****

I took this discovery of a hot doc as a good sign. I was finding both the original documents I had been looking for and the new outliers. It looked to me like I had succeeded in training and in broadening the scope of relevancy to its proper breadth. I might not be travelling a divine road to redemption, but it was clearly leading to better recall.

Since most of these last 42 documents were all close questions (some were part of the 10% random and were obvious), the review took longer than usual. The above tasks all took over 1.5 hours (not including machine search time or time to write this memo).

Good Job Robot!

My next task was to review the 51% predicted relevant set of 545 docs. One document was particularly interesting, control number 12004849, which was predicted to be 54.7% likely relevant. I had previously marked it Irrelevant based on my close call decision that it only pertained to voluntary terminations, not involuntary terminations. It was an ERISA document, a Summary Plan Description of the Enron Metals Voluntary Separation Program.



Enron Metals Voluntary Separation
Program Summary Description

October 2001

Consider Your Options Voluntary Separation

If you meet the following criteria, you may be eligible to apply for voluntary separation:

- You are a full time or part time US base employee of Enron Metals & Commodity Corp., or Enron Trading Services Inc.,
- You are based in New York, Chicago, St. Louis, or Montreal; and
- You are a regular, rather than temporary, employee.

You are not eligible to apply if you are employed in Enron North America, Enron Global Markets, Enron Industrial Markets, or Enron Energy Services, or if you are in the U.S. on an expatriate or short-term assignment.

All applications will be treated in the strictest confidence and will not be held in personnel files.

Acceptance of your application for voluntary separation will be entirely at the Company's discretion, and we anticipate that some applications will be rejected because of business need, an individual's skill set, and so on. Applying for voluntary separation is no guarantee of acceptance.

You will be required to sign a written Separation Agreement and Release to receive a voluntary separation payment. You will have up to 45 days to consider and sign the written Separation Agreement and Release. Once signed, the written Seperation Agreement and Release will be binding on the eighth day after signature.

You will not be eligible for severance benefits under the Enron Corp. Severance Pay Plan.

Closing date for applications is 6pm Wednesday, October 31, 2001.

This is a summary of the Enron Metals Voluntary Separation Plan. That plan is available for your review by contacting your Human Resources representative. The terms of the Enron Metal Voluntary Separation Plan control in the case of any conflict with this summary description.

Since the document on its face obviously pertained to voluntary separations, it was not relevant. That was my original thinking and why I at first called it Irrelevant. But my views on document characterizations on that fuzzy line between voluntary and involuntary employee terminations had changed somewhat over the course of the review project. I now had a better understanding of the underlying facts. The document necessarily defined both eligibility for this benefit, money when an employee left, and ineligibility. It specifically stated that employees of certain Enron entities were *ineligible* for this benefit. It stated that acceptance of an application was strictly within the company's discretion. What happened if even an eligible employee decided not

to *voluntarily quit* and take this money? Would they not then be terminated involuntarily? What happened if they applied for this severance, and the company said no? For all these reasons, and more, I decided that this document was in fact relevant to both voluntary and involuntary terminations. The relevance to involuntary terminations was indirect, and perhaps a bit of a stretch, but in my mind it was in the scope of a relevant document.

Bottom line, I had changed my mind and I now agreed with the computer and considered it Relevant. So I changed the coding to relevant and trained on it. Good call Inview. It had noticed an inconsistency with some of my other document codings and suggested a correction. I agreed. That was impressive. Good robot!

Looking at the New 51%+

Another one of the new documents that was in the 51%+ predicted relevant group was a document with 42 versions of itself. It was the Ken Lay email where he announced that he was not accepting his sixty-million dollar golden parachute. (Can you imagine how many law suits would have ensued if he took that money?) Here is one of the many copies of this email.

From: Ken Lay - Office of the Chairman [mbx_klayofficechair@ENRON.com]
Sent: Wednesday, November 14, 2001 12:18 AM
To: DL-GA-all_enron_worldwide2
Subject: Change of Control Provisions

As many of you know, I have a provision in my employment contract which provides for a payment of $20 million per year for the remaining term of my contract in the event of a change of control of Enron. The merger with Dynegy, or a similar transaction with any other company, would trigger this provision on closing. Assuming the merger with Dynegy is closed within 6-9 months, as we expect, this provision would entitle me to total payments of slightly more than $60 million.

Many CEOs have change of control provisions in their employment contracts and mine has been in place since 1989. But given the current circumstances facing the company and our employees, I have been giving a lot of thought these last few days to what to do about this payment. Initially, I thought I would use part of the funds for a foundation for our employees and take the remainder in stock and cash. However, after talking to a number of employees this afternoon, I have decided that the best course of action would be for me to waive my right to any of this payment. Therefore, at closing, I will receive no payments under this provision.

I know this action does not remedy the uncertainty that you and your families face. But please know that I will continue to do everything in my power to serve the best interests of Enron's employees and shareholders. I am still very proud of what we have built at Enron, and I want to continue working with all of you to correct the problems and restore Enron to its rightful place in the energy industry.

Thank you.

I had previously marked a version of this email as relevant in past rounds. Obviously the *corpus* (the 699,082 Enron emails) had more copies of that

particular email that I had not found before. It was widely circulated. I confirmed the predictions of Relevance. (Remember that this database was deduplicated only on the individual custodian basis, vertical deduplication. It was not globally deduplicated against all custodians, horizontal deduplication. I recommend full horizontal deduplication as a default protocol.)

I disagreed with many of the other predicted relevant docs, but did not consider any of them important. The documents now presenting as possibly relevant were, in my view, cumulative and not really new, not really important. All were fetched by the outer limits of relevance triggered by my previously allowing in as barely relevant the final day comments on Ken Lay's not taking a sixty-million dollar payment, and also allowing in as relevant general talk during bankruptcy that might mention layoffs.

Also, I was allowing in as relevant new documents and emails that concerned the ERISA plan revisions that were related to general severance. The SPD of the Enron Metals Voluntary Separation Program was an example of that. These were all fairly far afield of my original concept of relevance, which had grown as I saw all of the final days emails regarding layoffs, and better understood the bankruptcy and ERISA set up, etc.

Bottom line, I did not see much training value in these newly added docs, both predicted and confirmed. The new documents were not really new. They were very close to documents already found in the prior rounds. I was thinking it might be time to bring this search to an end.

Latest Relevancy Metrics

I ran one final search to determine my total relevant coded documents. The count was 659. That was a good increase over the last measured count of 545 relevant, but still short of my initial goal of 928, the point projection of *yield*. That is a 71% recall (659/928) of my target, which is pretty good, especially if the remaining relevant were just cumulative or otherwise not important. Considering the 3% confidence interval, and the range inherent in the 928 *yield* point projection because of that, from between 112 and 3,345 documents, it could in fact already be 100% recall, although I doubted that based on the process to date. *See* references to point projection, intervals, and William Webber's work on confidence intervals in *Day Two of a Predictive Coding Narrative: More Than A Random Stroll Down Memory Lane* and in Webber, W., *Approximate Recall Confidence Intervals*, ACM Transactions on Information Systems, Vol. V, No. N, Article A (2012 draft).

Enough Is Enough

I was pretty sure that further rounds of search would lead to the discovery of more relevant documents, but thought it very unlikely that any

more *significant* relevant documents would be found. Although I had found one hot doc in this round, the quality of the rest of the documents found convinced me that was unlikely to occur again. I had the same reaction to the grey area documents. The quality had changed. Based on what I had been seeing in the last two rounds, the relevant documents left were, in my opinion, likely cumulative and of no real probative value to the case.

In other words, I did not see value in continuing the search and review process further, except for a final null-set quality control check. I decided to bring the search to end. Enough is enough already. Reasonable efforts are required, not perfection. Besides, I knew there was a final quality control test to be passed, and that it would likely reveal any serious mistakes on my part.

Moving On to the *Perhaps-Final* Quality Control Check

After declaring the search to be over, the next step in the project was to take a random sample of the documents not reviewed or categorized, to see if any significant false-negatives turned up. If none did, then I would consider the project a success, and conclude that more rounds of search were not required. If some did turn up, then I would have to keep the project going for at least another round, maybe more, depending on exactly what false-negatives were found. That would have to wait for the next day.

But before ending this long day I ran a quick search to see the size of this null set. There were 698,423 docs not categorized as relevant and I saved them in a Null Set Folder for easy reference. Now I could exit the program.

Total time for this night's work was 4.5 hours, not including report preparation time and wait time on the computer for the training.

Day Nine of a Predictive Coding Narrative: A scary search for false-negatives, a comparison of my CAR with the Griswold's, and a moral dilemma

In this sixth installment I continue my description, this time covering day nine of the project. Here I do a quality control review of a random sample to evaluate my decision in day eight to close the search.

Ninth Day of Review (4 Hours)

I began by generating a random sample of 1,065 documents from the entire null set (95% +/- 3%) of all documents not reviewed. I was going to review this

sample as a quality control test of the adequacy of my search and review project. I would personally review all of them to see if any were *False Negatives*, in other words, relevant documents, and if relevant, whether any were especially significant or Highly Relevant.

I was looking to see if there were any documents left on the table that should have been produced. Remember that I had already personally reviewed all of the documents that the computer had predicted were like to be relevant (51% probability). I considered the upcoming random sample review of the excluded documents to be a good way to check the accuracy of reliance on the computer's predictions of relevance.

I know it is not the *only way*, and there are other quality control measures that could be followed, but this one makes the most sense to me. Readers are invited to leave comments on the adequacy of this method and other methods that could be employed instead. I have yet to see a good discussion of this issue, so maybe we can have one here.

If my decision in day eight to close the search was correct, then virtually all of the predicted irrelevant files should be irrelevant. For that reason I expected the manual review of the null set to go very fast. I expected to achieve speeds of up to 500 files per hour and to be able to complete the task in a few hours. Well, anyway, that was my hope. I was not, however, going to rush or in any way deviate from my prior review practices.

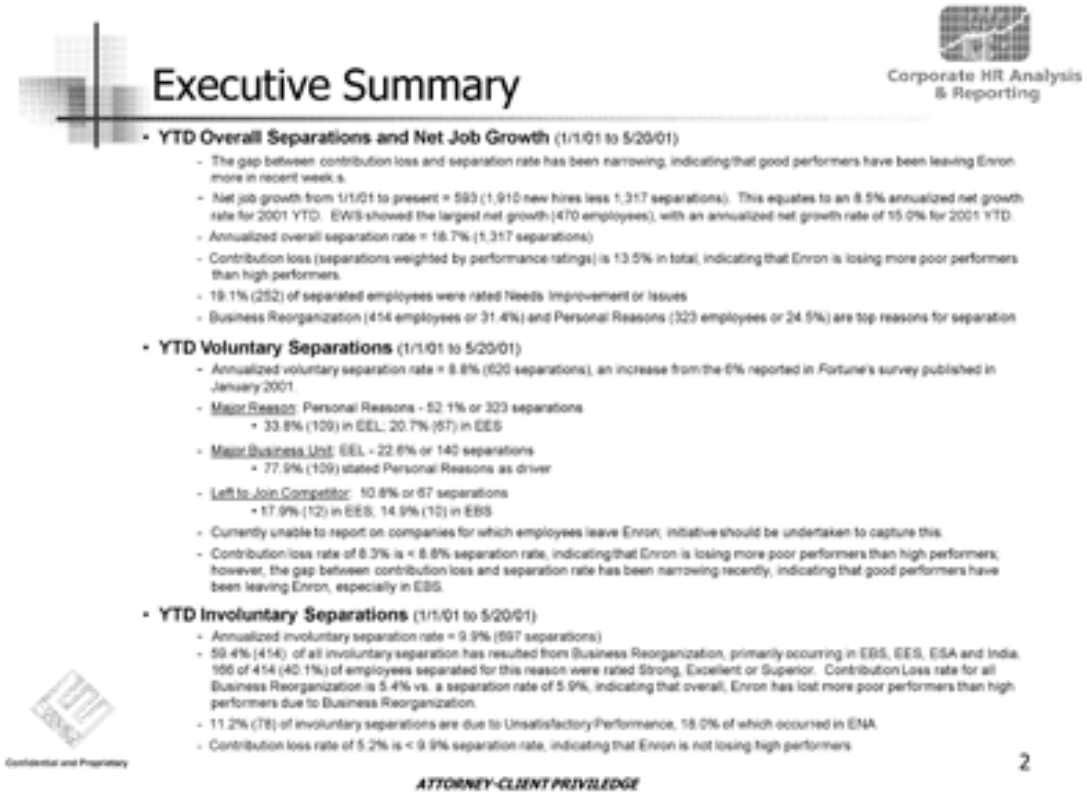
To be honest, I also hoped that I would not discover any Hot (Highly Relevant) documents in the null set. If I did, that would mean that I would have to go back and run more learning sessions. I would have to keep working to expand the scope so that the next time there would be no significant False Negatives. I was well aware of my personal prejudice not to find such documents, and so was careful to be brutally honest in my evaluation of documents. I wanted to be sure that I was consistent with past coding, that I continued the same evaluation standards employed throughout the project. If that led to discovery of hot documents and more work on my part, then so be it.

Scope of Null Set

I begin the Null Set review by noting that the random sample picked some that had already been categorized as Irrelevant as expected. I could have excluded them from the Null Set, but that did not seem appropriate, as I wanted the sample to be completely random from "all excluded," whether previously categorized or not. But I could be wrong on that principle and will seek input from information scientists on that issue. What do you think? Scientist or not, feel free to leave a comment below. Anyway, I do not think it makes much difference as only 126 of the randomly selected documents had been previously categorized.

Review of the Null Set

Next I sorted by file type to look for any obvious irrelevant I could bulk tag. None found. I did see one PowerPoint and was surprised to find it had slides pertaining to layoffs, both voluntary and involuntary, as part of the Enron bankruptcy, control number 12114291.



Executive Summary

Corporate HR Analysis & Reporting

- **YTD Overall Separations and Net Job Growth** (1/1/01 to 5/20/01)
  - The gap between contribution loss and separation rate has been narrowing, indicating that good performers have been leaving Enron more in recent weeks.
  - Net job growth from 1/1/01 to present = 593 (1,910 new hires less 1,317 separations). This equates to an 8.5% annualized net growth rate for 2001 YTD. EWS showed the largest net growth (470 employees), with an annualized net growth rate of 15.0% for 2001 YTD.
  - Annualized overall separation rate = 18.7% (1,317 separations)
  - Contribution loss (separations weighted by performance ratings) is 13.5% in total, indicating that Enron is losing more poor performers than high performers.
  - 19.1% (252) of separated employees were rated Needs Improvement or Issues
  - Business Reorganization (414 employees or 31.4%) and Personal Reasons (323 employees or 24.5%) are top reasons for separation
- **YTD Voluntary Separations** (1/1/01 to 5/20/01)
  - Annualized voluntary separation rate = 8.8% (620 separations), an increase from the 6% reported in *Fortune*'s survey published in January 2001.
  - Major Reason: Personal Reasons - 52.1% or 323 separations
    - 33.8% (109) in EEL; 20.7% (67) in EES
  - Major Business Unit: EEL - 22.6% or 140 separations
    - 77.9% (109) stated Personal Reasons as driver
  - Left to Join Competitor: 10.8% or 67 separations
    - 17.9% (12) in EES; 14.9% (10) in EBS
  - Currently unable to report on companies for which employees leave Enron; initiative should be undertaken to capture this.
  - Contribution loss rate of 8.3% is < 8.8% separation rate, indicating that Enron is losing more poor performers than high performers; however, the gap between contribution loss and separation rate has been narrowing recently, indicating that good performers have been leaving Enron, especially in EBS.
- **YTD Involuntary Separations** (1/1/01 to 5/20/01)
  - Annualized involuntary separation rate = 9.9% (697 separations)
  - 59.4% (414) of all involuntary separation has resulted from Business Reorganization, primarily occurring in EBS, EES, ESA and India. 166 of 414 (40.1%) of employees separated for this reason were rated Strong, Excellent or Superior. Contribution Loss rate for all Business Reorganization is 5.4% vs. a separation rate of 5.9%, indicating that overall, Enron has lost more poor performers than high performers due to Business Reorganization.
  - 11.2% (78) of involuntary separations are due to Unsatisfactory Performance, 18.0% of which occurred in ENA.
  - Contribution loss rate of 5.2% is < 9.9% separation rate, indicating that Enron is not losing high performers.

Confidential and Proprietary

2

ATTORNEY-CLIENT PRIVILEDGE

Following my prior rules of relevance I had to conclude this document was relevant under the expanded scope I had been using at the end, although it was not really important, and certainly not Highly Relevant. It looked like this might be a privileged document too, but that would not make any difference to my quality control analysis. It still counted.

By itself the document was not significant, but I had just started the review and already found a relevant document, a false-negative. If I kept finding documents like this I knew I was in trouble. My emotional confidence in the decision to stop the search had dropped considerably. I began bracing for the possibility of several more days of work to complete the project.

I then used a few other sort techniques for some bulk coding. The “From” field found a few obvious junk based on sender. Note that using the Short Cut Keys can help with speed. I especially like shifting into and out of *Power Mode* (for

review) with F6 and then the ALT Arrows keys on the keyboard for rapid movement, especially from one doc to the next. Keeping your hand positioned over the keys like a video game allows for very rapid irrelevancy tagging and movement from one doc to the next. You can do up to 20 individual docs per minute that way (3 seconds per doc), if the connection speed is good.

Most of these irrelevant docs are obvious and only a quick glance allows you to confirm this, so that is why you can get up to a 3 seconds per doc coding rate, even without mass categorization. Only a few in the null set required careful reading, where it may take a minute, but rarely more, to determine relevance.

This review took a bit longer than expected, primarily because I was in the office and kept getting interrupted. Starting and stopping always slows you down (except for periodic attention breaks, that actually speed you up). Not including the interruptions, it still took 4 hours to review these 1,065 documents. That means I "only" went about 260 files per hour.

The good news is I did not find another relevant document, or even arguable relevant document. One false negative out of 1,065 is an error of only .1% (actually .093%), and thus a 99.9% accuracy, a/k/a .1% *elusion* (the proportion of non-produced documents that are responsive). *See* Roitblat, H.L., *The process of electronic discovery.* Also, and this is very important to me, the one false negative document found was not important.

For these reasons, I declared the search project a success and over. I was relieved and happy.

Recap – Driving a CAR at 13,444 Files Per Hour

I searched an Enron database of 699,082 documents over nine days. That was a Computer Assisted Review ("CAR") using predictive coding methods and a hybrid multimodal approach. It took me **52 hours** to complete the search project.

(Day 1 – 8.5 hrs; day 2 – 3.5; day 3 – 4; 4 – 8; 5 – 4; 6 – 4; 7 – 7; 8 – 9; 9 – 4.) This means that my hybrid CAR cruised through the project at an average speed of *13,444 files per hour*.

That's fast by any standards. If it were a car going miles per hour, that is over seventeen times faster than the speed of sound.

This kind of review speed compares very favorably to the two other competing modes of search and review, manual linear review and keyword search. Both of these other reviews are computer assisted, but only marginally so.

The Model-T version of CAR is linear review. (It is *computer assisted* only in the sense that the reviewer uses a computer to look at the documents and code them.) A good reviewer, with average speed-reading capacities, can attain review speeds of 50 documents per hour. That's using straight linear review and the kind of old-fashioned software that you still find in most law firms today. You know, the inexpensive kind of software with few if any bells and whistles designed to speed up review. I have incidentally described some of these review enhancement features during this narrative. These enhancements, common to all top software on the market today, not just Kroll Ontrack's *Inview*, made it possible for me to attain maximum document reading speeds of up to 1,200 files per hour (3 seconds per document) during the final null-set review. I am a pretty fast reader, and have over 32 years of experience in making

relevancy calls on documents, but without these enhancements my review of documents can rarely go over 100 files per hour.

A contract review team performing a linear review at a rate of 50 docs per hours would take 13,982 hours to complete the project (699,082/50=13,982). As you have seen in this narrative, I completed the project in 52 hours. I did so by relying in a hybrid manner on my computer to work with me, under my direct supervision and control, to review most of the documents for me.

The comparison shows that manual review is *two-hundred and sixty-nine (269) times slower* than hybrid multimodal (13,982/52=269).

So much for linear review, especially when testing shows that such manual review over large scales is *not* more accurate. *See eg.* Roitblat, Kershaw, and Oot, *Document categorization in legal electronic discovery: computer classification vs. manual review*. Journal of the American Society for Information Science and Technology, 61(1):70–80, 2010. In fact, the Roitblat, *et al* study showed that a second set of professional human reviewers only agreed with the first set of reviewers of a large collection of documents 28% of the time, suggesting error rates with manual review of 72%!

Saving 93% (even with a billing rate *twenty times* as high)

Consider the costs of these CAR rides, which is central to my *bottom line driven proportional review* approach. It would be unfair to do a direct comparison and conclude that a linear review costs 269 times more than a predictive coding. Or put another way, that the state-of-the-art predictive coding CAR costs 269 time less than the old fashioned Model-T liner review method. It is an unfair comparison because the billing rate of a predictive coding skilled attorney would *not* be the same as a linear document reviewer, and the software costs would be higher.

Still, even if you assumed the skilled reviewer charged *twenty times as much*, the predictive coding review would still cost *over thirteen* times less.

Let's put some dollars on this to make it more real. For an old fashioned linear review utilizing a team of contract attorneys (often billed out anywhere from $45 - $80 depending on the market and complexity of the review), let's assume they were billed out at $50 per hour for their services. At 13,982 hours that would generate a fee of $699,100. On the other hand, at a twenty-times higher billing rate of $1,000 per hour, my 52 hours of work would cost the client $52,000. That represents a savings of $647,100. Not to mention the time savings - 13,892 hours would take a 20 person contract attorney team working 40 hours per week 17 weeks to complete, a 40 person contract attorney team working 40 hour weeks close to 9 weeks, etc. You get the idea.

My multimodal review utilizing predictive coding, even if billing at $1,000 per hour, still cost only 7.4% of what an team of contract review attorneys would have cost undertaking an old fashioned linear review. That is a 92.6% savings.

This is significantly more than the estimate of a 75% savings made in the *Rand Report*, but in the same dramatic-savings neighborhood. *Where The Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery* (2012); *also see* my blog on the Rand Report. I wonder when insurers are going to catch on to this?

Griswold's Keyword Car

But what about the reviewer driving the keyword search CAR, the gas-guzzler that seemed so cool in the 90s? What if contract reviewers were used for first review, and full-fee lawyers only used for heavy lifting and final review. Yes, it would be cheaper than all-manual linear review. But by how much? And, here is the most important part, *at what cost to accuracy*? How would the Griswold keyword wagon compare to the 2012 hybrid CAR with a predictive coding search engine?

First, let's give the Griswolds some credit. Keyword search was great when it was first used by lawyers for document review in the 1990s. It sure beat looking at everything. Use of keyword search culling to limit review to the documents with keyword hits limited the number of documents to be reviewed and thus limited the cost. It is obviously less expensive than linear review of all documents. But, it is still significantly more expensive than multimodal predictive coding culling before review. Importantly, keyword search alone is also far less accurate.

I have seen negotiated keyword search projects recently where manual review of the documents with hits showed that 99% of them were not relevant. In other words, the requesting parties keywords produced an astonishingly low precision rate of 1%. And this happened even though the keywords were tested (at least somewhat), hit-count metrics were studied, several proposed terms were rejected, and a judge (arbitrator) was actively involved. In other words, it was not a completely blind *Go Fish* keyword guessing game.

In that same case, after I became involved, the arbitrator then approved predictive coding (yes, not all such orders are published, nor the subject of sensationalist media-feeding frenzies). I cannot yet talk about the specifics of the case, but I can tell you that the precision rate went from 1% using keywords, to 68% using predictive coding. Perhaps someday I will be able to share the order approving predictive coding and my reports to the tribunal on the predictive coding search. Suffice it to say that it went much like this Enron search, but the *prevalence* and *yield* were much higher in that project, and thus the number of relevant documents found was also much higher.

But don't just take my word for it on cost savings. Look at case-law where keyword search was used along with contract reviewers. *In re Fannie Mae Securities Litigation*, 552 F.3d 814, (D.C. App. Jan. 6, 2009). True, the keyword search in the case was poorly done, but they did not review *everything*. The DOJ lawyers reviewed 660,000 emails and attachments with keyword hits at a cost of $6,000,000. The DOJ only did the second reviews and final quality control. Contract lawyers did the first review, and yet it still cost $9.09 per document.

Further, in the Roitblat, *et al* Electronic Discovery Institute study a review of 2.3 million documents by contract reviewers cost $14,000,000. This is a cost of $6.09 per document. This compares with my review of 699,082 documents for $52,000 (assuming a $1,000 per hour rate). The predictive coding review cost just over seven cents a document. *Also see* Maura Grossman & Gordon Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, Rich. J.L. & Tech., Spring 2011.

That is the bottom line: seven cents per document versus six dollars and nine cents per document. That is the power of predictive culling and precision. It is the difference between a hybrid, predictive coding, targeted approach with high precision, and a keyword search, gas-guzzler, shotgun approach with very low precision. The recall rates are also, I suggest, at least as good, and probably better, when using far more precise predictive coding, instead of keywords. Hopefully my lengthy narrative here of a multimodal approach, including predictive coding, has helped to show that. *Also see* the studies cited above and my prior trilogy *Secrets of Search: Parts One, Two, and Three.*

93% Savings Is *Not Possible* Under Real World Conditions

In future articles I may opine at length on how my review of the Enron database was able to achieve such dramatic cost savings, 93% ($52,000 vs. $699,100). For one thing, you would hope that attorneys would not review the entire set of documents, even though it had already been technically culled by deduplication, deNisting, and custodian limits. You would hope they would look for further culling alternatives to reduce the total file count. But, I am told by review companies that this kind of full linear review of full data sets still happens

everyday. So it is not far fetched to assume a full review of all 699,082 documents for comparison purposes. Even assuming the same number of documents are reviewed, I still do not think that this kind of 93% savings will often be possible in real world conditions, that 50%-75% is more realistic.

Putting aside the question of software costs, the 50%-75% savings assumes a modicum of cooperation between the parties. My review was done with maximum system efficiency, and thus resulted in maximum savings, because I was the requesting party, the responding party, the reviewer, the judge, and appeals court all rolled into one. There was no friction in the system. No vendor costs. No transaction costs or delays. No carrying costs. No motion costs. No real disagreements, just dialogue (and inner dialogue at that).

In the real world there can be tremendous transaction costs and inefficiencies caused by other parties, especially the requesting party's attorney, called *opposing counsel* for a reason. Often opposing counsel object to everything and anything without thinking, or any real reason, aside from the fact that if you want it, that means it must be bad for their client. This is especially true when the requesting party's legal counsel have little or no understanding of legal search.

Sometimes the litigation friction costs are caused by honest disagreements, such as good faith disagreements on scope of relevance. That is inevitable and should not really cost that much to work out and get rulings on. But sometimes the disagreements are not in good faith. Sometimes the real agenda of a requesting party is to make the other side's e-discovery as expensive as possible.

Unfortunately, anyone who wants to *game the system* to intentionally drive up discovery costs can do so. The only restraint on this is an active judiciary. With a truly dedicated obstructionist the 50%-75% savings from predictive coding could become far less, even nil. Of course, even without predictive coding as an issue, a dedicated obstructionist will find a way to drive up the costs of discovery. Discovery as abuse did not just spring up last year. *See* Judge Frank H. Easterbrook, *Discovery As Abuse*, 69 B.U. L. REV. 635 (1989). That is just how some attorneys play the game and they know a million ways to get away with it.

From my perspective as a practicing attorney it seems to be getting worse, not better, especially in high-stakes contingency cases. I have written about this quite a few times lately without dealing with case specifics, which, of course, I cannot do. *See eg.:*

- *Discovery As Abuse*
- *E-Discovery Gamers: Join Me In Stopping Them*
- *Judge David Waxse on Cooperation and Lawyers Who Act Like Spoiled Children*

These transaction costs, including especially the friction inherent in the adversarial system, explain the difference between a 93% savings in an ideal world, and a 75%-50% savings in a real world, under good conditions, or perhaps no savings at all under bad conditions. Still, as the software improves, and our review techniques improve, so will the review speeds, the average files per hour. For that reason the savings may continue to increase in spite of the transaction costs.

Even if we speed up the file review speeds, we must still also address the transaction costs that arise out of the adversarial system. Much of this arises from unnecessary friction between opposing counsel. Craig Ball, who, like me, is no stranger to high-stakes contingency litigation, recently made a good observation on human nature that sheds light on this situation in his LTN article *Taking Technology-Assisted Review to the Next Level*:

> It's something of a miracle that documentary discovery works at all. Discovery charges those who reject the theory and merits of a claim to identify supporting evidence. More, it assigns responsibility to find and turn over damaging information to those damaged, trusting they won't rationalize that incriminating material must have had some benign, non-responsive character and so need not be produced. Discovery, in short, is anathema to human nature.
>
> A well-trained machine doesn't care who wins, and its "mind" doesn't wander, worrying about whether it's on track for partnership.

What do you see as an option to our current adversarial-based system of e-discovery? What changes in our system might improve the efficiency of legal search and thus dramatically lower costs? Although I am grateful to the many attorneys and judges laboring over still more rule changes, I personally doubt that more *Band-Aid tweaks* to our rules will be sufficient. We are, after all, fighting against human nature as Craig Ball points out.

I suspect that a radical change to our current procedures may be necessary to fix our discovery system, that technology and rule tweaks alone may be inadequate. But I will save that thought for another day. It involves yet another paradigm shift, one that I am sure the legal profession is *not* yet ready to accept. Let's just say the Sedona Conference *Cooperation Proclamation* is a step in that direction. For more clues read my science fiction about what legal search might be like in 50 years. *A Day in the Life of a Discovery Lawyer in the Year 2062: a Science Fiction Tribute to Ray Bradbury*. In the meantime, I look forward to your comments, both on this overall search project, my final quality control check, and the implications for what may come next for legal search. Feel free to email me at Ralph.Losey@gmail.com.

In the Interests of Science

When I first wrote this narrative I planned to end at this point. The last paragraph was to be my last words on this narrative. That would have been in accord with real world practices in legal search and review where the project ends with final a quality control check and production. The 659 documents identified as relevant to involuntary employee termination would be produced, and, in most cases, that would be the end of it.

In legal practice you do not look back (unless the court orders you to). You make a decision and you implement. Law is not a science. It is a profession where you get a job done under tight deadlines and budgets. You make reasonable efforts and understand that perfection is impossible, that *perfect is the enemy of the good*.

But this is *not* a real world exercise. If it was, then confidentiality duties would not have allowed me to describe my work to begin with. This is an academic exercise, a scientific experiment of sorts. Its purpose is training, to provide the legal community with greater familiarity with the predictive coding process. For that reason I am compelled to share with you my thoughts and doubts of last week, in late July 2012, when I was rewriting and publishing Days Seven and Eight of the narrative.

I started to wonder in earnest whether my decision to stop after five rounds of predictive coding was correct. I described the decision and rationale in my Day Eight narrative. As I concluded in the *Enough Is Enough* heading: *I was pretty sure that further rounds of search would lead to the discovery of more relevant documents, but thought it very unlikely any more significant relevant documents would be found.* But now I am having second thoughts.

Troubling Questions

What if I was wrong? What if running another round would have led to the discovery of more *significant relevant documents,* and not just cumulative, insignificant relevant documents as I thought? What if a bunch of hot documents turned up? What if a whole new line of relevance was uncovered?

I also realized that it would only take a few more hours to run a sixth round of predictive and find out. Thanks to the generosity of Kroll Ontrack, the database was still online and intact. I could do it. But should I do it? Should I now take the time to test my decision? Was my decision to stop after five right, or was it wrong? And if it was wrong, how wrong was it?

I knew that if I now tested the decision by running a sixth round, the test would provide more information on how predictive coding works, on how a lawyer's use

of it works. It would lead to more pieces of truth. But was it worth the time, or the risk?

Chance and Choice

The personal risks here are real. Another round could well disprove my own decision. It could show that I was mistaken. That would be an embarrassing setback, not only for me personally, but also for the larger, more important cause of encouraging the use of advanced technology in legal practice. As I said in Day One of the narrative, *I took the time to do this in the hope that such a narrative will encourage more attorneys and litigants to use predictive coding technology*. If I now go the extra mile to test my own supposition, and the test reveals failure and delusion on my part, what would that do for the cause of encouraging others to take up the gauntlet? Was my own vanity now forcing me to accept needless risks that could not only harm myself, but others?

Of course, I could do the experiment and only reveal it if it was positive, or at least not too embarrassing, and hide it if it was. That way I could protect my own reputation and *protect* the profession. But I knew that I could never live with that. I knew that if I ran the experiment, then no matter how embarrassing the results proved to be, that there was no way I could hide that and still keep my self-respect. I knew that it would be better to be humbled than be a fraud. I knew that if I did this, if I took the time to go back and double-check my decision, that I would have to go all the way, pride and professional reputation be damned. I would have to tell all. If it was a story of delusion that discouraged other lawyers from adopting technology, then so be it. Truth should always triumph. Maybe other lawyers should be discouraged. Maybe I should be more skeptical of my own abilities. After all, even though I have been doing legal search in one form or another all my career, I have only been doing predictive coding for a little over a year.

Of course, I did not have to run the test at all. No one but a few folks at Kroll Ontrack would even know that it was still possible to do so. Everyone would

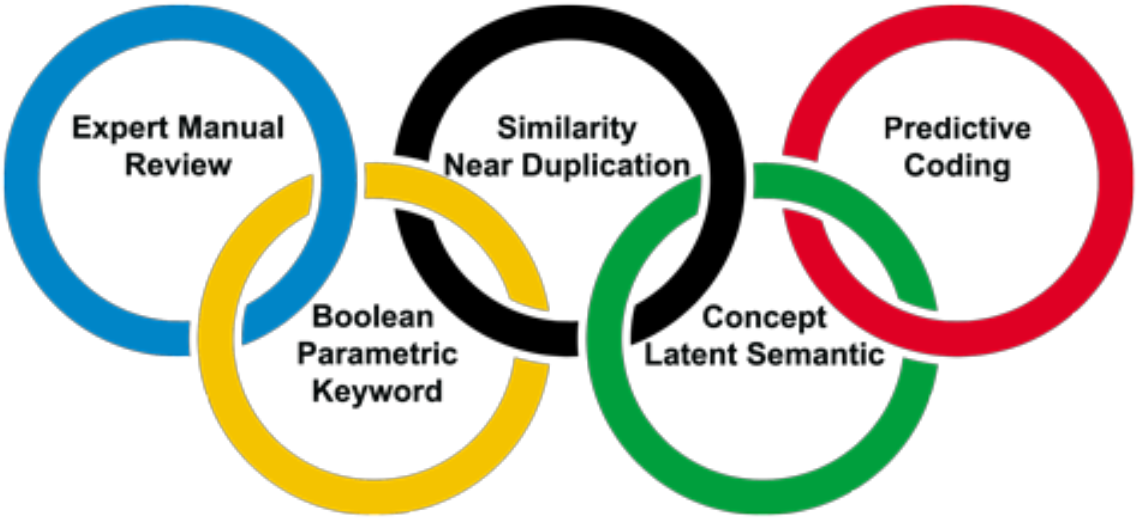
assume that the database had been taken down. By any logical analysis I should not run this test. I had little to gain if the test worked and confirmed my theory, and much to lose if it did not. Reason said I should just walk away and stick to my plan and end the narrative now. No one would ever know, except of course, I would know. Damn.

As I write this I realize that I really have no choice. I have to take the chance. A clean conscience is more important than a puffed ego, more important even than encouragement of the profession to adopt predictive coding. Anyway, what good is such encouragement if it is based on a lie, or even just an incomplete truth? I do not want to encourage a mistake. Yes, it means more work, more risk. But I feel that I have to do it. I choose to take a chance.

As I write this, I have not yet performed this experiment, and so I have no idea how it will turn out. But tomorrow is another day, the tenth day, wherein I will step outside of my normal protocol. I will run a sixth round of predictive coding to test and evaluate my decision to stop after five rounds.

Day Ten of a Predictive Coding Narrative: A post hoc test of my hypothesis of insignificant false negatives

This is the seventh and last narrative of my predictive coding search of 699,082 Enron emails. As you have seen, my legal search methodology is predictive coding dominant, but includes the four other basic types of search in a process I call *hybrid multimodal.* The five elements of *hybrid multimodal* search are shown below using the Olympic rings symbol in honor of the XXX Olympics in London that were going on when I finished this project.



Post Hoc Analysis

In Day Ten I subject myself to another quality control check, another hurdle, to evaluate my decision in day eight to close the search. My decision to stop the search in day eight after five rounds of predictive coding was based on the hypothesis that I had already found all *significant* relevant evidence. In my opinion the only relevant documents that I had not found, which in information science would be called *false-negatives*, were not important to the case. They would have some probative value, but not much, certainly not enough to continue the search project.

Put another way, my supposition was that the only documents not found and produced would be technically relevant only, and of no real value. They would certainly *not* be highly relevant (one of my coding categories). Further, the relevant documents remaining were probably of a type that I had seen before. They were cumulative in nature and thus not worth the extra time, money and effort required to unearth them. *See* my *Secrets of Search*, Part III, where I expound on the two underlying principles at play here: ***Relevant Is Irrelevant*** and ***7±2***.

This tenth day exercise was a *post hoc* test because I had already concluded my search based on my hypothesis that all significant relevant documents had been discovered. I confirmed this hypothesis to my satisfaction in the previously

described Day Nine *elusion* quality control test. This was a random sample test with a 99.9% accuracy finding. (This is to in no way intended to imply 99.9% *recall*. The *elusion* test is not intended to calculate recall.) In the *elusion* test I did a random sample test of all unreviewed documents to search for *significant* relevant evidence. Only one false negative out of a random sample of 1,065 was found and it was not significant. So I passed the test that was built into my quality control system. But would I now pass this additional *post hoc* test for significant false negatives?

Day Ten: 3 Hours

I start the day by initiating another round of predictive coding, the sixth round. It only takes a minute to start the process.

As I write this I am now waiting on *Inview* to *do its thing* and re-rank all 699,082 documents according to the new input I provided after the last session. This new input was described in Days Seven and Eight. It included my manual review and coding of two sets of 100 computer-selected training documents (total 200), plus review of all 51% plus predicted relevant documents.

At the end of day eight I had attained a count of 659 confirmed relevant documents and decided that enough was enough. I decided that any further rounds of predictive coding would likely just uncover redundant relevant documents of no real importance. I decided to stop the search, at least temporary, to see if I would pass a random sample *elusion* test for false negatives that I described in Day Nine.

As you know, the passed the test in Day Nine and so the project ended. And yet, here I am again, subjecting myself to yet another test. This Day Ten exercise is the result of my ethical wranglings described at the end of Day Nine.

Mea Culpa

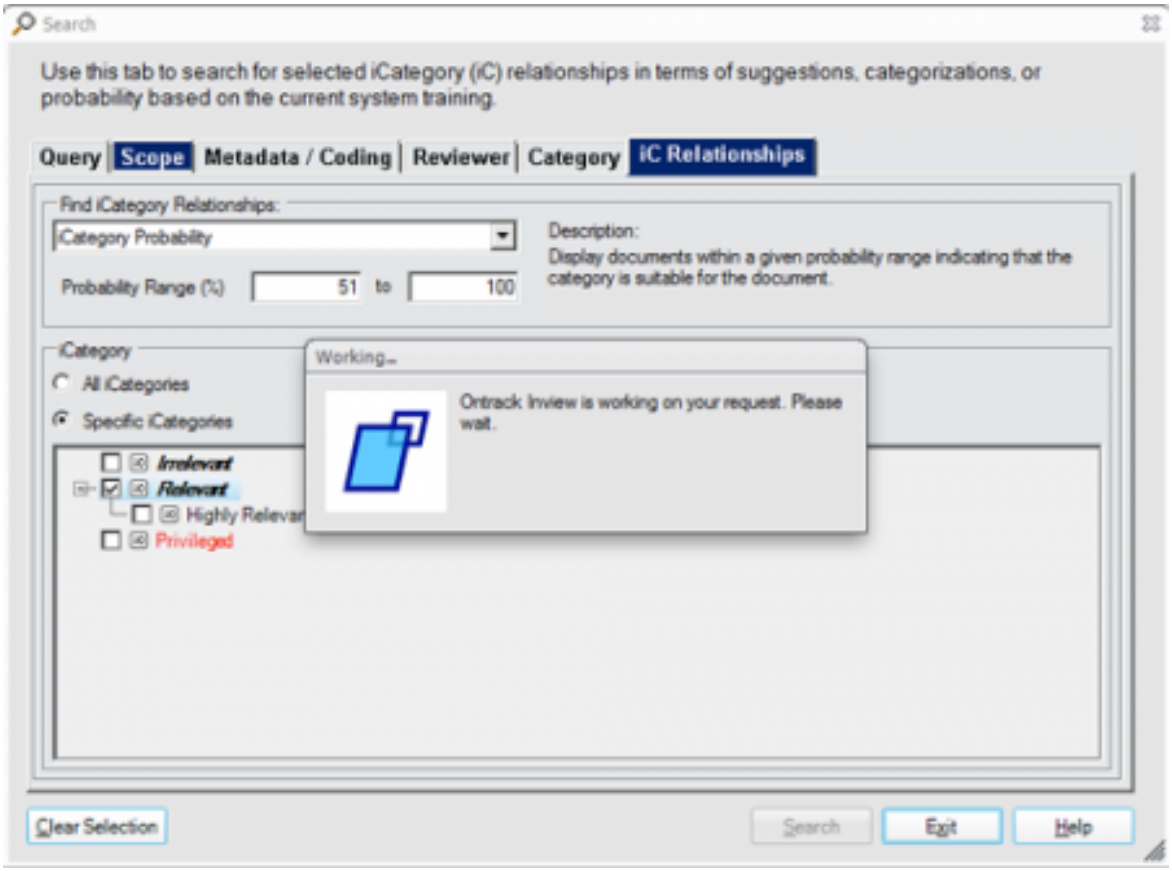
I am still waiting on Inview to give me more information, but whatever the findings, when I now look back on day eight, it seems to me like I made a mistake to stop the search when I did. Even if I pass this latest self-imposed test, and the decision is proven to be correct, it was still a mistake to have stopped there. Hopefully, just a slight mistake, but a mistake just the same. I had already trained 200 documents. I had found one new Highly Relevant document. I had provided new training information for the

computer. Why not just take a couple of more hours to see what impact this would have?

The lesson I learned from this, which I pass on to you here, is never to stop a project until you see the last report and results of training documents. Why guess that nothing of importance will come of the next training when it is easy enough to just run another round and find out? The answer, of course, is time and money, but here I guessed that only a few new relevant documents would be found, so the costs of the extra effort would be negligible. In retrospect, I think I was too confident and should have trusted by instincts less and my software more. But I will soon see for myself if this was harmless error.

Moment of Truth

The Moment of truth came soon enough on a Sunday morning as I logged back on to Inview to see the results of the Sixth Round. I began by running a search for all 51%+ predicted relevant documents. The search took an unusually long time to run. In the meantime I stared at this screen.



Call me over-dramatic if you will, but I was getting nervous. What if I made a bad call in stopping before the sixth round?

Finally it completed. There were 566 documents found. So far so good. Slight sigh of relief. If it were thousands I would have been screwed. Remember, I had already coded 659 documents as relevant. The computer's predicted relevant numbers were less than my last actuals.

After determining count, I sorted by categorization to see how many of the predicted relevant had not previously been categorized? In other words, how many of the 566 were new documents that I had not looked at before? Another slight sigh of relief. The answer was 51. These were 51 new documents that I would now need to look at for the final test. So far, this is all as predicted. But now to see if any of them were significant relevant. (Remember, I had predicted that some relevant would left, just not significant relevant.)

I noticed right away that 1 of the 51 documents had already been reviewed, but not categorized. I frequently did that for irrelevant documents of a type I had seen before. It was an Excel spreadsheet with voluntary termination payout calculations. I still thought it was irrelevant. Now on to the 50 documents that I had not reviewed before.

The 50 New Documents

Four of the fifty were the same email with the subject *Bids Open for Enron Trading Unit*. They had a 71.3% prediction of relevance. It was an AP news article. It had to do with an upcoming bankruptcy sale of Enron contracts. It included discussion of employees complaining about Enron's employee termination policy. Here is the relevant excerpt for your amusement. Note the reference to protesters carrying *Moron* signs.

```
.
Enron spokeswoman Karen Denne declined comment on the decision.
Dynegy sued Enron in Texas state court in Houston the day after Enron filed one of the
largest Chapter 11 bankruptcies in U.S. history Dec. 2 and sued Dynegy for $10 billion for
breach of contract in New York.
Dynegy and other creditors also have asked that the bankruptcy case be moved to Houston,
where Enron, Dynegy and many of Enron's 800 or so creditors are based. Gonzalez will consider
those requests Jan. 7.
Also Wednesday, nearly 40 of some 4,500 Enron employees laid off after the company filed for
bankruptcy gathered outside Enron headquarters here to discuss the company's severance policy
and sign a complaint to Enron on the lack of information.
Some carried signs that displayed the word ``Moron'' under Enron's logo and said ``What were
they thinking?'' Others wore T-shirts that said ``The Execs Who Stole Christmas.''
Gonzalez on Dec. 4 $1.5 billion in short-term financing to keep the company afloat and fund
$4,500 severance payments.
Former workers received those checks last week, but revelations that nearly 600 employees
deemed critical to Enron's survival received more than $100 million in retention payments
upset those entitled to more money under Enron's severance policy, they said.
```

It might be relevant, or might not. It was a newspaper article, nothing more. No comments by any Enron employees about it. I guess I would have to call it

marginally relevant, but unimportant. There were now only 46 documents left to worry about.

The next document I looked at was a three-page word document named *Retention program v2.doc*. It had to do with the payment of bonuses to keep employees from leaving during the Enron collapse. It had a 59.3% probable relevant prediction. I considered it irrelevant. There were several others like that.

Another document was an email dated November 15, 2001 concerning a rumor that Andy Fastow was entitled to a nine million dollar payout due to change in control of Enron. I remembered seeing this same email before. I checked, and I had seen and marked several copies of versions of this email before as marginally relevant. Nothing new at all in this email. There were several more document examples like that, about 25 altogether, documents that I had seen before in the exact same or similar form. Yes, they were relevant, but again duplicative or cumulative. It was a complete waste to time to look at these documents again.

I also ran into a few documents that were barely predicted relevant that had to do with voluntary termination and payment of severance for voluntary termination. The software was still having trouble making the differentiation between irrelevant voluntary and relevant involuntary. It was understandable in view of the circumstances. It was a grey area, but bottom line, none of these borderline documents presented were deemed relevant by me during this last quality control review.

One new relevant document was found, a two page spreadsheet named *Mariner events.xls* bearing control number 1200975. It had an agenda for a mass termination of employees on August 23, 2001. It apparently pertained to a subsidiary or office named Mariner. I had seen agendas like this before, but not this particular one for this particular office. I had called the other agendas relevant, so I would have to consider this one relevant too. But again, there was nothing especially important or profound about it.

In that same category as a new relevant document, but not important, I would include an email dated November 20, 2001, from Jim Fallon, bearing control number 11815873, who was trying to get his employment agreement changed to, among other things, provide benefits in case of termination.

The last document I considered seemed to address involuntary terminations and tax consequences of some kind concerning a so-called *clickathome* program. Frankly, I did not really understand what this was about from this email chain.

The last date in the chain was June 15, 2001. The subject line is *Clickathome – proposed Treatment for Involuntary Terminations – Business reorganizations*. It has control number 15344649 and is three pages long. It was predicted 66.9% likely relevant. The emails look like they pertain to employees who are transferred from one entity to another, and does not really involve employment termination at all. I cannot be sure, but it certainly is not important in my mind. Here is a portion of the first page.

Subject: RE: Clickathome - Proposed treatment for Involuntary Terminations - Business Reorganizations

All,

Please let me know by noon tomorrow if you have any issues with this approach. If not, we are going to roll out this change effective close of business tomorrow.

Steve/Cindy, I'd suggest that we simply make this change tomorrow. Do you feel the need to run it by any of the policy committee beforehand?

Kalen

From: James Sandt/ENRON@enronXgate on 06/14/2001 05:00 PM
To: Suzanne Brown/ENRON@enronXgate, Sharon Butcher/ENRON@enronXgate, Kalen Pieper/HOU/EES@EES, Gary P Smith/ENRON@enronXgate, Marla Barnard/Enron Communications@Enron Communications, David Oxley/ENRON@enronXgate, Robert W Jones/ENRON@enronXgate, Cindy Olson/ENRON@enronXgate, Drew Lynch/Enron@EUEnronXGate, Mary Joyce/ENRON@enronXgate, Cynthia Barrow/ENRON@enronXgate
cc: Sarah A Davis/ENRON@enronXgate, Marie Newhouse/ENRON@enronXgate, Elizabeth Boudreaux/ENRON@enronXgate
Subject: RE: Clickathome - Proposed treatment for Involuntary Terminations - Business Reorganizations

Suzanne:

It is my understanding that everyone is signed off on the suggested revisions to the Clickathome documents relating to redeployed individuals. Tax is.

In summary, if a person is terminated from Enron or one of its subsidiaries as a result of a business reorganization, the forfeiture penalty is waived.

1

I was kind of curious as to what the *clickathome* program was that the emails referred to, so I *Goggled* it. At page two I found an Enron document that explained:

clickathome is Enron's new program that gives eligible employees a computer and Internet connection (including broadband connectivity where available through program-approved vendors) for use at home.

Now I understood the reference in the email to a "PC forfeiture penalty." I suppose maybe this email chain worrying about tax consequences of PC forfeiture in the *clickathome* program might be technically relevant, but again, of no importance. Just to be sure I was not missing anything, I also keyword searched the Enron database for *clickathome* and found 793 hits. I looked

around and saw many emails and documents had been reviewed before and classified as irrelevant that pertained to the *clickathome program* where an Enron employee could get *free PC* from Dell. I was now comfortable that this email chain was also unimportant.

Hypothesis Tested and Validated

This meant that I was done. The second quality control test was over. Although I found 32 technically relevant documents as described above, no major relevant documents had been found. I had passed another test. (If you are still keeping score, the above additional review means I found a total of 691 relevant documents (659+4+25+1+1+1) out of my *yield point projection* at the beginning of the project of 928 likely relevant. That means a score of almost 75%. Not bad.)

It all went pretty much as expected and predicted at the end of Day Eight. I had wasted yet another perfectly good Sunday afternoon, but at least now I knew for sure that the sixth round was not necessary. My hypothesis that only marginally relevant documents would turn up in another round had been tested and validated.

I suppose I should feel happy or vindicated or something, but actually, *tired* and *bored* are the more accurate adjectives to describe my current mood. At least I am not feeling embarrassed, as I was concerned might happen.

By the way, the three hours that this last day took would have gone faster but for the many Internet disconnects I experienced while working from home. My three hours of reported time did not include the substantial write-up time, nor time waiting for the computer to train. Sigh. Test and writing is over. Time to jump in the pool!

Conclusion: *Come On In, The Water's Fine*

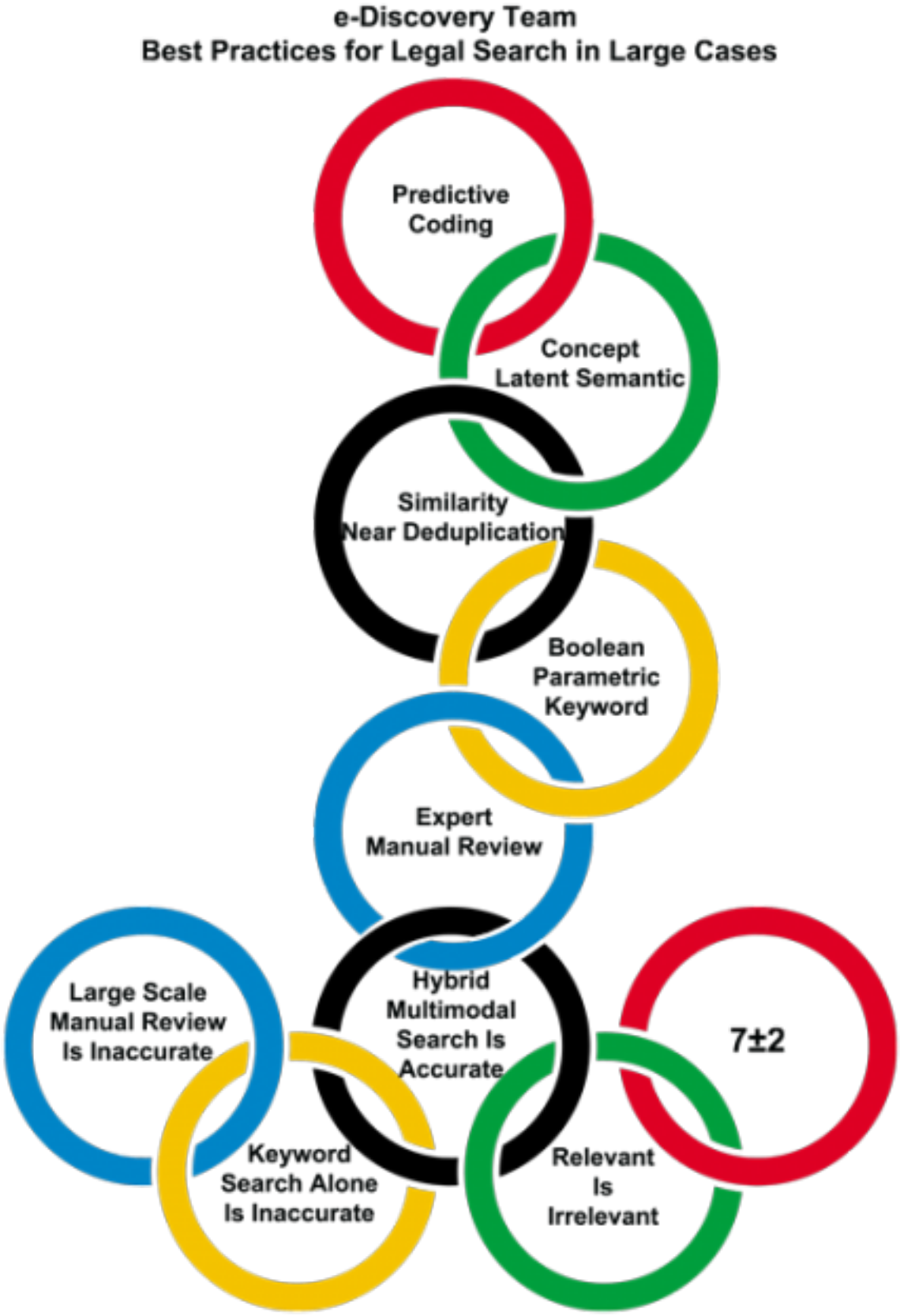
I hope this longer than intended narrative fulfills its purpose and encourages more lawyers to jump in and use predictive coding and other advanced technologies. The water is fine. True, there are sharks in some pools, but they are outside the pool too. They are a fact of life in litigation today. *Discovery As Abuse* is a *systemic* problem, inherent in the adversarial model of justice. The abuses are by both sides, including requesters who make intentionally over-broad demands and drive up the costs every chance they get, and responders who play *hide-the-ball*. Predictive coding will not cure the systemic flaws, but it will lessen the bite.

The multimodal hybrid CAR with a predictive coding search engine can mitigate your risks and your expenses. More often than not, it can save you anywhere from 50% to 75% in review costs and improve recall. The new technology is *win*

win for both requesting parties and responding parties. I urge everyone to give it a try.

When you go in and swim please remember the five rules of search safety. They were explained in my *Secrets of Search* trilogy in parts *One*, *Two*, *and* *Three* and are shown below in another version of the Olympic rings.

These five, when coupled with the five Olympic rings of multimodal search shown at the top of this essay, provide a blueprint for effective legal search. These ten, shown as one large symbol below, are a kind of *seed set* of best-practices principles. The legal profession can use them as a beginning to develop full peer-reviewed best practices for reasonable legal search. (A few months later I began this process by creating ***Electronic Discovery Best Practices***, found at EDBP.com.)



Feel free to contact the author with any comments you may have at ralph.losey@gmail.com.