e-Discovery Team at TREC 2015 Total Recall Track

Ralph C. Losey[*]
National e-Discovery Counsel
Jackson Lewis P.C.
e-DiscoveryTeam.com
Ralph.Losey@gmail.com

Jim Sullivan and **Tony Reichenberger**
Sr. Discovery Services Consultants,
Kroll Ontrack, Inc.
eDiscovery.com
JSullivan@kollontrack.com
TReichenberger@krollontrack.com

ABSTRACT

The 2015 TREC Total Recall Track provided instant relevance feedback in thirty prejudged topics searching three different datasets. The *e-Discovery Team* of three attorneys specializing in legal search participated in all thirty topics using Kroll Ontrack's search and review software, *eDiscovery.com Review* (*EDR*). They employed a *hybrid* approach to continuous active learning that uses both manual and automatic searches. A variety of manual search methods were used to find training documents, including high probability ranked documents and keywords, an *ad hoc* process the *Team* calls *multimodal*.

In the one topic (109) requiring legal analysis the *Team's* approach was significantly more effective than all other participants, including the fully automated approaches that otherwise attained comparable scores. In all topics the Team's *hybrid multimodal* method consistently attained the highest F1 values at the time of *Reasonable Call*, equivalent to a stop point. In all topics the *Team's* multimodal human machine approach also found relevant documents more quickly and with greater precision than the fully automated or other methods.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Search process, relevance feedback, supervised learning, best practices.

Keywords: Hybrid Multimodal; AI-enhanced review; predictive coding; predictive coding 3.0; electronic discovery; e-discovery; legal search; active machine learning; continuous active learning; CAL; Computer-assisted review; CAR; Technology-assisted review; TAR; relevant irrelevant training ratios.

1. INTRODUCTION

The *e-Discovery Team* participated in all thirty Total Recall Track topics in the *Athome* group where both manual and automatic methods were permitted. The *Team* is composed of three practicing attorneys who specialize in legal search. They used Kroll Ontrack's search and review software, *eDiscovery.com Review* ("*EDR*"), employing what they call a *hybrid multimodal* method.[1] They attained high recall and precision in most of the thirty topics. The few exceptions appear derived from the fact that the attorneys are accustomed to self-defining the ground truth, and, in some topics, their opinions on relevance differed significantly from the TREC assessors. In later topics the attorney *Team* learned to turn off their own judgments and rely primarily on their software's automated processes, which generally led to improved scores better matching the TREC relevance assessments. The Team's manual efforts, as measured by time expended and number of documents manually reviewed, were very low by legal search standards.

---

[*] The views expressed herein are solely those of the author, Ralph Losey, and should not be attributed to his firm or its clients.

The fully automatic methods employed by the *Sandbox* group participants in the Total Recall Track attained comparable high recall and precision in most topics. The *Team's hybrid multimodal* method did, however, consistently attain the highest F1 values at the time of *Reasonable Call*, equivalent to a training stop point, which is very important to legal search. One of the thirty topics, 109 - Scarlet Letter Law - required a small amount of legal knowledge and analysis to understand relevance (most of the others required none). On this topic our legal team, as you would expect, attained significantly better results than the fully automated methods that contained no base legal knowledge.

The *e-Discovery Team's hybrid multimodal* method is a type of continuous active learning text retrieval system that employs supervised machine learning and a variety of manual search methods.[2, 3] The *Team* attained very high recall and precision rates in most, but not all, of the thirty Total Recall topics. The Team's F1 scores at the time of *Reasonable Call* ranged from a perfect score of 100% in one topic (3484), to 91% to 99% in eight topics, and 82%-87% in five others. Although, of course, not directly comparable, these scores are far higher than any previously recorded in the six years of TREC Legal Track (2006-2011) or any other study of legal search. One reason for this may be that the thirty topics in the 2015 Total Recall track presented relatively simple information needs by legal search standards, with one exception (Topic 109 – Scarlet Letter Law). Another may be improved software and the Team's improved *hybrid multimodal* method that includes continuous active learning.

The *e-Discovery Team* was able to find the target relevant documents in all thirty topics with relatively little human effort and almost no legal analysis. Only Topic 109 required legal knowledge and analysis, with four others - 101, 105, 106, 107 - requiring some small measure of analysis.

A total of 16,576,798 documents were classified in thirty topics. Of these documents 70,414 were predetermined by TREC assessors to be relevant. The *e-Discovery Team* found these relevant documents by manual review of only 32,916 documents. The other 37,498 relevant documents were found with no human review of these documents.

1.1 Total Recall Track Description – *Athome* and *Sandbox*.

The Total Recall track offered 30 different pre-judged topics for search in two different divisions, *Athome* and *Sandbox.* Our Team only participated in the *Athome* experiments. In the *Athome* experiments the data was loaded onto the participants' own computers. There were no restrictions on the types of searches that could be performed. The setup allowed the *e-Discovery Team* to use a slightly modified version of our standard *Hybrid Multimodal* method, which, as mentioned, employs both *ad hoc* manual review and machine learning.

The *Sandbox* participants were only permitted to use fully automated systems and the data remained on TREC administrator computers. They searched the same three datasets as *Athome*, plus two more not included in the *Athome* division due to confidentiality restrictions. The Sandbox participants were prohibited from any manual review of documents or *ad hoc* search adjustments.[4] Even after the submissions ended, the *Sandbox* participants reported at the Conference that they *never looked at any documents*, even the unrestricted *Athome* shared datasets. They never made any effort to determine where their software made errors in predicting relevance, or for any other reasons. To these participants, all of whom were academic institutions, the ground truth itself was of no relevance.

Three different datasets were searched in both the *Athome* and *Sandbox* events, with the same ten topics in each. Even though the data searched and topics overlapped in the two divisions, none of the participants in one division participated in the other division. This is unfortunate because it makes direct comparisons problematic, if not impossible, especially as to

the software systems used. It is hope that some participants will participate in both events in future Total Recall tracks.

The e-Discovery Team participated in all thirty of the *Athome* topics. We were the *only* manual participant to do so, with all others completing ten or fewer topics. The lack of participation by others in the *Athome* group also make meaningful comparisons very difficult or impossible, but we note that the *e-Discovery Team's* scores were consistently higher than any other *Athome* participants.

At Home participants were asked to track and report their manual efforts. The *e-Discovery Team* did this by recording the number of documents that were *human reviewed* and classified. Virtually all documents human reviewed were also classified, although all documents classified were not used for active training of the software classifier. Moreover 53% of the relevant documents used for training were never human reviewed. We also tracked effort by number of attorney hours worked as is traditional in legal services.

The Team used Kroll Ontrack's software, known as *eDiscovery.com Review, or EDR*, which includes active machine learning features, a/k/a *predictive coding* in legal search. *EDR* employs a proprietary probabilistic type of logistic regression algorithm for document classification and ranking.

The *At Home* participants used their own computer systems and software for search, and then submitted documents to the TREC administrator that they considered relevant. TREC set up a "jig" whereby instant feedback was provided to a participant as whether each document submitted as relevant was in fact previously judged to have been relevant by TREC assessors. When a participant determined that a reasonable effort had been made to find all relevant documents required, which is important in legal search and represents a stopping point for further machine training and document review, they would notify TREC of this supposition and "Call Reasonable." Continued submissions were made after that point so that all documents were classified as either relevant or irrelevant. The goal as we understood it was to submit as many relevant documents as possible before the Reasonable call, and thereafter to have all false negatives appear in submissions as soon after the Reasonable Call as possible.

Most of the thirty topics presented only simple, single-issue *information needs* suitable for single-facet classification. Further, only a few of the topics required any legal analysis for relevance identification. These two factors, plus the omission of metadata, was, we think, a disadvantage to the *e-Discovery Team* of lawyers. Conversely, it appears that these same factors made it simpler for the academic *Sandbox* participants to perform well in most topics using fully automated methods. It should also be noted that although our lawyer Team was practiced and skilled in complex information needs requiring extensive legal analysis, and had long experience with projects using SME defined ground truths, none had any prior experience using machine learning for the types of searches presented in the 2015 Recall Track.

The one exception that brought in legal analysis with beneficial SME analysis, was Topic 109, Scarlett Letter Law. It required some legal knowledge, albeit very rudimentary, to begin locating relevant documents. The keywords alone - "Scarlett Letter Law" – would only find relevant documents with this word combination and similar text patterns. These words were just the *nickname* of the proposed and eventually enacted Florida Statute. Any attorney would know that to find relevant information they would not only have to search the name, but they would also have to search the various house and senate bill numbers for this law. These numbers would not often appear in the same document as the nickname, and since the machine did not know to search for these numbers, it did not realize the significance. Eventually the automated machine learning did see the connection, after many relevance feedback submissions. These

submissions and instant feedback of relevant, or not, would, of course, not happen in real legal search.

1.2 Governor Bush Email

The first set of *Athome* Topics searched a corpus of 290,099 emails of Florida Governor Jeb Bush. Most of the metadata of these emails and associated attachments and images had been stripped and converted to pure text files. This increased the difficulty of the Team's search, which normally includes a mixture of metadata specific searches.

A significant percentage of the Bush emails were *form type* lobbying emails from constituents, which repeated the same language with little of no variance. The unusually high prevalence of near-duplicate emails made search of many of the Bush topics easier than is typical in legal search.

The ten Bush email topics searched, and their names, which were the only guidance on relevance provided to either the *Athome* or *Sandbox* participants, are shown below.

Topic 100	**School and Preschool Funding**
Topic 101	**Judicial Selection**
Topic 102	**Capital Punishment**
Topic 103	**Manatee Protection**
Topic 104	**New Medical Schools**
Topic 105	**Affirmative Action**
Topic 106	**Terri Schiavo**
Topic 107	**Tort Reform**
Topic 108	**Manatee County**
Topic 109	**Scarlet Letter Law**

E-Discovery Team leader, Ralph Losey, a lifelong Florida native, personally searched each of these ten Topics. In about half of the topics his personal knowledge of the issues was helpful, but in several others it was detrimental. He had definite preconceptions of what emails he thought should be relevant and these sometimes differed significantly from the TREC assessors. In all of the Bush Topics Losey was at least somewhat assisted by a single "contract review attorney."[5] The contract attorneys in most of these ten Topics did a majority of the document review under Losey's very close supervision, but had only limited involvement in initial keyword searches, and no involvement in predictive coding searches or related decisions.

All participants in the 2015 Recall Track were required to complete all ten of the Bush Email Topics. Completion of the other twenty Topics in the two other data collections was optional. Several participants started review of the Bush Topics, but did not finish, and thus were not permitted to submit a report or attend the TREC Conference. Only one other *Athome* participant, Catalyst, completed all ten Bush Topics. No other *Athome* participants even attempted the other twenty topics, and thus comparisons with the *e-Discovery Team's* results are limited to the fully automatic participants.

1.3 Black Hat World Forums.

The second set of *Athome* Topics searched a corpus of 465,149 posts taken from *Black Hat World Forums*. Again, almost all metadata of these posts and associated images had been stripped and converted to pure text files. The ten topics searched, and their names, which again were the only guidance initially provided on relevance, are shown below.

Topic 2052	**Paying for Amazon Book Reviews**
Topic 2108	**CAPTCHA Services**
Topic 2129	**Facebook Accounts**
Topic 2130	**Surely Bitcoins can be Used**
Topic 2134	**PayPal Accounts**
Topic 2158	**Using TOR for Anonymous Internet Browsing**
Topic 2225	**Rootkits**
Topic 2322	**Web Scraping**
Topic 2333	**Article Spinner Spinning**
Topic 2461	**Offshore Host Sites**

The *Team* members again had expertise issues with some of these arcane topics that they happened to be familiar with. Their knowledge would sometimes prove detrimental. Again, as the review continued, the *Team* members learned to suspend their own knowledge and ground truth judgments and instead rely entirely on the automated ranking searches, much like the fully automated participants always necessarily did.

1.4 Local News Articles.

The third set of *Athome* Topics searched a corpus of 902,434 online *Local News Articles*, again in text only format. The ten topics searched, and their names, which again were the only guidance provided on relevance aside from the instant feedback, are shown below.

Topic 3089	**Pickton Murders**
Topic 3133	**Pacific Gateway**
Topic 3226	**Traffic Enforcement Cameras**
Topic 3290	**Rooster Turkey Chicken Nuisance**
Topic 3357	**Occupy Vancouver**
Topic 3378	**Rob McKenna Gubernatorial Candidate**
Topic 3423	**Rob Ford Cut the Waist**
Topic 3431	**Kingston Mills Lock Murders**
Topic 3481	**Fracking**
Topic 3484	**Paul and Cathy Lee Martin**

The Team found the *News Articles* less difficult to work with than our typical legal search of corporate ESI. Still, the same kind of ground truth validity and consistency issues were noted in some of the news topics, but to a lesser degree than the other two datasets.

1.5 *E-Discovery Team's* Three Research Questions.

Our first and primary question was to determine: *What Recall, Precision and Effort levels the e-Discovery Team would attain in TREC test conditions over all 30 Topics using the Team's*

Predictive Coding 3.0 hybrid multimodal search methods and Kroll Ontrack's software, eDiscovery.com Review (EDR).

Our secondary question was: *How will the Team's results using its semi-automated, supervised learning method compare with other Recall Track participants using semi automated supervised or fully automated unsupervised learning methods.*
Our last question was: *What are the ideal ratios, if any, for relevant and irrelevant training examples to maximize effectiveness of active machine learning with EDR.*

2. RELATED WORK

It is generally accepted in the legal search community that the use of *predictive coding* type search algorithms can improve the search and review of documents in legal proceedings.[6] The use of predictive coding has also been approved, and even encouraged by various courts around the world, including numerous courts in the U.S.[7]

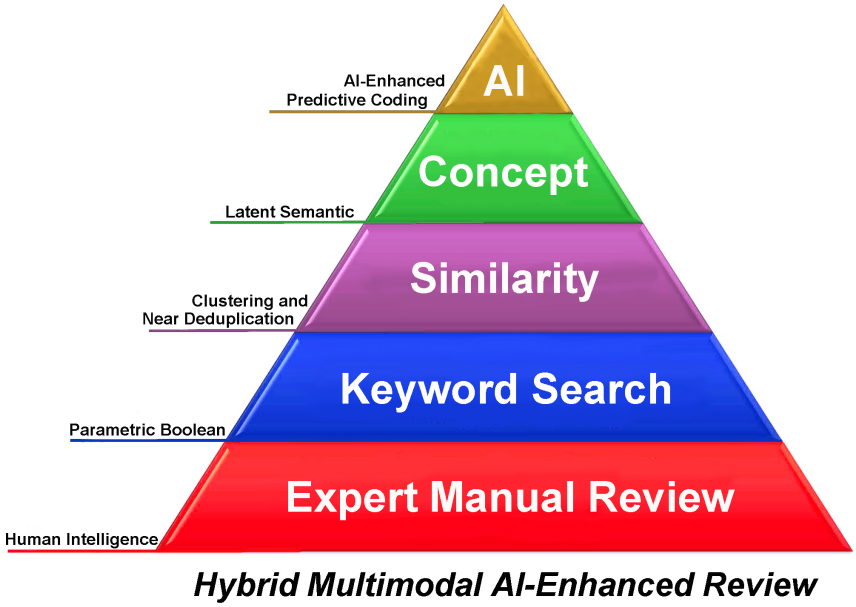
Although there is agreement on use of predictive coding, there is controversy and disagreement as to the most effective ***method****s* of use.[8] There are, for instance, proponents for a variety of different methods to find training documents for predictive coding. Some advocate for the use of chance selection alone, others for the use of top ranked documents alone, others for a combination of top ranked and mid-level ranked documents where classification is unsure, and still others, including Losey, call for the use of a combination of all three of these selection processes and more.[9] The latest respectful disagreement is between Losey's *e-Discovery Team*, and the Administrators of the Total Recall Track, Grossman and Cormack, concerning the advisability of: 1) keeping attorney search experts in the loop, the *hybrid* approach, as opposed to the *fully automated* approach; and 2) using a variety of search methods, the *multimodal* approach, as opposed to reliance on high ranking documents alone for machine training.[10]

Some attorneys, predictive coding software vendors, and, apparently, Grossman and Cormack, advocate for the use of predictive coding search methods alone, and forego other search methods when they do so, such as keyword search, concept searches, similarity searches and linear review. *E-Discovery Team* members reject that approach and instead advocate for a *hybrid multimodal* approach that they call *Predictive Coding 3.0*, further described below. It uses all methods. As discussed in Endnote 2, we reject the notion of inherent lawyer bias that underlies some experts' fully automated approaches, including, but to a lesser degree, Grossman and Cormack. We instead seek to augment and enhance attorney search experts, not automate and replace them. We do, however, favor certain safeguards against the propagation of errors, intentional or inadvertent, and advocate within the legal community for continuous active training of lawyers in search techniques and ethics.

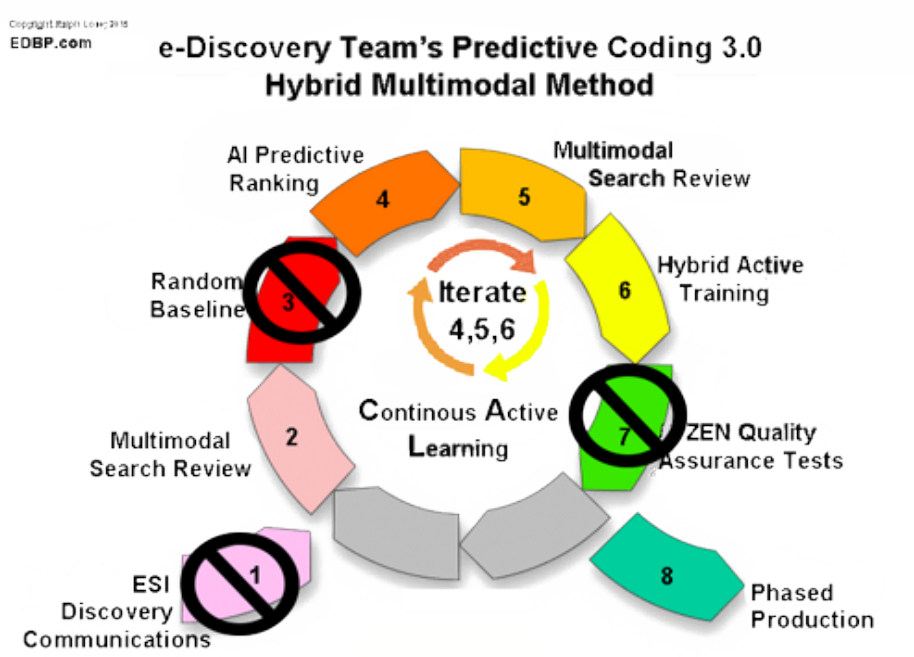
Our participation in the 2015 TREC Total Recall Track, the research questions we posed, and the experiments we performed, were not in any manner designed or intended to attempt to resolve this current methodology dispute with the Administrators of this Track. In fact, it was only at the 2015 Conference that we fully understood the extent of these differences. Although Grossman and Cormack did individually participate in this Track, as well as administrator it, and so too did other groups from Cormack's university, they did not participate in the manual Athome division that we did. To our knowledge the Total Recall track was not designed to address this newly emerging disagreement in preferred methodologies, nor advance any one particular methodology. Still, we would concede that, subject to normal caveats, some indirect lessons can be derived on this issue from the Total Recall Track results.

3. HYBRID MULTIMODAL APPROACH

The *e-Discovery Team* approach includes all types of search methods, with primary reliance placed on predictive coding and the use of high-ranked documents for continuous active training. In that way it is similar to the approach used by Grossman and Cormack,[11] but differs in that the *Team* uses a *multimodal* selection of search methods to locate suitable training documents, including high ranking documents, some mid-level ranked uncertain documents, and all other search methods, including keyword search, similarity search, concept search and even occasional use of linear review and random searches. The various types of searches usually included in the Team's multimodal approach are shown in the search pyramid, below.



The standard eight-step workflow used by the Team in legal search projects is shown in the diagram below. A step by step descriptions of the workflow can be found in e-Discovery Team writings.[12] The application of this methodology can be seen the Team's description of their work in each of the thirty Topics that is included in the Appendix. Our usual steps One, Three and Seven had to be omitted or severely constrained to meet the TREC experiment format.



Standard steps Three and Seven of the workflow were omitted to meet the time requirements of completing every review project in 1.5 days. Skipping these steps allowed us to complete 30 review projects in 45 days in the Team's *spare time*, but had a detrimental impact.

Our usual first step, *ESI Discovery Communications,* is where our information needs are established. This had to omitted to fit the format of the Recall Track *Athome* experiments. The only communication under the TREC protocol was a very short, often just two-word description of relevance, plus instant feedback in the form or yes or no responses as to whether particular documents submitted were relevant. In the *e-Discovery Team's* typical workflow discovery communications typically involve: 1) detailed requests for information contained in court documents such a *subpoenas* or *Request For Production*; 2) input from a qualified SME, who is typically a legal expert with deep knowledge of the factual issues in the case and how the presiding judge in the legal proceeding will likely rule on borderline relevant issues; and, 3) dialogues with the client, witnesses, and with the party requesting the production of documents to clarify the search target.

The *Team* never receives a request for production with just two or three word descriptions as encountered in the TREC experiments. When the *Team* receives vague requests, which is common, the *Team* seeks clarification in discussions (Step One). In practice if there is disagreement as to relevance between the parties, which is also common, the presiding judge is asked to make relevance rulings. Again, none of this was possible in the TREC experiments.

All of our usual practices in Step One had to be adjusted to the submissions format of the 30 *Athome* Topics. The most profound impact of these adjustments was that the attorneys on the *Team* often lacked a clear understanding as to the intended scope of relevance and the rationale behind the automated TREC relevance rulings on particular documents. These protocol changes had the impact of minimizing the importance of the SME role on the active machine learning process. Instead, this role was often shifted almost entirely to the analytics of the EDR software. The software analytics could often see patterns, and correctly predict relevance, that the human attorney reviewers could not (often, but not always, because the human reviewers disagreed

with the TREC assessors human judgment of ground truth in several topics, and otherwise could not follow or see any logic to the documents returned as relevant).

This minimization of the importance of the SME role is *not common* in legal search where attorney reviewers always have some sort of understanding of relevance. The role of the SME in the Team's decades of experience in legal search has always been important to help ensure high quality, trustworthy results. Contrary to the unfortunate popular belief among laypersons going back to the time of Shakespeare,[13] the vast majority of legal professionals maintain very high standards of ethics and trustworthiness. In spite of the alleged negative influences of the centuries old adversarial tradition of the common law, attorneys are dedicated to uncovering *the truth, the whole truth, and nothing but the truth*, regardless of the particular case impact. Any notion of inherent bias by attorneys is misplaced. It is, after all, attorneys who control the discovery process and define relevance, and attorneys, not robots or scientists, who make the production of relevant documents to the *other side*.[14]

Scientific research is better served when driven by reason and objective measurements, not prejudices and assumptions about an entire profession and our common law system of justice, based as it is on an adversarial truth seeking process. The *e-Discovery Team* will continue to look for ways to improve quality control, and guard against inadvertent errors, which always exists in any human endeavor, and identify intentional errors, which rarely exist in legal search, but, we concede may sometimes take place. For that reason we will explore greater reliance on automated process in our future research and other quality control techniques.[15] We will not, however, abandon a *hybrid* approach where a human remains, if not in control, then at least as an active partner, out of any subjective prejudices against lawyers. We also refuse to accept the unproven assumption that our adversarial system is inherently suspect, encourages bias, and otherwise requires that humans be removed from e-discovery and replaced by robots. Conversely, we do not naively assume lawyers are automatically superior to machines. We have long advocated against the current legal standard of only using manual review of every document. The Team's *hybrid* approach aims for a proportional balance.

4. EXPERIMENTS AND DISCUSSIONS

The *e-Discovery Team* sought to answer the three previously listed Research Questions in its experiments at the 2015 TREC Total Recall Track.

4.1 First and Primary Research Question.

What Recall, Precision and Effort levels will the e-Discovery Team attain in TREC test conditions over all 30 Topics using the Team's Predictive Coding 3.0 hybrid multimodal search methods and Kroll Ontrack's software, eDiscovery.com Review (EDR).

We primarily measured effort by the number of documents that were actually human-reviewed and coded relevant or irrelevant. The *Team human-reviewed* only 32,916 documents to classify 16,576,798 documents. As an additional measure of effort, we estimated our total time spent on all Topics. The *Team* spent 45 days doing all of the work, with an estimated average of 8 hours per day total expended by the *Team*. (All *Team* members carried on their normal employment activities on only a *somewhat reduced* basis during the 45 days of the review, and TREC work was also reduced on most weekends.) The estimated total hours spent by *Team* members for both analysis and review is thus approximately 360 hours.

It is typical in legal search to try to measure the efficiency of a document review by the number of documents classified in an hour. For instance, a typical contract review attorney can classify an average of 50 documents per hour. Here using *Predictive Coding 3.0* our Team classified 16,576,798 documents in 360 hours. That is an average speed of 46,047 files per hour.

In legal search it is also typical, indeed mandatory, to measure the costs of review and bill clients accordingly. If we here assume a high attorney hourly rate of $500 per hour, then the total cost of the review of all 30 Topics would be $180,000. That is a cost of less than $0.01 per document. In a traditional legal review, where a lawyer reviews one document at a time, the cost would be far higher. Even if you assume a low attorney rate of $50 per hour, and review speed of 50 files per hour, the total cost to review would be $16,576,798. That is a cost of $1.00 per document, which is actually low by legal search standards.[16]

Analysis of project duration is also very important in legal search. Instead of the 360 hours expended by our Team using *Predictive Coding 3.0*, traditional linear review would have taken 331,536 hours (16,576,798/50). In other words, what we did in 45 days, taking 360 hours, would have taken a team of two lawyers using traditional methods over 45 years.

Complete details and descriptions of the *ad hoc* methods employed in all thirty topics are included in the Appendix.

4.2 Research Question No. 2.

How will the Team's results using its semi-automated, supervised learning method compare with other Recall Track participants using semi automated supervised learning methods.

Unfortunately no other *Athome* participants completed all thirty topics and only one completed all ten Bush email topics. The lack of participation by others in the *Athome* group makes meaningful comparisons very difficult or impossible, but we note that the *e-Discovery Team's* scores were consistently higher than any other *Athome* participants.

The *Sandbox* participants' work included the same three datasets as *AtHome*, but none of them also participated in the *Athome* division. This is unfortunate because it makes direct comparisons problematic, if not impossible, especially as to the software systems used. Still, with some caveats, a few limited comparisons are possible between the two divisions because the same topics and datasets were searched.

4.3 Research Question No. 3.

What are the ideal ratios, if any, for relevant and irrelevant training examples to maximize effectiveness of active machine learning with EDR.

The Team experimented with various positive and negative training ratios using the predictive coding training features of their software. Most of these experiments were *post hoc*, but some were carried out during the initial TREC submissions. In some of the thirty topics our review work would have been concluded earlier but for these side experiments.

5. RESULTS

5.1 Research Question No. 1.

The TREC measured results demonstrated high levels of Recall and Precision with relatively little human review efforts using the *e-Discovery Team's* methods and *EDR*. The three-man attorney *Team* was able to review and classify 16,576,798 documents in 45 days under difficult TREC test conditions. They attained total Recall of all relevant documents in all 30 Topics by human review of only 32,916 documents. They did so with two-man attorney teams in the 10 Bush Email Topics, and one-attorney teams in the 20 other Topics. In Topic 3484, which searched a collection of 902,434 News Articles, the *Team* attained both 100% Recall and 100% Precision. On many other Topics the Team attained *near perfection* scores. In total, very high scores were recorded in 18 of the 30 topics with good results obtained in all, especially when considering the low human efforts involved in the supervised learning. Moreover, the *Team's* F1 scores at the time of *Reasonable Call* ranged from a perfect score of 100% in Topic 3484, to 91% to 99% in eight topics, and 82%-87% in five others.

Considering the limited human effort put into the reviews, and the speed of the reviews, we consider the results in all Topics to be excellent. As shown by the comparisons with traditional review discussed above, these results are far superior to the typical linear legal document review done by law firm attorneys and contract review attorneys.

The efforts by number of documents human reviewed in all thirty topics are shown in the below chart Figure 1. As you can see, the Team *reviewed* 32,916 documents to attain total recall of the 70,414 documents predetermined by TREC as relevant in all 30 Topics from out of a total of 16,576,798 documents. The average number of documents reviewed to attain total Recall in each topic was 1,097. The figure ranged from a low of 19 documents reviewed in Topic 2134 (PayPal), which had 252 relevant documents, to a high of 7,203 in Topic 103 (Manatee Protection), which had 5,725 relevant documents.

				Effort (Docs reviewed) by RECALL SCORES					
Topic	Need	Total Documents	Total Relevant	70%	80%	90%	95%	97.5%	100%
Topic 100	School and Preschool Funding	290,099	4,542	651	651	651	651	651	651
Topic 101	Judicial Selection	290,099	5,834	6,841	6,895	6,895	6,895	6,895	6,896
Topic 102	Capital Punishment	290,099	1,624	1,493	1,493	1,493	1,493	1,493	1,493
Topic 103	Manatee Protection	290,099	5,725	7,203	7,203	7,203	7,203	7,203	7,203
Topic 104	New Medical Schools	290,099	227	1,091	1,091	1,091	1,091	1,091	1,091
Topic 105	Affirmative Action	290,099	3,635	582	582	582	674	674	674
Topic 106	Terri Schiavo	290,099	17,135	831	1,987	1,995	2,005	2,025	2,226
Topic 107	Tort Reform	290,099	2,369	877	1,142	1,164	1,164	1,164	1,164
Topic 108	Manatee County	290,099	2,375	696	696	696	696	696	696
Topic 109	Scarlet Letter Law	290,099	506	491	496	639	753	753	753
Topic 2052	Paying for Amazon Book Reviews	465,147	265	1,842	1,960	2,213	2,325	2,325	2,325
Topic 2108	CAPTCHA Services	465,147	656	2,101	2,101	2,101	2,101	2,101	2,101
Topic 2129	Facebook Accounts	465,147	589	94	94	94	94	94	94
Topic 2130	Surely Bitcoins can be Used	465,147	2,299	283	283	285	285	285	285
Topic 2134	Paypal Accounts	465,147	252	19	19	19	19	19	19
Topic 2158	Using TOR for Anonymous Internet Browsing	465,147	1,261	1,332	1,332	1,332	1,332	1,332	1,335
Topic 2225	Rootkits	465,147	182	183	186	205	214	219	225
Topic 2322	Web Scraping	465,147	10,145	194	195	195	195	195	195
Topic 2333	Article Spinner Spinning	465,147	4,805	190	228	228	228	228	228
Topic 2461	Offshore Host Sites	465,147	179	32	32	32	32	32	32
Topic 3089	Pickton Murders	902,434	255	472	516	779	834	834	836
Topic 3133	Pacific Gateway	902,434	113	49	49	49	49	49	49
Topic 3226	Traffic Enforcement Cameras	902,434	2,094	18	18	18	78	81	81
Topic 3290	Rooster Turkey Chicken Nuisance	902,434	26	137	191	306	306	310	310
Topic 3357	Occupy Vancouver	902,434	629	751	751	920	920	920	920
Topic 3378	Rob McKenna Gubernatorial Candidate	902,434	66	79	161	200	200	200	200
Topic 3423	Rob Ford Cut the Waist	902,434	76	92	92	92	92	92	92
Topic 3431	Kingston Mills Lock Murders	902,434	1,111	272	272	272	272	272	302
Topic 3481	Fracking	902,434	1,966	31	236	367	367	367	367
Topic 3484	Paul and Cathy Lee Martin	902,434	23	22	22	22	22	73	73
Figure 1	TOTALS	16,576,800	70,964	28,949	30,974	32,138	32,590	32,673	32,916

The Team's attainment of high levels of Recall and Precision in multiple projects confirms the hypothesis that EDR software and the Team's *Predictive Coding 3.0 hybrid multimodal* methods are effective in most projects at attaining high levels of Recall and Precision with minimal human efforts.

The below charts summarize for each of the three datasets the Precision results obtained in each topic at 70% or higher Recall levels. Precision is shown on the left and Recall levels attained by submissions are shown on the bottom. A different colored line shows each Topic. Although Precision was not the focus of the efforts in the Team's Recall Track participation, instead the

focus was on Recall and effort, still the measurements of Precision across the Recall levels provide valuable insights into the overall work. Figure 2 below shows the results of the 10 Topics in *Jeb Bush Email* collection of 290,099 emails. Figure 3 shows the results of the 10 Topics in *BlackHat World Forum* collection of 465,149 posts, and Figure 4 shows the results of the *News Articles* collection of 902,434 articles.

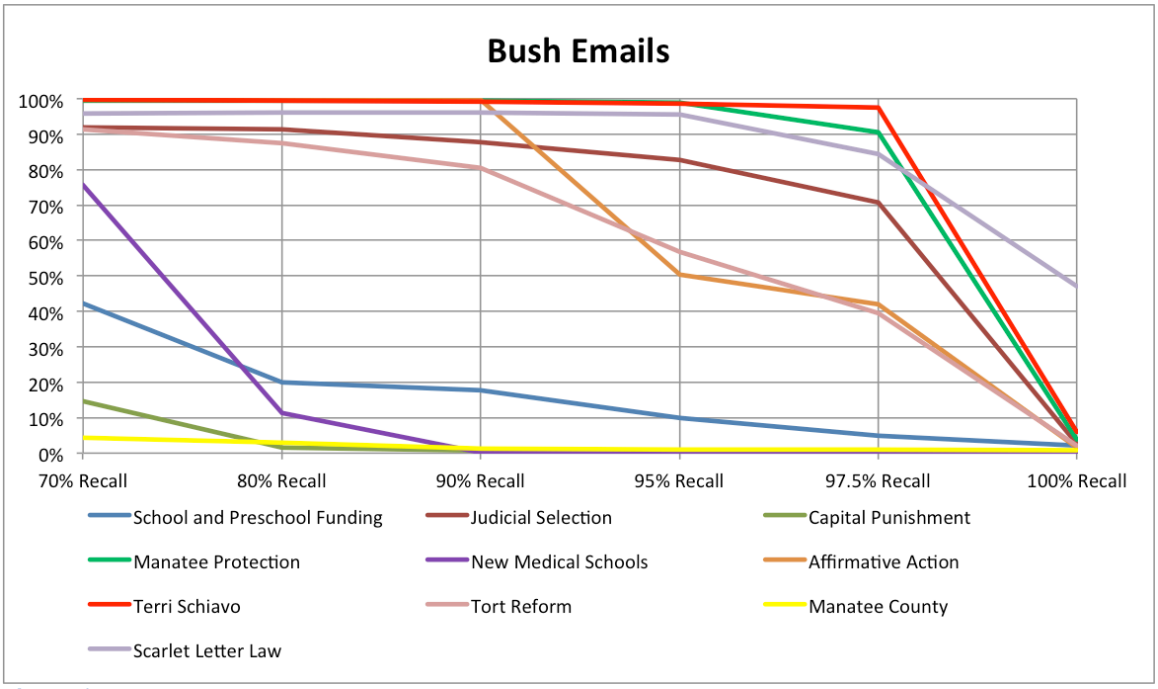


Figure 1

A quick exam of the results of the Bush Email Topics shows that four of the ten Topics had significantly less Precision in attaining 80% or higher Recall than the others. They are: Topic 104 New Medical Schools, shown in purple; Topic 100 School and Preschool Funding, shown in blue; Topic 102 Capital Punishment, shown in green; and, Topic 108 Manatee County. Topic 108 was probably the most error-filled of all of the Topic standards, and this may explain part of the outlier results for that topic and others in this low performing group. Investigation of the outliers showed that the primary cause of these results was disagreement by to the Team's lead attorney for the Bush email, a Florida life-long resident who is used to serving as the SME defining ground truth, and the TREC assessors' relevance determinations. Also, these ten Bush topics were carried out at the beginning of the project before the Team adopted mitigating counter strategies of greater reliance on machine ranking to mitigate the impact of the personal judgment disagreements.

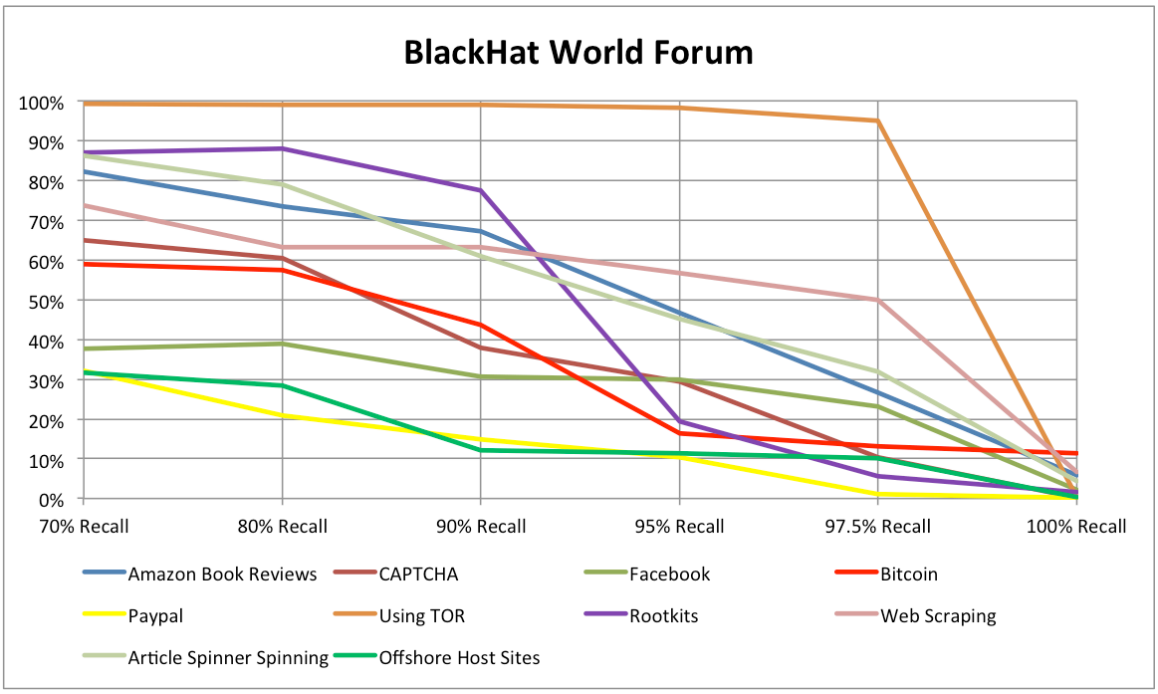


Figure 2

Analysis of the results of the ten Topics in *BlackHat World* also indicated that the relevance disagreements accounted for most of the discrepancies.

It appears that errors and inconsistencies in the TREC standard judging explain most of the Precision differences among the Topics, especially the Topics in the *BlackHat World* data set. In several of these Topics the Team often had difficulty detecting *any* logical pattern to the relevance scope. They instead, as mentioned, had to rely almost entirely on the EDR relevance predictions. Only the Team software in some of these Topics could detect any connectivity and pattern to the TREC relevant standards.

The results on the local News dataset of 902,434 articles (Figure 4 below) again shows significant divergences in Precision, although less than the differences seen in *Bush Email* or *BlackHat World* datasets. Analysis of the results of the ten *News Articles* Topics again shows considerable disagreement on relevance judgments in some topics. Inherent difficulty of the various issues in the Topics may also explain some of the differences. The size of the relevance pool also has a direct relationship on the Precision.

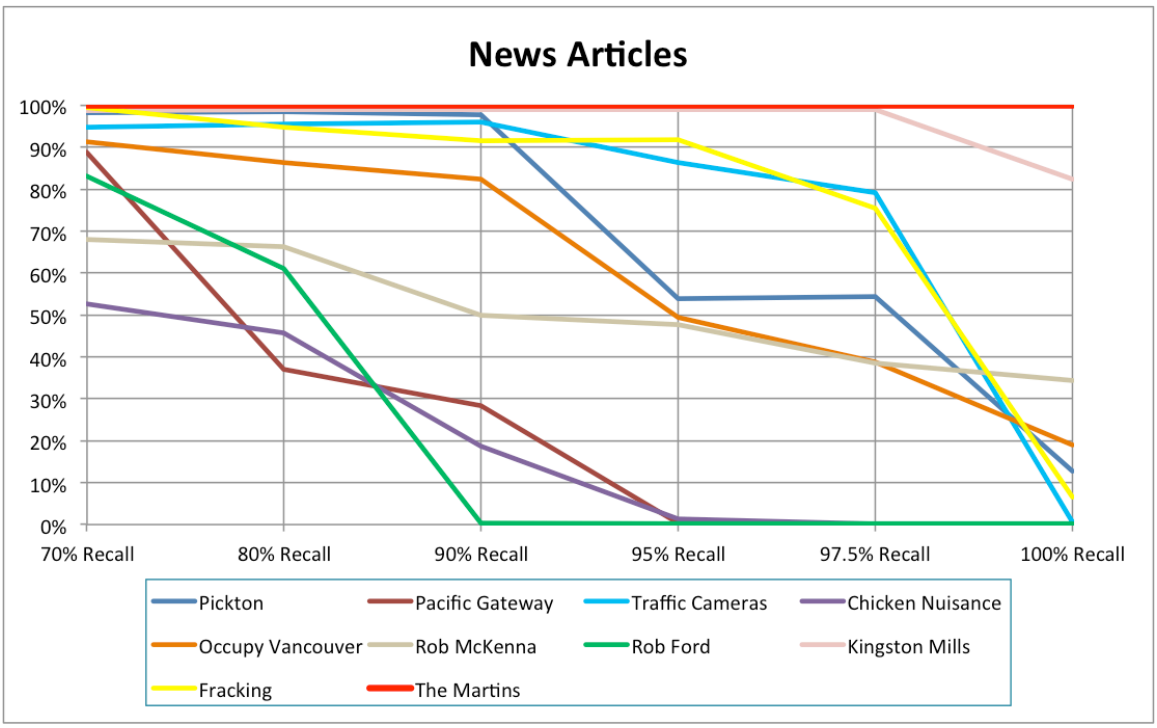


Figure 3

The following results are highlights of the Team's top 18 topics where at least seventy-five percent of the target documents (Recall 75%+) were found with a Precision rate of 80% or higher. The Top-18 Projects of the Team are ranked by us, somewhat arbitrarily, as follows, starting with a previously unheard of *perfect score*.

1. In Topic 3484 (Paul & Kathy Martin), the *e-Discovery Team* (Jim Sullivan) attained a perfect score of 100% Precision and 100% Recall. All 23 of the target documents were found in the first 23 documents submitted. Sullivan then *called Reasonable* after the 23rd relevant document was submitted and so played the *perfect game*. He predicted that the remaining 902,411 articles in the News collection would be irrelevant. Sullivan was right. The effort expended for perfection was his personal review of 73 news reports out of the total collection of 902,434. 100% Recall with 100% Precision in a large search project was previously thought impossible by most text retrieval experts.

2. In Topic 3431 (Kingston Mills Murders), 100% Recall was attained by the *Team* (Tony Reichenberger) with 82.3% Precision. He attained 97.5% Recall with a Precision of 98.9%, and 95% Recall with 99% Precision. The effort expended to reach 100% Recall was his personal review of 332 news reports out of the total collection of 902,434.

3. In Topic 106 (Terry Schaivo), which had the highest prevalence of any topic (5.9%), 98.47% Recall was attained by the *Team* (Ralph Losey) with 97.22% Precision. At that time, after submitting 2,025 documents, he called reasonable. The F1 measure then attained was 97.84%. The effort, or number of documents reviewed and coded by Losey to attain this result, was 2,025 Bush emails, out of the total collection of 290,099, and total relevant of 17,135. A contract review attorney, whose standard billing rate is one-tenth that of Losey's, assisted in the review effort. Losey also attained 99.7% Recall in this Topic with a Precision of 70%.

4. In Topic 2158 (Using TOR), the *Team* (Jim Sullivan) attained 97.5% Recall of the target while maintaining a Precision of 95%. He attained 95% Recall with a Precision of 98.4%, and 90%

Recall with 99% Precision. The effort expended to reach 97.5% Recall was his personal review of 1,332 BlackHat Forum posts, out of the total collection of 465,149.

5. Topic 103 (Manatee Protection), which had the third highest Prevalence of 1.97%, the *Team* (Ralph Losey) attained 97.5% Recall with a Precision of 90.6%, 95% Recall with a Precision of 98.8%, and 90% Recall with 99.3% Precision. The effort expended to reach 97.5% Recall was his personal review of 7,203 Bush emails, out of the total collection of 290,099. Again he was assisted by a contract review attorney. The high review count here is due to the fact this is one of two projects where the *Predictive Coding 3.0* second step of random sampling was included. This is also the first project undertaken.

6. In Topic 109 (Scarlett Letter Law), the *Team* (Ralph Losey) attained 97.5 % Recall with 84.4% Precision, 95% Recall with 95.4% Precision, and 90% Recall with 96% Precision. The effort expended to reach 97.5% Recall was his personal review of 753 Bush emails, again out of the total collection of 290,099. One contract review attorney assisted.

7. In Topic 3378 (Rob McKenna), the *Team* (Tony Reichenberger) attained 100% Recall after the submission of only 192 documents and review of only 200 documents. This was a low prevalence Topic with only 66 relevant out of the total collection of 902,434. For these reasons the Precision was 34.31%, even though only 192 documents were submitted to attain 100% Recall.

The *Team* results exceeded expectations, where our Recall goal was 90%, in many additional Topics:

8. In Topic 3481 (Fracking), the *Team* (Jim Sullivan) attained 95% Recall with 95.2% Precision by reviewing only 367 news articles.

9. In Topic 105 (Affirmative Action), the *Team* (Ralph Losey) attained 90% Recall with 99.7% Precision by reviewing only 582 mails (one contract review attorney assisted).

10. In Topic 3089 (Pickton Murders), the *Team* (Joe White) attained 90% Recall with 97.9% Precision by reviewing only 779 articles. A 99.61% Recall level was attained with 54.98% Precision, again with review of only 799 articles.

11. In Topic 3226 (Traffic Cameras), the *Team* (Jim Sullivan) attained 90% Recall with 95.9% Precision by his personal review only 18 forum posts.

12. In Topic 101 (Judicial Selection), which had the second highest Prevalence rate of 2%, the *Team* (Ralph Losey) attained 90% Recall with 87.8% Precision by reviewing 6,895 emails (one contract review attorney assisted).

13. In Topic 3357 (Occupy Vancouver), the *Team* (Tony Reichenberger) attained 90% Recall with 82.4% Precision by reviewing only 920 news articles.

14. In Topic 107 (Tort Reform), the *Team* (Ralph Losey) attained 90% Recall with 80.9% Precision by reviewing only 1,164 emails (one contract review attorney assisted).

Four additional Topics also did quite well, and attained Recall levels over 75% with high Precision rates:

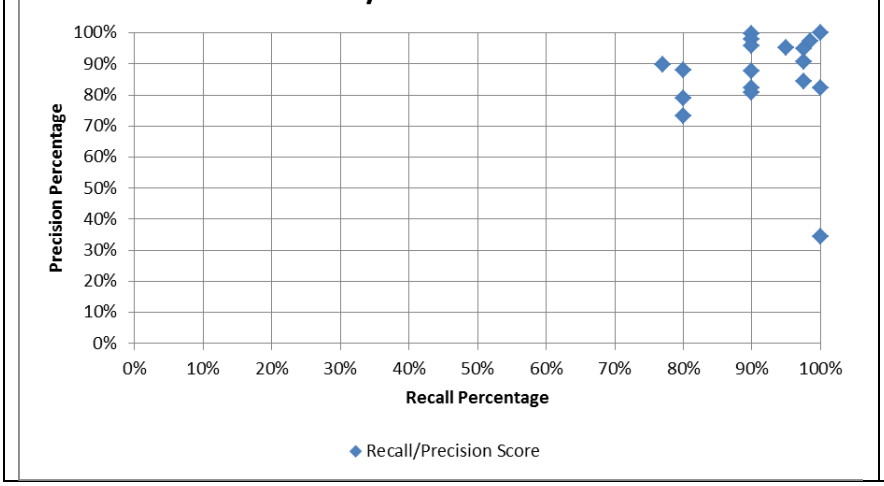
15. In Topic 2225 (Rootkits) the *Team* (Ralph Losey) attained 80% Recall with 88% Precision by reviewing only 186 forum posts.

16. In Topic 2333 (Article Spinner) the *Team* (Ralph Losey) attained 80% Recall with 79% Precision by reviewing only 228 forum posts.

17. In Topic 2052 (Paying for Book Reviews) the *Team* (Jim Sullivan) attained 80% Recall with 73.4% Precision) by reviewing 1,960 forum posts.

18. In Topic 3133 (Pacific Gateway) the *Team* (Ralph Losey) attained 76.99% Recall with 89.69% Precision by reviewing only 49 News Articles.

Figure 5 below shows the recall and precision of these top 18 projects.



Case	Recall	Precision
Topic 3484	100%	100%
Topic 3431	100%	82%
Topic 106	98%	97%
Topic 2158	98%	95%
Topic 103	98%	91%
Topic 109	98%	84%
Topic 3378	100%	34%
Topic 3481	95%	95%
Topic 105	90%	100%
Topic 3089	90%	98%
Topic 3226	90%	96%
Topic 101	90%	88%
Topic 3357	90%	82%
Topic 107	90%	81%
Topic 2225	80%	88%
Topic 2333	80%	79%
Topic 2052	80%	73%
Topic 3133	77%	90%

Figure 5

The Team's lower performance in the other 12 projects was, according to our analysis, primarily caused by the fact that the attorney *Team* members are accustomed to self-defining the ground truth, and their opinions on relevance differed significantly from the TREC assessors. In later topics the attorney *Team* learned to turn off their own judgments and rely primarily on their software's automated processes, at which point their scores improved. In all topics the machine learning of the Team's EDR software was able to find documents that TREC would consider relevant, even where the human team members could see no connection. But in some topics the human searchers would be completely bewildered by the *zig zag* relevance scope shown by TREC's response to submissions. The attorneys would not see any kind of logical connecting pattern to some of the documents that TREC determined to be relevant. Sometimes the attorneys only saw wrong answers and inconsistencies. Even though the attorneys could not see any pattern, they learned that their EDR software could often still find the patterns and correctly predict which documents TREC would label relevant. When this happened they would in effect turn all submission decisions over to EDR and only submit the highest-ranking documents. The cut-off point of ranking for submissions, be it top 5% or top 100 documents, or some other scheme, was still determined by the human in charge. That is part of the *Team's hybrid* design.

There are probably other explanations for the bottom twelve scoring topics aside from questionable TREC assessor adjudications, including: the data itself; the difficulty of the issues addressed in the Topic; relative performance of human reviewers; and, the impact of the omission of Steps Three and Seven from the *Team's* standard workflow to meet the 45 day time limitation, and the radical change to Step One. *See: Concept Drift and Consistency: Two Keys to Document Review Quality*, e-Discovery Team (Jan. 20, 2016). All of the *Team's* inconsistencies were not caused by differences of opinion on TREC relevance adjudications, only some. We appreciate the difficulty of creating interesting topics for such a diverse group of participants, most of whom used fully automated CAL approaches. We understand the inherent difficulties in setting a ground truth for prejudged relevance where the traditional TREC *pooling methods* could not be used.[17] In spite of our criticisms here, we overall have high praise and thanks for the TREC administrators' tireless efforts and agree with the majority of the assessments they made under difficult, time constrained conditions.

Regardless of these issues and metric inconsistencies, the *Team's* manual efforts, as measured by time expended and number of documents manually reviewed were consistently very low in all topics. More than half of the relevant documents found were not manually reviewed. Instead, the Team was routinely able to delegate relevance coding to the EDR software, either by choice and convenience, or sometimes, as discussed, by necessity in the topics where the ground truth of relevance was unknown and incomprehensible to the attorneys. This result should shatter once and for all the already weakened legal search *myth* that all documents must be manually reviewed for relevance.

Although not directly comparable due to different test conditions, different searches, etc., the *e-Discovery Team's* scores were far higher than any previously recorded in the six years of TREC Legal Track (2006-2011)[18] or any other study of legal search.[19] The results of Blair and Maron and TREC from 2007 to 2011 are summarized below in Figure 6 with F1 scores.

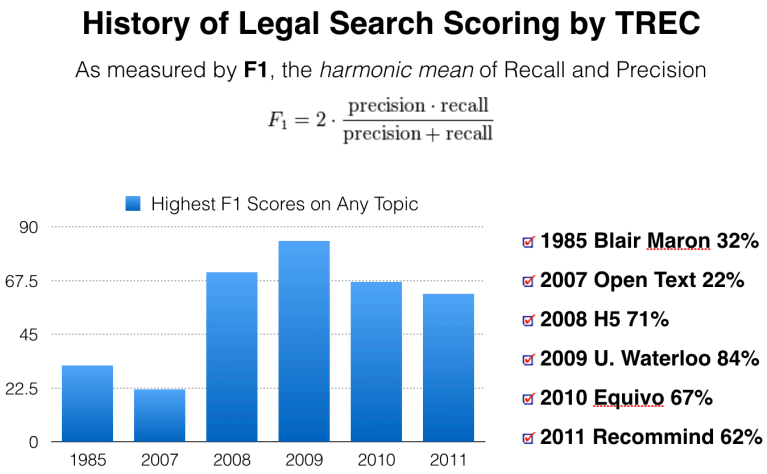


Figure 6

This is not a listing of the *average* score per year, such scores would be far, far lower. Rather this shows the very best effort attained by any participant in that year in any topic. These are the highest scores from each TREC year. Note how they compare with the Team's high scores in 2015, Figure 7.

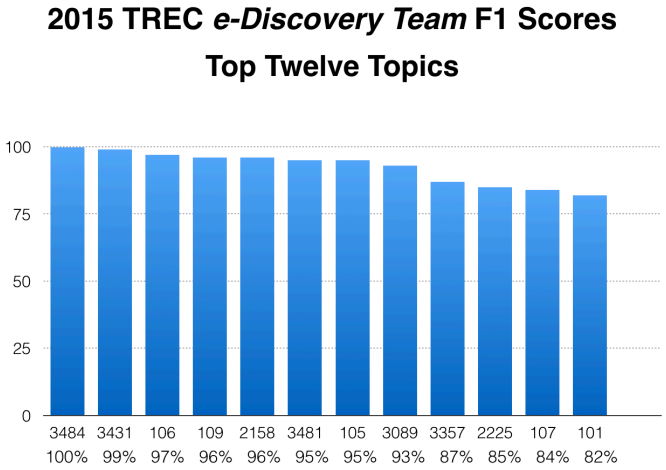


Figure 7

One reason for this significant jump in high scores may be that many of the thirty topics in the 2015 Total Recall Track presented relatively simple information needs by legal search standards, with one major exception, Topic 109 – Scarlet Letter Law. It required some legal knowledge and analysis. There were also four other minor exceptions – Topics 101, 105, 106, 107 – that required some measure of legal analysis. Another explanation may be improved software and the Team's *hybrid multimodal* method that includes continuous active learning. The later is strongly suggested because the results in Topic 109, as well as Topics 101, 105, 106 and 107, are close to typical legal search type projects and the Team's results in these topics were all consistently high: Topic 109 (Scarlett Letter Law) - 95% F1 at Reasonable Call; Topic 101 (Judicial Selection) - 87% F1 at Reasonable Call; Topic 105 (Affirmative Action) - 95% F1 at Reasonable Call; Topic 106 (Terri Schiavo)- 98% F1 at Reasonable Call; Topic 107 (Tort Reform) - 84% F1 at Reasonable Call. This is shown in Figure 8 below.

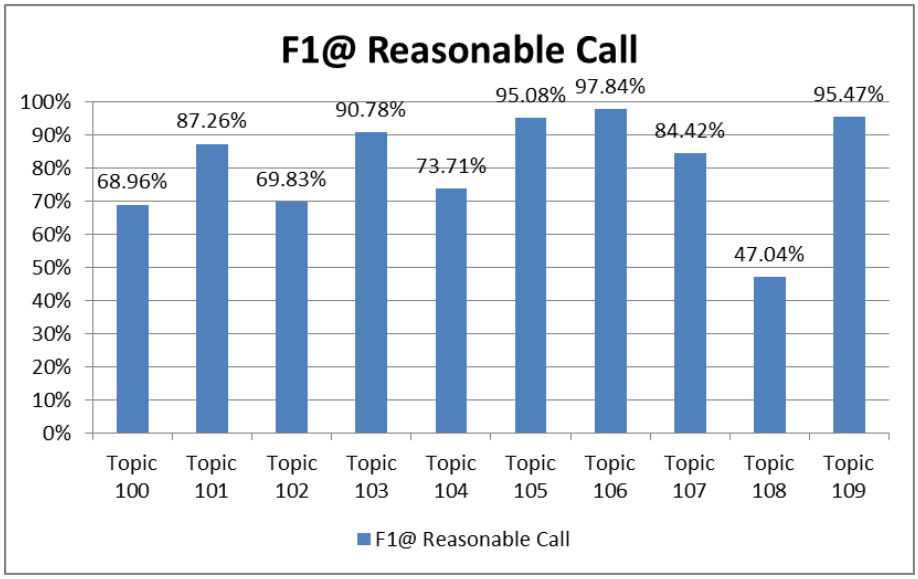


Figure 8

5.2 Research Question No. 2.

The *Team* attained very high recall and precision rates in most, but not all, of the thirty Total Recall topics. The Team's F1 scores at the time of *Reasonable Call* ranged from a perfect score of 100% in one topic (3484), to 91% to 99% in eight topics, and 82%-87% in five others.

Although, of course, not directly comparable, these scores are far higher than any previously recorded in the six years of TREC Legal Track (2006-2011) or any other study of legal search. One reason for this may be that the thirty topics in the 2015 Total Recall track presented relatively simple information needs by legal search standards, with one exception (Topic 109 – Scarlet Letter Law). Another may be improved software and the Team's *hybrid multimodal* method that includes continuous active learning.

Since most of the thirty topics presented only simple, single-issue information needs suitable for single-facet classification, they had somewhat limited value for purposes of legal search experimentation. Further, only a few of the topics required any legal analysis for relevance identification. This again limited the use of these experiments for purposes of legal search research. These two factors, plus the omission of metadata, was a disadvantage to the *e-Discovery Team* of lawyers who are practiced in more complex information needs requiring extensive legal analysis and SME defined ground truths. Further, their methods and *EDR*

software are designed to utilize full metadata derived from native files. Conversely, it appears that these same factors made it simpler for the *Sandbox* participants to perform well in most topics.

The one exception was Topic 109, Scarlett Letter Law, which, as mentioned, was the only topic requiring legal analysis and some very rudimentary knowledge to begin locating relevant documents. The keywords alone - "Scarlett Letter Law" – would only find relevant documents with this word combination and similar text patterns. These words were just the *nickname* of the proposed and eventually enacted Florida Statute. Any attorney would know that to find relevant information they would not only have to search the name, they would have to search the various house and senate bill numbers for this law. These numbers would not often appear in the same document as the nickname, and since the machine did not know to search for these numbers, it did not realize the significance. Eventually the automated machine learning saw the connection, after many relevance feedback submissions. These submissions would, of course, not happen in real legal search, and even if they did, this imprecision would equate to substantial additional human reviews and thus expense.

Somewhat surprisingly to us, the fully automatic methods employed by the *Sandbox* participants attained recall and precision scores comparable to that of the *e-Discovery Team* in most of the topics. Moreover, there were few differences between the various fully automated approaches. Still, the highest F1 values at the time of *Reasonable Call* were attained by the *e-Discovery Team* in twenty of the thirty topics, and the second or third best F1 scores in four others. This is shown in Figure 9 below. The *Team* F1 rankings for each topic are shown in the third column.

F1	pic 100	Rank	Topic 101	Rank	Topic 102	Rank	Topic 103	Rank	Topic 104	Rank	Topic 105	Rank	Topic 106	Rank	Topic 107	Rank	Topic 108	Rank	Topic 109	Rank
eDiscoveryTeam	68.96%	2	82.45%	4	69.88%	1	90.69%	1	73.53%	1	95.07%	1	97.38%	1	84.40%	1	47.03%	5	95.58%	1
NINJA	22.74%	8	79.17%	5	56.38%	5	83.79%	3	57.40%	4	77.24%	2	88.90%	5	50.89%	8	13.43%	11	48.79%	2
UvA.ILPS-baseline	73.55%	1	86.36%	1	56.38%	4	89.94%	2	10.27%	10	64.13%	5	95.87%	2	77.26%	4	64.47%	1	28.88%	3
UvA.ILPS-baseline2	45.56%	5	71.04%	7	42.42%	8	77.24%	6	2.42%	11	43.27%	7	84.67%	6	47.81%	9	35.13%	8	26.90%	6
WaterlooClarke-UWPAH1	11.95%	9	9.98%	11	32.16%	11	10.46%	11	68.51%	2	15.99%	10	3.61%	10	22.96%	11	21.61%	9	0.73%	8
WaterlooClarke-UWPAH2	10.37%	10	9.98%	10	32.16%	10	10.46%	10	65.93%	3	15.99%	11	3.54%	11	23.11%	10	21.54%	10	0.73%	9
WaterlooCormack-Knee100	45.02%	6	67.65%	9	42.32%	9	71.10%	9	28.49%	7	34.08%	8	77.03%	9	53.92%	7	42.65%	7	0.94%	7
WaterlooCormack-Knee1000	41.82%	7	67.67%	8	45.21%	7	71.11%	8	31.06%	5	33.90%	9	77.03%	8	57.79%	5	42.65%	6	27.17%	5
WaterlooCormack-stop2399	68.21%	3	72.02%	6	51.74%	6	75.55%	7	14.34%	9	58.92%	6	81.60%	7	57.77%	6	58.96%	2	27.17%	4
Webis-baseline	66.96%	4	83.87%	3	68.36%	2	82.42%	5	27.95%	8	64.91%	4	94.89%	4	79.24%	3	58.76%	3	0.00%	11
Webis-keyphrase	0.14%	11	85.21%	2	67.71%	3	83.15%	4	31.04%	6	65.13%	3	94.90%	3	79.24%	2	58.34%	4	0.33%	10

F1	Topic 2052	Rank	Topic 2108	Rank	Topic 2129	Rank	Topic 2130	Rank	Topic 2134	Rank	Topic 2158	Rank	Topic 2225	Rank	Topic 2322	Rank	Topic 2333	Rank	Topic 2461	Rank
eDiscoveryTeam	45.21%	1	53.99%	1	26.10%	6	64.31%	1	12.23%	6	95.61%	1	84.90%	1	72.60%	3	73.23%	1	16.68%	7
NINJA	58.13%	2	53.66%	2	49.22%	2	52.18%	2	39.70%	2	76.26%	2	39.43%	4	24.83%	9	62.65%	6	24.48%	5
UvA.ILPS-baseline	10.74%	3	22.74%	9	21.88%	7	41.12%	4	8.08%	7	42.02%	7	7.20%	9	73.20%	2	69.80%	2	7.33%	9
UvA.ILPS-baseline2	10.37%	4	22.45%	10	19.23%	8	30.88%	5	6.96%	8	22.47%	9	6.45%	10	48.11%	6	46.02%	9	6.53%	10
WaterlooClarke-UWPAH1	78.54%	5	52.20%	3	56.89%	1	13.42%	8	63.18%	1	40.08%	8	61.45%	2	5.85%	10	12.22%	10	49.90%	1
WaterlooCormack-Knee100	41.43%	6	33.89%	5	28.52%	5	19.49%	6	18.45%	3	16.15%	10	41.33%	3	47.39%	7	47.33%	7	43.87%	2
WaterlooCormack-Knee1000	38.10%	7	34.00%	4	30.91%	4	19.45%	7	18.45%	4	60.57%	5	27.02%	5	44.11%	8	47.30%	8	21.65%	6
WaterlooCormack-stop2399	16.94%	8	31.35%	7	31.01%	3	46.56%	3	15.51%	5	45.06%	6	11.84%	8	75.86%	1	68.87%	3	11.72%	8
Webis-baseline	13.24%	9	32.65%	6	7.73%	10	0.00%	10	2.21%	10	61.11%	4	18.36%	6	67.40%	5	68.07%	4	43.56%	3
Webis-keyphrase	10.53%	10	30.56%	8	8.29%	9	0.00%	9	2.21%	9	62.14%	3	12.97%	7	67.72%	4	68.04%	5	31.95%	4

F1	Topic 3089	Rank	Topic 3133	Rank	Topic 3226	Rank	Topic 3290	Rank	Topic 3357	Rank	Topic 3378	Rank	Topic 3423	Rank	Topic 3431	Rank	Topic 3481	Rank	Topic 3484	Rank
eDiscoveryTeam	93.28%	1	82.46%	1	55.39%	4	37.70%	2	86.70%	2	68.21%	1	58.12%	1	99.24%	1	95.48%	1	100.00%	1
NINJA	86.84%	2	67.97%	2	22.75%	9	38.98%	1	89.95%	1	67.88%	2	57.85%	2	74.67%	4	71.59%	2	100.00%	1
UvA.ILPS-baseline	5.47%	9	2.47%	9	37.25%	5	0.57%	9	12.75%	9	1.39%	9	1.26%	9	21.90%	7	35.00%	7	0.51%	9
UvA.ILPS-baseline2	5.35%	10	2.39%	10	34.75%	6	0.39%	10	11.82%	10	1.38%	10	0.74%	10	21.74%	8	29.19%	9	0.51%	10
WaterlooClarke-UWPAH1	76.14%	3	50.45%	3	24.73%	7	11.90%	5	62.65%	3	32.58%	4	18.65%	5	44.29%	6	26.87%	10	12.99%	6
WaterlooCormack-Knee100	57.66%	4	49.02%	4	64.61%	2	26.09%	3	55.57%	4	57.87%	3	30.70%	3	93.34%	3	53.62%	5	34.07%	4
WaterlooCormack-Knee1000	37.35%	5	18.38%	6	68.61%	1	4.59%	7	48.23%	5	11.26%	7	6.77%	7	93.77%	2	61.55%	4	4.07%	7
WaterlooCormack-stop2399	16.41%	7	8.43%	7	56.65%	3	2.01%	8	32.80%	6	5.01%	8	3.56%	8	44.78%	5	53.56%	6	1.78%	8
Webis-baseline	14.77%	8	47.06%	5	24.51%	8	19.31%	4	18.84%	7	27.37%	5	28.16%	4	19.71%	9	65.54%	3	34.59%	3
Webis-keyphrase	19.10%	6	6.40%	8	18.29%	10	10.22%	6	17.98%	8	18.23%	6	16.04%	6	19.19%	10	32.89%	8	30.08%	5

Figure 9

In Topic 109, Scarlet Letter Law, where some legal knowledge and analysis was required to understand relevance, the *Team* attained significantly better results - 96% F1 - at the time of Reasonable Call than did the automatic runs. In the *Sandbox* automatic runs the F1 values at the time of Reasonable Call ranged from 0% to 29%. Moreover, at the 1R point in Topic 109, the *e-*

Discovery Team had attained over 95% recall, whereas all of the automated methods were still less than 1% recall. This is shown in the chart below, Figure 10.

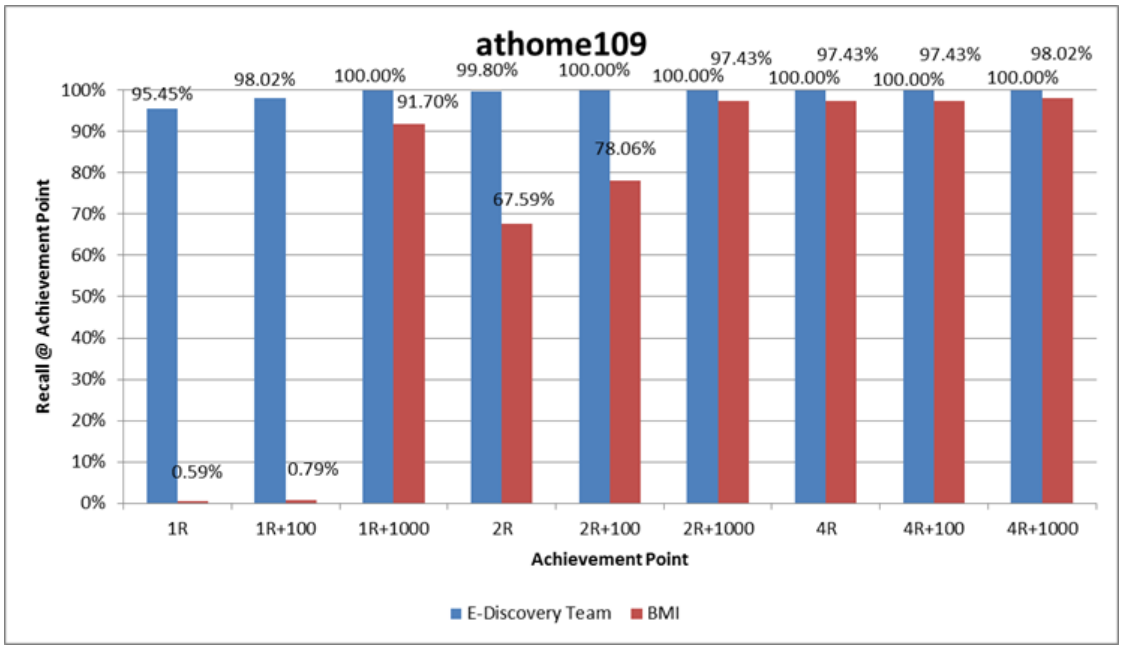


Figure 10

The *Team's* multimodal human machine approach also consistently found more relevant documents at the start of a search, and did so with greater precision than the fully automated approaches. Further, the *hybrid man-machine* approach was consistently more effective at determining a *stop point*, referred to by the Recall Track as a "*Reasonable Call*." An example of this is shown in the Figure 11 for Topic 109. The dark green line represents the Reasonable Call point, recall is shown in the vertical, and horizontal is the number of documents submitted.

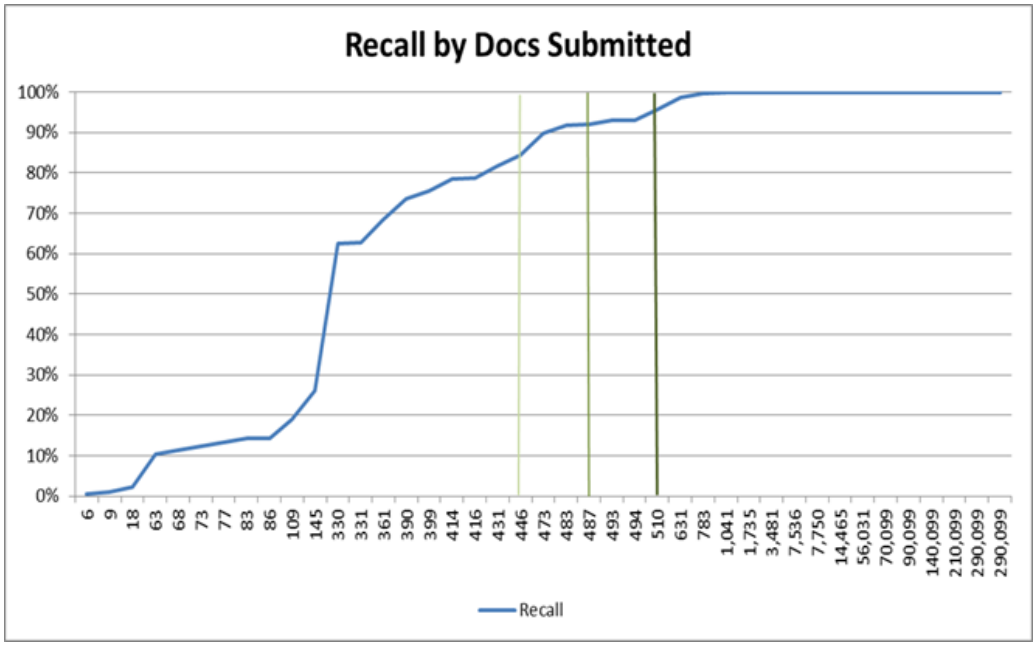


Figure 11

Another way to evaluate the performance of the multi-modal approach is to consider how precise the coding suggestions were during the course of review. This would indicate an efficient review, which is critical in legal search to cost savings. As to the *Athome109* topic, the below Figure 12 contrasts precision percentage on the Y-axis, with recall percentage on the X-axis. Precision does not begin to drop until approximately 95% Recall. Note that the green line representing percent of the database submitted barely moves off the baseline. Figure 13 shows the actual document counts reviewed and submitted in order to obtain the various precision thresholds.

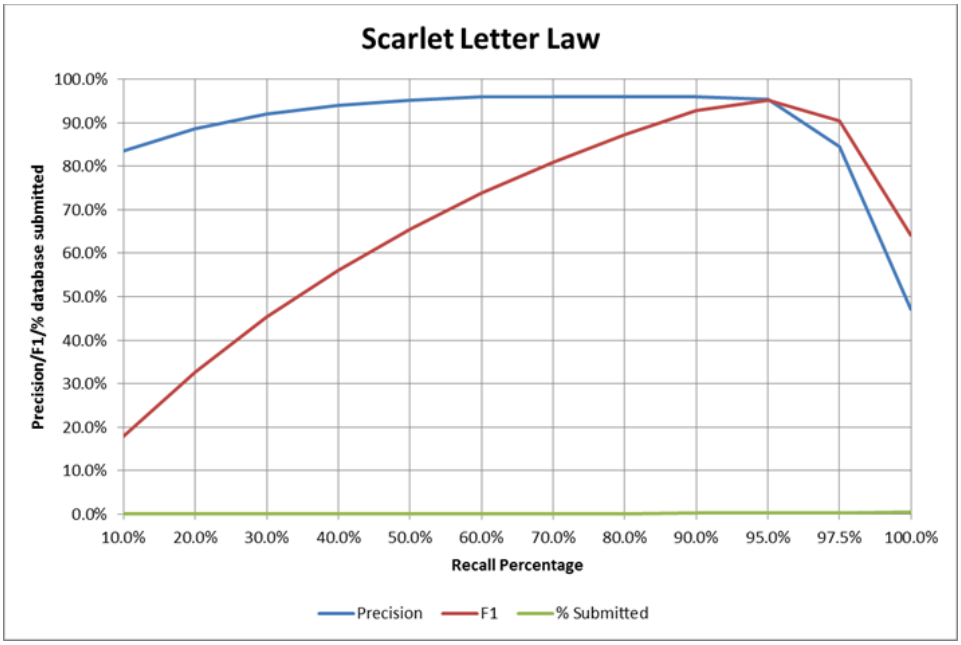


Figure 12

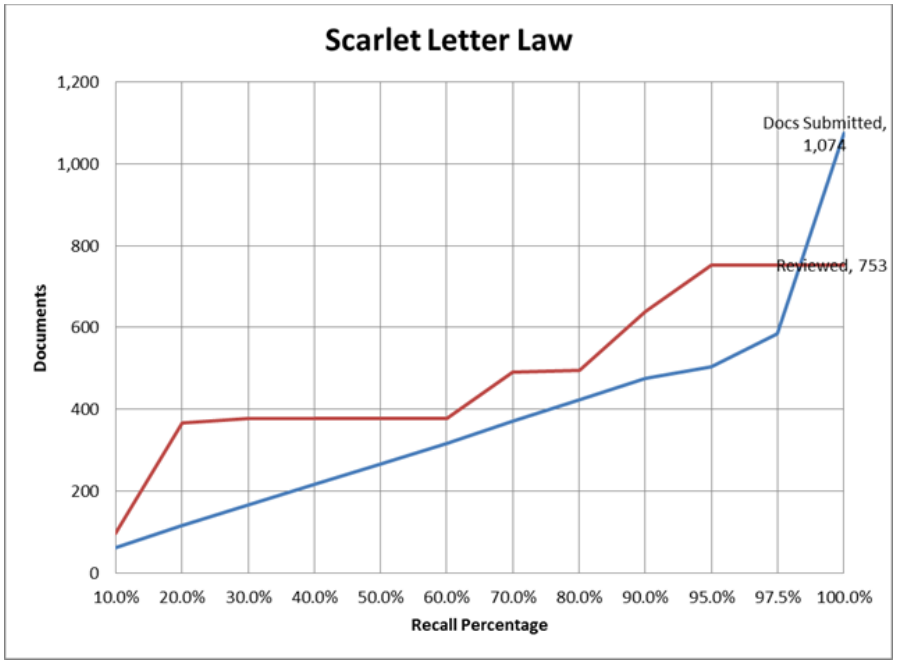


Figure 13

For further comparison Figure 14 below (prepared by the Total Recall administrators) plots the average *Athome3* precision by recall results. The *e-Discovery Team* results (barely visible on top) follow a curve very similar to the Athome109 topic. The Team's results outperformed the automated runs for most of the duration of the process, demonstrating a consistent efficiency in results. While various automated runs experienced comparable results in the Athome1 and Athome2 sets, the consistently high level of the multimodal approach corroborates a consistent efficient process across all data sets.

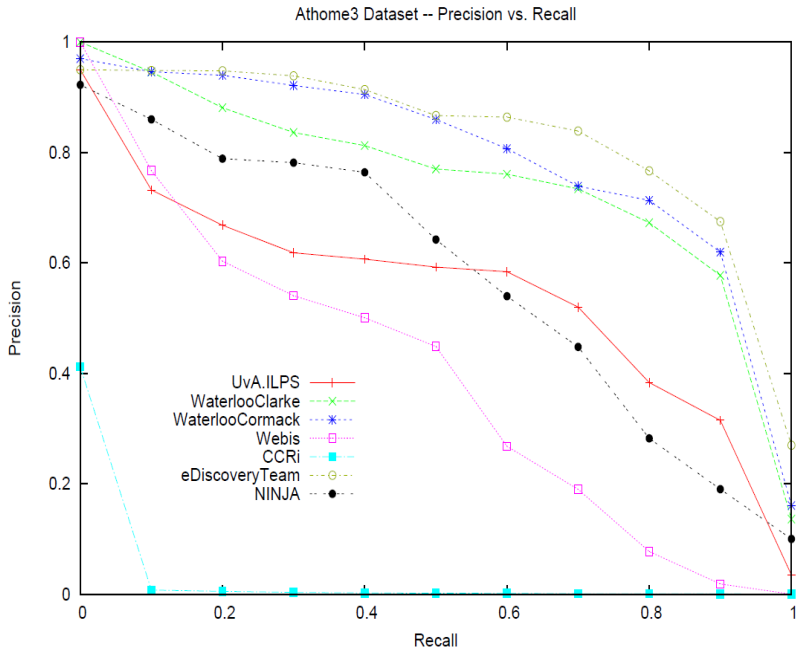


Figure 14

5.3 Research Question No. 3.

The Team's experiments with different positive negative training ratios showed that training using a 50/50 ratio of relevant to irrelevant documents performed consistently better than any other ratios. This result is believed to be specific to the proprietary type of logistic regression algorithm used in Kroll Ontrack's EDR. It may not have applications beyond this software, or even other more complex projects. Our work on this question continues.

6. CONCLUSIONS

The results in Topic 109 and other topics indicate that *hybrid* man-machine learning by skilled attorneys is, at the current time, significantly more effective at meeting complex legal search needs than fully automated approaches. This seems obvious, but more experiments on this issue are needed before this can be accurately quantified. The surprising success of the *Sandbox* participants using fully automated search, even though limited to non-legal topics and situations with only simple information needs, suggests that greater reliance on automated methods could be placed in legal search where the cases and needs are simple. The relatively low effort involved in automated learning, and thus low expense, is compelling, especially in view of the proportionality analysis required by law under the December 2015 Amendments to the *Federal Rules of Civil Procedure*. The *Team* has begun and will continue *post hoc* analysis and experiments using various hybrid methods that adjust the balance between man and machine.

We are experimenting with methods that place greater reliance on machine learning in all topics, including, but not limited to, topics with lesser complexity and information needs. We will also further investigate the use of both fully automated methods, and hybrid methods, in legal search quality control, fraud detection, and in the prediction of future wrongful conduct.[20]

The 2015 TREC Total Recall Track results also suggest that even when information needs are simple and require no complex analysis or background knowledge, as was true of most of the topics, that a hybrid method outperforms fully automated methods in two ways: one, at finding relevant documents quickly and with high precision; and two, at making better *stop decisions*. These two considerations are very important in legal search where attorneys must find a proportional balance between recall and effort/expense. The results in all topics, even the simple ones, thus caution against over-reliance at this time on machine learning alone without proper expert supervision.

7. ACKNOWLEDGMENTS

The *e-Discovery Team* would like to thank Kroll Ontrack, Inc. and Jackson Lewis P.C. for their generous support of this project. We would also like to thank the many employees at Kroll Ontrack who pitched in behind the scenes, often late at night and on weekends, to help make this happen.

8. REFERENCES (Endnotes)

[1] Losey, R., Predictive Coding 3.0, part two (e-Discovery Team, 10/18/15); also see *Predictive Coding Articles by Ralph Losey,* (collection of over 50 articles by Ralph Losey further describing the hybrid multimodal approach).

[2] The *e-Discovery Team's* hybrid multimodal approach is similar to the method promoted by the Total Recall Track administrators, Maura Grossman and Gordon Cormack, in that they both use continuous active learning (CAL) in legal search as part of a technology-assisted review (TAR). It is, however, fundamentally different from Grossman and Cormack's current methods in two ways.

First, our approach relies upon and encourages participation of skilled reviewers in the search process, the *hybrid* approach, whereas the Grossman and Cormack approach seeks to eliminate the role of the skilled user, namely trained attorneys. The rationale for their automation goal is the unsubstantiated claim that the adversarial context of legal search makes attorneys untrustworthy. They claim that inherent user bias means fully automated approaches are the only reliable methods of legal search. Grossman & Cormack, *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*, CoRR abs/1504.06868 at pg. 1 (2015) ("*In eDiscovery, the review is typically conducted in an adversarial context, which may offer the reviewer limited incentive to conduct the best possible search*.") Obviously the Team disputes this assumption and conclusion. We do not endorse the view of the inherent bias and untrustworthiness of attorneys. In Ralph Losey's experience as a practicing attorney since 1980 such bias is the rare exception, not the norm, and should not be the basis of a legal search strategy. The better solution to this minor issue of trustworthiness is educational, to train more attorneys in search and in professional ethics. Since our core assumptions on process and attorney honesty are fundamentally different, so too are our methods and goal. Our aim is *augmentation* of skilled attorneys to perform legal search, not *automation*, not replacement.

Second, our Team uses a variety of search methods, a *multimodal* approach, whereas the Grossman and Cormack approach relies solely upon the use of high-ranking

documents to train a classifier. This is consistent with their aim to fully automate and eliminate attorneys from the legal search process, again based on the premise we dispute of attorney bias. In their words: "*For the reasons stated above, it may be desirable to limit discretionary choices in the selection of search tools, tuning parameters, and search strategy*." *Id.* We disagree and seek to empower attorneys with a variety of search tools, including the one search method that they endorse of reliance on high-ranking documents. *Also see* and the discussion and citations in Endnote 19.

[3] In these respects the e-Discovery Team follows the teachings of Gary Marchionini, Dean of the School of Information and Library Sciences of U.N.C. at Chapel Hill, who explained in *Information Seeking in Electronic Environments* (Cambridge 1995) that information seeking expertise is a critical skill for successful search. Professor Marchionini argues, and we agree, that: "*One goal of human-computer interaction research is to apply computing power to amplify and augment these human abilities*." We also follow the teachings of UCLA Professor Marcia J. Bates who has advocated for a multimodal approach to search since 1989. Bates, Marcia J., *The Design of Browsing and Berrypicking Techniques for the Online Search Interface*, Online Review 13 (October 1989): 407-424. As Professor Bates explained in 2011 in *Quora*:

> "*An important thing we learned early on is that successful searching requires what I called "berrypicking." … Berrypicking involves 1) searching many different places/sources, 2) using different search techniques in different places, and 3) changing your search goal as you go along and learn things along the way. This may seem fairly obvious when stated this way, but, in fact, many searchers erroneously think they will find everything they want in just one place, and second, many information systems have been designed to permit only one kind of searching, and inhibit the searcher from using the more effective berrypicking technique.*"

Also see: White & Roth, *Exploratory Search: Beyond the Query-Response Paradigm* (Morgan & Claypool, 2009).

[4] The Total Recall Track fully automated method follows the Track Administrator's preferred methodology of fully automated monomodal search (high ranking only) and their recently announced goal to eliminate attorney review in favor of full automation. Grossman & Cormack, *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*, supra at pg. 1 (2015):

> "*Our goal is to fully automate these choices, so that the only input required from the reviewer is, at the outset, a short query, topic description, or single relevant document, followed by an assessment of relevance for each document, as it is retrieved*."

They call the method "*Autonomous TAR*." *Id.* at pg. 6. The protocols of the fully automated division of the Total Recall Track were apparently designed in part by Cormack and Grossman to test this premise, and the results they attained as participants in this division, along with all of the other fully automated participants from Universities around the world, are very impressive. Still, the e-Discovery Team, who did not participate in the 2015 automated division, notes that many of the protocols in this experiment are based on fictions and conditions not found in the real world of legal search, where the Team's methods were developed. The differences include, but are not limited to: the existence of an omnipotent SME that instantly provides perfectly correct judgmental feedback as to relevance of all documents selected by the automated processes as probable relevant; simple, single-facet issues; relatively simple datasets stripped of most native metadata; and, perhaps most importantly, issues

requiring little or no legal analysis or background legal knowledge. Note, in *post hoc* runs the e-Discovery Team ran a few fully automated runs on Kroll Ontrack systems and EDR. We used the same high ranking only *Autonomous TAR* training method and obtained the same results as all of the other fully automated division participants.

[5] "Contract review attorney," or simply "contract attorney," is a term now in common parlance in the legal profession to refer to licensed attorneys who do document review on a project-by-project basis. Their pay under a project contract is usually by the hour and is at a far lower rate than attorneys in a law firm, typically only $50 to $75 per hour. Their only responsibility is to review documents under the direct supervision of law firm attorneys who have much higher billing rates.

[6] *Predictive Coding* is defined by The Grossman-Cormack Glossary of Technology-Assisted Review, 2013 Fed. Cts. L. Rev. 7 (January 2013) (*Grossman-Cormack Glossary*) as: "An industry-specific term generally used to describe a Technology Assisted Review process involving the use of a Machine Learning Algorithm to distinguish Relevant from Non-Relevant Documents, based on Subject Matter Expert(s) Coding of a Training Set of Documents. " A Technology Assisted Review process is defined as: "A process for Prioritizing or Coding a Collection of electronic Documents using a computerized system that harnesses human judgments of one or more Subject Matter Expert(s) on a smaller set of Documents and then extrapolates those judgments to the remaining Document Collection. … TAR processes generally incorporate Statistical Models and/or Sampling techniques to guide the process and to measure overall system effectiveness." *Also see*: Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, Richmond Journal of Law and Technology, Vol. XVII, Issue 3, Article 11 (2011).

[7] Da Silva Moore v. Publicis Groupe 868 F. Supp. 2d 137 (SDNY 2012) and numerous cases later citing to and following this landmark decision by Judge Andrew Peck, including Judge Peck's own more recent *Rio Tinto v. Vale,* 2015 WL 872294 (March 2, 2015, SDNY).

[8] Grossman & Cormack, Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery, SIGIR'14, July 6–11, 2014; Grossman & Cormack, Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review", 7 Federal Courts Law Review 286 (2014); Herbert Roitblat, series of five OrcaTec blog posts (1, 2, 3, 4, 5), May-August 2014; Herbert Roitblat, Daubert, Rule 26(g) and the eDiscovery Turkey OrcaTec blog, August 11th, 2014; Hickman & Schieneman, The Implications of Rule 26(g) on the Use of Technology-Assisted Review, 7 FED. CTS. L. REV. 239 (2013); Losey, R. Predictive Coding 3.0, part one (e-Discovery Team 10/11/15).

[9] *Id*.; Webber, Random vs active selection of training examples in e-discovery (Evaluating e-Discovery blog, 7/14/14).

[10] *See* Endnote [2]. This disagreement is within a general framework of agreement on the superiority of computer assisted methods over traditional linear review, joint criticism of random selection methods and control sets in legal review, and agreement on the use of continuous active learning, as opposed to *one and done*, identified by Losey as *Predictive Coding Version 1.0*. Predictive Coding 3.0, part one (e-Discovery Team 10/11/15).

[11] Grossman & Cormack, *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*, CoRR abs/1504.06868 (2015); *Multi-Faceted Recall of*

Continuous Active Learning for Technology-Assisted Review, SIGIR'15, August 09-13, 2015, Santiago, Chile. (2015).

[12] Losey, R., Predictive Coding 3.0, part two (e-Discovery Team, 10/18/15).

[13] Shakespeare, W., *Henry VI, Pt II*, Act 4, Scene 2, 71-78 ("*The first thing we do, let's kill all the lawyers*."). This famous anti-lawyer line was spoken by "Dick the butcher," a traitor hoping to start a revolution and prop up his friend as an autocratic ruler.

[14] Losey, R., *Predictive Coding 3.0, part one)* (2015 e-Discovery Team), see the subsection therein, *Predictive Coding 1.0 and the First Patents*, discussing common prejudice against lawyers by academics and IT that drove the ill-advised imposition of *secret control* sets in the first versions of predictive coding software. The new drive by Cormack and Grossman to fully automate legal search and eliminate SMEs and attorney search expertise from legal search seems based, at least in part, on the same false premises. *Also see* Losey, R., *Mancia v. Mayflower Begins a Pilgrimage to the New World of Cooperation,* 10 Sedona Conf. J. 377 (2009 Supp.); Losey, R., *Lawyers Behaving Badly*, 60 Mercer L. Rev. 983 (Spring 2009).

[15] *See Zero Error Numerics* for a partial list of quality control and quality assurance methods endorsed by the *e-Discovery Team*, found at ZeroErrorNumerics.com (ZEN Document Review). *Also see: Concept Drift and Consistency: Two Keys to Document Review Quality*, e-Discovery Team (Jan. 20, 2016).

[16] The cost of traditional linear document review is often far higher than $1.00 per file in practice. In 2007 the U.S. Department of Justice spent $9.09 per document for review in the *Fannie Mae* case, even though it used contract lawyers for the review work. *In re Fannie Mae Securities Litig*., 552 F.3d 814, 817 (D.C. Cir. 2009) ($6,000,000/660,000 emails). At about the same time Verizon paid $6.09 per document for a massive *second review* project that enjoyed large economies of scale and, again, utilized contract review lawyers. 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 ($14,000,000 to review 2.3 million documents in four months).

[17] E. M. Voorhees, *Variations in relevance judgments and the measurement of retrieval Effectiveness*, Information Processing & Management, 36(5):697{716, 2000 (on pooling); Oard, Baron, Hedlin, lewis, Tomlinson, *Evaluation of Information Retrieval for E-Discovery*, Journal Artificial Intelligence and Law, Vol. 18 Issue 4, December 2010 Pgs. 347-386.

[18] *Autonomy and Reliability, supra* at pgs. 2-3 ("*This paper offers a historical review of research efforts to achieve high recall* ..." The paper also estimates the Blair Maron precision score of 20% and lists the top scores (without attribution) in most TREC years); Hedin, Tomlinson, Baron, and Oard, Overview of the TREC 2009 Legal Track (TREC 2009); Cormack, Grossman, Hedin, and Oard; *Overview of the TREC 2010 Legal Track* (TREC 2010); Grossman, Cormack, Hedin, and Oard, *Overview of the TREC 2011 Legal Track* (TREC 2011); *Evaluation of Information Retrieval for E-Discovery*, *supra* at pgs. 24-27. The top TREC results cited for the six years of Legal track are in the 60% to 70% F1 range with a couple of results in the low 80% F1 range. The Recommind participation in the last TREC Legal Track 2011, and their subsequent prohibited marketing advertisements claiming to "win," which led to their lifetime ban from TREC, only attained a Recall of 62.3% in one topic (403). *Overview of the TREC 2011 Legal Track* (TREC 2011) *supra*. Contrast all of the prior TREC results with the *e-Discovery Team* results in 18 topics in the 80% to 100% F1 range, with numerous topics in the mid to high 90% F1 range. Of

course, these different TREC events had varying experiments and test conditions and so direct comparisons between TREC studies are never valid, but general comparisons are instructive and frequently made in the cited literature.

[19] *See* the report on the Electronic Discovery Institute (EDI) Oracle legal search experiments involving the largest number of legal search participants to date where a member of the *e-Discovery Tea*m attained high scores. Bay, M., *EDI-Oracle Study: Humans Are Still Essential in E-Discovery: Phase I of the study shows that older lawyers still have e-discovery chops and you don't want to turn EDD over to robots* (11/20/13, LTN). Monica Bay, the Editor of *Law Technology News,* summarizes the conclusion of EDI from the study that: "*Conclusion: Software is only as good as its operators. Human contribution is the most significant element*." Patrick Oot, co-founder of the Electronic Discovery Institute presented the findings of Phase II of the Oracle Predictive Coding Survey at ILTACON Day 3, as reported in The Relativity Blog, 9/2/15: *"[W]hen it comes to what some vendors call Continuous Active Learning, Oot indicated the debate was somewhat of a red herring, adding, "Continuous Active Learning is just a buzzword." Oot summed up his thoughts by stressing the human component of technology-assisted review. Noting that the best performing technology in the Oracle study was the one used by a senior attorney, Oot said, "A good artist with a good brush is best."* Unfortunately the final results of the EDI Oracle study have not yet been published and, as participants in that study, we are currently constrained from any detailed reporting.

[20] *See* PreSuit.com where the *e-Discovery Team's* proposal is outlined to monitor the IT systems of large organizations with advanced analytics and other search methods to predict and avoid future illegal conduct. This man-machine hybrid type of early warning system includes safeguards to protect both individual privacy rights and confidential corporate information.

APPENDIX

E-Discovery Team *89-Page Narrative Report* of all 30 Topics

This Appendix Narrative Report describes the search of all thirty Total Recall topics in TREC 2015 using the *e-Discovery Team's* Hybrid Multimodal method. The report follows the chronological order in which the searches were conducted. The first project started on July 14, 2015. It was *Topic 103 Manatee Protection*. The last *Topic 3089 Pickton Murders* concluded on August 28, 2015. At the beginning of each Topic the results are reported for that Topic. Each has the same form and discloses metrics at the times when: (1) the Reasonable call was made; and, (2) the point where 97.5% Recall was attained. They are summarized along with a variation of a standard *Confusion Matrix*, a/k/a *Contingency Table* [1] The Confusion Matrix itself is highlighted in blue. It is followed by a list of the key the values attained: **Recall, Precision, F1 Measure, Accuracy, Error, Elusion and Fallout**.

Work on multiple topics was conducted at the same time. Sullivan, who worked on eight topics, Reichenberger, who worked on four, and White, who did one, each worked on a single topic at a time. They did, however, work concurrently with Losey and each other. Losey, who worked on seventeen topics, and had the assistance of a contract review attorney on the ten Bush Email Topics, typically worked concurrently on multiple topics at the same time. All Topics were a Team effort, but the attorneys identified as *running* each Topic controlled the review work for that Topic. Consultation was common, especially at first.

Topic 103 Manatee Protection

Confusion Matrix - Topic 103
Total Documents: 290,099
Total Relevant: 5,725
Total Prevalence: 1.97%

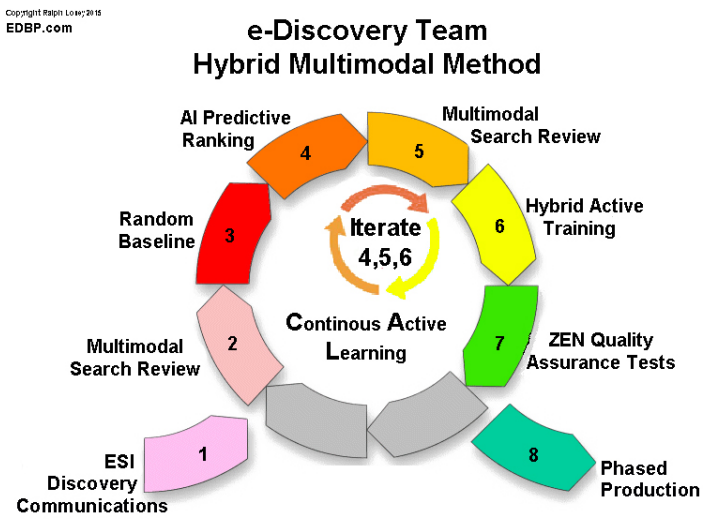
	@Reas. Call	@97.5% Recall
True Positives	4,780	5,582
True Negatives	284,348	283,793
False Positives	26	581
False Negatives	945	143
Recall	83.49%	97.50%
Precision	99.46%	90.57%
F1 Measure	90.78%	93.91%
Accuracy	99.67%	99.75%
Error	0.33%	0.25%
Elusion	0.33%	0.05%
Fallout	0.01%	0.20%

[1] *Grossman & Cormack Glossary, supra* FN 1 at pg. 6. The *Confusion Matrix* is also referred to as a *Contingency Table*.

The e-Discovery Team's TREC Total Recall project commenced on July 14, 2015 with work on Topic 103 Manatee Protection. This topic was run by Losey. He did not complete work until July 22, 2015. Although it may seem fast to see a review of 290,099 documents completed by one attorney in only eight days (with no breaks), there was more time spent on this topic than any of the others. But a significant amount of this time was spent on general set-up, procedures, contract reviewer training, project orientation, and communication protocols. Completion of this Topic was also delayed due to the availability of the contract review attorney, Anne Bottolene, who assisted Losey for the first part of the work on Topic 103, and due to some initial software configuration setup issues.

The Team found this Topic challenging for a variety of reasons, including the fact that the Bush collection of 290,099 emails had been stripped of its original metadata, images, and attachments. Further, we found some inconsistencies in judging this topic, although not many. Overall we found Topic 103 had one of the best *gold-standards* of the ten Bush Email Topics.

Ralph Losey is a native Floridian and Florida attorney for 35 years. He was somewhat knowledgeable about all of the Bush Email issues, certainly far more so than the average person, but he did not consider himself a *bona fide* subject matter expert (SME) on any of them. Losey's knowledge and interest on Manatee Protection issues was, however, higher than the other Bush Topics. For that reason it was chosen as the first topic. Losey's assistant, Bottolene, had lived in Florida for several years and also had some background with the Manatee Protection issue. They generally considered their familiarity with the issue to be an asset in the search of Topic 103. The same cannot be said of other Bush Email Topics.



The project commenced after initial orientation on July 14, 2015 with Losey beginning Step Two, *Multimodal Search Reviews*. Bottolene was assigned Step Three, Random Baseline. Due to various scheduling and implementation issues Bottolene did not complete her review of the sample until July 20, 2015, late afternoon. She reviewed and coded as either relevant or irrelevant a random sample of 1,534 Bush emails. This was one of only two Topics wherein Step Three was followed and a full random sample was taken. It proved very helpful.

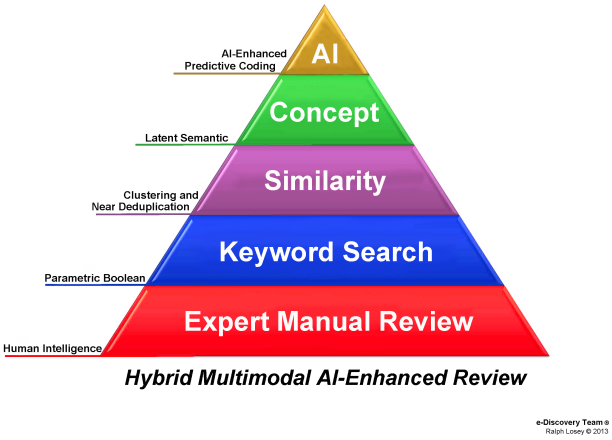
Based on the sample prevalence we predicted a spot projection for prevalence in Topic 103 of 5,175 documents (95% +/- 2.5% confidence levels). In fact, the total relevant documents in Topic 103 proved to be 5,725, well within the 2.5% margin of error. Based on the length of time needed for random sample review, and our desire to complete all thirty topics in 45 days, we decided to skip this step for ensuing reviews. (*Topic 101 Judicial Selection* was started shortly

after Topic 103, and also included Step Three Random Baseline.) As mentioned, we also skipped most of the procedures in Step 7- *" Zero Error Numerics"* concerning quality control in this and all 30 Topics.

After Bottolene completed the random sample review on July 20th she assisted Losey on July 21st and 22nd in his work on Step Five Multimodal Search Review. At that time submission to TREC had already begun and the Team was evaluating the confirmed relevant and irrelevant documents from TREC.

A total of 24 document submissions were made to TREC in this Topic: four document submissions on July 20th, one of July 21st, and the remaining nineteen submissions were made on July 22, 2015. In between most of these submissions the Team conducted Steps Four, Five and Six of its standard workflow. These are the predictive coding steps that iterate. In Step Four the software, Mr. EDR, analyzes the documents designated for training in Step Two in the seed set, and in Step Five thereafter. Mr. EDR then ranks the whole dataset according to probable relevance and irrelevance.

In Step Five the attorneys search for more documents to use to train Mr. EDR. It is essentially the same as Step Two, except now the attorneys can add probability and rank based searches to their multimodal toolkit. That is the Team's full search pyramid, shown right. The methods are used *ad hoc* according to what the attorney reviewer considers a promising method to find additional relevant documents based in part on the latest EDR rankings and TREC submission returns. Once new documents are found that are likely to be relevant, they are then designated in Step Six for Training. Not all documents are so designated. Again this is at the discretion of the attorneys as to what documents they think would best serve to train in the ongoing active learning process.



In Topic 103 the use of predictive coding ranked based searches was severely constrained. This was due to initial configuration setup errors, where input parameters for the learning engine were set incorrectly. These setup errors were detected and corrected by July 22, 2015, and thereafter Mr. EDR was of great assistance. Still, as a result of the delays and early errors, this Topic relied much more heavily than any other on keyword searches and human linear reviews. Similarity searches were also used extensively. Basically the predictive coding assistance in this Topic did not begin until the 14th submission. Losey *called Reasonable* after the 15th submission.

In the TREC experiments most, but not all, of the documents returned as relevant or irrelevant by TREC were included in training (Step Six). In that way their ranking impact was evaluated (Step Four) before the next submission. Training also included various irrelevant documents that were not TREC adjudicated, but were thought to be obviously irrelevant. Experiments were made as to the impact of varying the number of irrelevant documents in the hope that some

ideal range or ratio could be determined to maximize Mr. EDR efficiency. These experiments are still underway. Our conclusions as of late December 2015 are stated in the body of this report.

After a total of 15 submissions that presented 4,806 documents to TREC for adjudication, Losey *called Reasonable* and stopped work on July 22, 2015, a week after the Topic started. Thereafter an additional 9 submissions were made to TREC to submit the remaining 285,293 emails (98.34% of the 290,099 total). There was Training in between most of the remaining seven submissions based on the TREC adjudications, but no further human input. The first two post-call submissions were critical to the Team's excellent performance on this Topic.

Losey *called Reasonable* at the point he thought that a reasonable human effort had been made to find relevant documents. Losey and his assistant Bottolene had personally reviewed and coded as relevant or irrelevant 7,203 documents. (Additional documents had been coded without review.) In fact, by the time Losey had submitted 2,309 documents to TREC for adjudication (the 14th submission) he had completed all individual document review (7,203 documents), and had completed all searches other than predictive coding ranking searches where document content is not reviewed. At that time (after the 14th submission) he essentially turned the process over to Mr. EDR, who had by then just recovered from an earlier technical *illness* and had not been functional before.

At the time Losey *called Reasonable* he had submitted a total of 4,806 documents. Of those, 4,780 had been adjudicated as relevant. This was an incredible Precision rate of 99.46%. This was the most Precise production that Losey thinks he has ever made. He also thought that he may have attained as high as a 90% Recall, but, in fact the later submissions showed that at the time *Reasonable* was called he had attained a Recall of 83.5%. This is still considered a high Recall level in legal search, and the combined F1 measure of 90.8% is, in legal search, like any other, a very outstanding effort.

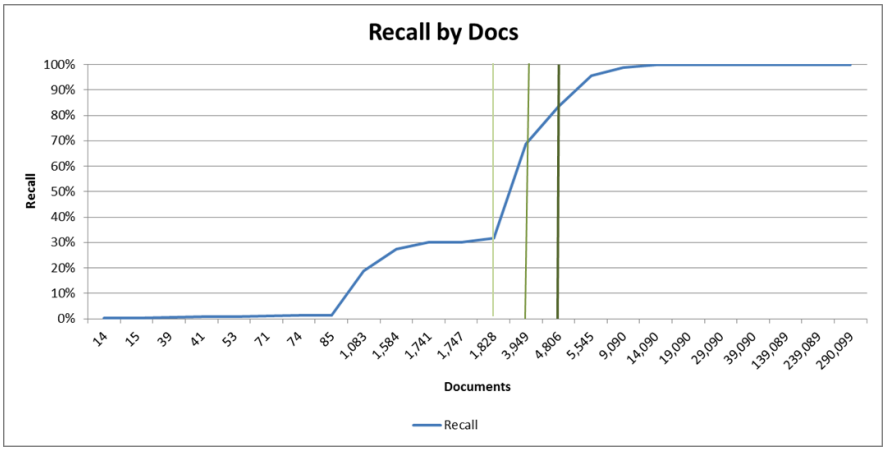
The next submissions after *Reasonable* was called were always the documents that were highest ranked by Mr. EDR, which is why we call this an automated function. As we understand the game set up by TREC for the Recall Track, the actual scoring is not impacted by the *Reasonable call*. The scoring continues for all submissions until all documents have been returned. The Reasonable call is merely an indication of efforts. The same goes for the 70%, 80% recall calls, when and if they are made before the *Reasonable effort* call, except they are of even less interest. These calls were not supposed to have an impact on scoring.

In the first two submissions after the call in Topic 103, the 16th and 17th submissions, Mr. EDR identified and highly ranked 661 additional relevant documents, bringing the total relevant found to 5,467 out of the total 5,725. We were thereby able to attain in that submission a Recall of 90% with Precision of 99.33%, a Recall of 95% with Precision of 98.8%, and **97.5% Recall with a Precision of 90.57%**! As far as Losey knows, these statistics represent his personal best efforts, especially considering that he did so with very little reliance on predictive ranking.

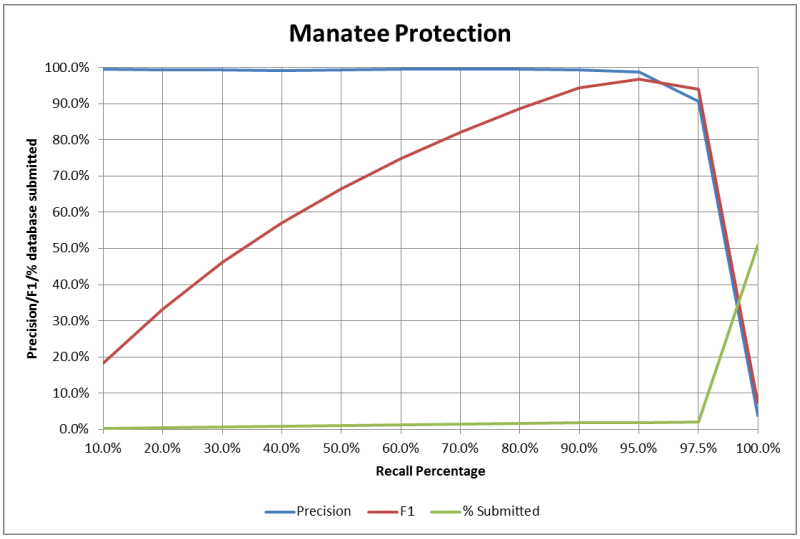
What makes this 97.5% Recall, 90.6% Precision all the more remarkable for legal search is that it was accomplished by only one expert attorney assisted by one contract review attorney. The measured effort to attain these high levels was remarkably low, especially considering that a significant amount of time in Topic 103 was spent reviewing the base line sample (Step Three). Together the two attorneys only reviewed 7,203 documents out of the total corpus of 290,099

emails (2.5%). In legal search it is common for attorney review teams to consist of dozens or even hundreds of attorneys. Moreover, even when predictive coding is used, a far higher percent of the corpus is typically reviewed than 2.5%, and Recall levels of 97.5% are unheard of, much less precision in excess of 90%.

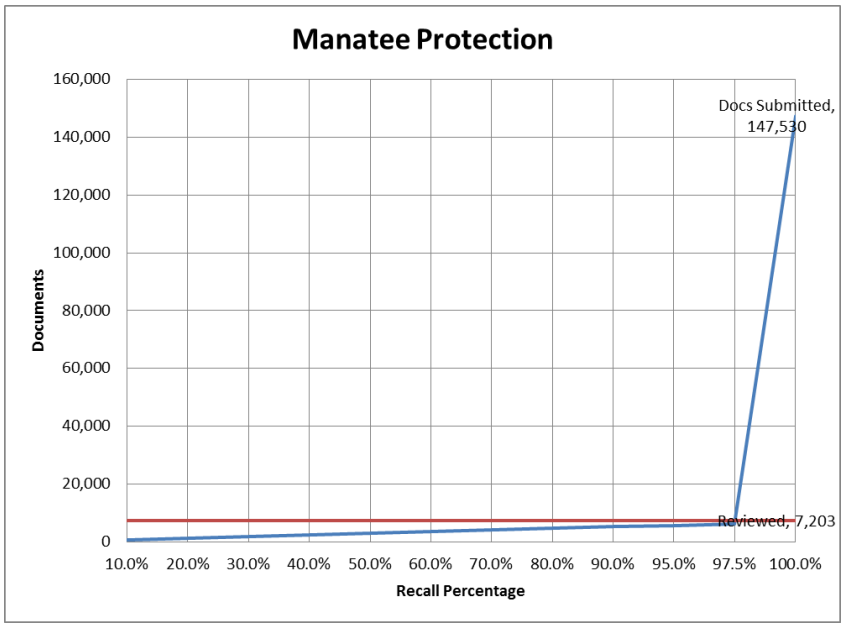
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call. (Please note, that the graph is not to scale as the graph is based on individual submissions. We thought this a better depiction than by proportionally showing progress because in most cases a proportional graph would be a line virtually straight up from the start and flat going over).



The next chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Manatee Protection topic, by the time 97.5% Recall had been attained only 2.12% of the corpus, 6,163 documents, had been submitted for adjudication. This is a triumph for the search pyramid foundation, especially keyword search, that supports AI training. The last portion of the graph thus represents the submission of the remaining 97.88% or 283,936 documents.



The chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 2108 CAPTCHA Services

Confusion Matrix- Topic 2108

Total Documents: 465,147
Total Relevant: 656
Total Prevalence: 0.14%

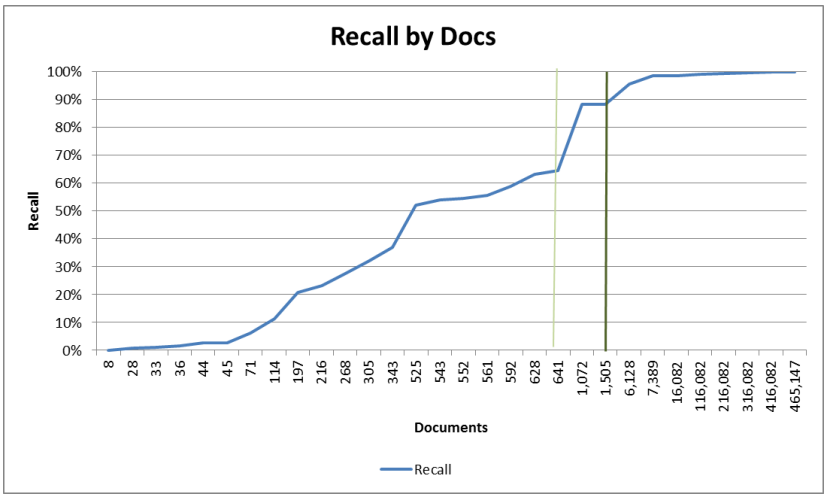
	@Reas. Call	@97.5% Recall
True Positives	580	640
True Negatives	463,566	458,906
False Positives	925	5,585
False Negatives	76	16
Recall	88.41%	97.56%
Precision	38.54%	10.28%
F1 Measure	53.68%	18.60%
Accuracy	99.78%	98.80%
Error	0.22%	1.20%
Elusion	0.02%	0.00%
Fallout	0.20%	1.20%

Topic 2108 was run by Losey without any assistance of a review lawyer. The work to search the 465,149 *BlackHat World Forum* posts started on July 16, 2015, but did not conclude until August 1, 2015. The reason for the delay in completion is that the Team encountered difficulties in understanding the initial TREC adjudications to their first submissions. Neither Losey, nor the other attorney Team members consulted, could understand the relevance pattern behind TREC's initial submission responses. Due to the initial EDR configuration error, predictive coding was not available to assist at first in ascertaining the relevance scope. After several days of struggling with this project, Losey put this Topic on hold until July 29th at which time Losey returned to the Topic to finish.

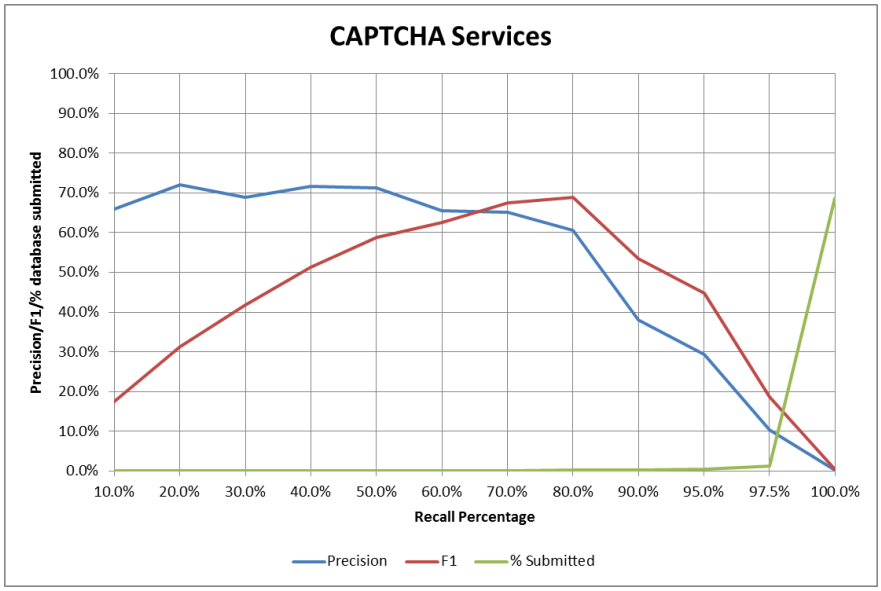
As a general comment the Team found all of the *BlackHat World Forum* posts challenging to search, more difficult than a typical search of corporate ESI. That is in part because almost all metadata of these posts, and all associated imagery, had been stripped by TREC and the ESI converted to text files. Also the language and issues (all non-legal) in the *Black Hat World Forums* were obscure. Even though our attorney searchers were all familiar with forums and had knowledge of most of the technologies and sometimes illegal, nearly always unethical, marketing practices discussed in *Black Hat World*, they still found the slang-filled posts difficult to review and analyze. The challenges were compounded by significant inconsistencies, and apparent illogic of the TREC judging in many of these topics. Still, the Team was able to overcome these challenges and, after we learned not to try to understand any relevance rules, we overall did quite well in review of the ten *BlackHat World Forum Topics*. Based on the elusive (to humans) relevance standard, we found that these topics required greater reliance on Mr. EDR than the *Bush Emails* and *News Articles*. Even though we continued to use a multimodal approach in *Forum* topics, our emphasis was on the AI features of ranking and probability. The Team readily admits that its own human intelligence, without the considerable AI enhancements of Mr. EDR, was not up to the task of matching TREC relevance calls for the *Forum* Topics. But with the help of predictive coding (Me. EDR) we overcame the difficulties and attained relatively high recall levels.

On July 31, 2015, after making 22 document submissions to TREC providing a total 1,505 documents, Losey had found a total of 580 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 2,101 documents. In fact, Losey had stopped document review after the 21st submission. His 22nd submission was entirely based on document rankings without review. After the 22nd TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 88.41%** had been attained. There were seven additional submissions to TREC after the *Reasonable* call point. In the next, 23rd submission, 95% Recall was attained after submitting only 2,130 additional documents.

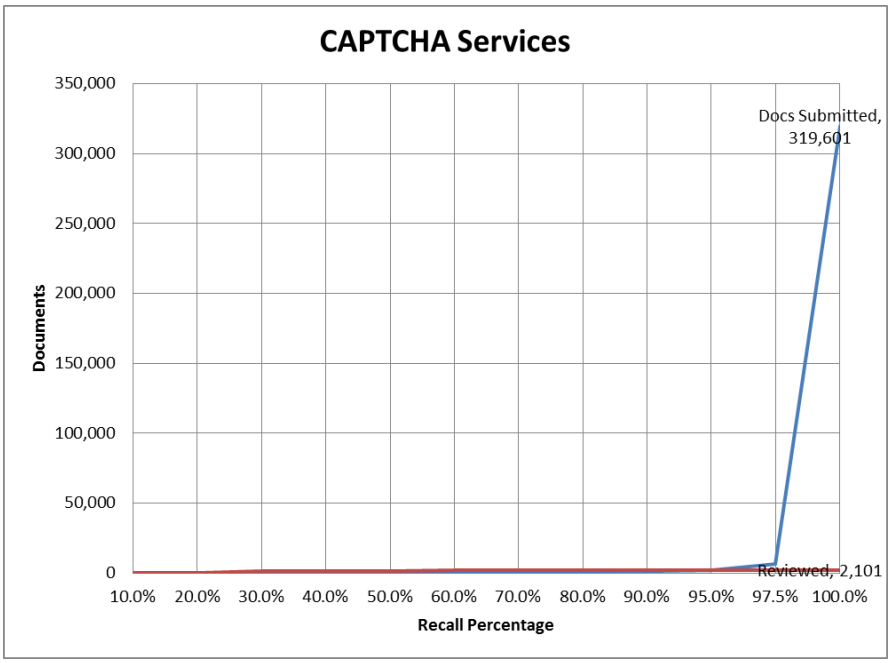
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call. (Please note, that this graph, and all others like it, are not to scale as the graphs are based on individual submissions. We thought this a better depiction than by proportionally showing progress because in most cases a proportional graph would be a line virtually straight up from the start and flat going over).



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the CAPTCHA Services topic, by the time 97.5% Recall had been attained only 1.34% of the corpus, 6,225 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 98.66% or 458,922 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 101 Judicial Selection

Confusion Matrix- Topic 101

Total Documents: 290,099
Total Relevant: 5,834
Total Prevalence: 2.01%

	@Reas. Call	@97.5% Recall
True Positives	5,026	5,688
True Negatives	283,608	281,901
False Positives	657	2,364
False Negatives	808	146
Recall	86.15%	97.50%
Precision	88.44%	70.64%
F1 Measure	87.28%	81.93%
Accuracy	99.49%	99.13%
Error	0.51%	0.87%
Elusion	0.28%	0.05%
Fallout	0.23%	0.83%

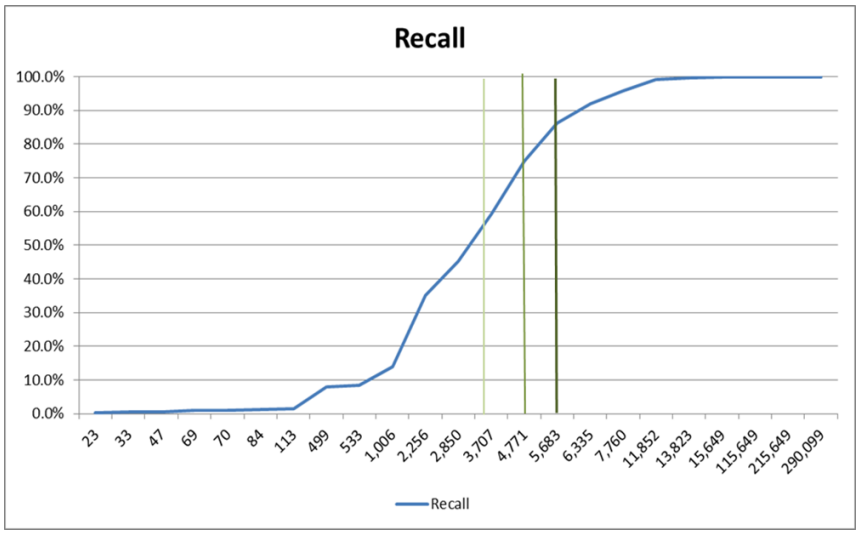
Topic 101 was run by Losey with the assistance of a review attorney, David Jensen. The work to search the 290,099 *Bush Emails* started on July 16, 2015 and concluded on July 26, 2015. The project commenced with Losey beginning Step Two, *Multimodal Search Reviews*, and Jensen assigned Step Three, Random Baseline. Jensen finished the random sample review the next day and began assisting Losey in Step Two, and after submissions began, the echo Step Five, multimodal. Losey handled all of the AI related searches in Step Five, including the probability and ranking related searches. Jensen focused on keyword searches and also made suggestions of documents to submit. Final decisions on submissions were always made by Losey on all Topics.

Due to the same mentioned initial configuration setup errors the AI features did not work until near the end of this Topic. Losey instead relied heavily on Keyword, linear, and a new type of Similarity search the Team invented out of necessity during TREC events. It is anticipated that the new similarity search feature will be included in future Mr. EDR releases.

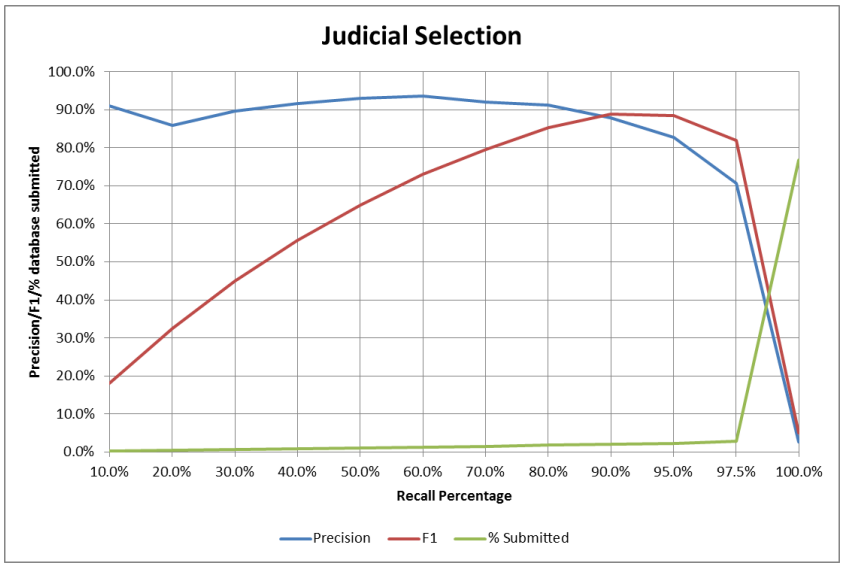
Review of the random sample of 1,534 Bush emails found 30 that were relevant. That suggested a prevalence of 1.96% and a spot projection of 5,673 documents. The actual relevant count of 5,834 and prevalence of 2.01% was very close to the projection. Note this is the second and last Topic in which a full Step Three random sample was implemented.

On July 25, 2015, after making 15 document submissions to TREC providing a total 5,683 documents, Losey had found a total of 5,026 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 6,895 documents. In fact, Losey had stopped document review after the 14th submission, as his 15th submission was entirely based on document rankings without review. After the 15th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 86.15%** had been attained with a **Precision of 88.44%**. There were an additional 8 submissions to TREC after the *Reasonable* call point. In the next, the 16th there was a submission of 652 documents, 345 of which were relevant. **95% Recall** with **82.7% Precision** was attained after submitting only 6,705 documents (1,022 after Reasonable call). **97.5% Recall** with **70.6% Precision** was attained after submitting only 8,052 documents (2,369 after *Reasonable* call).

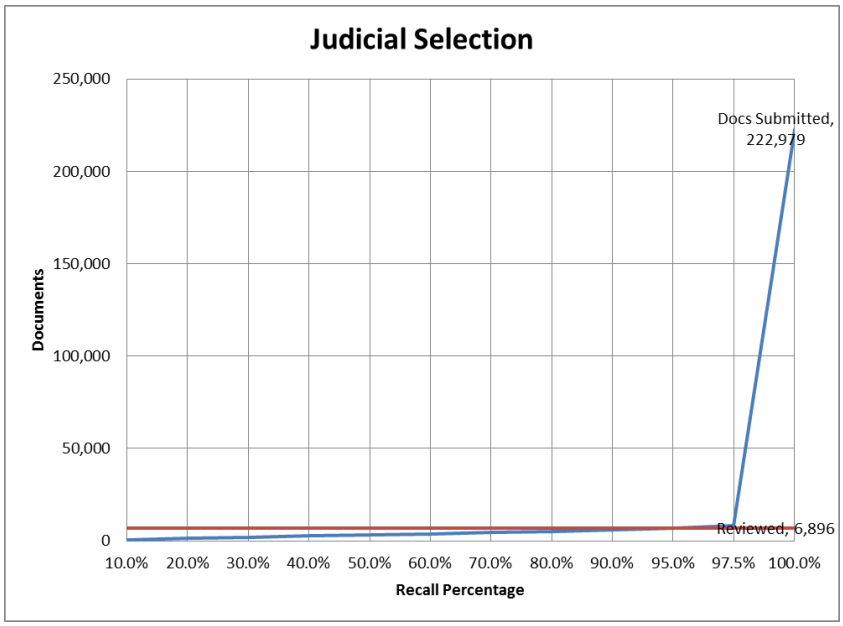
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Judicial Selection topic, by the time 97.5% Recall had been attained only 2.78% of the corpus, 8,052 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 97.22% or 282,047 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 108 Manatee County

Confusion Matrix- Topic 108
Total Documents: 290,099
Total TREC Relevant: 2,375
Total TREC Prevalence: 0.82%

Using TREC relevant calls	**@Reas. Call**	**@97.5% Recall**
True Positives	734	2,316
True Negatives	287,712	26,197
False Positives	12	261,527
False Negatives	1,641	59
Recall	30.91%	97.52%
Precision	98.39%	0.88%
F1 Measure	47.04%	1.74%
Accuracy	99.43%	9.83%
Error	0.57%	90.17%
Elusion	0.57%	0.22%
Fallout	0.00%	90.90%

Topic 108 was run by Losey with the assistance of a review attorney, Bottolene. The work to search the 290,099 *Bush Emails* also started on July 16, 2015 and concluded on July 24, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search*

Reviews. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal, was done by Losey with assistance at first of Bottolene. Losey handled all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made suggestions of documents to submit. All final submittal decisions were made by Losey.

Observations on the Errors of Relevance Judgments in This and Other Topics

This was the most frustrating of all of the TREC Recall Topics for the Team to work on because the judgments on relevance contained more obvious errors and inconsistencies than any other. This Topic was Manatee County, as opposed to Topic 103, which was Manatee Protection, which of course referred to the endangered mammal. Unfortunately, as a life long Florida attorney, Losey has substantial independent knowledge of Manatee County and manatees. Bottolene had also been a Florida resident for several years and an attorney. Their direct personal knowledge of Florida proved to be a significant disadvantage in this Track (and, to a lesser extent, in other Tracks, especially ones that contained obvious errors in relevance) because TREC adjudications were not tied to actual facts and reality (obviously no one at TREC was a Florida SME) and were otherwise surprising.

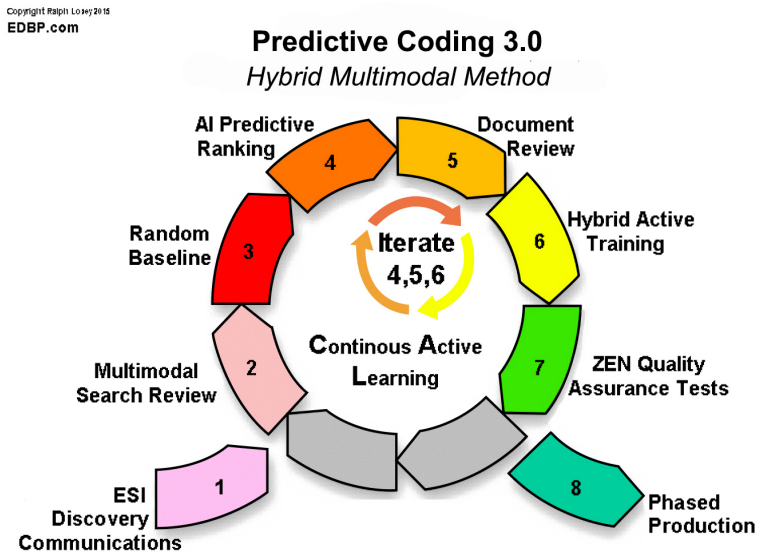
For instance in Topic 108, even though the subject was the *County* of Manatee, a political entity, sometimes, but not always, an email with mere mention of the *mammal* manatee would be considered relevant, even though there was no mention of location or the county. Also, many references to *Manatee Park* were considered relevant to TREC, even though that park is, as any Floridian would know, especially Losey who lives in Central Florida, not located in Manatee County and otherwise has no connection to the county. Also, almost all email addresses that had *manatee* in the name were called relevant by TREC, even if the email had nothing to do with the County of Manatee.

There may well be *some pattern* to the so-called *gold standard* used in this Topic, but if so, it was not logical and not known to Bottolene or Losey. It appeared to these Floridians, after the fact, to be lack of expertise on the part of TREC. Other team members reviewed these adjudications later agreed. One example we were later able to figure out: a well-known Florida law firm (Holland & Knight) has a home office in Bradenton, Florida, and the attorneys there would often write to the governor. As part of post-hoc analysis we saw that almost all of these emails were considered relevant by TREC assessors to this topic simply because the office city was in their standard signature line address, even though the content of the emails has nothing to do with Manatee County.

Since Losey is used to directing legal search as an SME, or direct SME surrogate, his usual approach to legal search involves using his knowledge and understanding to differentiate relevant from irrelevant. As mentioned, in legal search understanding of relevance is critical, in fact, it is a legal duty and responsibility of the attorney searchers. Thus his position as an actual Florida SME served as a disadvantage in many of the Bush email Topics, including this one.

The Team later encountered other Topics with inconsistencies and mistakes like Topic 108. In such cases we eventually learned to step out of the process and stop trying to understand or look for a rational basis for the TREC relevance calls. We would put aside our traditional SME role, which is otherwise the firmly established norm in legal search. Instead, when we found

ourselves in this situation (and this happened in a little less than half of the Topics), we would basically turn the search and submission decisions over to Mr. EDR. In those situations we did not even try to see any pattern or consistency to the adjudications. When we adopted this approach in later topics we did quite well, in spite of defects we saw in the TREC gold standards. This suggests that TREC's selection of relevant documents in some of the Topics suffered from over-delegation to computer selection without adequate SME based quality controls. It is unknown what software was used by TREC to create the relevant gold standard document set, but like any predictive coding software today, it obviously can be led astray without adequate human supervision and quality control safeguards. This is why the e-Discovery Team adopts a hybrid approach, computer and human, including SMEs, and why in normal circumstances Step Seven for quality control is so important under their Predictive Coding 3.0 method.



Topic 108 Description

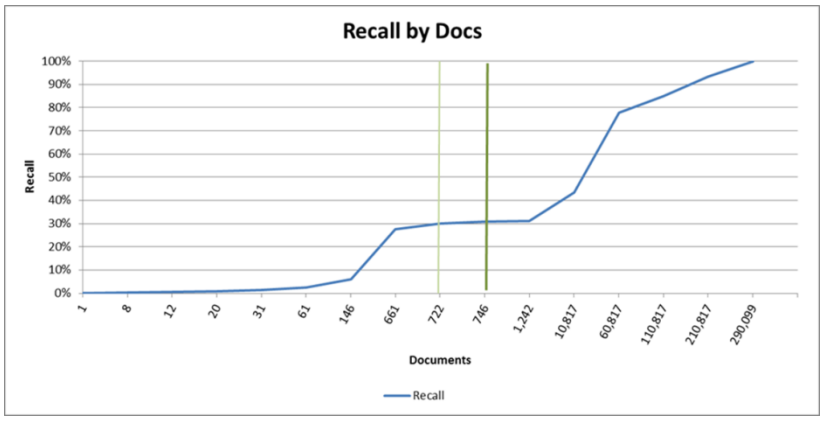
On July 23, 2015, after making 10 document submissions to TREC providing a total 746 documents, Losey had found a total of 734 relevant documents (Precision of 98.4%). The effort, or number of documents reviewed and coded by Losey to attain this result, was 696 documents. After the 10th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 31%** had been attained. The decision to call Reasonable proved to be a big mistake because the TREC adjudications were not limited to Manatee County relevance as the Team had assumed. As mentioned, the error was based upon the Team's construction of relevance in a much narrower manner than TREC. The divergence was not known because the Team did not do enough exploration of irrational constructions and so did not detect the, to our mind, outlier nature of TREC's approach to this Topic.

The Team should have been *less precise* (its submissions had a Precision of 98.4%), and should have presented more documents for submission, even though the Team did not personally consider them to be relevant. It should have better tested its relevance concept. But as mentioned, as an SME Losey was used to setting the scope of relevance, and as lawyers, the entire Team was used to rational adjudications of relevance along lines that make sense to them.

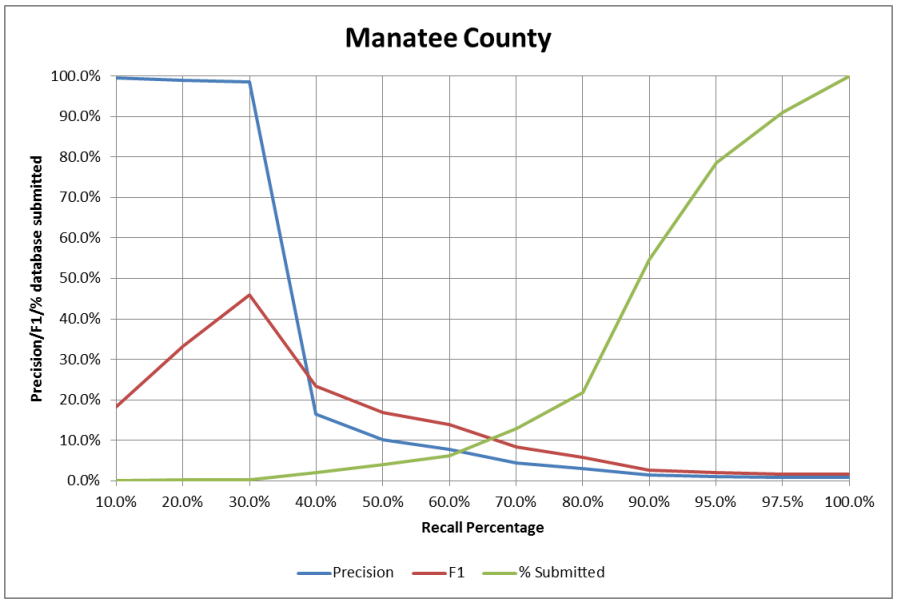
This was an early topic for us in the process and we had not yet learned to mistrust our own assessments.

There were 6 additional submissions to TREC after the *Reasonable* call point. In retrospect, this was also an error. The Team should have submitted multiple smaller submissions after they started to discover the outlier nature of the TREC adjudications, with training between each submission where Mr. EDR could take over in an automated fashion. This was another game-type lesson learned the hard way by this Topic, which proved to be the Team's worst performance. Even in the worst case with multiple mistakes the Team still managed to attain 78% Recall with review of only 696 documents, and submission of only 60,817 of the total 290,099 documents.

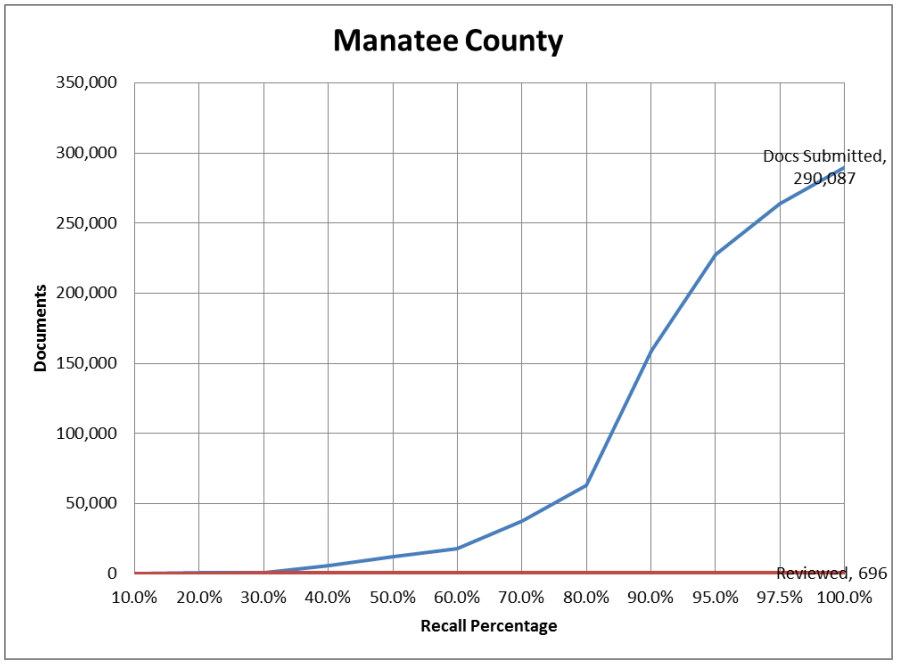
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the Reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Manatee County topic, by the time 97.5% Recall had been attained 90.95% of the corpus, 263,843 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 9.05% or 26,256 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Correction of the Gold Standard Relevance Set in Topic 108

Since the Team is considering use of the Bush email set in further testing, training and research, they wanted to try to correct the many deficiencies they saw in TREC's determination of the gold standard for this Topic. They also wanted to better understand why the score on this Topic was so out of range from their other scores. With this in mind they re-reviewed the TREC

adjudications and set up a three-attorney peer review of all errors spotted in the relevancy determinations. A conservative approach was taken and deference was given to the TREC adjudications where a rational, consistent basis could be found. Losey's personal, narrow view of what should be relevant was not followed, if there was a reason seen to follow TREC's adjudications. (Note, the Team and others in the filed of Legal Search, have observed over many projects that SMEs typically take a more narrow view of relevance than non-SMEs who, by definition, do not understand the subject as well.) Losey accepted all adverse rulings against his own positions as part of this process. Also note that suggestions to revise TREC adjudications came from all three Team members, not just Losey, and were all subject to multiple reviews and objections.

After the re-review and re-adjudication process was completed, 1,264 documents adjudicated as relevant by TREC were changed to Irrelevant. Further, 3 documents adjudicated as irrelevant by TREC were changed to relevant. Below are the corrected metrics of the Team's review under the improved adjudications.

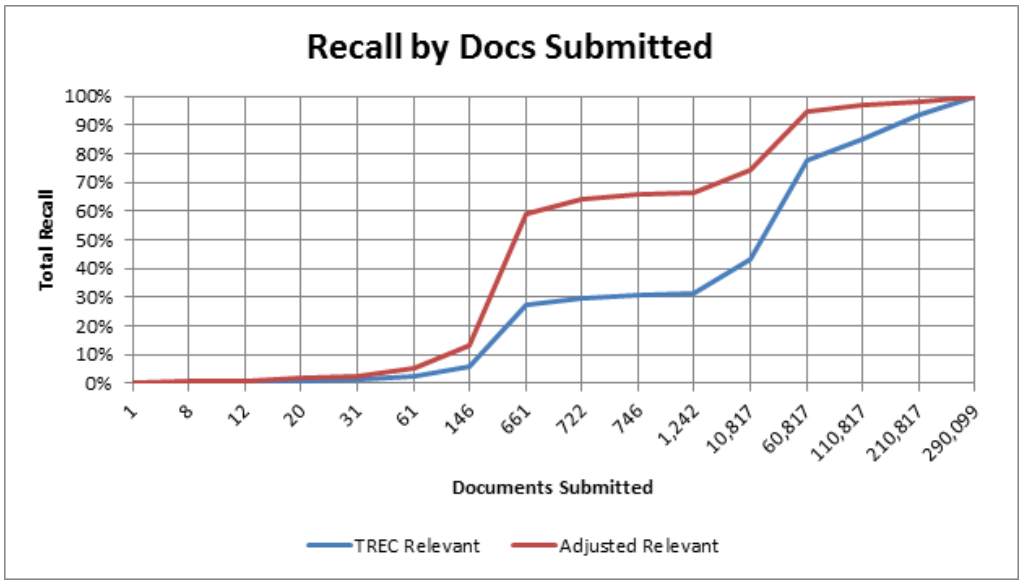
Confusion Matrix (Adjusted) - Topic 108

Total Documents: 290,099
Total Adjusted Relevant: 1,114 (was 2,375) (1,264 changed to Irrelevant, 3 Changed to Relevant)
Total Adjusted Prevalence: 0.38% (was 0.82%)

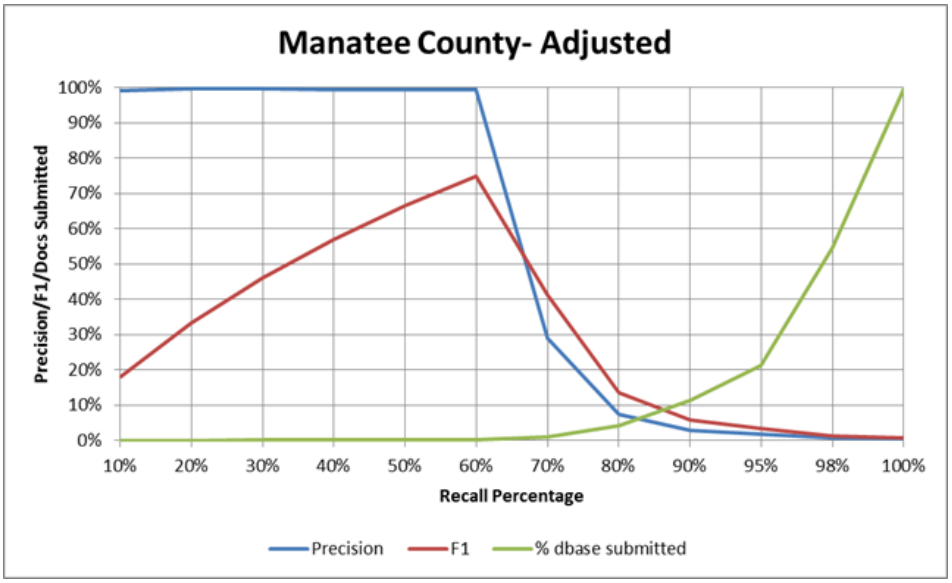
Using adjusted relevant calls	**@Reas. Call**	**@97.5% Recall**
True Positives	736	1,087
True Negatives	288,975	131,844
False Positives	10	157,141
False Negatives	378	27
Recall	66.07%	97.58%
Precision	98.66%	0.69
F1 Measure	79.14%	1.36%
Accuracy	99.87%	45.82%
Error	0.13%	54.18%
Elusion	0.13%	0.02%
Fallout	0.00%	54.38%

After the 10th TREC submission, when Losey decided to call *Reasonable,* Losey had found a total of 736 relevant documents (an increase of 2 documents) under the adjusted gold standard. This was a **Recall of 66.07**% and **Precision of 98.66**% under the adjusted standard. The F1 measure was 79.14%. Note that these metrics are much more inline with the other 29 projects, although the adjusted 66% Recall is still the Team's second to lowest Recall score at the Reasonable call point. Under the corrected standard the Team attained 94.43% Recall with review of only 696 documents, and submission of only 60,817 of the total 290,099 documents.

A graph mapping how the review by Recall attained after number of documents submitted is shown below with both the original TREC standard (blue) and the Team adjusted standard (red).



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds under the adjusted standard.



_____

Topic 2052 Paying for Amazon Book Reviews

Confusion Matrix- Topic 2052
Total Documents: 465,147
Total Relevant: 265
Total Prevalence: 0.06%

	@Reas. Call	**@97.5% Recall**
True Positives	257	259
True Negatives	464,364	464,165
False Positives	518	717
False Negatives	8	6
Recall	96.98%	97.74%
Precision	33.16%	26.54%
F1 Measure	49.42%	41.74%
Accuracy	99.89%	99.84%
Error	0.11%	0.16%
Elusion	0.00%	0.00%
Fallout	0.11%	0.15%

Topic 2052 was run by Sullivan, who started on July 20, 2015, and concluded July 22, 2015. This was Sullivan's first Topic. For that reason he spent more time than in his later reviews in trying to understand the dataset and processes.

Sullivan has a background in computers and programming. He has substantial experience in forums to understand the unique characteristics present in forum communications. While he considers himself far more knowledgeable than the average person, he has no experience with the unethical world of *Blackhat Forums* and does not consider himself to be a bona fide subject matter expert (SME) on any of them.

All forum topics presented a unique challenge of identifying variations of terms and understanding use of slang. While this proved to be easy to overcome, it certainly played a vital role in the process in a way not necessary in the News topics, where spelling errors were largely non-existent.

On the first day, Sullivan started with Step Three, Random Baseline and reviewed a random sample of 1,534 documents. This was used both as a method to estimate prevalence and a means of gaining better understanding of the dataset for this and future topics in AtHome2. This random sample yielded 1 relevant document. Based on the sample prevalence we predicted 303 relevant documents existed in the dataset (95% confidence level with 2.5% margin of error). We would later discover the dataset contained 265 relevant documents, which is well within the margin of error. Given the amount of time necessary to complete this random sample, and the little value gained, Step Three was omitted from all subsequent topics reviewed by Sullivan.

Day two was spent running keyword searches to find documents for seeding into the predictive coding algorithm and submitting documents to get a better understanding the TREC standard for relevance. At the end of day two, 273 documents had been submitted, with 204 being returned as relevant. This provided an adequate seed set to being relying more heavily on predictive coding.

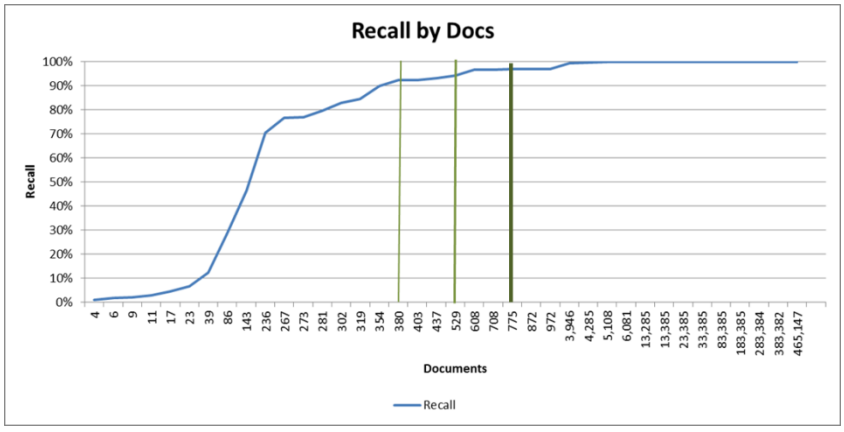
On day three, Sullivan developed a strategy which he relied heavily in future topics. Rather than relying on Mr. EDR alone and reviewing the documents that were given high scores by the machine, he used the multi-modal approach to prioritize documents for review. Starting with all variations of "Amazon" w/5 "Review," he worked down reviewing and categorizing the highest scoring documents first. When he hit a point where few relevant documents were being found, he iteratively expanded the scope of his review universe. He moved to all variations of "Amazon" w/10 "Review, then "Amazon" w/25 "Review," and "Amazon" AND "Review." He expanded into "Amazon" and ("Review" or "Book" or "Feedback" or "Purchase") and eventually to any document containing a variation of "Amazon."

As previously mentioned, the unique characteristics of the forums required more creative searches than necessary in other datasets. Using the Concept Searching tool as a guide, it was determined that almost all reasonable variations of "Amazon" could be found using the following search: ("amazon*" OR "@mazon" OR "@maz0n" OR "azmon*" OR "azmn*" OR "amzn*"). This method proved effective in eliminating issues of missed documents due to slang or misspelling.

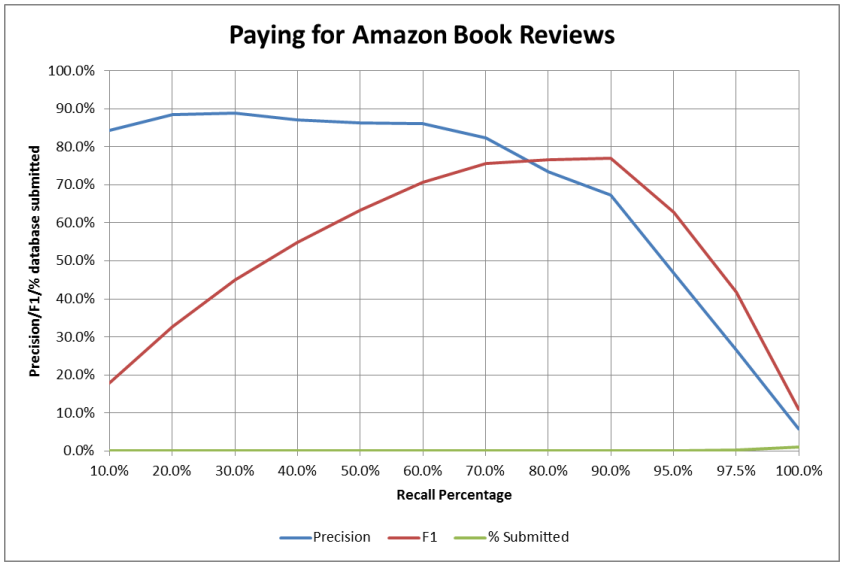
Using this method, Sullivan was able to identify 257 of the 265 relevant documents at the time he called Reasonable effort. 2,325 total documents had been reviewed, included the 1,534 documents in the initial random sample.

After calling Reasonable effort, Sullivan continued by submitting all documents that contained any variation of the term "Amazon" in order of priority score descending. 100% recall was obtained through this method. All remaining documents were then submitted in descending priority order, with no more relevant documents being returned.

A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, slightly darker line signifies 80% Recall call and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Paying for Amazon Book Reviews topic, by the time 97.5% Recall had been attained only 0.21% of the corpus, 976 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.79% or 464,171 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multimodal hybrid model of training EDR.

_____

Topic 2225 Rootkits

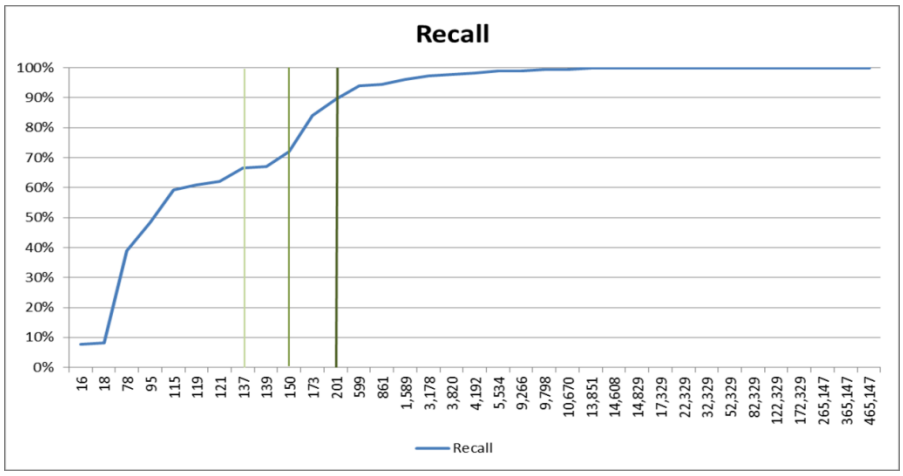
Confusion Matrix- Topic 2225
Total Documents: 465,147
Total Relevant: 182
Total Prevalence: 0.04%

	@Reas. Call	**@97.5% Recall**
True Positives	163	178
True Negatives	464,927	461,955
False Positives	38	3,010
False Negatives	19	4
Recall	89.56%	97.80%
Precision	81.09%	5.58%
F1 Measure	85.11%	10.56%
Accuracy	99.99%	99.35%
Error	0.01%	0.65%
Elusion	0.00%	0.00%
Fallout	0.01%	0.65%

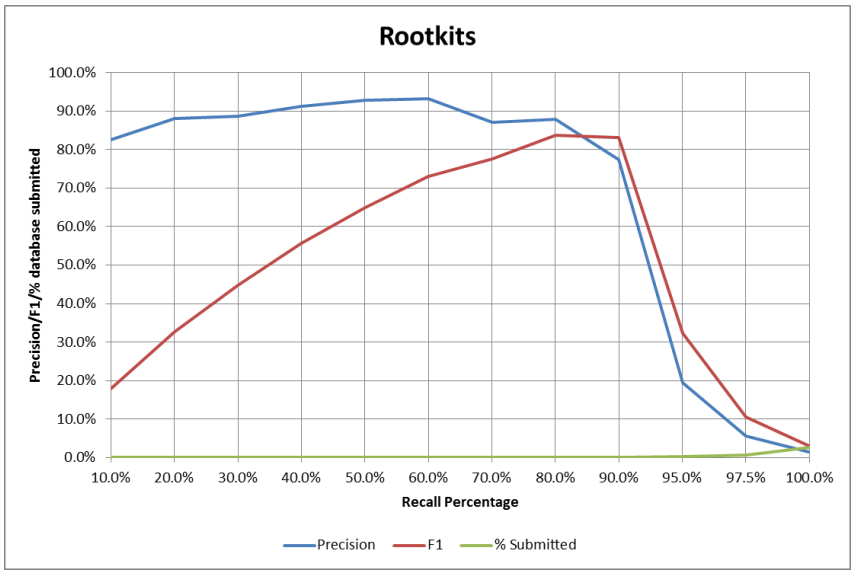
Topic 2225 was run by Losey who started the search of 290,099 *Black Hat Forum* posts on July 21, 2015 and concluded on August 18, 2015. Losey put aside work on this Topic several times while he gave priority to the *Jeb Bush Email Topics*. The project commenced as usual with Losey beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal search began, including predictive coding features, with iterated training.

On August, 2015, after making 12 submissions to TREC, and training after almost every submission, Losey had provided a total 201 documents to TREC and confirmed a total of 163 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 205 documents. After the 12th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 89.56%** had been attained with **a Precision of 81%.** There were 23 additional submissions to TREC after the *Reasonable* call point. A 90% Recall was attained after submitting only 212 documents. A 95% Recall was attained after submitting 891 documents, and 97.5% Recall attained after 3,188 documents. Total Recall was attained after submitting 12,109 documents out of the corpus total of 465,147.

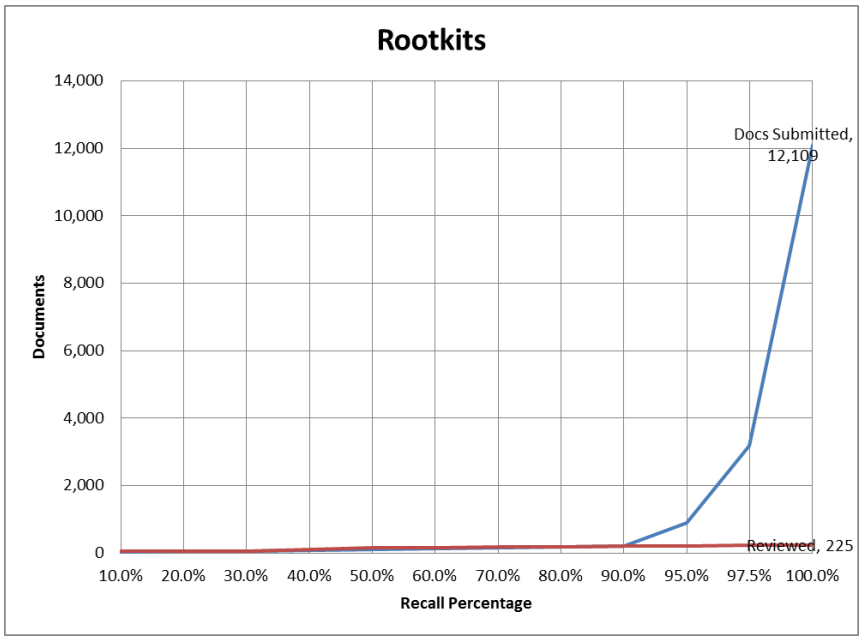
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the *Reasonable* Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Rootkits topic, by the time 97.5% Recall had been attained only 0.69% of the corpus, 3,188 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.31% or 461,959 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 102 Capital Punishment

Confusion Matrix- Topic 102 Capital Punishment

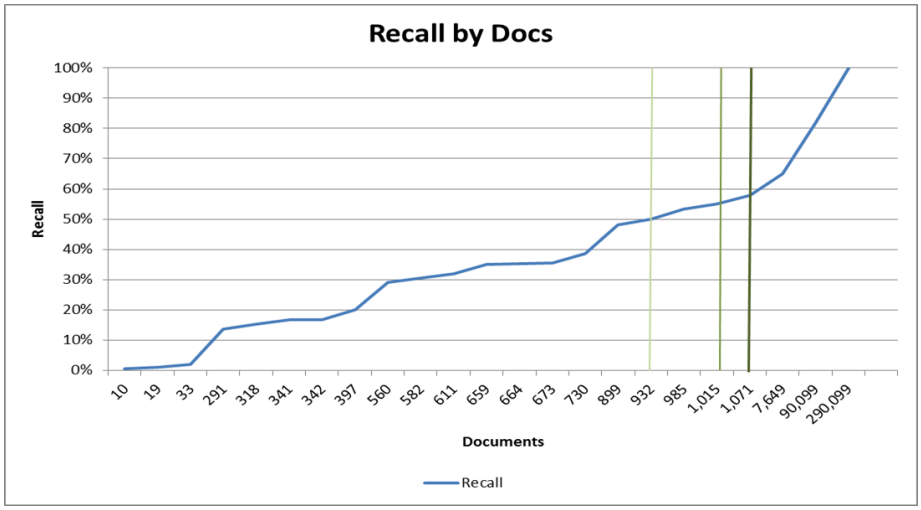
Total Documents: 290,099
Total Relevant: 1,624
Total Prevalence: 0.56%

	@Reas. Call	@97.5% Recall
True Positives	941	1,583
True Negatives	288,345	17,048
False Positives	130	271,427
False Negatives	683	41
Recall	57.94%	97.50%
Precision	87.86%	0.58%
F1 Measure	69.83%	1.15%
Accuracy	99.72%	6.42%
Error	0.28%	93.58%
Elusion	0.24%	0.24%
Fallout	0.05%	94.09%

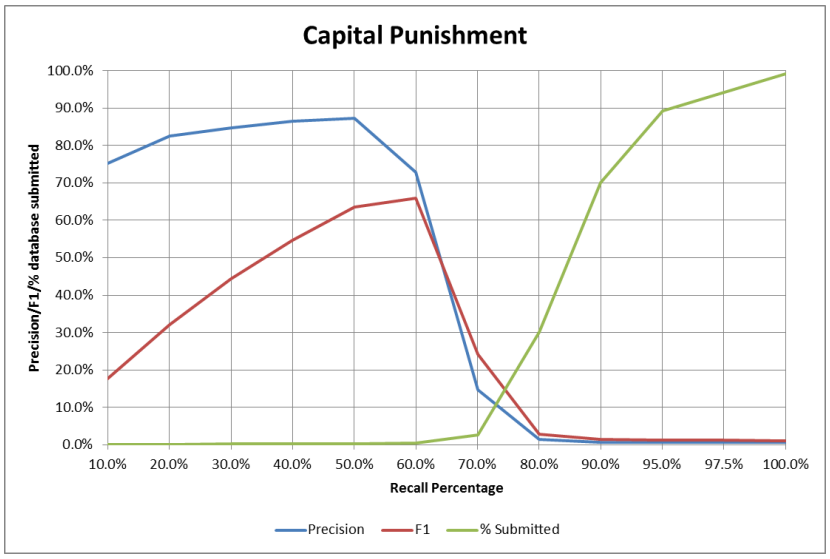
Topic 102 was run by Losey with the assistance of a review attorney, Jensen. The work to search the 290,099 *Bush Emails* started on July 26, 2015 and concluded on July 29, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal, was handled with the assistance, at first, of Jensen. Losey performed all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made suggestions of documents to submit. Again, all final decisions on submittal were made by Losey.

On July 28, 2015, after making 20 submissions to TREC, and training after almost every submission, Losey had provided a total 1,071 documents to TREC and confirmed a total of 941 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 1,493 documents. After the 20th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 57.94%** had been attained with a **Precision of 87.86%,** so his call proved to be early. There were only 3 additional submissions to TREC after the *Reasonable* call point, which we later learned was a mistake. We learned later that higher Recall and overall TREC scoring comes from multiple, smaller submissions, with training after each. This is another Topic in which we found many of the TREC judgments inconsistent and incomprehensible. Still, even with these problems and errors, a **Recall of 70%** was attained after a total of only 7,785 documents had been submitted out of 290,099, and only 1,493 documents had been reviewed.

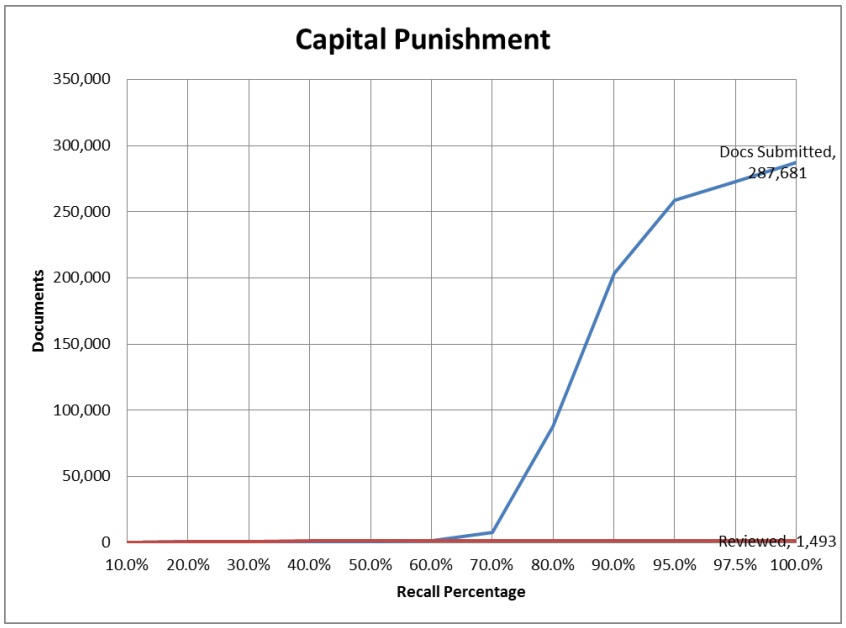
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall Call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Capital Punishment topic, by the time 97.5% Recall had been attained 94.11% of the corpus, 273,010 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 5.89% or 17,089 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



_____

Topic 106 Terri Schiavo

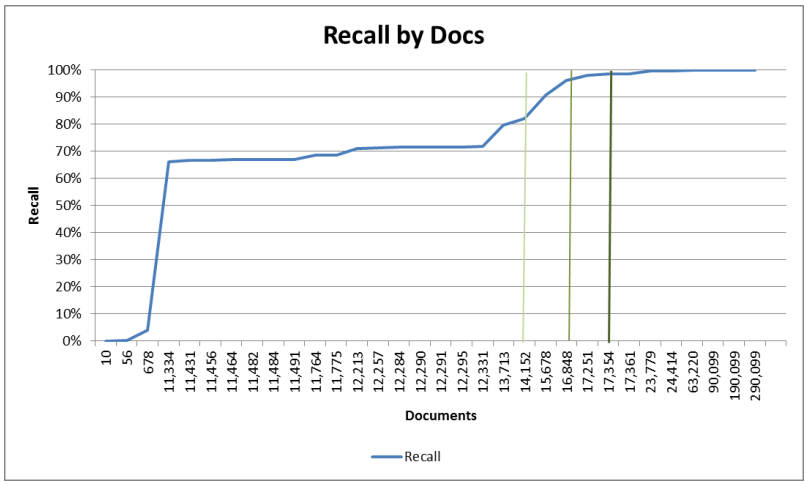
Confusion Matrix- Topic 106
Total Documents: 290,099
Total Relevant: 17,135
Total Prevalence: 5.91%

	@Reas. Call	@97.5% Recall
True Positives	16,872	16,707
True Negatives	272,482	272,551
False Positives	482	413
False Negatives	263	428
Recall	98.47%	97.50%
Precision	97.22%	97.59%
F1 Measure	97.84%	97.54%
Accuracy	99.74%	99.71%
Error	0.26%	0.29%
Elusion	0.10%	0.16%
Fallout	0.18%	0.15%

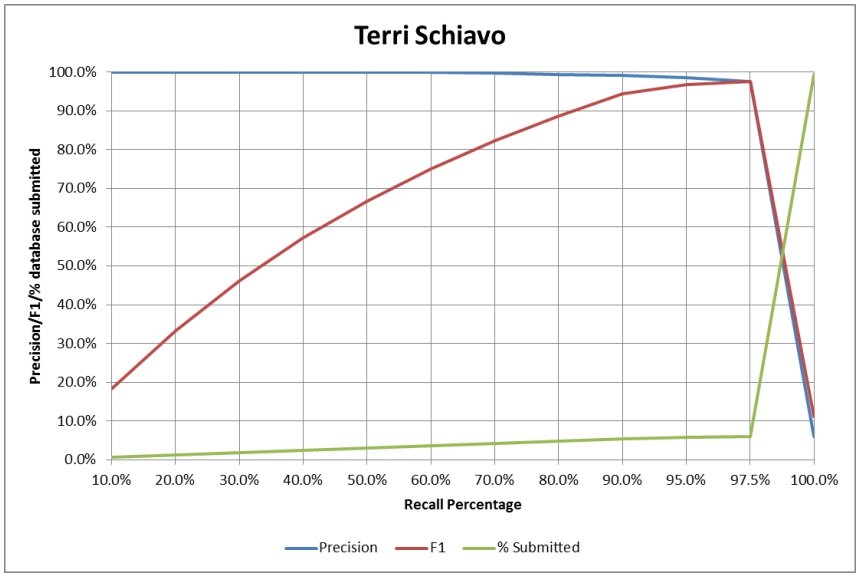
Topic 106 was run by Losey with the assistance of a review attorney, Bottolene. The work to search the 290,099 *Bush Emails* started on July 27, 2015 and concluded on August 2, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal was handled with the assistance at first of Bottolene. Losey performed all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made suggestions of documents to submit. Again, all final decisions on submittal were made by Losey. This review process went longer than other because this proved to be the highest prevalence Topic (5.91%).

On August 2, 2015, after making 25 submissions, with training after most of these, Losey had submitted a total 17,354 documents. A total of 16,872 of these submissions were confirmed relevant by TREC, for a **Precision rate of 97.22%**. The effort, or number of documents reviewed and coded by Losey to attain this result, was 2,025 documents. After the 25th TREC submission, Losey decided to call *Reasonable*. It was later determined that an incredible **Recall of 98.47%** had been attained. The **F1 measure was 97.84%**. That is the Team's best result on any of the Bush Email Topics. Further, Losey believes this may be a personal best for Recall and F1 scores. There were 7 additional submissions to TREC after the *Reasonable* call point. In the 29th submission, **99.7% Recall** was attained after submitting only 7,060 additional documents. The **Precision was 70%**.

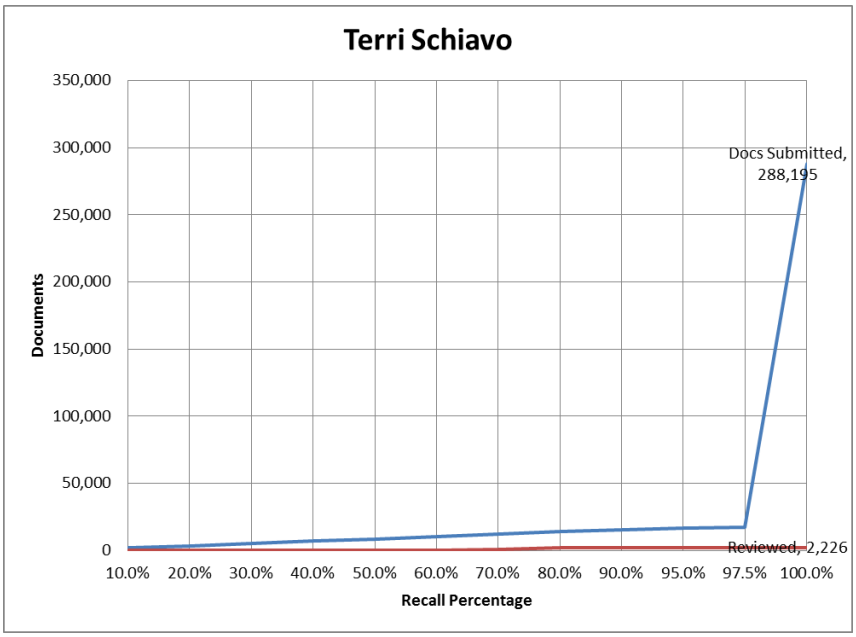
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the Reasonable Recall Call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Terri Schiavo topic, by the time 97.5% Recall had been attained only 5.90% of the corpus, 17,120 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 94.10% or 272,979 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 105 Affirmative Action

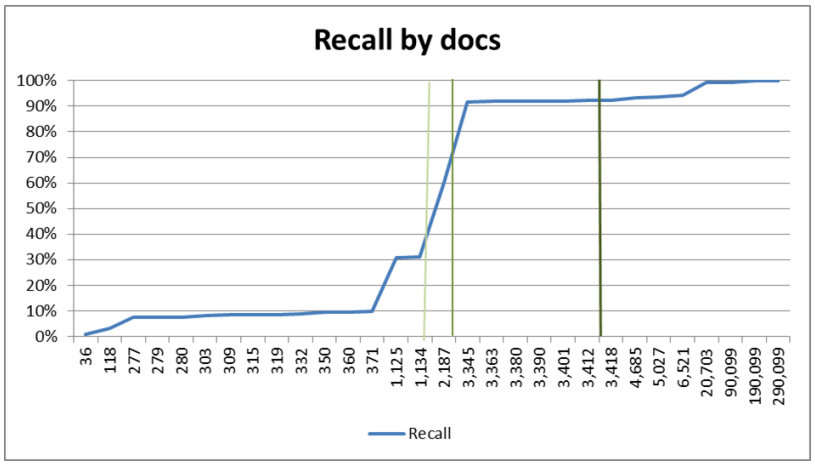
Confusion Matrix- Topic 105
Total Documents: 290,099
Total Relevant: 3,635
Total Prevalence: 1.25%

	@Reas. Call	**@97.5% Recall**
True Positives	3,353	3,544
True Negatives	286,399	281,585
False Positives	65	4,879
False Negatives	282	91
Recall	92.24%	97.50%
Precision	98.10%	42.08%
F1 Measure	95.08%	58.78%
Accuracy	99.88%	98.29%
Error	0.12%	1.71%
Elusion	0.10%	0.03%
Fallout	0.02%	1.70%

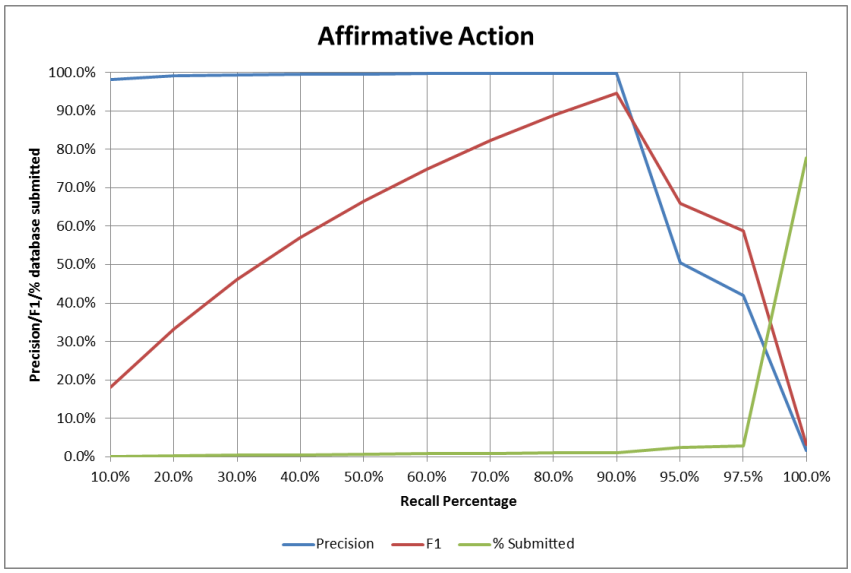
Topic 105 was run by Losey with the assistance of a review attorney, Jensen. The work to search the 290,099 *Bush Emails* started on July 29, 2015 and concluded on July 31, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal, was performed with the assistance at first of Jensen. Losey handled all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made suggestions of documents to submit. Again, all final decisions on submittal were made by Losey.

On July 30, 2015, after making 23 document submissions to TREC providing a total 3,418 documents, Losey had found a total of 3,353 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 674 documents. After the 23rd TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 92.24%** had been attained, with **Precision of 98.1%,** and **F1 of 95.08%**. There were 7 additional submissions to TREC after the *Reasonable* call point. In the 27th submission, after submitting only 3,427 additional documents (total 6,845), **95% Recall** was attained. This was attained after submission of only 2.36% of the total documents.

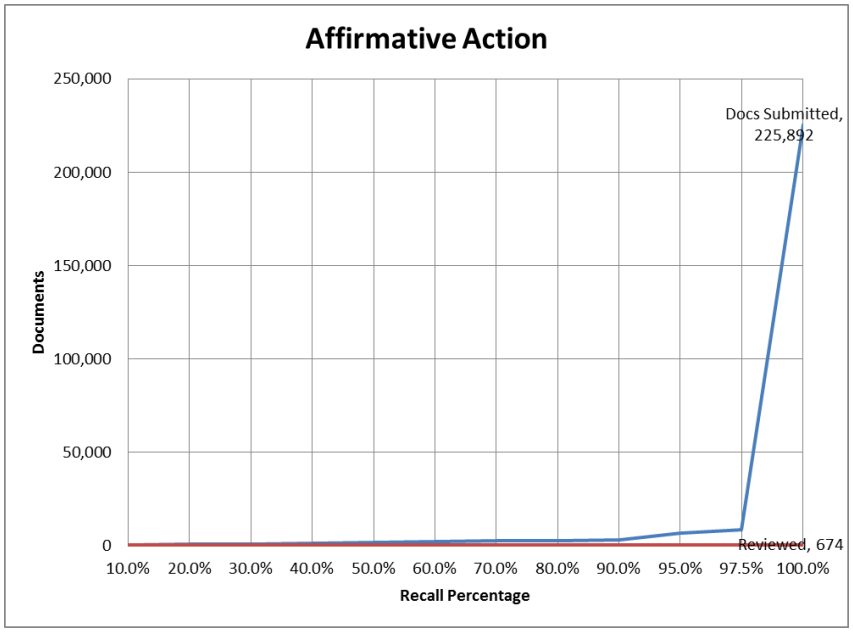
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the **Reasonable Recall call.**



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Affirmative Action topic, by the time 97.5% Recall had been attained only 2.90% of the corpus, 8,423 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 97.10% or 281,676 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



_____

Topic 3357 Occupy Vancouver

Confusion Matrix- Topic 3357
Total Documents: 902,434
Total Relevant: 629
Total Prevalence: 0.07%

	@Reas. Call	@97.5% Recall
True Positives	576	613
True Negatives	901,680	900,834
False Positives	125	971
False Negatives	53	16
Recall	91.57%	97.46%
Precision	82.17%	38.70%
F1 Measure	86.62%	55.40%
Accuracy	99.98%	99.89%
Error	0.02%	0.11%
Elusion	0.01%	0.00%
Fallout	0.01%	0.11%

Topic 3357 was run by Reichenberger. The work to search the 902,434 *News Articles database* started on July 29, 2015, and completed on July 30, 2015.

The initial submissions on the first day were to test the outlines of the category. The initial search of "Occupy" AND "Vancouver" identified a series of protests in Vancouver about economic income inequality. Documents were selected based on a varying of content, including "Occupy" movements in other cities, riots/protests that took place in the same area (but not same time) as the Occupy Vancouver protests, and generic stories about "Occupy" protests that reference protests in Vancouver but do not specifically name them as "Occupy Vancouver." Various sources were also tested, such as Letters to the Editor, stories sourced in other cities and so forth. Results helped formulate an anticipated rule on relevance.

After training EDR and receiving priority scores, relevant documents on subsequent submissions were confirmed by these rules and their priority scores. In fact, of the five irrelevant documents found in the last 2 submissions on July 29th, three scored over 97% and contained substantial and direct references to Occupy Vancouver; these may be TREC coding errors.

A modified Step Three, Random Sample of 1,000 documents was taken after Step Two was complete. The first 500 contained 50 "training" documents to focus on, while the second 500 documents contained 250. All documents hitting on "Occupy" OR "Vancouver" OR "Ashlie Gough" (a student who died at the protests) OR "Robson Square" (location of the protests) were reviewed, while all others mass trained as irrelevant. The last TREC submission on July 29th was from the 1,000 random documents. Of the 1,000 documents, 33 were identified as relevant, confirmed by submission.

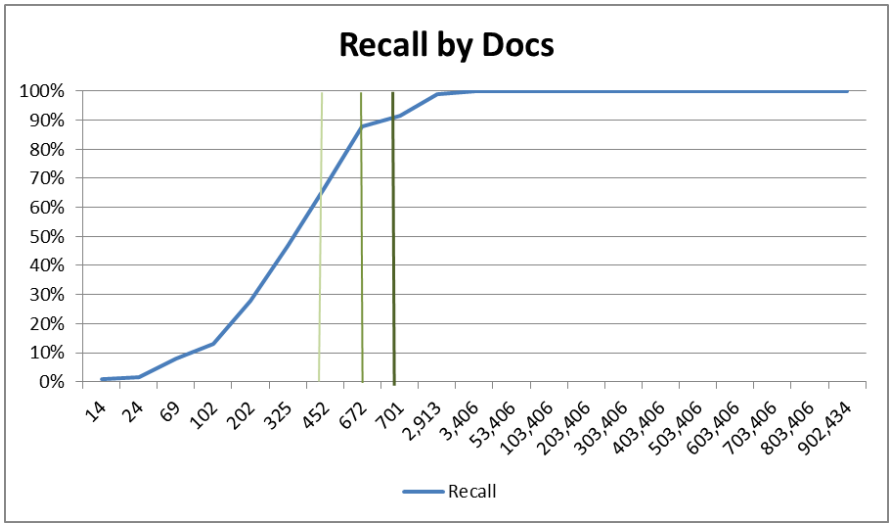
On the second day, the 30th, submissions by documents containing search terms and escalated as relevant were reviewed and submitted in priority order. In the first submission of the day, 123 were submitted as relevant and 118 came back as confirmed relevant. Of the five irrelevant in that set, four were documents that had the exact same relevant text as documents TREC previously confirmed as relevant. This is another example of the kind of "gold standard" inconsistencies the Team encountered in most of the Topics.

In the next set of submissions, documents escalated as relevant by Mr. EDR included stories sourced in the Vancouver paper on Occupy movements elsewhere, and sports stories with the word "occupy" in the article (e.g. "Another Vancouver player occupied the penalty box"). Once those documents were removed as irrelevant, all others were submitted and confirmed as relevant on submission. Some additional "gray area" documents were submitted (e.g. "Occupy Christmas" which was an offshoot of the protests, or campaign questions posed to candidates about the Occupy Vancouver protests).

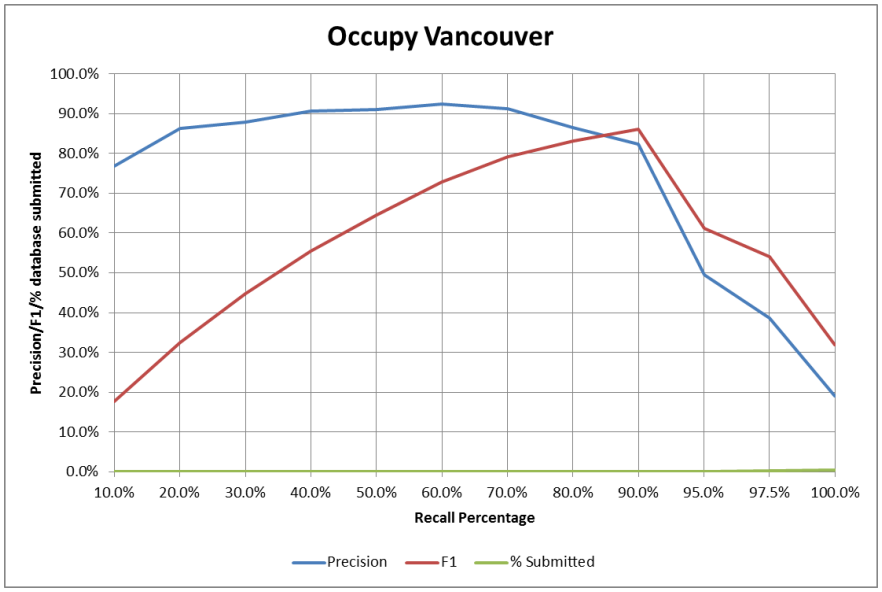
As the Mr. EDR ranking scores decreased, the precision dropped. Prior to the final submissions, all documents with "Occupy" and "Vancouver" with relevance probability scores over 0.1% had either been submitted or reviewed, and all documents with scores over 75% without those terms had also been reviewed.

After the final Reasonable call was made the remaining documents were submitted in the following groups in descending priority order: 1) all documents currently coded as irrelevant by the human reviewer not yet submitted (2,212 documents, of which 45 were found to be relevant); 2) anything remaining with "Occup!" AND "Vancouver" (493 documents, all these had scores below 0.1%, of which 8 were found to be relevant); and then 3) all else (no relevant documents found in this set).

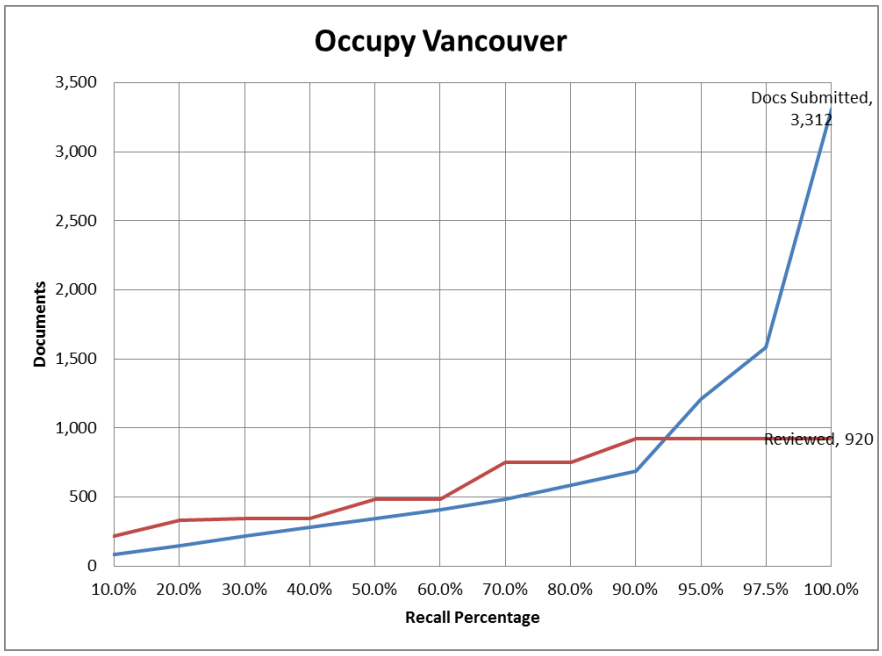
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Occupy Vancouver topic, by the time 97.5% Recall had been attained only 0.18% of the corpus, 1,584 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.82% or 900,850 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 2158 Using TOR for Anonymous Browsing on the Internet

Confusion Matrix- Topic 2158
Total Documents: 465,149
Total Relevant: 1,261
Total Prevalence: 0.27%

	@Reas. Call	@97.5% Recall
True Positives	1,243	1,230
True Negatives	463,793	463,824
False Positives	95	64
False Negatives	18	31
Recall	98.57%	97.54%
Precision	92.90%	95.05%
F1 Measure	95.65%	96.28%
Accuracy	99.98%	99.98%
Error	0.02%	0.02%
Elusion	0.00%	0.01%
Fallout	0.02%	0.01%

Topic 2158 was run by Sullivan who also started on July 29, 2015. He finished his review of 465,149 forum posts in *BlackHat World* on July 31, 2015

Sullivan's computer background proved to be helpful in another uncommon forum topic. He considers himself more knowledgeable on this topic than the average person, but does not consider himself to be a subject matter expert on TOR.

Day 1 of this topic started with concept searching to find other keywords relating to TOR and anonymous browsing. Many previously unknown terms came to light, such as vpn, torbrowser, proxy, and ip. This process of using concept searching at the beginning of every topic became standard process for all remaining reviews done by Sullivan. The results of this exercise were used in future keyword searches as well as database-wide keyword highlighting.

Next, Sullivan started manually reviewing some of the hits on terms he felt would be most likely to yield responsive documents. Starting with 102 documents that hit on "TOR" and "anonym*" and moving on to hits on "TOR Browser," then "TOR" and "Prox*." It was not difficult to find a relatively high quantity of relevant documents. 108 relevant documents and 100 irrelevant documents were trained for predictive coding when the first learning session was run.

After the first learning session completed, Sullivan manually reviewed the highest scoring documents that contained the term "TOR" and found almost all to be relevant. At the

conclusion of the first day, 214 documents had been submitted to TREC, with all 214 being returned as relevant.

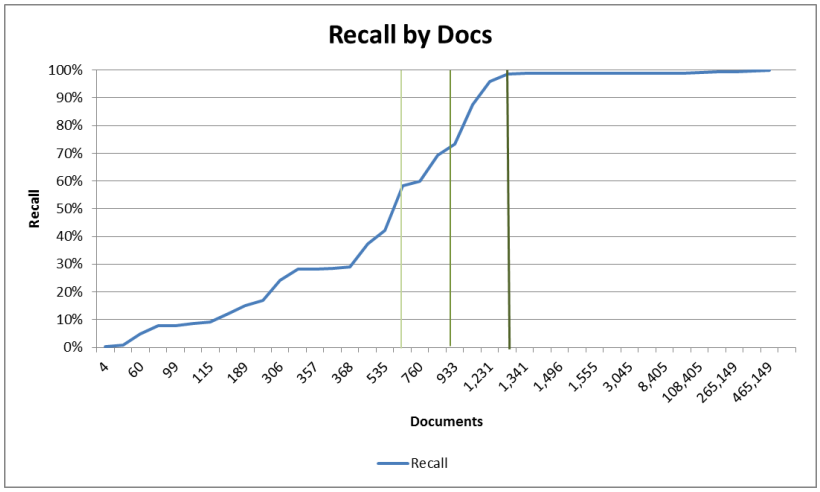
Day 2 consisted of many iterations of learning sessions and evaluating search results. Similar to how Sullivan reviewed Topic 2052, he started with a narrow list of keyword searches and broadened the terms iteratively. For each set, he reviewed the documents with the highest predictive coding scores. Starting the day with "TOR" and "prox*," he moved to "Try TOR," "Try using TOR," and "Use TOR." Eventually he moved to all documents that contained "TOR" or "T0R." Every document he determined to be relevant was submitted to TREC.

At the end of the exercise, Sullivan had submitted 1,339 documents, with 1,244 being returned as relevant and 95 being returned as not relevant according to the TREC standard. At this point he called his shot at Reasonable Recall.

Day 3 started with the submission of all remaining documents that contained the term "TOR" as a method to catch any documents potentially missed. No additional relevant documents were returned.

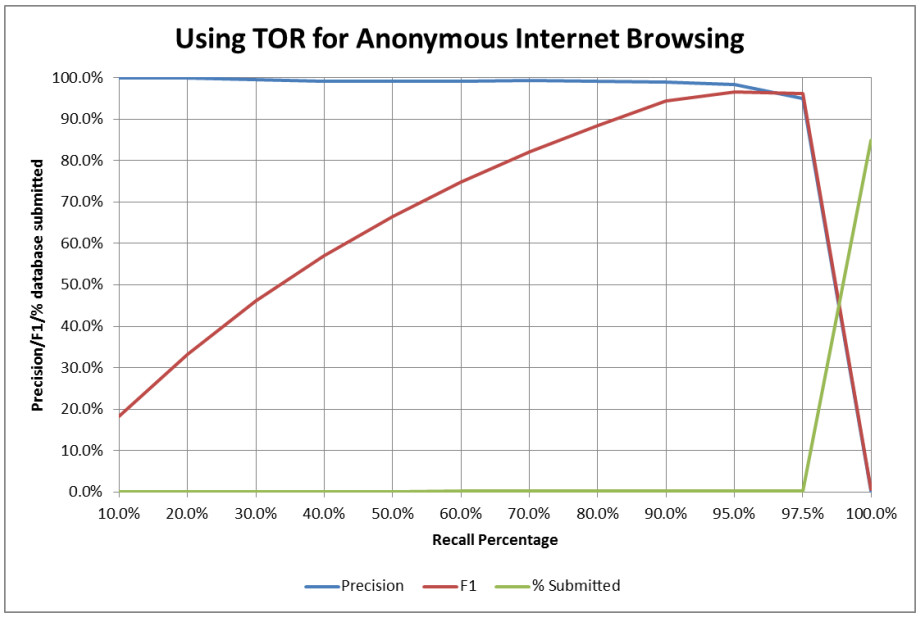
All remaining documents in the database were submitted in order of descending predictive coding score. 14 more relevant documents were returned. Evaluation of these documents led to finding spectacular errors in the TREC standard. All 14 contained "*tor*" in some context, but none had any even marginal links to the current topic. A majority of the missed documents contained the term "hostigator.com." Evaluation of these 14 documents resulted in a determination that all 14 were caused by an error in the TREC classification system.

A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable Recall call.

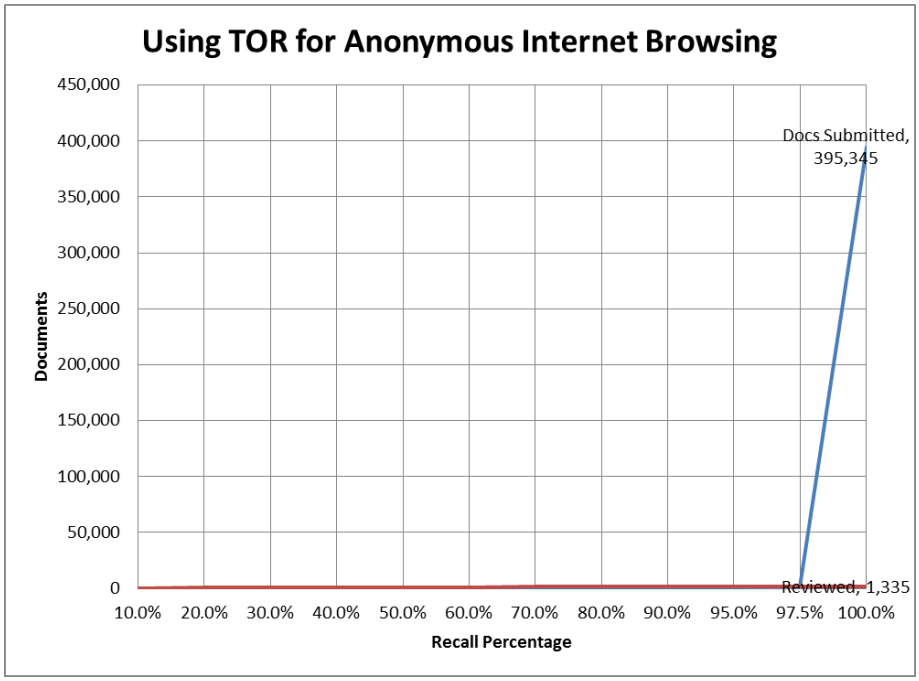


The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Using TOR for Anonymous Internet Browsing topic, by the time 97.5% Recall had been attained only 0.28% of the corpus,

1,294 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.72% or 463,855 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



TOPIC 104 New Medical Schools

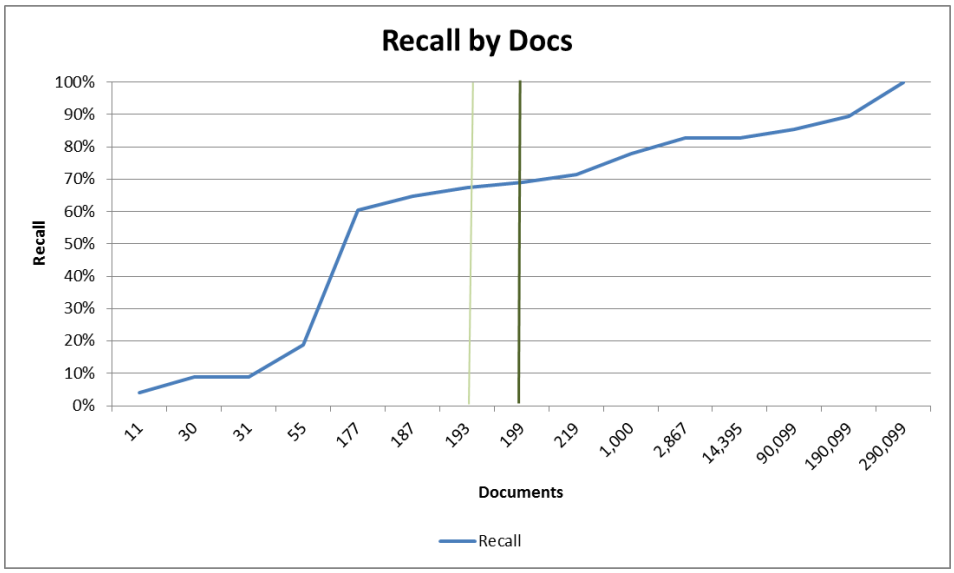
Confusion Matrix- Topic 104
Total Documents: 290,099
Total Relevant: 227
Total Prevalence: 0.08%

	@Reas. Call	@97.5% Recall
True Positives	157	222
True Negatives	289,830	51,763
False Positives	42	238,109
False Negatives	70	5
Recall	69.16%	97.80%
Precision	78.89%	0.09%
F1 Measure	73.71%	0.19%
Accuracy	99.96%	17.92%
Error	0.04%	82.08%
Elusion	0.02%	0.01%
Fallout	0.01%	82.14%

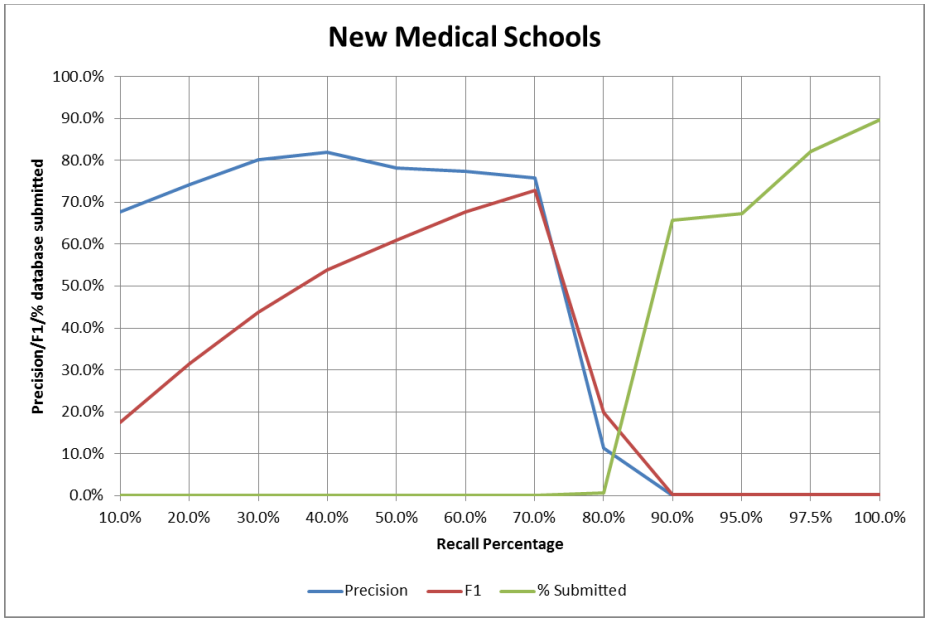
Topic 104 was run by Losey with the assistance of a review attorney, Jensen. The work to search the 290,099 *Bush Emails* started on July 31, 2015 and concluded on August 4, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal, was performed with the assistance at first of Jensen. Losey handled all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made suggestions of documents to submit. Again, all final decisions on submittal were made by Losey.

On August 3, 2015, after making 8 document submissions to TREC providing a total 199 documents, Losey had found a total of 157 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 1,091 documents. After the 8th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 69.16%** had been attained, with **Precision of 78.89%,** and **F1 of 73.71%**. He made the call decision a little prematurely on this Topic. In the next submission of only 20 documents, Losey brought the **Recall level up to 71.37% with Precision of 73.97%.** In the next submission of 781 documents he brought the **Recall level to 77.97%.** There were a total of 7 additional submissions to TREC after the *Reasonable* call point. After submitting a total of 1,611 documents, which is only 0.56% of the total documents, and reviewing only 1,091 documents, an **80% Recall** was attained.

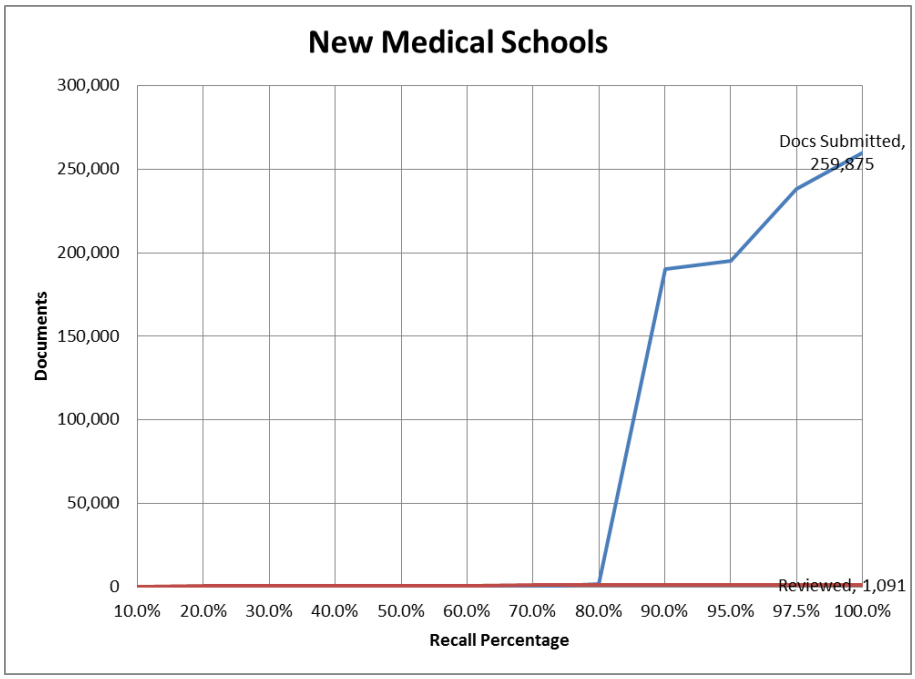
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the New Medical Schools topic, by the time 97.5% Recall had been attained 82.16% of the corpus, 238,331 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 17.84% or 51,768 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 109 Scarlet Letter Law

Confusion Matrix- Topic 109 Scarlet Letter Law

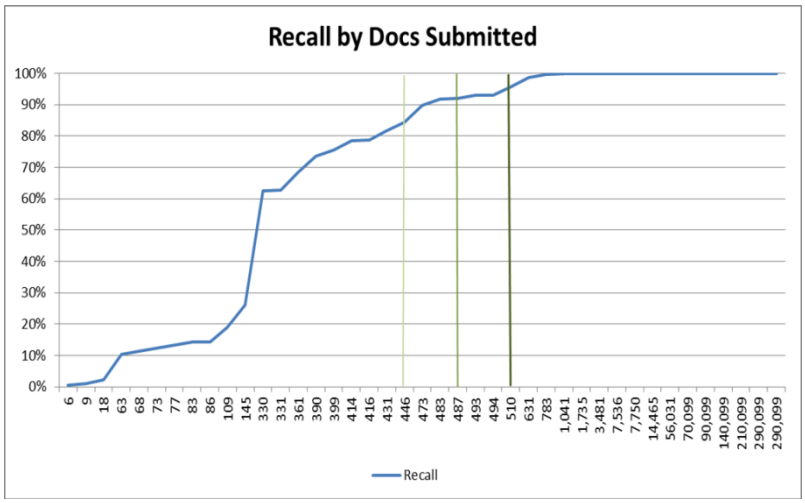
Total Documents: 290,099
Total Relevant: 506
Total Prevalence: 0.17%

	@Reas. Call	@97.5% Recall
True Positives	485	494
True Negatives	289,568	289,502
False Positives	25	91
False Negatives	21	12
Recall	95.85%	97.63%
Precision	95.10%	84.44%
F1 Measure	95.47%	90.56%
Accuracy	99.98%	99.96%
Error	0.02%	0.04%
Elusion	0.01%	0.00%
Fallout	0.01%	0.03%

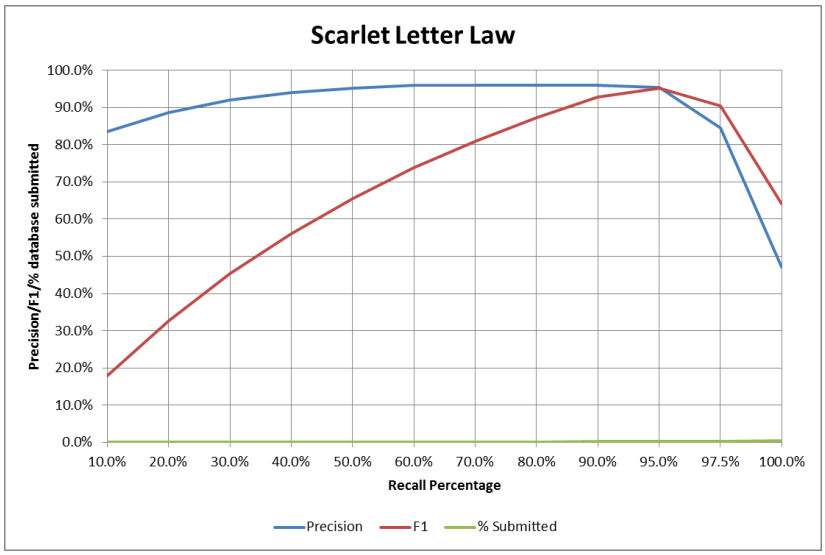
Topic 109 was run by Losey with the assistance of a review attorney, Bottolene. The work to search the 290,099 *Bush Emails* started on August 3, 2015 and concluded on August 11, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal, was performed with the assistance at first of Bottolene. Losey handled all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made suggestions of documents to submit. Again, all final decisions on submittal were made by Losey.

On August 11, 2015, after making 26 submissions to TREC providing a total 510 documents, Losey had found a total of 485 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 953 documents. After the 26th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 95.85%** had been attained, with **Precision of 95.1%**. There were 14 additional submissions to TREC after the *Reasonable* call point. In the next submission after the call of only 121 documents a Recall of 98.62% was attained. **Recall of 100%** was attained three submissions later after submitting only 1,074 documents, 0.37% of the total, and review of only 953 documents.

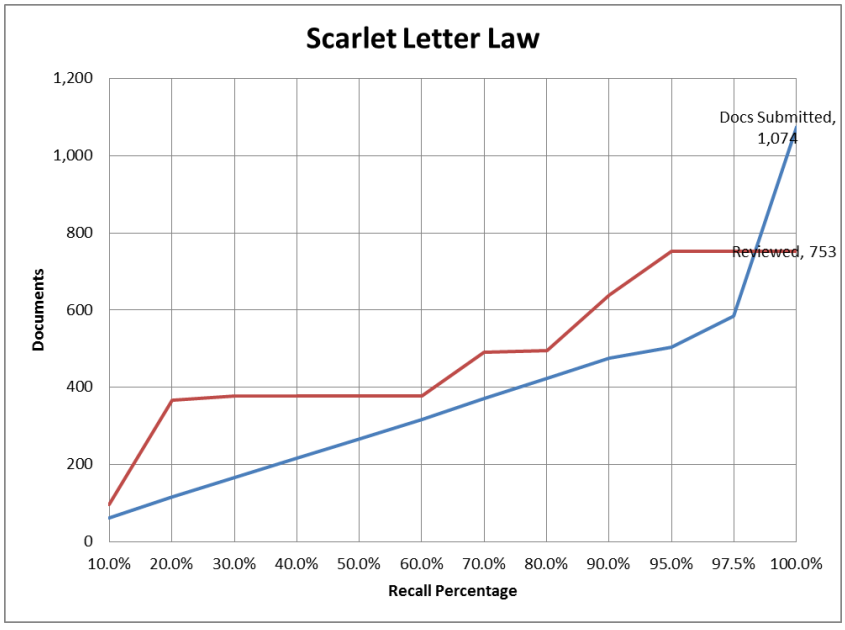
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Scarlet Letter Law topic, by the time 97.5% Recall had been attained only 0.20% of the corpus, 585 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.80% or 289,514 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 100 School and Preschool Funding

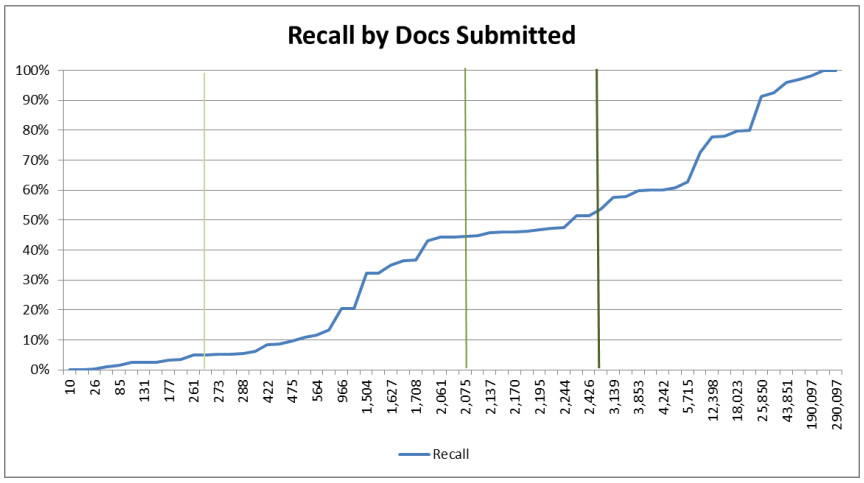
Confusion Matrix- Topic 100
Total Documents: 290,097
Total Relevant: 4,542
Total Prevalence: 1.57%

	@Reas. Call	@97.5% Recall
True Positives	**2,441**	**4,429**
True Negatives	**285,459**	**199,460**
False Positives	**96**	**86,095**
False Negatives	**2,101**	**113**
Recall	**53.74%**	**97.51%**
Precision	**96.22%**	**4.89%**
F1 Measure	**68.96%**	**9.32%**
Accuracy	**99.24%**	**70.28%**
Error	**0.76%**	**29.72%**
Elusion	**0.73%**	**0.06%**
Fallout	**0.03%**	**30.15%**

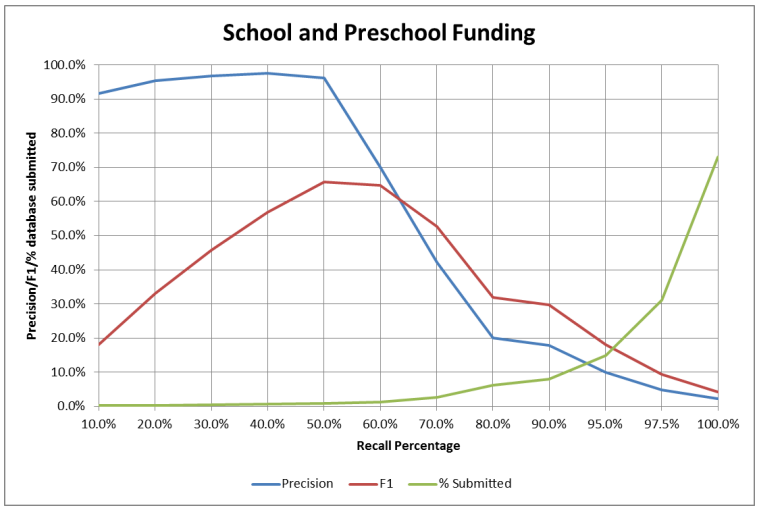
Topic 100 was run by Losey with the limited assistance of a review attorney, Jensen. The work to search the 290,099 *Bush Emails* started on August 4, 2015 and concluded on August 8, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal, was performed with some assistance at first of Jensen. Losey handled all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made a couple of suggestions of documents to submit. Again, all final decisions on submittal were made by Losey.

On August 6, 2015, after making 44 submissions to TREC providing a total 2,537 documents, Losey had found a total of 2,441 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 651 documents. After the 44th TREC submission, Losey decided to call *Reasonable*. This proved to be a premature call. It was later determined that a **Recall of 53.74%** had been attained, with **Precision of 96.22%,** and **F1 of 68.96%**. There were 19 additional submissions to TREC after the *Reasonable* call point. After submitting a total of 7,541 documents, which is only 2.6% of the total documents, and reviewing only 651 documents, a **70% Recall** level was attained. A **Recall of 80%** was attained after submitting 6.28% of the total documents, and **Recall of 90%** after submitting 7.92%.

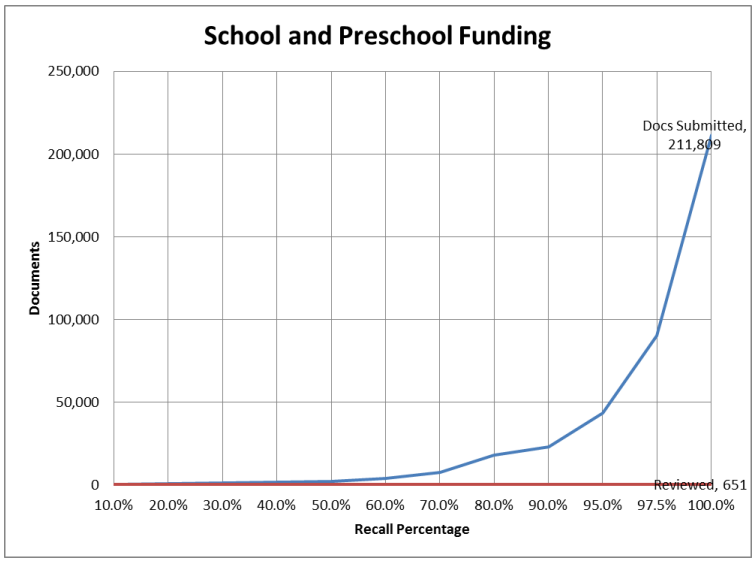
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the School and Preschool Funding topic, by the time 97.5% Recall had been attained only 31.20% of the corpus, 90,524 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 68.80% or 199,573 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 107 Tort Reform

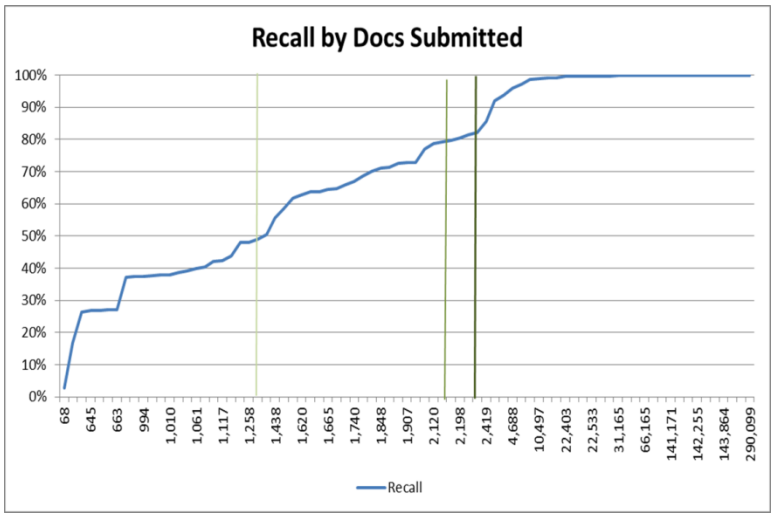
Confusion Matrix- Topic 107
Total Documents: 290,099
Total Relevant: 2,369
Total Prevalence: 0.82%

	@Reas. Call	@97.5% Recall
True Positives	**1,950**	**2,310**
True Negatives	**287,421**	**284,197**
False Positives	**309**	**3,533**
False Negatives	**419**	**59**
Recall	**82.31%**	**97.51%**
Precision	**86.32%**	**39.53%**
F1 Measure	**84.27%**	**56.26%**
Accuracy	**99.75%**	**98.76%**
Error	**0.25%**	**1.24%**
Elusion	**0.15%**	**0.02%**
Fallout	**0.11%**	**1.23%**

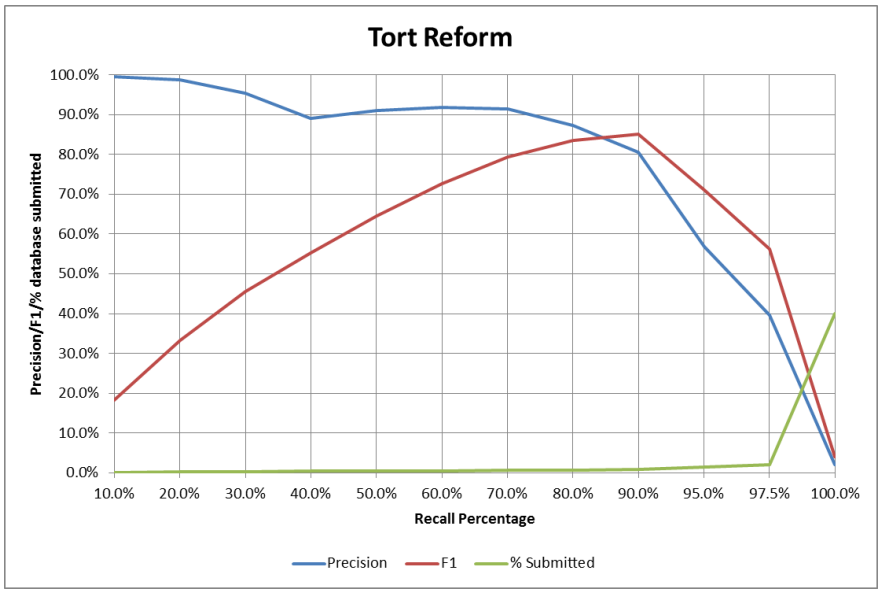
Topic 107 was run by Losey with the limited assistance of a review attorney, Jensen. The work to search the 290,099 *Bush Emails* started on August 5, 2015 and concluded on August 15, 2015. The project commenced with Losey and his assistant beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal, was performed with some assistance at first of Jensen. Losey handled all of the AI related searches in Step Five, including the probability and ranking related searches. His assistant focused on keyword searches and also made a couple of suggestions of documents to submit. Again, all final decisions on submittal were made by Losey.

On August 14, 2015, after making 48 submissions to TREC providing a total 2,259 documents, Losey had found a total of 1,950 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 1,164 documents. After the 48th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 82.31%** had been attained, with **Precision of 86.32%,** and **F1 of 84.27%**. There were 31 additional submissions to TREC after the *Reasonable* call point. After submitting a total of 2,648 documents, which is only 0.91% of the total documents, and reviewing only 1,164 documents, a **90% Recall** level was attained with **80.55% Precision**. Recall of 95% was attained after submitting 3,963 documents, 1.37% of total. Recall of 98% was attained after submitting 5,843 documents, 2.01% of total.

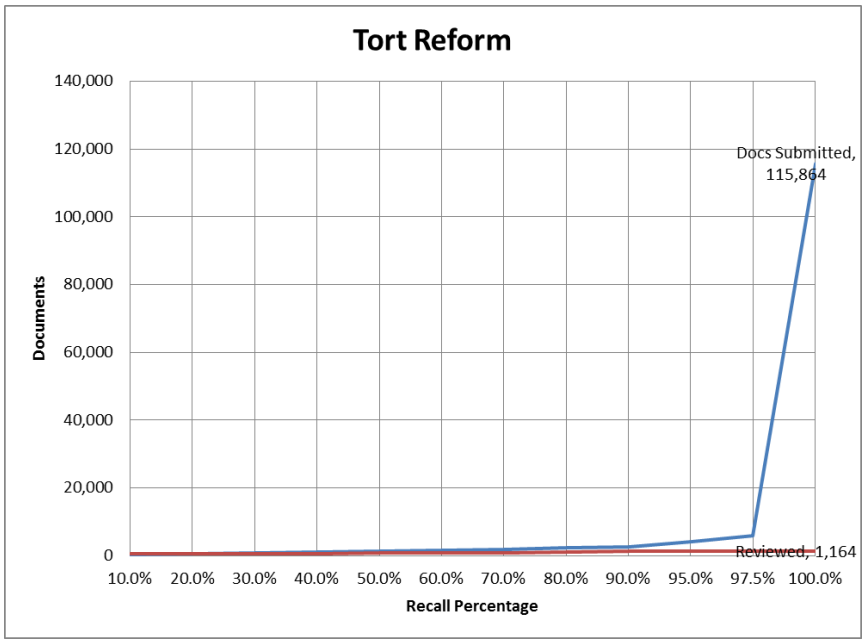
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Tort Reform topic, by the time 97.5% Recall had been attained only 2.01% of the corpus, 5,843 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 97.99% or 284,256 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



_____

Topic 3481 Fracking

Confusion Matrix- Topic 3481 Fracking
Total Documents: 902,434
Total Relevant: 1,966
Total Prevalence: 0.22%

	@Reas. Call	@97.5% Recall
True Positives	**1,893**	**1,917**
True Negatives	**900,284**	**899,841**
False Positives	**184**	**627**
False Negatives	**73**	**49**
Recall	**96.29%**	**97.51%**
Precision	**91.14%**	**75.35%**
F1 Measure	**93.64%**	**85.01%**
Accuracy	**99.97%**	**99.93%**
Error	**0.03%**	**0.07%**
Elusion	**0.01%**	**0.01%**
Fallout	**0.02%**	**0.07%**

Topic 3481 was run by **Sullivan** who started on August 4, 2015. He finished his review of 902,434 News Articles on Aug. 7, 2015 after 7 total hours of effort.

Sullivan had no background or knowledge of fracking prior to this exercise. While expert knowledge was not necessary, there were a few instances where some additional knowledge of the topic would have been helpful.

Sullivan had previously tackled topics in the forums data set, but this was his first topic in the News data set. He found the lack of spelling issues and overall consistency in the documents provided a much easier set of data to review. Much less manual review was necessary with the news topics.

On the first day, Sullivan used concept searching to identify similar topics, per his standard process. He created a list of most likely relevant keywords and used the list for searching and keyword highlighting. Both search and keyword highlighting lists were modified through the course of the review as new information was obtained.

Sullivan decided to go with a different approach to this topic. Rather than performing a manual review of documents to begin, he decided to submit as relevant any document that contained over 5 instances of the term "fracking" without review. 286 documents met this standard, and all were returned as relevant when submitted to TREC.

While the data used for this exercise did not contain any metadata, Sullivan determined any text that appeared in the first 2 lines of the document could be considered the document's title. He found 61 documents that contained "fracking" in the title and an additional instance of fracking elsewhere in the document. All 60 were returned as relevant, with 1 one not relevant. Further

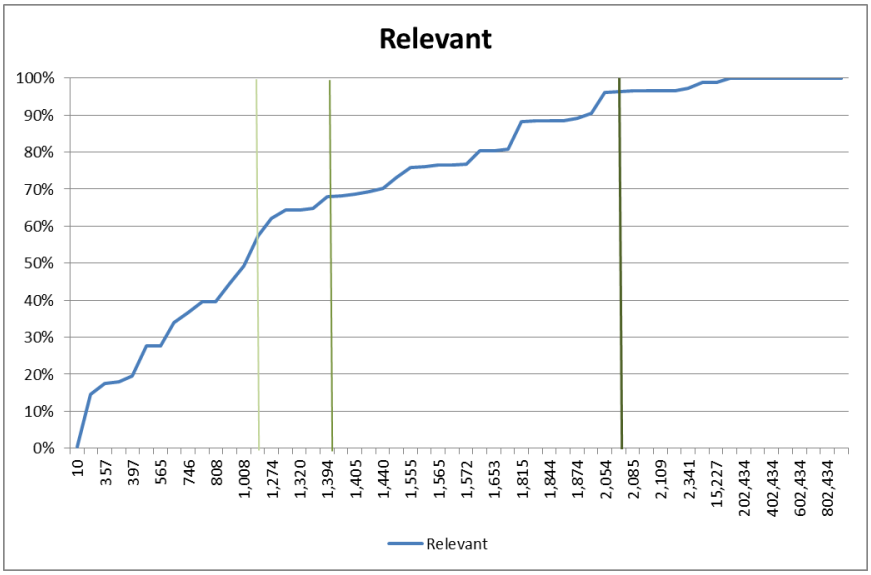
evaluation determined the not relevant document was an error in the TREC standard. Next, 9 documents were found which contained "hydrofracking" in the title. All 9 were returned as relevant. He then continued with slight variations until submitting all documents that contain 2 or more hits on the term "fracking." After 1 hour and manual review of 29 documents, 746 documents had been submitted with 745 being returned as relevant.

Sullivan continued manually reviewing the documents with a single hit on fracking to sort out the false positives. After reviewing a couple sets of documents, he initiated his first predictive coding learning session for this topic.

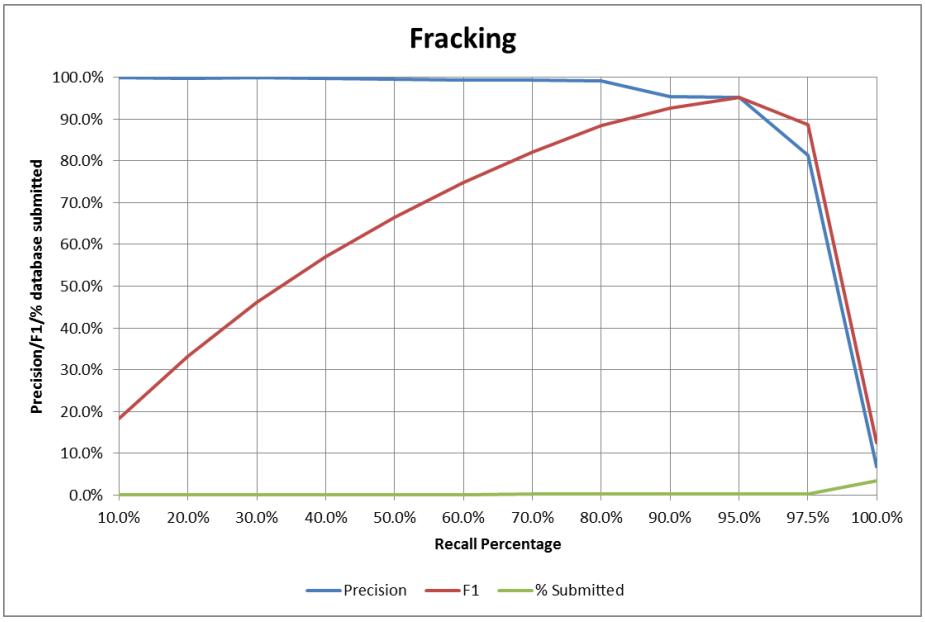
On the start of Day 2, Sullivan believed he had found nearly all relevant documents for this topic. However, after reviewing documents with high predictive coding scores, he quickly realized that "fracturing" was another key term he hadn't previously considered. The use of predictive coding helped him quickly find an additional 400 relevant documents that would have been lost if using keyword searching alone.

Reasonable Recall was called after submitting 2,077 documents, with 1,893 returned as relevant. The remaining documents were submitted in order of descending predictive coding scores, and 73 more relevant documents were returned. An evaluation of the returned documents contained many errors in the TREC standard, as well as a fair number of relevant documents that were not properly captured due to Sullivan's lack of knowledge of fracking and related mining terms.

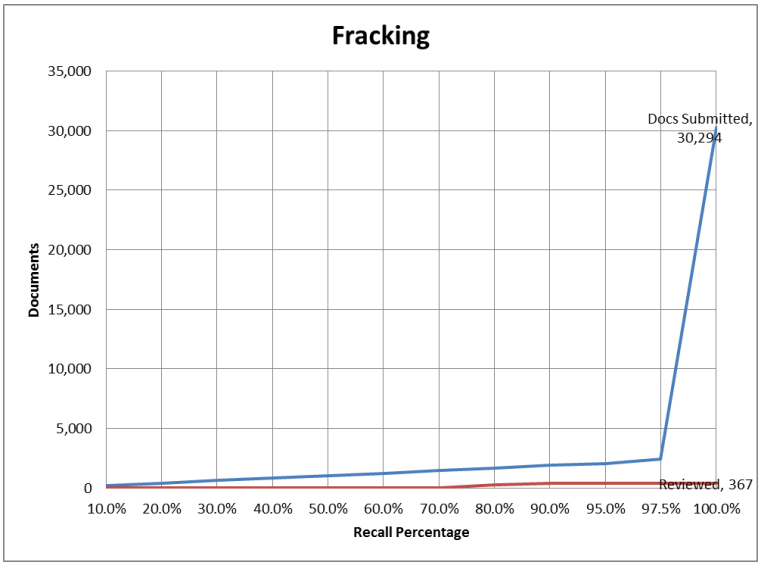
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Fracking topic, by the time 97.5% Recall had been attained only 0.27% of the corpus, 2,439 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.73% or 899,995 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multi-modal hybrid model of training EDR.



Topic 3431 Kingston Mills Lock Murders

Confusion Matrix- Topic 3431
Total Documents: 902,434
Total Relevant: 1,111
Total Prevalence: 0.12%

	@Reas. Call	@97.5% Recall
True Positives	**1,107**	**1,084**
True Negatives	**901,309**	**901,311**
False Positives	**14**	**12**
False Negatives	**4**	**27**
Recall	**99.64%**	**97.57%**
Precision	**98.75%**	**98.91%**
F1 Measure	**99.19%**	**98.23%**
Accuracy	**100.00%**	**100.00%**
Error	**0.00%**	**0.00%**
Elusion	**0.00%**	**0.00%**
Fallout	**0.00%**	**0.00%**

Topic 3431 was run by Reichenberger. The work to search the 902,434 *News Articles database* started on August 4, 2015, and was completed on August 5, 2015.

The initial submissions on the first day were to test the outlines of the category. The initial search of "Kingston" AND "murder" identified a sensationalized murder story about a man with the last name "Shafia" murdering his daughters in an "honor killing." Documents containing the information in various forms (headline, text, "clickbait" link reference at end of article) were submitted. Results helped formulate an anticipated rule on relevance.

After training Mr. EDR and receiving relevance priority scores, a search on the specific victim names or "Shafia" were sorted by prioritization order. Samples of 10 documents above 90%, 10 between 80-90%, 10 between 60-80%, 10 between 25-60% and 10 below 25% showed that documents above 60% were very likely relevant. In fact, documents scoring over 90% all had multiple name hits and were specifically on point; documents in the middle ranges were usually indirectly related (e.g. about "honor killing," or domestic abuse, or more of a casual reference to the Kingston Mills murders); and those documents below 5% were almost always irrelevant. As a test, the second submission contained all documents with a score over 90%, along with samples of several documents at various scores greater than 50%, cutting the submission off at 200 documents even. With only 111 documents reviewed eyes on to this point, Reichenberger had a 98.5% precision on 205 documents submitted.

Of the 205 documents submitted to this point, the only 3 irrelevant documents all had the same trait: "Shafia" appeared in the header but there was no reference to it in the text. Similar documents were mass coded as irrelevant going forward. Likewise, people with names similar to the victims were found in the 40-60% probability range but were "false positive" documents. These included an AP photographer, the President of Gambia, and protesters in Yemen with first

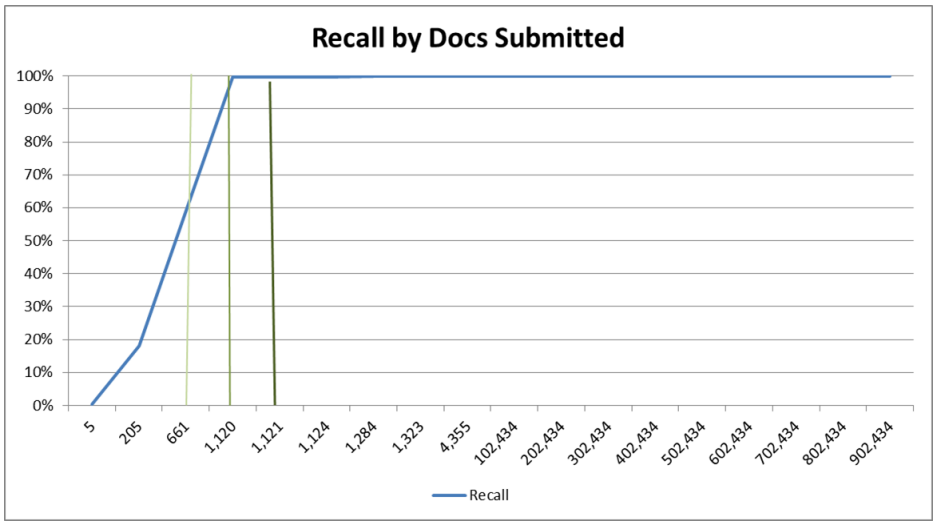
names the same as one of the victims. Searches were done on those specific names and mass-tagged as irrelevant. After a machine learning session, the scores adjusted dropping those false positive names to the bottom. At this point, a sampling of key term hits showed everything over 20% scores were relevant, and everything below 1% were irrelevant. Everything in between were low quality references to the murders with some irrelevant documents mixed in. As such, the next submission was for everything with a key term over 25% relevant score (456 documents) of which 449 were found relevant. The 7 documents found irrelevant were misclicks by Reichenberger (human error). In one case a document was primarily about a different murder, but later in the article there was relevant discussion of the target murder. Mr. EDR picked this up, but it was apparently missed by TREC's relevance scope adjudications. The 70% Recall call was then made having reviewed only 209 documents. It turned out that Recall was actually 58.6% with Precision at 98.5%.

The next submission consisted largely of documents containing a single line of "clickbait" link text found by TREC to be relevant. Other documents considered were documents with key terms that had scores raise above 20% following the machine learning session from the previous set and documents with scores above 50% with no key terms. While documents with key terms were largely found to be relevant, most of the documents without the terms were found to be irrelevant. In fact, documents scoring above 70% were often tangential to the issues in the murder (domestic violence mostly) but not relevant, while those 50-70% had no semblance of relevance at all, and were being escalated based on coincidental "clickbait" text advertisement lines at the end of the article. Another 459 documents were submitted with 456 were found relevant. The three irrelevant documents all were on the low end scores within the submission and were only passing references to the case. At this point the 80% recall call was made. **Recall was actually at 99.64% with a precision at 99.34%.** Only 272 documents were reviewed eyes on to this point, and 1120 relevant documents had been found. All documents with scores over 70% had been reviewed or submitted, and all those with key terms and scores over 20% had been reviewed or submitted.

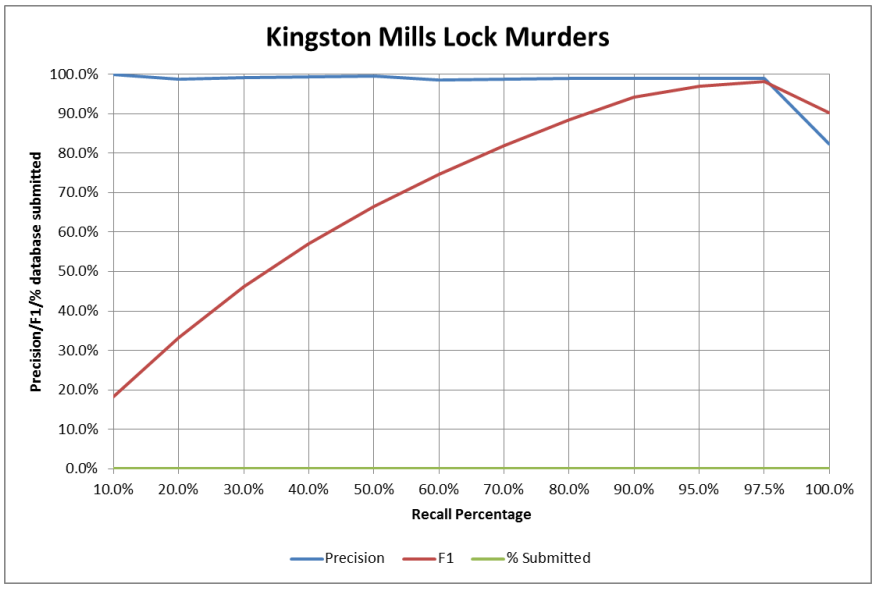
Following the subsequent machine learning session, 30 documents were escalated to consider. One borderline document was considered potentially relevant and submitted, returned as irrelevant, while the rest all marked irrelevant. The Reasonable call was made.

After the Reasonable call was made documents were submitted in the following groups in descending priority score order: 1) three documents potentially relevant found while pending results of the previous submission (one was found to be relevant) 2) all documents reviewed eyes on anticipated to be irrelevant, but not yet submitted (199 documents, of which two were relevant and the only relevant text within these two documents were contained in a document previously submitted to TREC and returned as irrelevant); 3) anything mass-coded as irrelevant (this resulted in one relevant document, of which there does not appear to be any relevant material within it and may be yet another TREC coding error); and 4) anything remaining (all irrelevant).

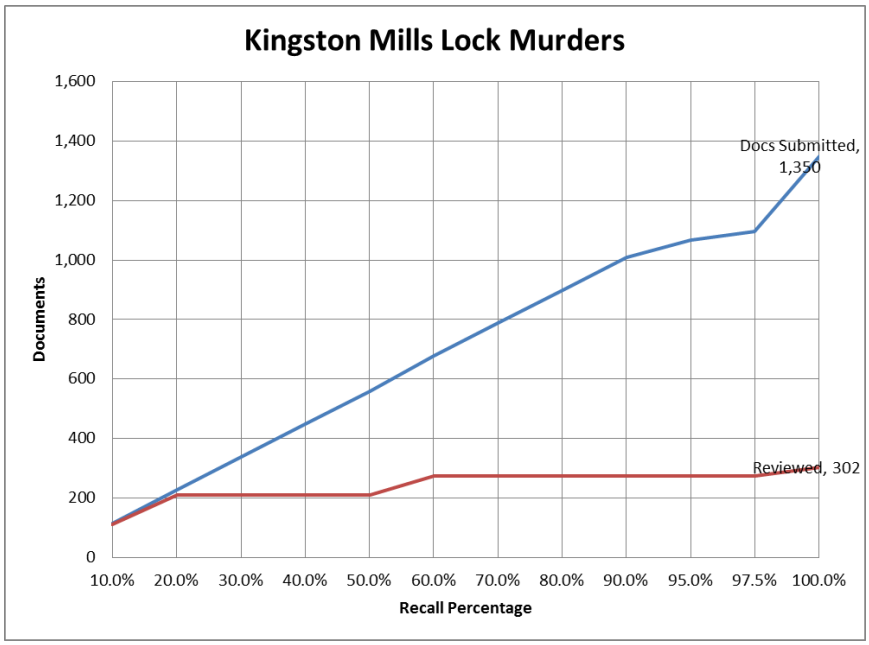
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the Reasonable call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Kingston Mills Lock Murders topic, by the time 97.5% Recall had been attained only 0.12% of the corpus, 1,096 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.88% or 901,338 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the Multimodal Hybrid model of training Mr. EDR.



Topic 2130 Surely Bitcoins Can Be Used

Confusion Matrix- Topic 2130

Total Documents: 465,147
Total Relevant: 2,299
Total Prevalence: 0.49%

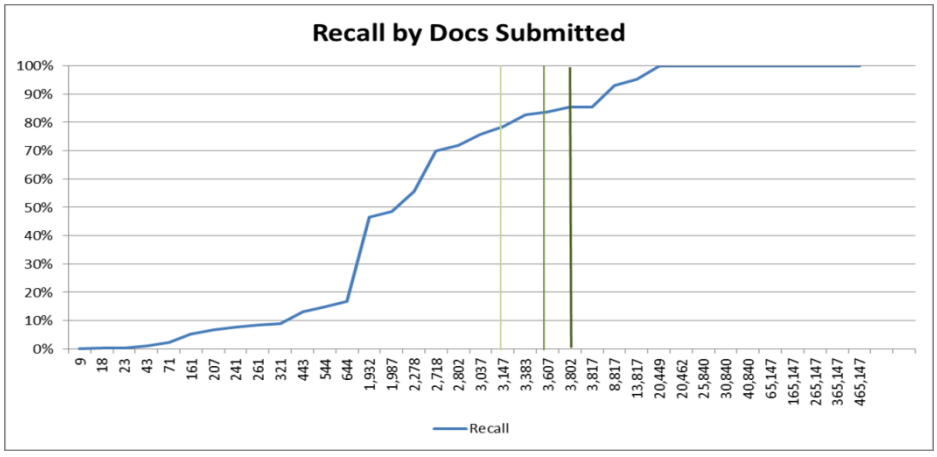
	@Reas. Call	@97.5% Recall
True Positives	1,961	2,242
True Negatives	461,007	448,083
False Positives	1,841	14,765
False Negatives	338	57
Recall	85.30%	97.52%
Precision	51.58%	13.18%
F1 Measure	64.29%	23.23%
Accuracy	99.53%	96.81%
Error	0.47%	3.19%
Elusion	0.07%	0.01%
Fallout	0.40%	3.19%

Topic 2130 was run by Reichenberger. The work to search the 465,147 documents in the *BlackHat World Forums* database started on August 7, 2015 and was completed August 13, 2015.

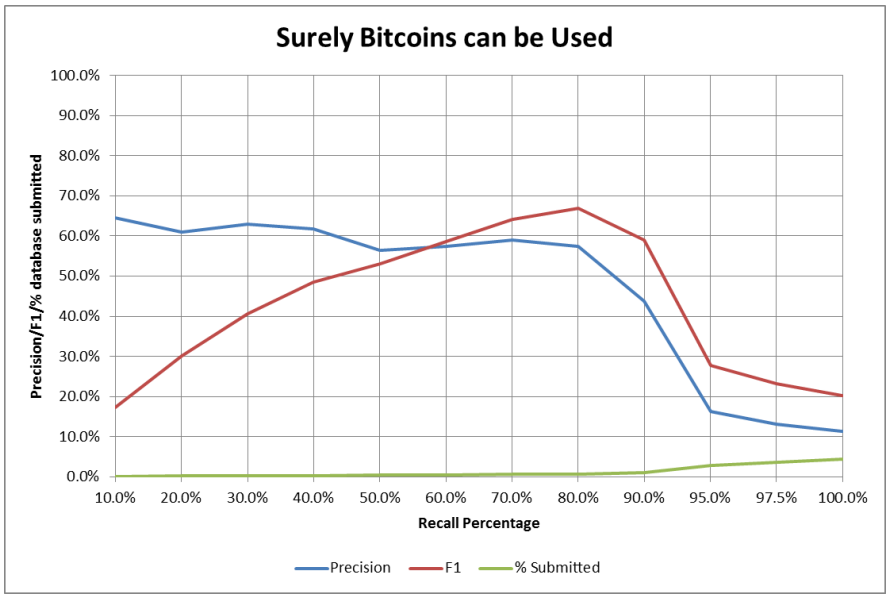
The initial submissions were to test the outlines of the category. The first submission was nine documents with varying discussions about Bitcoin (e.g. bitcoin exchanges, whether bitcoin was accepted, bitcoin mining, etc). All nine came back as irrelevant. A second submission of nine returned five relevant documents but no noticeable commonality among them except that "accept bitcoin" was relevant and "accept bitcoins" was not. The next 25 documents submitted also followed this trend, with singular "accept bitcoin" being relevant, those in the plural being irrelevant. All documents with "accept w/3 bitcoin" were submitted in the following two submission sets; however, having that text was not indicative of relevance, as some still came back irrelevant. Likewise, a variation of bitcoin ("BTC") was submitted (15 relevant, 5 irrelevant, no consistent thread).

After a machine learning session, the submitted documents were revisited and it appeared using bitcoin for legal activity or someone vouching for a forum user tended to be relevant, while illegal or immoral activity were irrelevant. For the next submission, the 60 highest scoring documents were submitted and anticipated as relevant/irrelevant based on the purpose of the transaction. While not perfect, this largely correlated with the results. (10 expected relevant, end result was 13). The next submission contained all documents with a 90% or higher probable relevant score and containing the term "vouch*". Of the 122 documents, 94 were relevant.

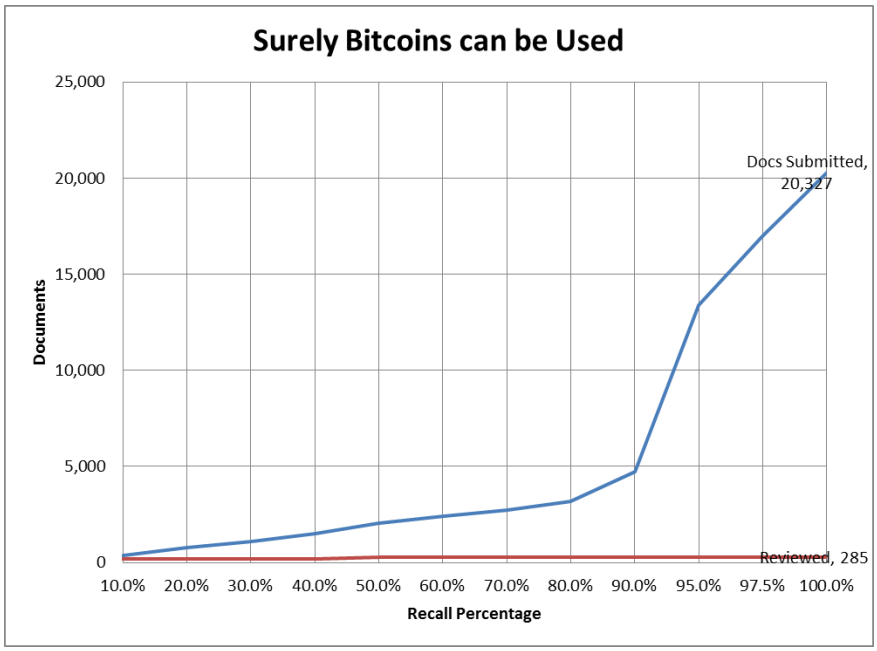
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Surely Bitcoins can be Used topic, by the time 97.5% Recall had been attained only 3.66% of the corpus, 17,007 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 96.34% or 448,140 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the Multimodal Hybrid model of training Mr. EDR.



Topic 3089 Pickton Murders

Confusion Matrix- Topic 3089
Total Documents: 902,434
Total Relevant: 255
Total Prevalence: 0.03%

	@Reas. Call	@97.5% Recall
True Positives	**236**	**249**
True Negatives	**902,164**	**901,971**
False Positives	**15**	**208**
False Negatives	**19**	**6**
Recall	**92.55%**	**97.65%**
Precision	**94.02%**	**54.49%**
F1 Measure	**93.28%**	**69.94%**
Accuracy	**100.00%**	**99.98%**
Error	**0.00%**	**0.02%**
Elusion	**0.00%**	**0.00%**
Fallout	**0.00%**	**0.02%**

Topic 3089 was run by Joe White. Work on this topic commenced on August 5, 2015 and concluded on August 28, 2015. Approximately 24 hours were spent on this topic, including a few hours up front researching the subject matter. This served as a proxy for the e-Discovery Team Hybrid Multimodal Model, Step 1, *ESI Discovery Communications*. Completion of this Topic was drawn out due to time conflicts including vacation.

The collection of 902,434 *News Articles* were generally easier to search than the *Bush Emails* or *Black Hat World Forum* posts, though the news articles contained many links, footers and subject matters that were shared with other news stories, creating the appearance of similarity. As would be expected with news articles, misspelled words and names seemed nonexistent, which was helpful. White did, however, find a few gold-standard inconsistencies in this topic.

White began Step Two, multimodal search, by creating several keyword lists based on his judgment and notes from the initial topic research. This research included events, names, locations, and other information related to the case. The keyword list goals were to: (a) to create a seed set to begin finding the potentially relevant documents and to begin training Mr. EDR; (b) to guesstimate how large the relevant document set would be a kind of rough substitute for Step Three Sample; and (c) to highlight relevant terms in the software to facilitate more effective review and training. (Note – all reviewers so highlighted certain keywords as a matter of course to speed up and improve review.)

When the initial keywords brought back only just over 220-some documents, while still cognizant of the limitations of keyword search, White believed this meant a relatively small potential data set existed. This afforded him the ability to perform a linear review of all of the keyword hits, but also meant that precision would be easily harmed by false positives. For that reason White knew that care would be needed in ascertaining true relevance. A normal Step 3,

initial *Random Baseline* sample, was omitted given the likely low prevalence and general time constraints for the work.

Based on the initial judgmental sample reviews in Step Two, White submitted initial sets of documents to TREC to establish relevance boundaries and begin whittling down on the set of relevant candidate documents. A minor loss of precision was anticipated on certain documents in exchange for knowledge that would guide subsequent submissions. Each time documents were determined to be relevant, White updated the training and predictive ranking, to facilitate priority-driven review that augmented the judgmental sampling work (*see* steps Four, Five and Six: *AI Predictive Ranking, Multimodal Search Review* & *Hybrid Active Training*). He also utilized conceptual search (predominantly Find Similar, via LSI) to branch off particularly interesting or novel documents to learn more. Although White, like all of the reviewers, did use concept search, and similarity search, he found that the predictive coding rankings (using a more robust technology) proved to be more effective overall. All reviewers had the same experience.

During the initial part of the submission process, White trained on all documents deemed relevant or irrelevant by TREC. This helped create additional separation in the model and rankings. In one instance he left one obvious TREC mistake trained as relevant (a duplicate of another document that had been adjudicated relevant) in order to ensure he would find any others like it.

During the predictive analysis and training, White found it was most helpful to review certain sets of documents from the bottom-up, to analyze the least-likely candidates in cases where relevance seemed clear. In other sets of documents, where relevance seemed less certain, White reviewed from the top-down. After additional analysis was completed and 99 documents had been submitted to TREC, White predicted there would be 200 – 250 relevant documents in total. (In the end, he would learn there were 255 total relevant documents in this topic, so the early prediction turned out to be quite close.) White also used random sampling in one instance, to train a set of 100 documents that seemed clearly irrelevant. These documents assisted Mr. EDR in separating irrelevant docs from relevant ones at a point early in the process when only relevant documents had been trained. This was part the Team's experimentation of the ideal ratios of irrelevant to relevant in training models.

As is almost always the case with an iterative training process, as the training and learning commenced, additional relevant subject areas came to light. While almost all of these areas were somewhat apparent from the start, fascinating and subtle nuances emerged. News stories on the case took little turns and spawned entirely new areas of relevance unto themselves. White thought the biggest challenge with these documents wasn't as much about whether they existed or how to locate them, but about whether TREC would see them as relevant or not. He found that it helped to track each pocket of relevance as a separate subject area, to utilize keywords for each subject area to create small seed sets, and to then utilize the predictive rankings within each subject area to dive deeper and ensure that each was adequately explored.

White made a total of 56 document submissions to TREC in this topic: 6 submissions between Aug. 6th and 12th, encompassing 184 documents, 22 submissions between Aug. 21 and 27th, encompassing 284 documents, and the remaining 28 submissions on Aug. 28th, encompassing 901,966 documents. In between most of these submissions he conducted iterative steps Four, Five and Six of the standard workflow, utilizing predictive ranking, search, and training.

After 218 documents had been submitted and additional priority-ranked documents and top keyword sets had been evaluated, White called 70%. There was still a fair quantity of suspected borderline documents in-hand, but his intuition was that he had probably surpassed 70% by a fair margin and so needed to call the shot. Actual Recall at this point turned out to be 83.53%.

White then studied closely the suspected borderline documents before he decided to submit them. He was attempting to determine the scope of relevance for these subject areas. After locating what he believed to be the full extent of the subject, and having found 23 more relevant documents, he called the 80% shot. White believed he was even farther along than 80%, given the ranked results he was seeing. As it turned out the actual Recall at this point was 92.55%.

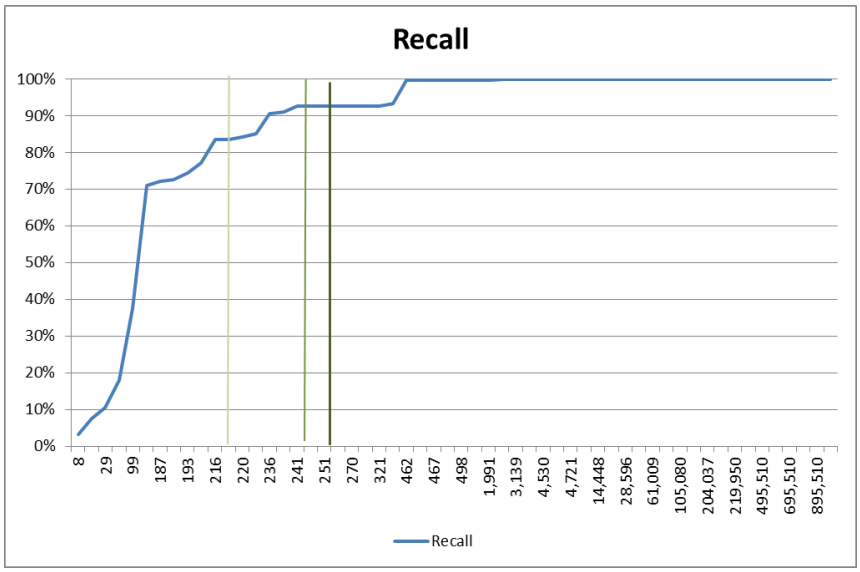
After submitting 8 more documents that he thought might be considered relevant, but were close questions and probably would not, White called *Reasonable*. This was with 251 total documents submitted, 236 of them relevant, and only 779 documents reviewed. Actual Recall at this point was still 92.55%.

Having called Reasonable and finding nothing new that looked relevant, White turned to his pool of remaining documents that looked irrelevant, to allow the predictive ranking to help him being submitting them. Indeed, Mr. EDR helped see things he could not, and soon found 18 additional documents that contained an oblique reference to a subject related to the case. While these documents seemed just as oblique as others that were deemed irrelevant, the fact that the predictive rankings caught them quickly was reassuring. After an additional round of training and predictive ranking turned up no additional documents, the submissions continued.

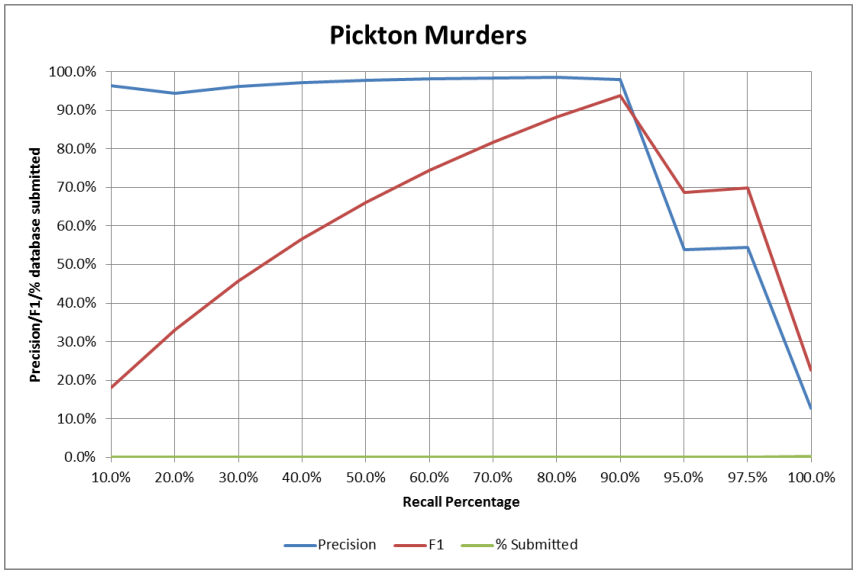
Finally, at the 2,000th document submitted, a "relevant" document was discovered that completed the 255-doc set. This document appeared to be a clear mistake, as it was only a reference to an unrelated London, UK murder. After that, all remaining documents submitted were confirmed as irrelevant.

On August 28, 2015, after making 19 submissions to TREC providing a total 251 documents, White had found a total of 236 relevant documents. The effort, or number of documents reviewed and coded by White to attain this result, was 834 documents. After the 18th TREC submission, White decided to call *Reasonable*. It was later determined that a **Recall of 92.55%** had been attained, with **Precision of 94.02%.** There were 37 additional submissions to TREC after the *Reasonable* call point. After submitting a total of 462 documents, which is only 0.05% of the total 902,434 documents, and reviewing only 834 documents, a **99.61% Recall** level was attained with **54.98% Precision**. **100% Recall** with **12.75% Precision** was attained after submission of 2,000 documents, which is 0.22% of the total.

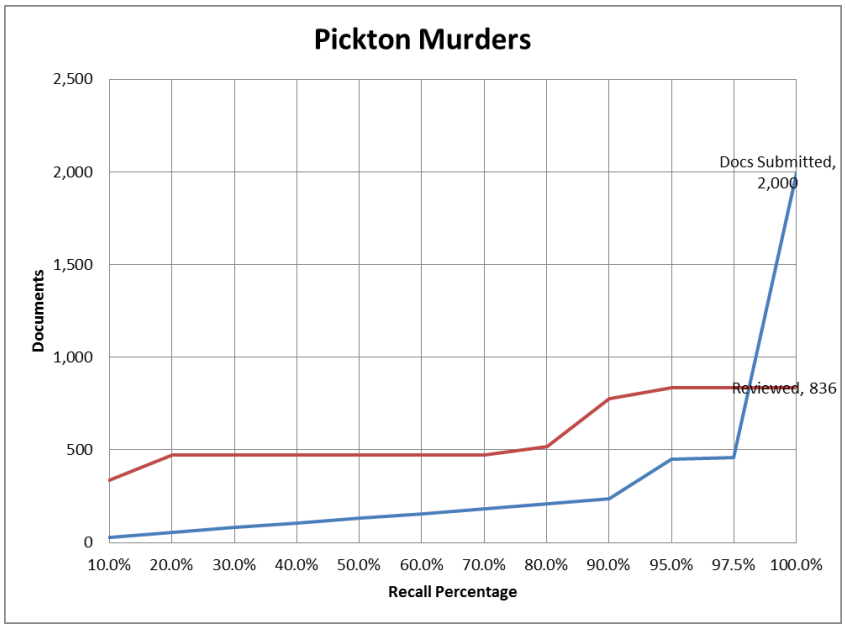
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Pickton Murders topic, by the time 97.5% Recall had been attained only 0.05% of the corpus, 457 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.95% or 901,977 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 2461 Offshore Host Sites

Confusion Matrix- Topic 2461 Offshore Host Sites
Total Documents: 465,147
Total Relevant: 179
Total Prevalence: 0.04%

	@Reas. Call	@97.5% Recall
True Positives	175	175
True Negatives	463,225	463,408
False Positives	1,743	1,560
False Negatives	4	4
Recall	97.77%	97.77%
Precision	9.12%	10.09%
F1 Measure	16.68%	18.29%
Accuracy	99.62%	99.66%
Error	0.38%	0.34%
Elusion	0.00%	0.00%
Fallout	0.37%	0.34%

Topic 2461 was run by Sullivan who started on August 14, 2015.

He finished his review of 902,434 News Articles on Aug. 15, 2015 after 5.0 total hours of effort.

Sullivan's background and knowledge in host sites was expected to be helpful in this topic, but in reality it worked against him. While he does not consider himself to be a subject matter expert on this topic, he has a solid level of knowledge with host sites. This proved difficult, because he thought he knew what documents should be considered relevant, but the TREC gold standard disagreed with most of his determinations.

Per his standard process, Sullivan started with concept searching to identify popular keywords to use as highlighting and future searches. This generated a long list of terms relating to different hosting sites and VPNs.

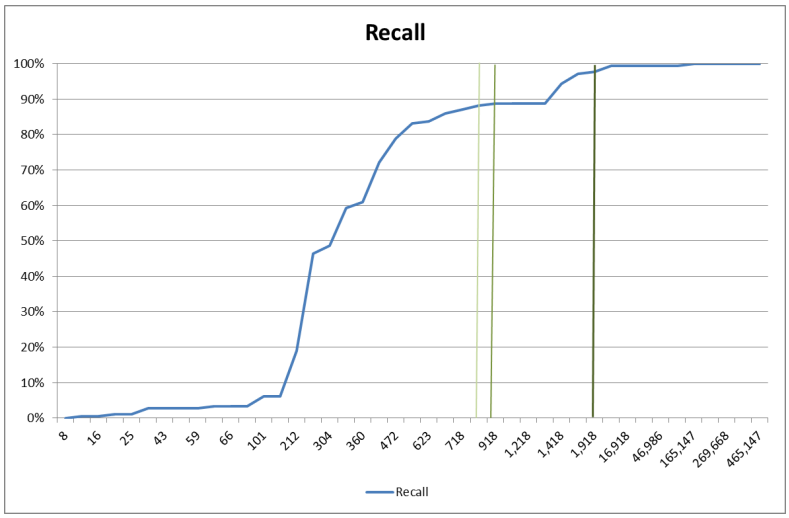
Sullivan continued with the next step of finding some documents to seed for predictive coding and get an understanding of the TREC line for relevance. He found 8 documents that hit on "offshore host* site*" and contained clearly relevant content by his definition. TREC determined all 8 to be not relevant. He then found 5 documents that relate to specific offshore hosting sites, such as *hosting panama* and *anon hoster*. TREC returned 1 relevant and 4 not relevant. He continued to try different variations of terms relating to hosting is specific countries and documents with different types of content and could not find any logic to the TREC relevance standard. Frustrated, he initiated a learning session and took a break.

Upon returning, he decided to try a test submission of 29 top scoring documents that contained the text "offshore" w/2 "host" without looking at any of the documents. To his surprise, 26 of the documents were returned by TREC as relevant. In a review of the documents, he saw no difference between the content of the TREC relevant documents and the documents he found and submitted that were returned as not relevant. The only general correlation he was able to identify is the TREC standard appeared to favor smaller sized documents with a higher proportion of content dedicated to offshore host sites. A document with a single line discussing offshore host sites was more likely to be relevant than a document with 50 lines and 10 references.

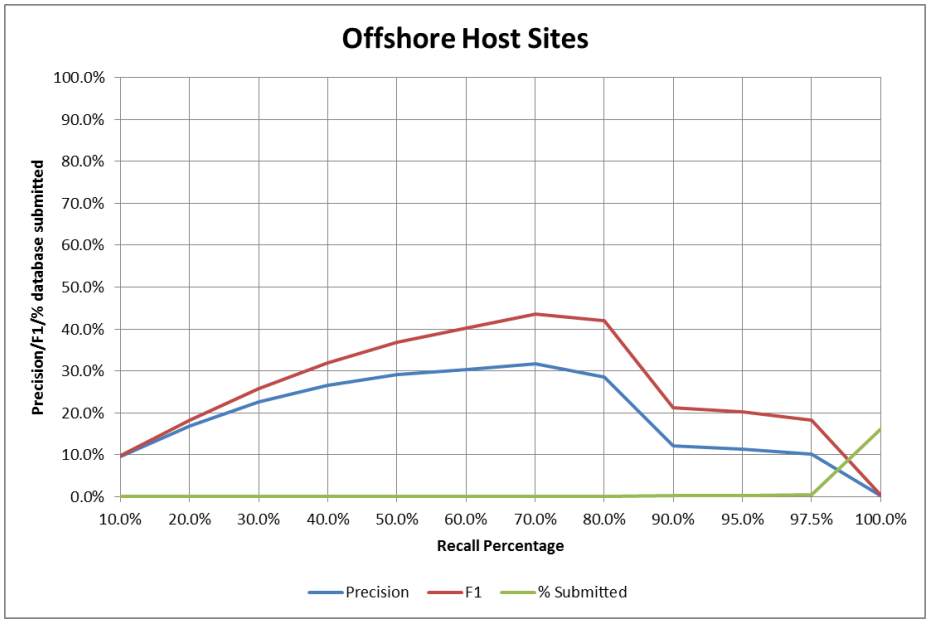
Being unable to determine any reasonable connection between content and relevance, Sullivan had no choice but to continue riding Mr. EDR's suggestions for documents to submit. This process consisted of many iterations of learning sessions and searching. Similar to how Sullivan reviewed Topic 2052 and 3481, he started with a narrow list of keyword searches and broadened the terms iteratively. For each set, he submitted the documents with the highest predictive coding scores. Starting with "offshore" w/2 "host*," he moved to "offshore" and "host," "offshore" and "web," and "offshore" and "vpn." Eventually he moved to all documents that contained "offshore" or "hosting." The difference between this process and what was used in prior reviews is Sullivan did not actually look at any of the documents. As he found his judgment to be out of line with the TREC standard, documents were submitted without review. Results of a search would be taken and the top documents would be submitted. If most were determined to be relevant, lower sets of documents from the result would be submitted until a low amount of relevant documents were returned. He would then move on to the next search and repeat.

After exhausting all of the all key terms, Sullivan submitted all remaining documents in descending priority order.

A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Offshore Host Sites topic, by the time 97.5% Recall had been attained only 0.37% of the corpus, 1,735 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.63% or 463,412 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multi-modal hybrid model of training EDR.

_____

Topic 3290 Rooster Turkey Chicken Nuisance

Confusion Matrix- Topic 3290
Total Documents: 902,434
Total Relevant: 26
Total Prevalence: 0.00%

	@Reas. Call	**@97.5% Recall**
True Positives	23	26
True Negatives	902,336	885,020
False Positives	72	17,388
False Negatives	3	0
Recall	88.46%	100.00%
Precision	24.21%	0.15%
F1 Measure	38.02%	0.30%
Accuracy	99.99%	98.07%
Error	0.01%	1.93%
Elusion	0.00%	0.00%
Fallout	0.01%	1.93%

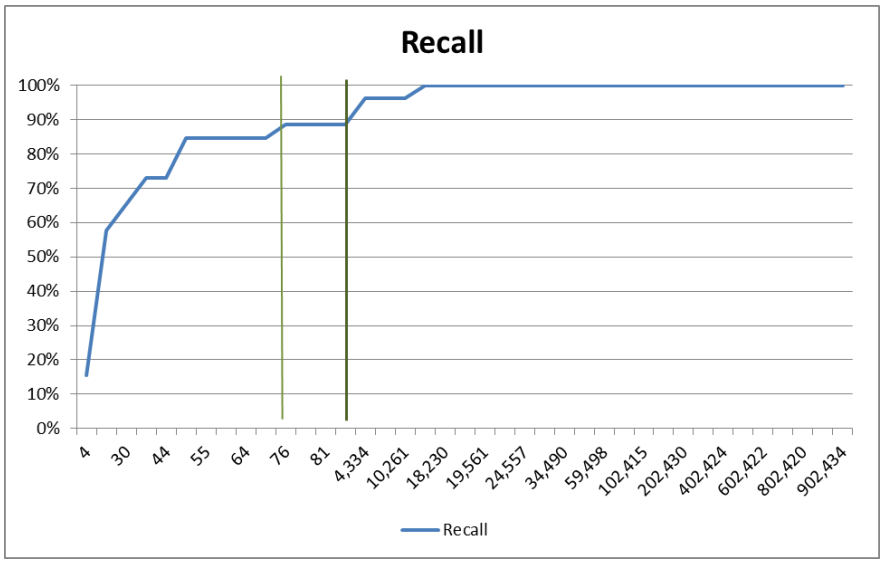
Topic 3290 was run by Losey alone who started on August 15, 2015 and concluded on August 23, 2015. The project commenced as usual with Losey beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal search began, including predictive coding features, with iterated training.

On August 22, 2015, after making 14 submissions to TREC, and training after almost every submission, Losey had provided a total of 95 documents to TREC and confirmed a total of 23 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 306 documents. After the 14th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 88.46%** was attained by submission of only 95 documents, which is 0.01% of the total 902,434 documents. This was accomplished by review of only 0.03% of the total collection.

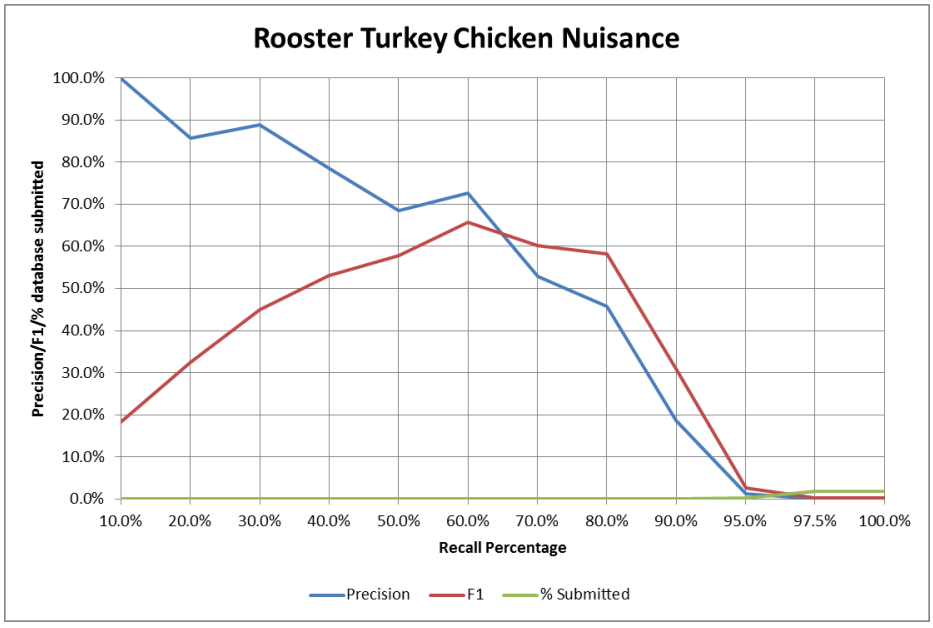
There were 23 additional submissions to TREC after the *Reasonable* call point. In the next submission after Reasonable call, the 15th, the Recall level rose to 96.15%. Recall of 100% was attained after submission of only 0.15%.

A 90% Recall was attained after submitting only 129 documents. A 95% Recall was attained after submitting 1,923 documents, and 97.5% Recall attained after 3,188 documents. Total Recall was attained after submitting 17,414 documents out of the corpus total of 902,43 (0.15%).

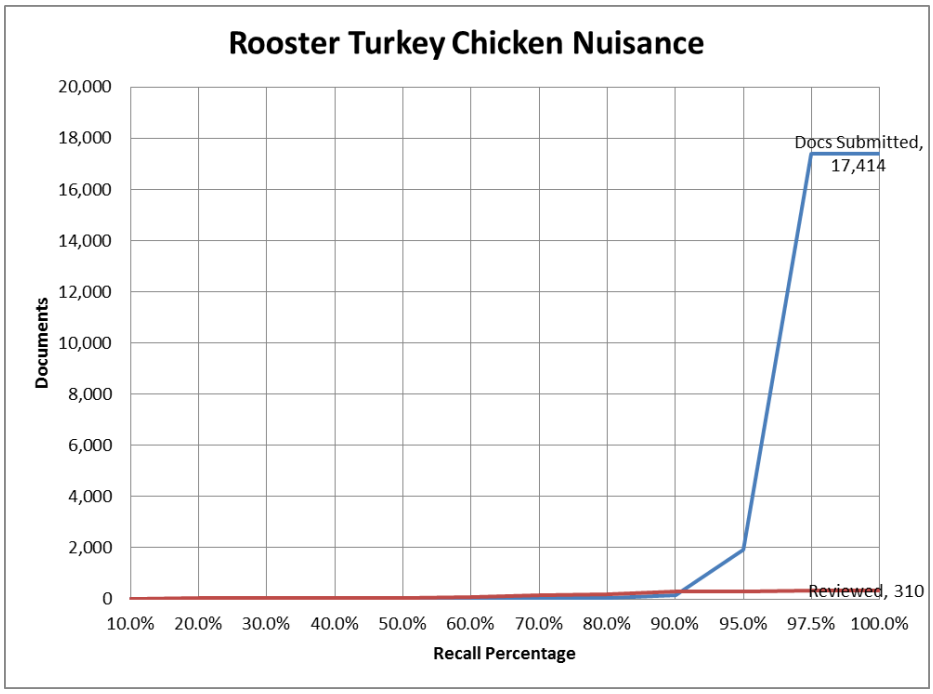
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Rooster Turkey Chicken Nuisance topic, by the time 97.5% Recall had been attained only 1.93% of the corpus, 17,414 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 98.07% or 885,020 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 2333 Article Spinner Spinning

Confusion Matrix- Topic 2333
Total Documents: 465,147
Total Relevant: 4,805
Total Prevalence: 1.03%

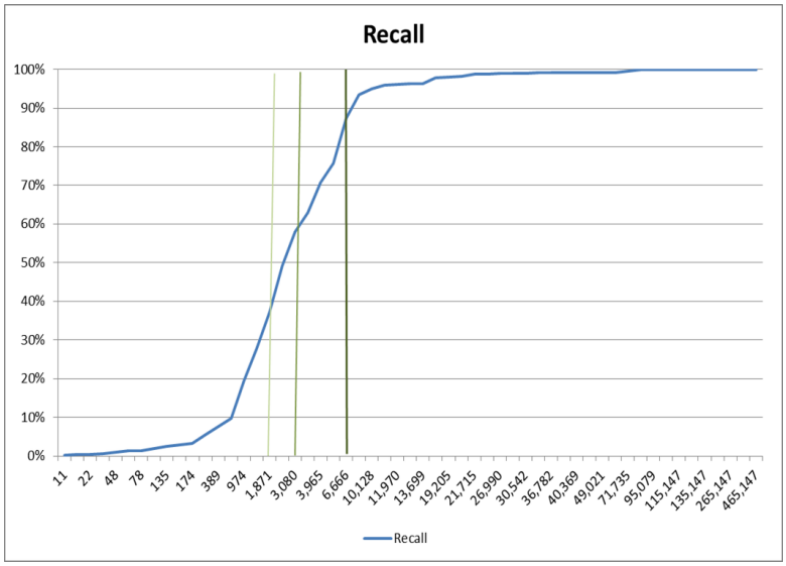
	@Reas. Call	@97.5% Recall
True Positives	**4,201**	**4,685**
True Negatives	**457,877**	**450,329**
False Positives	**2,465**	**10,013**
False Negatives	**604**	**120**
Recall	**87.43%**	**97.50%**
Precision	**63.02%**	**31.88%**
F1 Measure	**73.24%**	**48.04%**
Accuracy	**99.34%**	**97.82%**
Error	**0.66%**	**2.18%**
Elusion	**0.13%**	**0.03%**
Fallout	**0.54%**	**2.18%**

Topic 2333 was run by Losey who also started on August 19, 2015. He finished his review of 465,149 forum posts in *BlackHat World* on August 23, 2015. The project commenced as usual with Losey beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal search began, including predictive coding features, with iterated training.

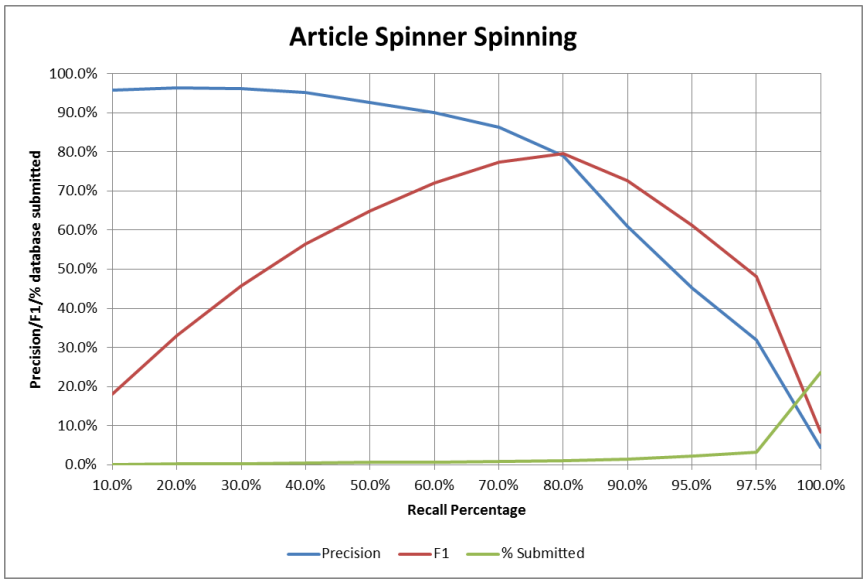
On August 21, 2015, after making 23 submissions to TREC, and training after almost every submission, Losey had provided a total of 6,666 documents to TREC and confirmed a total of 4201 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 228 documents. After the 23rd TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 87.43%** was attained by submission of only 6,666 documents, which is .043% of the total 465,147 documents. This was accomplished by personal review of only 228 documents, 0.05% of the total collection.

There were 32 additional submissions to TREC after the *Reasonable* call point. Recall of 90% was attained after submitting after submitting 7,091 documents, and 95% Recall after 10,931. Recall of 98% Recall was reached after submitting 14,698 documents, which was only 3.22% of total of 456,147 collection of *BlackHat World Forum* posts. Again, this was accomplished by personal review of only 228 documents, 0.05% of the total collection. In all topics we always stopped individual document review after the Reasonable call and relied on Mr. Robots automatic processes wherein the documents were submitted in order of highest ranking.

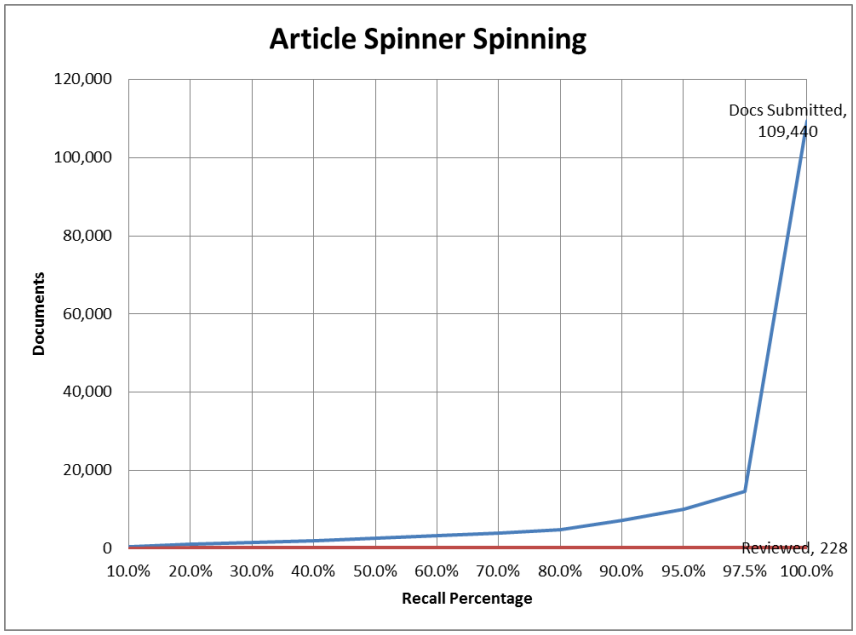
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Article Spinner Spinning topic, by the time 97.5% Recall had been attained only 3.16% of the corpus, 14,698 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 96.84% or 450,449 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 2129 Facebook Accounts

Confusion Matrix- Topic 2129
Total Documents: 465,147
Total Relevant: 589
Total Prevalence: 0.13%

	@Reas. Call	@97.5% Recall
True Positives	580	575
True Negatives	461,284	462,644
False Positives	3,274	1,914
False Negatives	9	14
Recall	98.47%	97.62%
Precision	15.05%	23.10%
F1 Measure	26.11%	37.36%
Accuracy	99.29%	99.59%
Error	0.71%	0.41%
Elusion	0.00%	0.00%
Fallout	0.70%	0.41%

Topic 2129 was run by Sullivan who started on August 21, 2015. He finished his review of 465,149 forum posts in *BlackHat World* on August 22, 2015.

While he counts himself among Facebook's 1.5 billion active users, Sullivan does not consider himself more knowledgeable on this topic than the average person.

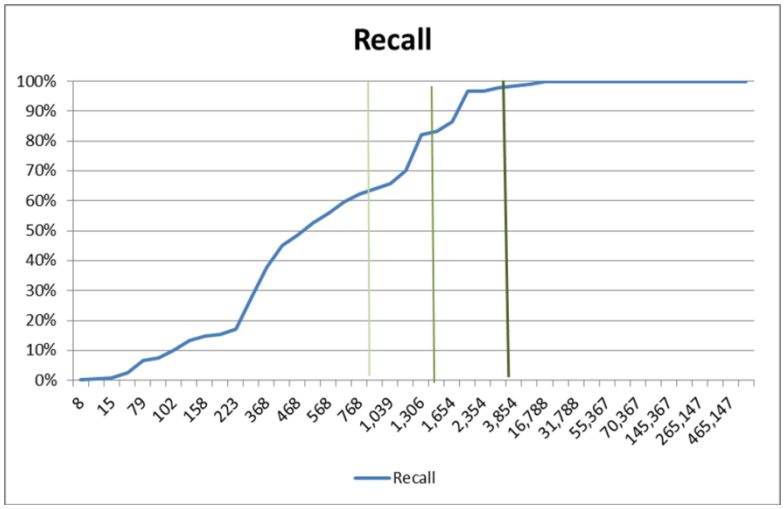
Day 1 on this topic started like all Sullivan topics with concept searching to find keywords relating to Facebook accounts for searching and highlighting. Specifically, variations of Facebook spelling and slang were investigated to ensure all common variants are identified. Many previously unexpected variations of facebook were identified, such as fbook. All variations were added to the highlighting list and documented for future searches.

Sullivan spent 2.5 hours on Day 1 trying to define relevance according to the TREC standard. He started with 8 documents that contained clear references to facebook accounts, and only 1 of the documents was returned as relevant according to the TREC standard. He continued by isolating documents that contained "Facebook account*" in the title as well as a number of common variants. At the end of the day, Sullivan was no closer to cracking the Facebook puzzle and was barely able to exceed 50% precision even though he was only submitting documents that were certain to be relevant by any objective standard.

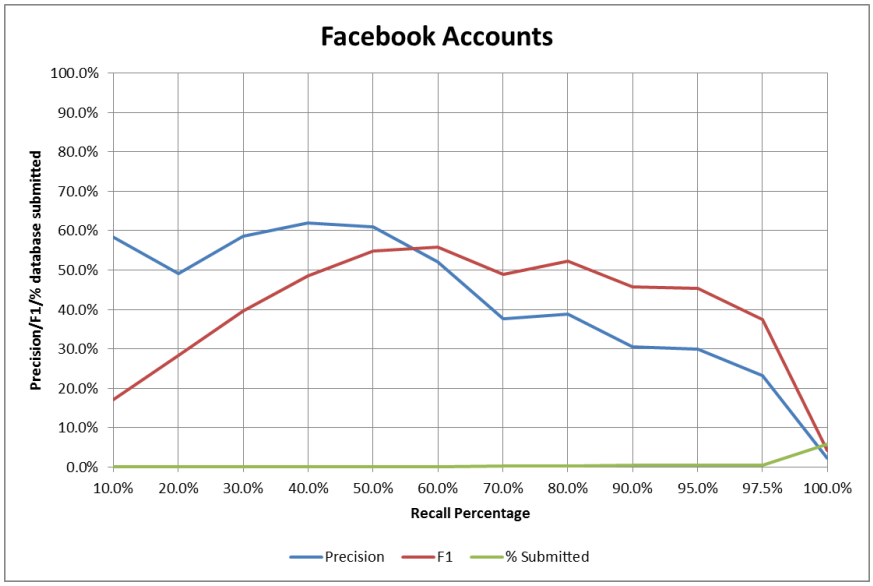
Facing what appeared to be a dead-end, Sullivan started Day 2 by relying on the priority scores generated by Mr. EDR, and started to see much better results. While Sullivan was unable to identify which documents would be returned as responsive by TREC, Mr. EDR seemed to be able to find the pattern. As such, he stopped looking at the documents, and just started submitting all documents that had a high priority score that contained the term Facebook or any known

variation, with learning sessions being run periodically to update the scores based on new learning. Once those documents were exhausted, all remaining documents were submitted in descending priority score order. He spent 2.75 hours submitting and evaluating the results, for a total of 5.25 hours spent on this topic.

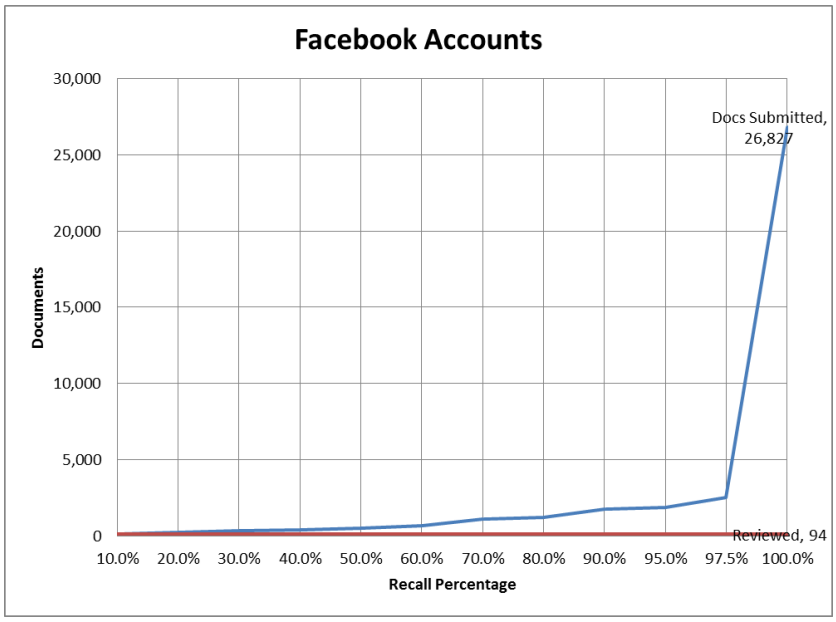
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Facebook Accounts topic, by the time 97.5% Recall had been attained only 0.54% of the corpus, 2,489 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.46% or 462,658 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multi-modal hybrid model of training EDR.



Topic 3378 Rob McKenna Gubernatorial Candidate

Confusion Matrix- Topic 3378
Total Documents: 902,434
Total Relevant: 66
Total Prevalence: 0.01%

	@Reas. Call	**@97.5% Recall**
True Positives	59	65
True Negatives	902,321	902,264
False Positives	47	104
False Negatives	7	1
Recall	89.39%	98.48%
Precision	55.66%	38.46%
F1 Measure	68.60%	55.32%
Accuracy	99.99%	99.99%
Error	0.01%	0.01%
Elusion	0.00%	0.00%
Fallout	0.01%	0.01%

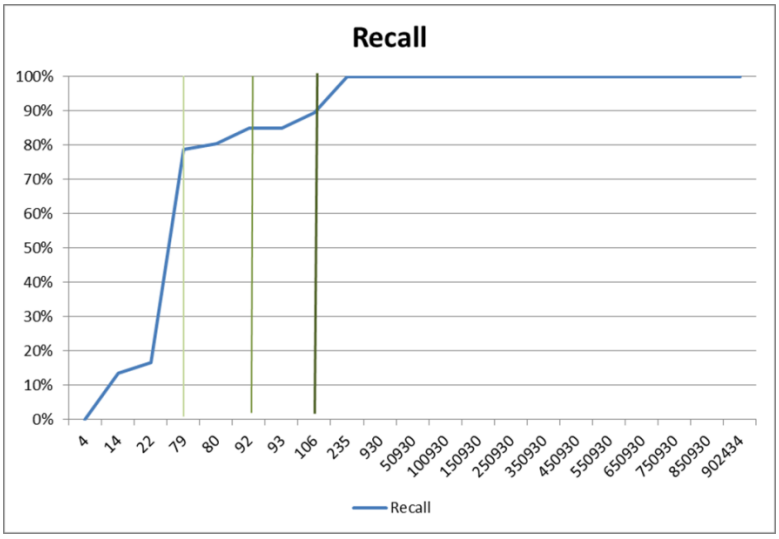
Topic 3357 was run by Reichenberger. The work to search the 902,434 *News Articles database* started on August 22, 2015, and was completed on August 23, 2015.

The initial submissions on the first day were to test the outlines of the relevance scope. It was ascertained in the first two submissions that documents relating to McKenna as a candidate were relevant, and those related to his job as Attorney General were irrelevant. Borderline documents were those associated with his Attorney General job that could be pretext to a political campaign (e.g. filing a suit related to Obamacare implementation). The third submission was made with the next 65 documents based on prioritization without looking at the content; the results largely confirmed the anticipated parameters (43 relevant, 22 irrelevant, with the borderline documents skewing to the irrelevant) The 70% call was made following the return of results.

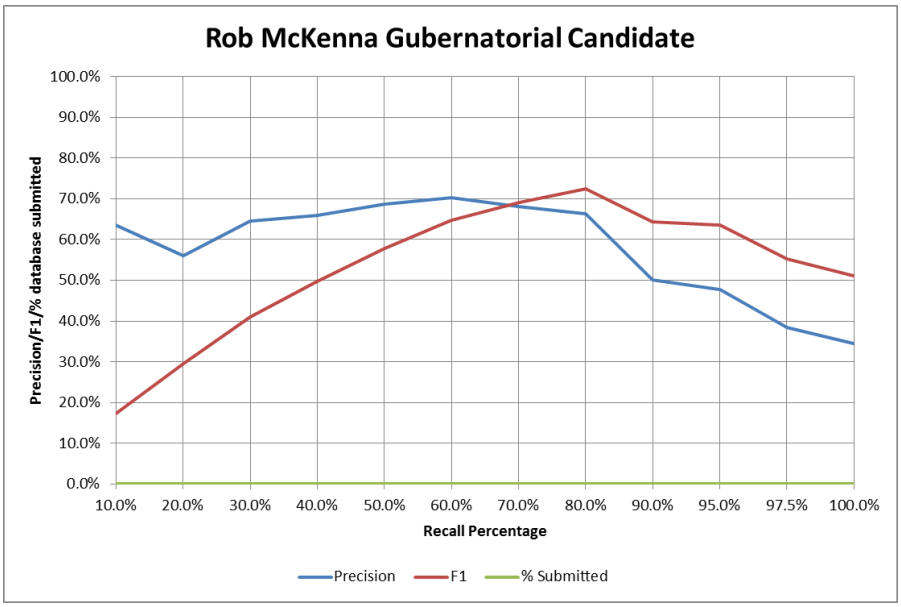
After looking at what was being promoted by prioritization and containing "McKenna," the next 13 documents were submitted. Most of these appeared to be borderline, only 4 were adjudicated relevant by TREC. The 80% recall call was made at that point. One more set of 14 documents was submitted and only 3 came back responsive. The decision was then made to call Reasonable, and thereafter the final submissions were made.

The post call submissions were made by the following groups in descending priority score order: 1) all documents reviewed that were currently anticipated to be irrelevant, but had now been submitted (129 documents, of which 7 were relevant); 2) anything remaining with "McKenna" (695 documents, all irrelevant; and then 3) all else (all irrelevant).

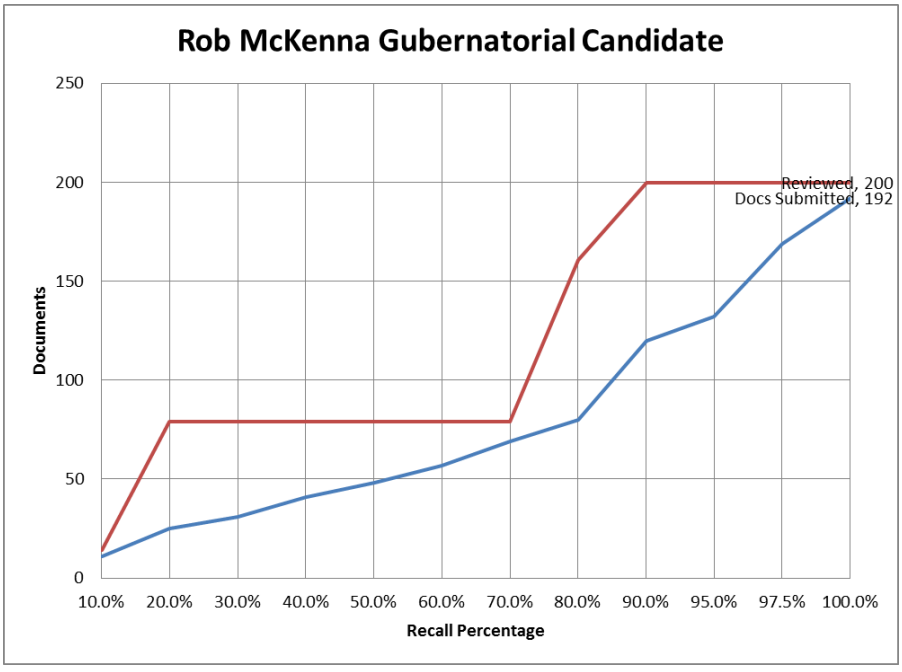
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying Recall thresholds. On the Rob McKenna Gubernatorial Candidate topic, by the time 97.5% Recall had been attained only 0.02% of the corpus, 169 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.98% or 902,265 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the Multimodal Hybrid model of training Mr. EDR.



_____

Topic 2322 Web Scraping

Confusion Matrix- Topic 2322 Web Scraping
Total Documents: 456,147
Total Relevant: 10,145
Total Prevalence: 2.22%

	@Reas. Call	**@97.5% Recall**
True Positives	8,060	9,892
True Negatives	441,263	436,073
False Positives	4,739	9,929
False Negatives	2,085	253
Recall	79.45%	97.51%
Precision	62.97%	49.91%
F1 Measure	70.26%	66.02%
Accuracy	98.50%	97.77%
Error	1.50%	2.23%
Elusion	0.47%	0.06%
Fallout	1.06%	2.23%

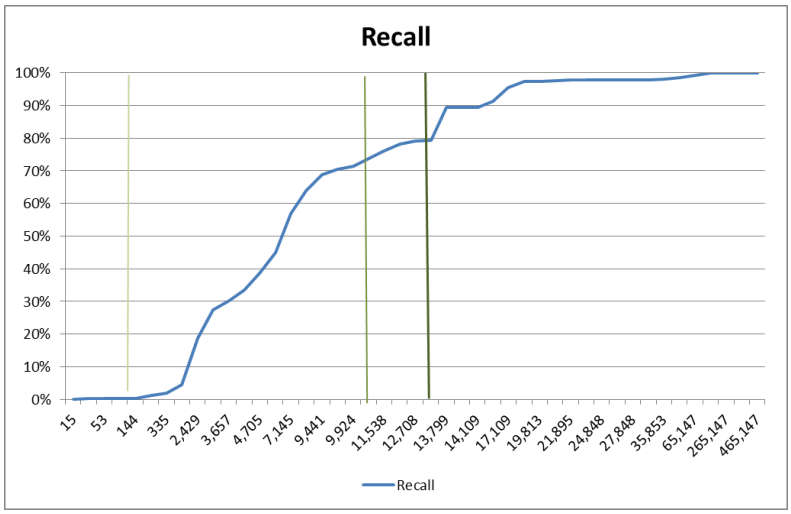
Topic 2322 was run by Losey who also started on August 22, 2015. He finished his review of 465,149 forum posts in *BlackHat World* on August 25, 2015. The project commenced as usual with Losey beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal search began, including predictive coding features, with iterated training.

On August 25, 2015, after making 24 submissions to TREC, and training after almost every submission, Losey had provided a total of 12,799 documents to TREC and confirmed a total of 8,060 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 195 documents. After the 24th TREC submission, Losey decided to call *Reasonable*. It was later determined that a **Recall of 79.45%** was attained by submission of only 12,799 documents, which is 2.8 % of the total documents. This was accomplished by review of only 0.04% of the total collection.

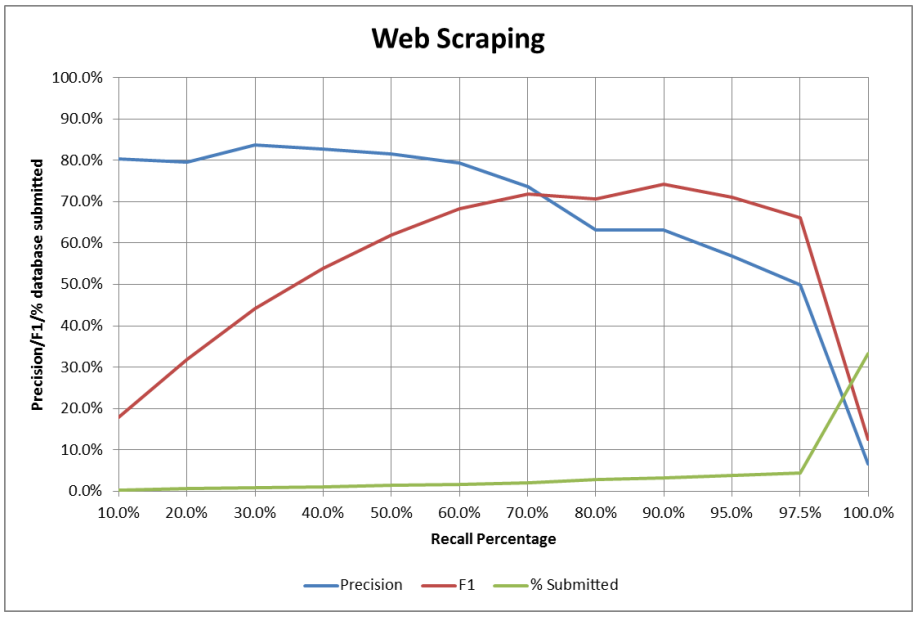
There were 21 additional submissions to TREC after the *Reasonable* call point. In the next submission after Reasonable call, the 25th, 1,000 documents were submitted and they all came back relevant. Obviously an error in gamesmanship had been made and the call was made a little too early. After that 25th submission, the Recall level rose to **89.31% and the Precision increased to 65.66%.**

A 90% Recall was attained after submitting 14,477 documents. A 95% Recall was attained after submitting 16,983 documents, and 97.5% Recall attained after 19,821 documents were submitted, which was only 4.35% of total of 456,147 collection of *BlackHat World Forum* posts.

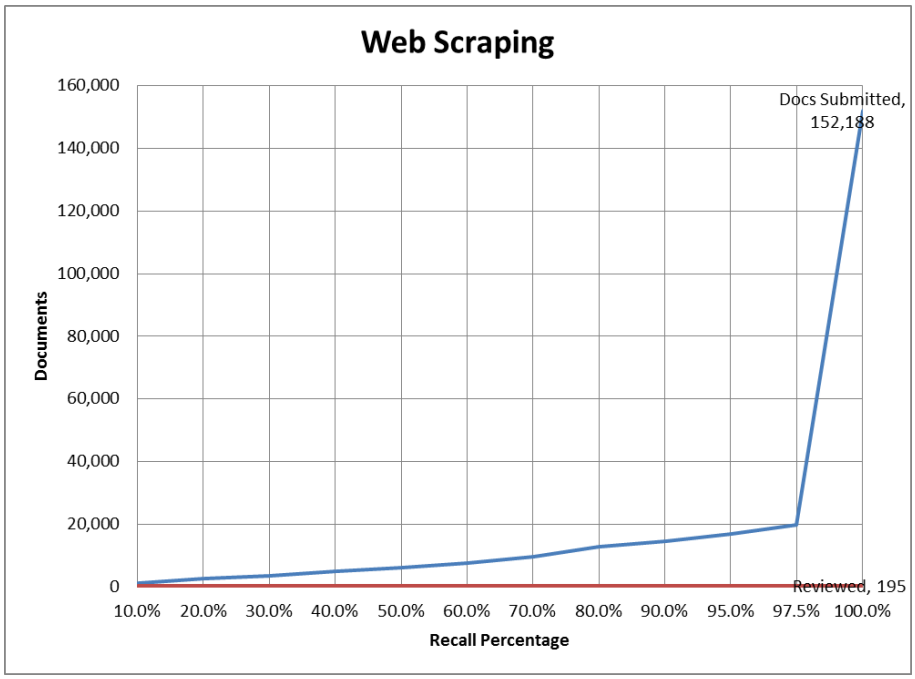
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Web Scraping topic, by the time 97.5% Recall had been attained only 4.35% of the corpus, 19,821 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 95.65% or 436,326 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 3484 Paul and Cathy Lee Martin

Confusion Matrix- Topic 3484

Total Documents: 902,434
Total Relevant: 23
Total Prevalence: 0.00%

	@Reas. Call	@97.5% Recall
True Positives	23	23
True Negatives	902,411	902,411
False Positives	0	0
False Negatives	0	0
Recall	100.00%	100.00%
Precision	100.00%	100.00%
F1 Measure	100.00%	100.00%
Accuracy	100.00%	100.00%
Error	0.00%	0.00%
Elusion	0.00%	0.00%
Fallout	0.00%	0.00%

This Topic was run by Sullivan who started on August 24, 2015. He completed his review of 902,434 documents on August 25, 2015. The entire Team observed his final submissions and cheered on his perfect handling of this search project.

This topic was completely unknown to Sullivan prior to this exercise. His only knowledge came from a quick Google search on the topic.

Sullivan started late on Day 1 and began with a simple search using the following keywords: ((martin w/3 paul) AND cathy) OR ((martin w/3 cathy) AND paul). This search returned 26 documents. A quick review of the documents yielded 22 clearly relevant documents and 1 marginally relevant. Sullivan submitted the 22 relevant documents, which were all returned as relevant by TREC and quit for the night after 15 minutes of effort.

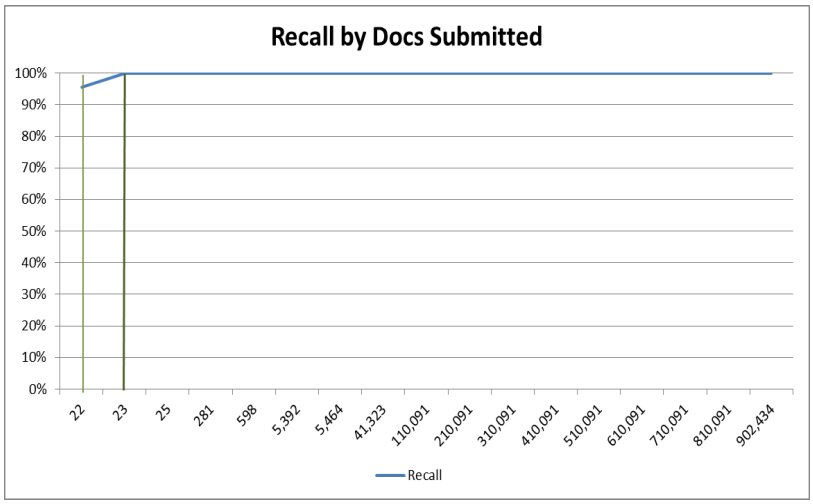
On Day 2, Sullivan went back to his standard process of using concept searching to find relevant keywords for highlighting and searches. As with all topics in dataset 3, spelling errors were non-existent, which removed the requirement of broad searching to account for slang or spelling issues.

Broad searches were run using all relevant keywords and the results were sampled. Next predictive coding scores were used to identify additional potentially relevant documents. A large number of false positives were encountered when it was discovered a popular hockey player and Prime Minister shared the same names as the parties. These were quickly identified and excluded from the potentially relevant set. After 90 minutes of work, Sullivan conceded that he was unable to find any additional relevant documents.

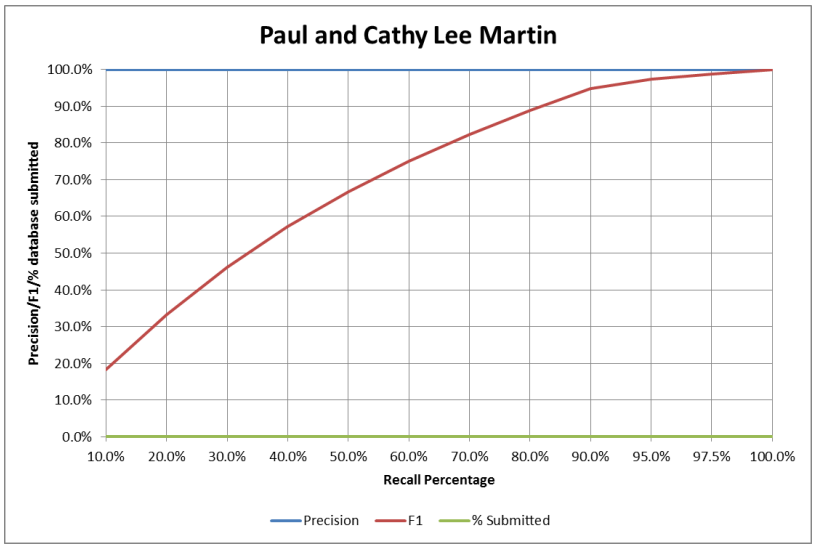
In reviewing the single marginally relevant document found on Day 1, it was determined this document was very likely to be relevant, so it was submitted to TREC and was in fact returned relevant. At this point, Sullivan called reasonable recall and submitted all remaining documents in descending order of priority score.

After all documents were submitted, it was discovered that Sullivan in fact had attained 100% recall and 100% precision at the point the reasonable call was made. Additionally, 95.7% recall was attained, with 100% precision, after only 15 minutes. In all, he was able to achieve a perfect game with only 1.75 hours committed to this topic!

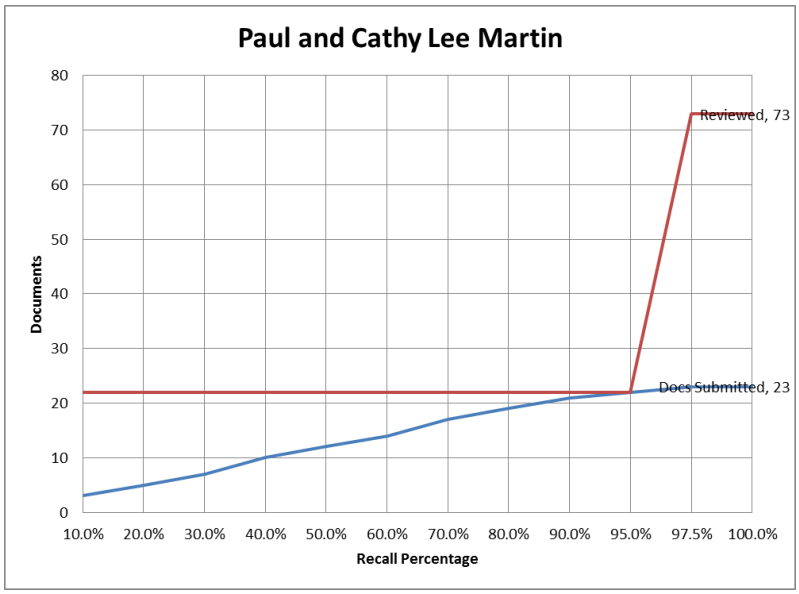
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Paul and Cathy Lee Martin topic, by the time 97.5% Recall had been attained only 0.00% of the corpus, 23 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 100.00% or 902,411 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multi-modal hybrid model of training EDR.



Topic 2134 Paypal Accounts

Confusion Matrix- Topic 2134
Total Documents: 465,147
Total Relevant: 252
Total Prevalence: 0.05%

	@Reas. Call	**@97.5% Recall**
True Positives	241	246
True Negatives	461,447	443,136
False Positives	3,448	21,759
False Negatives	11	6
Recall	95.63%	97.62%
Precision	6.53%	1.12%
F1 Measure	12.23%	2.21%
Accuracy	99.26%	95.32%
Error	0.74%	4.68%
Elusion	0.00%	0.00%
Fallout	0.74%	4.68%

Topic 2134 was run by Sullivan who started on August 26, 2015. He finished his review of 465,149 forum posts in *BlackHat World* on August 26, 2015.

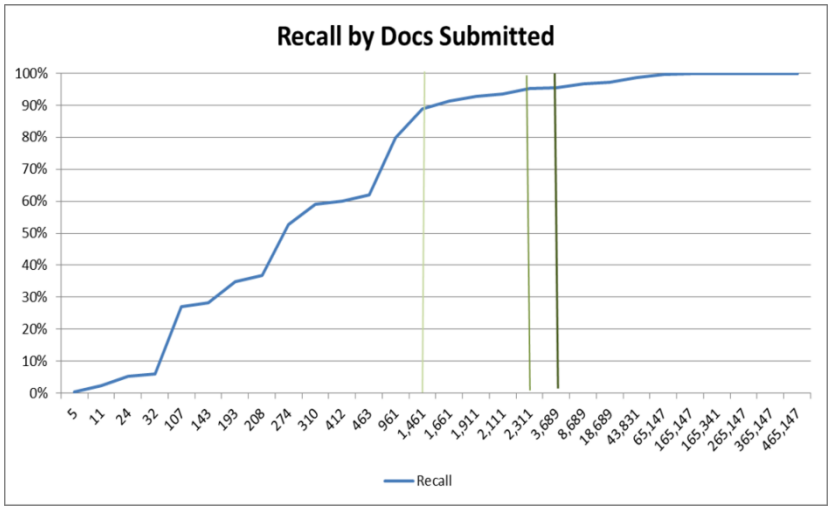
As a regular PayPal user for about 10 years, Sullivan has a high level of knowledge regarding this topic. This advanced knowledge proved to be a burden on this topic because his understanding of what should be relevant did not match with the TREC gold standard. He was able to overcome this burden by relying on a variety of advanced methods rather than using his own judgment in review of the documents.

Sullivan started this topic with his usual process of running concept searches to find similar and related keyword terms for highlighting and future searching. As with all forum topics, he spend some time identifying common variants based on misspelling or slang. All variations were added to the database for highlighting.

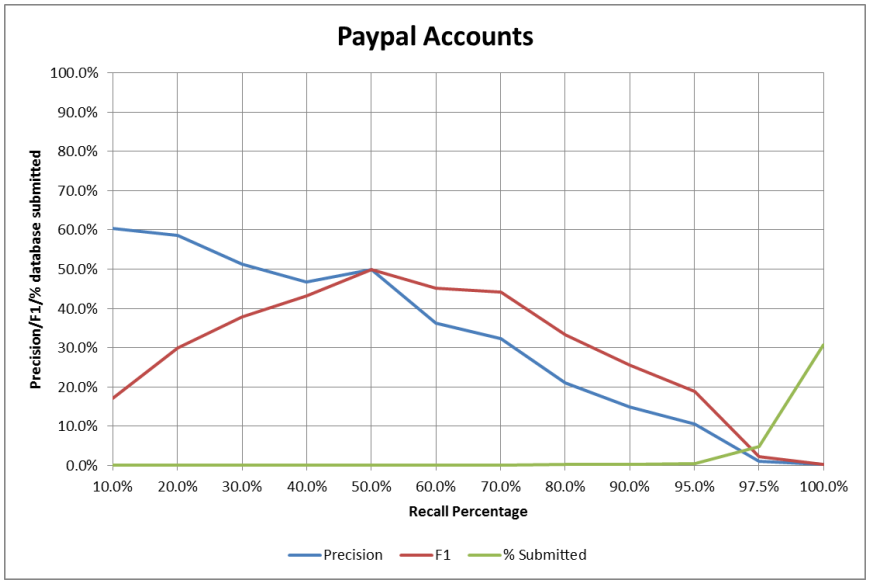
While using a number of methods to identify documents he felt were clearly relevant, Sullivan quickly realized he was unable to make any logic of the TREC relevance standard. Documents with similar or identical content were seemingly arbitrarily designated as relevant or not relevant. Rather than spend a considerable time evaluating the documents himself, as was done in Topic 2129 Facebook Accounts, he went straight to Mr. EDR for help.

Similar to the method developed in Topic 2129, Sullivan relied heavily on the predictive coding and did very little review on any documents. He would iteratively submit the highest scoring documents to TREC for analysis, and train the documents with the relevancy determination returned. In addition to using a continuous active learning approach, he started using the "Find Similar" feature much more to find documents that contained similar characteristics to documents already determined to be relevant. He started with documents that contained a variation of PayPal in the subject line, then moving to documents that contained the term anywhere in the text. Using this multimodal method he was able to work his way through the entire dataset with almost no actual review of the documents. In all, Sullivan was able to complete the review for this topic in less than 4 hours.

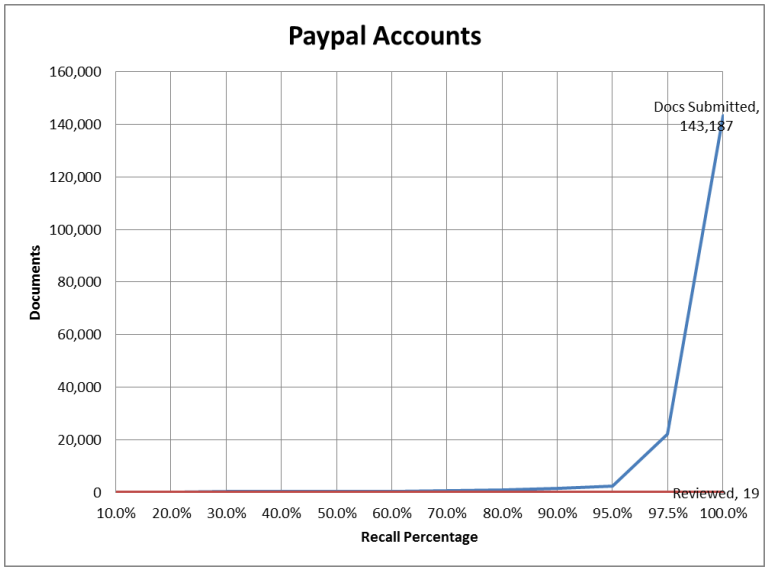
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Paypal Accounts topic, by the time 97.5% Recall had been attained only 4.73% of the corpus, 22,005 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 95.27% or 443,142 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multi-modal hybrid model of training EDR.



Topic 3423 Rob Ford Cut the Waist

Confusion Matrix- Topic 3423
Total Documents: 902,434
Total Relevant: 76
Total Prevalence: 0.01%

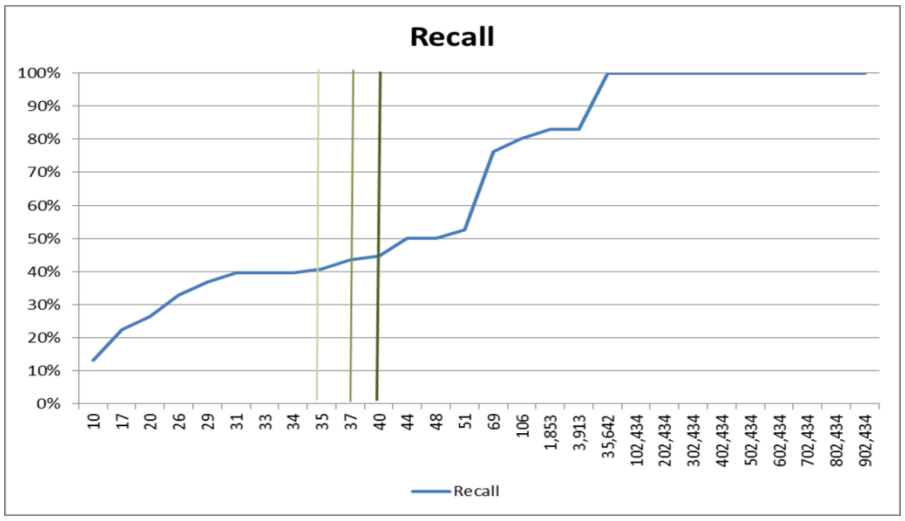
	@Reas. Call	**@97.5% Recall**
True Positives	34	75
True Negatives	902,352	867,337
False Positives	6	35,021
False Negatives	42	1
Recall	44.74%	98.68%
Precision	85.00%	0.21%
F1 Measure	58.62%	0.43%
Accuracy	99.99%	96.12%
Error	0.01%	3.88%
Elusion	0.00%	0.00%
Fallout	0.00%	3.88%

Topic 3423 was run by Losey who also started on August 26, 2015. He finished his review of 902,434 News Articles on August 27, 2015. The project commenced as usual with Losey beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal search began, including predictive coding features, with iterated training.

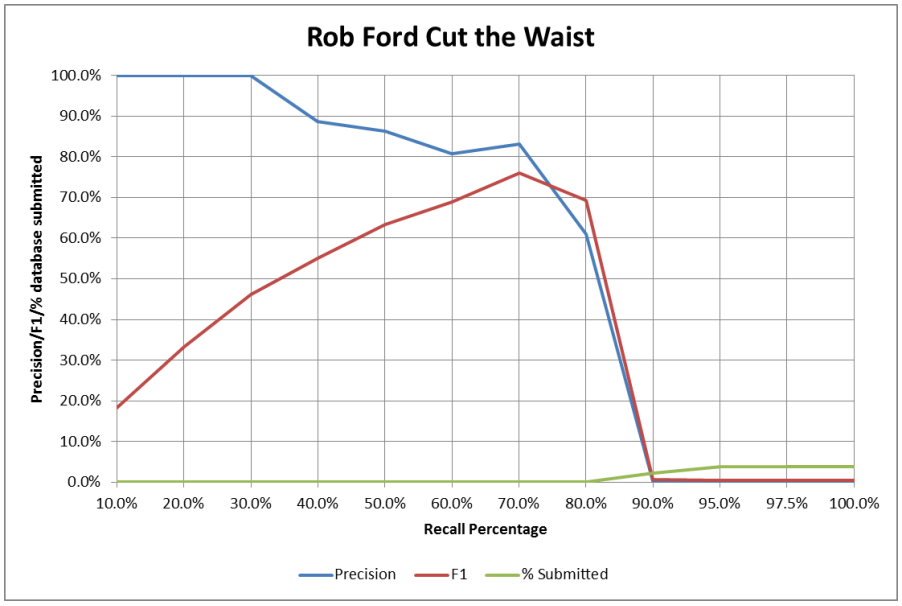
On August 26, 2015, after making 11 submissions to TREC, and training after almost every submission, Losey had provided a total of 40 documents to TREC and confirmed a total of 34 relevant documents. The effort, or number of documents reviewed and coded by Losey to attain this result, was 92 documents. After the 11th TREC submission, Losey decided to call *Reasonable*. This proved to be a premature call. It was later determined that a **Recall of 44.74%** was attained. In the 17 automatic submissions that followed, **Recall of 76.32% was attained with 84.06% Precision.** The 76.32% Recall was attained after submitting only 106 documents, which is 0.01% of the total of 902,434.

There were 17 submissions to TREC after the *Reasonable* call point. Total 100% Recall was attained after submitting only 35,193 documents, which is 3.9% of the total.

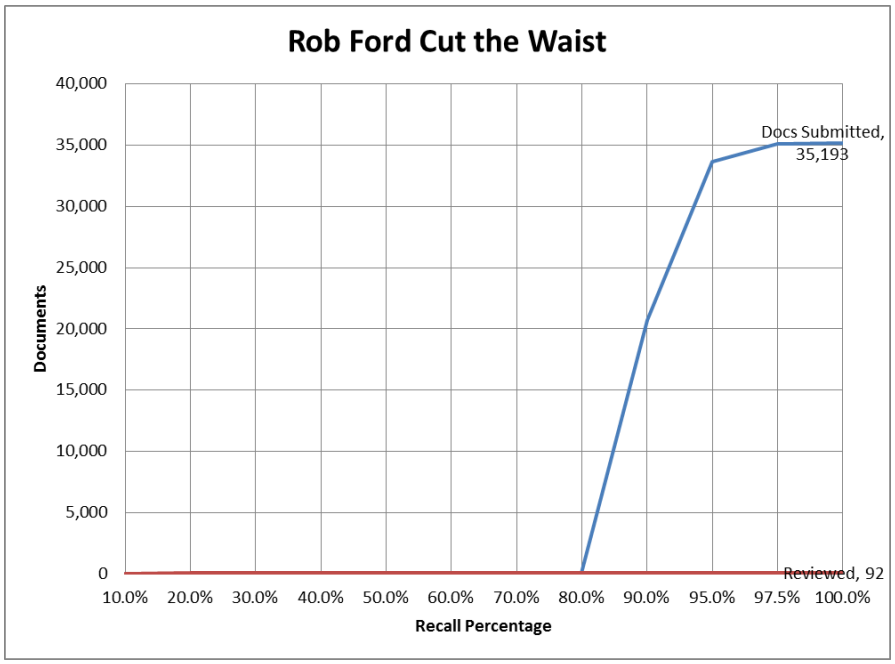
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% Recall call, and the dark green line the Reasonable Recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Rob Ford Cut the Waist topic, by the time 97.5% Recall had been attained only 3.89% of the corpus, 35,096 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 96.11% or 867,338 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% Recall using the multimodal hybrid model of search and training of Mr. EDR.



Topic 3133 Pacific Gateway

Confusion Matrix- Topic 3133
Total Documents: 902,434
Total Relevant: 113
Total Prevalence: 0.01%

	@Reas. Call	@97.5% Recall
True Positives	87	111
True Negatives	902,311	799,986
False Positives	10	102,335
False Negatives	26	2
Recall	76.99%	98.23%
Precision	89.69%	0.11%
F1 Measure	82.86%	0.22%
Accuracy	100.00%	88.66%
Error	0.00%	11.34%
Elusion	0.00%	0.00%
Fallout	0.00%	11.34%

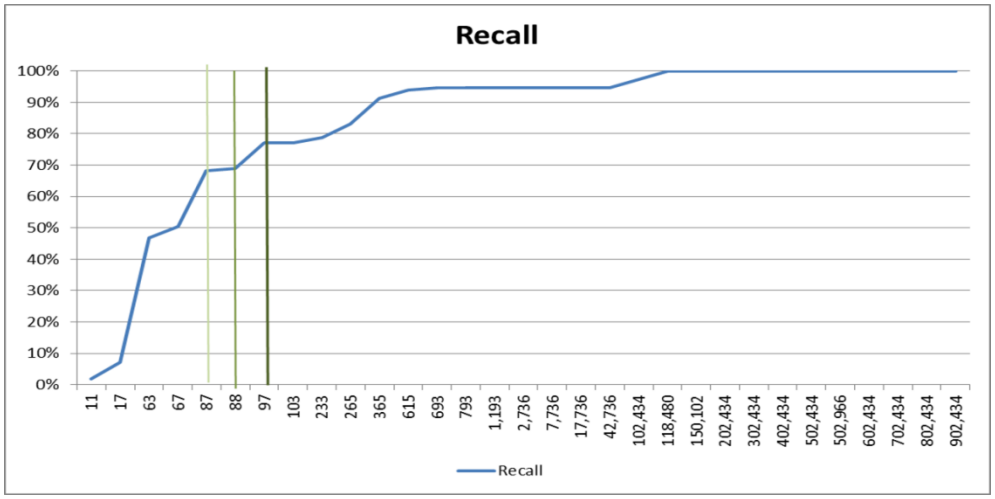
Topic 3133 was run by Losey who also started on August 27, 2015. He finished his review of 902,434 News Articles on August 28, 2015. The project commenced as usual with Losey

beginning Step Two, *Multimodal Search Reviews*. Step Three, Random Baseline, was omitted. After submissions began, the echo Step Five, multimodal search began, including predictive coding features, with iterated training.

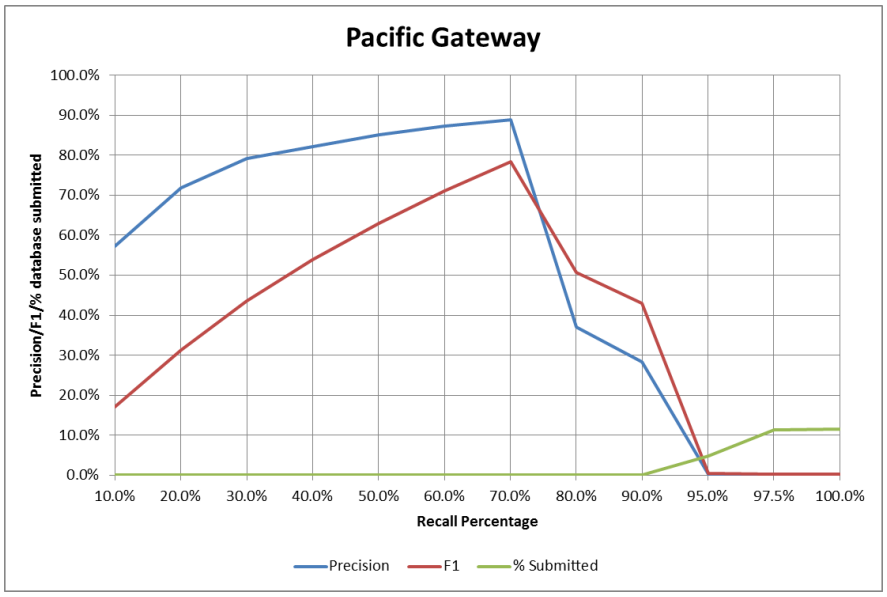
On August 28, 2015, after making 7 submissions to TREC, and training after almost every submission, Losey had provided a total of 97 documents to TREC and confirmed a total of 87 relevant documents. The effort, or number of documents individually reviewed and coded by Losey to attain this result, was 49 documents. After the 7th TREC submission, Losey decided to call *Reasonable*. That call proved to be a little premature. It was later determined that a **Recall of 76.99%** was attained with **Precision of 89.69%**. In the 6th automatic submission after the call, a **Recall of 94.69% was attained** after submitting only 693 documents total, which is 0.07% of the total of 902,434.

There were 24 submissions to TREC after the *Reasonable* call point. Total **100% Recall** was attained after submitting 103,189 documents, which is 11.43% of the total.

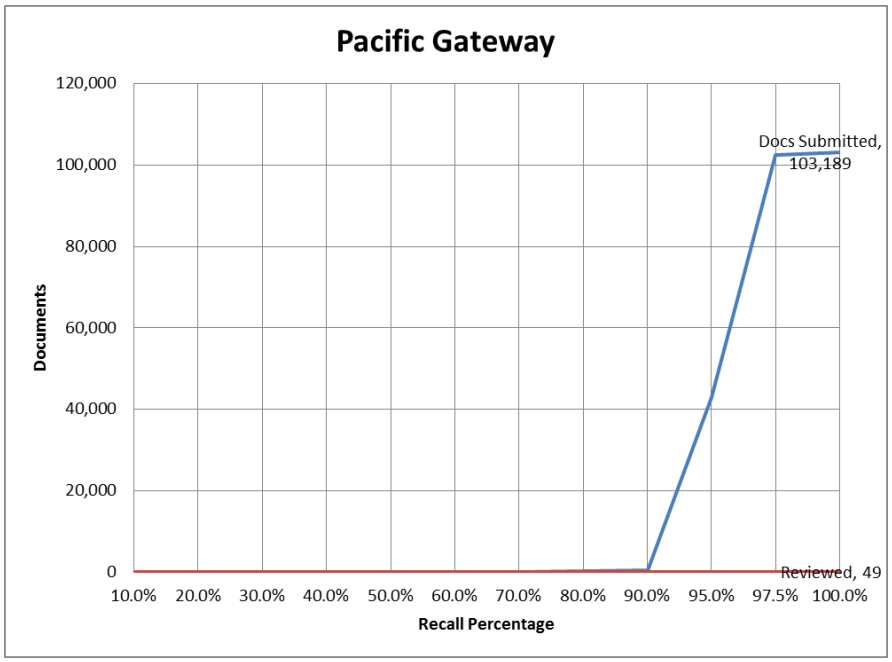
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Pacific Gateway topic, by the time 97.5% Recall had been attained only 11.35% of the corpus, 102,446 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 88.65% or 799,988 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multi-modal hybrid model of training EDR.



Topic 3226 Traffic Enforcement Cameras

Confusion Matrix- Topic 3226
Total Documents: 902,434
Total Relevant: 2,094
Total Prevalence: 0.23%

	@Reas. Call	@97.5% Recall
True Positives	2,061	2,042
True Negatives	897,054	899,807
False Positives	3,286	533
False Negatives	33	52
Recall	98.42%	97.52%
Precision	38.54%	79.30%
F1 Measure	55.39%	87.47%
Accuracy	99.63%	99.94%
Error	0.37%	0.06%
Elusion	0.00%	0.01%
Fallout	0.36%	0.06%

Topic 3226 was run by Sullivan who also started on August 27, 2015. He finished his review of 902,434 News Articles on August 28, 2015.

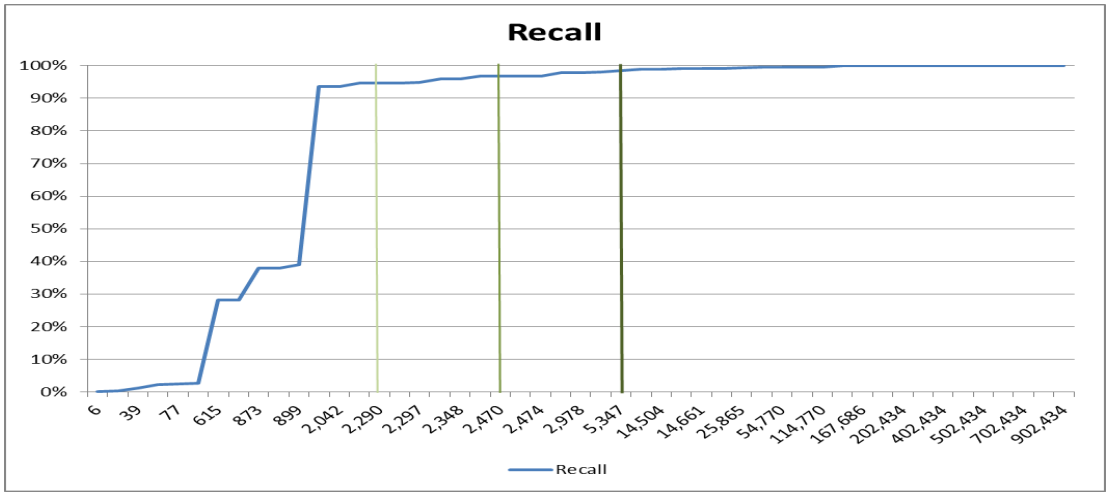
Sullivan has some prior experience as a criminal defense attorney, with experience with traffic laws, but he has no prior experience with traffic enforcement cameras, which were not in use at the time he was practicing.

As usual, Sullivan started his investigation with his standard process of using keyword and concept searches to formulate a list of related keywords for highlighting and future searching. For this exercise, nothing extraordinary was discovered, but he was able to generate a good list of terms relating to traffic cameras, red light cameras, and traffic tickets.

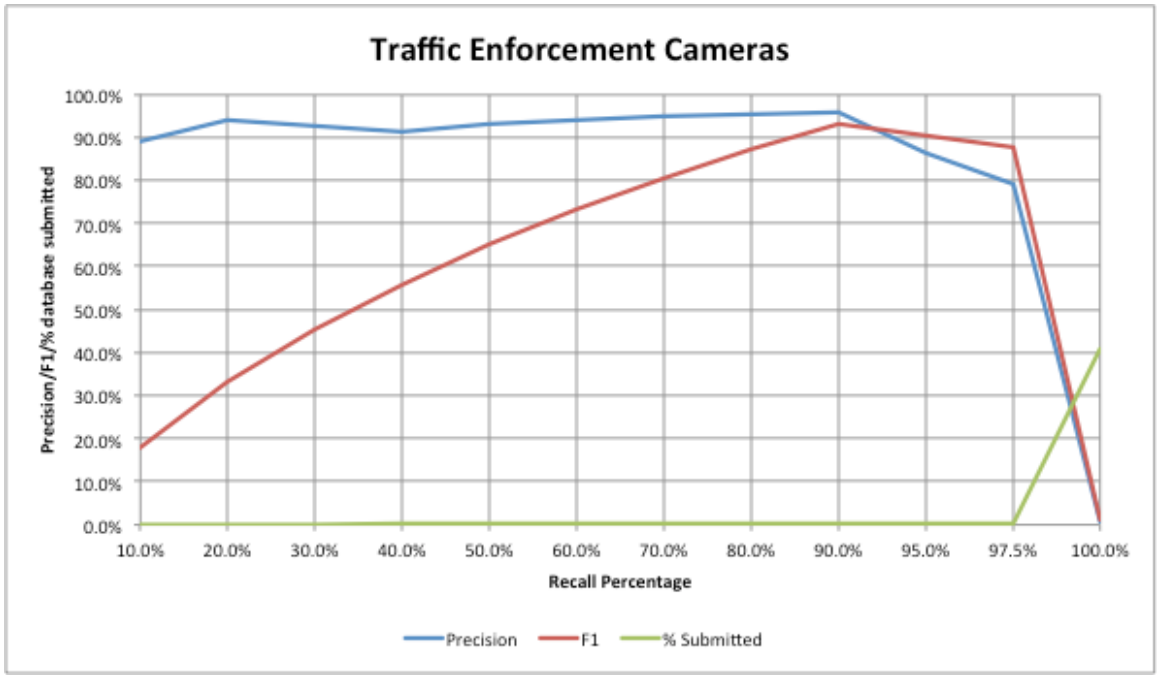
Day 1 was a short day and started with submitting the results of the most popular keyword searches with minimal review. After 30 minutes of work, 76 documents were submitted with 50 being returned as relevant.

Using the documents identified on Day 1, Sullivan was able to start utilizing the predictive coding to supplement his searches on Day 2. He was able to progressively make his way through the review set using a combination of predictive coding scores and keyword hits. He used this multimodal approach to submit large sets of documents with minimal, if any, manual review. He believed he had found all relevant documents after submitting only 5,347 total documents with 2,061 relevant. After submitting all of the remaining documents in descending order by predictive coding priority score, it was discovered he only missed 33 of the relevant documents in the dataset after submitting 0.6% of the documents! Because he minimized the amount of manual review on this topic, he was able to complete this topic after 3.0 hours on Day 2, for a total of 3.5 hours on this topic.

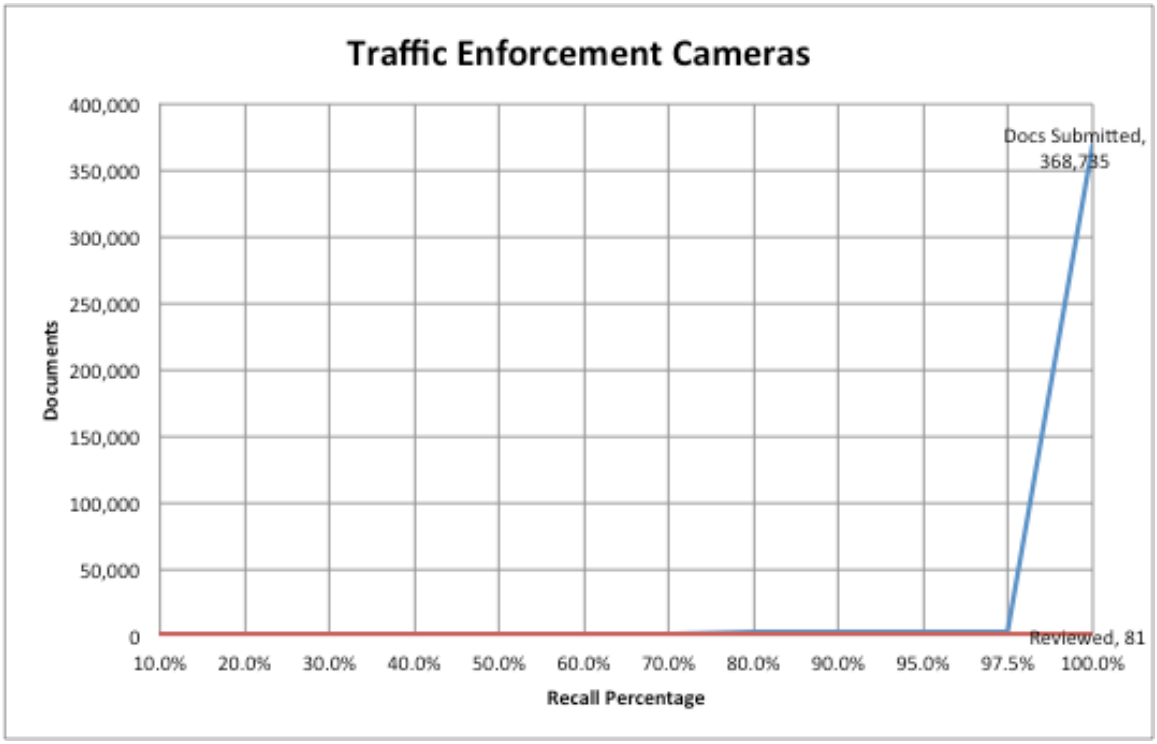
A graph mapping how the review was conducted appears below, with the light green line signifying the anticipated 70% recall call, and the dark green line the reasonable recall call.



The following chart shows Precision (left and blue line), F1 (red) and percent of documents submitted (green) as tracked across varying recall thresholds. On the Traffic Enforcement Cameras topic, by the time 97.5% Recall had been attained only 0.29% of the corpus, 2,575 documents, had been submitted for adjudication. The last portion of the graph thus represents the submission of the remaining 99.71% or 899,859 documents.



The last chart below represents the amount of effort in terms of documents reviewed to attain 100% recall using the multi-modal hybrid model of training EDR.


Traffic Enforcement Cameras
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Documents
Docs Submitted, 368,735
Reviewed, 81
10.0%
20.0%
30.0%
40.0%
50.0%
60.0%
70.0%
80.0%
90.0%
95.0%
97.5%
100.0%
Recall Percentage