Cooperation, Transparency, and the Rise of Support Vector Machines in E-Discovery: Issues Raised by the Need to Classify Documents as Either Responsive or Nonresponsive

Jason R. Baron, Esq.
University of Maryland, College Park, Maryland, USA
jason.baron@nara.gov

Jesse B. Freeman
Williams College, Williamstown, Massachusetts, USA
jesse.b.freeman@williams.edu[1]

Abstract: Exponential increases in the volume of electronically stored information are necessitating new thinking on the part of the greater legal community, including a movement away from linear or manual review, as well as away from reliance on keyword searching as the sole automated means to handle e-discovery search and document review requirements. Increasingly, lawyers are becoming more familiar with certain advanced forms of search techniques, including those utilizing machine learning. The landmark US opinion in *da Silva Moore v. Publicus Groupe SA,* issued in February 2012, giving a judicial imprimatur to use of "predictive coding" and other sophisticated iterative sampling techniques in satisfaction of discovery obligations, should assist in paving the way toward greater acceptance of these new methods. Almost all of these machine learning processes are based on support vector machines or related algorithms, which at first glance seem unapproachably complex. The basic intuitions behind their functionality are not nearly as daunting. After providing relevant background on traditional notions of the discovery process and the emergence of a need for more sophisticated forms of artificial intelligence to solve e-discovery challenges, this paper will explain the mathematical intuition behind support vector machines, so that lawyers can more fully grasp the implications of this new technology. In particular, this paper suggests that support vector machine technology necessarily requires lawyers paying heightened attention to notions of cooperation and transparency, in light of the collaborative, iterative interaction with coding software, and the need for sharing sets of non-responsive documents in order that use of the technology is optimized.

Keywords: e-discovery, support vector machines, electronically stored information, cooperation, transparency, Moore v. Publicus, iterative, responsive, nonresponsive.

1. Introduction

Since enactment of the 2006 US Federal Rules of Civil Procedure, lawyers in the United States increasingly have confronted the need to learn about a brave new world of "electronically stored

[1] Jason R. Baron is an Adjunct Faculty member of the University of Maryland College of Information Studies, and also serves as Director of Litigation at the National Archives and Records Administration. B.A., Wesleyan University; J.D., Boston University; Member of the District of Columbia and Massachusetts Bars. Jesse B. Freeman is a candidate for B.A. degrees in mathematics and economics, Class of '15, at Williams College, and interned in the Office of General Counsel at the National Archives during the Summer of 2012. The authors wish to kindly thank William Webber, Doug Oard, and Ralph Losey for their comments on earlier drafts. The views expressed here are the authors alone and do not necessarily represent the views of any institution, public or private, with which they are affiliated.

information" (ESI), including the need to be aware of tools and techniques borrowed from the realm of artificial intelligence that previously were unheard of in civil discovery practice prior to trial. The 2006 Rules anticipated that the profession would undergo a sea-change in practice, by requiring increased attention to preservation of and access to electronic evidence at the outset of litigation, in the form of increased awareness of the necessity of legal preservation holds [1], and the desirability of performing more advanced and efficient searches for relevant documents – beyond anything necessitated in an era of paper documents [2, 3]. Given the need to pay attention at the beginning of litigation to such highly technical issues, lawyers are beginning to embrace the notion of being more cooperative and transparent in their legal practice to conform to e-discovery demands [4].

Nevertheless, the legal profession as a whole is by no means aware of the latest, profound changes in discovery practice brought on by the emerging use of machine learning technologies in the cause of making document review more efficient. In particular, support vector machines (SVMs) have the potential to dramatically increase both the quality and efficiency of the search and document review functions in e-discovery. Unfortunately, the mathematical formulas used to describe SVMs are both technical and intimidating. This paper has two modest aims: first, we will show that the intimidating formulas that keep many from fully understanding how SVMs work are based on the much simpler mathematical notions of distance and separation. Hopefully, readers of this paper will develop greater understanding of SVMs, in order that they consider incorporating such promising new technologies in their everyday e-discovery practice. While SVMs are not the only predictive coding technology available, this paper focuses on SVMs for two reasons. First, SVMs are a highly popular form of predictive coding. Second, all predictive coding software maps documents based on specified characteristics and looks for those characteristics in unread documents in order to make similar classifications without the need for hands-on review. We focus on SVMs because the theoretical background on predictive coding involved in the explanation de-mystify the process for all users and the specific mechanism of the SVM should be directly relevant information to many.

A second aim is to preliminarily explore how growing and eventually widespread use of SVMs holds the potential to upset traditional notions of what it means to practice civil discovery. The paper will argue that optimum use of these technologies necessitates practicing a heightened level of cooperation and transparency between or among adversaries, at least with respect to the sharing of "nonresponsive" documents during the discovery process. The authors are well aware of how provocative these issues are; however, as described in detail below, starting with the *da Silva Moore v. Publicus Groupe SA* litigation in a US federal court in Manhattan, and in a select number of other cases, the parties are already largely on record as having embraced just such a level of cooperation -- thus making the positions taken in this paper somewhat easier to maintain, as at least not entirely speculative [5].

2. Traditional Means of "Cooperation" in US Discovery and E-Discovery Practice

Since 1938, with the adoption of the US Federal Rules of Civil Procedure, civil discovery practice, as ideally realized, has been grounded on notions of cooperation, transparency and fairness [6, 7]. The rules traditionally have assumed that lawyers will carry out their obligations on behalf of clients without need of active court supervision; however, in the age of ESI, judicial norms with regard to how active a court should be on the front end of litigation are, in many places, rapidly changing. Regardless, lawyers' obligations have been bounded, however, by at least one limiting condition that represents a fundamental aspect of practice, universally followed to date, namely: that due diligence involves the search for and production of any and all *nonprivileged, relevant* evidence requested by an opposing party. Thus, as early as 1946, the US Supreme Court held in the case of *Hickman v. Taylor* [8], that "[m]utual knowledge of all the *relevant* facts gathered by both parties is essential to proper litigation" (emphasis added). To that end, Rule 26(b)(1) states that "Parties may obtain discovery regarding any nonprivileged matter that is *relevant* to any party's claim or defense," and that "[f]or good cause, the court may order discovery of any matter *relevant* to the subject matter involved in the action" (emphasis added). The Rule goes on to add that "*relevant* information" need not be admissible at trial if discovery appears reasonably calculated to lead to the discovery of admissible evidence.

In the decades prior to the 2006 rules changes, for the most part the legal community met its obligations under the federal rules by performing reasonable searches for relevant documents in traditional folders, filing cabinets, and warehouses filled with records. The task at hand was to straightforwardly have one or more lawyers – sometimes in teams – work through a review of boxes of documents to cull out potentially relevant pieces of evidence, for a further decision on both relevance and privilege. Irrelevant or nonresponsive documents were left behind, and only in rare cases were there quality checks to determine if documents had been missed in the review. To the extent controversy existed with respect to the basic discovery protocol, it involved occasional albeit sometimes notorious cases where counsel (and their client) failed to make reasonably diligent efforts to comply with a legally proper discovery request by opposing counsel, resulting in sanctions in the most egregious cases of suppressed (i.e., known but not disclosed) evidence [9].

The past decade has seen the growing volume and complexity of evidence in the form of ESI. This in turn has led to a spotlight placed on the efficacy of keyword searching in lieu of wholesale reliance on manual or linear review, i.e., "eyes-on" review of every document by a team of attorneys [2]. In the paradigmatic case, counsel's initiation of search protocols centered around coming up with a limited number of keywords, with or without employment of Boolean operators, has been for some time the *de facto* standard for meeting legal requirements to perform reasonable searches for relevant documents. The Sedona Search Commentary went on to point out at length the known limitations of keyword searching based on the inherent ambiguities in written texts, citing to the important early work of Blair & Maron [10], and challenged the legal profession to recognize that more advanced means to perform searches of ESI held out the potential to increase both "recall" (the ratio of relevant documents obtained in a given search to the overall number of relevant documents in the repository subject to search), and "precision" (the ratio of relevant to irrelevant documents obtained in a given search). Accordingly, as Practice Pointer 1, the Commentary emphasized that

> In many settings involving [ESI], reliance solely on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary.

The Commentary went on to discuss alternative search methods, including use of techniques grounded in fuzzy search, concept search, latent semantic indexing, Bayesian belief networks, clustering and categorization techniques, and machine learning methods of various types [2]. The Commentary concluded with a call for research, to better evaluate known search methods in a legal context, and explicitly referenced the TREC Legal Track, run out of the US National Institute for Standards and Technology, as one such research effort underway [11].

In the years since the 2006 rules amendments, an explosion of case law and commentaries ensued, with increasing attention being paid to the importance of quality control, project management, and iterative sampling, to optimize completeness and accuracy in finding "relevant" documents in particular productions. (For a summary of cases and commentaries, see [12].) As part of this collective movement toward more sophisticated ways to perform quality control (QC) checks of results obtained, notions of how transparent the process should be to the "requesting" as opposed to "responding" party have come to be highlighted. Given the inherent asymmetry present in responding parties having unequal rights of access to and knowledge of their own data universe, in the pre-ESI era responding parties were comfortable in the expectation that they could perform reasonable searches of their client's records, without any *a priori* requirement imposed that the interim results of a given document production would be shared with opposing parties. The 2006 rules amendments, with an emphasis on early meet and confer conferences amongst parties to work through issues of preservation and access, somewhat undermined settled expectations. Against the backdrop of near-universal acceptance of the principle that lawyers should be more cooperative in negotiations involving their scope of ESI obligations, it was natural for the judiciary's expectations to be heightened with respect to the sophistication of would-be search protocols, including taking into account whether sufficient sampling of the "non-hit" population of documents had occurred to confidently say that all relevant documents had been found [3], [13].

3. A New Era: "Predictive Coding" Approved By Courts

Notwithstanding the growing sophistication in the legal space in the use of advanced search methods, not until the year 2012 had any reported judicial decision affirmatively ruled on whether the use of "predictive coding," as one form of software-assisted advanced search method, was justified. Everything has changed, however, with reported decisions out of New York [5], Virginia [13], and Louisiana [28], respectively, a further high-profile evidentiary proceeding pending in Illinois [14] -- all of which have involved various federal and state courts opining on the use of "predictive coding" in litigation to find relevant documents.

The term "predictive coding," as one of many labels describing partially automated software assisted review processes using support vector machines or related algorithms, involves (i) a set of preserved data, representing the entirety of what has been captured during a legal hold or culled down using filters for date ranges, custodians, or general subject areas; (ii) use of a random sample of seed documents, and/or a judgmental sample of documents obtained through prior coding, keyword searching, or known documents of particular high relevance to a particular discovery, coupled with a human-in-the-loop strategy of manually coding whatever seed set exists for relevance or privilege; (iii) employing machine learning software, including most notably support vector machines, to categorize similar documents; and (iv) using some kind of QC process to check for coding consistency [12].

In the much-cited case of *da Silva Moore,* a US federal magistrate judge held that the state-of-the-art in advanced search techniques had progressed to the point where the Court could "bless" the use of a predictive coding protocol in the litigation as submitted by one or both parties [5]. In his February 24, 2012 watershed opinion, Magistrate Judge Andrew Peck writes:

> In this case, the Court determined that the use of predictive coding was appropriate considering (1) the parties' agreement, (2) the vast amount of ESI to be reviewed (over three million documents), (3) the superiority of computer-assisted review to the available alternatives (*i.e.*, linear manual review or keyword searches), (4) the need for cost effectiveness and proportionality . . .; (5) the transparent process proposed by [defendants].
>
> This Court was one of the early signatories to The Sedona Conference Cooperation Proclamation, and has stated that 'the best solution in the entire area of electronic discovery is cooperation among counsel. . . .' *An important aspect of cooperation is transparency in the discovery process. [Defendants] transparency in its proposed ESI search protocol made it easier for the Court to approve the use of predictive coding. . . . [Defendants] confirmed that all of the documents that are reviewed as a function of the seed set, whether they are ultimately coded relevant or irrelevant, aside from privilege, will be turned over to plaintiffs. … If necessary, counsel will meet and confer to attempt to resolve any disagreements regarding the coding applied to the documents in the seed set. While not all experienced ESI counsel believe it necessary to be as transparent as [defendant] was willing to be, such transparency allows the opposing counsel (and the Court) to be more comfortable with computer-assisted review, reducing fears about the so called 'black-

> box' of the technology. This court highly recommends that counsel in future cases be willing to at least discuss, if not agree to, such transparency in the computer-assisted review process.

The magistrate judge's opinion allowing the use of "predictive coding" was subsequently affirmed by a federal district court judge [5]. An Order to the same effect also has been rendered in a state court proceeding in Virginia, where the Court issued a protective order allowing a responding party in discovery to use predictive coding over the objections of the requesting party [14]. In still another case in Illinois, multiple days of evidentiary hearings were held with expert testimony describing the pros and cons of using predictive coding, where the requesting party had moved to compel essentially "starting over" using such method --even after over a million documents have been located by a responding party using keyword searching and other traditional means [15]. The parties settled their search methods dispute in that case before an opinion was rendered.

The extraordinarily detailed protocol in *Moore,* attached as an appendix to the February 24, 2012 opinion [5], contains provisions for seed sets of documents generated through a combination of random and judgmental sampling, followed by up to *seven* iterative rounds of "training" the system, through a commitment by counsel to share both responsive and nonresponsive documents by "issue tag" categories. The protocol further provides for sampling at the back end of the initial training period to function as a QC check on excluded or irrelevant documents, to determine how well the trained system has done in coding accurately making those exclusions. (A similarly detailed joint protocol on predictive coding subsequently has been adopted in the *In re Actos* case out of Louisiana [28] .)

What the *Moore* protocol does not purport to explain, however, is the "black box" mathematical algorithms used in predictive coding or software-assisted method, which the judge in *Moore* more or less took on faith. It may be useful, therefore, to have an explanation at hand on what the mathematics of predictive coding entails, and why the protocol adopted by the Court in *Moore* does in fact represent best practice when using this technology, especially with respect to the issue of classifying documents as responsive or not.

4. Support Vector Machines: A Look Under The Hood

In order to develop a sense of how support vector machines (SVMs) and similar algorithms operate, one must at least consider the following questions. First, how do computers represent a lawyer's annotations of relevance on documents in a seed set? Second, how can annotations distinguishing relevant and irrelevant documents in the seed set enable the SVM to make the same distinction in a body of unread documents? Third, what are some complications that could arise in attempting to perform classifications between relevance and nonrelevance? After an elementary tutorial in section 4, we will go on in section 5 to ask are there ways in which legal professionals should alter traditional practices to achieve the full benefits of SVM-type technologies?

4.1 Separating Relevant and Irrelevant Data Using a Computer Algorithm

When a lawyer reviews potential documents in discovery, she is expected to have a good idea whether the document will be meaningful to the litigation or not – based on past legal experience and specific training on the issues arising in a particular case. For computers the process of determining relevance is less obvious. But, as shown by a growing number of studies, if trained by a lawyer and equipped with an SVM, a computer can estimate with remarkable accuracy whether or not a document will be relevant to a particular case, potentially saving legal professionals' valuable time [16]. To better understand how SVMs do this, we will start from a notion of documents as points in space, analyze how a computer could separate such points with a line, determine which separating line the computer could choose, and generalize our simple model to more complex searches.

SVMs can use the word content of documents to map each document within a corpus or seed set to a point in a coordinate space [17]. SVMs can also map documents using metadata [18] and relevant features derived from probabilistic latent semantic indexing [19].

For the sake of simplicity, suppose one is painting a house blue and only cares about the keywords "blue paint" and "maintenance." Place the frequency (representation as a percentage of total words) of the phrase "blue paint" on the X-axis and the frequency of "maintenance" on the Y-axis, such that both are increasing as one moves out from (0,0). Unless two documents are lexically equivalent up to the order of words, each document will correspond to a unique point in space. Figure 1 demonstrates this simplified model.

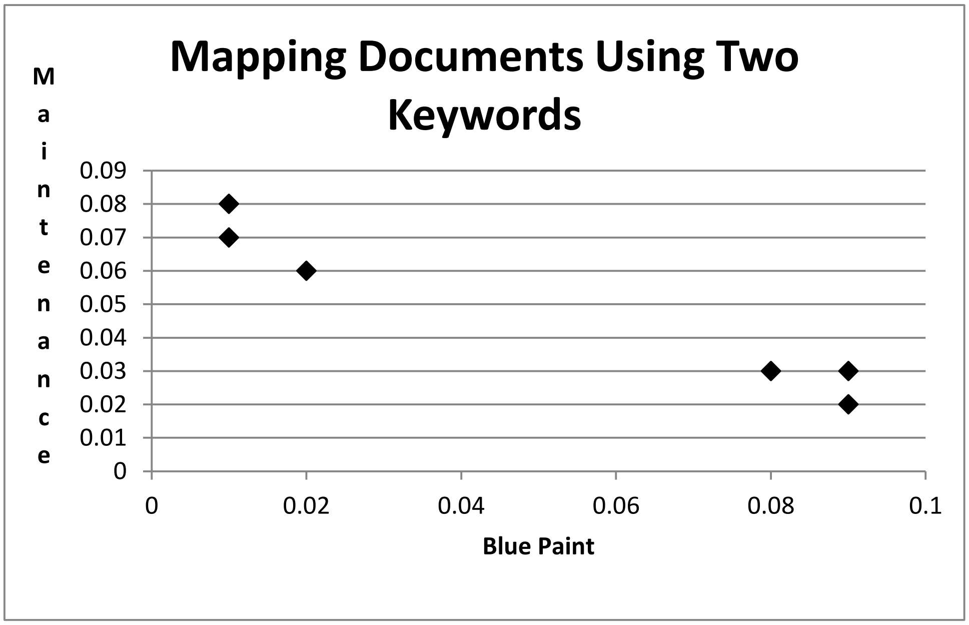

Fig 1. Documents mapped to points using word content

Now, one can understand how a lawyer would train an SVM. Out of potentially millions of articles, the SVM might give a lawyer seed sets of as few as fifty and as many as a few hundred at a time to analyze for relevance, up to some designated cumulative cap of several thousand documents to be judged overall. These documents are called the "seed set" [19]. Seed sets are often selected in one of two ways. The SVM might draw a random sample of documents from the entire body of documents. Or, the seed set could be selected from the results of a judgmental search performed within the corpus (e.g. using keywords). Using either way, or some combination of both, once a seed set is determined, the lawyer identifies or codes documents as either relevant or irrelevant. SVMs incorporate these annotations of relevance into their spatial representation of documents. In figure 2, we imitate this coding process by using clear squares to denote irrelevant documents and black diamonds to denote relevant documents.

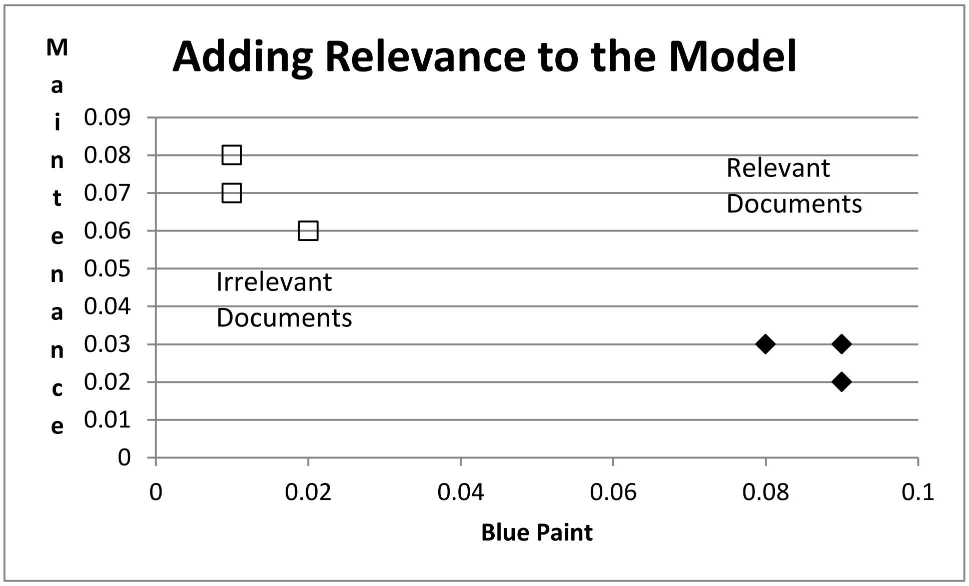

Fig 2. Figure 1 modified to incorporate relevance

Now we will develop an algorithmic notion of separation of articles based on relevance. In this case, relevant and irrelevant data are clustered together. Documents that disproportionately feature the word "maintenance" turn out to be about general home maintenance, and do not pertain to our research about maintaining the quality of a paint job. All other articles were helpful in some way. As figure 3 shows, there is more than one way to spatially divide these documents based on relevance. The divisions in figure 3 are clear because the data are nicely clustered. But, in fact, there is always more than one way to spatially divide coded documents no matter how entangled relevant and irrelevant documents are in the graphical space [20]. That process is explained later.

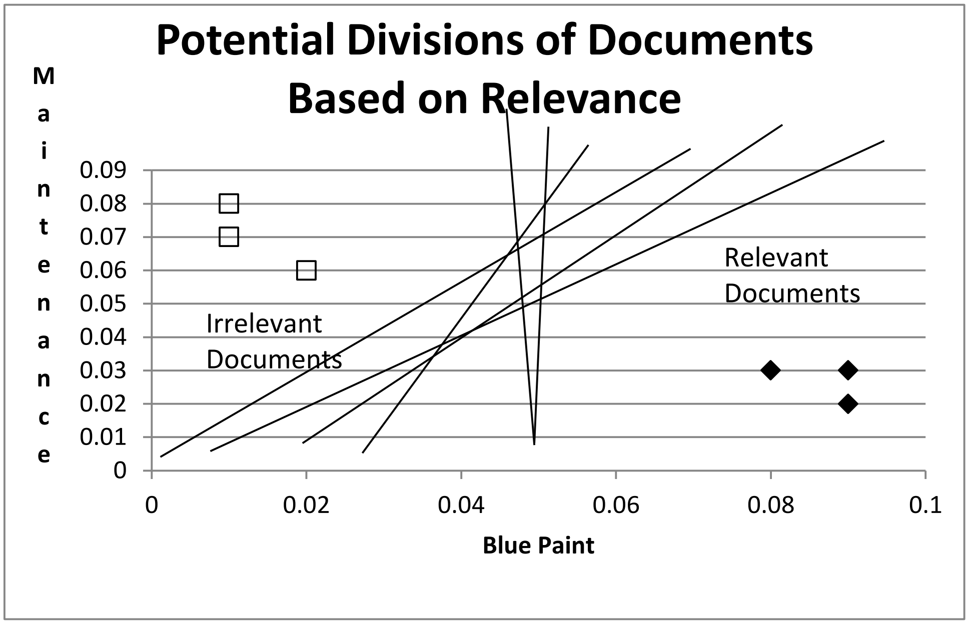

Fig 3. A subset of possible divisions of relevant and irrelevant documents

The non-uniqueness of the separating line presents a potential problem: which line should the computer choose? It should choose the line that preserves the maximum distance between both bodies of data. To see why, suppose it does not do so. Then, the computer line is fairly close to at least one of our clusters. For convenience, suppose it is closer to relevant documents. Now consider what happens when one uses the data on the opposite side of this line – data deemed irrelevant by the SVM. Note that under the specified mapping system, documents that are graphically proximate have similar lexical content. So, one might expect that a document that is spatially "close" to a relevant document to also be relevant. Therefore, a separating line unnecessarily close to the relevant cluster is more likely to place a potentially relevant document on the irrelevant side of the separating line. In this circumstance, the SVM might dismiss a relevant result as irrelevant, which neither counsel wants. To abate this problem, the SVM selects the line that maintains a maximum distance between both clusters of data [21]. The maximum distance criterion specifies a unique separation line. Figure 4 provides an example of a maximum margin solution.

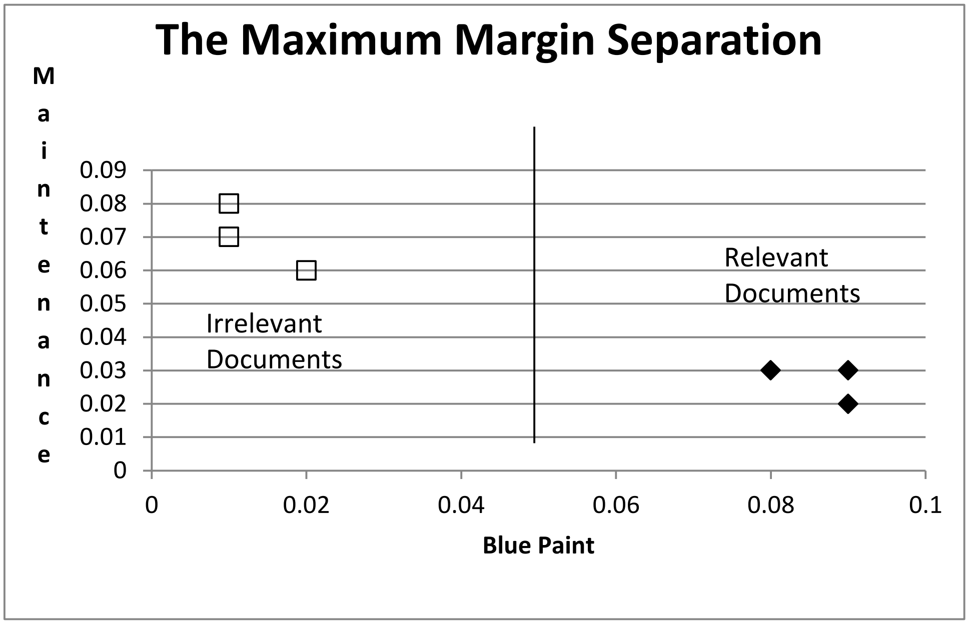

Fig 4. The maximum margin solution is least prone to error of all possible separating lines.

Models are rarely as simple as the artificial example provided above. There are three major generalizations of which one should be aware.

First, if one cares about more than two search terms, each point gains more coordinates and is thus positioned in a higher dimensional space. Suddenly, drawing a line is no longer an adequate way to separate two points. For example, if one wants to separate points in three dimensions, one uses a plane. Think of an umbrella as a small plane that separates points that are raindrops from points that are a person's skin, clothes, and hair. If the umbrella had no width, like a line or no dimension, like a point, it would not adequately separate the two sets of points in the three-dimensional universe. It needs to be at least two dimensional or the person carrying it will get soaked. So, the plane is the higher dimensional analogue of the line in terms of its ability to separate data in three dimensions. Yet, most searches will deal with more than three search terms and thus the input space for those searches will be higher than three dimensional. At this point, one loses the ability to easily visualize the space in which points representing documents lie. Moreover, as the space increases in dimension, one needs higher dimensional analogues of planes to separate points within the space. Mathematicians call these structures "hyperplanes" [20]. Visualizing hyperplanes is not important; having the intuition that hyperplanes perform the same function as separating lines in two dimensions is.

The second generalization is that sometimes the structure that separates clusters while maintaining maximum distance is a curve, rather than a straight line. In these cases, SVMs use so-called "kernel functions" to derive a curve that separates the sets of points [20]. This process will be explained *infra.*. Figure 5 gives an example of a separating curve.

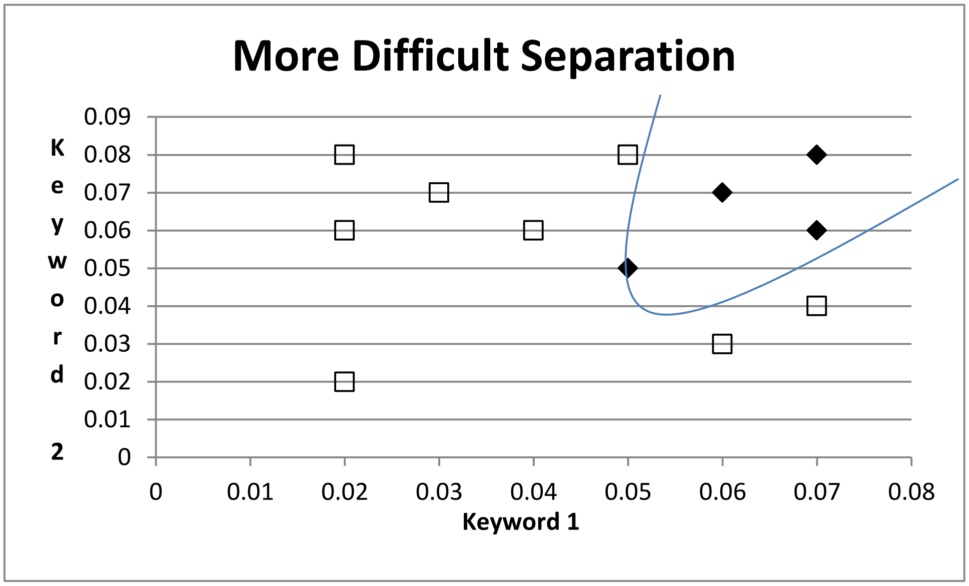

Fig 5. Using a curve to separate more entangled data

Third, in the case of both hyperplanes and separating curves, one still wants to maintain maximum distance from both clusters of data. Failure to do so has the same negative consequences in high and low dimensions: a high risk of obscuring desired results or signaling false positives[2].

4.2 Using Separating Spatial Constructs to Filter Future Results

SVMs are powerful because they can predict whether a document will be relevant even if no lawyer has performed "eyes-on" manual review of that document. This section explains how SVMs predict the relevance of unobserved documents.

An SVM can quickly map an unread document to a point in space by counting the keywords present in that document as a proportion of total words. This point will either lie on the relevant side of the line or the irrelevant side of the line. If the document falls on the relevant side of the line, the SVM will keep the document and notify the lawyer that it is relevant. If the document falls on the irrelevant side outside of the range of potential ambiguity, the SVM will discard it, reducing the lawyer's potential workload.

In higher dimensions, the position of the point with respect to the line might not be as obvious. So, SVMs use more general distance formulas. This will give the distance between an unobserved document and the hyperplane a positive or negative parity. The parity corresponds to which side of the hyperplane the document lies on. The side of the line on which the document

[2] Placing the separating nyperplane too close to the irrelevant cluster creates a risk of falsely identifying irrelevant documents as relevant.

lies informs the SVM about whether or not the document is likely to be relevant. So, even in higher dimensions SVMs can discern the relevance of a document using distance formulas. Distance formulas can even generate signed values of distance if the dividing hyperplane is curved.

4.3 Potential Complications and Their Solutions

Five potential complications arise in the use of SVMs to classify documents in the hyperplane: seemingly inseparable data; statistical outliers; data points that are close to or are contained in the separating hyperplane that divides relevant and irrelevant documents; the necessity of sorting documents into more than two categories; and the introduction of new documents.

Dealing with seemingly inseparable data. Sometimes, data will appear to be inseparable. These cases are best illustrated through an example. Suppose one is interested in a new tax law and that one only seeks to use the keyword “tax”. After parsing a set of seed documents, a lawyer finds that documents that contain “tax” as 0 – 3% of the total words are only tangentially related to his research and tend to be irrelevant. In contrast, documents in which “tax” represents 4 – 6% of the total words tend to be relevant. However, documents in which “tax” represents 7% or more of the total word count tend to be merely descriptive and do not provide the deep analysis the lawyer seeks. Figure 6 is a graphical representation of this apparent dilemma.

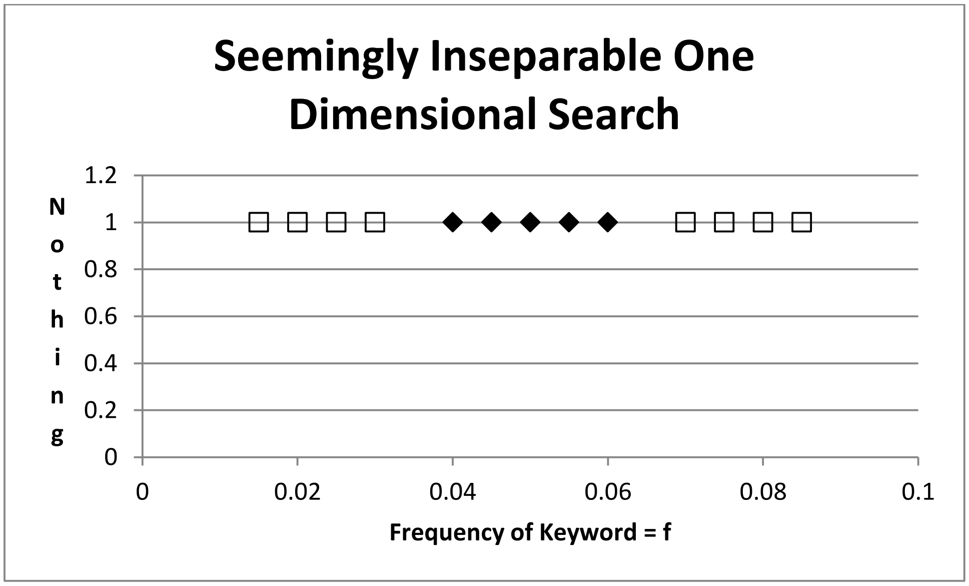

Fig 6. No single point can separate relevant data from irrelevant data well.

That there are two clusters of irrelevant documents on either side of the relevant documents makes it unclear where one should draw the separating line, which in this one-dimensional case would just be a point.

To solve this problem, SVMs use kernel functions. Kernel functions project data into higher dimensional spaces. Surprisingly, given a data set in which no two identical objects have opposite labels, there is always a kernel function that will allow the data to be linearly separated. In fact, this projection into higher dimensional space is equivalent to curving the separating hyperplane [20]. So, separation using a curved hyperplane is never necessary as a non-curved hyperplane can always separate the data in some dimension[3].

Consider the previous example. Suppose we projected our one-dimensional set of data into two dimensions. If f is the frequency with which “tax” appears in every hundred words, on average, then create a two dimensional graph mapping each document to f and $(f - .05)^2$. Graph the first dimension on the X-axis, the second on the Y-axis. Now, instead of a line, one has a parabola. Also, the model has become two dimensional. So, the separating geometric construct becomes a line instead of a point. Figure 7 shows that this new set of data can easily be separated with a line.

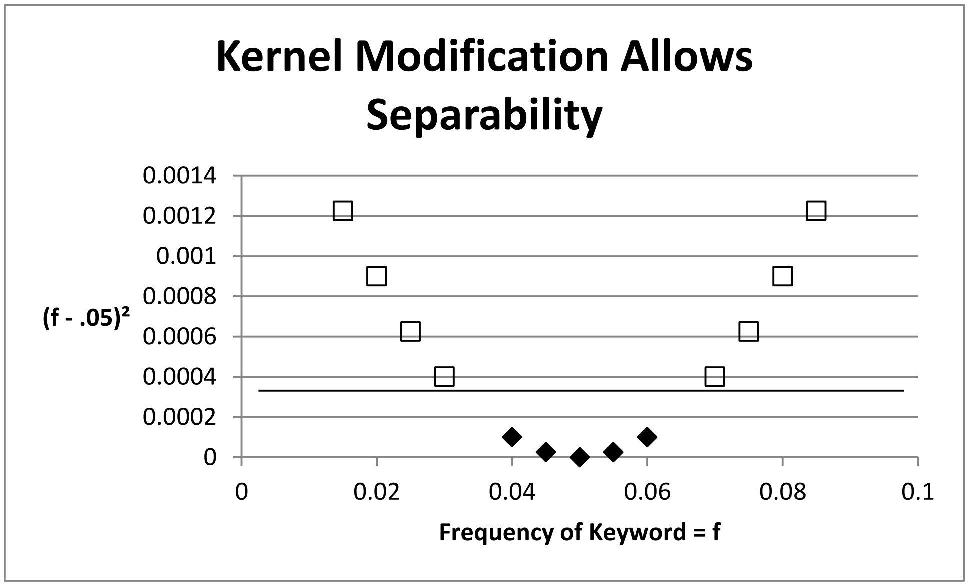

Fig 7. A solution to the problem posed by figure 6

By projecting points representing documents into higher dimensional space, it is always theoretically possible to linearly separate relevant from irrelevant documents using a non-curved hyperplane. Then, from the set of separating hyperplanes, an SVM could choose the one that maintains the maximum distance between both clusters of data. Although there is always a function that can separate relevant from irrelevant documents, some such functions are so complex that they are computationally intractable. In fact, most SVMs are only packaged with a few kernel algorithms to create kernel functions. In the cases that these packages fail to find a perfect separation function, the SVM will use a computationally feasibly kernel that separates most of the data with the maximum margin but accepts a "soft margin" of error.

Dealing with Outliers. There might be a few relevant documents that are surrounded by irrelevant documents or vice versa. This might for two reasons. The documents might be genuinely relevant (or irrelevant) even though their proportions of keywords do not match up with other documents of their type. Or, the documents could be false positive results; not even expert lawyers can separate relevant documents from irrelevant documents with anything approaching 100% accuracy [16, 22, 23].

To solve this problem, SVMs have a “soft margin” built into their algorithmic structure. This margin dictates how many outliers are allowed to lie on the opposite side of the hyperplane and how far they have to be from the hyperplane to be considered outliers [20].

Dealing with Documents that Lie Close to the Separating Hyperplane. Although most documents can be easily classified based on a lawyer’s coding annotations of the seed set, some classifications are not obvious. In particular, documents that lie close to or on the separating hyperplane are of ambiguous relevance. They are fairly close to both the cluster of relevant data and the cluster of irrelevant data. Thus, irrelevant documents that approach this hyperplane are more likely to be relevant than irrelevant documents that are farther away. The reverse is true for relevant documents. Therefore, this set of documents is most likely to be incorrectly classified by the SVM. A relevant document might be discarded or an irrelevant document might be labeled relevant, harming precision, recall, or both. To reduce the risk of false classification, an “active learning” SVM creates another seed set for the lawyer out of the documents that were left ambiguous by the previous filtering. After each seed set classification, the SVM uses the new inputs provided by the lawyer to create a more precise separation between the two classes of data [24]. In contrast, a “batch learning” SVM creates a new seed set out of random documents that were omitted from both the previous filtering and the previous seed set [24]. The SVM ends either of these iterative processes once it determines that the error that may result from automatic classification will be sufficiently small. In other words, the system “stabilizes” to an acceptable margin of error.

Relegating the task of classifying ambiguous documents to the lawyer means that the lawyer has to sift through more documents than are present in the initial seed set. However, on net, a lawyer who uses an SVM personally classifies significantly fewer documents than one who uses traditional review. In fact, lawyers do not even have to classify all of the documents of ambiguous relevance. If lawyers find more error acceptable, they can sift through smaller seeds of these documents, allow the SVM to record patterns in their classifications, and have the SVM classify the rest of the ambiguous documents.

Adapting SVMs to Sort Documents Into More than Two Categories. Standard SVMs are binary linear classifiers; they use lines (or their n-dimensional analogues, i.e., hyperplanes) to separate data into two categories. Yet, documents might need to be sorted by more than one criterion and divided into more than two sets. For example, lawyers may be interested in whether a document contains Personally Identifiable Information (PII) in addition to whether that document is relevant. To solve this problem, the SVM would simply make two binary classifications. One would separate the relevant documents from the irrelevant ones. The other would discern which documents are likely to have PII and which probably do not contain PII. Then, each document has two labels (or "issue tags," in the vernacular used in the *Moore* protocol), and the documents can be separated into four categories: PII relevant, PII irrelevant, non-PII relevant, and non-PII irrelevant. If there are n potentially important features a document can have, an SVM would do n binary classifications and use the results to create 2^n categories of documents [25].

Consider the following SVM: documents are mapped according to two keywords and then classified based on: (i) whether they are relevant; and (ii) whether they contain PII. Relevant documents are shaded black; irrelevant documents are clear. These two categories are separated by a vertical line. Documents containing PII are squares; documents without PII are diamonds. These two categories are separated by a horizontal line. Figure 8 depicts this dual division.

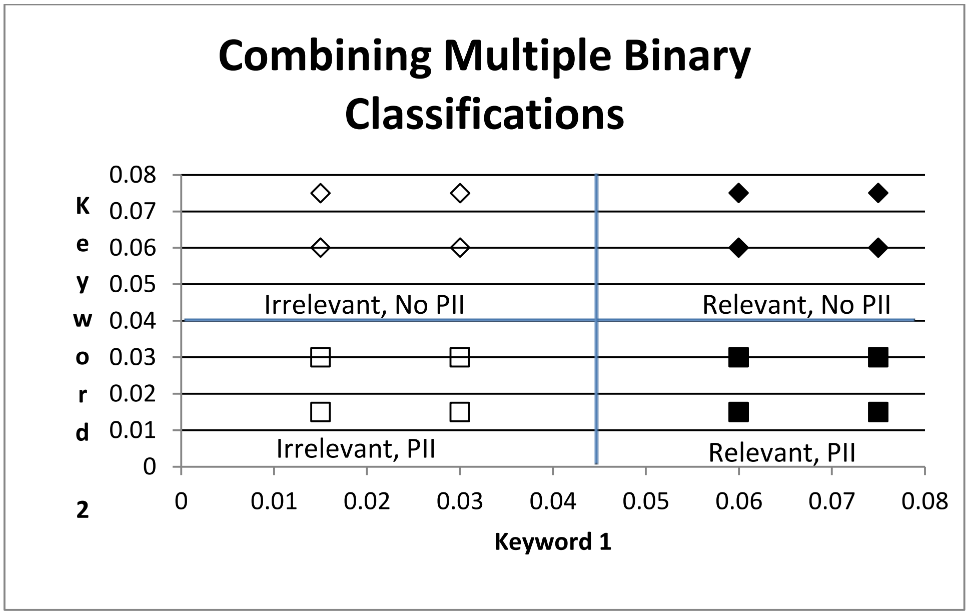

Fig 8. A basic example of division based on multiple criteria

Introduction of new documents. Finally, suppose new unlabeled documents are introduced. Then, cooperating counsel may agree to feed these new documents to an SVM, which has two benefits. First, after the SVM classifies these new documents, lawyers may program it to look for new "issue tags," that are highly correlated with relevance or irrelevance. Incorporating these tags as an additional proxy for relevance can improve both the current model and future filtering efforts.

This would allow both parties to channel the accuracy and efficiency of an SVM as new facts emerge to ensure the SVM best suits their needs. Second, independent of the chance of discovering a new, relevant issue tag, electronically sorting new documents will be faster and potentially more accurate than manual review [16].

5. Optimizing The Benefits of SVMs in Search Protocols

SVMs are useful because they hold out the potential to be more efficient and effective than other review methods. As the comprehensive RAND Study, “Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery” concludes, the answer is “not entirely clear” given the lack of present data points what the magnitude of savings there is to be achieved by using predictive coding methods as compared with other hybrid forms of automated and manual review [16] at p.66. However, as the RAND report also emphasizes, “predictive coding in large-scale discovery review has the potential to yield significant cost savings without compromising quality as compared with that provided by a human review.” [16] at p. 71.

This potential will, in our view, be more rapidly fulfilled as lawyers consider the benefits of greater cooperation and transparency, as Judge Peck and his colleagues have urged. [26] To this end, we make the following observations about process and protocols.

First, lawyers need to conceptualize the e-discovery process as involving multiple iterative feedback loops, where input from an opposing party is desirable in order to fine-tune the production of relevant documents. As first noted in [6], this process involves multiple meet and confers, in which sample sets are provided of the results of an automated search, with opportunity given for choices being made by opposing counsel on what constitutes the documents of greatest interest returned in the first, second, or subsequent sample.

Second, as set out in the *Moore* protocol, the SVM algorithm fairly demands that good exemplar candidate documents from both the “relevant” and “irrelevant” universes be agreed to, in order that the sophisticated machine learning techniques described above in section 4 can take place. Importantly, it turns out the computer achieves the greatest gains in learning through active learning processes such as re-seeding documents that are “closer” to the classifier hyperplane [27]. This represents a challenge, one that the parties in *Moore* may not have fully anticipated, when nominally agreeing to discuss the classification of documents into responsive and nonresponsive piles.

Unquestionably, the idea that a protocol would require the turning over nonprivileged, irrelevant documents, in order to optimize training of a machine learning algorithm, is fairly unprecedented outside of the *Moore* and *In re Actos* protocols. However, absent building in that specification, it is not difficult to imagine many situations where counsel for one party who may have insisted on using predictive coding (as in the case of the responding party in *Global Aerospace*), ends up over-training the system to fit a one-sided conception of “relevance” in the litigation. In other words, absent agreement on what is considered irrelevant, especially in hard cases, there is much greater potential for going off course. However, as Judge Peck anticipated, there will be participants in litigation that strongly object to the intentional turning over of any irrelevant documents, and/or a greater number of documents than absolutely required, regardless of circumstances. Over time, however, as more judges would be expected to adopt similar protocols urging cooperation between parties, resistance in the profession (and among clients) may lessen. A recent article in *Metropolitan Corporate Counsel* [29] observed:

> It remains to be seen whether corporations will embrace predictive coding with the levels of transparency involved in *Da Silva [Moore], Actos* and [*Global Aerospace]*. Some corporations will clearly be motivated by the potential cost savings. They may limit the matters they are willing to be transparent to those that they know are unlikely to involve the production of sensitive documents. Others may embrace transparency because they figure that the volume of irrelevant documents to be produced during the predictive coding training process will be relatively small and thus the risk low or they figure the problem of producing irrelevant documents can be controlled with a protective order or confidentiality agreement.

Given how novel the propositions discussed in this paper are, it is perfectly understandable that many lawyers will attempt to avoid any obligation that arises to engage with the other side in negotiations that include reaching agreement on the sharing of nonrelevant documents in connection with a protocol on advanced search techniques. See [30] for a further discussion of "forced" disclosure vs. voluntary disclosure of irrelevant documents when engaging in a predictive coding process. One day, however, courts may more routinely be in a position to rule that the failure to adopt such methods and protocols is unreasonable, i.e., that a process that goes so far as to transparently reveal both relevant and nonrelevant documents in the seed and training sets represents a benchmark of some kind for what is considered an "adequate" or "reasonable" response to a party's discovery obligations. If more lawyers take the time to understand the underlying mathematics, as well as the sophisticated joint protocols that have been proposed, they arguably will benefit from the realization that classification is a two-sided proposition, demanding appropriate attention to *all* documents in a given repository or data set in order that machine learning technologies can be fine-tuned or optimized appropriately.

References

1. Pension Comm.of Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC, 685 F. Supp. 2d, 456 (S.D.N.Y. 2010).
2. The Sedona Conference, The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery ("Sedona Search Commentary"), Sedona Conf. J. 8:189 (2007). https://thesedonaconference.org/publications
3. The Sedona Conference, The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process, Sedona Conf. J. 10:299 (2009). https://thesedonaconference.org/publications
4. The Sedona Conference, The Sedona Conference Cooperation Proclamation, Sedona Conf. J. 10:331 (2009). https://thesedonaconference.org/publications
5. Moore et al. v. Publicus Groupe SA, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012) (Peck., M.J.), aff'd, 2012 WL 1446534 (S.D.N.Y. April 26, 2012) (Carter, J.)

6. Paul, G.L., Baron, J.R., "Information Inflation: Can the Legal System Adapt?," Richmond J. Law & Tech., 13:10 (2007) http://jolt.richmond.edu/v13i3/article10.pdf
7. Beckerman, J.S., "Confronting Civil Discovery's Fatal Flaws," Minnesota L. Rev., 84:505 (2000). http://www.vallexfund.com/download/Confronting_Civil_Discovery_Fatal_Flaws_2000.pdf
8. Hickman v. Taylor, 329 U.S. 496 (1946).
9. Metropolitan Opera Ass'n, Inc. v. Local 100, Hotel Employees and Restaurant Employees Internat'l Union, 212 F.R.D. 178 (S.D.N.Y. 2003).
10. Blair, D.C., Maron, M.E., "An evaluation of retrieval effectiveness for a full-text document-retireval system," Communications of the ACM 289 (1985). http://opim-sun.wharton.upenn.edu/~sok/papers/b/blair-maron.pdf
11. Oard, D.W., Baron, J.R., Hedin, B., Lewis, D.D., Tomlinson, S., "Evaluation of information retrieval for E-discovery," Artificial Intelligence and Law, 18:347 (Springer 2010) (citing to research from the TREC Legal Track, http://trec-legal.umiacs.umd.edu/).
12. Baron, J.R., "Law in the Age of Exabytes: Some Further Thoughts on'Information Inflation' And Current Issues in E-Discovery Search," Richmond J. Law & Tech., 17:9 (2011). http://jolt.richmond.edu/v17i3/article9.pdf
13. Mt. Hawley Ins. Co. v. Felman Prod., Inc., 271 F.R.D. 125 (S.D. W.Va. 2010).
14. Global Aerospace Inc., et al. v. Landow Aviation, L.P., et al., 2012 WL 1431215 (Va. Cir. Cit. April 23, 2012) (order approving use of predictive coding in discovery)
15. Kleen Prods, LLC v. Packaging Corp. of Am., Docket 1:10-cv-05711 (N.D. Ill.) (plaintiffs' motion pending to compel use of predictive coding in discovery).
16. RAND Corporation, "Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery" (2012), http://www.rand.org/pubs/monographs/MG1208.html.
17. Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." Universitat Dortmund, 1998. Web. 07 Apr. 2012. http://www.cs.iastate.edu/~jtian/cs573/Papers/Joachims-ECML-98.pdf
18. Han, H., Manavoglu,E., Zha, H., Tsioutsiouliklis, K.,Giles,C.,and Zhang, X., "Rule-based Word Clustering for Document Metadata Extraction." Proc. of ACM Symposium on Applied Computing, Santa Fe, New Mexico. (2005). 1049-053. http://clgiles.ist.psu.edu/papers/SAC-2005-Document-Metadata-Extraction.pdf
19. Yan, H., "Support Vector Machines for Text Categorization Based on Latent Semantic Indexing." (2001). http://www.isn.ucsd.edu/courses/774/2001/lsa.pdf
20. Noble, William. "What is a Support Vector Machine." Nature Biotechnology 24.12 (2006): 1565-567. . http://www.broadinstitute.org/annotation/winter_course_2006/index_files/Noble%202006%20SVM%20tutorial%20Nat%20Biotech.pdf
21. Orbanz, Peter. "Support Vector Machines." Cambridge University, Cambridge. 09 Apr. 2012. Lecture. http://mlg.eng.cam.ac.uk/porbanz/teaching/slides_ml__svm.pdf
22. Roitblat, H.L, Oot, P., Kershaw. A., "Document Categorization in Legal Electronic Discovery: Computer Classification v. Manual Review," J. Am. Soc'y for Info. Sci. & Tech., 61:70 (2010). http://www.clearwellsystems.com/e-discovery-blog/wp-content/uploads/2010/12/man-v-comp-doc-review.pdf
23. Baron, J.R., Oard, D., Lewis, D., TREC Legal Track 2007 Overview, Proceeding of the 15th Annual Text Retrieval Conference, National Institute of Standards and Technology, http://trec-legal.umiacs.umd.edu
24. Burl, M.C.., Wang. E.,"Active Learning for Directed Exploration of Complex Systems." *Proceedings of the 26th International Conference on Machine Learning*. International Conference on Machine Learning, Montreal, Canada. 2009. 4. Web. http://cubs.buffalo.edu/govind/CSE705-SeminarPapers/2.pdf
25. Nayak, P. Raghavan, P., and Mooney.R., "Information Retrieval." Computer Science 276: Introduction to Information Retrieval. Stanford University, Palo Alto, CA. 06 Apr. 2012. Lecture. http://jolt.richmond.edu/v18i3/article8.pdf
26. Waxse, D.E.., "Cooperation--What Is It and Why Do It," Richmond J. Law & Tech (2012), 18:3.
27. Tong, S., Koller, D., "Support Vector machine Active Learning with Applications to Text Classification." Journal of Machine Learning Research (2001): 45-66. Print. http://www.ai.mit.edu/projects/jmlr/papers/volume2/tong01a/tong01a/pdf

28. In re Actos (Pioglitazone) Products, MDL No. 6-11-md-2299 (W.D. La. July 27, 2012)
29. Solomon, R., "Are Corporations Ready To Be Transparent And Share Irrelevant Documents With Opposing Counsel To Obtain Substantial Cost Savings Through The Use of Predictive Coding," Metropolitan Corporate Counsel 20:11 (Nov. 2012). Print. http://www.metrocorpcounsel.com/articles/21076/are-corporations-ready-be-transparent-and-share-irrelevant-documents-opposing-counsel
30. Losey, R., "Keywords and Search Methods Should Be Disclosed, But Not Irrelevant Documents" (May 26, 2013), http://e-discoveryteam.com/2013/05/26/keywords-and-search-methods-should-be-disclosed-but-not-irrelevant-documents/.