Predictive Coding:
Explanation and Analysis of Judicial Impact and Acceptance Compared To Established E-Discovery Methodologies

Dane Warren Henry

Dickinson School of Law
Pennsylvania State University
333 W. South Street
Carlisle, PA 17013
dwh190@law.psu.edu

ABSTRACT

Predictive Coding is a new and emerging method for searching and collecting discoverable data. Building upon previous methods, Predictive Coding is a stark advancement in the ability for computers to recognize discoverable relevant data from large data collections. As the quantity and complexity of Electronically Stored Information (ESI) increases, so, too, does the need for faster and more accurate tools to identify responsive materials for discovery. Predictive Coding uses a mixture of human oversight and various computer algorithmic processes to rapidly and accurately retrieve responsive documents from large corpuses. Unfortunately for its proponents, Predictive Coding often leaves much to the imagination in detailing and proving the processes by which it establishes which documents are responsive. Modern courts are only now beginning to address whether this 'black-box' approach will prove to be an insurmountable obstacle to the adoption of this new technology.

Outline

1. Introduction

Discovery, is "[t]he act or process of finding or learning something that was previously unknown... that relates to the litigation."[1] Where traditional discovery is the unearthing of physical materials such as papers or verbal accounts such as depositions, E-Discovery is the locating and collecting of electronically stored information, or ESI. Since the 1970's, ESI has become an increasingly important aspect of litigation as more and more litigants are utilizing computers for information production and retention. This prevalence has virtually guaranteed that almost every recent case is going to involve some sort of E-Discovery. Along with this increased usage of E-Discovery is the need for proper and efficient E-Discovery tools.

Predictive Coding is an E-Discovery tool that is coming into prominence rapidly due to its purported effectiveness and controversial nature. In its purest form, Predictive Coding is an implementation of machine learning, the beginnings of artificial intelligence. Machine learning is exactly what it sounds like: a computer is able to take a piece of disparate information and relate it to other pieces of information that is stored in its memory, adding that new information to its memory for use in future comparisons. The ability to use such a tool in the E-Discovery arena is a great addition to the arsenal of any litigator. Such high expectations do not come without shortcomings, however, and Predictive Coding is no exception. A general lack of a cohesive definition coupled with the fact that courts have been hesitant to allow for the use of Predictive Coding means that while Predictive Coding does have the capability to enhance E-Discovery for years to come, there is still substantial risk in the investment of such a tool.

The course of this discussion will provide understanding for the processes Predictive Coding follows, coupled with the guiding principles of discovery set forth by the Sedona Conference, to shed light on the reasoning behind Predictive Coding's support. The complex

[1] Black's Law Dictionary (9th ed. 2009), Link.

difficulties facing Predictive Coding will be addressed by laying an initial foundation of both paper and E-Discovery. Predictive Coding will then be explored in detail regarding both its strengths and its weaknesses. Particular attention will be paid to the idea that Predictive Coding appears to be too 'magical' to be allowed to assist in discovery, giving rise to the 'black box' argument.[2] Finally, an analysis of the judiciary's reception to Predictive Coding and the likely outcomes of future challenges to the technology will be discussed.

1.1 Discovery Rules and Authority

1.1.1 Federal Rules of Civil Procedure with Regards to E-Discovery

In general, the Federal Rules of Civil Procedure cover only the disclosure and production of ESI, without any direct input as to the methods by which that information is located or collected.[3] Fed. R. Civ. P. 26(b)(2)(B) and 34 specifically detail the expectations and requirements for the production of ESI,[4] such as requiring producing parties to convert 'legacy' information to a current readable format prior to producing the information.[5]

The Committee notes on the 2006 Amendments to Fed. R. Civ. P. 26 details the trouble the committee had in defining what is and is not reasonable in terms of electronically stored information.[6] The committee notes simply say that because the information is electronic, there is an inherent expectation that the information is searchable and therefore reasonably able to be produced.[7] Despite these comments, the committee notes still do not define how a responding party should go about actually searching for discoverable documents.

---

[2] *See* Craig Ball, Got TAR?, Ball In Your Court (September 4, 2012), Link.; Andrew Peck, Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, Law Technology News (Online), October 1, 2011, Link.
[3] Mark S. Sidoti, Wendy R. Steinand & Verne A. Pedro, Challenging 'Manual' ESI Collections, Law Technology News (Online), April 9, 2010, Link.
[4] Id.
[5] Fed. R. Civ. P. 34 Notes of Advisory Committee on 2006 amendments, Link
[6] Fed. R. Civ. P. 26 Notes of Advisory Committee on 2006 amendments, Link.
[7] Id.

It should be noted that the Federal Rules of Civil Procedure also touch on an area that is going to be important in determining whether a given search method, such as Predictive Coding, is going to be allowed for purposes of discovery. While not expressly stated in the rules, the Committee's notes on the Amendments once again provide illumination as to the purpose behind the language. The committee notes say that when it comes to discovery in modern litigation, the entire process should be fostered by a sense of cooperation.[8] Parties, typically during the Fed. R. Civ. P. 26(f) 'meet and confer' conference, need to come to an understanding on scope of discovery and the types of data that is going to be expected from either side. This mutual understanding of terms is an important factor in determining many aspects of discovery disputes, such as consideration for sanctions as well as a party's reasonableness regarding their actions.

1.1.2 The Sedona Conference

Picking up where the Federal Rules of Civil Procedure left off, the Sedona Conference outlined a series of important ideas for the modern litigator involved in discovery. The Sedona Conference is a nationally recognized group of intellectuals and professionals whose primary purpose is to bring a practical understanding of evolving issues of law by working closely with policy makers and active members of the profession.[9]

The Sedona Conference has published a number of articles that are intended to assist the profession by offering guidance for proper and ethical discovery procedures.

First, the Sedona Conference Cooperation Proclamation solidified the comments contained in the Federal Rules of Civil Procedure by calling for a formally signed and recognized agreement to 'play fair' when it comes to discovery, particularly E-Discovery.[10] The

[8] See, e.g., 1993, 2000, 2006 Amendments to Federal Rules of Civil Procedure, Advisory Committee Comments to Rules 26, 33, 34, 37.
[9] The Sedona Conference, About Us, https://thesedonaconference.org/aboutus (last visited Nov 6, 2012).
[10] The Sedona Conference, The Sedona Conference Cooperation Proclamation, 10 Sedona Conf. J. 331, 2009, Link.

Cooperation Proclamation recognized that there is a perceived conflict between the Model Rules of Professional Conduct Rule 1.3 (Diligence),[11] which in the comments calls for "zeal in advocacy of the client's interests,"[12] and the proposed cooperative discovery process. The Proclamation reconciled this conflict by noting the ever increasing cost associated with litigation, particularly E-Discovery, and how it is in the client's best interests to ensure that a fast, smooth, and open discovery is conducted in order to get to the true issue before the court.[13] For example, if the plaintiffs discover a document held by the defense that removes liability, the suit could be dropped or settled then and there. If, instead, a lawyer was to "hide the ball" when it came to discovery, not only would it increase the cost of the discovery itself, including the court costs if discovery were compelled, but the overall litigation costs would increase because the key information was never found and the case moved forward. At its root, the Sedona Conference knows that cooperation is not going to begin overnight but believes that with awareness, commitment, and conflict resolution tools, litigators can move in the right direction to open and efficient discovery.

The Sedona Conference has also published a number of articles through its working groups that more directly and concretely assist litigators in various aspects of discovery. Of particular interest is the Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery.[14] Despite having been published five years ago, this article remains astonishingly relevant as it breaks down each of the current methods for search and collection of discoverable ESI as well as the respective shortcomings of each. Most importantly,

---

[11] Model Rules of Prof'l Conduct R. 1.3, Link.
[12] Model Rules of Prof'l Conduct R. 1.3 cmt., Link.
[13] The Sedona Conference, The Sedona Conference Cooperation Proclamation, 10 Sedona Conf. J. 331, 2009, Link.
[14] A Project of The Sedona Conference Working Group on Best Practices for et. al., The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery August 2007 Public Comment Version, 8 Sedona Conf. J. 189, 200-202, 2007, Link.

this article lays out the quantitative processes that determine the accuracy of a search and collection technology. This process, using the metrics recall and precision, enables the corpus custodians to accurately measure the depth of penetration of the document set (recall) as well as the relevance of the documents that are retrieved (precision).[15] These quantitative measures are important in determining whether a particular search method is more than just words on paper. If a search process can comb through a collection of documents but not locate any responsive documents, there could be an issue with the recall aspect of the algorithm. Alternatively, if the search process is returning a huge amount of documents purporting to be relevant but in fact are not, then the precision of the algorithm needs improvement. Although a search process can sound great in the corporate-speak used by some product descriptions, the proof is in the measurable results produced.

2. Search and Collection Methods

Initially, discovery began with attorneys working with clients sifting through filing cabinets full of documents. It quickly changed to teams of attorneys going through literal warehouses filled with banker boxes full of documents. Clients could no longer afford to pay an attorney the hourly rates necessary to perform the inspection of the documents, so paralegals and temporary employees were utilized to rein in costs. Eventually, the amount of information stored by larger corporations still meant that to enter litigation and go through the paper discovery process was prohibitively expensive.

Computers were initially seen as a fix for the paper volume problem, as one employee could sit at a workstation and have access to any document within even the largest of multinational corporations regardless of where in the world the document was located. Where once there were warehouses full of boxes of paper, now there is a terminal with a connection to

[15] Id. at 205.

the company's central server hosting all the necessary files. As technology progressed, the price-to-storage ratio plummeted. Electronic storage space that used to cost hundreds if not thousands of dollars became obtainable for just a few pennies.[16] This explosion in the raw capacity storing electronic information did little to help the problems of the cash-strapped litigant, as it only provided more ESI that could be discoverable. What was already a burdensome process when involving paper now just became worse as more data was being stored in less space.

2.1 Manual Search & Collection

The idea of an explosion of data brings the manual search and collection process directly into the spotlight. Manual collection involves a human being, be it attorney, paralegal, or temporary worker, looking at each paper and file in the entire corpus. This process is extremely expensive, as it involves a person sitting for hours every day examining document after document to determine its relevance to the issue being litigated. Even after an initial review by an employee, the files are then further reviewed by a licensed attorney to both confirm the relevancy and to determine any type of privilege that might exist within the document. This process is expensive, time consuming, and inaccurate. Considering the sheer volume of data that is stored by not just corporations but even individuals, manual search and collection is becoming an increasingly unviable option.

Despite this, manual collection is still considered by many to be the gold standard to which all other discovery and review processes are compared. However recent studies have shown that not only is manual search falsely portrayed as the gold standard in the industry, it is inherently flawed in both its precision and recall rates.[17] These studies find that manual

[16] *See* Matthew Komorowski, A History of Storage Cost, Link (last visited Dec 8, 2012) (In 1980, the cost per gigabyte of storage was roughly $193,000.00. By 2009 that cost had fallen to $0.07 per gigabyte.).
[17] *See* Maura R. Grossman & Gordon V. Cormack (FNaa1), Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, 17 Rich. J.L. & Tech. 11, (2011), Link.

collection is far inferior to some of even the most basic automated methods being put into practice today.[18] Despite these studies, the idea of allowing anyone or anything other than an attorney, or at the very least, a human being, to determine the relevance of a document is unsettling to a fair percentage of the legal community which is why manual collection is still used to a large extent.

2.2 Traditional E-Discovery Methods

Once computers were used to store and view documents, it only made sense that tools would start to develop that would use the processing power of computers to assist in reviewing those documents. Some of the methods have been almost universally accepted as a tool for discovery, others still seem to need time in order to become understood, let alone accepted.

2.2.1 Keyword Search

Keyword searches are the *de facto* standard search method for ESI.[19] Keyword searches work by allowing a user to input one or more words into a search set which is then compared against the document set. Any documents that contain the word or words that are being searched are then marked as responsive and set aside for later review. While keyword searches are incredibly user-friendly because of their straightforward use, basic keyword searches also suffer from three critical shortcomings. The first problem with keyword searches is ambiguity, also known as polysemy, which effectively means multiple words can describe the same idea or product.[20] For example, the use of the word 'bow' in a document could refer to vastly different ideas depending on whether the document came from music, archery, shipping, or etiquette. This

---

[18] Id.

[19] A Project of The Sedona Conference Working Group on Best Practices for et. al., The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery August 2007 Public Comment Version, 8 Sedona Conf. J. 189, 200, 2007, Link.

[20] A Project of The Sedona Conference Working Group on Best Practices for et. al., The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery August 2007 Public Comment Version, 8 Sedona Conf. J. 189, 201, 2007, Link.

leads to many false positives, as there will be documents that contain the word for which the search was conducted but have nothing to do with the matter at hand.

The second issue with a basic keyword search is a kind of antithesis to polysemy, meaning that if a search is conducted for a particular word, there are often tens of other words that either individually or in sum represents the same idea as the searched-for word.[21] For example, if the search term 'government' is used, within the United States a document describing government could completely omit the phrase 'government' and instead use 'congress,' 'senate,' 'house,' 'legislature,' 'president,' 'executive,' and so forth. If that were the case, then a simple keyword search of that document would not flag it as responsive, and the information could be missed. While this is a fringe case due to the possibility of using multiple keywords, the fact still remains that a basic keyword search matches only the words exactly as they are input.

The last issue is that of simple misspellings either through typographical or transcription errors ('damag' instead of 'damage') or simply different spelling of the same word or name ('Steven' instead of 'Stephen').[22] Under a basic keyword search, the search for 'damage' under the first example would pass over the document containing the word 'damag.' Similarly, searching for 'Stephen' when the person involved spelled his name 'Steven' would overlook any documents regarding that individual. The following improvements or variations on basic keyword searches each have their own strengths and weaknesses, but they all seek to remedy the imperfections inherent in keyword searches.

2.2.2 Boolean Keyword Variation

The most common variation to basic keyword searches is a Boolean search. Boolean searches involve ideas borrowed from fuzzy logic, which means that a combination of words and

[21] Id.
[22] Id. at 202.

connectors can solve the issues of ambiguity and spelling.[23] Boolean operators provide the connectors that allow disparate words to provide context for one another, thereby limiting the scope in which the words can appear. The simplest example of Boolean operators is the '+,' 'and,' or '&' operator, which simply means finding the two terms together. Where there might be one hundred documents that respond to the term 'car' and another one hundred documents that are responsive to the term 'bat,' there might only be twenty documents that have 'car' and 'bat' together. Fuzzy logic attempts to capture the essence of human speech and processes word combinations similarly. Fuzzy logic assists in finding responsive documents by handling the innumerable permutations that words can possess. For example, using the wildcard symbol (*) in conjunction with the word damage will help to find all the variations of the word. Using 'damag*' as the search term will allow the search process to find 'damage,' 'damages,' 'damaging,' 'damaged,' etc. The combination of fuzzy logic with Boolean operators can effectively alleviate most of the problems previously mentioned.

2.2.3 Conceptual Mapping

Conceptual mapping involves more advanced aspects of Computer Science. While a computer cannot, at least for now, truly read a word with any level of comprehension like a human, computers can still be taught to form associations based on predetermined word sets. These word sets can be composed of replacement words with the same meaning (synonyms) or words that when taken in sum can mean the same thing (metonyms).[24] For instance, if a user searches for the word 'basketball,' based on conceptual mapping the computer will also look for the words 'hoop,' 'dribble,' 'shoot,' 'foul,' etc. in the document. This means that if a document

[23] A Project of The Sedona Conference Working Group on Best Practices for et. al., The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery August 2007 Public Comment Version, 8 Sedona Conf. J. 189, 202, 2007, Link.
[24] Id.

contains the words 'hoop,' 'dribble,' and 'shoot,' but the word 'basketball' itself does not, then the computer could be reasonably sure that the document is still responsive because taken in the sum, those words likely mean 'basketball.'

Conceptual mapping helps to alleviate the need to know each and every variation of a word within all the documents that are being searched. It can also find documents that are discussing the keyword without actually using the keyword. This helps to broaden the scope of the search to include what are in reality responsive documents but would otherwise be missed under a basic or Boolean keyword search.

2.2.4 Clustering

Clustering works much the same way that conceptual mapping does, by looking at other words rather than just the keyword, but on a more macro level. What this means is that when a search is executed, the computer will look in the document not just for the word itself but for the location of the word in the document as well as the proximity of other words in their relation to the keyword.[25] Going back to the basketball example, if the word 'basketball' appears in a document only once and there is no mention of any of the other associated words, then that document is not very likely to be responsive, and the word 'basketball' is likely to be in reference to something that has little to do with the litigation. Similarly, if the document references basketball only once but the keyword is surrounded by associated terms related to the keyword, then the document probably has something to do with the litigation.

Clustering is extremely helpful by examining the context in which a word appears in a document to determine relevance. This filtering helps to ensure that the document the reviewer is examining is at least more likely to be responsive than a document that simply matched a single keyword once.

[25] Id.

2.2.5 Deficiencies in Traditional Automated Tools

Traditional tools such as keyword searches and manual collection are the bread and butter for most attorneys. And in most instances, these methods do just fine for discovering responsive documents that can then be reviewed at a later point. Unfortunately, the exponential increase in data storage has overwhelmed the capacities of even the automated searches. As mentioned, the price of storing data has decreased so dramatically that companies and individuals are no longer even attempting to reign in the amount of information that is stored. Unfortunately, or fortunately, depending on how you look at it, almost all this information is subject to discovery during litigation. While a lot of the issues can be resolved through a combination of methods including keyword searching variants and deduplication, there are times when either the keywords are so unique, the concepts so abstract, or the corpus so large that these tools fail the attorney.

Such shortcomings have been noted in cases such as Victor Stanley Inc. v. Creative Pipe, Inc. [26] and United States v. O'Keefe.[27] Both cases involve the inadequacies of keyword searching. The party in Victor Stanley inadvertently disclosed privileged documents despite having searched the corpus for the keywords that would indicate privilege and then performing manual review. The court in Victor Stanley noted that because the party refused to agree to the non-waiver agreement and instead relied on keyword and manual review, the party could not then claim that the keywords were insufficient and that different keywords would have prevented the disclosure.[28] In contrast, O'Keefe was a case where the defendants claimed that the government failed to use proper keywords in its search of the corpus.[29] The defendants, however,

[26] Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, (D. Md. 2008), Link.
[27] United States v. O'Keefe, 537 F. Supp. 2d 14, (D.D.C. 2008), Link.
[28] Victor Stanley, 250 F.R.D. 262-263.
[29] O'Keefe, 537 F.Supp 2d 14.

had nothing more than their suspicions with which to back up the assertion and could not indicate specific keywords that would have been better in place of the ones the Government used.[30] This led the O'Keefe court to rule in favor of the Government, using language that has been often quoted in other cases including Victor Stanley that "for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread."[31]

When situations arise such as those that overwhelm the capabilities of manual collection or keyword searches, tools such as Predictive Coding, considered to be in their prime, often can do precisely the things that the traditional methods cannot.

3. Predictive Coding

Attempting to define current Predictive Coding under any one definition is an exercise in futility. Unfortunately, this is not because the technology as it exists is such a dynamic and powerful tool that it defies categorization. Predictive Coding cannot be properly defined because every company that is advertising or developing a product seems to view it as something different. Some companies are selling their Predictive Coding tool as a complete replacement for classic ESI search processes such as keyword or Boolean searches.[32] Other companies say that Predictive Coding is the platform that you enter all of the other search methods into to get more refined results.[33]

3.1 Non-Technical Explanation of the Predictive Coding Process

Using commonalities from each of the implementations studied, the best definition that can be offered is that the current Predictive Coding tools are a set of iterative steps that utilize

---

[30] Id. at 24.
[31] Id.
[32] Cost-Effective and Defensible Technology Assisted Review, Symantec Clearwell, 2, Link (last visited Oct 28, 2012).
[33] Jan Puzicha, Predictive Coding Explained, Recommind, 3 (Aug 30, 2012), Link.

human interaction to refine ever more accurate search criteria in order to pull relevant documents from a corpus. What that actually means is that generally each Predictive Coding program will ask for an initial source of relevant material. This can be a set of documents, a list of keywords, or some other starting point from which the software can base its search. This initial set of criteria is known as a seed list. The program takes the seed list and begins to search through the document collection looking for relevant results using a keyword search and its variants. Once the program has reached a certain threshold, typically based on the percentage of documents searched compared to the total number of documents, then the program will again ask for input from the operator. The operator will review the documents that the software has deemed responsive and will weigh them based on relevancy. For some programs it is a binary weight, either relevant or not relevant.[34] Other programs will ask the operator to weigh the relevancy on a scale from 1-100.[35] Once the human input is complete, the program takes this new information and begins to search again, having now adjusted the search criteria based on the responses from the human reviewer. This is where the idea of 'machine learning' comes into play that the program is learning what documents are and are not responsive. This process continues multiple times until another threshold is reached based on the overall number of documents searched as compared to the total number of documents in the collection. Once the multiple rounds of 'learning' have completed, the program should be able to produce all of the responsive relevant documents with upwards of 95% confidence.[36]

Certain implementations of machine learning are already in use quite successfully, the most referenced being that of e-mail spam filters. The filters are able to take the text from an email and process the words contained with it as they relate not just to each other but also to all

[34] Equivio Relevance, Relevance™ in Equivio Zoom, 3, Link (last visited Oct 28, 2012).
[35] Jan Puzicha, Finding Information: Intelligent Retrieval & Categorization, 8, (Sep 12, 2012), Link.
[36] Equivio Relevance, Relevance™ in Equivio Zoom, 5, Link (last visited Oct 28, 2012).

the other words in all the other emails that the filter has processed. As messages are marked as spam (think non-responsive), then the filter will use that information to figure out other similar messages in order to mark those as spam. If a message that was flagged as spam is marked as valid, the filter compares what was contained in the message to the contents of other messages that were left marked as spam and adjusts its filter accordingly.

3.2 Differentiation of Predictive Coding and Classic Keyword Searches

Much of the Predictive Coding process is encapsulated and hidden behind the veil of trade secrets. This means that most of the software on the market is considered to be 'black box,' whereby the inner workings are not known to the user.[37] Compared to the more straightforward approach provided by current search process such as keyword search, Predictive Coding is complex and enigmatic.

While it is true that Predictive Coding utilizes many of the aspects of Boolean keyword searches, the critical aspect of Predictive Coding is the iterative process under which Predictive Coding operates. This process, particularly when combined with other advances in alternative search and refinement technologies, gives Predictive Coding a distinct advantage over keyword searches.

For the sake of simplicity, assume that the Predictive Coding process starts with a seed list that contains a set of search terms. These search terms are exactly like those that would be entered into a Boolean keyword search. The Predictive Coding application performs the keyword search on a *subset* of the total documents and returns the set of responsive results from that subset. The human operator then reviews this small set of responsive documents and indicates whether or not the documents are relevant. If there are ten documents returned, and five are

[37] A Project of The Sedona Conference Working Group on Best Practices for et. al., The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery August 2007 Public Comment Version, 8 Sedona Conf. J. 189, 204, 2007, Link.

related to a car collision and five are related to a boating collision, the review might mark the five that are related to the boating incident as responsive. The Predictive Coding program uses that input in searching the next subset of documents to eliminate the documents that are related to a car crash as non-relevant or non-responsive which prevents the need for a human reviewer to sift through all of them. Effectively, the Predictive Coding program is appending the Boolean term '!car' (not car) to the search, without having to begin the search all over again. This process is then repeated using the next subset and the next, until the program is able to determine with a certain percentage of confidence that it would be able to find all the responsive documents in the entire collection.

While the boat and car illustration is an over simplified example, it is demonstrative of the adaptability of the Predictive Coding process. When dealing with a document corpus that is millions or tens of millions of documents large, the ability to continuously refine and re-evaluate the search at regular intervals provides a much more thorough and accurate result. Even more importantly, it provides the custodian with reasonable assurance that he or she located all of the responsive documents in the collection while having reviewed only a minute fraction.

3.3 Predictive Coding Shortcomings and Areas of Concern

As with any new and emerging technology, Predictive Coding is just beginning to see struggles that need to be overcome before it will be accepted as a widespread E-Discovery tool, particularly when the existing tools, such as keyword searches, have engendered such a confident and reliable reputation.

3.3.1. Lack of Unified Purpose and Definition

In looking at the differences and distinctions between keyword searches and Predictive Coding, one can't help but beg the question: is this really all that different from keyword searches

to be considered a revolutionary new product that could change the face of discovery? The short answer is no. The long answer is maybe.

While the theory behind Predictive Coding does in fact have the capability to radically alter the landscape of E-Discovery, in its current form Predictive Coding falls far short of the hype despite what the product brochures from various vendors would have you believing. Every company that has a Predictive Coding tool on the market says their implementation of Predictive Coding is the 'true' implementation. Unfortunately, as noted before, there is no one definition, even looking outside product brochures, that properly encapsulates what these companies are attempting to achieve.

Noted Forensic Examiner and E-Discovery Special Master, Craig Ball, recently raised a similar issue in his blog, Ball In Your Court. Professor Ball challenged the Predictive Coding vendors to form a cohesive group that can propose a single message that customers, educators, and most importantly the courts, can understand. [38]

Unless Predictive Coding vendors can push their software to a point where it is beyond the obvious comparisons to keyword search, Predictive Coding will likely never be anything more than yet one more 'enhancement' to the tried and true keyword search.

3.3.2. Lack of Understanding by Customers

As a byproduct of the individualized efforts put forth by the various Predictive Coding vendors, there is an increasing misunderstanding by those that the vendors need most, the customers. Customers, in beginning to research Predictive Coding, are at first amazed by the potential that this tool can provide for cutting costs and improving responsiveness to discovery requests. As more research is conducted, though, and product comparisons are introduced, the excitement is soon replaced by confusion.

[38] Craig Ball, Got TAR?, Ball In Your Court (September 4, 2012), Link.

Most of this is due to the specific implementations that vendors have released. Some have rushed to the scene, simply re-branding their existing keyword search products as Predictive Coding just to cash in on the latest and greatest.[39] Others do seem to actively be trying to make a product that utilizes 'machine learning,' but none have expressly proven to actually do so.

This is where the crucial misunderstanding comes into play. As alluded to earlier, from a strictly theoretical standpoint, Predictive Coding could forever alter the landscape of E-Discovery. This is because the theoretical implementations of Predictive Coding are designed around the inclusion of not just some of the alternatives to keyword searches, like Boolean operators, but all of them. Boolean operators work in conjunction with clustering and conceptual mapping to search a document not just for a word but for an *idea*. This is where the magic and mysticism surrounding Predictive Coding is coming from: that a machine could not only properly locate a series of words in a document but could understand the essence or purpose of that document and pair it with the intended goal of the search to determine whether the document is responsive. This is where the idea of machine learning took root in the Predictive Coding vernacular; over the course of multiple iterations and refinements, based on the responses of the human operator, the algorithm can successfully locate each and every document that has something to do with the *thing* the case is about. Unfortunately, this is nothing more than a pipe dream because, as stated above, the implementations most often seen are nothing more than a dressed up keyword search.

3.3.3. Black Box Approach

The implementations that currently exist for Predictive Coding seem to exist only to entice early-adopters away from their existing (or that company's previous E-Discovery) product. Using the guise that current Predictive Coding implementation is a fully developed 'machine

[39] Craig Ball, Gold Standard, Law Technology News (Online), April 1, 2012, Link.

learning' system, vendors attempt to lure potential clients with brochures so full of buzzwords that the meaning of what they are saying is anyone's guess.

For instance, Symantec Clearwell has developed what they are calling a 'transparent Predictive Coding' process that is fully defensible before the courts.[40] In their literature, Symantec lists the "lack of visibility into the predictive coding process" as one of the major challenges to Predictive Coding and that the Clearwell product is intended to help consumers to be able to properly defend the Predictive Coding process before any given court.[41] Immediately following these statements, however, under "How [Transparent Predictive Coding] Works," the document has the following: "Transparent Predictive Coding leverages intelligent training sets that are identified using sophisticated analytics, streamlining the selection of highly relevant training sets which are optimized for system training."[42] Yeah, I don't understand that either.

Obviously that is a very specific example. On a broader note, this idea of a 'magical' or 'black box' approach has attorneys concerned.[43] If an attorney cannot prove the methods by which they have complied with a discovery request, or they cannot defend the method adequately before the court, there is a chance of facing sanctions. If all an attorney can say in defense of their process is that they "leverage[d] intelligent training sets that [were] identified using sophisticated analytics,"[44] it is doubtful they will prevail. For the time being, it is enough to understand that what is given in the brochures simply is not going to cut it.

As evidenced by Clearwell's brochure, Predictive Coding companies are at least aware of the challenges their products face in gaining acceptance by the legal community. This, however,

---

[40] Cost-Effective and Defensible Technology Assisted Review, Symantec Clearwell, 2, Link (last visited Oct 28, 2012).
[41] Id. at 3-5.
[42] Id. at 5.
[43] A Project of The Sedona Conference Working Group on Best Practices for et. al., The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery August 2007 Public Comment Version, 8 Sedona Conf. J. 189, 204, 2007, Link.
[44] Id.

further underscores Professor Ball's call for a unification of the field so that these issues can be addressed by the trade as a whole.

3.3.4. Court's Acceptance of Predictive Coding

Tying in to the idea of defensibility, almost all would-be Predictive Coding supporters are grappling with the idea that this process has not truly been defended in front of a court. Predictive Coding is still so new that it is not as though someone has tried and *failed* to use Predictive Coding software, it is simply that it has never been challenged or endorsed by the judiciary.

The notion that Predictive Coding is a total unknown before the court changed earlier this year with cases that provide illumination into the future of Predictive Coding.

The first case, and the one that has thus far made the biggest waves, is the opinion from Da Silva Moore v. Publicis Groupe issued by Magistrate Judge Peck.[45] In addition to being the first known official endorsement of Predictive Coding, the Da Silva decision is also important because it is from a well known and well-respected Federal District and was authored by a technologically sound judge. Perhaps most importantly, however, Judge Peck does not simply endorse the use of Predictive Coding. Instead, he thoroughly explains not only why he thinks it is a good E-Discovery tool but also the reasoning behind why it should be permissible in a courtroom.

Judge Peck had already come out on the record as being in favor of judicial support for Predictive Coding. His article, Search Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, was published in Law Technology News and detailed Judge Peck's feeling on Predictive Coding. While it was important at the time to understand that Predictive Coding was at least being considered by the judiciary, Judge Peck

[45] Da Silva Moore v. Publicis Groupe, 2012 U.S. Dist. LEXIS 23350, (S.D.N.Y. Feb. 24, 2012), Link.

himself acknowledged that the article was little more than words on paper until such time as there was official judicial approval within the context of a case.[46] In Search Forward, Judge Peck outlined what he saw as the issues that were going to arise when (not if) Predictive Coding was challenged. Specifically, he concentrated on the overall admissibility of a search protocol, the fear of the 'black box' challenge, and whether or not Rule 702 of the Federal Rules of Evidence and Daubert[47] applied.[48]

With regards to overall admissibility of a search protocol, Judge Peck raised the point that despite there being quite a few cases by well respected judges that seriously questioned the validity of keyword searches, no one really seemed to think twice about using these methods as a search protocol.[49] Judge Peck further hypothesized that, perhaps fearing that no opinion is a dismissal of Predictive Coding, attorneys are hesitant to use Predictive Coding because they would not be able to break past the black box aspect where the parties and the court are unaware of the internal workings of the program. The article effectively stated that the internal workings are unimportant, so long as there is a sound, documented reason for each step of the process. The Federal Rules of Civil Procedure do not require that *every* responsive document be gathered. So long as there is enough proof to permit the court to evaluate the reasonableness of the decisions, the technical workings are immaterial.[50] Finally, addressing concerns that, like a black box evaluation, the entire process of Predictive Coding would be subject to a Daubert analysis to determine the procedure's admissibility, the article points out plainly that Daubert should not apply because the search protocol is not what is being admitted. Instead it is the documents that

[46] Andrew Peck, Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, Law Technology News (Online), October 1, 2011, Link.
[47] Daubert v. Merrell Dow Pharms., 509 U.S. 579 (U.S. 1993), Link.
[48] Andrew Peck, Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, Law Technology News (Online), October 1, 2011, Link.
[49] Id.
[50] Id.

the search protocol discovers that will be subject to the rules of admissibility.[51] In fact, other than the previously discussed O'Keefe [52] case, Judge Peck knows of no other case where a search protocol was subject to Rule 702 of the Federal Rules of Evidence and Daubert, and O'Keefe was related specifically to keyword searches, not Predictive Coding.[53]

Moving to the actual case discussion, the plaintiffs in Da Silva sued Publicis Groupe, asserting claims of employment and gender discrimination. The parties were referred to Judge Peck in an effort to resolve the remaining discovery issues, and after an initial meeting the parties had agreed to the use of Predictive Coding but were in disagreement over the methods that were to be used. Specifically, the plaintiffs contested that the defendants, after properly training the Predictive Coding program, wanted to produce only the top 40,000 responsive documents rather than evaluate the confidence levels and determine the appropriate number. The defendants chose this number strictly based on the financial impact of production rather relevancy. Plaintiffs further contended that the number of iterative rounds of 'learning' chosen by the defendants was insufficient. Judge Peck's ruling was to evaluate both the number of the results to be produced as well as the number of iterative rounds when necessary to determine whether the pre-determined number was reasonable under the conditions at the time.

Unsurprisingly, Judge Peck's opinion did not stray far from his Law Technology News article in establishing his reasoning, going so far as to quote much of the same language that was referenced above.[54] During his opinion, Judge Peck noted that Predictive Coding is not a one-size-fits-all solution.[55] Furthermore, just because the court was endorsing its use in no way meant

---

[51] Id.
[52] United States v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008), Link.
[53] Andrew Peck, Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, Law Technology News (Online), October 1, 2011, Link.
[54] Da Silva Moore v. Publicis Groupe, 2012 U.S. Dist. LEXIS 23350, 7-8 (S.D.N.Y. Feb. 24, 2012), Link.
[55] Id. at 27.

that a party should feel the need to use it every time discovery is requested nor were the parties compelled to use it.[56] In fact, Judge Peck had a rather easy time because the parties had already agreed to the use of Predictive Coding for the search but were only in dispute over aspects of the review.[57]

One of the more important aspects of the Da Silva opinion is when Judge Peck explores in *dicta* what exactly would be required of a party to prove that Predictive Coding would be allowable as a search protocol. He says that the crucial question to ask would be: if not Predictive Coding, what, if any, other search protocol would suffice?[58] Conventional manual review, also known as linear review, is completely out of the question when dealing with documents on a scale such as in Da Silva. Likewise, there are serious downsides to keyword searches, even with the use of Boolean operators, as had been discussed previously. Specifically, Judge Peck argues that attorneys using keyword searches is like a "child's game of Go Fish."[59] Plaintiff's attorneys often do not consult their clients regarding proper custodians or even basic keywords that could or should be used. Similarly, defense attorneys often do not meet with the custodians or IT professionals to attempt to assist the plaintiff's search.[60]

This leads to Judge Peck's two most crucial pieces of advice in ensuring that Predictive Coding be permitted into litigation: agree to the particular search protocol and be transparent throughout the entire process.[61] As one of the signatories of the aforementioned Sedona Conference's Cooperation Proclamation, Judge Peck is a firm believer in the use of bilateral exchanges to ensure that discovery disputes are resolved before becoming an issue.[62] Judge

---

[56] Id.
[57] Id.
[58] Id. at 28.
[59] Id. at 31. (Internal quotations omitted)
[60] Id. at 38.
[61] Id. at 35, 37.
[62] Id. at 36.

Peck's reasons that, regardless of his feeling about Predictive Coding, because the parties here had agreed to use Predictive Coding, he likely would have allowed it; much in the same way that without specifically endorsing them, judges all over the country nonetheless permit keyword searches in discovery.

A second aspect of the agreement to cooperate is transparency. One further reason that Judge Peck permitted the use of Predictive Coding is because the defendants in this case had provided thorough documentation of the seed set as well as which of the documents from the seed set were ultimately tagged as relevant or irrelevant and the full and complete list of what was being turned over to the requesting party.[63] Because the parties had worked at the onset of the litigation to develop the seed set based on emails that were located from keyword searches, it was obvious that the parties were at least in agreement about the use of Predictive Coding.[64]

In the end, Judge Peck provided his five reasons for deciding to permit Predictive Coding quite succinctly: 1) prior agreement, 2) quantity of ESI, 3) lack of alternative options, 4) superiority of Predictive Coding, and 5) the transparent processes established by the parties.[65] This reasoning coincided with Judge Peck's opinions from his Search Forward article previously discussed. Generally speaking, Judge Peck sees Predictive Coding as a superior technology not because it is perfect, but because it overcomes the obvious shortcomings of both manual collection and keyword searches.[66] Similarly, Judge Peck will likely be just as willing to accept the next viable technology that comes along and is able to fix the problems that are inherent in Predictive Coding.

---

[63] Id.
[64] Id. at 16-17.
[65] Id. at 36.
[66] Andrew Peck, Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, Law Technology News (Online), October 1, 2011, Link.

Following this opinion, the plaintiffs first appealed to District Judge Carter, who affirmed the opinion and order from Magistrate Judge Peck.[67] Plaintiffs additionally filed for Judge Peck to recuse himself,[68] citing multiple conflicts of interest including associating with the defendant's expert witness on numerous occasions.[69] Judge Peck denied the motion,[70] which was again appealed. District Judge Carter affirmed that there was no misconduct in Judge Peck's actions or opinions regarding Predictive Coding, securing Predictive Coding's use and Da Silva as valid precedent.[71]

Compared to Da Silva, the second notable case quite honestly has little impact. The main reason that it is worth noting is because it is authored by esteemed E-Discovery guru District Judge Shira Scheindlin (of Zubulake[72] fame). Her opinion in Nat'l Day Laborer Org. Network v. United States Immigration & Customs Enforcement Agency serves to add further legitimacy to the idea that Predictive Coding will, at the very least, not be turned down by a reasonable judge.[73] While much of the opinion is concerned with topics other than Predictive Coding, Judge Scheindlin did manage to slip in one specific paragraph that effectively mirrors Judge Peck's sentiments: Predictive Coding will be allowable so long as the parties agree to it just as they would any other search protocol like keyword searches.[74] While not expressly stated as such, Judge Scheindlin outlines the bases of The Sedona Conference's Cooperation Proclamation, emphasizing the need for communication and transparency with regards to the search

[67] Op. & Order, April 26, 2012, ECF No. 175, Link.
[68] Pl's Notice of Mot. for Recusal or Disqualification, April 13, 2012, ECF No. 169, Link.
[69] Pl's Mem. of Law in Supp. of Pl's Mot. for Recusal or Disqualification, 1-5 April 13, 2012, ECF No. 170, Link.
[70] Op. & Order, June 15, 2012, ECF No. 229, Link.
[71] Order, November 7, 2012, ECF No. 342, Link.
[72] *See* Zubulake v. UBS Warburg LLC, 217 F.R.D. 309 (S.D.N.Y. 2003), Link.
[73] Nat'l Day Laborer Org. Network v. United States Immigration & Customs Enforcement Agency, 2012 U.S. Dist. LEXIS 97863 (S.D.N.Y. July 13, 2012), Link.
[74] Id. at 57.

protocols.[75] Further, Judge Scheindlin specifies that the reasonableness of any request would be taken into account based on the cost and the reciprocal burdens placed on parties.[76]

While not nearly as impactful as Da Silva, Nat'l Day Laborer still is important for two reasons. First, as noted, Judge Scheindlin has a well-deserved reputation as an E-Discovery expert on the bench and provides a further air of legitimacy to the idea that Predictive Coding will be allowed. Second, the Southern District of New York is one of the most influential districts in terms of ESI, as it sees arguably the largest number of e-discovery related cases in the Federal Court system.[77]

Interestingly, the third and final case available for the approved use of Predictive Coding comes from Judge Chamblin of the Loudon County, Virginia, Circuit Court.[78] This is interesting because as commentators have noted, Loudon County is not a jurisdiction that is well known for its forward-thinking, precedent-setting decisions.[79] Odd jurisdiction aside, the defendants in Global Aerospace, Inc. v. Landow Aviation, L.P. produce a thorough and orderly argument for the use of Predictive Coding over manual linear review and keyword search.

The defendant, Landow Aviation, detailed that after culling over eight terabytes (TB) (8,000 gigabytes (GB)) of data, there remained approximately 250 GB of data that remained in the collection.[80] Landow estimated that manual linear review would take 20,000 man hours to review all approximately 2 million documents in the collection.[81] In addition to being prohibitively expensive and time consuming, the defendants pointed to various studies that

---

[75] Id. at 50.
[76] Id. at 57-58.
[77] WestlawNext Search, Link. (Jursidiction filter on the left indicates that the Southern District of New York has 848 cases with the next highest being the Northern District of California with 561 cases.).
[78] Order Approving the Use of Predictive Coding for Discovery, April 23, 2012, Link.
[79] Christopher Danzig, Virginia Judge Orders Predictive Coding, Despite Plaintiff Objections. Is This the Start of a New Era?, Above the Law, April 25, 2012, Link.
[80] Mem. in Supp. of Mot. for Protective Order Approving the Use of Predictive Coding, 3-4 April 9, 2012, Link.
[81] Id. at 4.

proved the accuracy of linear review is far below the level required to justify the exorbitant cost.[82] Landow additionally stated that keyword search would be faster than manual review but estimated that only 20% of the responsive documents would be located using that search methodology.[83] Finally Landow laid out its plans to use Predictive Coding. In addition to being 'orders of magnitude' less expensive, Predictive Coding would also be significantly more accurate, with almost 25% greater recall than manual linear review and upwards of 85% overall precision.[84] In light of Judge Peck and Judge Scheindlin's decision, perhaps the most critical argument that Landow made was their willingness for transparency and cooperation. Landow specifically detailed the information that would be provided to the opposing counsel and what remedies would be available in the event of a dispute.[85]

In opposition to the motion, Global Aerospace stated simply that there is no reason that the defendants should be able to skirt their obligation to inspect and review the documents for production.[86] Interestingly, Global Aerospace takes the position that Landow's proposition to produce 75% of the documents directly violates Landow's obligation to produce *all* the responsive documents.[87] Courts have repeatedly stated that discovery is not an obligation to produce *every* responsive document; perfection is not required under the Fed. R. of Civ. P.[88]

The Global Aerospace, Inc. decision is a full-fledged endorsement of Predictive Coding based on a brief containing solid arguments backed with facts and studies. The opposition brief attempted to address the weaknesses of Predictive Coding thoroughly, however, in the end, Judge Chamblin issued an order permitting the use of Predictive Coding. Unfortunately, because

---

[82] Id. at 6.
[83] Id. at 8.
[84] Id. at 11.
[85] Id. at 11-13.
[86] Mem. in Opp'n. to Pl's. Mot. for Protective Order Approving the Use of Predictive Coding, 1-2 April 16, 2012, Link.
[87] Id. at 2.
[88] Da Silva Moore v. Publicis Groupe, 2012 U.S. Dist. LEXIS 23350, 34 (S.D.N.Y. Feb. 24, 2012), Link.

of the nature of the state case, updates regarding appeals or other dispositions of the case are not available at this time.

One of the critical distinguishing points of the Global Aerospace, Inc. case is the idea that the plaintiffs are being compelled to accept the use of Predictive Coding. This is definitely a landmark shift because, as previously noted, the parties in Da Silva had already agreed to use Predictive Coding but were in contention with the specific implementation, and Judge Scheindlin in Nat'l Day Laborer merely gave permission if both parties agreed to its use.

Some are not convinced of the importance that is being placed on this case, at least from a far-reaching perspective. Craig Ball, the aforementioned E-Discovery Special Master, said that while it is a case to watch, there is really nothing remarkable about a judge permitting a party to use the search and collection method of their choice.[89]

Exploring Professor Ball's point for a minute, conventionally, so long as Landow produced proper results with adequate transparent documentation, Global Aerospace generally lacks grounds to question the method of collection.[90] This notion arose in Ford Motor Co. v. Edgewood Properties, Inc. when a party challenged the manual collection process.[91] Ford dealt with one party, Ford, who used manual collection methods while the other, Edgewood, used keyword searches to identify responsive ESI. Edgewood argued that the use of manual collection methods was inadequate for the collection of ESI and requested that a search specialist be brought in, paid for by Edgewood, to discover responsive documents from Ford.[92] The court in Ford found that a manual method was not improper in this situation, citing the Sedona Principle

[89] Evan Koblentz, Judge Orders Predictive Coding Over Plaintiff's Objections, Law Technology News (Online), April 24, 2012, Link.

[90]

[91] Mark S. Sidoti, et al, Challenging 'Manual' ESI Collections, Law Technology News (Online), April 9, 2010, Link.

[92] Id.

6, which states that absent prior agreement, the responding party is in the best position to determine the proper method for collection and production.[93]

Overall, thorough preparation of the briefs, coupled with the defensibility of a responding party being able to choose the method it desires, will likely make Judge Chamblin's ruling in Global Aerospace, Inc. an oft referenced decision by the champions of Predictive Coding, despite the commentary downplaying the importance of this decision and its conventionally non-binding precedential nature.

4. Questions and Analysis

4.1. Will Predictive Coding be Used for Future Document Collection?

Undoubtedly. As the quantity and diversity of ESI increases, there is going to be a greater need for more effective tools. E-Discovery specialists have seen and acknowledged the limitations that exist with keyword searches even when combined with other advanced search methodologies like Boolean operators and clustering.

The benefits of technologies such as Predictive Coding are well documented. As the science behind Predictive Coding matures, companies will be able to offer more truly machine learning based solutions. Even in its current state, often as little more than iterative keyword searches, Predictive Coding software provides stark improvements over the tried and true methods of manual review and traditional keyword searches.

Endorsements from such esteemed groups such as The Sedona Conference provide further credence to Predictive Coding's use. As judicial authority continues to be developed, the Cooperation Proclamation principles are going to play an increasingly important role in Predictive Coding's acceptance. If parties can work together and agree to use Predictive Coding, there is likely not a judge on the bench that would frown upon its use. Further acceptance will be

[93] Id.

seen so long as the processes are kept open and transparent with adequate quality control and dispute resolution between the parties and the court.

Where Predictive Coding might see a bumpy road is in the lack of quality control in the term itself. Without a cohesive idea of what Predictive Coding is, the chances of seeing widespread acceptance either from the courts or consumers is highly unlikely. If the industry behind Predictive Coding can band together, as Professor Ball suggests, and provide clarifications to the court and customers, the level of interest and, more importantly, trust in Predictive Coding is going to continue to rise.

4.2 Should a Firm Begin to Use Predictive Coding?

There's a familiar saying that it is better to ask for forgiveness than beg for permission. This seems to be the court's approach to the permissiveness of Predictive Coding, at least from the three examples that were outlined above. What is meant by this is that if a party wants to use a particular search method, such as Predictive Coding, then simply use it. The courts, assuming that the methodology is reasonable and proportional, will generally allow it, particularly if there is an agreement amongst the parties.

Even absent an agreement, a court is likely to endorse a particular method if it is chosen by the responding party based on the Sedona Principles.[94] Obviously this is not a clear cut rule because as Judge Peck noted, when there is no clear burden either via time or cost, a judge won't always agree to an unnecessary ESI search protocol just because a party requests it. Only when the search protocol can help "secure the just, speedy, and inexpensive determination of cases in our e-discovery world" will a judge permit the use of an unconventional or challenged search

[94] The (2004) Sedona Principles: Best Practices, Recommendations & Principles for Addressing Electronic Document Production the Sedona Conference Working Group on Electronic Document Retention and Production Sedona, Az, 5 Sedona Conf. J. 151, 162 (2004), Link.

methodology.[95] In such cases where there are millions of documents and hundreds if not thousands of gigabytes worth of data, as is becoming the norm, there is no doubt that the manual linear review simply is untenable and other search methods would be permitted.

5. Conclusion

We live in a digital world. Today there are very few homes without computers, and it is the cornerstone of a successful business to have electronic personnel records, inventory, payroll, etc. The idea of performing a discovery in modern litigation without running into some sort of ESI is almost nonexistent. At the same time, the pace of technology has dictated that more and more information can be stored smaller and cheaper than ever before in history. Storage devices with a capacity that cost thousands of dollars a mere ten years ago are now available in the $0.99 bins next to the checkout station at convenience stores. Parkinson's Law dictates that "data expands to fill the space available for storage"[96] and businesses and individuals are no exception. What this means for the modern litigator is that not only is a case more than likely to involve some sort of ESI, but typically the volume that must be handled for any given custodian is unfathomable.

Manual collection was becoming prohibitively expensive even before the exponential growth of ESI. Traditional automated systems are still viable solutions for most situations, however there is no doubt that the ESI burdens of the near future require a new solution to be developed to reduce time and costs, ensuring that the courts are still accessible to the average person. At the current rate of growth of data capacity and volume, litigation is rapidly becoming barred from any but those with the deepest pockets.

---

[95] Andrew Peck, Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, Law Technology News (Online), October 1, 2011, Link (internal quotation and citation omitted).
[96] Parkinson's Law, Wikipedia, Link (last visited Nov 20, 2012).

Thankfully, advances in data processing have also followed the advances that Computer Science has made in recent years with relation to storage capacity. This includes not only the algorithms and software languages to better handle huge data volumes but also the physical hardware advances that provide a platform from which to run these algorithms rapidly and efficiently. As can be imagined, one such advancement is the idea of Predictive Coding.

Predictive Coding, as an implementation of machine learning, truly does have the capacity to rapidly and accurately process absolutely astronomical amounts of data. The idea that an attorney can utilize small, iterative data sets to train the process to recognize responsive and non-responsive documents has the ability to bring even complex litigation discovery costs back to a reasonable level. The ability for a software platform to efficiently and, most importantly, accurately filter millions of documents through a machine learning process allows for attorneys to spend less and less time culling the corpus for relevant documents. This in turn allows attorneys to manually review fewer documents and produce the responsive documents with a greater confidence than ever before. As noted in the briefs in Global Aerospace, Inc., Predictive Coding is faster, cheaper, and more accurate than either manual collection or keyword searches.[97]

The Federal Rules of Civil Procedure are effectively silent on the manner in which documents are collected for discovery. Likewise, the Sedona Conference materials refrain from specifying a particular method or process that should be used. Instead, both sets of materials encourage the open communication between parties in the hopes that they will reach an agreement concerning the methods of discovery. When that occurs, the Rules, the Conference, and the courts are extremely unlikely to say that the parties reached that agreement in error. As

[97] Mem. in Supp. of Mot. for Protective Order Approving the Use of Predictive Coding, 3-4 April 9, 2012, Link.

noted in Da Silva [98] and Nat'l Day Laborer Org.,[99] in-depth judicial opinions have further vetted the budding technology, and as Judge Peck in Da Silva stated that while it may not be appropriate in every case, where the parties agree and the technology is reasonable, there should be little interference from the judges.[100]

Even when the parties are in dispute, the Sedona Conference concludes that the responding party is generally in the best position to determine the proper method for response.[101] Courts, like those in Global Aerospace, Inc. [102] and Ford Motor Co. v. Edgewood Props. [103] agreed, holding in each case that the use of Predictive Coding and manual collection, respectively, was proper under the circumstances, despite the objections of the opposing counsel.

The precedent has been set, the technology is ripe for use, and it is time to move forward. Given the rules, the expectation of cooperation, and the overall reasonableness standard to which parties are held, the fact that there are attorneys and law firms that are fighting this shift in technology seems entirely unreasonable. But for every generation, there are still those that will prefer the abacus to the calculator, the telegraph to the text message, and the permanent eye strain of manual review to that of a promisingly effective and accurate automated process.

---

[98] Da Silva Moore v. Publicis Groupe, 2012 U.S. Dist. LEXIS 23350 (S.D.N.Y. Feb. 24, 2012), Link.
[99] Nat'l Day Laborer Org. Network v. United States Immigration & Customs Enforcement Agency, 2012 U.S. Dist. LEXIS 97863 (S.D.N.Y. July 13, 2012), Link.
[100] Da Silva Moore v. Publicis Groupe, 2012 U.S. Dist. LEXIS 23350, 26-27 (S.D.N.Y. Feb. 24, 2012), Link.
[101] The (2004) Sedona Principles: Best Practices, Recommendations & Principles for Addressing Electronic Document Production the Sedona Conference Working Group on Electronic Document Retention and Production Sedona, Az, 5 Sedona Conf. J. 151, 162 (2004), Link.
[102] Order Approving the Use of Predictive Coding for Discovery, April 23, 2012, Link.
[103] Ford Motor Co. v. Edgewood Props., 257 F.R.D. 418, 427 (D.N.J. 2009), Link.