Vanderbilt Journal of Entertainment & Technology Law

Volume 15
Issue 2 *Issue 2 - Winter 2013*

Article 3

2013

Man Versus Machine Review: The Showdown between Hordes of Discovery Lawyers and a Computer-Utilizing Predictive-Coding Technology

Nicholas Barry

Man Versus Machine Review: The Showdown between Hordes of Discovery Lawyers and a Computer-Utilizing Predictive-Coding Technology

ABSTRACT

The discovery process is regularly capturing millions of pages of documents. Electronic storage is making storing documents cheaper and easier. When litigation begins, however, sorting through this massive amount of electronically stored information is costly and time intensive. Keyword searches are a start to managing the growing amount of electronic documents, but the discovery process is still falling behind in efficiency. Predictive coding could change all that.

Predictive coding is capable of solving the time-intensive nature (and resultant growing cost) of processing discovery documents. Predictive coding is faster, cheaper, and more accurate than traditional linear document review, the current "gold standard" of document review. It requires a senior attorney to code a small amount of the overall document pool, then the predictive-coding technology kicks-in and codes the rest of the documents based on the senior attorney's coding decisions. But the Federal Rules of Civil Procedure create obstacles for this new technology's adoption.

This Note examines the path other discovery technologies have taken before courts and practitioners have ultimately accepted them. It is those paths that offer insight into the path predictive coding should take to become accepted. Eventually, something will need to be done to reign in the increasing cost of discovery. This Note argues predictive coding is the answer and provides a pathway for its acceptance under the Federal Rules of Civil Procedure.

TABLE OF CONTENTS

I. BACKGROUND: E-DISCOVERY, KEYWORD-SEARCH TERMS, AND PREDICTIVE CODING 345
A. The "Reasonableness" Inquiry 346

B. What Is "E-Discovery" and How Has It Developed? 346
C. How Have the Courts Adapted E-Discovery to Keyword-Search Terms? 351
D. What Is Predictive Coding? 354
II. ACCEPTANCE OF E-DISCOVERY AND KEYWORD-SEARCH TERMS: WILL PREDICTIVE CODING TAKE A SIMILAR PATH? 357
A. The Development of E-Discovery 357
B. The Development of Keyword-Search Terms—A Possible Path for Predictive Coding 360
III. WHY COURTS SHOULD ADOPT PREDICTIVE CODING AND HOW IT CAN BE IMPLEMENTED 364
IV. CONCLUSION 372

Billions of dollars are at stake in a battle of man versus machine.[1] Typically, people think robotic efficiency destroys manufacturing jobs but not cerebral jobs, like a lawyer's. But with the explosion of electronic information and the massive amount of documents involved in a single case's discovery,[2] it is high time for a new technology to make managing this gargantuan amount of information feasible.

In comes our hero (or nemesis?): predictive coding. Predictive coding describes a computer program that predicts the relevance of discovery documents based on the prior coding of a small sample of discovery documents by an attorney.[3] Attorneys will no longer need to read thousands of pages of documents. Instead, after a senior attorney's initial review of a small document sample, the computer can then search through, categorize, and organize the rest of the documents for the attorney, rendering e-discovery more efficient.[4] It is a useful tool for an attorney and an aid in the litigation process, but it also may begin reducing the need for reviewing attorneys.[5] This transition could commence in the near future.

---

1. Joshua Bullough, *eDiscovery, Litigation, and Utah's Retention Schedules*, RECORDS KEEPERS—UTAH STATE ARCHIVES (Oct. 20, 2011), http://recordskeepers.wordpress.com/2011/10/20/ediscovery-litigation-and-utahs-retention-schedules.

2. George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, ¶ 23 (2007) (discussing the explosion of ESI and the difficulty of producing material, noting that some litigation has involved over one billion relevant electronic documents).

3. *See* Barry Murphy, *Is Predictive Coding the Future of Document Review?*, E-DISCOVERY J. (Oct. 28, 2010, 11:56 PM), http://ediscoveryjournal.com/2010/10/is-predictive-coding-the-future-of-document-review.

4. *Id.*

5. John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N.Y. TIMES, Mar. 4, 2011, http://www.nytimes.com/2011/03/05/science/05legal.html.

This transition will occur, however, only if courts decide that predictive coding is "reasonable" according to Federal Rule of Civil Procedure (FRCP) 26(b).[6] If courts find that predictive coding does not satisfy the reasonableness requirement, they will effectively relegate the technology to a substantially lesser role in litigation. But, since courts have accepted e-discovery and keyword-search terms as satisfying the "reasonableness" requirement, it is likely that similar acceptance of predictive coding is on the horizon.

This Note examines the relevant cases and FRCP addressing e-discovery and recommends that courts adopt predictive coding as reasonable under the FRCP. Part I outlines the development of e-discovery and its evolution, including the courts' acceptance of keyword-search terms. Part II analogizes courts' approaches to e-discovery and keyword-search terms to possible acceptance of predictive coding. Finally, Part III recommends that courts adopt predictive-coding methods that are more accurate than human linear document review.

I. BACKGROUND: E-DISCOVERY, KEYWORD-SEARCH TERMS, AND PREDICTIVE CODING

E-discovery significantly increased in the 1970s. Keyword-search terms (like those used in Google, LexisNexis, or Westlaw) developed later (by the late 1970s).[7] Finally, predictive coding, which uses algorithms to predict which electronic files are "responsive" and "non-responsive" to discovery requests and subpoenas, was developed only recently and is currently coming into mainstream use.[8] Each of these technological developments help lawyers to search more efficiently and accurately for important documents; however, they each also raise legal concerns.

---

6. FED. R. CIV. P. 26(b)(2)(B).

7. Edward F. Sherman & Stephen O. Kinnard, *The Development, Discovery, and Use of Computer Support Systems in Achieving Efficiency in Litigation*, 79 COLUM. L. REV. 267, 268 (1979).

8. Marilyn Odendahl, *Attorneys Discover Predictive Coding*, THE INDIANA LAW. (Oct. 10, 2012), http://www.theindianalawyer.com/attorneys-discover-predictive-coding/PARAMS/article/29842; Katey Wood, *Predictive Coding from Theory to Practice: kCura's Relativity Assisted Review*, ENTERPRISE STRATEGY GROUP 1, 2 (Oct. 2012), http://kcura.com/relativity/Portals/0/Documents/Predictive%20Coding%20from%20Theory%20to%20Practice%20-%20kCura's%20Relativity%20Assisted%20Review.pdf.

A. The "Reasonableness" Inquiry

The FRCP requires courts to balance cost and completeness when resolving discovery disputes.[9] Rule 26(b)(2)(C)(iii) requires the court, either "[o]n motion or on its own," to limit discovery if it finds that "the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issue at stake in the action, and the importance of the discovery in resolving the issues."[10] In addition, Rule 37(a)(4) requires that "an evasive or incomplete disclosure, answer or response must be treated as a failure to disclose, answer, or respond," which necessarily means discovery responses must be "complete."[11] Finally, attorneys must certify, to the best of their knowledge "formed after a reasonable inquiry,"[12] that "a discovery request, response, or objection"[13] is consistent with the FRCP, is not made for an improper purpose, and is "neither unreasonable nor unduly burdensome or expensive . . . ."[14] Thus, the court must balance these competing interests—completeness and cost—when deciding discovery disputes.[15] If a party can show the newer document-review processes are more accurate, more efficient, and more responsive than manual review, then these processes should also meet the standards set out by the FRCP.

B. What Is "E-Discovery" and How Has It Developed?

E-discovery stands for electronic discovery, the process of obtaining electronically stored information (ESI) from other parties involved in a lawsuit.[16] An amendment to the FRCP in 2006 explicitly made all electronic files discoverable.[17] The amended rules did not make ESI discoverable for the first time; courts had long held that electronic files were discoverable even without a specific grant in the

---

9. Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 11, ¶ 5 (2011).

10. FED. R. CIV. P. 26(b)(2)(c)(iii).

11. *Id.*

12. *Id.* at 26(g)(1).

13. *Id.* at 26(g)(1)(B).

14. *Id.* at 26(g)(1)(B)(iii).

15. Grossman & Cormack, *supra* note 9.

16. Working Grp. on Elec. Document Retention & Prod., Sedona Conference, *The (2004) Sedona Principles: Best Practices, Recommendations & Principles for Addressing Electronic Document Production*, 5 SEDONA CONF. J. 151, 151 (2004).

17. *See* FED. R. CIV. P. 26 advisory committee's note (2006).

rules.[18] Additionally, the 1970 amendment to FRCP 34 clarified that certain types of computer-stored information were discoverable.[19] Since that amendment, e-discovery has grown exponentially and now includes, *inter alia*, emails, word-processing files, spreadsheets, databases, video files, MP3 files, and virtually every other file now stored on computers and other electronic devices (such as PDAs, cell phones, flash drives, DVDs, etc.).[20]

The Sedona Principles[21] present several major differences between regular discovery and e-discovery that have had a strong impact on the courts and discovery rules and that courts see as persuasive authority on e-discovery issues.[22] Three are particularly relevant to predictive coding[23]: (1) volume and duplicability, (2) persistence, and (3) dispersion and searchability.[24] Volume and duplicability directly relate to the size of discovery.[25] As electronic files are more readily used, the discoverable information grows, and so does the cost of reviewing it for responsive documents. Predictive coding ameliorates this problem, for it is able to "weed out" duplicative files and reduce the volume of discovery to a manageable level.[26] Persistence refers to the way electronic documents cannot be

18. Bills v. Kennecott Corp., 108 F.R.D. 459, 461 (D. Utah 1985) ("It is now axiomatic that electronically stored information is discoverable under Rule 34 of the Federal Rules of Civil Procedure if it otherwise meets the relevancy standard prescribed by the rules . . . .").

19. *Id.*

20. WORKING GRP. ON ELEC. DOCUMENT RETENTION & PROD., SEDONA CONFERENCE, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION 1 (2d ed. 2007), *available at* https://thesedonaconference.org/download-pub/81.

21. The Sedona Principles were developed by a working group of the Sedona Conference. The Sedona Conference is a "non-partisan research and education institute dedicated to the advancement of law and policy . . . ." *Frequently Asked Questions*, THE SEDONA CONFERENCE, https://thesedonaconference.org/faq (last visited Sept. 4, 2012). Its focus is on antitrust, intellectual property and complex litigation. *Id.* A cross section of all three areas is electronic discovery. The Sedona Conference creates working groups of experts to tackle complex legal issues. *Id.* The working group then publishes best practices or guidelines. *See id.* The Sedona Conference has been called an "ALI-ABA on steroids." *Id.* (internal quotation marks omitted). But the Sedona Conference focuses on specific "crisis areas" or "bottlenecks" in the development of law rather than focusing on restatements or analyses of entire areas of law. *Id.* (internal quotation marks omitted).

22. *See* Victor Stanley, Inc. v. Creative Pipe, Inc. (*Victor Stanley I*), 250 F.R.D. 251, 262 (D. Md. 2008); Consol. Aluminum Corp. v. Alcoa, Inc., 244 F.R.D. 335, 345 n.18 (M.D. La. 2006); Treppel v. Biovail Corp., 233 F.R.D. 363, 374 (S.D.N.Y. 2006).

23. WORKING GRP. ON ELEC. DOCUMENT RETENTION & PROD., *supra* note 20, at 2-5.

24. *Id.* at 2 (looking at other issues such as scope, preservation obligations, how ESI should be preserved, how ESI should be produced, who should pay for production, and other e-discovery issues).

25. *Id.*

26. *See infra* Part III (discussing the reasons to use predictive coding).

destroyed easily, unlike paper documents.[27] Certain files may persist even after being "deleted" from the computer.[28] Predictive coding can detect these "deleted," but not destroyed, files for preservation or production.[29] Lastly, ESI is more difficult to search than paper because electronic files are more dispersed than paper files, which are normally located in one place.[30] Predictive coding assists by searching across multiple dispersed electronic components to quickly find relevant documents.[31] Thus, predictive coding helps to mitigate the new challenges presented by e-discovery.

Although the 1970 amendment to FRCP 34 formally allowed for e-discovery of certain types of electronic files, courts were reluctant to give opposing counsel access to the computer systems on which these files were stored.[32] As developments in technology made electronic files more common, lawyers began to use "computer-support systems"—computer systems created to support data storage—to collect discovery data.[33] But these computer-support systems were not always created with litigation in mind; some were simply created to store information.[34] Because these systems could contain privileged information and attorney work product, many courts required opposing counsel to show "good cause" before granting access to them.[35]

But some courts allowed limited discovery of systems. For example, some courts required responding parties to print out or transform the data into a useable electronic format.[36] Others allowed attorneys to use search terms to search the computer system.[37] Courts recognized that access to the computer system itself was necessary when parties lacked sufficient information to make specific discovery requests.[38]

---

27. WORKING GRP. ON ELEC. DOCUMENT RETENTION & PROD., *supra* note 20, at 3.

28. *Id.*

29. *Id.*

30. *Id.* at 5.

31. *Id.*

32. Sherman & Kinnard, *supra* note 7.

33. FED. R. CIV. P. 34 advisory committee's note (1970) ("The inclusive description of 'documents' is revised to accord with changing technology. It makes clear that Rule 34 applies to electronics data compilations from which information can be obtained . . . ."); Sherman & Kinnard, *supra* note 7.

34. Sherman & Kinnard, *supra* note 7, at 269-70.

35. *Id.* at 271-72, 291-94.

36. Adams v. Dan River Mills, Inc., 54 F.R.D. 220, 222 (W.D. Va. 1972) (requiring the responding party to allow access to computer cards and tapes and to print out the requested information).

37. Sherman & Kinnard, *supra* note 7, at 275.

38. *Id.*

In 2003, the US District Court for the Southern District of New York, in *Zubulake v. UBS Warburg LLC* (*Zubulake I*), determined the scope of electronic discovery.[39] Specifically, the court addressed whether it should force the responding party to pay to recover emails from backup tapes or if cost-shifting to the requesting party should take place instead.[40] The Supreme Court had previously held that the "presumption is that the responding party must bear the expense of complying with discovery requests" unless protection from "undue burden or expense" was granted under Rule 26(c).[41] Here, the district court held that whether electronic data is accessible or inaccessible turns largely upon the type of media on which it is stored, and accessibility determines whether cost-shifting is warranted.[42] To begin its accessibility analysis, the court identified five different types of data storage, ordered from most accessible to least accessible: (1) active, online data; (2) near-line data; (3) offline storage-archive data; (4) backup tapes; and (5) erased, fragmented, or damaged data.[43] The first three storage types are discoverable without cost-shifting,[44] while the last two types of data storage require a cost-shifting analysis because the data is not reasonably accessible, and accessing it may cause an "undue burden or expense."[45] The court then set out a seven-factor test to determine whether accessing the media is an undue burden or expense.[46] This is the analysis courts have typically applied when deciding accessibility issues.[47]

But in 2006, an amendment to the FRCP introduced a new standard for determining the scope of ESI discovery.[48] The amendment stated that a party did not have to provide ESI that was not "reasonably accessible" due to "undue burden or cost" unless the requesting party could show "good cause" for the discovery.[49] This is similar to the standards in *Zubulake I*, but some scholars have suggested that the newly amended rules should change the analysis for inaccessible data because it is not just a matter of cost shifting, but

39. Zubulake v. UBS Warburg LLC (*Zubulake I*), 217 F.R.D. 309 (S.D.N.Y. 2003).
40. *Id.* at 311-12.
41. Oppenheimer Fund, Inc. v. Sanders, 437 U.S. 340, 358 (1978).
42. *Zubulake I*, 217 F.R.D. at 318.
43. *Id.* at 318-19.
44. *Id.* at 319-20.
45. *Id.* at 316.
46. *Id.* at 322-23.
47. *See* Debra Lyn Bassett, *Reasonableness in E-Discovery*, 32 CAMPBELL L. REV. 435, 443-44 (2010).
48. *See* FED. R. CIV. P. 26(b)(2)(B).
49. *Id.*

also one of "reasonable accessibility."[50] In effect, this standard would reduce the scope of discovery because, under a *Zubulake I* framework, the inaccessible data is discoverable if cost shifting occurs, but under the new rules it is only discoverable if a party shows "good cause."[51] But the Advisory Committee Notes suggest that electronic storage systems can make data easier to access, and courts should take this into account when determining whether discovery of ESI poses an undue burden or expense.[52] Predictive coding makes it easier for attorneys to obtain and search electronic data, which reduces burden and cost. In other words, predictive coding can impact the result of the undue-burden-or-expense analysis for otherwise inaccessible data, because it makes ESI's discovery less expensive and could thereby broaden the scope of permissible discovery.

Like the 2006 amendment, Rule 26(b)(2)(C) also promotes acceptance of predictive coding. According to the rule, courts may limit discovery if "the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive."[53] In addition, courts must limit discovery if "the burden or expense of proposed discovery outweighs its likely benefit."[54] Predictive coding may be "more convenient, less burdensome, or less expensive."[55] It provides more benefits at a lower expense.[56]

E-discovery has come a long way since the 1970s, where courts were reluctant to find opposing counsels' computer systems discoverable. In 2003, the court in *Zubulake I* held that computer systems that were inaccessible could be discoverable if cost shifting occurred.[57] In addition, it undertook an in-depth analysis of various accessibility issues, along with how courts could apply "undue burden or expense."[58] The trend of presumptions deciding who bears the expense of discovery that occurred prior to *Zubulake I* began to shift toward an accessibility analysis after the case, showing that courts' understanding of electronic media and e-discovery had become more sophisticated; courts are now applying a more detailed analysis. A

---

50. Bassett, *supra* note 47, at 441 (arguing that courts need to update their analysis based on the newly amended FRCP).

51. FED. R. CIV. P. 26(b)(2)(B).

52. *Id.* at 26(b)(2) advisory committee's note (2006) (finding that Rule 26(b)(2)(C) "balance[s] the costs and potential benefits of discovery").

53. *Id.* at 26(b)(2)(C)(i).

54. *Id.* at 26(b)(2)(C)(iii).

55. *Id.* at 26(b)(2)(C)(i).

56. *See* Grossman & Cormack, *supra* note 9.

57. *Zubulake I*, 217 F.R.D. 309, 318-20 (S.D.N.Y. 2003).

58. *Id.*

court is now more likely to weigh the cost-savings and efficiencies offered by predictive coding when analyzing the application of the discovery rules because the reliance on presumptions is less likely.

C. How Have the Courts Adapted E-Discovery to Keyword-Search Terms?

Keyword-search terms for discovery purposes first appeared in the context of computer-information storage.[59] But while early computer systems were searchable, limitations existed based on the type of organization system the computer system employed.[60] Attorneys initially used keyword-search terms in early antitrust cases where the volume of information was extremely large.[61] But in those cases, parties would build their own databases so they could search and find their own relevant documents.[62] They did not use keyword-search terms to search the opposing party's computer systems.[63] In fact, although courts sometimes required a responding party to find relevant requested documents instead of allowing the party to respond by document dumping on the requesting party, they never permitted a requestor to search the responder's computers.[64] They did not even require a responding party to search its own computer database using keyword-search terms created by the requesting party, for they feared disclosure of privileged information.[65] Thus, early use of search terms was limited to a party's private search of its own database.[66] But keyword searches have substantially developed since.

First, parties now use keyword searches to preserve documents. For example, a party may have an obligation to run keyword searches on applicable data and preserve all documents the

---

59. *See* Sherman & Kinnard, *supra* note 7, at 267-70. Keyword-search terms are employed when using Google, Westlaw, or Lexis search functions. *See, e.g.*, Chris Wilde, *Google Keyword Tool*, GOOGLE KEYWORD TOOL BOX, http://www.googlekeywordtool.com (last visited Nov. 1, 2012).

60. Sherman & Kinnard, *supra* note 7, at 269. There were two main ways to develop these systems, "full text" and "indexing." *Id.* A full-text system included every word within its index, while the index method screened all the documents first and then coded in certain keywords that could be searched later. *Id.* at 269-70.

61. *Id.* at 268 n.6.

62. *Id.* at 267-68.

63. *See id.*

64. Budget Rent-A-Car of Mo., Inc. v. Hertz Corp., 55 F.R.D. 354, 357 (W.D. Mo. 1972) (finding that a party who is familiar with its own records has an obligation to find those requested documents rather than simply providing a mass of records for the requesting party to search through).

65. Sherman & Kinnard, *supra* note 7, at 271-77.

66. *Id.*

keyword searches return.[67] This does not require a litigant to review all those documents manually but simply to preserve those documents for possible future discovery requests.[68] In addition, if litigation is reasonably anticipated, giving rise to a preservation obligation, any entity could use a system-wide keyword search and retain all "hits" to meet their obligation.[69]

Second, by the 1990s courts permitted parties to supply the keywords to be used in keyword searches on their adversaries' databases to find documents that could be relevant to the litigation.[70] But courts also provide limiting guidelines. For example, courts suggest parties discuss keyword searches early in the litigation because, after substantial discovery has taken place, keyword searches may no longer be economical or possible.[71] Additionally, opposing parties should meet to discuss possible search terms, should communicate about which terms were effective and which were not, and should develop new terms as the case moves forward.[72]

Members of the Sedona Conference[73] have set forth "best practices" for developing keyword-search terms and interacting with opposing parties.[74] The best practices suggest that all parties cooperate with each other, which can help satisfy the goal of FRCP 1: a "just, speedy[,] and inexpensive determination of every action."[75] This includes exchanging information about which data sources opposing counsel needs to search, as well as aiding opposing counsel in crafting keyword searches.[76] But beyond the challenge of working with opposing counsel to develop keyword-search protocols, keyword searches themselves have limitations.[77]

---

67. Zubulake v. UBS Warburg LLC (*Zubulake II*), 229 F.R.D. 422, 432 (S.D.N.Y. 2004).

68. *Id.*

69. *Id.* at 431-32.

70. Procter & Gamble Co. v. Haugen, 179 F.R.D. 622, 632 (D. Utah 1998), *aff'd in part, rev'd on other grounds sub nom.* Proctor & Gamble Co. v. Haugen, 222 F.3d 1262 (10th Cir. 2000).

71. *In re* Prudential Ins. Co. Am. Sales Practice Litig. Agent Actions, 278 F.3d 175, 186 (3d Cir. 2002) (denying keyword searches to party after counsel failed to review the discovery that was already produced).

72. Ameriwood Indus., Inc. v. Liberman, No. 4:06CV524-DJS, 2007 WL 685623, at *1 (E.D. Mo. Feb. 23, 2007).

73. *See supra* note 21.

74. Jason R. Baron & Edward C. Wolfe, *A Nutshell on Negotiating E-Discovery Search Protocols*, 11 SEDONA CONF. J. 229, 229 (2010).

75. *Id.* at 230 (internal quotation marks omitted).

76. *Id.* at 231 (noting also that failure to help opposing council create keyword searches could be construed as an attempt to conceal relevant evidence because information about data is typically asymmetrical).

77. *Id.* (looking at search terms as potentially being both over- and under-inclusive because of the ambiguities of human language).

Courts have begun to recognize the difficulties and complexities that accompany keyword searches.[78] The keyword-search methodology necessarily involves computer science, statistics, and linguistics.[79] Because of these complexities, some courts have suggested that the methodology of developing keyword searches requires the use of an expert.[80] This would require an expert, in accordance with Federal Rule of Evidence (FRE) 702, to create the keyword search so that it would be defensible if opposing counsel challenged its sufficiency.[81] This complexity has caused at least one court to call for quality control and testing in managing keyword-search terms.[82]

Courts have also begun to recognize other ways to search through ESI beyond keyword searching.[83] Although court opinions mentioning other search methodologies are limited, at least two opinions seem to endorse the idea that parties can use other search methodologies.[84] These search methodologies can be incorporated into the predictive-coding process to further enhance results.[85] The court in *Victor Stanley, Inc. v. Creative Pipe, Inc.* (*Victor Stanley I*) discussed two in particular: clustering, where similar documents are placed together, and categorization, where a search captures documents that express the same thoughts in alternate ways.[86] Both of these processes are similar to predictive coding.

---

78. *Victor Stanley I*, 250 F.R.D. 251, 260 (D. Md. 2008).

79. *Id.* ("Given this complexity, 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.").

80. *E.g.*, United States v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008) (finding that a challenge to the keyword-search terms would require compliance with Rule 702 of the Federal Rules of Evidence).

81. *Id.*

82. William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 134 (S.D.N.Y. 2009) (serving as a "wake-up call to the Bar").

83. *See Victor Stanley I*, 250 F.R.D. 251, 259 n.9 (D. Md. 2008) (suggesting several other search methodologies that could help manage large amounts of discovery); Disability Rights Council of Greater Wash. v. Wash. Metro. Transit Auth., 242 F.R.D. 139, 148 (D.D.C. 2007) (noting that concept searching, rather than keyword searching, may be more efficient and gives more comprehensive results).

84. *See Victor Stanley I*, 250 F.R.D. at 259 n.9; *Disability Rights Council*, 242 F.R.D. at 148.

85. Caitlin Murphy, *5 Things You Should Know About Predictive Coding*, E-DISCOVERY INSIGHT (Jan. 25, 2011), http://ediscoveryinsight.com/2011/01/5-things-you-should-know-about-predictive-coding.

86. *Victor Stanley I*, 250 F.R.D. at 259 n.9.

D. What Is Predictive Coding?

Predictive coding is far more advanced than a simple keyword search. The process may include, but is not the same as, clustering, categorizing, culling, or threading.[87] Clustering and categorizing are processes that combine similar information into one data pile, but the document reviewer must still go through the documents page by page.[88] Culling removes documents from a set, which predictive coding does not do.[89] Finally, threading presents email conversations in one thread or as a conversation rather than as individual emails, which reduces duplicate emails.[90]

The predictive-coding process follows these subsequent steps: First, some other technology organizes the data, like concept searching, keyword searching, clustering or categorizing.[91] Second, a senior attorney receives this initial sample, manually reviews the documents, and begins coding them as responsive, non-responsive, privileged, or any other subcategory required.[92] Thus, predictive coding is not completely automated; it requires human input to "code" documents.[93] Third, the predictive technology comes into play: the computer software receives the coded documents and "learns" what is relevant.[94] The software sorts through the complete data set and separates the more relevant documents from the less relevant documents.[95] The computer software further refines the process by creating a new data set for manual review.[96] After this stage, the software will code the data in question—for example, as most relevant, least relevant, or somewhere in between.[97] In addition, the software can generate a confidence index by pulling a subset of random, irrelevant documents and sending it back for manual review.[98] If the reviewer finds too many relevant and irrelevant documents in the same pile (as determined by the parties), then a recoding takes place. But if the reviewer checks and finds enough accurate documents, it creates a statistically significant accuracy level

87. Murphy, *supra* note 85.
88. *Id.*
89. *Id.*
90. *Id.*
91. *Id.*
92. *See id.*
93. *See id.*
94. *See id.*
95. *See id.*
96. *See id.*
97. *See id.*
98. *See Predictive Coding Video*, RECOMMIND, http://www.recommind.com/resources/videos/predictive-coding-video (last visited Feb. 25, 2012).

that allows the reviewing attorney to say, with some level of statistical confidence, that the documents are accurately coded.[99]

While many vendors offer predictive-coding software, the processes each vary slightly.[100] In fact, the E-Discovery Institute[101] conducted a thorough survey that compared the different predictive-coding technologies currently available and outlined their differences.[102] These disparities create a uniformity problem among different types of predictive coding. For example, Catalyst Repository Systems explains its process as taking "initial coding decisions made by counsel during the initial document review" and then coupling these coding decisions with "weighted key concepts and search terms," which it then applies to the "non-reviewed documents."[103] This provides a final "Predictive Ranking" for responsiveness that establishes a threshold for what documents may be responsive or what documents may require further manual review.[104]

Comparing Catalyst's system to Xerox Litigation Services and its "CategoriX" document-review process shows the potential differences in processes.[105] Xerox describes CategoriX as automatically classifying documents "by learning from samples that have been reviewed by knowledgeable case attorneys."[106] It does this by combining "attorney-supplied document assessments, together with its own statistical analyses, to create a model that will accurately and consistently generalize the attorneys' assessments across the entire review population."[107] To further enhance performance, the software conducts several quality checks to "ensure the accuracy and consistency" of input.[108] Lastly, an attorney takes a final

---

99. *See* Murphy, *supra* note 3.

100. *See* E-Discovery Inst., *eDiscovery Institute Survey on Predictive Coding*, EDISCOVERY INST., 6-10 (Oct. 1, 2010), http://www.ediscoveryinstitute.org/images/uploaded/272.pdf.

101. E-Discovery Inst., *About Us*, EDISCOVERY INST., http://www.ediscoveryinstitute.org/aboutus (last visited Nov. 11, 2012).

102. *See* E-Discovery Inst., *supra* note 100.

103. *See id.* at 6.

104. *Id.* According to Catalyst, its process follows these steps: (a) the process starts with "a list of search terms that counsel believes are likely to find responsive documents"; (b) randomly sample generated responses searching for "false hit" terms; (c) refine terms based on sampling; (d) if certain phrases were found to create common "false hits," remove those terms; (e) assign each search term a score that represents its likelihood of returning a responsive document; (f) then assign each document a "responsiveness rank based on a combination of the search terms that hit and the scores of each search term;" (g) conduct additional samples to verify scoring; and (h) determine a "cut-off" score and remove documents that are ranked as "non-responsive." *Id.* at 7.

105. *See id.* at 6-7, 10.

106. *Id.* at 10.

107. *Id.*

108. *Id.*

quality-control sample and reviews it to determine the final set's quality.[109]

A third vendor, Recommind, responded to the survey with the following: "All software, processes and workflow are the proprietary intellectual property of Recommind and cannot, therefore, be disclosed."[110] This survey was released October 1, 2010,[111] and on June 8, 2011, Recommind was awarded a patent for its predictive coding process.[112] Essentially, the patented process follows these steps: (a) humans create a control set, (b) the technology analyzes the control set to create a "seed set parameter," (c) the technology then automatically codes a first portion of documents based on the initial control set and the seed set parameter, (d) the technology analyzes the first portion using an "adaptive identification cycle," and finally, (e) the technology retrieves a second portion of documents based on the analysis of the first portion taking into account the adaptive identification cycle, which produces the final document set.[113]

These three examples show the differences between each vendor's predictive-coding techniques. Despite the similarities that exist among the processes, the admitted differences in details, including the lack of disclosure, make creating an acceptable standard that complies with the FRCP and applies to all predictive-coding techniques difficult to determine.[114]

Even if courts do not consider predictive coding to be reasonable, the parties can still use it in a variety of ways, both before and during litigation.[115] Because of the speed at which predictive coding can work, possible litigants can get a good idea of their own data set before litigation begins, allowing them to better assess their own case.[116] In addition, litigants can get a better idea of their own data set before the discovery conference required by FRCP 26(f) and can generate better keyword searches.[117] These techniques can substantially reduce litigation costs when the data sets are very large,

---

109. *Id.*

110. *Id.* at 9.

111. *Id.* at i.

112. U.S. Patent No. 7,933,859 (filed May 25, 2010); Press Release, Recommind, Recommind Patents Predictive Coding (June 8, 2011), *available at* http://www.recommind.com/releases/recommind-patents-predictive-coding.

113. U.S. Patent No. 7,933,859 fig.4 (filed May 25, 2010).

114. *See infra* Part III.

115. *See* Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on "Information Inflation" and Current Issues in E-Discovery Search*, 17 RICH. J.L. & TECH. 9, ¶ 33 (2011).

116. *See id.*

117. *See* FED. R. CIV. P. 26(f); Baron, *supra* note 115.

which is common in complex litigation.[118] Thus, even before courts determine whether parties can use predictive coding to generate a discovery response, parties can effectively and efficiently use it to analyze their own data sets.[119] These searches will also allow parties to gain a better understanding of the merits of their case, which is helpful in potential settlement discussions.

II. ACCEPTANCE OF E-DISCOVERY AND KEYWORD-SEARCH TERMS: WILL PREDICTIVE CODING TAKE A SIMILAR PATH?

Predictive coding is likely to take a similar path through the courts that e-discovery and keyword-search terms took, because predictive coding is nothing more than a new e-discovery tool. Courts are likely to accept predictive coding just as they accepted keyword-search terms.

A. The Development of E-Discovery

Law is a conservative field; changes occur slowly. But technological development is only increasing, and it is putting pressure on the law to reflect these developments. As discussed above,[120] the Supreme Court promulgated the first federal electronic discovery rules in the 1970s,[121] but it did not update them until 2006.[122] It took over thirty years for the Court to implement a modern update. In 1990, a gigabyte of information cost about $20,000 to store, but today it costs less than $1.[123] This decrease in cost has increased the amount of data stored. But the cost to have a gigabyte of data analyzed can now "easily exceed $30,000."[124]

---

118. Baron, *supra* note 115.

119. *Id.* ¶ 34.

120. *See supra* Part I.B.

121. The Subdivision (a) amendment states:

> It makes clear that Rule 34 applies to electronics data compilations from which information can be obtained only with the use of detection devices, and that when the data can as a practical matter be made usable by the discovering party only through respondent's devices, respondent may be required to use his devices to translate the data into usable form.

FED. R. CIV. P. 34 advisory committee's note (1970).

122. *See* FED. R. CIV. P. 26 advisory committee's note (2006) (amending subdivision (a) to state: "Rule 26(a)(1)(B) is amended to parallel Rule 34(a) by recognizing that a party must disclose electronically stored information as well as documents that it may use to support its claims or defenses").

123. Search & Retrieval Scis. Special Project Team, Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 192 (2007).

124. *Id.* ("But, with billable rates for junior associates at many law firms now starting at over $200 per hour, the cost to review just one gigabyte of data can easily exceed $30,000.").

The inability to assess the legitimacy of this new technology is likely a major reason for its slow adoption. There is no reason for courts to expend judicial resources evaluating technology that will quickly fade or may never work. Before a court could reach the question of whether electronic files are discoverable, it would need to understand what electronic files are and the rights that are associated with that type of media. Luckily, courts have worked through these subjects over the last forty years and can now quickly adopt rules that accept this necessary technology, including predictive coding.[125]

Courts have already fully embraced e-discovery, recognizing its benefits to litigants and the legal system.[126] Courts have also found ways to protect privileged information,[127] to prevent placing exorbitant costs and burdens on a party,[128] and to deter unscrupulous litigants from deleting discoverable electronic files.[129] For example, in *Victor Stanley I*, a magistrate judge in the US District Court for the District of Maryland protected privileged information by applying traditional tests to electronic files. In *Victor Stanley I*, the defendants produced 165 electronic documents they later claimed were privileged.[130] Prior to that case, courts used three privilege tests to determine if the accidental production of privileged information should constitute waiver of the privilege.[131] The most lenient test states there is no waiver of privilege unless relinquishment of the privilege was done knowingly or intentionally.[132] The strictest test states that waiver has occurred once disclosure occurs, because "there can no longer be any expectation of confidentiality" once disclosed.[133] The intermediate test requires the court to balance "a number of factors to determine

125. *See generally* Canon U.S.A., Inc. v. S.A.M., Inc., Civil Action No. 07-01201, 2008 WL 2522087, at *3 (E.D. La. June 20, 2008) (finding that, when ESI is reasonably accessible, the responding party is required to pay the cost of searching and producing the electronic discovery); Bills v. Kennecott Corp., 108 F.R.D. 459, 461 (D. Utah 1985) ("It is now axiomatic that electronically stored information is discoverable under Rule 34 of the Federal Rules of Civil Procedure . . . ."); Adams v. Dan River Mills, Inc., 54 F.R.D. 220, 222 (W.D. Va. 1972) (finding that Rule 34 required discovery of electronic cards and tapes in addition to production of electronic files in a usable form).

126. *See supra* Part I.B-C.

127. *See Victor Stanley I*, 250 F.R.D. 251, 259 (D. Md. 2008) (finding that the attorney-client privilege had been waived by defendant regarding ESI provided inadvertently in a discovery response).

128. *See Zubulake I*, 217 F.R.D. 309, 324 (S.D.N.Y. 2003) (concluding that a cost-shifting test should be applied to determine which party should bear the ESI discovery cost).

129. Victor Stanley, Inc. v. Creative Pipe, Inc. (*Victor Stanley II*), 269 F.R.D. 497, 533 (D. Md. 2010) (finding that the court can sanction parties for deleting ESI).

130. *Victor Stanley I*, 250 F.R.D. at 258.

131. *Id.*

132. *Id.* at 257.

133. *Id.*

whether the producing party exercised reasonable care under the circumstances to prevent against disclosure" and, if the party did so, then there is no waiver.[134] The court looked at these tests, which were only ever used with hard-copy discovery, and applied them to electronic files, finding that the tests were equally as effective with electronic documents.[135] Thus, courts can draw parallels between hard-copy discovery and e-discovery to remedy problems with privileged documents that arise with e-discovery.[136]

Courts have also found ways to prevent e-discovery requests from placing an undue burden or cost on parties. In *Zubulake I*, the US District Court for the Southern District of New York found that e-discovery is no different than paper discovery and "the presumption is that the responding party must bear the expense of complying with discovery requests."[137] The court was concerned that imposing discovery costs on plaintiffs would end discovery prematurely or even prevent plaintiffs from filing meritorious claims.[138] It noted that large companies were increasingly moving to "entirely paper-free environments" and cost shifting could prevent courts from determining claims on their merits by increasing the litigation cost for plaintiffs.[139] Thus, the court found that cost-shifting should "be considered *only* when electronic discovery imposes an 'undue burden or expense' on the responding party."[140] In all other cases, respondents must cover the costs of discovery requests. This policy protects plaintiffs from large, deep-pocketed corporate defendants that use exclusively electronic document storage, and it promotes the public policy of resolving disputes on their merits.[141] Thus, *Zubulake I* demonstrates the courts can apply traditional policies, once used only for paper discovery, to e-discovery and still maintain fairness throughout the litigation.

Finally, in *Victor Stanley II*, the district court determined whether deleting ESI in response to litigation was a violation of the discovery rules, and if so, what sanctions were appropriate.[142] In *Victor Stanley II*, the defendants had intentionally deleted a

134. *Id.*
135. *Id.* at 259.
136. *See, e.g.*, Budget Rent-A-Car of Mo., Inc. v. Hertz Corp., 55 F.R.D. 354, 357 (W.D. Mo. 1972).
137. *Zubulake I*, 217 F.R.D. 309, 317 (S.D.N.Y. 2003).
138. *Id.* at 317-18.
139. *Id.*
140. *Id.* at 318.
141. *Id.*
142. *Victor Stanley II*, 269 F.R.D. 497, 500 (D. Md. 2010).

significant amount of ESI that was relevant to the litigation.[143] The court applied traditional discovery sanction rules to the ESI's destruction, implicitly analogizing electronic files to paper files and punishing the culpable party in the same way.[144] Since the defendant destroyed ESI in bad faith, the court applied the harshest sanctions possible.[145]

The application of these rules is relatively recent, but the rules have been in development since the first recognition of electronic discovery in the 1970s.[146] It took years and a substantial amount of litigation, but the courts became more comfortable with e-discovery and the technology behind it.[147] They developed a greater understanding of e-discovery, its purpose, its process, and what expectations arise when parties use it.[148] The courts were then able to extend traditional discovery rules to e-discovery to protect litigants' rights, promote public policy, and improve judicial economy.[149] Now that courts have accepted e-discovery, it will be easier for them to accept predictive coding.

B. The Development of Keyword-Search Terms—A Possible Path for Predictive Coding

The first keyword-search cases occurred when a party requested data that was on the opposing party's computer system.[150] Prior to the advent of e-discovery, to avoid the cost of producing discovery documents, a responding-party litigant would often offer its documents for inspection by the requesting party in order to shift the cost of discovery to the requester.[151] But a responding party was unlikely to open its computer system to an opposing party for inspection because it could reveal privileged information.[152] Even if the responding party was willing to allow access, such an arrangement still failed to shift costs because courts generally required the responding party to assist the requesting party (and therefore bear the search costs) due to the requester's lack of expertise with the

143. *Id.* at 532-33.
144. *Id.* at 533-34.
145. *Id.* at 532-33.
146. *See supra* Part I.B.
147. *See supra* Part II.A.
148. *See supra* Part II.A.
149. *See supra* Part II.A.
150. Sherman & Kinnard, *supra* note 7, at 271-72.
151. Bills v. Kennecott Corp., 108 F.R.D. 459, 462 (D. Utah 1985).
152. *See* Sherman & Kinnard, *supra* note 7, at 271-72.

responder's specific system.[153] As a result, the cost-shifting tactic was ineffective for e-discovery, and courts required parties to show "undue burden or expense" before shifting the cost of electronic discovery to the requesting party.[154]

Earlier cases, like *Budget Rent-A-Car of Missouri, Inc. v. Hertz Corp.*, held that a party could not dump a mountain of documents on opposing counsel in response to a discovery request; rather, the responding party had an obligation to search through the documents and find the relevant information for the requesting party.[155] Now, similarly, responding parties must use keyword searches to narrow down the amount of ESI they provide to the opposing counsel. For example, in *Haugen*, the US District Court for the District of Utah found that keyword-search terms that prevented the parties from generating an unwieldy volume of documents would be acceptable.[156] Consequently, the court allowed the requesting party to submit twenty-five search terms to the court for approval.[157] It also noted that it was the obligation of the producing party to execute the keyword searches and provide the relevant information, given that it was more familiar with the computer system.[158] Thus, *Haugen* followed precedent, particularly that of *Budget Rent-A-Car*, and held that the party familiar with the record system needed to provide a specific response rather than a document dump.[159]

The next case that supported keyword searches was *Zubulake II*. In *Zublake II*, the district court suggested that a party that reasonably anticipates litigation has a preservation obligation, and preserving documents located through a simple keyword search can meet this obligation.[160] The court did not require parties to review the documents; instead, it required only that the parties retain the electronic documents.[161] It stated: "[C]ounsel and client must take *some reasonable steps* to see that sources of relevant information are located."[162] Thus, the court intimated that keyword searching would

153. *Id.* at 278-79.

154. *Id.* at 296.

155. Budget Rent-A-Car of Mo., Inc. v. Hertz Corp., 55 F.R.D. 354, 357 (W.D. Mo. 1972) (finding that a responding party cannot give a mass of records when research of those records is feasible for only one familiar with them).

156. Procter & Gamble Co. v. Haugen, 179 F.R.D. 622, 632 (D. Utah 1998), *aff'd in part, rev'd on other grounds sub nom.* Proctor & Gamble Co. v. Haugen, 222 F.3d 1262 (10th Cir. 2000).

157. *Id.* at 633.

158. *Id.*

159. *Id.*; *see also supra* note 155 and accompanying text.

160. *Zubulake II*, 229 F.R.D. 422, 432 (S.D.N.Y. 2004).

161. *Id.*

162. *Id.*

have been sufficient to meet their common-law preservation obligation[163] and would have satisfied the reasonableness requirement in the discovery rules.[164] The court further determined that a party has an ongoing obligation to seek out and preserve electronic files from the "key players" on their side of the litigation.[165] But keyword searching may not reach these key players because they may store information in various locations; as such, the court required more than just keyword searching to show a reasonable effort to preserve electronic documents.[166]

Since *Zubulake II*, several other courts have explicitly adopted the idea that keyword searching could meet a party's preservation responsibility and satisfy the reasonability requirement imposed by the discovery rules.[167] These courts showed a strong understanding of technology and its ability to take a mountain of data and reduce it to those files that are relevant and important.[168] While validation by these courts does not establish that keyword searching is always accurate and effective, it does reiterate the fact that effective keyword searches have many advantages.

Courts have now begun to embrace keyword searches more fully.[169] But, as courts understand the technology behind electronic files, new issues and problems arise. One major problem is the efficacy of keyword-search terms.[170] Some courts suggest that a lay person is incapable of creating effective keyword-search terms and that only experts should be permitted to execute the searches.[171] This is because keyword-search terms involve the "interplay . . . of the

163. "Once a party reasonably anticipates litigation, it must suspend its routine document retention/destruction policy and put in place a 'litigation hold' to ensure the preservation of relevant documents." *Id.* at 431.

164. *Id.* at 432 ("It may be possible to run a system-wide keyword search; counsel could then preserve a copy of each 'hit.' Although this sounds burdensome, it need not be. Counsel does not have to review these documents, only see that they are retained.").

165. *Id.* at 433-34 (internal quotation marks omitted) (defining key players as those people initially identified in a party's disclosure, or "employees likely to have relevant information").

166. *Id.* at 432-34 (reasoning that keyword searches would not reach all documents from key players because some key players printed out emails and kept them in hardcopy only, while others used separate computer files).

167. *E.g.*, *In re* Seroquel Prods. Liab. Litig., 244 F.R.D. 650, 663 (M.D. Fla. 2007) (quoting *Zubulake II* about keyword-search terms helping to determine what electronic files to preserve).

168. *See supra* Part II.B.

169. *See Victor Stanley I*, 250 F.R.D. 251, 259-60 (D. Md. 2008); United States v. O'Keefe, 537 F. Supp. 2d 14, 23-24 (D.D.C. 2008).

170. *O'Keefe*, 537 F. Supp. 2d at 24.

171. *Id.*

sciences of computer technology, statistics and linguistics."[172] Any future motion arguing that search terms are insufficient must be in accordance with FRE 702.[173] This shows various courts' depth of understanding of keyword-search terms, and indicates they have become far more comfortable with current technology. As courts develop their understanding of technology, it will be easier for them to adapt to future technological introductions.

Keyword searching has now become a regular part of the discovery process, and courts are starting to fully accept keyword searches as reasonable under the discovery rules.[174] Attorneys recognized the cost savings that could be achieved with appropriate keyword searches and actively encouraged courts to recognize keyword searches as reasonable.[175] Academics and professional groups developed best practices for attorneys and wrote law review articles, and litigators educated the bench.[176] This multi-factored approach quickly garnered the ideas necessary to create a strong framework to develop rules on keyword searches in e-discovery.[177] It is this framework, an understanding of the whole system, which is necessary before courts will begin to adopt a new technology, such as predictive coding.

Courts will be able to look back to the rules that they developed for e-discovery and keyword searches and apply them to predictive coding.[178] Courts will need to determine what rules protect litigants' rights and promote judicial economy, as well as what the best practices are for predictive coding and how to implement them. A major concern will be how to differentiate between the various predictive-coding technologies. Some companies have already patented their processes of predictive coding.[179] By understanding how these processes work, courts will be able to categorize the different processes and develop appropriate rules for each. The academic community will need to do some of this work by conducting studies on the efficacy of the different predictive-coding technologies that are available.[180] Working groups will need to create best

172. *Id.*

173. *Id.*

174. *See supra* Part I.C.

175. *See Victor Stanley I*, 250 F.R.D. 251, 260 n.10 (D. Md. 2008); Baron & Wolfe, *supra* note 74, at 230; Search & Retrieval Scis. Special Project Team, *supra* note 123, at 200.

176. *See Victor Stanley I*, 250 F.R.D. at 260 n.10; WORKING GRP. ON ELEC. DOCUMENT RETENTION & PROD., *supra* note 20, at 8-11.

177. *See* Search & Retrieval Scis. Special Project Team, *supra* note 123, at 193-94.

178. *See supra* Part II.A-B.

179. *See supra* note 112 and accompanying text.

180. Some of this work has already begun. *See infra* Part III.

practices for how litigants should handle this new technology. Courts will need to understand the technology so they can create rules about which predictive-coding processes are acceptable, meaning reasonable under the rules, and which are not. The courts will need to ensure that new technology promotes judicial economy, rather than create a possible roadblock in the discovery process by allowing litigants to drag out the litigation.

It will be the legal community's responsibility to promote and educate the bench about these new technologies. The substantial time and cost savings predictive coding can offer is difficult to understate.[181] As discovery can now involve millions of documents, new technology will be necessary to reduce the document load to something lawyers can manage. Predictive coding is the technology capable of doing so.

III. WHY COURTS SHOULD ADOPT PREDICTIVE CODING AND HOW IT CAN BE IMPLEMENTED

Manual document review is seen as the "gold standard" of discovery review—that is, as the most effective form of document review and the standard against which all other standards are measured.[182] It is time for newer document-review processes to test this "myth" of perfection.[183] Manual review has flaws, such as increased human labor, fatigue, inattention, and boredom.[184] While technology-driven processes may suffer from these same flaws, as they still require human input, they are less likely to fall victim to them, as they require minimal human review. Predictive coding, for example, requires only a small pool of relevant documents, which entail less manual human review; thus, fewer human failures would occur.[185] If manual review is the current gold standard and considered reasonable according to the FRCP, but it is unable to handle the currently growing amount of ESI, then the legal field must adapt and find a way for predictive coding to comply with the FRCP.[186]

---

181. *See supra* Part I.D.

182. *See* Search & Retrieval Scis. Special Project Team, *supra* note 123, at 199 ("[T]here appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible—perhaps even perfect—and constitutes the gold standard by which all searches should be measured.").

183. Grossman & Cormack, *supra* note 9, ¶ 61.

184. *Id.* ¶ 58.

185. See *id.* ¶¶ 2, 58.

185. *See* Search & Retrieval Scis. Special Project Team, *supra* note 123 (noting the drastically increasing amount of reviewable data, and the concomitantly increasing cost of review).

186. *See supra* note 124 and accompanying text.

As discussed in Part I, the FRCP governs certain discovery practices. The discovery rules attempt to balance costs and completeness.[187] Discovery is limited by the court, or on a motion to the court, based on the expense of producing discovery and its benefit to the case.[188] In addition, discovery productions must be complete.[189] Finally, attorneys must certify, after reasonable inquiry, that their discovery request, response, or objection is consistent with the discovery rules.[190] Predictive coding can meet the FRCP standard if parties can show it is more accurate, efficient, and responsive than manual review.[191]

The transition to newer, more efficient, and more accurate modes of document review will require the legal field, as a whole, to undergo a process of education. Academia will need to conduct quantitative research studies to compare and contrast different document-review processes.[192] For example, legal researchers should conduct experiments comparable to those conducted by the Text Retrieval Conference (TREC).[193]

TREC has created TREC Legal Track with a goal to "develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections."[194] TREC Legal Track designed an "Interactive Task" to simulate real-world discovery requests by using actual cases and the corresponding discovery documents.[195] The participating teams included professional companies using manual-review processes as well as teams using technology-assisted review processes.[196] TREC Legal Track made this data available to researchers who then performed statistical analysis to determine which of the processes proved more accurate.[197] The study concluded that "technology-assisted review can achieve at least as high recall as manual review, and higher precision, at a fraction of

187. Grossman & Cormack, *supra* note 9.

188. FED. R. CIV. P. 26(b)(2)(C)(iii).

189. *Id.* at 37(a)(4); *see also* Grossman & Cormack, *supra* note 9.

190. FED. R. CIV. P. 26(g)(1).

191. Grossman & Cormack, *supra* note 9 ¶ 58.

192. *Id.* ¶ 1.

193. *Id.* ¶ 3.

194. *Id.* ¶ 29 (quoting *Overview*, TEXT RETRIEVAL CONFERENCE, http://trec.nist.gov/overview.html (last updated Aug. 10, 2010)).

195. *Id.* ¶ 30. The 2009 TREC Legal Track documents comprised "a collection of emails that had been produced by Enron in response to requests from the Federal Energy Regulatory Commission." Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, TEXT RETRIEVAL CONFERENCE (TREC) 2009 PROCEEDINGS, § 2.2.1, http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf (last visited Nov. 2, 2012).

196. Grossman & Cormack, *supra* note 9, ¶ 45.

197. *Id.* ¶¶ 3, 44.

the review effort, and hence, a fraction of the cost."[198] In addition, there was "*not one single measure for which manual review [was] significantly better than technology-assisted review.*"[199]

The researchers above concluded that the next question should not be "*whether* technology-assisted review *can* improve on manual review" but instead "*which* technology-assisted review process(es) will improve *most* on manual review."[200] How researchers should conduct these future tests is beyond the scope of this Note, but researchers should continue to compare manual document-review processes to new technology-assisted document-review processes. This way, researchers can quantitatively determine that the new technology-driven document-review processes are as good—if not better—than manual document review.[201]

Currently, ESI has grown to such massive proportions that the cost to identify relevant documents through manual review is dwarfing other costs in the e-discovery process.[202] Parties must decide whether the burden to identify and produce documents is worth the cost, given the issues, amount in controversy, and the other concerns discussed previously in this Note.[203] The solution to the growing size of ESI is to counter it with a new technology that can identify the same quantity of documents at a low cost.[204] But lawyers and courts are concerned with the effectiveness of the new processes. They wonder whether they can rely on these new processes and whether the processes satisfy the reasonable-inquiry requirement in the FRCP.[205]

Courts must solve the latter problem—whether certifying attorneys can rely on predictive coding as part of their reasonable inquiry. This problem requires a multi-faceted solution that involves all members of the legal field working together. But even if courts accept predictive coding generally, or other technology-assisted review, how can certifying attorneys be sure that the process they use in a particular matter satisfies the reasonable-inquiry requirement? A solution to this problem is simple: courts should require the producing party, or the party who is relying on predictive coding, to

198. *Id.* ¶ 55.

199. *Id.* ¶ 54.

200. *Id.* ¶ 61.

201. *See id.* ¶¶ 4, 61 (using statistical analysis to find that the technology-driven document-review processes in the study were better than manual document review).

202. *Id.* ¶ 6; *see* Search & Retrieval Scis. Special Project Team, *supra* note 123.

203. Grossman & Cormack, *supra* note 9, ¶ 6.

204. Baron, *supra* note 115, ¶ 6.

205. *See id.* ¶¶ 6-7 (arguing that technology-assisted review, including predictive coding, can fall within both the letter and the spirit of the FRCP).

supply statistically significant quantitative data regarding the quality of the discovery being produced.

Comparative studies are critical because, unlike familiar Westlaw and Lexis searching techniques, courts are less familiar with the new technology-assisted document-review processes.[206] Comparative studies between manual review and predictive coding can help show courts the reasonableness of relying on predictive coding by comparing it to something courts understand: manual review or keyword searches. Further, comparing the quality between manual review and predictive coding may provide the court with persuasive evidence of its accuracy and reliability.[207]

Professional groups such as the Sedona Conference[208] and the E-Discovery Institute[209] can, importantly, continue to educate the bench and bar about the possible issues that could arise when using processes such as predictive coding.[210] For example, professional groups can create best practices to help eliminate discovery disputes.[211] Best practices should, at a minimum, discuss the discovery processes that parties should employ, the quality assurance checks that parties should use, and the protocol for handling disclosures of privileged documents.[212] If the parties understand where the issues may arise under the new technology-assisted processes, they can deal with potential problems on the front end, saving valuable time for both the parties and the court.[213] Even if a dispute arises that a court must resolve, the publications from these groups, bolstered by academic research, will be an invaluable resource in helping a court determine that predictive coding complies with the requirements of the FRCP.[214] But "predictive coding" is a broad term, including many different processes.[215] How, then, should a court determine whether an attorney's discovery response, generated using predictive coding, constitutes a reasonable inquiry to a discovery request, as required under the FRCP?[216]

---

206. *See* Search & Retrieval Scis. Special Project Team, *supra* note 123, at 197.

207. *See* Grossman & Cormack, *supra* note 9, ¶ 8.

208. *See Frequently Asked Questions*, *supra* note 21.

209. *About Us*, ELECTRONIC DISCOVERY INST., http://www.ediscoveryinstitute.org (last visited Feb. 25, 2012).

210. *See* Grossman & Cormack, supra note 9, ¶¶ 5, 61.

211. Search & Retrieval Scis. Special Project Team, *supra* note 123.

212. *See id.* at 210-11.

213. *See id.* at 209.

214. *See id.* at 204.

215. E-Discovery Inst., *supra* note 100, at 5.

216. Working Grp. on Elec. Document Retention & Prod., *supra* note 16, at 214; *supra* note 49 and accompanying text.

The E-Discovery Institute compared the different predictive-coding technologies currently on the market, which highlighted the substantial differences in the processes between the various vendors' predictive-coding technologies.[217] These differences create a uniformity problem.[218]

Various professional industries, such as manufacturing, have addressed this uniformity problem.[219] For example, the Sedona Conference Working Group on Best Practices for Document Retention & Production, in its Commentary on Achieving Quality in E-Discovery Process, asserts there are at least five measures of quality: (a) judgmental sampling, (b) independent testing, (c) reconciliation techniques, (d) inspection to verify and report discrepancies, and (e) statistical sampling.[220] Judgmental sampling is already used in the traditional discovery process when a senior attorney randomly selects a batch of coded documents and reviews them for accuracy, determining whether the error rate is too high such that an additional round of review is necessary.[221] This quality measure is not statistically significant, however, as its effectiveness cannot be measured; thus, it may not offer predictive coding much help.[222]

Independent testing simply requires a third party to report on whether an approach's results can be repeated or replicated.[223] This approach is likely of little value to companies using predictive-coding techniques because they maintain the algorithms as proprietary.[224] Reconciliation techniques have a place in e-discovery, but their application is likely not helpful to courts in determining whether a predictive-coding process is reasonable because its results are not statistically verifiable.[225] Manual inspection and verification is a labor-intensive process in which one party inspects and another verifies, and it is often used in an apprentice setting.[226] This is unlikely to be of any help in whittling down massive quantities of ESI using predictive coding because it would still require manual review of every document.[227]

---

217. *See supra* Part I.D.

218. *See supra* Part I.D.

219. Working Grp. on Best Practices for Document Retention & Prod., *supra* note 21 at 311.

220. *Id.* at 300, 310-11.

221. *Id.*

222. *See id.* at 310-11.

223. *Id.*

224. *See id.*

225. *See id.* at 303, 310.

226. *See id.* at 311, 320-21.

227. *See id.* at 303, 319.

The final approach named by the Sedona Conference to measure the quality of e-discovery processes is statistical sampling, which manufacturers regularly use in quality-assurance checks by sampling a small portion of a manufacturing run to verify that the process is working correctly.[228] Statistical sampling is typically used when it is time and cost prohibitive to test each individual item, as sampling constitutes a "scalable solution" that works well, regardless of the size of the sampled population.[229] Statistical sampling provides statistical confidence about the sampled population and the accuracy of that population.[230] Statistical sampling is the best choice to assure the quality of predictive coding because manufacturers created it to check the quality of a large set of goods (here, documents) for which time and cost prohibit individual quality assurance.[231]

Statistical sampling offers a solution to the uniformity problem identified above.[232] Instead of testing the *process*, statistical sampling would test the *results* of that process.[233] Its application would extend to other technology-assisted discovery processes as well.[234] Statistical sampling's practical application to predictive coding would work by taking a random sample of documents and having a senior attorney code them.[235] If the senior attorney manually codes enough of the documents that the predictive-coding process has also coded, then statistical sampling can determine the error rate with statistical significance.[236]

The Advisory Committee contemplated sampling in connection with FRCP 26(b)(2)(C).[237] As discussed in the Advisory Committee's Notes, sampling may be necessary "to learn more about what burdens and costs are involved in accessing the information, what the information consists of, and how valuable it is for the litigation" when the court conducts a "good-cause" determination.[238] Courts have also used sampling to determine whether a party should restore all of their

---

228. *See id.* at 311-12.

229. *Id.* at 312.

230. *Id.*

231. *See id.*

232. *See supra* Parts I.D, III.

233. *See* Working Grp. on Best Practices for Document Retention & Prod., *supra* note 21 at 312.

234. *Id.*

235. *See id.* at 312 n.48.

236. *Id.*

237. *Id.* at 313-14.

238. FED. R. CIV. P. 26 advisory committee's note (2006) (commenting on amendment subdivision (b)(2)).

backup tapes based on the sample's cost and importance.[239] But statistical sampling has its flaws.[240]

Specifically, manufacturers designed statistical sampling to take a sample of a *homogenous* group of items.[241] Thus, if statistical sampling shows that a discovery response missed only a very small percentage of documents, statistical sampling does not account for the *importance* of the missed documents.[242] Maybe the documents missed were only "barely relevant," but maybe they were "smoking gun" documents.[243] There is a *variance* in the importance of the missed documents.[244] Thus, a proper quality check must control for both the number of missed documents and the importance of those documents.[245]

For example, if a quality check reveals a statistically significant error rate (that is, missed relevant documents) of 5 percent, the difficult question becomes: Is that good enough? A solution would look at the importance of the missed documents. Remember, the quality check would not have accounted for the missed documents' importance. At this point, it may be prudent for the requesting party to review the quality check and determine how important the missed documents were. If the requesting party can then show opposing counsel (or the court, if necessary) that these documents are smoking-gun relevant documents, then the responding party should update the predictive coding software to include them and create another document set. This way, the parties can begin to control for variance.

But most predictive-coding processes already have a quality-control check built in, sometimes at multiple stages.[246] Statistical sampling can provide statistically significant error rates, along with an appropriate probable variance.[247] Comparisons of the current manual-review standard and predictive coding could assure the court that the process of review conducted by the responding party was a "reasonable inquiry."[248] The court in *Victor Stanley I* noted how

239. *See Zubulake I*, 217 F.R.D. 309, 323-24 (S.D.N.Y. 2003); McPeek v. Ashcroft, 202 F.R.D. 31, 34-35 (D.D.C. 2001).

240. Working Grp. on Best Practices for Document Retention & Prod., *supra* note 21 at 327.

241. *Id.* at 312 n.49.

242. *See id.* at 328.

243. *See id.* at 312 n.49.

244. *See id.*

245. *See id.*

246. *See supra* Part I.D.

247. *See* Working Grp. on Best Practices for Document Retention & Prod., *supra* note 21 at 312.

248. *See id.* at 308 n.28 (internal quotation marks omitted).

important academics, law firms, corporate counsel, and companies providing ESI-discovery services could be in creating best practices.[249] If these groups created a set of best practices that counsel adhered to, it would "certainly . . . support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes."[250]

The legal profession as a whole will need to move predictive coding forward.[251] Professional institutes can establish best practices through working groups of active litigation specialists, researchers can conduct studies such as the TREC Legal Track that compare the current manual-review gold standard to predictive-coding process, and academics can begin finding ways that traditional and e-discovery rules are applicable to predictive coding. These efforts will educate the bar and bench, leading to courts' ultimate acceptance of predictive coding and a consequent decrease in costs to parties and the judiciary.[252]

Such acceptance has already begun. In *Moore v. Publicis Groupe SA*,[253] a district court judge affirmed Magistrate Judge Andrew Peck's order that predictive-coding technology be utilized in the discovery process.[254] The Plaintiff objected to this order as violating FRCP 26 and FRE 702.[255] But the district court held in favor of Judge Peck's order, comparing predictive coding to a "traditional keyword search."[256] The court found that it would be difficult to determine if predictive coding would be less reliable than keyword searches.[257] It also appears the gold standard of manual review took a hit when the district judge stated, "[E]ven if all parties here were willing to entertain the notion of manually reviewing the documents, such review is prone to human error and marred with inconsistencies from various attorneys' determination of whether a document is responsive."[258] This is likely the first case in what will become a deluge of court acceptance for predictive coding. Judge Peck was intimately familiar with predictive-coding technology, as were the

249. *See Victor Stanley I*, 250 F.R.D. 251, 260 n.10 (D. Md. 2008).
250. *Id.*
251. Paul & Baron, *supra* note 2, ¶ 6.
252. *Cf. Victor Stanley I*, 250 F.R.D. at 260 n.10.
253. 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).
254. *Id.* at *1-2.
255. *Id.* at *1.
256. *Id.* at *2-3.
257. *Id.* at *2.
258. *Id.* at *3.

parties.[259] As predictive coding becomes more "mainstream" and understood, more courts are going to be facing requests by parties to utilize this new technology. This case will likely set the precedent for a majority of future cases on whether predictive coding complies with the discovery rules.

IV. CONCLUSION

The battle between man and machine review is imminent. Predictive coding is not far from becoming a reality of mainstream e-discovery.[260] Innovative technology like predictive coding is helping drive estimates that the e-discovery market will become a billion-dollar industry within the next year,[261] but it is still unclear whether courts would allow a responding party to rely on predictive coding to generate a discovery response.[262] ESI's massive growth shows no plan of stopping, and our discovery system needs to adapt, or discovery costs will continue to skyrocket. While courts have adopted keyword searching, this stopgap measure will only take the e-discovery process so far.[263]

The FRCP does not demand perfection; the biggest hurdle is convincing the court to recognize that predictive coding is "reasonable" under the FRCP.[264] Given the courts' acceptance of keyword searches, courts will likely accept predictive coding, especially given the increasing prevalence and costs of e-discovery.[265] But the combined work of academia, professional working groups, industry, and litigators is critical to speed up the courts' acceptance of predictive coding. Because predictive-coding processes lack uniformity, predictive coding requires a quality-control solution that allows courts to compare across different predictive-coding methodologies. Currently, the best solution is to utilize statistical sampling to verify the accuracy of results and measure the variance of missed

259. *See, e.g.*, Andrew Peck, *Search, Forward; Will Manual Document Review and Keyword Searches Be Replaced by Computer-Assisted Coding?*, L. TECH. NEWS (Oct. 2011), http://www.recommind.com/sites/default/files/LTN_Search_Forward_Peck_Recommind.pdf.

260. *See Victor Stanley I*, 250 F.R.D. at 259 n.9 (discussing other technology-assisted discovery tools similar to predictive coding).

261. Press Release, Recommind, *supra* note 112.

262. *See* Working Grp. on Best Practices for Document Retention & Prod., *supra* note 21.

263. *See Victor Stanley I*, 250 F.R.D. at 260.

264. *See supra* Part I.A.

265. *See supra* Part I.C.

documents. With the proper studies, an educated bench and bar, and the proper methodologies, predictive coding has an opportunity to gain acceptance on its first discovery challenge.

Nicholas Barry[*]

* J.D. Candidate 2013, Vanderbilt University Law School. The Author would like to thank his wife for her support and grace through the publication process. The Author would also like to thank his editors: Frances Kammeraad, Michael Dearington, and Shane Valenzi. Their insightful comments and detailed editing were a great help in improving this Note.