South Carolina Law Review

Volume 64 | Issue 3 Article 6

Spring 2013

Predictive Coding: Emergin Questions and Concerns

Charles Yablon
Benjamin N. Cardozo School of Law

Nick Landsman-Roos
Standford Law School

PREDICTIVE CODING: EMERGING QUESTIONS AND CONCERNS

Charles Yablon* & Nick Landsman-Roos**

---

*Professor of Law, Benjamin N. Cardozo School of Law.
**J.D. Candidate, Stanford Law School, 2013.

The authors would like to thank Ron Dolin, William Kellerman, Matt Kesner, and Jonathan Land for discussing their firms' uses of predictive coding. While the discussion throughout this Article reflects information they provided, we have decided not to quote their comments directly so as to avoid any potential confidentiality concerns. We would also like to thank Maura Grossman and Richard Marcus for their comments on an earlier draft of this piece. All errors, of course, remain our own.

I. INTRODUCTION

Technology-assisted review—also referred to as "predictive coding," "computer-aided review," and "content-based advanced analytics"—is the most important development in e-discovery to have occurred in some time.[1] While technical characterizations of the process vary largely because of differences across software platforms, basically predictive coding is a process whereby computers are programmed to search large quantities of documents using complex algorithms to mimic the document selection process of a knowledgeable, human document review.[2] It is said to do such a review faster and without many of the dangers of human error.[3] Because of its speed and accuracy, it has been described as a fundamental change in the way discovery is conducted.[4] In fact, the popular legal press has predicted that technology-assisted review will ultimately end the armies of document-reviewing contract attorneys employed by law firms.[5] Law firms have scrambled to educate themselves about what predictive coding is, how it can be used, and whether it should be embraced or resisted.[6] Despite all this attention, however, predictive coding had never actually been used in any reported case until very recently. That changed in February 2012 when Judge Peck, in *Da Silva Moore v. Publicis Groupe*,[7] authorized the use of predictive coding for the first time in a reported federal case.

---

1. *See* 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 (noting that this year's hot topic in e-discovery is computer-assisted coding).

2. *See* NICHOLAS M. PACE & LAURA ZAKARAS, RAND CORP., WHERE THE MONEY GOES: UNDERSTANDING LITIGANT EXPENDITURES FOR PRODUCING ELECTRONIC DISCOVERY 59 (2012), *available at* http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf; *see also* Joe Palazzolo, *Why Hire a Lawyer? Computers Are Cheaper*, WALL ST. J., June 18, 2012, at B1 (describing how predictive-coding programs work).

3. *See, e.g.*, PACE & ZAKARAS, *supra* note 2, at 61–65 (citations omitted) (discussing the accuracy of predictive coding).

4. Scott Vernick, *Predictive Coding: Three Things You Need to Know About This Year's Biggest Legal Tech Trend*, HUFFINGTON POST (Aug. 15, 2012, 6:36 PM), http://www.huffingtonpost.com/scott-vernick/three-things-you-need-to-_b_1773959.html.

5. John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N.Y. TIMES, Mar. 5, 2011, at A1; Chris Opfer, *Rise of the Machines: New Technology May Spell the End for NYC's Bottom-Rung Lawyers*, N.Y. MAG. (Mar. 14, 2012, 3:45 PM), http://nymag.com/daily/intelligencer/2012/03/new-technology-may-spell-doom-for-new-lawyers.html; Kenneth Anderson, *Is Contract Lawyering Doomed by Algorithm?*, VOLOKH CONSPIRACY (May 5, 2012, 7:24 PM), http://www.volokh.com/2012/05/05/is-contract-lawyering-doomed-by-algorithm/.

6. *See* Palazzolo, *supra* note 2, at B5.

7. Da Silva Moore v. Publicis Groupe (*Da Silva Moore II*), No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *1, *12 (S.D.N.Y. Feb. 24, 2012).

Other cases are already beginning to emerge.[8] Predictive coding is too powerful and effective a technique to be ignored. It has the potential to be less costly and more accurate than manual human review, and it will undoubtedly be used increasingly by courts in coming years.[9] Yet many courts, commentators, and lawyers still know little about what predictive coding is or how it can be used. Even less consideration has been given to the new kinds of legal disputes and problems that may arise with the use of predictive coding. That preliminary inquiry is the subject of this Article.

One might think that predictive coding is nothing more than a technological improvement in current discovery practice. As such, it should have minimal impact on the underlying law. But experience has taught us that technological changes often put strains on existing legal norms, and the law of discovery is no exception.

Widespread use of predictive coding will raise numerous new legal questions that may well require reconsideration of some of the most basic principles of current discovery law. To start, predictive coding will once again require courts to make methodological decisions about how to apply the new technology within the existing discovery and e-discovery rules. Like keyword searching or document sampling, there are a great number of technical questions that courts must address regarding how parties carry out predictive coding, the methodological choices that are made, and which party decides those questions. Can (and should) the techniques of predictive coding be made sufficiently transparent and the technology made sufficiently accessible such that all parties and the judge will be able to provide effective legal oversight of its use?

More fundamentally, the very cost savings and increased accuracy available through predictive coding raise the question of how courts will apply it to

---

8. For a discussion of the unfolding case law, see *infra* Part III.

9. *See* 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, at 44 (2011), http://jolt.richmond.edu/v17i3/article11.pdf (finding that technology-assisted processes are more cost-effective and efficient and can yield results superior to those of exhaustive manual review, as measured by recall and precision); Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70, 79 (2010) (finding that computer classification is at least as accurate as manual review); BRUCE HEDIN ET AL., NAT'L INST. OF STANDARDS & TECH., OVERVIEW OF THE TREC 2009 LEGAL TRACK 16 & tbl.5 (2009), *available at* http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf; DOUGLAS W. OARD ET AL., NAT'L INST. OF STANDARDS & TECH., OVERVIEW OF THE TREC 2008 LEGAL TRACK 2 (2008), *available at* http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf; David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMM. OF THE ACM 289, 290 (1985), *available at* http://cacm.acm.org/magazines/1985/3/10387-an-evaluation-of-retrieval-effectiveness-for-a-full-text-document-retrieval-system/abstract (finding that keyword searching produces high false-negative rates). For a criticism of these studies, see Steve Green & Mark Yacano, *Computers vs. Humans? Putting the TREC 2009 Study in Perspective*, N.Y. L.J. (Oct. 1, 2012), *available at* http://www.newyorklawjournal.com/PubArticleNY.jsp?id=1202573060444.

discovery disputes under existing concepts of proportionality review. Will courts order the same discovery as previously, thereby allowing the discovered party to obtain the full benefits of the cost savings of predictive coding? Or, will they decide that the lower marginal cost of discovery available through predictive coding allows for broader discovery under the proportionality standard? Will they use the ability of predictive coding to rank documents in accordance with a preexisting standard of relevance to cut off or shift costs with respect to discovery of "less relevant" documents? Can predictive coding be made sufficiently flexible so that parties can apply its relevance standards in cases where there are shifting theories of liability? And what happens when one party vehemently objects to its use? Will courts order predictive coding over a party's objection? To what extent can courts order sua sponte that predictive coding be used, even when neither party has suggested such a possibility?

This Article provides a preliminary analysis of these questions. Part II provides a general description of the processes of predictive coding. We discuss some points of technical disagreement that have already arisen regarding implementation of predictive coding in order to highlight likely points of conflict that courts must address in the future. Part III then surveys the limited number of cases in which a court has addressed the use of predictive coding. We turn in Parts IV and V to emerging legal questions and concerns. In Part IV, we consider various ways in which predictive coding may change the conduct of proportionality review under Federal Rules of Civil Procedure 26(b)(2)(B) and (C)(iii). The issues are whether predictive coding may change the burden–benefit analysis set forth in those rules—perhaps by lowering the cost of discovery generally, by lowering the marginal cost of additional discovery, or by allowing for more accurate, cost-effective discovery—and how these technological benefits will be distributed between the parties.

In Part V, we consider how the use of predictive coding and court-ordered use of the process affects the traditional presumption that a responding party selects the means by which documents are collected and deemed relevant. We will discuss whether predictive coding can be ordered by a court at the request of one party over the complete objection of the other. May it be ordered sua sponte when neither party has suggested its use? And if courts begin ordering the use of predictive coding, to what extent will a refusal to participate in the formulation of a protocol constitute bad faith in the discovery process?

This Article will not provide definitive answers to the many methodological, legal, or philosophical questions raised by the new technology of predictive coding. At best, it will make tentative suggestions concerning some of them. Our primary goal is to at least identify many of the relevant questions, highlight several major areas of concern, and try to show what is at stake in the resolution of these important, emerging issues.

II. PREDICTIVE CODING: THE TECHNICAL PROCESS

A. What Is Predictive Coding: A General Overview

To a considerable extent, describing predictive coding as computer-generated or computer-conducted discovery is hyperbole, if not downright fantasy.[10] All of the technologies described as predictive coding still require a substantial level of human involvement, from collecting the documents that will be searched, to training the system to determine relevancy or privilege, to deciding what constitute relevant documents.[11] Moreover, the technologies that exist cannot assemble theories of a case, sort documents based on varying grounds of liability, or even decide whether a document is helpful or hurts a particular side's case. The technology that exists is more accurately described as technology-assisted review, because all that computers are capable of right now is making human-conducted discovery more efficient and accurate. Further complicating the description of technology-assisted review is the fact that the technical processes vary across proprietary systems.[12] Thus, no general description of technology-assisted review can perfectly describe the process of every software program.

The use of technology-assisted review began around 2008, when a small number of law firms started exploring ways in which they could use computers and sophisticated software to make the discovery review process more efficient.[13] The underlying technology, called machine learning, had been available for decades,[14] but only in about the last five years has the legal profession considered its use.[15] Initially, the technology was not very helpful; it could not simultaneously be hosted and be behind a firm's firewall,[16] and none

---

10. *See* Sharon D. Nelson & John W. Simek, *Predictive Coding: A Rose by Any Other Name*, LAW PRAC., July/Aug. 2012, at 20, 20 (noting the lack of agreement in defining predictive coding).

11. *See, e.g., id.* at 20, 22–24 (noting that predictive coding "doesn't take humans out of the equation").

12. *See generally* Jeffrey Parkhurst, *Bridging the Gap in E-Discovery: The Emergence of Conceptual Semantic Search*, ASS'N OF CERTIFIED E-DISCOVERY SPECIALISTS (Dec. 5, 2012), http://www.aceds.org/bridging-the-gap-in-e-discovery-the-emergence-of-conceptual-semantic-search/ (discussing the "virtually limitless variations of training algorithms used" for technology-assisted review).

13. *See* Johnathan Jenkins, *What Can Information Technology Do for Law?*, 21 HARV. J.L. & TECH. 589, 596 (2008) (discussing the use of technology to "streamline the discovery process").

14. *See, e.g.*, Tom M. Mitchell, *Does Machine Learning Really Work?*, 18 AI MAG., Fall 1997, at 11 (discussing the possibilities of machine learning).

15. *See, e.g.*, Jenkins, *supra* note 13, at 602 (noting the potential for machine learning in the legal profession).

16. *See generally* William W. Belt et al., *Technology-Assisted Document Review: Is It Defensible?*, 18 RICH. J.L. & TECH. 10, at 10 n.23 (2012), http://jolt.richmond.edu/v18i3/article10.pdf (stating that technology-review applications are "limited to instances where the technology is deployed behind the firewall").

of it was very affordable.[17] Early programs did not perform well with complex documents such as system logs that are voluminous and have complex terms.[18] But by 2010, use of computer systems utilizing machine learning began gaining momentum as an alternative to manual document review by humans.[19] This shift was especially true in securities cases where relevant documents were more readable—and less frequent in antitrust and intellectual property cases in which the document population was more technical and varied.[20]

The technology now being used is a type of machine-learning technology that allows a computer to assist in "predicting" how documents should be classified based on limited user input.[21] The process generally involves feeding a computer system with a small set of documents—called a "seed set"—that has been selected by attorneys with knowledge about the responsiveness of those documents.[22] Using this small set of documents and the coding of those documents determined by attorneys, the computer creates a model that then generates a prediction score for every document based on its degree of responsiveness.[23] The assignment of responsiveness scores "becomes increasingly accurate as the software continues to learn from human reviewers what is, and what is not," relevant or privileged.[24]

The first step, before any software is used, is culling the junk documents that are clearly nonresponsive or irrelevant.[25] Sometimes culling will also take the form of deduplication; that is, removing duplicate copies of a document from a document population.[26] This initial process of removing clearly irrelevant or duplicate documents is necessary because the licensing structure for many

17. *See generally* Jenkins, *supra* note 13, at 605 (discussing the prohibitively high costs of certain technologies).

18. *See, e.g.*, *id.* at 597 (noting the limitations of early machine learning systems).

19. *See, e.g.*, Palazzolo, *supra* note 2, at B1 (reporting an aviation hangar owner's use of a computer to conduct document review in 2010).

20. *See* Jim Eidelman, *Best Practices in Predictive Coding: When Are Pre-Culling and Keyword Searching Defensible?*, E-DISCOVERY SEARCH BLOG (Jan. 9, 2012), http://www.catalystsecure.com/blog/2012/01/best-practices-in-predictive-coding-when-are-pre-culling-and-keyword-searching-defensible/.

21. *See* Palazzolo, *supra* note 2, at B1.

22. PACE & ZAKARAS, *supra* note 2, at 59–60.

23. *Id.* at 60.

24. *Id.* at 59; *see also* EDISCOVERY INST., EDISCOVERY INSTITUTE SURVEY ON PREDICTIVE CODING 2 (2010), *available at* http://www.sfldata.com/wp-content/uploads/2012/07/2010_EDI_PredictiveCodingSurvey.pdf (explaining that predictive coding is "a combination of technologies and processes in which decisions pertaining to the responsiveness of records gathered or preserved for potential production purposes . . . are made by having reviewers examine a subset of the collection and having the decisions on those documents propagated to the rest of the collection without reviewers examining each record" (internal quotation marks omitted)).

25. *See* Rob McFarlane & Russell Petersen, *E-Discovery: Computer-Assisted Coding Is a Powerful Tool to Control Complex Case E-Discovery Costs*, INSIDE COUNS. (May 30, 2012), http://www.insidecounsel.com/2012/05/30/e-discovery-computer-assisted-coding-is-a-powerful?t=e-discovery.

26. *See* PACE & ZAKARAS, *supra* note 2, at 11 (defining deduplication).

predictive coding tools requires customers to pay higher fees to process files.[27] Culling reduces the per-document licensing costs and thereby reduces the overall costs of coding.[28]

Next, practitioners must train the system. The way in which a system is trained varies across software platforms.[29] One method is knowledge engineering, which entails the construction of linguistic and other models that replicate the manner in which humans approach document review.[30] Another approach employed by some technology-assisted-review tools is machine learning, which requires the creation of seed sets.[31] As already indicated, a "seed set" is the "initial [t]raining [s]et provided to the learning [a]lgorithm" of the coding software.[32] The seed set documents may be selected through random or judgmental sampling.[33] The random sampling approach requires the selection of a random sample from the total document population as a seed set, which is then coded by attorneys, either manually or with the assistance of a keyword search, as relevant or not relevant.[34] Judgmental sampling, on the other hand, requires that attorneys with knowledge of the case select documents—already uncovered through discovery—as "seeds" that they have determined are clearly fitting or not fitting a particular document category (e.g., a document is clearly relevant or not, privileged or not).[35] That seed set of documents is fed into the software to train it for assessing relevancy.[36] Moreover, one commentator has recently advocated yet another approach: the creation of a fake "perfect" document that can train software exactly what to look for.[37] The way in which this initial training set of documents is selected—whether through random or judgmental sampling—has become, as detailed below, a point of considerable debate.[38]

Training the system is an iterative process.[39] Regardless of whether knowledge engineering or machine learning is used or how the initial training set of documents is chosen, after attorneys code those documents, they are then

---

27. *See Early Case Assessment and Predictive Coding Technologies Often Share High ESI Processing Costs*, FLEX DISCOVERY, http://www.flexdiscovery.com/early-case-assessment-and-predictive-coding-technologies-often-share-high-esi-processing-costs/ (last visited Feb. 12, 2013).

28. *See id.*

29. *See generally* PACE & ZAKARAS, *supra* note 2, at 59–60 (detailing how predictive coding works).

30. *See* Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge*, 2013 FED. CTS. L. REV. 1, 9 [hereinafter *Grossman-Cormack Glossary*], *available at* http://www.fclr.org/fclr/articles/html/2010/grossman.pdf.

31. *See id.* at 22.

32. *Id.* at 29.

33. *Id.*

34. *See id.* at 27.

35. *See id.* at 21.

36. *See* PACE & ZAKARAS, *supra* note 2, at 60.

37. *See Predictive Coding Power User Panel from Carmel*, ESIBYTES (July 24, 2012), http://www.esibytes.com/predictive-coding-power-user-panel-from-carmel/.

38. *See infra* Part II.B.1.

39. *See* PACE & ZAKARAS, *supra* note 2, at 60.

analyzed by software that creates a model to be used to screen other documents and assign scores to each document to reflect the probability that it fits within the model.[40] The attorneys then take a random sample of documents coded at the outset of the review process to measure the effectiveness of the algorithm at various stages of training and to determine at what point training may cease.[41] Attorneys review these samples, make their own decisions about whether the document fits the characteristic for which it is being coded (e.g., relevancy or privilege), and then that document is fed back into the software to refine the model.[42] This sample is called a control set.[43] The size of the control set depends on the overall document population size as well as estimates as to the number of relevant documents that will appear within the overall population.[44] Too small a control set can result in a skewed sample and a high margin of error when the predictive coding ultimately occurs.[45] This process of drawing a control set continues until the software is optimized—it correctly predicts the coding of a document in a certain percentage of cases, as determined by attorneys.[46] The control set, which is used to measure accuracy, is separate from the seed set, which is used to train the system.[47]

Most software then applies a prediction to all the documents.[48] Some apply it just to the text of the document, while others also include the underlying metadata.[49] Each document is assigned a prediction score, expressed as a percentage, indicating the likelihood of responsiveness.[50] These prediction scores are often examined using a summary measure designed to assess the quality of prioritization.[51] That measure reflects the probability that a randomly chosen document is correctly ranked.[52] Some programs have built-in transparency features to help explain each document's prediction score; these features identify links between documents and summarize important words and

---

40. *See id.*

41. *See id.*

42. *See id.* Unlike traditional, manual document review, where review is often done by contract attorneys or junior associates, the review involved in predictive coding often involves senior partners or small teams with highly specialized knowledge of a case. *See* Peck, *supra* note 1.

43. *See Grossman-Cormack Glossary*, *supra* note 30, at 13 (defining "control set").

44. *See generally Predictive Coding-Measurement Challenges*, E-DISCOVERY 2.0 (July 6, 2012), http://www.clearwellsystems.com/e-discovery-blog/tag/margin-of-error/ (discussing control sets and confidence levels of samples).

45. *See id.*

46. *See* Peck, *supra* note 1.

47. *See Grossman-Cormack Glossary*, *supra* note 30, at 13, 29.

48. *See* Peck, *supra* note 1.

49. *See* Eidelman, *supra* note 20.

50. *See* PACE & ZAKARAS, *supra* note 2, at 62.

51. *See id.*; *see also Grossman-Cormack Glossary*, *supra* note 30, at 16, 28 (defining "F_1" and "Relevance Ranking").

52. *See* PACE & ZAKARAS, *supra* note 2, at 61–62.

phrases in each document.[53] A minority of programs perform binary (yes/no) coding and do not provide prediction scores;[54] however, those programs may be falling out of favor. The set of documents that are either not returned by the search process or are deemed not relevant is called the null set.[55]

A sample may once again be taken and then coded by a human to assess the reliability of the software's coding.[56] That sample is used to assess confidence levels—the measurement of the belief in the sample's reliability[57]—and margin of error or confidence intervals—a prediction of how precisely the sample estimates the true value of the whole population.[58] For instance, a 90% confidence interval means that in 90% of the samples with this confidence level, there is an expected true population value within the range specified by the experiment's confidence interval (e.g., ± 2%).[59] A "[c]onfidence [l]evel is not the [p]robability that the true value is contained in any particular [c]onfidence [i]nterval; [rather,] it is the [p]robability that the method of estimation will yield a [c]onfidence [i]nterval that contains the true value."[60]

"[T]he fraction of [d]ocuments indentified as [not r]elevant . . . that are in fact [r]elevant" is called elusion.[61] A low elusion value—determined after the use of the coding software—has been identified as evidence of an effective review, but that measure only quantifies the relevant documents missed, not the quantity found.[62] The overall error rate—"[t]he fraction of all [d]ocuments that are incorrectly coded"—is similarly often "advanced as evidence of an effective search."[63] But that too can be misleading because it is influenced by prevalence.[64] For instance, where there are 1 million documents in a document population, and 10,000 (1%) are relevant, a review that found no relevant documents would have a 1% error rate and would be 99% accurate despite missing 10,000 relevant documents.[65]

Once a model has been created and software has been trained, attorneys must then decide which documents to produce.[66] One possibility is to conduct manual review of all documents with prediction scores above a certain percentage threshold of relevancy or responsiveness and discard all other

---

53. *See* Hilary McQuaideon, *Falcon Discovery Ushers in Savings with Transparent Predictive Coding*, E-DISCOVERY 2.0 (Sept. 4, 2012), http://www.clearwellsystems.com/e-discovery-blog/2012/09/04/falcon-discovery-ushers-in-ediscovery-savings-with-transparent-predictive-coding/.
54. *See* Peck, *supra* note 1.
55. *See Grossman-Cormack Glossary*, *supra* note 30, at 25.
56. *See* PACE & ZAKARAS, *supra* note 2, at 60.
57. *See Grossman-Cormack Glossary*, *supra* note 30, at 25.
58. *See id.* at 12, 22.
59. *See id.*
60. *Id.* at 12.
61. *Id.* at 15.
62. *Id.*
63. *Id.* at 16.
64. *Id.*
65. *Id.*
66. *See* PACE & ZAKARAS, *supra* note 2, at 60.

documents.[67] Another option is to produce all documents meeting a certain threshold, exclude all documents below another percentage threshold, and then conduct manual review of the documents following within those two percentage thresholds.[68] A third approach is to produce all documents meeting a certain percentage threshold and then exclude the rest without conducting any manual review.[69]

B. Issues in Implementation

As Judge Peck predicted in an article predating the first judicial approval of the use of predictive coding,[70] much of the disagreement between lawyers has and will continue to concern the implementation and use of a predictive coding protocol. Because it is very unlikely that a court will ever issue an opinion choosing a particular software with a particular methodology, and as the facts of each case demand different uses of the predictive coding software, the decisions about how predictive coding will be used must be made on a case-by-case basis. In what follows, we summarize the particular implementation-related questions that may often be points of contention.

1. How the Seed Set Is Created

There is an ongoing debate about how best to create a seed set. The disagreement largely turns on whether random sampling or something approximating judgmental sampling is the best approach.[71] The division between predictive coding experts as to what is the proper approach was on full display in a recent panel discussion at the Carmel Valley eDiscovery Retreat.[72] There, Maura Grossman, an attorney and author of studies relating to the effectiveness of predictive coding, advocated for the use of predictive coding utilizing seed sets of predetermined responsive documents.[73] But Tom Gricks, one of the lead attorneys in *Global Aerospace Inc. v. Landow Aviation, L.P.*,[74] believes that random sampling is the proper approach in creating seed sets.[75] Similarly, this methodological disagreement is reflected within the literature on predictive coding as well.[76]

---

67. *Id.*
68. *Id.*
69. *Id.*
70. *See* Peck, *supra* note 1.
71. *See Predictive Coding Power User Panel from Carmel*, *supra* note 37.
72. *Id.*
73. *Id.*
74. No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012); *infra* Part III.C.
75. *See Predictive Coding Power User Panel from Carmel*, *supra* note 37.
76. *See, e.g.*, PACE & ZAKARAS, *supra* note 2, at 60 (listing several possible approaches).

This debate is really no different than other disagreements as to whether to use random or judgmental sampling. Creating a seed set through random sampling ensures that the sample being analyzed is representative of the larger document population.[77] A principal problem with random sampling is that it often takes longer to train the predictive coding system, or requires a larger sample, because the sample's richness—the number of relevant documents it contains—is lower.[78] On the other hand, judgmental sampling—the taking of a sample with a greater than average number of the most important documents—may be more helpful in identifying the best-case-scenario documents that might be turned up through predictive coding.[79] With more relevant documents in the sample, the system will train faster. There is, however, a risk of over-richness depending on the number of relevant documents included in the sample, or a risk of skew depending on how and why attorneys selected the documents to be in the sample.[80]

A related issue that is generally not discussed is which individuals are creating the seed set and training the system. A number of studies have noted judgment variation among human reviewers;[81] even if predictive coding removes much of that human input, there is still an initial review that must be done by humans, and that introduces the possibility of human review variation and error.[82] Because random sampling will often require a larger sample to establish a sufficient level of richness,[83] it runs the risk of introducing increased human review variation and error. On the other hand, judgmental sampling also carries that risk because the seed set is made up of documents hand selected by attorneys.[84]

We have previously argued that judgmental sampling is preferable to random sampling in the context of sampling documents before making cost-shifting decisions because an over-representative sample provides a court with more information about the particular documents most likely to be important to the case.[85] While we are not taking definitive positions on these methodological choices (which, as previously noted, are best determined on a case-by-case basis), our prior reasoning likely applies here as well. Using judgmental sampling to create a seed set means the predictive coding system will be trained

---

77. *See Grossman-Cormack Glossary*, *supra* note 30, at 27; *Predictive Coding Power User Panel from Carmel*, *supra* note 37.

78. *See id.*

79. *See id.*

80. *See id.*

81. *See* PACE & ZAKARAS, *supra* note 2, at 55–56.

82. *See, e.g.*, Roitblat et al., *supra* note 9, at 77 (discussing the variability in human relevance judgments due to random and systematic factors).

83. *See supra* notes 77–78 and accompanying text.

84. *See* PACE & ZAKARAS, *supra* note 2, at 60.

85. *See* Charles Yablon & Nick Landsman-Roos, *Discovery About Discovery: Sampling Practice and the Resolution of Discovery Disputes in an Age of Ever-Increasing Information*, 34 CARDOZO L. REV. 719, 768–69 (2012).

using documents that attorneys believe are the most relevant to the case and most representative of those for which each side is looking.[86] Thus, assuming the process is not skewed by adversarial or strategic considerations, it is likely to be most helpful in training the software what to look for. If the purpose of predictive coding is to identify those documents that are most relevant, why not give the system those documents that attorneys have already deemed most helpful or hurtful to the case?[87]

2. Whether the Seed Set Is Discoverable

A second debate is whether the seed set should be discoverable, and if it is, what information needs to be disclosed—just the documents used for the seed set, or does counsel need to explain why those documents were used and how they were coded? Judge Peck first raised the possibility that a requesting party will seek "the documents that were used to train the computer-assisted coding system."[88] In doing so, he suggested that "[c]ounsel would not be required to explain why they coded documents as responsive or non-responsive, just what the coding was."[89] Whether a seed set is discoverable may well depend on how it was formed. If the seed set is merely a random sample of the entire document population and is produced without coding as to whether documents are deemed responsive or not, the production is unlikely to concern a producing party.[90] If, on the other hand, the seed set is made up of documents selected or coded by a producing party as relevant, production of that seed set has a much higher probability of disclosing attorney impressions of the case.[91]

---

86. *See Grossman-Cormack Glossary*, *supra* note 30, at 21.

87. Craig Ball takes this reasoning further, advocating for the training software to use not actual documents but "contrived document fabrications" or "imagined evidence" to train the system in a way in which it will identify documents most sought after by attorneys. *See* Craig Ball, *Imagining the Evidence*, L. TECH. NEWS (Aug. 1, 2012), http://www.law.com/jsp/lawtechnology news/PubArticleFriendlyLTN.jsp?id=1202564068885&slreturn=20121117134031. While it is unlikely Ball's proposal will ever be implemented, it underscores the wide variety of opinions as to how a seed set should be built.

88. *See* Peck, *supra* note 1.

89. *Id.* The protocol Judge Peck approved in the *Da Silva Moore* case takes this approach. *See Da Silva Moore II*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *13–23 (S.D.N.Y. Feb. 24, 2012).

90. *See* Declaration of Paul J. Neale in Support of Plaintiffs' Rule 72(a) Objections to Magistrate Judge Peck's Feb. 8, 2012 Discovery Rulings at 8–9, *Da Silva Moore II*, No. 11 Civ. 1279(ALC)(AJP) (S.D.N.Y. Feb. 22, 2012), 2012 WL 607412.

91. *See* Yablon & Landsman-Roos, *supra* note 85, at 778; *see also* Wayne C. Matus & John E. Davis, *Does Your Search Pass Judicial Scrutiny?*, N.Y. L.J. (Oct. 27, 2008), *available at* http://www.newyorklawjournal.com/PubArticleNY.jsp?id=1202425538348&slreturn=20130115114204 (stating that by producing search keywords, to show the reasonableness of the search, a party can disclose their impressions and legal theories of a case).

In appropriate cases, possible solutions might be for the parties to collaborate on creating a seed set or allow the requesting party to create or to code a randomly generated seed set.[92]

3. *The Prediction Score Threshold for Production*

A third point of disagreement among attorneys is what prediction score threshold should exist for production.[93] Specifically, when predictive coding software produces responsiveness scores for an entire document population, at what percentage point is a document deemed nonresponsive? A party responding to document requests will generally be interested in setting a higher relevancy score, whereas the requesting party will want a lower relevancy score.[94] Thus, the prediction score threshold ultimately used will often be a product of attorney negotiation coupled with the particularities of a given case.

On a more general level, there is also the possibility that parties could agree on a percentage or number of the overall document population, rather than a particular prediction score. Because prediction scoring can order documents from most to least responsive, it is possible that instead of picking a relatively arbitrary prediction score, the parties could agree that only, for example, 60% of the documents, as selected by responsiveness, will be produced.[95] Of course, all of these decisions have profound implications for the overall cost of discovery.[96] The ability of predictive coding to rank documents with respect to the responsiveness also raises the concern that courts and litigants will confuse the "responsiveness" of a document with its "relevance" or "importance," distinctions which are discussed more fully in Part IV of this Article.

4. *The Appropriate Confidence Level and Interval*

A fourth consideration is what confidence level and interval should be used in determining the size of the seed set or other random sample to be used in assessing the accuracy characteristics of the document population.[97] Confidence levels are relevant to predictive coding because, as mentioned already, they are a standard by which to measure the probability that a random sample is

---

92. *See* Yablon & Landsman-Roos, *supra* note 85, at 772.

93. *See* PACE & ZAKARAS, *supra* note 2, at 60.

94. *See generally Grossman-Cormack Glossary*, *supra* note 30, at 13 (defining "cutoff").

95. In *Da Silva Moore II*, discussed in Part III.A *infra*, the defendants proposed to review and produce only the 40,000 most relevant documents in connection with the case.

96. *See generally* PACE & ZAKARAS, *supra* note 2, at 66–69 (citations omitted) (discussing the cost effectiveness of predictive coding).

97. *See supra* notes 57–60 and accompanying text. For a detailed analysis of confidence level issues, see William Webber, A Tutorial on Interval Estimation for a Proportion, with Particular Reference to E-Discovery (Aug. 2, 2012) (unpublished manuscript), *available at* http://www.umiacs.umd.edu/~wew/papers/sisa.pdf.

representative of the population of documents as a whole.[98] A recurring argument among lawyers "is whether it is most appropriate to use 95%, 99% or any other confidence level."[99] The party producing documents generally argues for a lower confidence level as a 99% level requires more documents to be in a seed set.[100] Conversely, a requesting party generally advocates for a higher confidence level, assuming that it will ultimately lead to more accurate coding and more accurate assessments of that coding.[101] But the question remains how important a 95% versus a 99% confidence level is. These are, after all, merely arbitrary thresholds that statisticians have standardized as a matter of custom.[102]

5. Where Subsequent Manual Review Is Appropriate

A final area of disagreement among attorneys is whether or to what extent a manual document review should occur after predictive coding is complete.[103] Some practitioners believe that after the parties negotiate a prediction-score threshold, all documents above that threshold must be produced and all documents below it are deemed irrelevant.[104] Others have advocated for a modified approach, where all documents above a particular threshold are produced, all documents below a different threshold are excluded, and the documents in the score range between those two thresholds are reviewed manually.[105]

* * *

In sum, there are many methodological choices associated with the use of predictive coding software—choices that attorneys are just now becoming aware of and on which courts have not yet ruled and are unlikely to resolve any time soon. That said, case law is beginning to emerge on the subject of predictive coding.

---

98. *See supra* notes 57–60 and accompanying text; *see also* David J. Kessler, *Debunking the Seven Biggest Myths of Predictive Coding*, LEGAL TECH. NEWSL. (ALM Law Journal Newsletters, New York, N.Y.), June 2012, at 2, *available at* http://pdfserver.amlaw.com/legaltechnology/LJN_Legal_Tech_Newsletter_0612.pdf (debunking "Myth 6" that confidence intervals measure reasonableness).

99. Kessler, *supra* note 98, at 2.

100. *See id.*

101. *See id.*

102. *See id.*

103. *See* Peck, *supra* note 1.

104. *See* PACE & ZAKARAS, *supra* note 2, at 60.

105. *See id.*

III. THE UNFOLDING CASE LAW

To date, five courts have issued opinions or orders regarding the use of predictive coding.[106] Together, they represent an evolution in the way courts have begun to treat predictive coding.

A. The Da Silva Moore *Case*

Da Silva Moore v. Publicis Groupe SA[107] is a federal employment discrimination case that was filed on February 24, 2011.[108] The complaint alleged that Publicis Groupe (Publicis), one of the largest advertising and public relations firms in the world, and MSLGroup Americas (MSL), the subsidiary that runs its public relations network in the United States, reserved its positions of power and influence for men only and that women are "rarely [able to] break through the glass ceiling" to the ranks of senior management.[109] Additionally, the complaint alleged that plaintiff Da Silva Moore was an employee of defendant MSL for six years, holding such titles as "director," "managing director," and "global director," yet never obtained any "real advancement."[110] Instead, she was terminated after returning from maternity leave.[111] The case was brought as a class action on behalf of "female [public relations] employees"[112] under Title VII of the Civil Rights Act, analogous state laws, and a local administrative ordinance.[113] The complaint was amended on April 14, 2011, to add four more named plaintiffs and to allege additional claims under the Equal Pay Act and the Fair Labor Standards Act.[114]

---

106. *See* discussion *infra* Parts III.A–E; *see also* Sheila Mackay, *Hooters! You're Ordered to Use Technology-Assisted Review*, XEROX E-DISCOVERY TALK BLOG (Dec. 3, 2012), http://ediscoverytalk.blogs.xerox.com/2012/12/03/hooters-youre-ordered-to-use-technology-assisted-review/ (listing the five cases).

107. No. 11 Civ. 1279(ALC)(AJP), 2012 WL 6082454 (S.D.N.Y. Dec. 3, 2012).

108. *See id.* at *1; Class Action Complaint at 36, *Da Silva Moore II*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412 (S.D.N.Y. Feb. 24, 2011) (No. 11 CIV 1279), 2011 WL 655226.

109. Class Action Complaint, *supra* note 108, at 1–2.

110. *See id.* at 6.

111. *See id.*

112. *See id.* at 16.

113. *See id.* at 26–30, 32 (citing 42 U.S.C. § 2000e-2(a) (2006); MASS. GEN. LAWS ch. 151B, § 4 (2004 & Supp. 2012); N.Y. EXEC. § 296 (McKinney 2010 & Supp. 2013); N.Y.C., N.Y. ADMIN. CODE § 8-107 (2010)). Plaintiff also brought individual claims under the Family and Medical Leave Act. *See id.* at 30–32 (citing Family and Medical Leave Act (FMLA) of 1993, 29 U.S.C. §§ 2601–2619 (2006 & Supp. V 2011)).

114. *See* Da Silva Moore v. Publicis Groupe SA (*Da Silva Moore VI*), No. 11 Civ. 1279(ALC)(AJP), 2012 WL 2574742, at *1 (S.D.N.Y. June 29, 2012) (citing Fair Labor Standards Act of 1938, 29 U.S.C. §§ 201–219 (2006 & Supp. V 2011)). Plaintiff also sought to bring the Fair Labor Standards Act claims as a "collective action" to which all current, former, and future female public relations employees could opt in. *See id.*

The case was originally assigned to Judge Richard Sullivan.[115] At a pretrial conference on July 21, 2011, Judge Sullivan established a deadline of June 30, 2012, for fact discovery.[116] On November 28, 2011, Judge Sullivan designated Judge Andrew Peck for general pretrial supervision.[117] Judge Peck is an experienced and well-regarded jurist who, as already noted, has spoken and written extensively on matters of e-discovery generally and predictive coding in particular.[118]

1. The Predictive Coding Protocol

Defendants had always expressed their intention to utilize predictive coding to handle e-discovery in this case, a proposal which had met some resistance from plaintiffs. At the first conference held before Judge Peck on December 2, 2011, defense counsel referred to plaintiffs' "reluctance to utilize predictive coding."[119] Judge Peck, speaking to defense counsel, commented, "You must have thought you died and went to Heaven when this was referred to me."[120] This statement later formed one of the primary bases for plaintiffs' motion to disqualify Judge Peck.[121]

The outlines of the predictive coding protocol that was ultimately ordered by the court took shape in the winter of 2012. Defendants had taken the position that they wanted to spend no more than $200,000 in producing relevant electronically stored information in accordance with plaintiffs' requests.[122] Defendants believed they could accomplish this task by using predictive coding

---

115. *See* Memorandum of Law in Opposition to Plaintiffs' Motion for Recusal or Disqualification at 2 n.1, Da Silva Moore v. Publicis Groupe, 868 F. Supp. 2d 137 (S.D.N.Y. 2012) (No. 11 Civ. 1279(ALC)(AJP)), 2012 WL 1687376. The case was subsequently transferred to Judge Andrew L. Carter, Jr. on January 9, 2012. *See id.*

116. *See Da Silva Moore VI*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 2574742, at *1 (S.D.N.Y. June 29, 2012) (citing Case Management Plan and Scheduling Order at 1, *Da Silva Moore VI*, 2012 WL 2574742 (No. 11-CV-1279(RJS)), ECF No. 35).

117. *See* Da Silva Moore v. Publicis Groupe (*Da Silva Moore V*), 868 F. Supp. 2d 137, 140 (S.D.N.Y. 2012).

118. *See, e.g.*, Hon. Andrew J. Peck, THE SEDONA CONF., https://thesedonaconference.org/bio/peck-andrew (last visited Feb. 16, 2013) (providing biographical information). In an October 2011 article, Judge Peck stated, "In my opinion, computer-assisted coding should be used in those cases where it will help 'secure the just, speedy, and inexpensive' (Fed. R. Civ. P. 1) determination of cases in our e-discovery world." Peck, *supra* note 1 (quoting FED. R. CIV. P. 1).

119. *See* Transcript of Dec. 2, 2011 Conference at 8, Da Silva Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 517207 (S.D.N.Y. Feb. 14, 2012), ECF No. 51.

120. *See id.*

121. *See* Memorandum of Law in Support of Plaintiffs' Motion for Recusal or Disqualification at 4, *Da Silva Moore V*, 868 F. Supp. 2d 137 (S.D.N.Y. June 15, 2012) (No. 11 Civ. 1279(ALC)(AJP)), 2012 WL 1421293.

122. *See Da Silva Moore II*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *3 (S.D.N.Y. Feb. 24, 2012).

software to locate and produce the 40,000 "most relevant" documents at a cost of approximately $5 per document.[123]

Plaintiffs' position on predictive coding was much more equivocal. While never stating that they were completely opposed to the use of predictive coding, the plaintiffs stated that they had "multiple concerns" about defendant's proposal and the way defendants intended to utilize predictive coding.[124] At the December 2, 2011 Conference, Judge Peck instructed the parties to continue to talk about these issues; however, he also expressed his generally positive views concerning predictive coding.[125]

By the February 8, 2012 Conference, defendants had presented plaintiffs and the court with a proposed predictive coding protocol to cover e-discovery in the case.[126] Although plaintiffs had agreed to certain elements of the proposed protocol,[127] they had many more objections.[128] At the February 8 hearing, Judge Peck dealt with those objections—sometimes by forging a compromise between the parties' positions, sometimes by ruling for defendants, and frequently by postponing the issue until after the document retrieval, produced by defendants' predictive coding software, had taken place.[129]

For example, the first discovery issue discussed at the hearing involved plaintiffs' request to include, among the email files searched, the files of certain individuals whom they called "comparators."[130] These "comparators" were male employees who, it was alleged, were either performing the same duties as plaintiffs for higher compensation or were receiving the same compensation as plaintiffs while performing less important or less skilled jobs.[131] Plaintiffs

---

123. *See id.*

124. *See id.* (quoting Transcript of Dec. 2, 2011 Conference, *supra* note 119, at 21).

125. *See id.* (quoting Transcript of Dec. 2, 2011 Conference, *supra* note 119, at 20–21).

126. *See id.* at *4 (citing Transcript of Feb. 8, 2012 Conference, *Da Silva Moore II*, 2012 WL 607412 (No. 11 Civ. 1279(ALC))). Although it appears that the plaintiffs had presented their own proposal for predictive coding discovery—which they allege Judge Peck ignored—by February 8, 2012, the court and both parties were focused on the defendants' proposed protocol. *See id.* at *3–5 (citing Transcript of Feb. 8, 2012 Conference, *supra* at 23–25, 27–39, 44–51).

127. For example, working backward from the desire to achieve a confidence level of 95% plus or minus 2% with respect to their determination of the number of responsive documents, defendants proposed that a random sample of 2,399 documents be examined to get a sense of what percentage of documents are likely relevant in the system. *Id.* at *5 (citing Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 58–59). Those 2,399 randomly selected documents would later be incorporated in the seed set used to train the system. *Id.* (citing Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 59–61). Plaintiffs apparently agreed to the use of the 2,399 random sample set, although there is some indication that they had earlier sought to make the confidence level 99%, which would have required examination of a larger set of randomly selected documents. *See* Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 58.

128. *See Da Silva Moore*, 2012 WL 607412, at *5 (citing Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 58–62).

129. *See generally* Transcript of Feb. 8, 2012 Conference, *supra* note 126 (taking up plaintiffs' objections to defendants' proposed protocol).

130. *See id.* at 28–29.

131. *See id.* at 26–30.

sought discovery of their files in order to compare their compensation and responsibilities with those of the plaintiffs.[132] Judge Peck denied any discovery of these comparators' files, at least in the first phase of discovery, holding that the information plaintiff sought from them was so different from the information sought in its other ESI requests that it could not easily be made part of the predictive coding protocol.[133] Instead, Judge Peck suggested that depositions of these comparators would be a better way of obtaining this discovery.[134]

Another question was whether the files of Olivier Fleurot, MSL's CEO, should be included in the documents searched by predictive coding.[135] The problem was that most of Mr. Fleurot's emails were in French, and it was unclear how effective defendants' predictive coding software would be in understanding and making relevance determinations with regard to French documents.[136] For these and other reasons, Fleurot's emails were also excluded from the phase one search.[137]

On the question of the predictive coding protocol itself, Judge Peck accepted its basic provisions but continued to note that the technology involved was unproven; the results of the search could not be known in advance; and that there would be time after the predictive coding software was run to observe its results, determine how well it worked, and consider various objections of, or modifications suggested by, the plaintiffs.[138] For this reason, he declined to consider what was probably plaintiffs' main objection to the protocol—the lack of any preestablished standard setting forth the degree of accuracy that had to be achieved by the system.[139] He concluded that such issues would be better considered and determined "'down the road' when real information [became] available to the parties and the Court."[140]

In an order dated April 26, 2012, Judge Carter affirmed Judge Peck's orders on these discovery matters, including his predictive coding protocol.[141] In that order, Judge Carter stressed the highly deferential nature of review of nondispositive orders under Rule 72(a) of the Federal Rules of Civil Procedure and stated that Judge Peck's rulings were "well reasoned" and "consider the potential advantages and pitfalls of the predictive coding software."[142]

---

132. *See id.* at 28–30.
133. *See id.* at 31.
134. *See id.* at 29–31.
135. *See id.* at 31–35.
136. *See id.* at 32–35.
137. *See id.* at 35.
138. *See id.* at 75–77, 83–84.
139. *See Da Silva Moore II*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *8 (S.D.N.Y. Feb. 24, 2012).
140. *Id.*
141. *See* Da Silva Moore v. Publicis Groupe SA (*Da Silva Moore III*), No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534, at *3 (S.D.N.Y. Apr. 26, 2012).
142. *See id.* at *1–2 (citing FED. R. CIV. P. 72(a)).

Meanwhile, on April 13, 2012, plaintiffs filed a formal motion seeking the recusal or disqualification of Judge Peck.[143] In early May, plaintiffs sought a stay of e-discovery and the predictive coding protocol pending Judge Carter's decisions on their motions to amend their complaint and certifying the opt-in action.[144] Although Judge Peck initially denied that request, he reconsidered after plaintiffs formally objected to his ruling, and on May 14, 2012, he issued an order granting a stay of MSL's production of ESI pending Judge Carter's decision.[145] Judge Carter decided those motions favorably for the plaintiffs in a decision dated June 28, 2012.[146] However, Judge Peck denied plaintiffs' recusal motion in a rather extensive opinion issued on June 15, 2012.[147] Likewise, Judge Carter denied plaintiffs' objections to Judge Peck's recusal decision in a November 7, 2012 order.[148]

2. *Holdings of the* Da Silva Moore *Opinions*

While the *Da Silva Moore* case has become instantly famous for being the first to "recognize[] that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases,"[149] it is a highly tentative decision in most respects, deferring many more issues than it resolves.[150] This lack of resolution reflects not only Judge Peck's awareness that defendants' technology was untried and unproven and the nature of the technological process itself—whose results cannot be known with any precision in advance—but it also reflects the procedural posture of the case.[151]

---

143. *See id.* at *1 (citing Plaintiffs' Notice of Motion for Recusal or Disqualification, *Da Silva Moore III*, 2012 WL 1446534 (No. 11 Civ. 1279(ALC)(AJP)), ECF No. 169).

144. *See* Plaintiffs' Rule 72(a) Objection to the Magistrate's Apr. 25, 2012 Discovery Rulings at 2–4, Da Silva Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1698980 (S.D.N.Y. May 14, 2012) (No. 11-CV-1279(ALC)(AJP)), 2012 WL 1677941 (citations omitted).

145. *See* Da Silva Moore v. Publicis Groupe (*Da Silva Moore IV*), No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1698980, at *1 (S.D.N.Y. May 14, 2012).

146. *See Da Silva Moore VI*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 2574742, at *1 (S.D.N.Y. June 29, 2012).

147. *See Da Silva Moore V*, 868 F. Supp. 2d 137, 140 (S.D.N.Y. June 15, 2012).

148. Da Silva Moore v. Publicis Groupe SA, No. 11 Civ. 1279(ALC)(AJP) (S.D.N.Y. Nov. 7, 2012), ECF No. 342.

149. *Da Silva Moore II*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *1 (S.D.N.Y. Feb. 24, 2012).

150. *See id.* at *8–11 (discussing how the decision to use computer-assisted review in this case was simple because the parties agreed to it and raising possible issues that could arise in future cases).

151. *See id.* The case was at an early stage, in which the identity of the parties (particularly the named plaintiffs) and the claims of the plaintiff were still in a state of flux. *See, e.g.*, *Da Silva Moore VI*, 2012 WL 2574742, at *1 (adding claims and named plaintiffs). Plaintiffs were resisting making a motion for class certification and were not seeking discovery in connection with their class claims. *See generally* Case Management Plan and Scheduling Order, *supra* note 116 (setting discovery deadlines and limiting parties to those before the court or represented by the named parties, unless the court granted leave to join additional parties). Finally, these initial battles were

Nonetheless, it is worth reviewing what the court in *Da Silva Moore* did rule on and what remains unclear in light of those rulings. First and foremost, it is now undeniably true that one federal court has held that predictive coding may be the basis for a protocol governing the discovery of ESI.[152] Judge Peck's and Judge Carter's rulings also strongly suggest that the consent of both parties is not a necessary condition for making such a ruling.[153] Although Judge Peck emphasized in his decision that the plaintiffs were not adamantly opposed to all forms of predictive coding, it is clear that his specific order regarding the use of predictive coding was made over plaintiffs' frequently repeated objections.[154]

It appears that the basis for the *Da Silva Moore* court's approval was a general finding that predictive coding was a more cost-effective search method than keyword or manual review.[155] The court's language implies, although does not quite state, that the parties' consent is irrelevant so long as the court determines that predictive coding is a more "appropriate" method than available alternatives like keyword search or manual review.[156] Given the court's prior comment about cost and human error, it would appear that "appropriateness" is to be judged by which method provides the most satisfactory results at the lowest cost. By these measures, predictive coding would appear to have a decided edge.[157]

---

the opening round in what was contemplated as phased discovery, and many difficult issues could be and were relegated to later phases of the process. *See Da Silva Moore II*, 2012 WL 607412, at *8–11.

152. *See Da Silva Moore II*, 2012 WL 607412, at *12.

153. *See id.* (concluding that computer-assisted review should be considered when it can save significant legal fees); *Da Silva Moore III*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534, at *3 (S.D.N.Y. Apr. 26, 2012) (adopting Judge Peck's orders in *Da Silva Moore II*).

154. *See Da Silva Moore II*, 2012 WL 607412, at *6–8 (commenting on and dismissing plaintiff's objections but noting that the objections were before District Judge Carter).

155. *See Da Silva Moore III*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534, at *3 (S.D.N.Y. Apr. 26, 2012). As Judge Carter noted in adopting Judge Peck's order:

> There simply is no review tool that guarantees perfection. The parties and Judge Peck have acknowledged that there are risks inherent in any method of reviewing electronic documents. Manual review with keyword searches is costly, though appropriate in certain situations. However, even 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 the various attorneys' determination of whether a document is responsive. Judge Peck concluded that under the circumstances of this particular case, the use of the predictive coding software as specified in the ESI protocol is more appropriate than keyword searching. The Court does not find a basis to hold that his conclusion is clearly erroneous or contrary to law.

Id.

156. *See Da Silva Moore II*, 2012 WL 607412, at *12. Although not mentioned by the *Da Silva Moore* court, there is also likely a strong presumption in favor of the collection method chosen by the responding party. *See infra* Part V.

157. Judge Peck made pretty much the same observations in his opinion and order dated February 24, 2012:

> The objective of review in ediscovery is to identify as many relevant documents as possible, while reviewing as few non-relevant documents as possible. Recall is the

Yet the court's order, while perhaps justifying the use of predictive coding in general, in no way justifies the particular predictive coding protocol adopted in the *Da Silva Moore* case. Many of plaintiffs' objections to the predictive coding protocol could be read not as an attack on predictive coding generally but as an argument that the particular method of predictive coding defendants sought to utilize was not the optimal search method.[158] Therefore, the relevant comparison was not between predictive coding and other document retrieval methods but between defendants' conception of predictive coding and a more stringent version advocated by plaintiffs that would perhaps be more costly but potentially yield more satisfactory results.[159] In essence, Judge Peck's response to that argument was to postpone its consideration to a later stage in the case.[160]

Indeed, some of the most interesting and controversial parts of Judge Peck's ruling are the decisions he makes with regard to the phasing of various aspects of the discovery process. The basic approach is to "code first and evaluate later."[161] Judge Peck takes a fairly deferential approach to defendants' plans to utilize predictive coding while making it clear that he is not necessarily endorsing the protocol as a correct or complete response to plaintiffs' e-discovery demands.[162] For example, defendants made various decisions in their predictive coding protocol which they frankly admitted were based largely on the desire to limit expense, including the decisions to use a 95% confidence level[163]—which reduced the size of the initial random sample set relative to a higher confidence level—and to run the predictive coding software process through seven iterations.[164] While accepting those features as part of the predictive coding

---

> fraction of relevant documents identified during a review; precision is the fraction of identified documents that are relevant. Thus, recall is a measure of completeness, while precision is a measure of accuracy or correctness. The goal is for the review method to result in higher recall and higher precision than another review method, at a cost proportionate to the "value" of the case.
>
> The slightly more difficult case would be where the producing party wants to use computer-assisted review and the requesting party objects. The question to ask in that situation is what methodology would the requesting party suggest instead? Linear manual review is simply too expensive where, as here, there are over three million emails to review. Moreover, while some lawyers still consider manual review to be the "gold standard," that is a myth, as statistics clearly show that computerized searches are at least as accurate, if not more so, than manual review.

Da Silva Moore II, 2012 WL 607412, at *9 (citing Grossman & Cormack, *supra* note 9, at 8–9).

158. *See, e.g.*, *id.* at *3 ("[The plaintiffs] expressed multiple concerns to defense counsel on the way in which they plan[ned] to employ predictive coding." (quoting Transcript of Dec. 2, 2011 Conference, *supra* note 119, at 21)).

159. *See id.* at *8.

160. *See id.* (stating that plaintiffs' concerns about the accuracy of the results are premature).

161. *See id.*

162. *See id.*

163. *See id.* at *5 (citing Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 58–60).

164. *See id.* at *6 ("The idea is to make it significantly better than the alternatives without nearly as much cost." (quoting Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 76)). Also

protocol, Judge Peck made it clear that his opinion was not a finding that the search protocol complied with the federal discovery rules.[165] To some extent, he was telling defendants they were acting at their peril. If he subsequently found that the results of the predictive-coding-based search were unsatisfactory, he could throw it out completely and require a new search utilizing a different methodology.[166]

Plaintiffs, in contrast, unsuccessfully argued for the establishment of accuracy criteria prior to or simultaneously with the adoption of the protocol.[167] Their papers in opposition to the adoption of the protocol included a declaration by their e-discovery expert which set forth the basic criteria for judging the validity of a predictive coding system: recall (a comparison of the number of documents the system predicts will be responsive compared to the actual number of responsive documents), precision (the percentage of actually responsive documents to the documents retrieved by the system as responsive), and F-measure (a method of combining the measured recall and precision of the system into a single numerical measure of accuracy).[168] The expert argued that industry best practices and the recommendations of other experts, such as those in The Sedona Conference, required that predictive coding systems be measured against

---

consider the following exchange between the plaintiffs' ESI expert and defense counsel, which took place at the February 8 hearing:

> [PLAINTIFFS' EXPERT]: [Recommind's] patent itself suggests that as a result of this process you should be reviewing 10 to 35 percent of your total document collection, which is supposed to indicate a significant savings, which in this case would be about 300 [thousand] to 1 million documents. They keep talking about 40,000 to 75 [thousand] as being burdensome and disproportional. If they don't understand the result of the system, what to expect, I don't understand why they are proposing it in the first place.
>
> [MSL COUNSEL]: Your Honor, one of the reasons why we developed this work flow was, again, this is not a case where we are prepared to review a million documents during this first phase. We worked with our vendor and came up with a modified work flow that we believe is defensible but is also reviewing a more reasonable number of documents for this case.
>
> THE COURT: We'll see. Make sure you're keeping track of your costs in ways that you will be able on both sides to present to the Court not for reimbursement but for proportionality as to where you draw the line. I'm not saying that there is a dollar number that I'm going to cut it off at or a percentage or where the cliff is. We are going to figure all that out. All of this, obviously at some expense, can be revisited if things are not working well.

Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 87–88.

165. *See Da Silva Moore II*, 2012 WL 607412, at *7 (citing FED. R. EVID. 702; Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993)) (stating that the Rule 702 of the Federal Rules of Evidence and the *Daubert* decision are not applicable to discovery search methods).

166. *See id.* at *8 (stating that reliability issues could be revisited "down the road").

167. *See id.* (citing Plaintiffs' Rule 72(a) Objection to the Magistrate's Feb. 8, 2012 Discovery Rulings at 13–18, *Da Silva Moore II*, 2012 WL 607412 (No. 11-CV-1279(ALC)(AJP)), ECF No. 93).

168. *See* Declaration of Paul J. Neale in Support of Plaintiffs' Rule 72(a) Objections to Magistrate Judge Peck's Feb. 8, 2012 Discovery Rulings, *supra* note 90, at 3–4.

such clear numerical standards of accuracy.[169] However, he was not able to cite any authorities which held that such accuracy measures had to be specified prior to the running of the software.[170]

Defendants, in contrast, took the position that assessing the accuracy of the predictive coding protocol after document retrieval had taken place was the only approach that made any sense.[171] Their protocol referred, somewhat vaguely, to the use of "judgmental and statistical sampling" during the review process to assess the accuracy of the results obtained.[172] Judge Peck clearly agreed on the timing issue.[173]

Plaintiffs made similar arguments about the failure of the protocol to set forth an "agreed-upon standard of relevance that is transparent and accessible to all parties."[174] Plaintiffs argued that "[w]ithout this standard, there [is] a high-likelihood of delay as the parties resolve disputes with regard to individual documents on a case-by-case basis" and that it would be "impossible for a third party investigator or auditor to replicate the results, should that be necessary."[175] Again, Judge Peck expressly rejected these arguments regarding timing.[176]

---

169. *See id.* at 5 (quoting The Sedona Conference, *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process*, 10 SEDONA CONF. J. 299, 320 (2009)).

170. *See id.* at 12. Plaintiffs' expert did, however, make an argument along those lines as he stated:

> (C) Defendants' Protocol Fails to Specify a Standard of Acceptance.
>
> 63. Finally, an adequate protocol will specify, in advance, a standard of acceptance. Whether that standard is expressed in terms of absolute levels of recall and precision or in terms of relative levels of recall and precision (e.g., the level that could be expected by a viable alternative approach) is a decision that should be made by the parties in advance.
>
> 64. A protocol that fails to specify a standard of acceptance in advance sets the stage for further disputes and acrimony. As a result, the implications of any measures of accuracy will remain open to interpretation and dispute.

Id.

171. *See Da Silva Moore II*, 2012 WL 607412, at *6, *21 (citing Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 74–75).

172. *Id.* at *18.

173. *See id.* at *12. As he stated in his opinion:

> [I]t is unlikely that courts will be able to determine or approve a party's proposal as to when review and production can stop until the computer-assisted review software has been trained and the results are quality control verified. Only at that point can the parties and the Court see where there is a clear drop off from highly relevant to marginally relevant to not likely to be relevant documents. While cost is a factor under Rule 26(b)(2)(C), it cannot be considered in isolation from the results of the predictive coding process and the amount at issue in the litigation.

Id.

174. *Id.* at *8.

175. Plaintiffs' Rule 72(a) Objection to the Magistrate's Feb. 8, 2012 Discovery Rulings, *supra* note 167, at 14.

176. *See Da Silva Moore II*, 2012 WL 607412, at *8. Judge Peck stated:

> Relevance is determined by plaintiffs' document demands. As statistics show, perhaps only 5% of the disagreement among reviewers comes from close questions of relevance, as opposed to reviewer error. The issue regarding relevance standards might be significant if MSL's proposal was not totally transparent. Here, however, plaintiffs will

A third issue, although one not much emphasized by plaintiffs, was the court's apparent willingness to base its decisions regarding discovery phasing on whether specific categories of information were well adapted to the use of predictive coding.[177] Judge Peck declined to include the files of plaintiffs' proposed "comparators" in the phase one documents to be searched and similarly excluded the files of MSL CEO Fleurot on the grounds that they were not well suited to a predictive-coding search.[178] This framework might appear to be putting the cart before the horse. After all, the search method should be adapted to the type of documents sought rather than the other way around. However, Judge Peck's opinion does provide a rationale (of sorts) for excluding such documents—at least in the initial phase of discovery. He states:

> [S]taging of discovery by starting with the most likely to be relevant sources (including custodians), without prejudice to the requesting party seeking more after conclusion of that first stage review, is a way to control discovery costs. If staging requires a longer discovery period, most judges should be willing to grant such an extension.[179]

Along these lines, defenders of the *Da Silva Moore* protocol could argue that it is designed to get the largest number of potentially relevant documents reviewed, in the most accurate manner, and at the lowest possible cost. If the protocol results in some potentially highly relevant documents being relegated to a second phase of discovery, it seems a small price to pay. The danger to plaintiffs of course is that the court, assuming that phase two contains primarily the more costly, less relevant documents, either cuts off discovery at phase one or imposes costs on plaintiffs in phase two.

B. Kleen Products LLC v. Packaging Corp. of America

Kleen Products LLC v. Packaging Corp. of America[180] is a case involving a Sherman Act antitrust class action related to the containerboard industry.[181] The e-discovery dispute in *Kleen Products* arose when the plaintiffs asked Judge Nan Nolan of the United States District Court for the Northern District of Illinois to

---

> see how MSL has coded every email used in the seed set (both relevant and not relevant), and the Court is available to quickly resolve any issues.

Id.

177. *See id.* at *4 (noting that the "search of the comparators' emails would be so different from that of the other custodians that the comparators should not be included in the emails subjected to predictive coding review" (citing Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 28, 30)).

178. *Id.* (citing Transcript of Feb. 8, 2012 Conference, *supra* note 126, at 31).

179. *Id.* at *12.

180. No. 10 C 5711, 2012 WL 4498465 (N.D. Ill. Sept. 28, 2012), *objections overruled by* No. 10 C 5711, 2013 WL 120240 (N.D. Ill. Jan. 9, 2013).

181. *Id.* at *1–2.

order defendants to use an alternative technology—specifically predictive coding—to produce documents, even though defendants had already produced a "significant amount of responsive information."[182] This "significant amount of responsive information" included over a million documents, and the defendants "had already spent thousands of hours reviewing and producing" these documents using keyword searches and related review techniques.[183] Plaintiffs argued that if the "[d]efendants had used '[c]ontent [b]ased [a]dvanced [a]nalytics'" tools—a term not defined by the plaintiffs—"then [defendants'] production would have been more thorough."[184] However, plaintiffs did not identify specific examples of documents missing from the defendants' production.[185] In opposition to the plaintiffs' request, the defendants argued that their use of keyword searches was sufficiently thorough.[186]

Judge Nolan heard opinions from expert witnesses regarding the adequacy of the initial production in two different hearings[187] and then asked the parties to try to reach a compromise on the keyword search approach.[188] She reasoned that a "mutually agreeable approach based on what [d]efendants had already implemented was preferable to" undertaking an entirely new mode of analysis.[189] Following the second hearing on the issue, held March 28, 2012, Judge Nolan stated: "the defendants had done a lot of work, the defendant under Sedona 6 has the right to pick the [discovery production] method. Now, we all know, every court in the country has used Boolean search, I mean, this is not like some freak thing that [the defendants] picked out . . ."[190]

---

182. *Id.* at *4 (citing Plaintiffs' Statement of Position with Respect to Disputed Items for Dec. 15, 2011 Status Conference at 4, 5, 8, *Kleen Products*, 2012 WL 4498465 (No. 1:10-cv-05711), ECF No. 266 [hereinafter Plaintiffs' Statement]) (criticizing defendants' Boolean search method and asking them to use "content-based advanced analytics").

183. *See id.* (citing Defendants' Statement of Position with Respect to Dispute Items for Dec. 15, 2011 Status Conference at 3, 8–10, *Kleen Products*, 2012 WL 4498465 (No. 1:10-cv-05711), ECF No. 267 [hereinafter Defendants' Statement]); Matthew Nelson, Kleen Products *Predictive Coding Update—Judge Nolan: "I Am a Believer of Principle 6 of Sedona,"* E-DISCOVERY 2.0 (June 5, 2012, 11:36 AM), http://www.clearwellsystems.com/e-discovery-blog/2012/06/05/kleen-products-ediscovery-predictive-coding-update-judge-nolan-i-am-a-believer-of-principle-6-of-sedona/.

184. Nelson, *supra* note 183.

185. *Id.*; *see also Kleen Products*, 2012 WL 4498465, at *4 (stating that plaintiffs' argument was simply that the defendants' process was "subject to . . . inadequacies" (quoting Plaintiffs' Statement, *supra* note 182, at 8)).

186. Nelson, *supra* note 183; *Kleen Products*, 2012 WL 4498465, at *4 (quoting Defendants' Statement, *supra* note 183, at 3).

187. Nelson, *supra* note 183; Transcript of Feb. 21, 2012 Proceedings Before the Honorable Magistrate Judge Nan R. Nolan, *Kleen Products*, 2012 WL 4498465 (No. 10 C 5711), ECF No. 304 (first hearing).

188. Nelson, *supra* note 183; Transcript of Mar. 28, 2012 Proceedings—Evidentiary Hearing Before the Honorable Magistrate Judge Nan R. Nolan Volume 2-A at 300, *Kleen Products*, 2012 WL 4498465 (No. 10 C 5711), ECF No. 319-1 (second hearing).

189. Nelson, *supra* note 183; Transcript of Mar. 28, 2012 Proceedings—Evidentiary Hearing Before the Honorable Magistrate Judge Nan R. Nolan Volume 2-A, *supra* note 188, at 299.

190. Nelson, *supra* note 183; Transcript of Apr. 2, 2012 Proceedings Before the Honorable Nan Nolan at 12, *Kleen Products*, 2012 WL 4498465 (No. 10 C 5711), ECF No. 319-1.

In so stating, Judge Nolan confirmed the presumption that responding parties generally may choose the method by which they produce documents. Quality and accuracy, according to Judge Nolan, are the key guideposts for assessing production, not the technology or process used.[191]

While there is some indication that the outcome in *Kleen Products* is largely a result of the perceived amount of time and money the defendants had already invested in document production, that assumption is not entirely correct; even though the search process was well underway, production had not yet occurred.[192] While some may read *Kleen Products* as leaving open the possibility of a different outcome if sunk costs were not involved and if the requesting party had sought the use of predictive coding at the outset, that misses the heart of the issue. Rather, *Kleen Products* largely turned on the Sedona 6 presumption, and the court's decision not to order predictive coding should be viewed as a failure of the objecting plaintiffs to show the inadequacy of the search and production, not as a success of the sunk cost defense.

C. Global Aerospace v. Landow Aviation

Global Aerospace Inc. v. Landow Aviation, L.P.[193] involved the collapse of three hangars at Dulles airport during a snowstorm in 2010.[194] A series of lawsuits were filed in the wake of that collapse, and the actions were consolidated in Virginia Circuit Court.[195]

In response to discovery requests, the defendant, Landow Aviation, indicated that it would use predictive coding to retrieve potentially relevant documents from a large electronic document population.[196] A number of the plaintiffs objected, and on April 9, 2012, Landow Aviation moved the court for a protective order authorizing its use of predictive coding.[197] Landow argued that a single manual review of the documents would cost over $2 million and would locate, at most, 60% of the responsive documents.[198] Similarly, it argued that keyword searching would be ineffective, producing possibly 20% of the responsive documents.[199] Citing a series of studies, Landow argued that

---

191. Transcript of Apr. 2, 2012 Proceedings Before the Honorable Nan Nolan, *supra* note 190, at 13.

192. *See* Defendants' Statement, *supra* note 183, at 3.

193. Order Approving the Use of Predictive Coding for Discovery, No. CL 61040, 2012 WL 1431215, at 1 (Va. Cir. Ct. Apr. 23, 2012).

194. Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding at 3, Global Aerospace Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012), 2012 WL 1419842 [hereinafter Landow Aviation Memorandum in Support].

195. *See* Order Approving the Use of Predictive Coding for Discovery, *supra* note 193, at 1.

196. *See* Landow Aviation Memorandum in Support, *supra* note 194, at 1–2.

197. *Id.* at 1.

198. *Id.* at 2.

199. *Id.*

predictive coding could locate up to 75% of potentially relevant documents "at a fraction of the cost and in a fraction of the time of linear review and keyword searching."[200]

Landow set forth in its motion a proposed predictive coding protocol.[201] Landow stated that it would first produce a full set of training documents—the seed set—to opposing counsel after enough iterations had occurred so that the set was reliable but before it was used to classify the complete document population.[202] Next, Landow would create a log for privileged and irrelevant, sensitive documents so that opposing counsel could determine whether there needed to be a "review [of] the documents to evaluate the coding decision and whether the coding decision appears to be correct."[203] After predictive coding had been completed and documents had been categorized, Landow proposed that it would "implement a statistically valid sampling program to establish that the majority of the relevant documents [had] been retrieved."[204] Landow proposed that an acceptable retrieval, or recall, rate was 75%—"predictive coding [would] conclude once the sampling program establishes that at least 75% of the relevant documents [had] been retrieved from the [document population]."[205]

The plaintiffs objected to this proposal, first informally to Landow and then in an opposition memorandum filed with the court.[206] They argued that "[t]here are no grounds justifying [a] departure" from the traditional approach to document production.[207] While the plaintiffs conceded that computer-assisted technologies can make the document review process more efficient, they argued that it should be used to "supplement" traditional manual review, not replace it.[208]

In a short order, Judge Chamblin approved the defendants' use of predictive coding "for purposes of the processing and production of electronically stored information."[209] But, like *Da Silva Moore*, Judge Chamblin left room for the

---

200. *Id.*; *see also id.* at 9–11 (citing Grossman & Cormack, *supra* note 9, at 22, 37 tbl.7; The 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, 218 (2007)).

201. *Id.* at 11–13.

202. *Id.* at 11.

203. *Id.*

204. *Id.* at 12.

205. *Id.*

206. *See* Opposition of Plaintiffs: M.I.C. Industries, Inc., Factory Mutual Insurance Co., Global Aerospace, Inc., and BAE Systems Survivability Systems, LLC to the Landow Defendants' Motion for Protective Order Regarding Electronic Documents and "Predictive Coding," Global Aerospace Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012), 2012 WL 1419848.

207. *Id.* at 2.

208. *Id.* at 4.

209. Order Approving the Use of Predictive Coding for Discovery, *supra* note 193, at 1.

plaintiffs subsequently to question the "completeness or the contents of the production or the ongoing use of predictive coding."[210]

This case is one of the first to raise the important question of when, if at all, a court may order the use of a particular technique for the production of documents. Relying heavily on the Sedona Principles, the court held that deference was due to the method selected by the responding party.[211] But the court's decision leaves open potentially different outcomes if, for example, the requesting, rather than the responding, party sought predictive coding.

D. In re Actos (Pioglitazone) Products Liability Litigation

In re Actos (Pioglitazone) Products Liability Litigation[212] is "a multidistrict litigation consolidating [eleven] civil actions for pretrial proceedings."[213] In this case, "[t]he plaintiffs allege[d] that Actos, a prescription drug for the treatment of diabetes type 2, increases a users' risk of developing bladder cancer," and that "defendants concealed and failed to adequately warn consumers about" these risks.[214]

"On July 27, 2012, United States Magistrate Judge Hanna Doherty of the Western District of Louisiana entered a Case Management Order outlining the electronically stored information (ESI) protocol the parties must follow during discovery."[215] The order lays out the detailed predictive coding protocol agreed upon by the parties.[216]

---

210. *Id.*

211. *See id.*; *see also* The Sedona Conference, *supra* note 200, at 204 ("Absence agreement, a party has the presumption, under Sedona Principle 6, that it is in the best position to choose an appropriate method of searching and culling data.").

212. Case Management Order: Protocol Relating to the Production of Electronically Stored Information ("ESI"), No. 6-11-md-2299, 2012 WL 3899669 (W.D. La. July 30, 2012).

213. Michael Roach, *Predictive Coding Watch: 'In re: Actos,'* EDD UPDATE (Aug. 17, 2012, 7:34 PM), http://www.eddupdate.com/2012/08/predictive-coding-watch-in-re-actos.html.

214. *Id.*

215. Matthew Nelson, In Re: Actos—*Does New Federal Litigation Clarify Predictive Coding in eDiscovery?*, FORBES (Aug. 27, 2012, 10:23 AM), http://www.forbes.com/sites/benkerschberg/2012/08/27/in-re-actos-does-new-federal-litigation-clarify-predictive-coding-in-ediscovery/ (citing Case Management Order, *supra* note 212, at 1).

216. Case Management Order, *supra* note 212, at 6–16. Specifically, one of the defendants, Epiq, "will collect [emails] from four key Takeda custodians, which will be combined to create the 'sample collection population.'" *Id.* at 7. "The [p]arties will meet and confer" as to which custodians' e-mail will be used. *Id.* Takeda will also "add a set of regulatory documents which have already been collected to the 'sample collection population.'" *Id.* This sample will then be used to build a control set to train the predictive coding system. *Id.* "At the conclusion of the training process and upon calculation of relevance scores, the [p]arties will [again] meet and confer regarding which relevance score will provide a cutoff for documents to be manually reviewed." *Id.* at 8–9. "In addition, . . . the [p]arties will collaboratively review . . . a random sample of documents in the sample collection population will relevance scores below the cut-off score set for establishing the review set." *Id.* at 14. The purpose of this step is to verify that the document population with

In *In re Actos* the parties appear to be in agreement as to the software's use, and to date, no discovery conflict has emerged.[217] Yet, it is an important decision because it is one of the most detailed orders to date about how the predictive coding process will actually occur and the steps that will be used to implement the process.

E. EORHB, Inc. v. HOA Holdings LLC

EORHB, Inc. v. HOA Holdings LLC,[218] or the "Hooters Case," is a commercial indemnity dispute concerning the sale of the Hooters restaurant chain.[219] On October 15, 2012, Vice Chancellor J. Travis Laster heard oral arguments on a motion for partial summary judgment and a motion to dismiss.[220] Vice Chancellor Laster went on to deny the motion to dismiss and plaintiff's motion for partial summary judgment,[221] but at the end of the hearing, he ordered the parties to use predictive coding to manage discovery in the case.[222] In so ordering, he stated:

> This seems to me to be an ideal non-expedited case in which the parties would benefit from using predictive coding. I would like you all, if you do not want to use predictive coding, to show cause why this is not a case where predictive coding is the way to go.
>
> I would like you all to talk about a single discovery provider that could be used to warehouse both sides' documents to be your single vendor. Pick one of these wonderful discovery super powers that is able to maintain the integrity of both side's documents and insure that no one can access the other side's information. If you cannot agree on a suitable discovery vendor, you can submit names to me and I will pick one for you.
>
> . . . The problem is that these types of indemnification claims can generate a huge amount of documents. That's why I would really encourage you all, instead of burning lots of hours with people reviewing, it seems to me this is the type of non-expedited case where we could all benefit from some new technology use.[223]

---

low relevancy scores does not contain a prevalence of relevant documents, and that "the proportionality assumptions underlying the cut-off decision are valid." *Id.*

217. *See id.* at 1.

218. Transcript of Motion for Partial Summary Judgment, Motion to Dismiss Counterclaim, and Ruling of the Court, EORHB, Inc. v. HOA Holdings LLC, No. 7409-VCL, 2012 WL 4896670 (Del. Ch. Oct. 15, 2012) (No. 7409-VCL).

219. *See id.* at 4.

220. *Id.* at 1, 4.

221. *See id.* at 45, 65.

222. *Id.* at 66.

223. *Id.* at 66–67.

Vice Chancellor Laster issued an order after the hearing indicating as much; he ordered:

> The parties shall confer regarding a case schedule. Absent a modification of this order for good cause shown, the parties shall (i) retain a single discovery vendor to be used by both sides, and (ii) conduct document review with the assistance of predictive coding. If the parties cannot agree on a single discovery vendor with expertise in predictive coding, the parties shall each submit up to two vendor candidates to the Court.[224]

Vice Chancellor Laster's order is the first instance in which a court has ordered sua sponte the use of predictive coding software. Moreover, it is a case in which both parties were likely to have substantial e-discovery obligations and both could, therefore, benefit relatively equally from the cost and accuracy advantages of predictive coding.[225]

F. The Unresolved Questions in the Case Law

With this context, we turn to the many issues that the decisions have not yet considered but that must be resolved if predictive coding is to become a standard method of e-discovery. First and foremost is the question of the appropriate application of proportionality review to predictive coding. None of the aforementioned judges evaluated the proposed protocols against the standards of proportionality review.[226] Such questions, if mentioned at all, were deferred by the courts to a later phase of discovery.[227] But every discovery protocol to some extent implicitly measured the costs and burdens of discovery against its proposed benefits. For example, the choice of the defendants in *Da Silva Moore* to utilize a 95% confidence level rather than a more stringent confidence interval—99%—which kept the randomly selected portion of the seed set smaller and therefore easier and cheaper for senior lawyers to review, involves

---

224. Order Granting Partial Summary Judgment, No. 7409-VCL, 2012 WL 4896670 (Del. Ch. Oct. 18, 2012).

225. Seizing on this decision, one commentator has argued that it may become an ethical obligation to use predictive coding. *See* Howard Sklar, *Legal Acceptance of Predictive Coding: A Journey in Three Parts*, RECOMMIND (Nov. 6, 2012), http://blog.recommind.com/legal-acceptance-of-predictive-coding-a-journey-in-three-parts/ ("In the future, we'll enter stage four: the decision by a state bar's ethics watchdog that failure to use predictive coding is ethically questionable, if not unethical. After all, purposefully using a less-efficient, less accurate, more expensive option is problematic. I think that's probably 18 months away. But given how fast we've gone through the first three states, stage four may come next week.").

226. *See supra* Part III.A–E.

227. *See, e.g.*, *supra* text accompanying note 140 (deferring questions related to proportionality review to "down the road").

such an implicit trade-off.[228] In short, if predictive coding allows a court to either order the same level of e-discovery at a much lower cost or order a fuller, more complete and accurate process of e-discovery at the same cost, which should the court order? This and related questions are the subject of Part IV.

Lastly, the orders in *Kleen Products*, *Global Aerospace*, and *EORHB* raise questions as to the extent to which a court may mandate the use of predictive coding. In *Kleen Products* and *Global Aerospace*, one party sought an order from the court allowing or requiring predictive coding over the objection of another party.[229] In *EORHB*, the Delaware Chancery Court required predictive coding sua sponte.[230] So, may a court force a party to agree to predictive coding? Does it matter if it is the party that is responding to the document request and seeking authorization for its use of predictive coding (as in *Global Aerospace*) or if it is the requesting party attempting to force the responding party to use the software (as in *Kleen Products*)? To what extent may a court order predictive coding when it believes that the software is particularly appropriate given the economics of the case, regardless of what positions the parties take? What would happen if both parties objected to such a sua sponte order? These questions are considered in Part V.

IV. PREDICTIVE CODING AND PROPORTIONALITY REVIEW

A. The Big Picture: Who Gets the Benefits of Predictive Coding?

Assume that the plaintiff in a large civil litigation served an initial demand for discovery of electronically stored information on the defendant. After substantial negotiation, the parties agreed to a discovery plan which provided for a certain number of files to be searched using specified keyword protocols to locate as many responsive documents as possible to the initial demand. Assume that the defendant then comes to court with an expert in computerized document retrieval systems who testifies that utilizing predictive coding will permit the

---

228. For instance, in *Da Silva Moore*, while plaintiffs apparently acquiesced in the 95% confidence level initially—at least with regard to determining the size of the seed set—in his Declaration opposing the protocol, plaintiffs' ESI expert argued that the sample derived from that confidence level, while adequate to determine the "yield" of the system, would be too small to provide an accurate measure of the "recall" of the system. *See* Declaration of Paul J. Neale in Support of Plaintiffs' Rule 72(a) Objections to Magistrate Judge Peck's Feb. 8, 2012 Discovery Rulings, *supra* note 90, at 9. As he stated, "using the sample size for the yield estimate in this case to also determine the recall estimate will result in an unacceptably large confidence interval as detailed below which would possibly result in as much as *60% of the responsive documents not being produced*."
Id. In his declaration, he included an example based on the recognized difficulty of accurately determining recall from a small sample size when the yield of the system is low (1.5% in his example). *Id.* Also, it was not clear from the *Da Silva Moore* protocol that defendants proposed to use the same randomly selected seed set to estimate both yield and recall. *See id.* at 11.

229. *See supra* text accompanying notes 182–89, 206–09.

230. *See supra* text accompanying notes 222–24.

same files to be searched for the same information at a lower cost and with greater accuracy. One might view this as an easy case in which a court should order predictive coding to lower discovery costs.

Suppose, however, the plaintiff argues that the initial discovery plan was based on the application of the principles of proportionality review to the issues in the case and represents an agreed-upon compromise between the costs of the keyword-search-based discovery and the needs of the plaintiff for the information. Therefore, the plaintiff argues that if the costs of that discovery are now reduced due to the use of predictive coding, the appropriate result should not be lower costs for the same discovery but rather increased discovery of additional files or changes in the predictive coding protocols to make the search more accurate and complete.

This argument is an issue that is generally not considered in the writing on predictive coding, but it seems to us to be the most fundamental legal, and even philosophical, issue raised by the new technology. Because it remains the case that the responding party bears the cost of discovery absent a court order shifting those costs,[231] it might be assumed that the cost benefits of predictive coding will automatically fall to the responding party in the form of lower discovery costs. However, a moment's reflection on the actual process of discovery, as currently conducted under the proportionality standard, reveals that the requesting party also bears some of the cost of discovery in the form of limitations on the information that can be requested when the additional costs of such information outweigh its benefits.[232] The cost and accuracy improvements represented by predictive coding, therefore, can be used to the advantage of either the responding or requesting party or can be divided in some way between them.

In essence, this distribution of the "efficiency" benefits resulting from the use of predictive coding is a new version of a problem quite familiar to economists—the problem of bilateral monopoly.[233] Both the requesting and

---

231. Oppenheimer Fund, Inc. v. Sanders, 437 U.S. 340, 358 (1978).

232. *See* FED. R. CIV. P. 26(b)(2)(C)(iii).

233. *See generally* Tom Campbell, *Bilateral Monopoly in Mergers*, 74 ANTITRUST L.J. 521 (2007) (arguing that mergers leading to monopoly could be socially desirable in certain circumstances); *see also* Richard D. Friedman, *Antitrust Analysis and Bilateral Monopoly*, 1986 WIS. L. REV. 873, 873–74 (citing Daniel Druckman & Thomas V. Bonoma, *Determinants of Bargaining Behavior in a Bilateral Monopoly Situation II: Opponent's Concession Rate and Similarity*, 21 BEHAV. SCI. 252, 252 (1976)) (discussing the definition of a bilateral monopoly). For example, assume that *A* and *B* have a contract under which *B* agrees to purchase all of *A*'s output of widgets for a fixed price of $10 per widget. Under the contract *B* also has the right to veto any material changes *A* makes in the manufacturing process. Assume that *A*, which has previously been manufacturing widgets at a cost of $8, finds a new technological process which will enable the same widgets to be produced for $6. If *A* makes the change, it will increase its profits by 100%, but *B* will get no benefit and has no incentive to agree to change. Alternatively, if *B* demands that the contract price be lowered to $8 before approving the change, A has no incentive to adopt new process. The obvious solution is for both parties to share the benefits of the new technology. However, from an efficiency perspective, there is no single correct way to split the benefit. Rather,

responding parties can potentially benefit from the technological superiority of predictive coding. Because they primarily would benefit in different ways, however—the responding party through decreased costs and the requesting party through increased, more accurate discovery—any increase in benefits to one decreases the benefits to the other, and there is no clearly "right" or most efficient way to allocate this benefit between the two parties.

It should be noted that this is not a serious problem in cases where both the plaintiff and defendant possess significant amounts of electronically stored information and both parties will, therefore, be requesting and responding to e-discovery requests. In such cases, both parties will benefit fairly equally from whatever determination a court makes in allocating the benefits of predictive coding. Such considerations may well underlie Vice Chancellor Laster's comment in the *EORHB* case: because such commercial indemnification cases "can generate a huge amount of documents" (presumably from both sides of the dispute), it is in precisely such cases that "we could all benefit from some new technology use."[234]

The real problem arises in those cases involving what we have previously labeled "asymmetric discovery,"[235] where one party, usually the defendant, has almost all of the relevant information and is, therefore, almost invariably the responding party.[236] Here the problem is presented in its purest form: should predictive coding be used to decrease the discovery costs of the responding party or to increase the scope and/or accuracy of the search available to the requesting party?

Obviously, we have no answer to this question, and the extensive work done by economists on the subject of bilateral monopolies strongly indicates that there is no correct answer to such a question.[237] Nonetheless, we have three observations that we think are relevant to this problem.

First, although we have presented the application of proportionality review to predictive coding as a broad legal, philosophical, and economic question, it will never appear that way to judges or lawyers engaged in actual discovery disputes. Rather, allocation of the benefits of predictive coding will take place through dozens of smaller decisions, many of a technical nature, such as: determining whether to use a 95% or 99% confidence level,[238] whether to order discovery of all documents having the responsiveness above a certain level,[239]

---

the actual result reached likely will be the result of the preexisting legal rules and the bargaining power of the parties involved.

234. Transcript of Motion for Partial Summary Judgment, Motion to Dismiss Counterclaim, and Ruling of the Court, *supra* note 218, at 67.

235. Yablon & Landsman-Roos, *supra* note 85, at 723, 730.

236. *See id.* at 726 (defining "asymmetric litigation").

237. *See* sources cited *supra* note 233.

238. *See supra* Part II.B.4.

239. *See supra* Part II.B.3.

how many iterations of the predictive coding software to run,[240] and what level of accuracy to require and how many files to review.[241] While all of these decisions may seem to be about discrete, technical subjects, it is important for decisionmakers to remember that each of these decisions fundamentally allocates the benefits of predictive coding between requesting and responding parties. Accordingly, these decisions should be decided on legal and equitable grounds and not purely technical ones.

The second observation comes from economic theories on bilateral monopolies that show that while there is no correct allocation of the bilateral benefit, there is one clearly incorrect allocation—giving it all to one party or the other.[242] In the context of predictive coding, such an all or nothing allocation would likely take the form of giving all the benefits to the responding party in the form of cost reduction, although the entire allocation could, in theory, be used to benefit the requesting party through the increased scope and accuracy of discovery as well. Both are bad solutions as they destroy the incentive of either party to agree to predictive coding and to seek to utilize it in the most effective manner.[243] Although, as Part V indicates, there may well be times when the court has the power to order predictive coding over the objections of one or even both parties, the preferred scenario, as in most discovery disputes, is to provide incentives for the parties to reach agreement on such protocols.[244] Allocating some portion of the benefit of predictive coding to both parties is the easiest way to create incentives for such cooperation.

Finally, as to precisely how those benefits should be allocated, we can only say that, like most discovery disputes, it should be determined on a case-by-case basis and be based on the cost–benefit analysis implicit in the rules governing proportionality review.[245] We have previously argued, and restate our position

---

240. For example, the *Da Silva Moore* protocol, while providing for seven iterations, also gave defendants discretion to utilize fewer iterations of review if "the change in the total number of relevant documents predicted by the system as a result of a new iteration, as compared to the last iteration, is less than five percent (5%), and no new documents are found that are predicted to be *hot* (aka *highly relevant*)." *Da Silva Moore II*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *20 (S.D.N.Y. Feb. 24, 2012). Here again, it would seem that the use of the 5% number is somewhat arbitrary.

241. *See supra* Part II.B.1, II.B.5.

242. *See generally* Friedman, *supra* note 233 (discussing various bilateral monopoly models).

243. *See, e.g.*, *id.* at 875 (stating that the "bargaining model" results in a Pareto equilibrium—the situation where "no other result can be better for one of the [parties] without being worse for the other").

244. In addressing the discovery dispute in *Kleen Products*, the court praised the parties for their efforts to cooperate. *See* Kleen Products LLC v. Packaging Corp. of Am., No. 10 C 5711, 2012 WL 4498465, at *1, *19 (N.D. Ill. Sept. 28, 2012). In that case, Judge Nolan advised: "Cooperation does not conflict with the advancement of their clients' interests—it enhances it. Only when lawyers confuse *advocacy with adversarial conduct* are these twin duties in conflict. . . . [T]his is a story as much about cooperation as dispute." *Id.* at *1. In concluding, the court "commend[ed] the lawyers and their clients for conducting their discovery obligations in a collaborative manner." *Id.* at *19.

245. *See* FED. R. CIV. P. 26(b)(2)(C)(iii), 34(a).

here, that such proportionality review can and should include a judge's preliminary consideration on the merits of the case.[246] That is, in weighing the "burden or expense" of any discovery request—or request for a particular technical parameter in a predictive coding protocol—against the "likely benefit" of such request—or parameter—pursuant to Rule 26(b)(2)(C)(iii) of the Federal Rules of Civil Procedure,[247] the court must apply some preliminary view as to the "benefits" to be obtained by such additional, broader, more accurate discovery. Such a view will frequently, but not always, be colored by the court's current view of the merits of the case, particularly its view as to the likelihood of the existence of documents or ESI that will substantiate plaintiff's claims.

We see nothing wrong with such prejudging in appropriate and limited circumstances and have argued that it is far better for courts to resolve discovery disputes informed by such considerations than to do so on purely "managerial" grounds.[248] We have argued for the use of certain techniques, such as sampling, which can enhance the accuracy of merits-based resolution of discovery disputes in some cases.[249] We believe that decisions regarding predictive coding protocols should also be informed by these merits-based considerations. The strong tendency that we see in the existing predictive coding cases is for courts to try to delay making definitive rulings on accuracy and similar parameters until they have reviewed preliminary results from the discovery software. This delay indicates to us that the courts are aware of the dangers of making these decisions during the early stages of the litigation and are seeking additional information from the discovery process itself, in part to make a more informed judgment about the underlying merits of the case. We include some suggestions for enhancing such merits-based consideration of predictive coding parameters in Part IV.C below.

B. Responsiveness, Relevance, and Importance

A different set of issues arises from the ability of many of the predictive coding software systems to rank responsive documents in accordance with their purported relevance.[250] Accordingly, judges could order disclosure of all documents ranked by the software as having a relevance level above 25%, 50%, or 60%.[251] Such rankings seem especially suited to proportionality review because they seem to ensure that the money spent on discovery results in

---

246. *See* Yablon & Landsman-Roos, *supra* note 85, at 723–24 (citing FED. R. CIV. P. 26(b)(2)(C)(iii)).

247. *See* FED. R. CIV. P. 26(b)(2)(C)(iii). It should also be noted that the limitations on ESI set forth in Rule 26(b)(2)(B) expressly reference the limitations on discovery, generally, in Rule 26(b)(2)(C). FED. R. CIV. P. 26(b)(2)(B).

248. *See* Yablon & Landsman-Roos, *supra* note 85, at 736–37.

249. *See id.*

250. *See* PACE & ZAKARAS, *supra* note 2, at 60.

251. *See id.*

production of the documents most relevant to the case and, therefore, the most likely to confer the greatest "benefit." This belief is, unfortunately, an overly simplistic view of what the software is doing and has the potential to lead courts into grave errors. To understand why this logic is misleading, we must distinguish between the "responsiveness," "relevance," and "importance" of ESI and other documents.

We can define "responsiveness" as the likelihood that a given document (or ESI) falls within the category of documents (or ESI) set forth in the document demand the requesting party served. "Relevance," in contrast, can be defined for our purposes as the probative value of a document (or ESI) to the disputed issues in the case. "Important" documents (or ESI) can then be considered those with a high degree of "relevance"—those that are likely to be actually used by the parties to prove or refute the allegations in the complaint.[252]

Given these definitions, a document can be simultaneously highly responsive, hardly relevant, and not at all important. Consider, for example, a hypothetical document request in the *Da Silva Moore* case seeking "all documents relating to the hiring of plaintiff Da Silva Moore." A document reflecting the fact that Da Silva Moore was hired by defendant for a particular position on a particular day, and nothing more than that, would obviously be a document called for by the document request. We might say that its probability of being "responsive" was close to 100%. Yet, it might also be the case that the document does not provide any information that was not already known to both sides and only demonstrates facts that are undisputed in the case. As such, its relevance would be very low and it would have no importance whatsoever.

The significance of these distinctions to predictive coding protocols is substantial. What the predictive coding software "learns" from the seed set that the document reviewer codes is how to identify a "responsive" document. What the software identifies when it ranks documents according to so-called "relevance" is really their likelihood of being "responsive." Effectively, it identifies the degree of similarity that exists between a document that the system coded and the documents that an expert reviewer has previously determined to be responsive to the document request. Documents with high "relevance" scores on such systems are merely documents very similar to those that reviewers previously coded as responsive and are not necessarily those that are most probative or important to the disputed issues in the case.

---

252. These distinctions are our own, devised to emphasize the distinctions made in this Part, and do not necessarily track the way the words are used in actual legal practice. For example, Rule 26(b) of the Federal Rules of Civil Procedure limits discovery to "relevant information." FED. R. CIV. P. 26(b). This rule is designed to include a rather broad definition of "relevance"—one that makes all responsive documents validly requested under Rule 26(b) "relevant" in some sense. *See id.* Yet anyone with any experience in the current discovery practice knows that the vast majority of documents (or ESI) in response to discovery requests have little or no probative value with respect to the disputed issues of the case.

The great danger of this relevance ranking software is that courts and parties will confuse the different concepts of "relevance" involved and assume that undiscovered or unproduced documents with lower relevance rankings are, for that reason alone, likely to be unimportant to the case. Thus, excluding such documents from discovery solely because of their low responsiveness ranking would be a grave error.

Yet, this reality is only part of the story. Responsiveness differs from relevance because requesting parties choose to define the categories of information requested far more broadly than the precise information in which they are most interested. For example, plaintiffs in *Da Silva Moore* might have limited their document request to information showing "why plaintiff was in fact fired," but they probably feared justifiably that such a narrowly targeted inquiry could easily be used to exclude documents that did not directly discuss her firing but might have hinted at its cause. Accordingly, lawyers drafting document requests have gotten into the habit of drafting them quite broadly and, therefore, expect to obtain large amounts of marginally relevant (but fully responsive) documents along with, they hope, a few highly relevant and important documents. The software utilized in predictive coding could, in theory, allow for more narrowly drawn, targeted document requests to be utilized and responded to accurately. Because such software analyzes documents in a more sophisticated way than just responding to particular keywords, it might enable the system to identify responsive documents through the use of more sophisticated concepts like causation, motive, etc.

The narrowness of a document request and the likelihood that it will, therefore, produce highly relevant, probative documents is an important factor in many of the standards courts have utilized in applying proportionality review.[253] Accordingly, once courts and litigants have sufficient familiarity with the capacities of the predictive coding software, it is possible that this familiarity will lead to a narrowing and focusing of discovery inquiries to the precise disputed issues most relevant to the case. This narrowing would lead to far greater efficiency and cost savings than what predictive coding could achieve alone. Although far in the future, it is worth keeping this utopian vision in mind as predictive coding software develops.

Another feature of predictive coding systems is that they provide a quantifiable estimate of the accuracy of the system, which necessarily also implies a quantifiable error rate.[254] In most current discovery practice, although it can be stated with assurance that some responsive and indeed even highly relevant documents are not produced due to human error, poorly selected keywords, misunderstandings and miscodings, etc., it is pretty much impossible

---

253. *See* Zubulake v. UBS Warburg LLC (*Zubulake III*), 216 F.R.D. 280, 287, 289 (S.D.N.Y. 2003); Yablon & Landsman-Roos, *supra* note 85, at 773.

254. *See, e.g.*, SYMANTEC CORP., PREDICTIVE CODING DEFENSIBILITY: THE SYMANTEC TRANSPARENT PREDICTIVE CODING WORKFLOW 1 (2013) (discussing the company's "Transparent Predictive Coding" software feature).

to quantify how many such documents are not produced due to those failings.[255] With predictive coding software, however, the error rate can be estimated from the results obtained by the system.[256]

Some judges may be reluctant in applying the proportionality standard to accept a recall rate of say, 80%—which implies that 20% of responsive documents have been missed. This reluctance would also be an error. The relatively greater accuracy of predictive coding software in identifying responsive documents means that in fact more, rather than fewer, responsive documents are being produced. All that has happened is that the number of such unproduced responsive documents has become, in Donald Rumsfeld's famous phrase, "known unknowns" rather than "unknown unknowns."[257]

C. Some Fairly Tentative Suggestions

Because we believe that predictive coding will become an increasingly important part of discovery practice and also that the proportionality standard will remain the dominant legal principle governing discovery disputes for the foreseeable future, the critical practical question is how best to accommodate predictive coding with proportionality review. What follows are some fairly tentative suggestions along those lines.

First, for the foreseeable future, predictive coding should be seen as an alternative method of obtaining certain types of discovery rather than a replacement for existing document retrieval methods. That is, rather than viewing the choice as whether to produce documents either through predictive coding or keyword review, the courts and parties should endeavor, at least in the initial stages of creation of the discovery plan, to determine first what information they most require and only secondly what technique is the best one for obtaining it. Obviously, this framework will be hard to use in cases where predictive coding software requires a major investment of time and money, but it is precisely in those cases where there is the greatest danger that important information will not be produced simply because it is difficult to code or locate under the predictive coding protocols.[258]

Information sought could be divided into two categories: (1) that best produced through predictive coding and (2) that best produced by other means.

---

255. However, it is worth noting that in *Global Aerospace*, defendants did cite a 40% nonproduction rate of responsive documents based on manual review and an 80% nonproduction rate based on keyword search alone. *See* Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding, *supra* note 194 at 2, 7–8 (citing Grossman & Cormack, *supra* note 9, at 18, 37 tbl.7). These rates, apparently based on prior studies, were obviously cited to make the 25% error rate of the predictive coding software seem small in comparison. *Id.* at 2.

256. *See* SYMANTEC CORP., *supra* note 254, at 1.

257. Donald H. Rumsfeld, U.S. Sec'y of Def., and Gen. Richard Myers, Chairman, Joint Chiefs of Staff, Department of Defense News Briefing (Feb. 12, 2012, 11:30 AM), *available at* http://www.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636.

258. *See supra* text accompanying notes 130–34.

The amount of information in each of those categories would help determine whether predictive coding should be utilized in an individual case. And even in cases where predictive coding is utilized for the majority of ESI production, the existence of this second category of information will remind the court and responding party of their obligation to provide some method for retrieval of this other information, assuming it otherwise meets the standard of proportionality review.

A second suggestion is to maintain maximum transparency, particularly regarding costs. While we understand that predictive coding software is proprietary and that companies are reluctant to disclose too much information publicly about the internal workings of the technology, their interest in maintaining confidentiality regarding their prices is much weaker. Rather, the courts, before ordering any predictive coding protocols, should be able to obtain complete and accurate pricing information not only about the particular protocols the responding party wishes to adopt but also of other relevant protocols that might provide broader or more accurate discovery to the requesting party. The party seeking to utilize predictive coding should be able to provide information not only about total cost but also about marginal costs. All of this information, of course, will be extremely useful to the courts in making their burden–benefit analysis under the proportionality standard.

Finally, we come to the question of timing. We have seen a strong tendency on the part of courts considering predictive coding to defer as many decisions as possible regarding the operation of such systems until after there has been some preliminary document review and there has been some experience utilizing the predictive coding software.[259] We believe this practice reflects not only the court's unfamiliarity with predictive coding systems but also two additional considerations implicit in creating a well-functioning pretrial discovery program. The first such consideration is not to commit the court or the parties to a discovery plan that cannot be changed or revised in light of subsequent developments. Such inflexibility is a great danger when using predictive coding where so much of the success of the system is based on the way the initial training set is selected and coded. Nonetheless, courts are, we believe, correctly trying to minimize the number of decisions that must be made at that early stage and preserving, as much as possible, the ability to change or limit the predictive coding protocol in light of subsequent information or developments. Courts are well aware of the danger of being "locked in" to a given discovery program or plan. The *Kleen Products* decision illustrates the courts' tendency not to revisit basic discovery questions once there have been "sunk costs" and substantial time committed to a prior discovery plan.[260] Nonetheless, the very uncertainty and unpredictability of litigation, and the fact that predictive coding protocols must be initially developed at an early stage in such litigation, make it important that

259. *See supra* text accompanying note 140.
260. *See supra* text accompanying notes 183, 189.

courts try to retain as much flexibility as possible to revise or alter predictive coding protocols in light of subsequent events. Once again, cost is a major factor. A predictive coding system that is structured in such a way that subsequent changes in the system are not priced at prohibitive levels is one that is likely to be, and should be, preferred by the courts.

The other reason for deferring major decisions regarding discovery is to obtain a better sense of the underlying merits of the case. Once the court has reviewed the training set and the documents that have been produced through the utilization of that training set, the parties will be in a better position to argue about what changes, if any, should be made in the predictive coding protocols to maximize the production of beneficial documents at the lowest cost. As previously noted, this back-and-forth between the parties may involve substantial merits-based arguments or inquiries.[261] For example, if the requesting party can show that the system excluded a highly relevant document as nonresponsive, that would constitute a strong argument for additional training or revision of the system. Indeed, it might even make sense to test the system by intentionally including some "made up" but highly relevant documents just to see how the system handles them.[262] On the other hand, if the system produces a reasonable number of documents that are responsive to the request but none that are particularly relevant or important, that occurrence might well justify an order limiting subsequent discovery or shifting costs.

While deferral of major decisions regarding discovery protocols is a good idea, it is impossible to defer all such decisions. In order to get a predictive coding system functioning and for it to be run, initial decisions must be made regarding the confidence levels used for the randomly selected seed set and various other initial parameters. Even here, the nature of the dispute, the amount at stake, and whatever other merits-based information is available should guide the court, to the extent possible. We expect that the courts will, and should rightly, be reluctant at early stages in the litigation to make decisions adopting initial predictive coding protocols that strongly favor one side or the other. In such circumstances, impartial expert bodies such as The Sedona Conference can play a very important and useful role. They can develop presumptive guidelines for the use of predictive coding protocols that analyze and consider, in depth, such complex, technical issues as confidence levels and intervals, training and control, set construction, and other matters. Moreover, these bodies can consider such guidelines not only from a technical point of view but also in light of the broader goals of the discovery process itself.

---

261. *See supra* p. 667.

262. This method would be a different version of the "artificially created" training document that others have proposed as the basis for a seed set. *See* Ball, *supra* note 87.

V. COURT-ORDERED CODING

As we have seen, courts are beginning to confront the question of whether they can order the use of predictive coding over the objection of one or even both parties. It is a question which raises fundamental issues about the role courts should play in the management of discovery. The limited number of predictive coding decisions already decided illustrates some of the ways courts may exercise coercive power in ordering predictive coding. Courts might: (1) grant a responding party's motion to utilize predictive coding over the opposition of the requesting party, as in *Global Aerospace*;[263] (2) grant a requesting party's motion to utilize predictive coding over the opposition of the responding party, a motion similar to that made by the plaintiffs in *Kleen Products*;[264] or (3) order or suggest the use of predictive coding sua sponte as was the case in *EORHB*.[265]

Each of these scenarios presents related, but different, issues of court ordered predictive coding. But, at their core, they raise a question of how much control courts may and should exercise over discovery practice. Much has been written about the changing role of judges in the discovery process. While the old judicial paradigm was that of judge-as-umpire,[266] judges are now largely conceived of as being managers of the judicial process,[267] and it appears that judges have largely embraced such a role. That said, questions still remain concerning what exactly that managerial role entails. For instance, on one end, managerial judging could be as simple as setting clear discovery deadlines and ensuring detailed initial disclosures to the court. On the other hand, a managerial judge could play a much more involved role, defining the parameters of discovery or even wading into the methodological questions concerning document sampling and keyword searches.

We have argued previously that judges should take an active role in managing e-discovery.[268] But that role should be adjudicative, not coercive, and informed by considerations of the nature and merits of the case, not merely managerial concerns. That is, while judges should avail themselves of their managerial powers to oversee discovery and make decisions to conduct sampling, review the use of keywords, shift costs, or cutoff discovery, they

---

263. *See supra* text accompanying notes 206–09.

264. *See supra* text accompanying note 182.

265. *See supra* text accompanying notes 224–25.

266. *See* Robert F. Peckham, *The Federal Judge as a Case Manager: The New Role in Guiding a Case from Filing to Disposition*, 69 CALIF. L. REV. 770, 770 (1981).

267. *See generally* Steven S. Gensler, *Judicial Case Management: Caught in the Crossfire*, 60 DUKE L.J. 669, 670–71 (2010) (discussing the change to judges taking an active role in managing the cases before them); James S. Kakalik et al., *Just, Speedy, and Inexpensive? An Evaluation of Judicial Case Management Under the Civil Justice Reform Act*, 49 ALA. L. REV. 17, 47–48 (1997) (discussing how judges are currently implementing case management policies and procedures into their courts); Judith Resnik, *Managerial Judges*, 96 HARV. L. REV. 374, 380 (1982) (discussing problems that have arisen with judicial management).

268. *See* Yablon & Landsman-Roos, *supra* note 85, at 736–37.

should not unnecessarily wade into the business of prescribing the methods by which document searching and production should occur.

That distinction may appear to be a subtle one, but it is grounded in sound managerial and efficiency concerns as well as the force of precedent. The common law and prevailing discovery practice under the Federal Rules of Civil Procedure leave the methodological choices relating to discovery production to the parties, in particular the responding party. Nothing in the Federal Rules suggests that a court may prescribe the manner in which documents are collected. Moreover, the Sedona Principles—reflecting best practices in the conduct of discovery—wisely explain that it is the producing party who is in the best position to determine the method by which documents are collected. Principle 6 states: "Responding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information."[269]

Principle 6 is generally based on the assumption that "[t]he standard form of production, in which the producing party identifies and produces responsive information, allows the party with the greatest knowledge of the computer systems to search and utilize the systems to produce responsive information."[270] Yet, while choices regarding the way in which documents are collected are discretionary, absent an agreement or objection, "the producing party under the Sedona Principles doesn't have carte blanche to specify the mode of collection."[271] That said, while the commentary to Principle 6 suggests that a producing party's discretion is not unlimited, nowhere in the Principle does it indicate that such discretion is to be automatically reviewed or limited by the court.[272] Indeed, if both parties agree to a discovery protocol, whether involving predictive coding or otherwise, a court will rarely have justifiable grounds to intervene. Rather, it is when a requesting party disrupts the presumption in favor of a responding party selecting the means of collection by objecting to the responding party's choice that the coercive powers of the court are properly invoked.

With these background principles in mind, what then is the proper approach in each of the above-described scenarios of court-ordered predictive coding?

---

269. THE SEDONA CONFERENCE, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION, at ii (2d ed. 2007), *available at* https://thesedonaconference.org/download-pub/81.

270. *Id.* at 39; *see also* Ford Motor Co. v. Edgewood Props., Inc., 257 F.R.D. 418, 427 (D.N.J. 2009) ("The producing party responding to a document request has the best knowledge as to how documents have been preserved and maintained.").

271. *Ford Motor Co.*, 257 F.R.D. at 427.

272. *See* THE SEDONA CONFERENCE, *supra* note 269, at 38–42 (citations omitted).

A. Responding Party Seeks to Use Predictive Coding over the Requesting Party's Objection

The easiest of the cases is where a responding party proposes the use of predictive coding and is met by the requesting party's objection. This scenario was the issue in *Global Aerospace*. There, the defendant sought to use predictive coding to respond to the plaintiffs' document requests, the plaintiffs balked, and, in response, the defendants went to the court for judicial authorization.[273] The court allowed the defendants to proceed with the use of predictive coding.[274] The court was not wrong to do so.[275] The general presumption is that the responding party may choose the means of document collection.[276] Because it is generally the case that predictive coding can serve as a reliable and efficient means of document collection, the defendants were justified in their proposed use of the technology.[277] Hence, a court order approving such use was justified because the responding party proposed the use and it had defensible grounds for its choice.

That said, it should not be a general rule that a responding party has "carte blanche" to use predictive coding in any instance, regardless of the objections of the requesting party or the concerns of the court.[278] Importantly, the requesting party in *Global Aerospace* did not identify a reason why the use of predictive coding would cause incomplete document production.[279] If a requesting party can detail specific ways in which a predictive coding protocol will not accurately find or code the requested documents, then an objection—and court order disapproving of the technology's use—may well be justified.

For instance, a court might reject a proposed use of predictive coding, or at least order modifications to a protocol, where legitimate objections are made regarding how the training set was assembled or the control set to be used or where the requesting party could demonstrate that the software was not retrieving significant numbers of responsive documents.

This course is fundamentally no different than how courts have treated other methods of document collection—whether the use of sampling, keyword searching, concept searching, or even manual review. Courts will not reject such methods in the abstract, but they will exercise managerial power where specific objections have been made as to the way in which technology or methods have been implemented. It is also in keeping with the principles stated in Part IV that predictive coding should be viewed as one possible method for the production of

---

273. *See supra* text accompanying notes 196–97.

274. *See supra* text accompanying note 209.

275. *See* THE SEDONA CONFERENCE, *supra* note 269, at 38; *see also Ford Motor Co.*, 257 F.R.D. at 427 (discussing the Sedona Principles).

276. THE SEDONA CONFERENCE, *supra* note 269, at 38.

277. *See supra* note 9 and accompanying text.

278. *See Ford Motor Co.*, 257 F.R.D. at 427.

279. *See supra* text accompanying notes 206–08.

ESI, not the only method, and that the degree to which it is used in any individual case should be determined by the nature of the information legitimately sought by the parties in that case.[280]

Relatedly, then, a party who refuses to agree to a discovery plan involving the use of predictive coding based on legitimate objections and proposes a reasonable plan based on keyword searches or another accepted methodology, should not be held to have failed to cooperate in discovery. While a blanket rejection of a form of document collection, without specific reasons, might be unduly obstructionist, a tailored and grounded objection based on the methodology selected is no less legitimate than an objection to any other form of document collection.

B. Requesting Party Seeks a Court Order Requiring the Producing Party's Use of Predictive Coding

The second case is trickier. This scenario—in which a requesting party seeks a court order requiring the producing party's use of predictive-coding technology—largely tracks the facts of *Kleen Products.* There, the plaintiffs moved to have the defendants use predictive coding to search and collect documents, many of which had already been collected through the use of keyword searches.[281] The court ultimately declined to order the defendants to use predictive coding, noting the presumption in Sedona Principle 6 that a responding party may generally elect the method by which documents are collected.[282] Also important to the court was that the defendants had already spent a considerable amount of time and financial resources on the review and collection of documents, and requiring re-review, using predictive coding, would be wasteful, especially in light of the fact that plaintiffs had not identified what documents they believed were missing from the defendants' production that would have been uncovered through the use of predictive coding.[283]

Kleen Products is, therefore, perhaps not the strongest case of a requesting party seeking a court order requiring use of predictive coding. A better case might be one in which the producing party had not undertaken any document review, and the requesting party had reason to believe that predictive coding would be more cost-effective and accurate in retrieving responsive documents. But, even in such an instance, an appeal to a court for an order requiring a producing party to use predictive coding would be difficult to obtain. Following the instruction of Sedona Principle 6, it is presumptively the responding party's decision to select the method by which documents are collected and produced.[284] As long as the producing party has selected a defensible method by which

280. *See supra* Part IV.C.
281. *See supra* text accompanying note 182.
282. *See supra* text accompanying notes 189–91.
283. *See supra* p. 658.
284. THE SEDONA CONFERENCE, *supra* note 269, at 38.

documents are collected, the presumptive rule has been to accept the choice of the responding party.

That said, there is a division among courts—outside the predictive coding context—as to whether a unilateral choice of a collection method, without any input from an opposing party, is a defensible collection method.[285] The question has arisen whether a responding party may unilaterally create and deploy keyword search terms to winnow down a pool of data, or whether there must be some form of agreement.[286] Following the Sedona Principle, a good faith, unilateral approach to the development of keywords for culling electronic documents has been considered defensible.[287] One court has noted that this "review may range from reading every word of every document to conducting a series of targeted key word searches," but regardless, "the producing party unilaterally decides on the review protocol."[288] However, other courts have required that the requesting party be allowed at least some input regarding the proposed search terms.[289] While there is no clear agreement, the keyword search cases seem to suggest a trend favoring a broad, but not unlimited, choice for the producing party.[290]

The teaching of the keyword search cases suggest that, consistent with Sedona Principle 6, a producing party's choice of a document collection method should be favored and that a requesting party's role should be limited to providing input into the way in which predictive coding or another document collection system is implemented. They also strongly suggest, as do the Sedona Principles, that the optimal discovery procedure is one in which there is agreement by both parties to a discovery plan and document retrieval protocols. In this context, it is worth remembering our previous suggestion that agreement on predictive coding methods is more likely to be obtained if both parties believe they will obtain some benefit from the use of predictive coding.[291]

---

285. *See* Ford Motor Co. v. Edgewood Props., Inc., 257 F.R.D. 418, 427 (D.N.J. 2009); *In re* Priceline.com Inc. Sec. Litig., 233 F.R.D. 88, 91 (D. Conn. 2005); *Zubulake III*, 216 F.R.D. 280, 290 (S.D.N.Y. 2003).

286. *See, e.g.*, *In re Priceline.com*, 233 F.R.D. at 91 ("Defendants shall . . . seek input from plaintiffs regarding proposed search terms.").

287. *See* Mia Mazza et al., *In Pursuit of FRCP 1: Creative Approaches to Cutting and Shifting the Costs of Discovery of Electronically Stored Information*, 13 RICH. J.L. & TECH. 11, at 27–33 (2007) (citations omitted), http://jolt.richmond.edu/v13i3/article11.pdf.

288. *Zubulake III*, 216 F.R.D. at 290.

289. *See, e.g.*, *In re Priceline.com*, 233 F.R.D. at 91 (requiring the defendants to seek the plaintiffs' input as to proposed search terms).

290. *See generally id.* (ordering cooperation as to the identification of keywords); *Ford Motor Co.*, 257 F.R.D. at 427 ("[T]he producing party under the Sedona Principles doesn't have carte blanche to specify the mode of collection . . . .").

291. *See supra* Part IV.A.

C. *Court-Ordered Use of Predictive Coding*

Vice Chancellor Laster's recent sua sponte order requiring both parties to use predictive coding raises a final question of judicial coercion.[292] Specifically, can or should a court order two parties to use predictive coding sua sponte? What if one or both parties object to its use?

This sort of judicial behavior is exceptional. Courts rarely ever wade into the selection of the means by which documents are collected and produced, and when judicial intervention does occur, it is almost always at the parties' request.[293] We believe the reluctance of courts to get involved in the methodological choices of document collection is not only understandable but, in most cases, normatively preferable. The parties have the best information about documents' characteristics, where they are located, and what may be the most efficient manner of collection. Courts should defer to the parties' judgment because they are in the best position to evaluate the cost effectiveness of various discovery methods in light of clients' needs and desires. While a particular form of document collection may seem to the court to be preferable, the court may not be aware of all the constraints and considerations which underlie the responding party's choice of a collection method. This scenario is especially the case where one or both parties objects to the use of predictive coding following a sua sponte order to show cause.

Having said that, however, we do not think it is inappropriate—particularly with respect to a new technology like predictive coding that is still unfamiliar to many lawyers and clients—for a court to *suggest*, as Vice Chancellor Laster effectively did, that it might be the most efficient method for conducting discovery.[294] We have not yet seen how the parties in Delaware will respond to Vice Chancellor Laster's order. But were either party to raise a legitimate objection to the court's order, even one based on sunk costs or unfamiliarity with the new technology, the court should be careful about forcing predictive coding on such a party, especially if that party is likely to have substantial document collection obligations.

VI. CONCLUSION

This past year, 2012, was the year in which predictive coding ceased to be a theoretical possibility and became an active part of discovery practice. The challenge now is to create appropriate legal rules that maximize the benefits of that technology to further the goals of modern discovery practice and create

292. *See supra* text accompanying notes 222–25.
293. *See* discussion *supra* Part III.B–C.
294. *See supra* text accompanying notes 222–25.

incentives to improve that technology to encourage the "just, speedy, and [relatively] inexpensive"[295] conduct of pretrial discovery.

295. FED. R. CIV. P. 1.

*