# PREDICTIVE CODING AND THE PROPORTIONALITY DOCTRINE: A MARRIAGE MADE IN BIG DATA

*Ralph C. Losey*[*]

TABLE OF CONTENTS



---

* Ralph C. Losey is an attorney, author, and educator in the field of electronic discovery, information technology law, and legal search. He is a partner at Jackson Lewis, LLP, where he serves as its *National E-Discovery Counsel* and Chair of its Electronic Discovery practice group, co-founder of the *IT-Lex* foundation and the Electronic Discovery Best Practices group (EDBP.com), and the developer of an online training course in e-discovery (e-DiscoveryTeamTraining.com). Mr. Losey is a frequent speaker and prolific author on e-discovery, having written five books, multiple law review articles, and over forty articles on AI-enhanced review ("predictive coding"). His latest book, *e-Discovery Stories from the Cutting Edge of Law and Technology* (2012), is available on iTunes, and his last paper book, *Adventures in Electronic Discovery* (West 2011), includes an often quoted chapter entitled *Child's Game of 'Go Fish' is a Poor Model for e-Discovery Search*. Mr. Losey is also the principle author and publisher of a popular weekly blog on e-discovery, *e-Discovery Team Blog* (e-discoveryteam.com), which is generally considered the leading source of commentary and analysis in the field, and he has over seventy published opinions to his credit, including some of the largest e-discovery cases in the country. He received his B.A. from Vanderbilt University in 1973, attained his J.D. with honors from the University of Florida School of Law in 1979, and entered private practice in Orlando, Florida, in 1980. Prior to becoming a partner at Jackson Lewis, Mr. Losey was a shareholder with Subin, Shams, Rosenbluth, Moran, Losey and Brennan, P.A., and then a shareholder of Akerman Senterfitt where he was the founding-chair of its e-discovery practice group. For Mr. Losey's full, detailed resume please see RalphLosey.com.

## INTRODUCTION

The search of electronic data to try to find evidence for use at trial has always been difficult and expensive. Over the past few years, the advent of *Big Data*, where both individuals and organizations retain vast amounts of complex electronic information, has significantly compounded these problems. The legal doctrine of proportionality responds to these problems by attempting to constrain the costs and burdens of discovery to what are reasonable. A balance is sought between the projected burdens and likely benefits of proposed discovery, considering the issues and value of the case. Several software programs on the market today have responded to the challenges of Big Data by implementing a form of artificial intelligence ("AI") known as *active machine learning* to help lawyers review electronic documents. This Article discusses these issues and shows that AI-enhanced document review directly supports the doctrine of proportionality. When used together, proportionality and predictive coding provide a viable, long-term solution to the problems and opportunities of the legal search of Big Data.

To demonstrate the combined effectiveness of proportionality and predictive coding, this Article is organized into four parts. Part I discusses how the rapid growth of electronic information drives the rising costs of civil litigation as discovery becomes increasingly expensive. This section also introduces proportionality and predictive coding as means of combating rising costs. Next, Part II explains how AI can be harnessed in document review, noting applicable case law and providing a detailed description of the predictive coding process. Then, Part III proceeds to consider the legal doctrine of proportionality—in other words, balancing the burden of e-discovery with its benefits—and considers relevant case law. Finally, Part IV concludes by demonstrating the close relationship between predictive coding and proportionality, observing that predictive coding allows one to fine-tune discovery in any case to the anticipated value of the suit against the projected costs of document review.

## I. THE HIGH COSTS OF LITIGATION ARISE PRIMARILY FROM EXPLODING VOLUMES OF DIGITAL INFORMATION

The volume of electronically stored information ("ESI") subject to discovery in litigation is growing at an explosive rate.[1] Every five minutes, today's brave new, computational world is said to create the digital equivalent of all of the information stored in the Library of Congress.[2] Put another way, we now create as much information in two days as we have from the dawn of man through 2003.[3]

The mind-boggling increase in the quantity of information is only part of the story. Consider also the impact of the changing form of our information. For millennia, writings were on paper. For centuries, the legal profession depended upon writings, referred to in the law as

---

[1] *See*, *e.g.*, Rowe Entm't, Inc. v. William Morris Agency, 205 F.R.D. 421, 429 (S.D.N.Y. 2002) (explaining that electronic data is so voluminous because, unlike paper documents, "the costs of storage are virtually nil[, and] [i]nformation is retained not because it is expected to be used, but because there is no compelling reason to discard it"); Kenneth Cukier, *Data, Data Everywhere*, ECONOMIST, Feb. 27, 2010, at 3, 3; Jason R. Baron & Ralph C. Losey, *E-Discovery: Did you Know?*, E-DISCOVERY TEAM (Feb. 4, 2010, 10:23 PM), http://e-discoveryteam.com/2010/02/04/baron-and-loseys-new-movie-e-discovery-did-you-know/ (providing video with graphic displays of data explosion and the law).

[2] DAVE EVANS & RICK HUTLEY, CISCO IBSG INNOVATIONS PRACTICE, THE EXPLOSION OF DATA: HOW TO MAKE BETTER BUSINESS DECISIONS BY TURNING "INFOLUTION" INTO KNOWLEDGE 1 (2010), *available at* http://www.cisco.com/web/about/ac79/docs/pov/Data_Explosion_IBSG.pdf.

[3] Marshall Kirkpatrick, *Google, Privacy and the New Explosion of Data*, TECHONOMY (Aug. 4, 2010, 8:57 PM), http://techonomy.typepad.com/blog/2010/08/google-privacy-and-the-new-explosion-of-data.html (reporting statistic from the speech of Eric Schmidt, former CEO of Google, at the Techonomy Conference in Lake Tahoe, CA).

*documents*, as the key evidence for resolving disputes in a fair and just manner.[4] Paper documents were well-known and mastered by every lawyer and judge who swore an oath to uphold the law. This all changed in a historical blink of the eye. In just one generation, documents have dematerialized and transformed into a dizzying array of ephemeral digital media, from email and texts, to Tweets and Facebook posts.

### *A. Paradigm Shift*

Many see this transformation of writing as a much more profound cultural revolution than that precipitated by Gutenberg, which took centuries to play out, not decades.[5] Legal thought-leaders Jason R. Baron and George L. Paul predicted in 2007 that the legal profession must significantly change and adopt new strategies of practice to cope with this information revolution.[6]

Documents originally created on paper still exist in our society, but are rare.[7] Most of the paper documents we see are merely printouts of one dimension (the text) of the original electronic information. The law recognized this transformation, and the Federal Rules of Civil Procedure were amended in 2006 to include ESI as information that can be discovered and used as evidence in lawsuits.[8] ESI is not specifically defined in the rules. The Rules Committee Commentary explained why: "The wide variety of computer systems currently in use, and the rapidity of technological change, counsel against a limiting or precise definition of electronically stored information. Rule 34(a)(1) is expansive and includes any type of information that is stored electronically."[9]

---

[4] *Cf.* RALPH C. LOSEY, ELECTRONIC DISCOVERY: NEW IDEAS, CASE LAW, TRENDS, AND PRACTICES 35–46 (2010) (discussing the comparative importance of paper and electronic records).

[5] George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH., no. 3, art. 10, Spring 2007, at 4–7, http://jolt.richmond.edu/v13i3/article10.pdf (explaining how writing co-evolved with civilization over the past fifty centuries or longer with a slow but steady increase in information as our writing technologies slowly improved, and pointing out that this all changed about twenty-five years ago when mankind invented a totally different form of electronic writing, free from physical confines, that triggered a Big-Bang-like explosion of a new universe of virtually unlimited information).

[6] *Id.* at 3; *see also* Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, 17 RICH. J.L. & TECH., no. 3, art. 9, Spring 2011, at 5, http://jolt.richmond.edu/v17i3/article9.pdf.

[7] *See* LOSEY, *supra* note 4, at 38; *see also* Zubulake v. UBS Warburg LLC, 217 F.R.D. 309, 311 & n.5 (S.D.N.Y. 2003) (citing Wendy R. Liebowitz, *Digital Discovery Starts to Work*, NAT'L L.J., Nov. 4, 2002, at C3 (reporting that in 1999, 93% of all information generated was in digital form)).

[8] *See* FED. R. CIV. P. 34(a)(1)(A) & advisory committee's note to 2006 amendment.

[9] FED. R. CIV. P. 34 advisory committee's note to 2006 amendment.

Even without specific amendments to rules, all state and federal courts today treat ESI as potentially admissible evidence subject to discovery.[10] The first Sedona Principle is now commonplace: "Electronically stored information is potentially discoverable under Fed. R. Civ. P. 34 or its state equivalents. Organizations must properly preserve electronically stored information that can reasonably be anticipated to be relevant to litigation."[11]

### *B. Lawyers Overwhelmed by Rapid Advances in Technology*

The legal profession has been severely stressed by the rapid, ever-accelerating advances in technology. The changes in writing and the resulting information explosion have been the key challenges.[12] ESI is not only changing and evolving into new forms every year, but, as mentioned, is now multiplying at an exponential rate that is almost beyond comprehension.[13]

Most lawyers are unfamiliar with ESI and the complex systems that store it. They prefer the familiar paper and alphabetical filing cabinets. They are paper lawyers living in a digital world. As a result, judges and juries today often do not see the key writings that they need to do justice. The fault lies with the lawyers who, in the U.S. system, are the ones charged with the duty of discovering the truth. They often fail in this duty, not for want of trying, but for the difficulty in finding the key documents. The evidence is lost in plain view, the signal is lost in the noise—hidden by too much data. The information explosion has made the traditional process of legal discovery "enormously expensive and burdensome," and many, including the venerable American College of Trial Lawyers, are implying that this is a crisis in our legal system that threatens our system of justice.[14]

The old methods of reviewing digital writings are too expensive. Few can afford the time and effort required to locate, review, and

---

[10] *See*, *e.g.*, FED. R. CIV. P. 34; N.C. R. CIV. P. 34; VA. CODE ANN. § 8.01-412.12 (Supp. 2013).

[11] THE SEDONA CONFERENCE, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION 11 (Jonathan M. Redgrave et al. eds., 2d ed. 2007), *available at* https://thesedonaconference.org/publication/The%20Sedona%20Principles.

[12] *See* Paul & Baron, *supra* note 5, at 1–2.

[13] *See supra* text accompanying notes 1–3.

[14] THE AM. COLL. OF TRIAL LAWYERS & THE INST. FOR THE ADVANCEMENT OF THE AM. LEGAL SYS., FINAL REPORT ON THE JOINT PROJECT OF THE AMERICAN COLLEGE OF TRIAL LAWYERS TASK FORCE ON DISCOVERY AND THE INSTITUTE FOR THE ADVANCEMENT OF THE AMERICAN LEGAL SYSTEM 16 (2009) ("Although electronic discovery is becoming extraordinarily important in civil litigation, it is proving to be enormously expensive and burdensome.").

produce all relevant evidence using those old methods. The costs and burdens incurred in following old methods can easily exceed the value of an entire case.[15] There is a real danger that the resolution of disputes in a court of law based on both testimony and writings will be a luxury available only to the wealthiest parties. Justice Stephen Breyer made a similar statement in his Preface to an issue of the Sedona Conference Journal:

> [Articles in this Supplement] suggest that if participants in the legal system act cooperatively in the fact-finding process, more cases will be able to be resolved on their merits more efficiently, and this will help ensure that the courts are not open only to the wealthy. I believe this to be a laudable goal, and hope that readers of this Journal will consider the articles carefully in connection with their efforts to try cases.[16]

The law remains as dependent as ever upon documents to prove the truth, but the vast majority of lawyers are untrained and unprepared to handle the electronic documents upon which the world is now built.[17] In fact, most lawyers, even those who specialize in litigation, dislike e-discovery and try their best to avoid it.[18] Lawyers are trained and prepared instead to handle paper documents following systems developed in the twentieth century.

### *C. Failure of Our Law Schools and Law Firm Training*

Even though many scholars, jurists, and practitioners recognize the problems created by the inability of lawyers to keep pace with technology, most law schools still only train students in paper evidence and discovery. Students graduate unprepared to handle ESI where the truth of past events is now stored.[19]

---

[15] *See* Mancia v. Mayflower Textile Servs. Co., 253 F.R.D. 354, 359–60 (D. Md. 2008).

[16] Justice Breyer, *Preface*, 10 SEDONA CONF. J., at i, i (2009 Supp.).

[17] LOSEY, *supra* note 4, at 355.

[18] *See* Ralph Losey, *Spilling the Beans on a Dirty Little Secret of Most Trial Lawyers*, E-DISCOVERY TEAM (Nov. 23, 2011, 8:54 PM), http://e-discoveryteam.com/2011/11/23/spilling-the-beans-on-a-dirty-little-secret-of-most-trial-lawyers/; Ralph Losey, *Tell Me Why?*, E-DISCOVERY TEAM (Dec. 6, 2011, 7:24 AM), http://e-discoveryteam.com/2011/12/06/tell-me-why/.

[19] LOSEY, *supra* note 4, at 328; William Hamilton, *The E-Discovery Crisis: An Immediate Challenge to Our Nation's Law Schools*, *in* ELECTRONIC DISCOVERY: NEW IDEAS, CASE LAW, TRENDS, AND PRACTICES 401, 402–04 (2010); Shannon Capone Kirk & Kristin G. Ali, *"Teach Your Children Well": A Case for Teaching E-Discovery in Law Schools*, *in* ELECTRONIC DISCOVERY: NEW IDEAS, CASE LAW, TRENDS, AND PRACTICES 394, 396 (2010); Judge Shira Scheindlin & Ralph Losey, *E-Discovery and Education*, *in* ELECTRONIC DISCOVERY: NEW IDEAS, CASE LAW, TRENDS AND PRACTICES 337, 343 (2010).

Novice lawyers are instead trained in law school, and as entry-level associates in most law firms, in paper-based legal search and review methods that are one-dimensional and linear in nature. They typically follow a sequential *Bates Stamp* organizational model created in the 1890s.[20] These linear systems, which were developed in the nineteenth and twentieth centuries for the discovery and production of documents, continue to be used today by most attorneys for both ESI and paper discovery.[21] Other experts and I have started training programs to address these problems that are related to, but still largely outside of, formal law school curriculum.[22]

### *D. Processes and Methods Designed for Search and Review of Paper Documents Do Not Work When Applied to High Volumes of ESI*

The old linear review methods involved serial culling of documents down to a final production set. The process generally required multiple reviews of the same document for different purposes. It was inefficient. It was expensive. Moreover, the quality control of human eyes on paper did not work with high volumes of documents. This is shown by the latest scientific experiments where the agreement rate in identifying relevant documents among professional legal reviewers was found to be around 50%.[23]

This tradition of multiple manual reviews, with only limited computer assistance and typically on a linear-based review platform, still continues today. But it is too expensive and inefficient with high volumes of ESI. This will only get worse as the amount of information continues to grow exponentially. Jason Baron, who served from 2000 to 2013 as the Director of Litigation at the United States National Archives and Records Administration, which is in charge of all federal records

---

[20] Ralph C. Losey, *Hash: The New Bates Stamp*, 12 J. TECH. L. & POL'Y 1, 4 (2007) ("A Bates machine uses a self-inking stamp and a mechanically advancing sequence of numbers. Each time the handle of the machine is pressed, a number is imprinted on the document below. With every press of the handle, the number advances sequentially and the next number is inked onto the document.").

[21] Consider the *D'Onofrio* saga, where Magistrate Judge John M. Facciola wrote four opinions describing the processes used in this case and many orders resolving discovery disputes, including an order requiring production of a sample of the 9,413 documents listed on the privilege log. D'Onofrio v. SFX Sports Grp., Inc., No. 1:06-cv-00687-JDB, 2010 WL 3324964 (D.D.C. Aug. 24, 2010); D'Onofrio v. SFX Sports Grp., Inc., 256 F.R.D. 277 (D.D.C. 2009); D'Onofrio v. SFX Sports Grp., Inc. 254 F.R.D. 129 (D.D.C. 2008); D'Onofrio v. SFX Sports Grp., Inc., 247 F.R.D. 43 (D.D.C. 2008).

[22] *See, e.g.*, GEORGETOWN UNIV. LAW CTR., THE EDISCOVERY TRAINING ACADEMY: THE INTERSECTION OF LAW AND IT (2013).

[23] GORDON V. CORMACK, MAURA R. GROSSMAN, BRUCE HEDIN & DOUGLAS W. OARD, OVERVIEW OF THE TREC 2010 LEGAL TRACK 30 (2012).

including White House email, explains this as a problem of scale.[24] He projects the number of White House emails will soon exceed one billion, if it has not done so already; moreover, he estimates it would cost over *$2 billion* to search that many emails.[25] That assumes a team of one hundred full-time lawyers working *over fifty-four years* at a very low billing rate of $100 per hour.[26] Although it also assumes computer-assisted review tools, they would follow the old paper-based linear review models.[27]

Moreover, too many mistakes are being made when these traditional linear review methods are applied to the astronomical volumes and new media of ESI.[28] For instance, in a large construction case in 2012 involving millions of documents reviewed for possible production, both sides inadvertently produced thousands of privileged documents.[29] They did so despite expenditures of tens of millions of dollars for traditional attorney review of each document before production.[30] The prevailing defendant in this case was awarded over $20 million in fees and costs.[31]

---

[24] *See* Paul & Baron, *supra* note 5, at 2.

[25] *Id.* at 12–13.

[26] *Id.*

[27] *Id.*; Jason R. Baron, E-Discovery and the Problem of Asymmetric Knowledge, Address at the Ninth Annual Georgia Symposium on Ethics and Professionalism: Ethics and Professionalism in the Digital Age (Nov 7, 2008), *in* 60 MERCER L. REV. 863, 868–69 (2008).

[28] *See, e.g.*, Mt. Hawley Ins. Co. v. Felman Prod., Inc., 271 F.R.D. 125, 135–36 (S.D. W. Va. 2010) (addressing a serious mistake made that resulted in waiver of privilege in spite of sophisticated counsel with very elaborate processes and safeguards); Diabetes Ctrs. of Am., Inc. v. Healthpia Am., Inc., No. 4:06cv-03457, 2008 WL 336382, at *2, *4 (S.D. Tex. Feb. 5, 2008) (denying sanctions against either party when both made material mistakes producing discovery, such as relying on an unsupervised junior associate or responding with incomplete information); Danis v. USN Commc'ns, Inc., 53 Fed. R. Serv. 3d (West) 828, 876–77, 897 (N.D. Ill. 2000) (recommending a $10,000 fine against a CEO personally when the inexperienced general counsel he hired to supervise ESI preservation was grossly negligent).

[29] Tampa Bay Water v. HDR Eng'g, Inc., No. 8:08-CV-2446-T-27TBM, 2012 WL 5387830, at *1, *15, *21 (M.D. Fl. Nov. 2, 2012). The plaintiff alone inadvertently produced 23,000 privileged documents. *Id.* at *15. The prevailing defendant in this case was awarded over $20 million in fees and costs. *Id.* at *1. Of this sum, $3,100,000 was awarded as a cost for e-discovery vendor processing and hosting of 2.7 million documents for review. *Id.* at *21; *see also* Ralph Losey, *$3.1 Million e-Discovery Vendor Fee Was Reasonable in a $30 Million Case*, E-DISCOVERY TEAM (Aug., 4, 2013, 9:46 PM), http://e-discoveryteam.com/2013/08/04/3-1-million-e-discovery-vendor-fee-was-reasonable-in-a-30-million-case/#comment-60139.

[30] Losey, *supra* note 29 (estimating $4,590,000 ($1.70 per file) to have been spent by one defendant in *attorney fees* to review the documents).

[31] *Tampa Bay Water*, 2012 WL 5387830, at *22.

### *E. Cheap Lawyers Are Not the Answer*

Some are looking for an answer to these expense issues by keeping the old processes, but outsourcing the work of manual review to less expensive contract lawyers.[32] They are called "contract lawyers" because the law firm that represents the client typically does not employ them.[33] Instead, they work for some other company under a contract to do review work. These contract lawyers may be located in India or other countries, or may be down the street from your office, or down the hall.[34] They are almost always paid far less than the first-year associates in most law firms, even less than paralegals or secretaries.[35]

Even assuming contract lawyers can adequately perform the task of the first-level relevance review, this is still just a stopgap measure based on old, linear paper-review methods. With ESI increasing so rapidly, outsourcing is futile as a long-term strategy. It merely attempts to tread water in the midst of a flood. An illustration of the futility of this outsourcing strategy is the attempt by the Department of Justice ("DOJ") to reduce the costs of a privilege review in the 2009 case *In re Fannie Mae Securities Litigation*.[36] Even though the DOJ used outside contract lawyers to do first-pass relevancy review to respond to a third party subpoena, the expenses still exceeded $6 million.[37] The district court's order denying the Government's motion for cost-shifting to the requesting party was upheld by the appellate court.[38]

### *F. The Answer Lies in Predictive Coding and Proportionality*

The answer does not lie in modifying the system somewhat to employ cheap labor to do manual review. Not only are the growing volumes of data too high for this to work, but this kind of manual review by teams of contract lawyers is remarkably inaccurate. The inconsistency rate between reviewers is typically as high as 70%, which means that different reviewers looking at the same documents would only agree with each other on the relevance of those documents an

---

[32] *See* Paul & Baron, *supra* note 5, at 3 & n.5.

[33] DEBORAH ARRON & DEBORAH GUYOL, THE COMPLETE GUIDE TO CONTRACT LAWYERING 7 (1999).

[34] Ralph Losey, *Perspective on Legal Search and Document Review*, E-DISCOVERY TEAM (Mar. 11, 2012, 4:51 PM), http://e-discoveryteam.com/2012/03/11/perspective-on-legal-search-and-document-review/.

[35] *See id.*

[36] *See In re* Fannie Mae Sec. Litig., 552 F.3d 814 (D.C. Cir. 2009).

[37] *Id.* at 817.

[38] *Id.* at 821, 824.

average of 30% of the time.[39] A recent study of a large contract review team project found an agreement rate of only 16%.[40]

The answer is a whole new system for e-discovery, a system based on the new doctrine of proportionality wedded to predictive coding, a new breakthrough, *disruptive technology*[41] for search and review. This Article will explain both the doctrine and technology, and show how their features reinforce each other to provide a viable solution to the problems of e-discovery. But first, here is more information on the problem from a recent study by the RAND Corporation.[42] The RAND Report concluded, consistent with this Article, that new predictive coding technologies, coupled with radical new legal methods, provide our best hope for the future.[43]

### *G. RAND Report on Litigation Expenses*

The RAND Corporation completed a study in 2012 on the high costs of electronic discovery entitled, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery* ("RAND Report").[44] The RAND Report concluded that the primary problem in e-discovery is the high cost of document review.[45] Based on corporate

---

[39] Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697, 701 (2000) (reporting that two retired intelligence officers agreed on responsiveness on only 45% of the documents, and that when three subject matter experts were considered they agreed on only about 30% of the documents); *see also* WILLIAM WEBBER, RE-EXAMINING THE EFFECTIVENESS OF MANUAL REVIEW (2011); 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., no. 3, art. 11, Spring 2011, at 10–11, http://jolt.richmond.edu/v17i3/article11.pdf; William Webber, *How Accurate Can Manual Review Be?*, EVALUATING E-DISCOVERY (Dec. 18, 2011, 6:41 AM), http://blog.codalism.com/?p=1549.

[40] Herbert L. Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70, 74 (2010); *see also* Grossman & Cormack, *supra* note 39, at 13–14 (applying Roitblat, Kershaw & Oot to suggest inconsistencies of 84% and agreement rates of 16%).

[41] *See* Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1, 6 (2013) [hereinafter *Grossman-Cormack Glossary*] (discussing how and why TAR is disruptive technology).

[42] The RAND Corporation is a well-known and prestigious non-profit institution. Its stated charitable purpose is to "improve policy and decisionmaking through research and analysis." *About RAND: History and Mission*, RAND CORP., http://www.rand.org/about/history.html (last updated Sept. 4, 2013).

[43] NICHOLAS M. PACE & LAURA ZAKARAS, RAND CORP., WHERE THE MONEY GOES: UNDERSTANDING LITIGANT EXPENDITURES FOR PRODUCING ELECTRONIC DISCOVERY 99–101 (2012) [hereinafter RAND REPORT].

[44] *Id.* at iii.

[45] *Id.* at 41–42.

surveys, the RAND Corporation found that document review constitutes 73% of the total cost of e-discovery.[46] For that reason, RAND focused its first e-discovery report on this topic, with side comments on the issue of preservation.[47]

The RAND Report not only analyzes the problem, it recommends a radical solution; namely, the adoption of new predictive-coding-type search and review methods.[48] The RAND Report also points out the resistance of the legal profession to taking the bold steps necessary to use such new methods:

> Despite the apparent promise of predictive coding and other computerized categorization techniques, however, the legal world has been reluctant to embrace the new technology. . . . [T]he key reason is the absence of widespread judicial approval of the methodology, specifically regarding any acknowledgment of the adequacy of the results in actual cases or whether the process was a reasonable way to prevent inadvertent privilege waiver. Without clear signs from the bench that the use of computer-categorized review tools should be considered in the same light as eyes-on review or keyword searching, litigants involved in large-scale reviews are unlikely to employ the technologies on a routine basis.
>
> . . . .
>
> The use of computerized categorization techniques, such as predictive coding, will likely become the norm for large-scale reviews in the future, given the likelihood of increasing societal acceptance of artificial intelligence technologies that might have seemed like

---

[46] *Id.*

[47] JAMES N. DERTOUZOS, NICHOLAS M. PACE & ROBERT H. ANDERSON, RAND CORP., THE LEGAL AND ECONOMIC IMPLICATIONS OF ELECTRONIC DISCOVERY: OPTIONS FOR FUTURE RESEARCH (2008).

[48] As the RAND Report states:

> To truly open the doors to more-efficient ways of conducting large-scale reviews in the face of ever-increasing volumes of digital information, litigants that have complained in the past about the high costs of e-discovery will have to take some very bold steps.
>
> . . . .
>
> The most promising alternative available today for large-scale reviews is the use of predictive coding and other computerized categorization strategies that can rank electronic documents by the likelihood that they are relevant, responsive, or privileged. Eyes-on review is still required but only for a much smaller set of documents determined to be the most-likely candidates for production. Empirical research suggests that predictive coding is at least as accurate as humans in traditional large-scale review. Moreover, there is evidence that the number of hours of attorney time that would be required in a large-scale review could be reduced by as much as three-fourths, depending on the nature of the documents and other factors, which would make predictive coding one answer to the critical need of significantly reducing review costs.

RAND REPORT, *supra* note 43, at 83, 97.

> improbable science fiction only a few decades ago. The problem is that considerable sums of money are being spent unnecessarily today while attitudes slowly change over time. New court rules might move the process forward, but the best catalyst for more-widespread use of predictive coding would be well-publicized instances of successful implementation in cases in which the process has received close judicial scrutiny. It will be up to forward-thinking litigants to make that happen.[49]

Since the RAND Report was issued in 2012, several courts have approved the use of predictive coding, which this Article will discuss, and this resistance factor has been greatly reduced. But the Report discusses other resistance factors as well, including an ethical issue that is rarely discussed:

> Another barrier to the widespread use of predictive coding could well be resistance to the idea of outside counsel motivated not so much by accuracy issues as by the potential loss of a historical revenue stream. Some interviewees reported grumblings from outside counsel when their companies decided to directly handle a fraction of the overall review process or to markedly reduce what was shipped out for review through the use of additional data processing.[50]

Another resistance factor implied by the RAND Report that remains a significant problem is the high prices charged by some vendors for the predictive coding features of their review software.[51] For this reason, predictive coding software use is typically limited to large cases. As the cost of the software inevitably comes down in the future, the use is likely to expand to medium and even small size cases where at least 25,000 to 50,000 documents have to be reviewed for possible relevance.[52]

### *H. Two-Fold Solution*

The RAND Report correctly concludes that the legal profession must now take bold steps to change our current system of discovery. The

---

[49] *Id.* at 98–99.

[50] *Id.* at 76.

[51] The RAND Report explains that ESI may be cost-prohibitive in smaller cases:

> Moreover, computer applications for conducting review are unlikely to be economically viable options when dealing with smaller document sets, in which any savings in attorney hours might be overwhelmed by vendor costs and machine-training requirements. Existing approaches, such as deduplication, cluster analysis, and email threading, may provide a more practical answer in these situations.

*Id.* at 98.

[52] *Cf.* Order at 2, 4, Northstar Marine, Inc. v. Huffman, CA 13-00037-WS-C (S.D. Ala. Aug. 27, 2013), ECF No. 28 (enforcing the parties' agreement to use predictive coding software and rejecting plaintiff's contention that it was "having difficulty locating an inexpensive provider of electronic search technology," which demonstrated a lack of "due diligence" on the part of plaintiff's counsel).

existing linear, confrontative,[53] one-dimensional, largely manual, costly, Bates Stamp approach to discovery must be replaced with a cooperative, iterative, largely automated, predictive-coding-based, proportionally cost-controlled, hash-value approach.[54]

Two ways to do this have been developing in the law for the past few years. The first is legal, involving amendments to rules[55] and development of a new body of law for e-discovery, and the second is technological-scientific. The legal approach has focused on the doctrine of proportionality,[56] combined with a new appreciation for legal ethics,[57] and the duty of attorneys to cooperate in e-discovery.[58] The technical

---

[53] Ken Withers, *When E-Mail Explodes*, SAN DIEGO LAW., Nov.–Dec. 2008, at 36, 36–38 (discussing confrontation and civility in e-discovery).

[54] *See* Losey, *supra* note 20, at 3, for more on hash values and e-discovery.

[55] The 2006 Amendments to the Federal Rules of Civil Procedure modified Rules 16, 26, 33, 34, 37, and 45, as well as Form 35, to include electronic discovery. Amendments to Federal Rules of Civil Procedure, 547 U.S. 1233 (2006). In particular, see FED. R. CIV. P. 16(b); 26(a)(1)(B); 26(b)(2)(B); 26(b)(5)(B); 26(f); 33(d); 34(a); 34(b); 37(f); 45(a)(1)(C). At the time of this writing, additional rule amendments are under consideration and in the final stages of public review. *See* ADVISORY COMMITTEE ON CIVIL RULES (2013), available at http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Agenda%20Books/Civil/CV2013-04.pdf. The adoption of these new rules sometime in 2014 appears probable, although, some modifications to the final language may be made. These rules will embody and strengthen the proportionality doctrine, especially as it pertains to sanctions. *See* Sekisui Am. Corp. v. Hart, No. 12 Civ. 3479 (SAS)(FM), 2013 WL 2951924, at *3 & n.3 (S.D.N.Y. June 10, 2013) (explaining pending rule revisions' impact on sanctions law); *see also* FED. R. EVID. 502.

[56] THE SEDONA CONFERENCE, THE SEDONA CONFERENCE COMMENTARY ON PROPORTIONALITY IN ELECTRONIC DISCOVERY 3 (Conor R. Crowley et al. eds., 2013) [hereinafter SEDONA, COMMENTARY ON PROPORTIONALITY (2013)], *available at* https://thesedonaconference.org/publication/The%20Sedona%20Conference%20Commentary%20on%20Proportionality. Moreover, consider the principles developed by a Seventh Circuit committee:

> Principle 1.03 (Discovery Proportionality)[:] The proportionality standard set forth in Fed. R. Civ. P. 26(b)(2)(C) should be applied in each case when formulating a discovery plan. To further the application of the proportionality standard in discovery, requests for production of ESI and related responses should be reasonably targeted, clear, and as specific as practicable.

SEVENTH CIRCUIT ELEC. DISCOVERY COMM., SEVENTH CIRCUIT ELECTRONIC DISCOVERY PILOT PROGRAM: INTERIM REPORT ON PHASE THREE 6 (2013).

[57] *See* Memorandum Opinion and Order, Kleen Prods. LLC v. Packaging Corp. of Am., No. 1:10-cv-05711 (N.D. Ill. Sept. 28, 2012), ECF No. 412; Ralph Losey, *Attorneys Admonished by Judge Nolan Not to "Confuse Advocacy with Adversarial Conduct" and Instructed on the Proportionality Doctrine*, E-DISCOVERY TEAM (Oct. 7, 2012, 4:40 PM), http://e-discoveryteam.com/2012/10/07/attorneys-admonished-by-judge-nolan-not-to-confuse-advocacy-with-adversarial-conduct-and-instructed-on-the-proportionality-doctrine/; *see also* MODEL CODE OF PROF'L CONDUCT R. 3.2–3.4 (2013).

[58] The lead article and summary on cooperation explains as follows:

> Lawyers have twin duties of loyalty: While they are retained to be zealous advocates for their clients, they bear a professional obligation to conduct

approach has been oriented toward software and specialist experts, and recognizes the growing importance of e-discovery vendors. The technical approach has recently culminated in the creation of electronic document review software that uses artificial intelligence to find the documents needed from Big Data in a very fast, efficient, and effective manner. This new technology is next described.

## II. THE USE OF ARTIFICIAL INTELLIGENCE IN DOCUMENT REVIEW

Predictive coding uses a type of AI programing that allows the computer, a/k/a the *machine*, to learn from attorney instruction. This is called active machine learning, which is one application of AI.[59]

---

> discovery in a diligent and candid manner. Their combined duty is to strive in the best interests of their clients to achieve the best results at a reasonable cost, with integrity and candor as officers of the court. 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.

The Sedona Conference, *The Sedona Conference Cooperation Proclamation*, 10 SEDONA CONF. J. 331, 331 (2009 Supp.).

The following cases also adopted the Cooperation Proclamation (or espoused similar principles). Capitol Records, Inc. v. MP3Tunes, LLC, 261 F.R.D. 44, 47–48 (S.D.N.Y. 2009); *In re* Direct Sw., Inc., Fair Labor Standards Act (FLSA) Litig., No. 2:08-cv-01984-MLCF-SS, 2009 WL 2461716, at *1 (E.D. La. Aug. 7, 2009); Wells Fargo Bank, N.A. v. LaSalle Bank Nat'l. Ass'n, No. 3:07-cv-449, 2009 WL 2243854, at *2 (S.D. Ohio July 24, 2009); Dunkin' Donuts Franchised Rests. v. Grand Cent. Donuts, Inc., No. CV 2007-4027(ENV)(MDG), 2009 WL 1750348, at *4 (E.D.N.Y. June 19, 2009); Ford Motor Co. v. Edgewood Props., Inc., 257 F.R.D. 418, 424, 426 (D.N.J. 2009); Newman v. Borders, Inc., 257 F.R.D. 1, 3 n.3 (D.D.C. 2009); Gipson v. Sw. Bell. Tel. Co., No. 2:08-cv-2017-EFM-DJW, 2009 WL 790203, at *20–21 (D. Kan. Mar. 24, 2009); William A. Gross Constr. Assocs. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 136 (S.D.N.Y. 2009); S.E.C. v. Collins & Aikman Corp., 256 F.R.D. 403, 415 (S.D.N.Y. 2009); Covad Commc'ns Co. v. Revonet, Inc., 254 F.R.D. 147, 149, 151 (D.D.C. 2008); Aguilar v. Immigration & Customs Enforcement, 255 F.R.D. 350, 359, 364 (S.D.N.Y. 2008); Mancia v. Mayflower Textile Servs. Co., 253 F.R.D. 354, 363–65 (D. Md. 2008); *see also* David J. Waxse, *Cooperation—What Is It and Why Do It?*, 18 RICH. J.L. & TECH., no. 3, art. 8, Spring 2012, at 5–6, http://jolt.richmond.edu/v18i3/article8.pdf. *But see* Bill E. Boie, *The Non-Cooperation Proclamation*, E-DISCOVERY TEAM (Oct. 25, 2009, 6:26 PM), http://e-discoveryteam.com/2009/10/25/the-non-cooperation-proclamation/.

Finally, consider a Seventh Circuit Committee's conclusion on this point: "An attorney's zealous representation of a client is not compromised by conducting discovery in a cooperative manner. The failure of counsel or the parties to litigation to cooperate in facilitating and reasonably limiting discovery requests and responses raises litigation costs and contributes to the risk of sanctions." SEVENTH CIRCUIT ELEC. DISCOVERY COMM., *supra* note 56, at 6.

[59] *See* Andrew Peck, *Search, Forward: Will Manual Document Review and Keyword Searches Be Replaced by Computer-Assisted Coding?*, L. TECH. NEWS, Oct. 2011, at 25, 29.

### *A. Active Machine Learning Explained*

In active machine learning, the machine learns in an interactive process with a human, preferably an attorney with special subject matter expertise[60] on the issues in the case. The machine learns from the subject matter expert ("SME") how documents in a particular case should be classified, such as relevant or irrelevant, privileged or nonprivileged. The machine extrapolates the input provided by the SME on a small subset of documents to (1) classify the complete collection, and (2) rank the probability of each document fitting into the classification.

In active machine learning, the SME's thinking and analysis is transferred to the computer where it is improved and enhanced through AI by the computer's own analysis of the documents.[61] The machine learning happens in a series of iterative steps where the SME confirms some of the computer's correct classifications and rankings and corrects some of its initial mistakes.[62] The human SME's intent is clarified and applied through the classification of repeated selections of new document subsets. The computer analysis includes not only the content of the documents but also the metadata.[63] The documents can be selected in

---

[60] Subject matter experts, known under the well-known acronym, "SME," are always preferred for any machine instruction based on another well-known principle and acronym, "GIGO," garbage in garbage out. *See* Ralph Losey, *Three-Cylinder Multimodal Approach to Predictive Coding*, E-DISCOVERY TEAM (Mar. 24, 2013, 9:04 PM), http://e-discoveryteam.com/2013/03/24/three-cylinder-multimodal-approach-to-predictive-coding/; *see also* Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 183–84, 192 & n.14 (S.D.N.Y. 2012) (Peck., Mag. J.), *aff'd*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).

[61] For insights into the mathematics behind machine learning and document classification, see JASON R. BARON & JESSE B. FREEMAN, COOPERATION, TRANSPARENCY, AND THE RISE OF SUPPORT VECTOR MACHINES IN E-DISCOVERY: ISSUES RAISED BY THE NEED TO CLASSIFY DOCUMENTS AS EITHER RESPONSIVE OR NONRESPONSIVE (2013), *available at* http://www.umiacs.umd.edu/~oard/desi5/additional/Baron-Jason-final.pdf.

[62] For a detailed, eighty-two page narrative description of an active-machine-learning-review project of 699,082 documents that was completed after five iterative steps, see RALPH C. LOSEY, PREDICTIVE CODING NARRATIVE: SEARCHING FOR RELEVANCE IN THE ASHES OF ENRON (2012), *available at* http://ralphlosey.files.wordpress.com/2013/04/predictive-coding-narrative_corrrected_3-21-13.pdf. For a description of the same search project that used slightly different search methods taking fifty iterations to complete in about the same time (52 hours), see Ralph Losey, *Borg Challenge: The Complete Report*, E-DISCOVERY TEAM (Apr. 18, 2013, 7:02 PM) [hereinafter Losey, *Borg Challenge*], http://e-discoveryteam.com/2013/04/18/borg-challenge-the-complete-report/. The latter source reports on my experimental review of 699,082 Enron documents using a semi-automated monomodal methodology, and is a five-part written and video series comparing two different kinds of predictive coding search methods.

[63] Douglas W. Oard & William Webber, *Information Retrieval for E-Discovery*, 7 FOUND. & TRENDS IN INFO. RETRIEVAL 99, § 3.3, at 129–35 (2013). This article is discussed and quoted at length in Ralph Losey, *The Many Types of Legal Search Software in the CAR*

three ways: (1) by the computer, (2) by the SME based on his or her judgmental sampling, and (3) by random chance.

1. Machine-Selected Sampling: In this key AI-based method, the computer selects documents for its own training. The selection is made from documents that the software classifier is uncertain of the correct classification. This typically involves documents ranked in the 40% to 60% probable relevant range.
2. Judgmental Sampling: This method includes in the training all other relevant documents that the skilled reviewer can find using a variety of search techniques. That may include some linear review of selected custodians or dates, parametric Boolean keyword searches, similarity searches of all kinds, concept searches, as well as several unique predictive coding probability searches. I call that a *multimodal approach*. The judgmental sampling will typically also include irrelevant documents.
3. Random Sampling: Some reasonable percentage of the documents presented for human review is selected at random. This helps maximize recall and premature focus on the relevant documents initially retrieved.[64]

Although documents can be selected for active machine learning in these three ways, some predictive coding review methods rely on some of the methods more than others, and some only use one or two of the methods and not all three.[65] Other experts in the field[66] and I[67] promote the use of all three but with only minimal reliance on the use of random chance for selection of training documents.

Information retrieval scientists Doug Oard and William Webber call this iterative process of machine learning, "Learning From Examples," and note that it requires both positive and negative input; in other

---

*Market Today*, E-DISCOVERY TEAM (Mar. 3, 2013, 8:39 PM), http://e-discoveryteam.com/2013/03/03/the-many-types-of-legal-search-software-in-the-car-market-today/.

[64] *See* CHRISTOPHER D. MANNING ET AL., INTRODUCTION TO INFORMATION RETRIEVAL, § 15.3, at 307–13 (2008) (examining the choice between the methods of classification); Oard & Webber, *supra* note 63, § 3.5, at 137 (discussing classification in e-discovery).

[65] *See* Losey, *supra* note 60.

[66] Jeremy Pickens, *Predictive Ranking: Technology Assisted Review Designed for the Real World*, E-DISCOVERY SEARCH BLOG (Feb. 1, 2013), http://www.catalystsecure.com/blog/2013/02/predictive-ranking-technology-assisted-review-designed-for-the-real-world/; J. William Speros, *Predictive Coding's Erroneous Zones Are Emerging Junk Science*, E-DISCOVERY TEAM (Apr. 28, 2013, 8:43 PM), http://e-discoveryteam.com/2013/04/28/predictive-codings-erroneous-zones-are-emerging-junk-science/.

[67] Losey, *supra* note 60 ("The exact mixture of the three types of [predictive coding search engine] cylinders—random, analytic, and judgmental—is where the *art of predictive coding search* comes in.").

words, examples of both relevant and irrelevant documents are required for proper training.[68] This kind of AI-enhanced legal review is typically described today in legal literature by the term *predictive coding*.[69] This is because the computer predicts how an entire body of documents should be coded (classified) based on how the lawyer has coded the smaller training sets.[70] The prediction places a probability ranking on each document, typically ranging from 0% to 100% probability. Thus, in a relevancy classification, each and every document in the entire dataset (the *corpus*) is ranked with a percentage of likely relevance and irrelevance. For example, a document could be ranked as 90% probable relevant and 90% probable irrelevant. They are not always ranked exactly synonymously as you might expect. In other words, a document could be ranked 90% probable relevant and 80% probable irrelevant. Typically, when searching for relevant documents, the focus is on relevancy ranking, and the counter-ranking on irrelevance prediction is given less weight.

If the predictive coding software ranks a document as having more than a 50% chance of probable relevance, then the software is predicting that it should be coded as relevant. For instance, in a million-document corpus, the software could, typically after several rounds of machine training, rank 100,000 documents as having a 50% or higher likelihood of relevance. You can then evaluate the ranking breakdown into any

---

[68] *See* Oard & Webber, *supra* note 63, § 3.4.2, at 136–37.

[69] These two terms, *predictive coding* and *machine learning*, will be used interchangeably in this article, along with the term *artificial intelligence* or *AI*, to refer to the same use of active machine learning. Note that there is a different type of *inactive* or *automatic machine learning* that is not intended to be included in this discussion. *See* Peck, *supra* note 59, at 26, 29.

[70] The RAND Report contains a helpful description of predictive coding:

> Predictive coding, sometimes referred to as *suggestive coding*, is a process by which the computer does the heavy lifting in deciding whether documents are relevant, responsive, or privileged. This process is not to be confused with keyword-based Boolean searches or the similarity-detection technologies described in Chapter Four. Near-duplication techniques, clustering, and email threading can help provide organizational structure to the corpus of documents requiring review but do not reduce the document set that has to be reviewed by attorneys for specific aspects, such as responsiveness or privilege. Predictive coding, on the other hand, takes the very substantial next step of automatically assigning a rating (or *proximity score*) to each document to reflect how close it is to the concepts and terms found in examples of documents attorneys have already determined to be relevant, responsive, or privileged. This assignment becomes increasingly accurate as the software continues to learn from human reviewers about what is, and what is not, of interest. This score and the self-learning function are the two key characteristics that set predictive coding apart from less robust analytical techniques.

RAND REPORT, *supra* note 43, at 59.

range you want. For instance, you could see that 75,000 of those 100,000 probable relevant documents were ranked as 90% or higher probable relevant. Documents in the 40% to 50% probable relevant range are ones where the algorithmic classifier is uncertain. Typically when the software itself selects documents for its own training, it selects documents that are within this uncertainty range.

As will be shown, this ranking feature is key to the use of the legal doctrine of proportionality. The ability to rank all documents in a corpus on probable relevance is a new feature that no other legal search software has previously provided.[71]

Predictive coding is one of several types of Technology Assisted Review ("TAR"), also known as Computer Assisted Review ("CAR"), in the market today. TAR is formally defined in the Grossman-Cormack Glossary of Technology-Assisted Review ("Grossman-Cormack Glossary") as follows:

> A process for Prioritizing or Coding a Collection of Documents using a computerized system that harnesses human judgments of one or more Subject Matter Expert(s) on a smaller set of Documents and then extrapolates those judgments to the remaining Document Collection. Some TAR methods use Machine Learning Algorithms to distinguish Relevant from Non-Relevant Documents, based on Training Examples Coded as Relevant or Non-Relevant by the Subject Matter Experts(s) [sic], while other TAR methods derive systematic Rules that emulate the expert(s)' decision-making process. TAR processes generally incorporate Statistical Models and/or Sampling techniques to guide the process and to measure overall system effectiveness.[72]

---

[71] Although many pre-predictive coding software programs would purport to rank documents, the ranking was not very reliable and did not include probabilities. Instead, it was merely indicative of the number of keywords in a search that were found in a document. The documents, then, were displayed in descending order of hit counts. This was useful to some extent, but it typically just showed the larger documents on top, as they usually had the higher hit counts. Also, this would only sometimes have any correlation with actual relevance. Although some software corrected for document size, the ranking still was just based on keyword hit counts, and this was often unreliable. Moreover, even with predictive coding software today, there seems to be a wide variance in the quality of ranking functions, and only a few programs now on the market do it well. Even with good AI-enhanced software, the ranking functions are very sensitive to the quality of the input, and knowledgeable SME input is required. Even then, it is not an exact measure of relevancy weight. Testing and quality controls should always be applied to know when and to what degree the ranking strata are reliable. Ralph Losey, Relevancy Ranking is the Key Feature of Predictive Coding Software, E-DISCOVERY TEAM (Aug. 25, 2013, 8:54 PM), http://e-discoveryteam.com/2013/08/25/relevancy-ranking-is-the-key-feature-of-predictive-coding-software/.

[72] *Grossman-Cormack Glossary*, *supra* note 41, at 32 (defining TAR).

As the definition indicates, some TAR methods use pattern recognition algorithms to harness the judgment of lawyers, and others do not. They instead use what are known as "rule-based" methods that rely on teams of human linguistic experts to design complex rules. Such rule-based work is labor-intensive and thus expensive.[73] Rule-based TARs are not a form of AI and are not included in this article as a type of active machine learning.[74] Still, the rule-based methods can also rank all documents in a corpus and can be effective. Thus, they can also be useful in proportionality-based document reviews, especially where the attorneys are not capable of performing machine-based active learning or otherwise prefer to delegate and depend on outside experts.

### *B. Predictive Coding Case Law*

The RAND Report concluded that the "key reason" for the slow adoption of predictive coding by the legal profession was "the absence of widespread judicial approval of the methodology."[75] Since then, several reported[76] decisions have come out with just the kind of judicial approval that the RAND Report said the profession needed. It all started with the opinion on February 24, 2012, by the leading judicial scholar on predictive coding, United States Magistrate Judge Andrew J. Peck of the Southern District of New York, in *Da Silva Moore v. Publicis Groupe*.[77] Judge Peck's opinion was discussed in the Report, but it was not affirmed and approved by the district court judge until after the Report's publication.[78]

---

[73] *See, e.g.*, Gabriel Techs. Corp. v. Qualcomm Inc., No. 08cv1992 AJB (MDD), 2013 WL 410103, at *10 (S.D. Cal. Feb. 1, 2013) (noting expenditure by parties of $2,829,349.10 for first-pass classification using rule-based technology to classify one million documents).

[74] *See Grossman-Cormack Glossary*, *supra* note 41, at 8 (defining *Active Learning* as "[a]n Iterative Training regimen in which the Training Set is repeatedly augmented by additional Documents chosen by the Machine Learning Algorithm, and coded by one or more Subject Matter Expert(s)"); *id.* at 28 (defining *Rule Base* as "[a] set of Rules created by an expert to emulate the human decision-making process for the purposes of Classifying Documents in the context of Electronic Discovery").

[75] RAND REPORT, *supra* note 43, at 98.

[76] I am sure there are many more unreported decisions, as I have been personally involved in at least one—a large arbitration proceeding.

[77] 287 F.R.D. 182 (S.D.N.Y. 2012) (approving use of predictive coding and listing justifications), *aff'd*, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).

[78] *Da Silva Moore*, 2012 WL 1446534, at *3 (affirming *Da Silva Moore*, 287 F.R.D. 182); RAND REPORT, *supra* note 43, at 78–80 (discussing *Da Silva Moore*); Press Release, RAND Corp., Predictive Coding Could Reduce E-Discovery Costs, but More Guidance Needed on Data Preservation (Apr. 11, 2012), *available at* http://www.rand.org/news/press/2012/04/11.html (announcing the release of the RAND Report). Interestingly, the final effect of this opinion was delayed for a year pending the plaintiffs' attempt to disqualify

Multiple other decisions and widely published hearings came in quick order after that.[79] One judge went so far as to *sua sponte* order both sides to use predictive coding and share the same vendor to save costs.[80] Others have considered the possible use of predictive coding technologies to make review less burdensome as a factor in rejecting protective orders.[81]

---

presiding Magistrate Judge Andrew Peck. *See* Da Silva Moore v. Publicis Groupe, 868 F. Supp. 2d 137, 140 (S.D.N.Y. 2012) (denying plaintiffs' motion for recusal), *cert. denied*, No. 13-51, 2013 WL 3489452 (U.S. Oct. 7, 2013); *see also* Motion Order, Da Silva Moore v. Publicis Groupe (*In re* Da Silva Moore), No. 12-05020, (2d Cir. Apr. 10, 2013), ECF No. 16 (denying petition to compel recusal of Judge Peck). Both Magistrate Judge Peck and the Second Circuit Court of Appeals rejected plaintiffs' arguments that voicing public support for ESI or appearing on a CLE panel with a lawyer constituted grounds for recusal or disqualification from this case. *See Da Silva Moore*, 868 F. Supp. 2d at 164; *In re Da Silva Moore*, No. 12-05020.

[79] *See* Gordon v. Kaleida Health, No. 08-CV-378S(F), 2013 WL 2250579, at *1 (W.D.N.Y. May 21, 2013) (referencing a judge's suggestion that the parties use predictive coding based on Judge Peck's opinion in *Da Silva Moore* and the parties' disagreement over methodology); *In re* Biomet M2A Magnum Hip Implant Prods. Liab. Litig., No. 3:12-MD-2391, 2013 WL 1729682, at *1, *3 (N.D. Ind. Apr. 18, 2013) (approving a multimodal predictive coding approach); Kleen Prods. LLC v. Packaging Corp. of Am., No. 10 C 05711, 2012 WL 4498465, at *5 (N.D. Ill. Sept. 28, 2012) (referencing a multi-day evidentiary hearing on plaintiffs' motion to compel use of predictive coding); *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2012 WL 7861249, at *1, *3–4 (W.D. La. July 27, 2012) (approving the use of predictive coding); Order Approving the Use of Predictive Coding for Discovery, Global Aerospace Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012).

[80] *See* Order Granting Partial Summary Judgment, EORHB, Inc. v. HOA Holdings LLC, No. 7409-VCL, 2012 WL 4896670 (Del. Ch. Oct. 15, 2012). Seven months later the judge backed off that order somewhat when the plaintiff showed good cause for not using predictive coding and sharing a vendor, but defendants complied and used predictive coding for their review:

> [F]or good cause shown, it is hereby ORDERED that: (i) Defendants may retain ediscovery vendor Kroll OnTrack and employ Kroll OnTrack and its computer assisted review tools to conduct document review; (ii) Plaintiffs and Defendants shall not be required to retain a single discovery vendor to be used by both sides; and (iii) Plaintiffs may conduct document review using traditional methods.

EORHB, Inc. v. HOA Holdings LLC, No. 7409-VCL, 2013 WL 1960621 (Del. Ch. May 6, 2013). The Court's good-cause analysis was primarily driven by an agreement among the parties that the cost of using predictive coding in this case would be outweighed by an expected low volume of relevant documents subject to discovery from the plaintiff. *Id.*

[81] *See* Chevron Corp. v. Donziger, No. 11 Civ. 0691(LAK), 2013 WL 1087236, at *32 & n.255 (S.D.N.Y. Mar. 15, 2013) (noting the potential effectiveness of predictive coding in reducing the burden of discovery); Harris v. Subcontracting Concepts, LLC, Civ. No. 1:12-MC-82 (DNH/RFT), 2013 WL 951336, at *5 (N.D.N.Y. Mar. 11, 2013) (stating that predictive coding and other technologies reduce the cost and time of producing large numbers of documents).

Most commentators agree that the main case in this area remains the first: *Da Silva Moore*.[82] The explanations, legal analysis, and detailed protocols provided in the opinion,[83] coupled with Judge Peck's reputation in the field, are a strong influence on other judges hearing the issue for the first time.[84] Here are a few illustrative excerpts from Judge Peck's opinion:

> In this case, the Court determined that the use of predictive coding was appropriate considering: (1) the parties' agreement, (2) the vast amount of ESI to be reviewed (over three million documents), (3) the superiority of computer-assisted review to the available alternatives (*i.e.*, linear manual review or keyword searches), (4) the need for cost effectiveness and proportionality . . . , and (5) the transparent process proposed by [Defendant].
>
> This Court was one of the early signatories to The Sedona Conference Cooperation Proclamation, and has stated that "the best solution in the entire area of electronic discovery is cooperation among counsel. . . ." An important aspect of cooperation is transparency in the discovery process. [Defendant's] transparency in its proposed ESI search protocol made it easier for the Court to approve the use of predictive coding. . . . [Defendant] confirmed that "[a]ll of the documents that are reviewed as a function of the seed set, whether [they] are ultimately coded relevant or irrelevant, aside from privilege, will be turned over to" plaintiffs. . . . ["]If necessary, counsel will meet and confer to attempt to resolve any disagreements regarding the coding applied to the documents in the seed set."[] While not all experienced ESI counsel believe it necessary to be as transparent as [Defendant] was willing to be, such transparency allows the opposing counsel (and the Court) to be more comfortable with computer-assisted review, reducing fears about the so-called "black box" of the technology. This Court highly recommends that counsel in future cases be willing to at least discuss, if not agree to, such transparency in the computer-assisted review process.[85]

---

[82] *See, e.g.*, Jacob Tingen, *Technologies-That-Must-Not-Be-Named: Understanding and Implementing Advanced Search Technologies in E-Discovery*, 19 RICH. J.L. & TECH., no. 1, art. 2, Fall 2012, at 11, 13, http://jolt.richmond.edu/v19i1/article2.pdf.

[83] An appendix to the February 24, 2012, *Da Silva Moore* opinion sets forth a detailed protocol that included (1) provisions for seed sets of documents generated through a combination of sampling methods, (2) up to seven iterative rounds of "training" the system, (3) a commitment by counsel to share both responsive and nonresponsive documents, and (4) sampling at the end of the initial training to function as a quality assurance check on excluded or irrelevant documents. *Da Silva Moore*, 287 F.R.D. app. at 199–202.

[84] *Da Silva Moore* is still an active case and my law firm is lead counsel for the defense on these issues; therefore, I do not comment on the case itself, but only provide these quotes.

[85] *Da Silva Moore*, 287 F.R.D. at 192 (fourth and fifth alterations in original) (footnote omitted) (citations omitted).

Many articles have been written subsequent to *Da Silva Moore* detailing the legal and scientific support now available for the use of predictive coding in legal search projects.[86] Judge Shira A. Scheindlin, who is perhaps the most influential judge in the e-discovery area as the author of the *Zubulake* opinions, a group of influential e-discovery cases,[87] has also joined in to approve and encourage the use of predictive coding.[88] Although the issue was not directly before her, her words in dicta are still influential:

> There are emerging best practices for dealing with these shortcomings [referring to keyword search] and they are explained in detail elsewhere. There is a "need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or 'keywords' to be used to produce emails or other electronically stored information." And beyond the use of keyword search, parties can (and frequently should) rely on latent semantic indexing, statistical probability models, and machine learning tools to find responsive documents. Through iterative learning, these methods (known as "computer-assisted" or "predictive" coding) allow humans to teach computers what documents are and are not responsive to a particular FOIA or discovery request and they can significantly increase the effectiveness and efficiency of searches. In short, a review of the literature makes it abundantly clear that a court cannot simply trust the defendant agencies' unsupported assertions that their lay custodians have designed and conducted a reasonable search.[89]

The main issue of debate currently focuses on the degree of disclosure that a court should require for use of the new technology, an issue Judge Peck specifically referenced in the earlier-quoted paragraph

---

[86] *See* RAND REPORT, *supra* note 43; *see also* Nicholas Barry, Note, *Man Versus Machine Review: The Showdown Between Hordes of Discovery Lawyers and a Computer-Utilizing Predictive-Coding Technology*, 15 VAND. J. ENT. & TECH. L. 343 (2013); Elle Byram, *The Collision of the Courts and Predictive Coding: Defining Best Practices and Guidelines in Predictive Coding for Electronic Discovery*, 29 SANTA CLARA COMPUTER & HIGH TECH. L.J. 675 (2013); Charles Yablon & Nick Landsman-Roos, *Predictive Coding: Emerging Questions and Concerns*, 64 S.C. L. REV. 633 (2013).

[87] Zubulake v. UBS Warburg LLC, 229 F.R.D. 422 (S.D.N.Y. 2004); Zubulake v. UBS Warburg LLC, 220 F.R.D. 212 (S.D.N.Y. 2003); Zubulake v. UBS Warburg LLC, 216 F.R.D. 280 (S.D.N.Y. 2003); Zubulake v. UBS Warburg LLC, 217 F.R.D. 309 (S.D.N.Y. 2003); *see also* Elaine Ki Jin Kim, Comment, *The New Electronic Discovery Rules: A Place for Employee Privacy?*, 115 YALE L.J. 1481, 1484 (2006) (stating that "*Zubulake* has had an impact far beyond the Southern District of New York" and that it is "influencing courts in other jurisdictions").

[88] *See* Nat'l Day Laborer Org. Network v. U.S. Immigration & Customs Enforcement Agency, 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012).

[89] *Id.* at 109–10 (footnotes omitted).

of his *Da Silva Moore* opinion.[90] Some argue for complete transparency and full disclosure as required in *Da Silva Moore*,[91] but others assert that no disclosure should be required and that everything should be protected as work product.[92]

I have taken a compromise position on the issue of disclosure in the past, arguing that keywords and search methods should be disclosed, but not the actual irrelevant documents, even if they were used as training documents.[93] I later revised my position on this issue somewhat to allow for limited disclosure of irrelevant documents when the SME considers them to be borderline-type documents.[94] Analysis of my Enron review experiment showed that inconsistencies by a single SME of these types of borderline documents occur at least 23% of the time, whereas inconsistencies in coding of all other irrelevant documents are extremely rare.[95]

---

[90] *See Da Silva Moore*, 287 F.R.D. at 192. For articles engaging in the debate, see Ronni Solomon, *Are Corporations Ready To Be Transparent and Share Irrelevant Documents with Opposing Counsel To Obtain Substantial Cost Savings Through the Use of Predictive Coding?*, METRO. CORP. COUNS., Nov. 2012, at 26; WILLIAM P. BUTTERFIELD, CONOR R. CROWLEY & JEANNINE KENNEY, REALITY BITES: WHY TAR'S PROMISES HAVE YET TO BE FULFILLED 8 (2013), *available at* http://www.umiacs.umd.edu/~oard/desi5/additional/Butterfield.pdf.

[91] BARON & FREEMAN, *supra* note 61, at 16.

[92] *See, e.g.*, Transcript of Discovery Dispute Hearing at 16, Robocast Inc. v. Apple Inc., No. 1:11-cv-00235-RGA (D. Del. Dec. 5, 2012), ECF No. 99; *Waiving Work Product with Predictive Coding*, ESIBYTES PODCAST (Sept. 17, 2012), http://www.esibytes.com/waiving-work-product-with-predictive-coding/ (recording of Karl Schieneman's interview of attorney Jeff Fowler). In *Robocast*, Judge Richard G. Andrews recognized that there was no more reason to require disclosure where documents were excluded by predictive coding than there would be to require disclosure of a sample of documents deemed nonresponsive as a result of linear review: "[W]hy isn't that something—you know, you answered their discovery however you answered it—why isn't it something where they answer your discovery however they choose to answer it, complying with their professional obligations? How do you get to be involved in the seed batch?" Transcript of Discovery Dispute Hearing, *Robocast*, *supra* at 16. The anti-disclosure arguments in predictive coding are an extension of an earlier argument opposing the disclosure of search terms in keyword searches. *See* David J. Kessler, Robert D. Owen & Emily Johnston, *Search Terms Are More Than Mere Words*, N.Y. L.J. (Mar. 21, 2011).

[93] Ralph Losey, *Keywords and Search Methods Should Be Disclosed, But Not Irrelevant Documents*, E-DISCOVERY TEAM (May 26, 2013, 4:44 PM), http://e-discoveryteam.com/2013/05/26/keywords-and-search-methods-should-be-disclosed-but-not-irrelevant-documents/.

[94] Ralph Losey, *A Modest Contribution to the Science of Search: Report and Analysis of Inconsistent Classifications in Two Predictive Coding Reviews of 699,082 Enron Documents*, E-DISCOVERY TEAM (June 11, 2013, 9:13 AM), http://e-discoveryteam.com/2013/06/11/a-modest-contribution-to-the-science-of-search-report-and-analysis-of-inconsistent-classifications-in-two-predictive-coding-reviews-of-699082-enron-documents/.

[95] Indeed, I provided a more detailed explanation:

The studies on inconsistent SME document classifications suggest that machine training can be made more reliable if clarifications are obtained on these borderline documents before machine training, analysis, and ranking are concluded.[96] This can be done by dialogue with opposing counsel where the types of documents under consideration are discussed without actually revealing the documents themselves.[97] Alternatively, limited disclosure can be made of the documents under special confidentiality restrictions or by in-camera submissions to the presiding judge.[98] This compromise position should address the legitimate confidentiality concerns of producing parties and still provide assurances to the requesting party that the AI has been properly trained to find the documents.

---

> The *inconsistencies* (opposite of *Jaccard* index) shown in this study of determinations of relevance, and excluding the classifications of irrelevant, were relatively small—23%, as compared to 55%, 70% and 84% in prior studies. Moreover, as mentioned, they were all derived from grey area or borderline type documents, where relevancy was a matter of interpretation. In the author's experience documents such as this tend to have low probative value. If they were significant to litigation discovery, then they usually would not be of a grey area, subjective type. They would instead be obviously relevance [sic]. I say *usually* because the author has seen rare exceptions, typically in situations where one borderline document leads to other documents with strong probative value. Still, this is unusual. In most situations the omission of borderline ambiguous documents, and others like them, would have little or no impact on the case.
>
> These observations, especially the high consistency of irrelevance classifications (98%+), support the strict limitation of disclosure of irrelevant documents as part of a cooperative litigation discovery process. Instead, only documents that a reviewer knows are of a grey area type or likely to be subject to debate should be disclosed. (The SME in this study was personally aware of the ambiguous type grey area documents when originally classifying these documents. They were obvious because it was difficult to decide if they were within the border of relevance, or not. The ambiguity would trigger an internal debate where a close question decision would ultimately be made.)
>
> Even when limiting disclosure of irrelevant documents to those that are known to be borderline, disclosure of the actual documents themselves may frequently not be necessary. A summary of the documents with explanation of the rationale as to the ultimate determination of irrelevance should often suffice. The disclosure of a description of the borderline documents will at least begin a relevancy dialogue with the requesting party. Only if the abstract debate fails to reach agreement would disclosure of the actual documents be required.

*Id.*

[96] *See, e.g.*, JIANLIN CHENG ET AL., SOFT LABELING FOR MULTI-PASS DOCUMENT REVIEW 10 (2013), *available at* http://www.umiacs.umd.edu/~oard/desi5/research/Cheng-final.pdf.

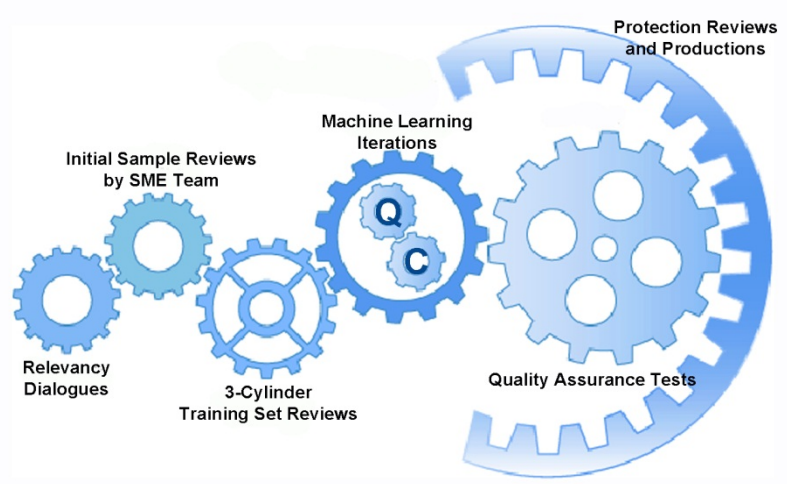
[97] Losey, *supra* note 94.

[98] *Id.*

### *C. Six-Step and Eight-Step Predictive Coding Work Flows*

Relevancy dialogues between the legal counsel for the requesting and responding parties are needed not only during the review itself for clarification of borderline documents, but also at the beginning of the review to clarify the basic *information need* to be fulfilled by the predictive coding search. The use of written requests for production with category lists is more of a starting point in a cooperative process, rather than the final word on the documents requested. That is why all of my models for predictive-coding-based document search begin with communications among all of the parties. The first model I use to describe the predictive coding process divides the work flow into six steps and is illustrated by the diagram below.

Diagram 1: Six-Step Predictive Coding Work Flow[99]




The first step is *Relevancy Dialogues* with opposing counsel. This is based on a cooperative approach to discovery required by both the rules of procedure and the rules of ethics.[100] The primary goal of these dialogues for predictive coding purposes is to clarify the e-discovery requests and reach agreement on the scope of relevancy and production. Searches depend upon the clarity of your information need.[101] Additional

---

[99] Copyright © Ralph Losey. The gears in the diagram indicate the interlocking nature of the ESI production processes used with predictive coding. In the next section the same process will be described in slightly greater detail using eight steps.

[100] *See* sources cited *supra* notes 55, 57–58.

[101] STEFAN BÜTTCHER, CHARLES L. A. CLARKE & GORDON V. CORMACK, INFORMATION RETRIEVAL: IMPLEMENTING AND EVALUATING SEARCH ENGINES § 1.2.1, at 5 (2010).

conferences to make disclosures designed to protect clients' interests are also sometimes needed for appropriate training and quality controls.

The additional disclosures will typically require some sharing of some of the ESI search techniques actually used, which is traditionally protected as work product. The disclosures may also sometimes include limited disclosure of some of the seed set documents used, both relevant and irrelevant.[102] Nothing in the rules requires disclosure of irrelevant ESI,[103] but if adequate privacy protections are provided, it may be in the best interests of all parties to do so. Such discretionary disclosures may be advantageous both for risk mitigation and efficiency (cost savings). If an agreement on search protocol is reached by the parties or imposed by the court, the parties are better protected from the risk of expensive motion practice and repetitions of discovery search and production.

Step two is *Initial Sample Reviews by the SME Team*. The use of SMEs is a critical aspect of predictive coding review. The samples reviewed are both random samples and judgmental samples. Judgmental samples use all of the various pre-predictive coding legal search methods, including parametric Boolean keyword searches, similarity searches, concept searches, and even strategic linear reviews of the documents of select custodians and date ranges. The random samples broaden the search and also make possible various types of random-sample-based statistical analysis. For instance, the random sample can provide a baseline of calculation of the prevalence of relevant documents in the corpus. This is very helpful for quality control purposes. The review by the SMEs of random and judgmental samples provides the first machine training input for the predictive coding software. The first round of machine training is also sometimes called the initial *Seed Set Build*.[104]

At the commencement of the project, but using different documents selected by random sample, good predictive coding software will also create what is called a *control set*. The SMEs code documents in both the first training set and the control set, and they may be unaware which of the documents selected by random for them to code are designated for

---

[102] *See* Oard & Webber, *supra* note 63, § 4.2.1, at 160–61.

[103] *See* FED. R. CIV. P. 26(b)(1) (limiting all discovery, electronic and otherwise, to relevant information).

[104] *Grossman-Cormack Glossary*, *supra* note 41, at 29 (defining *Seed Set* as "[t]he initial Training Set provided to the learning Algorithm in an Active Learning process. The Documents in the Seed Set may be selected based on Random Sampling or Judgmental Sampling. Some commentators use the term more restrictively to refer only to Documents chosen using Judgmental Sampling. Other commentators use the term generally to mean any Training Set, including the final Training Set in Iterative Training, or the only Training Set in non-Iterative Training.").

the control set as opposed to the initial training set. Typically, both the control and random documents used in the seed set are part of the first random sample. The control set is used solely for testing the SMEs' work during the iterative training process. It is not used for training. The control set documents test for SME consistency and for overall training effectiveness. The documents marked by the SMEs for the control set cannot also be used for training as this can introduce bias into the testing.

The next step, *Three-Cylinder Training Set Reviews*, continues the machine training in an iterative process. In an active-learning type of predictive coding, the machine selects documents for which it would like input. Typically that is a selection of documents whose classification is uncertain. As mentioned, two other document selection methods are also used—SME judgmental sampling and random sampling.

*Machine Learning Iterations*, the fourth step, is where the software takes the input provided by the SME team and extrapolates and applies those classifications to the entire collection of documents. The predictive coding software then ranks all documents in the collection from 100% probable relevant to 0% probable relevant. This key predictive coding step is repeated as needed for quality control purposes. These iterations continue until the training is complete within the proportional constraints of the case. At that point, the SME in charge of the search may declare the search complete and ready for the next quality assurance test.

In the fifth step, *Quality Assurance Tests* are based primarily on random sampling to verify the effectiveness of the final rankings. They are used to verify the reasonableness, or unreasonableness, of the search and the predictive coding parameters developed. If the tests are not passed, the review is reactivated for additional rounds of SME review and machine learning iterations. If the review passes the tests, then the first-pass relevancy review is complete, and the project moves to the next and final step.[105]

*Protection Reviews and Productions* is the last step. It comes only after the Quality Assurance Tests have been satisfied. Predictive coding

---

[105] For more on sampling and its use in both quality control and quality assurance, see Ralph Losey, *Review Quality Controls*, ELECTRONIC DISCOVERY BEST PRACTICES, http://www.edbp.com/search-review/review-quality-controls/ (last visited Oct. 30, 2013); see also MANFRED GABRIEL, KPMG, QUALITY CONTROL FOR PREDICTIVE CODING IN EDISCOVERY (2013), *available at* http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/quality-control-predictive-coding-ediscovery.pdf; CHRIS PASKACH ET AL., KPMG, THE CASE FOR STATISTICAL SAMPLING IN E-DISCOVERY (2012), *available at* http://www.kpmg.com/us/en/IssuesAndInsights/ArticlesPublications/Documents/case-for-statistical-sampling-e-discovery.pdf.

in this step is used to make efficient assignments to manual reviewers. They do the final reviews of the documents before production. They only review the documents that were coded relevant by the SME team in the prior steps. In this last step, SMEs are not required unless there are issues regarding any of their first-pass relevancy determinations. For instance, a SME would want to double-check any proposed promotion of a document from merely relevant to *Hot*. The SMEs would also want to double-check in some manner the proposed demotion to irrelevant of any documents predicted relevant. This final review step is done by attorneys knowledgeable about privilege and confidentiality issues in the case. In large projects, this final review is typically done by outside contract review attorneys.

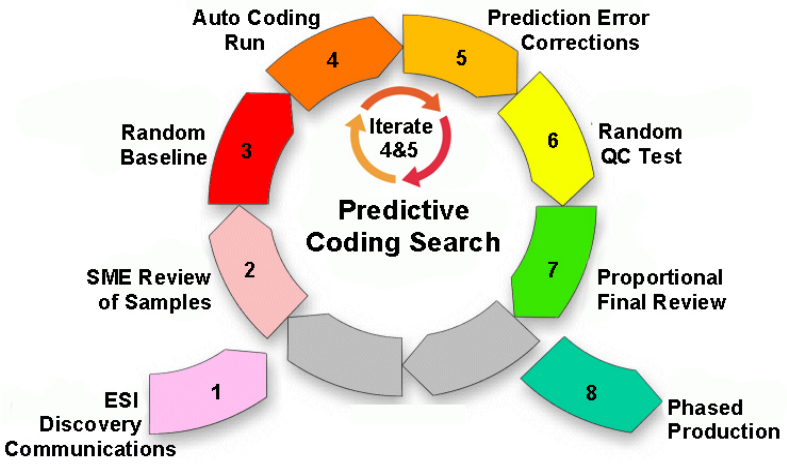
In this final step, the predicted relevant coding is confirmed, confidentiality redactions are made on the documents as needed, and privileged documents are identified and removed from production for logging. The documents culled by predictive coding, or other search methods, are culled out and not subject to expensive final manual review. So too are documents culled out that are predicted relevant but ranked below the budgeted amount under a *Bottom Line Driven Proportional Strategy* that will be discussed at the conclusion of this Article.[106] The last steps in *Final Review* are to spot-check the final production media before delivery.

The next chart is an eight-step summary of how to use predictive coding. It is still somewhat simplified, but it provides more detail than the prior six-step model. The circular flow depicts the iterative steps specific to the machine training features.

---

[106] *See infra* Part IV.B.

Diagram 2: Eight-Step Predictive Coding Work Flow[107]




In step one, the process starts with *ESI Discovery Communications,* or *Dialogues*, not only with opposing counsel or other requesting parties, but also with the client and within the e-discovery team assigned to the case. Good communications are critical to the success of all project management functions. The *ESI Discovery Communications* should be facilitated by the lead e-discovery specialist attorney assigned to the case and should include active participation by the team of trial lawyers.

In step two, the SMEs on the case (typically the partners, senior associates, and sometimes also the e-discovery specialist attorney assigned to the case), perform *Manual Reviews of Search Samples* of the data. The samples are not random but are selected by the SMEs' skilled judgments. The selections are made with the help of various software search features, including keyword search, similarities, and concept searches.

Step three in the diagram above, *Random Baseline*, is where statistically random sampling is used to establish a baseline for quality control purposes, the mentioned control set for testing. Most software also uses this random sampling selection and SME coding for initial machine training, so long as the documents do not overlay. In other words, the documents coded by the SMEs in the control sets cannot also be used in the training in the next or following steps.

Step four is the *Auto Coding Run* where the software's predictive coding calculations begin. This is also known as the first iteration of seed set training. Here the predictive coding software analyzes all of the

---

[107] Copyright © Ralph Losey 2012.

categorizations made by the SMEs in the prior steps so long as the documents were designated by them as training documents. Based on this input, the software scans all of the data uploaded onto the review platform (the corpus) and assigns a probable value of 0 to 100 to each document in the corpus. A value of 100 represents the highest probability (100%) that the document matches the category trained, such as relevant, or highly relevant. A value of 0 means no likelihood of matching, whereas 50% represents an equal likelihood. In the initial auto coding runs, the software predictions as to a document's categorization are often wrong, sometimes wildly so, depending on the kind of search and data involved. That is why spot-checking and further training are always needed for predictive coding to work properly. This is why it is an iterative process, not a *one-and-done* procedure.

Step five is where *Prediction Error Corrections* are made. Lawyers and paralegals find and correct the computer errors by a variety of methods. The predictive coding software learns from the corrections. Steps four and five then repeat as shown in the diagram. This iterative process is considered a *virtuous feedback loop* that continues until the computer predictions are accurate enough to satisfy the proportional demands of the case. This is a key point to understanding the perfect fit between proportionality and predictive coding.

Step six, *Random Quality Control* ("*QC*") *Test*, is where the reasonability of the decision to stop the training is evaluated by an objective quality control test. The test is based on a random sample of all documents to be excluded from the *Final Review* for possible production. The exclusion can be based on both category prediction (i.e., probable irrelevant) and/or probable ranking of documents with proportionate cut-offs. The focus is on a search for any false negatives (i.e., relevant documents incorrectly predicted to be irrelevant) that are hot or otherwise of significance. Perfect recall of all relevant documents is both scientifically impossible and legally unnecessary under best practices for proportional review. But the goal is to avoid all false negatives of hot documents. If this error is encountered, an additional iteration of steps four and five is usually required.

Step seven is where *Proportional Final Review* is performed and where the final decisions are made on the number of documents to be reviewed for possible production. Here the ranking feature of predictive coding makes the use of a proportionality analysis fairly easy and straightforward.[108] You only review the highest ranked documents—the

---

[108] *See* Ralph Losey, *Relevancy Ranking is the Key Feature of Predictive Coding Software*, E-DISCOVERY TEAM (Aug. 25, 2013, 8:54 PM), http://e-discoveryteam.com/2013/08/25/relevancy-ranking-is-the-key-feature-of-predictive-coding-software/.

ones most likely to be of any significance to the case—that your proportional budget allows. A decision is then made on the number of documents predicted to be relevant that fit within a reasonable budget for production. Based on prior experience, as of mid-2013, a standard cost of $1.00 to $4.00 per file is often used. Alternatively, specific calculations may be made based upon metrics gathered in that project as to what the per-document final review cost will be. This is accomplished by doing sample reviews and measuring how long the final review takes. After an agreement with the requesting party is reached or deemed unnecessary, or a court order is attained if there is disagreement, the Final Review is then completed, including redaction and logging of privileged documents. In large cases, the Final Review may be outsourced to a document review team to save time and money. The SMEs then play only a supervisory role.

Step eight, *Phased Production*, is where the documents are actually produced after a last quality control check of the media—typically CDs, DVDs, or FTP uploads—on which the production is made. The final work includes preparation of a privilege log, which is typically delayed until after production. Also, production is usually done in rolling stages as review is completed. This allows more time for an orderly process and creates good will with the requesting party and the court.

The selection of documents for training in step five uses all three selection methods (three-cylinders): judgmental sampling (multimodal search), random sampling, and machine-selected sampling. The selection of documents for the initial training (sometimes called the seed set) derives from steps two and three in the chart, with most documents in the first training round coming from step two. The second step uses only judgmental multimodal type searches. The first training round may, however, also include some documents from the random draw in step three. Steps three and six in the chart always use pure random samples and rely on statistical analysis.

My insistence on the use of multimodal judgmental sampling in steps two and five to locate relevant documents follows the consensus view of information scientists specializing in information retrieval[109] and

---

[109] *See* Marcia J. Bates, *The Design of Browsing and Berrypicking Techniques for the Online Search Interface*, 13 ONLINE INFO. REV. 407, 409–11, 414, 418, 421–22 (1989); *see also* MANNING ET AL., *supra* note 64, at 309 (explaining that a process is not a bona fide active learning search without including machine-selected sampling); GARY MARCHIONINI, INFORMATION SEEKING IN ELECTRONIC ENVIRONMENTS 5–6, 66–69 (1995); RYEN W. WHITE & RESA A. ROTH, EXPLORATORY SEARCH: BEYOND THE QUERY-RESPONSE PARADIGM 6, 15 (2009). Additionally, Professor Marcia Bates, in 2011, explained her prior article and her work on berrypicking; note the similarity to my *Multimodal* approach:

of many lawyers and courts,[110] but it is not followed by several prominent predictive coding software vendors in e-discovery. They instead rely entirely on machine selected documents for training, or even worse, on randomly selected documents to train the software. In my writings I call this process the *Borg Approach* for its overreliance on machines.[111] It unnecessarily minimizes the role of a legal expert's input and the usefulness of other types of searches to supplement an active learning process. I instead use a hybrid approach[112] where the expert reviewer remains in control of the process and his or her expertise is leveraged for greater accuracy and speed of review.

## III. PROPORTIONALITY

### *A. Origins of the Proportionality Doctrine*

The doctrine of proportionality as a legal initiative was launched by The Sedona Conference in 2010[113] as a reaction to the exploding costs of

---

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

Marcia J. Bates, *Online Search and Berrypicking*, QUORA (Dec. 21, 2011), http://www.quora.com/Marcia-J-Bates/Online-Search-and-Berrypicking/An-important-thing-we-learned-early-on-is-that-successful-searching-requires-what-I-called-berrypicking-It-is-usu-1.

[110] *See In re* Biomet M2A Magnum Hip Implant Prods. Liab. Litig., No. 3:12-MD-2391, 2013 WL 1729682, at *1, *3 (N.D. Ind. Apr. 18, 2013); Ralph Losey, *Reinventing the Wheel: My Discovery of Scientific Support for "Hybrid Multimodal" Search*, E-DISCOVERY TEAM (Apr. 21, 2013, 5:16 PM), http://e-discoveryteam.com/2013/04/21/reinventing-the-wheel-my-discovery-of-scientific-support-for-hybrid-multimodal-search; Speros, *supra* note 66.

[111] Losey, *Borg Challenge*, *supra* note 62; Ralph Losey, *Comparative Efficacy of Two Predictive Coding Reviews of 699,082 Enron Documents*, E-DISCOVERY TEAM (June 17, 2013, 9:28 AM), http://e-discoveryteam.com/2013/06/17/comparative-efficacy-of-two-predictive-coding-reviews-of-699082-enron-documents/; Ralph Losey, *Journey into the Borg Hive (Full Story Restatement),* E-DISCOVERY TEAM (Feb. 13, 2013, 8:03 AM), http://e-discoveryteam.com/2013/02/13/journey-into-the-borg-hive-full-story-restatement/.

[112] In the literature of information science, this hybrid approach is known as Human-Computer Information Retrieval (HCIR). *See* WHITE & ROTH, *supra* note 109, at 15 ("[I]nformation-seeking strategies need to be supported by system features and user interface designs, bringing humans more actively into the search process.").

[113] The Sedona Conference, *The Sedona Conference Commentary on Proportionality in Electronic Discovery,* 11 SEDONA CONF. J. 289, 292–94 (2010) [hereinafter Sedona,

e-discovery.[114] Proportionality requires the burdens of e-discovery, including production and preservation, to be reasonably balanced with the likely benefits.[115] The doctrine is intended to prevent litigants from using e-discovery and the expenses it can trigger as a weapon of extortion and a game to force inflated settlements,[116] instead of as a legitimate tool of discovery of the truth.[117]

The Sedona Commentary sets forth six principles of proportionality:

> 1. The burdens and costs of preserving potentially relevant information should be weighed against the potential value and uniqueness of the information when determining the appropriate scope of preservation.

---

*Commentary on Proportionality* (2010)]. I have been a member of The Sedona Conference since 2007. *See also* John L. Carroll, *Proportionality in Discovery: A Cautionary Tale*, 32 CAMPBELL L. REV. 455, 460 (2010) ("If courts and litigants approach discovery with the mindset of proportionality, there is the potential for real savings in both dollars and time to resolution."). *But see* Scott A. Moss, *Litigation Discovery Cannot Be Optimal but Could Be Better: The Economics of Improving Discovery Timing in a Digital Age*, 58 DUKE L.J. 889, 895–96 (2009) (criticizing proportionality limits as impossible to implement effectively).

[114] *See* RAND REPORT, *supra* note 43, at xiv, xvi.

[115] Patrick Oot, Anne Kershaw & Herbert L. Roitblat, *Mandating Reasonableness in a Reasonable Inquiry*, 87 DENV. U. L. REV. 533, 544 (2010).

[116] Maura Grossman & Gordon Cormack, *Some Thoughts on Incentives, Rules, and Ethics Concerning the Use of Search Technology in E-Discovery*, 12 SEDONA CONF. J. 89, 94–95, 101–02 (2011); Ralph Losey, *E-Discovery Gamers: Join Me in Stopping Them*, E-DISCOVERY TEAM (June 3, 2012, 6:01 AM), http://e-discoveryteam.com/2012/06/03/e-discovery-gamers-join-me-in-stopping-them/; *see also, e.g.*, Kassover v. UBS A.G., No. 08 Civ. 2753(LMM)(KNF), 2008 WL 5395942, at *3 (S.D.N.Y. Dec. 19, 2008) ("PSLRA's discovery stay provision was promulgated to prevent conduct such as: (a) filing frivolous securities fraud claims, with an expectation that the high cost of responding to discovery demands will coerce defendants to settle; and (b) embarking on a 'fishing expedition' or 'abusive strike suit' litigation."); Bondi v. Capital & Fin. Asset Mgmt. S.A., 535 F.3d 87, 97 (2d Cir. 2008) ("This Court . . . has taken note of the pressures upon corporate defendants to settle securities fraud 'strike suits' when those settlements are driven, not by the merits of plaintiffs' claims, but by defendants' fears of potentially astronomical attorneys' fees arising from lengthy discovery."); Spielman v. Merrill Lynch, Pierce, Fenner & Smith, Inc., 332 F.3d 116, 122–23 (2d Cir. 2003) ("The PSLRA afforded district courts the opportunity in the early stages of litigation to make an initial assessment of the legal sufficiency of any claims before defendants were forced to incur considerable legal fees or, worse, settle claims regardless of their merit in order to avoid the risk of expensive, protracted securities litigation."); Lander v. Hartford Life & Annuity Ins. Co., 251 F.3d 101, 107 (2d Cir. 2001) (citations omitted) ("Because of the expense of defending such suits, issuers were often forced to settle, regardless of the merits of the action. PSLRA addressed these concerns by instituting . . . a mandatory stay of discovery so that district courts could first determine the legal sufficiency of the claims in all securities class actions.").

[117] *See* Frank H. Easterbrook, *Discovery as Abuse*, 69 B.U. L. REV. 635, 636–37 (1989) (stating that discovery is "both a tool for uncovering facts . . . and a weapon capable of imposing large and unjustifiable costs on one's adversary").

> 2. Discovery should generally be obtained from the most convenient, least burdensome and least expensive sources.
>
> 3. Undue burden, expense, or delay resulting from a party's action or inaction should be weighed against that party.
>
> 4. Extrinsic information and sampling may assist in the analysis of whether requested discovery is sufficiently important to warrant the potential burden or expense of its production.
>
> 5. Nonmonetary factors should be considered when evaluating the burdens and benefits of discovery.
>
> 6. Technologies to reduce cost and burden should be considered in the proportionality analysis.[118]

The principles attempt to establish a balanced approach to proportionality that is not only fair to parties responding to requests for production of ESI but also to those requesting discovery.[119] The Sedona Commentary recognizes the inherent conflict between these positions, and, for that reason, the Commentary advises courts to use caution in its application:

> We recognize that some parties may inappropriately raise proportionality arguments, either as a sword to increase the burden on the producing party or as a shield to avoid legitimate discovery obligations. Courts must be wary of such abuses. In any event, the burden or expense of discovery is simply one factor in a proportionality analysis and may not be dispositive or even determinative in specific cases.[120]

The third principle especially is designed to protect against the unfair use of the doctrine to prevent discovery of relevant evidence based on the producing party's own negligence. The fifth and sixth principles are also intended to ensure a balanced approach that is fair to requesting parties. The sixth principle, mandating consideration of technologies to reduce costs and burdens, underlies the marriage of predictive coding and proportionality proposed in this Article.

The doctrine of proportionality is based on the well-established cost-burden analysis embodied in Federal Rule of Civil Procedure 26(b)(2)(C)(iii).[121] Similar analysis is also contained in Rules 26(b)(2)(B)

---

[118] SEDONA, COMMENTARY ON PROPORTIONALITY (2013), *supra* note 56, at 2.

[119] *See* Ralph Losey, *Why a Receiving Party Would Want to Use Predictive Coding?*, E-DISCOVERY TEAM (Aug. 12, 2013, 3:25 PM), http://e-discoveryteam.com/2013/08/12/why-a-receiving-party-would-want-to-use-predictive-coding/; John Tredennick, *Does Technology-Assisted Review Help in Reviewing Productions?*, CATALYST (Aug. 28, 2013), http://www.catalystsecure.com/blog/2013/08/does-technology-assisted-review-help-in-reviewing-productions/.

[120] SEDONA, COMMENTARY ON PROPORTIONALITY (2013), *supra* note 56, at 6 (footnotes omitted).

[121] FED. R. CIV. P. 26(b)(2)(C) provides:

and 26(g)(1)(B)(iii).[122] Magistrate Judge John M. Facciolla, of the United States District Court for the District of Columbia, is one of the leading experts on the proportionality doctrine. He addressed the doctrine in a 2011 opinion:

> All discovery, even if otherwise permitted by the Federal Rules of Civil Procedure because it is likely to yield relevant evidence, is subject to the court's obligation to balance its utility against its cost. . . .
>
> . . . .
>
> Without any showing of the significance of the non-produced e-mails, let alone the likelihood of finding the "smoking gun," the [party's] demands [for additional custodians] cannot possibly be justified when one balances its cost against its utility.[123]

### *B. Flexible Application of Cost-Burden Analysis*

Magistrate Judge Nan R. Nolan, of the United States District Court for the Northern District of Illinois, another proponent of the proportionality doctrine, applied the principle in a 2010 opinion, *Tamburo v. Dworkin*, calling it a "Rule 26 proportionality test."[124] For guidance on application of the test, Judge Nolan relied on The Sedona Conference Commentary:

> "The 'metrics' set forth in Rule 26(b)(2)(C)(iii) provide courts significant flexibility and discretion to assess the circumstances of the case and limit discovery accordingly to ensure that the scope and

---

> On motion or on its own, the court must limit the frequency or extent of discovery otherwise allowed by these rules or by local rule if it determines that: (i) the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive; (ii) the party seeking discovery has had ample opportunity to obtain the information by discovery in the action; or (iii) the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues.

[122] *See id.* 26(b)(2)(B), 26(g)(1)(B)(iii).

[123] U.S. *ex rel.* McBride v. Halliburton Co., 272 F.R.D. 235, 240–41 (D.D.C. 2011); *see also* Jones v. Nat'l Council of Young Men's Christian Ass'ns of the U.S., No. 09 C 6437, 2011 WL 7568591, at *2 (N.D. Ill. Oct. 21, 2011) ("The Court finds that Plaintiffs' untargeted, all-encompassing request fails to focus on key individuals and the likelihood of receiving relevant information."); Garcia v. Tyson Foods, Inc., No. 2:06-cv-02198-JWL-DJW, 2010 WL 5392660, at *14 (D. Kan. Dec. 21, 2010) (Waxse, Mag. J.) ("Plaintiffs present no evidence that a search of e-mail repositories of the 11 employees at issue is likely to reveal any additional responsive e-mails. . . . Plaintiffs must present something more than mere speculation that responsive e-mails *might* exist in order for this Court to compel the searches and productions requested.").

[124] Tamburo v. Dworkin, No. 04 C 3317, 2010 WL 4867346, at *3 (N.D. Ill. Nov. 17, 2010) (Nolan, Mag. J.).

> duration of discovery is reasonably proportional to the value of the requested information, the needs of the case, and the parties' resources."[125]

Judge Nolan adopted a phased approach to discovery in *Tamburo* to implement proportionality:

> Accordingly, to ensure that discovery is proportional to the specific circumstances of this case, and to secure the just, speedy, and inexpensive determination of this action, the Court orders a phased discovery schedule. . . . During the initial phase, the parties shall serve only written discovery on the named parties. Nonparty discovery shall be postponed until phase two, after the parties have exhausted seeking the requested information from one another.[126]

Judge Nolan then moved on to specific orders in *Tamburo* regarding what the parties must do, demonstrating an understanding that proportionality should have a space (scope) dimension and a time dimension.[127] She required discovery to be implemented in phases, not all at once, and she also understood that proportionality must be supported by cooperation, even if the cooperation is forced by court order.[128] A shotgun wedding is better than none.[129]

---

[125] *Id.* (quoting Sedona, *Commentary on Proportionality* (2010), *supra* note 113, at 294).

[126] *Id.* Other authorities reach similar conclusions. *See* Carroll, *supra* note 113, at 460–61 (internal quotation marks omitted) ("The proportionality concept also guides the court to use common sense techniques for managing discovery, like phased discovery or sequenced discovery. . . . Properly used, the proportionality tools available under the Federal Rules of Civil Procedure can go a long way toward reaching the long sought-after goal of Rule 1: securing the just, speedy, and inexpensive determination of every action and proceeding."); Sedona, *Commentary on Proportionality* (2010), *supra* note 113, at 297 ("Under [certain] circumstances, the court, or the parties on their own initiative, may find it appropriate to conduct discovery in phases, starting with discovery of clearly relevant information located in the most accessible and least expensive sources. Phasing discovery in this manner may allow the parties to develop the facts of the case sufficiently to determine whether, at a later date, further potentially more burdensome and expensive discovery is necessary or warranted.").

[127] *Tamburo*, 2010 WL 4867346, at *3.

[128] *Id.*

[129] Judge Nolan used this language to make all of these points:

> Within the next two weeks, the parties shall conduct an in-person meet and confer to prepare a phased discovery schedule. The parties are expected to be familiar with the Case Management Procedures regarding discovery on the Court's website, the Seventh Circuit's Electronic Discovery Pilot Program's Principles Relating to the Discovery of Electronically Stored Information, and the Sedona Conference Cooperation Proclamation. The parties are ordered to actively engage in cooperative discussions to facilitate a logical discovery flow. For example, to the extent that the parties have not completed their initial disclosures pursuant to Rule 26(a), or if their initial disclosures require updating, the parties should focus their efforts on completing their Rule 26(a) requirement before proceeding to other discovery requests. Second, the parties

In my opinion, electronic discovery production should almost always be conducted in phases. This is in accord with Sedona's second principle of proportionality, that, in general, litigants should seek discovery from the most convenient, least burdensome, and least expensive sources. As The Sedona Conference Proportionality Commentary indicates, parties should always focus first on the low-hanging fruit. In other words, they should focus first on evidence that is likely to have the most probative value and that is the most easily accessible.[130] Additionally, in my experience, requesting parties are open to phased discovery proposals so long as they are not asked to waive their right to additional, future discovery. Further, in most cases, after documents produced in the first round have been studied, the parties realize that they already have all they need to try the case. If there are any subsequent requests, they are usually very focused and constrained.

Where large amounts of ESI are involved, electronic discovery should always be phased, iterative, and fractal, just like the predictive coding process itself. I have found that this approach is the most effective and most efficient way to create order from today's near infinite and chaotic stores of ESI. It is the way to constrain electronic discovery in a just and efficient manner.

Judge Nolan addressed the proportionality doctrine again in her final *Kleen Products LLC v. Packaging Corporation of America* opinion when she considered an allegedly burdensome interrogatory, a motion to compel, and a counter-motion for a protective order.[131] Her analysis again relied on The Sedona Conference Proportionality Commentary. In her order partially granting the plaintiff's motion to compel, Judge Nolan explained that the defendants failed to back up their allegations of undue burden with specific facts:

> While a discovery request can be denied if the "burden or expense of the proposed discovery outweighs its likely benefit," a party objecting to discovery must specifically demonstrate how the request is burdensome. This specific showing can include "an estimate of the

---

> should identify which claims are most likely to go forward and concentrate their discovery efforts in that direction before moving on to other claims. Third, the parties should prioritize their efforts on discovery that is less expensive and burdensome. Finally, nothing in this Order shall prejudice the parties from conducting all forms of discovery after the pending motion to dismiss has been ruled upon.

*Id.* (footnote omitted) (citations omitted).

[130] SEDONA, COMMENTARY ON PROPORTIONALITY (2013), *supra* note 56, at 8–9 (describing the appropriateness in some cases of conducting discovery in phases, starting with the most obvious information located in the easiest-to-reach places).

[131] Kleen Prods. LLC v. Packaging Corp. of Am., No. 10 C 05711, 2012 WL 4498465, at *1, *7, *9–10 (N.D. Ill. Sept. 28, 2012) (Nolan, Mag. J.).

> number of documents that it would be required to provide . . . , the number of hours of work by lawyers and paralegals required, [or] the expense." Here, [the defendants'] conclusory statements do not provide evidence in support of their burdensome arguments.[132]

Judge Nolan makes an important point that proportionality is a mixed question of law and fact and often raises evidentiary issues concerning burden and benefit.

### *C. Importance of Early Assertion of Proportionality*

Courts have shown a preference for enforcing proportionality protection for parties responding to burdensome discovery when they raise the doctrine as early as possible. This is illustrated in three District Court cases that have recently considered the argument: *I-Med Pharma Inc. v. Biomatrix, Inc.* in New Jersey, *United States ex rel McBride v. Halliburton Company* in the District of Columbia, and *DCG Systems, Inc. v. Checkpoint Technologies, LLC* in California.[133]

#### 1. Very Late Assertion

When responding to discovery, the plaintiffs in *I-Med Pharma* waited far too long to raise the argument of the proportionality doctrine. They waited to raise the argument until the day before a stipulated and court-ordered deadline for production.[134] The underlying dispute concerned breach of contract, and the keyword list dreamed up by defense counsel, who apparently engaged in a rousing game of "*Go Fish*,"[135] included such zingers as the following (including variants of some of these terms): *contract, loss*, *profit*, *credit*, *refund*, *revenue*, *CL*, *HS*, *return*, *claim*, *FDA*, *HA*.[136] I could go on, but you get the picture. The opinion does not explain why plaintiff's counsel agreed to review and produce all non-privileged files that matched this ridiculously long list of keywords from opposing counsel.[137]

In *I-Med Pharma*, the attorneys not only used go fish keyword search, but also agreed to hire an expert to run the search and placed no

---

[132] *Id.* at *15, *17 (first two alterations in original) (citations omitted).

[133] I-Med Pharma Inc. v. Biomatrix, Inc., No. 03-cv-3677 (DRD), 2011 WL 6140658 (D.N.J. Dec. 9, 2011); DCG Sys., Inc. v. Checkpoint Techs., LLC, No. C-11-03792 PSG, 2011 WL 5244356 (N.D. Cal. Nov. 2, 2011).; U.S. *ex rel.* McBride v. Halliburton Co., 272 F.R.D. 235 (D.D.C. 2011).

[134] *See* Letter Motion Requesting Modification of Jan. 14, 2011 Discovery Order, *I-Med Pharma, Inc.*, 2011 WL 6140658, ECF No. 219 [hereinafter Letter Motion].

[135] *See* RALPH C. LOSEY, ADVENTURES IN ELECTRONIC DISCOVERY 204–06 (2011).

[136] So Ordered Stipulation at 5–6, *I-Med Pharma Inc.*, 2011 WL 6140658, ECF No. 182.

[137] *See I-Med Pharma Inc.*, 2011 WL 6140658.

limits on target custodians.[138] It was a search of the plaintiff's entire corporate computer system.[139] This is highly unusual. Not only that, the search was not restricted to any specific time periods; moreover, to make matters worse, they not only agreed to search the active files with word matches but also to search the *slack space* too.[140] *Slack space* is the so-called unallocated space files recovered by a forensic exam of plaintiff's computer system.[141] No wonder the wise judge presented with this conundrum, Senior United States District Court Judge Dickinson R. Debevoise, began his opinion with these words: "This case highlights the dangers of carelessness and inattention in e-discovery."[142]

Plaintiff's counsel finally woke up and discovered proportionality after the forensic expert searched the unallocated space of the client's computer system and found 64,382,929 hits covering the estimated equivalent of 95 million pages of documents![143] Given the complete failure to limit the search, this result, in Judge Debevoise's own words, "should come as no surprise."[144]

Since plaintiff's counsel by now probably had a pretty good idea of what a privilege review of another 95 million pages of mostly gibberish from slack space might cost, and since at this point the client probably did not want to pay for more, plaintiff's counsel finally said no. He asked defense counsel for a break on the prior agreement,[145] but defense counsel, perhaps sensing complete case victory, refused to modify the prior stipulation.[146] Then plaintiff requested relief from the prior

---

[138] *Id.* at *2.

[139] So Ordered Stipulation, *supra* note 136, at 1–2; *see also I-Med Pharma Inc.*, 2011 WL 6140658, at *2.

[140] Discovery Order at ¶ 5*, I-Med Pharma Inc.,* 2011 WL 6140658, ECF No. 211; *see also I-Med Pharma Inc.*, 2011 WL 6140658, at *2.

[141] *I-Med Pharma Inc.*, 2011 WL 6140658 at *5. "'Slack space' is the unused space at the logical end of an active file's data and the physical end of the cluster or clusters that are assigned to an active file." United States v. Triumph Capital Grp., Inc., 211 F.R.D. 31, 46 n.7 (D. Conn. 2002). Deleted data can be retrieved from slack space, but retrieval requires forensic tools. *Id.*

[142] *I-Med Pharma Inc.*, 2011 WL 6140658, at *1.

[143] *Id.* at *2; Brief in Support of Defendants' Appeal of Magistrate Judge Shipp's Discovery Order Dated September 9, 2011, at 7, *I-Med Pharma Inc.*, 2011 WL 6140658, ECF No. 240. The opinion does not say how many pages of documents with hits were found in the allocated spaces of the system, but it was probably millions more. Indeed, the opinion does not suggest that I-Med Pharma Inc., opposed the privilege review and production of these documents. In all likelihood they paid millions in vendor costs and attorney fees to comply with this portion of the stipulation.

[144] *I-Med Pharma Inc.*, 2011 WL 6140658, at *2.

[145] Letter Motion, *supra* note 134, at 1.

[146] *See I-Med Pharma Inc.*, 2011 WL 6140658, at *2; Letter Motion, *supra* note 134, at 6.

stipulation on discovery that had, as a matter of course, been converted to an order.[147] Plaintiff raised the doctrine of proportionality and suggested that the costs and burdens to review 64,382,929 hits from slack space would exceed any possible benefit from that exercise.[148] The magistrate assigned to hear the dispute, Judge Michael A. Shipp, agreed,[149] and the defendant, having little to lose (except, perhaps, credibility), appealed the decision to Judge Debevoise.[150]

Judge Debevoise, of course, affirmed his magistrate.[151] Judge Debevoise, a master of understatement, notes: "A privilege review of 65 million documents is no small undertaking. Even if junior attorneys are engaged, heavily discounted rates are negotiated, and all parties work diligently and efficiently, even a cursory review of that many documents will consume large amounts of attorney time and cost millions of dollars."[152]

Judge Debevoise granted a hearing on plaintiff's appeal. At the hearing, defense counsel argued that plaintiff's obligation to review 95 million pages need not really be that burdensome.[153] Judge Debevoise responded by asking defense counsel how *they* would do a privilege review of that many documents. Defendants' counsel said they would simply run a search for the word "privilege" and only review the documents with that word.[154] As Judge Debevoise observed, "In spite of the answer given, it is difficult to believe that lawyers from [their firm] regularly disclose large quantities of information from their client's files without examining it."[155]

Judge Debevoise, affirming the magistrate judge, let plaintiff's counsel off the hook and relieved them of elements of their prior e-discovery agreement.[156] But he had some choice words for them too, which provide good advice for all on a better way to do keyword search, going far beyond the simple guessing game the attorneys in this case had apparently been playing:

> While the precise number of hits produced was not known in advance and Plaintiff argues that it could not have predicted the volume of material that the search would uncover, it should have exercised more

---

147 Letter Motion, *supra* note 134, at 1.

148 *See I-Med Pharma Inc.*, 2011 WL 6140658, at *2.

149 *Id.*

150 *Id.* at *1–3.

151 *Id.* at *1, 6.

152 *Id.* at *5.

153 *Id.*

154 *Id.* at *5 & n.6.

155 *Id.* at *5 n.6.

156 *Id.* at *6.

> diligence before stipulating to such broad search terms, particularly given the scope of the search. In evaluating whether a set of search terms are [sic] reasonable, a party should consider a variety of factors, including: (1) the scope of documents searched and whether the search is restricted to specific computers, file systems, or document custodians; (2) any date restrictions imposed on the search; (3) whether the search terms contain proper names, uncommon abbreviations, or other terms unlikely to occur in irrelevant documents; (4) whether operators such as "and", "not", or "near" are used to restrict the universe of possible results; (5) whether the number of results obtained could be practically reviewed given the economics of the case and the amount of money at issue.
>
> . . . While Plaintiff should have known better than to agree to the search terms used here, the interests of justice and basic fairness are little served by forcing Plaintiff to undertake an enormously expensive privilege review of material that is unlikely to contain non-duplicative evidence.[157]

*I-Med Pharma* is a helpful case, not only for proportionality, but also for search. It is very telling that even though the case embodies the doctrine of proportionality, the keyword of "proportionality" itself is never used—even Rule 26(b)(2)(C) is never referred to. The opinion's omission of such key words demonstrates once again the limits of a keyword search.

### 2. Late Assertion

In *U.S. ex rel. McBride*, the party's timing in asserting the proportionality protection doctrine, though slightly better than in *I-Med Pharma*, was still late. Although not raised until after discovery had closed, at least protection was sought before stipulation to an order.[158] Fortunately for the responding party, the judge, United States Magistrate Judge John M. Facciola, is a strong advocate of proportionality. Although Judge Facciola is an expert and strong proponent of proportionality, my keyword search of the opinion shows that he too never once used the word in this opinion.[159] He cites Rule 26(b)(2)(C) several times,[160] but never says proportionality, showing once again the limits of keyword search.

The proportionality argument was raised in this case by the defendant in opposition to the plaintiff's motion to compel production.[161]

---

[157] *Id.* at *5–6.

[158] U.S. *ex rel.* McBride v. Halliburton Co., 272 F.R.D. 235, 236, 238 (D.D.C. 2011).

[159] *See U.S. ex rel. McBride*, 272 F.R.D. 235; *see also id.* at 240 (discussing the obligation to balance utility against cost without using the term "proportionality").

[160] *Id.* at 240–42.

[161] *Id.* at 240.

Defendant Halliburton had already reviewed and produced relevant emails of 230 custodians.[162] Discovery had closed, but the plaintiff wanted Halliburton to search and produce still more email from an additional 35 custodians.[163] These additional custodians were now targeted by the plaintiff, McBride, because they were carbon copied on emails transmitting relevant documents that were already produced as part of the 230.[164] No other reason was provided. Defendant opposed plaintiff's motion with a lengthy and detailed affidavit showing why this supplemental request would be burdensome.[165] Judge Facciola noted the affidavit, and then relied on Rule 26(b)(2)(c) to deny plaintiff's motion to compel:

> While the present record does not permit a precise conclusion, I can presume, given the numbers of hours for which the defendants billed and the period of time at issue, that the amount in controversy is great and that the defendants' resources are greater than the relator's. Claims of fraud in providing services to military personnel raise important, vital issues of governmental supervision and public trust. Thus, these factors might weigh in favor of the discovery sought.
>
> On the other hand, the defendants protest, and relator does not deny, that they have already spent a king's ransom on discovery in this case—$650,000—without the addition of attorneys' fees. They have produced more than two million paper documents, thousands of spreadsheets, and over a half a million e-mails.
>
> Given the discovery that relator has had, what defendants have already spent, and the detailed showing made of how much more time and money will likely have to [be] spent to search an additional thirty-five custodians, surely relator has to make a showing that the e-mails not produced are crucial to her proof. She has not made such a showing, and they are not. . . .
>
> In this context, it is telling that relator does not show from the e-mails she has received that there is good reason to believe that the ones she claims are missing are highly probative of some fact. Indeed, there is no showing whatsoever from what has been produced that those e-mails not produced will make the existence of some crucial fact more likely than not. It is, after all, unlikely that a transmitting e-mail will do any more than transmit attached information and, by copy, alert others of that transmittal.
>
> Without any showing of the significance of the non-produced e-mails, let alone the likelihood of finding the "smoking gun," the search

---

[162] *Id.* at 239–40.

[163] *Id.* at 240.

[164] *Id.*

[165] *Id.*; *see also* Defendants' Opposition to Relator's Motion to Compel Production of Documents, U.S. *ex rel.* McBride v. Halliburton Co., 272 F.R.D. 235 (D.D.C. 2011) (No. 1:05-cv-00828-FJS-JMF), ECF No. 108.

> relator demands cannot possibly be justified when one balances its cost against its utility. The motion will be denied.[166]

Therefore, although the timing of the argument was imperfect, proportionality prevailed against an unjustifiable search.

### 3. Timely Assertion

In *DCG Systems* the timing was right. The issue of proportionality was raised at the Rule 26(f) conference and 16(b) hearing as part of discovery plan discussions.[167] That is what the rules intend. Proportionality protection requires prompt, diligent action, as seen in this case.

*DCG Systems* was a patent case between two companies with competing patent rights.[168] The defendant wanted to have a *Model Order Limiting E-Discovery in Patent Cases* ("Model Order") entered in the case and thereby limit the initial scope of both sides' e-discovery.[169] The Model Order is not mandatory in patent cases but may be adopted upon court order or the parties' stipulation. The Model Order limits initial e-discovery to email from five custodians and five keywords per custodian.[170] This represents the Patent Bar's first attempt at a procedure to implement proportionality in e-discovery. The parties may jointly agree to modify these limits or request court modification for good cause, but even if they do not agree, or there is no order permitting more email discovery, a requesting party may obtain more discovery *if they pay for it*.[171]

Here the plaintiff objected to the defendant's request to have the Model Order entered in this case, and so they brought the issue to the judge at the Rule 16(b) hearing.[172] United States Magistrate Judge Paul S. Grewal agreed with the defendant and adopted the Model Order, reasoning:

> Critically, the email production requests must focus on particular issues for which that type of discovery is warranted. The requesting party must further limit each request to a total of five search terms and the responsive documents must come from only a defined set of

---

166 *United States ex rel. McBride*, 272 F.R.D. at 241 (citations omitted).

167 DCG Sys., Inc. v. Checkpoint Techs., LLC, No. C-11-03792, 2011 WL 5244356, at *1 (N.D. Cal. Nov. 2, 2011).

168 *Id.* at *1–2.

169 *Id.* at *1 (citing Advisory Council for the U.S. Court of Appeals for the Fed. Circuit, An E-Discovery Model Order ¶¶ 10–11 (last visited Oct. 29, 2013) [hereinafter E-Discovery Model Order], *available at* http://www.cafc.uscourts.gov/images/stories/announcements/Ediscovery_Model_Order.pdf).

170 E-Discovery Model Order, *supra* note 169.

171 *Id.* ¶ 11.

172 *DCG Sys., Inc.*, 2011 WL 5244356, at *1.

> five custodians. These restrictions are designed to address the imbalance of benefit and burden resulting from email production in most cases. As Chief Judge Rader noted in his recent address in Texas on the "The State of Patent Litigation" in which he unveiled the Model Order, "[g]enerally, the production burden of expansive e-requests outweighs their benefits. I saw one analysis that concluded that .0074% of the documents produced actually made their way onto the trial exhibit list—less than one document in ten thousand. And for all the thousands of appeals I've evaluated, email appears more rarely as relevant evidence."[173]

Judge Grewal concluded with a cautionary note, however, and left the door open for the plaintiff to return seeking more discovery.[174]

### *D. Proportionality Requires Justice, as Well as Speed and Efficiency: Criticisms of* DCG Systems *and the Patent Bar Model Order*

*DCG Systems* is, by far, the best of the three cases here examined for applying proportionality, but it is still far from perfect. It embraces proportionality, and will no doubt save the parties money in e-discovery, but at what cost? Litigation is about finding justice. If you lose that, you lose everything.

Rule 1 of the Federal Rules of Civil Procedure requires that litigation be "speedy" and "inexpensive."[175] Limiting discovery to five keywords and five custodians will get you that. But Rule 1 also requires litigation to be "just."[176] That is, after all, the whole point of litigation. In America, like most of the civilized world, we do not just go through the motions of legal process in a fast and cursory manner. Court systems are not just an empty charade. The heart of law as we know it is due process. We decide cases on the merits, on the facts, on the evidence, not just on the whim of judges or juries. That is what justice means to us. For those reasons, we should all be concerned about placing on e-discovery arbitrary limits designed to save money, and speed things along, if the tradeoff is justice.

Judge Grewal, who decided *DCG Systems*, shares these concerns.[177] So too does the Patent Bar who adopted this Model Order, and Chief

---

[173] *Id.* (alteration in original).

[174] *Id.* at *2 ("Perhaps the restrictions of the Model Order will prove undue. In that case, the court is more than willing to entertain a request to modify the limits. But only through experimentation of at least the modest sort urged by the Chief Judge will courts and parties come to better understand what steps might be taken to address what has to date been a largely unchecked problem.").

[175] FED. R. CIV. P. 1.

[176] *Id.*

[177] *See DCG Sys., Inc.*, 2011 WL 5244356, at *2.

Judge Randall Rader who promotes it.[178] They are trying hard to find a proportional balance between benefit and burden, to know when *enough is enough* in the search for evidence. They do not want too much, like some unscrupulous attorneys for whom e-discovery is little more than a legal tool in a game to extort settlement. They also do not want too little, like some equally unscrupulous attorneys who play hide-the-ball. Good attorneys are like *Goldilocks*; they are looking for the *just-right* amount of e-discovery. They are looking for proportionality.

The patent judges show this concern in the pains they take to say that the five-and-five rule is just a "starting point."[179] They make clear that more e-discovery outside of these limits may be appropriate, and that parties can always move the court for additional discovery. For instance, in *DCG Systems*, Judge Grewal acknowledged that the restrictions of the Model Order might prove too onerous and said he was "more than willing to entertain a request to modify the limits."[180] The Model Order shows the same concern that justice not be sacrificed at the altar of efficiency.[181]

The intent to preserve justice apparent in *DCG Systems* and the Model Order is, however, frustrated by the order's reliance on go-fish-type keyword search.[182] It is not so much the arbitrary limit to five keywords that is troubling, nor the initial limit to five custodians, which is fine. What is troubling about the Model Order to search experts is the reliance on keyword search alone, and *blind-pick* keyword search at that,[183] which should bother anyone who has read the scientific studies.[184] The Model Order is promoting the worst kind of search: the blind-keyword-guessing kind. That is probably an inadvertent error. The lawyers and judges behind the Model Order were apparently not aware

---

[178] *See* Advisory Council for the U.S. Court of Appeals for the Fed. Circuit, *Introduction* to An E-Discovery Model Order 3–4 (last visited Oct. 29, 2013), *available at* http://www.cafc.uscourts.gov/images/stories/announcements/Ediscovery_Model_Order.pdf.

[179] *Id.* at 2–3.

[180] *DCG Sys., Inc.*, 2011 WL 5244356, at *2.

[181] E-Discovery Model Order, *supra* note 169, ¶ 10 ("The Court shall consider contested requests for up to five additional custodians per producing party, upon showing a distinct need based on the size, complexity, and issues of this specific case.").

[182] *See* LOSEY, ADVENTURES IN ELECTRONIC DISCOVERY, *supra* note 135, at 204–06; E-Discovery Model Order, *supra* note 169, ¶ 11 (requiring the use of a limited number of search terms for email production requests without requiring the requesting party to reveal what they are looking for).

[183] *See* Daniel B. Garrie & Yoav M. Griver, *Unchaining E-Discovery in the Patent Courts*, 8 WASH. J.L. TECH. & ARTS 487 (2013) (discussing the debate over the effectiveness of keyword searching); *see also* LOSEY, ADVENTURES IN ELECTRONIC DISCOVERY, *supra* note 135, at 204–06.

[184] *See* sources cited *supra* notes 39–40.

of the limits of blind-guessing-based keywords. When they do become aware, I assume they will consider appropriate revisions to the Model Order, including revisions to the use of keywords to include metrics and information sharing[185] and provisions pertaining to the use of predictive coding. Accordingly, they should consider using language similar to that found in the Commentary to newly enacted Rule 502 of the Federal Rules of Evidence.[186]

The Model Order of the Patent Bar is a good start, but it needs revision so that keyword searches can be more effective, and the use of predictive coding can be encouraged. As is shown in the concluding section, the relevance-ranking features of predictive coding make it easier to adapt to the proportionality doctrine than keyword searches that have no such ability.[187] For that reason, when predictive coding is used, it is much easier to attain efficient cost savings without omission of key relevant documents. Thus, a fair balance can be reached between the seemingly contradictory dictates of Rule 1 to be efficient and inexpensive, but also to be just.

### *E. The Growing Influence of the Proportionality Doctrine*

Even without buttressing the proportionality doctrine with a marriage to predictive coding technology as here proposed, the doctrine is growing in popularity.[188] In addition to the legal opinions already discussed, I have identified 16 other district court opinions which are of

---

[185] The Model Order should be reformed to require that basic metrics be shared on proposed keywords. It should require enough disclosure so that the keyword picks are not blind. A requesting party should be permitted some keyword testing before five terms are settled upon.

[186] *See* FED. R. EVID. 502 advisory committee's note to subdivision (b) ("Depending on the circumstances, a party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken 'reasonable steps' to prevent inadvertent disclosure.").

[187] *See infra* Part IV.

[188] *See generally* Philip J. Favro & the Hon. Derek P. Pullan, *New Utah Rule 26: A Blueprint for Proportionality Under the Federal Rules of Civil Procedure*, 2012 MICH. ST. L. REV. 933 (2012) (proposing amendments to Utah's rules that emphasize proportionality and citing Ralph Losey for his discussion of proportionality in the *DCG Systems* case); Theodore C. Hirt, *The Quest for "Proportionality" in Electronic Discovery—Moving from Theory to Reality in Civil Litigation*, 5 FED. CRTS. L. REV. 171 (2011) (discussing the development of, the challenges in, and the recent support for the application of proportionality principles); Brian C. Vick & Neil C. Magnuson, *The Promise of a Cooperative and Proportional Discovery Process in North Carolina: House Bill 380 and the New State Electronic Discovery Rules*, 34 CAMPBELL L. REV. 233 (2012) (examining the cooperation and proportionality principles included in North Carolina's new e-discovery rules).

some importance to the growth of the doctrine.[189] Most were written since the Sedona Proportionality paper was first published in 2010.

---

[189] Although not a complete list, the 16 cases are helpful to the proportionality doctrine. *See* Apple Inc. v. Samsung Elecs. Co., No. 12-CV-0630-LHK(PSG), 2013 WL 4426512, at *3 (N.D. Cal. Aug. 14, 2013) (footnote omitted) ("[T]he court is required to limit discovery if 'the burden or expense of the proposed discovery outweighs its likely benefit.' This is the essence of proportionality—an all-to-often ignored discovery principle."); Tucker v. Am. Int'l Grp., Inc., 281 F.R.D. 85, 90, 95, 98 (D. Conn. 2012) ("[Plaintiff requested] essentially carte blanche access to rummage through Marsh's electronically stored information, purportedly in the hope that the needle she is looking for lurks somewhere in that haystack. . . . [T]he burdens of plaintiff's proposed inspection upon Marsh outweigh the benefits plaintiff might obtain were she to obtain the emails through a Datatrack inspection. Plaintiff seeks to search, *inter alia*, the mirror images of eighty-three laptops—in effect, to dredge an ocean of Marsh's electronically stored information and records in an effort to capture a few elusive, perhaps non-existent, fish. . . . Courts are obliged to recognize that non-parties should be protected with respect to significant expense and burden of compelled inspections under Fed. R. Civ. P. 45(c)(2)(B)(ii). . . . Moreover, courts have focused on the importance of the Rule 26(b)(2)(C) proportionality limit to implement fair and efficient operation of discovery. . . . Balancing the prospective burden to Marsh against the likely benefit to plaintiff from the proposed inspection, the Court concludes that the circumstances do not warrant compelling Marsh to endure inspection of its computer records by Datatrack."); Madere v. Compass Bank, No. A-10-CV-812 LY, 2011 WL 5155643, at *2 (W.D. Tex. Oct. 28, 2011) ("As the cost to restore Compass Bank's backup tapes 'outweighs its likely benefit,' especially in light of the amount in controversy, the Court DENIES Madere's request for production."); Gen. Steel Domestic Sales, LLC v. Chumley, No. 10-cv-01398-PAB-KLM, 2011 WL 2415715, at *2–3 (D. Colo. June 15, 2011) (rejecting defendant's request for the production of every recorded sales call on plaintiff's database for around a two-year period because it would take four years to listen to the calls to identify potentially responsive information); Thermal Design, Inc. v. Guardian Bldg. Prods., Inc., No. 08-C-828, 2011 WL 1527025, at *1 (E.D. Wis. Apr. 20, 2011) (refusing to approve plaintiff's electronic fishing expedition simply because the defendant had the financial resources to pay for the searches, and holding the financial resources of the defendant are not tantamount to good cause under Federal Rule of Civil Procedure 26(b)(2)(C)); Wood v. Capital One Servs., LLC, No. 5:09-CV-1445 (NPM/DEP), 2011 WL 2154279, at *9 (N.D.N.Y. Apr. 15, 2011) (holding the "rule of proportionality" dictated that the plaintiff's motion be denied "without prejudice to his right to renew the motion to compel in the event he is willing to underwrite the expense associated with any such search"); Call of the Wild Movie, LLC v. Does 1–1,062, 770 F. Supp. 2d 332, 352, 354 (D.D.C. 2011) (granting a motion to compel because the request was narrow and the ESI requested was important, compared with an insufficient showing of undue burden); Hock Foods, Inc. v. William Blair & Co., No. 09-2588-KHV, 2011 WL 884446, at *9 (D. Kan. Mar. 11, 2011) (denying in part a motion to compel in light of estimated costs between $1.2 and $3.6 million dollars to search 12,000 gigabytes of data in order to answer an overbroad interrogatory); Diesel Mach., Inc. v. Manitowoc Crane, Inc., No. CIV 09-4087-RAL, 2011 WL 677458, at *3 (D.S.D. Feb. 16, 2011) (denying a motion to compel the production of documents in native format because no explanation was provided on why information contained in native format was necessary to the facts of the case when those same documents had already been produced as PDFs); Daugherty v. Murphy, No. 1:06-cv-0878-SEB-DML, 2010 WL 4877720, at *7–8 (S.D. Ind. Nov. 23, 2010) (holding that the cost and burden of the additional production outweighed the benefit, and the defendant's sworn testimony on burden and cost was credible); Willnerd v. Sybase, Inc., No. 1:09-cv-500-BLW,

## IV. HOW PREDICTIVE CODING SUPPORTS PROPORTIONALITY

In most lawsuits, the focus of proportionality efforts is on document review. That is appropriate because document review typically constitutes between 60% and 80% of the total e-discovery expense.[190] The use of the latest AI-based review technologies can significantly reduce these costs as shown,[191] and for this reason alone predictive coding is the best tool we have for proportionality. But there is more to it than that. What makes this a marriage truly made in heaven is the document-ranking capabilities of predictive coding. This allows parties to limit the documents considered for final production to those that the computer determines have the highest probative value. This key ranking feature of AI-enhanced document review allows the producing party to provide the requesting party with the *most bang for the buck*. This not only saves the producing party money, and thus keeps its costs proportional, but it saves time and expenses for the requesting party. It makes the production much more precise, and thus faster and easier to review. It avoids what can be a costly exercise to a requesting party to wade

---

2010 WL 4736295, at *3 (S.D. Idaho Nov. 16, 2010) ("[A] search of the employees' e-mails would amount to the proverbial fishing expedition—an exploration of a sea of information with scarcely more than a hope that it will yield evidence to support a plausible claim of defamation. In employing the proportionality standard of Rule 26(b)(2)(C), as suggested by Willnerd, the Court balances Willnerd's interest in the documents requested, against the not-inconsequential burden of searching for and producing documents."); Moody v. Turner Corp., No. 1:07-cv-692 (S.D. Ohio filed Sept. 21, 2010) ("[T]he mere availability of such vast amounts of electronic information can lead to a situation of the ESI-discovery-tail wagging the poor old merits-of-the-dispute dog."); Bassi & Bellotti S.p.A. v. Transcon. Granite, Inc., No. 08-cv-1309-DKC, 2010 WL 3522437, at *3 (D. Md. Sept. 8, 2010) ("[T]he Federal Rules do impose an obligation upon courts to limit the frequency or extent of discovery sought in certain circumstances, such as when the discovery requested is unreasonably duplicative or cumulative, or the burden or expense of the proposed discovery outweighs the likely benefit, considering the needs of the case, the importance of the issues at stake in the action, and the importance of the discovery in resolving those issues."); Rimkus Consulting Grp. v. Cammarata, 688 F. Supp. 2d 598, 613 (S.D. Tex. 2010) (requiring the parties engage in *reasonable* efforts, and stating that what is reasonable "depends on whether what was done—or not done—was *proportional* to that case"); Rodríguez-Torres v. Gov't Dev. Bank of P.R., 265 F.R.D. 40, 44 (D.P.R. 2010) ("[T]he Court determines that the ESI requested is not reasonably accessible because of the undue burden and cost. The Court finds that $35,000 is too high of a cost for the production of the requested ESI in this type of action. Moreover, the Court is very concerned over the increase in costs that will result from the privilege and confidentiality review that Defendant GDB will have to undertake on what could turn out to be hundreds or thousands of documents."); Dilley v. Metro. Life Ins. Co., 256 F.R.D. 643, 644 (N.D. Cal. 2009) (citation omitted) ("The court must limit discovery if it determines that 'the burden or expense of the proposed discovery outweighs its likely benefit,' considering certain factors including 'the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues.'").

190 *See* RAND REPORT, *supra* note 43, at 41 (finding a 73% average).

191 *See supra* notes 43, 52, 63 and accompanying text.

through a document dump,[192] a production that contains a high number of irrelevant or marginally relevant documents. Most importantly, it gives the requesting party what it really wants—the documents that are the most important to the case.

### *A. Two Stages of Document Review Using Predictive Coding*

To understand the full value of document ranking, it is necessary to understand how document review using predictive coding is now typically conducted in two stages. The first stage is identification of the likely responsive or relevant documents, which is also known as *first-pass review*. The second is study of the selected likely relevant documents to verify relevancy and determine which relevant documents must nevertheless be withheld, logged, redacted, and/or labeled to protect a client's confidential information. The second stage can also include tagging specific issues unrelated to confidentiality concerns.

There is no need for the second-pass review of any documents determined to be irrelevant, since such documents will not be produced, and thus, there is no need to implement such protections. This second-pass final review is an enormous problem in litigation for a variety of reasons, not just cost, especially as it concerns attorney-client privileges.[193] Therefore, the ability to limit the number of documents passed through to second review is critical to effectuating both cost and risk efficiencies.

---

[192] See Branhaven, LCC v. Beeftek, Inc., 288 F.R.D. 386, 392–93 (D. Md. 2013), where the plaintiff's document dump in response to a request for production led to the imposition of sanctions upon both plaintiff and its attorneys. As the court explained:

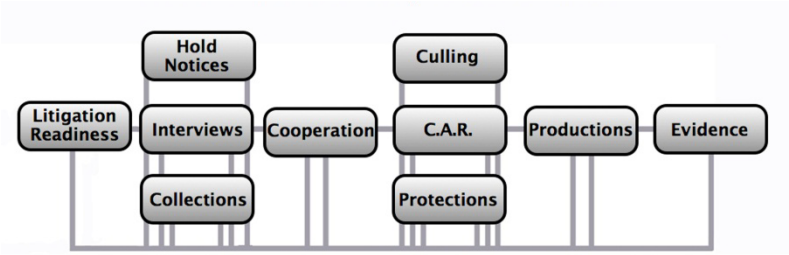
> As plaintiff's counsel has an affirmative duty to assure that their client responds completely and promptly to discovery requests [sic]. Their inaction seriously frustrated the defense of this case. The record here demonstrates a casualness at best and a recklessness at worst in plaintiff's counsel's treatment of their discovery duties. I agree with defense counsel that the attorneys abdicated their responsibilities while representing that they had not. If all counsel operated at this level of disinterest as to discovery obligations, chaos would ensue and the orderliness of the discovery process among counsel in federal courts, which is exquisitely dependent on honorable attorney self-regulation, would be lost.

*Id. See also* Ralph Losey, *The Increasing Importance of Rule 26(g) to Control e-Discovery Abuses*, E-DISCOVERY TEAM (Feb. 24, 2013, 6:10 PM), http://e-discoveryteam.com/2013/02/24/the-increasing-importance-of-rule-26g-to-control-e-discovery-abuses/.

[193] *See* Anonymous, *An Open Letter to the Judiciary—Can We Talk? (Part One)*, E-DISCOVERY TEAM (Sept. 11, 2011, 10:09 PM), http://e-discoveryteam.com/2011/09/11/an-open-letter-to-the-judiciary-%E2%80%93-can-we-talk-part-one/; Anonymous, *An Open Letter to the Judiciary—Can We Talk? (Part Two)*, E-DISCOVERY TEAM, (Sept. 18, 2011, 8:04 PM), http://e-discoveryteam.com/2011/09/18/an-open-letter-to-the-judiciary-%E2%80%93-can-we-talk-part-two/.

The second stage review is still computer-assisted, but it primarily involves human study of the documents identified in the first pass as likely relevant. The second stage does not consider the documents rejected in the first-pass review.[194] In the *Electronic Discovery Best Practices* ("EDBP") work flow diagram used to explain all of the legal work involved in e-discovery,[195] not just document review, this second stage of document review is called *Protections*. That is because its primary focus is on protection of attorney-client and work-product privileges. Protections is step number eight in the EDBP ten-step model shown below. The binary relevancy identification work, first-pass review, is step seven, labeled as C.A.R. for computer-assisted review, in the EDBP diagram. This step can be understood as containing the first five steps in the six-step model of predictive coding, or the first six steps in the eight-step model of predictive coding.

Diagram 3: Electronic Discovery Best Practices[196]




In the vast majority of cases, litigants do not dispense with final manual review of documents or rely *solely* on automated software in the Protections step.[197] The likelihood of error is simply still too high at this point in AI-enhanced software development for this to be an acceptable risk, at least in most cases for most clients. In some cases, for some clients, the risk of waiver may be acceptable. But in most cases the damage caused by disclosure of some privileged communications cannot be fully repaired by clawback agreements and orders, even when they

---

[194] The second-stage review is identified as the last step in the six-step model of predictive coding described above where it is labeled *Protection Reviews and Productions*. *Supra* Diagram 1. In the eight-step predictive coding model, it is step number seven, called *Proportional Final Review*. *Supra* Diagram 2.

[195] See ELECTRONIC DISCOVERY BEST PRACTICES, www.EDBP.com for a complete explanation of the Electronic Discovery Best Practices model and its ten steps.

[196] Created by e-Discovery Team, Copyright © Ralph Losey 2012.

[197] This is based on my informal polling and questions of leaders of e-discovery departments of many large law firms and corporate law departments and hundreds of participants in CLE events around the country.

are enforced.[198] That is the primary reason litigants are unwilling to rely on technology and clawback agreements alone.

The only exceptions routinely encountered at this point are non-litigation circumstances, such as internal investigations, or in some productions, such as various productions to the government or second reviews in merger approvals. The second review may also sometimes be waived in a litigation context when the client has little choice but to save expenses or where the client thinks the risk of disclosure is very low. This later scenario typically arises when the data under review is very unlikely to contain confidential information that the client cares about, such as old data of a company that it acquired or where the data has been previously viewed by the requesting party.

Since second-pass review is required in most cases to preserve client secrets and confidential data, the reduction of the number of documents subject to second review by elimination in the first-pass review as probable irrelevant has a direct impact on the cost of the project. This is where the ranking features of AI-based search come in. Only documents determined appropriate according to a ranking system are subject to the second review. All others are culled out of consideration for production.

The ranking cut-off point is within the reviewer's control. Like most things in the law, the appropriate number depends on the case. The most common threshold is a simple probable relevance point where only documents ranked as 50% or higher probable relevance are subject to second review. Most predictive coding software can easily determine and segregate these documents and channel them for second review.[199] The documents with 50% or lower ranking are automatically classified as irrelevant and not produced although they should be subject to some quality control verifications. Alternatively, a higher or lower probability level could be used as a threshold, such as 75% or higher probable relevance. Using this higher threshold would typically reduce the number of documents subject to second review.

The selection of a gatekeeper probability level is dependent on a number of factors, including quality controls, special sampling, and

---

[198] *See* Brookfield Asset Mgmt., Inc. v. AIG Fin. Prods. Corp., No. 09 Civ. 8285(PGG)(FM), 2013 WL 142503 (S.D.N.Y. Jan. 7, 2013). The privileged documents produced in *Brookfield* because of a vendor error were ordered returned to defendant, but plaintiff's counsel and plaintiff still were able to read the documents and become aware of the secrets. *Id. Neuralyzer* devices to erase human memory are only fictional; thus, the bell sounding client secrets cannot be un-rung.

[199] I am aware that some predictive coding software does not so categorize and rank the document collection. A percentage-probable-ranking feature is an essential feature to my proportional review methods here discussed. For that reason, I do not recommend those vendors, or their software, but instead encourage these companies to enhance their products to include this key feature.

many different project metrics. No particular probability percentage is appropriate for all cases. Instead, this is dependent on the data itself, the functioning of the machine learning software and ranking, the number of files in each ranking category, and, as will be shown next, on the overall proportionality analysis of the case.

### *B. Bottom-Line-Driven Proportional Review and Production*

The new method of review and production analysis that I developed over the past 7 years is called *Bottom-Line-Driven Proportional Review*. The bottom line in e-discovery production is what it costs. Despite what some lawyers and vendors may say, total cost is not an impossible question to answer. It takes an experienced lawyer's skill to answer, but, after a while, you can get quite good at such estimation. It is basically a matter of estimating attorney billable hours plus vendor costs. With practice, cost estimation can become a reliable art, a projection that you can count on for budgeting purposes, and, as we will see, for proportionality arguments. Furthermore, as better technological tools are developed in the future to assist in this process I expect cost estimation to become much more of a science than an art.

Total cost projections may never be exact, but the ranges can usually be predicted, subject, of course, to the target changing after the estimate is given. If the complaint is amended or different evidence becomes relevant, then, just like a construction project, a change order may be required for the new specifications.[200]

Price estimation is an obvious thing to do before you begin work on any big project, especially complex projects, such as building construction or large e-discovery document reviews. Estimating legal review costs is basically the same thing as construction estimating—projecting materials and labor costs. In construction you calculate prices per square foot. In e-discovery you estimate prices per file.

The new strategy and methodology is based on a bottom line approach where you estimate what review costs will be, make a proportionality analysis as to what should be spent, and then engage in defensible culling to bring the review costs within the proportional budget. The producing party determines the number of documents to be subjected to final review by calculating backwards from the bottom line of what they are willing, or required, to pay for the production.

The defensible culling aspect of the method has been significantly buttressed by predictive coding software, especially the new ranking abilities. As discussed, predictive coding software evaluates the strength

---

[200] I had two years of experience in the 1970s before law school as a construction estimator.

of relevance and irrelevance of every document in the data set analyzed. Before probability ranking, although parties always tried to cull out the least likely relevant documents when using Bottom-Line-Driven Proportional Review, there was considerable guesswork involved.

No legal review software existing before the new AI-enhanced predictive coding versions had any real document ranking abilities. Lawyers would instead rely on their own judgment and experience, coupled with sampling. There was too much room for human error. Only lawyers with extremely high skill and experience levels could cull accurately. There was too much art, and not enough science. But all of that changed with AI-enhanced document ranking. Now it is much easier to accurately focus your review on the documents most likely to have probative value to the case. With this new technology, we can, for the first time, confidently attain our proportional budget goals by culling out documents that are the *least likely* to be relevant.

1. Setting a Budget Proportional to the Case

The process begins by the producing party calculating the maximum amount of money appropriate to spend on ESI production. This is typically a range rather than one specific number. The budget range is usually tied to a number of different conditions and assumptions. This kind of budgetary analysis requires not only an understanding of the ESI production requests, but also a careful and realistic evaluation of the merits of the case. This is where the all important *proportionality* element comes in.

The amount selected for the budget should be proportional to the monies and issues in the case. As shown in the discussion on the proportionality doctrine, a producing party is not required to assume excessive, disproportional expenses, but it is required to pay for proportional discovery. The art is in knowing where to draw the line.

The budget becomes the bottom line that drives the review and keeps the costs proportional. The producing party seeks to keep the total costs within that budget. The budget should either be by agreement of the parties after some discussion at a Rule 26(f) conference, or at least without objection, and, failing that, by court order that in some way protects the producing party from excessive expense. If a party chooses not to disclose the restraints they have decided to utilize, they risk later second-guessing and an expensive do-over. The proportional approach is, as we have seen in the case law, necessarily based on a cooperative approach[201] and some disclosure.[202]

---

[201] *See supra* notes 58, 128–29 and accompanying text.

[202] *See supra* notes 90–94, 102–03 and accompanying text.

The failure of most practicing attorneys to estimate and project future costs and decide in advance to conduct the review so as to stay within budget, is one of the primary reasons that e-discovery costs today are so high. Once you spend the money, it is very hard to have additional costs shifted to the requesting party.[203] But if you raise objections and argue proportionality before the spend, then you will have a much better chance.[204]

Under the Bottom-Line-Driven Proportional approach, after analyzing the case merits and determining the maximum proportional expense, the responding party makes a good faith estimate of the likely maximum number of documents that can be reviewed within that budget. The document count represents the number of documents that you estimate can be reviewed for final decisions of relevance, confidentiality, privilege, and other issues and still remain within budget. The review costs you estimate must be based on best practices, which in all large review projects today means predictive coding, and the estimates must be accurate (i.e., no puffing or mere guesswork).

The producing party then uses predictive coding techniques and quality controls to find the documents most likely to be responsive within the number of documents the budget allows. Since predictive coding is based on document relevancy ranking, it is the perfect tool to facilitate bottom-line-driven review.

By using best methods with predictive coding search[205] and taking advantage of the relevancy ranking features, you can get the most bang for your buck, arriving at the core truth. That in turn helps persuade the requesting party or court to go along with your budgetary limits. That is the essential reason I consider predictive coding to be a great facilitator of the Bottom-Line-Driven Proportional Review method.

### 2. Small Case Example

A few examples may help clarify how this method works. Assume a case where you determine a proportional cost of production to be $50,000, and estimate, based on sampling and other hard facts, that it will cost you $1.25 per file for both the automated and manual review before production of the ESI at issue (steps seven and eight of the EDBP

---

[203] *See, e.g.*, *In re* Fannie Mae Sec. Litig., 552 F.3d 814, 822 (D.C. Cir. 2009) (discussing the lower court's denial of request for cost-shifting after expenses were incurred).

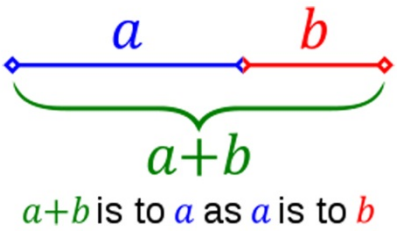
[204] Boeynaems v. LA Fitness Int'l, LLC, 285 F.R.D. 331, 340, 342–43 (E.D. Pa. 2012) (granting a defense motion to shift costs to plaintiff based on estimates that the requested search for emails would cost $219,000 and the member notes search would cost $360,000).

[205] This requires careful use of SMEs and a hybrid multimodal approach. *See supra* notes 60–67, 109–10, 112 and accompanying text.

flow chart).[206] Then you can review no more than 40,000 documents and stay within budget. It is that simple. No higher math is required.

The most difficult part is the legal analysis to determine a budget proportional to the real merits of the case. But that is nothing new. What is the *golden mean* in litigation expense? How to balance just, with speedy and inexpensive?[207] The essence of the ideal proportionality question has preoccupied lawyers for decades. Proportionality has also preoccupied scientists, mathematicians, and artists for centuries. Many mathematicians and artists claim to have found a mathematical and aesthetic answer that they call the *golden mean* or *golden ratio*, shown below.

Diagram 4: Golden Mean[208]




In law, this is the perennial *Goldilocks* question. How much is too much? Too little? Just right? How much is appropriate to spend to produce documents? The issue is old. I have personally been dealing with this problem since 1980. What is new is applying this legal analysis to a modern-day, high-volume-ESI search and review plan. Unfortunately, unlike art and math, there is no generally accepted golden ratio in the law, so it has to be recalculated and reargued for each case.[209]

---

[206] *Supra* Diagram 3.

[207] FED. R. CIV. P. 1.

[208] This graphic is open-source.

[209] Ralph Losey, *My Basic Plan for Document Reviews: The "Bottom Line Driven" Approach (2013 Second Updated Version)*, E-DISCOVERY TEAM (Oct. 1, 2013, 4:47 PM), http://e-discoveryteam.com/2013/10/01/my-basic-plan-for-document-reviews-the-bottom-line-driven-approach/ ("If the golden ratio [of art and science] were accepted in law as an ideal proportionality, the number [would be] 1.61803399, aka *Phi*. That would mean 38% is the perfect proportion. I have argued that when applied to litigation that means the total cost of litigation should never exceed 38% of the amount at issue. In turn, the total cost of discovery should not exceed 38% of the total litigation cost, and the cost of document production should not exceed 38% of the total costs of discovery (as opposed to our current 73% reality). (It's like Russian nesting dolls that get proportionally smaller.) Thus for a $1 million case you should not spend more than $54,872 for document productions (1,000,000 – 380,000 – 144,400 – 54,872)."). Perhaps someday a judge will agree and at least refer to the *golden mean* in math and nature as part of a proportionality analysis. In the

Estimation for bottom-line-driven review is essentially a method for marshaling evidence to support an undue burden argument under Rule 26(b)(2)(C). It is basically the same thing we have been doing to support motions for protective orders in the paper production world for over 60 years. The only difference is that now the facts are technological, the numbers and variety of documents are enormous, sometimes astronomical, and the methods of review, especially the preferred predictive coding methods, are complex and not yet standardized.

3. Estimate of Projected Costs

The calculation of projected cost per file to review can be quite complicated, and is frequently misunderstood or is not based on best practices. Still, in essence, this cost projection is also fairly simple. Parties project how long it will take to do the review and the total cost of the time. (The materials costs, i.e., software usage fees, may also have to be factored in.)

Thus, for example (and this is an over-simplification), assume again the review project of 40,000 documents. Note that it probably started as 100,000 or 200,000 documents, but it is bulk-culled down[210] before beginning review by making such legal decisions as custodian ranking and phasing, date ranges, and file types. In other words, irrelevant date ranges, file types (such as music or graphics), and custodians are culled out.

Your next step is to identify the relevant documents from the 40,000 remaining after bulk culling. This is the previously described first-pass relevancy review where predictive coding is primarily used. It sets the stage for the protections review, where documents that were coded likely relevant, and only those documents, are then re-reviewed for privilege and confidentiality, redacted, labeled, and logged. They are often also issue-tagged at this stage for the later use and convenience of trial lawyers. Mistakes in first-pass relevancy review are also corrected; for example, an attorney may find that a document predicted to be relevant is not relevant in that attorney's judgment. Some mistakes will always be made by the machine in the probability projection process, no matter how many iterations there are. But it is not uncommon to reduce the errors to 20% or less, depending on the difficulty of the search.

The first-pass relevancy review used to be done (and still is as of late 2013 by most lawyers and review companies) by having a lawyer actually look at—meaning skim or read—each of the 40,000 documents.

---

meantime, the 38% ratio it is at least an interesting starting point for analysis and discussion.

[210] *Culling* is step six in the EDBP. *Supra* Diagram 3.

Using low paid contract lawyers, this kind of first-pass relevancy review typically goes at a rate of 50 to 100 files per hour. But by using predictive coding, a skilled search expert, who must also be an SME for predictive coding to work, can attain speeds in excess of 10,000 files per hour for first-pass review. A good SME, therefore, can use machine training and determine file relevancy at a speed at least 1,000 times faster than a contract lawyer, and with far more accurately. That is why the SME with good software can charge 20 times as much as a contract lawyer, if not more, and still do the first-pass review at a fraction of the cost.[211]

In my experimental review of the 699,082 Enron documents for evidence concerning involuntary employee terminations, a fairly simple relevancy determination, my first-pass review was completed at an average speed of 13,444 files per hour.[212] Speeds such as this are common in many types of employment law issues, but similar speeds are attainable in other types of cases as well.[213]

Returning to the small case example of only 40,000 documents, let us assume a modest, AI-enhanced, first-pass review speed of 2,000 files per hour. That means a SME could complete the review in 20 hours.[214] It would probably take the SME about 3 hours to master the particular factual issues in the case, so let us assume a total time of 23 hours and a review rate for this SME of $550 per hour (in a small case like this, SMEs at relatively low rates are common, whereas the SME rates can be much higher in larger cases, but the speed of review and savings realized can also be much larger). That means an expense for first-pass review (excluding software charges) of $12,650, which is still less than half the cost of traditional manual review. Under a traditional contract lawyer review, where we assume a very fast speed (for them) of 75 files per hour, and a low, unmotivated lawyer rate of $50 per hour, you have a projected fee of $26,666.67.

---

[211] *See* Gabriel Techs. Corp. v. Qualcomm Inc., No. 08cv1992 AJB (MDD), 2013 WL 410103, at *9–10 (S.D. Cal. Feb. 1, 2013).

[212] *See* Ralph Losey, *Predictive Coding Narrative: Searching for Relevance in the Ashes of Enron (Restatement)*, E-DISCOVERY TEAM (Mar. 18, 2013, 2:27 PM), http://e-discoveryteam.com/2013/03/18/predictive-coding-narrative-searching-for-relevance-in-the-ashes-of-enron-restatement/.

[213] For instance, I recently completed another more complex fraud case review of over 1.5 million documents at an average speed of 35,831 files per hour. I did this review myself in one week's time, as I happened to be the only SME available for this project.

[214] Typical predictive coding review projects involve far more documents than this to review and so are able to attain faster speeds; still, I have done it in small cases with only 40,000 documents before. The math and cost savings still work with small projects like this if the predictive coding software cost is not too high.

Thus, even though the SME's $550 rate is 11-times higher than the contract lawyer's rate, since the SME is 26.67 times faster, the net savings are still greater than 50%. That is because it would take the contract lawyers 533.33 hours to complete the project, and, importantly, they would necessarily do so with a far lower accuracy rate.[215] They are likely to find far fewer relevant documents than the automated SME approach. This makes clear the power and importance of SMEs doing predictive coding work, and why, along with their current scarcity, they are now in such demand.[216]

Again returning to the example, the slower protections review comes after the first-pass review. Now the highly skilled SMEs are no longer required. The lower-paid contract lawyers can do the review on the documents the SMEs have determined to be relevant. Assume that the first-pass review found that 10,000 of the 40,000 documents were relevant. This means that 10,000 documents are subject to confidentiality protections review.[217] Let us assume this work goes at an average rate of 50 files per hour. This means a final pass review should be completed in 200 hours at a cost of $10,000. So the *base minimum* review cost for both passes is $22,650.

I say *base minimum* because there are additional expenses beyond just contract reviewer time, including the expense of partner and senior associate management time, direct supervision of contract lawyers, quality control reviews, et cetera, plus software costs, which, depending on the vendor and the particular deal, can sometimes be very high. Let us assume that there is another $7,000 cost here, for a total expense of $29,650. You would then have completed your review of 40,000 documents at a cost of $0.74 per document. That is pretty good. But in larger projects, where millions of documents are involved with more realistic prevalence rates, frequently less than 5%, the savings are even higher, and the per-document rate even lower, sometimes much lower.

All of these costs could be estimated in advance by having a bank of experience to draw upon to know the likely costs-per-file range. Still,

---

[215] *See, e.g.*, Maura R. Grossman & Gordon V. Cormack, *Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?*, 32 PACE L. REV. 267, 287–88 (2012); Grossman & Cormack, *supra* note 39, at 14–17, 24; Roitblat, Kershaw & Oot, *supra* note 40, at 79; Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697, 714–15 (2000).

[216] *See* David Cowen, *Job Market Heating Up for E-Discovery Technologists, Managers, and Attorneys*, E-DISCOVERY TEAM (Feb. 17, 2013, 8:30 PM), http://e-discoveryteam.com/2013/02/17/job-market-heating-up-for-e-discovery-technologists-managers-and-attroneys/.

[217] *Protections Review* is step eight in the EDBP. *Supra* Diagram 3.

practitioners should remember that even in the world of repeat litigation, like many employment law claims, all projects are different. All document sets are different. They have to, as I like to say, get their hands dirty in the digital mud.[218] Practitioners have to know their ESI collection, which they can only accomplish by spending time reading sample documents themselves. Even, for example, in the type of ESI most common in e-discovery today—email and attachments—the variances in email collections can be enormous.

The review speeds and thus review costs depend on the *density*[219] of the documents, the type of documents, and the general difficulty in understanding the documents. For example, emails are easier to read than spreadsheets, and shorter documents are generally easier to review than longer documents. The difficulty of the relevancy determinations also has a major impact on the speed of review. That is where the art of estimation comes in, and success will depend on your comprehension and detailed understanding of the project. Just as in building cost estimation, the practitioner must understand the blueprints and specifications of any project before having the capacity to make a valid estimate.

This is especially true of the SME work. You need to do some sampling to see what review rates apply. How long will it take these particular SMEs or contract reviewers to do the tasks assigned to them in this case with this data? Sampling is the only reliable way to answer that, especially when it comes to the all-important prevalence calculations.

### 4. A Big Data Example

Let us change the scenario somewhat for a final example. Assume there are 10,000,000 documents *after culling* for the SMEs to review. Assume sampling by an SME showed a prevalence of 10% (somewhat high), and a predictive coding review rate of 10,000 files per hour (somewhat slow for Big Data reviews). This means that only around 1,000,000 documents will need final protection review.[220] More sampling shows the contract reviewers using advanced AI-based techniques

---

[218] *See* Ralph Losey, *"The Hacker Way"—What the E-Discovery Industry Can Learn From Facebook's Management Ethic*, E-DISCOVERY TEAM (Aug. 18, 2013, 9:10 PM), http://e-discoveryteam.com/2013/08/18/the-hacker-way-what-the-e-discovery-industry-can-learn-from-facebooks-management-ethic/.

[219] *Density*, *yield*, and *prevalence* are terms that all refer to the percent of relevant documents in larger collection. In raw, unfiltered data such as email collections, the percent of relevant documents is usually less than 5% and often far less than 1%.

[220] 10% of 10,000,000 = 1,000,000.

(smart routing, et cetera) will be able to review 1,000,000 documents at the rate of 100 files per hour.

With this information from sampling, you can now estimate a total first-pass review cost of $1,000,000 ($1,000 per hour SME fee x 1,000 hours). Note that this $1,000,000 charge compares very well to the actual $2,829,349.10 charge approved in one large case in 2013 as a costs award for computer assisted review.[221] Next you can estimate a total final-pass protection review cost of $250,000 ($25 per hour contract lawyer fee x 10,000 hours).

Also assume, from experience, that other supervision fees and software costs are likely to total another $150,000. The total cost estimate for the project would thus be $1,400,000. That represents a cost to review the total corpus of 10,000,000 documents of only $0.14 a document.[222]

Too high you say? Perhaps it is not proportionate to the value of the case? Maybe it is not proportionate to the expected probative value in this case from these 10,000,000 documents, which is something your sampling can indicate and can provide evidence to support? Then use ranking to further limit the review costs.

If the SME's identification of 1,000,000 likely relevant documents was based on a 50% or higher probability ranking using predictive coding, then try a higher ranking cut-off. Again, with experience this becomes fairly easy to do using sampling and good software. Maybe a 75% or higher ranking cut-off will bring the document count down from 1,000,000 to 250,000. Or maybe you just arbitrarily decide to use the top ranked 200,000 documents because that is all you can afford, or all you think is proportionate for this data and case. That may result in only reviewing documents ranked 79% or higher. Either way, you are now only passing the strongest documents along for second-pass review. You are only producing the documents most likely to have the strongest probative value.

Using the higher cut off, the cost for second-pass protection review would then be 25% of what it was, reduced from $250,000 for review of 1,000,000 documents, to $62,500 to review 250,000 documents. The other fees and costs also drop in your experience by 50%, from $150,000 to $75,000. The total estimate is now $1,137,500, instead of $1,400,000. It

---

[221] This cost is significantly less than the fee approved in *Gabriel Technologies Corp. v. Qualcomm, Inc.* No. 08cv1992 AJB (MDD), 2013 WL 410103 (S.D. Cal. Feb. 1, 2013); Defendants Qualcomm, Inc., Snaptrack Inc., & Norman Krasner's Motion for Attorneys' Fees at 26, Gabriel Techs. Corp. v. Qualcomm Inc., No. 08cv1992 AJB (MDD) (S.D. Cal. Oct. 12, 2012), ECF No. 332-1.

[222] $1,400,000 / 10,000,000 = 0.14.

has gone down to just over $0.11 a document.[223] Assume this $1,137,500 number is now within your legal judgment to be proportional to this document request. It is now within your budget. You are done, and you now try to implement it within projected costs. Sometimes you succeed and the total costs are almost exactly what you projected. Other times you will go over, or sometimes maybe even come in under budget. With experience, your estimates become more reliable. Typically, a good estimator will estimate slightly on the high side so as to be more likely to surprise with savings.

If the $1,137,500 number was still not proportional in your judgment or your client's opinion, there are many other things to try. Typically I would focus on the bulk culling before the SME first-pass relevancy review. Change the custodian count or date range (but please, do not filter using keyword search). Bring the initial 10,000,000 documents down to 5,000,000 documents, then do the math. Thus, you may be talking about around $700,000, back to fourteen cents per document. Is that within the budget? Is that an amount that a court is likely to force you to spend anyway?

Another approach, one you have to take if further bulk culling is not possible, is to only review a smaller top range of the probable relevant documents. For instance, just review the top 10%, the documents with a probable-relevant ranking of 90% or higher. In some cases, it may even be appropriate and reasonable to only review the top 1%, those with a 99% or higher probable-relevant ranking. The quantity and quality of the top 1% may be so good that you do not need to see any additional documents. After all, sometimes it only takes one smoking-gun-type document to win or lose a case.

---

[223] The costs of review have come way, way down in the past few years for those who are using AI-based methods. For some context on the $0.14 and $0.11 per document numbers used in this example, back in 2007 the Department of Justice spent $9.09 per document for review in the *Fannie Mae* case, even though it used contract lawyers for the review work. *In re* Fannie Mae Secs. Litig., 552 F.3d 814, 817 (D.C. Cir. 2009) ($6,000,000/660,000 emails). There were no comments by the court that this price was excessive when the government later came back and sought cost-shifting. At about the same time, Verizon paid about $6.11 per record for a massive second-review project that enjoyed large economies of scale and, again, utilized contract review lawyers. Roitblat, Kershaw & Oot, *supra* note 40, at 73, 79 ($14,000,000 to review 2.3 million documents in four months). A large construction case that went to trial in 2012 incurred a charge per file of $2.85 to process, host, and review 2,700,000 files comprising more than 17 million pages using contract lawyers paid $85 per hour. Tampa Bay Water v. HDR Eng'g, Inc., No. 8:08-CV-2446-T-27TBM, 2012 WL 5387830, at *2, *15 (M.D. Fla. Nov. 2, 2012); s*ee also* Losey, *supra* note 29. In 2011, before AI-enhanced software started to become available, I still saw an average cost of $5.00 per file for reviews.

### 5. All Review Projects Are Different

In order to make a valid estimate for bottom-line-driven proportional review, you must closely study the case and review project goals. It is not enough to have handy per-file cost estimates. This move to actual examination of the ESI at issue, and study of the specific review tasks, is equivalent to the move in construction estimation from rough estimates based on average per square foot prices to a careful study of the building's plans and specifications and a site visit with inspection and measurements of all relevant conditions. No builder would bid on a project without first doing the detailed, real-world estimation work. Lawyers must do the same for this method to succeed.

Even in the same organization, when just dealing with email, the variances between custodians can be tremendous. Some, for instance, may have large amounts of privileged communications. This kind of email takes the most time to review, and if relevant, to log. High percentages of confidential documents, especially partially confidential ones, can also significantly drive up the costs of the second-pass review. All of the many unique characteristics of ESI collections can affect the speed of review and total costs of review. That is why parties must look at the data and test-sample the emails in the collection to make accurate predictions. Estimation in the blind is never adequate. It would be like bidding on a building without first reading the plans and specs.

Even when you have dealt with a particular client's email collection before, a repeat customer so to speak, the estimates can still vary widely depending on the type of lawsuit, the issues, and the amount of money in controversy or the general importance of the case. The opposing counsel and judge can also have a big impact on your analysis. The less sophisticated they are on these subjects, the more difficult the task, and the more important it is to engage in fair and respectful education efforts.

Although this may seem counter-intuitive, it is easiest to conduct e-discovery in complex, big-ticket cases, especially if the goal is to do so in a proportional manner. If there is a billion dollars at issue, a reasonable budget for ESI review is fairly large. On the other hand, proportional e-discovery in small cases is a real challenge, no matter how simple they supposedly are. Many cases that are small in monetary value are still very complex. And complex or not, all cases today have a lot of ESI. The medium to small size cases are where bottom-line-driven proportional review has the highest application for cost control and the greatest promise to bring e-discovery to the masses.

### *C. The More-Bang-for-the-Buck-Bottom-Line-Ranked Approach Is Good for Both the Requesting Party and the Producing Party*

When you are able to use ranking and predictive coding in a bottom-line-driven proportional review, it is much easier to persuade the requesting party to accept your proposed budgetary constraints. Failing that, it is much easier to persuade the court. The use of AI and predictive-coding ranking so that you only review and produce the best documents, the ones with the highest relevancy ranking, is a win/win proposal. It gives everyone the most truth for the dollar. This benefits both the producing party, who can thereby budget and avoid disproportionate burdens, and the requesting party. The requesting party benefits by a smart search system that finds more relevant documents—indeed, the most important documents. They benefit by not wasting their valuable time and resources reviewing irrelevant or marginally relevant documents. They are not overburdened by a document dump, an overly large production where they have to sort through thousands of barely relevant documents to find a few gems. The plaintiffs in the large, multi-district, class-action case, *Kleen Products*, reached the same conclusion, which is one reason why they tried to force the defendants to use predictive coding in their productions.[224]

In spite of the *Kleen Products* precedent, a producing party will often need to sell the benefits of these new methods to the requesting party. The requesting party will be more likely to cooperate if they understand the value to *them* of these methods. This often requires the producing party to provide some reasonable degree of transparency for the proposed review processes. For instance, tell them if you have an experienced, high quality SME lined up to direct the machine learning; share the SME's qualifications and experience.

As discussed, it is important to also engage the requesting party in relevancy dialogues. Make sure you are training the machine to find the documents that are really wanted. Clarify the target. If need be, share some examples early on of the relevant documents to be used in the training. Invite them to provide documents they consider relevant to use in the training. In some cases it may even make sense to invite them to fabricate documents to use for training. You can do that yourself as well, with or without their participation, or even knowledge. It makes a powerful persuasive tool to document your good faith attempts to try to find documents the requesting party is looking for, even if they would be seriously damaging to your case should they exist.

---

[224] Kleen Prods. LLC v. Packaging Corp. of Am., No. 10 C 5711, 2012 WL 4498465, at *5 (N.D. Ill. Sept. 28, 2012).

Try to make the AI search at least a somewhat collaborative effort. Input on gray-area documents, where relevance is uncertain, can often have a big effect on machine learning.[225] If the requesting party refuses to cooperate, for instance, by refusing to give a clear idea of what they are looking for, then document your efforts. As shown by the three proportionality cases, you need to take any disputes to the judge as early as possible.[226]

Use of a collaborative approach, even if it is largely unreciprocated and only partial, is the best way to convince a requesting party that your estimates and proportionality positions are reasonable. It is the best way to show the requesting party that you are not still stuck in the old paradigm of *hide-the-ball* discovery games. I cannot overstate how important it is to develop *trust* between opposing counsel on discovery. Often, the only way to do that is through some level of transparency. You do not have to disclose all of your search secrets, but you may have to keep the requesting party at least partially informed and involved in the process. That is what cooperation looks like. It involves honest, good-faith communications. That builds trust and so makes it easier to represent your client's interests. It also makes it easier to fulfill the Rule 1 dictates of speedy, inexpensive, and just litigation.

## CONCLUSION

The future of discovery involves new methods of technology-assisted discovery where Man and Machine work together to find the core truth. This day will come; in fact, it is already here. As the science fiction writer William Gibson said: "The future is already here. It's just not evenly distributed yet."[227] The key facts needed to try a case and to do justice can be found in any size case, big and small, at an affordable price, but you have to embrace change and adopt new legal and technical methodologies. The Bottom-Line-Driven Proportional Review method is part of that answer, and so too is advanced-review software at affordable prices. When the two are used together, it is a marriage made in heaven.

---

[225] *See supra* notes 93–95 and accompanying text.

[226] *See supra* Part III.C.

[227] Pagan Kennedy, *William Gibson's Future Is Now*, N.Y. TIMES, Jan. 15, 2012, at BR1.