FINDING THE SIGNAL IN THE NOISE: INFORMATION GOVERNANCE, ANALYTICS, AND THE FUTURE OF LEGAL PRACTICE

Bennett B. Borden[*] and Jason R. Baron[**]

Cite as: Bennett B. Borden & Jason R. Baron, *Finding the Signal in the Noise: Information Governance, Analytics, and the Future of Legal Practice*, 20 RICH. J.L. & TECH. 7 (2014), http://jolt.richmond.edu/v20i2/article7.pdf.

INTRODUCTION

[1] In the watershed year of 2012, the world of law witnessed the first concrete discussion of how predictive analytics may be used to make legal practice more efficient. That the conversation about the use of predictive analytics has emerged out of the e-Discovery sector of the law is not all that surprising: in the last decade and with increasing force since 2006—with the passage of revised Federal Rules of Civil Procedure that

[*] Mr. Borden is a partner in the Commercial Litigation section at Drinker Biddle & Reath, LLP, Washington, D.C., where he serves as Chair of the Information Governance and e-Discovery Group. He is Co-Chair of the Cloud Computing Committee and Vice Chair of the e-Discovery and Digital Evidence Committee of the Science and Technology Law Section of the ABA. He is also a founding member of the steering committee for the Electronic Discovery Section of the District of Columbia Bar. B.A., with highest honors, George Mason University; J.D., *cum laude*, Georgetown University Law School.

[**] Mr. Baron serves as Of Counsel in the Information Governance and e-Discovery Group, Drinker Biddle & Reath, LLP, Washington, D.C, and is on the Adjunct Faculty at the University of Maryland. He formerly served as Director of Litigation at the National Archives and Records Administration, and is a former steering committee Co-Chair of The Sedona Conference Working Group 1 on Electronic Document Retention and Production. B.A., *magna cum laude,* Wesleyan University; J.D., Boston University School of Law. The authors wish to thank Drinker Biddle & Reath associates Amy Frenzen and Nicholas Feltham for their assistance in the drafting of this article. The views expressed are the authors' own and do not necessarily reflect the views of any institution, public or private, that they are affiliated with.

expressly took into account the fact that lawyers must confront "electronically stored information" in all its varieties—there has been a growing recognition among courts and commentators that the practice of litigation is changing dramatically. What needs now to be recognized, however, is that the rapidly evolving tools and techniques that have been so helpful in providing efficient responses to document requests in complex litigation may be used in a variety of complementary ways to the discovery process itself.

[2] This Article is informed by the authors' strong views on the subject of using advanced technological strategies to be better at "information governance," as defined herein. If a certain evangelical strain appears to arise out of these pages, the authors willingly plead guilty. One need not be an evangelist, however, but merely a realist to recognize that the legal world and the corporate world both are increasingly confronting the challenges and opportunities posed by "Big data."[1] This Article has a modest aim: to suggest certain paths forward where lawyers may add value in recommending to their clients greater use of advanced analytical techniques for the purpose of optimizing various aspects of information governance. No attempt at comprehensiveness is aimed for here; instead, the motivation behind writing this Article is simply to take stock of where the legal profession is, as represented by the emerging case law on predictive coding represented by *Da Silva Moore*,[2] and to suggest that the expertise law firms have gained in this area may be applied in a variety of related contexts.

[3] To accomplish what we are setting out to do, we will divide the discussion into the following parts: first, a synopsis of why and how predictive coding first emerged against the backdrop of e-Discovery. This discussion will include a brief overview of predictive coding with

[1] *See infra* text accompanying notes 47-49 for a definition.

[2] Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 192 (S.D.N.Y. 2012), *aff'd sub nom*. Moore v. Publicis Groupe SA, 2012 U.S. Dist LEXIS 58742 (S.D.N.Y. Apr. 26, 2012) (Carter, J.).

references to the technical literature, as the subject has been recently covered exhaustively elsewhere. Second, we will define what we mean by “Big data,” “analytics,” and “information governance,” for the purpose of providing a proper context for what follows. Third, we will note those aspects of an information governance program that are most susceptible to the application of predictive coding and related analytical techniques. Perhaps of most value, we wish to share a few “early” examples of where we as lawyers have brought advanced analytics, like predictive coding, to bear in non-litigation contexts and to assist our clients in creative new ways. We fully expect that what we say here will be overrun with a multitude of real-life use cases soon to emerge in the legal space. Armed with the knowledge that we are attempting to catch lightning in a bottle and that law reviews on subjects such as this one have ever decreasing “shelf-lives”[3] in terms of the value proposition they provide, we proceed nonetheless.

A. The Path to *Da Silva Moore*

[4] *The Law of Search and Retrieval.* In the beginning, there was manual review. Any graduate of a law school during the latter part of the twentieth century who found herself or himself employed before the year 2000 at a law firm specializing in litigation and engaged in high-stakes discovery remembers well how document review was conducted: legions of lawyers with hundreds if not thousands of boxes in warehouses, reviewing folders and pages one-by-one in an effort to find the relevant needles in the haystack.[4] (Some of us also remember “Sheparding” a case to find subsequent citations to it, using red and yellow booklets, before automated key-citing came along.) Although manual review continues to remain a default practice in a variety of more modest engagements, it is

[3] We recognize the paradox of articles living “forever” on the Internet, especially when published in online journals such as this one, while at the same time ever more rapidly becoming obsolete and out of date.

[4] *See generally* The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery,* 8 SEDONA CONF. J. 189, 198 (2007) [hereinafter *Sedona Search Commentary*].

increasingly the case that all of discovery involves "e-Discovery" of some sort—that the world is simply "awash in data"[5] (starting but by no means ending with email, messages and other textual documents of all varieties), and that it will increasingly be the unusual case of any size where documents in paper form still loom large as the principal source of discovery.

[5] At the turn of the century, the dawning awareness of the need to deal with a new realm of electronically stored information ("ESI") led to burgeoning efforts on many fronts, including, for example, the creation of The Sedona Conference working group on electronic document retention and production, members of which drafted *The Sedona Conference Principles: Addressing Electronic Document Production* (2005; 2d ed. 2007) and its "prequel," *The Sedona Guidelines: Best Practice Guidelines and Commentary for Managing Records and Information in the Electronic Age* (2005; 2d ed. 2007). These early commentaries, including a smattering of pre-2006 case law,[6] recognized that changes in legal practice were necessary to accommodate the big changes coming in the world of records and information management within the enterprise. Subsequent developments would constitute various complementary threads leading to the greater use of analytics in the legal space.

[6] First, part of that early recognition was that in an inflationary universe of rapidly expanding amounts of ESI, new tools and techniques would be necessary for the legal profession to adapt and keep up with the times.[7] By the time of adoption of the revised Federal Rules of Civil Procedure in 2006, which expressly added the term "ESI" to supplement "documents" in the rule set applicable to discovery practice, the legal

[5] THOMAS H. DAVENPORT & JINHO KIM, KEEPING UP WITH THE QUANTS: YOUR GUIDE TO UNDERSTANDING AND USING ANALYTICS 1-2 (2013).

[6] *See Sedona Search Commentary*, *supra* note 4, at 200-201 nn.16-19.

[7] *See, e.g.*, George L. Paul & Jason R. Baron, *Information Inflation: Can The Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, ¶ 2 (2007), http://law.richmond.edu/jolt/v13i3/article10.pdf.

profession was well aware of the need to perform automated searches in the form of keyword searching within large data sets as the only realistically available means for sorting information into relevant and non-relevant evidence in particular engagements, be they litigation or investigations. So too, it was recognized early on in commentaries[8] and followed by case law[9] that keyword searching, as good a tool as it was, had profound limitations that in the end do not scale well. At the end of the day, even being able to limit or cull down a large data set to one percent of its original size through the use of keywords leaves the lawyer with the near impossible task of manually reviewing a very large set of documents at great cost.[10]

[7] Second, in evolving e-Discovery practice after 2006, a growing recognition also occurred around the idea that e-Discovery workflows are an "industrial" process in need of better metrics and measures for evaluating the quality of productions of large data sets. As recognized in *The Sedona Conference Commentary on Achieving Quality in E-discovery* (Post-Public Comment Version 2013):

> The legal profession has passed a crossroads: When faced

[8] *Id.*; *see Sedona Search Commentary*, *supra* note 4, at 201-202; Mia Mazza, Emmalena K. Quesada, & Ashley L. Stenberg, *In Pursuit of FRCP1: Creative Approaches to Cutting and Shifting Costs of Discovery of Electronically Stored Information*, 13 RICH. J.L. & TECH. 11, ¶ 46 (2007), http://jolt.richmond.edu/v13i3/article11.pdf.

[9] *See* Victor Stanley v. Creative Pipe, 250 F.R.D. 251, 256-7 (D. Md. 2008); *see also* United States v. O'Keefe, 537 F. Supp. 2d 14, 23-24 (D.D.C. 2008); William A. Gross Const. Ass'n v Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 135 (S.D.N.Y. 2009); Equity Analytics, LLC v. Lundin, 248 F.R.D. 331, 333 (D.D.C. 2008); *In re* Seroquel Prod. Liab. Litig., 244 F.R.D. 650, 663 (M.D. Fla. 2007). *See generally* Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, 17 RICH. J.L. & TECH. 9, ¶ 11 n.38 (2011), http://jolt.richmond.edu/v17i3/article9.pdf.

[10] *See* Paul & Baron, *supra* note 7, at ¶ 20; *see also* Bennett B. Borden, *The Demise of Linear Review*, WILLIAMS MULLEN E-DISCOVERY ALERT, Oct. 2010, at 1, http://www.clearwellsystems.com/e-discovery-blog/wp-content/uploads/2010/12/E-Discovery_10-05-2010_Linear-Review_1.pdf.

> with a choice between continuing to conduct discovery as it had "always been practiced" in a paper world—before the advent of computers, the Internet, and the exponential growth of electronically stored information (ESI)—or alternatively embracing new ways of thinking in today's digital world, practitioners and parties acknowledged a new reality and chose progress. But while the initial steps are completed, cost-conscious clients and over-burdened judges are increasingly demanding that parties find new approaches to solve litigation problems.[11]

[8] The Commentary goes on to suggest that the legal profession would benefit from greater

> awareness about a variety of processes, tools, techniques, methods, and metrics that fall broadly under the umbrella term "quality measures" and that may be of assistance in handling ESI throughout the various phases of the discovery workflow process. These include greater use of project management, sampling, machine learning, and other means to verify the accuracy and completeness of what constitutes the "output" of e-[D]iscovery. Such collective measures, drawn from a wide variety of scientific and management disciplines, are intended only as an entry-point for further discussion, rather than an all-inclusive checklist or cookie-cutter solution to all e-[D]iscovery issues.[12]

[9] Indeed, more recent case law has recognized the need for quality control, including through the use of greater sampling, iterative methods,

[11] THE SEDONA CONFERENCE, THE SEDONA CONFERENCE COMMENTARY ON ACHIEVING QUALITY IN E-DISCOVERY 1 (Post-Public Comment Version 2013), *available at* www.thesedonaconference.org/publications (for publication 15 SEDONA CONF. J. ___ (2014) (forthcoming)).

[12] *Id.*

and phased productions in line with principles of proportionality.[13] Still other case law has emphasized the need for cooperation among parties in litigation on technical subjects, especially at the margins of, or outside the range of, lawyer expertise if not basic competence.

[10] Active or supervised "machine learning," as referred to here in the context of e-Discovery, refers to a set of analytical tools and techniques that go by a variety of names, such as "predictive coding," "computer-assisted review," and "technology assisted review." As explained in one helpful recent monograph:

> Predictive coding is the process of using a smaller set of manual reviewed and coded documents as examples to build a computer generated mathematical model that is then used to predict the coding on a larger set of documents. It is a specialized application of a class of techniques referred to as supervised machine-learning in computer science. Other technical terms often used to describe predictive coding include document (or text) "classification" and document (or text) "categorization."[14]

[13] *See, e.g.*, *William A. Gross Constr.*, 256 F.R.D. at 136; *Seroquel*, 244 F.R.D. at 662. *See generally* Bennett B. Borden et al., *Four Years Later: How the 2006 Amendments to the Federal Rules Have Reshaped the E-Discovery Landscape and Are Revitalizing the Civil Justice System*, 17 Rich. J.L. & Tech. 10, ¶¶ 30-37 (2011), http://jolt.richmond.edu/v17i3/article10.pdf; Ralph C. Losey, *Predictive Coding and the Proportionality Doctrine: A Marriage Made in Big Data*, 26 REGENT U. L. REV. 7, 53 n.189 (2013) (collecting cases on proportionality).

[14] RAJIV MAHESHWARI, PREDICTIVE CODING GURU'S GUIDE 21 (2013); *see also* Baron, *supra* note 9, at ¶ 32, n.124 (stating predictive coding and other like terminology as used by e-Discovery vendors); Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1, 4 (2013), http://www.fclr.org/fclr/articles/html/2010/grossman.pdf; Nicholas M. Pace & Laura Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*, RAND INSTITUTE FOR CIVIL JUSTICE 59 (2012), *available at* http://www.rand.org/pubs/monographs/MG1208.html (defining predictive coding).

[11] And as stated in *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* (Post-Public Comment Version 2013):

> Generally put, computer- or technology-assisted approaches are based on iterative processes where one (or more) attorneys or [Information Retrieval] experts train the software, using document exemplars, to differentiate between relevant and non-relevant documents. In most cases, these technologies are combined with statistical and quality assurance features that assess the quality of the results. The research . . . has demonstrated such techniques superior, in most cases, to traditional keyword based search, and, even, in some cases, to human review.
>
> The computer- or technology-assisted review paradigm is the joint product of human expertise (usually an attorney or IR expert working in concert with case attorneys) and technology. The quality of the application's output, which is an assessment or ranking of the relevance of each document in the collection, is highly dependent on the quality of the input, that is, the human training. Best practices focus on the utilization of informed, experienced, and reliable individuals training the system. These individuals work in close consultation with the legal team handling the matter, for engineering the application. Similarly . . . the defensibility and usability of computer- or technology-assisted review tools require the application of statistically-valid approaches to selection of a "seed" or "training" set of documents, monitoring of the training process, sampling, and quantification and verification of the results.[15]

[15] THE SEDONA CONFERENCE, THE SEDONA CONFERENCE BEST PRACTICES COMMENTARY ON THE USE OF SEARCH AND INFORMATION RETRIEVAL METHODS IN E-DISCOVERY (Post-Public Comment Version 2013), *available at* www.thesedonaconference.org/publications (for publication in 15 SEDONA CONF. J. ___

A discussion of the mathematical algorithms that underlie predictive coding is beyond the intended scope of this Article, but the interested reader should refer to references cited at the margin to understand better what is "going on under the hood" with respect to the mathematics involved.[16]

[12] *The* Da Silva Moore *Precedent.* The various threads in search and retrieval law, including the need for advanced search methods applied to document review in a world of increasingly large data sets, were well known by 2012. In February 2012, drawing on recent research and scholarship emanating out of the Text Retrieval Conference (TREC) Legal Track[17] and the 2007 public comment version of The Sedona Conference

(2014)). For an excellent, in-depth discussion of how a practitioner may use predictive coding in e-Discovery, with references to experiments by the author, see Losey, *supra* note 13, at 9.

[16] *See, e.g.*, *Sedona Search Commentary*, *supra* note 4, at app. 217-223 (describing various search methods); Douglas W. Oard & William Webber, *Information Retrieval for E-Discovery*, 7 FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL 100 (2013), *available at* http://terpconnect.umd.edu/~oard/pdf/fntir13.pdf; 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), http://www.umiacs.umd.edu/~oard/desi5/additional/Baron-Jason-final.pdf. For good resources in the form of information retrieval textbooks, see GARY MINER, ET AL., PRACTICAL TEXT MINING AND STATISTICAL STRUCTURED TEXT DATA APPLICATIONS (Elsevier: Amsterdam) (2012); CHRISTOPHER D. MANNING, PRABHAKAR RAGHAVAN, & HINRICH SCHUTZE, INTRODUCTION TO INFORMATION RETRIEVAL (2008).

[17] *See TREC Legal Track*, U. MD., http://trec-legal.umiacs.umd.edu (last visited Feb. 23, 2014) (collecting Overview reports from 2006-2011) (as explained on its home page, "[t]he goal of the Legal Track at the Text Retrieval Conference (TREC) [was] to assess the ability of information retrieval techniques to meet the needs of the legal profession for tools and methods capable of helping with the retrieval of electronic business records, principally for use as evidence in civil litigation."); *see also* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 11, ¶¶ 3-4 (2011), http:/jolt.richmond.edu/v17i3/article11.pdf; Patrick Oot, et al., *Mandating Reasonableness in a Reasonable Inquiry*, 87 DENV. U.L. REV. 533, 558-559 (2010); Herbert Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70,

Search Commentary,[18] Judge Peck approached the *Da Silva Moore* case as an appropriate vehicle to provide a judicial blessing for the use of predictive coding in e-Discovery. In doing so, however, Judge Peck's opinion may also be viewed as setting the stage for greater use of analytics generally in the information governance practice area, beyond "mere" e-Discovery.

[13] Plaintiffs in *Da Silva Moore* brought claims of gender discrimination against defendant advertising conglomerate Publicis Groupe and its United States public relations subsidiary, defendant MSL Group.[19] Prior to the February 2012 opinion issued by Judge Peck, the parties had already agreed that defendant MSL would use predictive coding to review and produce relevant documents, but disagreed on methodology.[20] Defendant MSL proposed starting with the manual review of a random sample of documents to create a "seed set" of documents that would be used to train the predictive coding software.[21] Plaintiffs would participate in the creation of the "seed set" of documents by offering keywords.[22] All documents reviewed during the creation of the "seed set," relevant or irrelevant, would be provided to plaintiffs.[23]

77-79 (2010), *available at* http://onlinelibrary.wiley.com/doi/10.1002/asi.21233/full; *see generally* Pace & Zakaras, *supra* note 14, at 77-80.

[18] *Sedona Search Commentary*, *supra* note 4, at 192-193.

[19] Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 183 (S.D.N.Y 2012), *aff'd sub nom*. Moore v. Publicis Groupe SA, 2012 U.S. Dist LEXIS 58742 (S.D.N.Y. Apr. 26, 2012) (Carter, J.).

[20] *Id.* at 184-87.

[21] *Id.* at 186-87.

[22] *Id.* at 187.

[23] *Id.*

[14] After creation of the seed set of documents, MSL proposed using a series of "iterative rounds" to test and stabilize the training software.[24] The results of these iterative rounds would be provided to plaintiffs, who would be able to provide feedback to further refine the searches.[25] Judge Peck accepted MSL's proposal.[26] Plaintiffs filed objections with the district judge on the grounds that Judge Peck's approval of MSL's protocol unlawfully disposed of MSL's duty under Federal Rule of Civil Procedure 26(g) to certify the completeness of its document collection, and the methodology in MSL's protocol was not sufficiently reliable to satisfy Federal Rule of Evidence 702 and *Daubert*.[27]

[15] Judge Peck found the plaintiffs' objections to be misplaced and irrelevant.[28] With respect to Federal Rule of Civil Procedure 26(g), Judge Peck commented that no attorney could certify the completeness of a document production as large as MSL's. Moreover, Federal Rule of Civil Procedure 26(g) did not require the type of certification plaintiffs described.[29] Further, Federal Rule of Evidence 702 and *Daubert* are applicable to expert methodology, not to methodologies used in electronic discovery.[30] Judge Peck went on to note that the decision to allow computer-assisted review in this case was easy because the parties agreed

[24] *Da Silva Moore*, 287 F.R.D. at 187.

[25] *Id.*

[26] *Id.*

[27] *Id.* at 188-89.

[28] *Id.*

[29] *Da Silva Moore*, 287 F.R.D. at 188.

[30] *Id.* at 188-89 (citing Daubert v. Merrell Dow Pharms., 509 U.S. 579, 585 (1993)). *But cf.* David J. Waxse & Benda Yoakum-Kris, *Experts on Computer-Assisted Review: Why Federal Rule of Evidence 702 Should Apply to Their Use*, 52 WASHBURN L.J. 207, 219-23 (2013) (arguing that the *Daubert* standard should be applied to experts presenting evidence on ESI search and review methodologies)

to this method of document collection and review.[31] While computer-assisted review may not be a perfect system, he found it to be more efficient and effective than using manual review and keyword searches to locate responsive documents.[32] Use of predictive coding was appropriate in this case 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 under Rule 26(b)(2)(C), and (5) the transparent process proposed by MSL.[33]

[16] In issuing this opinion, Judge Peck became the first judge to approve the use of computer-assisted review.[34] He also stressed the limitations of his opinion, stating that computer-assisted review may not be appropriate in all cases, and his opinion was not intended to endorse any particular computer-assisted review method.[35] However, Judge Peck encouraged the Bar to consider computer-assisted review as an available tool for "large-data-volume cases" where use of such methods could save significant amounts of legal fees.[36] Judge Peck also stressed the importance of cooperation, or what he called "strategic proactive disclosure of information." If counsel is knowledgeable about the client's key custodians and fully explains proposed search methods to opposing

[31] *Id.* at 189.

[32] *Id.* at 190-91; *see* Grossman & Cormack, *supra* note 17, at ¶ 61.

[33] *Da Silva Moore*, 287 F.R.D. at 192.

[34] *Id.* at 193.

[35] *Id.*

[36] *Id.*

counsel and the court, those proposed search methods are more likely to be approved. To sum up his opinion, Judge Peck noted that "[c]ounsel no longer have to worry about being the 'first' or 'guinea pig' for judicial acceptance of computer-assisted review. . . . Computer-assisted review now can be considered judicially-approved for use in appropriate cases."[37] In the two years since *Da Silva Moore*, in addition to cases in which the parties have agreed upon a predictive coding methodology,[38] courts have confronted the issue of having to rule on either the requesting or responding party's motion to compel a judicial "blessing" of the use of predictive coding (however termed). In *Global Aerospace*,[39] the responding party asked that the court approve its own use of such technique; in *Kleen Products*, the requesting party made an ultimately unsuccessful demand for a "do-over" in discovery, where the responding party had used keyword search methods and the plaintiffs were demanding that more advanced methods be tried.[40] In the *EOHRB* case, the Court *sua sponte* suggested that the parties consider using predictive coding, including the same vendor.[41] And in the *In re Biomet* case,[42] the court approved a predictive coding methodology over the objections of the requesting party. These cases represent only some of the reported

[37] *Id.*

[38] *See, e.g.*, *In re* Actos (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2012 U.S. Dist. LEXIS 187519, at *20 (W.D. La. July 27, 2012).

[39] Global Aero. Inc. v. Landow Aviation, No. CL 61040, 2012 Va. Cir. LEXIS 50, at *2 (Apr. 23, 2012).

[40] Kleen Products, LLC v. Packaging Corp., No. 10 C 5711, 2012 U.S. Dist. LEXIS 139632, at *61-63 (N.D. Ill. Sept. 28, 2012).

[41] EORHB v. HOA Holdings, Civ. Ac. No. 7409-VCL (Del. Ch. Oct. 15, 2012), 2012 WL 4896670, *as amended in a subsequent order*, 2013 WL 1960621 (Del. Ch. May 6, 2013).

[42] *In re* Biomet M2a Magnum Hip Implant Prods. Liab. Litg., No. 3:12-MD-2391, 2013 U.S. Dist. LEXIS 84440, at *5-6, *9-10 (N.D. Ind. Apr. 18, 2013).

decisions to date, and we suspect that there will be dozens of reported cases and many more unreported ones in the near term.

[17] As recognized in these cases (implicitly or explicitly), as well as in a growing number of commentaries,[43] predictive coding is an analytical technique holding the promise of achieving much greater efficiencies in the e-Discovery process. Notwithstanding *Da Silva Moore's* call to action, it needs to be conceded, however, that the research has not proven that active machine learning techniques will *always* achieve greater scores than keyword search or manual review.[44] Additionally, we bow to the reality that in a large class of cases the use of predictive coding is currently infeasible or unwarranted, especially as a matter of cost.[45]

[18] Nevertheless, it seems apparent that the legal profession finds itself in a new place—namely, in need of recognizing that artificial intelligence techniques are growing in strength from year to year—and thus it appears to be only a matter of time until a much greater percentage of complex cases involving a large magnitude of ESI will constitute good candidates for lawyers using predictive coding techniques, both as available currently and as improved with future technological progress. As William Gibson once put it, "the future is here, it's just not evenly distributed."[46]

[43] *See, e.g.*, 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, 344-345 (2013); Harrison M. Brown, Comment, *Searching for an Answer: Defensible E-Discovery Search Techniques in the Absence of Judicial Voice*, 16 CHAP. L. REV. 407, 407-409 (2013); Jacob Tingen, *Technologies-That-Must-Not-Be-Named: Understanding and Implementing Advanced Search Technologies in E-Discovery*, 19 RICH. J.L. & TECH 2, ¶ 63 (2012), http://jolt.richmond.edu/v19i1/article2.pdf.

[44] *See* Pace & Zakara, *supra* note 14, at 61-65.

[45] *Cf.* Losey, *supra* note 13, at 68.

[46] Pagan Kennedy, *William Gibson's Future is Now,* N.Y. TIMES (Jan. 13, 2012), www.nytimes.com/2012/01/15/books/review/distrust-that-particular-flavor-by-william-gibson-book-review.html?pagewanted=all&_r=0.

B. Information Governance and Analytics in the Era of Big Data

[19] We are now in a post-*Da Silva Moore*, "Big data" era where lawyers are on constructive (if not actual) notice of a world of technology assisted review techniques available at least in the sphere of e-Discovery. The proposition being advanced is that the greater revelation of *Da Silva Moore* is how similar the techniques being put forward as best practices in e-Discovery fit a larger realm of issues familiar to lawyers, many of which fall within what is increasingly being recognized as "information governance" practice. It is here where we can break new ground in our legal practice by recommending the use of these advanced techniques to solve real-world problems of our clients. First, however, some definitions are in order to better frame the legal issues that will follow in Section C.

[20] *Big data.* It has been noted that "Big data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software."[47] Alternatively, "Big data" is a term that "describe[s] the technologies and techniques used to capture and utilize the exponentially increasing streams of data with the goal of bringing enterprise-wide visibility and insights to make rapid critical decisions."[48]

[21] The fact that the data encountered within the corporate enterprise increasingly is indeed "big" means, at least according to Gartner, that it not only has volume, but velocity and complexity as well.[49] As Bill

[47] Chris Snijders, Uwe Matzat, & Ulf-Dietrich Reips, *"Big Data": Big Gaps of Knowledge in the Field of Internet Science*, 7 INT'L J. INTERNET SCI. 1 (2012), http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf.

[48] Daniel Burrus, *25 Game Changing Trends That Will Create Disruption & Opportunity (Part I),* DANIEL BURRUS, http://www.burrus.com/2013/12/game-changing-it-trends-a-five-year-outlook-part-i/ (last visited Feb. 24, 2014).

[49] BILL FRANKS, TAMING THE BIG DATA TIDAL WAVE: FINDING OPPORTUNITIES IN HUGE DATA STREAMS WITH ADVANCED ANALYTICS 5 (John Wiley & Sons, Inc. ed., 2012)

Franks has put it, "What this means is that you aren't just getting a lot of data when you work with big data. It's also coming at you fast, it's coming at you in complex formats, and it's coming at you from a variety of sources."[50] These elements all significantly contribute to the challenge of finding signals in the noise.

[22] These definitions seem to get us closer to what makes Big data a new and interesting phenomenon in the world: it is not its volume alone, but the fact that we are able to "mine" large data sets using new and advanced techniques to uncover unexpected relationships, patterns and categories within these data sets, that makes the field potentially exciting. Indeed, "it is tempting to understand big data solely in terms of size. But that would be misleading. Big data is also characterized by the ability to render into data many aspects of the world that have never been quantified before; call it 'datafication.'"[51]

[23] *Analytics.* Second, we need to place "predictive coding" as one form of active machine learning in the context of the broader realm of "analytics." In their book, *Keeping Up With the Quants: Your Guide To Understanding and Using Analytics*,[52] authors Thomas Davenport and Jinho Kim provide a useful construct in categorizing the newly emergent field of "analytics": they define analytics to mean "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and add value," going on to say that "[a]nalytics is all about making sense of big data, and

(citing STEPHEN PRENTICE, CEO ADVISORY: 'BIG DATA' EQUALS BIG OPPORTUNITY (2011)).

[50] *Id.* at 5.

[51] Kenneth Neil Cukier & Viktor Mayer-Schoenberger, *The Rise of Big Data: How It's Changing the Way We Think About the World*, COUNCIL ON FOREIGN RELATIONS (Apr. 3, 2013), http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data.

[52] DAVENPORT & KIM, *supra* note 5.

using it for competitive advantage." The authors divide the world of analytics into three categories:

> (i) descriptive analytics – gathering, organizing, tabulating and depicting data;
> (ii) predictive analytics – using data to predict future courses of action; and
> (iii) prescriptive analytics – recommendations on future courses of action.[53]

[24] To the extent that "predictive coding" has been used to date to have machines "predict" relevancy in large ESI data sets, the term comfortably can be said to fall within category (ii). But the world of analytics is a larger universe, encompassing a greater number of mathematical magic tricks,[54] and this should be kept in mind as we choose to limit our discussion here to a few examples of how predictive coding as one form of analytics may be usefully applied in non-traditional contexts.[55]

[25] Corporations (much ahead of the legal profession) have rushed headlong during the past half-decade to use a variety of analytics to understand the Big data they increasingly hold, to add value, and to

[53] *Id.* at 3.

[54] *See id.* at 4-5 (providing a listing of various fields of research that make up a part of and comfortably fit within the broader term "Analytics," including statistics, forecasting, data mining, text mining, optimization and experimental design).

[55] For additional titles in the popular literature, see THOMAS H. DAVENPORT & JEANNE G. HARRIS, COMPETING ON ANALYTICS: THE NEW SCIENCE OF WINNING (2007); FRANKS, *supra* note 49; THORNTON MAY, THE NEW KNOW: INNOVATION POWERED BY ANALYTICS (John Wiley & Sons, Inc. ed., 2009); MICHAEL MINELLI, MICHELE CHAMBERS & AMBIGA DHIRAJ, BIG DATA ANALYTICS: EMERGING BUSINESS INTELLIGENCE AND ANALYTIC TRENDS FOR TODAY'S BUSINESSES (John Wiley & Sons, Inc. ed., 2013); ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE (John Wiley & Sons, Inc. ed., 2013).

improve the bottom line.[56] A 2013 AIIM study indicates that corporations find analytics to be useful in a variety of settings.[57]

[26] *Information Governance.* "Information governance," as defined in The Sedona Conference's recently published Commentary on the subject, means:

> an organization's coordinated, interdisciplinary approach to satisfying information legal and compliance requirements and managing information risks while optimizing information value. As such, Information Governance encompasses and reconciles the various legal and compliance requirements and risks addressed by different information focused disciplines, such as records and information management ("RIM"), data privacy, information security, and e-[D]iscovery.[58]

Or, as highlighted by the seminal law review article devoted to information governance written by Charles R. Ragan who quotes Barclay Blair in defining information governance as a "'new approach' that

[56] *See* DAVENPORT & KIM, *supra* note 5.

[57] *See* AIIM, BIG DATA AND CONTENT ANALYTICS: MEASURING THE ROI 9 (2013), *available at* http://www.aiim.org/Research-and-Publications/Research/Industry-Watch/Big-Data-2013. In a questionnaire asking "What type of analysis would you like to do/already do on unstructured/semi-structured data?", respondents identified over a dozen uses for analytics which they would consider of high value to their corporation, including: Metadata creation; Content deletion/retention/duplication; Trends/pattern analysis; Compliance breach, illegality; Fraud detection/prevention; Security re-classification/PII (personally identifiable information) detection; Predictive analysis/modeling; Data visualization; Cross relation with demographics; Incident prediction; Geo-correlation; Brand conformance; Sentiment analysis; Image/video recognition; and Diagnostic/medical. *Id*.

[58] THE SEDONA CONFERENCE, THE SEDONA CONFERENCE COMMENTARY ON INFORMATION GOVERNANCE 2 (2013), *available at* https://thesedonaconference.org/publication [hereinafter *Sedona IG Commentary*].

"builds upon and adapts disciplines like records management and retention, archiving business analytics, and IT governance to create an integrated model for harnessing and controlling enterprise information . . . [I]t is an evolutionary model that requires organizations to make real changes."[59]

[27] As the Sedona IG Commentary highlights, "many organizations have traditionally used siloed approaches when managing information."[60] The "core shortcoming" of this approach is "that those within particular silos are constrained by the culture, knowledge, and short-term goals of their business unit, administrative function, or discipline."[61] This leads in turn to key actors within the organization having "no knowledge of gaps and overlaps in technology or information in relation to other silos. . . ."[62] In such situations, "[t]here is no overall governance or coordination for

[59] Charles R. Ragan, *Information Governance: It's a Duty and It's Smart Business*, 19 RICH. J.L. & TECH. 12, ¶ 32 (2013), http://jolt.richmond.edu/v19i4/article12.pdf (internal quotation marks omitted) (quoting Barclay T. Blair, *Why Information Governance, in* INFORMATION GOVERNANCE EXECUTIVE BRIEFING BOOK, 7 (2011), *available at* http://mimage.opentext.com/alt_content/binary/pdf/Information-Governance-Executive-Brief-Book-OpenText.pdf). For additional useful definitions of what constitutes information governance, see *The Generally Accepted Recordkeeping Principles*, ARMA INT'L, http://www.arma.org/r2/generally-accepted-br-recordkeeping-principles (last visited Feb. 24, 2014) (setting out eight principles of IG, under the headings Accountability, Integrity, Protection, Compliance, Availability, Retention, Disposition and Transparency); Debra Logan, *What is Information Governance? And Why is it So Hard?*, GARTNER (Jan. 11, 2010), http://blogs.gartner.com/debra_logan/2010/01/11/what-is-information-governance-and-why-is-it-so-hard/ (defining IG on behalf of Gartner to be "the specification of decision rights and an accountability framework to encourage desirable behavior in the valuation, creation, storage, use, archival and deletion of information. It includes the processes, roles, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals.").

[60] *Sedona IG Commentary*, *supra* note 58, at 5.

[61] *Id.*

[62] *Id.*

managing information as an asset, and there is no roadmap for the current and future use of information technology."[63]

[28] The Sedona IG Commentary goes on to provide eleven principles of what constitutes good IG practices, of which Principle 10 is of special relevance to our discussion here: "An organization should consider leveraging the power of new technologies in its Information Governance program."[64] As stated therein,

> Organizations should consider using advanced tools and technologies to perform various types of categorization and classification activities. . . such as machine learning, auto-categorization, and predictive analytics to perform multiple purposes, including (i) optimizing the governance of information for traditional RIM [records and information management]; (ii) providing more efficient and more efficacious means of accessing information for e-discovery, compliance, and open records laws, and (iii) advancing sophisticated business intelligence across the enterprise.[65]

With respect to the latter category, the Commentary goes on to specifically identify areas where predictive analytics may be used in compliance programs "to predict and prevent wrongful or negligent conduct that might result in data breach or loss," as a type of "early warning system."[66] It is precisely this latter type of conduct that we wish to primarily explore in the next section, along with a few final words on using analytics with auto-categorization for the purpose of records classification and data remediation.

[63] *Id.*

[64] *Id.* at 25.

[65] *Sedona IG Commentary*, *supra* note 58, at 25.

[66] *Id.* at 27.

C. Applying the Lessons of E-Discovery In Using Analytics for Optimal Information Governance: Some Examples

[29] Advanced analytics are increasingly being used in the e-Discovery context because the legal profession has begun to realize the limitations of manual and keyword searching, while at the same time seeing how advanced techniques are at least as efficacious and far more efficient in a wide variety of substantial engagements. But more efficient and at least as equally effective at doing what, precisely? In e-Discovery, the primary information task involves separating relevant from non-relevant, and to a secondary degree, privileged from non-privileged information, in documents and ESI. Indeed, lawyers are under a duty to make "reasonable"—not perfect—efforts to find *all* relevant documents within the scope of a given discovery request.[67] The illusiveness of this quest in an exponentially expanding data universe is becoming increasingly apparent to many.[68]

[30] Moreover, the degree of success in being able to either find or demand substantial amounts of relevant information is not (nor should it be) the fundamental goal or point of engaging in e-Discovery.[69] Rather,

[67] *See* Pension Comm. of Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC, 685 F. Supp. 2d 456, 461 (S.D.N.Y. 2010). The information task in e-Discovery is therefore very unlike the user experience with the leading, well-known commercial search engines on the Web in, for example, finding a place for dinner in a strange city. For the latter project, few individuals religiously scour hundreds of pages of listings even if thousands of "hits" are obtained in response to a select set of keywords; instead they browse only from the first few pages of listings. Yet the lawyer is tasked with making reasonable efforts to credibly retrieve "the long tail" represented by "any and all" documents in response to document requests so phrased under Federal Rule of Civil Procedure 34.

[68] *See, e.g.*, Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 191 (S.D.N.Y. 2012), *aff'd sub nom*. Moore v. Publicis Groupe SA, 2012 U.S. Dist LEXIS 58742 (Apr. 26, 2012) (Carter, J.); *Pension Comm.*, 685 F. Supp. 2d at 461.

[69] *See* Bennett B. Borden et al., *Why Document Review Is Broken*, EDIG: E-DISCOVERY AND INFORMATION GOVERNANCE, May 2011, at 1, *available at* http://www.umiacs.umd.edu/~oard/desi4/papers/borden.pdf.

the liberal discovery rules that at least U.S. lawyers operate within have as their underlying purpose the ferreting out of important, material facts to the case at hand. The increasingly overwhelming nature of ESI poses clear technological obstacles to a lawyer en route to efficiently engaging in developing facts from all those relevant documents to determine what happened and why.[70] The promise of using an advanced analytical method such as predictive coding is its ability to quickly find and rank-order the *most* relevant documents for answering these questions. For once we determine how something happened and why, it is relatively straightforward to figure out the parties' respective rights, responsibilities, and even liability. That is precisely the point of litigation, and the purpose of the Rules that govern it.[71] And, facts drive it all.

[31] Given our increasing ability in litigation in finding the most relevant needles (i.e., facts) in the Big data haystack, it stands to consider whether similar methods may be successfully applied in non-litigation contexts. Somewhat paradoxically, however, experience indicates that there are advantages to dealing with *larger* volumes of data when applying analytical tools and methods to solve corporate legal issues. That is, while a vast amount of data residing in corporate networks and repositories admittedly poses complex information governance challenges, the volume of Big data also may be a boon to the investigator simply trying to figure out what happened. This is the case because there are simply many more data points from which to derive facts. One can liken the phenomenon to the difference in quality of a one-megapixel versus a ten-megapixel

[70] *An Insider's Look at Reducing ESI Volumes Before E-Discovery Collection*, EXTERRO, http://www.exterro.com/ondemand_webcast/an-insiders-look-at-reducing-esi-volumes-before-e-discovery-collection/ (last visited Feb. 24, 2014); Andrew Bartholomew, *An Insider's Perspective on Intelligent E-Discovery*, E-DISCOVERY BEAT (Sept. 11, 2013), http://www.exterro.com/e-discovery-beat/2013/09/11/an-insiders-perspective-on-intelligent-e-discovery/.

[71] *See* FED. R. CIV. P. 1 ("These rules . . . should be construed and administered to secure the *just, speedy, and inexpensive* determination of every action and proceeding.") (emphasis added).

picture: the difference in the quality of the image is a function of the greater density of points of illumination.

[32] Big data is more data, and more data means the potential for a more complete picture of what happened in a given situation of interest, assuming of course that the facts can be captured *efficiently*. The problem is not one of volume, but of visibility. In the era of Big data, the investigator with the more powerful analytical methods, who can search into vast repositories of ESI to draw out the facts that are critical to the question at hand, is king (or queen). This is where the skillful application of advanced analytics to Big data can bring about some remarkable results. The true strategic advantage of advanced analytics is the *speed* with which an accurate answer can be ascertained.[72]

[33] *True Life Example #1*.[73] A corporate client is being sued by a former employee in a whistleblower *qui tam* action.[74] Because of the False Claims Act allegations, the suit represented a significant threat to the company. The corporation retains counsel to understand the client's information systems as well as its key players, and to assist in the implementation of a litigation hold. Counsel strategically targets the data most likely to shed light on the facts. The law firm's Fact Development Team applies advanced analytics to 675,000 documents, and within four days knows enough to defend the client's position that the allegations are

[72] Borden et al., *supra* note 69, at 3.

[73] All of the "True Life Examples" referred to in this article are "ripped from" the pages of the author's legal experience, without embellishment.

[74] A qui tam suit is a lawsuit brought by a "private citizen (popularly called a 'whistle blower') against a person or company who is believed to have violated the law in the performance of a contract with the government or in violation of a government regulation, when there is a statute which provides for a penalty for such violations." *Qui Tam Action*, THE FREE DICTIONARY, http://legal-dictionary.thefreedictionary.com/qui+tam+action (last visited Feb. 24, 2014); *see also* United States *ex rel.* Eisenstein v. City of New York, 556 U.S. 928, 932 (2009) (defining a qui tam action as a lawsuit brought by a private party alleging fraud on behalf of the government) (internal citations omitted).

indisputably baseless. All of this is done before the answer to the Complaint was due.

[34] Armed with this information, counsel for the corporation approached plaintiff's counsel and asked to meet. Prior to the meeting, the corporation voluntarily produced 12,500 documents that laid out the parties' position precisely. Counsel then met with plaintiff's counsel and walked them through the evidence, laying out all the facts. The case ended up being settled within days for what amounted to nuisance value based on a retaliation claim—without any discovery, and at a small fraction of the cost budgeted for the litigation.

[35] This example indicates that the real power of advanced analytics is not merely in potentially reducing the cost of vexatious litigation, but rather the strategic *advantage* that comes with counsel getting to an answer quickly and accurately. This precise strategic advantage has many applications outside of litigation, each of which involves an aspect of optimizing information governance.

[36] Only a short step away from the direct litigation realm is using advanced analytics for investigations, either in response to a regulatory inquiry or for purely internal purposes. As we have already seen, corporate clients are often faced with circumstances where determining whether an allegation is true, and the scope of the potential problem if it is, is critically important. Often, management must wait, unsure of their company's exposure and how to remediate it, while traditional investigation techniques crawl along. However, with the skillful application of advanced analytics upon the right data set, accurate answers can be determined with remarkable speed.

[37] *True Life Example #2.* A highly regulated manufacturing client decided to outsource the function of safety testing some of its products. A director of the department whose function was being outsourced was offered a generous severance package. Late on a Friday afternoon, the soon-to-be former director sent an email to the company's CEO demanding four times the severance amount and threatened to go to the

company's regulator with a list of ten supposed major violations that he described in the email if he did not receive what he was asking for. He gave the company until the following Monday to respond.

[38] The lawyers were called in. They analyzed the list of allegations and determined which IT systems would most likely contain data that would prove their veracity and immediately pulled the data. Applying advanced analytics, the law firm's Fact Development Team analyzed on the order of 275,000 documents in thirty-six hours. By that Monday morning, counsel was able to present a report to the company's board indisputably proving that the allegations were unfounded.

[39] *True Life Example #3*. A major company received a whistleblower letter from a reputable third party alleging that several senior personnel were involved with an elaborate kickback scheme that also involved FCPA violations. If true, the company would have faced serious regulatory and legal issues, as well as major internal difficulties. Because of the extremely sensitive nature of the allegations, a traditional investigation was not possible; even knowing certain personnel were under investigation could have had immense consequences.

[40] The lawyers were tasked with determining whether there was any information within the company's possession that shed any light on the allegations. If there were, the company would proceed to take whatever steps were required. The investigation was of such a secret nature that no one was authorized to involve the internal IT staff. Fortunately, counsel knew the company and its information systems well. Over a weekend, they were able to pull 8.5 million documents from relevant systems using the law firm's personnel. This turned out to be a highly complex investigation involving a number of potential subjects, where the task involved tracking the subject's travel, meetings with suppliers, subsequent sales orders and fulfillments, rebates and promotions, all across several years.

[41] Again, applying advanced analytics, the law firm's Fact Development Team analyzed the 8.5 million documents in ten days. They

were able to prove that the allegations were largely baseless, and precisely where there were potential areas of concern. Counsel also was able to make clear recommendations for areas of further investigation and for modifying compliance tracking and programs. The company was able to act quickly and with certainty. These real-life use cases illustrate how the power of analytics enhances the ability of lawyers to provide legal advice under conditions of "certainty" previously unobtainable, at least in the past few decades of the digital era. "Certainty" is a somewhat foreign concept in the law—lawyers tend to be a conservative and caveating bunch, largely because certainty has historically been hard to come by, or at least prohibitively expensive. With advanced analytics and good lawyers who know how to use these new tools, that is no longer necessarily the case. There is so much data that if one cannot, after a reasonable effort, find evidence of a fact in the vastness of a company's electronic information (as long as you have the right information), the fact most likely is not true. Such has been illustrated, proving a negative is particularly useful in investigations.

[42] Using advanced analytics (and good lawyering) for investigations is not that far removed from using it for litigation: one is still attempting to find the answer to the question of what happened and why. But there are many other questions that companies would like to ask of their data. And indeed, both the analytics tools and the fact development techniques used in litigation and investigations can be "tuned" to solve a variety of novel issues facing our clients.

[43] For example, analytics can be used to vet candidates for political appointments as well as candidates for senior leadership positions. Due to the candid nature of the medium, providing access to corporate email coupled with using analytic capabilities allows for an accurate picture to be drawn *before* a decision is made with regard to making a candidate your next CEO or running mate. Analytics can be used to analyze business divisions to identify good and bad leaders, how decisions are made, why a division is more successful than another, and many more similar applications.

[44] Quite simply, a company's data is the digital imprint of the actions and decisions of all of its managers and employees. Having insight into those actions and decisions can be immensely valuable. That value has lain largely fallow, hidden in plain sight because the valuable wheat could not effectively be sifted from the chaff. With the proper application of advanced analytics, that is no longer the case. The answers we can obtain are limited only by the creativity of management in asking the right questions.

[45] *True Life Example #4.* Advanced analytics used upon the major acquisition of another company by a corporate client. As with most acquisitions, the client undertook traditional due diligence, gathering information from the target regarding its financial performance, customers, market share, receivables, potential liabilities, and came up with a valuation, an appropriate multiplier, and a final purchase price. Also as is typical, the acquisition agreement contained a provision such that if the disclosures made by the target were found to be off by a certain margin within thirty days of the acquisition, the purchase price would be adjusted.

[46] The moment the acquisition closed, the corporate client then owned all of the target's information systems. Having some concern about the bases for some of the target's disclosures, at the client's request counsel proceeded to use analytics on those newly acquired systems to determine what we could about those disclosures. Preparing a company for sale is a complicated affair, with many people involved in gathering information to present to the acquirer to satisfy due diligence. This gathering and presentation of information is done primarily through electronic means—and leaves a trail.

[47] Using advanced analytics, the law firm's Fact Development Team traced the compilation of the target's due diligence information, including all of the discussion that went along with it. They were able to understand the source of each disclosure, the reasonableness of its basis, and any weaknesses within it. They uncovered disagreements within the target over such things as what the right numbers were, or how much of a

liability to disclose. Using this information, counsel prepared a claim in accord with the adjustment provision seeking twenty-five percent of the purchase price totaling millions of dollars. The claim was primarily composed using quotes from their own documents. It is difficult to argue with yourself.

[48] As demonstrated, using advanced analytics in the form of predictive coding and similar technologies can accomplish some notable aims. But each of the prior examples uses data to look back to determine what has already occurred: the descriptive use of analytics.[75] This is extremely valuable. But for many of a law firm's clients, it would be even more useful to be able to catch bad actors while the misconduct was occurring, or even to predict misconduct before it happens.

[49] Based on the anecdotal experience gathered from many past investigations, the authors believe that certain kinds of misconduct follow certain patterns, and that when bad actors are acting badly, they tend to undertake the same kinds of actions, or are experiencing similar circumstances. For example, in our experience the primary factors that pertain to a person committing fraud are personal relationship problems, financial difficulties, drug or alcohol problems, gambling, a feeling of under appreciation at work, and unreasonable pressure to achieve a work outcome without a legitimate way to accomplish it (and so they attempt illegitimate ways to do so). These factors are often detectable in the electronic information the subject creates. Similarly, a person who is harassing or discriminating against others also tends to undertake specific actions and use particular language in communications. All of these indicia of misconduct are detectable using advanced analytics and skillful strategy.

[50] Lawyers have gotten quite good at finding this information when looking back in time. We thought, then, that it should not be too difficult to find this information while the misconduct is unfolding, or to identify warning signs that misconduct is likely to occur, and seek to provide relief of certain factors where possible or take corrective action when needed

[75] *See* DAVENPORT & KIM, *supra* note 5, at 3.

and as early as possible. So, we put this to the test, developing Early Warning Systems ("EWS") for some of our clients.

[51] The idea for an EWS first occurred to one of the authors when working on a pro bono matter with the ACLU in a case against the Baltimore Police Department ("BPD") alleging unconstitutional arrest practices in its Zero Tolerance Policing policies.[76] As a result of the case, the BPD agreed to, among other things, implement a tracking system whereby certain data points were collected regarding police officer conduct and arrest practices that research had proven were warning signs of potential problem officers.[77] The accumulation of certain data points with respect to an officer triggered a review of the officer's conduct, with various remediation outcomes.[78] We thought that a similar approach could be used for our clients.

[52] An EWS is a tricky thing to implement, and requires careful consideration of many factors, employee privacy at the forefront. However, with careful planning, policy development, and training, an effective EWS can be designed and implemented. Predictive analytics applications can be trained to search for indicia of the conduct, language, or factors across information systems. The specific systems to be targeted will vary depending on what is being sought and the systems most likely to contain it and will vary greatly from company to company. But, when properly trained and targeted, we have found these systems to be very

[76] *See* Amended Complaint and Demand for Jury Trial, NAACP v. Balt. City Police Dep't, No. 06-1863 (D. Md. Dec. 18, 2007), *available at* http://www.aclu-md.org/uploaded_files/0000/0205/amended_complaint.pdf.

[77] *See* CHARLES F. WELLFORD, JUSTICE ASSESSMENT AND EVALUATION SERVICES, FIRST STATUS REPORT FOR THE AUDIT OF THE STIPULATION OF SETTLEMENT BETWEEN THE MARYLAND STATE CONFERENCE OF NAACP BRANCHES, ET. AL. AND THE BALTIMORE CITY POLICE DEPARTMENT, ET. AL. 2 (2012), *available at* http://www.aclu-md.org/uploaded_files/0000/0207/first_audit_report_april_30.pdf; *see also Plaintiffs Win Justice in Illegal Arrests Lawsuit Settlement with the Baltimore City Police Department*, ACLU (June 23, 2010), https://www.aclu.org/racial-justice/plaintiffs-win-justice-illegal-arrests-lawsuit-settlement-baltimore-city-police-depar.

[78] *See* WELLFORD, *supra* note 77, at 2, 14.

effective in detecting and even preventing misconduct. We believe that this use of predictive analytics will become one of the most powerful applications of this technology in the near future.

[53] Moving from the business intelligence aspects of information governance to the arguably more prosaic field of records and information management, the authors also count themselves as true believers in the power of analytics to optimize traditional RIM (records and information management) functionality. A full discussion of archival and records management practices in the digital age is beyond the scope of this Article, but the interested reader will find a wealth of scholarly literature in the leading journals discussing how the traditional practice of records management is being transformed in the digital age. One of the authors has argued that predictive coding and like methods are the most promising way to open up "dark archives" in the public sector, such as digital collections of data appraised as permanent records (mostly consisting of White House email at this point), that for reasons of privacy or privilege will be otherwise inaccessible to the public for many decades to come.[79]

[54] In the authors' experience, email archiving using auto-categorization for recordkeeping purposes is available using existing software in the marketplace. In such instances, email is populated in specific "buckets" in a repository depending on how it is characterized, based on either the position of the creator or recipient of the email, the subject matter, or based on some other attribute appearing as metadata.[80] In the most advanced versions of auto-categorization software, the system "learns" as it is trained using exemplars in a seed set selected by subject matter experts (i.e., records managers or expert end users), via a protocol highly reminiscent of the methods adopted by the parties in *Da Silva*

[79] *See* Jason R. Baron & Simon J. Attfield, *Where Light in Darkness Lies: Preservation, Access and Sensemaking Strategies for the Modern Digital Archive*, *in* THE MEMORY OF THE WORLD IN THE DIGITAL AGE CONFERENCE: DIGITALIZATION AND PRESERVATION 580-595 (2012), http://www.ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf.

[80] *See id.* at 587.

Moore and similar cases. It is only a matter of time before predictive analytics is more widely used to optimize auto-classification while reducing the burden on end users to perform manual records management functions.[81]

[55] In similar fashion, the power of predictive analytics to reliably classify content after adequate training makes such tools optimal for data remediation efforts. The problem of legacy data in corporations is well known, and only growing over time with the inflationary expansion of the ESI universe.[82] Using advanced analytics to classify low value data, the chaos that is the reality of most shared drives and other joint data repositories, may potentially be reduced by orders of magnitude. The challenge of engaging in defensible deletion is one important aspect of optimizing information governance.[83]

CONCLUSION

[56] As was made clear at the outset, it is the authors' intent merely to scratch the surface of what is possible in the analytics space as applied to matters of importance for corporate information governance. No one has a one hundred percent reliable crystal ball, but it seems evident that as computing power increases, those forms of artificial intelligence that we have referred to here as analytics will themselves only grow in importance in both our daily and professional lives. By the end of this decade, we would be surprised if the following do *not* occur: pervasive use of business intelligence software; the use of more automated decision-making (also known as "operational business intelligence"); the use of alerts in the form of early warning systems including the type described above; much greater

[81] *See id.* at 588; *see also* Ragan, *supra* note 59, at ¶ 6.

[82] *See, e.g.*, THE SEDONA CONFERENCE, THE SEDONA CONFERENCE COMMENTARY ON INACTIVE INFORMATION SOURCES 2, 5 (2009), *available at* https://thesedonaconference.org/publication/The%20Sedona%20Conference®%20Commentary%20on%20Inactive%20Information%20Sources.

[83] *See Sedona IG Commentary*, *supra* note 58, at 20-22.

use of text mining and predictive technologies across a variety of domains.[84]

[57] All of these developments dovetail with the expected demand on the part of corporate clients for lawyers to be familiar with state of the art practices in the information governance space, as already anticipated by the type of technology that *Da Silva Moore* and related cases suggest. As best said in *The Sedona Commentary on Achieving Quality in E-Discovery*, "[i]n the end, cost-conscious firms, organizations, and institutions of all types that are intent on best practices . . . will demand that parties undertake new ways of thinking about how to solve e-[D]iscovery problems. . . ." [85] The same holds true for the greater playing field of information governance. Lawyers who have embraced analytics will have a leg up on their competition in this brave new space.

[84] *See* DAVENPORT & HARRIS, *supra* 55, at 176-78.

[85] The Sedona Conference, *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process*, 10 SEDONA CONF. J. 299, 325 (2009).