Introduction

Jason R. Baron

"There [were] 5 exabytes of data created between the dawn of civilization through 2003, but that much information is now created every two days, and the pace is increasing. People aren't ready for the technology revolution."

—Eric Schmidt, CEO of Google[1]

Each of the three editors of this volume graduated law school in 1980, which has meant that we have been firsthand witnesses to the transformation of legal practice and especially discovery practice during the past few decades. There was a time when discovery meant searching only through boxes containing paper files, where the big case simply meant searching through more boxes in the client's warehouse.

Discovery did not yet need an "e" as a prefix, and manual searches for relevant documents sufficed. Judge Andrew J. Peck notes this, as well, in his Foreword to this volume.

Fast forward to the present, and how the world of lawyering has changed. The present "inflationary" period of information exploding has been built on copying machines and personal computers in the 1970s, e-mail beginning widespread use in the late 1980s, and the opening of the desktop to the Internet and especially the World Wide Web in the 1990s. The pace of change has

1. *See* Marshall Kirkpatrick, *Google CEO Schmidt: "People Aren't Ready for the Technology Revolution,"* ReadWrite (Aug. 4, 2010), http://readwrite.com/2010/08/04/google_ceo_schmidt_people_arent_ready_for_the_tech/.

only continued to accelerate since the turn of the century, with the emergence of social media and mobile devices in the last decade transforming what it means to conduct business. As this book goes to print, we are on the cusp of the Internet of Things, with smart devices proliferating and generating new data streams and new forms of evidence to search.

Today, every lawyer conducting "discovery" in civil litigation needs to confront the fact that—no matter how large or small the case may be—it is insufficient to simply define the search task as being limited to finding relevant documents in traditional paper files. The legal profession lives and breathes in a world of "electronically stored information" (ESI), a term of art introduced into legal practice by virtue of the 2006 amendments to the Federal Rules of Civil Procedure. But what constitutes our doing a "reasonable" job in finding relevant evidence in a world exploding in data?

The initial approach lawyers took (and still take) to confronting large volumes of ESI is to rely on keyword searching, supplemented by manual searches, to cull out relevant and privileged material before a production is made to opposing counsel. Although these "time-tested" approaches have their defenders, simple reliance on manual and keyword searching increasingly is seen as inadequate to the task at hand, both on grounds of accuracy and efficiency, as compared with more advanced search techniques.

The editors of this book are readily willing to stipulate in advance that they have a strong bias in favor of advancing the cause of computer-assisted review and educating the profession on how more advanced search techniques work. In one way or another, they have spent the better part of the last 15 years engaged in initiating and participating in research projects[2] and academic conferences,[3] joining think tanks,[4] communicating through online media platforms,[5] writing law

2. *See, e.g.*, the TREC Legal Track, http://trec-legal.umiacs.umd.edu/.

3. *See, e.g.*, the DESI (Discovery of ESI) international workshop series, http://www.umiacs.umd.edu/~oard/desi6/.

4. *See, e.g.*, The Sedona Conference, www.thesedonaconference.org.

5. *See, e.g.*, e-Discovery Team blog, www.e-discoveryteam.com.

reviews,[6] authoring e-discovery books,[7] and teaching e-discovery in law and graduate schools, in evangelizing on the topic of how lawyers may conduct "better" searches of electronic evidence using smarter methods than manual and keyword searching. Along the way, we have been fortunate to encounter a number of brilliant lawyers and scholars at the cutting edge of e-discovery and information science, many of whom we are grateful to for their contributions to this volume.

This book is an attempt to catch lightning in a bottle; namely, to provide a set of perspectives on predictive coding and other advanced search techniques, as they are used today by lawyers in pursuit of e-discovery, in investigations, and in other legal contexts, such as information governance. We are painfully aware that the shelflife of publications such as the present work is not long. Nevertheless, we trust that a cross-section of related—and sometimes differing—perspectives on how today's advanced search methods at the cutting-edge of legal practice will prove illuminating to a greater legal audience.

The book is divided into four subparts, under the headings "Searching for ESI: Some Preliminary Perspectives," "Practitioner Perspectives," "Information Retrieval Perspectives; E-Discovery Standards," and "Analytics and the Law." As discussed at more length below, the chapters provide insights into predictive coding and other advanced search methods from the perspectives of the judiciary, from requesting and responding parties in litigation, and from information scientists who have been engaged in intensive study of the field of e-discovery search over the past decade.

The book is meant to appeal both to practitioners who are seeking knowledge of what predictive coding and other advanced search methods are all about, as well as to those members of the legal community who are "inside the bubble" of e-discovery already and wish

6. *See, e.g.*, Ralph C. Losey, *Predictive Coding and the Proportionality Doctrine: A Marriage Made in Big Data*, 26 Regent U. L. Rev. 7 (2013–14), https://ralphlosey.files.wordpress.com/2013/12/law_review_pcandpropor.pdf; Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on "Information Inflation" and Current Issues in Legal Search*, 17 Rich. J.L. & Tech. 9 (2011), http://jolt.richmond.edu/v17i3/article9.pdf.

7. *See, e.g.*, Michael D. Berman, Courtney Ingraffia Barton, & Paul W. Grimm, Managing E-Discovery and ESI From Pre-Litigation Through Trial (ABA 2012).

to be exposed to the latest, cutting-edge techniques. We would like to imagine that the book may also be read by lawyers who do not consider themselves litigators or e-discovery practitioners, but who wish to apply a knowledge of smart analytics in other legal contexts.

The reader should be aware that given the relative novelty of predictive coding and other advanced search methods, there have been and will continue to be disagreements over what constitutes "best practices" in the space, and the editors of course have their own preferences and biases. However, the book attempts to be inclusive of a range of views, not always necessarily our own. What follows is a summary of the contents of this volume, representing an attempt to key the reader into recurrent themes, open issues, and present-day controversies.

Searching for ESI: Some Preliminary Observations

Many of us litigators have been on a journey these past two decades, learning as we go about how difficult it is to find all relevant ESI through existing methods in what are increasingly large data haystacks. In Chapter 1, "The Road to Predictive Coding: Limitations on the Defensibility of Manual and Keyword Searching," Tracy Drynan and I "serve notice to the everyday practitioner of the issues surrounding still widely used manual and keyword search methodologies." The chapter consists of a literature review and a basic tutorial on the subject of keyword searching, as well as an overview of certain electronic tools and processes (e.g., deduplication) often used in connection with all search methods. The authors also discuss notions of defensibility, noting the important point to the practitioner that reasonableness, not perfection, is required when conducting a search for relevant ESI.

Part of the journey has been the recently emergent acceptance of predictive coding in the courts, following a half-decade of research studies, law reviews, and commentaries pointing to the deficiencies of keyword searching and the possibility of parties using more advanced techniques in pursuit of justice. A watershed moment was reached when Magistrate Judge Peck issued his decision in 2012 in *Da Silva Moore v. Publicis Groupe*.[8] In Chapter 2, "The Emerging Acceptance

8. 287 F.R.D. 182 (S.D.N.Y. 2012), *aff'd*, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012) (Carter, J.).

of Technology-Assisted Review in Civil Litigation," Alicia L. Shelton and Michael D. Berman discuss *Da Silva Moore* and survey its progeny, including the *In re Actos* and *Global Aerospace* cases, through to Judge Peck's later decision in *Rio Tinto.* The authors discuss the tension inherent in the judiciary permitting liberal discovery in an age of Big Data, while attempting to adhere to the goal of ensuring the "just, speedy, and inexpensive" determination of actions in accordance with Federal Rules of Civil Procedure 1.

Practitioner Perspectives

Maura R. Grossman and Gordon V. Cormack, the authors of Chapter 3, "A Tour of Technology-Assisted Review," have been thought leaders in the area of e-discovery search for a long time, even before publication of their seminal 2011 law review piece *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*,[9] prominently cited by Judge Peck in *Da Silva Moore.* In their chapter, the authors provide the practitioner with an overview of the distinctions among various automated tools and methods that are legitimately grouped within the "TAR" label, as distinguished from other aspects of search, analysis, and review. In so doing, the authors explain in clear language the differences between passive and active machine-learning techniques, as well as between simple versus continuous machine-learning. One of the hottest topics in e-discovery today is the efficacy of what the authors refer to as "continuous active learning" (CAL), and their article makes the business case for CAL methods being presently superior to all others.

In Chapter 4, Vincent M. Catanzaro, Samantha Green, and Sandra Rampersaud provide useful guidance on "The Mechanics of a Predictive Coding Work Flow." While recognizing that "one size does not fit all" in e-discovery, they argue for "commonalities among the various predictive coding applications that allow the user to begin a customized work flow" by starting at a general level. These include beginning with assessment or evaluation of your case as a good candidate for the use of advanced search methods, followed by training and validation of the tools and methods used.

9. 17 Rich. J.L. & Tech. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf.

Chapter 5 consists of Ralph C. Losey's "Reflections on the Cormack and Grossman SIGIR Study: The Folly of Using Random Search for Machine Training." Likening random search to searching only under a spotlight because it is easy to do, Ralph discusses why the justifications that have been offered to date on its continued behalf (most prominently, the introduction of lawyer bias), are found wanting. He goes on to describe the Cormack-Grossman study, which was designed to answer the question "Should training documents be selected at random, or should they be selected using one or more non-random methods, such as keyword search or active learning?" In accord with what the study has found, Ralph believes that employing continuing active learning constitutes a "superior method to quickly find the most relevant documents" in a large collection. Along the way, he observes the benefits of using "multimodal" methods of search to optimize the process even further.

Chapter 6 may be of special interest to the greater community of legal practitioners in state courts as well as federal. In the chapter "TAR for the Small and Medium Case," William F. (Bill) Hamilton makes the argument that in a large variety of smaller, more routine cases, counsel may profit from using forms of technology-assisted review and other automated methods, inexpensively, for such purposes as early case assessment, analyzing the opponent's document production, or even preparing for a deposition. The author makes the case that the "hidden promise" of advanced search methods will increasingly be seen "as a critical tool for the 99 percent, not just the 1 percent" of cases (and litigators).

Chapters 7, 8, and 9 present differing perspectives from the plaintiffs and defendants bar, as well as from the judiciary, on the hot button issue of how much transparency and cooperation and required to faithfully execute and get agreement on a given predictive coding method.

William P. Butterfield and Jeannine M. Kenney contribute a "plaintiffs" perspective in Chapter 7, "Reality Bites: Why TAR's Promises Have Yet to Be Fulfilled." They begin with the question, why, given the advantages of TAR, "has it not been more widely adopted by parties in appropriate cases?" After a survey of relevant case law, including the protocols used in *Da Silva Moore* and *In re Actos*, they make the

case that a greater level of transparency and cooperation is necessary to build a requesting party's trust that TAR-like methods have resulted in a better (meaning richer, not just speedier) result in productions. In noting the present-day objections of many in the defense bar to disclosure of details of the process used in conjunction with advanced search technologies, the authors nevertheless remain optimistic that when "lawyers and judges become better educated about the processes needed to employ TAR effectively, agreement about the specifics of TAR protocols should become easier to achieve."

Chapter 8 in turn contains an analysis of "Predictive Coding from a Defense Perspective: Issues and Challenges," authored by Ronni D. Solomon and three of her colleagues, Rose J. Hunter-Jones, Jennifer A. Mencken, and Edward T. Logan. Here, in making the case for more widespread use of predictive coding, the authors focus on the cost-saving potential when corporate defendants are required to produce large volumes of ESI in discovery, drawing from their own use case experiences. They also review what cases are a good fit for predictive coding, as well as the challenges associated with the training process for implementing predictive coding in the workflow. The authors go on to provide a defense bar perspective on the issues of transparency and cooperation, arguing for what they consider to be appropriate limitations on transparency to protect client's interests.

In Chapter 9, "Safeguarding the Seed Set: Why Seed Set Documents May Be Entitled to Work–Product Protection," the Hon. John M. Facciola and Philip J. Favro further weigh in on issues of cooperation and transparency in connection with whether "seed sets," that is, the initial subset of documents selected to train software to recognize and distinguish what constitutes a relevant document from ones that are not, are appropriately shielded from production under the work–product doctrine. As reflected in the tension between the positions advocated in Chapters 7 and 8, *supra*, the authors report on the diversity of opinions regarding disclosure of seed sets and responsiveness decisions, and suggest a nuanced approach. They argue that the common-law, paper-days' work–product doctrine holding that the selection and ordering of documents is work product provides the proper rule to apply, but they note significant limitations on its application in this unique context.

Another open issue in current day e-discovery practice is the judiciary's take on whether the Supreme Court's *Daubert* standard for the use of expert testimony in the courtroom applies in the discovery context, in connection with a court's evaluation of the propriety of using predictive coding or other advanced search methods. In Chapter 10, the Hon. David J. Waxse and Brenda Yoakum-Kriz set out the case in favor of a *Daubert* standard, in "Experts on Computer-Assisted Review: Why Federal Rule of Evidence 702 Should Apply to Their Use." Noting the clear disagreement "about whether electronic searching of ESI should be considered an expert process subject to the requirements of Rule 702 and *Daubert*-style challenges," the authors conclude "the better view is that search methodologies such as computer-assisted review should be treated as an expert process subject to Rule 702 and *Daubert* challenges." We know from Judge Peck's decision in *Da Silva Moore*, as well as his Foreword in this volume, that he has weighed in on the contrary side, finding that *Daubert* is inapplicable. This is an area of the law to watch.

Ralph Losey provides a deeper dive into predictive coding in Chapter 11, "License to Cull: A Two-Filter Document Culling Method That Uses Predictive Coding and Other Search Tools." The author is a leading proponent of what he has coined "multimodal" search and culling techniques, and in this piece he steps the reader through his recommended approaches in filtering documents during the collection and processing phases of e-discovery, as well as in using predictive coding techniques. In his words, "[T]he basic idea behind the two-filter method is to start with a very large pool of documents, reduce the size by a coarse first filter, then reduce it again by a much finer second filter." The author makes the important point that "[t]here is much more to efficient, effective review than just using software with predictive coding features. The methodology of *how* you do the review is critical" (emphasis in original).

Information Retrieval Perspectives; Standards in E-discovery

Preceding chapters have cited to a body of research that began around 2006 with the TREC Legal Track, and continued through later research, that acted to support the claims made that keyword searching

has limitations, and that more advanced search methods may in fact be more efficient and effective than either manual review or keyword searching.[10] In these evaluation studies, it has been of paramount importance to be able to measure how well one is doing when performing searches for ESI—something that lawyers historically have shied away from.

In this subpart, we first present leading experts in the burgeoning field of information retrieval (IR) taking on the job of explaining the metrics that we as lawyers need to understand to be able to measure or evaluate how well that predictive coding and other advanced search technologies are in fact doing. This is especially the case where the software algorithms that are at the heart of performing advanced searches are "black box" technologies—not easily or intuitively understood (at least by lawyers and judges). The need or desire for better measurement in turn has led to an increased focus on what the standard for measurement should be, and what kind of quality controls should be put into place. Ultimately, these lines of inquiry lead to a discussion of whether the legal profession can and should arrive at a standard for judging e-discovery search efforts.

In Chapter 12, "Defining and Estimating Effectiveness in Document Review," Dr. David D. Lewis cogently makes the case for seeing the e-discovery search problem through the lens of "text classification"—a well-known approach in which there has been "extraordinary progress in computer science, statistics, and related fields in recent decades." He goes on to explain (we might better say, "demystify") the quantitative, statistical measures that function as important tools in achieving, assuring, and demonstrating a high quality and cost-effective review. They are used to both manage the process and justify the results. He explains, for example, the lawyer's role in making the tradeoff between different measures (e.g., recall and precision) to suit the needs of the case. Along the way, he provides the reader with an in-depth, mini "textbook"-like course in IR.

Continuing with looking through the lens of IR in viewing e-discovery, in Chapter 13 Drs. Douglas W. Oard and William Webber approach the subject of "Metrics in Predictive Coding," by assuming

10. *See* Chapter 3, *supra*, authored by Maura R. Grossman & Gordon V. Cormack (and the research cited therein).

that a "core task of discovery" is "to find the highest proportion of relevant documents in the collection at the least cost." To that end, they set out a practitioner's guide to how to measure what they term the "cost-for-completion tradeoff," that is, "how standard metrics and evaluations help" in guiding lawyers to that goal, and "how a misunderstanding of those metrics" act as an obstacle to achieving that goal. The chapter is filled with practical descriptions of what recall, precision, and more arcane IR metrics mean, and goes on to address the potential pitfalls and shortcomings of predictive coding. The authors point to the system's dependency on the accuracy of the subject-matter expert's decision-making, noting, importantly, that "we don't yet fully understand the consequences" of errors in that process.

In Chapter 14, "On the Place of Measurement in E-Discovery," Dr. Bruce Hedin, Dan Brassil, and Amanda Jones further discuss measurement, in the form of sampling and estimation protocols, as an integral part of an e-discovery quality management regimen. As an initial matter, they discuss other essential building blocks of a sound review process, including the need for advanced planning and thorough topic analysis, the proper use of technology, the appropriate use of expertise, the importance of being adaptable, and the need to provide for clear and complete documentation of one's efforts. They then proceed to provide a comprehensive discussion of what they term "general principles governing the use of measurement in e-discovery," what the benefits are of employing rigor to measurement, and the reasons for continued resistance by the legal profession in doing so.

Gilbert S. Keteltas, Karin S. Jenson, and James A. Sherer, eloquently grapple with questions of standards in Chapter 15, "A Modest Proposal for Preventing e-Discovery Standards from Being a Burden to Petitioners, Clients, the Courts, or Common Sense." The chapter discusses a number of existing and proposed standards, some well-known and others that are novel to most practitioners (e.g., an ISO-based standard for e-discovery), why there are so many competing standards efforts, and what kind of standards operating outside the Federal rules can be fashioned that are acceptable to e-discovery practitioners. Along the way, they tackle such issues as whether standards should be about process or result, and how to build flexibility into any contemplated standards—including acknowledgement of the

principle that perfection is not required in practicing e-discovery or performing searches.

Analytics and the Law

With 20-20 hindsight, it seems almost self-evident that the advent of the use of advanced search techniques in the e-discovery context would come into being contemporaneously with the emergence of Big Data and the wholesale adoption of new forms of analytics across a variety of disciplines. In these chapters, we survey how advanced search methods of a similar sort are being integrated into a number of aspects of legal practice closely related to, but distinct from litigation and e-discovery.

A step away from the litigation arena, modern antitrust practice has been undergoing a similar transformation in its embrace of predictive analytics. In Chapter 16, "Algorithms at the Gate: Leveraging Predictive Analytics in Mergers, Acquisitions and Divestitures," Jeffrey C. Sharer and Robert D. Keeling explain how predictive analytics "has applications for both buyers and sellers and in all phases of the deal lifecycle." The authors also focus attention on how predictive analytics is being used when corporations must quickly respond to Hart-Scott-Rodino "second requests." As they point out, "the same technologies and workflows that have gained acceptance in the litigation context can be deployed much earlier in the information lifecycle to improve information governance and drive cost savings and productivity gains across the organization." These would include software being trained to make distinctions as to "whether a document is a contract or not, contains intellectual property or not, is a financial report or not, is a personnel record or not, and so on."

The authors believe that predictive coding "is already delivering significant reductions" in the time and the cost of productions in the antitrust space, without a reduction in quality.

Another application of the software analytics behind predictive coding has been their use in the emerging discipline of "information governance," or "IG." IG has been broadly defined by one entity as "[t]he activities and technologies that organizations employ to maximize the value of their information while minimizing associated risks

and costs."[11] The next three chapters describe how corporations, law firms, and legal practitioners can all benefit from incorporating—with appropriate customization—the kind of advanced analytics we have been discussing in the e-discovery arena.

In Chapter 17, Sandra Serkes presents "The Larger Picture: Moving Beyond Predictive Coding for Document Productions to Predictive Analytics for Information Governance." She acknowledges what she terms a "fundamental difference" as between IG and e-discovery, namely, that the latter task is consumed with (or limited to) the "safe" culling of relevant, nonprivileged documents for purposes of production—whereas in her view the "hallmark" of IG's use of predictive analytics is moving "*beyond* simple culling, into areas such as classification, organization, trendlining, and forecasting, and modeling past or future behaviors" (emphasis in original). The author shows how businesses, litigators, and law firms may all harness the power of many of the same forms of analytics used in e-discovery to perform better enterprise searching and internal case management.

Continuing in the same vein, Leigh Isaacs in Chapter 18 discusses "Predictive Analytics for Information Governance in a Law Firm: Mitigating Risks and Optimizing Efficiency." Law firms of course have significant security concerns, but they also experience personnel changes, and frequently must import or export volumes of carefully screened law firm data. In short, all of the questions facing American industry are magnified in the unique context of a law firm with its need to preserve client confidences and other fiduciary duties. The author shows how predictive coding may help supply defensible solutions, and how a business case for "return on investment" may be made for the use of such advanced tools and techniques.

In Chapter 19, "Finding the Signal in the Noise: Information Governance, Analytics, and the Future of Legal Practice," Bennett B. Borden and I set out the case for the legal profession embracing both IG and analytics, across a range of legal practice areas, as a way to "break new ground . . . to solve real-world problems of our clients." Use case examples are provided that demonstrate the power of predictive

11. *See* Information Governance Initiative, *Annual Report 2014: Information Governance Goes to Work*, http://www.iginitiative.com.

coding in the evaluation of large data sets handed over by clients for multiple nonlitigation purposes, including, for example, a law firm's evaluation of whistleblowing allegations, or whether a party received full information on the value of an acquired company in a merger and acquisition context. The authors go on to discuss the possible deployment of software as a form of "early warning system" to guard against the loss of trade secrets or even the filing of discrimination claims. In a new supplement to the original article as published, the continuing trend lines recognizing the importance of IG, data science, and the law are emphasized.

Next, in Chapter 20, "Preparing for the Near Future: Deep Learning and Law," author Kathryn Hume takes us on a journey beyond currently used search methods and algorithms, to a time when even more advanced methods springing from the field of "neural networks" and "deep learning" may yet have a place in e-discovery and the law. Neural networks are a type of machine learning using multiple computing layers designed to better mimic the brain. For our purposes, deep learning using these techniques holds out the promise of lawyers being better able to analyze multiple data feeds, not only from traditional texts, but also from audio and video sources as well. The author challenges us to think about what the "practice of law" may mean in a world of machine learning techniques.

The Grossman-Cormack Glossary of Technology-Assisted Review

As an Appendix to this volume, Maura R. Grossman and Gordon V. Cormack graciously have allowed the reprinting here of their Grossman-Cormack Glossary of TAR with a Foreword by Judge John M. Facciola. Since its publication in the *Federal Courts Law Review* in 2013, this important work has greatly contributed to the understanding by practitioners of the many and varied technical terms used by information retrieval experts and increasingly by the e-discovery community when discussing technology-assisted review methods and protocols.

As this book goes to print, there appear to be voices in the profession questioning whether predictive coding has been oversold or

overhyped, and pointing to resistance in some quarters to wholesale embrace of the types of algorithmics and analytics on display throughout this volume. Notwithstanding these critics, the editors of this volume remain serene in their certainty that the chapters in this book represent the future of e-discovery and the legal profession as it will come to be practiced into the foreseeable future, by a larger and larger contingent of lawyers. Of course, for some, the prospect of needing to be technically competent in advanced search techniques may lead to considerations of early retirement. For others, the idea that lawyers may benefit from embracing predictive coding and other advanced technologies is exhilarating. We hope this book inspires the latter feelings on the part of the reader.