PREDICTIVE CODING: ARTIFICIAL INTELLIGENCE COMES TO DISCOVERY

Ashley S. Lipson, Esq.*

I. INTRODUCTION

On a massive international scale, electronically stored information, universally labeled "ESI", has been replacing paper at an exponential rate. To the delight of tree lovers, but to the dismay of document producing litigants, our once simple, paper based discovery process has now become an electronic nightmare. Referring to discovery as "simple" might be an overstatement; admittedly, pre-ESI paper production was no picnic. Depending upon the size and nature of the litigation, it could have involved hundreds of hours of sorting, Bates-stamping, and photocopying, not to mention physical transportation. Nevertheless, as time consuming as the paper chase may have been, it pales in comparison to the *years* required to examine, analyze, and select documents from a small jump drive, now capable of holding several gigabytes, of ESI.[1]

---

* Ashley S. Lipson is a computer programmer and Professor of Law at the University of La Verne College of Law. He has been a trial lawyer for more than thirty years and has published many practice-related books (Prentice Hall, Times Mirror, Matthew Bender, Lexis/Nexis, and James Publishing) along with more than sixty articles, in addition to co-authoring a casebook on video game law. He is the inventor of the *Lawyers' Comprehensive Computer Document System* (see ASHLEY S. LIPSON, LAW OFFICE AUTOMATION (1986)) and the author/programmer of the popular *Objection!* computer game series. His works also include: *Guerrilla Discovery* (James Publishing Company) and *Mathematics, Physics and Finance for Lawyers* (Carolina Academic Press). Mr. Lipson has a B.A. in Telecommunications from Michigan State University; an M.A. in Mathematics from Wayne State University; a J.D. from St. Johns University School of Law; an L.L.M. in Tax Law from Wayne State University School of Law; an As.D in Computer Science; and a Post Degree Studies in Physics from University of Michigan.

[1] *See* Global Aerospace Inc. v. Landow Aviation, L.P., No. CL 61040 (Va. Cir. Ct., Apr. 23, 2012) (wherein it was effectively argued that responding to a document production request would have required

Physically, we can now store the Library of Congress in our back pockets and pass it to other individuals in seconds. In most litigation settings, the physical aspects of discovery are thus quite simple. The real problem lies in the review and selection process. Responding to a discovery request is no longer a simple matter of flipping through papers in a file cabinet. On the contrary, ESI, which comprises more than 90 percent of all documents presently produced, possesses some explosively expanding characteristics.

First and foremost of these characteristics is *reproducibility*, which permits us to effortlessly copy large documents at the speed of light. Think of how often we respond to a simple email by including the text of the original transmission along with all other prior communications. Second, and closely related, is *duplicability*, which allows us to rapidly create new documents by cutting and pasting (control+C/control+V) old ones, a plagiarist's dream come true. Thirdly, is the characteristic known as *persistence*. Like a holdover guest, ESI never goes away. Bad guests eventually die; ESI does not. The "delete key" is merely an illusion, a temptress designed to beckon clients onto the rocks of spoliation sanctions.[2] Finally, consider *variation.* The number of ESI producing devices increases as we speak. Gone are the days of the weary typist and the

---

reviewing over 2,000,000 documents, consuming ten years of billable time).

[2] Sanctions for deleting or destroying potentially relevant writings or evidence can be extreme. *See* Zubulake v. USB Warburg, LLC, 382 F. Supp. 2d 536 (S.D.N.Y. 2005); U.S. v. Phillip Morris USA, Inc., 327 F. Supp. 2d 21 (D.D.C. 2004) (involving a fine of $2,995,000); *see also* ASHLEY S. LIPSON, GUERRILLA DISCOVERY §§ 3.13, 3.76, 3.8.60, 8.86 (2003); ASHLEY S. LIPSON, DOCUMENTARY EVIDENCE § 3.08 (1985).

Underwood; ESI can now be found in e-mails, instant message conversations, jump/flash/thumb drives, mobile devices, CDs, DVDs, voice mail, digital photographs, iPods, iPads, and an unspeakable number of other products.

The problem, therefore, is not tangible transmission, but rather—once all of the documents are physically located and assembled, who then is going to read through the potential Grand Canyon of words and determine what should be produced and what should be withheld? For cases involving massive documentation, humans alone are not capable; computer assistance is, therefore, inevitable.

II. RESPONDING TO REQUESTS FOR PRODUCTION[3]

Were it not for matters such as privilege, relevance, secrecy, work product protection, embarrassment, and a general aversion to spreading all of one's private records about the planet, a party responding to a document request would have a relatively simple task. Merely turn everything over to the opposition; this, of course, would be subject to the minor inconvenience of gathering the data from our briefly mentioned variety of sources and devices. Reality, however, requires a prior review of documents to determine which are to be produced and which are to be withheld. For cases involving sizeable numbers, the physical handling tasks pale in comparison to the review part of the process. On the receiving end, some or many poor souls are then going to be required to analyze and determine which

[3] Rule 34 of the Federal Rules of Civil Procedure (along with its state counterparts) governs the laws pertaining to document discovery and production. FED. R. CIV. P. 34.

documents are relevant and usable for the subject litigation.

Determining precisely which documents to withhold and which to provide to the opposition is a high wire balancing act that becomes more difficult as the numbers increase. Falling to either side of the wire by producing either too few[4] or too many[5] documents has its own set of unpleasant consequences (scary citations relating to sanctions omitted).

III. TRADITIONAL SEARCH METHODS

Searching and gathering paper for the benefit of someone whom you or your client does not like (*i.e.*, opposing litigants) has never been a pleasant task. The search methodologies vary, depending primarily upon the number of documents involved.[6] Predictive coding, our target procedures, employs all of the following more traditional tools but in varying degrees. Familiarization with them is a prerequisite for understanding predictive coding.

A. *MANUAL SEARCHING*

For matters involving hundreds of writings or less, a simple manual review may be warranted. At best, a Boolean or keyword search might be considered, but nothing approaching the magnitude of work involved to justify predictive coding.

---

[4] The Federal Rules of Civil Procedure provides a wide range of sanctions for withholding documents, including fines, adverse inferences, defaults, and dismissals. FED. R. CIV. P. 37.

[5] Producing too many documents is also dangerous, even though clawback agreements purport to undo inadvertent disclosures of privileged material and pretend to un-ring the errant bell. *See* FED. R. EVID. 502(b).

[6] When determining "numbers," the use of the term "document" can range from something as voluminous as a detailed congressional report to a simple "*OMG*" reply to a text message.

B. *BOOLEAN SEARCHING*

For most electronic discovery, the simple Boolean search (often referred to as a "keyword search") still reigns supreme. The elementary basics that one learned in order to cope with Westlaw searches provide the basics. Three fundamental skills are required: (1) Familiarity with the documents *typically* generated by entities involved in the particular litigation; (2) familiarity with the terms, slang, and code words employed by the subject parties;[7] and, (3) an ability to use and manipulate strings (pieces of text) using your Boolean operators (and, or, not, etc.).

C. *TECHNOLOGY ASSISTED REVIEW*

Technology Assisted Review ("TAR") utilizes varying methods of document control and classification. In addition to linear methodologies (used in Boolean and keyword searches), TAR employs non-linear methodologies such as clustering, grouping, and de-duplication.

D. *E-VENDOR ASSISTED SEARCHING*

Numerous companies claim to provide solutions to the ESI search problems with a variety of services. These services include software, training, and personnel assistance.

A few years ago, a group of technicians and lawyers joined forces to study

[7] In the classic 1990 film *Goodfellas,* two gangsters committed an unforgivable sin by killing Benny Batts (a "made" man). Thereafter, they never referred to Benny by name, but rather as "that problem upstate." They had buried the body in upper New York. GOODFELLAS (Warner Bros. 1990).

the mounting ESI discovery problems and the effectiveness of the e-discovery vendors. "*The Text Retrieval Conference Legal Track*," a project jointly created by those diverse individuals, was designed to help resolve the e-discovery problems.[8] On average, the group's study concluded that, regardless of the claims or methodologies employed by a wide variety of vendors, the accuracy or "hit rate" was fairly consistent—between 22 percent and 57 percent of all relevant documents were identified. The various search engines, software products, and methodologies proved to be no more accurate than a standard Boolean search. But that was then. In the past few years, the number of e-discovery vendors has increased and so too have their skill levels and effectiveness.

IV. COMPUTER ASSISTED REVIEW

For massive quantities of ESI, typical of today's employment discrimination cases and actions involving SEC matters, attorneys are going to require assistance from all of the previously discussed methodologies and then some. We are now focusing on matters that humans, unassisted by computers, are incapable of handling. But, even computers require more than standard string search capabilities. That is where predictive coding enters the picture.

A. *PREDICTIVE CODING DEFINED*

There are as many definitions for predictive coding as there are e-discovery vendors. Most definitions serve as sales pitches, avoiding undue

[8] *See* Jason Kraus, *In Search of the Perfect Search*, A.B.A. J., Apr. 2009, at 38.

technicality.[9] For a non-commercial definition, consider Judge Peck's ruling in *Da Silva Moore v. Publicis Groupe*,[10] which provides the following: "[T]ools (different vendors use different names) that use sophisticated algorithms to enable the computer to determine relevance, based on interaction with (i.e., training by) a human reviewer." Alternatively, consider the following definition, designed to provide a more specific description of the procedure:

> A computer based methodology designed to accelerate the discovery process for a large group of electronically stored documents by using (1) humans ("coders") to segregate and mark ("code") a sample subset of the group, which (2) the computer would then record and analyze, (3) thereby enabling the analysis of that sample to predict the relevance and discoverability of the non-coded remainder of the group.[11]

B. *How Predictive Coding Works*

Predictive coding does not pretend to completely replace human analysis. Rather, it is a tool that greatly accelerates the review process. That process entails assembling all of the documents related to a particular issue and ranking and tagging them so that a human reviewer can control the process and later confirm its accuracy and relevance. From the multitude, a small manageable sample, referred to as "the seed set" is extracted and scrutinized by people who undergo at

---

[9] Consider, for example, the definition provided a leader in the field, Recommind: "Predictive Coding is a court-endorsed process that combines people, technology[,] and workflow to find key documents quickly, irrespective of keyword." *A Message from Recommind: How is Predictive Coding Revolutionizing the Economics of eDiscovery?*, ABAJOURNAL.COM, Jan. 22, 2013, http://www.abajournal.com/advertising/article/how_is_predictive_coding_revolutionizing_the_economics_of_ediscovery/. The term "court endorsed" may be overstated.

[10] Da Silva Moore v. Publicis Groupe, 2012 WL 607412 (S.D.N.Y., Feb. 24, 2012); *see also* Andrew Peck, *Search, Forward*, LAW TECH. NEWS, Oct. 2011, at 25, 29.

[11] Author's definition.

least a week of training. Technical and paralegal skills are preferred for these trainees. Their decisions (coding) regarding the relevance of the documents in the seed set will then instruct the computer as to the potential relevance of all documents containing or lacking similar strings. The key feature of predictive coding, therefore, is the ability of the software, assisted by coders (humans), to teach the computer to become more accurate as the process continues.

Predictive coding software does not entail any mysterious powers and abilities. It involves string comparisons (*i.e.*, letters, words, etc.) and search engines—nothing more, nothing less. The term "coding" or "tagging" refers to designating a document as either having or lacking a certain characteristic. For example, the "coder" (a human) could instruct the program to tag or mark every writing containing a particular name, date, or range of dates. In other words, the process involves classification. At its simplest, we might only code a document as "responsive," *i.e.*, discoverable or otherwise necessary to produce, or "non-responsive." If enough documents are coded into the system, the computer software will then, by comparing the most common strings that are deemed responsive, be able to "predict" those strings that are most likely to be found in the remaining responsive documents.

There are two basic methodologies for the predictive coding process: *sampling* and *observation*. For sampling, the computer will randomly select a relatively small subset of documents from the masses to form the seed set. The

computer software will then analyze the decisions of the coder with respect to common document characteristics and relevance. The software will simultaneously record and note matters such as author, recipient, date, titles, headings, subject matter, and keywords. Based upon this collected information, the software will then attempt to "predict" the value of the remaining documents, identifying and rating those writings that are likely to be of greatest relevance.

For observation predictive coding, the process begins with the human coder, who will select, analyze, and code the seed set. As before, the computer software will then analyze and begin to predict the value of the remaining documents. As the human coder continues, the computer program will refine and improve its predictions (in much the same fashion as a Google search attempts to predict your final search goal before you have completed typing your entry).

In both methods, sampling and observation, it is important to note that humans, not computers, make the ultimate decisions as to relevance. The computer software can only attempt to predict the value of the masses based upon human selection of the few. If the process performs correctly, the remaining documents will receive scores that should indicate the probability of their relevance and significance to the subject litigation. Vast amounts of work hours would then be spared.

C. *Judicial Acceptance*

Courts have gradually moved from the sidelines onto the battlefront over

the use of predictive coding. They prefer, of course, that the parties sort out their own discovery disputes by way of stipulation.[12] After all, at present, predictive coding as an "expertise," does not satisfy the requirements of Rule 702 of the Federal Rules of Evidence,[13] nor has it yet satisfied *Daubert*'s[14] gatekeeping function for expert testimony. Nor should it, until it becomes a substantive issue in the litigation as opposed to a matter that solely involves discovery. Trying to predict predictive coding's move from the sidelines into the gavels of the judges is another matter.[15]

When might a judge order the parties to engage in predictive coding? It is certainly expedient, at least from the court's perspective. It provides a rational solution to an otherwise overwhelming problem. It is far less expensive to spend

---

[12] Nat'l Day Laborer Org. Network v. U.S. Immigration & Customs Enforcement Agency, 877 F. Supp. 2d 87 (S.D.N.Y. 2012) (involving a massive request by plaintiffs under the *Freedom of Information Act* (*FOIA*)). The court said:

> The parties will need to agree on search terms and protocols—and, if necessary, testing to evaluate and refine those terms. If they wish to and are able to, then they may agree on predictive coding techniques and other more innovative ways to search. Plaintiffs will need to be reasonable in their demands—aware of the real cost that their massive FOIA request has imposed on the agencies—and will be restricted to seeking records from only the most important custodians on only the most important issues. Defendant agencies, in turn, will need to cooperate fully with plaintiffs. As in the past, the Court will supervise this process and provide a variety of mechanisms for resolving any disputes.

[13] Rule 702 states:

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if: (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.

FED. R. EVID. 702.

[14] Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

[15] In *Da Silva Moore v. Publicis Groupe,* 868 F. Supp. 2d 137 (S.D.N.Y. 2012), a discrimination action by female employees, the court determined that predictive coding was at least "judicially acceptable." For a more pessimistic stance, however, see *Da Silva Moore v. Publicis Groupe*, 2012 WL 1446534 (S.D.N.Y., Apr. 26, 2012): "Perhaps they are looking for an opinion concluding that: 'It is the opinion of this court that the use of predictive coding is a proper and acceptable means of conducting searches under the Federal Rules of Civil Procedure . . . .' If so, it will be a long wait."

several days to a week training coders than spending thousands of hours for manual searches. When an order for predictive coding is issued, presumably, any dissenting party would have the option of conducting his or her own manual search and, of course, paying the added cost.[16]

Da Silva Moore v. Publicis Groupe[17] involved a federal lawsuit by five named female plaintiffs against the defendant, Publicis Groupe, a large advertising conglomerate, alleging that it had a "glass ceiling" limiting women to entry level positions by use of systemic, company-wide gender discrimination against female employees. Estimates from the defendant's custodians suggested that there were over 3,000,000 electronic documents to be provided to the plaintiffs. In ordering the use of computer assisted coding with respect to jurisdictional discovery, Judge Peck provided additional insight into the process:

> Unlike manual review, where the review is done by the most junior staff, computer-assisted coding involves a senior partner (or [small] team) who review and code a "seed set" of documents. The computer identifies properties of those documents that it uses to code other documents. As the senior reviewer continues to code more sample documents, the computer predicts the reviewer's coding. (Or, the computer codes some documents and asks the senior reviewer for feedback.)
>
> When the system's predictions and the reviewer's coding

[16] One bold Virginia judge did it in *Global Aerospace Inc. v. Landow Aviation, L.P.*, No. CL 61040 (Va. Cir. Ct., Apr. 23, 2012). After hearing cries by defendants that reviewing over 2,000,000 documents would require ten years of billable time, Judge James Chamblin ruled: "[I]t is hereby ordered Defendants shall be allowed to proceed with the use of predictive coding for purposes of processing and production of electronically stored information." The judge acknowledged that the receiving party would still have the opportunity to question "the completeness of the contents of the production or the ongoing use of predictive coding." The court order allowed sixty days for processing, and an additional sixty days for production. Chamblin's order was in response to the defendant's motion requesting either that predictive technology be allowed or that the parties demanding a more expensive method pay the added costs.

[17] Da Silva Moore v. Publicis Groupe, 2012 WL 607412 (S.D.N.Y., Feb. 24 2012).

> sufficiently coincide, the system has learned enough to make confident predictions for the remaining documents. Typically, the senior lawyer (or team) needs to review only a few thousand documents to train the computer.
>
> Some systems produce a simple yes/no as to relevance, while others give a relevance score (say, on a 0 to 100 basis) that counsel can use to prioritize review. For example, a score above 50 may produce 97 [percent] of the relevant documents, but constitutes only 20 [percent] of the entire document set.
>
> Counsel may decide, after sampling and quality control tests, that documents with a score of below 15 are so highly likely to be irrelevant that no further human review is necessary. Counsel can also decide the cost-benefit of manual review of the documents with scores of 15–50.[18]

For the record, the plaintiff, who claimed that only a manual search could assure full compliance, objected to the process. Therefore, the verdict with respect to Peck's predictive prognosticative process is still pending.

V. CONCLUSION

Computers are taking over. What else is there to say? Today's crude and perhaps questionable processes will continue to improve: tomorrow they naturally will become an unimpeachable standard. Will some form of predictive coding, therefore, be judicially accepted at some point in the future? Definitely, at least until *its* inevitable replacement comes along.

[18] Da Silva Moore v. Publicis Groupe, 2012 WL 607412 (S.D.N.Y., Feb. 24 2012) (Peck, Mag. J.) (citing Andrew Peck, Andrew Peck, *Search, Forward*, LAW TECH. NEWS, Oct. 2011, at 25, 29).