Vetting and Validation of AI-Enabled Tools for Electronic Discovery, Chapter 13 from Litigating Artificial Intelligence (May 2020)

Maura R. Grossman and Gordon V. Cormack

13 Vetting and Validation of AI-Enabled Tools for Electronic Discovery

Maura R Grossman and Gordon V Cormack

I. Preface

The use of technology-assisted review (TAR) in electronic discovery (e-discovery) is a well-developed application of artificial intelligence (AI) in law. It has been in use for over a decade. Unfortunately, along with its growing use has come well-developed "received wisdom" that is at best oversimplified and at worst misleading. That has necessitated the need for a more technical treatment, which this chapter supplies. We advise the reader that this chapter is written for a specialized audience that is well versed in e-discovery. For others, the major take-aways from the chapter are the two model validation protocols supplied in the appendixes, which stand alone and can be adapted for most matters where TAR is employed. The chapter itself provides an explanation and a justification for the proposed validation protocols and counters a number of prevalent but misleading notions associated with TAR and the statistics used to evaluate TAR efforts.

II. Introduction

The process of document discovery in litigation has been almost entirely supplanted by e-discovery. In e-discovery, responsive electronically stored information (ESI)[1] must be identified and, unless withheld for privilege, produced to the requesting party. Although diverse sources and formats of ESI—such as email, text messages, social media, and scanned images subject to optical character recognition (OCR) to make them searchable—have replaced filing cabinets and banker's boxes filled with hard-copy documents, the term "document" has survived to describe any unit of ESI subject to production that is both relevant to the claims and defences of the matter at hand and can be found and produced without disproportionate burden and cost.

Along with the emergence of ESI, there has been an explosion in the number of potentially relevant "documents," along with the cost and burden of examining them to determine which are responsive and, among those that are responsive, which may be withheld for privilege. It is typically not feasible to review every email for a single

1 The term "electronically stored information" was first introduced in the 2006 amendments to the US *Federal Rules of Civil Procedure* as an umbrella term for computerized data, regardless of their format. See Maura R Grossman & Gordon V Cormack, "The Grossman-Cormack Glossary of Technology-Assisted Review" (2013) 7:1 Fed Cts L Rev 1 at 15 [TAR Glossary] (ESI is "[u]sed in Federal Rule of Civil Procedure 34(a)(1)(A) to refer to discoverable information 'stored in any medium from which the information can be obtained either directly or indirectly or, if necessary, after translation by the responding party into a reasonably usable form.' Although Rule 34(a)(1)(A) references 'Documents or Electronically Stored Information,' individual units of review and production are commonly referred to as Documents, regardless of medium."). The term "relevant" refers to ESI that is pertinent to a particular inquiry (e.g., an information request or a claim or defence in litigation); the term "responsive" refers to ESI that meets the criteria set forth in a request for production or subpoena: *ibid* at 28. Although the notion of relevance is more general than that of responsiveness, within the context of this article and in e-discovery practice more generally, the terms are often used interchangeably.

individual let alone email for the key custodians of an organization, along with other sources of ESI, such as their word-processing files, text messages, and social media. Unlike filing cabinets and banker's boxes, email folders and other electronic documents are seldom well enough organized that just a few folders that are highly likely to contain the documents of interest can be identified.

Imagine a hypothetical "ideal" situation in which cost and timeliness are non-issues and in which you have an army of competent reviewers available to examine every potentially responsive document, sorting those that are responsive from those that are not. AI methods—collectively dubbed TAR—can effectively emulate this ideal even when there are thousands or millions of potentially responsive documents to be labelled. In 2011, we published the results of an experiment showing that two different TAR methods were at least as effective as exhaustive manual review, with a tiny fraction of the effort.[2] The two TAR methods we found to be more effective than exhaustive manual review were a *rule-based* method and a *supervised machine learning* method.[3]

Citing our work either directly or by reference, courts in the United States, Ireland, the United Kingdom, and Australia have approved the use of TAR for e-discovery.[4] Its use is now generally accepted, in principle, in these and other jurisdictions, including Canada. Controversy remains, however, as to precisely what constitutes a reasonable TAR process and, after following such a process, whether an acceptable result has been achieved.

In a perfect world, a recognized body would set standards for the application of TAR technology in e-discovery and would certify particular tools and protocols for their adherence to these standards. The practitioner could then enjoy a reasonable probability of success by properly applying a certified tool and could confirm success by applying a certified validation protocol. At the time of writing (January 2021), however, no organization had taken up the mantle.[5] Therefore, for now—and for the

2 Maura R Grossman & Gordon V Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review" (2011) 17 Rich JL & Tech 1 [JOLT study].

3 *Ibid* at 29-34. A rule-based TAR method relies on a set of rules created by one or more experts to emulate the human decision-making process for the purposes of classifying documents. A supervised machine learning TAR method is one in which an algorithm learns to distinguish between relevant and non-relevant documents using a training set and labelled human judgments. See TAR Glossary, *supra* note 1 at 28, 31. Supervised machine learning methods are discussed more fully in Section IV of this chapter.

4 See e.g. *Da Silva Moore v Publicis Groupe*, 287 FRD 182 (SDNY 2012); *Irish Bank Resolution Corp & Ors v Quinn & Ors*, [2015] IEHC 175 (HC); *Pyrrho Investments Ltd v MWB Property Ltd*, [2016] EWHC 256 (Ch); *McConnell Dowell Constructors (Aust) Pty Ltd v Santam Ltd & Ors (No 1)*, [2016] VSC 734.

5 The nascent ISO/IEC 27050 standard prescribes standards for the conduct of e-discovery but is currently silent on how to measure the effectiveness and reliability of e-discovery technology, including TAR. See "ISO/IEC 27050:2016+—Information Technology—Security Techniques—Electronic Discovery (Parts 1-3 Published, Part 4 DRAFT)" (last visited 24 January 2021), online: *IsecT* <https://www.iso27001security.com/html/27050.html>.

foreseeable future—practitioners must take it upon themselves to identify reasonable tools, procedures, and validation protocols based on the best available information and augmented by their own vetting and evaluation efforts.

This chapter summarizes the available evidence concerning how AI-based TAR (and some non-TAR) e-discovery tools and processes work. It surveys which tools and processes have (and have not) been shown to be effective and how to measure their effectiveness. Furthermore, it offers practical guidance and model protocols covering both how to select and use a TAR tool and how to evaluate its efficacy—before, during, and after a particular e-discovery review effort.

III. TAR and Non-TAR Methods and Their Effectiveness

Exhaustive manual review typically entails too much time, effort, and cost to be a viable strategy for all but the smallest matters. Prior to the advent of TAR, keyword culling was traditionally used to reduce the size of the document collection; it functions by excluding from review all documents that do not match a particular set of search terms. The excluded documents (the "null set") are presumed to be non-responsive, whereas the matching documents are subject to manual review and labelling for responsiveness, privilege, confidentiality, and sometimes specific issues related to the claims or defences in the litigation.

Before we examine the different TAR and non-TAR methods, we must first define the measures and related terms we will use to assess their efficacy. The *end-to-end recall* of a review effort (a measure of completeness) is the fraction of responsive documents in the collection that are correctly labelled as responsive. For the keyword-culling method described above, a document can be labelled responsive only when it is matched by the search terms *and* correctly labelled by the reviewer. The *end-to-end precision* of a review effort (a measure of accuracy) is the fraction of documents labelled responsive that are, in fact (i.e., truly), responsive.[6] It is possible to compute only an *estimate* of recall and precision by comparing the results to a gold standard. A "gold standard" is the best available determinant of relevance or non-relevance of all (or a sample) of a document collection, used as a benchmark to evaluate the effectiveness of a search and review effort.[7] It is also referred to as "ground truth."

It is important to distinguish the end-to-end recall of the review effort from the recall of the keyword culling alone (i.e., search-term recall). Search-term recall is the fraction of responsive documents that are matched by the keywords regardless of whether they are correctly labelled by the reviewer. Similarly, search-term precision is the fraction of documents matched by the keywords that are actually responsive, irrespective of human review. In general, search-term recall will be higher than end-to-end recall because

6 See TAR Glossary, *supra* note 1 at 27 and 25 for definitions of "recall" and "precision," respectively.

7 *Ibid* at 18 for the definition of "gold standard."

manual review is imperfect, and reviewers will incorrectly exclude some relevant documents.[8] Search-term precision will generally be lower than end-to-end precision because manual reviewers will correctly exclude some non-relevant documents during the review process. In assessing the effectiveness of any review effort, it is important to consider end-to-end recall and precision so as to capture all sources of error—both technological and human. In unusual circumstances, the results of a keyword search may be produced without subsequent manual review. Only in such circumstances would the search-term recall and the end-to-end recall be one and the same.

▶ Beware of keyword culling.

In a seminal 1985 study, Blair and Maron[9] had lawyers and paralegals employ iterative searches and then review the resulting documents until they believed they had achieved at least 75 percent recall, which they considered to be adequate. Fifty-one different information needs (i.e., topics or requests for production [RFPs]) relating to a Bay Area Rapid Transit (BART) train accident were employed in the study.[10] These search and retrieval efforts achieved, on average, only 20 percent recall, at the same time achieving 79 percent precision.[11] From this result—which has not been contradicted in 35 years—we can conclude that it is difficult to achieve both high recall and high precision using search terms; that it is difficult to know whether high recall has been

8 See JOLT study, *supra* note 2 at 37, table 7 (showing recall estimates for human reviewers ranging from 25.2 percent to 79.9 percent, with an average of 59.3 percent); Herbert L Roitblat, Anne Kershaw & Patrick Oot, "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review" (2010) 61 J American Society for Information Science & Technology 70 at 76, table 2 (showing recall estimates for human reviewers ranging from 45.8 percent to 53.9 percent, with an average of 50.3 percent). See also Scott M Cohen, Elizabeth T Timkovich & John J Rosenthal, "The Tested Effectiveness of Equivio>Relevance in Technology Assisted Review," Winston & Strawn e-discovery & Information Management White Paper (February 2014) at 6, figure 5, 7, online (pdf): *Equivio* <http://www.equivio.com/files/files/White%20Paper%20-%20Winston%20and%20Strawn.pdf> (showing a recall estimate of 52.4 percent for human reviewers). An example of the confusion between search-term recall and end-to-end recall can be seen in *In re Biomet M2a Magnum Hip Implant Prods Liab Litig (MDL 2391)*, Case No 3:12-MD-2391, 2013 WL 1729682, at *2 (ND Ind 18 April 2013), where the Court concluded that "a comparatively modest number of documents" had been missed by the TAR process when the search-term recall estimate was approximately 60 percent and the end-to-end recall estimate (including TAR and manual review) would necessarily have been substantially lower than that. See William Webber, "What Is the Maximum Recall in re Biomet?" (24 April 2013), online (blog): *Evaluating E-Discovery* <http://blog.codalism.com/index.php/what-is-the-maximum-recall-in-re-biomet/>.

9 David C Blair & ME Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," (1985) 28:3 Communications of ACM 289.

10 *Ibid* at 291.

11 *Ibid* at 293.

achieved; that searchers overestimate the recall of their searches;[12] and that high precision is often mistakenly construed to be an indicator of high recall.

Keyword culling yields dramatically inferior results to exhaustive manual review, but prior to the introduction of TAR, keyword culling was (and sometimes still is) considered a necessary evil. Even so, it was commonly recognized that some keywords were better than others and that some sort of validation of search terms was necessary.[13] This observation, in part, gave impetus to the National Institute of Standards and Technology's (NIST) Text REtrieval Conference (TREC) Legal Track, which, from 2006 through 2011, measured the effectiveness of various search methodologies by comparing their results to a gold standard created for that purpose.[14] The TREC Legal Track showed results for keyword culling consistent with those of Blair and

12 See also Maureen Dostert & Diane Kelly, "Users' Stopping Behaviors and Estimates of Recall" in Mark Sanderson et al, eds, *SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: Association for Computing Machinery [ACM], 2009) 820 (showing that most subjects in an interactive information retrieval experiment reported that they had found 51 to 60 percent of the relevant documents when, on average, recall was only 7 percent).

13 See e.g. *William A Gross Constr Assocs, Inc v American Mfrs Mutual Ins Co*, 256 FRD 134 at 134 (SDNY 2009): "This Opinion should serve as a wake-up call to the Bar in this District about the need for careful thought, quality control, testing … in designing search terms or 'keywords' to be used to produce emails or other electronically stored information ('ESI')." The Court concluded (at 136): "[W]here counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians … and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of 'false positives'"; *Victor Stanley, Inc v Creative Pipe, Inc*, 250 FRD 251 at 260 (D Md 2008): "While keyword searches have long been recognized as appropriate and helpful for ESI search and retrieval, there are well-known limitations and risks associated with them, and proper selection and implementation obviously involves technical, if not scientific knowledge." The Court continued (at 262): "Selection of the appropriate search and information retrieval technique requires careful advance planning by persons qualified to design effective search methodology. The implementation of the methodology selected should be tested for quality assurance"; *In re Seroquel Prods Liab Litig*, 244 FRD 650 at 662 (MD Fla 2007): "[W]hile key word searching is a recognized method to winnow relevant documents from large repositories, … [c]ommon sense dictates that sampling and other quality assurance techniques must be employed to meet requirements of completeness."

14 "The goal of the Legal Track at the Text Retrieval Conference (TREC) is 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. In the USA, this problem is referred to as 'e-discovery.' Like all TREC tracks, the Legal Track seeks to foster the development of a research community by providing a venue for shared development of evaluation resources ('test collections') and baseline results to which future results can be compared": (last modified 10 May 2012), online: *TREC Legal Track* <https://trec-legal.umiacs.umd.edu/>.

Maron[15] and dramatically superior results (compared to manual review) for two specific TAR methods.[16] To be clear, it also showed inferior results for the majority of TAR methods evaluated at TREC and inconsistent results for others.[17] Thus, the use of TAR is likely to yield an acceptable result only when TAR tools and methods that have been shown to be effective are properly applied and validated.

▶ Not all "TAR" is effective.

Myriad service providers were quick to repurpose (or simply relabel) their existing search and analytics offerings as "TAR," pointing to TREC and our 2011 JOLT study as (specious) evidence of the efficacy of these offerings. Along with these offerings came a number of misconceptions and prescriptions—many of which still persist today—about how TAR should be conducted and its results should be validated.[18] Many practitioners were disappointed by the results of using these "TAR" offerings. The adoption of TAR was hindered either because inferior results were obtained or because the prescribed rituals for the use of these tools were too complicated and burdensome.

15 See Douglas W Oard et al, "Overview of the TREC 2008 Legal Track" in Ellen M Vorhees & Lori P Buckland, eds, *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings*, NIST Special Publication: SP 500-277 (August 2009) at 4, 8-9, online (pdf): *National Institute of Standards and Technology* <https://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf> (showing an average recall for search terms of 24 percent over 45 topics); S Tomlinson et al, "Overview of the TREC 2007 Legal Track" in Ellen M Vorhees & Lori P Buckland, eds, *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, NIST Special Publication: SP 500-274 (August 2008) at 4, 10, online (pdf): *National Institute of Standards and Technology* <https://trec.nist.gov/pubs/trec16/papers/LEGAL.OVERVIEW16.pdf> (showing an average recall for search terms of 22 percent over 50 topics); Jason R Baron, David D Lewis & Douglas W Oard, "TREC—2006 Legal Track Overview" in Ellen M Vorhees & Lori P Buckland, eds, *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, NIST Special Publication: SP 500-272 (October 2007) at 4-5, 11, online (pdf): *National Institute of Standards and Technology* <https://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf> (showing an average recall for search terms of 57 percent over 43 topics).

16 See Bruce Hedin et al, "Overview of the TREC 2009 Legal Track" in Ellen M Vorhees & Lori P Buckland, eds, *The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings*, NIST Special Publication: SP 500-278 (August 2010) at 17, table 6, online (pdf): *National Institute of Standards and Technology* <https://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf>.

17 *Ibid.*

18 See e.g. Karl Schieneman & Thomas C Gricks III, "The Implications of Rule 26(g) on the Use of Technology-Assisted Review" (2013) 7 Fed Cts L Rev 239. But see Maura R Grossman & Gordon V Cormack, "Comments on 'The Implications of Rule 26(g) on the Use of Technology-Assisted Review" (2013) 7 Fed Cts L Rev 285 (criticizing Schieneman and Gricks) [Grossman & Cormack, Comments on Rule 26(g)].

More recently, the machine learning method we have found to be most effective—which has been trademarked as "Continuous Active Learning" or "CAL"[19] and is known generically in the e-discovery industry as "TAR 2.0,"[20]—has come to be recognized as the industry standard and the method of choice in most cases. A number of service providers have claimed to implement similar methods to CAL, typically branding their products either as "TAR 2.0" or with two of the three words "continuous," "active," or "learning." At the time of writing, no independent evaluation of these commercial offerings had been conducted.

The rule-based method that was found to be effective at the TREC Legal Track is a trade secret of H5.[21] What we do know is that it involves extensive interplay among subject-matter experts, linguists, and statisticians.[22] Some service providers have claimed to implement rule-based methods that are just as good as, if not better than, H5's version, but, again, as of the time of writing, there has been no independent verification of such claims. Overall, rule-based TAR methods have attracted less attention in e-discovery than supervised machine learning TAR methods, perhaps because they are more time and resource intensive and can be applied only by a highly trained team.

▶ Is keyword culling obviated by TAR?

TAR plays a similar role to keyword search in that it identifies a subset of documents that are likely to be responsive. Typically, all documents identified as potentially responsive by TAR are labelled by reviewers, in which case, TAR fulfills the same role as keyword culling: to identify for review only those documents that are likely to be responsive. As with keyword culling, it is important to consider the end-to-end recall and precision that take into account the accuracy of the TAR as well as the subsequent human review of the documents identified by TAR as potentially responsive.

19 "Continuous Active Learning" and "CAL" are registered trademarks of Maura R Grossman and Gordon V Cormack. See "Continuous Active Learning—Trademark Details" (last visited 29 January 2021), online: *Justia* <https://trademarks.justia.com/866/34/continuous-active-86634255.html>, and "CAL—Trademark Details" (last visited 29 January 2021), online: *Justia* <https://trademarks.justia.com/866/34/cal-86634265.html>, respectively.

20 See John Tredennick et al, *TAR for Smart People, Expanded and Updated*, 3d ed (Catalyst, 2018) at 21–22.

21 H5 is a San Francisco–based e-discovery service provider. See online: *H5* <https://www.h5.com/about-us/>.

22 See JOLT study, *supra* note 2 at 29-31, n 145. See also Christopher Hogan et al, "H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement" in Ellen M Vorhees & Lori P Buckland, eds, *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings*, NIST Special Publication: SP 500-277 (August 2009), online (pdf): *National Institute of Standards and Technology* <https://trec.nist.gov/pubs/trec17/papers/h5.legal.rev.pdf>.

When using TAR, the primary reason for keyword culling—to save manual review effort by excluding likely non-responsive documents from review—no longer exists because TAR eliminates the need to review the vast majority of non-responsive documents. Despite this fact, many e-discovery practitioners believe it may still be necessary to use keyword culling to reduce the size of the collection prior to TAR, either because search is the only practical way to retrieve documents for review from the cloud or elsewhere or because of capacity limits or the costs of processing and ingesting documents into the TAR platform. In any event, the number of documents that can be reviewed by TAR is orders of magnitude greater than what can be manually reviewed.

Setting aside cost considerations, the better practice is *not* to use keyword culling before TAR. When search terms, TAR, and manual review are used together, the estimated end-to-end recall is equal to the product of the search-term recall, the TAR recall, and the reviewer recall. If it is possible to achieve 70 percent search-term recall, 70 percent TAR recall, and 70 percent reviewer recall, the estimated end-to-end recall will be $70\% \times 70\% \times 70\% = 34\%$. On the other hand, if it were possible to achieve 85 percent search-term recall, 85 percent TAR recall, and 85 percent reviewer recall, then the estimated end-to-end recall would be $85\% \times 85\% \times 85\% = 61\%$. Under the same assumptions, but without the keyword culling, estimated end-to-end recall for the TAR plus manual review would be $85\% \times 85\% = 72\%$.

If keyword culling must be used before TAR, it is desirable to use very broad keywords to achieve higher recall at the expense of more documents than would otherwise be practical for a manual review.

Regardless of whether keyword culling or TAR is used, end-to-end recall should be estimated with respect to the gold standard of an exhaustive manual review of the entire collection had that been performed. To do so, we recommend using a *stratified sample*,[23] with separate strata for (1) documents labelled responsive by reviewers, (2) documents labelled non-responsive by reviewers, (3) documents excluded from review by TAR, and (4) documents excluded from review by keywords. In this way, the gold standard provides an independent assessment to gauge effectiveness regardless of the method of search and review. The steps for taking and using stratified samples are described in detail in the appendixes to this chapter.

IV. Machine Learning Methods

The term "machine learning" encompasses two vastly different mechanisms with different purposes: (1) *supervised* machine learning, in which an algorithm is "taught" by showing it examples of human decision-making, from which it learns to emulate

23 A "stratified sample" is a sample formed as the aggregate of separate random samples for distinct subpopulations. For example, a stratified sample of Canadians might be formed by combining samples from each of the 13 Canadian provinces and territories in proportion to their respective populations. See Adam Hayes, "Stratified Random Sampling" (3 March 2020), online: *Investopedia* <https://www.investopedia.com/terms/stratified_random_sampling.asp>.

that decision-making on new examples, and (2) *unsupervised* machine learning, in which the algorithm, without being taught, learns to organize or to draw inferences from the data without examples.

▶ Logistic regression and support vector machines are state-of-the-art TAR algorithms.

Although certain supervised machine learning algorithms and protocols have been shown to be effective for TAR,[24] certain unsupervised machine learning methods have been referred to as (or included with) "TAR" offerings, but their effectiveness has not been established.

Within the context of TAR, a supervised machine learning algorithm is "taught" using documents that have been labelled as responsive or non-responsive by a human reviewer. The algorithm learns to emulate the coding decisions of that reviewer either by predicting categorically whether the reviewer would label an as-yet-unlabelled document as responsive or by predicting the likelihood that the human reviewer would label the document responsive and then ranking the documents in the collection from most to least likely to be responsive. When using supervised machine learning for TAR, the most important choices are (1) which particular machine learning algorithm to use and (2) the protocol for selecting the documents from which the algorithm will learn to make distinctions between responsive and non-responsive documents. Common supervised machine learning algorithms for TAR include *support vector machines* (SVMs), *logistic regression*, and *nearest neighbour* (NN or 1-NN).[25]

Before applying a supervised machine learning method, it is necessary first to decompose each document into *features*,[26] which may be words, phrases, or word fragments, or *latent features* (e.g., concepts), generated using an unsupervised machine

24 See JOLT study, *supra* note 2; Gordon V Cormack & Maura R Grossman, "Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery" in *SIGIR '14: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York: ACM, 2014) 153 [SIGIR 2014]. See also Maura R Grossman, Gordon V Cormack & Adam Roegiest, "TREC 2016 Total Recall Track Overview" in Ellen Voorhees & Angela Ellis, eds, *The Twenty-Fifth Text REtrieval Conference (TREC 2016) Proceedings*, NIST Special Publication: SP 500-321 (December 2016), online (pdf): *National Institute of Standards and Technology* <https://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf>; Adam Roegiest et al, "TREC 2015 Total Recall Overview" in Ellen Voorhees & Angela Ellis, eds, *The Twenty-Fourth Text REtrieval Conference (TREC 2015) Proceedings*, NIST Special Publication: SP 500-319 (December 2015), online (pdf): *National Institute of Standards and Technology* <https://trec.nist.gov/pubs/trec24/papers/Overview-TR.pdf>; Hedin et al, *supra* note 16 at 17, table 6.

25 See TAR Glossary, *supra* note 1 at 31, 22, and 24, respectively, for definitions of "support vector machine," "logistic regression," and "nearest neighbour."

26 *Ibid* at 17 for a definition of "features."

learning method. Based on a (manually) labelled set of training documents, the supervised machine learning algorithm constructs a *model*[27] to predict whether a document would likely be labelled as responsive based on its features. Within the context of TAR, the two algorithms that have been demonstrated to be most effective are SVM and logistic regression, using words, phrases, or word fragments as features. A common but much less effective method applies NN to latent features (i.e., concepts). A number of other methods, including "random forests," "k-nearest neighbour (k-NN)," "neural networks," and "deep learning," have shown promise in the scientific literature but have yet to be applied successfully to TAR.

SVMs employ a geometric interpretation of the document features, representing each document as a point in space—a space with many more than three dimensions—where each dimension represents the presence or absence of a feature. Think, for example, of what makes spam spam. It might be the use of caps or punctuation, the sender's email address, or an incorrect time zone (e.g., EDT instead of EST). Each of these (and other things) might be features that distinguish spam from non-spam. In the geometric representation of the document, each possible feature represents a dimension in space.

Responsive documents will tend to be nearer to one another in this space than to non-responsive documents, and vice versa. SVMs attempt to find a separating hyperplane in this space that best separates responsive from non-responsive documents. The distance from the hyperplane indicates the likelihood that a document is responsive: if the document is far from the hyperplane on the side with mostly responsive documents, it is likely to be responsive; if the document is far on the opposite side, it is very unlikely to be responsive; if the document is near the hyperplane, its responsiveness is uncertain.

Logistic regression employs a probabilistic interpretation of the document features, determining how well each feature distinguishes responsive from non-responsive documents and combining these estimates for all of the features in a document to determine the probability that it is responsive or not. Logistic regression has been used for nearly a century as a data analysis tool (e.g., to estimate the risk of cancer given a combination of environmental factors). For text categorization in general and TAR in particular, the scientific literature shows its effectiveness to be comparable or superior to the newer SVMs.

- Nearest neighbour is an inferior TAR algorithm.
- In e-discovery vernacular, "latent semantic analysis" and "latent semantic indexing" are synonyms for "nearest neighbour."

27 *Ibid* at 23, 30 for a definition of "model," which is also referred to as a "statistical model."

Like SVM, NN considers each document to be a point in hyperspace. The label for each document is deemed to be the same as the label of the nearest document in the training set (i.e., its "nearest neighbour"). Because responsive documents tend to be more similar to each other than to non-responsive documents, and vice versa, chances are that the correct label for a document is the same as its nearest neighbour. NN does not work well for the high-dimensional spaces arising from the use of words, phrases, or word fragments as features[28] and therefore is generally applied to latent features derived from an unsupervised machine learning algorithm, such as latent semantic analysis (LSA).[29] In the e-discovery literature, the method of NN with latent features is often referred to by the unsupervised latent-feature algorithm, meaning that when an article refers to LSA as a TAR method, it almost certainly means "NN with LSA features." In e-discovery, latent feature analysis is also sometimes referred to as "clustering"[30] or "concept clustering."

NN with LSA features falls short of the state of the art for TAR, and its shortcomings have engendered commonly held beliefs that do not apply to state-of-the-art TAR methods—for example, that TAR cannot be applied to short documents or spreadsheets, that not all labelled documents should be used as training examples, that the training set must be randomly selected, that the training set must be carefully hand-picked by a senior lawyer, or that one or a few mislabelled training documents can result in a cascade of TAR errors (i.e., "garbage in, garbage out").[31]

Other supervised machine learning methods, notably those involving deep learning, have received much attention of late. At the time of writing, none of these methods have been shown to be effective for TAR or to be efficient enough for the size of the document collections that typically arise in e-discovery.

28 See Rosalind B Marimont & Monte B Shapiro, "Nearest Neighbour Searches and the Curse of Dimensionality" (1979) 24 IMA J Applied Mathematics 59.

29 LSA is also referred to as latent semantic indexing (LSI). Other methods include probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA). See TAR Glossary, *supra* note 1 at 22 and 26, respectively, for definitions of LSA, LSI, and PLSA.

30 *Ibid* at 11 for a definition of "clustering."

31 The "garbage in, garbage out" (GIGO) problem is specific to NN, although in the e-discovery literature, it has been widely but inaptly offered as evidence to be wary of all TAR methods. See Adam Roegiest & Gordon V Cormack, "Impact of Review Set Selection on Human Assessment for Text Classification," in *SIGIR '16: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2016) 861.

V. TAR Protocols

- Continuous Active Learning (CAL or TAR 2.0) is state of the art in circumstances in which every document is to be reviewed before production.
- Simple Active Learning (SAL) is state of the art in circumstances in which documents are to be produced without prior review.
- Simple Passive Learning (SPL) is inferior, whether or not every document is to be reviewed before production.
- The generic term "TAR 1.0" encompasses both SAL and SPL.

Supervised machine learning protocols determine the method by which training examples are selected and labelled, how a supervised machine learning algorithm is applied, and how the results are harvested. The primary protocols in use today are Simple Passive Learning (SPL), Simple Active Learning (SAL), CAL, or a hybrid approach involving two or more of these protocols.[32] The machine learning literature in the domain of information retrieval (i.e., outside the field of e-discovery) refers to SPL as "supervised learning" and to SAL as "active learning," without qualification. Little non-TAR literature concerning CAL exists; however, in the information-retrieval literature, the term "relevance feedback" generally refers to a similar protocol used to refine keyword searches.[33]

A TAR tool may be used for one of two distinct purposes:

1. where the TAR tool identifies substantially all potentially responsive documents, which are then manually reviewed and labelled by human reviewers. Documents not identified by the TAR tool are deemed non-responsive following a validation process.
2. where the TAR tool classifies every document as responsive or not based on a limited training set and then labels the remaining documents according to the TAR tool's ranking or classification process, without any further human review.

32 See, generally, Grossman & Cormack, Comments on Rule 26(g), *supra* note 18; SIGIR 2014, *supra* note 24. See also Maura R Grossman & Gordon V Cormack, "Continuous Active Learning for TAR" (April/May 2016) Practical LJ 32 at 36, table.

33 See e.g. Christopher D Manning, Prabhakar Raghavan & Hinrich Schütze, *Introduction to Information Retrieval*, online ed (Cambridge: Cambridge University Press, 2009) at 178, online: *Stanford Natural Language Processing Group* <https://nlp.stanford.edu/IR-book/html/htmledition/relevance-feedback-and-pseudo-relevance-feedback-1.html>; see also TAR Glossary, *supra* note 1 at 28 (defining "relevance feedback" as a process in which the documents with the highest likelihood of relevance are labelled by a reviewer and added to the training set).

In the traditional context of a review for production in litigation, the first purpose is the most common: the producing party typically wants to review every document to be produced in order to label those that are confidential, to avoid producing non-responsive or privileged documents, and to gain an understanding of their case. In some instances (e.g., "second requests,"[34] or matters where parties are cost sensitive or the collection is likely to contain few privileged or confidential documents), the burden of reviewing every document may outweigh the risk of producing non-relevant or privileged/confidential documents, occasioning the second approach.

For the first purpose, CAL is the most effective protocol of which we are aware. CAL is quite simple. At the outset, one or more responsive documents (typically real but possibly synthetic or contrived) are used to train the algorithm, which scores the remaining documents as to their likelihood of responsiveness. Some of the top-scoring documents are reviewed and coded as responsive or not and are used to train the algorithm. This process continues until very few of the top-scoring remaining documents are responsive, and those few responsive documents that remain are neither novel nor important.[35] Once this threshold is met, post hoc validation will very likely show high recall.[36] In the unlikely event that validation reveals that a substantial number of novel or important responsive documents have been missed, those documents can be used to train the algorithm, and the CAL process can be resumed and continued for as long as necessary.

For the second purpose, SAL is the most effective commercially available protocol of which we are aware. SAL starts out like CAL, using one or more responsive documents or synthetic documents to initiate the process, although some users prefer to begin with randomly selected documents, which we believe is an inefficient training strategy, especially when responsive documents are sparse in the collection. Then the machine learning algorithm chooses from all the yet-to-be-reviewed documents the ones from which it will likely learn the most. Two strategies that have been shown to be effective for SAL are *uncertainty sampling*[37] and *Scalable Continuous Active Learning*

34 Second requests are document requests propounded by the Antitrust Division of the US Department of Justice or the US Federal Trade Commission to investigate mergers and acquisitions that may have anti-competitive effects. Typically, second requests seek to gather information about the markets, sales, facilities, assets, and structures of the businesses that are parties to the intended transaction.

35 For a description of the CAL process, see Grossman & Cormack, "Continuous Active Learning for TAR," *supra* note 32. See also Grossman & Cormack, Comments on Rule 26(g), *supra* note 18 at 289-91; SIGIR 2014, *supra* note 24 at 154.

36 See Gordon V Cormack & Maura R Grossman, "Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review" in *SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2015) 763.

37 With "uncertainty sampling," the machine learning algorithm selects the documents the relevance about which it is least certain for review by a human reviewer. See TAR Glossary, *supra* note 1 at 33-34.

(S-CAL).[38] The selected documents are reviewed and labelled and used to train the machine learning algorithm. This process continues until a fixed number of documents (typically several thousand) have been reviewed and labelled or until "stabilization" occurs, indicating that further training examples would not substantially improve the model constructed by the algorithm. At this point, the model is used to label all the as-yet-unlabelled documents. Post hoc validation is conducted, as for any review effort. If validation shows that the estimated recall is too low, it can typically be increased (at the expense of decreased precision) by altering the threshold score above which the documents are presumptively labelled responsive.

Although we do not recommend SAL when all documents to be produced will be subject to human review, it has been and continues to be used. Its use follows the same course as if the documents were not subject to review in that the TAR tool is used to label every document responsive or not. But the documents labelled responsive by the TAR tool (said to constitute the "review set"), rather than being produced, are reviewed and labelled by a human reviewer. Only those documents deemed responsive by the human reviewer are produced. When used in this manner, estimated recall and precision must be calculated based on the final labels coded by the human reviewer, not the presumptive labels assigned by the TAR tool. Otherwise, they do not reflect the recall and precision of the end-to-end review effort but only that of a single phase (i.e., the TAR component, not including the manual review that follows).[39]

In the academic literature, the term "supervised learning" typically refers to the process we denote as SPL to distinguish it from SAL and CAL. In SPL, training examples are chosen not by the machine learning algorithm but solely by random selection and/or a human reviewer. SPL is not an efficient training strategy, especially when responsive documents are sparse in the collection, and has yielded results inferior to SAL and CAL in studies of TAR-protocol effectiveness.[40] The use of SPL has engendered much controversy as to how to choose the training examples, with requesting parties often insisting on micromanaging the process because of the belief that this method can easily be gamed. There is no empirical evidence that justifies ad hoc efforts to choose the "best" training documents or to use heroic efforts to ensure agreement on the labels used to teach the TAR algorithm. Nonetheless, if parties insist on using SPL (or any other TAR method for which evidence of efficacy is lacking), time is likely

38 For a description of S-CAL, see Gordon V Cormack & Maura R Grossman, "Scalability of Continuous Active Learning for Reliable High-Recall Text Classification" in *CIKM '16: Proceedings of the 25th ACM International Conference on Information and Knowledge Management* (New York: ACM, 2016) 1039. S-CAL is a patented method of Gordon V Cormack and Maura R Grossman. See "Systems and Methods for a Scalable Continuous Active Learning Approach to Information Classification," Patent Application Pub No US 2016/0371262 A1 (22 December 2016 [pub date]; allowed 2020), online (pdf): *FPO* <http://www.freepatentsonline.com/20160371262.pdf>.

39 See Grossman & Cormack, Comments on Rule 26(g), *supra* note 18 at 300-1.

40 See e.g. SIGIR 2014, *supra* note 24.

better spent ensuring valid computation of recall and precision estimates rather than in arguing about the mechanics of the process.

VI. Measures of Effectiveness

▶ Relevance is in the eye of the beholder.

One of the most difficult challenges of estimating review effectiveness is that *relevance* is a subjective concept.[41] Regardless of how clearly the criteria for responsiveness are set forth in an RFP or in a subpoena, equally knowledgeable and well-intentioned expert reviewers—whether lawyers, state archivists, or intelligence analysts—will disagree on the relevance of a surprisingly large number of documents.[42] Suppose that two separate, equally competent teams—Team A and Team B—were to independently review the same set of potentially responsive documents for responsiveness to an RFP or a subpoena, labelling each document as responsive or not. Scientific studies indicate that no more than about 70 percent of documents labelled "responsive" by Team A will also be labelled "responsive" by Team B, and vice versa.[43] This observation is a consequence of the fact that relevance is subjective, not that either labelling is necessarily defective.

▶ Effectiveness is quantified by "recall" and "precision" measured with respect to an independent gold standard.

Notwithstanding the subjectivity of relevance, the results of any competent review effort (e.g., Review A) can be used as a "gold standard" or "ground truth" against

41 See Tefko Saracevic, "The Notion of Relevance in Information Science: Everybody Knows What Relevance Is. But What Is It Really?" (September 2016) 8 Synthesis Lectures on Information Concepts, Retrieval, & Services; Peter Bailey et al, "Relevance Assessment: Are Judges Exchangeable and Does It Matter?" in Sung-Hyon Myaeng et al, eds, *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2008) 667.

42 See e.g. Gordon V Cormack & Maura R Grossman, "Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me" in *SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2017) 5 (the expert in this article being a senior state archivist); JOLT study, *supra* note 2 at 10-14; Roitblat, Kershaw & Oot, *supra* note 8 (the experts in this article being lawyers); Ellen M Voorhees, "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness" (2000) 36 Information Processing & Management 697 (the experts in this article being intelligence analysts). See also Adam Roegiest & Anne McNulty, "Variations in Assessor Agreement in Due Diligence" in *CHIIR '19: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (New York: ACM, 2019) 243.

43 *Ibid*.

which to evaluate another review effort (e.g., Review B), provided that *the review to be evaluated is independent of the review from which the gold standard is derived*. The two most common measures of effectiveness in the science of information retrieval are *recall* and *precision*, which are predicated on the convenient fiction that the gold standard is infallible and therefore denotes "true relevance." In theory, recall is the proportion of "truly relevant" documents that are identified by a review effort out of all of the "truly relevant" documents in the collection (i.e., a measure of the *completeness* of the review effort), whereas precision is the proportion of documents identified by the review effort that are "truly relevant" (i.e., a measure of the *accuracy* of the review effort). In practice, recall is estimated as the proportion of documents labelled "relevant" in the gold standard that are also labelled "relevant" in the review effort being evaluated, whereas precision is estimated as the proportion of documents labelled "relevant" in the review effort being evaluated that are also labelled "relevant" in the gold standard. Assuming that both the gold standard and the review effort being evaluated are conducted by "ideal" but independent, exhaustive manual reviews, we would not expect either estimate to exceed the typical positive agreement[44] between two independent reviews, or about 70 percent.[45]

- Exact recall and precision scores can never be known because they rely on an infallible gold standard that cannot be achieved.
- Recall and precision estimates fall short of the exact values.
- Recall and precision estimates are incomparable unless calculated according to precisely the same independent gold standard.
- Seventy percent recall (or any other recall score) is meaningless in isolation and inappropriate as an absolute standard of review effectiveness.

The scientific observation—that the recall and precision of a high-quality review effort, estimated with respect to an independent gold standard, are typically in the neighbourhood of 70 percent—has been widely misconstrued in the legal community to imply that 70 percent or even 75 percent recall (without specifying the recall of what and how it is measured) is necessary and sufficient to establish the adequacy of an e-discovery review effort.[46] This mantra is espoused by requesting and producing parties alike, either to set an unreasonably high and potentially unachievable recall target or to provide cover for inadequate or incomplete results.

44 "Positive agreement" is the proportion of documents labelled "relevant" by one review and labelled "relevant" by a second review. See TAR Glossary, *supra* note 1 at 25.

45 *Supra* note 41.

46 See Grossman & Cormack, Comments on Rule 26(g), *supra* note 18, at 303.

As a preliminary matter, "70 percent recall" can never be established; at best, a party can demonstrate an *estimated recall* of 70 percent. If the estimate is conducted with respect to an ideal, independent gold standard, 70 percent estimated recall is indeed consistent with—but not proof of—an effective review effort.

A recall estimate is meaningless unless the gold standard is independent of the review effort for which it is calculated; an estimate of about 70 percent is merely the best that is likely to be achieved by *a single qualified reviewer, unassisted by quality control or by technology*.

Taking the following as an absurd example, an exact copy of the gold standard would achieve an estimated recall of 100 percent and an estimated precision of 100 percent when evaluated with respect to itself as the gold standard. But it would achieve no more than 70 percent estimated recall and 70 percent estimated precision when correctly evaluated with respect to an independent gold standard. Similarly, a copy of the gold standard as to which 30 percent of the documents labelled responsive were relabelled non-responsive would achieve an estimated 70 percent recall compared to the original gold standard—and might be offered as evidence of an adequate production—but recall correctly estimated *with respect to an independent gold standard* would show the recall estimate of the revised gold standard to be at most $70\% \times 70\% = 49\%$.

> ▶ Recall and precision can be used to compare review efforts only if calculated according to a common gold standard that is independent of the review efforts being compared.

We can compare the effectiveness of Reviews B and C by calculating their estimated recall and precision according to an independent gold standard, Review A. If Review B yields substantially higher estimated recall and precision than Review C, according to gold-standard Review A, we can say with some confidence that Review B is more effective than Review C. But what if Review B yields higher estimated recall, whereas Review C yields higher estimated precision? Which is more effective depends on the relative importance of finding *all* responsive documents (i.e., achieving high recall) versus finding *only* responsive documents (i.e., achieving high precision). In e-discovery, recall is generally the more important of the two measures, considering the aspirational goal of most productions: to find "all" responsive documents. However, if precision is very low (e.g., much less than 50 percent), the responsive documents that are produced can be overwhelmed and therefore obscured by the non-responsive documents.[47]

47 For example, this was precisely what happened in *In re Domestic Airline Travel Antitrust Litig*, MDL No 2656, Misc No 15-1404 (CKK), 2018 WL 4441507 (DDC 13 September 2018), where one of the defendants' productions had a recall of 97 percent but a precision of only 17 percent, meaning that only about 600,000 of the 3.5 million documents produced were responsive, or that five out of every six documents produced were non-responsive.

- There is a trade-off between recall and precision, in both the conduct and the evaluation of a review effort.

The human reviewer (or the TAR tool) can be instructed to construe responsiveness broadly, thus increasing recall at the expense of precision, or to construe responsiveness narrowly, increasing precision at the expense of recall. For the purposes of e-discovery, Review B is generally considered superior to Review C if it achieves substantially higher recall while still achieving similar precision, or if it achieves similar recall and substantially higher precision.

- F_1 quantifies the extent to which a review effort achieves high recall and high precision.
- F_1 offers no insight into the recall–precision trade-off and should not be used as the sole measure to evaluate review effectiveness.

F_1 is the harmonic mean of recall and precision.[48] Unlike the traditional arithmetic mean, F_1 emphasizes the lesser of the two, so a high F_1 score is achieved only when recall and precision are both high and approaches zero as either recall or precision approaches zero. When recall and precision are equal, so is F_1. The same F_1 score can be achieved if recall is slightly higher and precision is slightly lower, or vice versa, but if either is substantially lower than the other, F_1 will fall. These properties justify the observation that an exhaustive manual review is unlikely to yield F_1 greater than 70 percent, which would be achieved if recall and precision were both 70 percent, or thereabouts, as measured with respect to an independent gold standard.

VII. The Use and Misuse of Statistics

Statistics play an essential role in many human endeavours, including e-discovery, but are easily miscalculated, misapplied, or misinterpreted. A common use of statistics is to estimate a quantity that is too large to be counted, such as the number of people in a country who hold a particular opinion. Another common use is to estimate a quantity that can only be known in the future, such as how many people will contract a particular disease.

48 See TAR Glossary, *supra* note 1 at 16 for a definition of "F_1." A harmonic mean, "unlike the more common arithmetic mean (i.e., average), falls closer to the lower of the two quantities. As a summary measure, a Harmonic Mean may be preferable to an arithmetic mean because a high Harmonic Mean depends on both high Recall and high Precision, whereas a high arithmetic mean can be achieved with high Recall at the expense of low Precision, or high Precision at the expense of low Recall" (*ibid* at 18).

▶ Numbers or percentages?

Statistical estimates may be expressed as numbers—for example, "approximately 760,000 Canadians believe the world is flat," or "[o]n average, 617 Canadians will be diagnosed with cancer every day."[49] Or they may be expressed as percentages—for example, approximately 2 percent of Canadians believe the world is flat,[50] or about 0.0017 percent of Canadians will be diagnosed with cancer tomorrow. To convert from one to the other, it is necessary to have an estimate of the size of the population of Canada, which, as of 2020, was about 38 million. Thus, the estimated number of Canadian flat-earthers is $2\% \times 38$ million $\approx 760{,}000$, whereas the estimated daily percentage of Canadians that will be diagnosed with cancer is $617 \div 38$ million $\approx 0.0017\%$. More generally, statistical estimates—whether numbers or percentages—tell an incomplete story absent an estimate of the size of the population from which they are derived.

▶ Counting found versus missed responsive documents.

In e-discovery, the principal application of statistics is to estimate *how many responsive documents have been identified for production and how many have been missed*. If the number missed is large compared to the number found, the review is incomplete; if the number missed is small compared to the number found, the review may be complete depending on the novelty and importance of the missed responsive documents.

From these estimates, we can derive an estimate of recall, which is the estimated number of responsive documents that are identified for production, as a percentage of the estimated overall number of responsive documents in the collection.[51] A recall of substantially greater than 50 percent indicates that the number found is greater than the number missed. For example, an estimate of 75 percent recall indicates that three times as many responsive documents were found than were missed.

Precision is the estimated number of truly responsive documents identified for production as a percentage of the total number of documents labelled responsive by the review effort. Precision greater than 50 percent indicates that the production

49 See "Cancer Statistics at a Glance" (last visited 31 January 2021), online: *Canadian Cancer Society* <https://www.cancer.ca/en/cancer-information/cancer-101/cancer-statistics-at-a-glance>.

50 See "What Canadians Believe: From Science and Spirituality to Conspiracies and the Supernatural" (2019), online (pdf): *Pollara* <https://www.pollara.com/wp-content/uploads/2017/12/Pollara-Beliefs2019-RptF2.pdf>.

51 The estimated overall number of responsive documents in the collection is simply the estimated number of responsive documents found plus the estimated number of responsive documents missed.

contains more responsive than non-responsive documents. For example, an estimate of 75 percent precision indicates that three-quarters of the documents identified as responsive were actually responsive.

Remember that valid estimates of recall and precision rely on an independent gold standard of relevance—that is, they must be based on comparison to a *ground truth*. It is generally impractical to conduct an independent review of the entire collection of documents in order to form such a gold standard. Statistical sampling methods allow us to estimate, based on an independent review of many fewer documents, what the result would be were all the documents to be independently reviewed to establish the ground truth.

▶ Statistical sampling and blind review yield an independent gold standard.

To estimate the number of truly responsive documents identified for production, it is necessary to draw a random sample of the production set. To estimate the number of responsive documents that were missed, it is necessary to draw one or more random samples of all documents excluded from the production set for any reason—whether by keyword culling, by TAR, or by manual review. The documents labelled non-responsive by a search and review process are typically referred to as the "null set" (or sets). These different samples must be scrambled together and labelled by an independent reviewer, *who can have no knowledge of whether a document has been included or excluded from the production or why*. This is typically referred to as a "blind review."[52] The results of this independent review can be used to properly estimate end-to-end recall, precision, and other measures.

▶ Sample size, margin of error, and confidence level.

Estimates may differ from the hypothetical but unknowable true values for a variety of reasons. The first reason, uncertainty in the conception of relevance, is inherent in any recall or precision calculation, whether statistical or not. This error is mitigated by the use of an independent gold standard labelled diligently by one or more knowledgeable reviewers. It could be—but in practice rarely is—mitigated by labelling each document according to the majority vote of a panel of three (or more) reviewers.

52 In a blind review, certain information is withheld from the reviewer to reduce bias and error—for example, the population from which a sample was drawn or the prior responsiveness label given to a document. See Robert MacCoun & Saul Perlmutter, "Blind Analysis: Hide Results to Seek the Truth" (2015) 526 Nature 187. Blind review is discussed further in para F of Appendix A to this chapter.

Compounding the error arising from uncertainty in relevance is the *error arising from statistical estimation*. The terms of art used in information retrieval to describe this error are "bias"—the extent to which the estimate systematically misses the mark—and "random error"—the extent to which the estimate differs from the true value by chance. Bias is mitigated by random sampling and blind independent review. Random error is mitigated by increasing sample size: the larger the sample, the smaller the random error.

The terms of art used to describe the magnitude of random error are "margin of error" and "confidence interval," which are always qualified by a "confidence level."[53] For example, we might say that "760,000 Canadians believe the earth is flat, with a margin of error of ± 76,000 and a confidence level of 95 percent." Equivalently, we might say that "the number of Canadians who believe the earth is flat falls in the confidence interval of 684,000 to 836,000 people, with a 95 percent confidence level." Either statement may be interpreted as strong evidence (but never proof) that the true value is somewhere between 684,000 and 836,000 and likely nearer to 760,000 than the extremes of the interval. The confidence level of 95 percent has been widely, but somewhat arbitrarily, adopted as a standard for reporting results in the scientific literature.

Extreme caution must be used when expressing margin of error as a percentage and combining it with a numerical value. If we say "760,000 Canadians believe the earth is flat, with a margin of error of ± 10 percent," it is ambiguous whether we mean ± 10 percent of the 760,000 Canadians or ± 10 percent of all Canadians. When we translate these to numbers, we see that the first interpretation means 760,000 ± 76,000 people, whereas the second means 760,000 ± 3.8 million people (which would be a rather useless estimate). Yet e-discovery practitioners often commit the fallacy of plugging the numbers into a statistical calculator that uses the latter interpretation and representing the result as if it were the former.

- Margin of error is commonly miscalculated and misinterpreted.
- For estimates of a small percentage of the population, margin of error cannot be predetermined.
- For estimates of a very small percentage of the population, no reasonable estimate can be achieved by sampling.

The fallacy of misinterpreting margins of error has engendered a pervasive school of specious "learned wisdom" in e-discovery circles, which can be identified by its vernacular. The typical e-discovery practitioner is likely to encounter articles, briefs, and

53 See TAR Glossary, *supra* note 1 at 22-23, 12, and 12, respectively, for definitions of "margin of error," "confidence interval," and "confidence level."

protocols calling for a "sample with a 2 percent margin of error at a 95 percent confidence level," or simply a "± 2 percent sample," or a "statistically significant sample." Depending on what statistical calculator is used, the size of such a sample turns out to be 2,395, 2,399, or 2,405. This sample size ensures a margin of error of no more than ± 2 percent of the Canadian population; in our flat-earth example, that would be ± 760,000 Canadians (i.e., $2\% \times 38$ million ≈ 760,000 Canadians). Thus, the margin of error would be 100 percent of the number of Canadian flat-earthers (760,000 ± 760,000), not 2 percent of the flat-earthers as implied. To achieve a margin of error of at most ± 2 percent of Canadian flat-earthers (15,200 people or 0.04 percent of all Canadians) would, according to the same assumptions and the same statistical calculator, require an absurdly large sample size of 518,000.

A more sophisticated analysis using a binomial calculator[54] yields a somewhat smaller margin of error, which can be calculated only after the sample has been reviewed. Suppose that in a sample of 2,400 Canadians, 48 were to believe the earth was flat. From the fact that 2 percent of the sample believed that the earth was flat, we can estimate that about 2 percent of all Canadians (i.e., 760,000 Canadians) also believe the earth is flat. Using a binomial calculator, we derive a confidence interval of 1.48 percent to 2.65 percent (or between 562,000 and 1,007,000) of Canadians who believe the earth is flat, with 95 percent confidence.

Suppose that 2,400 Canadians were surveyed on a particular day to determine whether they had received a cancer diagnosis that day. In all likelihood, none of them would respond affirmatively. From this result, we can conclude that being diagnosed with cancer on a given day is unlikely, but not much else. According to the binomial calculator, this result yields a confidence interval of 0 percent to 0.15 percent (or between zero and 57,000) of Canadians who receive such a diagnosis on a given day. The sample provides strong evidence that fewer than 57,000 Canadians received a positive diagnosis on that particular day but conveys no other information about the number of diagnoses on that particular day or on any future day. Clearly, a method other than simple sampling of a single day's data would be necessary to yield a meaningful estimate in this example. The reader is urged to eschew the received wisdom in e-discovery that any quantity can be estimated "± 2 percent" with a sample size of approximately 2,400. When necessary, hire a statistical expert to assist in calculating margins of error, confidence intervals, confidence levels, and appropriate sample sizes. Do not do this yourself.

54 A binomial calculator is a statistical method used to calculate confidence intervals based on the binomial distribution (as opposed to the Gaussian or normal distribution). When there are few relevant documents in the sample, the binomial estimation is more accurate than the Gaussian or normal estimation; when there are many relevant and non-relevant documents in the sample, then the binomial and the Gaussian (or normal) estimates are nearly identical. See TAR Glossary, *supra* note 1 at 9-10. An example of an online binomial calculator can be found at "Exact Binomial and Poisson Confidence Intervals" (last modified 25 May 2009), online: *StatPages* <https://statpages.info/confint.html>.

▶ Statistics for vetting and for validation.

The purpose of vetting a TAR tool is to estimate, in advance, how well it will work for a given review effort by examining how well it has worked for one or more past review efforts. For these purposes, it is feasible to spend considerably more time and effort in constructing a gold standard and in sampling to achieve a small margin of error compared to evaluating a single review effort. It is also feasible to repeat the process for many past reviews so as to aggregate the results. For example, the TREC 2016 Total Recall Track simulated 34 reviews for different information needs using the email collection from Jeb Bush's administration when he was the governor of Florida.[55] Six professional reviewers were employed for six weeks to create substantially complete gold-standard labels for each of the information needs.

Using such test collections, it is possible to assess the relative effectiveness and reliability of the methods employed by the participants at TREC 2016, or to assess the effectiveness and reliability of new methods, relative to those employed at TREC 2016. Such an assessment yields a reasonable prediction as to how effective the methods would likely be when applied to a new matter. At the time of writing, very few commercially available TAR tools have been vetted in this manner.

In contrast, the purpose of validation is to estimate the effectiveness of a particular review effort. The sample size that is used is limited by time and cost; a sample size of about 2,400 is often chosen for the specious reasons described above. Although a sampling strategy involving the review of 2,400 documents cannot possibly yield an estimate of recall or precision with a margin of error of ± 2 percent, it can yield a sufficiently precise estimate that can, along with other evidence, help confirm the effectiveness of a particular review effort.

The validation strategy that we propose is to use blind review of a combined stratified sample to compute separate estimates of (1) the number of truly responsive documents that are correctly coded relevant by the end-to-end review process, (2) the number of responsive documents incorrectly coded non-relevant by human reviewers, (3) the number of responsive documents incorrectly excluded from review by the TAR process, and, if employed, (4) the number of responsive documents incorrectly excluded by keyword culling or by any other culling method that may have been used. When combined, these estimates can provide an end-to-end estimate of the recall and precision of the entire search and review effort rather than just search-term recall or TAR recall, which can be misleading because they both make it appear as if the recall of the production set is higher than it actually is.[56]

55 Grossman, Cormack & Roegiest, "TREC 2016 Total Recall Track Overview," *supra* note 24.

56 A step-by-step guide for taking a blind stratified sample to compute end-to-end recall and precision for an individual review effort is provided in Appendix A to this chapter.

The number of truly responsive documents identified in the production set will be a fairly large proportion—perhaps 70 percent—of all documents identified for production.[57] A sample of 400 documents is sufficient to estimate this proportion ± 5 percent with 95 percent confidence, which is sufficient for these purposes. The number of responsive documents incorrectly excluded by reviewers will be a smaller but still substantial proportion—perhaps 10 percent—of the total number of documents excluded by the reviewers. If it is 10 percent, a sample of 400 documents is sufficient to estimate this proportion ± 3 percent with 95 percent confidence, which is also sufficient for these purposes. The number of responsive documents incorrectly excluded by an effective TAR process is likely to constitute a very small proportion of the total number excluded—that is, the vast majority of excluded documents will be non-responsive. A sample of 1,600 may well reveal no responsive documents. In this event, we can conclude that, with 95 percent confidence, no more than 0.23 percent of the documents excluded by TAR are responsive. If search terms are used before TAR, an additional sample of 1,600 documents excluded by the keywords should be drawn. As with the TAR sample, this will, with high confidence, reveal whether an inordinate number of responsive documents have been missed by the keywords, but it can only provide a coarse upper limit of that number.

Assuming that no keyword culling was performed before TAR, we employ a combined sample size of 2,400 documents for the purposes of validation because that number is the one most often used in e-discovery today to evaluate review efforts (albeit for the wrong reasons, as described above) and because it should be sufficient and proportionate for most matters. The total size of the three samples—400 + 400 + 1,600 = 2,400 (or 2,400 + 1,600 = 4,000 if keywords were used)—can be adjusted on a case-by-case basis to balance the tension between validation review *effort* and the *precision of the estimates* achieved through the validation process. The estimates may be combined to yield estimates of recall and precision, which summarize the effectiveness of the end-to-end review effort. However, the separate estimates, and the samples themselves, can also offer deeper insights into potential shortcomings of various aspects of the review effort and opportunities for mitigation. A model protocol for determining the effectiveness of an individual review effort using this method is provided in Appendix A.

57 It will not be all (i.e., 100 percent) of the documents identified for production because the reviewer and the gold standard will sometimes disagree and because often document families are produced in whole, such that non-documents associated with responsive documents (i.e., attachments or "family members") are often produced automatically (i.e., along with the associated responsive parent email, or vice versa, document) even though they are not responsive in their own right. Such documents should not be counted when estimating recall and precision because they were not identified as responsive through the search and review effort that is being measured; their production is fortuitous.

▶ Statistics such as accuracy, elusion, and F_1 do not tell the whole story.

Some statistics may convey the illusion of—but no actual insight into—review effectiveness. "Accuracy" is simply the overall proportion of documents that are correctly labelled as either responsive or non-responsive by the review effort (when combined together).[58] Suppose that 1 percent of the documents in a collection are responsive and that a vacuous review labels every one of them non-responsive. The accuracy of this review is 99 percent, although it does not identify a single responsive document. Accuracy is an uninformative measure of review effectiveness.

"Elusion" is the percentage of excluded documents that are responsive (i.e., the percentage of responsive documents found in the null set).[59] A small elusion number (e.g., 1 percent) is commonly touted as evidence of the effectiveness of a review effort. But a vacuous review that identifies no responsive documents could still achieve an elusion of 1 percent even though no responsive documents were identified. By itself, elusion conveys no useful information. Combined with the size of the excluded set, elusion can be used to estimate the number (as opposed to the percentage) of excluded responsive documents. This number can be compared to the number of produced responsive documents, but only if both are sampled and measured with respect to the same gold standard, derived from a blind review, as proposed above and in the appendixes to this chapter.

F_1 combines recall and precision. Arguably, a high F_1 score is evidence that both precision and recall are high, but a low F_1 score gives no indication as to which of the two is low or how to remedy the problem.

VIII. Establishing the Effectiveness of TAR Tools and Review Efforts

As noted in the previous sections, estimates of effectiveness scores such as recall, precision, and F_1 can be used to compare the efficacy of review efforts given an independent gold standard. For the purpose of establishing the reasonableness of a particular review effort, it would be desirable to establish beforehand that the effort would be likely to achieve—and afterward that it did achieve—recall and precision comparable to or surpassing that of accepted practice or, better still, the hypothetical "ideal" of exhaustive manual review.

58 See TAR Glossary, *supra* note 1 at 8.

59 *Ibid* at 15.

▶ The effectiveness of TAR tools and protocols should be established in advance; the effectiveness of review efforts should be validated after the fact. Neither is a substitute for the other.

The same approach may be used to establish the effectiveness of a surgical procedure. First and foremost, it is necessary to use tools, procedures, and surgeons whose outcomes have been validated for similar patients with similar conditions; second, it is necessary to verify for every case that post-operative tests yield results consistent with a successful surgery. Unfortunately, few e-discovery service providers have subjected their tools, procedures, or experts to anything resembling a "clinical trial" or to any sort of rigorous evaluation. In the future, a consortium of service providers, regulators, and/or practitioners should conduct such product testing. In the meantime, practitioners are on their own to vet the TAR tools and methods they choose to use and to establish the effectiveness of their individual review efforts after the fact. A model protocol for vetting the effectiveness of a TAR tool or protocol in advance is provided in Appendix B.

▶ TREC offers standard test collections that can be used as a benchmark to assess the effectiveness and reliability of TAR methods.

TREC is an annual conference that evaluates information retrieval methods using test collections consisting of a common set of documents, information needs (i.e., topics or RFPs), and independent gold-standard relevance assessments.[60] Each year, academic, government, and industry participants test various approaches to search and review on the test collections, and the results are reported in the TREC proceedings.

60 TREC was initiated in 1992. According to NIST's overview of TREC,

> [i]ts purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, the TREC workshop series has the following goals:
>
> - to encourage research in information retrieval based on large test collections;
> - to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
> - to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
> - to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.
>
> For more information about TREC, see online: *NIST* <https://trec.nist.gov/>.

The test collections are also made publicly available so that they can be used to conduct experiments on methods and tools that were not represented at TREC.

Of particular interest here are the Legal Track, an evaluation campaign that ran at TREC from 2006 through 2011, and the Total Recall Track, which ran at TREC in 2015 and 2016. From 2006 through 2008, the Legal Track employed a collection of seven million documents from the tobacco litigation[61]; from 2009 through 2011, it employed a collection of about 700,000 documents captured from Enron at the time of its collapse.[62] Each year, information needs were specified using a mock complaint and mock RFPs. A gold standard was created for a statistical sample of the document collection and was used to estimate recall, precision, F_1, and other effectiveness measures for the TREC participants' retrieved results. The results are published in the TREC proceedings;[63] the collections and tools, which can be used to evaluate the results of future experiments, are available from NIST, subject to a usage agreement.[64]

▶ The Jeb Bush collection from the TREC 2015 and 2016 Total Recall Tracks offers 290,000 documents, 44 topics, and independent gold standards for research into the effectiveness of TAR tools and methods.

The TREC Total Recall Track employed six different collections: (1) a set of 290,000 emails from Jeb Bush's tenure as governor of Florida; (2) a set of 500,000 postings from two "black hat" web forums; (3) a set of 900,000 news articles from the Columbia region of North America; (4) a set of 90,000 clinical records from a hospital intensive care unit; (5) a set of 400,000 emails from Tim Kaine's tenure as governor of Virginia; and (6) a set of 300,000 emails from Rod Blagojevich's tenure as governor of Illinois.[65] Of these collections, the first three are publicly available for research purposes. Of most interest to practitioners is the Jeb Bush collection, for which 44 information needs

61 David D Lewis et al, "Building a Test Collection for Complex Document Information Processing" in Susan Dumais et al, eds, *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2006) 665.

62 Gordon V Cormack, "TAR Evaluation Toolkit Release 1.0.0" (2013), online: *University of Waterloo* <https://cormack.uwaterloo.ca/tar-toolkit/>.

63 "TREC Proceedings" (last modified 21 February 2020), online: *NIST*, <https://trec.nist.gov/proceedings/proceedings.html>. See also TREC Legal Track, *supra* note 14.

64 See "Data" (last modified 20 March 2020), online: *NIST* <https://trec.nist.gov/data.html> (for most TREC data sets, including the Legal Track), and "TREC Total Recall Information-Retrieval Text Research Collections" (last visited 31 January 2021), online: *Programming Languages Group* <https://plg.uwaterloo.ca/~gvcormac/total-recall/TR-group.html> (for the TREC Total Recall data sets).

65 Grossman, Cormack & Roegiest, "TREC 2016 Total Recall Track Overview," *supra* note 24. See also Roegiest et al, *supra* note 24.

(i.e., topics or RFPs) and full gold-standard relevance assessments are available, as well as three alternate, mutually independent sampled sets of gold-standard assessments for the 34 TREC 2016 Total Recall topics.

We have used these collections—and similar collections from other evaluation efforts—in our research. Our paper comparing the effectiveness of rule-based and supervised machine learning methods for TAR to exhaustive manual review was based on a retrospective analysis of the TREC 2019 Legal Track results.[66] Our paper comparing the effectiveness of different supervised machine learning protocols for TAR reported a simulation study using the TREC 2019 Legal Track collection.[67] Our paper measuring the reliability of TAR methods used the TREC Total Recall collections, as well as several others.[68] Practitioners may use the same collections to evaluate commercial TAR tools and protocols.

▶ The results of a prior review effort can be used as the basis for a test collection to evaluate new tools and methods.

Many practitioners also have access to the results of a prior review that they believe was competently performed. The documents, RFPs, and relevance assessments for that review can be used to compare the effectiveness of two new methods—for example, TAR versus a traditional manual review or two TAR methods. For this purpose, the prior relevance assessments are used as the gold standard; in other words, the results of the prior review (presumably representing accepted practice) may be used to assess the effectiveness of the two new methods.

In the alternative, it is possible to compare the effectiveness of the prior review to that of a new review by creating a new independent gold standard. In this case, it is necessary to create the gold standard by conducting a blind review of a stratified sample of documents so as to populate a confusion matrix,[69] from which recall, precision, F_1, and other measures can be estimated.

66 See JOLT study, *supra* note 2.

67 See SIGIR 2014, *supra* note 24.

68 See Gordon V Cormack & Maura R Grossman, "Engineering Quality and Reliability in Technology-Assisted Review," in *SIGIR '16: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2016) 75.

69 A confusion matrix (also sometimes referred to as a "contingency table") is a two-by-two table listing the values for the number of true positives (i.e., documents labelled responsive that are truly responsive), true negatives (i.e., documents labelled non-responsive that are truly non-responsive), false positives (i.e., documents labelled responsive that are truly non-responsive), and false negatives (i.e., documents labelled non-responsive that are truly responsive). Virtually all of the standard evaluation measures in information retrieval are algebraic combinations of the four values in the confusion matrix. For an example of a confusion matrix and the formulas for the information-retrieval measures that may be computed from it, see TAR Glossary, *supra* note 1 at 12.

To compare review methods, A and B, the stratified sample must contain a number of documents selected at random from each of these four strata, as detailed in Appendix B, paragraph B:

1. documents labelled responsive by Review A and responsive by Review B;
2. documents labelled responsive by Review A and non-responsive by Review B;
3. documents labelled non-responsive by Review A and responsive by Review B; and
4. documents labelled non-responsive by Review A and non-responsive by Review B.

The documents in the samples are shuffled together and labelled by an expert reviewer, who is given no information regarding the stratum from which each document was derived. This blinding process is essential to the validity of any estimate derived from the gold standard. It is well established that, regardless of the motivation of the reviewer, they would be influenced by knowing from which stratum each document was derived or even by knowing some factor associated with the stratum—for example, the order in which the documents are presented for review.[70]

From the samples and the gold-standard assessments, we can estimate the number of documents labelled responsive by Review A that are also labelled responsive according to the gold standard (i.e., precision) and the number of documents that are labelled responsive according to the gold standard that are also labelled responsive according to Review A (i.e., recall). Precision and recall for Review B can be estimated in a similar manner and compared to those for Review A. A similar strategy can be employed to compare more than two methods, using strata to capture all combinations of relevance (and non-relevance) labels among the systems. Such an approach was used for the TREC 2008 through 2011 Legal Tracks.[71]

---

70 Adam Roegiest & Gordon V Cormack, "Impact of Review-Set Selection on Human Assessment for Text Classification" in *SIGIR '16: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2016) 861; Falk Scholer et al, "The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment" in *SIGIR '13: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2013) 623; Mark D Smucker & Chandra P Jethani, "Human Performance and Retrieval Precision Revisited" in Hsin-Hsi Chen et al, eds, *SIGIR '10: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM, 2010) 595; William Webber et al, "Assessor Error in Stratified Evaluation" in Xiangji Jimmy Huan et al, eds, *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (New York: ACM 2010) 539; Mu-hsuan Huang & Hui-yu Wang, "The Influence of Document Presentation Order and Number of Documents Judged on Users' Judgments of Relevance" (2004) 55 J American Society Information Science & Technology 970; Michael Eisenberg & Carol Barry, "Order Effects: A Study of the Possible Influence of Presentation Order on User Judgments of Document Relevance" (1988) 39 J American Society Information Science & Technology 293.

71 *Supra* note 14.

- One information need for one collection is sufficient as a proof of concept; multiple information needs and multiple collections are necessary to show reliability.

Each TREC collection entails a massive volunteer effort to construct a gold standard for each of many information needs. Given the results of a method over all of these information needs, it is possible to estimate reliability—the percentage of the time that the method succeeds, for some definition of success—in other words, a measure of consistency. We might say, for example, that a method is reliable if its estimated recall equals or exceeds that of standard practice for at least 95 percent of all information needs. Further evidence of reliability can be demonstrated by showing that a method is reliable using different test collections.

Although the literature provides evidence of the effectiveness and reliability of some search and review methods,[72] it is largely silent on the effectiveness and reliability of most commercially available TAR tools. We strongly advocate standardized testing to address these questions, but at the present time and for the foreseeable future, the practitioner's best recourse is to consider the available empirical evidence regarding the effectiveness of various methods, to consider critically service providers' claims as to which tools and methods they employ, to conduct proofs of concepts where feasible, to monitor the progress of their ongoing review efforts, and to evaluate their own results after the fact using stratified sampling and blind review, as set forth in the appendixes to this chapter.

At the very least, after-the-fact validation of a putative review effort should be based on the blind review of a stratified sample of documents, as detailed in the model validation protocol set forth in Appendix A, including both documents labelled responsive by the review effort and documents labelled non-responsive by the review effort.

From this blind review, estimates of end-to-end recall and precision—according to the gold-standard blind review—can be derived. As a rule of thumb, an estimate of 70 percent end-to-end recall and 70 percent end-to-end precision is consistent with an adequate review when estimated under these circumstances; however, cases of disagreement between the review result and the gold standard should be examined carefully to determine whether they indicate a systemic problem with the review effort. Moreover, responsive documents in the sample that were missed by the review effort should be examined for their novelty and importance to determine whether an identifiable class of responsive documents has been left behind.

A more convincing, but heretofore unused, method for after-the-fact validation would be to have the sample reviewed by two independent, equally qualified, teams—one arbitrarily chosen to serve as the gold standard and one as a control. If the estimated

72 *Supra* note 24.

recall and precision of the actual review effort equal or exceed those of the control (when judged according to the gold standard), we can reasonably conclude that the review was as effective as the control, which represents exhaustive manual review.

Finally, we note that recall and precision are not the sole determinants of effectiveness. For a typical information need, the majority of responsive documents—perhaps 70 percent or more—will be similar and perhaps only marginally important. A minority of the documents may be dissimilar to the majority but very important. A review that systematically excludes these different-but-important documents is deficient, regardless of what estimate of recall and precision it achieves. As a preliminary matter, documents for which the review effort and the gold standard disagree should be examined for instances of excluded categories of responsive documents. Better still, auxiliary keyword searches of the collection and/or of the null set(s) should be conducted to try to find as many diverse examples of responsive documents as possible to determine if any have mistakenly been excluded by the review effort. This approach is akin to audit sampling in accounting. It is not the same as random sampling, but it can be very effective when the prevalence or richness of the collection is low (i.e., when there are few responsive documents to be found) and the nature of difficult-to-find responsive documents can be anticipated.

IX. A TAR Checklist and Parting Words

What follows is a checklist of our key take-aways from this chapter:

- *Not all search and review efforts are created equal.* In choosing a TAR tool and protocol, the user should rely first and foremost on independent, controlled studies as an indicator of which algorithms and protocols are most effective and reliable. To the extent that such independent product evaluations are unavailable, users should test the efficacy and efficiency of TAR tools and protocols themselves, in advance of their use, so they are not first learning about their (in)effectiveness after they complete a review effort. The TREC test collections, the user's own prior competent review efforts, and the procedures described in Appendix B provide users with all they need to vet search and review methods. Users should not settle for scripted searches demonstrated by the service provider's marketing team to make purchasing decisions. Most service providers use some version of the Enron data set for the purposes of their product demonstrations. At the very least, in vetting search tools, users can consult the TREC 2009 to 2011 Legal Track Overview papers[73] for possible information needs (i.e., topics or RFPs) to use for testing the system beyond the canned searches used in provider product demonstrations.
- *Search-term recall, TAR recall, and reviewer recall are not the same things as end-to-end recall.* Measuring only a single component of a review effort and presenting

73 See NIST, "TREC Proceedings," *supra* note 63.

it as an estimate of the quality of the production set misrepresent the recall estimate so that it appears better than it actually is. Each culling or review step introduces its own error and reduces the overall (i.e., end-to-end) recall estimate, the combined effect of which can be substantial. Therefore, unless it is infeasible or too costly to do so, do not use keywords to cull before TAR. Your end-to-end recall will be considerably higher that way.

- *Know your TAR algorithms.* If at all possible, use an SVM or logistic regression. Do not confuse concept-clustering methods such as LSI and LSA with TAR.
- *Know your TAR protocols.* Use a CAL protocol when you intend to review all potentially responsive documents before production; use a SAL or S-CAL protocol when you do not intend to do so. If you are using a SAL protocol, make sure that the two components that make up the F_1 score (i.e., recall and precision) are both acceptable before you stop training. Begin the training process with examples of known or synthetic (i.e., contrived) responsive documents (as well as some known non-responsive documents) rather than random samples, which decrease efficiency, particularly when responsive documents are sparse in the collection.
- *Manage expectations about outcome.* Relevance is subjective, and two equally qualified reviewers will label the same documents differently about 30 percent of the time. Thus, without the use of quality control measures and technology, it is challenging to obtain recall estimates above 70 percent for manual review since that is the level at which reviewers typically agree. All review efforts entail a trade-off between recall and precision; the higher of one that is achieved, the lower the other will be. F_1 summarizes the two but does not show which is lower or what can be done about that.
- *Always measure recall and precision using a blind review of stratified samples that include examples of documents labelled responsive for any reason and examples of documents excluded or labelled non-responsive for any reason.* Only when measured in this manner, as described in Appendix A and according to an independent gold standard, can an end-to-end recall on the order of 70 percent to 80 percent be said to demonstrate an adequate review effort (assuming that no novel or important documents are identified during the validation process). We provide model protocols with step-by-step guidelines for vetting a TAR tool, a TAR protocol, or other review method (see Appendix B) and for evaluating an individual review effort (see Appendix A).
- *Statistics and online statistical calculators are not toys.* Beware of "learned wisdom" regarding margins of error, confidence intervals, confidence levels, and sample sizes; it is often wrong. For example, there is no such thing as a "statistically significant training set." There is no such thing as a guarantee that you have found 95 percent of all relevant documents or that you will find all of the relevant documents 95 percent of the time.

Statistical terms such as "margin of error," "confidence interval," and "confidence level" apply only to the estimate of a particular quantity from a particular sample—for example, the number of relevant documents found or the number of relevant documents missed. The mantra that a sample of 2,395 (or any other magic number) documents will give you an estimate of these numbers ± 2 percent is misleading. The margin of error for the estimate of a *number* should be expressed as a *number*; the margin of error for the estimate of a *percentage* should be expressed as a *percentage*. The notion that a recall estimate with "a margin of error of ± 2 percent, with a confidence level of 95 percent," can be derived from a simple sample of 2,395 from the collection is simply false. Such a sample can yield nothing more than a coarse estimate of the proportion of a certain type of document (i.e., the "richness" or "prevalence" of responsive documents) in a collection, which, like the proportion of Canadian flat-earthers, is typically low. Compared to this low richness, a margin of error of ± 2 percent is enormous.

Appendix A

A Model Validation Protocol for Examining the Effectiveness of an Individual Review Effort[1]

A. The review effort should be conducted using tools and protocols whose effectiveness has been established prior to the review effort under consideration. The review process should incorporate quality control and quality assurance measures to ensure that the selected review tool and protocol are performing as expected during the course of the review effort. Once the producing party reasonably believes that it has identified for production substantially all responsive non-privileged documents, it should conduct validation according to the sampling protocol described below. This Validation Protocol should apply to the review process regardless of whether keyword culling, manual review, or TAR (or any combination of them) was used by the producing party.

B. The Document Collection (Collection) is defined as including all documents identified for review for responsiveness. This Validation Protocol assumes that the completeness or adequacy of the Collection has already been established.

C. The Collection shall be partitioned into the following Subcollections as appropriate:

   1. documents identified by the review effort as responsive to at least one request for production (RFP), including any privileged documents but not including family members of responsive documents, unless those family members are deemed to be responsive in their own right (Subcollection C(1));
   2. documents coded as non-responsive by a human reviewer, regardless of how the documents were selected for review (Subcollection C(2));
   3. documents excluded from manual review as the result of a TAR process (Subcollection C(3)). If the review process did not employ TAR, the Collection will not include Subcollection C(3); and
   4. documents excluded from manual review by keyword culling (Subcollection C(4)). If the review process did not employ keyword culling, the Collection will not include Subcollection C(4).

1 This validation protocol was adapted from the Order Regarding Search Methodology for Electronically Stored Information in *In re Broiler Chicken Antitrust Litig*, Case No 1:16-cv-08637, 2018 WL 1146371 (ND Ill 3 January 2018) (Special Master Maura R Grossman).

D. A sample shall be drawn consisting of the following:
   1. 400 documents selected at random from Subcollection C(1) (Subsample D(1));
   2. 400 documents selected at random from Subcollection C(2) (Subsample D(2));
   3. 1,600 documents selected at random from Subcollection C(3) if TAR was used (Sample D(3)). If TAR was not used, there will be no Subsample D(3); and
   4. 1,600 documents selected at random from Subcollection C(4) if keyword culling was used (Subsample D(4)). If keyword culling was not used, there will be no Subsample D(4).

E. The sample sizes in paragraph D above are chosen so as to estimate the proportion of responsive documents in Subcollections C(1) and C(2) with a margin of error of no more than ± 5 percent and to determine if collections C(3) and C(4) contain more than 0.23 percent responsive documents, at the 95 percent confidence level.

   Should smaller margins of error be required, sample sizes may be increased as follows. Quadrupling the size of D(1) and D(2) from 400 to 1,600 will halve the size of the confidence interval, from ± 5 percent to ± 2.5 percent. Doubling the size of D(3) and D(4) from 1,600 to 3,200 will halve the proportion of responsive documents that can be detected, from 0.23 percent to 0.12 percent.

   Conversely, quartering the size of D(1) and D(2) from 400 to 100 will double the margin of error from ± 5 percent to ± 10 percent, whereas halving the size of D(3) and D(4) from 1,600 to 800 will double the proportion of responsive documents that can be detected, from 0.23 percent to 0.46 percent.

F. The sample of documents composed of the documents from Subsamples D(1), D(2), and, if present, D(3) and/or D(4) shall be combined into a single Validation Sample.

   The Validation Sample shall be reviewed and coded by a subject-matter expert (SME) who is knowledgeable about the subject matter of the litigation. This should be a lawyer who is familiar with the RFPs and the scope of relevance (i.e., the claims and defences at issue in the litigation) but need not be a senior partner. The documents shall be presented to the reviewer in random or arbitrary order (e.g., by MD5 hash value). During the course of the review of the Validation Sample, the SME shall not be provided with or have available to them any information concerning the Subcollection or Subsample from which any document was derived, the prior coding of any document, or the score afforded to any document by the TAR tool. The intent of this requirement is to ensure that the review of the Validation Sample

is blind, as necessary to form an independent gold standard for the purpose of evaluating the production.

G. Once the coding in paragraph F has been completed, the producing party shall prepare a table listing each of the documents in the Validation Sample. For each document, the table shall include:
   1. the Bates number of the document (for any documents previously produced) or a control/identification number (for any non-produced documents);
   2. the Subsample from which the document came (i.e., D(1), D(2), and, if present, D(3) or D(4));
   3. the SME's responsiveness coding for the document (i.e., responsive or non-responsive); and
   4. the SME's privilege coding for the document (i.e., privileged or not privileged). If the document is coded as non-responsive, a privilege determination need not be made for that document. All documents in the Validation Sample coded as responsive and privileged may be withheld from production (subject to logging or other stipulated requirements) but shall be counted as responsive for the purpose of this Validation Protocol.

H. The following items shall be provided to the requesting party:
   1. the table described in paragraph G;
   2. a copy of each responsive, non-privileged document in the Validation Sample that was identified for production but was not previously produced to the requesting party; and
   3. the statistics and recall estimate detailed in paragraph J below.

I. The parties shall meet and confer to determine whether they agree that the recall estimate and the quantity and nature of the responsive documents identified through the Validation Protocol indicate that the review is substantially complete. If the recall estimate and the samples indicate that Subcollections C(2) and/or C(3) and/or C(4) still contain a substantial number of non-marginal, non-duplicative responsive documents compared to Subcollection C(1), the review and quality control processes should continue, and the Validation Protocol should be repeated, as warranted. If the parties are unable to agree on whether the review is substantially complete or whether the Validation Protocol should be repeated, the parties shall seek the court's intervention.

J. Method for Estimating Recall and Precision:

   An estimate of recall and precision shall be computed to inform the decision-making process described in paragraph H above; however, the absolute numbers in their own right shall not be dispositive of

whether a review is adequate or substantially complete. Also of concern is the novelty and materiality (or, conversely, the duplicative or marginal nature) of any responsive documents identified in Subsamples D(2) and/or D(3) and/or D(4). The estimates of recall and precision shall be derived as described below. It should be noted that, when conducted by an SME pursuant to paragraph F of this Validation Protocol, a recall estimate on the order of 70 percent to 80 percent is consistent with, but not the sole indicator of, an adequate (i.e., high-quality) review. A recall estimate somewhat lower than this does not necessarily indicate that a review is inadequate, nor does a recall in this range or higher necessarily indicate that a review is adequate; the final determination will also depend on the quantity and nature of the documents that were missed by the review effort.

Recall Estimation Calculation:

The number of responsive documents found ≈ the size of Subcollection C(1) × the number of responsive documents found in Subsample D(1) ÷ 400.

The number of responsive documents coded incorrectly ≈ the size of Subcollection C(2) × the number of responsive documents found in Subsample D(2) ÷ 400.

The number of responsive documents excluded by TAR ≈ size of Subcollection C(3) × the number of responsive documents found in Subsample D(3) ÷ 1,600 (if TAR is used; otherwise, 0).

The number of responsive documents excluded by keyword culling ≈ size of Subcollection C(4) × the number of responsive documents found in Subsample D(4) ÷ 1,600 (if keyword culling is used; otherwise 0).

Estimated recall ≈ the number of responsive documents found ÷ (the number of responsive documents found + the number of responsive documents coded incorrectly + the number of responsive documents excluded by TAR + the number of responsive documents excluded by keyword culling).

Precision Estimation Calculation:

Estimated precision ≈ the number of responsive documents found ÷ the size of Subcollection C(1).

Appendix B

Model Validation Protocol for Vetting a TAR Tool or Comparing Two Different Review Methods

A. One method to establish the effectiveness of a new review tool or protocol is to compare it to the results of a prior review that is known to be of high quality. To this end, it is necessary to estimate the recall and precision of both the prior review (Review Effort A) and the new review (Review Effort B) for the same information need(s), according to the same independent gold standard.

B. The first step in the process is to identify the following four Subcollections:

   1. documents identified as responsive by both Review Effort A and Review Effort B (Subcollection C(1));
   2. documents identified as responsive by Review Effort A but not Review Effort B (Subcollection C(2));
   3. documents identified as responsive by Review Effort B but not Review Effort A (Subcollection C(3)); and
   4. documents not identified as responsive by Review Effort A or by Review Effort B (Subcollection C(4)).

C. From each of these Subcollections, a random sample of 600 documents is drawn, comprising Subsamples D(1), D(2), D(3), and D(4), respectively. These Subsamples are combined and reviewed blind by a subject-matter expert, as described above in Appendix A, paragraph F.

D. A sample size of 600 yields an estimate of each proportion with a margin of error of no more than ± 4 percent, with 95 percent confidence. Quadrupling the sample size will halve the margin of error, whereas quartering the sample size will double the margin of error.

E. ***Recall, Precision, and F_1* Estimation:**

   The number of responsive documents found by Review Effort A and Review Effort B ≈ the size of Subcollection C(1) × the number of responsive documents found in Subsample D(1) ÷ 600.

   The number of responsive documents found by Review Effort A but not by Review Effort B ≈ the size of Subcollection C(2) × the number of responsive documents found in Subsample D(2) ÷ 600.

   The number of responsive documents found by Review Effort B but not Review Effort A ≈ the size of Subcollection C(3) × the number of responsive documents found in Subsample D(3) ÷ 600.

The number of responsive documents found by neither Review Effort A nor Review Effort B ≈ the size of Subcollection C(4) × the number of responsive documents found in Subsample D(4) ÷ 600.

Estimated Recall for Review Effort A ≈ (the number of responsive documents found by Review Effort A and Review Effort B + the number of responsive documents found by Review Effort A but not Review Effort B) ÷ (the number of responsive documents found by Review Effort A and Review Effort B + the number of responsive documents found by Review Effort A but not Review Effort B + the number of responsive documents found by Review Effort B but not Review Effort A + the number of responsive documents found by neither Review Effort A nor Review Effort B).

Estimated Recall for Review Effort B ≈ (the number of responsive documents found by Review Effort A and Review Effort B + the number of responsive documents found by Review Effort B but not Review Effort A) ÷ (the number of responsive documents found by Review Effort A and Review Effort B + the number of responsive documents found by Review Effort A but not Review Effort B + the number of responsive documents found by Review Effort B but not Review Effort A + the number of responsive documents found by neither Review Effort B nor Review Effort A).

Estimated Precision for Review Effort A ≈ (the number of responsive documents found by Review Effort A and Review Effort B + the number of responsive documents found by Review Effort A but not Review Effort B) ÷ (the number of documents—whether responsive or not—in Subcollections C(1) + C(2)).

Estimated Precision for B ≈ (the number of responsive documents found by Review Effort A and Review Effort B + the number of responsive documents found by Review Effort B but not Review Effort A) ÷ (the number of documents—whether responsive or not—in Subcollections C(1) + C(3)).

Estimated F_1 (for Review Effort A or Review Effort B) ≈ 2 × (estimated recall × estimated precision of the review effort) ÷ (estimated recall + estimated precision of the review effort).

F. If Review Effort B (the new tool or protocol) has comparable or superior recall and comparable or superior precision to Review Effort A, it is reasonable to conclude that Review Effort B is at least as effective as A. F_1 combines recall and precision into a single effectiveness measure, which is commonly reported, but may obscure the fact that Review Effort B has inferior recall or inferior precision to Review Effort A.