Chapter to be published in Ed Walters (ed.), *Data Analysis in Law*
(Taylor & Francis Group, forthcoming 2018)

TECHNOLOGY-ASSISTED REVIEW IN ELECTRONIC DISCOVERY

Maura R. Grossman and Gordon V. Cormack
University of Waterloo

I. Technology-Assisted Review and the Role of Measurement

Electronic discovery ("eDiscovery") is "[t]he process of identifying, preserving, collecting, processing, searching, reviewing, and producing electronically stored information ["ESI"] that may be relevant to a civil, criminal, or regulatory matter."[1] Review for production ("review") concerns a particular phase of eDiscovery: the identification of documents from a specific collection, which meet certain criteria, typically set forth by an adversary in the form of requests for production ("RFPs"). Documents that meet the criteria are generally referred to as "responsive," and those that do not, as "non-responsive."

Technology-assisted review ("TAR") is the process of using computer software to categorize each document in a collection as responsive or not, or to prioritize the documents from most to least likely to be responsive, based on a human's review and coding of a small subset of the documents in the collection.[2] In contrast, the more familiar and widely accepted practice of manual review involves human review and coding of each and every document in the collection,[3] usually following the application of keywords or other forms of culling, such as limiting the collection to certain custodians or file types, or applying date restrictions.

Manual review is an expensive, burdensome, and error-prone process. Scientific evidence suggests that certain TAR methods offer not only reduced effort and cost, but also improved accuracy, when compared to manual review.[4] This evidence has been derived using experimental methodologies from the domain of information retrieval ("IR") research, which can be impenetrable to the average, non-technical, legal practitioner.

---

[1] Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1, 15 (2013), http://www.fclr.org/fclr/articles/html/2010/grossman.pdf (hereinafter "Glossary").
[2] Glossary, *supra* n.1, at 32.
[3] Glossary, *supra* n.1, at 22.
[4] *See*, *e.g.*, Gordon V. Cormack and Maura R. Grossman, *Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me*, in PROCEEDINGS OF THE 40TH INT'L ACM SIGIR CONFERENCE ON RESEARCH AND DEV. IN INFO. RETRIEVAL ___ (2017), http://dx.doi.org/10.1145/3077136.3080812; Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf (hereinafter "2011 JOLT Study").

Because TAR is new and unfamiliar, it has been necessary to demonstrate its efficacy to clients, their counsel, opposing parties, and the courts, often using arcane concepts and terms from IR research, such as "recall," "precision," "F_1," "margin of error," and "confidence level," to name but a few. A popular misconception has emerged that these terms and concepts are uniquely associated with TAR and must be mastered in order to use TAR, but can be avoided through the use of "tried-and-true" manual review. Recall, precision, F_1, margin of error, and confidence level, however, relate to scientific methods for measuring the efficiency and effectiveness of *any* review method, whether TAR or manual. They do not concern how to conduct a review, any more than Terry Newell's *Carbon Balance and Volumetric Measurements of Fuel Consumption*[5] concerns how to drive a fuel-efficient automobile.

Measurements of recall—or fuel consumption—can inform a user's choice of a review—or travel—method, insofar as they predict how well a particular method—or model of automobile—will meet the user's requirements. To this end, it is worthwhile to appraise the reliability and accuracy of the measurement techniques, as well as how closely the measured quantities reflect the user's actual needs and requirements.

Measurements of recall—or fuel consumption—may also be helpful during or after a review process—or road trip—to verify that the method for reaching one's destination is performing (or has performed) as expected, and, if not, to take remedial action. The measurement techniques used in this circumstance might be vastly different from those used beforehand: Our driver would likely consult the fuel gauge rather than visiting the EPA testing laboratory to conduct a carbon-balance test; the legal team would similarly use a metric commensurate with the requirements of the review task at hand.

This chapter sets forth the distinctions between different review methods, summarizes a body of scientific research that compares these different review methods, and describes various approaches to track the progress or quality of particular review efforts.

II. Review Objectives

The objective of any review effort, whether manual or TAR, is to identify, as nearly as practicable, *all* and *only* the documents that satisfy certain criteria. Following IR practice, we call documents that satisfy the criteria "relevant," and documents that do not satisfy the criteria "not relevant" or "non-relevant."

[5] U.S. Environmental Protection Agency Technical Report EPA-AA-SDSB-80-05 (Apr. 1980), https://goo.gl/F2x6Qr.

While it is obvious that the meaning of "as nearly as practicable" is open to interpretation, it may be less apparent that the meaning of "all and only [relevant] documents" is equally inscrutable.

It is well known that the notion of "relevance" is subjective, and that no two reviewers will identify exactly the same set of relevant documents within a collection. This observation applies regardless of the knowledge, skill, and diligence of the reviewers, and regardless of how precisely the relevance criteria are specified.[6] The sets of relevant documents identified by two reviewers—or by the same reviewer on two different occasions—are remarkably dissimilar. Suppose two reviewers each deem 100 documents to be "all and only the relevant documents" from a collection. The IR literature suggests that these two sets would be unlikely to have more than about 67 documents in common—documents that both reviewers deemed relevant.[7] An additional 67 documents would be deemed relevant by one reviewer and non-relevant by the other.[8] Which reviewer, we might ask, came closer to identifying "all and only the relevant documents"?

It is generally accepted that, for most practical purposes, the set of relevant documents identified by *either* reviewer is sufficiently close to the ideal of "any and all," provided that each reviewer is informed, competent, diligent, and operating in good faith. Absent supplemental evidence, there is no basis to say that one set is "closer to the ideal," or that one reviewer is "better" than the other.

If we were to consider the set of 100 documents deemed relevant by a third reviewer, we would expect to find about 67 in common with the set returned by the first reviewer, and about 67 in common with the set returned by the second reviewer.[9] Even fewer—about 45—would be in common among all three.[10]

Given two or more reviews for the same set of documents, it is possible to "triangulate," using statistical methods, to deduce the relative accuracy of each reviewer, and thus, which review is "closer to the ideal."[11]

---

[6] *See, e.g.*, Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 AM. SOC'Y FOR INFO. SCI. AND TECH. 70 (2010); Peter Bailey et al., *Relevance Assessment: Are Judges Exchangeable and Does It Matter?*, in PROCEEDINGS OF THE 31ST ANNUAL INT'L ACM SIGIR CONFERENCE ON RESEARCH AND DEV. IN INFO. RETRIEVAL 667 (2008); Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697 (2000).

[7] *See generally* Voorhees, *supra* n.6.

[8] *See generally id.*

[9] *See generally id.*

[10] *See generally id.*

[11] *See* Pavel Metrikov et al., *Aggregation of Crowdsourced Ordinal Assessments and Integration with Learning to Rank: A Latent Trait Model*, in PROCEEDINGS OF THE 24TH ACM INT'L CONFERENCE ON INFO. AND KNOWLEDGE MGMT. 1391 (2015), http://dl.acm.org/citation.cfm?doid=2806416.2806492. A more complete explanation of statistical methods such as this are beyond the scope of this chapter.

The same statistical methods can be used to estimate the accuracy of manual review and TAR alike. If the accuracy of a TAR method compares favorably with, or is indistinguishable from, that of manual review, and manual review is considered "close enough" in practice, shouldn't TAR also be considered "close enough"?

III. Review Methods

A. Exhaustive Manual Review

Exhaustive manual review involves having a human reviewer examine every document in a collection and code each document as relevant or non-relevant, and perhaps apply additional labels such as "privileged" or not, "confidential" or not, "hot" or not, and sometimes, specific issue tags. We say that the coding is *positive* when the reviewer deems the document to be relevant, and *negative* when the reviewer deems the document to be non-relevant. As noted above, positive coding is evidence—but not proof—of relevance, whereas negative coding is evidence of non-relevance.

Manual review is often accompanied by some sort of quality control process in which a portion of the documents is re-reviewed and, where indicated, re-coded by a second, more authoritative reviewer. Where the coding decisions disagree disproportionately often, action may be taken to diagnose and mitigate the cause; notwithstanding this process, the vast majority of documents in the collection are reviewed only once, and the original reviewer's coding is the sole determinant of the disposition of the document.

Post-hoc validation or acceptance testing may employ similar methods: Some of the documents in the collection may be reviewed and, where necessary, re-coded, and the result of the review is deemed acceptable if the first and second coding decisions agree sufficiently often, or if the second review does not identify a substantial number of relevant documents that were missed by the first review. When there is an insufficient level of agreement, or discrepancies are found, corrective action may be taken.

B. Culling or Narrowing the Collection

Exhaustive manual review is seldom employed in practice except in the smallest of matters. Typically, the collection of documents identified for review is first culled to include only documents belonging to certain custodians, documents created or modified within a specific time frame, or documents containing one or more search terms thought likely to appear in relevant documents. Only documents from the narrowed collection are manually reviewed, and only the documents deemed by the reviewer to be responsive and non-privileged are produced.

This culling process substantially decreases the size of the collection, and hence the burden of manual review, at the cost of excluding some difficult-to-quantify number of relevant documents from review, and hence from production. Even so, the vast majority of documents presented for review are non-relevant—often ten times as many as relevant ones.

In some very weak sense, this type of culling might be considered a form of TAR, because computer software is being employed to make coding decisions (*i.e.*, non-relevant) on the group of documents excluded from review based on some criterion, such as the lack of occurrence of any of the search terms. However, we reserve the term "TAR" to refer only to computer methods that *affirmatively categorize each document* as relevant or not, or *prioritize the entire collection* from most to least likely to be relevant. The reader should be aware, however, that many commentators and software providers assume a vacuously broad definition of "TAR," using it to refer to any of a number of processes which use a computer for narrowing, navigating, or searching a collection, or for organizing or grouping documents within a collection (*e.g.*, "email threading," "near-deduplication," or "clustering").[12] Regardless of what it is called, the culling process imposes a fundamental limit on how close to *all* relevant documents can be identified by any subsequent review effort.

All too often, quality control and validation methods are limited to the review phase, and are disregarded with respect to the documents excluded by earlier culling efforts. This omission is illogical in light of the review objective, which is to find as nearly as practicable *all* (and only) the relevant documents in the collection, not just the relevant documents in the narrowed collection, which may be substantially fewer than *all*.

C. Rule-Based TAR

A "rule base" is a set of rules—akin to a checklist, decision tree, or flow chart—that determines how to decide whether a document is relevant or not.[13] Rule bases are typically constructed by a specialized team with expertise in the subject matter(s) of

---

[12] *See*, *e.g.*, KrollDiscovery, *Defining Technology Assisted Review*, Ediscovery.com (2017), http://ediscovery.com/infobite-tar-umbrella/#.WVFvGlGQz3h ("The term Technology Assisted Review (TAR) encompasses many forms of document review technology. Under the Technology Assisted Review Umbrella are some of the following ediscovery technologies: deduplication, visual analytics, predictive coding, workflow, reporting, and searching."); Herbert L. Roitblat, *Introduction to Predictive Coding* (OrcaTec LLC 2013), at 15, http://theolp.org/Resources/Documents/Introduction to Predictive Coding - Herb Roitblat.pdf (defining TAR as "[a]ny of a number of technologies that use technology, usually computer technology, to facilitate the review of documents for discovery").

[13] Glossary, *supra* n.1, at 28 (defining Rule Base as "[a] set of rules created by an expert to emulate the human decision-making process for the purposes of classifying documents in the context of electronic discovery.").

the RFP(s), rule-base construction, linguistics, and statistics. While the construction of a rule base is labor-intensive, it can involve substantially less effort than the manual review of collections of hundreds of thousands or millions of documents, which are often encountered in major litigation or regulatory matters. Research has shown that at least one rule-based TAR method can achieve results that compare favorable to exhaustive manual review.[14]

D. Supervised Machine Learning for TAR

Supervised machine-learning methods (*i.e.*, "learners") infer how to distinguish relevant from non-relevant documents by analyzing training examples—documents that are coded (*i.e.*, labeled) as relevant or non-relevant by a human teacher. In 2014, the authors proposed the taxonomy set forth below for describing TAR methods using supervised machine learning.[15] This taxonomy has since been widely adopted in the legal industry to characterize the TAR offerings in the marketplace.[16]

In simple passive learning ("SPL") methods,[17] the *teacher* (*i.e.*, human operator) selects the documents to be used as training examples; the learner is trained using these examples, and once sufficiently trained, is used to label every document in the collection as relevant or non-relevant. Generally, the documents labeled as relevant by the learner are re-reviewed manually. This manual review represents a small fraction of the collection, and hence a small fraction of the time and cost of an exhaustive manual review.

In simple active learning ("SAL") methods,[18] after the initial training set, the *learner* selects the documents to be reviewed and coded by the teacher, and used as training examples, and continues to select examples until it is sufficiently trained. Typically, the documents the learner chooses are those about which the learner is *least certain*, and therefore from which it will learn the most. Once sufficiently trained,

---

[14] 2011 JOLT Study, *supra* n.4.
[15] Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, in PROCEEDINGS OF THE 37TH INT'L ACM SIGIR CONFERENCE ON RESEARCH AND DEV. IN INFO. RETRIEVAL 153 (2014), http://dx.doi.org/10.1145/2600428.2609601 (hereinafter "SIGIR 2014 Paper"). S*ee also* Maura R. Grossman & Gordon V. Cormack, *Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review*," 6 FED. CTS. L. REV. 285 (2014), http://www.fclr.org/fclr/articles/pdf/comments-implications-rule26g-tar-62314.pdf (hereinafter "Comments Paper"); Maura R. Grossman & Gordon V. Cormack, *Continuous Active Learning for TAR*, PRACTICAL LAW J. 32 (Apr./May 2016), at 36 (hereinafter "Practical Law Article").
[16] *See*, *e.g.*, Supreme Court of Victoria [Australia], *Practice Note SC Gen 5 – Technology in Civil Litigation* (Jan. 30, 2017), http://assets.justice.vic.gov.au/supreme/resources/fba6720a-0cca-4eae-b89a-4834982ff391/gen5useoftechnology.pdf, at 6 (approving CAL, SAL, and SPL TAR protocols).
[17] SIGIR 2014 Paper, *supra* n.15; *see also* Practical Law Article, *supra* n.15, at 36.
[18] SIGIR 2014 Paper, *supra* n.15; *see also* Practical Law Article, *supra* n.15, at 36.

the learner is then used to label every document in the collection. As with SPL, the documents labeled as relevant are generally re-reviewed manually.

In continuous active learning ("CAL")[19]—the TAR method the authors developed, use, and advocate—after the initial training set, the *learner* repeatedly selects the *next-most-likely-to-be-relevant* documents (that have not yet been considered) for review, coding, and training, and continues to do so until it can no longer find any more relevant documents. There is generally no second review, because by the time the learner stops learning, all documents deemed relevant by the learner have already been identified and manually reviewed.

In the marketplace, the term "predictive coding" has been used to describe the use of supervised machine learning for TAR, but not to distinguish between SPL, SAL, or CAL. Recently, CAL methods have been promoted under the moniker "TAR 2.0," while SPL and SAL methods have been grouped together and referred to as "TAR 1.0."[20]

E. How to Start?

Two important issues that must be addressed in any supervised machine-learning TAR method are: How to start, and when to stop?

The learner needs examples of *both* relevant and non-relevant documents in order to infer the characteristics that distinguish one from the other. Finding non-relevant examples to begin the process is easy; in most situations, the vast majority of documents in the collection are non-relevant. A random sample of documents from the collection can be expected to contain mostly or entirely non-relevant documents, which may be used as negative training examples.

Finding relevant examples can be more challenging, as they are usually less frequent—if not rare—in the collection. A random sample of documents may contain few or no relevant documents. If one document of every N in the collection is relevant, it is necessary to examine, on average, N random documents to find a single relevant one, and to examine kN random documents to find k relevant ones, as may be needed to start the learning process. As that k increases, so too will the burden of training a system that relies on many positive training examples.

---

[19] SIGIR 2014 Paper, *supra* n.15; *see also* Practical Law Article, *supra* n.15, at 36.
[20] *See, e.g.*, John Tredennick et al., *TAR for Smart People: Expanded and Updated Second Ed.* (Catalyst 2016), *available at* www.catalystsecure.com/TARforSmartPeople.

A more efficient method to find one or more positive training examples is to use a search engine—particularly one that employs *relevance ranking*[21]—to find one or more relevant documents. Given a simple query consisting of a few search terms, a search engine using relevance ranking can present to the user a set of likely relevant documents, which may be used as training examples. It is important to note that the use of search terms to identify training examples is entirely different from the use of search terms for culling or narrowing the collection. In the former case, the search terms are used to *include* documents for review, not to *exclude* them.

F. When to Stop?

For SPL and SAL, it is necessary to estimate when the learner has been sufficiently trained, a point that is often referred to as "stabilization."[22] For many SPL and SAL methods, it is further necessary to adjust the sensitivity of the learner: The higher the sensitivity, the more nearly *all* relevant documents are identified for subsequent manual review; the lower the sensitivity, the more nearly *only* relevant documents are identified. These two decisions—when stabilization has occurred, and the sensitivity of the learner—effect a multi-dimensional tradeoff among the amount of effort required for training, the amount of effort required for the subsequent manual review, and how nearly all, and how nearly only, relevant documents will be identified by the review process. These decisions are typically informed by estimates derived from the manual review of a separate random sample of documents—typically referred to as a "control set"[23]—over and above those used for training the learner.

For CAL, the decision of when to stop is deferred until evidence suggests that substantially all relevant documents have been reviewed.[24] Several methods have been proposed and evaluated for determining when a CAL review is complete.[25] Among the simplest and most effective is the following: A CAL review may be considered complete when the total number of negative coding decisions for the

---

[21] Relevance ranking is "[a] search method in which the results are ranked from the most likely to the least likely to be relevant to an information need. . . . Google Web Search is an example of relevance ranking." Glossary, *supra* n.1, at 28.

[22] *See, e.g.*, Chris Dale, *Far From the Black Box: Explaining Equivio Relevance to Lawyers* (Equivio undated white paper), http://www.equivio.com/files/files/White Paper - Far from the Black Box – Explaining Equivio Relevance to Lawyers.pdf, at 9.

[23] A "control set" is "[a] random sample of documents coded at the outset of a search or review process that is separate from and independent of the training set. Control sets are used in some technology-assisted review processes. They are typically used to measure the effectiveness of the machine learning algorithm at various stages of training, and to determine when training may cease." Glossary, *supra* n.1, at 13.

[24] SIGIR 2014 Paper, *supra* n.15, at 160; Practical Law Article, *supra* n.15, at 36.

[25] *See* Gordon V. Cormack & Maura R. Grossman, *Engineering Quality and Reliability in Technology-Assisted Review*, PROCEEDINGS OF THE 39TH INT'L ACM SIGIR CONFERENCE ON RESEARCH AND DEV. IN INFO. RETRIEVAL 75 (2016), http://dx.doi.org/10.1145/2911451.2911510, and the discussion on Quality Assurance *infra* section IV.H.

documents reviewed thus far exceeds the number of positive decisions, plus 1,000.[26] At the outset, most documents presented for review will be relevant, and hence labeled positive; the stopping criterion will not be met, and the review will continue. Eventually, unreviewed relevant documents will become more and more scarce, with the consequence that most documents selected for review will be non-relevant, and hence labeled negative. Eventually the number of negatives will exceed the number of positives by 1,000 or more, the stopping criterion will be met, and the review can cease. There are other, more formal methods for determining when to stop a CAL review,[27] but the authors have found this one to be easy to implement and effective.

IV. Measuring Success

Choosing an appropriate method to employ for review involves weighing tradeoffs among a number of considerations, including the (i) effectiveness, (ii) efficiency, (iii) cost, (iii) availability, (iv) familiarity, and (v) general acceptance of candidate methods. Effectiveness and efficiency are amenable to scientific inquiry, while the other considerations depend on social, legal, and market factors that, while influential, are difficult to measure, and beyond the scope of this chapter.

The most commonly used measures of effectiveness are *recall* and *precision*. Recall quantifies how nearly *all* the relevant documents are found[28]; precision quantifies how nearly *only* the relevant documents are found.[29] Unfortunately, for reasons previously discussed in section II on "Review Objectives," recall and precision can never be known with certainty, and can only be estimated. Moreover, the manner in which recall and precision are estimated has a profound effect, such that different recall and precision estimates are incomparable unless they are calculated under precisely the same conditions.

The net effect is that *naked recall and precision numbers are essentially meaningless*. A claim of "70% recall" is more properly described as a recall *estimate*, and whether or not it indicates that an acceptable proportion of the relevant documents have been found by a review process depends on how the estimate was derived (as well as other legal considerations related to whether a court or regulator might deem that proportion as indicative of a "reasonable" or acceptable review). If the estimate is derived from the coding of an *independent* reviewer, 70% recall is at or

---

[26] *See* Maura R. Grossman et al., *TREC 2016 Total Recall Track Overview*, in PROCEEDINGS OF THE 25TH TEXT RETRIEVAL CONFERENCE (NIST 2016), http://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf, at 5. Another way to phrase this stopping criterion is: when the total number of documents reviewed exceeds twice the number of responsive documents, plus 1,000.

[27] *See generally* Cormack & Grossman, *supra* n.25.

[28] "Recall" is "[t]he fraction of relevant documents that are identified as relevant by a search or review method," *i.e.*, a measure of completeness. Glossary, *supra* n.1, at 27.

[29] "Precision" is "[t]he fraction of documents identified as relevant by a search or review effort, that are in fact relevant," *i.e.*, a measure of accuracy. Glossary, *supra* n.1, at 25.

near the upper limit of what could be achieved by exhaustive manual review,[30] which, in most contexts, should represent a *de facto* standard of acceptable effectiveness.

If a second review were to achieve a 60% recall estimate according to the same independent reviewer, we might reasonably conclude that the second review found fewer relevant documents than the first, provided that we could exclude the possibility that the difference was a fluke, the product of chance, or the result of some confounding factor. Similarly, we might reasonably conclude that a third review achieving an estimated 80% recall found more relevant documents than the first, subject to the same caveats.

It would *not* be appropriate to conclude that the three manual reviews described above found 70%, 60%, or 80% of the relevant documents, respectively, or, conversely, that they missed 30%, 40%, or 20% of the relevant documents. All that can be said is that they found a certain proportion of the documents *that an independent reviewer would have coded positive*. Almost certainly, some—perhaps even a substantial—fraction of the independent reviewer's positive coding decisions would be wrong (or at least disputable), resulting in an *underestimate* of the proportion of relevant documents found, and an *overestimate* of the number of relevant documents missed.

The bottom line is that recall and precision estimates convey little information as an *absolute indicator* of how nearly all and only the relevant documents have been identified by a particular review effort. When estimated by reference to an independent review, 65% recall and 65% precision are close to the best that can be achieved,[31] and to demand or promise higher is unrealistic. As Ellen Voorhees noted in her seminal 2000 study, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, "[t]he [recall and precision estimates] for the two sets of secondary judgments imply [that] a practical upper bound on [estimated] retrieval system performance is 65% precision at 65% recall since that is the level at which humans agree with one another."[32]

At the same time, challenges in estimating recall provide no license to willfully exclude 35%—or any other specific number—of relevant documents. The objective of review remains unchanged: to identify, as nearly as practicable, all and only the relevant documents. Recall and precision estimates approaching or exceeding 65% may provide evidence of a satisfactory result, if those estimates are derived from an *independent* coding effort, rather than the same review team that performed the original manual review.

---

[30] *See* Voorhees, *supra* n.6.
[31] *See id.*
[32] Voorhees, *supra* n.*6,* at 701.

In practice, seldom are the resources available to conduct a separate, independent review, over and above the original review, for the purpose of estimating the recall and precision of the original review. At best, a separate (but rarely independent) review is conducted on a *random sample* of the documents in the collection. Over and above the uncertainties in relevance determinations we have previously discussed, sample-based estimates are also subject to random error. This random error is typically quantified by the statistical terms "margin of error,"[33] "confidence interval,"[34] and "confidence level,"[35] which are the source of much confusion—and many misconceptions and ill-conceived practices—in eDiscovery circles.[36]

A third source of confusion regarding recall and precision concerns the particular phase of the review process that is being measured. Recall and precision estimates are most informative when they measure the *end-to-end* effectiveness of the review process, including culling efforts and other activities that precede the selection of documents for review, as well as the ultimate coding decision of the reviewers, as amended by any quality control processes. All too often, however, recall and precision estimates are calculated only for the document-selection component of the review (*i.e.*, the application of TAR alone), under the tacit assumption that the antecedent culling and subsequent manual review processes are flawless.

For nearly the last decade, the authors have conducted a comprehensive program of experimental research evaluating the end-to-end effectiveness of review methods

---

[33] A "margin of error" is "[t]he maximum amount by which a point estimate might likely deviate from the true value, typically expressed as 'plus or minus' a percentage, with a particular confidence level. For example, one might express a statistical estimate as '30% of the documents in the population are relevant, plus or minus 3%, with 95% confidence.' This means that the point estimate [of the prevalence or richness of the collection] is 30%, the margin of error is 3%, the confidence interval is 27% to 33%, and the confidence level is 95%." Glossary, *supra* n.1, at 22.

[34] A "confidence interval . . . [a]s part of a statistical estimate, [is] a range of values estimated to contain the true value, with a particular confidence level." Glossary, *supra* n.1, at 12.

[35] The "confidence level . . . [a]s part of a statistical estimate, [is] the chance that a confidence interval derived from a random sample will include the true value. For example, '95% confidence' means that if one were to draw 100 independent random samples of the same size, and compute the confidence interval from each sample, about 95 of the 100 confidence intervals would contain the true value." Glossary, *supra* n.1, at 12.

[36] *See generally* Comments Paper, *supra* n.15. By way of example, the following assertions involving statistics are typical, but, unfortunately, incorrect: "The confidence tests Biomet ran as part of its process suggest a comparatively modest number of documents would be found." *In Re: Biomet M2a magnum Hip Implant Prods. Liab. Litig.*, No. 3:12-MD-2391, Order Regarding Discovery of ESI (N.D. Ind. Apr. 18, 2013), at 5, *available at* http://www.ctrlinitiative.com/wp-content/uploads/2014/Predictive Coding Opinions/Biomet_1_DiscoveryOrder_April18.pdf ; "[O]ne can avoid reviewing 80% or more of the collection and still be 95% confident of finding every relevant document." Andy Kraftsow, *Comment: When is Litigation Like Las Vegas?*, LEGAL INSIDER (Jan. 13, 2013), https://www.legaltechnology.com/latest-news/comment-when-is-litigation-like-las-vegas/; "[T]he overturn rate for non-responsive documents was only 2 percent. . . . At this point, we felt confident we had identified all potentially responsive documents." *How CDS Saved Hundreds of Attorney Hours with Assisted Review*, Relativity – Customer Wins (kCura LLC 2012), https://www.kcura.com/relativity/ediscovery-resources/customer-wins/cds-assisted-review/.

using CAL and other TAR technologies, as well as manual review. Our experimental results have led to enhancements to the CAL process that we have employed in practice on hundreds of reviews since 1999; at the same time, our practical experience, as well as concerns that have been raised in the eDiscovery community, have guided our choice of questions to address in our empirical research.

V. Research Results

A. Assessor Disagreement

The issue of relevance assessment has challenged researchers since computers were first used for information retrieval. Because "those who cannot remember the past are condemned to repeat it,"[37] we defer to IR pioneer Tefko Saracevic to summarize the first 50 years of research in IR:

> In the mid 1950s there was an attempt to test the performance of two competing IR systems developed by separate groups . . . each group searched 98 requests using the same 15,000 documents, indexed separately, in order to evaluate performance based on relevance of retrieved documents. *However, each group judged relevance separately.* Then, not the systems' performance, but their relevance judgments became contentious. The first group found that 2,200 documents were relevant to the 98 requests, while the second found that 1,998 were relevant. There was not much overlap between groups. The first group judged 1,640 documents relevant that the second did not, and the second group judged 980 relevant that the first did not. You see where this is going. Then they had reconciliation, considered each other's relevant documents, and again compared judgments. Each group accepted some more as relevant, but at the end, they still disagreed; their rate of agreement, even after peace talks, was 30.9%. That did it. The first ever IR evaluation did not continue. It collapsed. *Because of relevance assessments.* Moreover, it seems that the rate of human agreement on relevance assessment hovers indeed around that figure. . . .[38]

---

[37] This famous statement, which has many variants and paraphrases, has been attributed to George Santayana. https://en.wikiquote.org/wiki/George_Santayana.

[38] Tefko Saracevic, *Why is Relevance Still the Basic Notion in Information Science? (Despite Great Advances in Information Technology)*, in REINVENTING INFO. SCI. IN THE NETWORKED SOC., PROCEEDINGS OF THE 14TH INT'L SYMPOSIUM ON INFO. SCI. 26 (May 2015), https://zenodo.org/record/17964/files/keynote2.pdf (emphasis in original).

"Peace talks" involving the coding of disputed documents were similarly contentious and unproductive in *Da Silva Moore* v. *Publicis Groupe*,[39] the 2012 federal case of first impression approving the use of TAR, and in other cases since then.[40]

B. The Roitblat, Kershaw and Oot "EDI Study"

A 2010 study by Herbert Roitblat, Patrick Oot, and Anne Kershaw—cited in *Da Silva Moore* v. *Publicis*[41] as one of the authorities showing the superiority of TAR over manual review—observed similarly low rates of agreement between a pair of qualified human reviewers recruited for the study, and an even lower rate of agreement between those reviewers and the earlier exhaustive manual review conducted by a team of 225 attorneys to meet the requirements of a U.S. Department of Justice "second request" involving the acquisition of MCI by Verizon.[42]

Fortunately, research suggests that it is not necessary for the parties to agree on the relevance of every document in order to determine the relative effectiveness of two different IR approaches.[43] In general, if review #1 achieves a higher effectiveness score than review #2 in the eyes of competent, independent review #3, we can infer that method #1 is likely more effective than method #2, even is #3 is imperfect.

Roitblat et al. used the prior production as "review #3" to evaluate the relative effectiveness of the reviews conducted by their two experts ("review A" and "review B"). According to review #3, reviews A and B achieved 49% and 54% recall, and 20% and 18% precision, respectively—an insubstantial and statistically insignificant difference.[44]

Again according to review #3, Roitblat et al. further evaluated the effectiveness of reviews C and D, which were conducted using undisclosed commercial TAR methods. These methods achieved 46% recall and 53% recall, respectively—an insubstantial and statistically insignificant difference from human reviews A and B. On the other hand, reviews C and D achieved substantially and significantly higher precision: 27% and 29%, respectively.[45]

---

[39] *See*, *e.g.*, *Da Silva Moore* v. *Publicis Groupe SA*, No. 11 Civ. 1279 (ALC) (AJP), Tr. (S.D.N.Y. May 7, 2012).

[40] *See*, *e.g.*, Joint letter to Hon. Andrew J. Peck, ECF Doc. No. 398, filed in *Rio Tinto PLC* v. *Vale SA*, No. 14-cv-3042 (RMB) (AJP) (S.D.N.Y. Nov. 12, 2015), at 24-25, *available at* http://ctrlinitiative.com/wp-content/uploads/2016/01/Rio-Tinto-Status-Update-Incl.-Predictive-Coding-ECF-398-11-12-2015-1.pdf (advising the Court that "[a]fter a series of meet and confers to discuss coding challenges, the parties were still unable to resolve coding disputes for a handful of documents and agreed to submit a handful of disputed documents to Special Master Grossman for resolution.").

[41] *Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182, 190 (S.D.N.Y. 2012).

[42] Roitblat et al., *supra* n.6.

[43] *See*, *e.g.*, Voorhees and Bailey, *supra* n. 6.

[44] Roitblat et al., *supra* n.6.

[45] *Id.*

As we previously noted, these recall numbers should not be interpreted to mean that the manual or TAR reviews missed half of the relevant documents, that only one-fifth of the documents identified by the manual reviews were relevant, or that only three-tenths of the documents identified by the TAR reviews were relevant. We can, however, say that the manual and TAR reviews identified about the same number of relevant documents, and that the TAR reviews identified substantially fewer non-relevant documents.

C. TREC: The Text REtrieval Conference Legal Track Interactive Task

The Text REtrieval Conference ("TREC"), co-sponsored by the National Institute of Standards and Technology ("NIST") and the U.S. Department of Defense, is an annual workshop and conference that has, since its inception in 1992, been one of the premier venues for IR research. Its stated purpose is:

> [T]o 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.[46]

From 2006 through 2011, the TREC Legal Track addressed the application of advanced search technology to several aspects of eDiscovery. In particular, the TREC Legal Track Interactive Task, which ran from 2008 through 2010, evaluated the end-to-end effectiveness of various review strategies carried out by participating teams.

In each year, the Interactive Task required participants to identify, as nearly as they could, from a large publicly available document collection, all and only the documents responsive to one or more mock RFPs. In 2008, the collection consisted

[46] National Institute of Standards and Technology, *Text Retrieval Conference (TREC) Overview*, http://trec.nist.gov/overview.html.

of seven million documents previously collected in connection with the tobacco litigation that culminated in the Master Settlement Agreement among 49 state and territorial jurisdictions and four tobacco manufacturers.[47] In 2009 and 2010, respectively, the collection consisted of 847,791 and 685,592 email messages and attachments collected from Enron Corporation by the Federal Energy Regulatory Commission in the course of its investigation of Enron's failure.[48]

Interactive Task participants were provided with a mock complaint, and one or more requests for production concerning subject matters to be found in the document collection, both of which were composed by Track coordinators and other volunteers. For each RFP (referred to as a "topic" in TREC parlance), a volunteer "Topic Authority" ("TA") was assigned. The TA was a senior lawyer who provided consultation to the participants during the course of their review and acted as the final arbiter of relevance during the subsequent evaluation process.

Relevance assessment for the purposes of evaluation was accomplished using a novel, three-phase approach. In the "first-pass review," volunteer reviewers—supplied either by law school *pro bono* programs or eDiscovery contract-review service providers—coded a statistical sample of documents as relevant or non-relevant. These coding decisions were released to TREC participants who were invited to "appeal" those decisions with which they disagreed. The Topic Authority reviewed all documents whose coding was appealed, and rendered a final relevance determination for each.

For the purpose of calculating the recall and precision of the participants' efforts, where relevance determinations were not appealed, the first-pass reviewer's coding was taken to be correct; where relevance determinations were appealed, the TA's final coding determination was taken to be correct.

The 2008 Legal Track reported—and introduced to the eDiscovery lexicon—a summary measure known as F_1.[49] F_1 combines recall and precision into a single summary measure, with the lesser of the two given more weighting. Thus, in order to achieve high F_1, it is necessary to achieve both high recall (approaching the ideal

---

[47] Douglas W. Oard et al., *Overview of the TREC 2008 Legal Track*, in PROCEEDINGS OF THE 17TH TEXT RETRIEVAL CONFERENCE (NIST 2008), at 3, http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf.

[48] Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, in PROCEEDINGS OF THE 18TH TEXT RETRIEVAL CONFERENCE (NIST 2009), at 4-5, http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf, and Gordon V. Cormack, *Overview of the TREC 2010 Legal Track*, in PROCEEDINGS OF THE 19TH TEXT RETRIEVAL CONFERENCE (NIST 2010), at 2-3, http://trec.nist.gov/pubs/trec19/papers/LEGAL10.OVERVIEW.pdf, respectively.

[49] Oard et al., *supra* n.47, at 7-8. "F_1" is defined as "[t]he harmonic mean of recall and precision, often used in information retrieval studies as a measure of the effectiveness of a search or review effort, which accounts for the tradeoff between recall and precision. In order to achieve a high F_1 score, a search or review effort must achieve *both* high recall and high precision." Glossary, *supra* n.1, at 16 (emphasis in original).

of all relevant documents) *and* high precision (approaching the ideal of only relevant documents).

Four teams—two from universities and two from eDiscovery service providers—participated in the 2008 Interactive Task. The team from one of the service providers (H5) achieved remarkably high recall, precision, and F_1 scores of 62%, 81%, and 71%, respectively, using a rule-based TAR approach.[50] By comparison, no other team achieved scores recall, precision, and F_1 scores higher than 16%, 80%, and 39%, respectively.[51]

In 2009, the H5 team achieved similarly high scores for the review they conducted (topic 204; 80% F_1), as did a team from the University of Waterloo (led by the second author) for each of four reviews that the team conducted (topics 201, 202, 203, and 207; 84%, 76%, 77%, and 83% F_1).[52] A second industry team (Cleary/Backstop) achieved 80% F_1 on one of the three reviews they conducted (topic 207); a third industry team (Equivio) achieved 61% and 58% F_1 on the two reviews it conducted (topics 205 and 207); and a fourth industry team (Clearwell) achieved 62% F_1 on one of the two reviews it conducted (topic 202).[53] The remaining 15 of 24 reviews—from eight of the eleven participating teams—achieved F_1 scores between 2% and 43%.[54]

D. The 2011 JOLT Study

While the results from the TREC 2008 and 2009 Legal Track Interactive Tasks were remarkable, they left unanswered the question of how the well-performing TAR processes employed by industry participants and the University of Waterloo would compare to exhaustive manual review. While the results reported at TREC were numerically greater than those reported by Roitblat et al. for human review,[55] and greater than Voorhees' observed "upper bound on retrieval performance,"[56] they were incomparable, as they came from different review tasks and reflected different methods of assessing relevance.

The Interactive Task was designed to compare the effectiveness of the review strategies implemented by participating teams, none of which employed exhaustive manual review. For the purposes of evaluation, a manual review—the first-pass assessment—had been conducted, albeit only for a statistical sample of the documents in the collection. Disagreements between the first-pass assessment and

---

[50] Oard et al., *supra* n. 47, Table 15 at 30.
[51] *Id.*
[52] Hedin et al., *supra* n.48, Table 6 at 15.
[53] *Id.*
[54] *Id.*
[55] *See* Roitblat et al., *supra* n.6.
[56] Voorhees, *supra* n.6, at 701.

participating teams were anticipated in the experimental design; such disagreements were adjudicated by the Topic Authority.

The purpose of this adjudication was to achieve the most accurate possible relevance determination for use in evaluating and comparing the effectiveness of the participants' reviews, all of which employed some form of TAR. In their 2011 JOLT Study, the authors employed the adjudicated relevance determinations for a different purpose not anticipated at the time: to evaluate and compare the effectiveness of the manual first-pass review with the results achieved by the most consistently effective TAR reviews.

The results indicated that the manual reviews achieved, on average, 59% recall, 32% precision, and 36% F_1, while the TAR reviews achieved, on average, 77% recall, 85% precision and 80% F_1.[57] While each measure is higher for TAR than for manual review, the difference in recall was not statistically significant, while the differences in precision and recall were.[58]

These results were consistent with those reported by Roitblat et al.: In terms of recall, there was little to choose between the TAR and manual review results; in terms of precision (and, consequently, F_1), the TAR results were vastly superior. At the same time, the TAR reviews involved human review of only 2% of the collection—or fifty times less effort than an exhaustive manual review would entail—a very substantial difference.[59]

It is important to note that the studies by Roitblat et al. and by the authors compared specific TAR methods to reasonably well-conducted manual reviews under laboratory conditions. The results suggest that methods similar to those tested can, in practice, achieve superior results to manual review. The results cannot, however, be interpreted to suggest that methods dissimilar to those tested—whether labeled as "TAR" or otherwise—improve on manual review.

E. Comparing TAR Methods

The 2011 JOLT Study has been cited, either directly or by reference, in cases of first impression approving the use of TAR in the United States, Ireland, the United Kingdom, and Australia.[60] An apt characterization of the evidence is offered by Master Matthews in the High Court of Justice Chancery Division (U.K.):

---

[57] 2011 JOLT Study, *supra* n.4, Table 7 at 37.
[58] *Id.*
[59] *Id.* at 43.
[60] *See*, *e.g.*, *McConnell Dowell Constructors (Austl.) Pty Ltd* v. *Santam Ltd & Ors (No 1)*, [2016] VSC 734 (Austl.); *Pyrrho Inv. Ltd.* v. *MWB Prop. Ltd.*, [2016] EWHC (Ch) 256 (Eng.); *Irish Bank Resol. Corp.* v. *Quinn*, [2015] IEHC 175 (H. Ct.) (Ir.); *Rio Tinto PLC* v. *Vale S.A.*, 306 F.R.D. 125 (S.D.N.Y. 2015); *Progressive Casualty Ins. Co.* v. *Delaney*, Case No. 2:11-cv-00678, 2014 WL 3563467 (D. Nev. July 18, 2014); *Fed. Hous. Fin. Agency* v. *HSBC North Am. Holdings Inc.*, No. 1:11-cv-06188-DLC, 2014 WL

> There is no evidence to show that the use of [TAR] software leads to less accurate disclosure being given than, say, manual review alone or keyword searches and manual review combined, and indeed there is some evidence (referred to in the US and Irish cases to which I referred above) to the contrary.[61]

More sweeping generalizations of the Roitblat et al. and 2011 JOLT Study results have been advanced, both to promote so-called TAR methods that bear little resemblance to those tested, and as straw men to impugn the studies and all TAR.[62] At the same time, a number of burdensome practices associated with untested TAR methods, as well as the statistical apparatus of laboratory IR evaluation, have erroneously been associated with TAR in general.[63]

In order to investigate the relative effectiveness of different TAR methods, in 2014, the authors introduced a taxonomy of supervised machine-learning methods for TAR representative of the three basic approaches to TAR taken by eDiscovery service providers in the market: (i) Simple Passive Learning ("SPL"), (ii) Simple Active Learning ("SAL"), and Continuous Active Learning ("CAL").[64]

Our taxonomy excluded rule-based TAR methods that relied on opaque or ill-specified techniques that were difficult to characterize, as well as methods that the authors did not consider to be TAR, which are marketed under names such as "concept search," "clustering," "concept clustering," "find similar," "visualization," "deduplication," "near-deduplication," and "email threading." We have since

---

584300 (S.D.N.Y. Feb. 14, 2014); *Nat'l Day Laborer Org. Network* v. *U.S. Immigr. & Customs Enf't Agency*, 877 F. Supp. 2d 87 (S.D.N.Y. 2012); *Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012).

[61] *Pyrrho Inv. Ltd.* v. *MWB Prop. Ltd.*, [2016] EWHC (Ch) 256 (Eng.), at 14.

[62] *Compare*, *e.g.*, *Visualize a New Concept in Document Decisioning*, OrcaTec – FAQ (Internet Archive Oct. 1, 2011), https://web.archive.org/web/20111001071436/http://orcatec.com/index.php/resources/faq (stating "Can OrcaTec provide any scientific evidence concerning such processes as predictive coding? Grossman & Cormack in the Richmond Journal of Law & Technology" (with link), when the OrcaTec tool bore no resemblance to the TAR methods studied by Grossman and Cormack) *with* Bill Speros, *Despite Early Success, Technology Assisted Review's Acceptance Is Limited by Lack of Definition*," News & Press: ACEDS News (Aug. 31, 2016), http://www.aceds.org/news/3059301 (stating that the court in *Da Silva Moore* "misperceive[ed] the [2011 JOLT] article upon which the court relied as being proof-of-capability rather than as proof of concept" and concluding that "until TAR consolidates definitions about what it is, its capabilities and its limitations, and specifies any underlying science and all necessary protocols, TAR will face meaningful criticism about its reliability. And it should.")

[63] *See generally*, *e.g.*, Karl Schieneman & Thomas C. Gricks III, *The Implications of Rule 26(g) on the Use of Technology-Assisted Review*, 7 FED. CTS. L. REV. 239 (2013), http://www.fclr.org/fclr/articles/html/2010/Gricks.pdf. *Cf.* Comments Paper, *supra* n.15 (responding to Schieneman & Gricks' article).

[64] SIGIR 2014 Paper, *supra* n.15; *see also* Comments Paper, *supra* n.15.

published a broader taxonomy of TAR tools, as well as non-TAR tools, which we characterize as tools for search and analysis.[65]

To measure the relative effectiveness of supervised machine-learning methods for TAR, we created an open-source "TAR Evaluation Toolkit"[66] that simulates SPL, SAL, and CAL in a laboratory environment. Using data collected from TREC 2009, as well as four legal matters in which the authors had been involved, we found that, for a given level of review effort, CAL achieved the highest recall (and, as a consequence, the highest precision and F_1) of the three methods.[67] We found that, *given the correct parameter settings*, SAL could achieve recall comparable to CAL, but only for one particular level of effort.[68] We found that SPL yielded results substantially inferior to those achieved by CAL or SAL.[69] This peer-reviewed study was presented at *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*.[70]

F. Autonomy and Reliability of CAL

A commonly expressed view in the legal community has been that TAR requires exceptional skill on the part of an operator; for example, to select the appropriate training documents and operating parameters for the learning method.[71] In *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*[72] the authors evaluated "AutoTAR," an enhancement of CAL that has no parameters to set, and requires at the outset only a single relevant document, or in the alternative, a fragment of text containing relevant content. Given this initial input, AutoTAR presents documents in sequence for review, and the coding is returned to AutoTAR. The process continues until evidence suggests that substantially all relevant documents have been presented for review.

---

[65] *See* Maura R. Grossman & Gordon V. Cormack, *A Tour of Technology-Assisted Review*, ch. 3 in Jason R. Baron et al. (eds.), PERSPECTIVES ON PREDICTIVE CODING AND OTHER ADVANCED SEARCH METHODS FOR THE LEGAL PRACTITIONER (ABA Publishing 2016).
[66] http://cormack.uwaterloo.ca/tar-toolkit/.
[67] SIGIR 2014 Paper, *supra* n.15.
[68] *Id.*
[69] *Id.*
[70] *See* http://sigir.org/sigir2014/; *see also id.*
[71] *See*, *e.g.*, Ralph C. Losey, *Why the 'Google Car' Has No Place in Legal Search*, e-Discovery Team Blog (Feb. 24, 2016), https://e-discoveryteam.com/2016/02/24/why-the-google-car-has-no-place-in-legal-search/ (regarding selection of training documents); Rishi Chhatwal et al., *Empirical Evaluations of Preprocessing Parameters' Impact on Predictive Coding's Effectiveness*, in PROCEEDINGS OF THE 2016 IEEE INT'L CONFERENCE ON BIG DATA 1394 (2016), *available at* https://www.navigant.com/-/media/www/site/insights/legal-technology/2017/predictive-codings-effectiveness.pdf (regarding selection of operating parameters).
[72] Gordon V. Cormack & Maura R. Grossman, *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*, https://arxiv.org/abs/1504.06868 [cs.IR] (Apr. 15, 2015).

Our results show that, *regardless of what initial input is chosen*, AutoTAR finds substantially all relevant documents with less review effort than the CAL method we had previously evaluated.[73] We observed the same results for a wide variety of publicly available IR benchmarks, including the 103 subjects of the Reuters RCV1-v2 dataset, the 50 topics of the TREC 6 AdHoc Task, and the 50 topics of the TREC 2002 Filtering Track, as well as for datasets from four actual legal matters.[74]

An open-source implementation of AutoTAR was subsequently used as the "Baseline Model Implementation" ("BMI")[75] for the TREC Total Recall Tracks in 2015[76] and 2016.[77] As in the earlier TREC Legal Track Interactive Task, participants were asked to find, as nearly as they could, all and only the relevant documents in the collection. In contrast to the Legal Track, however, Total Recall participants submitted documents incrementally to a Web server for assessment, and received a relevance label (derived automatically from a prior labeling of the entire collection) for each document immediately when it was submitted. This architecture allowed for precise tracking of each team's recall as a function of the number of documents submitted.

Participants could use fully automated strategies like BMI, or manual strategies involving any combination of human and computer input, including keyword search, manual review, and hand-selected training documents. While some participants achieved higher recall than BMI on some topics, at some levels of review effort, no participant at TREC 2015 or 2016—whether automatic or manual—achieved consistently higher recall than BMI, for the same level of effort.[78]

The 2015 and 2016 Total Recall Tracks evaluated TAR systems with respect to a diverse set of datasets and topics. In 2015, systems were evaluated with respect to 53 different topics and five datasets: Ten topics were developed for a collection of approximately 290,099 emails from Jeb Bush's administration as Governor of Florida; ten topics were developed for a collection of 465,147 postings from Blackhat World and Hacker Forum; ten topics were developed for a collection of 902,434 on-line news clippings from the northwestern United States and southwestern Canada; four preexisting topics reflecting statutory definitions of various types of records and non-records were used in connection with a collection of 401,953 emails from Tim Kaine's administration as Governor of Virginia; and nineteen preexisting topics reflecting ICD-9 codes were used in connection with the

---

[73] *Id.*
[74] *Id.* at 2.
[75] Baseline Model Implementation for Automatic Participation in the TREC 2015 Total Recall Track, http://cormack.uwaterloo.ca/trecvm/.
[76] *See* Adam Roegiest et al., *TREC 2015 Total Recall Track Overview*, in PROCEEDINGS OF THE 24TH TEXT RETRIEVAL CONFERENCE (NIST 2015), http://trec.nist.gov/pubs/trec24/papers/Overview-TR.pdf.
[77] See Maura R. Grossman et al., *supra* n.26.
[78] *See* Roegiest et al., *supra* n.76; Grossman et al., *supra* n.26.

MIMIC II clinical dataset, consisting of 31,538 medical records from an intensive care unit.[79]

In 2016, systems were evaluated with respect to an additional 34 topics developed for the Jeb Bush collection; six topics developed for a collection of 2.1M emails from the administrations of Illinois Governors Rod Blagojevich and Pat Quinn; and four preexisting topics were used in connection with a collection of 800,000 Twitter tweets.[80]

Overall, the results indicate that fully autonomous TAR systems can achieve very high recall levels, for reasonable effort, for a wide variety of datasets and relevance criteria. Results on the Tim Kaine and MIMIC II datasets are of particular interest, because relevance was formally defined, and relevance determinations were rendered by independent professionals (the Virginia Senior State Records Archivist and physicians, respectively) in the course of their employment.

G. Facets of Relevance

It has been suggested by some that TAR may exhibit "blind spots" in that a TAR review may miss certain kinds of relevant documents, either because those documents have an unusual format or because they pertain to an obscure aspect of relevance.[81] Our peer-reviewed study, *Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review*, presented at *The 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*,[82] indicates that CAL, when it has found nearly all relevant documents overall, has also found nearly all relevant documents for each facet of relevance, whether those facets are defined as file types or as substantive sub-topics.[83] It may be that CAL identifies certain types of documents or documents representing certain aspects of relevance sooner than others, but once such documents become scarce, it identifies other facets, and so on, until all facets have been identified.

This result was reaffirmed at the TREC 2016 Total Recall Track, where assessors were asked to sort relevant documents into subfolders, based on the particular

---

[79] *See* Roegiest et al., *supra* n.76, at 3-5.
[80] See Grossman et al., *supra* n.26, at 3-5.
[81] See, e.g., *Proper Use of Predictive Coding Technology* (Inspired Review Blog Jan. 7, 2104), http://www.inspiredreview.com/blog5.html ("Some questions have already arisen regarding the ability of predictive coding algorithms to properly address terse documents (documents that do not contain abundant text for language based analysis such as spreadsheets or short documents) and 'novel content' documents."). *See also* Comments Paper, *supra* n.15, at 304-05.
[82] *See* http://sigir2015.org/; Gordon V. Cormack & Maura R. Grossman, *Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review*, in PROCEEDINGS OF THE 38TH INT'L ACM SIGIR CONFERENCE ON RESEARCH AND DEV. IN INFO. RETRIEVAL 763 (2015), http://dl.acm.org/citation.cfm?doid=2766462.2767771.
[83] Cormack & Grossman, *supra* n.82.

subject matter they contained. When the recall for each subfolder was considered separately, participating systems that achieved high recall overall, also achieved high recall for the documents in each subfolder.[84]

Finally, the same result also was recently reproduced through an independent research effort.[85]

H. Quality Assurance

Recall, precision, and F_1 are commonly used to measure the average effectiveness of IR systems and methods. When we say that a particular method achieves 65% recall and 65% precision, we are generally referring to the average recall and precision achieved by applying the same method to a set of information needs (*i.e.*, queries) that are representative of those that might be encountered in practice. An average provides no guarantee that, for any given retrieval effort, any particular level of recall or precision will be achieved.

This concern has led parties to use sampling in an often-futile effort to estimate recall and precision for particular review efforts, and to set thresholds as standards of acceptability. For example, in *Global Aerospace* v. *Landow Aviation*,[86] based on the recall level reported in our 2011 JOLT Study, the producing party promised to achieve at least 75% recall, and subsequently represented to the court, after the fact, that 81% recall had been achieved.[87] What was achieved was, in fact, a coarse *estimate* of the recall of only the TAR (document-selection) component of the review process. The estimate itself had a margin of error such that the true value could easily have been less than 75%. More importantly, however, the estimate did not account for the fact that the documents selected by the TAR system were reviewed manually, and only the documents coded relevant and non-privileged by the reviewers were produced. Thus, only if we assumed that the manual review was perfect—that is, achieved 100% recall—would the estimate of 81% recall apply to the end-to-end review. More likely—as determined by an independent

---

[84] *See* Grossman et al., *supra* n.26, at 5.

[85] *See* Thomas Gricks, *Does Recall Measure TAR's Effectiveness Across All Issues? We Put It To The Test*, Catalyst E-Discovery Search Blog (Mar. 23, 2017), https://catalystsecure.com/blog/2017/05/does-recall-measure-tars-effectiveness-across-all-issues-we-put-it-to-the-test/.

[86] Karl Schieneman & Thomas C. Gricks III, *supra* n.63, at 259.

[87] Letter from Gordon S. Woodward, Att'y for the Landow Entities, to All Counsel in *Global Aerospace Inc.* v. *Landow Aviation, L.P.*, Consol. Case No. CL601040 (Va. Cir. Ct. Loudoun Cty. Nov. 30, 2012) ("At the end of the predictive coding process, we conducted a statistically valid sampling program to establish that an acceptable level of document recall had been achieved. As we indicated in our motion to the court . . . the Landow Entities proposed that 75% recall would be adequate. Below is a report reflecting our final analysis with respect to document recall. The report indicates that we achieved 81%.").

assessment—the manual review achieved recall on the order of 70%, for a net end-to-end recall estimate of 57%.[88]

It is not our intent to impugn the adequacy of production in *Global Aerospace*—we have no reason to doubt its quality—but rather, to illustrate the fallacy of relying on ill-specified recall thresholds as acceptance criteria. On the other hand, as has been suggested[89]—it is also not our intent to suggest that, because relevance is difficult to define, and because recall is difficult to estimate on a case-by-case basis, that all measurement should be eschewed and that producing parties should be absolved of any and all responsibility of ensuring the adequacy of the production.

The first step in assuring the adequacy of a production, we believe, is to use a method that has previously been shown to *reliably* achieve high recall.[90] Reliability is the probability that, for any given application, a high-quality result will be achieved. High average recall does not, in itself, imply *reliably* high recall. A method that, for example, achieved 100% recall 80% of the time, and 0% recall 20% of the time, would achieve an apparently high level of 80% recall, on average, but poor reliability, since 20% of the time the method could not be counted on to find anything. One would not likely consider a one-in-five chance of complete failure to be an acceptable risk. On the other hand, a one-in-twenty chance of achieving 74% recall, when 75% recall was deemed acceptable, might be acceptable.

Inextricably intertwined with the notion of reliability is the question of "when to stop?" For CAL, one can select and review documents indefinitely. For SPL and SAL, one can select training documents indefinitely, and when training ceases, one can adjust the sensitivity of the resulting classifier so as to review documents indefinitely. At some point, the decision must be made that enough responsive documents have been found, and that further review is disproportionate. We would like to ensure that, when that decision is reached, with high probability, high recall has been achieved.

In support of this goal, we investigated three methods of achieving high reliability using CAL.[91] One method—the "Target Method"—provably achieved a recall target of 70%, with 95% reliability, at the cost of reviewing a large random sample of documents, over and above those selected by the TAR system as relevant. A second method—the "Knee Method"—achieved better reliability on a wide variety of datasets and information needs with less effort than the Target Method. A third method—the "Budget Method"—achieved vastly superior reliability on the same

---

[88] If 81% of the relevant documents were identified by the TAR system, and 70% of those were correctly coded relevant by the manual review, the end-to-end recall would be 81% x 70% = 56.7%.
[89] *See*, *e.g.*, Herbert L. Roitblat, *Daubert, Rule 26(g) and the eDiscovery Turkey,* OrcaBlog (Aug. 11, 2014), https://web.archive.org/web/20140812155631/http:/orcatec.com/2014/08/11/daubert-rule-26g-and-the-ediscovery-turkey/.
[90] Comments Paper, *supra* n.15, at 305.
[91] Cormack & Grossman, *supra* n.25.

datasets, when the same number of documents as the Target Method were reviewed, but the documents were selected by the TAR system (not through random sampling). At TREC 2016, we investigated a fourth method to achieve high reliability—the simple method described in the When to Stop? section (III.F) above.[92] To our surprise, it worked as well as the more complex Knee Method,[93] but more research is certainly needed in this area.

While a fuller discussion of these stopping criteria is beyond the scope of this chapter, they show that it is possible to reliably determine when to stop a TAR review without resorting to large random samples and faulty statistics.

I. TAR vs. Manual Review Redux

We have recently had occasion to reconfirm the results of our 2011 JOLT Study by comparing the results of using CAL to an exhaustive manual review of 401,960 email messages from the administration of Virginia Governor Tim Kaine, which was previously reviewed by the Virginia Senior State Records Archivist Roger Christman ("Roger"). We showed, using subsequent blind assessments rendered by Roger, that Roger could have achieved the same recall and higher precision, for a fraction of the effort, had he employed CAL to review the 401,960 email messages.[94]

Prior to our study, Roger had rendered decisions for each of three topics, seriatim, as follows: First, "Virginia Tech" documents subject to a legal hold were identified; second, documents not subject to the hold were classified as either "archival records" or "non-records"; finally, documents classified as archival records were categorized as "restricted" or "open" records. Open records are available to the public.[95] As a consequence, the document collection diminished for each subsequent topic.

CAL was run on the same dataset, using Roger's prior decisions to simulate user feedback for the purposes of training the learner. When the CAL run was complete, cases of disagreement between the CAL system and Roger's prior coding were identified, and Roger rendered a second relevance determination for a sample of these documents in a double-blind review, where neither Roger nor the authors were aware of Roger's previous determinations. The overlap[96] between Roger's

---

[92] Gordon V. Cormack & Maura R. Grossman, *"When to Stop": Waterloo (Cormack) Participation in the TREC 2016 Total Recall Track*, in PROCEEDINGS OF THE 25TH TEXT RETRIEVAL CONFERENCE (NIST 2016), http://trec.nist.gov/pubs/trec24/papers/WaterlooCormack-TR.pdf.
[93] *See id.*
[94] Cormack & Grossman, *supra* n.4.
[95] *See* http://www.virginiamemory.com/collections/kaine/. *See also* http://cormack.uwaterloo.ca/kaine (the authors' CAL demonstration using the Kaine open records).
[96] "Overlap" or "Jaccard Index" is "[a] measure of the consistency between two sets (e.g., Documents Coded as Relevant by two different reviewers). . . . Empirical studies have shown that expert

first and second determinations was 80.6%, 60.2%, and 64.2%,[97] for each of the three classifications—at the high end of what one might expect for independent reviewers, but far from perfect. Two months later, Roger conducted a third relevance determination in every case where his first and second determinations had been inconsistent, again blind to his previous determinations.

According to Roger's final determinations, we calculated recall and precision for Roger's original review, and for CAL. Roger's recall ranged from 89% to 97%, while CAL's ranged from 90% to 96%—not a significant difference.[98] Roger's precision ranged from 75% to 91%, while CAL's ranged from 80% to 96%—a significant difference *in favor of CAL*.[99] F_1 similarly favored CAL by a significant margin.[100]

Overall, the Roitblat, Kershaw & Oot study, our 2011 JOLT Study, and our SIGIR 2017 *Roger and Me* study, all show the same result: There is no significant difference in the recall achieved by the TAR systems studied and manual review, and significantly superior precision for the TAR systems. This should reaffirm the reasonableness of using at least some forms of TAR.

VI. The Future

Review for production is a difficult problem. Conventional review methods using keyword culling and manual review are burdensome and far from perfect, as shown by the scientific literature, both within the context of eDiscovery and within the context of IR in general. Methods for measuring review effectiveness are similarly burdensome and imperfect.

Vendors, service providers, and consumers need to gather evidence that the review methods they use—whether manual or TAR—work effectively. Doing so is far more challenging than merely reviewing a "statistically significant sample [sic]"[101] of documents for the purpose of training the system or calculating recall.

---

reviewers commonly achieve Jaccard Index scores of about 50%, and that scores exceeding 60% are rare." Glossary, *supra* n.1, at 20, 25.

[97] Cormack & Grossman, *supra* n.4, Table 5 at 7.

[98] *Id.*, Table 3 at 7.

[99] *Id.*, Table 3 at 7.

[100] *Id.*

[101] *See*, *e.g.*, Tracy Greer, *Electronic Discovery at the Antitrust Division: An Update*, U.S. Dep't of Justice (June 25, 2015), https://www.justice.gov/atr/electronic-discovery-antitrust-division-update (suggesting that quality assurance "could be accomplished by the producing party providing [the Division with] a statistically significant sample of both relevant and non-relevant documents."); Alison Nadal et al, *E-discovery: The Value of Predictive Coding in Internal Investigations*, INSIDE COUNSEL (Aug. 13, 2013), at 1, http://www.insidecounsel.com/2013/08/13/e-discovery-the-value-of-predictive-coding-in-inte ("Executing an internal investigation using predictive coding begins with generation of a randomly selected, statistically significant seed set of documents."); Bill George, *Predictive Coding Primer Part II: Key Variables in a Predictive Coding Driven Review*, Tanenholz & Associates, PLLC News (May 8, 2013), http://tanenholzlaw.com/predictive-coding-primer-part-two

Challenges in measurement provide no license to continue to use keyword culling and manual review just because "that is the way it has previously been done." There is ample evidence that those methods are flawed, and there is no evidence that they are superior to certain TAR alternatives. At the same time, there is a growing body of evidence that certain TAR methods can improve on manual review.

We have worked, and continue to work to contribute to that body of evidence, while at the same time improving the state of the art in TAR. We have no reason to think that our Continuous Active Learning™ method is the best that can possibly be achieved, but it is the best of which we are aware at this time, and we continue to work to improve it. We have made an implementation via the TREC Baseline Model Implementation available under the GPL 3.0 public license,[102] and invite researchers and practitioners alike to try it and to work to find more effective and efficient methods to review ESI "to secure the just, speedy, and inexpensive determination of every action and proceeding" as envisioned by Federal Rule of Civil Procedure 1.

("After the control set has been reviewed, the subject matter experts will then need to train the predictive coding model further through review of a statistically significant sample of documents."). The phrase "statistically significant sample" is a non sequitur. *See* Bill Dimm, *TAR 3.0 and Training of Predictive Coding Systems*, Presentation materials from ACEDS Webinar (Dec. 15, 2015), at 12, *available at* http://www.cluster-text.com/papers/TAR_3_and_training_predictive_coding.pdf ("Training set size should never involve phrases like . . . [s]tatistically significant sample (this isn't even a thing!)").

[102] *See* Baseline Model Implementation, *supra* n.75.