©iStockphoto.com/alexandercreative

Continuous Active Learning for TAR

The vast array of technology-assisted review (TAR) tools available in the marketplace, along with their associated jargon, can seem daunting to counsel. But using TAR in litigation and regulatory matters need not be. By implementing a continuous active learning (CAL) protocol, with a TAR tool that uses a state-of-the-art machine-learning algorithm, responding parties can quickly and easily identify substantially all of the relevant documents in a collection and minimize the potential for disputes with requesting parties.

MAURA R. GROSSMAN
OF COUNSEL
WACHTELL, LIPTON, ROSEN & KATZ

Maura focuses her practice on legal, technical, and strategic issues involving domestic and international e-discovery and information governance. She has served as a court-appointed special master, neutral, and expert on search-related issues, and her legal and scientific publications on TAR have been cited as authorities by federal, state, and international courts. Maura also teaches e-discovery courses at Columbia Law School and Georgetown University Law Center, and is a coordinator of the Total Recall Track at the Text REtrieval Conference (TREC).

GORDON V. CORMACK
PROFESSOR
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE,
UNIVERSITY OF WATERLOO

Gordon is a professor of computer science and serves as an independent e-discovery consultant. His primary research focus is on high-stakes information retrieval. Gordon is the co-author of *Information Retrieval: Implementing and Evaluating Search Engines* (MIT Press 2010), as well as more than 100 scholarly articles. He serves on the Program Committee of TREC, and also is a coordinator of its Total Recall Track.

The legal marketplace offers many tools, methods, and protocols purporting to employ technology-assisted review (TAR), under names like predictive coding, assisted review, advanced analytics, concept search, and early case assessment. Yet adoption of TAR has been remarkably slow, considering the amount of attention these offerings have received since the publication of the first federal opinion approving TAR use (see *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012)). The complex vocabulary and rituals that have come to be associated with TAR, including statistical control sets, stabilization, F_1 measure, overturns, and elusion, have dissuaded many practitioners from embracing TAR.

However, none of these terms, or the processes with which they are associated, are essential to TAR. Indeed, none of them apply to continuous active learning (CAL), the TAR protocol that has achieved the best results reported in the scientific literature to date (see Maura R. Grossman & Gordon V. Cormack, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, 2014 Proceedings of the 37th Ann. Int'l ACM SIGIR Conf. on Research & Dev. in Info. Retrieval, 153-62 (2014)).

CAL is a method for finding substantially all relevant information on a particular subject within a vast sea of electronically stored information (ESI). At the outset, CAL resembles a web search engine, presenting first the documents that are most likely to be of interest, followed by those that are somewhat less likely to be of interest. Unlike a typical search engine, however, CAL repeatedly refines its understanding about which of the remaining documents are most likely to be of interest, based on the user's feedback regarding the documents already presented. CAL continues to present documents, learning from user feedback, until none of the documents presented are of interest.

Counsel interested in using CAL should first become familiar with:

- The promise of TAR.
- The differences between CAL and other TAR protocols.
- How the CAL process works.
- How to determine when a TAR review is complete.
- How to evaluate the success of a TAR review.

THE PROMISE OF TAR

In e-discovery, the problem of finding everything that can reasonably be found in response to a request for production has traditionally been addressed through a manual review process, in which every one of a large collection of documents, whether paper or electronic, is reviewed for responsiveness. Often, documents that are unlikely to be responsive are removed using a variety of filters, such as keywords, file types, or date restrictors, before beginning the manual review process.

Scientific research indicates that manual review is neither particularly effective nor efficient, and TAR-based methods, where a computer selects only a fraction of the available ESI for review, can be more effective and efficient than traditional manual review methods (see generally Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 Rich. J.L. & Tech. 11 (2011); Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. Am. Soc'y for Info. Sci. & Tech. 70 (2010)).

The Text REtrieval Conference (TREC)

TREC is an annual workshop, co-sponsored by the National Institute of Standards and Technology (NIST) and the US Department of Defense. Its goals are to:

- Encourage research in information retrieval using large test datasets.
- Increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas.
- Speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems.
- Increase the availability of appropriate evaluation techniques for industry and academic use, including developing new evaluation techniques that are more applicable to current systems.

(TREC, *Overview*, available at *trec.nist.gov*.)

The TREC Total Recall Track evaluates TAR systems using several benchmark document collections, including the Jeb Bush email dataset referred to in this article (see *How CAL Works*). Practitioners, service providers, and researchers may vet their TAR tools or protocols by downloading these collections, or by participating in the Total Recall Track (see *trec-total-recall.org*).

Most TAR tools use supervised machine learning, where a computer algorithm ranks or classifies an entire collection of documents by analyzing the features of training documents previously classified by the user. The supervised machine-learning algorithms used for TAR should not be confused with unsupervised machine-learning algorithms used for clustering, near-duplicate detection, and latent semantic indexing, which receive no input from the user and do not rank or classify documents.

Search E-Discovery Glossary for more on clustering, near-duplicate detection, and other e-discovery terms of art.

Supervised machine-learning algorithms that have been shown to be effective for TAR include:

- **Support vector machines.** This algorithm uses geometry to represent each document as a point in space, and deduces a boundary that best separates relevant from not relevant documents.

- **Logistic regression.** This algorithm estimates the probability of a document's relevance based on the content and other attributes of the document.

Popular, but generally less effective, supervised machine-learning algorithms include:

- **Nearest neighbor.** This algorithm classifies a new document by finding the most similar training document and assuming that the correct coding for the new document is the same as its nearest neighbor.
- **Naïve Bayes (Bayesian classifier).** This algorithm estimates the probability of a document's relevance based on the relative frequency of the words or other features it contains.

(See Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 Fed. Cts. L. Rev. 1, 9, 22, 24 (2013).)

CAL VERSUS OTHER TAR PROTOCOLS

A TAR protocol determines how the learning algorithm is used to select documents for review. TAR providers generally employ one of three protocols: CAL, SAL (simple active learning), or SPL (simple passive learning) (see *Box, Comparing CAL, SAL, and SPL Protocols for TAR*). CAL is much simpler than SAL or SPL, and may be used with any TAR tool, provided that the tool incorporates a supervised machine-learning algorithm that can rank documents by the likelihood that they are relevant.

A number of burdensome steps commonly associated with TAR are absent from CAL, such as:

- Careful crafting of seed sets.
- Determining when to cease training.
- Selecting and reviewing large random control sets, training sets, or validation sets.

(See, for example, *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125, 128 (S.D.N.Y. 2015) (noting that where the TAR methodology uses CAL rather than SAL or SPL, the seed set is much less significant) (see below *Completion of a TAR Review*).)

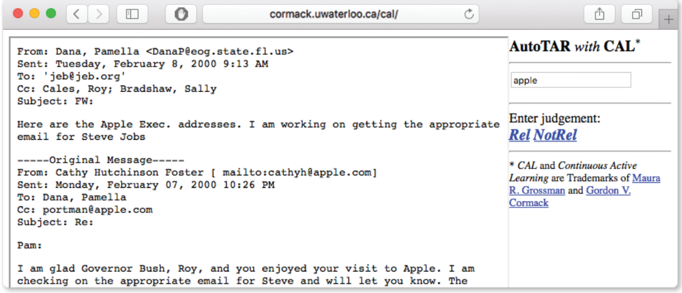
Moreover, CAL should achieve superior results, with less review effort than the other protocols. However, it will yield the best possible results only if the TAR tool incorporates a state-of-the-art learning algorithm.

HOW CAL WORKS

To better understand how to conduct a CAL review, Practical Law readers can access a free model CAL system supplied by the authors at *cormack.uwaterloo.ca/cal*. It contains as a dataset the recently released collection of 290,099 email messages from Jeb Bush's administration as governor of Florida.

As a running example, consider the task of finding "all communications relating to Apple Inc." within that dataset. The model CAL system suggests the document shown in Figure 1 as the most-likely relevant document to an initial query using the term "apple." Other CAL implementations might identify an initial batch of documents containing those shown in Figures 1 and 2, rather than a single document.

Figure 1: Document concerning "Apple Exec. addresses" identified first in response to a CAL search for "apple."



```
From: Dana, Pamella <DanaP@eog.state.fl.us>
Sent: Tuesday, February 8, 2000 9:13 AM
To: 'jeb@jeb.org'
Cc: Cales, Roy; Bradshaw, Sally
Subject: FW:

Here are the Apple Exec. addresses. I am working on getting the appropriate
email for Steve Jobs

-----Original Message-----
From: Cathy Hutchinson Foster [ mailto:cathyh@apple.com]
Sent: Monday, February 07, 2000 10:26 PM
To: Dana, Pamella
Cc: portman@apple.com
Subject: Re:

Pam:

I am glad Governor Bush, Roy, and you enjoyed your visit to Apple. I am
checking on the appropriate email for Steve and will let you know. The
```

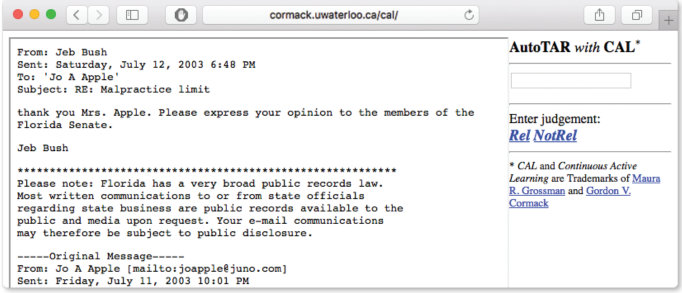
AutoTAR *with* CAL*

apple

Enter judgement:
Rel NotRel

* *CAL* and *Continuous Active Learning* are Trademarks of Maura R. Grossman and Gordon V. Cormack

Figure 2: Document from "Jo A Apple" identified in an early batch of responses to a CAL search for "apple."



```
From: Jeb Bush
Sent: Saturday, July 12, 2003 6:48 PM
To: 'Jo A Apple'
Subject: RE: Malpractice limit

thank you Mrs. Apple. Please express your opinion to the members of the
Florida Senate.

Jeb Bush

***********************************************************
Please note: Florida has a very broad public records law.
Most written communications to or from state officials
regarding state business are public records available to the
public and media upon request. Your e-mail communications
may therefore be subject to public disclosure.

-----Original Message-----
From: Jo A Apple [mailto:joapple@juno.com]
Sent: Friday, July 11, 2003 10:01 PM
```

AutoTAR *with* CAL*

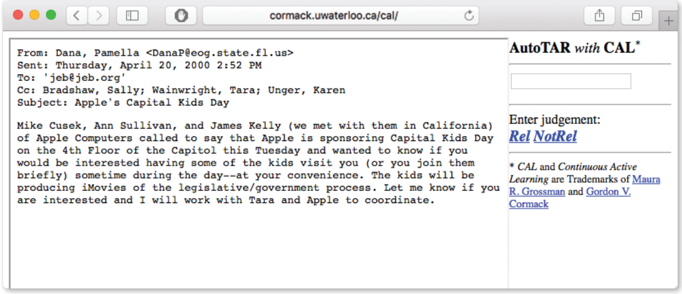
Enter judgement:
Rel NotRel

* *CAL* and *Continuous Active Learning* are Trademarks of Maura R. Grossman and Gordon V. Cormack

In this example, Figure 1 is relevant, while Figure 2 is not. Counsel relays this information to the CAL system, which employs a machine-learning algorithm to determine which characteristics of the two documents render Figure 1 relevant and Figure 2 not relevant.

In the model CAL system, this relevance feedback is provided by clicking either the link labeled "Rel" or the link labeled "NotRel" for each document. The learning algorithm might infer that a reference to "Steve Jobs" is evidence of relevance, while a reference to "Mrs. Apple" is evidence of non-relevance. Combining this and other evidence from the two documents and, more generally, from all of the documents reviewed thus far, the CAL system might determine the next most-likely relevant documents are those shown in Figures 3 and 4.

Figure 3: "Apple's Capital Kids Day" document.



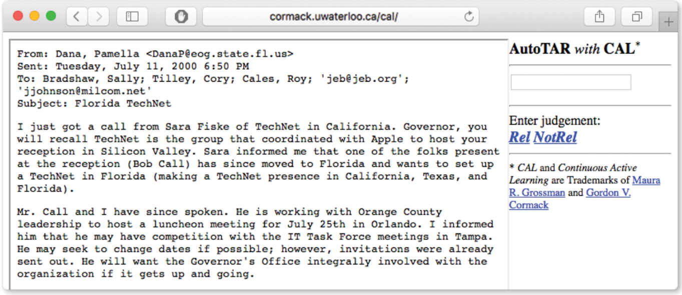
```
From: Dana, Pamella <DanaP@eog.state.fl.us>
Sent: Thursday, April 20, 2000 2:52 PM
To: 'jeb@jeb.org'
Cc: Bradshaw, Sally; Wainwright, Tara; Unger, Karen
Subject: Apple's Capital Kids Day

Mike Cusek, Ann Sullivan, and James Kelly (we met with them in California)
of Apple Computers called to say that Apple is sponsoring Capital Kids Day
on the 4th Floor of the Capitol this Tuesday and wanted to know if you
would be interested having some of the kids visit you (or you join them
briefly) sometime during the day--at your convenience. The kids will be
producing iMovies of the legislative/government process. Let me know if you
are interested and I will work with Tara and Apple to coordinate.
```

AutoTAR *with* CAL*

Enter judgement:
Rel NotRel

* *CAL* and *Continuous Active Learning* are Trademarks of Maura R. Grossman and Gordon V. Cormack

Figure 4: "Florida TechNet" document.



```
From: Dana, Pamella <DanaP@eog.state.fl.us>
Sent: Tuesday, July 11, 2000 6:50 PM
To: Bradshaw, Sally; Tilley, Cory; Cales, Roy; 'jeb@jeb.org';
'jjohnson@milcom.net'
Subject: Florida TechNet

I just got a call from Sara Fiske of TechNet in California. Governor, you
will recall TechNet is the group that coordinated with Apple to host your
reception in Silicon Valley. Sara informed me that one of the folks present
at the reception (Bob Call) has since moved to Florida and wants to set up
a TechNet in Florida (making a TechNet presence in California, Texas, and
Florida).

Mr. Call and I have since spoken. He is working with Orange County
leadership to host a luncheon meeting for July 25th in Orlando. I informed
him that he may have competition with the IT Task Force meetings in Tampa.
He may seek to change dates if possible; however, invitations were already
sent out. He will want the Governor's Office integrally involved with the
organization if it gets up and going.
```

AutoTAR *with* **CAL**[*]

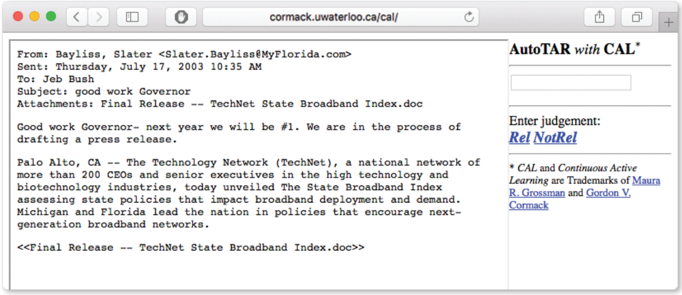
Enter judgement:
Rel ***NotRel***

[*] *CAL* and *Continuous Active Learning* are Trademarks of Maura R. Grossman and Gordon V. Cormack

Presumably, these two documents are relevant to the Apple Inc. search, and counsel would provide the CAL system with additional relevance feedback by clicking on the "Rel" links. Generally, the CAL system continues to present similar documents if they are judged to be of interest, learning suggestive terms like "iMovies" as indicia of potential relevance, along with less obvious cues like "Dana Pamella," "Capital Kids Day," and "TechNet."

However, "TechNet" turns out to be a poor choice for indicating relevance. The term eventually prompts the CAL system to suggest the document shown in Figure 5.

Figure 5: "TechNet State Broadband" document.



```
From: Bayliss, Slater <Slater.Bayliss@MyFlorida.com>
Sent: Thursday, July 17, 2003 10:35 AM
To: Jeb Bush
Subject: good work Governor
Attachments: Final Release -- TechNet State Broadband Index.doc

Good work Governor- next year we will be #1. We are in the process of
drafting a press release.

Palo Alto, CA -- The Technology Network (TechNet), a national network of
more than 200 CEOs and senior executives in the high technology and
biotechnology industries, today unveiled The State Broadband Index
assessing state policies that impact broadband deployment and demand.
Michigan and Florida lead the nation in policies that encourage next-
generation broadband networks.

<<Final Release -- TechNet State Broadband Index.doc>>
```

AutoTAR *with* **CAL**[*]

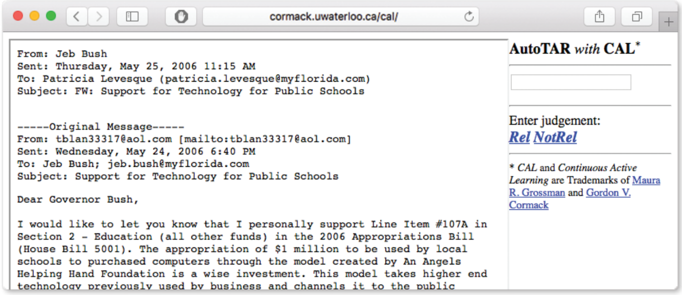
Enter judgement:
Rel ***NotRel***

[*] *CAL* and *Continuous Active Learning* are Trademarks of Maura R. Grossman and Gordon V. Cormack

By receiving a "NotRel" coding, CAL learns of its poor choice, discounts the term "TechNet," and avoids further presentation of similar documents.

After suggesting 70 relevant and 38 not relevant documents, the model CAL system presents the document shown in Figure 6, which concerns appropriations for technology in schools. Although Apple Inc. may have been involved in this project behind the scenes, there is no specific reference to the company, its products, or its personnel in this document. Counsel may tentatively deem this document to be not relevant, and conduct further research to confirm that decision. Alternatively, counsel might adopt an expansive interpretation of relevance and code the document as relevant, which would cause the CAL system to present additional documents concerning technology in schools, even without explicit references to Apple Inc.

Figure 6: "Support for Technology for Public Schools" document.



```
From: Jeb Bush
Sent: Thursday, May 25, 2006 11:15 AM
To: Patricia Levesque (patricia.levesque@myflorida.com)
Subject: FW: Support for Technology for Public Schools


-----Original Message-----
From: tblan33317@aol.com [mailto:tblan33317@aol.com]
Sent: Wednesday, May 24, 2006 6:40 PM
To: Jeb Bush; jeb.bush@myflorida.com
Subject: Support for Technology for Public Schools

Dear Governor Bush,

I would like to let you know that I personally support Line Item #107A in
Section 2 - Education (all other funds) in the 2006 Appropriations Bill
(House Bill 5001). The appropriation of $1 million to be used by local
schools to purchased computers through the model created by An Angels
Helping Hand Foundation is a wise investment. This model takes higher end
technology previously used by business and channels it to the public
```

AutoTAR *with* **CAL**[*]

Enter judgement:
Rel ***NotRel***

[*] *CAL* and *Continuous Active Learning* are Trademarks of Maura R. Grossman and Gordon V. Cormack

In this same way, the model CAL system will continue to explore the boundaries of relevance until it exhausts all avenues of potential relevance, and suggests increasingly fewer documents of interest. At this point, counsel can have reasonable confidence that they have seen substantially all of the relevant Apple Inc. documents in the dataset.

To increase counsel's confidence in the quality of the review, they might:

- Review an additional 100, 1,000, or even more documents.
- Experiment with additional search terms, such as "Steve Jobs," "iBook," or "Mac," and examine the most-likely relevant documents containing those terms.
- Invite the requesting party to suggest other keywords for counsel to apply.
- Review a sample of randomly selected documents to see if any other documents of interest are identified.

At some point, however, counsel must decide that they have undertaken reasonable efforts to identify the relevant documents.

COMPLETION OF A TAR REVIEW

A problem common to all TAR methods is determining whether or not the review is sufficient.

The CAL protocol has been compared to popping popcorn. After an initial warming period, the kernels begin to pop at a high rate, but the popping eventually slows down and nearly stops. At that point, it is a reasonable assumption that substantially all of the kernels have popped. If the kernels do not begin popping rapidly after a reasonable amount of time following placement of the bag in the microwave, or many unpopped kernels are discovered in the bag after the fact, it is likewise reasonable to assume that something has gone wrong with the popping process and remedial action is needed. However, the process is considered successful if only a small residual of unpopped kernels is left behind.

The same is true for relevant documents that are left behind following a TAR review. The two most critical issues are:

- How many relevant documents have been missed.
- Whether any of those relevant documents are novel or important.

Comparing CAL, SAL, and SPL Protocols for TAR

This table summarizes the three most common protocols employed by major TAR providers.

CAL PROTOCOL	SAL PROTOCOL	SPL PROTOCOL
STEP 1: Find one or more relevant documents by any means, or create a hypothetical relevant document, known as a synthetic document. **STEP 2:** Use a machine-learning algorithm to suggest the next most-likely relevant documents. **STEP 3:** Review the suggested documents and provide relevance feedback to the learning algorithm, indicating whether each suggested document is actually relevant or not. **STEP 4:** Repeat Steps 2 and 3 (and, optionally, Step 1) until very few, if any, of the suggested documents are relevant.	**STEP 1:** Remove a certain number of randomly selected documents from the collection (for example, 500 documents), and label them as the control set. **STEP 2:** Review and code the documents in the control set as relevant or not relevant. **STEP 3:** Repeat Steps 1 and 2 until the control set contains a sufficient number of relevant documents (for example, at least 70 documents). **STEP 4:** Without any knowledge of the control set, choose a seed set using random sampling or any other means. **STEP 5:** Review and code the documents in the seed set as relevant or not. **STEP 6:** Use a machine-learning algorithm to suggest documents from which the algorithm will learn the most. This typically consists of documents that are marginally likely to be relevant. **STEP 7:** Review and code the newly suggested documents and add them to the seed set. **STEP 8:** Repeat Steps 6 and 7 until the learning algorithm determines that stabilization has occurred, and that further training will not improve the algorithm. This is based on the accuracy of the algorithm's predictions of relevance for the documents in the control set. **STEP 9:** Use the learning algorithm to categorize or rank all documents in the collection. **STEP 10:** Review all documents categorized as relevant, or ranked above a predetermined cut-off score.	**STEP 1:** Choose a seed set using random sampling or any other means. **STEP 2:** Review and code the documents in the seed set as relevant or not relevant. **STEP 3:** Evaluate the effectiveness of the training so far, typically by counting the number of overturns, and the number of documents that could not be classified by the learning algorithm. **STEP 4:** If the result from Step 3 is deemed to be insufficient, repeat Steps 1, 2, and 3 with a larger seed set, until the effectiveness of the training is deemed sufficient. **STEP 5:** Use the learning algorithm to categorize or rank all documents in the collection. **STEP 6:** Review the documents categorized as relevant, or ranked above a predetermined cut-off score.

SAL and SPL protocols raise a compound problem that CAL does not: when to stop training the machine-learning algorithm, and how many documents to review once training is complete. To address this problem:

- SAL protocols typically use a randomly selected control set to track the progress of the review.
- SPL protocols typically use ad hoc sampling methods, such as counting the number of instances in which the learning algorithm and the reviewer disagree (overturns), as the basis for determining when to stop.

(See Box, *Comparing CAL, SAL, and SPL Protocols for TAR*.)

Determining whether a learning algorithm is adequately trained has generated considerable controversy. This issue often prompts requesting parties to demand either to participate in or monitor the TAR training process, or to compel the disclosure of both relevant and not relevant seed set or training documents. Both options tend to be unpalatable to responding parties.

Whether counsel starts out using CAL, SAL, or SPL, counsel should consider training the algorithm using every document that has been subject to human review, after the document has been reviewed. In a SAL or SPL process, a requesting party would be better served by having the responding party feed the results of the final review back into the learning algorithm to identify additional relevant documents, than by litigating the disclosure of the seed set, training documents, and stabilization criteria. This additional training effectively transforms a SAL or SPL protocol into a CAL protocol, thereby providing additional feedback to the learning algorithm.

MEASURES OF SUCCESS

After a TAR review, counsel may wish to determine how many relevant documents were missed in proportion to the total number of relevant documents in the collection. For example, if there were 300 documents concerning Apple Inc. in the Jeb Bush email dataset, and the CAL protocol failed to find 30 of them, the failure rate would be 10%, and the success rate would be 90%. In the information retrieval space, this success rate is known as recall.

Unfortunately, recall is difficult to measure. Indeed, if counsel knew which documents were missed, they would not have been missed in the first place. Where there are few relevant documents in a dataset, it is exceedingly difficult to find them through random sampling, as required for a statistical estimate. It is even more challenging to do so following a thorough review process. (See Maura R. Grossman & Gordon V. Cormack, *Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review,"* 7 Fed. Cts. L. Rev. 285, 293, 300-12 (2014) (discussing various TAR validation methods).)

Moreover, reasonable minds can differ on the relevance or non-relevance of specific missed documents, particularly on the margins. If, during the validation process, counsel identifies some missed documents, and achieves consensus that they are indeed relevant, counsel can resume the CAL review using the newly identified documents for additional relevance feedback training and thus improve the recall of the review.

Recall may be estimated through the following steps:

- Count the number of relevant documents found by the CAL system.
- Estimate the number of relevant documents missed by the CAL system by drawing a random sample of the documents not selected for review, commonly known as the null set or the discards.
- Calculate an estimate of the total number of relevant documents, which is the number of relevant documents found plus the estimate of the number missed.
- Estimate recall by dividing the number of relevant documents found by the estimate of the total number of relevant documents.

For example, if the CAL system identified 270 relevant documents and, through sampling, counsel estimates that 30 have been missed, counsel can estimate that there are 300 relevant documents in total and that recall is 270 out of 300, or 90%.

Counsel should beware of using measures that fail to account for both the number of relevant documents found and the number that were missed. Measures counsel should avoid include:

- **Accuracy.** This refers to the proportion of all documents that are correctly classified as either relevant or not relevant.
- **Elusion.** This refers to the proportion of documents in the null set that are in fact responsive.
- **Overturns.** This refers to the number of documents in a sample that are incorrectly classified by the TAR algorithm and are corrected by a manual reviewer.

These metrics are uninformative because it is possible to achieve apparent success while failing to identify meaningful numbers of relevant documents. Using the Apple Inc. example, counsel could achieve accuracy of 99.9%, elusion of 0.1%, and zero overturns — seemingly stellar results — while failing to find a single relevant document.

TRY CAL AT HOME

There is no better way to learn CAL than to use it. Counsel may use the online model CAL system to see how quickly and easily CAL can learn what is of interest to them in the Jeb Bush email dataset. As an alternative to throwing up their hands over seed sets, control sets, F_1 measures, stabilization, and overturns, counsel should consider using their preferred TAR tool in CAL mode on their next matter.

The author, Maura R. Grossman, served as a court-appointed special master in the Rio Tinto *case referenced in this article.*

The views expressed in this article are those of the authors and should not be attributed to Wachtell, Lipton, Rosen & Katz or its clients.