The AccessData Approach to Technology Assisted Review
Executive Summary

No controversy remains regarding the effectiveness of Technology-Assisted Review (“TAR”). When executed as a complement to human review, research and court holdings will support any user’s decision to engage automated review technologies in the culling, organization, review and production of responsive documents. By analyzing the coding choices made by subject-expert reviewers, a TAR-enabled review platform will write rules to determine its own coding choices. It will, as it tests its own choices against the representative sample, report its accuracy rates to the human review team. The team will then decide, ideally in conversation with opposing counsel, what accuracy rate is acceptable. The choices made by the TAR apparatus will, at this stage, be limited only to those choices that merit automatic review: questions of responsiveness, for instance. These choices, however limited, constitute a massive portion of the work currently done by human reviewers. As multiple studies have shown, this process not only decreases time and cost of review but increases overall accuracy.

This paper will begin by briefly providing two definitions of this particular form of TAR. Then, a hypothetical review process in the absence of TAR will be described, outlining the current business problem. Following, the technology supporting AD Summation 5’s TAR feature will be explained. Finally, a solution will be proposed in the form of a workflow suggested by the Summation TAR.

Defining Technology Assisted Review

Grossman and Cormack, in the first glossary dedicated only to terminology relating to TAR, define predictive coding as:

An industry-specific term generally used to describe a Technology-Assisted Review process involving the use of a Machine Learning Algorithm to distinguish Relevant from Non-Relevant Documents, based on Subject Matter Expert(s)’ Coding of a Training Set of Documents.

While all TAR applications will share these characteristics, each will approach the task in its own way. The “Machine Learning Algorithm” will differ, necessarily, between products. Also, the technology capabilities will not be limited to relevancy determinations. It should be noted, that – with the great interest shown in TAR— many differing tools are being offered by many service and technology providers who use a shared vocabulary but not necessarily a shared glossary.

With this understanding in mind, a second definition may also prove useful. Chuck Rothman of eDiscovery Journal defines predictive coding as “a process whereby a definition, made up of various rules, is created. Records in a collection are then evaluated to determine how well they match the definition.” This defines a process, rather than a technology. Taking TAR as a process, assisted by technology, reminds us that human engagement remains essential. This being so, each provider owes its clients a duty to explain, on a functional level, how its proprietary technology operates and with which workflows. Armed with this understanding, each client organization can then customize and develop its own workflows and realize the maximum return on their product investment.

The primary purpose of TAR is to assist users by automating the culling, coding and categorization of large document collections, thus speeding up the review process. Beyond this assistance, a properly trained application will also provide improved review consistency and, in many instances, increased overall accuracy.

Traditional Document Review v. TAR

Traditionally, when incoming evidence exits a processing workflow, it brings very little value with it other than its potential for review. In order to realize this potential value and after the enormous expense already accrued in the collection, sequestration, conversion, decryption and promotion to review, the document must undergo at least a first-pass review. This review typically begins with a request from an attorney to a supervising paralegal and a team of non-specialist paralegals, contract attorneys or junior associates. Ideally, this collection will already have undergone a mostly automated filtration process (often referred to as “Early Case Assessment” (“ECA”)) which has attempted to identify and to cull the raw data of duplicates, corrupted files, unsupported file types and system files. However, particularly with large...
case files with hundreds of thousands of documents exchanged, ECA is only one preliminary step towards the first-pass review. After ECA has been performed, the review team supervisor parses out the documents in equal numbers to the team. Each reviewer then, under considerable time pressure and with considerable overtime hours billed, reviews hundreds if not thousands of documents each day and "codes" each one individually according to a predetermined categorization scheme. In essence, these first-pass reviewers are not dissimilar to a TAR engine. They are, however, much more human and vulnerable to subjective inconsistencies and concentration fatigue. And, as document productions continue to increase in size, these reviewers are likewise increasingly over-worked and under-stimulated as their projects stretch further out into the indefinite future. Meanwhile, each reviewer bills an hourly rate to the client and represents ongoing overhead for the attorney. The end result is a process that perpetually increases expenses while concurrently decreasing in efficacy.

TAR provides an iterative workflow in which high-level reviewers examine and code a representative portion of the documents, thereby “teaching” the system how to code other documents. The system then takes documents coded by the user, applies similar coding to documents with similar content, and waits for the reviewer to perform quality assessment on its work. Again, this is not dissimilar, in function, to the work of a traditional review team.

TAR does differ from traditional review, however, by automating a process that has always been better suited to automation, resulting in welcome savings of both time and money. Grossman and Cormack's analysis of the TREC Legal 2009 Interactive Task Test Collection, compares the results of a team of attorneys and law students, managed by a document review vendor, with the results of two different TAR engines. Their quantitative results demonstrate that both TAR technologies achieved greater recall and precision than manual review. Moreover, the qualitative analysis finds that the human reviewers even mischaracterize documents that humans would presumably be better suited to answer; for instance, concepts relying on context for meaning. These results unequivocally demonstrate that “[t]echnology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort.”

Once expert reviewers train the TAR engine with a sufficient and representative sample of documents, the process can work as fast as its hardware will allow and achieve a consistent degree of accuracy across its review. Humans, on the contrary, will make more errors as they respond to the stressful demand to work faster. They will necessarily pay less attention to the details before them. The TAR engine does not experience this pressure and, therefore, saves time while increasing accuracy. This, as a result, saves money by both reducing the number of reviewers required (along with related overhead) and reducing the amount of time for which those reviewers are billing.

Summation 5's proprietary TAR function has been designed with precisely this objective.

How Summation 5’s TAR Feature Works

As discussed above, all TAR engines share characteristics associated with supervised learning. The process basically is: (1) a “seed set” of documents, representative of the larger collection, undergoes review by one or more subject-matter experts, (2) the engine “observes” the choices made by the team and, through an iterative process, uses its algorithm to create a decision tree for itself,
(3) it reviews and categorizes the seed set itself in parallel, and (4) reports its accuracy rate as a "confidence score". This process is repeated as often as necessary until the confidence score meets the review team's approval. Of course, while this is the general approach, each design will differ in varying degrees on the detail.

**AccessData has developed its own TAR feature in Summation. Here is how it works**:10

1) **How to Instruct the System**

First, in order for the system to learn the parameters of an assignment, a representative "seed set" of documents must be defined by the reviewer. By using Summation's many filters, facets and advanced search features, the reviewer identifies and marks an appropriate subset of documents. AccessData suggests that roughly 10% of the document collection should be included in this set. Once this set has been created, the team should review and code it while the "Predictive Coding Layout" is in effect. As of this release, Summation 5's TAR focuses solely on "Responsive" determinations. The reviewer codes each document as in traditional manual review, marking whether the document is "Responsive" or "Non-Responsive" and entering keywords important to the choice. This process is repeated until the "seed set" is complete.

2) **How the System Learns**

As the reviewer works in the "Predictive Coding Layout", the system records and analyzes the decisions made. The Iterative Dichotomiser 3 (ID3) algorithm, a trusted and well-tested method for creating a decision tree based on a collection of attributes identified by its user, serves as a starting point for AccessData's TAR engine.11,12 Summation 5 utilizes ID3 to create a unique decision tree, correlating the seed set reviewers' input with input attributes generated during processing by AccessData's proprietary content-based clustering feature.13 The resulting decision tree provides the framework by which Summation will execute its TAR.

While the system default processes all document attributes in the set, the user can refine the list of input attributes using feature selection algorithms. The more technical user can further refine the system by altering the configuration file.

3) **How the System Executes**

The system begins to execute by randomly partitioning the manually-coded set into two groups, a validation set and a training set. The system then performs its decision tree analysis on the training set, cross-validating its results with the decisions made by the reviewers. This process is repeated multiple times, using different subsets of the validation set to test the results. The results from the tests are averaged to produce a single estimation of the results.

4) **How the System Determines its Confidence Level**

Finally, the system provides a Confidence Level score, defined using the statistical calculation known as the $F_1$ score.14 $F_1$ is calculated using precision rate (true positive count over total positive labeled) and recall rate (true positive count over total positive count). This confidence score gauges the system's readiness to perform analysis on the remaining documents.

10 All actions taken in this section are taken through the Summation 5 user interface.
12 http://en.wikipedia.org/wiki/ID3_algorithm
13 This feature is an option used by the evidence processing portion of AccessData product(s) to logically group objects together using only the text extracted from those objects.
Workflow

In Summation, the TAR workflow has three basic phases: teach the system, apply the system's learning, and perform quality control. Reports generated during the process provide information about system performance and ensure that the process remains iterative.

1) **Teach the System**

As described above, the first phase requires training the system to review this unique universe of documents. The reviewer begins by selecting a set of documents representative of the entire document universe. Once the training set is identified, the reviewer evaluates each document individually. This process leverages Summation's coding and tagging process while using the pre-defined "Predictive Coding Layout", discussed above. The primary function of the reviewer here is to determine responsiveness to the pre-determined subject matter. After reviewing a subset of documents, the reviewer may test the system to obtain a confidence level of how well the system has learned. The confidence level is generated and displayed by the "Predictive Coding Confidence" layout panel.

2) **Apply the System Learning**

Once the reviewer reaches an acceptable "Confidence Level", he or she defines a set of documents for the system to code. Again using any available filter and search features, the user marks documents for predictive review. The process can be initiated on all, checked or unchecked items in the document list. If documents included in this set that were already manually coded and used to train the system, they will not be re-coded by the system. Once the TAR process is complete, a report is generated detailing results of the process. This report can be presented to either the judge or opposing counsel if the process should be challenged. The system reports the number of documents in the learning set, the number coded by the system, a breakdown of documents coded as responsive or non-responsive and information about the confidence level at the beginning of the TAR process.

3) **Perform Quality Control**

Finally, the reviewer can perform quality control on the TAR-coded documents by applying a filter to list only those documents. Performing a manual review of a statistical sample, the reviewer evaluates these documents to verify that the code applied was correct. If the code is incorrect, the reviewer can override the system code. The reviewer would repeat this process of reviewing until they are satisfied the results are sufficient or determine that enough codes have been overridden that the system needs to be re-taught and re-applied.
Conclusion

A TAR-enabled review platform can efficiently analyze coding choices made by subject-expert reviewers and convert its analysis into a powerful decision-tree, able to conduct analysis on its own. As this is an iterative process, the platform will report evolving accuracy rates to its users so that any necessary adjustments can be made to its training. While the choices made by the Summation 5 TAR apparatus will, at this stage, be limited only to the question of responsiveness, future development will allow the system to perform other checks including privilege review, issue relevancy and other choices that are essentially binary in nature and answer the question: is it or is it not? Traditional review has required humans to make these limited choices according to prescribed rules, not choices that would require – or benefit from – the human ability to interpret or describe complex processes. These choices are routine and call for automation. They remain, when tied to traditional solutions, sources of great expense as reviewers continue to face exponentially increasing numbers of documents. Those who utilize the new generation of solutions TAR provides, on the other hand, will realize a measured and significant return on their investment, both in time and money.