Factors to Consider When Deciding Whether to Use TAR

A. Data Types and TAR

When considering whether to use TAR, one must consider the data set one wants to analyze before deciding whether and how to proceed. While TAR can be very effective at sorting through a custodian' email, TAR may not be an effective tool for organizing or categorizing the spreadsheets or videos attached to those same emails. Though the various algorithms used by each TAR platform differ, all such tools are at least somewhat dependent on the semantic content of the document population being analyzed. It's important that users consider this content and their goals in order to determine whether TAR is a reasonable workflow and, congruently, to what extent TAR actually will help reduce the costs of linear review.

Generally speaking, TAR is most effective when applied to high-content, user-generated documents. This includes emails, electronic documents such as Word files and searchable PDFs, high-content presentation slides, etc. Such documents have sufficient semantic content for the TAR algorithm to effectively analyze each document's terms and finding semantic patterns it can apply to other documents in the population. In contrast, extremely short, low-content documents such as an email that says nothing more than "Please see attached," lack sufficient content and cannot be effectively analyzed. These documents generally make more training examples and/or may not be correctly categorized, if they are categorized at all. Such materials are often excluded from TAR project automatically and are almost always excluded from training populations.

In addition, it is generally preferred that documents submitted to TAR be user-created, meaning they are considered "unstructured data," as opposed to formalized tables or charts of structured data extracted from a database such as Excel. Even when such spreadsheets are language-based (contra numeric), there generally is little or no relationship between words across multiple cells, meaning such documents have little semantic content for TAR to analyze. In contrast, user-created emails and documents, where language is used naturally and words and sentences relate and build on one another, are more readily analyzed by Natural Language Processing and similar TAR algorithms.

Additional data types one should consider/evaluate before using TAR include:

- Spreadsheets and similar exports from structured databases, particularly those with little semantic or user-generated language content;
- Outlook Calendar Invitations unless they include extensive semantic content in the body of the invite;
- Hard Copy documents with less-than-perfect OCR results;
- Audio/video files generally lack any semantic content;
- Foreign language/ESL documents can be analyzed by TAR, but generally require separate training sets for each language (N.B., mixed-language documents may cause additional issues); and/or
- Mobile Data/Chat/MSM/Slack/Social Media/IoT/Real Big Data.

These data types are often set aside or dealt with concurrently while applying TAR to the appropriate population, however as one evaluates whether to use TAR, one should consider what percentage of the target population might be excluded due to the above factors.

B. Data Volume

As discussed, predictive coding lends itself to document populations that contain emails, MS Word or Power Point files, PDF files and text-heavy MS excel files as the population which is subject to TAR analysis. Documents with little conceptual content, such as multi-media files, images, java scripts, system files and encrypted documents, should be culled and removed from the TAR process and analyzed separately, by attorneys, for responsiveness and production.

With ongoing data collection, the introduction of additional documents will not require starting the process from the beginning but it will be necessary to train additional documents to update the predictive model until all documents have been processed to ensure there are representative documents from the newly processed data in the training documents. Additional documents should be trained until you have a final static population.

Conceptual analysis is used to rank documents based on their likelihood of being responsive. The variables associated with each document (such as combinations of words, document features, concepts and metadata values) predicts a binary output variable (responsive or non-responsive). The variables identified are evaluated along with coding determinations of the training set, to ascertain commonalities and concepts that distinguish responsive documents from non-responsive documents. Based on these variables, an equation is developed that allows the classifier to map concepts to a resulting responsive or non-responsive designation. The classifier also applies optimal weights for the features based on the coding decisions associated with each document.

Every document in the Predictive Coding Population is assigned a responsive/non-responsive probability score based on the combination of features found in each document.

Utilizing a Responsive Prioritization workflow, the responsive probability scoring is used to batch documents from highest to lowest likelihood of responsiveness, so that responsive documents are front loaded for review. As the documents that are predicted responsive are reviewed, they are periodically incorporated into the model, at which point, the model recalibrates the scoring based on the additional input of information with more coded documents. The scoring is refined further by the model as more documents are reviewed. The review continues until the client and their counsel determine the percentage of responsive documents has dropped to an acceptable level or all documents are reviewed.

C. Timing and the TAR Workflow

Reviewing documents predicted highly responsive can start as early as the first week of training to meet early production deadlines.

While the volume of data and deadlines are key factors in determining a project's timeline and staffing, the increasing complexity of many projects requires managing the project with forethought to completing various workflows. If the final review population is unknown as a result of ongoing collections and/or processing, an estimation of additional data that will be included in the review population is necessary.

Under a traditional predictive coding TAR 1.0 approach, the machine learning process must be factored in when determining the length of time it will take to complete the review. The start of a full substantive review can be delayed for a week or longer waiting for the machine training to reach stability. To avoid a delay in commencing review, any segment of the population that will require review can be started while stability is achieved. Documents that should be identified for review are documents, as described above, with little or no conceptual content. The TAR 2.0 "Continuous Active Learning (CAL)" approach prioritizes documents for review based on a relevancy ranking. The initial prioritization can be commenced based on documents counsel identifies prior to the review or learns of during an ICA. If counsel has not identified any relevant or key documents to assist in the prioritization, this can be accomplished by review of a sample set of documents which will act as the seed set to create the initial predictive rankings. Both TAR 1.0 and 2.0 have timing considerations that must be factored in to the project timeline.

Additional workflows that should be considered include second level review by counsel, foreign language review, redactions (including whether the redactions are tiff or native redactions) and the delivery of a privilege log. It is also important to understand the vendor's production turnaround time, from approval of a production submission to the time the production is available for delivery, and account for this time in the project schedule. The size of the data being produced will have implications on how the production is received by counsel (e.g. hard drive or FTP) and may affect when the production must be submitted. Creating a project schedule from the outset is the pathway to a successful project as it focuses attention on potential workflows and the establishment of deadlines which provides clarity and sets expectations from the start of the review.

D. TAR Expertise [Forthcoming]

- Ethical and knowledge requirements/competency?
- Ability to have someone who can understand and evaluate
- Vendor analysis and product offerings
- Ability to understand options, workflows, capabilities
- Project management and ability to use the tools

E. The Forum & Your Adversary

In calculating whether to use TAR, you should consider the forum and your adversary. As an initial matter, you may need to obtain prior approval or cooperation to use machine learning. You should consider the likelihood that the requesting party or agency or the court will agree with your approach, and how much time and resources you may have to spend in obtaining an agreement to use TAR. Engaging in a protracted battle with opposing counsel, spending time educating your adversary about TAR, or involving the court may not be worth the cost savings otherwise afforded by TAR.

Another forum-related problem is that some courts and practitioners demand that the producing party disclose not only its assisted review process but also its full training sets, including both responsive and non-responsive documents. This requirement can be especially onerous where the non-responsive documents contain material that must be redacted, usually at an hourly rate. For this reason, consideration of the forum's approach to disclosure about these processes must be included in calculating the overall risks and costs associated with TAR and whether other measures of the effectiveness of TAR (e.g., recall) will satisfy the court. As courts and practitioners become more familiar with statistically evaluating TAR, this concern should diminish.

When responding to government subpoenas and requests for information, carefully consider the agency's requirements and policies – and the reasons behind those policies – before embarking on any type of machine learning to limit production. The Antitrust Division of the Department of Justice, for example, requires prior approval of not only the format but also the method of production. "Before using software or technology (including search terms, predictive coding, de-duplication, or similar technologies) to identify or eliminate documents, data, or information potentially responsive," a Company responding to the request must submit a written disclosure of the process, including details regarding the particulars of the proposed process.

G. The Cost of TAR vs. Traditional Linear Review

TAR is generally considered to be a faster and cheaper process than a traditional linear document review, at least in the context of litigation. There are a number of factors that impact the relative costs of a TAR project however, and one's goals, workflow, and time line are all factors that must be considered when deciding whether to use TAR.

Document review is often considered to be the single largest expense of litigation-related discovery – frequently estimated at 60 to 70% of the total cost. Clients obviously welcome any tools or methods that can reduce this line item, but they should be aware of the initial set-up and training expenses that come with a TAR workflow, as well as the relative uncertainty about the outcome and timeline involved in the process. In addition, rarely does TAR eliminate the need for any document review, and users should not be surprised when TAR shifts the ratio of first level to second level review towards the more-expensive second-level/QC/Privilege review side of the equation.

Unsurprisingly, cost-savings resulting from the use of TAR vary considerably from case-to-case. Factors discussed elsewhere in this primer, such as data quality, the types of data included in the population, the breadth or complexity of responsiveness, the richness of the data being analyzed, and the statistical thresholds agreed to with opposing counsel, as well as costs associated with the service provider or software (this is frequently a per document or per gigabyte fee); hourly consulting and/or project management fees related to the TAR process (including possible Expert fees related to oral or written testimony) all affect the overall cost of the TAR project and the cost savings realized in comparison to a traditional linear review.

In addition, clients should be mindful of additional costs resulting from use the Subject Matter Experts required to train TAR, the cost of TAR-related negotiations with opposing counsel, and any QC or validation review that might not otherwise be part of a traditional linear review

TAR is also not an over-night process and clients and courts should be aware that, unlike traditional linear review, a TAR review may involve an initial lag of several days or weeks as the system is set up and trained. For example, the availability and experience of Subject Matter Experts and the sophistication/automation of the tool being used can both result in training taking days or even weeks. While some linear review can often begin in parallel, clients need to understand that kicking off a TAR process often causes delays on the front-end of a review that are not typical of a traditional linear review where reviewers can start going through emails as soon as they are trained in the tool and the matter.

Reducing Costs through Data Re-Use

The costs of training TAR can often be significantly reduced by re-using previously reviewed and coded materials. Known-responsive documents and files can often serve as seed sets for training purposes, allowing one to both jump-start the TAR categorization while also reviewing fewer documents. However, even in these situations, some care needs to be given to determining whether any individual document is a good example for TAR training. Even in tools that automatically filter out large or complicated examples from the seed set, care must often be taken to identify wholly responsive text from documents with mixed responsive/non-responsive content. In addition, it is important in these situations to ensure that the responsiveness standard used in the initial review is the same as in the TAR project. For example, if the definition of responsiveness has since expanded, documents previously coded as Non Responsive may actually be Responsive at the time of review.

However, even in situations where all previously reviewed seed documents need to be re-reviewed, this can be an efficient place to start the training process as it frequently generates a highly rich set of documents for initial training/categorization, thereby quickly identifying more responsive documents for review and production.

The Cost of TAR vs. Proportionality

As of this writing, courts have yet to fully address the question of whether and how the use of analytics and TAR may impact proportionality. It is foreseeable however that a document request that would be considered unduly burdensome if the respondent were to review a large population linearly, will be considered reasonable if the respondent uses some form of analytics and TAR.

Even though the responding/producing party is generally considered to be best-positioned to evaluate the best way to identify and produce requested materials, courts may one day consider whether that party is acting in good faith in resisting the use of TAR as a means of reducing the anticipated burden of a particular request.

Suggested Metrics for Evaluating the Cost of TAR

As discussed above, clients may want to consider the following line items in their evaluation of TAR:

- Cost of TAR Tool/Index
- Savings from anticipated data reduction
  - Data reduction from non-TAR analytics, including email threading, de-duplication, near-duplicate identification, structured data analytics, etc.
  - Estimated data reduction due to TAR alone
- The potential cost of negotiating and possibly defending a TAR workflow
- Initial TAR Investment/Set Up
- Cost of TAR training/Subject Matter Experts
- Additional TAR-related consulting fees
- The cost of any additional QC/Validation review not otherwise contemplated under a traditional linear review

H. Third Party Data vs. Client Data?

- Initial knowledge/understanding of data set may affect TAR decision.

- Quality of Third Party Data: Is Third Party Data highly- responsive vs a data dump? Is Third Party Data complete with all meta-data?
- Case Strategy- How much do you actually need to know about the different data sets?

TAR may be an effective tool for the review and analysis of data collected from other parties to a litigation or investigation. In such situations, data quality is a major consideration, but TAR may help to quickly identify highly-relevant documents for motion practice or deposition

J. Data Testing and Sampling

Sampling helps you assess whether the data you have collected will work with the review tools at your disposal. Prior to committing to using TAR, you should sample your dataset to determine whether the documents contain the type of data suitable for the particular assisted review tool you are using and to calculate the ratio of responsive to non-responsive documents ("richness"). For example, if your documents are comprised mostly of numbers or contain very little text, you may not want to use latent syntax indexing. If you have a very large volume of data of low richness, you may want to test responsiveness rates returned through search terms or other culling methods. This process may be effective enough that you and your adversary agree that it is sufficient for the purposes of the responses. Use of search terms may enable you to produce documents sooner and more efficiently than possible through an iterative TAR process. Alternatively, use of search terms may limit the dataset to a size that your machine can process. Although some oppose limiting a dataset before using TAR, pre-TAR culling may nonetheless be reasonable and desirable under certain circumstances.

K. Data Source Considerations [Forthcoming]

- Custodian/Departmental Distinctions
- Disparate Data Sources
- Email vs. Network vs. Social Media vs. ____

L. TAR Project Goals [Forthcoming]

- Defensible litigation production
- Investigative
- Find the story in the data/Narrative
- Completeness/Deficiency Testing
- Best/Hot Docs
- Deposition Preparation