TECHNOLOGY ASSISTED REVIEW (TAR) GUIDELINES

January 2019

BOLCH
JUDICIAL INSTITUTE

DUKE LAW

FOREWORD†

In December 2016, more than 25 EDRM/Duke Law members volunteered to develop and draft guidelines providing guidance to the bench and bar on the use of technology assisted review (TAR). Three drafting teams were formed and immediately began work. The teams gave a progress status report and discussed the scope of the project at the annual EDRM May 16-17, 2017, workshop, held at the Duke University campus in Durham, N.C. The number of team volunteers swelled to more than 50.

The augmented three teams continued to refine the draft during the summer of 2017 and presented their work at a Duke Distinguished Lawyers' conference, held on September 7-8, 2017, in Arlington, Virginia. The conference brought together 15 federal judges and 75-100 practitioners and experts to develop separate "best practices" to accompany the TAR Guidelines. An initial draft of the best practices is expected in summer 2019. While the EDRM/Duke "TAR Guidelines" are intended to explain the TAR process, the "best practices" are intended to provide a protocol on whether and under what conditions TAR should be used. Together, the documents provide a strong record and roadmap for the bench and bar, which explain and support the use of TAR in appropriate cases.

The draft TAR Guidelines were revised in light of the discussions at the September 2017 TAR Conference, which highlighted several overriding bench and bar concerns as well as shed light on new issues about TAR. The Guidelines are the culmination of a process that began in December 2016. Although Duke Law retained editorial control, this iterative drafting process provided multiple opportunities for the volunteers on the three teams to confer, suggest edits, and comment on the Guidelines. Substantial revisions were made throughout the process. Many compromises, affecting matters on which the 50 volunteer contributors hold passionate views, were also reached. But the Guidelines should not be viewed as representing unanimous agreement, and individual volunteer contributors may not necessarily agree with every recommendation.

James Waldron
Director, EDRM

John Rabiej, Deputy Director
Bolch Judicial Institute

ACKNOWLEDGEMENTS

The *Technology Assisted Review (TAR) Guidelines* is the work product of more than 50 experienced practitioners and experts, who devoted substantial time and effort to improve the law. Three of them assumed greater responsibility as team leaders, including:

TEAM LEADERS

Mike Quartararo eDPM Advisory Services	Matt Poplawski Winston & Strawn	Adam Strayer Paul, Weiss, Rifkind, Wharton & Garrison

The following practitioners and ediscovery experts helped draft particular sections of the text:

CONTRIBUTORS

Kelly Atherton NightOwl Discovery	Doug Austin CloudNine	Ben Barnett Dechert	Lilith Bat-Leah BlueStar
Chris Bojar Barack Ferrazzano	Michelle Briggs Goodwin Procter	Jennifer Miranda Clamme Keesal Young & Logan	David Cohen Reed Smith
Xavier Diokno Consilio	Tara Emory Driven Inc.	Brian Flatley Ellis & Winters	Paul Gettmann Ayfie, Inc.
David Greetham RICOH USA, Inc.	Robert Keeling Sidley Austin	Deborah Ketchmark Consilio	Jonathan Kiang Epiq
John Koss Mintz Levin	Jon Lavinder Epiq	Brandon Mack Epiq	Rachi Messing Microsoft
Michael Minnick Brooks Pierce	Connie Morales Capital Digital and Califorensics	Lynne Nadeau-Wahlquist Trial Assets	Tim Opsitnick TCDI
Constantine Pappas Relativity	Chris Paskach The Claro Group	Donald Ramsey Stinson Leonard Street	Niloy Ray Littler Mendelson
Philip Richards Consilio	Bob Rohlf Exterro	Herbert Roitblat Mimecast	John Rosenthal Winston & Strawn
Justin Scranton Consilio	Dharmesh Shingala Knovos	Michael Shortnacy King & Spalding LLP	Mimi Singh Evolver Inc.

CONTRIBUTORS (CONT.)

Clara Skorstad Kilpatrick Townsend	Harsh Sutaria	Hope Swancy-Haslam Stroz Friedberg	Tiana Van Dyk Burnet Duckworth & Palmer
Patricia Wallace Murphy & McGonigle	Ian Wilson Servient	Carolyn Young Consilio	

We thank Patrick Bradley, Leah Brenner, Matthew Eible, and Calypso Taylor the four Duke Law, Bolch Judicial Institute Student Fellows, who proofread, cite checked, and provided valuable comments and criticisms. George Socha, co-founder of EDRM and BDO managing director, provided rock-solid advice and overview guidance.

In particular, we gratefully acknowledge the innumerable hours that Matt Poplawski, Tim Opsitnick, Mike Quartararo, and James Francis, Distinguished Lecturer at the City University of New York School of Law and former United States Magistrate Judge, devoted to reviewing every public comment submitted on the draft. Their dedication markedly improved the document's clarity and precision. We also want to single out the helpful and extensive suggestions made during the public-comment period by Tara Emory, Xavier Diokno, Bill Dimm, and Trena Patton. We are indebted to them for their input, along with comments and suggestions submitted by many others during the public-comment period.

The feedback of the judiciary has been invaluable in exploring the challenges faced by judges and the viability of the proposed guidelines. The ways in which these guidelines have benefitted from the candid assessment of the judiciary cannot be understated. It is with the greatest of thanks that we recognize the contributions of the 14 judges, who attended the conference and the six judges who reviewed early drafts and provided comments and suggestions.

EDRM/Duke Law School
January 2019

PREFACE

Artificial Intelligence (AI) is quickly revolutionizing the practice of law. AI promises to offer the legal profession new tools to increase the efficiency and effectiveness of a variety of practices. A machine learning process known as technology assisted review (TAR) is an early iteration of AI for the legal profession.

TAR is redefining the way electronically stored information (ESI) is reviewed. Machine learning processes like TAR have been used to assist decision-making in commercial industries since at least the 1960s leading to efficiencies and cost savings in healthcare, finance, marketing, and other industries. Now, the legal community is also embracing machine learning, via TAR, to automatically classify large volumes of documents in discovery. These guidelines will provide guidance on the key principles of the TAR process. Although these guidelines focus specifically on TAR, they are written with the intent that, as technology continues to change, the general principles underlying the guidelines will also apply to future iterations of AI beyond the TAR process.

TAR is similar conceptually to a fully human-based document review; the computer just takes the place of much of the human-review work force in conducting the document review. As a practical matter, in many document reviews, the computer is faster, more consistent, and more cost effective in finding relevant documents than human review alone. Moreover, a TAR review can generally perform as well as that of a human review, provided that there is a reasonable and defensible workflow. Similar to a fully human-based review where subject-matter attorneys train a human-review team to make relevancy decisions, the TAR review involves human reviewers training a computer, such that the computer's decisions are just as accurate and reliable as those of the trainers.

Notably, Rule 1 of the Federal Rules of Civil Procedure calls on courts and litigants "to secure the just, speedy, and inexpensive determination of every action and proceeding." According to a 2012 Rand Corporation report, 73% of the cost associated with discovery is spent on review.

The potential for significant savings in time and cost, without sacrificing quality, is what makes TAR most useful. Document-review teams can work more efficiently because TAR can identify relevant documents faster than human review and can reduce time wasted reviewing nonrelevant documents.

Moreover, the standard in discovery is reasonableness, not perfection. Traditional linear or manual review, in which teams of lawyers billing clients review boxes of paper or countless online documents, is an inefficient method. Problems with high cost, exorbitant time to complete review, fatigue, human error, disparate attorney views regarding document substance, and even gamesmanship are all

associated with manual document review. Studies have shown a rate of discrepancy as high as 50% among reviewers who identify relevant documents by linear review. The TAR process is also imperfect, and although no one study is definitive, research suggests that, in some contexts, TAR can be at least as effective as human review. Indeed, judges have accepted the use of TAR as a reasonable method of review, and importantly, no reported court decision has found the use of TAR invalid.[1]

The most prominent law firms in the world, on both the plaintiff and the defense side of the bar, are using TAR. Several large government agencies, including the DOJ, SEC, and IRS, have recognized the utility and value of TAR when dealing with large document collections. But in order for TAR to be more widely used and accepted in discovery, the bench and bar must become more familiar with it, and certain standards of validity and reliability should be considered to ensure its accuracy. These guidelines will not only demonstrate the validity and reliability of TAR but will also demystify the process.

The TAR GUIDELINES reflect the considered views and consensus of the participants. They may not necessarily reflect the official position of Duke Law School or the Bolch Judicial Institute as an entity or of Duke Law's faculty or any other organization, including the Judicial Conference of the United States.

One final note. There are several different variations of TAR software in the marketplace. TAR 1.0 and TAR 2.0 are the two most commonly marketed versions. Although one or the other version may be more prevalent, both continue to be widely used. These Guidelines are intended to provide guidance to all users of TAR and apply across the different variations of TAR. These Guidelines assiduously take no position on which variation is more effective, which may depend on various factors, including the size and richness of the TAR data population.

[1] As a further example of its reasonableness and legitimacy as a review process, the committee note to F. R. Evid. 502, states that "Depending on the circumstances, a party that uses advanced analytical software and linguistic tools in screening for privilege and work product may be found to have taken 'reasonable steps' to prevent inadvertent disclosure."

TECHNOLOGY ASSISTED REVIEW (TAR) GUIDELINES
EDRM/DUKE

CHAPTER ONE
DEFINING TECHNOLOGY ASSISTED REVIEW

CHAPTER ONE

DEFINING TECHNOLOGY ASSISTED REVIEW

A. INTRODUCTION

Technology assisted review (referred to as "TAR," and also called predictive coding, computer assisted review, or supervised machine learning) is a review process in which humans work with software ("computer") to train it to identify relevant documents.[2] The process consists of several steps, including collection and analysis of documents, training the computer using software, quality control and testing, and validation. It is an alternative to the manual review of all documents in a collection.

Although there are different TAR software, all allow for iterative and interactive review. A human reviewer[3] reviews and codes (or tags) documents as "relevant" or "nonrelevant" and feeds this information to the software, which takes that human input and uses it to draw inferences about unreviewed documents. The software categorizes documents in the collection as relevant or nonrelevant, or ranks them in order of likely relevance. In either case, the number of documents reviewed manually by humans can be substantially limited while still identifying the documents likely to be relevant, depending on the circumstances.

---

[2] In fact, the computer classification can be broader than "relevancy," and can include discovery relevance, privilege, and other designated issues. For convenience purposes, "relevant" as used in this paper refers to documents that are of interest and pertinent to an information or search need.

[3] A human reviewer is part of a TAR team. A human reviewer can be an attorney or a non-attorney working at the direction of attorneys. They review documents that are used to teach the software. We use the term to help keep distinct the review humans conduct versus that of the TAR software.

B. THE TAR PROCESS

The phrase "technology assisted review" can imply a broader meaning that theoretically could encompass a variety of nonpredictive coding techniques and methods, including clustering and other "unsupervised"[4] machine learning techniques. And, in fact, this broader use of the TAR term has been made in industry literature, which has added confusion about the function of TAR, defined as a process. In addition, the variety of software, each with unique terminology and techniques, has added to the confusion by the bench and bar in how each of these software works. Parties, the court, and the service provider community have been talking past each other on this topic because there has been no common starting point to have the discussion.

These guidelines are that starting point. As these guidelines make clear, all TAR software share the same essential workflow components; it is just that there are variations in the software processes that need to be understood. What follows is a general description of the fundamental steps involved in TAR.[5]

1. ASSEMBLING THE TAR TEAM

A team should be selected to finalize and engage in TAR. Members of this team may include: service provider; software provider; workflow expert; case manager; lead attorneys; and human reviewers. Chapter Two contains details on the roles and responsibilities of these members.

2. COLLECTION AND ANALYSIS

TAR starts with the team identifying the universe of electronic documents to be reviewed. A member of the team inputs documents into the software to build an analytical index. During the indexing process, the software's algorithms[6] analyze each document's text. Although various algorithms work slightly differently, most analyze the relationship between words, phrases, and characters, the frequency and pattern of terms, or other features and characteristics in a document. The software uses this features-and-characteristics analysis to form a conceptual representation of the content of each document, which allows the software to compare documents to one another.

---

[4] Unsupervised means that the computer does not use human coding or instructions to categorize the documents as relevant or nonrelevant.

[5] Chapter Two describes each step in greater detail.

[6] All TAR software has algorithms. These algorithms are created by the software makers. TAR teams generally cannot and do not modify the feature extraction algorithms.

3. "Training" The computer using software To Predict Relevancy

The next step is for human reviewers with knowledge of the issues, facts, and circumstances of the case to code or tag documents as relevant or nonrelevant. The first documents to be coded may be selected from the overall collection of documents through searches, identification through client interviews, creation of one or more "synthetic documents" based on language contained, for example, in document requests or the pleadings, or the documents might be randomly selected from the overall collection. In addition, after the initial-training-documents are analyzed, the TAR software itself may begin selecting documents that it identifies as: (i) most helpful to refine its classifications; or (ii) most relevant, based on the human reviewer's feedback.

From the human reviewer's relevancy choices, the computer learns the reviewer's preferences. Specifically, the software learns which combinations of terms or other features tend to occur in relevant documents and which tend to occur in nonrelevant documents. The software develops a model that it uses to predict and apply relevance determinations to unreviewed documents in the overall collection.

4. Quality Control and Testing

Quality control and testing are essential parts of TAR, which ensure the accuracy of decisions made by a human reviewer and by the software. TAR teams have relied on different methods to provide quality control and testing. One popular method is to identify a significant number of relevant documents from the outset and then test the results of the software against those documents. Other software test the effectiveness of the computer's categorization and ranking by measuring how many individual documents have had their computer-coded categories "overturned" by a human reviewer. Yet other methods involve testing random samples from the set of unreviewed documents to determine how many relevant documents remain. Methods for quality control and testing continue to emerge and are discussed more fully in Chapter Two.

5. Training Completion And Validation

No matter what software is used, the goal of TAR is to effectively categorize or rank documents both quickly and efficiently, i.e., to find a reasonable number of relevant documents while keeping the number of nonrelevant documents to be reviewed by a human as low as possible. The heart of any TAR process is to categorize or rank documents from most to least likely to be relevant. Training completion is the point at which the team has identified a reasonable amount of relevant documents proportional to the needs of the case.

How the team determines that training is complete varies depending upon the software, the number of documents reviewed, and the results targeted to be achieved after a cost benefit analysis. Under the training process in software commonly marketed as TAR 1.0,[7] the software is trained based upon a review and coding of a subset of relevant and nonrelevant documents, with a resulting predictive model that is applied to all nonreviewed documents. Here, the goal is not to have humans review all predicted relevant documents during the TAR process, but instead to review a smaller proportion of the document set that is most likely to help the software be reasonably accurate in predicting relevancy on the entire TAR set. The software selects training documents either randomly or actively (i.e., it selects the documents it is uncertain about for relevancy that it "thinks" will help it learn the fastest), resulting in the predictive model being updated after each round of training. The training continues until the predictive model is reasonably accurate in identifying relevant and nonrelevant documents. At this point, all documents have relevancy rankings, and a "cut-off" point is identified in the TAR set, with documents ranked at or above the cut-off point identified as the predicted relevant set, and documents below the cut-off point as the nonrelevant set.

In many TAR 1.0 processes, the decision whether the predictive model is reasonably accurate is often measured based on the use of a control set, which is a random sample taken from the entire TAR set, typically at the beginning of training, and is designed to be representative of the entire TAR set. The control set is reviewed for relevancy by a human reviewer and, as training progresses, the computer's classifications of relevance of the control set documents are compared against the human reviewer's classifications. When training no longer substantially improves the computer's classifications of the control set documents, training is viewed as having reached completion. At that point, the predictive model's relevancy decisions are applied to the unreviewed documents in the TAR set. Under TAR 1.0, the parameters of a search can be set to target a particular recall rate. It is important to note, however, that this rate will be achieved regardless of whether the system is well trained. If the system is undertrained, an unnecessarily large number of nonrelevant documents will be reviewed to reach the desired recall, but it will be reached. Ceasing training at the optimal point is not an issue of defensibility (achieving high recall), but rather a matter of reasonableness, minimizing cost of reviewing many extra nonrelevant documents included in the predictive relevant set.[8]

---

[7] It is important to note that the terms TAR 1.0 and 2.0 can be seen as marketing terms with various meanings. They may not truly reflect the particular processes used by the software, and many software use different processes. Rather than relying on the term to understand a particular TAR workflow, it is more useful and efficient to understand the underlying processes, and in particular, how training documents are selected, and how training completion is determined.

[8] In many TAR 1.0 workflows, this point of reaching optimal results has been known as reaching "stability." It is a measurement that reflects whether the software was undertrained at a given point during the training process. The term "stability" has multiple meanings. The term "optimum results" is used throughout to eliminate potential confusion.

Compare this process with software commonly marketed as TAR 2.0. Here, the human review and software training process are melded together; review and training occur simultaneously. From the outset, the software continuously analyzes the entire document collection and ranks the population based on relevancy. Human coding decisions are submitted to the software, the software re-ranks the documents, and then presents back to the human additional documents for review that it predicts as most likely relevant. This process continues until the TAR team determines that the predictive model is reasonably accurate in identifying relevant and nonrelevant documents, and that the team has identified a reasonable number of relevant documents for production. There are at least three indicators of when completeness has been reached. The first is when a reasonable recall rate is reached (the human review team has reviewed a set of documents that reached a certain level of recall rate, which is calculated/tracked by the TAR software or by a TAR team member during the review). The second is the point at which the software appears to be offering up for review only nonrelevant or a low number of marginally relevant documents. The third is the point at which the human review team has identified an expected, pre-calculated number of relevant documents. In other words, the team took a sample before review started to estimate the number of relevant documents in the TAR set, and then the human team reviewed documents until it reached approximately that number. When training is complete, the human reviewers will have reviewed all the documents that the software predicted as relevant up to that point of the review. If the system is undertrained, then the human reviewers will not have reviewed a reasonable number of relevant documents for production, and the process should continue until that point is reached.

Before the advent of TAR, producing parties rarely provided statistical estimates as evidence to support the effectiveness of their document reviews and productions. Only on a showing that the response was inadequate did the receiving party have an opportunity to question whether the producing party fulfilled its discovery obligations to conduct a reasonable inquiry.

But when TAR was first introduced to the legal community, parties provided statistical evidence supporting the TAR results, primarily to give the bench and bar comfort that the use of the new technology was reasonable. As the bench and bar become more familiar with TAR and the science behind it, the need to substantiate TAR's legitimacy in every case should be diminished.[9]

Nonetheless, because the development of TAR protocols and the case law on the topic is evolving, statistical estimates to validate review continue to be discussed. Accordingly, it is important to understand the commonly cited statistical metrics and

[9] The Federal Rules of Civil Procedure do not specifically require parties to use statistical estimates to satisfy any discovery obligations.

related terminology. At a high level, statistical estimates are generated to help the bench and bar answer the following questions:

- How many documents are in the TAR set?
- What percentage of documents in the TAR set are estimated to be relevant, how many are estimated to be nonrelevant, and how confident is the TAR team in those estimates?
- How many estimated relevant documents did the team identify out of all the estimated relevant documents that exist in the review set, and how confident is the team in that estimate?
- How did the team know that the computer's training was complete?

TAR typically ends with validation to determine its effectiveness. Ultimately, the validation of TAR is based on reasonableness and on proportionality considerations: How much could the result be improved by further review and at what cost? To that end, what is the value of the relevant information that may be found by further review versus the additional review effort required to find that information?

There is no standard measurement to validate the results of TAR (or any other review process). One common measure is "recall," which measures the proportion of truly relevant documents that have been identified by TAR. However, while recall is a typical validation measure, it is not without limitations and depends on several factors, including consistency in coding and the prevalence of relevant documents. "Precision" measures the percentage of actual relevant documents contained in the set of documents identified by the computer as relevant.

The training completeness and validation topics will be covered in more detail later in these guidelines.

CHAPTER TWO
TAR WORKFLOW

A. INTRODUCTION

TAR can be used for many tasks throughout the Electronic Discovery Reference Model (EDRM), from information governance to deposition and trial preparation, which are discussed in Chapter Three. This chapter focuses on the use of TAR to determine relevancy of documents. To be more specific, the chapter focuses on a suggested workflow by which a human reviewer works with a computer that can be taught to classify relevant and nonrelevant documents in support of document production obligations. When the human training and computer review are complete, the documents capable of being analyzed will be classified into two piles: the predicted relevant set, which may have been reviewed by humans or may be unreviewed but predicted to be relevant (i.e., documents subject to potential production) and the predicted nonrelevant set, which are typically not reviewed by humans (i.e., documents not subject to potential production).[10]

Under this workflow, a human reviewer will have reviewed, or will have the option to review, the predicted relevant set prior to production. The documents in the predicted nonrelevant set typically are omitted from human review based on the classification decisions made by the computer.[11] From this perspective, the computer is supplementing the need to have humans engage in first-pass review of the documents for relevancy.

The resulting benefits are often that: (i) the first-pass review can be completed faster; (ii) the amount of human resources required to conduct the first-pass review is substantially less; (iii) the overall cost of the review is lower (although there is debate in the industry regarding the amount of those savings); and (iv) industry experience and evidence from experimental studies suggest that TAR can make relevance determinations as accurately as human review teams, provided that a reasonable workflow is applied to suitable data.

The TAR workflow, to date,[12] can work to fully meet Fed. R. Civ. P. 26 discovery obligations (and their state equivalents), often with lower cost and in shorter times

[10] Please see the Appendix for further information on the definitions of predicted relevant set and predicted nonrelevant set.

[11] This workflow does not apply to any other use cases, such as using TAR to simply prioritize documents for human review (this means the entire review set will still be reviewed by humans, but the computer makes the review more efficient by prioritizing the most likely relevant documents to be reviewed; this review can be done in support of production obligations), early case assessment, opposing party production analysis, or fact/investigatory research.

[12] We note "to date," as it is fully anticipated that technology will continue to evolve, and new workflow components will be incorporated into standard TAR workflows.

than linear review. A party should consider the workflow components herein when formulating a final workflow to satisfy Rule 26 obligations.[13]

To that end, there are a variety of software that can be used as part of this workflow, each with its own unique terminology and a set of distinguishing competitive advantage features. These guidelines provide a framework to address the approaches that different software use. Workflow considerations are identified throughout to help explain the differences among software.

B. FOUNDATIONAL CONCEPTS & UNDERSTANDINGS

1. KEY TAR TERMS

To avoid confusion and miscommunication, it is important to explain basic definitions and concepts relevant to the discussion.[14] Definitions of the key terms can be found in the Appendix.

2. TAR SOFTWARE: ALGORITHMS

TAR is a review process. In order to engage in TAR, software is required. There are numerous software available that may be used as part of a Rule 26(g) reasonable inquiry that leads to a defensibly sufficient production. Drastically simplified, the software applies a set of instructions and rules ("algorithms") to a data set. Generally, there are two main algorithms that the software uses to review documents: (1) feature extraction algorithms, which allow the software to identify content in documents, and thus establish relationships among documents in the TAR set; and (2) supervised machine learning algorithms, which use the organized set of features to infer relationships between documents and thus classify documents in the data set pursuant to criteria such as relevance.

a) Feature Extraction Algorithms

At a high level, feature extraction algorithms: (i) analyze each document within the TAR set; (ii) extract meaningful values, sometimes referred to as feature values, from each document; and (iii) store these values.[15] After analyzing all documents in

[13] Not all components may be needed to satisfy a Rule 26 obligation, which will depend on the specific facts and circumstances of each matter. Note that references to the Federal Rules of Civil Procedure throughout are intended to include their state law equivalents when relevant.

[14] The Appendix contains the most technical language on statistics in these guidelines. The purpose of the guide is NOT to educate on the minutiae of how to do statistical calculations and the differences in approaches of statistical calculations; rather, it is to note that statistical calculations may occur through the use of the TAR workflow, and the types of statistics (like recall) that are referred throughout the guidelines need to be understood.

[15] For example, if a document is about a blueberry pancake eating competition, one feature may be blueberry pancakes.

the TAR set, the computer can then organize the TAR set according to the values of each document's features.

All TAR software has feature extraction algorithms. The feature extraction algorithms are created by the software makers. TAR teams generally cannot and do not modify the feature extraction algorithms.

b) Supervised Machine Learning Algorithms (Supervised Learning Methods)

Whereas a feature extraction algorithm allows the TAR software to develop a representation of the content of documents and relationships among them, a supervised machine learning algorithm allows a human reviewer to train the software to recognize relevance. For the software to begin classifying documents as to relevance, documents that are representative of relevant content must be identified and submitted to the computer. For many supervised machine learning methods, documents that are representative of nonrelevant content must also be identified and submitted. Once a set of relevant and nonrelevant examples have been submitted, the software analyzes their features and builds a predictive model, a classification system that categorizes or ranks documents in the TAR set.[16] This process of submitting representative examples and having the software analyze the examples to build the model is often referred to as "training."

Overall, supervised machine learning methods allow for a training process that is iterative and interactive, when the human reviewer and software provide feedback to each other to improve the software's ability to analyze and classify documents. The software will rank or classify the documents within the TAR set, and the team will use the rankings or classifications to determine which documents are likely relevant, and which are not. A more detailed discussion on training processes and variations is found in Section C (5).

c) Varying Industry Terminology Related to Various Supervised Machine Learning Methods

Supervised machine learning methods utilize iterative training processes. Human reviewers code documents in multiple rounds and submit them to the TAR software to fine-tune the software's ability to classify relevant documents.

---

[16] The terms "classifies," "ranks," and "categorizes" are used in this document. In the context of this workflow, TAR software creates a predictive model or classifies documents in the TAR set as likely relevant or nonrelevant. This classification can be expressed in various ways, depending on the software. For example, some software rank documents based upon a scoring system from 0 – 100, with 0 being nonrelevant and 100 being relevant. Other systems do not use scores but categorize documents as relevant and nonrelevant. However, even when categorization like this occurs, there is still an underlying measure that the system is using to determine relevancy.

C. THE TAR WORKFLOW

A defensible TAR workflow addresses the following components:

- Identify the team to finalize and engage in the workflow
- Select the software
- Identify, analyze, and prepare the TAR set
- Develop project schedule and deadlines
- Human reviewer prepares for engaging in TAR
- Human reviewer trains the computer to detect relevancy, and the computer classifies the set documents
- Implement review quality control measures during training
- Determine when computer training is complete and validate
- Final identification, review, and production of the predicted relevant set

1. IDENTIFY THE TEAM TO ENGAGE IN THE TAR WORKFLOW

Tasking the appropriate people, process, and technology to engage in the workflow is critical to satisfying production obligations. With respect to the people, a team should be identified to finalize and engage in TAR. Typically, this team may include (in smaller-size actions, a single individual can serve multiple roles):

- SERVICE PROVIDER. The service provider provides access to the TAR software. The service provider can describe the workflow and support the process once it begins. The service provider can be a client, law firm, e-discovery service provider, or TAR software provider. Selection of the service provider is discussed in Section C (2).
- SOFTWARE PROVIDER. The software provider is the creator of the software. Some service providers create their own software (and, thus, are also the software provider), while others license it from software providers.
- WORKFLOW EXPERT. A workflow expert or litigation support project manager advises the team on the design and implementation of the workflow, and if necessary, supports the defensibility of the process.
- CASE MANAGER. The case manager is essential to every discovery project and is often responsible for managing the data. This may include keeping track of several items, such as: (a) the data that was collected and processed; (b) the data that survived any culling criteria, including date or search term limitations; (c) documents that were both included and excluded from the TAR set; and (d) the predicted relevant set and predicted nonrelevant set that result from the workflow.
- LEAD ATTORNEY. There must be at least one lead attorney engaged in the workflow who fully understands the scope of relevancy at issue. The lead

attorney is sometimes known as the subject matter expert on the case, or someone who is most familiar with the claims and defenses of the case. The lead attorney must work to ensure that every human reviewer and the software are engaging in accurate document review. A lead attorney sometimes engages in the actual review and training process of the workflow, and thus can also act as a human reviewer.

- HUMAN REVIEWER. A human reviewer reviews documents for relevancy, and these relevancy determinations are used to train the software. A human reviewer may also review documents that are predicted to be relevant to confirm relevance and check for privilege before production. As such, every human reviewer must be educated on the scope of relevancy to ensure reasonably accurate and consistent training of the software.

2. SELECT THE SERVICE PROVIDER AND SOFTWARE

In order to engage in the workflow, the producing party needs access to TAR software. The decision on what software to use goes hand-in-hand with the service provider selection. A key element to ensuring a successful project is a service provider who will be assisting or managing the process. The producing party needs to perform due diligence on the service provider selection. The service provider should have an expert who can describe the process in a meaningful and understandable way, including the steps that the team will need to take to ensure a reasonable review. Other topics that the producing party might consider discussing with the service provider are:

- Does the service provider have a written TAR guide?
- Which TAR software does the service provider have?
- Can the service provider demonstrate by using measurable verification methods that the software they use works for the particular assigned task?
- How many TAR-based reviews in support of production obligations has the service provider completed in the past six months or year? What were the results?
- Has the service provider ever provided affidavits or declarations in support of the workflow?
- How does the service provider report on the progress or provide updates on the workflow?
- What level of training and support will the service provider provide to the team?
- Does the service provider have an expert that is able to support or participate in discussions with the opposing party or the court on the use of TAR?
- If supplemental collections or rolling productions are anticipated throughout TAR, how will that impact the workflow?

- If foreign language is at issue, how will foreign language documents be handled?
- Who will be reviewing and coding the training documents, and where does that review take place?
- What factors or criteria are assessed to determine whether the workflow is reasonable?
- Is the TAR software actively supported? (Does the software provider periodically engage in upgrades, updates, and bug fixes to improve the software and workflow?)

3. IDENTIFY, ANALYZE, AND PREPARE THE TAR SET

Each document review requires the producing party to first identify the document set subject to review. This may involve "relevance culling criteria" that will limit the document collection and review to what is potentially relevant to the case. Typically, the relevance culling criteria will be based on custodians/document repositories, date ranges, and file types, and may also involve search terms.[17] The relevance culling criteria are often addressed during the Rule 26(f) meet and confer meetings.

After applying the relevancy culling criteria to the collected documents, an appropriate member of the TAR team analyzes the resulting document set and identifies problematic documents that the software will not be able to review. Software predominantly analyzes a document's text.[18] Documents with minimal or too much textual content can be problematic because there is either too little or too much information for the software to analyze.

Most TAR software prescribes a list of parameters to assist in identifying problematic documents. For example, documents to be excluded from the workflow are often based on file type, such as audio, video, and image files, as well as text size, such as documents containing more than a set number of megabytes of text.[19] TAR software vary in the types of files they are suited to analyze. The parameters used to exclude documents from the TAR set should be discussed with the service provider.

---

[17] If there is a very large volume of data of low richness, the relevance rates returned through search terms or other culling methods should be tested. Alternatively, use of search terms may limit the dataset to a size that TAR can manage. Although some oppose limiting a dataset before using TAR, pre-TAR culling may nonetheless be reasonable and desirable under many circumstances.

[18] Most software analyzes a document's text. However, some software may analyze other document metadata, such as email header fields and file names, and others may analyze a document's visual appearance independently of whether text is present.

[19] For example, the TAR set may be limited to emails, documents, presentations, or spreadsheets; and documents with a text size of less than a set number of megabytes. Any document not falling within these limitations is excluded from the TAR set.

After this analysis of the document set is complete, any documents that were excluded from the final TAR set should be tracked, and if necessary, sent through an alternate review workflow.

A TAR set should also be analyzed for foreign languages. Most software can analyze and review documents containing a mixture of human languages. Even so, a separate TAR workflow may be necessary for handling documents from each language. If documents from multiple languages are expected, the process for identifying and handling these documents should be discussed with the service provider.

Finally, once the TAR set is identified, it must be submitted to the software for review preparation. The software will typically perform a "build" over the TAR set.[20] As described earlier, this building process involves the computer analyzing each document's text (and potentially some metadata), extracting certain features, and organizing the TAR set according to these features. Typically, the native file types (for example, Word, Excel, or PowerPoint) do not matter for building the index (the build does not occur on the document's native file format, but on each native file's extracted text, to the extent that text exists).

After the build is completed, the team is ready to engage in training the software to identify what is relevant to the case.

a) Timing and the TAR Workflow

Although 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, an estimate of additional data that will be included in the review population is helpful.

Both TAR 1.0 and TAR 2.0 have timing considerations that must be factored into the project timeline. Generally, under a typical TAR 1.0 approach, the time it takes to train the computer to reach optimal results must be factored in when determining the length of time it will take to complete the review. To avoid a delay in commencing review, any segment of the document population that will require some level of review outside of TAR can be started.

Generally, under a typical TAR 2.0 approach, documents that are likely relevant are prioritized for review based on a continuous update of the relevancy

[20] Some software can support multiple builds, allowing for multiple workflows to be run simultaneously.

rankings throughout review. The initial prioritization can be commenced based on documents counsel identifies prior to the review. 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.

It is also important to understand the service provider'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. 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.

4. The Human Reviewer Prepares for Engaging in TAR

There are a couple of key preparation items that must be undertaken before a human reviewer can start to train the computer. Importantly, the scope of relevancy must be defined, and the lead attorney must train any other human reviewer on that scope. However, many times the scope of relevancy may evolve after the early stages of a matter. Although the scope of discovery is typically defined through the complaint and discovery request process, requesting and receiving parties frequently disagree on the scope of discovery requests, which may cause delays (sometimes substantial) in the final agreement or order on the scope of discovery for the matter. Motion practice, including motions to dismiss, can affect when the final scope of discovery is known. Ultimately, to use TAR, the lead attorney must be comfortable defining the scope of relevancy to be applied to the workflow. The team should discuss any negative consequences of engaging in the workflow prior to a reasonably defined discovery scope.

Once the lead attorney determines the scope of relevancy, the human reviewer or reviewers must be trained so that they may analyze and code training examples accurately and consistently.

> **Workflow Consideration**: Selecting the Human Reviewer to Train TAR. The human reviewer performing the training may be a single lawyer, a small group of attorneys, or a larger group of attorneys. Selecting the team that will be reviewing and coding the training examples can be dependent upon several factors, including production deadlines, the scope of relevancy, the complexity of the subject matter, the anticipated size of the training set, and the software to be used. For example, a team of 15 human reviewers may generate more inconsistent coding results to train the software in comparison to a team of two lead attorneys, and thus the quality control review of those human reviewers

may require a greater effort than if the lead attorneys trained the software.

Workflow Consideration: Establishing the TAR Process Coding Field(s). Most TAR workflows will use one tag (sometimes called "global relevance" or "universal relevance" or "super relevance" tag), which covers the entire scope of relevance. Under this approach, the same global relevance tag would be used regardless of whether the document being used to train is relevant to 1, 10, or 15 out of 15 relevant topics.

Some TAR software allow for the human reviewer to train the computer to recognize more than one relevant topic, allowing training on sub-topics or sub-issues of relevance, or on topics that overlap with relevance (such as privilege). Other software allow the human reviewer to train the computer to recognize all topics at the same time. Others may require a separate training session for each topic. Introducing multiple topics, if done carefully, can reduce the time for the review by reducing the complexity of distinctions to be learned by the computer and allowing adaptation to late-negotiated changes in the definition of relevance. Commonly, however, more topics require more review time and effort. If the lead attorney is considering training on more than global relevancy, the pros and cons should be discussed with the team.

Finally, the workflow expert should set expectations for the lead attorney and human reviewer on the training method experience, including what the workflow components are, any key decision points, and estimated times for training the computer.

5. Human Reviewer Trains the Computer to Detect Relevancy, and the Computer Classifies the TAR Set Documents

Now that the scope of relevancy and the structure for coding documents is in place, the human reviewer must engage in a process of conveying the decisions on the scope of relevancy to the computer, with the computer using that training to distinguish between relevant and nonrelevant documents in the remainder of the TAR set. The iterative training steps include:

- Identify training documents (selection of documents for human reviewer review);
- The human reviewer codes the training documents for relevancy;
- The human reviewer's relevancy decisions are submitted to the software;

- The software uses the training documents' relevancy decisions to build a predictive model, and applies the model to rank or classify all documents in the TAR set; and
- Repeat steps (1)-(4) (add more documents for human review and computer training) until further review is no longer needed because review's goals have been met.

Workflow Consideration: Selecting Training Examples. Supervised machine learning methods vary in how they select training examples for the human reviewer.[21] Training examples can be chosen based on human judgment, randomly, or by the computer based upon its analysis of the current training set and the TAR set ("computer feedback," also known as active learning).[22] Which form of training example identification should be used may depend upon: (1) the size and nature of the TAR set; (2) the review goals; (3) the software used; (4) the service provider's recommendations; and (5) any party agreement or court order.

In selecting training examples by human judgment, the team finds examples of relevant and nonrelevant documents that exist in the TAR set to train the computer. The team may find these examples through the use of relevant key words, clusters, concept searches, custodian information, or other metadata. The human reviewer reviews those documents and codes for relevancy, and then submits those examples to the computer to train it.

Training examples may also be selected randomly.[23] This means that the training examples are selected without concern for document content or based on any prior rounds of training.

Finally, the software may take into account prior training-round information to make selections of training examples,[24] which allows the computer to provide feedback based upon its categorization of documents after each round of training.

---

[21] As a practical note, all TAR training methods can be effective in classifying documents. Some methods may be more efficient (take less time) to achieve the review goals, but it is largely dependent on the nature of the data set and circumstances of the case.

[22] The first set of training examples is called a "Seed Set." This set of training examples cannot be selected by the computer based upon prior rounds of training, as there are no prior rounds of training at that point.

[23] Human selection and random selection have been traditionally known as "passive" selection methods for training. This is because the computer is not involved in making subjective decisions on which documents should be used as training examples.

[24] This has been traditionally known as "active" training method. This is because the computer is actively involved in identifying documents that should be used as training examples, which is done based upon past training rounds.

The computer's choice of which documents to put before the human reviewer depends on the weighting of various factors by the machine learning algorithms used. One method of this type of selection, used in TAR 2.0, is called "relevance feedback," whereby the software attempts to identify for review only those documents that are most likely to be relevant. Other methods, which may be used in TAR 1.0 and also sometimes in TAR 2.0, also take into account factors such as how different the potential training examples are from each other and from previously coded examples, as well as how unsure the software is about the examples.[25]

There are differing views in the e-discovery industry as to the best method for selecting training examples. Some of these differing views reflect preferences for different types of workflows, and the fact that different workflows (and cost structures) are well-suited to different ways of choosing training data. Other views result from differing levels of concern for possible biases introduced when selecting documents by human judgment or differing preferences by human reviewers. It is important to recognize that any approach to selecting training data will produce an effective predictive model if it is used to produce a sufficiently broad training set. Thus, differing views over selection of training data are less about whether an effective predictive model can be produced, than about how much work it will take to do so.

Some general characteristics of different selection methods should be noted:

- The quality of training documents selected by manual judgment depends on the skill of the team, their knowledge of the TAR set, and the relevance definition. Selection by manual judgment will typically improve the computer's ability to find relevant documents, but it may require more time and cost to ensure enough relevant samples are identified that span the entire scope of relevancy to be used to train the system. For example, if the team's scope of relevancy spans ten document production requests, but the team only finds training-set examples that cover five out of ten requests, then the computer may not identify documents relevant to the other five requests. Concerns about such omissions and other forms of potential bias often lead to decisions to combine selection by human judgment with other training document selection methods.
- Random selection is a rapid method of choosing training documents that is supported by most software. It is independent of human judgment and gives every document in the TAR set an equal chance to be selected. It requires no human effort and is immune to concerns about biased selection. But training sets produced by random sampling may need to be larger than those produced by other methods, particularly when richness is low.

[25] This is uncertainty feedback, where the computer attempts to present examples it is least certain about for relevancy. The computer will avoid presenting documents for which it is most certain about relevancy.

- Computer feedback/active learning methods are also an efficient and automated way of choosing training documents, though some require substantially more computer time (as the computer processes decisions and identifies new documents to review) than simple random sampling. Computer feedback methods may require less training as compared with random selection to produce an effective predictive model, particularly for low richness TAR sets. Some TAR software may combine various types of computer feedback learning to achieve the review goals.
- If some documents in previous batches were chosen by manual judgment, those choices may influence later choices made by computer feedback methodologies, potentially leading to concerns about bias. On the other hand, some computer feedback methods are designed to choose documents different from those in previous training batches, and thus can help mitigate concerns of bias.

6. Implement Review Quality Control Measures During Training

An important function of any document review is to ensure that relevancy decisions are reasonably accurate and consistent.[26] Because a human reviewer analyzes documents and applies their own understanding of the scope of relevancy, there is variation in how documents are coded, which in turn causes variation in how the software classifies documents for relevancy. This challenge is commonly addressed by engaging in review quality control.

Many review quality-control measures that are applied in non-TAR workflows can also be applied in TAR workflows and aid in ensuring reasonable review is taking place. Some of these review quality-control options are discussed below.

a) Decision Log

For medium-to-large sized human reviewer teams, a common method to assist those reviewers with their understanding of the matter and the scope of relevancy is to create a decision log. A decision log is a record of relevancy questions made by the lead attorney, which provide guidance to the human reviewer. The lead attorney answers the questions and provides any needed clarification on the relevancy scope. A question may touch on an issue that is not addressed in the current scope of relevancy, resulting in the update of the scope. As more entries are added to the decision log, the more valuable it becomes as a reference for the human reviewer.

[26] Note that the larger the training-attorney team, the greater the chance for inconsistent coding.

b) SAMPLING

Another long-established method to ensure quality is to use samples of documents to both measure and improve the quality of coding by a human reviewer. For example, a sample of a human reviewer's coding decisions can be generated and reviewed, in many instances, by the lead attorney, which ensures that the coding values are in-line with the scope of relevancy. While reviewing the samples, a record of the lead attorney's overturns of the human reviewer's relevancy decisions can be maintained. This record of overturns can be used to re-educate the team on the scope of relevancy by reinforcing the correct relevancy scope that should be applied. Although using sampling in quality control for a human reviewer has similarities to the use of coded data to evaluate (e.g., control sets) and train (e.g., training sets) software, the fact that one is evaluating and aiding humans in their coding can substantially change the priorities in sampling. Factors considered when developing a sampling methodology for quality control of a human reviewer include: (a) who will review the samples; (b) how to keep track of the sampling process; (c) how often will documents be sampled; (d) how many documents will be re-reviewed; and (e) how will the samples be selected.

In TAR workflows, the sampling of documents for review quality control can also be based on the predictions of the software. For instance, sampling may be focused on documents that the software ranks or categorizes as most likely to be relevant. If the human reviewer identifies a large percentage of documents to be nonrelevant, this may suggest an issue either with the human reviewer's coding decisions in the training set or with the effectiveness of the current predictive model.

c) REPORTS

Some TAR tools provide the ability, sometimes in the form of a report, to identify documents for which the software's classification and the human reviewer coding disagree on relevancy. Using these tools, the team can easily identify training set documents that: (1) the software considers as likely relevant and the human reviewer coded as nonrelevant; or (2) the software considers as likely nonrelevant and the human reviewer coded as relevant. This analysis can be done on a regular basis, with a human reviewer re-reviewing inconsistently classified documents for final resolution, with any changes resulting in updating of the software's classification decisions, and also, if need be, a continued re-education of the human reviewer on the scope of relevancy.

7. DETERMINE WHEN COMPUTER TRAINING IS COMPLETE AND VALIDATE

In a TAR workflow, a major decision is when to stop the training process. In practice, this usually means trying to quantify the percentage of relevant documents identified and validate the success and reasonableness of the review, the adequacy of

which is assessed under Rule 26 proportionality and reasonableness factors. There is currently no black letter law or bright-line rule as to what constitutes a reasonable review; rather, each workflow must be analyzed for reasonableness based upon the circumstances of the matter and the proportional needs of the case.

a) Training Completion

There are several indicators that provide information to allow the team to make reasonable decisions on training completion. These measurements are directed toward understanding whether the review process achieved optimal results. An optimal review result will vary from case to case and does not have a technical definition. In many instances, the training process is considered optimal if review of further training documents is unlikely to substantially improve the results. Some software provides measurements or indicators of training completion, while other measurements can be derived by the team from estimates of effectiveness. There are three broad approaches to understanding whether training results are optimal: (1) tracking of sample-based effectiveness estimates; (2) observing sparseness of relevant documents returned by the computer during active learning; and (3) comparing the predictive model behaviors.

(i) Tracking of Sample-Based Effectiveness Estimates

By comparing how sample-based effectiveness estimates (e.g., recall at a fixed cost level) change over time, the team can get a sense of whether further training is of value. Two types of samples may be used. A control set is a random sample taken from the entire TAR set,[27] typically at the beginning of training. The control set is reviewed for relevancy by a human reviewer and, as training progresses, the computer's classifications of relevance of the control set documents are compared against the human reviewer's classifications.[28] When training no longer substantially improves the computer's classifications, this is seen as a point of optimal training. An alternative to drawing a control set at the beginning of training is to draw a random sample only when it is believed that training or review should be stopped. This measurement is commonly taken in certain TAR 1.0 workflows.

(ii) Observing Sparseness of Relevant Documents Returned by the Computer During Active Learning

Another measurement of training completion involves the human reviewer continuing to train the computer on relevancy until the number of relevant documents presented by the computer for human review is too low to justify

[27] Note that as the control set is a random sample from the TAR set, it can be used to calculate various statistical estimates, namely recall, richness, and precision.

[28] Accurate coding of the control set by the human reviewer is very important since the coding is used as the "gold standard" to measure how well the computer's learning is progressing.

additional review. When the review reaches this point, it is seen as an indicator of optimal training, and the training is complete, because no additional training is anticipated to improve the predictive model or identify more relevant documents. This measurement is common in TAR 2.0 workflows that use relevancy feedback learning.

(iii) Comparison of Predictive Model Behaviors

Another approach to monitoring training completion is to directly compare the rankings of classified documents across different training-round iterations. A wide variety of approaches is possible, and the details are not always revealed by the software provider. As an example, predictive models generated during different iterations could be used to rank, score, or classify the entire TAR set. If those predictions are becoming largely static (e.g., document ranks or categorizations are not changing), then the team may be able to conclude that further training likely will have diminishing benefits, because the behavior of the TAR predictive model is not meaningfully changing at that point. The team may view this as the point of optimal training and stop training.

(iv) Comparing Typical TAR 1.0 and TAR 2.0 Training Completion Processes

The following example clarifies a variation between TAR 1.0 and 2.0 processes.[29] In a TAR set of 200,000 documents, 20,000 of the documents are relevant (10% richness). For purposes of this example, the team intends to use the workflow in an attempt to achieve a recall of at least 80%, i.e., identify at least 16,000 of the 20,000 relevant documents.

Many TAR 1.0 workflows typically start by randomly selecting 400 to 2,000 documents, which serve as a control set. A human reviews and identifies the relevant documents in the control set.[30] The resulting percentage of relevant documents in relation to all reviewed documents in the control set is a benchmark percentage, defined as "richness," which is used to evaluate the TAR review of the entire predicted

[29] Note that the illustrations are provided at a high level and only reflect two possible TAR sets and processes. The illustrations should not be used to conclude that situations similar to these are or are not reasonable and proportionate. Each TAR project must be conducted on a case-by-case basis. In addition, there may be variations in service providers' TAR 1.0 and TAR 2.0 processes. This illustration only provides two common examples of TAR (TAR 1.0 using a control set with uncertainty feedback, and TAR 2.0 with no control set but using relevancy feedback only).

[30] Again, this is a broad generalization of certain traditional TAR 1.0 processes, and there may be variations in TAR 1.0 software. The size of the control set is influenced by the richness of the TAR set and the margin of error for recall estimates. The lower the desired margin of error, the higher the number of control set documents are required to reach that level of estimated certainty. Very low richness (very small numbers of responsive documents in the TAR set) will also require a larger control set in most instances.

relevant document set. The benchmark percentage of relevant documents in the control set (richness) is a key element in analyzing when TAR has attained optimal results.

After the benchmark percentage of relevant documents has been determined, the software begins the TAR training by randomly or actively selecting training documents, which will include and identify both relevant and nonrelevant documents (contrast with the TAR 2.0 example below). General experience has shown that 400 to 2,000 documents are used for training. These training documents are reviewed and used to build the predictive model.

When review and processing of the training documents achieves "optimal results," the predictive model's rankings are locked, and the final predictive model is applied to the entire set of documents and used to identify the predicted relevant document set. Determining when TAR has reached "optimal results" applies a reasonableness test. It takes into account not only the percentage of identified relevant documents compared with the control-set benchmark percentage, but also a cost-benefit analysis. The cost-benefit analysis weighs the likelihood of identifying significant relevant documents missed in the TAR results and the added costs and burdens, which would be incurred in further processing and review to identify them. To recall every relevant document, the TAR results would have to include a large percentage of documents that are not relevant, which would increase the costs and burdens of review. Depending upon the size of the TAR document set, every percentage point could represent a very large number of predicted relevant documents that are actually nonrelevant, which would result in unnecessary review costs and burden. General experience has shown that achieving a recall rate of 75% to 85% has been a good balance in many cases, but the facts and circumstances of each case are different, and no rigid standards are appropriate. In our example, we used an 80% recall rate for optimal results, which would require review of at least 16,000 documents identified as relevant, but likely more, as the TAR model predictions cannot perfectly select only relevant documents for review.[31]

For some traditional TAR 2.0 software, the workflow might start by generating a random sample of the review population to get a sense of the richness. This allows the team to estimate the number of relevant documents in the review set (here, 20,000 documents). Thus, for a recall goal of 80%, a team can estimate upfront the need to identify at least 16,000 documents. Those relevant examples found in the richness sample, along with any other relevant samples identified by the team, are then submitted to the software to start learning relevancy. As each document is reviewed and submitted to the software, the software continues to re-rank the document set

[31] Although failing to identify 4,000 relevant documents may appear troubling, studies have consistently shown that lawyers reviewing every document manually identify the same or fewer number of relevant documents.

and present only those documents it believes are relevant (i.e., engages in relevancy feedback).[32] This process continues until optimal results are reached, which is when the computer is returning a very low number of relevant documents for review. At that point, all predicted relevant documents will have been used as training examples. In this TAR 2.0 example, at least 16,000 documents will both be reviewed by a human reviewer and used to train the computer to achieve the recall goal. In practice, more than 16,000 documents will need to be reviewed, as neither human selection nor TAR model predictions can perfectly select only relevant documents for review.

b) Validation

Whatever software is utilized, it must generate, or allow for the generation of metrics or effectiveness measures, which allow the team to evaluate the workflow and determine if the review goals have been met. In many instances, this means the team will need to be able to measure the recall achieved. Estimates of other effectiveness measures besides or in addition to recall may also be used. These methods are not mutually exclusive. To the contrary, all these approaches may play a role in a reasonable inquiry, consistent with Rule 26(b) proportionality considerations and Rule 26(g).

Recall measurements are statistical in nature and involve random sampling, which has underlying parameters that determine the sample size: richness, confidence level, and confidence interval. These parameters dictate how many sample documents need to be reviewed in order to achieve a certain comfort level with the recall estimate. If the team achieves a recall level that it believes is reasonable for the workflow and matter, then training can be considered complete and the predicted relevant set is validated as reasonable. If the team does not achieve a reasonable level of recall, it may need to go back and conduct additional training to further identify more relevant documents to be added to the predicted relevant set.

> **Workflow Consideration**: Identifying a target recall level prior to the start of training. Some software requires the team to identify the desired estimated recall level before the start of training. In these situations, the human reviewer continues to review training examples and use them in the training set until that estimated recall level is achieved.
>
> One challenge that may occur with this process is that it is unknown how long training will take, or how many documents will be needed in

[32] It presents only those that are ranked or categorized as predicted relevant. Again, not all of these documents will be relevant. As the review continues, the human reviewer will see lower numbers of relevant documents.

the training set, to achieve the estimated recall level. This makes review judgments difficult if the targeted recall is set too high, because it may require unreasonable and disproportionate human review to train the computer to be able to achieve that targeted recall. If that occurs (the training goes beyond the reasonable and proportional review of documents), the team may need to lower the targeted recall level so that a reasonable and proportionate predicted relevant set can be identified (e.g., a sliding scale based upon the document rankings).[33]

At a particular recall percentage, there is an associated predicted relevant set that is identified. The ability to determine the size of the relevant set for various recall percentage options is a feature found in some software. As noted above, in some TAR 2.0 software, the entire predicted relevant set has already been reviewed by humans, and all that is left at that point is to calculate the recall achieved by that predicted relevant set.

Workflow Consideration: How Recall Is Estimated. There are two primary approaches to estimating the extent to which TAR has found relevant documents.[34] One common approach involves taking a random sample of the TAR set, reviewing it for relevancy, identifying the relevant documents, and then determining the percentage of documents that are relevant from this sample. By examining how this sample of relevant documents is categorized by TAR, the recall of the review can be estimated.

The second method of determining recall involves drawing a random sample from the documents in the predicted nonrelevant set. The sample is used to estimate the richness of the predicted nonrelevant set, sometimes called the "elusion rate" of responsive documents in the predicted nonrelevant set. The elusion rate can be used to estimate the number of relevant documents in the nonrelevant set. This measurement can be used to calculate recall when used with other measurements, such as the number of estimated or actual responsive documents in the responsive TAR set (depending on workflow).[35]

Even if an acceptable recall level is attained, attorneys may further check TAR performance by evaluating the importance of relevant documents found in validation that were categorized nonrelevant by TAR. If the missed documents are especially

[33] If the recall goals are not being met, it may also mean that the training itself may be at issue. If that is suspected, the team may also need to engage in remedial measures (*See* Section 8(e)).
[34] These examples are NOT intended to educate on the minutiae, or all variations, of how to do statistical calculations.
[35] Consistency of review decisions in the predicted nonrelevant set and the predicted relevant set is important for this approach.

significant to the case and contain evidence that may not be contained in the relevant set, the attorneys should consider whether additional training is needed to enable the computer to identify other relevant documents in the predicted nonrelevant set.[36]

There are variations in how recall is estimated, so it is important to understand how the service provider or workflow expert calculates recall.

8. Final Identification, Review, and Production of the Predicted Relevant Set

After the team trains and validates, the TAR process will have separated the TAR set into the predicted relevant set and a predicted nonrelevant set. In some workflows, the predicted relevant set will be a combination of documents the human reviewer classified as relevant, along with documents not reviewed by the human reviewer but the computer determined as relevant. In other workflows, the predicted relevant set will be all documents reviewed and identified as relevant by the human reviewer.

The predicted nonrelevant set will also be identified. Some of these documents will have been reviewed by the human reviewer and used in the training set. But, the vast majority of these documents will not have been reviewed by the human reviewer, only the computer.[37]

In addition to the predicted relevant and nonrelevant sets, a third group of documents may also emerge from the TAR process—documents that could not be categorized by the TAR software. For example, documents containing illegible text, solely comprised of numbers, or from which conceptually relevant content may not be gleaned, will likely not be categorized. These documents should still be considered a part of the review set and will need to be investigated for relevance outside of the TAR process.

Finally, in every workflow leading up to a document production, steps should be taken to address family members, privileged documents, confidentiality, redaction, and other issues that fall outside the workflow process. An attorney should assess

---

[36] For example, an attorney may believe 80% recall to be sufficient and proportional in a matter given the needs of the case and the results of the TAR process. That view would be confirmed if the relevant validation documents missed by TAR are relevant but of low value in context of the documents TAR found. On the other hand, the attorney may reconsider TAR's effectiveness if the missed documents contain evidence that is key to the matter.

[37] During the process of engaging in the TAR workflow, some documents will remain uncategorized, because the computer was not able to make decisions on relevancy or because the documents are not similar to any training documents or do not contain enough meaningful content. These documents should be identified by the team and addressed through additional searching, sampling, or review, depending on the case circumstances.

whether to fully review, partially review, or simply produce any documents that have not yet had human review and that are in an unreviewed predicted relevant set and family members. Considerations should include costs of further review, weighed against the risks of producing nonrelevant documents that could include confidential, private, or privileged[38] client information.

9. WORKFLOW ISSUE SPOTTING

There are certain challenges that may arise during the workflow. Common challenges include: (1) extremely low or high richness of the TAR set; (2) supplemental collections; (3) changes to the scope of relevancy; and (4) unreasonable training results. These challenges should be identified as early as possible in the process and discussed with the service provider.

a) Extremely Low or High Richness of the TAR Set

The team will want to identify whether the TAR set at issue has extremely low richness (a very small percentage of the TAR set is relevant) or extremely high richness (a very high percentage of the documents in the TAR set is relevant).

If the TAR set's richness is extremely high, it could undermine the utility of using TAR to assist in identifying documents for relevancy. The team may want to consider other measures to prioritize or organize the TAR set for review.

If the TAR set's richness is extremely low, human reviewers may have a difficult time training the software on what is relevant, because examples may be scarce or difficult to come by in the TAR set. The team should discuss how to resolve the training of the low richness TAR set, which will depend on the training method used by the service provider.

The amount of training required under TAR 1.0 and TAR 2.0 depends upon the TAR set's richness and the statistical certainty required, such as margin of error and confidence level. If richness is low, TAR 1.0 workflows, which use a control set, may require more review. Conversely, if richness is high, TAR 2.0 workflows may require more review.

[38] Parties should consider protecting privileged documents from waiver, regardless of the circumstances under which they were produced, with an order under Federal Rule of Evidence 502(d) or similar order when available. Such an order should not, however, prevent a party from conducting an appropriate privilege review if that party chooses. *See* The Sedona Conference, *Commentary on Protection of Privileged ESI* (2014). Absent a Rule 502(d) order, Rule 502(b) only prevents waiver if the producing party used reasonable procedures to identify and avoid producing privileged documents.

b) Supplemental Collections

If the team introduces new documents into the TAR set after the training has already started, the human reviewer and the software may need to learn more about the new documents in order to ensure reasonable training and review. There are two main questions to ask the service provider when supplemental collections are contemplated: (1) should the new documents be merged with the original TAR set, or treated separately;[39] and (2) if merged, how or when does the team introduce these new documents to the computer and ensure that the software has analyzed them properly?

Overall, many supervised machine learning methods use the human and software knowledge already acquired to train and categorize the supplemental collection being added to the original TAR set. In other words, the software tries to avoid starting from scratch by leveraging prior training applied to the TAR set. Ultimately, the team will utilize core workflow components in an attempt to update the education of the human reviewer and the software to complete training and review of the documents, to conduct new statistical estimates of completion of review, and to validate the training and review.[40] This leads to the identification of an updated predicted relevant set and predicted nonrelevant set.

c) Changing Scope of Relevancy

Another challenge that may arise occurs when the scope of relevancy expands or contracts during the TAR training, or at some point after the review has been completed. In these situations, the team may need to go back and update the human reviewer and the software on what is relevant (in other words, they may need to conduct a "re-review" of documents to identify the reasonable predicted relevant set). The team will need to assess how different the original review scope of relevancy was from the new scope of relevancy.

If there are multiple new issues, or very broad new issues, the team will most likely need to update the training and review to reflect this scope. If the differences are discrete or narrow in nature, the team may be able to use other strategies to identify those discrete topics for further review and avoid updating the training and

[39] If the new documents are not merged with the original TAR set, then the new documents could go through a separate, parallel TAR workflow. This would result in two different TAR exercises.

[40] Note that if the team merges the supplemental collection with the original TAR set, any statistical calculations on how well the review of the original TAR set was performed may be stale (due to the supplement, the updated TAR set has new properties and may have new richness, recall, and precision). In addition, the TAR workflow may be adjusted when supplemental documents are expected. For example, limiting an upfront control set may reduce the total number of documents in a review when dealing with supplemental collections, because there is no need to update a large control set before continuing with training.

review. The team should work with the service provider and workflow expert to understand what steps need to be taken to reasonably deal with this challenge.

d) Unreasonable Training Results

A challenge may also arise when, upon conducting validation of the training, the team believes the review results are not reasonable. This determination can often be made based on the quantity and quality of documents the software incorrectly categorized. In these situations, there are several actions the team may take to improve the review results, which will largely depend on what the issue is and what software was used. Any remedial measures should be discussed with the service provider and workflow expert to ensure defensibility of process, which may include the following:

- Confirm that the sampling techniques used were statistically appropriate, including that the correct set of documents was sampled and a sufficient number of documents was sampled;
- Confirm that the control set and validation-set documents are coded correctly by the human reviewer, if applicable;
- Engage in additional control set document review to reduce the uncertainty of effectiveness estimates;[41]
- Re-review training set documents to confirm the human reviewers' relevancy decisions and modify them as necessary;
- Engage in additional review of training set documents to improve the training results;[42]
- Review documents that the human reviewer and computer classified differently to correct any inconsistencies or to evaluate whether certain types of documents create problems for categorization by concept;[43] or
- Identify any large quantities of problematic documents in the TAR set that the computer is having difficulties making relevancy classifications on, remove them from the TAR set, and review them outside the workflow.

[41] The greater the certainty/lower margin of error used to create the control set, the less likely it is that additional review will change the metrics.

[42] For example, this may involve using relevant documents identified in the predicted nonrelevant set as training documents.

[43] This may require creating a new control or validation set afterwards.

CHAPTER THREE
ALTERNATIVE TASKS FOR APPLYING TAR

A. INTRODUCTION

TAR can be an effective tool to identify relevant documents and respond to discovery requests. But TAR can also be useful to handle other discrete discovery tasks. Several examples of alternative tasks follow.

B. EARLY CASE ASSESSMENT/INVESTIGATION

Early Case Assessment (ECA)[44] is one efficient way to get a high-level view of the overall makeup of the documents. From here, counsel will have some understanding of the content of the documents and can better assist with development of legal strategy.

In an appropriate case, a practitioner may use TAR to assist in the identification of the ESI that should be reviewed. Sample documents may be used to identify conceptually similar documents and to build a general understanding of the overall document collection. [45] Alternatively, finding documents that are conceptually

[44] The concept of Early Case Assessment (ECA) used herein is limited to the analysis of data content of documents.

[45] In cases involving a large volume of ESI, practitioners may first use unsupervised machine learning methods (e.g., clustering, concept search, near-duplicate detection, and visualization) early in the

dissimilar to sample nonrelevant documents can assist to identify documents that do not need further review.

ECA may also bring any missing ESI to the forefront early in a matter rather than later in the review process. This applies to both unexpected documents and to documents that were expected but are missing from the collection. For this reason, ECA tools can significantly assist in scope and cost containment.

C. PRIORITIZATION FOR REVIEW

TAR is an effective tool for prioritizing and organizing documents for attorneys to focus their initial discovery review, increase reviewer efficiency, assist with reviewer training at the start of the case (TAR 1.0), or facilitate consistency in human-coding decisions. Counsel has two ways to leverage TAR in this context: targeted review and full review.

Targeted review uses TAR on a subset of similar documents, which can be identified often from knowledge gleaned from any ECA on an ESI collection. Alternatively, unsupervised machine learning tools can be used to prioritize documents in a targeted review. Email threading identification, communication analysis, and topical clustering can group documents containing similar concepts into review batches.

A full human review may follow the completion of the TAR process. The documents identified by TAR as most likely to contain relevant content are prioritized and reviewed initially. Such prioritization can help inform early development of legal strategy.

Counsel can also exclude documents with a low relevance score from manual review by creating and reviewing only a sample set of these documents to verify that they are in fact not relevant.

D. CATEGORIZATION (BY ISSUES, FOR CONFIDENTIALITY OR PRIVACY)

TAR is an effective tool for categorizing documents. The most common workflow involves categorizing documents by relevance. But TAR can also be used to identify documents by specific category such as privileged, confidential, or "hot" documents, and by issues germane to the case. In these scenarios, the software is trained in the same way as when categorizing and ranking for relevance. However, reviewers might isolate as training exemplars discrete concepts, words or phrases, or

---

litigation so that they can gain objective insight into what the ESI collection includes. These early case assessment (ECA) tools analyze and index the content of electronic documents without any input by a human reviewer and separate the documents into conceptually similar groupings. The results often give insight into the ESI collection, particularly when examining ESI produced from opposing parties.

even excerpts from documents. These examples are provided to the software for training and then the categorization process identifies similar documents.

E. PRIVILEGE REVIEW

Privilege review is one area where existing permutations of TAR face significant challenges that may make them less valuable to clients and counsel. This is partly attributable to the fact that software analysis of text-based documents cannot reliably account for legal context and nuance. In addition, the danger of disclosing privileged documents, in contrast to merely non-responsive ones, influences the risk analysis.

Although TAR can sometimes play a role in privilege review, it is essential to understand the current significant limitations and risks of employing TAR in a defensible privilege review. First, the standards that apply to privilege are highly variable and subject to dispute among counsel. There are a variety of privileges that protect information from disclosure, each with specific legal standards. While TAR can determine the topic of a document, the topic alone may not determine whether the document is privileged. Second, privileged information in a document may have little traditional indicia signaling that the information might be privileged. In fact, the same exact content may be privileged in one document and not in another. Third, the content of a document alone does not determine whether a document meets the legal standards governing privilege. Privileged documents are often about the same topic as nonprivileged documents, but many TAR tools would tend to categorize together documents about the same topic. There are myriad other factors impacting that determination.

Current TAR processes may not overcome these challenges. The richness of privileged materials in most cases will be relatively low, which presents a challenge for any review. Moreover, recall rates will likely be unsatisfactory for privilege review. Given that attorneys on the same review team may strongly disagree about whether a document is privileged, it is not surprising that software struggle to properly categorize documents as privileged or not privileged. The software cannot account for the events surrounding the creation or dissemination of a document that might render an otherwise privileged document not privileged.

Employing TAR in privilege review can sometimes be helpful in terms of timing, prioritization of review, and coding consistency. Any discussion regarding the use of TAR for privilege review, however, should begin with understanding the nature of the privileged documents and the client's concerns regarding the documents.

Depending on the circumstances, in some cases, it may be appropriate to use TAR in conjunction with human/linear review prior to production of any documents. In some cases, linear review can be skipped at the initial stages to expedite production

of non-privileged documents, leaving for later the work required for redactions, privilege logging, and claw-backs/downgrades.

No matter the decision regarding whether and when to utilize TAR, strong claw-back agreements or provisions should be negotiated and in place prior to any production to foreclose a waiver argument.

F. QUALITY CONTROL AND QUALITY ASSURANCE

TAR can be effectively used for quality control (QC): (1) during document review to assess reviewer accuracy; (2) as a quality assurance checkpoint at the completion of a specific review phase; (3) during the production preparation phase; and (4) to complement other privilege screens.

TAR can be used effectively to assess the review team's coding accuracy by comparing the coding decisions from the human document review with the categorization or ranking scores assigned by the algorithm and revisiting documents when discrepancies exist. Depending on the number of discrepancies identified and time or budget restraints, QC is typically limited to the discrepancies identified at the very top and bottom of the ranked relevance scores. Based on the results, the review manager can adjust the coding protocol, perform supplemental training, or reassign members of the review team.

TAR can also be used to assess the overall quality of human coding at a specific review phase, such as first-pass review. This approach helps to measure the overall quality of the work product created by the human review, which is especially critical when documents identified as relevant during one review phase need to move to a second phase based on the decisions applied. This approach can identify relevant documents that were wrongly tagged as nonrelevant. It can also find sets of documents that were tagged relevant in the first-pass review but for which the additional cost of second-pass review is not warranted; once identified, these documents can be addressed together.

Finally, once review coding is complete, rankings can be applied during the production-preparation phase as a final check to identify and correct coding discrepancies and as an additional privilege screen to ensure only the intended documents are disclosed to the requesting party.

G. REVIEW OF INCOMING PRODUCTIONS

TAR is an efficient tool for a responding party to identify and produce relevant documents from large sources of ESI. But it is increasingly used by requesting parties to efficiently review and analyze voluminous document productions.

TAR offers the ability to streamline a review of a "data dump" and zero in on relevant data quickly based on documents that are relevant or key, to the extent they have already been identified, or by sampling and review of the production to develop relevant and nonrelevant coding decisions to train the TAR software.

The goal of reviewing incoming productions is to prepare key evidence and understand noteworthy content, including developing timelines, assessing case strengths and weaknesses, and understanding witness knowledge. TAR can aid those goals with issue categorization and key document analysis. TAR can also be used to effectively demonstrate gaps and to aid in motion practice (e.g., evaluate the sufficiency of the incoming production and potential spoliation issues, and perform a responsiveness analysis to evaluate whether the opposing party produced a data dump replete with non-responsive content).

Using TAR on incoming productions generates very little controversy in terms of disclosure, transparency, and opposing party challenges. Accordingly, techniques used for incoming data can be broader in scope than produced data and need not be limited by the terms of governing ESI stipulations. Many approaches are available for review of incoming productions, with the order and combination dictated by team preference, type of produced data, and importance of the produced data. TAR may be used, for example, to categorize or rank documents by relevance or issues in the matter.[46]

H. DEPOSITION/TRIAL PREPARATION

TAR can be a powerful and effective tool to identify key documents for witness interviews, depositions, and trial. Historically, the process of identifying key documents to conduct substantive witness interviews or to examine or defend a witness at deposition or trial involved search term and linear reviews. The review would normally start with the witness's custodial file followed by a broader search among other documents collected or produced in the case. This process was time-consuming, resource intensive, and susceptible to missing key information if a witness used "code" terms.

TAR offers some significant advantages over key word searching and linear review. Categorization and ranking allow counsel to identify key documents and issues that apply to a specific witness. Categorization aids understanding of the types of documents in the dataset and of key dates that can be converted into an interview outline.

[46] Unsupervised machine learning tools might also be used on incoming ESI productions to cluster documents for case analysis or to identify key documents or good example documents for TAR analysis. Alternatively, it may be useful to perform communication analysis using email threading.

TAR also equips attorneys to prepare for deposition and trial witnesses by re-purposing previously reviewed documents under any review model. It allows a more comprehensive analysis of documents that the witness may face or need to be questioned on during deposition or trial (this is particularly true for Rule 30(b)(6) witnesses) and should focus on finding and categorizing documents that counsel has already determined to be critical for a witness. TAR accelerates the speed and accuracy of this process over a larger number of documents and can identify holes in a key document collection or witness narrative.

Even with these advantages, it is important to recognize that TAR may not identify every document key to an individual witness. Not all ESI can be categorized by TAR, and the success of TAR is dependent on the content of the documents and the quality of the training and QC rounds. Still, TAR can help counsel prepare for more effective witness interviews earlier in a litigation or investigation and significantly improve the speed at which witness preparation materials can be assembled.

I. INFORMATION GOVERNANCE AND DATA DISPOSITION

TAR can be used to help manage organizations' ever-growing volume of electronic information. Although machine learning can be incorporated into enterprise-content-management software, TAR can be leveraged to perform episodic electronic discovery and knowledge management tasks, including the identification, preservation, or disposition of discovery data.

Among other things, TAR tools can prove valuable in:

- Identifying data that is subject to retention under an organization's information management policy;
- Assessing legacy data that may be appropriate for defensible deletion;
- Segregating data that contains protected information, such as personally identifiable information ("PII"), medical information, or other information subject to privacy protections;
- Capturing corporate records that may contain contractual obligations or confidentiality clauses, including third-party notification provisions; and
- Isolating potentially privileged, proprietary, or business-sensitive content (e.g., intellectual property, product development, or merger and acquisition data).

The same techniques and approaches for leveraging TAR in electronic discovery apply when using TAR for information governance. Consideration should be given to litigation-hold requirements, regulatory records-retention requirements, internal audit and compliance needs, other legal obligations (such as contractual requirements), and business operational needs.

1. Records Management Baseline

A records-retention schedule provides a good starting point for using TAR for information governance. Many records management systems are rules-based, categorizing records according to defined characteristics. Applying TAR to information governance practices relies on users training the computer to identify specific records or content that need to be isolated and preserved. Exemplar documents can be identified through Boolean searches or targeted sample collection, and additional exemplars can be found by running conceptual clusters and sampling documents that reside in the same cluster as the identified exemplars. TAR offers the added benefit of being language-agnostic, aiding in challenges associated with search term translation and language identification. Given the wide variety of records covered by corporate records retention schedules, a categorization approach can prove more promising than a relevant/not relevant approach.

2. Assessing Legacy Data – Data Disposition Reviews

TAR may be used to manage legacy data, including backup tape contents and "orphaned" data associated with departed employees. Legacy data is often viewed with an eye to preserving only: (i) data subject to a regulatory preservation requirement or legal hold; (ii) data subject to records-retention requirements; or (iii) data that has lasting IP or strategic value to the company. When using TAR for these purposes, it is advisable to use a layered, multi-featured approach (i.e., using both TAR and other search strategies), combined with rigorous statistical sampling, to ensure adequate capture before data is potentially destroyed.

3. Isolating Sensitive Content – PII/PHI/Medical/Privacy/ Confidential/ Privileged/Proprietary Data

As with the approach to records management, isolating protected or sensitive material often begins with basic search strategies, including pattern-based searching to identify credit card numbers, bank accounts, dates of birth, and other content that follows a regular pattern. Once good exemplars are identified, TAR can be leveraged to identify documents with similar content.

As with all machine learning, it is important to perform a file analysis at the outset to isolate documents that might not be readily susceptible to TAR, including handwritten or numerical documents that are likely to contain protected information. Once the data is identified, it can be segregated for appropriate treatment, which may include limited-access secure storage, redaction, or structured content management.

CHAPTER FOUR
FACTORS TO CONSIDER WHEN DECIDING WHETHER TO USE TAR

A. INTRODUCTION

In any particular matter that involves document classification, questions can arise early regarding appropriate tasks for TAR as well as the factors that might enhance or diminish its value in a particular case. In other words, should the legal team use TAR, and if so, which TAR review process? While the following sections provide insight into how to assess these questions, it is not an exhaustive analysis. Any use case must be analyzed by the facts and circumstances facing the legal team.

B. SHOULD THE LEGAL TEAM USE TAR?

The threshold question of whether TAR should be used can be answered by understanding a few key decision points, which relate to an assessment of the cost and risk threshold in a particular case:

- Are the documents appropriate for TAR?
- Are the cost and use reasonable?
- Is the timing of the task/matter schedule feasible?
- If applicable, is the opposing party reasonable and cooperative?
- If applicable, are there considerations related to the forum or venue in which the case is based that influence the decision?

1. ARE THE DOCUMENTS APPROPRIATE FOR TAR?

The types of documents that potentially will be subject to TAR is an important factor to consider when deciding whether to use TAR. TAR software requires text to work, and thus is at least somewhat dependent on the semantic content of the

document population being analyzed. However, documents with no text, too little text (not enough meaningful content), or too much text (too much content to analyze) should be set aside, because the software will not be able to classify them well, or at all. With this in mind, TAR can generally be very effective at sorting through a custodian's email files, but may not be an effective tool for organizing or categorizing the spreadsheets or videos attached to those same emails.

TAR is most effective when applied to text based, user-generated documents. This often includes emails, electronic documents such as Word files and searchable PDFs, presentation slides, etc. Such documents have sufficient semantic content for the software to effectively analyze their characteristics and find meaningful patterns it can apply to other documents in the population. Documents with little text or substantive discussion (such as an email that says nothing more than "Please see attached") lack sufficient content and cannot be effectively analyzed. These documents generally require more training examples or may not be correctly categorized, if they are categorized at all.

TAR may not work well with the following additional data types:

- 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 invitation;
- Hard copy documents with less-than-desirable OCR results (although results may be better than results using other search methodologies, like keyword searching);
- Audio/video/image files, which generally lack any semantic content;
- Foreign language/ESL documents, which can be analyzed by TAR, but may require separate training sets for each language (N.B., mixed-language documents may cause additional issues); or
- Structured or semi-structured data stored in databases, such as Mobile Data/Chat/MSM/ Social Media/IoT/Real Big Data.

2. ARE THE COST AND USE REASONABLE?

Legal teams typically engage in a cost and risk–benefit analysis when deciding whether to use TAR or conduct a full manual review of documents. The team must be cognizant of the costs of access and use of the TAR software. For one, the volume of documents at issue should be considered. If the volume of documents is small, the cost of use[47] may be higher than if a different review method is used to identify

[47] There are unique costs associated with using TAR, including cost of access to the TAR application, development and implementation of workflow, and project management time. There are also other cost risks identified further in this Chapter below.

relevant documents. In addition, with a very small document collection, the risk that the collection is very rich (mostly relevant documents) also may negate the value of the use of TAR.

3. IS THE TIMING OF THE TASK / MATTER SCHEDULE FEASIBLE?

As seen in Chapter 2, TAR is a more complex review than just releasing all documents to a review team for review. As such, TAR is not an overnight process. A TAR review may involve an initial delay in days or weeks as the documents are indexed, the workflow is finalized, and the TAR process is set-up. Document production deadlines or deposition schedules may impact the decision to use TAR. For example, if two custodians will be deposed within two weeks of the first document production, the team would need to ensure at a minimum that those two custodians' documents are targeted for TAR. If those two custodians' documents have not been fully collected, supplemental collections and their impact on TAR must be considered.

4. IS THE OPPOSING PARTY REASONABLE AND COOPERATIVE?

This issue only relates to the use of TAR for relevancy determinations, when disclosure concerns exist. A key consideration here is whether the opposing party will be reasonable and cooperative in the use of TAR. Engaging in a protracted battle with opposing counsel, spending time educating an adversary about TAR, or involving the court may not be worth the cost savings otherwise afforded by TAR. With respect to government agencies, a party relying on TAR to respond to subpoenas and requests for information must carefully abide by the agency's requirements and policies. 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 party responding to the request must submit a written disclosure of the particulars of the proposed process.

5. ARE THERE JURISDICTIONAL CONSIDERATIONS THAT INFLUENCE THE DECISION?

Like Item 4 above, this issue only relates to using TAR for relevancy determinations. Just like the bar at large, the bench's support and understanding of TAR vary. There is insufficient guidance and conflicting case law relating to the extent of disclosure about TAR the producing party must provide to the requesting party. When disclosure occurs, and parties cannot agree about the TAR process to be used, a judge may need to determine whether the producing party's proposal is reasonable. This is a risk for a producing party, as judges vary in their familiarity and views on different approaches to TAR. For these reasons, the forum's approach

to disclosure must be considered in calculating the overall risks and costs associated with TAR.

C. THE COST OF TAR VS. TRADITIONAL LINEAR REVIEW

TAR can be a faster and cheaper process than a traditional linear document review. There are several factors that impact the relative costs of TAR, however, and one's goals, workflow, and timeline are all pertinent.

Document review is often the single largest expense of litigation-related discovery – frequently estimated at 60% to 70% of the total cost. TAR can reduce costs dramatically, but the upfront costs incurred in the initial set-up and training expenses, as well as the relative uncertainty about the outcome and timeline, are important factors to consider. In addition, TAR does not eliminate the need for any document review, and users should not be surprised when costs shift away from the first level review to 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 such as data quality, the types of data included in the population, the breadth or complexity of relevance, the richness of the data, and the statistical thresholds applied, as well as costs associated with the service provider or software (this is frequently a per document or per gigabyte fee); hourly consulting or project management fees related to the TAR process (including possible expert fees) affect the overall cost of the TAR project and the cost savings realized in comparison to a linear review.

The costs of training the computer may be significantly reduced by re-using previously reviewed and coded documents. Known-relevant documents and files can often serve as training documents, 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.

D. THE COST OF TAR AND PROPORTIONALITY

Parties and courts are wrestling with addressing the question of whether and how TAR can impact proportionality.

To that end, even though the responding/producing party is generally considered to be best-positioned to evaluate the best way to identify and produce requested materials, a court in the future may, under a Rule 26 proportionality analysis, question a party's decision not to use TAR, when substantial cost savings and effectiveness appear clear.

APPENDIX

KEY TERMS

- CONFIDENCE INTERVAL (MARGIN OF ERROR) AND CONFIDENCE LEVEL. The confidence interval and confidence level characterize the certainty of the point estimate.[48] For example, the recall point estimate of 80% can be combined with a margin of error of 5%, allowing for a confidence interval of 75% (5% below 80%) and 85% (5% above 80%). Moreover, a confidence interval is meaningful only if accompanied by a confidence level, which is a measure of how likely the sample is to represent the true set, within the confidence interval. Continuing the example used here, a confidence level of 95% means that 95 times out of 100, the result of the sample will include the actual recall within its confidence interval. Put another way, there is a 95% chance that the actual recall is between 75% and 85%.

- CONTROL SET. A control set is a random sample taken from the entire TAR set that acts as a relevancy truth set against which the computer's decisions can be judged. It is used to estimate the computer's effectiveness in classifying documents during TAR. It may also be used to estimate the richness of the TAR set. Not all workflows use a control set.

- ELUSION. Elusion estimates how many relevant documents were missed and are in the nonrelevant set. In the example used below in the recall definition, the computer identified 800,000 documents as potentially nonrelevant. Because there are a total of 100,000 relevant documents and 80,000 documents were identified within the 100,000 potentially relevant documents, 20,000 relevant documents were missed. The elusion of the TAR predictive model is therefore 20,000 / 800,000 = 0.025 or 2.5%.

- ESTIMATE OR ESTIMATION. Knowing the exact value of an effectiveness measure (such as recall) would require knowing the true relevancy status of every document in the TAR set. In practice, therefore, one must estimate the effectiveness using sampling techniques. These estimates allow for a statistical certainty that the estimated values are close to the true value.

- PRECISION. Precision measures the percentage of documents that are truly relevant among all the documents the computer identified as potentially relevant. Using the example in the recall definition, the computer identified 200,000

[48] POINT ESTIMATE. A point estimate is an estimate that is a single value. Based on the recall definition example below, the point estimate for recall is the single value of 0.80 (80%), since the computer correctly identified 80,000 of the 100,000 total relevant documents. However, as provided in the confidence interval and level definitions, a point estimate alone is of limited use, and therefore should be combined with how confident we are in the point estimate.

documents as potentially relevant, of which 80,000 were identified as relevant by human-review, resulting in a precision of 40% (80,000/200,000).

- PREDICTED NONRELEVANT SET. The predicted nonrelevant set is a subset of documents in the TAR set. It contains those documents in the TAR set that are predicted as nonrelevant by the software and thus would be excluded from further review or production efforts workflow.[49]

- PREDICTED RELEVANT SET. The predicted relevant set is a subset of documents in a TAR review set. These are the documents that the software identifies as relevant and subject to potential production as a result of the TAR process. No matter what form of TAR used, the identification of the potential production set is derived from the software's predictions on what is relevant and nonrelevant. In some workflows, the entire predicted relevant set is reviewed by humans during the TAR training process. For example, in common TAR 2.0 workflows, the software is only trying to return relevant documents to the humans, and the humans review all the documents returned by the computer as predicted relevant. In other workflows, the predicted relevant set is not reviewed in its entirety during the TAR training process. For instance, in common TAR 1.0 workflows, the TAR process is designed to build a predictive model to make relevancy calls on the entire TAR set, and after TAR is complete, the human review team makes the decision to review the entire relevant review set or to simply accept the software's relevancy decisions. In any event, documents originally predicted to be relevant can be subsequently reviewed and determined actually to be relevant or nonrelevant under both TAR 1.0 or TAR 2.0 workflows. Despite no longer being a "prediction" at that point, these documents continue to be classified as part of the "predictive relevant set" to eliminate confusion that would otherwise arise.

  With this in mind, it is important to note that, like manual reviews, TAR classifications are not perfect. The "predicted relevant set" will not contain all the relevant documents from the TAR set: its recall will not be 100%. Nor will it contain only relevant documents: its precision will not be 100%. Any documents in the predicted relevant set that are subsequently determined to be nonrelevant by a human reviewer can always be excluded from production (insofar as they are not part of a family that includes relevant documents).

[49] Just as there will be nonrelevant documents in the predicted relevant set, there will be some estimated number of relevant documents in the "predicted nonrelevant set." But for simplicity purposes, we identify this as the predicted nonrelevant set because most of these documents have been identified by the computer as nonrelevant, and thus will be excluded from further human review.

- RECALL. Recall measures the percentage of documents found to be relevant. Consider a workflow in which a TAR set of one million documents are collected, of which 100,000 are relevant.[50] The computer identifies 200,000 documents as potentially relevant and 800,000 documents as potentially nonrelevant. A human review of the 200,000 potentially relevant documents shows that 80,000 are relevant. Therefore, the effectiveness of the classification system, when measured using recall is 80%, since the computer identified 80,000 of the 100,000 relevant documents. The producing party may represent that their workflow achieved an 80% recall, i.e., the documents being produced represent 80% of the relevant population prior to any possible privilege review.

- REVIEW QUALITY CONTROL. During a document review, the team may engage in quality control efforts to ensure the human reviewer's and computer's relevancy decisions are as accurate as reasonably possible.

- RICHNESS. Richness is the estimated proportion of documents in a data set that are relevant. For example, if a set of one million documents contains 100,000 relevant documents, it has 10% richness. Richness is also known as prevalence.

- TAR SET. This is the total set of documents that the workflow (the document review) will be conducted on.

- TRAINING SET. The training set is the subset of documents in the TAR set that the human reviewer reviews to train the software on what is relevant. The training set will contain relevant and nonrelevant documents. The software uses the training set to produce a predictive model, and the predictive model will be used to define the predicted relevant set. The number of relevant and nonrelevant documents necessary to produce a predictive model with good effectiveness will depend on the nature of the documents in the TAR set, the difficulty of the relevance definition, and the particular TAR software and method used.

[50] In order to estimate recall, the total number of relevant documents in the TAR set must be known. Because the only way of identifying the total number of relevant documents in a set is to review the entire TAR set, the total number of relevant documents must also be estimated.

****

THANK YOU TO OUR SPONSORS!

Bolch Judicial Institute at Duke Law gratefully acknowledges the financial support of its EDRM sponsors.

FOUNDING SPONSORS

Dechert
Exterro
Knovos

GOLD SPONSORS

AccessData
BDO
Relativity
TCDI
UnitedLex

SILVER SPONSORS

CloudNine
Everlaw
Fronteo
FTI Consulting
KLDiscovery
Morae Global Corporation
NightOwl Discovery
Nuix
Zapproved