**CALL FOR PUBLIC COMMENT**

**(comments due no later than july 16, 2018)**

TECHNOLOGY ASSISTED REVIEW (TAR)

GUIDELINES

EDRM/DUKE LAW SCHOOL

EDRM/Duke Law School

March 18, 2018

**Foreward**†

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 for the first time in EDRM’s new home on 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 and draft “best practices” using TAR. An initial draft of the best practices is expected in summer 2018. 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 legitimizes and supports 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 during 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.

After the expiration of the public-comment review, the teams will make appropriate revisions. The approved document will be posted on the Institute’s web site and made available to the bench and bar.

James Waldron

Director, EDRM

John Rabiej, Deputy Director

Bolch Judicial Institute

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

† Copyright © 2018, All Rights Reserved. This document does not necessarily reflect the views of Duke Law School or its faculty, or any other organization including the Judicial Conference of the United States or any other government unit.

**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  Stroock & Stroock & Lavan | 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  DiscoverReady | Bob Rohlf  Exterro | Herbert Roitblat  Mimecast | John Rosenthal  Winston & Strawn |
| Justin Scranton  Consilio | Dharmesh Shingala  Knovos | Michael Shortnacy  Sidley Austin | Mimi Singh  Evolver Inc. |
| **Contributors (Cont.)** | | | |
| Clara Skorstad  Kilpatrick Townsend | Harsh Sutaria  Knovos | Tiana Van Dyk  Burnet Duckworth & Palmer | Patricia Wallace  Murphy & McGonigle |
|  | Ian Wilson  Servient | Carolyn Young  DiscoverReady |  |
|  |  |  |  |

We thank Patrick Bradley, Leah Brenner, Matthew Eible, and Calypso Taylor the four Duke Law, Bolch Judicial Institute Fellows, who edited, proofread, cite checked, and provided valuable comments and criticisms. In particular, we gratefully acknowledge the editing suggestions of Tim Opsitnick (TCDI) and James Francis, United States magistrate judge (retired), which markedly improved the document’s clarity.

The feedback of the judiciary has been invaluable in identifying best practices, 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 May 18, 2018

**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 automate 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, they 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, the computer is faster, more consistent, and more cost effective than Human Review teams. 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 non-relevant 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 imperfect method. Problems with fatigue, human error, disparate attorney views regarding document substance, and even gamesmanship are all associated with manual document review. Multiple studies have shown a significant rate of discrepancy among reviewers who identify relevant documents by linear review — as much as 50%. The TAR process is similarly imperfect, but studies show that the use of TAR is equally as accurate, if not more accurate, than humans performing document-by-document review.

Importantly, no reported court decision has found the use of TAR invalid. Scores of decisions have permitted the use of TAR, and a handful have even encouraged its use.

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 in discovery, the bench and bar must become more familiar with it, and certain standards of validity and reliability must be met to ensure its accuracy.

Validity means that the documents that a TAR process says are relevant, after using any particular TAR engine to implement that process, actually *are* relevant. Reliability means that the TAR process is consistent and “like” documents are categorized similarly. These guidelines will not only demonstrate the validity and reliability of TAR but will also demystify the process.

**TECHNOLOGY ASSISTED REVIEW (TAR) GUIDELINES**

**EDRM/DUKE**

**CHAPTER ONE**

**DEFINING TECHNOLOGY ASSISTED REVIEW**

|  |
| --- |
| **Chapter One**  **Defining Technology Assisted Review** |
| 1. Introduction…………………………………………………………………………. 1 |
| 1. The Tar Process……………………………………………………….…………….. 2 |
| 1. Assembling The TAR Team…………………………………………………….. 2 2. Collection And Analysis……………………………………………………… 2 3. “Training” The Computer Using Software To Predict Relevancy…... 3 4. Quality Control And Testing ……………………………………………….. 3 5. Training Completion And Validation………………………..…....……….. 3 |

* 1. Introduction

Technology assisted review (referred to as “TAR,” and also called predictive coding, computer assisted review, or machine learning) is a review process in which humans work with software (“computer”) to teach it to identify relevant documents.[[1]](#footnote-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[[2]](#footnote-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 each document 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 to those likely to be relevant, depending on the circumstances.

* 1. 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”[[3]](#footnote-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 vendor 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.[[4]](#footnote-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 vendor; workflow expert; case manager; lead attorney; and human reviewer. Chapter Two contains details on the roles and responsibilities of these members.

1. Collection and Analysis

TAR starts with the team identifying the universe of electronic documents to be reviewed. The case manager inputs the documents into the software to build an analytical index. During the indexing process, the software’s algorithms[[5]](#footnote-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.

1. “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, thorough client interviews, by creating 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 most helpful to refine its classifications 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 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.

1. Quality Control and Testing

Quality control and testing are essential parts of TAR, which ensure 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. The most 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, by how many documents have moved up and down in their rankings, or by measuring and tracking the known relevant documents until the algorithm suggests that few if any relevant documents remain in the collection. Yet other methods involve labeling 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.

1. 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 the maximum number of relevant documents possible 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 maximized its ability to find 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. Under the training process in software commonly marketed as TAR 1.0,[[6]](#footnote-7) the software is trained based upon a review and coding of a subset of the document collection that is reflective of the entire collection (representative of both the relevant and non-relevant documents in the population), with a resulting predictive model that is applied to all nonreviewed documents.  The predictive model is updated after each round of training until the model is reasonably accurate in identifying relevant and nonrelevant documents, i.e., reached a stabilization point, to be applied to the unreviewed population.  This stability point is often measured through the use of a control set, which is a random sample taken from the entire TAR set, typically at the beginning of training, and can be seen as representative of the entire review 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, this is seen as a point of reaching training stability. At that point, the predictive model’s relevancy decisions are applied to the unreviewed documents.

Under software commonly marketed as TAR 2.0, the human review and software training process is melded together. The software from the outset continuously searches the entire document collection and identifies the most likely relevant documents for review by a human. After each training document’s human coding is submitted to software, the software re-categorizes the entire set of unreviewed documents, and then presents back to the human only those documents that it predicts as relevant. This process continues until the number of relevant documents identified by the software after human feedback becomes small. At this point, the TAR team determines whether stabilization has been reached or whether additional re-categorization (i.e., more training) is reasonable or proportional to the needs of the case.

Before the advent of TAR, parties did not provide statistical evidence evaluating the results of their discovery. Only on a showing that the discovery 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 compared to human-based reviews. As the bench and bar have become more familiar with TAR and the science behind it, the need to substantiate TAR’s legitimacy in every case has diminished.[[7]](#footnote-8)

Nonetheless, because the current state of TAR protocols and the case law on the topic is limited, statistical estimates to validate review continue to be discussed. Accordingly, it is important to understand the commonly cited statistical metrics and 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, and how many are estimated to be nonrelevant, and how confident is the TAR team in those estimates?
* As a result of the workflow, how many estimated relevant documents did the team identify, and how confident is the team in that estimate?
* How did the team know 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? 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 definition of what level of accuracy is sufficient 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 topic will be covered in more detail later in these guidelines.

**CHAPTER TWO**

**TAR WORKFLOW**

|  |
| --- |
| * 1. Introduction……………………………………………………..…………………… 7 |
| * 1. Foundational Concepts & Understandings…………..………………………..8 |
| * + 1. Key Terms…………………………………………..............................................8 |
| * + 1. TAR Software: Algorithms……………..………………...............................8 |
| * + - 1. Feature Extraction Algorithms…………………….…………………...8       2. Supervised Learning Algorithms (Supervised Learning Methods) ………………………………………………………………..………...9       3. Varying Industry Terminology Related To Various Supervised Learning Methods ……………………………………………..…………........9 |
| * 1. The Tar Workflow ……………………………………………………..………….. ..9      1. Identify The Team To Engage In The Tar Workflow………..…..…..... 10      2. Select The Service Provider And Software…………………..……….... 11      3. Identify, Analyze, And Prepare The Tar Set……………...……..…….... 12         1. Timing And The Tar Workflow…………………………………...…...… 13      4. The Human Reviewer Prepares For Engaging In Tar……………..…….14      5. Human Reviewer Trains The Computer To Detect Relevancy, And The Computer Classifies The Tar Set Documents…………..…………..15      6. Implement Review Quality Control Measures During Training….…………………...………………………………………. ……….... 18         1. Decision Log…..……………………………………..…………………….. 18         2. Sampling……………………..………………………..………………...…….18         3. Reports………………………………………………….…..………………. 19      7. Determine When Computer Training Is Complete And Validate………………………………………………………………….………..19         1. Training Completion………………………………………………....…….20            1. Tracking Of Sample-Based Effective-Ness Estimate……………………………………………………………….......20            2. Observing Sparseness Of Relevant Documents Returned By The Computer During Active Learning……….………..………… 20            3. Comparison Of Classifier Behaviors…………..…………..……….21            4. Comparing Traditional Tar 1.0 And Tar 2.0 Training Completion Processes………………………………………....…….. . 21         2. Validation………………………………………………………………….....22      8. Final Identification, Review And Production Of The Predicted Relevant Set ………………………………….…..…………………………….. 24         1. Workflow Issue Spotting………….……….…………..……………… 25         2. Extremely Low Or High Richness Of The Tar Set………...………...25         3. Supplemental Collections…………………….………......................... 25         4. Changing Scope Of Relevancy………………...………….……………..26         5. Unreasonable Training Results……………...………………… ……...27 |

1. 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 is complete, documents will be classified into two piles: the predicted relevant set (i.e., those documents that will be subjected to potential production) and the predicted nonrelevant set (i.e., those documents that will not be subjected to potential production).

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 is omitted from human review based on the classification decisions made by the computer.[[8]](#footnote-9) 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 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 smaller (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 generally make relevance determinations as accurately as human review teams, provided that a reasonable workflow is applied to suitable data.

The essential components of the TAR workflow, to date,[[9]](#footnote-10) can work to fully meet Fed. R. Civ. P. 26 discovery obligations (and their state equivalents) with as little cost and judicial time as possible. A party should consider the components when formulating a final workflow to satisfy Rule 26 obligations.[[10]](#footnote-11) 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 consideration are identified throughout the workflow to help explain the differences among software.

* 1. Foundational Concepts & Understandings
     1. Key Terms

Part of the confusion regarding TAR, types of TAR, and TAR concepts involve a “failure to communicate.” To avoid confusion and miscommunication, it is important to explain basic definitions and concepts relevant to the discussion.[[11]](#footnote-12) Definitions of the key terms can be found in Appendix A.

* + 1. 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 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 responsiveness).

* + - 1. 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.[[12]](#footnote-13) After analyzing all documents in 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.

* + - 1. Supervised 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 learning algorithm allows a human reviewer to teach 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 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 attempts to categorize or rank documents in the TAR set.[[13]](#footnote-14) 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 learning methods allow for a training process that is iterative and interactive whereby the human reviewer and software provide feedback to each other to improve the software’s ability to review 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).

* + - 1. Varying Industry Terminology Related to Various Supervised Learning Methods

All supervised learning methods used by the various software utilize iterative training processes: multiple rounds of a human reviewer coding documents and submitting them to the TAR software to fine-tune the software’s decision-making abilities, and thus to increase the software’s accuracy and consistency in returning relevant documents. This is also known as building a predictive model or “classifier.”

* 1. 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 vendor, or TAR software provider. Selection of the service provider is discussed in Section C (2).
* Software Vendor. The software vendor is the creator of the software. Some service providers create their own software (and, thus, are also the software vendor), while others license it from software vendors.
* 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 those who are 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 and classifies documents for relevancy in the training set, and this classification is used to train the software on what is relevant. As such, every human reviewer must be educated on the scope of relevancy to ensure reasonably accurate and consistent training of the software.
  + 1. Select the Service provider and Software[[14]](#footnote-15)

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 the service provider that will be assisting or managing the process. The producing party needs to do due diligence on service provider selection. The service provider should have an expert that 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 due diligence topics to consider discussing with the service provider are:

* Does the service provider have a written TAR guide?
* Which TAR software does the service provider have?
* Does the service provider have demonstrable, measurable verification 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 vendor periodically engage in upgrades, updates, bug fixes to improve the software and workflow?)
  + 1. 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 define the “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 relevant custodians/document repositories, date ranges, and file types, and may also involve search terms.[[15]](#footnote-16) 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, the human reviewer 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.[[16]](#footnote-17) 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 review and learn.

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 greater than 30 megabytes of text.[[17]](#footnote-18) Some TAR software have the capacity to analyze and identify problematic documents. Ultimately, the parameters used to exclude documents from the TAR set should be discussed with the service provider. 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.[[18]](#footnote-19) 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 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). In other words, for each document, the software does not recognize each as an email, Word file, Excel, or PowerPoint presentation; rather, it just recognizes the text drawn from each of these native file types. After the build is completed, the team is ready to engage in teaching the software what is relevant to the case.

* + - 1. 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 necessary.

Both TAR 1.0 and TAR 2.0 have timing considerations that must be factored in to the project timeline. Under a traditional TAR 1.0 approach, the machine learning process must be factored in when determining the length of time it will take to complete the review. 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. Documents that should be identified for separate analysis are documents, as described above, with little or no conceptual content.

Under a traditional TAR 2.0 approach, documents are prioritized for review based on a continuous update of the relevancy 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 vendor’s production turnaround time, from approval of a production submission to the time the production is available for delivery, and account for this time in the project schedule. The size of the data being produced will have implications on how the production is received by counsel (e.g. hard drive or FTP) and may affect when the production will be submitted. Creating a project schedule from the outset is the pathway to a successful project as it focuses attention on potential workflows and the establishment of deadlines, which provides clarity and sets expectations from the start of the review.

* + 1. 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. To that end, the scope of relevancy must be defined, and the lead attorney must train any other human reviewer on that scope. However, many times it is difficult to get a final scope of relevancy in place during 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 Rule 34 discovery requests, which may cause delays (sometimes substantial) in the final agreement or order on the scope of discovery for the matter. There may also be motion practice, including motions to dismiss claims, which can affect the timing of when the scope of discovery is solidified. Ultimately, the lead attorney must reach a point where they are comfortable defining the scope of relevancy to be applied to the workflow and the team should discuss any negative consequences of engaging in the workflow prior to a reasonably defined discovery scope.

Once the lead attorney has comfort about 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**: In most instances, 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 teach 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**: 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 teach 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 teach the computer 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 responsiveness. More 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.

* + 1. 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 their 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:

1. Identify training documents (selection of documents for human reviewer review);

2. Have the human reviewer code the training documents for relevancy;

3. Have the human reviewer’s relevancy decisions submitted to the software;

4. Have the software use the training documents’ relevancy decisions to build a predictive model, and apply the model to rank or classify all documents in the TAR set; and

5. Repeat steps (1)-(4) (add more documents to the training set) until further training is no longer needed because review’s goals have been met.

**Workflow Consideration**: Supervised learning methods vary[[19]](#footnote-20) in how they select training examples for the human reviewer. 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 marketed as active learning).[[20]](#footnote-21) 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 may find examples of relevant and nonrelevant documents that exist in the TAR set to feed to 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.[[21]](#footnote-22) 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[[22]](#footnote-23), which allows the computer to provide feedback based upon its categorization of documents after each round of training. The computer’s choice of what 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 is called “relevance feedback,” whereby the software attempts to identify for training only those documents that are most likely to be relevant. Other methods 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.[[23]](#footnote-24)

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 classifier if it is used to produce a large enough training set. Thus, differing views over selection of training data are less about whether an effective classifier can be produced, than about how much work it will take to do so.

Some general tendencies 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 lead to a rich set of documents being reviewed (high percentage of relevant documents in the TAR set), but it may require more time and money to ensure enough relevant samples are identified that span the entire scope of relevancy to be used to teach 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 completely independent of human judgment (including choices made in selecting training documents in previous batches) and gives every document in the TAR set an equal chance to be selected. It therefore 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.
* Computer feedback/active learning methods are also an efficient and automated way of choosing training documents, though some require substantially more computer time than simple random sampling. Computer feedback methods may require less training data as compared with random selection to produce an effective classifier, 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.
  + 1. 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.[[24]](#footnote-25) 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.

Successful review quality control involves not only correcting relevancy-call errors, but also continued educating of the human reviewer, which in turn will improve the quality of review from which the software learns. 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 measures are discussed below.

* + - 1. 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.

* + - 1. 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 (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 predicts 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 classifier.

* + - 1. 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 responsive and the human reviewer coded as nonresponsive; or (2) the software considers as likely nonresponsive and the human reviewer coded as responsive. 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.

* + 1. Determine When Computer Training Is Complete and Validate

In a TAR workflow, a major decision is when to stop the training process (when does the team determine that the software has identified an adequate percentage of relevant documents). In practice, this usually means trying to quantify 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.

* + - 1. 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 the stability of the training process. Stability does not have a precise technical definition. Roughly speaking, however, the training process is stable if review of further training documents is unlikely to substantially improve the effectiveness or cost-effectiveness of the trained classifier. Some software directly provide stability measures, and other indicators of stability can be derived by the team from estimates of effectiveness.[[25]](#footnote-26) There are three broad approaches to understanding stability: (1) tracking of sample-based effectiveness estimates; (2) observing sparseness of relevant documents returned by the computer during active learning and (3) comparing the classifier behaviors.

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,[[26]](#footnote-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.[[27]](#footnote-28) When training no longer substantially improves the computer’s classifications, this is seen as a point of reaching training stability. 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 stability measurement is usually only taken in certain TAR 1.0 workflows.

Observing Sparseness of Relevant Documents Returned By the Computer During Active Learning

Another measurement of stability involves the human reviewer continuing to train the computer on relevancy until they reach the point where the number of relevant documents presented by the computer for human reviewer review is extremely low or none. When the human reviewer reaches this point, it is seen as an indicator of stability and the training is complete, as no additional training is anticipated to improve the classifier or identify more relevant documents. This stability measurement is usually only taken in certain TAR 2.0 workflows that only use relevancy feedback learning.

Comparison of Classifier Behaviors

Another approach to monitoring stability is to directly compare the behavior of classifiers produced on different training iterations. A wide variety of approaches is possible, and the details are not always revealed by the software provider. As an example, classifiers produced on different iterations could be used to rank, score, or classify the entire TAR set. If those predictions are static, then the team may be able to conclude that further training likely will have no benefit, because the behavior of the TAR classifier is not meaningfully changing at that point. The team may view this as the point of training stability and stop training.

Comparing Traditional TAR 1.0 and TAR 2.0 Training Completion Processes

The following example clarifies this variation between traditional TAR 1.0 and 2.0 processes.[[28]](#footnote-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 to achieve a recall of at least 80%, i.e., identify at least 16,000 of the 20,000 relevant documents. 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 and present only those documents it believes are relevant (i.e., engages in relevancy feedback).[[29]](#footnote-30) This process continues until stability is 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.

In the same example, some traditional TAR 1.0 start by building a control set of 1,800-3,500 documents that is used to measure stability.[[30]](#footnote-31) Then, the software might use 1,400-3,000 documents for training (or however many documents is enough to achieve stability or the reasonable identification of the predicted relevant set). The training documents are selected by the computer to identify both relevant and nonrelevant documents (contrast with the TAR 2.0 example above). These 1,400-3,000 documents are used to build the predictive model that is seen as an estimated representative of the review set. When stability is reached (the software is reasonably accurate at predicting the relevancy of the review set), the predictive model’s rankings are locked, and the final predictive model is used to identify the predicted relevant document set (at least 16,000, but will be more, as the TAR model predictions cannot perfectly select only relevant documents for review). At this point, the TAR team can determine what, if any, additional human review will take place on the unreviewed portion of the predicted relevant set.

* + - 1. 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 once the training has concluded. In most 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 inputs that feed into the sample size: richness, confidence level, and confidence interval. These inputs 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 completed 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 require 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 the training set, to achieve the estimated recall level. This makes reasonable and proportional review judgments difficult if the targeted recall is set too high, as it may require unreasonable and disproportionate human reviewer 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[[31]](#footnote-32)), 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).

At a particular point on the scale, there is an associated predicted relevant set that is identified, with a corresponding estimated recall percentage. This feature is found in TAR 1.0 and some TAR 2.0 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 recall during TAR.[[32]](#footnote-33) The first involves taking a simple random sample from the set of documents of interest (e.g., the TAR set), reviewing it, and then using only the relevant documents from this sample. These relevant documents constitute a random sample from the (unknown) population of relevant documents. By examining how this sample of relevant documents is treated by the review or the component, the recall of that review or component can be estimated.

The second method of estimating recall involves drawing a random sample only from the documents in the predicted nonrelevant set for the review. The sample is used to estimate the richness of the predicted nonrelevant set, and thus the number of relevant documents in that set. The team then counts the number of relevant documents that a review found in the predicted relevant set. Adding these two values gives an estimate of the total number of relevant documents in the set of interest, and thus allows an estimate of recall of the review.[[33]](#footnote-34)

In conjunction with this effort, the team can also assess how important those relevant documents located in the nonrelevant document sample set are to the claims and defenses in the matter. If the number of relevant documents found in the sample is higher than anticipated, or if the relevant documents found are important to the matter, the team will need determine whether additional training is needed to enable the computer to identify other relevant documents in the predicted nonrelevant set.

It is important to know that there are variations in how recall is estimated and to understand how the service provider or workflow expert calculates recall.

* + 1. Final Identification, Review, and Production of the Predicted Relevant Set

After the team trains and validates measurements, the TAR process will have separated the TAR set into the predicted relevant set and 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.[[34]](#footnote-35)

Finally, in every workflow leading up to a document production, steps should be taken to address family members, privileged documents, confidentiality, and other issues that fall outside of the workflow process. An attorney should assess whether to fully review, partially review, or simply produce any documents that have not yet had human review and that are in the predicted relevant set and family members. Considerations should include costs of further review, weighed against the risks of producing nonresponsive documents that could include confidential, private, or privileged[[35]](#footnote-36) client information.

* + - 1. Workflow Issue Spotting

There are certain challenges that may arise during the workflow, namely: (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.

* + - 1. 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 (for instance, 3 percent or less of TAR set is relevant) or extremely high richness (the vast majority of the documents in the TAR set, 85% or more, are relevant ). With respect to the latter, if the TAR set is mainly relevant documents, it may mean that developing, implementing, and engaging in the workflow will be a waste of time, money, and resources, because it is anticipated that almost the entire TAR set is relevant, thus defeating the purposes of using TAR to assist in reviewing documents for relevancy. The team may want to consider other measures to prioritize or organize the TAR set for review.

With respect to the former challenge of low richness, the human reviewers may have a difficult time training the software on what is relevant, as 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. Low richness TAR sets may require more time to teach the software about what is relevant, and thus to identify the predicted relevant set and predicted nonrelevant set. Finally, the reasonableness of the sample-based effectiveness estimates will need to be assessed on a case-by-case basis.

* + - 1. 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 three 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;[[36]](#footnote-37) (2) if merged, how/when does the team introduce these new documents to the computer; and (3) if merged, how does the team ensure the software has reviewed these new documents properly (how do we ensure successful training on these new documents)?

Overall, many supervised learning methods look to try to use the human and software knowledge gained to date to start the training and review of the supplemental collection being added to the original TAR set. In other words, the team tries to avoid starting from scratch by leveraging all the 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.[[37]](#footnote-38) This leads to the identification of an updated predicted relevant set and predicted nonrelevant set.

* + - 1. 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 targeted key word searches to identify those discrete topics for further review and avoid updating the training and 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.

* + - 1. Unreasonable Training Results

A final challenge may 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 machine wrongly 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[[38]](#footnote-39) and validation-set documents are coded correctly by the human reviewer;
* Engage in additional control set document review to reduce the uncertainty of effectiveness estimates;[[39]](#footnote-40)
* Re-review training set documents to confirm that the human reviewer’s relevancy decisions were correct and fix any relevancy decisions;
* Engage in additional review of training set documents to improve the training results;[[40]](#footnote-41)
* 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;[[41]](#footnote-42) or
* Identify any large quantities of problematic documents in the TAR set that the computer is having difficulties making relevancy classifications on, and if so, remove those from the TAR set and review outside of the workflow.

**CHAPTER THREE**

**Alternative Tasks for Applying TAR**

|  |
| --- |
| 1. Introduction……………………..…………………………………………..…...... 28 2. Early Data Analysis/Investigation……..……………………….....................28 |
| 1. Prioritization for Review……………………………………………………..… .29 |
| 1. Categorization (by Issues, for Confidentiality or Privacy)…………..….29 |
| 1. Privilege Review………………………………………………………………….... 30 |
| 1. Quality Control and Quality Assurance…………………………………….. 31 |
| 1. Review of Incoming Productions………………………………………………..31 |
| 1. Deposition/Trial Preparation……………………………………………………32 2. Information Governance and Data Disposition…………………………….33 |
| 1. Records Management Baseline ……………………………………………. 33 2. Assessing Legacy Data – Data Disposition Reviews…………………… 34 3. Isolating Sensitive Content – PII/PHI/Medical/Privacy/Contractual/Confidential/ Privileged/Proprietary Data………………………………………………...34 |

1. 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.

1. Early Data Analysis/Investigation

Early Data Analysis (EDA) 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. Alternatively, finding documents that are conceptually dissimilar to sample documents can assist to identify documents that do not need further review.[[42]](#footnote-43)

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

C. Prioritization for Review

TAR is an effective tool for prioritizing and organizing batches of documents for attorney review to yield consistent and accurate work product. Counsel has two ways to leverage TAR in this context: targeted review or full review.

Targeted review uses TAR on a subset of documents that have been identified as having similarities. If counsel has performed any EDA on an ESI collection, knowledge gleaned from that analysis can be used to create a targeted sample to feed documents most relevant to the case to the TAR software and jumpstart TAR. Of course, unsupervised machine learning tools can also be used to evaluate and prioritize documents. Email threading identification, communication analysis, and topical clustering can group documents containing similar concepts into review batches to increase reviewer efficiency, assist with reviewer training at the onset of the case, and facilitate consistency in human coding decisions.

In a full review, a random sample of documents representing the entire data set can be fed to the algorithm for learning purposes. TAR then assigns relevance scores to documents across the collection based on their similarity to reviewed sample documents. Once the documents have been assigned a relevance score, review can be prioritized based on the documents most likely to contain relevant content.

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 irrelevant.

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 categorize privileged, confidential, or “hot” documents and to categorize documents 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 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 types of TAR face significant challenges and may be of less value to clients and counsel. This has more to do with the law, procedures, and risks surrounding privilege review rather than with TAR.

Although TAR can play a role in privilege review, it is essential to understand the 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. 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. 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, acceptable recall rates will likely fall short of that required 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 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 client’s concerns regarding the documents and a cost-benefit analysis. If the concerns are high, TAR can be used in conjunction with human/linear review prior to production of any documents. If client concerns are low or production timing is an issue, linear review could 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) and is commonly used: (1) during document review to assess the review team’s understanding of the review protocol and 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.

During a review, TAR can be used to assess the review team’s understanding of the coding protocol and 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 where 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, revisit training, or reassign members of the review team as needed.

TAR is also an effective QC tool to assess the overall quality of categorization of documents at the completion of 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 categorized during one review phase need to move to a second phase based on the decisions applied.

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 offers an efficient tool for a responding party to identify and produce relevant documents from large sources of ESI. But it is increasingly being used by requesting parties to efficiently review and analyze voluminous document productions. Search terms on and metadata searching of a large document production have their limits. 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 responsive and non-responsive 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 (e.g., evaluate the sufficiency of the incoming production and potential spoliation issues) and to aid in motions’ practice (e.g., 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 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 (e.g., linear progression or adoption of several approaches) 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.[[43]](#footnote-44)

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 or if all documents were not currently available for review, as is common in large litigation or investigative matters.

TAR offers some significant advantages over that method. Categorization and ranking, as opposed to a traditional TAR relevance review, allows counsel to identify key documents and issues that apply to a specific witness. Categorization delivers a high-level understanding of the types of documents in the dataset and of key dates that can be converted into an interview outline.

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 cannot 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 the 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 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. Notably, defensibility issues are often less pronounced, so the tools can be used more aggressively, with a possible exception when dealing with litigation hold or regulatory records-retention requirements associated with defensible deletion.

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 principles 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.

1. 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 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.

1. 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**

|  |
| --- |
| 1. Introduction……………………………………………........................................35 |
| 1. Should The Legal Team Use Tar?…………………………...……................... 35 |
| * + - 1. Are The Documents Appropriate For Tar?………………….…............... 35       2. Is The Cost And Use Reasonable?........................................................... 36       3. Is The Timing Of The Task/Matter Schedule Feasible?....................... 37       4. Is The Opposing Party Reasonable And Cooperative ?.........................37       5. Are There Jurisdictional Considerations That Influence The Decision?………………………………………………………………..…………37  1. The Cost Of Tar Vs. Traditional Linear Review………………..……….......38 2. The Cost Of Tar And Proportionality………………………….…………….. 39 |
| Appendix – Key Terms…………….………..………………..................................... 40 |
| Sponsors………………………….………….............................................................. 44 |

1. 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.

1. 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?
* Is 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 will be subject potentially to TAR is an important factor to consider when deciding whether to use TAR. TAR software require text to work, and thus are 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, as they will not be able to be classified well, or at all, by the software. With this in mind, TAR can generally be very effective at sorting through a custodian’s email, 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 high-content, user-generated documents. This includes emails, electronic documents such as Word files and searchable PDFs, high-content presentation slides, etc. Such documents have sufficient semantic content for the software to effectively analyze each document’s characteristics and find meaningful patterns it can apply to other documents in the population. In contrast, extremely short, low-content documents such as an email that says nothing more than “Please see attached,” lack sufficient content and cannot be effectively analyzed. These documents generally require more training examples or may not be correctly categorized, if they are categorized at all. Such materials are often excluded from a TAR project automatically and are almost always excluded as examples to be used for training purposes.

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

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

1. Is 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, however, before committing to it. For one, the volume of documents at issue should be considered. If the volume of documents is small, the cost of use[[44]](#footnote-45) may be higher than if a different review method was used to identify 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 In addition, the frequency of use must be analyzed.

1. Is the Timing of the Task / Matter Schedule Feasible?

As is 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, and, unlike traditional linear review, a TAR review may involve an initial lag of several days as the documents are indexed, the workflow is finalized, and the TAR process is set-up and the computer is trained. Tight document production deadlines or accelerated 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, training the computer become a challenge. Supplemental collections and their impact on TAR must be considered in these situations.

1. 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 governmental opponents, 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 process, including details regarding the particulars of the proposed process.

1. 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 widely vary. There is conflicting case law across some federal circuits relating to the extent of information about TAR to make relevancy determinations that the producing party must provide to the requesting party. Most judges have not ruled on the use of TAR, nor have they ruled on a similar issue. Some judges that have ruled do not require any level of disclosure, while some others have approved parties’ agreements that require a very high level of disclosure not only of the TAR process employed, but also production of non-responsive documents used during the TAR process. This requirement can be onerous and concerning, particularly when the non-responsive documents contain sensitive information or contain material that must be redacted, usually at an hourly rate. For this reason, consideration of the forum’s approach to disclosure must be included in calculating the overall risks and costs associated with TAR.

C. The Cost of TAR vs. Traditional Linear Review

TAR is 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 time line are all pertinent.

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

The costs of training the computer can often be significantly reduced by re-using previously reviewed and coded materials. Known-responsive documents and files can often serve as seed sets for training purposes, allowing one to both jump-start the TAR categorization while also reviewing fewer documents. However, even in these situations, some care needs to be given to determining whether any individual document is a good example for TAR training.

1. 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 may, under a Rule 26 proportionality analysis, question a party’s decision one day 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 help characterize the certainty of the point estimate. 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 conservative we want the estimation procedure to be.[[45]](#footnote-46) A confidence level of 95% has become widely used for estimation in e-discovery, and there is little reason to use other confidence levels.
* 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.
* Effectiveness Measures. Review workflows, including their individual components (software, keyword queries, attorneys,…) can be viewed as a classification system that determines whether documents are “relevant” or “not relevant.” If the results of the classification system are compared against a gold standard (e.g., a qualified attorney’s assessment), numerical effectiveness measures, such as recall and precision, can determine the system’s effectiveness. In addition to numerical measures, there are several other methods for evaluating a classification systems effectiveness.
* Elusion.  Another way of measuring effectiveness is to ask: how many relevant documents were missed?  In the example used below in the recall definition, the computer identified 800,000 documents as potentially nonrelevant.  Elusion estimates how many relevant documents were missed and are in the 800,000 nonrelevant set. Since there are a total of 100,000 responsive 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 classifier 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 or guarantee that the estimated values are close to the true value.
* 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.
* Precision. Precision is another effectiveness measure that answers the question: of all the documents the computer identified as potentially relevant, how many are truly relevant?  Using the example in the recall definition, the computer identified 200,000 documents as potentially relevant, of which 80,000 were identified as relevant by human-review, resulting in a precision 40% (80,000/200,000).
* Predicted Nonrelevant Set. The predicted nonrelevant set is one of two sub-sets of documents in the TAR set. It contains those documents in the TAR set that are not selected for the predicted relevant set, and thus would be excluded from further review or production efforts in this workflow.[[46]](#footnote-47)
* Predicted Relevant Set. The predicted relevant set is one of two subsets of documents in the TAR set. It is the set of documents that the human reviewers and the computer identify as potentially relevant and would achieve the goals for the matter as a result of the review. It is important to note that, like manual reviews, TAR classifications are not perfect.  The “predicted relevant set” will not contain all of 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%.   The contents of the predicted relevant set is dependent on several factors, such as the nature of documents in the TAR set, the definition of relevance, the training of human reviewers, the design and management of the workflow, the choice of software, and the settings used with that software.
* Recall. Recall is an effectiveness measure that answers the question: how much of the relevant material has the classification system found? Consider a workflow in which a TAR set of one million documents are collected, of which 100,000 are relevant.[[47]](#footnote-48)[2]  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 and computer’s relevancy decisions are as accurate as reasonably possible.
* Richness. Richness is defined to be the proportion of relevant documents in a data set. 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 teach the software what is relevant. The training set will contain relevant and nonrelevant documents. The software uses the training set to produce a classifier, and the classifier will be used to define the predicted relevant set. The number of relevant and nonrelevant documents necessary to produce a classifier with good effectiveness will depend on the nature of the documents in the TAR set, the difficulty of the responsiveness definition, and the particular TAR software and method used.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***Thank you to our Sponsors!**

Duke Law gratefully acknowledges the financial support of its EDRM sponsors.

|  |  |
| --- | --- |
| **Founding Sponsor**  Knovos  Exterro | |
| **Gold Sponsors**  AccessData  BDO  Dechert  TCDI  UnitedLex  Zapproved | **Silver Sponsors**  Everlaw  Fronteo  Garden City Group  KLDiscovery  Morae Global Corporation  NightOwl Discovery  Relativity |

1. In fact, the computer classification can be broader than “relevancy,” and can include discovery responsiveness, 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. [↑](#footnote-ref-2)
2. 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. [↑](#footnote-ref-3)
3. Unsupervised means that the computer does not use human coding or instructions to categorize the documents as relevant or nonrelevant. [↑](#footnote-ref-4)
4. Chapter Two describes each step in greater detail. [↑](#footnote-ref-5)
5. 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. [↑](#footnote-ref-6)
6. It is important to note that the terms TAR 1.0 and 2.0 can be seen as a 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 process, 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. There are a limited number of ways to select training documents, and a limited number of ways to determine training completion. [↑](#footnote-ref-7)
7. The Federal Rules of Civil Procedure do not specifically require parties to use statistical estimates to satisfy any discovery obligations. [↑](#footnote-ref-8)
8. 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. [↑](#footnote-ref-9)
9. 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. [↑](#footnote-ref-10)
10. Not all components may be needed to satisfy a Rule 26 obligation, which will depend on the specific facts and circumstances of each matter. [↑](#footnote-ref-11)
11. Appendix A 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. [↑](#footnote-ref-12)
12. For example, if a document is about a blueberry pancake eating competition, one feature may be blueberry pancakes. [↑](#footnote-ref-13)
13. The terms “classifies,” “ranks,” and “categorizes” are used in this document. Overall, in the context of this workflow, TAR software 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. [↑](#footnote-ref-14)
14. It should be noted that the workflow assumes that TAR is appropriate for the need. However, the team should always first determine if TAR is appropriate for a particular need. Chapter 4 is devoted to this issue. [↑](#footnote-ref-15)
15. If there is a very large volume of data of low richness, the responsiveness rates returned through search terms or other culling methods should be tested. Use of search terms may be used to jump-start TAR. 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 most circumstances. [↑](#footnote-ref-16)
16. Most software solely analyze a document’s text. However, some software may analyze other document metadata, such as email header fields and file names. [↑](#footnote-ref-17)
17. For example, the TAR set may be limited to emails, documents, presentations, or spreadsheets; and documents with a text size of less than 16 megabytes. Any document not falling within these limitations is excluded from the TAR set. [↑](#footnote-ref-18)
18. Some software can support multiple builds, allowing for multiple workflows to be run simultaneously. [↑](#footnote-ref-19)
19. 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. [↑](#footnote-ref-20)
20. 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. [↑](#footnote-ref-21)
21. Human selection and random selection have been traditionally known as “passive” training methods. This is because the computer is not involved in making subjective decisions on which documents should be used as training examples. [↑](#footnote-ref-22)
22. 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. [↑](#footnote-ref-23)
23. 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. [↑](#footnote-ref-24)
24. Note that if the lead attorneys (typically 1-3 people) are coding the training sets to teach the computer, there is less concern of inaccurate and inconsistent coding. The larger the training attorney team, the greater chance for inaccurate and inconsistent coding. [↑](#footnote-ref-25)
25. Note that “stable” is distinct from “acceptable.” Stability is an evaluation of how much the training process is altering the algorithm and related classifications. If additional training does not materially alter the results, yet those results are poor, TAR is arguably complete but a failure. Remediation may be possible through changes in population, criteria, coders, or other means. [↑](#footnote-ref-26)
26. 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. [↑](#footnote-ref-27)
27. 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. [↑](#footnote-ref-28)
28. Note that these illustrations are provided at a high level, and only reflect two possible TAR sets and processes. These should not be used to conclude that situations similar to these are or are not reasonable and proportionate TAR project. 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 1.0 (using a control set with uncertainty feedback, and TAR 2.0 with no control set but using relevancy feedback only). [↑](#footnote-ref-29)
29. 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 volumes of relevant documents. [↑](#footnote-ref-30)
30. Again, this is a broad generalization of certain traditional TAR 1.0 processes, and there may be variations in TAR 1.0 software. [↑](#footnote-ref-31)
31. 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 J (5)). [↑](#footnote-ref-32)
32. Again, the purpose of these guidelines is NOT to educate on the minutiae, or all variations, of how to do statistical calculations. [↑](#footnote-ref-33)
33. Consistency of review decisions in the predicted nonrelevant set and the predicted relevant set is important for this approach. [↑](#footnote-ref-34)
34. During the process of engaging in the TAR workflow, some documents will remain uncategorized, either because the human reviewer did not review the documents, the human reviewer reviewed the documents but could not make decisions on relevancy due to a lack of meaningful content, or the computer was not able to make decisions on relevancy 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. [↑](#footnote-ref-35)
35. 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). Without a 502(d) order, Rule 502(b) only prevents waiver if the producing party used reasonable procedures to identify and avoid producing privileged documents. [↑](#footnote-ref-36)
36. If the new documents are not merged with the original TAR set, then the new documents would go through a separate, parallel TAR workflow utilizing the same components herein. This would result in two different TAR exercises. [↑](#footnote-ref-37)
37. 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 would be stale (due to the supplement, the updated TAR set has new properties (may have new richness, recall, precision). In addition, the TAR training method in use may have more or less impact on the workflow. For example, TAR training methods that do not use an upfront control set may be more beneficial to use when dealing with supplemental collections, as there is no need to update a control set before continuing with training. [↑](#footnote-ref-38)
38. This would only apply to software that use a control set. [↑](#footnote-ref-39)
39. The greater the certainty/lower margin of error was used to create the control set, the less likely it is that additional review will change the metrics. [↑](#footnote-ref-40)
40. Again, selections of training examples could be made randomly or subjectively. This may involve using the relevant documents identified in the predicted nonrelevant set as training documents. [↑](#footnote-ref-41)
41. This may require creating a new control or validation set afterwards. [↑](#footnote-ref-42)
42. 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 litigation so that they can gain objective insight into what the ESI collection includes. These early data analysis (EDA) 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. [↑](#footnote-ref-43)
43. 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. [↑](#footnote-ref-44)
44. 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. [↑](#footnote-ref-45)
45. Unfortunately, but for good statistical reasons, the definition of confidence level is unintuitive. The confidence level is the probability that a particular method for producing confidence intervals (for instance, drawing a sample of a certain size, coding it, and applying an estimation formula) will generate a confidence interval containing the true value. However, once that procedure is used to produce a particular confidence interval from a particular sample, there is no longer any probability involved: the interval either contains the true value or it doesn’t. [↑](#footnote-ref-46)
46. 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. [↑](#footnote-ref-47)
47. [2] 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. [↑](#footnote-ref-48)