Volume 102 Number 2 | Summer 2018

Judicature
BOLCH JUDICIAL INSTITUTE

Published by the Bolch Judicial Institute at Duke Law. Reprinted with permission. 

JUDICATURE.DUKE.EDU

# Unlocking the e-discovery TAR blackbox

**EDRM at Duke Law has published a proposed set of e-discovery guidelines** that explain technology assisted review (TAR), also known as predictive coding and computer assisted review, and is now seeking public comments on the guidelines from judges and practitioners. An editable version of the guidelines is available for download on the EDRM website (*see* EDRM.net or http://bit.ly/EDRM-TARcomment).

More than 50 volunteer judges, practitioners, and e-discovery experts have been working on the project since December 2016. A companion set of "best practices" is being developed by 20 other judges and practitioners to provide protocols on whether and under what conditions TAR should be used. Together, the guidelines and best practices will provide a record and roadmap for the bench and bar, which legitimize and support the use of TAR in appropriate cases.

TAR is a machine-learning process and an early iteration of artificial intelligence (AI) for the legal profession. AI is quickly revolutionizing the practice of law and will continue to generate a steady stream of new tools designed to increase the efficiency and effectiveness of the practice of law. To date, the legal profession has been a reluctant suitor of technological assistance in e-discovery.

Machine-learning processes like TAR have been used to automate decision-making in industries since at least the 1960s, leading to efficiencies and cost savings in healthcare, finance, marketing, and other industries. But it is only now that segments of the legal community have begun to accept machine learning, via TAR, to automate the classification of large volumes of documents in discovery. These guidelines provide guidance on the key principles of the TAR process. Although the guidelines focus specifically on TAR, they are written with the intent that, as technology continues to change, the general principles will also apply to future iterations of AI beyond the TAR process.

TAR is similar conceptually to a fully human-based document review — but the computer replaces the human reviewer 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 a human review, provided that there is a reasonable and defensible workflow. Similar to a fully human-managed review where subject-matter attorneys train a human review team to make relevancy decisions, the TAR process involves human reviewers training a computer so that the computer's decisions are just as accurate and reliable as those of the trainers.

The potential for significant savings in time and cost — without sacrificing quality — is what makes TAR most useful. According to a 2012 Rand Corp. report, 73 percent of the cost associated with discovery is spent on review. Document-review teams can work more efficiently because TAR can identify relevant documents faster than human review and can reduce or eliminate time wasted reviewing nonrelevant documents. TAR promotes Rule 1 of the Federal Rules of Civil Procedure, which calls on courts and litigants "to secure the just, speedy, and inexpensive determination of every action and proceeding."

Traditional linear or manual review, in which teams of lawyers — billing clients — review boxes of paper or count-

less 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 significant discrepancy rates in the determinations of reviewers charged with identifying relevant documents by linear review — as much as 50 percent or more. TAR is similarly imperfect, but studies show that TAR is at least equally accurate, if not more accurate, than humans performing document-by-document review. Such review meets the overarching legal standard in discovery, which requires reasonableness, not perfection.

Importantly, no reported court decision has found the use of TAR invalid. Scores of decisions have permitted TAR, and a handful have even encouraged its use. The most prominent law firms in the world, on both the plaintiff and the defense sides 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.

In order for TAR to be more widely used in discovery, however, the bench and bar must become more familiar with it. These guidelines and the soon-to-be-issued best practices demystify the process and, more importantly, establish a logical framework for the bench and bar to accept future technological breakthroughs without interminable delay.

The leaders of the teams that drafted the guidelines are **Matt Poplawski** (Winston & Strawn); **Mike Quartararo** (eDPM Advisory Services); and **Adam Strayer** (Paul, Weiss, Rifkind, Wharton & Garrison) with **Tim Opsitnick** (TCDi). **James Francis**, retired United States magistrate judge, provided general editorial assistance. Following is the first chapter of the proposed 40-page TAR guidelines, which provides a good executive summary.

# Proposed Technology Assisted Review Guidelines

**EDRM at Duke Law – May 2018**

## CHAPTER ONE: *Defining Technology Assisted Review*

### A. 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] 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] 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.

## B. THE TAR PROCESS

The phrase "technology assisted review" can imply a broader meaning that theoretically could encompass a variety of nonpredictive coding techniques and methods, including clustering and other "unsupervised"[3] 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]

### *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.

### *2. 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] 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.

### *3. "TRAINING" THE COMPUTER USING SOFTWARE TO PREDICT RELEVANCY*

The next step is for human reviewers with knowledge of the issues, facts, and circumstances of the case to code or tag documents as relevant or nonrelevant. The first documents to be coded may be selected from the overall collection of documents through searches, 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.

### *4. 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.

*5. TRAINING COMPLETION AND VALIDATION*

No matter what software is used, the goal of TAR is to effectively categorize or rank documents both quickly and efficiently, i.e., to find 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] 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 nonrelevant 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]

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.

## ABOUT EDRM

EDRM is a Duke Law-based community of e-discovery and legal professionals who create practical resources to improve e-discovery and information governance. As technology radically transforms litigation and the legal profession, EDRM members collaboratively develop frameworks, standards, educational tools, and other resources to guide the adoption and use of e-discovery technologies. Learn more or get involved at **EDRM.net**.

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.

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.

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

4 Chapter Two describes each step in greater detail.

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.

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.

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