WHITE PAPER

Technology-Assisted Review

An In-depth look into TAR 1.0 and TAR 2.0

HAYSTACK™

Technology-Assisted Review (TAR) Overview

What is Technology-Assisted Review?

Technology-Assisted Review is a term Maura R. Grossman and Gordon V. Cormack first coined and introduced to the legal industry in 2011:

> *"A technology-assisted review process involves the interplay of humans and computers to identify the documents in a collection that are responsive to a production request.... A human examines and codes only those documents the computer identifies – a tiny fraction of the entire collection. Using the results of this human review, the computer codes the remaining documents in the collection for responsiveness...."*(1)

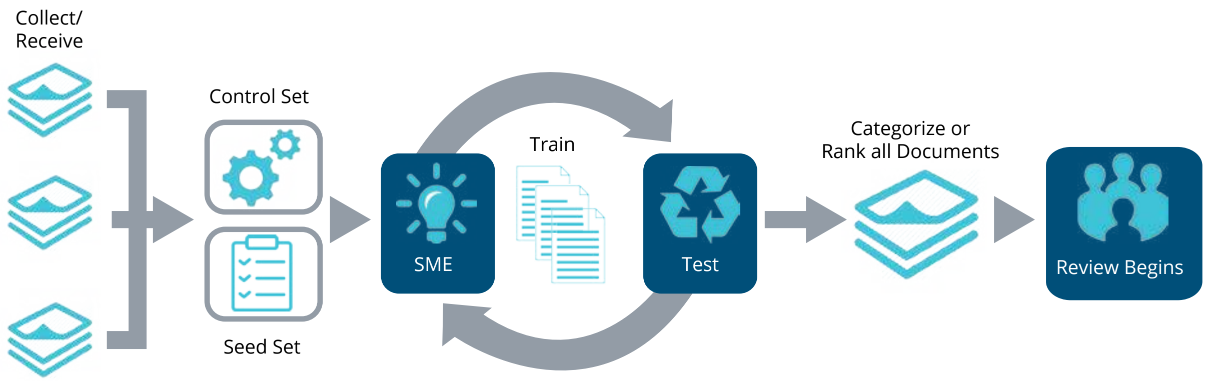
Although the technology and terminology are relatively new to the legal industry, the underlying technology itself, supervised machine learning, has been around for over fifty years—first in the field of information retrieval and later in areas such as digital marketing, online-sales, and the financial industry. Supervised machine learning is a different approach to creating computer software: the machine learns from examples, rather than being explicitly programmed for a particular outcome.

Particularly in the last five to ten years, technologists have been developing, refining, and marketing to the legal industry various TAR engines and workflows. The legal industry has taken notice of the savings in time, effort, and resources needed to separate relevant from non-relevant documents during discovery. In 2012, the Southern District of New York published the first opinion recognizing TAR as an "acceptable way to search for relevant ESI in appropriate cases."(2) Since then, other courts in the United States, Ireland, United Kingdom, and Australia have approved of and encouraged the use of TAR in appropriate cases, commenting on its reliability and availability to reduce cost and burden in the discovery process.(3)

Whether coined or marketed as predictive coding, computer-assisted review, or supervised machine learning, all are encompassed by the term TAR and possess the same underlying principle: a human reviews and codes electronically stored documents as either relevant or not relevant examples and submits the examples to the computer. The computer analyzes the features in the text that make a document relevant or not relevant. Usually, these features are words, though they can be word combinations or mathematical values related to groups of words. The computer learns which features are related to documents in each category (relevant and not relevant) and which distinguish between the categories. The computer starts to build a predictive model that, with sufficient and appropriate training, can be applied to and categorize or rank other unreviewed documents in the collection as relevant or not.

There are different ways to build such a predictive model. Over the years, experts and enthusiasts in the TAR field have coined various terms to describe the methodologies used to build a predictive model, the protocols surrounding its use, and to help legal practitioners differentiate between the various TAR tools available in the market. This is where buzzwords like TAR 1.0, TAR 2.0, Continuous Active Learning® (CAL®), Simple Active Learning (SAL), and Simple Passive Learning (SPL), to name a few, come into play. Keep in mind most of the TAR tools available today do not fall neatly into one of these categories but are a good starting point to better understand TAR. In the following sections of this paper, we will examine what these terms mean.

What is TAR 1.0?



While different products that employ a TAR 1.0 workflow vary in the details of how they work and the terminology used, all follow the same workflow whereby the training of the predictive model is one time and, once training is complete and the predictive model has been applied to the rest of the documents to categorize or rank them as relevant or not, the review process begins. You are aiming to build one ultimate classifier to differentiate between relevant and not relevant documents.

At the outset, a Subject Matter Expert(s) (SME) reviews and tags example documents in a randomly selected control set, which generally consists of at least 500 documents (and often more depending on the richness of the data set), as relevant or not. Most of the documents reviewed in the control set will be non-relevant as most data sets contain far less than 50% relevant documents. The documents in the control set are not used to train the classifier how to differentiate between relevant and not-relevant documents but are used as a "truth set" to measure how well the classifier is performing throughout the training rounds by comparing how the computer classified the document (relevant or not relevant) to how the human coded the document.

Once the control set is complete, the next step is to train the classifier to learn to differentiate between relevant and not relevant documents based on the coding decisions of the SME(s). The system learns which terms or other features tend to occur in relevant documents and which tend to occur in non-relevant ones. Training generally takes place in rounds. The number of documents coded per round varies by product and can include a pre-set number of documents configured by the TAR tool, a set number chosen by the end-user, or a variable number of documents selected using random or judgment sampling with the round size determined by the chosen confidence level and margin of error.

How the documents are selected for each training round can also vary from product to product. Documents used for training can be selected by the human reviewer, either in a targeted manner or by random selection. Or, the algorithm can choose future documents to train the system on, usually from among the documents about which it is least certain.

The training rounds generally will be a mix of relevant and not-relevant documents, but the make-up can vary based on how and who selects the next round of documents for training.

Most TAR systems also employ various quality control measures in between training rounds as part of its iterative, ongoing process to gauge how training is progressing. Such measures include comparing the human coding of the documents in the initial control set to the classifier's coding of the same documents to test the classifier's accuracy; measuring the number of documents where the human and classifier disagree as to whether a document was relevant or not; or the number of documents changing each round from one category to the other (relevant to not relevant and vice versa).

The SME(s) continue to train and develop the model to help it predict the relevance of other documents in the data set until the classifier is deemed sufficiently accurate. TAR systems vary as to how the end-user knows it's time to stop training. Some will advise the end-user that training has "stabilized" and further training of the system would not further help the classifier differentiate between relevant and not relevant documents. Other systems leave it to the discretion of the end-user to determine when the classifier is sufficiently accurate based on the outcomes of the between-rounds quality control measures.

The training process typically ends with validation to determine its effectiveness. There are various approaches to validating the accuracy of the classifier at differentiating between relevant and not relevant documents, but most commonly, validation consists of reviewing a random sample of documents that were categorized as not relevant or below a certain cut-off score to verify that you are not finding more relevant documents than expected or key documents amongst the documents categorized or scored as not relevant.

Ultimately, however, when to stop the TAR training process is based on legal judgment of reasonableness and proportionality considerations: How much could the result be improved by further review? Does the case justify further review? 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?

Once validation is complete, the classifier is applied across the entire data set and the documents are categorized as relevant or not relevant or assigned a ranking score from most to least likely to be relevant. From there, document review begins. Most commonly, the TAR results will be used to prioritize the documents for review from most to least likely relevant or to eliminate from review those documents categorized or ranked as not relevant.

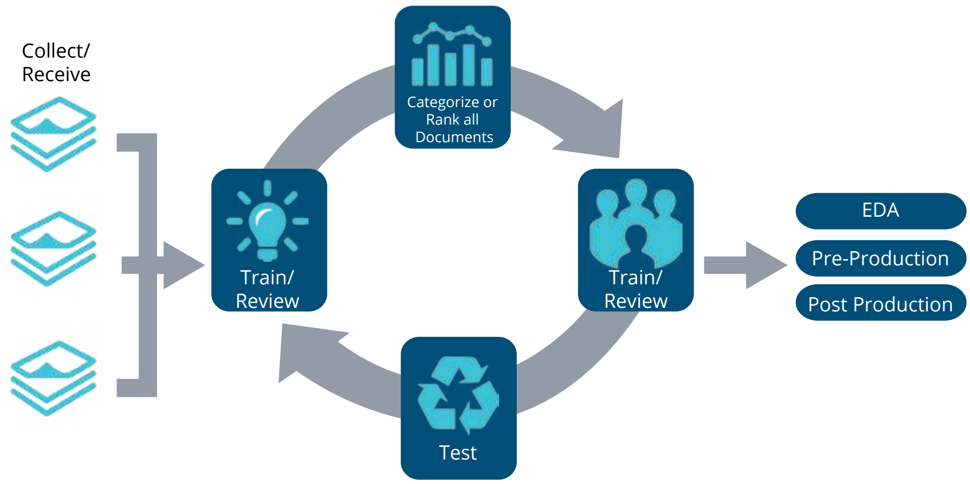
What is Simple Passive Learning and Simple Active Learning?

Broadly speaking and greatly oversimplifying the intricacies of the available TAR tools on the market, most TAR algorithms that utilize a TAR 1.0 workflow fall into one of two categories, or a variation thereof: Simple Passive Learning (SPL) or Simple Active Learning (SAL). The differentiators here are whether the computer or the human is choosing the training documents and whether the training is one time or continuous.

In an SPL model, the training is "one time" in that once the training is considered complete (i.e., the algorithm has stabilized), the training is complete and the review process begins. SPL is a passive model in that all the training documents are selected by the human reviewer; the algorithm plays no role in selecting the training examples.[4]

In SAL models, again the training is "one time" in that once the training is considered complete (i.e., the algorithm has stabilized), the training is complete and the review process begins. SAL is active, however, in that instead of the human selecting all the training documents, after the initial seed set the algorithm chooses additional documents for the human to review. Usually, the algorithm chooses the documents it is least certain how to classify and, therefore, from which it will learn the most.[5]

What is TAR 2.0?



In a TAR 2.0 workflow, the training of the classifier and the review process are one in the same. The training and classification of the documents are performed concurrently and on a continuous basis throughout the review process.

Prior to starting the training and review process, it is useful to code a random sample of documents to get an estimate on the number of relevant documents in the collection that will need to be reviewed to budget for the time and resources needed to complete the project on time.

To start the training and review, the user generally submits at least one relevant document to the system, located either through keyword searching, concept searching, clustering, or some other means. The system then submits for review the document it predicts to be the next most likely to be a relevant document. The user codes this document as relevant or not and submits the coded document to the system. Some tools will also periodically submit for review those documents it is unsure of how to categorize. Especially early in the process, the user will review some non-relevant documents. As the training and review progress and the classifier can better predict whether a document is relevant or not, the reviewer should review progressively fewer non-relevant documents.

The coded documents are continuously fed to the algorithm to rank and re-rank documents based on the previous coding decision. This review and training process continues until the number of relevant documents significantly decreases or runs out and further training and review to find the next relevant document would be disproportionate—again involving a legal judgment of reasonableness and proportionality considerations.

The training process typically ends with validation to determine its effectiveness. Most commonly, validation consists of reviewing a random sample of unreviewed documents that were categorized as not relevant to verify that there are not more relevant documents than expected or key documents amongst the documents categorized as not relevant.

The tools available on the market vary as to how continuous the training decisions are submitted to the classifier and how frequently the classifier is updated. All tools, however, share the common trait that they create a series of disposable classifiers—each with the sole purpose of identifying more relevant documents for review.

In some tools, training and review takes place one document at a time and the classifier is updated and documents re-ranked after each coded document is submitted. The next document presented to the reviewer is the next most likely to be relevant. Other tools serve up the next document to be reviewed by working from a review queue consisting of a local list of a couple hundred documents predicted most likely to be relevant. The training model, and thereby the review queue, is updated at set intervals of time. There are yet other tools that work from traditional review batches—both static and dynamic. After a set number or percentage of documents are reviewed, the coding decisions are submitted to the system, the classifier updates and re-ranks the documents, and the administrator or TAR tool creates new batches for review consisting of all or some of the documents predicted most likely to be relevant.

What is Continuous Active Learning®?

TAR 2.0 workflows are associated with Continuous Active Learning (CAL). The system is active in that instead of the human selecting the training documents, after the initial seed set the algorithm chooses the additional documents for review. The learning is continuous in that instead of creating one "best" classifier, it creates a series of disposable classifiers—each with the sole purpose of identifying more relevant documents for review—and continues to construct new classifiers trained on all documents reviewed to date, until substantially all relevant documents in the data set have been reviewed.(6)

Both protocols aim to find the greatest number of relevant documents while reviewing the least amount of non-relevant documents to do so.

How is TAR used in the review process?

The most discussed and analyzed way to use TAR is to eliminate from human review and production those documents categorized or ranked by the system as non-relevant. This use case is the focus of the case law surrounding the use of TAR and receives the most attention due to defensibility concerns. The savings in review costs can be substantial.

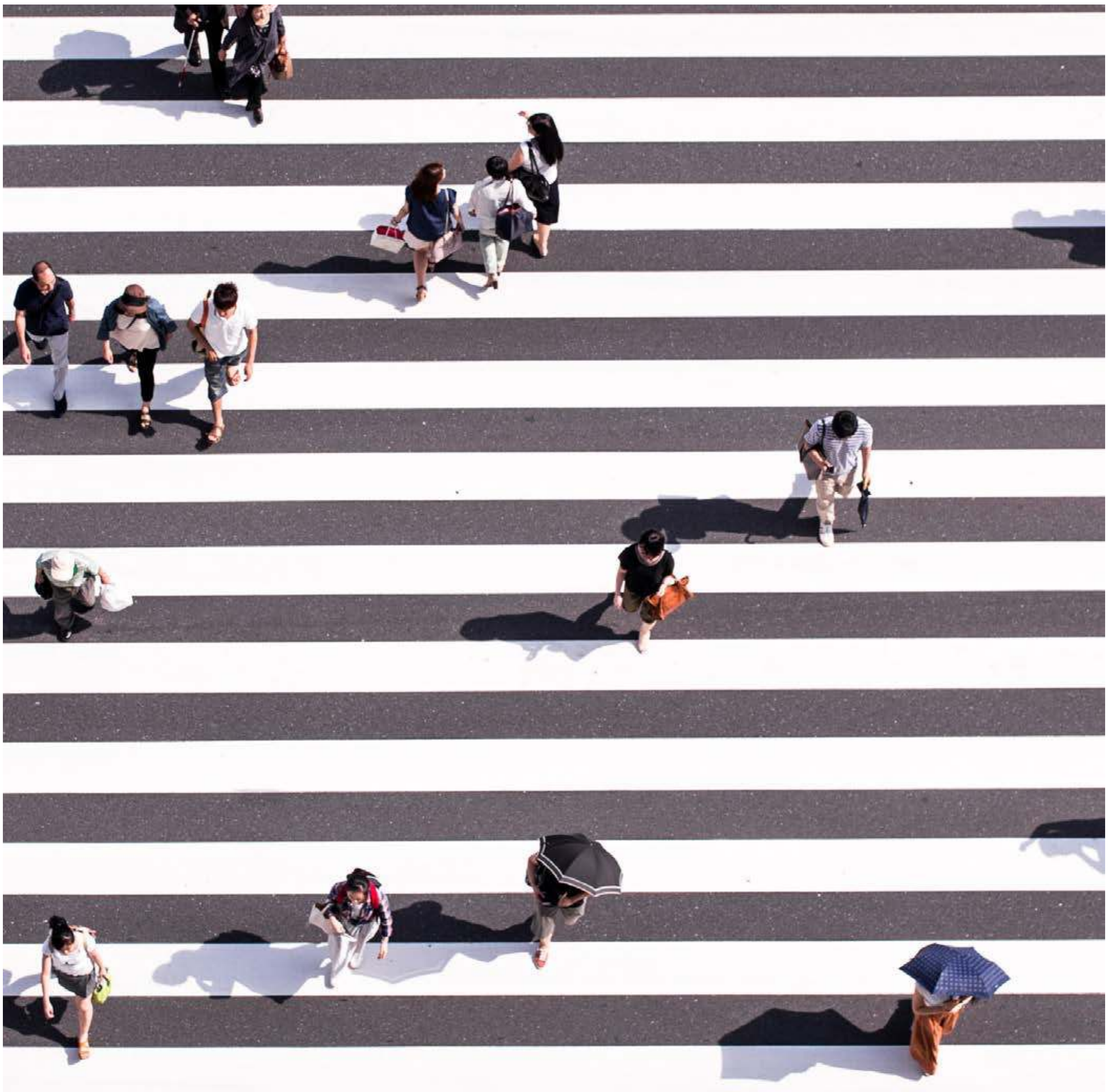
Another way to use TAR is to prioritize your review. The review team will still review the entire document population but will review and provide to the case team those documents most likely to be relevant at the start of the review. Key documents found early in the process can help the legal team identify the strengths and weaknesses of their case, better prepare for depositions, and shape case strategy early on. Because all documents will be reviewed, often less stringent validation measures, if any, are needed.

TAR can also be used to identify and review only the documents most likely to be relevant in an incoming production. If the producing party takes a broad view of relevance in responding to document requests or produces relevant documents with their families (many of which will be not relevant), most of the documents in an incoming production will be of little interest to the receiving party and can be culled from review without defensibility concerns.

In addition to, and usually in conjunction with one of the above use cases, TAR can be used for quality control during a review to assess the individual and overall review team's coding accuracy and understanding of the review protocol. The review manager can compare the coding decisions of the human document review with the categorization or ranking scores assigned by the TAR algorithm and use the results to elevate for second pass review those documents where discrepancies exist or to inform areas of the coding protocol that require further explanation or re-training on.

What kinds of cases should TAR be used for?

TAR algorithms analyze the extracted text of the documents. Accordingly, data sets where most of the documents have accurate extracted text and are rich in human authored text work best. Plan on employing an alternative workflow for human review of those documents that do not fall into this category, such as scanned documents with poor OCR, photos, videos, voice recordings, construction and architect plans, or spreadsheets consisting primarily of numbers.



How are TAR 1.0 and TAR 2.0 similar?

- Both can offer significant savings in the time and resources required to separate relevant and not relevant documents during document review.

- Both protocols aim to find the greatest number of relevant documents while reviewing the least amount of non-relevant documents to do so.

- Both training protocols are iterative and interactive between the human and the computer.

- The use cases are the same—culling non-relevant documents from review, prioritizing documents for review, and quality control.

How are TAR 1.0 and TAR 2.0 different?

In TAR 1.0, you train to build one ultimate classifier to differentiate between relevant and not relevant documents. In TAR 2.0, you train to build many, disposable classifier based on the latest coding decision(s) to differentiate between relevant and not relevant documents.

In TAR 2.0, there is no control set to code. As such, there is no arm twisting involved to get an SME(s) to carve out the time to review and tag these documents, most of which are not relevant, before starting the review process.

In TAR 1.0, the review team needs to wait for the SME(s) to finish training the classifier before review can start. In TAR 2.0, training and review are one and the same.

TAR 1.0 does not necessarily require a large set of documents to work, but its value tends to increase disproportionately as the size of the document collection grows because the effort typically required to train a system does not increase (or does not increase as quickly) as the size of the document collection grows. Small collections can require almost the same level of training effort as large collections do.(7) For example, if your collection is small, say 10,000 – 20,000 documents, and you must code 3,000 – 7,000 documents to complete the control set and training rounds before you can even begin review of a segment of the remaining documents categorized as relevant, your savings on time and cost as compared to traditional human linear review may not be that great, particularly after you factor in the upfront costs for the technology, the hourly fees for SME(s) training the system, any delays caused in having SME(s) train the system, and the expert guidance needed to administer the technology.

In TAR 2.0, because training and review are one and the same and there is no lag time in starting the review, your savings on time and cost, regardless of collection size, should always exceed your investment in the upfront costs for the technology and administration and provide savings as compared to traditional human linear review.

Some TAR 1.0 tools can accommodate evolving definitions of relevance and the addition of incremental loads of data; others cannot. Given the nature of TAR 2.0 to build and train several disposable classifiers and to continuously re-categorize and rank the documents, this limitation is not a factor.

In TAR 1.0, Subject Matter Expert (SME) handles all training and review team judgments are not used to further train the system, which could provide the benefit of precise and narrowed relevance. In TAR 2.0, review teams train the system as they review, working alongside SME for maximum effectiveness. This is beneficial in cases where the relevance evolves during the case, for example in investigative matters.

Is TAR 2.0 better than TAR 1.0?

Studies exist comparing TAR products and protocols and reasonable minds can and do differ on the "best" tools and protocols available.[8] While some tools may be slightly more or less accurate and require more or less training than others, one point most can agree on is that when used correctly and for the right cases, all TAR tools and protocols can offer significant cost and time savings and are at least as accurate, if not better, than traditional human linear review.

Conclusion

After many years and hundreds of TAR projects, we have seen countless matters positively affected using TAR. We have seen the benefits this technology can provide our clients and believe the savings are worth the investment. To that end, HaystackID has an Analytics division to educate, advise, implement, and support TAR projects on behalf of our clients.

Works Cited

(1) Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11, p. 3-4 (2011), http://jolt.richmond.edu/v17i3/article11.pdf; Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 1, 15 (2013), http://www.fclr.org/fclr/articles/html/2010/grossman.pdf.

(2) Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 183 (S.D.N.Y. 2012).

(3) *See generally Id.* at 193; *National Day Laborer Organizing Network v. U.S. Immigration & Customs Enforcement Agency*, 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012); *Rio Tinto PLC v. Vale S.A.*, F.R.D. (S.D.N.Y. 2015), *citing Dynamo Holdings Ltd. P'Ship v. Comm'r of Internal Revenue*, 2014 WL 4636526 at *3; *McConnell Dowell Constructors v. Santam Ltd Ors* 2016 VSC 734 (2016); *Irish Bank Resolution Corporation Ltd & Ors v Quinn & Ors* [2015] IEHC 175; *Pyrrho Investments Ltd. V. MWB* [2016] EWHC 256 (Ch).

(4) Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, in PROCEEDINGS OF THE 37TH INT'L ACM SIGIR CONFERENCE ON RESEARCH AND DEV. IN INFO. RETRIEVAL 153 (2014), http://dx.doi.org/10.1145/2600428.2609601.

(5) *Id.*

(6) *Id.*

(7) Herbert L. Roitblat, *Introduction to Predictive Coding* (2013), https://theolp.wildapricot.org/Resources/Documents/Introduction%20to%20Predictive%20Coding%20-%20Herb%20Roitblat.pdf.

(8) See Maura R. Grossman and Gordon V. Cormack, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery* (2014) Maura R. Grossman and Gordon V. Cormack, *Technology-Assisted Review in Electronic Discovery, in Data Analysis in Law* (2017); Maura R. Grossman and Gordon V. Cormack, *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review* (2015); Bill Dimm, *The Single Seed Hypothesis* (2015), https://blog.cluster-text.com/2015/04/25/the-single-seed-hypothesis/.

Learn More. Today.

For more information on our TAR or other analytics services, please reach out to HaystackID for more information.

About HaystackID

HaystackID is a specialized eDiscovery services firm that helps corporations and law firms find, understand, and learn from data when facing complex, data-intensive investigations and litigation. HaystackID mobilizes industry-leading computer forensics, eDiscovery, and attorney document review experts to serve more than 500 of the world's leading corporations and law firms from North America and Europe. Serving nearly half of the Fortune 100, HaystackID is an alternative legal services provider that combines expertise and technical excellence with a culture of white glove customer service. For more information about its suite of services, go to HaystackID.com.