Why Continuous Active Learning

Our Algorithm Continuously Learns and Adjusts, Improving Results and Reducing Costs

Insight Predict was the first predictive ranking engine to employ Continuous Active Learning. Predict continuously integrates the judgments made by the review team and then uses those judgments to constantly rerank the entire document collection.

Unlike earlier forms of predictive ranking, no control set is required, no expert training is needed and rolling uploads are no problem. With Continuous Active Learning, you find more relevant documents more quickly, with less effort and at lower cost.

CAL’s Proven Performance Superiority

The superiority of Continuous Active Learning (CAL) over other forms of technology assisted review was established in a 2014 peer-reviewed study conducted by two of the world’s leading experts on e-discovery, Maura R. Grossman and Gordon V. Cormack.

Comparing CAL against Simple Passive Learning (SPL) and Simple Active Learning (SAL)—two TAR protocols associated with early approaches to predictive coding, or TAR 1.0—they concluded that CAL demonstrated superior performance, while avoiding many problems associated with the TAR 1.0 protocols.

The Limitations of TAR 1.0 Protocols

TAR 1.0 protocols require a cumbersome process that limits their adaptability to real-world contexts.

- A subject matter expert (SME)—often a senior lawyer—reviews and tags a random sample (500+ documents) to use as a control set for training.
- The SME then begins a training process using SPL or SAL, reviewing documents and tagging them relevant or non-relevant.
Using these judgments, the TAR engine builds a classification algorithm that will find and rank other relevant documents.

The algorithm is tested against the control set and, depending on the results, the SME may be required to do more training to help improve the algorithm.

Training and testing continue until the classifier is “stable.” Then, the TAR engine runs its algorithm against the entire document population. The SME can then review a random sample of ranked documents to determine how well the algorithm did in pushing relevant documents to the top. The sample will help the review administrator decide how many documents will need to be reviewed to reach the appropriate recall rate.

Even though training is initially iterative, the process is finite. Once the classifier has learned all it can about the 500+ documents in the control set, that is the end. You simply turn it loose to rank the larger population (which can take hours to complete) and then divide the documents into categories to review or not review.

**Practical Problems with TAR 1.0**

These TAR 1.0 protocols present a number of practical problems when applied to “real world” discovery.

- **Only One Bite at the Apple**: Once the team starts to review documents, there is no way to feed back their judgments and improve the classification/ranking algorithm.
- **SMEs Required**: Requiring a senior lawyer to review thousands of documents to build a control set, to train and then test the results is expensive and a frequent cause of delay.
- **Rolling Uploads**: TAR 1.0 is unable to handle rolling uploads, which are common in e-discovery. New documents render the control set invalid, requiring new training rounds.
- **Low Richness**: TAR 1.0 is ineffective with low-richness collections, sometimes requiring review of thousands of documents just to train the system.

**TAR 2.0 and Continuous Active Learning**

These real-world problems disappear with Insight Predict, our TAR 2.0 engine that uses continuous ranking and Continuous Active Learning to reduce review time and costs while also making the process more fluid and flexible.

**Continuous Ranking**

Insight Predict can rank millions of documents in minutes. It ranks every document in the collection every time, continuously integrating new judgments by the review team into the algorithm as the work progresses. This solves the problems that characterized TAR 1.0 systems:

- **Eliminates need for control set.** Training success is based on ranking fluctuations across the entire set, rather than a limited set of randomly selected documents. When document rankings stop changing, the classification/ranking algorithm has settled, at least until new documents arrive.
- **Allows rolling uploads.** Because Predict does not use a control set for training, it can integrate rolling document uploads. When new documents are added, they simply become part of the continual ranking process.
- **Works well with low richness collections.** Start the training with any relevant documents you can find. As the review progresses, more relevant documents rise to the top of the rankings, letting your trial team can get up to speed more quickly.
As the new documents are reviewed, they integrate further into the ranking.

This example from Insight Predict illustrates the initial fluctuation when new documents were added to the collection midway through the review process. Initially the rankings fluctuated to accommodate the newcomers. Then, as representative samples were identified and reviewed, the population settled down to stability.

**Continuous Active Learning**

Continuous Active Learning has two aspects. First, it is “continuous”; training doesn’t stop until the review finishes. Second, it is “active”; the computer feeds documents to the review team with the goal of making the review as efficient as possible (minimizing the total cost of review).

As the reviewers progress through documents, their judgments are fed back to the system to be used as seeds in the next ranking process. When the reviewers ask for a new batch, the documents are presented based on the latest completed ranking. To the extent the ranking has improved from the additional review judgments, they receive better documents than they would have had the learning stopped after “one bite at the apple.”

In effect, the reviewers become the trainers and the trainers become reviewers. Training is review, we say. And review is training.
How the Continuous Learning Process Works

The process of Continuous Active Learning is simple and flexible. Here are the basic steps:

1. Start by finding as many relevant documents as possible. Feed them to the system for initial ranking. You can even start with no relevant documents and build off of the review team’s work.

2. Let the team begin review. They get an automated mix including highly relevant documents and others selected by the computer based on contextual diversity and randomness to avoid bias.

3. As the review progresses, QC a small percentage of the documents at the senior attorney’s leisure. Our QC algorithm will present documents that are most likely mistagged.

4. Continue until you reach the desired recall rate. Track your progress through our progress chart and an occasional systematic sample, which will generate a yield curve.

The process can proceed in almost any way you want. Start with tens of thousands of tagged documents if you have them, or with just a few or none at all. Just let the review team get going and let the system balance the mix of documents included in the dynamic, continuously iterative review queue.

As reviewers finish batches, the ranking engine keeps getting smarter. If you later find relevant documents through whatever means, simply add them. It just doesn't matter when your goal is to find relevant documents for review rather than train a classifier.

Key Differences Between TAR 1.0 and 2.0

<table>
<thead>
<tr>
<th>TAR 1.0</th>
<th>TAR 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. One Time Training before assigning documents for review. Does not allow training or learning past the initial training.</td>
<td>1. Continuous Active Learning allows the algorithm to keep improving over the course of review, improving savings and speed.</td>
</tr>
<tr>
<td>2. Trains Against Small Reference Set, limiting ability to handle rolling uploads; assumes all documents received before ranking. Stability based on training against reference set.</td>
<td>2. Ranks Every Document Every Time, which allows rolling uploads. Does not use a reference set but rather measures fluctuations across all documents to determine stability.</td>
</tr>
<tr>
<td>3. Subject Matter Expert handles all training. Review team judgments not used to further train the system.</td>
<td>3. Review Teams Train as they review, working alongside expert for maximum effectiveness. SME focuses on finding relevant documents and QCing review team judgments.</td>
</tr>
<tr>
<td>4. Uses Random Seeds to train the system rather than key documents found by the trial team.</td>
<td>4. Uses Judgment Seeds so that training begins with the most relevant documents, supplementing training with active learning to avoid bias.</td>
</tr>
<tr>
<td>5. Doesn’t Work Well with low richness/prevalence collections; impractical for smaller cases because of stilted workflow.</td>
<td>5. Works Great in low richness situations; ideal for any size case from small to mega because of flexible workflow.</td>
</tr>
</tbody>
</table>

What Happens to the SME?

Rather than require SMEs at the outset, as TAR 1.0 systems do, CAL allows the process to start immediately and frees the SMEs for other tasks. The SMEs can focus on finding (through search or otherwise) relevant documents to help move the training forward as quickly as possible. They can also be used to monitor the review team, using our QC algorithm designed to surface documents likely to have been improperly tagged.
What are the Savings?

In another Grossman and Cormack study, they quantified the differences between the TAR 1.0 and 2.0 protocols by measuring the number of documents a team would need to review to get to a specific recall rate. This chart from their study shows the difference in the number of documents a team would have to review to achieve a 75% level of recall comparing continuous active learning and simple passive learning:

<table>
<thead>
<tr>
<th>Matter</th>
<th>CAL Total Effort</th>
<th>SPL Total Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>6,000</td>
<td>44,000</td>
</tr>
<tr>
<td>202</td>
<td>11,000</td>
<td>7,000</td>
</tr>
<tr>
<td>203</td>
<td>6,000</td>
<td>9,000</td>
</tr>
<tr>
<td>207</td>
<td>11,000</td>
<td>9,000</td>
</tr>
<tr>
<td>A</td>
<td>11,000</td>
<td>27,000</td>
</tr>
<tr>
<td>B</td>
<td>8,000</td>
<td>7,000</td>
</tr>
<tr>
<td>C</td>
<td>4,000</td>
<td>3,000</td>
</tr>
<tr>
<td>D</td>
<td>18,000</td>
<td>6,000</td>
</tr>
</tbody>
</table>

The results showed that the review team would have to look at substantially more documents using the SPL (random seeds) protocol than CAL. For matter 201, the difference would be 50,000 documents. At $2 a document for review and QC, that would be a savings of $100,000. For matter 203, which is the extreme case here, the difference would be 93,000 documents. The savings from using CAL based on $2 a document would be $186,000.

Here is another chart from the Grossman and Cormack study that compares all three protocols over the same test set. In this case, they varied the size of the training sets for SAL and SPL to see what impact it might have on the review numbers. The results for both of the TAR 1.0 protocols improve with additional training but at the cost of requiring the SME to look at as many as 8,000 documents before beginning training. Even using what Grossman and Cormack called an “ideal” training set for SAL and SPL (which cannot be identified in advance), CAL beat or matched the results in every case, often by a substantial margin.
Our research similarly finds that there are substantial savings to be had by continuing the training through the entire review. You can see it in this example:

![Graph showing savings through continuous training]

**What about Review Bias?**

Some fear that Continuous Active Learning is tainted by “review bias.” If the ranking is based on documents you found through keyword search, what about the relevant documents you didn’t find? They argue that random selection of training seeds reduces the likelihood of review bias.

Insight Predict combats review bias by using a proprietary method known as contextual diversity sampling. Contextual diversity sampling uses an algorithm we developed to present the reviewers with documents that are very different from what the review team has already seen.

Because we rank all the documents, we know something about the nature of the documents already seen by the reviewers and the documents not yet reviewed. The contextual diversity algorithm clusters unseen documents and then presents a representative sample of each group as the review progresses. And, like our relevance and QC algorithms, contextual diversity keeps learning and improving as the review progresses.

This image shows side-by-side comparisons of the coverages achieved using random
sampling and contextual diversity sampling. You can see that contextual diversity sampling achieved much broader coverage.

Ultimately, the review teams get a mix of documents selected through relevance feedback and those selected for their contextual diversity. Doing so helps better train our algorithm and combats the possibility of unwanted bias.

**The Best is Now Even Better**

When Insight Predict debuted as the first commercial review product to use an advanced CAL protocol, it won “New Product of the Year” in the 2015 Legaltech News Innovation Awards. Since then, we’ve made the best even better. Enhancements to Predict’s CAL algorithm expand its machine-learning capabilities to include multi-word bigrams and trigrams, further improving Predict’s ability to differentiate relevant documents and prioritize them for review. These enhancements can increase review efficiency 20 to 40 percent. For a set of 250,000 documents, that could mean savings of $50,000 to $100,000.

**Superior Performance, Superior Savings**

Insight Predict is the only TAR engine to offer continuous ranking, Continuous Active Learning and contextual diversity sampling. These protocols have been proven to reduce the effort, time and cost of review. Keep learning, get smarter and save more. That is a winning combination.