Realizing they had neither the time nor money to review all 2.1 million documents, client and counsel turned to Insight Predict, Catalyst’s unique technology assisted review engine.

Client Snapshot: Financial Corp.
- Loan fraud investigation
- 2.1 million documents to search
- Innovative tool jump-starts review
- Catalyst cuts review by 94%
Our client was a large banking institution embroiled in nasty litigation with a now-defunct borrower. The bank alleged it lost millions due to the borrower’s principals’ accounting fraud. Legal shots were fired, excuses ran rampant and the parties went hard at each other to see where the blame would end up. Bring on the discovery.

**The Problem**

Responding to a production request, our client conducted an extensive investigation to find responsive documents. Even after using a variety of techniques to cull those that it found, it was still left with over 2.1 million that needed consideration. Further keyword searching might have resulted in more reductions but the team wasn’t comfortable with what that process might miss.

Realizing they had neither the time nor money to review all 2.1 million documents, client and counsel turned to Insight Predict, Catalyst’s unique technology assisted review (TAR) engine. The plan was to employ Predict’s Continuous Active Learning (CAL) protocol and see if TAR might be effective in further reducing the population in a defensible manner.

**Step One: Building a Seed Set**

In this case, counsel had already reviewed and identified approximately 50,000 relevant documents for a previous production based on a similar request. Because our predictive ranking engine has no effective limit on the amount of training seeds it can handle, we used these documents as initial seeds to start the ranking process. Almost immediately, relevant documents from the larger collection were pushed to the front of the line for review.

**Step Two: Immediate Review**

Reviews using first-generation TAR 1.0 systems cannot begin until senior lawyers (aka subject matter experts) train the system. With Insight Predict’s advanced TAR 2.0 technology, this is not necessary. Rather, the review team could immediately begin requesting batches of documents to review.

Predict’s algorithm provided batches made up primarily of the documents it most highly ranked. This ensured that the review team was productive immediately, because they were focused on the...
documents that were most likely relevant. In turn, it enabled the trial team to quickly get their hands on the most important documents to help sharpen their analysis of the case.

The review batches also included a mix of documents chosen based on their “contextual diversity.” This unique feature of Insight is designed to solve the problem of “you don’t know what you don’t know.” Specifically, our contextual diversity algorithm chooses documents for review that are different than those already reviewed.

In effect, the algorithm clusters unseen documents by their common themes. It then pulls the most relevant examples from each cluster and presents them to reviewers as part of the batch. If the reviewer tags an example as relevant, the ranking engine is cued to promote similar documents. If the example is not relevant, the ranking engine learns that this cluster is of lesser interest.

The ultimate goal of the CAL protocol is to feed reviewer judgments back to the system to improve the training and thereby the responsiveness rate for subsequent review assignments. As the reviewers release their batches, Insight Predict adds their judgments to further its training. The net result is that the algorithm gets smarter and smarter about finding and promoting relevant documents. This is in sharp contrast to the “one-time” training used by the earlier TAR 1.0 systems.

**Step Three: Using Keyword Search to Further Improve the Training**

Predict’s ability to use flexible inputs allowed the review team to take a multimodal approach to finding responsive documents. As the review progressed, Predict identified promising search terms as well as custodians who held the most likely relevant documents. This enabled the trial team independently to run searches with these key terms and then tag the relevant documents found through their searches.

As with the regular review, these tagged documents could then be fed into Predict’s ranking engine to further improve the training. This way, every attorney judgment on a document was used by Predict, no matter where that judgment was made.
Step Four: Completing the Review

Early on in the review, we took a richness sample to get a feel for the number of relevant documents to expect in the collection. That sample suggested we would find one relevant document out of every 100, which translates to a richness estimate of 1%.

As the review progressed, we tracked the number of relevant documents found by the team. Toward the beginning, the team tagged 10% relevant, representing a ten-fold increase in review productivity. Over time, that figure rose to 25% and sometimes as high as 35%. Through their tagging, the review team showed that the predictive ranking process was paying dividends.

Eventually, the relevant documents petered out, dropping down to and below base richness. At this point the team decided to stop the review and measure their progress.

Step Five: Measuring the Results

To determine how many relevant documents the team had found, we ran a systematic random sample against the entire document population. In this case the team chose to sample just under 6,000 documents, which is larger than the typical discovery sample. The goal was to present results with a high level of confidence and a narrow margin of error. Senior attorney reviewers manually reviewed these documents for added credibility.

“To determine how many relevant documents the team had found, we ran a systematic sample against the entire document population. Senior attorneys manually reviewed these documents for added credibility.”
Conclusion: 98% Recall After Reviewing Just 6.4% of the Total Population

The sample suggested that the team had found and reviewed 98% of the documents relevant to the production. This conclusion was based on a sample confidence level of 95% and a 2% margin of error. Even taking the lower end of the margin-of-error range, we estimated that the team had found at least 92% of the relevant documents, still well beyond levels previously approved by the courts.

All of this was accomplished through a CAL workflow that put attorney reviewers’ eyes on every document produced, yet still required a total review effort of only 6.4% of the total reviewable population of 2.1 million documents. That was 1.97 million documents the attorneys were saved from having to review.