Predictive Coding Defensibility: The Symantec Transparent Predictive Coding Workflow

Who should read this paper

Predictive coding is one of the most promising technologies to reduce the high cost of review by improving the efficiency of the review process. Limitations of available technology to provide a repeatable and well-defined predictive coding workflow have led to questions of its defensibility in court, and slow acceptance within legal departments. Symantec Transparent Predictive Coding provides a new level of visibility into the predictive coding process, allowing corporations, law firms, and government agencies to achieve the benefits of this promising technology while ensuring the defensibility of their document review process.
# Predictive Coding Defensibility: The Symantec Transparent Predictive Coding Workflow

## Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Symantec Transparent Predictive Coding</td>
<td>1</td>
</tr>
<tr>
<td>Early Case Assessment</td>
<td>2</td>
</tr>
<tr>
<td>Train</td>
<td>3</td>
</tr>
<tr>
<td>Predict</td>
<td>5</td>
</tr>
<tr>
<td>Test</td>
<td>6</td>
</tr>
<tr>
<td>Iterate</td>
<td>9</td>
</tr>
<tr>
<td>Conclusion</td>
<td>9</td>
</tr>
</tbody>
</table>
Introduction

Predictive coding is a type of machine-learning technology that enables a computer to help "predict" how documents should be classified based on limited human input. The technology is exciting for organizations attempting to manage skyrocketing legal budgets because the ability to automatically predict document responsiveness has the potential to save organizations millions in document review costs. Instead of paying lawyers and legal teams to review and code large numbers of potentially responsive documents, predictive coding technology allows a fraction of the documents to be reviewed by humans and results in a fraction of the review costs.

While incredibly promising, the industry has been slow to adopt predictive coding technologies largely because of the lack of clarity on whether predictive coding is defensible. The question of defensibility has focused in a few areas including differences in the underlying technology amongst vendors, the accuracy of predictive coding, the testing process to understand that accuracy, and the ability to present a repeatable workflow to the courts.

The debate has been driven mainly by corporations, law firms, and judges who have begun to recognize that predictive coding technologies offer a promising approach to achieve both cost savings and higher accuracy than traditional review. Although the interpretation of case law decisions related to defensibility remain in flux, judicial opinions and the industry discourse do make it clear that continued adoption of predictive coding technology requires a workflow that aligns with the way that attorneys conduct eDiscovery.

From a product perspective, this means new solutions that are easy to integrate into existing review tools, can provide better visibility into the technology processes, and have clearly defined repeatable steps will help drive a more efficient review process and substantiate defensibility in court.

Symantec Transparent Predictive Coding

The Clearwell eDiscovery Platform from Symantec™ is the leading electronic discovery software solution that enables enterprises, governments, and law firms to manage legal, regulatory, and investigative matters using a single application. Clearwell eDiscovery software makes it easy for organizations to cost effectively and defensibly solve real-world challenges across the entire eDiscovery lifecycle from legal hold and collections through analysis, review, and production.

Within the review and production module of Clearwell, Symantec has introduced a fully integrated predictive coding feature called Transparent Predictive Coding. As a form of technology assisted review, Transparent Predictive Coding works by leveraging algorithms that are created as the technology learns from tagging criteria of a manually-reviewed training set, enabling the system to automatically generate accurate coding predictions for the remaining documents. Unique in the marketplace, the Symantec solution provides visibility into the prediction process to enable more informed decisions and facilitate greater review accuracy.

In response to the challenges identified by enterprises, law firms, and legal service providers, Symantec designed a simple and flexible predictive coding tool that is embedded within the Clearwell eDiscovery Platform. Transparent Predictive Coding is seamlessly integrated, and works in conjunction with, the many other Clearwell technology tools. As a result, litigators and their staff have the power and flexibility to use any or all of the tools in the Clearwell eDiscovery Platform, including Transparent Predictive Coding, at any stage of their case. The technology can provide value in multiple ways by quickly prioritizing documents based on relevance for early case assessment and prioritized

---

linear review, or by automatically identifying clearly irrelevant documents — thereby eliminating the need to review those documents at all.

Litigator’s demand the flexibility to utilize predictive coding to address specific case needs, while the judicial system requires that users understand and present a repeatable process for defensibility. Symantec has integrated the flexibility to address multiple predictive coding use cases, but within a clearly defined workflow. There are five critical steps in the Transparent Predictive Coding workflow: Early Case Assessment, Train, Predict, Test, and Iterate.

![Figure 1. Symantec Transparent Predictive Coding provides a simple, repeatable, and flexible workflow.](image)

**A note about this paper:** For the simplicity of this white paper, it discusses a review workflow for finding responsive documents. Much in the same way traditional reviews are conducted for first pass review, for responsiveness, for hot documents, or for privilege, predictive coding can be used to address many real world eDiscovery issue codes. It is written assuming there is some familiarity with the eDiscovery process and predictive coding, and subsequently some terms and processes may not be explained fully. Where available, there are cited sources for a more in-depth explanation.

**Early Case Assessment**

Predictive coding is just one tool available to help litigators conduct eDiscovery in a defensible and cost-effective manner. Many entrants into the predictive coding marketplace have not realized that attorneys can derive more value from using multiple eDiscovery tools and techniques together than using predictive coding alone. Some providers recommend entirely separate predictive coding modules or workflows from typical linear or Technology Assisted Review (TAR) tools and processes. This is a critical failure because it abandons the tools and processes with which customers are familiar and already using to successfully reduce their eDiscovery expenditure. Tools like Early Case Assessment filters and searches not only reduce overall eDiscovery costs but make the latter steps of the predictive coding workflow more effective. Using predictive coding alone negatively impacts a customer’s bottom line while ignoring their demand for the flexibility to be able to use multiple technology tools together to address the specific needs of each user and case.

Using predictive coding as a standalone product or process also introduces risk into the review. For instance, the export from processing tools to predictive coding solutions and the import from predictive coding solutions into linear review tools makes chain of custody difficult to track, maintain, and present. Managing a coherent workflow across the eDiscovery process is required for defensibility in court.
Symantec Transparent Predictive Coding is seamlessly integrated within the larger Clearwell eDiscovery Platform and used in conjunction with other capabilities including advanced search, filtering, Early Case Assessment, and review features. Early Case Assessment in particular is critical to the eDiscovery process because it eliminates clearly irrelevant documents in the data set driving immediate cost savings. Additionally it is critical to making the predictive coding process more efficient.

Actions in the Early Case Assessment step:
- Using the advanced Early Case Assessment capabilities from Clearwell, customers drastically reduce the data set by 90 percent, eliminating those documents that are clearly irrelevant.

Symantec Transparent Predictive Coding is seamlessly integrated within the larger Clearwell eDiscovery Platform and used in conjunction with other capabilities including advanced search, filtering, Early Case Assessment, and review features. Early Case Assessment in particular is critical to the eDiscovery process because it eliminates clearly irrelevant documents in the data set driving immediate cost savings. Additionally it is critical to making the predictive coding process more efficient.

Using Early Case Assessment in advance of predictive coding helps increase the "yield," or percentage distribution, of responsive documents vis a vis the total population of documents. Excluding the profoundly irrelevant documents that users easily identify in Early Case Assessment increases the percentage of responsive documents vs. non-responsive documents (the yield) in the data set. As explained later in this paper in the Test step, the "yield" of the total document population has a direct effect on the sample size of the control-set that is used for testing. The higher "yield" created by the Early Case Assessment step in Clearwell's workflow translates into smaller sample sizes required for defensible testing and thus less documents that require manual review for predictive coding.

Train

Training the predictive coding system entails relying on human reviewers to teach the predictive coding tool to identify the difference between responsive and nonresponsive documents. The system studies the document coding decisions made by human reviewers on a small set of documents (often referred to as the training or seed set) in conjunction with the characteristics of the selected documents. This training process enables the computer to learn how human reviewers distinguish responsive from nonresponsive documents.

---

4 Clearwell customers often experience cull rates above 90 percent as documented in multiple case studies http://www.symantec.com/ediscovery-platform

5 A full explanation of eDiscovery sampling requirements can be found in the Symantec whitepaper; Predictive Coding Defensibility: Measuring accuracy with Random Sampling.
Many studies have demonstrated the effectiveness of the machine learning technologies used for predictive coding. So as knowledge of predictive coding has evolved, the minor differences in the underlying technology have become less relevant. But apart from underlying algorithms, the ability to train the predictive coding system effectively depends on whether legal users have the tools at their disposal to identify quality documents for the “training set”. If done incorrectly, the process to achieve an acceptable level of accuracy can incur more time and review cost—for with many solutions the general rule of thumb “garbage in, garbage out” applies.

Clearwell users benefit from a workflow that enables them to utilize a bevy of search and TAR tools like keywords, concepts, discussion threads, and custodian filters—to quickly identify the best training documents that are likely to be representative of the entire document population. The ease with which users identify documents helps train the system faster, and the ability to identify the correct “training set” means the initial predictions will be more accurate.

**Actions in the Train step:**

- Using their knowledge of the case, attorneys rely on their judgment to select a small number of documents that are representative of responsive documents in the case using searching, discussion threads, concept searching, and various other tools in the Clearwell eDiscovery Platform from Symantec™.
- These expert reviewers review and tag this “training set.”
- The “training set” information is fed into the system and the system uses this information to construct a profile of responsive documents.
- Based on the profile the system generates a mathematical prediction model that can be applied to other documents.

*Figure 3. Users easily progress the workflow forward as Clearwell’s action menu integrates both linear review and predictive coding functionality.*

The unique ability of the Clearwell eDiscovery Platform to use linear review and other TAR tools interchangeably with predictive coding is instrumental in helping organizations adopt predictive coding for their eDiscovery needs. Those knowledgeable about the case can easily perform each step of the workflow. This ensures that important case details or changes incorporated into the Early Case Assessment and manual review steps are represented in the predictive coding phase of the workflow.

The Transparent Predictive Coding functionality is seamlessly embedded within the larger eDiscovery workflow, offering a simple drop down menu to tag documents, identify the “training set” and advance the workflow. As a result elements from the onset of the case such as document sets, tags, training intelligence, and folders never require import/export to and from a separate predictive coding tool, module, or

---

Predict

Machine-learning technology used in a legal context for predictive coding has the ability to convert attorneys’ tagging intelligence into a prediction model, and use that intelligence to predict how a human attorney would tag documents for case responsiveness. However, for many tools on the market concerns persist regarding this “predicting” capability because it is difficult to understand why a particular document is considered to be by the system as responsive or non-responsive. This is required to demonstrate that the review process is accurate and defensible. It is also critical to efficiently move the case forward and improve the accuracy of the process. A lack of visibility into how predictions are made makes it difficult to decide case direction, determine whether training documents are effective or what type of additional review will be required.

In what Symantec refers to as the Predict step, the Clearwell system leverages the coding intelligence applied to the “training set” by the human reviewers to train the system and generate a computer-based mathematical prediction model. That model is then applied to the remaining documents in the population (or any designated subset of documents, for instance, a particular custodian) so that prediction scores can be assigned to each document based on their degree of responsiveness. In other words, the percentage likelihood that a particular document is responsive will now accompany each document in the total population.

Figure 4. Each document’s prediction rank is associated with it and available to users to filter, batch, or access additional Prediction Insight.
At this point in the Clearwell workflow, users have options for how they continue with their review. Depending on their objectives Transparent Predictive Coding can be used in various ways. The system automatically creates prediction filters which batch the documents by their prediction probability scores. Leveraging these filters attorneys can use Transparent Predictive Coding similar to a very advanced and automated culling tool— with the lowest percentage documents immediately discarded. Or Transparent Predictive Coding can be used to prioritize a manual review from this point onward, with lowest percentage documents discarded, those documents where decisions between responsive and non-responsive are unclear submitted to manual review and the top tier batched to be reviewed as critically responsive to the case.

To address concerns around the lack of visibility into this prediction process, Clearwell provides unparalleled analytics into understanding the underlying prediction rationale for a particular document. Prediction Insight transparently displays content and metadata relevant to each prediction. Users can quickly assess how the prediction was generated and use that insight to improve prediction accuracy and make consistent review decisions.

Figure 5. Clearwell provides analytics on the content and metadata related to the particular prediction rank, helping users make more informed decisions throughout the workflow.

Test

In many cases, the goals of the particular review may not require that the user undertake a full, predictive coding quality control process. As discussed in the context of the Clearwell workflow, many organizations can benefit from conducting the Early Case Assessment through Predict phase once, using the results to initiate a more informed and efficient linear review effort from that point on.

But when organizations conduct a review intending to leverage coding predictions to produce documents or exclude for privilege, they must test the accuracy of the system to be confident that they have not missed critical items or overproduced items of privilege or non-relevance. To test the system many solutions use sampling — taking a statistically valid amount of the total population of data, reviewing it, and testing predictions against reviewer’s codes. The document set selected via sampling to test the accuracy of the system’s predictions is called the “Control Set.” Comparing predictions against reviewer’s codes in the Control Set results enables attorneys to examine recall, precision, and the balance of the two, f-measure. F-measure is what we will discuss as the measurement of the “accuracy” of the review.

Recall refers to the proportion or percentage of truly responsive documents identified within a defined document population that are identified as responsive.

Precision refers to the proportion or percentage of documents identified within a defined document population that are truly responsive.

7- Recall refers to the proportion or percentage of truly responsive documents identified within a defined document population that are identified as responsive.

8- Precision refers to the proportion or percentage of documents identified within a defined document population that are truly responsive.
Sampling has been an area of major confusion in the predictive coding marketplace and an area where courts have increased scrutiny. A critical issue is that most predictive coding solutions use basic sample size calculators. These calculators assume that the characteristic being measured is present in the majority of documents across the population. For example, the percentage of cars in a car manufacturing plant that pass quality control is a large majority of the total population. In eDiscovery, however, the characteristic being measured is responsive documents, and in most matters, responsive documents are not the majority of documents on the case. Often they comprise only a small percentage of documents, and would be characterized as having a "low yield." In this situation a basic calculator will not work.

Early predictive coding tools have either failed to recognize the need to adjust for the yield or have lacked an integrated sampling workflow to enable legal users to confidently test and measure the accuracy of predictive coding. Not appropriately adjusting for yield can lead to significant variances in the margin of error of the accuracy calculation—which has not been accepted for litigation and presentation in court.

The Clearwell eDiscovery Platform from Symantec™ removes the guesswork from the sampling process to ensure that users are confident in the accuracy indicated by the system. Being directly integrated into the rest of the Clearwell Platform, users benefit from leveraging advanced Early Case Assessment techniques to improve the likelihood that the "yield" of the total population will be higher and therefore minimize the need for increasing the size of the control set. When a sample size adjustment is needed, the calculations are directly available in the tool and the standard workflow integrates the adjustment required.

### Actions in the Test step:

- The system generates a statistically valid random sample (Control Set).
- Users review, and tag the Control Set documents from within the normal review interface.
- The system automatically calculates the yield of the population given responsive vs. non-responsive information is available with the tags applied to the Control Set documents.
- Based on the yield, the system suggests additional samples (if required), which are then reviewed and added to the Control Set.
- The system compares predictions on the total population of the case to the tags applied to the Control Set.
- After comparison, the system provides an f-measure, or accuracy of the predictions after one cycle of training.

---

9 Andrew Peck, Search, Forward, October 2011 issue of Law Technology News
10 For an extensive explanation of statistical sampling for eDiscovery please see Symantec’s technical whitepaper - Predictive Coding Defensibility: Measuring accuracy with Random Sampling
Figure 6: Symantec makes complex statistics easy, automatically conducting sampling calculations and adjustments.

Users are presented with a sampling workflow, in which they input the desired margin of error and confidence level to generate a statistically valid random sample as the Control Set. Iterating directly into a manual review process, users review and tag the Control Set documents. The system immediately calculates the "yield" of the random sample given the tagging decisions. If necessary, the system automatically adjusts to the "yield" by suggesting the size of an appropriate additional sample to be reviewed.

Since attorneys have directed the system to apply the prediction model to the document population, the Clearwell eDiscovery Platform compares those coding predictions to the Control Set and automatically calculates accuracy metrics including recall, precision, and f-measure. Once accuracy has been established using the Review Quality Control capabilities in Clearwell, review teams can export reports detailing all steps used in testing accuracy and associated metrics. These reports are useful for demonstrating the results of predictive coding to the court or opposing counsel. They can also be used to evaluate the cost that may be incurred to achieve higher accuracy levels with additional training.
Iterate

First generation predictive coding tools often lack the ability to improve accuracy with confidence. Early movers in the predictive coding market have suggested varying options to try and improve accuracy, such as an arbitrary number of random samplings or statisticians painstakingly identifying areas from which to sample new training documents. The challenge with these approaches is that legal teams cannot estimate the cost of improving the accuracy of the predictions, leaving attorneys less informed in potential proportionality arguments.

This is a step within the predictive coding workflow that benefits from automation and visibility into the definitive amount of new training required to improve. Directed Training, a key element of Transparent Predictive Coding as part of the Clearwell platform, uses patent-pending active learning technology to automatically optimize each subsequent training set after the initial training set, selecting documents for training which will result in the highest improvement in prediction accuracy.

Actions in the Iterate step:
- The system identifies and presents the additional documents required to improve the accuracy.
- Reviewers tag the additional Training Set accordingly.
- The system utilizes the new training documents to improve its algorithm for identifying a responsive document.
- Users leverage the improved prediction model to predict codes on the total population set of documents.
- The testing process provides the details of improved recall, precision, and f-measure.

Conclusion

As the eDiscovery market better understands predictive coding, the promise of drastically reduced review costs and shortened review timeframes is fueling increased interest this new technology. Legal teams at corporations, government agencies, and law firms will justify predictive coding usage with cost savings. But just as importantly they will require a workflow that supports actual case needs, multiple eDiscovery goals, and confidence in its defensibility. Symantec offers an easy to follow predictive coding workflow directly integrated into the existing eDiscovery process that organizations follow today. The visibility and flexibility of Transparent Predictive Coding will enable organizations to adapt their process to specific cases and discovery goals while ensuring the defensibility of their eDiscovery workflow.
About Symantec
Symantec protects the world’s information, and is a global leader in security, backup, and availability solutions. Our innovative products and services protect people and information in any environment – from the smallest mobile device, to the enterprise data center, to cloud-based systems. Our world-renowned expertise in protecting data, identities, and interactions gives our customers confidence in a connected world. More information is available at www.symantec.com or by connecting with Symantec at go.symantec.com/socialmedia.

For specific country offices and contact numbers, please visit our website.

Symantec World Headquarters
350 Ellis St.
Mountain View, CA 94043 USA
+1 (650) 527 8000
1 (800) 721 3934
www.symantec.com