

Symantec Transparent Predictive Coding

Cost-Effective and Defensible Technology Assisted Review

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. However, until now there were limitations with available technology which slowed its adoption in the eDiscovery field. 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.

Cost-Effective and Defensible Technology Assisted Review

Symantec Transparent Predictive Coding

CONTENT

Introduction 1

Limitations of Linear Review 1

What is Predictive Coding? 2

Challenges with Today’s Predictive Coding Technology. 2

 The Need for Visibility into the Predictive Coding Process 3

 Defending the Review Methodology in Court 3

 Workflow Ease-of-Use and Flexibility 4

Transparent Predictive Coding. 5

 How it Works 5

 Use Cases for Transparent Predictive Coding 7

Conclusion 8

Introduction

The rapid growth of electronically stored information (ESI) has accelerated in recent years, prompting organizations to seek new ways of meeting their eDiscovery requirements cost-effectively. Since document review is the most time consuming and expensive aspect of eDiscovery, many of these efforts have focused on reducing review cost – either indirectly by reducing the amount of irrelevant data sent to the review team or directly through practices that improve review efficiency. One of the most promising developments in the area of review efficiency is predictive coding. Predictive coding uses software that assists reviewers in classifying documents according to criteria such as responsiveness and privilege.

Academic research has shown that predictive coding has the potential to help reviewers achieve accurate results at significantly reduced time and cost. However, this technology has not been widely adopted in the legal community due to limitations with early solutions that made it challenging to use predictive coding as part of a defensible review process. These early solutions often used a “black box” approach which obscured the prediction process from the reviewer. Without visibility into the prediction process, legal teams found it challenging to achieve high levels of review accuracy and were not confident in their ability to explain and defend their review methodology in court.

As a result, a more transparent approach to predictive coding, which takes into account the unique requirements of eDiscovery, is needed. 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.

Limitations of Linear Review

Many legal teams now routinely use eDiscovery search tools to help them identify and prioritize responsive documents, mitigate the risk of producing privileged information, and reduce the volume of data sent to review. While “eyes-on” linear review has traditionally been considered to be the most accurate way of assessing responsiveness and privilege, it is also time consuming and expensive. Gartner estimates the cost of reviewing one gigabyte of ESI at \$18,750.¹ In light of exponentially growing data volumes, searching and culling ESI before review is not in itself sufficient to adequately contain review costs. Improving the review process is seen as the next step in achieving further cost savings and perhaps addressing other drawbacks of linear review.

While linear review has long been considered the gold standard in discovery, recent research is beginning to call into question the results of traditional, “eyes-on” review methods. One of the problems is that across a review team it is virtually impossible to achieve consistent and accurate tagging decisions because different reviewers often do not categorize documents in the same way. This human error can be attributed to reviewers interpreting relevancy differently as well as reviewer fatigue, boredom, and inattention. A well-known study demonstrated that even under ideal conditions, reviewers only tag a document the same way 65 percent of the time.² While there has been a great deal of focus on measuring the recall and precision of search technology, until now there has been less attention on the accuracy limitations of linear review.

In search and information retrieval terminology, “recall” is the measure of all responsive documents across the organizations which are identified and produced. “Precision” measures the percentage of documents in the production set that are responsive. In other words, high recall means that there is less risk of under-inclusiveness, while high precision means there is less risk of over-inclusiveness and privilege waiver. The accuracy limits of traditional linear review inhibit recall and precision, regardless of the effectiveness of the search methodology. Since the eDiscovery process can be thought of as an exercise in maximizing both recall and precision, achieving lower review accuracy at high cost is not an ideal long term solution. Accordingly, many organizations are looking at ways of improving the review process to reduce costs and improve accuracy.

1-Debra Logan, John Bace, Gartner, E-Discovery: Project Planning and Budgeting 2008-2011, Feb. 2008.

2-Ellen M. Voorhees, Variations in relevance judgments and the measurement of retrieval effectiveness, 36:5 Information Processing & Management 697, 701 (2000)

What is Predictive Coding?

Over the last several decades there have been significant advances in the area of computing called machine learning. The goal of this research is to improve the ability of software to learn from inputs in the environment and use this information to make decisions. This technology is now being used in a variety of different ways ranging from filtering email spam to generating personalized recommendations on shopping websites based on an individual's purchase history. As an example, machine learning is also employed by the US Postal Service to sort and deliver 703 million letters and packages each day using handwriting recognition software.

This same machine learning technology can also be used to help make the review process more cost effective and accurate. The application of this technology to eDiscovery is commonly referred to as "predictive coding." It works by having software interact with human reviewers in order to learn the review criteria in the case. As reviewers tag documents in a sample set, the software learns the criteria for assessing documents and can generate accurate estimates of which tags should be applied to the remaining documents. As a result, reviewers can tag documents more efficiently, and it's even possible that fewer documents would need to be reviewed manually, which together could lower review costs considerably.

Although predictive coding is not a replacement for human reviewers, studies show that augmenting the review process with predictive coding can result in greater accuracy than traditional linear review. For example, relying on data from the TREC Legal Track, Maura Grossman and Gordon Cormack recently demonstrated in the *Richmond Journal of Law* that automated review approaches could yield more accurate results with less manual effort.³ A similar study comparing the automatic classification of documents to linear review concluded that "On every measure, the performance of the two computer systems was at least as accurate (measured against the original review) as that of a human re-review."⁴

Based in part on the results of academic research, members of the bench have begun taking note of this technology. Judge Andrew Peck, United States Magistrate Judge for the Southern District of New York, has commented about the potential benefits of predictive coding. He states "[i]n my opinion, computer-assisted coding should be used in those cases where it will help 'secure the just, speedy, and inexpensive' (Fed. R. Civ. P. 1) determination of cases in our e-discovery world."⁵ Internationally, predictive coding has even been referenced in court decisions. In a key legal opinion in the UK, Master S. D. Whitaker, Senior Master of the Supreme Court in the Queen's Bench Division, made reference to "software that will effectively score each document as to its likely relevance and which will enable a prioritisation of categories within the entire document set."⁶ However, despite promising research and interest from the judiciary, predictive coding has not yet been widely adopted for eDiscovery.

Challenges with Today's Predictive Coding Technology

Although predictive coding has a number of potential benefits, legal teams have found it challenging to achieve these benefits in real world cases due to limitations with the existing technologies. Early predictive coding solutions used approaches that are similar to the way machine learning technology has been applied in other domains, without addressing the unique requirements of legal review. As a result, organizations have often been hesitant to fully embrace predictive coding, typically confining its usage to a narrow range of matters like second requests.

There are three key challenges with first generation predictive coding technologies:

- Lack of visibility into the predictive coding process
- Difficult to defend black box processes in court
- Complex and non-intuitive workflow

3-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 (2011), <http://jolt.richmond.edu/v17i3/article11.pdf>.

4-Documents Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review, *Journal of Am. Society for Information Science & Technology* 61(1) (2010).

5-Search, Forward, October 2011 issue of Law Technology News

6-Goodale v Ministry of Justice 2009 EWHC 3834 (QB) Handled Down: 5 November 2009 Claim No HQ06X03876

The Need for Visibility into the Predictive Coding Process

The objective of a document review process during discovery is to ensure that information is accurately assessed for responsiveness and privilege to meet the production requirements of the court, government regulator or other third party. Failure to produce documents within scope (under-inclusiveness) can lead to sanctions and other penalties while producing too many documents outside scope (over-inclusiveness) increases the risk of releasing privileged or confidential information.

Numerous high profile cases involving sanctions have highlighted the issue of under-inclusive production, while recent cases such as the malpractice suit⁷ against the law firm of McDermott Will & Emery have highlighted the issue of over-inclusive production. In the McDermott case, the client alleged that a botched review resulted in the inadvertent production of privileged documents, bringing the issue of privilege waiver and the limitations of claw back agreements into the spotlight. In light of the risks of under-inclusiveness and over-inclusiveness, it's clear that any technology used to improve review efficiency cannot do so at the expense of accuracy.

Organizations found that first generation predictive coding solutions utilized a “black box” approach that obscured how predictions were generated. This opaque approach worked when machine learning technology was applied to other domains, but falls short of what is needed in eDiscovery. For example, it's often unnecessary to understand the criteria used to classify an email as spam or how a suggested book title on a shopping website is generated. However, in the eDiscovery realm, reviewers need to understand what information led to a particular prediction and how likely the prediction is to be accurate so they can confidently review the document in context and apply the appropriate responsiveness or privilege tag. In some cases, if a prediction is incorrect, it may also be helpful to understand the error in order to improve prediction accuracy for the rest of the case.

Regardless of the review methodology employed, legal teams require visibility into the accuracy of the document review at a case-level in order to make informed decisions on whether the document set is ready to be produced. One of the most efficient ways of measuring recall and precision is statistically random sampling. In addition to making informed judgments as to whether sufficient review quality has been achieved to permit production, sampling also enables informed arguments for proportionality. However, early predictive coding solutions did not provide an intuitive way for organizations to perform sampling according to EDRM⁸ and Sedona⁹ best practices. Lacking visibility into the prediction process, organizations have been concerned about the risk of under-inclusive and over-inclusive production, as well the defensibility of the process.

Defending the Review Methodology in Court

Over the last several years, the number of eDiscovery sanction awards has increased dramatically, nearly tripling since 2005.¹⁰ A recent study published in the Duke Law Journal analyzed sanctions, finding they span all types of cases and included sizable monetary awards, adverse jury instructions, and even outright dismissals. There is also a growing trend of sanctions in cases that involve mere negligence without any malicious intent, showing that the judiciary is losing patience with parties that have not taken appropriate steps to avoid incomplete or incorrect eDiscovery production. As a result, it's more important than ever to ensure adequate steps are being taken to avoid under-inclusiveness/over-inclusiveness and to demonstrate to the court that the underlying technology and processes are reasonable.

As legal teams look to new technologies like predictive coding to enhance the review process, a common question in the legal community is whether these technologies can be used in a defensible manner. If their use is challenged by opposing counsel in court, parties will likely have to explain and defend their approach. To minimize the risk of disagreements, the Sedona Conference Cooperation Proclamation¹¹ encourages parties to cooperate with opposing counsel during the initial stages of the case and reach an agreement on the methods and technology that will be used. While cooperation reduces the risk of disagreements, parties will likely need to be prepared to explain and defend their approach, in the event they can't obtain consensus.

7-J-M Manufacturing Co., Inc. v. McDermott Will & Emery, Los Angeles Superior Court, Case No. BC462832 (subsequently removed to US Dist Ct, C.D. Cal.)

8-EDRM Search Guide <http://www.edrm.net/resources/guides/edrm-search-guide>

9-Sedona Conference* Commentary on Achieving Quality in the E-Discovery Process

10-Sanction awards have increased 271% since 2005. Sanctions For E-Discovery Violations: By The Numbers in the Duke Law Journal 15 November 2010

In commenting on the defensibility of predictive coding, Judge Andrew Peck has outlined three key questions that parties should be able to address:

1. What was done?
2. Did the process produce defensible results?
3. Did the process produce “responsive documents with reasonably high recall and high precision”?

While these questions also apply to other review approaches, they are particularly important in the context of predictive coding because early solutions lacked the visibility needed to ensure the defensibility of review.

When using first generation prediction coding solutions, review teams found that the technology largely obscured the process used to generate predictions. In some cases, these solutions provided reviewers suggested tags without guidance on the likelihood of the tag matching the document, or even why the tag was suggested in the first place. In other cases, early technologies automatically applied tags without allowing reviewers to cross-check prediction accuracy. As a result, legal teams found it impossible to explain how documents were assessed for responsiveness and privilege and why specific documents were tagged the way they were in the production set. Additionally, organizations did not find the statistical sampling in these tools to be intuitive or robust enough to accurately measure recall and precision. Due to the lack of capabilities in early predictive coding solutions to address these questions, many legal teams did not feel comfortable they would be able to explain or defend their approach in court.

Workflow Ease-of-Use and Flexibility

The widespread adoption of search technology in eDiscovery was partly driven by the familiarity of using keywords to search the web and personal computers. Early predictive coding solutions, on the other hand, introduced a novel technology that was unfamiliar to reviewers. Because predictive coding relies on sophisticated computer algorithms to predict review tags, it is important to perform steps correctly and in the right order to achieve accurate results. However, organizations that adopted first generation predictive coding technologies found they required a high degree of manual intervention to use, making it difficult to ensure predictive coding best practices were being followed. As a result, legal teams often needed extensive training to use the software, causing delays that impacted court deadlines.

The manually-intensive workflow of early solutions not only increased the risk of errors, but also increased the likelihood that these errors would go undetected because of the difficulty of performing sampling. As discussed above, sampling is an effective technique for measuring review accuracy because random samples are able to serve as proxies for measuring the characteristics of a larger population, such as a large document set. However, sample sizes must be chosen carefully to produce accurate estimates, and documents must be selected at random. Once documents in the sample set have been reviewed, tags must be compared to tags in the original review set in order to calculate accuracy estimates. While these steps can be performed manually, this requires a higher level of experience and effort, making it more challenging for legal teams to use sampling to routinely measure review accuracy.

Early adopters of first generation solutions also discovered that their predictive coding workflows did not easily adapt to different types of cases. Early solutions were designed around the idea that predictive coding would be used the same way regardless of the type of case. However, there are many factors such as the type of case, the deadline, and budget which require different applications of predictive coding. For instance, in cases where review work was distributed across different teams of reviewers, the organization may seek to use predictive coding as a quality control measure to ensure the integrity of review already performed. These alternate use cases were not easily accommodated using the rigid workflows of early solutions.

Taken together, organizations are in need of an intuitive predictive coding solution which delivers a higher level of transparency, defensibility, and flexibility than existing technology.

Transparent Predictive Coding

Symantec's Transparent Predictive Coding is the first technology to open the black box of predictive coding by providing visibility into the prediction process, enabling more informed decisions and facilitating greater review accuracy. The solution provides an intuitive workflow that adapts to the unique requirements of each case, allowing reviewers to begin using predictive coding immediately and achieve optimal results. Also, each step in the workflow is documented by comprehensive reporting to help demonstrate the integrity of review to the court. Finally, as an integrated part of the Clearwell eDiscovery Platform, Transparent Predictive Coding was designed to simplify the use of predictive coding in conjunction with other steps in the eDiscovery lifecycle. The net result is a more cost effective, streamlined, and defensible predictive coding process.

How it Works

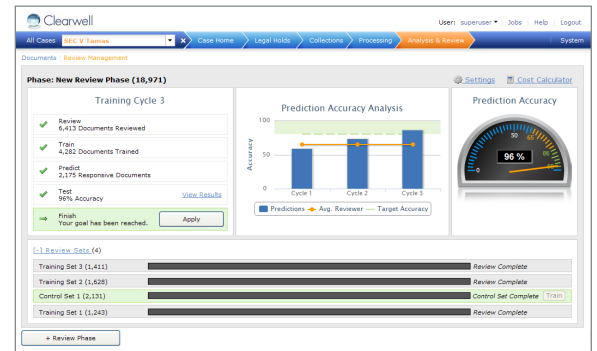
The Transparent Predictive Coding process begins with a set of case documents. Using Clearwell's powerful collection, processing, and culling capabilities, legal teams typically reduce data volumes by up to 90 percent before review. Once these steps are complete, case administrators manage the document review process from a centralized console that walks through each step in the Transparent Predictive Coding workflow, provides real time updates on review accuracy, and automatically indicates detailed next steps. This intuitive workflow ensures that each step is performed according to predictive coding best practices, and that reviewers can begin using predictive coding immediately while achieving accurate and defensible results.

The Transparent Predictive Coding process occurs in three phases:

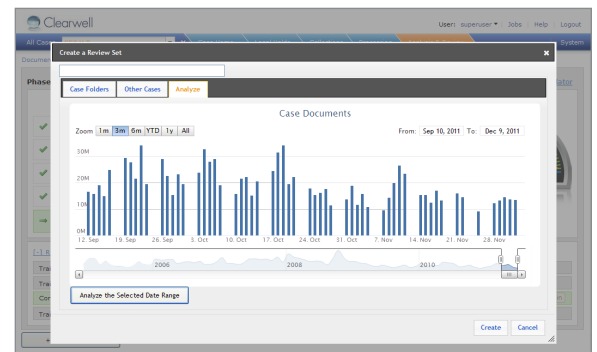
1. System Training
2. Applying Predictions
3. Quality Control

During the training phase, reviewers provide guidance on tagging criteria to the system so that Clearwell can learn review criteria and make accurate predictions. To perform training, reviewers tag documents in a training set, which is a small subset of the case documents. Transparent Predictive Coding leverages intelligent training sets that are identified using sophisticated analytics, streamlining the selection of highly relevant training sets which are optimized for system training. Next, reviewers log in and review documents by highlighting the specific metadata and content supporting their tagging decision, improving the accuracy of predictions and minimizing the number of training cycles.

As expert reviewers tag documents in the training set, the software identifies tagging criteria common across those documents, enabling it to "predict" the reviewers' tagging decisions for all documents in the case. Once training is complete, Clearwell builds a mathematical prediction model using the metadata and content of each document, taking into account the specific sections highlighted and tags selected by reviewers during training. Using



Prediction Workflow Management: Provides a review management console with step-by-step guidance that automates the predictive coding workflow.



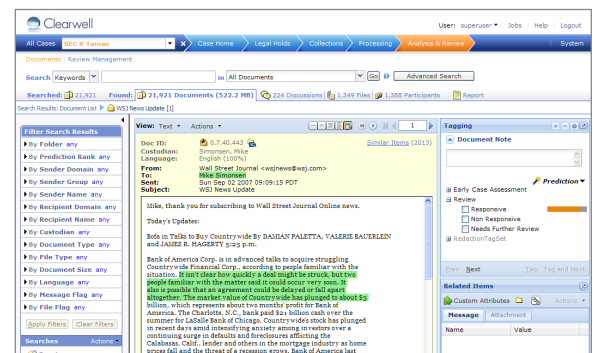
Smart Sampling: Offers sophisticated analytics to ensure the selection of highly relevant training samples.



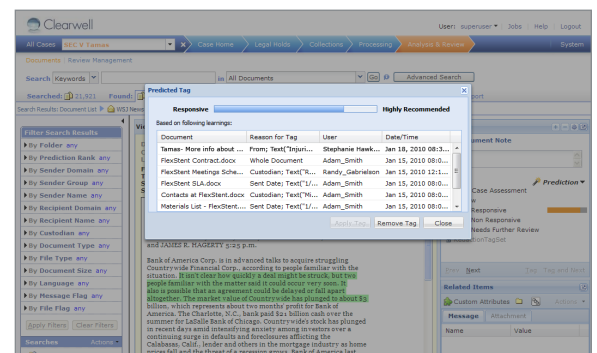
metadata in addition to content enables Clearwell to understand the full context of the document, such as the email sender or recipient, date, or file name which often hold importance for responsiveness and privilege.

Next, administrators apply the model to case documents and Clearwell automatically generates tagging predictions. Each tagging prediction has an associated prediction probability score, providing visibility into the likelihood of the tag's accuracy. At this stage, administrators may choose to bulk tag documents that have a high probability of a specific tag, or assign documents for review. During review, reviewers can leverage predictions to make faster, more informed tagging decisions. Using Prediction Insight, reviewers have visibility into why a tag was predicted for the document under review, allowing them to drill down and view highlighted sections that support the prediction. This visibility helps ensure reviewers make accurate decisions and legal teams can defend review workflows that leverage predictive coding.

During the quality control phase, administrators specify the accuracy requirements for a case and Clearwell automatically generates the appropriate random sample for accuracy testing, taking the guesswork out of random sampling. Once reviewers have reviewed these documents, Clearwell compares tagging in the quality control sample against predictions and generates an interactive set of charts and reports. Using these tools, administrators have visibility into the review's recall and precision, and the cost of achieving higher levels of accuracy, enabling them to make informed



Smart Tagging: Allows reviewers to highlight the metadata and content relevant to a tagging decision for more granular training.



Prediction Insight: Automatically provides a prediction probability score for the document under review and highlights content and metadata relevant to the prediction.

decisions and minimize the risk of under-inclusive and over-inclusive production. If the required level of accuracy is not met, Clearwell suggests specific next steps to improve accuracy, repeating phases 1 through 3 iteratively until the desired level of review quality is achieved.

Use Cases for Transparent Predictive Coding

For each different type of case across an organization there will be different review requirements. Transparent Predictive Coding offers the flexibility to leverage technology assisted review in a variety of ways depending on specific requirements of the case, such as the budget, timeline, and risk profile of the organization. For instance, organizations can use Transparent Predictive Coding to perform more effective culling, augment the linear review process through batching and quality control, or fully replace linear review with a complete automated workflow. Although not detailed here, Transparent Predictive Coding also delivers the ability to identify evidence during investigations and analyze productions from opposing counsel more effectively.

Culling

Before review begins, Prediction Templates can be used to apply prediction models across cases for more effective culling. For example, by leveraging prediction models created for other matters, case administrators can accurately identify and immediately set aside non-responsive items such as junk emails, system messages, and other profoundly irrelevant files. This capability supplements the powerful search and culling capabilities already available in Symantec's Clearwell eDiscovery Platform. Prediction Templates can also be used to set aside attorney-client or work product privileged documents before review begins in cases where privilege is similar across matters, resulting in reduced cost and risk.

Batching Strategy for Prioritized Linear Review

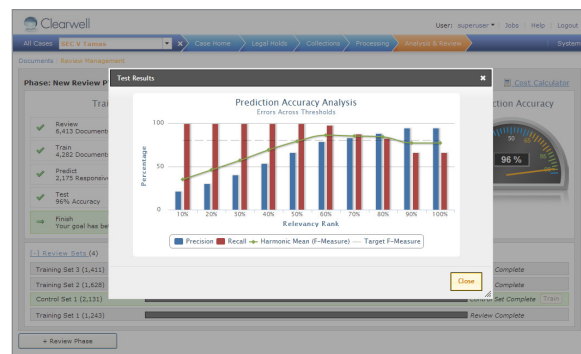
Utilizing Transparent Predictive Coding for first pass review enables review teams to segment and assign documents to different reviewers based on prediction probability score. For example, administrators can assign documents with a very high or very low probability of a tag to less experienced reviewers while assigning documents with greater ambiguity to more senior reviewers. This enables an efficient allocation of review resources leading to better utilization of reviewers and more intelligent linear review.

Quality Control

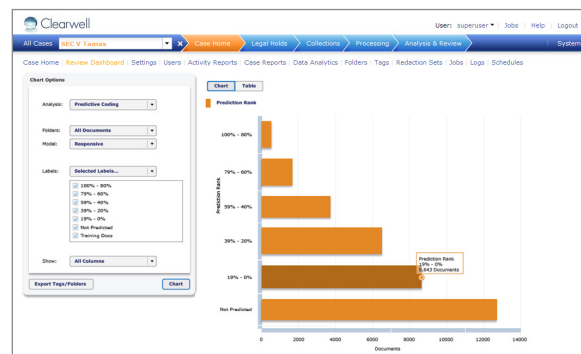
After traditional linear review has been completed, Transparent Predictive Coding can be used as a final quality control measure to ensure that no privileged documents are included in the production set. For example, a prediction model can be built using documents marked as privileged in the case, and applied across the documents in the production as a final step to confirm no errant privileged documents were missed during review. Given the accuracy limitations of linear review, this capability helps reduce the risk of privilege waiver and ensures higher review accuracy.

Complete Replacement of Linear Review

Depending on the case, deadline, budget, and risk profile of the organization, Transparent Predictive Coding may also be used as a comprehensive review workflow, allowing organizations to review a smaller number of the total case documents manually. In this case, the



Review Quality Control: Provides a comprehensive quality assurance workflow leveraging statistically valid sampling to assess and improve review accuracy.



Prediction Analytics: Delivers a set of interactive charts and reports that enables reviewers to measure prediction accuracy and analyze documents by probability score.

workflow described above would be performed iteratively until the desired level of accuracy is achieved. Using the review quality control mechanisms in Clearwell, review teams can leverage intuitive statistical sampling features to ensure a high degree of accuracy before production without manually reviewing all documents. Deploying the solution as a complete technology assisted review workflow allows organizations to achieve the highest level of cost savings over traditional linear review.

Conclusion

Predictive coding is one of the most promising technologies to reduce the high cost of review by improving the efficiency of the review process. However, until now there were limitations with available technology which slowed its adoption in the eDiscovery field. Transparent Predictive Coding introduces a new level of transparency, defensibility, and ease of use to technology assisted review, allowing organizations to confidently take advantage of this capability in the real-world. Whether the goal is to improve the culling of documents before review, augment linear review, or replace linear review using a complete predictive coding workflow, Transparent Predictive Coding delivers a more cost effective, streamlined, and defensible predictive coding process. While Transparent Predictive Coding is not a replacement for human reviewers, it can help review teams leverage machine learning to reduce the risks of under-inclusiveness and over-inclusiveness with more cost-effective and defensible results.

We plan to make the Clearwell Transparent Predictive Coding feature set for the Review & Production Module available in a future release on a when and if available basis in a general release to all customers.

Forward-looking Statement

Any forward-looking indication of plans for products is preliminary and all future release dates are tentative and are subject to change. Any future release of the product or planned modifications to product capability, functionality, or feature are subject to ongoing evaluation by Symantec, and may or may not be implemented and should not be considered firm commitments by Symantec and should not be relied upon in making purchasing decisions.

About Symantec

Symantec is a global leader in providing security, storage, and systems management solutions to help consumers and organizations secure and manage their information-driven world. Our software and services protect against more risks at more points, more completely and efficiently, enabling confidence wherever information is used or stored.

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

Symantec helps organizations secure and manage their information-driven world with [IT Compliance, discovery and retention management, data loss prevention, and messaging security](#) solutions.

Copyright © 2011 Symantec Corporation. All rights reserved. Symantec, the Symantec Logo, and the Checkmark Logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.
1/12