Relativity®

White Paper

Understanding the Components of Computer-Assisted Review and the Workflow that Ties Them Together

Contents

Executive Summary

In a computer-assisted review workflow, analytics technology leverages the expertise of humans to suggest coding decisions on all documents in a universe, basing its decisions off of a random sample set of documents coded by expert reviewers and lawyers. Reviewers code documents as they always have, while computer-assisted review works behind the scenes to propagate the decisions of the reviewers and decide if the remaining documents in a document universe are responsive or non-responsive. Human reviewer decisions are captured, analyzed, and used to teach the system to apply the reviewer's decisions to conceptually similar documents found elsewhere in the database. Studies have determined that this process can be more effective than linear review[1], while reducing the time and cost.

This paper will explore the computer-assisted review process, and take a look at one example of an effective computer-assisted review workflow.

1 Roitblat, H. L., Kershaw, A. and Oot, P. (2010), Document categorization in legal electronic discovery: computer classification vs. manual review. J. Am. Soc. Inf. Sci., 61: 70–80. doi: 10.1002/asi.21233.

Foreword

People used to say that "paper is patient." Words moved at human speed, no faster than we could read, write, or speak them—and compared to the writer at least, they lasted forever. The digital data we generate today may still outlive us all, but it's now much faster, larger, and a less-obedient servant. Data no longer waits around for us, or travels at human speed and scale. Instead, it's created spontaneously and orbits the globe electronically, often distributed to billions of other people who access it on any variety of devices and formats.

The accelerated pace, volume, distribution, and transformation of electronic data is leading to new challenges—and, subsequently, new approaches in dealing with it in discovery. In reviewing paper evidence, an attorney could once simply start at page one and methodically read document by document to the end. But investigators 'born with ink in their blood' need a new toolbox for navigating digital data.

> **"...investigators 'born with ink in their blood' need a new toolbox for navigating digital data."**

Computer-assisted techniques like keyword search are commonplace today in litigation support; new techniques leveraging machine learning, mathematical algorithms, and statistical sampling are emerging to generate greater power and consistency in helping machines retrieve documents with a higher likelihood of relevance. This is vital for some of today's biggest matters, in which large-scale document review may be too vast to complete manually in line with case or regulatory timeframes.

Computer-assisted review is not a matter of replacing attorneys, but arming them with the tools they need to practice effectively—because, while machines can offer greater speed and consistency than humans, they lack human creativity and the ability to make fine-grained distinctions about relevance. In fact, like a racecar with a teenage driver, they may wreck on the curves, or simply transport passengers to the wrong place faster.

Attorneys must learn to leverage this new technology successfully—not only to harness its power for the benefit of its speed and efficiency, but to defend and negotiate its use in court. While the law isn't known for keeping pace with changes in technology, the bench has shown a growing acceptance of (and even insistence on) the use of computer-assisted review in certain cases, as witnessed most recently in matters like *Monique da Silva Moore, et al. v. Publicis Group SA, et. al. and Kleen Products, LLC, et. al. v. Packaging Corporation of America, et. al.*

Digital data is not patient; it is oblivious. It won't stop growing, and it won't slow down to accommodate the slow-turning wheels of justice. Computer-assisted review is a vital tool in the litigator's belt for coping with the growing volume and velocity of discovery—one that's becoming increasingly risky for litigants to ignore. Without a good knowledge of it, litigants risk bringing a knife to a gun fight in court, or simply having a poor understanding of their available weapons.

Katey Wood, Analyst
Enterprise Strategy Group

Definition of Computer-Assisted Review

Computer-assisted review is more about process than product. The actual product is often a workflow tied to a text analytics system. This analytics technology leverages the expertise of humans to suggest coding decisions on all documents in a universe, basing its decisions off of a seed set of documents selected and coded by expert reviewers and lawyers.

This process is also referred to as technology-assisted review, predictive coding, etc. These terms refer to the same process of training a computer to group documents based on responsiveness using examples provided by humans—process plus technology.

Why There is a Need

The need for computer-assisted review begins with big data—an important trend in the corporate space. For this paper, we'll define big data as all of the electronically stored information being created in the enterprise. This includes the structured data that's being created in databases, as well as all unstructured data, such as emails, Word documents, and mobile info. Due to all of these sources of content generation, some 90 percent of the world's information was created in the last two years, according to Duke Chang, program director for Watson Solutions within IBM's Software Group.[2] However, this means the enterprise is seeing an overload of content, taking up the vast majority of their data storage.[3] Of course, much of this data is waste; for example, an email may be sent to 10 different people in an organization who make 1-2 minor changes to it before sending it to 10 others. Now, instead of a single email, an enterprise may have to deal with 50-60 emails that are basically identical.

There is a downstream effect here when it comes to e-discovery. When litigation or an investigation begins, there are now more data streams than ever to wade through. The following statistics are based on data mined from Relativity and analyzed by Detroit-based Strait & Associates, pointing to this increase in big data and the need for a review workflow to comb through documents faster than standard linear review.

- Comparing the 100 largest cases hosted in Relativity, the median case size grew from 2.2 million documents in 2010 to 7.5 million in 2011.
- Comparing the largest 5 percent of Relativity cases, the median case size grew from 2.1 million documents in 2010 to 3.5 million in 2011.
- Comparing the largest 10 percent of Relativity cases, the median case size grew from 700,000 documents in 2009 to 1.3 million in 2010 to 1.8 million in 2011.

This increase in data means that there may be significant time and money wasted having reviewers going through these documents, many of which may be irrelevant or duplicative.

“Comparing the 100 largest cases hosted in Relativity, the median case size grew from 2.2 million documents in 2010 to 7.5 million in 2011.”

2 Schifrin, Matt. (2012, May 9). Will Supercomputer Watson Be a Superhero for Banking? Forbes. Retrieved from http://www.forbes.com/sitesschifrin/2012/05/09/will-supercomputer-watson-be-a-superhero-for-banking/

3 Economist Intelligence Unit. (2011). Big data: Harnessing a game-changing asset. The Economist (September 2011). Retrieved from http://www.sas.com/resources/asset/SAS_BigData_final.pdf.

The Element of Computer-Assisted Review

A computer-assisted review system requires three main components—a **domain expert**, an analytics engine, and a method for validation.

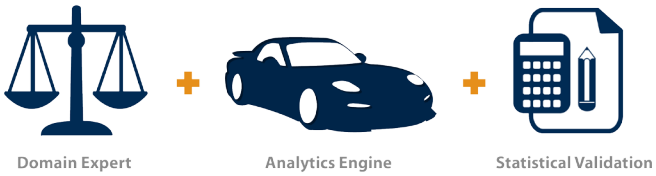

Figure 1: The Three Elements of Assisted Review

The domain expert plays an important role in the front-end of the computer-assisted review process. What computer-assisted review accomplishes is taking the expertise of these individuals and populating it to the rest of the document universe. As opposed to a review team made of up of 100 human reviewers—each with their own interpretation of the relevance of the case—the computer consistently applies the teaching of the domain expert. The system, combined with the domain expert, becomes a "force multiplier"—a military term describing how a highly trained team of few can defeat a loosely grouped team of many.

> **"The system, combined with the domain expert, becomes a "force multiplier"—a military term describing how a highly trained team of few can defeat a loosely grouped team of many."**

A number of different indexing and retrieval technologies can make up the analytics engine behind computer-assisted review. This technology is what sorts documents as responsive or not based on examples provided by the domain experts. Without the knowledge of reviewers, the computer has no way to learn and the engine is rendered useless.

Finally, the validation component is critical, as it allows lawyers to provide an audit trail for the efficacy of the computer-assisted review system. Lawyers can take a sample of the document universe and use **statistical sampling** to report that the system is actually achieving the results desired by the review team. Ultimately, it needs to be demonstrated that statistical evidence exists to prove the effectiveness of the review, and that the review is grounded in the same expertise used in linear review.

Rooted in Text Analytics

The computer-assisted review workflow we'll be discussing in this white paper uses a specific type of text analytics called **latent semantic indexing (LSI)**. In LSI, the analytics engine is provided with sample text for the system to identify the "latent" concepts in our documents. Once completed, a query can be mapped to these concepts. Ultimately, similar documents are identified based on context and concept rather than keyword. If "New York" and "Big Apple" occur together in many documents, a concept will be identified that includes both of these two phrases. If all goes well, for example, the computer will see the word "green" within an email about golf, and LSI will group documents that are about the sport, rather than documents exclusively about the color green. It is important to remember that LSI solely uses text in its decision-making; numbers, symbols, and images are not considered during the computer training process. After providing the computer with some example documents, the act of matching a new document to this sample set for a given topic is called **categorization**. Therefore, a few experts can identify documents relevant to a case so new documents

can then be "categorized" as relevant or not relevant based on the sample set. When computer-assisted review is tagging documents as responsive, it's simply categorizing all responsive documents.

Statistical Sampling and Binary Classification

Statistical sampling has been accepted as a validation technique in a number of industries. Since 1981, it has been an acceptable accounting practice to sample transactions to meet specific standards.[4] In manufacturing, sampling for defects—with techniques such as six sigma—has been used for years. Conducting surveys is another example, as survey results often determine a percentage approval rate with a defined margin of error. Within legal, judgmental sampling has been used in finding key custodians during identification and preservation. In computer-assisted review, a statistical sampling methodology can help validate a review for defensibility needs.

Fundamentally, categorization during the computer-assisted review workflow treats each document as a classification problem of sorting documents into similar groups. The groups can be identified as responsive vs. non-responsive, privileged vs. non-privileged, or one topic vs. another.

When sorting documents for computer-assisted review—or in any process where items are classified based on two possible choices—**binary classification** comes into play. In binary classification, there are always four possible outcomes based on the two possible choices in the decision. For computer-assisted review, this means a document could be classified as responsive or non-responsive, and that this coding decision could be right or wrong. In other words, every document will fit into one of the four categories in the diagram below.

True Positive: These are documents classified as responsive—or "true" regarding the document type being pursued—and that are indeed responsive.

False Positive: These are documents classified as responsive, but that are not actually responsive.

True Negative: These documents are classified as non-responsive, and are non-responsive.

False Negative: These are the documents classified as being non-responsive, but that are actually responsive and relevant to the case.

When looking at these four possible outcomes, it's clear that the true positives are where you ideally want your computer-assisted review workflow to take you. In a perfect system, computer-assisted review would code everything as true positives or true negatives and be correct every time, making it easy to reject all non-responsive docs. However, with four possible outcome scenarios, much goes into training the system to correctly categorize documents during computer-assisted review.

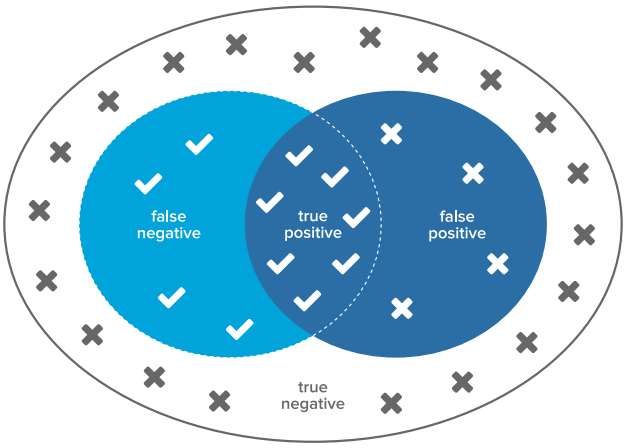

Figure 2: Results of Binary Classification

4 Akresh, Abraham D., "Statistical Sampling in Public Accounting", 50 The CPA Journal 20 (July 1980).

The Computer-Assisted Review Workflow

The specific workflow for computer-assisted review can be best explained with a simple example, which we will analyze as we progress. Let's envision a relatively small case of 500,000 documents.

In the first phase of the project, we start off by providing an initial set of human-coded documents to train the analytics engine on how to apply the computer tagging to the other documents in this review universe. These training documents are typically referred to as **seed documents.** Seed documents can be found by using previously coded documents, judgmental sampling techniques like keyword searching, or a random sampling methodology. In some cases, a mixture of all three approaches may be appropriate.

For this project, we'll say we have very limited time and budget, and that we have not begun a formal review. In this case, we would utilize random sampling methodology to make sure we have proper coverage of the types of documents in our case. In this case, we ask the computer to select a fixed amount of 500 randomly selected documents. (Many workflows today utilize a keyword search approach upfront to quickly identify potentially relevant documents; computer-assisted review can work in conjunction with this process or in lieu of it altogether.)

At this point, the computer randomly selects the sample set for our domain experts to code, which when coded becomes our seed set of documents. Once those 500 documents have been coded, the computer uses its text analytics engine to view and analyze the coded documents, and tags the remaining documents in the universe of 500,000. The computer then returns its results. Based on how the 500 sample documents were coded, the computer may suggest that 150,000 documents are responsive, 300,000 documents are not responsive, and 50,000 documents can't yet be determined based on the data provided. This process of coding sample documents followed by the computer coding the document universe accordingly is called a **round**. We then continue with additional training rounds until a marginal amount of documents remain uncategorized.

Once training is complete, there will be a validation round which utilizes statistical sampling of the documents coded by the computer. The review team can fine tune the statistical sampling parameters to meet the quality standards they set as a review team. In this case, let's say we selected a confidence level of 95 percent plus or minus 2.5 percent. If this is unacceptable to the review team, they can easily set the confidence level to 99 percent instead.

Once the sampling parameters are set, the computer will randomly select documents for the domain experts to review. The experts will either agree or disagree with the computer's coding decisions. At the end of this round, an overturn report is produced, allowing you to see how well the computer did, i.e. how many of the computer's decisions were overturned (false positives and false negatives), which documents they were, and which original document in the coded document universe led the computer to make this incorrect decision. In addition, the computer will also provide statistical information and project what percentage of the universe has been coded correctly by the engine.

Based on this overturn report, adjustments can be made—such as recoding the document that caused the computer to incorrectly categorize a different document—and the computer will then suggest domain experts code another sample set of documents to train the computer to more accurately code the rest of the universe. The computer will still base the size of this second sample on a 95 percent confidence level plus or minus 2.5 percent. In this case, the computer suggests another 1,500 documents to be reviewed.

This iterative process will continue until enough rounds have been completed for the computer to achieve its designated confidence level. The overall workflow will follow this pattern, starting with the seed set:

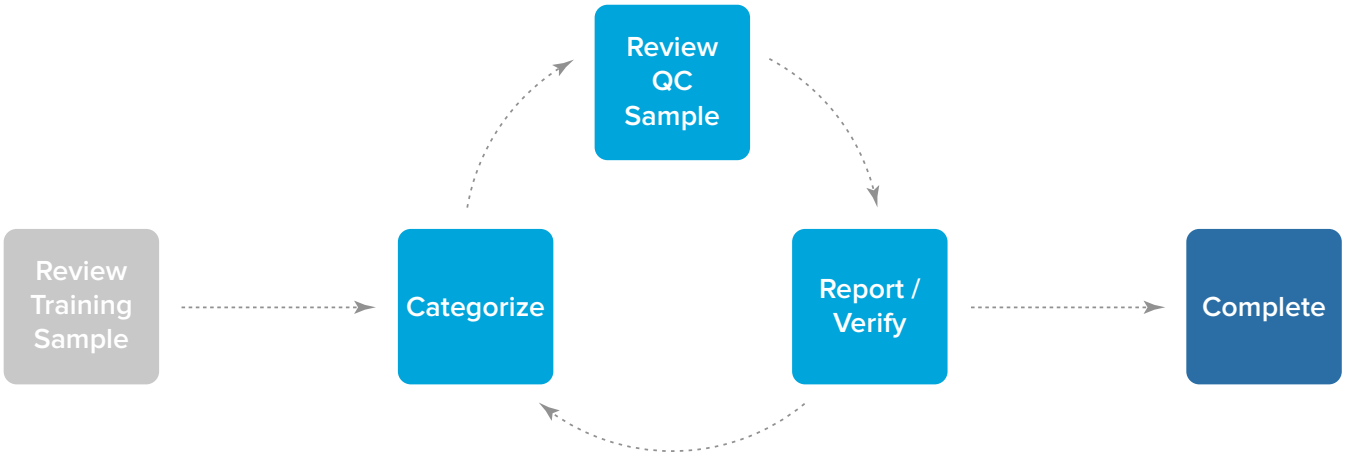

Figure 3: Assisted Review Workflow

At any point, lawyers may choose to end the review, understanding that the statistical sampling methodology has helped produce defensible results. In the end, the case team has reviewed 10,000 of the documents in the universe of 500,000, but may have produced results with the same level of accuracy as having reviewers spend months thoroughly examining all 500,000 documents. This is where the time and cost savings from computer-assisted review are manifested.Beyond this, lawyers can attack a margin of error with additional techniques, such as keyword searching or further computer-assisted review rounds. Lawyers may want to keep validating the documents and performing rounds to train the computer until it gets to the point where there is no incremental benefit from additional training and the review is complete, per the following graph:

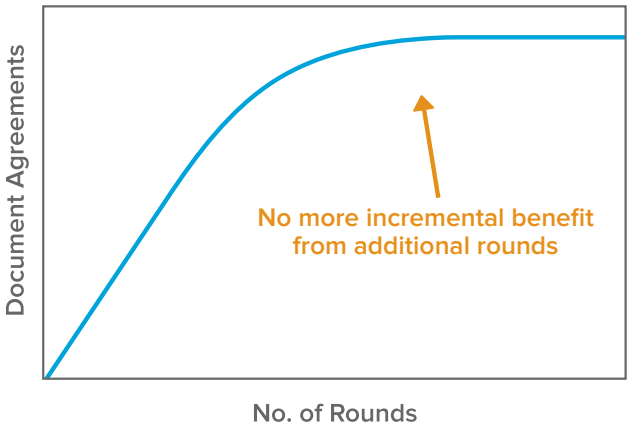

Figure 4: Plateau Effect of Assisted Review

Takeaways

A solid analytics engine coupled with statistical sampling can make for a successful computer-assisted review workflow. However, it's the human element—the lawyer reviewer providing true documents in the form of seed sets and managing the computer-assisted review workflow for consistency—that ensures computer-assisted review is an effective and defensible process. In the end, this workflow can speed up review, cull down the non-responsive "noise" documents in a universe from the get-go, and allow human reviewers to prioritize what's suggested as the most relevant documents early on in a case. Ultimately, immense cost and time savings can be realized by allowing the computer to become a "force multiplier" for domain experts—taking the expertise of these individuals and populating it to the rest of the document universe.

kCura®

231 South LaSalle Street, 8th Floor, Chicago, IL 60604
T: 312.263.1177 • F: 312.263.4351
sales@kcura.com • www.relativity.com