

# Intelligent Review Technology: Improving the Practice of Document Review in Legal Discovery

---

An examination of Intelligent Review Technology's application to the legal practice of document review



An Altegrity Company

<b>3</b>	The Need for Intelligent Review Technology
<b>4</b>	Intelligent Review Technology . . . What Is It?
<b>4</b>	Automating the Document Review Workflow
<b>5</b>	Prioritizing Documents Most Likely to Be Responsive
<b>5</b>	Technology-Aided Categorization
<b>6</b>	“Old School” Prioritization and Categorization
<b>7</b>	Supervised Learning: Ground-Breaking Technology to Fuel Intelligent Prioritization and Intelligent Categorization
<b>9</b>	The Application of IRT (Workflow, iP and iC) in Document Review
<b>11</b>	Sampling and the Evaluation of Classification Effectiveness
<b>12</b>	Variations on Workflow with iP and iC
<b>14</b>	Conclusion
<b>14</b>	About the Authors
<b>15</b>	Suggested Reading
<b>16</b>	About Kroll Ontrack

## The Need for Intelligent Review Technology

The uncontrolled proliferation of electronic communication and record keeping has transformed the practice of legal discovery. Today, legal professionals conduct discovery in a whirlwind environment, where legal and IT teams identify, preserve, review and produce responsive, non-privileged data to opposing parties, government agencies, and investigatory bodies. Before the Federal Rules of Civil Procedure were amended in 2006 to specifically provide for the discovery of Electronically Stored Information (ESI), lawyers and judges could resist the onslaught of technology in discovery, favoring traditional exchange of paper documents. Since 2006, federal courts require parties to fully comply with requests for production of ESI and leverage technological tools that can reduce the burden and expense of discovery. State and local courts are following closely behind. Thankfully, the legal community has discovered that technology has the power to improve the discovery process – and thus reduce the expense of litigation overall – by getting to the merits of a case early on, encouraging collaboration, and promoting principles of proportionality.

Despite these advancements, document review continues to be the most costly, complicated and time-consuming part of discovery. The vast volume of data encountered in discovery makes exhaustive linear review economically untenable, even when typical filtering by custodian, date, and keywords is first applied. Fortunately, a new generation of Intelligent Review Technology (IRT) promises to further reduce the expense and unpredictability of document review by identifying, prioritizing, routing, and categorizing documents that are most likely responsive.

A new generation of Intelligent Review Technology (IRT) promises to reduce the expense and unpredictability of document review.

Much like the skeleton that provides the structure for human organs, automated workflow provides the framework that supports iP and iC.

## Intelligent Review Technology . . . What Is It?

Intelligent Review Technology (IRT) attacks the bottlenecks of traditional review by automating the workflow, prioritization and categorization of documents for review. The three components of IRT are:

- » **Workflow Automation:** Automatically distributes documents for review while sampling them for quality control. Automating the process of document review minimizes manual effort and inconsistency in document staging, distribution, routing, assessment, and quality control.
- » **Intelligent Prioritization (iP):** Learns from reviewer decisions to identify and promote documents likely to be responsive. This leads to a virtuous cycle of reviewers seeing more important documents earlier in review, resulting in better learning and prioritization.
- » **Intelligent Categorization (iC):** Learns from reviewer categorization of small samples of documents and applies category suggestions for an entire collection. Categories can be used to organize and route documents, support better querying, and provide sophisticated analytics.

The remainder of this whitepaper examines the components of Workflow Automation, iP and iC; the role they play in document review; the technologies they are based upon; and the benefits they yield in comparison with traditional document review.

## Automating the Document Review Workflow

Until recently, document review relied upon human review administrators to batch and route documents to individual reviewers or review teams. This manual process required many hours of administrative work to create, maintain and oversee the process of document review, in addition to the time spent actually reviewing content. Progress reports and spreadsheets were maintained manually.

Today, advanced document review technologies incorporate automated distribution and quality control of documents during review and production. Automating the process or “workflow” of document review reduces the need for human administration, ensures an efficient and repeatable process, and provides for real-time reporting and metrics to display progress.

In addition, automating workflow enables heightened oversight and management that would have been impractically costly with manual routing. For example, automated workflow allows documents requiring specialized review (such as foreign-language documents, highly technical documents, or highly sensitive documents) to be routed to reviewers with appropriate skills or responsibilities. In addition, unbiased random samples may be continuously drawn and routed for assessment as new documents enter the system, enabling real-time estimation of not just what proportion of documents have been reviewed, but of what proportion of responsive documents have been found.

Finally, the automation of document review workflow provides a technological framework within which the other aspects of IRT can function. Much like a skeleton that provides the structure for human organs, automated workflow provides the framework that supports iP and iC.

## Prioritizing Documents Most Likely to Be Responsive

Before IRT, the standard practice was a “linear review,” i.e., documents were reviewed in the order they were received, or crudely prioritized by manually constructed filters (e.g., custodian). The luck of the draw determined whether important documents were found early or late in the process, and smoking gun documents might be missed completely during the hurried end-stages of a review. A number of studies have also shown disturbingly high levels of disagreement among reviewers on which documents are responsive when review is done in the traditional fashion<sup>1</sup>.

Intelligent Prioritization (iP) technology revolutionizes review by learning from the reviewer decisions as they occur and using automated workflow to route for early review the documents most likely to be responsive. This approach allows for earlier identification and review of case-critical documents and/or review by the most capable reviewers. This technology can also highlight manual review choices that are most likely to be incorrect or inconsistent.

## Technology-Aided Categorization

Responsiveness is never the only distinction of interest in review. Privilege must always be evaluated, lists of documents for each document request should be produced, and a review team will frequently wish to organize documents by important people, products, legal theories, or other topics relevant to a matter. Unfortunately, manual organization and categorization of documents has traditionally been time-consuming and error-prone. And, as with review for responsiveness, numerous studies have shown high degrees of disagreement among categories assigned by different manual assessors<sup>2</sup>.

The same supervised learning technology that enables iP to learn from review decisions can be used to enable Intelligent Categorization (iC). An entire collection can be automatically and consistently categorized based on learning from a modest number of manual category assignments. Categorization can thus be used for a number of purposes, including to aid in the search for particular documents, route documents to particular reviewers, or produce analytics enabling better understanding of the collection.

<sup>1</sup>

Roitblat, H.L., Kershaw, A., and Oot, P. (2010) Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61, 1, pp. 70-80. See also, Hedin, B., Tomlinson, S., Baron, J. R., and Oard, D. W. (2010) Overview of the TREC 2009 Legal Track. *The Eighteenth Text Retrieval Conference (TREC 2009) Proceedings*. NIST Special Publication SP 500-278. [http://trec.nist.gov/pubs/trec18/t18\\_proceedings.html](http://trec.nist.gov/pubs/trec18/t18_proceedings.html).

<sup>2</sup>

Cleverdon, C. W. (1991). *The significance of the Cranfield tests on index languages*. *Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '91)*. ACM, New York, pp. 3--12. See also, Lewis, D.D., Yang, Y., Rose, T.G., and Li, F. (2004) *RCV-1: A new benchmark collection for text categorization research*. *Journal of Machine Learning Research*, 5, 361-397.

Relying on an individual is counter to the implementation of repeatable discovery processes with predictable costs and quantifiable effectiveness.

## “Old School” Prioritization and Categorization

Automated prioritization and categorization both rely on *predictive models*, i.e., mathematical functions that take a document as input and produce a predicted class label (or probabilities or degrees of class membership) as output. The traditional approach to prioritization and categorization has been to build predictive models (typically called *queries* or *profiles* for prioritization, and *rules* or *classifiers* for categorization) manually, then apply them automatically.

Manual construction of predictive models in e-discovery requires a person who is expert in both searching text and metadata, as well as knowledgeable about the issues in the legal matter at hand. He or/she builds queries or rules that attempt to capture the classes of interest, tests them on case data, perhaps reviews results with attorneys, and attempts to understand the linguistic and statistical properties of the collection that possibly led to mistakes. This is an interactive process.

Relying on expert searchers to build models for prioritization or categorization poses obvious problems. Training personnel to the necessary level of expertise is time-consuming, and hiring experts is expensive. Further, the expert is needed not just once, but at many points during the review process. A batch of documents with new characteristics may arrive at any time, requiring substantial modifications to any previously developed model. The definition of responsiveness and of categories may also change with increased understanding of the case, receipt of new document requests, or negotiations among the parties. Further, the searcher ideally needs to be proficient with searching and quickly assessing class membership; the process is greatly slowed by constant consultation with lawyers and other experts. Unfortunately, deciding whether a document belongs to a class may require specialized legal knowledge, the ability to read foreign languages, or other familiarity with the matter at hand. There may simply be no single person with the mix of skills necessary to be an ideal interactive searcher and model developer.

Despite these challenges, skillful searchers (or knowledge engineers in some contexts) do sometimes achieve great results in building predictive models. But relying on the skill of a unique talented individual for prioritization and categorization is counter to the implementation of formalized, repeatable e-discovery processes with predictable costs and quantifiable effectiveness. Regrettably, litigation is not an occasional, unpredictable event for many organizations. Large corporations are routinely involved in dozens, even hundreds, of ongoing lawsuits at any moment in time. At this scale, document review needs to have more in common with industrial processes than it does with one-off Web searches or electronically digging for references for an academic paper.

Notably, courts have been increasingly scornful of informal, inconsistent, or under-inclusive approaches to document review and production, particularly when they have led to failures to turn over responsive material or flag information that is privileged and should not be produced. For example, in the well-publicized 2008 case of *Victor Stanley, Inc. v. Creative Pipe, Inc.*, Magistrate Judge Paul Grimm found that defendants’ reliance on an insufficient keyword search did not constitute a reasonable precaution to prevent the disclosure of privileged information. Case law instructs that the best defense is a systematic, well-documented process that can be equally applied in a range of cases.

## Supervised Learning: Ground-Breaking Technology to Fuel Intelligent Prioritization and Intelligent Categorization

An alternative approach to manually building predictive models relies upon machine learning, and in particular *supervised learning*. Supervised learning has become a dominant technology in many fields in recent years, including content management, predictive advertising, and customer relationship management. A supervised learning algorithm takes as input a set of documents which has been manually labeled with some distinction of interest: responsive vs. non-responsive, privileged vs. non-privileged, category A vs. B vs. C, etc. The algorithm outputs a predictive model that, when given documents as input, can be used to rank (prioritize) the documents and/or assign them to categories. Of particular interest are predictive models that estimate the probability that a document belongs to each class, since these probabilistic models can be used for both prioritization and categorization, and can be easily tuned to optimize effectiveness and minimize cost.

Supervised learning has profound advantages over manual construction of predictive models, particularly in e-discovery settings:

### **a. Focusing Efforts of Human Reviewers on Substantive Case Issues**

First, by separating the assessment of responsiveness from the creation of a prioritization model, there no longer needs to be a single person who both understands all the aspects of responsiveness and is expert at using search software. Legal teams can rely on technology to continually improve the prioritization model as more documents are received and reviewed. Human reviewers are free to focus their efforts on the content and substance of the documents, responding only to the substantive issues in the case. This approach improves a lawyer's ability to staff and train review teams and allows the simultaneous efforts of many reviewers to be leveraged to improve predictions of responsiveness.

The benefits for categorization are even larger. There is no longer a need for each reviewer to keep in mind, or even be aware of, all categories of interest when reviewing documents. When categorization is desired, small machine-selected samples of documents can be sent to a specific reviewer or reviewers for manual categorization, and then a learned classifier can be used to categorize all remaining documents. With this flexibility, categories can be added, deleted, or modified at low cost, and thus used more flexibly to aid filtering, routing, search, and analytics.

Supervised learning has profound advantages over manual construction of predictive models.

## **b. Building More Effective Predictive Models**

Second, supervised learning can, given sufficient training data, produce predictive models that are *more* effective than those produced by expert searchers. Human searchers must determine how to make use of the vast vocabulary of natural languages, including specialized uses within an organization. They must account for ambiguity, synonymy, and figurative and indirect expressions at all levels of linguistic structure.

Further, the behavior of a predictive model on actual documents is affected by unanticipated data anomalies of every sort, from simple typographical errors and spelling glitches, to the distortions introduced by automated speech recognition, optical character recognition, and machine translation. These difficulties are compounded by the fact that legal requests for documents are themselves posed in broad and often ambiguous natural language. Further, as documents are reviewed, the understanding of the case changes, in turn changing the definition of responsiveness.

Converting these murky, shifting distinctions into an effective predictive model is arguably easier for an automated system, which can examine massive amounts of changing feedback, than it is for a human searcher. Further, supervised learning algorithms can more easily differentially weigh large numbers of document features (each of which may contribute only a small amount to the accuracy of the prediction) than can a human searcher, as shown by the effectiveness of relevance feedback in improving search systems.<sup>3</sup>

A large data set at issue in discovery can easily contain hundreds of millions of distinct words, phrases, metadata attributes, and other features. Further, supervised learning allows the model to be updated on short notice whenever new or modified training data becomes available, properties of documents change, or criteria for effectiveness change.

It should also be noted that using supervised learning does not rule out benefiting from human search expertise when available. All learning algorithms can be aided by expert selections of particularly good or bad features, as well as more complex engineering of features. Indeed, any search query created by a human expert can itself be used as a feature in a predictive model.

## **c. Achieving Consistency in Prioritization and Categorization**

Neither predictive models nor human reviews are perfect in the judgments made for responsiveness or other classifications. Predictive models have the advantage, however, that their decisions are consistent; the same model given the same input will always produce the same prediction. The model can also be rerun at any time, allowing its predictions to be compared with human assessments to find conflicts between the two. Both the data that the model used to make predictions and the labeled data used to train that model can be examined at any time, if justification of review and categorization decisions is necessary.

## The Application of IRT (Workflow, iP and iC) in Document Review

The full power of Intelligent Review Technology is found in combining prioritization and categorization based on supervised learning with automated workflow. The process begins when the first documents are received at the start of a case. This initial delivery of documents to the review platform becomes the start of a growing document corpus. Typically no review has been done at this preliminary stage, so supervised learning is not yet applicable. The first documents to be reviewed might therefore be chosen through random sampling, by running a manually produced query or by utilizing other criteria already known to the attorneys in the case. Additional documents are always selected by random sampling, for use in producing statistically valid measurements of the progress of the review process and of the effectiveness of the predictive models.

The first documents selected for review are automatically batched and routed to human reviewers by the workflow system. The reviewers assess whether each document is responsive and, optionally, to which of the initially defined categories it belongs. In situations where a client did some review prior to documents entering the review platform, those initial assessments may be loaded and used as well.

Documents that have been reviewed for responsiveness are fed back to the supervised learning algorithm to train a prioritization model. Similarly, documents for which category membership has been determined are fed back to train classifiers for each category discrimination. The prioritization model is used to reprioritize the remaining documents, and a new set of high priority documents is queued for review. Classifiers for categories, once they have reached sufficient effectiveness, are used to recategorize the entire collection on each iteration, allowing categories to be used in searching, generating reports, or routing documents for specialized review. The random sample of reviewed documents continually generates the estimate of how many responsive documents from the collection remain unreviewed, along with statistics on prioritization and categorization effectiveness.

This process continues iteratively, with new deliveries of documents occasionally being received. With each new session of supervised learning, a new prioritization model is produced and used to re-prioritize all unreviewed documents. This allows for new aspects of responsiveness to be continually identified and applied to the documents. As documents are reviewed they continue to be fed back in to the prioritization model, improving both the training of the prioritization model and the estimation of the number of unreviewed responsive documents.

Note that the estimated (and actual) number of unreviewed responsive documents can increase as new documents come into the system, but then always decreases as responsive documents are found and reviewed. The estimated number of unreviewed responsive documents is continuously updated and can be used to guide decisions on timelines and staffing. After all documents have been received, decisions can be made about terminating or scaling down review, based on unbiased statistical estimates of how many unreviewed responsive documents remain.

Incorporation of iP and iC with automated workflow can greatly increase the efficiency of review. By finding more responsive documents earlier, iP helps reviewers, and the legal team, better understand the case early on. Similarly, iC

As documents are reviewed, they are fed back into the prioritization model, improving the training of the model.

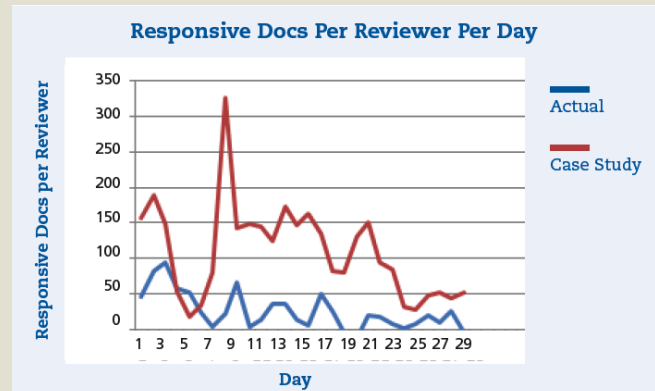
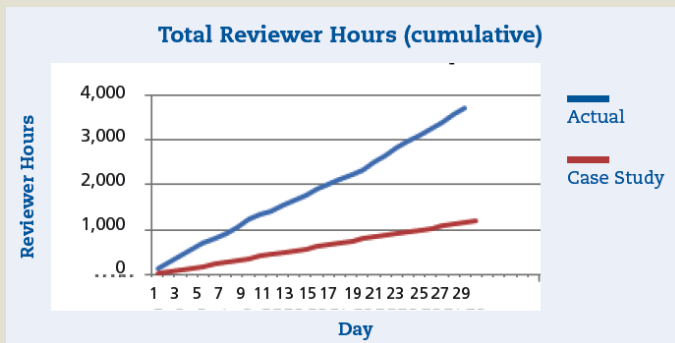
allows related documents to be dealt with more efficiently as a group, potentially by reviewers with particular expertise in that category of documents. Categories assigned by iC also reduce time spent searching for documents and preparing lists of documents responsive to particular requests.

A case study authored by Kroll Ontrack comparing and contrasting the application of Intelligent Prioritization to a document review project is discussed below. Preliminary findings (soon to be released) show that the application of iC would further decrease review hours and improve the consistency of the category determinations.

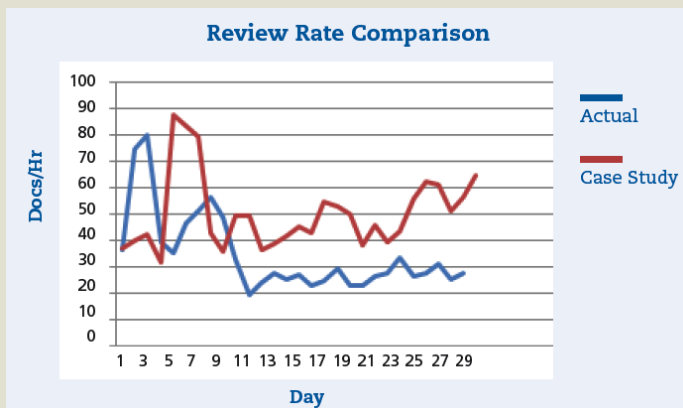
**Original Review Project:** Involved more than 100,000 documents using three responsiveness categories, several privilege categories and additional confidential and redaction categories. First-pass review was completed within 29 days and 21,000 documents were produced.

**Reenactment:** This reenacted project used a staff one-third the size of the original review team. On day 25 of the review, reviewers coded more than 21,000 documents as responsive, using only 28% of the labor hours utilized on the original project. Ultimately, the review was completed using less than half (about 46%) of the original review hours.

The number of responsive documents coded per reviewer each day improved. By day 7 of the review, the number of likely responsive documents routed to reviewers dramatically increased and then maintained a constantly higher level than the original review.



Average review rates also increased. As seen in the chart below, the number of documents reviewed per hour spiked early in the project and maintained a consistently higher rate when compared with the original review.



## Sampling and the Evaluation of Classification Effectiveness

Neither manual reviewers nor predictive models are perfect in their judgments. The advantage of IRT is that potential errors can be measured, monitored, controlled, and acted upon in a fashion difficult (if not impossible) in traditional linear review. The key here is statistical sampling. By randomly selecting a small subset of documents for careful and exhaustive review, the effectiveness of predictive models for responsiveness, privilege, and all other classification decisions can be estimated and tracked over time. Automated workflow enables this sampling to be done in an easy and unbiased fashion.

Such sampling must take into account the fact that documents to be reviewed are rarely received all at once. Ten documents might be delivered to the review platform on the first day, and 10,000 the next day. The composition and proportion of responsive documents in deliveries also may vary substantially. Statistical sampling methods implemented for IRT should therefore be adaptively tuned to achieve estimates that have the desired confidence at all times, while remaining unbiased by the order, size, and composition of document deliveries.

Note that sampling is desirable in review regardless of whether supervised learning or any other particular technology is used. Any classifications imposed in document review, even (perhaps especially!) routing or filtering done by manually written keyword queries, should be evaluated by sampling. The inevitable complexities of human language and human judgments suggest no such classification should be accepted on its face. As illustrated in the 2010 case from the Southern District of West Virginia, *Mt. Hawley Ins. Co. v. Felman Prod., Inc.*, the court instructed that practitioners should not wait for all documents to be received before starting sampling. Relying on *Victor Stanley* (discussed earlier), the *Mt. Hawley* court found the plaintiff failed to perform critical quality control sampling during review and concluded the plaintiff did not take reasonable steps to prevent disclosure.

How estimates of classification effectiveness are used varies somewhat between iP and iC. In iP, the central question at any moment is how many responsive documents have not yet been identified. By continually providing an estimate of how many unreviewed responsive documents remain in a collection (and updating that estimate as new batches of documents arrive), sampling gives a review manager tools for producing accurate and cost-effective reviews. Depending on resources available, manual review might be suspended and restarted at several points during a project, based on estimates of the completeness of review and information on the timing of future deliveries of documents. Such estimates might also be used to make a proportionality argument that review should be suspended because the number of responsive documents remaining does not merit the corresponding cost of “completing” the review. Finally, the concreteness of such estimates may aid collaboration among the litigant parties to agree on when the review should be considered complete. (We return to these issues in a later section on selective review.)

Sampling  
is desirable  
in review  
regardless  
of whether  
supervised  
learning is  
used.

# IRT platforms allow for flexible configuration of workflow.

For iC, the effectiveness measures of interest are typically focused on the classifiers themselves. A variety of effectiveness measures have been developed in information retrieval and machine learning, including recall (the proportion of documents belonging to a category that a classifier says belong to the category) and precision (the proportion of documents that a classifier says belong to a category that actually do). Other measures, including utility measures that explicitly put a financial cost on each classifier mistake, can be used as well. Estimates of classifier effectiveness can be used in deciding whether to do more training of a classifier, change the definitions of categories, or put the classifier into use.

## Variations on Workflow with iP and iC

The above section discusses one possible workflow incorporating iP and iC, but IRT platforms typically allow flexible configuration of workflow as appropriate for the individual review project. A few alternative workflows are discussed below.

### a. Predictive Models to Assess Categories to Determine Workflow

Predictive models can be used to determine who sees a document during review. For example, all documents that are assessed as responsive may also be reviewed for additional categories (privilege, in particular), with classifiers for these categories being continually retrained in the same fashion as the responsiveness model. Then, for instance, documents determined likely to be privileged might be routed to different reviewers than those merely likely to be responsive.

Categories could also be used more selectively. Consider a set of categories corresponding to key contractual clauses in a commercial litigation case. Only a limited set of reviewers might be asked to evaluate documents with respect to these categories. Once the classifiers for these categories reach a desired level of effectiveness, all human review on the categories might be halted (until and unless changes in the document population require more training to be done). The classifiers could then categorize the remainder of the document collection. The machine-assigned categories can then be used for a number of purposes including: to route documents for review, as search terms in queries, in driving graphical analytics displays or report generation, or for any other purpose that a manually assigned category in metadata might be used. The same technology can actually be used pre-review, for instance to support filtering in early case assessment.

### b. Predictive Models to Double-Check Human Review

While human judgment is always the final word on responsiveness or category membership, predictive models can play an important role in improving the consistency of review. By the end of the review process, one typically has trained highly effective predictive models for responsiveness and categories of interest. These models can be used on the manually reviewed documents to look for documents where the predictive model is highly confident in one classification, but the human reviewer made another. Some of these documents may simply exhibit weaknesses in the predictive model, and others may be instances where the correct classification is

unclear. But instances where the predictive model and human reviewer differ may be easily observed and corrected. Reconciling these instances is an easy and inexpensive way to improve the quality of the overall review. If desired, the process can be iterated, retraining the predictive models and doing additional rounds of correction.

### c. Predictive Models to Perform Selective Review

A natural question, given the effectiveness of predictive models produced by supervised learning, is whether it is plausible to skip review of some documents altogether. The fundamental questions here are legal: Under what circumstances would judges accept such procedures? When can they be negotiated with opposing counsel? When will clients be accepting of such procedures? Unfortunately, at the writing of this paper the authors are unaware of cases or statutes that provide legal guidance regarding these questions.

From a purely technical standpoint, however, IRT systems can support a range of approaches to such selective review. Three of the many hypothetical possibilities are outlined below:

- » Documents to which the prioritization model assigns a sufficiently low probability of responsiveness could be omitted from review. Alternatively, documents could be reviewed in order of probable responsiveness until sampling determines with high confidence that all but a specified number or proportion of responsive documents have been reviewed. A more radical possibility would be to automatically produce some documents without review if the probability of responsiveness is sufficiently high, relying on clawback to recover from mistakes.
- » Predictive models could be used, not for skipping documents but for guiding a review to read just the most important portions of long documents. This is essentially a turbocharged version of the keyword highlighting mechanisms used in many review interfaces.
- » A common concern in discussions of selective review is that some key or “smoking gun” documents might be missed even if few responsive documents are missed overall. Categorization can play a role in limiting this danger, by focusing extra review on documents likely to belong to sensitive categories or combinations of categories.

IRT systems  
can support  
a range of  
approaches  
to selective  
review.

## Conclusion

Today's unyielding siege of electronic information has forced litigants involved in document-intensive litigation to rely more heavily on technology during document review, largely considered to be the most expensive portion of discovery. IRT has the ability to help litigants better estimate and contain costs and uncover critical information early on, enabling a more informed case strategy earlier in the review.

## About the Authors

**David D. Lewis, Ph.D.** ([www.DavidDLewis.com](http://www.DavidDLewis.com)) is a consulting computer scientist working in the areas of information retrieval, data mining, natural language processing, and the evaluation of complex information systems. Previously he held research positions at AT&T Labs, Bell Labs, and the University of Chicago. He has published more than 50 scientific papers and seven patents, and was elected a Fellow of the American Association for the Advancement of Science in 2006. Dr. Lewis was one of the founders of the TREC Legal Track, the first large-scale open evaluation of information retrieval technology for electronic discovery in legal cases. As part of this work, he led the creation of the first large-scale test collection for experimentation on e-discovery.

**Regina A. Jytla Hagen, Esq.** is a managing staff attorney at Kroll Ontrack. In this role, Ms. Hagen tracks the evolving common and statutory law in the areas of corporate information management, electronic discovery, and computer forensics. Through the design and implementation of educational, thought-leadership programming, in addition to writing and speaking, she helps attorneys, litigation support professionals, and technicians understand the pressing legal and practical issues involved in electronic discovery. Prior to joining Kroll Ontrack, Ms. Hagen worked as an insurance defense trial lawyer, defending first and third-party claims in the areas of negligence, products liability and insurance coverage.

## Suggested Reading

A good introductory text on machine learning is: Mitchell, Tom M. *Machine Learning*. New York: McGraw-Hill, 1997.

A more mathematical treatment can be found in: Bishop, Christopher M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

Uses of supervised learning in document prioritization and classification are discussed in: Manning, Christopher D.; Raghavan, Prabhakar; and Schütze, Hinrich. *Introduction to Information Retrieval*. New York: Cambridge Univ. Press, 2008. See also, Croft, W. Bruce; Metzler, Donald; and Strohman, Trevor. *Search Engines: Information Retrieval in Practice*. Boston: Addison Wesley, 2010.

A good textbook on sampling is: Thompson, Steven K. *Sampling, 2nd edition*. New York: Wiley-Interscience, 2002.

The TREC (Text REtrieval Conference) evaluations at the National Institute of Standards have been at the forefront of evaluation methods for information retrieval systems. The early years of TREC and the nature of TREC evaluations are discussed in: Voorhees, Ellen M. and Harman, Donna K. *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge: MIT Press, 2005.

More recent TREC evaluations (since 2006) have included simulated e-discovery tasks. The overview papers on the TREC Legal Track in the TREC proceedings, available at <http://trec.nist.gov/proceedings/proceedings.html>, provide a good introduction.



For more information, call or visit us online.

**800.347.6105** in the U.S. and Canada

**+1.952.937.1107**

**[www.krollontrack.com](http://www.krollontrack.com)**

Copyright © 2012 Kroll Ontrack Inc. All Rights Reserved.  
Kroll Ontrack, Ontrack and other Kroll Ontrack brand and product names referred to herein are trademarks or registered trademarks of Kroll Ontrack Inc. and/or its parent company, Kroll Inc., in the United States and/or other countries. All other brand and product names are trademarks or registered trademarks of their respective owners.