FOMO and eDiscovery: a Producing Party's Perspective

Robert D. Keeling[1]
Michael B. Buschbacher[2]

* * *

Dr. Roitblat's article provides an important corrective to one of the most persistent eDiscovery misconceptions: that an 80% recall rate means that 20% of the relevant information has been missed. In fact, as Dr. Roitblat's analysis demonstrates, it means almost the opposite: in a review with a well-trained TAR model that generates an 80% recall rate, it is very unlikely that any important facts were passed over. This validates what many experienced eDiscovery practitioners have known for some time: that rarely, if ever, is there anything of significance to be found in the "null set"—*i.e.*, the set of documents that are deemed not responsive by the TAR model. If anything, Dr. Roitblat's paper overstates the likelihood of finding a key or "consequential" document in the set of unproduced materials.

This has important implications for how parties and courts should go about applying Federal Rule of Civil Procedure 26(b)(1)'s requirement that discovery be proportional to the needs of the case. In particular, Dr. Roitblat's article introduces a crucial distinction that transcends the specifics of his calculations and that should inform any discussion about whether a given recall rate is sufficient, namely, the distinction between responsive documents and consequential facts – *i.e.* facts that actually impact the dispute. While the number of responsive documents has typically been much greater than the number of consequential facts, this ratio has grown dramatically with the rise of the data economy. This is because, while the amount of data involved in litigation has skyrocketed, the number of documents ultimately used at trial has remained relatively unchanged. In other words, the number of needles has remained relatively constant over time, but the size of the haystack has grown much, much larger.

All else being equal, the increasing amount of data means that parties have to review more and more documents to arrive at the same number of facts. This, combined with the well-ingrained habit of drafting overbroad document requests has greatly increased the burdensomeness of discovery for producing parties relative to the size and complexity of the underlying dispute. Further, in addition to reviewing documents for responsiveness, producing parties bear the increased burdens of reviewing for attorney-client privilege and work-product. Privilege review and privilege log drafting, moreover, are the most complex aspects of production and the most difficult to outsource to TAR models.

---

[1] Robert Keeling is a partner at Sidley Austin LLP. He is an experienced litigator whose practice includes a special focus on electronic discovery matters. Robert is co-chair of Sidley's eDiscovery Task Force, and he represents both plaintiffs and defendants in civil litigation and conducts internal investigations in the U.S. and throughout the world.

[2] Michael Buschbacher is an associate at Sidley Austin LLP who represents corporate and individual clients at all phases of litigation with a particular focus on critical motions, appeals, and complicated legal matters arising in discovery. He is a member of the firm's eDiscovery task force.

This new normal is both time consuming and expensive. A 2012 RAND Report—*Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*—estimated that collection, processing, and review of data costs around $18,000 per gigabyte. To put this in perspective, many cases now often involve more than a terabyte (1,000 gigabytes) of data, and review costs in large matters often reach into the millions or tens of millions of dollars.

Today's colossal productions also pose risks to producing parties that have little or nothing to do with the case itself. Even with careful review protocols, large productions will inevitably contain substantial amounts of irrelevant proprietary business information and sensitive personal data, such as social security and credit card numbers, employee and patient health data, customer financial records, and intimate and potentially embarrassing details, such as family photos and personal medical information. The duplication, transfer, storage, and use of such documents greatly increases the risk that irrelevant private information will be inappropriately exposed, whether inadvertently or in a data breach.

Thoughtful use of eDiscovery tools can significantly reduce such burdens, and in fact often allow parties to develop a more accurate and complete understanding of the facts than was possible before the advent of eDiscovery. Technological advances—even those functions now thought of as "basic," such as search terms—are far more powerful at finding responsive documents than any number of human reviewers paging through documents in a warehouse full of bankers' boxes. Unlike human reviewers, machine learning programs do not get distracted or apply inconsistent standards.

But even with these developments, producing parties still struggle to find cost-effective solutions for dealing with the ever-growing sea of data. Dr. Roitblat's analysis points to a better way forward by undermining the premise that broad document requests are necessary to ensure that relevant information is not missed. By narrowing relevance criteria to focus more closely on uncovering facts that are likely to actually shape how a case plays out, the gap between technically responsive documents and those that contain consequential facts can be significantly decreased, ameliorating the burden and expense on both requesting and producing parties with little or no harm to the requesting party.

The need for narrower document requests is pressing and may be clearer if we give a concrete example. We were recently involved in a very large, multi-stage proceeding that went on for nearly two years, culminating in a lengthy trial. All told, we collected over 2 Terabytes of data from a long list of custodians. During the review of these documents, we relied heavily on a TAR model calibrated with a target recall rate of 75%. Over 7 million documents were deemed non-responsive without eyes-on review because they were below the TAR cutoff score—*i.e.* part of the "null set."

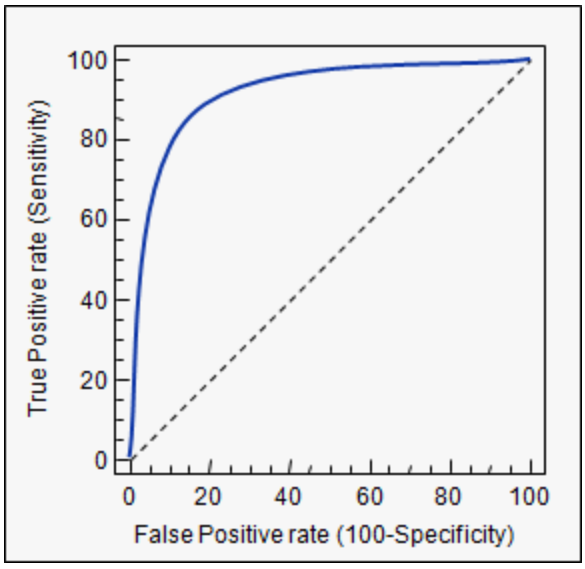
In order to ensure the effectiveness of this process, we performed two null-set samples. The first involved a total population (after privilege review) of 1,583 documents. Of these only 61 were deemed responsive and produced. The second set involved a total population (again after privilege review) of 3,167 documents, of which 167 were deemed responsive and produced. This works out to low total responsiveness rate of 4.8%, drawn from a quite substantial sample size. On the one hand, this suggests that that 4.8% of responsive documents

were not produced. But to Dr. Roitblat's point, none of the responsive documents found in the sampling were used as evidence at trial. In fact, fewer than fifty of the roughly five million documents eventually produced were actually used as trial exhibits—0.001%. Further, the documents that were used at trial all scored above the responsive cut-off.

This of course is just one case—it is not statistically significant on its own. But the lack of meaningful documents in the null set accords with our experience in numerous other matters. It also shows just how far apart meaningful factual development can be from technical responsiveness and how disproportionate (and enormous) the burdens of modern discovery can be on producing parties.

Those who would use TAR (or human reviewers) to cast an even wider discovery net in order to allay their fears of "missing out" draw exactly the wrong conclusions from these realities. Far from improving proportionality, seeking higher recall rates typically results in a "boil the ocean" review that imposes additional burdens on producing parties with little or no corresponding benefit. The reasons for this are rooted in the way that machine learning programs work. Generally, a well-trained TAR model will have to review an increasing proportion of non-responsive documents to find the responsive ones. As the ratio of relevant documents to non-relevant documents decreases, the number of false positives (*i.e.*, documents that are identified as relevant but in fact are not) inevitably increases. In other words, as the recall rate (True Positives / True Positives + False Negatives) increases, the precision rate (True Positives / True Positives + False Positives) will begin to decrease, often at a roughly exponential rate. Today's advanced machine learning tools allow this relationship—sometimes called the Precision-Recall Tradeoff—to be optimized to fit the needs of a given review, but even sophisticated models inevitably reach the point where the benefit of further review is outweighed by the rapidly increasing burden of finding additional technically relevant documents.

These realities can be graphically represented in what is known as a receiver operating characteristic curve—"ROC curve." An example of such a curve is shown below:



The greater the area under the ROC curve, the better the overall performance of the TAR model with respect to technical responsiveness. As our case suggests, increasing precision rather than recall may often be the more appropriate focus in order to achieve a discovery process that is proportional to the needs of the case. While there may be valid reasons in specific cases for a requesting party to sacrifice precision and overall performance in order to increase the recall rate, such decisions should not be based on unfounded fears or anecdotes but on a rigorous assessment of the actual costs and benefits to further discovery. If receiving parties wish producing parties to engage in further burdensome review to find the last theoretical (and likely non-existent) needle in the haystack, there is a solution: they can pay for it.

Dr. Roitblat has provided a useful theoretical touchstone for this discussion, one that we hope will continue to be refined through further research and analysis. Employing machine learning on eDiscovery matters offers the promise of substantial benefits and cost savings. Focusing on the documents in the null set threatens to significantly undermine those benefits and cost savings. The reality is that the overwhelming majority of documents that are responsive to broad discovery requests have no bearing on a case. By focusing on identifying documents that are actually consequential to a matter, we have a better chance of realizing what TAR can (and cannot) do effectively. And more importantly, focusing on the search of consequential documents will better allow courts and litigants to fulfill Rule 26(b)(1)'s proportionality mandate.