Retrospective and Prospective Statistical Sampling in Legal Discovery[1]

Richard T. Oehrle, Cataphora Legal, a division of Cataphora, Inc.(rto@cataphora.com)

DESI IV Workshop

Statistical sampling can play an essential double role in defining document sets in response to legal discovery requests for production. Retrospectively, looking backward at existing results, statistical sampling provides a way to measure quantitatively the quality of a proposed production set. Prospectively, looking forward, statistical sampling (properly interpreted) shows how the quality of a proposed production set can be improved. The proposed improvements depend on transparency of classification: in order to correct clashes between human judgments of sampled data and a proposed hypothesis, one must know the source of the misclassification. And in correcting such misclassifications, one must take care to avoid standard dangers like overfitting.

Section 1 below sets the stage by presenting the most basic material on statistical sampling and introducing the concept of data profiles. Section 2 argues that statistical sampling in retrospective mode is the only practical way to assess production quality. Section 3 offers several reasons why statistical sampling assessment has failed to become the standard practice it deserves to be in the field of legal discovery. Section 4 focuses on the use of statistical sampling in prospective, forward-looking mode, to drive iterative improvement. Section 5 describes at a high level some of our practical experience at Cataphora using iterative statistical sampling to force rapid convergence of an evolving responsiveness hypothesis with very high quality standards of review assessment. (In fact, the intrinsic role that statistical sampling plays in this process described in this section—a process that has consistently yielded measurably high quality—is one reason to consider the issues that arise in the preceding sections.) Along the way, we offer a variety of questions for discussion in a series of footnotes.

1 Background

1.1 statistical sampling, confidence intervals, confidence levels

Statistical sampling starts with a sample drawn from a data set. We cannot be certain that the sample is representative of the data as a whole. But we can estimate the likelihood that it is. This estimate takes the form of two hedges—a confidence interval and a confidence level. The confidence interval pads the particular results derived from the sample with room for error on both sides (say +/- 5%). The confidence level states how probable it is that any sample drawn from the data will fit within this interval. Intuitively, think of any distribution as being roughly like a bell curve, which prototypically has a central axis (the vertical line that goes through the top of the bell), with distribution falling off symmetrically on either side. Then think of a sample as a subset of the region between the horizontal x-axis and the bell curve. If the sample is randomly selected, because of the shape of the bell curve, most of the items in the sample fall within a relatively small interval flanking the central axis symmetrically on either side. This interval is represented by the confidence interval, and when the bell curve is relatively normal—not too flat—most of the points are not far from the central axis. Most is not the same as all, of course. And the confidence level is added to

[1]I'd like to thank the anonymous DESI IV referees for their constructive comments. Of course, any errors in this paper are my responsibility, not theirs.

deal with the outliers on either side that don't make it into the interval. (These outliers form the tails on either side of the bell that trail off on either side getting closer and closer to the x-axis the further away they are from the central axis.) If we claim a 95% confidence level, the claim is roughly that at least 95% of the points under the bell curve fall within the window and less then 5% of the points under the bell curve fall within the outlying tail on either side outside the window. This is why we need both a confidence interval and a confidence level.

1.2 data profiles

Many discussions of information retrieval and such basic concepts as *recall* and *precision* assume a binary distinction between *responsive* (or *relevant*) and *non-responsive* (or *non-relevant*). As anyone with any practical experience in this area knows, this is quite an idealization. To get a grasp on the range of possibilities that emerges from combining a responsiveness criterion with a dataset, it is useful to introduce the concept of a *data profile*.[2] Suppose we are given a dataset D and some omniscient being or oracle has the quick-wittedness and charity to rank every document on a scale from 0 (the least responsive a document could possibly be) to 10 (the most responsive a document could possibly be), and to provide us with a list of documents ranked so that no document is followed by a document with a higher rank. Here are some illustrative pictures, where documents are represented as points on the x-axis (with document d1 corresponding to a point to the left of the point corresponding to document d2 if document d1 precedes document d2 in the oracle's list), and the degree of responsiveness of a document represented by points on the y axis from 0 (least responsive) to 10 (most responsive).

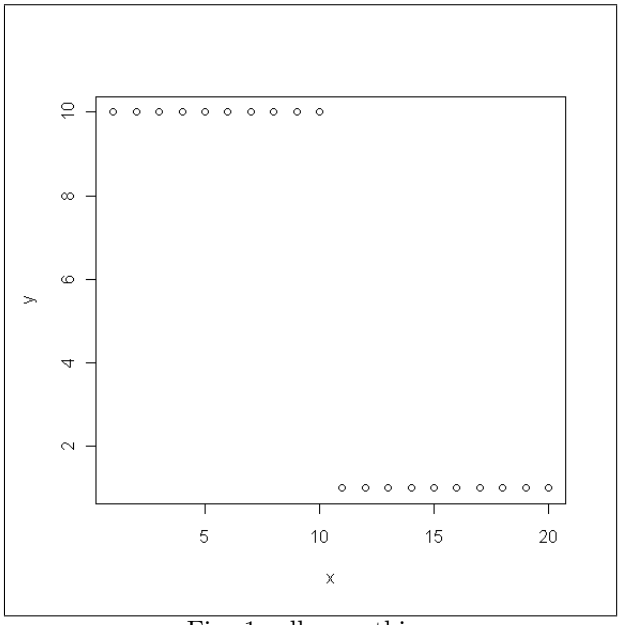

Fig. 1: all-or-nothing

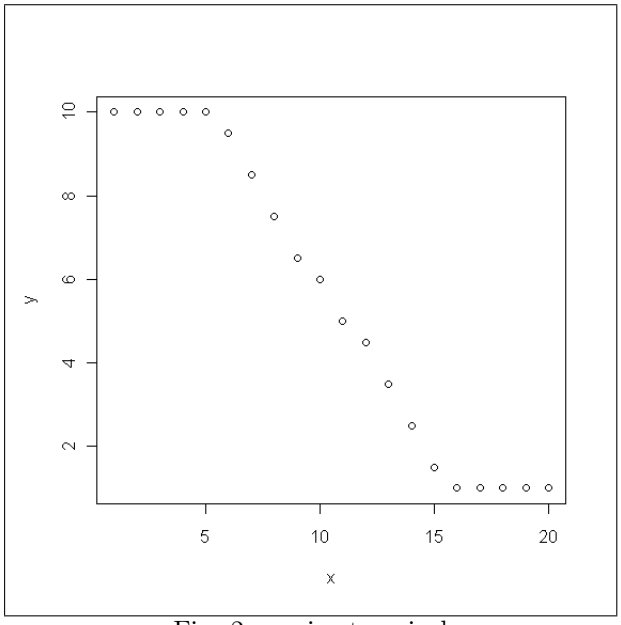

Fig. 2: semi-categorical

[2]The intuitions behind this concept have affinities with the AUC Game described by Gordon Cormack & Maura Grossman (2010).

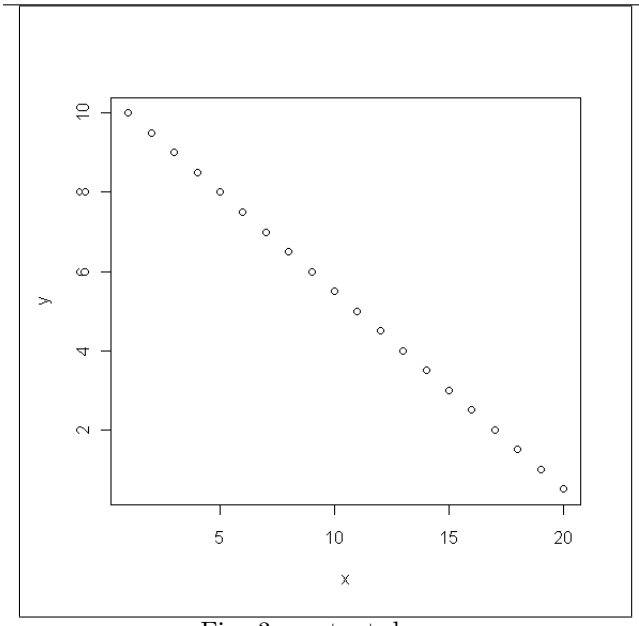

Fig. 3:constant decay

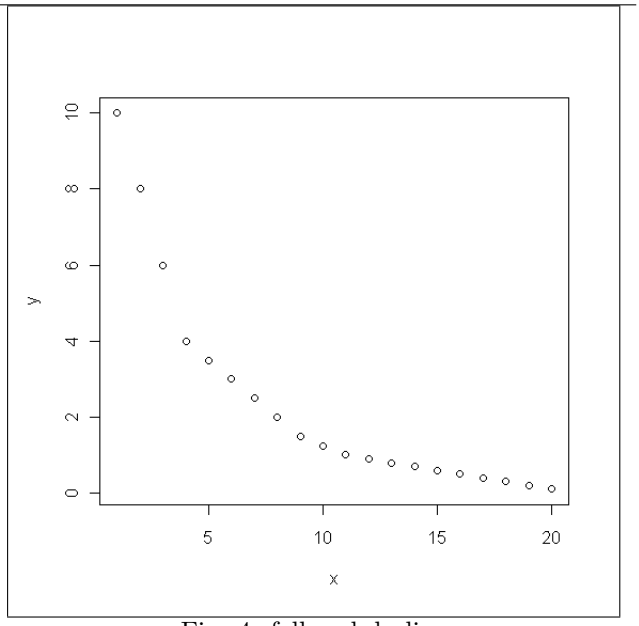

Fig. 4: fall-and-decline

Technically, the fundamental property of these graphical displays is that the relations they depict are weakly decreasing: if a point (x_1, y_1) is to the left of a point (x_2, y_2) (so that $x_1 < x_2$)), then y_2 cannot be greater than y_1. The point of introducing them is simple: they make it possible to apprehend and explore a landscape of theoretical possibilities which illuminates the practical questions that practitioners face.[3]

2 Statistical Sampling is the only practical way to assess production quality

Legal Discovery requires a specification (at some level of detail) of what is regarded as Responsive and what is regarded as Non-Responsive with respect to a particular document collection. A production set or potential production set drawn from this collection can be regarded as a hypothesis about which documents satisfy the Responsive specification.

When both the underlying dataset and the proposed production set are relatively large,[4] there is only one practical way to assess the quality of such an hypothesis: statistical sampling. Statistical sampling relies on a solid and well-understood mathematical foundation. It has been employed extensively across a broad range of subject matters. It is quantitative, amazingly efficient, replicable, and informative.

[3] **Question:** Given a data profile associated with a responsive criterion R associated with a request for production and a dataset D, what portion of the dataset should be produced?

[4] A referee has noted the potential and importance of hot-document searches, whose relative rarity may insulate them from the representative sweep of statistical sampling. We will come back to this point briefly in the final section.

2.1 human categorization does not in and of itself entail quality results

It is sometimes assumed that a quantitative assessment of production quality is unnecessary, on the grounds that the method used to define the candidate production entails its high quality. But assessment is completely independent of this process of definition. If we define a candidate production set by flipping a fair coin, we should still be able to assess the quality of the result. Historically, human manual review of the entire document collection has served as a benchmark of sorts, based on the assumption that human manual review must be correct. But there have always been skeptics who have doubted the efficacy of human manual review. And empirical investigations, which are not easy to arrange in practice, are beginning to show this assumption is in fact incorrect: defining a potential production set by human manual review does not guarantee a high quality result. (See, for example, Roitblat, Kershaw, and Oot (2010).)

Recently, there has been another version of this argument applied to automated methods of review. This version takes a form like the following: if a method can be shown to be consistently accurate across a diverse population of document collections, then we can assume that it will be consistently accurate when applied to a new collection that it has never been tested on. This formulation involves some delicate conditions concerning the properties of the document collections that form the provisional testing set. How could one be sure in practice that these same conditions actually hold when we move to a new document collection? The simplest way is to measure the quality of the results by statistical sampling. But in this case, it isn't necessary to rely on the delicate conditions at all: the quantitative assessment will provide the information needed.

2.2 human re-review: expensive, inefficient

One conceivable way to test quality is to re-review the entire document collection manually. This approach would be expensive and time-consuming. Furthermore, recent empirical research (such as Roitblat, Kershaw, and Oot (2010)) shows that multiple human reviews of the same document set yield astonishingly large disagreements in judgments.[5] In other words, apart from its expense and inefficiency, this approach is unlikely to provide a true assessment of the quality of a proposed production set.

Similarly, suppose the candidate production set was defined by a fully automated process. We can't test the process by re-running the fully automated process. If the process is consistent, then a second run will replicate the results of the earlier run, without providing any information about quality. Again, the remedy is to submit the results to statistical sampling.

2.3 informal QC vs. statistical sampling

Statistical sampling is sometimes replaced by informal browsing through a candidate production set. This differs from the statistical approach in a number of ways. For example, the sample set is not always selected appropriately. Moreover, quantitative results are not always tabulated. While the results of this seat-of-the-pants QC can be better than nothing, they do not provide the detailed insights available from statistical sampling.

[5] **Question:** if we consider a data profile associated with the responsive criterion and data set of the Verizon study, what is the corresponding *error profile*: that is, are clashes in judgment randomly distributed across the x-axis values? are they concentrated at the extremes of responsiveness / nonresponsiveness (represented by the left end and right end, respectively? are they concentrated in the middle? ...

2.4 if human review is fallible in general, why is it effective in sampling?

There are two practical reasons to distinguish the general properties of human review in large linear reviews and the general properties of human review in sampling reviews. First, because sampling review is remarkably efficient, it makes sense to employ senior attorneys with knowledge of both the details of the case at hand and the underlying law, rather than junior associates or contract attorneys or paralegals. (Compare the role of the *Topic Authority* in recent TREC Legal rounds.) In other words, the population is different, in a way that should (in principle) tilt the balance toward improved results. Second, our knowledge of the inconsistencies of multiple human reviews is based on large datasets with thousands of judgments. Since sampling reviews involve a much smaller dataset, clashes between a given reasonable hypothesis concerning responsiveness and actual expert reviewer judgments tend in practice to be even smaller. In fact, they are small enough to be subjected to individual examination, which sometimes confirms the expert reviewer, but at other times confirms the given hypothesis. This kind of detailed examination provides a highly valuable constraint on the quality of information provided by human reviewers, a constraint absent in the large scale multiple reviews whose differences have been studied. Finally, sampling review occurs over a time-span that lessens the risks of fatigue and other vicissitudes.

2.5 summary

In summary, if you want to know how good your proposed production set is, statistical sampling provides a quantitative, replicable, efficient, informative, practical, defensible answer. No other method known to us comes even close.[6]

3 Why isn't statistical sampling the de facto standard in legal discovery?

Properly conducted statistical sampling answers basic questions about the quality of legal discovery productions (up to approximations represented by confidence interval and confidence level). Why doesn't statistical sampling play a more central role when issues concerning discovery arise? Why isn't it regarded as reasonable and customary?

3.1 is no quantitative check needed?

One possible answer to this question (already discussed above) is that no quantitative check on quality is needed and if it isn't needed, it poses an additional and unnecessary burden. The primary justification for this answer is that the particular method chosen (manual, technology assisted, or fully automatic) serves as a guarantee of production quality. But empirical studies of manual review consistently show that manual review does not support this justification. And there is little reason to think that automated forms of review fare better. Moral: skepticism is called for.

[6] **Question:** where in the E-Discovery process should statistical sampling be employed? Example: if one side proposes to use *keyword culling* to reduce the size of the data and the associated costs of discovery, should the other side be provided with quantitative measures of the impact of this procedure on the responsive and non-responsive populations before and after culling?

3.2 what is a practical standard?

A related myth is that on the assumption that human review is perfect (100% recall and 100% precision), revealing actual sampling results will introduce quantitative figures that can never meet this perfect standard. It's true that statistical results always introduce intervals and confidence levels. And while such results can approach 100%, sampling can never guarantee 100% effectiveness. The practical impact of these facts is that some may feel that introducing statistical sampling results can only serve to illuminate defects of production. But if the introduction of statistical sampling results were the accepted practice, whether for manual or automated forms of review and production, it would very quickly become clear what the acceptable numbers for review quality actually are, what numbers require additional work, and what numbers are of high enough quality that further improvements would require increasing amounts of work for decreasing rewards.[7]

3.3 ignorance may be preferable to the consequences of knowledge

There is another possible factor which may have contributed to the failure of statistical sampling to be regarded as a reasonable and customary part of discovery. This factor has nothing to do with disclosing such results to the court or to other parties. Rather, it involves fear that the results will not satisfy one's own standards. Weighed in the balance, fear and ignorance trump knowledge and its consequences. Suppose you conduct a traditional linear review on a large document set. At the end of the review, you sample appropriately across the dataset as a whole to estimate the recall and precision of your candidate production. What if you were aiming for 90% at a minimum (with a confidence interval of 5% and a confidence level of 95%), but your sampling review shows that the recall is 75%. What choices do you face? Do you certify in some way a review that is plainly deficient in its results (even though it may have been conducted flawlessly)? Do you launch the manual review again from scratch, with all the attendant costs in time, effort, and money—and no guarantee in advance that the results of the second round of review will outperform the unsatisfactory results of the first review? One way to avoid this dilemma is to refrain from a quantitative estimate of production quality. The resulting shroud of ignorance obscures the painful choice.

This situation is not restricted to cases involving human manual review. For example, a similar dilemma would arise in circumstances in which the initial results depended on a black-box algorithm—that is, an automated approach that offers a hypothesis about how documents are to be sorted in a Responsive set and a Non-Responsive set, but does not reveal the details of how the hypothesis treats individual documents. For example, think of clustering algorithms that can be adjusted to bring back smaller or larger clusters (by strengthening or relaxing the similarity parameters). In the face of unsatisfactory recall results, one might be able to adjust the algorithm to drive recall numbers. Typically, however, this very adjustment adversely affects precision numbers, because additional documents that are in reality non-responsive may be classified as Responsive.

[7] **Question:** who should have access to sampling numbers? Example: does the counsel for the sampling side want to know that the sampling numbers are not perfect—they never are, for reasons discussed above—even they may far exceed contemporary standards? **Question:** what role do sampling measures play in defending production quality before a judge? Example: can the opposing side reasonably demand statistical measures of recall and precision when the quality of a production to them is in question?

3.4 making reality your friend

Not every method of defining a production set faces this dilemma. In the next section, we discuss the conditions needed to leverage statistical sampling results to improve review quality. And subsequently, because the necessary conditions are somewhat abstract, we discuss our experience at Cataphora using this iterative review model over the past seven years.

4 Leveraging statistical sampling results prospectively for hypothesis improvement

If the results of statistical sampling can be used to improve a hypothesis about a potential production set—that is, improve recall and improve precision—then a system based on successive rounds of sampling and hypothesis refinement can return better and and better results.[8] Before considering how quickly this convergence takes place in practice in the next section, we focus first on exactly how sampling can be leveraged for hypothesis improvement.

Suppose you review 800 documents selected to test the recall of your current hypothesis. This is a test of completeness, whose goal is to ascertain whether the current hypothesis mischaracterizes Responsive documents as Non-Responsive. Suppose that the resulting review judgments are as follows:

	hypothesized Responsive	*hypothesized NonResponsive*
judged Resp	80	40
judged NR	40	640

This sampling review thus confirms that the current hypothesis is underperforming with respect to recall: 40 documents that were hypothesized to be NonResponsive were judged Responsive. We might think that 40 is a relatively small number: only 5% of the 800 document sample. But this represents a third of all the documents judged Responsive in the sample. Suppose that the document set as a whole contains 800,000 items. If the sample is representative, 120,000 of them are responsive and our current hypothesis only identifies 80,000 of them. This is clearly unacceptable. The hypothesis needs to be revised and substantially improved.

Let's assume that all the clashes between review judgments and the current categorization hypothesis are settled in favor of the review. (In practice, it happens that further scrutiny of clashes can lead to a resolution that favors the categorization hypothesis rather than the human review.) Having determined the identity of 40 false negatives, the least we can do is to ensure that these 40 are re-categorized in some way so that they are categorized as Responsive. But this is obviously insufficient: the 40 false negatives are representative of a much larger class. It's this larger class that we must be concerned with and we want to use information extractable from the 40 false negatives to improve our hypothesis. What we seek is a way of categorizing these 40 false negatives that generalizes appropriately over the data as a whole. Two basic cases arise.

In the first case, the 40 false negatives are categorized by the current hypothesis, but the combination of the categorization components involved are incorrectly associated with NonResponsiveness, rather than Responsiveness. By adjusting the way in which such combinations of categorization

[8] See the work by Grossman & Cormack, cited earlier, and Büttcher, Clarke, and Cormack (2010), especially section §8.6 *Relevance Feedback*. What we describe here is a form of iterated supervised feedback involving both relevant and non-relevant information.

components determine Responsiveness or NonResponsiveness, the performance of the current categorization hypothesis can be improved in a suitably general way. (This is the very situation that Cataphora's patented Query Fitting Tool is designed to address, particularly when the number of categorization components is so high that finding a near-optimal combination of them manually is challenging.)

In the second case, the 40 false negatives are not categorized at all or are categorized in a way that is overly specific and not suitable for generalization. In this case, we seek to divide the 40 documents into a number of groups whose members are related by categorization properties (topics, subject line information, actors, time, document type, etc.). We next add categorization components sensitive to these properties (and independent of documents known to be NonResponsive) and assign documents satisfying them to the Responsive class.

The result of these two cases is a revised categorization hypothesis. It can be tested and tuned informally during the course of development. But to determine how well performance has improved, it is useful to review an additional sample drawn in the same way. If the results confirm that the revised categorization hypothesis is acceptable, the phase of hypothesis improvement (for recall, at least) can be regarded as closed. (When all such phases are closed, a final validation round of sampling is useful to ensure overall quality.) On the other hand, if the results suggest that further improvements are indicated, we repeat the tuning of the categorization hypothesis as just outline and test a subsequent sample. In this way, we get closer and closer to an ideal categorization hypothesis. In practice, we get a lot closer with each round of improvement. (Details in the next section.)

4.1 transparency

There is one critical point to note about this iterative process: it depends critically on the transparency of categorization. In order to improve a categorization hypothesis, we need to know how particular documents are categorized on the current hypothesis and we need to know how these documents will be categorized on a revised hypothesis. If we cannot trace the causal chain from categorization hypothesis to the categorization of particular documents, we cannot use information about the review of particular documents to institute revisions to the categorization hypothesis that will improve results not only for the particular documents in question but also for more general sets of documents containing them.

5 Cataphora's practical experience: empirical observation on the success of iterative sampling

We've shown above how statistical sampling results can be used to both measure performance and drive hypothesis improvement. If we follow such a strategy, results should improve with each iteration. But how much better? And how many iterations are required to reach high-quality results? In this section, we address these questions from a practical, empirically-oriented perspective. Cataphora Legal has been successfully using statistical sampling to measure performance and to improve it for almost a decade. In what follows, we draw on this experience, in a high-level way (since the quantitative details involve proprietary information). Our goal is not to advertise Cataphora's methods or results, but to document the effectiveness of statistical sampling review in the development of high-quality hypotheses for responsiveness categorization.

Before discussing project details, it is worth pointing out that statistical sampling can be integrated with other forms of review in many ways. As an example, it may be desirable and prudent to review the intersection of a responsive set of documents and a set of potentially privileged documents manually, because of the legal importance of the surrounding issues. As an other example, it may be desirable to isolate a subpopulation of the dataset as a whole to concentrate for manual hot-document searches. In other words, different techniques are often appropriate to different subpopulations. Overall, such mixed methods are perfectly compatible.

In a recent project, we developed a hypothesis concerning responsiveness for a document set of over 3 million items. This work took approximately 2 person-weeks, spread out over 3 months (not related to our internal time-table). Attorneys from the external counsel reviewed a randomly selected sample of the dataset. The recall of our hypothesis exceeded 95%. Precision exceeded 70%.

Not long after this sampling review occurred, we received additional data, from different custodians, as well as some modifications in the responsiveness specification. After processing the new data, we arranged a sampling review, using the previously developed responsiveness hypothesis. In this second sample, involving significant changes to the population, the performance of the original responsiveness hypothesis declined considerably: recall dropped to about 50% from the previous high on the original data. We spent days, not weeks, revising and expanding the responsiveness hypothesis in ways that generalized the responsive review judgments in the sampling results. At the end of this process, the attorneys reviewed a fresh sample. Results: the recall of the revised hypothesis exceeded 93%; precision exceeded 79%.

These numbers compare favorably with publicly available estimates of human manual review performance. The dataset involved was large. The overall process was efficient. In the present context, what is most notable is that the convergence on high quality results was extremely quick and the role played in this convergence by statistical sampling was significant.

References

Büttcher, Stefan, Charles L. A. Clarke, and Gordon V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*, Cambridge: The MIT Press, 2010.

Gordon V. Cormack and Maura R. Grossman, *TREC Legal Track—Learning Task Draft Guidelines*, `http://plg.uwaterloo.ca/~gvcormac/legal10/legal10.pdf`, (2010).

Roitblat, H., A. Kershaw, and P. Oot, '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, 2010.