THE ISSUES BEHIND THE HEADLINES

ESTIMATING THE COST BURDEN OF E-DISCOVERY

A New and Better Method

James. M. Wright, P.E.

TABLE OF CONTENTS

On Dec 1, 2006 the Federal Rules of Civil Procedure [FRCP] were modified, having a profound impact on litigation in U.S. Federal Courts. Specifically, the rules relating to the discovery of Electronically Stored Information [ESI] were enhanced and expanded, largely codifying the trend in case law that had been developing for several years. The effects of these new rules are still emerging, and it would be impossible to predict the full impact for many years, which is complicated by the advancements in technology that are arising to deal with the enormous volumes of ESI now potentially in play in litigation. State courts are largely following suit, adopting similar rules, or even the Federal Rules themselves [e.g. New Jersey]. Local Rules are also becoming common, generally following the same guidelines as the FRCP. Even Canada and the UK are following suit, although 2-3 years behind the USA.

The biggest impact on litigation is a direct result of the huge volume[s] and wide variety of types of ESI, which require different methods and protocols to collect, analyze, search, process, etc. for discovery. It is common now to find thousands of different file types on a single hard drive, many of which are difficult [or nearly impossible] to identify the application that created them, and many of which are not electronically searchable, necessitating costly human review. Add to this confusion the many and increasing types of non-PC generated ESI such as Instant Messages, Text Messages, Cell Phones, VOIP [internet phones], digital photos, voice mail, Blackberry data, etc., etc., etc. and you begin to grasp the scope of the problem. ESI is everywhere and it's discoverable.

The impact of all this on litigation is staggering. The cost of discovery is now potentially order[s] of magnitude greater than with the old days of paper. This has drastically changed the nature of smaller matters as well as David-Goliath disputes, as the relative discovery burdens have become much more disproportionate. It is widely recognized in corporate law departments that the discovery of ESI is the greatest, largely uncontrolled cost growth area for the foreseeable future. In short, Electronic Discovery is one of, if not THE biggest negative impacts to the bottom line and it's only expected to get worse.

An industry of technology service providers [aka "Vendors"] has arisen to assist with the processing, searching, review and production of this ESI. In 2007 this industry had revenues in excess of $2BB, and as 2008 began, there were over 500 such vendors offering a dizzying array of solutions and services ranging from complete in-house enterprise applications to ED-in-a-box products, to full soup-to-nuts outside service offerings, each and every one assured to facilitate the ED process, saving time and money. It would take small team of full time specialists just to keep track of all the offerings.

Despite the fact that all this technology is helpful, its purpose is to facilitate the processing of ESI that is subject to discovery. It would be even more helpful if the amount of ESI subject to discovery were limited in the first place, before sending it to a vendor. Fortunately, the new FRCP contain provisions for just such limitations.

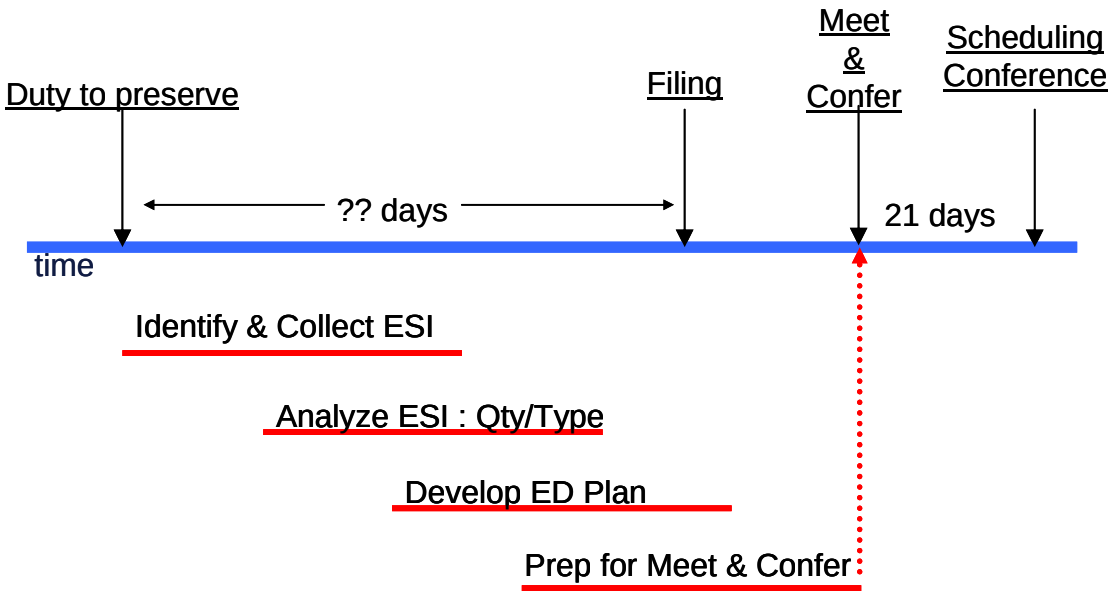
LIMITING THE SCOPE OF E-DISCOVERY

One of the more important provisions of the new FRCP is the establishment of a two-tiered categorization of ESI as it relates to discovery. While essentially all ESI is potentially subject to discovery, producing parties can now endeavor to exclude certain portions of their ESI from the obligations of discovery by designating it as "Not Reasonably Accessible [NRA] due to undue burden or cost" under Rule 26[b][2][B]. It is important to note that unlike before, this designation must be affirmatively made in order to be valid. Previously, parties could exclude ESI from discovery by deeming it to be beyond the "reasonable search" standard, but had no obligation to identify its existence. Requesting parties could only learn of it by interrogatory or deposition.

The rules further provide the process for dealing with this NRA designation. If the requesting party disagrees with the NRA claim, the matter goes to the court for resolution under the Proportionality Rule, wherein the court weighs the stated need against the claimed burden, relying on several factors to make the decision. These factors include: the specificity of the request, the availability of the ESI from other sources, the cost of the discovery vs. the amount in dispute, the resources of the parties, and their incentives to control costs, etc. Suffice it to say this is not an easy decision for a court to make. The FRCP do not state when the NRA claim must be made, but it is stated that the burden claim is to be defended whenever a motion is filed.

Another key new provision of the FRCP is the 26[f] rule requiring the parties to "Meet and Confer" prior to the Rule 16 Scheduling Conference with the court. The rules require the parties to discuss E-Discovery issues, hopefully to reach broad agreement thereon, which the court can use to develop a mutually-agreeable scheduling order. It therefore stands to reason that the best opportunity to present an NRA claim is at the "Meet and Confer". It follows that if the opposing party disagrees with the NRA claim this will become a topic at the Rule 16 conference, where both parties should be prepared to state their respective positions re: need and burden, thereby avoiding the time and expense of filing motions.

To be able to develop a reasonable quantitative ED burden estimate and be ready for the Meet and Confer, a lot needs to be accomplished, and time may be short.



When making an NRA claim "due to undue burden or cost" the party doing so is faced with a significant challenge. Unless there is a substantial business disruption burden claim available [e.g. shutting down I.T. systems, confiscating cell phones, etc.] the only remaining burden claim available is disproportionate cost. And the information needed to develop a credible cost estimate is not east to develop, since at this point in the discovery process there isn't much quantity information available. The factors involved in developing a cost estimate are many, varied, very difficult to subjectively estimate, and the bottom line is highly

sensitive to relatively small fluctuations in these factors, which include:

- Gross Volume of ESI
- Portions thereof which are known not to contain User-Created information [e.g. system and application files]
- Portion thereof within date range[s] of interest
- Portion thereof which are file types that may "reasonably" contain discoverable ESI
- Portions thereof which are not electronically searchable
- Portions thereof that require native production to be "reasonably usable"
- Number of files/GB of ESI
- Portion of ESI that will be responsive to electronic searches using search criteria that has yet to be established [i.e. "hit rate"]
- Plan for responsive and privilege review[s]-number and resources
- Number of files reviewed per hour [Review rate]
- Form[s] of production

While it may be possible to improve the estimates of some of these factors by cataloging and or sampling, it is always the case that most of these factors have a range of "reasonable" values. And with the bottom line highly sensitive to some of these factors, it follows that individual calculations can result in values that may differ widely. It would be possible to run a series of calculations using the "reasonable" ranges of the input factors, but the result will be a range of values that can be quite great.

This creates an unusual and awkward situation for the court, who is asked to judge the proportionality of the dispute. It can [and often currently does] result in experts, each relying on their "reasonable" assumptions, providing testimony to their estimates, which are vastly different. Judges, who often lack technical knowledge and experience with the nuts and bolts of E-Discovery, are left to judge the credibility of such testimony, or, as is increasingly the case, ordering the parties to a Special Master or Neutral Expert. This adds time and expense to both parties and further delays progression of the case to the merits, which is what the court [and presumably the parties] want to do. What is needed is a better method for estimating cost that will acknowledge that a range of values is possible, and further identifies the values within that range that are most likely to occur.

A BETTER METHOD – PROBABILITY ANALYSIS

This situation is not uncommon in estimating. For example, when bidding for a construction project, it's important for bidders to be able to determine how much variability there could be in the primary factors they use in estimating, e.g. labor and material costs, productivity, weather, disruptions, etc. This is also important for those who must appropriate funds for such activities, as their return on investment may be critical to the success of the project [and perhaps their continued employment]. In summary, it can be critically important to understand how much the final cost can vary, and what the most probable outcome is. Professional Estimators and Economists, aided by Statisticians, have developed method[s] to address these situations.

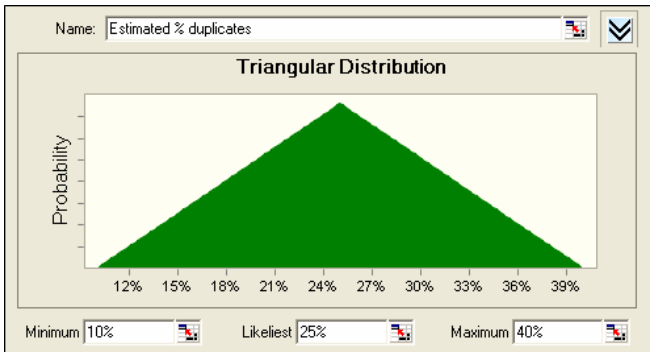
The method consists of two basic steps: [1] defining the range[s] of the input variables and the most likely values therein, and [2] performing a series of calculations encompassing these input ranges and arranging the results in a manner [graph] illustrating the range of outcomes and their respective probability of occurrence. The appeal of this method lies in its inherent defense against manipulation of the results. Although the results can be somewhat sensitive to any particular subjective input range, the output results won't be significantly changed without some heavy-handed gross skewing of multiple input criteria, which are relatively easy to identify and challenge. In a sense, it forces the estimating experts to be honest.

INPUT RANGES – PROBABILITY DISTRIBUTIONS

Let's look at how such input ranges might be developed. These are subjectively created, and therefore subject to challenge, so it's important to understand what they are. Let's take a simple example from the E-Discovery model: Duplicates. It is well established that it isn't necessary to produce duplicates of ESI, and it's also well known that a lot of them exist. ED Vendors routinely filter out duplicates by use of a special calculation algorithm resulting in a "Hash Value". This value is considered so unique that it is equivalent to a fingerprint. In other words, the likelihood of any two different files having the same hash value is infinitesimally small.

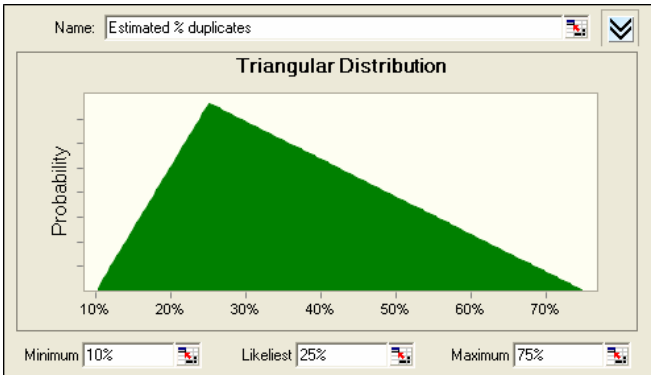
There should be some discussion about how the de-duplication process is to be applied: [i.e. across an individual "custodian" or across the entire ESI collection], and there also might be discussion on "near-duplicates", but since these are unlikely to have been discussed at this early time, the estimator should define his assumptions. In each case, there are three key data points to define: Minimum, Maximum, and Most Likely.

In our example, let's assume the Estimator believes, based on his experience that duplicates can range from 10% to 40% of the whole collection. Further, based on his experience the most likely is 25%. If we present these graphically with the % values on the "X" [horizontal] axis, and a probability of occurrence on the "Y; axis, and graph all the possible outcomes, without any additional subjective input, we get this:

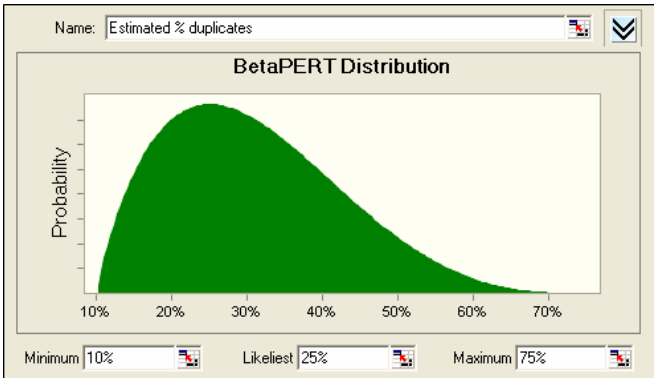


Called a "Triangular Distribution" for obvious reasons, this represents the simplest type of input distribution. As you can see, the greatest probability occurs at the "Most Likely" input point, but the probabilities are automatically established by the definition of the Min and Max values. So there's no possibility to skew the probability with this type of distribution.

But what if the Estimator's experience was that duplicates could range up to as high as 75%, although it's extremely unlikely that anything greater than, say, 40% ever occurs. If he used a triangular distribution, the distribution would look like this:



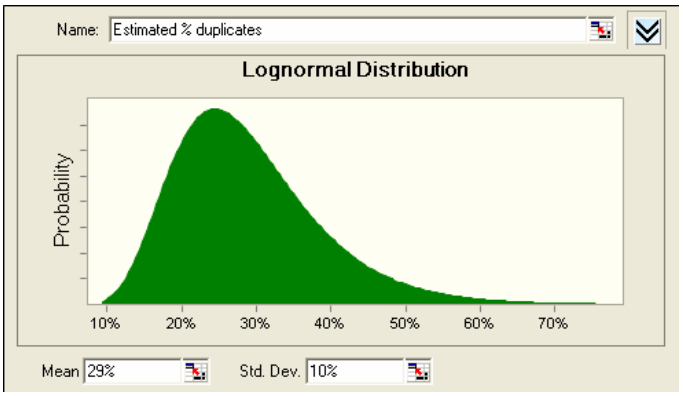
The problem here is that the probabilities to the right of the "Most Likely" are all too high until you get way out near to 75%. In this case the Estimator would need to use a different type of distribution, such as this:



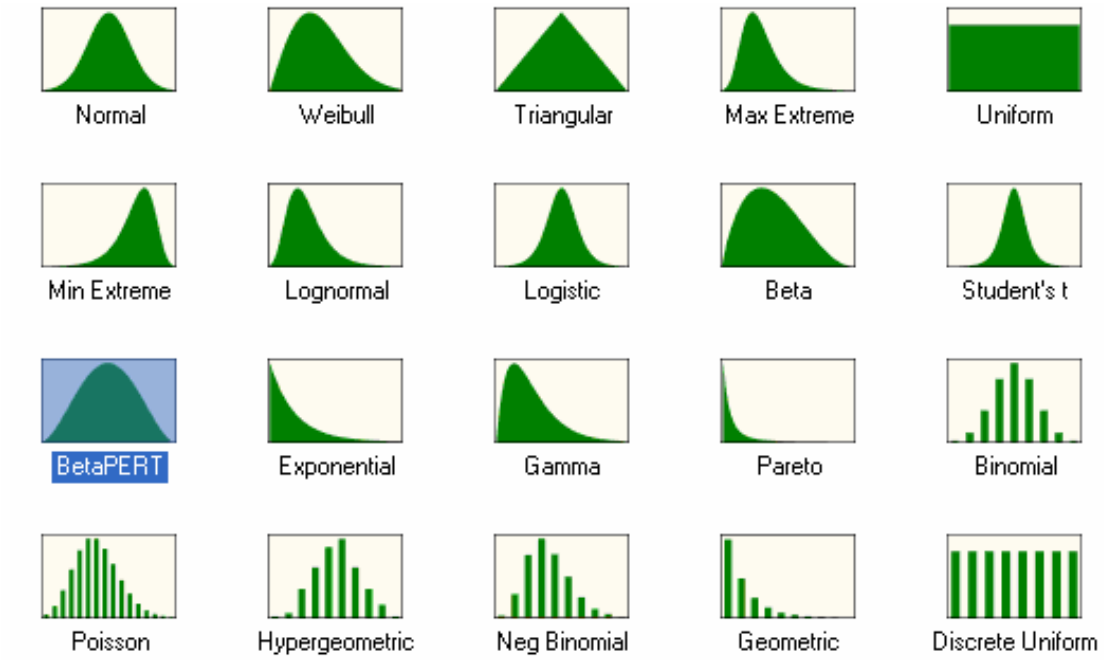
This type distribution [Beta-Pert] has somewhat limited the probabilities of the higher percentages.

But if the Estimator still feels that the outer probabilities are too high, he could select a different distribution such as:

Note that the input criteria here are necessarily different, but would be chosen by the Estimator to be consistent with his Min/Max/Most Likely choices. NOTE: It is also very important to note that any individual distribution can be "tightened" by the use of sampling].



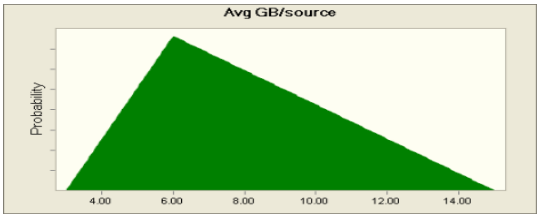
There are many, many types of probability distributions that can be employed in this method as shown below. The selection of distributions is a specialized skill, beyond the scope of this article.

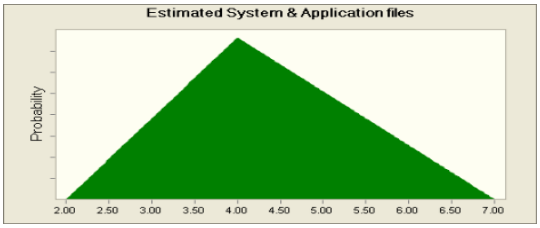


THE DISCOVERY BURDEN ESTIMATE

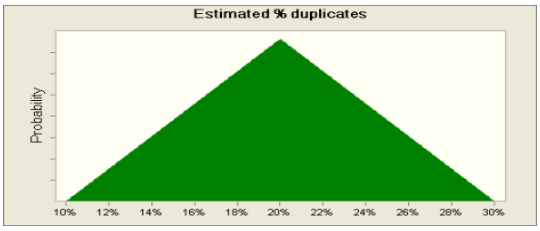
There are numerous factors involved in estimating the E Discovery Burden. The mathematical model to be employed will utilize these factors in the calculation, which must be adjusted to be consistent with the Discovery Plan. These factors would include those listed earlier and the calculation model would be set up to address any assumptions and/or agreements between the parties on such criteria as: data ranges of interest, file types to be included/excluded from discovery, definition of duplicates, specificity and/or number of search terms, etc. It is becoming increasingly common for parties to agree to begin discovery with a subset of the ESI collection [e.g. E-mails for Top Tier custodians] – often referred to as "Low-Hanging Fruit", with an agreement to discuss the scope of additional searches/productions after the review of the initial productions are made. [Note: It is important to ask the Court to accept this approach and allow for it in the Scheduling Order]. To be consistent with this approach, it is prudent to develop separate cost estimates for different subsets of the ESI collection to facilitate subsequent discussions.

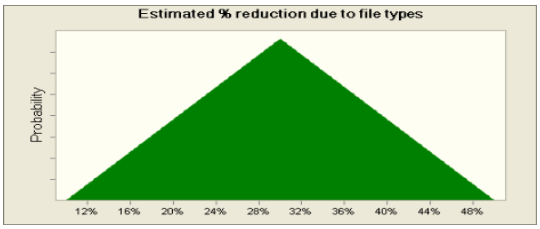
The calculation model for a typical Discovery effort might begin with development of input distributions and a calculation model like these examples:

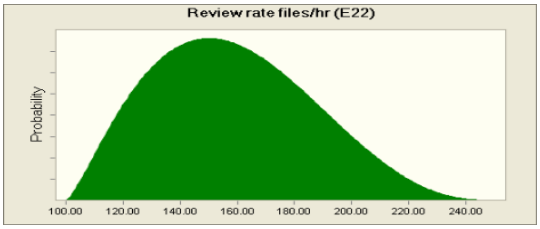


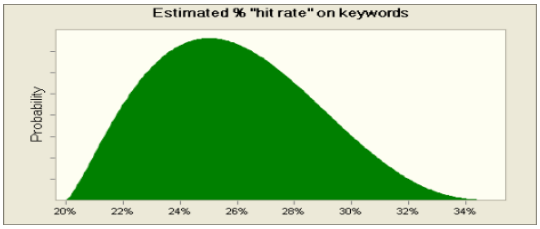


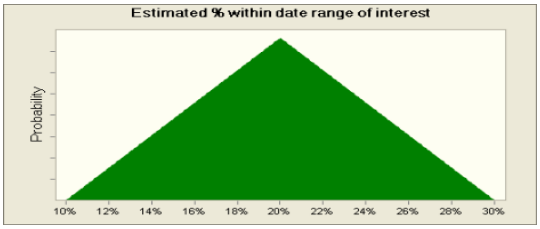
:

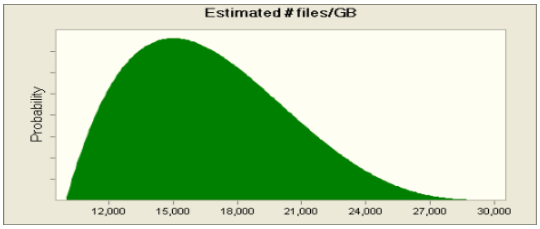


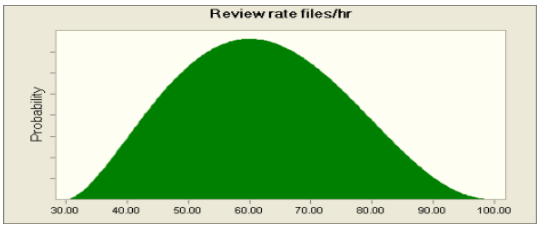


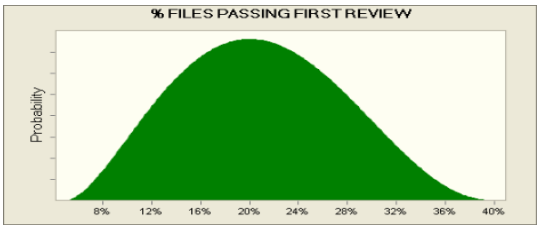


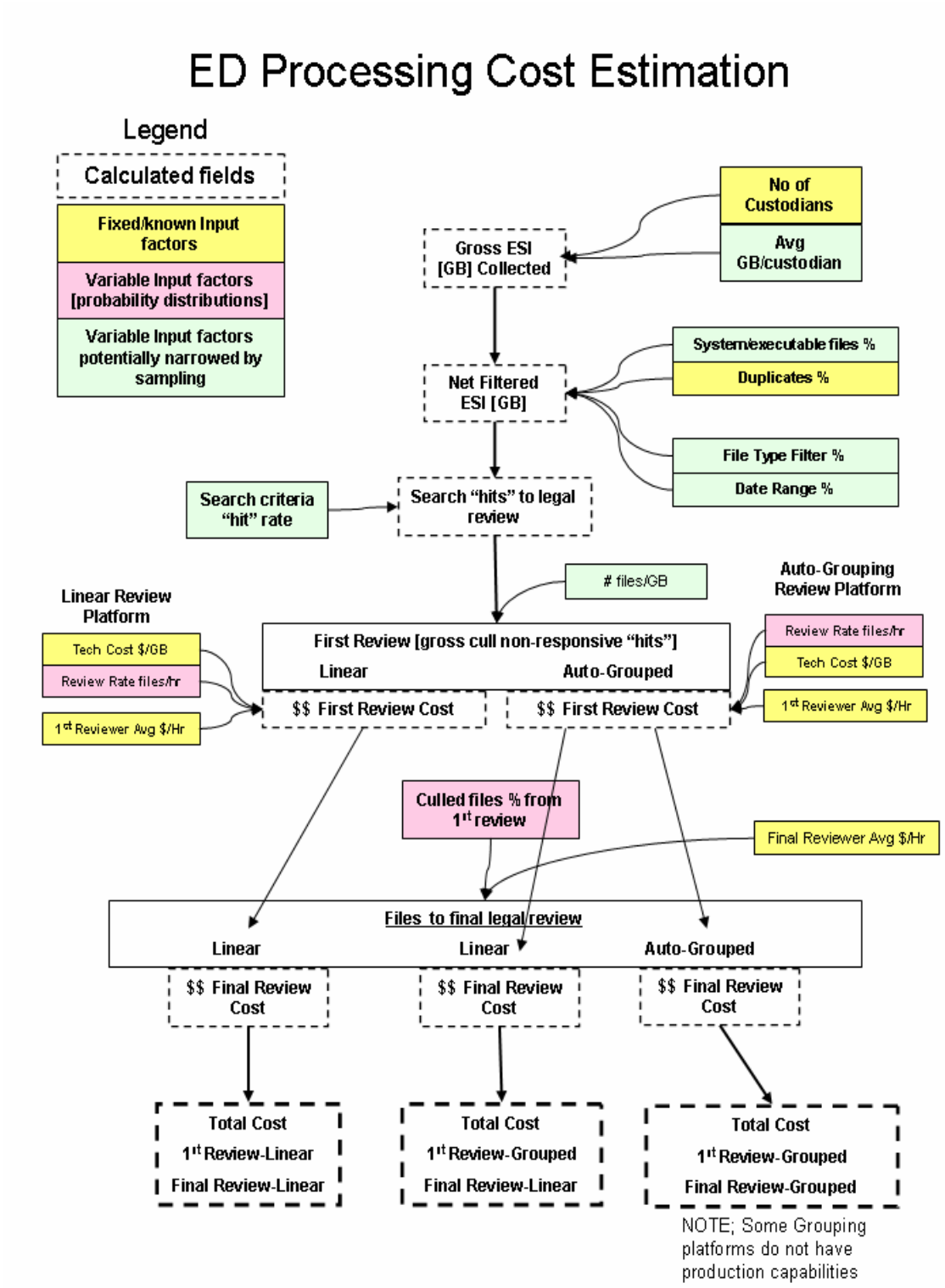








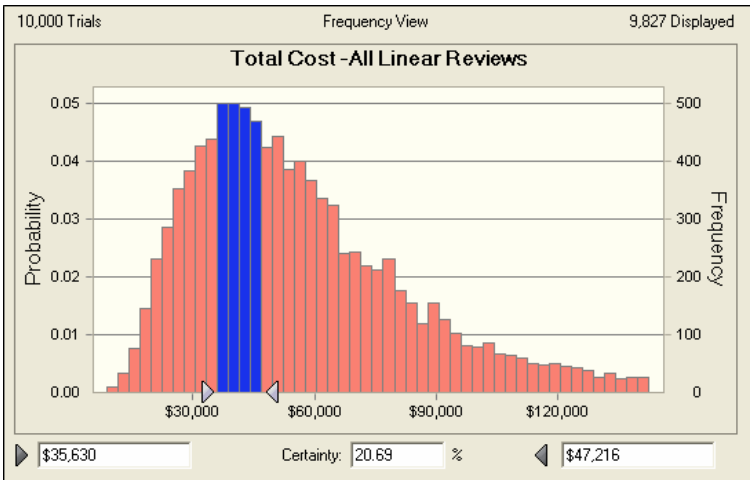




Now with the model fully prepared a simulation run can be performed, wherein the simulation will randomly select input values from each input distribution and calculate a result. This process is repeated thousands of times to assure that input values have been selected across each entire input distribution, thereby compiling a large collection of possible results. As many such calculations can be run as desired, but experience shows that more than a few thousand will not cause any significant change to the results.

Since these results are generated using randomly selected input values, the results are immune to intentional manipulation. With the great number of sample results, the collection can be assumed to be reliably representative of the entire spectrum of possible results.

The next step is to display the results in a meaningful manner that illustrates the relative probabilities of the values. To do this, the range of results is divided into small, discrete sub-ranges, and the number of results within each range is calculated and shown graphically, with each range illustrating the percentage of the whole, which represents the probability of the actual result falling within that range. For this example, the results would appear as follows: [only one of the results in the model is shown].



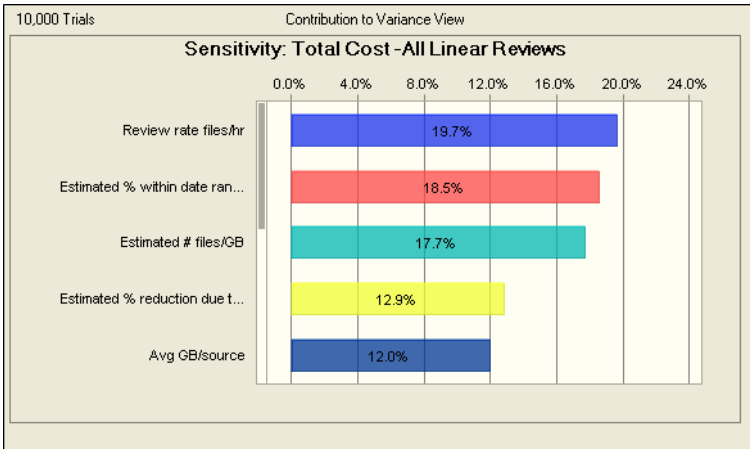
As it shows, the results range from less than $10,000 to almost $150,000, a wide range indeed. But more importantly, it is clearly shown that the results with the greatest probability of occurrence are in the $35,000 - $47,000 range, represented by the blue area [this area is subjectively selected by the Estimator].

This range encompasses just over 20% of the results values, but certainly illustrates that these have the highest probability. This range of “Most Likely Results” could be adjusted as the user[s] deem appropriate.

This type of analysis output is a great improvement over the single-value method of estimating, as it provides so much more information for decision-making by the parties, and, subsequently, should it be necessary, by the court.

THE BONUS – SENSITIVITY

This analytical method also produces a bonus: Sensitivity analysis. This is a chart showing which input factors are having the greatest effect on the results. For the example, this would appear as follows:



As shown, the factors contributing the most to the results are review rate [files/hr] , date range, and estimated no of files/GB. Note that with sampling these input factors could perhaps have their distributions “tightened”, thereby producing more definitive results [i.e. a narrower, taller, results distribution.]. With this information as a guide, any further actions to improve the estimate accuracy can be prioritized. In this case, for example, using a technology that improves review rates [e.g., an auto-grouping review platform] might be considered to reduce the bottom line.

SUMMARY AND CONCLUSIONS

Under the revised Federal Rules of Civil Procedure, the amount of ESI subject to discovery in litigation is greater than ever before, and it is growing as technology makes it so easy to retain ESI. This in turn creates the potential for staggeringly increased costs unless means to limit the ESI in discovery are employed. Fortunately, the developers of the new FRCP recognized this problem and included provisions to allow parties to argue to exclude portions of their ESI from discovery by designating it a “Not Reasonably Accessible due to undue burden or cost”. The FRCP also provides a process to allow requesting parties who disagree with such NRA designation[s] by producing parties to dispute the NRA claim and send it to the court to rule under the Proportionality Rule. In making such arguments, the means to quantitatively estimate the cost of discovery are essentially important. The factors involved in making such estimates are many and varied, and the resulting cost estimates can be extremely sensitive to relatively minor fluctuations in the input factors. This in turn results in widely varying cost estimates thereby creating a difficult situation for Courts who are required to make ruling on these disputes.

This newly developed, proprietary methodology for generating cost estimates using probability analysis produces outputs which enable the parties and the Court to make more informed decisions, which will facilitate the discovery process, enabling the matter to proceed to arguments on the merits, and save the litigants [and the court] time and expense.

About the Author

Jim Wright is a Director in the Technology Segment of FTI Consulting in Houston, Texas. His practice focuses on in-house electronic discovery, including litigation readiness, expert analysis and testimony. Prior to joining FTI, he was the Director of E-Discovery for Halliburton’s Litigation Group, where he gained broad experience in E-Discovery matters for complex multi-national litigation & investigations. In 2006, He was a co-founder of the Corporate E-Discovery Forum, a non-profit organization currently with over 350 members from over 165 corporations working collectively to develop best practices for in-house litigation and records management programs. His background is in Engineering and Construction, including Expert Testimony in Construction Disputes, and areas of expertise include Capital Cost Estimating, and Critical Path Scheduling and analysis.