WHITE PAPER

Metrics that Matter

What do Van Halen, M&Ms and metrics in e-discovery have to do with one-another? More than you might think.

Gareth Evans, Partner, Gibson, Dunn & Crutcher

David Grant, Senior Managing Director, FTI Consulting

GIBSON DUNN

FTI CONSULTING™ | TECHNOLOGY

What do Van Halen, M&Ms and metrics in e-discovery have to do with one another? More than you might think.

In Van Halen's heyday, the band toured with an elaborate stage show that included heavy sound and light equipment, and harnesses so that David Lee Roth could fly over the audience.

In each city, the band depended upon different local companies to unload and set up the equipment. High-quality work was paramount. If the local production company overlooked something, it could mean serious injury for band members or the audience.

How could the band ensure their instructions were followed? The band devised a strategy that would enable them to quickly assess whether the local company was likely following instructions.

Van Halen's 53-page tour rider, required a bowl of M&Ms backstage with all of the brown M&Ms had to be removed. Failure to comply would result in the concert promoter's forfeiture of the entire show at full pay.

If Van Halen arrived on the day of the concert and found brown M&Ms backstage, they knew that their instructions weren't followed closely. So, they would conduct a thorough quality review of every step of the stage setup. If the brown M&Ms were removed, however, the band had greater comfort that their instructions had been followed, and they could focus on testing a few key pieces of equipment.

Just as the colored M&Ms provided a quick visual metric of quality, we can measure the effectiveness of search and review. It's possible both to have effective statistical arguments for minimizing the burden of discovery (for instance by proving the wasted effort caused by a too broad keyword/s) and to understand whether you are on the right track in document review.

Doing so is increasingly important as large document volumes are changing the approaches litigants must take if they wish to avoid being overwhelmed with e-discovery burden and expense. In short, using the right metrics can help ensure defensibility and cost-effectiveness.

Key Metrics for Search and Review

Key metrics for use in search and review include:

- Prevalence (also called "Richness")
- Recall
- Precision
- Depth for Recall
- Confidence Level
- Confidence Interval (*i.e.*, margin of error)

These metrics all involve statistical sampling, which allows one to use a representative random sample to affordably draw reliable conclusions about the overall document population without the time and expense of reviewing the entire document set.

The size of the sample needed is not directly proportional to the population size, and so generally requires review of a relatively small number of documents.[1]

We refer to "responsive" or "relevant" documents in the examples below, but the same metrics can be used with respect to other aspects of a document population, such as attorney-client privileged documents, key documents, foreign language documents and so on.

[1] The population from which the sample is drawn should exclude the documents used to develop the search methodology – such as documents used to train a predictive coding model. Otherwise the sample may not properly represent how the process would work against documents not seen before by the process.

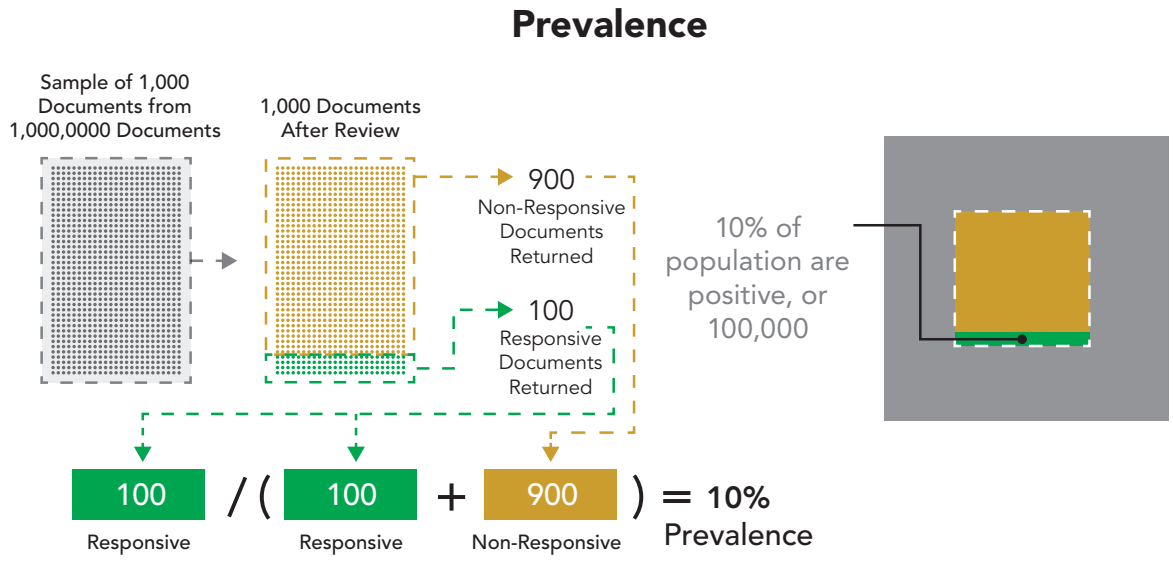
Prevalence / Richness – *"How many are there?"*

Imagine knowing in advance roughly how many responsive documents there will be, so that you can ensure the right resources are in place. "Prevalence" will tell us.

Prevalence measures the proportion of the population that has a particular characteristic based on the proportion found in the sample. If 10% of the sample is responsive, then we can project that 10% of the population from which the sample is drawn will be responsive (within a certain margin of error).

Or, to go back to the M&Ms example, if a statistical sample of M&Ms produces 10% brown M&Ms, we can project that 10% of the total M&Ms are brown.

Knowing Prevalence or Richness not only enables planning for the right resources and employing the best search and review strategy, but it also can help you establish your "goal" for retrieval processes (for example, how many responsive documents exist that your keywords, technology assisted review process or reviewers are trying to find).



Recall – *"How many of (the Brown M&Ms) am I finding?"*

Imagine knowing with a high level of confidence how well your review process is working to find responsive documents. In other words, what proportion of the total number of brown M&Ms are we finding? "Recall" will tell us.

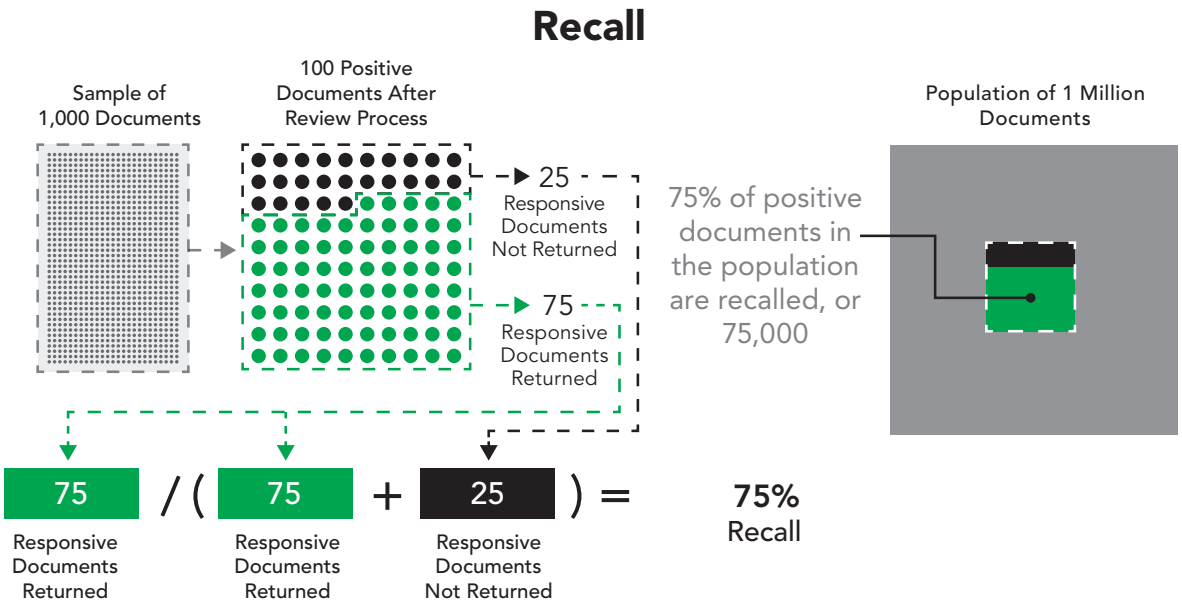
Recall is the percentage of the total responsive documents in a document population that a search or review process actually finds. It is probably the most important search and review metric. To get Recall, divide the number of responsive documents in a sample that a process finds by the actual total number of responsive documents in the sample.

For example, if you have a very large bag of M&Ms with 100 of each color, and you are searching for all of the brown M&Ms, and you find 80 of the M&Ms, that would mean your Recall on brown M&Ms is 80%.

Recall permits one to reliably evaluate the effectiveness of proposed search terms—*i.e.*, how well they perform in actually finding relevant and responsive documents. This information can be critical in negotiating search terms or ensuring those that you are using are defensible.

Recall is the key metric used to target and evaluate the performance of predictive coding protocols. Predictive coding protocols often target a particular Recall rate, such as 80%. A validation process involving the review of a statistical sample can confirm that the predictive coding process achieved the targeted Recall.

Recall can evaluate the effectiveness of a combined search and review process, such as using search terms to initially cull the document set and then applying predictive coding to the search term "hits." Recall can also assess reviewer effectiveness.



Precision – *"How much effort am I wasting to find (the brown M&Ms)?"*

Imagine knowing how efficient your search and review process is currently and quantifying the burden of any wasted effort involved. What percentage of the documents it finds are "false positives." In other words, when you grab some M&Ms from the bowl, what portion are not brown M&Ms?

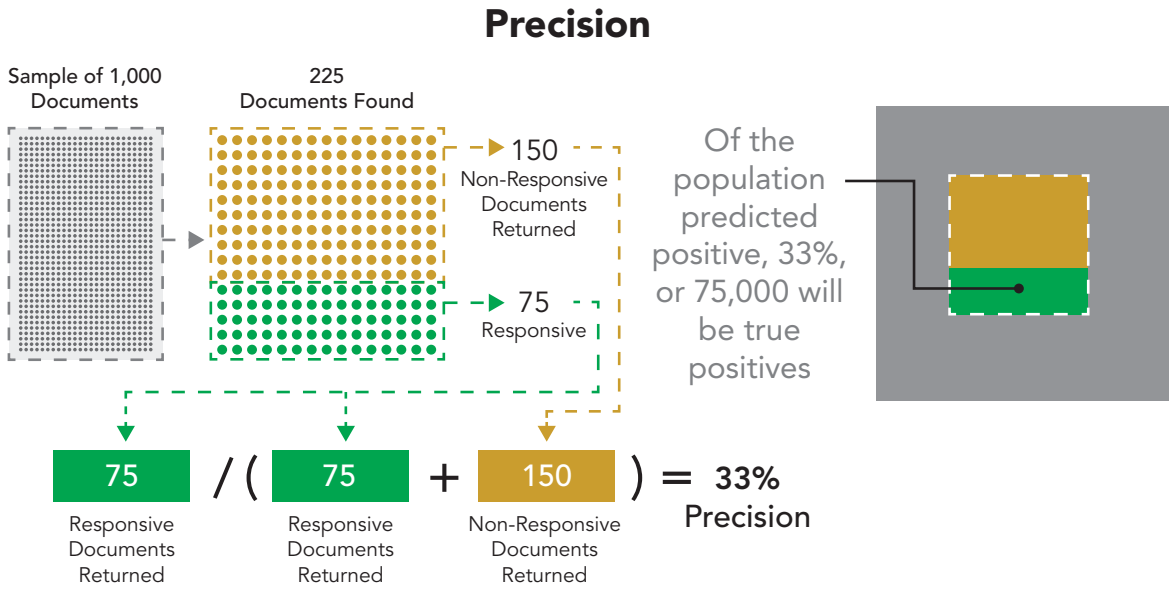
Precision measures how well a search or review process yields only responsive documents. It is determined using statistical sampling methods similar to those used for Recall. The total number of responsive documents identified by the search/retrieval process is divided by the total number of documents that the search/retrieval process retrieved. Many will be false positives, *i.e.*, "negatives" that the process suggested would be "positives." For example, 5% Precision means 19 irrelevant or nonresponsive documents are retrieved for every 1 relevant or responsive document retrieved

Having a search process that yields high levels of Precision is key to avoiding the costly inefficiencies of reviewing large numbers of irrelevant documents.

Being able to measure Precision can be critical for a responding party in negotiating an acceptable set of search terms or a reasonable Recall level for predictive coding. Achieving higher levels of Recall is often done at the expense of lower Precision. Measuring Precision can also be important in persuading the court that the opposing party's proposed search terms or Recall level are overly broad.

In M&M terms, let's say you have a bag of hundreds of assorted M&Ms and want to find the 100 brown M&Ms. If you were to pour out some of the M&Ms in the bag and get a total of 130 M&Ms, including all 100 brown M&Ms, that means that you have an additional 30 M&Ms beyond your goal of 100 brown M&Ms—*i.e.*, 30 "false positive" hits. That would be a Precision rate of 77% (*i.e.*, 100 brown M&Ms divided by the 130 total M&Ms poured out of the bag).

The goal of search is to have best of both worlds with high Recall and high Precision. A carefully tested and calibrated set of search terms and an effective predictive coding process may achieve that goal. The two, however, are often inversely related—*i.e.*, achieving higher Recall may mean lower Precision and vice versa. Having both Recall and Precision statistics can be very useful in negotiations seeking proportionality. Precision, fore example, allows you to quantify the increased burden associated with higher levels of Recall.



Depth for Recall – *"How much work is needed to find (the brown M&Ms)?"*

How many documents will you need to review using a particular search process? Depth for Recall measures the proportion of the document population that you must review using a particular search process to achieve a given Recall level. In other words, using our search process, how many M&Ms do I need to look at to confirm that I have found the brown M&Ms at my targeted level of Recall? What trade-offs are involved between different levels of Recall?

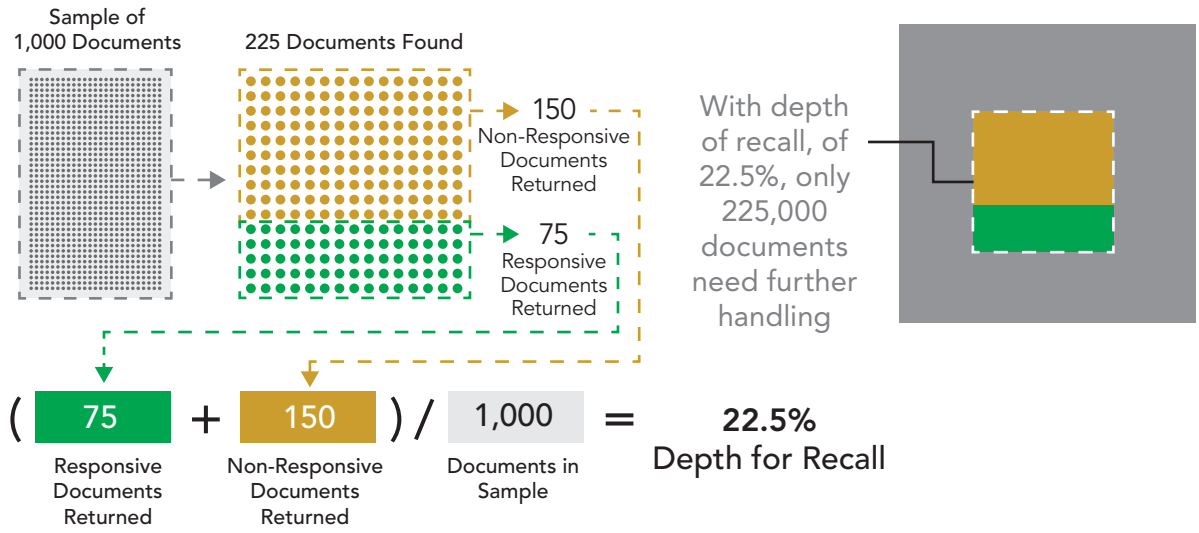
One way to obtain the Depth for Recall figure is to multiply the Prevalence of responsive documents by the targeted Recall, then divide by the Precision of the search process at that Recall level.

Let's say we have a population of 100,000 documents. If only 1% (or 1,000 documents) is projected to be responsive (*i.e.*, 1% Prevalence), then a target of 75% Recall using our search process means that we have to find 750 responsive documents to achieve the target. If the Precision of our search process is only 5%, how many documents will we need to review to find 750 responsive documents? Even though our search process yields only 5% Precision, the Depth for Recall metric (Prevalence x Recall divided by Precision) tells us that, using our search process, we'll only need to review 15% (*i.e.*, 1% x 75% divided by 5%) of the population (*i.e.*, 15,000 documents) to find 750 responsive documents (*i.e.*, 75% Recall).

Depth for Recall is important in two different ways. First, Precision alone doesn't actually indicate the effort required, because in cases of very low Prevalence even a low Precision can lead to a small proportion of the population requiring review. The 15% Depth of Recall number tells us that even though we have a very low Precision figure, our process culls out 85% of the irrelevant documents, leaving only 15% remaining for review.

Depth for Recall enables comparison between various achievable Recall levels using predictive coding (or other search methods). It allows the additional effort required to achieve higher Recall levels to be quantified, along with any benefits gained in terms of the number of additional responsive documents found. In other words, how many more documents will we need to review to achieve higher Recall levels? Depth for Recall thus enables analysis of the proportionality of the different Recall levels.

Depth for Recall



Confidence Level and Confidence Interval (Margin of Error)

Imagine knowing how accurate your key metrics are. The measurements are not 100% accurate when reviewing only a sample of the population. Perfection could be theoretically accomplished by reviewing an entire document population with perfect reviewers. This is neither humanly possible nor required by the Federal Rules of Civil Procedure. With respect to Prevalence, or Precision, for example, a 2,000 document sample may provide 99% confidence (the Confidence Level) with approximately a +/- 3% margin of error (the Confidence Interval), or better. (As outlined below, for Recall the measurements depend on the number of positive documents in the sample).

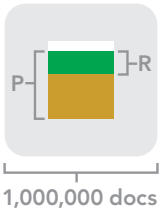
The accuracy of Prevalence, Recall and Precision measurements are related to the size of the sample used to generate the metric. So, for prevalence, if a 2,000 document sample shows that 50% of the documents are responsive, we could say we are 99% confident that between 47% and 53% of the documents in the population are responsive.

A few things to keep in mind about Confidence Levels and Confidence Intervals: A common mistake is to assume that the sample size for a given Confidence Level and Confidence Interval is proportional to the population size. In reality, the sample size required for a given Confidence Level and Confidence Interval is not significantly affected by the population size. Sampling can be relatively inexpensive even for very large document populations. While many might be reluctant to introduce sampling into their search and review process because of concerns about additional burden, in most cases the burden may be relatively small.

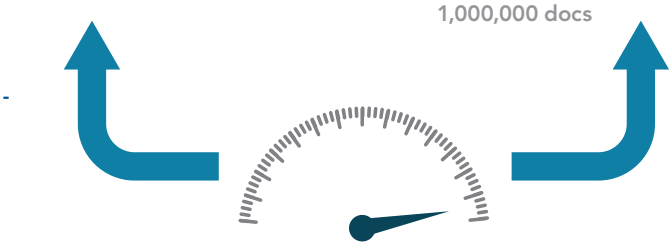
The Confidence Interval (or margin of error) is a crucial part of your measurement. For example, if the Confidence Interval for the 15% Depth for Recall example above is +/- 5%, that means you will actually need to review anywhere between 10% to 20% of the document population to achieve the targeted Recall level (in the example of a population of 100,000 documents, that means a range of between 10,000 to 20,000 documents requiring review to achieve the targeted Recall).

The relevant sample size required to achieve a given Confidence Level and Confidence Interval is tied to what you are measuring. When measuring Recall, and seeking a particular Confidence Level and Interval, the necessary sample size will be impacted by the Prevalence of responsive documents in the sample. This can sometimes mean the sample must be larger to find enough responsive documents in the sample to match your goals for the Confidence Level and Confidence Interval of your Recall measurements.

Simple distribution measurements measure the percentage of documents which have a quality



Recall/precision measurements measure the percentage of documents with a quality which are retrieved by some process for finding them



Confidence levels and margin of errors tell us how accurate our measurements of distribution and recall/precision are

To confirm a Recall level of 50% with a Confidence Level of 95% and a Confidence Interval/margin of error of +/- 5%, you would need 385 responsive documents in your sample. If the responsive documents are 10% of the population, you would need a sample of 3,850 documents to get 385 responsive documents, but if the responsive documents are 33% of the population you would need a sample of 1,155 documents to get 385 responsive documents. In other words, a higher Prevalence rate will result in needing a smaller sample size than if you had a lower Prevalence rate.

Don't forget that Prevalence, Recall and Precision measure actual performance, while the Confidence Level and Confidence Interval are simply measuring the accuracy of those metrics.

What do brown M&Ms have to do with it again?

Metrics generated from review of small document samples can make a world of difference in helping you to develop a defensible and cost-effective document review process.

Having counsel and service providers involved in your document search and review process who are familiar with and can accurately generate such metrics can yield significant cost savings and verify the efficacy of the process. After all, Van Halen used a convenient metric to help avoid a disaster befalling band members and the audience in their elaborate stage show. You can do so, too, in your document search and review process.

Gareth Evans is a Partner at Gibson, Dunn & Crutcher LLP.

David Grant is a Senior Managing Director at FTI Consulting.

Gibson Dunn and FTI Technology help clients manage the risk and complexity of e-discovery. For more information, please visit: **www.gibsondunn.com** and **www.ftitechnology.com**.

GIBSON DUNN

FTI CONSULTING | TECHNOLOGY