FOMO and eDiscovery: A Receiving Party's Perspective and Response

David R. Buchanan[1]
Douglas E. Forrest
William Webber

> *"Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know."* – Donald Rumsfeld[2]

FOMO and eDiscovery ("*FOMO*") attempts to quell concern about the "known unknowns" in an e-discovery process. In *FOMO*'s example process, the "known unknowns" comprise the estimated 20% of relevant documents left in the null set—the (generally large) set of collected documents that are not returned as relevant through a search or review process. In modern e-discovery TAR processes, the null set is probabilistically suggested by computerized text classification engines. *FOMO* asserts that the "short answer" to the question of whether any notable facts are missed is "no." In fact, when *FOMO*'s simplifying assumptions and model preferences are more reasonably characterized (and certain seeming calculation errors remedied), the short answer for most moderately complex cases appears to be the reverse.[3]

Dr. Roitblat correctly points out that missing 20% of potentially relevant documents does not necessarily mean missing 20% of relevant information, due to the potential for redundant information between different documents. Though such an observation is intuitively obvious, and a model could be useful in guiding and quantifying our intuition, we must remain alert to the model's simplifications and limitations before attempting to extend its results to the real world.

As a threshold matter, the simplification of the model from documents to "facts" is definitionally problematic. Facts—as discerned from the *words* in documents (which is what predictive coding classification engines *actually* evaluate)—have many shades. The same or similar collection of words generate different facts and inferences in different document contexts. Identical words in a document years before a product recall can lead to far different inferences when appearing in documents shortly after it. So too when those words appear in the files of particular decision makers or departments. Any trial lawyer who has spent hours sifting similar documents for the one sent or received by the key decision maker, in the right temporal and broader context, can attest that sameness of words in no sense means sameness of fact or inference.[4] But *FOMO*'s model assumes far more broadly.

---

[1] David Buchanan is a trial lawyer and litigator with Seeger Weiss LLP; Douglas E. Forrest is the Vice President, eDiscovery Analytics, for ILS; William Webber, PhD, is an independent statistical consultant.

[2] Judge Johnston's notable e-discovery decision in *City of Rockford v. Mallinckrodt ARD, Inc.,* Civ. No. 3:17-cv-50107 (N.D. Ill. Aug. 7, 2018), begins its sampling and validation discussion by harkening back to Secretary Rumsfeld's memorable 2002 quote (https://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636).

[3] Though we take exception to *FOMO*'s conclusion that there is little practical risk of "missing out" on important facts by foregoing review, let alone analysis, of the null set, those criticisms do nothing to diminish the value of TAR itself. A well-constructed TAR e-discovery process, with sound validation and discretionary supplementation, is a great leap forward over the attempts to intuit responsiveness from a document corpus, with no validation, through rote Boolean keyword searches.

[4] *FOMO*'s adoption of a new analytical framework—"facts" vs. documents—counsels caution until there has been far greater consideration. As Thoreau analogously put it, "Beware of all enterprises that require new clothes."

The keystone assumption—which underlies *FOMO*'s "Approach 1" (a model that assumes an unlimited number of "facts") and *FOMO*'s repeated invocation of a 2.6% probability—is that the appropriate frame of analysis is whether some single "fact" could have been missed.[5] But why should the analysis be restricted to only a single missed fact? What happens to the analysis when there is more than one "fact" that could be missed (a likelihood, given a more appropriate appreciation of facts, not addressed by Approach 1)? For example, if there are 100 possibly missed "facts," each with a 2.6% chance of being missed, then the probability that at least one would be missed in the reviewed set and found in "known unknowns" would be 92.8% (that is, $1 - (1 - 0.026)^{100}$).

FOMO's "Approach 2" assumes a known number of "consequential" facts, sets that number "somewhat arbitrarily" at 300, and recognizes that some will appear more frequently than others. Working with the assumed 300 facts (a light estimate given that separate contexts spawn separate facts), with a power law distribution having exponent $\beta = 1.07$, the most frequent fact will occur in around 19% of relevant documents (almost 1 in 5), while the least frequent 44 will each occur in less than 0.05% (under 1 in 2,000).[6] Nevertheless, *provided that those documents are a simple random sample of the relevant set,* we could expect every fact to be represented in a set of between 6,959[7] and 17,097 relevant documents.

In reality, the relevant documents identified using predictive coding are far from a simple random sample of the relevant set. Indeed, the only way to draw a simple random sample containing 80% of the relevant documents would be to draw a sample containing 80% of the full collection, and review all documents drawn—no great gain in efficiency! Modern TAR processes enable users to sidestep review of all collected documents by focusing on likely relevant documents. This is done by building a statistical model of what a relevant document looks like, based upon the relevant documents that have already been located. In one common approach, known as continuous active learning (CAL), the most-likely relevant documents are reviewed at each iteration, and these reviewed documents are fed back into the predictive model.

Predictive coding boosts efficiency by concentrating on documents that look like relevant documents already seen. This also implies that if there are multiple relevant "facts", the predictive coding model will tend to be biased towards recommending documents containing already seen facts. This bias will tend to favor more frequent over less frequent facts, increasing the likelihood that one or more of the rarer facts will be missed from the identified set.

We can illustrate the effect of such a bias by stochastic simulation. Under Dr. Roitblat's assumption of 300 topics, with exponent $\beta = 1.07$, and simple random sampling of relevant documents, simulation indicates that the median case requires around 10,500 relevant documents to provide an example of each fact. To reflect bias towards already seen facts, let's assume a CAL environment where the model is updated every time a relevant document is located. Modify Equation 3 as follows:

$$w_i = \frac{1.0}{i^{\beta}} + \frac{n_i}{100}$$

---

5 Nor is 2.6% the maximum probability of missing a single fact. A fact appearing in 0.0001 of documents would have a 63% chance of being absent from the identified set, and a 22% chance of being present in the missed set, for a combined "miss" probability of 14%.

6 Dr. Roitblat's Equation 3 provides weights, not direct probabilities; to derive probabilities, we must divide by the sum of all the weights.

7 We believe that Dr. Roitblat's figure of 2,394 for the lower bound is a calculation error, and the correct lower bound from Equation 4 is 6,959.

where n_i is the number of documents containing fact i that have been found to date, and $n_i/100$ represents the bias toward already seen facts (arbitrarily chosen, for the sake of illustration). The more documents that are found containing a given fact, the more likely that the next relevant document to be selected for review will also contain that fact. In this biased model, simulation indicates that the median number of facts that are not represented in a set of 10,500 relevant documents is 70, or almost a quarter of the total.

FOMO submits that there is little to be missed among the "known unknowns." However, as shown herein, modelling that pushes in the direction of reality guides otherwise. Hands-on experience with large litigation document collections as a receiving party has similarly shown that the cautions underlying e-discovery FOMO are rational. As a receiving party employing predictive coding around production set "issue builds," the known unknowns are never ignored. Rather, lower ranked/low probability documents are further "challenged" by other search techniques, including keyword searches, concept and similarity searches, and focused linear searches (key people, dates, document types, etc.) to supplement predictive coding. Such supplemental efforts virtually always yield fruit.[8] Indeed, such multi-modal techniques are now advocated by attorneys and prominent e-discovery vendors in comprehensive and iterative TAR production workflows, with the promise of superior recall and precision over prior TAR processes.[9]

While models that can help us understand and characterize what we are missing may prove useful, there should be no debate that continued development of tools and workflows that can efficiently get to the novel documents or facts—and further validate our success at doing so—must always be the first priority.

---

[8] A stroll through the UCSF litigation archive reveals many documents—with telltale signs that humans recognized their litigation significance (trial exhibit numbers)—that would likely find themselves in the "known unknowns" bucket today. See, for example, the email exchange between Merck's President of science (and member of the Board of Directors) and its commercial President, rebutting Merck's trial defense that it worked cooperatively with FDA to warn about Vioxx's heart risks years before it was withdrawn: "Be assured we will not accept this label [proposed by FDA]"; "We knew it would be UGLY and it is. We'll fight back…."; "it is ugly cubed. thye [sic] are bastards." (https://www.industrydocuments.ucsf.edu/docs/#id=fphw0217, last accessed April 26, 2019). That undoubtedly "lower relevance" document—as discerned by text classification—had reach beyond the dozen or more trials in which it was used; it was featured in Congressional hearings that yielded legislation providing FDA with new powers to help overcome drug company obstinance.

[9] KLDiscovery, *The Ultimate Predictive Coding Handbook*, at 12 (2018) ("Hybrid Multimodal Review … yields the highest quality of output of any method (as defined by Recall and Precision) on most cases"), last accessed April 26, 2019, https://assets.krollontrack.com/hv4/pdf/BRO_KLD_US_Mastering_Predictive_Coding_Handbook_Jan2018.pdf.