



WHITE PAPER

# Back to Basics - What is Concept- Based Analytics and Why Should I Care?

By Trevor J. Morgan, PhD,  
VP and General Manager, FastLine Technologies,  
Inc., and

**Rich Turner**  
Vice President of Marketing at Content Analyst  
Company.

Concept-based advanced analytics isn't anything new, and many of the technologies incorporating conceptual analytics are very mature. Numerous case studies illustrate how these technologies help manage costs, trim schedules, and increase quality throughout the eDiscovery process. Justices, too, are pressing attorneys to use advanced techniques to avoid keyword pitfalls and to accelerate the discovery process of ever-larger electronic document sets. Newer software offerings such as predictive coding and automated first-pass review depend heavily on conceptual analytics. Seemingly all the recent innovations in eDiscovery stem from conceptual analytics capabilities.

One would expect that advanced analytics would be widely embraced as a requisite part of most eDiscovery workflows, but that isn't always the case—for a number of reasons. We think, therefore, it's time to go "back to basics" and look at these reasons—and why you need to think again about concept-based analytics.

### **I don't understand what concept analytics is.**

All search engines "index" the set of searchable documents to discover their contents. Keyword-based engines create a lookup list of all known words in the entire document set. When a user submits a query, the keyword engine returns documents based on word matches between the query and the searchable documents.

Concept-based analytics engines take this indexing process one step further, by determining the conceptual content of the searchable documents, not just the individual words found in them. The analytics engine returns a document not because of any matched words with the query but rather because the query and the document share the same (or highly similar) concept.

### **Does this replace keyword search?**

Absolutely not! Keyword-based technologies (including fielded searches and metadata searches) are very useful, and most eDiscovery efforts start off with a list of key terms. Looking for concepts is simply the next logical step: if you find documents using keywords that describe fraud, concept-based analytics makes the process of finding other documents that also describe

fraud very straightforward—as easy as "find me everything else like this."

### **How defensible is concept-based analytics?**

Every credible concept-based analytics engine incorporates well-defined algorithms to carry out conceptual indexing of documents, and there are only a few different technical approaches to accomplish this. Semantic-based engines "deconstruct" sentences into words and use dictionaries to "look-up" their meaning. Mathematics-based engines, which arguably are more sophisticated and powerful, use either probabilistic algorithms (typically Bayesian) or deterministic algorithms (typically LSI) that compare words and documents on a mathematical basis. The algorithms for these techniques are well-documented, and all three techniques (semantic, probabilistic, and deterministic) are going to yield results that are repeatable and are based on conceptual relevance rather than shared keywords.

### **How is concept-based analytics used?**

Concept-based analytics drives many powerful capabilities; two of the most popular are organizing documents based on conceptual content, and concept search (finding documents based on conceptual content).

Document organization can be automatic or user-driven. Clustering is the automatic version: the engine groups all documents together which share a predominant concept or theme without user guidance. Clustering is a very "non-intrusive" technology – it's the familiar review paradigm, only in this case the documents are all conceptually similar, making review much faster.



User-driven categorization, on the other hand, starts with human input. The user designates categories and then “defines” those categories by identifying example documents. The engine uses these examples to “teach itself” what belongs in each category. It’s a more targeted approach than clustering, and it is the basic technology behind all “predictive coding” solutions.

Concept searching – the other most popular capability of concept-based analytics - is simply a more natural way to find relevant documents. Rather than trying to devise a group of keywords that will return a relevant result, users can submit phrases, paragraphs, or even whole documents as their “query.” The engine figures out the concepts in the query, and then finds documents with similar concepts. It is very similar to human thought processes.

**I’m still not convinced: analytics is expensive, and I don’t see how to justify it.**

Time is money in any business, and clients are more cost-conscious than ever. There are only two real choices when cost becomes a factor: you either participate in a “race to the bottom” with ever-decreasing prices, or you figure out how to become more efficient.

Concept-based analytics provides tremendous efficiency. The cost of using analytics, whether it’s licensing an analytics module for current software or deploying new analytics-driven products, is magnitudes less than applying more “human-power” toward any aspect of eDiscovery. Cost isn’t the only benefit that analytics provides. Even with unlimited funding, time constraints put pressure on cases that only automation can address. Worse, the mind-numbing task of reviewing hundreds of thousands of documents virtually guarantees mistakes and flawed judgments. Concept-based analytics helps navigate around such mistakes and avoid the huge financial penalties that can accompany them.

Most concept-based analytics providers have a wealth of tools to help illustrate and justify costs, including “calculators” or spreadsheets that look at the total cost of review. Specific solutions like assisted document coding systems have even more obvious returns on investment.

**Conclusion.**

When you look at concept-based analytics, the real question should not be one of “I don’t understand,” or “I’m not comfortable...” but rather how can you justify not deploying analytics. Courts rarely tolerate eDiscovery failures due to flawed keyword searches, and cases like *Spieker v Quest* remind litigants that lack of technological prowess isn’t tolerated either. When pressed on whether advanced techniques like predictive coding are “defensible,” judges have responded that attorneys need to be mindful that their eDiscovery plans are what need to be defensible. FRE-502 goes further: the use of “advanced techniques” can be a defense against waiver of privilege.

Concept-based analytics leverage proven technology. They address the very real issues of cost control and time constraints. They help avoid expensive inadvertent spoliation and help protect against inadvertent production. In the highly-competitive yet rigorous world of litigation, concept-based analytics makes a real difference.

