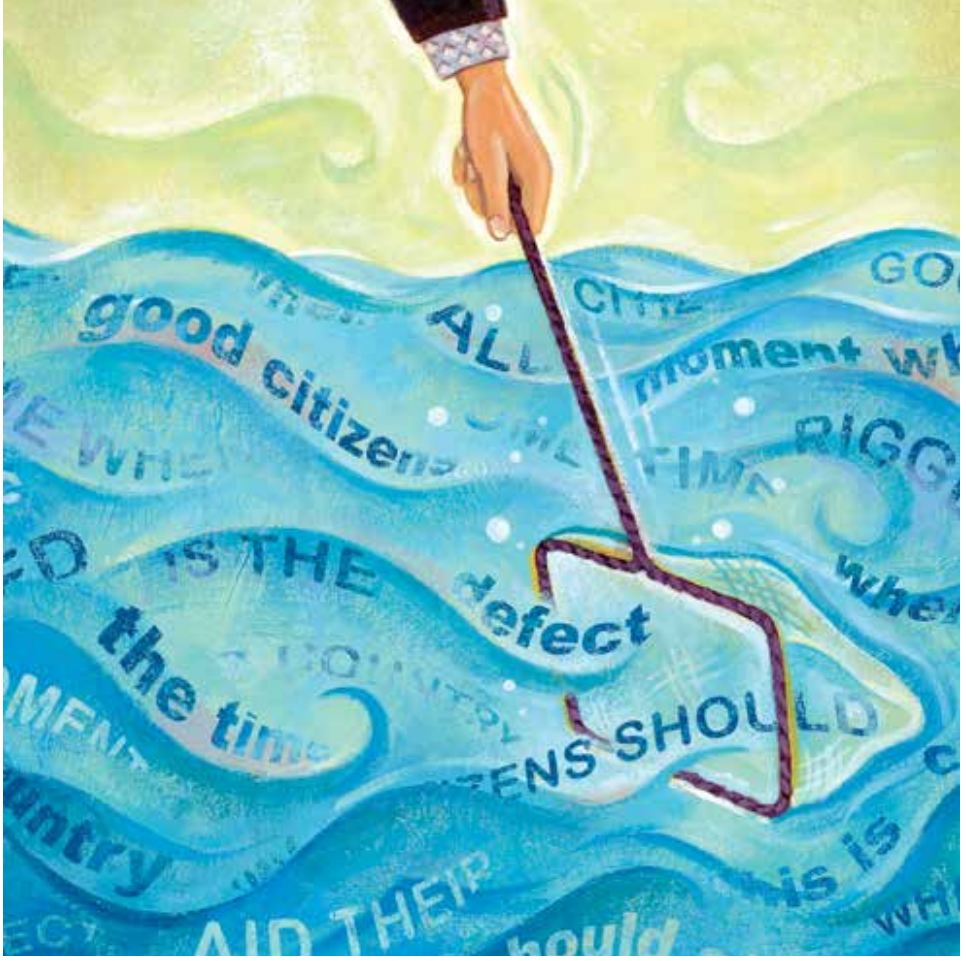
HEADNOTES



TRIAL PRACTICE

Analyze This!

WILLIAM F. HAMILTON

The author is the executive director of the UF Law E-Discovery Project at the Levin College of Law, Gainesville, Florida.

Legal cases begin with questions needing quick, accurate answers: What happened? Whom did it happen to? How did it happen? What relief does the client need? How quickly does the client need it?

Litigation, like any human drama, begins with a surge of emotions. Clients care passionately about the case: I was wronged! I am innocent! And, of course, whichever side they are on: Make the other side pay!

Like any good therapist, the litigator has to get past emotionally driven witness accounts that unconsciously add, embellish, hide, conceal, and distort, to find answers needed to develop strategies to solve the problems. Unfortunately, too many litigators bluster into the early court scheduling hearings and conferences with only the client's story. Often, that story is incomplete—or, worse yet, belied by the evidence.

How can we increase our confidence that witnesses are reporting their perceptions accurately? A few quick steps—enabled by wonderful digital properties of electronically stored information—will help us quickly get below the surface.

First, do a targeted collection from your key protagonists: Grab the basic office documents generated around the dispute. All the preserved data need not be collected. What we want for now are the data needed to unearth a reliable picture of events.

Next, load the data into an e-discovery tool that provides analytical features, for example, document similarity, near duplication, and relevance ranking. These tools will help you find patterns below the surface.

Thanks to the incredible processing power of today's computers, we can quickly build mathematically complex document indexes. Advanced analytical indexing provides a map of the below-surface data geography. With e-discovery analytical indexing, we can reveal word and letter relationships and frequencies in documents and across entire document collections.

So how does it work? Finding similar and related documents can reveal some untoward associations. Analytical indexing maps how often various words appear near one another (called word "co-incidence"), thereby revealing latent keywords, those mysterious characters hanging out together in heretofore unseen documents.

For example, with an analytical index, a keyword search for the word "defect" will also find words in your targeted collection in frequent proximity with "defect"—for example, "rigged." "Rigged" then is recognized as an important term based on its co-incidence with "defect."

Illustration by Sean Kane

Published in *Litigation*, Volume 43, Number 1, Fall 2016.

Suddenly, your search displays relevant documents that do not contain your original keywords; they contain close "friends" of those words.

"Friends"—words mathematically co-incident—of my keywords are my friends too, especially disreputable friends that I need to know about—whichever side I'm on!

Near-duplicate documents can have a critical impact. A slightly changed document often tells the hidden story of negotiations. To get below the surface of initial case documents, we need to find the near-duplicates that sometimes reveal unpleasant family secrets.

One near-duplicate analytical method looks for patterns of overlapping consecutive word groupings, called "n-grams." The greater the number of common "n-grams," the more the nonidentical documents resemble one another.

Let's take, for example, two very small documents, each composed of one sentence.

1. This is the time when all good citizens should aid their country.

2. This is the moment when all good citizens should aid their country.

We start by dividing these sentences into overlapping chunks (n-grams) of two words each:

this is	***this is***
is the	is the
the time	the moment
time when	moment when
when all	***when all***
all good	***all good***
good citizens	***good citizens***
citizens should	***citizens should***
should aid	***should aid***
aid their	***aid their***
their country	their country

These two documents have 9 similar n-grams out of a total of 13 unique n-grams. That's a score of 9/13, or .69. A score of 1.0 would be a perfect match. We can set our n-gram near-duplicate retrieval score higher or lower. We can thus find documents that are nearly identical to one another, or we can choose to find documents whose resemblance is not so close. Document changes and modifications—what is different—are often what is important for unfolding a case story.

Document importance also is important. We want to see the most important documents first. No one has time to spend years in therapy in superficial conversation and chitchat. Get serious and go to the important stuff first.

Basic keyword searches produce a flat, democratic, surface universe: Every retrieved document equals every other retrieved document. Floating below the surface are the really important key documents—but where?

Using a variety of technologies with exotic names such as "term frequency/inverse document frequency" and "vector space cosine value," our analytical tools can mathematically crunch the retrieved documents to rank the search results and put those with the highest likely relevance at the top.

With the right analytical tools, an early targeted collection—without a full-blown review—will give you increased comfort that your client's story will not give way to a troubling nightmare. Let your data tell you the real story early—don't be surprised later. ■