In fraud and corruption investigations, artificial intelligence and data analytics save time and reduce client costs

27 June 2018

When a legal team needs to find the facts behind fraud and corruption allegations in a government investigation, technology can drive substantial new efficiencies. By filtering and evaluating vast amounts of information, artificial intelligence (AI) can effectively sort text messages, audio files, e-mail, and other unstructured data into manageable groups; identify potential relationships between parties accused of fraud or corruption; and recognize patterns of frequency or timing, which may support a client's defense. Technology-assisted data analysis can provide the diligence and reliable quality control needed to provide the government with conclusions they can trust.

In this hoganlovells.com interview, Peter Spivack, a partner at Hogan Lovells in Washington, D.C., explains how the process of gathering, sorting, and evaluating enormous volumes of data has changed, and why skilled human intelligence is likely to remain a required component of an accurate analysis.

Where does the data come from in a government investigation, what are you looking for in that data, and how do you use it in your case?

Peter Spivack: When we do investigations for a client, we're trying to determine the facts. There's usually a very vague allegation that comes in, maybe through an internal whistleblower hotline or a subpoena from the government. We have a list of documents that a government entity is requesting, a complaint that somebody called in to a hotline, an anonymous e-mail that reports some allegation, or a news article.

So there's a certain amount of information that may show us the ballpark, for example, but we don't know what row and what seat we're sitting in. We're trying to determine if there is an issue, and whether it's the same issue that's been identified. If it's a government investigation, what's the government looking at? What's the strength of the evidence? What are the legal or regulatory defenses that we can use to advocate? What's the client's exposure? And how do we explain this to the in-house general counsel, chief compliance officer, board, audit committee of the board, and outside auditors, to give them an assessment of what the risk is?

And if we're going in to see the government, how do we explain that we've done an investigation that they can rely on and found the relevant facts, or have taken sufficient steps that, based on the sources of information available to us, we can rule things out? Because they're not going to simply say, we trust you and if you tell us you've looked at five e-mails, that's all you needed to do. We need to be able to tell them that we've looked at the whole picture.

Then there are other constituencies that drive investigations, especially for big public companies. Are they trying to get a line of credit? Are they looking at a possible merger where someone may ask them, as part of due diligence, do you have any issues? If you do, what have you done to look at the issues, what steps have you taken to resolve them, and what confidence do we have in the result?

There's a variety of things that drive us to try to determine what the facts are, depending on the situation, and sometimes they are present all at once.

After determining the approach to the investigation and the data you need, you then have to review the data sources. Where is that data stored?

Spivack: The way companies keep data is basically structured and unstructured. Structured data is essentially kept in an accounting or enterprise resource planning (ERP) system, such as SAP or Oracle. The data housed there is a record of all the transactions they've undertaken, and we'll work with a forensic accounting firm to define a set of data analytic tests that we can run.

Those tests can be a variety of different parameters we can flag that can be used to show basic fraud or corruption criteria. For example, are there round-number transactions? Are there sequential invoices to the same vendor? It may seem strange that the vendor is getting Invoices 1, 2, and 3, when it's a vendor that ostensibly would have many other customers. Is there a mismatch in the location of the work and the actual route of payment? Maybe the work is being done in Colombia, but the payment is going to France. We may need to ask for an explanation.

You run those parameters across structured data and come up with transactions that can be tested by taking what's been journaled in the accounting system and looking at the underlying documents. If you have a contract, does the contract description match the payment description in the system? If there are deliverables under the contract, are they general and vague, or measurable and specific? Can we determine that the transaction has actually taken place?

You can narrow the scope of those data analytics if you've got a specific question. For example, we think that this consultant is allegedly paying bribes to government people, so we're going to look at that consultant, the contract, the signatories on the contract, and the description of work under the contract. We're going to see what evidence there is of the work. We're going to look at the payment terms and say, does this seem like something that is commensurate with the value

being delivered? Is it a fair market value? We're trying to hone in and test the bona fides, so to speak, of that contractor business arrangement.

Take an example: one of our clients paid a lot of money to hire a well-known lawyer from another firm. But if you looked at that lawyer, you'd say, he doesn't really seem to have expertise in that area. So how do you explain that? Maybe there's a legitimate explanation. But it's something that comes up and so we say, we want to look at that further. It doesn't necessarily give you the answer, but it focuses things for you to look at.

What is unstructured data, and how do you use them?

Spivack: That's basically the way that people use communications systems. It's text messages, e-mails, messaging apps like Viber and WhatsApp, and other types of point-to-point encrypted communications. There's been an explosion of unstructured data — so much more than there used to be.

My first investigation at Hogan Lovells was for a company that was under investigation for promoting its product off label. They had human growth hormone that had a very specific use, and the government was concerned that they were promoting it widely for other unapproved uses. I was literally looking at hard-copy documents and putting them in Redwelds, depending on which paragraph they were responsive to. That was in 1998. Twenty years later, we'd never do something like that, because we've gone from 100 boxes of documents to 200 terabytes of data, and one terabyte is enough to fill the U.S. Library of Congress. So there's got to be some way to manage all that data and filter it.

How do you start narrowing down data?

Spivack: The first step is collection: you've got to go out and actually get it. That means, looking at the e-mail system. If we think of unstructured data as a series of concentric circles, it also means going out one ring and getting devices that people use, like laptops, and imaging their hard drives. And going out another ring if you can, depending on data rules, and collecting peripheral devices — smart phones, external drives, USB sticks — that store data.

So now you have this immense amount of data, more than any team of lawyers could review if they reviewed every single document for the rest of their lives, their children's lives, and their children's children's lives. It's clearly an unmanageable amount. The only way to address that is to try to process it and get it all in a form that can be managed and filtered. You try to exclude things that may be very data heavy but are of little value: program files, photographs, things that are really large that suck up data storage space. Then you have a set of data that you try to filter.

What techniques do you use in filtering?

Spivack: The most basic technique is search terms. You come up with a list of words related to

the investigation and apply them across the data to see if there are hits for documents that contain those words. Then you're reviewing them at first and second levels to see if they're relevant to the investigation. That sounds good, except you may only have narrowed your ballpark to the club-level seats; you're getting a tighter ring, but it's still an enormous amount of data.

There are other techniques that can be used as well, such as an algorithm. It's called technology-assisted review. You're taking a set of documents and reviewing that set with a subject matter expert on the investigation. They're going through a thousand documents and saying, this one is relevant, this one is not. You're essentially training an algorithm and honing it on the computer so that it can then give you a probability-based set of outcomes for the potential relevance of documents. The probability stratifications can be in 10 percent levels, so you have buckets, from a very unlikely probability bucket to a highly likely probability. You might be able to review the first two buckets, so you screen out a portion of your documents that way.

There is nothing available yet that is really AI, but there are ways of doing concept searching with certain applications. One that we use is called Brainspace, and it's basically a sophisticated form of the algorithm that groups concepts. You can run a set of documents through Brainspace and decide what concepts to look for. If you want to look for "office leases," it will group documents around that. You get a set of documents that you can then review for the concept of "office leases," whereas if it's payments to a particular third party, you can group them around that concept as well. That gives you more ability to target and focus.

A lot of times we'll run different techniques as a way to cross reference. That helps get through larger amounts of data at a higher and more efficient rate. But at the end of the day, it still depends on human evaluation and intelligence to look at a document and say, this is important, as it's related to things that we're talking about, or there's a particular issue here.

While we're doing that, you have to remember it's a dynamic situation, so there may be something that comes out of the transaction and data analytics that then says, wait a minute, we really want to look at this company, so let's run that as a new term through whichever technique we're using. Or we may put these documents together for interviews. We have a set of documents, we go to talk to a witness and say, what happened? And they tell us but then they raise another issue. That gets fed back into the review of documents and transactions to see if there is anything here that we have to be concerned about or if a new issue has come up.

Or there might be another whistleblower e-mail that comes in, competitor complaint, or newspaper article. So a lot of times, it's a dynamic process. You don't have just a static set of issues that you're looking at. One of the fun things about it is that there's this constant evolution.

Who is a typical client in this scenario, and what's the primary benefit of this approach?

Spivack: This is a way of making investigations more efficient, and efficiency means cost effectiveness. Clients are getting more and more comfortable with data and techniques of analyzing data, to the point where some clients in their compliance programs not only have lawyers and accountants, but also data scientists. A big multinational client, with tens or hundreds of thousands of employees, has people on staff who can design the most current state-of-the-art search engines, train algorithms, and use them as a way to leverage resources.

It's the outer edge of clients who have that capability, and they have to be big enough and operate in enough countries that they'll use it. If they're not, then we work with forensic technologists, both in-house and at forensic consulting companies. They're very familiar with different search techniques and technologies and the way to leverage them to process and filter larger and larger amounts of data.

There have been significant advances in technology in the last few years, and more and more interest in it. A lot of it is because of the ever-increasing amounts of data, and as a result, ever-increasing cost. There have to be ways to get costs to a controllable, reasonable level. So we have to know how to do that and work with people who understand and are comfortable with the concepts. We have to be able to articulate both to the client — if they're unfamiliar with it — and to the government — to defend it — what we're doing, how we're doing it, and why it's reliable. The government uses these techniques as well, so most of them are very familiar with it. They just want to make sure there's a sufficient level for liability.

About Peter S. Spivack

Peter Spivack is one of the most experienced members of the Investigations, White Collar, and Fraud practice area, and served as a global co-leader of the practice for six years. His experience in the criminal arena includes antitrust, environmental, Foreign Corrupt Practices Act (FCPA), government contract, and healthcare matters. Peter has three decades of experience working with multijurisdictional investigations, including matters involving allegations of bribery and corruption under the FCPA, the UK Bribery Act, and other anti-bribery laws.

Contacts

Peter S.
Spivack

Partner
Washington,
D.C.

> Read the full article online