WHITE PAPER

ADVICE FROM COUNSEL: CAN PREDICTIVE CODING DELIVER ON ITS PROMISE?

By Ari Kaplan, Principal, Ari Kaplan Advisors
Joe Looby, J.D., CFE, FTI Technology

FTI CONSULTING | TECHNOLOGY

INTRODUCTION

While the promise of predictive coding is alluring, many questions remain for corporations and law firms. Where does the software end and the importance of workflow begin? What can lawyers do to effectively defend its use? Are companies using it successfully? How much money can it save the client? Is the technology well-suited for all types of legal matters? Do law firms and corporations have similar perceptions of the technology? What is the perspective of the courts? Will future rulings impede or accelerate adoption of machine learning and classification technology in legal review?

Predictive coding has sparked discussion among e-discovery practitioners as no other technology has. From conference panels to blog posts, attorneys debate whether it will eliminate the role of human reviewers and question its capacity to defensibly automate and reduce the cost of e-discovery. Numerous providers of e-discovery software and services add to the industry hype by touting competitive software classifiers. Yet for all of the excitement surrounding predictive coding, the technology itself is not new. Patents for machine-assisted document classification tools have existed since the 1970's, and machine learning patents relating to e-discovery are a decade old.

Predictive Coding:
Predictive coding uses classifiers to extrapolate human coding decisions made on a subset of materials to a larger data set. It requires an iterative approach that includes statistical sampling and quality assurance to refine and improve the classifier. Predictive coding (*a.k.a. supervised learning*) relies on known properties derived from the sample set to identify matches in a larger data set. Predictive coding is not concept clustering or keyword search.

FTI Technology commissioned an interdisciplinary survey of law firm leaders and senior corporate counsel to identify key trends and perspectives on the emergence of predictive coding. This report summarizes the findings collected via phone interviews in April and May 2012. Thirteen in-house counsel from Fortune 1000 companies and 11 Am Law 200 (*American Lawyer*) law firm partners and senior counsel were interviewed for this report. More than half of the respondents have used some type of predictive coding technology, and all expressed an interest in the industry discussion around predictive coding.

Respondents were asked a wide range of questions relating to predictive coding, including their thoughts on high-profile court rulings, cost savings estimates, and adoption inhibitors. The feedback reveals widespread interest in the technology, evidence of successful predictive coding pilot projects, and optimism about its ability to reduce e-discovery costs in the long term. However, skepticism and uncertainty remains. Respondents expressed a healthy amount of concern about the reality of cost savings, the defensibility of the technology and its usefulness for certain types of legal matters.

PARTICIPANTS

FTI Technology focused on the perspectives of large corporations and law firms. The 13 participating in-house lawyers represented Fortune 1000 companies within the manufacturing, insurance, life sciences, energy, financial services, retail, technology, and transportation sectors. All 11 of the law firm participants practice at Am Law 200 firms. All responses were provided under condition of anonymity.

Number of litigation events reported by corporate counsel in the past 12 months:

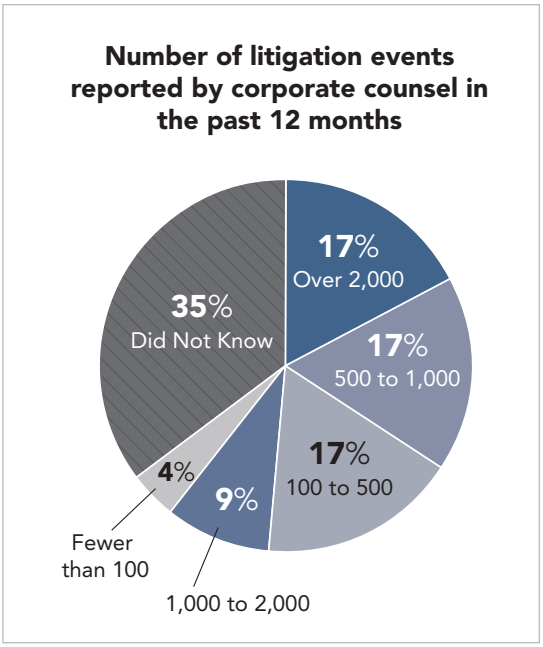
- 17% reported over 2,000 litigation events
- 9% reported between 1,000 and 2,000 litigation events
- 17% reported between 500 and 1,000 litigation events
- 17% reported between 100 and 500 litigation events
- 4% reported fewer than 100 litigation events
- 35% did not know the number of litigation events

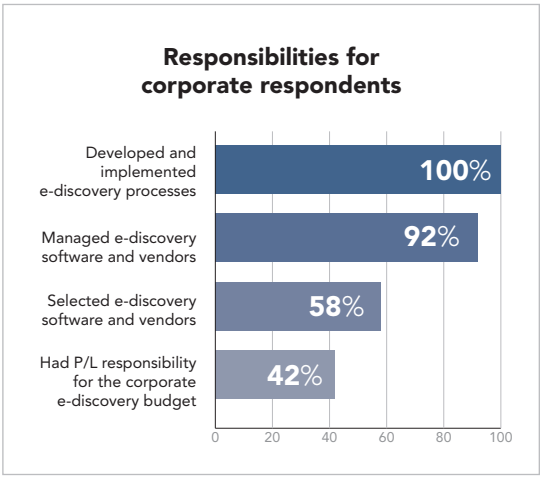
Responsibilities for corporate respondents:

- 100% developed and implemented e-discovery processes
- 92% managed e-discovery software and vendors
- 58% selected e-discovery software and vendors
- 42% had P/L responsibility for the corporate e-discovery budget

Responsibilities for law firm respondents:

- 100% developed and implemented e-discovery processes
- 46% recommended e-discovery software and services for corporate clients
- 42% selected e-discovery software and vendors





KEY FINDINGS

Among all of the survey responses, a number of recurring themes emerged:

LAWYERS ARE TRYING PREDICTIVE CODING WITH SOME PROMISING RESULTS

A lack of predictive coding case studies or first-hand testimonials at conferences may lead to the assumption that adoption has been light to date. The survey results partially support this, indicating that predictive coding may not be ready for "prime time" use but there are positive signs for future adoption. While 54% of participants reported use of predictive coding, the majority of these appear to be through pilot projects with predictive coding vendors. For those that have yet to use it, many are considering pilot programs and anticipate making a decision about its use in the coming year.

Of the 54% that have used predictive coding, 91% of them described the project as successful, for a variety of reasons. One respondent explained how predictive coding helped on an internal investigation in which 1.6 million documents needed to be reviewed in three weeks. "It could not have been done without a technology-assisted review."

> "It is not a magic bullet on its own."

Another in-house respondent discussed how predictive coding not only helped with a review of 100,000 documents, but also was used to help the company's outside counsel assemble deposition kits. "We got a lot out of predictive coding other than straight document review."

A third discussed how predictive coding helped ensure more consistent and higher quality review results. "I think predictive coding is superior to the alternatives because in a linear review model with contract reviewers, you are asking the least knowledgeable people in the case to review the documents. You are subjecting the documents to review by multiple people with divergent understanding or views on the case."

That said, "it is not a magic bullet on its own," according to one of the predictive coding users. "There was manual review beyond the seed set. That is the challenge in defining all of this. You could rely on the seed set and coding, but as the producing party you would not have learned your own documents. There is an additional cost."

FOR MOST RESPONDENTS, PREDICTIVE CODING REALLY MEANS FAST CULLING AND PRIORITIZING

As the above respondent indicated, predictive coding is not a silver bullet for automatically coding an entire data set. In fact, very few respondents appeared to use the technology for actual document coding. The majority discussed using predictive coding technology as advanced keyword search functionality so that reviewers could focus in on important materials faster.

"For years, we have tried to narrow the corpus of documents that we need to review by hand. We have not used predictive coding without further validation or checks, but have used it to narrow down what humans need to review," said an in-house attorney with a manufacturing company.

"With predictive coding, you are getting your hands on the key documents much earlier in the process than in more traditional approaches," said a law firm respondent.

As a result, several mentioned the ability for predictive coding to support early case assessments (ECA). "Predictive coding makes the review of the mass body of information faster and more efficient; you can get to an early case assessment conclusion much more quickly than linear review," noted one lawyer. "The success is almost like ECA success where we get to the facts quickly and end the case quickly," claimed another law firm respondent.

THE VERDICT IS STILL OUT ON COST, AS WELL AS POTENTIAL SAVINGS

Perhaps because the majority of predictive coding projects were conducted on a trial basis, respondents had a hard time quantifying the costs on a real-world matter, or even potential cost savings. 27% of those using predictive coding could not assess the costs, in fact. Of the remaining respondents, 27% estimated the costs savings between $25,000 and $250,000, 18% estimated savings between $250,000 and $500,000, and another 27% said they probably saved over $500,000.

Only one respondent was able to assess the organization's savings in terms of an exact figure ($580,000). "It does pay to run the numbers. Otherwise, you are making it up as you go along," the lawyer said. It seemed, however, from many responses that while not 'making it up', a number of lawyers are still struggling with concrete methods of determining their cost savings.

The respondents provided general estimates of cost savings. One estimated saving $25,000-$50,000 because "the predictive coding tool was able to filter out 80-90% of the documents, allowing the company to avoid reviewing 80,000-90,000 records." Another highlighted that while the law department did not calculate an amount, "30% is the right order of magnitude [since] we are reducing the costs to 60-75% of what they would have been." One law firm partner noted, "We have achieved savings of over 60%."

> "It is not just the technology you are using, it is how you employ it."

In lieu of calculations, others simply remarked that they must have achieved cost savings. "There were some huge data sets and we made substantial inroads on them," said one. "We have collected all of the documents and reviewed them; we use predictive coding to be thorough," said another, who added, "We don't necessarily use it to compare a delta of costs."

Despite the allure of streamlining spending, it seems that both law departments and law firms are still trying to balance the mere usage of the tool with its potential. They are also trying to determine the sizes and types of cases for which it is most appropriate.

"GARBAGE IN, GARBAGE OUT"

Much of the dialogue surrounding predictive coding relates to its potential for eliminating or reducing human review, the most expensive e-discovery phase. However, with predictive coding, the respondents overwhelmingly agreed that humans become an even more important part of the process since they determine the reference set, test and refine the software, and conduct additional quality control to ensure defensibility. "It is not just the technology you are using, it is how you employ it," said one.

Or, put another way it is "Garbage in, garbage out. It depends on how good the humans are in giving you the seed set. You can't just push a button and a machine will provide results," advised one law firm partner. Quite simply, the quality of the predictive coding results is dependent upon the quality of the predictive coding input, as the software will propagate bad coding decisions as well as good coding decisions.

Given the requirements for creating a seed set to properly execute the predictive coding model, many described it as an interdependent relationship. "I don't think you can ultimately replace humans; you will still need some combination," said one attorney.

SOME MATTERS ARE BETTER SUITED TO PREDICTIVE CODING THAN OTHERS

Although predictive coding is not necessarily designed for a specific type of matter or case size, 88% of respondents agreed that there are certain parameters that may make its usage successful. "There is a threshold below which predictive coding doesn't make financial sense or it doesn't work the way it might in *Da Silva Moore*," said one lawyer. To provide a bright line, the participants seemed to agree that the minimum number of documents for which it is appropriate is 100,000. And, "If the total case value is less than two hundred thousand dollars, it may not be worth it," said one in-house respondent.

Collectively, the respondents listed the following instances where a legal team should consider using predictive coding:

Cases of a certain size:

- Any matter of at least 100,000 documents, which could require 10-20 attorneys to review the material in a minimum of two weeks.
- Cases involving 50 or more custodians.

Particular types of matters:

- Commercial litigation with similar or consistent data types.
- Class actions.
- Cases with straightforward definitions of relevance.
- Multilingual cases, assuming the software classifier has multi-lingual capabilities. If so, multi-lingual software may be more reliable than an assortment of native speaking reviewers.
- FTC or DOJ requests requiring large-scale review in a relatively short amount of time.

The participants also highlighted that predictive coding technology may be less effective for:

Cases of a certain size:

- Cases with low volume, unless there is a budget that works.
- Small single plaintiff cases.
- Medium and small matters involving few custodians with limited document sets.

Particular types of matters:

- Employment cases, or any case with a large population and a small number of responsive documents.
- Patent litigation because key documents are often drawings or visuals, which are not typically conducive for the technology to review.
- Cases with a high number of documents that are not text based (e.g., photographs, images, audio files).

On a related theme, the respondents were asked to rank the top reasons they would or would not use predictive coding. The following list represents the top four ranked items for both categories, compiling the percentages each answer was given as one of the top four benefits and concerns around predictive coding.

TOP-RANKED BENEFITS OF PREDICTIVE CODING

1. **Prioritize documents to be reviewed by humans**
   As discussed earlier in the paper, current predictive coding users and those performing pilot programs are using predictive coding to focus in on the most important documents first. It's no surprise, then, that 92% of respondents listed this as one of the top four reasons to use predictive coding.

2. **Cost-effectively eliminate irrelevant documents**
   78% of respondents gave this as a top reason to use predictive coding. Comments included:

   - "It really does ferret out the irrelevant documents."
   - "We need to get down to the really relevant materials. This helps you get to the heart of the matter faster."
   - "You want to be able to fulfill your obligation of duty and candor to the tribunal. You care about the quality, but you just want to meet your obligation. You don't want to do the best job in the world; document review is a commodity."

3. **Test the results of human review to ensure accuracy of coding decisions**
   57% of respondents suggested that predictive coding would be an effective tool for quality assurance on human reviewers:

   - "I think people make mistakes and eyes get tired."
   - "We had to develop our own protocol around this."
   - "There are errors no matter what systems you use. You'll never get 100% precision and 100% recall."

4. **Find more responsive documents**
   Predictive coding could be an important tool to ensure a thorough review, according to 55% of the respondents:

   - "In utopia, we don't want to produce anything, but we will need to do so and want to find the responsive documents."
   - "This is especially important if we are talking about the opposing party's production."

Those reasons not ranking in the top four of predictive coding benefits included replacing costly human review with machine review, a belief that machine review could get better results than human review, and that predictive coding results were more defensible than human review.

TOP-RANKED CONCERNS ABOUT PREDICTIVE CODING

1. **It is a "black box" technology**
   More than two-thirds (74%) of the respondents expressed concerns about the "black box" nature of predictive coding technology and its propagated outputs. Some of this was due to the technical nature of the software and process, which will be discussed in the below "experts needed" section.

For others, though, there was a sense that predictive coding providers were not being transparent with the respondents. "In many cases, it is impossible or difficult to lift the hood on predictive coding," said one. "There are some techniques that are a little 'black box' and about which I'm not fond," said another. Connected to this, respondents expressed frustration trying to test and compare predictive coding solutions from various providers. "Different vendors have different technology so you are not getting the same thing. Depending upon who you go to, one tool may be better than another," said one.

"In many cases, it is impossible or difficult to lift the hood on predictive coding."

"We are most scared of the fact that we cannot explain the functionality of some of the tools."

Several more expressed general frustration that "black box" predictive coding would be difficult to explain in court. "Look at the hassle in *Kleen*, which required two full days of testimony to inform the court about predictive coding," reported one attorney. As another stated, "We are most scared of the fact that we cannot explain the functionality of some of the tools. This is the biggest reason that corporations or law firms have been afraid to move forward. There has been an unwillingness to explain it."

2. **It is not well-suited for investigations or finding the "needle in the haystack"**
   70% of respondents gave this as a key reason why they would not use predictive coding. While several commented that most forms of e-discovery don't perform well for "needle in the haystack" investigations, a few commented that this was especially true for predictive coding. As an example:

   > "The fundamental lynchpin of predictive coding is being able to validate the null set and decide that there are not responsive documents in that set to some degree of confidence. It is really hard to validate that null set and you would need to create a sample size that is so large that you would have to essentially review the documents manually. It is a terrible system for finding documents when only 1% of the material is responsive. You would need an enormous sample size so it is not a good system for finding needles."

3. **Fear of inadvertent productions**
   66% stated that risk of inadvertent productions would inhibit their use of predictive coding. A few stated that they "always fear inadvertent production" and others stated that this fear is the reason why they probably would never use predictive coding for anything more than prioritizing documents for review.

4. **Respondents do not want to be early adopters**
   Perhaps due to the risk-averse nature of the legal industry, 64% of respondents stated that they would rather not be an early adopter of predictive coding:

   - "We don't want to spend the money to be the guinea pig; we would rather have other people go through it and learn from their experience."
   - "We try to be on the cutting edge, but we don't want to be the first ones out there using something the courts find has no validity."

"It is a terrible system for finding documents when only 1% of the material is responsive."

Those reasons not ranking in the top four of predictive coding concerns included waiting for clear Federal case law on the use of predictive coding, as well its cost, which will be discussed in the next section.

IN CERTAIN CIRCUMSTANCES, PREDICTIVE CODING CAN COST MORE

While cost concerns did not rank in the top four among respondents, the topic was definitely a running theme throughout the survey results. In fact, 62% of respondents expressed concerns about the cost of predictive coding, narrowly missing the top four areas of concerns. While many of these qualified their cost concerns by saying it depended on the urgency of the case, a few common themes emerged:

> "It saves money in the long-term, but it is very expensive in the short term because you need senior level reviewers at the outset."

It is a big cost up-front but in the long-term it reduces cost.
According to one respondent, "I'm not convinced that it is more expensive. I wouldn't be considering it if I didn't think it saved us money. It costs more in the beginning, but you shouldn't be considering it if it doesn't save you money." Another echoed this sentiment by saying "it saves money in the long-term, but it is very expensive in the short term because you need senior level reviewers at the outset." Interestingly, a law firm respondent discussed how this proposition – larger upfront costs but more long-term savings – was particularly difficult to sell to the corporate client.

It is dependent upon the scope of the matter.
"I don't think it makes sense for every case; only for big data cases."

> "Integrating predictive coding with an attorney review workflow is absolutely key as workflows are one of the most difficult things."

EXPERTS NEEDED

Because predictive coding is just as much a process as a technology, there was widespread acknowledgement that experts would be an integral part of the process. "With respect to predictive coding, there is a small group of attorneys who know what they are doing. I would predict because of the way technology works that the market for predictive coding of discovery materials consolidates to a fairly small group of technology providers that become the industry standard. You will still need lawyers who run those machines correctly."

> "With respect to predictive coding, there is a small group of attorneys who know what they are doing."

In addition to understanding how the software classifier works, respondents listed other responsibilities for predictive coding experts: collaborate with the review team to create a statistically valid reference set, monitor and test predictive coding results through statistical sampling and validation, integrate the technology with human review workflow, and be able to testify about the technology and process in court.

"There is a whole workflow that makes predictive coding more complex than linear review," said one. "Integrating predictive coding with an attorney review workflow is absolutely key as workflows are one of the most difficult things," highlighted another participant.

In addition, many of the attorneys interviewed were blunt in their lack of technical knowledge – or need to understand the technology or process. "With respect to understanding different predictive coding algorithms and classifiers, there are people smarter than me. We could run a test case and I don't know that it would be worth my time to understand the technology at the level of the algorithms." Or as another said, "Understanding the different predictive coding algorithms and classifiers is for the technology geeks to deal with."

EMERGING CASE LAW IS HAVING AN IMPACT

100% of the respondents have heard of and are following developments around two predictive coding matters, *Da Silva Moore v. Publicis Groupe*; 11 Civ. 1279 (S.D.N.Y. Feb. 24, 2012) and *Kleen Products. LLC v. Packaging Corp. of America*; 1:10-CV-05711 (N.D. Ill.). 58% said they were more likely to use predictive coding because of these cases.

The difference between the impact on those at law firms and those in-house is particularly striking. Of the lawyers who admitted to being influenced by the positive judicial commentary associated with predictive coding, 69% practice in-house while only 31% are with law firms.

In fact, many of the law firm partners interviewed noted that the decisions do not change their approach because of their pre-existing belief in the technology. "The decisions have not affected our use of predictive coding as we started using the technology two years ago," said one law firm partner. "It is nice to see it is an industry-accepted practice, which is a positive thing, but we use it when it makes sense," the lawyer added.

There was not, however, universal acceptance of the decisions. "*Da Silva Moore* is good in that it popularized predictive coding, but bad because of the misinterpretation of statistical sampling," commented one lawyer. "I don't think you need a court case to bless it; if we were to get a court case that said don't use it, it would be a concern," added another.

From a law firm perspective, "It endorses the decision to use predictive coding enunciating what the law already was; it was a fairly predictable outcome." Also, there was more than one comment highlighting that the ultimate use of the technology is a client decision.

The majority felt that it was bound to happen at some point. "The dam was going to break as soon as a judge issued a holding establishing the propriety of using the technology." An in-house participant seemed to affirm this view, highlighting "We try to be on the cutting edge, but we always feel more comfortable investing in it and using it once a court has blessed it."

CORPORATE AND LAW FIRM ATTORNEYS ARE LARGELY IN ALIGNMENT

As mentioned above, corporate and law firm attorneys differed on whether current predictive coding cases and rulings would impact their predictive coding use. Interestingly, that was one of the few areas where corporate and law firm responses diverged. All survey participants generally agreed on the majority of issues raised, including key benefits, ideal matter types and chief concerns. In addition, there was little disparity between the two groups in other survey areas, including the percentage piloting predictive coding and those able to assess costs.

LOOKING AHEAD

Based upon the survey responses, it is possible to forecast coming trends that legal teams should be aware of:

PREDICTIVE CODING EXPERTISE WILL BE IN HIGH DEMAND

Predictive coding is a complex process that combines technology, workflow, legal review, project management and statistical sampling. The quality of predictive coding results is dependent upon the expert orchestration and coordination of all these components. The growing need for this very specific skill set will lead legal teams to seek outside expertise for developing and implementing defensible predictive coding programs.

PREDICTIVE CODING WILL BE LESS BLACK BOX

Lawyers unsatisfied with the lack of complete transparency into predictive coding technology, workflow, and overall defensibility will push for additional technology or methodology, such as visual verification, statistical sampling and quality assurance guidelines. Classification inputs, processes and outputs will need to be sufficiently documented and explained to clients and courts. Any adverse court decisions on predictive coding will likely accelerate this trend.

MULTIPLE CLASSIFIERS FOR MULTIPLE MATTERS

A number of different classifiers exist in the market today, and this number is likely to grow. Each is different, and each may perform better or worse depending on the data set or review objectives. Due to the number of different classifiers, as well as the number of differing variables in each matter (e.g. language, investigations versus litigation, deadlines and size), corporations are likely to test several classifiers and opt for the one that works best for that particular matter. Thus, it may be prudent for corporations and law firms to avoid "purchasing a predictive coding technology" and instead to "select the right tool for each job" to avoid investing in yesterday's technology.

SUMMARY

Corporations and law firms are taking a measured and pragmatic approach to the adoption of predictive coding technology and workflow. Pilot projects, a reliance on "techno-lawyer" experts, as well as a focus on the types of matters best suited for predictive coding, demonstrate a keen understanding of the technology's promise and potential hazards. As organizations transition from pilot programs to regular use, the industry as a whole will benefit from broader focus on and awareness of these potential hazards, especially around cost calculations and defensibility. With this focus, predictive coding can deliver on its promise.

ABOUT THE AUTHORS

JOE LOOBY

Joe Looby is a Senior Managing Director with FTI Consulting, Inc. ("FTI"), and has more than 20 years of combined experience in the military, regulatory enforcement, investigations and technology, much of which has involved the research, documentation, explanation and validation of proprietary software technology, computer models, and technical processes. Joe is a certified fraud examiner ("CFE"), with a Bachelors degree in Economics, a Juris Doctorate, and is listed as the co-inventor of U.S. Patent "System, Software and Method for Examining a Database in a Forensic Accounting Environment." From participation in the National Institute of Standards and Technology's (NIST) Text Retrieval Conference (TREC), to co-chairing The Sedona Conference's first Panel (circa. 2007) Proposing the Use of Statistics to Defend Reasonable eDiscovery efforts; co-authoring The Sedona Conference's Best Practices on the Use of Search & Information Retrieval Methods in eDiscovery; and, Achieving Quality in the eDiscovery Process. Joe is a regular speaker, author and consultant to law firms and corporations. A former U.S. Navy JAG Lieutenant, an experienced regulator, and published software developer, Joe has appeared before regulatory agencies and provided expert testimony.

ARI KAPLAN

The New York Law Journal called Ari Kaplan's first book, *The Opportunity Maker: Strategies for Inspiring Your Legal Career Through Creative Networking and Business Development* (Thomson-West, 2008), a "must-have treasure box of marketing ideas," and CEOs have described his new book, *Reinventing Professional Services: Building Your Business in the Digital Marketplace* (Wiley, 2011), as "an essential guide" that "expertly showcases the multitude of opportunities the digital age has brought to the professional services market." After nearly nine years practicing with large law firms in Manhattan, Kaplan, named to the inaugural Fastcase 50 list of innovators in the law, has become the leading copywriter and industry analyst in the legal community. He was the keynote speaker for the 2010 ABA Techshow, and has shared ideas with students and professionals throughout North America and the United Kingdom.

The author of over 200 articles, Kaplan has received multiple Apex Awards for feature writing. He has been recognized in the Wall Street Journal's Law Blog, the Chicago Tribune, the Miami Herald, the New York Post, the ABA Journal, Above the Law, the National Jurist, the Chicago Lawyer, and the California Recorder, among other publications. He also served as a legal commentator for CNET Radio and has been interviewed on CNN, WGN-TV Chicago and Good Morning San Diego. Named a "Law Star" by LawCrossing, Kaplan provides law-related ghostwriting for a number of companies, firms and individuals in the legal industry. He also provides consulting to individuals and organizations interested in creating deeper connections with law students, lawyers, legal administrators, and other legal professionals.