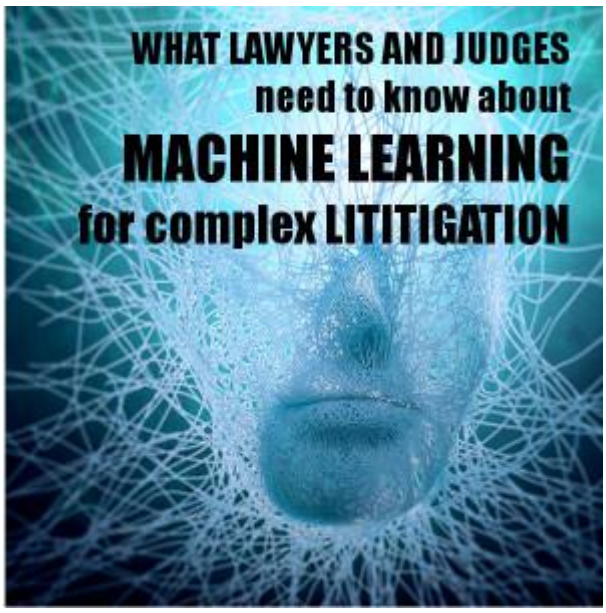
What Lawyers and Judges Need to Know About Machine Learning for Complex eDiscovery

Michael McCreary | March 20, 2013 | 21 Comments



Many judges and lawyers might start by asking whether they can trust machine learning to replace traditional legal review by contract or staff lawyers in complex litigation. In fact, the first question should be: how can we allow the current culling and manual review process to continue when we know that it cannot begin to fulfill legal obligations to the courts and to clients?

The reality is that the current process consistently and demonstrably fails to identify most of the relevant documents. Holds are often inadequate and the continuing obligation to produce is usually ignored because it is simply too painful to go back after the initial collection. Structured data is rarely given ample consideration due to the specialized knowledge required. Keyword searches used to cull the population have been proven to identify only a fraction of relevant documents, as well as a massive volume of irrelevant content. Contract lawyer review is hardly more accurate than a coin flip. Reviewers are poorly incentivized and often inadequately trained to understand the complexities of the case. The task itself is tedious, making consistency throughout a long day of review difficult with oneself, let alone with the many other reviewers over the course of weeks and months of review.

In the end, the current process is expensive for all parties, disruptive to the enterprises involved, and time consuming for the courts as they try to resolve discovery disputes. It is little more than a ritual, the legitimacy of which derives from a time when relevant data was more manageable and review was done by tenure-tracked lawyers who were better incentivized; conscious of their obligation to the firm, their clients, and the courts; and, most significantly, more aware of the issues in the case.

Not All Machine Learning is Created Equal

Even though machine learning has the potential to increase productivity and accuracy and reduce the overall cost of eDiscovery far beyond that offered by current document review processes, if it is not properly applied, it may fall well short of delivering on its promise. Many of the predictive coding and TAR solutions commercially available to today's eDiscovery market have several fundamental limitations that judges, lawyers, and the corporations footing the bill need to understand.

Machine learning has been used to interpret data since the 1950s and, as such, there exists a substantial base of high quality academic research bolstered by years of applied experience in healthcare and other industries. However, the prevalence of machine learning outside of eDiscovery in no way means that its application within eDiscovery is consistent or equal.

The first limitation relates to machine learning's reliance on mathematical algorithms and models for classification. There are currently over 30 different types of classifiers capable of machine-learning-based text classification, and

each one returns different results depending upon its configuration and the data set at hand. Given that most predictive coding and TAR solutions use only a single, preconfigured classifier, they are, at best, only able to provide high quality results when applied to data sets for which they are optimally suited. Thus, when the classifier and data are misaligned, results can be less accurate than even manual review.

The second limitation of most commercially available solutions is that they are essentially black boxes; they provide little transparency into the underlying algorithms and associated technology, and are unable to output a human-understandable translation of their results. As such, it is impossible for judges, lawyers, and experts to examine, question, and compare the selected approach and associated results.

The third limitation is an extension of the first two, i.e., many commercially available solution providers are hiding behind the opinion of Judge Peck in *Da Silva Moore*, where he suggested that the *Daubert* Standards do not apply to machine learning, and as such, vendors need not provide machine learning subject-matter experts. In *Da Silva Moore* the witness from Recommind clearly lacked an in-depth understanding of machine learning theory and could not discuss, nor demonstrate the validity of its application. Contrary to Judge Peck, we believe that per the *Daubert* Standard, machine learning providers are indeed producing evidence and as such should be expected to provide credible experts who can detail the inner workings and statistical validity of their approach, and successfully address the concerns of judges and opposing counsel under examination. Machine learning involves math and science that can make a significant difference to the evidence presented in complex litigations involving significant financial and legal issues and must be treated accordingly.

The final limitation is that both vendors and lawyers are almost exclusively focused on unstructured data. They think of it terms of documents, not data. The truth is that documents only tell part of the story. Often, millions or even billions of structured data records are, or should be, produced in discovery, e.g., financial transactions. Only through advanced analytics grounded in machine learning can we bring to light the hidden patterns and systematic behaviors locked within the data. By combining structured and unstructured data, we can identify both what happened, i.e., the structured financial transactions, and why it happened, i.e., the associated communications and decision-making. Moreover, this level of analysis leads logically to the inclusion of a new step in the process – validation. Using advanced analytical tools it is now possible to evaluate the whole of a production and identify logical gaps and missing content. This step is essential given the inherent requirement that discovery processes must limit the scope of enterprise data under consideration at the outset, typically by custodian, date range and data source.

With roots in automating the paper processes of the past, it is no surprise that most eDiscovery vendors face challenges with respect to structured data. In addition, most of the early entrants into machine learning for the eDiscovery market are relying on rudimentary applications of the technology, so collecting and analyzing structured data from enterprise applications and relational databases has not been a priority, or even an option.

If the currently available predictive coding vendors only provide algorithms that perform well with certain data sets, are opaque, do not offer experts to attest to the validity of their technology, and cannot address structured data, they are failing to provide a truly useful and defensible solution to the eDiscovery market. We should not replace one flawed process with another.

What Lawyers and Judges Need to Know About Machine Learning for Complex eDiscovery

Strategic Use and Proper Application of Machine Learning

The current uncertainty in the market surrounding machine learning has given rise to an arbitrage opportunity, where lawyers and clients who embrace the new technology will be at a significant strategic advantage over those who do not. Smart counsel knows that if they can convince their adversary to agree to a list of keywords and a manual review of documents, the opposition will not receive the majority of the documents to which they are entitled. Similarly, lawyers providing the best representation to their clients will not agree to keyword culling for documents produced to them; and they will not agree to a manual review of documents by poorly trained and inexperienced lawyers.

The current outdated methods of collecting and reviewing ESI have been tolerated because there was, until recently, no better alternative. New technology provides an alternative that is more comprehensive, more accurate, and less expensive. It will ensure compliance, transparency, and sound practice. As the application of machine learning becomes industry standard, courts and lawyers must rigorously examine and challenge the experts who have developed and applied machine learning to eDiscovery. Courts and lawyers need to demand machine learning solutions that have the flexibility to match algorithms to the data on a project-by-project basis. They need to demand transparency from machine learning vendors and abide by the *Daubert*Standard by offering expert witnesses to explain in a legally defensible and academically sound manner how their solutions produced evidence. Ultimately, a neutral third party is needed to help ensure that the right tools and processes for eDiscovery activities are put in place at the outset of a case to establish fairness for both sides and enforce cooperation. If these changes take place, the industry will see a paradigm shift in eDiscovery with machine learning technology supporting unprecedented accuracy rates at considerably less cost than current eDiscovery processes.

About the Author
Michael McCreary is the President and CEO of Rational Retention. With over 20 years of experience working at the intersection of business and technology with deep experience in highly regulated and litigious environments, he spent 12 years at Pfizer in various management positions within R&D, HR and Legal. Most recently he was CIO of Pfizer's Legal and Public Affairs divisions. As a member of Pfizer's IT and Legal Leadership teams he led the corporate technology strategy for eDiscovery, IP, Security, Privacy, Records Management and Information Risk Management. Prior to joining Pfizer he was a partner in a software development and consulting firm working for various clients in the manufacturing, energy and professional sports industries. Michael's team was honored by Law Technology News for the most innovative use of technology by a law department and he is a regular speaker and writer on topics including information life cycle management, retention risk management, auto-categorization and eDiscovery preparedness. Michael holds dual BA's from Union College.

Comments:

S. N. - "Interesting thoughts Don Quixote!!"

B. W. - "This article is spot on. It's about time that some had the guts to standup and speak the truth about what's going on in litigation technology."

B. T. - “Spot on! There is no doubt that the current system is broken and produces less than adequate results. Unfortunately, there is lots of money backing the status quo. Will take some major technology wins and embarrassing losses for predictive coding to take off.”

B. J. - “This article is nothing more than a marketing piece. And, Recommind has it all over Rational Retention – better technology, larger installed base (do you even have any customers?) and better sales and marketing organization.”

@ediscoverygroup - “Sounds like you must work for Recommind. Actually, Rational Retention utilizes machine learning technology from the Evidence-Based Medicine Information Retrieval and Scientometrics Laboratory (EBMIRSL) in the Center for Health Informatics and Bioinformatics (CHIBI at New York University). The technology has 9 patents and 8 provisional patents. It is currently being utilized at some of the largest corporations in the world to tackle some of the most complex information technology problems.”

P. Z. - “Timely topics and great article. The current eDiscovery systems are nothing more than an extension of the paper process and therefore it doesn’t work. It produces inadequate results and has grown up to do nothing more than increase billable hours.”

OrcaTec User - “I suspect your article is a little biased in favor of Rational Retention. However, I still agree with most of the points. Unfortunately for you, I have known Dr. Roitblat for over 20 years and believe that he understands predictive coding and litigation technology better than anyone in the industry. Our firm has used OrcaTec on 3 cases now and it has worked great. Will be using Orcatec for the foreseeable future. Best of luck with Rational.”

R. B. - “Really nicely stated!!”

T. N. - “You are barking up the wrong tree on this one. There is no way that lawyers and judges are going to allow a technology into the litigation process that replaces them. Not that it’s a bad idea. Just not going to happen.”

E. J. - “I think that this post is a complete self-serving exaggeration of the facts. Collection is fine, culling is fine, document review is fine. Predictive Coding is a black box no matter what you contend.”

T. J. - “It is really sad that our judicial system is so uneducated that they don’t even understand how evidence is produced.”

P. H. - “Great article. Predictive Coding is definitely the future. We just need to get the judges and the lawyers to accept that fact that manually reviewing every document is inefficient and costly. It certainly does help with billable hours!!”

S. W. - “This article is ridiculous. You have no idea what you are talking about. Predictive Coding is a scam and will never be mainstream.”

C. N. - “You have no idea what you are talking about. The current system and process is fine. It just needs a little tweak here and there.”

Young Technology Savvy Lawyer - “As a young lawyer that grew up with technology, I believe that machine learning or predictive coding is the next big thing in litigation. Unfortunately, my boss still uses a yellow pad and is only focused on increasing billable hours. I don’t think he really cares if we do a good job with eDiscovery or if our clients get a fair trial.”

"In the end, it is going to take a changing of the guard to enable technologies like machine learning to really take hold."

C. N. - "I have been involved in 5 cases in the last year in which we tried to utilize Predictive Technology. Each time the judge allowed opposing counsel to contend that it was overly burdensome to require both sides to understand and accept the results. Anyone have any suggestions in regards to how we can get judges to accept this technology?"

T. B. - "Tremendous article. Predictive coding is definitely the leap frog / paradigm shift technology that is needed to get us past all of this manual processing."

A. W. - "Nice article. However, the fact of the matter is that predictive coding isn't ready for prime time because the system has refused to take the time to understand the technology and the vendors have gone to great lengths to mask the rule truth about their predictive coding capabilities. It will be years before any of this works."

T. G. - Really well done. I think that predictive technology is going to have a very positive and long term impact on the quality of the eDiscovery and litigation. However, the market has been corrupted by vendors that are publishing crap like "Predictive Coding for Vendors". Unfortunately, it will take a while for this corruption to be uncovered and to filter through. Hopefully and ultimately, the best technologies will win.

T. N. - "Tremendous article. Thanks!! And Good luck changing the industry!!"

M. J. - "It's about time to have this discussion. You rock Rational Retention!!"

* * *

A Rational Reaction to your Comments

Michael McCreary | March 27, 2013

We were delighted by the responses to "How Judges and Lawyers Should Think about Machine Learning in Complex Litigation." And not just with these responses that were favorable, but also with those suggesting we were wrong about the transformative potential of machine learning in e-Discovery and the expert based model. Of course, we are committed to the expert-based model -- not only in the cases where competitive products are currently being applied (post collection), but in circumstances where litigation is not even anticipated. In our view the current technology, like the current e-Discovery model, is being applied too late to make a meaningful impact. It is in the upstream application of Machine Learning where a real difference can be made, but also where the current products have limited utility.

The typical application of Machine Learning today occurs after documents have been collected and culled or produced. The shrink-wrapped products will work better or worse depending on the particular collection. Every collection will be different, and every product will work differently on each collection. The best use of Machine Learning at that point will be driven by deep understanding the collection and applying the right model after testing and tailoring. But that is not the most interesting nor useful application of Machine Learning, or of the expertise needed to apply it properly.

A Rational Reaction to your Comments

A more interesting and important use of Machine Learning occurs before data is collected, or even subpoenaed. Indeed, the same Machine Learning tools and techniques used in predictive coding may be applied to content in-place, to both understand and control where, and for how long content exists in the enterprise. In the content aware enterprise, we can move towards real Information Governance where risky, (e.g., Intellectual Property, Confidential or Personally Identifiable Information), is identified and controlled; content that needs to be kept is retained pursuant to policy while transient and obsolete content is destroyed. By applying Machine Learning to content in place we can build a true information governance platform. However, the real transformation is in discovery.

Combining the ability to control (preserve, destroy and collect), content in place with machine learning will fundamentally and irrevocably change legal discovery. The current waterfall EDRM process is essentially obsolete. By combining real experts with access to content in place we will no longer be bound by the inherent biases and limited scope of the current process. Direct access enables a facile, iterative approach. Going back to the source is no longer a burden. Deeper analysis, high quality sampling and focused and relevant productions are all available directly from within the enterprise.

E-Discovery is no longer a factory manned by semi-skilled workers, but a profession led by real experts who have domain knowledge of the industry, understand technology and law, and can apply the next generation tools the complex problems they face.

It is important to note that the above is not merely a vision of what could be, but is rather a description of what can be done today. Clearly this approach makes obsolete much of the current e-discovery landscape; however this is the nature of progress and competition. In the end the customer wins, and while the costs savings will accrue to the companies footing the bill, the most important return is that e-discovery can finally facilitate rather than impede justice.