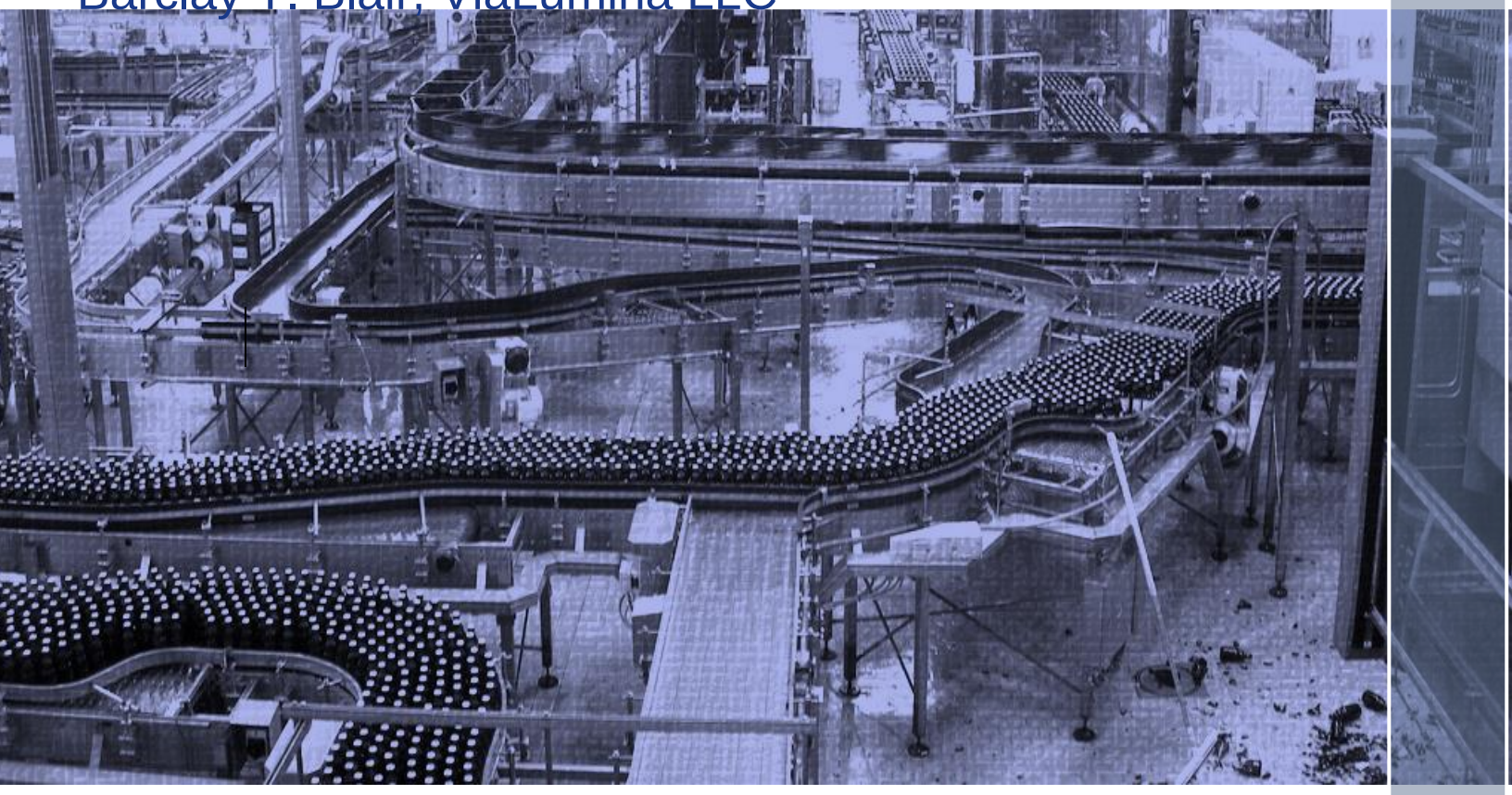
Predictive Coding: Making it Work

An exploration of predictive coding in practice, identifying key considerations for practitioners using predictive coding in their electronic discovery process

Barclay T. Blair, ViaLumina LLC



Published as part of Equivio's *Predictive Coding Minus the Hype* Educational Series

ViaLumina
INFORMATION GOVERNANCE ILLUMINATED

Introduction

> *"What is really astonishing is how quickly predictive coding is gaining acceptance. Only a couple of years ago, everyone was saying 'it's great, but it will never be broadly used.' How things have changed."*
>
> *Bennett B. Borden, Chair Information Governance and eDiscovery Group, Drinker Biddle & Reath LLP*

When I was a boy, growing up on a prairie farm, my father taught me that cutting the first round of a hay field was the most important, because every pass after that would follow the same path, and imperfections would only grow with each successive round. These errant wobbles and warps would make it harder to later bale the hay, wasting time and fuel.

Besides, the neighbors were watching.

I suppose my dad had learned these techniques from his father, and then perfected them on mile after mile around our fields. As a *practitioner* of farming, he had put in the time, and learned how to use the tools and technology at his disposal most effectively.

Technology, on its own, does nothing. It is only when technology is put to work in a context that its value is realized. Technology truly comes alive in the hands of practitioners who have a specific problem to solve. People with a problem to solve – especially problems that have deadlines and budgets – refine the technology in the foundry of real-world struggle. Practitioners develop techniques, shortcuts, and workarounds that take advantage of the technology's strengths and that sidestep its weaknesses.

This is exactly what is happening in the world of predictive coding today. Practitioners of predictive coding, i.e., the lawyers, technologists, paralegals and others who use this technology every day, are developing a book of practical expertise on putting predictive coding to work.

And these practitioners have been on an accelerated path. Although the science behind predictive coding can be measured in decades, the first federal case in the US substantively addressing predictive coding only appeared in 2012. In an incredibly short period of time for the legal world, predictive coding has gone from being cautiously viewed as something we will do in some uncertain future (like fly in our cars) to something that most e-discovery practitioners are either using or evaluating. It also appears that e-discovery service providers and law firms are accelerating their

use of the technology. A 2012 survey by analyst firm eDJ Group found, for example, that only 33% of practitioners had used predictive coding (Q1 2013 Predictive coding Survey). One year later, the survey found that only about 20% were **not** using or planning to use predictive coding at all.

Will 2014 be the tipping point for predictive coding and its mainstream, systematic use by most firms? It seems likely. The same survey found that nearly 50% of practitioners already consider predictive coding to be a core part of their e-discovery workflow for certain kinds of matters.

So, what have practitioners learned about using predictive coding in this relatively short time? A lot. In this paper we share what we have learned from working with our clients, conducting research into predictive coding, and talking with other subject matter experts.

1. If You Are Just Getting Started with Predictive Coding, Pick the Right Case.

Although there is no universal "right" size or kind of case for predictive coding, the pure economic benefits of reducing review costs with predictive coding will be more pronounced in cases with greater volumes of information. Indeed, very large cases are often what drive many firms to look at predictive coding in the first place. Although overwhelming volume may be what drives firms to predictive coding initially, it is not what keeps them coming back. In fact, it is often the more strategic benefits that cement predictive coding as part of their e-discovery toolbox.

In theory, predictive coding can be used to support any case that involves discovery of a lot of text-based information. In reality, like other tools in a toolbox, predictive coding technology excels at solving certain kinds of problems. For example, traditional, manual, linear review processes may be more cost effective for cases with smaller pools of potentially responsive information. Some practitioners use 10,000 documents as a rule of thumb for the threshold at which predictive coding becomes more cost effective and efficient, but the absolute number is less important than the overall context of the case, which includes schedule, resources, budgets, and the richness of the corpus.

Carpenters do not start out building houses – they start by building birdhouses and stepstools, then work their way up to more complicated structures that require a blend of professional skill, experience, and creativity. Similarly, predictive coding practitioners are not born – they must learn their trade by practicing their trade. As

such, practitioners getting started with predictive coding should start with a matter that enables that learning process to happen productively.

For example, we have seen practitioners use predictive coding on matters that had already been resolved, which provided the incredibly valuable benefit of enabling them to validate and cross-check their process and results with the results of their traditional linear review process. This also helped them build support for predictive coding with clients and with attorneys inside the firm.

Other factors to consider when getting started with predictive coding include:

- **Timeline**. Choose a matter with a generous production timeline, so that practitioners have time to explore the capabilities and limitations of the product, and to try different techniques. The pressure of aggressive timelines will reduce the opportunities to learn and improve.
- **Tenor of the case**. Choose a matter involving an opponent who is relatively cooperative regarding discovery, and who is open to the use of predictive coding, rather than a matter where the minutia of each e-discovery decision will be scrutinized or opposed. This is not to suggest that the predictive coding process should be completed in anything other than a completely professional and defensible manner, but merely that a matter where discovery is a battleground may not be the best proving ground.
- **Complexity**. Matters with multiple issues and other complicating factors may not be the best choice. Simpler, more contained cases are more appropriate.

2. There is an Art to Predictive Coding. Master That Art.

> *"There is an art to this. This is not rocket science. Predictive analytics has been around for years, but it is relatively new to the science of information retrieval. There is an art to interpreting the results, just like a doctor interpreting the science."*
>
> *Tom Groom, VP and Sr. Discovery Engineer, D4 LLC*

When we are sitting in the doctor's office wearing only a backless green gown and our socks, we like to believe that our doctor is some kind of supercomputer who gathers the data on our symptoms, and then calculates a perfectly logical, binary diagnosis – the only possible diagnosis. This is not reality. Like diagnosing and treating medical conditions, there is both an art and a science involved in predictive coding.

The science part has been around for decades, but has been encoded and productized for e-discovery relatively recently. It is the practitioners – the predictive coding "doctors" in the emergency room of law firms and e-discovery providers across the country that now need to develop and apply their knowledge and experience to this well-established science and make it art.

The art of predictive coding lies in several critical areas, including:

- **Transparency**. Practitioners must learn the art of communicating effectively with the other side about how and where predictive coding is used. Some attorneys choose to focus on the e-discovery result rather than the process and communicate very little about whether or how they are using predictive coding. Others take the opposite approach and are very transparent with the other side throughout the process (i.e., in the spirit of the Sedona Cooperation Proclamation and similar trends). The art is in understanding which of the two is the right approach, at the right time. Litigants vary widely in their level of sophistication regarding predictive coding, and too much information can unnecessarily overwhelm, especially if they do not understand what to do with the information. On the other hand, a spirit of transparency with the other side regarding predictive coding can help to reduce the cost and complexity of the e-discovery process.
- **Translation**. Although predictive coding in practice can be quite straightforward for practitioners, the science and mathematics behind the technology are not. Predictive coding involves complex statistical concepts and algorithms that are difficult for even very sophisticated legal professionals – including judges – to understand. However the enervating question in e-discovery is quite simple: was the responsive evidence found and produced? Translating between that simple question and the complexity that underlies predictive coding is an art that practitioners need to master – an art that is also central to defensibility.
- **Technique**. Predictive coding technology is a tool. Tools only come alive and provide value when they are used as part of a technique. Like most tools used to accomplish something complex, there is no “right” technique for predictive coding. The art of predictive coding is developing and applying the right techniques at the right time. For example, should predictive coding be used on the entire corpus, or should it first be culled using keywords and other search techniques? It depends. Should a training set be developed randomly, or developed purposefully by a practitioner? It depends. Successfully navigating these and countless other decisions is at the heart of the art of predictive coding.
- **Technology**. The science behind predictive coding has been productized into many different software products, and not all of these products are

created equal. Evaluating and choosing the right product for the job is a critical art for e-discovery practitioners to master. There are significant objective differences among predictive coding products in terms of usability, efficiency, complexity, quality, and potential defensibility. Each product has its own strengths and weaknesses. In addition, not every product can support every use case effectively.

3. Your E-Discovery Team Must Evolve.

In large cases it used to be customary for e-discovery teams to work for months before substantially involving the lawyer in charge. In a world of predictive coding, this approach is no longer desirable. Predictive coding changes e-discovery strategy.

Predictive coding can help attorneys establish the relative merits of the other side's claims – as well as the strength of their own position – nearer the beginning of the e-discovery process, rather than after a months-long document review cycle. This capability should change the way that firms tackle e-discovery, and change the makeup of the teams involved in the e-discovery process. Specifically, it means that predictive coding teams should have much greater – and earlier – participation from a senior attorney who shares responsibility for litigation strategy and who understands the legal issues.

Predictive coding teams often find that, during the process of training the software for a specific matter, the software trains the team as much as the team trains the software. In other words, as the software starts to "understand" the types of information the team is looking for, it will begin to suggest topically relevant but unanticipated documents. With well-designed software and workflow, this is happening *before* document review has even begun. Having a senior attorney involved in the training process enables a faster and more organic process where he/she can react and respond to the responsive documents and issues, focus the team, and generate strategic insight weeks earlier than traditional approaches.

A representative from the client should also be on the team. This should be someone who is familiar with the organization and its relevant business activities and practices – someone who can guide the team on what the client views as relevant. Practitioners who are actually doing the coding and working with the system cannot work in a vacuum. Rather, they need to understand what the experts are looking for and what they care about.

This team should be in place right from the beginning of the predictive coding process, especially given how critical the training process is to a supervised learning

system. Spending the time upfront to get the training process right will pay off exponentially downstream in the form of better results gained more quickly. Building consensus and understanding across the team right from the beginning is essential.

4. Tailor Predictive Coding Techniques To Fit the Case

Experienced practitioners bring predictive coding technology into the e-discovery process at different stages, depending on the needs of the case. Learning how and when to employ the technology is a critical part of making predictive coding work. Insight that practitioners have developed in this area include:

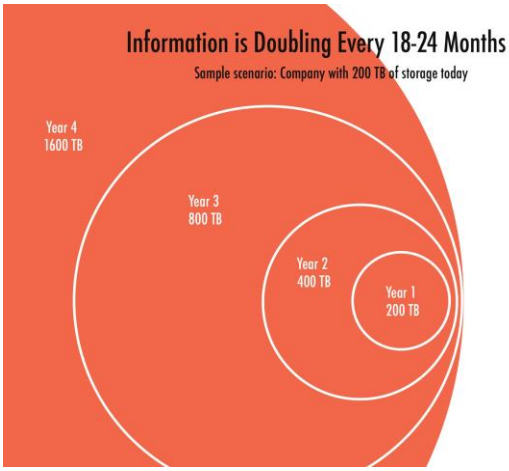
- **Culling**. Many practitioners are using predictive coding as a means to cull the potentially responsive document population prior to conducting attorney document review. In fact, a recent survey found that 80% of respondents were using predictive coding in this way (eDJ Group's Q1 2013 Predictive coding Survey). When executed well, and with the right technology, this use case can help to avoid problems associated with keyword-only culling (such as introducing bias, missing critical responsive documents, time to negotiate and execute, and so on). The right approach depends on a variety of case-specific factors, including the size of the document set, the relative richness of responsive information, and so on.
- **Prioritized review**. Predictive coding enables practitioners to conduct prioritized document review, where the document review process starts with the documents most likely to be responsive. This can generate critical insight and reveal strategic case issues much earlier in the review process. Prioritized review is a common way for practitioners to get started with predictive coding, as the review process is an internal practice that is rarely if ever examined or challenged by the court or the other side in the matter.
- **Fact development**. Some practitioners employ predictive coding to assess their own information early in a matter to develop the essential facts of the case and to develop an initial case strategy. Predictive coding enables this process to occur much earlier and more quickly than with traditional e-discovery processes.
- **Training**. Practitioners take a variety of approaches to training predictive coding software. Some begin with a random sample from the document set and start there. This approach has the advantage of being less susceptible to the introduction of bias, but tends to be inefficient and it may also fruitlessly ignore real insight that exists about the case. Some create a training set by assembling known responsive documents. Others use active learning, a capability supported by some predictive coding technologies in which the

system selects training samples for practitioners during the training process. Active learning melds human knowledge and statistical techniques in a way that drives efficiency and avoids the bias associated with practitioner-selected training sets. Practitioners should endeavor to select an approach that is reasonable, efficient, and defensible given the totality of the circumstances surrounding the matter.

5. Use Predictive Coding Techniques to Solve Upstream Business Problems

In the Sturm und Drang of litigation, it is easy to miss the big picture. Litigation is like firefighting in that the goal (if not the method) is simple: put out the fire. Bring all your tools, your energy, and your resources to bear on the problem, because the fire will continue to burn until you put it out. However, your client is not in the firefighting business. They produce auto parts, market medical devices, or manage money. As a business, if they even think about fires at all, they are mostly thinking about how to prevent them and how to contain and limit the damage if and when the next fire starts.

At the end of the e-discovery process, e-discovery practitioners often understand the organization better than it understands itself. E-discovery reveals the strengths and weaknesses of a client's existing information governance (IG) program. It can reveal, for example, that the organization has little idea what information it has, where that information is stored, or even if that information has business value.



Human based-classification and management methods alone do not solve the information governance problem for most organizations. There is simply too much information.

Today, we have the opportunity to apply predictive coding software to the information governance problem. In the e-discovery context, predictive coding helps us find the right documents, to separate wheat from chaff, in a highly automated and efficient manner. Information governance needs this capability. In fact, in large organizations it is difficult to see how information governance can be made real without this capability. As such, there

is a tremendous opportunity for predictive coding practitioners to bring their knowledge of predictive coding to bear on the information governance problem. Some examples of how predictive coding can support information governance include:

- Classification of **business records** for retention, knowledge management, privacy and other requirements.
- Identification of documents that are critical to a merger, joint venture, or **other major business activity.**
- Identification of documents that are **eligible for disposition** according to corporate retention policies.
- Finding and collecting content **that will generate value** through business analytics and other data-focused activities.

Conclusion

Predictive coding is a powerful tool, especially in the hands of experienced practitioners who understand the art and science of predictive coding. This entails both a practical understanding that enables efficient and defensible workflows, as well as a strategic understanding that informs case strategy. Key insights that practitioners have gained in putting predictive coding to work include:

- Pick the right case to get started with predictive coding.
- Not all predictive coding software is created equal. Choose the right tool for the job, i.e., the software that will deliver high-quality and defensible outcomes for the use cases you care about.
- Senior attorneys who understand the legal issues of a matter should be involved earlier in the e-discovery process when predictive coding is being used.
- Where and how predictive coding is used in a case should be tailored to the needs of the case.
- Predictive coding tools and techniques can also help to solve the difficult information governance problem, particularly when practitioners bring what they learn about a client's information environment to the table.

Endnotes

This work should be cited as: Barclay T. Blair, “Predictive Coding: Making It Work,” December 2013, ViaLumina LLC.