Predictive Coding: What You Need To Know Now

May 14, 2013

By: Karl Schieneman & Barry Murphy

eDiscoveryJournal

Table of Contents

Predictive Coding is a type of Technology-Assisted Review (TAR) that propagates known decisions about a sample of documents (e.g. responsive, privileged) to the rest of the documents in a corpus.

This eDiscoveryJournal research brief explores the survey results and market research into Predictive Coding and its impact on eDiscovery and Information Governance practices. The brief is aimed at eDiscovery professionals seeking to understand how Predictive Coding works in the review process and what to consider before making decisions on which Predictive Coding solutions to utilize. Specifically, readers of this report will get:

- eDJ's analysis of our February 2013 Predictive Coding Survey
  - Predictive Coding Adoption (and comparison to adoption versus our 2012 survey)
  - Reasons To *Not Use* Predictive Coding
  - How Users Leverage Predictive Coding When Using It
  - How Users Source Predictive Coding Solutions
- A Framework for Analyzing Predictive Coding Solutions
  - Defensibility and Transparency
  - User Experience and Workflow Support
  - Platform Support and Architecture
  - Pricing

Future research reports in this series will focus on taking the covers off Predictive Coding and how to validate the results of Predictive Coding.

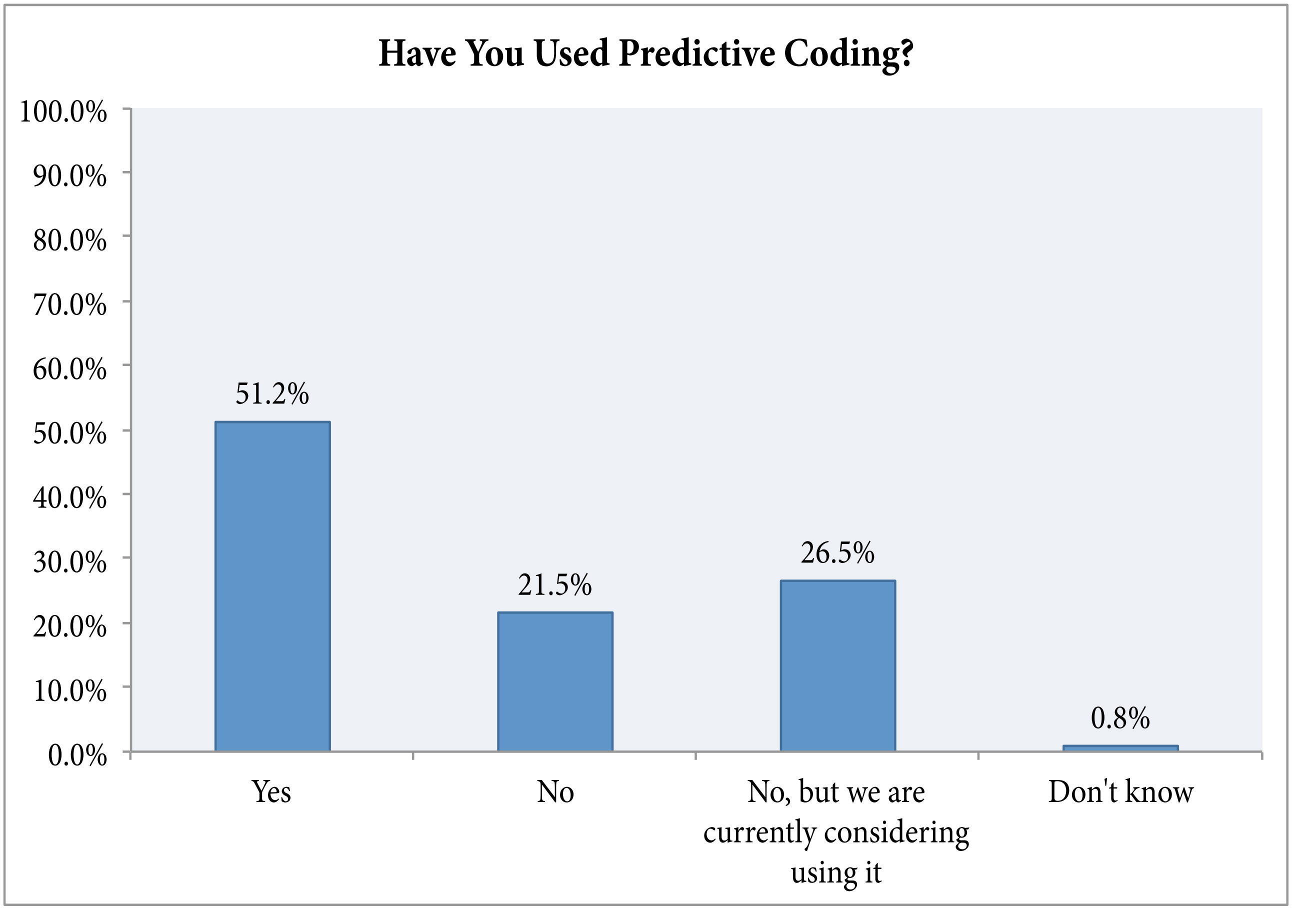
Predictive Coding Has Applications With Both Near- and Long-Term Impact

The benefits of Predictive Coding (PC) technology are limitless. PC has proven to make Legal Review simultaneously more accurate and less expensive[1]. In the litigation context, Predictive Coding is a form of Technology-Assisted Review (TAR) that leverages advanced analytics technologies, algorithms, and/or machine learning to augment human legal reviewers' knowledge. Considering that review makes up the bulk of total eDiscovery costs, PC is hugely impactful just within that one small niche. Beyond Legal Review, PC promises to improve information governance (IG) activities through more effective and automated records and information classification, better defensible deletion projects, and the ability to address the real challenge of Big Data – analyzing unstructured content in a fast, efficient way.

Use of PC for IG projects – defensible deletion, automated information classification – is not even in the anecdotal stages yet. There is a ton of interest in how PC can improve or kick-start IG projects, but precious little activity on the ground. This is not surprising given that PC just for Legal Review is still in the very early stages of adoption. Recent eDiscoveryJournal research shows that just about half of respondents have used Predictive Coding.

[1] Cormac, Gordon and Grossman, Maura. Technology-Assisted Review in E-Discovery Can Be More Effective and Efficient Than Exhaustive and Manual Review, XVII Rich. J.L. & Tech. 11 (2011), http://jolt.richmond.edu/v1713/articlee11.pdf

Predictive Coding Usage In Legal Review Is Picking Up



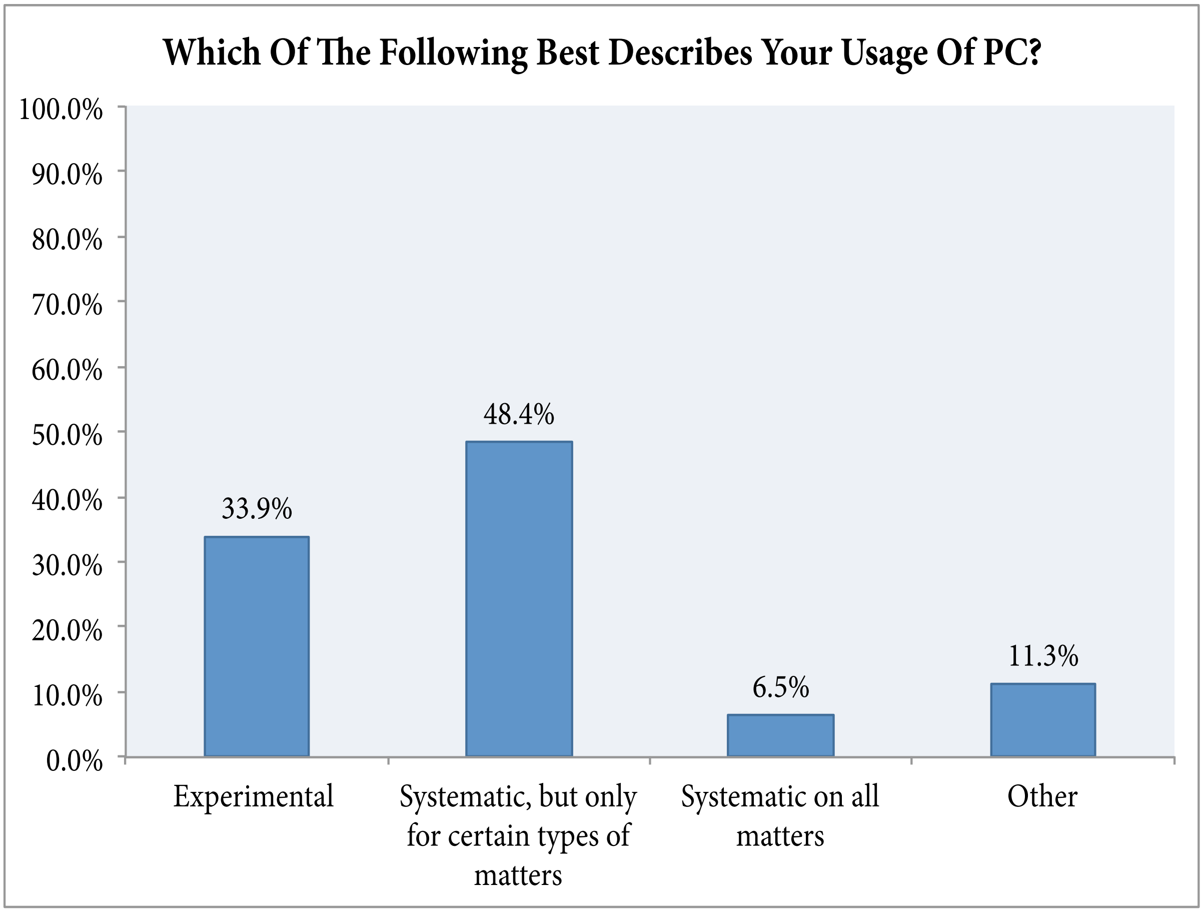
Source: eDJ Group's Q1 2013 Predictive Coding Survey, February 2013, N = 121

With only about 20% of respondents not using or planning to use PC, it is clear that this technology is here to stay. Some might argue that its usage is mainstream, but in reality, PC usage is just starting to scratch the surface, in the Legal Review / Litigation use-case. While total PC use percentages are large, the numbers would likely drop substantially if the number of cases handled by a firm were compared to the number of cases using PC. Conversely, we expect rapid growth in the number of cases utilizing PC once corporations push for more use or Law Firm early adopters fully embrace the use of this technology. Anecdotally, we see the emergence of both in the marketplace.

In the litigation context, PC can help to more automatically, cost-effectively, and efficiently determine the responsiveness or privileged nature of a document collected as part of a specific matter. As eDiscoveryJournal's survey data shows, more and more respondents are using PC; in a similar survey conducted in 2012, only 33% of respondents had used PC. This growth in PC usage is due at least partly to massive data growth in general. Combine that with the ability to do in-place eDiscovery and organizations face larger collections. With very tight deadlines for conducting eDiscovery, there is a need to smarter review, not harder review.

As adoption of PC grows, usage of it is beginning to mature at least somewhat. Whereas in 2012, a majority of PC usage was experimental, in 2013 almost half of respondents report a more systematic approach to PC, at least for some matters. The PC technologies continue to evolve and mature, but workflow and user education remain fairly immature – not surprising for a relatively new market.

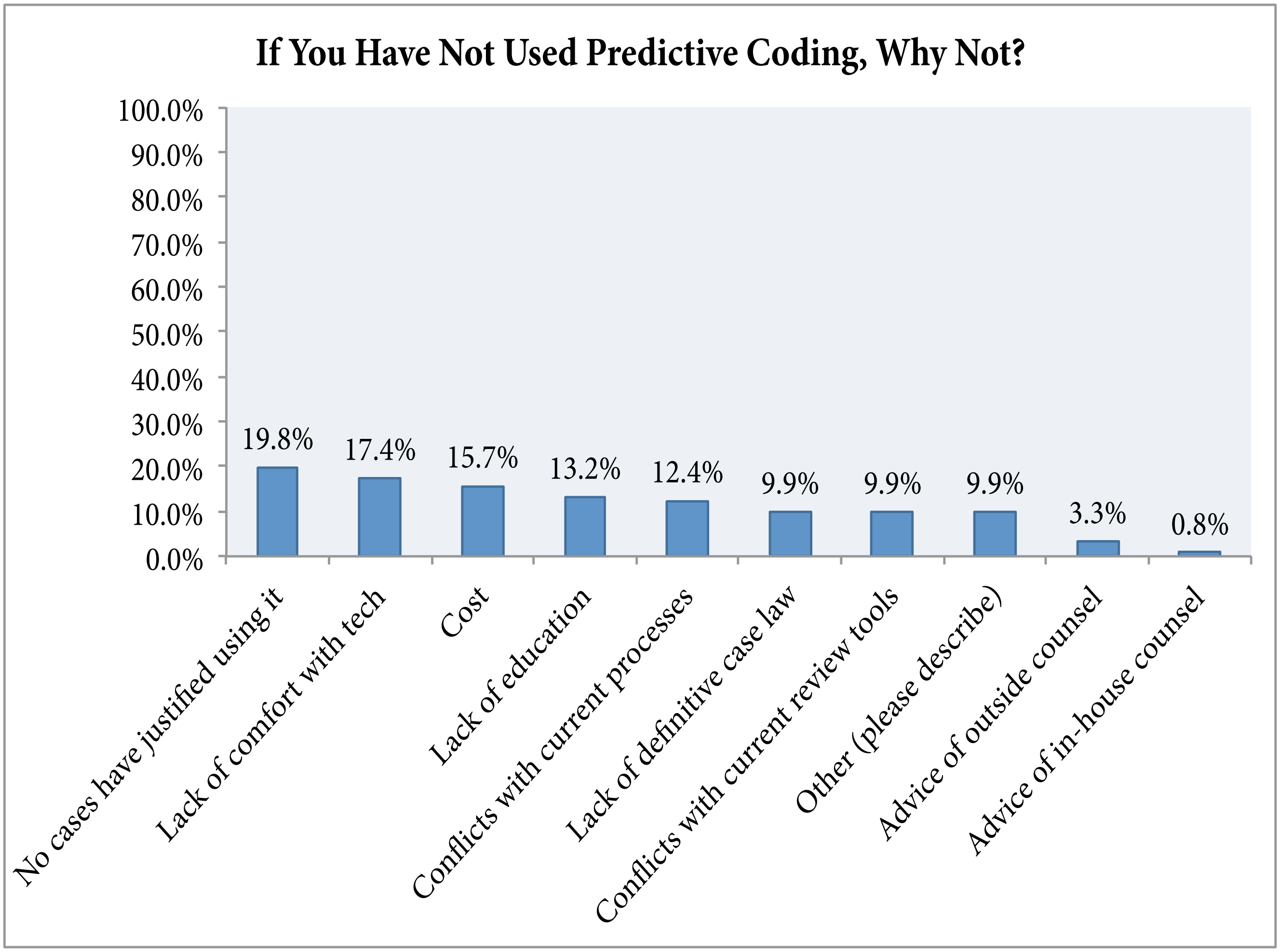
Predictive Coding Usage Evolving From Experimental To Systematic



Source: eDJ Group's Q1 2013 Predictive Coding Survey, February 2013, N = 121

Usage and maturity are growing, so those that do not get onboard and start making use of PC for the litigation use-case are in danger of being left behind. For survey respondents that report not having used Predictive Coding, there is not one dominant reason that keeps them away.

There Are No Dominant Reasons *Not* To Use Predictive Coding



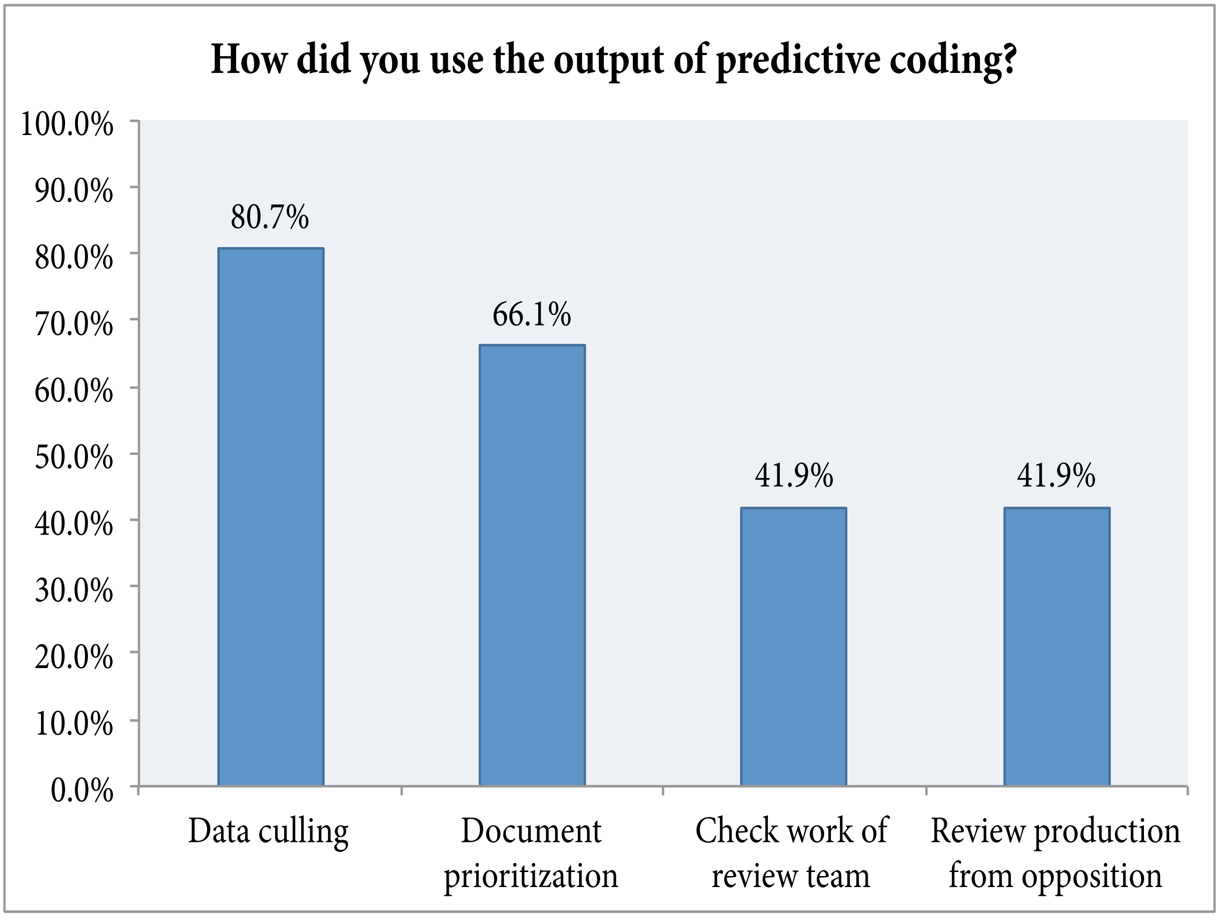
Source: eDJ Group's Q1 2013 Predictive Coding Survey, February 2013, N = 66

Certainly, there is a lack of comfort with technology and a lack of education about how to use PC that can get in the way of adoption, but those reasons will go away over time as PC usage becomes mainstream and lawyers become more accustomed to industry norms and best practices. Already there are numerous educational offerings springing up that can get lawyers more comfortable with PC processes, best practices, and technologies including a successful vendor-neutral boot camp series being offered by eDJ and Review Less to grow the pool of predictive coding users in each market and to highlight judicial receptivity to the use of these tools. That will serve to keep PC on the track to mainstream adoption and, as a result, all Legal teams – whether law firm or corporate – should make getting up to speed on PC a priority sooner than later. If that is not enough of a reason, consider that more and more cases involve decisions on the use of PC and two of note - *Da Silva Moore v. Publicis Groupe* and *Global Aerospace v. Landow Aviation* - featured Judge-approved usage of PC protocols.

The Simple Goal of Predictive Coding

In all the analysis of case law around PC and arguments over whether its usage will take off or not, it is easy to overlook the astoundingly straightforward benefit that PC provides: the ability to better cull data and create a validation record for the decisions made based on statistical sampling. It is a simplification to say that PC is just another culling tool, but in reality that is one of the main uses of PC at present. Yes, PC can and will offer additional benefits as use matures – for example, the ability to set case strategy more quickly and efficiently – but, right now, PC is primarily used as a culling tool and a good one at that when speed, cost and accuracy are considered.

Predictive Coding Is Currently Used Primarily As A Culling Tool

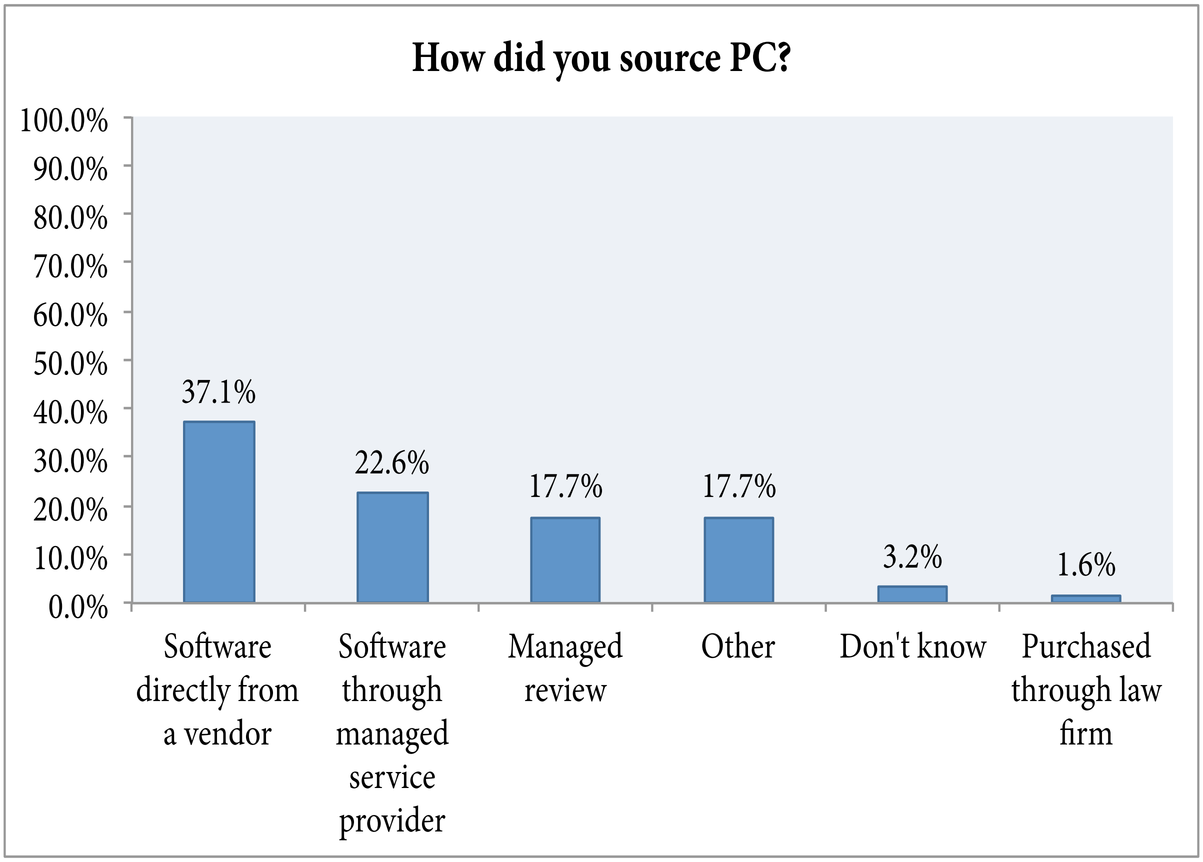


Source: eDJ Group's Q1 2013 Predictive Coding Survey, February 2013, N = 62

Analyzing Predictive Coding Alternatives

Gaining experience with PC requires purchasing a solution. While PC technology and its underlying components – machine learning, advanced clustering algorithms and analytics – have been around for quite some time, PC applications are still relatively new. As such, there does exist some confusion over the best way to consume a PC solution. There is no shortage of options; organizations can buy on-premise software solutions, software-as-a-Service (SaaS) solutions direct from software providers, SaaS solutions from service providers, or managed services. The best way to deploy a PC solution depends on the situation. A highly litigious corporation with a large Legal team and sufficient IT resources may want to purchase on-premise software or SaaS directly from a software provider whereas a Law Firm with limited IT resources may want to purchase SaaS through a trusted service provider. According to our survey research, there is not yet a dominant PC purchase method.

There Is No One Dominant Method Of Sourcing Predictive Coding Solutions



Source: eDJ Group's Q1 2013 Predictive Coding Survey, February 2013, N = 62

When sourcing a PC solution, buyers should consider how the solution meets criteria across four broad categories:

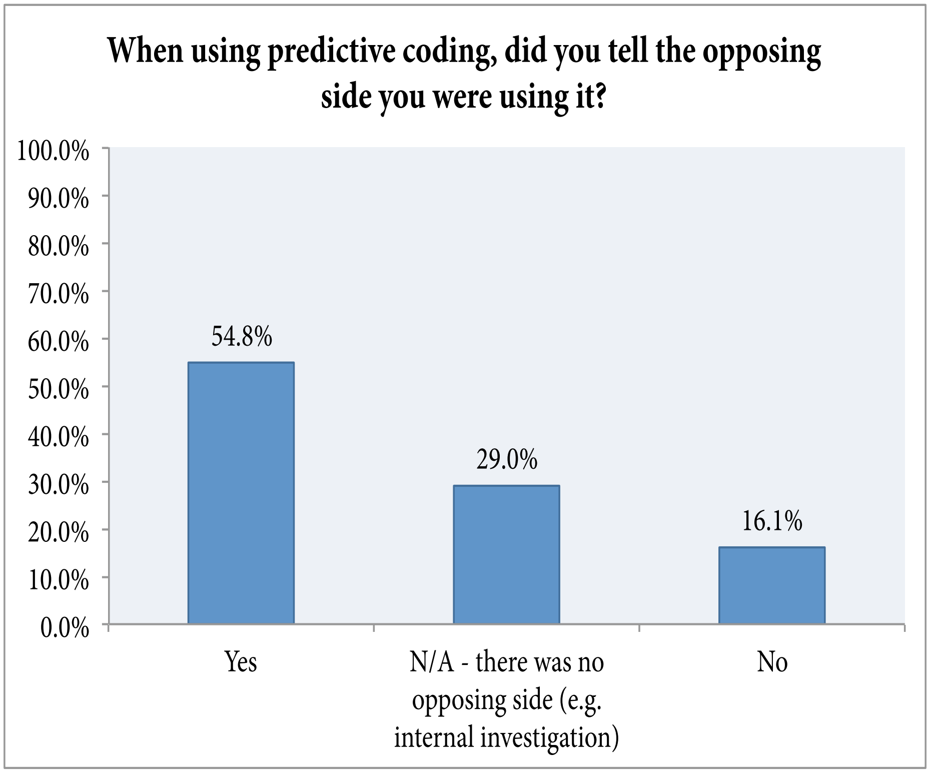
- **Defensibility**. How do you explain usage of PC, get agreement on the process/solutions, and/or validate the results of PC?
- **User experience**. Is the solution easy to use and learn? Does the solution support your primary litigation goals, e.g. identification of mass culling/filters, applying PC to raw collection or already used search criteria to raise richness? Does the solution support the workflows you need? Such workflows include:
  - Recommendation (predict relevance and/or privilege)
  - Decision expansion (similarity/cluster)
  - Profile extraction (group characteristics for decisions)
  - Analyze incoming production
  - Document prioritization

- **Platform support**. Does the solution fit into (or need to fit into) existing IT infrastructure(s)?
- **Pricing**. Does the solution fit into various budgets, e.g. Litigation, IT? Does the solution make economic sense for a given matter?

Defensibility

For PC to work effectively, its results must be defensible. Users must be able to statistically prove that the process and technology worked in a reasonable way. In the litigation use-case, defensibility must be established on a case-by-case basis. The reality is that different tools operate differently with different operators and different data collections. There is no trial speedway on which to test these tools and measure performance. However, there are always sampling protocols, which can provide snapshot comparisons with confidence levels and margins of error[2] to give users comfort in the PC output. While different tools offer different metrics for measuring defensibility, just about all have precision and recall[3] covered as statistics to gauge the process based on sampling which can provide a form of uniformity across tools the industry is looking for to speed up acceptance. eDiscovery professionals take note, however: such statistical tools work best when there is cooperation and transparency between litigation opponents. Examples of this cooperation do exist and are becoming more common, as our survey results show.

Cooperation Becoming More Common When Predictive Coding In Play



Source: eDJ Group's Q1 2013 Predictive Coding Survey, February 2013, N = 62

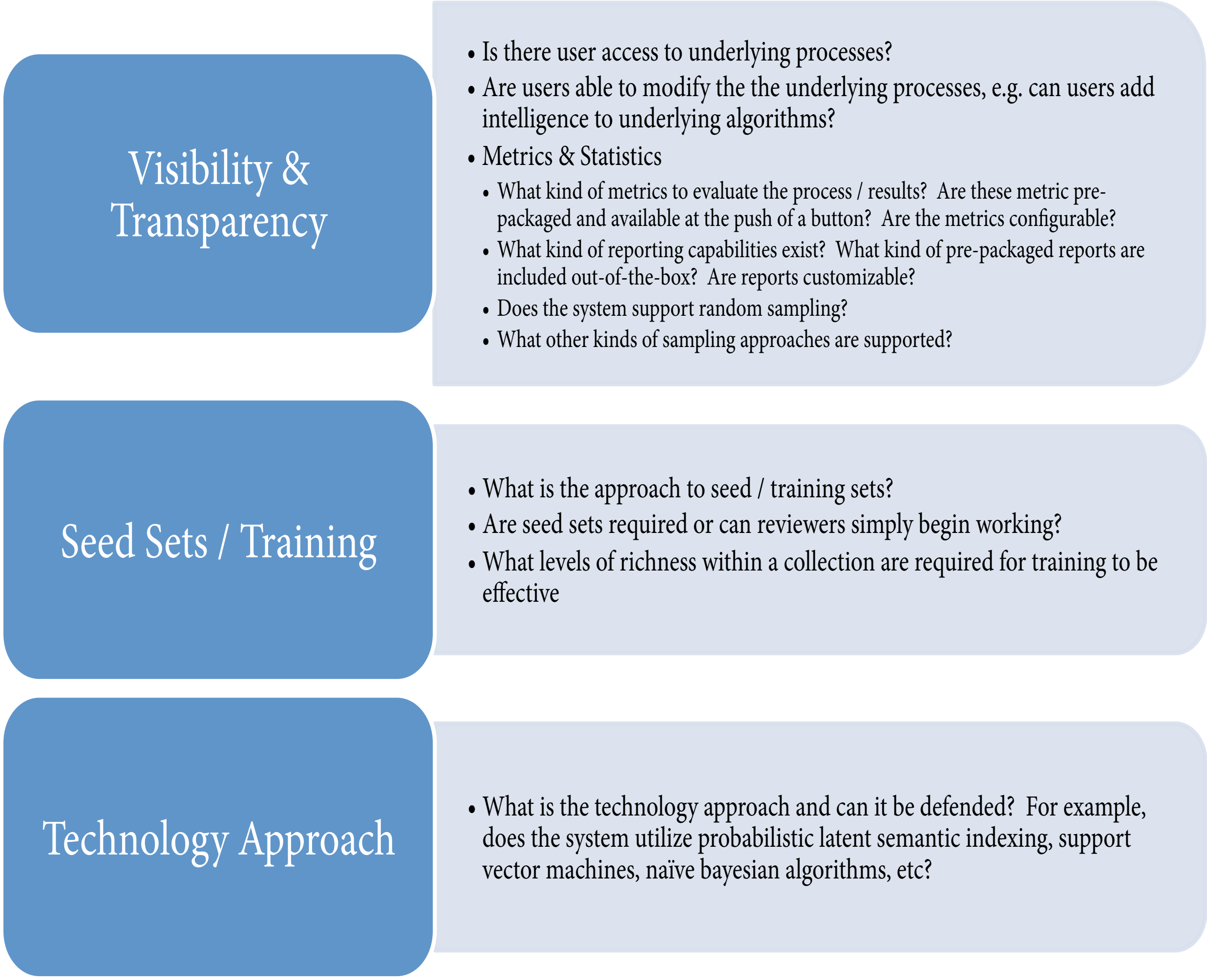
[2] Confidence level is a random interval constructed from data in such a way that the probability that the interval contains the true value of the parameter can be specified before the data are collected; margin of error is A measure of the uncertainty in an estimate of a parameter. (Source: UC Berkely Department of Statistics, http://www.stat.berkeley.edu/~stark/SticiGui/Text/gloss.htm)

[3] Precision is a measure of the ability of a system to present only relevant items; recall is a measure of the ability of a system to present all relevant items. (Source: TREC, http://trec.nist.gov/pubs/trec10/appendices/measures.pdf)

Defensibility takes on a slightly different meaning for the IG use-case. In litigation, the precision and recall focus relate to knowing what you are looking for, e.g. what is potentially responsive and/or privileged. With IG, the goal is different: to to keep only the information an organization needs for business purposes or litigation and/or regulatory obligations. For now, it is unclear what metrics will be best to defend that task. Anecdotally, however, Legal contacts tells us that it is unlikely that IG policies will be challenged in court because they are outside the context of a given matter. It is more likely that challenges would come in the preservation and Litigation use-case only and look at the use of PC in a specific matter on data already collected.

The following charts depicts some of the key considerations for eDiscovery professionals when evaluating how a PC solution will help with measuring and proving defensibility:

Determining How A PC Solution Supports Defensibility



Visibility & Transparency

- Is there user access to underlying processes?
- Are users able to modify the the underlying processes, e.g. can users add intelligence to underlying algorithms?
- Metrics & Statistics
    - What kind of metrics to evaluate the process / results? Are these metric pre-packaged and available at the push of a button? Are the metrics configurable?
    - What kind of reporting capabilities exist? What kind of pre-packaged reports are included out-of-the-box? Are reports customizable?
    - Does the system support random sampling?
    - What other kinds of sampling approaches are supported?

Seed Sets / Training

- What is the approach to seed / training sets?
- Are seed sets required or can reviewers simply begin working?
- What levels of richness within a collection are required for training to be effective

Technology Approach

- What is the technology approach and can it be defended? For example, does the system utilize probabilistic latent semantic indexing, support vector machines, naïve bayesian algorithms, etc?

In addition to these questions, buyers should also consider what level of advanced support a PC solution comes with. Because PC is still relatively new, users will need project management support and help with metrics, measurement, and workflow issues. Such support will be critical to making usage of PC effective. In addition, buyers should look for the availability of expert witnesses for defending results of the tool. Many organizations will not have the wherewithal to find internal experts to play that role. It is critical that any users of PC be able to document how the system was used and explain to both litigation opponents and the Court why the results are valid.

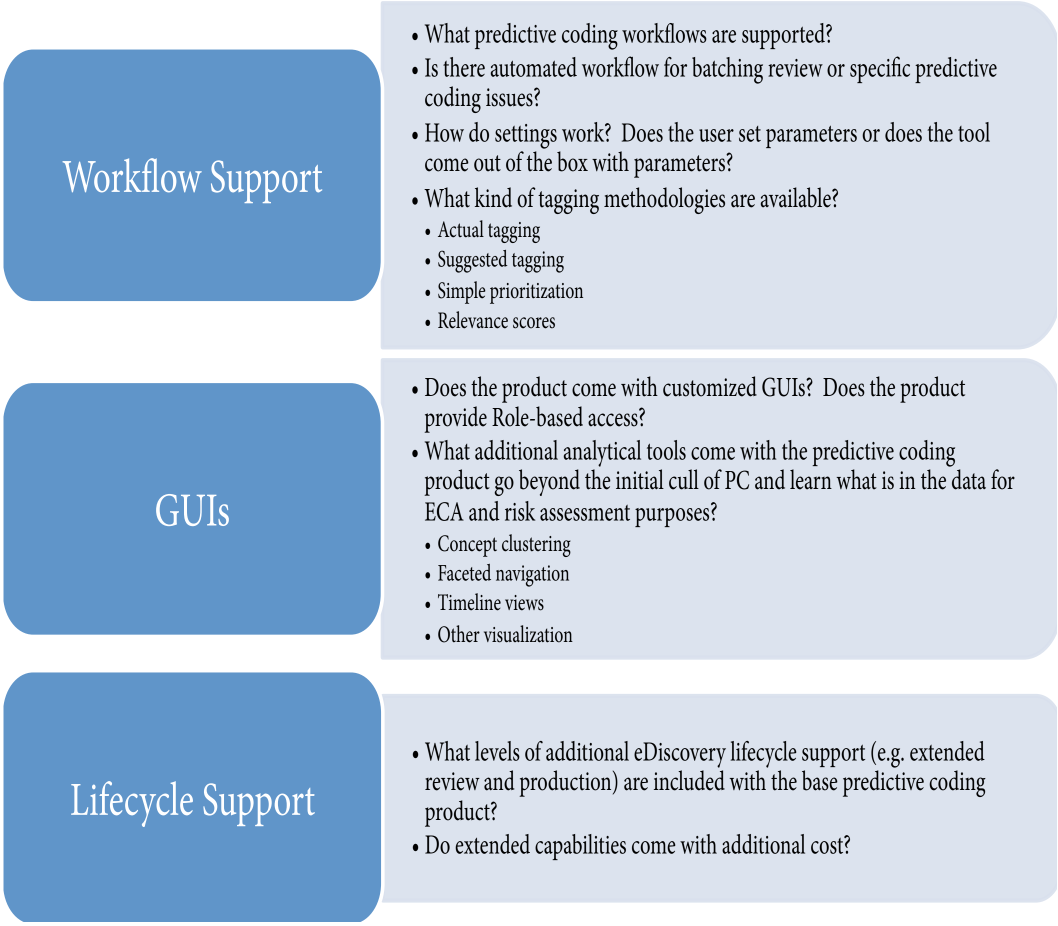
In addition, users should consider the scalability of the PC offering given the size of both many cases and the IG problem. For litigation, the Federal Rules of Civil Procedure support proportionality arguments which have been used successfully to reduce the size of data which is ingested into a PC tool. However, for IG there is no built in filter. Some organizations who have over-retained data may seek to evaluate the scalability of the tool and what types of IG workflows are offered to tackle the IG problem of saving too much old data which serves no business purpose. In that case, PC may be applied to ESI in-place (within the source repository) as opposed to in a collection for a given matter.

User Experience

A PC solution supports a human – in the litigation use-case, a Legal reviewer – conducting a process. To that end, it must help the Legal reviewer be as effective and efficient as possible in conducting review. The application should be easy to learn, simple, and intuitive, even if the process itself and accompanying need for defensibility can seem complicated. PC solutions should also be flexible, though, as users may be trying to accomplish different things in different cases. Buyers should be aware that there is going to be a trade-off between simplicity and flexibility. The solutions that are less customizable also tend to offer the best graphical use interface (GUI), with less clutter and more simplicity. The trade-off is that the simple GUI's workflow might not be the right one for a given case.

The following chart depicts some of the key considerations for eDiscovery professionals when evaluating the user experience of a PC solution:

Evaluating The User Experience Of A PC Solution



Workflow Support

- What predictive coding workflows are supported?
- Is there automated workflow for batching review or specific predictive coding issues?
- How do settings work? Does the user set parameters or does the tool come out of the box with parameters?
- What kind of tagging methodologies are available?
  - Actual tagging
  - Suggested tagging
  - Simple prioritization
  - Relevance scores

GUIs

- Does the product come with customized GUIs? Does the product provide Role-based access?
- What additional analytical tools come with the predictive coding product go beyond the initial cull of PC and learn what is in the data for ECA and risk assessment purposes?
  - Concept clustering
  - Faceted navigation
  - Timeline views
  - Other visualization

Lifecycle Support

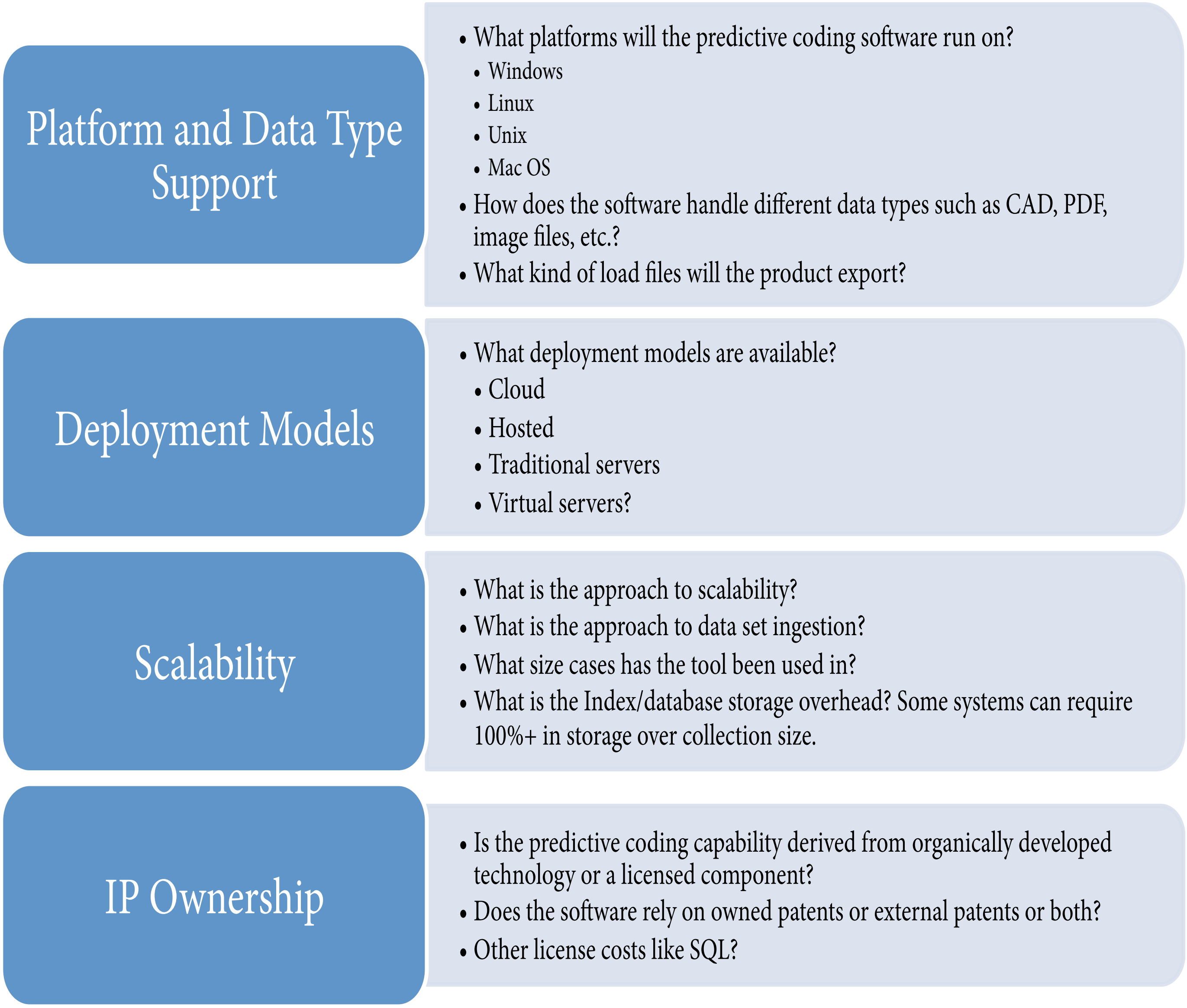
- What levels of additional eDiscovery lifecycle support (e.g. extended review and production) are included with the base predictive coding product?
- Do extended capabilities come with additional cost?

Platform Support And Architecture

For those looking to bring PC technology in-house or ensure that a given PC solution will work their environment and/or data, it is important to examine the platform support a PC solution provides and to understand its technology architecture. This can be especially important when making a strategic investment. After all, PC solutions are rooted in technology and how they are built will matter greatly in both the efficacy and value of the solution.

The following chart depicts some of the key considerations for eDiscovery professionals when evaluating the technological components of a PC solution:

Evaluating The Technological Components Of A PC Solution



Platform and Data Type Support

- What platforms will the predictive coding software run on?
  - Windows
  - Linux
  - Unix
  - Mac OS
- How does the software handle different data types such as CAD, PDF, image files, etc.?
- What kind of load files will the product export?

Deployment Models

- What deployment models are available?
  - Cloud
  - Hosted
  - Traditional servers
  - Virtual servers?

Scalability

- What is the approach to scalability?
- What is the approach to data set ingestion?
- What size cases has the tool been used in?
- What is the Index/database storage overhead? Some systems can require 100%+ in storage over collection size.

IP Ownership

- Is the predictive coding capability derived from organically developed technology or a licensed component?
- Does the software rely on owned patents or external patents or both?
- Other license costs like SQL?

Pricing

For any buyer, the price of a solution will be a very important factor in the buying decision. There are myriad pricing models for PC that exist:

- All you can ingest (enterprise)
- Project-based (Per case)
- Per CPU core
- Per user
- Subscription
- Term license
- Volume-In based
- Volume-Out based
- Per Item

Each buyer will have to consider its own unique circumstances to determine the pricing model that provides the best return on investment. For example, many litigious corporations can justify spending heavily on PC because of a desire to reduce Legal Review costs; an expensive "all you can ingest" license for PC software can thus make sense because the cost of the software can be spread across multiple cases. Likewise, a smaller company without much litigation may only consider PC on a case-by-case basis and therefore prefer subscription or volume-based pricing to avoid over-spending and under-utilizing the solution.

We do expect that the competitive situation in the PC solution market will create downward pricing pressures. Many eDiscovery tools with upstream lifecycle support – archiving, processing, and preservation – have added PC functionality and offer that functionality at no additional cost (above and beyond the license for the tool). While PC is not free in this scenario, it is likely to be marketed that way and give buyers pause for thought when comparing such a solution with a stand-alone PC solution. The solutions that include PC functionality in a broader eDiscovery platform will also have a leg up in addressing the IG use-case when that becomes more mainstream in the next 18-24 months.

This report provides a framework for evaluating PC solutions most specifically for the litigation use-case. The next reports in this series will focus on validating PC results and a deeper look under the covers of the PC technology.

About The eDJ Group

eDJ Group offers unbiased information and pragmatic advice, based on years of experience and proven industry best practices. Whether researching a technology or service solution, conducting an eDiscovery Bootcamp or finding the right expertise to answer your specific questions, eDJ Group is the source for all eDiscovery professionals.

We are committed to helping eDiscovery professionals get the information necessary to excel in their professions, rather than offering legal advice or counsel. We operate with the utmost integrity and commitment to our clients on these guiding principles:

- Independence – All research, reports, advice and services are agnostic and conducted independently without influence by sponsors.
- Highest Ethical standards – All content is honest perspective based on real experience and interactions with thousands of practitioners; detailing both successes and failures without favoritism.
- Pragmatic, Experienced Expertise – All services are conducted by industry experts with decades of experience

in eDiscovery and strictly vetted by the eDJ Group founders.

For further information about the eDJ Group and their research, please contact Barry Murphy (barry@edjgroupinc.com) or Jason Velasco (jason@edjgroupinc.com).