ICAIL 2013 Workshop on Standards for Using Predictive Coding, Machine Learning and Other Advanced Search and Review Methods in E-Discovery
(DESI V Workshop)

June 14, 2013, *Consiglio Nazionale delle Ricerche*, Rome, Italy

The Challenge and Promise of Predictive Coding for Privilege

Manfred Gabriel
Principal
Forensic Technology
KPMG LLP (US)
345 Park Avenue
New York, NY 10154
USA
+1 (212) 954-3656
mjgabriel@kpmg.com

Chris Paskach
Partner
Forensic Technology
KPMG LLP (US)
6032 Katella Avenue
Cypress, CA 90630
USA
+1 (714) 934-5442
cpaskach@kpmg.com

David Sharpe
Manager
E-Discovery Services
KPMG LLP (Canada)
333 Bay Street, Suite 4600
Toronto, ON M5H 2S1
Canada
+1 (416) 777-3738
davidsharpe@kpmg.ca

1. INTRODUCTION[*]

The stakes in privilege review in eDiscovery are high. Given that privilege review is often the most expensive part of document review, there are significant economic implications. Successful privilege prediction through advanced text analytics could reduce the volume of attorney reviewed documents by almost 70%.

Current eDiscovery practice spends significant resources to find relevant and responsive documents from the ever-expanding populations of electronically-stored information ("ESI"). Attorney review of documents makes up nearly three-quarters of that cost.[1] Because of the significant risk presented by inadvertent production of critical privileged documents, a large portion of the attorney review cost is spent to identify privileged documents. The chosen methods and sequence of document review processes or workflows can have a significant impact on the cost, timing, and risk of producing ESI.

This paper describes the challenges of finding privileged information in documents during eDiscovery review, and the methods, and their sequencing to optimize the balance between review cost and risk reduction for a particular matter—including the use of predictive coding and sampling technologies.

Based on our analysis, the most productive approach to performing privilege review follows three phases:

1. Identification of the producible document set with predictive analysis;
2. Identification of privileged content in the producible set with predictive text analysis; and
3. Confirmation of the limited privileged distribution of the documents – utilizing attorney review and metadata analysis, as well as the use of sampling to test the non-privileged produced documents to avoid inadvertent inclusion of privileged documents.

We further contend that the use of next-generation computer-assisted review technologies can be critical to improving the speed and accuracy of identifying privileged information, using tools such as statistical sampling and advanced text analytics that use both metadata and rules-based classification.

Studies have long suggested that human review is highly error-prone. Recall rates for responsiveness of 50% have been reported,[2] although it is not clear that such rates are typical for human review. For privilege, such error rates are unacceptable and counsel today relies on Boolean searches, advanced analytics, and multiple review passes to minimize the risk of inadvertent disclosure of privileged information. For responsiveness-review, predictive coding already offers benefits of increased consistency, improved recall, and cost savings. If successful prediction for privilege can be achieved, the risk of inadvertent disclosure will be reduced in workflows combining predictive coding with accepted practices, as well as enabling potentially significant cost savings.

We are aware of one published test of predictive-coding technology to identify privileged information in eDiscovery, the 2010 TREC Legal Track. Interactive task 304 was introduced to test whether participants were able to "identify any and all

[*] This article represents the views of the authors only, and does not necessarily represent the views or professional advice of KPMG LLP. The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act upon such information without appropriate professional advice after a thorough examination of the particular situation.

documents or communications in the TREC Legal Track Enron Collection that are subject to a claim of privilege, attorney work-product, or other applicable protection, regardless of whether they are responsive…"[3] The tests took a very broad view of what constituted "potentially privileged" documents.[4] The two runs that scored the highest achieved 63.3% to 71.5% recall, levels that will be insufficient in practice to protect against inadvertent disclosure.

2. PRIVILEGE IN EDISCOVERY

During legal document review, as part of the U.S. system of electronic discovery (or similar rules in other jurisdictions, e.g. e-disclosure in the U.K.), counsel pursues three goals:

2.1 Disposition

First, each document that is being considered must be determined to be responsive or non-responsive to the eDiscovery request. Responsive documents will typically be produced. Family members of responsive documents, that is, cover email or attachments, will be included.

2.2 Knowledge Construction

Second, the construction of knowledge: eDiscovery is, like all discovery, designed to develop the facts of the dispute and to streamline civil procedure by revealing the incontrovertible facts contained in the objective documents relating to the case. The challenge therefore is to find the small set of documents (among the population of potentially millions) that is not only responsive but truly interesting.

2.3 Protection of Information, Including Privilege

Finally, document review in eDiscovery must protect certain information from disclosure. The most obvious example is privileged information, specifically the attorney-client privilege recognized under U.S. law. There is also the work-product doctrine, which protects materials prepared in anticipation of litigation from discovery. Additional types of information may be protected and must be identified: PII (personally identifiable information, PHI (protected health information), confidential documents, trade secrets, source code, etc. The specific type of information that requires protection from disclosure will vary from matter to matter.

The remainder of this paper will focus on the third goal (protection of information) and on the challenge of identifying information subject to the attorney-client privilege and the work-product doctrine ("privileged" information). We discuss traditional techniques as well as predictive-coding strategies used to identify privileged information. The goals of disposition and knowledge construction still need to be achieved. Will the next-generation predictive-coding software permit successful prediction for privilege, if used in the context of an appropriate workflow, and what will be the economic impact?

3. THE CHALLENGES OF IDENTIFYING PRIVILEGED INFORMATION

Greater automation in legal services is seen as inevitable. In fact, experts anticipate a "1% crisis," meaning that eventually, even after applying automated keyword searching to eliminate 99% of a dataset, the remaining one percent of large datasets will still be too large for reasonably cost-effective manual review.[5] The need for greater automation, however, comes with its own difficulties—particularly in the case of privilege review. In addition to the human challenge of knowing a privileged document[6] text when one sees it, there are specific challenges to finding privileged documents using predictive coding:

- Designations as "privileged and confidential" themselves are not a definitive basis for a claim of privilege.
- The determination whether legal advice was sought or rendered may be nuanced and subtle.
- Unlike for most other review tasks, the sender and recipient of the communication may help to determine whether privilege existed (disclosure to a third party will waive privilege, legal advice rendered by or sought from non-attorneys is not privileged).
- Joint-Defense agreements may preserve privilege despite disclosure to a party that is not the counsel for the disclosing party.
- Facts not apparent in the text of the document may affect privilege. For example, privilege can be waived for *all* copies of a document even if only one copy was disclosed to a third party. This waiver may extend to the subject matter of the disclosed document, and thus all documents dealing with the same subject matter.
- Communications between non-attorneys may be privileged. For example, if communications are between employees of the attorney's client and the communication conveys legal advice rendered to the client, the communication may be privileged.
- While under U.S. rules in-house lawyers can render legal advice and their communications may be subject to privilege, their specific role in the communication must be considered. A specific lawyer who is employed by the client may act in their role as in-house counsel (making privilege available) or in a business role (making privilege unavailable).

Detail on US privilege rules is provided in Appendix I.

Privileged documents tend to occur less frequently in review populations, as an analysis of document repositories showed. [7] The analyzed repositories represent both litigation and investigation matters, comprising over 13 million reviewed documents. Only 2.7% of the total reviewed documents were marked privileged. This corresponds to 11.7% of all documents marked "Responsive." Anecdotal evidence suggests prevalence of privilege may be much lower in some cases, as low as 0.5% of all documents. Such low richness has implications for predictive coding workflows. Sample sizes will need to be increased, potentially significantly, to estimate privilege distribution and to attain sufficient numbers of privileged training documents to train a classifier.

Finally, given the high stakes of privilege review, recall levels typically considered acceptable for predictions on responsiveness will be too low for privilege purposes. (See Appendix I for more detail on the stakes in privilege review.) Since the inadvertent disclosure of privileged documents may lead to the waiver of privilege for the entire subject matter, recall levels of 95-99% will likely be desirable.[8] Such recall levels tend to have an adverse impact on precision, so that significant numbers of documents

ultimately not deemed privileged will be swept up for manual review.

4. WHAT WILL THE ECONOMIC IMPACT OF SUCCESSFUL PRIVILEGE PREDICTIONS BE?

4.1 Change in Workflow

Currently, because no predictive coding solution has proven fully effective for privilege classification, counsel tends to use predictive coding technology to cull non-responsive documents. The remaining responsive population is still subjected to human review, for two purposes: to identify privilege, and to understand the substance of the documents (the construction of knowledge). Thus a significant reviewable population remains. The human review is augmented and supported through the use of a set of tools:

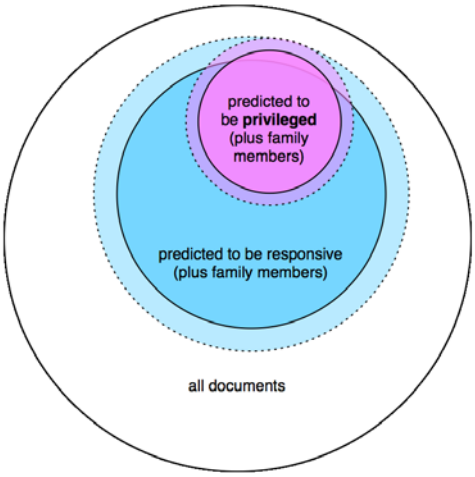
- search terms indicative of privilege ("law, legal, investigation, court, privileged, confidential, complaint, plaintiff, defendant, regulatory, attorney, draft" and so forth);
- search terms for names of attorneys, law firms, law firm URLs, etc.;
- iteration of search terms and continual expansion of search terms as additional attorney names or case-specific indications of privilege are found;
- advanced analytics tools, such as concept clustering, visualization, similarity analysis and hash-based de-duplication; thread analysis for email, as well as social-network data;
- multiple layers of quality control using full re-review, judgmental and statistical sampling, error analysis, and continual reviewer training;

In the context of review for privilege, there are now two approaches to integrating predictive coding:

First, predictive coding can be used as part of the quality control step in privilege review. In this model, all producible documents are still reviewed by attorneys. Either during the initial training, or in a second phase, a classifier for the identification of privileged content is trained, and each (producible) document is scored for its likelihood to contain privileged information. Recall targets should be higher than for responsiveness review, given the higher stakes for inadvertently producing a privileged document. But even if desired high levels of privilege recall cannot be achieved, a backstop exists because all producible documents will be reviewed by an attorney.[9] In effect, this approach permits a prioritization or stratification of the review of the producible population: documents most likely to be privileged are either reviewed first or are assigned to a dedicated team of highly-experienced reviewers who form a privilege-review team.

Second, predictive coding technology can be used to limit the portion of the producible population actually submitted to attorney review. In other words, predictive coding is used to define a part of the producible population that will either not be reviewed or be reviewed in a different, more cost-effective workflow. Specifically, that population could be subject to a quality control or sampling workflow. This chart demonstrates the stratification of populations:

Figure 1



The outer circle shows the overall population. The blue circle shows all documents predicted to be responsive through predictive coding. The penumbras shown by the dotted lines and lighter color indicate family members: cover emails or attachments of a responsive or privileged document or email. (A "document family" consists of an email and its attachments.[10]) The penumbra around the blue circle shows the family members of documents predicted to be responsive,[11] and, together with the blue circle, defines the producible population. The purple circle shows documents that are (a) producible, and (b) predicted to be privileged.[12] The penumbra around the purple circle shows family members of those documents.

This second approach permits the highest level of efficiency in workflow, if predictions perform successfully, that is, if recall and precision targets are met for predictions of responsiveness and privilege. Four resulting categories of documents can be stratified into three workstreams:

Figure 2

Possible Workflows for successful prediction on Privilege	Predicted **Privileged** (Classifier B)	Predicted **Not Privileged** (Classifier B)
Predicted **Responsive** (Classifier A)	Workstream 3: Attorney Review	Workstream 2: Sample
Predicted **Non-Responsive** (Classifier A)	Workstream 1: Cull	

The listed workstreams are examples of workflows applied to each resulting category of documents. Workstream 1 (Cull) means that these documents are treated as non-responsive and non-producible (except family members of documents predicted R). Workstream 2 refers to documents that were predicted R but were not predicted to be privileged. While the current dominant practice is to review these documents, this is done mainly to avoid

the inadvertent disclosure of protected information. Therefore, if we assume successful prediction of privilege, this category of documents could potentially be produced without human review. There should, however, be rigorous quality control performed on this category of documents through statistical sampling, taking the form of, for example, elusion sampling. This approach will provide comfort that no privileged documents will be produced (within the set tolerances of the sampling workflow).

Workstream 3 refers to documents predicted to be both responsive and privileged. This is the smallest number of documents requiring review, which is conducted for two purposes: First, to identify false positives for privilege, which should be produced (recall that predictions for privilege should be optimized for recall, at the inevitable expense of precision), and second, to enable the preparation of a privilege log.[13]

The proposed workflow, which relies on the successful prediction of privileged information in the producible population, means that counsel cannot rely on human review to achieve the second goal of document review: the construction of knowledge. Given the difficulties of capturing the knowledge about the documents gained during large-scale human reviews, this may not be a significant loss. In fact, highly targeted reviews of documents identified through iterated rounds of advanced analytics (predictive coding focused on core documents, concept clustering, similarity analysis, sampling, and visualization) are likely to be as effective as human review at capturing meaningful insights into the substance of the documents.

4.2 Modeling Financial Impact

There are significant economic implications of successful prediction for privilege. The measure of success must be that a very high level of confidence exists that a producible document population contains no privileged documents. Counsel would thus have assurance that the risk of inadvertent disclosure of privileged documents is controlled and minimized.

As described in 4.1, human review ought to be restricted to producible, potentially privileged documents, in appropriate cases. The chart below demonstrates the potential economic impact.

	Assumption	% of Pop.	Documents
Only Responsiveness Prediction			
Hypothetical Document Population		100%	500,000
Actual Responsive %		20.0%	100,000
Docs selected (TP) (assumed Recall)	85%		85,000
Increase (FP) (assumed Precision)	75%		113,333
Add family members (producible pop.)	150%	34.0%	170,000
Training, sampling, etc. documents		4.0%	20,000
Total to be reviewed by attorneys		38.0%	190,000
Documents not requiring review:		62.0%	310,000
With Effective Privilege Prediction			
Producible Population per above			170,000
Actual Privileged %		2.5%	4,250
Docs selected (TP) (assumed Recall)	99%		4,208
Increase (FP) (assumed Precision)	50%		8,415
Add family members (review context)	150%	2.5%	12,623
Training, sampling, etc. documents *(incl. for Responsiveness classifier)*		9.0%	45,000
Total to be reviewed by attorneys		11.5%	57,623
Documents not requiring review		88.5%	442,378
Net documents NOT requiring attorney review with effective privilege prediction		26.5%	132,378

This hypothetical calculation shows that, even using highly conservative assumptions (such as a total of 45,000 documents reviewed for training two classifiers for Responsiveness and for Privilege, sampling, and so forth), a net gain of over 25% would be achieved compared to reviewing all documents predicted to be Responsive. In other words, review effort would be reduced by almost 70% (57,623 documents reviewed using effective prediction for privilege, compared to 190,000 documents reviewed).

5. TESTING PREDICTIONS FOR PRIVILEGE: PRELIMINARY RESULTS

In testing that is still underway, KPMG LLP Canada has been working with Porfiau Inc. ("Porfiau"), a Canadian company that develops advanced text analytics software. Initial results, discussed below, suggest that Porfiau's technology can detect potentially privileged text with 90%+ recall without generating excessive numbers of false positives.[14] Additional testing is needed to generate more robust results across a larger number of cases, but these initial findings suggest that specific applications of advanced text classification technology can, if deployed as part of an appropriately designed review protocol, provide predictions for privilege at high recall levels at acceptable precision.

5.1 Research Context

This research initiative continues the work of others in recent years in the use of computer technology and, particularly, pattern-recognition, text-analytics and machine-learning technologies, to assist in legal document review.[15] Most of this work has been at the level of general relevance or issue-relevance. The most recent research into the use of technology-assisted review in the detection of privileged documents was the TREC 2010 project, Interactive Task 304. Results ranged from an F_1 of 0.408 (Recall of 0.633, Precision of 0.302) down to an F_1 of 0.126 (Recall of 0.072, Precision of 0.494).[16]

The results reported in this paper suggest that the Porfiau technology, still being developed, can achieve F_1 scores in the range of 0.53, with Recall of approx. 0.90 and Precision of approximately 0.37.[17]

5.2 Porfiau

Using a form of machine learning with statistical pattern recognition, Porfiau deploys proprietary algorithms with finite state machines ("FSMs").[18] This novel application produces two methodological advantages that allow it to identify unusual and nuanced text characteristics. First, it is able to search for representative patterns in 100% of available text characters, including all alphanumeric characters, punctuation and special characters. Only after identifying the representative signal is non-informational data suppressed, both achieving strong signal-to-noise ratios and identifying subtle signals. Second, the methodology classifies new text sequentially, treating each individual character and recognizing signals at the most granular level.

Porfiau's automated text classification method is similar to current technology-assisted review methods in that it must be given a training data, supplied in seed and teaching sets of text units that are strongly representative of the kind of text being sought in the larger population – in this case, privileged text.[19] Text units are typically paragraphs or phrases identified by the reviewer, not

entire documents. A selected "privileged document" is therefore more accurately described as a document containing one or more privileged text units.

5.3 The Origin of the Test

The KPMG/Porfiau test worked with documents from two closed Canadian litigation matters. These two cases provided over 345,273 relevant documents. Of these, lawyers had marked a total of 1,572 documents as privileged. At this stage, "privileged" was accepted as meaning "text that a mid-level lawyer performing privilege review in a standard litigation context had deemed privileged." After being trained on a subset of lawyer-coded privileged documents, Porfiau successfully identified as privileged, within the dataset that had not been part of its training set, more than 90% of the documents that the lawyers had coded as privileged.

A review of the results revealed, however, that, at least in the eyes of a team member with over 12 years of experience as a document review lawyer on both the plaintiff and defendant side (the team's "lawyer-reviewer" or "LR"), many of the documents that the lawyers had deemed privileged were not privileged.[20] This was not merely a matter of business-confidential materials being deemed privileged; often the only indicator of potential privilege was a lawyer's name or a reference to a meeting. In many cases there was nothing remotely privileged about the document.

Meanwhile, the LR found many documents treated as Non Privileged by the review lawyers that were clearly privileged. In many of these cases, the classifier properly identified them as Privileged.

As a result, the team faced the following challenges:

1. To try to disprove the classifier's results, perhaps by finding that certain kinds of text content were causing spurious correlations;
2. Decide whether relying on the initial reviewers' decisions as the best available indicator of privileged content (the "gold standard") would prevent the tool from ever learning how to detect *true privilege*[21]; and
3. Find a way to detect truly privileged content regardless of how the initial reviewers coded the material.

The team thus shifted its focus from the initial question ("Can we find the same kind of text that human reviewers deemed privileged?") to the following exploratory questions:

1. Can we see what kind of text might have caused the human reviewers to misclassify something as privileged?
2. Can we eliminate this kind of text, as well as other kinds of text that might be causing erroneous and/or overly cautious coding in the lawyers' privilege review?
3. Can we identify and eliminate particular kinds of text that might be causing the tool itself to see privilege where there is none, or not to see it where it exists?

… and eventually to the following experimental challenge:

4. Can we develop a tool that is not fooled by *apparently but not actually* privileged text, detect truly "potentially privileged" text with a very high degree of recall, and not pull in too many false positives?

5.4 General Approach and Adaptations in Response to Findings

Test phase I was aimed at determining if the automated classifier ("classifier") could, once trained on a sample of documents deemed privileged by law firm review attorneys, find in the larger review population all of the other documents that the lawyers had tagged as Privileged.

From a case comprising 213,224 documents (Case A), the team created a working dataset consisting of the complete text files from (a) 623 documents that review lawyers had identified as Privileged and (b) 12,450 documents that they considered non privileged.[22] The 623 Privileged documents constituted the entirety of all such documents in the case; the 12,450 Non Privileged documents constituted approximately 5.9 % of all such documents in the case.

To build the classifiers that would find privileged documents, the team selected, judgmentally, two sets of data, each containing at least 200,000 characters: (1) Privileged ("P") and (2) Non Privileged ("NP"). Each of these was then split in two, to allow the team to create a pair of classifiers within the P and NP populations. This allowed for a cross-validation of each classifier against text from the same case that had not been used to teach that particular classifier. The initial phase looked at whether a classifier trained on text type "P" (privileged) could find other P text that wasn't part of its own teaching set. Classifier P1 was tested against text from P2, and *vice versa*; classifier NP1 (non-privileged) was tested against text NP2, and *vice versa*.

Text units from	used to build Classifier	used to find	in text units from
P1	P1	Privileged	P2
P2	P2	Privileged	P1
NP1	NP1	Non Privileged	NP2
NP2	NP2	Non Privileged	NP1

Initial results were in the range of 93-95% accuracy in matching the review lawyers' decisions.[23]

Throughout the experiment, data analysis was performed by Porfiau personnel. All assessments of documents and text to make privilege determinations were by the LR.

The overall approach is similar to the iterative approach used in now-standard technology-assisted review (predictive coding) workflows,[24] although the team used more judgmental sampling and less statistical sampling to assess text units and identify those that were problematic and therefore in need of assessment.[25]

After these initial results (using the complete, document-level text files from Case A), the team collected and added to its analysis pool a set of files from a second case (Case B).[26] By the time these Case B text units were added, the team had learned that the teaching sets would work more effectively if they consisted, not of entire document files, but of selected text strings strongly representative of Privilege and Non Privilege content.[27]

The following discussion summarizes lessons learned following Case A analysis, Case B analysis, and combined Case A+B analysis.[28]

5.4.1 Some text units appeared in both seed sets

The same text is often coded differently by different reviewers, even when the reviewers are working at full attention and with full awareness of the issues in the case and the relevant rules.[29] The team found evidence of such inconsistent coding: a reviewer coded a document Privileged and the same reviewer, later, or another reviewer, did not. The team found all such documents. The LR reviewed them, determined their correct status (P, NP), and selected from them text strings that appeared representative of potentially privileged information. Only these text strings were included in the appropriate teachings set.

5.4.2 Initial reviewer decisions fell short of the gold standard

During the review of initial results to identify spurious correlations and arbitrate the status of text units with contradictory coding, the LR discovered that many lawyer-determined Privileged documents were not in fact privileged, and could not be considered a gold standard. At the same, time, it became clear just how limited the privileged content in a document can be. To select candidates for teaching sets, it was necessary to isolate the most privileged content and discard the rest of the document.

5.4.3 Steps taken to cleanse the teaching-set populations

The challenge of isolating high-quality, strong-signal text strings for use in teaching sets involved two related steps: First, the team developed scripts to identify various kinds of often-repeated text. The assumption here is that any text constituting a privileged communication or work product will, by its very nature, not be replicated across multiple documents.[30] Common forms of repeated text include: email addresses; domain names; application header information (often extracted by processing tools); email signature blocks, boilerplate text such as "Confidentiality" addenda to emails; and embedded scripts, hyperlinks, attachment links and similar coding. These scripts brought the team close to isolating for analysis and possible use as teaching-set units only those strings of text that were written by the person who created the document or communication.[31] With much of the noise content removed, the remaining text (to be used in the teaching sets) was effectively the sub-strings of text that the team wanted to score. The second step was to develop scripts that allowed the classifier to generate scores at the phrase level.

While these steps are useful in enhancing the isolation of candidate text for inclusion in the teaching sets, they also allow for the words or phrases that score highest for potential privilege to be isolated within a document and brought forward for review, using a graphical user interface in which each phrase can be appropriately coded.

5.4.4 Special classes of documents that warrant separate workflow

Most privilege reviews are aimed at finding privileged text that was written only once and whose author and recipient(s) never intended the text to be shared outside the scope of the privilege. In these communications, much of the language used is not clearly legal—at least, not denotatively so. However, many privilege reviews now rely at least to some extent, and even heavily, on keyword searches in which the keywords that are not attorney and law firm names tend to be highly legal in their dictionary meaning: law, legal, lawsuit, law firm, plaintiff, defendant, damages, claim, appeal …. However, whether or not they are effective at finding important attorney-client emails, they are guaranteed to find a particular kind of document which, in its early phases, is most likely highly privileged but which, in its final phase, is *deliberately* made non-privileged. These are contracts and court filings.[32] For this reason, sophisticated review teams have ways of identifying all contracts, draft contracts, filings, draft filings, reports of decisions, and similar documents so that they can be held back from standard relevance and privilege review. The Porfiau/KPMG team built a separate classifier to identify these kinds of documents so that they can be lifted out of the target population and handed over for special treatment.

5.4.5 Whether to try to detect waiver

The team drew a clear line between (1) the detection of the kind of text that is likely to be privileged by virtue of *the text itself* (its wording, its phrasing, its meaning, its content); and (2) the numerous extraneous reasons that exist other than in the words themselves for which privilege may not have in fact existed (there was a third party in the room) or may have been waived (the draft was later shared with a third party). The focus of the project was the detection within the text itself of the *kind* of meaning that *can* make a piece of text privileged. This has obvious practical consequences: metadata and other factors can be ignored during this phase of analysis but must be brought to bear, later or in tandem, as part of a rigorous review workflow.

5.5 Rebuilding of Classifiers and Analysis of Full Population

Using the trained and cleansed classifiers against the full population of 345,273 documents, the team identified a total of 2,823 text units as showing some level of privilege and a total of 8,724 as showing some level of non-privilege. They then selected a sample of 1,737 units from both populations, with the sampling weighted in favor of the privileged population. This sample of 1,737 documents was submitted for review by the LR to assess each text unit and determine whether it could be considered privileged.

5.5.1 Assessment of results

The results were queued up for review by the LR. The LR assessed the text units in a blind test which placed the text units in random order with no indication of how the manual reviewers or the classifier had coded them. Working without time pressure and for only a few hours at a time, the LR declared all 1,737 text units to be either Privileged or Non Privileged.[33] The team treated the LR decisions as the "correct" privilege ruling against which the results of the original manual review and the automated review were to be measured.[34]

The following charts show the penultimate results for manual review versus the automated text classifier.

Lawyer / Manual Review
(pre-adjustment)

	Coded Privileged	**Coded Non Privileged**	
Privileged (LR)	239	255	
Non Privileged (LR)	369	874	
			1,737

Recall	48%
Precision	39%
F_1	43%

Lawyer / Manual Review
(post-adjustment)

	Coded Privileged	**Coded Non Privileged**	
Privileged (LR)	239	170	
Non Privileged (LR)	369	874	
			1,652

Recall	58%
Precision	39%
F_1	47%

Porfiau
(pre-adjustment)

	Coded Privileged	**Coded Non Privileged**	
Privileged (LR)	444	50	
Non Privileged (LR)	616	627	
			1,737

Recall	90%
Precision	42%
F_1	57%

Porfiau
(post-adjustment)

	Coded Privileged	**Coded Non Privileged**	
Privileged (LR)	370	39	
Non Privileged (LR)	616	627	
			1,652

Recall	90%
Precision	38%
F_1	53%

5.5.2 Adjustment for possible source of bias

The team recognized that, in the specific case of documents deemed Non Privileged by lawyers but identified as Privileged by the classifier, the lawyers may have originally detected privilege but then determined that privilege had been subsequently waived. In such a case, the classifier's detection of privilege could not be considered a correction of the case lawyer's decision—that is, as a successful retrieval of a privileged document that lawyers had missed. To assess the extent of this possibility, the team selected the 255 documents that had been tagged as non-privileged in the manual review but identified as privileged by the automated classifier. These documents were reassessed by the LR with the benefit of the full document, associated metadata, distribution lists and family group relationships.

Of the 255 documents, 85 did contain potentially privileged text but any such privilege had been vitiated or waived. The team adopted the conservative approach and restated the results as if these 85 documents had never existed (i.e. they were entirely suppressed from the assessment process). Thus the number of privileged documents in the LR results decreased by 85 to 409, and the number of privileged documents "missed" in manual review dropped by 85 to 170. The adjusted results for classifier-identified documents became 370 for privileged and 616 for non-privileged.

With these adjusted results, the Porfiau classifier recall rate remains at 90% and the lawyers' recall rate rises from 48% to 58%.

5.6. CONCLUSION

The identification of privileged documents using machine-learning technology presents several challenges. Given the high stakes of inadvertent disclosure of privileged information in US legal proceedings, significant economic benefits will accrue to the successful prediction for privilege. The KPMG/Porfiau team tested advanced text classification techniques to retrieve privileged information in eDiscovery cases. Additional tests are in progress. The preliminary results are encouraging and indicate that it should be possible to improve the identification of privileged text using automated means as part of a well-planned and managed review protocol, thus significantly improving the efficiency and reliability of the privilege review process.

7. APPENDIX I: WHAT IS PRIVILEGED INFORMATION UNDER U.S. LAW AND WHY IS IT HARD TO FIND?

1. Types of Privilege

The determination whether a document may be withheld on the basis of privilege can be nuanced and subtle. There are two relevant types of "privilege":

1.1 Attorney-client privilege

Attorney Client Privilege (ACP) protects confidential communications between privileged persons (most often a client and his or her attorney) for the purpose of seeking or giving legal advice.[35] In the context of eDiscovery, the documents subject to ACP are typically communications between privileged persons. Assuming that the client is a corporation or similar legal entity, the privileged communication may be between the client's employees and the client's attorneys (including the attorneys' agents), between the client's attorneys, between the client's employees, or between the client's attorney and an expert retained by the client to assist in understanding information so that legal advice can be given.

In addition, the communication must be for the purpose of seeking or giving legal advice. Communications from an attorney to his or her client are therefore not privileged if they deal with matters other than legal advice, and neither are requests from the client to an attorney requesting information other than legal advice. The fact that legal advice was sought or rendered is itself not privileged. There are a significant number of cases in U.S. jurisdictions dealing with the reach and scope of the privilege, and variations can be observed in practice in how U.S. attorneys interpret the law. This leads to a degree of ambiguity and uncertainty in the identification of privilege.

1.2 Attorney work product doctrine

The attorney work product doctrine protects materials prepared by a client or for a client by its representatives (such as attorneys, consultants, or agents), if the materials were prepared in anticipation of litigation.[36] The doctrine grants immunity from discovery that is potentially broader than the attorney-client privilege, because it goes beyond the attorney-client relationship. Examples of work product are a memorandum drawn up by an attorney containing a summary of facts or plans on how to conduct a trial, or a report prepared by a consulting expert to aid understanding facts related to an anticipated litigation. Disclosure to third parties leads to a loss of the immunity granted by the attorney work product doctrine only if the document is voluntarily or intentionally disclosed to an adverse party. As there is for attorney-client privilege, there is variation in how counsel applies the doctrine in specific cases.

2. Risks of Inadvertent Disclosure

The identification of privilege is a high-stakes part of eDiscovery for two reasons. First, disclosure, even inadvertent disclosure, can lead to a waiver of privilege. The waiver can be limited to the disclosed document, or can extend to the subject matter of the document. If subject-matter waiver occurs, all documents dealing with the same subject matter (and not otherwise privileged) must be turned over to the other side in discovery and may be used by the other side as evidence. Waiver typically extends beyond the litigants in the case where the waiver occurred and formerly privileged documents become discoverable in other cases.

2.1 Subject matter waiver and claw-backs

While any waiver is potentially damaging in litigation, subject-matter waiver is particularly dreaded by U.S. litigants. Rules about when inadvertent disclosure leads to subject-matter waiver differ from jurisdiction to jurisdiction. A rule enacted in 2006 governs the question in federal courts, and grants some specific and beneficial protections, by allowing parties to enter into a court-sanctioned agreement to "claw back" inadvertently disclosed privileged documents and prevent subject-matter waiver (Fed. R. Evidence 502). A court order may also prevent inadvertent disclosure from waiving privilege for other proceedings (F.R.E. 502(d)).

Rule 502 is specifically designed to address the challenges of identifying privilege, and the cost associated with privilege review. The Rule's purpose of limiting the burden of privilege review is addressed in the notes as follows:

> [Rule 502] responds to the widespread complaint that litigation costs necessary to protect against waiver of attorney-client privilege or work product have become prohibitive due to the concern that any disclosure (however innocent or minimal) will operate as a subject matter waiver of all protected communications or information. This concern is especially troubling in cases involving electronic discovery. See, e.g., Hopson v. City of Baltimore, 232 F.R.D. 228, 244 (D.Md. 2005) (electronic discovery may encompass "millions of documents" and to insist upon "record-by-record pre-production privilege review, on pain of subject matter waiver, would impose upon parties costs of production that bear no proportionality to what is at stake in the litigation").[37]

In practice, entering into claw-back agreements in litigation has become routine, but Rule 502 has not led to a marked reduction in the effort litigants devote to the identification of privilege, partially because of some uncertainty around the application of Rule 502, and because even with the protections of Rule 502, the disclosure of privileged documents may be harmful in litigation.

2.2 Privileged document waiver

The second reason the stakes around privilege review are high is that, quite apart from questions of subject-matter waiver and claw-backs, knowledge of privileged documents may give the opposing side an advantage in litigation. For example, a privileged document may lay out an attorney's assessment of the chances of success in litigation, weaknesses in the client's case, unfavorable case law, litigation strategy and the arguments to be relied on at various points in the litigation.

A privileged document may also induce an attorney to go digging deeper into non-privileged documents, where otherwise he or she might not have. While U.S. civil procedure is designed to eliminate surprises and gamesmanship in litigation, in an adversarial system knowledge of the other lawyer's mind confers an advantage. This point is implicitly acknowledged by the existence of attorney-client privilege, which aims at permitting client and attorney to communicate freely and confidentially.

2.3 Can't un-ring the bell

This second concern is commonly referred to with the phrase: "You cannot un-ring the bell." The important point here is that even if Rule 502 and comparable state or common law rules worked perfectly, the full and effective protection of privileged information would still be desirable. At least some of the privileged documents in a case will go specifically to the core of the matter being litigated and the other side should have no access to them, even temporarily. On the other hand, there tend to be privileged documents that are ultimately meaningless to the matter being litigated, and for these documents procedural protections offered by Rule 502 and similar will be sufficient.

To further put privilege review in context, it can be contrasted with review for responsiveness. The stakes in responsiveness review are significantly lower than they are for privilege review. The determination of responsiveness tends to be easier—it is a factual determination that in the paper days was routinely delegated to paralegals. Determining that a document may be withheld on a claim of privilege, however, is a legal determination reserved to counsel, and it can be a complex, nuanced decision. In practice this means that human privilege review tends to be more time-consuming than responsiveness review.

Privilege review is also more costly; rates reported for responsiveness review are often $110 and $175 per hour for less expensive first-pass attorneys, while privilege review can run between $225 and $300 (and often more) for second-pass lawyers.[38] Privilege review is the most expensive part of discovery. Reported estimates are of $9[39] or more per privileged document. In a document population of 500,000 documents at an expected prevalence of 5.0% documents requiring review for privilege (in a second phase after general review), privilege review alone would therefore cost about $225,000.

There is also a practical distinction between responsiveness and privilege review: While a party may face sanctions for failing to produce responsive documents, the other side tends not to be aware of documents that were not produced to it (which is why U.S. eDiscovery requests tend to be broad). It is only in unusual circumstances therefore that failure to produce a responsive document because of simple reviewer error leads to sanctions for the producing party.

3. Current Approaches to Privilege Review

In practice, sophisticated document review projects will rely on four lines of attack to identify privilege during human review.

3.1 Boolean search terms indicating privileged content

These are search terms designed to retrieve passages in documents that request or render legal advice. Terms commonly included are "law, legal, investigation, court, privileged, confidential, complaint, plaintiff, defendant, regulatory, attorney, draft" and so forth. These search terms tend to perform poorly both in terms of recall and precision. As an example, the term "confidential" tends to show very low precision. It is therefore often omitted or restricted to the phrase "privileged AND confidential," a designation often used by attorneys to signal privileged content. The phrase, however, is widely overused and often appears in email footers indiscriminately attached to all communications. Although indexing engines can now remove such "boilerplate" text from the index, generally, Boolean search terms of this type are regarded as necessary but clearly insufficient.

3.2 Boolean search terms indicating privileged persons

Attorney client privilege requires a confidential communication between privileged persons, typically an attorney and his or her client (and their agents). An additional group of search terms is used that contains attorneys' names, of both in-house and outside counsel. The list is often supplemented with URLs for law firms. In practice, the two types of search terms are often combined. Since documents are not privileged simply because they were sent to or by attorneys, search terms indicating potentially privileged persons also tend to show very low precision scores.

3.3 Iteration of Boolean search terms

As more information becomes known about the case and privileged documents in the corpus, search terms are iterated and continually improved. There are two main strategies: the first is to add substantive terms indicative of privileged documents in the corpus. For example, in an antitrust case these might be "antitrust", "monopoly" or "dominant." The second strategy lies in refining the search terms to improve precision. For example, the optimal search might be for "[name of counsel AND antitrust]" to stop news stories that mention antitrust from being returned. Significant experience and skill is required to optimize Boolean search terms in this manner.

3.4 Advanced Analytics and Quality Control

The final line of attack to identify privilege is the use of advanced analytic technologies offered by state-of-the-art eDiscovery tools. For example, concept clustering permits attorneys to use unsupervised learning techniques to review documents that are similar to documents already identified as privileged, but that did not hit on search terms. A vital step here is quality control both of the search term results (for example, using statistical sampling[40]) and of the human review of potentially privileged documents. This quality control step should use all advanced techniques, including clustering, visualization, sampling, similarity analysis and hash-based de-duplication, thread analysis for email, and social-network data (that is, to/from metadata).

---

8. ENDNOTES

[1] RAND Institute for Civil Justice publication: Where the Money Goes; Understanding Litigant Expenditures for Producing Electronic Discovery, 2012; page xiv.

[2] Grossman, M. R. & Cormack, G.V. 2011. Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, XVII RICH. J.L. & TECH. 11 (2011). http://jolt.richmond.edu/v17i3/article11.pdf.

[3] Cormack, G.V., Grossman, M.R., Hedin, B., Oard, D.W. 2011. Overview of the TREC 2010 Legal Track, 23. http://trec.nist.gov/pubs/trec19/papers/LEGAL10.OVERVIEW.pdf

[4] *Ibid*, 35.

[5] The Sedona Conference, The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process, 10 SEDONA

CONF. J. 299, 302 (2009), as quoted in Baron, J.R. 2011. Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search, XVII RICH. J.L. & TECH. 9 (2011). http://jolt.richmond.edu/v17i3/article9.pdf

[6] Here and elsewhere in this article, the term "privilege" generally refers to both the attorney-client privilege and the work product doctrine.

[7] KPMG selected a judgmental sample of 51 hosted document repositories which included more than 27 million documents and ranging in size from 4,417,957 to 4,231 documents. These repositories represent a wide variety of industries and matters including civil litigation and investigations dating from 2007 to the present. The following table summarizes the sampled repositories and the aggregate numbers of documents by general category. Please note that review strategies and coding instructions vary from engagement to engagement, and that these numbers were not normalized to account for differences between matters.

Total Documents (51 engagements)	27,785,549
Total Review Corpus	13,796,558
Documents marked Responsive	2,705,977
Documents marked Privileged	316,123
Reviewed to Responsive Ratio	19.61%
Privileged to Responsive Ratio	11.68%
Privileged to Reviewed Ratio	2.29%

[8] In KPMG's 2011 paper: "Software-assisted document review: An ROI your GC can appreciate," we determined that software assisted review achieved consistently higher recall than multi-stage human review on three out of three matters tested, comprising more than one-hundred thousand documents. And, in preliminary tests of advanced document analytics technology that was applied to two previously manually coded privileged document sets, the analytical technology was significantly, (almost 50%) better at identifying privileged text than the human reviewers. This improvement may not, however, be sufficient in practice to rely on privilege predictions.

[9] Whether this is an *effective* backstop depends on the expected error rate for human review.

[10] There are other types of document families, for example, a word-processing document with embedded objects. Such document families tend to be artifacts of data processing.

[11] Despite some differing preferences among law firms, the generally preferred approach in eDiscovery document review is to make coding decisions per document, whether an email, attachment, or loose file, rather than by document family, and to rely on family-relationship metadata to construct the final production set. Predictive coding solutions currently available are implemented to work on a document-by-document level (and not family-by-family), and therefore fit well into the workflow.

[12] Privileged documents are likely to exist among the non-responsive documents, but unless producible as family members of responsive documents, their review for privilege is not required since these documents will not be produced.

[13] A privilege log lists all documents withheld on the basis of privilege, lists relevant meta data (author, recipient, etc.), and must identify the subject matter of the communication as well as basis for the claim of privilege. *See*, for example, F.R.C.P 26(b)(5)(A).

[14] The Porfiau/KPMG team is working to develop an approach to privilege detection with a very high degree of recall for user-generated text strings that are neither boilerplate nor legal-document legalese, but that are likely to be privileged in any litigation—without addressing whether such privilege was later waived or vitiated.

[15] Landmark papers in this field include Grossman & Cormack, "Technology-Assisted Review," *supra* note 2; Cormack et al., "Overview," *supra*, note 3.

[16] *See* Cormack et al., "Overview," at 35.

[17] For the TREC 304 task, the participants assessing attorney-client privilege were given only communications, while, for the assessment of work product, they were assured that the documents being reviewed were material to the litigation. These screening steps allowed the reviewers to focus on the words themselves and ignore extraneous factors. In the Porfiau experiment described here, there was no culling or categorization by communication/ non-communication andthe documents used in the Porfiau analysis had not been deemed Relevant by review attorneys.

[18] The technology needed to build the FSMs used in this study is not public domain; it is owned by Porfiau. Notwithstanding this, it is possible, for testing or academic purposes to have FSMs built to recognize a given sample of text. This is a web service offered under http://www.porfiau.com to allow the general community to test and experiment with this technology.

[19] In this paper, "seed" set refers to the text units used in the first round of analysis and "teaching" set refers to the more-refined collections of text units used in later rounds of analysis.

[20] The LR is a KPMG manager with a Canadian law degree and several years of experience as a document review attorney in the United States, on both the defense and the plaintiff side. Although overall he applied a strict standard when assessing privilege (with the instincts of a plaintiffs' attorney looking at a defendant's privilege log), he erred on the side of finding privilege (agreeing with the initial reviewer's privilege call) whenever there was a colorable case to be made for a claim of privilege or at least for seeking a second opinion. Text discussing business-sensitive material, misdeeds or embarrassing facts was not treated as privileged, nor was text merely evidencing that a request for legal advice had been made or that a legal discussion was going to occur.

[21] As in Grossman & Cormack, "Technology-Assisted Review," *supra* note 2, relevance calls of junior reviewers were compared to (a) the relevance calls of a team using predictive-coding technology *and to* (b) the decisions of a subject-matter expert or topic authority. Thus the predictive-coding tools were not assessed on whether they successfully predicted what the first human reviewer did; they were assessed on whether they predicted the subject-matter expert's decision. This decision is the standard against which both the first reviewers and the tools were judged.

[22] The word "considered" is deliberate. Documents referred to in this paper as "Non Privileged" are those documents that were *not affirmatively coded by a reviewer as Privileged*. Privilege review is usually restricted to documents retrieved by Boolean searches. It is therefore typically inaccurate to suggest that a "Non Privileged" document was *declared by a reviewer* to be Non Privileged; rather it has simply not been declared Privileged. Thus, the document is "considered" Non Privileged.

[23] Measured as follows: For any 100 documents that the lawyers had coded "Privileged" but which the classifier had not been given to train on, 93 to 95 of them were coded by the classifier as

Privileged. The same results were observed on the Non Privileged side.

[24] *See*, for example, D. Grossman, "Measuring and Validating the Effectiveness of Relativity Assisted Review," February 2013, available at http://www.edrm.net/resources/edrm-white-paper-series/measuring-and-validating.

[25] By problematic, we mean (a) text units that appeared in both of the initial teaching sets (Privileged and Non Privileged) due to inconsistent review-team coding; and (b) text units on which the initial lawyer Privilege call and the Porfiau classifier Privilege call differed. It was these units that received the most attention during the phase devoted to refining the teaching sets. The LR, when assessing these problematic units, either (a) confirmed a Porfiau Privileged call for the whole unit, (b) confirmed a Porfiau Non Privileged call for the whole unit; (c) overturned a Porfiau Privileged call for the whole unit, (b) overturned a Porfiau Non Privileged call for the whole unit; (d) replaced the text unit that had been used with a refined/edited text unit, while also assigning the appropriate Privileged call; or (e) declared the text unit to be inappropriate as a teaching-set text unit (e.g. boilerplate).

[26] From a total of 132,059 files, 949 Privileged (100 % of all Privileged) and 7,536 Non Privileged (5.7% % of all Non Privileged).

[27] Not entire document files, but selected paragraphs, sentences, even phrases. Each was coded so as to be traceable back to its original document. The "larger context" of the text string was **not** taken into account in order to overturn the tool's assessment. The algorithm was designed to score the text it was actually "looking at" (x characters at a time), not to score it in light of text it had analyzed 100, 500 or 1,000 characters earlier, or would soon be analyzing lower down in the document.

[28] By "combined," we refer to the fact that the classifiers can be taught with text strings from multiple sources. (Some predictive coding protocols allow for, even encourage, the use of text from third party sources, such as newspaper articles, to prime the pump. This was not done here. All text strings came from Case A or Case B.) As the experiment proceeded, the team used text from both cases as it developed more refined teaching-set inputs and also used text from both cases as the target population for analysis.

[29] *See* Grossman & Cormack, "Technology-Assisted Review," *supra* note 2, at 7-15; Cormack et al., "Overview," *supra* note 3, at 20.

[30] Where a piece of privileged text *is* replicated (e.g. in an email thread), it only needs to be captured and used in the teaching set once; repeating the same text in the teaching set adds nothing but risks creating inaccurate signals.

[31] The team did not engage in this kind of cleansing of the target population, i.e. the universe of documents to be searched for potentially privileged content. Cleansing was only performed within the population of text units that were candidates for teaching-set status, and these had been copied out of the larger universe. This decision reflects two judgments: first, an appropriately sensitive and precise classifier will be able to detect privileged content even amidst a lot of noise; second, in real-word applications, it will likely be infeasible to first cleanse a target population before running the tool against it.

[32] Draft contracts alternate between privileged and waived: they lose their privileged status whenever they are shared with the other side in the negotiation, but quickly regain it in the next round of mark-ups. Court filings often make their way into court decisions; judges sometimes lift entire passages from litigants' submissions. But the drafts of those filings likely remain privileged.

[33] *See* note 19, above, on the LR's sensitivity to plausible claims of privilege.

[34] Any assessment of the ability of technology to improve on human privilege review requires a human to make the final determination as to the technology's performance.

[35] *See* Fed. R. Evid. 503(b)

[36] Fed. R. Civ. P. 26(b)(3)

[37] Explanatory note on Evidence Rule 502.

[38] RAND Institute for Civil Justice publication: Where the Money Goes; Understanding Litigant Expenditures for Producing Electronic Discovery, 2012

[39] Losey, R. 2013. Bottom-Line Driven Proportional Review (2013 Updated Version). http://e-discoveryteam.com/2013/03/14/bottom-line-driven-proportional-review/, citing "In re Fannie Mae Securities Litig., 552 F.3d 814, 817 (D.C. Cir. 2009)"

[40] *See* Paskach, C., Carter, M., Strauss, P. 2012. The Case for Statistical Sampling in eDiscovery, KPMG LLP, January 2012.