Ten Foundational Flaws in Key Word Search, Prioritized Review and Predictive Coding

(A draft for comments, version 1.02)

Jianqing Wu, Ph.D.

In this paper, I will prove that three lines of technologies: search methods (when it is used to capture documents to create a review pool), prioritized review methods, and predictive coding methods are invalid because they are developed on a foundation with ten flaws.

A. Ten Foundational Flaws in The Questioned Technologies

Before I discuss ten foundational flaws, I first present informal survey results on how document reviewers generally feel about the performance of those technologies based upon their first-hand experience. Among those who have solid litigation background, they would express the view that those technologies are largely useless and pose unreasonable risks to litigants' causes.

I also would like to conduct a survey on how many times law firms abandoned the use of predictive coding after they found that the work product of the algorithm could not meet the lowest requirements, as they promised. Why litigation experts could not accept those technologies while the information world has published a large number of studies to show the validity of those technologies?

I can show that the questioned technologies do not deliver claimed benefits, and are not only useless, but also responsible for the bottom-line quality of e-discovery performance. To prove my point, I ask you to review one law review article on this subject: Technology-assisted review in e-dis-

covery can be more effective and more efficient than exhaustive manual review"[1] This article concluded that "technology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort." I can show that the conclusion is patently wrong. I show you ten foundational flaws which are revealed in those technologies discussed in cited articles.

1. First Foundational Flaw: All Performance Criteria Derived From Tag Counts and Document Numbers

In essentially all studies in cited studies, tag counts and document numbers are used as statistics. They shared a first common foundation flaw of using coding consistency, accurate rate, precision rate, percents using tag counts and document numbers as performance criteria. Wrong conclusions were reached because tag counts and documents are meaningless statistics.

All articles cited in this law review article used coding consistency or comparison with imagined correct answers. In applying those criteria, they actually imitate widget quality evaluation method. To make their error crystal clear, one needs to think how factories evaluate "thing". They routinely compare a TV set with identical TV set in quality comparison. If anyone tries to compare a 35-inch TV set with a 50-inch LCD TV, we all can see the flaw. If one tries to compare a TV set with a fighter carrier, we probably will be shocked. This is exactly what happens in e-discovery quality evaluation. They not only treat different contracts as same, but even treat different documents such as 100-billion acquisition agreement and a trivial email as same. In the cited studies, they assigned all issue tags with equal weights and treated all documents including email, contracts, counsel opinions, and

1 Cite as: Maura R. Grossman & Gordon V. Cormack, 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.ppdf.

board's resolutions as same even though they have several-magnitude different impacts on litigation and litigant's other interests. The difference between two documents can be as much as the difference between any two things we can see. Evil intents, for example, can range from steeling a few cents, beating up a person, embezzling a million dollars, embezzling 30 billions, killing a person, bringing down a plane, and poisoning a community.... Some studies have skillfully avoided addressing document nature and turned documents into “abstract numbers” for comparison. Skillful treatment of data cannot avoid the absurd logic that they treat paper clips and fighters carriers as same.

Using coding consistencies, accurate rates, precision rates and all number-based similar quantities as performance evaluation criteria is a fatal flaw. All validation studies cited in this law review article and other publications are invalid. Since this foundational error cannot be cured, and all technologies that have been validated by using this and other similar criteria are invalid.

2. Second Foundational Flaw: Misuse of Sample and Population for Performance Evaluation

In conducting a performance study, a sample is taken from a population for study and inference is made from the result of the sample to the population. If a sample and a statistical population are TV sets, the inference is valid. By extending this widget model, the proponents regard a document as one member in a document pool and thus achieved a perceived number-based frequency distribution. This sampling method is invalid even if the inference is made for document's printing costs because some documents have only one page and others have hundreds of pages. In litigation, document numbers have no meaning to case disposition. The nature of information, carried risks, and potential impacts on the case can be dramatically different. Any inference from this frequency distribution to litigation

performance is meaningless because 99% of the documents are meaningless to litigation outcome.

The essence of this foundational flaw is swapping two distributions for making performance inference, as shown in the appended figure. The simple frequency distribution for documents is shown in the top figure a, which is widely used for identical physical things such as TV units. From this distribution, a sample can be taken to represent the population. It does not have problems in sampling method, risk tolerance, and work performance when it is used in the widget world. Due to unit uniformity and value proportionality, percentage measures and most statistical analysis methods can be used without violating basic assumptions. That is why good and bad widget unit numbers in any sample can be extrapolated to the population, which may represent production performance, labor performance, risk of loss, and revenue gain.

For a given document pool, documents can be classified according to information important to litigation, we would conceptually see a totally different distribution, known as litigation-significant document distribution, as shown in the figure d. Documents in this distribution may comprise one or more smoking gun documents (HC and HC'), a few critical documents (C and C' in figure b), some documents that affect weights of evidence on critical issues (W and W'), some documents that are required for essential issues (E and E'), and basic relevant documents (B and B'). Each class of documents may have two opposite roles. Critical documents include those which are not directly relevant, but can change or have a severe impact on case outcome. Those documents may include subjects revealing company's culture of fabricating documents, past similar violations or acts, and actions or statements which discredit its own legal theories. Many duplicate relevant documents and a good portion of marginal documents are also mean-

ingless. We can classify those documents in different ways, but their roles are reflected in court opinions for closed cases.

A document pool contains a large number of non-relevant documents. Many of them contain sensitive business information (figure b), and some of them may carry damaging information. The effects of non-responsive information on produced responsive documents and non-responsive documents are shown in figure b. Producing non-responsive documents is not good to the litigant but some of the documents may have various impacts on litigants' long-term interest.

The document pool also contains some documents which may reveal litigant other liabilities in any of potentially hundreds of areas of law under any of many jurisdictions the litigant does its business (shown in figure c). Those risks are revealed in relevant documents and inadvertently produced non-responsive documents. While it is hard to find those documents, the odds for finding those documents are always more than a plurality of units. Most of the times, litigants do not know that certain things they are doing are illegal under specific law.

Litigation-significant document distribution is dynamic, and dependent of legal theories of two sides. This distribution is totally different from the simple frequency distribution that is widely used in widget production. The population of documents which really affect case disposition is often less than 0.1%, and in criminal cases, it is a few in a million. It is a blunder to draw a document sample from this simple frequency distribution, evaluate the sample, and then make an inference to litigation performance. By using this frequency distribution, the proponents also make another blunder of giving improper weights to highly critical documents. Unlike a widget, even relevant documents have totally different roles and different impacts. Even certain non-relevant documents can alter litigation course. One single document to address a highly contentious issue can alter case disposition, a

few documents which can change evidentiary strength or change witness credibility can give a litigant a final victory, a scandal-carrying document may put a litigant in a defenseless position, and a document in support of willfulness of a person may support punitive damages.

The significance of a large number of documents also depends upon how other related documents are processed if legal theories are fixed. If thousands of documents are probative of a necessary issue, and if all of them are unavailable or lost, discovering anyone of them would be critical. After one or several of them are available, the rest of them become irrelevant numbers. If several different documents can be used to prove a same element with different forces and impacts, some of them may be used to replace others but not vice versa. There are all kinds of information interactions which can be seen only in specific factual environment. While marginally responsive documents are meaningless, some of them may be useful for resolving side issues.

Information requirement is highly specific. If a party is required to prove a particular critical element, only documents concerning this element will do. When those critical documents and highly relevant documents are settled for the legal theory, all the vast number of other documents, including those hundreds of copies of duplicate responsive and marginally responsive documents, are meaningless. For all those obvious reasons, litigation performance cannot be evaluated in any method used in a widget production shop, where a defect TV unit can be replaced by another unit.

Given the huge difference between the document frequency distribution and the litigation-significant documents distribution, and practical inability to adjust the weights of different documents, we MUST reject all performance data that have been based upon the document frequency distribution. In addition, we must reject performance evaluation and current quality control method because it is not entitled to make any inference from sam-

pled documents to the documents in a pool concerning performance. This error cannot be cured, and all technologies which have been validated by using this sampling method in more than a decade are invalid.

3. Third Foundational Flaw: Use of Improper Risk Model

At a document production site, one has to consider what kinds of errors a litigant can tolerate. At a widget production site, production risk is assessed by analyzing a small percentage of units and making a reference from a measured sample to the whole batch of produced units. One important thing for using this quality control method is that loss is proportional to defective units and thus any prediction inaccuracy can result in only a tolerable loss. However, the sampling method cannot be used in cases where damages are beyond the unit value. A defect of a plane component can not only bring down the plane but kill passengers. Sample-based quality control method cannot be applied to the plane component.

Risks of harm from improperly processed documents are closer to the aviation risks than production risks. When a critical document is missed, mishandled, or improperly processed, its impact is not limited to the document itself. The mishandling of the document may doom the case, harm the company by exposing other unrelated liabilities, and cause the company to lose its competitive standing. The impact can be far more than the stake in dispute. It can cause the company to fail and thus injure shareholders, pension holders, retirees and all kinds of persons in between. It may also affect criminal liabilities of responsible persons.

It only takes one or more documents to alter litigation course and final result. For a million of documents, a 15% percent quality control sample means that 85% of documents are not assessed. Those 85% unchecked documents (850,000) can easily accommodate a few, hundreds or few thousand of documents and could be thousands of times of a “room” required to hold the devils. The odds for the devils to fall within the sample is so small that it

must be regarded as a non-happening event. For this reason, any sample-based quality control process should be regarded as no quality control having been done. It is a flawed application of the widget risk controlling model. Small probability theory, which is widely used in production setting and research, may not be used to address this type of problems.

This sampling-based quality control method has no real utility in controlling litigation risks except that it may be used to check out whether reviewers have systematic different coding habits or misunderstood coding guideline. In reality, it is not used for this purpose because quality control is performed after the first review is finished.

4. Fourth Foundational Flaw: Incomplete Benefit and Risk Analysis

Most of the cited articles and other studies fail to conduct a complete benefit and risk analysis. When the results of incomplete analyses are presented to litigants, they are not fully informed of risks beyond their pending cases.

In most cases, cost-benefit analysis include only the pending case even though exposure of secondary liability and loss of business information are generally known in the legal profession. Documents can affect not only the litigant's chance to win in the pending case but also its future competitive standing in the business environment. Nearly all court rulings consider only the stake in current case and ignore the fact that a review production can affect more than the litigant. Production quality in criminal cases can affect civil fines, criminal fines, restitution and forfeiture, decisions for prosecuting, selection of defendants for criminal charges, and entities' right to exist. In addition, litigation may affect the jobs of the litigant's employees, pension of retirees, investment of shareholders, and rights and obligations of many related parties. No one can say that a huge amount of information including trade secrets, business plans, pipeline products and services, customer lists, research and development plans, and

everything designated by computer algorithms is harmless.

As a result of using the incomplete risk-benefit analysis in evaluating those technologies, litigants are not informed what a large number of produced documents might contain, where their critical information will end up, and how they can be harmed many years later. While the legal profession generally regards such information as given-away public information, many litigants are not told of the truth at the time when they were asked to use those technologies. The stake from the information in review production could be far more than the litigant's stake in the pending case, and many litigants have made wrong decisions, specially if the business information and technologies are important to their continued success.

A litigant in defense has an incentive to produce fewer documents. Unfortunately, this goal cannot be achieved by using computer coding methods. The litigant cannot assume that those technologies definitely favor defense. Random drawing methods cut in both ways. When those technologies designate responsive documents, they are unable to recognize favorable documents, scandal-carrying documents, and other risky documents. There are all kinds of potential issues that may affect the litigant's defense in unpredictable ways. Produced documents may contain information that damages the litigant public images, affect its competitiveness, and increase the risk of unrelated lawsuits. The technologies are not in a position to accurately assess if documents carry stakes which are larger than what they might have in their pending cases.

5. Fifth Foundational Flaw: Bias, Unfair, and Useless validation Methods Designed to Achieve Fixed Results

It can be shown that a large number of factors work together to degrade review quality so seriously that human review cannot be used as a comparative standard.

Under the network-based distributed review model, each of N reviewers, who can get 1/N case information in the whole review life, reads each document in isolation from other documents, analyzes substances by relying upon limited personal knowledge out of language and business contexts, and codes documents by exercising subjective personal judgments according to complex instructions which are often poorly communicated due to all kinds of practical reasons. In nearly all cases, document context deficiency, reviewers' knowledge limitations, differences in reviewers judgment, inherent communication mishaps, and constant changes together with hundreds to thousands of other variables to make accurate review impossible. Under the review method, every reviewer does his own work, the whole team may do N-fold duplicate efforts, and all reviewers code documents by making “the best calls” on the basis of what they are able to understand from the four-corner view of each document.

Due to this model defect, the current review model is completely incompatible with the nature of discovery tasks. It can be easily seen that the authors of corporate documents are not only experts in the fields of their writings but also experts in particular subjects of their writings. The reviewers are confronted with an overwhelming number of unknown including business background, business environment, relevant events, company organization, governing law, and an overwhelming number of specific case facts including person names, entity names, product names (product codes, serial numbers, model numbers, and common names), transaction names, special terms, implied time, contextual information, information passed by understanding, unstated facts or events, and material assumptions. The task of document reviewers is to investigate facts, but not just scan documents to appreciate general ideas. Despite extreme difficulties, the reviewers have to understand documents.

The proponents not just failed to do anything to remedy this serious problem, but seized the review model problems to promote flawed, defective, and useless technologies for their financial gain. First, they develop a method for pulling only a small or fraction of documents for review. For example, they may pull only 5% to 20% by using poorly formulated search keys. Thus, a first document may be “Good job, we signed the agreement. Jack loved the deal.” The last document might be a document that describes details about the deal. In this example, even a reviewer can read every word, the reviewer could NOT see legal significance because most of words lack antecedents. The search methods also affect human review by a separate reason: most reviews are finished in short time windows before reviewers have acquired enough case information for conducting meaningful review. Thus, review quality is set to the lowest quality that are ordinarily seen in the first a few weeks.

Destroying document reviewability was not the only thing the proponents did. By leveraging litigation cost pressure, they skillfully worked in concert to introduce practice of imposing review quotas such as hourly quota at 40, 60, 100, and even 150 documents per hour, and turned document review into a computer game of reading subject lines and highlighted words. They achieved this by providing flawed performance-evaluation tool, and exerting cost pressure by showing how faster, cheaper and more accurate computers are. At many review sites, reviewers have to work like Charlie Chaplin in a modern assembly line doing simple tasks they are deemed to fail because they can never beat computers. In playing this twisted game, human's inconsistencies must be higher than computer algorithm's because the computer algorithms can “see” all documents, while reviewers can see only the current document under review. They then showed human's higher inconsistencies as proof that computer methods are more accurate. Essentially, by manipulating those factors, they created unfair and inaccurate vali-

dation results that technologies-assisting review is faster, cheap, and more accurate. They then published such unfair performance data. The technologies accuracy claims essentially degrade the value of professional skills to the computer processing price of about $400-500 per GB data, and is responsible for devaluing legal professional skills.

In parallel to the development of this unfair, invalid, and misleading validation methods, proponents also developed practices to reduce reviewer qualifications to the minimum. They have achieved the objective by using improper performance metrics and taking advantage of many interference factors. Fast buzzword-based coders become super stars, first-day reviewers can top the whole team due to their high daily numbers, and lucky junk-document reviewers are rewarded due to high numbers. When no body is able to see performance differences, site managers naturally retain reviewers of lowest qualifications at lowest pay rates. This consistencies-based evaluation method has guided the industry to form a retention culture that everyone is good for each review job and very review job is open to everyone. Reviewer qualifications are lowed to the minimum: a license and an ability to walk to a review site. By engaging in all those activities, they have further damaged the legal profession by retaining untrained, unqualified and misfitted reviewers. Inflated technological accuracy, low human review speeds, high computer coding speeds, and litigation cost pressure work together naturally to force litigants to reach an unstated notion that contract attorneys deserve a minimum pay. By running document review in this race-to-the-bottom spiral process, those technologies are actually responsible for low review quality, low reviewer qualifications, and low pay rates.

I can further show that coding variances or coding conflicts from many factors are so large relative to those affecting true performance that all human validation and comparison analysis are meaningless. Factors that affect variances include document context deficiency, reviewer knowledge

limitations, practical communication problems, reviewers' different judgment, and routine changes. Factors that affect performance more seriously include review quotas, unworkable reviewers-and-law firms relationship, depressed pay rates, and retention of mismatched reviewers. Those factors are responsible for the failure of e-discovery in all three objectives: capturing winning odds in the current case, controlling secondary or unrelated risks, and controlling the litigant's sensitive business information.

6. Sixth Foundational Flaw: Omitting Concept-Concept Interactions Within Documents, Between Documents, and with Non-documents.

By extending the widget production model to e-discovery, those technologies have treated all documents as being independent of each other. However, corporate documents generally concern a series of transactions. They always incorporate prior events, terms, subjects, concepts and anything in later documents. In addition, documents routinely incorporate any concepts from business environment. They also incorporate anything existing and being commonly known among themselves. Documents falling within the litigation-significant document distribution are intertwined with legal theories and the strategies of the adversary. Document significance changes with their relative numbers. When a litigant has only one document to prove one required element, it is a critical document. The interactivity is more than language context that is generally known in linguistics. Interactions exists in many ways. Prioritized review methods and search key methods can dramatically change document appearance orders, making accurate review impossible, especially in the first a few weeks of review.

7. Seventh Foundational Flaw: Treat All Documents as Equal in Allocating Reviewing Time

While the proponents of those technologies did not directly use this foundational flaw in their core software, they have skillfully provided such tools for others to abuse. This tool may be called as analytic, performane

evaluation tool, or other names. Their start point is using unfair and misleading performance results to generate pressure to reduce discovery costs. It is clear that computers are capable of doing a few simple things at very high speeds. A strategy to win market share in this service market is to change the nature of discovery from investigative tasks into a simple computer game. This requires a new way of measuring amount of work. Thus, they come up with quota. Review quota is similar to a production quota for *anything* in the physical world. They have successfully turned investigative tasks into the game to race with computers in reading buzzwords and highlighted words. In performing such simple tasks, computers are always the winners. By using such misleading results, they harmed the reputation of licensed attorneys and naturally invite unthinking managers to demand higher review speeds by using meaningless analytic tools they have already provided.

At a widget production site, uniformity of the amount of labor is beyond question. However, review speed in e-discovery is a function of a long list of variables such as reviewer's skills and reviewer's background, litigation subjects, task nature, document length and complexity, issue complexity and their interactions, document reviewability, network delivery speed, file format, risk levels and stake involved.... There are many more variables. None of those variables are fixed and capable of being fixed. Document lengths alone can differ by a thousand times. This is same as talking about weighing weights without a weighing standard. The comparison is exactly same as demanding assembly workers to build a fixed number of anything. Such absurd quotas can never be accepted at widget production sites, but can actually be carried out at many document review sites. Ironically, for the same reason that metrics are detached from performance, it is easy for contract attorneys to defeat it. Document reviewers are forced to use the same amount of time for reviewing a junk email or a document carrying bil-

lions of dollar stake, and they are forced to use the same amount of time to review a two-thousand-page document and a one-line email. Quotas force them to skip reading up to 99% of texts and make review quality approach to zero. In the end, the litigant actually produces documents where a majority of portion of written materials are not read.

Any quota can also directly ruin contract attorneys and management relationship by creating constant friction. It forces contract attorneys to raise speeds. However, the review team always has review speed distribution. If the management terminates a few slowest reviewers, it would create new slowest reviewers next day. Quotas viciously push reviewers to raise speeds in perpetual cycles. In the end, reviewers have to work together to fix their hourly numbers. Thus, a reviewer who could code a hundred of junk documents per hour would just code 40 junk documents per hour. Performance gain is achieved by reducing review quality, but performance loss is actually lost. If such as quota is used in a widget production, all produced products would be paper clips rather than air planes.

8. Eighth Foundational Flaw: Assuming Every Document Has Only One Right and Fixed Answer

Most of the validation studies are based upon an assumption that there must be one right answer. This may be so when all surrounding parameters are fixed. It is improper during the discovery stage when none of the parameters can be fixed. For example, conflicts of coding may arise from different understandings between the drafters of document requests and the attorneys who produce responses. Coding conflicts also arise from original documents which reflect internal disputes, fraud, mistakes, and omission. Coding conflicts may also arise from different document usage histories and distribution histories. An identical document may be properly coded as confidential, public or privileged and non-privileged. No argument can be made that a later disclosure must change its confidential or privi-

leged status retroactively in all cases, especially if this timing fact is significant to the investigative purpose.

For internal investigation, coding conflicts due to personal conduct, business transactions, and statements may carry important investigative information for finding truth. The consistency rule forces reviewers to eliminate the most important, perhaps, the only leads prematurely. After those conflicts are removed by simple-minded quality control staff, there is little chance to find reasons for the conflicts.

9. Ninth Foundational Flaw: Misuse of Percentage and Statistical Analysis

In the physical world, percentage and statistical comparison are often the best ways to make comparative analysis of performance or quality. However, when one sees the litigation-significance document distribution, all of those analysis methods are meaningless and improper. The flaw is that one uses a method which is incapable of differentiating real quality. When a toxin can kill human by part per million, a percentage measure would be meaningless.

Each document pool has a large number of substantially duplicate documents and marginally relevant documents. For example, the number of documents that may be characterized as invoices can be of nearly hundreds. Additional duplicates by file backup practice would make several hundreds. This makes a vast of their tag numbers meaningless in most cases. Two best reviewers can easily have more than 50% coding conflicts. Coding inconsistencies are measured by tagging counts or variances of assigned values for conflicts. The variances from those two sources alone can be much larger than what are required to accommodate the difference in coding litigation-significant documents. Any validation studies using percentage differences and statistical comparisons such as variance analysis, significance test, and

confidence intervals are meaningless. The world is still busy with conducting validation tests using those methods.

Small probability theory and percent expressions have no utility for addressing this kind of low frequencies or very low frequency problems.

10. Tenth Foundational Flaw: Ignore the Effects of Many Factors on Nature of Errors.

I can show that nature of errors depends upon many factors. Reviewer qualification and background is the first factor. Other factors include review speed, document composition, review platform, and reviewer's prior experience. This problem makes most performance comparisons meaningless because errors defined by a same error rate can mean totally different things.

The cited studies fail to consider how characteristics of corporate documents affect nature of errors. Corporate employees generally perform repetitive functions and routinely discuss business matters with a team or a group of employees. This practice results in a large number of duplicates documents, but the numbers of duplication depend upon distribution scope defined by a team, a group, division, and company. Employees also routinely amend the same documents or discuss the same subject matters many times thus results in substantially duplicates documents. Email becomes longer and longer with dialog continuing.

Each of original documents becomes a large number of copies as a result of routine backup. Thus, most transaction documents have tens, hundreds, or even thousands copies. The document pool also has low-frequency documents which contain only one copy or limited copies. Those documents are called singlets, doublets, triplets, quadruplet etc. Documents concerning personal matters, highly sensitive matters, crimes, bad conduct, unpleasant incidents tend to be in low frequencies. Employees are less likely to distribute sensitive documents to a team, a group, a division, or depart-

ment, or the whole company. Documents for expressing official approval of a matter are generally in lower frequencies. When an executive expresses an opinion about an important corporate matter, the executive would not distribute it to the whole company. An employee would not tell everyone that he has committed crime, looted the company, stole company property, or did something bad. Documents such as customer complaints and notices from outside sources may have limited copies. Those documents may be subject to fewer backup operations. Therefore, a document pool comprises a large number of series of duplicate or substantially duplicate transaction documents, plus a large number of unpredictable and uncertain singlets, doublets, triplets, and low-frequency documents.

The routine transaction documents not only have a large number of variants or copies but also tend to contain “expected” search keys. Those large series of documents can be easily captured by random drawing or any search methods. Chances to capture documents by drawing decrease with their duplicate count. It would have thousands times more likely to capture a large serial documents than a singlet. When a search method is used, the duplicate nature affects capture yield. If a search key hits one word in a document, it would get all similar documents and copies. Capturing the large serial duplicate documents would give impression of getting a great deal out of the document population. This explains why even a 1% seed documents can lead to a super majority of documents in a short time. Now, recall rate is used to measure responsive documents captured by algorithm relative to all responsive documents. A 80% recall rate means that the algorithm can capture 80% of potentially responsive documents and miss 20% potentially responsive documents. One cannot take this number as a meaningful bench mark. The claimed number is highly questionable because it would completely depend upon document characteristics, legal matter, selected search keys, employee's draft, file backup practice, and luck. For a

given search method and key matrix, the population will quickly vanish. This creates a deceptive phenomenon: the algorithm can get nearly all of the responsive documents by a few iterations. This merely means that documents containing those search keys can be quickly captured. When only 1% seed documents are used, the chance to get a particular singlet or doublet is very small. The chance of getting all of critical singlets, doublets, triplets and low-frequency documents are smaller than the chances to win a lottery or land a doomed plane. In reality, critical documents often comprise only 0.1-0.001% of the document population. Thus, such a recall rate cannot be used to predict critical documents. Smoking gun, reasons for doing a bad thing, approval of questionable transactions, personal reflection of an incident... most probably fall within this hard-to-draw and hard-to-search population. Due to how probabilities work, even some big series of documents may be missed due to bad luck in selecting search keys. The algorithm can capture a large number of document series. The portion of unreachable and highly uncertain documents are mixed with the bulk volume of non-relevant documents. Nothing can be done to those documents. Such a production is useless in litigation. Thus, for technology-assisted review, high consistencies, recall rates and high precision rates comprise primarily the contribution of coding consistencies for the large number of duplicate and marginal documents, while the errors and conflicts in human review comprise primarily the coding differences for a large number of duplicates, marginally responsive, and meaningless documents. In other words, regardless of performance numbers, human review can do much better for coding low-yield, large-to-read, and hard-to-search documents.

Due to fundamental differences in coding methods between technologies-assisted review and human review, all performance metrics acquired for technology-assisted review cannot be compared with those obtained from pure human review. As discussed above, increasing review speeds will

increase coding errors primarily for long, difficult and critical documents. Based upon dramatic differences in document drafting practice, information nature, and reviewer skills, their differences are more than those between Apple and Orange.

B. Additional Specific Defects in Each of Questioned Technologies

Many of the ten foundational flaws are sufficient grounds for invalidating the three questioned technologies. However, each of the technologies has additional defects, as described below.

1. Search Key Methods

All search methods share two common problems: search cannot capture all responsive, useful, and helpful documents, and when a method is used to build a review pool, it severely interferes with document reviewers by changing the scope of the document pool.

Search methods generally are incapable of reliably capturing all documents important to litigation course and final result. They are incapable of reliably capturing short documents, troublesome documents, documents intended to keep secret, documents with characters-encoding problems, dialog-like communication, partial communication with information passed by context, complex documents, scanned graphs and images, and documents written in combination of two or more languages.

After the use of search methods is a public knowledge, everyone knows how to defeat this investigation method. As a result, this method has no utility for investigating white collar crimes and bribery anymore. It is very easy for people to avoid those keys. It is also easy to pass part of communication by using ambiguous terms, words, and names to avoid scrutiny.

If search methods are used in due diligence reviews for foreign language documents, they cannot properly address mixed use of two or more languages. It is not simply a mixture of two sets of vocabulary. They actually apply one language's rules onto words from another language. Mixed text cannot be searched by using ordinary search methods.

Both methods can negatively affect the performance of document reviewers primarily through disrupting document context. They may indirectly increase risk of exposure through over-inclusive production of non-responsive documents, which could be used by skilled persons in the subject to be combined with other known information to arrive at confidential business information.

2. Prioritized Review Method

Prioritized review method is an extension of a search method. It inheres all foundational flaws and all specific search defects, but also present its own problems of using improper document review orders and review schedules. While an early case assessment is important, a wrong assessment will serve no useful purpose. However, proponents fail to realize that case merit can change for any of a large number of possibilities. Case merit may hinge on one document, one statement in a document, one phrase in a statement, and one or more words in a definition in a document. Documents reviewed at a disorganized order and out of context may have little value for case assessment. A cursory review may cause attorneys to settle on a legal theory that the litigant cannot win. It is unrealistic to think that reviewing a few documents can provide meaningful information for case assessment.

If such review can provide useful information, it should be done dependent from a first review.

3. Predictive Coding

The validity of predictive coding depends upon many assumptions it uses. Proponents have to treat all documents with three properties: uniformity, independence, and representativeness. Documents are dramatically different in their sizes, subjects, technical types, difficulty levels, relevancy to disputed issues, privilege statuses, confidential natures, carried risks, and stakes to litigants and related persons. Considering differences between individual documents, “documents” is equivalent to “things” in the physical world. Things may comprise paper clips, cellular phones, TV units, airplanes, buildings, and carriers. Based this obvious analog, all methods used in validating predictive coding are invalid. The following is a summary of additional defects.

1. The ability to capture 75% documents mean that it may miss 25% of potentially relevant documents. It is common knowledge, a case may be decided by relying upon less than 0.1% documents. The uncertainty would provide 250 times of the room required for holding critical documents.

2. Initial coding seed documents were created by using only about 5% documents. For one single document to be drawn is only about 5% chance. The odds for a plurality of critical documents to be drawn in seed documents are very small.

3. Review of initial seed documents is performed in the first one or two weeks, it has the worse quality index because reviewers cannot acquire enough case information for understanding documents.

4. Patterns matching algorithms such as latent semantic indexing are used to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. Those algorithms are based on the principle that words that are used in the same contexts tend to have similar meanings. This main assumption cannot hold well in corporate documents because documents are created by a large of number of diverse employees who often do not have anything in common including way of ex-

pression, language patterns, and vocabulary. In addition documents are not created in the same context, It cannot achieve enough accuracy to have meaningful utility. It has more problems when a pattern-match algorithm is used in cases involving white collar crimes such as rooting, stealing, embezzlement, and bribery. No distribution theory can be used to model normal corporate behaviors and criminal conduct.

5. Computer algorithms do not have proper way for handling coding conflicts. Factors that affect coding include (1) constant changes leading to coding conflicts, (2) different review standards leading to different coding decisions for same materials, (3) differentiating coding thresholds for protecting non-responsive sensitive information, (4) coding conflicts from avoiding excessive number of duplicates, (5) differential coding due to document quality problems, language encoding issues, and different risk levels, (6) different coding precedence for different portions of materials, (7) potentially multiple ways of coding complex documents, (8) coding decisions that are based upon other factors such as custodians, time of review, external facts, and sufficiency of found similar documents.

6. A search method may be unable to capture (1) documents that have all kinds of text problems; (2) documents that are created in image files, especially, scanned image files; (3) documents that incorporate one or more unspecified antecedents; (4) documents that convey at least one piece of critical information by context or understanding; (5) documents with character encoding schemes that are different from that used in search keys; (6) documents that are intended to preserve secrecy; (6) documents that are intended to be partial communication; and (7) documents created for a complex and sophisticated scheme.

7. It is presumed that a distinctive concept may appear in a document, appear in a document which contains other independent concepts, interact with other concepts, and be modified by other concepts. An algorithm

could not precisely understand (1) strengthening a concept, (2) weakening a concept, (3) corroborating a concept, (4) rejecting a concept, (5) denying a concept or some aspect of it, and (6) alter some aspects of a concept such as its significance, strength, or truth, or associated properties (e.g., someone else is responsible for doing something defined by the concept).

8. Its learning capability is very limited and cannot have a fraction of human intelligence. It is unable to deal with changes that are inherent in litigation. It cannot “understand” any things which are not reflected in documents such as drafting histories and disclosing histories. In many times, coding decisions can be changed back and forth without following any rule. In such situations, the algorithm cannot understand reasons for changes because they are not on the documents. If a change is to reverse a document definition scope back to an old scope as a result of a motion's ruling, some reviewers may continue to use prior definition scope. In another situation, tens of similar documents are coded as non-privileged, but a reviewer finds all previously coded tags are improper, and thus the determination of a correct coding decision cannot be based upon tag plurality. Some uninformed reviewers may continue to code documents using an obsolete definition, and thus the algorithm cannot choose coding decisions by coding times. The tags with highest frequencies can also be wrong because all reviewers can make same error. Tags coded at latest times may be wrong too if the last reviewer did not get an update. Predictive coding probably can make awkward imitations.

9. Predictive coding methods are validated under degraded coding environments intended for their acceptance. They have created unreviewable conditions so that reviewers had no choice but to accept preexisting tags. Bias can be found in the following situations: (1) a reviewer can make a safe decision by accepting preexisting tags; (2) when reviewers are driven by daily quota, the easiest way of meeting the quota is to accept preexisting

tags; (3) when context is deficient and defensible coding decisions are impossible, the safest option is not to disturb preexisting tags; (4) when reviewers lack knowledge, skills, patience, or time to analyze documents, they just do not touch preexisting tags. Some validation designs lack control.

10. Computer algorithm lacks utility for performing complex functions such as (1) understanding background technologies and technical matters, (2) associating various parts of documents in language context to make an intelligent judgment, (3) associating different parts of matters in the same document, (4) properly treating assumptions, obvious expression errors, obvious omissions, and improper abbreviations, (5) appreciating informal expressions and all kinds of secondary meanings such as connotation, implication, sarcasm, and hidden messages, (6) detecting inherent properties of matters, things, and events, (7) understanding human emotion, malice, and intention, (8) making connection between two or more things by times, persons, events, or concepts, and (9) recognizing handwritten notes and drawings. Human reviewers experienced or well trained in the subject can handle any of those mentioned discovery tasks and any of documentary issues. It is impossible for a simple algorithm to have those functions.

C. The Technologies Lack Functional Capability for Claimed Use

The proponents of those technologies take advantage of the nature of litigation: document production cannot be evaluated at the time of review and a long time lapse exists between document review, and case disposition and risk realization.

The legal profession has long used a completely different approach to addressing professional performance. First, most of state bars require adequate legal education. Required education and licensing schemes provide

some assurance that licensed professionals know those concepts and have the skills to address litigation problems. However, the profession also requires more than just inherent capability to address legal problems. It also make comprehensive ethical rules to bar conflicting dealings and require attorneys to do work with zeal and passion. Those measures are necessary because work product often cannot be seen, accurately evaluated, and final result may not be known many years later. Moreover, litigants generally understand the importance of relevant experience in the field. They know that finding a criminal lawyer for a criminal matter, finding an experienced patent lawyer for their inventions, and finding an experienced personal injury lawyer for their personal injuries.

When those technologies enter the profession to practice law, they must prove their capability. As I have shown that due to all of the foundational flaws, they have not shown real capability of performing claimed discovery tasks. Thus, the question is whether they can show any specific functions for addressing litigation-significant discovery issues. They have not.

(1) There is no method for determining important documents and unimportant documents. I can show there is no method for achieving what a licensed attorney can do in dealing with important discovery issues.

(2) There is no method for identifying and processing sensitive business information. All problems in search methods still exit.

(3) There is no method for identifying and processing risk-carrying documents. All problems in search methods still exist.

I can show how those technologies fail in each of a large number of discovery tasks.

Many reviews have been tainted by misuse of those technologies and their review products should be assumed to be invalid. For each tainted review, a corrective review, as conducted by current method, cannot remedy

prior deficiency. Factors that preclude possible fixes include context deficiency, division of case information, problems of new reviewers, and lost information on document-document interactions. Breaking up a review into three smaller reviews can screw up final work product. When documents are drawn into two or more document pools and reviewed by two or more teams at different times, many documents will be coded as non relevant and lost in forest. For a defective review, the only solution is to conduct a thorough first review without regarding prior work product.

D. Court Rulings Are Based upon Junk Science

Many courts have ruled on some aspects of those invalid technologies. None of the opinions has found that those technologies have gained general acceptance. Those rulings thus have limited impacts, and are not entitled to control other cases as precedents. Some opinions reflect limited challenges, some reflect failure of parties to raise critical issues, and others reflect judges' poor understanding of the subject matter. Nothing can change the truth that those claimed misused uses are built upon a foundation with many layers of flaws. The entire technological infrastructure sooner or later will collapse. Nothing can change the truth that those technologies are unable to perform nearly all critical discovery tasks. Nothing can change the reality that reviewers are able to capture only the documents that are relevant on face under the current review model. Nothing can change the reality that misuse of those technologies is responsible for unacceptable review quality. A super majority of contract attorneys, who actually review documents, would agree to those assessments.

One argument is always how litigants and judges have gotten tied of human coding inconsistencies and “errors.” They do not know what are re-

ally human errors, judgment difference, and inherent problems of the discovery system. This argument reflects their poor understanding of the discovery system, which has dynamical, uncertain, heterogeneous, different, troublesome, and highly interactive properties. Consistencies are something they can find in widget production shops where they can pick up any unit to measure it.

E. Technology-assisted Review Dramatically Degrades Quality Index

We must realize that discovery for litigation is entirely different from any widget production system in all aspects except number count. Considering litigation dynamics, legal theory and document interactions, document characteristics, information nature, risk-realizing probability, stake of impacts, information specificity, human knowledge, document reviewability etc, a rational approach is to simulate the behaviors of hypothetical documents which would affect case outcome.

1. Simulations of Human Review Performance For a Hypothetical Model

In contrast to the whole review industry, the number-one factor I must consider is reviewer skills and training. I choose to use “quality index” as a measure of litigation performance or review quality which was once well known in the legal profession. Two qualities that are most important to review quality index are review experience and level of review. If time is fixed, a new team with no experience may ruin the case at the highest probability, a team of moderately experienced reviewers may improve quality index, and a team of highly experienced reviewers may get the best result. Experience levels may affect performance in all three risk areas.

While the knowledge of a reviewer relative to the subject of discovery is the most important factor, the review quality of a given reviewer depends upon time the reviewer spends. The reviewer may read only subject lines and highlighted words, selectively read text surrounding some highlighted words, read the entire document, read the document plus conduct some researches, read the entire documents, plus conduct some research, and plus reread some parts. Review time can dramatically differ between buzzword-focused cursory review and a thorough review. Quality index is not correlated to review time by a simple linear equation. At the time of review, extra time always seems to be wasted because this unique distribution pattern of the litigation-significant documents. The reviewer must eliminate junk, but cannot do so by a cursory review of each document. For example, for a large number of cases, only one or more documents for every one thousand may have real impact on the current case (however, business and secondary risks are entirely different matters). The whole review team must properly take care of all of those significant documents. If the entire review team actually reviews all documents in the pool, their attempts are definitely rewarded by a final work product which properly addresses all those critical documents.

Now, I conduct a simple simulation to see how different review teams affect the probabilities to capture a few hypothetical critical documents. The model has the following parameters: three critical documents that are independent of each other: one concerning subject of litigation, one containing company's policy that authorizes employees to create any false stories for external communications (“policy fraud”), and one containing a statement which is in direct conflict with one element in the legal theory. Each document occurs in extremely low frequency. Assume that the three documents are reviewed by three reviewers independently, respectively, by five review teams consisting of untrained, inexperienced, experienced, expert,

and perfect reviewers. Capturing probabilities of each reviewer are, respectively, 0.00, 0.25, 0.50, 0.90 and 1.0 for the five review teams. I compute the outcomes for getting zero document, one document, two documents, and three documents for each team. I also computer mathematical expectations for each review team. Results are shown in the following table.

Table 1. Effects of Individual Reviewer Abilities on Capturing Three Critical Documents:

Catch Doc No.	Untrained Reviewers (p=0.00)	Inexperienced Reviewers (p=0.25)	Experienced Reviewers (p=0.50)	Expert Reviewers (p=0.90)	“Perfect” Reviewer (p=1.00)
0	1.00	0.4219	0.125	0.001	0
1	0	0.4219	0.375	0.027	0
2	0	0.1406	0.375	0.243	0
3	0	0.0156	0.125	0.729	1.00
Math Exp.	0	0.75	1.50	2.70	3.00

The probabilities in the above table are exactly like what I would expect. The untrained layperson reviewers miss all three, but the “perfect” reviewers, which do not exist in the world, would capture all three with certainty. Anyone would expect is that the probabilities for capturing all three dramatically increase with their individual ability to recognize them. For example, inexperienced review team most probably get nothing or one, while an experienced review team at proper review pace get two or three in the highest probabilities. The mathematical expectations are from 0 to 3 from no experience to completely experienced, as expected. This seems to sup-

port the approach that law firms and attorneys historically paid a great attention to their staff skills.

Several extensions can be made from those simple simulations. Mathematical expectation is the product of a particular document number by the capturing probability of each reviewer. If a review team comprises a plurality of reviewers with differing capture probabilities from 0.25 to 0.90, the mathematical expectation of this team for three documents would be a value between the lowest number 0.75 to the highest number of 3.00. Also, the distribution rule is equally applicable to any reasonable number of critical documents. For 30 critical documents, mathematical expectations for the five hypothetical review teams would be, respectively, 0, 7.5, 15, 27 and 30.

One can also explore the effects of two rounds of full scale review on the capturing probabilities. This would depend upon review rule used at the review site, as shown below.

Case 1. Reviewers are allowed to tag critical documents but are not allowed to revert critical decisions back to lower significance tags. Thus, a reviewer may be able to re-capture critical documents from previously marked as non-relevant document pool. Capturing probability for each critical document is simply p=1-(1-p)^N, where N is the rounds of reviews. For review team with p=0.90, the capturing probability of two rounds of reviews would be 0.99. For N=5, p=0.99999. When a law firm conducts five rounds of review, it has near unit probability to capture each of the critical documents. This result tends to support the wisdom that law firms love to keep reviewing documents. They are fighting for the best chance.

Case 2. In reality, a site review rule is that reviewers can change prior coding tags back and forth freely. The final round of review can overrule all prior tags, it seems that such a review wastes the client's money because only the final round of review counts. This is overly simplified state-

ment. Prior coded tags do affect the coding decisions of next round reviewers. If the second-round reviewer has to overturn a prior coding decision, the second reviewer would have to find more convincing reason for doing so. If the first-round reviewer and the second-round reviewer have a similar skill level, the additional challenges forced upon the second-round reviewer would raise the quality of coding decisions. However, real benefits are that all reviewers can gain more case knowledge by reading documents on the second round, and thus can see same documents in ways they would not in the first-round review.

In the first-round review, reviewers can learn very little about the case before the review is completed. If they have stayed on the case for some time, they know much more. If they have stayed in the case for six months, they know persons, transactions, incidences, and even personalities and their stories. After they stayed in a case for two to three years, they know all kinds of subtle details. When a review team has reached that level understanding of case knowledge, the law firm can be very confident to bring the case for trial. If anything comes up, any document reviewer can provide required help such as finding evidence for new issues, identifying required documents, finding newly raised facts, or helping the law firm to formulate supplemental theories. Review duration is one of the most important quality index for contentious cases and it is directly proportional to the rounds of reviews. Problem is how to use extended review in a way most favorable to client returns. Understanding the difficulties of document review and the huge stake of certain cases, ten rounds of review are not excessive, but costs can be very high.

2. The Seed Generation Step Seriously Degrades Human Review

Now, I do simple simulations for predictive coding and human review. Assuming that seed documents are captured at 1% frequency, the probabil-

ity can be estimated by frequency (due to the large population, the effect of sampling size can be ignored) rather than full permutations.

Table 2. Effects of Random Selection of Seed Documents on the Chance to Capture Three Hypothetical Critical Documents:

Catch Doc No.	Select 10, 000 Docs from A Million Doc (p=0.01)	Probabilities Split by* Human Reviewers (each with p)	Moderately Experienced Reviewers (p=0.50)	Expert Reviewers (p=0.90)	"Perfect" Reviewers (p=1.00)
0	0.970299	NA	0.98507488	0.97324227	0.9703
1	0.029403	P; (1-p)	0.01485038	0.02651619	0.02940
2	0.000297	(1-p)(1-p); 2p(1-p); p*p	0.00007463	0.00024081	0.000297
3	0.000001	(1-p)(1-p)(1-p); 3*p(1-p)(1-p); 3*p*p(1-p); p*p*p	0.00000013	0.00000073	0.000001
Math Exp.	0.03097	NA	0.01599	0.02797	0.03097

The first column shows the probability for each of four possible outcomes of getting 0, 1, 2 and 3 documents. The result is terrible. The most likely outcome is missing all. The mathematical expectation is 0.03. This step has already made the whole review a total waste.

The low capturing probabilities in the first column from the seed generation step are then split as follows. If the algorithm has captured one doc-

ument, the reviewer may capture it with p or miss it with (1-p). If the algorithm captures two documents, the review team may miss all, capture one, or two. If the algorithm captures three documents, the review team may miss all, capture one, two or three. No matter how well the reviewers are, the final performance must be worse than what the seed step gives. Thus, the combination effect is that overall probabilities would be far worse than the result from human review alone (compared with the data in Table 1). It would be lucky to get just one of the three documents. The mathematical expectation indicates that a random drawing would get only 0.03097 documents. When the generated pool is further reviewed by a review team, the final mathematical expectations would be further lowered to the range from 0.01599 to 0.03097. Due to the interference, human review can never be better than this miserable performance. When the upstream has such a severe restriction, the skill levels of reviewers no longer have meaningful effect. It is reasonably expected that the process is going to screw up a super majority of critical documents. This is virtually consistent with the hiring trend of finding anyone who can read buzzwords and click tags at the lowest pay rates. Also, additional rounds of review could not change the performance in a meaningful way.

When the number of critical documents increases, the performance of the model becomes worse and worse. Assuming the document pool has 100 critical singlet documents, each reviewer of a moderately experienced team has a missing rate of 30%, the team still can capture 70 (since each document has 70% independent chance to get caught). If 1% documents are drawn as seed documents, the probability of capturing any one of those documents by random drawing could be only 1% (neglect the drawing effects on the pool size). The chance to get 70 can only be a dream. In a random drawing case, it is a typical ball-drawing model, where drawing a small number (a fraction) of balls from a million balls, the chance of capturing N

red balls will quickly reduce with the value of N. When a search method is used, it only gets the documents that contain at least one key, as shown below.

3. Search Methods also Seriously Degrades Human Review

In practice, the initial seed documents are generated by using search methods. The above simulations are not applicable to the cases where a search method is used. In this model, seed documents are captured by searching using a key matrix. Captured documents are then reviewed by humans. The probabilities entirely depend upon search algorithm, selection of keys, original authors' attempts to avoid those keys, the nature of the three critical documents, plus a long list of other factors. When authors have made an effort to avoid scrutiny, the search algorithm has no utility. Only God knows the right keys for such documents. Also, for a good population of highly critical documents, it is impossible to formulate keys for capturing them.

Table 3. Effects of a Search on the “Probabilities” to Capture the Three Critical Documents:

Catch Doc No.	The Actual Results From Searches	Probabilities Split by* Reviewers, each with p	Human Reviewers (p=0.50)	Human Reviewers (p=0.90)	Perfect Reviewers (p=1.00)
0	0	NA	0.50	0.10	0
1	1	P; (1-p)	0.50	0.90	1
2	0	(1-p)(1-p); 2p(1-p); p*p	0	0	0

3	0	(1-p)(1-p)(1-p); 3*p(1-p)(1-p); 3*p*p(1-p); p*p*p	0	0	0
Math Exp.	NA (1)	NA	0.50	0.90	1.00

Use of a search method can produce highly unpredictable results. Among the three hypothetical documents, one concerns the matter of litigation, it must have some search keys. However, the other two documents do not contain any expected keys for capture. Of course, this table can be easily modified to include estimated probabilities from search results rather the actual outcomes (use the method in Table 2). One should expect (1) it is highly unpredictable, (2) it is impossible to capture certain critical documents, and (3) human review will further lower the probabilities as in the second case. Search key deficiency and authors' effort to avoid commonly expected keys could totally defeat the effort finding right keys. This example also shows that humans cannot do better than what they have in their start point.

While a search method for generating seed documents do improve its chance for getting most obvious relevant documents, it would be a miracle to get all critical documents. The chance of getting all critical documents would be much lower than the chance at which humans can get in linear review.

4. Contextual Deficiency, Review Order, and Short Review Duration Further Degrade the Quality of Human Review

In addition to the effect of the seed selection step, predictive coding, prioritized review methods, and search methods also have additional impacts. Context deficiency and document review orders can further reduce review quality. Even if some of the critical documents are in the seed documents, the chance for reviewers to properly code is substantially diminished due to poor reviewability. Reduced project duration also degrades review quality. All studies have neglected the human' factor which is most important. They assumed that humans can do anything at computer set speeds. None of reviewers has the God's ability to read documents and code them perfectly. All reviewers are in two disadvantaged situations. They generally lack general knowledge in the field of litigant's technologies and lack specific knowledge of everything mentioned in any documents. It requires time for each reviewer to change their understanding from interpreting terms by generic meanings to interpreting teams by context-specific meanings. Only at that level of understanding, reviewers can reliably relate document contents to legal issues. Many projects are reviewed in a duration shorter than the minimum time required for them to learn basic case facts. Thus, they can only conduct a cursory review.

Now, wish was that predictive coding had a magic power to capture the documents in iteration reviews by concept co-existence and luck. Unfortunately, even if the algorithm captures a few documents by luck, no reviewer may be aware of their significance. All reviewers are gone in a few weeks. In the end, the documents that are caught by luck are still lost in forest.

5. Wasted Attempts by Humans Are Not Really Wasted

One argument for reducing review documents is that most documents are irrelevant. The reality is that each of the critical documents is captured by making hundreds of “wasted” attempts. Wasted attempts are not what we can dispense with. In dealing with critical documents, review perfor-

mance depends upon the team's efforts and the team's performance. In a case with only one smoking gun, one cannot foretell which reviewer, among a team of reviewers, will encounter it. If each reviewer can recognize it, the team would have a unity probability to get it. If none of the reviewers can recognize it, the team will miss it with certainty. If half reviewers in the team can recognize it, but half cannot, the chance of capturing it may be about 50%. The precise probability depends upon if same number of documents are allocated to each reviewer (Reviewers may work various hours in the project duration). In order to get it, everyone has to review each document carefully and cannot blindly code any of them. One cannot use hindsight result to prove that work by all reviewers are unimportant because they did not capture this document. The required capability of the team and joint efforts are the guarantee for best outcome.

The argument on futile attempts seems to be in conflict with that the tag counts and document numbers. However, there is a clear difference. It is impossible to tell which attempts are worthy tries and which are not. In comparison, the effects of documents on litigation course and final results are definite in at least some point of litigation. The impacts of those documents follow their own rules, as shown in a large number of court disposition opinion.

There is no method for getting those critical documents by any technologies. Human review can deliver unsurpassed review quality. In protecting litigant's business information, a team's effort of capturing a particular type documents can be defeated by one reviewer's omission if quality control is conducted by drawing a sample. This is why linear quality control is a must. This ensures that each document is in front of two reviewers. When a predictive coding method is used, the leaking-out of privileged documents are presumed because of the use of flawed quality control method and lack of human attention to production.

6. Results of Simple Simulations Are Consistent with Reality

Those simple simulations are explanatory of the general impression of legal professionals. For litigation-significant documents, full human review is by far more reliable than predictive coding or search methods. Even if a review team is inexperienced and each reviewer has a high missing rate, it still has a fair chance to capture some critical documents. This general observation is equally applicable to other documents such as privileged documents, and documents that contain trade secrets and implicate secondary liabilities. I also firmly believe that this pattern is equally applicable to other discovery tasks and can be approved. In additional, full human review has a built-in self-correction mechanism. When a critical document has a limited duplicates, some reviewers may capture it while other reviewers miss it. Thus, errors or omissions are harmless as they can found those missed. Since human review depends upon individual ability and personal attention, each document can be captured at a reasonable chance. However, drawing documents by a small percent is an entirely different thing. The chances of getting a set of critical documents are dramatically reduced, as the number of critical documents increases. Even more interesting, since a computer algorithm treats all duplicate documents as same, there is no cross-checking capability. The algorithm either captures all duplicates or misses all. Those peculiar facts show why review quality of predictive coding is so low that litigation attorneys can note the differences by drawing a few documents at review sites. So, this is by no means a few percentage difference, which would require a control study to see. While they have not proved the problem, I have.

7. “Cost-of-Saving” Is A Bad Deal for Litigants

By misusing those technologies, a litigant can indeed save 90% costs that would be required by human review. However, the litigant does not get the same quality index. While I cannot quantify exact impact, those proba-

bility computations tend to show that technologies-assisted review could get only a fraction of quality index that would be achieved by an experienced review team. Given the critical importance of those documents, the impact is not proportional. What a litigant can get, unfortunately, is not even 10% return as in widget production, but perhaps a total loss plus collateral damages. There is no proportionality in e-discovery as in widget production: when a worker produces 10% widget volume, the factory gets 10% revenue. Those technologies are so simple, and there is no basis to think that they somehow have a magic power to draw 1-10% for human review, and can "amplify" human analysis to reach beyond quality and scope that are embodied in the seed documents. The "magic" capability of capturing up to 80% documents in a matter of hours is simple phenomenon of capturing a large number of series of duplicate documents that contain those search keys. This power is so deceptive that it forces the industry to form a region-like belief and makes everyone to chant after the crowd. After those documents are captured, nothing can be done to get those highly important, unique, short, and distinctive documents that are mixed with the bulk volume of non-responsive documents.

Reviewing 80% of non-responsive documents cannot be regarded as waste of time. For investigative reviews and discovery reviews for contentious cases, reviewers must acquire and understand (1) general knowledge of litigant's technologies and business practices, (2) specific knowledge of players, groups, divisions, things, transactions, and products and services..., (3) any aspect of business practices as references knowledge for understanding facts surrounding legal issues (without this reference baseline, reviewers cannot see fraud, bias, questionable conduct, and looting etc.), and (4) detailed information that may be used for reviewing other documents. Looting, embezzlement, and bribery are closely integrated with litigant's culture and business practices. Successful investigation depends

upon reviewers' actual ability to detect subtle deviations of business behaviors from routine business activities. Acquisition of this ability depends upon not only professional knowledge but also detailed case knowledge, relevant background information, and a understanding of reference practices. Successful use of this ability also requires endless tries which should never be measured by percents. It is not like a chip solder who can do the job with little to learn.

Given the overwhelming foundational flaws and contrary reality we see at review sites, no credable argument can be made that somehow those technologies can improve review quality. Technologies-assisting review ruins every aspect of quality index because they eliminate the chance for reviewers to read and force the entire review team to use a “lazy review method”. In addition, those technologies remove helpful documents, disrupt document review context, and shorten review period to a point that no meaningful review is possible. Other poor practices include imposition of quota, careless elimination of coding conflicts, and limited quality control, all further reducing review quality.

8. Comparison between Human Review and Technologies-assisted Review

While the actual outcome of human review depends upon reviewer experience levels and training, their performance in correctly recognizing documents can range from missing all to capturing all. Based upon capturing probabilities for various scenarios, deep-level two-round review by an experienced review team can properly capture 90% litigation significant documents; a moderate review by fairly experienced team may get most of those documents, but a cursory review driven by predictive coding may capture very few and will miss them even after reviewers capture them. The capturing mathematical expectation by intermediate skills is still about 50%. Two-round human review gives higher mathematical expectation. With two

rounds of review, a team of intermediate skills can capture any document at p=0.75.

The overall performance by predictive coding in capturing singlets, doublets, triplets, and low-frequent critical documents that contain only distinctive, unique, and hard-to-search concepts is EXTREMELY low or most probably non-existing. In a case where predictive coding is used, a perfect review team could do no better than what has been captured in the drawing or search step. The impact of subsequent human review is same: the team may miss all or capture all in the seed documents.

The chance of capturing critical documents is degraded by the seed generation step by one to two magnitudes, depending upon the percent of seed documents and selection method. Other factors that further degrade review quality include context deficiency, bad review orders, review quota, and lack of attention to captured documents. Chances are that real difference may be far more than what I can show in the numbers for those models. I have not considered the effects of influence of preexisting tags, distrusted relationship, abusive mistreatment, and low-pay rates. I would not count on unhappy reviewers to control future risks.

E. Conclusion

All e-discovery technologies are validated on the basis of two foundational flaws: use of consistency and tagging count to evaluate discovery performance when they have no utility to evaluate performance. By extending widget production models to document review, the technologies contain many additional foundational flaws.

Those technologies fail to take into account of nature of litigation, interactions between documents and legal theories, unique distribution of litigation-significant documents, which can be characterized by different information natures, different impacts, different risk-realization probabilities,

very high specificity, and many document characteristics. Those technologies improperly apply to documents all widget properties such as uniformity in amount of work, independence, and representativeness.

When those technologies are used to process documents in e-discovery, they result in very poor performance due to ten foundational flaws in sampling, measuring performance criteria, validation analysis, comparison tests, risk control, hypothesis tests, confidential intervals, and quality control. Beyond the magic power of being fast, cheap and capable of capturing up to 80% potentially responsive documents in short times is lack of predictable capability to process documents that are unique or distinctive in concept, contain short texts, contain only a single copy or limited duplicates, or are drafted to avoid search keys.

Technologies have the effects of dramatically degrade human review quality by one to three magnitudes. It has the effects of withholding evidence, interfering with government investigation, failing due diligence review, disseminating litigant's privileged documents, leaking out litigant business information, exposing secondary liabilities, and ruining prospective claims or defenses unpredictably. Overall, they can hurt both sides in unpredictable ways.

(a) Widget model for responsive and non-responsive documents