THE GROSSMAN-CORMACK GLOSSARY OF TECHNOLOGY ASSISTED REVIEW

(Version 1.02, November 2012)

Preamble

"Disruptive technology" is a term that was coined by Harvard Business School professor Clayton M. Christensen, in his 1997 book *The Innovator's Dilemma*, to describe a new technology that unexpectedly displaces an established technology. The term is used in the business and technology literature to describe innovations that improve a product or service in ways that the market did not expect, typically by designing for a different set of consumers in the new market and later, by lowering prices in the existing market. Products based on disruptive technologies are typically cheaper to produce, simpler, smaller, better performing, more reliable, and often more convenient to use. Technology assisted review ("TAR") is such a disruptive technology. Because disruptive technologies differ from sustaining technologies – ones that rely on incremental improvements to established technologies – they bring with them new features, new vernaculars, and other challenges.

The introduction of TAR into the legal community has brought with it much confusion because different terms are being used to refer to the same thing (*e.g.*, "technology assisted review," "computer-assisted review," "computer-aided review," "predictive coding," and "content based advanced analytics," to name but a few), and the same terms are also being used to refer to different things (*e.g.*, "seed sets" and "control sample"). Moreover, the introduction of complex statistical concepts, and terms-of-art from the science of information retrieval, have resulted in widespread misunderstanding and sometimes perversion of their actual meanings.

This glossary is written in an effort to bring order to chaos by introducing a common framework and set of definitions for use by the bar, the bench, and service providers. The glossary endeavors to be comprehensive, but its definitions are necessarily brief. Interested readers may look elsewhere for detailed information concerning any of these topics. The terms in the glossary are presented in alphabetical order, with all defined terms in capital letters.

In the future, we plan to create an electronic version of this glossary that will contain live links, cross references, and annotations. We also envision this glossary to be a living, breathing work that will evolve over time. Towards that end, we invite our colleagues in the industry to send us their comments on our definitions, as well as any additional terms they would like to see included in the glossary, so that we can reach a consensus on a consistent, common language relating to technology assisted review. Comments can be sent to us at mrgrossman@wlrk.com and gvcormac@uwaterloo.ca.

We hope you will find this glossary useful.

Maura R. Grossman
Wachtell, Lipton, Rosen & Katz
New York, New York

Gordon V. Cormack
University of Waterloo
Waterloo, Ontario

October 2012

Accept on Zero Error: A technique in which the training of a Machine Learning method is gauged by taking a sample after each training step, and deeming the training process complete when the learning method codes a sample with 0% Error (*i.e.*, 100% Accuracy).

Accuracy: The fraction of documents that are correctly coded by a search or review effort. Note that Accuracy + Error = 100%, and that Accuracy = 100% – Error. While high Accuracy is commonly advanced as evidence of an effective search or review effort, its use can be misleading because it is heavily influenced by Prevalence. Consider, for example, a document collection containing one million documents, of which ten thousand (or 1%) are Relevant. A search or review effort that identified 100% of the documents as non-Relevant and therefore, found *none* of the Relevant documents, would have 99% Accuracy, belying the failure of that search or review effort.

Active Learning: An Iterative Training regimen in which the Training Set is repeatedly augmented by additional documents chosen by the Machine Learning Algorithm, and coded by one or more Subject Matter Expert(s).

Agreement: The fraction of all documents that two reviewers code the same way. While high Agreement is commonly advanced as evidence of an effective review effort, its use can be misleading, for the same reason that the use of Accuracy can be misleading. When the vast majority of documents in a collection are Not Relevant, a high level of Agreement will be achieved when the reviewers agree that these documents are Not Relevant, irrespective of whether or not they agree that any of the Relevant documents are Relevant.

Algorithm: A formally specified series of computations that, when executed, accomplishes a particular goal. The Algorithms used in E-Discovery are implemented as computer software.

Area Under the ROC Curve ("AUC"): From Signal Detection Theory, a summary measure used to assess the quality of Prioritization. AUC is the Probability that a randomly chosen Relevant document is given a higher priority than a randomly chosen Non-Relevant document. An AUC score of 100% indicates a perfect ranking, in which all Relevant documents have higher priority than all Non-Relevant documents. An AUC score of 50% means the Prioritization is no better than random.

Artificial Intelligence: An umbrella term for computer methods that emulate human judgment. These include Machine Learning and Knowledge Engineering, as well as Pattern Matching (*e.g.*, voice, face, and handwriting recognition), robotics, and game playing.

Bag of Words: A Feature Engineering method in which the Features of each document comprise the set of words contained in that document. Documents are determined to be Relevant or Not Relevant depending on what words they contain. Elementary Keyword Search and Boolean Search methods, as well as some Machine Learning methods, use the Bag of Words model.

Bayes / Bayesian / Bayes' Theorem: A general term used to describe Algorithms and other methods that estimate the overall Probability of some eventuality (*e.g.*, that a document is Relevant), based on the combination of evidence gleaned from separate observations. In Electronic Discovery, the most common evidence that is combined is the occurrence of particular words in a document. For example, a Bayesian Algorithm might combine the evidence gleaned from the fact that a document contains the words "credit," "default," and "swap" to indicate that there is a 99% Probability that the document concerns financial derivatives, but only a 40% Probability if the words "credit" and "default," but not "swap," are present. The most elementary Bayesian Algorithm is Naïve Bayes, however most Algorithms dubbed "Bayesian" are more complex. Bayesian Algorithms are named after Bayes' Theorem, coined by the 18th century mathematician, Thomas Bayes. Bayes' Theorem derives the Probability of an outcome, given the evidence, from the Probability of the outcome absent the evidence, plus the Probability of the evidence, considering the outcome.

Bayesian Classifier / Bayesian Learning / Bayesian Filter: A colloquial term used to describe a Machine Learning Algorithm that uses a Bayesian Algorithm resembling Naïve Bayes.

Binomial Distribution: The Probability that a random sample from a large population will contain any particular number of Relevant documents, given the Prevalence of Relevant documents in the population. Used as the basis for Binomial Estimation.

Binomial Estimation / Binomial Calculator: A statistical method used to calculate Confidence Intervals, based on the Binomial Distribution, that models the random selection of documents from a large population. Binomial Estimation is generally more accurate, but less well known, than Gaussian Estimation. A Binomial Estimate is substantially better than a Gaussian Estimate (which, in contrast, relies on the Gaussian or Normal Distribution) when there are few (or no) Relevant documents in the sample. When there are many Relevant and many Non-Relevant documents in the sample, Binomial and Gaussian Estimates are nearly identical.

Blair and Maron: Authors of an influential 1985 study that showed that skilled paralegals, when instructed to find 75% or more of the Relevant documents from a document collection using search terms and iterative search, found only 20%. That is, the searchers believed they had achieved 75% Recall, but had in fact achieved only 20%. In the Blair and Maron study, the paralegals used an iterative approach, examining the retrieved documents and refining their search terms until they believed they were done. Many current commentaries incorrectly distinguish the Blair and Maron study from current iterative approaches, failing to note that the Blair and Maron subjects did in fact refine their search terms based on their review of the documents that were returned from the searches.

Boolean Search: A Keyword Search in which the Keywords are combined using operators such as "AND," "OR," and "[BUT] NOT." The result of a Boolean Search is precisely determined by the words contained in the documents. (*See also* Bag of Words method.)

Bulk Coding: The process of coding all members of a group of documents (identified, for example, by De-duplication, Near-deduplication, Email Threading, or Clustering) based on the review of only one or a few members of the group. Also referred to as Bulk Tagging.

Bulk Tagging: *See* Bulk Coding.

Classical or Gaussian Estimation / Classical, Normal, or Gaussian Calculator: A method of calculating Confidence Intervals based on the assumption that the quantities to be measured follow a Gaussian (Normal) Distribution. This method is most commonly taught in introductory statistics courses, but yields unreasonably large Confidence Intervals when the proportion of items with the characteristic being measured is small. (*C.f.* Binomial Estimation / Binomial Calculator.)

Clustering: An Unsupervised Learning method in which documents are segregated into categories or groups so that the documents in any group are more similar to one another than to those in other groups. Clustering involves no human intervention, and the resulting categories may or may not reflect distinctions that are valuable for the purpose of search and review.

Computer-Aided Review: *See* Technology Assisted Review.

Computer-Assisted Review: *See* Technology Assisted Review.

Concept Search: A marketing term used to describe Keyword expansion techniques, which allow search methods to return documents beyond those that would be returned by a simple Keyword or Boolean Search. Methods range from simple techniques like Stemming, Thesaurus Expansion, and Ontology search, through statistical Algorithms like Latent Semantic Indexing.

Confidence Interval: In Statistics, a range of values estimated to contain the true value, with a particular Confidence Level.

Confidence Level: In statistics, the chance that a Confidence Interval derived from a Random Sample will include the true value. For example, "95% Confidence" means that if we were to draw 100 independent Random Samples of the same size, and compute the Confidence Interval from each sample, 95 of the 100 Confidence Intervals would contain the true value. It is important to note that the Confidence Level is *not* the Probability that the true value is contained in any particular Confidence Interval; it is the Probability that the method of estimation will yield a Confidence Interval that contains the true value.

Confusion Matrix: A two by two table listing values for the number of True Negatives ("TN"), False Negatives ("FN"), True Positives ("TP"), and False Positives ("FP") resulting from a search or review effort. As shown below, all of the standard evaluation measures are algebraic combinations of the four values in the Confusion Matrix. Also referred to as a Contingency Table. An example of a Confusion Matrix (or Contingency Table) is provided immediately below.

	Truly Non-Relevant	Truly Relevant
Coded Non-Relevant	True Negatives (“TN”)	False Negatives (“FN”)
Coded Relevant	False Positives (“FP”)	True Positives (“TP”)

Accuracy = 100% – Error = (TP+TN) / (TP+TN+FP+FN)
Error = 100% – Accuracy = (FP+FN) / (TP+TN+FP+FN)
Elusion = 100% – Negative Predictive Value = FN / (FN+TN)
Fallout = False Positive Rate = 100% – True Negative Rate = FP / (FP+TN)
Negative Predictive Value = 100% – Elusion = TN / (TN+FN)
Precision = Positive Predictive Value = TP / (TP+FP)
Prevalence = Yield = Richness = (TP+FN) / (TP+TN+FP+FN)
Recall = True Positive Rate = 100% – False Negative Rate = TP / (TP+FN)

Content Based Advanced Analytics (“CBAA”): *See* Technology Assisted Review.

Contingency Table: *See* Confusion Matrix.

Control Set: A Random Sample of documents coded at the outset of a search or review process, that is separate from and independent of the Training Set. Control Sets are used in some Technology Assisted Review processes. They are typically used to measure the effectiveness of the Machine Learning Algorithm at various stages of training, and to determine when training may cease.

Culling: The practice of narrowing a large data set to a smaller data set for purposes of review, based on objective or subjective criteria, such as file types, date restrictors, or Keyword Search terms. Documents that do not match the criteria are excluded from the search and any further review.

Cutoff: A given score or rank in a prioritized list, resulting from a Relevance Ranking search or Machine Learning Algorithm, such that the documents above the Cutoff are deemed to be Relevant and documents below the Cutoff are deemed to be Non-Relevant. In general, a higher cutoff will yield higher Precision and lower Recall, while a lower cutoff will yield lower Precision and higher Recall. Also referred to as a Threshold.

Da Silva Moore: *Da Silva Moore* v. *Publicis Groupe*, Case No. 11 Civ. 1279 (ALC) (AJP), 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012), *aff’d* 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012). The first federal case to recognize Computer Assisted Review as “an acceptable way to search for relevant ESI in appropriate cases.” The opinion was written by Magistrate Judge Andrew J. Peck and affirmed by District Judge Andrew L. Carter.

Decision Tree: A step-by-step method of distinguishing between Relevant and Non-Relevant documents, depending on what combination of words (or other Features) they contain. A Decision Tree to identify documents pertaining to financial derivatives might first determine whether or not a document contained the word "swap." If it did, the Decision Tree might then determine whether or not the document contained "credit," and so on. A Decision Tree may be created either through Knowledge Engineering or Machine Learning.

De-duplication: A method of replacing multiple identical copies of a document by a single instance of that document. De-duplication can occur within the data of a single custodian (also referred to as Vertical De-Duplication), or across all custodians (also referred to as Horizontal De-duplication).

Document Population: The collection of documents or electronically stored information about which a Statistical Estimation may be made.

Electronic Discovery / E-Discovery: The process of identifying, preserving, collecting, processing, searching, reviewing and producing electronically stored information ("ESI") that may be Relevant to a civil, criminal, or regulatory matter.

Elusion: The fraction of documents identified as Non-Relevant by a search or review effort, that are in fact Relevant. Elusion is computed by taking a Random Sample from the Null Set and determining how many (or what Proportion of) documents are actually Relevant. A low Elusion value has commonly been advanced as evidence of an effective search or review effort (*see*, *e.g.*, *Kleen*), but that can be misleading because absent an estimate of Prevalence, it conveys no information about the search or review effort. Consider, for example, a document collection containing one million documents, of which ten thousand (or 1%) are Relevant. A search or review effort that found *none* of the Relevant documents would have 1% Elusion, belying the failure of the search. Elusion = 100% – Negative Predictive Value.

Email Threading: Grouping together email messages that are part of the same discourse, so that they may be understood, reviewed, and coded consistently, as a unit.

EORHB: *EORHB* v. *HOA Holdings LLC*, Civ. Action No. 7409-VCL, tr. and slip op. (Del. Ch. Oct. 19, 2012). The first case in which a court *sua sponte* directed the parties to use Predictive Coding as a replacement for Manual Review (or to show cause why this was not an appropriate case for Predictive Coding), absent any party's request to do so. Vice Chancellor J. Travis Laster also ordered the parties to use the same e-discovery vendor and to share a document repository.

Error / Error Rate: The fraction of all documents that are incorrectly coded by a search or review effort. Note that Accuracy + Error = 100%, and that 100% – Accuracy = Error. While a low Error Rate is commonly advanced as evidence of an effective search or review effort, its use can be misleading because it is heavily influenced by Prevalence. Consider, for example, a document collection containing one million documents, of which ten thousand (or 1%) are relevant. A search or review effort that found *none* of the relevant documents would have 1% Error, belying the failure of the search or review effort.

F1: The Harmonic Mean of Recall and Precision, often used in Information Retrieval studies as a measure of the effectiveness of a search or review effort, which accounts for the tradeoff between Recall and Precision. In order or achieve a high F1 score, a search or review effort must achieve both good Recall and good Precision.

Fallout: *See* False Positive Rate.

False Negative ("FN"): A Relevant document that is missed (*i.e.*, incorrectly identified as Non-Relevant) by a search or review effort. Also known as a Miss.

False Negative Rate ("FNR"): The fraction (or Proportion) of Relevant documents that are missed (*i.e.*, incorrectly identified as Non-Relevant) by a search or review effort. Note that False Negative Rate + Recall = 100%, and that 100% – Recall = False Negative Rate.

False Positive ("FP"): A Non-Relevant document that is incorrectly identified as Relevant by a search or review effort.

False Positive Rate ("FPR"): The fraction (or Proportion) of Non-Relevant documents that are incorrectly identified as Relevant by a search or review effort. Note that False Positive Rate + True Negative Rate = 100%, and that 100% – True Negative Rate = False Positive Rate. In Information Retrieval, also known as Fallout.

Feature Engineering: The process of identifying Features of a document that are used as input to a Machine Learning Algorithm. Typical Features include words and phrases, as well as metadata such as subjects, titles, dates, and file formats. One of the simplest and most common Feature Engineering techniques is Bag of Words. More complex Feature Engineering techniques include Ontologies and Latent Semantic Indexing.

Features: The units of information used by a Machine Learning Algorithm to classify documents. Typical Features include text fragments such as words or phrases, and metadata such as sender, recipient, and sent date. *See* also Feature Engineering.

Find Similar: A search method that identifies documents that are similar to a particular exemplar. Find Similar is commonly misconstrued to be the mechanism behind Technology-Assisted Review.

Gain Curve: A graph that shows the Recall that would be achieved for a particular Cutoff. The Gain Curve directly relates the Recall that can be achieved to the effort that must be expended to achieve it, as measured by the number of documents that must be reviewed and coded.

Gaussian Estimation / Gaussian Calculator: A method of Statistical Estimation based on the Gaussian Distribution.

Global Aerospace: *Global Aerospace Inc.* v. *Landow Aviation*, Consol. Case No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012). The first State Court Order approving the use of Predictive Coding by the producing party, over the objection of the requesting party, without prejudice to the requesting party raising an issue with the Court as to the completeness or the contents of the production, or the ongoing use of Predictive Coding. The Order was issued by Loudoun County Circuit Judge James H. Chamblin.

Harmonic Mean: The reciprocal of the average of the reciprocals of two or more quantities. If the quantities are named *a* and *b*, their Harmonic Mean is $\frac{2}{\frac{1}{a}+\frac{1}{b}}$. In information Retrieval, F1 is the Harmonic Mean of Recall and Precision. The Harmonic Mean, unlike the more common arithmetic mean (*i.e.*, average), falls closer to the lower of the two quantities.

Horizontal De-duplication: De-duplication of documents across multiple custodians. (*Cf.* Vertical De-duplication)

Index: A list of Keywords in which each Keyword is accompanied by a list of the documents (and sometimes the positions within the documents) where it occurs. Manual indices have been used in books for centuries; automatic indices are used in Information Retrieval systems to identify Relevant documents.

Indexing: The manual or automatic process of creating an Index. In Electronic Discovery, Indexing typically refers to the automatic constriction of an electronic Index for use in an Information Retrieval System.

In Re: Actos: *In Re: Actos (Pioglitazone) Products Liability Litigation*, MDL No. 6:11-md-2299 (W.D. La July 27, 2012). A product liability action with a Case Management Order ("CMO") that memorializes the parties' agreement on a "search methodology proof of concept to evaluate the potential utility of advanced analytics as a document identification mechanism for the review and production" of ESI. The search protocol provides for the use of a technology assisted review tool on the email of four key custodians. The CMO was signed by District Judge Rebecca F. Doherty.

Information Need: In Information Retrieval, the information being sought in a search or review effort. In E-Discovery, the Information Need is typically to identify documents responsive to a request for production, or to identify documents that are subject to privilege or work-product protection.

Information Retrieval: The science of how to find information to meet an Information Need. While modern Information Retrieval relies heavily on computers, the discipline predates the invention of computes.

Internal Response Curve: From Signal Detection Theory, a method of estimating the number of Relevant and Non-Relevant documents in a population, or above and below a particular

Cutoff, assuming that the scores yielded by a Learning Algorithm for Relevant documents obey a Gaussian Distribution, and the scores for Non-Relevant documents obey a different Gaussian Distribution.

Issue Codes: Subcategories of the overall Information Need to be identified in a search or review effort.

Iterative Training: The process of repeatedly augmenting the Training Set with additional examples of coded documents until the effectiveness of the Machine Learning Algorithm reaches an acceptable level. The additional examples may be identified though Judgmental Sampling, Random Sampling, or by the Machine Learning Algorithm, as in Active Learning.

Jaccard Index: The number of documents identified as Relevant by two reviewers, divided by the number of documents identified as Relevant by one or both. Used as a measure of consistency among review efforts. Also referred to as Overlap or Mutual F1. Empirical studies have shown that expert reviewers commonly achieve Jaccard Index scores of about 50%, and scores exceeding 60% are very rare.

Judgmental Sample / Judgmental Sampling: A method in which a sample of the document collection is drawn subjectively, so as to include the "most interesting" documents by some criterion. Unlike a Random Sample, the statistical properties of a Judgmental Sample may not be extrapolated to the entire collection. However, an individual (such as a quality assurance auditor or an adversary) may use judgmental sampling to attempt to uncover defects. The failure to identify defects may be taken as evidence (albeit not statistical evidence, and certainly not proof) of the absence of defects.

Keyword: A word that is used as part of a query for a Keyword Search.

Keyword Search: A search in which all documents that contain one or more specific Keywords are returned.

Kleen: *Kleen Products LLC* v. *Packaging Corp. of Am.*, Case No. 1:10-cv-05711 (N.D. Ill.) (Nolan, M.J.). A federal case in which plaintiffs sought to compel defendants to use Content Based Advanced Analytics ("CBAA") for their production, after defendants had already employed a complex Boolean Search to identify Responsive documents. Defendants advanced Elusion scores of 5%, based on a Judgmental Sample of custodians, to defend their use of the Boolean Search. After two days of evidentiary hearings before, and many conferences with, Magistrate Judge Nan R. Nolan, plaintiffs withdrew their request for CBAA, without prejudice.

Knowledge Engineering: The process of capturing the expertise of a Subject Matter Expert in a form (typically a Rule Base) that can be executed by a computer to emulate the human's judgment.

Latent Semantic Analysis: *See* Latent Semantic Indexing.

Latent Semantic Analysis ("LSA") / Latent Semantic Indexing ("LSI"): A Feature Engineering Algorithm that uses linear algebra to group together correlated Features. For example, "Windows, Gates, Ballmer" might be one group, while "Windows, Gates, Doors" might be another. Latent Semantic Indexing underlies many Concept Search tools. While Latent Semantic Indexing is used for Feature Engineering in some Technology Assisted Review tools, it is not, *per se*, a Technology Assisted Review method.

Linear Review: A document-by-document Manual Review in which the documents are examined in a prescribed order, typically chronological.

Logistic Regression: A state-of-the-art Supervised Learning Algorithm for Machine Learning that estimates the Probability that a document is Relevant, based on the Features that it contains. In contrast to the Naïve Bayes, algorithm, Logistic Regression identifies Features that discriminate between Relevant and Non-Relevant documents.

Machine Learning: The use of a computer Algorithm to organize or classify documents by analyzing their Features. In the context of Technology Assisted Review, Supervised Learning Algorithms (*e.g.*, Support Vector Machines, Logistic Regression, Nearest Neighbor, and Bayesian Classifiers) are used to infer Relevance or Non-Relevance of documents based on the Coding of documents in a Training Set. In Electronic Discovery generally, Unsupervised Learning Algorithms are used for Clustering, Near-Duplicate Detection, and Concept Search.

Manual Review: The practice of having human reviewers individually read and code a collection of documents for Responsiveness, particular issue codes, privilege, and/or confidentiality.

Margin of Error: The maximum amount by which a Statistical Estimate might likely deviate from the true value, typically expressed as "plus or minus" a percentage, with a particular Confidence Level. For example, one might express a Statistical Estimate as "30% of the documents in the population are Relevant, plus or minus 3%, with 95% confidence." This means that the Statistical Estimate is 30%, the Confidence Interval is 27% to 33%, and the Confidence Level is 95%. Using Gaussian Estimation, the Margin of Error is one-half of the size of the Confidence Interval. It is important to note that when the Margin of Error is expressed as a percentage, it refers to a percentage of the Population, not a percentage of the Statistical Estimate. In the current example, if there are one million documents in the Population, the Statistical Estimate may be restated as "300,000 documents in the Population are Relevant, plus or minus 30,000 documents, with 95% confidence." The Margin of Error is commonly misconstrued to be a percentage of the Statistical Estimate. However, it is incorrect to interpret the Confidence Interval in this example to mean that "300,000 documents in the Population are Relevant, plus or minus 9,000 documents." The fact that a Margin of Error of "plus or minus 3%" has been achieved is not, by itself, evidence of an accurate Statistical Estimate when the Prevalence of Relevant documents is low.

Miss: A Relevant document that is not identified by a search or review effort. Also referred to as a False Negative.

Miss Rate: The fraction (or proportion) of truly Relevant documents that are not identified by a search or review effort. Miss Rate = 100% – Recall. Also referred to as the False Negative Rate.

Mutual F1: *See* Jaccard Index.

Naïve Bayes: A Supervised Machine Learning Algorithm in which the relative frequency of words (or other Features) in Relevant and Non-Relevant training examples is used to estimate the likelihood that a new document containing those words (or other Features) is Relevant. Naïve Bayes relies on the simplistic assumption that the words in a document occur with independent Probabilities, with the consequence that it tends to yield extremely low or extremely high estimates. For this reason, Naïve Bayes is rarely used in practice.

NDLON: *National Day Laborer Organizing Network* v. U.S. *Immigration and Customs Enforcement Agency*, Case No. 10-Civ-3488 (SAS), 2012 WL 2878130 (S.D.N.Y.), a Freedom of Information Act ("FOIA") case in which District Judge Shira A. Scheindlin held that "most custodians cannot be 'trusted' to run effective searches because designing legally sufficient electronic searches in the discovery or FOIA contexts is not part of their daily responsibilities," and stated (in *dicta*) that "beyond the use of keyword search, parties can (and frequently should) rely on latent semantic indexing, statistical probability models, and machine learning to find responsive documents. Through iterative learning, these methods (known as 'computer-assisted' or 'predictive' coding) allow humans to teach computers what documents are and are not responsive to a particular FOIA or discovery request and they can significantly increase the effectiveness and efficiency of searches."

Near-Duplicate Detection: A method of grouping together "nearly identical" documents. A variant of Clustering in which the similarity among documents in the same group is very strong. Near-Duplicate Detection is typically used to reduce review costs, and to ensure consistent coding.

Nearest Neighbor: A Supervised Machine Learning Algorithm in which a new document is classified by finding the most similar document in the Training Set, and assuming that the correct coding for the new document is the same as the most similar one in the Training Set.

Negative Predictive Value ("NPV"): The fraction (Proportion) of documents in a collection that are not identified as Relevant by a search or review effort, and which are indeed Not Relevant. Akin to Precision, when the meanings of Relevant and Non-Relevant are transposed.

Non-Relevant / Not Relevant: In Information Retrieval, a document is considered Non-Relevant (or Not Relevant) if it does not meet the Information Need of the search or review effort. The synonym "irrelevant" is rarely used in Information Retrieval.

Normal / Gaussian Distribution: The "bell curve" of classical statistics. The number of Relevant documents in a sample tends to obey a normal distribution, provided the sample is large enough to contain a substantial number of Relevant and Non-Relevant documents. In this situation, Gaussian Estimation is reasonably accurate. If there are few Relevant documents in the sample, the Binomial Distribution better characterizes the number of Relevant documents in the sample, and Binomial Estimation is more appropriate.

Null Set: The set of documents that are not identified to be potentially Relevant by a search or review process; the set of documents that have been labeled as Not Relevant by a search or review process.

Ontology: A representation of the relationships among words and their meanings that is richer than a Taxonomy. For example, an Ontology can represent the fact that a wheel is a part of a bicycle, or that gold is yellow, and so on.

Overlap: *See* Jaccard Index.

Pattern Matching: The science of designing computer Algorithms to recognize natural phenomena like parts of speech, faces, or spoken words.

Positive Agreement: The Probability that, if one reviewer codes a document as Relevant, a second independent reviewer will also code the document as Relevant. Empirical studies show that Positive Agreement Rates of 70% are typical, and Positive Agreement rates of 80% are rare. Positive Agreement should not be confused with Agreement (which is a less informative measure) or Overlap (which is a numerically smaller but similarly informative measure). Under the assumption that the two reviewers are equally likely to err, Overlap is roughly equal to the square of Positive Agreement. That is, if Positive Agreement is 70%, Overlap is roughly 70% x 70% = 49%.

Positive Predictive Value ("PPV"): *See* Precision. Positive Predictive Value is a term of art in Signal Detection Theory; Precision is the equivalent term in Information Retrieval.

Precision: The fraction of documents identified by a search or review effort that are Relevant to the information request. Precision must be balanced against Recall. Also referred to as Positive Predictive Value.

Predictive Coding: A Technology Assisted Review process involving the use of a Machine Learning Algorithm to distinguish Relevant from Non-Relevant documents, based on Subject Matter Expert(s) coding of a Training Set of documents. *See* Supervised Learning and Active Learning.

Prevalence: The fraction of documents in a collection that are Relevant to an Information Need. Also known as Yield or Richness.

Probabilistic Latent Semantic Analysis: A variant on Latent Semantic Analysis based on conditional Probability rather than correlation.

Probability: The fraction (proportion) of times that a particular outcome would occur, should the same action be repeated under the same conditions an infinite number of times. For example, if one were to flip a fair coin, the Probability of it landing "heads" is one-half, or 50%; as one repeats this action indefinitely, the fraction of times that the coin lands "heads" will become indistinguishable from 50%. If one were to flip two fair coins, the Probability of both landing "heads" is one-quarter, or 25%.

Proportion: The fraction of a set of documents having some particular property (typically Relevance).

Proportionality: Pursuant to Federal Rule of Civil Procedure 26(b)(2)(C), the legal doctrine that electronically stored information may be withheld from production if the cost and burden of producing it exceeds its potential value to the resolution of the matter.

Quality Assurance: A method to ensure, after the fact, that a search or review effort has achieved reasonable results.

Quality Control: Ongoing methods to ensure, during a search or review effort, that reasonable results are being achieved.

Random Sample / Random Sampling: A subset of the Document Population selected by a method that is equally likely to select any document from the population for inclusion in the sample. Random Sampling is the basis of Statistical Estimation.

Recall: The fraction of Relevant documents that are identified by a search or review effort. Recall must be balanced against Precision.

Recall-Precision Curve: The curve representing the tradeoff between Recall and Precision for a given search or review effort, depending on the chosen Cutoff value.

Receiver Operating Characteristic Curve ("ROC"): In Signal Detection Theory, a graph of the tradeoff between True Positive Rate and False Positive Rate, as the Cutoff is varied.

Relevance Feedback: An Active Learning process in which the documents with the highest likelihood of Relevance are coded by a human, and added to the training set.

Relevance Ranking: A search method in which the results are ranked from the most likely through the least likely to be Relevant to an Information Need. Google search is an example of Relevance Ranking.

Relevant / Relevance: In Information Retrieval, a document is considered Relevant if it meets the Information Need of the search or review effort.

Responsiveness: A document that is Relevant to the Information Need expressed by a particular request for production or subpoena in a civil, criminal, or regulatory matter.

Richness: The fraction (or Proportion) of documents in a collection that are Relevant. Also known as Prevalence or Yield.

Rule Base: A set of rules created by an expert to emulate the human decision-making process for the purposes of classifying documents in the context of e-discovery.

Sample Size: The number of documents drawn at random that are used to calculate a Statistical Estimate.

Search Term: *See* Keyword.

Seed Set: The initial Training Set provided to the learning Algorithm in an Active Learning process. The documents in the Seed Set may be selected based on Random Sampling or Judgmental Sampling. Some commentators use the term more restrictively to refer only to documents chosen using Judgmental Sampling. Other commentators use the term generally to mean any Training Set, including the final Training Set in Iterative Training, or the only Training Set in non-iterative training.

Signal Detection Theory: Co-invented at the same time and in conjunction radar, the science of distinguishing true observations from spurious ones. Signal Detection Theory is widely used in radio engineering and medical diagnostic testing. The terms True Positive, True Negative, False Positive, False Negative, and Area Under the ROC Curve ("AUC") all arise from Signal Detection Theory.

Significance [of a statistical test]: The confirmation, with a given Confidence Level, of a prior hypothesis, using a Statistical Estimate. The Statistical Estimate is said to be "statistically significant" if all values within the Confidence Interval for the desired Confidence Level (typically 95%) are consistent with the hypothesis being true, and inconsistent with it being false. For example, if our hypothesis is that fewer than 300,000 documents are Relevant, and a Statistical Estimate shows that, 290,000 documents are Relevant, plus or minus 5,000 documents, we say that the result is significant. On the other hand, if the Statistical Estimate shows that 290,000 documents are Relevant, plus or minus 15,000 documents, we say that the result is not significant, because the Confidence Interval includes values (*i.e.*, the values between 300,000 and 305,000) that contradict the hypothesis.

Statistical Estimate / Statistical Estimation: The act of estimating the Proportion of a Document Population that have a particular characteristic, based on the Proportion of a Random Sample that have the same characteristic. Methods of Statistical Estimation include Gaussian Estimation and Binomial Estimation.

Statistical Sampling: A method in which a sample of the Document Population is drawn at random, so that statistical properties of the sample may be extrapolated to the collection.

Stemming: In Keyword or Boolean Search, or Feature Engineering, the process of equating all forms of the same root word. For example, the words "stem," "stemming," "stemmed," and "stemmable" would all be treated as equivalent, and would each yield the same result when used as a Search Terms In some search systems, stemming is implicit and in others, it must be made explicit through particular search syntax.

Subject Matter Expert(s): An individual (typically, but not necessarily, a senior attorney) who is familiar with the Information Need and can render an authoritative opinion as to whether a document is Relevant or not.

Supervised Learning: A Machine Learning method in which the learning Algorithm infers how to distinguish between Relevant and Non-Relevant documents using a Training Set. Supervised Learning can be a stand-alone process, or the final step in Active Learning,

Support Vector Machine: A state-of-the-art Supervised Learning Algorithm for machine learning that separates Relevant from Non-Relevant documents using geometric methods (*i.e.*, geometry). Each document is considered to be a point in [hyper]space, whose coordinates are determined from the Features contained in the document. The Support Vector Machine finds a [hyper]plane that best separates the Relevant from Non-Relevant Training documents. Documents outside the Training Set (*i.e.*, uncoded documents from the Document Population) are then classified as Relevant or not, depending on which side of the [hyper]plane they fall on. Although a Support Vector Machine does not calculate a Probability of Relevance, one may infer that the classification of documents closer to the [hyper]plane is more doubtful than those that are far from the [hyper]plane. In contrast the Naïve Bayes, and similar to Logistic Regression, Support Vector Machines identify Features that discriminate between Relevant and Non-Relevant documents.

Taxonomy: A hierarchical organization scheme that arranges the meanings of words into classes and subclasses. For example, Fords and Chryslers are cars; cars, trucks, and bicycles are vehicles, and vehicles, aircraft, and ships are modes of transportation.

Technology Assisted Review ("TAR"): A process for prioritizing or coding a collection of electronic documents using a computerized system that harnesses human judgments of one or more Subject Matter Expert(s) on a smaller set of documents and then extrapolates those judgments to the remaining Document Population. Some TAR methods use Algorithms that determine how similar (or dissimilar) each of the remaining documents is to those coded as Relevant (or Non-Relevant, respectively) by the Subject Matter Experts(s), while other TAR methods derive systematic rules that emulate the expert(s) decision-making process. TAR systems generally incorporate statistical models and/or sampling techniques to guide the process and to measure overall system effectiveness.

Term Frequency and Inverse Document Frequency ("TF-IDF"): An enhancement to the Bag of Words model in which each word has a weight based on term frequency – the number of times the word appears in the document – and inverse document frequency – the number of documents in which the word appears.

Thesaurus Expansion: In Keyword or Boolean Search, replacing a single Search Term by a list of its synonyms, as listed in a thesaurus.

Threshold: *See* Cutoff.

Training Set: A sample of documents coded by one or more Subject Matter Expert(s) as Relevant or Non-Relevant, from which a Machine Learning Algorithm then infers how to distinguish between Relevant and Non-Relevant documents beyond those in the Training Set.

TREC: The Text REtrieval Conference, sponsored by the National Institute of Standards and Technology ("NIST"), which has run since 1992 to "support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, the TREC workshop series has the following goals: to encourage research in information retrieval based on large test collections; to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas; to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems."

TREC Legal Track: From 2006 through 2011, TREC included a Legal Track, which sought "to assess the ability of information retrieval techniques to meet the needs of the legal profession for tools and methods capable of helping with the retrieval of electronic business records, principally for use as evidence in civil litigation."

True Negative ("TN"): A Non-Relevant document that is correctly identified as Non-Relevant by a search or review effort.

True Negative Rate ("TNR"): The fraction (or Proportion) of Non-Relevant documents that are correctly identified as Non-Relevant by a search or review effort.

True Positive ("TP"): A Relevant document that is correctly identified as Relevant by a search or review effort.

True Positive Rate ("TPR"): The fraction (or Proportion) of Relevant documents that are correctly identified as Relevant by a search or review effort.

Uncertainty Sampling: An Active Learning approach in which the Machine Learning

Algorithm selects the documents as to which it is least certain about Relevance, for coding by the Subject Matter Expert(s), and addition to the Training Set.

Unsupervised Learning: A Machine Learning method in which the learning Algorithm infers categories of similar documents without any training by a Subject Matter Expert. Examples of Unsupervised Learning methods include Clustering and Near-Duplicate Detection.

Validation: The act of confirming that a process has achieved its intended purpose. Validation may be statistical or judgmental.

Vertical De-duplication: De-duplication within a custodian; identical copies of a document held by different custodians are not De-duplicated. (*Cf.* Horizontal De-duplication)

Yield: The fraction (or Proportion) of a Document Population that are Relevant. Also known as Prevalence or Richness.