FEDERAL COURTS LAW REVIEW

Volume 7, Issue 1 2013

THE GROSSMAN-CORMACK GLOSSARY OF TECHNOLOGY-ASSISTED REVIEW

with

Foreword by John M. Facciola, U.S. Magistrate Judge

FOREWORD

"When *I* use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean—neither more nor less." "The question is," said Alice, "whether you *can* make words mean so many different things."

Lewis Carroll, *Through the Looking Glass, What Alice Found There*[1]

"A word is not a crystal, transparent and unchanged, it is the skin of a living thought and may vary greatly in color and content according to the circumstances and the time in which it is used."

Justice Oliver Wendell Holmes Jr., *Towne* v. *Eisner*, 245 U.S. 418, 425 (1918)

In the heels of the higgling lawyers, Bob,
Too many slippery ifs and buts and howevers,
Too much hereinbefore provided whereas,
Too many doors to go in and out of.

When the lawyers are through
What is there left, Bob?
Can a mouse nibble at it
And find enough to fasten a tooth in?

Carl Sandburg, *The Lawyers Know Too Much*[2]

1. THE COLLECTED STORIES OF LEWIS CARROLL 238 (Citadel Press 1994).
2. THE COMPLETE POEMS OF CARL SANDBURG 189 (Harcourt, Brace and Co., rev. ed. 1970).

It is always the words. Humpty Dumpty, Holmes, and Sandburg, who sensed the power of language instinctively, knew how quickly their meanings can slip away. What seemed clear when the contract or statute was drafted is now recondite. When the parties contracted to buy and sell a horse, did they mean a mare or a stallion? When the statute required that a lawyer be disbarred upon conviction of a crime of moral turpitude, did the legislature mean a lawyer who gets convicted of assault after a brawl in a bar? Lawyers and judges are mocked for their continued use of Latin but they know that it is so much easier to say "res judicata" and take advantage of the encrusted meaning of those words than to start fresh and try to improve on what they convey. Law books come and go but Black's Law Dictionary will always be around. The words, as Paul Simon might put it, keep "slip, sliding away."[3]

It is hard enough when the world in which the words are used remains static, like the farm on which the horse, be it mare or stallion, lives. But, what happens when the movement of technology radically transforms what a word might have once meant? What is the "original" of an e-mail? Is another e-mail a copy of it when the visible text is the same but the metadata created in its production by a computer, rather than a human being, is entirely different? What happens when the meaning of the words in a statute applied to a process that was in existence when the statute was enacted but now no longer exists? Some of the definitions in the Stored Communications Act,[4] enacted in 1986, may drive judges to distraction since they were premised on technology used in 1986 but is no longer and must be applied to new processes that no one knew would exist when the statute was enacted. What kind of words can be used in a statute or a rule that are capacious enough to hold their meaning despite unknown technological change but precise enough to convey a definite meaning? It may be hard to believe, but there was not even a clear indication in the Federal Rules of Civil Procedure until 2006 that "electronically stored information" was within the scope of what a party had to produce in

---

3. PAUL SIMON, *Slip Slidin' Away*, on THE ESSENTIAL PAUL SIMON (Warner Bros. 2007).

4. *See, e.g.*, The definition of "remote computing service" in 18 U.S.C. § 2711(2) ("the provision to the public of computer storage or processing services by means of electronic communications systems"). As Professor Orin Kerr explains, the statute "freez[es] into law the understanding of computer network use as of 1986." Orin S. Kerr, *A User's Guide to the Stored Communications Act, and a Legislator's Guide to Amending It*, 72 GEO. WASH. L. REV. 1208, 1214 (2004). In 1986, users would use remote computing services to outsource computing tasks such as storing extra files or processing data when doing so was beyond the capacity of their computers. *Id*. Given the storage and processing capacities of new computers and tablets, this kind of distant processing capacity has disappeared. In its wake, however, are the difficult questions of the application of the definitions in the Stored Communications Act to "cloud computing." *See* William Jeremy Robison, *Free at What Cost? Cloud Computing Privacy under the Stored Communications Act*, 98 GEO. L.J. 1195, 1210, 1212-13 (2010).

discovery.[5] Indeed, until 2006, the word "phonorecords" appeared in Federal Rules of Civil Procedure 34(a)[6] which must have, at the time,[7] mystified anyone under 35.

The pace of technological change makes the situation worse. That pace is astonishing. What was thought to be the impossible crossing of the Atlantic Ocean in a plane and the landing of a man on the moon occurred in my father's lifetime. Yet, George Washington's troops moved no faster than Caesar's, and it took thousands of years before human beings discovered how to transmit messages by the telegraph using electricity.[8] Until then, a messenger, whether from Marathon to Athens, or through the colonial towns of Massachusetts, had to deliver them by hand.

The legal system deals in words and the pace of technological process is creating billions of them on a nearly daily basis,[9] creating a set of problems that were unimaginable a few years ago.

The first problem is that this explosion of words has been matched by the ever-increasing capacity of machines to capture and preserve them. It is a simple fact that an iPod has much more memory, i.e., capability to store information indefinitely, than the first computer[10] which took up an entire room. Additionally, the cost of storage is diminishing. A one-terabyte drive can be purchased for about $100[11] and can hold what would otherwise be hundreds of thousands of pages of paper. Indeed, now the information can be kept on a distant server with the space rented from a vendor for that purpose and retrieved by use of the Internet ("cloud computing"). For the first time, it is cheaper for human beings to buy a new file cabinet and keep more paper than to clean the useless clutter out of the old file cabinet to make room for the new information.

---

5. *Compare* FED. R. CIV. P. 34(a)(1)(A) (West 2006 rev. ed.) (describing "any designated documents or electronically stored information—including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations—stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form" as discoverable) *with* FED. R. CIV. P. 34(a)(1)(A) (West 2006) (mentioning only "any designated documents (including writings, drawings, graphs, charts, photographs, phonorecords, and other data compilations from which information can be obtained, translated, if necessary, by the respondent through detection devices into reasonably usable form").

6. FED. R. CIV. P. 34(a)(1)(A) (West 2006).

7. Vinyl records are making a comeback; eight track tapes will not. *See, e.g.*, Brian Passey, *Vinyl Records spin back into vogue*, USATODAY.COM (Feb. 26, 2011).

8. *See* TOM STANDAGE, THE VICTORIAN INTERNET (Walker & Co. 1998).

9. *See* George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, 14-23 (2007).

10. *See, e.g.*, *What is a Mainframe Operating System?*, WISEGEEK, http://www.wisegeek.com/what-is-a-mainframe-operating-system.htm (last visited Nov. 19, 2012).

11. *See* AMAZON, http://www.amazon.com/s/ref=nb_sb_ss_i_1_7?url=search-alias%3Daps&field-keywords=1+terabyte+hard+drive&sprefix=1+terab%2Caps%2C227 (listing numerous options under $100 for a one-terabyte external hard drive).

This phenomenon has led to the consequence that litigants are confronted with the often horrifying costs of searching through immense amounts of data to find what they need. The Federal Rules of Civil Procedure contemplate a system of demand and production; plaintiff asks for all documents pertaining to the merger of companies A and B and defendant either objects or produces them. But, when there are now hundreds of thousands of documents that may meet the definition of "pertinent," how can defendant find them and not go bankrupt in the process? Of course, defendant may be the victim of its own failure to maintain a responsible record-keeping process in which a principled decision-making process guides what will be kept and what will be thrown out. And, as anyone knows who has ever cleaned out a closet or an old hard drive, keeping everything is no solution. It only increases the expense and cost of finding what you want or need. Nevertheless, the affordability of cheap storage has led too many entities in our society to be quickly overwhelmed by their inability to search for what they need, whether because they need it to run their business or because they must produce it in discovery. They may find that the cost of searching and producing is so great that settling the lawsuit may be the only way out of an otherwise impossible situation.

It is understandable that, nature and technology abhorring a vacuum, a new scientific methodology has emerged to aid in the collection and searching process. "Technology-Assisted Review," called by its nickname "Predictive Coding," describes a process whereby computers are programmed to search a large amount of data to find quickly and efficiently the data that meet a particular requirement. Computer science and the sciences of statistics and psychology inform its use. While it bruises the human ego, scientists, including the authors of this glossary we are publishing, have determined that machines are better at the task of making such discoveries than humans.[12] Lawyers love to think that there is no substitute for their reviewing each document page by page. Not only is there a substitute, but an improvement. It is now indubitable that technology-assisted review is an appreciably better and more accurate means of searching a set of data.[13] That is hardly surprising news to those judges and lawyers who have experienced the mind-numbing tedium of reviewing large data sets only to find that one is seeing the same e-mail

12. *See* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and Efficient Than Exhaustive and Manual Review*, XVII RICH. J.L. & TECH. 11 (2011).

13. *Id.*

chain again and again, or worse, wading through mountains of data and finding nothing of any pertinence to the case being litigated.

The great benefits of technology-assisted review, however, bring in new concerns and questions for judges and lawyers. In a paper universe, the manner in which a party searched through a file cabinet hardly raises any significant issues. If what was produced appeared to be what was demanded and there were no inexplicable gaps, that was that and the court and parties moved on to other things. Now, the methodology of the use of technology-assisted review may itself be in dispute, with the parties controverted to each other's use of a particular method or tool. Those controversies have already lead to judicial decisions that have to grapple with a wholly new way of searching and with scientific principles derived from the science of statistics or other disciplines.[14] Lawyers and judges once again have to learn a whole new vocabulary to resolve the emerging and inevitable battle of the "experts."

To aid in the creation of that vocabulary, we publish with pride the glossary created by Maura R. Grossman and Gordon V. Cormack who are two of the most respected and acknowledged experts in this field. We agree with them that the creation of a clear and common vocabulary is essential to a comprehension of the legal issues at stake.

We are particularly gratified that the authors announced that they intend the glossary to be interactive so that others can suggest additional clarifications, revisions, and additions. We are certain that, if the experience of the courts in the first few years of the information technology revolution is any guide, the learning curve will be steep and that it must be climbed quickly if courts are going to be able to resolve promptly the controversies before them at the least expense. We are equally certain that the bench and bar will find this glossary useful as this new science develops and grows.

The Editors of the Federal Courts Law Review

By John M. Facciola, U.S. Magistrate Judge

---

14. *See*, *e.g.*, EORHB, Inc. v. HOA Holdings, Civ. Ac. No. 7409-VCL (Del. Ch. Oct. 19, 2012); Kleen Prods. LLC v. Packaging Corp., Civ. No. 10C 5711, 2012 WL 4498465 at *84-85 (N.D. Ill. Sept. 28, 2012); In re Actos (Pioglitazone) Prods. Liab. Litig., MDL No. 6:11-md-2299 (W.D. La. July 27, 2012); Global Aerospace Inc. v. Landow Aviation, L.P., No. CL 61040 (Va. Cir. Ct. Apr. 23, 2012); Moore v. Publicis Groupe & MSL Group, No. 11 Civ. 1279 (ACL) (AJP), 2012 WL 607412 (S.D.N.Y. Feb. 24, 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 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 set" 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 intended 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 bench, the bar, and service providers. This 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 this glossary are presented in alphabetical order, with defined terms in capital letters.

We 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 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 TAR. Comments can be sent to us at mrgrossman@wlrk.com and gvcormac@uwaterloo.ca. Subsequent versions of this glossary will be available online at http://cormack.uwaterloo.ca/targlossary/.

The authors would like to acknowledge the helpful comments provided by Craig Ball, Michael Levine, Ralph Losey, Amir Milo, and

Keith Roland on an earlier draft of this work. We are very grateful to Magistrate Judge John M. Facciola for his enthusiastic support.

We hope that 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

January 2013

* The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

THE GLOSSARY

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 Population containing one million Documents, of which ten thousand (or 1%) are Relevant. A search or review effort that identified 100% of the Documents as Not 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).

Actos: *See* In Re: Actos.

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 Population 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 chance.

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: (i) the probability of the outcome, independent of the evidence; (ii) the probability of the evidence, given the outcome; and (iii) the probability of the evidence, independent of the outcome.

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

Bigram: An N-Gram where N = 2 (i.e., a 2-gram).

Binomial Calculator / Binomial Estimation: 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.

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 Estimate: A Statistical Estimate of a Population characteristic using Binomial Estimation. It is generally expressed as a Point Estimate accompanied by a Margin of Error and a Confidence Level, or as a Confidence Interval accompanied by a Confidence Level.

Blair and Maron: Authors of an influential 1985 study (David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMC'NS ACM 289 (1985)), showing that attorneys supervising skilled paralegals believed they had found at least 75% of the Relevant Documents from a Document Collection, using search terms and iterative search, when they had in fact found only 20%. That is, the searchers believed they had achieved 75% Recall, but had achieved only 20% Recall. In the Blair and Maron study, the attorneys and paralegals used an iterative approach, examining the retrieved Documents and refining their search terms until they believed they were done. Many current commentators incorrectly distinguish the Blair and Maron study from current iterative approaches, failing to note that the Blair and Maron searchers did in fact refine their search terms based on their review of the Documents that were returned in response to their queries.

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.)

Bulk Coding: The process of Coding all members of a group of Documents (identified, for example, by Deduplication, 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, Gaussian, or Normal Calculator / Classical, Gaussian, or Normal Estimation: 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 inaccurate Confidence Intervals when the Prevalence of items with the characteristic being measured is low. (*Cf.* Binomial Calculator / Binomial Estimation.)

Classifier / Classification / Classified / Classify: An Algorithm that Labels items as to whether or not they have a particular property; the act of Labeling items as to whether or not they have a particular property. In Technology-Assisted Review, Classifiers are commonly used to Label Documents as Responsive or Non-Responsive.

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 a search or review effort.

Code / Coded / Coding: The action of Labeling a Document as Relevant or Non-Relevant, or the set of Labels resulting from that action. Sometimes interpreted narrowly to include only the result(s) of a Manual Review effort; sometimes interpreted more broadly to include automated or semi-automated Labeling efforts. Coding is generally the term used in the legal industry; Labeling is the equivalent term in Information Retrieval.

Collection: *See* Document Collection.

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

Computer-Assisted Review (CAR): *See* Technology-Assisted Review.

Concept Search: An industry-specific term generally 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 such as Stemming, Thesaurus Expansion, and Ontology search, through statistical Algorithms such as Latent Semantic Indexing.

Confidence Interval: As part of a Statistical Estimate, a range of values estimated to contain the true value, with a particular Confidence Level.

Confidence Level: As part of a Statistical Estimate, the chance that a Confidence Interval derived from a Random Sample will include the true value. For example, "95% Confidence" means that if one were to draw 100 independent Random Samples of the same size, and compute the Confidence Interval from each Sample, about 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.

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

Accuracy = 100% – Error = (TP + TN) / (TP + TN + FP + FN)
Elusion = 100% – Negative Predictive Value = FN / (FN + TN)
Error = 100% – Accuracy = (FP + FN) / (TP + TN + FP + FN)
Fallout = False Positive Rate = 100% – True Negative Rate = FP / (FP+TN)
False Negative Rate = 100% – True Positive Rate = FN / (FN+TP)
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 = Sensitivity = TP / (TP+FN)
True Negative Rate = Specificity = TN / (TN + FP)

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.

Crossover Trial: An Experimental Design for comparing two search or review processes using the same Document Collection and Information Need, in which one process is applied first, followed by the second, and then the results of the two efforts are compared. (*Cf.* Parallel Trial.)

Culling: The practice of narrowing a larger data set to a smaller data set for the purposes of review, based on objective criteria (such as file types or date restrictors), or subjective criteria (such as Keyword Search Terms). Documents that do not match the criteria are excluded from the search and from 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 the word "credit," and so on. A Decision Tree may be created through Knowledge Engineering or Machine Learning.

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

Dimensionality Reduction: A Feature Engineering method used to reduce the total number of Features considered by a Machine Learning Algorithm. Simple Dimensionality Reduction methods include Stemming and Stop Word elimination. More complex Dimensionality Reduction methods include Latent Semantic Indexing and Hashing.

Document: In the context of Electronic Discovery, a discrete item of Electronically Stored Information that may be the subject or result of a search or review effort.

Document Collection: The process of gathering Electronically Stored Information for search, review, and production; the set of Documents resulting from such a process. In many cases, the Document Collection and Document Population are the same; however, it is important to note that Document Population refers to the set of Documents over which a particular Statistical Estimate is calculated, which may be the entire Document Collection, a subset of the Document Collection (e.g., the documents with a particular file type or matching particular Search Terms), a superset of the Document Collection (e.g., the universe from which the Document Collection was gathered), or any combination thereof.

Document Population: The set of Electronically Stored Information or Documents about which a Statistical Estimate may be made.

Early Case Assessment (ECA): An industry-specific term generally used to describe a variety of tools or methods for investigating and quickly learning about a Document Collection for the purposes of estimating the risk(s) and cost(s) of pursuing a particular legal course of action.

EDI Study: *See* JASIST Study.

EDI-Oracle Study: An ongoing initiative (as of January 2013) of the Electronic Discovery Institute to evaluate participating vendors' search and document review efforts using a Document Collection contributed by Oracle America, Inc.

Electronic Discovery / E-Discovery: The process of identifying, preserving, collecting, processing, searching, reviewing, and producing Electronically Stored Information that may be Relevant to a civil, criminal, or regulatory matter.

Electronically Stored Information (ESI): Used in Federal Rule of Civil Procedure 34(a)(1)(A) to refer to discoverable information "stored in any medium from which the information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form." Although Rule 34(a)(1)(A) references "Documents or Electronically Stored Information," individual units of review and production are commonly referred to as Documents, regardless of the medium.

Elusion: The fraction of Documents identified as Non-Relevant by a search or review effort that are in fact Relevant. Elusion is estimated 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 it quantifies only those Relevant Documents that have been *missed* by the search or review effort; it does not quantify the Relevant Documents *found* by the search or review effort (i.e., Recall). Consider, for example, a Document Population containing one million Documents, of which ten thousand (or 1%) are Relevant. A search or review effort that returned 1,000 Documents, none of which were Relevant, would have 1.001% 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 either party's request to employ Predictive

Coding. 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 Population 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.

ESI: *See* Electronically Stored Information.

Experimental Design: A standard procedure accepted in the scientific community for the evaluation of competing hypotheses. There are many valid experimental designs. Some that can be appropriate for evaluating Technology-Assisted Review processes include Crossover Trials and Parallel Trials.

F_1: 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 to achieve a high F_1 score, a search or review effort must achieve *both* high Recall and high 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, dates, and file types. One of the simplest and most common Feature Engineering techniques is Bag of Words. More complex Feature Engineering techniques include the use of Ontologies and Latent Semantic Indexing.

Features: The units of information used by a Machine Learning Algorithm to Classify or Prioritize 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 Calculator / Gaussian Estimation: *See* Classical, Gaussian, or Normal Calculator / Classical, Gaussian, or Normal Estimation.

Gaussian Distribution: *See* Normal Distribution.

Gaussian Estimate: A Statistical Estimate of a Population characteristic using Gaussian Estimation. It is generally expressed as a Point Estimate accompanied by a Margin of Error and a Confidence Level, or as a Confidence Interval accompanied by a Confidence Level.

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 Court Judge James H. Chamblin.

Global Deduplication: Deduplication of Documents across multiple custodians. Also referred to as Horizontal Deduplication. (*Cf.* Vertical Deduplication.)

Gold Standard: The best available determination of the Relevance or Non-Relevance of all (or a sample) of a Document Population, used as benchmark to evaluate the effectiveness of a search and review effort. Also referred to as Ground Truth.

Goodhart's Law: An observation made in 1975 by Charles Goodhart, Chief Adviser to the Bank of England, that statistical economic indicators, when used for regulation, become unreliable. Restated and generalized in 1997 by University of Cambridge Professor Marilyn Strathern as "When a measure becomes a target, it ceases to be a good measure." Within the context of Electronic Discovery, Goodhart's Law suggests that the value of Information Retrieval measures such as Recall and Precision may be compromised if they are prescribed as the definition of the reasonableness of a search or review effort.

Grossman and Cormack: Authors of the JOLT (a.k.a. Richmond Journal) Study.

Ground Truth: *See* Gold Standard.

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, F_1 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. As a summary measure, a Harmonic Mean may be preferable to an arithmetic mean because a high Harmonic Mean depends on both high Recall and high Precision, whereas a high arithmetic mean can be achieved with high Recall at the expense of low Precision, or high Precision at the expense of low Recall.

Hashing / Hash / Hash Value: A statistical method used to reduce the contents of a Document to a single, fixed-size, alphanumeric value, which

is, for all intents and purposes, unique to a particular Document; the single, fixed-size alphanumeric value resulting from Hashing a particular Document. Common Hashing Algorithms include, but are not limited to, MD5, SHA-1, and SHA-2. Hashing and Hash Values are typically used for Document identification, Deduplication, or ensuring that Documents have not been altered.

Horizontal Deduplication: *See* Global Deduplication. (*Cf.* Vertical Deduplication.)

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 the Documents that contain particular Search Terms.

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

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 computers.

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 Electronically Stored Information. The search protocol provides for the use of a Technology-Assisted Review tool on the email of four key custodians. The CMO was issued by District Judge Rebecca F. Doherty.

Internal Response Curve: From Signal Detection Theory, a tool for estimating the number of Relevant and Non-Relevant Documents in a Population, or the number of Documents that fall above and below a particular Cutoff. The use of Internal Response Curves for this purpose

assumes that the scores yielded by a Machine Learning Algorithm for Relevant Documents obey a Gaussian Distribution, and the scores for Non-Relevant Documents obey a different Gaussian Distribution. These distributions are then used to predict the number of Relevant and Non-Relevant Documents in any given range of scores.

Interval Sample / Interval Sampling: *See* Systematic Sample / Systematic Sampling.

Issue Code(s) / Issue Coding: One or more subcategories of the overall Information Need to be identified in a search or review effort; the act of generating such subcategories of the overall Information Need. Examples include specification of the reason(s) for a determination of Relevance or Non-Relevance, Coding of particular subcategories of interest, and Coding of privileged, confidential, or significant ("hot") Documents.

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 through Judgmental Sampling, Random Sampling, or by the Machine Learning Algorithm, as in Active Learning.

Jaccard Index: A measure of the consistency between two sets (e.g., Documents Coded as Relevant by two different reviewers). Defined mathematically as the size of the intersection of the two sets, divided by the size of the union (e.g., the number of Documents coded as Relevant by both reviewers, divided by the number of Documents identified as Relevant by one or the other, or both reviewers). It is typically used as a measure of consistency among review efforts, but also may be used as a measure of similarity between two Documents represented as two Bags of Words. Jaccard Index is also referred to as Overlap or Mutual F_1. Empirical studies have shown that expert reviewers commonly achieve Jaccard Index scores of about 50%, and that scores exceeding 60% are rare.

JASIST Study: A 2009 study (Herbert L. Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y. FOR INFO. SCI. & TECH. 70 (2010)), showing that the Positive Agreement between each of two Technology-Assisted Review methods, and a prior production to the Department of Justice, exceeded the Positive Agreement between each of two Manual Review processes and the same production. Also referred to as the EDI Study.

JOLT Study: A 2011 study (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)), *available at* http://jolt.richmond.edu/v17i3/article11.pdf, that used data from TREC 2009 to show that two Technology-Assisted Review processes (one using Machine Learning and one using a Rule Base) generally achieved better Recall, better Precision, and greater efficiency than the TREC Manual Review process. Also known as the Richmond Journal Study, or the Richmond Study.

Judgmental Sample / Judgmental Sampling: A method in which a Sample of the Document Population is drawn, based at least in part on subjective factors, so as to include the "most interesting" Documents by some criterion; the Sample resulting from such method. Unlike a Random Sample, the statistical properties of a Judgmental Sample may not be extrapolated to the entire Population. 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 (or Search Term) that is used as part of a Query in a Keyword Search.

Keyword Expansion: *See* Query Expansion.

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

Kleen: *Kleen Prods. LLC v. Packaging Corp. of Am.*, Case No. 1:10-cv-05711, various Pleadings and Tr. (N.D. Ill. 2012). 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 complex Boolean Searches to identify Responsive Documents. Defendants advanced Elusion scores of 5%, based on a Judgmental Sample of custodians, to defend the reasonableness 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.

Label / Labeled / Labeling: *See* Code / Coded / Coding.

Latent Semantic Analysis (LSA): *See* Latent Semantic Indexing.

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. Also referred to as Latent Semantic Analysis.

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

Logistic Regression: A state-of-the-art Supervised Learning Algorithm that estimates the Probability that a Document is Relevant, based on the Features it contains.

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 the Documents in a Collection for Responsiveness, particular issues, privilege, and/or confidentiality.

Margin of Error: The maximum amount by which a Point 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 Point Estimate is 30%, the Margin of Error is 3%, 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 to a

percentage of the Point Estimate. In the current example, if there are one million Documents in the Document 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"; or, alternatively, "between 270,000 and 330,000 Documents in the Population are Relevant, with 95% confidence." The Margin of Error is commonly misconstrued to be a percentage of the Point Estimate. However, it would be 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 a precise Statistical Estimate when the Prevalence of Relevant Documents is low.

Miss / Missed: A Relevant Document that is not identified as Relevant 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 as Relevant by a search or review effort. Miss Rate = 100% – Recall. Also referred to as the False Negative Rate.

Model: *See* Statistical Model.

Mutual F_1: *See* Jaccard Index.

Naïve Bayes: A Supervised 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.

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. July 13, 2012), 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: An industry-specific term generally used to describe a method of grouping together "nearly identical" Documents. Near-Duplicate Detection is a variant of Clustering in which the similarity among Documents in the same group is very strong. It is typically used to reduce review costs, and to ensure consistent Coding. Also referred to as Near-Deduplication.

Near-Deduplication: *See* Near-Duplicate Detection.

Nearest Neighbor: A Supervised 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 that are identified as Non-Relevant by a search or review effort, that are in fact Non-Relevant. The complement of Precision; that is, Negative Predictive Value is computed the same way as Precision when the definitions of Relevant and Non-Relevant are transposed.

N-Gram: N consecutive words or characters treated as a Feature. In the phrase, "To be or not to be," a word Bigram (i.e., 2-gram) would be "to be"; a word Trigram (i.e., 3-gram) would be "to be or"; a Quad-Gram (i.e., 4-gram) would be "to be or not"; and so on. *See also* Shingling.

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 Distribution: The "bell curve" of classical statistics. The number of Relevant Documents in a Sample tends to obey a Normal (Gaussian) Distribution, provided the Sample size is large enough to capture a substantial number of Relevant and Non-Relevant Documents. In this situation, Gaussian Estimation is reasonably accurate. If the Sample size is insufficiently large to capture a substantial number of both Relevant and Non-Relevant Documents (as a rule of thumb, at least 12 of each), the Binomial Distribution better characterizes the number of Relevant Documents in the Sample, and Binomial Estimation is more appropriate. Also referred to as a Gaussian Distribution.

Null Set: The set of Documents that are not returned by a search process, or that are identified as Not Relevant by a 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, that gold is yellow, and so on.

Overlap: *See* Jaccard Index.

Parallel Trial: An Experimental Design for comparing two search or review processes using the same Document Collection and Information Need, in which both processes are applied concurrently but independently, and then the results of the two efforts are compared. (*Cf.* Crossover Trial.)

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

Point Estimate: The most likely value for a Population characteristic. When combined with a Margin of Error (or Confidence Interval) and a Confidence Level, it reflects a Statistical Estimate.

Population: *See* Document Population.

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 measure that conveys similar information). 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% × 70% = 49%.

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

Precision: The fraction of Documents identified as Relevant by a search or review effort, that are in fact Relevant. Also referred to as Positive Predictive Value.

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

Precision-Recall Tradeoff: The notion that most search strategies can be adjusted to increase Precision at the expense of Recall, or vice versa. At one extreme, 100% Recall could be achieved by a search that returned the entire Document Population, but Precision would be low (equal to Prevalence). At the other extreme, 100% Precision could be achieved by a search that returned a single Relevant Document, but Recall would be low (equal to 1/N, where N is the number of Relevant Documents in the Document Population). More generally, a broader search returning many Documents will have higher Recall and lower Precision, while a narrower search returning fewer Documents will have lower Recall and higher Precision. A Precision-Recall Curve illustrates the Precision-Recall Tradeoff for a particular search method.

Predictive Coding: An industry-specific term generally used to describe 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 Population that are Relevant to an Information Need. Also referred to as Richness or Yield.

Prioritization / Prioritized: *See* Relevance Ranking.

Probabilistic Latent Semantic Analysis: A variant of Latent Semantic Analysis based on conditional Probability rather than on 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 Rules of Civil Procedure 26(b)(2)(B), 26(b)(2)(C), 26(g)(1)(B)(iii), and other federal and state procedural rules,

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. Proportionality has been interpreted in the case law to apply to preservation as well as production.

Quad-Gram: An N-Gram where N = 4 (i.e., a 4-gram).

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.

Query: A formal search command provided as input to a search tool.

Query Expansion: The process of adding Search Terms to a Query to improve Recall, often at the expense of decreased Precision.

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

RAND Study: A 2012 study (Nicholas M. Pace & Laura Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*, RAND Institute for Civil Justice (2012)), indicating that Document review accounts for 73% of Electronic Discovery costs, and concluding that "[t]he exponential growth in digital information, which shows no signs of slowing, makes a computer-categorized review strategy, such as predictive coding, not only a cost-effective choice but perhaps the *only* reasonable way to handle many large-scale productions."

Ranking: *See* Relevance Ranking.

Recall: The fraction of Relevant Documents that are identified as Relevant by a search or review effort.

Recall-Precision Curve: *See* Precision-Recall Curve.

Recall-Precision Tradeoff: *See* Precision-Recall Tradeoff.

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 to the least likely to be Relevant to an Information Need; the result of such ranking. Google Web Search is an example of Relevance Ranking.

Relevance / Relevant: 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 an Information Need expressed by a particular request for production or subpoena in a civil, criminal, or regulatory matter.

Richness: *See* Prevalence or Yield.

Richmond Journal Study / Richmond Study: *See* JOLT Study.

Roitblat, Kershaw, and Oot: Authors of the JASIST (a.k.a. EDI) Study.

Rolling Collection / Rolling Ingestion: A process in which the Document Collection is periodically augmented as new, potentially Relevant Documents are identified and gathered. Whenever the Document Collection is augmented, the results of prior search or review efforts must be supplemented to account for the new Documents.

Rolling Production: A process in which Responsive Documents are delivered incrementally to a requesting party to provide timely, partial satisfaction of a Document request.

Rule: A formal statement of one or more criteria used to determine a particular outcome, e.g., whether to code a Document as Relevant or Non-Relevant.

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 Electronic Discovery.

Sample / Sampling: A subset of the Document Population used to assess some characteristic of the Population; the act of generating such a subset of the Document Population. *See* Interval Sample, Judgmental Sample, Random Sample, Statistical Sample, or Systematic Sample.

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

Search Term: *See* Keyword.

Sedona / Sedona Conference: The Sedona Conference® (https://thesedonaconference.org) is a nonprofit, 501(c)(3) research and educational institute, founded in 1997 by Richard G. Braman, dedicated to the advanced study of law and policy in the areas of antitrust, complex litigation, and intellectual property rights. Sedona sponsors a preeminent think-tank in the area of Electronic Discovery known as Working Group 1 on Electronic Document Retention and Production. Sedona is well known for its thoughtful, balanced, and free publications, such as *The Sedona Conference® Glossary: E-Discovery & Digital Information Management* (Third Edition, Sept. 2010), *The Sedona Principles Addressing Electronic Document Production, Second Edition* (June 2007), and *The Sedona Conference® Cooperation Proclamation* (July 2008).

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.

Sensitivity: *See* True Positive Rate.

Shingling: A Feature Engineering method in which the Features consist of all N-Grams in a text, for some number N. For example, the Trigram Shingling of the text "To be or not to be" consists of the features "to be or"; "be or not"; "or not to"; and "not to be." Note that the Features overlap one another in the text, suggesting the metaphor of roof shingles.

Signal Detection Theory: Invented at the same time and in conjunction with 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, Sensitivity, Specificity, Receiver Operating

Characteristic Curve, Area Under the ROC Curve, and Internal Response Curve, all arise from Signal Detection Theory.

Significance / Significant: The confirmation, with a given Confidence Level, of a prior hypothesis, using a Statistical Estimate. The result 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 the 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.

Specificity: *See* True Negative Rate.

Statistical Estimate: A quantitative estimate of a Population characteristic using Statistical Estimation. It is generally expressed as a Point Estimate accompanied by a Margin of Error and a Confidence Level, or as a Confidence Interval accompanied by a Confidence Level.

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

Statistically Significant / Statistical Significance: *See* Significance.

Statistical Model: A mathematical abstraction of the Document Population that removes irrelevant characteristics while largely preserving those of interest for a particular purpose. For the purpose of computing Recall, a Statistical Model need only consider whether or not the Documents are Relevant, and whether or not the Documents are Coded Relevant, not any other characteristics of the Documents.

Statistical Sample / 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 entire Document Population; the Sample resulting from such action.

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 Search Terms in a Query. In some search systems, stemming is implicit, and in others, it must be made explicit through particular Query syntax.

Stop Word: A common word that is eliminated from Indexing. Eliminating Stop Words from Indexing dramatically reduces the size of the Index, while only marginally affecting the search process in most circumstances. Examples of Stop Words include "a," "the," "of," "but," and "not." Because phrases and names such as "To be or not to be," and "The Who," contain exclusively Stop Words that would not be Indexed, they would not be identified (or identifiable) through a Keyword Search.

Subject Matter Expert(s): One or more individuals (typically, but not necessarily, attorneys) who are familiar with the Information Need and can render an authoritative determination 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 used repeatedly in an Active Learning process.

Support Vector Machine: A state-of-the-art Supervised Learning Algorithm 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 Relevant from Non-Relevant Training Examples. Documents outside the Training Set (i.e., uncoded Documents from the Document Collection) 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 less certain than for those that are far from the [hyper]plane.

Synthetic Document: An industry-specific term generally used to describe an artificial Document created by either the requesting party or the producing party, as part of a Technology-Assisted Review process, for use as a Training Example for a Machine Learning Algorithm. Synthetic Documents are contrived Documents in which one party imagines what the evidence might look like and relies on the Machine Learning Algorithm to find actual Documents that are similar to the artificial Document.

Systematic Sample / Systematic Sampling: A Sampling method in which every Nth Document (for some fixed number N) is selected, when the Documents are considered in some prescribed order; the Sample resulting from such action. A Systematic Sample is random (and hence a true Statistical Sample) only when the prescribed order is itself random. Sometimes referred to as an Interval Sample / Interval Sampling.

TAR: *See* Technology-Assisted Review.

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

Technology-Assisted Review (TAR): A process for Prioritizing or Coding a Collection of 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 Collection. Some TAR methods use Machine Learning Algorithms to distinguish Relevant from Non-Relevant Documents, based on Training Examples Coded as Relevant or Non-Relevant by the Subject Matter Experts(s), while other TAR methods derive systematic Rules that emulate the expert(s)' decision-making process. TAR processes generally incorporate Statistical Models and/or Sampling techniques to guide the process and to measure overall system effectiveness.

Term Expansion: *See* Query Expansion.

Term Frequency and Inverse Document Frequency (TF-IDF): An enhancement to the Bag of Words method 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 reciprocal of the number of Documents in which the word occurs.

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 Example: One Document from a Training Set.

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."

Trigram: An N-Gram where N = 3 (i.e., a 3-gram).

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. True Positive Rate is a term used in Signal Detection Theory; Recall is the equivalent term in Information Retrieval.

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 involve Statistical or Judgmental Sampling.

Vertical Deduplication: Deduplication within a custodian; identical copies of a Document held by different custodians are not Deduplicated. (*Cf.* Horizontal Deduplication.)

Yield: *See* Prevalence or Richness.