proofpoint.

# E-DISCOVERY ANALYTICS GLOSSARY

## AN EXPLANATION OF COMMONLY USED E-DISCOVERY TERMS

Legal pressures are evolving in the face of today's data growth. With new content types to support and ever-evolving court decisions, it's difficult to keep up with what you need to know. Legal and IT teams often speak different languages, contributing to the chaos. At Proofpoint, we understand that cross-collaboration is a key part of building a defensible litigation readiness strategy. This document will help bridge the gap between legal and IT teams by providing explanations of common terminology.

## ACCEPT ON ZERO

A statistical sampling procedure (an acceptance sampling procedure) that draws a random sample of objects from a population and checks each one to determine whether it is a defect. If none of the objects in the sample is found to be defective, then we can conclude with a specifiable level of confidence that there were no more than a specifiable proportion of defects in the original population. Finding zero defects in the sample does not mean that there were zero defects in the population, only that there were no more than a specifiable percentage. One application of this procedure in e-discovery is to draw a random sample from the population of documents determined by the review to be nonresponsive. The size of the sample is determined by your specified confidence level and by the maximum acceptable percentage of responsive documents that were not retrieved. If none of the documents in the sample is found to be responsive, then we can say with confidence X% that there were no more than Y% responsive documents left behind.

## ACCEPTANCE SAMPLING

Any of a family of sampling procedures used to evaluate whether a batch is of acceptable quality. In e-discovery, a batch could be a collection of documents that have been coded for responsiveness. Acceptance sampling determines whether the coding was adequately accurate.

## ACTIVE LEARNING

A form of supervised machine learning that presents for review or human categorization the documents with the highest current uncertainty, those documents that will be most informative about how to update the learning process.

## ALGORITHM

A specific set of steps that when accurately executed leads to a specific outcome. Algorithms can be created for many different kinds of processes, including calculation, data processing, automated reasoning, and mathematical computations. Algorithms should be distinguished from heuristics. The word “algorithm” is often misused to refer to any computer-implemented process.

## AUTO-CATEGORIZATION

The process of using machine learning or other rule-based systems for categorizing documents without direct human intervention. For example, emails may be auto-categorized as they arrive at an archive as to their retention period. The categories may be based on a taxonomy or ontology.

## BAG OF WORDS

An approach to modeling documents as unordered collections of words. In these models, the distribution, not the order, of words is important. For example, how many instances of each word type were contained in a document irrespective of order.

## BAYES' THEOREM

A mathematical formula or construction that relates prior probabilities and observed evidence to posterior probabilities. Let’s say that we know that about 7% of the documents in a collection are truly responsive. If we knew nothing else, we would infer that a randomly selected document would have a (prior) probability of 0.07 of being responsive. If we knew, though, that a given word in this document occurred twice as often in responsive as in nonresponsive documents, then the (posterior) probability that the document is responsive roughly doubles to 0.13.

## BAYESIAN CATEGORIZER

An information retrieval tool that computes the probability that a document is a member of a category from the probability that each word is indicative of each category. These estimates are derived from example documents. Uses the probability of each word given each category to compute the probability of each category given each word.

## BAYESIAN

Refers to statistical approach of Thomas Bayes, an 18th century mathematician and clergyman. Bayes wrote a theorem which shows how to calculate conditional probabilities from the combinations of observed events and prior probabilities. Many information retrieval systems implicitly or explicitly use Bayes’ probability rules to compute the likelihood that a document is relevant to a query.

## BIG DATA

An ill-defined term for large collections of data of various sorts. Big data may be big because it includes a large number of records (such as, all of the transactions on Amazon), because it includes a large number of variables (all of the characteristics or features that a bank knows about each customer), or both. Big data often suffers from the 4-Vs: Velocity, Variety, Volume, and Veracity. Big data accumulate very rapidly, they consist of many different kinds of data, at high volume, and the quality of the data often presents a challenge. Big data can also refer to very large collections of electronically stored information.

## BOOLEAN QUERY

An information retrieval query that employs Boolean logic. Queries are described in terms of logical relations among search terms. The basic relations are "AND," "OR," and "NOT." A query involving A AND B will retrieve documents that contain both terms. A query involving A OR B will return documents containing either term. A query involving A NOT B will return documents that contain A, but omit those documents that also contain B. Many systems allow longer expressions and some allow other relations (such as proximity operators) to be used.

## CAR

Computer assisted review. Using machine learning or other computational techniques to make the review of documents more accurate and faster. See also TAR, Machine Learning.

## CATEGORIZATION

The process of organizing objects (such as, documents) into specific classes or categories, such as responsive vs. nonresponsive.

## CLUSTERING

Organizing documents by similarity. Several information retrieval systems use computer algorithms to group similar documents together. Clustering usually uses unsupervised machine learning to identify similar documents. Many different clustering algorithms are available, most of them slow down significantly as the size of the document set gets larger.

## COLLECTION

A group of documents. These can be documents gathered for a particular matter or purpose. Information retrieval scientists tend use several well-known document collections (such as, RCV1) for testing and comparison purposes.

## COGNITIVE COMPUTING

A style of computing intended to mimic the way the human mind works. It is intended to address the kinds of problems where human judgment has previously been required, problems involving high amounts of ambiguity and uncertainty. Cognitive computing typically involves sophisticated forms of natural language processing and automatic reasoning.

## CONCEPT SEARCH

Being able to search for documents that are about the query, rather than just those that contain the query word. There are a number of technologies that can be used to capture the meaning. These include language modeling, latent semantic analysis, thesauri, ontologies, and taxonomies. All of these work by translating the query a user enters into an expanded query reflecting what is known about the meaning of the words in that query. They differ in how they perform this expansion.

## CONFIDENCE INTERVAL

The expected range of results. The confidence interval is used as an estimate of the uncertainty of a sample. One can use a sample to estimate some feature of a population. For example, the percentage of voters supporting a candidate. Because the sample is different from a complete measure of the population (the election result), it is an estimate and, as an estimate, is susceptible to error. One poll may estimate that a candidate has 45%, another random sample might estimate that the candidate is supported by 40% of voters. The true proportion of voters supporting a candidate will not be known until the actual votes are counted, but the sample is an estimate of that proportion. Generally, the larger the sample, the narrower the expected range of error in the sample.

### CONFIDENCE LEVEL

How often we would achieve a similar result if we repeated the same sampling process many times. If we repeatedly drew a random sample from the same population more than once, the confidence level would tell us how often we would get a result that is within a certain range (the confidence interval). Most scientific studies employ a minimum confidence level of 0.95, meaning that 95% of the time when you repeated the experiment, you would find a similar result. The higher the confidence level the larger the sample size that is required. Confidence level and confidence interval tell us about the quality of our sample estimate, not about the quality of the process they are measuring.

### CONJUNCTIVE NORMAL FORM

A formula for writing Boolean queries where clauses contain OR relations (called disjunctions) and are connected by AND relations (called conjunctions). Any query can be organized into conjunctive normal form, the form does not limit the expressions that can be represented. Using a normal (standard) form helps to simplify and organize the query, it does not limit the kinds of searches that can be done using it.

### CORRECT REJECTION

A term of information retrieval based on signal detection theory. A non-relevant document that has been categorized as non-relevant by the information retrieval system. A non-relevant document that has not been retrieved.

### CORPUS

A collection (body) of documents.

### CUSTODIAL DEDUPE

Removing all file duplicates from within an individual custodian. Contrast this with global dedupe and hyperdupe.

### DARK DATA

Data that are stored, perhaps in a data lake, in the hope that someday it might be useful to the organization. Dark data is typically big data that are unstructured, unanalyzed, uncategorized, and, most importantly, unused for any valuable business activity. Hoarded data, also called dusty data.

### DATA MINING

The process of extracting useful data from a volume of unstructured information. Data mining is used to search for patterns and systematic relationships in big data collections to extract other useful pieces of information from these collections.

### DEEP LEARNING

An approach to building and training neural networks. Deep learning typically involves a hierarchical neural network where each of the levels in the hierarchy is trained separately.

### DUPLICATE

A document that is identical to a specific document being considered. Hash functions are used to derive a digital digest of a document, which is usually a very large number. If two documents differ by even a single letter, then their digital digests will be very different. Therefore, two documents with the same digital digest are almost certain to be copies of one another.

### EARLY CASE ASSESSMENT

From a legal perspective, early case assessment means evaluating a case for its risks, merits, and value. Identifying the main case characteristics and the main significant participants. From a data analysis perspective, early case assessment has come to mean taking a quick look at the kind of data that are available for the case and using those data to begin preliminary aspects of discovery and estimate costs.

### EDRM MODEL

A map of the electronic discovery process from data preservation and collection to production. This map was designed through the collective effort of the Electronic Discovery Reference Model group.

### ELECTRONICALLY STORED INFORMATION

Information stored on computer disks or similar media consisting of the electronic representation of documents and other information types.

## ELUSION

An information retrieval measure of the proportion of responsive documents that have been missed. Most often used as a quality assurance measure in which a sample of non-retrieved documents is evaluated to determine whether a review has met reasonable criteria for completeness.

## ELUSION SAMPLING

A form of accept-on-zero testing, which is widely used in industrial quality control and in health evaluations. The basic idea is that one draws a random sample of documents from a population. If none of the items in that random sample is defective then we can say with confidence that there were not more than an acceptable number of defects in the population. When applied to e-discovery, this test is intended to ensure that the methods used to distinguish responsive from nonresponsive documents did not miss an unreasonable number of responsive documents.

## ESI

Electronically stored information.

## EXACT MATCH

Search results that correspond precisely to the query as specified. Contrast with fuzzy match and wild cards.

## F

Van Rijsbergen's F. A formula for combining precision and recall into a single number to make it easier to compare the information retrieval accuracy of different systems.

## F1

One form of van Rijsbergen's F formula for combining precision and recall into a single number to make it easier to compare the information retrieval accuracy of different systems. F1 is the weighted harmonic mean of precision and recall = 2 × precision × recall / (precision + recall). See F.

## FACETED QUERY

A search query wherein the system returns a set of alternative values that are typically subcategories of the original query. For example, on a website selling television sets, a user might enter an initial query consisting of the word "TV." The system will then present a list of various subclasses of TVs, categorized, for example, by the size of the screen. In e-discovery, a system might return a list of email senders in response to an initial query. A faceted query is a type of query expansion, where the expanded queries are categorized and can be selected.

## FEATURE

A characteristic of an item. In text, a feature is usually a word, but it could be a phrase or other grouping of words. In a search engine, features are the items that are specifically indexed.

## FALSE ALARM

An information retrieval term derived from signal detection theory. A nonresponsive document that is mistakenly assigned to the responsive group.

## FUZZY MATCH

Search results that correspond approximately to the query as specified. Fuzzy search is sometimes used to describe the use of wild cards, spelling variations, stemming, and other operations that result in approximate matches.

## GLOBAL DEDUPE

Removing all file duplicates, no matter where they occur in the collection. Contrast this with custodial dedupe and hyperdupe.

## HEURISTIC

A general practical approach to solving a problem that is useful to address the problem, but whose result is not guaranteed. Examples of heuristics include mental shortcuts, rules of thumb, and general strategies. Unlike an algorithm, a heuristic is not guaranteed to produce a specific result.

### HIT
An information retrieval term derived from signal detection theory. A hit is a responsive document that has been identified by the retrieval system as responsive.

### HYPERDUPE
The process of providing hyperlinks to all file duplicates wherever they occur in a collection. Rather than removing duplicates, hyperduping simply links all of the duplicates together so that they can be treated consistently. Compare to custodial and global deduping.

### INFORMATION
Signs, symbols, text, and so on about something. In the sense of information theory, information is the opposite of entropy. It is the reduction in uncertainty. Generally used in a generic sense to convey knowledge about something.

### INFORMATION RETRIEVAL
A generic term for using systems to find information. Typically used to refer to computational systems that help users to find information.

### INTERNET OF THINGS
The emerging idea that every-day things like refrigerators, garage door openers, and thermostats can and should be connected to the internet where they can communicate with one another, be remotely controlled, and do other tasks such as ordering ice cream when the current supply in the freezer is nearly gone. In order to be connected usefully to the internet, things need sensors, network connectivity, software, and processing capability to collect and exchange data.

### JUDGMENT SAMPLING
A method of sampling where the examiner's judgment is used to select objects for examination. This approach has unknown statistical characteristics so there is no way to use it to estimate characteristics of the total population with any reliability. On the other hand, it can potentially be effective at identifying some responsive documents that can be used as training examples for finding other responsive documents.

### KULLBACK-LEIBLER DIVERGENCE
A method of computing the similarity between two distributions. One distribution is derived from word frequencies in a document and the other distribution is derived from the word frequencies in the identified responsive or non-responsive (respectively) documents in the training set.

### LANGUAGE MODELING
Computing a model of the relationships among words in a collection. Language modeling is used in speech recognition to predict what the next word will be based on the pattern of preceding words. Language modeling is used in information retrieval and predictive coding to represent the meaning of words in the context of other words in a document or paragraph.

### LATENT SEMANTIC ANALYSIS
Latent semantic indexing (LSA) is a statistical method for finding the underlying dimensions of correlated terms. For example, words like law, lawyer, attorney, lawsuit, and so on all share some meaning. The presence of any one of them in a document could be recognized as indicating something consistent about the topic of the document. Latent Semantic Analysis uses statistics to allow the system to exploit these correlations for concept searching and clustering.

### LATENT SEMANTIC INDEXING
Latent semantic indexing (LSI) is the use of latent semantic analysis to index a collection of documents.

### LEXICOGRAPHIC ORDER
The order in which a collection of terms would appear in a dictionary.

### MACHINE LEARNING
A branch of computer science that deals with designing computer programs to extract information from examples. For example, properties that distinguish between responsive and nonresponsive documents may be extracted from example documents in each category. The goal is to predict the correct category for future untagged examples based on the knowledge extracted from the previously classified examples. Example approaches include neural networks, support vector machines, Bayesian classifiers and others.

### MISS
A term derived from signal detection theory, which refers to a target item that was not identified as a target item. An undetected responsive document.

### NATURAL LANGUAGE PROCESSING
A family of approaches to processing textual data according to the structure of language. Most often, natural language processing refers to the syntactic (grammatical) processing of language similar, for example, to having the computer diagram the sentences and then using these diagrams to infer something about the meaning of the sentence. Natural language processing can also refer to the semantics (meaning) of the words in a sentence.

### NEAR DUPE
Near dupe (or near duplicate) is a document that is almost the same as a specific target document. A small change to an email, such as sending back a copy with a new short message at the top, is a near dupe or near duplicate of the original email. Two versions of a contract may also be near duplicates. Unlike exact duplicates, near dupes do not yield the same or similar hash values.

### NEAREST NEIGHBOR CLASSIFICATION
A statistical procedure that classifies objects such as documents, according to the most similar item that has already been assigned a category label. This approach uses a set of labeled examples to classify subsequent unlabeled items, by choosing the category assigned to the most similar labeled example (its nearest neighbor) or examples. K-nearest neighbor classification uses the k most similar classified objects to determine the classification of an unknown object.

### N-GRAM
An information retrieval term. Statistics are often computed based on N-grams where an N-gram is a combination of N words or sometimes letters. When N = 1, it is also called a unigram. Statistics are computed for each word independently. When N = 2, it is also called a bigram. Statistics are computed for each adjacent pair of words.

### NEURAL NETWORK
A kind of machine learning where the basic forms of computation are derived from the characteristics of neurons, the cells that do the computations in the brain. Neural networks are used for a variety of functions, including organizing documents into categories based on the words they contain.

### NON-NEGATIVE MATRIX FACTORIZATION
A statistical technique related to neural networks, LSA, and PLSA for finding the underlying dimensions of meaning in a document collection. One of a family of statistical/mathematical techniques for finding simplified relationships between words and documents or between words and other words.

### OCR
Optical character recognition. The process of converting the optical image of the characters in a document into the computer codes for those characters.

### ONTOLOGY
A categorical or conceptual structure that may not be strictly hierarchical (cf. taxonomy). Concepts can be related to one another in complex ways. For example, an ontology may represent that lawyers, paralegals, and judges are associated with one another (one is not strictly a subset of the other).

## PLSA

See Probabilistic Latent Semantic Analysis.

## PRECISION

Precision is the proportion of retrieved documents that are responsive. See Recall.

## PREDICTIVE CODING

A family of evidence-based document categorization technologies that are used to put documents or ESI into matter-relevant categories, such as responsive/nonresponsive or privileged/nonprivileged. The underlying concept of using evidence to categorize objects has been around since the 18th century. Similar ideas have been applied to document classification or categorization since 1961 or earlier. The evidence that is used to categorize documents is typically the words in the documents.

## PRESERVATION

Maintaining documents in a usable form, preventing their destruction.

## PRINCIPAL COMPONENT ANALYSIS

A mathematical technique that summarizes the correlation between items. One of the techniques used in e-discovery as the basis for concept search, where the items are words.

## PROBABILISTIC LATENT SEMANTIC ANALYSIS

A statistical procedure for finding the underlying dimensions of correlated terms. Like Latent Semantic Analysis, this procedure attempts to capture the meaning shared by multiple terms to provide a concept search capability. It differs some from LSA in that it involves a slightly different statistical model. Also called probabilistic latent semantic indexing.

## PROXIMITY

Nearness. Search proximity operators, for example, find documents that contain words that are near to one another.

## QUERY

The set of search terms and operators submitted to a search system. A description of the required information, in the terms allowed by the system.

## QUERY BY EXAMPLE

The capability of searching a collection of documents to find more documents like the example.

## QUERY EXPANSION

The process of adding terms to those specified by the user in a query to give a broader range of terms. A thesaurus can be used, for example, to add synonyms to a query to increase the range of documents retrieved.

## RANDOM SAMPLING

The statistical process of choosing objects randomly, meaning that each object has an equal chance of being selected. There are several method to choosing a random sample.

## RECALL

Recall is the proportion of responsive documents in the entire collection that have been retrieved.

## RECALL-PRECISION GRAPH

A graph that shows the trade-off between precision and recall. Typically, the higher the recall level, the lower the precision level. In order to get more of the responsive documents, one usually has to accept more irrelevant documents.

## REDUPLICATE

The process of restoring duplicate files to the produced set. Duplicates may be removed from the review set using custodial or global deduplication.

### RELEVANCE
The degree to which a document is related to a query. Some kinds of search systems include information about the rank or intensity of the relation between a document and a query, some only provide binary information about whether a document is or is not related.

### RELEVANCE FEEDBACK
A class of machine learning techniques where users indicate the relevance of items that have been retrieved for them and the machine thereby learns to improve the quality of its recommendations.

### RELIABILITY
A statistical idea related to consistency and repeatability. The same measure taken in the same kind of situation should yield the same results if the system is reliable.

### RESULTS LIST
The items returned as the result of a query. Results lists are often ranked by relevance.

### RETRIEVAL
Identifying documents in the collection as potentially responsive. Generating a results list in response to a query.

### SAMPLING
The process of choosing objects from a population of objects where the chosen objects are intended to represent the whole population.

### SEMANTIC WEB
A proposal to equip the world wide web with metadata and tools so that information in documents can be usefully identified as to its content and intent. For example, appointment and calendar documents can have metadata that describe the appointment in a way that scheduling software could use it to set up conferences. The semantic web relies on ontologies to describe the usable categories of information.

### SEMANTIC
Having to do with meaning.

### SEMI-STRUCTURED DATA
Data in documents such as emails, that contain both structured and unstructured information. The metadata in an email are structured in that there are specific fields that are expected to contain specific information, such as the sender of the email. The text body of an email, however, is unstructured in that there is no specified organization for the content of the email message. May also include other kinds of documents with both fielded and free-text data.

### SENTIMENT ANALYSIS
A process to identify the sentiment in a text (for example, a Tweet, blog post, or document). Typically sentiment analysis identifies whether the text expresses a positive (such as, happiness) or negative (such as, anger) emotion, though more subtle distinctions are also possible.

### SINGULAR VECTORS
A statistical term. Singular vectors summarize highly complex data along a relatively small number of dimensions. Singular vectors represent the association structure of a set of the variables. In latent semantic analysis, for example, the singular vectors represent the fact that some words are more likely to occur with certain other words. For example, “lawyer,” “law,” “attorney,” and “counsel” are all related to one another. If one of them appears in a document, some of the others are also likely to appear. All of these words represent to varying degrees a single underlying concept having to do with the practice of law. One singular vector could represent this underlying concept.

### SINGULAR VECTOR DECOMPOSITION
Singular vectors are the basis of latent semantic indexing. They allow documents to be retrieved by meaning rather than just by whether they contain a keyword or not.

### SOCIAL NETWORK ANALYSIS
Investigations of who in an organization is communicating with whom. These connections are often displayed as a network diagram, with individuals as nodes and the emails or other communications between them as links. Social networks are often useful to determine how information has been flowing through an organization. They can also help to identify individuals with specific kinds of knowledge.

### SQL
Structured query language, an international standard language for querying databases.

### STEMMING
The process of removing prefixes and suffixes from words before indexing them and as part of query processing. For example, the word "swimming" could be stemmed to "swim." If words are stemmed as they are indexed, the query must also stem the words so that the query can match the index. In a system that uses stemming, several word forms can be indexed identically, for example, "swimmer" and "swimming" would both be indexed as "swim."

### STOPWORDS (ALSO STOP WORDS)
Very common words that usually have little value in a query. English words, such as "the," "and," "I," and others are so common, that they do not differentiate among documents. Many systems simply ignore these words when they are entered in a query. Some systems do not index these words.

### SUGGESTIVE CODING
See Predictive Coding.

### STRATIFIED SAMPLING
A form of random sampling in which the population is formed into subgroups or "strata." Objects in each group are sampled in the same proportion as the size of the group is to the whole population. Each object has an equal chance of being sampled, but stratified sampling also ensures that each group is included in the sample.

### STRUCTURED DATA
Data arranged in specific fields. The meaning of an entry can be discerned from the field in which it appears. Database records are usually structured data (though certain fields of a database record may contain unstructured data).

### SUPERVISED LEARNING
A kind of machine learning where the objects are labeled by an exterior source, typically, a subject matter expert. The goal of supervised learning is typically to replicate the decision pattern of the outside expert and apply the same patterns to previously unseen objects.

### SUPPORT VECTOR MACHINE (SVM)
A machine-learning approach, used for categorizing data. The goal of the SVM is to learn the boundaries that separate two or more classes of objects. Given a set of already categorized training examples, an SVM training algorithm identifies the differences between the examples of each training category and can then apply similar criteria to distinguishing future examples.

### TAR
Technology-assisted review. See also CAR, Machine Learning.

### TAXONOMY
A hierarchical representation of the concepts in a particular domain. In biology, the system of kingdoms, phyla, and so on, constitute a taxonomy. A taxonomy consists of a hierarchy of categories and subcategories and, perhaps, sub-subcategories, and so on.

### TF-IDF
In information retrieval, a weighting procedure so that some words in a query or document get emphasized more than others. A document is ranked higher using TF-IDF when it has more occurrences of the query term (TF or term frequency) and ranks lower when the word occurs in more documents (IDF or inverse document frequency). There are different rules for deciding how to combine TF with IDF. One common rule is to rank the documents based on the ratio of TF to log (IDF).

### THREADING
Organizing emails into conversational groups. For example, if John sends an email to Mary and she replies, both emails are part of the same conversational thread.

### TREC
The Text REtrieval Conference, an annual conference run by the National Institute of Standards and Technology to study information retrieval technologies and tasks. Since 2006, TREC has included a legal track intended to provide information about information retrieval in the legal space, primarily e-discovery.

### UNCERTAINTY SAMPLING
A machine learning process by which the user is presented objects to judge that are most likely to be uncertain, for example those that are equally likely to be classified as one category vs. another. How the user classifies these uncertain examples helps to identify where the boundary is between the two classes. In short, the system focuses on the difficult to classify objects during each stage of training.

### UNSUPERVISED LEARNING
A kind of machine learning where the objects are not labeled by an exterior source. Instead, the machine learning system organizes the objects based on implicit criteria that it derives. The selection of criteria is a function of the specific learning methods that are employed, the nature of the objects, and the way in which features of the object are represented. Clustering is an example of an unsupervised machine learning method. The goal of unsupervised learning is typically to identify hidden structure in unlabeled data, to summarize key features of the data.

### VALIDITY
A statistical term concerning the strength or truth of inferences or propositions. Valid measures of some characteristic are those that actually reflect the underlying difference effectively. For example, if we want to distinguish between responsive and nonresponsive documents, a valid measure is one that effectively makes this distinction (such as, the presence of certain words might reliably indicate a responsive document).

### VECTOR SPACE MODEL
In information retrieval, the vector space model is a method of representing documents and queries as an ordered list of numbers. Each word in the vocabulary of the collection has a position in this list. If the word is present in the document or query, then this position is set to a nonzero value. If the word is not present, then that position is set to zero. Related to the bag-of-words approach to representing documents. The order of the words in the document does not matter, but the number of times the word occurs may matter.

### WILD CARDS
Operators that allow query terms to be expanded, usually through completion. For example, in many systems, an asterisk signals that a word can be completed in any way. For example, “Plan*” would match “plan,” or “planning,” or “plant,” and others. See Fuzzy Match.

### WORD CLOUD
A method of displaying the most frequent words in a set of documents where the size of the word in the display corresponds to its frequency in the set of documents.

### WORDNET
An electronic thesaurus developed by George Miller and his students at Princeton University. Used by some systems to provide synonyms for query expansion.

## ABOUT PROOFPOINT

Proofpoint, Inc. (NASDAQ:PFPT), a next-generation cybersecurity company, enables organizations to protect the way their people work today from advanced threats and compliance risks. Proofpoint helps cybersecurity professionals protect their users from the advanced attacks that target them (via email, mobile apps, and social media), protect the critical information people create, and equip their teams with the right intelligence and tools to respond quickly when things go wrong. Leading organizations of all sizes, including over 50 percent of the Fortune 100, rely on Proofpoint solutions, which are built for today's mobile and social-enabled IT environments and leverage both the power of the cloud and a big-data-driven analytics platform to combat modern advanced threats.