Symantec™ Data Loss Prevention Vector Machine Learning Best Practices Guide

Version 11.1

Symantec

Symantec™ Data Loss Prevention Vector Machine Learning Best Practices Guide

Documentation version: 11.1, last updated June 9, 2011

Legal Notice

Symantec Corporation
350 Ellis Street
Mountain View, CA 94043

http://www.symantec.com

Technical Support

Symantec Technical Support maintains support centers globally. Technical Support's primary role is to respond to specific queries about product features and functionality. The Technical Support group also creates content for our online Knowledge Base. The Technical Support group works collaboratively with the other functional areas within Symantec to answer your questions in a timely fashion. For example, the Technical Support group works with Product Engineering and Symantec Security Response to provide alerting services and virus definition updates.

Symantec's support offerings include the following:

- A range of support options that give you the flexibility to select the right amount of service for any size organization
- Telephone and/or web-based support that provides rapid response and up-to-the-minute information
- Upgrade assurance that delivers automatic software upgrades protection
- Global support purchased on a regional business hours or 24 hours a day, 7 days a week basis
- Premium service offerings that include Account Management Services

For information about Symantec's support offerings, you can visit our web site at the following URL:

www.symantec.com/business/support/

All support services will be delivered in accordance with your support agreement and the then-current enterprise technical support policy.

Contacting Technical Support

Customers with a current support agreement may access Technical Support information at the following URL:

www.symantec.com/business/support/

Before contacting Technical Support, make sure you have satisfied the system requirements that are listed in your product documentation. Also, you should be at the computer on which the problem occurred, in case it is necessary to replicate the problem.

When you contact Technical Support, please have the following information available:

- Product release level

- Hardware information
- Available memory, disk space, and NIC information
- Operating system
- Version and patch level
- Network topology
- Router, gateway, and IP address information
- Problem description:
  - Error messages and log files
  - Troubleshooting that was performed before contacting Symantec
  - Recent software configuration changes and network changes

Licensing and registration

If your Symantec product requires registration or a license key, access our technical support web page at the following URL:

www.symantec.com/business/support/

Customer service

Customer service information is available at the following URL:

www.symantec.com/business/support/

Customer Service is available to assist with non-technical questions, such as the following types of issues:

- Questions regarding product licensing or serialization
- Product registration updates, such as address or name changes
- General product information (features, language availability, local dealers)
- Latest information about product updates and upgrades
- Information about upgrade assurance and support contracts
- Information about the Symantec Buying Programs
- Advice about Symantec's technical support options
- Nontechnical presales questions
- Issues that are related to CD-ROMs or manuals

Support agreement resources

If you want to contact Symantec regarding an existing support agreement, please contact the support agreement administration team for your region as follows:

Asia-Pacific and Japan	customercare_apac@symantec.com
Europe, Middle-East, and Africa	semea@symantec.com
North America and Latin America	supportsolutions@symantec.com

Additional enterprise services

Symantec offers a comprehensive set of services that allow you to maximize your investment in Symantec products and to develop your knowledge, expertise, and global insight, which enable you to manage your business risks proactively.

Enterprise services that are available include the following:

Managed Services	These services remove the burden of managing and monitoring security devices and events, ensuring rapid response to real threats.
Consulting Services	Symantec Consulting Services provide on-site technical expertise from Symantec and its trusted partners. Symantec Consulting Services offer a variety of prepackaged and customizable options that include assessment, design, implementation, monitoring, and management capabilities. Each is focused on establishing and maintaining the integrity and availability of your IT resources.
Education Services	Education Services provide a full array of technical training, security education, security certification, and awareness communication programs.

To access more information about enterprise services, please visit our web site at the following URL:

www.symantec.com/business/services/

Select your country or language from the site index.

Contents

Chapter 1

Introduction

This chapter includes the following topics:

What is Vector Machine Learning

Vector Machine Learning (VML) helps automate the detection of confidential information in your organization. To use VML, you train Symantec Data Loss Prevention to detect content similar to an example set of documents you provide. VML then performs statistical analysis to determine if unstructured data in your organization is similar to the data in the training set of documents.

You implement VML by training the system against positive content you want to protect and negative content you want to ignore. During the training process, the system selects features and generates a statistical model. The system then applies the model to detect content with features similar to the positive content.

About this guide and attachment

This guide complements the VML product documentation available in the *Symantec Data Loss Prevention Administration Guide* and the *Symantec Data Loss Prevention Release Notes* for version 11.1. This guide assumes that you are familiar with the information related to VML that is contained in those documents. In addition, this guide assumes that you understand policy concepts, including the various detection technologies and rule conditions. Refer to the *Symantec Data Loss Prevention Administration Guide* for more information.

This guide will be updated periodically and made available at the DLP Knowledgebase (https://kb-vontu.altiris.com), article number 54340. You can register at the Knowledgebase to be notified when the document is updated.

Attached to this article at the Knowledgebase is the archive file `DLP_Wikipedia_sample.zip`. This file includes 10,000 text documents that contain generic or neutral content. You can use some of these documents to seed your negative training sets, and the entire archive to test and tune your VML profiles. Where appropriate the best practices in this guide describe how you should use these documents. These documents are provided under the Creative Commons license (http://creativecommons.org/licenses/by-sa/3.0/).

Summary of VML best practices

The following table provides a summary of the VML best practices discussed in this guide, with links to individual topics for more in-depth recommendations.

Table 1-1 Summary of VML best practices

Functional area	Best practice
Recommended uses for VML	Use VML to protect unstructured, text-based content. Do not use VML to protect graphics, binary data, or personally identifiable information (PII). See "When to use VML" on page 13.
Category of content	Define the VML profile based on a single category of content that you want to protect and that is derived from a specific business use case. Narrowly defined categories are better than broadly defined ones. See "Recommendations for training set definition" on page 15.
Positive training set	Archive and upload the recommended (250) number of example documents for the positive training set, or at least the minimum (50). See "Guidelines for training set sizing" on page 16.
Negative training set	Archive and upload the example documents for the negative training set. Ideally the negative training set contains a similar number of well-categorized documents as the positive training set. In addition, add some documents containing generic or neutral content to your negative training set. See "Guidelines for training set sizing" on page 16.

Table 1-1 Summary of VML best practices *(continued)*

Functional area	Best practice
Profile sizing	Consider adjusting the memory allocation to low. Internal testing has shown that setting the memory allocation to low may improve accuracy in certain cases. See "Guidelines for profile sizing" on page 18.
Training set quality	Reject the training result and adjust the example documents if either of the base accuracy rates from training are more than 5%. See "Recommendations for accepting or rejecting a profile" on page 18.
Profile tuning	Tune the VML profile by performing negative testing using a corpus of testable data. See "Testing and tuning VML profiles" on page 21.
Profile deployment	Remove accepted profiles not in use by policies to reduce detection server load. Tune the Similarity Threshold before deploying a profile into production across all endpoints to avoid network overhead. See "Recommendations for deploying profiles" on page 19.

Chapter 2

VML profile recommendations

This chapter includes the following topics:

When to use VML

VML is designed to protect unstructured content that is primarily text-based. VML is well-suited for protecting sensitive content that is highly distributed such that gathering all of it for fingerprinting is not possible or practical. VML is also well-suited for protecting sensitive content that you cannot adequately describe and achieve high matching accuracy.

The following table summarizes the recommended uses cases for VML.

Table 2-1 Recommended uses for VML

Use VML when	Explanation
It is not possible or practical to fingerprint all the data you want to protect.	Often collecting all of the content you want to protect for fingerprinting is an impossible task. This situation arises for many forms of unstructured data: marketing materials, financial documents, patient records, product formulas, source code, and so forth. VML works well for this situation because you do not have to collect all of the content you want to protect, only a smaller set of example documents.
You cannot adequately describe the data you want to protect.	Often describing the data you want to protect is difficult without sacrificing some accuracy. This situation may arise when you have long keyword lists that are hard to generate, tune, and maintain. VML works well in these situations because it automatically models the features (keywords) you want to protect, and lets you easily manage and update the source content.
A policy reports frequent false positives.	Sometimes a certain category of information is a constant source of false positives. For example, a weekly sales report may consistently produce false positives for a Data Identifier policy looking for social security numbers. VML may work well here because you can train against the content that causes the false positives and create a policy exception to ignore those features. **Note:** The false positive contents must belong to a well-defined category for VML to be an effective solution for this use case. See "Recommendations for training set definition" on page 15.

When not to use VML

VML is not designed to protect structured data, such as Personally Identifiable Information (PII), or binary content, such as documents that contain mostly graphics or image files.

The following table summarizes the non-recommended uses of VML.

Table 2-2 Non-recommended uses for VML

Do not use VML to	Explanation
Protect personally identifiable information (PII).	Exact Data Matching (EDM) and Data Identifiers are the best option for protecting the common types of PII.
Protect binary files and images.	Indexed Document Matching (IDM) is the best option to protect content that is largely binary, such as image files, CAD files, etc.

Recommendations for training set definition

A VML category is the specific business use case from which you derive your example documents for training the VML profile. The more specific the category the better the detection results. For example, the category "Financial Documents" is not recommended because it is too broad. A better category classification is "Sales Forecasts" or "Quarterly Earnings" because each is particular to a specific business use case.

A VML category contains two sets of training content: positive and negative. The positive training set contains content you want to protect; the negative training set contains content you want to ignore. You should derive both the positive and negative training sets from the same category of content such that all documents are thematically related.

While it is possible to use entirely generic content for the negative training set, this is not recommended. A completely generic negative training set may produce good design-time training accuracy rates, but it is likely that at runtime you will not be able to detect the content you want to protect with sufficient accuracy.

Note: While a completely generic negative training set is not recommended, seeding the negative training set with some neutral-content documents does have value. See "Guidelines for training set sizing" on page 16.

The following table provides some example categories and possible positive and negative training sets comprising those categories.

Table 2-3 Some example categories and training sets

Category	Positive training set	Negative training set
Product Source Code	Proprietary product source code	Source code from open source projects

Table 2-3 Some example categories and training sets *(continued)*

Category	Positive training set	Negative training set
Product Formulas	Proprietary product formulas	Non-proprietary product information
Quarterly Earnings	Pre-release earnings; sales estimates; accounting documents	Details of published annual accounts
Marketing Plans	Marketing plans	Published marketing collateral, advertising copy
Medical Records	Patient medical records	Healthcare documents
Customer Sales	Customer purchasing patterns	Publicly available consumer data
Mergers and Acquisitions	Confidential legal documents; M&A documents	Publicly available materials; press releases
Manufacturing Methods	Proprietary manufacturing methods and research	Industry standards

Guidelines for training set sizing

VML is only as accurate as the example content you train. Unlike other detection technologies, to use VML you do not have to locate all the data you want to protect, nor do you have to describe it. But, you must select example documents that accurately represent the type of content you want to protect, as well as content you want to ignore that is thematically related to the positive content.

The more example documents you collect for training the more accurate the VML profile will be. A well-defined category of content contains 500 example documents: 250 positive and 250 negative. The minimum number of documents per training set is 50.

Ideally you will collect for training a similar number of negative documents as positive. However, this is not always possible. Regardless of how many negative documents you collect, you should seed the negative training set with generic or neutral-content documents. The archive file `DLP_Wikipedia_sample.zip` that is attached to this guide at the Knowledgebase is provided for this purpose. For example, if your positive training set contains the recommended number of example documents (250), and the negative training set contains 150 documents, you could add 100 to 200 generic documents to your negative training set from the `DLP_Wikipedia_sample.zip` archive file. Internal testing has shown that

adding generic content to complement an otherwise well-defined negative training set can improve accuracy for VML.

If you cannot collect enough positive documents to meet the minimum requirement, you can upload the under-sized training set multiple times. For example, consider a case where you have the category of content "Sales Forecasts." For this category you have collected 25 positive spreadsheets and 50 negative documents. In this case, you could upload the positive training set twice to reach the minimum document threshold and equal the number of negative documents. Note that you should use this technique for development and testing purposes only. Production profiles should be trained against at least the minimum number of documents for both training sets.

The table below lists the optimal, recommended, and minimum number of documents to include in each training set.

Note: These training set guidelines assume an average document size of 3 KB. If you have larger-sized documents, fewer in number may be sufficient.

Table 2-4 Training set size guidelines

Training set	Minimum	Recommended
Positive example documents	50	250
Negative example documents	50	250
Total number of documents for the category	100	500

Recommendations for uploading documents for training

While you can upload individual documents to the Enforce Server for training, it is recommended that you upload a document archive (ZIP, RAR, TAR) that contains the example documents for each training set. The maximum upload size is 30 MB. There is no training set size limit.

To gather the documents for training, it is recommended that you create a staging area. For example, consider a category called "Sales Reports." In this case you would create a folder called `\VML\training_stage\sales_reports` that represents the category. Within this folder you would create two subfolders, one for the positive training set and the other for the negative training set (for example: `\VML\training_stage\sales_reports\positive`). When you are ready to train

the profile, you compress the positive subfolder and the negative subfolder into separate document archives. You can partition the training set across archives if you have more than 30 MB of data to upload for a training set. Do not embed an archive within an archive.

Guidelines for profile sizing

Before you train a VML profile, you can adjust the amount of memory allocated to the profile. The amount of memory you allocate determines how many features the system models, which in turn affects the size of the profile. The higher the memory allocation setting, the more in-depth the feature extraction and the plotting of the model, and the larger the profile. In general, for server-based policy detection, the recommended memory allocation setting is high, which is the default setting.

On the endpoint, the VML profile is deployed to the host computer and loaded into memory by the DLP Agent. (Unlike EDM and IDM, VML does not rely on two-tier detection for endpoint policies.) Because memory on the endpoint is limited, the recommendation is to allocate low or medium memory for endpoint policies. Internal testing has shown that reducing the memory allocation does not reduce the accuracy of the profile and may improve accuracy in certain situations.

Table 2-5 Memory allocation recommendations

Memory allocation	Description
High	Default setting generally appropriate for server-based detection.
Medium	Use this setting to reduce the size of the profile.
Low	Use this setting for endpoint detection.

Recommendations for accepting or rejecting a profile

When you train a VML profile against the category content, the system selects features, creates the model, and calculates the base accuracy rates for false positives and negatives. Base accuracy rates are calculated using a standard and generally accepted process called k-folds evaluation. The base accuracy rates provide you with an early indicator of the quality of your category training sets.

To illustrate how the k-folds evaluation process works, assume that you have a category with 500 total example documents: 250 positive and 250 negative. During the training run, the system divides the training set into 10 folds, each of which

are distinct subsets of the overall training set and contain both positive and negative example documents. The system uses nine folds to generate a VML profile, and one fold to test the profile. Any of the folds can become the test fold for the first round of evaluation. For the next round, the next fold in the queue becomes the test fold. This process repeats for all 10 folds. The system performs a final training run called the cross-fold, averages the results of all folds, and generates the final model.

On successful completion of the training process, the system displays the averaged accuracy rates and prompts you to accept or reject the training profile. The false positive accuracy rate is the percentage of negative test documents misclassified as positive. The false negative rate is the percentage of positive test documents that are misclassified as negative. As a general guideline, you should reject the training profile if either rate is more than 5%.

Note: You can use the log file `machinelearning_training.log` to evaluate per-fold training accuracy rates.

See "Log files for troubleshooting training and detection" on page 26.

Recommendations for deploying profiles

Accepted VML profiles are transferred to every detection server and Symantec DLP Agent even if those profiles are not required by the active policies on that server or endpoint. Detection servers load all VML profiles into memory regardless of whether or not any associated VML policies are deployed to those servers. DLP Agents only load the VML profiles that are required by an active policy. To optimize server performance, it is recommended not to deploy (accept) unnecessary VML profiles and remove any accepted (deployed) VML profiles that are not required by active policies.

In addition, when you change the Similarity Threshold, the system re-syncs the entire profile with the detection servers and DLP Agents. If you have a large VML profile and possible bandwidth limitations (for example, deployment to many endpoints), this may cause network congestion. In this case you should test and tune the profile at a select few endpoints before deploying the profile into production at every endpoint on your network.

Chapter 3

VML testing and tuning recommendations

This chapter includes the following topics:

Testing and tuning VML profiles

You tune a VML profile by testing it with the Similarity Threshold set to 0. Once you determine the possible range of Similarity Scores for false positives, you adjust the Similarity Threshold to be just above the highest Similarity Score reported by false positives. This is referred to as negative testing.

A good training set has a well-defined range where the Similarity Threshold is set to achieve the best accuracy rates. A poor training set yields poor accuracy results regardless of the Similarity Threshold. A Similarity Threshold that is set too high or too low can result in a large number of false positives or false negatives.

To determine the proper Similarity Threshold setting, the recommendation is to perform negative testing as described in the following steps.

Table 3-1 Steps for tuning VML profiles

Step	Action	Description
Step 1	Train the VML profile.	Follow the recommendations set forth in this guide for defining the category and uploading the training set documents. Adjust the memory allocation before you train the profile. Refer to the *Symantec Data Loss Prevention Administration Guide* for help performing the tasks involved.
Step 2	Set the Similarity Threshold to `0`.	The default Similarity Threshold is 10. At this value the system does not generate any incidents. A setting of 0 produces the most amount of incidents, many of which are likely to be false positives. The purpose of setting the value to 0 is to see the entire range of potential matches and to tune the profile to be just above the highest false positive score.
Step 3	Create a VML policy.	Create a policy that references the VML profile you want to tune. The profile must be accepted to be deployable to a policy.
Step 4	Test the policy.	Test the VML policy using a corpus of test data. For example, you can use the `DLP_Wikipedia_sample.zip` file to test your VML policies against. Create some mechanism to detect incidents, such as a Discover scan target of a local file folder where you place the test data, or a DLP Agent scan of a copy/paste operation.
Step 5	Review any incidents.	Review any matches at the Incident Snapshot screen. Verify a relatively low Similarity Score for each match. A relatively low Similarity Score indicates a false positive. If one or more test documents produce a match with a relatively high Similarity Score, you have a training set quality issue. In this case you need to review the content and if appropriate add the document(s) to the positive training set. You then need to retrain and retune the profile. See "Log files for troubleshooting training and detection" on page 26.
Step 6	Adjust the Similarity Threshold.	By reviewing the incidents you should now be able to determine the highest Similarity Score among the detected false positives that you have tested the profile against. At this point you can adjust the Similarity Threshold for the profile to be just above the highest Similarity Score for the false positives. For example, if the highest detected false positive has a Similarity Score of `4.5`, set the Similarity Threshold to `4.6`. This will filter the known false positives from being reported as incidents.

Properties for configuring training

VML includes several property files for configuring VML training and logging. The following table lists and describes relevant VML configuration properties.

Table 3-2 Property files for VML

Property file at \Protect\config\	Description
`MLDTraining.properties`	Main property file for configuring VML training settings. See Table 3-3 on page 23.
`Manager.properties`	Property file for the Enforce Server; contains 1 VML setting. See Table 3-4 on page 25.
`MLDTrainingLogging.properties`	Properties file for configuring VML logging. See "Log files for troubleshooting training and detection" on page 26.

The following table lists and describes the VML training parameters available for configuration in properties file `MLDTraining.properties`.

Table 3-3 Relevant configuration parameters for VML training

Parameter	Description
`minimum_documents_per_category`	Specifies the minimum number of documents required for each training set (positive and negative). The default setting is `50`. Reducing this number below 50 is not recommended or supported. See "Recommendations for training set definition" on page 15.

Table 3-3 Relevant configuration parameters for VML training *(continued)*

Parameter	Description
`mld_num_folds`	Specifies the number of folds to use for the k-fold evaluation process. The default is `10`. Reducing this value will speed up the time the system takes to train against the content because less folds will be evaluated, but potentially at the sacrifice of visibility into profile quality. There is no need to change this value, unless you have a large number of example documents (and thus the training sets are very large), and you know for certain that you have a well-categorized overall training set. See "Recommendations for accepting or rejecting a profile" on page 18.
`minimum_features_to_keep`	Specifies the minimum number of features to keep for the profile. The default setting is `1000`. Lowering this value can help reduce the size of the profile. However, adjusting this setting is not recommended. Instead, use the memory allocation setting to tune the size of the profile. See "Guidelines for profile sizing" on page 18.
`significance_threshold`	Specifies the minimum number of times a word must occur before it is considered a feature. The default is `2`. Increasing this value (to 3 or 4, for example) may help reduce the size of the profile because fewer words will qualify as features. In general you should not adjust this setting unless setting the memory allocation to "Low" does not produce a small enough profile for your deployment requirements. See "Guidelines for profile sizing" on page 18.

Table 3-3 Relevant configuration parameters for VML training *(continued)*

Parameter	Description
`stopword_file`	Specifies the default stopword file `\config\machinelearningconfig\stopwords.txt`. Stopwords are common words, such as articles and prepositions. During training the system ignores (does not consider for feature extraction) any word contained in the stopwords file. If you add words to be ignored, you must use all lower case because VML feature extraction normalizes the content to lower case for evaluation.
`logging_config_file`	Specifies the configuration file for standard VML logging. See "Log files for troubleshooting training and detection" on page 26.
`native_logging_config_file`	Specifies the configuration file for native VML logging. See "Log files for troubleshooting training and detection" on page 26.

The following parameter is available for configuration in properties file `MLDTraining.properties`.

Table 3-4 Configuration parameter for VML profiles

Parameter	Description
`DEFAULT_SIMILARITY_THRESHOLD`	Establishes the default value for the Similarity Threshold, which is `10`. Changing this value affects the default value only. You can adjust the value using the Enforce Server administration console. See "Testing and tuning VML profiles" on page 21.

Log files for troubleshooting training and detection

The system provides debug log files for troubleshooting the VML training process and policy detection. The following table lists and describes the debug log files.

Table 3-5 Debug log files for VML

Log file	Description
`machinelearning_training.log`	Records the accuracy from training percentage rates for each fold of the evaluation process for each VML profile training run. This log file is useful for examining the quality of each training set at a granular, per-fold level. See "Recommendations for accepting or rejecting a profile" on page 18.
`machinelearning_native_filereader.log`	Records the "distance," which is expressed as a positive or negative number, and the "confidence," which is a similarity percentage, for each message evaluated by a VML policy. This log file is useful for examining all messages or documents evaluated by VML policies, including positive matches with similarity percentages beneath the Similarity Threshold, or messages the system has categorized as negative (expressed as a negative "distance" number). See "Testing and tuning VML profiles" on page 21.

Table 3-5 Debug log files for VML *(continued)*

Log file	Description
machinelearning_training_native_manager.log	Records the total number of features modeled and the number of features kept to generate the profile for each training run. The total number of features modeled versus the number of features kept for the profile depends on the memory allocation setting: ■ If "high" the system keeps 80% of the features. ■ If "medium" the system keeps 50% of the features. ■ If "low" the system keeps 30% of the features. See "Guidelines for profile sizing" on page 18.

Glossary

accuracy rates from training	Percentages reported by the system that provide an early indicator of the overall quality of a training set.
base false negative rate	Average percentage of positive training content that would not generate an incident with Similarity Threshold set to 0.
base false positive rate	Average percentage of negative training content that would generate an incident with Similarity Threshold set to 0.
category	Single, specific business use case from which you derive the example documents to train the system against.
cross-fold	Eleventh or last fold of the k-fold evaluation process that averages the results of all folds and produces the final accuracy rates from training.
current profile	Accepted, read-only version of the VML profile; the deployable instance.
example documents	Documents containing primarily text-based content that you upload to the system for training against.
false negative	Message or document that does not match but is a violation.
false positive	Message or document match that is not a true violation.
features	Keywords extracted from the example positive and negative documents; normalized to lower case.
k-fold evaluation	Standards-based process for evaluating and modeling the features and determining base accuracy rates. By default the system performs 10 distinct folds when creating a profile, as well as the final cross-fold.
memory required	Minimum amount of memory (in kilobytes) required by the server and the endpoint to load the profile into memory for policy detection.
message	Data evaluated for policy detection. Depending on the type of policy and server or endpoint, the message may be a document or a message component such as a header, body, or attachment.
model	Statistical results of a proprietary algorithm that plots the frequency of features, positive and negative, within and across the example documents in each training set.
negative training set	Example documents containing content you want the system to ignore; thematically related to the positive content.

never accepted	Profile state indicating the profile cannot be deployed for policy detection.
positive training set	Example documents containing content you want to protect that you upload to the Enforce Server for training.
profile	Data profile that the system generates from the model based on the positive and negative features.
similarity score	Number between 0 and 10 that indicates how similar the detected message is to the positive content in the profile.
similarity threshold	Configurable profile parameter between 0 and 10 that is used to make adjustments for training set quality.
temporary profile	Editable version of the profile; has not been trained, accepted, or both; cannot be deployed.
training process	Separate system process that evaluates the contents of the training sets, builds the model, and generates the profile for acceptance or rejection.
training set	The entire collection of example documents comprising the category of content; includes both the positive and negative training sets.
Vector Machine Learning	Protects unstructured data by performing statistical analysis to determine if content is similar to an example set of documents you train against.

Index

V