

# Machine Learning Sets New Standard for Data Loss Prevention: Describe, Fingerprint, Learn

# Machine Learning Sets New Standard for Data Loss Prevention: Describe, Fingerprint, Learn

## Contents

<b>The Challenge of Finding Intellectual Property Hidden in a Sea of Unstructured Data .....</b>	<b>1</b>
<b>Current Data Loss Prevention Detection Technologies: Describe and Fingerprint .....</b>	<b>1</b>
<b>A New Way to Find and Protect Data: Vector Machine Learning .....</b>	<b>2</b>
<b>How Vector Machine Learning Works .....</b>	<b>2</b>
<b>Putting Vector Matching Learning into Practice.....</b>	<b>3</b>
<b>Automated, Zero-day Protection for Data .....</b>	<b>4</b>
<b>The New Model for Data Loss Prevention Detection: Describe, Fingerprint and Learn .....</b>	<b>4</b>
<b>How to Learn More About Vector Machine Learning .....</b>	<b>5</b>

## **The Challenge of Finding Intellectual Property Hidden in a Sea of Unstructured Data**

Many organizations today are implementing Data Loss Prevention (DLP) programs to identify sensitive information and create policies aimed at controlling where data should and shouldn't go, as well as how it should get there. But in a world where cyber threats continue to increase almost daily, DLP products and programs must meet ever evolving challenges, especially from thieves targeting valuable intellectual property (IP). According to a May 2009 US federal government report, between 2008 and 2009 American business losses due to cyber attacks had grown to more than \$1 trillion worth of intellectual property. <sup>1</sup>

Thus the task of protecting intellectual property and sensitive information contained in data such as Microsoft® Word™ documents, spreadsheets, and image files, is more important than ever before. Yet locating such data throughout the organization has become much more difficult. That's because sensitive information is often buried in a sea of unstructured data that proliferates throughout the organization at multiple locations and in various devices.

As one industry publication recently noted, "The challenge for IT is that unstructured data is growing at a breakneck pace-- a compound annual growth rate of 61 percent, according to the International Data Corporation (IDC), almost three times the growth rate of structured data. It's also scattered throughout the enterprise: in folders on file servers, on laptops, and tucked inside USB drives." <sup>2</sup>

## **Current Data Loss Prevention Detection Technologies: Describe and Fingerprint**

Protecting sensitive information through deep content inspection and analysis using DLP is usually the first step to preventing data loss or misuse. Current DLP detection technologies typically rely on multiple methods for content analysis ranging from identifying keywords, dictionaries, and regular expressions, to partial document matching and fingerprinting. They can be classified into two major categories:

***Describing Technology*** - Protects confidential data by looking for matches to keywords, expressions or patterns, file type recognition, and other signature-based detection techniques.

***Fingerprinting Technology*** - Works by looking for exact matches of whole or partial files. Data to be protected is first collected in a variety of formats such as Microsoft Word files, Excel® files, and PDFs, and is then fingerprinted with a hashing algorithm to produce an index that can be deployed as part of a DLP policy.

While effective in protecting much of an organization's sensitive information, Fingerprinting and Describing technologies have limitations when addressing growing amounts of unstructured data and intellectual property such as product formulas, sales and marketing reports and source code.

That's because collecting all of the data that needs to be protected and fingerprinted can be challenging for organizations with limited resources or highly dispersed data. Thus, fingerprinting is most useful with highly specific and centralized sources of data. For unstructured textual data, keyword lists are typically used to find sensitive information. This approach, however, can be time consuming since generating and tuning keyword lists must be conducted continuously to ensure accuracy.

1-"The Financial Management of Cyber Risk," published by Internet Security Alliance (ISA) / American National Standards Institute (ANSI), 2010, p.10 Download at [www.isalliance.org](http://www.isalliance.org) or [www.ansi.org](http://www.ansi.org)

2-"A Strategy for Protecting Unstructured Data," Adam Ely, InformationWeek.com, Sept. 10,2010 [http://www.informationweek.com/news/business\\_intelligence/information\\_mgt/showArticle.jhtml?articleID=227500068](http://www.informationweek.com/news/business_intelligence/information_mgt/showArticle.jhtml?articleID=227500068)

## A New Way to Find and Protect Data: Vector Machine Learning

Recently, a new category of DLP detection technology has emerged that enables organizations to use software that learns to detect the types of confidential data that require protection. Through training, this approach continuously improves the accuracy and reliability of finding sensitive information. By applying the concept of machine learning to DLP, **Vector Machine Learning (VML)** helps to quickly and efficiently protect IP and confidential information among increasing amounts of unstructured data.

While machine learning as a concept has been around for decades and has been used in everything from anti-spam engines to Google™ algorithms for translating text, it is only now being applied to DLP content analysis. As a DLP detection technology, Vector Machine Learning learns to recognize sensitive information that must be protected using algorithms applied to a set of given example documents.

## How Vector Machine Learning Works

Figure 1 illustrates the components of VML whereby positive and negatives examples of sensitive data are provided at the “training” stage. During training, features are extracted to build a statistical profile that is used to classify unstructured textual data that should be protected.

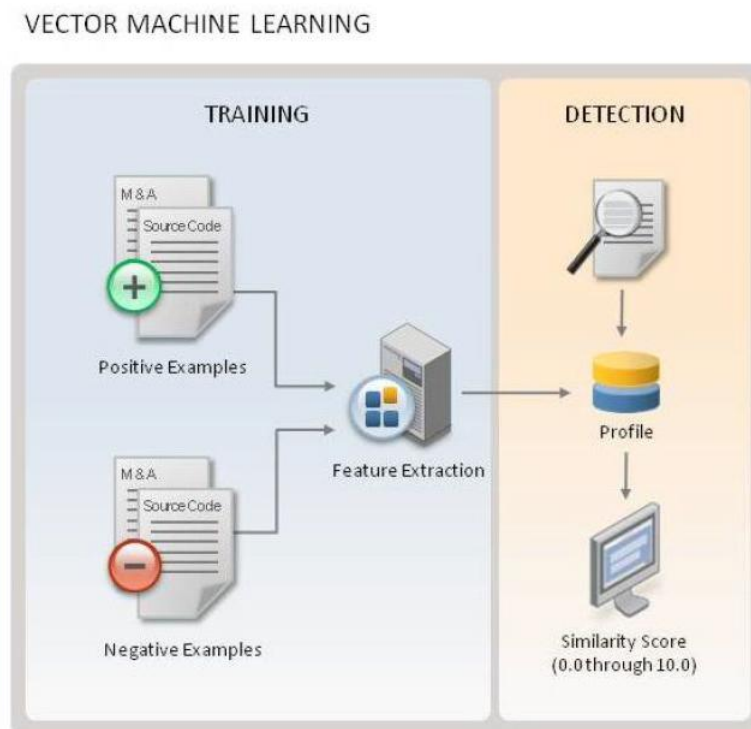


Figure 1 - How Vector Machine Learning works

## Machine Learning Sets New Standard for Data Loss Prevention: Describe, Fingerprint, Learn

Training defines the category of data to be protected through example documents. These include positive examples of data to be protected such as intellectual property or M&A documents, as well as negative examples of data that should be ignored. Positive examples could be documents containing proprietary source code. Negative examples could be an open source project downloaded from the Web. Both training sets are necessary to extract the key features that form a VML profile which will optimize accuracy during detection.

The process for implementing a Vector Machine Learning policy is straightforward. The user creates a VML profile by uploading positive and negative example documents. The VML engine performs training using the example documents and generates a statistical model, or profile, that is deployed after the user reviews and accepts training. For detection, the VML profile is used as part of a policy to classify any unknown document or message. If the data is similar to the positive example documents, then an "incident" is generated.

During detection the VML profile assigns a "similarity score" to the unknown document or message as part of classification. A similarity score of ten indicates that the examined data looks exactly like the example documents supplied in training. A score of zero means the data examined looks nothing like the example data from training.






Whenever false positives are generated, they can serve as feedback into the Training Set to help fine-tune the profile and improve its accuracy over time. In one case, a large computer hardware company using VML found that a single iteration of learning (by feeding false positives back into the negative training set) reduced the false positive rate to less than 4 percent---an accuracy rate on par with fingerprinting technologies.

### **Putting Vector Maching Learning into Practice**

Several use case examples are provided in figure 2 to illustrate the advantages of VML technology in detecting sensitive information among unstructured data. Keep in mind that the "sweet spot" for VML technology comes from protecting data that is usually difficult to access accurately with fingerprinting or information that might not be readily identified through a set of keywords or phrases.

For example, VML is well suited to protecting data such as proprietary source code for products, trading models for financial service firms or actuary algorithms for insurance companies. In the case of source code, VML would be able to provide coverage for completely new source code written by a developer on an endpoint such as a laptop computer (and therefore never seen before), and ensure that the new code would be subject to DLP policy enforcement.

VML is also useful in detecting sensitive data such as sales reports that may change frequently and exist in various formats such as Excel, Word, or email documents. By collecting examples of these kinds of reports for training, VML creates a profile that would be able to identify and enforce protection policies for the distribution of new sales reports each week regardless of their format.

Vector Machine Learning: Example Data Sets	
 <b>Source code</b>	Protect proprietary source code for a product, trading models, or actuary algorithms
 <b>Reports and forms</b>	Monthly or weekly sales reports, loan applications, and resumes
 <b>Legal contracts</b>	Licensing, partnerships, and sales agreements
 <b>HIPAA and HITECH</b>	Protected Health Information in the form of insurance claims, billing and procedure codes, emails to patients
 <b>ITAR (International Traffic in Arms Regulations)</b>	Intellectual Property and unstructured data that may be restricted

**Figure 2- Examples of data sets ideal for protection with VML technology**

**Automated, Zero-day Protection for Data**

Vector Machine Learning has specific advantages that complement existing describing and fingerprinting technologies, improving the ability of organizations to protect sensitive information especially for unstructured data that resides in highly dispersed and diverse locations.

**Automated processes help streamline set up and management** – By automating the policy definition and tuning process, VML significantly reduces staff time required to set up and maintain DLP technologies. Since training requires only examples of data to be protected, set up can be achieved quickly and efficiently. Many manual tasks such as maintaining keyword lists or trying to collect all data for fingerprinting are eliminated, and the incidence of false positives and tuning is minimized as the technology learns to recognize targeted information and improves in accuracy over time.

**Dynamic learning improves Accuracy and Timely Protection** – Much like zero-day protection with antivirus software, Vector Machine Learning is capable of delivering “zero-day protection” for confidential data with the accuracy of fingerprinting. The dynamic learning characteristics of VML make it possible to recognize newer or never seen before information more easily and accurately and therefore provide coverage for sensitive data that has yet to be created. Given the accelerating growth of unstructured data, therefore, VML complements the content analysis of both fingerprinting and described content technologies to enhance enforcement of DLP policies.

**The New Model for Data Loss Prevention Detection: Describe, Fingerprint and Learn**

Vector Machine Learning marks the introduction of a new category of deep content analysis that complements and improves existing DLP technologies designed to protect proprietary or confidential information. As shown in a figure 3, Vector Machine Learning, combined with currentdescribing and fingerprinting technologies, provides a new model for improving the efficiency and performance of DLP products and programs. Organizations with highly dispersed, growing data sets of unstructured proprietary and confidential information will want to examine and evaluate VML more closely.

DLP Technology	Description	Type of Data	Advantages/Challenges
<b>Describe</b>	<i>Describing</i> looks for data matches using keywords, signatures and patterns	All types of data	<ul style="list-style-type: none"> <li>• Most flexibility</li> <li>• Difficult to achieve very high accuracy</li> </ul>
<b>Fingerprint</b>	<i>Fingerprinting</i> looks for exact and partial data matches using hashing algorithms	Centralized structured and unstructured data	<ul style="list-style-type: none"> <li>• Highest accuracy</li> <li>• Difficult to access all data</li> </ul>
<b>Learn</b>	<i>Learning</i> looks for data by building a statistical model based on example documents	<ul style="list-style-type: none"> <li>• Unstructured textual data</li> <li>• Dynamic and/or highly distributed data</li> </ul>	<ul style="list-style-type: none"> <li>• Best coverage for new and never seen before data</li> <li>• Very high accuracy</li> <li>• Limited to unstructured data</li> </ul>

**Figure 3 - The new model for DLP detection technologies: Describe, Fingerprint, Learn**

### How to Learn More About Vector Machine Learning

To learn more about how your organization can take advantage of Vector Machine Learning to enhance existing or anticipated DLP program and technology investments, visit the Symantec web site at <http://www.symantec.com/business/products/family.jsp?familyid=data-loss-prevention> or contact the Symantec representative in your area.

#### To speak with a Product Specialist in the U.S.

Call (415) 829-5013

#### To speak with a Product Specialist outside the U.S.

For specific country offices and contact numbers, please visit our website at [www.symantec.com](http://www.symantec.com)

## About Symantec

Symantec is a global leader in providing security, storage, and systems management solutions to help consumers and organizations secure and manage their information-driven world. Our software and services protect against more risks at more points, more completely and efficiently, enabling confidence wherever information is used or stored. Headquartered in Mountain View, Calif., Symantec has operations in 40 countries. More information is available at [www.symantec.com](http://www.symantec.com).

For specific country offices and contact numbers, please visit our website.

Symantec World Headquarters  
350 Ellis St.  
Mountain View, CA 94043 USA  
+1 (650) 527 8000  
1 (800) 721 3934  
[www.symantec.com](http://www.symantec.com)

Symantec helps organizations secure and manage their information-driven world with IT Compliance, discovery and retention management, data loss prevention, and messaging security solutions.

Copyright © 2010 Symantec Corporation. All rights reserved. Symantec and the Symantec Logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Google is a registered trademark of Google and its affiliates in the U.S. and other countries. Microsoft Word and Microsoft Excel are registered trademarks of Microsoft and its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.  
11/2010 21158455