Enterprise Strategy Group by TechTarget

TECHNICAL FIRST LOOK

JUNE 2024

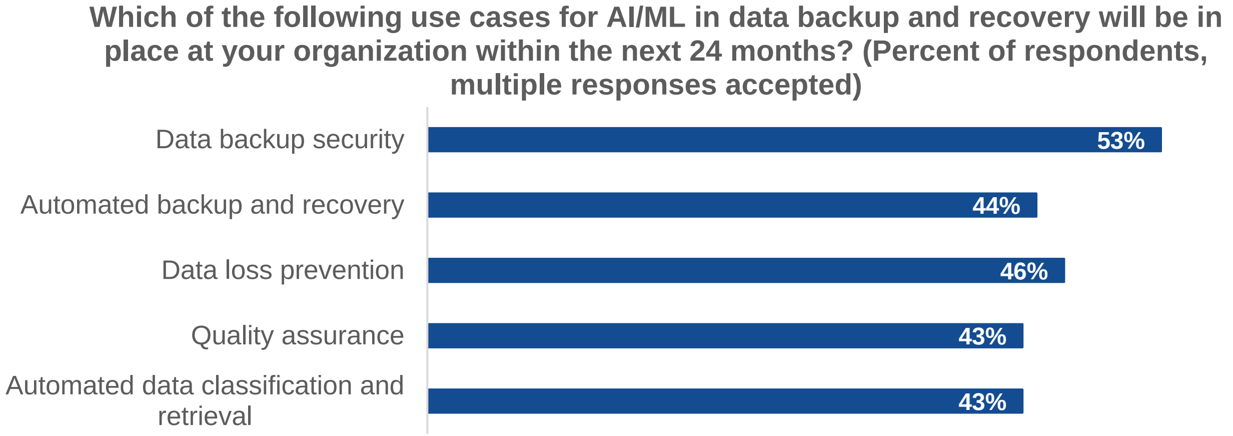
Index Engines' CyberSense Validated 99.99% Effective in Detecting Ransomware Corruption

Alex Arcilla, Senior Analyst – Validation Services

Cybersecurity Challenges When Protecting Data

Ransomware continues to plague organizations. According to research from TechTarget's Enterprise Strategy Group, 75% of respondents to a recent survey experienced at least one ransomware attack in the past year. As a result, organizations cited data loss, operational disruption, and financial loss to be among the top five negative impacts experienced.[1] To better recover from ransomware attacks, organizations have been adopting AI and machine learning (ML) into their data backup and recovery processes. In fact, 53% of respondents cited data backup security as the primary use case for AI/ML in backup and recovery (see Figure 1).[2]

Figure 1. Most Cited Use Cases for Data Backup and Recovery



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Recovering data quickly and effectively from any disruptive event requires organizations to ensure that data backup copies are free of ransomware and ransomware corruption. One approach is to use AI/ML to detect and verify the presence of corrupted data caused by ransomware in backups. However, the effectiveness of using any AI/ML process is dependent on how rigorously and continuously the supporting AI/ML models are trained to detect to most up-to-date and sophisticated ransomware attack patterns with a high degree of accuracy.

[1] Source: Enterprise Strategy Group Research Report: *Ransomware Preparedness: Lighting the Way to Readiness and Mitigation*, December 2023.
[2] Source: Enterprise Strategy Group Complete Survey Results: *Reinventing Backup and Recovery With AI and Machine Learning*, April 2024.

Index Engines CyberSense

Index Engines CyberSense detects ransomware corruption within backups and snapshots with 99.99% effectiveness. CyberSense analyzes data in backups or snapshots and when an attack is detected, it issues timely alerts to notify organizations of the attack type and blast radius and creates detailed post-attack forensic reports to support curated and quicker recovery efforts. The key to CyberSense's accuracy is its proprietary AI/ML engine trained with petabytes of data from both real-world attacks and attacks simulated in Index Engines' Research Lab, using thousands of known ransomware variants.

CyberSense employs AI-based data analysis that inspects the full content of files within data backups or snapshots to detect changes over time and reveal patterns of ransomware data corruption. Deep inspection of files in storage platforms, databases, core infrastructure (e.g., router configurations), as well as metadata-based and—critical to CyberSense—content-based analytics predicts if detected changes are due to a ransomware attack. Central to CyberSense is its ability to—beyond just looking at changes in metadata, thresholds, or compression rates—recognize how contents of backup files change over time and precisely detect data corruption due to cyberattacks.

Compatible with multiple backup platforms, organizations large and small can enable direct scanning of their backups without the need to rehydrate data. CyberSense has proven its value to customers based on its experience in the field, partnering with top storage vendors to provide their cybersecurity capabilities.

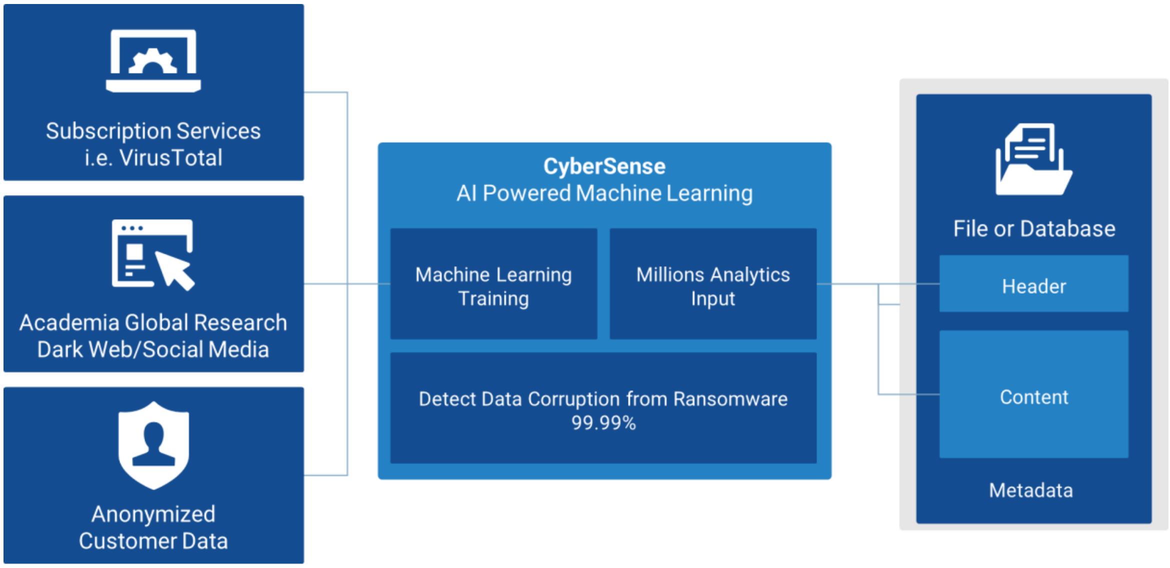
First Look

To provide the detection that organizations demand, Index Engines continually and rigorously trains CyberSense's AI/ML engine to recognize current and emerging patterns of ransomware attacks. Currently, Index Engines estimates its accuracy rate to be 99.99%. To validate that the quoted 99.99% is a realistic estimate, Enterprise Strategy Group evaluated the process that Index Engines uses to train and test CyberSense's ML engine—specifically, how Index Engines acquires data, generates data sets, trains the supervised ML models, and calculates the accuracy rate.

Data Acquisition

We first reviewed how the raw data is acquired to create the training data set for CyberSense's ML engine. As shown in Figure 2, Index Engines acquires data for creating the training data sets from three main sources: subscriptions services (e.g., Virustotal.com for ransomware executables, Wayback Machine for examples of clean files and associated changes over time); public sources (e.g., academia, global research from third-party organizations, the dark web, social media); and anonymized Index Engines customer data. A portion of the customer base of varying sizes and verticals opt in to supply data daily through the Index Engines' private cloud.

Figure 2. CyberSense's AI-powered Data Analysis and Machine Learning



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Enterprise Strategy Group took note of the variety and volume of raw data acquired, as these factors affect the quality and completeness of the training data set. We saw that:

- Index Engines detonates executables, both manually and automatically, via scripting to reveal ransomware attack patterns that the ML engine can use to learn.
- To minimize false negatives, Index Engines adds in its own files from backups of both clean and infected hosts to the training set.
- Index Engines leverages over 200 statistics—such as file properties and number of files added, deleted, or modified over a given period—to characterize how the population of files in all backups change over time.
- Over 7,000 ransomware variants have been identified to train the ML engine, categorized into three behaviors: files in which data has changed with no file name preserved, files in which the file name has changed with known ransomware extension, and files in which the file name has been obfuscated or modified.

Training Data Set Creation

The effectiveness of CyberSense in detecting ransomware corruption relies on the data set used to train the ML engine. To evaluate how the training set is created, Enterprise Strategy Group evaluated Index Engines' Statistical Analytic Generation (SaGen) process.

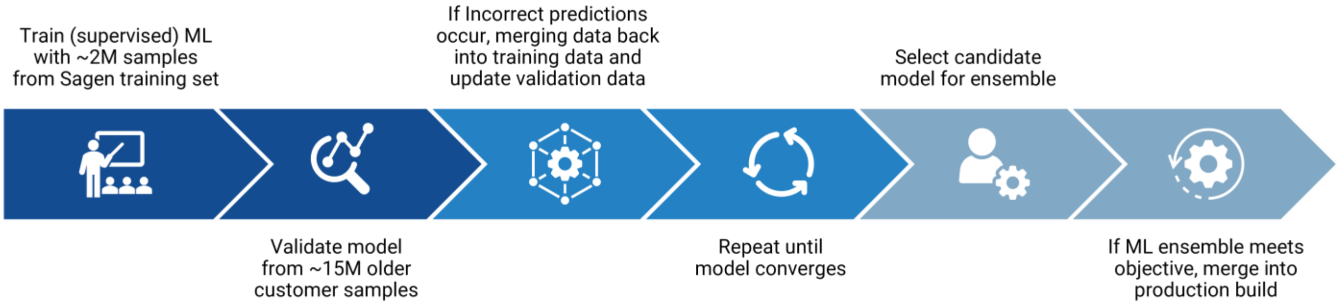
Using Index Engines' native indexing and query capabilities, the ML engine first learns and analyzes how files change from clean to infected state using clean/detonated backups. Using this knowledge, Index Engines artificially creates millions of different backup scenarios (performing incremental and full backups, varying operating systems and file types). Events that identify unique sequences of changes that can occur within backup files are classified as either normal operations or attacks. Millions of virtual backups representing the artificial backup scenarios are generated; up to 1 million can be created per day. CyberSense then processes and analyzes these backups to create the training data set.

Enterprise Strategy Group noted how comprehensive the SaGen process is for creating a data set for training CyberSense to uncover ransomware corruption in IT production networks.

ML Engine Training and Validation

To ensure that the ML engine is properly trained, Enterprise Strategy Group noted the thorough and sound approach taken by Index Engines (see Figure 3).

Figure 3. Training and Verifying CyberSense's ML Model Ensemble


Train (supervised) ML with ~2M samples from Sagen training set
Validate model from ~15M older customer samples
If Incorrect predictions occur, merging data back into training data and update validation data
Repeat until model converges
Select candidate model for ensemble
If ML ensemble meets objective, merge into production build

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

CyberSense's ML engine is an ensemble of 10 AI/ML models, each building off the results of the others during the training process until results converge. ML training is first conducted with approximately 2 million samples generated from the training set. Index Engines then validates the ML engine from approximately 15 million customer samples. Any incorrect predictions are merged back into training data, which will update the validation data. This repeats until the ensemble model converges.

For the ML engine's most recent training, approximately 6,000 samples were merged back into the training data. Subsequent passes quickly reduced it to a few (less than 10) samples after three to four iterations.

Once converging, the candidate model group was selected to test on customer scenarios that were deemed difficult (based on specific customer scenarios that were not initially detected on previous releases), then tested on an additional 30 million new customer samples (i.e., actual backups occurring in the field).

Accuracy is currently estimated to be 99.997% based on the fraction of true positives identified from the total true positives and false negatives identified within 125,000 data samples used for model validation.

Conclusion

Ransomware prevention strategies are critical to implement but rarely 100% effective. Detecting ransomware corruption in data backups is paramount to ensure that the business minimizes any operational risk when prevention fails. While adopting AI/ML can help to bolster data backup security, specifically in detecting ransomware corruption and the presence of ransomware, organizations must ensure that the supporting models are continuously trained with data reflecting existing and emerging ransomware corruption patterns. Enterprise Strategy Group validated that the approach to creating the data sets and training the ML models for Index Engines' CyberSense are as complete and thorough as possible, justifying the quoted 99.99% accuracy rate. For organizations seeking to incorporate AI/ML in improving its accuracy in detecting ransomware corruption, we strongly suggest closely looking at CyberSense.

Information contained in this publication has been obtained by sources TechTarget considers to be reliable but is not warranted by TechTarget. This publication may contain opinions of TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

✉ contact@esg-global.com
🖥 www.esg-global.com