CYERA

PLATFORM WHITEPAPER

Redefining Data Classification

Leveraging AI and LLMs for Precision and Speed in Sensitive Data Protection

Abstract

As organizations expand their digital footprints, the challenge of locating, managing and protecting sensitive data has become increasingly complex. Traditional data classification methods have long been the backbone of data protection strategies, but they have proven to be inadequate, typically generating too many false positives—a weak backbone—especially in the face of today's data sprawl and dynamic environments. This white paper explores how advancements in Artificial Intelligence (AI) and Large Language Models (LLMs) are revolutionizing the automatic classification of sensitive data, enabling greater accuracy, efficiency, speed and contextual understanding. It also delves into how Cyera’s innovative approach leverages these technologies to enhance data security without compromising privacy.

1. The Evolution of Automatic Data Classification: From Traditional Methods to AI

1.1 Reversing the Business Paradigm of Data Ownership

For years, the industry operated under the assumption that business teams within organizations inherently own and understand their data, and would therefore inform security teams about which data is sensitive and requires protection. This approach has proven to be fundamentally flawed; self-attestation of data by the business often fails due to numerous challenges, including limited visibility and a lack of comprehensive understanding of the data landscape. It's time for a shift in this paradigm. Security teams must be empowered to take the lead, proactively identifying sensitive data and guiding the business in determining what needs protection and how to best achieve it.

1.2 Traditional Approaches to Data Classification and Their Limitations

Data classification is a fundamental element of any robust data protection strategy. The process involves identifying, categorizing, and protecting sensitive information across various locations within an organization’s digital landscape. Generally, this process involves scanning files and databases to look for specific keywords, which are then used to classify those objects as sensitive.

Data classification has generally been viewed narrowly as identifying the sensitivity level of data—such as public, internal, confidential, or restricted. However, effective data classification should go beyond these basic sensitivity labels to provide a more granular and enriched understanding of the data. This level of granularity allows organizations to pinpoint specific data characteristics, such as data types and data compliance with certain industry regulations, and to assess its criticality to the business. By classifying data with a highly detailed approach, organizations can better target their protection efforts, ensuring that sensitive and regulated data is safeguarded in line with both organizational and legal requirements.

Traditional data protection solutions rely heavily on static content-based detection algorithms, such as regular expressions (regex). While weak patterns, like names and addresses, often produce poor results, even stronger patterns (such as credit card numbers) may yield only acceptable results, still prone to false positives without additional context. This lack of contextual understanding significantly limits the accuracy of static detection methods, underscoring the need for more sophisticated solutions that can interpret data within its specific context.

Although these techniques can identify specific patterns, they are often limited in scope and lack the necessary context, leading to inefficiencies and inaccuracies. For example, let’s consider the word ‘Jordan.’ A regex can be built to match the word ‘Jordan,’ but without context, it’s impossible to determine what 'Jordan' represents in that specific instance. Is it a person’s name, the country, or the shoe brand?

Static data classification methods also struggle to adapt to evolving data environments, necessitating continuous manual tuning to remain effective. They often miss vast amounts of sensitive data, especially new data types they have not encountered before and are not yet equipped to detect. Additionally, these methods frequently generate false positives—incorrectly flagging non-sensitive data as sensitive—leading to business disruptions and overwhelming incident response teams with unnecessary alerts.

Other traditional algorithms, like file fingerprinting and Exact Data Matching (EDM), are resource-intensive and computationally expensive, particularly when dealing with voluminous datasets, such as large data repositories. These methods focus solely on predefined, known data types and often require significant time—usually weeks or months—to complete the scanning of multiple cloud datastores for example. As a result, such approaches are often avoided, such as for endpoint data discovery where quick and efficient scans are essential. These methods are also burdensome for large-scale environments, such as cloud or on-premises datastores, where rapid and efficient scanning has become crucial. Avoiding these methods leaves potentially sensitive data undetected and, therefore, unprotected.

Fundamentally, legacy data protection solutions are all constrained by their lack of contextual understanding necessary to accurately assess data sensitivity. These systems are rule-based and static, unable to interpret the nuanced contexts that human analysts can easily discern. Their rigidity makes them ill-suited for modern, dynamic data environments, where data formats and contexts are continually changing.

This is not to say that these methods are ineffective. On the contrary, they can be very effective in basic situations and use cases where sensitive data is easy to identify. However, in today's world of data sprawl and ubiquity, they are no longer sufficient. Traditional methods need to be augmented by more intelligent approaches that can accurately and efficiently classify sensitive data.

2. The Advent of AI and LLMs in Data Classification

2.1 A Paradigm Shift with AI and LLMs

The introduction of AI and Large Language Models (LLMs) has marked a significant turning point in the field of data classification. Unlike traditional methods, AI-powered classification systems can understand and interpret data contextually, much like a human analyst. These systems leverage vast amounts of training data to develop a nuanced understanding of language, patterns, and context, enabling them to classify data and assess its sensitivity with unprecedented accuracy.

LLMs are incrementally trained on diverse datasets, allowing them to recognize and understand a wide range of data types and structures, even auto-learn new data. These capabilities enable LLMs to detect sensitive information in any form, also when it is embedded in complex, unstructured formats, such as emails, reports, or free-text fields, and even in hybrid semi-structured formats.

2.2 Addressing Privacy Concerns with Secure AI Implementations

A common concern with leveraging AI and LLMs for data classification is the potential risk to data privacy. Many organizations fear that using external AI models could expose their sensitive data to third parties, undermining the very security measures they seek to enhance. However, advancements in secure, private AI implementations have addressed these concerns. Cyera's AI models for example are trained, fine-tuned, and deployed in-house and are proprietary to the organization, are mainly trained on public datasets, and leverage customers' isolated environments.

3. Cyera’s Advanced Data Classification: A Comprehensive Approach

3.1 Datastore Discovery: The Foundation

How can sensitive data be found if you don’t know where to look? Datastore discovery is a fundamental yet often overlooked aspect of data at-rest discovery and classification, particularly in cloud-native environments. Traditional approaches have significant limitations; they typically require users to manually onboard specific datastores they are aware of, leaving unknown or unmanaged datastores unscanned and potentially unprotected. This creates critical gaps in data visibility and security.

Datastores include a wide range of data repositories such as buckets, object stores, unmanaged, semi-managed, and managed databases, as well as data lakes and data management platforms. Examples include S3 buckets, Virtual Machines, O365 OneDrives, SharePoint drives, and various databases.

Cyera addresses this issue by proactively discovering all datastores within an organization and its cloud accounts, including unknown or obsolete ("ghost") datastores that might otherwise go unnoticed. Cyera automatically identifies their locations, general details, security configurations, data classifications, and any issues requiring attention. Metadata and security configurations for each datastore are continuously reassessed. This proactive and comprehensive approach ensures that no sensitive data remains unclassified or unprotected.

3.2 Data Scanning and Sampling: Structured vs. Unstructured Data

As a baseline, Cyera leverages traditional out-of-the-box data detection methods for quick and easy recognition of sensitive data, using data identifiers and optical character recognition (OCR), combining regular expressions and rich contextual information around data and files. But it doesn’t stop there. Cyera augments traditional detection methods based on pre-defined data identifiers with advanced data-centric AI and LLMs to offer a robust, accurate, and context-aware data classification solution. Cyera handles structured, unstructured and even semi-structured (i.e. JSON, NDJSON, JSONL) data types. Cyera’s data classification process begins with comprehensive data scanning and sampling, tailored to the specific characteristics of both structured and unstructured data.

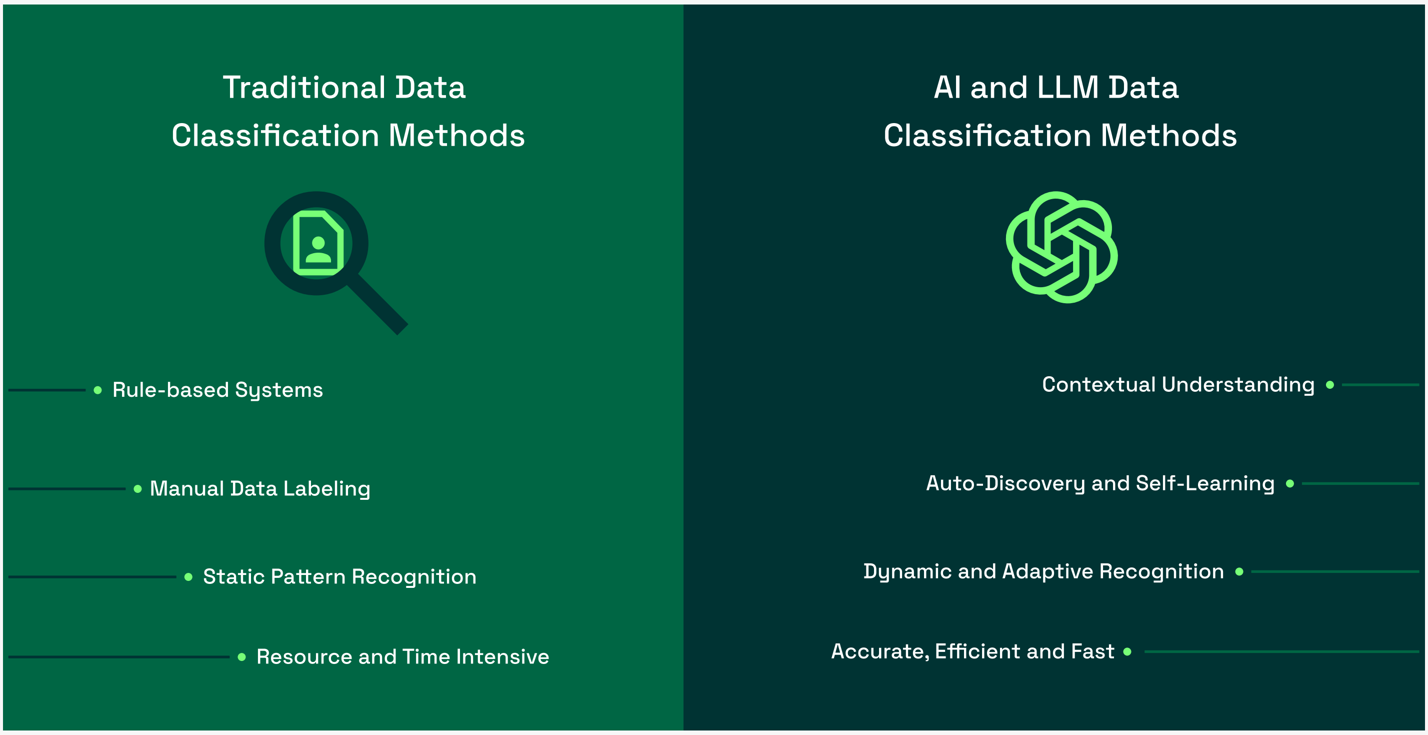
- **Structured Data Scanning**: Cyera supports a wide range of structured file formats, including AVRO, XLSX, XLSB, XLS, XLSM, XLTX, XLT, PARQUET, PARQ, CSV, TSV, ORC, FLAT, LST, PSV, SSV and many others, across multi-cloud datastores and on-premises repositories. For databases like MS SQL and PostgreSQL, Cyera creates a local snapshot of the database within the customer’s environment for analysis, ensuring data remains secure and within the customer's control during the scanning process. This process is highly optimized for speed, involving the sampling of rows and columns to create a statistically significant dataset for further analysis. The system examines database and structured file schemas, counts distinct values, and pulls diversified data samples, ensuring that the classification engine receives a comprehensive view of the data. This approach dramatically reduces scan time while maintaining high classification accuracy.

- **Unstructured Data Scanning:** Unstructured data (i.e. DOCX, PPTX, GDOC, GSLIDES, PDF, PNG, JPG, JPEG, BMP, GIF, TXT, CERT, PDF, compressed files and many others), such as textual and image files stored in cloud environments or on-premises systems, requires a different approach. Cyera clusters similar files together and samples a representative subset for analysis. This method reduces the computational load while maintaining high accuracy in classification. During this process, Cyera also gathers metadata—such as file size, type, and access permissions—to enrich the classification results.

3.3 File-Level Classification

While content-level data detection focuses on identifying specific information within files, such as phone numbers or full names, Cyera also excels in object-level (file) classification, which considers the entire file as a whole. This approach is particularly useful for identifying sensitive files based on their overall characteristics, not just their content. Object-level classification enables Cyera to classify files like as sensitive such as sales agreements, credit reports, financial statements (tax, earning, invoices, loan applications etc.) purchase orders, healthcare documentation (disability forms, medical records etc.), intellectual property (software design documents, recipes etc.), subscription and services agreements and many others, in addition to recognizing specific data and fields within them, such as credit card numbers, invoice IDs, PHI or employee numbers. File-level classification is particularly valuable for customers who may not handle much PII but prioritize the protection of their intellectual property, confidential documents and other secrets.

This approach allows Cyera to detect and classify the broadest pool of sensitive data in customer environments, including new and previously unknown types of data, that traditional solutions often miss due to their strict reliance on full content-level inspections. By leveraging multiple levels of data scanning, Cyera not only enhances the speed and accuracy of data classification but also addresses newer and more specific customer needs more effectively. This tailored approach ensures that critical data, such as proprietary information, is precisely identified and safeguarded, providing a more impactful and relevant solution compared to traditional methods.



Traditional Data Classification Methods

- Rule-based Systems
- Manual Data Labeling
- Static Pattern Recognition
- Resource and Time Intensive

AI and LLM Data Classification Methods

- Contextual Understanding
- Auto-Discovery and Self-Learning
- Dynamic and Adaptive Recognition
- Accurate, Efficient and Fast

3.4 AI-Powered Classification: Precision and Contextual Understanding

Not all data needs to be protected, and certainly not all data requires the same level of protection. Legacy data security solutions are often known to hinder legitimate business processes by incorrectly labeling data as sensitive, blocking access or sharing when it is unnecessary. Additionally, some data is more sensitive or confidential than others, necessitating tailored responses that may vary based on sensitivity levels. This is why the data classification process must produce precise outcomes. Relying solely on user-driven data tagging is insufficient, as users may overlook or dismiss this process. Therefore, automatic classification is fundamental. To do so Cyera leverages a powerful combination of out-of-the-box data identifiers and advanced AI models.

Cyera’s classification engine leverages proprietary AI models to achieve remarkable accuracy in data classification. The system is pre-trained and continually updated by Cyera using large datasets and is also capable of auto-learning from each customer’s unique environment without requiring re-training. It adapts to specific data formats and patterns encountered across different industries and geographies. This adaptive learning capability enables Cyera’s solution to enhance and refine sensitive data classification outcomes beyond what basic methods provide, and to identify sensitive data that traditional methods miss.

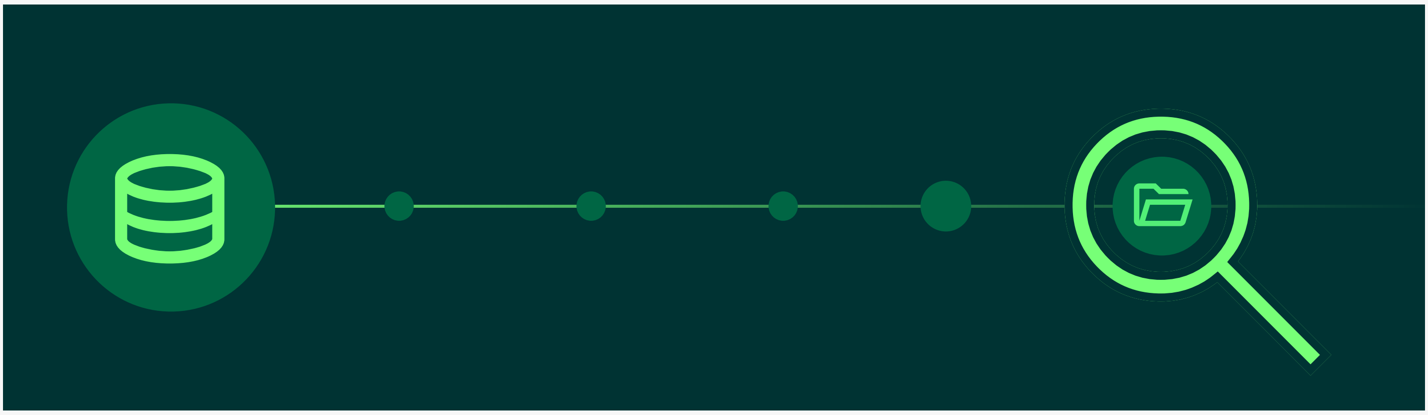
The classification process goes beyond simple pattern recognition. Cyera’s AI models analyze the context surrounding the data, including factors such as the role of the data subject (e.g., customer, employee, contractor, patient), the geographic location and regional context (e.g., within the EU or outside the EU, and within specific privacy legislation like those in California, the UK, or Canada), the applicable compliance framework, and the specific security levels associated with different data types (e.g., plain text, hashed, masked, encrypted). This contextual understanding is crucial for accurately classifying data and applying the appropriate security measures.

3.5 Continuous Learning and Automatic Fine-Tuning

At the core of Cyera’s AI and LLM capabilities is a continuous learning process. The models are pre-trained on vast amounts of data to create robust, out-of-the-box data classifiers. These classifiers are capable of identifying common sensitive data types such as credit card numbers, social security numbers, and personal identifiable information (PII) right from the start.

However, what sets Cyera apart is the ability of its models to auto-learn from customer-specific data during production runtime (inference). This means that unique identifiers such as employee IDs, product SKUs, lot numbers, claim numbers, and specific statements are incorporated into the learning process. Over time, the model adapts to the unique data environment of each customer, enhancing its accuracy and relevance.

Typically, about 70% of Cyera’s classified data is attributed to 'learned classification,' which is often missed by traditional data security solutions. In fact, these traditional solutions frequently overlook a substantial portion of the data or require extensive tuning and effort to achieve comparable coverage.



3.6 Cyera's AI and LLM: Achieving High Data Classification Accuracy

Cyera's AI and Large Language Models (LLM) are meticulously designed to achieve exceptional accuracy in data classification, a critical requirement in today's data-driven world. Once the models are trained —an ongoing process—they are integrated into Cyera's Data Insights Service. Here, the models classify data by analyzing database metadata, file contents, and other contextual information. Cyera ensures that only high-precision classifications, supported by a large amount of training data, are presented within the platform, minimizing the risk of false positives. The journey to this high level of precision begins with the foundational architecture of Cyera's proprietary models.

Let's revisit the example mentioned in section 1.2, focusing on the word 'Jordan.' While a regex can be designed to match the term 'Jordan', it lacks the context needed to determine its specific meaning. Without additional information, it's unclear whether 'Jordan' refers to a person's name, the country, or a popular shoe brand. Regex alone cannot resolve this ambiguity, but AI, particularly with advanced language models, can easily discern the context. You probably use generative AI applications like ChatGPT, and are aware of how well LLMs understand context. Similarly, AI-powered data classification can identify nuanced data roles and adhere to data sovereignty requirements with precision.

Beyond classification, Cyera enriches the data by identifying contextual factors that impact its sensitivity. For instance, the system can differentiate between a customer's phone number and an employee's phone number, applying different security protocols based on the sensitivity level. This nuanced approach reduces the risk of over-protection, which can lead to unnecessary business interruptions.

Cyera also tracks data access patterns, assigning trust levels to both human and non-human identities. This capability helps organizations granularly enforce Zero Trust policies at the data level by ensuring that sensitive data is only accessible by authorized users or systems.

3.7 Proprietary Development and Model Ownership

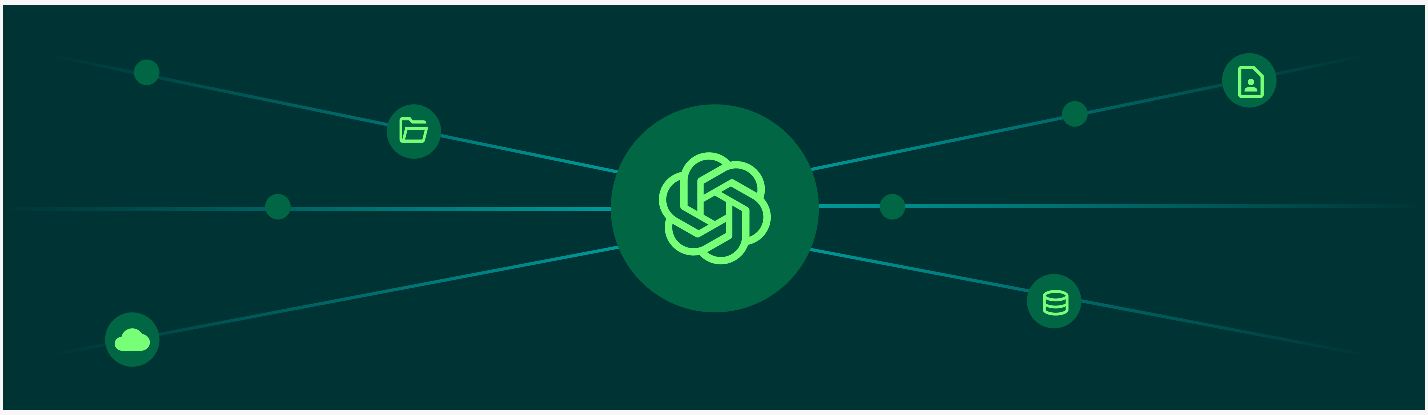
Cyera's AI and LLM models are entirely proprietary, developed in-house to maintain control over their capabilities and security. While Cyera does leverage common open-source foundation models, such as FLAN-T5 , these are just the starting points. The real value comes from how Cyera uniquely fine-tunes these models to optimize them for specific data classification tasks such as NER (Named Entity Recognition).

This fine-tuning is performed within Cyera's isolated, secure environment, ensuring that the models are uniquely tailored to Cyera's data classification needs. This exclusive development process, protected as a trade secret, allows Cyera to deliver a level of precision and performance that is unmatched.

3.8 Privacy and Security: Protecting Customer Data

Cyera takes customer data privacy seriously. Cyera's commitment to data privacy extends to its AI model development and training processes. Cyera's AI models are developed in-house and are proprietary to the organization. These models are trained on large volumes of public datasets and can also be automatically trained on minimal customer samples—such as when classifying novel data types, all within a secure, isolated environment. The samples are embedded, irreversible, and segregated to prevent any possibility of data spillage. For classifications of unstructured files and structured data, Cyera collects only a small fraction of the environment volume. These samples are used for classification purposes. Cyera's architecture ensures tenant isolation, meaning that no customer data is ever exposed, shared externally, or mixed with that of another customer. While our robust data protection measures make opting out of training an unnecessary precaution, we do offer the option for customers who prefer it, while still benefiting from the ML-based classifications provided by Cyera. This flexibility ensures that organizations can maintain control over their data while still leveraging the powerful capabilities of Cyera's AI and LLM.

This rigorous approach to data privacy allows organizations to benefit from the advanced capabilities of AI and LLMs, ensuring they can confidently use Cyera's AI-powered classification engine without risking data leakage or exposure.



4. The Future of Data Classification: AI and LLMs at the Forefront

4.1 Adapting to Data Sprawl with AI

As data sprawl continues to accelerate, the need for advanced, accurate, and context-aware data classification becomes increasingly critical. AI and LLMs, alongside traditional data identification methods, are uniquely positioned to address these challenges by offering a level of precision and contextual understanding that traditional methods alone cannot match.

Cyera's AI-driven approach represents the future of data classification. By integrating advanced AI models with traditional detection methods, Cyera provides a comprehensive solution that enhances data protection while supporting business agility. Organizations can reduce false positives, ensure compliance, and protect all sensitive data—structured, semi-structured, and unstructured—across all environments—cloud IaaS, DBaaS and SaaS and on-premises—while achieving unprecedented data scanning performance and minimizing costs related to deployment, computing resources, and storage.

4.2 Conclusion: Embracing the AI-Driven Future of Data Security

The age of AI has arrived, bringing with it transformative capabilities for data classification and protection. Cyera is at the forefront of this revolution, offering a solution that not only meets the demands of today's complex data environments but also anticipates the challenges of tomorrow.

By harnessing the power of AI and LLMs, organizations can achieve a new level of accuracy and efficiency in data classification, ensuring that sensitive data is protected without compromising business operations. In this new era, Cyera is leading the charge in redefining how sensitive data is detected, classified, and safeguarded.

About Cyera

Cyera is reinventing data security. Companies choose Cyera to improve their data security and cyber-resilience, maintain privacy and regulatory compliance, and gain control over their most valuable asset: data. Cyera instantly provides companies with a holistic view of their sensitive data and their security exposure, and delivers automated remediation to reduce their attack surface.

Learn more at **www.cyera.io**, or follow Cyera on **LinkedIn.**

Trusted by:

LifeLabs Takeda ACV Cboe ARMIS GRANICUS