SYMMETRY

WHITE PAPER

Benchmarking of Data Classification for Healthcare

EXECUTIVE SUMMARY

Data classification has long been a cornerstone policy for modern security teams, underpinning the strategies that protect an organization's most valuable assets. The importance of data classification extends beyond mere organization of information; it is a critical process that ensures data is appropriately secured based on regulatory requirements, as well as its sensitivity, value, and relevance.

> Classification is the **starting point, not the final destination for data security.**
>
> Ensuring that data is handled in accordance with it's classification is the ultimate goal.

Claude Mandy
Chief Evangelist

The integration of generative AI into this ecosystem as a strategic enabler of enhanced productivity further highlights the importance of robust data classification. Generative AI offers unprecedented opportunities for businesses to optimize operations. However, leveraging this potential safely and effectively requires a deep understanding of the underlying data, necessitating precise classification. Symmetry's approach to data classification and performance provides the necessary precision and coverage to deal with this challenge. By reducing operational overhead, increasing confidence in data security, and providing consistent multi cloud support, Symmetry empowers organizations to protect their crown jewel data assets more effectively.

Key Challenges

- **Dynamic Nature of Data:** Data constantly evolves, requiring continuous reevaluation and reclassification to maintain alignment with its current state.
- **Multi-Cloud Complexity:** Cloud-native classification services are tightly coupled with their platforms, creating challenges for organizations with multi-cloud or hybrid environments.
- **Non Deterministic:** Determining sensitivity of data is often subtle and context- and usage- dependent.
- **False Positives:** Even a low false positive rate can lead to millions of incorrectly flagged items, overwhelming security teams and diverting resources from genuine threats.

Symmetry Systems' Approach & Performance

Symmetry leverages advanced AI and a unique deployment model to offer a compelling solution. In addition, Symmetry offers this classification across all major clouds and on-premises environments, at lower costs compared to volume-based pricing of cloud-native solutions, and allows customers to create their own identifiers and leverage context for robust detection. A detailed comparison with Amazon Macie and Google Sensitive Data Protection (gSDP) using a synthetic healthcare dataset revealed:

a. **Comprehensive Coverage:** Symmetry identified **2x more sensitive data types accurately than gSDP** and **6x more sensitive data types than Amazon Macie**.

b. **High Precision and Reduction of False Positives:** 92-100% precision across all data types at different sampling sizes, reducing false positives. For an average hospital, this would result in up to **73 million more sensitive data matches or up to 24 million less false positives to deal with than Amazon Macie and gSDP** respectively.

These results show that while traditional tools struggle with complex data environments, Symmetry effectively addresses modern data classification challenges.

TABLE OF CONTENTS

DATA CLASSIFICATION: A CORNERSTONE OF MODERN DATA SECURITY

Data classification has long been a cornerstone policy for modern security teams, underpinning the strategies that protect an organization's most valuable assets. The importance of data classification extends beyond mere organization of information; it is a critical process that ensures data is appropriately secured based on regulatory requirements, as well as its sensitivity, value, and relevance.

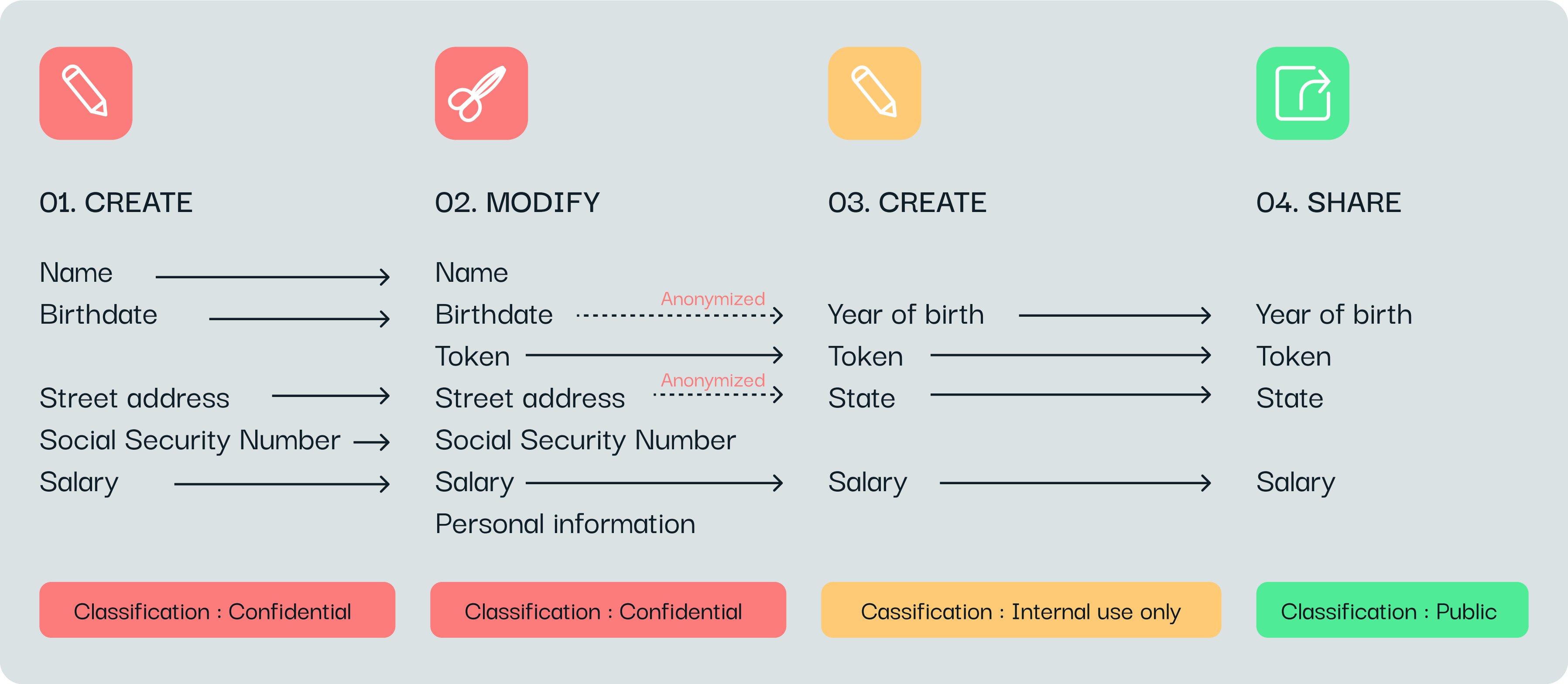
The integration of generative AI into this ecosystem as a strategic enabler of enhanced productivity further highlights the importance of robust data classification. Generative AI offers unprecedented opportunities for businesses to optimize operations, innovate product offerings, and personalize customer experiences through its ability to create new content, analyze vast datasets, and automate complex processes. However, leveraging this potential safely and effectively requires a deep understanding of the underlying data, necessitating precise classification.

The Challenges: The Costs Of Data Classification & False Positives

The challenges associated with data classification are multifaceted and complex, making it a near Sisyphean task for organizations. All of the difficulties stem from the dynamic nature of data itself.

DATA IS DYNAMIC

Data is never static, unless it has lost its usefulness; it evolves continuously. As information is created, modified, shared, and stored, its context–and, by extension, its classification–can change.



01. CREATE

Name
Birthdate
Street address
Social Security Number
Salary

Classification : Confidential

02. MODIFY

Name
Birthdate
Token
Street address
Social Security Number
Salary
Personal information

Anonymized
Anonymized

Classification : Confidential

03. CREATE

Year of birth
Token
State
Salary

Cassification : Internal use only

04. SHARE

Year of birth
Token
State
Salary

Classification : Public

Data is generated by a myriad of sources at an unprecedented rate, and the systems put in place to classify this data must keep pace. This dynamic nature necessitates a system capable of not just initial classification but constant reevaluation and reclassification to ensure data handling remains aligned with its current state.

MULTI-CLOUD CHALLENGES

Cloud-native data classification services such as Amazon Macie and Google Sensitive Data Protection (gSDP) present significant challenges for organizations trying to avoid cloud lock-in. These services are tightly coupled with their respective cloud platforms, which complicates operations in multi-cloud or hybrid environments.

Organizations face inconsistent data classification coverage across different platforms, leading to potential security gaps. Each cloud provider has its own methods and standards for data classification, resulting in variability in how sensitive information is identified and protected. This inconsistency can cause security teams to miss critical vulnerabilities or misclassify sensitive data, exposing the organization to potential data breaches and compliance issues.

Additionally, security teams must manage increased operational complexity due to the need for multiple tools and platforms. In a multi-cloud or hybrid environment, teams often have to juggle various security solutions, each with its own interface, functionality, and configuration requirements. This fragmentation can lead to inefficiencies, and human error. It also demands more resources in terms of training and staffing, as personnel need to be proficient in multiple systems.

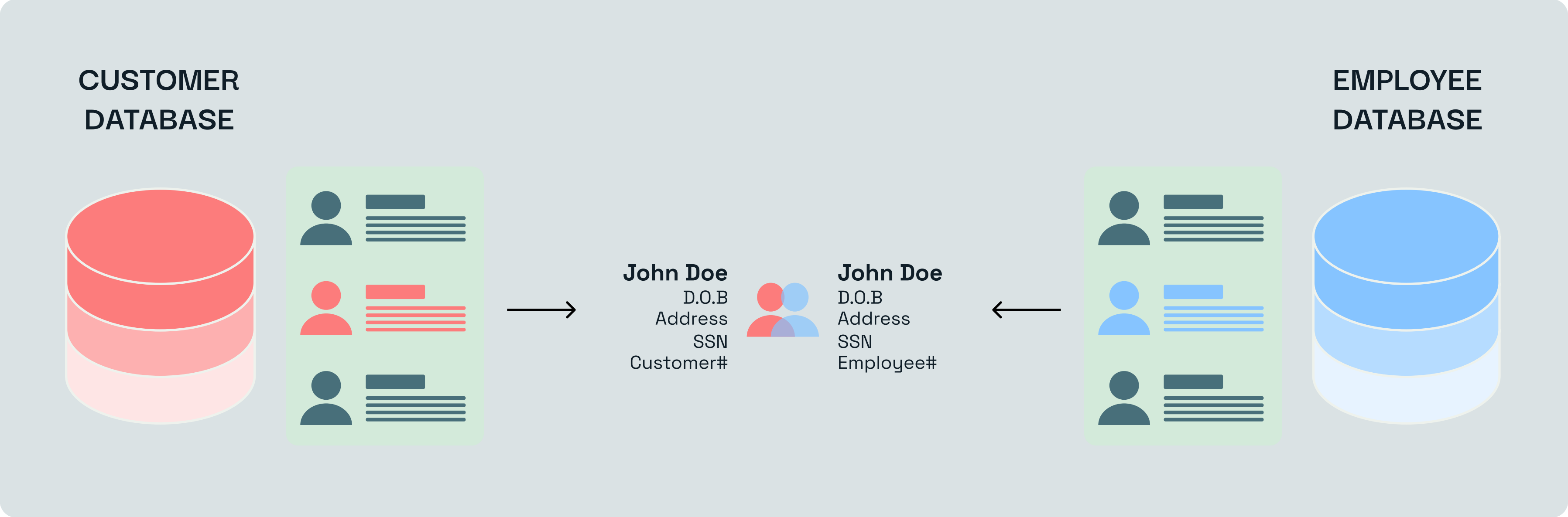
Aggregating security insights and metrics also becomes difficult, hindering a comprehensive understanding of the organization's security posture. With data spread across various platforms and tools, compiling and correlating security events and metrics is a daunting task - driving security teams in search of a mythical single pane of glass.

SENSITIVITY IS NEVER REALLY DETERMINISTIC

The nuances that distinguish sensitive from non-sensitive data are often subtle and context-dependent. However the type of data can play an outsized role in accurate classification. It is easier to classify data in structured data formats such as RDBs or REST APIs because the rigidity associated with the structure allows finding high-confidence contexts faster that can be used deterministically.

As an example, column names within RDB tables provide greater context as compared to a free-flowing text file where several strings are candidates for being the relevant context. Even with appropriate context in structured data formats, differentiating between sensitivity levels of various data types, such as synthetic vs. training data or employee vs. customer data, can be impossible to determine entirely deterministically.

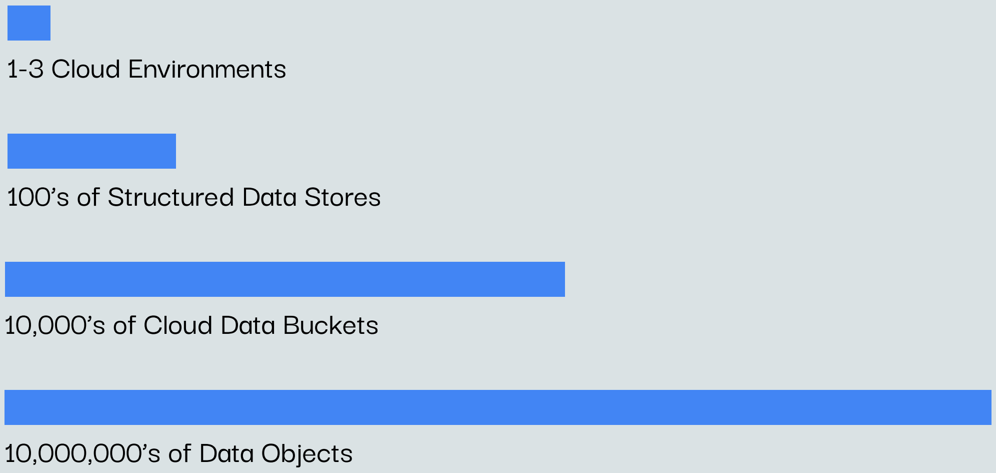
Synthetic data, for instance, may not carry the same privacy implications as real customer data, but in certain contexts, it could appear to be just as sensitive. Similarly, distinguishing between employee and customer data requires a deep understanding of both the legal landscape and the organizational context, making sensitivity determination an area where complexity and ambiguity are the norm.

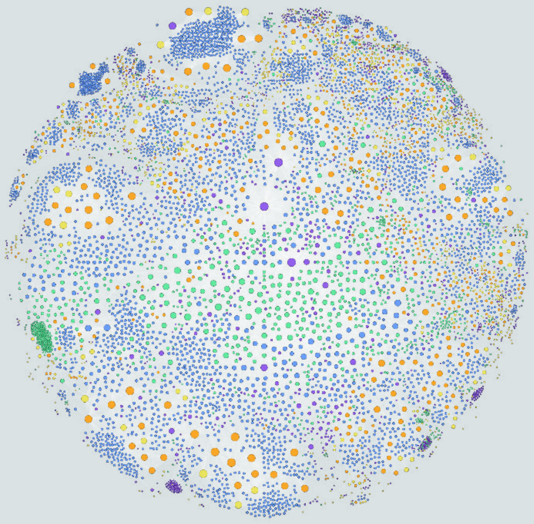


OVERHEAD CREATED BY FALSE POSITIVES AT SCALE

One of the most daunting challenges in data classification is the inevitability of false positives during the data identification phase. Even with advanced algorithms and sophisticated detection methodologies, no data classification system is immune to inaccuracies. At the scale of data objects we've seen in large organizations, where data often reaches into the hundreds of millions of data objects, and billions of identifiers within those objects, a false positive rate of even 0.001% can result in a million potential false positives that require manual review.

THE SIZE AND SHAPE OF A TYPICAL ORGANIZATION





This not only imposes a substantial resource burden on the organization but also increases the risk of overlooking genuinely sensitive or critical data amidst the volume of false positives.

When a system inaccurately flags data as sensitive or misclassifies it, each instance then requires manual intervention to correct. This necessitates dedicated personnel who must sift through potentially millions of inaccuracies, a task that is not only time-consuming but also diverts skilled resources away from other critical functions. The time spent reviewing and correcting these false positives is time not spent on proactive data security measures, strategic analysis, or innovation. Moreover, the cognitive load and decision fatigue associated with continuous manual review can lead to errors, further exacerbating the issue by potentially overlooking actual sensitive data amidst the sea of false alarms.

The energy expended in managing false positives also has broader implications for an organization's security. It can slow down addressing real security issues, leading to delays in decision-making and a decrease in agility.

Furthermore, the inefficiencies bred by handling false positives can further escalate costs. Beyond the direct costs associated with increased labor for manual reviews, there are indirect costs tied to delayed projects, slower time-to-market for products and services, and the potential for decreased morale among teams burdened with tedious, repetitive tasks. This not only affects the bottom line but can also impact an organization's competitive edge and innovation capacity.

In addressing the challenge of false positives, organizations must seek a balance between sensitivity in data classification and practicality in its application. This involves selecting the best performing classification tool not only in number of matches identified, but the type of data identified and the precision of the classification.

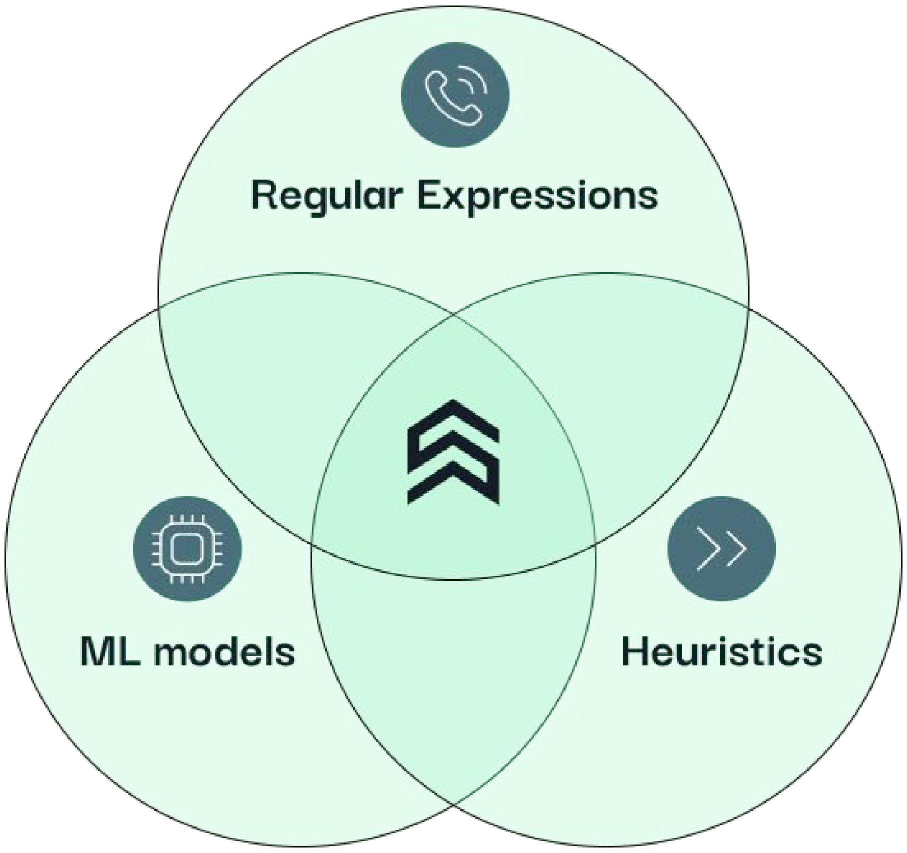
Organizations should also seek to continually refining the precision of classification algorithms, implementing tiered review processes to efficiently manage potential false positives, and continuously training machine learning models with feedback from manual reviews to improve over time.

Reducing the incidence and impact of false positives is pivotal in streamlining data classification processes, ensuring that the time and energy of security teams are invested in enhancing the organization's data security posture rather than correcting missteps.

Symmetry’s Approach

Organizations must balance the imperative of thorough data classification with the operational costs, seeking efficient ways to minimize expenses without compromising on the rigor of their data classification practices. This necessitates the deployment of highly efficient, automated systems capable of processing and classifying vast datasets at scale without sacrificing accuracy, a requirement that can strain even the most advanced data management infrastructures.

Data Classification at Symmetry is accomplished by finding pattern matches and leveraging context to justify the match. Determining pattern matches uses several techniques which include:



ML models: We use properties of the target match to build a custom machine learning model or leverage existing open-source models for detection. ML-based models are particularly useful where the pattern is hard to capture with regular expressions or heuristics.

Regular expressions: We use Regex to identify unique patterns such as a phone-number or a Credit Card CVV can be easily and definitively defined. The precision of Regex is increased by combining additional context.

Heuristics: A heuristics-based approach is also used to detect sensitive data which needs a deeper validation than a simple regular expression match but a machine learning model may be an overkill in terms of both computation and model maintenance.

With Symmetry, customers can quickly create their own identifiers using regular expressions and also leverage context specification for a robust detection mechanism. Moreover Symmetry’ multicloud-native approach addresses these issues by providing a single, centralized platform to classify, monitor, and protect data consistently across all major public clouds and on-premises environments, enabling organizations to unlock the benefits of a multicloud strategy without compromising on the security and governance of their most sensitive data assets.

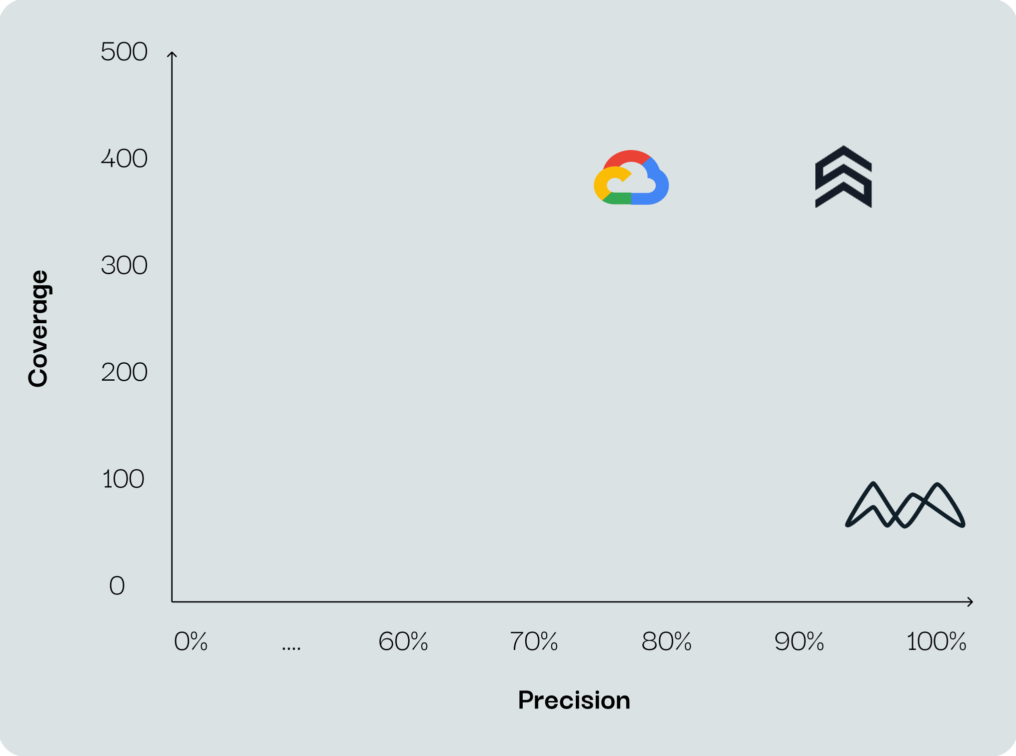
How We Perform Against Others

We evaluate our approach against two of the most well-known data classification services in the cloud: Amazon Macie and Google Sensitive Data Protection (gSDP). To do the evaluation, we use Synthea, a synthetic data generation tool which generates sensitive data especially for the Healthcare domain. Using Synthea, we generate 100 JSON files, each containing sensitive data for a fictitious person. The total size of the dataset is approx 300MB. To do the evaluation, we focused on easily verifiable US-specific and dataset-relevant identifiers that are supported by the three data classification services. This includes identifiers such as: Person's name, Social Security Number (SSN), Income Tax Identification Number (ITIN), Credit Card numbers, US-National Drug Code (NDC), and others. We'll explore more of these identifiers in future iterations of this research.

We evaluated each of the services on two axes:

Coverage - A measure of how effective the services are at identifying matches. Calculated by counting the volume of matches to sensitive data types across the files in the dataset at various sampling sizes.

Precision - A measure of the accuracy of the classification. Determined by anomaly selecting 100 matches from the previous exercise and evaluating the correctness. Precision is then defined as the percentage of correct matches divided by the total number of matches.



COVERAGE

For each service, we ran the classification engine against 100 files to determine how many matches each Classification Service would identify out of the box. At this stage of the comparison, we were not looking to determine the precision of these classifiers. This means that we made no comparison to how many identifiers were expected to be included in the samples. This is the same circumstances our customers would operate in and we wanted to replicate this experience.

We also leveraged four different file sampling strategies where we sample 10KB, 100KB, 1MB, or the entire file, to identify the matches and demonstrate the impact of a sampling approach. Table 4 outlines the coverage observed for each of the three services.

TABLE 1 IDENTIFIED CLASSIFICATION MATCHES (BASED ON FILE SIZE)

Sampling size (100 files)	SYMMETRY	Google Cloud	Amazon Macie
10KB	268	327	99
100KB	316	349	100
1MB	374	386	100
Entire file	397	400	100

Digging deeper into the coverage of identifiers identified by the services, the primary difference was in the types of data (or labels) that each service was finding matches for across 100 files.

TABLE 2

IDENTIFIED CLASSIFICATION QUANTITY BY TYPE (10KB)

	Phone #	Name	Birthdate	Taxpayer ID	Medical Institution	VIN	Credit Card #	US Driver's License	Medical Illness
	99	0	99	29	3	0	0	25	13
	86	89	99	49	0	2	2	0	0
	0	0	99	0	0	0	0	0	0

The varied classification coverage yielded some intriguing discrepancies. Google identified person names, vehicle identification numbers (VINs) and Credit Card Numbers. However on closer inspection, this wasn't a gap in coverage from the others, but rather false positives. Conversely, only Symmetry detected medical illnesses, Medical Institutions and US Driver's License. This highlights that Symmetry's data classification service is better suited for identification and classification of healthcare data - identifying twice as many data identifiers accurately within the dataset.

PRECISION

For each service, we randomly chose 100 matches across all matches identified by the service to assess the accuracy of the classification. This was performed manually to emulate the pain that a customer would need to experience in evaluating the accuracy of these classification results.

With regards to evaluating the Precision for each service, we note the following:

- gSDP identifies a large number of false positives with respect to Person-Names.
- When analyzing entire files, gSDP resulted in 50+ false positives for Credit Card numbers which are absent in any of the files in the dataset. Symmetry and Macie do not show any such matches.
- Symmetry identifies Medical Institutions and Illnesses, while neither Macie nor gSDP has any such data type.

TABLE 3

PRECISION OF DATA CLASSIFICATION

Sampling Size	SYMMETRY	Google Cloud	Amazon Macie
10KB	100%	74%	100%
100KB	99%	78%	100%
1MB	95%	72%	100%
Entire file	92%	82%	100%

The Data Classification Workload At Scale

Extrapolating the results to the number of files that a midsize hospital would expect to create every year, the challenge facing security analysts dealing with these results is mind-blowing

EXAMPLE

Okay, let's combine and simplify the calculations for the number of patients and documents a midsize hospital would have in a year:

Total annual patient days

Assumptions:

35 admissions = annual avg. per bed

5.8 days = avg. length of stay

150 beds = avg. midsize size hospital

Calculations:

150 beds x 35 admissions = 5,250 admissions

5,250 admissions x 5.8 days = **30,450 patient days**

Total annual documents

Assumptions:

5,250 admissions = annual avg. per hospital

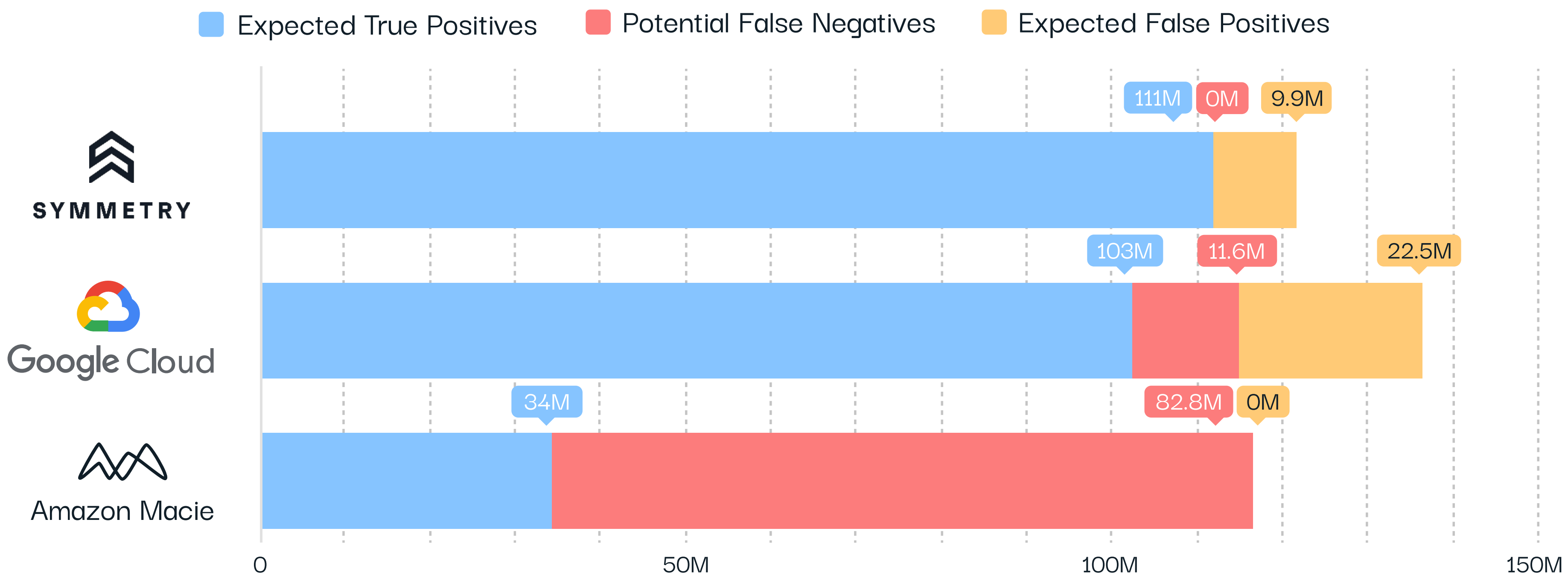
50 documents = avg. per admission

50,000 documents = avg. annual administrative docs

Calculations:

5,250 admissions x 50 documents = 262,500 documents

262,500 + 50,000 = **312,500 total annual documents**



Other Considerations: We did note that the classification run on the dataset took considerably longer with a single instance of Symmetry compared to the time each of the cloud services take. As a cloud service provider, both Amazon and Google are backed by an infinite scale compute power, so it is hardly surprising that their classification throughputs are much faster. Symmetry is however designed to scale horizontally and therefore derive greater throughput with appropriate resource allocation.

Conclusion

Effective data classification is crucial for modern organizations, but it presents significant challenges. The rapid growth of data, the nuanced nature of data sensitivity, and the high volume of false positives create a complex problem that traditional tools struggle to solve.

Symmetry Systems' data classification approach offers a compelling solution. By leveraging advanced AI and a unique deployment model, Symmetry delivers comprehensive coverage, high precision, and seamless scalability - while integrating with an organization's existing security controls.

As demonstrated in this white paper, Symmetry outperforms data classification services from both Amazon Macie and Google Data Sensitive Data Protection services in both accuracy and breadth of coverage. This translates to tangible benefits for Symmetry customers, including reduced operational overhead, faster time-to-value, and increased confidence in the security of their most sensitive data assets.

Symmetry's innovative approach positions it as a leader in the evolving data security landscape. By empowering security teams to focus on strategic initiatives rather than getting bogged down in data identification and protection, Symmetry enables organizations to truly unlock the potential of their data and innovate with confidence.

About Symmetry Systems

Symmetry Systems is the Data+AI Security company. Our platform is engineered specifically to address modern data security and privacy challenges at scale from the data out, providing organizations the ability to innovate with confidence. With total visibility into what data you have, where it lives, who can access it, and how it's being used, Symmetry safeguards your organization's data from misuse, insider threats, and cybercriminals, as well as unintended exposure of sensitive IP and personal information through use of generative AI technologies.

Symmetry works with structured and unstructured data in all major clouds (AWS, GCP, Azure), SaaS storage services (e.g. OneDrive), and on-premise databases and data lakes. It is deployable in the most strictly regulated environments; as a read-only service, it inherits all your security and compliance controls (e.g. FedRamp). That's why the most innovative Fortune 50 financial service providers, manufacturers, pharmaceutical companies, and federal agencies rely on Symmetry to protect their crown jewel data.

Powered by best-in-class AI, Symmetry provides organizations with the necessary toolkit to minimize data posture risks, demonstrate compliance, and react to threats and policy violations in real time. Symmetry solves challenging problems for customers with ease, ranging from classifying custom data types, reducing data blast radius and attack surface, detecting ransomware attacks, enforcing least-privilege access, and more.

Born from the award-winning and DARPA funded Spark Research Lab at UT Austin, Symmetry is backed by leading security investors like ForgePoint, Prefix Capital, and others.

Learn more about how Symmetry Systems can enable you to bring "symmetry" to the asymmetry between cyber attacks and defense at www.symmetry-systems.com or follow us on Twitter or LinkedIn.

Innovate with confidence with Symmetry.