How AI Uncovers the Human Profiles in Your Cluttered Data

NetApp

Executive Summary

Data privacy is more than just complying with regulations like the European Union's General Data Protection Regulation (GDPR)—it's about protecting the people who rely on your organization and your business. The problem is finding a way to sort out the human profiles from the rest of the noise, often in extremely cluttered data environments. Data privacy in this sense is as much a data governance challenge as it is a security concern.

Thankfully, there are new technologies to help address these concerns. Powerful artificial intelligence (AI) tools can help discover, categorize, classify, and report on the personal data creating human profiles.

This ebook will explain how AI is providing new ways to help organizations to discover their hidden human data, compare the different AI methods being used to categorize this data, and discuss the breakthroughs in AI technology that are making this technology more accurate than ever to protect both your organization and the people at the heart of it.

Table of Contents

Introduction

Companies store enormous amounts of data. Where operations are dispersed, that data can be stored in many locations across multiple systems. Managing and securing structured databases and semi-structured data is one thing, but dealing with petabytes of unstructured documents requires a whole new level of data governance. Moreover, that data has begun to reach out beyond the enterprise data center: the cloud is now a mandatory extension of almost every organization's IT infrastructure, with more and more companies relying on services from multiple cloud providers in addition to on-prem environments. The challenge of gaining visibility and control over such sprawling hybrid data sets is daunting.

With the expanded scale of data governance requirements, these days organizations are pressed more than ever to fulfill their fundamental need to manage and secure personal information. There is more at stake than just the fines and data privacy regulations. Behind the data that is regularly being collected or processed are real people—the same people who organizations rely on.

Organizations and people depend on each other. If organizations can't ensure that they can manage their users' information intently, they won't just face serious fines, they risk losing the trust of the very people who make their business possible.

Companies and organizations now realize the value of personal data.

For the security and privacy teams responsible for ensuring this data is protected the job can sometimes seem impossible. There is just too much data to responsibly monitor manually. The only real solution is to turn to technology that can sort out the human data from the rest of the clutter. The solution is in artificial intelligence.

This ebook will demonstrate how artificial intelligence can be used to map and secure sensitive data, from specifying several types of sensitive information to concrete methods of detecting their occurrences in organizational data. First, let's start with a simple question: What is sensitive/personal information? And how would you map it?

Part One

How Privacy Works in the New Age of Data Management

Mapping Personal Information

Three main questions pave the way to face this challenge efficiently:

1. **Which types of information could be relevant?**
2. **In what form or shape do they occur?**
3. **How to detect their occurrences accurately?**

File-Level Data Management

Off the top of your head, can you remember all the personal information that exists on your own PC? A brief glimpse at your Documents or Downloads folders may give you a hint of how quickly random documents and forms containing personal information can pile up. Legal agreements, medical examinations, and personal letters are all laying there comfortably in your hard drive, totally exposed. Remember, this is just your information.

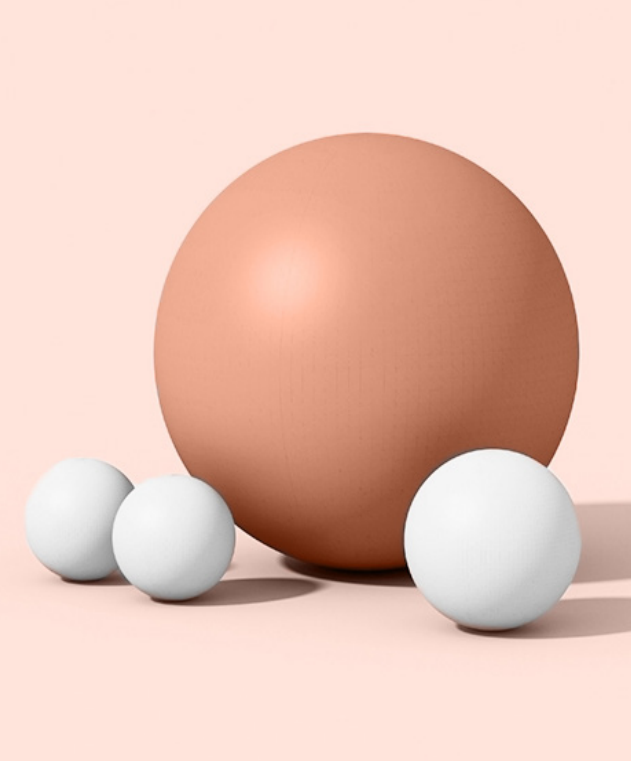
Now, imagine what a challenge it is for enterprises to manage documents across the entire organization, within every data silo: end-points, servers, and cloud storage. Organizations not only need to take proper care of personal information, but also need to secure sensitive business-related information.

For an AI to begin to sort through all of this information, the abstract term “sensitive data” must therefore be broken down into more concrete, basic components:

Reason for sensitivity
Can be perceived as independent layers: personal data, confidential by organization/ department or generally sensitive (i.e. legal agreements)

Common form
Different layers of types of data interchangeably emerge in patterns and contents of varying sizes: single word, sentence, paragraph, or an entire document.

Once a clear definition of sensitive data is established, the process of detecting files and mapping sensitive data becomes more intuitive. For example, a given file stored in a cloud repository may contain a single data set (a Social Security Number for example) that signals its potential sensitivity—disclosure of personal data. Another file on a salesperson’s PC might contain a paragraph that reveals a company’s confidential operations process.



Methods for Detection

Manual Labeling: An Enormous Effort

Technically speaking, the most precise way to find sensitive information would probably be to manually label relevant occurrences. In other words, every file created will have to be passed through the eyes of a **domain expert**: a professional capable of determining whether the document may be sensitive, with respect to its context and domain. Ideally, this person would mark relevant words or sections, and specify precise reasons. For example, this section contains the customer's full name, as well as their home address.

In practice, this approach is unrealistic because it requires:

- Massive dedication of time and attention to label documents.
- High familiarity with data protection regulations and enterprise confidentiality.
- Old, received, or automatically generated documents all require examination by a relevant domain expert.

As a reference, try to think of organizations that deliberately and dedicatedly label each and every piece of data manually upon creation. The immediate examples coming to mind are usually national security agencies, military units, and perhaps embassies. The importance of data protection for them is very clear, often with specific labels required by law, and the risk posed by information leakage is severe enough to demand the time and effort required for manual labelling.

AI Labelling: Rule-Based Vs. Machine Learning

A more scalable and generalized approach can involve the use of automatic methods that would require a lesser amount of human involvement in the process of detecting relevant information. Some methods are based on exact matches of terms or series of characters. Other methods rely on statistical models based on occurrences and adjusted frequencies of terms or entities within a document or paragraph. More advanced methods even attempt to utilize the context of every word in a sentence to predict whether it's relevant or not, using deep learning and natural language processing.

Different methods can solve different tasks:

- **Rule-based** methods can be applicable to detect and match the occurrences of a variety of IDs, credit cards, and email addresses.
- **Statistical based, machine learning** methods are useful to categorize documents according to their content.
- **Context-aware, deep learning** methods can aid in extracting personal names from sentences or detecting context-dependent phrases.

Protecting Privacy Is Changing How Companies Work

Managing and securing sensitive data is an ambitious and challenging goal. To gain more insight and control over this data with a view to complying with data privacy regulations, organizations can attempt to map the sensitive information that exists in various forms and shapes in their data sets. Manually, it's an impossible task, but with the right set of advanced artificial intelligence tools, this goal may be achieved much faster.

Now that you have some insight into the concepts that lay the foundation of mapping sensitive information, let's dive deeper into the practical methods that can be used towards protecting the sensitive personal data in your repositories.

Part Two

Document Categorization: Machine Learning Vs. Rule-Based Methods

Data is growing at a phenomenal rate, for every type of organization. With that expansion comes a renewed need to make sure that sensitive personal and business information isn't in risk of being exposed.

The preceding section introduced the challenge of identifying and understanding the sensitivity of information within unstructured and structured data across enterprise storage silos. How are companies defining what sensitive data is? What kind of technologies can they rely on to identify this information to help make sure that they aren't exposing intimate details about people?

As you can guess, some pieces of information are more difficult to detect than others. The following sections will illustrate this point more precisely by outlining some approaches to detect, map, and categorize certain sensitive data. These approaches include rule-based methods and two forms of machine learning methods: supervised learning and unsupervised learning (clustering).

Document Classification with Rule-Based Methods

When it comes to detecting sensitive data, the straightforward approach is based on defining exact patterns of numbers, characters, or specific terms, also known as regular expressions. Once those definitions are in place, every piece of data containing such patterns can be detected directly. This practice is quite common when it comes to matching ID numbers, credit card numbers, and email addresses, for instance.

However, this method by itself is usually insufficient, for several reasons:

- **Manual effort**
  Tailoring terms and rules requires a lot of human labor.
- **False positives**
  Not every string of nine digits is a Social Security Number.
- **Lack of context**
  The surrounding sentences are ignored when determining the validity of any given match. For example, a rule-based method will identify a telephone number, even one that appears in an ad for office furniture and is obviously not a threat to anyone's privacy.

Some data management solutions rely solely on this type of partial detection, while claiming to use "advanced artificial intelligence." That's not necessarily accurate.

Document Classification with Machine Learning Methods

Document Categorization with Supervised Learning

To achieve a holistic and meaningful data mapping, the ability to automatically categorize files according to their content is a huge milestone. This can be possible with the help of a specific type of machine learning: supervised learning.

Supervised Learning is the machine learning task of inferring a mapping between data inputs and outputs based on ground-truth samples of input-output pairs.

How it works in a nutshell: Natural Language Processing techniques such as bag of words and word embedding enable the transformation of each arbitrary piece of text into a fixed-size numerical vector representation. Then, machine and deep learning algorithms are trained with the input of labeled data samples—sets of texts (documents) and their corresponding labels (categories).

In other words, a supervised machine learning pipeline uses a small set of labeled data in an attempt to generalize the essence of the categorization task in order to correctly classify even never-seen-before sets of documents.

Out-of-the-Box Categorization
Many types of documents are common in organizations: invoices, NDAs, resumes etc. For an AI system to detect and identify such documents correctly:

- The documents' content (text or image) can be scanned and interpreted.
- No custom specification is required from the organization.

For this purpose, a ready-to-use machine learning based categorization mechanism can be trained to detect dozens of predefined categories.

Custom Categorization
In addition to typical document categories, every organization is also likely to have its own unique way to divide categories of interest. There's a flexible solution to do this that only requires each organization to initially train a supervised learning model that learns to differentiate and categorize according to the unique organizational policy, using uploaded documents as data samples for the model. Later, more categories and samples can be added to update the model's predictive ability over time. Yet, such a solution may introduce some further challenges for the organization:

- How to figure out which categories exist?
- Where to obtain training samples?

Any experienced data scientist would tell you that collecting and labeling training data is one of the most crucial yet demanding steps in developing machine learning solutions. Not surprisingly, even with frameworks that only require data samples, people in charge of managing and protecting their users' data sometimes find it difficult to gather a representative set of samples of documents from different departments.

But what if there was a way to more easily detect and group together potential samples that exist in the actual data? Just a peek at the results enables companies to identify document groups that they never knew they had to protect.

Document Clustering with Unsupervised Learning

There is another unguided document grouping made possible with machine learning methods. A holistic and meaningful mapping of a vast number of instances can be achieved using clustering, a form of unsupervised learning.

While in supervised learning the training data is labeled with relevant classifications, in unsupervised learning models are to learn relationships between data points and classify the raw data without guidance or “ground truth” provided.

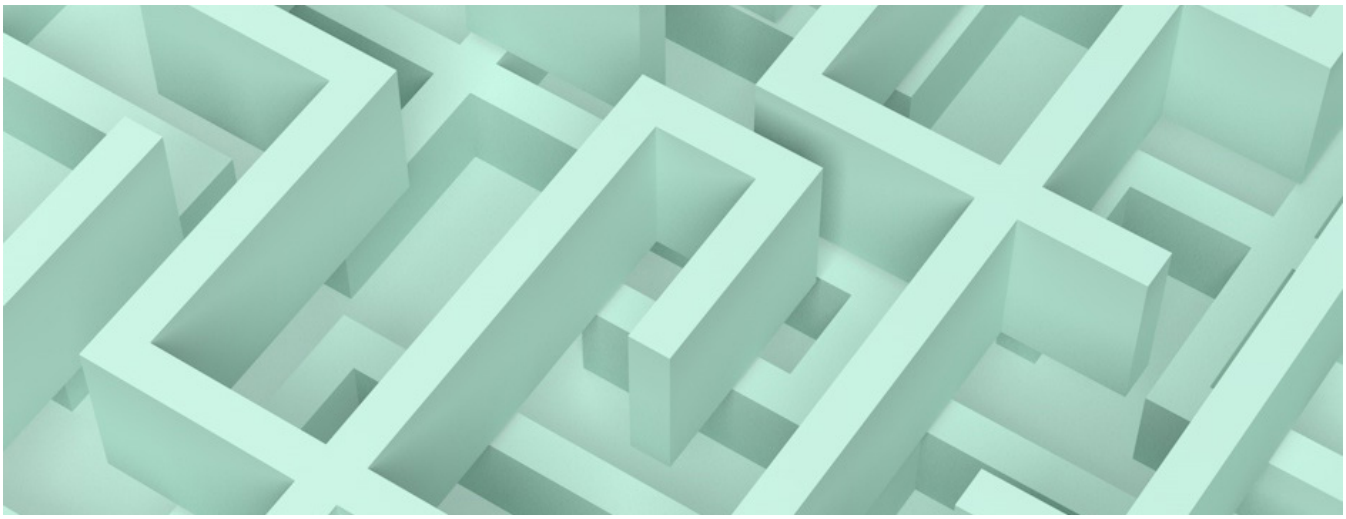
To illustrate how this mechanism works in our domain, think for a moment about all the millions, or perhaps billions of documents in your organization. If all documents were represented by points in a vector space, so that similar documents are within a close distance from one another, a set of statistical algorithms can be used to group together such similar documents.

This is just a simplified look at this method, and of course it takes a delicate process to represent documents’ meaning in numerical vectors, so that the proximity between vectors genuinely reflects the similarity between the documents’ meaning.

The Final Word in Machine Learning Vs. Rule-Based Categorization

Machine learning offers methods that can greatly improve on rule-based methods to detect, map, and categorize sensitive documents on an organizational level. It should be noted that when mapping sensitive information at enterprise scale with any of these methods, there are many instances that require organizations to take into account the different possible contexts that can be attached to pieces of information. For instance, not all sensitive pieces of information come in the size of a full document.

There’s a lot more to dive into here. The next part will introduce how machine and deep learning are beginning to utilize context towards automating sensitive data classification.



Part Three

Data Privacy and Compliance - It's All About Context

Data is changing. Companies no longer just generate data as an output of their activities, but in many cases, also manage the data—including sensitive personal data—of their entire user bases. This growth, largely made possible because of the increased storage scale that the cloud provides, comes with all new concerns. The major concern for companies today is how to ensure personal data is identified and treated appropriately.

This section addresses the data identification tasks that are considered more challenging to automate, since they are located on the edge of artificial intelligence abilities: understanding context.

Privacy regulations such as the GDPR and CCPA define personal information quite broadly, thereby making context crucial to be able to accurately identify relevant data within the mountains of information that organizations store.

Sentence-Level Data Management: Utilizing Context

There are a few questions that enterprises can ask themselves to gain intuition on how AI is going to help in data management compliance:

1. **Can you estimate how many names of people are mentioned in all the documents on your storage devices?**

2. **Which documents mention a specific company or project?**

3. **Can you immediately retrieve documents that mention a specific person?**

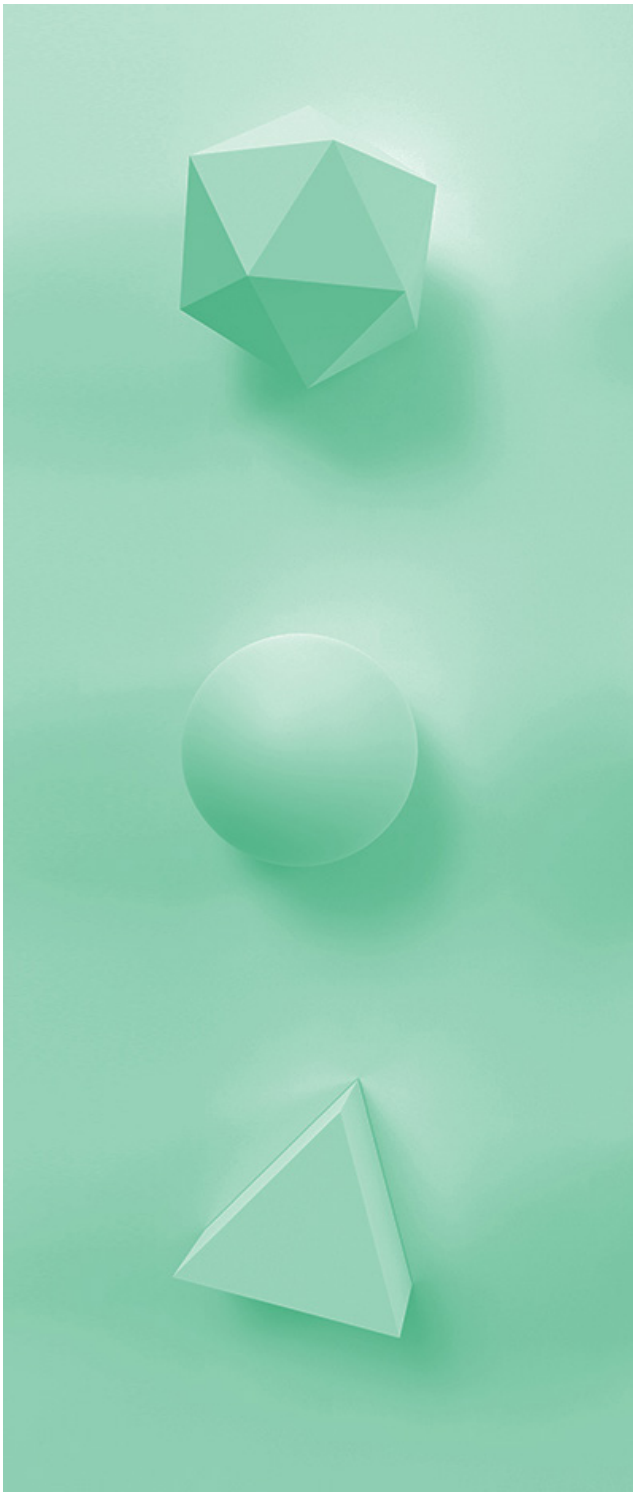
Mapping Personal Data: Named-Entity Recognition

The ability to retrieve information quickly using search engines involves some form of indexing the searchable domain. Mapping content from petabytes of free-text, out of files which are not always accessible (for example when a server is down), is an ambitious engineering challenge. Should a search engine index and store every single word appearing in every document into huge data structures just to retrieve given names?

What if you could somehow label entity names during parsing (person/organization/location), then index just the relevant entities with the documents referring to them? This is where **named-entity recognition** comes in handy.

Named-entity recognition is an AI method of extracting any specific mention of a named entity within a set of unstructured text. The named entities are classified into categories, such as personal names, organizations, and locations, according to the context in which they appear.

Using deep learning models, an named-entity recognition (NER) process can be automated with impressive precision and recall scores: each sentence is converted to a sequence of vectors, which are then passed forward through recurrent neural networks that were trained to locate and classify named entities using the entire sentence's context. The end goal is for meaningful names to be detected, indexed, and, ultimately, queried effortlessly.



Detecting names is useful—what about the accompanying context?

The context in which a name appears reveals more relevant information about the personal data that may or may not coexist with the mentioned name. Whether it's a spreadsheet of contacts, an employee performance review, or a consumer credit risk assessment, it's the context that tells the full story. Understanding context is also important when analyzing whether a company maintains inferences considered personal information under privacy laws.

Contextual Personal Information Detection

Among the data protection regulations, there is one that can be especially hard to comply with. This regulation specifies several categories of personal information for which processing is prohibited, including ethnicity, sexual preferences, political or religious views, and health background. In the GDPR it is defined as such:

In other words, any organization that wishes to comply with GDPR must find and treat accordingly the occurrences of such personal information residing in its data. Yet in practice, such pieces of information do not generally fit neatly in structured tables, but surface in unstructured texts in free form. Moreover, they can't be detected easily—can you think of a mechanism to find, for example, descriptions of a person's ethnicity or religious views?

Which of the following sentences might contain personal information about a person's ethnicity?

- Joshua has Italian origins.
- Joshua has Italian restaurants.

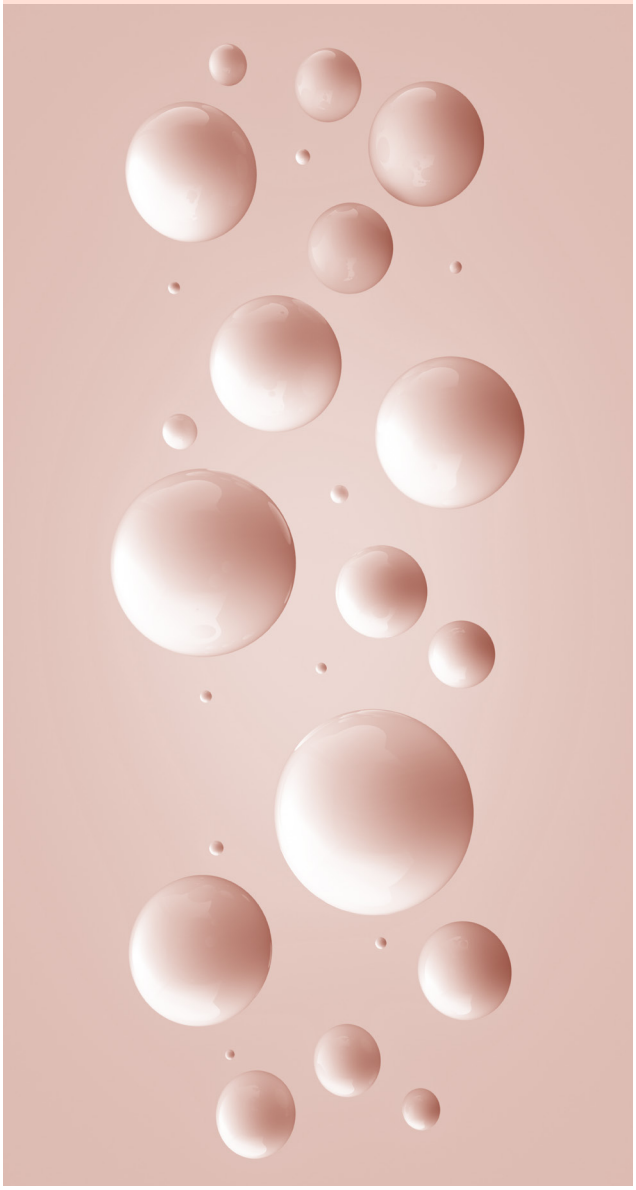
Of course, every English-speaking human would be able to figure this out immediately. However, training machines to differentiate between the contextual information in these types of sentences correctly requires not only linguistic understanding and the ability to parse sentences, but also some utilization of context.

Recent breakthroughs in natural language processing research have made it possible to extract context from text more effectively. For instance, take the concept of Word2Vec—a fixed meaningful vector representation for each word in a text. With this technology, every sentence can then be represented as a sequence of vectors. Applying deep learning methods, such as sequence-based neural networks, on those vector representations can be useful to achieve contextual representations of text: a deep learning model is trained to generate some "summarized" encoding for every input text, based on the sequence of word-vectors it contained.

As a result, deep learning models enable the automation of detecting personal information—all while taking word-level context into account.

"Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited. "

GDPR, Article 9

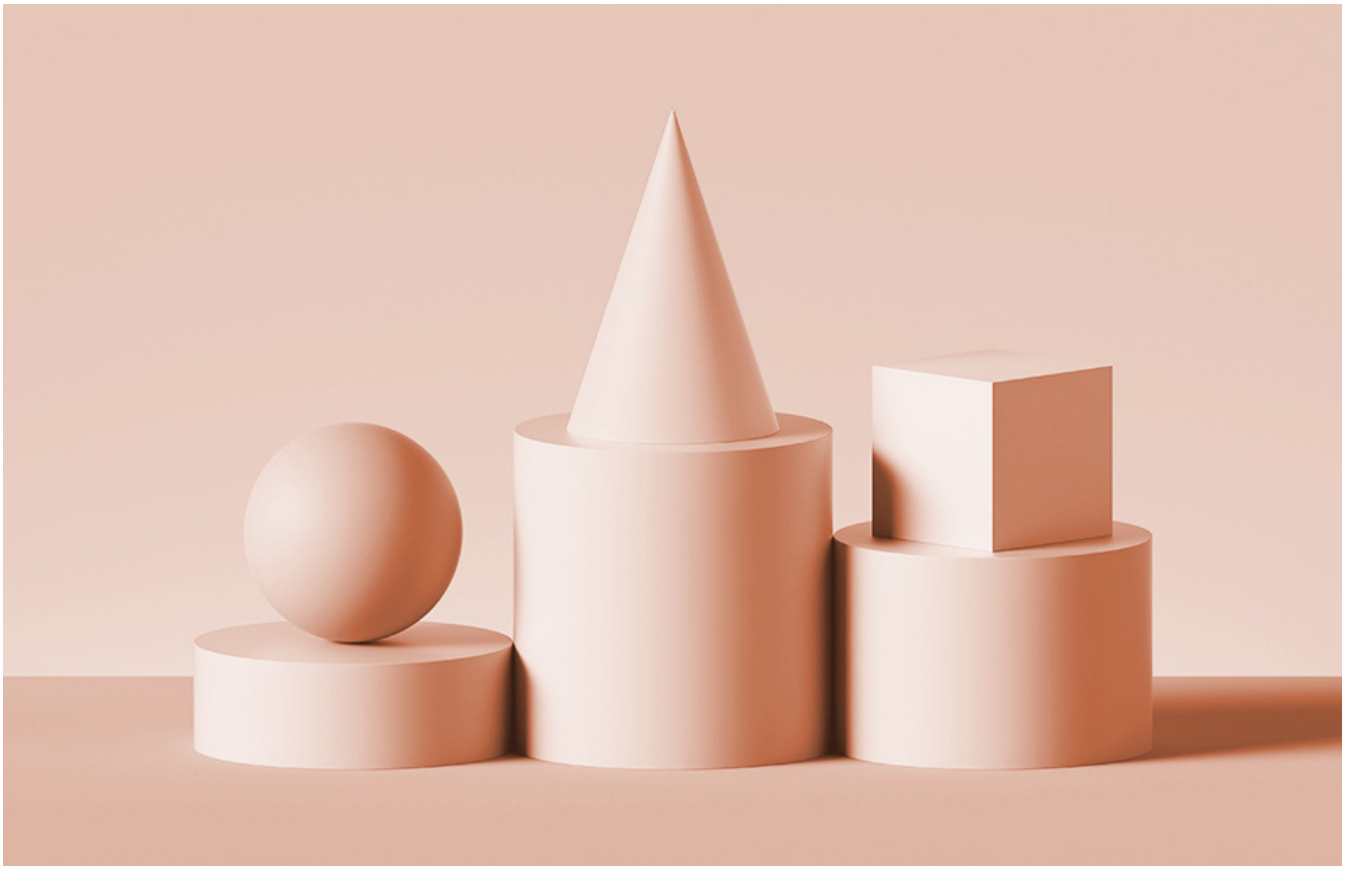


Personal Cross-Reference Resolution: The Missing Link in Data Compliance

GDPR also introduced an unprecedented right of access to data subjects. As GDPR Recital 63 states: “a data subject should have the right of access to personal data which have been collected concerning him or her, and to exercise that right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing”

That means that once a data subject submits a request for their personal data, an enterprise may need to expend an unbearable amount of effort to accurately retrieve the relevant information in order to comply with the regulation. Effective preparations for automated responsiveness would require a complete and efficient mapping of all relevant information as mentioned above in the section on mapping personal data.

Furthermore, often a data subject is referred to not by their name but by some kind of ID number(s) that could be stored in internal databases. In such cases, to retrieve all the relevant information, all the different personal references pointing at the same person need to be linked. Utilizing such a co-reference mechanism enables the company to search for documents using a person’s name, giving them the ability to reach not only documents that contain the full name, but also ones that refer to the person by their relevant ID.



Final Thoughts

Data governance is becoming harder to achieve via manual processes given the new, data-driven landscape that almost every organization faces today. The urgent need to identify and manage personal data of users makes this challenge even more important.

Since data governance at this scale is beyond human control, this ebook has detailed some of the AI technologies at work today to help map data subjects and identify pieces of personal data in today's sprawling IT environments. These solutions must be context-aware, taking into account accurate identification, and advancements and thorough language interpretation. And while achieving data governance is still going to be challenging for enterprises, eventually an AI with the right set of tools will make it much easier and faster.

NetApp is doing just that with NetApp Cloud Compliance, the new AI-driven data mapping tool that intelligently maps, classifies, and reports on data in your system so you can help protect the people who count the most in your data.

AI is changing the face of data management compliance. The pressure that enterprises are under to secure and define their sensitive information has lead to refining the different methods of data categorization, and given a new push to develop an AI that can understand context better than ever, so that complying with important data privacy regulations can be done automatically.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NA-000-0221