



---

# Making sense of your sensitive unstructured data risk: 5 questions to ask and answer

UNSTRUCTURED DATA IS EVERYWHERE. IT'S EITHER A BIG OPPORTUNITY OR A HUGE OBSTACLE TO SECURITY AND COMPLIANCE. WHICH WAY WILL IT GO? THAT'S UP TO YOU.

# Making sense of your sensitive unstructured data risk: 5 questions to ask and answer

UNSTRUCTURED DATA IS EVERYWHERE. IT'S EITHER A BIG OPPORTUNITY OR A HUGE OBSTACLE TO SECURITY AND COMPLIANCE. WHICH WAY WILL IT GO? THAT'S UP TO YOU.

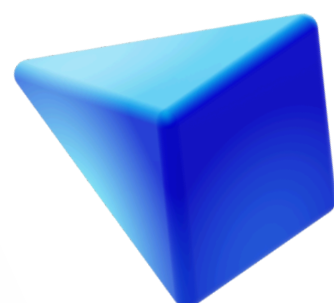
## Unstructured doesn't mean unprotectable

Up to 80% of your business data is unstructured, and this volume continues to grow daily. Every document, presentation, and PDF adds up fast. These files have always been more challenging to govern, manage, and secure, even though they may contain your most valuable and sensitive data.

Valuable data that's difficult to secure? As you can guess, attackers absolutely love it. And frankly, regulators and everyone else don't care about how difficult it is; they have high expectations around security, compliance, and privacy.

## What's this all about?

Confidential experts put this guide together to help decision-makers get the answers they need to improve their sensitive unstructured data security and usability.



# Who we are

Confidential was born from DARPA-funded research at SRI, developed by a team of cybersecurity engineers and PhDs dedicated to unlocking the massive potential of unstructured data—safely and securely. Built to meet the military's need for protecting sensitive and proprietary data, our technology was forged under the highest standards of resilience and control. With millions invested across two DARPA projects, SRI recognized the broader market opportunity and spun out Confidential in 2021, exclusively licensing all related IP to us. Today, we're bringing that same battle-tested security to organizations seeking to harness AI's power—without compromise, only control.

# Our Mission

Our mission is to be woven into the fabric of how organizations secure sensitive information within unstructured data. Confidential bridges the critical gap between identifying sensitive data at a granular level and truly protecting it—supporting and accelerating the entire journey from discovery to control. By making data secure and usable, we empower organizations to confidently unlock the full potential of their most valuable information, enabling AI workflows and pipelines to act responsibly and intelligently.



# Introduction

02

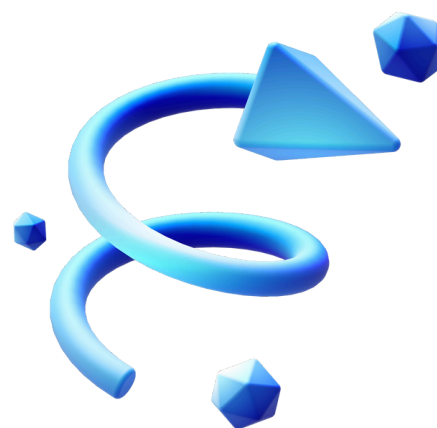
## Welcome to a great big pile of stress and opportunity

The first step towards assessing sensitive unstructured data exposure is finding and understanding all your unstructured data. It's a much harder job than scanning structured data for some reasons that also help us understand what makes unstructured data different. We'll take a minute to compare the two now.

## Compare and Contrast

**The easiest way to think about structured data is an Excel worksheet.** The data is organized in rows, neatly labeled, and easy to export and import. The modern 'relational' database is essentially a large set of these worksheets, but all the underlying data remains neatly organized and comprehensible.

**Unstructured data refers to information that is not contained within these structured documents.** You may have a structured worksheet containing 400 Social Security numbers (SSNs). You could also have 400 contracts in PDF format, each containing a Social Security number (SSN). Same risk, much different problem. You're not managing a list, but files that don't fit easily into a relational database.





# Everything gets **harder** with unstructured data

As you might guess, this makes managing unstructured data relatively challenging. However, it also means that finding and securing sensitive data will be much more complicated for a couple of reasons.

## More Locations

Unstructured formats are stored outside of centralized databases, sometimes per application. You may also have unstructured data stored in both on-premises and cloud environments.

## More Formats

Unstructured formats can range from PDFs and presentations to images, as well as audio and video. Scanning this information is a lot more complicated than ingesting a CSV file.

## More Value

Unstructured documents are often valuable precisely because they contain sensitive information. This may include contracts, presentations, technical documentation, financial records, and other key documents.

*This last point is important because it means the very documents you need to secure are the same ones that really bring critical information richness to other workflows, especially AI models and agents.*

*This makes this early identification and assessment more than a compliance exercise – it's where your AI pipeline begins.*

# The Questions

## The 5 Questions You Need to Ask & Answer

With so much at stake around finding and securing your sensitive unstructured data, where do you start? We'll go through five questions and examine why the question and its answer are so important.

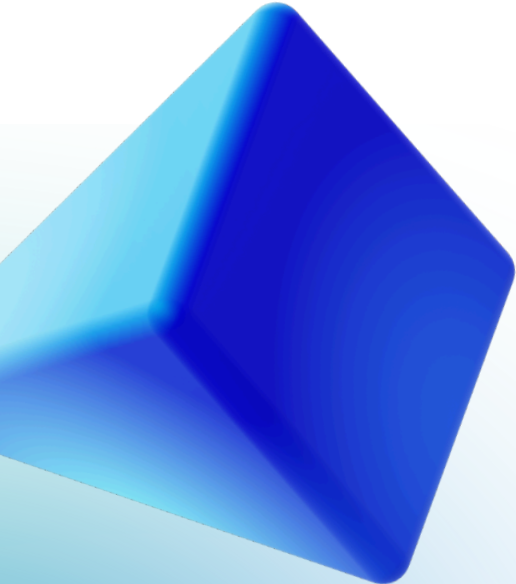
This is only a thought starter exercise – the real work begins once you're in the know. There are numerous tools available to make the job easier, but more on that later.



## QUESTION 1

# Where is it?

Finding sensitive, unstructured data across your environment can be the most complex step in the process. Once the data is located, you can point the tooling at it to explore and assess. This is where you get your first view of all the places and formats where your unstructured data lives.

- 
- ★ **This discovery might also reveal gaps between governance best practices and operational reality.** In theory, data governance strategies should be built on complete visibility. Operationally, especially with SaaS and shadow IT, that doesn't always happen. Clarity on your sensitive, unstructured data helps you align reality with best practices.
  - ★ **It also helps you understand which systems store, move, and share this data.** This is especially important because more and more compliance frameworks have requirements specifically designed to build and protect a purpose-built sensitive data infrastructure, and to document and demonstrate those assets, networks, and controls.

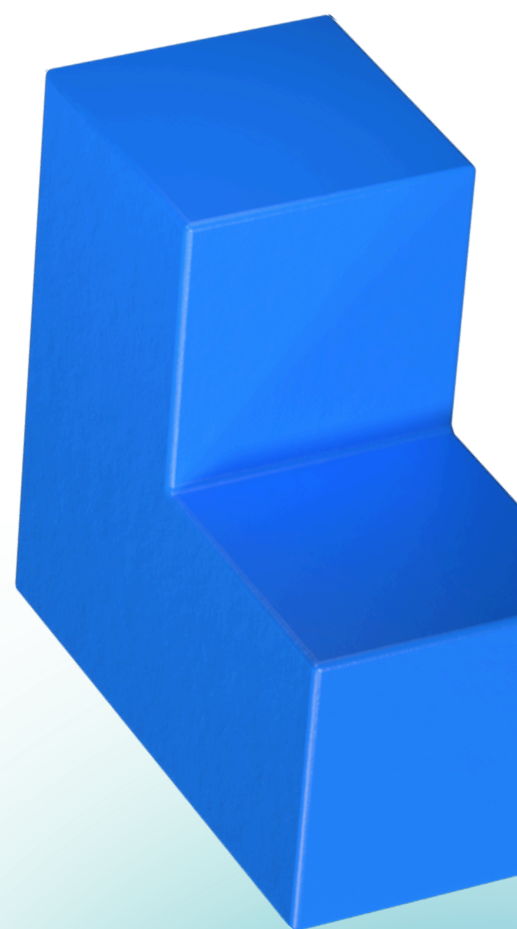
## QUESTION 2

# What's the artifact?

Sensitive unstructured data is always embedded inside that document, and this is what the business consumes. But what fields inside that document matter a lot to both security and compliance teams. Is it an employment application? A customer data file? The business purpose has both technical and compliance implications.

★ **The artifact determines how the data is used in business and is governed by regulators.** Those rules trigger specific requirements around everything from retention to encryption. For instance, EO 14117 requires securing employment and investment contracts to protect sensitive data. HIPAA, PCI, and the rest of the alphabet soup all focus on how data is used.

★ **It's also where AI planning and preparation begin.** Secure AI only occurs with a compliant information pipeline in place, covering both structured data and unstructured documents, which are lumpy, unpredictable, and oh-so-valuable. None of your AI plans can be realized until you secure your sensitive unstructured data.



## QUESTION 3

# What's the format?

While the artifact helps us understand how data is used, the format helps us understand how it's stored. Whether the file is in HTML, PowerPoint, PDF, or even JSON format, the file type determines how the information is read and managed.

- ★ **The format will also determine the tooling, if any, required to get at and extract the data.** This is especially true for LLMs and AI agents, as well as any other data-focused platform that requires configuration for new inputs. Data classification, tagging, and metadata might also change based on document type.
- ★ **The format must be mastered to enable continuous scanning going forward.** Meaningful compliance around sensitive unstructured data requires more than a snapshot in time. Scanning must be performed at regular intervals to ensure that new files containing sensitive unstructured data are appropriately identified, documented, protected, and eventually leveraged.





## QUESTION 3

# What's the age?

As data storage has become cheaper and easier, we are storing more and more data, especially within data lakes, which are frequently used for unstructured data. For many organizations, this can mean years of accumulated data, all stored and waiting for something to do. But understanding data age matters.

- ★ **First and foremost, retention periods matter (a lot).** Compliance requirements are heavily impacted by retention rules. This means that the age of an unstructured data file determines the level of effort required to secure it, often triggering the implementation of specific technical controls. Last, but not least, data deletion schedules, where they exist, must be maintained.
- ★ **Age also impacts how secure AI data pipelines are built.** When considering how information can be leveraged by AI models and agents, relevance is crucial, and this often means age. The same tiered data temperature strategies used to manage storage matter when building your AI data supply chains are also applicable.





## QUESTION 3

# What's the value?

Assigning specific dollar values to sensitive data is part art, part science. However, risk-based security and compliance strategies require this in order to prioritize defenses and allocate finite security resources and attention. It's a solid addition to the risk assessments nearly every regulator and underwriter now requires.

- ★ **80% of your data is unstructured**, and IBM and Ponemon helpfully remind us that the average cost of a data breach last year was almost \$5M. It's also fairly easy to estimate the value of certain data on the web. For example, "Fullz", or a full set of stolen consumer credentials, run from \$20-120.
- ★ **It's also harder to calculate opportunity costs if you can't use the data securely, especially with Gen AI.** Those amazing models that will churn through contracts and financials and assess versus deep industry benchmarks can't get started if there's no way to protect sensitive data along the way.



# Getting the answers

**Moving from uncertainty to confidence around sensitive, unstructured data begins with understanding what needs to be protected and doing so without disrupting data operations or AI initiatives.** Only technology can help teams discover at the scale required by the enormous volume and variety of the average enterprise data store.

However, not all platforms are created equal, especially when it comes to securing sensitive, unstructured data. They should all be able to give you answers to our questions, but that's simply a baseline. **You should also look for platforms that:**

- Combine discovery and classification with proactive threat mitigation
- Scan across most unstructured file formats, on-prem and in the cloud
- Offer customizable search capabilities that expand past embedded support for PCI, HIPAA, and other frameworks
- Let you evaluate sensitive unstructured data at the field level

**No matter which platform you choose, the breadth and depth of visibility must be your top priority, especially with unstructured data.**



# See what Confidencial.io can do

To help organizations gain better visibility into their unstructured data, Confidencial is offering a free, no-obligation scan of your documents to report on:

- Types and volume of sensitive and regulated data discovered
- Classes
- Specific sensitive data fields and their locations
- Document age, format analysis, and exposure risks
- Economic impact calculations, including estimated breach costs and the potential value of exposed data on the dark web

**Let's get the answers you need to find and protect your sensitive unstructured data. Get in touch today.**



[confidencial.io](https://confidencial.io)



[linkedin.com/company/confidencial-inc/](https://linkedin.com/company/confidencial-inc/)



[hello@confidencial.io](mailto:hello@confidencial.io)

