



ICDIGITAL



# BETTER DATA CLASSIFICATION FOR BETTER DATA SECURITY

## > SUMMARY

1. Data classification is foundational for data protection. Without understanding your data, trying to apply the right policies and controls to protect it will be challenging.
2. Data classification programs require several key elements for success: Classification Policy, Scope, Discovery, Solutions, Analysis and Audit.
3. Effective data classification doesn't have to be complicated. Many effective classification programs rely on only 3 levels and start with data in active use.
4. Digital Guardian gives you flexible data classification optimized for the most common use cases, eliminating the need to modify your business processes to protect your most critical asset.

## > WHAT IS DATA CLASSIFICATION

Data classification is a process to categorize documents based on specific criteria driven by data governance, business and regulatory compliance (PCI, HIPAA, and ITAR), protection of intellectual property (IP), or simply based on business requirements including NDAs. Those levels/categories are typically based on risk or impact to the business, but ultimately it defines groups of documents that you expect to be handled in similar ways from a policy perspective. There are a few basic questions organizations may ask to help define classification categories:



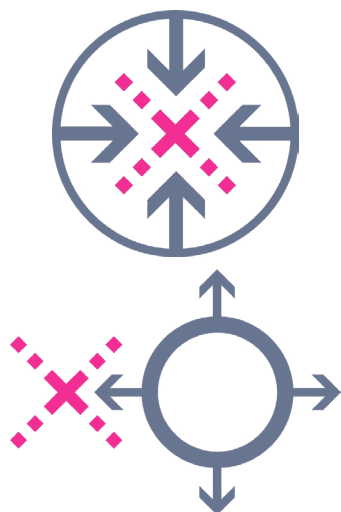
- What types of data do we have?
- What regulated data do we have?
- Who has access to the data?
- Who should have access to the data?
- What is the risk if the data were exposed?
- What controls are in place to protect the data?

Once the classification categories are defined, organizations ideally end up with 3-5 categories; classification definitions may include: *confidential* (i.e. *privileged access reserved to a limited group*), *private* (i.e. *internal sharing only*), and *public data* (i.e. *can be shared with those outside the organization*).

## > WHAT IS THE PURPOSE OF A DATA CLASSIFICATION PROGRAM?

The purpose of data classification is to establish a framework for categorizing data based on its level of sensitivity, value and criticality to the organization as required by the Organization's Information Security Policy. Classification of data will aid in determining security controls for the protection of data.

## > WHY DATA CLASSIFICATION IS FOUNDATIONAL



Data classification provides a documented map of what is important in your organization, this is what ultimately drives enhanced data security. Whether it be internal or external threats, a data classification program can deliver the consistency and guidance needed.

### INSIDER AND OUTSIDER THREAT

While insider threats are a longstanding risk infosec professionals have contended with, the outsider threat has emerged as a source of data exfiltration risk too. Data classification programs can enable better data protection from both the internal and the external threat.

Internal actors can cause both malicious and unintentional data loss; regardless of how the data gets out, it is out. With a data classification program in place the mistyped email address in a message with sensitive data is flagged before leaving your organization. Classified files that are intentionally being leaked get the attention of security solutions, such as Data Loss Prevention (DLP), prior to an incident.

External actors seek data that can be monetized. Understanding what data within your organization has the greatest value and the greatest risk for theft is where classification comes in to play. Through this program you evaluate your data then prioritize protection of it based on their value; information that is most critical can be given the greatest monitoring and controls.

### CONSISTENCY AND GUIDANCE

Classification provides basis and consistency for your data protection rules with pre-defined, minimal, and simple categories for the organization's data. Consistency avoids confusion about what is and isn't sensitive data across the organization, each department is using the same rules and language. Without this framework all information is by default valued equally; there is no distinction between a public facing document and restricted data.

In the event of an incident where data is exposed, knowing the full extent of what has been exposed guides post-breach planning. Correctly stating that "no sensitive data was exposed" with the documentation to prove this can minimize the repercussions of a breach. With a classification program in place your organization knows what information was exposed, and what wasn't. This knowledge guides incident response teams, allowing them to focus on only what was compromised reducing time, complexity, and cost.

## > CLASSIFICATION CHECKLIST

The most successful data classification programs are those thorough enough to deliver business insights but do not overburden the organization. Five elements are instrumental in programmatic and successful data classification: Classification Policy, Scope, Discovery, Solutions, Analysis and Audit

1. **Classification Policy:** This is your instruction sheet and reference guide to the classification program. In this you outline the ground rules data categories using examples for each, provide a real view of the potential risk if this data were to leak, and document the data protection controls.
2. **Scope:** As with any project, clearly outlining the boundaries ensures you don't take on too much, yet don't focus it so narrowly that there is no appreciable benefit to the organization. Starting with data in active use and in a high value function such as CAD drawings used by manufacturing gets the project underway, shows a quick win, and delivers learnings upon which to base expansion.
3. **Discovery:** With your policy and scope in place you are ready to take action. Data discovery involves



identifying and locating sensitive or regulated data in order to adequately protect it or securely remove it; you have to find it in order to protect it.

4. **Solutions:** These range from simple procedural methods, to internally developed and built solutions, all the way up to enterprise class, 3<sup>rd</sup> party software packages. Each has a tradeoff, and your organizations must decide upon the benefits vs costs to determine what to deploy. DLP solutions are commonly deployed in conjunction with, or as a fast follower to, a classification program as a way to monitor and enforce data movement.
5. **Analysis & Audit:** The final item is key for the ongoing, and necessary, modifications to the program over time as your business evolves or the external market evolves.

## > DATA CLASSIFICATION CONSIDERATIONS

Information security professionals and attorneys have defined several types of restricted data based on state and federal regulatory requirements. They're often defined as follows:

- Authentication Verifier
- Covered Financial Information
- Electronic Protected Health Information ("EPHI")
- Export Controlled Materials
- Federal Tax Information ("FTI")
- Payment Card Information
- Personally Identifiable Education Records
- Personally Identifiable Information
- Protected Health Information ("PHI")
- Controlled Technical Information ("CTI")


### DETERMINING THE CORRECT CLASSIFICATION LEVEL

In some situations, the appropriate classification may be obvious, such as when federal laws or compliance regulations require the organization to protect data (e.g. personally identifiable information and credit card data). If the appropriate classification is not inherently obvious, consider each security objective using the following table as a guide (the table is an excerpt from *Federal Information Processing Standards ("FIPS") publication 199* published by the *National Institute of Standards and Technology*):

SECURITY OBJECTIVE	POTENTIAL IMPACT		
	LOW	MODERATE	HIGH
CONFIDENTIALITY	Limited Adverse Effect	Serious Adverse Effect	Severe or catastrophic adverse effect
INTEGRITY	Limited Adverse Effect	Serious Adverse Effect	Severe or catastrophic adverse effect
AVAILABILITY	Limited Adverse Effect	Serious Adverse Effect	Severe or catastrophic adverse effect

ESTABLISHING DATA HANDLING REQUIREMENTS AND CONTROLS

For each classification, data handling requirements should be defined to appropriately safeguard the information. It’s important to understand that overall sensitivity of organizational data encompasses not only its confidentiality but also the need for integrity and availability. The following table provides guidance on the required safeguards for protecting data and data collections based on their classification and/or the federal/state laws or regulations:

CONTROL CATEGORY	PUBLIC DATA	PRIVATE DATA	RESTRICTED DATA
ACCESS CONTROLS			
COPYING/PRINTING (PAPER AND ELECTRONIC)			
NETWORK SECURITY			
SYSTEM SECURITY			
REMOTE ACCESS TO SYSTEMS HOSTING THE DATA	<div>No, or very limited controls</div>	<div>Strictest controls</div>	
DATA STORAGE			
TRANSMISSION			

CLASSIFICATION SUCCESS FACTORS

**Discovery:** In today’s era of remote workers, business is frequently conducted in the cloud and file sharing and storage are the norm. To perform an effective data classification knowing where your information lives, and how that changes, is necessary.

**Simplify:** Often organizations look at their data types and feel they are unique and that there is no way to rely on only 3 classification categories. Leading Security Consultancy, PricewaterhouseCoopers, recommends three sensitivity levels, these three provide the most efficient and effective starting point:

PUBLIC DATA	PRIVATE DATA	RESTRICTED DATA
Data should be classified as Public when the unauthorized disclosure, alteration or destruction of that data would results in little or no risk to the organization and its affiliates.	Data should be classified as Private when the unauthorized disclosure, alteration or destruction of that data would results in a moderate level of risk to the organization and its affiliates.	Data should be classified as Restricted when the unauthorized disclosure, alteration or destruction of that data could cause a significant level of risk to the organization or its affiliates.

Over time, if these prove insufficient more granular classification can be introduced to users familiar with the process. More categories do not always lead to better results; too many can lead to frustration, inconsistent application of categories, and ultimately a failed project. Global organizations where regions have their own official and unofficial language can be particularly problematic.

**Time to Value:** Legacy data classification can be a daunting task; rather than tackle it all, target data in active use and classify going forward. If an analysis shows volumes of data are untouched for years, classification efforts would be better spent on either secure archiving or data destruction. However, when required, even large pools of data can be effectively classified by using automated classification approaches. This legacy data may still contain sensitive items, thus early successes should be used to gain support for broader analysis.

**Evolve with the Business:** It is also important to periodically reevaluate the classification of data to ensure it is still appropriate based on changes to legal and contractual obligations as well as changes in the use of the data or its value to the organization. This evaluation should be conducted by the appropriate data owners annually unless otherwise determined. If the classification of a certain data set has changed, an analysis of security controls should be performed to determine whether existing controls are consistent with the new classification. If gaps are found in existing security controls, they should be corrected in a timely manner, commensurate with the level of risk presented by the gaps.

**Communicate:** To ensure success of any corporate wide initiative communication is paramount, this needs to happen during the plenary process, the roll-out, and as an ongoing program.

- While planning the data classification program listen to multiple stakeholders.
- Education before and during roll-out helps users know what is expected and program value.
- Periodic reminders drive awareness and education, reinforcing and encouraging desired actions.

As part of classification process, an organization may choose to deploy automated tools and/or employ manual processes to achieve data classification goals. Regardless of how the data classification is added to the document, leveraging either approach drives more effective data protection.

- Automated tools help gather content-specific and contextual information about the data and provide automatic data classification based on that information.
- Manual or user classification can be used to enhance a classification framework and/or create a bridge between technical controls and data/business owners.

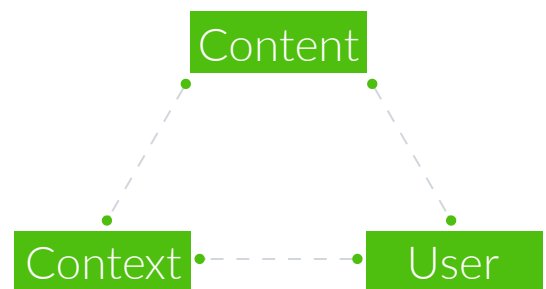
## > DATA CLASSIFICATION METHODS

Data classification can be handled via several methods, each has benefits to the organization but there are considerations that drive the proper method. Content-based, context-based, and user driven are three of the more common techniques today, a discussion of each follows.

**Content-Based Classification:** Classifying via content requires analysis and interpretation of the data to decide where it falls into the predetermined categories.

Within content based classification are multiple techniques, each can be used individually or combined for a hybrid approach:

1. **Database Fingerprinting:** Relies on accessing a database such as client or medical records and cataloging the information based on a policy to match only the sensitive data you must protect. A hash of this information is created and serves as the basis for a policy and any automated actions when a match is detected.
2. **Regular Expression:** RegEx for short, is a sequence of structured text designed to define a search pattern. These can be used for structured data such as credit card numbers, social security numbers, or even patient numbers.
3. **File Fingerprinting:** Similar to database fingerprinting this technique looks at specific files rather than entire databases, then creates a hash of that file.
4. **Partial Matching:** Where database fingerprinting and file fingerprinting look at the database or document as a whole, partial matching looks at pieces of the document and creates multiple hashes, then uses these as the basis for a potential match to policy.



**Machine Learning:** This method relies on training the DLP content inspection engine with a set of documents and then any document to be inspected is given a statistical probability that it is the same type of document in the initial training set.

**Context-Based Classification:** Context-based classification operates wholly independent of the data contained in the documents, file, or email; a context-based classification strategy looks at characteristics as a proxy to determine the level of sensitivity. Context-based approaches can look at application used, storage location, and who created it (both by role using an Active Directory integration, or simply a user list) among other options to make the classification. For example an automotive company may dictate that any CAD files stored on the “Engineering” server is by default sensitive and apply rules to those files.

Contextual classification operates based on several parameters, such as:

- |                                                                 |                                                                         |
|-----------------------------------------------------------------|-------------------------------------------------------------------------|
| Identity of the user who created the data                       | • Drive type: Bus/IO, ATAPI, SCSI, SATA, IDE, ATA, Bluetooth, IrDA, USB |
| Network source & destination                                    |                                                                         |
| File type, name, extension and path                             | • Application: Image Name, Parent process and version, etc.             |
| Operation: read, write, open, save, copy, move, recycle, delete | • Email: To, From, cc, Body, Attachments                                |
| Buffer: Copy, paste, print screen                               | • Time of use                                                           |

**User-Based Classification:** User-based classification is on one hand the simplest, but also the most complex for the same reason – it is user driven and dependent. At the creation of a document, or upon editing, the user designates which of the pre-defined categories it belongs to. Because it is a manual process the users must remember to perform the task, then consistently apply a policy. As they are the closest to the data, user based decisions can deliver superior accuracy over automated approaches.

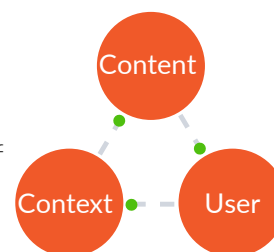
## CHOOSING THE RIGHT METHOD

The decision around which classification method to choose is more often a prioritization than a choice of only one method.

Content-based analysis is implemented when detailed granularity into the file itself is needed.

When this detailed inspection is unavailable or impractical context based classification relies on analyzing the “container” around the data, no file analysis is required.

Finally, if users have the ability to make the classification decision, enable them. The users want to do the right thing for their organization in most cases, they also know the data intimately.

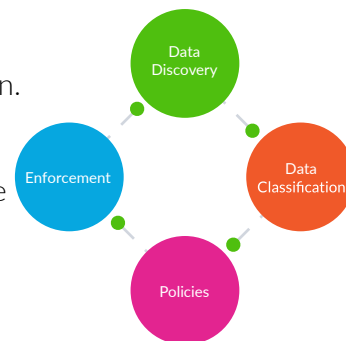


Each of these on its own provides insight, combining them provides further detail; organizations must be able to process this greater volume of information about the data to drive insights and actions.

## > DIGITAL GUARDIAN FOR THE MOST COMPLETE DATA CLASSIFICATION AND DATA PROTECTION SOLUTION

Digital Guardian relies on a flexible approach to data classification to enable organizations to align their needs with our solutions. For that, Digital Guardian delivers upon the 3 techniques for data classification outlined previously.

Digital Guardian's data classification is integrated into our data protection solution. This connection, and the built in automation, delivers a more accurate DLP program to limit false positives and false negatives. By combining data discovery, classification, policies, and enforcement Digital Guardian provides comprehensive data protection.



## > ABOUT DIGITAL GUARDIAN

The company's flagship product, the Digital Guardian Data Protection Platform, is purpose built to stop data theft and performs across the network, on endpoints, in the cloud and in data storage to make it easier to see and stop all threats to your sensitive data. This platform is recognized as a leader in the 2016 Gartner Magic Quadrant for Enterprise Data Loss Prevention, and is #1 for intellectual property protection per Gartner Critical Capabilities for Enterprise Data Loss Prevention.

For more than 10 years, Digital Guardian has enabled data-rich organizations to protect their most valuable assets with an on premise deployment or an outsourced managed security program (MSP). The company operates in more than 60 countries while servicing one out of five of the Fortune 500, seven of the top ten global patent holders and, seven of the ten largest global automobile manufacturers.

Digital Guardian delivers anytime, anywhere data security.



**ICDIGITAL**

[advisor@icdigital.com](mailto:advisor@icdigital.com)  
[www.icdigital.com/digitalguardian](http://www.icdigital.com/digitalguardian)