

Creating Classifiers from Scratch to Identify Sensitive Data with Microsoft Purview



Myron Helgering

<< Security >>

#ScottishSummit26



Sponsors (1)

PARTNERS

Platinum



Gold



Sponsors (2)

PARTNERS

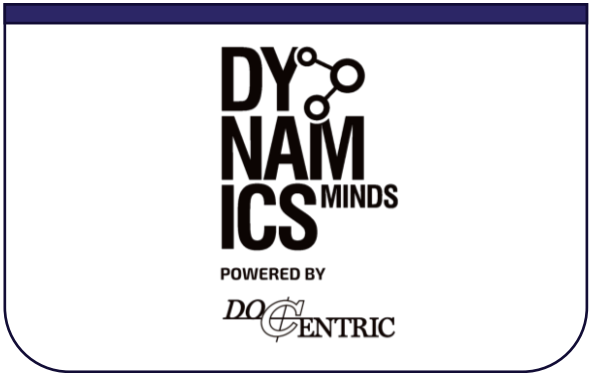
Silver



Media



Community



Charity Partner





Myron Helgering

Tech Lead Security // Pink Elephant

Microsoft Security MVP

Blogger and speaker

Fantasy & sci-fi geek

Focus

Microsoft Purview

Microsoft Defender

Microsoft Entra

Awards

Microsoft MVP

Connect

 [in/myronhelgering](https://www.linkedin.com/in/myronhelgering)

 myronhelgering.com

Agenda

- ◆ Introduction and scenario
- ◆ Classifiers
 - ◆ Sensitive Information Types
 - ◆ Fingerprint SITs
 - ◆ Trainable Classifiers
 - ◆ Exact Data Match Classifiers
- ◆ Takeaways & Q&A

Meet Minerva McGonagall



- Professor at Hogwarts School
- Chief Information Security Officer (CISO)
- Her goal is to **identify** and **protect** sensitive school data



What does sensitive data look like?

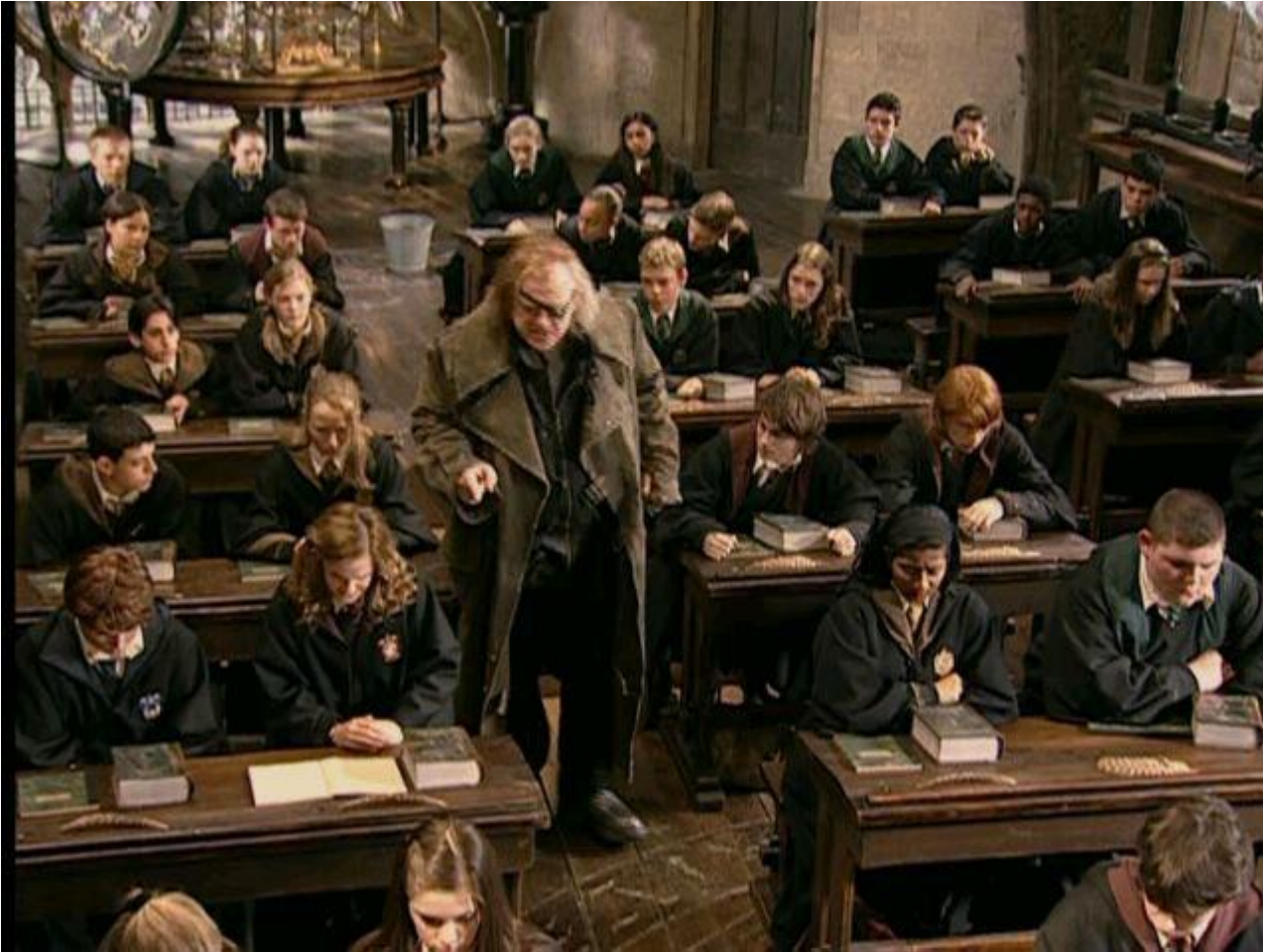


Sensitive information for healthcare

- ◆ Patient number
- ◆ Insurance number
- ◆ Medical record
- ◆ Treatment plan
- ◆ Lab rapport
- ◆ Medical terms



Sensitive information for education



- ◆ Student number
- ◆ Student PII
- ◆ Student grades
- ◆ Student database
- ◆ Education records
- ◆ Class schedules

How to start, step by step

- ◆ Talk to the security stakeholders and data owners
- ◆ Determine your top 5 most sensitive information types
- ◆ Ask for examples, formats and templates to work with
- ◆ Start creating, testing and finetuning classifiers

Hogwarts Sensitive Information Types (Top 5)

1. Passport numbers
2. StudentID
3. Student Information Forms
4. Potion Recipes
5. Student Record Database



Built-in Sensitive Information Types

327 Built-in Sensitive Information Types;

- ◆ Credit card
- ◆ Passport number
- ◆ National Insurance number (UK)
- ◆ SSN number (USA)
- ◆ Physical address
- ◆ Ip-address
- ◆ Driver's license number
- ◆ Password

Classifiers		
Trainable classifiers	<u>Sensitive info types</u>	EDM classifiers
The sensitive info types here are available to use in your security and compliance policies. These include a large collection of types we		
+ Create sensitive info type + Create Fingerprint based SIT 🔄 Refresh		
Name ↑	Type	Publisher
☐ ABA Routing Number	🔗 Entity	Microsoft Corporation
☐ ASP.NET Machine Key	Credential	Microsoft Corporation
☐ All Credential Types	BundledCredential	Microsoft Corporation
☐ All Full Names	BundledEntity	Microsoft Corporation
☐ All Medical Terms And Conditions	BundledEntity	Microsoft Corporation
☐ All Physical Addresses	BundledEntity	Microsoft Corporation
☐ Amazon S3 Client Secret Access Key	Credential	Microsoft Corporation
☐ Argentina National Identity (DNI) Number	🔗 Entity	Microsoft Corporation
☐ Argentina Unique Tax Identification Key (CUIT/CUIL)	🔗 Entity	Microsoft Corporation
☐ Australia Bank Account Number	🔗 Entity	Microsoft Corporation
☐ Australia Driver's License Number	🔗 Entity	Microsoft Corporation
☐ Australia Medical Account Number	🔗 Entity	Microsoft Corporation

Demo: Sensitive Information Types & Data explorer



I want to know how many
and where (student)
passport numbers are
being stored

Overview | Microsoft Purview

https://purview.microsoft.com/informationprotection/purviewmipoverview?tid=97f8ab89-ccf3-4f29-a961-c4e40d453141

Microsoft Purview

Search

Copilot

Home

Solutions

Agents

Learn

Settings

Information Protection

Information Protection

Overview

Reports

Recommendations

Sensitivity labels

Policies

Classifiers

Explorers

Data explorer

Content explorer (classic)

Activity explorer

Diagnostics

Related solutions

Data Lifecycle Management

Data Loss Prevention

Information Protection

Discover, label, and protect sensitive and business-critical info across your multicloud data estate.

Setup tasks

1

Register and scan multicloud data sources

Bring in data from platforms like Azure and AWS so we can start detecting sensitive data.

Highly recommended

30 minutes

2

Scan registered data sources

Scan your registered data sources so we can start detecting sensitive data.

Required

Varies

3

Get to know the new Information Protection

Educate yourself on the new Information Protection capabilities in Purview.

Optional

7 minutes

Recommendations

No recommendations right now

We're constantly gathering insights into your current configuration, so look out for new recommendations that'll help enhance and improve your Information Protection deployment.

Learn more about recommendations

Custom Sensitive Information Types

StudentID number

- ◆ Number format is 1234-5678
- ◆ Wordlist contains StudentID

Example

- ◆ Doc 1: Harry has ID number 389-9468
- ◆ Doc 2: Harry has StudentID number 0397-3845



Learn the Basics
Regular Expressions
<https://regexone.com/>

Lesson 1½: The 123s

Characters include normal letters, but digits as well. In fact, numbers 0-9 are also just characters and if you look at an [ASCII table](#), they are listed sequentially.

Over the various lessons, you will be introduced to a number of special metacharacters used in regular expressions that can be used to match a specific type of character. In this case, the character `\d` can be used in place of **any digit from 0 to 9**. The preceding slash distinguishes it from the simple `d` character and indicates that it is a metacharacter.

Below are a few more lines of text containing digits. Try writing a pattern that matches all the digits in the strings below, and notice how your pattern matches **anywhere within the string**, not just starting at the first character. We will learn how to control this in a later lesson.

Exercise 1½: Matching Digits

Task	Text	
Match	abc123xyz	✓
Match	define "123"	✓
Match	var g = 123;	✓

[Continue >](#)

Solve the above task to continue on to the next problem, or read the [Solution](#).

123...	Digits
\d	Any Digit
\D	Any Non-digit character
.	Any Character
\.	Period
[abc]	Only a, b, or c
[^abc]	Not a, b, nor c
[a-z]	Characters a to z
[0-9]	Numbers 0 to 9
\w	Any Alphanumeric character
\W	Any Non-alphanumeric character
{m}	m Repetitions
{m,n}	m to n Repetitions
*	Zero or more repetitions
+	One or more repetitions
?	Optional character
\s	Any Whitespace
\S	Any Non-whitespace character
^...\$	Starts and ends
(...)	Capture Group
(a(bc))	Capture Sub-group
(.*)	Capture all
(abc def)	Matches abc or def



Expression

```
/^(?:[A-Z][0-9]{8}|[A-Z]{2}[0-9]{7}|[A-Z]{2}[0-9]{6}[A-Z])$/g
```

Text

Tests

C12345678

Tools

Replace

List

Details

Explain



Roll-over elements below to highlight in the Expression above. Click to open in Reference.

^ Beginning. Matches the beginning of the string, or the beginning of a line if the multiline flag (m) is enabled.

(?: Non-capturing group. Groups multiple tokens together without creating a capture group.

[Character set. Match any character in the set.

A-Z Range. Matches a character in the range "A" to "Z" (char code 65 to 90). Case sensitive.

]

Test and Create
Regular Expressions
<https://regexone.com/>

Demo: Custom Sensitive Information Types

A woman with short blonde hair, wearing a black high-collared dress, is holding a thin wooden wand. Above her head is a large, white, cloud-like thought bubble with a dark red outline. The background is dark blue with faint, stylized white geometric shapes.

I want to be able to
classify data that
contains a StudentID

The sensitive info types here are available to use in your security and compliance policies. These include a large collection of types we provide, spanning regions around the globe, as well as any custom types you have created.

333 items

Supported platforms: **Any**

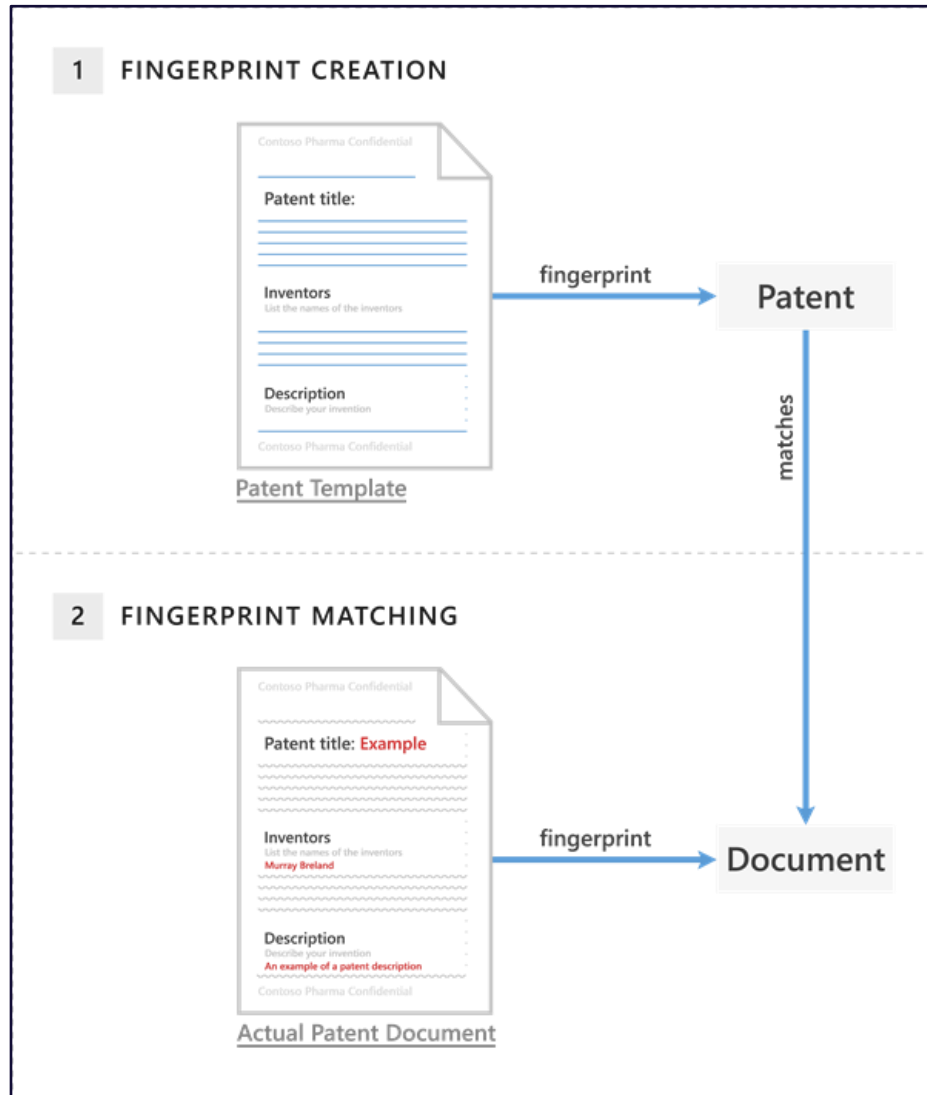
Type: Any

Publisher: Any

 **Add filter**

	Name ↑ ↓		Supported platforms ↓	Type ↓	Publisher ↓
☐	ABA Routing Number		All	Entity	Microsoft Corpor...
☐	All Credential Types		Microsoft 365	BundledCred...	Microsoft Corpor...
☐	All Full Names		Microsoft 365	BundledEntity	Microsoft Corpor...
☐	All Medical Terms And Conditions		Microsoft 365	BundledEntity	Microsoft Corpor...
☐	All Physical Addresses		Microsoft 365	BundledEntity	Microsoft Corpor...
☐	Amazon S3 Client Secret Access Key		Microsoft 365	Credential	Microsoft Corpor...
☐	Argentina National Identity (DNI) Number		All	Entity	Microsoft Corpor...
☐	Argentina Unique Tax Identification Key (CUIT/CUIL)		All	Entity	Microsoft Corpor...

Fingerprint Sensitive Information Types



- ◆ Employee information forms
- ◆ Customer information forms
- ◆ Medical records
- ◆ Contracts
- ◆ Invoices
- ◆ Non-Disclosure Agreements (NDA)
- ◆ Or any other custom template file

Demo: Fingerprint Sensitive Information Types

A woman with short blonde hair, wearing a black high-collared dress, is pointing a thin wooden wand towards a large, white, cloud-shaped thought bubble. The background is dark blue with faint, stylized white geometric shapes. Three small white circles lead from the woman's head to the thought bubble.

I want to be able to
classify Student
Information Forms

Home

Solutions

Agents

Learn

Settings

Information Protection

Information Protection

Overview

Reports

Recommendations

Sensitivity labels

Policies

Classifiers

Trainable classifiers

Sensitive info types

EDM classifiers

On-demand classification

Collection policies

Preview

Explorers

Diagnostics

Related solutions

Microsoft Purview

Search

Copilot

Sensitive info types

The sensitive info types here are available to use in your security and compliance policies. These include a large collection of types we provide, spanning regions around the globe, as well as any custom types you have created.

+ Create sensitive info type

+ Create Fingerprint based SIT

Refresh

333 items

Search

Filters:

Supported platforms: Any

Type: Any

Publisher: Any

Add filter

	Name	Supported platforms	Type	Publisher
☐	ABA Routing Number	All	Entity	Microsoft Corpor...
☐	All Credential Types	Microsoft 365	BundledCred...	Microsoft Corpor...
☐	All Full Names	Microsoft 365	BundledEntity	Microsoft Corpor...
☐	All Medical Terms And Conditions	Microsoft 365	BundledEntity	Microsoft Corpor...
☐	All Physical Addresses	Microsoft 365	BundledEntity	Microsoft Corpor...
☐	Amazon S3 Client Secret Access Key	Microsoft 365	Credential	Microsoft Corpor...
☐	Argentina National Identity (DNI) Number	All	Entity	Microsoft Corpor...
☐	Argentina Unique Tax Identification Key (CUIT/CUIL)	All	Entity	Microsoft Corpor...

Built-in Trainable Classifiers

- ◆ Trainable classifiers identify a “type” of content
- ◆ 134 pre-trained classifiers are ready-to-use
- ◆ 14 different languages



Custom Trainable Classifiers

- ◆ 50-500 positive and negative samples (English only)
- ◆ Microsoft trains the classifier
- ◆ Review false negatives
- ◆ Retrain for higher accuracy
- ◆ Publish, test and use the classifier

Demo: Custom Trainable Classifiers



I want to be able to
classify potion recipes

Trainable classifiers | Microsoft Pu xPotion Recipe Documents (positiv xPotion Recipe Documents (negati x+

https://purview.microsoft.com/informationprotection/dataclassification/trainableclassifiers?tid=97f8ab89-ccf3-4f29-a961-c4e40d453141

Microsoft Purview

Search

Copilot

1

?

Home

Solutions

Agents

Communi... Compliance

eDiscovery

Data Loss Prevention

Information Protection

Compliance Manager

...

Information Protection

Overview

Reports

Recommendations

Sensitivity labels

Policies

Classifiers

Trainable classifiers

Sensitive info types

EDM classifiers

On-demand classification

Collection policies

Preview

Explorers

Diagnostics

Related solutions

Trainable classifiers

Use built-in or custom classifiers to identify specific categories of content based on existing items in your organization. Once created, classifiers can be used in several compliance solutions to detect related content and classify it, protect it, retain it, and more. [Learn more](#)

+ Create trainable classifier

Refresh

135 items

Group

Filters:

Language: Any

Type: Any

Name: Any

Status: Any

Filters

	Name	Accuracy	Status	Type	Language	Created by	
	Published (135)						
☐	Actuary reports		-	Ready to use	Built-In	English	Microsoft
☐	Agreements		-	Ready to use	Built-In	English	Microsoft
☐	Asset Management		-	Ready to use	Built-In	English	Microsoft
☐	Bank statement		-	Ready to use	Built-In	English	Microsoft
☐	Bank statement		-	Ready to use	Built-In	German	Microsoft
☐	Bank statement		-	Ready to use	Built-In	Spanish	Microsoft

Waiting..... (maximum of 48 hours)



Trainable classifiers | Microsoft Pu x

Trainable classifiers | Microsoft Pu x

+

←

↺

https://purview.microsoft.com/informationprotection/dataclassification/trainableclassifiers?tid=97f8ab89-ccf3-4f29-a961-c4e40d453141

🔍

☆

☆

👤

...

🌐

Microsoft Purview

🔍

Search

🚀

Copilot

🔔

🔔1

🗣️

⚙️

?

👤

🏠

Home

🧩

Solutions

👤

Agents

📖

Learn

⚙️

Settings

🔒

Information Protection

☰

Information Protection

Overview

Reports

Recommendations

Sensitivity labels

Policies

Classifiers

Trainable classifiers

Sensitive info types

EDM classifiers

On-demand classification

Collection policies

Preview

Explorers

Diagnostics

Trainable classifiers

Use built-in or custom classifiers to identify specific categories of content based on existing items in your organization. Once created, classifiers can be used in several compliance solutions to detect related content and classify it, protect it, retain it, and more. [Learn more](#)

+ Create trainable classifier

🔄 Refresh

136 items

📑 Group

Filters:

Language: Any

Type: Any

Name: Any

Status: Any

🔍 Filters

	Name	Accuracy	Status	Type	Language	Created by
▼	Published (135)					
☐	Actuary reports	🔗 -	Ready to use	Built-In	English	Microsoft
☐	Agreements	🔗 -	Ready to use	Built-In	English	Microsoft
☐	Asset Management	🔗 -	Ready to use	Built-In	English	Microsoft
☐	Bank statement	🔗 -	Ready to use	Built-In	English	Microsoft
☐	Bank statement	🔗 -	Ready to use	Built-In	German	Microsoft
☐	Bank statement	🔗 -	Ready to use	Built-In	Spanish	Microsoft



Exact Data Match Classifiers

#ScottishSummit26



StudentID	First name	Last name	Date of Birth
0019-3947	Ron	Weasley	01-03-1980
0010-2394	Harry	Potter	31-06-1980
0015-5934	Hermione	Granger	19-09-1979



To

Cc

Subject

Hello Albus,

I found one of our students misbehaving and brought him in for detention.
Hereby I share with you his personal information.

StudentID: 0019-3947

Name: Ron Weasley

Date of Birth: 01-03-1980

Greetings,

Severus Snape

Demo: Exact Data Match Classifiers

A woman with short blonde hair, wearing a black high-collared dress, is shown from the waist up. She is holding a thin wooden wand in her right hand. Above her head are three small white circles of increasing size, leading to a large white thought bubble with a dark red outline. The background is dark blue with faint, large, stylized letters 'S' and 'U' in a lighter blue color.

I want to be able to
classify data that exactly
matches the student
record database

EDM classifiers | Microsoft Purview

https://purview.microsoft.com/informationprotection/dataclassification/exactdatamatch?tid=97f8ab89-ccf3-4f29-a961-c4e40d453141

Microsoft Purview

Search

Copilot

1

?

Home

Solutions

Agents

Learn

Settings

Information Protection

Information Protection

Overview

Reports

Recommendations

Sensitivity labels

Policies

Classifiers

Trainable classifiers

Sensitive info types

EDM classifiers

On-demand classification

Collection policies

Preview

Explorers

Diagnostics

EDM classifiers

New EDM experience ☒ On

Learn the end-to-end workflow

Industry-specific sample files

[Why are there two experiences?](#)

Exact data match (EDM) classifiers use exact values from your org's data to detect matches instead of generic patterns. They can then be included in several compliance solutions to classify and protect sensitive data. [Learn more about EDM](#)

+ Create EDM classifier

0 items

Search

Name ↑	Created by	Status
--------	------------	--------

Create your first exact data match

Preparation for next step

- ◆ Create the “EDM_DataUploaders” security group
- ◆ Install the Exact Data Match (EDM) agent
- ◆ Separate the hashing and uploading process
- ◆ Automate the process of uploading live data

Create exact data match sensitive information type in the New Experience workflow



EDM classifiers | Microsoft Purview

https://purview.microsoft.com/informationprotection/dataclassification/exactdatamatch?tid=97f8ab89-ccf3-4f29-a961-c4e40d453141

Microsoft Purview

Search

Copilot

Home

Solutions

Agents

Learn

Settings

Information Protection

Information Protection

Overview

Reports

Recommendations

Sensitivity labels

Policies

Classifiers

Trainable classifiers

Sensitive info types

EDM classifiers

On-demand classification

Collection policies

Preview

Explorers

Diagnostics

EDM classifiers

New EDM experience

On

Learn the end-to-end workflow

Industry-specific sample files

Why are there two experiences?

Exact data match (EDM) classifiers use exact values from your org's data to detect matches instead of generic patterns. They can then be included in several compliance solutions to classify and protect sensitive data. [Learn more about EDM](#)

+ Create EDM classifier

1 item

Search

Name ↑	Created by	Status
☐ Student Database Information	Helgering	Source file not uploaded yet How to upload

Waiting..... (maximum of 1 hour)



EDM classifiers | Microsoft Purview

https://purview.microsoft.com/informationprotection/dataclassification/exactdatamatch?tid=97f8ab89-ccf3-4f29-a961-c4e40d453141

Microsoft Purview

Search

Copilot

1

?

Home

Solutions

Agents

Learn

Settings

Information Protection

Information Protection

Overview

Reports

Recommendations

Sensitivity labels

Policies

Classifiers

Trainable classifiers

Sensitive info types

EDM classifiers

On-demand classification

Collection policies

Preview

Explorers

Diagnostics

EDM classifiers

New EDM experience

On

Learn the end-to-end workflow

Industry-specific sample files

Why are there two experiences?

Exact data match (EDM) classifiers use exact values from your org's data to detect matches instead of generic patterns. They can then be included in several compliance solutions to classify and protect sensitive data. [Learn more about EDM](#)

+ Create EDM classifier

1 of 1 selected

Search

Name ↑	Created by	Status
Student Database Information	Helgering	Index complete

Goal achieved

- ◆ Identify where passport numbers were stored with the **Data Explorer**
- ◆ Classify data that contains a StudentID with a **Custom SIT**
- ◆ Classify student information forms with a **Fingerprint SIT**
- ◆ Classify potion recipes with a **Trainable Classifier**
- ◆ Classify data containing information that exactly matches the student record database with an **Exact Data Match Classifier**



License and permissions

- ◆ Access Data Explorer
- ◆ Access to Endpoint DLP
- ◆ Auto-classify data based on classifiers → ◆ M365 E5 or Purview suite
- ◆ Use advanced classifiers (e.g. trainable and EDM classifiers)
- ◆ Create classifiers and policies → ◆ Compliance Administrator role
- ◆ View content in Data Explorer → ◆ Content Explorer Content viewer

Takeaways

- ◆ Involve the business before creating classifiers.
- ◆ Let classifiers be part of your classification process, but don't fully rely on them; involve your users.
- ◆ Use a classifier that fits the type of data.
- ◆ The task of managing classifiers is never done.



Classifier Simulation Mode

#ScottishSummit26

Microsoft Purview: Purview Information Protection | Classifier Simulation Mode (Health Monitoring)

Microsoft Purview

Microsoft Purview introduces the Classifier Simulation Mode, the first phase of the broader Classifier Health Monitoring Platform. This capability enables you to test, analyze, and optimize custom classifiers on production data before publishing them, hence offering a reliable means of verifying the effectiveness of classifiers before committing them to live environments. Custom classifiers are powerful tools but can sometimes include inefficient regex patterns or overly broad logic, leading to Noisy classifier due to high matches, False positives and High scanning latency Simulation Mode mitigates these risks by allowing you to validate classifier quality and performance before publishing, helping ensure healthy, performant, and more precise classification outcomes.

Roadmap ID	Cloud instances(s)	Platform(s)	Release phases(s)
523201	Worldwide (Standard Multi-Tenant)	Web	General Availability

Added to roadmap: 11/05/2025 | Last modified: 07/14/2026

SHARE



■ □ □ IN DEVELOPMENT

PREVIEW AVAILABLE
October 2026

ROLLOUT START
November 2026



<https://techcommunity.microsoft.com/blog/microsoft-purview-blog/meet-the-new-microsoft-purview-classifiers-experience-your-admin-playbook-for-pu/4561324>

Q&A and

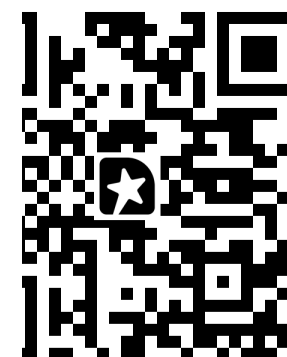
Please submit your questions
All feedback is used to improve
what works and what doesn't



Creating Classifiers from Scratch to Identify Sensitive Data
with Microsoft Purview



Feedback



Classifiers from Scratch to Identify Sensitive Data
with Microsoft Purview