

Adventures in Data Labelling

From Concepts to Implementation in Microsoft Purview

Don Mallory
BSidesTO - Oct 19, 2024

whoami

- 30+ years in IT, mostly for critical infrastructure
- Healthcare security professional
- CISSP, GSEC, GCED, GCIH, and others
- Volunteer:
 - Healthcare Infosec Group - Moderator
 - C3X - Builder & Mentor (2018-2020)
 - Hak4Kidz Toronto (2019)
 - B&W Photography Lead (since 2007)



Disclaimer

The thoughts and opinions shared throughout this presentation are mine alone and not those of my past, present, or future employers

Agenda

- Overview
- Data Asset Inventory
- Labelling Models
- Implementing Labelling
- M365 and Purview
 - Sensitivity / Retention
 - IAP & Trainable Classifiers
 - Auto Labelling
 - DLP
- Conclusions
- Resources & links



No need to take photos
Wait for the **last** QR code



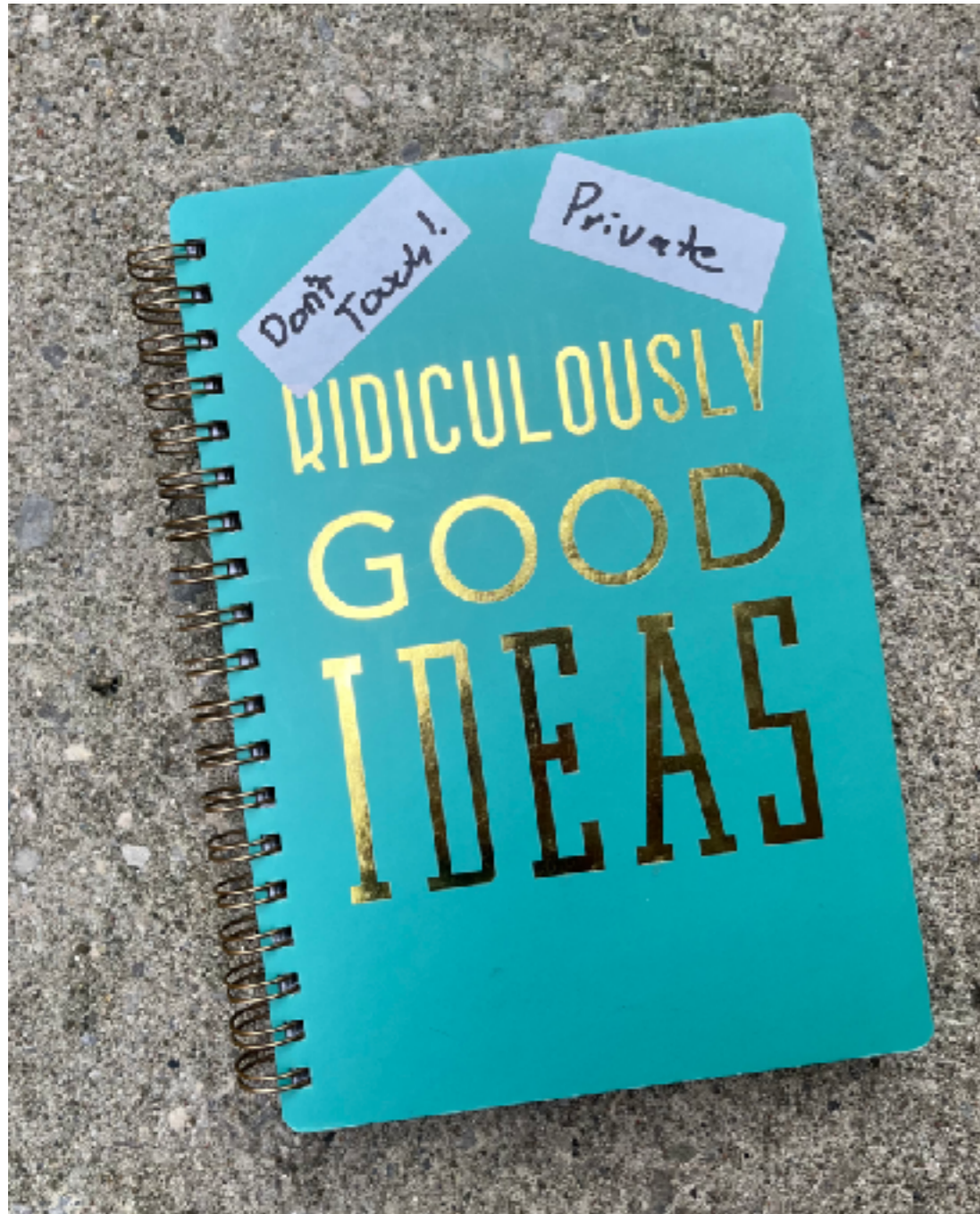
Assumptions

- There is no such thing as one size fits all
- This doesn't have to be perfect
- **We will not cover**
 - Everything
 - There will always be more
 - Embrace the rate of change
 - Pricing - that's between you and your sales rep
 - How long will it take to implement - talk to your stakeholders

What is this not?

- Ransomware protection
- Magical AI Unicorns
- False positive and false negative free
- Free - there is always a cost

What is data labelling?



- Is it important?
- Should I keep it?
- Do I want to share it?
- Who do I want to share it with?
- Why should I keep it?
- How long should I keep it?
- When should I throw it out?

Why label your data?

Questions:

- What is the impact or injury if the data is lost or accessed?
- Do you have any legal, regulatory, or moral obligations?
- How much data do you have of a given type?

Defines:

- Where should you store it?
- How should you store it?
- How should you audit it?
- Retention obligations?
- Disclosure obligations?
- Who do you have to tell when things go wrong?

Outcomes:

- Data inventory aligns to governance frameworks
- Helps with discoverability
- Reduces costs in the event of an incident



Data Asset Inventory

Know your data

- Scan your data with automated tools
- Manual review process

What are you looking for?

- Data type
- Location
- RACI
- Backups
- Retention
- Disposition
- Destruction
- Contractual / Legal / Regulatory obligations
- Controls applied

Why?

- Informs how you will label and classify data
- Allows you to train machine learning models
- Supports eDiscovery, Data Governance, building DLP rules, etc.

Metadata

- Document name
- Document format
- Author
- Date created
- Version
- Size
- Summary
- Content tags
- Retention
- Data sensitivity

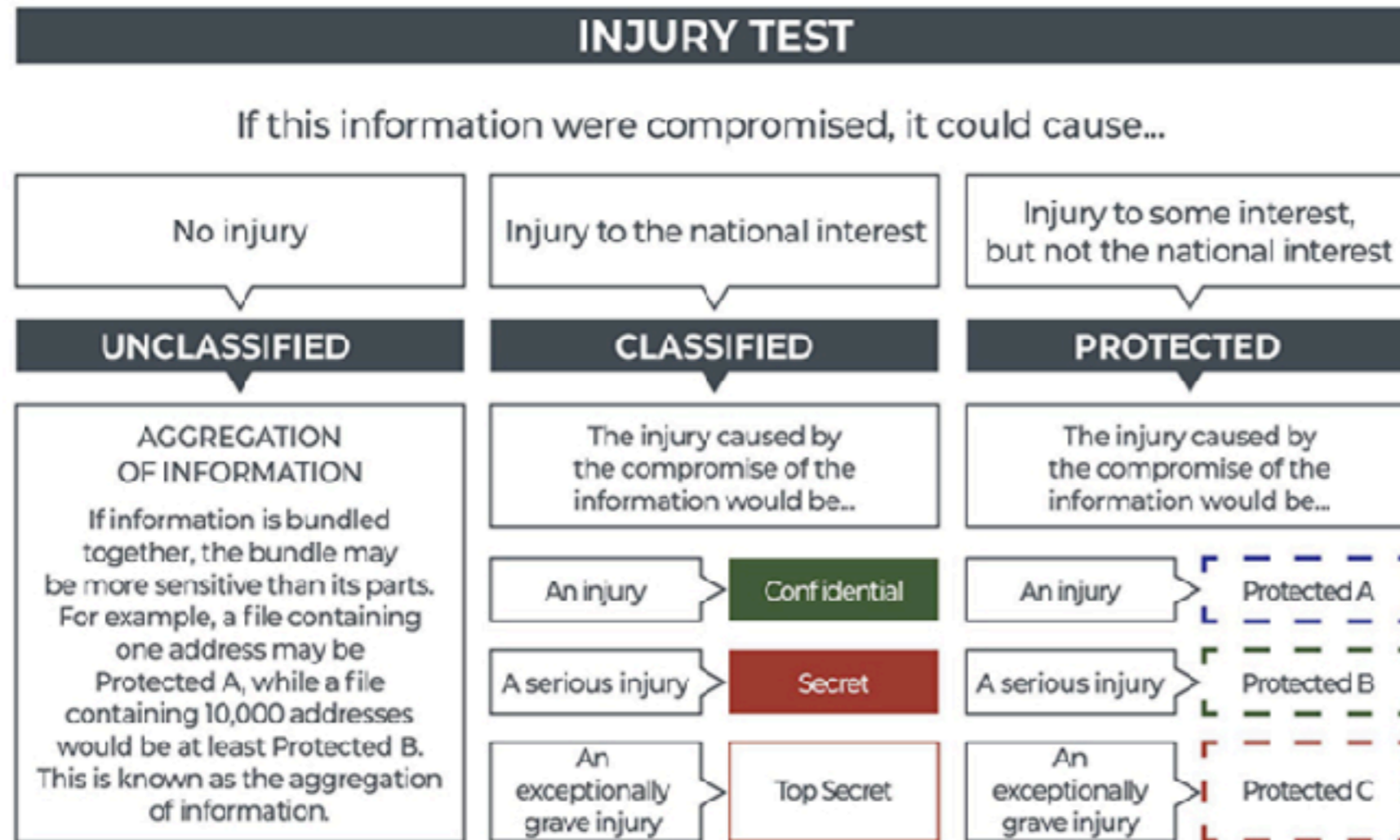
Naming standards - based on:

- Department
- Content focus
- Regulatory / Legal obligations

Data Lifecycle Management

Labelling for Sensitivity - Models

Is your information Classified or Protected? Try this test



Commercial

- Public
- Sensitive
- Proprietary / Confidential / Private

Canadian Gov

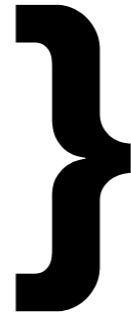
- Unclassified
- Protected A1
- Protected B1/B2
- Protected C1/C2
- Confidential
- Secret
- Top Secret

Healthcare

- Public
- Internal
- Confidential
- PHI
- Restricted

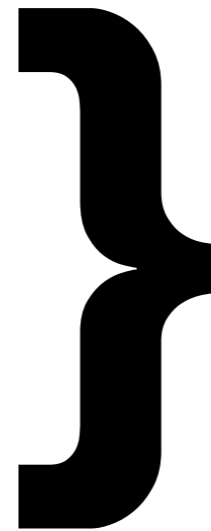
Sensitivity - Alternative Models

- Confidential + PII
- Confidential + PHI
- Confidential + PCI



Defined for regulated data

- Confidential-Financial
- Confidential-HR
- Confidential-Legal
- Confidential-Technical
- Confidential-PII-Identifier
- Confidential-PII-Identifying
- Confidential-PII-Confriming



**Increased granularity
for data subject rights**

Labelling for Retention

- HR Records (3 yrs after employee departure)
- Financial (7 years)
- Corporate Records (10 years)
- Patents (20 years)
- Patient records
 - Under 18 (18+25 years)
 - Over 18 (25 years)
- Engineering drawings (Life of subject+15 years)
- Nuclear safety data
 - 30 years beyond decommissioning
(5+5+5+25+5+25+5+25+10+10+30 = 150 years)



Retention also means disposition, disposal, and destruction.

Implementation - Sponsors & Stakeholders

Executive sponsorship

Stakeholders

- IT
- Legal
- Privacy & FOI
- Records Management
- Data Governance
- Finance & Audit
- HR
- Organizational Development & Training
- Public Relations

Special client groups

- Sales
- Accounting
- Engineering
- Research
- Clinical leadership & informatics
- Patient Experience
- Partners and Peers



Implementing Data Labelling

Initiative Support

- Executive sponsorship
- Engage stakeholders

Policy

- Scoped to include
 - Data in all forms
 - Data in all locations
- Descriptions
- Clear examples
- Impact of loss/disclosure
- Storage and controls
- Transport, and handling
- Disposal, declassification

Appendix A: [the EHR Solution] Information and Asset Classification

Classes		Description	Examples Assets
Confidentiality	Integrity and Availability		
Public	LOW	Information or assets that are used in the normal course of business and that are unlikely to cause harm. Available to the public.	<ul style="list-style-type: none"> • Inform external website • External
Internal		Information or assets that have a low sensitivity outside of [the EHR Solution] and could have low levels of impact on service levels or performance, or result in low levels of financial loss. Available to all agents of [the EHR Solution], and Electronic Service Providers of [the EHR Solution] and HICs with a need to know.	<ul style="list-style-type: none"> • User n • Solution • High-le • planni • High-le • inform • effecti • EHR Sc
Confidential	MEDIUM	Information or assets that have a moderate to high sensitivity within [the EHR Solution] and outside of [the EHR Solution], and could have a moderate impact to service levels or performance, or result in moderate levels of financial loss.	<ul style="list-style-type: none"> • Person • includi • identifi • pay • Inform • disclos • Financ

Implementing Data Labelling

Communications plan

- Continuous...
 - Town halls, meetings, Intranet, emails, FAQ
 - Educational materials
 - Quick reference cards
 - Mandatory eLearning module
 - Reminders delivered through multiple methods

Apply controls

- Implement and test in IT systems
- Start small in monitor mode
- Plan for physical data assets
- Unlabelled data is public / unrestricted



Implementation Challenges

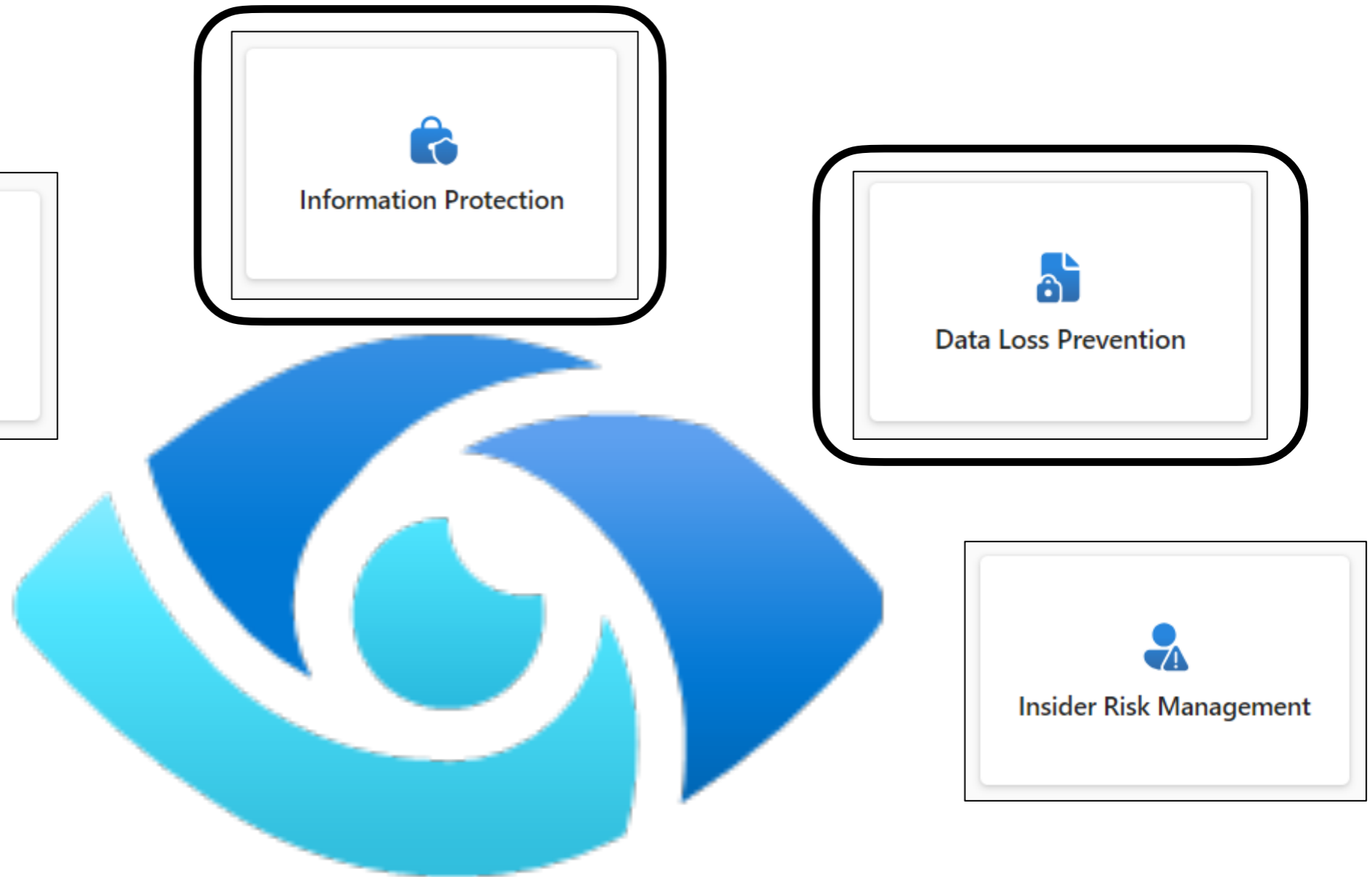
- Did you forget to do an asset inventory?
- Still no executive sponsor?
- Policy stuck in review cycles
- Bad or missing examples
- Education materials are not mandatory
- Nobody read any of it
- Communications plan delayed or missing
- Too many labels - Do not exceed 5
- Tech doesn't work as expected
- You missed an important business process
- Everyone hates watermarking

Implementing the business side was easy...

Is everybody with me so far?



Microsoft Purview



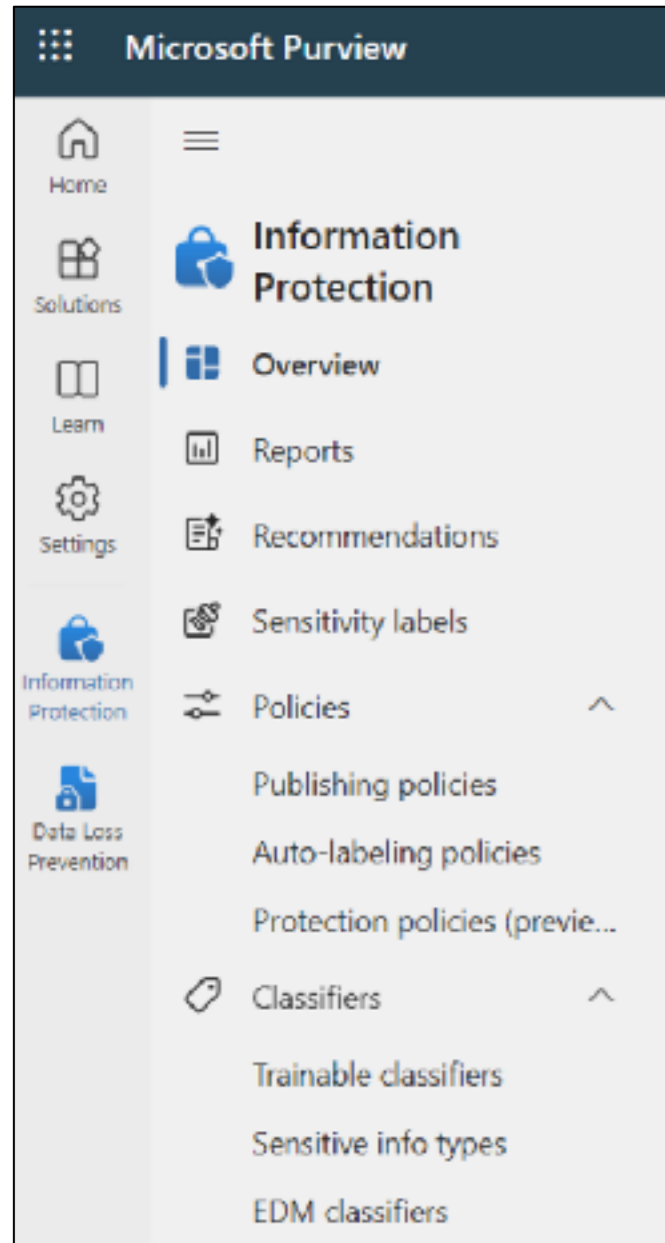
Data Governance
Govern data seamlessly to empower your organization.

- Data Catalog**
Find and create data across your org with this searchable inventory of data assets and metadata.
- Data Lifecycle Management**
Manage your content lifecycle so you can keep what you need and delete what you don't.

Risk & Compliance
Manage critical risks and regulatory requirements.

- Communication Compliance**
Capture, analyze, and manage risk to help reduce communication risks and take steps to minimize harm.
Powered by Copilot
- Compliance Manager**
Get insight into your employees' posture and reduce risks with built-in assessments and recommended improvement actions.
- eDiscovery**
Identify, preserve, and export data in response to legal discovery requests and eDiscovery cases.
Powered by Copilot
- Information Barriers**
Enable two-way communication and collaboration to avoid conflicts of interest and safeguard internal info.
- Records Management**
Automate and simplify the retention schedule for regulatory, legal, and business-critical records.

Sensitivity Labels



Before you start

- Enable for M365 Groups and Sites
- Enable Co-Authoring with Sensitivity Labels before you start
- Teams needs a Premium License



Implementation

- Plan your labels
 - Order matters
 - Scope
 - Content marking
 - Access Controls
 - Protection of Sites & Groups
 - External Sharing
-
- Don't delete a label that's been used. **EVER.**

Sensitivity Labels

Sensitivity labels

Sensitivity labels are used to classify email messages, documents, sites, and more. When a label is applied (automatically content marking, and control user access to specific sites. [Learn more about sensitivity labels](#)

 Create a label  Publish labels  Export  Refresh

<input type="checkbox"/>	Name		Priority	Scope
<input type="checkbox"/>	Public	⋮	0 - lowest	File, Email, Meetings, S
<input type="checkbox"/>	Internal	⋮	1	File, Email, Meetings, S
<input type="checkbox"/>	Confidential	⋮	2	File, Email, Meetings, S
<input type="checkbox"/>	PHI	⋮	3	File, Email, Meetings, S
<input type="checkbox"/>	Restricted	⋮	4	File, Email, Meetings, S

Sensitivity Label Policy

- Publish your labels
- Require labels for everything
- Define a default label - Internal
- Inherit labels from highest labelled attachments
- Users must provide justification to lower a label or remove a label
- Link to your FAQ

Policy settings

Configure settings for the labels included in this policy.

- Users must provide a justification to remove a label or lower its classification**
Users will need to provide a justification before removing a label or replacing it with one that has a lower classification and justification text.
- Require users to apply a label to their emails and documents**
Users will be required to apply labels before they can save documents or send emails (only if these labels are published to Exchange email). [Support and behavior for this setting varies across apps and platforms. Learn more about managing labels.](#)
- Require users to apply a label to their Fabric and Power BI content**
Users will be required to apply labels to unlabeled content they create or edit in Fabric and Power BI.
- Provide users with a link to a custom help page**
If you created a website dedicated to helping users understand how to use labels in your org, enter the URL here.

Corporate Sensitivity Labelling

[Edit policy](#) [Delete policy](#)

Name
Corporate Sensitivity Labelling

Description

Published labels
Public
Internal
Confidential
PHI
Restricted

Admin units
None

Published to
Exchange email - All accounts

Policy settings
Label is mandatory for: documents, emails, sites & groups, meeting
Default label for documents is: Internal
Default label for emails is: Internal
Default label for meetings is: Internal
Users must provide justification to remove a label or lower its classification
Use custom URL to provide more information

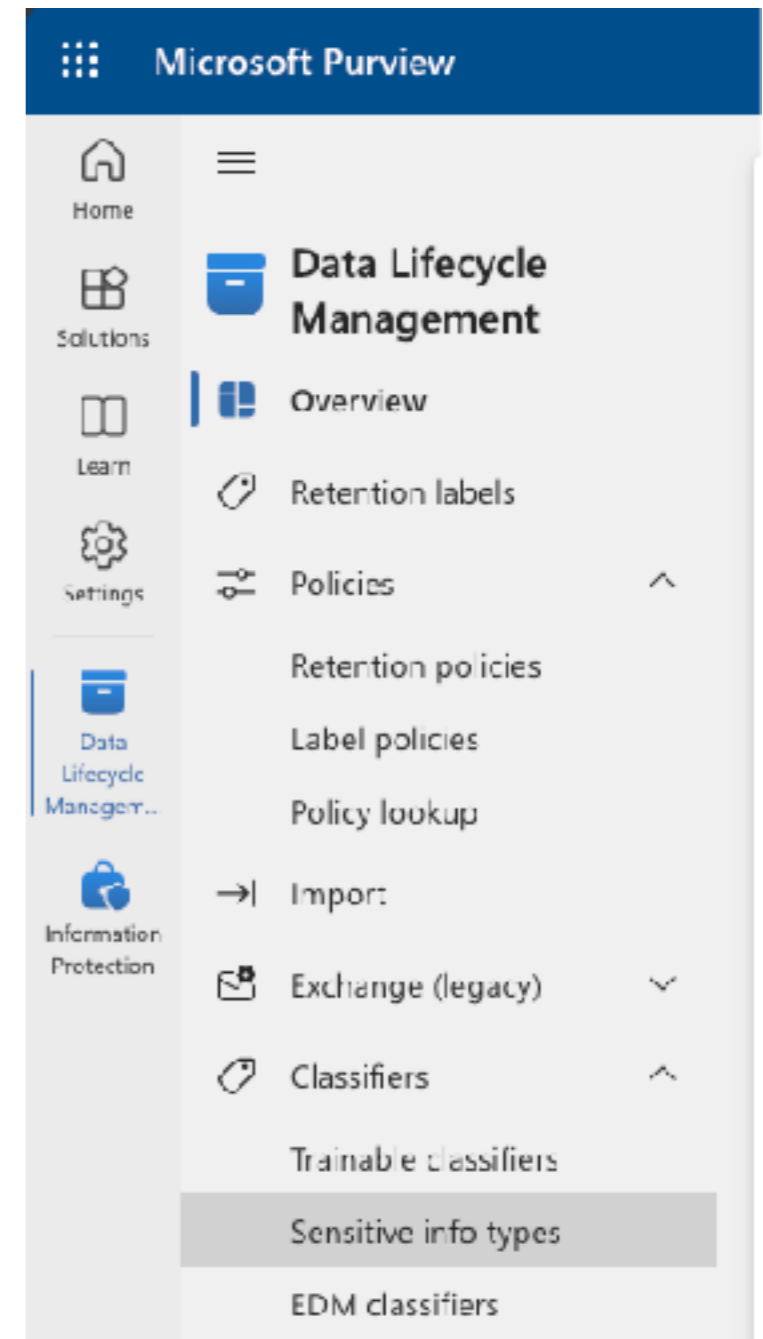
Retention Policies

Apply default policies based on scope

- Exchange mailboxes
- SharePoint classic and communication sites
- OneDrive accounts
- Microsoft 365 Group mailboxes & sites
- Skype for Business
- Exchange public folders
- Teams channel messages
- Teams chats and Copilot interactions
- Teams private channel messages
- Yammer community messages
- Yammer user messages

Policy Options:

- Retain forever
- Delete after a period
- Retain for a period
 - Then delete or do nothing



Retention Policies

Requirements

- Legal
- Regulatory
- eDiscovery or FOI impacts

Separate default policies based on type

- Teams Channels - retain 365d /delete
- Teams Private Channels - retain/delete
- Teams Chat & Copilot - delete after 7 days
- Yammer - delete after 1 day
- OneDrive/SharePoint - retain
- Exchange - retain

The screenshot shows the 'Retention policies' management page. At the top, there is a title 'Retention policies' and a descriptive text: 'Your users create a lot of content every day, from emails'. Below this is a grey informational box with an 'i' icon and the text: 'If your role group permissions are restricted to a specific se'. Underneath the box are three action buttons: '+ New retention policy', '↓ Export', and '↻ Refresh'. The main content area is a table with a 'Name' header. The table lists five default retention policies, each with an unchecked checkbox to its left:

Name
<input type="checkbox"/> Default Teams Channel (365 days)
<input type="checkbox"/> Default Teams Chat (14 days)
<input type="checkbox"/> Default M365 Retention Policy
<input type="checkbox"/> Default Teams Private Channels (365 days)
<input type="checkbox"/> Default Yammer Policy (delete)

Retention Labels

Align labels to requirements

- Ask Legal, Privacy, FOI
- Finance - 7 years
- HR - 3 years after depart
- Corp Records - 10 years
- Align to industry:
 - Patents - 25 years
 - Patient < 18 - 43 years
 - Patient > 18 - 25 years
- Give staff options - 1, 3, 5 yrs
- Users can adjust in OneDrive
- Use auto-label policies to align by type
- No label means retain forever.

Define the period

Choose how long the period is and when it begins.

How long is the period?

7 years

When should the period begin?

When items were created

When items were created

When items were last modified

When items were

Choose what happens after the retention period

These settings determine what happens to items when the retention period ends.

- Delete items automatically
We'll permanently remove labeled items from wherever they're stored.
- Start a disposition review
Let the disposition reviewers you assign in the next step decide if items can be safely deleted or what actions should be taken. [Learn more](#)
- Change the label
You can extend the period by choosing an existing label to replace this one with. [Learn more](#)
- Run a Power Automate flow
Customize what happens to labeled items with a Power Automate flow. You can run a flow to move items to a certain location or sending email notifications. [Learn more about running a Power Automate flow](#)
- Deactivate retention settings
Labeled items won't be retained or deleted when their retention settings are deactivated.

SITs & Trainable Classifiers

Sensitive Info Types

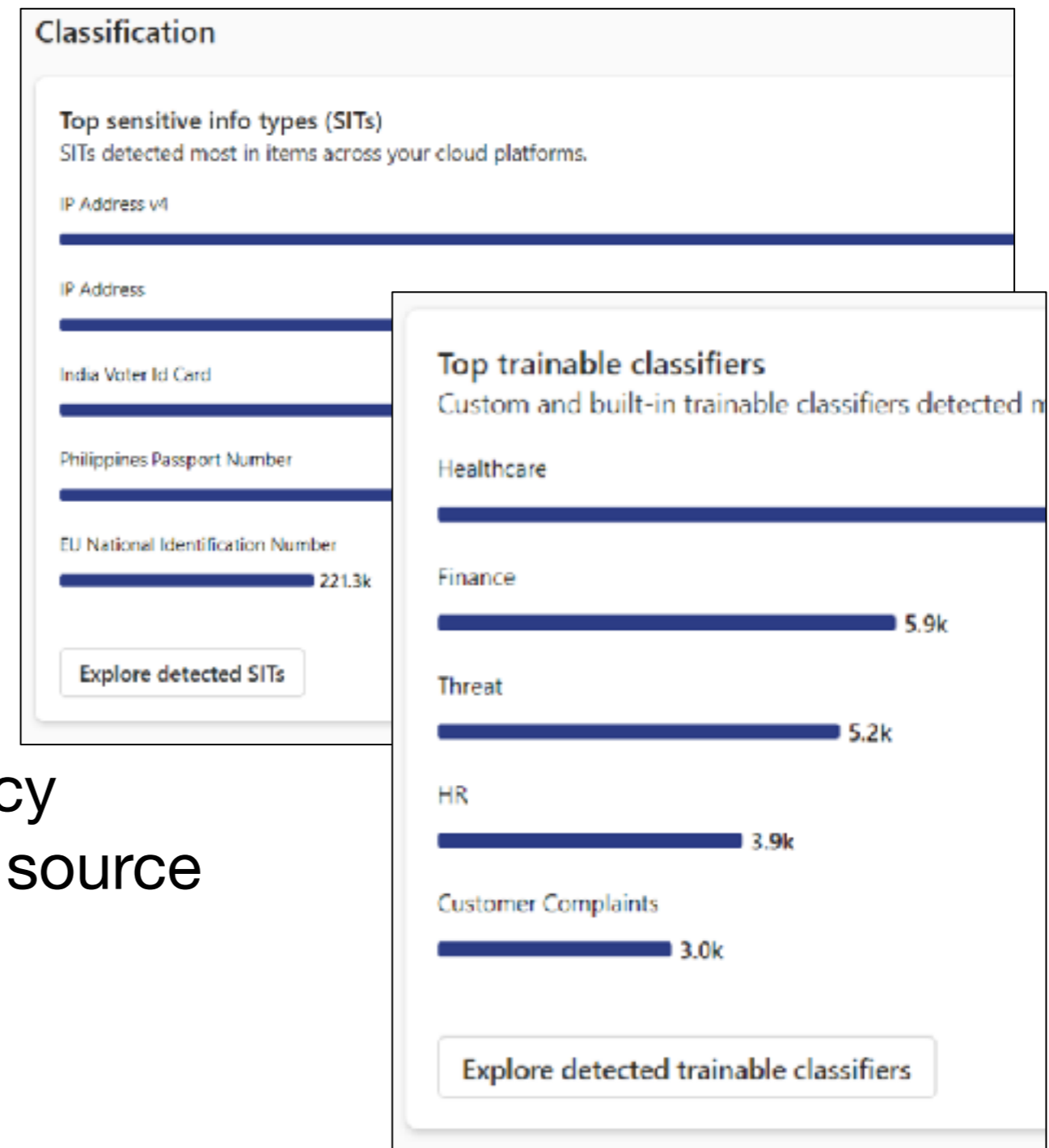
- Predefined RegEx filters
- Custom RegEx, keyword lists
- Eg. IP address, credit cards, Canadian passport numbers, addresses, phone numbers

Trainable classifiers

- Pre-canned data models
- Review / adjust match accuracy
- Create your own - SharePoint source

Exact Data Match

- Train with exact data



Auto Labelling for Retention

Retention policies work best with auto-labelling

- Target selection with:
 - Trainable Classifiers
 - Sensitive Info Types
 - Keywords
 - OneDrive, SharePoint, or M365 Groups

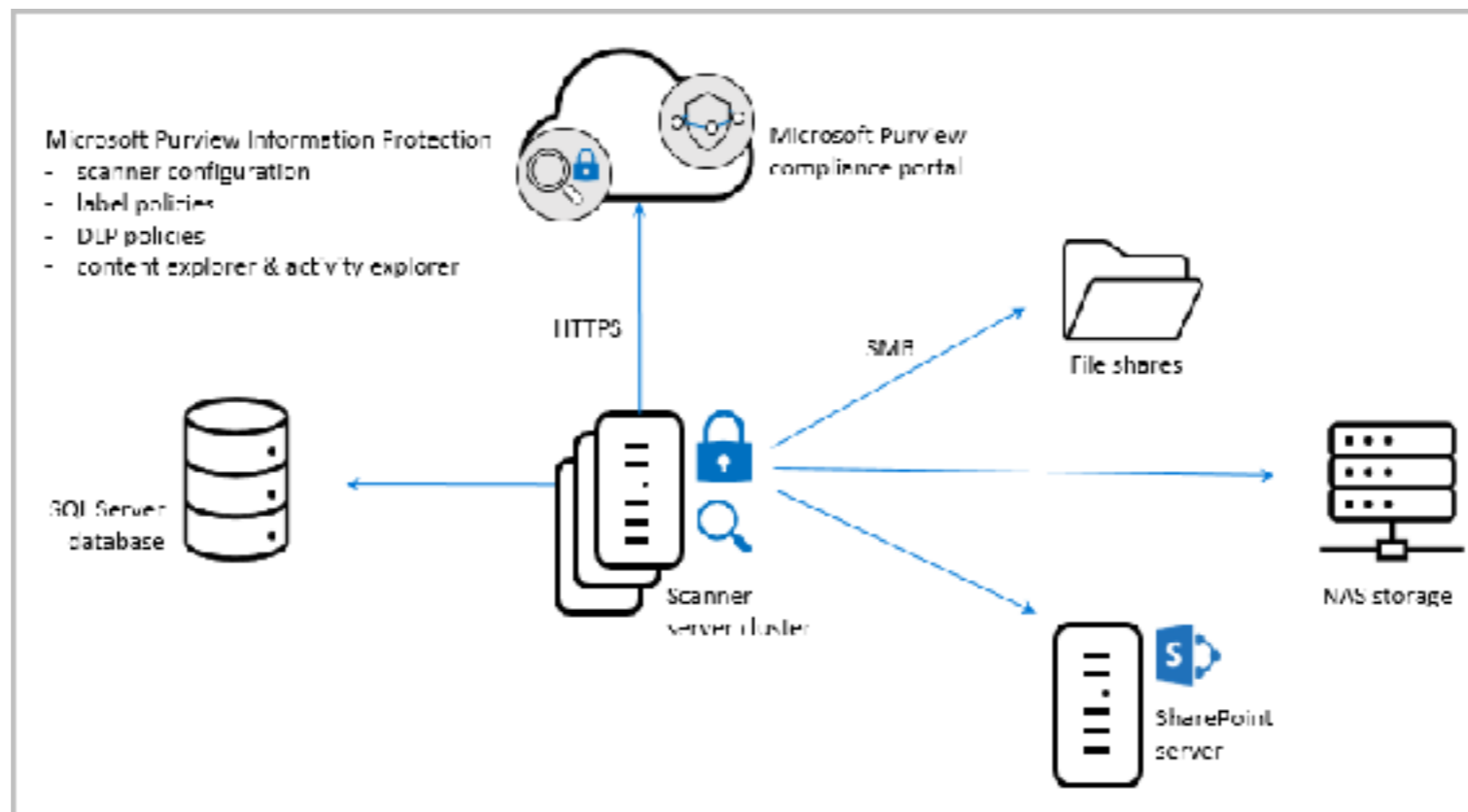
The screenshot shows the Microsoft Purview auto-labeling policy configuration interface. At the top, there are buttons for '+ Create auto-labeling policy' and 'Refresh'. Below this is a table with columns: Name, Locations, Label applied, and Last modified. The table is expanded to show a 'Simulation (4)' section. A context menu is open over the 'Financial Retention' policy, showing options: 'Publish labels', 'Auto-apply a label', and 'Refresh'. The context menu also displays a table with columns: Name, Status, and Type. The 'Financial Retention' policy is shown with a checkbox, a vertical ellipsis menu, 'Enabled' status, and 'Auto-apply' type. The main table shows the 'Financial Retention' policy is applied to 'Exchange, SharePoint, OneDrive' and labeled 'PHI' on 'Sep 25, 2024 10:39 AM'.

Name	Locations	Label applied	Last modified
Simulation (4)			
<input type="checkbox"/> Canada Personal Health Act (PHIPA) - Ontario			
<input type="checkbox"/> Canada Financial Data			
<input type="checkbox"/> PHIPA Data in ODB & SPO			
<input type="checkbox"/> Healthcare Trainable Classifier	Exchange, SharePoint, OneDrive	PHI	Sep 25, 2024 10:39 AM

Name	Status	Type
<input type="checkbox"/> Financial Retention	Enabled	Auto-apply

Information Asset Protection Scanner

- Scan on prem SMB servers and SharePoint servers
 - Requires service account with read access
- Deploy on a local server
- Requires a SQL server back-end
- Scans are slow and detailed
- Reports in CSV & Purview Data Explorer (when it works)
 - C:\Users\aipscanneraccount\AppData\Local\Microsoft\MSIP\Scanner\Reports



Data Loss Prevention

Use

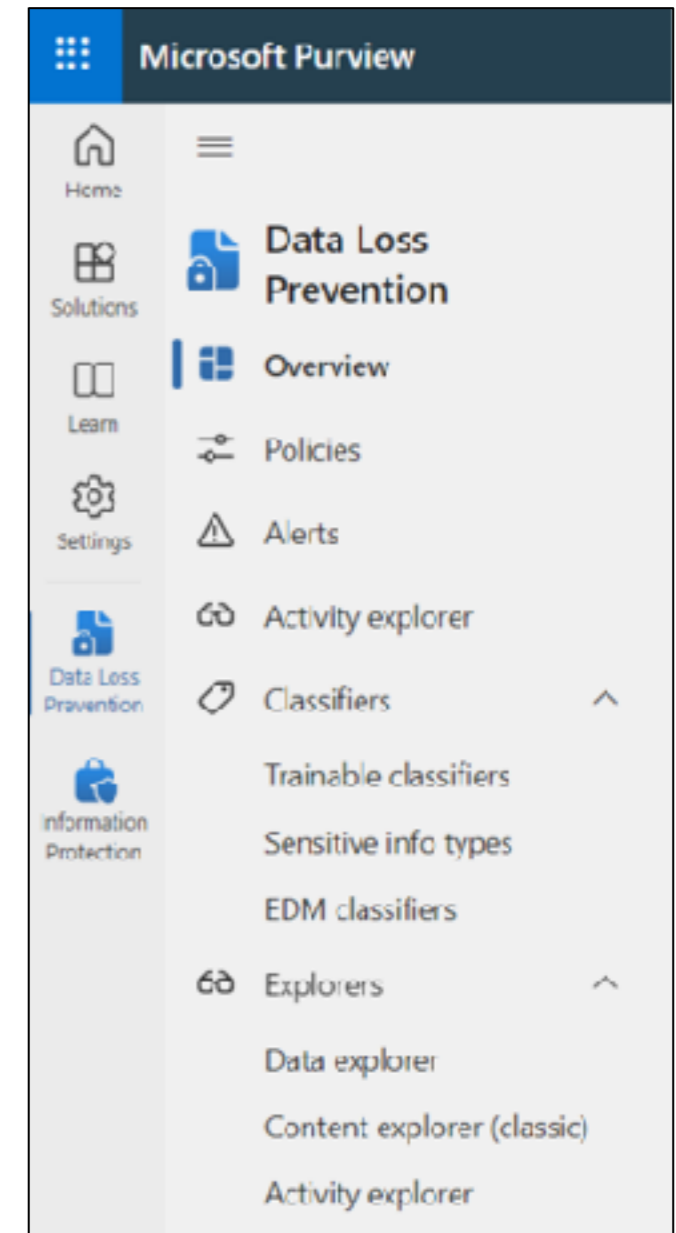
- Trainable classifiers
- Sensitive Info Types
- Exact Data Match
- Sensitivity Labels

Match based on

- Location
- Source type - determines match options
- Shared externally or internal only
- Add in Insider Risk Management

Actions

- Send a nudge
- Send email
- Alert
- Email reports
- Block access or encrypt data



Data Loss Prevention - Conditions

EXO	OD	SP	Teams	Instances	On Prem Repos	Power BI	EDR	Condition
X	X	X	X	X	X	X	X	Content contains
X							X	Content is not labeled
X	X	X	X	X				Content is shared from Microsoft 365
X	X	X					X	Document could not be scanned
	X	X						Document created by
	X	X						Document created by member of
X	X	X					X	Document didn't complete scanning
	X							Document is shared
X	X	X					X	Document name contains words or phrases
X							X	Document name matches patterns
X	X	X					X	Document or attachment is password protected
X	X	X			X		X	Document property is
X	X	X					X	Document size equals or is greater than
X	X	X			X		X	File extension is
X			X				X	Insider risk level for Adaptive Protection is
X			X					Recipient domain is
X			X					Recipient is
X			X					Sender domain is
X			X					Sender is

Data Loss Prevention

Name

^ Low volume of content detected DLP - Canada Financial Data

Conditions

Content contains any of these sensitive info types:

Credit Card Number
Canada Bank Account Number

Content is shared from Microsoft 365
with people outside my organization

Exceptions

Except if content is shared from Microsoft 365
only with people inside my organization

Actions

Notify users with email and policy tips
Restrict access to the content for external users
Send alerts to Administrator
Restrict third-party apps

v High volume of content detected DLP - Canada Financial Data

Policies

Use data loss prevention (DLP) policies to help identify and protect your organization's sensitive info. For ex

i If your role group permissions are restricted to a specific set of users or groups, you'll only be able to manage policies

+ Create policy ↓ Export ↻ Refresh

<input type="checkbox"/>	Name		Priority
<input type="checkbox"/>	Allow T4	⋮	0
<input type="checkbox"/>	Disable Auto-Shared Recording	⋮	1
<input type="checkbox"/>	DLP - Canada Personal Health Act (PHIPA) - Ontario	⋮	2
<input type="checkbox"/>	DLP - Canada Financial Data	⋮	3
<input type="checkbox"/>	DLP - Possible Credential Leak	⋮	4
<input type="checkbox"/>	DLP - Potential Generic Credential Leak	⋮	5
<input type="checkbox"/>	DLP - All Healthcare Terms	⋮	6
<input type="checkbox"/>	DLP - Block sharing sensitive docs - Email	⋮	7
<input type="checkbox"/>	DLP - Block sharing sensitive docs - SP/OD	⋮	8

Data Loss Prevention

Adjust

- Auto-Label credentials as Restricted
- Auto-label PII, Legal Contracts, NDAs, SIN numbers as Confidential, Medical Forms as PHI
- Confidential, PHI - Encrypt and Do Not Forward

Allow

- PII - if labeled as Confidential, PHI if labelled as PHI
 - SIN, Passport Number, Drivers License
- The Employee Discount book - it has plaintext passwords... seriously.
- Sharing externally for authorized users - use Admin Groups

Nudge

- ToolTip popup for sensitivity data types - should you be doing this?
 - Warn - Credentials, SIN, Passport, Credit Card, Bank Accounts, SWIFT codes
 - PHI in Teams
- Enhance email notifications to educate users of proper handling of sensitive info

Deny

- Restricted/Top Secret labelled
- Classifier detected PHI labelled as Public or Internal
- Block unlabelled mail
- Sharing credentials, credit cards, bank account numbers through Teams



Limitations and Frustrations

- Sensitivity Label watermarks are global
- Sensitivity Label identifiers are different for each tenant
 - You have to build incoming auto-classification rules based on markings
- Retention policies have a limited range (1000 items)
 - <https://learn.microsoft.com/en-us/purview/retention-limits>
- Tuning trainable classifiers is slow methodical work
 - You have to read things you never wanted to know
- Microsoft wants to sell you PS to do this
- Every other vendor wants to sell you PS to do this
- Nobody is sharing any of this stuff
- The documentation for the IAP scanner is disjointed and horrible
- Very limited tuning options in DLP rules
- DLP is limited in actions you can take
- DLP alerting during simulation is noisy
- The alert management interface in Defender and Purview are awful
- Everything in M365 is slow

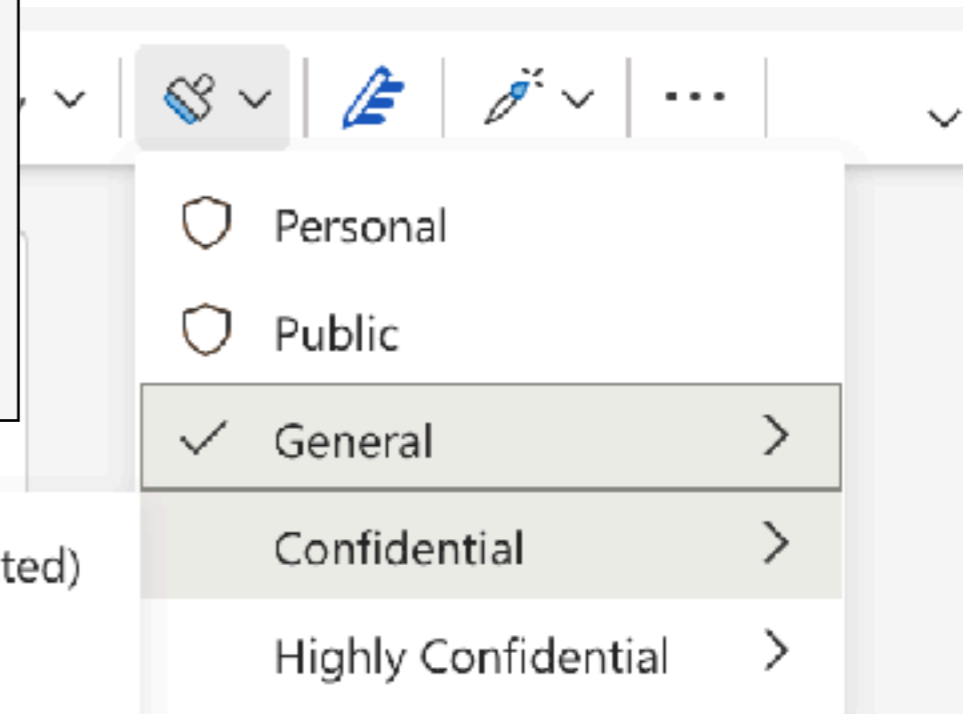
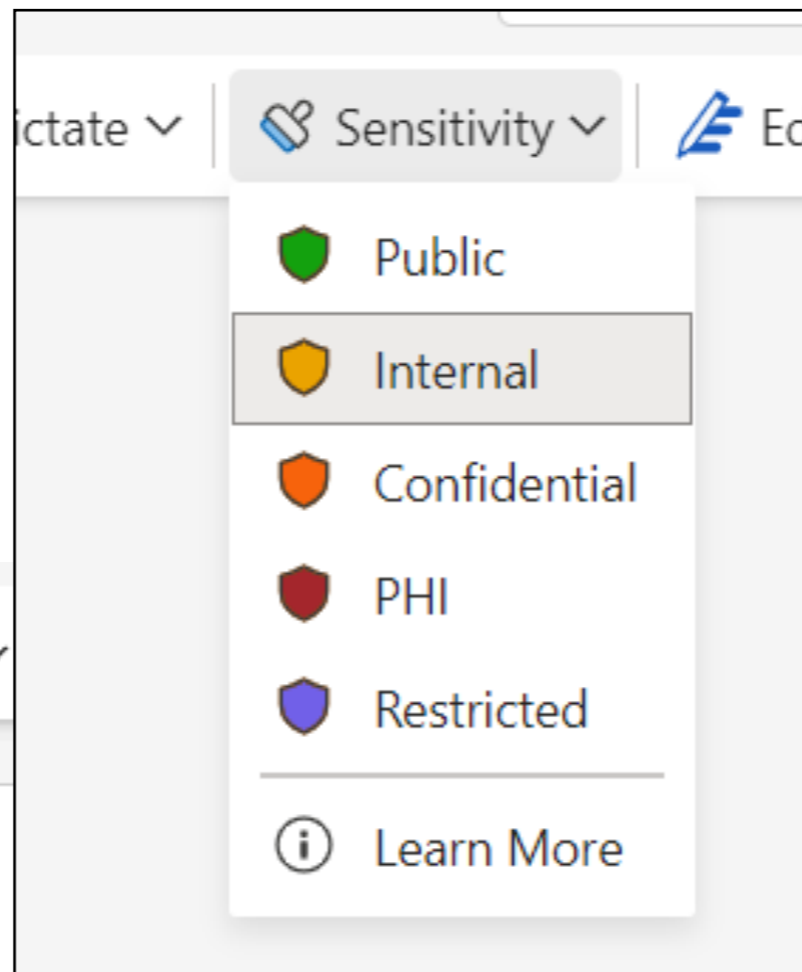
What could go wrong?

- Too many labels = Too complicated
- Lack of support

- Lack of education
 - Poorly delivered
 - Unclear examples
 - Staff are confused

- Under classification
 - Not enough controls
 - Improper management
 - Increased risk
 - Regulatory fines

- Over classification
 - Too many controls
 - Increased friction
 - Blocked workflows
 - Excessive alerts
 - Increased cost



Confidential data for internal/external sharing that can be reshared by trusted recipients.

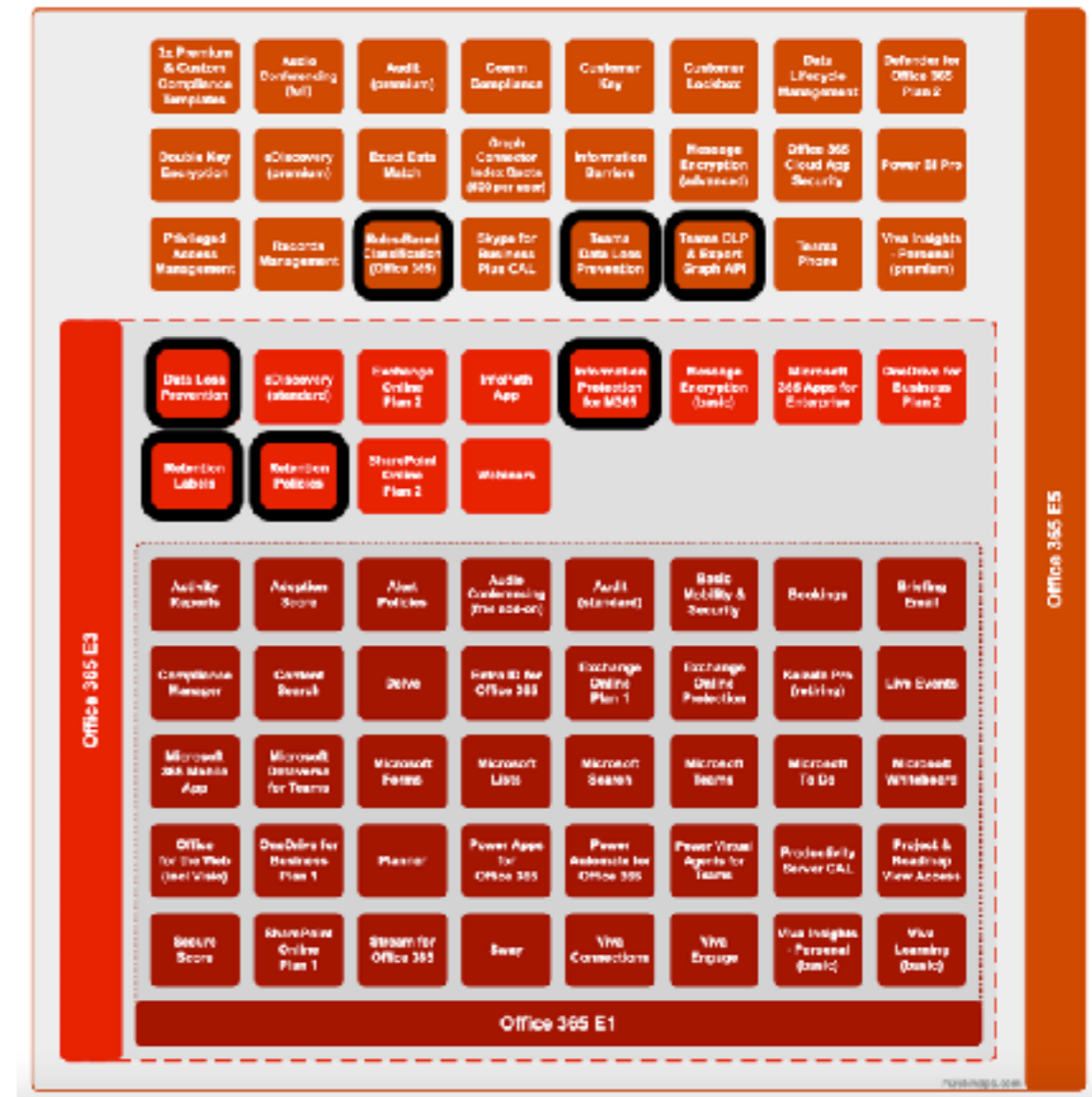
What about licensing?

E3 includes:

- Information Protection for M365
- Retention Labels
- Retention Policies
- DLP

E5 Compliance:

- Trainable Classifiers
- Rules-Based Classification
- Endpoint DLP
- Teams DLP
- Teams DLP & Export Graph API



- Everything is in M365 F5/A5/E5 except Teams Premium support for Sensitivity Labels

But wait, there's more



Conclusions

Start with

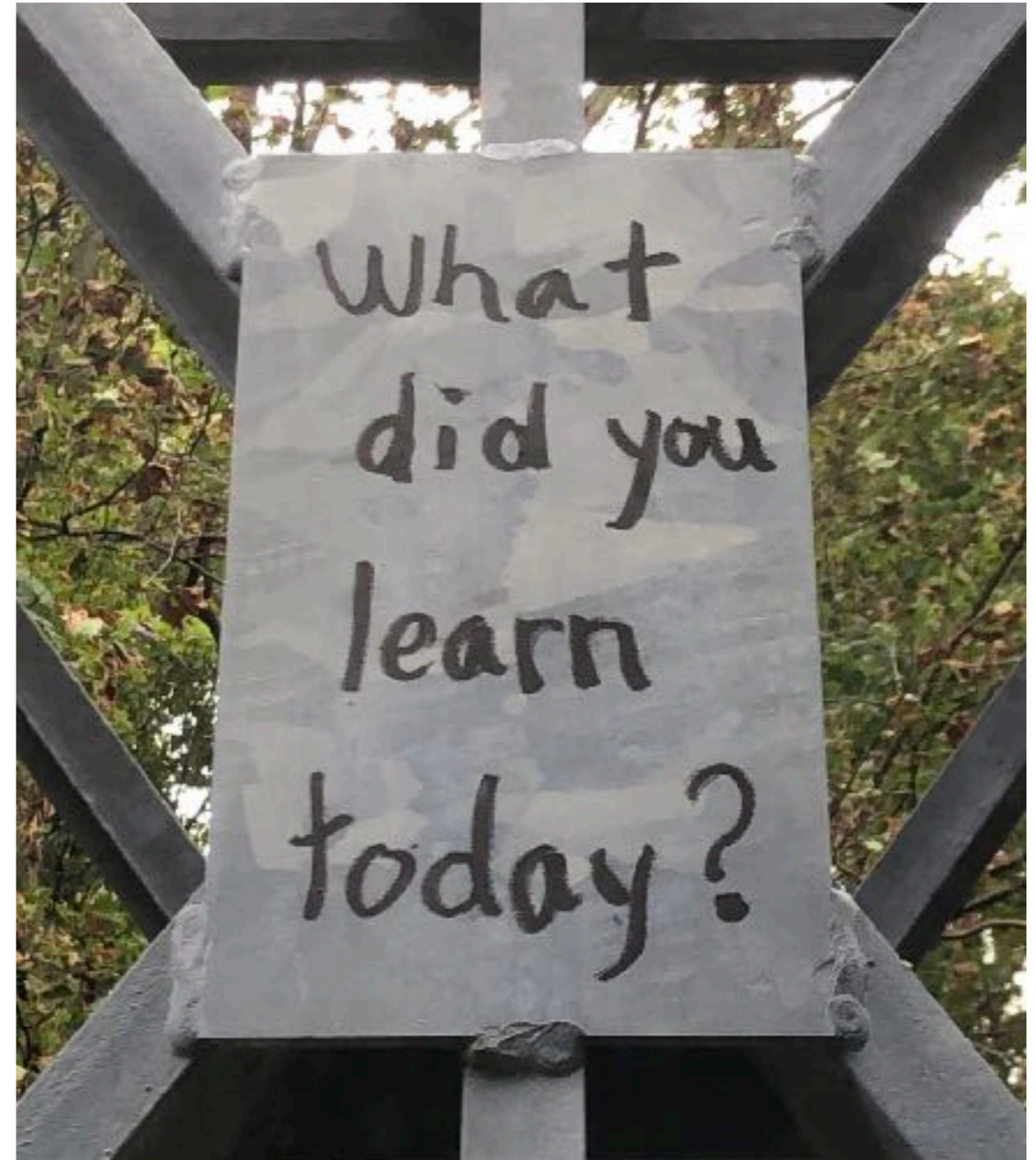
- A plan
- A policy
- Stakeholder engagement
- Education
- Awareness

Then deploy

- Sensitivity Labels
- Retention Policies
- Retention Labels
- IAP Scanner
- Trainable Classifiers
- Auto Labelling
- DLP

Follow with

- Run policies in simulation mode
- Iterate, iterate, iterate
- Use DLP as a nudge before you use it as a cudgel
- Be ready with pre-canned responses to alerts



Resources & Links

M365Maps - Enterprise Landscape

<https://m365maps.com/files/Microsoft-365-Enterprise-Landscape.htm>

Australian Government - Digital Transformation Authority - Protected Utility Blueprint

<https://blueprint.asd.gov.au>

<https://blueprint.asd.gov.au/configuration/purview/>

eHealth Ontario - Information Asset Management Standard

https://ehealthontario.on.ca/files/public/support/Security/Security_Toolkit/Information_and_Asset_Management_Policy_EN.pdf

National Defence - Working with Sensitive Information Infographic - Injury Test

<https://www.canada.ca/en/department-national-defence/maple-leaf/defence/2020/12/working-with-sensitive-information.html>

National Defence - Levels of Security

<https://www.tpsgc-pwgsc.gc.ca/esc-src/documents/levels-of-security.pdf>

<https://www.tpsgc-pwgsc.gc.ca/esc-src/protection-safeguarding/niveaux-levels-eng.html>

DCMA - Classifying Information - Section 4

<https://www.dcma.mil/Portals/31/Documents/Policy/DCMA-MAN-3301-08.pdf>

Australian Government - Classification System

<https://www.protectivesecurity.gov.au/publications-library/policy-8-classification-system>

NIST - Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) - Section 3 - Impact Levels

<https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf>

Canada - Treasury Board - Standard Classes of Records

<https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information/info-source/standard-classes-records.html>

Ontario Hospital Association - Record Retention Toolkit

https://www.oha.com/Legislative_and_Legal_Issues_Documents1/Records_Retention_Toolkit_September_2022.pdf

Data Labelling: The Authoritative Guide (for ML)

<https://scale.com/guides/data-labelling-annotation-guide>

Not horrible example blogs showing setting up DLP policies:

<https://www.gitbit.org/course/ms-500/learn/preventing-accidental-and-malicious-data-loss-with-dlp-policies-ispgsme8w>

<https://alberthoitingh.com/2023/11/10/running-the-aip-scanner-in-detect-only-mode/>

Questions?

Thank you

Contact Info:

singleusermode@infosec.exchange
nixuser23@gmail.com

Do not contact me on LinkedIn unless
you talk to me first.



**This QR
code is safe**

<https://github.com/nixy23/bsidesto2024>

