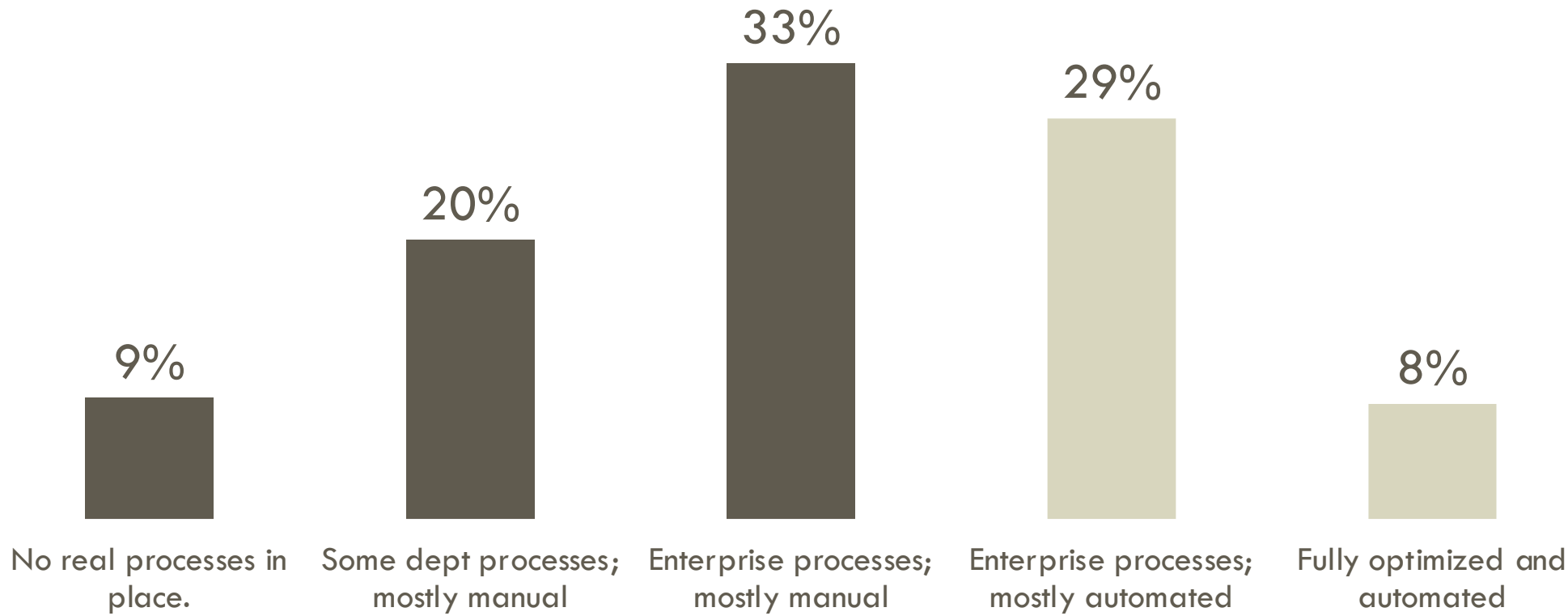
M365 AUTO-CLASSIFICATION

Infotechtion

3 SIGNS THAT YOU HAVE A CLASSIFICATION PROBLEM

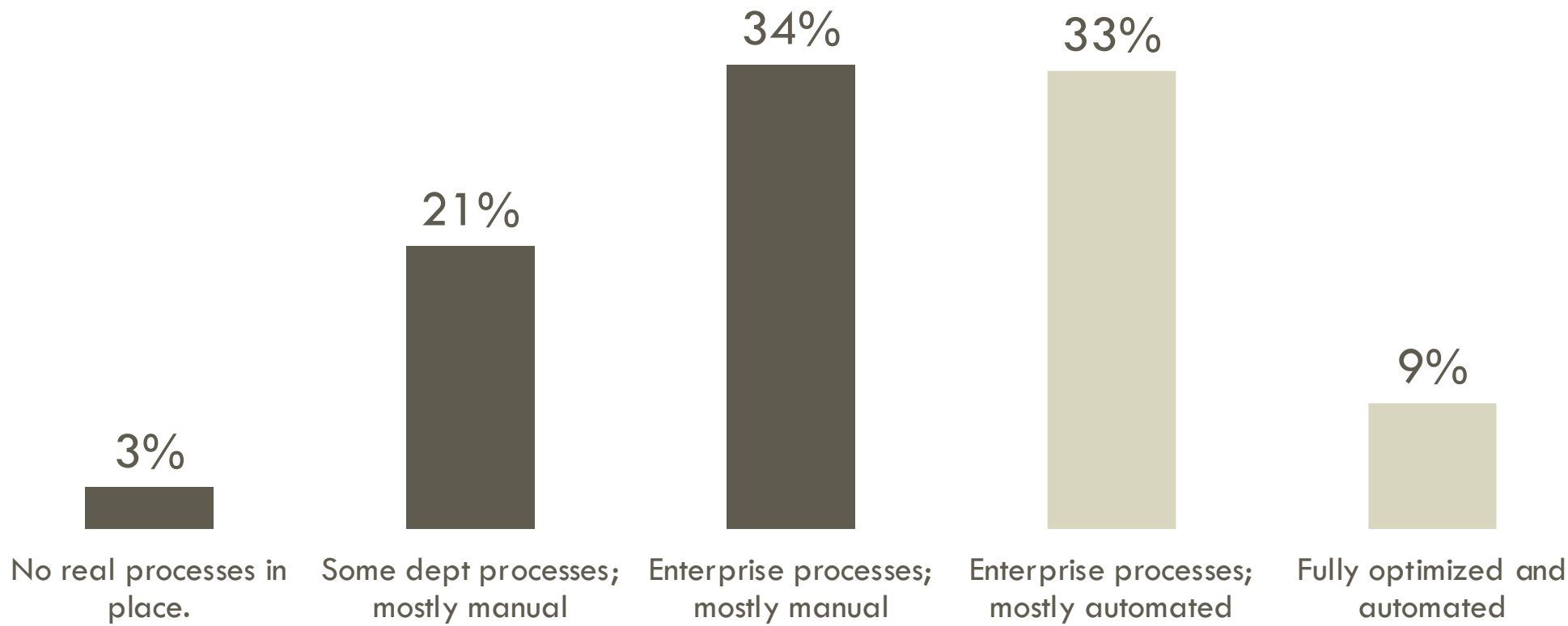
"DISPOSITION OF REDUNDANT, OBSOLETE, AND TRIVIAL CONTENT IS AN ACCEPTED AND REGULAR ORGANIZATIONAL DISCIPLINE."



Source: AIIM, 2020

Infotechtion

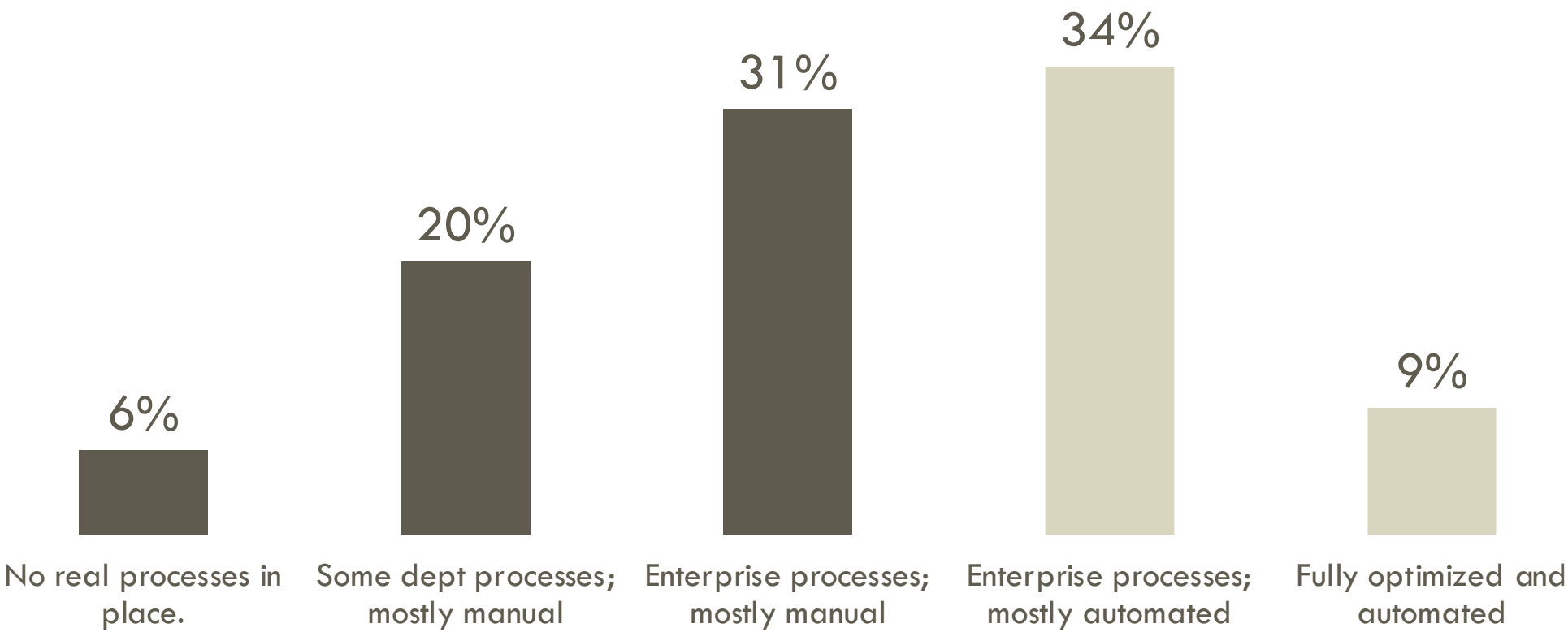
"POLICIES AND PROCESSES EXIST TO IDENTIFY AND PROTECT PERSONALLY IDENTIFIABLE INFORMATION."



Source: AIIM, 2020

Infotechtion

"A CLEAR ORGANIZATION-WIDE STRATEGY EXISTS FOR METADATA."

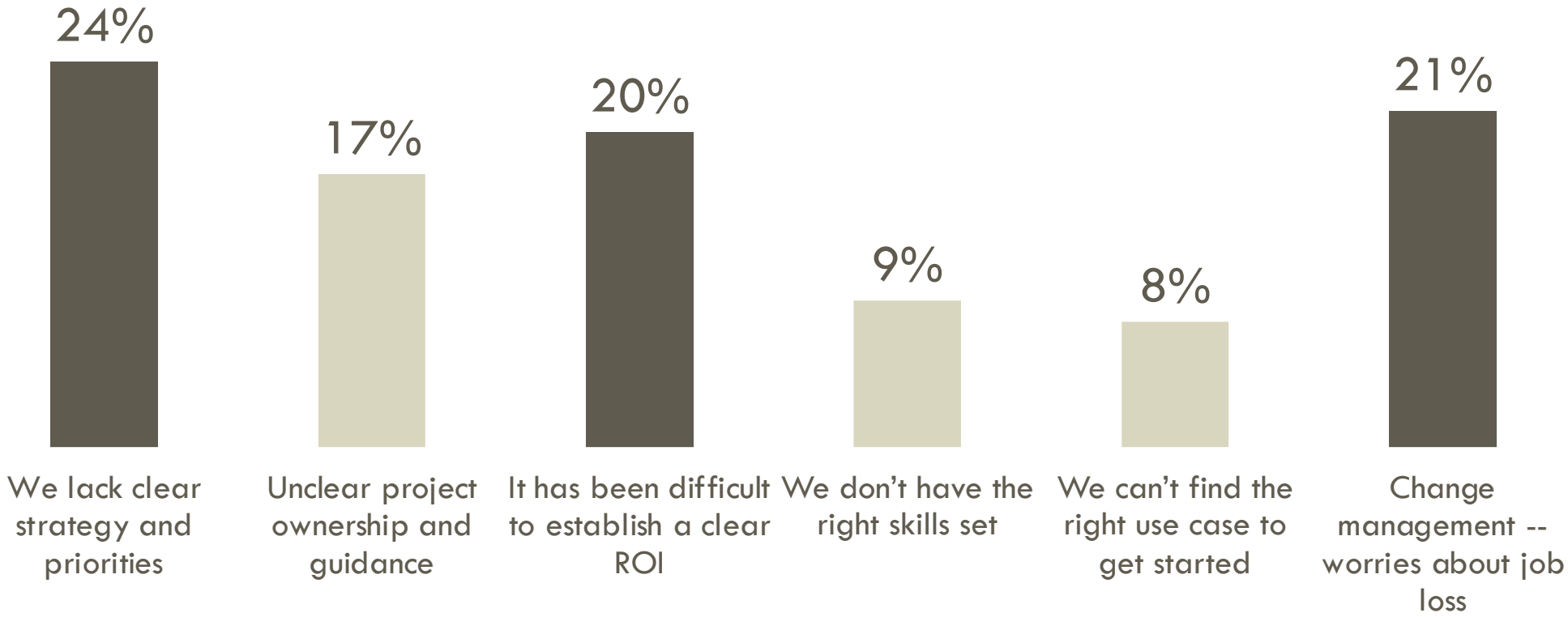


Source: AIIM, 2020

Infotechtion

WHAT BARRIERS DO YOU NEED TO ADDRESS AS YOU CONSIDER AUTO-CLASSIFICATION?

WHICH OF THE FOLLOWING IS THE MOST IMPORTANT BARRIER TOWARD MOVING FORWARD AGGRESSIVELY WITH GOVERNANCE PROJECTS?



Source: AIIM, 2020

INFOTECHTION INFORMATION GOVERNANCE VISION

Value: Worker

AI based Content Discovery

- Microsoft Viva Topics
- Enterprise / Cognitive Search

Value: Work Group

AI / ML based Auto classification

- SharePoint Syntex
- Trainable classifiers
- Azure AI / Pipelines

Value: Organization

Governance & Compliance

- Retention
- Protection and Privacy
- Knowledge Management

Architecture Foundation

PURPOSE OF AUTOCLASSIFICATION

Search

Investigation

eDiscovery

IG Insights

Type identification

Function

Workflow

Integration

Context and Value

Quality

Risk

Sensitivity

Metadata Extraction

Protection

Reporting

Integration

WHY

Challenges to Manual Classification

- Staff don't remember what they have
- Staff don't know what a record is
- More expense / less available staff have more files to classify (because they have been there longer)
- Sometimes, classification is the only way to achieve compliance

Typically:

30 seconds to figure out what it is

30 seconds to tag with simple metadata

1,500 files per person = 1.5GB

= 3 days for every person

Microsoft has 166,475 staff @$500 per day = $250M

MANAGING INFORMATION – WHY WE CLASSIFY

Indexing and classification impact on information management (ISO 15489)

- Linking individual records
- Consistent naming of records over time
- Classification assists in retrieving records relating to a particular function, topic, or activity

Also

- Helps users find content when it is not where it is expected (They can follow the same logic as the classification and track it down)
- It can be used to link (integrate) systems of similar purpose
- It is evidence of the logic behind why something was done
- User permissions: facilitates managing user permissions for access to, or action on, particular groups of records

CLASSIFICATION SOURCES

Content	Format	Metadata (Context)	Workflow
• Words • Word number patterns • Phrases, topics, lists • Similarity	• Template • Rendition • Image • Version • Utility	• Names, acronyms • Authorship or responsibility • Transmission / distribution • Age relevance • Lifecycle status	• Time • Authority • Status • System • Process

Content indicates topic, similarity, sentiment
Context indicates function, status, ownership

LIFECYCLE CONTEXT

Data in Motion

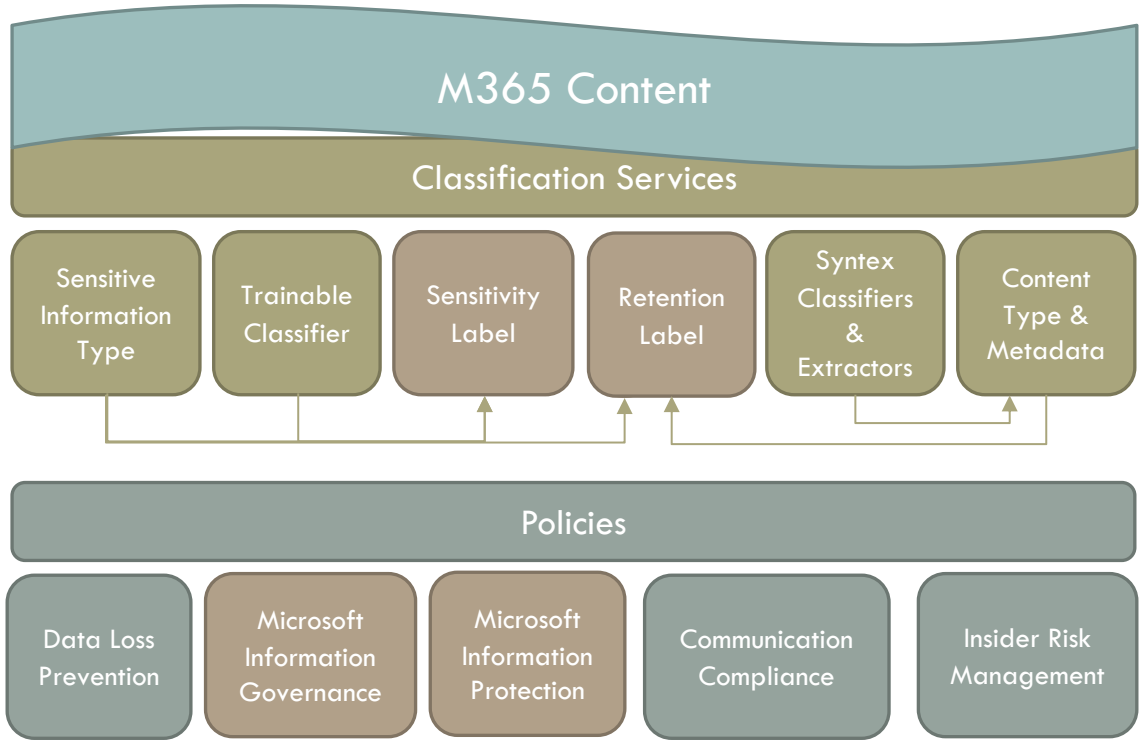
- Content has capturable value from the activities around its creation
- System used to create it, the workflow being run, the transaction supporting it, the user writing it, the web page transmitting it, etc.

Data at Rest

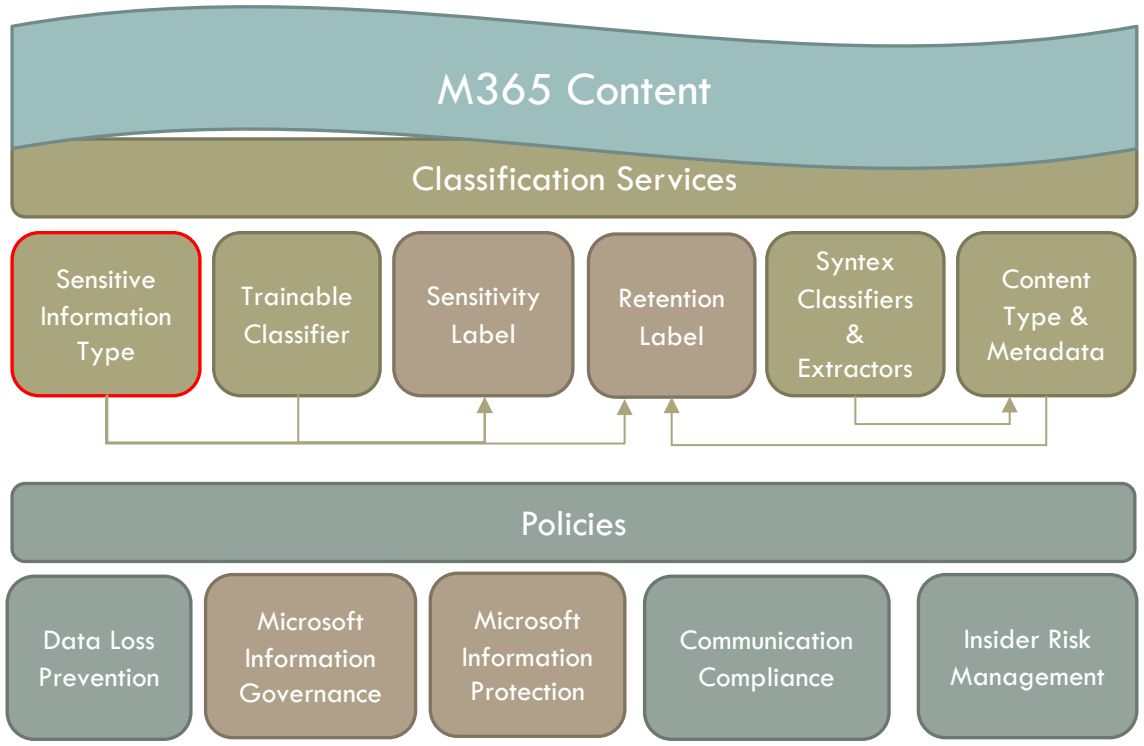
- Much of this context goes away once completed
- Where it was saved or archived, what it was named, what attributes are fixed.

The content doesn't change between these states.

CLASSIFICATION ECOSYSTEM



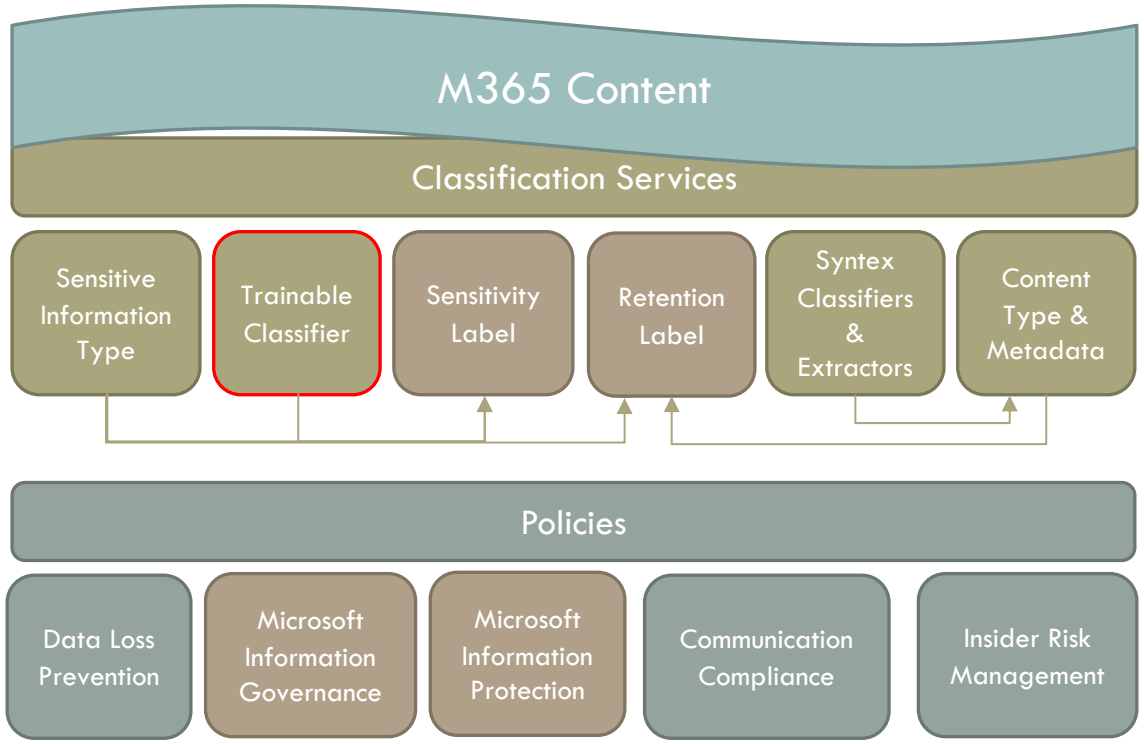
CLASSIFICATION ECOSYSTEM



Sensitive Information Type

- Built-in
  - Credit Card,
  - Social Security
  - Number,
  - IBAN,
  - …
- Custom
  - RegEx
  - Dictionary
  - Fingerprint
  - Supporting Elements with proximity
- Exact Data Match

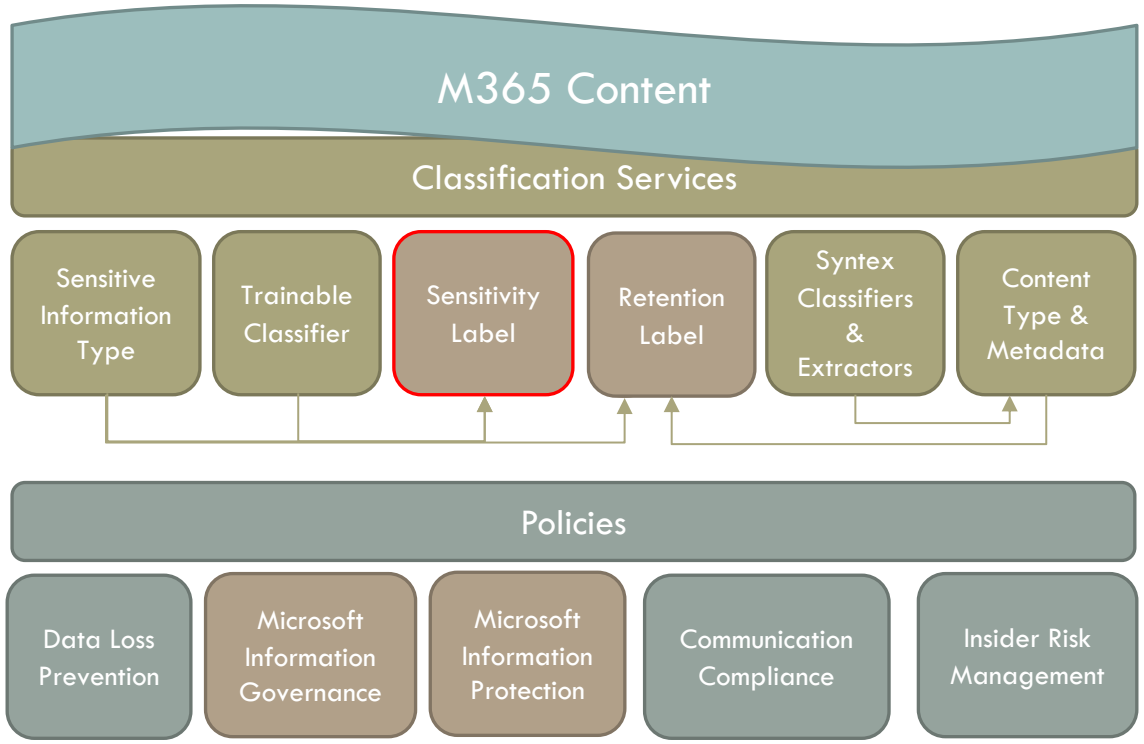
CLASSIFICATION ECOSYSTEM



Trainable Classifier

- Built-in
  - Profanity
  - Resumes
  - Source Code
  - Threat
  - Target Harassment
- Custom
  - Seed
  - Test
  - Validate
  - Publish

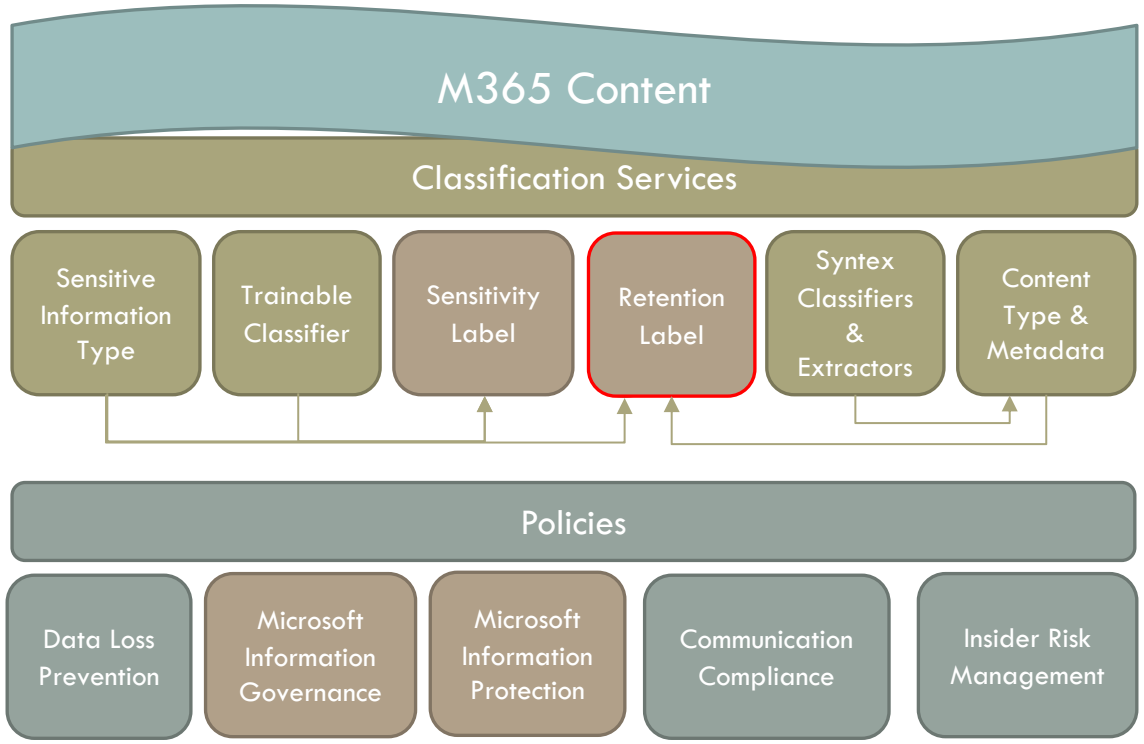
CLASSIFICATION ECOSYSTEM



Sensitivity Label

- Not only a classification label
- Encryption
- Content Marking
  - Watermark
  - Footer
  - Header
- Groups & Sites (out of scope here)
- Applied
  - Manually
  - Suggestion
  - Automatically

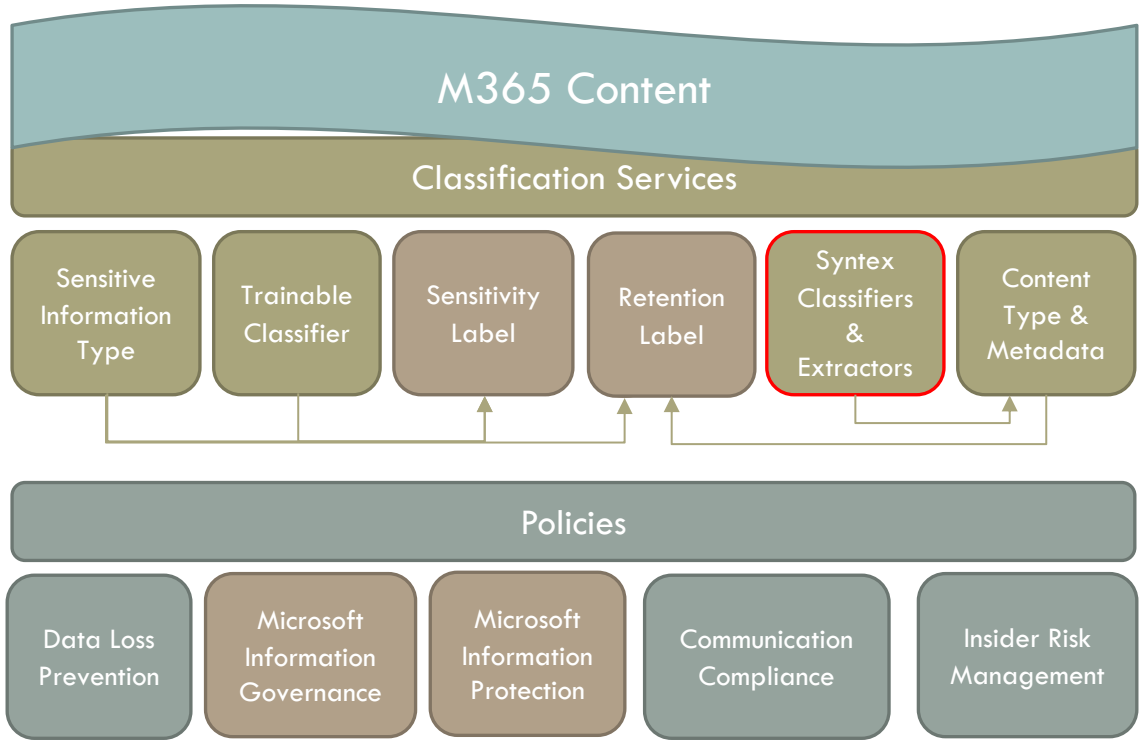
CLASSIFICATION ECOSYSTEM



Retention Label

- Not only a classification label
- Retention
- Deletion
- Record Management
- Disposition
- Trigger
- Applied
  - Manually
  - Default
  - Automatically

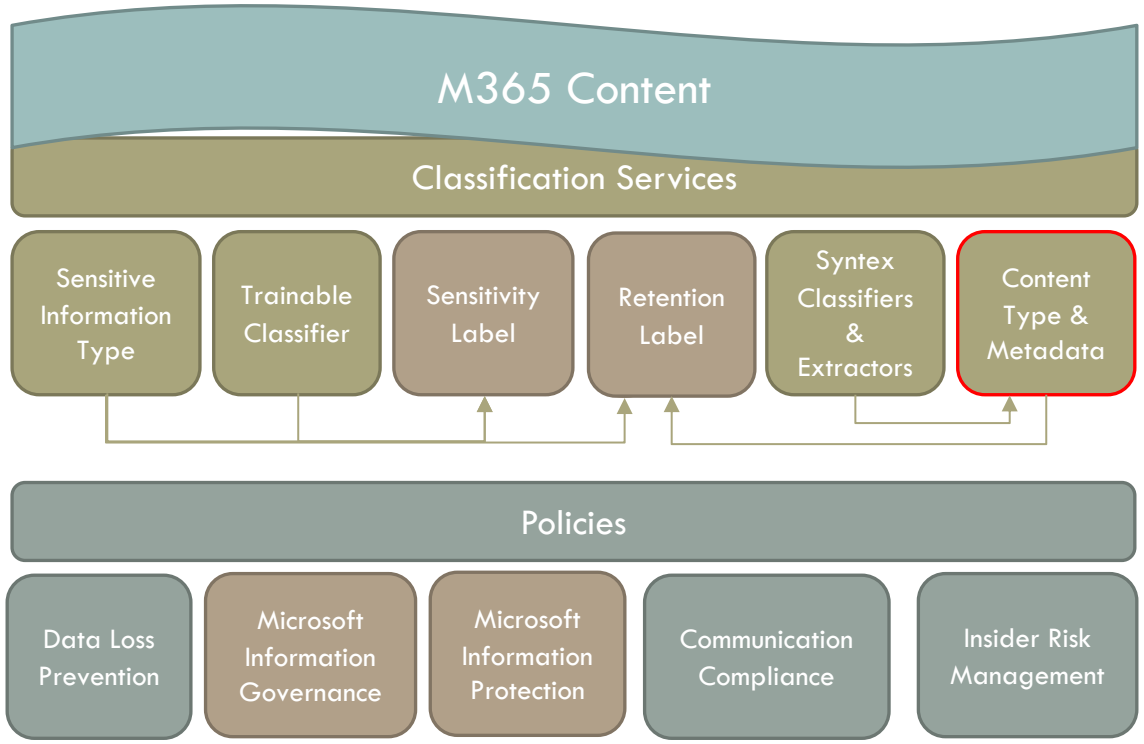
CLASSIFICATION ECOSYSTEM



Machine Teaching – Providing Explanations

- Classifiers – Recognize type of content
  - → Content Type
- Extractors – Extract entities/concept
  - → Site Column
- Requires (very) structured documents

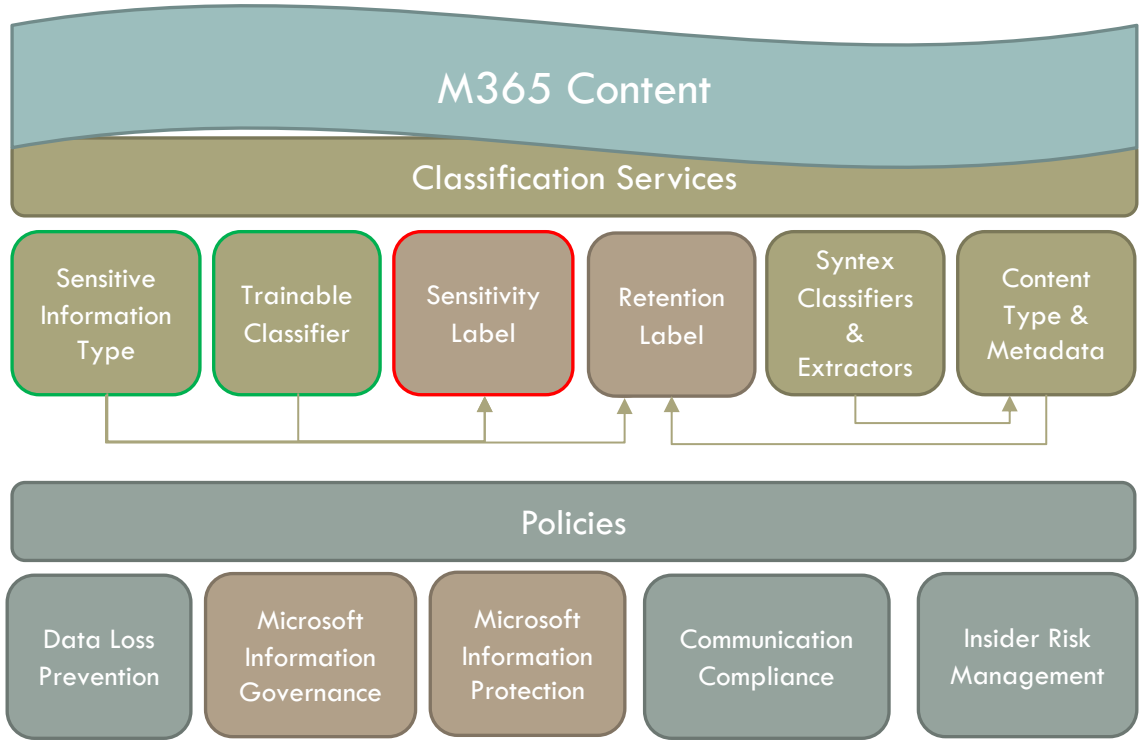
CLASSIFICATION ECOSYSTEM



Content Type and Metadata

- Typical Information Architecture Artifacts
- Content Type define types of document
- Site Column define metadata
  - Taxonomy/folksonomy

AUTO-LABELING: SENSITIVITY LABELS



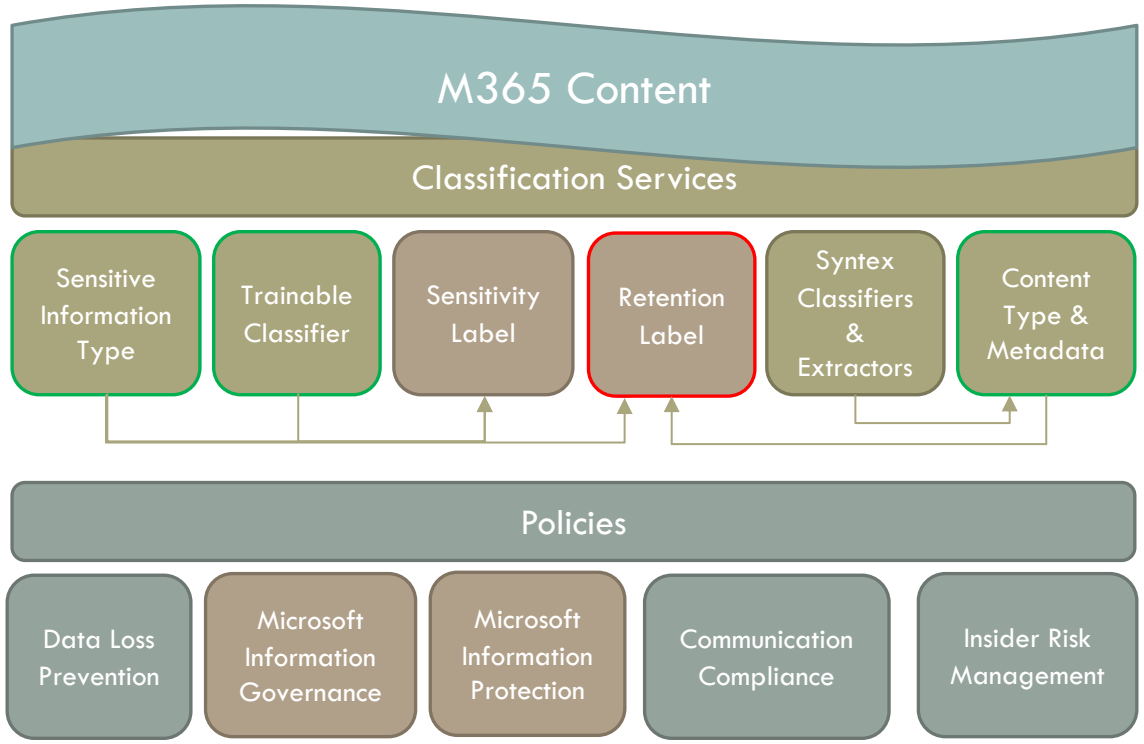
Client-side

- Edit Office files or compose, reply to, or forward emails from Outlook
- Apply or Recommend label
- Conditions
  - SITs
  - Trainable Classifiers
- Requires
  - AIP unified labeling client
  - Some version of Office

Service-side

- OneDrive, SPO at rest
- Email in-transit
- Conditions
  - SITs
  - Mail specific (recipient and attachment)
- Limit
  - Maximum of 25,000 automatically labeled files in your tenant per day.
  - Maximum of 10 auto-labeling policies per tenant, each targeting up to 10 sites (SharePoint or OneDrive).
- Existing values for modified, modified by, and the date are not changed as a result of auto-labeling policies—for both simulation mode and when labels are applied.
- When the label applies encryption, the Rights Management issuer and Rights Management owner is the account that last modified the file.

AUTO-APPLY: RETENTION LABELS



Service-side

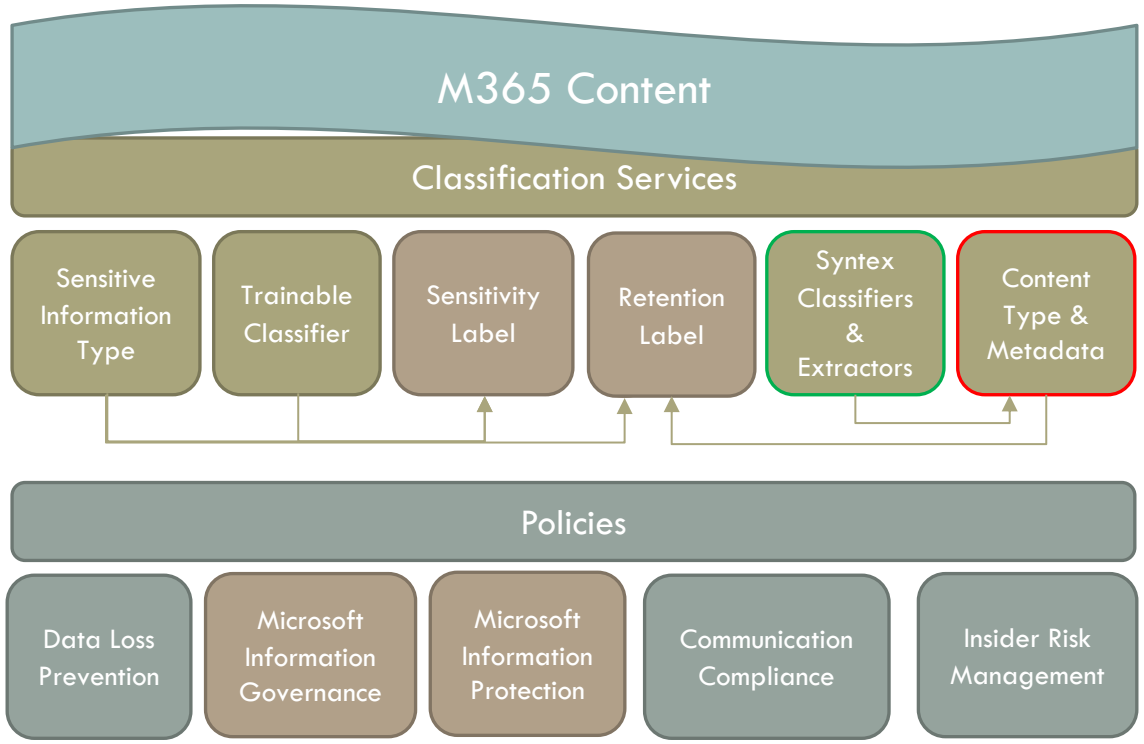
- OneDrive, SPO at rest
- Email in-transit

- Conditions
  - SITs
  - Trainable Classifiers
  - KQL
    - Keywords
    - Properties

KQL: Properties

- Some properties are workloads specific
  - E.g.: recipients, subject, contenttype
- Custom Managed Properties are not supported
  - RefinableStringxx, RefinableDatexx are supported
- Use SPO search or Content Search to validate your KQL conditions

CONTENT TYPE AND METADATA EXTRACTION



Classifier and Extractor

- Classifier are represented by Content Type
- Extractors are represented by Site Column

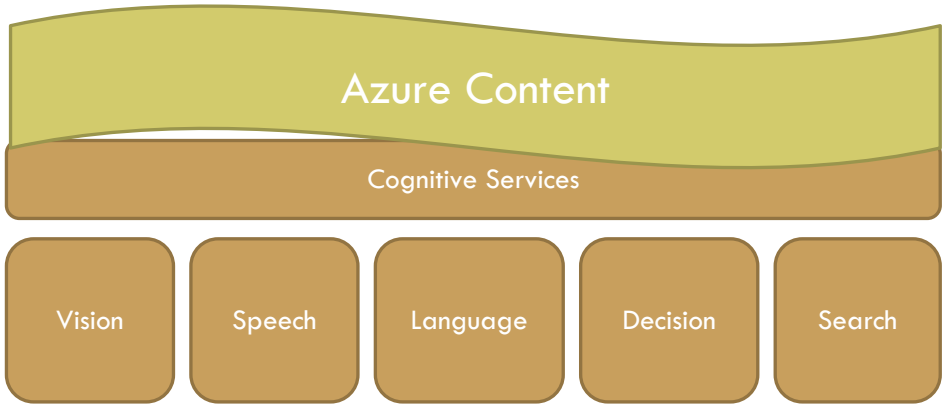
CLASSIFICATION ECOSYSTEM

Cognitive Services – set or services to build cognitive intelligence

- Vision
- Speech
- Language: amongst other Text Analytics
- Decision
- Search

Text Analytics – Natural Language Processing capabilities to build insights

- Language Detection
- Entity Extraction
  - Key Phrases, Named Entities, PII Entities
- Sentiment Analysis
- Medical Terminology



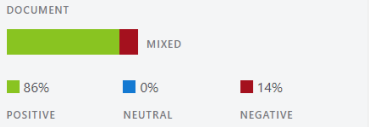
AZURE COGNITIVE SERVICES

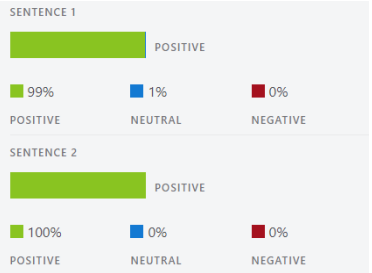
We went to Contoso Steakhouse located at midtown NYC last week for a dinner party, and we adore the spot! They provide marvelous food and they have a great menu. The chief cook happens to be the owner (I think his name is John Doe) and he is super nice, coming out of the kitchen and greeted us all. We enjoyed very much dining in the place! The Sirloin steak I ordered was tender and juicy, and the place was impeccably clean. You can even pre-order from their online menu at www.contososteakhouse.com, call 312-555-0176 or send email to order@contososteakhouse.com! The only complaint I have is the food didn't come fast enough. Overall I highly recommend it!

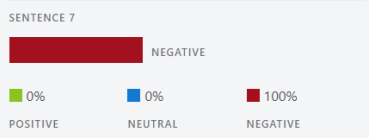
LANGUAGES: English (confidence: 100 %)

KEY PHRASES: place, online menu, great menu, marvelous food, midtown NYC, week, dinner party, Contoso Steakhouse, pre-order, John Doe, Sirloin steak, chief cook, owner, kitchen, spot, dining, complaint, email

SENTIMENT:







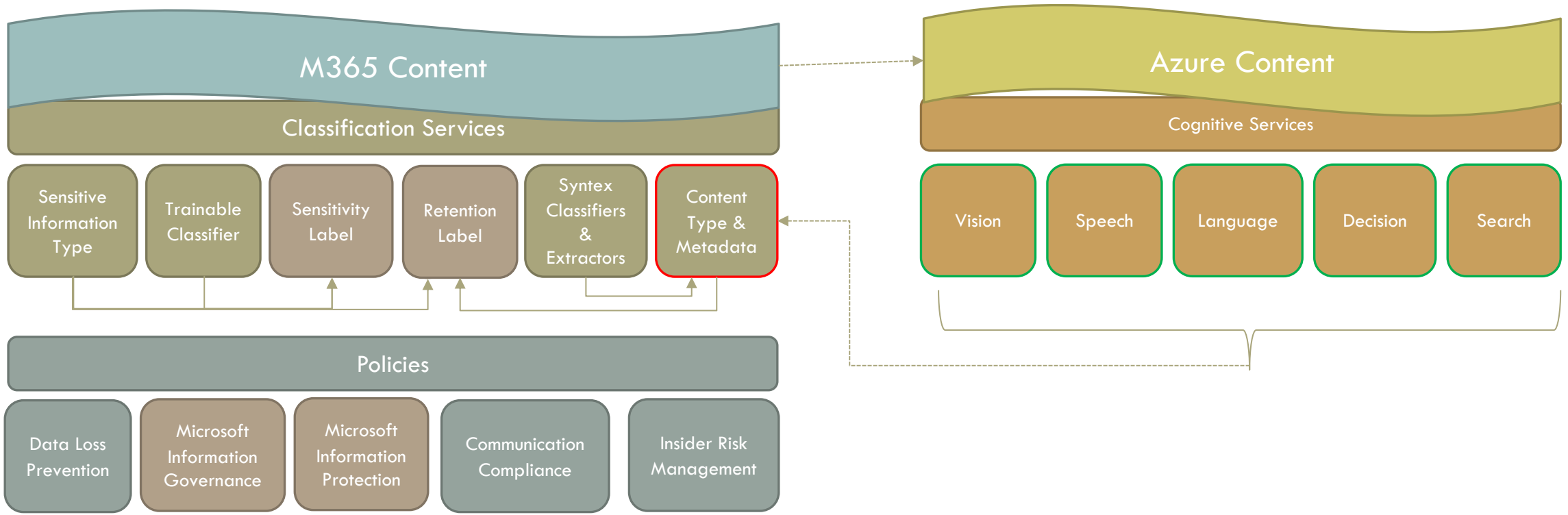
NAMED ENTITIES:
Contoso [Organization]
Steakhouse [Location]
NYC [Location-GPE]
last week [DateTime-DateRange]
dinner party [Event]
chief cook [PersonType]
owner [PersonType]
John Doe [Person]
kitchen [Location-Structural]
Sirloin steak [Product]
www.contososteakhouse.com [URL]
312-555-0176 [Phone Number]
email [Skill]

PII ENTITIES:
Type: Person
Value: John Doe

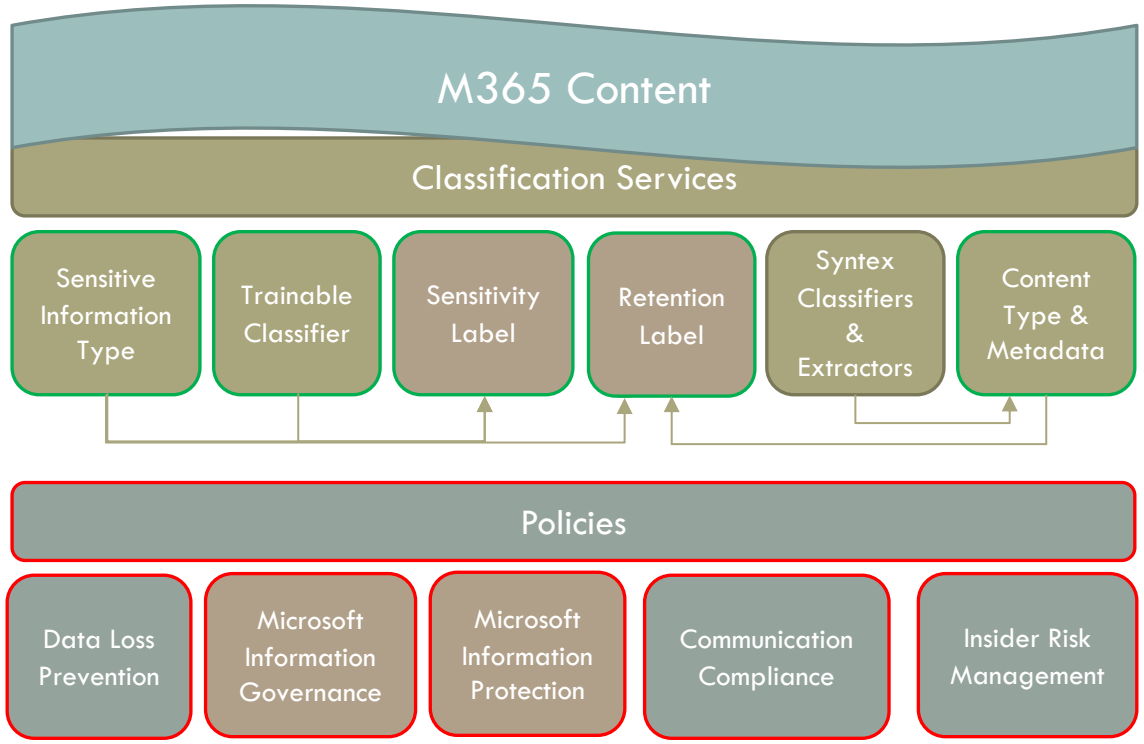
Type: URL
Value: www.contososteakhouse.com

Type: Phone Number
Value: 312-555-0176

CLASSIFICATION ECOSYSTEM



LEVERAGING CLASSIFICATION TO ENFORCE BEHAVIOR



- DLP allow you to use SITs and Labels (both retention and sensitivity) to avoid sharing sensitive content
- MIP allow you to enforce document protection even outside M365 (encryption, action restriction)
- MIG enforce document retention and deletion
- Communication Compliance allows you to supervise employees' communication
- Insider Risk Management allows you to monitor employees' behavior

STRATEGIES FOR SUCCESS

THE 4 STEP CLASSIFICATION ACTIVITY

1. Identify Sources

- The value of a good map
- Which of your 5,000 shares and 60,000 on-prem SharePoint sites have XYZ data

2. Index/Extract content and metadata

- Extract necessary data
- Understand and Determine Issues and Goals

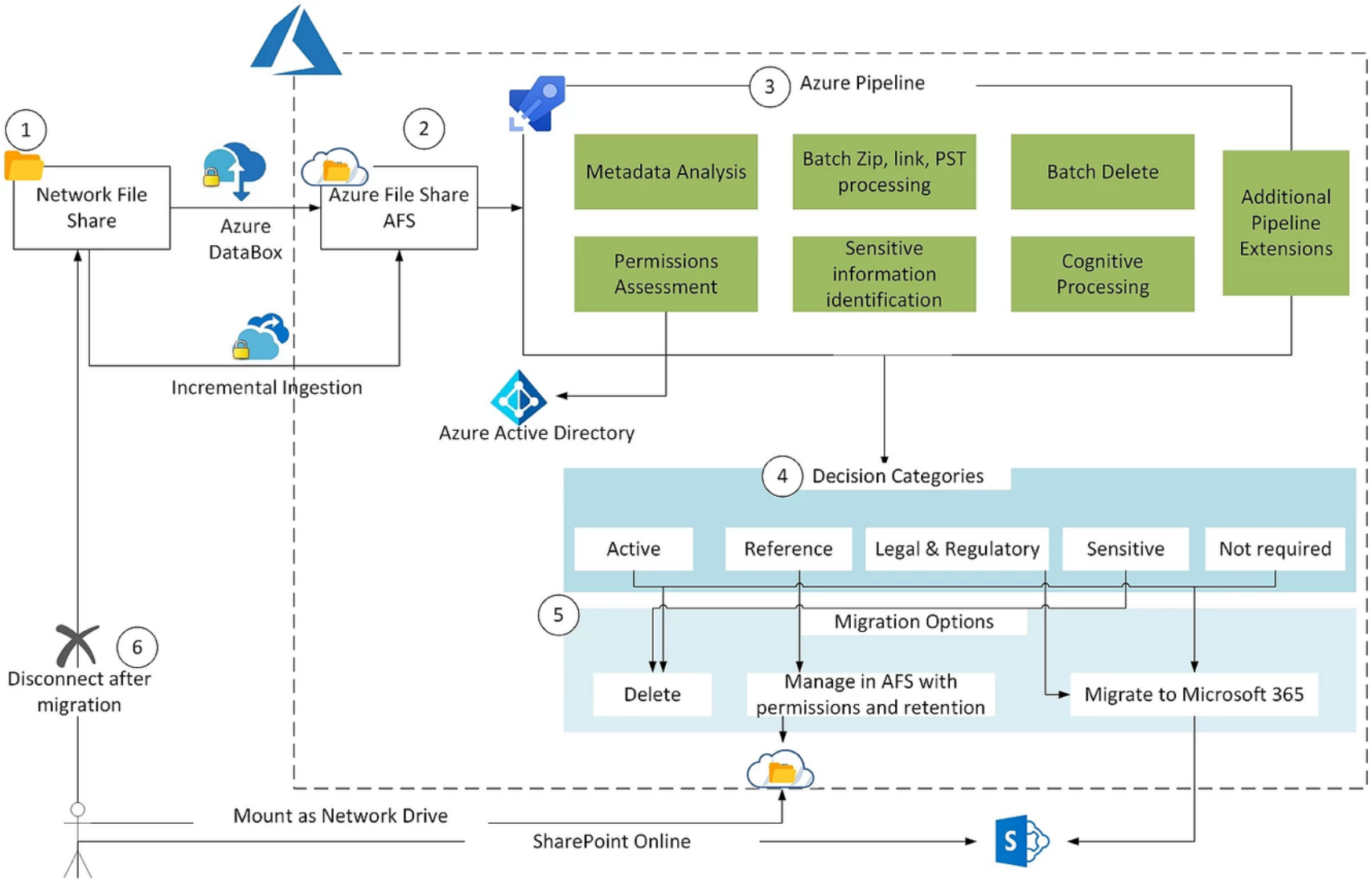
3. Classification Features and Functions

- Metadata
- Multi-element classifications
- Multi-faceted taxonomies
- Training classifiers

4. Action and purpose

- Indexing or migrating, tagging, deleting, moving, copying, locking, encrypting, OCRing, etc.

DEALING WITH SHARED DRIVES



L&S THROUGH AZURE

Challenges:

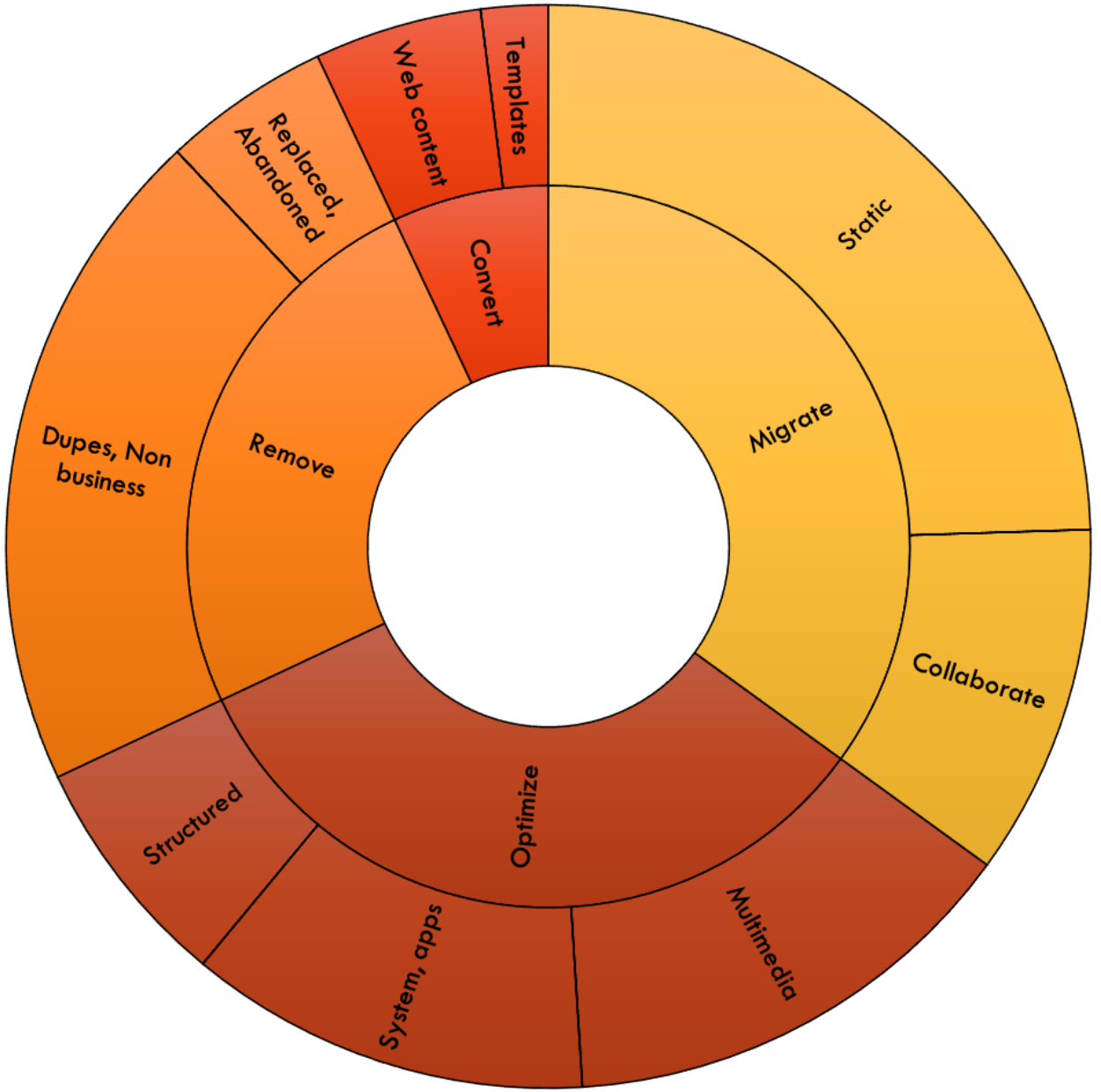
Need to review prior to commit

Time delays mess up change management

Broken links

Duplicate storage

Wasted time and resources on content that cannot/should not move



TRAINING CLASSIFIERS


1. Identify what type you want to classify


2. Build a classifier


3. Send it off to look at other content in scope


4. Examine results to see what need to be excluded


5. Iterate


6. Extract metadata

Challenges:

Function is different than topic

Function changes with time/value

Different spreadsheets are similar

Process can be cumbersome

Context and format aren't used

GRANULARITY

Team
Accuracy: High
Speed: High

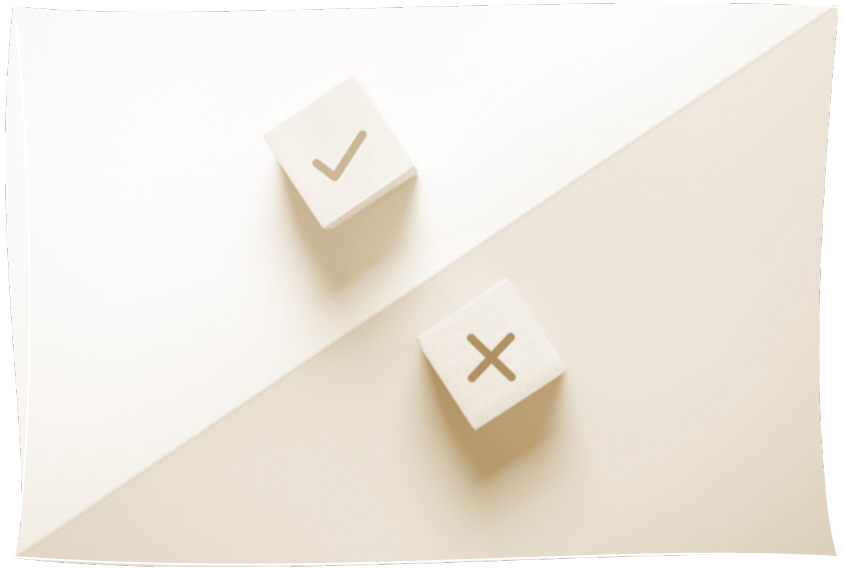
Work group
Accuracy: Varies
Speed: Medium

Organization
Accuracy: Overlapping
Speed: Low

FALSE POSITIVES

Ways to reduce false positives:

- Only look in the possible locations (start with a data map): pre-migrate index
- Judicial selection of classification criteria – Pre and post index
- Regex Validation – Pre index
- Risk Counting and Calculations – Pre and Post index
- Positive and negative qualifying queries – Post Index
- Leverage multiple properties and attributes – Post index
- Visualizations – Post Index
- Data relationships (use a specific person name)



LOGICAL KQL EXAMPLES

1. If it was based on a personnel action template, was saved by a manager in the HR department this past year, classify it as a 7 year retention
2. If it contains a valid contract number near the words “contract no.” and it is not in the IT department “sample data folder”, classify it as a Contract and apply the contract number as metadata
3. If it was authored by someone in marketing and it has one of our secret project names, mark it as sensitive and make sure it cannot be emailed
4. If it contains a 16-digit credit card number, but its format is a DLL system file, don’t classify it

VALIDATE TAXONOMY

Managing the rules – know what criteria and rules you need to comply with:

- Scope out the classification facets
- Identify criteria search and classification rules
- More rules is easier to troubleshoot that bigger rules
- Test against real data
- Clarify location and type
- Reduce false positives
- Iterate

Low	**Medium**	**High**	**Critical**
Garbage	Isolated	Undefined	Compliant
Duplicate	Un-owned	Potential	Valuable
Outdated	Non-text	Important	Risky

DEFINING GOALS AND TAXONOMIES

Records	~150 content type collections based on function and modified dates. Some types are sensitive, some will contain PII
Privacy	Potentially 2-300 (Depending on regulation and countries) personal identifiers in rough collections.
Divestitures	Split of company information assets by owning company
Health Regs	Lists of medical conditions
Payment Cards	Multiple card number and card identifying details with validation
Information Security	Custom internal definition ranging from public to confidential, to top secret often tied to subparts of documents
Business Value	Project, case, part, product, contract, matter, value

IGRM

Information Governance Reference Model (IGRM)

Linking duty + value to information asset = efficient, effective management

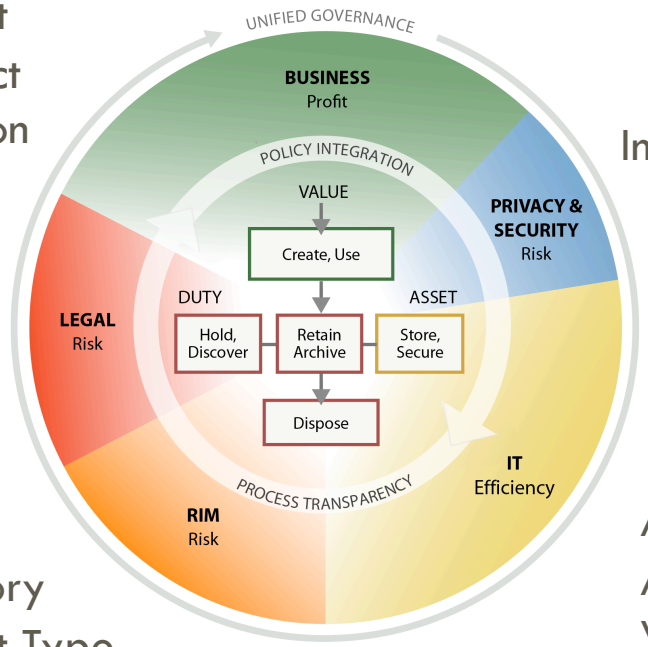
Project
Product
Function

Info Security

Case
Matter

Category
Content Type

Access
Age
Value



Duty: Legal obligation for specific information

Value: Utility or business purpose of specific information

Asset: Specific container of information

1

Choose big buckets for retention

- Classify “Accounting documents” rather than “Invoices, Credit Memos, Debit Memos, Adjustments”
- Classify low value, not just records – often as “everything else”

2

Adjust categories and processes to enhance (not replicate) a paper based classification

3

Avoid event-based triggers where possible

Use expected life +

4

Align classifications to other systems, processes

Leverage integration capabilities for event triggers

TO AUTOMATE RECORDS MANAGEMENT

TO PROTECT SENSITIVE DATA

Your most important detail about your most important asset

- Set policy, train, audit
- It is not primarily about removing it or keeping people away, it is about being a useful, responsible steward

Leverage overlapping taxonomies

Keep like-content together in deployments

Leverage regexes and compound queries for accuracy

Sensitive data is OK to have if it is in the right place

Most of it is in structured locations

https://www.sarahderemer.com/hybrid-animals

AUTO DATA-MINIMIZATION

Type	Definition	Example
Enterprise rules	Based on approved policy and approved rules	Temporary files, duplicates, expired records, unsubscribed
Departmental Review Rules	Based on rules, sampling for accuracy	Christmas party photos, expired programs, non business content
Sites, shares, groups	Activity and content rules	Project X team site, non-accessed – non sensitive
Individual Rules/Folder	Folder Name	App Development versions, Identified low value
Individual Files	File Name or Format	Large duplicate backups, Zip containers

For disposition review, granularity, accuracy, and process should be similar effort to paper files or other data collections.

FILES VERSUS EMAILS

Limitations – Classifying emails is different than file shares.

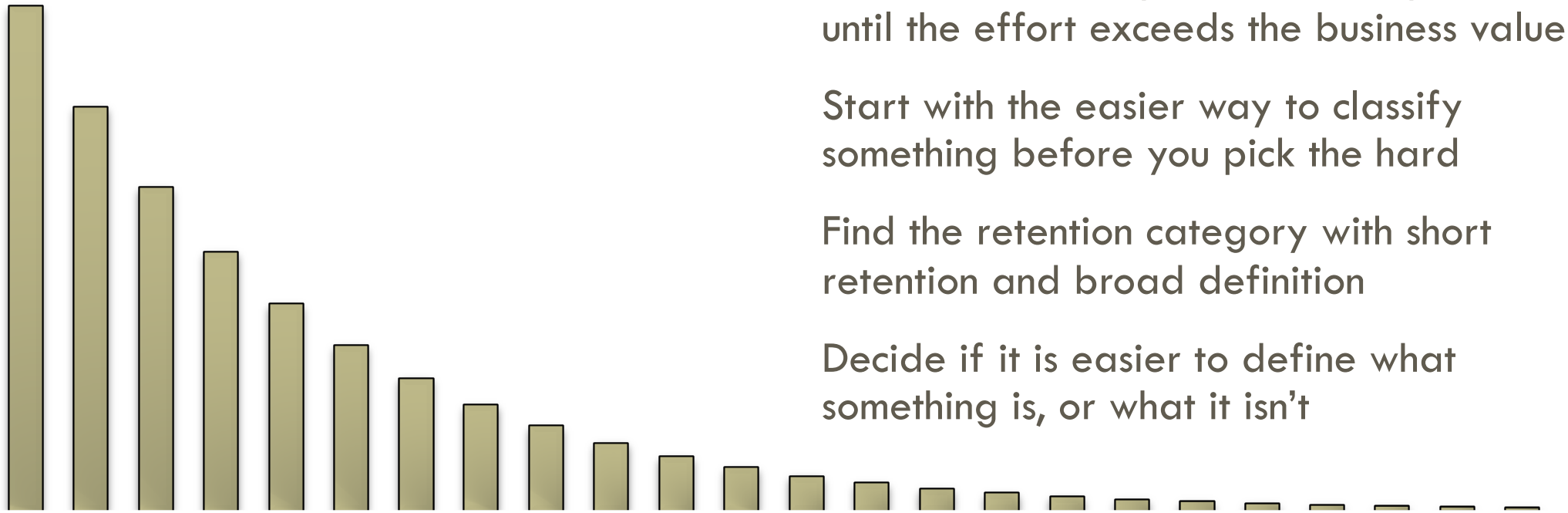
You lose:

- File formats (they are all emails)
- File folders (everybody's is unique)
- Usable date comparisons (there is only one date)
- Templates or forms
- Non-content (no applications, databases, web content, temp files)
- Context for attachments

You gain:

- A subject matter (providing consistency and topics)
- A sender and recipient (showing content and intent)
- Domains (who is involved in the discussion)

CATCHING THE SLOW FAT RABBIT



Start with the easy wins and keep working until the effort exceeds the business value

Start with the easier way to classify something before you pick the hard

Find the retention category with short retention and broad definition

Decide if it is easier to define what something is, or what it isn't

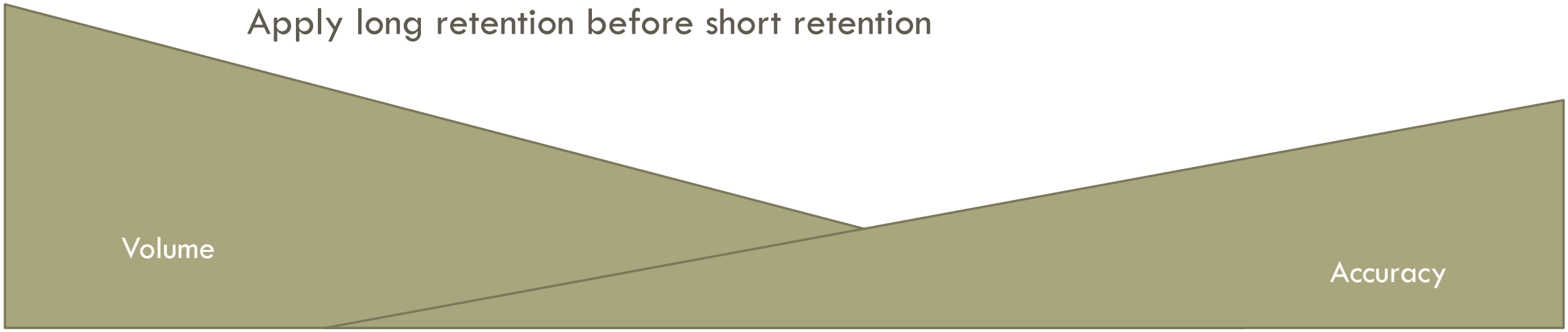
CLASSIFY OVERLAP FROM LOW TO HIGH

It is more effective to define what to get rid of and get rid of it before you then decide it matches a classification for being valuable

Example, the second, third, and forth copy of an important document WILL be auto classified as being important.

Throw away, then put away, then classify

Apply long retention before short retention



TEST, TEST, TEST

Make a copy of sample data that you know has positive examples in it to see what issues, problems, accuracy before you commit

Sample results before committing to labels

Approve and accept rules, not results

Autoclassification is 100% accurate. Your content might not be.

PRESERVE AND REUSE KNOWLEDGE

Classifying legal content as privileged, or responsive to a litigation case is very expensive process when done manually.

Sensitive data has value for protection, but also for data minimization

Transitioning to big bucket retention uses the same queries as traditional schedule

Mergers, divestitures and reorganizations are all easier with good classification

Re-use classifications if you can. (Don't just "find and exclude" but add persistent labels and metadata.)

GUIDING PRINCIPLES

Important or voluminous information has been foldered, named, nicknamed, acronymed, numbered, templated, or isolated

Leverage policy, communication, and SMEs for review to maximize benefit and minimize user impact

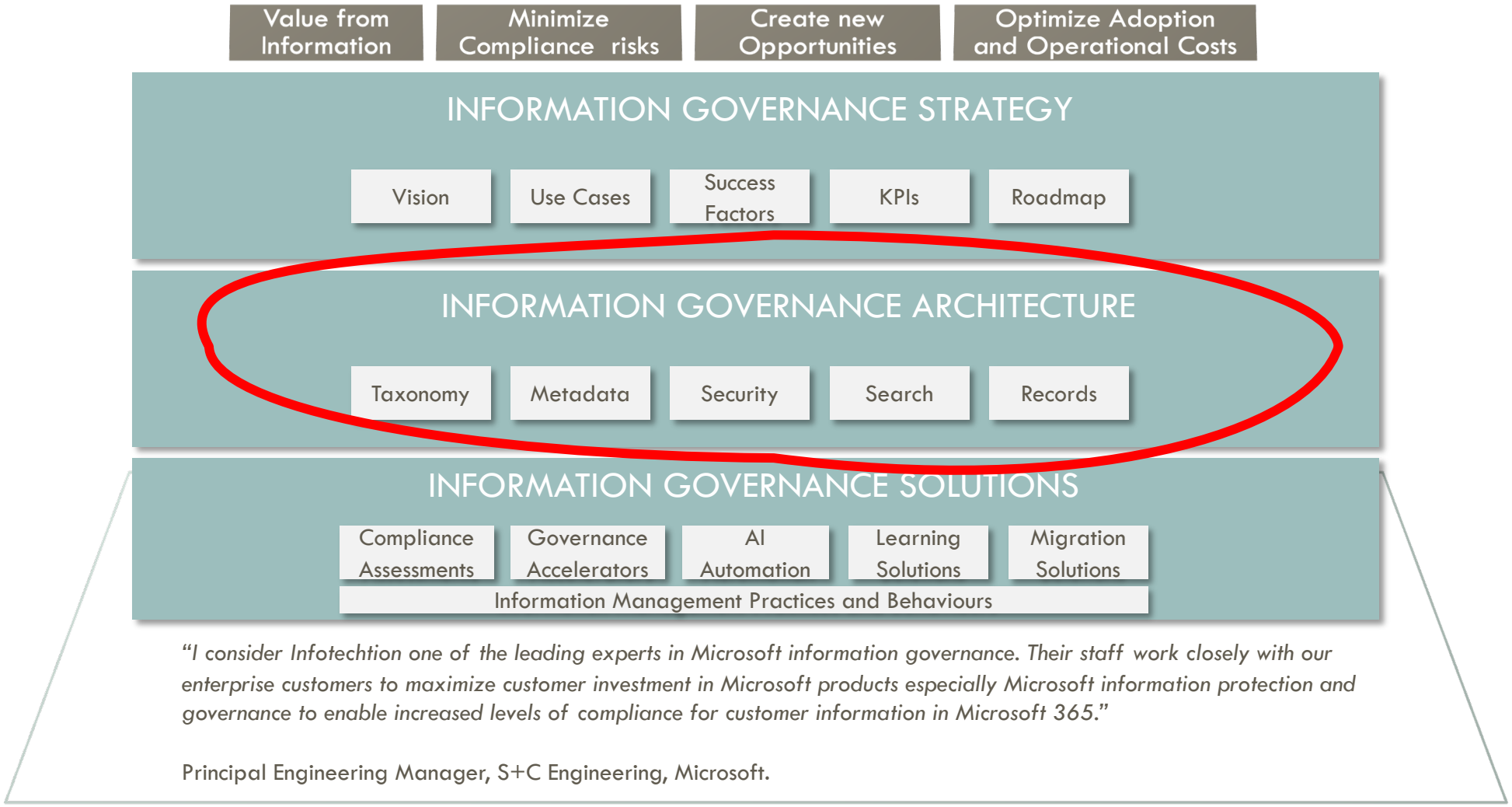
Consistent approach and less accuracy is better than individual interpretation

- "Don't let perfection be the enemy of good"

A semi-automated process, not just a tool

Throw away, put away, then organize so you don't spend time analyzing data that should not be there

INFOTECHTION SERVICES



"I consider Infotechtion one of the leading experts in Microsoft information governance. Their staff work closely with our enterprise customers to maximize customer investment in Microsoft products especially Microsoft information protection and governance to enable increased levels of compliance for customer information in Microsoft 365."

Principal Engineering Manager, S+C Engineering, Microsoft.

Infotechtion

John.M@infotechtion.com

Brian.T@Infotechtion.com

THANK YOU