Scalable Document Classification

Malek Ben Salem, PhD
Malek.ben.salem@accenture.com
@MalekCyber
Protecting Big Data adequately is a creeping problem for organizations.

40-50% CAGR

$250 billion/year
The traditional way of protecting data focuses on control:

**Networks**
We have locked the data away in the data center

**Devices**
We have enabled AES-256 encryption on all mobiles and encrypted the full disks on Windows

**Apps**
Everyone uses our container solutions

**Services**
We only give authorized people access to the application
Need for Data-Centric Security

The traditional way of protecting data focuses on control:

- **Networks**: We have locked the data away in the data center.
- **Devices**: We have enabled AES-256 encryption on all mobiles and encrypted the full disks on Windows.
- **Apps**: Everyone uses our container solutions.
- **Services**: We only give authorized people access to the application.

Identity and Access  Data-Centric Security
DATA-CENTRIC SECURITY STARTS WITH INFORMATION CLASSIFICATION. ACCURATE AND SCALABLE INFORMATION CLASSIFICATION IS CRITICAL TO PROTECT AN ORGANIZATION’S MOST VALUABLE ASSETS.
Content producers have to label files that they are authoring.
Text Document Classification using ML and NLP

New/Modified Document → Generate N-Gram Profile → Document Profile

Category Profiles → Generate N-Gram Profile → Measure Profile Distance

Profile Distances → Find Minimum Distance and Match

Sensitive? → Sensitive?

Classify Data → Use Encryption → Apply Access Control

Allow Access

Category Samples → Generate N-Gram Profile → Measure Profile Distance

Profile Distances → Find Minimum Distance and Match

Sensitive? → Sensitive?

Classify Data → Use Encryption → Apply Access Control

Allow Access
Use data classification to enable security controls

- Access Controls
- Data Access Monitoring
- Enterprise Data Rights Management
- Multi-Level Security
- Encryption
- Data Loss Prevention (DLP)
- Data Masking
- Data Tokenization
USE CASES AND BUSINESS VALUE

LOW INFORMATION RISK
OFFLINE/LEGACY CLASSIFICATION

DATA PROTECTION
ONLINE/REAL-TIME CLASSIFICATION

COMPLIANCE
CLASSIFICATION SCHEME AUDITING

DATA INSIGHTS
CLASSIFICATION POLICY INFERENCE
Value Proposition

• Use information classification to reduce information risk and IT costs.

INFORMATION RISK

Use your controls effectively

- Enhance Access Controls
- Implement Security Controls
- Institute Reclassification Practices
- Decrease Storage Costs

IT OPERATIONS

- Improve IT Security Spend and Reduce Storage Costs
- Ensure Compliance
- Incorporate Retention Schedule
Scalable Document Classification

Malek Ben Salem, PhD
Malek.ben.salem@accenture.com
@MalekCyber