



Education

# **Information Classification: The Cornerstone to Information Management**

Sheila Childs, EMC Corporation

# SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced without modification
  - ◆ The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.

*The key to managing information is to understand it. Given the huge volumes of information that organizations are dealing with, that understanding requires new classification technology. Being able to classify unstructured data as well as more structured data types greatly expands information's value to business, and allows companies to cost-effectively meet the demands of retention requirements, discovery, and swelling storage needs. As a result, businesses can make low-cost management possible for all levels of information. This presentation will demonstrate how classification can help organizations that are experiencing explosive growth in unstructured information.*

# What's Driving The Need for Classification *TODAY*

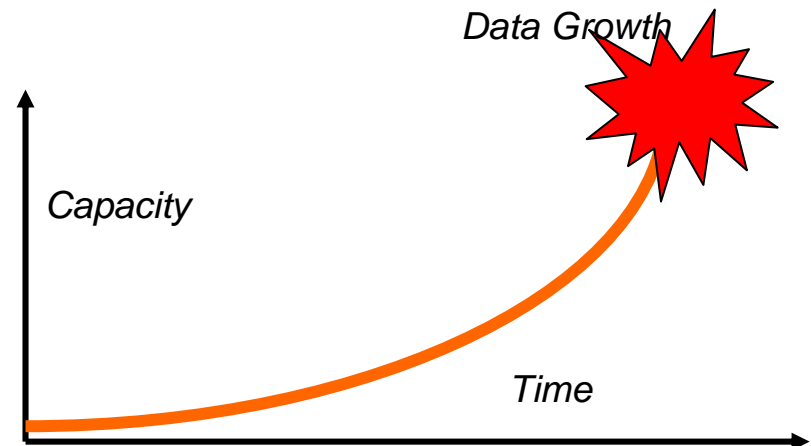
Corporations are *saving everything*, because

- ◆ They are unsure about the value of their information
- ◆ They are being litigated
- ◆ They are complying with government regulations

Resulting in

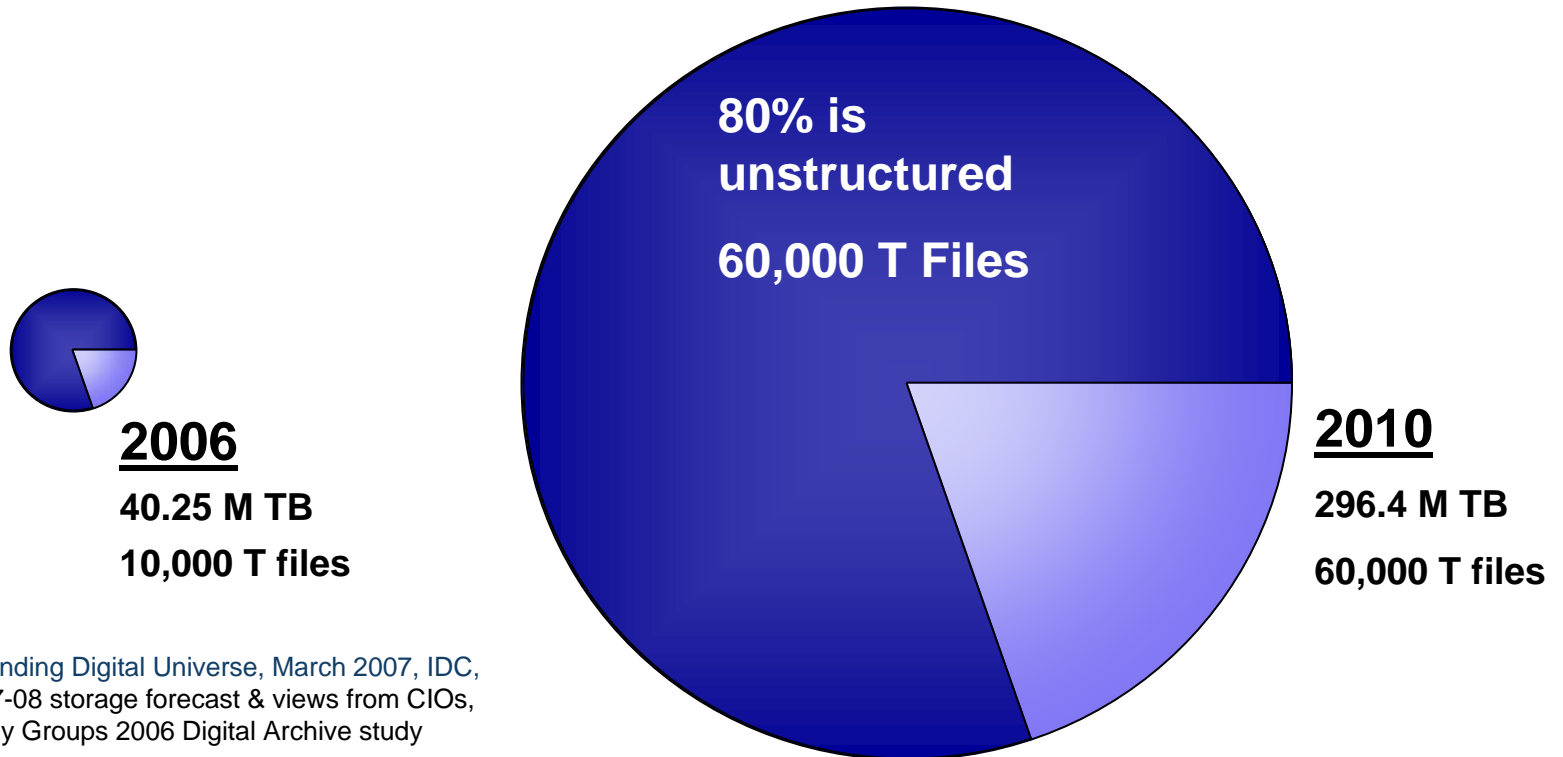
- ◆ Massive amounts of information growing at fantastic rates
- ◆ Information security breaches
- ◆ Lots of money being spent for governance and compliance

Corporations are balancing *IT Infrastructure and Management Costs* against *Information Risk Management*



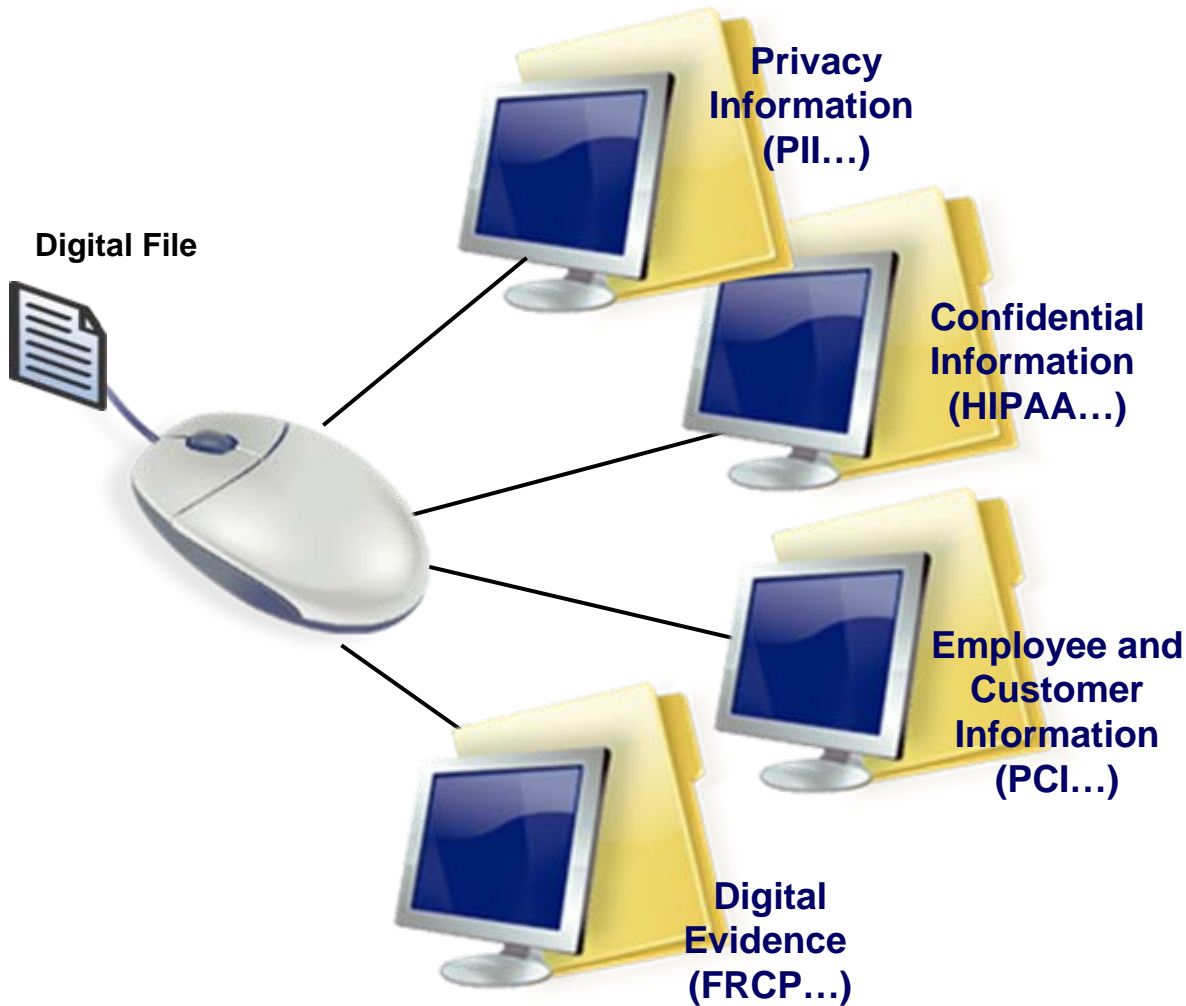
## Storage TCO

- External disk storage purchase projected to grow at 52% annually
- Capacity is #1 storage issued driven by email, unstructured data
- Significant transition to disk-based archival storage
- Digital archive capacity will increase nearly tenfold between 2005 and 2010



Source: The Expanding Digital Universe, March 2007, IDC, Merrill Lynch 2007-08 storage forecast & views from CIOs, Enterprise Strategy Groups 2006 Digital Archive study

# Information Risk



## Inappropriate Access

Law suits, jail time, penalties, reputation, employee morale, market share

## Inability to Find & Produce

Productivity losses, legal fines, loss of legal cases, jail time

## Inability to Retain

Federal penalties, law suits, business effects

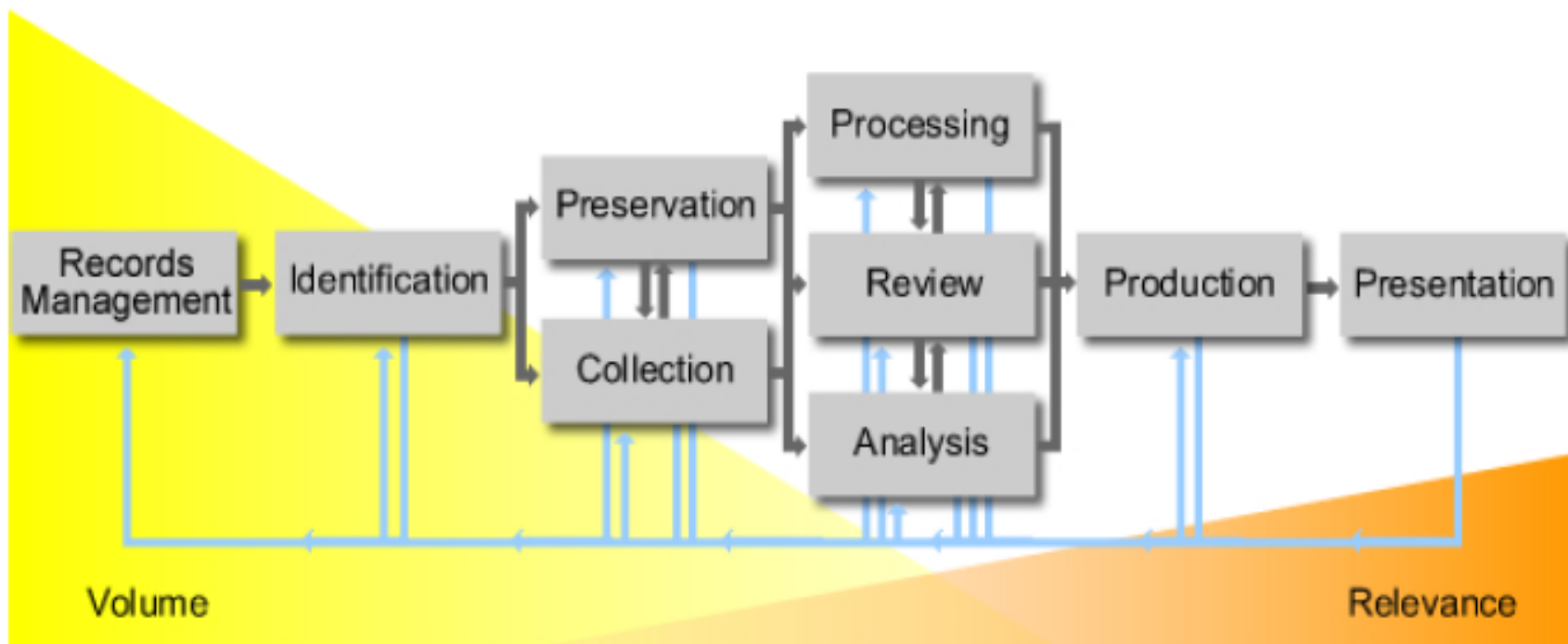
## ➤ eDiscovery and records management coming together

- ◆ Driven by huge costs and risks
- ◆ Changes to the Federal Rules of Civil Procedure
  - › Electronically Stored Information (**ESI**) is subject to production (the way it is managed from cradle to grave will impact costs and risks of eDiscovery)
  - › There will be an early “**meet and confer**”
  - › Word “**preserving**” appears in the rules for the first time
  - › There is a need to understand the “**sources**” of ESI
- ◆ Average eDiscovery costs can run into the millions of dollars per event



# eDiscovery Value Chain

- Driven by high cost of litigation support and new Fed. Rules of Civil Procedure (Rule 26)
- IDC estimates eDiscovery at \$1.6BN market in 2006, growing to **\$4.2BN in 2007**



Source: EDRM.net

## Improved productivity

- The average knowledge worker spends *six hours per week* searching for information
  - ◆ 50% of all searches fail to locate desired information
  - ◆ 15% of the average knowledge worker's time is spent recreating existing information
  
- **Need**
  - ◆ Better organization of information
  - ◆ Accurate search
  - ◆ Consistent management of information
  - ◆ Shortened “time-to-information”



# A Brief History of Information Classification

660 B.C. Archives and library are organized by King Ashurbanipal in Nineveh, marking the first systematically organized library of the ancient Middle East. Some 20,000 tablets from it survive today.

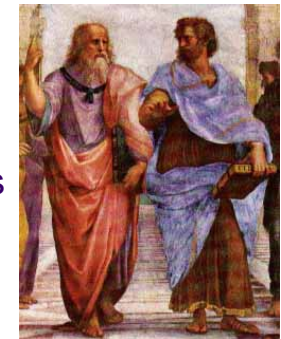


620 B.C. Wei Cheng writes the bibliographic section of the official Sui Dynasty History, dividing the books into four categories: Confucian classics, historical records, philosophical writings, and miscellaneous works.

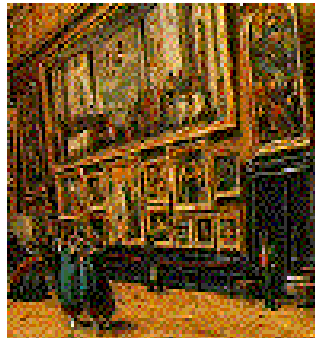


On the Parts of Animals  
 By Aristotle  
 Written 350 B.C.

The earliest known system of classification is that of Aristotle, who attempted to group animals according to such criteria as mode of reproduction and possession or lack of red blood.



1373 The first inventory is made of the massive book collection housed in the Louvre, which later forms the basis for the Bibliotheque Nationale.



Source: Geoffrey Nunberg, Timeline of the History of Information, <http://www.ischool.berkeley.edu>

# A Brief History of Information Classification

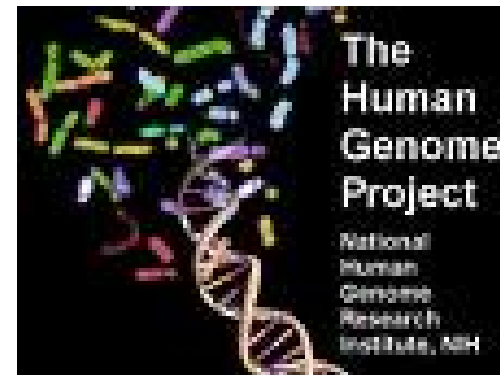


1735 Carolus Linnaeus “The Father of Taxonomy”, offers first systematic organizational schema to understand the variety of life in the natural order, which is the basis of taxonomical nomenclature.

1876 Melvil Dewey outlines the classificatory system later known as Dewey Decimal.



1990 Start of the project of mapping the location of all genes on every chromosome in human beings, the Human Genome Project.



Source: Geoffrey Nunberg, Timeline of the History of Information, <http://www.ischool.berkeley.edu>

# Today: Classification Has Moved Into the Corporation – Education

## Who is Responsible?

### ➤ Traditionally – Records Information Managers

- ◆ Records retention for regulatory compliance
- ◆ Coordinated records management across the enterprise

### ➤ Now – Collaborative effort

- ◆ **Information Technology:** provides service delivery for ESI
- ◆ **Information-centric support**
  - › Information Security
  - › Litigation support (Legal)
  - › Corporate Governance
- ◆ **Line of Business**
  - › Responsible for the business of the corporation

Role	Risk Management			TCO
	Security and Risk Mitigation	Litigation Support	Records Mgmt	Cost Mgmt
Chief Security Officer	X			
Chief Legal Officer	x	X	X	
Corporate Counsel		X	X	
Records Managers		X	X	
Chief Compliance Officer	X	X	X	
Chief Risk Officer	X			
Chief Financial Officer	X			X
Line of Business	X	X	X	X
Chief Information Officer (I.T.)	X	X	X	X

# The Dilemma: At The Crossroads

## ➤ Multiple Groups Converging

- ◆ Records managers have certain requirements
- ◆ IT has different/additional requirements
- ◆ Security has different/additional requirements

## ➤ Coming Together

- ◆ Corporations are beginning to establish cross-functional groups that will be responsible for information management
- ◆ Starts with understand of requirements for corporate information
- ◆ **Classification** is the first step

## ➤ Organizations report - *this is difficult*

390 B.C. Augustine frames a system which dominates the structure of encyclopedias. It is based on the ordering of human knowledge of the world and human customs as they pertain to salvation.



# The Benefits of Coming Together: Example #1

Legal, Records Manager, I.T collaborate to save money by employing automated classification of eRecords of Exited Employees

<i>Case Study Measurements and Assumptions</i>	<i>Results</i>
Average number of electronic files per exited employee	601
Average number of emails per exited employee	7418
Average time to manually classify each document	2.25 minutes
Number of documents manually classified per day	180
Annual Salary of a qualified records technician (loaded)	\$55,000
Cost to manually classify per exited employee	\$12,473
Cost to outsource the autotclassification of documents	\$800-1200 per exited employee

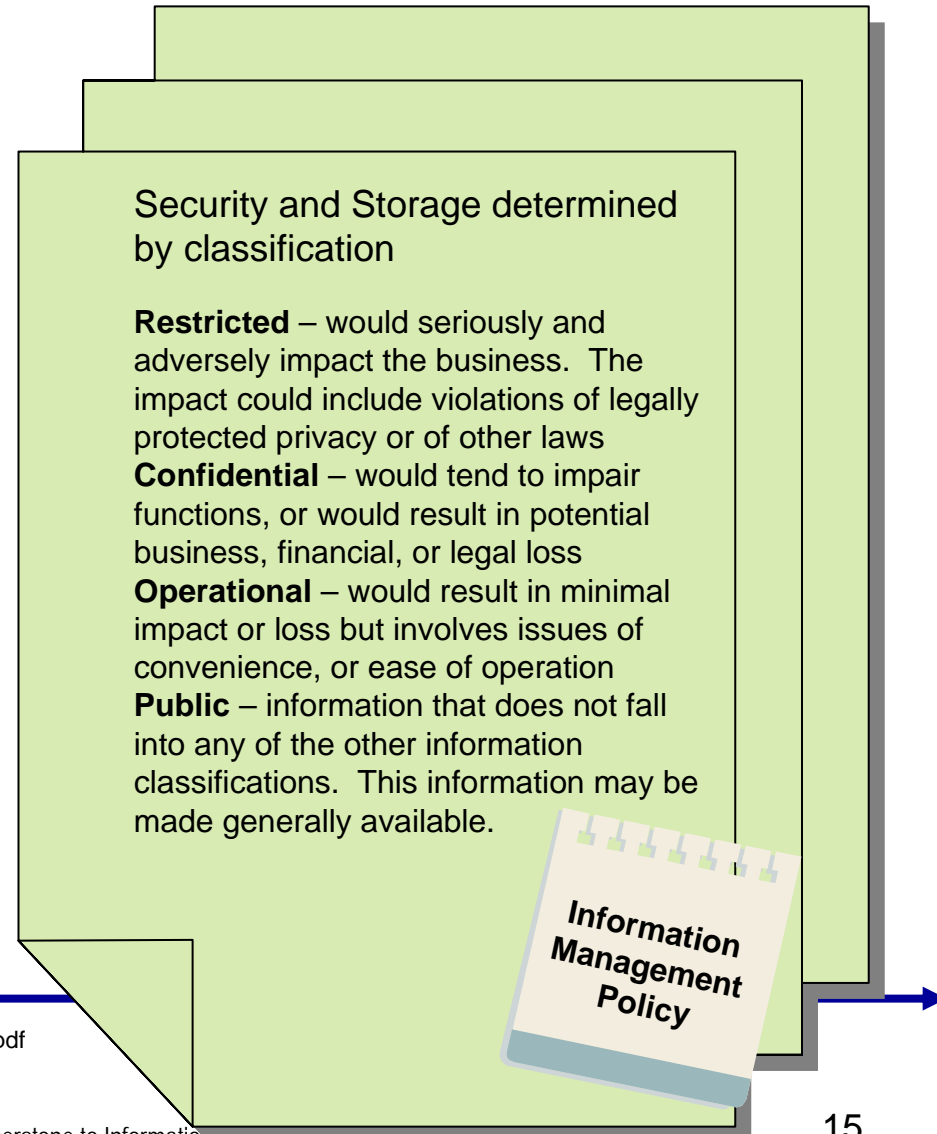
Source: ARMA 2005 International Conference  
 "Classifying e-Records of Exited Employees: Case Study Using an autoe Classification Tool", Session Number T027

# The Benefits of Coming Together: Example #2

Hospital establishes cross-functional group that will be responsible for information management

➤ Information is *Restricted, Confidential, Operational, Public*

- ◆ *How important is it to keep this information confidential?*
- ◆ *How important is it's integrity?*
- ◆ *How important is it's availability?*



Source: <http://www.psychiatry.ufl.edu/security/docs/InformationClassificationDept.pdf>

# The Benefits of Coming Together: Example #2

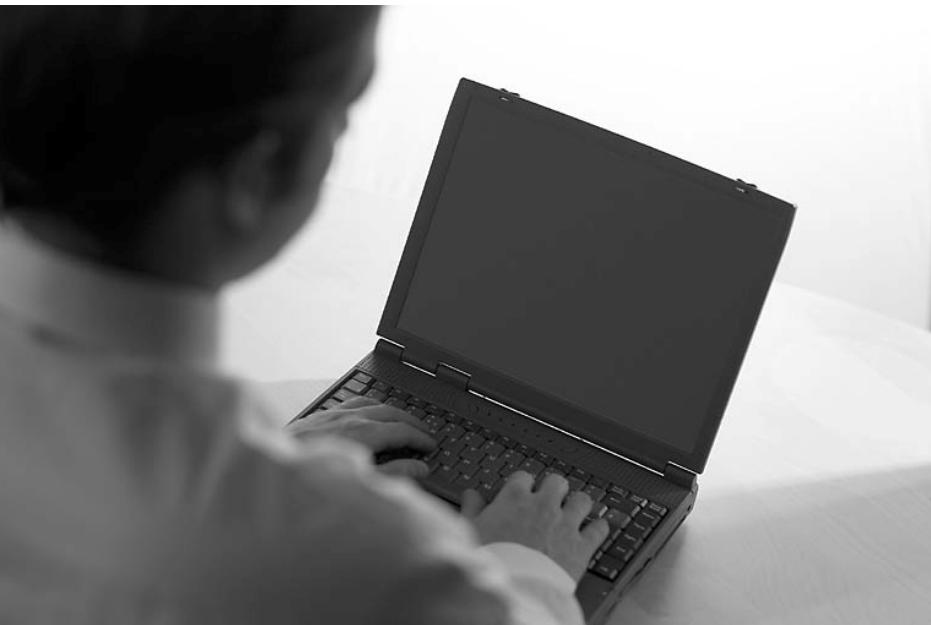
Information Type	Confidentiality	Integrity	Availability	Classification
Medical record	High	High	High	Restricted
Administrative Documents with Private Health Information (PHI)	High	Medium	Medium	Restricted
Email with PHI	High	High	High	Restricted
Patient claims or billing information with PHI	High	Medium	High	Restricted
Research information with PHI	High	High	Medium	Confidential
Budget Information	Medium	Medium	Medium	Confidential
Financial Reports	Low	High	Low	Confidential
Personnel and fiscal operations records without PII	Medium	Medium	Medium	Confidential
Research Proposals	Medium	High	Medium	Operational
Memos without PII	Low	Medium	Medium	Operational



Source: <http://www.psychiatry.ufl.edu/security/docs/InformationClassificationDept.pdf>

➤ **FILE ATTRIBUTE** classification is largely based on file attributes and **access patterns**

- What is file named?
- What is the file type?
- Who owns the data?
- Where is it located?
- When was it created?



- Classification based on **CONTENT** makes use of *indexes, lexicons and taxonomies*



- What keywords?
- How is this data related to other data?
- How should data be retained/disposed of for compliance or otherwise used by the business?



## ➤ **Taxonomy**

- ◆ A hierarchical structure used for categorizing a body of information or knowledge, allowing an understanding of how that body of knowledge can be broken down into parts, and how its various parts relate to each other. Taxonomies are used to organize information in systems, therefore helping users to find it
  - Related terms: ontology, categories, evidence structures

## ➤ **Folk Taxonomy**

- ◆ A taxonomy that expresses a part of the groups' social knowledge and has real meaning to those who use it. They evolve, and are generated and communicated as a result of a social or group dynamic. They are used in everyday speech

## ➤ **Folksonomy**

- ◆ A social, cultural and tag-based taxonomy. User driven approach to organizing information, with no real structure. The meaning of the words are specific to the author who writes the tag

1086 William the Conqueror undertakes the first complete government census of land, possessions, and inhabitants, leading to the establishment of public archive.



# Classification Terminology: Semantic Definitions

- **Semantics** is devoted to the study of meaning, utilizing syntactic levels of words, phrases, sentences and paragraphs
  - ◆ **Homonymy:** a group of words that share the same spelling or pronunciation (or both) but have different meanings
    - › Examples: *tramp*, and the words *dear and deer*
  - ◆ **Synonymy:** Different words with similar or identical meanings
    - › Example: *forest and woods*
    - › Example: *long time* and an *extended time*, *long* and *extended* become synonyms.
  - ◆ **Antonyms:** Word pairs that are opposite in meaning
    - › Examples: hot and cold, *tall and short*
  - ◆ **Hypernymy** is all encompassing term. "Word A" is a hypernym of "Word B" if A's meaning encompasses the meaning of B.
    - › Example: *vehicle* denotes all the things that are separately denoted by the words train, chariot, dogsled, airplane and automobile and is therefore a hypernym of each of those words.
  - ◆ **Meronymy:** denotes a constituent part of, or a member of something.
    - › Example: 'bathroom' is a meronym of 'house' because a bathroom is part of a house



- All content-based classification is based on “natural language”
- Two general types:

## Rules-based content classification algorithms

- ◆ Conceptual and Semantic Analysis
- ◆ Keywords
- ◆ Term frequency
- ◆ Pattern matching
- ◆ Stemming
- ◆ Compound terms
- ◆ Semantic analysis

## “Learning”-based content classification algorithms

- ◆ Neural Networks
- ◆ Probabilistic modeling
- ◆ Bayesian Inference
- ◆ Shannon’s Information Theory



## **Automated Classification speeds time-to-information effectiveness**

### **Automated Content Classification make sense**

- ◆ When multiple classification options results in confusion
- ◆ When there is an overwhelming volume of items to classify
- ◆ When some documents require time-consuming review by subject matter experts
- ◆ When there are a large number of non-business documents
- ◆ When you don't want to have idiosyncratic results

“The highest quality and accuracy occurs when records management is as non-intrusive as possible to the desktop end users and does not interfere with the normal work routines of professional staff in the enterprise”\*\*

\*\* Timothy J. Sprehe and Charles R. McClure, “Lifting the Burden.”  
*Information Management Journal*, Vol.39 Issue 4 (Jul/Aug 2005), 475

1555 Conrad Gesner, Swiss naturalist, completes his *Bibliotheca Universalis*, a classification of all past and present writers.



# Challenges of Content Classification: What is “Good enough?”

## ***Challenges of classification – various types***

- Some human intervention always required to review results of classification
  - ◆ Automated tools improve efficiency
- Documents with little text – how are these classified?
  - ◆ Pictures, Music, Power point slides, email, etc.
  - ◆ Metadata classification might be better in this case
- Lack of consistency in naming, structure, format
  - ◆ Metadata classification may be best

## ***Factors affecting accuracy***

- Document consistency / naming consistency
- The strength of the taxonomy (content)
- Applicability of classification algorithms to specific content

What is a reasonable cost per document?  
What is the cost, or risk, if a document is incorrectly classified?  
*Does value to the organization outweigh the cost?*

*Millions of Files*

File Attribute Classification

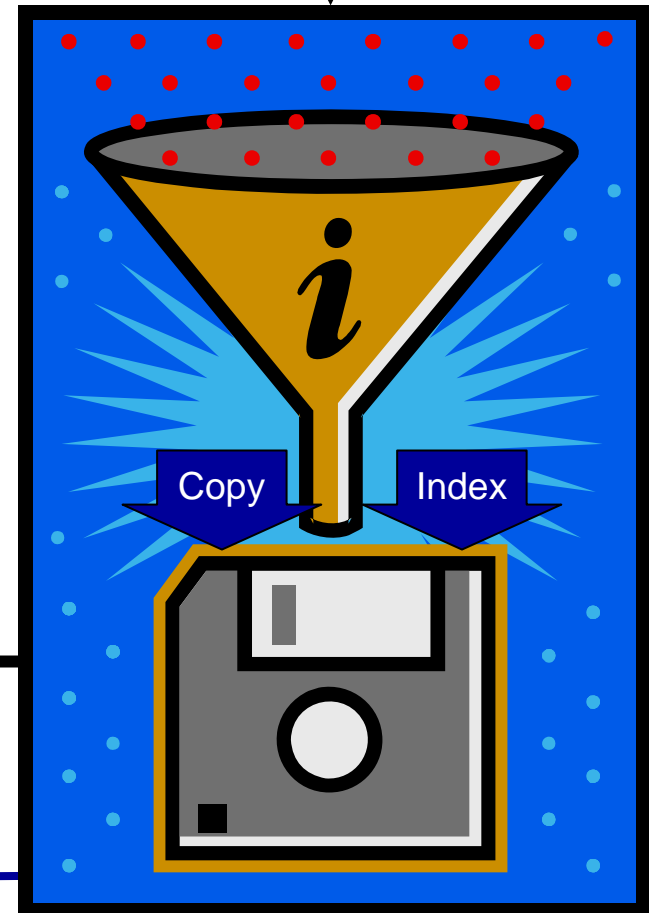
Group/Owner Classification

Content Classification

"acquisition"

*Relevant Files*

Legal Matter  
Repository



# Use Case #2: Classification for Security and Privacy

## Pattern Matching

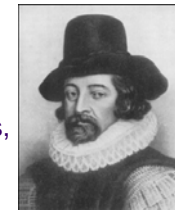


- ▶ **Personally identifiable/identifying information (PII)**
  - ◆ **Any** piece of information which can potentially be used to uniquely identify, contact, locate, stalk or steal an identity
- ▶ **Payment Card Industry (PCI) Data Security Standard**
  - ◆ Protects PII: Forbids retailers from storing credit and debit card data on point-of-sale systems

## **Classification** becomes the cornerstone for identification of PII

- ◆ Example: Major bank
  - › Formal security policies include identifying categories of information: “bank-confidential”, “bank customer confidential”, “other”
  - › New patterns include customer credit card numbers per PCI standards

1620 Francis Bacon's *Great Instauration* published, the first comprehensive plan for organizing knowledge around the human sciences, separating external nature from man.



# Use case #2: Classification For Security and Privacy

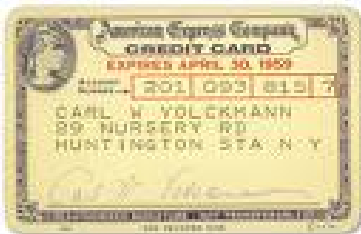
## ➤ How to find content that contains credit card numbers?

- ◆ *Automated Classification can help*

pattern template is equal to ▼ Please select... ▼

Please select...

- Social Security numbers with dashes (123-45-6788)
- Social Security numbers with dashes or spaces (123-45-6788, 123 45 6789)
- Diners Club/Carte Blanche Card Numbers: 14 digits with optional spaces or dashes
- American Express Card Numbers: 15 digits with optional spaces or dashes
- VISA Card Numbers: 13 digits with no spaces or dashes
- VISA Card Numbers: 16 digits with optional spaces or dashes
- MasterCard Numbers: 16 digits with optional spaces or dashes
- Discover Card Numbers: 16 digits with optional spaces or dashes
- Email address (someone@someaddress.com)
- Date in the form of MM/DD/YYYY
- Http based internet address (http://www.emc.com)
- 5 or 5+4 US ZIP code (27709)
- US phone numbers in the form of (999) 999-1234**



1923 Credit Cards are introduced in United States at hotels and gasoline stations.

- Classification is the First Step
  - What comes next?
    - ◆ *Alignment of classified information to IT infrastructure*
    - ◆ *Policy-based management*
- Define policies for classified information based on business requirements – *Service level objectives*
    - ◆ “Is my data receiving the service levels I pay for?”
    - ◆ “Am I compliant with regulation *xyzy*?”
    - ◆ “Is there any PII data where it shouldn’t be?”
    - ◆ “Is my information recoverable in the event of a disaster?”
    - ◆ “What data is under litigation hold – and until when?”

- **Classify for Records Management**
  - ◆ Policy: copy classified information to Enterprise Records Management Repository
- **Classify for Security**
  - ◆ Policy: Encrypt sensitive data
- **Classify for eDiscovery**
  - ◆ Policy: move information to legal matter repositories to be ready for litigation and to meet FRCP requirements
- **Classify for TCO**
  - ◆ Policy: archive classified information to less expensive storage; identify and delete duplicate files



1885 Francis Galton introduces the first system for classifying fingerprints, after proving that each set is unique

## ➤ What should you do next?

- ◆ Establish cross-functional group that will be responsible for information management
- ◆ Understand your requirements for classification based on risk and cost
- ◆ Begin to evaluate classification tools and technologies – get help if needed!

*Information Classification – done well - provides the foundation for cost-effective information risk management and IT infrastructure cost reduction*

*Corporations CAN balance IT Infrastructure and Management Costs against Information Risk Management*

- Please send any questions or comments on this presentation to SNIA: [trackdatamgmt@snia.org](mailto:trackdatamgmt@snia.org)

**Many thanks to the following individuals  
for their contributions to this tutorial.**

*SNIA Education Committee*

**Rob Peglar  
Bob Rogers  
Edgar St. Pierre  
Jeff Porter**