



Education

# **Active Archive - Data Protection for the Modern Data Center**

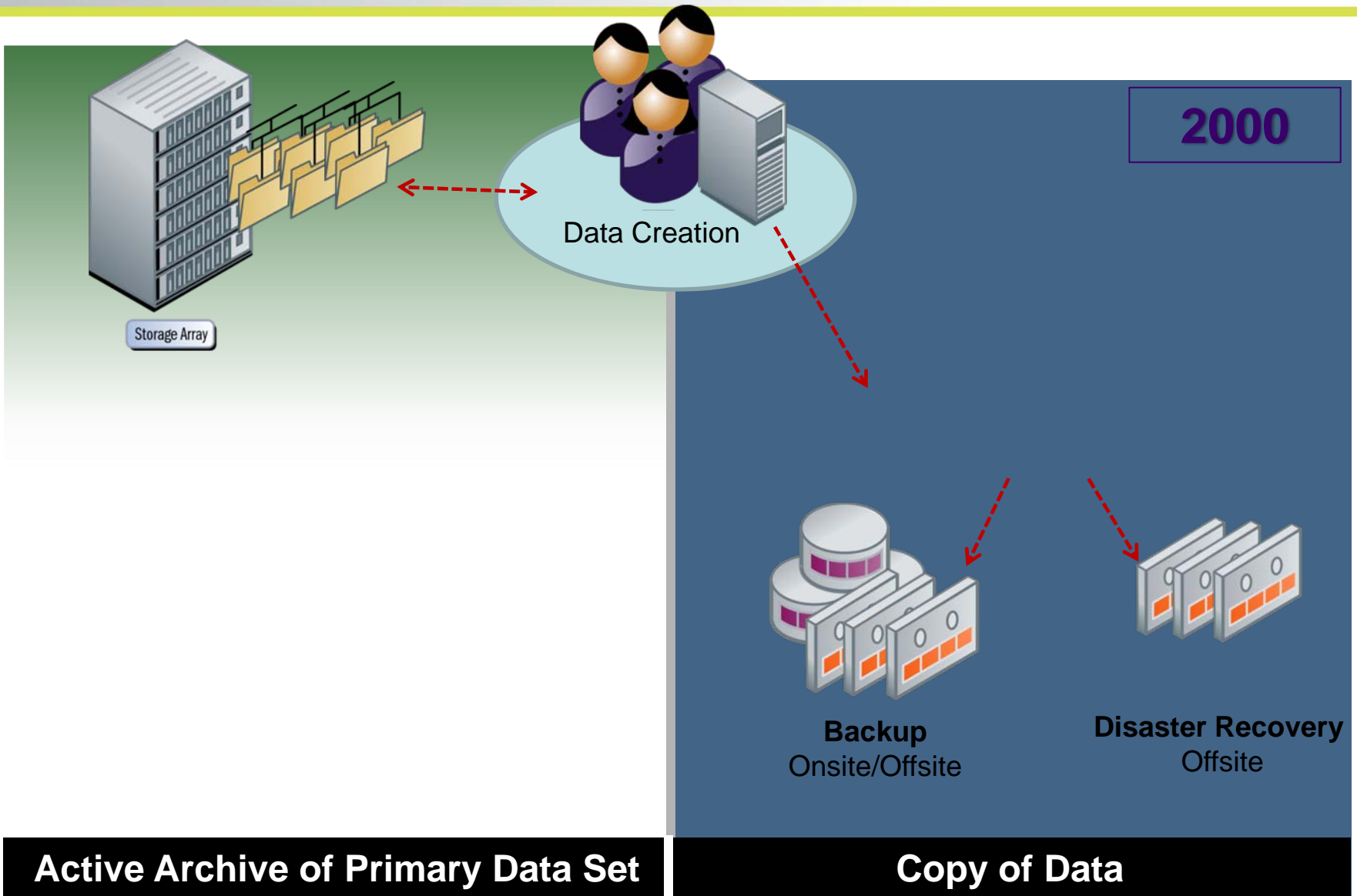
Molly Rector, Spectra Logic  
Dr. Rainer Pollak, DataGlobal

- The material contained in this tutorial is copyrighted by the SNIA.
  - Member companies and individual members may use this material in presentations and literature under the following conditions:
    - ◆ Any slide or slides used must be reproduced in their entirety without modification
    - ◆ The SNIA must be acknowledged as the 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.
  - Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
  - The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.
- NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.**

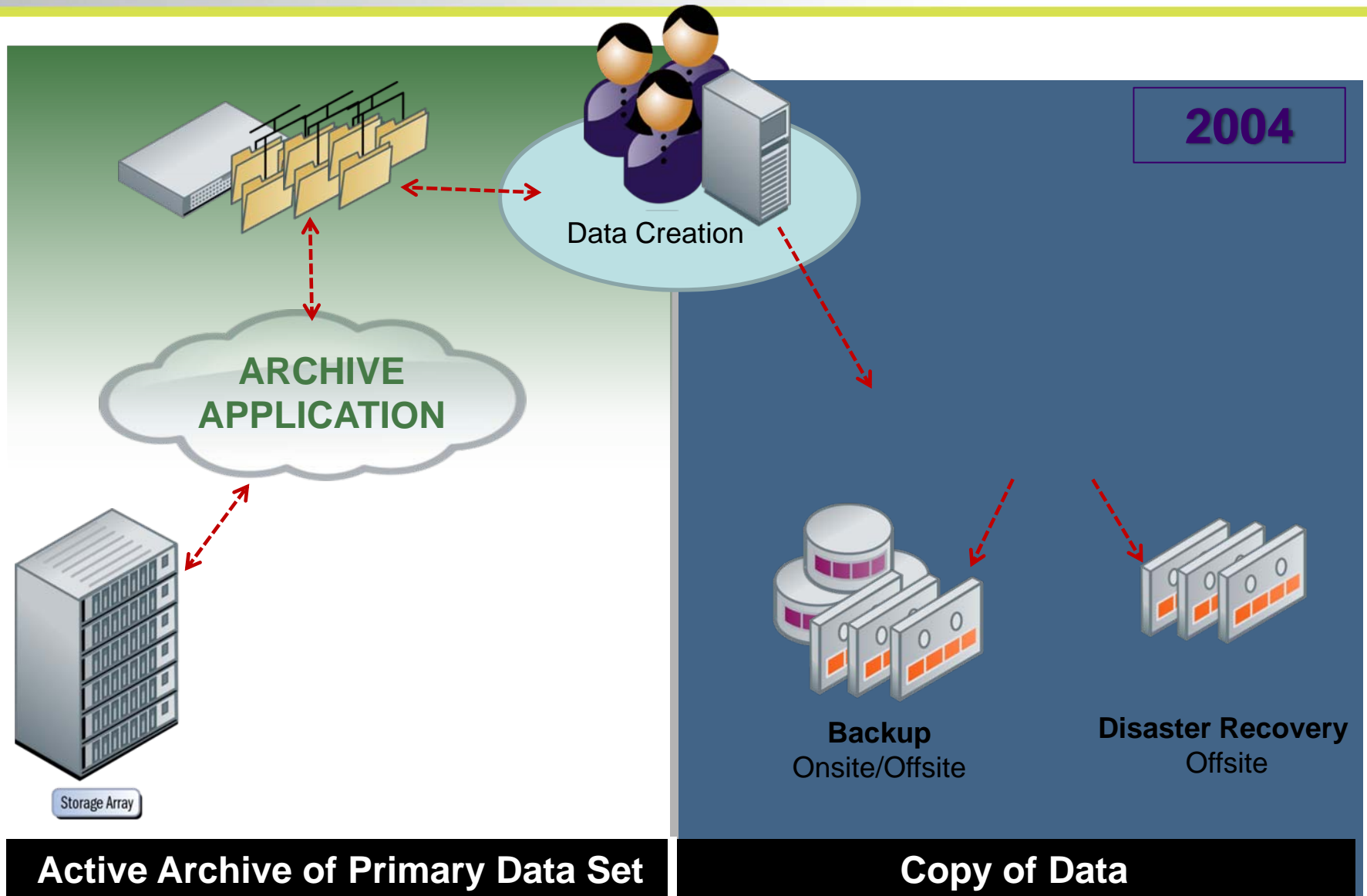
## Active Archive - Data Protection for the Modern Data Center

- Backup has long been thought of as a key technology for data protection. However, in today's modern data centers, the fastest moving and most fundamental changes are actually in archive technologies which enable affordable, online, long-term data access. Today's presentation focuses on key technologies used to manage data archives and how data classification can enable a fast and effective archival strategy. By focusing on managing the unstructured data within the data center, and using the right storage platforms designed to support the needs of the active archive, today's companies can gain unprecedented insight into their data to reduce the cost of storing, maintaining, and managing data while mitigating data leakage risks.

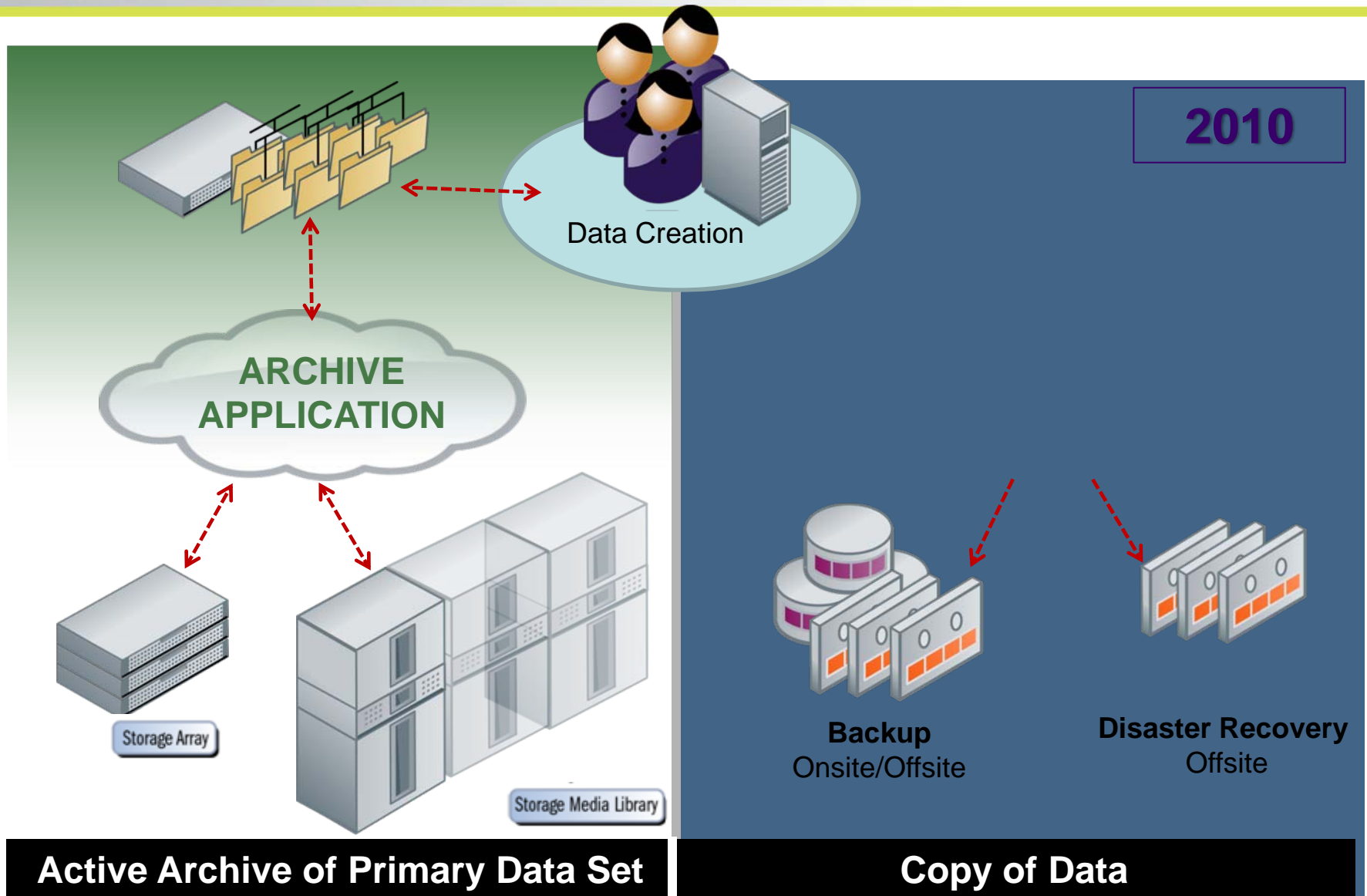
# The Development of Active Archive



# The Development of Active Archive



# The Development of Active Archive



**Active Archive of Primary Data Set**

**Copy of Data**

## ➤ Active archive

- ◆ A set of unstructured data such as office files and documents, video/audio files, email PST files and CAD/CAM files, that contains production data, no matter how old or infrequently accessed, that can be accessed online.
  - > Fueled initially by introduction of high density, lower power disk drives
  - > Momentum continued to build with release of power efficient disk arrays and high density, lower power disk drives.
  - > Next generation archives can leverage the latest in automated tape technologies offering high density, low power, cost effective archive storage

# Why Now?

- The market is at a tipping point with:
  - New application development and availability
  - Economic conditions
  - Explosive data growth
- Tape addresses all the needs of active archive:
  - Must integrate easily with archive management application making it simple access data (regardless of the storage type)
  - Power efficient
  - Very High Density
  - Reliable (built in data integrity validation)
  - Low cost
  - Fast throughput



Digital data being produced had risen to 281 exabytes (EB,  $10^{18}$ ) in 2007.

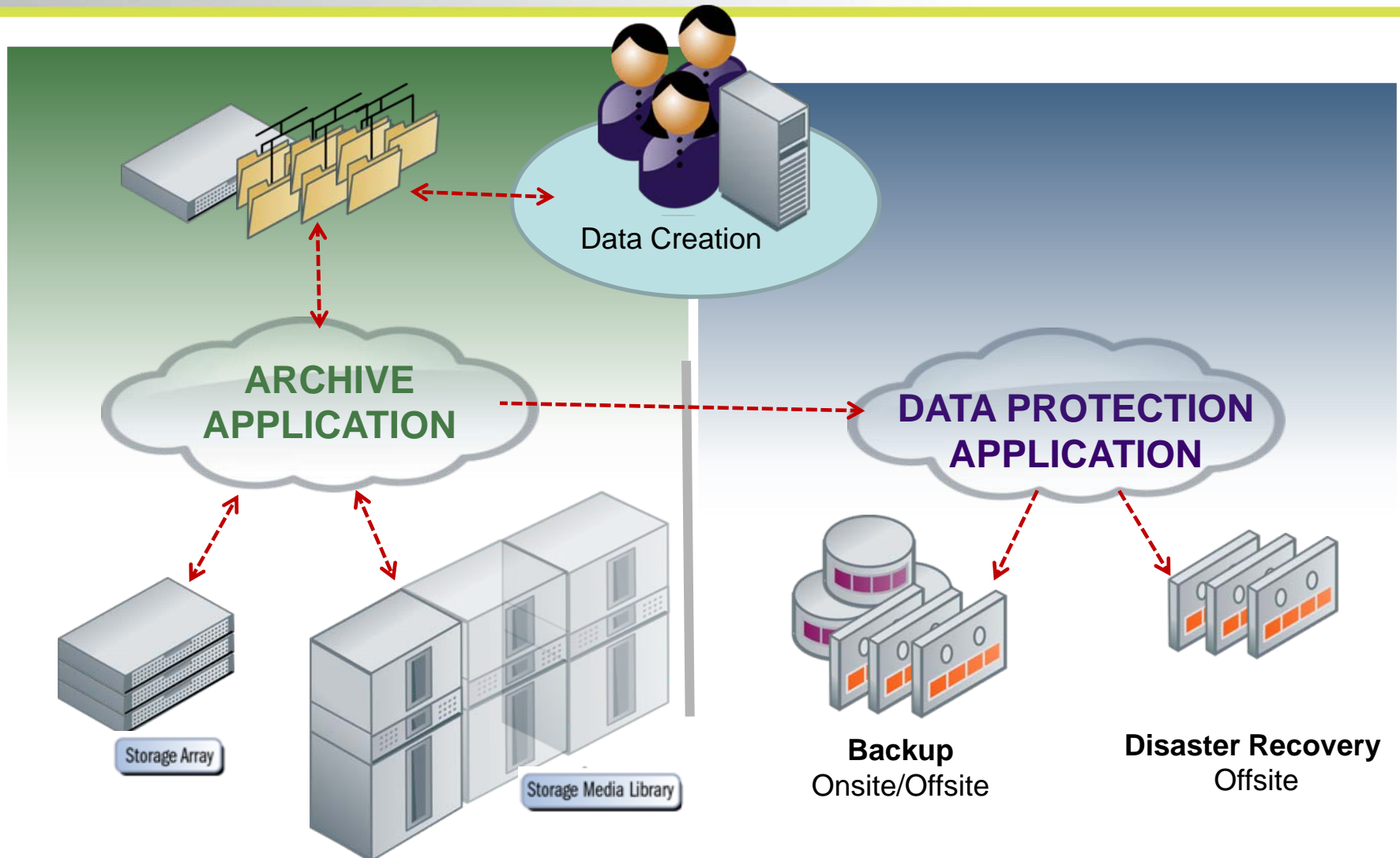
Estimates are the total amount of digital information will grow at a rate of 58% per year, reaching 1610 EB by 2011! Each EB is equivalent to 50,000 times the entire U.S. Library of Congress printed collection.\*

An active archive can provide an affordable, online solution to access and store all this newly created data!



Sources: \* International Data Corporation 2010

# The Ideal Modern Data Center



**Active Archive of Primary Data Set**

**Copy of Data**

- Attributes of active storage
  - Must integrate easily with archive management application making it simple access data (regardless of the storage type)
  - Power Efficient
  - Very high density
  - Reliable (built in data integrity validation)
  - Low cost
  - Fast throughput

# Addressing the Needs of Active Archive



***“The disk systems cost 25 times more to power and cool than the tape system...”***

The disk systems cost 25 times more to power and cool than the tape system...a data center needed to store 150 TBs of data that was growing at 30% per year. The 15 disk systems configured will cost about \$109,745 in electricity in one year. The electric bill for the automated tape library will only be \$4,238 a year.

—Source: The Clipper Group Report #TCG200701

**energy**  
ENERGY ACCOUNTABILITY  
**Demand it.**



Yearly Electrical Costs with a 30% Increase in Data Each Year

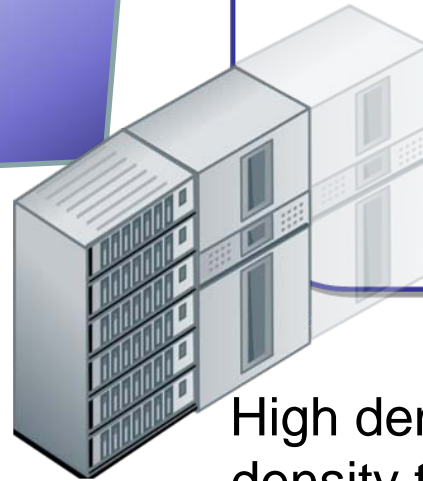
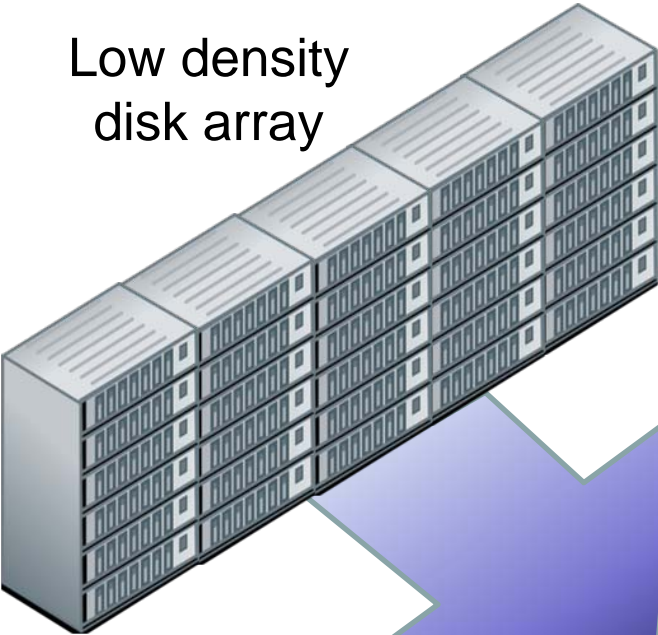
Time	Tape Systems	Disk Systems
Year One	\$4,238	\$109,745
Year Two	\$4,662	\$120,720
Year Three	\$5,128	\$132,792
Year Four	\$5,640	\$146,071
Year Five	\$6,205	\$160,678

Source: Clipper analysis

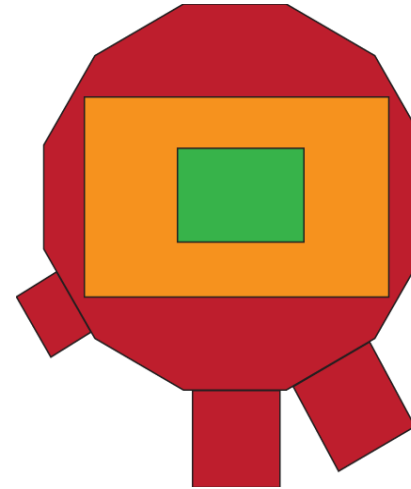
Active Archive – Data Protection for the Modern Data Center  
© 2010 Storage Networking Industry Association. All Rights Reserved.

# Addressing the Needs of Active Archive

Low density  
disk array

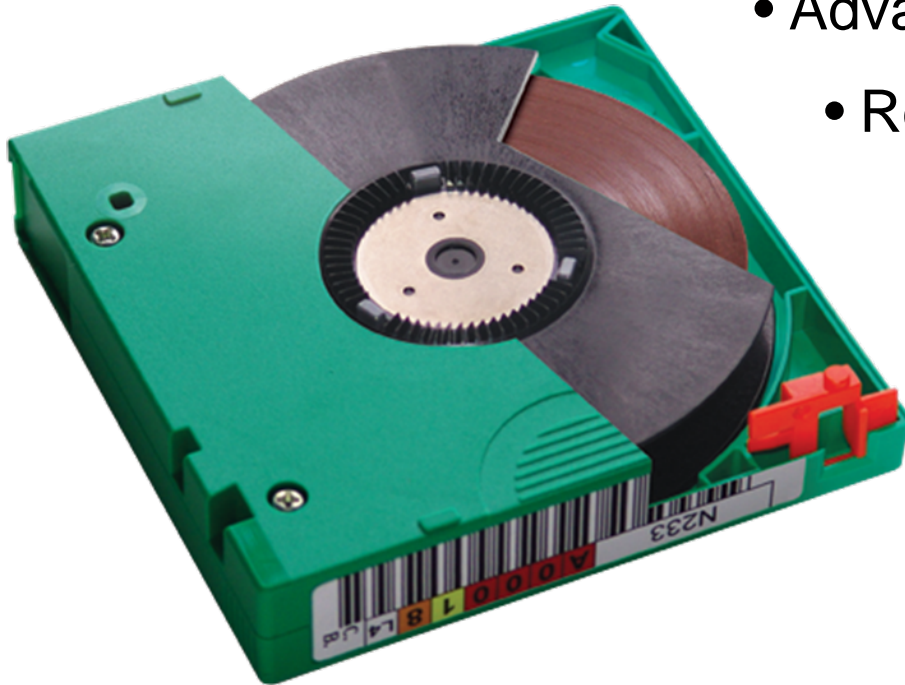


High Density Tape Reclaim  
space: 44 – 218%



High density disk array + high  
density tape = same capacity,  
smaller footprint

Tape media reliability has increased 700% over the technology available a decade earlier



- Advances in the coating of tape film
- Read-after-write data verification
- Error correction codes
- Drive technology features simplified tape paths and servo tracking systems

Beech, Debbie. "Best Practices for backup and long-term data retention" Sylvatica Whitepaper. The evolving role of disk and tape in the data center. June 2009

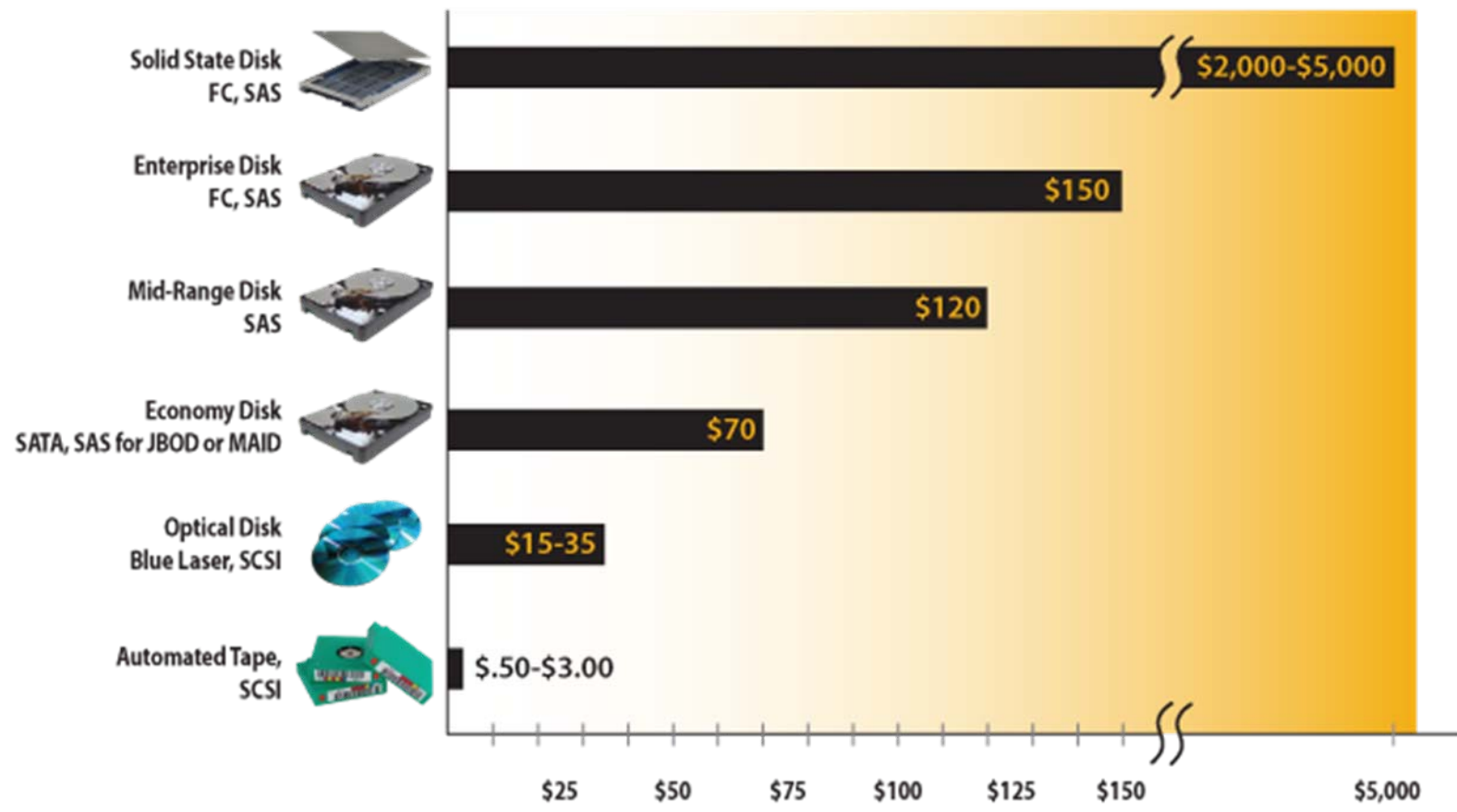
- Tape libraries today have the intelligence to proactively alert if a media or hardware issue is developing
- Expect the same reliability from disk and tape archive storage.  
Proactively be alerted when:
  - ◆ A tape or disk has been used beyond any manufacturer thresholds
  - ◆ There has been environmental damage to media
  - ◆ There are correlations between drive or media failures
  - ◆ Data is at risk and needs a duplicate copy



- ◆ Drive Lifecycle Management (DLM)
  - › Improves operational and support efficiency
  - › Reduces the risk of error during operational windows
  - › Improves ability to resolve, quickly and easily, error situations
- ◆ Library Lifecycle Management (LLM)
  - › Proactive notification of upcoming service events
  - › Auto-save of configurations for recovery / rollback
  - › Simple capacity expansion with system online
- ◆ Media Lifecycle Management (MLM)
  - › Prevents tape-related failures by alerting you when to replace tapes that pose a risk to your data
  - › Pre- and post-scan on tapes to ensure data integrity
  - › Tracking of cleaning media to avoid overuse

# Why Now?

# Today's Tape Technologies are Affordable...SNIA



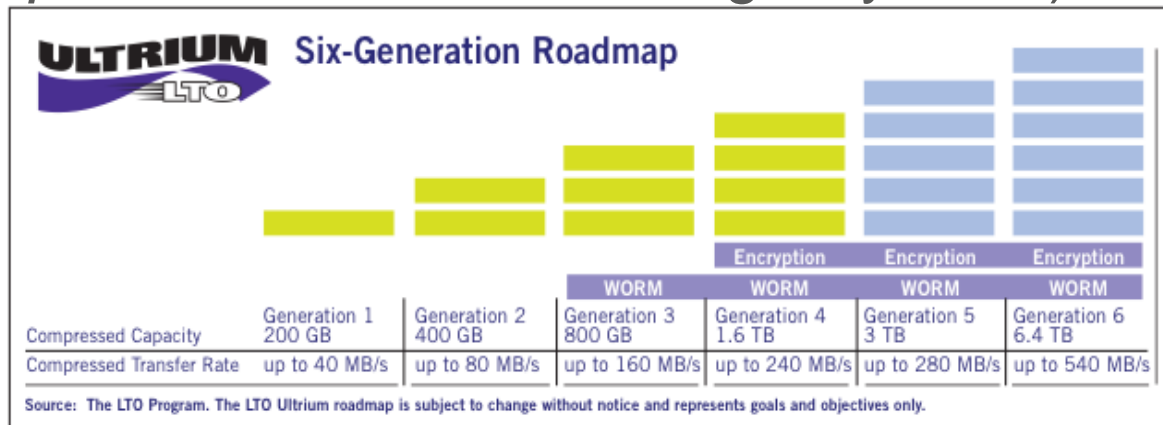
Storage Cost per GB

Source: IDC 2010

## Today's Tape Technologies are Fast...

### ...really fast

- Average file access time\*: 65-75 seconds\*\*
- Streamed data throughput\*:
  - ◆ 240 MB/second compressed transfer rate per drive; libraries scale to 480 tape drives (that's 240 *uncompressed TB/hour* to a single system!)



\* Based on benchmarked data

\*\* Times vary based on library and tape drive in use

Active Archive – Data Protection for the Modern Data Center

© 2010 Storage Networking Industry Association. All Rights Reserved.

# Active Archive Customer Case Study

# Achieving Efficiency with Active Archive

## NASA Ames Case Study

- Pain Points
  - Lack of roadmap for future product that mapped to data growth needs
  - Downtime due to media issues
  - Maxed out data center floor space utilization
- Goals
  - Ensure media and data integrity
  - Better manage media within the library
  - Improve storage density and foot print
  - Move to production quickly



## NASA AMES case study

### ► Benefits of migration to an active archive:

- ◆ Extended file system capacity on tape
- ◆ Reclaimed 1400 sq. ft. of data center space (that is the size of an average American's home!)
- ◆ Increased online archive capacity 12 PB to 32 PB
- ◆ Increased data storage reliability





Education

## **Data Classification**

A basic technique for Storage and Information Management

Dr. Rainer Pollak, DataGlobal

- External IT compliance requirements include:
  - ◆ PCI DSS  
Payment Card Industry Data Security Standard
  - ◆ EUDPD  
European Union Data Protection Directive
  - ◆ HIPAA  
Health Insurance Portability and Accountability Act
  - ◆ SOX  
Sarbanes-Oxley
  - ◆ GLBA  
Gramm-Leach-Bliley

# Compliance Recommendation!

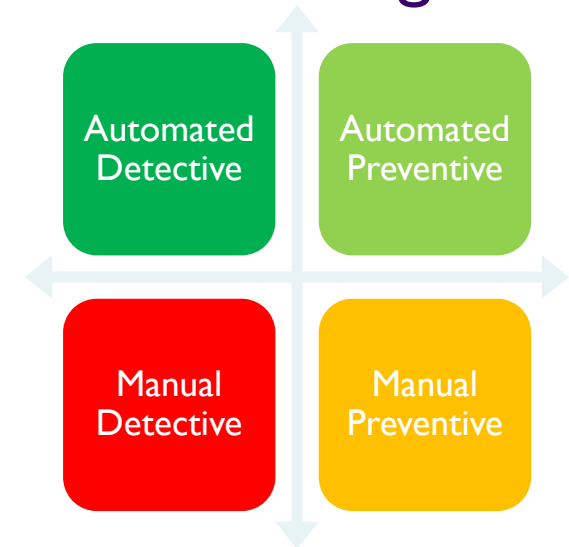
- If the driving force for data classification is external, do not handle it as an *internal* IT project!
- Consult legal counsel, a GRC subject matter expert or auditor to discuss specifics for your type of organization!

## ➤ Internal reasons to adopt data classification:

- ◆ Reduce cost of storage
  - > hierarchical storage management (HSM)
  - > binary classification criteria (active data vs. stale data)
- ◆ Data protection
  - > encryption
  - > confidentiality
- ◆ Cloud storage decisions
- ◆ Capacity management
- ◆ eDiscovery
- ◆ Knowledge capture
- ◆ Storage reclamation

➤ In the context of technical controls, the following classification scheme is often used.

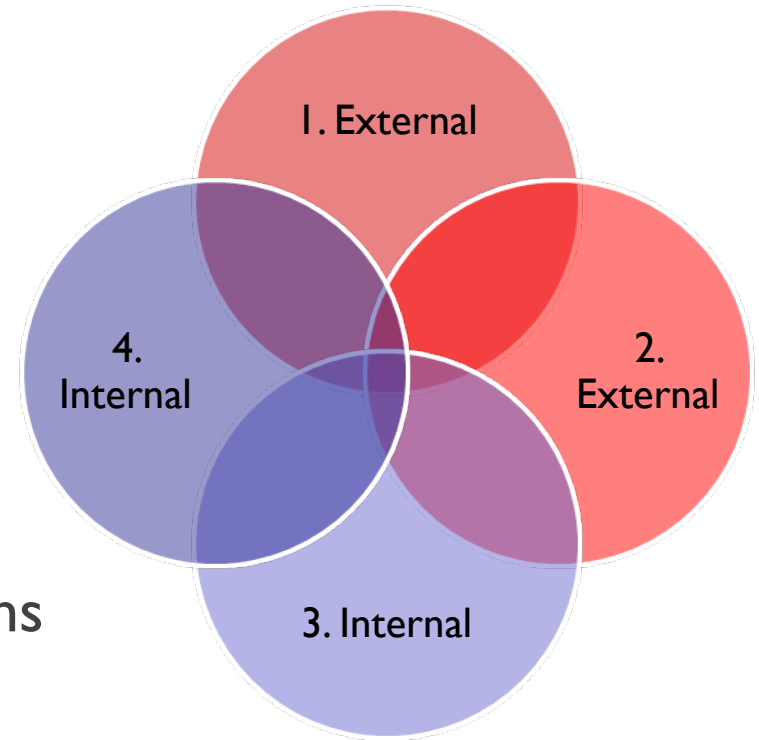
- ◆ Automated vs. manual
- ◆ Detective vs. preventive



➤ *Preventive* = classifying data at origination

➤ *Detective* = classifying the data after the fact

- Multiple external and internal requirements
- Requirements aren't static
  - ◆ New compliance needs arise
  - ◆ Existing compliance regulations are legally reduced
  - ◆ The nature of certain regulations change constantly

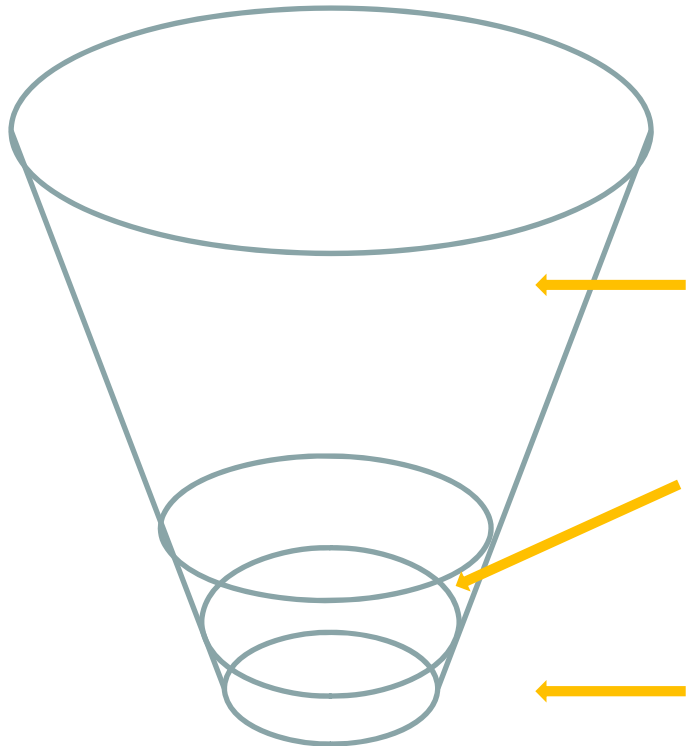


- 80% unstructured data is legacy that must be classified after the fact

⇒ **An automated detective solution is critical!**

## ➤ Every single file has to be classified

- To meet a combination of criteria
- Not only once, but continuously
- Reclassification of data is not optional, it is mandatory



Files have to be classified by their attributes and metadata!

- location, name, file extension, ACL's, ..

Just a **small subset** of the data can/should be indexed for more detailed analysis!

- full text, OCR, ..

The manual analysis of single files should be avoided!

- Data Classification by itself is not the goal
- Data Classification provides data needed to perform another process
  - ◆ Migration
  - ◆ Archival
  - ◆ Storage Tiering
  - ◆ Compliance
- You must know how the results are to be handled and stored to ensure the Data Classification results are useful

## ➤ Paper reports are useless

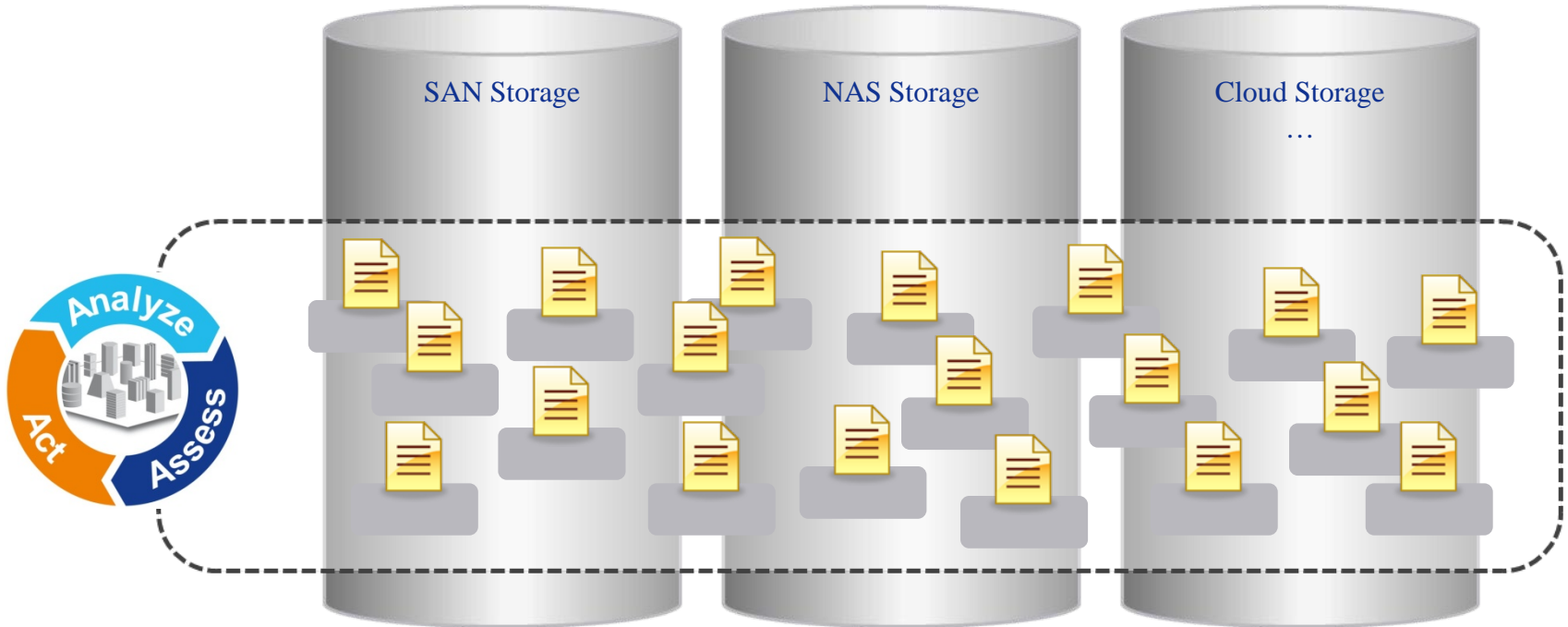
- ◆ Obsolete as soon as it is printed
- ◆ Further processing is extremely labor intensive

## ➤ Database storage

- ◆ CAN be processed further by 3rd party applications, but
- ◆ Impossible to keep the results synchronized with file system

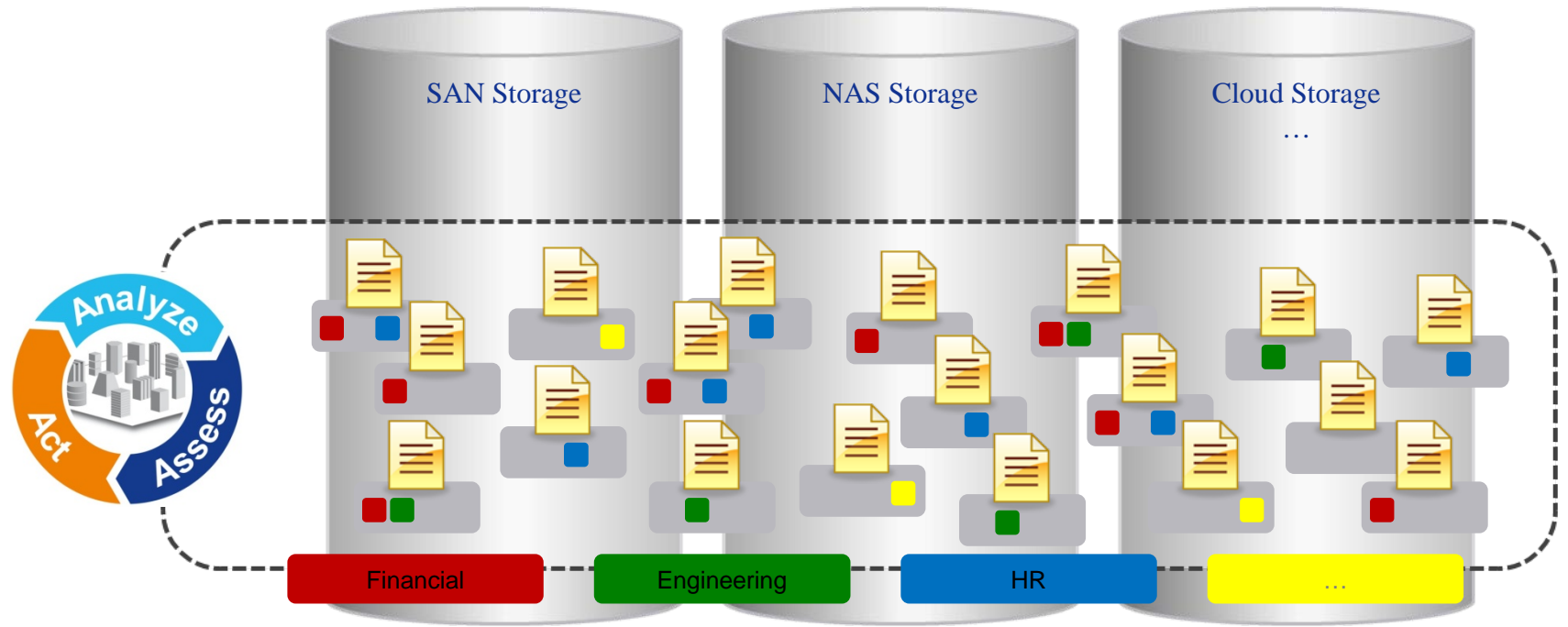
## ➤ Directly attached to the file

- ◆ Storing metadata directly on the file is critical
- ◆ Ensures classification results remain with the file regardless of location
- ◆ Results in synchronized storage that can always be accessed and protected



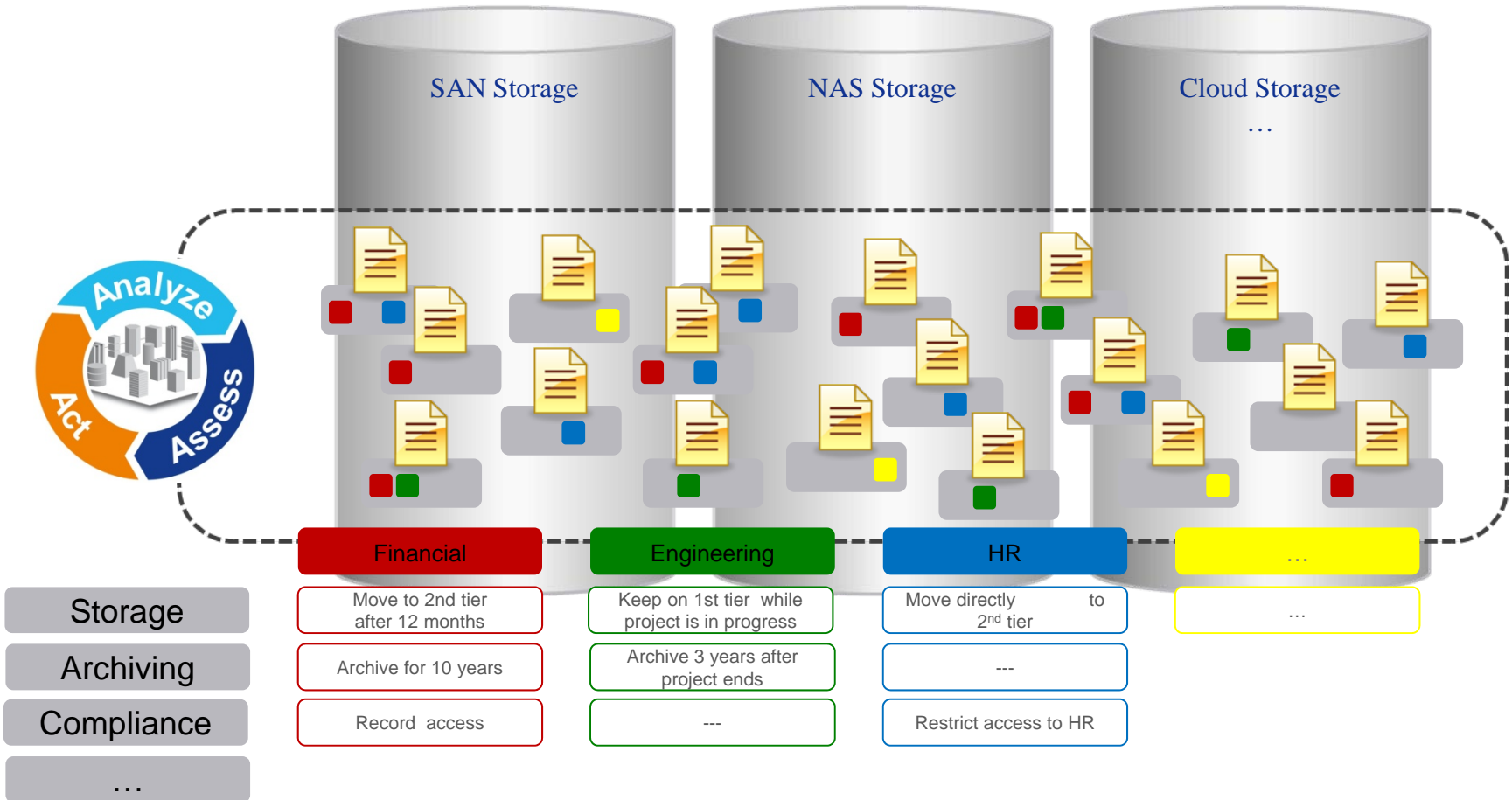
Scanning (attributes, metadata, ...) all storage resources independent of the sourcing application and location.

# Basic Workflow



Organizations have to use **enterprise-wide rules** to classify all their information assets

# Basic Workflow



- 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**

**Molly Rector  
Robin Lutchansky  
Michael Schwend**

**Dr. Rainer Pollak  
Michael Fishman**