Long-term retention and access to data in the life-sciences: beyond storage

Dr Matthew Addis

Arkivum Ltd

NGS 2012 London

Contents

- Why archive data?
- The challenges of long-term data retention
- Recommendations
- The Arkivum solution
- Questions

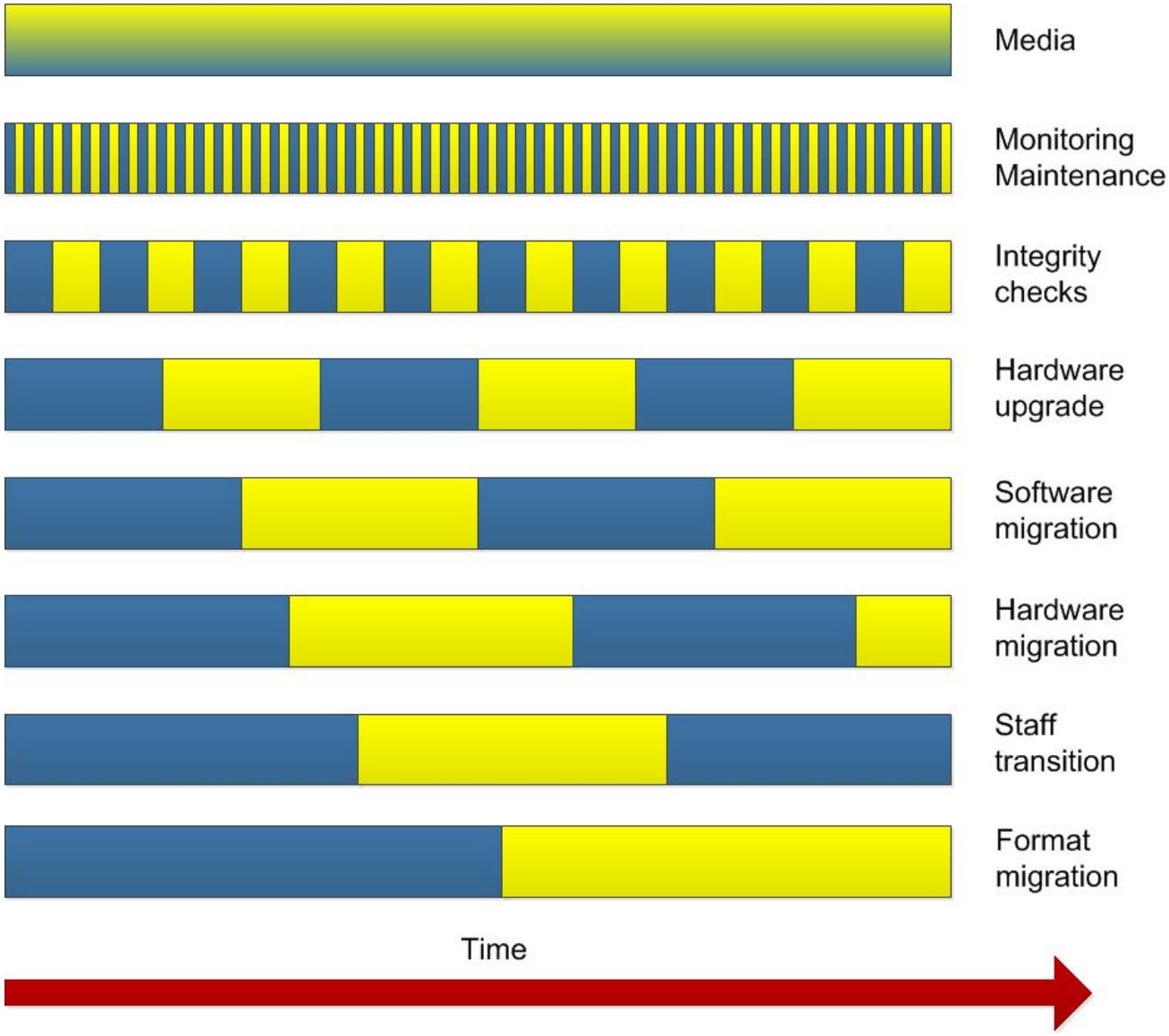
"Security is of key importance to our business, Arkivum's A-Stor Pharma service allows us to store our encrypted data for the long term in a cost efficient way that is entirely scalable and reduces pressure on our internal IT infrastructure."

Dan Watkins, Oxford Fertility Clinic

Why *archive*?

- Regulation and compliance
- Intellectual Property
- Reuse
- Save money
- Better use of in-house resources

20 years of keeping content alive



30,000 years

2,200 years

1,200 years

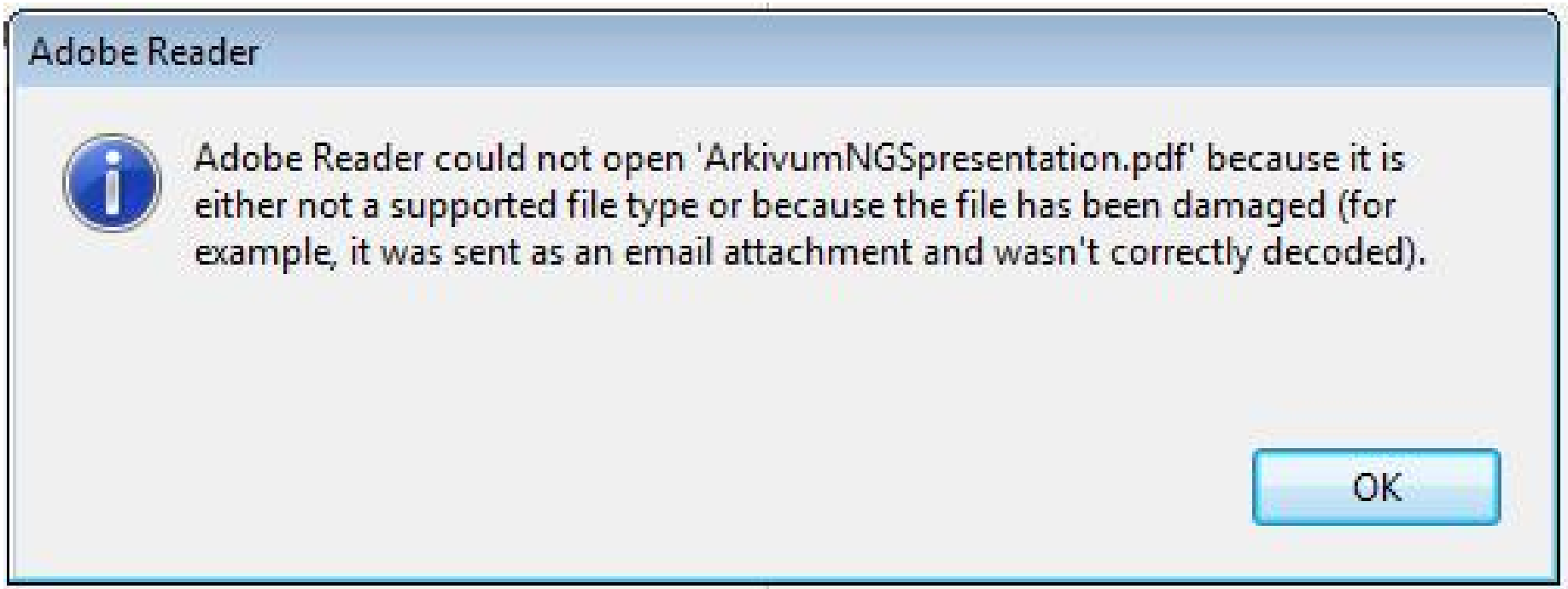
200 years

Each new generation of technology:

1000x times denser
lasts 1/10th as long

20 years

Copyright Barney Livingstone



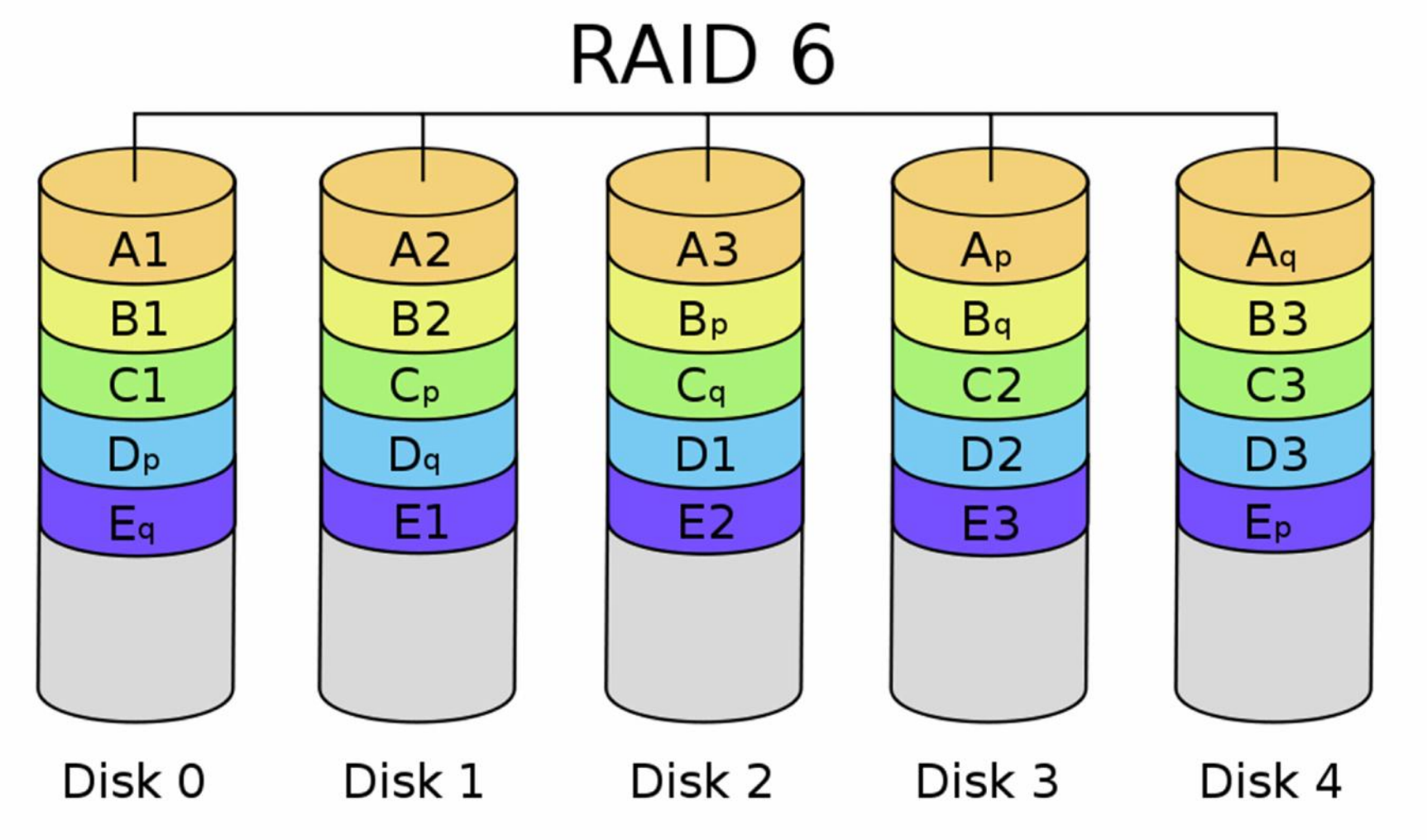
It's not getting any better!

- 1000 times more HDD capacity over last 15 years
- Only 10 times lower Bit Error Rates (BER)

- HDD BER = 10^{-14}
- 1 TB = 10^{13} bits
- 10% chance of an error when reading all of a HDD

- *Within a few years, you are more likely than not to get a read error when copying a HDD*

The IT Industry knows this already



Systems bring their own problems

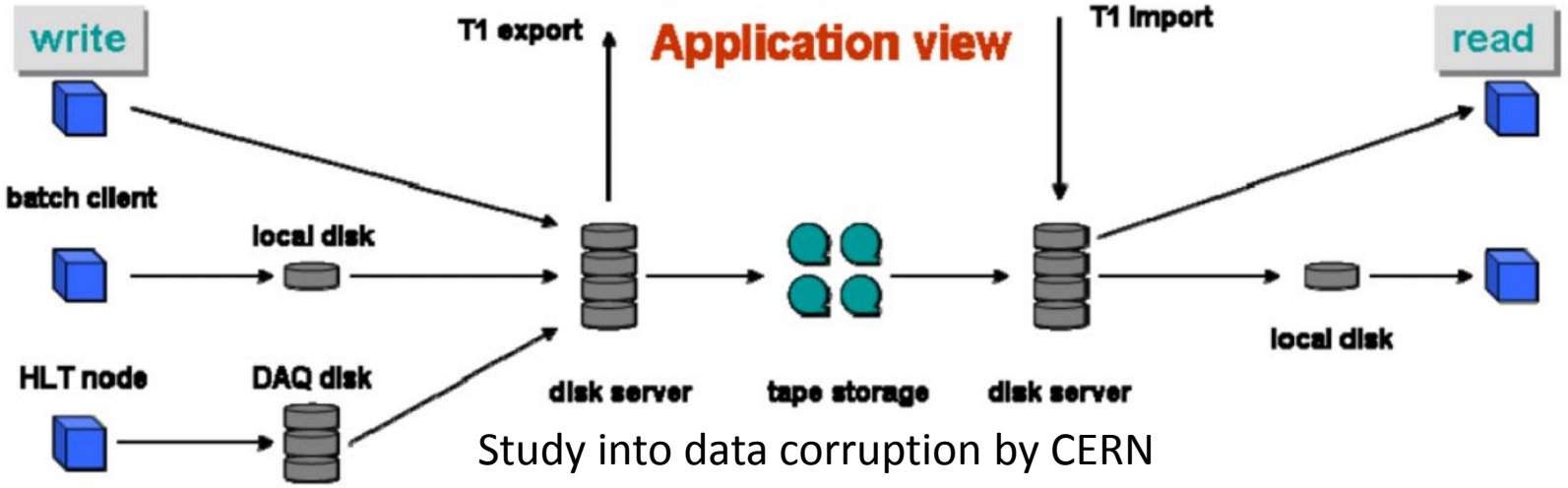
"Disk failures are not always a dominant factor of storage subsystem failures, and a reliability study for storage subsystems cannot only focus on disk failures. Resilient mechanisms should target all failure types"

2008 NetApp study of 1.8M HDD in 155,000 systems

‘bit rot’

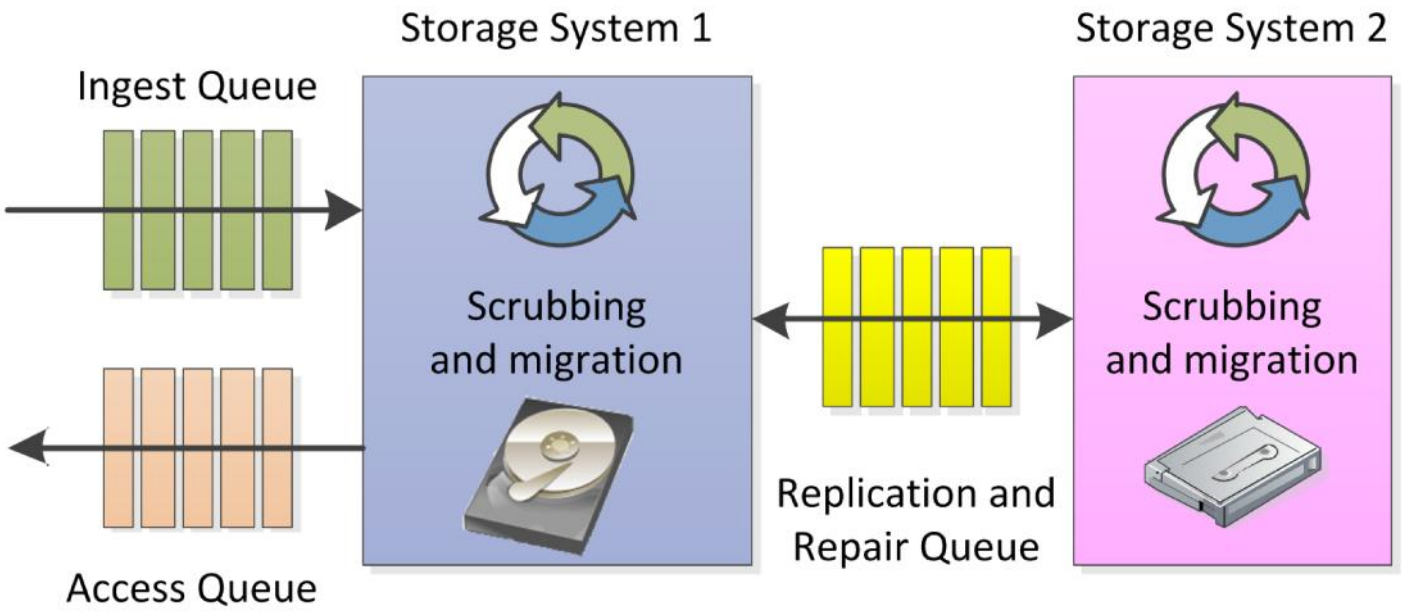
David Rosenthal’s blog
http://blog.dshr.org/

- Errors can be silent (latent)
  - Permanent and undetected corruption of data
  - Deeply worrying for archives
  - Seen in the field (if you know how to look)



Study into data corruption by CERN

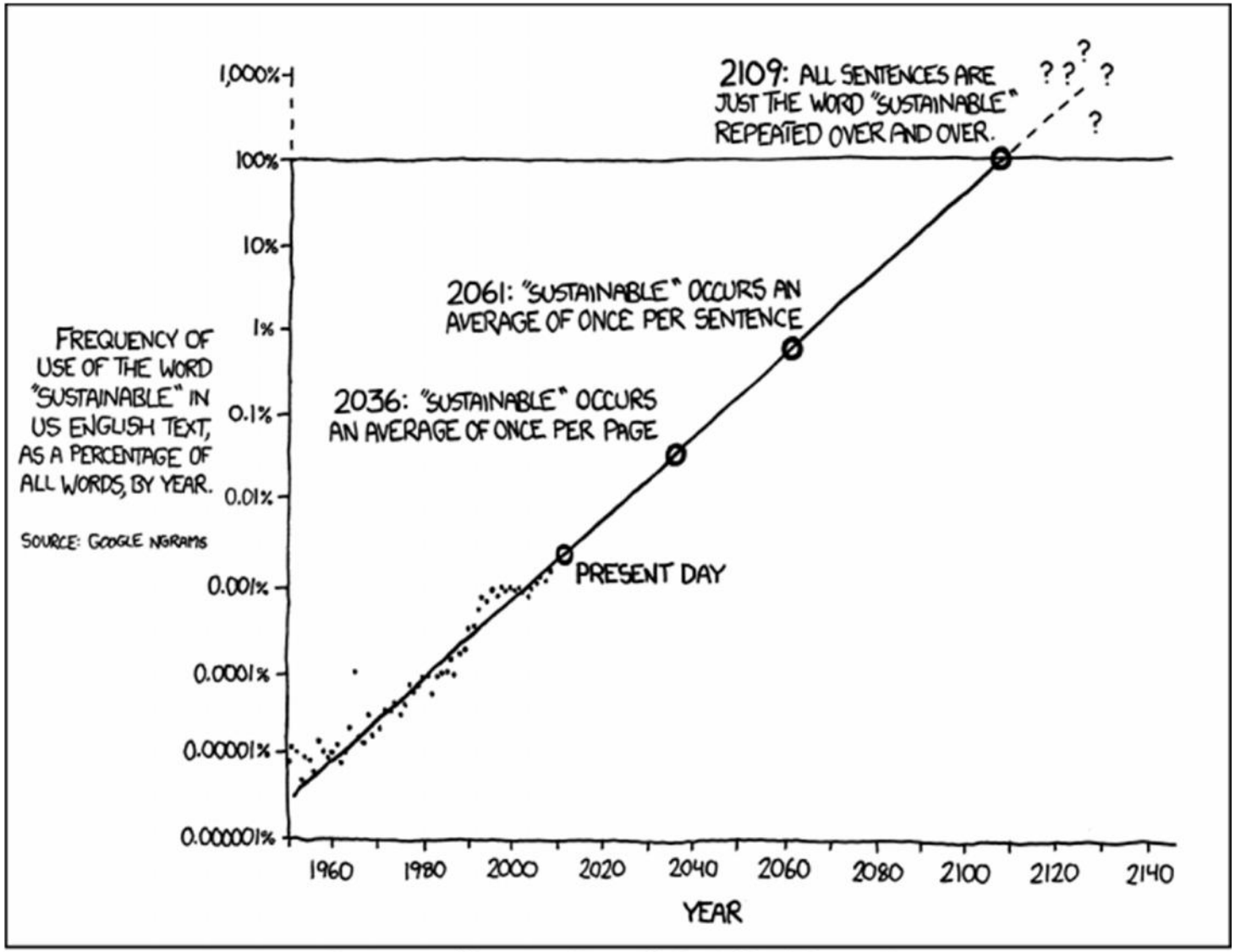
So what do you do?



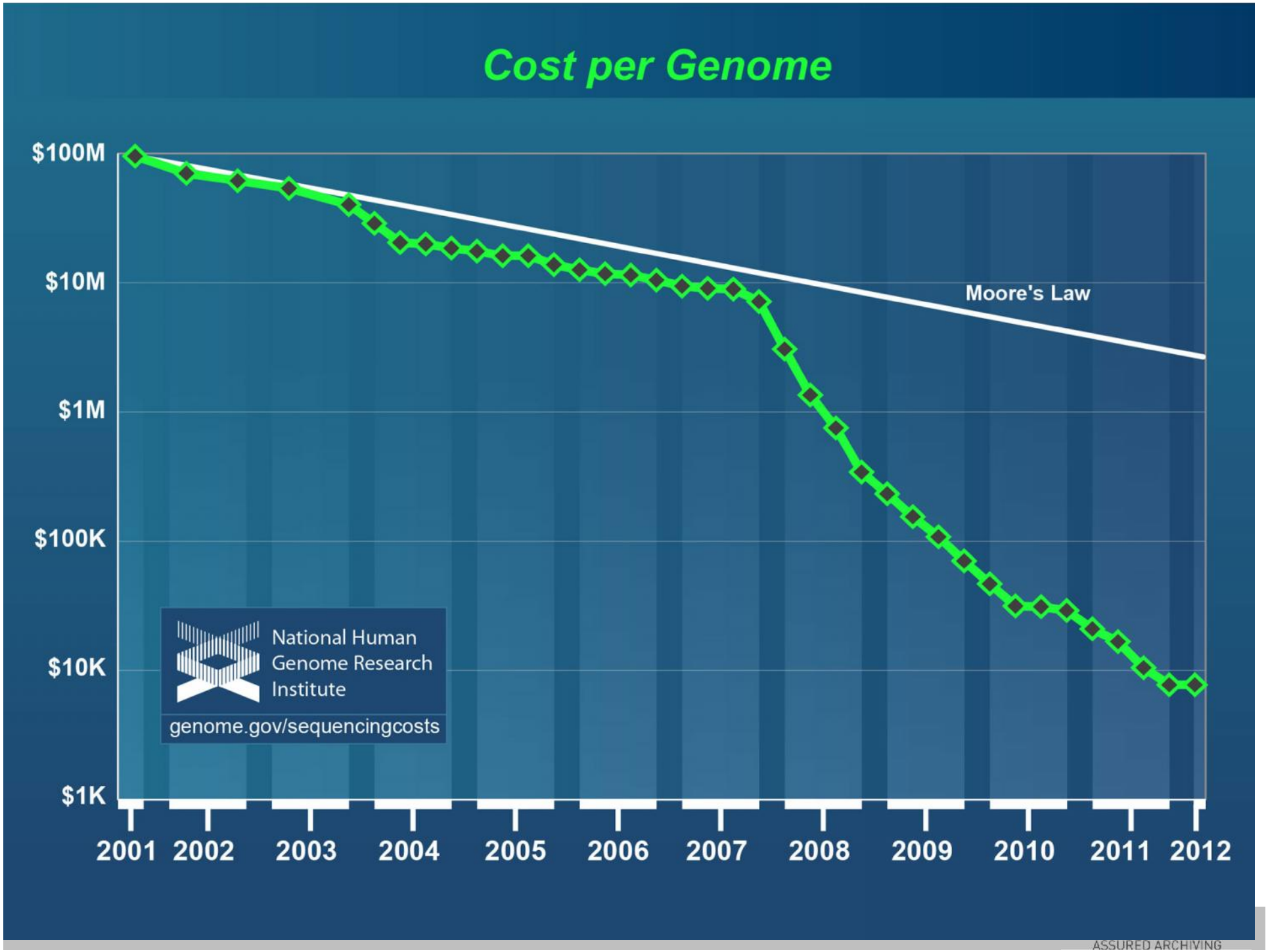
- Preservation best practice (diversity, intervention)
  - Multiple copies in different locations
  - Different technologies and different people
  - Active management: migration, integrity

Why not make more copies?

Why not make more copies?



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Something has to give

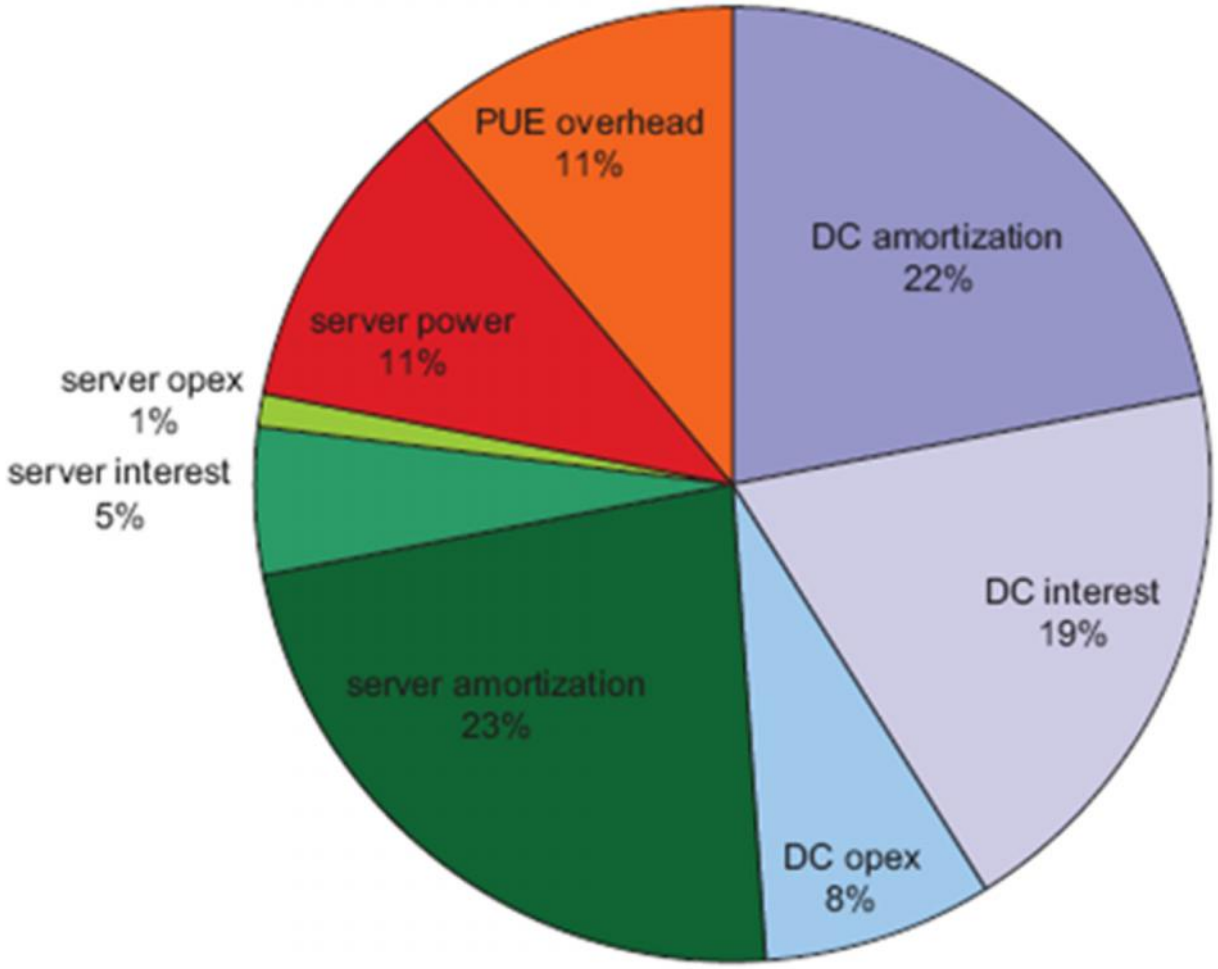
- >100% CAGR in data volumes
- 20% CAGR in HDD capacity per £ IHS iSuppli
- 2% CAGR in IT Budgets ComputerEconomics

→ Store less

→ Compress

→ Find a way to lower the costs

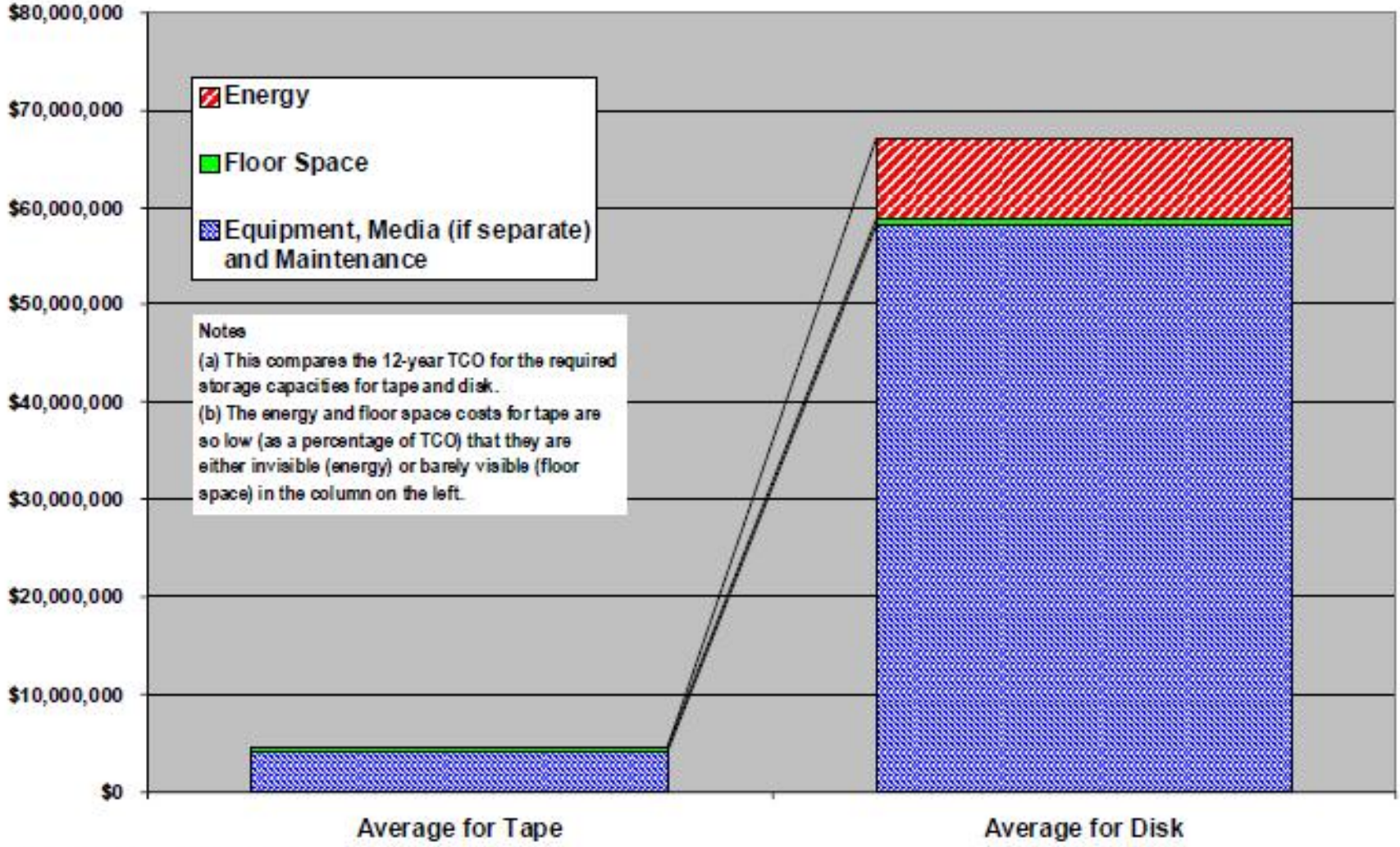
Storage TCO



Google

Exhibit 1 —
Comparing 12-Year TCO for Tape to Disk for Long-Term Archived Data

TCO for Disk is Approximately 15 Times Tape Using Clipper's Case Study Model



Source: The Clipper Group

"Storing infrequently-accessed data on disk is equivalent to keeping your car running constantly in the driveway - it wastes energy and it costs money."

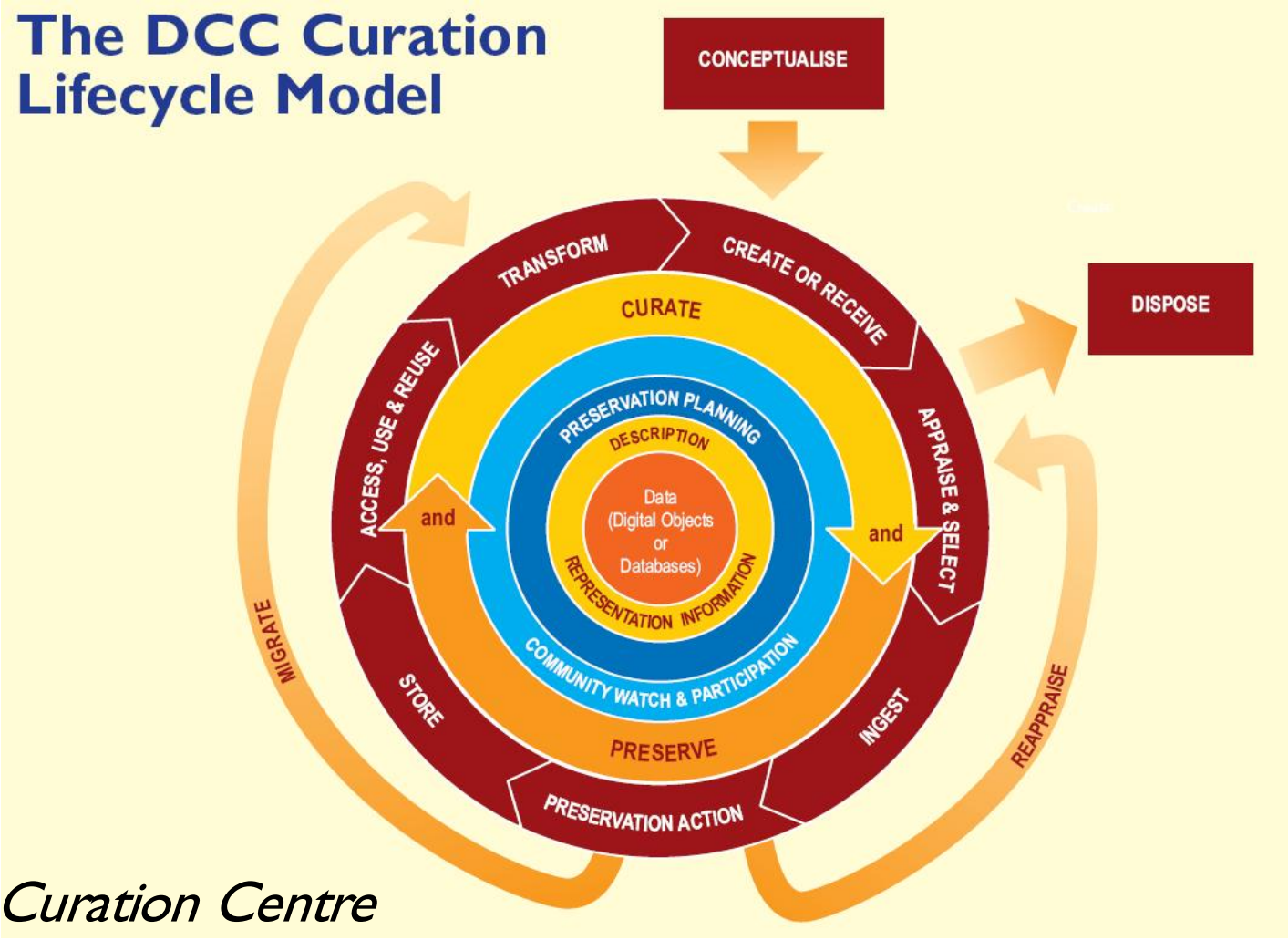
The Clipper Group

Orion

Cost v.s. safety v.s. access

	Data tape on shelves	HDD in servers	Storage as a Service
Storage Cost	Low (media, shelves, climate control)	High (servers, power, cooling, maintenance)	High (fully managed service)
Access Cost	High (people retrieve and load media)	Low (internal network, automated)	High (bandwidth, charges for i/o)
Latent Failures	Low (data tape is reliable)	Med ('bit rot')	Low (replication and monitoring)
Access Failures	Medium (people handle tapes)	Low/Medium (depends on system)	Low (automated checks)

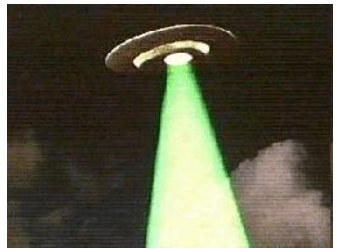
Preservation and curation



Digital Curation Centre

Risks

- Technical obsolescence, e.g. formats and apps
- Hardware failures, e.g. digital storage systems
- Loss of staff, e.g. skilled archive and IT staff
- Insufficient budget, e.g. storage too expensive
- Accidental loss, e.g. human error
- Selection, e.g. don't retain what you should
- Stakeholders, e.g. retention no longer a priority
- Underestimation of resources or effort
- Fire, flood, meteors, aliens…



Manage the risks

- Do nothing
- Do the wrong thing
- Do it in-house
- Use a service provider

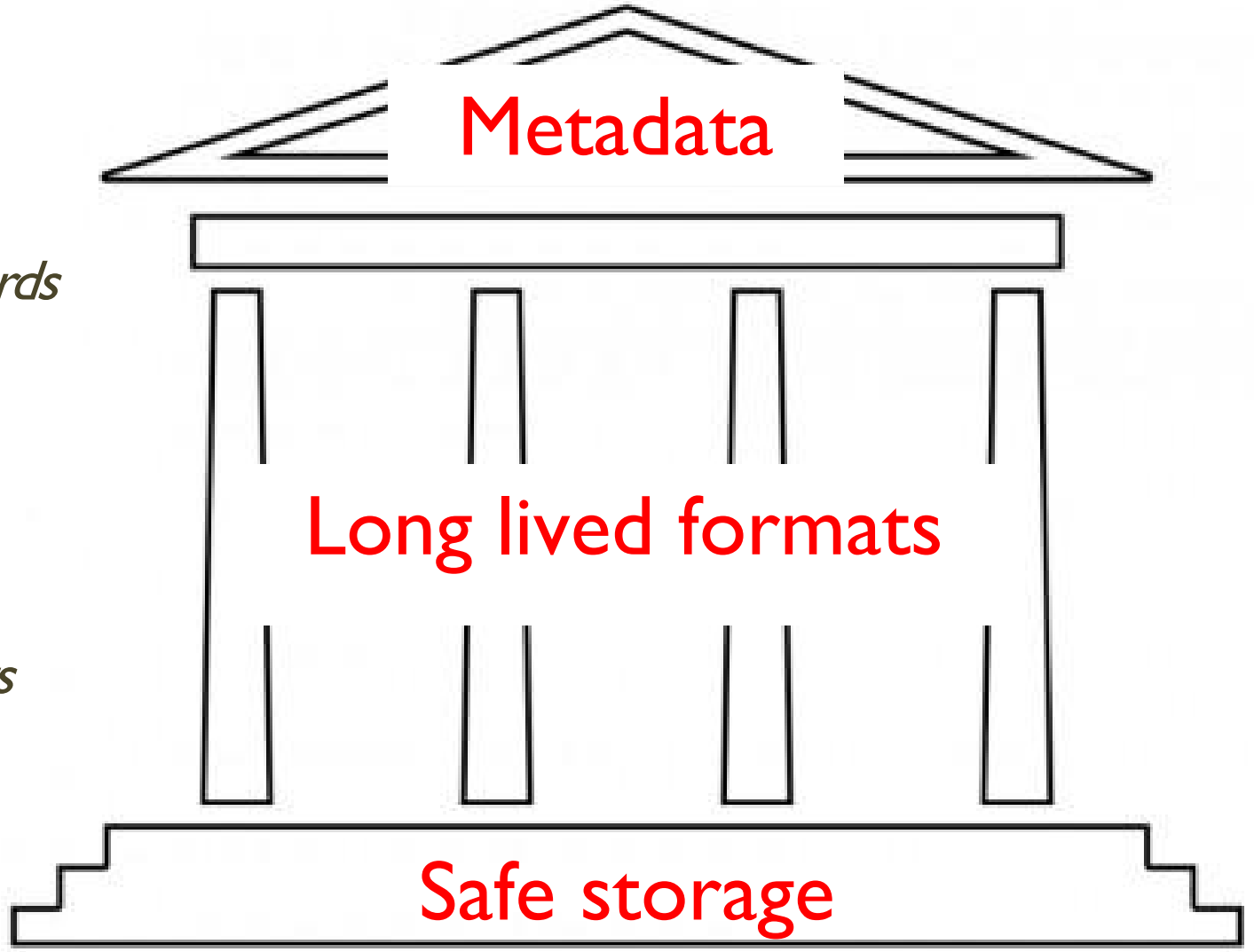
- ISO27001 Information Security Management
- ISO16363 Trusted Digital Repositories

Simple strategy

Descriptive
Technical
Preservation
Rights
Metadata Standards

Creation
Preservation
Access
Canonical Formats

Optimise for archive



Seven questions

Integrity	proving data hasn't changed
Authenticity	proving where the data came from
Availability	access when you need it
Confidentiality	control over who can read it
Possession	knowing who has it
Utility	knowing you can still use it
Provenance	chain of custody

Three recommendations

1. Risk management
2. Start somewhere, start now
3. Build a preservation house

Arkivum

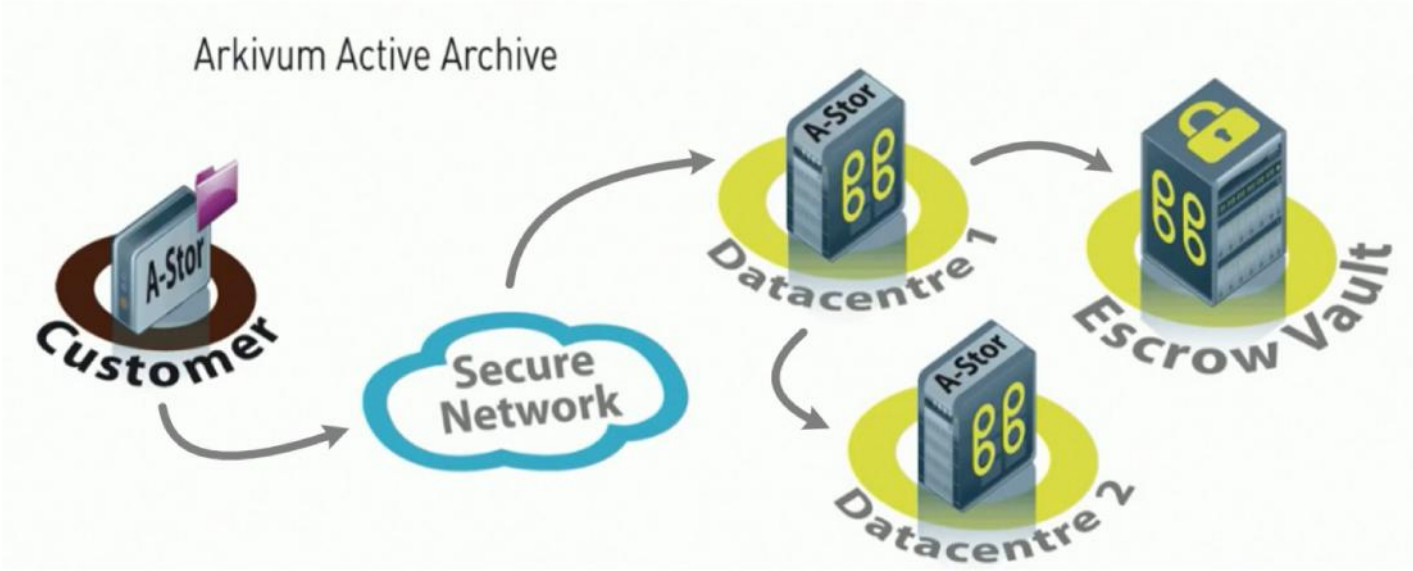
- Online data archiving as a service
- Founded in 2011
- Decade of know-how working with archives
- Safe, secure, accessible long-term data storage

Fundamentals

- Risk assessment and management
  - ISO27001 (information security)
  - ISO16363 (trusted digital repositories)
  - ISO14721 (open archival systems)
  - 78 risks, 7 categories, each one mitigated

- Ground up engineering of safety and security
  - Integrity, encryption, media handling
  - Staff, policies, procedures

How does it work?



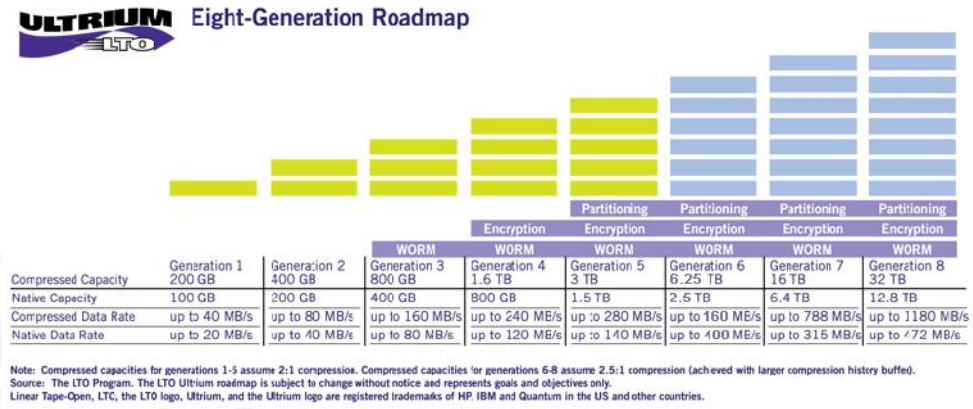
- 3 copies, 3 locations, online and offline
- Active management: integrity, obsolescence
- Standard filesystem, global persistent namespace

100% data integrity guarantee

- All data is returned ‘bit perfect’
- No restriction on time
- No restriction on volume
- Included in the SLA
- Worldwide insurance backed: £5M per loss event
- Supported by ISO27001 certification

Data escrow

- Copy of data offline at third-party escrow site
- LTFS on LTO tape with open source tools
- Customer has access to escrow copy:
  - If we fail to provide the service
  - If the customer decides to leave



Compressed Capacity	Generation 1 200 GB	Generation 2 400 GB	Generation 3 800 GB	Generation 4 1.6 TB	Generation 5 3 TB	Generation 6 6.25 TB	Generation 7 16 TB	Generation 8 32 TB
Native Capacity	100 GB	200 GB	400 GB	800 GB	1.5 TB	2.5 TB	6.4 TB	12.8 TB
Compressed Data Rate	up to 40 MB/s	up to 80 MB/s	up to 160 MB/s	up to 240 MB/s	up to 280 MB/s	up to 160 MB/s	up to 788 MB/s	up to 1180 MB/s
Native Data Rate	up to 20 MB/s	up to 40 MB/s	up to 80 MB/s	up to 120 MB/s	up to 140 MB/s	up to 400 MB/s	up to 315 MB/s	up to 472 MB/s

Note: Compressed capacities for generations 1-5 assume 2:1 compression. Compressed capacities for generations 6-8 assume 2.5:1 compression (achieved with larger compression history buffer).
Source: The LTO Program. The LTO Ultrium roadmap is subject to change without notice and represents goals and objectives only.
Linear Tape-Open, LTC, the LTO logo, Ultrium, and the Ultrium logo are registered trademarks of HP, IBM and Quantum in the US and other countries.

Transparent pricing

- *Long-term* retention and access
- Pay as you go
- 10 year fixed price
- 25 year fully Paid-Up

Thank you

- Arkivum Stand
- www.arkivum.com
- matthew.addis@arkivum.com

"Arkivum has helped us to create a robust archiving solution that will allow us to focus our budget on the business rather than yet more storage"

"Archived documents can then be seamlessly accessed from within the document management system, in the same way current documents are".