College of Information Studies
University of Maryland Hornbake Library Building College Park, MD 20742-4345

Storage and Preservation

Week 3

LBSC 671

Creating Information Infrastructures

Physical Storage

- Segregate by:
  - Users (e.g., Chemistry library)
  - Type (e.g., audiovisual materials)
  - Usage frequency (e.g., offsite storage)
  - Size (e.g., folios)
- Arrange in a way that facilitates access
  - Topical shelf order (e.g., Dewey Decimal System)
- Foster preservation
  - Environment (temperature, humidity, light)
  - Access controls (closed stacks, gloves, …)

High-Density Shelving

http://www.kmhsystems.com/high-density-storage.html

Compact Storage Robot

Kyushu University, Japan

Closed Stacks

University of Education, Ghana

Preservation

c. 3000 BCE

Organic Decay

- Rag paper: 300-2,000 years
- Acidic paper: 25-50 years
- Acetate: 40 years
- Nitrate: 40-1-00 years

ISO 11799:2003

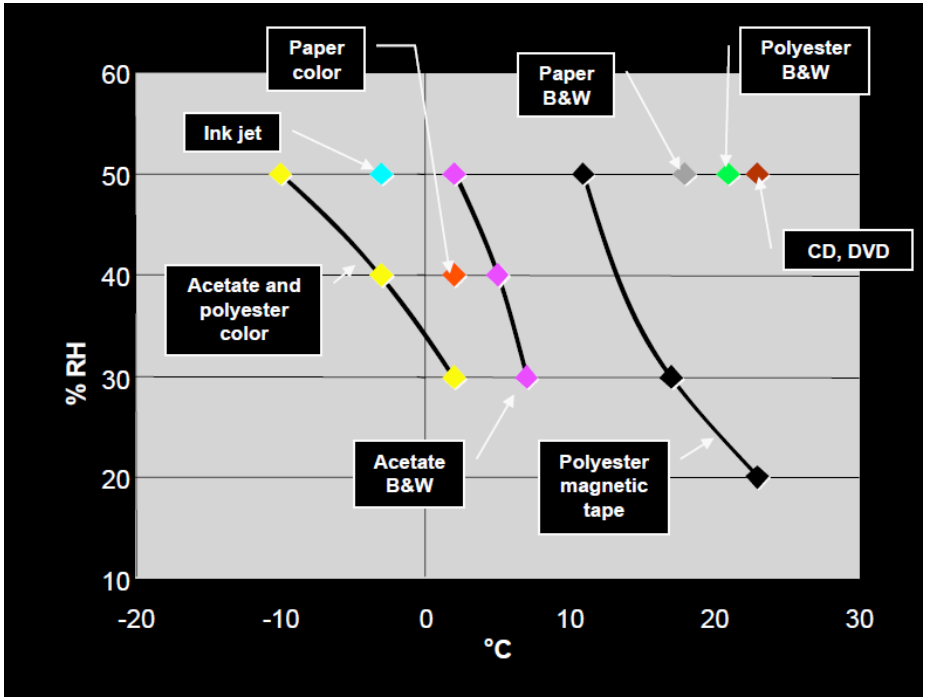


Image Permanence Institute, 2012

Threats to Physical Collections

- Organic decay
- Intentional actions
  - Pilferage and vandalism
  - Official actions
- Disasters
  - Accidents
    - Fire, sprinkler malfunction, …
  - Natural disasters
    - Flood, tornado, earthquake, …
  - Armed conflict

Disaster Mitigation Examples

- Flood:
  - Decide quickly what to freeze
    - And know where you can vacuum freeze dry
    - Air dry or dehumidify the rest
  - Immerse wet+muddy tape or film in water
    - Then air dry or dehumidify
  - Replace wet archival boxes immediately
- Fire:
  - Handle as fragile, wrap in clean paper
  - Pack between cardboard to stiffen

http://matrix.msu.edu/~disaster/balcplan.php

Digital Preservation

- Preservation of born-digital materials
  - Appearance
  - Behavior

- Digitization for preservation
  - Scanning (paper, microfilm)
  - Audio digitization
  - Video digitization
  - Volumetric imaging
    - Digital holography, computational tomography

Binary Data Representation

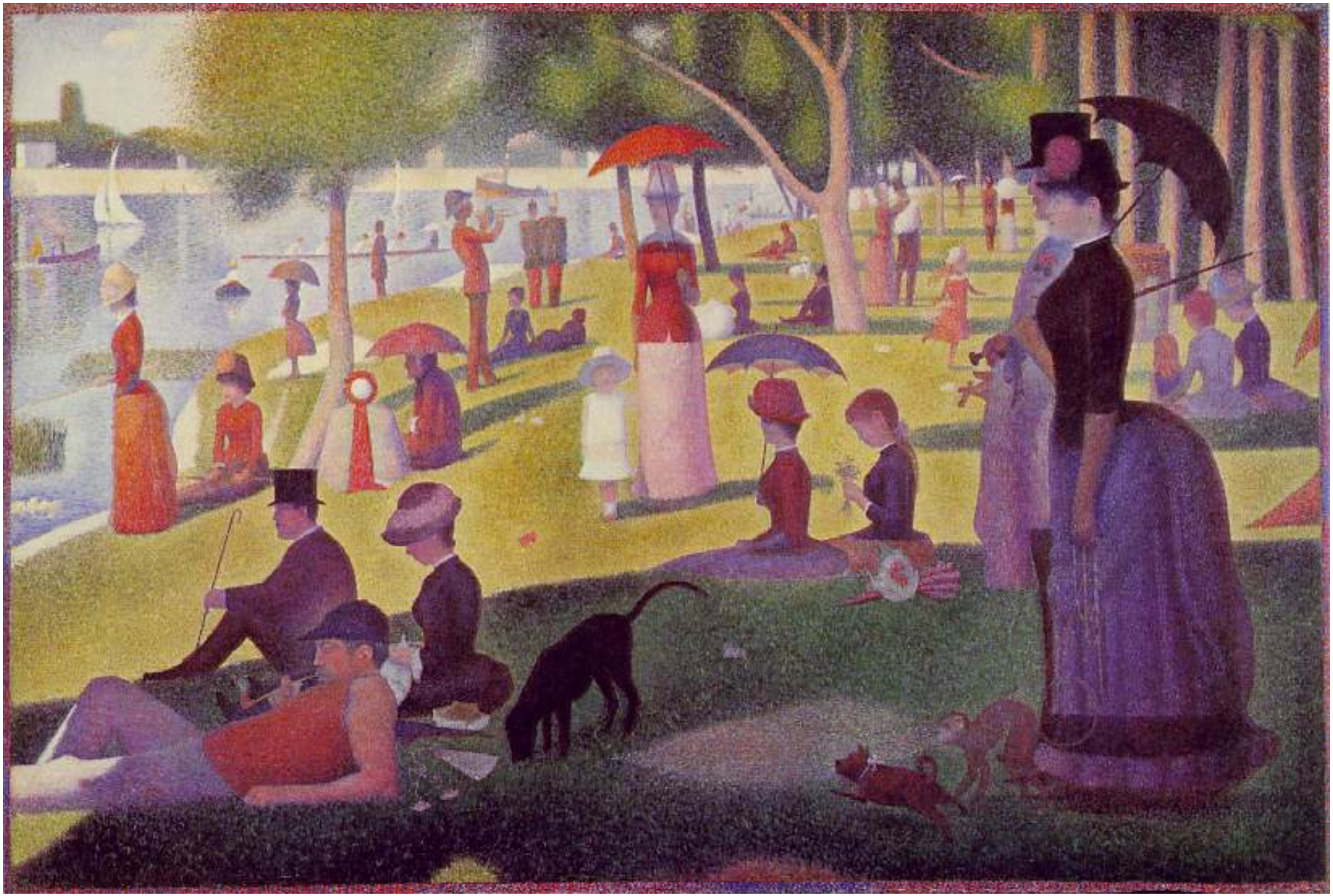
Example: American Standard Code for Information Interchange (ASCII)

01000001 = A	**01100001 = a**
01000010 = B	**01100010 = b**
01000011 = C	**01100011 = c**
01000100 = D	**01100100 = d**
01000101 = E	**01100101 = e**
01000110 = F	**01100110 = f**
01000111 = G	**01100111 = g**
01001000 = H	**01101000 = h**
01001001 = I	**01101001 = i**
01001010 = J	**01101010 = j**
01001011 = K	**01101011 = k**
01001100 = L	**01101100 = l**
01001101 = M	**01101101 = m**
01001110 = N	**01101110 = n**
01001111 = O	**01101111 = o**
01010000 = P	**01110000 = p**
01010001 = Q	**01110001 = q**
…	**…**

Units of Size

Unit	**Abbreviation**	**Size (bytes)**
bit	b	1/8
byte	B	1
kilobyte	KB	$2^{10} = 1024$
megabyte	MB	$2^{20} = 1,048,576$
gigabyte	GB	$2^{30} = 1,073,741,824$
terabyte	TB	$2^{40} = 1,099,511,627,776$
petabyte	PB	$2^{50} = 1,125,899,906,842,624$

Nothing new…



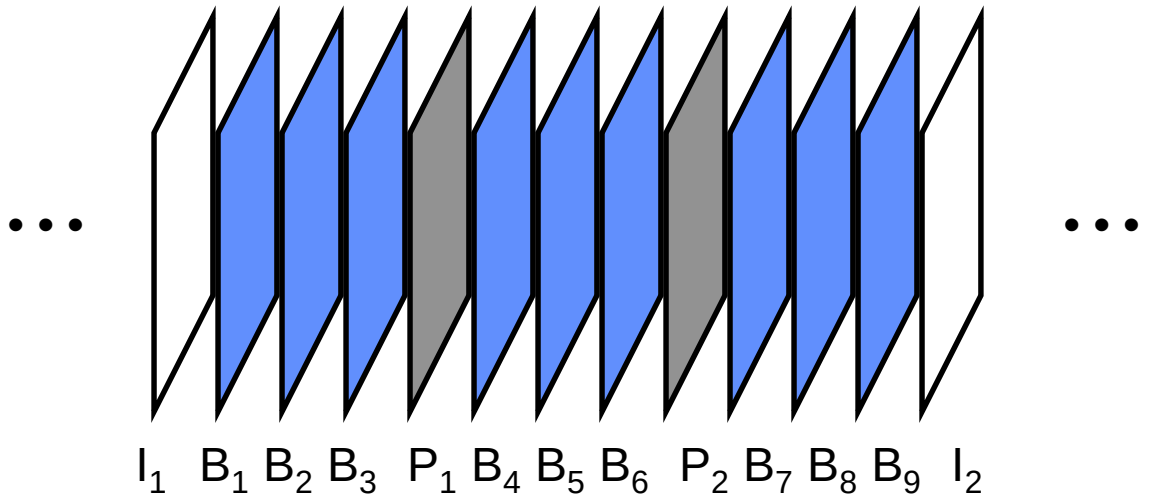
Georges Seurat, A Sunday Afternoon on the Island of La Grande Jatte

Basic Audio Coding

- Sample at twice the highest frequency
  - 8 bits or 16 bits per sample

- Speech (0-4 kHz) requires 8 kB/s
  - Standard telephone channel (1-byte samples)
- Music (0-22 kHz) requires 172 kB/s
  - Standard for CD-quality audio (2-byte samples)

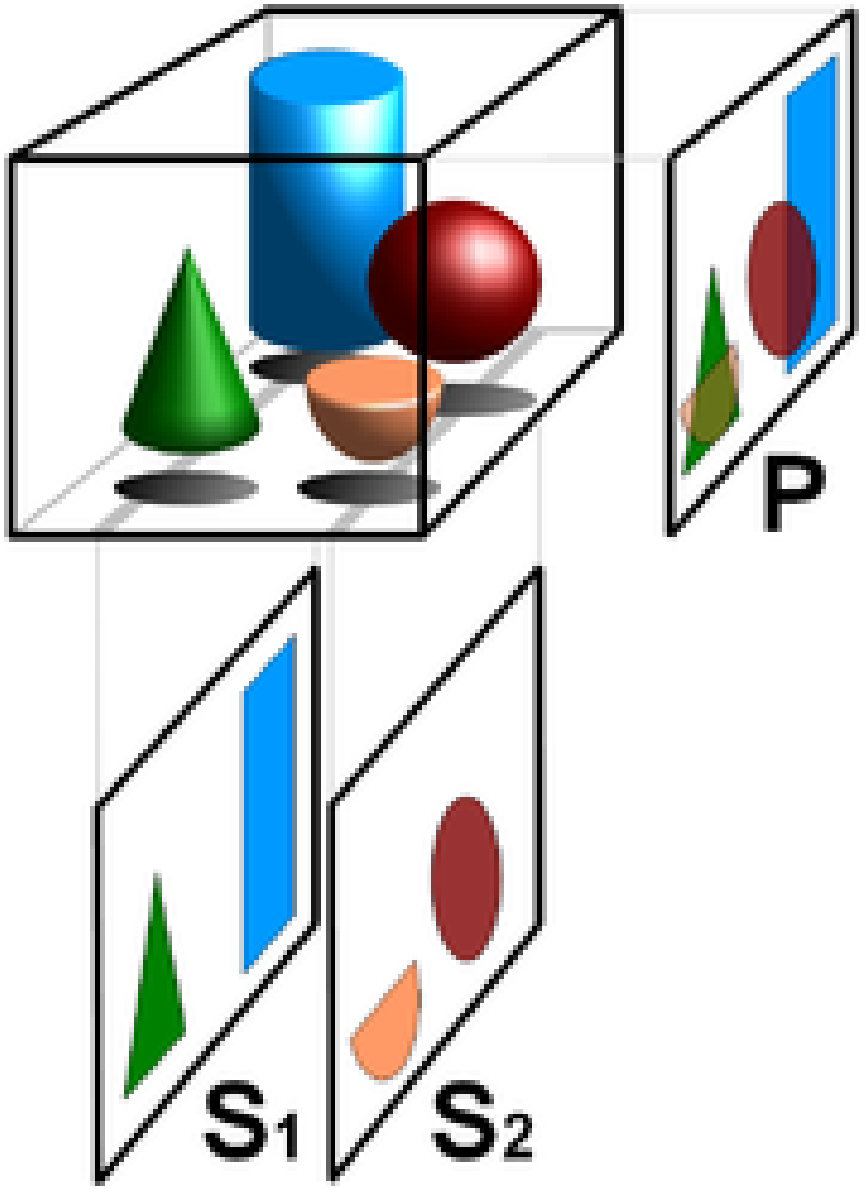
MPEG Encoding



Frame Types

I	Intra	Encode complete image, similar to JPEG
P	Forward Predicted	Motion relative to previous I and P's
B	Backward Predicted	Motion relative to previous & future I's & P's

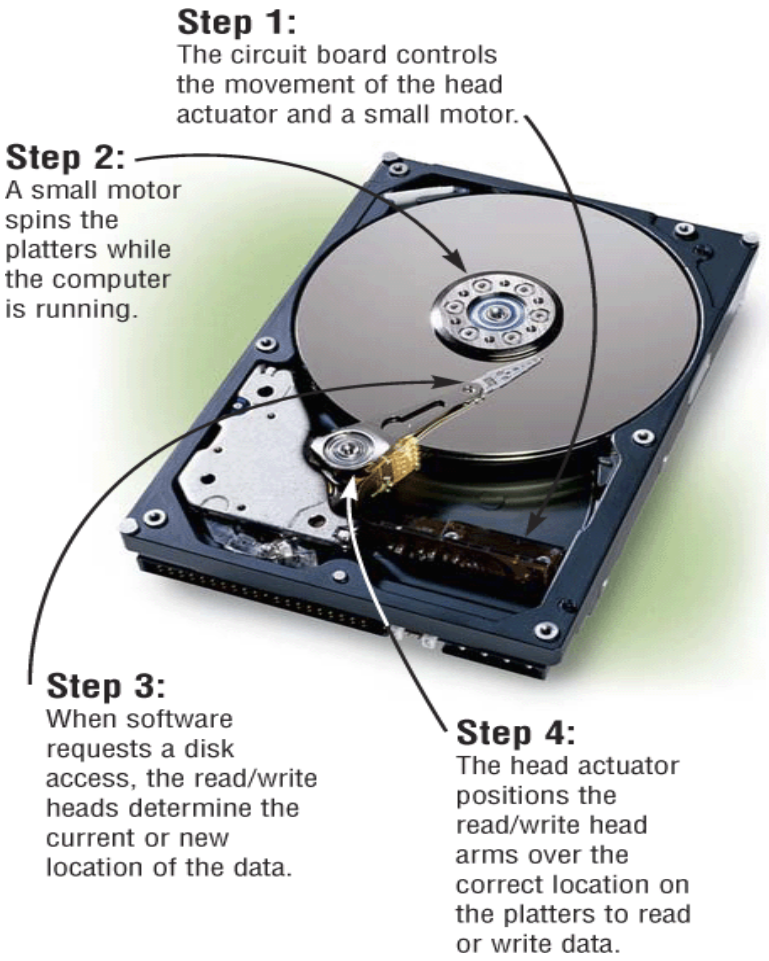
Volumetric Imaging

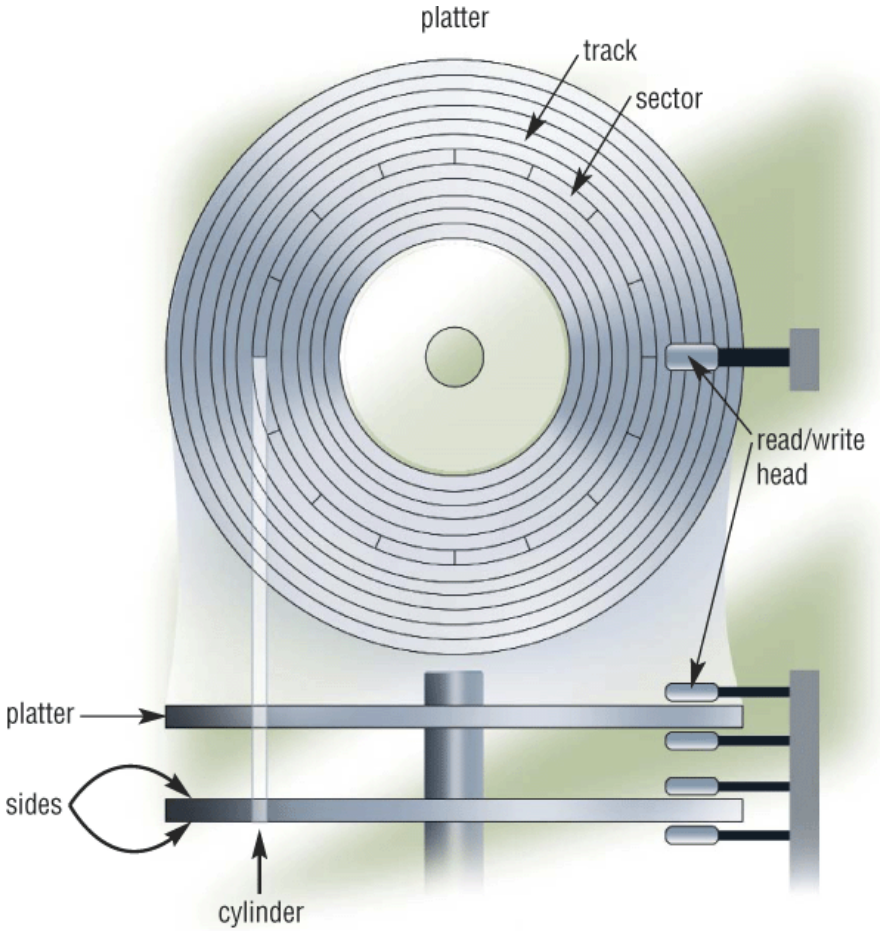


Rotating Storage Media

- Fixed magnetic disk
  - Hard drives
- Removable magnetic disk
  - Floppy disk
- Removable optical disc
  - CD, DVD, Blu-ray

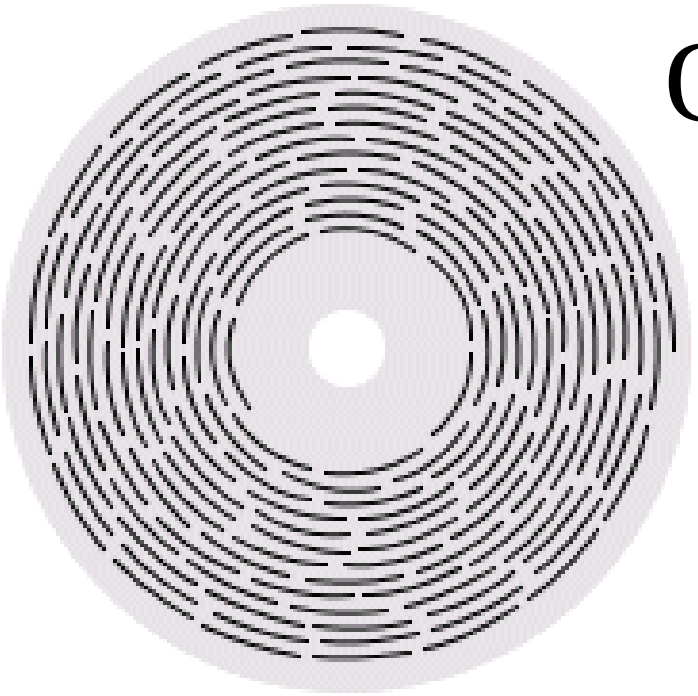
Magnetic Disk (Hard Drive)

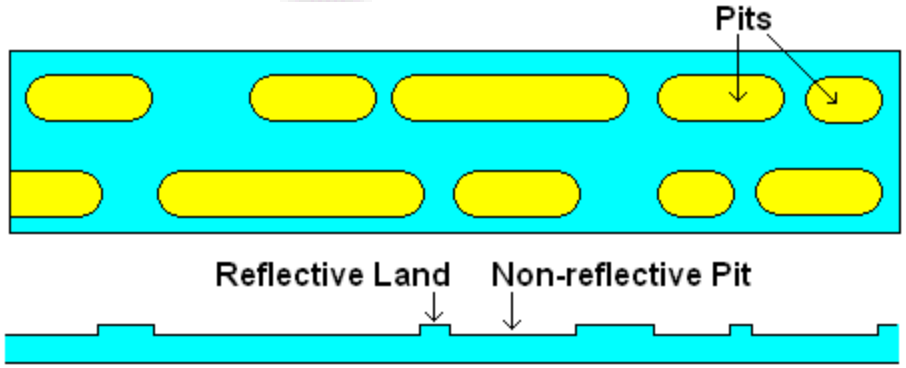


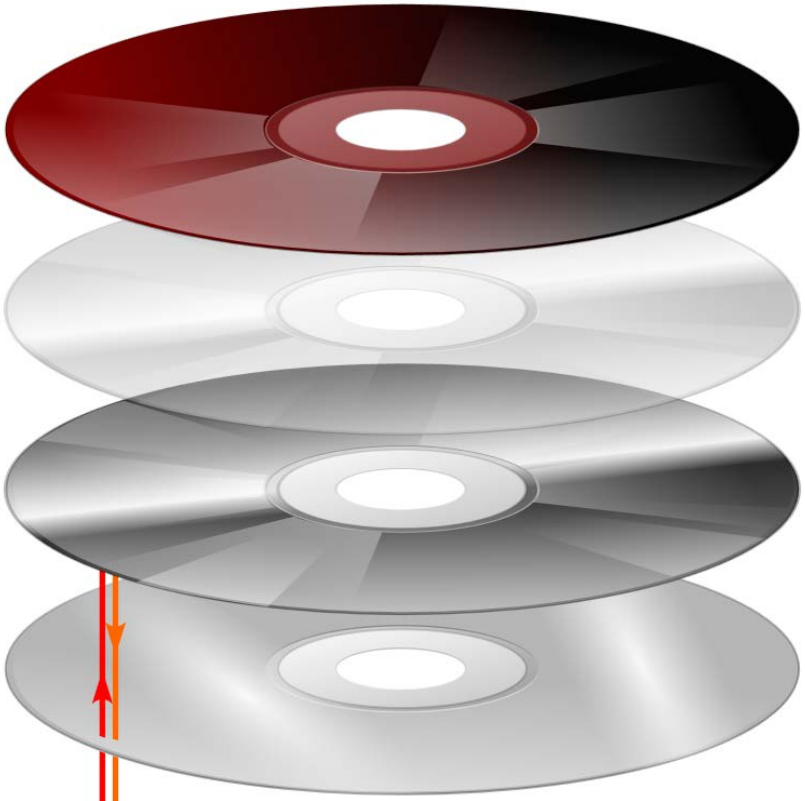


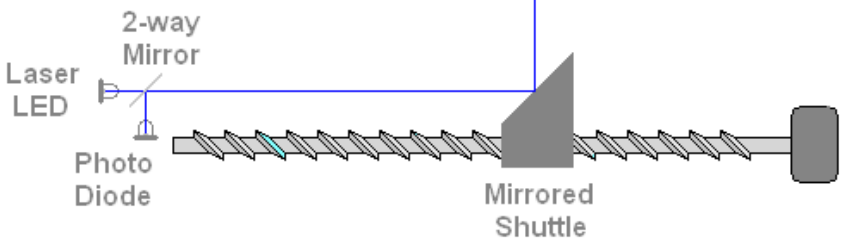
Shelly, Cashman and Vermatt, *Discovering Computers*, 2004

Optical Disc

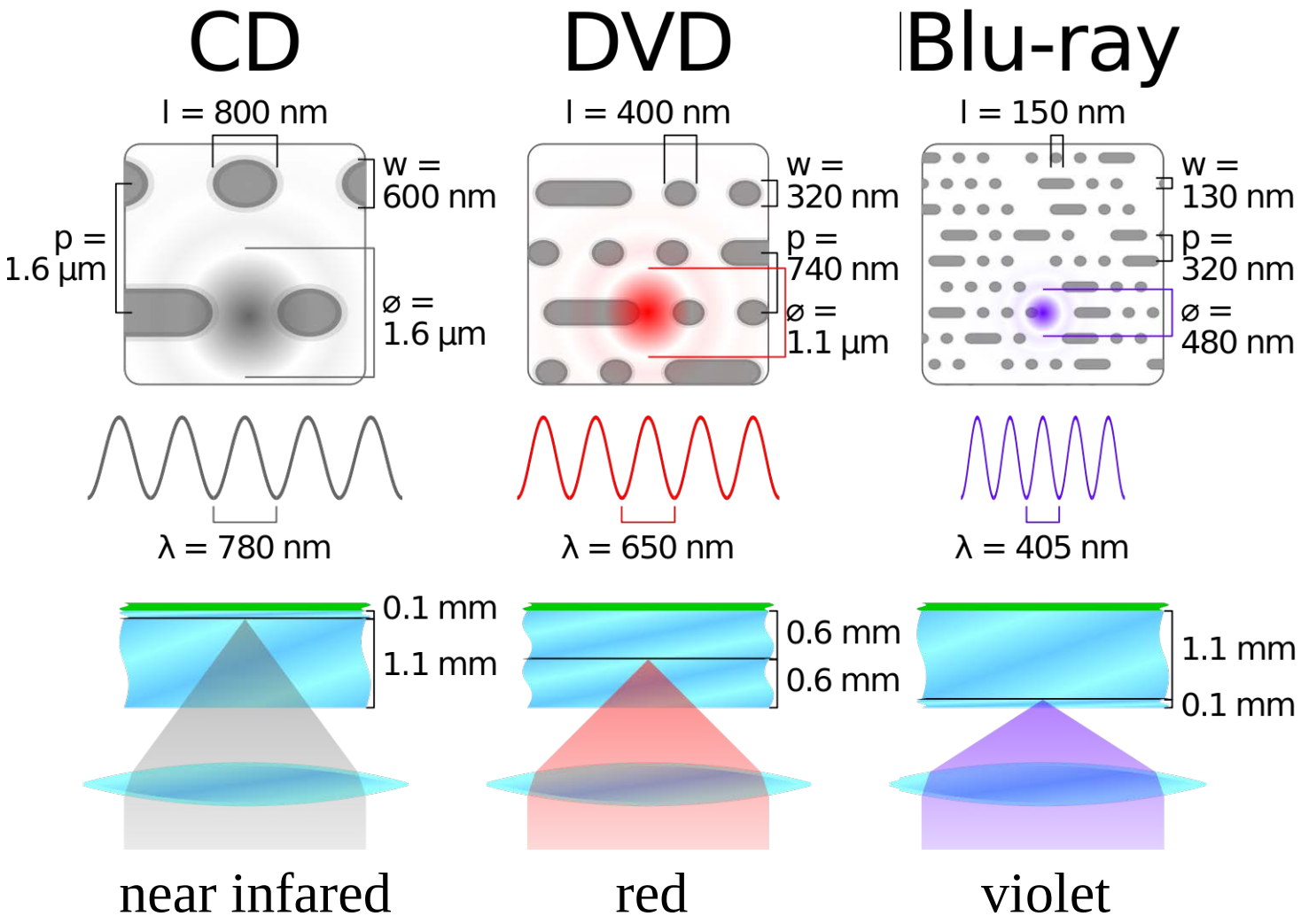








Optical Disk Technologies



Magnetic Tape

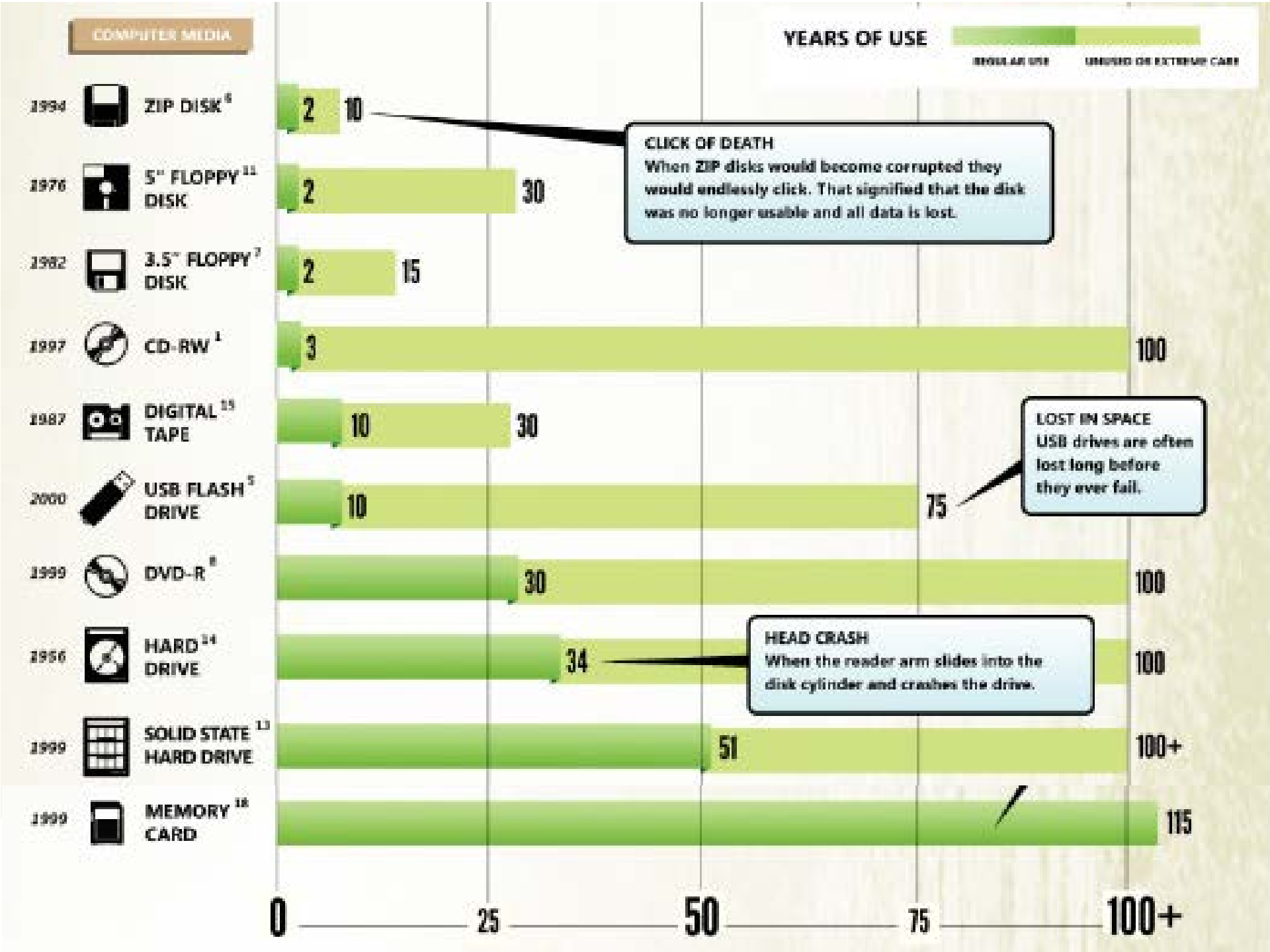
- Tapes store data sequentially
  - Fast transfer, but no practical “random access”
- Used only for low-use storage
  - Disaster recovery, offline storage

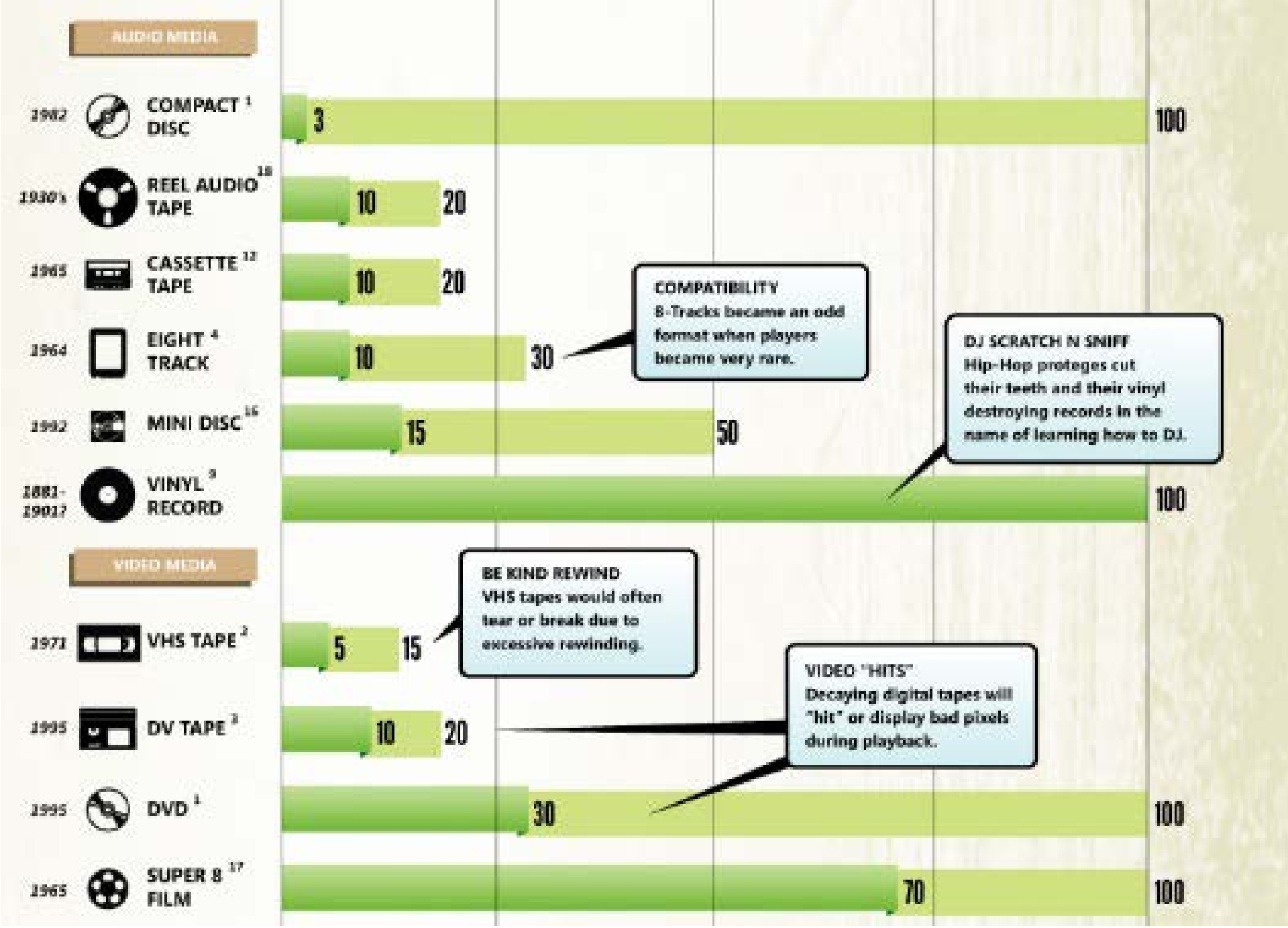
Solid-State Memory

- ROM
  - Does not require power to retain content
  - Used for “Basic Input/Output System” (BIOS)
- RAM
  - Cheap and fast, but works only while power is on
- Flash memory (Solid State Disk, memory sticks)
  - Much faster “random access” than rotating disk
    - On average, 10,000 times faster
    - About 10 times the cost per bit
  - Limited number of lifetime write operations (~5000)
    - But Zipf’s law permits “wear leveling”

Threats to Digital Collections

- Business decisions
  - Termination of service
  - Termination of infrastructure support
    - e.g., reading Amiga files, displaying Word Perfect
- Malfunctions
  - Hardware failure, operator error, software bugs, …
- Vandalism (hackers)
- Disasters
  - Physical risks to servers
  - Electromagnetic pulse


COMPUTER MEDIA
YEARS OF USE
REGULAR USE
UNUSED OR EXTREME CARE
1994 ZIP DISK [6]
2 10
1976 5" FLOPPY DISK [11]
2 30
1982 3.5" FLOPPY DISK [7]
2 15
1997 CD-RW [1]
3 100
1987 DIGITAL TAPE [15]
10 30
2000 USB FLASH DRIVE [5]
10 75
1999 DVD-R [8]
30 100
1956 HARD DRIVE [14]
34 100
1999 SOLID STATE HARD DRIVE [13]
51 100+
1999 MEMORY CARD [18]
115
CLICK OF DEATH
When ZIP disks would become corrupted they would endlessly click. That signified that the disk was no longer usable and all data is lost.
LOST IN SPACE
USB drives are often lost long before they ever fail.
HEAD CRASH
When the reader arm slides into the disk cylinder and crashes the drive.
0
25
50
75
100+



http://www.crashplan.com/medialifespan/

Media Migration

- What format should old tapes be converted to?
  - Newer tape
  - Rotating media
  - Solid state disks
- How often must we “refresh” these media?

Risk Management

- Redundancy drives down uncorrelated risk
  - Let *p* be the probability of loss of one copy
  - Then *p*p*p* is the chance of loss for 3 sites
  - Example: if *p=0.01* then *p*p*p=0.000001*
- Two fundamental problems:
  - Unanticipated correlation
    - For example, an operating system bug
  - Underestimated “black swan” probabilities

Layered Defense

- Good storage practices
  - Offline: Media migration
  - Online: uninterruptable power, RAID, backups
- Distributed storage
  - Storage Resource Broker (SRB), LOCKSS, …
- Air gaps
  - Interrupt unexpected correlation

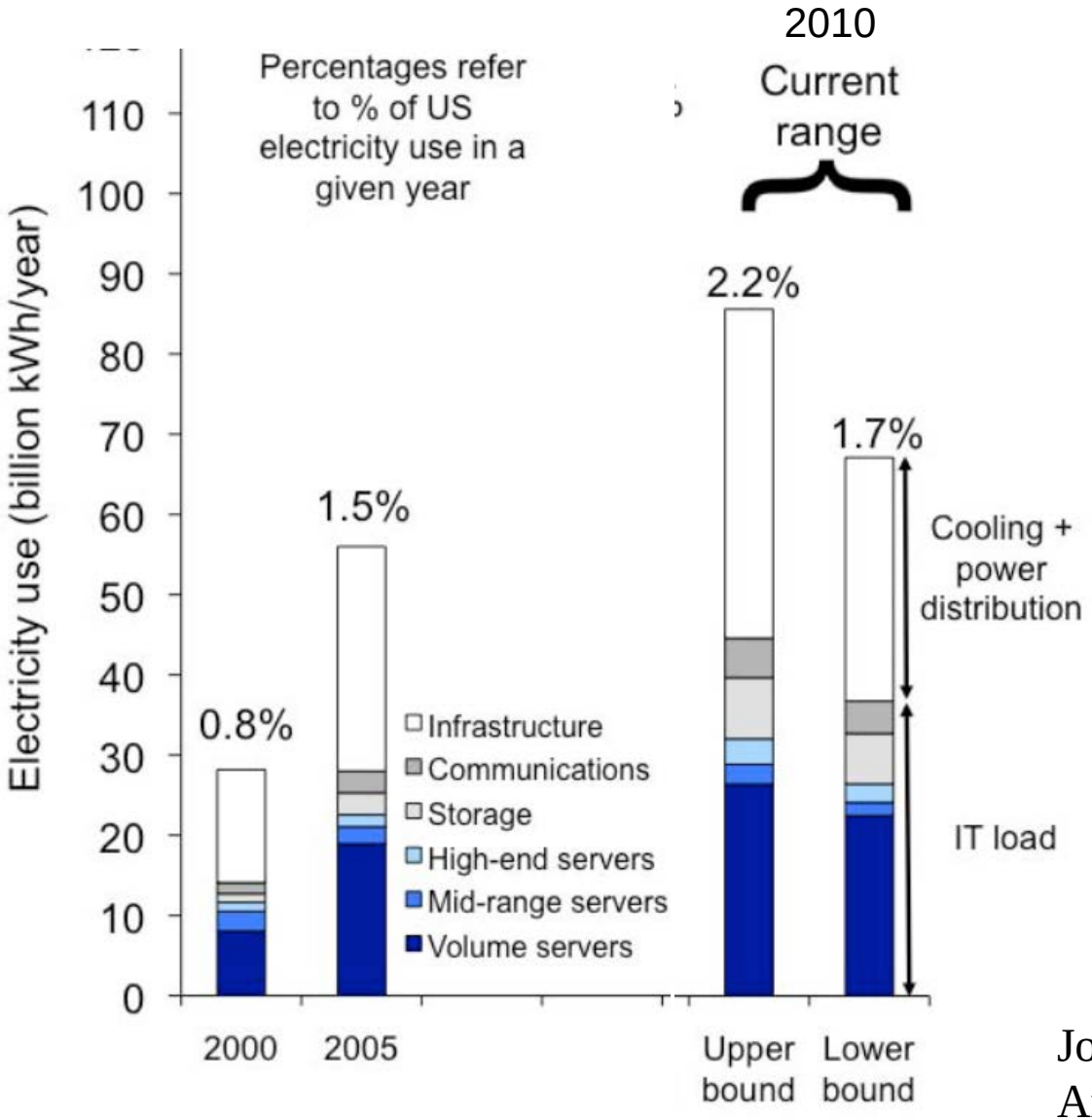
Data Centers

Source: Wikipedia

Shared Data Center Locations

http://www.datacentermap.com/usa/datacenters.html

Data Center Electricity Use (USA)

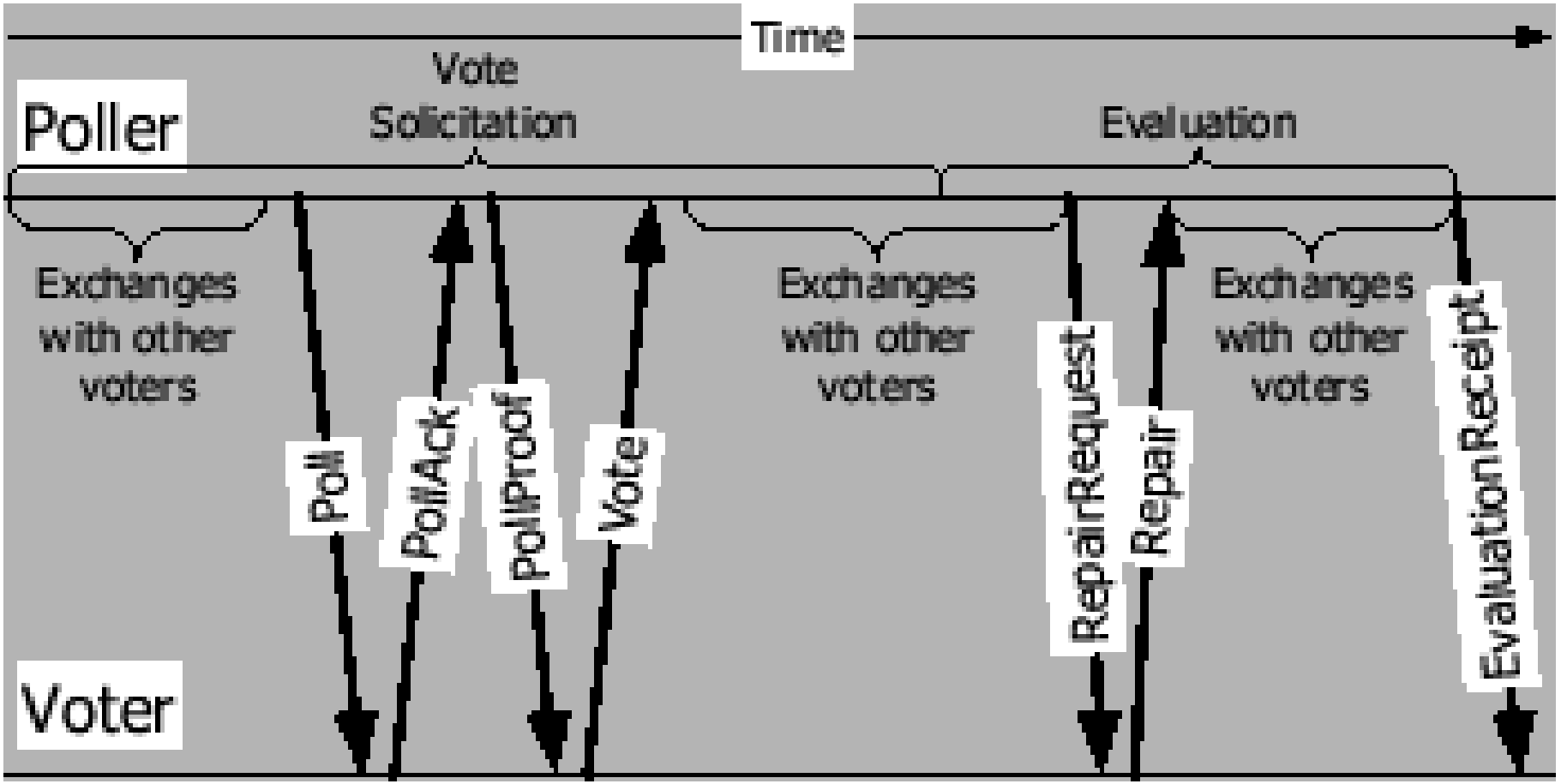


Jonathan Koomey, Analytics Press, 2010

Digital Federal Depository Library

http://lockss-usdocs.stanford.edu

LOCKSS Distributed Repair

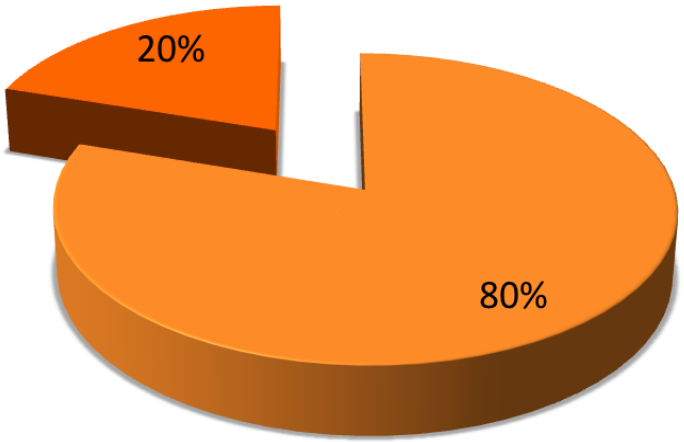


ITHAKA

- JSTOR digitization
  - Back runs of journals
  - Recently expanded to books
- Portico preservation
  - Centralized management, originally for journals
    - Release triggers: discontinuation, loss of access
  - Also service for books and datasets

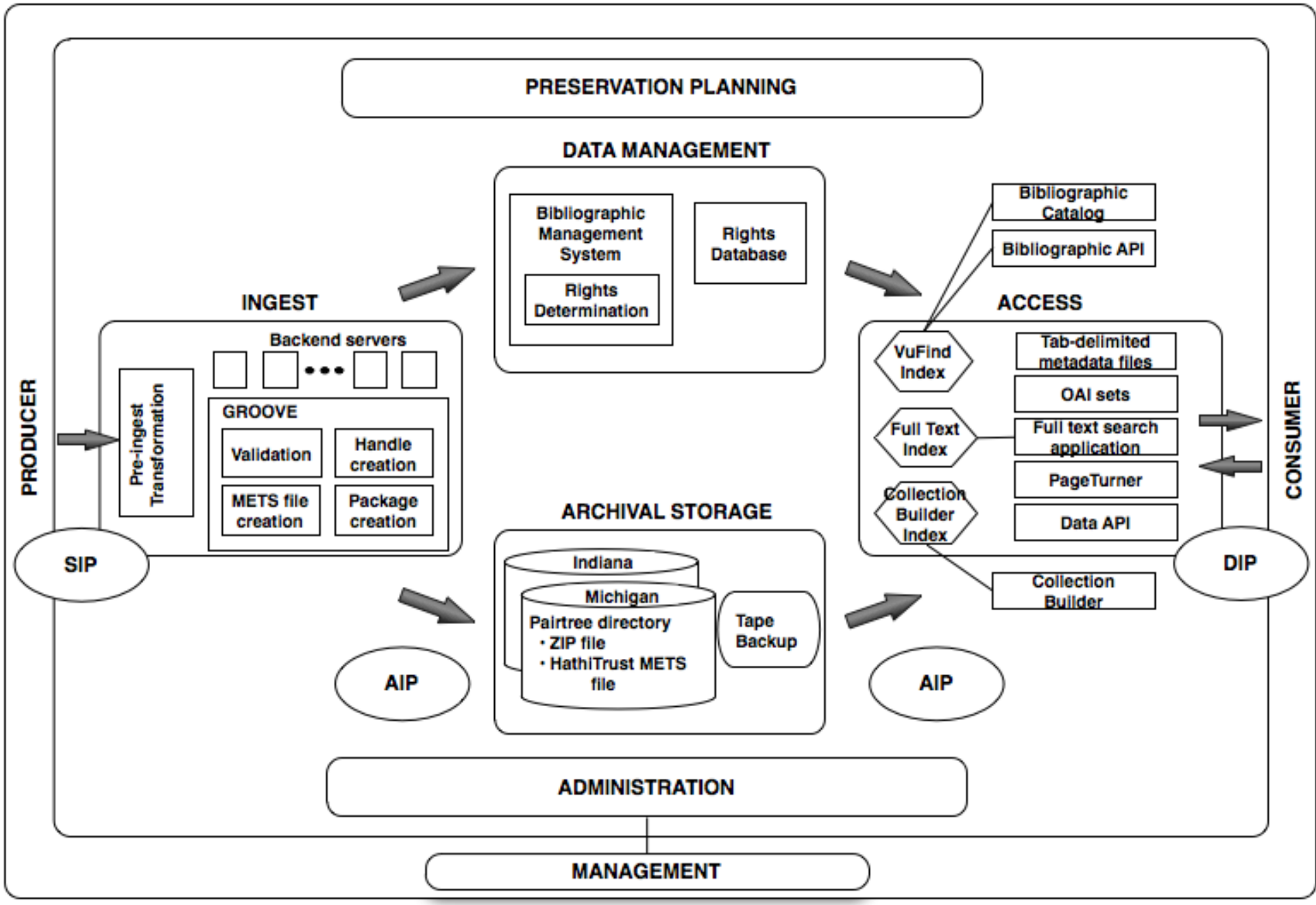
HathiTrust

- Centralized repository for digitized books
  - Google Books digitization (via owning libraries)
  - Microsoft book search (ran from 2006-2008)
  - Internet Archive
    - Million book project, project Gutenberg, contributions, …
  - Cooperative digitization



As of August 13, 2010

6,549,680 – Total volumes
1,300,896 – Public Domain
3,798,116 Book titles
153,311 Serial titles



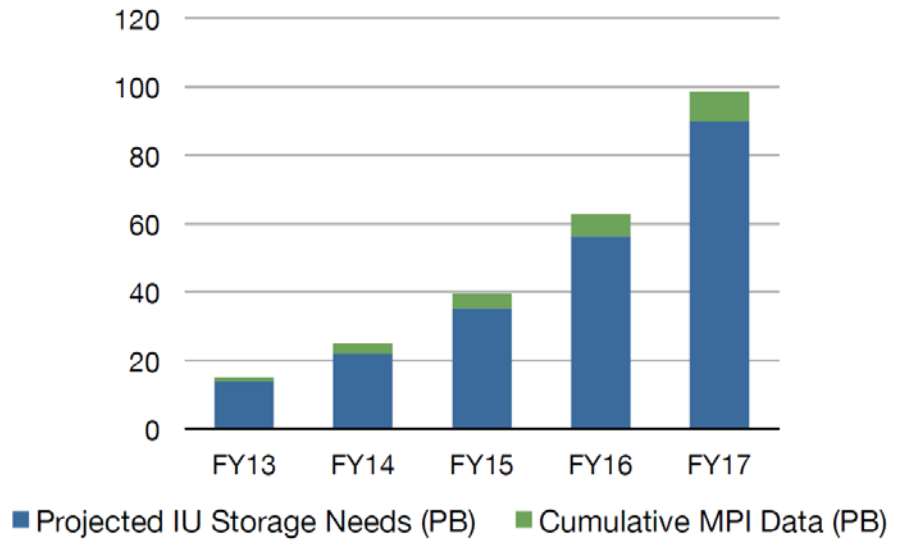
Jeremy York, IFLA 2010

Indiana University Digitization

Table 6: Media Preservation Targets, 2013-2027

Target	Hours	Objects	% of Total Holdings
15 Years— all media types	317,000	408,000	71%
Audio	207,000	284,000	82%
Video	83,000	66,000	53%*
Film (access digitization)	27,000	58,000	69%

**IU Bloomington video holdings include a large number of non-archival, commercial VHS tapes and DVDs that circulate primarily to students. These are not included here.*



Preserving Behavior

- Word processors
  - Formatting, track changes, undo deleted text
- Spreadsheets
  - Formulas, visualizations
- Databases
  - Queries, forms, derived values
- Computer-Assisted Design (CAD)
  - Display, modification, manufacturing
- Software
  - Simulation, games, embedded systems, …

Behavior Preservation Strategies

- Format migration
  - For example, convert Word Perfect to PDF
- Emulation
  - Allows running old software on newer systems

Apollo Guidance Computer Emulation

AGC Simulation Type

Guidance Computer (AGC) software

- ○ Apollo 1 Command Module
- ○ Apollo 7 Command Module
- ◉ Apollo 8 Command Module
- ○ Apollo 9 Command Module
- ○ Apollo 9 Lunar Module
- ○ Apollo 10 Command Module
- ○ Apollo 10 Lunar Module
- ○ Apollo 11 Command Module
- ○ Apollo 11 Lunar Module
- ○ Apollo 12 Command Module
- ○ Apollo 12 Lunar Module
- ○ Apollo 13 Command Module
- ◉ Apollo 13 Lunar Module
- ○ Apollo 14 Command Module
- ○ Apollo 14 Lunar Module
- ○ Apollo 15-17 Command Module
- ○ Apollo 15-17 Lunar Module
- ○ Apollo Skylab/Soyuz Command Module
- ○ Validation Suite
- ○ Custom: [ ] [...]

Interfaces

- ☑ Guidance Computer
- ☑ DSKY (AGC display and keypad)
- ☑ Attitude Controller Assembly [Handler]
- ☑ Telemetry Downlink Monitor
- ☑ LM Abort Computer (AEA)
- ☑ DEDA (AEA display and keypad)
- ☐ AGC CPU Bus/Input/Output Monitor
  - ☐ Inertial Monitor Unit / FDAI (8-ball)
  - ☐ Discrete Outputs
  - ☐ Discrete Inputs (crew)
  - ☐ Discrete Inputs (LM system)
  - ☐ Propulsion/Thrust/Fuel Monitor

[Novice] [Expert]

Browse Source Code

[AGC] [AEA]

Options

AGC Startup

- ◉ Restart program, wiping memory
- ○ Restart program, preserving memory
- ○ Resume from ending point of prior run
- ○ Custom: [ ] [...] [Save]

Interface styles

DSKY:	◉ Full	○ Half	○ "Lite"
Downlink:	◉ Normal	○ "Retro"	
DEDA:	◉ Full	○ Half	

Use AGC/AEA debugger?

AGC code:	◉ Normal	○ Debugger
AEA code:	◉ Normal	○ Debugger

LM Abort Computer (AEA) software

- ○ Apollo 9 (Flight Programs 3, 4)
- ○ Apollo 10 (Flight Program 5)
- ◉ Apollo 11 (Flight Program 6)
- ○ Apollo 12-14? (Flight Program 7)
- ○ Apollo 15-17 (Flight Program 8)
- ○ Custom: [ ] [...]

[Run!] [Defaults] [Exit]

http://www.ibiblio.org/apollo/

An Integrated Strategy

- Delay decay of organic materials
  - But balance costs and benefits
- Balance quality and scale
  - Preservation: rescue at-risk collections
  - Access: Quantity has a quality all its own
- Design in diversity
  - Technologies, risk exposure, institutions
- Adequately resource the process

Before You Go!

- On a sheet of paper (no names), answer the following question:

  What was the muddiest point in today's class?