Sustainability Issues in Digital Preservation

Krishna Kant
George Mason University
National Science Foundation

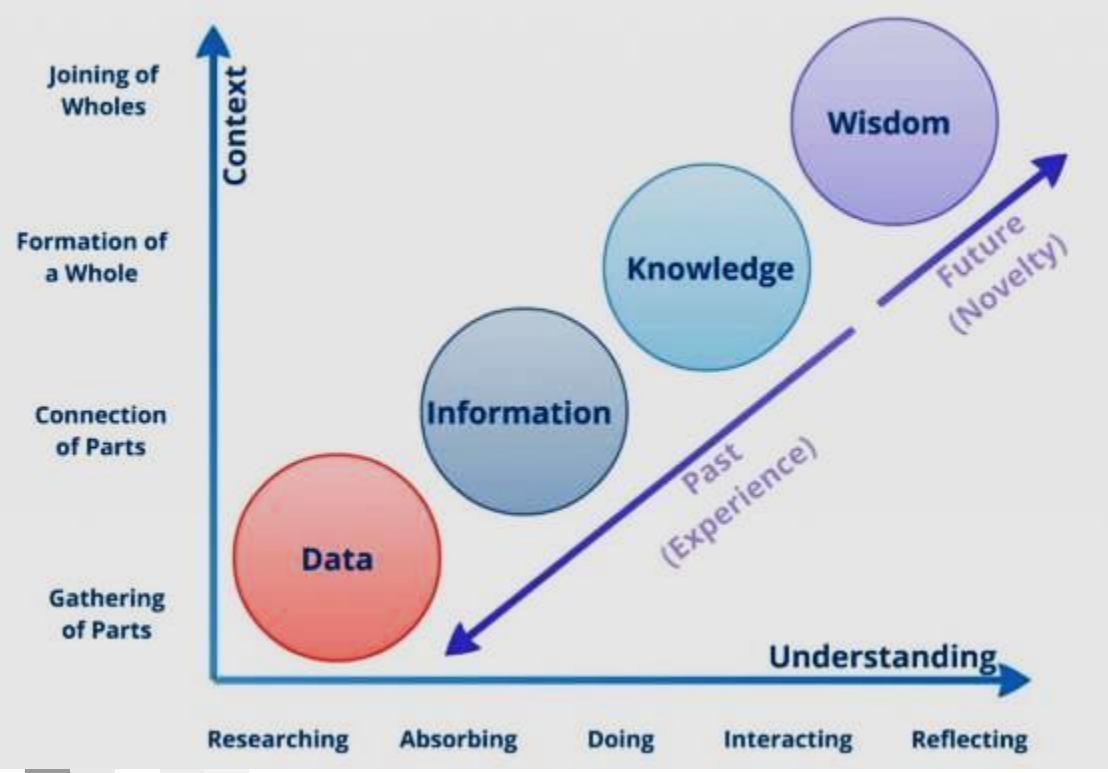
Digital Preservation 2013
Alexandria, VA, July 24, 2013

Can Preservation be Sustainable?

- Significant environmental footprint.
  - Storage media life-cycle (mfg, distribution, …)
  - Housing & access management of media
  - Processing & IO infrastructure (Data Centers)

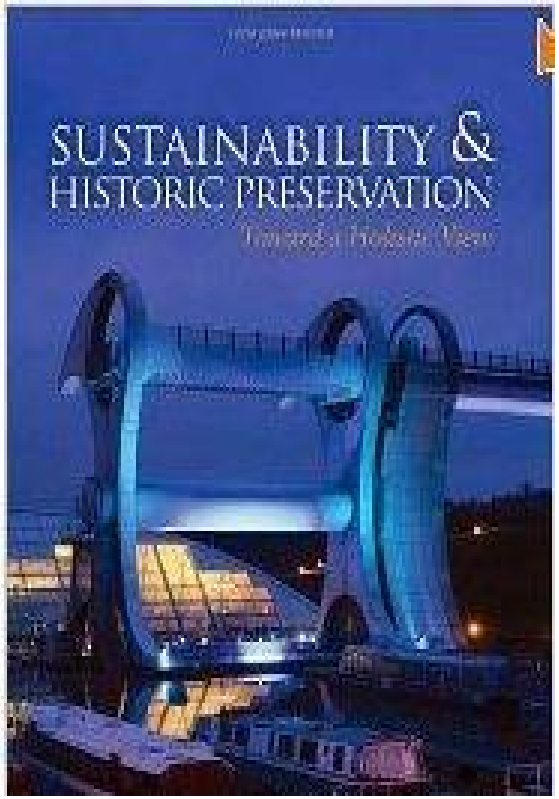
What do we want to preserve?

- Not data, but information
- Ideally, knowledge



Sustainability Issues

- Make knowledge derivation more sustainable
  - Minimize environmental impact of data centers
- Retain only essential data
  - Remove duplicate, inessential data
- Storage vs. reprocessing
  - Has sustainability tradeoffs

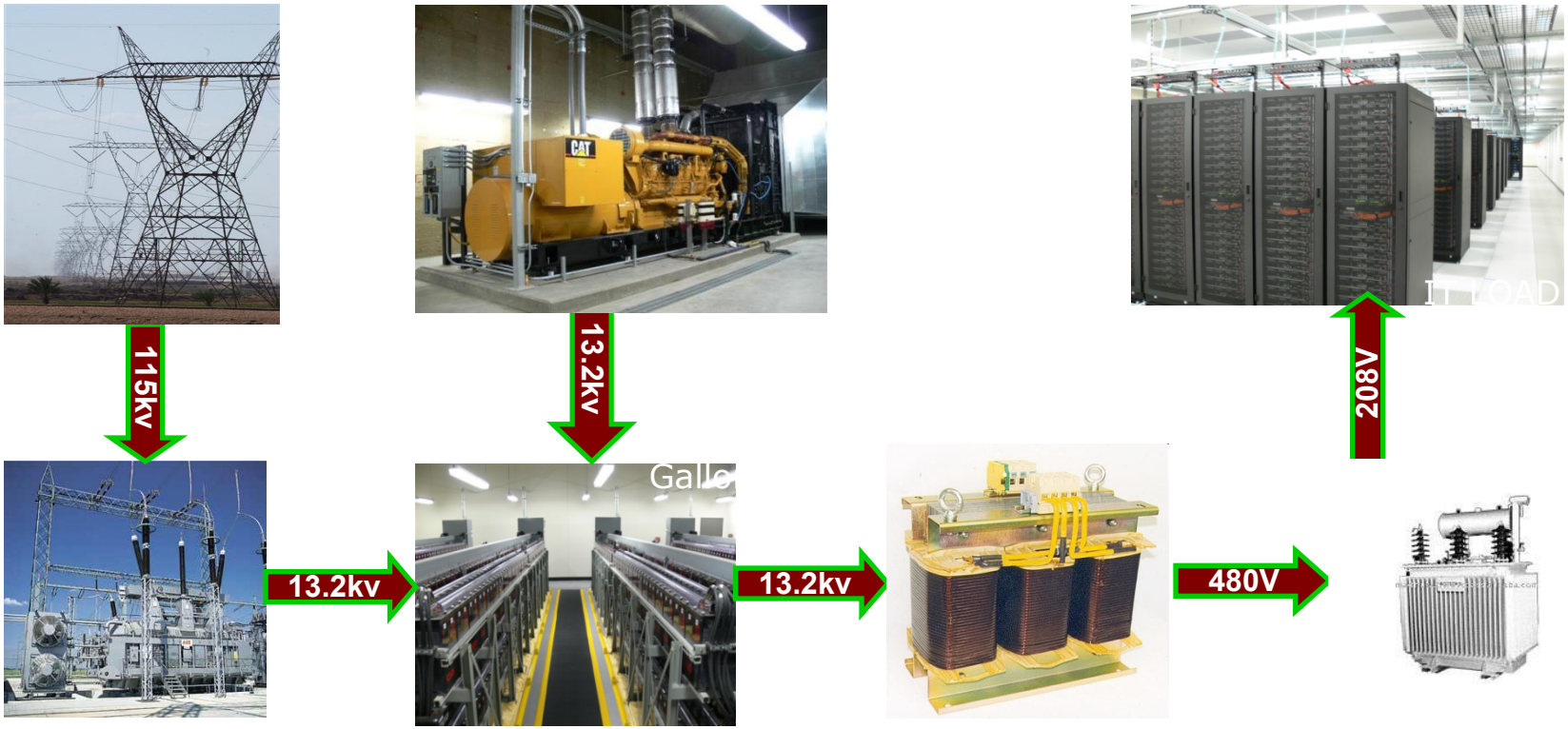


Data Center Impact

- Power consumption rising.
- Most of it wasted:
  - Power distribution,
  - Cooling
  - Idle Machines

- But impact is more than power…
  - Materials, water, manufacturing, …
- Sustainability perspective
  - Energy doesn't matter, its carbon footprint does

Power Distribution Infrastructure



- 9-10% distribution loss at power source
- Lots of earth's resources used (metals, rare earths, …)

Renewable Energy Powered IT?

- Limit energy draw from grid
  - Less infrastructure & losses, but variable supply
  - Impact on performance, QoS, SLA, …
- Challenges
  - Variability at multiple time scales
  - Reliability issues

Need better power adaptability

Cooling Infrastructure

- Cooling is very resource intensive
  - Lot of materials
  - Water, much of which evaporates

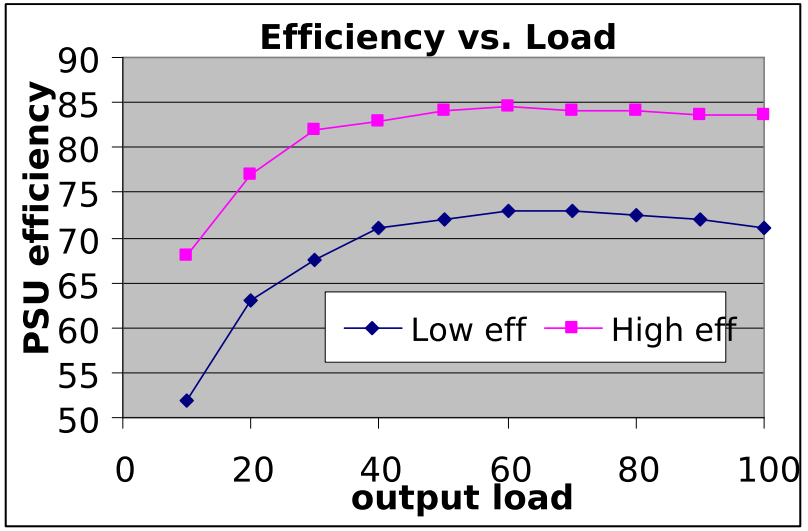
High Temperature Operation

- Chiller-less data centers
  - Less energy/materials, but space inefficient
- High temperature operation of comm. /computing equipment
  - Smaller $T_{outlet} - T_{inlet}$
  - Deal with occasionally hitting temp. limits.

Need smarter thermal adaptability

Overdesign

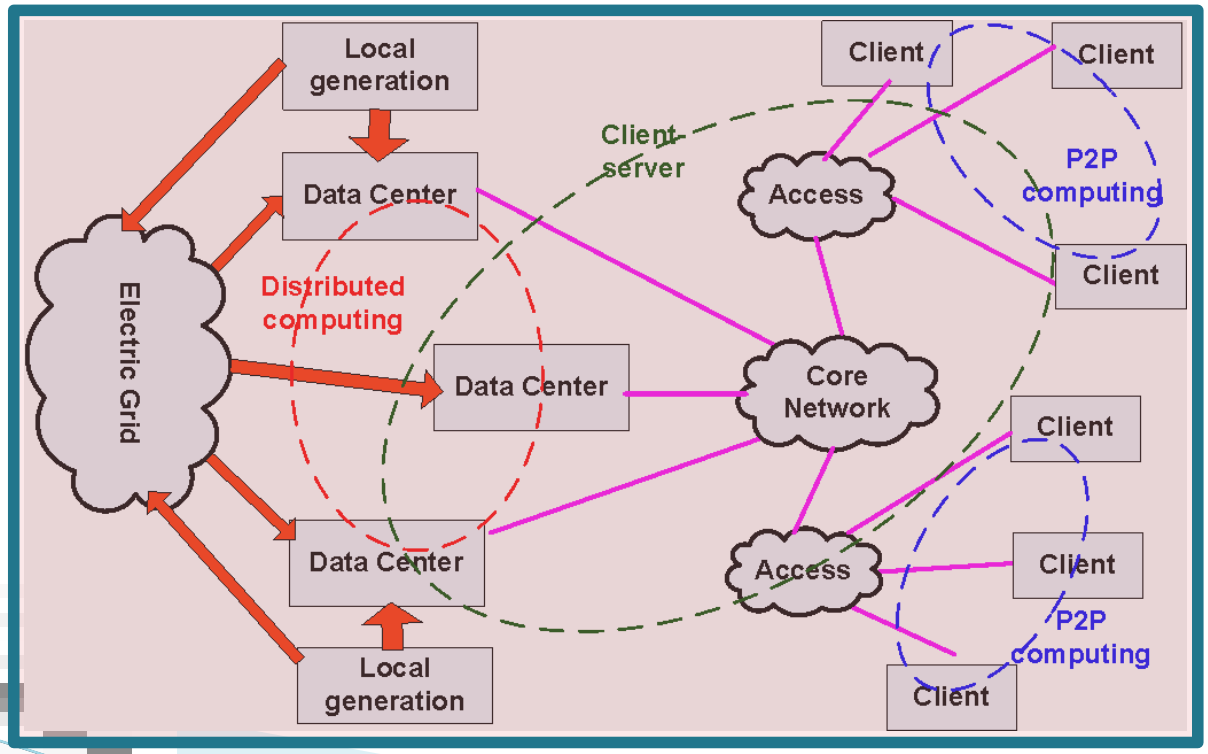
- Huge UPS, Generators, dist. frames, power supplies, fans, …
- Engineered for worst case
  - Huge waste of power, materials, …
- Power Supply & VRs
  - Low utilizations ➔ Low efficiency



Need Better Power Infra. adaptability

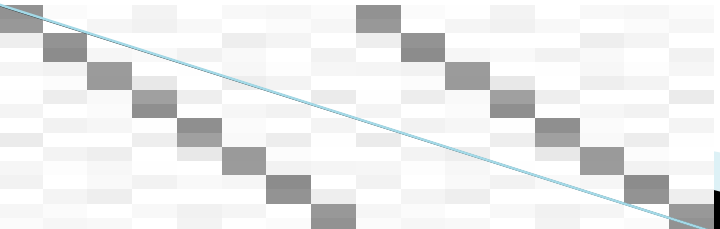
Energy Adaptive Computing

- Overdesign ➔ Rightsizing + smart adaptation
  - Adaptation to energy/power/thermal/cooling limitations.
- Dynamic adaptation of infrastructure & workloads
  - Need coordination across compute, network & storage.



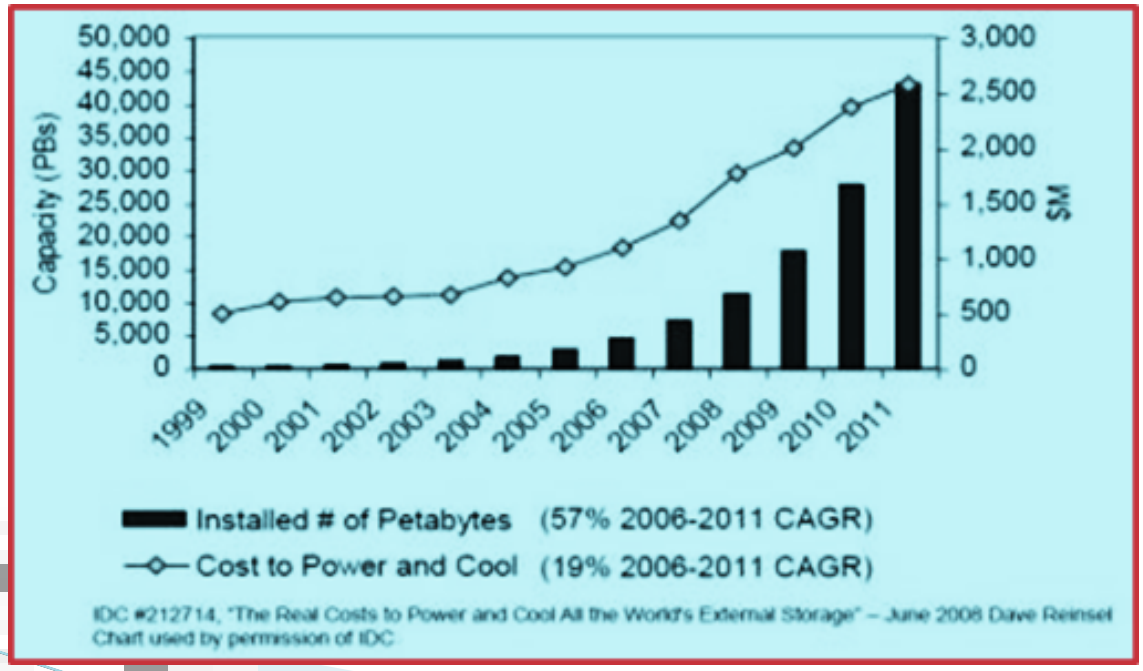
Data Growth

- Exponential growth in both generation & retention
- Data vs. useful data
  - More data ➔ More junk (Less information)
  - Duplication, Keep just in case, too lazy to purge, …
- But, can we define useful?

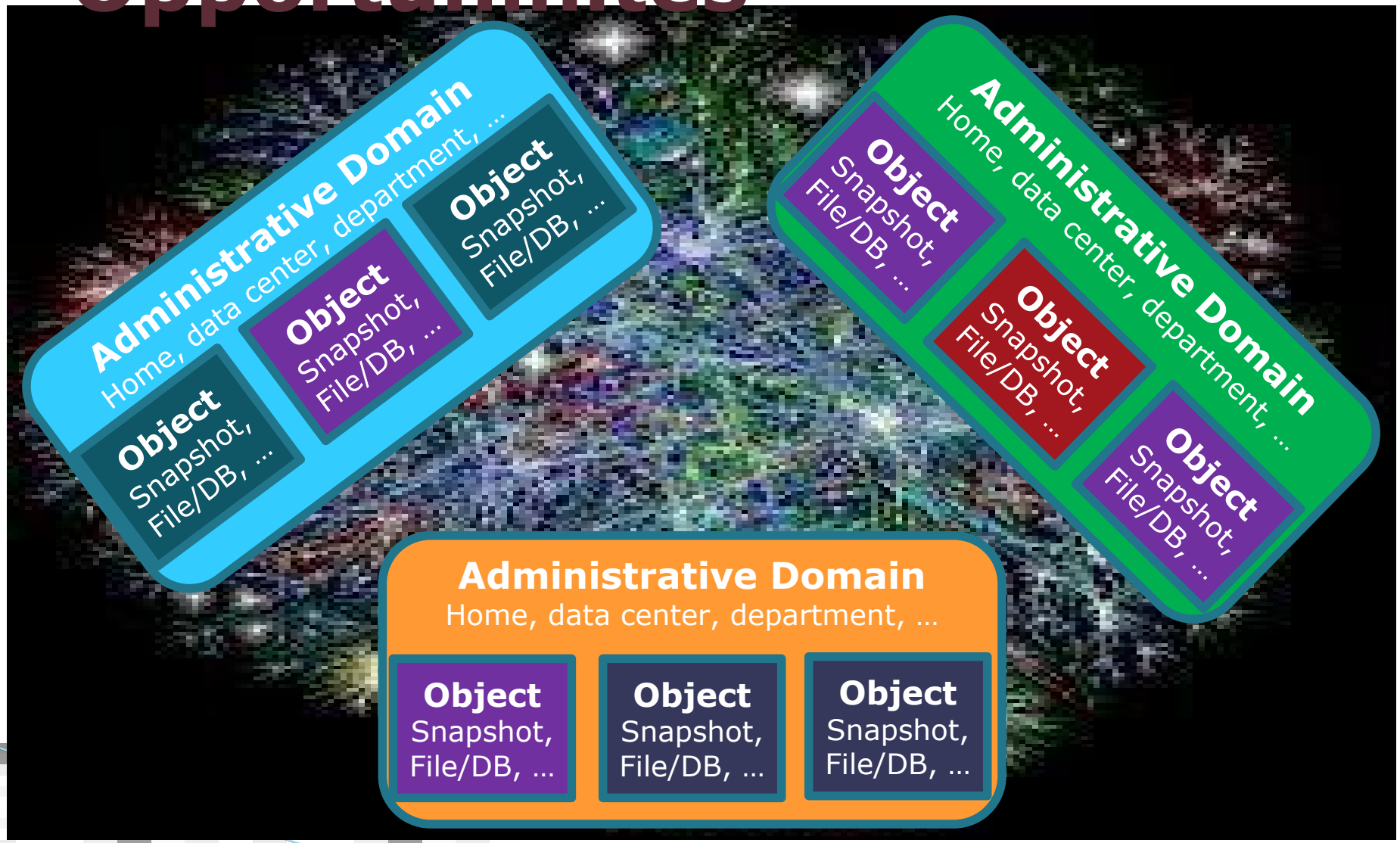


Sustainability Impact of Data

- Increased power consumption – 10-40% of total power
- Insatiable drive demand ➔ Life cycle impact
- Cumulative impact because of little deletion
- Not sustainable!



Data Reduction Opportuninites



Data Reduction

- Within an object
  - Compression, compressive sampling, delta encoding, remove bundled VM
- Within and across administrative domain
  - Deduplication across objects & storage nodes
  - Cloud based storage/access/deduplication
  - Only a few copies (collectively) across nodes
- Tradeoffs
  - Storage vs. data movement vs. processing
  - Fidelity vs. cost (reduced representations)
    - Similar, filtered, derived, …
  - Cross domain access/privacy/security issues

Role of Content Creators

- Best Practices
  - Send link instead of content
  - Don’t create local copy
  - Purge obsolete, defective, unneeded data

- Data is valuable, meta-data is precious!
  - Designed, not an afterthought!
  - Strong association with data
  - Must reflect data quality
  - Preservation more crucial than for data

Thank you!

Other Sustainability Aspects

- Physical degradation of media
  - Will require keeping media healthy for exponentially increasing data ➔ Unsustainable
- Media obsolescence
- HW & SW obsolescence
- Increasing amount of materials to manufacture storage media
- Unable to power the media
- Loss of meta-data or inadequate meta-data
  - Have the data but don’t know how to use it/