Storing Research Data "Forever"

Funding Long-Term Preservation of Research Data

CNI: 12/13/2010

Why Store Data "Forever"

- Because we have to:
  - Funding agencies want data "sharing" plans
  - NIH Data Sharing Policy (2003):

  http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

  "all investigator-initiated applications with direct costs greater than $500,000 in any single year will be expected to address data sharing in their application."

NIH Data Sharing Policy

- "Applicants may request funds for data sharing and archiving. The financial issues should be addressed in the budget section of the application."
- Specifics depend on grant, published in RFP, RFA or PA

NSF Data Sharing Policy

Beginning January 18, 2011, proposals submitted to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. See Grant Proposal Guide (GPG) Chapter II.C.2.j for full policy implementation.

Other agency Policies

- See Gary King's Page on "Data Sharing and Replication"
- http://gking.harvard.edu/replication.shtml
- See National Academy of Sciences "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age", July, 2009
- http://www.nap.edu/catalog/12615.html

What is Data?

- Numbers?
  - Recorded? Collected? Generated?
- Images? Video? Audio?
  - Shoah
  - In what format?
- Code?
- Publications/Text?
  - In what format?
- Transcription service
- Is pure "raw" data useful
  - May require extensive meta-data to be useful

What is "Forever"?

- Longer than a typical project?
- Longer than a typical career?
- Longer than a typical institution?
- 5 years, 10 years, 25 years, 100 years?
- Suggestion: treat data same way library treats books
    - Intent is to preserve indefinitely
    - As long as practical, feasible
    - Cannot be precisely defined

Why Save Data "Forever"

- Because we want to:
  - Available to ourselves and our students and colleagues
    - Where are the data sitting today? On a departmental server? On a computer under your desk? On a CD or DVD somewhere?
    - Where is your dissertation data?
  - Available to future scholars, including ourselves

Why Save Data "Forever"

- Because we need to:
  - Encourage honesty?
    - Gregor Mendel probably cheated
  - Like open-source: help uncover mistakes, bugs?
  - Open Data Movement
    - Mostly library/catalog data, map data, WordNet
  - Open Access Movement
    - Mostly publications
- Because it's not "our" data

Current Storage Models

- Let someone else do it
  - Government agency/lab/bureau
    - NOAA National Geophysical Data Center
    - GenBank (DNA data)
    - fMRIDC (fMRI publications and data)
    - NCSA Astronomy Digital Image Library

Current Storage Models

- Professional society/Journals
  - Global Ocean Observing System: coordinates distributed data
  - Dryad: ecology/evolutionary biology
- Nice folks at another University
  - ICPSR, University of Michigan (political/social)
  - Dryad: ecology/evolutionary biology
  - Protein Data Bank (PDB): 3-D protein data
  - NCSA Astronomical Image Library
  - Sloan Digital Sky Survey
- The "Cloud"

Current Funding Models

- Institution/department pays
- Grants pay monthly/yearly
- Haphazard
  - Some grant money
  - Some departmental money
  - Use whatever is available
  - Don’t worry, someone will pay

Current Funding Models

- Most require some form of on-going payment
- Advantages
  - Capitalist approach to data storage
  - If someone wants to pay, data gets saved
  - "Natural" expiration process
- Disadvantages
  - Capitalist approach to data storage
  - Who pays to save rarely used data?

Different Approach

PAY ONCE, STORE ENDLESSLY (POSE)

Why Pay Once?
- Grants expire often and quickly
- Researchers expire pretty often

How Store Forever?
- Administrators expire slowly
- Institutions expire rarely

The Business Model (1)

- I = Initial cost of storage
- D = rate at which storage costs decrease yearly, expressed as a fraction (e.g., 20% would be 0.2)
- R = How often, in years, storage is replaced
- T= Cost to store the data "forever"

$$T = I + (1-d)^{r} * I + (1-d)^{2r} * I + \ldots.$$

If d=20%, r = 4:

$$T = I + (.8^{4})* I + (.8^{8})* I + \ldots.$$

The Business Model (2)

If d >0,

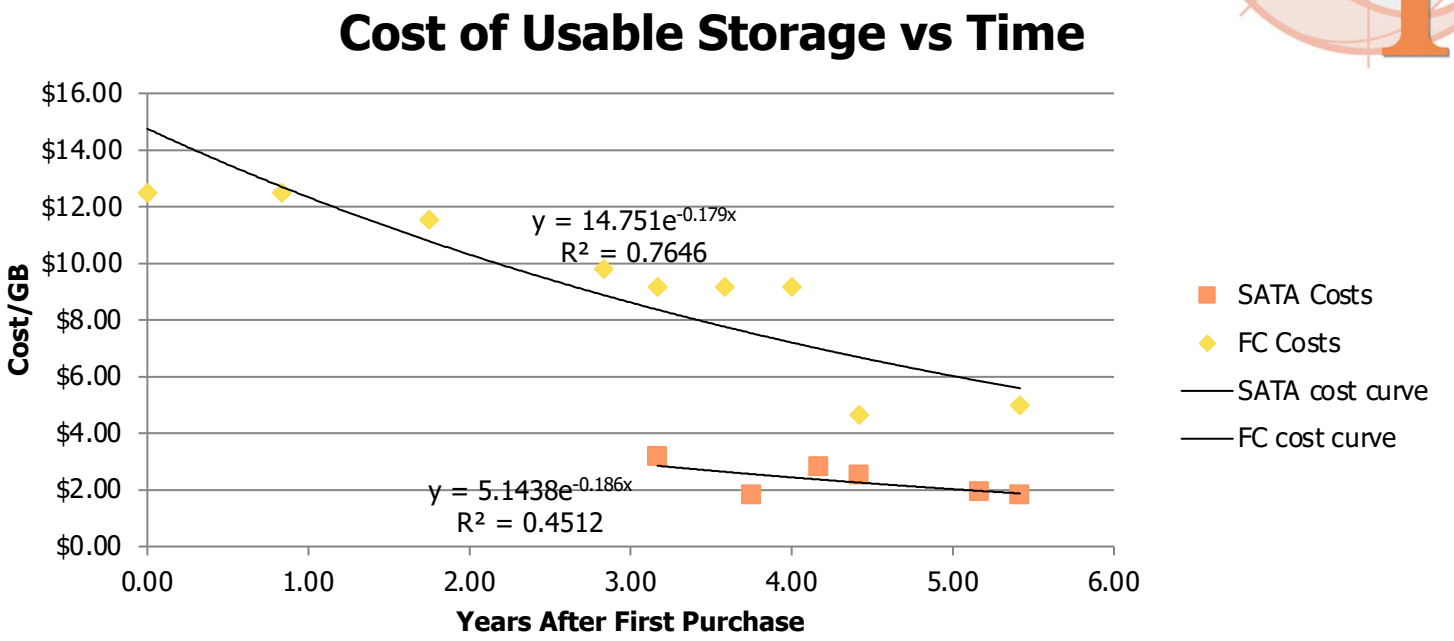
$T = I + (1-d)^r * I + (1-d)^{2r} * I + \ldots$
$= I/(1-d)^r$

The series CONVERGES!

For d=20%, r = 4: T=I * 2

Charge 2x initial storage cost, save half, store forever!

An Example: DataSpace at Princeton



- FC costs decrease by about 16% per year
- SATA costs decrease by about 17% per year
- Additional savings every few years from new storage

CNI: 12/13/2010

The Model for DataSpace

- SATA cost = $1.81/gb
- Replace every four years
- Costs decrease by 20% year

T = 1.81/(1-.8 **4) = $3/gb

Adding tape backup jumps this to $5/gb

$5K one-time to store a terabyte forever.

What about People?

- Studies: 5% of total data preservation costs for disk drives.
- Bulk of cost is for "people" (staff).
- True only if people costs are added to every storage request.
- "Marginal" people costs decrease as quickly as disk drives.
- Staff 20 years ago->1 gig; today-> 1 petabyte

Operational Model

- Must minimize ancillary costs
- Keep these to a minimum
  - Minimal boutique" services"
  - Customer pays for data curating or delivery if not web based.
- No re-use of disk space.
- ALL data public; no special handling.

Want to know more?

- **http://arks.princeton.edu/ark:/88435/dsp01w6634361k**
- **dataspace.princeton.edu**
- **serge@princeton.edu**