

ISTeC Data Management Forum #3 'Big Data Miasma'

ISTeC Data Management Forum #3

Pat Burns

Dean of CSU Libraries & VP for IT

Friday, May 2, 2014

Miasma

- ▶ “A dangerous, foreboding, or deathlike influence or vapor.”

Theme

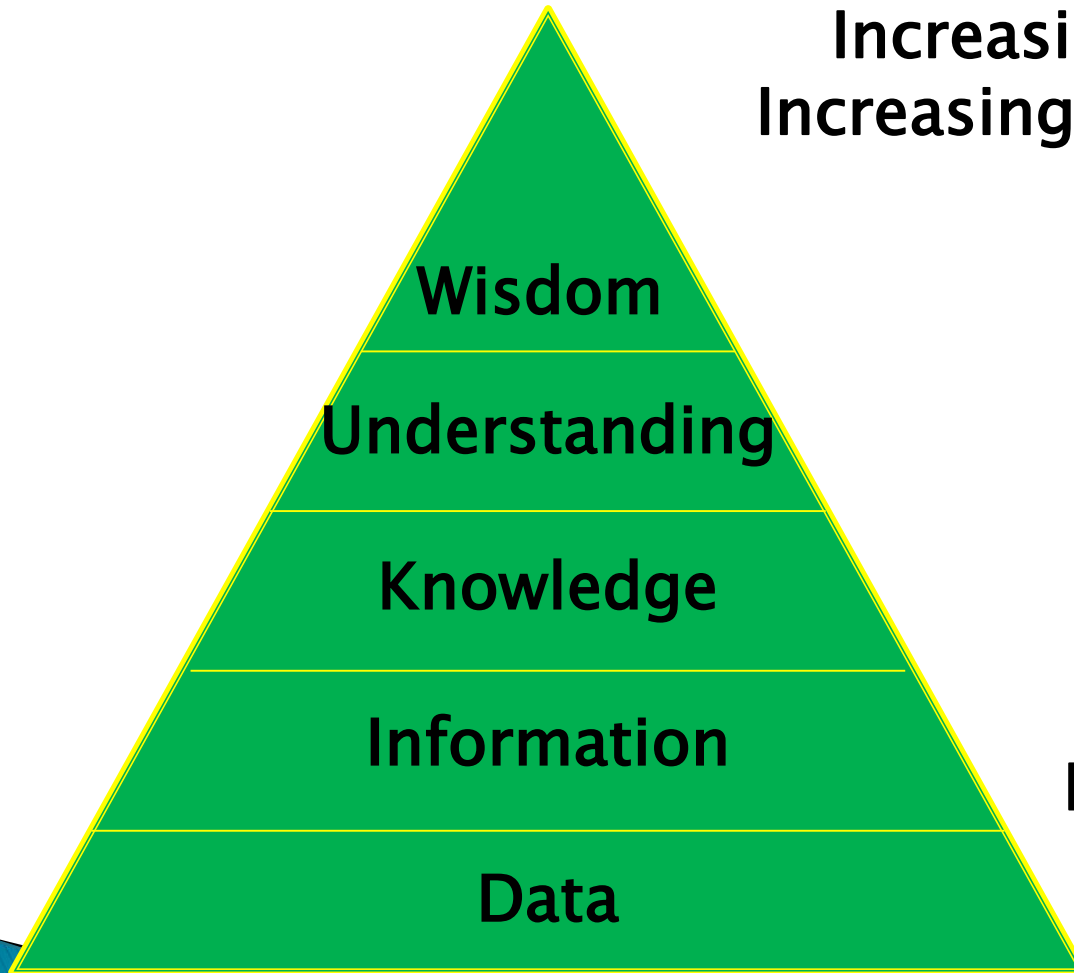
- ▶ “We are drowning in information and starving for knowledge.” – Rutherford D. Rogers

Intended Outcomes Today

- ▶ Understand needs for data management
- ▶ Understand support for data management
 - Local IT infrastructure
 - Central IT infrastructure: Globus
 - Libraries' Digital Repository, DM support
 - Other, external
- ▶ Develop an institutional approach/strategy?

The Information Pyramid

Innovation!



**Increasing access to data =
Increasing research productivity?**

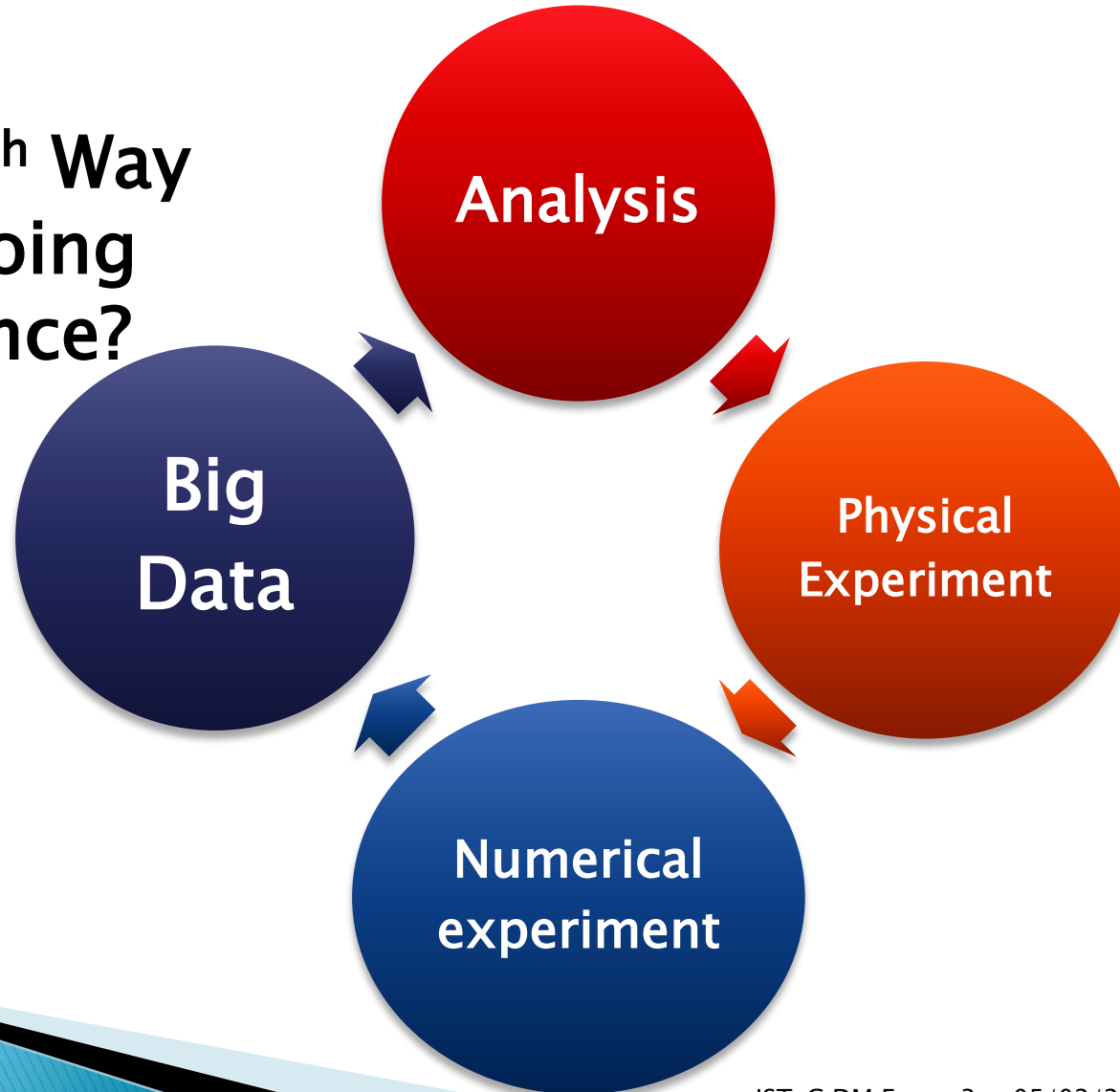


**IT Systems:
'Doing'**



Data are Key to Discovery

The 4th Way
of Doing
Science?

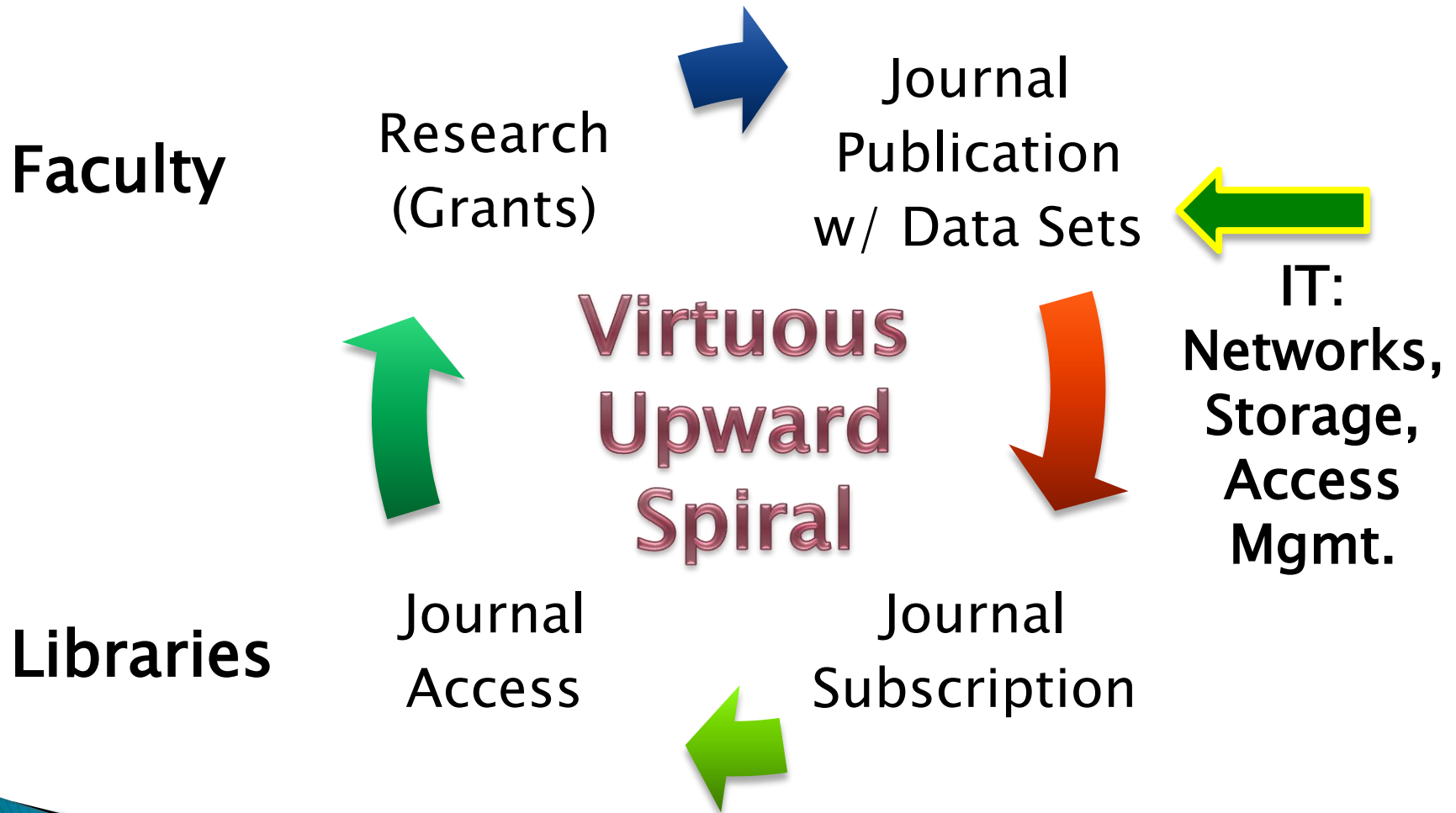


Open Government Initiative

- ▶ Agencies with > \$100m research funding must require federally-funded research to be made publically available, including data sets



Knowledge Ecosystem



Sharing Data/Information

What to Share



Locality

**Scholarly
Data Sets
(linked to
pubs)**



Remember

- ▶ Data sets, by themselves, are not “creative works,” and are therefore not copyrightable
 - But, publications and user’s manuals describing them are

One Faculty's Reaction to Sharing

- ▶ “I’d sooner show someone my underwear, than share my data sets with them!”



So, what's Wrong with this picture?



Yet Another

- ▶ Unfunded mandate for us!
 - Faculty
 - RA's
 - Students
 - IT
 - Libraries

Possible INCOMING!

NSF is SERIOUS About
Data Sharing



NSF Data Management Plans



**Required as
of Jan. 18,
2011**

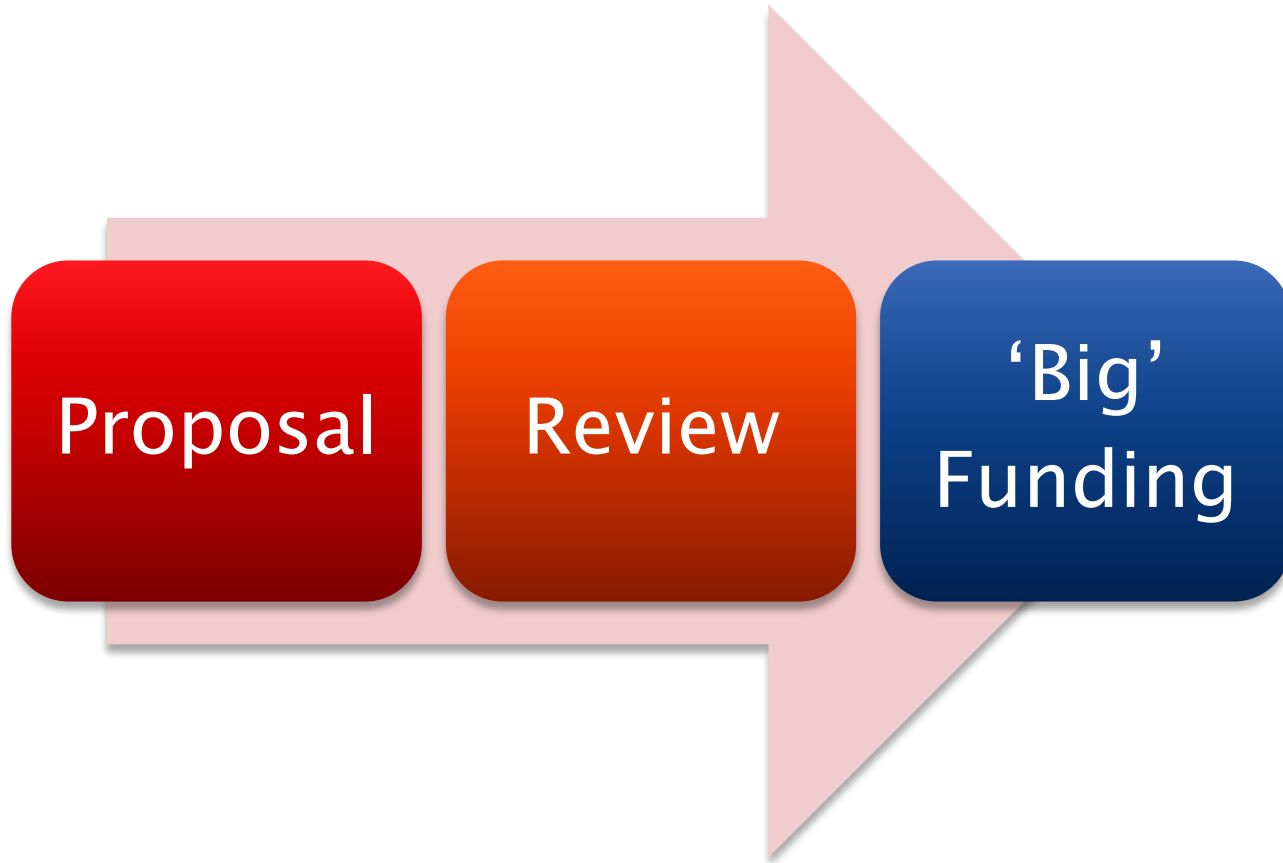
**Sharing data will
speed research
and economic
development!**

NSF and ‘Data Sharing?’

- ▶ ARL Meeting Oct. 2012
 - Myron Gutmann, Assistant Director, National Science Foundation, Directorate for Social, Behavioral & Economic Sciences



Current Research Funding – “My Data Are Strategic”



Future Research Funding? – “Prove Your Data Are Strategic”

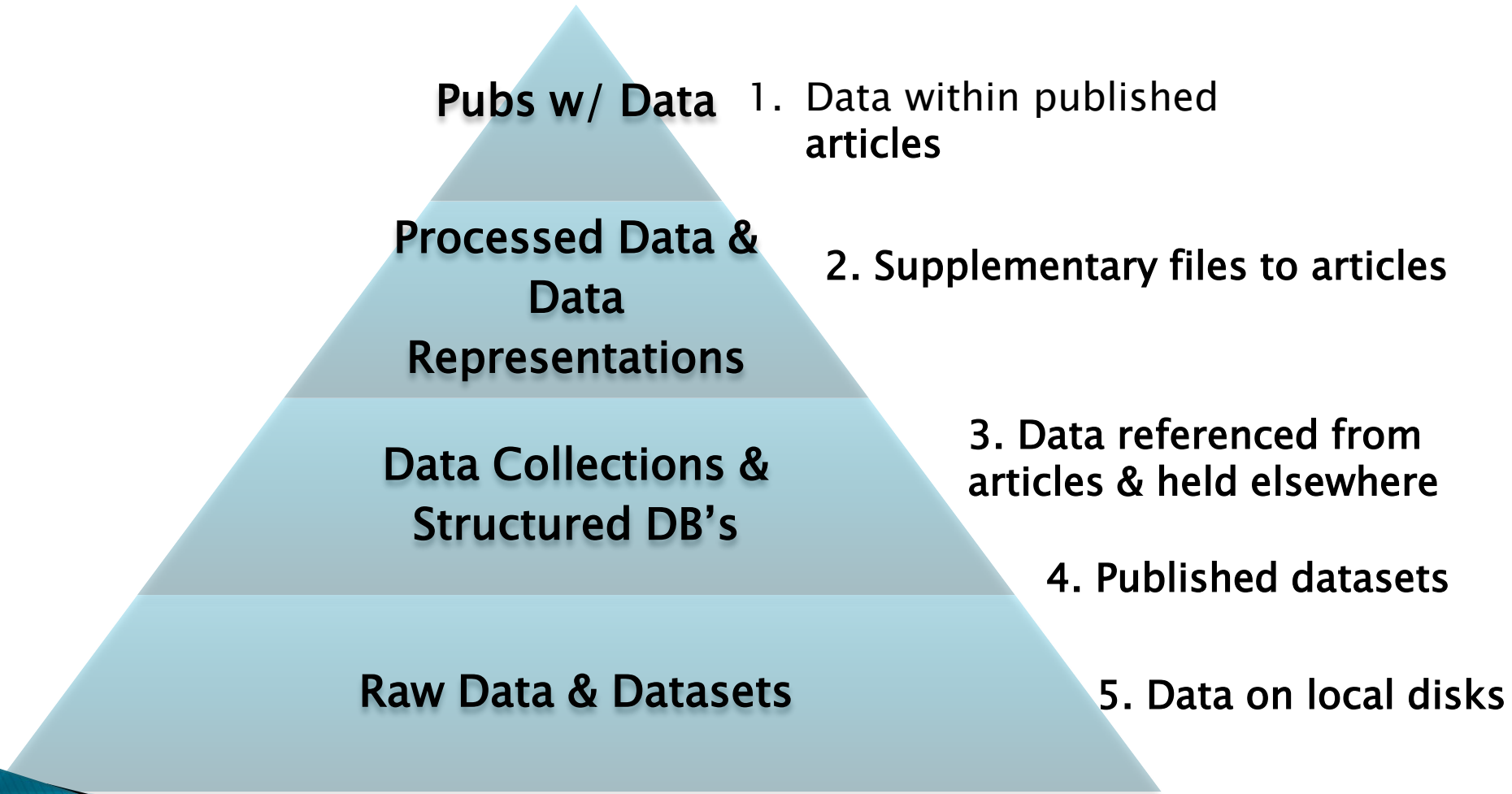




OUCH!

The Uncertain Data Landscape

(Ref: Eefke Smit, Opportunities for Data Exchange, Data Enters Scholarly Communication, Oct. 2011)



Data Sets Need to Be 'Citable'


- ▶ Discoverable
 - Crawled by Google and others
- ▶ Accessible
 - Robust infrastructure, accessibility
- ▶ Organized/searchable
 - Metadata
 - Catalogued
- ▶ Persistent
 - Persistent Digital Object Identifier (DOI)
 - Preserved and transcoded as needed

**> 25% Failure
Rate of URL's!!**

How Do I Share Scholarly Data?

1. Federal agency systems: NIH PubMed Central
 - NASA only other federal agency deploying a DR?
2. Disciplinary repositories
3. Local repositories: CSU's DigiTool
 - Data Management Plan templates
 - Discoverability, accessibility, & preservation
 - Usage stats
 - Linked to pubs
4. Local file share files from a data store
 - Granular access control via a common infrastructure, e.g. Globus

Two Key Questions

1. Where do I put my scholarly data?
 - Wherever it's free!
 - Locality? Can I move it around?
 - Repository: persistence/longevity
 - Globus-enabled file share
2. How do I expose my data?
 - PubMed Central, NASA repository
 - Disciplinary repository
 - Local repository (metadata) 

ARL's SHARE Initiative

- ▶ Association of Research Libraries (ARL)
SHARE initiative: local repositories
 - “Shared Access Research Ecosystem”
 - For preservation and access of scholarly works and data sets
 - Not for storage of working files
- ▶ CSU's DigiTool Digital Repository (DR)

[DCC Home](#) [Search / Browse](#) [Results](#) [Previous Searches](#) [My Account](#) [About](#) [Login](#) [End Session](#) [Help](#)

Guest

[Simple Search](#) [Advanced Search](#)

A word or phrase: Contains Exact Starts With

Search

Select collection: ▼

[Browse All Collections](#) > [Colorado State University, Fort Collins](#)



[CSU Departments and Schools](#)
(3708)
[Agricultural and Resource Economics](#)
[Animal Sciences](#)
[Anthropology](#) ...



[CSU Colleges](#) (2015)
[University Libraries](#)
[Veterinary Medicine and Biomedical Sciences](#)
[Warner College of Natural Resources](#)



[CSU Centers and Research Institutes](#) (2151)
[Center for Collaborative Conservation](#)
[Colorado Institute of Public Policy](#)
[Colorado Water Institute](#) ...



[CSU Theses and Dissertations](#)
(1847)
2011-
1980-2010
1950-1979 ...



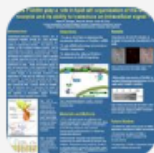
[CSU Archives and Special Collections](#) (37874)
[Agricultural and Natural Resources Archive](#)
[Manuscripts](#)
[Special Collections](#) ...



[CSU Conference Proceedings and Events](#) (44)
[Bridging the Gap Conference](#)
[CI Days: Cyberinfrastructure 2010 in the Rockies](#)
[Forest Biomass Conference](#)



[CSU Journal Publications](#) (28)
[Furthering Perspectives](#)
[Journal of Student Affairs](#)
[Journal of Undergraduate Research and Scholarly Excellence](#)



[CSU Student Research Projects](#) (43)
[Celebrate Undergraduate Research and Creativity \(CURC\) Showcase](#)



[Books](#) (3)

The Digital Collections of Colorado include publications, theses and dissertations, presentations, historical materials and other scholarly works in various formats.

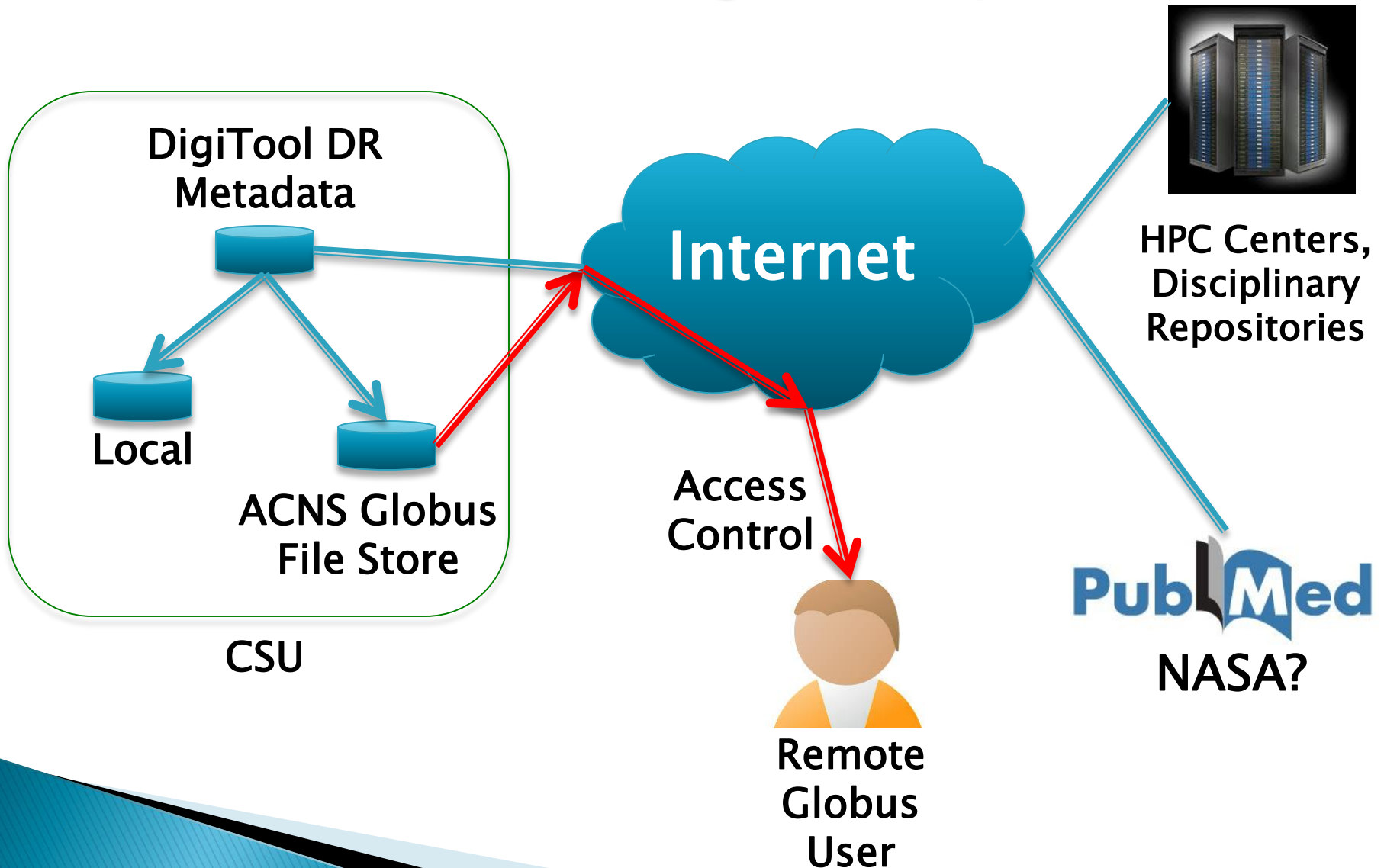
[Share](#) • [Bookmarkable URL](#) • [Statistics for this collection](#) [About the Digital Collections of Colorado](#) • [Contact Us](#) • [Home](#)

© Copyright 2013 [Colorado State University](#), the [University of Colorado](#), [Colorado School of Mines](#) and [Colorado Mesa University](#)

Another (Unattractive) Option



The Data Sharing Ecosystem

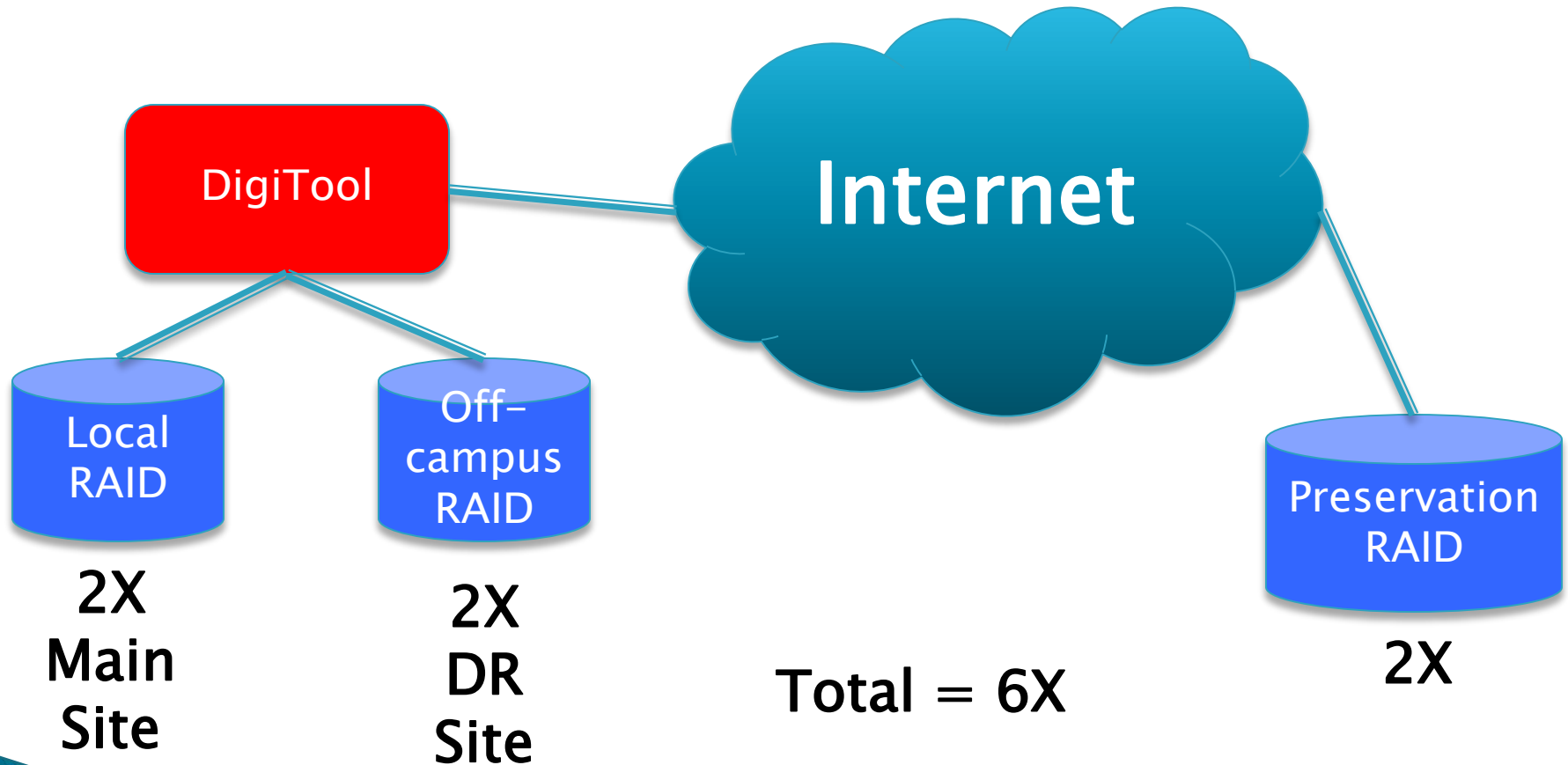


Approach

- ▶ If files stored in (any) digital repository
 - ~OK for metadata and discoverability
- ▶ If files stored elsewhere
 - Metadata should be stored on a digital repository for discoverability, accessibility, citability, and persistence; and “point” to the data
- ▶ If files stored locally
 - Can use Globus to manage access granularly
 - What about backup and preservation?

DigiTool Storage/Preservation

X = Storage Size



CSU's DigiTool DR Attributes

- ▶ NSF likes:
 - Managed jointly by Librarians and IT Professionals
 - Connected to ultrahigh-speed research network
 - The best metadata, standards,...
 - Discoverability (crawled by all services), accessibility
 - For preservation of scholarly data and pubs
- ▶ Limitations
 - Very limited access control (NSF likes this too)
 - Does not support structured data (e.g. databases)
 - Not for working data sets

Challenges

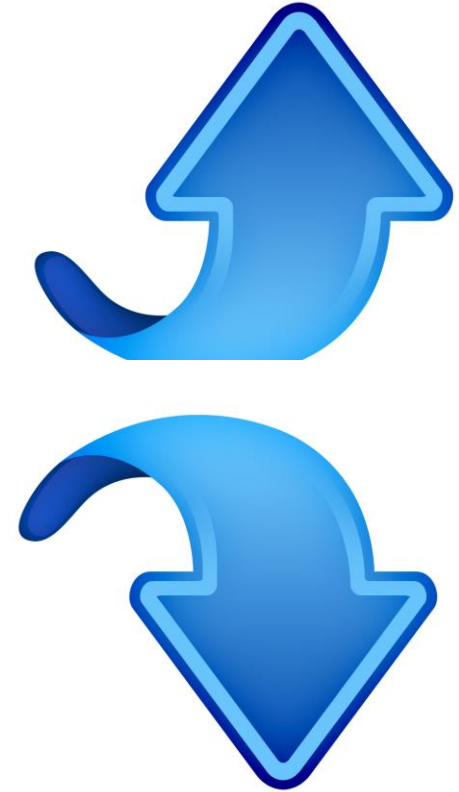
1. Cultural change
2. More work – ough!
3. Greater capital cost
 - The cost of additional storage required could be extreme

The Cost of Storage: Up or Down?

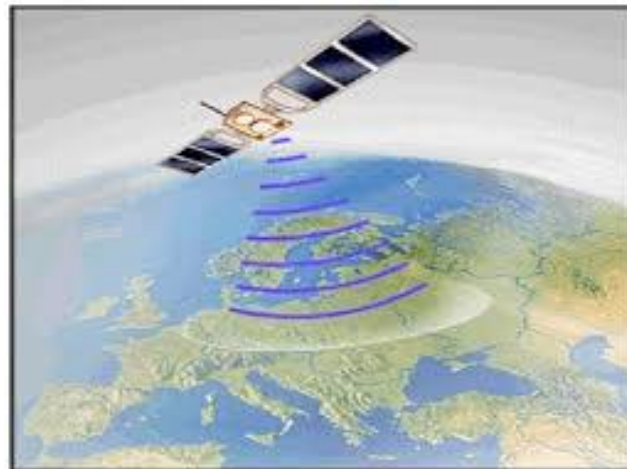
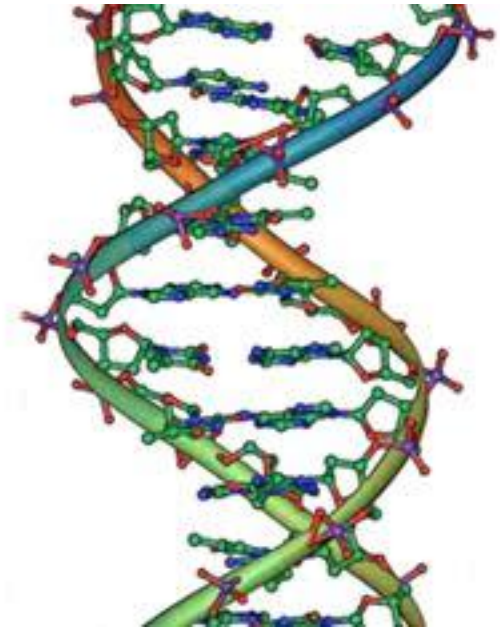
▶ **It depends!**

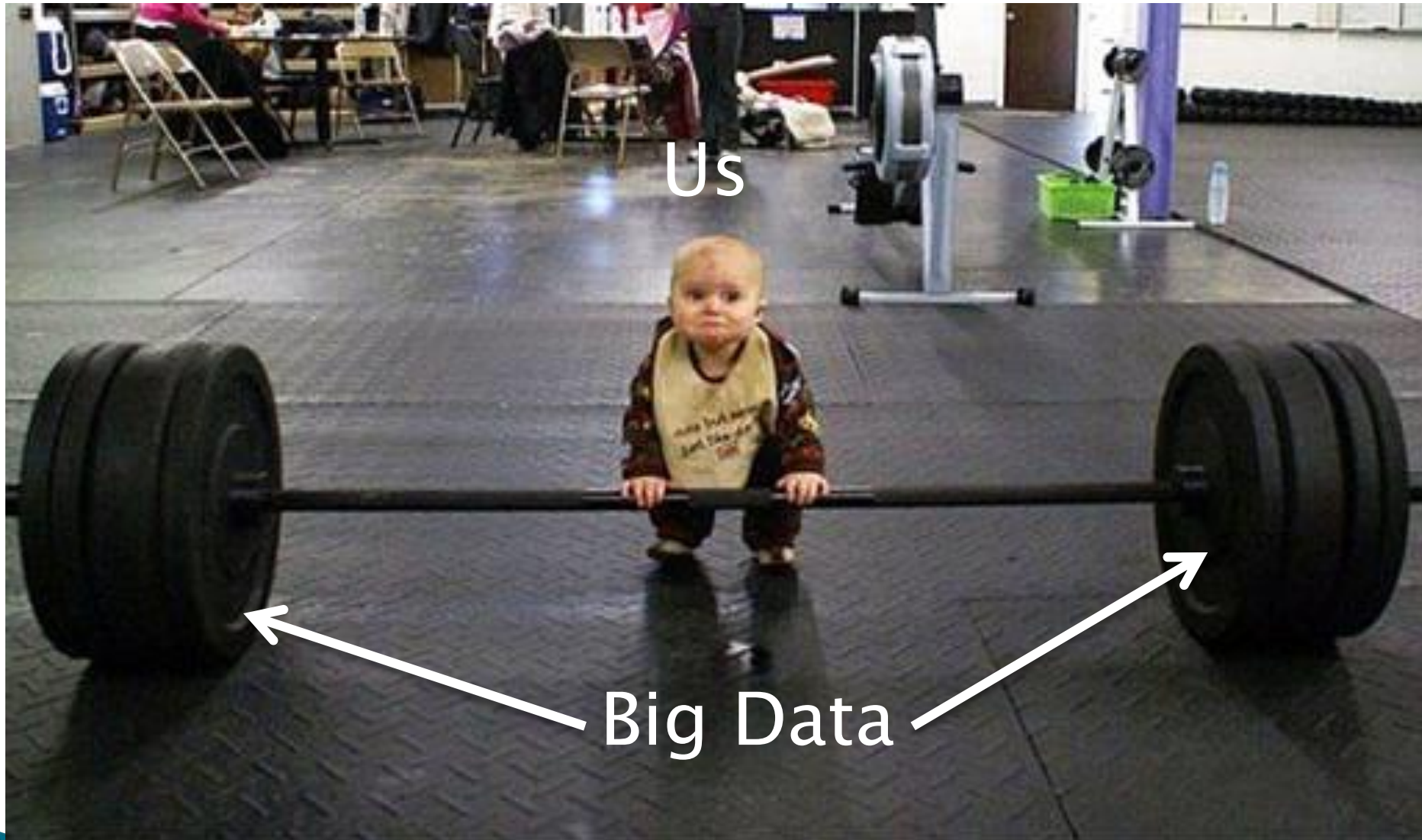
The Cost of Storage: Up or Down?

- ▶ SSD's are required for performance, but expensive
- ▶ Amazon announced ~80% cost reduction in storage – \$10/TB–mo!
 - We are exploring!!!



Big Data (Sets)





Us

Big Data

Funding Storage Costs in the DR

- ▶ If “incidental,” can be included as a direct cost in the grant
- ▶ Some storage for scholarly data sets “free” to CSU Faculty
 - First 100 GByte free, thereafter \$4k/TB
- ▶ Storage of pubs+ always “free” to CSU faculty and students
 - Assumes no “big” embedded data
- ▶ Exploring Amazon as an option **Game Changer!**

What's 'Big?' It Keeps Changing

- ▶ For a single file
 - Small: 1 GByte
 - Medium: 10–100 GBytes
 - Big: >100 GBytes
- ▶ For a set of files
 - 100X that of a single file?
- ▶ How many files is 'big?'



**1 full rack stores:
~350 Tbytes usable
& Costs ~\$150k
(~\$1,000/TB)**

IT Infrastructure = ACNS and Local IT



**Local
Storage**



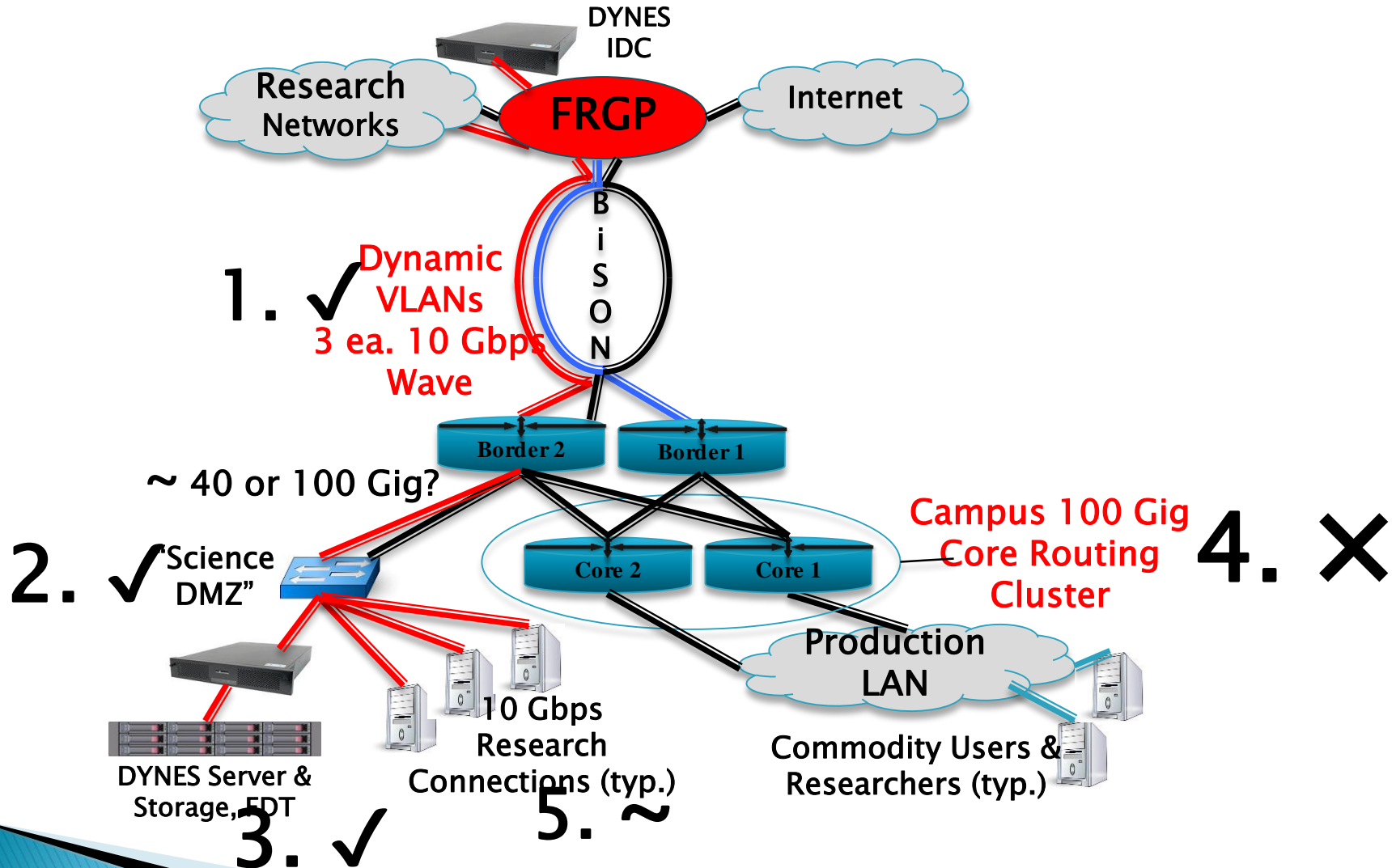
**HPC
& Other Devices**



**ACNS
Storage**

Ultrahigh-speed Research Network

3. ✓



LIBRARIES, IT, & Data Management

- ▶ Later in this forum, you will hear more from our consummate experts

An Emerging Numeric Identifier

- ▶ ORCID – the Organizational Researcher Contributor ID
- ▶ An unique numeric identifier to “Connect Research and Researchers”
- ▶ Makes publication data gathering much easier
- ▶ Complements the DOI

- ▶ See <http://orcid.org/>

Make Work as Easy as Possible



So much of what we call management consists of making it difficult for people to work.

-- Peter Drucker

Today's Desired Outcome

- ▶ Do we need a standing Data Management Committee?



Move in any direction, as long as its somewhat positive!



Finis



**Sometimes it is better to
journey hopefully, than to
arrive!!!**